Importação de dados para o banco de dados do projeto LACOS Web

Alex Leonardo Dalcin

Orientador: Prof. Dr. Daniel Luis Notari

Curso de Bacharelado em Sistemas de Informação - Universidade de Caxias do Sul (UCS - CARVI)

{aldalcin, dlnotari}@ucs.br

Resumo. *O objetivo deste trabalho é apresentar a análise de ferramentas de integração de banco de dados e efetuar a unificação das informações do projeto LACOS Web para uma base única, estruturada e organizada para que seja possível a análise das informações de forma correta e satisfatória. LACOS Web é um portal desenvolvido para estruturar e armazenar coletas de informações e características de lagoas e rios da costa do Rio Grande do Sul.*

Palavras-chave: *Integração; Banco de dados; LACOS Web*

1. Introdução

Em meados de 2007, a Universidade de Caxias do Sul (UCS), em parceria com a Petrobras Ambiental, iniciou o Projeto Lagoas Costeiras (LACOS), que possui por objetivo mapear e analisar as características de lagoas e rios da costa do Rio Grande do Sul, bem como o uso turístico destas áreas e o desenvolvimento socioambiental da região.

O projeto LACOS foi dividido em três etapas, sendo que cada etapa foi direcionada à uma região, fragmentando os trabalhos para melhorar a coleta dos dados e a análise das informações geradas. Todas as coletas foram documentadas em planilhas eletrônicas ou em meio físico e depois transpostas para um meio eletrônico. A estrutura do arquivo em cada uma das etapas documentadas é diferente das demais, fazendo com que os dados a serem analisados não estejam todos no mesmo local.

Os dados biológicos das lagoas do Rio Grande do Sul estão sendo coletados desde meados dos anos 1980, antes mesmo da criação do projeto LACOS, e diversas ferramentas foram usadas para isso, como planilhas eletrônicas e pequenas bases de dados (Microsoft Access). A dificuldade é que são muitos dados fragmentados e, possivelmente, há mais informação dispersa em outros locais ou outras bases de dados. Informações duplicadas também surgem quando há uma base não unificada de informação, pois não há tratamento dos dados. Para construir uma base mais segura e confiável, disponível para todos que necessitarem fazer uso destas informações, uma base de conhecimento unificada e estruturada se torna a solução mais adequada. Com esse intuito, o LACOS Web foi criado por alunos da Universidade de Caxias do Sul. Este projeto foi desenvolvido em etapas, que contemplam um portal Web (podendo ser acessado pelo link http://bioinfoucs.com/lacosweb/). Este portal possui um banco de dados padronizado que armazena todas as informações coletadas, tanto da avaliação da qualidade da água quanto

da fauna existente na região. Complementando os dados registrados, foi adicionada a função de georreferenciamento que indica a localização exata da medição e um sistema de administração das operações para controle de acesso e permissões.

As informações do banco de dados que estavam em outras fontes ainda não foram unificadas no mesmo local, possuindo somente dados novos. Esta falta de integração faz com que os dados sejam difíceis de serem analisados e a falta de informações centralizadas fazem com que as análises de uma mesma amostra fiquem discrepantes entre si. Esta análise errada pode ser a causa de ações equivocadas, projetos mal planejados, orçamentos de soluções mal dimensionados, podendo causar falhas enormes em projetos nestes locais. Além disso, uma base única das informações possui melhor gerenciamento para o administrador do banco de dados, simplifica o desenvolvimento das ferramentas, facilita a gestão da segurança dos dados, continuidade e disponibilidade das informações. Mantendo as informações dispersas, há a possibilidade de perda de dados, informações duplicadas, dados divergentes entre si, dentre outras ocorrências que podem fazer com que o conhecimento gerado não seja de acordo com a realidade.

Como as bases também foram criadas com arquivos de planilhas, como o Microsoft Excel ou LibreOffice, a aplicação do portal não consegue utilizar estes arquivos como base de dados, pois não possuem uma estrutura correta para isso. Estes arquivos não podem ser importados diretamente para o banco de dados, pois as informações que nele constam não condizem com os dados das tabelas do banco relacional que foram criadas. Além de arquivos de planilhas, também foram usados bancos simples, criados com o uso do Microsoft Access, onde estes também possuem uma estrutura não relacional para armazenagem dos dados.

Na maioria dos casos, refazer a documentação se torna inviável, e em alguns casos impossível pela quantidade de dados. Por isso, a migração dos dados de diversas fontes para uma só é a melhor opção. Isso caracteriza a criação de um Data Warehouse, que "consiste em agregar informação proveniente de uma ou mais Bases de Dados (BD), ou de outras fontes, para posteriormente a tratar, formatar e consolidar numa única estrutura de dados" (Ferreira et al., 2010, p. 1). Este armazém de dados deverá ser povoado com as informações transformadas das diversas fontes diferentes, com o processo de ETL.

