UNIVERSIDADE FEDERAL DO PAMPA

LUCIANO MORAES DA LUZ BRUM

APLICAÇÃO DE TÉCNICAS DE BUSINESS INTELLIGENCE EM SISTEMAS DE APOIO À TOMADA DE DECISÃO DE PRODUTORES RURAIS

Bagé
2019

LUCIANO MORAES DA LUZ BRUM

APLICAÇÃO DE TÉCNICAS DE BUSINESS INTELLIGENCE EM SISTEMAS DE APOIO À TOMADA DE DECISÃO DE PRODUTORES RURAIS

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Computação Aplicada da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Mestre em Computação Aplicada.

Orientador: Sandro da Silva Camargo

Coorientador: Vinicius do Nascimento Lampert

Bagé
2019

Ficha catalográfica elaborada automaticamente com os dados fornecidos pelo(a) autor(a) através do Módulo de Biblioteca do Sistema GURI (Gestão Unificada de Recursos Institucionais).

B893a Brum, Luciano Moraes da Luz
Aplicação de técnicas de business intelligence em sistemas de apoio à tomada de decisão de produtores rurais / Luciano Moraes da Luz Brum.
201 f.: il.

Dissertação (Mestrado) - Universidade Federal do Pampa, Campus Bagé, MESTRADO EM COMPUTAÇÃO APLICADA, 2019.
"Orientação: Sandro da Silva Camargo".

1. Agronegócio. 2. Data warehouse. 3. Integração de dados. I. Camargo, Sandro (Orient.). II. Título.

APLICAÇÃO DE TÉCNICAS DE BUSINESS INTELLIGENCE EM SISTEMAS DE APOIO À TOMADA DE DECISÃO DE PRODUTORES RURAIS

Dissertação de mestrado apresentada ao Programa de Pós-graduação em Computação Aplicada da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Mestre em Computação Aplicada.

Área de concentração: Tecnologias para Produção Agropecuária

Dissertação de mestrado defendida e aprovada em: 18/03/2019

Banca examinadora:

Prof. Dr. Sandro da Silva Camargo
Orientador
UNIPAMPA

Prof. Dra. Ana Paula Lüdtke Ferreira
UNIPAMPA

Prof. Dr. Marcelo Benevenga Sarmento
URCAMP

Prof. Dr. Fabiano Nunes Vaz
UFSM

RESUMO

Através de parcerias entre instituições de ensino superior do município de Bagé e a Empresa Brasileira de Pesquisa Agropecuária, estão sendo desenvolvidos Sistemas de Apoio à Decisão que visam auxiliar o produtor rural na tomada de decisões na pecuária de corte. A dificuldade surge quando se torna necessário realizar análises e estabelecer relações entre as variáveis desses sistemas de forma automática. Da forma como estão sendo desenvolvidos, os sistemas podem acabar fornecendo diferentes versões de uma verdade da pecuária de corte, acarretando em maiores dificuldades. A proposta deste trabalho é integrar os dados de dois Sistemas de Apoio à Decisão, através de técnicas de *Business Intelligence*. O objetivo é permitir que produtores rurais possam realizar análises a partir da integração de dados de indicadores ambientais, econômicos e produtivos, e proporcionar meios para visualização integrada e simplificada destas informações. Para implementação da solução, foram utilizadas tecnologias gratuitas e de código aberto. Como resultados, foi elaborado um modelo dimensional esquema estrela, no qual foi modelada uma tabela fato com indicadores econômicos, produtivos e ambientais da pecuária de corte. Tais indicadores podem ser analisados no tempo, por localidade, faixas de área dos estabelecimentos rurais, entre outras dimensões. O processo de ETL foi realizado, permitindo a integração dos dados dos sistemas em um único repositório, com uma versão única da verdade das diferentes faces da pecuária de corte. A ferramenta de visualização dos dados Saiku Analytics, integrada na suíte Pentaho, permite a análise das informações no estilo OLAP, por meio de tabelas e gráficos, além de permitir exportar os dados em relatórios e planilhas eletrônicas. A solução pode auxiliar na resposta de questões como “a média de produtividade de estabelecimentos rurais por localidade”, “custos totais de estabelecimentos rurais por faixas de área”, “variação dos custos e indicadores zootécnicos por ano”, entre outros. Foram detectadas diversas dificuldades, também apresentadas na revisão da literatura, nos processos que envolvem integração de dados de fontes heterogêneas. Destacaram-se problemas de granularidade dos dados, sistemas desenvolvidos sem o viés de integração, dificuldades para o entendimento claro de todos os aspectos do domínio do problema e ausência de documentação dos sistemas legados.

Palavras-Chave: Agronegócio. Data warehouse. Integração de dados.

ABSTRACT

Through partnerships between higher education institutions in the city of Bagé and the Brazilian Agricultural Research Corporation, Decision Support Systems are being developed that aim to assist the rural producer in making decisions on beef cattle. The difficulty arises when it becomes necessary to perform analyzes and establish relations between the variables of these systems automatically. In the way they are being developed, systems may end up supplying different versions of a true breeding herd, leading to greater difficulties. The purpose of this paper is to integrate data from two Decision Support Systems, using Business Intelligence techniques. The objective is to enable farmers to carry out analyzes based on the integration of environmental, economic and productive indicators data, and to provide means for integrated and simplified visualization of this information. For the implementation of the solution, only free and open source technologies were used. As results, a star schema dimensional model was elaborated, in which one fact table was modeled with economic, productive and environmental indicators of the cattle ranching. These indicators can be analyzed in time, by location, area ranges of properties, among other dimensions. The ETL process was carried out, allowing the integration of the data of the systems into a single repository, with a unique version of the truth of the different faces of the cattle ranching. The Saiku Analytics data visualization tool, integrated in the Pentaho suite, allows the analysis of OLAP-style information through tables and graphs, as well as allowing the export of data in reports and in spreadsheets. The solution can help answer questions such as "the average productivity of farms by location", "total costs of properties by area ranges", "variation costs and zootechnical indicators per year", among others. Several issues were detected, also presented in the literature review, in the processes that involve integration of data from heterogeneous sources. Problems of data granularity, systems developed without the integration bias, difficulties for a clear understanding of all aspects of the problem domain and no available legacy systems documentation were highlighted.

Keywords: Agribusiness. Data warehouse. Data integration.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

3NF - Terceira Norma Formal

BI - *Business Intelligence*

BSD - *Berkeley Software Distribution*

CE - *Community Edition*

CIF - *Corporate Information Factory*

CRM - *Customer Relationship Management*

DM - *Data Mart*

DRLS - *Dynamic Row Level Security*

DW - *Data Warehouse*

EDW - *Enterprise Data Warehouse*

EE - *Enterprise Edition*

EIS - *Environmental Information System*

EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária

ER - Entidade-Relacionamento

ERP - *Enterprise Resource Planning*

ETL - *Extract, Transform and Load*

FGC - Ferramenta de Gestão de Custos

HOLAP - *Hybrid Online Analytical Processing*

HTML - *Hypertext Markup Language*

GPL - *General Public License*

IBGE - Instituto Brasileiro de Geografia e Estatística

IDC - *International Data Corporation*

IES - Instituições de Ensino Superior

IFSUL - Instituto Federal Sul-Rio-Grandense

KPI - *Key-Performance Indicator*

LGPL - *Lesser General Public License*

LS - *Livestock Sustainability*

MAPA - Ministério da Agricultura, Pecuária e Abastecimento

MDX - *Multidimensional Expressions*

MOLAP - *Multidimensional Online Analytical Processing*

MPL - *Mozilla Public License*

ODS - *Operational Data Store*

OLAP - *Online Analytical Processing*

OLTP - *Online Transaction Processing*

PBAS - *Pentaho Business Analytics Server*

PDF - *Portable Document Format*

PDI - *Pentaho Data Integration*

PIB - Produto Interno Bruto

PPGCAP - Programa de Pós-Graduação em Computação Aplicada

PSW - *Pentaho Schema Workbench*

RAM - *Random Access Memory*

RF - Requisitos Funcionais

RNF - Requisitos Não-Funcionais

ROLAP - *Relational Online Analytical Processing*

RS - Rio Grande do Sul

SAD - Sistema de Apoio à Decisão

SCD - *Slowly Chaging Dimension*

SCM - *Supply Chain Management*

SDW - *Spatial Data Warehouse*

SGBD - Sistema de Gerenciamento de Banco de Dados

SMTP - *Simple Mail Transfer Protocol*

SNPA - Sistema Nacional de Pesquisa Agropecuária

SOLAP - *Spatial Online Analytical Processing*

SQL - *Structured Query Language*

TAM - *Technology Acceptance Model*

TI - Tecnologia da Informação

UFV - Universidade Federal de Viçosa

UNIPAMPA - Universidade Federal do Pampa

URCAMP - Universidade da Região da Campanha

WEKA - *Waikato Environment for Knowledge Analysis*

XML - *Extensible Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Atualmente, a tecnologia faz parte do cotidiano da maioria das pessoas, assim como está cada vez mais incorporada em empresas, indústrias e organizações. Em 2002, o número de micro e pequenas empresas brasileiras informatizadas já era de aproximadamente 50% (LUNARDI; DOLCI; MAÇADA, 2010). Um estudo realizado em 2010 pelo Instituto Brasileiro de Geografia e Estatística (IBGE) apontava que 80,8% das empresas brasileiras já utilizavam computador e 76,9% já utilizavam a internet (IBGE, 2012).

O aumento da adoção de tecnologias, principalmente das que capturam dados, acarretou no crescimento do volume de dados gerados diariamente. Vieira (2012) afirma que o aumento do número de dispositivos para captação de dados, o aumento da capacidade de armazenamento e da velocidade de transmissão nas redes são alguns dos principais fatores que explicam o crescimento no volume de dados gerados e coletados. De acordo com a *International Data Corporation*[1], em 2013, o volume de dados produzidos somou 4,4 *zettabytes* e estima-se que, em 2020, o volume de dados que será utilizado em todo o mundo alcançará 40 *zettabytes*.

Além do suporte dos dispositivos tecnológicos atuais para coleta, processamento e armazenamento dos dados, é necessário também extrair informações úteis destes dados, visando dar suporte para a tomada de decisão nos mais diversos ambientes de negócio. Em outras palavras, agregar valor a todos estes dados disponíveis. Para isso, inicialmente é necessário sistematizar, organizar e manter um banco de dados adequado às necessidades da empresa. Se devidamente analisados, os dados poderão ser utilizados para reduzir os riscos e incertezas relacionadas aos processos decisórios, visto que uma decisão errônea pode comprometer o futuro de uma organização ou empresa.

Em um contexto semelhante das demais organizações, produtores rurais e profissionais dos setores da agricultura e pecuária também necessitam de tecnologias para subsidiar de uma forma mais adequada suas decisões. Porém, ao contrário da alta adesão à tecnologia pelas empresas urbanas, a realidade que o Brasil enfrentava em 2006 era o baixo nível tecnológico dos estabelecimentos agropecuários.

Os dados do Censo Agropecuário realizado pelo IBGE em 2006 e 2017 (dados preliminares), conforme a Tabela 1, apresentam o número e percentual de estabelecimentos agropecuários que possuíam computadores e acesso à internet:

Tabela 1 - Estabelecimentos agropecuários com computador e acesso à internet.

Eletrodoméstico utilizado	Variável			
	Número de estabelecimentos agropecuários (Unidades)		Percentual dos estabelecimentos agropecuários (%)	
	2006	2017	2006	2017
Computador	183.623	Não Informado	3,55	Não Informado
Acesso à internet	75.407	1.425.323	1,46	28,10

Fonte: Instituto Brasileiro de Geografia e Estatística (2017).

Mendes, Buainain e Fasiaben (2014) afirmam que o Sudeste e o Sul do Brasil são as regiões com maior adoção de Tecnologias da Informação (TI) por estabelecimentos agropecuários, considerando-se os dados do censo agropecuário de 2006 do IBGE. Mais recentemente, de acordo com o censo agropecuário de 2017, no Sul e Sudeste do Brasil somam-se 51,30% do total de estabelecimentos que fazem uso de internet no país (14,42% do total de estabelecimentos rurais do Brasil). Estas mesmas regiões também tiveram maior participação no Produto Interno Bruto (PIB) brasileiro em 2015 (IBGE, [2016?]). Portanto, a presença das tecnologias e do acesso à informação nos estabelecimentos rurais tem a sua importância, em parte, demonstrada nos indicadores econômicos do país.

Visando auxiliar produtores rurais e demais clientes nos processos decisórios, surge uma iniciativa conjunta para o desenvolvimento de sistemas para este fim. Estão sendo desenvolvidos diferentes sistemas para o suporte à tomada de decisão de produtores rurais pela Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) e Instituições de Ensino Superior (IES) do município de Bagé. Esses sistemas, de modo geral, visam auxiliar produtores rurais, gerentes, consultores e pesquisadores no entendimento dos indicadores envolvidos na atividade pecuária e no processo de tomada de decisão (EMBRAPA, [2015?]). Este auxílio será subsidiado através da coleta de dados que sejam preponderantes e influentes nos diferentes tipos de decisões envolvidas em um sistema produtivo da pecuária de corte, como organização dos diferentes tipos de custos da propriedade, análise de indicadores que permitem otimizar a taxa de desmame, indicadores presentes em sistemas produtivos com produtividade por área mais alta, entre outros. Abaixo são detalhados dois dos sistemas:

- Ferramenta de Gestão de Custos (FGC): Solução web voltada para produtores rurais de pecuária de corte. Este serviço web deve permitir sua utilização inserindo poucas informações sobre os custos da propriedade e posteriormente ir migrando para os demais níveis à medida que o usuário for qualificando seu processo de registro de dados e de gerenciamento de sua propriedade rural. A solução também deve gerar relatórios e gráficos simplificados utilizando as informações inseridas pelos usuários, além de permitir realizar análises comparativas entre produtores.
- *Livestock Sustainability* (LS): A ideia do aplicativo é fornecer ao produtor uma forma de estimar a emissão da pegada de carbono do sistema de produção, utilizando variáveis que sejam conhecidas por ele. Será coletado um conjunto de dados que permitirá realizar mineração de dados, de forma a tentar identificar quais variáveis mais influenciam na emissão, de forma a subsidiar alternativas para sua redução. É um sistema de informação que estima a pegada de carbono e sugere alternativas de redução. Este sistema engloba o cálculo de indicadores produtivos a partir do modelo proposto pela tese de Lampert (2010).

É previsto que estes sistemas sejam utilizados por produtores rurais e empresas de consultoria. Com o crescimento da utilização destes sistemas, serão gerados e coletados grandes volume de dados e informações sobre os estabelecimentos rurais cadastradas nos sistemas, além dos indicadores e variáveis sobre os estabelecimentos rurais. Também é prevista a necessidade de integração e análises dos dados de todos os sistemas, por pesquisadores e produtores rurais. São necessários métodos efetivos de coleta, integração, visualização e análises destes dados. O desafio desta proposta é construir um ambiente que permita realizar a integração dos dados destas fontes de dados, que utilizam diferentes tecnologias e formas de armazenamento das informações, organizá-los e disponibilizá-los de forma que seja possível visualizá-los, realizar análises e auxiliar nas tomadas de decisão com base nas informações dos sistemas FGC e LS.

Um conceito que relaciona as tomadas de decisão, grandes volumes de dados, integração de dados de diferentes fontes e extração de informações úteis destes dados é o de *Business Intelligence* (BI). O BI, como um conjunto de técnicas e métodos, permite que os dados operacionais de diferentes repositórios possam ser integrados para serem explorados de forma estratégica, visando agregar novos conhecimentos ou descoberta de padrões que estavam ocultos nos dados. Estes

novos conhecimentos podem subsidiar e otimizar os processos decisórios, reduzindo riscos e incertezas e possibilitando prever situações futuras em menos tempo e com maior precisão.

Apresentados alguns conceitos e os desafios futuros, a proposta deste trabalho é aplicar métodos de integração de dados nos dois sistemas de apoio à decisão para a pecuária de corte, em desenvolvimento, com o objetivo de viabilizar a análise e visualização das informações de forma integrada, subsidiando assim, os processos decisórios.

1.1 Justificativa

Cada um dos sistemas em desenvolvimento, LS e FGC, possui um banco de dados relacional com a estrutura necessária para armazenar as informações dos produtores rurais, sejam elas virtuais (oriundas de simulações), ou reais. Neste cenário, se os dados de ambos os sistemas forem necessários para relacionamentos e análises, estes deverão ser extraídos de cada sistema e posteriormente integrados (em um banco de dados organizado em esquemas relacionais, por exemplo) para serem utilizados no processo de tomada de decisão ou obtenção de conhecimento. Esta pode não ser a melhor alternativa para a extração de informações relevantes para a tomada de decisão porque pode haver redundância de informações entre os sistemas. Processos manuais envolvendo dados redundantes são mais suscetíveis a falhas (WIJAYA; PUDJOATMODJO, 2015). Realizar processos analíticos nas próprias bases de dados pode requerer um alto poder de processamento, pois depende do volume de dados envolvido e da complexidade das consultas. Isso pode comprometer o desempenho dos sistemas, tanto para os usuários das bases de dados para operações de banco de dados operacionais (inserção, remoção, atualização e seleção) como para o próprio usuário que deseja realizar o processo analítico nos dados (HAN; KAMBER; PEI, 2011). Ainda, os bancos de dados relacionais são otimizados para o processamento de transações. Processos de análises de dados e extração de conhecimento não são adequados para este tipo de sistemas de banco de dados (HAN; KAMBER; PEI, 2011). O conteúdo dos bancos de dados relacionais é muito detalhado para ser imediatamente utilizado em processos de análise (HAN; KAMBER; PEI, 2011). Muitas vezes será necessário um resumo dos dados para

auxiliar nas decisões (WIJAYA; PUDJOATMODJO, 2015). Por fim, os dados em bancos de dados relacionais podem estar incompletos, incorretos ou inconsistentes.

Visando facilitar o processo de análise, integração e visualização dos dados destes sistemas, surge o conceito de *Data Warehouse* (DW). Inmon (2002) afirma que um DW se difere de outros sistemas repositórios de dados devido a 4 conceitos-chave: orientação a assunto, integração, variação no tempo e não-volatilidade dos dados. O DW é uma estrutura que armazena dados analíticos preponderantes para a solução de BI e seus principais objetivos são armazenar, integrar e disponibilizar informações para oferecer suporte ao processo decisório na gerência de negócio (JUNIOR, 2004).

O DW mostra-se como parte da solução das dificuldades apresentadas, permitindo a criação de um repositório que contenha dados de fontes heterogêneas de forma estruturada e otimizada para processos analíticos. O processo que permite que o DW possua todas as informações preponderantes na tomada de decisão é o ETL (*Extract, Transform and Load* – Extração, Transformação e Carga*)*, que é dividido nas fases de extração dos dados dos repositórios de dados, transformação dos dados em um formato apropriado para análise e carga dos dados no DW, sendo este um processo de integração dos dados. Um DW também permite o uso de ferramentas OLAP (*Online Analytical Processing*), que oferecem suporte à análise e visualização eficiente de informações. Técnicas de descoberta de conhecimento também podem ser utilizadas através de mineração de dados, que é facilitada com a aplicação dos conceitos de BI.

A modelagem do DW, a construção e execução do processo de ETL e utilização de ferramentas OLAP fazem parte da solução de BI proposta neste trabalho, além da realização de uma avaliação crítica da solução desenvolvida.

As principais contribuições esperadas com este trabalho são as seguintes:

- Contribuições Tecnológicas: Um ambiente que permite a realização de análises históricas, comparações entre informações e indicadores de dois sistemas em desenvolvimento, de forma a subsidiar os processos decisórios de técnicos e produtores rurais.
- Contribuições Científicas: Elaboração de um instrumento que permite a integração de dados da pecuária de corte de fontes heterogêneas.

1.2 Objetivos

O objetivo geral deste trabalho é realizar a integração de dados de dois sistemas de apoio à decisão heterogêneos, de forma que seja possível realizar a análise e visualização destas informações de forma integrada, para subsidiar os processos decisórios de produtores e consultores rurais da pecuária de corte.

Os objetivos específicos são os seguintes:

- Definir os requisitos funcionais e não-funcionais da solução proposta.
- Definir a arquitetura do DW e realizar a modelagem e implementação.
- Viabilizar a integração dos dados dos sistemas e disponibilizar informações de apoio à decisão.
- Avaliar as técnicas de DW/BI como recurso para integração, análise e visualização de dados da pecuária de corte, específicos deste estudo.

1.3 Organização do texto

O restante do trabalho está organizado da seguinte forma: No capítulo 2 é abordada a revisão da literatura e o estado da arte. No capítulo 3 é apresentada a caracterização da pesquisa e a proposta metodológica para atingir os objetivos do projeto. No capítulo 4 é apresentada a execução da proposta metodológica. No capítulo 5 são apresentados os resultados e discussões. No final, o capítulo 6 apresenta as considerações finais e perspectivas de trabalhos futuros.

2 REVISÃO DE LITERATURA

Neste capítulo é apresentada a revisão bibliográfica dos conceitos teóricos e tecnologias que subsidiaram o desenvolvimento deste trabalho. A seção 2.1 aborda conceitos e dificuldades da tomada de decisão no agronegócio. A seção 2.2 trata sobre o surgimento e a importância do Sistema de Apoio à Decisão (SAD) no contexto da agropecuária. A seção 2.3 apresenta as teorias relacionadas ao tema *Business Intelligence* e cada uma das etapas e tecnologias envolvidas neste processo. A seção 2.4 apresenta as principais tecnologias utilizadas em soluções de BI. Por fim, a seção 2.5 aborda o estado da arte do tema pesquisado, citando diferentes trabalhos que relacionem as técnicas de BI ou DW com a solução de problemas na agricultura, pecuária e áreas diretamente relacionadas com as mesmas.

2.1 Tomada de Decisão no Agronegócio

Para Chiavenatto (2003, p. 348), a decisão “é o processo de análise e escolha entre as alternativas disponíveis de cursos de ação que a pessoa deverá seguir”. Para Oliveira (2004), a tomada de decisão é a conversão da informação disponível em ações. Para este trabalho, julga-se mais adequada a definição defendida por Oliveira, por considerar a importância da informação, que são os dados tratados e organizados de forma adequada, e seu uso para conversão em ações. Tal definição se aproxima dos objetivos do BI.

Bazerman e Moore (2010) afirmam que o processo de tomar uma decisão se divide nas seguintes etapas: definição do problema, identificação de critérios, ponderação de critérios, geração de alternativas, classificação das alternativas seguindo algum critério e identificação da solução ideal.

O produtor rural constantemente realiza a tomada de decisões com o objetivo de escolher corretamente uma dentre várias alternativas na produção, visando a que melhor se adequar em seu estabelecimento para melhorar sua eficiência produtiva e financeira. Porém, Simon (1945 *apud* LAMPERT, 2014, p. 132) afirma que o ser humano possui racionalidade limitada e é influenciado por seus valores e extensão de conhecimento, o que impede a racionalidade objetiva da melhor escolha. No modelo de racionalidade limitada, o que o indivíduo faz é tomar decisões satisfatórias. Na administração dos estabeleciementos rurais, não é possível obter todas as

possibilidades de ação e mensurar todos fatores envolvidos devido à limitações de tempo, custo e conhecimento sobre todas estas informações (LAMPERT, 2014).

Para subsidiar os processos decisórios de forma adequada, é necessário que estejam disponíveis dados e informações sobre o processo em questão. Hoje, a informação é considerada um bem das organizações e está relacionada ao poder de competitividade das empresas no mercado. Com as constantes variações no cenário mundial, considerando-se questões mercadológicas, políticas e ambientais, emergem diversas consequências e podem afetar diretamente os setores produtivos da pecuária. Portanto, além do conhecimento necessário para produzir, é essencial que os produtores rurais tenham ciência de como os fatores externos podem exercer impactos no sistema produtivo. Apesar de não poder controlá-los, o produtor deve conhecê-los para ajustar as suas decisões.

Três trabalhos apresentam alguns dos fatores externos que afetam a atividade no campo. Machado, Oliveira e Schnorrenberger (2006) mostram que a atividade agropecuária possui riscos adicionais em comparação com outros negócios. Tais riscos estão associados a fatores externos, como: sazonalidade da produção, observância de ciclos, variações climáticas, perecibilidade dos produtos, necessidades próprias de processamento e transformação das matérias-primas, influência de fatores biológicos, indicadores sociais, ambientais e econômicos (LAMPERT *et al*., 2017), entre outros. Com relação aos aspectos mercadológicos e políticos, são apontados os seguintes fatores externos (SANTOS; MARION; SEGATTI, 2002): preços dos produtos, existência de mercado para o produto, políticas de crédito e financiamentos, transporte e disponibilidade de mão de obra na região.

No cenário ideal, portanto, caberia aos produtores administrarem seus estabelecimentos rurais de forma que fosse possível a obtenção e geração de dados e informações sobre seus sistemas produtivos, de uma forma geral. Em um segundo momento, seria necessário realizar processos analíticos, estatísticas, integração e cruzamento destas informações com a realidade externa à propriedade. Por fim, dadas as informações disponíveis e o conhecimento inerente ao produtor, as decisões e ações seriam tomadas, visando a otimização dos processos produtivos, de forma que a eficiência e eficácia fossem maximizadas e os riscos inerentes a atividade, minimizados, de forma a garantir a sua permanência no campo. Cabe ressaltar que a

gestão econômica da propriedade, para o pecuarista de corte brasileiro, é a principal preocupação na atualidade (EMBRAPA, 2018).

Um desafio apontando por Mayer e Werlang (2016) para os produtores rurais é o de não apenas gerar informações, mas também apoiar os processos decisórios com base nelas. Em outras palavras, gerar informações úteis na tomada de decisão. Chaves *et al*. (2010, p. 5) complementam:

> [...] a limitação organizacional e estrutural inerentes ao ambiente do empreendedor rural dificulta a tarefa de gerar informações gerenciais que permitam a tomada de decisão, com base em dados consistentes e reais. Dessa forma, o processo decisório no meio rural é muito mais baseado na criatividade, julgamento, intuição e experiência do administrador do que em métodos analíticos e quantitativos com suporte científico, não considerando praticamente nenhuma estatística dos dados disponíveis e muito menos a forma ideal para maximizar o lucro [...]

Essas questões são parcialmente corroboradas pelo trabalho de Hofer *et al.* (2011), no qual os autores concluem que a maioria dos pequenos e médios produtores da região oeste do Paraná gerencia os processos produtivos e atividades de forma informal e através de anotações escritas em papel. Essa realidade dificulta ainda mais a execução dos processos analíticos, devido ao trabalho adicional de inserir essas informações em um computador. Por estarem anotadas em papel, a probabilidade de erro nos dados, devido ao método de coleta, é mais alta, podendo comprometer decisões baseadas nestes dados.

Uma informação interessante apresentada nos dados do censo agropecuário de 2006 mostra que a adoção de tecnologias como a internet é maior entre produtores com nível de ensino superior. Em uma pesquisa mais recente com os dados preliminares do censo agropecuário de 2017, é possível observar que o percentual de estabelecimentos rurais com acesso à internet aumentou de 1,87% para 28,10% no período 2006-2017 (aumento de 1790%). Observou-se um grande aumento na adesão de tecnologias de acesso à informação por produtores rurais neste período (IBGE, 2017). Isso demonstra o potencial da utilização de tecnologias de suporte à decisão por estes.

2.2 Sistemas de Apoio à Decisão

Dadas as dificuldades apresentadas nas seções anteriores, constata-se a necessidade de tecnologias que, além de possuírem facilidade de uso, acessibilidade

e baixo custo, apresentem para o produtor as informações necessárias para apoiar e otimizar os processos decisórios e, com isso, revelando aos mesmos a utilidade da adoção de tecnologias nos sistemas produtivos. Neste contexto, surgem os SAD.

Um SAD é uma área de Sistemas de Informação que se concentra no apoio e na melhoria da tomada de decisões gerenciais (ARNOTT; PERVAN, 2008). Para Keen e Scott-Morton (1978, *apud* HUNG *et al*., 2007, p. 2093), estes sistemas oferecem acesso à informação, análise de modelos e ferramentas de suporte.

Um SAD deve possibilitar a geração de maior lucratividade, menores custos e melhoria dos processos relacionados aos produtos e serviços (STAIR; REYNOLDS, 2011). No caso do setor agropecuário, por exemplo, isso seria possível por meio da coleta e uso conjunto dos dados dos sistemas produtivos, informações externas e modelos de simulação através de um sistema de informação. É conveniente ressaltar que o objetivo do SAD não é substituir o ser humano no processo de tomada de decisão. O SAD deve permitir que os dados sejam manipulados para melhor subsidiar as decisões.

As principais características de um SAD são: capacidade de manipular grandes volumes de dados; obter e processar dados de fontes diferentes; proporcionar flexibilidade de relatórios e de apresentação; possuir orientação textual e gráfica; executar análises e comparações complexas; oferecer suporte às abordagens de otimização e satisfação; executar análises de simulações e baseada em metas (STAIR; REYNOLDS, 2011). Os principais componentes de um SAD são: banco de dados; software de um SAD; interface de usuário (LAUDON; LAUDON, 2004).

HUNG *et al.* (2007) afirmam que o uso dos SAD gera decisões de maior qualidade. Porém, mensurá-las, assim como o sucesso dos SAD, é algo não trivial. Os mesmos autores apresentam um resumo de 18 estudos sobre a mensuração das medidas de sucesso de um SAD, que visam atingir a satisfação do usuário e performance das decisões.

Apesar da proposta do uso de SAD na agropecuária ser interessante, existem diversos desafios que devem ser superados para consolidar e ampliar a sua utilização por produtores rurais. Mesmo com os grandes avanços técnicos e tecnológicos nos SAD, em questões de armazenamento, desempenho, funcionalidades e projeto das interfaces, poucas melhorias tem sido feitas na eficácia e na extensão de uso destes sistemas (BEYNON; RASMEQUAN; RUSS, 2002).

Abaixo, alguns fatores apontados como preponderantes no baixo nível de adoção destes sistemas no meio agropecuário para tomada de decisões (EASTWOOD; CHAPMAN; PAINE, 2012; KERR, 2004; LINDBLOM *et al.*, 2017; ROSSI *et al*., 2014; VAN MEENSEL *et al*., 2012):

- SAD existentes são baseados no que os cientistas e desenvolvedores de sistemas consideram como conhecimento necessário que deve ser implementado no apoio a decisão. A realidade é que estes falham em capturar o conhecimento tácito e as necessidades práticas dos produtores rurais.
- Problema de percepção dos desenvolvedores e cientistas da complexidade de uso do sistema.
- Falta de confiança dos usuários em soluções baseadas em SAD.
- Nível de conhecimento dos usuários, muitas vezes associado à escolaridade.
- Design da interface de usuário pobre.
- Requerimentos de entrada de dados tediosas.
- Baixa adaptação à situação dos estabelecimentos rurais.
- Atualizações de informações infrequentes.
- Falta de incentivo para aprender e adotar novas práticas.
- Insegurança de substituir consultores.

2.3 *Business Intelligence*

Muitos autores utilizam definições diferentes para o termo *Business Intelligence*, conforme também afirmado e apresentado por Botelho e Razzolini Filho (2014). Foi feita uma investigação na literatura para esclarecer e apresentar algumas das mais utilizadas.

Para Wu, Barash e Bartolini (2007) BI é um termo que descreve um conjunto de tecnologias e aplicações que são utilizadas para coletar, prover acesso e analisar dados e informações sobre uma determinada empresa ou organização, visando auxiliá-los no processo de tomada de decisão. Para Turban, Sharda e Delen (2011), BI possui uma definição mais ampla e genérica: “é um termo "guarda–chuva" que inclui arquiteturas, ferramentas, banco de dados, aplicações e metodologias”. Chaudhuri, Dayal e Narasayya (2011) afirmam que BI é um conjunto de tecnologias de suporte à decisão para empresas que visam auxiliar profissionais do conhecimento para uma tomada de decisão mais rápida e precisa. Por fim, para Díaz (2012) BI é o

conjunto de metodologias, aplicações, práticas e capacidades focadas na criação e administração da informação, permitindo assim, aos funcionários de uma organização, tomar melhores decisões.

Entre as definições de BI apresentadas, observa-se que algumas delas são mais objetivas e, de certa forma, restritas. Outras são mais abstratas, e não delimitam as possibilidades. Há também definições com um caráter técnico, mais direcionadas à importância da informação e dos dados. Apesar das diferenças, a maioria dos autores conclui com a importância de subsidiar e otimizar os processos decisórios de uma organização utilizando tais técnicas e métodos. É importante ressaltar que o BI não é uma ferramenta, e sim, um conceito. O BI também não é sinônimo de tecnologia, apesar de hoje ser dependente da mesma em vários aspectos. A concepção utilizada para BI neste trabalho é a seguinte: conjunto de métodos e tecnologias que permitem extrair, integrar, organizar e gerenciar dados e informações de diferentes fontes de forma estratégica, possibilitando o acesso flexível destes dados por usuários autorizados, para fornecer um melhor suporte aos seus processos decisórios.

Os objetivos do BI são: “permitir o acesso interativo aos dados (às vezes, em tempo real), proporcionar a manipulação desses dados e fornecer aos gerentes e analistas de negócios a capacidade de realizar a análise adequada” (TURBAN; SHARDA; DELEN, 2011, p. 19, tradução nossa).

Abaixo, algumas das razões para o uso de BI (DÍAZ, 2012):

- Quando a tomada de decisão é realizada de forma intuitiva na organização;
- Quando há problemas na qualidade da informação;
- Uso do software Excel como repositório de informações corporativas;
- Quando há a necessidade de cruzar dados de diferentes repositórios;
- Excesso de dados na organização para serem analisados de forma convencional;
- Necessidade de automatizar os processos de extração e distribuição de informações.

De acordo com Kimball e Ross (2013), um dos precursores dos conceitos de DW, o BI possui alguns componentes que auxiliam a otimizar processos decisórios. Abaixo são apresentados tais componentes:

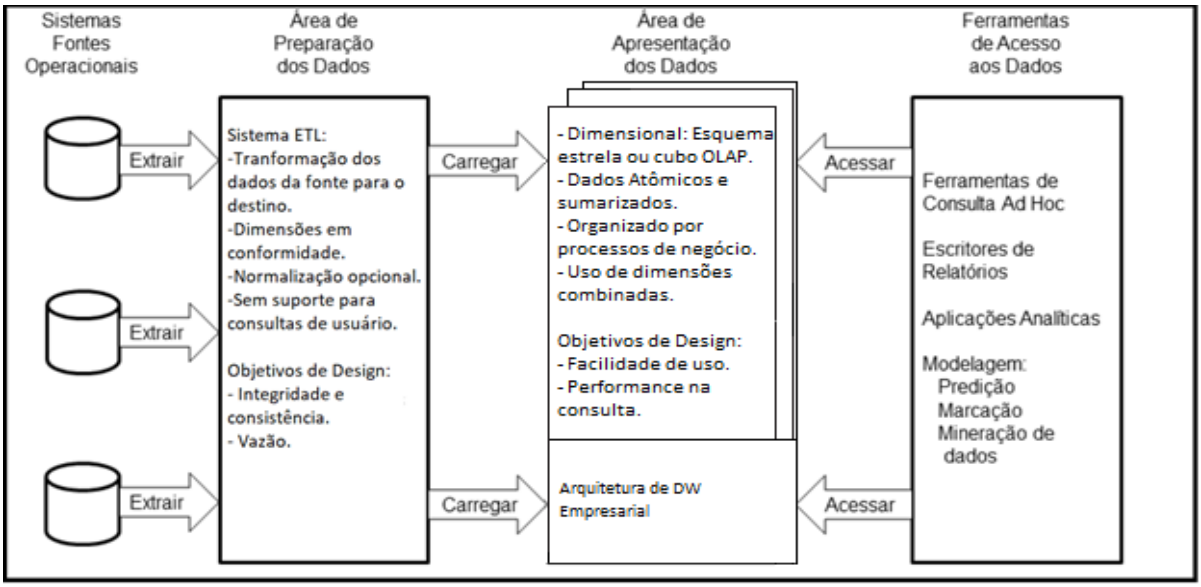
- Fontes de informação: São os repositórios de dados, internos ou externos da organização, que vão alimentar o DW. Os mais utilizados são os bancos de

dados gerenciados por um SGBD (Sistema de Gerenciamento de Banco de Dados).

- Sistema de ETL: Antes dos dados serem armazenados no DW, eles devem ser extraídos dos repositórios de dados. Posteriormente, tais dados devem ser transformados, filtrados, tratados e redefinidos. Por fim, os dados estarão prontos para serem armazenados no repositório de informações para o subsídio de tomadas de decisão.
- Área de apresentação: Através do OLAP e outros métodos, provê capacidades de cálculos, consultas, prognósticos e análises de cenários com grandes volumes de dados. Considera-se este componente como essencial, pois é o local onde os dados estão organizados, armazenados e disponíveis para consulta direta para usuários, relatórios e outras aplicações. É o DW em si, o repositório de dados preponderantes para processos decisórios.
- Aplicações de BI: São aquelas que acessam os dados da área de apresentação para entregar aos usuários as informações de forma organizada. Entende-se por consulta aos dados aplicações *ad-hoc* (consultas sob demanda), *data mining* e aplicações com modelos.

A Figura 1 apresenta uma ilustração dos componentes de uma solução de BI.

Figura 1 – Elementos principais de uma solução de BI/DW.



Fonte: Traduzido e adaptado de Kimball e Ross (2013).

Um maior detalhamento será realizado para cada um destes componentes nas próximas seções, visando proporcionar o entendimento necessário sobre como uma solução completa de BI pode otimizar processos de tomada de decisão.

2.3.1 Fontes de informação

Antes da difusão das tecnologias e popularização dos computadores e repositórios de dados, muitas organizações guardavam os seus registros históricos em arquivos de papel, o que dificultava de forma significativa a análise rápida e eficiente das informações. Hoje, a maioria das empresas possui repositórios digitais de informações, citando como exemplo um dos mais difundidos, os bancos de dados.

No contexto do BI, as fontes de informação são os locais dos quais serão extraídos os dados. Dependendo do ambiente de aplicação das técnicas de BI, diferentes podem ser as fontes de dados. Hoje, existem diversas técnicas e tecnologias utilizadas para armazenagem de dados. Giner (2007) e Turban, Sharda e Delen (2011) mencionam as seguintes possíveis fontes de informação, no contexto empresarial:

- Sistemas operacionais ou transacionais, citando como exemplo ERP (*Enterprise Resource Planning* - Sistema de Gestão Empresarial), CRM (*Customer Relationship Management* – Gestão de relacionamento com o cliente), SCM (*Supply Chain Management* – Gestão da cadeia de suprimentos), sistemas OLTP (*Online Transaction Processing* - Processamento de Transações em Tempo Real), entre outros.
- Sistemas de informação departamental.
- Fontes de informações externas (dados mercadológicos, do governo, de institutos de pesquisa, etc.) que, se relacionadas com os dados internos, podem enriquecer os processos analíticos.

Giner (2007) menciona a dificuldade e complexidade de se extrair dados de diferentes e numerosas fontes de dados. Para o autor, um dos aspectos-chave é conhecer o modelo de informação transacional e cada um dos seus elementos. A situação se torna mais complicada ainda quando as aplicações não são devidamente documentadas. Outra dificuldade que surge é quando essas fontes de dados são modificadas com o tempo, por diferentes programadores, que não atualizam a

documentação. O gerenciamento da documentação, porém, é uma peculiaridade que não é relacionada com a tecnologia.

Outro aspecto importante relacionado às fontes de informação é a qualidade dos dados. Subsidiar as decisões com dados errôneos ou inconsistentes podem acarretar em previsões imprecisas e decisões precárias. Eckerson (2002) aponta alguns aspectos que os dados devem ter para serem considerados de qualidade: precisos, íntegros, consistentes, completos, válidos, disponíveis e acessíveis.

Batini, Palmonari e Viscusi (2012) apresentam o ciclo de vida dos dados e uma série de fatores, chamados de *clusters* de dimensões, que afetam a qualidade dos dados. Por fim, Rodic e Baranovic (2009) apontam como implementar regras de qualidade dos dados no processo de ETL, que é o assunto da próxima subseção.

2.3.2 Processo ETL

Integração de dados é “o conjunto de aplicações, produtos, técnicas e tecnologias que permitem uma visão única consistente dos nossos dados de negócio” (DÍAZ, 2012, p. 56, tradução nossa). Não é uma tarefa trivial para desenvolvedores realizar a integração de dados de diferentes fontes externas (NILAKANTA; SCHEIBE; RAI, 2008). Uma das tecnologias existentes que permitem a integração de dados é o processo de ETL.

Para Kimball e Ross (2013) o processo de ETL é dividido em três etapas: extração, transformação e carga dos dados. Este processo ocorre na área de preparação dos dados, entre as fontes de informação e a área de apresentação.

O primeiro passo do ETL é a extração dos dados necessários das diferentes fontes de informação para posterior manipulação (DÍAZ, 2012; KIMBALL; ROSS, 2013; TURBAN; SHARDA; DELEN, 2011). Estes dados são armazenados em uma base de dados temporária, chamada na literatura por *stage area* ou área de preparação dos dados.

Posteriormente, poderão ser necessárias transformações nos dados, visando corrigir erros de digitação, conflitos de domínios, dados faltantes ou incompletos, conversão do formato do dado, combinação de dados de diferentes fontes, correção de dados duplicados, entre outras possibilidades (KIMBALL; ROSS, 2013). Essa é a etapa de transformação dos dados.

A última etapa do processo ETL é a carga dos dados para os modelos dimensionais da área de apresentação. Nesta etapa é realizado o processamento das dimensões, como: atribuição de chaves substitutas (*surrogate keys)*, fornecimento de descrições apropriadas para os atributos das dimensões, repartições ou combinação de colunas para apresentação dos valores, entre outros. Com relação às tabelas fato, apesar de grandes e consumirem muito tempo para carregar, sua preparação para a área de apresentação é, em geral, direta (KIMBALL; ROSS, 2013). Após este processo, os dados ficam disponíveis na área de apresentação dos dados.

Na visão de Giner (2007) o processo ETL se subdivide em cinco processos: extração, limpeza, transformação, integração e atualização. A extração se refere a captura dos dados das diferentes fontes de informação. O processo de limpeza se refere à correção dos dados, tornando-os limpos e de alta qualidade. A etapa de transformação estrutura os dados para adequá-los aos modelos de análises. A integração é o momento em que os dados são carregados no DW, assegurando a consistência destes com o formato e definições do DW, além de integrá-los nos diferentes modelos de cada área de negócio previamente definidos. Por fim, a etapa de atualização é o momento em que é definida a periodicidade de carga dos dados no DW, que irá garantir dados atualizados em relação às fontes de dados.

2.3.3 *Data Warehouse*

Nesta subseção, serão abordados os conceitos e tecnologias relacionadas ao DW. A subseção foi dividida nos seguintes assuntos: definição e características, arquiteturas de DW e modelagem dimensional de dados.

2.3.3.1 Definição e Características de um DW

O principal componente de um sistema de BI, para Díaz (2012), é o *Data Warehouse* (em português, Armazém de Dados). Os autores mencionam a seguinte definição para um DW:

> Um *data warehouse* é um repositório de dados que proporciona uma visão global, comum e integrada dos dados de uma organização – independentemente de como sejam utilizados posteriormente pelos consumidores ou usuários, com as seguintes propriedades: estável, coerente, confiável e com informação histórica (DÍAZ, 2012, p. 32, tradução nossa).

Na Tabela 2, Han, Kamber e Pei (2011) apresentam as principais diferenças entre bancos de dados convencionais, conhecidos como sistemas OLTP, e DW, conhecidos como sistemas OLAP.

Tabela 2 - Diferenças entre sistemas OLTP e OLAP.

Atributo	OLTP	OLAP
Característica	Processamento operacional	Processamento informacional
Orientação	Transações	Análises
Usuário	Administradores e profissionais de bancos de dados	Gerentes, executivos e analistas
Função	Operações rotineiras	Apoio de decisão de requisitos informativos a longo prazo
Design do Banco de Dados	Modelo ER e orientado a aplicações	Modelo estrela/floco de neve, orientado a assuntos
Dados	Atuais, atualização garantida	Históricos, acurácia mantida com o tempo
Sumarização	Primitiva, altamente detalhada	Sumarizados, consolidados
Visão	Detalhados, plano relacional	Sumarizados, multidimensionais
Unidade de trabalho	Curta, transação simples	Consultas complexas
Acesso	Leituras/Escritas	Grande parte são leituras
Foco	Entrada de dados	Saída de informações
Operações	Indexação/*hash* com chaves-primárias	Muitas varreduras
Número de registros acessados	Dezenas	Milhões
Número de usuários	Milhões	Centenas
Tamanho do banco de dados	*Gigabytes* (GB)	≥ *Terabytes* (TB)
Prioridade	Alta performance, alta disponibilidade	Alta flexibilidade, autonomia para o usuário final
Métricas	*Throughput* de transações	*Throughput* de consultas, tempo de resposta

Fonte: Adaptado de Han, Kamber e Pei (2011, p. 130).

Outra definição clássica de DW é a de Inmon (2002, p. 31): “O *Data Warehouse* é uma coleção de dados orientados a assunto, integrados, não-voláteis e variantes no tempo que apoiam decisões de gerenciamento”. Abaixo, segue uma descrição detalhada destas características, elaboradas por Inmon (2002):

- Orientação a assunto: as informações são organizadas ao redor de um assunto central, que em geral é relevante no processo decisório.
- Integrado: Contém todos os dados das diferentes fontes de informação da organização, ou seja, dados relativos ao processo de negócio. Tais dados são obtidos, organizados, transformados e corrigidos no processo de ETL.

- Variante no tempo: É necessário um componente temporal que permita a avaliação dos dados sob uma perspectiva histórica.
- Não voláteis: Os dados devem ser estáveis, sem a possibilidade de exclusões, disponíveis apenas para leitura dos usuários.

Díaz (2012) também descreve alguns elementos associados ao DW:

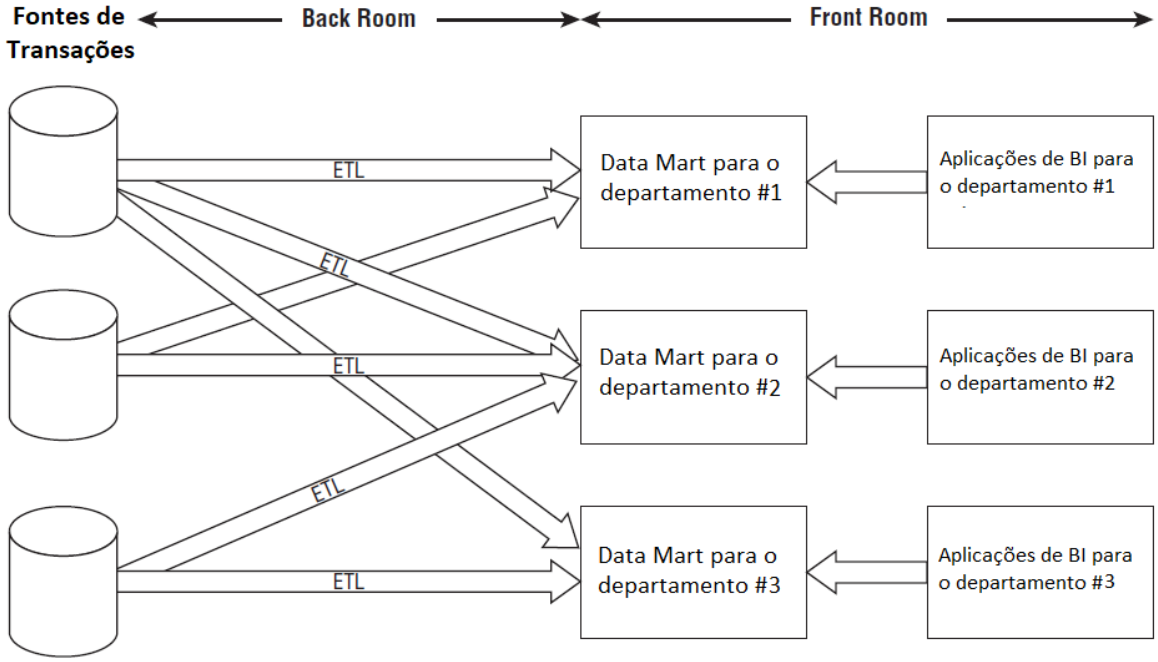
- *Data Mart* (DM): é um subconjunto dos dados do DW, após transformados e filtrados no processo de ETL. O objetivo é atender demandas ou necessidades de um conjunto específico de usuários ou um departamento específico de uma organização. No contexto da pecuária de corte, seria possível termos um DM específico para cada setor da pecuária: gestão econômica e financeira, questões produtivas do rebanho, questões ambientais e biológicas, entre outras.
- Armazém de dados operacionais (*Operational Data Store* – ODS): É um repositório de dados que fornece os valores mais atuais das informações, e não o histórico.
- Área de preparação dos dados (*Stage Area*): Repositório de dados intermediário. Tal repositório, considerando a arquitetura de uma solução de BI, fica entre as fontes de dados operacionais e o DW. Tal repositório possui os seguintes objetivos:
    - Auxiliar no processo de ETL, processo complexo que envolve extração, transformação e carga de grandes volumes de dados de fontes possivelmente heterogêneas;
    - Auxiliar na melhoria da qualidade dos dados;
    - Funciona como uma cache dos dados operacionais, usado no processo de carga para o DW;
    - Utilizada para acesso em detalhe de dados que possivelmente não estão no DW.
- Metadados: dados estruturados e codificados que visam auxiliar na descrição, organização, identificação, descoberta e administração de instâncias ou dados. Em resumo, são dados sobre os dados.

2.3.3.2 Arquiteturas de DW

Existem diferentes arquiteturas possíveis para uma solução de DW envolvendo os elementos anteriormente apresentados. A seguir, são apresentadas algumas arquiteturas comuns encontradas na literatura:

Arquitetura de *Data Marts* Independentes (*Independent Data Marts Architecture*) (KIMBALL; ROSS, 2013; TURBAN; SHARDA; DELEN, 2011): Os dados analíticos são desenvolvidos em uma base departamental, sem a preocupação de compartilhamento ou integração de informações com outros setores de uma empresa. Em outras palavras, os dados analíticos nos DM são desenvolvidos para operar independentemente um do outro. A Figura 2 ilustra essa arquitetura.

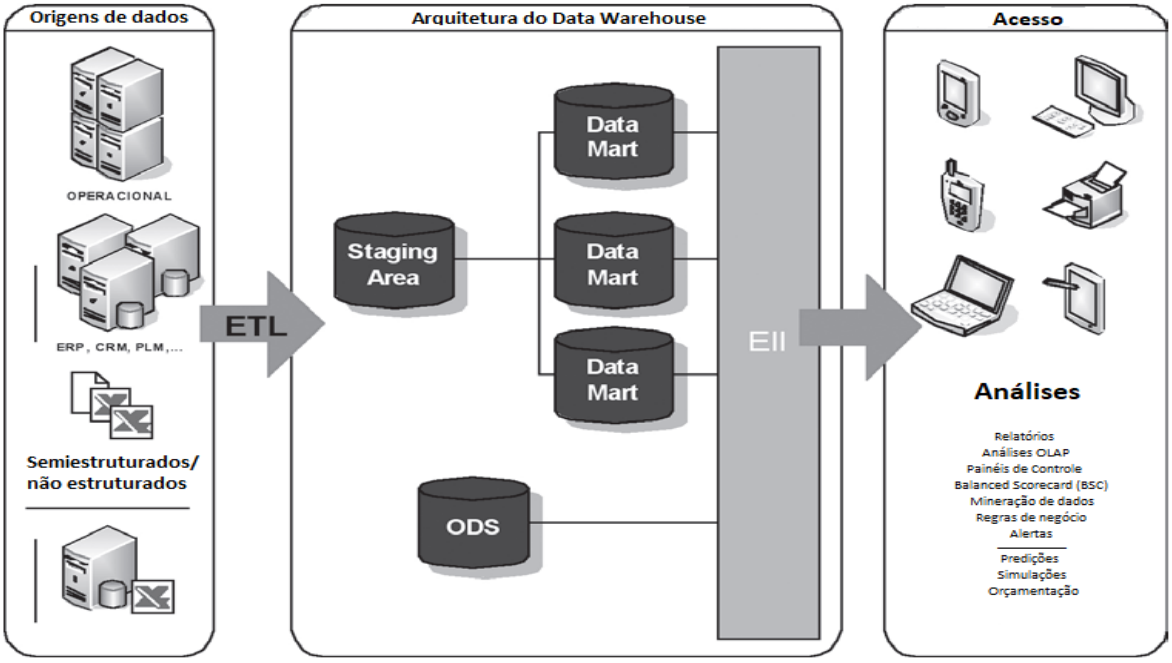
Figura 2 - Arquitetura de DM independentes.



Fonte: Adaptado de Kimball e Ross (2013, p. 27).

Enterprise Bus Architecture ou ***Data Warehouse*** **Virtual/Federado** (DÍAZ, 2012): arquitetura baseada em DM independentes federados que podem fazer uso de uma área de preparação dos dados, caso necessário. Se faz uso de uma ferramenta de integração de informações empresariais para realizar consultas como se existisse um único DW. Se necessário, também pode ser utilizado um ODS. A Figura 3 ilustra essa arquitetura.

Figura 3 - *Enterprise Bus Architecture*.

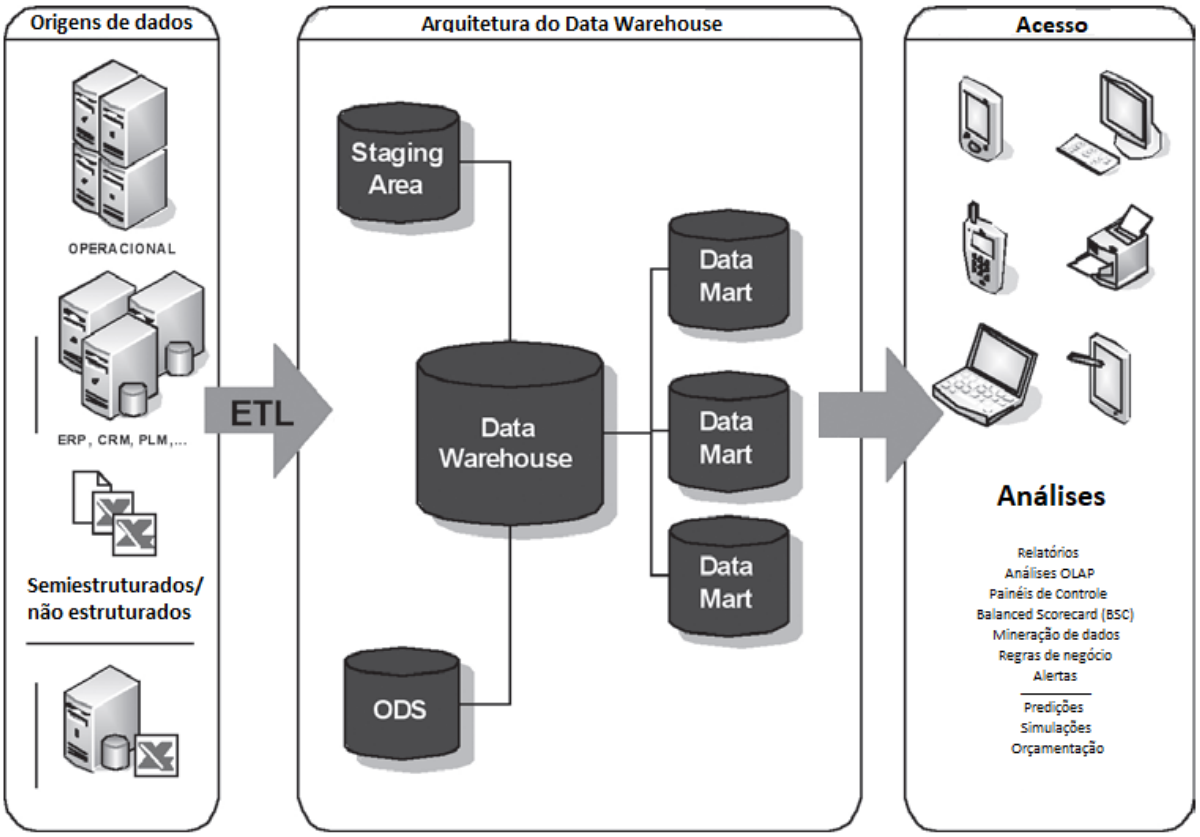


Fonte: Adaptado de Díaz (2012, p. 39).

Fábrica de informações corporativas (*Corporate Information Factory - CIF*) (DÍAZ, 2012; KIMBALL; ROSS, 2013): abordagem conhecida e defendida por Inmon, consiste na existência de um DW corporativo e DM dependentes deste. Na visão de Díaz (2012), o acesso aos dados é feito através dos DM (ou ODS, caso existam), e não pelo DW. Se necessária, pode existir uma área de preparação dos dados. Já para Kimball e Ross (2013), os dados são extraídos das fontes de dados operacionais, processados pelo ETL e, por fim, os dados atômicos resultantes são armazenados em um banco de dados normalizado pela norma 3NF. Este repositório é o próprio DW Empresarial. As Figuras 4 e 5 apresentam este modelo de arquitetura de DW, nas visões de Díaz (2012) e Kimball e Ross (2013), respectivamente.

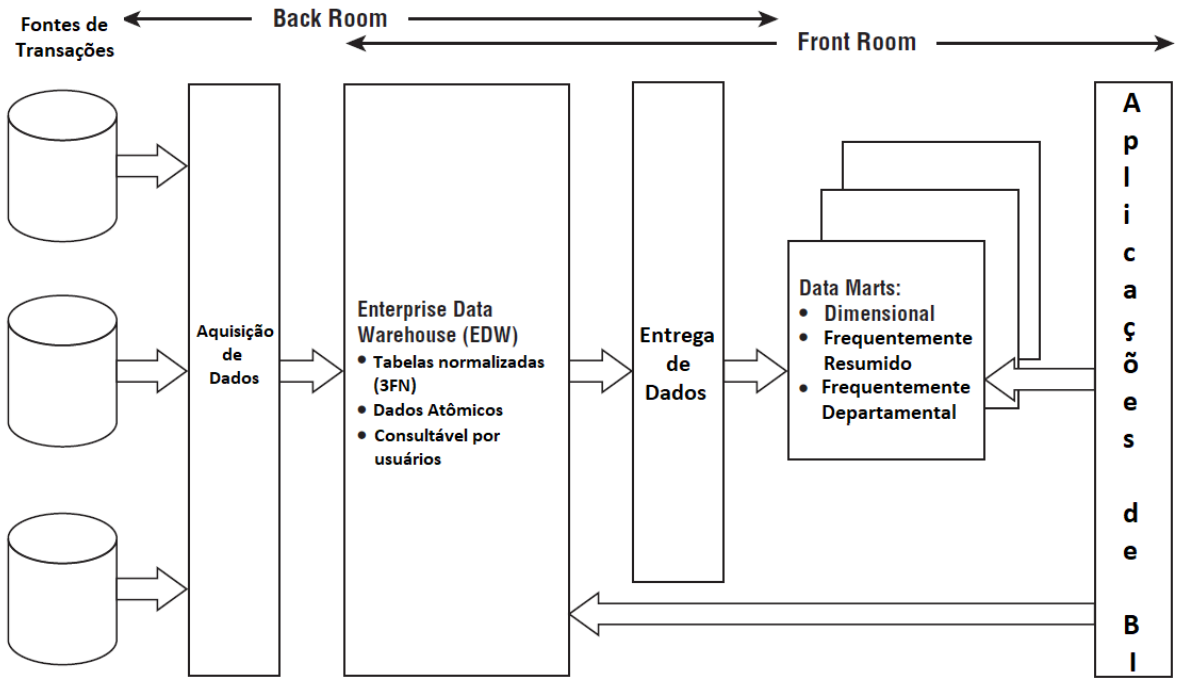
Arquitetura de *Data Warehouse* Centralizado (ou também chamado de *Enterprise Data Warehouse – EDW*) (TURBAN; SHARDA; DELEN, 2011): Está arquitetura de DW é semelhante à CIF, porém não existem DM dependentes. Nesta arquitetura, existe apenas um DW centralizado em que todos os usuários tem total acesso aos dados. Usuários não são limitados ao acesso de dados de apenas um departamento, como seria em um DM. É uma abordagem defendida pela corporação *Teradata*. A Figura 6 demonstra este modelo de arquitetura abordada por Turban.

Figura 4 - *Corporate Information Factory* na visão de Díaz.



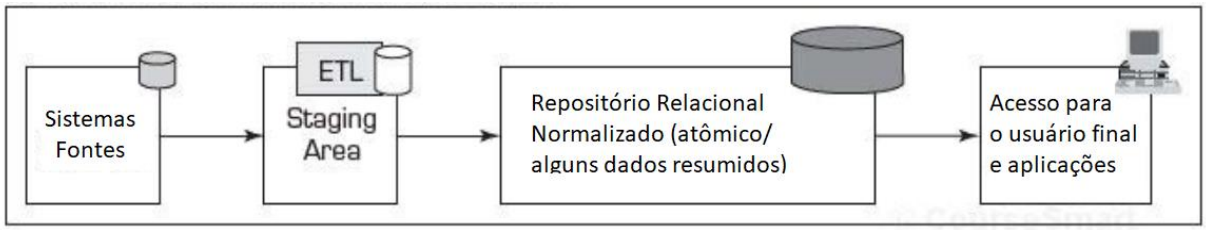
Fonte: Adaptado de Díaz (2012, p. 40).

Figura 5 - *Corporate Information Factory* na visão de Kimball e Ross.



Fonte: Adaptado de Kimball e Ross (2013, p. 28).

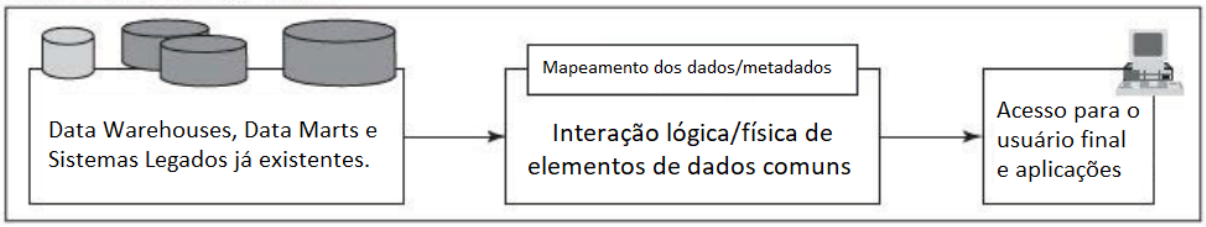
Figura 6 – Arquitetura de um DW centralizado.



Fonte: Adaptado de Turban, Sharda e Delen (2011, p. 338).

Arquitetura Federada (TURBAN; SHARDA; DELEN, 2011): Tal arquitetura utiliza todos os meios possíveis para integrar recursos analíticos de múltiplas fontes de forma que vá de encontro às necessidades dos negócios ou de mudanças. Sistemas existentes são mantidos e seus dados são acessados, conforme necessário. Tal abordagem é suportada por vendedores de *middlewares* que oferecem consultas distribuídas e capacidades de junções (*joins*). A Figura 7 ilustra essa arquitetura.

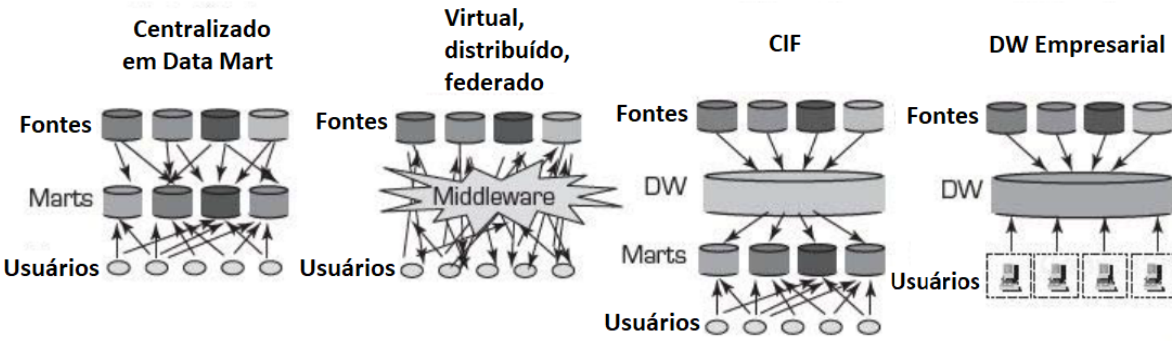
Figura 7 – Arquitetura federada.



Fonte: Adaptado de Turban, Sharda e Delen (2011, p. 338).

Turban, Sharda e Delen (2011) apresentam uma comparação de quatro das arquiteturas apresentadas (DM independentes, CIF, DW centralizado e a federada), com suas vantagens e desvantagens, conforme mostra a Figura 8.

Figura 8 - Comparação entre quatro arquiteturas de DW.



	Data Marts Independentes	Deixe os dados onde eles estão	Data Marts Dependentes	Dados centralizados e integrados com acesso direto
Prós	• Fácil de construir organizacionalmente. • Fácil de construir tecnicamente.	• Sem necessidade de ETL. • Sem necessidade de plataformas adicionais.	• Permite fácil customização com a interface do usuário e relatórios.	• Visão do negócio da empresa. • Consistência do design e qualidade dos dados. • Reusabilidade dos dados.
Contras	• Empresa de negócio vê como indisponível. • Custos com dados redundantes. • Alto custo de ETL. • Alto custo das aplicações. • Altos custos operacionais.	• Somente viável para baixos volumes de dados. • Problemas com metadados. • Problemas com largura de banda de rede e complexidade de *joins*. • Carga de trabalho típicamente colocada em uma estação de trabalho.	• Empresa de negócio vê como desafiador. • Custos com dados redundantes. • Altos custos operacionais. • Latência dos Dados.	• Requer liderança corporativa e visão.

Fonte: Adaptado de Eckerson (2003, p. 46-49), *apud* Turban; Sharda; Delen (2011, p. 339).

Foi realizada uma análise de sucesso das arquiteturas de DM independentes, *Enterprise Bus Architecture*, CIF, DW centralizado e a federada, com 454 empresas variando em tamanho, localidades e em percentuais de adoção destas arquiteturas. Através de quatro medidas de sucesso das arquiteturas, foi constatada uma pontuação inferior para as arquiteturas de DM independentes e a federada, nesta ordem, fato já conhecido e apontado na literatura. As arquiteturas restantes tiveram pontuações superiores e similares, sendo constatado que a CIF é a mais utilizada nas implementações coorporativas e em DW maiores, ao mesmo tempo que é a solução mais cara e a que consome mais tempo de implementação (ARIYACHANDRA; WATSON, 2006).

Não existe uma metodologia rigorosa que permita determinar com alta precisão a arquitetura ideal para ser utilizada em uma solução qualquer. Cada negócio tem as suas particularidades, objetivos, metas e limites que devem ser considerados no momento da definição da arquitetura da solução do problema. Ariyachandra e Watson (2005) apontam alguns como fatores determinantes na escolha de uma solução:

interdependência de informação entre as unidades organizacionais, necessidades de informação de gerentes superiores, urgência ou necessidade de um DW, natureza das tarefas para o usuário final, restrições de recursos, visão estratégica prévia à implementação do DW, compatibilidade entre os sistemas existentes, capacidade percebida da equipe de Tecnologia da Informação (TI) interna, questões técnicas e fatores sociais e políticos.

