

Rafael Liberal Ferreira

GEOKETTLE: APLICAÇÃO DE FERRAMENTAS DE ANÁLISE
ESPACIAL EM REGISTROS DE ESPÉCIES DA FAUNA

XIV Curso de Especialização em Geoprocessamento



UFMG

Instituto de Geociências

Departamento de Cartografia

Rafael Liberal Ferreira

**GEOKETTLE: APLICAÇÃO DE FERRAMENTAS DE ANÁLISE ESPACIAL EM REGISTROS DE
ESPÉCIES DA FAUNA**

Monografia apresentada ao Curso de Pós-Graduação em Geoprocessamento da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Especialista em Geoprocessamento.

Orientador: Clodoveu Augusto Davis Junior

Belo Horizonte - MG

2013

F383g Ferreira, Rafael Liberal.
2013 Geokettle [manuscrito] : aplicação de ferramentas de análise espacial em registros de espécies da fauna / Rafael Liberal Ferreira. – 2013. vii, 40 f. : il. (color.)

Monografia (especialização em Geoprocessamento) – Universidade Federal de Minas Gerais, Instituto de Geociências, 2013.

Orientador: Clodoveu Augusto Davis Junior.

Bibliografia: f. 39-40.

1. Análise espacial (Estatística). 2. Análise espacial - Modelos. 3. Animais. I. Davis Junior, Clodoveu Augusto. II. Universidade Federal de Minas Gerais, Instituto de Geociências. III. Título.

CDU: 502:519.2



Universidade Federal de Minas Gerais
Instituto de Geociências
Departamento de Cartografia
Curso de Especialização em Geoprocessamento

Monografia defendida e aprovada em 04 de Dezembro de 2014 pela
banca examinadora constituída pelos seguintes professores:

A handwritten signature in blue ink, appearing to read "Clodoveu A Davis Jr.", is written over a horizontal line.

Prof. Dr. Clodoveu A Davis Jr. – DCC/UFMG - Orientador

Profa. Dra. Karla Albuquerque de Vasconcelos Borges – PRODABEL

Agradecimentos

Agradeço aos colegas de sala pelo companheirismo durante o curso. Aos professores, que pela dedicação nos proporcionou o crescimento acadêmico.

Agradeço ao Prof. Dr. Clodoveu Augusto Davis Junior, pela dedicação e orientação para o desenvolvimento deste trabalho, fundamental para o bom resultado.

Sumário

1. Introdução.....	8
2. Objetivo Geral	10
Objetivos Específicos	10
3. Referencial Teórico	11
4. GEOBI e suas Funcionalidades	12
Data Warehouse	13
Granularidade	14
Data Warehouse Geográfico.....	15
<i>Data Marts</i> Espaciais (DME)	16
Tipos de Dados Espaciais	17
Consultas SOLAP	18
Range Queries.....	19
Pentaho	19
Geokettle	20
5. Metodologia.....	24
6. Estudo de Caso.....	27
7. Considerações Finais.....	37
8. Referências Bibliográficas	39

Lista de Figuras

Figura 1 - Arquitetura de um DW	14
Figura 2 - ARQUITETURA DE UM DWG.....	16
Figura 3 - Exemplo de Feições	17
Figura 4 - Representação dos tipos de dados espaciais	18
Figura 5 - INTERSECTION RANGE QUERY	19
Figura 6 - CONTAINMENT RANGE QUERY.....	19
Figura 7 - ENCLOSURE RANGE QUERY	19
Figura 8 - EXEMPLO DE ETAPAS DO ETL NO GEOKETTLE	21
Figura 9 - ESTRUTURA DE ORGANIZAÇÃO DO ETL.....	21
Figura 10 - CONFIGURAÇÃO DO GEOKETTLE	23
Figura 11 - EXEMPLO DE CONFIGURAÇÃO DE UM MODELO DIMENSIONAL.....	24
Figura 12 - FLUXOGRAMA DA METODOLOGIA APLICADA.....	26
Figura 13 - MODELO LÓGICO DO BANCO DE DADOS DA FAUNA.....	27
Figura 14 - FERRAMENTAS UTILIZADAS NO ETL	28
Figura 15 - FLUXOGRAMA DE DESENVOLVIMENTO DO ETL	31
Figura 16 - FERRAMENTAS DA PRIMEIRA ETAPA DE EXTRAÇÃO DE DADOS	32
Figura 17 - DETALHE DO SISTEMA DE PROJEÇÃO SAD69(DEG)	33
Figura 18 - EXEMPLO DA ETAPA DE TRANSFORMAÇÃO ATRAVÉS DO JOIN.....	34
Figura 19 - EXEMPLO DA ETAPA DE TRANSFORMAÇÃO ATRAVÉS DO MERGE JOIN – INPUT SHP E OUTPUT EXCEL	35

Figura 20 - EXEMPLO DA ETAPA DE TRANSFORMAÇÃO ATRAVÉS DO MERGE JOIN – INPUT E OUTPUT EXCEL	36
Figura 21 - PLANILHA RESULTANTE DOS CRUZAMENTOS	36

Resumo

O presente estudo tem como objetivo aplicar as principais ferramentas de análise espacial da extensão Geokettle em registros de espécies da fauna. Tendo como metodologia a pesquisa bibliográfica sobre GeoBI e suas funcionalidades; análise e descrição das principais ferramentas disponíveis de análise espacial do Geokettle; aplicação de um estudo de caso utilizando as ferramentas estudadas e avaliação da eficiência das ferramentas em relação aos resultados obtidos.

O desafio de integrar SIG e BI, conhecendo as potencialidades e aplicabilidades, a fim de possibilitar a melhor opção para a tomada de decisão, apoiada por técnicas de identificação, extração e análise de dados, é proposta desse estudo. O Geokettle é uma poderosa aplicação do PDI, possibilitando diversas modelagens, principalmente, análises espaciais. A sua manipulação e entendimento se torna fácil quando se tem o conhecimento do Kettle.

A falta de uma documentação detalhada e de conhecimento do funcionamento da aplicação inviabilizou o uso de algumas ferramentas. Os diversos formatos de *input* e *output* aceitáveis, conexões avançadas com banco de dados e ferramentas diversas, possibilitam o uso do Geokettle como um potencial para processo de ETL geográfico. Trabalhos futuros podem ser desenvolvidos, analisando as potencialidades das ferramentas e possibilitando trabalhar melhor as análises espaciais. Deste modo, o conhecimento adquirido com este estudo de caso aplicado possibilitou uma série de questionamentos que ficou sem um resultado.

Palavras chave: Análise Espacial, GeoBI, Geokettle.

1. Introdução

O desenvolvimento de tecnologias aplicadas às questões espaciais e a disseminação de ferramentas de armazenamento, tratamento, análise e disponibilização de dados chama a atenção para novas tendências do mercado em *softwares open source*. Os aplicativos GIS fornecem técnicas avançadas para tratamento de dados espaciais junto com a utilização das funcionalidades e aplicabilidades.

A proposta deste trabalho está relacionada com a dinâmica da evolução tecnológica e da crescente demanda por dados geoespaciais e a evolução de *softwares* que integrem, de forma prática, possibilidades de análises espaciais. As ferramentas de GIS e GeoBI serão aqui pesquisadas e trabalhadas, viabilizando um estudo de caso.

Segundo Alonso *et. al.* (2011), a tecnologia da informação geográfica relacionada com as ciências, possibilita a exploração de dimensões temporais e espaciais, elevando a capacidade de processamento e troca da informação, resultando em melhor divulgação, tomada de decisão e ação. Esses resultados fazem com que o planejamento tenha maior evolução no resultado, incentivando a utilização do SIG por diversas partes e setores.

As fases de planejamento, monitoramento e gestão podem ser integrados a partir da utilização de modelos e aplicações quando da aferição e uso das informações geográficas. Tais fases podem necessitar da aquisição, análise e divulgação de uma grande quantidade de dados para tanto, a integração entre o GIS e GEOBI se torna uma ferramenta estratégica, moderna e capaz de mostrar resultados temporais de uma vasta quantidade de dados.

Utilizar de habilidades, conhecimento, tecnologias, qualidade, risco e questões de segurança são alguns dos objetivos do BI. Tais objetivos são alcançados a partir da coleta, integração, interpretação e apresentação de informações através das aplicações de BI. Para Percivall & Singh (2011), BI é a aquisição de informação a partir de dados ou o conhecimento resultante das análises dos mesmos.

A pesquisa pela utilização de ferramentas que possibilitem integrar tecnologia, opções de análises espaciais e softwares *open source* é de grande importância para a evolução das questões geoespaciais.

Para a execução deste estudo, foi proposta uma investigação a cerca do tema, realizada em etapas, alcançando os objetivos específicos supracitados. A primeira etapa foi a realização de pesquisa bibliográfica acerca do tema GeoBI, focando na utilização de ferramentas de conexão, análise espacial e tratamento espacial de dados. Esta etapa possibilitou o entendimento da tecnologia de *Business Intelligence* (BI) e suas aplicações espaciais em GeoBI, possibilitando a implementação de um sistema que integre o BI e o SIG, utilizando de ferramentas de análise espacial.

O GeoBI integra algumas funcionalidades do GIS e do BI, possibilitando análises multidimensionais, geração de relatórios em tabelas, diagramas e gráficos, relacionados com os mapas, além da extração, transformação e carga no *Data Warehouse* (DW) de vários tipos de dados, incluindo a sua geometria.

Como segunda etapa, análise e descrição, de forma sucinta, das ferramentas de análises espaciais disponíveis na aplicação Geokettle, que integra o pacote de ferramentas do Pentaho *Data Integration* (PDI).

A terceira etapa aplicou as etapas anteriores em um estudo de caso utilizando as ferramentas de análises espaciais encontradas e verificou na prática as suas potencialidades. O estudo de caso desenvolveu uma aplicação de ETL (*Extract, Transform and Load*), utilizando um banco de dados composto por registros de espécies da fauna a fim de auxiliar a análise de uma grande quantidade de dados disponíveis, aumentando a eficiência de relatórios, com menor tempo de resposta, facilitado o acesso às informações.