O processo de Extrair, Transformar e Carregar, (ETL, sigla do inglês "Extract Transform Load"), é definido por Abreu (2008, p.1) como um processo que "destina-se à extração, transformação e carga dos dados de uma ou mais bases de dados de origem para uma ou mais bases de dados de destino", tratando as informações de modo que somente os dados corretos sejam incluídos na nova base. Este tratamento é constituído pela limpeza de informações não corretas ou duplicadas, verificações de consistências com o novo banco e padronizando os dados para serem utilizados por alguma aplicação.

Este projeto tem por objetivo principal avaliar uma ferramenta para migração dos dados de diversos arquivos para um banco de dados unificado e estruturado, e utilizar desta ferramenta para efetuar a unificação dos dados do projeto LACOS Web.

O artigo está estruturado da seguinte forma: na Seção 2 são abordadas a teoria de transformação de dados em conhecimento e a importância da transformação dos dados antes da análise, além dos trabalhos relacionados ao assunto. Na Seção 3 é apresentada a

metodologia. Na Seção 4 são aplicados e detalhados os resultados da aplicação e na Seção 5 são apresentadas as considerações finais.

2. Referencial Teórico

Ao longo dos anos, muitos dados são criados e armazenados com diversas finalidades. Os mesmos dados, por muitas vezes, são armazenados em diversas ferramentas digital e fisicamente. Com isso, a informação fica dividida em diversas fontes, fazendo com que seja mais difícil gerar conhecimento satisfatoriamente.

A transformação de dados em conhecimento é um processo crucial para a análise dos dados a fim de uma tomada de decisão. Prass (2012, p1) cita que "visando transformar estes dados em conhecimento, surge o processo chamado de Descoberta de Conhecimento em Bancos de Dados", ou KDD. Esse processo é comumente confundido, erradamente, com a mineração de dados ou data mining. Para isso, o KDD é dividido em fases, como destacadas na Figura 1, e brevemente descritas na sequência.

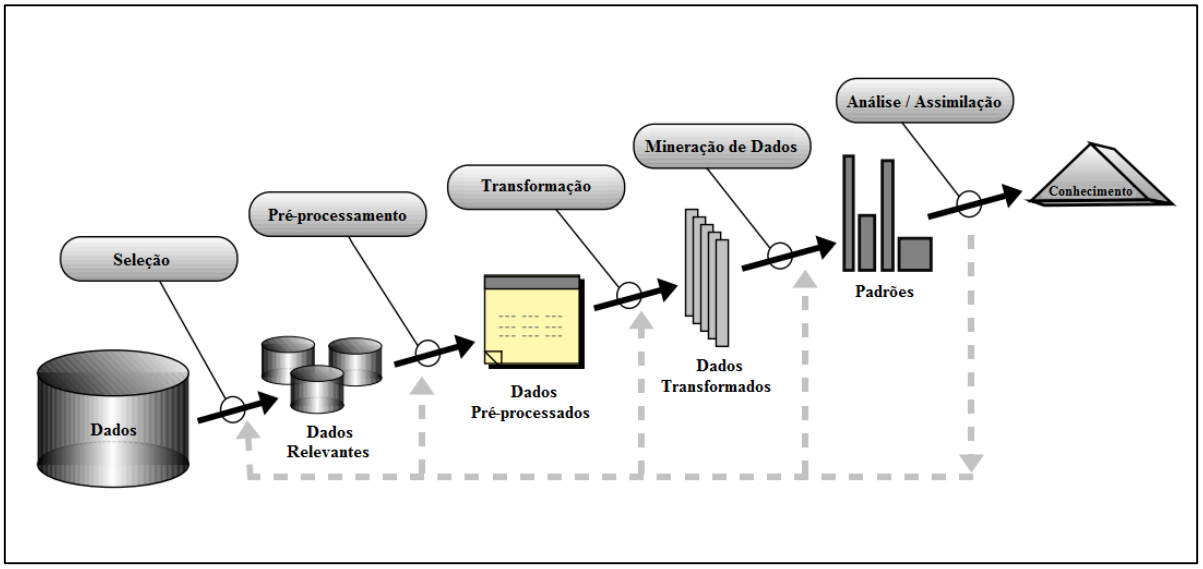


Figura 1. Fases do KDD

A etapa de seleção dos dados é caracterizada pelo descobrimento dos dados, com a pesquisa e mapeamento de todas as fontes, atributos e registros para a análise. O pré-processamento e limpeza faz a limpeza dos dados antes da transformação e mineração, removendo duplicações de dados, dados inconsistentes e dados faltantes.