2.3.3.3 Modelagem Dimensional

Os bancos de dados relacionais utilizam modelos do tipo Entidade-Relacionamento (ER), em que temos diferentes entidades, que representam objetos do mundo real, e seus respectivos relacionamentos. O modelo ER também visa eliminar redundâncias de dados. Estes modelos de dados são otimizados para processamentos de transações *online* (OLTP). Tais modelos de dados são muito complicados para consultas de BI, que são muitas vezes imprevisíveis e complexas, sobrecarregando os otimizadores dos bancos de dados, resultando em consultas com um péssimo desempenho (KIMBALL; ROSS, 2013). Isso pode comprometer um dos principais aspectos da área de apresentação de um DW, que é a entrega das informações com alta performance e de forma intuitiva (KIMBALL; ROSS, 2013).

Para solucionar os problemas acima, surge o conceito de modelagem dimensional (ou multidimensional). A modelagem dimensional é uma técnica que permite que os bancos de dados sejam simplificados, visando facilitar o entendimento dos dados para usuários e também permitir que softwares entreguem resultados de consultas de forma rápida e eficiente (KIMBALL; ROSS, 2013).

A implementação de modelos dimensionais em bases de dados relacionais é em geral chamada de esquema estrela, devido à sua semelhança com a estrutura de uma estrela. A implementação de modelos dimensionais em ambientes de banco de dados multidimensionais são referidos como cubos OLAP. Apesar de utilizarem os mesmos conceitos, a implementação física de cada uma é diferente (KIMBALL; ROSS, 2013).

A seguir, alguns dos elementos chaves dos modelos dimensionais (DÍAZ, 2012; KIMBALL; ROSS, 2013):

- **Tabela fato**: Cada registro na tabela fato representa um evento que foi medido. Todos os dados nesta tabela devem possuir a mesma granularidade, ou seja,

o mesmo nível de detalhamento. Um exemplo de tabela fato poderia ser as vendas de bovinos de uma determinada empresa rural. Essas tabelas possuem dois tipos de atributos: as medidas dos processos de negócio e as chaves estrangeiras para registros em tabelas de dimensões.

- **Tabela dimensão**: representam o contexto textual dos eventos registrados nas tabelas fato. A tendência é que tais tabelas possuam mais colunas ou atributos, porém, menos registros que as tabelas fato. Cada dimensão possui uma chave-primária única que serve para a integridade referencial com quaisquer tabelas fato associadas à mesma. Um exemplo de tabela deste tipo seria a dimensão “animal”, que contém diversos atributos sobre um determinado bovino, como raça, sexo, lote, descrição, entre outros.
- **Medidas ou métricas**: medidas de performance resultantes dos processos de negócio da organização. Exemplos de medidas de negócio que podem ser parte de uma tabela fato seriam o número de bovinos vendidos e o valor total de vendas de uma determinada raça de bovinos.

Com relação as métricas e os tipos de informações que estas armazenam, Díaz (2012) fazem uma separação entre dois tipos: métricas e indicadores-chave. Há dois tipos de métricas:

- **Métricas de realização de uma atividade**: medem a realização de uma atividade qualquer no contexto do negócio. Exemplo: participação dos empregados em algum evento.
- **Métricas de resultado de uma atividade**: medem o resultado de uma atividade qualquer no contexto do negócio. Exemplo: média de ganho de peso diário de um bovino.

Já os indicadores-chave são informações relacionadas ao que se pretende alcançar e mostram o grau de aceitação dos objetivos. Tais indicadores informam o rendimento das atividades que visam atingir as metas estabelecidas. Temos dois tipos destes indicadores (DÍAZ, 2012):

- **Indicador Chave de Performance** (*Key Performance Indicator* – KPI): Métricas do processo que nos informam o desempenho e eficiência necessários para atingir os objetivos.

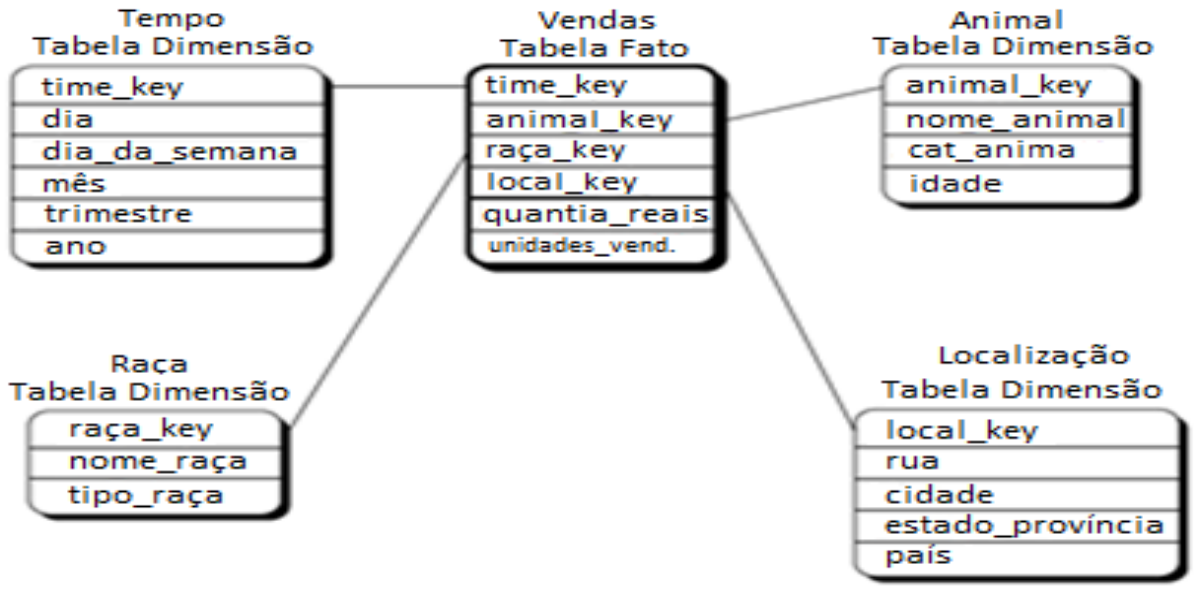
- **Indicador Chave de Metas** (*Key Goal Indicator* – KGI): Definem as medições para informar a gerência se um processo de TIC atingiu seus requisitos de negócios. Geralmente são expressos em termos de critérios de informação.

Apresentados os elementos básicos dos modelos dimensionais, abaixo são abordadas as diferentes formas de modelagem dimensional para DW e DM implementadas em bancos de dados relacionais:

- **Esquema Estrela (*Star Schema*)**: Este esquema possui uma tabela fato centralizada com a maior parte dos dados, sem redundâncias, e um conjunto de tabelas auxiliares, uma para cada dimensão. Cada dimensão possui uma tabela e cada uma destas possui um conjunto de atributos (HAN; KAMBER; PEI, 2011). Um exemplo de tabela dimensão seria a localização, com os atributos rua, cidade, estado, país. Essa tabela introduz redundâncias, como: Dom Pedrito e Bagé fazem parte do estado do Rio Grande do Sul (RS) e do Brasil (BR), portanto, gerando redundâncias nos atributos estado e país.
- **Esquema Floco de Neve (*Snowflake Schema*)**: Este esquema possui algumas tabelas de dimensões normalizadas, fazendo com que os dados sejam divididos e mais tabelas sejam geradas (HAN; KAMBER; PEI, 2011). Para Díaz (2012, p. 34), temos dois tipos de esquemas deste tipo:
    - **Completo**: Todas as tabelas de dimensão estão normalizadas.
    - **Parcial**: Apenas algumas das tabelas de dimensão estão normalizadas.
- **Constelação de Fatos (*Fact Constellations* ou *Galaxy Schema*)**: Para aplicações mais sofisticadas, poderá ser necessário um conjunto de tabelas fato compartilhando várias dimensões. Neste esquema, é possível que diversas tabelas de dimensões sejam compartilhadas com diferentes tabelas fato (HAN; KAMBER; PEI, 2011).

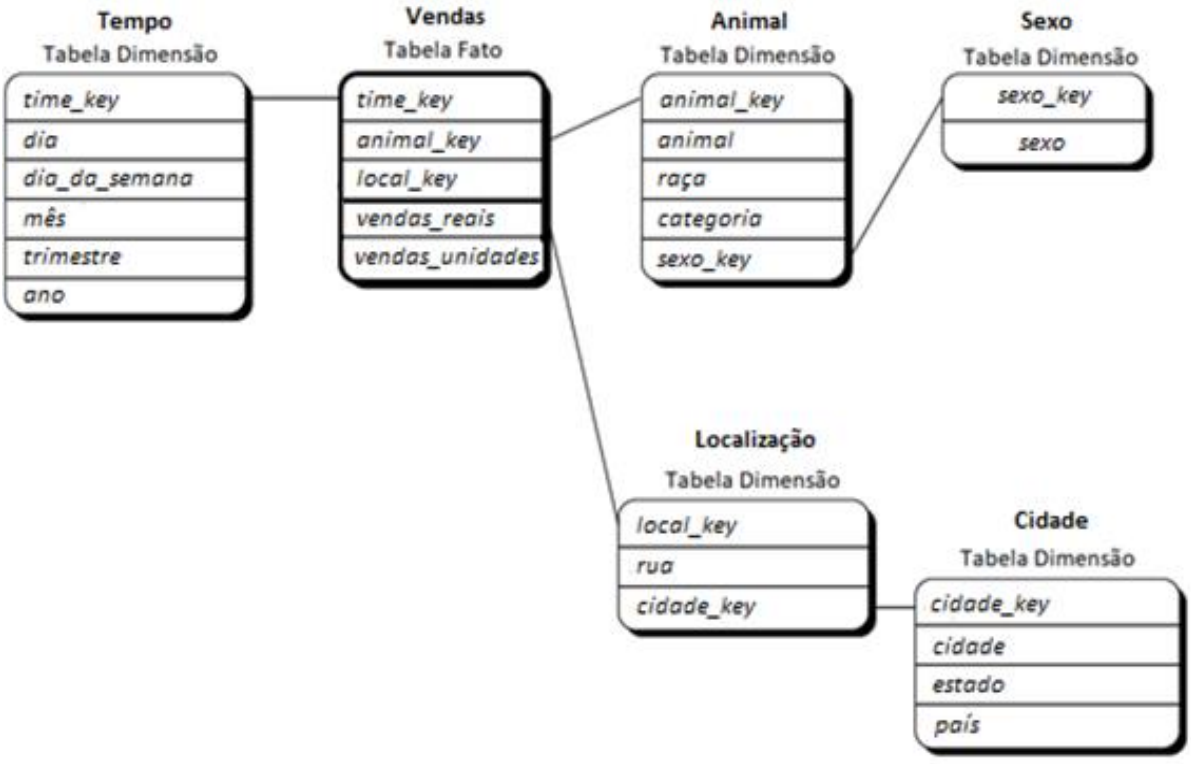
Han, Kamber e Pei (2011) afirmam que o esquema de Constelação de Fatos é geralmente utilizado porque o DW possui assuntos que remetem a toda a organização. Já os esquemas Estrela e o Floco de Neve são mais utilizados em DM, pois cada um trata de um assunto específico na organização, em nível de departamentos. As Figuras 9, 10 e 11 apresentam exemplos destes modelos de dados dimensionais aplicados em um contexto de vendas de bovinos.

Figura 9 - Esquema Estrela para vendas de bovinos.



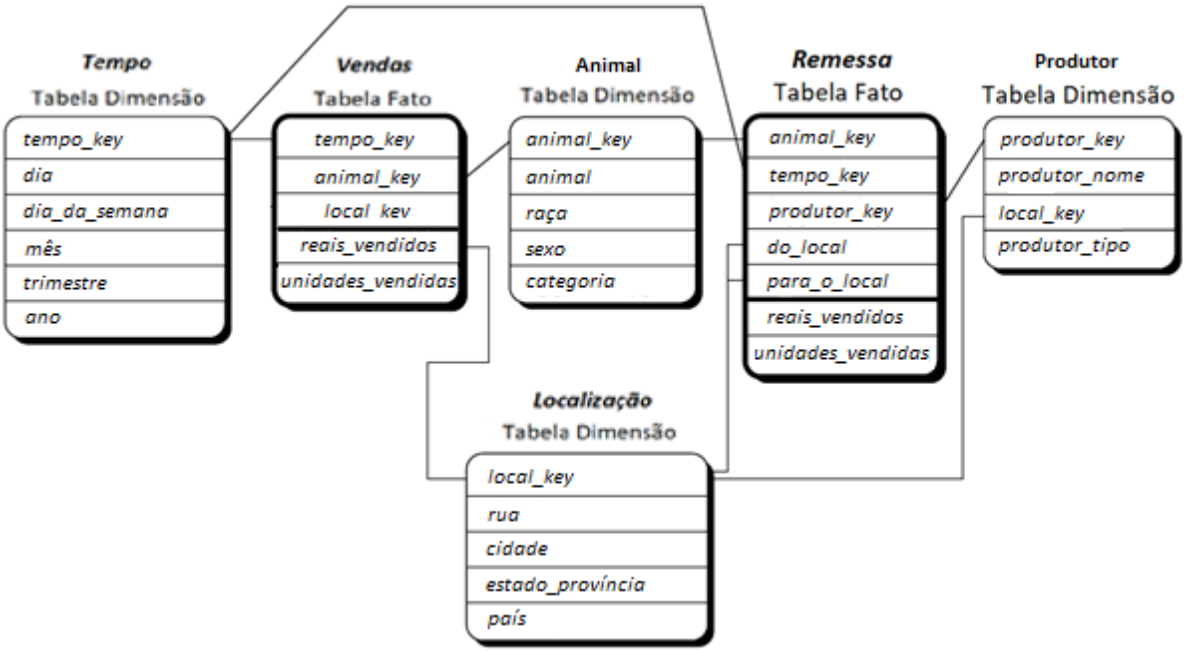
Fonte: Adaptado de Han, Kamber e Pei (2011, p. 140).

Figura 10 - Esquema Floco de Neve para vendas de bovinos.



Fonte: Adaptado de Han, Kamber e Pei (2011, p. 141).

Figura 11 - Esquema Constelação de Fatos para vendas de bovinos.



Fonte: Adaptado de Han, Kamber e Pei (2011, p. 142).

Outro aspecto importante na modelagem dimensional é a granularidade dos dados. A granularidade significa o nível de detalhe dos dados que serão disponibilizados no modelo dimensional do DW. Quanto menor for a granularidade dos dados, maior o nível de detalhe dos dados, portanto, mais dados serão armazenados no DW. Um exemplo de decisão envolvendo granularidade é se os dados (métricas e medidas do negócio) serão armazenados por hora, diariamente, semanalmente ou mensalmente. A decisão do nível de granularidade tem impacto nos tipos de consultas que podem ser realizadas, no volume de dados armazenados no DW, no desempenho das consultas e abrangência de informações disponíveis (INMON, 2002).

2.3.4 Área de Apresentação

A área de apresentação é onde os dados são armazenados, organizados e disponibilizados para consultas diretas por usuários, analistas e executivos, além de aplicações de BI analíticas. É tudo que a empresa pode ver e acessar através das aplicações e ferramentas de BI (KIMBALL; ROSS, 2013).

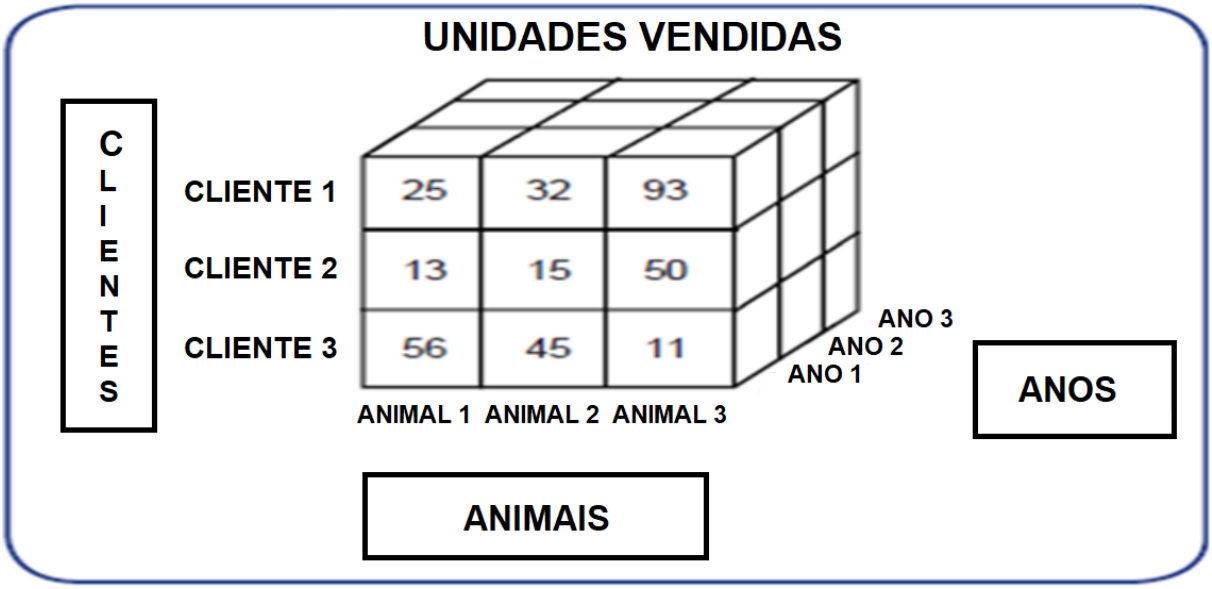
Os mesmos autores apresentam opiniões a respeito deste componente:

- A apresentação, armazenamento e acesso aos dados devem ser feitos em esquemas dimensionais ou cubos OLAP.
- A área de apresentação devem conter dados detalhados e atômicos.
- Deve ser estruturada ao redor das medidas dos processos de negócio.

Na subseção anterior, foram demonstradas diferentes possibilidades para a modelagem dimensional dos dados no DW ou nos DM. Os dados são organizados em múltiplas dimensões, sendo que cada uma possui um determinado nível de abstração. Tal organização dos dados permite que os usuários possam visualizar as informações por diferentes perspectivas com flexibilidade.

Apesar do autor (Giner, 2007) utilizar outra nomenclatura para este componente do DW, uma das tecnologias mais difundidas que permitem a análise dos dados armazenados e estruturados no DW ou nos DM é a OLAP. Na literatura, também é muito comum o termo cubo OLAP, pois a visão multidimensional das informações proporcionada por tecnologias OLAP pode ser apresentada em uma estrutura com formato de cubo. A Figura 12 apresenta um exemplo de cubo.

Figura 12 - Exemplo de cubo OLAP.



Fonte: Adaptado de Giner (2007, p. 127).

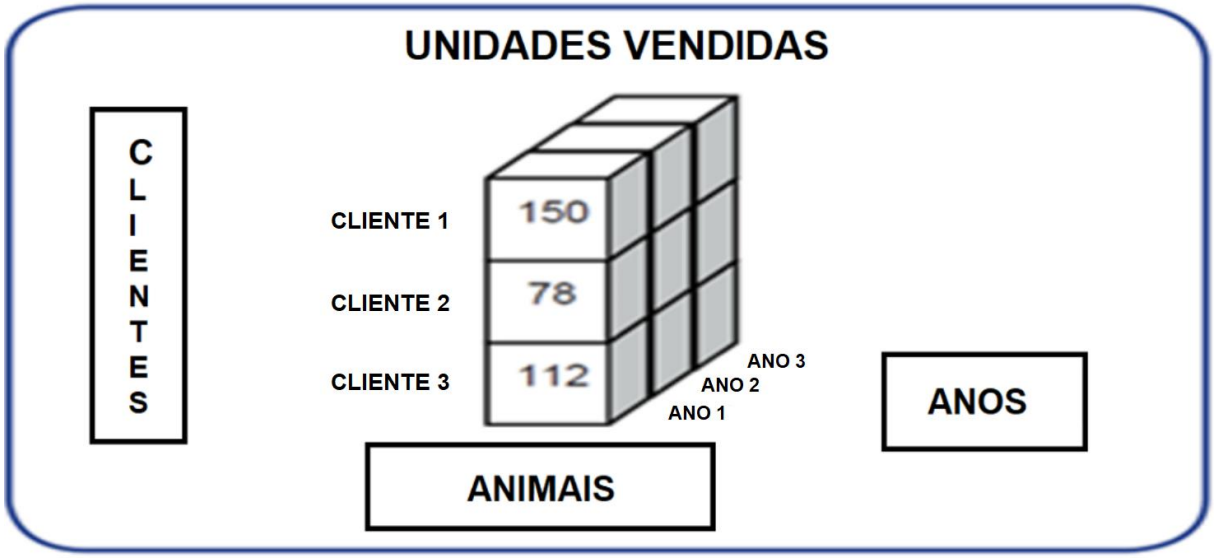
Neste exemplo de cubo, temos as dimensões de cliente, animais e anos. A informação contida em cada cubo (as unidades vendidas dos animais, neste caso) é chamada de fato. No cubo, temos as informações de unidades vendidas de cada

animal, para cada cliente e em cada ano. Portanto, podemos acessar informações como o número de animais de determinado tipo vendidos, para determinado cliente, e em determinado ano. Isso é possível através de diversas operações com o cubo.

Existem diferentes operações OLAP para dados multidimensionais que viabilizam a flexibilidade no acesso aos dados do DW/DM para os usuários. Abaixo segue uma explicação de cada uma delas, usando o exemplo do cubo da Figura 13 como base (HAN; KAMBER; PEI, 2011; GINER, 2007):

- *Roll-up (*também chamada de *Drill-up)*: Operação que permite uma agregação ou sumarização dos dados. Essa operação promove uma maior granularidade dos dados, um menor detalhamento. Isso pode ser feito de duas formas: através do aumento do nível de hierarquia de uma dimensão ou através da redução de dimensionalidade. Um exemplo do primeiro caso: Se o total das vendas estivessem armazenadas mensalmente, através de uma operação *roll-up* poderíamos agregar e apresentar o total de vendas dos animais de mês para ano. Aumenta-se a granularidade dos dados sem remover a dimensão relacionada ao tempo. Um exemplo do segundo caso: Se interessar apenas o total de vendas de todos os animais, é possível fazer uma operação de *roll-up* para apresentar os totais de vendas para cada cliente em cada ano, não importando mais qual o animal vendido. A Figura 13 apresenta este exemplo.

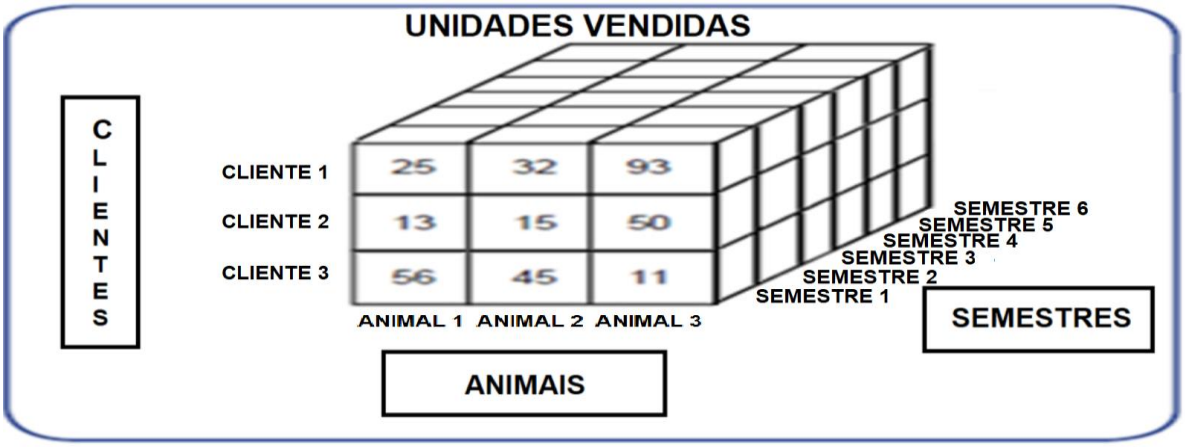
Figura 13 - Exemplo de operação *roll-up* com o cubo da Figura 12.



Fonte: Adaptado de Giner (2007, p. 129).

- *Drill-down*: é a operação inversa do *roll-up*. Permite que a apresentação dos dados seja mais detalhada, ou seja, com um menor nível de granularidade. Isso pode ser feito pela diminuição no nível de hierarquia de alguma dimensão ou através da adição de novas dimensões nas análises. Um exemplo de operação *drill-down* é apresentar o total das vendas por semestre, em vez de ano, caso houvessem disponíveis essas informações neste detalhe. Com isso, apresenta-se os dados em um maior detalhamento. A Figura 14 apresenta esse exemplo.

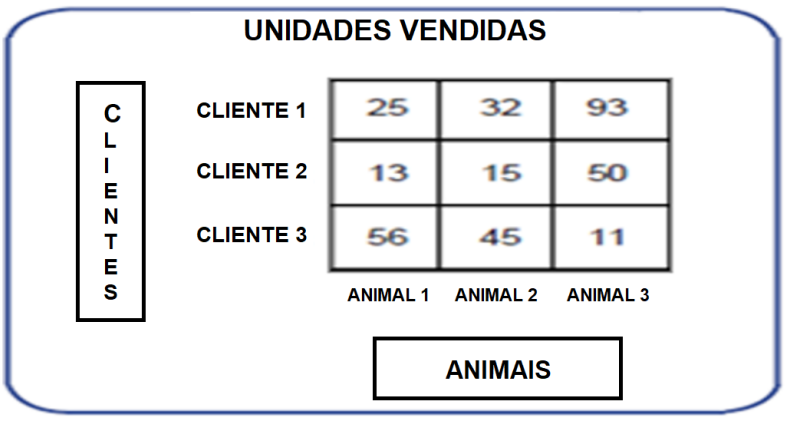
Figura 14 - Exemplo de operação *drill-down* com o cubo da Figura 12.



Fonte: Autor (2019).

- *Slice*: Esta operação é a seleção de uma dimensão do cubo, resultando em um subcubo. Um exemplo seria considerar apenas as vendas do ano 1. A Figura 15 apresenta o exemplo desta operação.

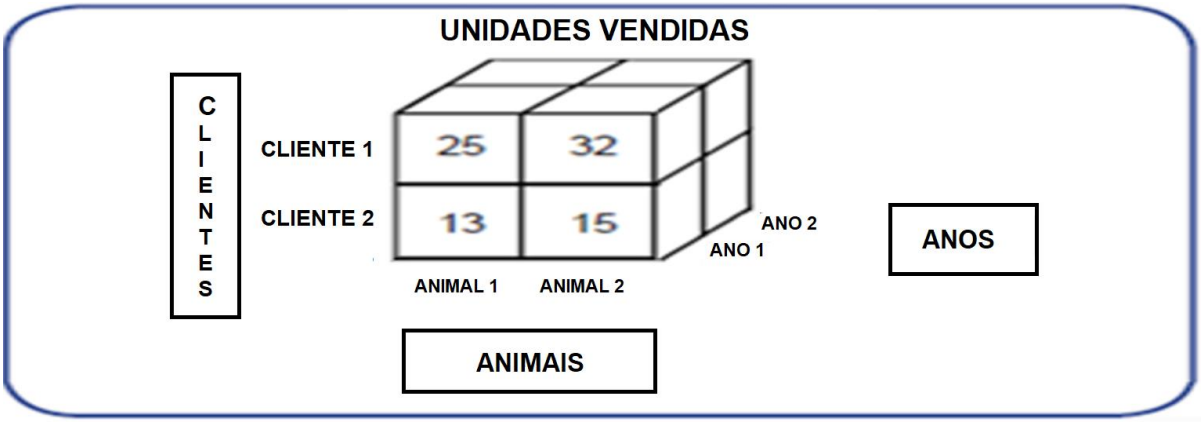
Figura 15 - Exemplo de operação *slice* com o cubo da Figura 12.



Fonte: Autor (2019).

- *Dice*: Esta operação é a seleção de duas ou mais dimensões do cubo para análises, resultando em um subcubo. Um exemplo seria considerar apenas as vendas do ano 1 ou 2, para os clientes 1 ou 2 e dos animais 1 ou 2. A Figura 16 apresenta o exemplo desta operação.

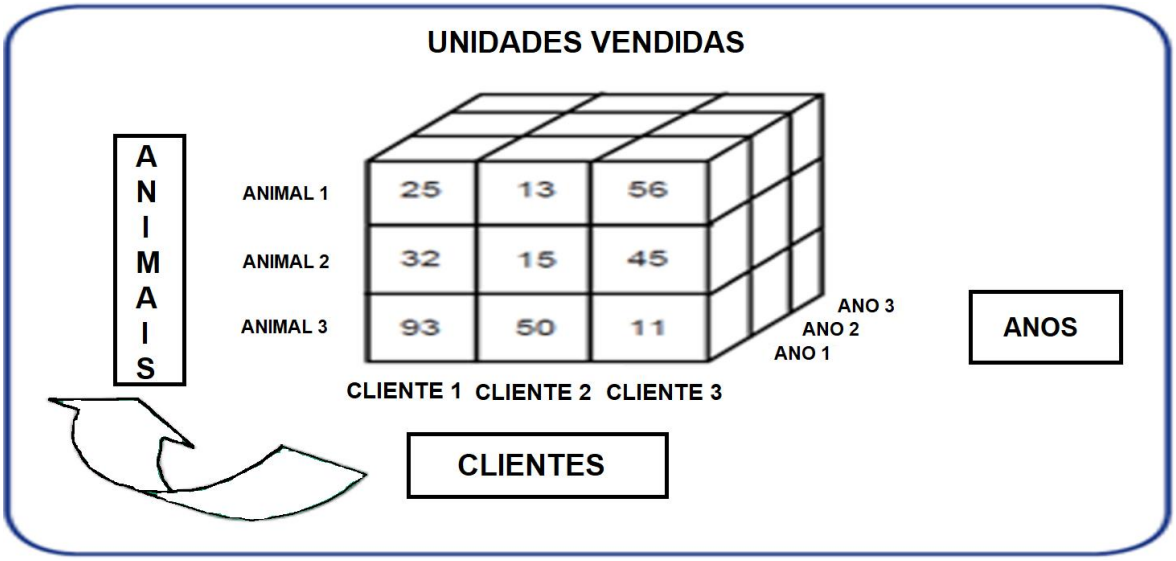
Figura 16 - Exemplo de operação *dice* com o cubo da Figura 12.



Fonte: Autor (2019).

- *Pivot* (também chamada de *rotate* ou rotação): É uma operação de visualização que realiza uma rotação dos dados entre os eixos. É uma apresentação alternativa dos dados para o usuário. Um exemplo desta operação seria, em vez de visualizar as vendas por cliente, visualizá-las por tipo de animal. A Figura 17 apresenta uma execução desta operação sobre o cubo da Figura 12.

Figura 17 - Exemplo de operação *pivot* com o cubo da Figura 12.



Fonte: Adaptado de Giner (2007, p. 129).

- Outras operações OLAP: Há ainda a operação *drill-across*, que envolve consultas com mais de uma tabela fato. A operação *drill-through* usa recursos do SQL para detalhar o nível mais baixo de um cubo de dados. Tem-se ainda outras possibilidades, como: listagem dos 'N' valores mais altos ou mais baixos e medidas estatísticas, como médias e medianas.

Além das operações, temos diferentes ferramentas e tecnologias OLAP, que se diferenciam na forma de acesso aos dados e forma de armazenamento. As mais utilizadas são (GINER, 2007; DÍAZ, 2012; HAN; KAMBER; PEI, 2011):

- ROLAP (OLAP Relacional): Utiliza os bancos de dados relacionais para armazenar os dados em um formato dimensional, através de tabelas fato e dimensões. São realizados acessos em bases de dados relacionais através de consultas SQL (*Structured Query Language*) para gerar as informações.
- MOLAP (OLAP Multidimensional): Utiliza estruturas de bancos de dados otimizadas para recuperação de informações. São conhecidas como bases de dados multidimensionais. O armazenamento dos dados é otimizado para uma maior velocidade no cálculo das consultas. Os dados são organizados em cubos que o usuário pode rotacionar e realizar operações OLAP.
- HOLAP (OLAP Híbrido): Os dados estão divididos em armazenagem relacional e multidimensional. O acesso aos dados de alto nível é realizado na base multidimensional e o acesso aos dados atômicos e mais detalhados é realizado na base relacional.

A Tabela 3 apresenta as vantagens e desvantagens das tecnologias ROLAP e MOLAP (GINER, 2007; DÍAZ, 2012; HAN; KAMBER; PEI, 2011).

Tabela 3 - Vantagens e desvantagens das diferentes tecnologias OLAP.

Tecnologia	Vantagens	Desvantagens
ROLAP	• Solução mais escalável que MOLAP. • Sem limitações de tamanho. • Habilidade de visualizar os dados quase em tempo real.	• O pré-processamento de grandes volumes de dados é difícil de implementar com eficiência. • Mais lento que MOLAP. • Consultas limitadas pelas funcionalidades do SQL.
MOLAP	• Formato de armazenamento otimizado para consultas e cálculos de agregações. • Otimizados para recuperação em padrões de acesso hierárquico. • Necessita de menos espaço de armazenamento para dados esparsos. • Cálculos complexos são facilitados.	• Se quisermos criar novas dimensões, deveremos recarregar o cubo de dados. • Em geral, são tecnologias proprietárias, requer investimentos. • Portabilidade de aplicações de BI é menor em ferramentas OLAP do que em bancos de dados relacionais.

Fonte: Autor (2019).

Com isso, está finalizada a seção de fundamentação teórica de BI. Na próxima seção serão analisadas ferramentas e tecnologias disponíveis que viabilizam a implementação dos métodos abordados nesta seção.

2.4 Ferramentas de BI

Para viabilizar a implementação das técnicas e conceitos de BI abordados anteriormente, são utilizados determinados conjuntos de ferramentas e tecnologias. A modelagem do DW, implementação das técnicas de integração, armazenamento, visualização e análise dos dados não são tarefas simples, portanto, é necessária uma arquitetura de software que dê suporte para a solução desde o nível do banco de dados até a interface do usuário.

A intenção desta seção é apresentar as ferramentas de BI mais utilizadas, já consolidadas e que sejam gratuitas e, preferencialmente, de código aberto (do inglês *open source*). Também serão apresentadas tecnologias auxiliares que podem contribuir no processo de desenvolvimento da solução.

2.4.1 Sistemas Gerenciadores de Bancos de Dados

Um Sistema Gerenciador de Banco de Dados (SGBD), para Elmasri e Navathe (2006, p. 4):

> [...] é uma coleção de programas que permite aos usuários criar e manter um banco de dados. O SGBD é, portanto, um sistema de software de propósito geral que facilita os processos de definição, construção, manipulação e compartilhamento de bancos de dados entre vários usuários e aplicações.

Em um estudo realizado pela empresa de consultoria da Austrália *solid IT*[1], dos SGBD relacionais mais famosos e utilizados, temos: *Oracle*, *MySQL*, *SQL server* e *postgreSQL*. Apesar de todos possuírem versões gratuitas para uso limitado, o *MySQL*, o *postgreSQL* e o *mariaDB* são os SGBD de código aberto. Porém, recentemente adquirido pela *Oracle*, o *MySQL* está sob uma licença GPL versão 2.0 - *General Public License* - que permite o uso não-comercial gratuito. Porém, o *MySQL* só pode ser usado e distribuído gratuitamente se o produto ligado a ele também estiver sob a licença GPL, além do código-fonte conter as modificações realizadas. Para uso comercial, o produto também deriva as restrições da licença GPL, ou seja, o código-fonte deve ser disponibilizado (MYSQL, 2019). O mesmo caso se aplica ao *mariaDB*, que possui uma licença GPL versão 2.0 (MARIADB, 2019). O *postgreSQL* está sob a licença BSD (*Berkeley Software Distribution*), que possui menos restrições que a GPL, no sentido de que não há a obrigatoriedade de disponibilização do código-fonte do produto final, apesar de encorajar-se essa prática (POSTGRESQL, 2019).

A escolha do SGBD relacional é uma etapa importante, pois é a tecnologia utilizada para armazenar os dados do DW/DM em sistemas ROLAP. As informações apresentadas anteriormente foram preponderantes na determinação da ferramenta utilizada.

2.4.2 Ferramentas para BI

Nesta subseção serão analisadas as tecnologias disponíveis que permitem acessar, organizar, modelar, disponibilizar, visualizar e analisar dados para um melhor entendimento dos negócios. Uma suíte de BI deve possuir meios para executar o

[1]SOLID IT. DB-Engines Ranking. 2018. Disponível em: https://db engines.com/en/ranking. Acesso em: 15 abr. 2018.

processo ETL com dados de fontes distintas e viabilizar análises que forneçam informações relevantes para suportar processos decisórios.

Serão analisadas apenas as ferramentas de BI *open source* e gratuitas. Existem diversos trabalhos na literatura científica que trazem comparações entre suítes de BI de código aberto (GAJENDRAGADKAR *et al*., 2016; GOLFARELLI, 2009; JUNIOR, 2010; MARINHEIRO; BERNARDINO, 2013). Marinheiro e Bernardino (2013) realizam comparações entre cinco suítes de BI de código aberto: *Jaspersoft*, *Palo*, *Pentaho*, *SpagoBI* e *Vanilla*. Os autores concluem afirmando que *SpagoBI* é a ferramenta que possui mais funcionalidades, sendo a melhor classificada, porém a *Pentaho* seria a mais utilizada. A Tabela 4 apresenta os resultados do autor, baseado na aplicação de sua metodologia de avaliação. Em seu trabalho, são descritas todas as definições dos termos na coluna 'capacidades'.

Tabela 4 - Comparação entre cinco suítes de BI de código aberto.

Capacidades	**Suítes de BI de Código Aberto**				
	Jaspersoft	***Palo***	***Pentaho***	***SpagoBI***	***Vanilla***
ETL	**X**	**X**	**X**	**X**	**X**
Colaboração		**X**		**X**	**X**
Reporting	**X**	**X**	**X**	**X**	**X**
Dashboards	**X**	**X**	**X**	**X**	**X**
Consultas *ad-hoc*		**X**	**X**	**X**	**X**
Integração com Office	**X**	**X**	**X**	**X**	
Mobilidade BI	**X**		**X**	**X**	**X**
OLAP	**X**	**X**	**X**	**X**	**X**
Visualização Interativa	**X**	**X**	**X**	**X**	**X**
Data Mining			**X**	**X**	**X**
Scorecards			**X**	**X**	**X**

Fonte: Adaptado de Marinheiro e Bernardino (2013).

Na análise dos resultados dos trabalhos anteriores, foi possível observar que as ferramentas *Pentaho* e *SpagoBI* receberam um maior destaque com avalições positivas. Uma característica da *SpagoBI*, em comparação com a *Pentaho*, é que ela não possui versões comerciais, é totalmente gratuita. Ambas são suítes de código aberto. Porém, uma das críticas para o *SpagoBI* é a documentação escassa, que pode

dificultar o processo de desenvolvimento e implementação de uma solução de BI em algum momento.

Apesar das diferenças, a escolha da ferramenta vai depender do contexto de aplicação da mesma, dos objetivos analíticos e das necessidades do negócio, visto que cada ferramenta possui alguma vantagem em algum aspecto que pode ser preponderante em determinada solução (GAJENDRAGADKAR *et al*., 2016). Outro ponto notório é a maturidade das ferramentas de BI de código aberto, que hoje já possuem alta confiabilidade e seu uso pode ser considerado em vez de soluções proprietárias, principalmente se o volume de dados e a carga de trabalho não forem fatores críticos na organização (GOLFARELLI, 2009).

Além das ferramentas anteriores, recentemente foi disponibilizada a suíte de BI *Knowage*[2]. *Knowage* é uma evolução da ferramenta *SpagoBI* (a partir da 6° versão), sendo a versão *Community Edition* (CE) uma alternativa adicional a ser considerada neste trabalho, por ser gratuita e *open source*.

Com isso, está finalizada a seção de Ferramentas de *BI*. Na próxima seção serão analisados os trabalhos correlatos e o estado da arte relacionados aos temas BI e DW na agricultura e pecuária.

2.5 Trabalhos Relacionados

Os trabalhos encontrados e selecionados serão apresentados na sequência. Os critérios de seleção dos mesmos é apresentado na metodologia deste trabalho.

2.5.1 Aplicações de BI/DW no exterior

Wijaya e Pudjoatmodjo (2015) abordam o desenvolvimento de um DW para o armazenamento de informações do Ministério da Agricultura da Indonésia. O ministério possui diferentes sistemas de informação para processar os dados de cada domínio em suas divisões. Os dados eram extraídos de cada sistema manualmente e utilizados no processo de tomada de decisão. Os bancos de dados eram gerenciados separadamente e localmente em cada divisão e possuíam formatos de dados não-uniformes (SGBD e planilhas eletrônicas). Essa realidade dificultava o processo de tomada de decisão dos executivos. Com o DW, os executivos puderam obter os

[2]KNOWAGE. Community edition: open source for innovation. 2018. Disponível em: <https://www.knowage-suite.com/site/licensing/community-edition/>. Acesso em: 10 abr. 2018.

relatórios desejados em um período curto de tempo além de visualizar as informações em vários formatos. O sistema carrega dados de três domínios (produção, financeiro e recursos humanos), os transforma e os armazena em um único banco de dados (DW) e, por fim, permite a visualização das informações em forma de relatórios para os usuários, através de acessos aos DM. Os autores não apresentam o processo de extração dos dados do DW para os DM e nem exemplos da aplicação que faria acesso aos dados dos DM.

Nilakanta, Scheibe e Rai (2008) abordam sobre desafios enfrentados durante o desenvolvimento das dimensões de tempo e localidade em um DW para o setor de agricultura da Índia, o *INARIS warehouse*. Na granularidade de local tem-se no mesmo nível de hierarquia: aldeias, mercados agrícolas e estações agrometeorológicas. Para operações de *roll-up*, facilmente agregam-se aldeias para o próximo nível da hierarquia (distrito) mas para mercados agrícolas, onde a condição é o preço, agregações simples não funcionariam. Ainda, nem todos distritos possuem mercados agrícolas e estações. Também, nem todos distritos produzem todos produtos. A disponibilidade dos produtos comercializados e mercados agrícolas são locais, pois dependem da área, produção e consumo. Portanto, pode não ser possível agregar estes tipos de dados para níveis superiores na hierarquia. Outro problema é a granularidade de tempo. Na Índia, existem três tipos de anos: civil, agrícola e financeiro, e cada um possui um período de início e fim. Portanto, a dimensão de tempo deverá ter diferentes hierarquias (NILAKANTA; SCHEIBE; RAI, 2008). Com isso, haverá vários períodos sobrepostos, dificultando ainda mais o processo de integração destes dados. Foi realizado o desenvolvimento e a implementação do *INARIS warehouse* e os problemas anteriores foram resolvidos usando diferentes tabelas fato para cada tipo de dimensão associada a localidade e suas hierarquias. A proposta foi avaliada através de um questionário aplicado para os usuários deste sistema e verificou-se que houve uma avaliação positiva dos mesmos com relação à satisfação com a solução. É possível realizar análises espaciais através de um SIG (Sistema de informação Geográfico), mineração de dados e consultas *ad-hoc* para um conjunto selecionado de usuários e apresentam figuras mostrando a capacidade dos relatórios gerados pela solução (NILAKANTA; SCHEIBE; RAI, 2008).

Traub *et al.* (2017) apresentam o design e arquitetura do Inventário Nacional de Florestas e Sistema de Análises (*NAFIDAS*), da Suíça. Nesta solução, os autores utilizam uma arquitetura de DM dependentes e repositório de dados operacionais

(ODS). A solução consiste em um ODS que possui uma interface com as fontes de dados internas e externas, uma área de armazenamento de dados e metadados (DW) e ferramentas de apresentação para usuários finais. O sistema é acessível através de duas aplicações web hospedadas em dois sites distintos: o website interno permite total acesso ao DW e seus conjuntos completos de ferramentas de gerenciamento e análise dos dados e o website público oferece consultas nos DM e provê um conjunto considerável dos dados do Inventário Nacional de Florestas, sem limitações de acesso, precisando apenas de conexão na internet e um navegador web. O NAFIDAS, após dez anos de desenvolvimento e melhorias, atingiu um estado estável e maduro, possibilitando consultas em múltiplos idiomas e o sistema de filtro na web provê acesso à grandes volumes de dados (TRAUB *et al*., 2017). O sistema é classificado como um sistema de apoio à informação e decisão devido à confiabilidade e ao suporte sustentável de estimativas de população florestal de alta qualidade. Por fim, são apontados como fatores chave para o sucesso dessa solução: a escolha da arquitetura de DW, controle de qualidade do armazenamento das informações e o suporte do gerenciamento (TRAUB *et al*., 2017).

No trabalho de Vernier *et al*. (2013) foi desenvolvido um sistema de informação ambiental (*Environmental Information System* – EIS) para avaliação do uso de pesticidas, em que foi utilizada a tecnologia de DW espacial (*Spatial* DW - SDW). O objetivo era fornecer uma ferramenta para avaliar os impactos dos pesticidas da agricultura em diferentes escalas em bacias hidrográficas. O estudo de caso foi feito com o Rio Charente, do sudoeste da França. Foram utilizadas bases de dados de pesquisas da bacia hidrográfica deste rio, assim como dados estatísticos de outras fontes para caracterizar a agricultura local e condições ambientais. As dimensões e métricas foram definidas, o SDW foi modelado, preenchido com os dados, e por fim, foi utilizada uma ferramenta OLAP espacial (*Spatial* OLAP - SOLAP) para análise dos dados. Os autores concluem afirmando que a tecnologia de DW pode ser útil no auxílio à tomada de decisão para planos de ação agroambientais e na avaliação dos impactos potenciais da evolução da agricultura na área, permitindo que os responsáveis pelas políticas públicas possam tomar decisões mais informadas ao gerenciar áreas de risco ou ao implementar medidas de mitigação.

Nielsen (2011) descreve o propósito do uso de um DW na área da saúde e bem-estar animal. Para o autor, a Dinamarca seria o primeiro país a relacionar todos os dados relativos à pecuária, saúde e bem-estar dos animais de produção

dinamarqueses entre si em um DW. É ressaltada a importância do acesso aos dados para pesquisadores de estudos epidemiológicos em saúde e bem-estar animal e o gerenciamento de risco e comunicação de risco pelas autoridades competentes. Os dados utilizados seriam oriundos de bases de dados do ministério da alimentação, agricultura e pesca, e de outras fontes com dados relacionados ao bem-estar animal. Dados de produção e bem-estar animal de entidades privadas poderão ser agregados futuramente. Também é apresentada uma preocupação sobre a validade dos dados, apontando como soluções relatórios contínuos dos dados, controle sistemático das possibilidades de entrada dos dados e controle cruzado com os dados de outras fontes. O autor não apresenta nenhuma referência sobre a arquitetura de DW que seria utilizada, nem as tecnologias adicionais sobre os repositórios ou análise de dados que viabilizariam a proposta.

No trabalho de Rai *et al*. (2008) é abordada a construção de um DM para recursos animais na Índia. As informações disponíveis foram extraídas de vários sistemas: Banco de dados das raças de gado, da população pecuária, de produção e infraestrutura de gado, de produtos da pecuária e de importação e exportação de gado. Foram apresentados os diagramas das três dimensões do DM: animal, tempo e localização. Também foi apresentado o diagrama da tabela fato e a medida "quantidade de produtos pecuários" como leite, carne, lã, ovos, entre outros. Os dados das diferentes fontes são extraídos para a *stage area,* onde é tratada a consistência e uniformidade dos dados antes de carrega-los para o DW. É apresentado como as informações de população pecuária podem ser acessadas através de cubos multidimensionais (OLAP) em um navegador. Também existem diversas operações OLAP disponíveis para os usuários, assim como filtros, geração de relatórios nos formatos PDF (*Portable Document Format* – Formato de Documento Portátil) e HTML (*Hypertext Markup Language* – Linguagem de Marcação de Hipertexto) além de análise espacial das informações no DM através do acesso em um SIG, semelhante ao apresentado em Nilakanta, Scheibe e Rai (2008). Por fim, é mencionado como um dos fatores mais influentes no processo de desenvolvimento da solução a diversificação e complexidade do setor na Índia.

2.5.2 Aplicações de BI/DW no Brasil

No trabalho de Tech *et al*. (2010) é abordado um modelo de gestão para aplicação no agronegócio baseado em *e-science* e DW. Na solução, é apresentado

um processo de coleta e monitoração de rebanhos, permitindo que gestores possam controlar o rebanho de maneira segura e confiável sem causar desconforto ao animal. Os dados coletados são armazenados no banco de dados do *e-science* zootécnico, que futuramente poderá ser explorado para a utilização de DW para os dados serem disponibilizados para análises para pesquisadores, através de acesso remoto. Apesar dos autores mencionarem no trabalho, não é apresentado nenhum aspecto da modelagem e implementação do DW, assim como os seus componentes ou tecnologias utilizadas. A implementação do DW ficou como sugestão futura.

Ferreira e Camargo (2013) apresentam um estudo de caso da implementação de um DW em uma cooperativa na Região Sul do Rio Grande do Sul (RS), onde 44% do faturamento advém de produtos agrícolas. Foram utilizados e analisados os dados de recebimento de cereais, que possuíam potencial para auxiliar em processos decisórios. Foi realizado um pré-processamento dos dados, através de transformações, correções e eliminação de atributos desnecessários. Posteriormente os dados foram extraídos do ambiente transacional para o dimensional, através de um aplicativo na linguagem *Delphi*. O ambiente do DW foi baseado na tecnologia de banco de dados *MySQL* 5.6 e a ferramenta de BI utilizada foi a *SpagoBI*. Os autores apresentam os modelos relacional e dimensional da solução, assim como todos os relatórios gerados pela solução de BI proposta, concluindo que estes relatórios viabilizaram mudanças e a elaboração de políticas mais agressivas pelos gestores da cooperativa.

No artigo de Mota *et al*. (2017), foi desenvolvida uma solução de análise de dados através do uso de DW, consultas OLAP e mineração de dados, de forma integrada, para subsidiar decisões na pecuária de corte. O objetivo era tentar predizer o melhor momento de abate, com base nos dados disponíveis. A solução é segmentada em quatro etapas. A primeira é a extração dos dados de três bases de dados (*PostgreSQL*): banco de dados de abate, de dados financeiros e uma planilha de regras técnicas. O dito processo ETL foi executado usando a ferramenta *Pentaho Data Integration*. A segunda etapa foi a modelagem dimensional, definição das dimensões, métricas e fatos do modelo estrela do DW, que foi hospedado no SGBD *PostgreSQL*. Na terceira etapa foram definidas as ferramentas de acesso aos dados do DW: *Pentaho Business Intelligence Server*, *Schema Workbench* e *Saiku Analytics.* Diversas formas de relatórios, painéis e gráficos foram apresentadas com o uso de tais ferramentas, permitindo uma rica experiência visual para os usuários. A etapa final

abordou o processo de mineração de dados do DW através da ferramenta *WEKA*, com o objetivo de predizer o melhor momento de abate. Os autores concluem afirmando como efetiva a utilização de um portal Web para consumo e exploração dos dados, através das técnicas e tecnologias abordadas anteriormente, assim como deve ser considerada a qualidade dos dados, relacionado aos aspectos de precisão e consistência (MOTA *et al*., 2017).

Em Moreira, Martinhago e Drummond (2015) é abordado o desenvolvimento de um sistema de gestão para apoio à tomada de decisão no agronegócio, da região do Alto Paranaíba, através de técnicas de BI. Na solução, foram extraídos dados de duas fontes. Uma das fontes é a estação meteorológica instalada no Campus I UFV Rio Paranaíba (dados exportados em forma de planilhas eletrônicas). Cada registro possuía informações de data, hora, temperatura máxima, temperatura mínima, velocidade do vento, direção do vento, humidade relativa do ar e evapotranspiração. A outra fonte de são os dados quantitativos da produtividade de culturas de cenoura e beterraba de fazendas da zona rural dos municípios Rio Paranaíba e São Gotardo, ambos de Minas Gerais, cedidos pelo Grupo *Sekita* Agronegócios. Foi realizado o tratamento dos dados para correção de incompletudes. Nesta fase, foi utilizada a ferramenta *Pentaho Data Integration*. Na sequência, foi modelado o *Data Mart* para receber estes dados. Foi adotado o modelo *star schema* e a ferramenta de modelagem foi o *MySQL Workbench*. O SGBD utilizado foi o *PostgreSQL*, porque que o mesmo possui compatibilidade com a suíte e os *plug-ins* do *Pentaho,* além de ser *open source* e ter um bom desempenho. Na etapa final do trabalho, foi configurado o ambiente para visualização e análise dos dados. Foram utilizadas as ferramentas *Pentaho Business Analytics* para configuração do ambiente, o módulo *Mondrian* da suíte *Pentaho* foi usado para criação dos cubos OLAP e o *Pentaho BI Server* foi utilizado para a publicação dos cubos, para posterior análise e visualização das informações. Por fim, os autores apresentam o modelo dimensional final em detalhes, o processo de ETL e os relatórios gerados pelo *Pentaho BI Server* para dados de produção da cenoura e algumas associações com área plantada e dados climáticos. Os autores concluem que o sistema desenvolvido pode auxiliar gestores rurais da região do Alto Paranaíba nos processos decisórios (MOREIRA, MARTINHAGO e DRUMMOND, 2015).

Correa *et al*. (2009) aborda a criação de um DW para o mercado de grãos do Brasil, usando informações sobre preços e regiões da soja e milho. Foi criado um modelo dimensional baseado no *star schema* para estas informações no SGBD

MySQL. Posteriormente foi executado o processo ETL para carga dos dados das duas fontes operacionais para o DW. O ETL foi realizado através de scripts na linguagem *Perl*. Por fim, foi utilizada a ferramenta OLAP *Mondrian* para a análise e visualização das informações. É apresentado um gráfico gerado pela ferramenta, que mostra os preços da soja e milho em três diferentes regiões do Brasil. Os autores concluem afirmando que tecnologias de DW permitem cruzamentos e navegação dos dados de forma fácil e flexível, permitindo a obtenção de dados dinâmicos de séries históricas dos preços do mercado de grãos. O sistema não está concluído, porém é previsto que sejam acrescentadas outras cadeias do mercado agrícola no modelo dimensional (CORREA *et al*., 2009). Também é prevista a utilização do software *Talend* para o processo ETL e aplicação de técnicas de mineração de dados no futuro (CORREA *et al*., 2009).

2.5.3 Síntese

Comparado ao que era esperado, foram encontrados poucos trabalhos na literatura científica que utilizem de tecnologias de *DW* e *BI* para soluções na agropecuária e áreas afins, de uma forma geral. Acredita-se que isso se deve ao fato de que o setor ainda está em fase inicial de adoção de tecnologias contemporâneas para soluções de problemas rotineiros do agronegócio. Dentre estes, subsidiar os processos decisórios com dados e informações. Estes dados e informações devem ser de qualidade, característica que deve ser garantida em todo o ciclo de vida do dado, desde o momento de coleta dos dados, até o momento de disponibilizar as informações para os usuários. Observou-se, ainda, uma predominância maior da menção às tecnologias de DW sobre as de BI na área agro.

A maioria dos trabalhos citados mencionou alguma preocupação com aspectos de qualidade dos dados. Foram mencionados, em sua maioria, problemas de inconsistência e incompletude dos dados nos processos de integração dos dados. Estes problemas estão relacionados ao aspecto da qualidade dos dados. Dados de péssima qualidade podem comprometer todas as etapas futuras do desenvolvimento de uma solução para o suporte à decisão, já que estes perderiam o seu propósito.

Outro aspecto percebido foi a diferença de maturidade e consolidação do uso de repositórios de dados preponderantes para processos analíticos na área agro no Brasil e no exterior. Muitos dos trabalhos realizados no Brasil apresentam DM/DW com apenas uma área do negócio em análise, sendo assim, soluções pontuais, com

pouco suporte à análises holísticas. Já nos trabalhos desenvolvidos no exterior, percebe-se, pelos trabalhos publicados, que os repositórios de dados para análises já abrangem uma grande parte ou todas as áreas de negócio de determinados setores.

Uma das limitações das análises realizadas é que não foram investigadas soluções da iniciativa privada neste trabalho, que podem estar apontando para o uso destas tecnologias no setor. Apenas foram investigadas as bases de dados científicas. Sugere-se, como trabalho futuro, a investigação da aplicação destas soluções tecnológicas pela iniciativa pública e privada, em escala nacional e internacional, no setor agropecuário.

Nas subseções anteriores, foram apresentados dez trabalhos, sendo cinco deles realizados em países no exterior e cinco realizados no Brasil. A Tabela 4 apresenta um resumo das informações destes artigos. Ainda, foi realizada uma pesquisa complementar com um maior nível de profundidade e detalhamento sobre os artigos que abordam a temática “DW e BI aplicados no setor agropecuário”, que pode ser conferida em Brum, Lampert e Camargo (2019), onde resultados interessantes poderão ser encontrados. Tal pesquisa também serviu de fundamentação teórica para o desenvolvimento desta proposta.

Por fim, encerrada a seção de trabalhos correlatos, encerra-se o capítulo de referencial teórico. No próximo capítulo será descrita a metodologia que será adotada nesta proposta para atingir os objetivos propostos.

Tabela 5 - Resumo dos trabalhos encontrados na investigação literária.

Artigo	Qualis	Foco	Síntese
WIJAYA; PUDJOATMODJO, 2015	*	Agricultura	Uso um DW para prover dados atualizados e integrados de três fontes de dados para subsidiar decisões de executivos do departamento de agricultura da Indonésia.
NIKALANTA; SCHEIBE; RAI, 2008	A2	Agricultura	Desenvolver um DW governamental para o setor da agricultura da Índia.
TRAUB et al., 2017	A2	Silvicultura	Uso de DW e sistemas analíticos para o inventário nacional de florestas da suíça.
VERNIER et al., 2013	A1	Agricultura	Desenvolver uma ferramenta efetiva para avaliar os impactos da agricultura em diferentes bacias hidrográficas na França.
NIELSEN, 2011	B1	Pecuária	Uso de DW com dados integrados de dez bases de dados para otimizar o bem-estar animal da bovinocultura e suinocultura na Dinamarca.
RAI et al., 2008	A2	Pecuária	Uso de DM para integração de dados de cinco bases de dados de recursos animais na Índia.
TECH et al., 2010	B2	Pecuária	Uso de DW e *e-Science* para gestão focada na qualidade e produtividade animal.
FERREIRA; CAMARGO, 2013	B5	Agronegócio	Implementar um DW em uma cooperativa de produtos agrícolas no sul do RS.
MOTA et al., 2017	B5	Pecuária	Suportar a tomada de decisão no setor da pecuária de corte através de DW, consultas OLAP e mineração de dados integrados.
MOREIRA; MARTINHAGO; DRUMMOND, 2015	B5	Agricultura	Uso de DM para armazenar dados climáticos e de produção de culturas para facilitar a análise e visualização das informações e auxiliar em processos decisórios.
CORREA et al., 2009	*	Agronegócio	Uso de DW para o mercado de grãos do Brasil, usando informações sobre preços e regiões da soja e milho

Fonte: Autor (2019).

*Trabalhos publicados em conferências internacionais (ICoICT e EFITA).

3 METODOLOGIA

Neste capítulo são descritos os procedimentos metodológicos adotados no trabalho. Na seção 3.1 é realizada a caracterização da pesquisa em diferentes aspectos, assim como a definição de suas fases, e na seção 3.2 é apresentada a sequência de etapas de desenvolvimento da solução e uma descrição e discussão dos aspectos gerais de cada uma delas.

3.1 Caracterização da pesquisa

- Quanto à natureza: a pesquisa é de natureza aplicada. Existem diversas fontes de dados que foram desenvolvidas para dar suporte a problemas específicos e que não permitem uma visão ampla da realidade. Portanto, deve-se permitir que os clientes possam obter uma visão holística do seu estabelecimento rural. Isto será realizado através da integração dos indicadores relacionados com os custos (por exemplo: custos de produção, total de receita e custo de atividade de cria), produção (por exemplo: taxa de desfrute, produtividade, idade de abate, taxa de mortalidade) e meio ambiente (por exemplo: emissão de dióxido de carbono), que são as especialidades de cada um dos sistemas.
- Quanto ao tipo: a pesquisa é do tipo exploratória. Um maior conhecimento sobre o fenômeno em estudo é propiciado ao pesquisador, permitindo elaborar hipóteses e realizar pesquisas futuras mais precisas (LAKATOS; MARCONI, 2003).
- Quanto à origem dos dados: o FGC, por estar em fase de desenvolvimento, atualmente não possui dados coletados. Ainda, gestão de custos é a principal preocupação do pecuarista brasileiro neste momento (EMBRAPA, 2018), refletindo a necessidade de compreender melhor a sistemática de registro de receitas e despesas da propriedade rural. Portanto, foi realizada a simulação dos dados de custos. Para isso, serão utilizados parâmetros da realidade da pecuária de corte brasileira. Os dados relacionados aos indicadores produtivos e ambientais foram coletados de produtores rurais da pecuária de corte, através de um questionário *online* na ferramenta *Google Docs*. Porém, este processo não foi parte desta pesquisa.
- Unidade observacional: é o sujeito de pesquisa, o objeto de investigação. Os dados obtidos foram coletados por meio de questionários, para produtores

rurais, em que é posto em foco o sistema produtivo. Portanto, a unidade observacional é o sistema produtivo da pecuária de corte.

As fases de uma pesquisa, para Lakatos e Marconi (2003) envolvem a definição do tema de pesquisa, levantamento de dados, formulação do problema, definição dos termos, construção de hipóteses, indicação de variáveis, delimitação de pesquisa, amostragem, seleção de métodos e técnicas, organização do instrumental de pesquisa e teste de instrumentos e procedimentos. São detalhadas na sequência as três primeiras fases. Na próximas seções são abordadas as etapas metodológicas para o desenvolvimento deste trabalho. A construção de hipóteses e indicação de variáveis não foram realizadas por não serem adequadas para esta natureza de pesquisa.

Definição do tema de pesquisa: Em reuniões conjuntas com os orientadores, o tema escolhido foi a construção de um processo de integração de dados de dois diferentes SAD para produtores rurais. Cabe mencionar que este é um Mestrado Acadêmico de Computação Aplicada na resolução de problemas das Ciências Agrárias. Logo, o tema deve, de alguma forma, envolver algum aspecto emergente da agricultura, pecuária ou áreas relacionadas.

Levantamento de dados: Foi realizada uma extensa investigação na literatura científica sobre os conceitos e tecnologias envolvidas com integração de dados de fontes distintas e heterogêneas. Em particular, repositórios com dados de atividades de agricultura, pecuária ou áreas afins. Esta foi a fase de construção do referencial teórico da proposta, através de pesquisa bibliográfica.

Formulação do problema: o problema que motivou o desenvolvimento desta pesquisa é a seguinte pergunta: Como realizar o processo de integração de dados de fontes distintas e heterogêneas de forma a garantir a qualidade dos dados e para fornecer aos usuários (produtores rurais e consultores) informações que subsidiem os processos decisórios?

Com isso, o próximo passo que deve ser realizado é o planejamento do desenvolvimento da solução do problema.

3.2 Definição das etapas de desenvolvimento da solução

A Figura 18 apresenta graficamente cada uma das etapas de desenvolvimento propostas, com o subsídio dos conceitos apresentados na revisão de literatura, para

a solução dos problemas apontados na introdução. As etapas destacadas com a fonte na cor amarela farão parte de um ciclo de iterações. Serão abordadas, em detalhes, cada uma destas etapas logo a seguir, assim como as iterações.

Figura 18 - Etapas para o desenvolvimento da solução.

Fonte: Autor (2019).

Delimitação do Problema

A EMBRAPA teve sua Lei de criação n° 5851/72 decretada em 7 de dezembro de 1972 e foi criada em 26 de abril de 1973 e é vinculada ao Ministério da Agricultura, Pecuária e Abastecimento (MAPA). A EMBRAPA, em conjunto com o Sistema Nacional de Pesquisa Agropecuária (SNPA), assumiu o desafio de desenvolver um modelo de agricultura e pecuária tropical genuinamente brasileiro, superando as barreiras que limitavam a produção de alimentos, fibras e energia no Brasil[3]. A EMBRAPA proporciona treinamentos, consultorias, mapeamentos e serviços web; Desenvolvimento de procedimentos, protocolos e metodologias; Soluções tecnológicas estruturadas fisicamente oferecidas ao mercado ou à sociedade em geral.

[3]EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. Quem somos. Brasília, [s.d]. Disponível em: https://www.embrapa.br/quem-somos. Acesso em: 15 ago. 2018.

Uma das unidades descentralizadas da EMBRAPA é a EMBRAPA Pecuária Sul, localizada no município de Bagé-RS, que há mais de três décadas de existência vem trabalhando para a agropecuária brasileira, disponibilizando tecnologias nas áreas de bovinocultura de corte, de leite e ovinos, buscando o bem-estar socioeconômico do homem, com o foco no agronegócio[4].

Um dos projetos atuais da EMBRAPA Pecuária Sul é o *MyBeef.* O objetivo deste projeto é proporcionar aos pesquisadores das ciências agrárias e do setor rural um maior conhecimento sobre a cadeia produtiva da pecuária e dos territórios da região Sul do Brasil. A ideia é que sejam incorporadas e integradas diversas funções e funcionalidades, conforme mencionado no site do projeto da EMBRAPA:

> [...] espera-se produzir boletins, periódicos com informações sobre a cadeia produtiva, website com um banco de dados com softwares nacionais sobre a pecuária de corte e leite, matriz de indicadores de sustentabilidade para a pecuária, software para análise econômica e ambiental, método para análise espaço-temporal de indicadores sociais, econômicos, ambientais e produtivos e uma plataforma para abrigar sistemas de apoio à decisão para a pecuária, tendo como público-alvo prioritário o produtor rural. (EMBRAPA, [2015?]).

A EMBRAPA e as IES parceiras estão atualmente desenvolvendo diferentes SAD que visam dar suporte a dados que atendem o agronegócio sob diferentes perspectivas. Porém, os SAD estão em diferentes fases de desenvolvimento. Para esta proposta, foram selecionados os dois SAD que estão em um estágio mais avançado de desenvolvimento, a FGC e o LS. Em vista disso, a solução de integração de dados é direcionada para esses sistemas, que possuem um modelo mais consolidado da base de dados.

Construção do Arcabouço Teórico

Foi realizada uma investigação na literatura científica seguindo uma metodologia baseada nos seguintes passos:

1. Foram investigados periódicos com estrato igual ou superior ao B1 e conferências internacionais, nas áreas de ciência da computação e ciências agrárias.

[4]EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. Embrapa pecuária sul: apresentação. Brasília, [2018?]. Disponível em: <https://www.embrapa.br/pecuaria-sul/apresentacao>. Acesso em: 18 jul. 2018.

2. Definição das palavras-chave nas buscas por trabalhos: *Data Warehouse*, *Business Intelligence*, *Agriculture*, *Data Integration* e *Livestock*.

3. Critérios: Seleção de trabalhos por título, que abordem o uso de tecnologias de suporte à decisão que envolvam BI ou DW na agricultura, pecuária ou áreas afins no exterior. Foram selecionados apenas os trabalhos realizados nos últimos dez anos (2008-2018).

4. Mesma investigação do passo anterior, porém para soluções no Brasil.

5. Caso não sejam encontrados pelo menos 5 trabalhos nas etapas 3 e 4, serão investigados periódicos e conferências de extratos inferiores, do B2 ao B5.