Por fim, avaliou a eficiência das ferramentas de análises espaciais utilizadas, em relação aos resultados obtidos com o estudo de caso, possibilitando propostas futuras de trabalhos.

2. Objetivo Geral

Aplicar as principais ferramentas de análise espacial da extensão Geokettle em registros de espécies da fauna.

Objetivos Específicos

- Realizar pesquisa bibliográfica sobre GeoBI e suas funcionalidades;
- Analisar e descrever as principais ferramentas disponíveis de análise espacial do Geokettle;
- Aplicar um estudo de caso utilizando as ferramentas estudadas;
- Avaliar a eficiência das ferramentas em relação aos resultados obtidos.

3. Referencial Teórico

Por se tratar de aplicação utilizando o GeoBI, entender o que é um banco de dados é de fundamental importância. Segundo MySQL (2011), banco de dados é a estruturação de uma coleção de dados que pode ser de qualquer tema que seja importante para o usuário, por exemplo, um cadastro de empresas, uma lista de espécies de fauna. A edição, o acesso e o processamento dessas informações necessitam de um sistema de gerenciamento de banco de dados como o Servidor MySQL. A manipulação de grandes volumes de dados é facilitada pelo uso de computadores, e o sistema de gerenciamento de dados tem a função de ser o núcleo de todas as manipulações, podendo ser independente ou compor um conjunto de aplicações.

O banco de dados da fauna foi concebido em ambiente de linguagem de programação PHP com banco de dados MySQL. PHP, segundo *The PHP Group* (2011), *PHP Hypertext Preprocessor*, originalmente denominado *Personal Home Page*, é caracterizado por ser uma linguagem *Open Source* de *script*, utilizada para diversos fins, principalmente, no desenvolvimento de aplicações web, juntamente com o formato HTML (*HyperText Markup Language*). A linguagem PHP facilita a manipulação, inserindo no seu código comandos para mostrar HTML.

MySQL é um programa cliente/servidor robusto de bancos de dados SQL (*Structured Query Language*), ágil, multitarefa e multiusuário. O Servidor MySQL é utilizado para sistemas de produção com grande volume de dados, inserido em programas de uso por muitos usuários, de acordo com MySQL (2011). A licença utilizada é *Open Source*, sob os termos da GNU (*General Public License*), assim como o PHP. O MySQL é considerado o sistema de gerenciamento de banco de dados SQL *Open Source* mais popular, principalmente, pelas suas funcionalidades e características.

O MySQL é um sistema de gerenciamento de banco de dados relacional, que consiste em armazenar individualmente as tabelas, evitando a aglomeração dos dados em apenas um local. Esta organização eleva a velocidade e a flexibilidade de acesso, possibilitando o relacionamento entre as diferentes tabelas, resultando em respostas

específicas e detalhadas dos dados armazenados. Um *software Open Source* (Código Aberto) possibilita que o usuário modifique o programa da melhor forma, dentro de suas necessidades, para a estruturação de seu banco de dados.

Outro conceito importante a ser tratado, foco principal deste estudo, é o de geoprocessamento, definido pelo Instituto Nacional de Pesquisas Espaciais (2007) como um conjunto de métodos e técnicas computacionais realizado em etapas, direcionados a coleta e tratamento de dados espaciais. A primeira etapa refere-se à aquisição de dados, a segunda ao tratamento dos dados como meio de agregar e compatibilizar informações de diferentes características, com as respectivas posições geográficas por meio de pares de coordenadas, em um Banco de Dados Geográficos (BDG), e a terceira etapa trata-se da representação dos dados através do Sistema de Informações Geográficas (SIG), que permite, a partir dos dados capturados e tratados, perceber a distribuição espacial do tema em estudo, representados em mapas para melhor interpretação e análise.

4. GEOBI e suas Funcionalidades

Business Intelligence (BI) possibilita a tomada de decisão apoiada por técnicas de identificação, extração e análise de dados. Algumas das funções mais comuns em BI, segundo Percivall & Singh (2011) são: *dashboards*, *scoreboarding*, processamento analítico *online* (OLAP), mineração de dados, mineração de processos, análise de negócios de gestão de desempenho e análise preditiva, que fornecem análises de cenários históricos, atuais e futuros.

Data Warehouse

Data Warehouse é conceituada por alguns autores de destaque como William Inmon e Ralph Kimball como um repositório de informação que agrupa dados históricos de forma organizada, em uma base de dados, caracterizado por ser orientado a assunto, integrado, histórico, não-volátil e acessível. O armazenamento possibilita a execução de tarefas como análises e elaboração de relatórios de forma mais prática e fácil.

Por definição, sua estrutura de organização é orientada a assuntos, compartimentada por temas relevantes, de acordo com as necessidades dos usuários finais. Integrado por compor uma única fonte, onde os usuários acessam dados completos sobre o negócio, sem a necessidade de buscar, em outras bases de dados, composto por séries históricas, viabilizando análises de cenários pretéritos, atuais e, possivelmente, prognósticos. O *Data Warehouse* não é volátil por permanecer com os dados da forma como entraram, não sendo apagados ou alterados, mantendo a informação estável. Por fim, a premissa de um *Data Warehouse* é proporcionar o acesso ágil à informação, possibilitando transformar dados de fontes distintas em informações que podem ser analisadas para responder às necessidades do negócio.

Segundo DIAS (s/d),

“Estes dados são obtidos através de um processo chamado de Extração, Transformação e Carga sobre as bases de dados dos sistemas transacionais. Após os dados estarem carregados no Data Warehouse os mesmos estarão disponíveis apenas para leitura, e a atualização desta base de dados se dará de forma incremental.”

A criação de um *Data Warehouse*, para SEZÕES, OLIVEIRA & BAPTISTA (2006), é motivada pela necessidade de agregar dados armazenados em diferentes estruturas, visando análises globais, e a necessidade de tomar decisões selecionando dados para processos usuais de análises e relatórios. Em determinados casos, pode-se utilizar

conjuntos menores de dados agrupados por assuntos ou áreas, de acordo com a necessidade da instituição, denominados de *Data Marts*.

Granularidade

O conceito de granularidade é descrito por Inmon (2005) apud DIAS (s/d) como o nível de detalhe dos dados de um *Data Warehouse*. Quanto maior a granularidade menor são os detalhes, portanto, quanto mais baixa a granularidade maiores serão os detalhes. A granularidade está relacionada com a escala dos dados geográficos, que é dado pelo nível de detalhamento do dado no momento de sua aquisição. Uma escala grande possui um denominador pequeno, mais são os detalhes, e uma escala pequena tem um denominador com valores altos, menores são os detalhes, conforme a figura 01.

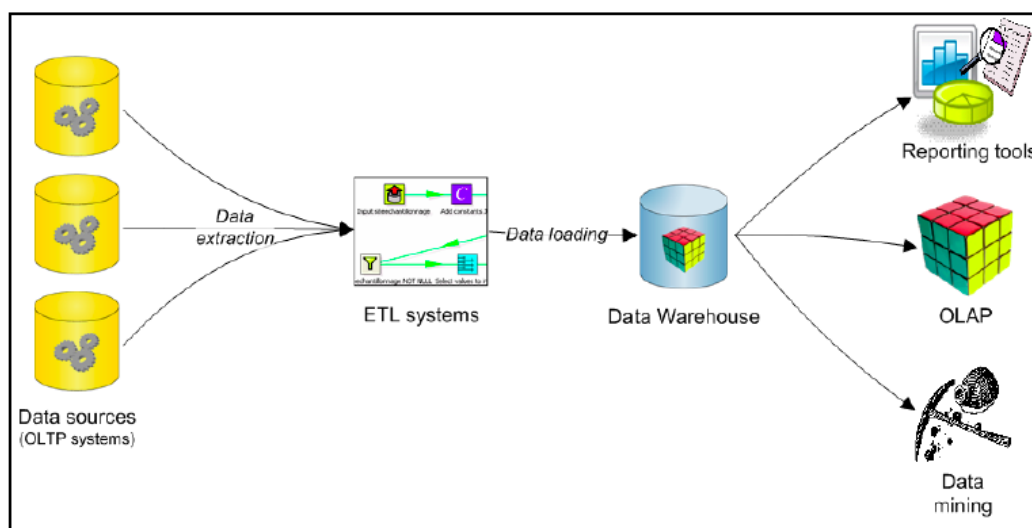


Figura 1 - Arquitetura de um DW

Fonte: Adaptado de BADARD, 2009 apud ANDRADE, 2011

A granularidade interfere no volume de dados e no desempenho do *Data Warehouse*.

A definição do nível de análise é fundamental para que se tenha a definição da granularidade dos dados, pois uma vez com os dados obtidos com granularidades altas, esses não mais poderão ser transformados para níveis de granularidades mais baixas.

Data Warehouse Geográfico

Um *Data Warehouse Geográfico* (DWG), segundo BRITO (2012), ou *Data Warehouse Espacial* (DWE), para DEGGAU & FILETO (s/d), distingue de um DW convencional ou tradicional no armazenamento dos dados de características espaciais, contendo atributos de dimensão ou medidas nas tabelas fatos, com suporte ao tratamento de objetos geográficos, representado pela figura 02. Um DWG combina dados e operadores do modelo dimensional com representações geométricas e operações para manipulação de dados geográficos. Os atributos podem conter suas hierarquias, que são comumente classificadas por *ad hoc* e predefinidas. Segundo BRITO (2012), a hierarquia espacial *ad hoc* agrega objetos espaciais de diversas granularidades, a partir da utilização de um índice espacial que os agrupam, indicando a agregação das medidas numéricas correspondentes, por exemplo, o R-Tree. A hierarquia espacial predefinida condiz com a associação 1:N entre atributos espaciais com granularidades mais elevada e mais baixa, definida por um relacionamento topológico, como contém, está contido, intersecta, etc.