A etapa de transformação dos dados se faz necessária para agrupar todos os dados em um repositório único. O processo de ETL é o exemplo que temos de transformação dos dados, que é o objetivo deste estudo. Essa etapa é essencial para uma boa análise dos dados no final do processo.

Na mineração de dados, os mesmos já estão em local único e disponíveis para o trabalho de processamento de informações e transformação dos dados em conhecimento.

Na etapa de interpretação e avaliação, o "conhecimento adquirido através da técnica de data mining deve ser interpretado e avaliado para que o objetivo final seja

alcançado” (Prass, 2012, p6). A análise deve ser feita em conjunto com especialistas no assunto para a tomada da decisão para qual o projeto foi desenvolvido.

Segundo a ideia de Fayyad et al. (1996), a literatura existente que cita o processo de KDD dá muita ênfase na mineração dos dados. Porém, outros passos também são importantes (e provavelmente mais importantes) para o sucesso do KDD. Tendo isso em mente, o ETL é o processo ligado à etapa de transformação e integração dos dados para ampliação da vida útil das informações, dentro de uma empresa ou companhia.

2.1 ETL

O ETL é definido por Abreu (2008, p.1) como um processo que “destina-se à extração, transformação e carga dos dados de uma ou mais bases de dados de origem para uma ou mais bases de dados de destino”, tratando as informações de modo que somente os dados corretos sejam incluídos na nova base. Este tratamento pode ser a limpeza de informações não corretas ou duplicadas, verificações de consistências com o novo banco e padronizando os dados para serem utilizados por alguma aplicação.

O processo de ETL é tão importante para o negócio que pode ajudar ou destruir o banco de dados (Kimbal e Caserta, 2004). Ainda de acordo com Kimbal e Caserta (2004), esse processo pode consumir em até 70% do tempo de implantação de um banco de dados ou Data Warehouse. Já para (Ferreira et al, 2010), este processo pode consumir 80% do tempo necessário em um projeto de implantação de um Data Warehouse.

Segundo Eckerson e White (2003), os principais usos do ETL nas empresas são bases relacionais e arquivos, mas uma significante porcentagem já está utilizando deste método para SAP, arquivos XML e bases de dados web, dentre outros.

Na Figura 2, podemos identificar um exemplo das operações do ETL. Neste exemplo existem três tabelas diferentes, com alguns dados que são repetidos, mas possuem estruturações diferentes. Com estas características não é possível importar os dados para o novo banco de dados sem serem transformados ou trabalhados. Na transformação, os dados são padronizados para remoção de duplicidade, padronização de campos como altura, estes possuindo diversos formatos de armazenar os dados, além da organização das colunas de acordo com a necessidade. A última tabela mostra o banco de dados final, com os dados padronizados, estruturados e limpos, podendo ser utilizada para a análise e mineração de dados.

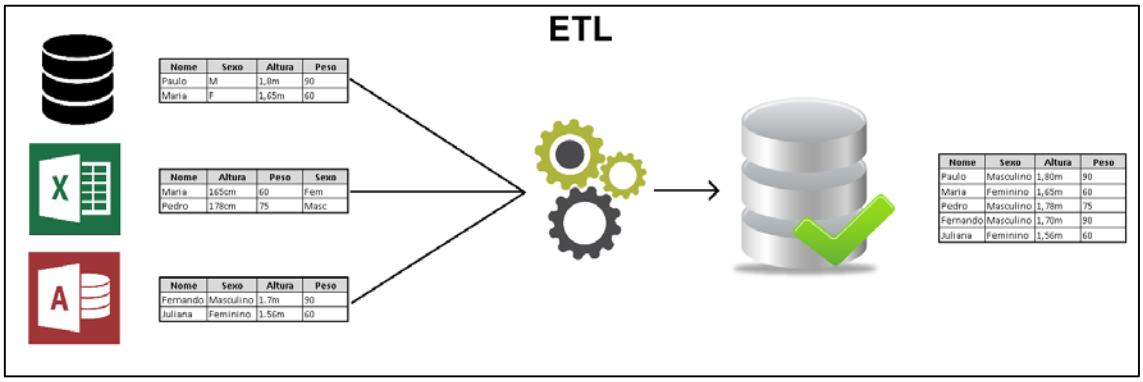


Figura 2. Exemplo de ETL

2.2 Trabalhos Relacionados

O processo de ETL é de extrema importância para diversas aplicações. Como caso de estudo feito por Ferreira et al. (2010, p6), onde havia a "necessidade de validar os dados dos recursos humanos de um centro hospitalar", o que fez com que fosse analisado o processo de integração. "A informação encontrava-se dispersa em mais de uma centena de tabelas com registos processados e a processar", segundo os mesmos autores, e havia a necessidade de "garantir que toda a informação estava correta e consistente" para evitar que "dados incorretos pudessem conduzir a erros críticos de tomada de decisão". Para isso, foram testadas diversas ferramentas e escolhida a plataforma de desenvolvimento Oracle Warehouse Builder para a construção de um Data Warehouse.