6. Leitura do resumo dos artigos encontrados. Selecionar os mais relevantes e correlacionados com os objetivos desta proposta.

7. Por fim, leitura dos artigos selecionados para fomentar a elaboração e discussão do estado-da-arte do tema pesquisado.

Esta metodologia foi adotada visando garantir que fossem abordados trabalhos com a maior relação possível com o problema de pesquisa, além de contemporâneos e de qualidade. Essa é a razão do estabelecimento dos critérios.

Com o propósito de se ter uma visão geral sobre as abordagens adotadas para a solução de problemas correlatos, foram também investigadas as soluções sob duas perspectivas: trabalhos realizados no exterior e trabalhos realizados no Brasil, com o objetivo de verificar se haviam diferenças significativas nas tecnologias ou métodos utilizados para resolver os problemas de integração de dados de agricultura/pecuária. Não foram detectadas diferenças relevantes nestes aspectos. Grande parte das soluções encontradas convergiram para o uso de técnicas de BI, através de DW, ETL e OLAP, para realizar integração de dados de fontes distintas e heterogêneas, além de permitir a análise e visualização estratégica das informações, visando o suporte de processos decisórios.

Definição dos *Stakeholders*

O *stakeholder* é a pessoa ou grupo de pessoas que exerce influência direta ou indireta sobre os requisitos do sistema (SOMMERVILLE, 2011). Ainda, existem dois tipos de *stakeholders*. O *stakeholder* interno é a pessoa de interesse na solução que é interna à organização. Logo, o pesquisador Dr. Vinicius do Nascimento Lampert é o *stakeholder* interno, pois é o responsável pelo desenvolvimento dos SAD na

EMBRAPA. Também são considerados *stakeholders* internos os desenvolvedores dos sistemas fontes. Os *stakeholder*s externos são as pessoas de interesse na solução que são externas à organização. Neste caso, os *stakeholders* externos são os produtores e consultores rurais, que serão os principais usuários dos SAD.

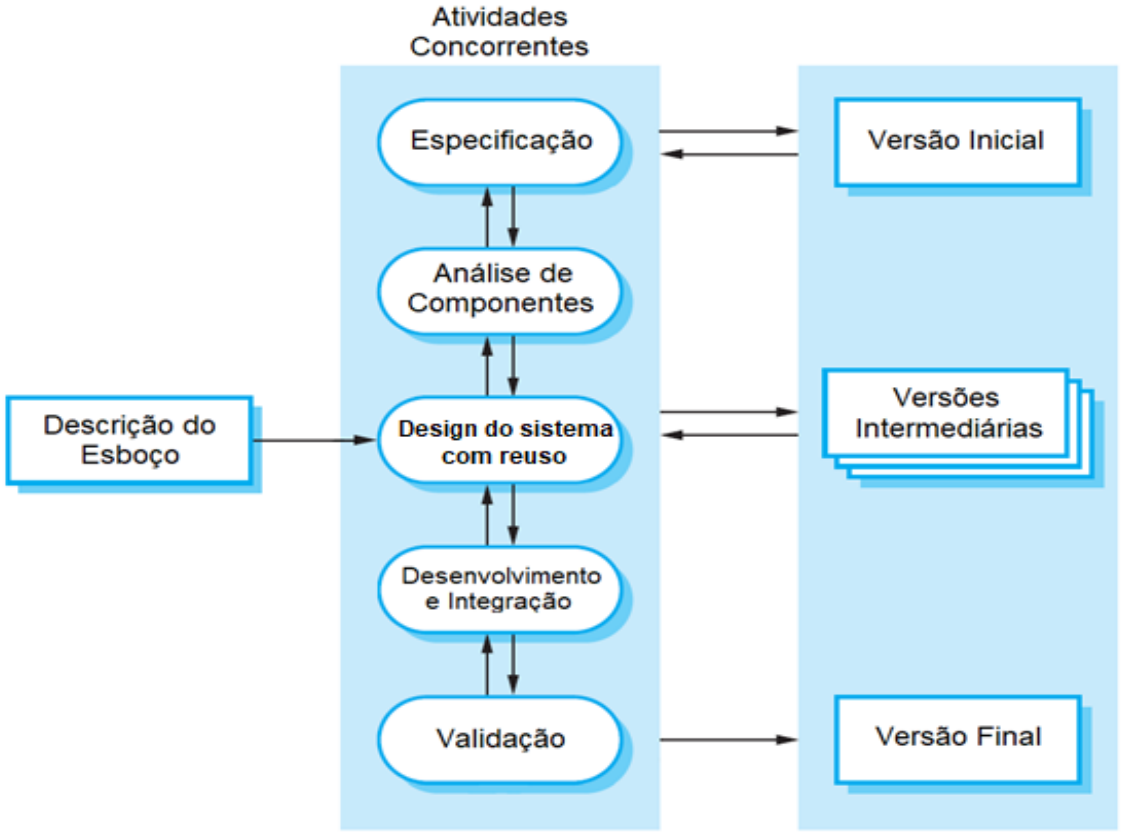
Convém mencionar que na execução das etapas posteriores, foi mantido um cronograma de encontros presenciais ou não-presenciais, minimamente quinzenal, com os *stakeholders* internos, de acordo com a disponibilidade de cada um. O objetivo foi garantir o desenvolvimento de uma solução, além de cientifica e tecnicamente correta, que atenda às necessidades da EMBRAPA. Tais encontros são necessários devido à grande quantidade de atributos e relacionamentos que o especialista e os desenvolvedores responsáveis pelos SAD possuem domínio e o completo entendimento.

Definição do Processo de Engenharia de Software

Os modelos de processos de software utilizados como base para o desenvolvimento da solução foram os processos interativos e engenharia de software orientada ao reuso (SOMMERVILLE, 2011). Modelos como o cascata, em que são desenvolvidas atividades de forma sequencial, não são adequados para esta natureza de projeto, em que os sistemas que irão abastecer a solução com os dados estão em pleno desenvolvimento e implementação, acarretando em mudanças nos requisitos.

Por que o projeto foi baseado em duas metodologias? A natureza deste projeto envolve a adaptação e desenvolvimento de soluções de forma incremental, com as atividades de especificação, desenvolvimento e validação (com *stakeholder* interno) de forma paralela. Ao mesmo tempo, o projeto envolve a reutilização de componentes e softwares *open source* e gratuitos que já atendem de forma parcial os requisitos definidos pelo cliente. As vantagens de ambas abordagens são o baixo custo de implementação de mudanças nos requisitos, obtenção mais simples e direta do *feedback* do cliente para cada incremento desenvolvido, entrega e desenvolvimento mais rápido de software útil ao cliente (mesmo que o sistema completo ainda não tenha sido desenvolvido), economia de tempo em relação à quantidade de software desenvolvida e redução de custos e riscos. A Figura 19 apresenta o modelo de software proposto e adaptado para esse projeto.

Figura 19: Modelo de processo de software utilizado.



Fonte: Autor (2019).

A descrição do esboço da solução foi realizada na introdução e justificativa deste projeto. A etapa de especificação do software e respectivas mudanças nos requisitos é abordada e detalhada no documento de requisitos elaborado, apresentado em maiores detalhes na seção 4.2 e no apêndice A. A etapa de análise de componentes, que envolve a pesquisa por ferramentas e tecnologias para BI e DW que atendam as especificações e requisitos de software, é abordada na seção 4.3. Usualmente, não existe uma combinação exata de componentes de software que atendam todas as especificações, portanto, eventualmente, adaptações foram realizadas. A etapa de design do sistema com reuso é apresentada na subseção 4.4.2. O desenvolvimento e a integração de componentes são abordadas em detalhes nas seções 4.3, 4.4 e 4.5. A análise das funcionalidades de acesso aos dados do DW é abordada na seção 4.6 e a realização da análise crítica da solução, é abordada no capítulo 5.

Análise dos Modelos dos Bancos de Dados dos SAD

Nesta fase, são obtidos os modelos conceituais, lógicos ou físicos dos bancos de dados dos SAD com o *stakeholder* interno. O objetivo desta etapa é realizar um estudo preliminar das tabelas, atributos e relacionamentos existentes nestes sistemas para subsidiar as discussões e decisões das próximas etapas. Além dos modelos, caso exista alguma documentação dos sistemas, essa também deverá ser analisada.

Definição de Requisitos

Através de conversas com o *stakeholder* interno, foram definidos os requisitos funcionais e não-funcionais da solução de BI proposta. O artefato resultante desta atividade é um documento formal de requisitos de software, para fins de registro e auxílio para futuros profissionais que venham a contribuir na continuação ou manutenção do sistema. Estes requisitos norteiam o processo de escolha das ferramentas e tecnologias de BI utilizadas para o desenvolvimento da proposta, assim como a execução dos métodos específicos de desenvolvimento nas fases de implementação.

Análise e Seleção de ferramentas para implementação da solução

Com a definição dos requisitos da proposta elaborada, são investigadas as principais tecnologias e ferramentas do mercado que viabilizam o desenvolvimento e a implementação de projetos de BI.

Na seção 2.4 foram apresentados estudos recentes sobre as tecnologias de SGBD e suítes de BI de código aberto e comparações. Logo, dados os resultados apresentados, foi investigada a aplicabilidade das ferramentas *Pentaho* e *SpagoBI* na modelagem e desenvolvimento do DW e etapas posteriores, assim como a utilização de ferramentas complementares que auxiliem na implementação da solução.

Das alternativas de SGBD analisadas, optou-se pela utilização do *PostgreSQL* por ser, além de código-aberto, gratuito e também oferece uma extensão espacial (*PostGIS*), caso seja necessário lidar com dados espaciais no futuro do projeto. Limitações de desempenho e outras características técnicas dos SGBD foram desconsideradas neste momento, pois a prioridade é verificar como técnicas de BI podem resolver o problema existente.

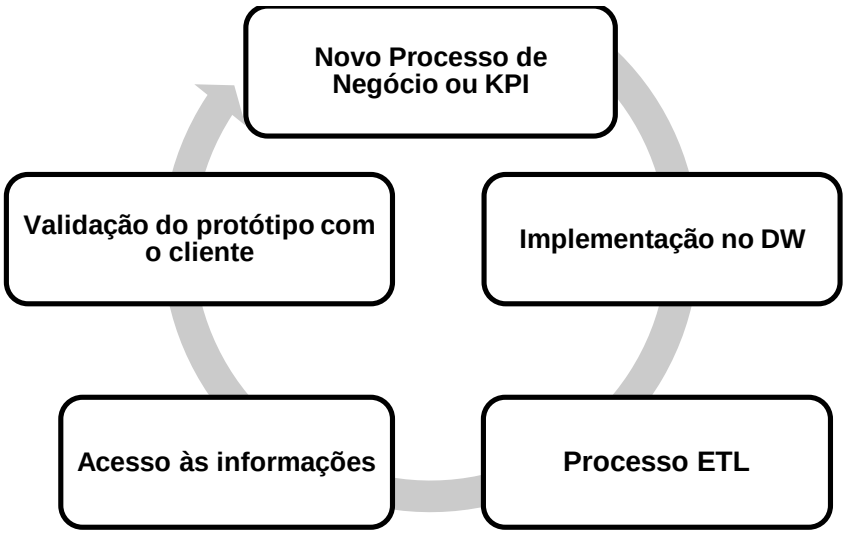
Definição dos Processos de Negócio

O processo de desenvolvimento de uma solução que envolve o projeto de um DW é iterativo. O objetivo de projetos com esta característica é a rápida entrega de funcionalidades para o *stakeholder* (CORR; STAGNITTO, 2011), neste caso, as funcionalidades do BI através do projeto e implementação do DW e execução do ETL. Isso permite a utilização imediata da solução, ao mesmo tempo em que componentes adicionais são projetados em paralelo pela equipe de desenvolvimento. Outra característica deste processo é a avaliação imediata dos clientes, se é o produto certo que está sendo desenvolvido. Isso permite que erros sejam corrigidos logo nas fases iniciais de implementação.

Logo, nesta etapa inicia-se o processo interativo. Foi adotada a metodologia de desenvolvimento ágil de DW, abordada por Corr e Stagnitto (2011). O mínimo que deve ser entregue por iteração é um *star schema* simples, os processos de ETL que o preenchem, uma ferramenta de BI configurada para acessa-lo e validar o protótipo entregue. A Figura 20 ilustra os passos iterativos adaptados à proposta:

A definição dos processos de negócio ou KPI foi realizada em conjunto com o *stakeholder* interno, em paralelo com a realização da análise das estruturas dos bancos de dados dos sistemas fontes. Os processos de negócio vão, além de orientar o processo de modelagem dimensional do DW, evidenciar para os *stakeholders* externos a distância da empresa rural de atingir os seus objetivos.

Figura 20 – Etapas do processo de desenvolvimento da solução.



Fonte: Autor (2019).

Seguindo o método de desenvolvimento ágil, inicialmente foram definidas as métricas ou conjuntos de KPI, usando como base as informações e atributos já existentes nos sistemas legados (SAD). Isso dará subsídios para a execução das próximas etapas do projeto.

Definição da Arquitetura e Modelagem do DW

Nesta etapa, com base nos requisitos estabelecidos, nas análises dos sistemas fontes e encontros com o *stakeholder* interno, foi definida a arquitetura de DW mais adequada para esta solução. Diferentes arquiteturas e suas características foram apresentadas na subseção 2.3.3.2.

Por fim, com base nas métricas selecionadas, iniciou-se a modelagem do DW. Este processo é dividido nas seguintes etapas:

- Modelagem conceitual: descrição do DW de forma independente da tecnologia de SGBD utilizada, que deverá incluir apenas os nomes das tabelas, as métricas do negócio e as perspectivas de análise destas métricas.
- Projeto Lógico: Descrição de como o DW será implementado no SGBD relacional selecionado. Neste projeto, será realizada a modelagem dimensional no *PostgreSQL*. Neste ponto, podem ser utilizados os esquemas estrela, floco de neve ou constelação de fatos. Após a escolha do esquema, são definidas as suas tabelas fato e dimensões e seus relacionamentos.
- Projeto Físico: Serão considerados aspectos de indexação, redundância dos dados, materialização das visões, particionamento, paralelismo, entre outros, que são importantes no projeto físico de um DW. Além destes, também serão avaliadas restrições do negócio e limitações, decididas conjuntamente com o *stakeholder* interno.

A granularidade dos dados é outro importante aspecto que foi definido na etapa de modelagem do DW. Os níveis de granularidade possíveis foram analisados após a análise das estruturas das bases de dados dos sistemas legados.

Após a execução destas etapas, o repositório de dados ficou disponível para consultas por ferramentas OLAP. Porém, ainda era necessário extrair os dados dos sistemas legados para o DW, que foi o foco da próxima etapa.

Processo ETL

Nesta etapa foram realizados todos os processos de extração dos dados dos sistemas legados para a *stage area (*que por sua vez também foi criada), transformações e eventuais correções nos dados, a carga dos mesmos para o DW e, por fim, a definição da periodicidade de carga dos dados. A *stage area* possui um modelo correspondente ao dos bancos de dados de origem, porém guarda só os atributos relevantes para análise e necessários para o DW. Com o DW povoado com os dados necessários, ferramentas OLAP podem acessá-lo na próxima etapa.

Uso de Ferramentas OLAP

Nesta etapa foi realizada a exploração dos dados do DW através de ferramentas de acesso à informação, como OLAP e relatórios. Diferentes tecnologias serão utilizadas e analisadas, com o objetivo de entregar aos *stakeholders* internos e externos uma solução que permita um suporte eficiente aos processos decisórios.

Com o atendimento da métrica modelada no DW, iniciou-se o processo de avaliação do protótipo desenvolvido, pelo *stakeholder* interno. Caso o protótipo tenha sido validado, inicia-se um novo ciclo iterativo, a partir da modelagem de outro esquema para uma nova métrica. Este ciclo se repete para que sejam atendidas todas as métricas de negócio necessárias para a solução de BI, considerando as limitações dos prazos estipulados para a execução desta e outras etapas do projeto. Após o final deste ciclo, iniciou-se o processo de avaliação da solução desenvolvida.

Como decisão de projeto, apesar das desvantagens relacionadas à ausência de clientes durante o desenvolvimento de um software/sistema, optou-se por não realizar a avaliação de cada protótipo desenvolvido com os *stakeholders* externos devido às restrições de tempo para a elaboração do projeto, desenvolvimento e avaliação da proposta. Outro fator preponderante nesta decisão foi que, para viabilizar a análise dos *stakeholder* externos para cada protótipo ou para a solução integral, seriam necessários dados suficientes de produtores que tenham utilizado tanto o sistema LS como o FGC, ambos em processo de desenvolvimento. Obter o contato e dados de produtores rurais em diversas oportunidades para ambos os sistemas para cada protótipo desenvolvido levaria um tempo considerado proibitivo, dado o tempo disponível para desenvolver o projeto da dissertação. O *stakeholder* interno responsável pelo projeto participou do processo de consolidação dos processos de

negócio sugeridos e limitação do escopo das métricas de negócio para cada processo na modelagem do DW.

Avaliação Crítica da Proposta

A avaliação crítica da proposta foi a última etapa deste trabalho, onde foram avaliados os resultados atingidos com as técnicas e tecnologias adotadas. Também foram apontados os principais problemas encontrados durante o desenvolvimento da solução e propostas sugestões de melhorias. A ideia desta avaliação é permitir que, após a correção e adequação da solução, a ferramenta possa ser avaliada pelos *stakeholders* externos, através de questionários que utilizem as teorias do TAM (*Technology Acceptance Model*) (VENKATESH; BALA, 2008).

Finalizada a descrição da metodologia do trabalho, o próximo capítulo aborda o processo de execução e implementação da proposta.

4 ESTUDO DE CASO

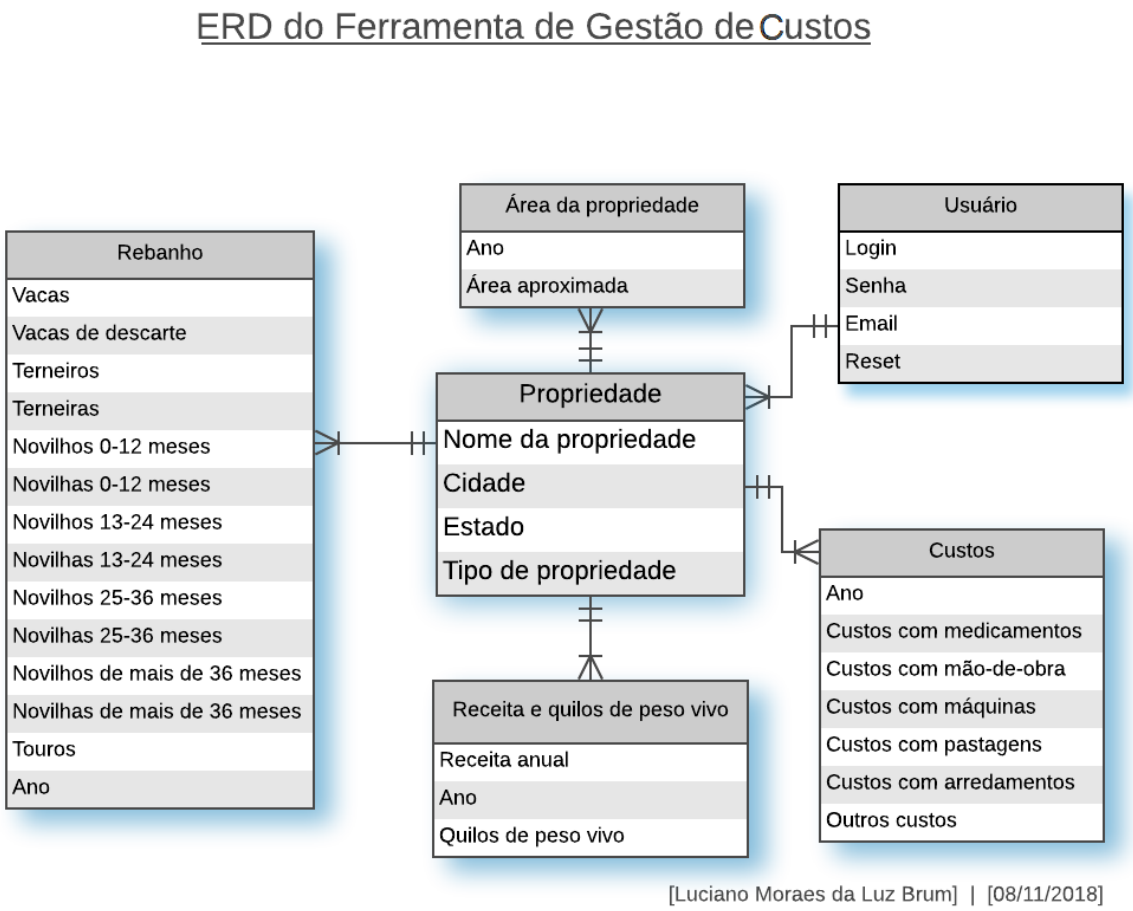
Neste capítulo é abordada a execução das etapas propostas na metodologia. Na seção 4.1 são apresentadas informações e características técnicas sobre os SAD que vão fornecer os dados para o DW. Na seção 4.2 é abordada a elaboração do documento de requisitos para o sistema desenvolvido. Na seção 4.3 é apresentada e justificada a escolha das ferramentas que viabilizaram a execução do projeto. Na seção 4.4 é apresentado o projeto do DW, onde são apresentadas a metodologia de aquisição de requisitos informacionais, a definição da arquitetura de DW utilizada, as métricas de negócio e respectivos descritores utilizados para a modelagem do DW e a criação e apresentação dos modelos lógicos do DW. A seção 4.5 apresenta a solução de ETL desenvolvida. A seção 4.6 apresenta as ferramentas utilizadas para visualização das informações do DW.

4.1 Sistemas fontes de dados: Descrição

Ferramenta de Gestão de Custos: este sistema visa auxiliar o produtor rural a compreender melhor os seus custos na pecuária de corte. O sistema está em desenvolvimento por fases. Cada fase é caracterizada pela inserção de um conjunto específico de informações sobre os custos da propriedade, informações quantitativas da produção, entre outros. Conforme o produtor conclui e avança nas fases, mais dados e informações sobre os custos e produção de sua propriedade serão requisitados. Com isso, o retorno de informações, gráficos, relatórios e *dashboards* sobre os custos serão cada vez mais enriquecidos e detalhados com as mesmas informações, de forma organizada. No atual momento, apenas as variáveis da fase 1 do sistema estão consolidadas. O SGBD utilizado para o armazenamento e gerenciamento dos dados deste sistema é o *PostgreSQL*. A Figura 21 apresenta o modelo ER conceitual da FGC na notação de Martin.

As variáveis de entrada da FGC utilizadas para simulação e visualização dos dados são os atributos de todas as tabelas (exceto a “Propriedade” e “Usuário”). Produtores precisam criar uma conta para utilizar o sistema informando apenas alguns dados pessoais e credenciais, apresentadas na tabela “Usuário”. Produtores podem registrar um ou vários estabelecimentos rurais, fornecendo os dados necessários apresentados na tabela “Propriedade”. É possível inserir diferentes registros para uma mesma propriedade, porém para diferentes anos.

Figura 21 – Modelo ER conceitual do FGC.



Fonte: Autor (2019).

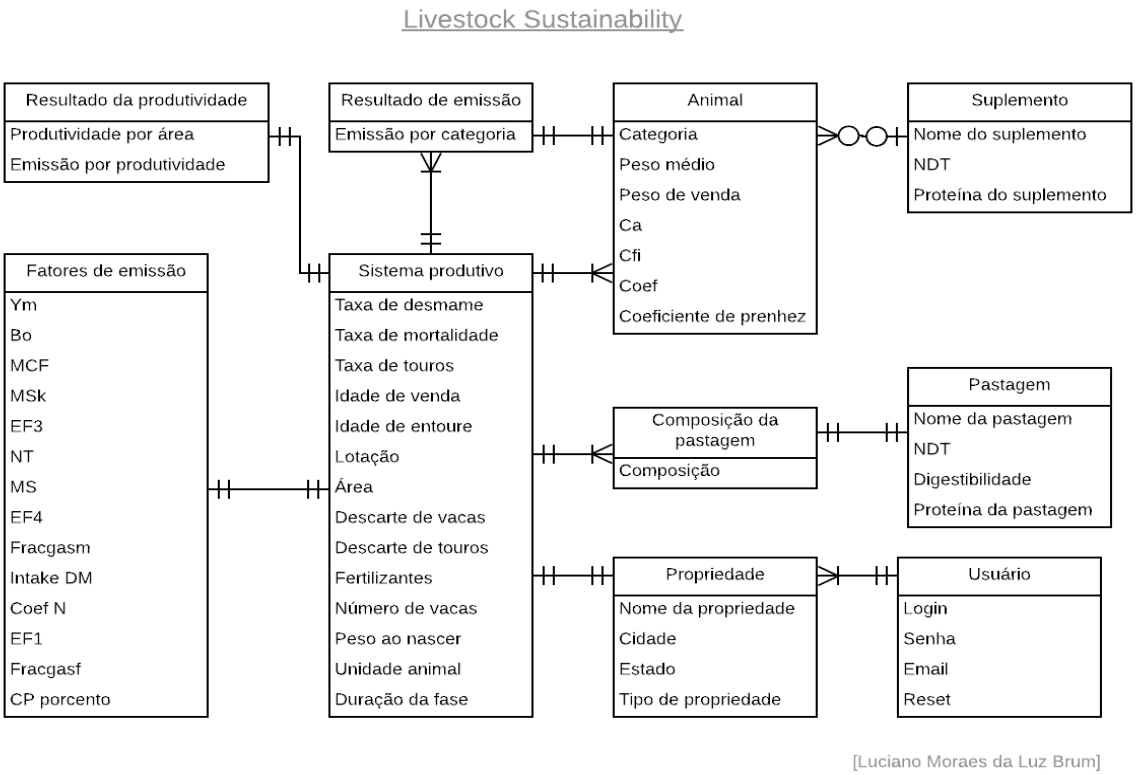
Inicialmente, o produtor deve informar os dados relativos aos custos, rebanho e propriedade, apresentados nas tabelas da Figura 22, para viabilizar as análises dos dados de sua propriedade. Os resultados são apresentados através de uma interface web com informações resumidas e organizadas sobre os custos produtivos. Gráficos simples são apresentados para fornecer um melhor entendimento da contribuição dos custos por tipo, estoque animal, custos por hectare, receita por hectare, custo por quilo de peso vivo produzido, lucro por hectare, entre outras métricas. Também é possível realizar análises comparativas com outros estabelecimentos rurais cadastrados no FGC.

Livestock Sustainability: é um aplicativo que visa fornecer ao produtor uma forma de estimar a emissão da pegada de carbono do sistema de produção (pecuária de corte), utilizando variáveis que sejam conhecidas por ele. A pegada de carbono é

uma medida da quantidade total de dióxido de carbono causada por alguma atividade ou acumulada durante os estágios da vida de um produto (WIEDMANN; MINX, 2007). A partir do sistema de informação, será coletado um conjunto de dados que permitirá realizar mineração de dados, de forma a tentar identificar quais variáveis mais influenciam na emissão e possíveis alternativas de reduzir a emissão. É um sistema de informação que estima a pegada de carbono e sugere alternativas de redução. O modelo que permite o cálculo da estimativa da emissão da pegada de carbono foi baseado em três teses de doutorado e faz parte de um projeto de mestrado em andamento.

Com relação à estrutura de armazenamento e gerenciamento de dados, o SGBD que será utilizado no servidor do LS é o *MySQL*, enquanto que para gerenciar e armazenar os dados locais do aplicativo foi utilizado o *SQLite*. Na Figura 22, é apresentado o modelo ER conceitual deste sistema.

Figura 22 - Modelo ER conceitual do sistema LS.



Fonte: Autor (2019).

O sistema utiliza informações zootécnicas da produção e da propriedade (taxa de natalidade (1) e mortalidade (2) do rebanho, taxa de desmame (3), idade de acasalamento e abate do rebanho e área disponível em hectares) combinada com informações consideradas padrão para a região, para mensurar a produtividade (4) da propriedade naquela região. Os cálculos são derivados do modelo proposto pela tese de Lampert (2010). Na sequência, são descritas as fórmulas destas métricas.

$$\text{Taxa de natalidade} = \left(\frac{\text{terneiros nascidos}}{\text{fêmeas acasaladas}}\right) * 100 \quad (1)$$

$$\text{Taxa de mortalidade} = \left(\frac{\text{número de animais que morreram}}{\text{total de animais}}\right) * 100 \quad (2)$$

$$\text{Taxa de desmame} = \left(\frac{\text{número de terneiros desmamados}}{\text{fêmeas acasaladas}}\right) \quad (3)$$

$$\text{Produtividade} = \left(\frac{\text{quantidade de quilos produzidos}}{\text{área}}\right) \quad (4)$$

Sobre o modelo conceitual da Figura 22, cabe mencionar que:

- Ao cadastrar cada pastagem, o usuário deve informar a sua composição, ou seja, o percentual utilizado desta pastagem.
- As variáveis de entrada peso médio, peso de venda e suplemento devem ser inseridas uma vez para cada categoria animal, que são dez (vacas, novilhas de 3 anos, novilhas de 2 anos, novilhas de 1 ano, novilhos de 3 anos, novilhos de 2 anos, novilhos de 1 ano, bezerras, bezerros e touros).
- As constantes (*Cfi*, *Coef* e *Cpregnancy*) devem ser inseridas uma vez para cada categoria animal.
- As demais variáveis de entrada do usuário são informadas uma única vez.

4.2 Requisitos do sistema: Descrição

Para delimitar o desenvolvimento da solução, em relação as funcionalidades e aspectos técnicos, e subsidiar a continuação do desenvolvimento da proposta em trabalhos futuros para outros profissionais e para a EMBRAPA, foi elaborado o documento de requisitos do sistema. O documento é um norteador para o desenvolvedor do sistema, de forma que o mesmo possa desenvolver o sistema correto para os usuários corretos. Também serve como um registro histórico das mudanças dos requisitos de usuário e do sistema, todos propostos pelo *stakeholder* interno.

O documento apresenta informações descritivas dos sistemas fontes e seus respectivos modelos conceituais, a arquitetura do sistema proposto e também os

Requisitos Funcionais (RF) e Não-Funcionais (RNF). Cada requisito possui seu respectivo identificador (composto pelo tipo de requisito seguido por um número inteiro de dois algarismos), descrição do caso de uso, prioridade (essencial, importante ou desejável), entradas e pré-condições e saídas e pós-condições. Cabe mencionar que todos RF definidos são de prioridade essencial. O documento de requisitos completo pode ser encontrado no Apêndice A deste trabalho.

4.3 Ferramenta de BI selecionada

Analisados os trabalhos da literatura científica em paralelo com a utilização e teste das ferramentas *Pentaho* e *SpagoBI*, foi consolidada a escolha da suíte *Pentaho*. O *SpagoBI*, apesar de estar bem colocado como solução gratuita e *open source* para BI, recentemente foi atualizado para um novo produto, chamado *Knowage*. A ferramenta *Knowage* não foi analisada por ser uma solução ainda muito recente no mercado de suítes de BI, diferente do *Pentaho*. A desvantagem imediata em consequência disto é a existência de um universo reduzido de usuários, consequentemente, pouco material para sanar eventuais dúvidas sobre a ferramenta.

A suíte *Pentaho* foi desenvolvida na linguagem de programação *Java*, com o propósito de auxiliar equipes de TI e gestores no gerenciamento de informações e tomadas de decisão. A suíte possui uma versão gratuita (*Community Edition* - CE) e uma versão paga (*Enterprise Edition* - EE). A versão gratuita está sob as licenças LGPL (*Lesser General Public License*) versão 2.0, GPL (*General Public License*) versão 2.0 e MPL (*Mozilla Public License*) versão 1.1. A suíte pode ser utilizada nos ambientes *Linux*, *Windows* e *MacOS*. Foi utilizada a versão CE do *Pentaho*.

A suíte conta com componentes desenvolvidos individualmente, cada um para atender um determinado tipo de demanda relacionada aos processos intrínsecos de um projeto de BI. Abaixo são detalhados os componentes utilizados neste trabalho[5].

- PDI (*Pentaho Data Integration*): é o recurso da suíte que permite desenvolver o processo de ETL. O módulo *spoon* permite realizar diversas tarefas básicas necessárias para um processo de ETL, como leitura (de planilhas, bancos de dados, arquivos JSON, etc.), processamento (filtros, normalização e desnormalização de tabelas, cruzamentos, etc.) e escrita (em planilhas, arquivos

[5]SOURCEFORGE. Hitachi Ventara | Pentaho. 2018. Disponível em: https://sourceforge.net/projects/pentaho/files/. Acesso em: 24 jul. 2018.

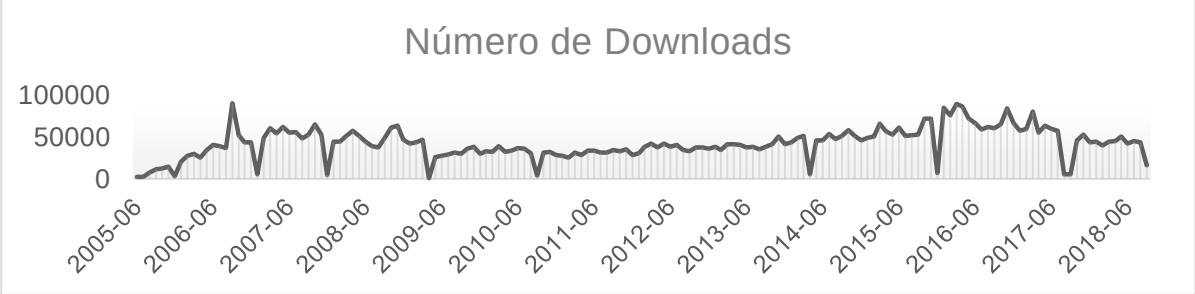
de texto, bancos de dados, etc.) de dados através de uma interface gráfica simples e intuitiva. Ainda, é compatível com várias linguagens e tecnologias gratuitas e *open source*, incluindo as que foram utilizadas neste trabalho;

- *Pentaho Business Analytics Server* (PBAS): O PBAS, versão 8.1.0.0, foi utilizado como servidor web, no nível de aplicação, para: configurar as fontes de dados, gerenciar o controle de acesso ao DW para os usuários, realizar as junções (*joins*) necessárias entre as tabelas fato e dimensões, gerenciar a interface web, configurar o acesso e consumo dos dados do DW para permitir que os usuários possam visualizar as informações de interesse. Também é possível, para o administrador do sistema, instalar *plug-ins* adicionais para serem integrados com o PBAS de forma simples e intuitiva.
- *Pentaho Schema Workbench* (PSW): o PSW possui o servidor OLAP *Mondrian* embutido para o processamento de consultas aos cubos de dados. O *Mondrian* foi escrito em Java, existe como projeto desde 2003 e foi adquirido pelo *Pentaho* em 2005. Permite executar consultas escritas na linguagem MDX (*Multidimensional Expressions*) para leitura dos dados de bases de dados baseadas em tecnologias relacionais. É *open source,* gratuito e permite construir estruturas lógicas no topo de um banco de dados com base em um arquivo XML específico chamado *Schema*. O *Schema* provê definições XML para os cubos de dados, dimensões, hierarquias, indicadores analíticos e seus mapeamentos para as estruturas de dados do DW. Também é possível definir regras de acesso para os usuários através do *Schema*. Em resumo, o *Mondrian* se encarrega de receber consultas dimensionais a um cubo por MDX, sendo o cubo um conjunto de metadados que definem o mapeamento das consultas SQL para as bases de dados que realmente contém os dados (DÍAZ, 2012).
- *Saiku Analytics*[6]*:* A aplicação de BI e cliente OLAP utilizado como meio para acessar, interagir e visualizar os dados do DW, permitindo gerar e baixar relatórios nos formatos .pdf, .png, .xls e .csv foi o *Saiku Analytics*. A ferramenta possui funcionalidades OLAP, funcionalidades do tipo *drag and drop* (arrastar e soltar), funciona bem junto com o PBAS como um *plugin*, é *open source* e possui uma versão disponível para download e uso gratuito.

[6]SAIKU ANALYTICS. Meteorite bi: saiku analytics. [2018?]. Disponível em: <http://meteorite.bi/saiku>. Acesso em: 18 jul. 2018.

As Figuras 23 e 24 mostram, respectivamente, o número de downloads da suíte *Pentaho* desde a criação do projeto, incluídos seus componentes e o número de downloads do *PBAS* versão 8.1, disponibilizado em maio de 2018. Os números demonstram o interesse das pessoas e empresas na utilização de soluções *open source* e gratuitas de suítes de BI, como o *Pentaho* e seus componentes.

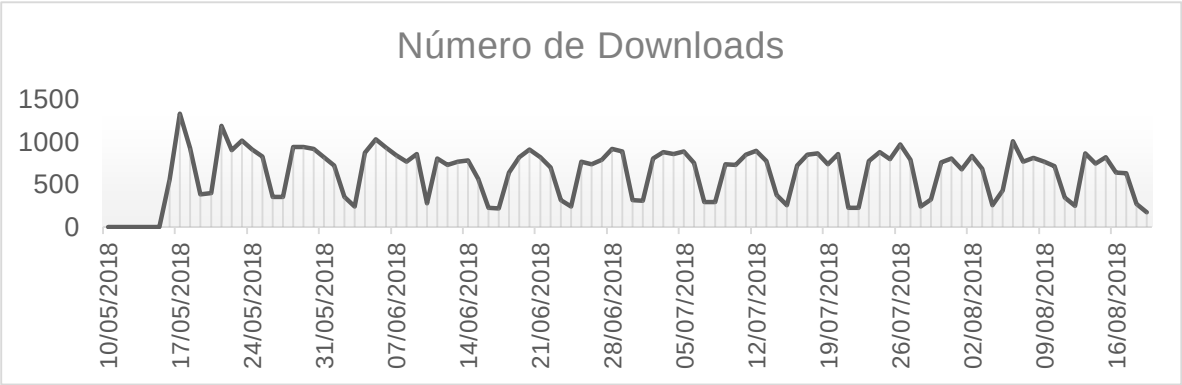
Figura 23: Downloads da suíte *Pentaho* e seus componentes desde 2005.



Fonte: Adaptado de SOURCEFORGE (2018).

Os principais requisitos de hardware e software para os usuários da solução PBAS é possuir um dispositivo com um navegador web instalado com acesso à internet. O sistema pode ser utilizado em computadores e dispositivos móveis. Os requisitos mínimos de hardware e software para a instalação dos componentes do Pentaho na máquina servidora podem ser encontrados no site da documentação do *Pentaho*[7], versão 8.1. A próxima seção aborda o projeto, desenvolvimento e implementação do DW na máquina servidora.

Figura 24: Downloads do *Pentaho BI Server* v. 8.1 desde maio de 2018.



Fonte: Adaptado de SOURCEFORGE (2018).

[7]Disponível em: https://help.pentaho.com/Documentation/8.1/Setup/Components_Reference. Acesso em: 18 jul. 2018.

4.4 Projeto do *Data Warehouse*

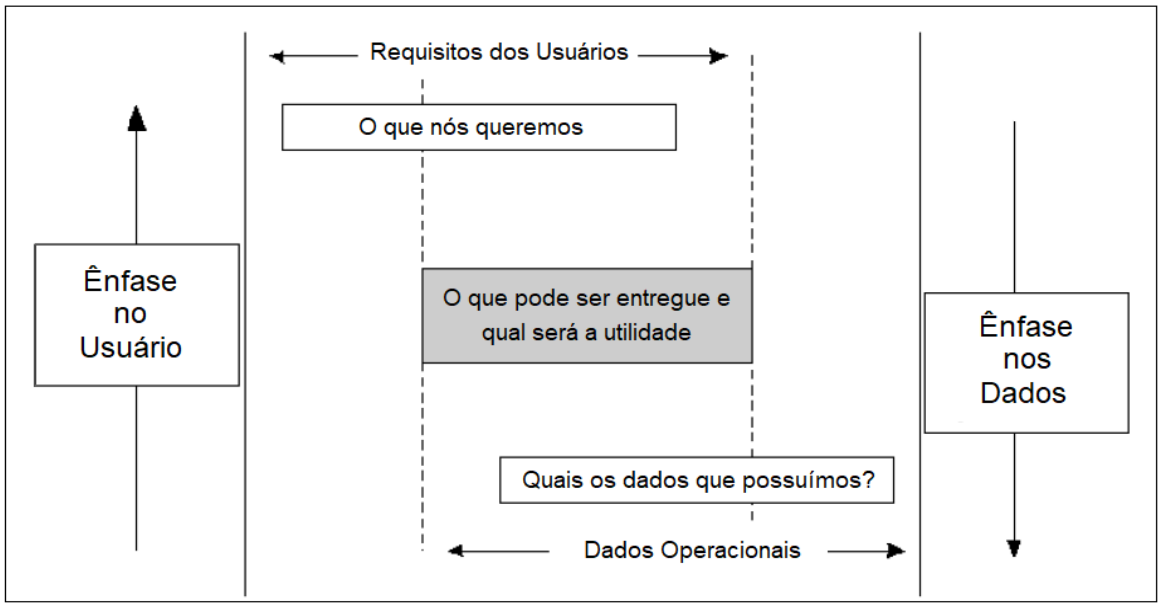
Sobre o projeto do DW, na seção 4.4.1 é abordada a metodologia de aquisição de requisitos informacionais e necessidades do usuário e do sistema. Na seção 4.4.2 é abordada, de forma genérica, a arquitetura geral da solução de DW proposta. Na seção 4.4.3 são abordados os diversos aspectos da modelagem dimensional, assim como são apresentados os modelos conceitual e físico do DW.

4.4.1 Metodologia de aquisição de requisitos informacionais

Existem diferentes abordagens e metodologias para o design multidimensional de DW. Metodologias direcionadas às necessidades do usuário são extremamente importantes, especialmente quando os sistemas fontes de dados possuem uma variedade muito grande de dados, no qual os desenvolvedores e projetistas não possuem o domínio ou entendimento mínimo suficiente do problema. Porém, utilizar essa abordagem, de forma exclusiva, não é adequada no caso deste trabalho, em que os requisitos de dados do usuário (EMBRAPA) podem não estar alinhados com o estado atual de cada sistema fonte, em fase de desenvolvimento. Em outras palavras, é possível que parte dos dados que o cliente deseja podem não estar disponíveis, por estes não estarem modelados ou conformados nos sistemas fontes de dados. Ao mesmo tempo, utilizar abordagens direcionadas exclusivamente a dados podem acarretar na modelagem de um DW que reunirá estes dados, porém que não serão preponderantes para o auxílio decisório dos usuários do sistema, os produtores rurais, causando a insatisfação dos clientes internos e externos com o produto final. Este poderia ser o resultado, caso não houvesse interação frequente com os especialistas do domínio do problema.

Portanto, para esse trabalho, foi adotada a metodologia mista, em que são considerados ambos: os requisitos informacionais do cliente e o estado atual dos sistemas fontes de dados. Esta é uma metodologia que tem se mostrado mais efetiva para a concepção de modelos conceituais para DW (ZAZA *et al*., 2018). Tal metodologia foi adaptada na Figura 25.

Figura 25: Metodologia com ênfase nos requisitos do usuário e nos dados.



Fonte: Autor (2019).

Foram realizadas discussões presenciais e não-presenciais com o cliente interno com frequência, na maioria dos casos, diária, em que foram abordados os dados que seriam necessários para a solução de DW. Ainda, para obter um melhor embasamento sobre quais métricas e indicadores poderiam ser apresentados, foi utilizado como base o trabalho de Costa *et al*. (2018), e algumas métricas produtivas e econômicas apresentadas neste trabalho foram utilizadas. Após as reuniões, foi realizado apenas um contato presencial e não-presencial (por e-mail) com o desenvolvedor da FGC, visando sanar eventuais dúvidas relacionadas às estruturas relacionais da base de dados desse sistema. O mesmo contato foi realizado com o desenvolvedor do sistemas LS, porém, foram mantidos encontros diários, pelo período de um mês. A diferença no número de interações foi devido à proporcional diferença no nível de complexidade dos sistemas, visto que a FGC, diferentemente do LS, não possui equações matemáticas complexas integradas. Todos os encontros com os desenvolvedores foram realizados após o acesso e análise das estruturas e funções dos sistemas.

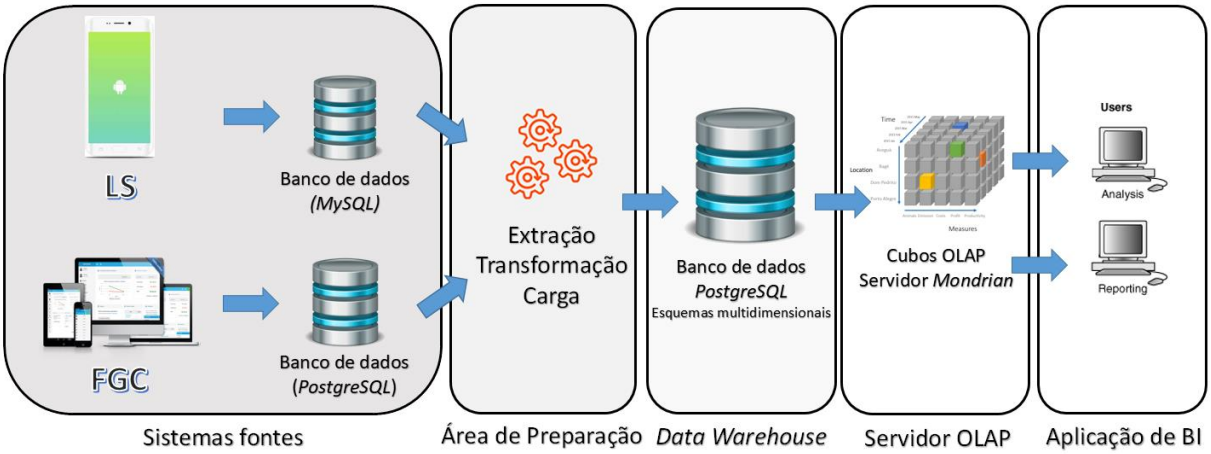
Foram identificados diversos problemas relacionados ao formato e a estrutura no qual estavam armazenados os dados dos sistemas fontes, de forma a não só acarretar na entrega de dados com pouca qualidade e confiabilidade, mas também inviabilizando a integração dos dados dos sistemas. Estes problemas serão discutidos

em detalhes na seção 4.5.1, no processo de ETL, onde também são sugeridas as soluções para atender os requisitos da proposta.

4.4.2 Arquitetura

A primeira etapa do projeto foi a definição da arquitetura geral da solução de DW. A arquitetura foi baseada na metodologia defendida por Kimball e Ross (2013), baseado na arquitetura *Enterprise Bus Architecture* (sem EII ou ODS), conforme apresentada e adaptada na Figura 26. Esta solução é baseada em tecnologia ROLAP, implementada com tecnologias gratuitas e *open source*.

Figura 26: Arquitetura geral da solução, baseada no método de Kimball e Ross.



Fonte: Autor (2019).

Optou-se por este modelo de arquitetura por haver menor replicação dos dados em comparação com a abordagem de Inmon (CIF). Não se mostrou necessário realizar a separação entre os dados para uma visão geral da empresa rural e dados para uma visão setorial. Neste sentido, o DW nesta solução é representado por um conjunto de *star schemas* conectados por dimensões conformadas. Como o grão da informação é o mesmo em todas, foi possível estabelecer apenas a tabela fato em apenas um *star schema* com todas as faces da pecuária de corte.

A seguir, são detalhados cada um dos componentes da arquitetura:

Sistemas fontes: são os componentes que possuem os dados de negócio e necessitam ser integrados, visando prover uma única versão da verdade de diferentes facetas da pecuária de corte para produtores rurais e pesquisadores. Os dados coletados são sobre aspectos produtivos, ambientais e econômicos de

estabelecimentos rurais que tem como atividade foco a pecuária de corte. Conforme apresentado anteriormente na seção 4.2, os dados de dois sistemas serão integrados para fornecer uma visão multidimensional das métricas de ambas soluções.

Área de preparação dos dados: é onde o processo de ETL é executado, sendo um armazenamento intermediário entre os sistemas fontes e o DW. Atualmente, os sistemas fontes estão sendo desenvolvidos (interface da aplicação) e, no caso da FGC, nenhum dado foi coletado ainda. Portanto, esses dados serão simulados com base em parâmetros conhecidos da pecuária de corte brasileira, definidos conjuntamente com o *stakeholder* interno. Já no caso do LS, alguns dados reais de pecuaristas de corte foram coletados através do uso de um questionário *online* no *Google Docs*. Algumas regras de negócio e checagens sobre a qualidade dos dados, assim como uma descrição detalhada do processo ETL, são discutidas na seção 4.7.

Data Warehouse: O DW é onde os dados extraídos, transformados e processados no ETL são armazenados. O DW foi implementado utilizando o SGBD *PostgreSQL*, e consiste em vários cubos de dados, cada um implementado através do modelo *star schema*.

Servidor *Mondrian*: foi utilizado para executar consultas aos dados do DW e apresenta-los no formato multidimensional, disponível na ferramenta PSW.

Aplicação de BI: As ferramentas utilizadas nesta camada foram o *plugin Saiku Analytics* e o PBAS. O foco nesta camada é fornecer ao usuário as funcionalidades necessárias para acesso e visualização dos dados do DW.

Na próxima seção, é detalhado o processo de modelagem, desenvolvimento e implementação do modelo dimensional do DW.

4.4.3 Modelagem, desenvolvimento e implementação do DW

A primeira atividade neste processo é identificar os processos de negócio e dados de interesse para o *stakeholder*, consonantes às realidades dos sistemas em desenvolvimento. O objetivo é fornecer o máximo de informações relevantes com o mínimo de dados disponíveis para os produtores rurais. Os processos de negócio são elucidados de forma a orientar o processo de definição das métricas do negócio. Foram identificados três processos de negócio e cinco dimensões sob o qual é possível realizar análises.

Foram identificas as seguintes dimensões:

Tempo: a dimensão temporal é essencial em projetos de DW, pois permite realizar análises históricas das mudanças nas métricas de negócio. Na pecuária de corte, considerando-se os processos de compra e venda de animais, período de gestação das vacas, época e tempo de desmame e a experiência do *stakeholder* interno, considerou-se que as métricas devem ser registradas por ano. A maioria dos processos da pecuária de corte são analisados neste período de tempo e a grande maioria dos produtores rurais já não realizam uma gestão em um nível de detalhe maior das informações de sua propriedade, conforme apresentado na revisão da literatura. Portanto, o nível de detalhe das informações será anual, para convencer o produtor, se possível, da importância de se obter os dados das diferentes dimensões sobre o sistema produtivo com uma maior frequência.

Localidade: por meio da dimensão de localidade que é possível analisar os indicadores econômicos, ambientais e produtivos por propriedade individualmente. Foram definidos vários níveis hierárquicos para esta dimensão: país, região, estado, mesorregião, microrregião, cidade e propriedade. É possível analisar as métricas pelos níveis acima, e por estabelecimento rural individualmente, sendo este identificado pelo seu nome, anteriormente cadastrado nos sistemas fontes de dados.

Área: a dimensão de área permite a análise dos indicadores por agrupamentos de estabelecimentos rurais de acordo com sua área. Inicialmente foi proposta a divisão das faixas de áreas de acordo com a realizada pelo IBGE, porém, foi definido com o *stakeholder* interno uma divisão menor, de até quatro faixas, para evitar a poluição de informações na interface para os usuários do sistema. As faixas definidas foram de 0 a 300 hectares, de 300 a 600 hectares, de 600 a 1000 hectares e mais de 1000 hectares.

Suplemento: a dimensão de suplemento permite que as métricas sejam analisadas pelo uso ou não de suplementos nos estabelecimentos rurais. Em um nível de detalhe maior, também é possível analisar as métricas por NDT (nutrientes digestíveis totais) do suplemento, quando utilizado. O NDT foi utilizado como grão de informação porque, no contexto das métricas, faz mais sentido analisá-las pela sua contribuição através de nutrientes digestivos em vez de analisá-las pelo nome do suplemento, que possui uma variedade muito grande e, ainda, um mesmo suplemento pode ter composições diferentes, além de não trazer contribuições no entendimento do seu impacto nas métricas de negócio. Tal decisão foi realizada em conjunto com o *stakeholder* interno.

Pastagem: a dimensão de pastagem permite realizar análises das métricas por categoria de pastagem predominante (natural, nativa ou artificial, sendo esta última de verão ou inverno), por nome da pastagem predominante utilizada e pela descrição de todas pastagens e respectivas composições. O grão da informação é a pastagem predominante, que pode variar nos estabelecimentos rurais.

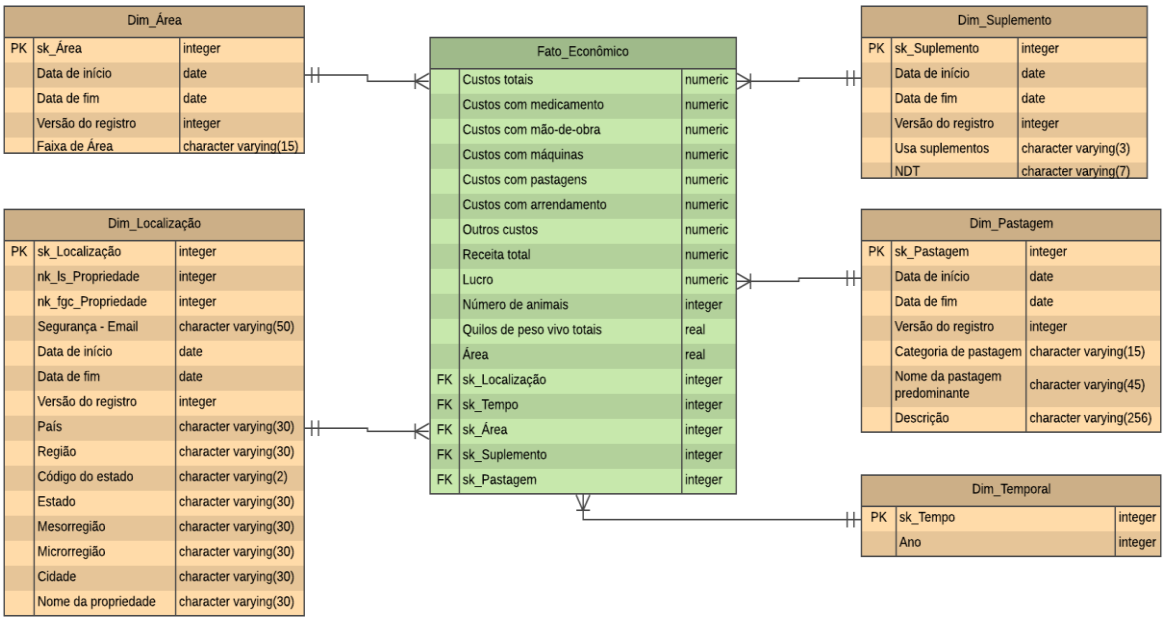
Foram identificados três processos de negócio, em que convencionou-se chamá-los de econômico, produtivo e ambiental. Os processos são detalhados abaixo.

Econômico: as métricas de negócio presentes neste processo são relacionadas aos indicadores econômicos da propriedade. O grão da informação associado com as métricas deste processo é a propriedade rural em relação à localidade, anual em relação ao tempo, por faixas de área dos estabelecimentos rurais em hectares em relação à área, por pastagem predominante em relação às pastagens e por NDT do suplemento utilizado em relação aos suplementos. Na Figura 27 é possível observar o DM gerado a partir deste processo de negócio.

Produtivo: As métricas de negócio presentes neste processo são relacionadas aos indicadores produtivos da propriedade. O grão da informação associado com as métricas deste processo é o mesmo das métricas do processo Econômico. Na Figura 28 é possível observar o DM gerado a partir deste processo de negócio.

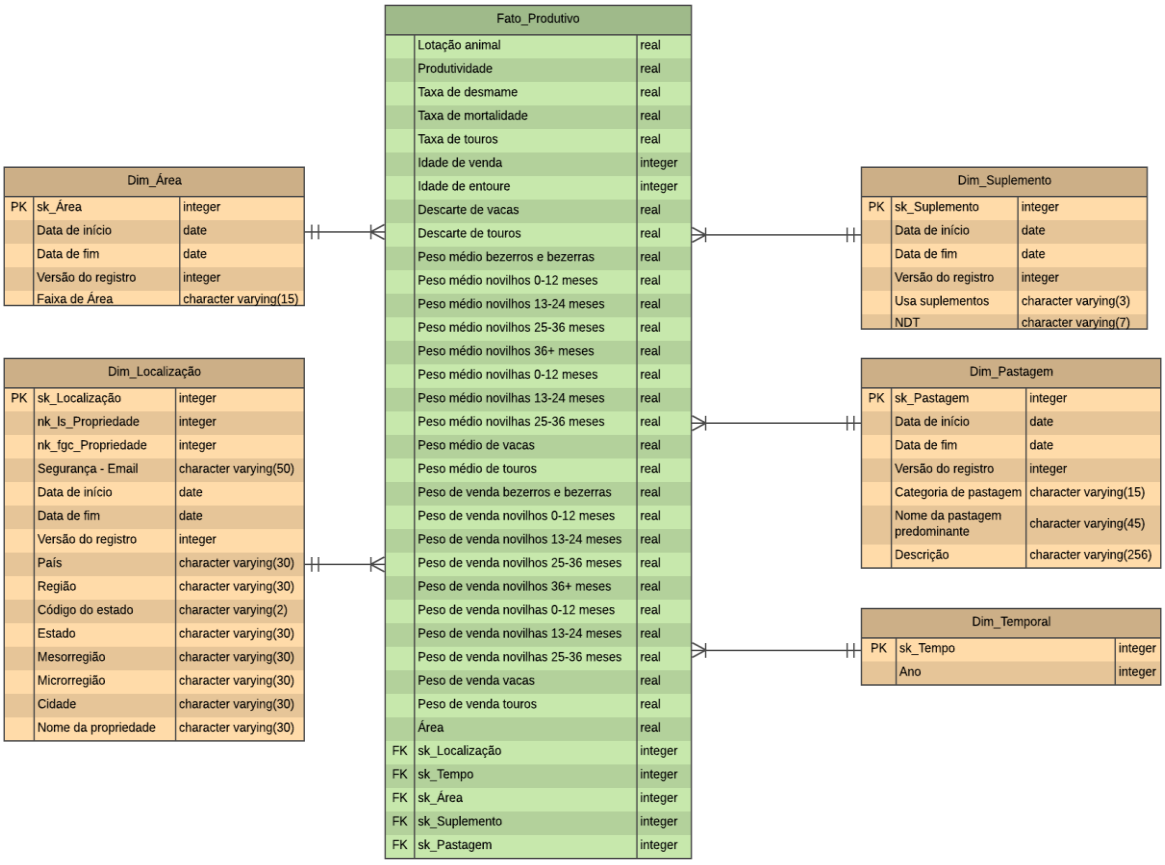
Ambiental: As métricas de negócio presentes neste processo são relacionadas aos fatores ambientais. O grão da informação associado com as métricas deste processo é o mesmo das métricas dos processos Econômico e Produtivo. Na Figura 29 é possível observar o DM gerado a partir deste processo de negócio.

Figura 27: Modelo lógico do DM fato_economico.



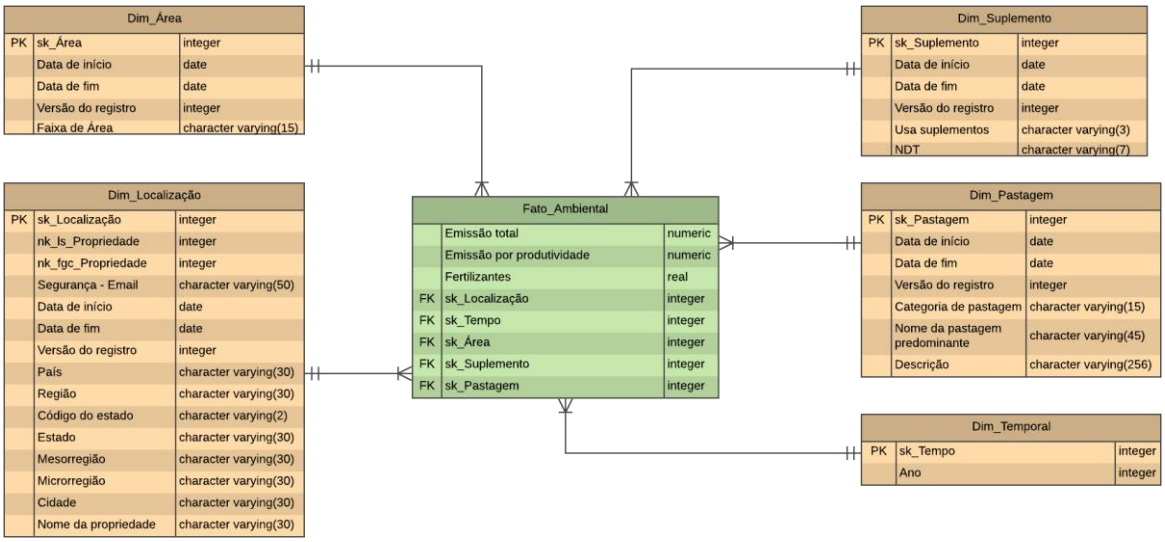
Fonte: Autor (2019).

Figura 28: Modelo lógico do DM fato_produtivo.



Fonte: Autor (2019).

Figura 29: Modelo lógico do DM fato_ambiental.



Fonte: Autor (2019).

Na Tabela 6, são apresentados os processos de negócio e suas dimensões associadas. Este tipo de tabela é chamada de matriz de barramento (do inglês *Bus Matrix*), que visa facilitar a identificação dos processos, dimensões e dimensões comuns entre os processos. O objetivo é fornecer um melhor entendimento do DW ou conjunto de DM que serão modelados e implementados.

Tabela 6: *Bus Matrix* com os processos de negócio e dimensões comuns.

	Dimensões em comum				
Processos de Negócio	**Localidade**	**Tempo**	**Suplemento**	**Área**	**Pastagens**
Econômico	X	X	X	X	X
Produtivo	X	X	X	X	X
Ambiental	X	X	X	X	X

Fonte: Autor (2019).

Na coluna processos de negócio, é possível observar os três processos de negócio anteriormente apresentados e as dimensões pelos quais é possível analisar as métricas destes processos. Projetou-se os DW de forma que todas as dimensões fossem comuns para todos os processos (dimensões conformadas), sendo estes separados em três tabelas fato. Como a granularidade das informações é a mesma

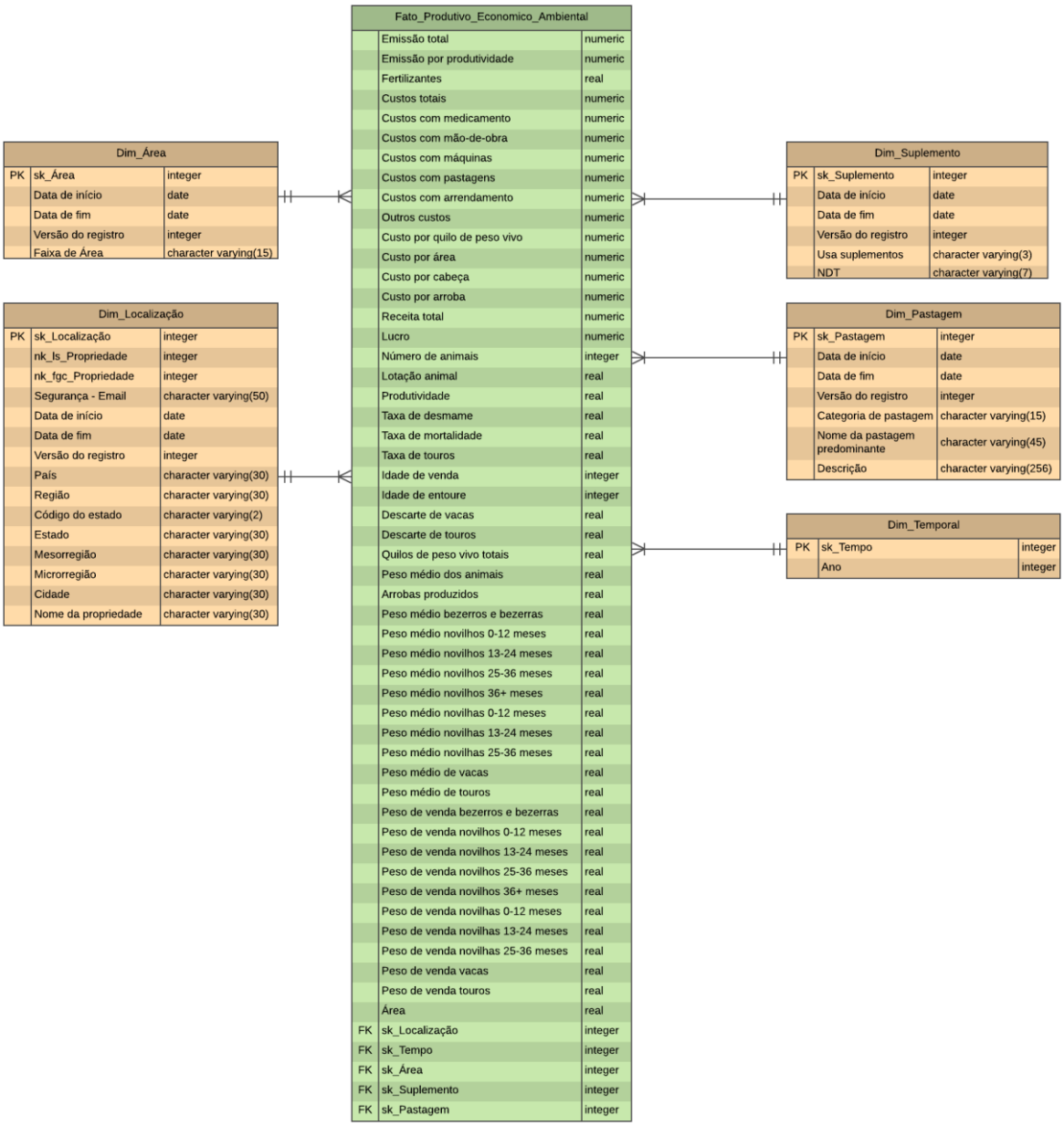
para as três dimensões de análise da pecuária de corte, é possível implementar apenas uma tabela fato contendo todas os indicadores do sistema produtivo.

Uma peculiaridade adicional deve ser considerada com relação às possíveis mudanças nas dimensões. Por exemplo, é possível que um produtor rural mude de propriedade, e neste contexto, surgirão inconsistências nas análises se tais casos não forem considerados na modelagem e tratados no processo ETL. O mesmo tratamento cuidadoso deve ser realizado com as outras dimensões, pois além de propriedade, o produtor pode expandir a área de sua propriedade, modificar as composições e uso de suas pastagens, assim como suplementos.

Uma alternativa muito utilizada para lidar com estes casos são as dimensões *Slowly Changing Dimension (SCD)* do tipo 2. As dimensões SCD são aquelas em que considera-se futuras mudanças nas dimensões, visando mantê-las atualizadas e sincronizadas com os sistemas fontes (KIMBALL; ROSS, 2013). Todas as dimensões, exceto a tempo, são SCD do tipo 2. A dimensão de tempo, SCD de tipo 0, é carregada apenas uma vez no DW, portanto, não são necessárias mudanças futuras. Três atributos são inseridos adicionalmente na dimensão: a data no qual o registro foi inserido na dimensão (chamada de data efetiva ou data de início), a data no qual ocorreu a mudança no registro da dimensão (data de expiração dos valores daquela linha da dimensão ou data de fim) e uma *flag* indicando se o registro daquela dimensão é válido ou expirado.