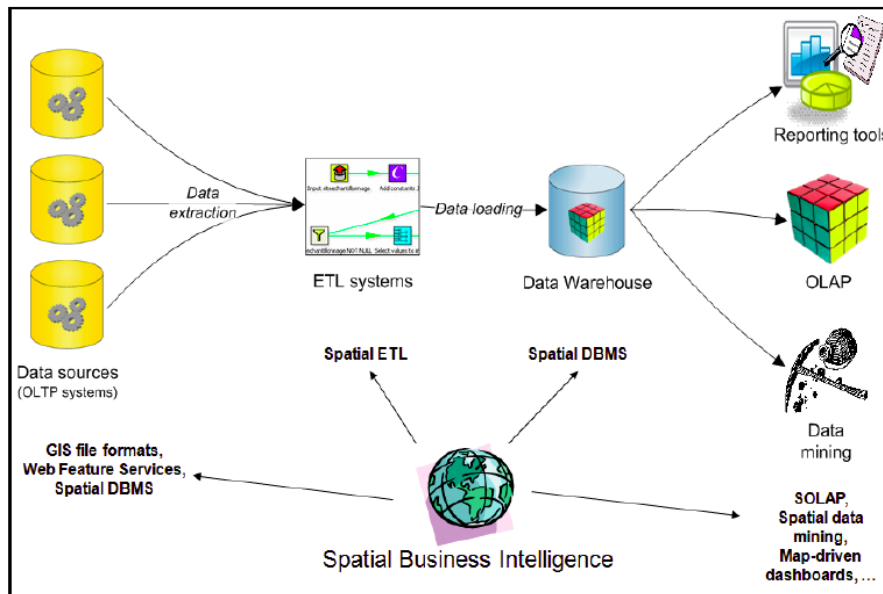


Figura 2 - ARQUITETURA DE UM DWG
 Fonte: Adaptado de Badard (2009) apud ANDRADE (2011)

De acordo com DEGGAU & FILETO (s/d), os DWG apresentam alguns problemas, como integração do modelo de dados dimensional com o espacial, levando em conta características do modelo, operadores e implementação de sistemas integrados; extração, transformação e carga de dados espaciais e; apoio ao processo de geração de *Data Marts* espaciais (DME).

Data Marts Espaciais (DME)

Data Marts Espaciais (DME), segundo DEGGAU & FILETO (s/d), podem ser criados e o seu conjunto pode fazer parte de um DWG, compartilhando dimensões e medidas, mesmo as que referenciam objetos geográficos. Um DME é criado a fim de atender objetivos ou assuntos específicos de análises de dados, incluindo um ou vários cubos de análise de dados, que são modelos multidimensionais, contendo suas tabelas fato e dimensões, com a possibilidade de analisar os dados, utilizando operadores e funções de agregação de dados escalares e espaciais.

Tipos de Dados Espaciais

Os tipos de dados espaciais usados na representação gráfica do mundo real utilizam sistemas de coordenadas onde todas as feições são descritas por pontos, linhas ou polígonos (figura 03), denominadas entidades geométricas. Ex: limites municipais, estaduais, hidrografia, vegetação, sedes municipais, bacias hidrográficas, etc.

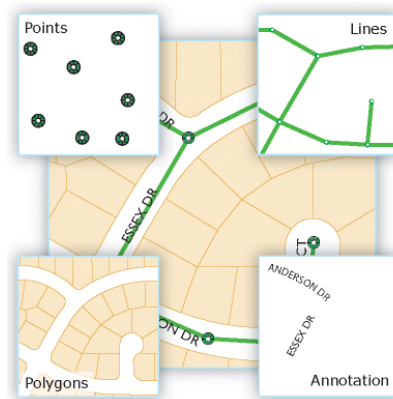


Figura 3 - Exemplo de Feições
Fonte: ESRI, 2011

Ponto é considerado a entidade mais primitiva na representação de objetos espaciais, por não possuir dimensão (0-dimensional), representado por um par de coordenadas (x, y) .

Linha é formada pelo conjunto de dois ou mais pontos (1-dimensional). Os pontos que iniciam e terminam uma linha são os nós, e os pontos intermediários da linha são os vértices. Polígono é um conjunto de pontos conectados, que se inicia e termina no mesmo ponto (2-dimensional.). (Ver figura 04)

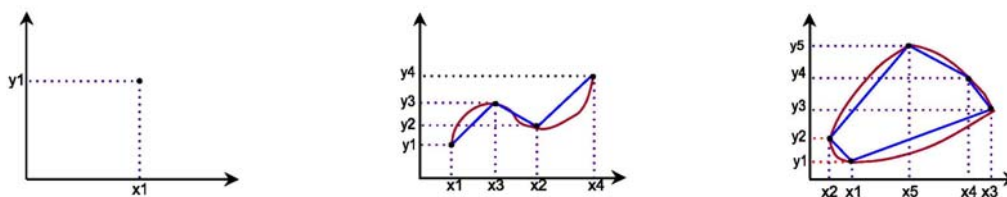


Figura 4 - Representação dos tipos de dados espaciais

Segundo BRITO (2012), para analisar um predicado espacial a consulta deve ser realizada em duas fases: filtragem e refinamento, pois os objetos geográficos podem conter uma centena de pontos. A primeira fase consiste em diminuir a quantidade de dados para análise, denominada filtragem, utilizando, por exemplo, o *minimum bounding rectangle* (MBR), seleciona objetos candidatos. Por fim, é realizado o refinamento, onde os candidatos selecionados são analisados de acordo com a exata geometria e descartados os candidatos que não estão no relacionamento topológico, podendo ser aplicados vários refinamentos. Estas fases possibilitam um maior desempenho.

Consultas SOLAP

No DWG, além da utilização de operadores OLAP (*Online Analytical Processing*) tradicionais, possibilitando filtros e agrupamentos de informações, permite também o uso de funções de agregação de dados, denominadas SOLAP (*Spatial Online Analytical Processing*), aplicados a operadores geográficos e funções de agregação de dados espaciais nas análises, possibilitando a representação em mapas. As consultas espaciais são classificadas com base nas restrições, como *point queries*, *region queries*, *range queries* e consultas com junção espacial.

Range Queries

As consultas espaciais por *range queries* selecionam, a partir de um intervalo, podendo ser o retângulo envolvente do objeto espacial de referência, objetos espaciais.

Segundo BRITO (2012), existem três subtipos principais de classificação de *Range Queries*: *intersection range query* (figura 05), seleciona os objetos espaciais que tocam o intervalo consultado; *containment range query* (figura 06), seleciona apenas os objetos que estão totalmente dentro do intervalo consultado; e por fim, *enclousure range query* (figura 07), que seleciona os objetos que contém o intervalo.

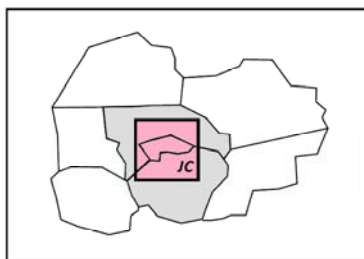


Figura 5 - INTERSECTION RANGE QUERY

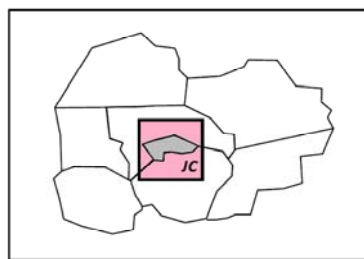


Figura 6 - CONTAINMENT RANGE QUERY

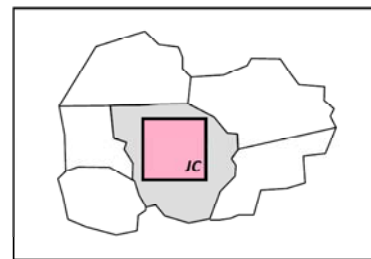


Figura 7 - ENCLOUSURE RANGE QUERY

Fonte: BRITO (2012)

Pentaho

O Pentaho *Open BI* é uma plataforma composta por um conjunto de aplicativos *open source* para criar soluções de *Business Intelligence* (BI). O Pentaho possui módulos com ferramentas específicas e os mais importantes são: *Pentaho BI Platform*, *Pentaho Data Integration* (PDI) (Kettle), *Analysis View* (Mondrian), *Pentaho Reporting* (Reporting) e Weka (Data Mining).

Os tipos de licenciamento do Pentaho são *Community* e *Enterprise*, através do *Pentaho Public License* (PPL). A versão *Community*, com ferramentas básicas e essenciais, é

gratuita e pode ser baixada do site da Pentaho, com constantes atualizações de versões. Para a versão *Enterprise*, a licença é comercializada e disponibiliza um conjunto de ferramentas integradas e melhor desenvolvidas, contando com suporte técnico, manutenções e atualizações frequentes.

Neste estudo de caso, será enfatizado o uso da extensão Geokettle, integrada ao *Pentaho Data Integration* (PDI). O PDI tem como objetivo transformar dados a partir do banco de dados relacional para um banco de dados dimensional, de forma automatizada, através de sua modelagem.

A modelagem da PDI é realizada a partir de um conjunto de operações executadas nos fluxos dos dados. As operações são representadas por blocos e a conexão dos dados entre dois blocos são identificadas por setas, ligando o bloco fornecedor e o bloco receptor. Segundo GUTUYAMA & SACILOTTO (2009), na PDI, para a elaboração de um projeto, são necessárias operações de: *Table input* (entrada do tipo tabela), realiza a leitura de uma tabela em uma base de dados relacional e insere o seu conteúdo ao fluxo de dados; *Lookup* (consulta), de acordo com um campo existente no fluxo, realiza consulta na fonte de dados para buscar um valor semelhante a um do fluxo; *Table output* (saída do tipo tabela), salva uma tabela de um banco de dados relacional ao conteúdo do fluxo; *Lookup/Update Combination* (combinação de consulta e atualização), busca em uma base de dados relacional e verifica o valor encontrado com o fluxo, havendo diferença, atualiza a base de dados.