Outro caso de estudo que pode ser citado, feito por Florea et al. (2015, p1), que compara a evolução tecnológica das aplicações Pentaho Data Integration e Oracle Data Integrator. Segundo os autores, ambas as aplicações são poderosas, altamente customizáveis e possuem compatibilidade com diversos bancos de origem e destino. Para tomar esta decisão, foram avaliados a facilidade de uso, compatibilidade com demais bancos, atualizações na aplicação e custo.

Na abordagem feita por Bansal e Kagemann (2014), foi a avaliação da ideia principal das funções e operações do ETL em si para a criação de um Big Data. A integração teve por objetivo avaliar a efetividade do ETL na integração de um conjunto de dados educacionais de mais de 7 milhões de usuários, pesquisa nacional de viagens dos Estados Unidos e dados de economia de combustível da Agência de Proteção Ambiental Americana. Todos estes dados foram compilados e testados em ferramentas para a criação de um Big Data.

Já para Degan (2005, p15), "A necessidade da integração de dados em ambientes empresariais, apesar de antiga, ainda é um problema crucial a ser resolvido para a maioria das empresas". Por isso, a autora propôs uma arquitetura para integração dos dados para melhorar a gestão, análise e aumentar a vida útil da informação, com soluções no campo de integração dos dados, banco de dados e o uso destas informações em sistemas Web.

Em seu trabalho, Ferrandin (2002) descreve os métodos e ferramentas de integração, tendo como base de estudo, bancos heterogêneos e padrão XML para integração dos dados. Alguns modelos também foram apresentados e suas vantagens e desvantagens no processo também são relatadas.

2.3 Ferramentas analisadas

Como ponto de partida na pesquisa, foram identificadas ferramentas existentes no mercado, baseando-se em experiências documentadas em outros trabalhos acadêmicos e em artigos onde possuem como objetivo a integração de bancos de dados. Neste estudo, foram identificadas 4 ferramentas para o comparativo de características e funcionalidades existentes em cada uma delas. As ferramentas selecionadas são Apatar, Geokettle, Pentaho e Talend. Cada uma destas ferramentas possui características em comum e outras que possuem funcionalidades únicas.

Apatar é uma ferramenta de integração de dados *open source* que possui objetivo de ligação entre bancos de dados, aplicações e arquivos estruturados ou não estruturados.

A ferramenta é atualmente utilizada em mais de 10 mil organizações, integrando serviços e bancos de dados dos mais variados. Mais informações podem ser acessadas pelo site oficial http://www.apatar.com/.

Geokettle é uma ferramenta baseada na ferramenta genérica de integração Kettle, com alterações para abranger dados geográficos, arquivos, mapas estruturados, bancos de dados em nuvem e serviços web. A carga dos dados também pode ser feita nestes serviços. Possui foco em dados espaciais, e pode ser melhor conhecido através do site http://www.spatialytics.org/projects/geokettle/.

O Pentaho Data Integration é uma ferramenta que pertence ao grupo Hitachi, desenvolvido com o intuito de integrar dados. Na opção Data Analytics pode ser utilizado também para integrar e analisar dados de quaisquer fontes de maneira mais visual. O site oficial da ferramenta é https://www.hitachivantara.com/go/pentaho.html.

Como última ferramenta, o Talend Open Studio também é especializado em integração de aplicações e bancos de dados, podendo utilizar como base dados de serviços em nuvem e aplicações diversas. Os dados podem ser compartilhados com outras ferramentas para análise de dados e auxílio na tomada de decisões. Para conhecer a ferramenta, pode-se acessar o site https://www.talend.com/.

3. Metodologia

Buscando otimizar os processos do projeto, o mesmo foi dividido em seis grandes etapas a fim de estruturar a solução do problema. Este processo pode ser visto na Figura 3.