Por fim, o modelo lógico do DW pode ser conferido na Figura 30. A Tabela 7 apresenta em detalhes as descrições de cada uma das colunas presentes na tabela fato e as tabelas 8, 9, 10, 11 e 12 apresentam as descrições de cada uma das colunas presentes nas tabelas dimensões.

Figura 30: Modelo lógico do DW no formato *star schema.*



Fonte: Autor (2019).

Tabela 7: Descrição da tabela ‘Fato_Produtivo_Economico_Ambiental’.

(continua)

Atributo	Tipo	Tamanho	Chave	*Descrição ou equação
Sk_Localização	INTEGER	4 bytes	Estrangeira	Código que associa o registro com a dimensão localização.
Sk_Tempo	INTEGER	4 bytes	Estrangeira	Código que associa o registro com a dimensão tempo.
Sk_Área	INTEGER	4 bytes	Estrangeira	Código que associa o registro com a dimensão área.
Sk_Suplemento	INTEGER	4 bytes	Estrangeira	Código que associa o registro com a dimensão suplemento.
Sk_Pastagem	INTEGER	4 bytes	Estrangeira	Código que associa o registro com a dimensão pastagem.
Número de animais	INTEGER	4 bytes		N° total de animais.
Lotação animal	REAL	4 bytes		$\frac{\text{UA}}{\text{Área (ha)}}$
Produtividade por área	REAL	4 bytes		$\frac{\text{Quilos de PV (kg)}}{\text{Área (ha)*ano}}$
Taxa de desmame	REAL	4 bytes		$\frac{\text{N° de bezerros desmamados*100}}{\text{N° de fêmeas}}$
Taxa de mortalidade	REAL	4 bytes		$\frac{\text{N° de animais mortos ou perdidos}}{\text{N° total de animais}}$
Taxa de touros	REAL	4 bytes		Percentual de touros em relação ao total de animais.
Idade de venda	INTEGER	4 bytes		Idade de venda dos novilhos(as).
Idade de entoure	INTEGER	4 bytes		Idade de entoure dos novilhos(as).
Descarte de vacas	REAL	4 bytes		Percentual de vacas descartadas.
Descarte de touros	REAL	4 bytes		Percentual de touros descartados.
Quilos de peso vivo totais	REAL	4 bytes		Total de quilos produzidos.
Peso médio dos animais	REAL	4 bytes		Peso médio de todos animais.
Arrobas produzidos	REAL	4 bytes		Valor produzido em arrobas.
Peso médio (todas categorias)	REAL	4 bytes		Peso médio de cada categoria animal.
Peso de venda (todas categorias)	REAL	4 bytes		Peso de venda de cada categoria animal.
Área	REAL	4 bytes		Área da propriedade em ha.
Custos totais	NUMERIC	Variável		Total de custos em reais.
Custos com medicamento	NUMERIC	Variável		Custos com medicamentos em reais.
Custos com mão-de-obra	NUMERIC	Variável		Custos com mão-de-obra em reais.
Custos com máquinas	NUMERIC	Variável		Custos com máquinas em reais.
Custos com pastagens	NUMERIC	Variável		Custos com pastagens em reais.
Custos com arrendamento	NUMERIC	Variável		Custos com arrendamento em reais.

Tabela 7: Descrição da tabela ‘Fato_Produtivo_Economico_Ambiental’.

(conclusão)

Atributo	Tipo	Tamanho	Chave	*Descrição ou equação
Outros custos	NUMERIC	Variável		Custos que não estão nas categorias anteriores, em reais.
Custos por quilo de PV produzido	NUMERIC	Variável		$\frac{\text{Custos totais (R\$)}}{\text{Quilos de PV totais (kg)}}$
Custo por área	NUMERIC	Variável		$\frac{\text{Custos totais (R\$)}}{\text{Área (ha)}}$
Custo por cabeça	NUMERIC	Variável		$\frac{\text{Custos totais (R\$)}}{\text{N° de animais}}$
Custo por arroba	NUMERIC	Variável		$\frac{\text{Custos totais (R\$)}}{\text{Arrobas (@)}}$
Receita total	NUMERIC	Variável		Total de receita gerada em reais.
Lucro	NUMERIC	Variável		Receita total-Custos totais
Emissão total	REAL	4 bytes		Emissão total de CO_2 pelos animais (kg).
Emissão por produtividade	REAL	4 bytes		$\frac{\text{Emissão total de } CO_2 \text{ (kg)}}{\text{Quilos de PV totais (kg)}}$
Fertilizantes	REAL	4 bytes		Litros de fertilizantes utilizados.

Fonte: Autor (2019).

Notas: PV = Peso Vivo, ha = hectares, kg = quilos, UA = Unidade Animal (cada UA corresponde à uma vaca de 450 kg), @ = arroba (cada @ corresponde à 15 quilos de carcaça de um animal, sendo a carcaça o peso da carne e o osso do animal apenas, em média representando 50% do peso total do animal), CO_2 = Dióxido de Carbono.

Tabela 8: Descrição das colunas da tabela ‘Dim_Temporal’.

Atributo	Tipo	Tamanho	Chave	*Descrição ou equação
Sk_Tempo	INTEGER	4 bytes	Primária	Chave primária da dimensão que é igual ao ano do registro.
Ano	INTEGER	4 bytes		Ano do registro.

Fonte: Autor (2019).

Tabela 9: Descrição das colunas da tabela ‘Dim_Área’.

Atributo	Tipo	Tamanho	Chave	*Descrição ou equação
Sk_Área	INTEGER	4 bytes	Primária	Chave primária da dimensão.
Data de início	DATE	4 bytes		Data em que foi inserido o registro da propriedade.
Data de fim	DATE	4 bytes		Data em que deixou de ser atual o registro da propriedade.
Versão do registro	INTEGER	4 bytes		Realiza o versionamento das mudanças na dimensão.
Faixa de área	CHARACTER VARYING (15)	Variável		Faixa de área ao qual a propriedade pertence.

Fonte: Autor (2019).

Tabela 10: Descrição das colunas da tabela ‘Dim_Suplemento’.

Atributo	Tipo	Tamanho	Chave	*Descrição ou equação
Sk_Suplemento	INTEGER	4 bytes	Primária	Chave primária da dimensão.
Data de início	DATE	4 bytes		Data em que foi inserido o registro da propriedade.
Data de fim	DATE	4 bytes		Data em que deixou de ser atual o registro da propriedade.
Versão do registro	INTEGER	4 bytes		Realiza o versionamento das mudanças na dimensão.
Usa suplementos?	CHARACTER VARYING (3)	Variável		Registro que identifica se a propriedade utiliza suplementos.
NDT	CHARACTER VARYING (7)	Variável		Nutrientes digestíveis totais médios dos suplementos utilizados.

Fonte: Autor (2019).

Tabela 11: Descrição das colunas da tabela ‘Dim_Localização’.

Atributo	Tipo	Tamanho	Chave	*Descrição ou equação
Sk_Localização	INTEGER	4 bytes	Primária	Chave primária da dimensão.
Nk_ls_propriedade	INTEGER	4 bytes	Natural	Chave primária no sistema LS que identifica a propriedade.
Nk_fgc_propriedade	INTEGER	4 bytes	Natural	Chave primária no sistema FGC que identifica a propriedade.
Segurança (*e-mail*)	INTEGER	4 bytes		*E-mail* do usuário. Será utilizado para implementação do DRLS.
Data de início	DATE	4 bytes		Data em que foi inserido o registro da propriedade.
Data de fim	DATE	4 bytes		Data em que deixou de ser atual o registro da propriedade.
Versão do registro	INTEGER	4 bytes		Realiza o versionamento das mudanças na dimensão.
País	CHARACTER VARYING (30)	Variável		País onde se localiza a propriedade.
Região	CHARACTER VARYING (30)	Variável		Região onde se localiza a propriedade.
Estado	CHARACTER VARYING (30)	Variável		Unidade federativa onde se localiza a propriedade.
Código do estado	CHARACTER VARYING (2)	Variável		Código da unidade federativa.
Mesorregião	CHARACTER VARYING (30)	Variável		Mesorregião onde se localiza a propriedade.
Microrregião	CHARACTER VARYING (30)	Variável		Microrregião onde se localiza a propriedade.
Cidade	CHARACTER VARYING (30)	Variável		Cidade onde se localiza a propriedade.
Nome da propriedade	CHARACTER VARYING (30)	Variável		Nome da propriedade.

Fonte: Autor (2019).

Tabela 12: Descrição das colunas da tabela ‘Dim_Pastagem’.

Atributo	Tipo	Tamanho	Chave	*Descrição ou equação
Sk_Pastagem	INTEGER	4 bytes	Primária	Chave primária da dimensão.
Data de início	DATE	4 bytes		Data em que foi inserido o registro da propriedade.
Data de fim	DATE	4 bytes		Data em que deixou de ser atual o registro da propriedade.
Versão do registro	INTEGER	4 bytes		Realiza o versionamento das mudanças na dimensão.
Categoria da pastagem	CHARACTER VARYING (30)	Variável		Categoria da pastagem predominante no estabelecimento rural.
Nome da pastagem	CHARACTER VARYING (45)	Variável		Nome da pastagem predominante no estabelecimento rural.
Descrição	CHARACTER VARYING (256)	Variável		Pastagens e respectivas composições no estabelecimento rural, em porcentagem.

Fonte: Autor (2019).

Um problema observado na solução proposta é o de permitir que produtores acessem as métricas de estabelecimentos rurais no qual eles não sejam os proprietários. Este problema é incomum em projetos de DW, pois, em geral, são gerentes ou funcionários da alta hierarquia de uma empresa que acessam as informações, sem restrições de segurança (exceto quando tem-se diversos departamentos, sendo que cada gerente só pode acessar dados relativos ao seus processos). Para solucionar este problema, em que o usuário só pode acessar determinadas linhas de uma dimensão, são utilizadas técnicas de segurança em nível de linha dinâmica (do inglês *Dynamic Row Level Security* - DRLS). O meio de garantir que o produtor possa apenas visualizar dados de suas estabelecimentos rurais por DRLS é através do uso de um atributo pessoal e único, que neste caso, foi utilizado o *e-mail*. Apenas registros com o *e-mail* do produtor logado no sistema poderão ser acessados no nível de detalhe de nome de propriedade. Não foi possível implementar em tempo hábil tal funcionalidade nas ferramentas de BI utilizadas devido à limitação de tempo da execução da proposta, sendo esta uma proposta para trabalhos futuros.

Todas decisões realizadas no projeto, desenvolvimento e implementação do DW afetarão o nível de complexidade do ETL. A próxima seção aborda este processo, considerado o mais complexo e demorado em projetos de DW.

4.5 Sistema ETL

Todo o processo ETL foi realizado com base no modelo dimensional implementado na subseção 4.5.3, apresentado na Figura 30, e implementado com o uso do módulo *spoon* da ferramenta PDI, apresentada na seção 4.3. O ETL inicia-se pela construção da *stage area*, local onde os dados extraídos dos sistemas fontes são armazenados. Este processo ocorre para não sobrecarregar os sistemas fontes com processamentos e análises intrínsecas deste processo, que visam garantir a qualidade e consistência dos dados, além da carga destes no DW. Outra questão é a de que os sistemas fontes estão disponíveis para extração dos dados em determinados períodos de tempo, que nem sempre são os mesmos. Ainda, são necessárias avaliações de determinadas condições que exigem a execução de junções (*joins*) entre informações de dois ou mais sistemas, sendo isto possível apenas se os dados estiverem no mesmo banco de dados físico. Portanto, a *stage area* será o local onde os dados serão tratados, após armazenados os dados de todos os sistemas.

A sequência desta seção está dividida em duas subseções. A subseção 4.5.1 aborda todos os problemas encontrados nos modelos e dados dos sistemas fontes de dados e propostas de mudanças para viabilizar a integração dos dados. A subseção 4.5.2 aborda a implementação do processo de ETL na máquina servidora.

4.5.1 Problemas encontrados

O principal problema detectado nas reuniões com o *stakeholder* interno e o desenvolvedor do sistema LS, que inviabilizaria o processo de integração de dados é a inexistência de tabelas com informações da propriedade e do usuário. O sistema LS foi concebido como uma solução prática e rápida para o cálculo da produtividade, da emissão da pegada de carbono total e a emissão em função da produtividade do sistema produtivo. Na concepção do projeto do LS, não foi prevista a necessidade de informações de cadastro do usuário ou da propriedade. Então, a primeira sugestão feita para o *stakeholder* interno e analisada conjuntamente com o desenvolvedor responsável pelo LS foi o acréscimo de tabelas no banco de dados que permitiriam identificar o usuário e sua propriedade unicamente, de forma que fosse possível alinhar com as informações existentes no sistema FGC. O resultado deste processo foi a criação das tabelas usuário e propriedade no sistema LS, idênticas às existentes

no sistema FGC. A variável utilizada para alinhar e integrar os dados dos sistemas é o *e-mail* do usuário, que é uma forma única de identificá-lo (independentemente de onde reside). Como um usuário pode ter um ou mais estabelecimentos rurais, o nome da propriedade será o atributo utilizado para alinhar as métricas dos estabelecimentos rurais entre os sistemas.

A segunda sugestão foi a alteração do relacionamento entre as tabelas ‘propriedade’ e ‘sistema produtivo’ de 1:1 para 1:N, além da inserção da coluna ‘ano’ na tabela ‘sistema produtivo’. Isto se deve ao fato de que, apesar dos registros dos estabelecimentos rurais dos usuários do sistema LS serem anuais, não há referências ao ano no modelo do banco de dados. Portanto, haverá uma coluna na tabela ‘sistema produtivo’ referente ao ano dos dados inseridos para cada simulação. No momento, é possível realizar várias simulações para uma mesma propriedade, desde que o ano para o registro seja diferente.

Outro problema identificado foi nos dados recebidos para serem utilizados como base para o teste do ETL e visualização. Os dados utilizados como base para o sistema LS são parte de uma dissertação de mestrado em andamento no PPGCAP e foram originados de duas fontes: questionários *online* aplicados para produtores da pecuária de corte e dados disponíveis na dissertação de mestrado de Marques (2011). Os dados utilizados da dissertação de Marques (2011), apesar de serem úteis, não são suficientes para o preenchimento de todos os requisitos informacionais do sistema LS, portanto, parte dos dados advindos desta dissertação foram simulados. Ainda, os dados de entrada para o sistema produtivo e as categorias de animais servem como subsídio para um modelo de estimativa da pegada de carbono e cálculo de produtividade da propriedade. Se o produtor não conhece todos os dados necessários, podem ser gerados e utilizados valores considerados padrões, de acordo com diferentes parâmetros da pecuária de corte. Portanto, os valores resultantes de emissão e produtividade podem ser reais ou simulados, dependendo no nível informacional do produtor que preencheu as informações no questionário para o sistema LS. Tal problemática, sobre o que é dado real ou simulado para tratamento no ETL, deverá ser abordada em trabalhos futuros, visto que os sistemas que abastecem o DW não estão consolidados.

No caso do sistema FGC, não há dados disponíveis, visto que o sistema se encontra em fase de testes, não tendo sido coletados dados reais de produtores. Como alternativa, definida conjuntamente com o *stakeholder* interno, para testar e

validar o processo de ETL e subsidiar o processo de visualização de dados com custos incluídos, sugeriu-se utilizar parâmetros internos do *MyBeef*, os mesmos que foram anteriormente utilizados para gerar as saídas do sistema LS, para inclusão no sistema FGC. O parâmetro do *MyBeef* utilizado foi o número de animais por categoria animal, informação que é relevante não só para gerar e calcular métricas para o DW, mas também para estimar os custos por tipo de custo do sistema produtivo.

Considerando-se um sistema de produção de ciclo completo com 500 animais em pastagem nativa com 66% de taxa de desmame e idade de abate de 30 meses, o custo anual por animal foi de R$ 415,00, conforme apontado em um cenário específico do ANUALPEC (2015). Este foi um valor similar encontrado e foi utilizado como base para simular os valores de custos dos registros simulados para o sistema FGC, com o objetivo de testar e validar a solução de ETL e analisar as possibilidades de visualização destes dados.

4.5.2 Desenvolvimento e implementação do processo ETL

A *stage area* foi implementada no SGBD *PostgreSQL*, no mesmo servidor do DW, porém em um *schema* diferente. Todo o processo de ETL foi desenvolvido na ferramenta PDI, apresentada anteriormente na seção 4.3.

Foram utilizados padrões de nomenclatura para o projeto físico da *stage area*, assim como o projeto da arquitetura do ETL, visando simplificar o entendimento da solução implementada para futuros desenvolvedores. Os nomes de todas as tabelas relacionadas com a *stage area* seguem o padrão ST_Y_W_Z, onde:

- ST é o prefixo, sigla para *Stage Area*, indicando que a tabela é parte dela;
- Y refere-se ao processo realizado que originou os dados da tabela. Tem-se duas possibilidades, sendo ‘extracao’ para o processo de extração e ‘transformacao’ para o processo de transformação dos dados;
- W refere-se ao nome da(s) tabela(s) relacionada(s) ao sistema fonte de dados do qual serão utilizados ou tratados os dados;
- Z refere-se ao sistema do qual os dados foram originados. Tem-se duas possibilidades, sendo ‘ls’ para o sistema LS e ‘fgc’ para o sistema FGC.

Para a nomeação das colunas das tabelas, foi utilizado o padrão X_Y_Z, onde:

- X é o prefixo, indicando o significado da coluna, em termos de ETL. Os valores possíveis são ‘pk’ (*primary key* – chave primária), ‘nk’ (*natural key* – chave natural), ‘fk’ (*foreign key* – chave estrangeira), ‘nm’ (nome, variável do tipo *string*), ‘nr’ (número, variável do tipo inteiro), ‘vl’ (valor, variável do tipo real ou *numeric*) e ‘per’ (percentual, variável do tipo real ou *numeric*);
- Y é a variável intermediária que indica o significado semântico, em termos de métrica de negócio, daquela coluna. Alguns exemplos como ‘ndt_suplementos_touros’, ‘peso_medio_vacas, ‘peso_venda_vacas’, entre outros, foram utilizados com frequência;
- Z é a variável que indica o sistema fonte dos dados daquela coluna, sendo ‘ls’ para o sistema LS e ‘fgc’ para o sistema FGC;

Para uma análise mais detalhada destes padrões no projeto físico do banco de dados, sugere-se a análise do script de criação da *stage area*, documentado no apêndice B deste projeto, onde é detalhado todo o processo de ETL.

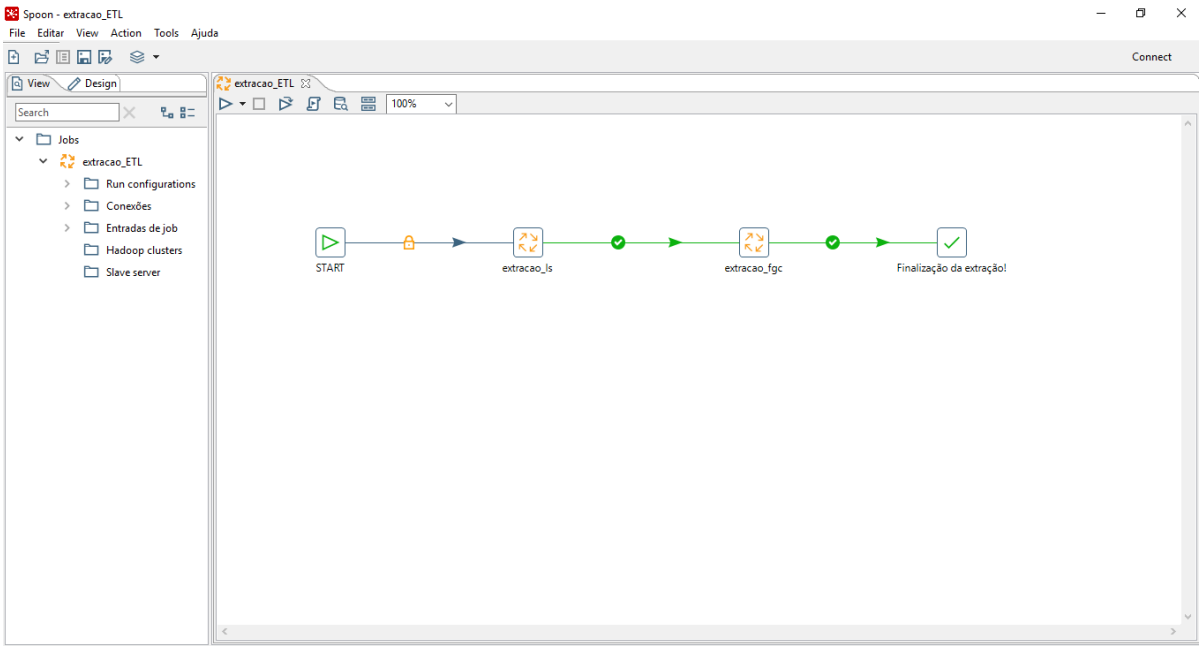
Para fins de simplificação do projeto de ETL, o processo foi dividido em três conjuntos de etapas: extração; transformação e integração; e carga. Dois conjuntos de tabelas foram utilizados para a *stage area*, sendo um deles para armazenar os dados extraídos dos sistemas fontes e o outro para armazenar os dados transformados, tratados e integrados na etapa de transformação. Tal separação foi realizada para evitar a mistura de dados tratados e validados com dados ainda não avaliados. Ainda, na etapa de extração, foram derivados dois subconjuntos de tabelas, um para armazenar dados do sistema LS e outro para os dados do sistema FGC. A etapa de carga é o momento em que os dados são, de fato, consolidados no DW.

Os processos do módulo *spoon* do PDI se dividem em Tarefas (*Jobs,* arquivos com extensão .kjb) e Transformações (*Transformations,* arquivos com extensão .ktr). As Tarefas atuam em um nível de abstração maior, sendo possível agendar a execução de uma ou várias Tarefas e Transformações (por exemplo, diariamente, semanalmente ou mensalmente), enviar *e-mail* em caso de falha ou sucesso de determinada Tarefa ou Transformação, gerenciar arquivos e conexões com as bases de dados necessárias para o ETL, checagem de condições, execução de scripts, entre outras funções. Já as transformações são funções direcionadas ao processamento de dados, como leitura de dados de planilhas e bases de dados, realização de operações nas bases de dados com scripts SQL, uso de filtros, realização de cálculos, entre

outros. As operações realizadas pelas transformações são chamadas de passos (*steps*). A transferência dos dados entre os passos se dá através dos saltos (*hops*) que representam o fluxo de dados entre uma fonte e um destino, que podem ser saltos, Transformações ou Tarefas. Ainda, é possível definir destinos distintos para um passo (válido também para transformações e tarefas) quando o mesmo obtém sucesso ou falha, fornecendo maior flexibilidade no processo de ETL.

As Tarefas e Transformações foram alocadas em três diretórios de acordo com o processo relacionado: extração, transformação e integração, e carga. Ainda, os arquivos relacionados foram nomeados com o padrão X_Y_Z, onde X representa a parte do processo (extração, transformação, integração ou carga), Y representa o processo de negócio, tabela ou parte específica do ETL com o qual são movimentados os dados e Z representa o sistema fonte dos dados daquela tarefa. Padrões semelhantes também foram utilizados na nomeação de cada um dos passos dentro de cada transformação e tarefa, conforme pode ser conferido a seguir, na Figura 31, onde é realizado o processo de extração dos dados dos sistemas fontes de dados para as tabelas da *stage area*.

Figura 31: Processo de extração dos dados na interface do PDI.

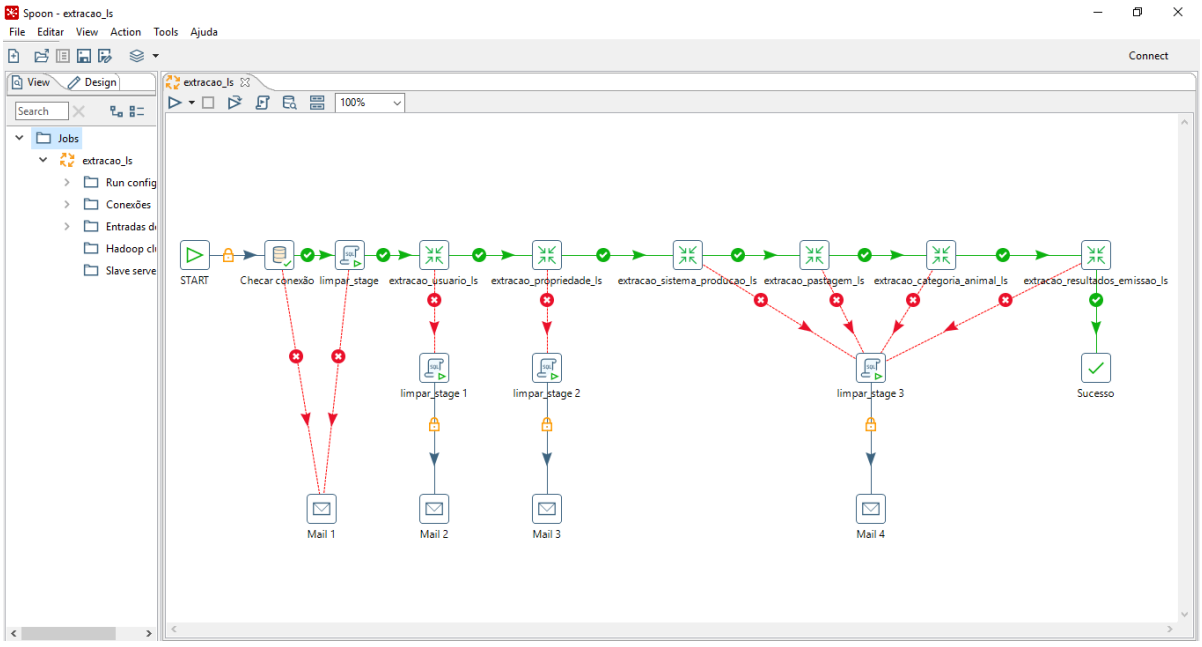


Fonte: Autor (2019).

Processo de extração dos dados

A Figura 31 apresenta duas tarefas, cada uma para extrair dados do respectivo sistema fonte de dados (LS e FGC). Devido à semelhança do processo de extração entre os sistemas, será apresentada apenas a tarefa de extração de dados do LS, que pode ser analisada em detalhes na Figura 32.

Figura 32: Processo de extração dos dados do LS na interface do PDI.



Fonte: Autor (2019).

É possível observar, na Figura 32, o seguinte fluxo de etapas para a extração:

1. **‘checar conexão’**: verifica a conexão com o banco de dados do sistema LS (configurado localmente) e com o banco de dados da *stage area* (configurado localmente). O mesmo procedimento foi realizado no processo de extração de dados do FGC. Em caso de falha na conexão, é disparado um *e-mail* para o administrador do sistema alertando o dia e horário da falha, assim como o passo, tarefa e transformação que originou a falha, e a Tarefa interrompe sua execução. O nível de detalhe dos erros pode ser modificado pelo usuário no passo de envio do *e-mail*. Em caso de sucesso, o fluxo da Tarefa prossegue para a próxima etapa. Para o envio do *e-mail*, foi utilizado o servidor SMTP (*Simple Mail Transfer Protocol*) padrão do *gmail*, com as respectivas informações de *login* e senha do

administrador do sistema. A restrição para o uso do servidor SMTP do *gmail* é o envio de, no máximo, 100 *e-mails* por dia.

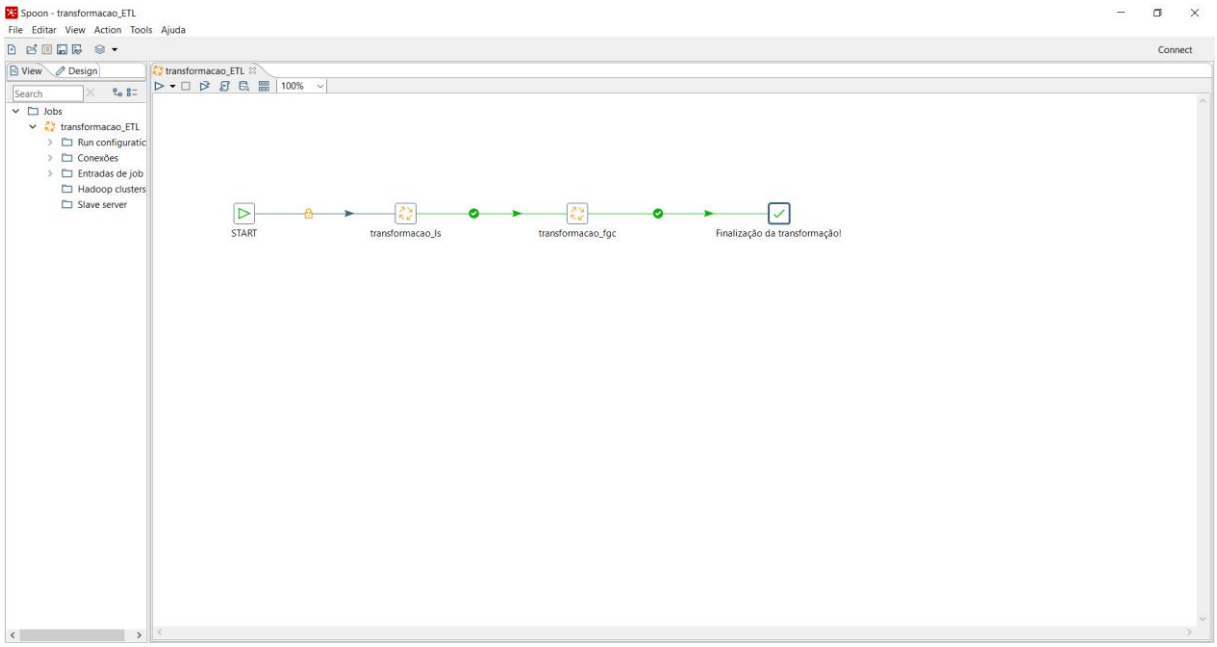
2. **‘limpar_stage’**: é realizado o truncamento das tabelas da stage_area que receberão fluxos de dados do sistema LS. Este passo é executado através de um comando SQL (TRUNCATE TABLE "stage_area"."st_usuario_extracao_ls" CASCADE) que deleta todos os dados contidos de outras execuções prévias em todas as tabelas relacionadas ao sistema LS no processo de extração. O comando CASCADE foi utilizado para garantir a limpeza das tabelas de usuário e todas tabelas que possuam uma chave ligada à ela. Este passo é necessário para garantir que apenas os dados extraídos do LS daquele dia serão processados. Se este passo falhar, o mesmo processo descrito no passo 1, em caso de falha, ocorrerá.
3. **‘extracao_usuario_ls’**: é uma transformação que realiza a extração dos dados da tabela de usuário do LS e os carrega para a tabela de usuário do LS na *stage area*. Em caso de falhas nesta ou nas próximas transformações: qualquer dado anterior extraído é deletado da *stage area* (através do mesmo comando do passo 2) e o processo do passo 1 em caso de falha é executado (envio de *e-mail* para administrador).
4. **‘extracao_propriedade_ls’**: é uma transformação que realiza a extração dos dados da tabela de propriedade do LS e carrega-os para a tabela de propriedade do LS na *stage área.*
5. **‘extracao_sistema_producao_ls’**: é uma transformação que realiza a extração dos dados da tabela de sistema de produção do LS e os carrega para a tabela de sistema de produção do LS na *stage area.*
6. **‘extracao_pastagens_ls’**: é uma transformação que realiza a extração dos dados das tabelas de composições e pastagens do LS e os carrega para a tabela de sistema de produção do LS na *stage area.*
7. **‘extracao_categoria_animal_ls’**: é uma transformação que visa realizar a extração dos dados das tabelas de animais e suplementos do LS e carregar para a tabela de sistema de produção do LS na *stage area.*
8. **‘extracao_resultado_emissao_ls’**: é uma transformação que visa realizar a extração dos dados das tabelas ‘resultados de produtividade’ e ‘resultados de emissão’ do LS e carregar para a tabela de sistema de produção do LS na *stage area.*

Cabe mencionar que, devido à inexistência de colunas que indiquem a data e hora de inserção ou atualização dos registros nos sistemas fontes, foi adotado um esquema de carga completa (*full load*) em que todos os dados sempre serão extraídos de ambos sistemas fontes. Com o sucesso da execução das tarefas e transformações anteriores, inicia-se o próximo passo do ETL, a transformação dos dados, onde são executadas tarefas de pré-processamento dos dados como preenchimento de algumas informações faltantes, correção e conversão de tipos dos dados e validação de informações. Cabe mencionar que checagens de integridade entre chaves primárias e estrangeiras (de entidade e referencial) das tabelas são realizadas automaticamente, pois no processo de extração, se ocorrer de algum registro possuir chave estrangeira que seja inexistente na tabela relacionada ou não possuir chave primária, o processo de extração é imediatamente interrompido.

Processo de transformação dos dados

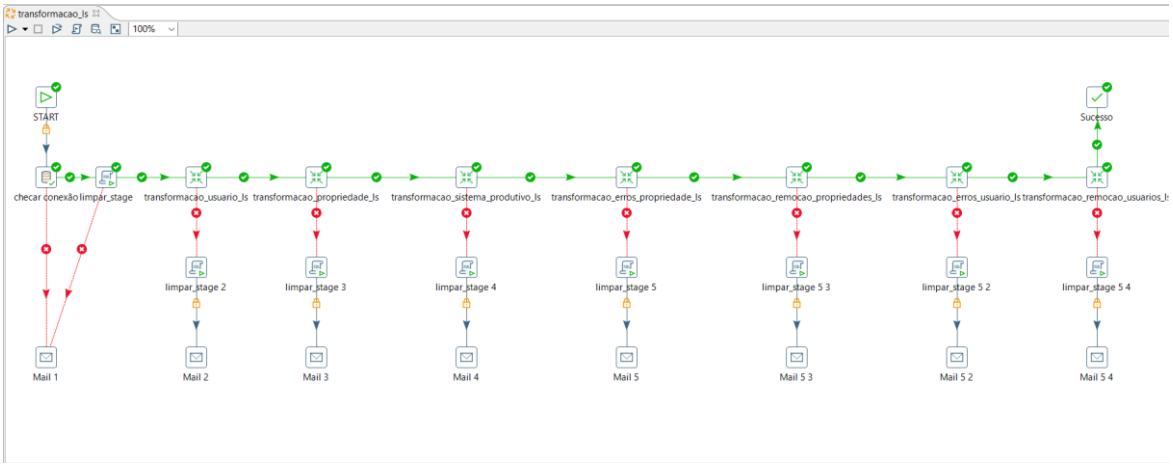
A Figura 33 apresenta duas tarefas, cada uma para transformar, processar e integrar os dados do respectivo sistema fonte de dados (LS e FGC). A tarefa de transformação de dados específica do LS e a ordem de seus respectivos passos podem ser analisados na Figura 34. O processo de transformação da FGC ocorreu de maneira semelhante, porém também foi realizado o processo de integração com os dados do sistema LS. No apêndice C deste projeto é possível analisar em detalhes cada um dos passos e o significado dos mesmos no contexto das transformações e tarefas do ETL.

Figura 33: Tarefa de transformação dos dados na interface do PDI.



Fonte: Autor (2019).

Figura 34: Tarefa de transformação dos dados específica do LS.



Fonte: Autor (2019).

Todos os registros que por quaisquer razões não tenham sido inseridos nas tabelas de transformação da *stage area* são inseridos em um outro conjunto de tabelas que se convencionou chamar de '**stage_erro_usuario**', '**stage_erro_propriedade**', e '**stage_erro_sistema_produtivo**'. São tabelas semelhantes às do processo de transformação, porém com duas colunas adicionais em cada tabela, sendo uma para o registro da data e hora da ocorrência de determinada violação de restrição e outra para identificar a razão do erro. Convencionou-se a existência de três tipos de erros,

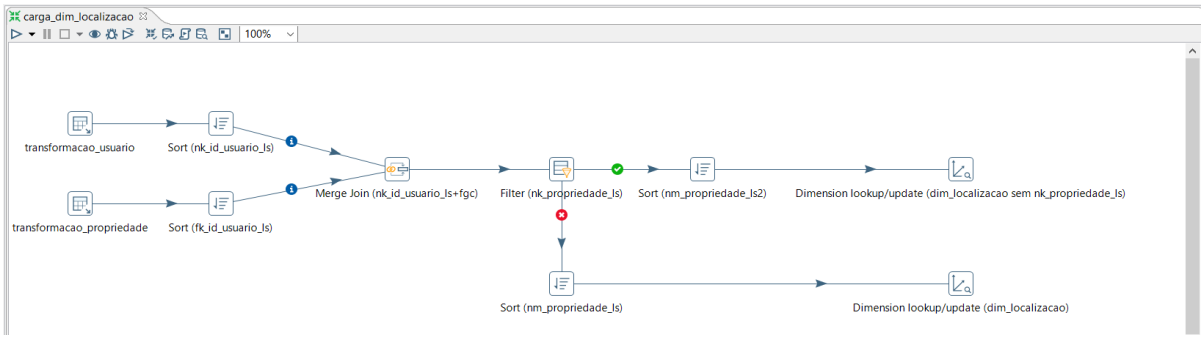
sendo atribuído o valor ‘1’ quando o erro for relacionado à violação de regras de negócio e valores inválidos, incorretos ou nulos, o valor ‘2’ quando o erro envolver violação de restrições de integridade referencial e o valor ‘3’ quando o registro na tabela não possuir dados nas tabelas de sistema produtivo. Este conjunto de tabelas de erros são persistentes, caso o administrador do sistema necessite realizar correções manuais, ele poderá localizar, acessar e corrigir tais registros, seja no processo de extração ou nos próprios sistemas fontes, caso possua autorização. Ainda, tais registros subsidiarão ações de correções e melhorias nos próprios sistemas fontes de dados, visando obter melhorias na qualidade dos dados extraídos para os tratamentos no ETL.

Processo de carga dos dados

O processo de carga do ETL inicia-se a partir da carga dos dados para as dimensões. Inicialmente, é realizado o processo de *lookup* nas dimensões (verificação dos dados já presentes na dimensão através de uma chave). Se as dimensões já possuem dados relacionados ao registro que seria inserido, nenhuma ação é realizada. Caso um novo registro na dimensão seja detectado, é realizada a inserção (*insert*) na dimensão. Conforme mencionado anteriormente, todas as dimensões, com exceção da temporal, possuirão o histórico das mudanças (SCD tipo II), com um versionamento das mesmas. Esse fato visa atender eventos como: o produtor mudar o nome da propriedade, mudar de localização ou a propriedade passar a possuir um novo dono, uma área maior, pastagens diferentes, entre outros. A dimensão de tempo possuirá apenas registros dos anos, não sendo necessários versionamentos. A dimensão de faixas de área só vai mudar caso ocorram alterações no tamanho da propriedade, implicando na atualização dos registros nas tabelas fato. Por fim, caso ocorram mudanças nas dimensões de pastagens e suplementos, o registro é sobrescrito.

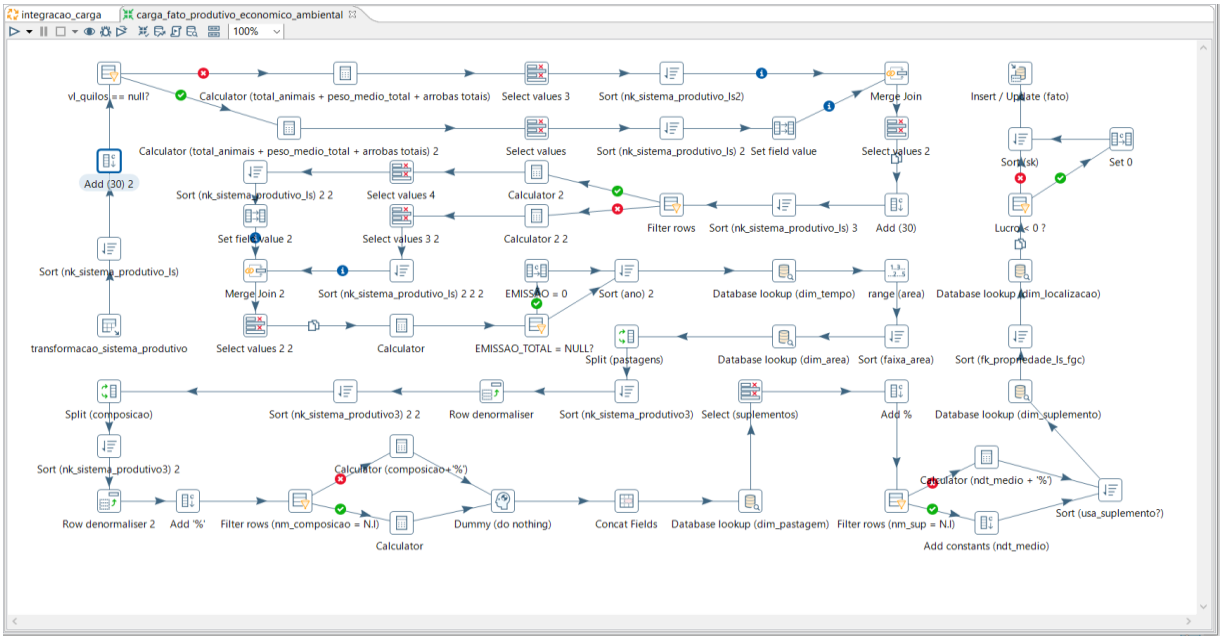
A Figura 35 ilustra os passos da carga da dimensão de localização do DW. Processos semelhantes foram realizados para todas as outras dimensões, diferindo apenas na complexidade de obtenção das chaves e atributos. A Figura 36 ilustra os passos para a carga da tabela ‘**fato_produtivo_economico_ambiental**’. Os detalhes técnicos da implementação do processo de carga do ETL podem ser conferidos no Apêndice B deste trabalho.

Figura 35: Tarefa de carga dos dados na dimensão ‘dim_localizacao’.



Fonte: Autor (2019).

Figura 36: Carga na tabela ‘fato_produtivo_economico_ambiental’.



Fonte: Autor (2019).

Após a realização dos testes com todas as tarefas e transformações do processo de ETL para sua verificação e validação, foram computados os tempos de execução de cada parte do processo. A Tabela 15 possui a compilação destes resultados, onde observa-se que o gargalo do processo é a transformação dos dados, visto que muitas verificações e validações de dados são necessárias consumindo a maior parte do tempo do ETL.

Mesmo com o tempo total consumido (22.6 segundos para 100 registros, sendo 4 deles para testes de erros) e supondo a não ocorrência de falhas que interrompam o ETL, é possível fornecer ao DW dados atualizados de ambos os sistemas legados com uma frequência, pelo menos, diária, resolvendo assim um dos principais

problemas enfrentados: implementação de um processo de integração de dados de fontes de dados heterogêneas. Finalizado o ETL, na próxima seção é abordado o processo de exploração de dados com o servidor OLAP *Mondrian*.

Tabela 13: Resultados de desempenho para o processo de ETL.

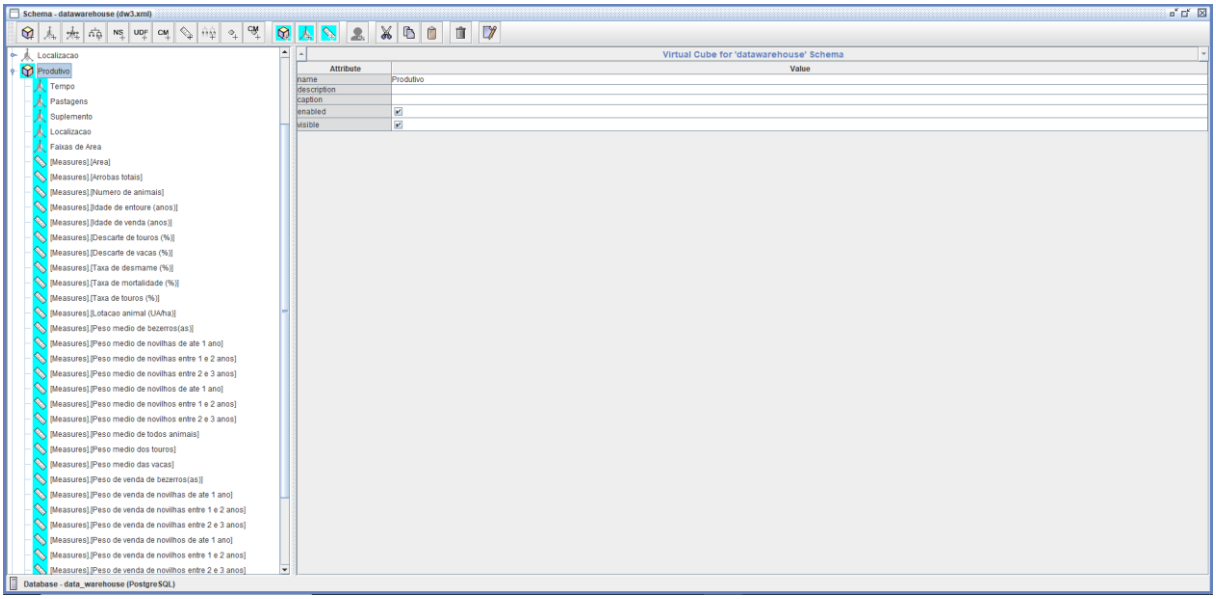
Processo do ETL	Percentual do tempo total de execução	Transformação/Tarefa	Tempo (segundos)	% de tempo
Extração	**13%**	extracao_truncate_ls	0,27	15%
		extracao_usuario_ls	0,10	6%
		extracao_propriedade_ls	0,12	6%
		extracao_sistema_produtivo_ls	0,15	8%
		extracao_pastagens_ls	0,51	28%
		extracao_categoria_animal_ls	0,50	27%
		extracao_resultados_emissao_ls	0,17	9%
		extracao_ls	**1,82**	**100%**
		extracao_truncate_fgc	0,28	15%
		extracao_usuario_fgc	0,22	19%
		extracao_propriedade_fgc	0,15	13%
		extracao_n1p1_fgc	0,14	12%
		extracao_n1p2_fgc	0,10	8%
		extracao_n1p3_fgc	0,18	15%
		extracao_n1p4_fgc	0,10	8%
		extracao_fgc	**1,16**	**100%**
Transformação e Integração	**51%**	transformacao_truncate_ls	0,22	4%
		transformacao_usuario_ls	0,21	4%
		transformacao_propriedade_ls	2,19	38%
		transformacao_sistema_produtivo_ls	2,59	45%
		transformacao_erros_remocao_ls	0,60	10%
		transformacao_ls	**5,81**	**100%**
		transformacao_usuario_fgc	1,02	18%
		transformacao_propriedade_fgc	3,01	53%
		transformacao_sistema_produtivo_fgc	1,01	18%
		transformacao_erros_remocao_fgc	0,64	11%
		transformacao_fgc	**5,67**	**100%**
Carga	**36%**	carga_dim_tempo	0,15	8%
		carga_dim_area	0,22	13%
		carga_dim_pastagem	0,23	13%
		carga_dim_suplemento	0,26	15%
		carga_dim_localizacao	0,90	51%
		carga_dimensões	**1,76**	**100%**
		carga_fatos	**6,35**	**100%**

Fonte: Autor (2019).

4.6 Configuração do servidor OLAP

Com os dados tratados armazenados no DW, o próximo passo é realizar a configuração dos cubos de dados para apresentação na interface da aplicação. Para realizar essa configuração foi utilizado o PSW, que permite configurar os cubos através de uma interface gráfica. A Figura 37 apresenta a interface do PSW e o exemplo de configuração do cubo '**Produtivo**', onde é possível observar no lado esquerdo da imagem as dimensões deste cubo (tempo, faixas de área, pastagens, suplemento e localização) e logo abaixo, as métricas produtivas. Os outros cubos foram configurados de maneira semelhante.

Figura 37: PSW e configuração do cubo 'Produtivo'.



Fonte: Autor (2019).

Com o PSW, é possível atribuir nomes para as colunas das dimensões e fatos (diferente dos nomes na tabelas físicas), realizar diferentes tipos de operações com os dados, quando agregados na aplicação (média, soma, mínimo, máximo ou contagem), definir dimensões conformadas, atribuir regras de acesso com base nas credenciais do usuário, entre diversas outras possibilidades. Tais funções podem ser realizadas no PSW, após realizar a sua conexão com a base de dados do DW.

Como resultado da configuração dos cubos, são gerados arquivos em XML que possuem a estrutura do cubo através do uso de tabelas fatos e dimensões encontradas no servidor do DW. Estes metadados em XML gerados são interpretados pelo *Mondrian* em conjunto com a base de dados do DW. A Figura 38 apresenta um

exemplo de arquivo XML gerado para o cubo '**fato_produtivo**'. O mesmo procedimento é realizado para os cubos '**fato_economico**' e '**fato_ambiental'**, e com isso, a configuração dos cubos para a apresentação dos dados na interface da aplicação é concluída.

Figura 38: Arquivo XML com parte das informações do cubo 'Produtivo'.

Fonte: Autor (2019).

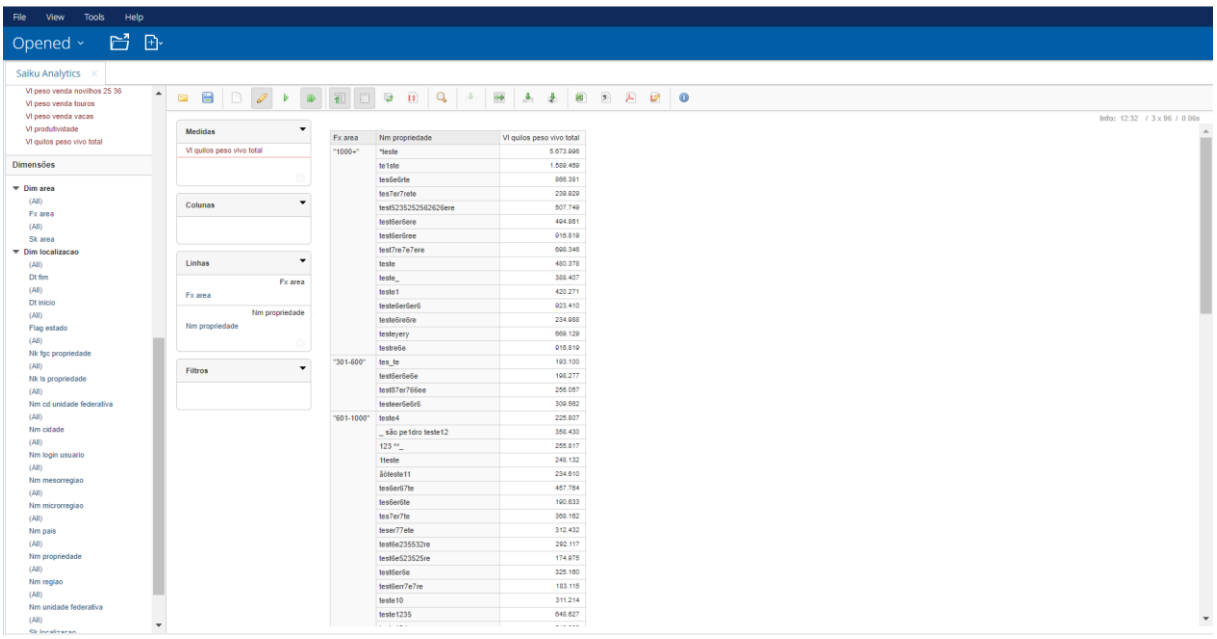
5 RESULTADOS E DISCUSSÕES

Neste capítulo serão abordados os resultados atingidos e discussões críticas sobre a solução. A seção 5.1 apresenta os resultados atingidos com a proposta, através da visualização dos dados do DW, tratados previamente no ETL, na ferramenta de BI *Saiku Analytics*. A seção 5.2 apresenta uma avaliação crítica das limitações e possibilidades da solução proposta.

5.1 Exploração dos dados

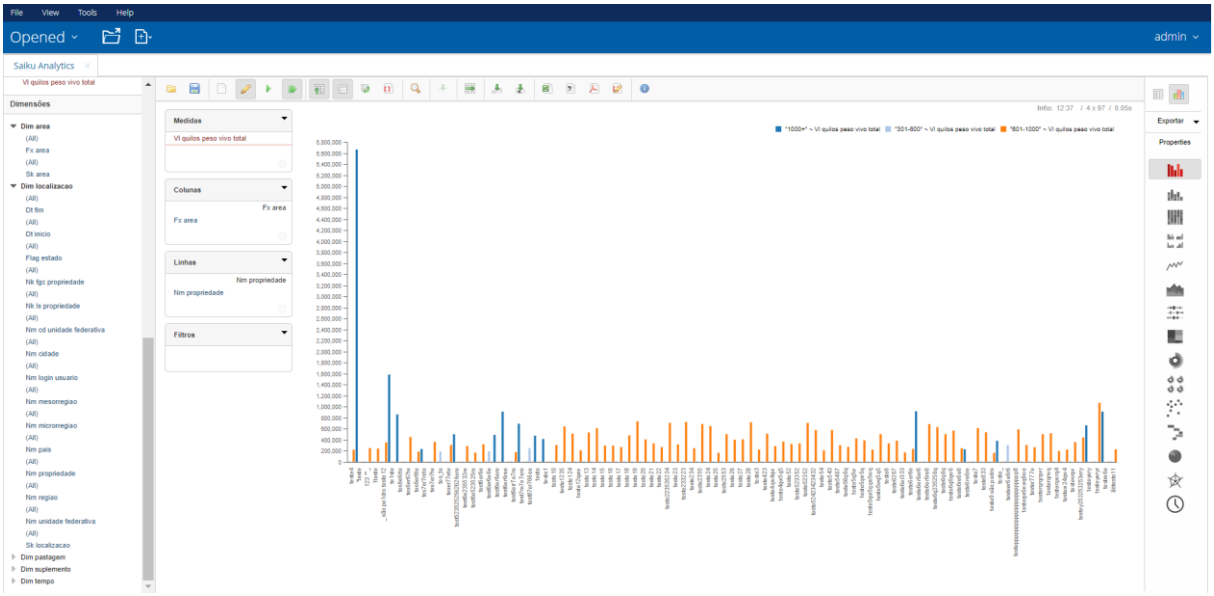
Para realizar a exploração e análise dos dados, foram utilizadas as ferramentas Pentaho BI server versão 8.1 como servidor da interface da aplicação e o *plugin Saiku Analytics* para exploração dos cubos de dados nesta interface. No *Saiku*, é possível realizar diferentes operações com os cubos de dados e apresentar os resultados na tela tanto em formato de tabelas (Figura 39) como gráficos (Figura 40), além de permitir exportar estes resultados em relatório no formato PDF e em planilhas eletrônicas, como o Excel.

Figura 39: Exemplo de consulta no cubo Produtivo no *Saiku Analytics.*



Fonte: Autor (2019).

Figura 40: Exemplo de consulta com gráficos no *Saiku Analytics.*



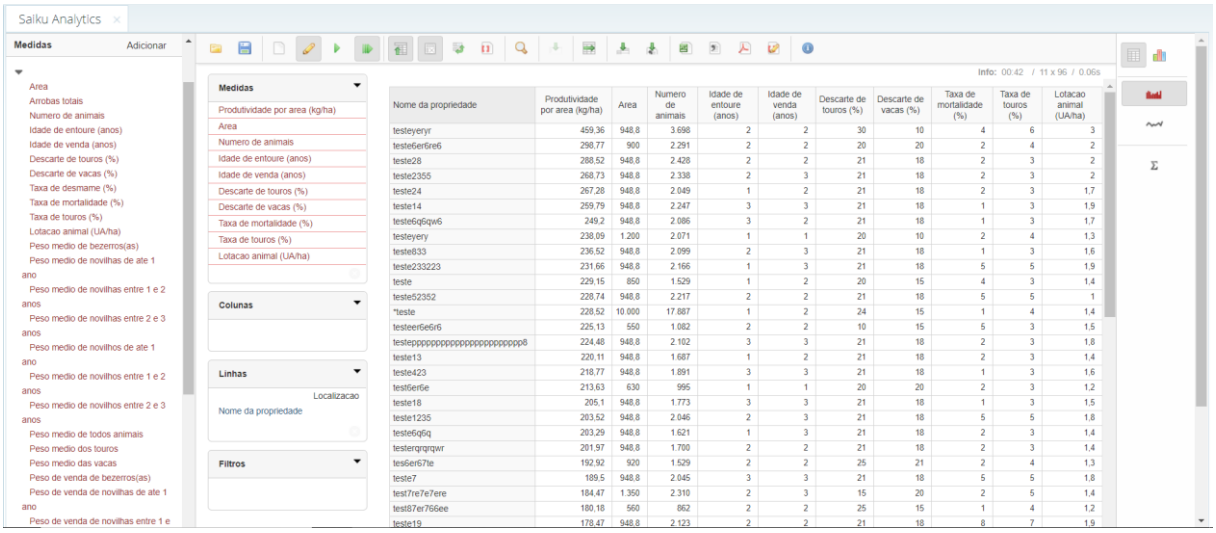
Fonte: Autor (2019).

Na Figura 39 é evidenciado o problema de apresentar o nome da propriedade e suas respectivas métricas. Em vez de apresentar o nome da propriedade, podem ser apresentados valores numéricos sem significado, com o objetivo de preservar a privacidade dos indicadores produtivos e econômicos de estabelecimentos rurais do qual o usuário não é o proprietário. Ao mesmo tempo, deve ser possível que o usuário possa identificar a sua propriedade para viabilizar o processo de análise dos dados.

Com os dados reais coletados, já é possível iniciar algumas análises, como:

1. Quais as características dos sistemas produtivos com maior produtividade? (Figura 41)
2. É possível observar alguma relação entre taxa de desmame e produtividade? (Figura 42)
3. Qual a relação entre a área e os diferentes tipos de custos dos estabelecimentos rurais? (Figura 43)
4. A emissão média de dióxido de carbono por produtividade é maior em estabelecimentos rurais maiores? (Figura 44)
5. Estabelecimentos rurais com maior área possuem maior produtividade? (Figura 45)

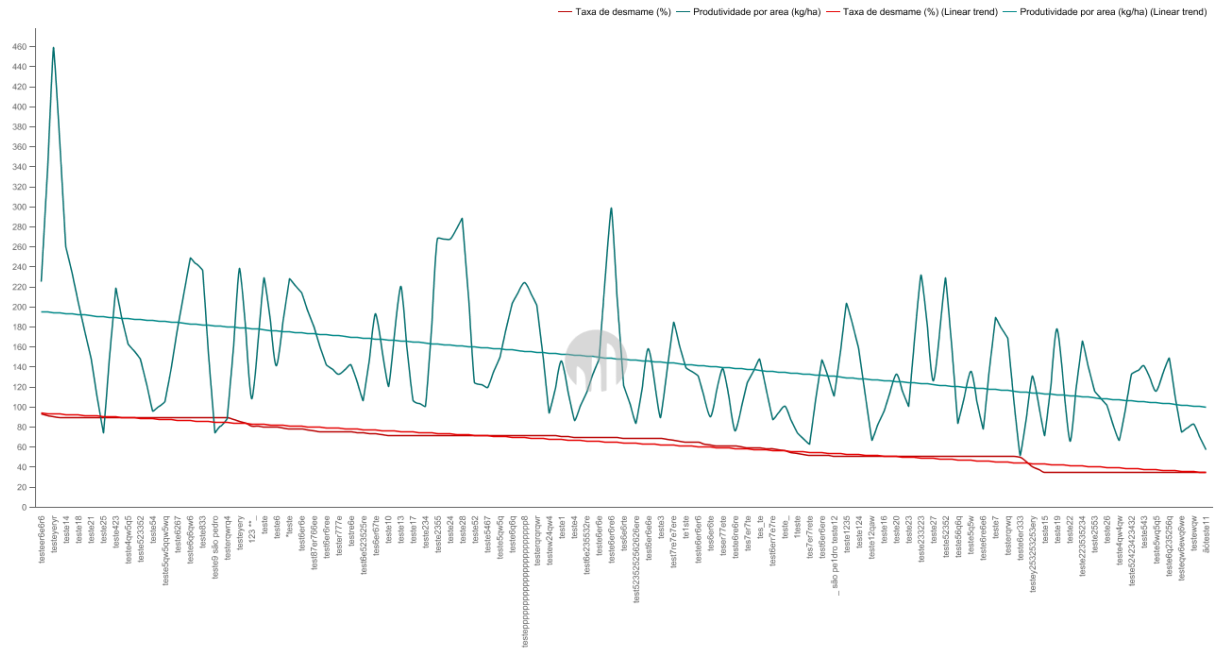
Figura 41: Dados em ordem decrescente por produtividade.



Fonte: Autor (2019).

Nos casos em que é necessário analisar um conjunto muito grande e detalhado dos dados, sugere-se a utilização do formato tabular. A utilização de gráficos neste contexto pode ser de difícil entendimento, principalmente para usuários que não possuem experiência prévia com a utilização de softwares de análise de dados. É possível observar que os estabelecimentos rurais com produtividade (quilos por área ao ano) mais alta são as que possuem a idade de entoure mais baixa, possuem mais animais, tem o percentual de desama mais alto, uma taxa de mortalidade reduzida e lotação animal maior se comparada com o restante dos estabelecimentos rurais. Portanto, fica evidente para o usuário que alternativas devem ser tomadas para realizar a mudança nos seus próprios indicadores zootécnicos visando incrementar a produção e, por consequência, o retorno financeiro.

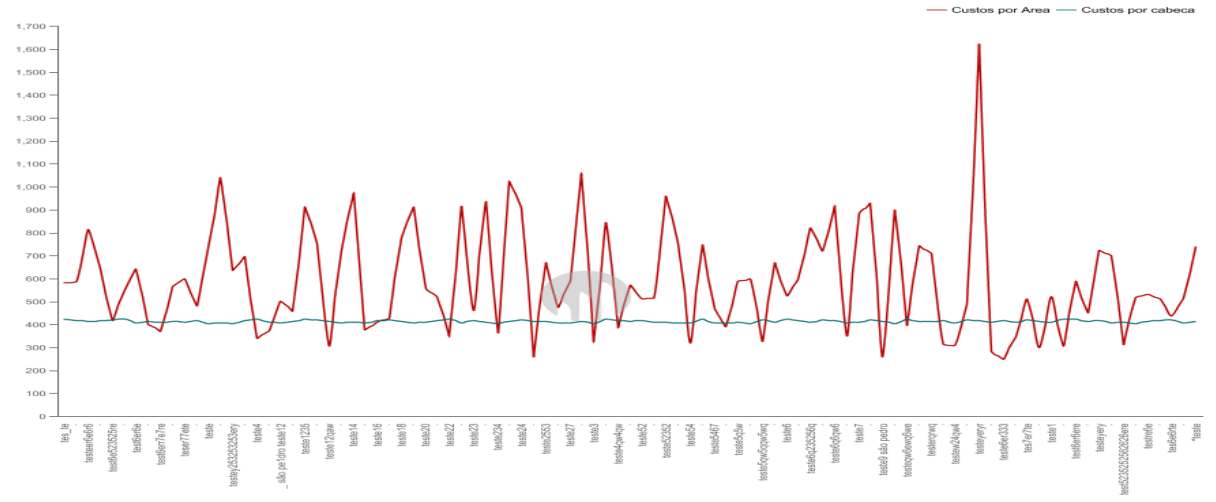
Figura 42: Relação entre desmame e produtividade.



Fonte: Autor (2019).

Na figura 42 é possível observar que, ao realizar a ordenação dos dados por taxa de desmame em ordem decrescente, observamos um padrão de produtividade por área aproximadamente decrescente ocorrendo simultaneamente. Isso sugere para o produtor que um desmame maior pode levar à uma maior produção de carne por área por ano e, consequentemente, ganhos financeiros. Portanto, o sistema aponta qual indicador o produtor necessita melhorar para otimizar o sistema produtivo.

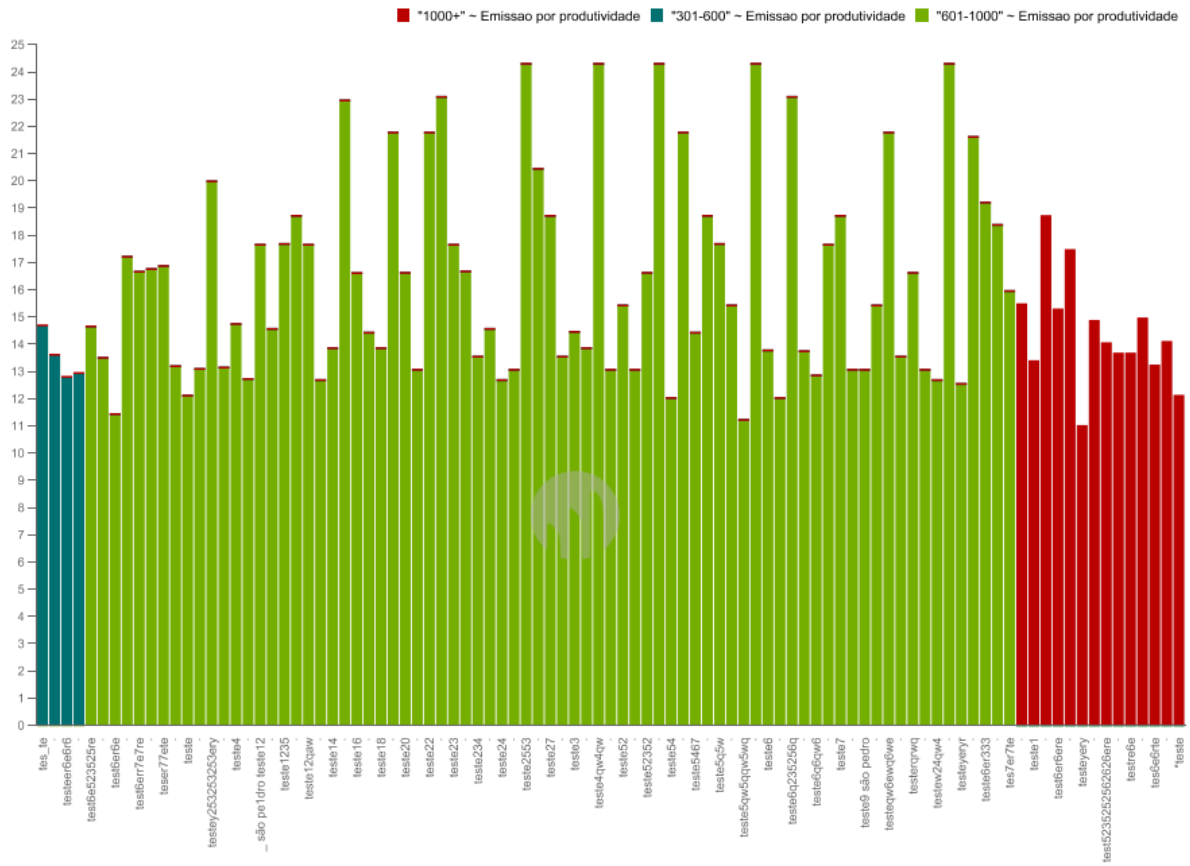
Figura 43: Relação entre custos e área da propriedade.



Fonte: Autor (2019).

Neste cenário (Figura 43) observa-se que a área, em ordem crescente, não exerce influência preponderante no custo por cabeça, porém isto se deve ao fato dos dados de custos terem sido simulados com base no número de animais. Já o custo por área (R$/ha) possui uma variação muito grande entre estabelecimentos rurais menores e maiores. Com os dados disponíveis, foi possível observar apenas que estabelecimentos rurais maiores tem uma tendência de custos por área menor. A variabilidade maior do custo por área é que aproximadamente 30% dos registros de área são dados reais, e para os registros restantes que estavam ausentes foi realizado o cálculo da média simples para preencher essas informações.