Geokettle

O GeoKettle é uma aplicação ETL (*Extract, Transform and Load*), direcionada para metadados e disponibiliza ferramentas espaciais voltadas à integração de diferentes fontes para a construção e atualização de bases de dados geoespaciais, armazenamento de dados e serviços da web., conforme figuras 08 e 09.

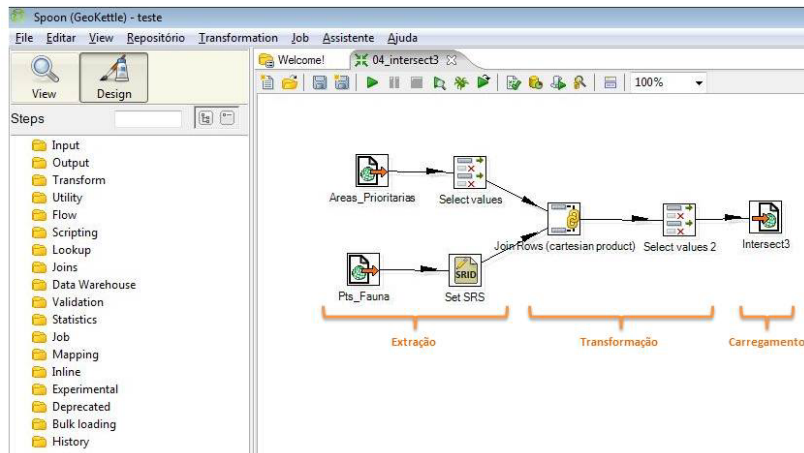


Figura 8 - EXEMPLO DE ETAPAS DO ETL NO GEOKETTLE

É uma extensão do Pentaho *Data Integration* (PDI), também denominado Kettle, que permite extrair dados de diversas fontes, transformar, corrigir erros, modificar a estrutura, deixando compatível com padrões de dados e carregar os dados transformados em um sistema de gerenciamento de bases de dados (*DataBase Management System* - DBMS), um arquivo para o SIG ou para serviços de web geoespacial.

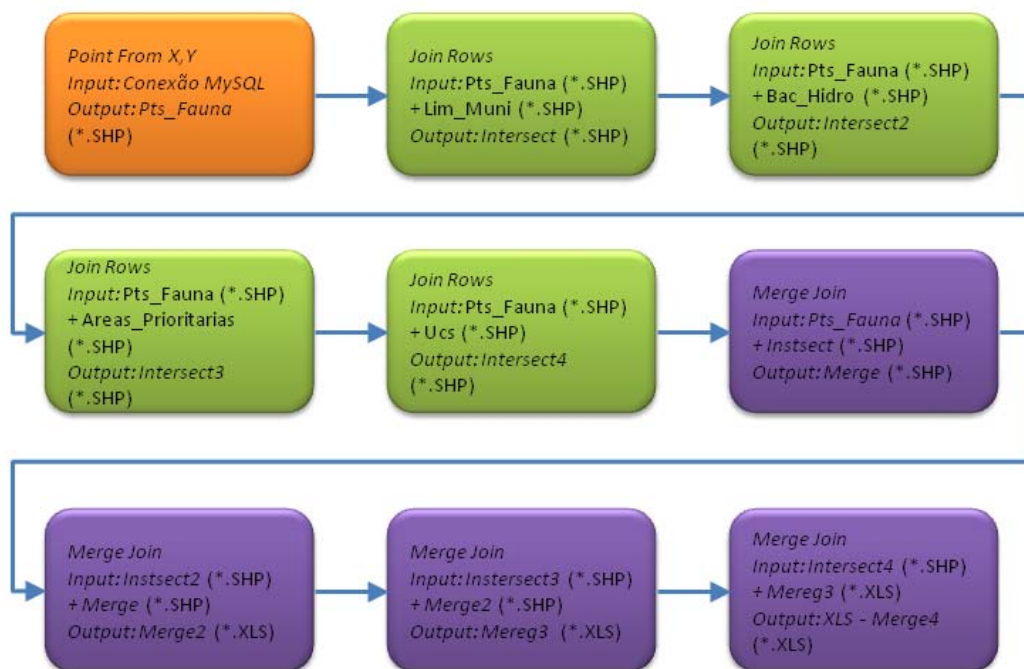


Figura 9 - ESTRUTURA DE ORGANIZAÇÃO DO ETL

A automatização de processos completos e repetitivos de forma prática, sem a necessidade de utilizar códigos específicos, são algumas das vantagens do GeoKettle.

Geokettle possibilita extrair dados de mais de 35 tipos de bases de dados, formatos de dados geoespaciais, SOLAP, ferramentas de transformação de dados como: *select*, *split*, *merge*, *join*, *duplicate* e *cluster*. Leitura de dados em diferentes formatos de bases de dados como MySQL, PostgreSQL/PostGIS, Oracle; formatos de dados geoespaciais como *Shapefile*.

O tipo de licenciamento do Geokettle é pela *Lesser General Public License* (LGPL), com linguagem de programação e bibliotecas Java, JTS, GeoTools. Possibilita o uso de vetores, sendo quatro tipos de arquivos SIG, mas não suporta arquivos raster. Disponibiliza mais de 25 operações de análise espacial, mas não existe uma visualização integrada dos resultados cartográficos gerados.

No Geokettle existem alguns conceitos fundamentais para iniciar o seu uso, *Transformation* que corresponde aos passos para realização em uma sequência lógica da transformação, e o *Job* que controla a execução de uma ou várias sequências de *Transformation*, podendo ser iniciada manualmente ou de forma automática.

Na elaboração do ETL, os *Steps* e *Hops* fazem parte da estrutura, a ligação entre duas transformações é realizada pelos *Steps*, e cada transformação é denominada por *Hops*, conforme figura 10.

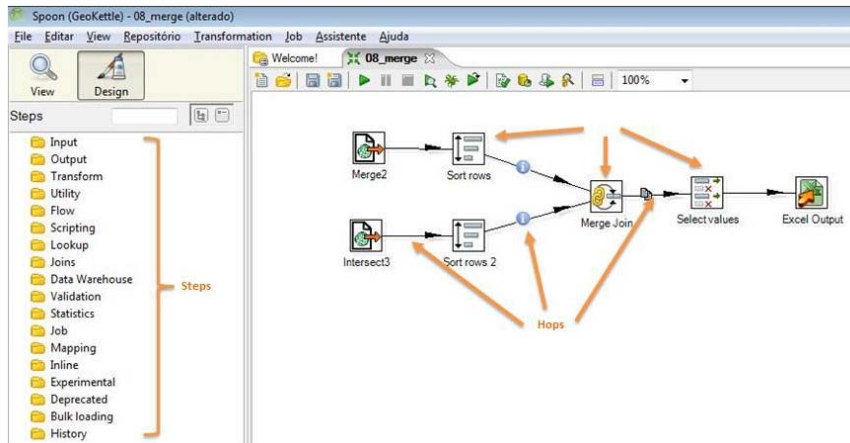


Figura 10 - CONFIGURAÇÃO DO GEOKETTLE

As principais *Steps* do Geokettle são: *Input* (entrada dos dados), *Output* (saída dos dados), *Transform* (transforma os dados), *Flow* (fluxo de dados), *Utility* (utilitários), *Scripting* (executa *scripts*), *Lookup* (consulta dos dados), *Joins* (junções entre fontes), *Data warehouse* (armazenamento de dados) e *Validation* (validação),.

5. Metodologia

A fim de aplicar os conceitos dispostos anteriormente, a modelagem dimensional dos dados é parte integrante ao processo de execução das etapas. O Modelo Dimensional, segundo GUTUYAMA & SACILOTTO (2009), é uma modelagem particular, onde organiza os dados numéricos, agregando-os conforme as suas semelhanças, considerando as agregações como dimensões, conforme figura 11. Primeiramente, foram identificados os dados para a execução dos eventos, fundamentais nas regras de negócio, agrupados na tabela fato, modelados nas dimensões que contém seus atributos. Finalizado o levantamento dos atributos, representa-se através do modelo dimensional seus respectivos atributos detalhados.

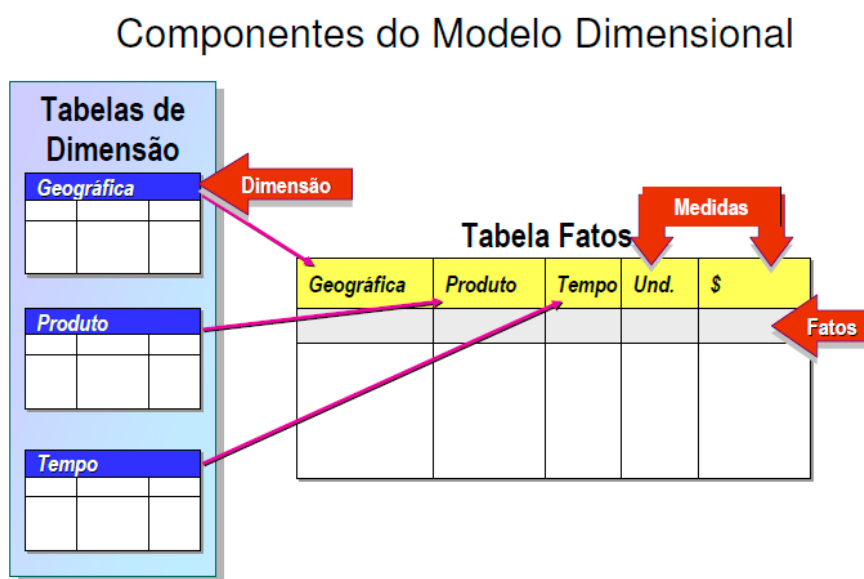


Figura 11 - EXEMPLO DE CONFIGURAÇÃO DE UM MODELO DIMENSIONAL

Selecionou-se o modelo estrela como padrão para elaboração do modelo dimensional, segundo GUTUYAMA & SACILOTTO (2009), em projetos com poucas dimensões hierarquizadas, não há prejuízo na normalização dos dados. A principal vantagem do modelo estrela está na simplicidade de elaboração e acesso rápido aos atributos de

uma determinada dimensão, apesar do baixo nível de normalização.