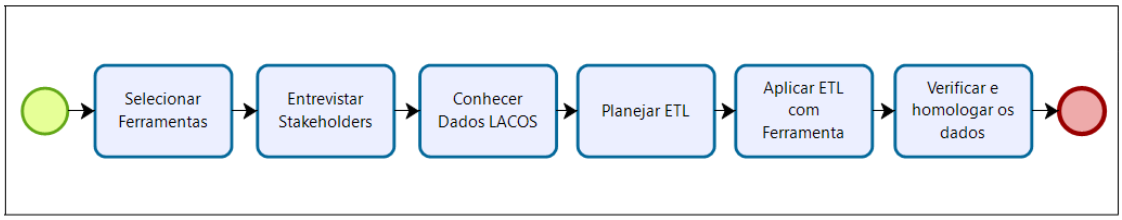


Figura 3. Modelo BPMN da solução

Na primeira etapa foram selecionadas as ferramentas existentes no mercado e analisadas as funcionalidades de cada uma. Também foi feito o comparativo das características de todas a fim de identificar qual melhor se adapta ao projeto.

A segunda etapa teve o foco no conhecimento do projeto, assim como as necessidades dos interessados pelo projeto. Nessa etapa foram identificadas as demandas que o projeto deve suprir e explicadas as limitações e planos para o resultado final.

Na terceira etapa houve a análise dos dados existentes, tanto as planilhas de origem quanto o banco de dados criado, sendo este já com a estruturação final e possuindo informações cadastradas.

A quarta etapa foi elaborada a modelagem do processo de ETL. Nessa modelagem foi criado um mapeamento com a estrutura que serve como modelo para a transformação dos dados. Essa estruturação é importante a fim de aproximar-se ao máximo da estruturação final para evitar imprevistos com dados não analisados ou confusão com o uso da ferramenta.

Na quinta etapa foi criada a modelagem da transformação utilizando a ferramenta selecionada na primeira etapa. Essa modelagem segue a modelagem ETL pois esta já foi analisada e elaborada para suprir a necessidade do projeto.

Na sexta e última etapa os dados importados foram analisados para garantir o sucesso da importação, homologando se todos os dados existentes foram importados. O portal LACOS Web também foi apresentado para a equipe, mostrando os registros importados e apresentando como foi feita a migração.

4. Resultados

Com o planejamento feito, foi iniciada a implementação das etapas do projeto, como detalhado nos tópicos a seguir.

4.1. Seleção de ferramenta

Para a definição do método de avaliação das ferramentas foi utilizada uma compilação de características presentes nos diversos trabalhos analisados, e as mais relevantes para o projeto foram destacadas e avaliadas. Para outros projetos, com demandas diferentes, podem ser utilizados parâmetros diferentes de avaliação.

4.1.1 Método de avaliação

Baseado em demais trabalhos já feitos sobre este assunto, foram identificadas características básicas para a escolha da ferramenta. As características são a leitura de planilhas Excel, escrita no banco de dados PostgreSQL e integridade dos dados carregados. Com estas três características, foram selecionadas apenas 4 ferramentas para seguir na busca da melhor ferramenta para o projeto.

A facilidade de uso demonstra a capacidade da ferramenta ser modelada de acordo com a necessidade do usuário, sem a necessidade de conhecimento técnico para utilizar a mesma. Na curva de aprendizagem foi avaliado o tempo para aprender a utilizar a ferramenta e ter sucesso na operação. A documentação auxilia no conhecimento da ferramenta e das suas funcionalidades, fazendo com que seja muito mais simples o desenvolvimento da solução proposta. Uma comunidade ativa demonstra que a ferramenta está sendo utilizada por diversas outras pessoas, que muitas vezes disponibilizam suas soluções, mesmo não fazendo parte de uma empresa, o que facilita no entendimento das soluções e casos de estudo. Atualizações demonstram que a ferramenta possui uma melhoria nas tecnologias e soluções, o que fazem com que entregam novas funções e melhorias de performance. Os cases de sucesso demonstram o uso da ferramenta em soluções reais, que servem como exemplo na escolha da ferramenta que melhor se adapta ao projeto.

Após instaladas, cada ferramenta foi submetida aos testes e os requisitos foram avaliados de acordo com uma simulação simples do projeto. Nessa simulação os dados forma exportados para uma planilha, a fim de agilizar o processo, visto que não foi avaliada a performance na execução da importação. Após os testes, uma nota de 0 a 10 foi dada para cada ferramenta, conforme relatado na Tabela 1.