Figura 44: Relação entre emissão por produtividade e área.

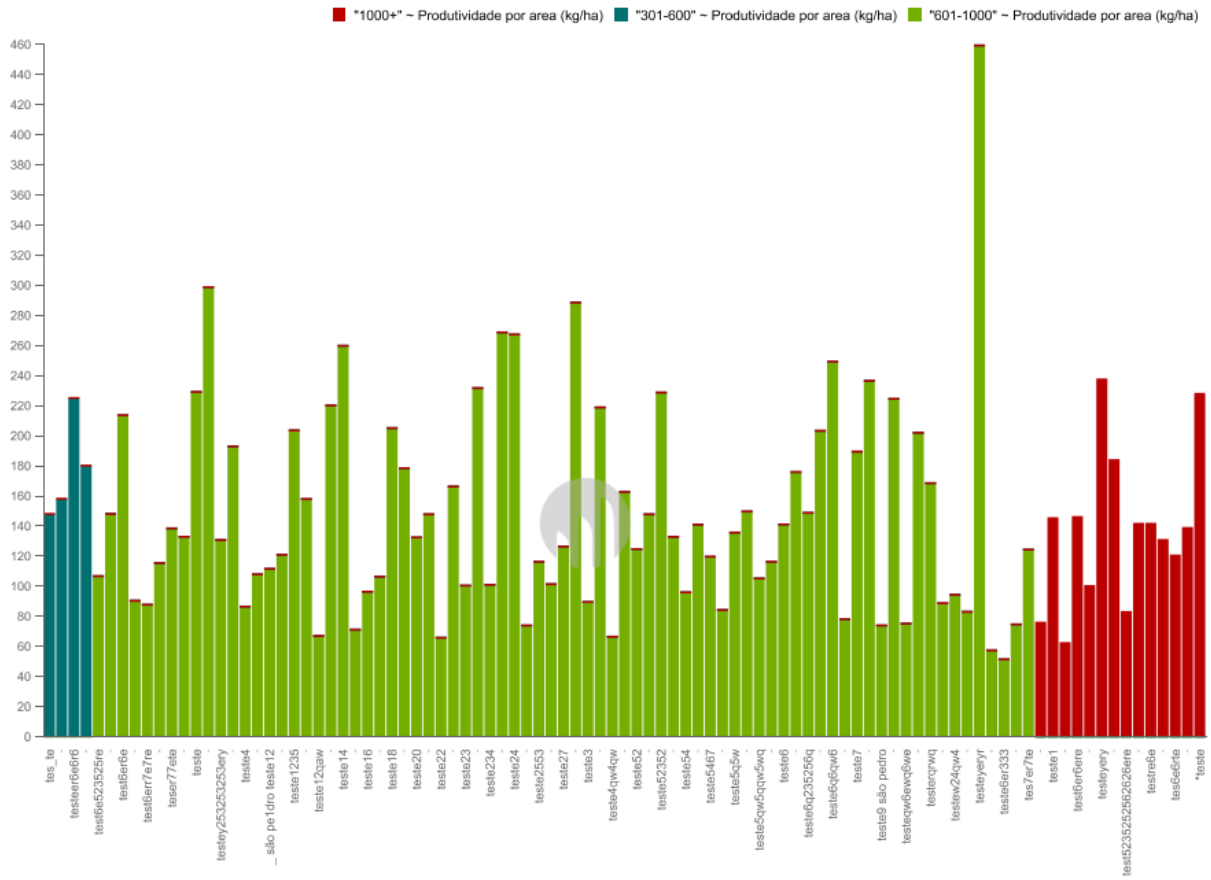


Fonte: Autor (2019).

Neste cenário (Figura 44), com os estabelecimentos rurais ordenados pela área, é possível observar que estabelecimentos rurais maiores não necessariamente causam um maior impacto ambiental por quilo de peso vivo produzido (emissão de

CO_2 por quilo de peso vivo produzido, ou emissão por produtividade). Porém, estabelecimentos rurais menores causam menos impacto (valor absoluto de emissão por produtividade é menor em estabelecimentos rurais menores).

Figura 45: Relação entre área e produtividade.



Fonte: Autor (2019).

Neste cenário (Figura 45), com a produtividade em ordem crescente, observa-se que estabelecimentos rurais maiores não necessariamente produzem mais quilos de peso vivo por hectare. Observa-se mais estabelecimentos rurais menores do que maiores com produtividade alta, o que faz sentido, pois um pequeno aumento da produção em uma propriedade pequena tem um impacto muito maior na produtividade do que na propriedade maior.

Com relação às funcionalidades da aplicação, há possibilidades de geração de gráficos de linha, de barras simples, de barras empilhadas, barras 100% (dados normalizados), de barras múltiplo, de linha, de área, *tree maps*, *sunbursts*, com

pontos, em formato de pizza e radial. É possível customizar as cores dos gráficos, inserir linhas de tendência em gráficos de linha e exportar os gráficos em formatos png e pdf. Também é possível exportar os dados em planilhas no formato .csv e .xls, comuns em aplicações como o Excel, e também em formato .pdf. Tais funcionalidades de exportação podem ser utilizadas caso o usuário fique sem acesso à internet em determinado momento, mas deseja ter acesso permanente aos dados e relatórios. Já o *Pentaho Server* possui um componente chamado de CDE (*Community Dashboard Editor*) para a criação de *dashboards,* outra funcionalidade que pode ser explorada com os dados.

5.2 Discussão dos resultados

A solução desenvolvida possui potencial de utilização, ao mesmo tempo em que existem diversos pontos de otimização que devem ser considerados. A solução de *Business Intelligence*, como um todo, foi capaz de solucionar o problema existente de realizar a integração de dados de dois sistemas de informação com projetos, desenvolvedores, objetivos e até mesmo públicos distintos. A solução, além de realizar a integração e disponibilizar estes dados através de uma interface gráfica com diversas possibilidades de análises, interfere muito pouco nos sistemas legados (interfere apenas no momento da extração dos dados, que pode ser executada durante um período de pouco acesso aos sistemas fontes), não prejudicando o desempenho destas aplicações.

O repositório DW foi capaz de armazenar os dados destes sistemas e fornecer uma visão única das diferentes faces da pecuária de corte (produtiva, econômica e ambiental), resolvendo parte dos problemas de redundância e inconsistências apresentados inicialmente, problemas estes que além de contemporâneos, já foram reportados em diversos trabalhos presentes na revisão da literatura, especialmente nos resultados de um artigo publicado recentemente (BRUM, LAMPERT e CAMARGO, 2019). A solução, além de ser de grande valor para os futuros usuários produtores rurais, pela grande gama de possibilidades de análises com os dados, também se torna um repositório rico de informações para pesquisadores do setor rural, tornando processos analíticos como os de mineração de dados e estatísticas mais simples, pois os dados no DW já passaram por processos de correção e validação no ETL. Cabe mencionar que o sucesso obtido no projeto e implementação

da solução de BI foi em grande parte devido às constantes reuniões e interações entre desenvolvedor e especialista do domínio do negócio, que neste caso também é o *stakeholder* interno, pois a integração envolve um número expressivo de variáveis da pecuária de corte do qual o desenvolvedor da área de computação não possuí o conhecimento necessário a ponto de realizar validações e derivações de outras métricas a partir destes indicadores. O cenário piora se for considerado que não havia nenhuma documentação disponível dos sistemas legados. A necessidade de constante interação entre especialistas do domínio e desenvolvedor da área de computação para projetos de DW já foi apontado nos resultados de uma revisão literária e bibliométrica (BRUM, LAMPERT e CAMARGO, 2019). Também cabe destacar que foram utilizadas apenas ferramentas gratuitas e *open source* para o desenvolvimento do sistema de BI, permitindo uma maior liberdade e flexibilidade para os desenvolvedores implementarem mudanças no próprio software de acordo com suas necessidades.

Está prevista a realização da avaliação da solução com especialistas do setor rural, visando obter críticas e contribuições externas com o objetivo de direcionar os próximos esforços para otimizar o produto em desenvolvimento. Devido às restrições de tempo, não foi possível realizar tal avaliação no prazo de entrega do texto da dissertação, porém a mesma deverá ser feita no futuro, utilizando as teorias de facilidade de uso e utilidade percebida do TAM, através de um questionário avaliativo.

Com relação ao processo de desenvolvimento da solução, foram detectados muitos problemas que inviabilizariam a integração dos dados. Este projeto se tornou particularmente desafiador por lidar com sistemas legados em pleno desenvolvimento, que acarretavam em constantes mudanças no planejamento e execução da proposta. Por outro lado, existe a flexibilidade de se realizar mudanças nos bancos de dados dos sistemas legados que estão menos próximos dos testes de usabilidade, visando facilitar o processo de ETL.

Uma das mudanças realizadas foi a de integrar a FGC com o LS, pois o desenvolvimento da solução do *MyBeef* foi interrompido no meio do ano de 2018. Durante o desenvolvimento e as buscas por soluções alternativas, surgiu a oportunidade de realizar a integração dos dados de custos do FGC com a ferramenta LS, que estava em desenvolvimento por outro mestrando do PPGCAP. Foi acertado com o *stakeholder* interno esta mudança, que seria interessante e desafiadora, pois apesar do *MyBeef* (sistema) ter saído da integração, o sistema LS englobou

internamente os processos de simulação de dados produtivos do mesmo. Portanto, a solução de BI a ser desenvolvida passou a englobar indicadores de três dimensões da pecuária de corte: produtivos, econômicos e ambientais.

Conforme mencionado nos principais problemas para o ETL, o sistema LS não previa tabelas para registro de usuários e estabelecimentos rurais. O processo de integração acabou orientando mudanças positivas no projeto do banco de dados do sistema LS.

O processo de ETL tomou a maior parte do tempo de execução deste projeto, o que já era esperado e apontado por diversos trabalhos da revisão literária, porém houve um tempo maior que teve de ser dedicado para preparar os ambientes de testes do ETL. O processo de aquisição dos dados para os testes vieram no formato de planilha eletrônica, portanto foi utilizado o software *RStudio* para organizar estes dados e inseri-los nas bases de dados locais que simulavam os sistemas legados. Além deste processo, o banco de dados do sistema LS teve de ser implementado localmente, pois não havia ainda uma solução consolidada do mesmo. Ainda, não foi possível obter dados de custos em tempo hábil, e acabaram por ser simulados com base em um cenário bem específico obtido pelo ANUALPEC (2015). Como o objetivo era testar as transformações e tarefas do ETL, assim como analisar as possibilidades de visualização dos dados, não houve prejuízos significativos com o ocorrido.

O ETL desenvolvido executa em um tempo aceitável com os dados disponíveis no momento, porém o cenário ideal de ETL é aquele que realiza cargas incrementais em vez de completas. Isso porque, a carga completa sempre irá realizar o ETL com todos os dados, não importando se foram modificados ou não nos sistemas legados. Com isso, o tempo do ETL se torna extremamente alto com o crescimento do volume de dados. Como atualmente os sistemas legados não possuem suporte nenhum para obter a data ou horário da última atualização das tabelas ou registros, optou-se pela carga completa. Sugere-se, portanto, a implementação de mecanismos nos sistemas legados que permitam identificar o quão atuais ou antigos são os registros presentes nas bases de dados, com o objetivo de reduzir o tempo de execução total do ETL.

Em vários momentos da modelagem do DW foram detectados problemas na granularidade dos dados. A solução final para o DW visou tornar todas as dimensões conformadas, ou seja, todas as métricas de negócio analisáveis por todas as dimensões existentes e utilizou-se como grão da informação o sistema produtivo. Com o amadurecimento dos sistemas legados e inserção de outras fontes, será possível

aumentar o universo de possibilidades de análises, seja com a inserção de novas dimensões, novos indicadores ou novos DM com granularidades distintas e com maior detalhamento.

Com relação ao processo de visualização de dados com o *Saiku Analytics*, é possível observar nas figuras da seção de exploração dos dados alguns nomes incorretos nas variáveis e colunas. Este foi um problema que ocorreu devido à diferença das versões do Mondrian presentes no *Saiku Analytics* e no PSW. O cubo publicado pelo PSW não estava sendo reconhecido pelo *Saiku*, acarretando no uso dos nomes físicos das colunas do DW. Tal fato prejudicou o entendimento dos gráficos e tabelas, porém é um detalhe que deve ser logo corrigido. Uma das deficiências detectadas no *Saiku* é que não há gráficos ou tabelas que realizem a padronização dos dados de atributos com magnitudes muito distintas (por exemplo: lotação animal e quilos de peso vivo). Isso torna inviável analisar indicadores de magnitudes distintas nos gráficos. Portanto esta é outra oportunidade de melhoria na solução, implementar processos de normalização automática nos dados que serão utilizados em gráficos visuais, preponderantemente no caso dos atributos com magnitudes muito distintas.

Por fim, visando dar um suporte para a continuação deste projeto de pesquisa, foi elaborado, além do documento de requisitos com versionamentos (Apêndice A), a documentação completa do sistema de BI (Apêndice B) e cada uma das etapas que permitiram integrar e visualizar os dados dos sistemas legados. Também são fornecidos códigos fontes e orientações para o uso das ferramentas abordadas neste projeto. Por fim, também foi elaborado um guia do usuário, para facilitar os processos de análise de dados com a solução proposta para usuários menos experientes nestes processos.

6 CONSIDERAÇÕES FINAIS

Com o desenvolvimento deste projeto foi possível verificar os principais desafios no desenvolvimento de uma solução completa de BI para produtores rurais no setor da pecuária de corte. A solução desenvolvida atinge o objetivo de se realizar a integração de dados com foco em métricas econômicas, produtivas e ambientais dos sistemas produtivos de dois sistemas legados heterogêneos. Além da integração, foi possível avaliar alguns cenários de análise de dados da pecuária com o *Saiku Analytics,* onde foi possível demonstrar como o sistema pode orientar determinadas decisões com os dados organizados da forma adequada. Por fim, esta solução computacional é um produto que pode ser utilizado tanto por produtores e consultores para o auxílio em decisões e obtenção de informações, como também por pesquisadores que possuem interesses em realizar análises aprofundadas com dados da pecuária de corte do Brasil.

Grande parte das dificuldades enfrentadas foram previamente identificadas na revisão de literatura, como: integração de sistemas heterogêneos, ausência de documentação dos sistemas, modelagem de soluções no qual o desenvolvedor não possui o total domínio das métricas envolvidas, diferentes granularidades dos dados e sistemas desenvolvidos sem o viés nenhum de integração. Também foi percebida uma relativa dificuldade para integrar todo o conjunto de ferramentas gratuitas e *open source* selecionadas para esta solução de BI.

Sugere-se como trabalhos futuros:

- Investigar outras soluções gratuitas e *open source* para BI;
- Implementação de DRLS;
- Avaliar separadamente no ETL os dados simulados e reais.
- Atualizar os modelos dimensionais, em conformidade com o avanço e integração de novos sistemas fontes de dados;
- Integrar regras de mensuração da qualidade dos dados no ETL;
- Integrar novas funcionalidades para visualização de dados no sistema (*dashboards*, novos KPI, entre outros);
- Atualizar os sistemas legados e integrar dados espaciais para a localização dos estabelecimentos rurais.
- Obter um conjunto maior de dados reais para subsidiar de uma melhor maneira as análises dos dados.

REFERÊNCIAS

ANUALPEC. **Anuário da pecuária brasileira**. 22. ed. São Paulo: Instituto FNP, 2015.

ARIYACHANDRA, Thilini; WATSON, Hugh J. Key factors in selecting a data warehouse architecture. **Business Intelligence Journal***, [S.l.],* v. 10, n. 2, p. 19-26, 2005.

ARIYACHANDRA, Thilini; WATSON, Hugh J. Which data warehouse architecture is most successful? **Business Intelligence Journal**, [*S.l.*], v. 11, n. 1, p. 4-6, 2006.

ARNOTT, David; PERVAN, Graham. Eight key issues for the decision support systems discipline. **Decision Support Systems**, [*S.l.*], v. 44, n. 3, p. 657-672, 2008. Disponível em: https://www.sciencedirect.com/science/article/pii/S016792360700169. Acesso em: 15 fev. 2018.

BATINI, Carlo; PALMONARI, Matteo; VISCUSI, Gianluigi. The many faces of information and their impact on information quality. *In:* INTERNATIONAL CONFERENCE ON INFORMATION QUALITY, Paris. **Anais eletrônicos...** Paris: Curran Associates, Inc., 2012, p. 212-228. Disponível em: http://mitiq.mit.edu/ICIQ/2012/2012%20ICIQ%20CDproceedings%20final.pdf. Acesso em: 10 mar. 2018.

BAZERMAN, Max Hal; MOORE, Don. **Processo decisório**. 7. ed. Rio de Janeiro: Elsevier, 2010.

BEYNON, Meurig; RASMEQUAN, Suwanna; RUSS, Steve. A new paradigm for computer-based decision support. **Decision Support Systems**, [*S.l.],* v. 33, n. 2, p. 127-142, 2002. Disponível em: https://www.sciencedirect.com/science/article/pii/S016792360100140. Acesso em: 18 fev. 2018.

BOTELHO, Fernando Rigo; RAZZOLINI FILHO, Edelvino. Conceituando o termo business intelligence: origem e principais objetivos. **Revista Iberoamericana de Sistemas, Cibernética e Informática**, [*S.l.*], v. 11, n. 1, p. 55-60, 2014. Disponível em: http://www.iiisci.org/journal/CV$/risci/pdfs/CB793JN14.pdf. Acesso em: 14 fev. 2018.

BRUM, Luciano Moraes da; LAMPERT, Vinícius do Nascimento; CAMARGO, Sandro da Silva. Business intelligence and data warehouse in agrarian sector: a bibliometric study. **Journal of Agricultural Science**, Richmond Hill, v. 11, n. 2, p. 353-368, 2019. Disponível em: https://doi.org/10.5539/jas.v11n2p353. Acesso em: 12 fev. 2019.

CHAUDHURI, Surajit; DAYAL, Umeshwar; NARASAYYA, Vivek. An overview of business intelligence technology. **Communications of the ACM**, New York, v. 54, n. 8, p. 88-98, 2011. Disponível em: https://doi.org/10.1145/1978542.1978562. Acesso em: 13 fev. 2018.

CHAVES, Roselene de Queiroz et al. Tomada de decisão e empreendedorismo rural: um caso da exploração comercial de ovinos de leite. **Revista Brasileira de Gestão e Desenvolvimento Regional**, Taubaté, v. 6, n. 3, p. 3-21, 2010. Disponível em: http://www.rbgdr.net/revista/index.php/rbgdr/article/view/291/203. Acesso em: 11 jan. 2018.

CHIAVENATTO, Idalberto. **Introdução à teoria geral da administração**. 7. ed. Rio de Janeiro: Elsevier, 2003.

CORR, Lawrence; STAGNITTO, Jim. **Agile data warehouse design:** collaborative dimensional modeling from whiteboard to star schema. Leeds: DecisionOne Press, 2011.

CORREA, Fernando Elias et al. Data warehouse for soybeans and corn market on Brazil. *In:* EUROPEAN FEDERATION FOR INFORMATION TECHNOLOGY IN AGRICULTURE, FOOD AND THE ENVIRONMENT, 2009, Wageningen. **Anais eletrônicos…** Wageningen: Wageningen Academic Publishers, 2009, p. 675-681. Disponível em: https://www.informatique-agricole.org/download/Efita-site/EFITA%20Congresses/EFITA%202009/Section%2013/Elias%20Correa,%20F/Correa.pdf. Acesso em: 15 mar. 2018.

COSTA, Fernando Paim et al. **Indicadores de desempenho na pecuária de corte:** uma revisão no contexto da Plataforma +Precoce. Campo Grande: Embrapa Gado de Corte, 2018. Disponível em: https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1090951/1/Indicadoresdedesempenhonapecuariadecorte.pdf. Acesso em: 20 dez. 2018.

DÍAZ, Josep Curto. **Introducción al business intelligence**. Barcelona: Editorial UOC, 2012.

EASTWOOD, Callum; CHAPMAN, David F.; PAINE, Mark S. Networks of practice for co-construction of agricultural decision support systems: Case studies of precision dairy farms in Australia. **Agricultural Systems**, [*S.l.*], v. 108, p. 10-18, abr. 2012. Disponível em: https://doi.org/10.1016/j.agsy.2011.12.005. Acesso em: 18 abr. 2018.

ECKERSON, Wayne W. Data Quality and the Bottom Line. **TDWI report series**, 2002. Disponível em: http://download.101com.com/pub/tdwi/Files/DQReport.pdf. Acesso em: 10 mar. 2018.

ELMASRI, Ramez; NAVATHE, Shamkant B. **Sistemas de banco de dados:** fundamentos e aplicações. 4. ed. Rio de Janeiro: LTC, 2006.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. **Desenvolvimento de sistemas de apoio à decisão e de métodos de coleta, análise de dados e monitoramento da pecuária na região Sul do Brasil**. Brasília, [2015?]. Disponível em: https://www.embrapa.br/busca-de-projetos/-/projeto/210797/desenvolvimento-de-sistemas-de-apoio-a-decisao-e-de-metodos-de-coleta-analise-de-dados-e-monitoramento-da-pecuaria-na-regiao-sul-do-brasil. Acesso em: 09 set. 2018.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. **Gestão de custos é a principal preocupação do pecuarista brasileiro**. Brasília, 2018. Disponível em: https://www.embrapa.br/busca-de-noticias/-/noticia/36645433/gestao-de-custos-e-a-principal-preocupacao-do-pecuarista-brasileiro. Acesso em: 12 set. 2018.

_____. **Embrapa pecuária sul: apresentação**. Brasília, [2018?]. Disponível em: https://www.embrapa.br/pecuaria-sul/apresentacao. Acesso em: 18 jul. 2018.

_____. **Quem somos**. Brasília, [s.d]. Disponível em: https://www.embrapa.br/quem-somos. Acesso em: 15 ago. 2018.

FERREIRA, Rafael dos Santos; CAMARGO, Sandro da Silva. Construindo um data warehouse para o agronegócio. *In:* CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 2013, Cuiabá. **Anais eletrônicos...** Cuiabá: Embrapa Informática Agropecuária, 2013. Disponível em: http://200.129.241.80/sbiagro/anais/artigos-completos. Acesso em: 02 fev. 2018.

GAJENDRAGADKAR, Madhura et al. Analysis of business intelligence tools. **International Education and Research Journal**, [*S.l.*], v. 2, n. 12, p. 45-46, 2016. Disponível em: http://ierj.in/journal/index.php/ierj/article/view/589/559. Acesso em: 03 abr. 2018.

GINER, Josep Lluís Cano. **Business Intelligence: competir con información**. Barcelona: Banesto Fundación Cultural, Banespyme, Esade, 2007.

GOLFARELLI, Matteo. Open source BI platforms: a functional and architectural comparison. *In:* INTERNATIONAL CONFERENCE ON DATA WAREHOUSING AND KNOWLEDGE DISCOVERY, 2009, Linz. **Anais eletrônicos...** Linz: Springer, 2009, p. 287-297. Disponível em: https://dblp1.uni-trier.de/db/conf/dawak/dawak2009.html. Acesso em: 05 fev. 2018.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data mining: concepts and techniques**. 3. ed. Waltham: Elsevier/Morgan Kaufmann Publishers, 2011.

HITACHI VANTARA. **Pentaho documentation: components reference**. 2018. Disponível em: https://help.pentaho.com/Documentation/8.1/Setup/Components_Reference. Acesso em: 18 jul. 2018.

HOFER, Elza et al. A relevância do controle contábil para o desenvolvimento do agronegócio em pequenas e médias propriedades rurais. **Revista Contabilidade e Controladoria**, Curitiba, v. 3, n. 1, p. 27-42, 2011. Disponível em: http://dx.doi.org/10.5380/rcc.v3i1.21490. Acesso em: 05 fev. 2018.

HUNG, Shin-Yuan et al. Regret avoidance as a measure of DSS success: An exploratory study. **Decision Support Systems**, [*S.l.],* v. 42, n. 4, p. 2093-2106, 2007. Disponível em: https://doi.org/10.1016/j.dss.2006.05.006. Acesso em: 28 mar. 2018.

INSTITUTO BRASILEIRO DE PESQUISA E ESTATÍSTICA. **Pesquisa sobre o uso das tecnologias da informação e comunicação nas empresas - 2010**. Rio de Janeiro: IBGE, 2012. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv62955.pdf. Acesso em: 21 abr. 2018.

_____. Censo agropecuário 2017: resultados preliminares. **Censo agropecuário**, v. 7, p. 1-108, 2017. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/periodicos/3093/agro_2017_resultados_preliminares.pdf. Acesso em: 24 ago. 2018.

_____. **Produto interno bruto dos municípios**. [2016?]. Disponível em: https://sidra.ibge.gov.br/pesquisa/pib-munic/tabelas. Acesso em: 22 out. 2018.

INMON, Willian Harvey. **Building the data warehouse**. 3. ed. New York: JohnWiley & Sons, 2002.

JUNIOR, Edgar Macari. **Análise de suítes de ferramentas integradas para a construção de data warehouses espaciais**. 2010. Trabalho de Conclusão de Curso (Bacharel em Ciências da Computação) — Departamento de Informática e Estatística, Universidade Federal de Santa Catarina, Florianópolis. Disponível em: https://repositorio.ufsc.br/bitstream/handle/123456789/184088/EdgarMacariJunior-Projetos_FINAL.pdf?sequence=-1&isAllowed=y. Acesso em: 25 abr. 2018.

JUNIOR, Methanias Colaço. **Projetando sistemas de apoio à decisão baseados em data warehouse**. 1. ed. Rio de Janeiro: Axcel Books, 2004.

KIMBALL, Ralph; ROSS, Margy. **The data warehouse toolkit: the definitive guide to dimensional modeling**. 3. ed. Indianapolis: John Wiley & Sons, Inc., 2013.

KERR, Don. Factors influencing the development and adoption of knowledge based decision support systems for small, owner-operated rural business. **Artificial Intelligence Review**, [s. l.], v. 22, p. 127-147, 2004. Disponível em: https://doi.org/10.1023/B:AIRE.0000045503.74951.7a. Acesso em: 10 mar. 2018.

KNOWAGE. **Community edition: open source for innovation**. 2018. Disponível em: https://www.knowage-suite.com/site/licensing/community-edition/. Acesso em: 10 abr. 2018.

LAMPERT, Vinicius do Nascimento. **Produtividade e eficiência de sistemas de ciclo completo na produção de bovinos de corte**. 2010. Tese (Doutorado em Zootecnia) — Faculdade de Agronomia, Universidade Federal do Rio Grande do Sul, Porto Alegre. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/28644/000770640.pdf?sequence=1&isAllowed=y. Acesso em: 15 dez. 2017.

LAMPERT, Vinicius do Nascimento. Simulando a escolha da estratégia para o sistema de produção. *In:* JORNADA DO NÚCLEO DE ESTUDOS EM SISTEMAS DE PRODUÇÃO DE BOVINOS DE CORTE E CADEIA PRODUTIVA, 2014, Porto Alegre. **Anais eletrônicos**... Porto Alegre: NESPRO/UFRGS, 2014, p. 131-147. Disponível em: http://www.ufrgs.br/nespro/arquivos/anais_jornadas/anais_ix_jornada_nespro_2014.pdf. Acesso em: 28 jan. 2018.

LAMPERT, Vinicius do Nascimento et al. Matriz de indicadores de sustentabilidade para produção de bovinos de corte no Rio Grande do Sul. *In:* V SIMPÓSIO DA CIÊNCIA DO AGRONEGÓCIO, 2017, Porto Alegre. **Anais eletrônicos...** Porto Alegre: CEPAN/UFRGS, 2017, p. 243-248. Disponível em: https://www.ufrgs.br/cienagro/wp-content/uploads/2015/11/Anais-CIENAGRO_2017.pdf. Acesso em: 13 jun. 2018.

LAKATOS, Eva Maria; MARCONI, Marina de Andrade. **Fundamentos de metodologia científica**. 5. ed. São Paulo: Atlas, 2003.

LAUDON, Kenneth C.; LAUDON, Jane Price. **Sistemas de informação gerenciais:** administrando a empresa digital. 5. ed. São Paulo: Prentice Hall, 2004.

LINDBLOM, Jessica *et al.* Promoting sustainable intensification in precision agriculture: review of decision support systems development and strategies. **Precision Agriculture**, [*S.l.*], v. 18, p. 309-331, 2017. Disponível em: https://doi.org/10.1007/s11119-016-9491-4. Acesso em: 12 mar. 2018.

LUNARDI, Guilherme Lerch; DOLCI, Pietro Cunha; MAÇADA, Antônio Carlos Gastaud. Adoção de tecnologia de informação e seu impacto no desempenho organizacional: um estudo realizado com micro e pequenas empresas. **Revista de Administração**, São Paulo, v. 45, n. 1, p. 5-17, 2010. Disponível em: https://doi.org/10.1016/S0080-2107(16)30505-2. Acesso em: 15 fev. 2018.

MACHADO, João A. Dessimon; OLIVEIRA, Lessandra Medeiros de; SCHNORRENBERGER, Adalberto. Compreendendo a tomada de decisão do produtor rural. *In:* CONGRESSO DA SOCIEDADE BRASILEIRA DE ECONOMIA E SOCIOLOGIA RURAL, 2006, Fortaleza. **Anais eletrônicos**... Fortaleza: Sociedade Brasileira de Economia e Sociologia Rural, 2006. Disponível em: http://www.sober.org.br/palestra/5/316.pdf. Acesso em: 18 fev. 2018.

MARIADB. **MariaDB license - MariaDB server license**. 2019. Disponível em: https://mariadb.com/kb/en/library/mariadb-license/. Acesso em: 21 fev. 2019.

MARINHEIRO, António; BERNARDINO, Jorge. Analysis of open source business intelligence suites. *In:* IBERIAN CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGIES, 2013, Lisboa. **Anais...** Lisboa: IEEE, 2013, p. 1-7. Disponível em: https://ieeexplore.ieee.org/document/6615764. Acesso em: 10 abr. 2018.

MARQUES, Pedro Rocha. **Avaliação da competitividade dos sistemas de produção de bovinos de corte da fronteira oeste do Rio Grande do Sul**. 2011. Dissertação (Mestrado em Agronegócio) — Centro de Estudos e Pesquisas em Agronegócios, Universidade Federal do Rio Grande do Sul, Porto Alegre. Disponível em: https://lume.ufrgs.br/bitstream/handle/10183/28654/000771727.pdf?sequence=1&isAllowed=y. Acesso em: 15 dez. 2018.

MAYER, Carlise Eyng; WERLANG, Nathalia Berger. O processo de tomada de decisão em propriedades rurais de Itapiranga – SC**.** *In:* WORKSHOP DE PRÁTICAS TECNOLÓGICAS NO AGRONEGÓCIO E MOSTRA DE EMPREENDEDORISMO, 2016, Itapiranga. **Anais eletrônicos...** Itapiranga: FAI Faculdades, v. 1, 2016, p. 1-17. Disponível em: https://eventos.uceff.edu.br/eventosfai_dados/artigos/inovaagro2016/585.pdf. Acesso em: 18 fev. 2018.

MENDES, Cássia Isabel Costa; BUAINAIN, Antônio Márcio; FASIABEN, Maria do Carmo Ramos. Heterogeneidade da agricultura brasileira no acesso às tecnologias da informação. **Espacios**, Caracas, v. 35, n. 11, 2014. Disponível em: http://www.revistaespacios.com/a14v35n11/14351111.html. Acesso em: 20 fev. 2018.

MOREIRA, Rodrigo; MARTINHAGO, Adriana Zanella; DRUMMOND, Luis César Dias. Desenvolvimento de um sistema de gestão para apoio à tomada de decisão no agronegócio da região do Alto Paranaíba. *In:* ESCOLA REGIONAL DE BANCO DE DADOS, 2015, Caxias do Sul. **Anais eletrônicos…** Caxias do Sul: Universidade de Caxias do Sul, 2015. Disponível em: http://www.lbd.dcc.ufmg.br/colecoes/erbd/2015/003.pdf. Acesso em: 16 abr. 2018.

MOTA, Fernando Maia da *et al*. BovReveals: uma plataforma OLAP e data mining para tomada de decisão na pecuária de corte. *In:* CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 2017, Campinas. **Anais eletrônicos…** Campinas: Universidade Estadual de Campinas, out. 2017, p. 29-39. Disponível em: https://www.sbiagro.org.br/sbiagro2017/wp-content/uploads/2017/12/anais-sbiagro2017-1ed.pdf. Acesso em: 16 abr. 2018.

MYSQL. **Licensing information user manual**. 2019. Disponível em: https://downloads.mysql.com/docs/licenses/mysqld-8.0-gpl-en.pdf. Acesso em: 06 fev. 2019.

NIELSEN, Annette Cleveland. Data warehouse for assessing animal health, welfare, risk management and communication. **Acta Veterinaria Scandinavica**, *[S.l.],* v. 53, jun. 2011. Suplemento 1. Disponível em: https://doi.org/10.1186/1751-0147-53-S1-S3. Acesso em: 22 abr. 2018.

NILAKANTA, Sree; SCHEIBE, Kevin; RAI, Anil. Dimensional issues in agricultural data warehouse designs. **Computers and Electronics in Agriculture**, [*S.l.*], v. 60, n. 2, p. 263-278, 2008. Disponível em: https://doi.org/10.1016/j.compag.2007.09.009. Acesso em: 07 mar. 2018.

OLIVEIRA, Djalma de Pinho Rebouças. **Sistemas de informações gerenciais:** estratégias, táticas, operacionais. 9. ed. São Paulo: Atlas, 2004.

POSTGRESQL. **License**. 2019. Disponível em: https://www.postgresql.org/about/licence/. Acesso em: 21 fev. 2019.

RAI, Anil et al. Design and development of data mart for animal resources. **Computers and Electronics in Agriculture**, [*S.l.],* v. 64, n. 2, p. 111-119, 2008. Disponível em: https://doi.org/10.1016/j.compag.2008.04.009. Acesso em: 01 mar. 2018.

RODIC, Jasna; BARANOVIC, Mirta. Generating data quality rules and integration into ETL process. *In:* PROCEEDINGS OF THE ACM TWELFTH INTERNATIONAL WORKSHOP ON DATA WAREHOUSING AND OLAP, 2009, Hong Kong. **Anais eletrônicos…** Hong Kong: ACM, Inc., 2009, p. 65-72. Disponível em: https://doi.org/10.1145/1651291.1651303. Acesso em: 14 fev. 2018.

ROSSI, Vittorio et al. Addressing the implementation problem in agricultural decision support systems. **Computers and Electronics in Agriculture**, [*S.l.],* v. 100, p. 88-99, jan. 2014. Disponível em: https://doi.org/10.1016/j.compag.2013.10.011. Acesso em: 26 fev. 2018.

SAIKU ANALYTICS. **Meteorite bi: saiku analytics**. [2018?]. Disponível em: http://meteorite.bi/saiku. Acesso em: 18 jul. 2018.

SANTOS, Gilberto José dos; MARION, Jose Carlos; SEGATTI, Sonia. **Administração de custos na agropecuária**. 3. ed. São Paulo: Atlas, 2002.

SOLID IT. **DB-Engines Ranking**. 2018. Disponível em: https://db-engines.com/en/ranking. Acesso em: 15 abr. 2018.

SOMMERVILLE, Ian. **Software engineering**. 9. ed. São Paulo: Pearson, 2011.

SOURCEFORGE. **Hitachi Ventara | Pentaho**. 2018. Disponível em: https://sourceforge.net/projects/pentaho/files/. Acesso em: 24 jul. 2018.

_____. **Hitachi Ventara | Pentaho: download statistics**. 2018. Disponível em: https://sourceforge.net/projects/pentaho/files/stats/timeline. Acesso em: 22 set. 2018.

STAIR, Ralph M.; REYNOLDS, George. **Princípios de sistemas de informação:** uma abordagem gerencial. 9. ed. São Paulo: Cengage Learning, 2011.

TECH, Adriano Rogério Bruno et al. Um modelo de gestão baseado em conceitos de e-science e data warehouse para aplicação no agronegócio da pecuária. **Archivos de Zootecnia**, Córdoba, v. 59, n. 226, p. 161-168, 2010. Disponível em: http://www.uco.es/organiza/servicios/publica/az/php/img/web/05_09_48_01UmmodeloTechpruebaimagenesjpegGus.pdf. Acesso em: 27 abr. 2018.

TRAUB, Berthold *et al.* The data storage and analysis system of the Swiss National Forest Inventory. **Computers and Electronics in Agriculture**, [*S.l.*], v. 132, p. 97-107, jan. 2017. Disponível em: https://doi.org/10.1016/j.compag.2016.11.016. Acesso em: 06 abr. 2018.

TURBAN, Efrain; SHARDA, Ramesh E.; DELEN, Dursun. **Decision support and business intelligence systems**. 9. ed. New Jersey: Pearson Education, Inc., 2011.

VAN MEENSEL, Jef et al. Effect of a participatory approach on the successful development of agricultural decision support systems: The case of Pigs2win. **Decision Support Systems**, [*S.l.*], v. 54, n. 1, p. 164-172, 2012. Disponível em: https://doi.org/10.1016/j.dss.2012.05.002. Acesso em: 05 fev. 2018.

VENKATESH, Viswanath; BALA, Hillol. Technology acceptance model 3 and a research agenda on interventions. **Decision Sciences**, [*S.l.*], v. 39, n. 2, p. 273-315, 2008. Disponível em: https://doi.org/10.1111/j.1540-5915.2008.00192.x. Acesso em: 21 abr. 2018.

VERNIER, Françoise et al. EIS pesticides: an environmental information system to characterize agricultural activities and calculate agro-environmental indicators at embedded watershed scales. **Agricultural Systems**, [*S.l.*], v. 122, p. 11-21, nov. 2013. Disponível em: https://doi.org/10.1016/j.agsy.2013.07.005. Acesso em: 02 mar. 2018.

VIEIRA, Marcos Rodrigues et al. Banco de Dados NoSQL: conceitos, ferramentas, linguagens e estudos de casos no contexto de Big Data. *In*: SIMPÓSIO BRASILEIRO DE BANCOS DE DADOS, 2012, São Paulo. **Anais eletrônicos...** São Paulo: Sociedade Brasileira de Computação, 2012. Disponível em: http://data.ime.usp.br/sbbd2012/artigos/pdfs/sbbd_min_01.pdf. Acesso em: 06 fev. 2018.

WIEDMANN, Thomas; MINX, Jan. **A definition of "Carbon Footprint"**: Relatório Técnico. Durham, 2007.

WIJAYA, Rahmadi; PUDJOATMODJO, Bambang. An overview and implementation of extraction-transformation-loading (ETL) process in data warehouse (Case study: Department of agriculture). *In:* INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGY*,* 2015, Nusa Dua, Bali, Indonésia. **Anais eletrônicos...** [*S.l.*]: IEEE, 2015, p. 70-74. Disponível em: https://doi.org/10.1109/ICoICT.2015.7231399. Acesso em: 06 mar. 2018.

WU, Liya; BARASH, Gilad; BARTOLINI, Claudio. A service-oriented architecture for business intelligence. *In:* IEEE INTERNATIONAL CONFERENCE ON SERVICE-ORIENTED COMPUTING AND APPLICATIONS, 2007, Newport Beach, CA, USA. **Anais eletrônicos...** [*S.l.]:* IEEE, 2007, p. 279-285. Disponível em: https://ieeexplore.ieee.org/document/4273437/. Acesso em: 08 mar. 2018.

ZAZA, Claudio *et al*. A new decision-support system for the historical analysis of integrated pest management activities on olive crops based on climatic data. **Computers and Electronics in Agriculture**, [*S.l.*], v. 148, p. 237-249, 2018. Disponível em: https://doi.org/10.1016/j.compag.2018.03.015. Acesso em: 06 abr. 2018.

APÊNDICE A – Documento de Requisitos do Sistema

Documento de Requisitos do Sistema *De Business Intelligence* para Produtores Rurais v. 1.2

Bagé, 09 de novembro de 2018.

Prefácio

Este documento foi elaborado para os clientes (especialistas da Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA), desenvolvedores de software, profissionais de Tecnologia da Informação, Engenheiros e Cientistas de Computação e demais profissionais da tecnologia. Com este documento, espera-se dar subsídios para o desenvolvimento da proposta de integração de dados de diferentes Sistemas de Informação voltados à pecuária de corte, com o objetivo de subsidiar as decisões de pecuaristas brasileiros através das tecnologias de *Data Warehouse* e *Business Intelligence*.

Histórico de Alterações

Data	Versão	Descrição	Autor
13/09/2018	1.0	Criação do documento de requisitos, descrição das fontes de dados (*myBeef*, Gestão de Custos e *livestock sustainability*), inclusão dos modelos conceituais (*myBeef* e Gestão de Custos), identificação dos requisitos, descrição dos requisitos funcionais e não-funcionais.	Luciano Moraes da Luz Brum
26/09/2018	1.1	Alteração do nome do sistema de Emissão para *Livestock Sustainability*. Mudança e ajustes nos modelos de banco de dados dos sistemas e acréscimo do modelo do *Livestock Sustainability*.	Luciano Moraes da Luz Brum
08/11/2018	1.2	Remoção do sistema *MyBeef* da integração. Inserção do *ETL* nos requisitos não-funcionais. Atualização dos modelos conceituais dos sistemas fontes. Inserção de uma visão geral da arquitetura da solução.	Luciano Moraes da Luz Brum

1. Introdução

Este documento especifica os requisitos do Sistema de *Business Intelligence* (BI) para produtores rurais, fornecendo aos desenvolvedores as informações necessárias para o projeto, implementação, realização de testes e validação do sistema.

1.1 Visão geral do documento

Além desta seção introdutória, as seções seguintes estão organizadas como descrito abaixo.

Seção 2 – Descrição geral do sistema: apresenta uma visão geral do sistema de BI (e das fontes de dados), caracterizando qual é o seu escopo e descrevendo seus usuários.

Seção 3 – Requisitos funcionais: especifica as funcionalidades do sistema, descrevendo os fluxos de eventos, prioridades, atores, entradas e saídas de cada caso de uso a ser implementado.

Seção 4 – Requisitos não-funcionais: especifica todos os requisitos não funcionais do sistema, divididos em requisitos de usabilidade, confiabilidade, desempenho, segurança e requisitos de hardware e software.

1.2 Convenções, termos e abreviações

A correta interpretação deste documento exige o conhecimento de algumas convenções e termos específicos, que são descritos a seguir.

1.2.1 Identificação dos requisitos

Os requisitos devem ser identificados com um identificador único. A numeração inicia com o identificador [RF001] ou [NF001] e prossegue sendo incrementada à medida que forem surgindo novos requisitos.

1.2.2 Prioridades dos requisitos

Para estabelecer a prioridade dos requisitos, nas seções 4 e 5, foram adotadas as denominações “essencial”, “importante” e “desejável”.

Essencial é o requisito sem o qual o sistema não entra em funcionamento. Requisitos essenciais são requisitos imprescindíveis, que têm que ser implementados impreterivelmente.

Importante é o requisito sem o qual o sistema entra em funcionamento, mas de forma não satisfatória. Requisitos importantes devem ser implementados, mas, se não forem, o sistema poderá ser implantado e usado mesmo assim.

Desejável é o requisito que não compromete as funcionalidades básicas do sistema, isto é, o sistema pode funcionar de forma satisfatória sem ele. Requisitos desejáveis podem ser deixados para versões posteriores do sistema, caso não haja tempo hábil para implementá-los na versão que está sendo especificada.

2. Descrição geral do sistema

2.1 Problema Existente

Atualmente, estão sendo desenvolvidos diferentes sistemas com o propósito de auxiliar profissionais do campo e consultores nos seus processos decisórios. Cada um dos sistemas possui um propósito específico e auxilia na solução de um conjunto específico de problemas enfrentados pelos produtores rurais e consultores.

Em um futuro próximo, quando os sistemas estiverem disponíveis para o público geral, percebe-se a necessidade de realizar a integração dos dados destes sistemas para a realização de análises, tanto por profissionais do campo e consultores como pelos pesquisadores responsáveis pelos sistemas.

Viabilizar e facilitar o acesso e a análise das variáveis dos sistemas de forma integrada torna-se um desafio, pois os sistemas estão sendo desenvolvidos por equipes distintas de desenvolvedores, podem possuir diferentes significados para variáveis em comum e diferentes Sistemas Gerenciadores de Banco de Dados (SGBD). Ainda, visando entregar informações úteis e de qualidade para os clientes, se faz necessário um processo eficiente de integração de dados destes sistemas, visando garantir não só o acesso às informações, mas também a qualidade dos dados disponíveis para análise. Os sistemas Ferramenta de Gestão de Custos (FGC) e *Livestock Sustainability* (LS) estão em um estágio de desenvolvimento mais avançado, próximos a etapa de testes de usabilidade com os usuários, portanto sendo estes os alvos iniciais do processo de integração dos dados.

Portanto, o problema existente é **a não disponibilidade de uma ferramenta capaz de integrar os dados e viabilizar o acesso e análise dos dados dos sistemas FGC e LS em um único ambiente, através de consultas *ad-hoc*, com uma interface simples que permita a geração flexível de gráficos, relatórios e *dashboards***.

2.2 Descrição dos Sistemas Fontes e do Sistema de BI

O principal propósito do Sistema de *Business Intelligence* para Produtores Rurais é auxiliar os seus clientes no acesso e análise das informações disponíveis nos repositórios dos sistemas já existentes, visando subsidiar os seus processos decisórios relacionados à pecuária de corte. É um BI que tem por objetivo permitir análises comparativas entre estabelecimentos rurais através das métricas definidas nos sistemas fontes dos dados (FGC e LS).

Abaixo, uma breve descrição dos sistemas fontes dos dados da solução de BI:

Ferramenta de Gestão de Custos: Esta solução está no escopo do projeto *MyBeef*, projeto da EMBRAPA Pecuária Sul que visa proporcionar aos pesquisadores da EMBRAPA um maior conhecimento sobre a cadeia produtiva da pecuária e dos territórios da região Sul do Brasil. Esta é a versão atual do projeto, futuramente serão incorporadas e integradas outras funções, conforme mencionado no site do projeto da EMBRAPA:

> [...] espera-se produzir boletins, periódicos com informações sobre a cadeia produtiva, website com um banco de dados com softwares nacionais sobre a pecuária de corte e leite, matriz de indicadores de sustentabilidade para a pecuária, software para análise econômica e ambiental, método para análise espaço-temporal de indicadores sociais, econômicos, ambientais e produtivos e uma plataforma para abrigar sistemas de apoio à decisão para

> a pecuária, tendo como público-alvo prioritário o produtor rural. (EMBRAPA, 2015?).

Este sistema web *online* visa auxiliar o produtor rural a compreender melhor os seus custos dentro da atividade pecuária. O sistema está em desenvolvimento por fases. Cada fase é caracterizada pela inserção de um conjunto específico de informações sobre os custos da propriedade. Conforme o produtor conclui e avança nas fases, mais dados e informações sobre os custos de sua propriedade e produção serão requisitados. Com isso, o retorno de informações, gráficos, relatórios e *dashboards* sobre os custos serão cada vez mais enriquecidos e detalhados com essas mesmas informações de forma organizada. Até o momento, foi concluída a fase 1 do sistema. A Figura 1 apresenta o modelo conceitual do sistema:

Figura 1: Modelo conceitual do FGC.

Fonte: Autor (2019).

Livestock Sustainability: Este sistema é um aplicativo que visa fornecer ao produtor uma forma de estimar a emissão da pegada de carbono do sistema de produção, utilizando variáveis que sejam conhecidas por ele. A partir do sistema de informação, será coletado um conjunto de dados que permitirá realizar mineração de dados, de forma a tentar identificar quais variáveis mais influenciam na emissão e como será possível reduzir a emissão. Ou seja, é um sistema de informação que estima a pegada de carbono e sugere alternativas de redução.

O sistema também recebe de entrada indicadores zootécnicos e informações relacionadas relevantes da pecuária de corte, para viabilizar seus impactos em indicadores econômicos e produtivos da propriedade. Estas informações são as mesmas contidas no sistema *MyBeef*. Com o uso de algumas informações da produção e propriedade (taxa de natalidade (1) e mortalidade (2) do rebanho, taxa de desmame (3), idade de acasalamento e abate do rebanho e área disponível em hectares) combinada com informações consideradas padrão para a região, é possível verificar a produtividade (4) desta propriedade.

$$Taxa\ de\ natalidade\ =\ ((n°\ de\ terneiros\ nascidos\ )/(n°\ de\ fêmeas\ acasaladas))\ *\ 100 \quad (1)$$

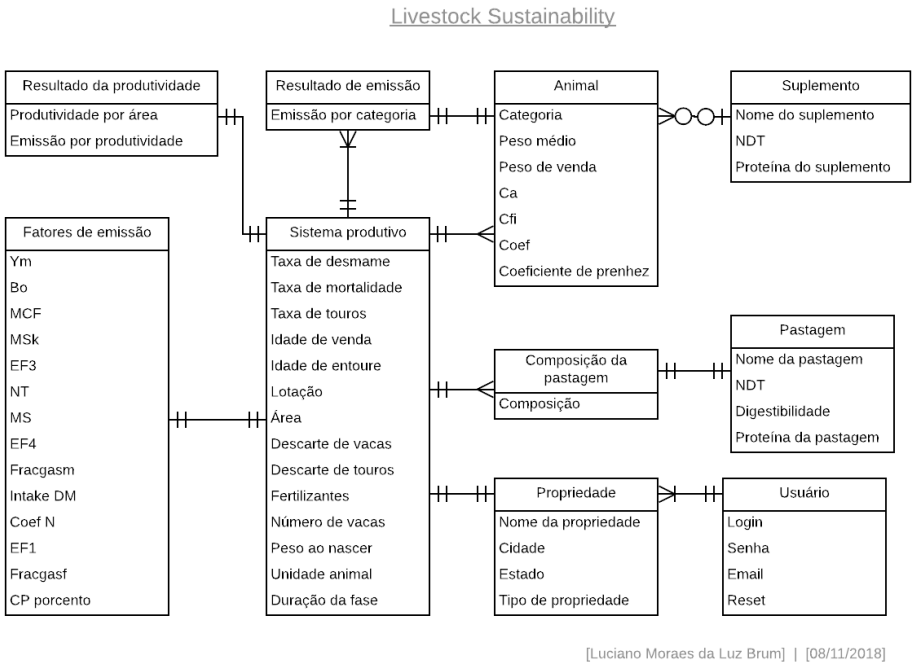
$$Taxa\ de\ mortalidade\ =\ ((n°\ de\ animais\ que\ morreram)/(n°\ total\ de\ animais))\ *\ 100 \quad (2)$$

$$Taxa\ de\ desmame\ =\ \left(\frac{n°\ de\ terneiros\ desmamados}{n°\ de\ fêmeas\ acasaladas}\right) \quad (3)$$

$$Produtividade\ =\ ((quantidade\ de\ quilos\ produzidos)/área) \quad (4)$$

A Figura 2 apresenta o modelo conceitual do sistema LS:

Figura 2: Modelo conceitual do LS.



Fonte: Autor (2019).

Sobre as variáveis:

- O sistema permite que o produtor coloque até quatro pastagens e para cada pastagem ele deve dizer a composição, ou seja, o quanto ele possui dessa pastagem.

- As variáveis de entrada do usuário, peso médio, peso de venda e suplemento, devem ser inseridas uma vez para cada categoria animal, que são dez (vacas, novilhas de 3 anos, novilhas de 2 ano, novilhas de 1 ano, novilhos de 3 anos, novilhos de 2 anos, novilhos de 1 ano, bezerras, bezerros e touros).

- As variáveis de entrada fixa *Cfi*, *Coef* e *Cpregnancy* devem ser inseridas uma vez para cada categoria animal.

- As demais variáveis de entrada do usuário e variáveis de entrada fixa são informadas uma única vez.

A solução de BI proposta engloba: o desenvolvimento/adaptação de suítes e tecnologias de *Business Intelligence* existentes (*Pentaho/SpagoBI/Knowage*), criação do processo de Extração, Transformação e Carga dos dados (*Extract, Transform and Load - ETL*) das fontes de dados (FGC - SGBD *PostgreSQL* e LS - *MySQL*), com o objetivo final de auxiliar o público-alvo (produtores rurais e consultores) nas suas

tomadas de decisão, através do acesso simples e eficiente às informações produtivas, econômicas e ambientais com garantias da sua qualidade.

2.3 Funções do Sistema

O sistema deve ter funções que permitam o acesso às informações dos sistemas FGC e LS de forma expressiva e eficiente, sem comprometer a usabilidade e simplicidade.

O sistema deve possuir uma tela inicial para *login* através da utilização de *e-mail* e senha. A conta deve ter sido previamente criada e cadastrada nos sistemas anteriormente especificados. O sistema de BI deve fornecer acesso integrado às informações dos sistemas, com distinção entre os acessos ao nível de propriedade em concordância com a identificação do usuário. A ferramenta deve permitir a seleção flexível das variáveis de interesse (métricas) e seus descritores (dimensões), geração de diferentes tipos de gráficos (linha, barras horizontais/verticais, pizza), visualização das informações através de tabelas e geração de relatórios para o acesso *off-line* às informações.

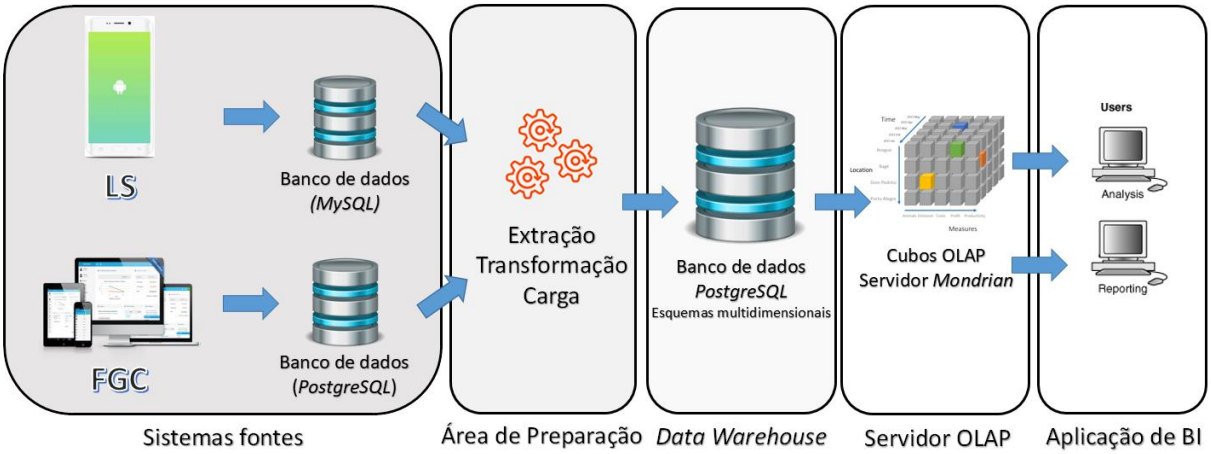
Os principais usuários finais deste sistema são produtores rurais e especialistas. Um cuidado adicional deve ser tomado em relação aos aspectos de usabilidade e facilidade de uso do sistema. Produtores rurais podem não possuir o tempo e conhecimento necessários para o aprendizado na utilização do sistema, portanto, este deve ser simplificado ao máximo, sem comprometer suas funções e funcionalidades, ao mesmo tempo que deve fornecer informações que sejam úteis para as suas decisões no ramo pecuário.

Cabe ressaltar que, conforme são entregues novas versões do sistema, surge a necessidade de adaptação ou criação de novas funcionalidades e requisitos, sejam funcionais ou não-funcionais. Os requisitos são mutáveis ao longo do tempo, conforme o entendimento das funções do sistema fica mais claro para desenvolvedores e clientes. A ideia deste documento, portanto, é também registrar essas mudanças nas necessidades e requisitos.

2.4 Visão Geral da Arquitetura da Solução Proposta

A Figura 3 apresenta a arquitetura geral da solução proposta e é detalhada logo abaixo.

Figura 3: Visão geral da Arquitetura da solução proposta.



Fonte: Autor (2019).

Sistemas fontes: são os sistemas de informação atualmente em desenvolvimento e em fase de testes. A interação dos produtores rurais e especialistas com estes sistemas (via aplicativo para *android* no LS e navegador web *online* no FGC), deverá gerar um volume de dados que será armazenado em bancos de dados relacionais.

Stage Area (Área de preparação dos dados): local onde serão armazenados os dados extraídos dos sistemas fontes. É onde serão tratadas todas as questões relacionadas à qualidade dos dados (precisão, completude, consistência) e às regras de negócio do cliente. É onde ocorre o ETL.

Data Warehouse: local onde serão armazenados os dados tratados na *Stage Area*. Os dados neste ambiente estarão armazenados em um banco de dados relacional, porém, organizados em esquemas multidimensionais.

Servidor OLAP: *Mondrian* é o servidor que vai processar todas as consultas aos dados do DW.

Aplicação de BI: é a camada onde ocorre o acesso aos dados do DW pelo cliente, através de aplicações específicas de BI.

3. Requisitos funcionais

[RF001] Acesso ao sistema

Descrição do caso de uso: O sistema deve permitir que o usuário faça o *login* no sistema.

Prioridade: ■ Essencial ☐ Importante ☐ Desejável

Entradas e pré-condições: Para realização do *login* são necessários o ***e-mail*** e a **senha** do usuário. O usuário deve possuir uma conta previamente cadastrada nos sistemas **FGC ou LS** para a utilização deste sistema.

Saídas e pós-condição: O usuário terá acesso às funcionalidades no sistema relativas ao seu nível de acesso.

[RF002] Acesso às informações do *Data Warehouse*

Descrição do caso de uso: O sistema deve permitir que o usuário possa selecionar as variáveis de interesse para posterior uso. As variáveis de seleção devem ser as que estão disponíveis no *Data Warehouse* (DW), de acordo com o nível de acesso do usuário. Através da seleção das variáveis e seus descritores, será possível utilizar as funcionalidades de visualização destas informações.

Prioridade: ■ Essencial ☐ Importante ☐ Desejável

Entradas e pré-condições: Para o acesso às variáveis do sistema, é necessário que o usuário esteja logado no sistema, através do uso do seu ***e-mail*** e **senha**.

Saídas e pós-condição: O usuário poderá utilizar as informações contidas no DW de acordo com o seu interesse.

[RF003] Visualização das informações do *Data Warehouse*

Descrição do caso de uso: O sistema deve permitir que o usuário possa visualizar os dados de interesse. Os dados selecionados devem estar associados às variáveis disponíveis no *Data Warehouse* (DW), de acordo com o nível de acesso do usuário. Deve ser possível que o usuário possa visualizar as informações em um formato de tabela, de forma que seja possível visualizar separadamente os dados por linhas e colunas, sendo cada uma destas relacionadas com as métricas e seus descritores.

Prioridade: ■ Essencial ☐ Importante ☐ Desejável

Entradas e pré-condições: Para o acesso às variáveis do sistema, é necessário que o usuário esteja logado no sistema, através do uso do seu ***e-mail*** e **senha**.

Saídas e pós-condição: O usuário poderá utilizar as informações contidas no DW de acordo com o seu interesse.

[RF004] Geração de gráficos

Descrição do caso de uso: O sistema deve permitir que o usuário, após a seleção das variáveis de interesse, possa optar pela geração de diferentes tipos de gráficos para a análise detalhada das informações. O sistema deve permitir a geração de gráficos simples, expressivos e efetivos que permitam o entendimento da relação entre os valores do DW. Deve ser tomado um cuidado adicional no quesito usabilidade, pois o público-alvo (produtores rurais) não é habituado com a utilização de sistemas de informação para este fim. Sugere-se, pelo menos, os gráficos de barras, linha e de pizza.

Prioridade: ■ Essencial ☐ Importante ☐ Desejável

Entradas e pré-condições: Para a geração dos gráficos, é necessário que o usuário realize o *login*, através do seu ***e-mail*** e a **senha e** selecione as informações de interesse para análises (métrica numérica (fato) e descritores (dimensões)). Um conhecimento mínimo necessário sobre as informações (métricas e descritores das informações) é necessário para a sua efetiva utilização

Saídas e pós-condição: O usuário poderá visualizar as informações selecionadas do DW de acordo com o seu interesse e necessidade.

[RF005] Geração de relatórios

Descrição do caso de uso: O sistema deve permitir que o usuário, após a seleção das variáveis de interesse, possa optar pela geração de relatórios, por exemplo em .pdf (*portable document format*), para a análise detalhada das informações. O sistema deve permitir a geração de relatórios que possam incluir, por exemplo, os gráficos gerados no **[RF004]**.

Prioridade: ■ Essencial ☐ Importante ☐ Desejável

Entradas e pré-condições: Para a geração dos relatórios, é necessário que o usuário realize o *login*, através do seu ***e-mail*** e a **senha e** selecione as informações de interesse para análises (métrica numérica (fato) e descritores (dimensões)).

Saídas e pós-condição: O usuário poderá acessar as informações detalhadas do DW através de um relatório, por exemplo em formato .pdf, sem que seja necessária a conexão com a internet para a leitura após o *download*.

4. Requisitos não-funcionais

[NF001] Usabilidade

A interface com o usuário é de vital importância para o sucesso do sistema, principalmente porque será utilizado por produtores e consultores rurais. Sistemas voltados para o suporte de processos decisórios nos setores da agricultura, pecuária e agronegócio tem uma alta taxa de desistência dos usuários. Ainda, nem todos os usuários possuem tempo disponível para aprender sobre como utilizar o sistema, portanto este deve ser simples e intuitivo. O sistema deve possuir uma interface amigável, de forma que não se torne cansativa aos usuários mais experientes.

Portanto, define-se o seguinte indicador: Os usuários devem estar hábeis a utilizar todas as funções do sistema após 2 horas de treinamento ou após a leitura completa do guia do usuário.

Prioridade: ■ Essencial ☐ Importante ☐ Desejável

[NF002] Desempenho

Este é um aspecto que deve ser considerado por corresponder a um fator de qualidade de software. Especificamente neste sistema, o resultado das consultas aos dados do *Data Warehouse* deve ser rápido de tal forma que não prejudique a experiência de uso na geração de gráficos e relatórios.

Até o momento, não há previsão do número de acessos de usuários simultâneos. Portanto, não será definida uma restrição neste quesito.

O mesmo em relação à utilização de recursos físicos (memória, disco, processamento, ...), pois os sistemas fontes ainda não foram disponibilizados para o público em geral.

Quanto ao tempo de resposta de uma transação ou conjunto de transações, este deve ser priorizado, pois as informações devem ser fornecidas em tempo próximo ao real para os usuários. Um usuário não deve esperar mais do que 10 segundos para que o relatório ou gráfico seja gerado e disponibilizado.

Prioridade: ☐ Essencial ■ Importante ☐ Desejável

[NF003] Disponibilidade

Descrição do caso de uso: O sistema deve estar disponível para o acesso dos usuários na maior parte do tempo. Na ocorrência de erros ou falhas, a manutenção deve ser precisa e rápida, para que não prejudique a experiência dos usuários com a solução. Para mensurar este requisito, sugere-se o cálculo do TMF (Tempo Médio

entre Falhas), do TMR (Tempo Médio para Reparo) e da própria disponibilidade geral do sistema. A seguir, as fórmulas para o cálculo destes indicadores.
TMF = (Tempo total disponível – Tempo perdido) / (Número de paradas)
TMR = (Tempo total de reparo) / (quantidade de falhas)
Disponibilidade = TMF/(TMF+TMR)
Como as versões 1.x do sistema são uma proposta, este requisito não é prioridade no momento.

Prioridade: ☐ Essencial ☐ Importante ■ Desejável

[NF004] Segurança

Descrição do caso de uso: O sistema não deve permitir o acesso de pessoas não autorizadas às informações dos produtores rurais. Usuários sem uma conta previamente cadastrada no FGC ou LS não podem utilizar o sistema. Não deve ser permitido modificar ou atualizar o conteúdo do DW sob nenhuma forma. O uso de criptografia nas senhas no processo de autenticação é desejado.

Prioridade: ■ Essencial ☐ Importante ☐ Desejável

[NF005] Hardware e Software

O SGBD utilizado como base para o *Data Warehouse* na versão 1.0 será o *PostgreSQL*. Este SGBD é gratuito, confiável e de código aberto, com uma ampla comunidade de usuários e possuí disponível sua documentação na internet. As limitações deste SGBD não serão um empecilho para a implementação da primeira versão do sistema.
A máquina servidora do banco de dados no momento para a execução dos testes é um notebook *Dell* de processador *Intel Core i7-7500U*, frequência de 2.70 *Ghz* – 2.90 *Ghz*, memória RAM (*Random Access Memory*) de 16 *Gygabytes,* sistema operacional *Windows 10* de 64 bits.

Prioridade: ■ Essencial ☐ Importante ☐ Desejável

[NF006] Sistema *ETL*

Foi definido que, com a periodicidade mensal, será realizado o processo de extração, transformação e carga dos dados dos dois sistemas fontes para o DW. Este processo será elaborado e testado no mesmo ambiente de software/hardware apresentado em [**NF005]**. Será utilizado um banco de dados intermediário (diferente do DW) para realizar as transformações nos dados, visando garantir sua qualidade, confiabilidade e consistência.

Prioridade: ■ Essencial ☐ Importante ☐ Desejável

5. Gerenciamento de Requisitos

Para alteração ou inclusão de requisitos no projeto da solução deverão ser seguidas as etapas descritas a seguir:

- O cliente solicita uma mudança de requisito ao projetista;
- O projetista recebe a mudança sugerida;
- O projetista juntamente com o orientador do projeto analisarão tal mudança e avaliarão o impacto da mesma no sistema;
- O projetista juntamente com o cliente negociam a mudança pretendida;
- Como resultado dessa negociação ocorrerá ou não a mudança solicitada.

6. Referências

EMBRAPA. Desenvolvimento de sistemas de apoio à decisão e de métodos de coleta, análise de dados e monitoramento da pecuária na região Sul do Brasil. Brasília, [2015?]. Disponível em: <https://www.embrapa.br/busca-de-projetos/-/projeto/210797/desenvolvimento-de-sistemas-de-apoio-a-decisao-e-de-metodos-de-coleta-analise-de-dados-e-monitoramento-da-pecuaria-na-regiao-sul-do-brasil>. Acesso em: set. 2018.

Assinaturas

Cliente/Orientador Bagé, 08 de novembro de 2018

_____

Vinícius do Nascimento Lampert
Empresa Brasileira de Pesquisa Agropecuária – Embrapa Pecuária Sul

_____

Luciano Moraes da Luz Brum
Programa de Pós-Graduação em Computação Aplicada - Unipampa

APÊNDICE B – Documentação do sistema

Documentação do Sistema *De Business Intelligence* para Produtores Rurais v. 1.1 (para desenvolvedores)

Bagé, 25 de fevereiro de 2019.

Prefácio

Este documento foi elaborado para os desenvolvedores de software, profissionais de Tecnologia da Informação, Engenheiros e Cientistas de Computação e demais profissionais da tecnologia. Espera-se dar subsídios para a continuação do desenvolvimento da proposta de integração de dados de diferentes Sistemas de Informação voltados à pecuária de corte, com o objetivo de subsidiar as decisões de pecuaristas brasileiros através das tecnologias de *Data Warehouse* e *Business Intelligence*.