A ligação dos registros da tabela fato às suas dimensões é realizada pelos campos denominados de chave estrangeira que, segundo GUTUYAMA & SACILOTTO (2009), apesar de não ter significado de negócio, possui uma importância de associação entre entidades de forma mais rápida e simplificada. As tabelas auxiliares são as tabelas dimensão, relacionadas à tabela fato, com os atributos de cada dimensão do negócio.

A elaboração do ETL utilizou de ferramentas dos *Steps de Input, Flow, Transform Join Output* da aplicação Geokettle. O modelo lógico do banco de dados da fauna foi desenvolvido no aplicativo BRModelo, de licença *open source*.

As bases de dados vetoriais utilizadas foram (figura 12): limite municipal do Instituto Brasileiro de Geografia e Estatística (IBGE, 2010) na escala de 1:250.000, bacia hidrográfica estadual (unidades de planejamento do IGAM – Instituto Mineiro de Gestão das Águas) e bacia hidrográfica federal do IGAM (2002) na escala de 1:100.000, áreas prioritárias para a conservação da Fundação Biodiversitas (DRUMMOND *et al.*, 2005), e unidades de conservação do Instituto Estadual de Florestas de Minas Gerais (IEF, 2013). A projeção de todos os arquivos é do sistema de coordenadas geográficas, datum SAD69.

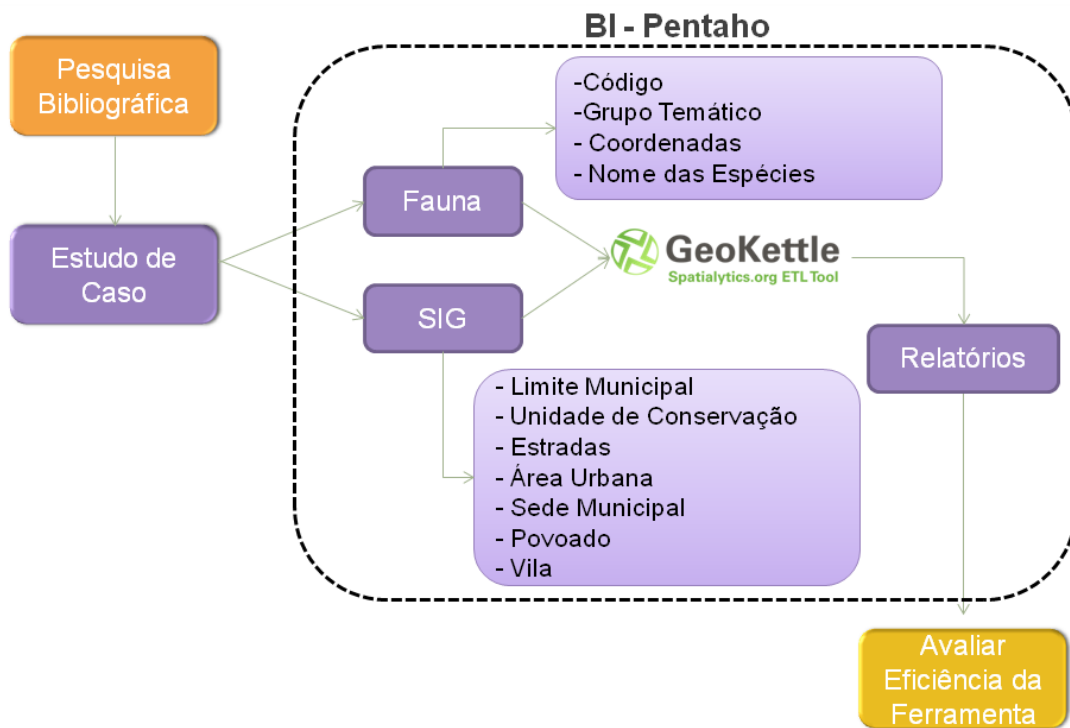


Figura 12 - FLUXOGRAMA DA METODOLOGIA APLICADA

6. Estudo de Caso

O estudo de caso utilizou ferramentas de análises espaciais e verificou na prática as suas potencialidades. O estudo de caso desenvolveu um ETL utilizando um banco de dados composto por registros de espécies da fauna, envolvendo os grupos temáticos de Invertebrados, Ictiofauna, Herpetofauna, Avifauna e Mastofauna.

A aplicação visa auxiliar e facilitar as análises de uma grande quantidade de dados, aumentando a eficiência de relatórios, com menor tempo de resposta, agilizando o acesso às informações de nome municipal, bacia hidrográfica estadual (unidades de planejamento do IGAM – Instituto Mineiro de Gestão das Águas), bacia hidrográfica federal, áreas prioritárias para a conservação e unidades de conservação.

A inserção correta e validação dessas informações se tornou um ponto de relevância, a partir do momento em que dados eram extraídos do banco de dados utilizando como filtros tais campos. A veracidade do dado só pode ser consolidada a partir do momento em que a inserção não tenha nenhum erro ou haja validações dos dados. Uma das formas de validação é através do ETL criado através do Geokettle.

O banco de dados da fauna, de onde forma extraídos os dados da fauna tem a seguinte estrutura, de acordo com o modelo lógico apresentado na figura 13.

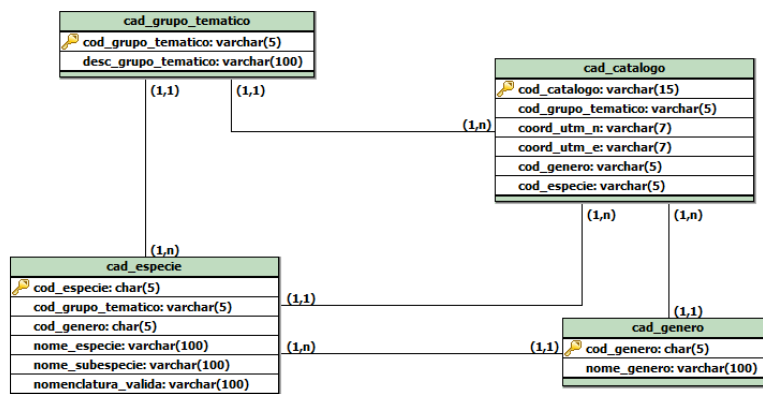


Figura 13 - MODELO LÓGICO DO BANCO DE DADOS DA FAUNA

A tabela 'cad_grupo_tematico' contém os nomes dos grupos temáticos e seus respectivos códigos (cod_grupo_tematico), neste caso é chave primária. O 'cad_genero' contém os nomes dos gêneros e seus códigos (cod_genero) como chave primária. No 'cad_especie' além dos nomes das espécies e subespécies, e da chave primária (cod_especie), ainda tem o relacionamento com o grupo temático, gênero e a nomenclatura válida, que é a junção do nome do gênero com o nome da espécie mais a subespécie, esse campo foi utilizado no ETL para extração dos dados. Por fim, no banco de dados da fauna temos o código do 'cod_catalogo' (chave primária), o grupo temático, as coordenadas lat/long, o código do gênero e da espécie.

As ferramentas utilizadas na estruturação do modelo seguiu a organização da extração, transformação e carregamento dos dados. Na extração utilizou os Steps de *Input: Table Input, Shapefile file Input e Excel Input*. Para a transformação utilizou do Step *Flow o Filter rows*; do Step *Transform o Select Values, Calculator, Set SRS e Sort rows*, além do Step *Join* as transformações de *Join Rows (cartesian product)* e *Merge Join*. E por fim, o carregamento dos dados extraídos e transformados empregou do Step *Output o Excel Output, Shapefile file Output e XML output*, conforme figura 14.

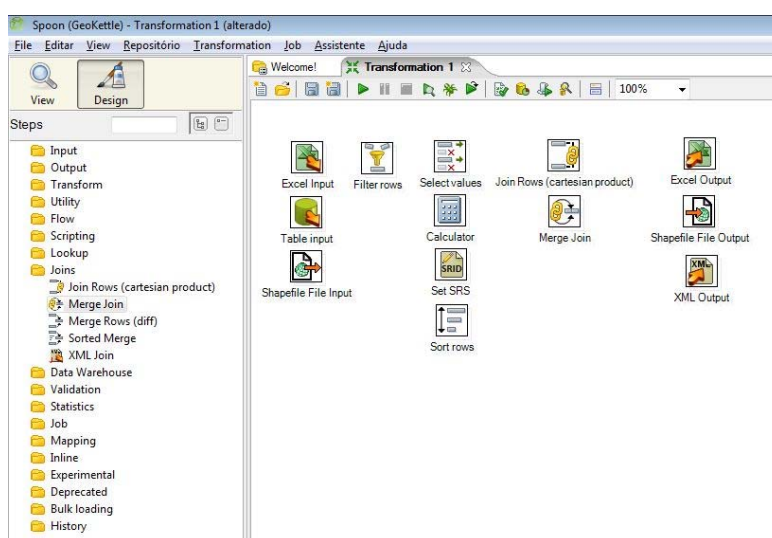


Figura 14 - FERRAMENTAS UTILIZADAS NO ETL

Table Input realiza a leitura de tabelas de um banco de dados convencional ou geográfico, utilizando conexão SQL. A busca de geometrias podem ser ativadas a partir da ativação de versões espaciais de banco de dados, e as declarações do SQL são geradas de forma automática.

Shapefile file Input permite a leitura e visualização da tabela de atributos (DBF) e das feições espaciais, em um visualizador primitivo, com a possibilidade de *zoom* e alteração das cores das simbologias.