Tabela 1. Comparativo das ferramentas analisadas

	Apatar v1.12.23	**Geokettle v2.5**	**Pentaho Data Integration v7.1.0.0-12**	**Talend Open Studio v7.0.1**
Facilidade de uso	3/10	4/10	8/10	9/10
Curva de aprendizagem	4/10	4/10	9/10	8/10
Documentação em português	2/10	5/10	8/10	7/10
Comunidade ativa	1/10	2/10	9/10	9/10
Atualizações	3/10	2/10	10/10	9/10
Cases de sucesso	1/10	2/10	10/10	8/10
Média	**2,3/10**	**3,16/10**	**9/10**	**8,3/10**

Os testes foram realizados durante duas semanas, aonde todos os pontos foram analisados para a tomada da decisão das notas em cada requisito. Ao final, a média de pontos decidiu qual das ferramentas foi a melhor escolha para o projeto.

4.1.2 Ferramenta selecionada

Todas as ferramentas listadas suprem a necessidade básica para a solução do problema de integração dos dados, sendo que a que teve um melhor desempenho na função foi o Pentaho Data Integration (PDI). As demais ferramentas também possuem desempenho semelhante, porém, os fatores mais críticos em relação à escolha da ferramenta foram a base de conhecimento já existente e a facilidade de aprendizado na ferramenta.

Os testes para a escolha da ferramenta nesse projeto, por se tratar de diversos bancos de dados pequenos, não levou em consideração a performance de processamento dos dados. Para projetos grandes, com necessidade de processamento e transformação constante dos dados (como em caso de processamento de dados para transferência de informações de bancos para um Data Warehouse ou ferramentas de BI), a performance deve ser considerada com mais prioridade na escolha da ferramenta.

O PDI obteve boas notas em todos os requisitos, sendo que os pontos que mais contaram foram as atualizações que o software recebe e os cases de sucesso, relatados tanto em casos de estudos em trabalhos acadêmicos quanto no livro Pentaho Data Integration 4 cookbook, onde mostra mais de 70 casos de sucesso utilizando a ferramenta.

4.2. Entrevista com Stakeholders

Stakeholder, ou partes interessadas, é o conjunto de uma ou mais pessoas que está ligada direta ou indiretamente ao projeto. Na primeira entrevista com a equipe, feita no dia 06/08/2018, foram feitos os levantamentos das demandas necessárias para a conclusão do projeto de maneira satisfatória. Nesta reunião a grande preocupação foi com a integridade dos dados, facilidade do uso do portal e a possibilidade de importação dos dados futuros quando necessário. Para isso, foi acordado que o foco será as medições físico/químicas dos lagos.

Em meio à estas demandas, foram feitas algumas sugestões. A primeira das sugestões foi utilizar um banco de dados paralelo para a importação dos dados, antes de importar os mesmos para o banco de dados oficial do portal. Isso foi sugerido para evitar

que ocorra algum problema com a importação dos dados e para validar a integridade das informações importadas.

Como segunda sugestão foi desenvolvido um arquivo padrão para futuras importações dos dados físico/químico dos lagos. Estas importações podem ser feitas em caso de algum arquivo documentado manualmente ou outras planilhas eletrônicas sejam encontrados e devam também ser importadas. Essa planilha modelo foi criada a partir da planilha base e acordada com a equipe, pois atende as demandas das medições feitas até então.

4.3. Conhecer os dados do projeto LACOS

Na mesma entrevista citada no tópico anterior, a equipe do projeto fez uma explicação melhor de como iniciou o projeto LACOS, como eram feitas as coletas e qual era o intuito. Também foram apresentadas as informações originais e como deveriam ser os dados importados. Mediante essa reunião com as partes envolvidas, houve um crescimento no entendimento do projeto e em como estas informações coletadas e a serem importadas podem contribuir com a comunidade acadêmica, além da sociedade que também pode usufruir do conhecimento gerado durante a pesquisa. Esse crescimento no entendimento do projeto auxiliou no planejamento da transformação dos dados.

Os dados utilizados foram disponibilizados em uma planilha eletrônica, com a estruturação própria das informações. Nesta planilha foram armazenados dados das coletas como data e hora, posição geográfica velocidade do vento e profundidade da coleta. Em cada coleta são analisados diversas características físicas e químicas da água, como a concentração e saturação de oxigênio na água, índice pH, transparência, temperatura, além de diversas composições químicas como Clo-a fluoroprobe, Clo-a, DBO-5, NNO3, NNH3, dentre outros. No total, podem ser medidas 13 características químicas e físicas da água, não sendo obrigatória a coleta de todos os índices.