Histórico de Alterações

Data	Versão	Descrição	Autor
19/01/2019	1.0	Criação do documentação do sistema. Inclusão da descrição técnica dos sistemas fontes de dados, arquitetura de ETL e DW.	Luciano Moraes da Luz Brum
25/02/2019	1.1	Alteração do modelo do DW para *star schema*, em vez de *galaxy schema*. Alteração do ETL e outros para adaptar mudanças	Luciano Moraes da Luz Brum

Lista de Figuras

Lista de Tabelas

Sumário

1. Introdução

Este documento especifica os detalhes técnicos do Sistema de *Business Intelligence* para produtores rurais, fornecendo aos desenvolvedores as informações necessárias para a continuação do projeto, desenvolvimento, implementação, realização de testes e validação do sistema.

1.1 Visão geral do documento

Além desta seção introdutória, as seções seguintes estão organizadas como descrito abaixo.

Seção 2 - Descrição geral do sistema: apresenta uma visão geral do sistema de *Business Intelligence* e das fontes de dados, caracterizando o seu escopo e apresentando informações técnicas essenciais para o entendimento do projeto de integração.
Seção 3 - *Data Warehouse*: especifica o modelo dimensional implementado e demais características e especificidades do projeto.
Seção 4 - Processo ETL (*Extract, Transform and Load*): especifica a arquitetura de ETL implementada, características e detalhes técnicos.
Seção 5 - OLAP (*Online Analytical Processing*): especifica os arquivos relacionados aos cubos de dados gerados com o acesso ao DW.
Seção 6 - *Pentaho BA Server* e *Saiku Analytics*: apresenta as características e detalhes técnicos da interface da aplicação que fornece acesso aos cubos de dados.

1.2 Convenções, termos e abreviações

A correta interpretação deste documento exige o conhecimento de algumas convenções e termos específicos muito utilizados em soluções de *Business Intelligence*, que são descritos a seguir.

BA – *Business Analytics*
Definição: conjunto de métodos e tecnologias que permitem realizar diferentes tipos de análises em conjuntos de dados.

BI – *Business Intelligence*
Definição: conjunto de métodos e tecnologias que permitem extrair, integrar, organizar e gerenciar dados e informações de diferentes fontes de forma estratégica, possibilitando o acesso flexível destes dados por usuários autorizados, para fornecer um melhor suporte aos seus processos decisórios.

DW – *Data Warehouse*
Definição: Um *Data Warehouse* é um repositório de dados que proporciona uma visão global, comum e integrada dos dados de uma organização – independentemente de como sejam utilizados posteriormente pelos consumidores ou usuários, com as seguintes propriedades: estável, coerente, confiável e com informação histórica (DÍAZ, 2012, p. 32, tradução nossa).

ETL - *Extract, Transform and Load*
Definição: o processo de ETL é dividido em três etapas: extração, transformação e carga dos dados. O primeiro passo do ETL é a extração dos dados necessários das diferentes fontes de informação para posterior manipulação. Estes dados são

armazenados em uma base de dados temporária, chamada na literatura por *stage area* ou área de preparação dos dados. Posteriormente, poderão ser necessárias transformações nos dados, visando corrigir erros de digitação, conflitos de domínios, dados faltantes ou incompletos, conversão do formato do dado, combinação de dados de diferentes fontes, correção de dados duplicados, entre outras possibilidades. Essa é a etapa de transformação dos dados. A última etapa do processo ETL é a carga dos dados para os modelos dimensionais da área de apresentação. Nesta etapa é realizado o processamento das dimensões, como: atribuição de chaves substitutas (*surrogate keys*), fornecimento de descrições apropriadas para os atributos das dimensões, repartições ou combinação de colunas para apresentação dos valores, entre outros (KIMBALL; ROSS, 2013; TURBAN; SHARDA; DELEN, 2011, DÍAZ, 2012).

NK – Natural Key
Definição: É a chave que identifica unicamente um registro em uma tabela no sistema legado.

OLAP – ***Online Analytical Processing***
Definição: É uma forma de acesso aos dados que permite analisa-los de várias formas e em diferentes profundidades, tendo como características a visão multidimensional dos dados, vários modos de visualização, acesso aos dados com alta flexibilidade e performance.

SA – ***Stage Area***
Definição: Repositório de dados intermediário que, considerando a arquitetura de uma solução de BI, fica entre as fontes de dados operacionais e o DW. Tal repositório visa auxiliar no processo de ETL e na melhoria da qualidade dos dados, funciona como uma cache dos dados operacionais usado no processo de carga para o DW (quando a mesma for persistente, caso contrário é transiente) e, ainda, utilizada para acesso em detalhe de dados que possivelmente não estão no DW.

SK – ***Surrogate Key***
Definição: É a chave primária das tabelas que representam as dimensões. Estas chaves não possuem um significado próprio, geralmente são números inteiros e servem para estabelecer o relacionamento com as chaves estrangeiras presentes nas tabelas fato.

2. Descrição geral do sistema

2.1 Problema Existente

Estão sendo desenvolvidos diferentes sistemas com o propósito de auxiliar profissionais do campo e consultores nos seus processos decisórios. Cada um dos sistemas possui um propósito específico e auxilia na solução de um conjunto específico de problemas enfrentados pelos produtores rurais e consultores. Percebe-se a necessidade de realizar a integração dos dados destes sistemas para a realização de análises, tanto por profissionais do campo e consultores como pelos pesquisadores responsáveis pelos sistemas.

Viabilizar e facilitar o acesso e a análise das variáveis dos sistemas de forma integrada torna-se um desafio, pois os sistemas estão sendo desenvolvidos por equipes distintas de desenvolvedores, podem possuir diferentes significados para variáveis em comum e diferentes Sistemas Gerenciadores de Banco de Dados (SGBD). Ainda, visando entregar informações úteis e de qualidade para os clientes, se faz necessário um processo eficiente de integração de dados destes sistemas, visando garantir não só o acesso às informações, mas também a qualidade dos dados disponíveis para análise. Os sistemas Ferramenta de Gestão de Custos (FGC) e *Livestock Sustainability* (LS) estão em um estágio de desenvolvimento mais avançado, próximos a etapa de testes de usabilidade com os usuários, portanto sendo estes os alvos iniciais do processo de integração dos dados.

Em suma, o problema existente é a não disponibilidade de uma ferramenta capaz de integrar os dados e viabilizar o acesso e análise dos dados dos sistemas FGC e LS em um único ambiente, através de consultas *ad-hoc*, com uma interface simples que permita a geração flexível de gráficos, relatórios e *dashboards*.

2.2 Descrição dos Sistemas Fontes e do Sistema de BI

O propósito do Sistema de *Business Intelligence* para Produtores Rurais é auxiliar os seus clientes no acesso e análise das informações disponíveis nos repositórios dos sistemas já existentes, visando subsidiar os seus processos decisórios relacionados à pecuária de corte. É um BI que tem por objetivo permitir análises comparativas entre propriedades através das métricas definidas nos sistemas fontes dos dados (FGC e LS).

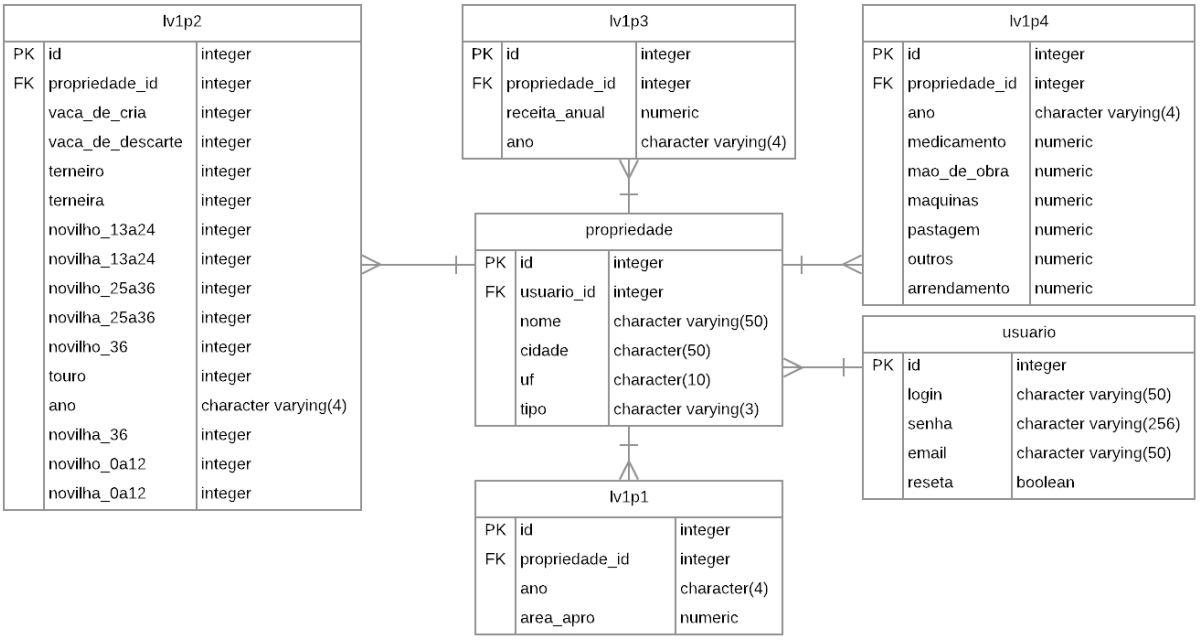
Abaixo, uma descrição técnica dos sistemas fontes dos dados da solução de BI:

Ferramenta de Gestão de Custos (FGC): Esta solução está no escopo do projeto *MyBeef*, projeto da Embrapa Pecuária Sul que visa proporcionar aos pesquisadores da Embrapa um maior conhecimento sobre a cadeia produtiva da pecuária e dos territórios da região Sul do Brasil. Esta é a versão atual do projeto, futuramente serão incorporadas e integradas outras funções, conforme mencionado no site do projeto da Embrapa:

> [...] espera-se produzir boletins, periódicos com informações sobre a cadeia produtiva, website com um banco de dados com softwares nacionais sobre a pecuária de corte e leite, matriz de indicadores de sustentabilidade para a pecuária, software para análise econômica e ambiental, método para análise espaço-temporal de indicadores sociais, econômicos, ambientais e produtivos e uma plataforma para abrigar sistemas de apoio à decisão para a pecuária, tendo como público-alvo prioritário o produtor rural. (EMBRAPA, 2015?).

Este sistema web online visa auxiliar o produtor rural a compreender melhor os seus custos dentro da atividade pecuária. O sistema está em desenvolvimento por fases. Cada fase é caracterizada pela inserção de um conjunto específico de informações sobre os custos da propriedade. Conforme o produtor conclui e avança nas fases, mais dados e informações sobre os custos de sua propriedade e produção serão requisitados. Com isso, o retorno de informações, gráficos, relatórios e *dashboards* sobre os custos serão cada vez mais enriquecidos e detalhados com essas mesmas informações de forma organizada. Até o momento, foi concluída a fase 1 do sistema. A Figura 1 apresenta o modelo lógico da FGC:

Figura 1: Modelo lógico do FGC.



Fonte: Autor (2019).

Detalhes: a tabela usuário remete às informações credenciais do usuário, a tabela propriedade remete às informações da propriedade rural e sua localização e tipo (cria, recria, engorda ou sistema completo) e as tabelas restantes possuem os dados e métricas de interesse para registro. Um usuário pode ter uma ou várias propriedades, e cada propriedade pode ter um ou vários registros de simulações. Cada tabela (exceto usuário e propriedade) refere-se a uma página na interface da FGC. Os dados das páginas conectam-se entre si através da FK contida nelas e do ano, pois não é permitido realizar duas simulações para a mesma propriedade e, simultaneamente, no mesmo ano. Portanto, a FK e o ano são os atributos que interligam as tabelas.

Livestock Sustainability: Este sistema é um aplicativo que visa fornecer ao produtor uma forma de estimar a emissão da pegada de carbono do sistema de produção, utilizando variáveis que sejam conhecidas por ele. A partir do sistema de informação, será coletado um conjunto de dados que permitirá realizar mineração de dados, de forma a tentar identificar quais variáveis mais influenciam na emissão e como será possível reduzir a emissão. Ou seja, é um sistema de informação que estima a pegada de carbono e sugere alternativas de redução.

O sistema também recebe de entrada indicadores zootécnicos e informações relacionadas relevantes da pecuária de corte, para viabilizar seus impactos em

indicadores econômicos e produtivos da propriedade. Com o uso de algumas informações da produção e propriedade (taxa de natalidade (1) e mortalidade (2) do rebanho, taxa de desmame (3), idade de acasalamento e abate do rebanho e área disponível em hectares) combinada com informações consideradas padrão para a região, é possível verificar a produtividade (4) desta propriedade.

$$Taxa\ de\ natalidade\ =\ ((n°\ de\ terneiros\ nascidos\ )/(n°\ de\ fêmeas\ acasaladas))\ *\ 100 \quad (1)$$

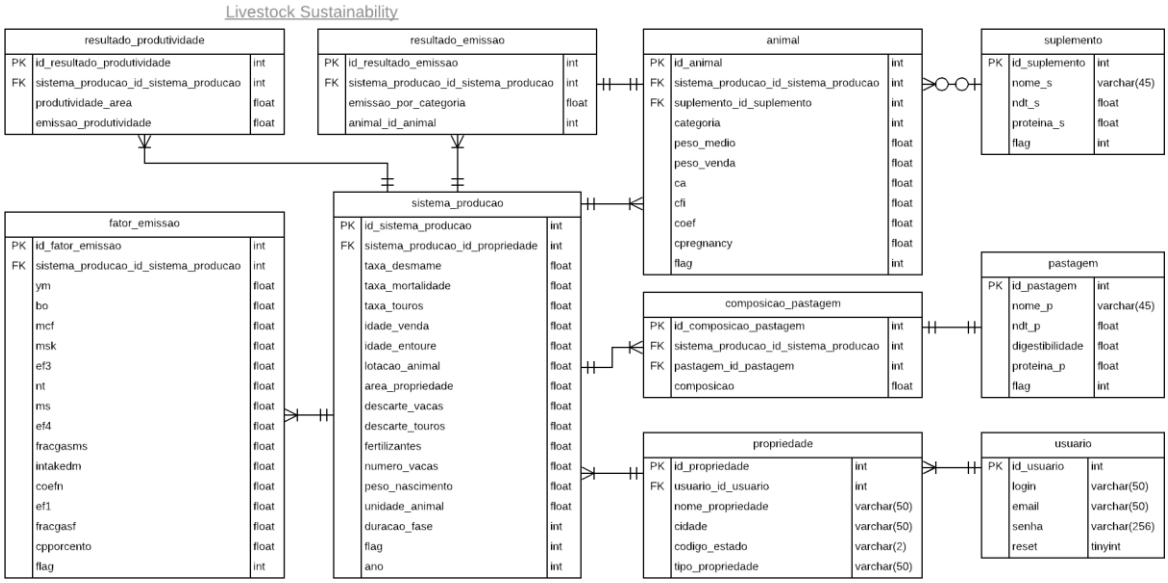
$$Taxa\ de\ mortalidade\ =\ ((n°\ de\ animais\ que\ morreram)/(n°\ total\ de\ animais))\ *\ 100 \quad (2)$$

$$Taxa\ de\ desmame\ =\ ((n°\ de\ terneiros\ desmamados)/(n°\ de\ fêmeas\ acasaladas)) \quad (3)$$

$$Produtividade\ =\ ((quantidade\ de\ quilos\ produzidos)/área) \quad (4)$$

A Figura 2 apresenta o modelo lógico do sistema LS:

Figura 2: Modelo lógico do LS.



Fonte: Autor (2019).

Detalhes: O sistema permite que o produtor coloque até quatro pastagens e para cada pastagem ele deve informar a sua composição. As variáveis da tabela ‘animal’ (peso médio e peso de venda) e ‘suplemento’ (nome, ndt, etc) devem ser inseridas uma vez para cada categoria animal, que são dez (vacas, novilhas de 3 anos, novilhas de 2 ano, novilhas de 1 ano, novilhos de 3 anos, novilhos de 2 anos, novilhos de 1 ano, bezerras, bezerros e touros). Os valores para as colunas *Cfi*, *Coef* e *Cpregnancy* da tabela ‘animal’ devem ser inseridas uma vez para cada categoria animal. No geral, são consideradas constantes para cada categoria. As demais variáveis de entrada do usuário e constantes (tabelas ‘sistema_producao’ e ‘fator_emissao’) são informadas uma única vez.

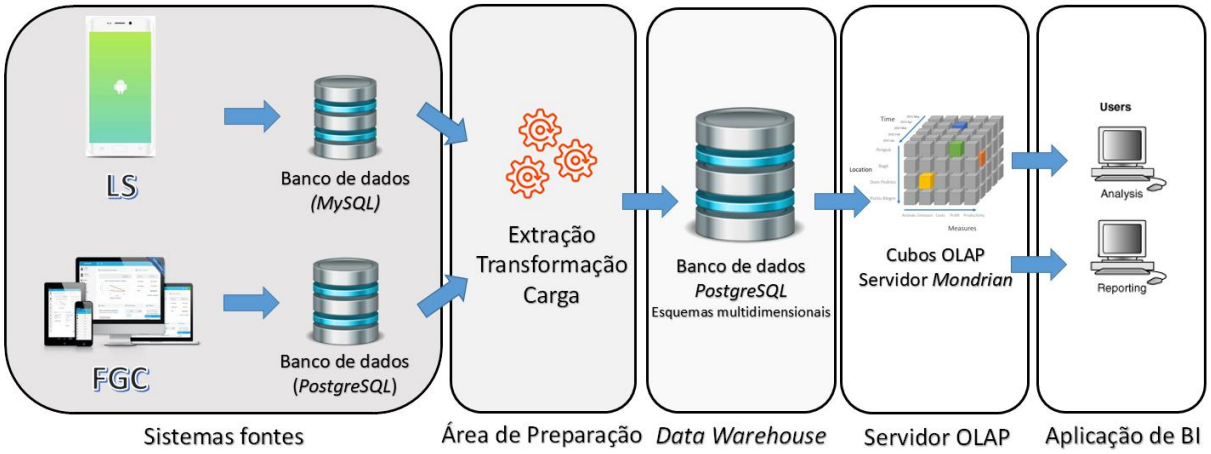
Com relação aos usuários e propriedades: a tabela usuário remete às informações credenciais do usuário, a tabela propriedade remete às informações da propriedade rural e sua localização e tipo (cria, recria, engorda ou sistema completo) e as tabelas restantes possuem os dados e métricas de interesse para registro e simulação. Um usuário pode ter uma ou várias propriedades, e cada propriedade pode ter um ou vários registros de simulações. Os registros das simulações de cada propriedade são salvos na tabela ‘sistema_producao’, que tem uma chave para a propriedade para o qual foi realizada a simulação, assim como um registro pro ano da simulação. Não é

permitido realizar duas simulações para a mesma propriedade e, simultaneamente, no mesmo ano.

2.3 Visão Geral da Arquitetura da Solução Proposta

A Figura 3 apresenta a arquitetura geral da solução proposta e é detalhada logo abaixo.

Figura 3: Visão geral da Arquitetura da solução proposta.



Fonte: Autor (2019).

Sistemas fontes: são os sistemas de informação atualmente em desenvolvimento e em fase de testes. A interação dos produtores rurais e especialistas com estes sistemas (via aplicativo para *android* no LS e navegador web online no FGC), deverá gerar um volume de dados que será armazenado em bancos de dados relacionais.

Stage Area (Área de preparação dos dados): local onde serão armazenados os dados extraídos dos sistemas fontes. É onde serão tratadas todas as questões relacionadas à qualidade dos dados (precisão, completude, consistência) e às regras de negócio do cliente. É onde ocorre o ETL.

Data Warehouse: local onde serão armazenados os dados tratados na *Stage Area*. Os dados neste ambiente estarão armazenados em um banco de dados relacional, porém, organizados em esquemas multidimensionais.

Servidor OLAP: *Mondrian* é o servidor que vai processar todas as consultas aos dados do DW.

Aplicação de BI: é a camada onde ocorre o acesso aos dados do DW pelo cliente, através de aplicações específicas de BI.

A solução de BI proposta engloba: o desenvolvimento/adaptação de suítes e tecnologias de *Business Intelligence* existentes (*Pentaho*), criação do processo de Extração, Transformação e Carga dos dados (*ETL*) das fontes de dados (FGC - SGBD *PostgreSQL* e *LS - MySQL*), com o objetivo final de auxiliar o público-alvo (produtores rurais e consultores) nas suas tomadas de decisão, através do acesso simples e eficiente às informações produtivas, econômicas e ambientais com garantias da sua qualidade.

3. Data Warehouse

3.1 Descrição das tabelas dimensões e fatos do DW

Na sequência, são detalhadas cada uma das tabelas do *Data Warehouse*: as dimensões e as fatos.

Dimensão Tempo: a dimensão temporal é essencial em projetos de DW, pois permite realizar análises históricas das mudanças nas métricas de negócio. Na pecuária de corte, considerando-se os processos de compra e venda de animais, período de gestação das vacas, época e tempo de desmame e a experiência do *stakeholder* interno, considerou-se que as métricas devem ser analisadas apenas anualmente. A maioria dos processos da pecuária de corte são analisados neste período de tempo e a grande maioria dos produtores rurais já não realizam uma gestão em um nível de detalhe maior das informações de sua propriedade, conforme apresentado na revisão da literatura.

Dimensão Localidade: a dimensão de localidade é de extrema importância neste projeto, pois através dela que é possível analisar os indicadores econômicos, ambientais e produtivos por propriedade individualmente. Foi definida apenas uma hierarquia para esta dimensão com sete níveis: país, região, unidade federativa (ou código da unidade federativa), mesorregião, microrregião, município e propriedade. É possível, portanto, analisar as métricas por estes níveis individualmente, sendo a propriedade identificada pelo seu nome, anteriormente cadastrado nos sistemas fontes de dados.

Dimensão Área: a dimensão de área permite a análise dos indicadores por agrupamentos de propriedades de acordo com sua área. Inicialmente foi proposta a divisão das faixas de áreas de acordo com a realizada pelo IBGE, porém, foi definido com o *stakeholder* interno uma divisão menor, de até 4 faixas, para evitar a poluição de informações na interface para os usuários do sistema. As faixas definidas foram de 0 à 300 hectares, de 300 à 600 hectares, de 600 à 1000 hectares e mais de 1000 hectares.

Dimensão Suplemento: a dimensão de suplemento permite que as métricas sejam analisadas pelo uso ou não de suplementos nas propriedades. Em um nível de detalhe maior, também é possível analisar as métricas por NDT (nutrientes digestíveis totais) do suplemento, quando utilizado. O NDT foi utilizado como grão de informação porque, no contexto das métricas, faz mais sentido analisá-las pela sua contribuição através de nutrientes digestivos em vez de analisá-las pelo nome do suplemento, que possui uma variedade muito grande e, ainda, um mesmo suplemento pode ter composições diferentes, além de não trazer contribuições no entendimento do seu impacto nas métricas de negócio. Tal decisão foi tomada em conjunto com o *stakeholder* interno.

Dimensão Pastagem: a dimensão de pastagem permite realizar análises das métricas por categoria de pastagem predominante (natural, nativa ou artificial, sendo esta última de verão ou inverno), por nome da pastagem predominante utilizada e pela descrição de todas pastagens e respectivas composições. O grão da informação é a pastagem predominante, que pode variar nas propriedades.

Foram identificados três processos de negócio, em que convencionou-se chamá-los de econômico, produtivo e ambiental. Estes, na sequência, serão as próprias métricas da tabela fato do *Data Warehouse*. Os processos são detalhados abaixo.

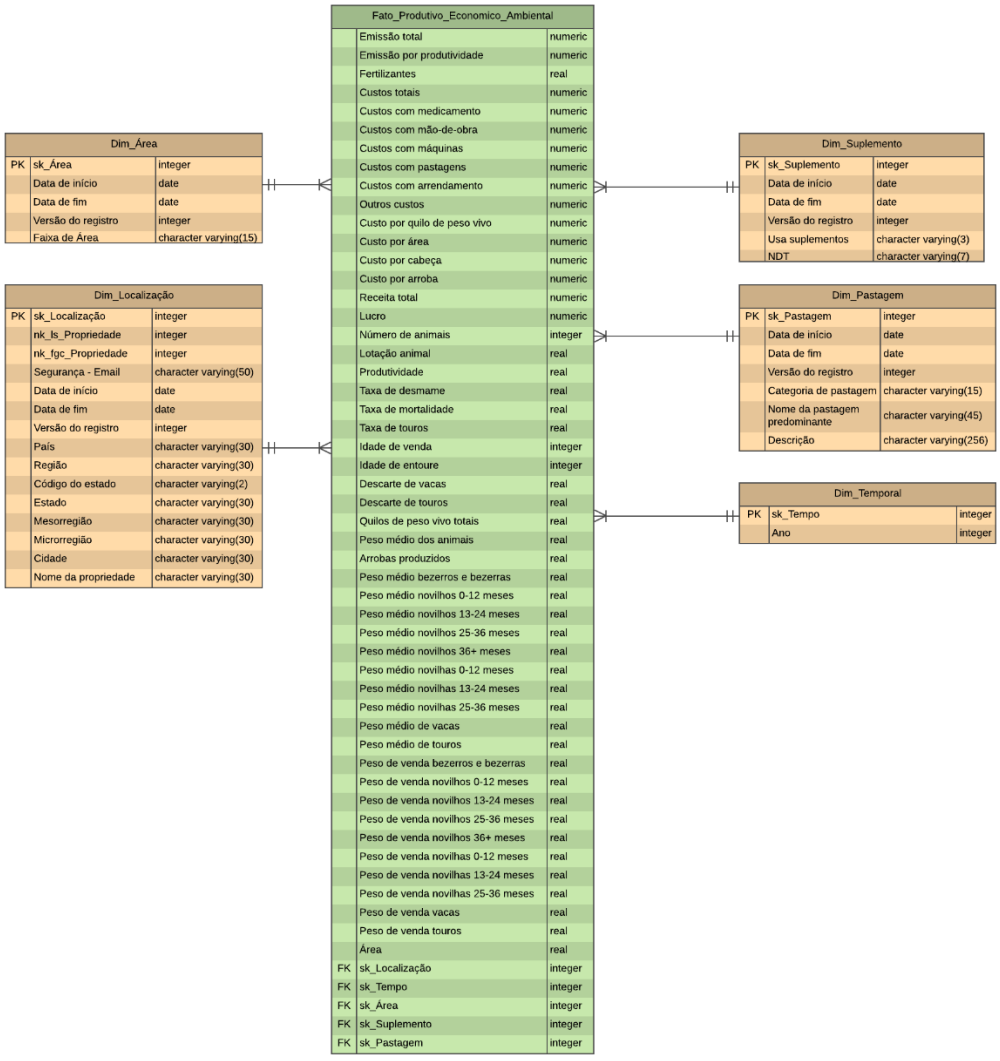
Econômico: as métricas de negócio presentes neste processo são relacionadas aos indicadores econômicos da propriedade. O grão da informação associado com as métricas deste processo é a propriedade rural em relação à localidade, anual em relação ao tempo, por faixas de área das propriedades em hectares em relação à área, por pastagem predominante em relação às pastagens e por NDT do suplemento utilizado em relação aos suplementos.

Produtivo: As métricas de negócio presentes neste processo são relacionadas aos indicadores produtivos da propriedade. O grão da informação associado com as métricas deste processo é o mesmo das métricas do processo **Econômico**.

Ambiental: As métricas de negócio presentes neste processo são relacionadas aos fatores ambientais. O grão da informação associado com as métricas deste processo é o mesmo das métricas dos processos **Econômico** e **Produtivo**.

Por fim, o modelo lógico do DW pode ser conferido na Figura 4. A Tabela 1 apresenta em detalhes as descrições de cada uma das colunas presentes nas tabelas fato e as tabelas 2, 3, 4, 5 e 6 apresentam as descrições de cada uma das colunas presentes nas tabelas dimensões. Na sequência, é possível verificar o script para implementação do DW no banco de dados.

Figura 4: Modelo lógico do DW no formato *star schema.*



Fonte: Autor (2019).

Tabela 1: Colunas da tabela ‘Fato_Produtivo_Economico_Ambiental’.

(continua)

Atributo	Tipo	Tamanho	Chave	Descrição ou equação
Sk_Localização	INTEGER	4 bytes	Estrangeira	Código que associa o registro com a dimensão localização.
Sk_Tempo	INTEGER	4 bytes	Estrangeira	Código que associa o registro com a dimensão tempo.
Sk_Área	INTEGER	4 bytes	Estrangeira	Código que associa o registro com a dimensão área.
Sk_Suplemento	INTEGER	4 bytes	Estrangeira	Código que associa o registro com a dimensão suplemento.
Sk_Pastagem	INTEGER	4 bytes	Estrangeira	Código que associa o registro com a dimensão pastagem.
Número de animais	INTEGER	4 bytes		N° total de animais.
Lotação animal	REAL	4 bytes		$\frac{UA}{Área\ (ha)}$
Produtividade por área	REAL	4 bytes		$\frac{Quilos\ de\ PV\ (kg)}{Área\ (ha) * ano}$
Taxa de desmame	REAL	4 bytes		$\frac{N°\ de\ bezerros\ desmamados * 100}{N°\ de\ fêmeas}$
Taxa de mortalidade	REAL	4 bytes		$\frac{N°\ de\ animais\ mortos\ ou\ perdidos}{N°\ total\ de\ animais}$
Taxa de touros	REAL	4 bytes		Percentual de touros em relação ao total de animais.
Idade de venda	INTEGER	4 bytes		Idade de venda dos novilhos(as).
Idade de entoure	INTEGER	4 bytes		Idade de entoure dos novilhos(as).
Descarte de vacas	REAL	4 bytes		Percentual de vacas descartadas.
Descarte de touros	REAL	4 bytes		Percentual de touros descartados.
Quilos de peso vivo totais	REAL	4 bytes		Total de quilos produzidos.
Peso médio dos animais	REAL	4 bytes		Peso médio de todos animais.
Arrobas produzidos	REAL	4 bytes		Valor produzido em arrobas.
Peso médio (todas categorias)	REAL	4 bytes		Peso médio de cada categoria animal.
Peso de venda (todas categorias)	REAL	4 bytes		Peso de venda de cada categoria animal.
Área	REAL	4 bytes		Área da propriedade em ha.
Custos totais	NUMERIC	Variável		Total de custos em reais.
Custos com medicamento	NUMERIC	Variável		Custos com medicamentos em reais.
Custos com mão-de-obra	NUMERIC	Variável		Custos com mão-de-obra em reais.
Custos com máquinas	NUMERIC	Variável		Custos com máquinas em reais.
Custos com pastagens	NUMERIC	Variável		Custos com pastagens em reais.
Custos com arrendamento	NUMERIC	Variável		Custos com arrendamento em reais.

Tabela 1: Colunas da tabela ‘Fato_Produtivo_Economico_Ambiental’.

(conclusão)

Atributo	Tipo	Tamanho	Chave	*Descrição ou equação
Outros custos	NUMERIC	Variável		Custos que não estão nas categorias anteriores, em reais.
Custos por quilo de PV produzido	NUMERIC	Variável		$\frac{\text{Custos totais (R\$)}}{\text{Quilos de PV totais (kg)}}$
Custo por área	NUMERIC	Variável		$\frac{\text{Custos totais (R\$)}}{\text{Área (ha)}}$
Custo por cabeça	NUMERIC	Variável		$\frac{\text{Custos totais (R\$)}}{\text{N° de animais}}$
Custo por arroba	NUMERIC	Variável		$\frac{\text{Custos totais (R\$)}}{\text{Arrobas (@)}}$
Receita total	NUMERIC	Variável		Total de receita gerada em reais.
Lucro	NUMERIC	Variável		Receita total-Custos totais
Emissão total	REAL	4 bytes		Emissão total de CO_2 pelos animais (kg).
Emissão por produtividade	REAL	4 bytes		$\frac{\text{Emissão total de } CO_2 \text{ (kg)}}{\text{Quilos de PV totais (kg)}}$
Fertilizantes	REAL	4 bytes		Litros de fertilizantes utilizados.

Fonte: Autor (2019).

Nota: PV = Peso Vivo, ha = hectares, kg = quilos, UA = Unidade Animal (cada UA corresponde à uma vaca de 450 kg), @ = arroba (cada @ corresponde à 15 quilos de carcaça de um animal, sendo a carcaça o peso da carne e o osso do animal apenas, em média representando 50% do peso total do animal), CO_2 = Dióxido de Carbono.

Tabela 2: Descrição das colunas da tabela ‘Dim_Temporal’.

Atributo	Tipo	Tamanho	Chave	*Descrição ou equação
Sk_Tempo	INTEGER	4 bytes	Primária	Chave primária da dimensão que é igual ao ano do registro.
Ano	INTEGER	4 bytes		Ano do registro.

Fonte: Autor (2019).

Tabela 3: Descrição das colunas da tabela ‘Dim_Área’.

Atributo	Tipo	Tamanho	Chave	*Descrição ou equação
Sk_Área	INTEGER	4 bytes	Primária	Chave primária da dimensão.
Data de início	DATE	4 bytes		Data em que foi inserido o registro da propriedade.
Data de fim	DATE	4 bytes		Data em que deixou de ser atual o registro da propriedade.
Versão do registro	INTEGER	4 bytes		Realiza o versionamento das mudanças na dimensão.
Faixa de área	CHARACTER VARYING (15)	Variável		Faixa de área ao qual a propriedade pertence.

Fonte: Autor (2019).

Tabela 4: Descrição das colunas da tabela ‘Dim_Suplemento’.

Atributo	Tipo	Tamanho	Chave	*Descrição ou equação
Sk_Suplemento	INTEGER	4 bytes	Primária	Chave primária da dimensão.
Data de início	DATE	4 bytes		Data em que foi inserido o registro da propriedade.
Data de fim	DATE	4 bytes		Data em que deixou de ser atual o registro da propriedade.
Versão do registro	INTEGER	4 bytes		Realiza o versionamento das mudanças na dimensão.
Usa suplementos?	CHARACTER VARYING (3)	Variável		Registro que identifica se a propriedade utiliza suplementos.
NDT	CHARACTER VARYING (7)	Variável		Nutrientes digestíveis totais médios dos suplementos utilizados.

Fonte: Autor (2019).

Tabela 5: Descrição das colunas da tabela ‘Dim_Localização’.

Atributo	Tipo	Tamanho	Chave	*Descrição ou equação
Sk_Localização	INTEGER	4 bytes	Primária	Chave primária da dimensão.
Nk_ls_propriedade	INTEGER	4 bytes	Natural	Chave primária no sistema LS que identifica a propriedade.
Nk_fgc_propriedade	INTEGER	4 bytes	Natural	Chave primária no sistema FGC que identifica a propriedade.
Segurança (*e-mail*)	INTEGER	4 bytes		*E-mail* do usuário. Será utilizado para implementação do DRLS.
Data de início	DATE	4 bytes		Data em que foi inserido o registro da propriedade.
Data de fim	DATE	4 bytes		Data em que deixou de ser atual o registro da propriedade.
Versão do registro	INTEGER	4 bytes		Realiza o versionamento das mudanças na dimensão.
País	CHARACTER VARYING (30)	Variável		País onde se localiza a propriedade.
Região	CHARACTER VARYING (30)	Variável		Região onde se localiza a propriedade.
Estado	CHARACTER VARYING (30)	Variável		Unidade federativa onde se localiza a propriedade.
Código do estado	CHARACTER VARYING (2)	Variável		Código da unidade federativa.
Mesorregião	CHARACTER VARYING (30)	Variável		Mesorregião onde se localiza a propriedade.
Microrregião	CHARACTER VARYING (30)	Variável		Microrregião onde se localiza a propriedade.
Cidade	CHARACTER VARYING (30)	Variável		Cidade onde se localiza a propriedade.
Nome da propriedade	CHARACTER VARYING (30)	Variável		Nome da propriedade.

Fonte: Autor (2019).

Tabela 6: Descrição das colunas da tabela ‘Dim_Pastagem’.

Atributo	Tipo	Tamanho	Chave	*Descrição ou equação
Sk_Pastagem	INTEGER	4 bytes	Primária	Chave primária da dimensão.
Data de início	DATE	4 bytes		Data em que foi inserido o registro da propriedade.
Data de fim	DATE	4 bytes		Data em que deixou de ser atual o registro da propriedade.
Versão do registro	INTEGER	4 bytes		Realiza o versionamento das mudanças na dimensão.
Categoria da pastagem	CHARACTER VARYING (30)	Variável		Categoria da pastagem.
Nome da pastagem	CHARACTER VARYING (45)	Variável		Nome da pastagem predominante.
Descrição	CHARACTER VARYING (256)	Variável		Pastagens e suas respectivas composições na propriedade em relação ao total, em porcentual.

Fonte: Autor (2019).

3.2 Script SQL para criação do Data Warehouse

```
-- Data Warehouse --
-- Banco de dados - PostgreSQL --

create sequence datawarehouse.dim_faixa_area_id_seq increment by 1 minvalue 1 maxvalue 999999 start 1;
create sequence datawarehouse.dim_pastagem_id_seq increment by 1 minvalue 1 maxvalue 999999 start 1;
create sequence datawarehouse.dim_suplemento_id_seq increment by 1 minvalue 1 maxvalue 999999 start 1;
create sequence datawarehouse.dim_localizacao_id_seq increment by 1 minvalue 1 maxvalue 999999 start 1;

DROP TABLE IF EXISTS datawarehouse.dim_area;
CREATE TABLE datawarehouse.dim_area
(
        sk_area INTEGER PRIMARY KEY NOT NULL default nextval('datawarehouse.dim_faixa_area_id_seq'),
        fx_area VARCHAR(15) NOT NULL,
         dt_inicio DATE,
         dt_fim DATE,
         flag_estado INTEGER
);

DROP TABLE IF EXISTS datawarehouse.dim_tempo;
CREATE TABLE datawarehouse.dim_tempo (
         sk_tempo INTEGER NOT NULL,
         nr_ano INTEGER NOT NULL,
         CONSTRAINT dim_tempo_pk PRIMARY KEY (sk_tempo)
);

DROP TABLE IF EXISTS datawarehouse.dim_suplemento;
CREATE TABLE datawarehouse.dim_suplemento (
         sk_suplemento INTEGER primary key NOT NULL default nextval('datawarehouse.dim_suplemento_id_seq'),
         nm_usa_suplemento VARCHAR(3) NOT NULL,
```

```
        nm_per_ndt VARCHAR(7) NOT NULL,
        dt_inicio DATE,
        dt_fim DATE,
        flag_estado INTEGER
);
DROP TABLE IF EXISTS datawarehouse.dim_pastagem;
CREATE TABLE datawarehouse.dim_pastagem (
        sk_pastagem INTEGER primary key NOT NULL default
nextval('datawarehouse.dim_pastagem_id_seq'),
        nm_cat_pastagem VARCHAR(30) NOT NULL,
        nm_pastagem VARCHAR(45) NOT NULL,
        nm_descricao VARCHAR(256) NOT NULL,
        dt_inicio DATE,
        dt_fim DATE,
        flag_estado INTEGER
);

DROP TABLE IF EXISTS datawarehouse.dim_localizacao;
CREATE TABLE datawarehouse.dim_localizacao (
        sk_localizacao INTEGER primary key NOT NULL default
nextval('datawarehouse.dim_localizacao_id_seq'),
        nm_propriedade VARCHAR(30),
        nm_cidade VARCHAR(30),
        nm_cd_unidade_federativa VARCHAR(2),
        nm_microrregiao character varying(30),
        nm_mesorregiao character varying(30),
        nm_unidade_federativa character varying(30),
        nm_regiao character varying(30),
        nm_pais character varying(30),
        dt_inicio DATE,
        dt_fim DATE,
        flag_estado INTEGER,
        nm_login_usuario VARCHAR(50),
        nk_ls_propriedade INTEGER,
        nk_fgc_propriedade INTEGER
);

DROP TABLE IF EXISTS datawarehouse.fato_produtivo_economico_ambiental;
CREATE TABLE datawarehouse.fato_produtivo_economico_ambiental (
        nr_animais INTEGER,
        nr_idade_venda INTEGER,
        nr_idade_entoure INTEGER,
        vl_lotacao_animal REAL,
        vl_produtividade REAL,
        vl_quilos_peso_vivo_total REAL,
        vl_peso_medio_total REAL,
        vl_peso_medio_novilhos_0_12 REAL,
        vl_peso_medio_novilhos_13_24 REAL,
        vl_peso_medio_novilhos_25_36 REAL,
        vl_peso_medio_novilhas_0_12 REAL,
        vl_peso_medio_novilhas_13_24 REAL,
        vl_peso_medio_novilhas_25_36 REAL,
        vl_peso_medio_vacas REAL,
        vl_peso_medio_touros REAL,
        vl_peso_medio_bezerros_bezerras REAL,
        vl_peso_venda_novilhos_0_12 REAL,
        vl_peso_venda_novilhos_13_24 REAL,
        vl_peso_venda_novilhos_25_36 REAL,
        vl_peso_venda_novilhas_0_12 REAL,
        vl_peso_venda_novilhas_13_24 REAL,
```

```
    vl_peso_venda_novilhas_25_36 REAL,
    vl_peso_venda_vacas REAL,
    vl_peso_venda_touros REAL,
    vl_peso_venda_bezerros_bezerras REAL,
    per_desmame REAL,
    per_mortalidade REAL,
    per_touros REAL,
    per_descarte_vacas REAL,
    per_descarte_touros REAL,
    der_arrobas_totais REAL,
    area REAL,

    vl_custos_medicamentos NUMERIC,
    vl_custos_mao_de_obra NUMERIC,
    vl_custos_maquinas NUMERIC,
    vl_custos_pastagens NUMERIC,
    vl_custos_arrendamento NUMERIC,
    vl_custos_outros NUMERIC,
    vl_receita_total NUMERIC,
    der_custos_totais NUMERIC,
    der_custos_por_quilo_peso_vivo NUMERIC,
    der_custos_por_area NUMERIC,
    der_custos_por_cabeca NUMERIC,
    der_custos_por_arroba NUMERIC,
    der_lucro NUMERIC,

    vl_emissao_total REAL,
    vl_emissao_por_produtividade REAL,
    vl_fertilizantes REAL,

    sk_tempo INTEGER,
    sk_localizacao INTEGER,
    sk_area INTEGER,
    sk_suplemento INTEGER,
    sk_pastagem INTEGER,
    CONSTRAINT fato_produtivo_economico_ambiental_dim_localizacao_fk FOREIGN KEY
(sk_localizacao) REFERENCES datawarehouse.dim_localizacao (sk_localizacao) MATCH SIMPLE,
    CONSTRAINT fato_produtivo_economico_ambiental_dim_tempo_fk FOREIGN KEY
(sk_tempo) REFERENCES datawarehouse.dim_tempo (sk_tempo) MATCH SIMPLE,
    CONSTRAINT fato_produtivo_economico_ambiental_dim_area_fk FOREIGN KEY (sk_area)
REFERENCES datawarehouse.dim_area (sk_area) MATCH SIMPLE,
    CONSTRAINT fato_produtivo_economico_ambiental_dim_suplemento_fk FOREIGN KEY
(sk_suplemento) REFERENCES datawarehouse.dim_suplemento (sk_suplemento) MATCH SIMPLE,
    CONSTRAINT fato_produtivo_economico_ambiental_dim_pastagem_fk FOREIGN KEY
(sk_pastagem) REFERENCES datawarehouse.dim_pastagem (sk_pastagem) MATCH SIMPLE
);
```

4. Processo ETL

A *stage area* foi implementada no SGBD *PostgreSQL*, no mesmo servidor do DW, porém em um *schema* diferente. Todo o processo de ETL foi desenvolvido na ferramenta PDI.
Foram utilizados alguns padrões para o projeto físico da *stage area*, assim como o projeto da arquitetura do ETL, visando simplificar o entendimento da solução implementada para futuros desenvolvedores. Os nomes de todas as tabelas relacionadas com a *stage area* seguem o padrão ST_Y_W_Z, onde:

- ST é o prefixo, sigla para *Stage Area*, indicando que a tabela é parte dela;
- Y refere-se ao processo realizado que originou os dados da tabela. Tem-se duas possibilidades, sendo 'extracao' para o processo de extração e 'transformacao' para o processo de transformação dos dados;
- W refere-se ao nome da(s) tabela(s) relacionada(s) ao sistema fonte de dados do qual serão utilizados ou tratados os dados;
- Z refere-se ao sistema do qual os dados foram coletados. Tem-se duas possibilidades, sendo 'ls' para o sistema LS e 'fgc' para o sistema FGC.

Para a nomeação das colunas das tabelas, foi utilizado o padrão X_Y_Z, onde:

- X é o prefixo, indicando o significado da coluna, em termos de ETL. Os valores possíveis são 'pk' (*primary key* – chave primária), 'nk' (*natural key* – chave natural), 'fk' (*foreign key* – chave estrangeira), 'nm' (nome, variável do tipo *string*), 'nr' (número, variável do tipo *integer*), 'vl' (valor, variável do tipo *real* ou *numeric*) e 'per' (percentual, variável do tipo *real* ou *numeric*);
- Y é a variável intermediária que indica o significado semântico, em termos de métrica de negócio, daquela coluna. Alguns exemplos como 'ndt_suplementos_touros', 'peso_medio_vacas, 'peso_venda_vacas', entre outros, foram utilizados com frequência;
- Z é a variável que indica o sistema fonte dos dados daquela coluna, sendo 'ls' para o sistema LS e 'fgc' para o sistema FGC. Quando não especificado, refere-se à ambos;

O resultado deste processo foi a criação das seguintes tabelas para a parte principal do ETL:

a) **stage_area.st_extracao_usuario_fgc**: refere-se aos dados de usuário extraídos da FGC.
b) **stage_area.st_extracao_propriedade_fgc**: refere-se aos dados de propriedade extraídos da FGC.
c) **stage_area.st_extracao_fases_fgc**: refere-se aos dados de cada página da fase 1 extraídos da FGC.
d) **stage_area.st_extracao_usuario_ls**: refere-se aos dados de usuário extraídos do LS.
e) **stage_area.st_extracao_propriedade_ls**: refere-se aos dados de propriedade extraídos do LS.
f) **stage_area.st_extracao_sistema_produtivo_ls**: refere-se aos dados de cada sistema produtivo extraído do LS.
g) **stage_area.st_transformacao_usuario**: refere-se aos dados de usuário de ambos sistemas, após processados, tratados e integrados.

h) **stage_area.st_transformacao_propriedade**: refere-se aos dados de propriedade de ambos sistemas, após processados, tratados e integrados.
i) **stage_area.st_transformacao_sistema_produtivo**: refere-se aos dados do sistema produtivo de ambos sistemas, após processados, tratados e integrados.
j) **stage_area.st_erros_usuario**: refere-se aos dados de usuário de ambos sistemas no qual possa ter ocorrido algum erro no processo de transformação de dados.
k) **stage_area.st_erros_propriedade**: refere-se aos dados de propriedade de ambos sistemas no qual possa ter ocorrido algum erro no processo de transformação de dados.
l) **stage_area.st_erros_sistema_produtivo**: refere-se aos dados de sistema produtivo de ambos sistemas no qual possa ter ocorrido algum erro no processo de transformação de dados.

Também foram criadas as seguintes tabelas auxiliares:

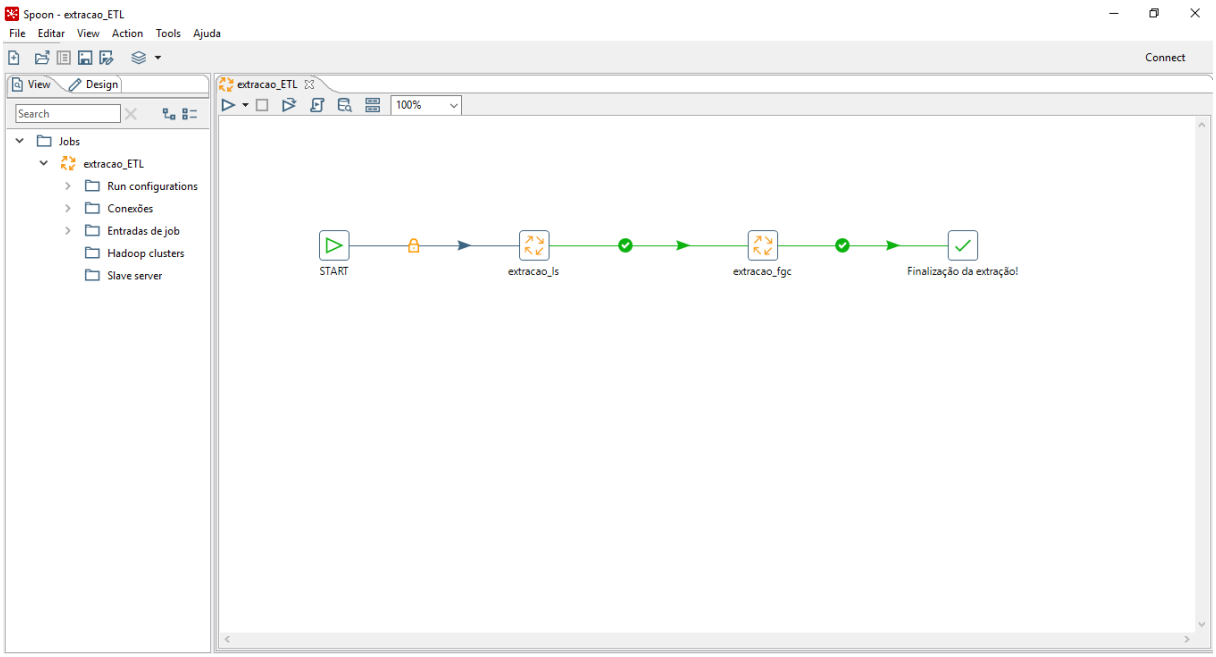
a) **public.categoria_pastagem**: tabela com registro dos possíveis tipos de pastagens (natural, nativa ou artificial).
b) **public.pastagens**: tabela com registro das possíveis pastagens (nomes), com relacionamento n:1 com a tabela **public.categoria_pastagem**.
c) **public.paises, public.regioes, public.mesorregioes, public.microrregioes, public.municipios, public.distritos** e **public.subdistritos**: tabelas sobre os diferentes níveis hierárquicos de localizações no Brasil com informações pré-cadastradas do IBGE sobre as mesmas.

Para fins de simplificação do projeto de ETL, o processo foi dividido em três conjuntos de etapas: extração; transformação e integração; e carga. Dois conjuntos de tabelas foram utilizados para a *stage area*, sendo um deles para armazenar os dados extraídos dos sistemas fontes e o outro para armazenar os dados transformados, tratados e integrados na etapa de transformação. Tal separação foi realizada para evitar a mistura de dados tratados e validados com dados ainda não avaliados. Ainda, na etapa de extração, foram derivados dois subconjuntos de tabelas, um para armazenar dados do sistema LS e outro para os dados do sistema FGC. A etapa de carga é o momento em que os dados são, de fato, consolidados no DW.
As Tarefas e Transformações foram alocadas em três diretórios de acordo com o processo relacionado: extração, transformação e integração, e carga. Ainda, os arquivos relacionados foram nomeados com o padrão X_Y_Z, onde X representa a parte do processo (extração, transformação, integração ou carga), Y representa o processo de negócio, tabela ou parte específica do ETL com o qual são movimentados os dados e Z representa o sistema fonte dos dados daquela tarefa. Padrões semelhantes também foram utilizados na nomeação de cada um dos passos dentro de cada transformação e tarefa.

4.1 Processo de extração dos dados

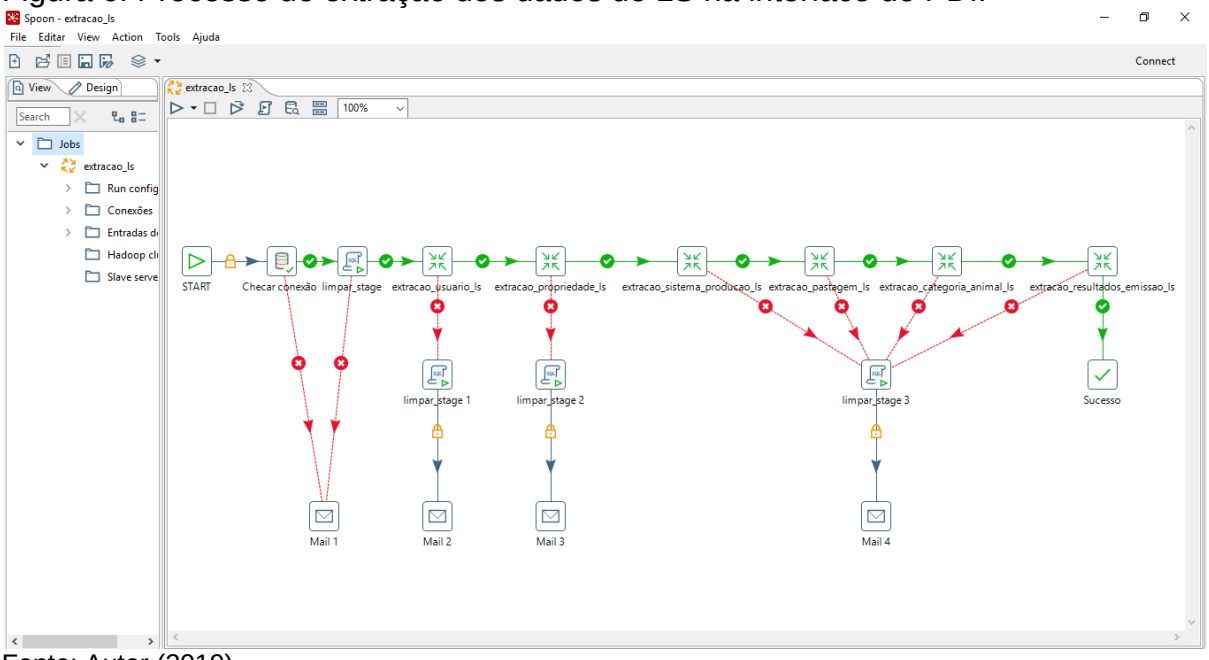
A Figura 5 apresenta duas tarefas, cada uma para extrair dados do respectivo sistema fonte de dados (LS e FGC). Na sequência, é apresentada a tarefa de extração de dados do LS, que pode ser analisada em detalhes na Figura 6.

Figura 5: Processo de extração dos dados na interface do PDI.



Fonte: Autor (2019).

Figura 6: Processo de extração dos dados do LS na interface do PDI.



Fonte: Autor (2019).

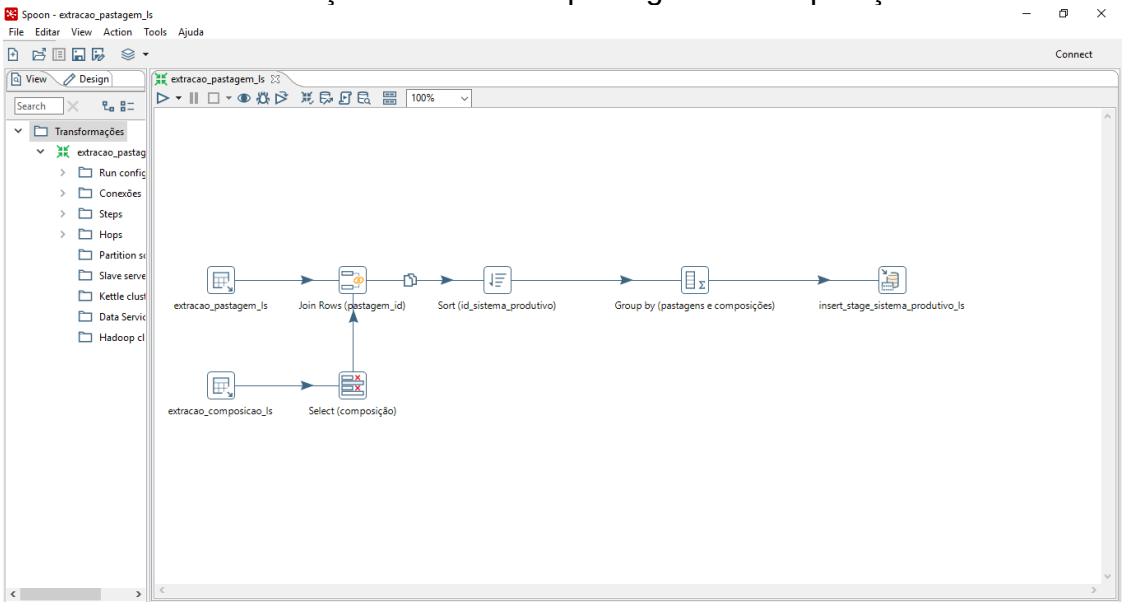
É possível observar, na Figura 6, o seguinte fluxo de etapas para a extração:

9. **‘checar conexão’**: verifica a conexão com o banco de dados do sistema LS (configurado localmente) e com o banco de dados da *stage area* (configurado localmente). O mesmo procedimento foi realizado no processo de extração de dados do FGC. Em caso de falha na conexão, é disparado um *e-mail* para o administrador do sistema alertando o dia e horário da falha, assim como o passo, tarefa e transformação que originou a falha, e a Tarefa interrompe sua execução. O nível de detalhe dos erros pode ser modificado pelo usuário no passo de envio do *e-mail*. Em caso de sucesso, o fluxo da Tarefa prossegue para a próxima etapa.

Para o envio do *e-mail*, foi utilizado o servidor SMTP (*Simple Mail Transfer Protocol*) padrão do *gmail*, com as respectivas informações de *login* e senha do administrador do sistema. A restrição para o uso do servidor SMTP do *gmail* é o envio de, no máximo, 100 *e-mails* por dia.

10. **‘limpar_stage’**: é realizado o truncamento das tabelas da stage_area que receberão fluxos de dados do sistema LS. Este passo é executado através de um comando SQL (TRUNCATE TABLE "stage_area"."st_extracao_usuario_ls" CASCADE) que deleta todos os dados contidos de outras execuções prévias em todas as tabelas relacionadas ao sistema LS no processo de extração. O comando CASCADE foi utilizado para garantir a limpeza das tabelas de usuário e todas tabelas que possuam uma chave ligada à ela. Este passo é necessário para garantir que apenas os dados extraídos do LS daquele dia serão processados. Se este passo falhar, o mesmo processo descrito no passo 1, em caso de falha, ocorrerá.
11. **‘extracao_usuario_ls’**: é uma transformação que realiza a extração dos dados da tabela de usuário do LS e os carrega para a tabela de usuário do LS na *stage area*. Em caso de falhas nesta ou nas próximas transformações: qualquer dado anterior extraído é deletado da *stage area* (através do mesmo comando do passo 2) e o processo do passo 1 em caso de falha é executado (envio de *e-mail* para administrador).
12. **‘extracao_propriedade_ls’**: é uma transformação que realiza a extração dos dados da tabela de propriedade do LS e carrega-os para a tabela de propriedade do LS na *stage área.*
13. **‘extracao_sistema_producao_ls’**: é uma transformação que realiza a extração dos dados da tabela de sistema de produção do LS e os carrega para a tabela de sistema de produção do LS na *stage area.*
14. **‘extracao_pastagens_ls’**: é uma transformação que realiza a extração dos dados das tabelas de composições e pastagens do LS e os carrega para a tabela de sistema de produção do LS na *stage area.* Esta transformação é detalhada na Figura 7 e apresenta os seguintes passos:

Figura 7: Passos de extração de dados de pastagens e composições do LS.



Fonte: Autor (2019).

a. **‘extracao_pastagens_ls’**: é realizada a importação dos dados das pastagens do LS através de um comando SQL.
b. **‘extracao_composisao_ls’**: é realizada a importação dos dados das composições das pastagens do LS através de um comando SQL.
c. **‘Select (Composição)’**: é realizado um arredondamento dos valores de composição para duas casas decimais.
d. **‘Join Rows (pastagen_id)’**: é realizada uma junção (ou produto cartesiano) das informações das pastagens com suas respectivas composições neste passo, através do uso da chave que liga ambas as informações (pastagem_id). Cabe mencionar que uma propriedade pode ter entre uma e quatro pastagens, e respectivas composições.
e. **‘Sort (id_sistema_produtivo)’**: as informações são ordenadas de acordo com a chave para o respectivo sistema produtivo (id_sistema_produtivo).
f. **‘Group By (Pastagens e Composições)’**: as informações das pastagens e composições, por propriedade, são agrupadas em duas colunas, sendo essa uma operação de concatenação de *strings*.
g. **‘insere_stage_sistema_produtivo_ls’**: as informações de pastagens e composições, por propriedade, concatenadas no passo anterior, são finalmente inseridas na *stage area*, na tabela de sistema de produção do LS.

15. **‘extracao_categoria_animal_ls’**: é uma transformação que visa realizar a extração dos dados das tabelas de animais e suplementos do LS e carregar para a tabela de sistema de produção do LS na *stage area.* Esta transformação é detalhada na Figura 8 e apresenta os seguintes passos.

Figura 8: Passos de extração das categorias animais e suplementos do LS.

Fonte: Autor (2019).

a. **‘extracao_cat_animal_ls’**: é realizada a importação dos dados da tabela ‘animal’ do LS através de um comando SQL.
b. **‘extracao_suplemento_ls’**: é realizada a importação dos dados da tabela ‘suplementos’ do LS através de um comando SQL.

c. **‘Join Rows (animal_id)’**: é realizada uma junção (ou produto cartesiano) das informações da categoria animal com seus respectivos suplementos neste passo, através do uso da chave que liga ambas as informações (animal_id). Cabe mencionar que cada categoria animal pode ter entre zero e um suplemento.
d. **‘Sort (id_sistema_produtivo)’**: as informações são ordenadas de acordo com a chave para o respectivo sistema produtivo (id_sistema_produtivo).
e. **‘Row denormalizer (categoria)’**: as informações reunidas de categoria animal e suplementos são desnormalizadas neste passo. A coluna categoria animal, que armazenava um inteiro que indicava a categoria animal de cada registro, foi utilizada como chave para criar 27 novas colunas, baseadas nos atributos peso médio, peso de venda e NDT do suplemento de cada categoria animal.
f. **‘insere_stage_sistema_produtivo_ls’**: as informações por categoria animal do passo anterior são inseridas na *stage area*, na tabela de sistema de produção do LS.

16. **‘extracao_resultado_emissao_ls’**: é uma transformação que visa realizar a extração dos dados das tabelas ‘resultados de produtividade’ e ‘resultados de emissão’ do LS e carregar para a tabela de sistema de produção do LS na *stage area.* Esta transformação é detalhada na Figura 9 e apresenta os seguintes passos.

Figura 9: Passos de extração dos resultados de emissão e produtividade do LS.

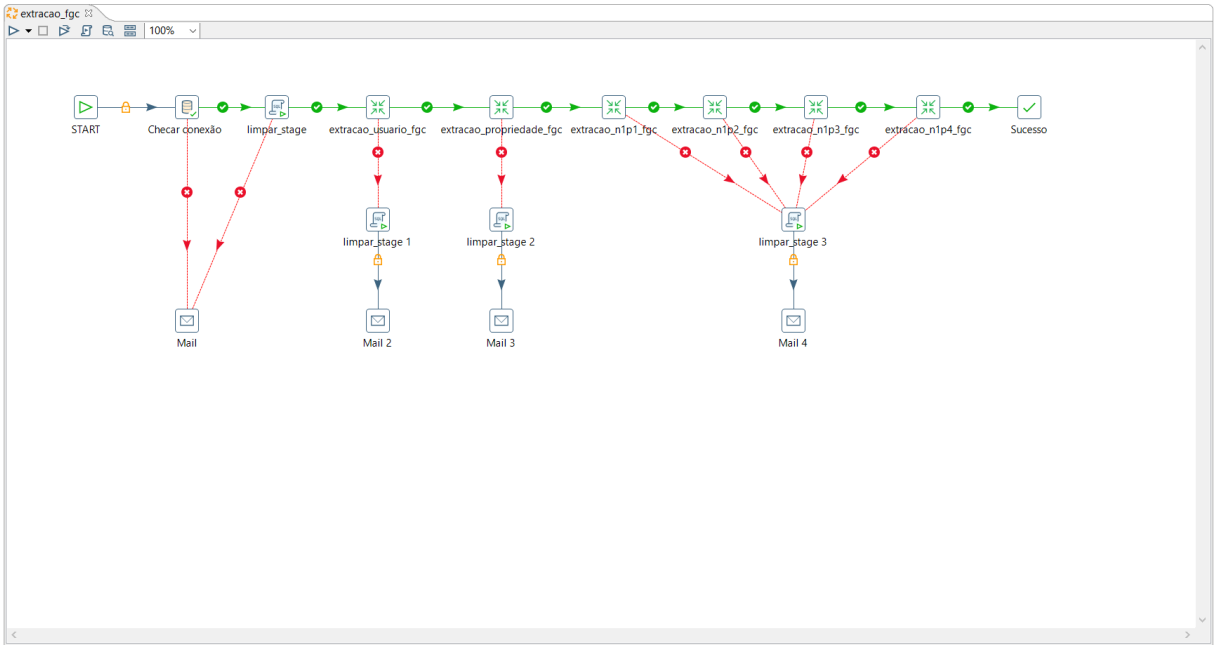
Fonte: Autor (2019).

a. **‘extracao_emissao_ls’**: é realizada a importação dos dados da tabela ‘resultados de emissão’ do LS através de um comando SQL.
b. **‘Sort (id_sistema_produtivo)’**: as informações são ordenadas de acordo com a chave para o respectivo sistema produtivo (id_sistema_produtivo).
c. **‘Row denormalizer (categoria)’**: as informações reunidas de emissão por categoria animal são desnormalizadas neste passo. A coluna categoria animal, que armazena um inteiro que indica a categoria animal de cada registro, foi utilizada como chave para criar 9 novas colunas, baseadas no atributo emissão de cada categoria animal.

d. **‘insere_stage_sistema_produtivo_ls’**: as informações de emissão por categoria animal do passo anterior são inseridas na *stage area*, na tabela de sistema de produção do LS.
e. **‘extracao_resultados_ls’**: é realizada a importação dos dados da tabela ‘resultados de produtividade’ do LS através de um comando SQL.
f. **‘insere_stage_sistema_produtivo_ls 2’**: as informações de produtividade e emissão por produtividade, por propriedade, são inseridas na *stage area*, na tabela de sistema de produção do LS.

Na sequência, é apresentada a tarefa de extração de dados da FGC, que pode ser analisada em detalhes na Figura 10.

Figura 10: Processo de extração dos dados do sistema FGC.