O *Excel Input* possibilita realizar a leitura de um ou mais arquivos nas extensões XLS, XLSX e ODS (do *OpenOffice Workbooks*). Por padrão, a extensão da planilha está definida como Excel 97-2003 (XLS), no caso da necessidade em ler outras extensões, as mesmas devem ser selecionadas no momento da inserção. Podem ser utilizadas planilhas protegidas, com funções especiais, arquivos com várias planilhas, mas as características devem ser alteradas na aba *Content* e *!Sheets*.

Filter rows filtra linhas baseado em comparações convencionais e espaciais. Esta ferramenta só é possível ser utilizada a partir da inserção de alguma entrada de dados, onde tenha pelo menos um *Hope*. Existe a possibilidade de trabalhar com mais de uma condição, selecionando o campo de referência, a condição e o campo de comparação. Existem dezenas de condições disponíveis, sem as básicas para qualquer tipo de consulta.

Select Values transforma os dados através da renomeação, alteração do tipo de configuração, além de configurar o comprimento e precisão dos campos no fluxo (aba *Select & Alter*), exclusão de campos da tabela (aba *Remove*) e altera o nome, tipo, tamanho e precisão do metadado (aba *Meta-data*).

A ferramenta *Calculator* disponibiliza algumas funções predefinidas para serem executadas nos valores dos campos de entrada, possibilita também o uso de funções para operações geométricas mais comuns, como *buffer*, *centroid*, *distane*, etc. A execução de funções nesta ferramenta se torna mais ágil se comparada com funções executadas em *scripts* na linguagem JavaScript.

Uma ferramenta importante no Geokettle é o *Set SRS*, onde pode escolher o sistema

de referência das coordenadas geográficas de arquivos vetoriais, como em *shapefiles*. São disponibilizados códigos das projeções seus respectivos datum. Para a identificação do código correto e os detalhes do datum pode utilizar a ferramenta *SRS Transformation*, do *Step Transform*, para verificar as configurações através do botão *Search*, e do *Details*, após a seleção de algum sistema de referência espacial.

Sort rows ordena os campos de forma ascendente ou descendente, classificando as linhas dos dados. A ordenação utiliza de arquivos temporários, que são configurados como padrão até 1 milha de linhas.

Join Rows (cartesian product) combina as linhas das entradas utilizando de operadores espaciais como: *intersects*, *equals*, *crosses*, *disjoint*, *within*, *overlaps*, *touches*, etc.

Merge Join possibilita mesclar um conjunto de dados de duas entradas distintas, juntando através de campos chaves, que devem ser ordenadas de forma ascendente, utilizando da ferramenta *Sort rows*, através de um *Hope*. Se ocorrer a duplicidade dos nomes dos campos chaves das duas etapas o nome do campo resultante terá a formatação \$key_1. Estão disponíveis quatro tipos de *Join*: *INNER*, *LEFT OUTER*, *RIGHT OUTER* e *FULL OUTER*. *INNER join* extrai apenas as linhas das entradas que possuem a mesma chave; *LEFT OUTER* extrai todas as linhas da primeira etapa selecionada e *RIGHT OUTER* extrai todas as linhas da segunda etapa selecionada, ambas disponibilizando valores nulos para linhas cuja a chave não correspondentes a etapa selecionada. Por fim, o *FULL OUTER*, independente da etapa, de haver ou não a chave da linha, todas serão disponibilizadas nos resultados, com a valores nulos para aquelas que não tiverem relacionamento com sua respectiva chave.

As ferramentas *Excel Output*, *Shapefile file Output* e *XML output* permitem salvar um ou mais arquivos em suas respectivas extensões. Para os arquivos *shapefile* é necessário inserir o código da extensão, após o nome digitado do novo arquivo, para que a ferramenta funcione (*.SHP).

O desenvolvimento do ETL utilizou de ferramentas de extração (*extract*), transformação (*transform*) e carregamento dos dados (*load*). O fluxograma com as principais ferramentas utilizadas, arquivos e suas respectivas extensões de entrada e

saída, são apresentados na figura 15. Ressalta-se que as *Steps*, *Hopes* e etapas propostas podem ser variáveis para cada usuário, podendo ser mais direta ou mais extensas, de acordo com os conhecimentos de cada um.

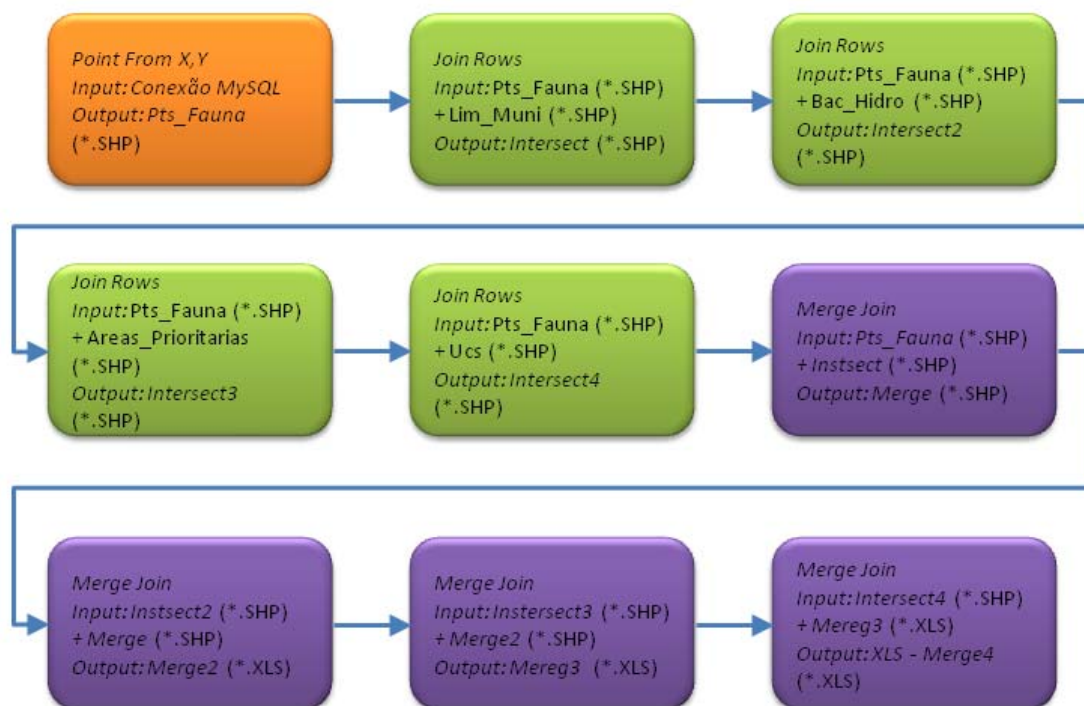


Figura 15 - FLUXOGRAMA DE DESENVOLVIMENTO DO ETL

A ferramenta utilizada de entrada dos dados foi a *Table input*, que busca o conteúdo na base de dados relacional denominada 'BD_Fauna' e realizada a leitura da tabela 'cad_catalogo', adicionando ao fluxo de dados. A seleção dos campos necessários utilizou da linguagem SQL para executar um *Select* dos campos 'Código', que contém a codificação única para cada registro, o 'Grupo temático', que distingue cada registro entre os sete grupos faunísticos. Foram selecionados também os campos 'Espécie', que descreve a nomenclatura taxonômica de cada registro, e por fim as informações de coordenadas 'X' e 'Y'.

O *Filter rows* foi utilizado para criar condições de filtros como ausência de coordenadas. Com o *Select Values* foram renomeados os campos selecionados, no *Calculator* foi criado o campo do tipo *geometry* onde calculou, através da expressão

Point from X,Y as coordenadas geográficas da tabelas 'BD_Fauna'. Neste caso foram utilizadas coordenadas geográficas geodésicas, no formato graus decimais, mas podem ser utilizadas coordenadas geográficas planas, no formato métrico, mas é necessária atenção ao realizar transformações destes dados com outros em formato geodésico. As ferramentas da etapa de extração dos dados podem se observadas na figura 16.

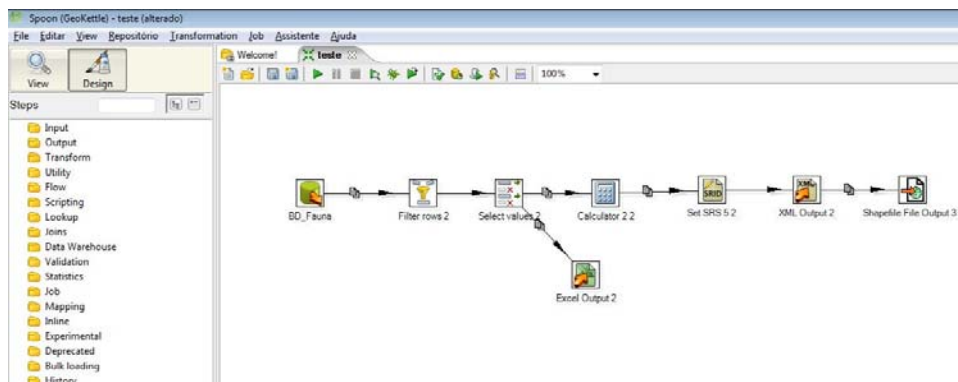


Figura 16 - FERRAMENTAS DA PRIMEIRA ETAPA DE EXTRAÇÃO DE DADOS

No *Set SRS* configura o sistema de projeção geográfica do dado de saída, através do uso do código do EPSG, para o campo geometria criado, neste caso denominado por 'Ponto'. Finalmente, a saída no XML para exportar no formato *shapefile*.