Para o conhecimento do banco de dados, foi replicado o banco em um computador para teste e entendimento das ferramentas utilizadas, juntamente com a diagramação das tabelas. Além disso, foi feito contato com outro colega que ajudou a desenvolver parte do LACOS Web, e foram tiradas algumas dúvidas em relação à estrutura e acesso ao servidor e banco de dados. Para melhorar o mapeamento das informações, a Figura 4 apresenta o diagrama de relacionamento (ER) das tabelas do banco de dados.

O diagrama ER apresentado representa a modelagem do banco de dados existente com suas ligações entre as tabelas. Essas tabelas armazenam dados distintos das coletas, como as medições das coletas, as características do local, informações de cidades, dentre outras. Nesse mesmo diagrama também são identificadas as ligações entre as tabelas, mostrando quais delas são utilizadas para vincular cada uma das informações ali armazenadas. Essa estrutura auxilia na organização da informação e na gestão dos dados armazenados e do banco de dados, utilizado para o desenvolvimento do portal.

Além disso, o diagrama também é uma informação base para a estruturação de novos serviços e novas ferramentas adicionais à aplicação, como ferramenta de análise de dados, mineração de dados, exportação de informações para outras bases, dentre outras tantas.

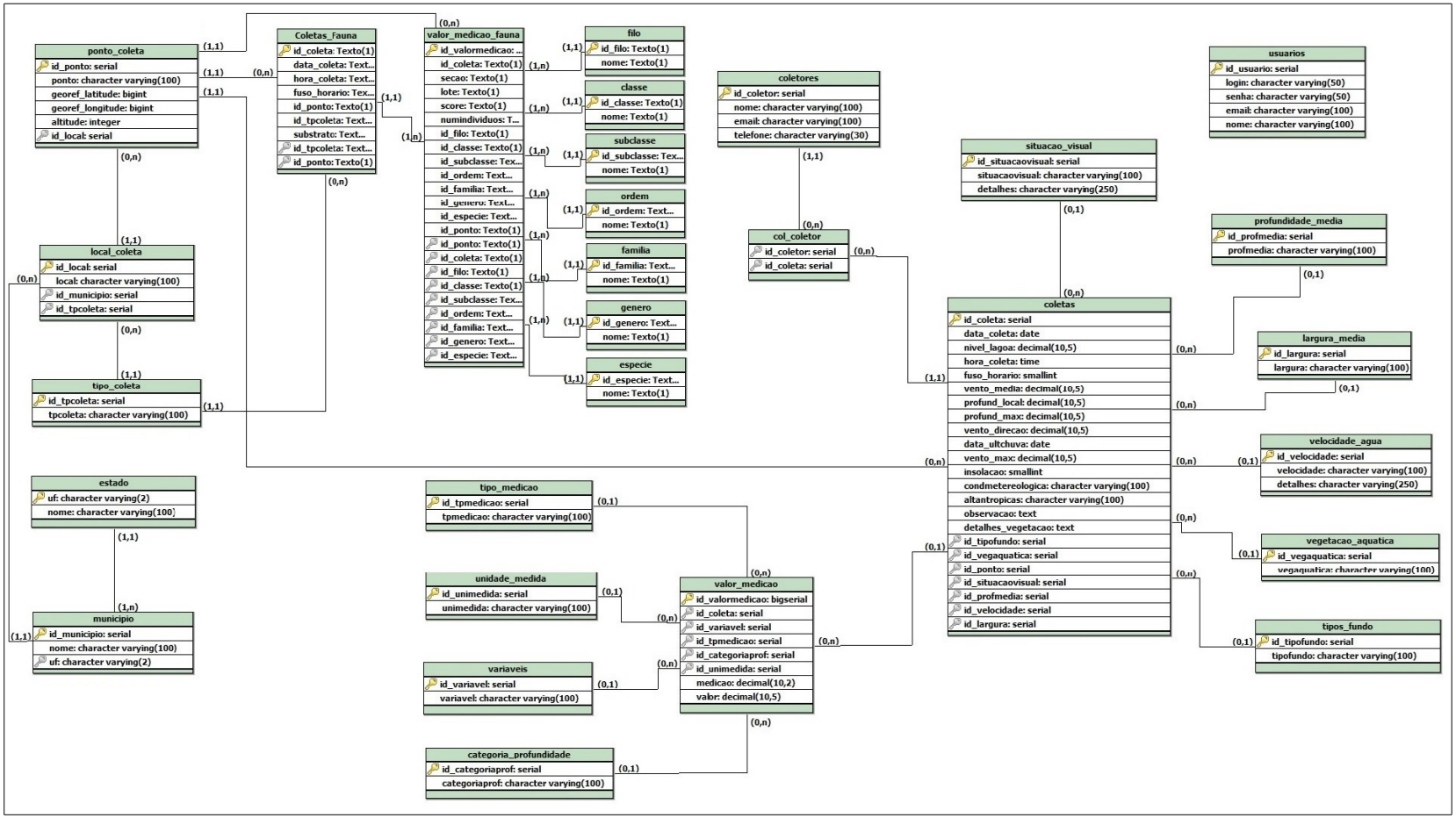


Figura 4. Diagrama ER do LACOS Web