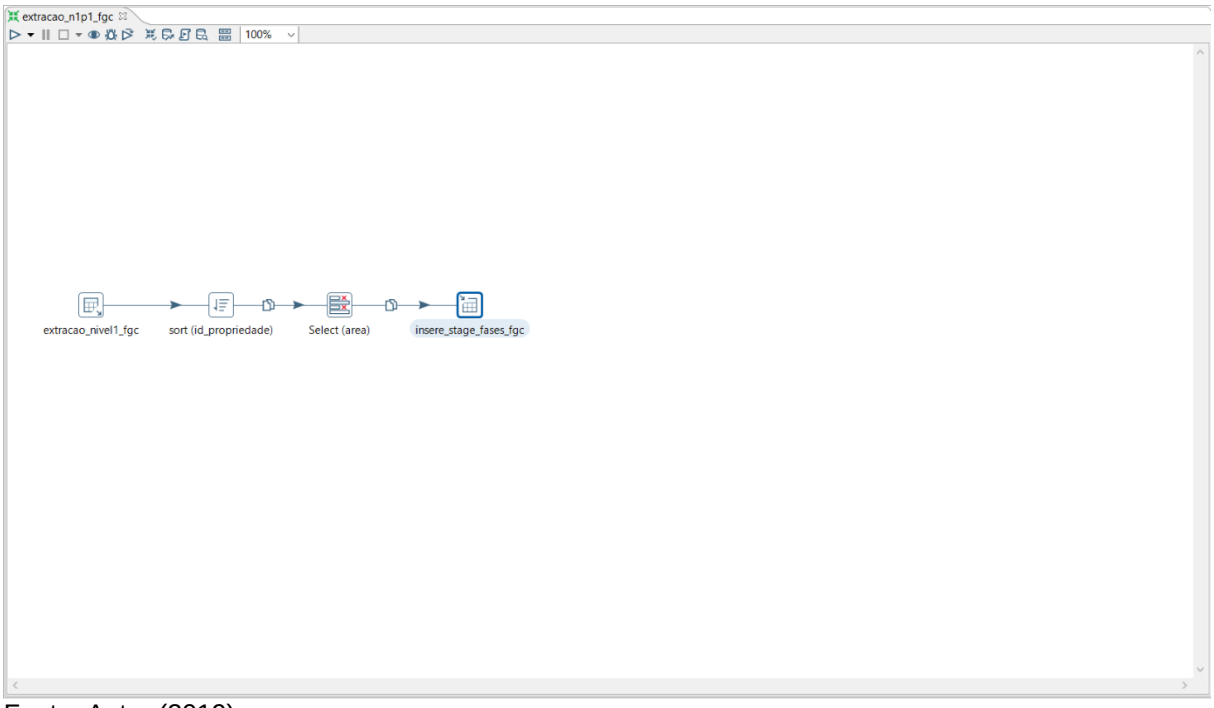


Fonte: Autor (2019).

É possível observar, na Figura 6, o seguinte fluxo de etapas para a extração:
1. **‘checar conexão’**: verifica a conexão com o banco de dados do sistema FGC (configurado localmente) e com o banco de dados da *stage area* (configurado localmente). O mesmo procedimento foi realizado no processo de extração de dados do LS. Em caso de falha na conexão, é disparado um e-mail para o administrador do sistema alertando o dia e horário da falha, assim como o passo, tarefa e transformação que originou a falha, e a Tarefa interrompe sua execução. O nível de detalhe dos erros pode ser modificado pelo usuário no passo de envio do e-mail. Em caso de sucesso, o fluxo da Tarefa prossegue para a próxima etapa.
2. **‘limpar_stage’**: é realizado o truncamento das tabelas da *stage area* que receberão fluxos de dados do sistema LS. Este passo é executado através de um comando SQL (TRUNCATE TABLE "stage_area"."st_extracao_usuario_fgc" CASCADE) que deleta todos os dados contidos de outras execuções prévias em todas as tabelas relacionadas ao sistema FGC no processo de extração. O comando CASCADE foi utilizado para garantir a limpeza das tabelas de usuário e todas tabelas que possuam uma chave ligada à ela. Este passo é necessário para garantir que apenas os dados

extraídos do FGC daquele dia serão processados. Se este passo falhar, o mesmo processo descrito no passo 1, em caso de falha, ocorrerá.
3. **‘extracao_usuario_fgc’**: é uma transformação que realiza a extração dos dados da tabela de usuário do FGC e os carrega para a tabela de usuário da FGC na *stage area*. Em caso de falhas nesta ou nas próximas transformações: qualquer dado anterior extraído é deletado da *stage area* (através do mesmo comando do passo 2) e o processo do passo 1 em caso de falha é executado (envio de e-mail para administrador).
4. **‘extracao_propriedade_fgc’**: é uma transformação que realiza a extração dos dados da tabela de propriedade da FGC e carrega-os para a tabela de propriedade da FGC na *stage area*.
5. **‘extracao_n1p1_fgc’**: é uma transformação que realiza a extração dos dados da tabela relacionada à página 1 do nível 1 da FGC e os carrega para a tabela de fases da FGC na *stage area*. O processo pode ser visualizado na Figura 11 e possui os seguintes passos:

Figura 11: Processo de extração dos dados da página 1 do nível 1 da FGC.



Fonte: Autor (2019).

a. **‘extracao_nivel1_fgc’**: é realizada a extração dos dados da tabela relacionada à página 1 do nível 1 da FGC através de um comando SQL.
b. **‘Sort (id_propriedade)’**: as informações são ordenadas de acordo com a chave para a respectiva propriedade (id_propriedade).
c. **‘Select (area)’**: as informações de interesse da tabela da página 1 no nível 1 são selecionadas e o ano é convertido para o tipo de dado adequado (*integer*).
d. **‘insere_stage_fases_fgc’**: as informações do passo anterior são inseridas na *stage area*, na tabela de fases da FGC.

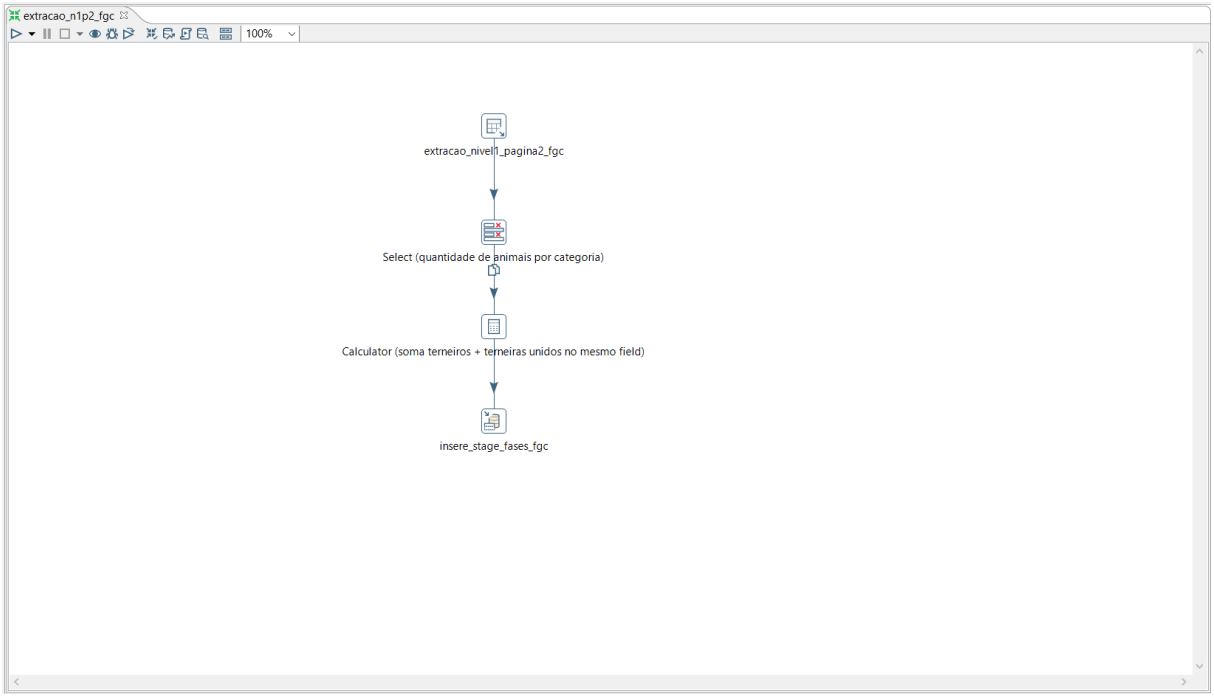
As transformações ‘**extracao_n1p3_fgc**’ e ‘**extracao_n1p4_fgc’** da Figura 10 são idênticas às apresentadas na Figura 11. Já a transformação ‘**extracao_n1p2_fgc**’, é apresentada na Figura 12 e explorada em maiores detalhes na sequência.

a. **‘extracao_nivel2_fgc’**: é realizada a extração dos dados da tabela relacionada à página 2 do nível 1 da FGC através de um comando SQL.
b. **‘Select (quantidade de animais por categoria)’**: as informações são selecionadas e o ano é convertido para o tipo de dados adequado (*integer*).
c. **‘Calculator (soma terneiros + terneiras unidos no mesmo field)’**: as quantidades de terneiros e terneiras são somadas em um único campo.
d. **‘insere_stage_fases_fgc’**: as informações do passo anterior são inseridas na *stage area*, na tabela de fases da FGC.

Cabe mencionar que, devido à inexistência de colunas que indiquem a data e hora de inserção ou atualização dos registros nos sistemas fontes, foi adotado um esquema de carga completa (*full load*) em que todos os dados sempre serão extraídos de ambos sistemas fontes.

Com o sucesso da execução das tarefas e transformações anteriores, inicia-se o próximo passo do ETL, a transformação dos dados, onde são executadas tarefas de pré-processamento dos dados como preenchimento de algumas informações faltantes, correção e conversão de tipos dos dados e validação de informações. Cabe mencionar que checagens de integridade entre chaves primárias e estrangeiras (de entidade e referencial) das tabelas são realizadas automaticamente, pois no processo de extração, se ocorrer de algum registro possuir chave estrangeira que seja inexistente na tabela relacionada ou não possuir chave primária, o processo de extração é imediatamente interrompido.

Figura 12: Processo de extração dos dados da página 2 do nível 1 da FGC.



Fonte: Autor (2019).

4.2 Processo de transformação dos dados

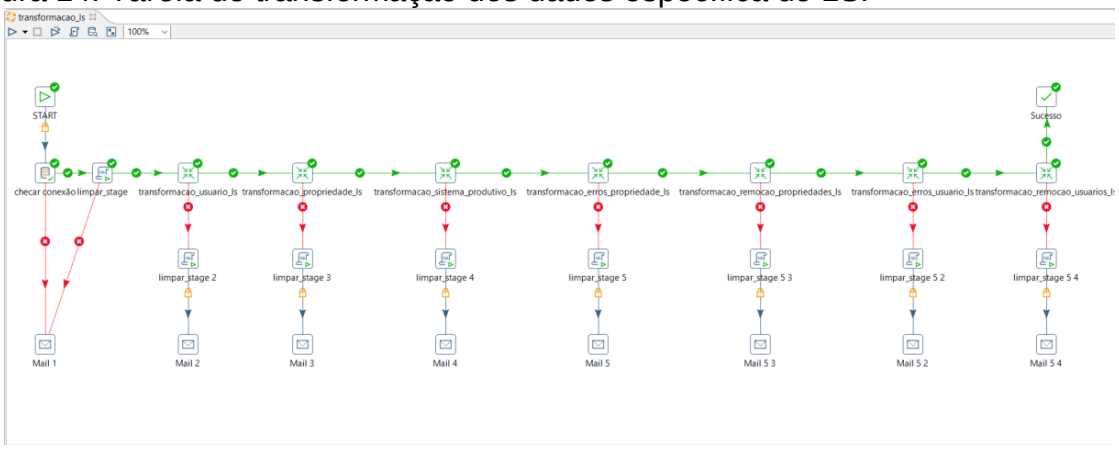
A Figura 13 apresenta duas tarefas, cada uma para transformar, processar e integrar os dados do respectivo sistema fonte de dados (LS e FGC). A tarefa de transformação

de dados específica do LS e a ordem de seus respectivos passos podem ser analisados na Figura 14.

Figura 13: Tarefa de transformação dos dados na interface do PDI.

Fonte: Autor (2019).

Figura 14: Tarefa de transformação dos dados específica do LS.



Fonte: Autor (2019).

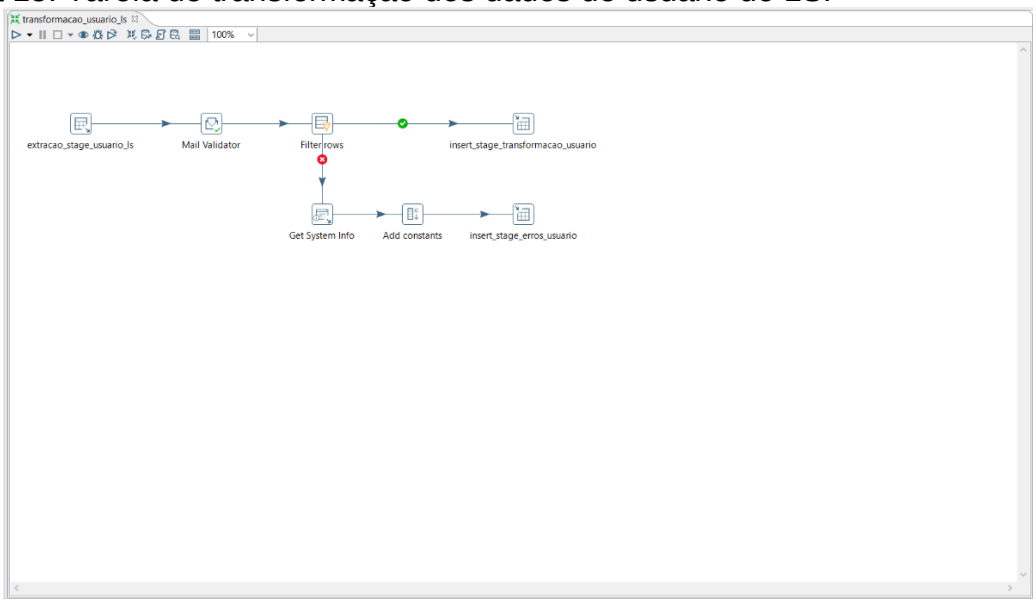
Todos os registros que por quaisquer razões não tenham sido inseridos nas tabelas de transformação da *stage area* são inseridos em um outro conjunto de tabelas que convencionou-se chama-las de '**stage_erro_usuario**', '**stage_erro_propriedade**', e '**stage_erro_sistema_produtivo**'. São tabelas semelhantes às do processo de transformação, porém com duas colunas adicionais em cada tabela, sendo uma para o registro da data e hora da ocorrência de determinada violação de restrição e outra para identificar a razão do erro. Convencionou-se a existência de três tipos de erros, sendo atribuído o valor '1' quando o erro for relacionado à violação de regras de negócio e valores inválidos, incorretos ou nulos, o valor '2' quando o erro envolver violação de restrições de integridade referencial e o valor '3' quando o registro na tabela não possuir dados nas tabelas de sistema produtivo. Este conjunto de tabelas de erros são persistentes, caso o administrador do sistema necessite realizar correções manuais, ele poderá localizar, acessar e corrigir tais registros, seja no

processo de extração ou nos próprios sistemas fontes, caso possua autorização. Ainda, tais registros subsidiarão ações de correções e melhorias nos próprios sistemas fontes de dados, visando obter melhorias na qualidade dos dados extraídos para os tratamentos no ETL.

Na sequência, é realizado o detalhamento dos passos da transformação dos dados do LS:

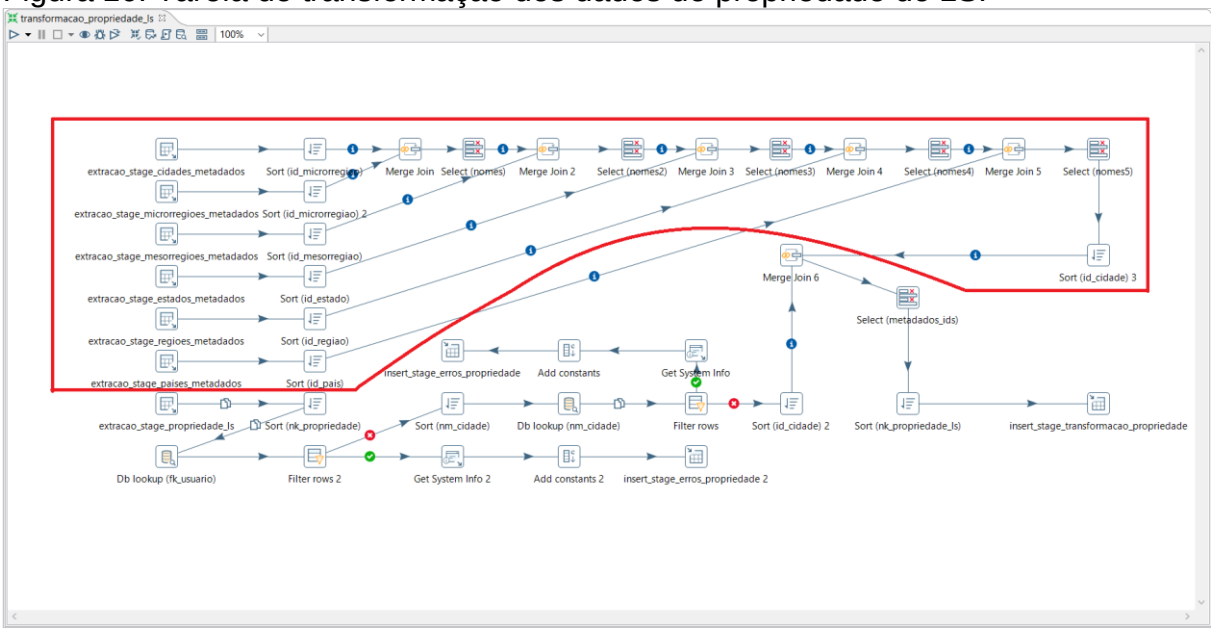
1. **‘checar conexão’** e ‘**limpar_stage’** são idênticos aos mencionados nos passos 1 e 2 da etapa de extração de dados do LS e FGC.
2. **‘transformacao_usuario_ls’** simplesmente move os dados da tabela de usuários do LS da stage extração para a stage transformação. Os únicos dados de interesse são o identificador (chave primária) único da tabela e o *e-mail* do usuário. É realizado um teste para verificar se o e-mail do usuário possui um formato válido, e caso seja, o registro é inserido na tabela **st_transformação_usuario**, caso contrário, o registro com identificação do tipo de erro (1) e data de ocorrência do erro é salvo na tabela de erros **st_erros_usuario**. A transformação pode ser analisada na Figura 15.
3. **‘transformacao_propriedade_ls’:** esta transformação não só move os dados da stage extração para a stage transformação, como realiza a validação dos dados de localização informados. A Figura 16 apresenta os passos desta transformação, conforme detalhado na sequência. Os passos presentes na região demarcada em vermelho fazem parte de um processo de busca por informações de localização previamente cadastradas em um banco de metadados, com tabelas normalizadas. Nestas tabelas, presentes em um *schema* “metadados”, no mesmo banco de dados do DW, há as seguintes informações sobre o Brasil: regiões, estados, mesorregiões, microrregiões e municípios. Também foram cadastradas informações de distritos e subdistritos, porém no momento, tais informações não são utilizadas, pois o grão da informação cadastrado nos sistemas fontes é o município. Caso, no futuro, as informações sejam mais detalhadas, o ETL já estará preparado. Os dados foram obtidos do site do IBGE e da base de dados SIDRA. Os passos abaixo da região vermelha são detalhados na sequência:

Figura 15: Tarefa de transformação dos dados de usuário do LS.



Fonte: Autor (2019).

Figura 16: Tarefa de transformação dos dados de propriedade do LS.



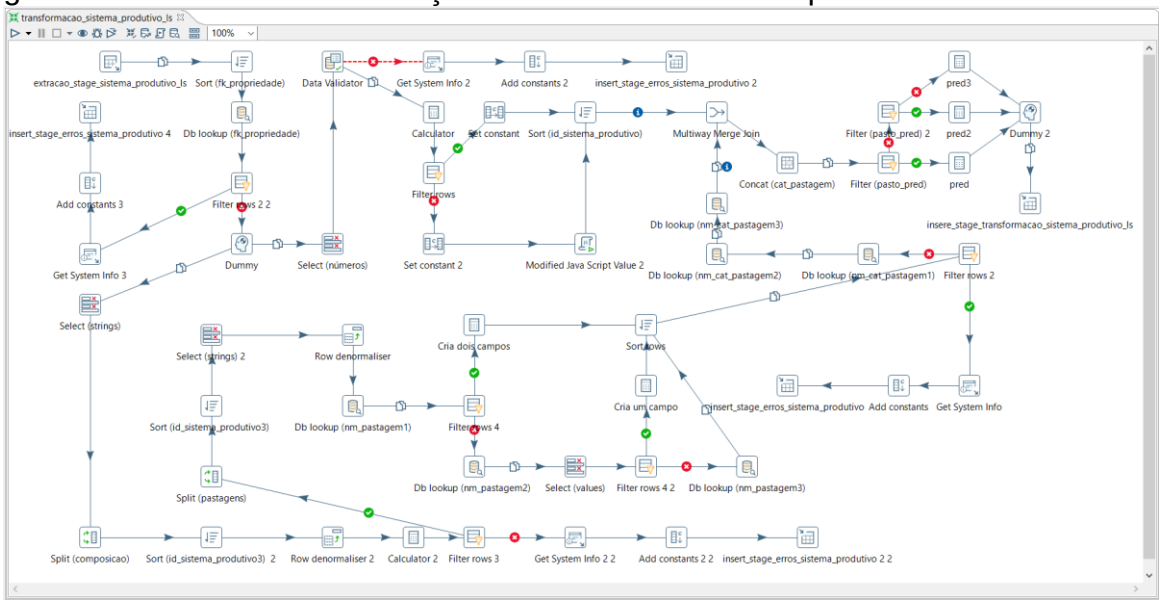
Fonte: Autor (2019).

a. ‘**extracao_stage_propriedade_ls**’: passo que realiza a extração dos dados de propriedades do LS obtidos no processo de extração.
b. ‘**Sort (nk_propriedade)**’: passo que ordena os registros em ordem pelo atributo ‘nk_propriedade’. Sempre antes de realizar um passo de *lookup* em um banco de dados, é necessário realizar a ordenação da chave de busca.
c. ‘**Db lookup (fk_usuario)**’: este passo verifica se os registro das propriedades possuem um usuário relacionado. Pode ocorrer este caso devido à transformação anterior, onde podem ter sido descartados registros de usuários. É retornado o identificador do usuário, caso exista, ou NULL caso contrário.
d. ‘**Filter rows 2**’: filtra os registros de propriedades com e sem usuário. Se não houver usuário, o registro é salvo em uma tabela de erros com registro do tipo de erro (**Add constants 2**), data e horário da ocorrência (**Get System info 2**), caso contrário, a execução prossegue.
e. ‘**Sort (nm_cidade)**’: passo que ordena os registros em ordem alfabética pelo nome da cidade.
f. ‘**Db lookup (nm_cidade)**’: este passo verifica se a cidade cadastrada para a propriedade no LS existe no Brasil, através da verificação nas tabelas com informações do IBGE. Se existe, é retornado o identificador desta cidade, caso contrário, nenhum valor é retornado para o campo identificador.
g. ‘**Filter rows**’: Se houve erros de digitação ou acentuação no nome da cidade, o registro é salvo em uma tabela de erros com registro do tipo de erro (**Add constants**), data e horário da ocorrência (**Get System info**), caso contrário, a execução prossegue.
h. ‘**Sort (id_cidade) 2**’: passo que ordena os registros em ordem crescente pelo identificador da cidade no banco de metadados.
i. ‘**Merge join 6’**: é realizada a junção do registro da propriedade com as informações adicionais do IBGE através do uso do identificador da cidade como chave.

j. ‘**Select_metadados_ids**’: seleciona apenas os dados de interesse para a inserção na *stage*.
k. ‘**Sort (nk_propriedade_ls)**’: passo que ordena os registros em ordem crescente pelo identificador da propriedade do LS.
l. ‘**insert_stage_transformação_propriedade**’: este passo realiza a inserção dos dados da propriedade e sua respectiva localização, assim como chaves do sistema fonte (LS), na respectiva tabela na stage transformação.

4. **‘transformacao_sistema_produtivo_ls’**: essa é a transformação mais complexa do ETL por exigir um grande número de verificações em grandes quantidades de colunas do sistema LS. A Figura 17 apresenta os passos dessa transformação, que é detalhada na sequência. Devido ao grande número de passos, essa transformação será explicada de uma forma mais genérica e sucinta.

Figura 17: Tarefa de transformação dos dados do sistema produtivo do LS.



Fonte: Autor (2019).

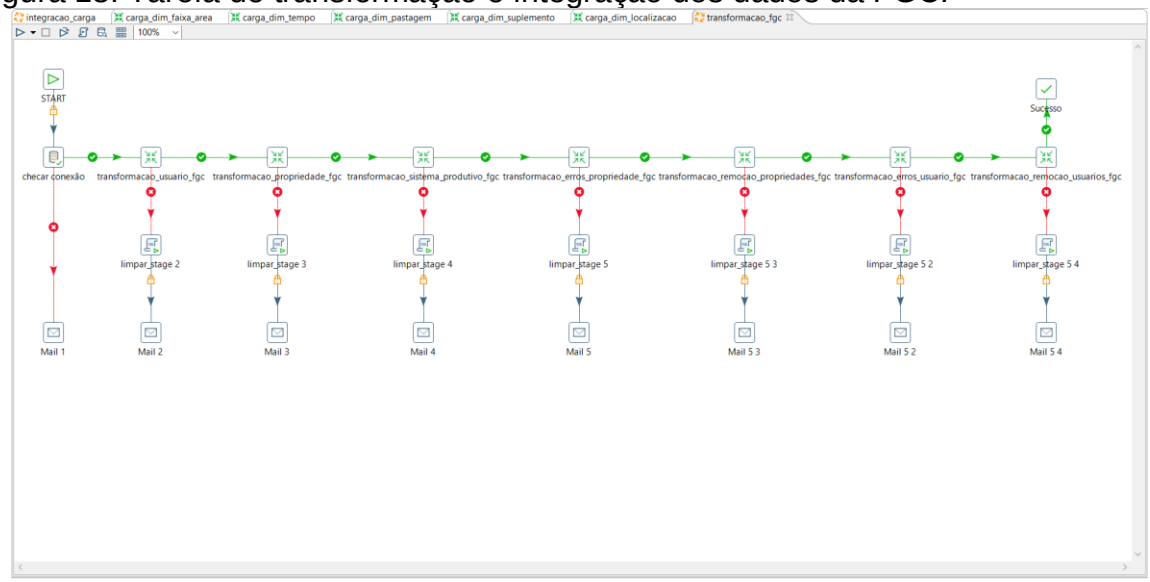
A transformação primeiramente verifica se os registros de sistema produtivo possuem uma propriedade relacionada. Pode ocorrer este caso devido à transformação anterior, onde podem ter sido descartados registros de propriedades. Os registros de sistemas sem propriedade são salvos em uma tabela de erros, com identificação do tipo de erro e horário da ocorrência do erro. Os registros com propriedade relacionada prosseguem em dois fluxos distintos, um para o tratamento de valores numéricos e outro para o tratamento de textos. Para o tratamento dos valores numéricos foram realizadas tarefas como: verificações de tipo de dados, verificação de faixas de valores para todos atributos, prévia geração de chaves substitutas com base na área da propriedade e uso de suplementos para a carga do DW, cálculo do NDT médio para uso na carga (DW) e filtro para determinar se o sistema produtivo utilizou algum suplemento ou não com base no somatório dos NDT. Os valores de texto existentes eram os nomes das pastagens cadastradas por propriedade, assim como suas respectivas composições. Para validar as composições, foi verificado se a soma das composições é 100, e caso seja, o fluxo de execução prossegue normalmente, caso contrário, houve um erro no preenchimento dos dados de composições, e neste caso o registro é salvo em uma tabela de erros com registro do tipo de erro, data e horário

da ocorrência. Para validar as pastagens, foi criado um banco de metadados com informações corretas e atualizadas sobre pastagens e categorias de pastagens para ser utilizado como meio de validação das informações preenchidas ou selecionadas por usuários do LS. Portanto, além de validar a pastagem informada pelo usuário, nesta transformação também é criada uma nova coluna para armazenar a categoria de cada pastagem, informação que será necessária a posteriori no DW. Todos os registros, após verificados, validados e processados, são inseridos na *stage area*, nas tabela de transformação relacionada ao sistema produtivo.

As quatro transformações seguintes da Figura 14 (**transformacao_erros_propriedade_ls**, **transformacao_remocao_propriedade_ls**, **transformacao_erros_usuario_ls**, **transformacao_remocao_usuario_ls**) visam detectar e deletar registros de propriedades e usuários que ficaram sem dados de sistema produtivo. O fato da transformação anterior poder descartar registros de sistemas produtivos não acarreta no descarte de registros das propriedades e usuários associados a este. Este conjunto de quatro transformações, portanto, serve para verificar a integridade referencial dos registros válidos já inseridos. Ainda, antes de deletar os registros com problemas de integridade, os mesmos são salvos nas tabelas de erros associadas ao tipo do registro (usuário, propriedade ou sistema produtivo), com identificação do tipo de erro e horário da ocorrência do problema.

A tarefa de transformação de dados específica do FGC ocorre após o processo de transformação de dados do sistema LS, sendo também um processo de integração dos dados de ambos os sistemas. Tal tarefa pode ser analisada em detalhes na Figura 18. Devido à semelhança com a transformação dos dados do LS, serão abordadas as diferenças adicionais que este processo possui devido à integração dos dados. Os detalhes são discutidos na sequência:

Figura 18: Tarefa de transformação e integração dos dados da FGC.

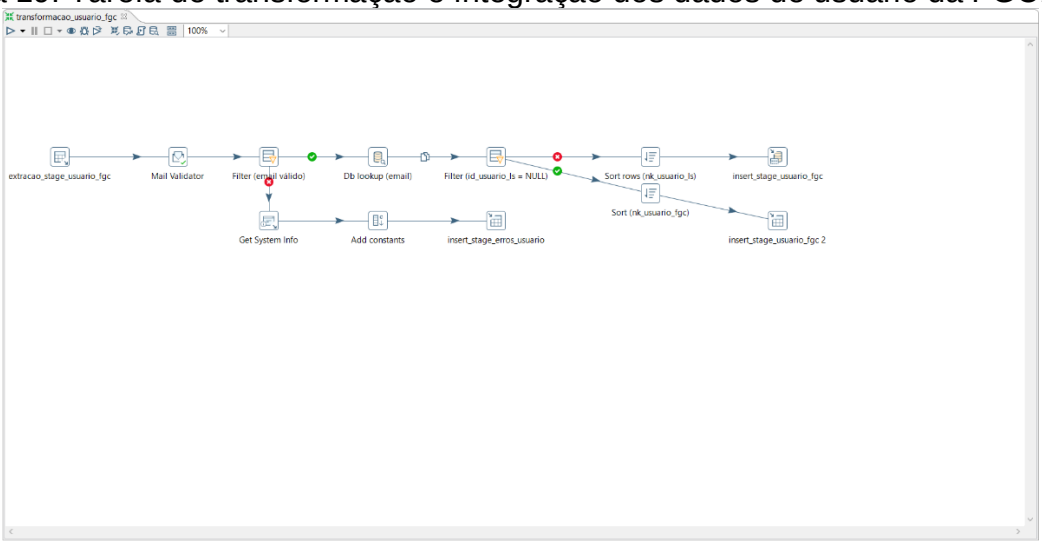


Fonte: Autor (2019).

1. '**transformacao_usuario_fgc**': idêntica a respectiva transformação no LS, porém, antes de realizar a inserção, é verificado se o *e-mail* que será inserido já existe nas tabelas de transformação. Caso exista, é porque há registros do sistema LS com o mesmo *e-mail*, portanto, o mesmo usuário possui registros em ambos sistemas

e a ação realizada é a atualização do registro com a inserção da chave primária do usuário da FGC. Caso não exista, é porque este usuário apenas possui registros na FGC, e a ação realizada é apenas a inserção de ambas chave primária e *e-mail* do usuário (neste caso, o dado armazenado no campo de chave primária do usuário do LS será nulo). Cabe mencionar que na criação das tabelas da stage relacionadas às transformações, não foram implementadas restrições de chave primária e estrangeira em nenhuma coluna. O processo fluxo pode ser visualizado na Figura 19.

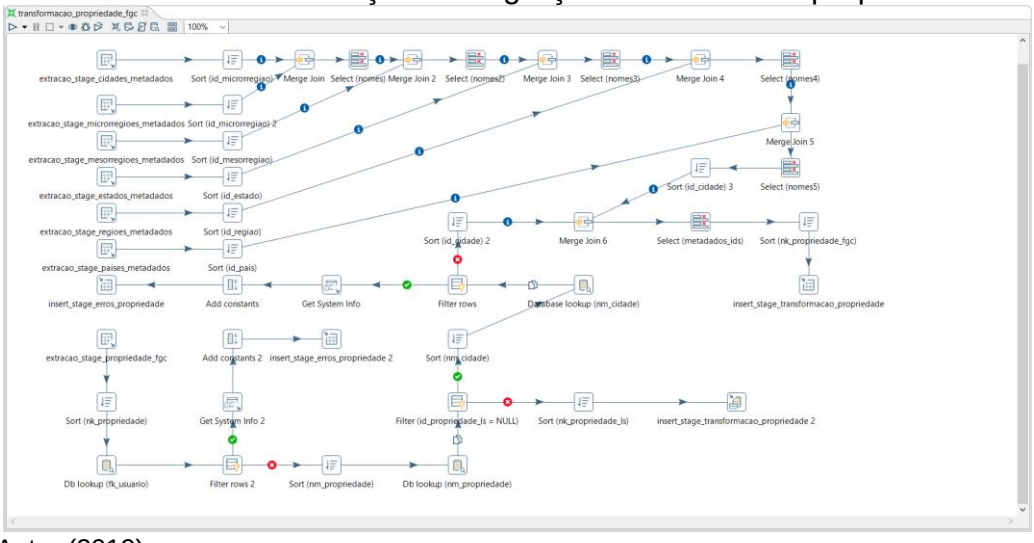
Figura 19: Tarefa de transformação e integração dos dados de usuário da FGC.



Fonte: Autor (2019).

2. '**transformacao_propriedade_fgc**': idêntica a respectiva transformação no LS, porém, antes de realizar a inserção, é verificado se o nome da propriedade que será inserida já existe nas tabelas de transformação. Em caso afirmativo, apenas serão inseridas as chaves de usuário (estrangeira) e da propriedade (primária) da FGC na tabela de propriedades da *stage area*. Caso contrário, será realizado o mesmo processo descrito na transformação '**transformacao_propriedade_ls'**. A transformação pode ser visualizada na Figura 20.

Figura 20: Tarefa de transformação e integração dos dados de propriedade.

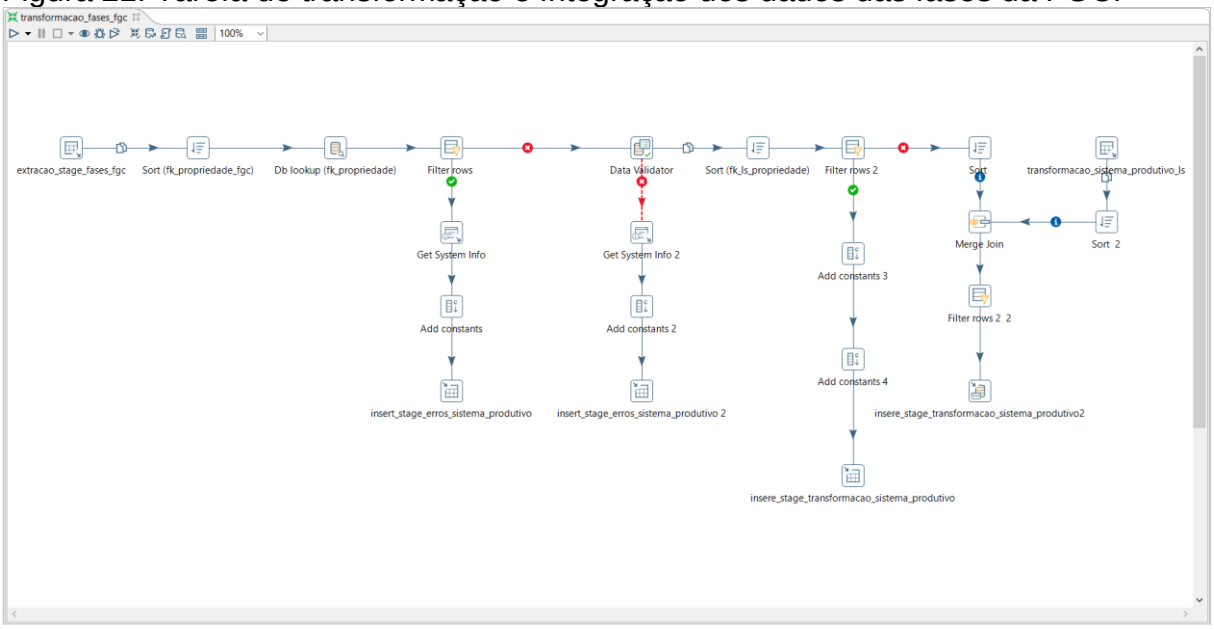


Fonte: Autor (2019).

3. ‘**transformacao_sistema_produtivo_fgc**’: semelhante a respectiva transformação no LS, porém, com as informações das fases integradas na tabela de sistema produtivo da *stage area*. São realizadas verificações de tipos dos dados, se os mesmos estão em faixas de valores aceitáveis e consistência de valores redundantes. A Figura 21 apresenta o fluxo desta transformação.

As quatro transformações seguintes da Figura 18 (**transformacao_erros_propriedade_fgc**, **transformacao_remocao_propriedade_fgc**, **transformacao_erros_usuario_fgc**, **transformacao_remocao_usuario_fgc**) visam detectar e deletar registros de propriedades e usuários que ficaram sem dados de sistema produtivo na *stage area,* relativa ao processo de transformação de dados da FGC. São verificados se os registros das propriedades nos sistemas produtivos existem nas tabelas de propriedades, através das chaves primárias e estrangeiras, e se os registros de usuários nas propriedades existem nas tabelas de usuários, por meio de uma verificação similar à anterior. Avaliações adicionais são necessárias para verificar se os registros não possuem dados do sistema LS já integrados. Neste caso, os dados não são deletados.

Figura 21: Tarefa de transformação e integração dos dados das fases da FGC.



Fonte: Autor (2019).

4.3 Processo de carga dos dados

O processo de carga do ETL inicia-se a partir da carga dos dados das tabelas de transformação, local onde estão os dados tratados, para as dimensões. Após a carga das dimensões, são realizadas as cargas das tabelas fato. A Figura 22 apresenta a tarefa de carga em um nível maior de abstração, assim como o fluxo das atividades deste processo.

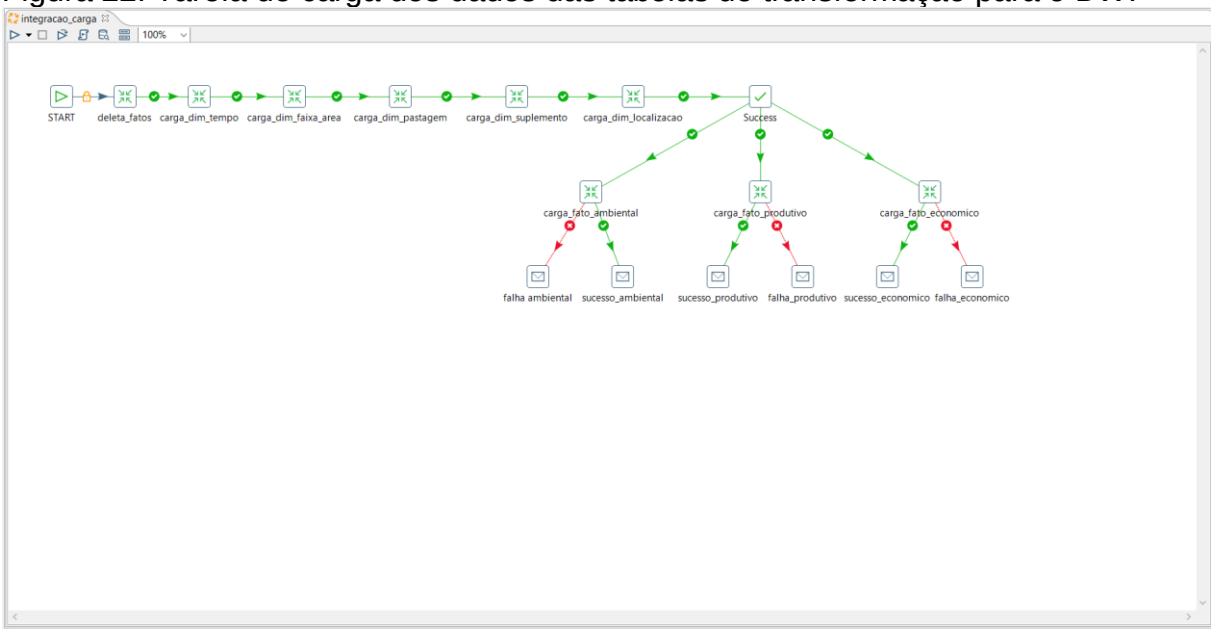
1. ‘**deleta_fatos**’: as tabelas fato são sempre inicialmente truncadas, pois optou-se pela abordagem de carga completa.
2. ‘**carga_dim_tempo**’: a Figura 23 ilustra o fluxo desta transformação. Simplesmente é realizada a inserção dos anos nas dimensões. Outra forma

possível e a mais convencional de realizar este processo é inserir de antemão o conjunto necessário de informações temporais. Como é necessário apenas o ano, optou-se por realizar a verificação dos registros e, caso exista algum ano ainda não cadastrado na dimensão, o mesmo será inserido.

3. '**carga_dim_faixa_area**': a Figura 24 ilustra o fluxo desta transformação. São calculadas as faixas de área com base nos valores das áreas dos sistemas produtivos. O resultado deste processo é um novo campo do tipo *string* que contém a faixa de área da propriedade em questão ('**number range (area_ls -> faixa)'**). Por fim, as faixas são ordenadas de forma crescente (**'sort (range_ls)'**), valores duplicados destas faixas são eliminados ('**Unique rows**') e os restantes são inseridos na dimensão **dim_faixa_area**.
4. '**carga_dim_pastagem**': a Figura 25 ilustra o fluxo desta transformação. Processamentos são necessários para realizar a separação das *strings* que contém o nome das pastagens e respectivas composições ('**split_pastagem'** e '**split_composicao'**) para tarefas como inserção do símbolo '%' na composição ('**Add '%''**), concatenação de pastagens e respectivas composições ('**Concat Fields**'), eliminação de dados de pastagens e composições duplicados (**'Unique rows'**) e inserção destas informações na **dim_pastagem**.
5. '**carga_dim_suplemento**': a Figura 26 ilustra o fluxo desta transformação. Processamentos são necessários para gerar a informação que será inserida na dimensão de suplementos. O símbolo de porcentagem '%' ('**Add '%''**) é concatenado com a informação de NDT médio do suplemento ('**Calculator (ndt_medio + '%')'**), quando disponível. Em casos desta informação não estar disponível, é gerada a *string* '**N.I**', que significa **Não Informado ('Add constants (ndt_medio)')**. Por fim, ocorre a eliminação de dados de NDT médio duplicados ('**Unique rows**') e a inserção destas informações na **dim_suplemento**.
6. '**carga_dim_localizacao**': a Figura 27 ilustra o fluxo desta transformação. São extraídas informações tanto das tabelas de propriedade como das tabelas de usuário, relacionadas ao processo de transformação de dados. Estas informações são unificadas por meio da chave primária da tabela de usuário com as chaves estrangeiras de usuário na tabela de propriedades. A verificação da chave primária da tabela de propriedade, relacionada ao sistema LS, é realizada antes da inserção. Por fim, estes registros são ordenados e inseridos na **dim_localizacao**.
7. '**carga_fato_produtivo_economico_ambiental**': a Figura 28 ilustra o fluxo desta transformação. A Figura 28 foi subdividida em X conjuntos de passos com propósitos específicos para a realização da carga da tabela. No passo 1 é realizada uma série de cálculos de métricas produtivas da pecuária de corte que não estão disponibilizadas previamente nos sistemas fontes, como número total de animais, peso médio total e arrobas. No passo 2 é realizada uma série de cálculos de métricas econômicas da pecuária de corte que não estão disponibilizadas previamente nos sistemas fontes, como custo total, custos por quilos de peso vivo, custos por área, custos por cabeça, custo por arroba e lucro total. No passo 3 é simplesmente realizada uma verificação da emissão total do sistema produtivo analisado, e caso seja nulo, o valor de emissão zero é atribuído ao mesmo. No passo 4 é realizada a busca pela chave substituta na dimensão de tempo. No passo 5 é realizada a busca pela chave substituta na dimensão de faixas de área. No passo 6 é realizada a busca pela chave substituta na dimensão de pastagens, processo que envolve um número maior de passos devido à manipulação de *strings*. No passo 7 é realizada a busca pela chave substituta na dimensão de suplementos. No passo 8 é realizada a busca pela chave substituta na dimensão

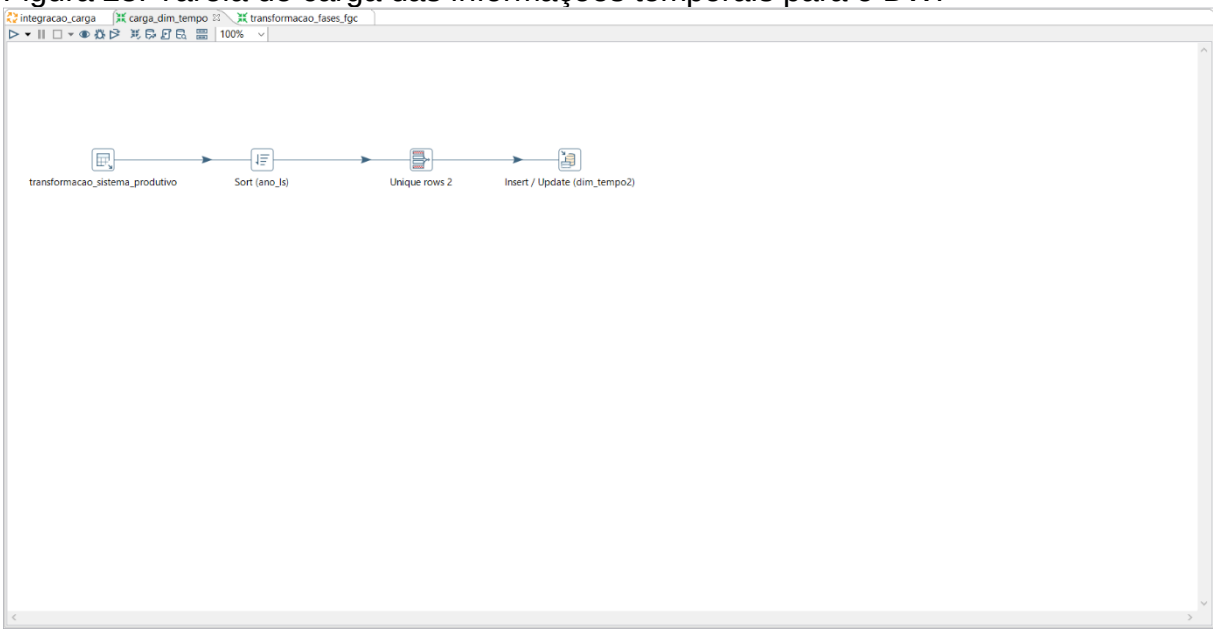
de localização. Por fim, o passo 9 ordena, em ordem crescente, os registros de acordo com as chaves substitutas, faz verificação se o lucro for negativo para atribuir valor zero e insere os dados de interesse na tabela **fato_produtivo_economico_ambiental**.

Figura 22: Tarefa de carga dos dados das tabelas de transformação para o DW.



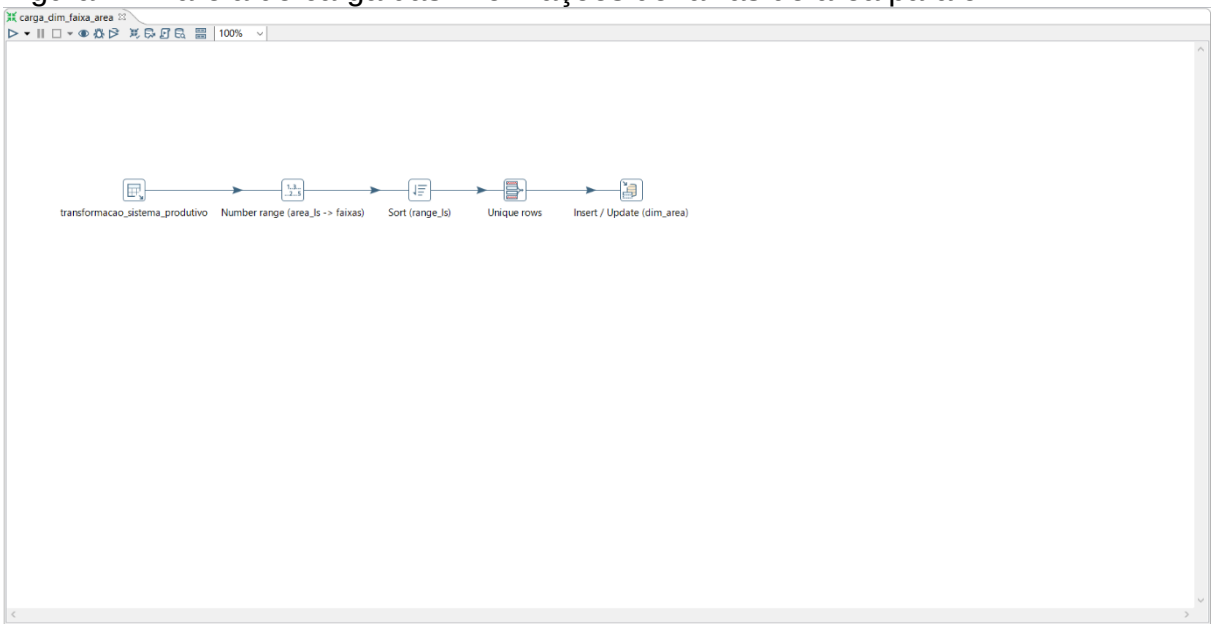
Fonte: Autor (2019).

Figura 23: Tarefa de carga das informações temporais para o DW.



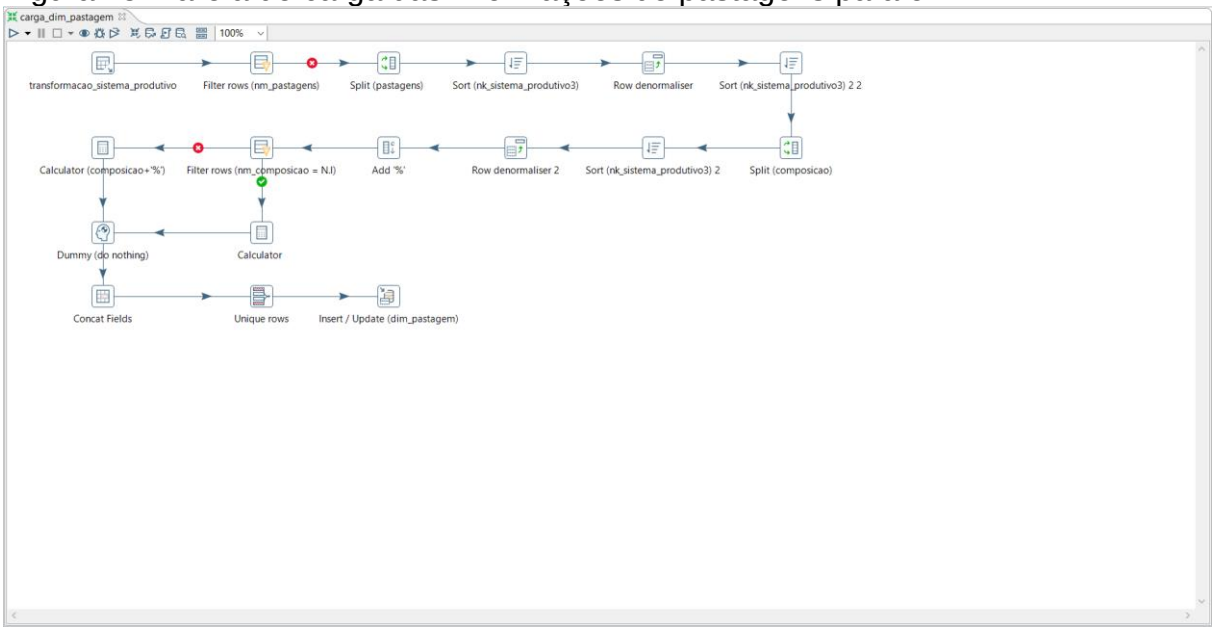
Fonte: Autor (2019).

Figura 24: Tarefa de carga das informações de faixas de área para o DW.



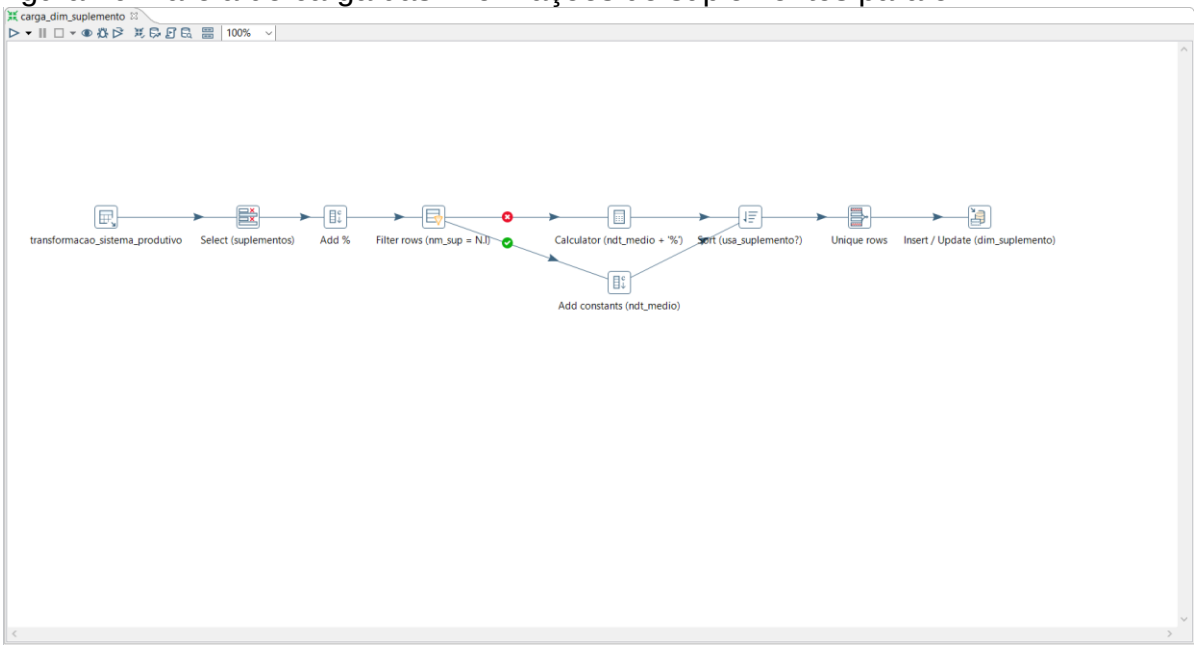
Fonte: Autor (2019).

Figura 25: Tarefa de carga das informações de pastagens para o DW.



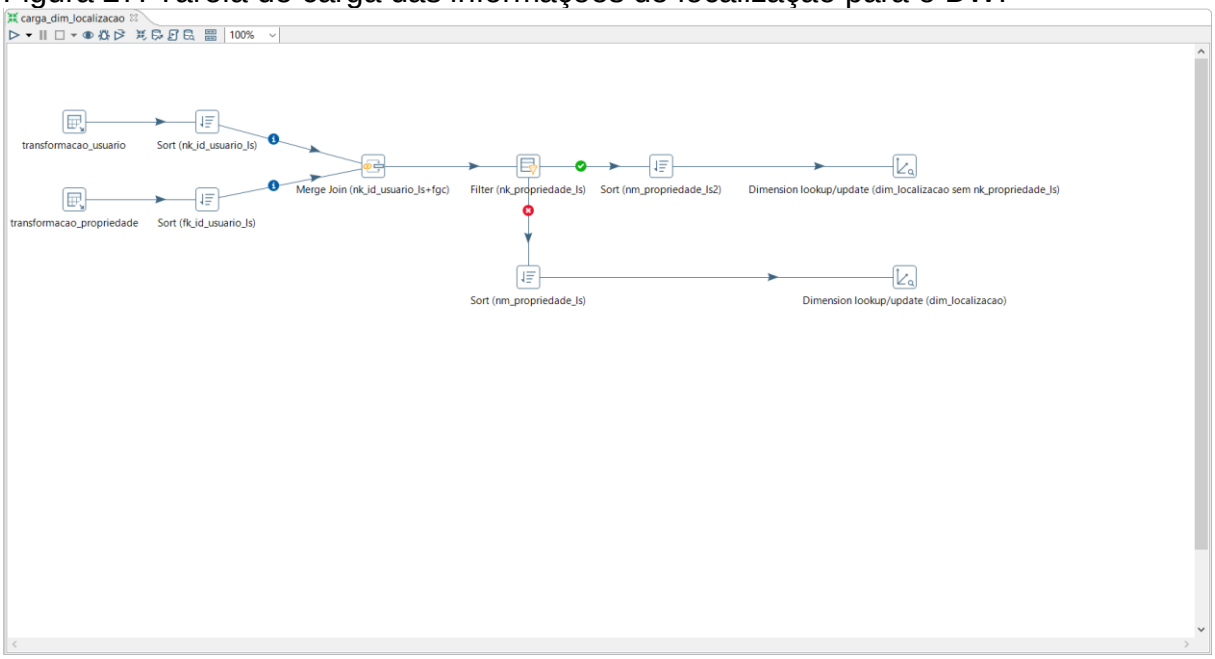
Fonte: Autor (2019).

Figura 26: Tarefa de carga das informações de suplementos para o DW.



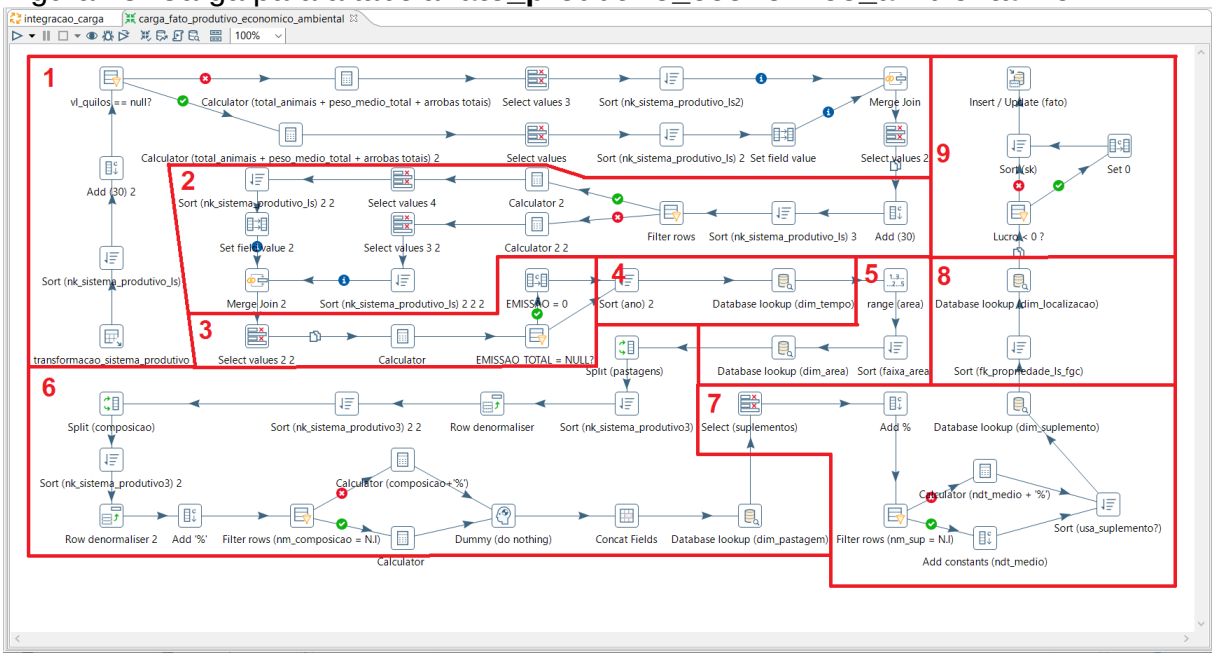
Fonte: Autor (2019).

Figura 27: Tarefa de carga das informações de localização para o DW.



Fonte: Autor (2019).

Figura 28: Carga para a tabela **fato_produtivo_economico_ambiental** no DW.



Fonte: Autor (2019).

4.4 Execução das tarefas do ETL

Após a conclusão da implementação do processo de ETL, é possível realizar a execução das tarefas de três formas:

1. Execução manual utilizando o próprio *Pentaho Data Integration*, módulo *spoon*.

    a. O primeiro passo é localizar o diretório dos arquivos do *Pentaho Data Integration*.
    b. Executar o módulo *spoon* (spoon.bat no Windows e spoon.sh no Linux).
    c. Localizar o diretório das tarefas do ETL para abrir as tarefas.
    d. Executar as seguintes tarefas, nesta ordem: **Extração/extracao_ETL.kjb**, **Transformação e Integração/transformacao_ETL.kjb** e **Carga/integração_carga.kjb.**

2. Execução manual utilizando o *prompt* de comando do Sistema Operacional.

O primeiro passo é abrir o *prompt* de comando do Sistema Operacional a partir do diretório de arquivos do *Pentaho Data Integration*. Por fim, após abrir o *prompt*, é necessário digitar os seguintes comandos:

No Windows:

Kitchen.bat /file:"DIRETÓRIO_DO_ETL/jobs e steps/extração/extracao_ETL.kjb”
Kitchen.bat /file:"DIRETÓRIO_DO_ETL/jobs e steps/Transformação e Integração/transformacao_ETL.kjb”
Kitchen.bat /file:"DIRETÓRIO_DO_ETL/jobs e steps/Carga/integração_carga.kjb”

No Linux:

sh kitchen.sh /file:"DIRETÓRIO_DO_ETL/jobs e steps/extração/extracao_ETL.kjb”

```
sh kitchen.sh /file:"DIRETÓRIO_DO_ETL/jobs e steps/Transformação e
Integração/transformacao_ETL.kjb”
sh kitchen.sh /file:"DIRETÓRIO_DO_ETL/jobs e steps/Carga/integração_carga.kjb”
```

3. Execução automática através do agendamento de tarefas.

A versão *Community Edition* do PDI não possui o agendador de tarefas embutido. Portanto, torna-se necessário utilizar os recursos do Sistema Operacional para realizar esta tarefa. No Windows 10, são necessários os seguintes procedimentos:

1. Criar três arquivos do tipo .bat (extracao.bat, transformacao.bat e carga.bat), um para cada processo do ETL, no diretório principal das tarefas do ETL.
2. Nos arquivos .bat, escrever os seguintes comandos (de acordo com o processo do ETL):

I. Extracao.bat

```
@echo off
TITLE Extracao
SET currentdir=%~dp0
SET kitchen=C:\Users\Dell\Desktop\pdi-ce-8.1.0.0-365\Kitchen.bat
SET logfile="%currentdir%\Logs\log1.txt"
echo. >> %logfile%
"%kitchen%" /file:"%currentdir%/Extração/extracao_ETL.kjb" /level:Basic >> %logfile%
```

II. Transformacao.bat

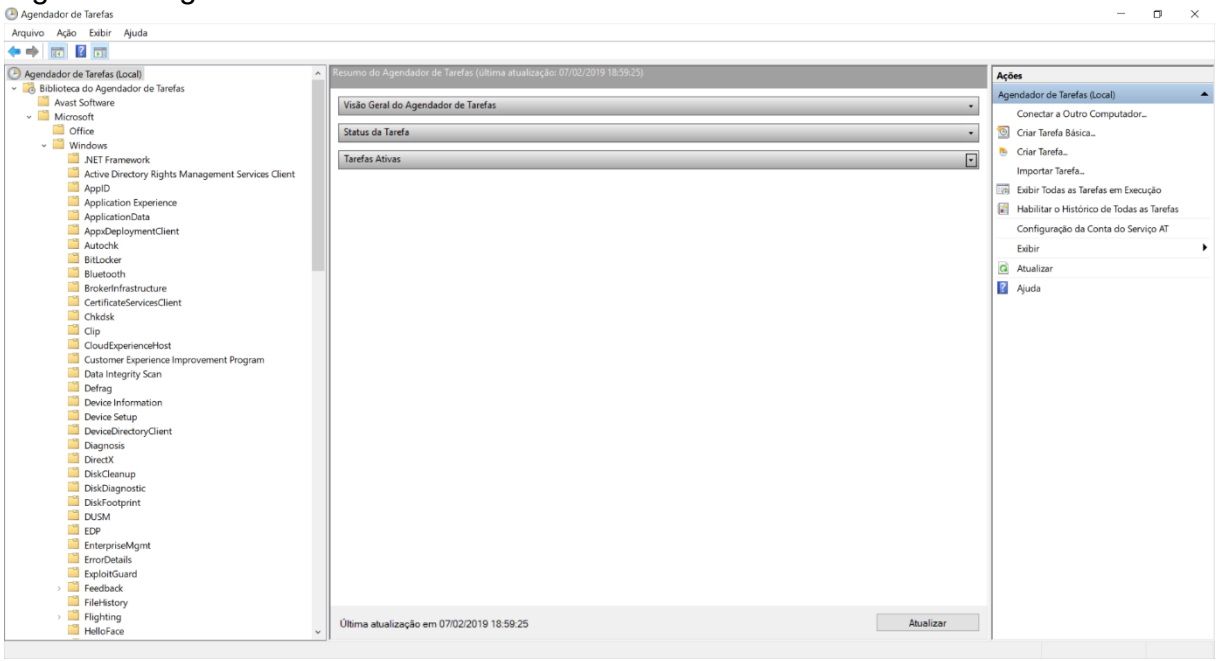
```
@echo off
TITLE Transformacao
SET currentdir=%~dp0
SET kitchen=C:\Users\Dell\Desktop\pdi-ce-8.1.0.0-365\Kitchen.bat
SET logfile="%currentdir%\Logs\log2.txt"
echo. >> %logfile%
"%kitchen%" /file:"%currentdir%/Extração/transformacao_ETL.kjb" /level:Basic >>
%logfile%
```

III. Carga.bat

```
@echo off
TITLE Carga
SET currentdir=%~dp0
SET kitchen=C:\Users\Dell\Desktop\pdi-ce-8.1.0.0-365\Kitchen.bat
SET logfile="%currentdir%\Logs\log3.txt"
echo. >> %logfile%
"%kitchen%" /file:"%currentdir%/Extração/integracao_carga.kjb" /level:Basic >>
%logfile%
```

3. Por fim, basta utilizar o agendador de tarefas do Sistema Operacional Windows 10, conforme Figura 29.

Figura 29: Agendador de tarefas do Windows 10.