O resultado nesta etapa foi um arquivo vetorial, da feição pontos, com as colunas selecionadas do 'BD_Fauna', no sistema de projeção SAD69 (deg), conforme detalhes apresentados na figuras 17.

```

WKT details
GEOGCS["SAD69 (deg)",
  DATUM["South American Datum 1969",
    SPHEROID["GRS 1967 Modified", 6378160.0, 298.25, AUTHORITY["EPSG","7050"]],
    TOWGS84[-66.87, 4.37, -38.52, 0.0, 0.0, 0.0, 0.0, 0.0],
    AUTHORITY["EPSG","6618"]],
  PRIMEM["Greenwich", 0.0, AUTHORITY["EPSG","8901"]],
  UNIT["degree", 0.017453292519943295],
  AXIS["Geodetic longitude", EAST],
  AXIS["Geodetic latitude", NORTH],
  AUTHORITY["EPSG","66186405"]]
  
```

As etapas de transformação serão descritas a partir de dois exemplos, um de junção e outro de merge, pois utilizam as mesmas ferramentas, alterando apenas os arquivos de saída, que são consecutivamente inseridos nas etapas posteriores como entrada de dados.

As quatro etapas que realizaram junção entre os dados de entrada, usaram como padrão o arquivo de pontos da primeira etapa, com os dados de espécie, coordenadas, código e grupo temático, e individualmente usaram os arquivos de polígono de limite municipal, bacia hidrográfica estadual e federal, áreas prioritárias para a conservação e as unidades de conservação. Com o *Select values* foi alterado o nome do campo 'the_geom' para Geometria, pois é o mesmo nome do campo do arquivo de pontos e na *Hope Join rows (cartesian product)* o Geokettle não separa quais são dos dados de entrada, nem consegue distinguir campos com nomes iguais. Em *Join rows (cartesian product)* foram selecionados os campos de geometria 'the_geom' e 'Geometria', utilizando a função 'GIS_INTERSECTS', para a condição de comparação e junção dos campos dos arquivos de entrada. Existem diversas funções que podem ser utilizadas, conforme descritas anteriormente.

Após o *join*, para não ter uma lista grande de campos desnecessários ao processo, que vieram com os arquivos de polígono, foram deletados, realizando uma limpeza antes de salvar os arquivos. Neste caso é indispensável apontar os campos que devem permanecer, através da aba *Select & Alter*, apenas indicar os campos que deseja remover na aba *Remove*. As ferramentas da etapa de transformação dos dados podem ser observadas na figura 18.

Como resultado, foram gerados quatro arquivos contendo o código dos catálogos, e, de acordo com cada arquivo de entrada, as informações de nome de município, nome das bacias hidrográficas estaduais e federais, áreas prioritárias e unidades de conservação, que intersectam os pontos de fauna. A quantidade de pontos gerados em cada um dos arquivos é variável, de acordo com a espacialização dos pontos, por isso teve valores iguais para os cruzamentos com os limites municipais e bacias, por serem dados

contínuos, ou seja, todo catálogo está inserido em um único município e bacia, e uma variação em relação às áreas prioritárias e unidades de conservação, pois os pontos podem ter nenhuma ou várias interseções. Este resultado afetará as etapas posteriores e, principalmente, as funções a serem utilizadas.

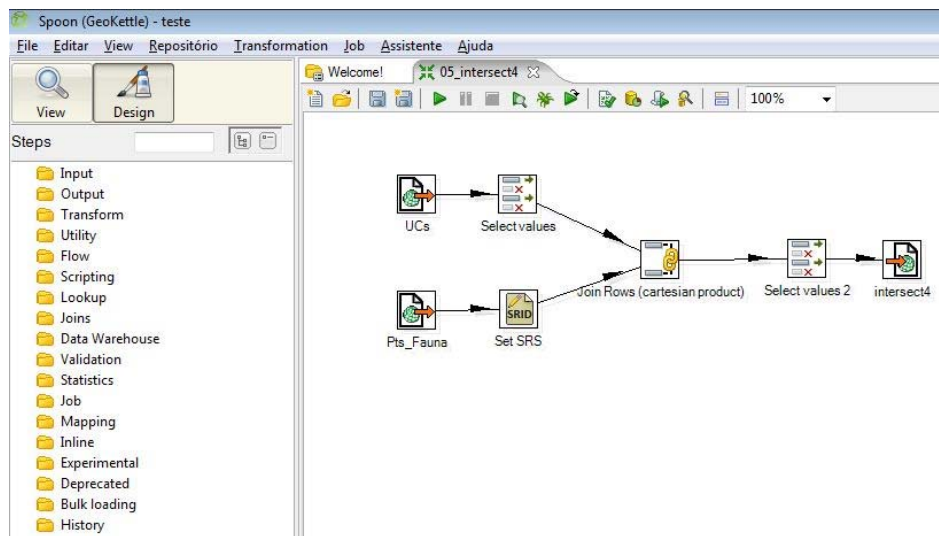


Figura 18 - EXEMPLO DA ETAPA DE TRANSFORMAÇÃO ATRAVÉS DO JOIN

As etapas finais de transformação dos dados utilizaram a ferramenta *Merge join* para a junção das informações e os arquivos de *input* e *output* variaram para estas etapas. Os arquivos vetoriais no formato *shapefile* de ponto foram inseridos utilizando a ferramenta *Shapefile Input*. Os campos utilizados como chave, código do registro, foram classificados através da ferramenta *Sort rows*, na ordem ascendente, possibilitando o uso da ferramenta *Merge Join*. (Ver figura 19)

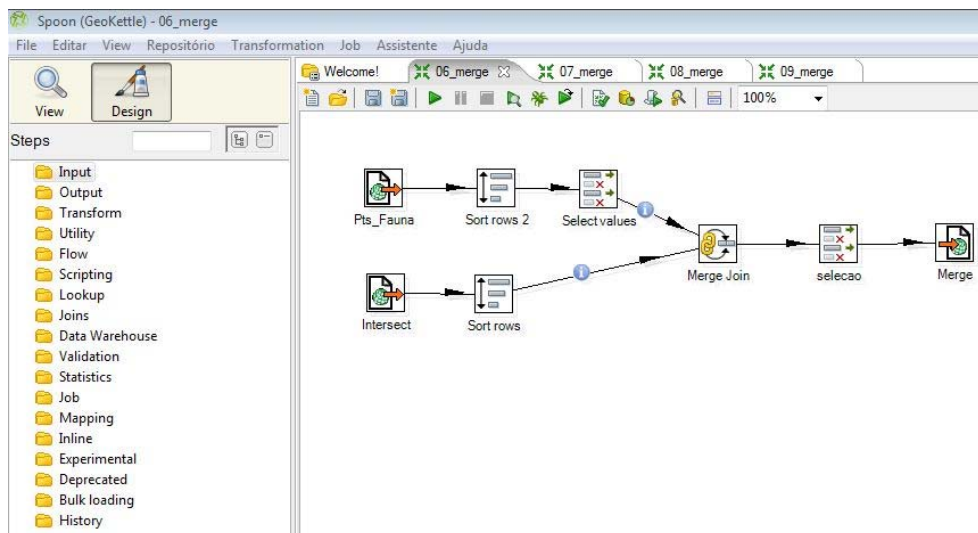


Figura 19 - EXEMPLO DA ETAPA DE TRANSFORMAÇÃO ATRAVÉS DO MERGE JOIN – INPUT SHP E OUTPUT EXCEL

Na primeira transformação utilizando a ferramenta *Merge join* foi necessário aplicar uma alteração no nome do campo 'Cat' para 'Catalogo', pois o mesmo nome foi utilizado no *shapefile Intersect*, e para selecionar os campos chave devem ter nomenclaturas distintas. Após a seleção dos campos e da seleção do tipo de *join*, que utilizou o *INNER join* e o *RIGHT OUTER*, resultando, respectivamente, apenas linhas em que os valores dos códigos eram iguais; e todas as linhas de referência as quais estão na primeira entrada como referência, desconsiderando a existência ou não deles na segunda tabela, inserindo linhas com campos em branco. Em duas etapas foram utilizadas tabelas de *input* na extensão Excel, pois não foi encontrada uma solução para o erro encontrado na exportação em *shapefile*, e as tabelas de *output* destas etapas também foram em Excel.

Este processo de *input* e *output* com arquivos Excel foi necessário, pois os arquivos de referência tinham números distintos de linhas (registros) devido aos cruzamentos anteriores, que foram gerados utilizando as áreas prioritárias para a conservação e unidades de conservação. Com os arquivos em Excel gerados poderiam ser transformados em *shapefile*, através da ferramenta *Shapefile file output*, mas neste estudo de caso não interferiu nos resultados para responder aos objetivos. (Ver figura 20)

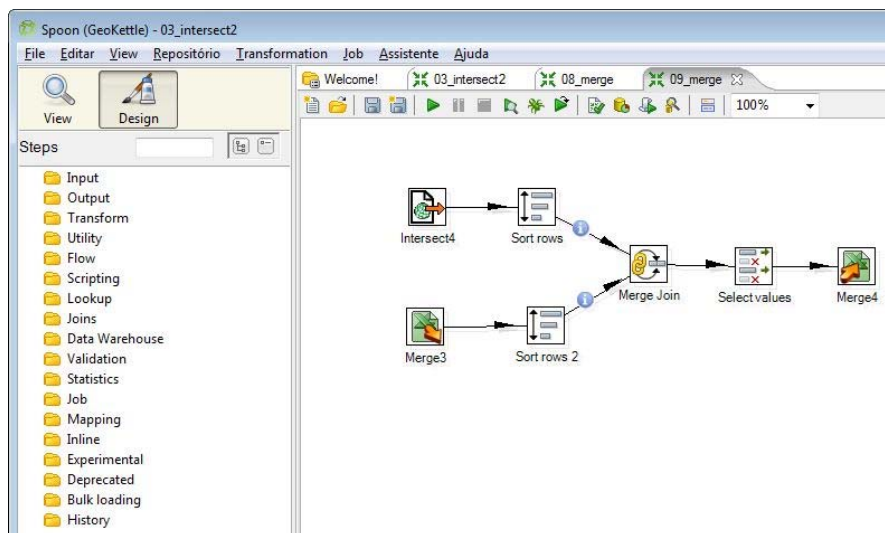


Figura 20 - EXEMPLO DA ETAPA DE TRANSFORMAÇÃO ATRAVÉS DO MERGE JOIN – INPUT E OUTPUT EXCEL

Como resultado final, foi gerada uma planilha (figura 21) contendo o número de linhas (registros) idêntico à tabela original extraída do Banco de Dados. Nem todos os campos foram preenchidos, pois os registros não correspondiam à distribuição espacial das bases de dados utilizadas. As perguntas levantadas inicialmente foram respondidas, com os dados de nome do município, bacias hidrográficas estadual e federal, áreas prioritárias para a conservação e unidades de conservação.