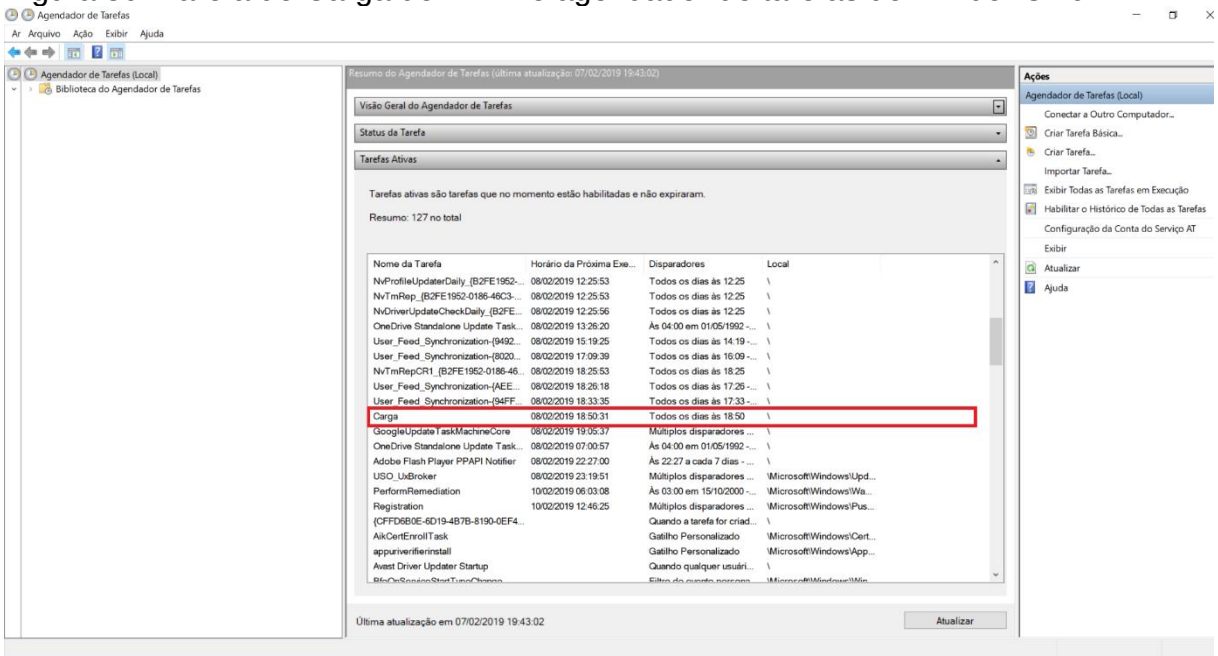


Fonte: Autor (2019).

Passos: clicar em Criar Tarefa..., na aba Geral é necessário dar um nome para a Tarefa, na aba Disparadores é possível definir a frequência, dia e horário da Tarefa, na aba Ações é possível definir o caminho do script que deverá ser executado e na aba Configurações é possível alterar outras configurações e detalhes da execução da tarefa.

Seguindo estes passos, a nova Tarefa aparecerá na lista de tarefas ativas do Sistema Operacional e o processo de agendamento do ETL é concluído (Figura 30).

Figura 30: Tarefa de Carga do ETL no agendador de tarefas do Windows 10.



Fonte: Autor (2019).

Por fim, abaixo é apresentado o script SQL utilizado para a criação das tabelas pertencentes à *stage area*.

4.5 Script SQL para criação da *Stage Area*

```
-- Stage Area --
-- Banco de dados - PostgreSQL --
-- Última edição: 01/02/2019 --

CREATE SCHEMA stage_area AUTHORIZATION postgres;

DROP TABLE IF EXISTS stage_area.st_extracao_usuario_fgc;
CREATE TABLE stage_area.st_extracao_usuario_fgc
(
	nk_id_usuario_fgc INTEGER,
	nm_login_fgc VARCHAR(50),
	nm_senha_fgc VARCHAR(256),
	nm_email_fgc VARCHAR(50),
	CONSTRAINT usuario_fgc_pk PRIMARY KEY (nk_id_usuario_fgc)
);

DROP TABLE IF EXISTS stage_area.st_extracao_propriedade_fgc;
CREATE TABLE stage_area.st__extracao_propriedade_fgc
(
	fk_id_usuario_fgc INTEGER,
	nk_propriedade_fgc INTEGER,
	nm_cidade_fgc VARCHAR(30),
	nm_cd_unidade_federativa_fgc VARCHAR(2),
	nm_propriedade_fgc VARCHAR(30),
	CONSTRAINT propriedade_fgc_pk PRIMARY KEY (nk_propriedade_fgc),
	CONSTRAINT usuario_fgc_fk FOREIGN KEY (fk_id_usuario_fgc) REFERENCES
stage_area.st_extracao_usuario_fgc (nk_id_usuario_fgc) MATCH SIMPLE
);

DROP TABLE IF EXISTS stage_area.st_extracao_fases_fgc;
CREATE TABLE stage_area.st_extracao_fases_fgc
(
	fk_propriedade_fgc INTEGER,
	nr_animais_vacas_fgc INTEGER,
	nr_animais_novilhos_0_12_fgc INTEGER,
	nr_animais_novilhos_13_24_fgc  INTEGER,
	nr_animais_novilhos_25_36_fgc  INTEGER,
	nr_animais_bezerros_bezerras_fgc  INTEGER,
	nr_animais_novilhas_0_12_fgc  INTEGER,
	nr_animais_novilhas_13_24_fgc  INTEGER,
	nr_animais_novilhas_25_36_fgc  INTEGER,
	nr_animais_touros_fgc INTEGER,
	vl_area_fgc REAL,
	vl_receita_anual_fgc NUMERIC,
	vl_quilos_peso_vivo_fgc NUMERIC,
	vl_custos_medicamentos_fgc NUMERIC,
	vl_custos_mao_de_obra_fgc NUMERIC,
	vl_custos_maquinas_fgc NUMERIC,
	vl_custos_pastagens_fgc NUMERIC,
	vl_custos_arrendamento_fgc NUMERIC,
	vl_custos_outros_fgc NUMERIC,
	nr_ano_fgc INTEGER,
```

```
	CONSTRAINT propriedade_fgc_fk FOREIGN KEY (fk_propriedade_fgc) REFERENCES
stage_area.st_extracao_propriedade_fgc (nk_propriedade_fgc) MATCH SIMPLE
);

DROP TABLE IF EXISTS stage_area.st_extracao_usuario_ls;
CREATE TABLE stage_area.st_extracao_usuario_ls
(
	nk_id_usuario_ls INTEGER,
	nm_login_ls VARCHAR(50),
	nm_senha_ls VARCHAR(256),
	nm_email_ls VARCHAR(50),
	CONSTRAINT usuario_ls_pk PRIMARY KEY (nk_id_usuario_ls)
);

DROP TABLE IF EXISTS stage_area.st_extracao_propriedade_ls;
CREATE TABLE stage_area.st_extracao_propriedade_ls
(
	fk_id_usuario_ls INTEGER,
	nk_propriedade_ls INTEGER,
	nm_cidade_ls VARCHAR(30),
	nm_cd_unidade_federativa_ls VARCHAR(2),
	nm_propriedade_ls VARCHAR(30),
	CONSTRAINT propriedade_ls_pk PRIMARY KEY (nk_propriedade_ls),
	CONSTRAINT usuario_ls_fk FOREIGN KEY (fk_id_usuario_ls) REFERENCES
stage_area.st_extracao_usuario_ls (nk_id_usuario_ls) MATCH SIMPLE
);

DROP TABLE IF EXISTS stage_area.st_extracao_sistema_produtivo_ls;
CREATE TABLE stage_area.st_extracao_sistema_produtivo_ls
(
	nk_sistema_produtivo_ls INTEGER,
	fk_propriedade_ls INTEGER,
	per_desmame_ls REAL,
	per_mortalidade_ls REAL,
	per_touros_ls REAL,
	per_descarte_vacas_ls REAL,
	per_descarte_touros_ls REAL,
	nr_idade_venda_ls INTEGER,
	nr_idade_entoure_ls INTEGER,
	vl_unidade_animal_ls REAL,
	vl_lotacao_animal_ls REAL,
	vl_area_ls REAL,
	vl_fertilizantes_ls REAL,
	vl_produtividade_ls REAL,
	vl_emissao_por_produtividade_ls REAL,
	vl_emissao_novilhos_0_12_ls REAL,
	vl_emissao_novilhos_13_24_ls REAL,
	vl_emissao_novilhos_25_36_ls REAL,
	vl_emissao_bezerros_bezerras_ls REAL,
	vl_emissao_novilhas_0_12_ls REAL,
	vl_emissao_novilhas_13_24_ls REAL,
	vl_emissao_novilhas_25_36_ls REAL,
	vl_emissao_vacas_ls REAL,
	vl_emissao_touros_ls REAL,
	vl_peso_medio_novilhos_0_12_ls REAL,
	vl_peso_medio_novilhos_13_24_ls REAL,
	vl_peso_medio_novilhos_25_36_ls REAL,
	vl_peso_medio_bezerros_bezerras_ls REAL,
	vl_peso_medio_novilhas_0_12_ls REAL,
	vl_peso_medio_novilhas_13_24_ls REAL,
```

```
		vl_peso_medio_novilhas_25_36_ls REAL,
		vl_peso_medio_vacas_ls REAL,
		vl_peso_medio_touros_ls REAL,
		vl_peso_venda_novilhos_0_12_ls REAL,
		vl_peso_venda_novilhos_13_24_ls REAL,
		vl_peso_venda_novilhos_25_36_ls REAL,
		vl_peso_venda_bezerros_bezerras_ls REAL,
		vl_peso_venda_novilhas_0_12_ls REAL,
		vl_peso_venda_novilhas_13_24_ls REAL,
		vl_peso_venda_novilhas_25_36_ls REAL,
		vl_peso_venda_vacas_ls REAL,
		vl_peso_venda_touros_ls REAL,
		vl_ndt_suplemento_novilhos_0_12_ls REAL,
		vl_ndt_suplemento_novilhos_13_24_ls REAL,
		vl_ndt_suplemento_novilhos_25_36_ls REAL,
		vl_ndt_suplemento_bezerros_bezerras_ls REAL,
		vl_ndt_suplemento_novilhas_0_12_ls REAL,
		vl_ndt_suplemento_novilhas_13_24_ls REAL,
		vl_ndt_suplemento_novilhas_25_36_ls REAL,
		vl_ndt_suplemento_vacas_ls REAL,
		vl_ndt_suplemento_touros_ls REAL,
		nm_todas_pastagens_ls VARCHAR(100),
		nm_composicoes_pastagens_ls VARCHAR(50),
		nr_ano_ls INTEGER,
		CONSTRAINT sistema_produtivo_ls_pk PRIMARY KEY (nk_sistema_produtivo_ls),
		CONSTRAINT propriedade_ls_fk FOREIGN KEY (fk_propriedade_ls) REFERENCES
stage_area.st_extracao_propriedade_ls (nk_propriedade_ls) MATCH SIMPLE
);

DROP TABLE IF EXISTS stage_area.st_transformacao_usuario;
CREATE TABLE stage_area.st_transformacao_usuario
(
		nk_id_usuario_ls INTEGER,
		nk_id_usuario_fgc INTEGER,
		nm_email_ls VARCHAR(50)
);

DROP TABLE IF EXISTS stage_area.st_transformacao_propriedade;
CREATE TABLE stage_area.st_transformacao_propriedade
(
		fk_id_usuario_ls INTEGER,
		nk_propriedade_ls INTEGER,
		fk_id_usuario_fgc INTEGER,
		nk_propriedade_fgc INTEGER,
		nm_cidade_ls VARCHAR(30),
		nm_cd_unidade_federativa_ls VARCHAR(2),
		nm_propriedade_ls VARCHAR(30),
		nm_microrregiao_ls character varying(30),
		nm_mesorregiao_ls character varying(30),
		nm_unidade_federativa_ls character varying(30),
		nm_regiao_ls character varying(30),
		nm_pais_ls character varying(30)
);

DROP TABLE IF EXISTS stage_area.st_transformacao_sistema_produtivo;
CREATE TABLE stage_area.st_transformacao_sistema_produtivo
(
		nk_sistema_produtivo_ls INTEGER,
		fk_propriedade_ls INTEGER,
		per_desmame_ls REAL,
```

```
per_mortalidade_ls REAL,
per_touros_ls REAL,
per_descarte_vacas_ls REAL,
per_descarte_touros_ls REAL,
nr_idade_venda_ls INTEGER,
nr_idade_entoure_ls INTEGER,
vl_unidade_animal_ls REAL,
vl_lotacao_animal_ls REAL,
vl_area_ls REAL,
vl_fertilizantes_ls REAL,
vl_produtividade_ls REAL,
vl_emissao_por_produtividade_ls REAL,
vl_emissao_novilhos_0_12_ls REAL,
vl_emissao_novilhos_13_24_ls REAL,
vl_emissao_novilhos_25_36_ls REAL,
vl_emissao_bezerros_bezerras_ls REAL,
vl_emissao_novilhas_0_12_ls REAL,
vl_emissao_novilhas_13_24_ls REAL,
vl_emissao_novilhas_25_36_ls REAL,
vl_emissao_vacas_ls REAL,
vl_emissao_touros_ls REAL,
vl_peso_medio_novilhos_0_12_ls REAL,
vl_peso_medio_novilhos_13_24_ls REAL,
vl_peso_medio_novilhos_25_36_ls REAL,
vl_peso_medio_bezerros_bezerras_ls REAL,
vl_peso_medio_novilhas_0_12_ls REAL,
vl_peso_medio_novilhas_13_24_ls REAL,
vl_peso_medio_novilhas_25_36_ls REAL,
vl_peso_medio_vacas_ls REAL,
vl_peso_medio_touros_ls REAL,
vl_peso_venda_novilhos_0_12_ls REAL,
vl_peso_venda_novilhos_13_24_ls REAL,
vl_peso_venda_novilhos_25_36_ls REAL,
vl_peso_venda_bezerros_bezerras_ls REAL,
vl_peso_venda_novilhas_0_12_ls REAL,
vl_peso_venda_novilhas_13_24_ls REAL,
vl_peso_venda_novilhas_25_36_ls REAL,
vl_peso_venda_vacas_ls REAL,
vl_peso_venda_touros_ls REAL,
vl_ndt_suplemento_novilhos_0_12_ls REAL,
vl_ndt_suplemento_novilhos_13_24_ls REAL,
vl_ndt_suplemento_novilhos_25_36_ls REAL,
vl_ndt_suplemento_bezerros_bezerras_ls REAL,
vl_ndt_suplemento_novilhas_0_12_ls REAL,
vl_ndt_suplemento_novilhas_13_24_ls REAL,
vl_ndt_suplemento_novilhas_25_36_ls REAL,
vl_ndt_suplemento_vacas_ls REAL,
vl_ndt_suplemento_touros_ls REAL,
nm_categoria_pastagens_ls VARCHAR(50),
nm_todas_pastagens_ls VARCHAR(100),
nm_composicoes_pastagens_ls VARCHAR(50),
sk_faixa_area_ls INTEGER,
nm_usa_suplemento_ls VARCHAR(3),
vl_ndt_medio_suplementos_ls REAL,
nr_ano_ls INTEGER,
fk_propriedade_fgc INTEGER,
nr_animais_vacas_fgc INTEGER,
nr_animais_novilhos_0_12_fgc INTEGER,
nr_animais_novilhos_13_24_fgc  INTEGER,
nr_animais_novilhos_25_36_fgc  INTEGER,
```

```
	nr_animais_bezerros_bezerras_fgc  INTEGER,
	nr_animais_novilhas_0_12_fgc  INTEGER,
	nr_animais_novilhas_13_24_fgc  INTEGER,
	nr_animais_novilhas_25_36_fgc  INTEGER,
	nr_animais_touros_fgc INTEGER,
	vl_area_fgc REAL,
	vl_receita_anual_fgc NUMERIC,
	vl_quilos_peso_vivo_fgc NUMERIC,
	vl_custos_medicamentos_fgc NUMERIC,
	vl_custos_mao_de_obra_fgc NUMERIC,
	vl_custos_maquinas_fgc NUMERIC,
	vl_custos_pastagens_fgc NUMERIC,
	vl_custos_arrendamento_fgc NUMERIC,
	vl_custos_outros_fgc NUMERIC,
	nr_ano_fgc INTEGER,
	nm_pastagem_predominante character varying(30),
	nm_tipo_pastagem_predominante character varying(30)
);

DROP TABLE IF EXISTS stage_area.st_erros_usuario;
CREATE TABLE stage_area.st_erros_usuario
(
	nk_id_usuario_ls INTEGER,
	nk_id_usuario_fgc INTEGER,
	nm_email_ls VARCHAR(50),
	id_erro INTEGER,
	data_carga TIMESTAMP WITHOUT TIME ZONE
);

DROP TABLE IF EXISTS stage_area.st_erros_propriedade;
CREATE TABLE stage_area.st_erros_propriedade
(
	fk_id_usuario_ls INTEGER,
	nk_propriedade_ls INTEGER,
	fk_id_usuario_fgc INTEGER,
	nk_propriedade_fgc INTEGER,
	nm_cidade_ls VARCHAR(30),
	nm_cd_unidade_federativa_ls VARCHAR(2),
	nm_propriedade_ls VARCHAR(30),
	nm_microrregiao_ls character varying(30),
	nm_mesorregiao_ls character varying(30),
	nm_unidade_federativa_ls character varying(30),
	nm_regiao_ls character varying(30),
	nm_pais_ls character varying(30),
	id_erro INTEGER,
	data_carga TIMESTAMP WITHOUT TIME ZONE
);

DROP TABLE IF EXISTS stage_area.st_erros_sistema_produtivo;
CREATE TABLE stage_area.st_erros_sistema_produtivo
(
	nk_sistema_produtivo_ls INTEGER,
	fk_propriedade_ls INTEGER,
	per_desmame_ls REAL,
	per_mortalidade_ls REAL,
	per_touros_ls REAL,
	per_descarte_vacas_ls REAL,
	per_descarte_touros_ls REAL,
	nr_idade_venda_ls INTEGER,
	nr_idade_entoure_ls INTEGER,
```

```
vl_unidade_animal_ls REAL,
vl_lotacao_animal_ls REAL,
vl_area_ls REAL,
vl_fertilizantes_ls REAL,
vl_produtividade_ls REAL,
vl_emissao_por_produtividade_ls REAL,
vl_emissao_novilhos_0_12_ls REAL,
vl_emissao_novilhos_13_24_ls REAL,
vl_emissao_novilhos_25_36_ls REAL,
vl_emissao_bezerros_bezerras_ls REAL,
vl_emissao_novilhas_0_12_ls REAL,
vl_emissao_novilhas_13_24_ls REAL,
vl_emissao_novilhas_25_36_ls REAL,
vl_emissao_vacas_ls REAL,
vl_emissao_touros_ls REAL,
vl_peso_medio_novilhos_0_12_ls REAL,
vl_peso_medio_novilhos_13_24_ls REAL,
vl_peso_medio_novilhos_25_36_ls REAL,
vl_peso_medio_bezerros_bezerras_ls REAL,
vl_peso_medio_novilhas_0_12_ls REAL,
vl_peso_medio_novilhas_13_24_ls REAL,
vl_peso_medio_novilhas_25_36_ls REAL,
vl_peso_medio_vacas_ls REAL,
vl_peso_medio_touros_ls REAL,
vl_peso_venda_novilhos_0_12_ls REAL,
vl_peso_venda_novilhos_13_24_ls REAL,
vl_peso_venda_novilhos_25_36_ls REAL,
vl_peso_venda_bezerros_bezerras_ls REAL,
vl_peso_venda_novilhas_0_12_ls REAL,
vl_peso_venda_novilhas_13_24_ls REAL,
vl_peso_venda_novilhas_25_36_ls REAL,
vl_peso_venda_vacas_ls REAL,
vl_peso_venda_touros_ls REAL,
vl_ndt_suplemento_novilhos_0_12_ls REAL,
vl_ndt_suplemento_novilhos_13_24_ls REAL,
vl_ndt_suplemento_novilhos_25_36_ls REAL,
vl_ndt_suplemento_bezerros_bezerras_ls REAL,
vl_ndt_suplemento_novilhas_0_12_ls REAL,
vl_ndt_suplemento_novilhas_13_24_ls REAL,
vl_ndt_suplemento_novilhas_25_36_ls REAL,
vl_ndt_suplemento_vacas_ls REAL,
vl_ndt_suplemento_touros_ls REAL,
nm_categoria_pastagens_ls VARCHAR(50),
nm_todas_pastagens_ls VARCHAR(100),
nm_composicoes_pastagens_ls VARCHAR(50),
sk_faixa_area_ls INTEGER,
nm_usa_suplemento_ls VARCHAR(3),
vl_ndt_medio_suplementos_ls REAL,
nr_ano_ls INTEGER,
fk_propriedade_fgc INTEGER,
nr_animais_vacas_fgc INTEGER,
nr_animais_novilhos_0_12_fgc INTEGER,
nr_animais_novilhos_13_24_fgc  INTEGER,
nr_animais_novilhos_25_36_fgc  INTEGER,
nr_animais_bezerros_bezerras_fgc  INTEGER,
nr_animais_novilhas_0_12_fgc  INTEGER,
nr_animais_novilhas_13_24_fgc  INTEGER,
nr_animais_novilhas_25_36_fgc  INTEGER,
nr_animais_touros_fgc INTEGER,
vl_area_fgc REAL,
```

```
            vl_receita_anual_fgc NUMERIC,
            vl_quilos_peso_vivo_fgc NUMERIC,
            vl_custos_medicamentos_fgc NUMERIC,
            vl_custos_mao_de_obra_fgc NUMERIC,
            vl_custos_maquinas_fgc NUMERIC,
            vl_custos_pastagens_fgc NUMERIC,
            vl_custos_arrendamento_fgc NUMERIC,
            vl_custos_outros_fgc NUMERIC,
            nr_ano_fgc INTEGER,
            id_erro INTEGER,
            data_carga TIMESTAMP WITHOUT TIME ZONE
);

DROP TABLE IF EXISTS public.categoria_pastagem;
CREATE TABLE public.categoria_pastagem
(
    id integer NOT NULL,
    nm_categoria_pastagem character varying(30) NOT NULL,
    PRIMARY KEY (id)
)

DROP TABLE IF EXISTS public.pastagens;
CREATE TABLE public.pastagens
(
    id integer NOT NULL,
    nm_pastagem character varying(30) NOT NULL,
    fk_categoria_pastagem integer NOT NULL,
    PRIMARY KEY (id),
    CONSTRAINT pastagens_pkey FOREIGN KEY (fk_categoria_pastagem) REFERENCES
public.categoria_pastagem (id) MATCH SIMPLE
)

INSERT INTO public.categoria_pastagem VALUES (1,'Natural'),(2,'Nativa'),(3,'Artificial');
INSERT INTO public.pastagens VALUES (1,'campo nativo',1),(2,'azevém forragem',3),(3,'sorgo
forrageiro',3);

DROP TABLE IF EXISTS paises;
CREATE TABLE paises (
  id SMALLINT NOT NULL,
  nome VARCHAR(32) NOT NULL,
  CONSTRAINT pk_paises PRIMARY KEY (id)
);

DROP TABLE IF EXISTS regioes;
CREATE TABLE regioes (
  id SMALLINT NOT NULL,
  id_pais SMALLINT NOT NULL,
  nome VARCHAR(32) NOT NULL,
  CONSTRAINT pk_regioes PRIMARY KEY (id)
);

DROP TABLE IF EXISTS estados;
CREATE TABLE estados (
  id SMALLINT NOT NULL,
  id_pais SMALLINT NOT NULL,
  id_regiao SMALLINT NOT NULL,
  nome VARCHAR(32) NOT NULL,
  sigla VARCHAR(2) NOT NULL,
  CONSTRAINT pk_estados PRIMARY KEY (id)
);
```

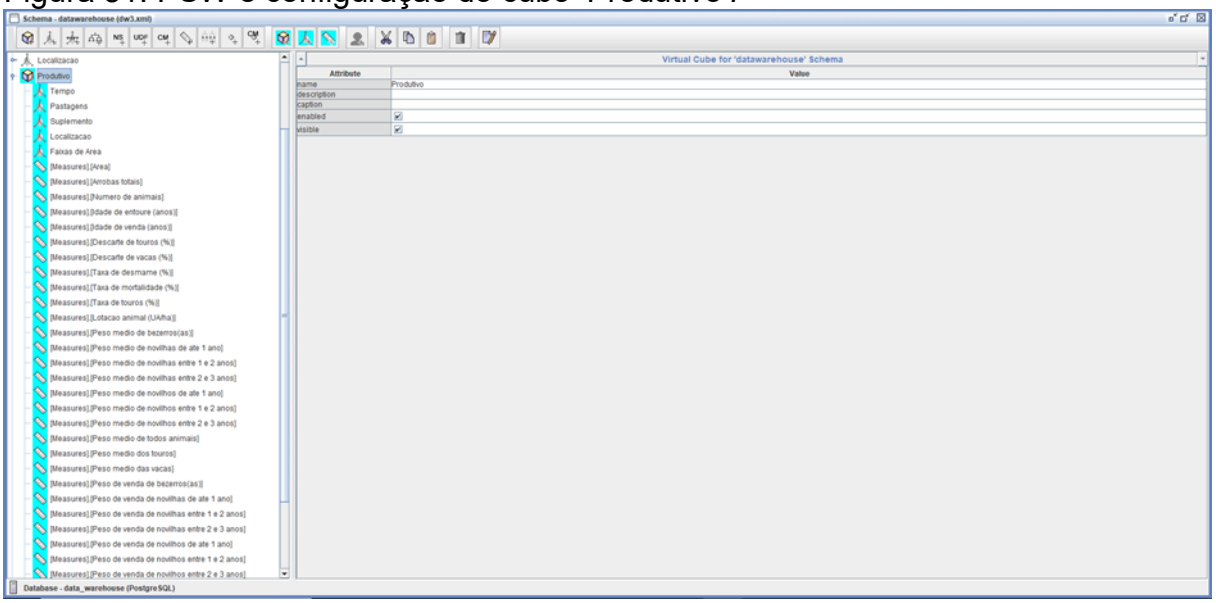
```
DROP TABLE IF EXISTS mesorregioes;
CREATE TABLE mesorregioes (
  id SMALLINT NOT NULL,
  id_estado SMALLINT NOT NULL,
  nome VARCHAR(64) NOT NULL,
  CONSTRAINT pk_mesorregioes PRIMARY KEY (id)
);

DROP TABLE IF EXISTS microrregioes;
CREATE TABLE microrregioes (
  id INTEGER NOT NULL,
  id_mesorregiao SMALLINT NOT NULL,
  id_estado SMALLINT NOT NULL,
  nome VARCHAR(64) NOT NULL,
  CONSTRAINT pk_microrregioes PRIMARY KEY (id)
);

DROP TABLE IF EXISTS municipios;
CREATE TABLE municipios (
  id INTEGER NOT NULL,
  id_microrregiao INTEGER NOT NULL,
  id_mesorregiao SMALLINT NOT NULL,
  id_estado SMALLINT NOT NULL,
  nome VARCHAR(64) NOT NULL,
  CONSTRAINT pk_municipios PRIMARY KEY (id)
);

DROP TABLE IF EXISTS distritos;
CREATE TABLE distritos (
  id INTEGER NOT NULL,
  id_municipio INTEGER NOT NULL,
  id_microrregiao INTEGER NOT NULL,
  id_mesorregiao SMALLINT NOT NULL,
  id_estado SMALLINT NOT NULL,
  nome VARCHAR(64) NOT NULL,
  CONSTRAINT pk_distritos PRIMARY KEY (id)
);

DROP TABLE IF EXISTS subdistritos;
CREATE TABLE subdistritos (
  id BIGINT NOT NULL,
  id_distrito INTEGER NOT NULL,
  id_municipio INTEGER NOT NULL,
  id_microrregiao INTEGER NOT NULL,
  id_mesorregiao SMALLINT NOT NULL,
  id_estado SMALLINT NOT NULL,
  nome VARCHAR(64) NOT NULL,
  CONSTRAINT pk_subdistritos PRIMARY KEY (id)
);

ALTER TABLE regioes ADD CONSTRAINT fk_regioes_id_pais FOREIGN KEY (id_pais)
REFERENCES paises (id);
ALTER TABLE estados ADD CONSTRAINT fk_estados_id_pais FOREIGN KEY (id_pais)
REFERENCES paises (id);
ALTER TABLE estados ADD CONSTRAINT fk_estados_id_regiao FOREIGN KEY (id_regiao)
REFERENCES regioes (id);

ALTER TABLE mesorregioes ADD CONSTRAINT fk_mesorregioes_id_estado FOREIGN KEY
(id_estado) REFERENCES estados (id);
```

```
ALTER TABLE microrregioes ADD CONSTRAINT fk_microrregioes_id_mesorregiao FOREIGN KEY (id_mesorregiao) REFERENCES mesorregioes (id);
ALTER TABLE microrregioes ADD CONSTRAINT fk_microrregioes_id_estado FOREIGN KEY (id_estado) REFERENCES estados (id);

ALTER TABLE municipios ADD CONSTRAINT fk_municipios_id_microrregiao FOREIGN KEY (id_microrregiao) REFERENCES microrregioes (id);
ALTER TABLE municipios ADD CONSTRAINT fk_municipios_id_mesorregiao FOREIGN KEY (id_mesorregiao) REFERENCES mesorregioes (id);
ALTER TABLE municipios ADD CONSTRAINT fk_municipios_id_estado FOREIGN KEY (id_estado) REFERENCES estados (id);

ALTER TABLE distritos ADD CONSTRAINT fk_distritos_id_municipio FOREIGN KEY (id_municipio) REFERENCES municipios (id);
ALTER TABLE distritos ADD CONSTRAINT fk_distritos_id_microrregiao FOREIGN KEY (id_microrregiao) REFERENCES microrregioes (id);
ALTER TABLE distritos ADD CONSTRAINT fk_distritos_id_mesorregiao FOREIGN KEY (id_mesorregiao) REFERENCES mesorregioes (id);
ALTER TABLE distritos ADD CONSTRAINT fk_distritos_id_estado FOREIGN KEY (id_estado) REFERENCES estados (id);

ALTER TABLE subdistritos ADD CONSTRAINT fk_subdistritos_id_distrito FOREIGN KEY (id_distrito) REFERENCES distritos (id);
ALTER TABLE subdistritos ADD CONSTRAINT fk_subdistritos_id_municipio FOREIGN KEY (id_municipio) REFERENCES municipios (id);
ALTER TABLE subdistritos ADD CONSTRAINT fk_subdistritos_id_microrregiao FOREIGN KEY (id_microrregiao) REFERENCES microrregioes (id);
ALTER TABLE subdistritos ADD CONSTRAINT fk_subdistritos_id_mesorregiao FOREIGN KEY (id_mesorregiao) REFERENCES mesorregioes (id);
ALTER TABLE subdistritos ADD CONSTRAINT fk_subdistritos_id_estado FOREIGN KEY (id_estado) REFERENCES estados (id);
```

5. Configuração dos cubos OLAP

O próximo passo é realizar a configuração dos cubos de dados para apresentação na interface da aplicação. Para realizar essa configuração foi utilizado o *Pentaho Schema Workbench* (PSW), que permite configurar os cubos através de uma interface gráfica. A Figura 31 apresenta a interface do PSW e o exemplo de configuração do cubo '**Produtivo**', onde é possível observar no lado esquerdo da imagem as dimensões deste cubo (tempo, faixas de área, pastagens, suplemento e localização) e logo abaixo, as métricas produtivas. Os outros cubos foram configurados de maneira semelhante.

Figura 31: PSW e configuração do cubo 'Produtivo'.



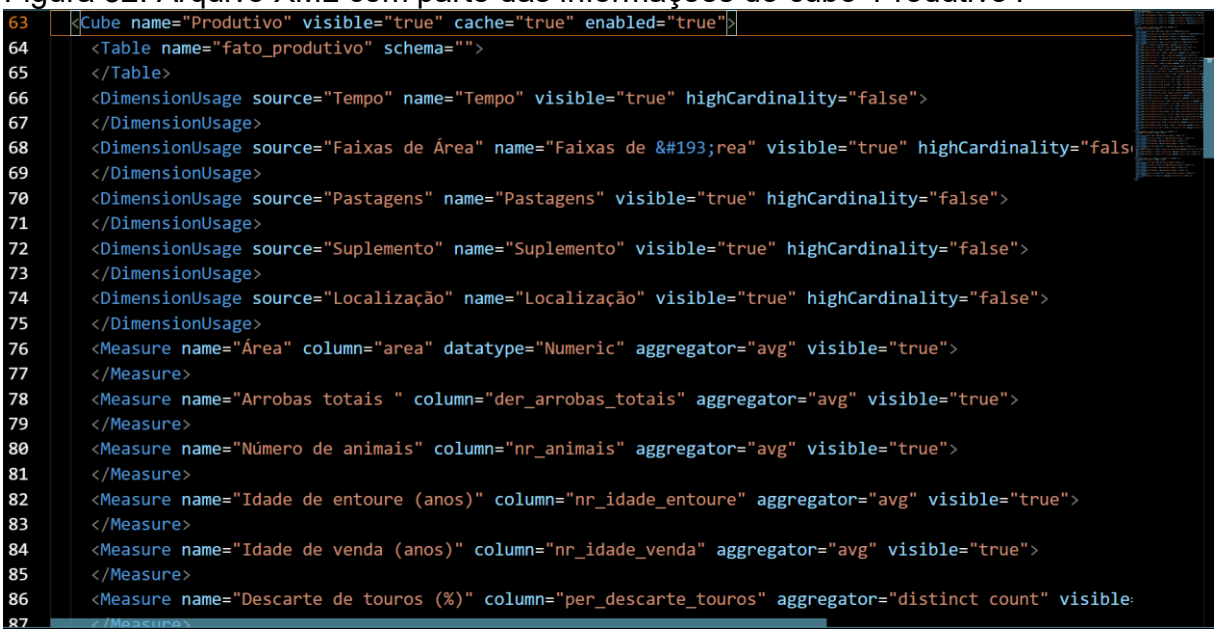
Fonte: Autor (2019).

Com o PSW, é possível atribuir nomes para as colunas das dimensões e fatos (diferente dos nomes na tabelas físicas), realizar diferentes tipos de operações com os dados, quando agregados na aplicação (média, soma, mínimo, máximo ou contagem), definir dimensões conformadas, atribuir regras de acesso com base nas credenciais do usuário, entre diversas outras possibilidades. Tais funções podem ser realizadas no PSW, após realizar a sua conexão com a base de dados do DW.

Como resultado da configuração dos cubos, são gerados arquivos em que possuem a estrutura do cubo através do uso de tabelas fatos e dimensões encontradas no servidor do DW. Estes metadados gerados no formato XML (*Extensible Markup Language*) são interpretados pelo *Mondrian* em conjunto com a base de dados do DW. O arquivo .xml gerado foi salvo em um diretório com o nome de *Schemas*. A Figura 32 apresenta um exemplo de arquivo XML gerado para o cubo 'Produtivo'. O mesmo procedimento é realizado para os cubos 'Economico' e 'Ambiental', e com isso, a configuração dos cubos para a apresentação dos dados na interface da aplicação é concluída.

Para realizar modificações nos cubos, é possível modificar o arquivo XML gerado ou utilizar a interface gráfica do PSW.

Figura 32: Arquivo XML com parte das informações do cubo ‘Produtivo’.



Fonte: Autor (2019).

Após configurar o cubo, é necessário realizar as seguintes etapas, para a publicação do mesmo no *Pentaho BA Server*:

1. Na interface gráfica do PSW, com o arquivo do cubo que deve ser publicado aberto, clicar em *File -> Publish*...
2. No campo *server* URL -> http://localhost:8080/pentaho, nos campos *user* e *password*, utilizar as informações credenciais de *login* do *Pentaho BA Server*, no campo *Pentaho or JNDI Data Source* é necessário informar o nome do *Data Source* (fonte de dados) do *Pentaho BA Server* no qual o cubo será publicado (*data_warehouse*, em nosso caso). O procedimento para criação e configuração do *Data Source* é explanado entre as páginas 56-58.
3. Certificar-se que o *Pentaho BI Server* está em execução.
4. Clicar em *Publish*.
5. Se não aparecer nenhuma mensagem, é porque o cubo foi publicado com sucesso no servidor da aplicação. Se aparecer uma janela com o Título *Overwrite*? é porque já existe um *schema* para este mesmo *Data Source*. Basta clicar em *Yes* para sobrescrever. Se aparecer uma janela com o título *Publish Error*, é porque ocorreu um erro ao publicar o cubo no Pentaho BI Server. Neste caso, você deve certificar-se que o *Pentaho BA Server* está em execução e se os parâmetros informados no passo 2 estão corretos.

Com isso, os próximos passos são relacionados ao *Pentaho BA Server* e o *plugin Saiku Analytics*.

5.1 Arquivo XML dos cubos do PSW

```
<Schema name="datawarehouse">
  <Dimension type="StandardDimension" visible="true" highCardinality="false" name="Tempo">
    <Hierarchy visible="true" hasAll="true" primaryKey="sk_tempo">
      <Table name="dim_tempo" schema="datawarehouse">
      </Table>
      <Level name="Ano" visible="true" table="dim_tempo" column="nr_ano" type="Integer"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension type="StandardDimension" visible="true" highCardinality="false" name="Faixas de
Area">
    <Hierarchy visible="true" hasAll="true" primaryKey="sk_area">
      <Table name="dim_area" schema="datawarehouse">
      </Table>
      <Level name="Faixa de Area (ha)" visible="true" column="fx_area" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension type="StandardDimension" visible="true" highCardinality="false" name="Pastagens">
    <Hierarchy visible="true" hasAll="true" primaryKey="sk_pastagem">
      <Table name="dim_pastagem" schema="datawarehouse">
      </Table>
      <Level name="Tipo da pastagem predominante" visible="true" column="nm_cat_pastagem"
type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
      <Level name="Pastagem predominante" visible="true" column="nm_pastagem" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
      <Level name="Descricao de todas pastagens" visible="true" column="nm_descricao" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension type="StandardDimension" visible="true" highCardinality="false" name="Suplemento">
    <Hierarchy visible="true" hasAll="true" primaryKey="sk_suplemento">
      <Table name="dim_suplemento" schema="datawarehouse">
      </Table>
      <Level name="Usa suplementos?" visible="true" column="nm_usa_suplemento" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
      <Level name="Nutrientes Digestivos Totais Medio" visible="true" column="nm_per_ndt"
type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
    </Hierarchy>
  </Dimension>
  <Dimension type="StandardDimension" visible="true" highCardinality="false" name="Localizacao">
    <Hierarchy visible="true" hasAll="true" primaryKey="sk_localizacao">
      <Table name="dim_localizacao" schema="datawarehouse">
      </Table>
      <Level name="Nome da propriedade" visible="true" column="nm_propriedade" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
      <Level name="Municipio" visible="true" column="nm_cidade" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
```

```
      <Level name="Microrregiao" visible="true" column="nm_microrregiao" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
      <Level name="Mesorregiao" visible="true" column="nm_mesorregiao" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
      <Level name="Estado" visible="true" column="nm_unidade_federativa" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
      <Level name="Sigla do estado" visible="true" column="nm_cd_unidade_federativa" type="String"
uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
      <Level name="Regiao" visible="true" column="nm_regiao" type="String" uniqueMembers="false"
levelType="Regular" hideMemberIf="Never">
      </Level>
      <Level name="Pais" visible="true" column="nm_pais" type="String" uniqueMembers="false"
levelType="Regular" hideMemberIf="Never">
      </Level>
    </Hierarchy>
  </Dimension>
  <Cube name="Integrado" visible="true" cache="true" enabled="true">
    <Table name="fato_produtivo_economico_ambiental" schema="datawarehouse">
    </Table>
    <DimensionUsage source="Tempo" name="Tempo" visible="true" foreignKey="sk_tempo"
highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="Faixas de Area" name="Faixas de Area" visible="true"
foreignKey="sk_area" highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="Pastagens" name="Pastagens" visible="true"
foreignKey="sk_pastagem" highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="Suplemento" name="Suplemento" visible="true"
foreignKey="sk_suplemento" highCardinality="false">
    </DimensionUsage>
    <DimensionUsage source="Localizacao" name="Localizacao" visible="true"
foreignKey="sk_localizacao" highCardinality="false">
    </DimensionUsage>
    <Measure name="Area" column="area" datatype="Numeric" aggregator="avg" visible="true">
    </Measure>
    <Measure name="Arrobas totais" column="der_arrobas_totais" aggregator="avg" visible="true">
    </Measure>
    <Measure name="Numero de animais" column="nr_animais" aggregator="avg" visible="true">
    </Measure>
    <Measure name="Idade de entoure (anos)" column="nr_idade_entoure" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Idade de venda (anos)" column="nr_idade_venda" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Descarte de touros (%)" column="per_descarte_touros" datatype="Numeric"
aggregator="avg" visible="true">
    </Measure>
    <Measure name="Descarte de vacas (%)" column="per_descarte_vacas" datatype="Numeric"
aggregator="avg" visible="true">
    </Measure>
    <Measure name="Taxa de desmame (%)" column="per_desmame" datatype="Numeric"
aggregator="avg" visible="true">
    </Measure>
```

```
   <Measure name="Taxa de mortalidade (%)" column="per_mortalidade" datatype="Numeric"
aggregator="avg" visible="true">
   </Measure>
   <Measure name="Taxa de touros (%)" column="per_touros" datatype="Numeric" aggregator="avg"
visible="true">
   </Measure>
   <Measure name="Lotacao animal (UA/ha)" column="vl_lotacao_animal" datatype="Numeric"
aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso medio de bezerros(as)" column="vl_peso_medio_bezerros_bezerras"
aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso medio de novilhas de ate 1 ano" column="vl_peso_medio_novilhas_0_12"
aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso medio de novilhas entre 1 e 2 anos"
column="vl_peso_medio_novilhas_13_24" aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso medio de novilhas entre 2 e 3 anos"
column="vl_peso_medio_novilhas_25_36" aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso medio de novilhos de ate 1 ano" column="vl_peso_medio_novilhos_0_12"
aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso medio de novilhos entre 1 e 2 anos"
column="vl_peso_medio_novilhos_13_24" aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso medio de novilhos entre 2 e 3 anos"
column="vl_peso_medio_novilhos_25_36" aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso medio de todos animais" column="vl_peso_medio_total" aggregator="avg"
visible="true">
   </Measure>
   <Measure name="Peso medio dos touros" column="vl_peso_medio_touros" aggregator="avg"
visible="true">
   </Measure>
   <Measure name="Peso medio das vacas" column="vl_peso_medio_vacas" aggregator="avg"
visible="true">
   </Measure>
   <Measure name="Peso de venda de bezerros(as)" column="vl_peso_venda_bezerros_bezerras"
aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso de venda de novilhas de ate 1 ano"
column="vl_peso_venda_novilhas_0_12" aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso de venda de novilhas entre 1 e 2 anos"
column="vl_peso_venda_novilhas_13_24" aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso de venda de novilhas entre 2 e 3 anos"
column="vl_peso_venda_novilhas_25_36" aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso de venda de novilhos de ate 1 ano"
column="vl_peso_venda_novilhos_0_12" aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso de venda de novilhos entre 1 e 2 anos"
column="vl_peso_venda_novilhos_13_24" aggregator="avg" visible="true">
   </Measure>
   <Measure name="Peso de venda de novilhos entre 2 e 3 anos"
column="vl_peso_venda_novilhos_25_36" aggregator="avg" visible="true">
   </Measure>
```

```
    <Measure name="Peso de venda dos touros" column="vl_peso_venda_touros" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Peso de venda das vacas" column="vl_peso_venda_vacas" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Produtividade por area (kg/ha)" column="vl_produtividade" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Quilos de peso vivo totais" column="vl_quilos_peso_vivo_total" aggregator="avg"
visible="true">
    </Measure>

    <Measure name="Emissao por produtividade" column="vl_emissao_por_produtividade"
datatype="Numeric" aggregator="avg" visible="true">
    </Measure>
    <Measure name="Emissao total" column="vl_emissao_total" datatype="Numeric" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Fertilizantes" column="vl_fertilizantes" datatype="Integer" aggregator="avg"
visible="true">
    </Measure>

    <Measure name="Custos por Area" column="der_custos_por_area" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Custos por arroba" column="der_custos_por_arroba" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Custos por cabeca" column="der_custos_por_cabeca" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Custo por quilo de peso vivo" column="der_custos_por_quilo_peso_vivo"
aggregator="avg" visible="true">
    </Measure>
    <Measure name="Custos totais" column="der_custos_totais" aggregator="avg" visible="true">
    </Measure>
    <Measure name="Lucro Bruto" column="der_lucro" aggregator="avg" visible="true">
    </Measure>
    <Measure name="Custos com arrendamento" column="vl_custos_arrendamento" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Custos com mao-de-obra" column="vl_custos_mao_de_obra" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Custos com maquinas" column="vl_custos_maquinas" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Custos com pastagens" column="vl_custos_pastagens" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Outros custos (impostos, etc)" column="vl_custos_outros" aggregator="avg"
visible="true">
    </Measure>
    <Measure name="Receita total" column="vl_receita_total" aggregator="avg" visible="true">
    </Measure>
  </Cube>

  <VirtualCube name="Produtivo">
    <VirtualCubeDimension name="Tempo"/>
```

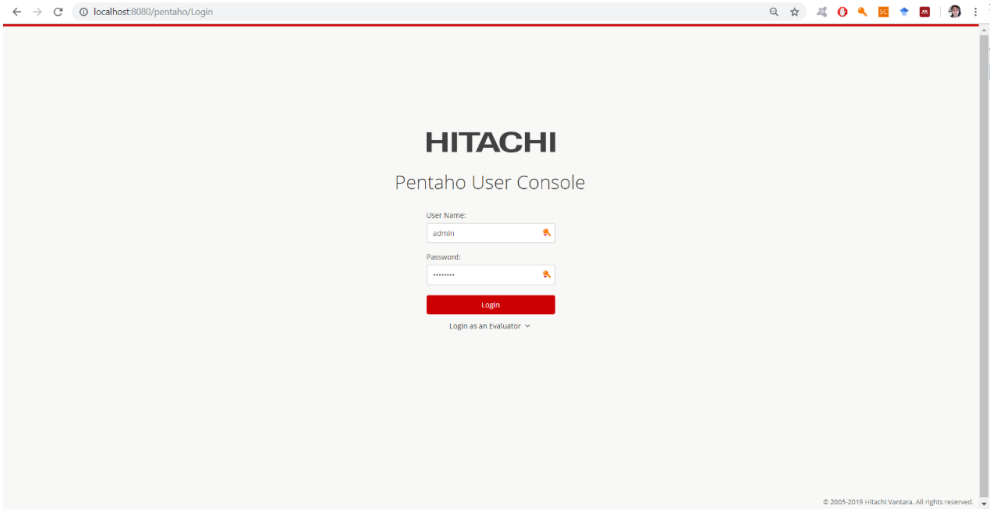
```
    <VirtualCubeDimension name="Pastagens"/>
    <VirtualCubeDimension name="Suplemento"/>
    <VirtualCubeDimension name="Localizacao"/>
    <VirtualCubeDimension name="Faixas de Area"/>

    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Area]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Arrobas totais]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Numero de animais]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Idade de entoure (anos)]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Idade de venda (anos)]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Descarte de touros (%)]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Descarte de vacas (%)]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Taxa de desmame (%)]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Taxa de mortalidade (%)]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Taxa de touros (%)]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Lotacao animal (UA/ha)]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso medio de bezerros(as)]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso medio de novilhas de ate 1
ano]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso medio de novilhas entre 1
e 2 anos]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso medio de novilhas entre 2
e 3 anos]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso medio de novilhos de ate 1
ano]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso medio de novilhos entre 1
e 2 anos]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso medio de novilhos entre 2
e 3 anos]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso medio de todos animais]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso medio dos touros]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso medio das vacas]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso de venda de
bezerros(as)]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso de venda de novilhas de
ate 1 ano]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso de venda de novilhas entre
1 e 2 anos]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso de venda de novilhas entre
2 e 3 anos]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso de venda de novilhos de
ate 1 ano]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso de venda de novilhos entre
1 e 2 anos]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso de venda de novilhos entre
2 e 3 anos]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso de venda dos touros]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Peso de venda das vacas]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Produtividade por area
(kg/ha)]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Quilos de peso vivo totais]"/>
  </VirtualCube>
  <VirtualCube name="Ambiental">
    <VirtualCubeDimension name="Tempo"/>
    <VirtualCubeDimension name="Pastagens"/>
    <VirtualCubeDimension name="Suplemento"/>
    <VirtualCubeDimension name="Localizacao"/>
    <VirtualCubeDimension name="Faixas de Area"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Emissao por produtividade]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Emissao total]"/>
```

```
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Fertilizantes]"/>
  </VirtualCube>
  <VirtualCube name="Economico">
    <VirtualCubeDimension name="Tempo"/>
    <VirtualCubeDimension name="Pastagens"/>
    <VirtualCubeDimension name="Suplemento"/>
    <VirtualCubeDimension name="Localizacao"/>
    <VirtualCubeDimension name="Faixas de Area"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Custos por Area]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Custos por arroba]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Custos por cabeca]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Custo por quilo de peso vivo]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Custos totais]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Lucro Bruto]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Custos com arrendamento]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Custos com mao-de-obra]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Custos com maquinas]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Custos com pastagens]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Outros custos (impostos, etc)]"/>
    <VirtualCubeMeasure cubeName="Integrado" name="[Measures].[Receita total]"/>
  </VirtualCube>
</Schema>
```

6. *Pentaho BA Server e Saiku Analytics*

Para iniciar a execução do *Pentaho BA Server*, basta executar o arquivo start-pentaho.bat (Windows) ou start-pentaho.sh (Linux). Para inicializar o *Pentaho BA Server* na máquina servidora local, basta abrir um navegador de internet de sua preferência e digitar o seguinte endereço: http://localhost:8080/pentaho/Login
Para acessar de outro dispositivo conectado na rede local, basta digitar o seguinte endereço: http://IP_DA_MAQUINA_SERVIDORA:8080/pentaho/Login
Se tudo ocorrer conforme o previsto, será apresentada a seguinte tela ao usuário (Figura 33).

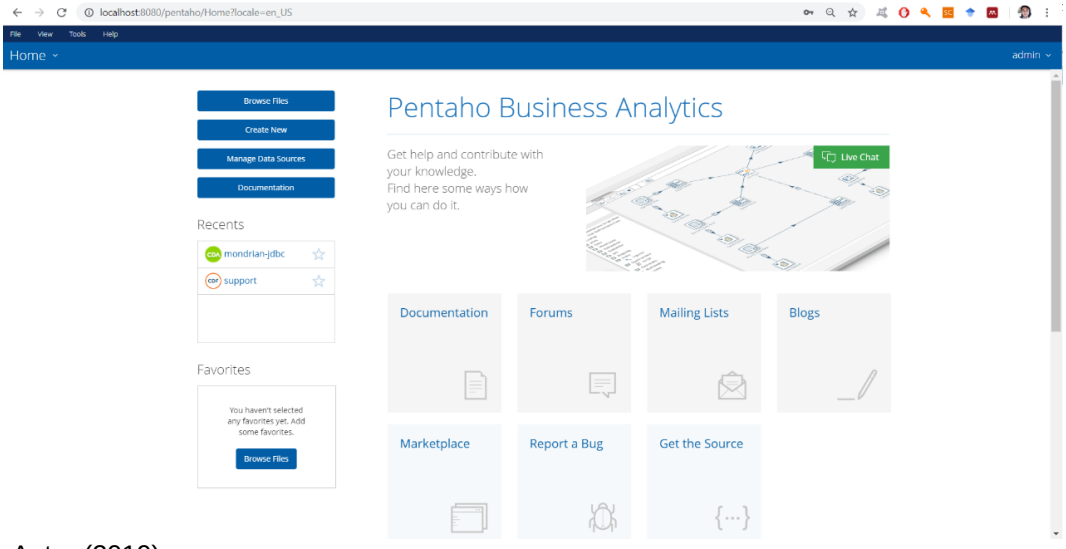
Figura 33: Interface gráfica da ferramenta *Pentaho BA Server.*



Fonte: Autor (2019).

Por padrão, é possível realizar a autenticação como administrador com o *User Name* admin e o *Password* password. Com isso, será exibida a tela da Figura 34.

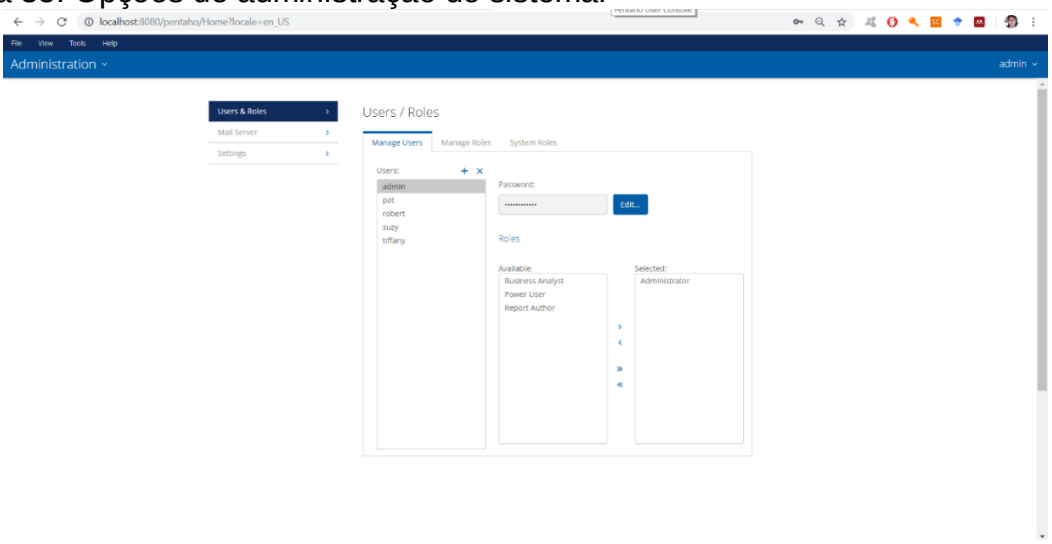
Figura 34: Interface gráfica da ferramenta *Pentaho BA Server.*



Fonte: Autor (2019).

Para alterar os usuários que vem por padrão na aplicação, basta clicar em *Home -> Administration* no menu localizado no canto superior esquerdo da tela. Na opção *Users and Roles* é possível não só modificar *login* e senha dos usuários como também regras ou níveis de acesso às funcionalidades da aplicação. Também é possível configurar um servidor de email SMTP e agendar ou deletar arquivos do sistema, nas opções *Mail Server* e *Settings*, respectivamente (Figura 35).

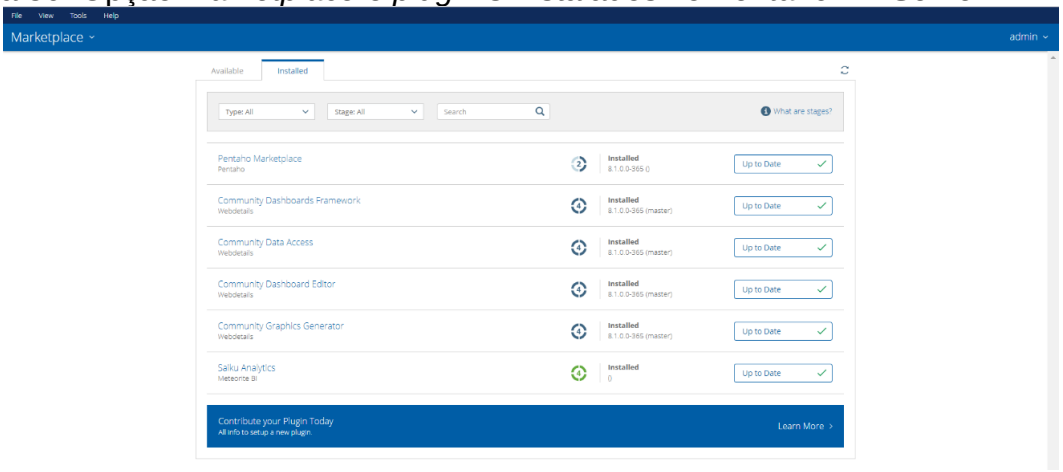
Figura 35: Opções de administração do sistema.



Fonte: Autor (2019).

Na opção *Home -> Marketplace*, é possível verificar *plug-ins* já instalados na aplicação e também instalar funcionalidades adicionais. Foi instalado o *plug-in Saiku Analytics* para acessar os dados do DW e realizar análises nos cubos OLAP, de diferentes formas. O processo de instalação é intuitivo e requer a reinicialização do servidor para que as mudanças sejam percebidas na aplicação. A Figura 36 apresenta a interface do *Marketplace*, com o *Saiku Analytics* já instalado.

Figura 36: Opção *Marketplace* e *plug-ins* instalados no *Pentaho BA Server.*



Fonte: Autor (2019).

Na opção *Manage Data Sources*, na interface principal da aplicação, é possível modificar, configurar e excluir fontes de dados, gerenciar conexões com bancos de dados, importar arquivos gerados pelo Mondrian e importar metadados.

Passo a passo da configuração: é necessário, inicialmente, criar uma conexão com a base de dados do DW (*PostgreSQL*). No menu *Manage Data Sources*, clicar na opção *New Data Source.* No campo *Data Source Name*, é necessário dar um nome para a fonte de dados (*data_warehouse* em nosso caso). No campo *Source Type*, selecionar a opção *database table(s)* (no caso, as fontes dos dados são tabelas de um banco de dados). Na sequência, clicar no botão *Add Conection*, representado pelo símbolo '**+**', para criar uma conexão com um banco de dados. Selecionar o *PostgreSQL* no campo *Database Type*, Nos campos restantes, é necessário preencher informações sobre a conexão:

- *Connection name:* nome da conexão com o banco de dados *(data_warehouse).*
- *Host name*: endereço IP da máquina servidora do banco de dados *(localhost).*
- *Database name*: nome do banco de dados que será acessado *(data_warehouse).*
- *Port number:* número da porta do servidor do banco de dados (5432).
- *User name*: nome de usuário do banco de dados *(postgres).*
- *Password:* senha do usuário do banco de dados (12345).

Após preenchidos, clicar no botão *Test* para checar se a conexão foi bem sucedida. Se a mensagem "*Connection to database [Connection name] suceeded*" aparecer, é porque a conexão com a base de dados foi realizada com sucesso. Ao clicar em *Ok*, o usuário volta na tela anterior, onde é necessário selecionar a opção "*Reporting and Analysis (Requires Star Schema)*" e clicar em *Next*.

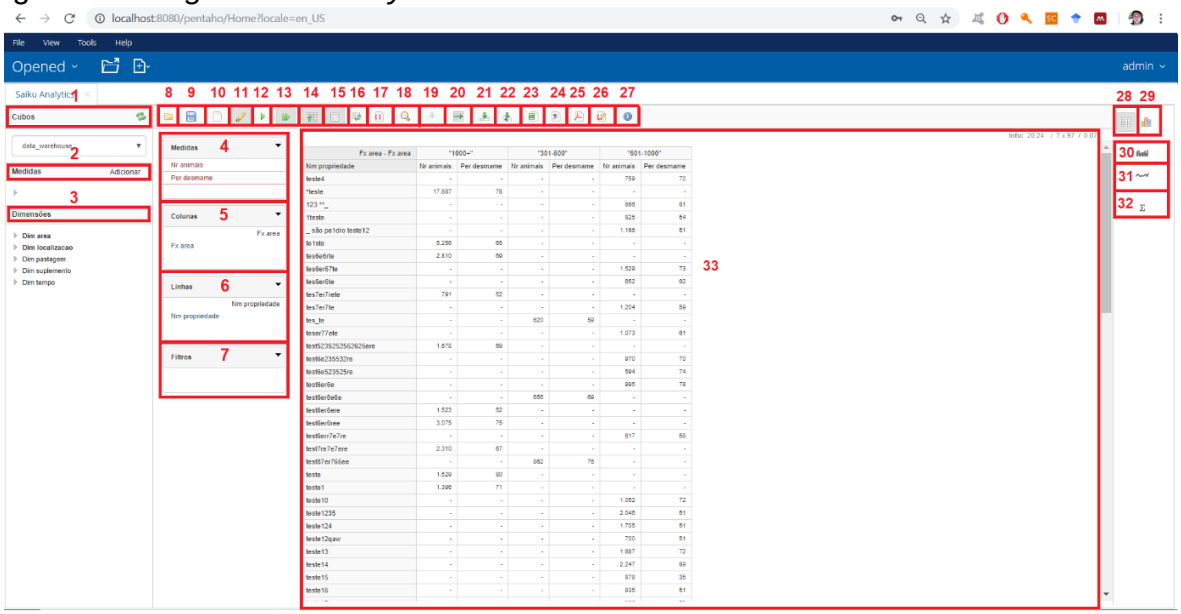
Nesta nova janela, primeiro é necessário selecionar o *schema* correto (*data_warehouse*) do banco de dados. Após a seleção, são selecionadas as dimensões e a tabela fato correspondente à estas dimensões. No menu *Fact Table*, selecionar a tabela que representa as métricas do negócio, e por fim, o usuário pode clicar em *Next*.

Nesta última janela, é necessário criar as junções entre as dimensões e a tabela fato, através das chaves substitutas (*surrogate keys*). No campo *Left Table* deve aparecer selecionada a tabela fato, enquanto que no campo *Right Table* deve aparecer uma das dimensões. O usuário deve selecionar a chave estrangeira da tabela fato que faz referência à chave primária da tabela dimensão do lado direito, e clicar em Create Join. Este processo deve ser realizado cinco vezes, pois existem cinco dimensões para cada tabela fato. Após concluído este processo, clicar em *Finish*, e o processo de criação de um *Data Source* (fonte de dados) estará concluído. Ao clicar novamente em *Manage Data Sources*, na interface principal da aplicação, aparecerá dois novos *Data Sources*: *data_warehouse* (JDBC, a conexão com a base de dados *PostgreSQL*) e *data_warehouse* (*Data Source Wizard*, a configuração das tabelas fato e dimensões do *star schema*).

Para realizar a etapa de acesso aos dados, na interface principal do *Pentaho BA Server*, é necessário clicar em "*Create New*" -> "*Saiku Analytics*". Na nova janela, clicar em "Criar Nova Consulta", e a interface principal do *plugin Saiku Analytics* aparecerá (Figura 37).

Na sequência, serão abordadas cada uma das funcionalidades do *plugin Saiku Analytics*, conforme cada uma das regiões demarcadas na Figura 37.

Figura 37: *Plugin Saiku Analytics.*



Fonte: Autor (2019).

1) Região de seleção do cubo de dados a ser analisado. O botão verde localizado ao lado direito serve para atualizar os cubos de dados, caso estes tenham sido alterados no PSW ou diretamente no arquivo XML.
2) Região de seleção das métricas de interesse para serem analisadas. É possível selecionar uma ou várias métricas. Tais dados aparecem tanto nas tabelas como nos gráficos.
3) Região de seleção das dimensões e descritores de interesse das informações do DW. É possível selecionar um ou mais descritores de diferentes dimensões para a realização de análises.
4) As métricas selecionadas em (2) aparecem neste campo. É possível modificar a forma como esses dados são apresentados, no campo *DETAILS*, representado por uma seta para baixo.
5) Local onde os descritores das dimensões aparecem. É possível aplicar filtros, ordenar de forma crescente e decrescente, apresentam os registros *top* 10 em determinada métrica, entre outras possibilidades. No caso de tabelas, serão os descritores das colunas.
6) Local onde os descritores das dimensões aparecem. É possível aplicar filtros, ordenar de forma crescente e decrescente, apresentam os registros top 10 em determinada métrica, entre outras possibilidades. No caso de tabelas, serão os descritores das linhas.
7) Campo onde é possível aplicar filtros personalizados nos dados exibidos.
8) Botão que permite abrir consultas previamente salvas (não funcional).
9) Botão que permite salvar consultas (não funcional).
10) Botão que permite remover a consulta previamente realizada. Com isso, as métricas, descritores das dimensões e filtros aplicados são removidos das áreas 4, 5, 6 e 7.
11) Botão que habilita ou desabilita a edição das consultas. Ao desabilitar, as áreas 1, 2, 3, 4, 5, 6 e 7 deixam de ser exibidas.
12) Botão para executar uma consulta, após a seleção das métricas, dimensões e filtros de interesse. Ao executar a consulta, os dados serão apresentados, seja nas tabelas ou nos gráficos.

13) Botão que torna as consultas automáticas, ou seja, sempre que o usuário selecionar uma nova métrica, descritor ou filtro, a consulta será realizada de forma instantânea.
14) Botão que permite ocultar níveis superiores.
15) Botão que permite mostrar apenas registros com valores não-nulos, quando habilitado.
16) Botão que alterna os descritores das linhas para as colunas, e vice-versa, quando pressionado.
17) Botão que permite apresentar a consulta realizada na linguagem MDX (*Multidimensional Expressions*), sendo possível também modificá-la.
18) Botão que permite dar zoom na tabela. Ao habilitar o botão, basta selecionar o registro na tabela para realizar uma análise em maiores detalhes.
19) Botão cenário query (Desabilitado, não funcional).
20) Botão que permite realizar a operação *Drill Across on Cell*. Esta operação permite realizar um maior detalhamento da célula selecionada, através da inserção de novas métricas ou descritores das dimensões.
21) Botão que permite realizar a operação *Drill Through* (detalhar célula). Permite analisar todos os registros para a célula selecionada, de acordo com os descritores e métricas escolhidas.
22) Botão que permite exportar o detalhamento da célula (operação 21) para uma planilha .csv. O usuário deve selecionar a célula da tabela, as métricas e descritores de interesse para serem exportados.
23) Botão que permite exportar os dados da tabela em exibição na interface gráfica para uma planilha eletrônica .xls (Excel).
24) Botão que permite exportar os dados da tabela em exibição na interface gráfica para uma planilha eletrônica .csv (arquivo Excel de valores separados por vírgula).
25) Botão que permite exportar a tabela em exibição na interface gráfica (ou o gráfico) para um arquivo em formato de relatório .pdf.
26) Botão que permite executar as consultas diretamente com a linguagem MDX.
27) Botão de sobre com as informações de licença e site oficial da ferramenta *Saiku Analytics*.
28) Botão que permite exibir os dados das consultas no modo tabular.
29) Botão que permite exibir os dados na forma de gráficos. Os possíveis gráficos são: barras, barras empilhadas, barras 100%, múltiplas barras, linha, área, *gride* de temperatura, *tree maps*, *sunburst*, *multisunburst*, pontos, pizza, radar, cascata e *time wheel*. O usuário pode alternar no tipo de gráfico com apenas um clique no gráfico de interesse. Também é possível interagir com os gráficos, clicando nos mesmos para realizar maiores detalhamentos nos mesmos.
30) Caso o modo tabular esteja selecionado, este botão permite apresentar um pequeno gráfico de barras para cada registro da tabela.
31) Caso o modo tabular esteja selecionado, este botão permite apresentar um pequeno gráfico de linhas para cada registro da tabela.
32) Caso o modo tabular esteja selecionado, para as métricas selecionadas são exibidas cinco medidas estatísticas básicas: mínimo, máximo, média, soma e desvio padrão.
33) Região onde são apresentados os dados, tanto no formato tabular, como no formato de gráficos.

Referências

DÍAZ, Josep Curto. **Introducción al Business Intelligence**. Barcelona: Editorial UOC, 2012.

EMBRAPA. **Desenvolvimento de sistemas de apoio à decisão e de métodos de coleta, análise de dados e monitoramento da pecuária na região Sul do Brasil**. Brasília, [2015?]. Disponível em: <https://www.embrapa.br/busca-de-projetos/-/projeto/210797/desenvolvimento-de-sistemas-de-apoio-a-decisao-e-de-metodos-de-coleta-analise-de-dados-e-monitoramento-da-pecuaria-na-regiao-sul-do-brasil>. Acesso em: set. 2018.

KIMBALL, Ralph; ROSS, Margy. **The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling**. 3° ed. Indianapolis: John Wiley & Sons, Inc., 2013.

TURBAN, Efrain; SHARDA, Ramesh E.; DELEN, Dursun. **Decision Support and Business Intelligence Systems**. 9° ed. New Jersey: Pearson Education, Inc., 2011.