Catalogo	Grup	SSP	SIGLA_CATE	NOME	NOME_CATEG	DA_ÁREA	CATEGORIA	Ação_PRIOR
23	CT	Rhamdia quelen	APA	SUL RMBH	Area de Proteção Ambiental	Quadrilátero	ESPECIAL	Investigação Científica
33	CT	Poecilia reticulata	APA	SUL RMBH	Area de Proteção Ambiental	Quadrilátero	ESPECIAL	Investigação Científica
32	CT	Oreochromis niloticus	APA	SUL RMBH	Area de Proteção Ambiental	Quadrilátero	ESPECIAL	Investigação Científica
38	CT	Oligosarcus argenteus	APA	Piracicaba	Area de Proteção Ambiental			
41	CT	Oligosarcus argenteus	APA	SUL RMBH	Area de Proteção Ambiental	Quadrilátero	ESPECIAL	Investigação Científica
19	CT	Leporellus vittatus	APA	SUL RMBH	Area de Proteção Ambiental	Quadrilátero	ESPECIAL	Investigação Científica
22	CT	Imparfinis cf. minutus	APA	SUL RMBH	Area de Proteção Ambiental	Quadrilátero	ESPECIAL	Investigação Científica

Catalogo	Grup	SSP	Geocódigo	Município	FEDERAL	Estadual
23	CT	Rhamdia quelen	310900	Brumadinho	Rio São Francisco	Paraopeba
33	CT	Poecilia reticulata	310900	Brumadinho	Rio São Francisco	Paraopeba
32	CT	Oreochromis niloticus	310900	Brumadinho	Rio São Francisco	Paraopeba
38	CT	Oligosarcus argenteus	313170	Itabira	Rio Doce	Piracicaba
41	CT	Oligosarcus argenteus	310900	Brumadinho	Rio São Francisco	Paraopeba
19	CT	Leporellus vittatus	310900	Brumadinho	Rio São Francisco	Paraopeba
22	CT	Imparfinis cf. minutus	310900	Brumadinho	Rio São Francisco	Paraopeba

Figura 21 - PLANILHA RESULTANTE DOS CRUZAMENTOS

7. Considerações Finais

A integração entre SIG e BI é um dos desafios que ainda devemos superar para conhecer as potencialidades e aplicabilidades, proposta desse estudo, a fim de possibilitar a melhor opção para a tomada de decisão, apoiada por técnicas de identificação, extração e análise de dados.

O Geokettle é uma poderosa aplicação do PDI, possibilitando diversas modelagens, principalmente, análises espaciais. A sua manipulação e entendimento se torna fácil quando se tem o conhecimento do Kettle. Porém, usuários que tenham pouco domínio não conseguem um referencial de tutorial para explicar minuciosamente as funcionalidades de cada ferramenta e seus relacionamentos. O SPATIALYTICS DOCUMENTATION CENTER (2013), disponibiliza uma documentação simplória e que descreve sucintamente as ferramentas disponíveis, mas individualmente, e ao executar qualquer planejamento de ETL, não se sabe quais as particularidades e configurações necessárias para prosseguir com cada *Step*.

Existem diversas apresentações e trabalhos que utilizam do Geokettle, mas o processo é omitido e ao tentar replicar tal exemplo, particularidades de configuração são necessárias. Uma boa referência de aplicação do Kettle encontrada foi GUTUYAMA & SACILOTTO (2009), onde descreve todo o processo de planejamento, execução e implementação do ETL e DW.

A falta de uma documentação detalhada e de conhecimento do funcionamento da aplicação inviabilizou o uso de ferramentas como *Sextante Module* e *Spatial Analysis*, além de determinadas funções utilizadas nas condições de ferramentas como *Join Rows (cartesian product)* e *Filter rows*. Por não utilizar todo o potencial de tais ferramentas, o resultado esperado foi alcançado parcialmente, e o objetivo de aplicar em um estudo de caso, avaliando a eficiência das ferramentas utilizadas ficou dentro do esperado, superando a expectativa em relação à quantidade de ferramentas e possibilidade de utilizar *scripts* em java.

Todo o processo executado no ETL poderia ser realizado em diversos *softwares* de geoprocessamento *open source* ou não. Sobretudo, a eficiência, apesar de não ter sido mensurada, superou a expectativa, principalmente, no processamento. Este resultado se deve às configurações das *Steps* para seleção, exclusão, filtros, fluxos, etc, disponíveis, agilizando todo o processo de execução e alcançando resultados ótimos. Uma das vantagens mais consideráveis é a possibilidade de rodar os *Hops* em arquivos temporários.

Os diversos formatos de *input* e *output* aceitáveis, conexões avançadas com banco de dados e ferramentas diversas, possibilitam o uso do Geokettle como um potencial para processo de ETL geográfico. Trabalhos futuros podem ser desenvolvidos, analisando as potencialidades das ferramentas descritas anteriormente e possibilitando trabalhar melhor as análises espaciais. Deste modo, o conhecimento adquirido com este estudo de caso aplicado possibilitou uma série de questionamentos que ficou sem um resultado, muitos deles pela falta de conhecimento do funcionamento da aplicação, outros pelo curto tempo de desenvolvimento.

8. Referências Bibliográficas

ANDRADE, Ludymilla Marianne Guimarães de. 2011. **Data Warehouse Geoespacial aplicado à Gestão estratégica na Prefeitura de Belo Horizonte**. XIII Curso de Especialização em Geoprocessamento, Instituto de Geociências, Departamento de Cartografia. Monografia, 47p.

BRITO, J. J. 2012. **Processamento de consultas SOLAP *drill-across* e com junção espacial em *data warehouses* geográficos**. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 110 p.

DEGGAU, Renato & FILETO, Renato. **Enriquecendo Data Warehouses Espaciais com Descrições Semânticas**. XXIV SBBB, p. 61-66.

DIAS, Braian O.. **Projeto de Data Mart utilizando ferramentas Open Source** Curso de Ciência da Computação, Universidade de Passo Fundo (UPF). Passo Fundo, RS.

DRUMMOND, Gláucia Moreira, et al.. 2005. **Síntese das Áreas Prioritárias de Minas Gerais**. In. Biodiversidade em Minas Gerais: um atlas para sua conservação. 2ª Ed, Belo Horizonte: Fundação Biodiversitas, 162-178 pp. Disponível em: <http://www.biodiversitas.org.br/atlas/sintese.asp>. Acesso em: 15 de jan. de 2010.

GeoKettle. Disponível em: <<http://www.spatialytics.org/projects/geokettle/#>>. Acesso em: 22 out. de 2013.

GUTUYAMA, Fábio Hideki & SACILOTTO, Renan Lotto. 2009. **Desenvolvimento de um data warehouse para o processo de decisão em uma empresa de telecomunicações**. Monografia, Escola Politécnica da Universidade de São Paulo, Engenharia de Computação, São Paulo, 80 p.

GUTUYAMA, Fábio Hideki & SACILOTTO, Renan Lotto. 2009. **Desenvolvimento de um data warehouse para o processo de decisão em uma empresa de telecomunicações**. São Paulo. Escola Politécnica, Universidade de São Paulo, Curso de Engenharia de Computação. Monografia, 80 p.

IBGE - Instituto Brasileiro de Geografia e Estatística. 2010. **Malha Municipal 2010**. Disponível em: <http://servicodados.ibge.gov.br/Download/Download.ashx?u=geoftp.ibge.gov.br/malhas_digitais/municipio_2010/mg.zip>. Acesso em: 20 de fev. de 2013.

IEF – Instituto Estadual de Florestas. 2013. Arquivo Vetorial das Unidades de Conservação do Estado de Minas Gerais.

IGAM – Instituto Mineiro de Gestão das Águas. 2002. **Unidades de Planejamento e Gestão dos Recursos Hídricos do Estado de Minas Gerais**. Disponível em: <[http://www.igam.mg.gov.br/images/stories/arquivos/ArcView/Unidades de Planejamento.zip](http://www.igam.mg.gov.br/images/stories/arquivos/ArcView/Unidades_de_Planejamento.zip)>. Acesso em 18 de out. de 2011.

Instituto Nacional de Pesquisas Espaciais - INPE. Disponível em: <http://www.dpi.inpe.br/spring/portugues/tutorial/introducao_geo.html>. Acesso em: 26 març. 2007.

OSGeo Live. Disponível em: <http://live.osgeo.org/en/overview/geokettle_overview.html>. Acesso em: 22 out. de 2013.

Pentaho. Disponível em: <<http://www.pentaho.com/>>. Acesso em: 22 out. de 2013.

PULVIRENTI, Adrián Sergio & ROLDÁN, María Carina. 2011. **Pentaho Data Integration 4 Cookbook**: Over 70 recipes to solve ETL using Pentaho Kettle. Packt Publishing, 352 p.

SEZÕES, Carlos; OLIVEIRA, José & BAPTISTA, Miguel. 2006. **Business Intelligence**. Sociedade Portuguesa de Inovação, Consultadoria Empresarial e Fomento da Inovação. Porto. 160 p.

SPATIALYTICS DOCUMENTATION CENTER. 2013. **Spatialytics ETL**. Disponível em: <http://docs.spatialytics.com/doku.php?id=en:spatialytics_etl>. Acesso em: 06 de Nov. de 2013.

Thierry Badard (co-ed),. 2012. **Geospatial Business Intelligence (GeoBI)**. OGC White Paper, Open Geospatial Consortium, 25 p.