O banco de dados utilizado no portal é o PostgreSQL, um banco de dados bastante utilizado no mundo para soluções em empresas, e também bastante utilizados como objeto de ensino por se tratar de uma aplicação simples e robusta. Alguns dados já foram adicionados para demonstração da ferramenta durante o desenvolvimento da mesma. Os novos dados serão acrescentados na base já existente.

Após isso, foi possível identificar a estrutura existente no banco de dados e como a importação poderia ser feita, de acordo com o arquivo modelo e o banco de dados estruturado, criado para o portal.

4.4. Planejamento do ETL

Após a escolha da ferramenta e entendimento dos dados a serem trabalhados, iniciou-se o processo de mapeamento dos dados. Os dados armazenados, dispostos digitalmente em uma planilha, foram analisados e homologados pela equipe do LACOS, que disponibilizou as informações. Após a homologação dos dados, iniciou-se o processo de mapeamento de entrada e saída das informações a serem processadas. Este mapeamento pode ser visto na Figura 5.

O mapeamento dos dados foi dividido em duas partes, uma contendo as informações das coletas, que levam em consideração os dados gerais da lagoa e a outra contendo as medições feitas em cada uma das coletas. Essas medições levam em consideração diversas informações físico/química da água. Os valores das medições são vinculados todos no campo de valor, e o nome da categoria define qual é o tipo da medição.

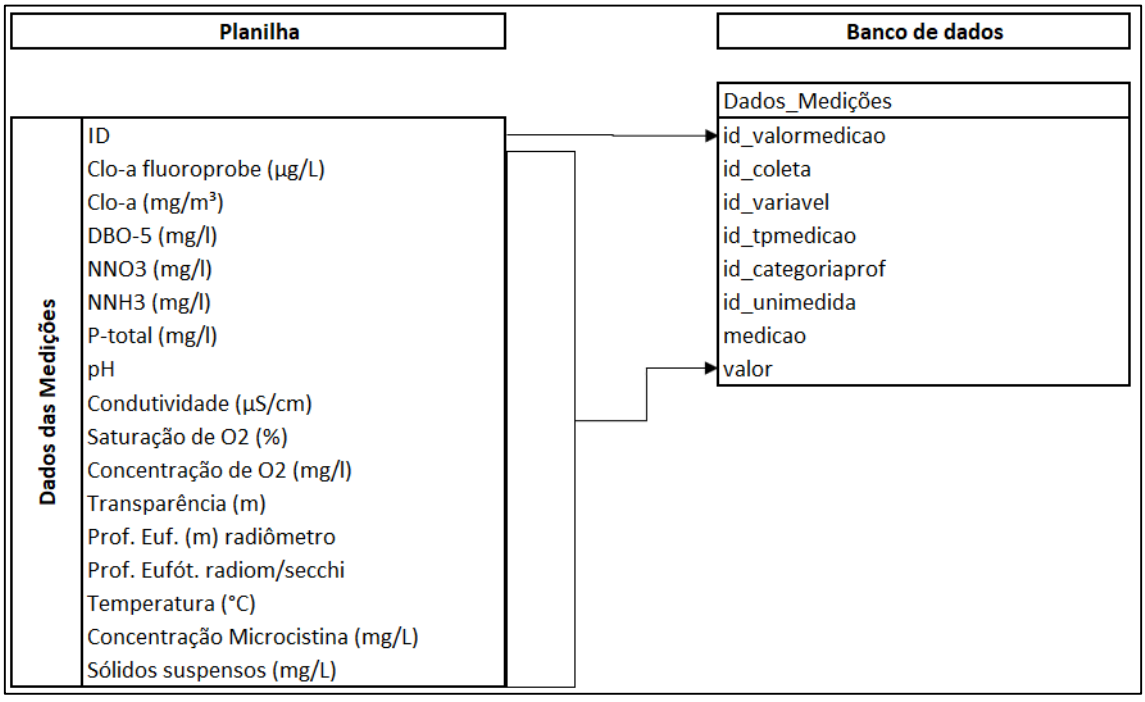


Figura 5. Mapeamento dos dados de entrada e saída das medições

Na Figura 6 é apresentado o estudo da transformação dos dados das coletas, onde os dados informados nas medições sempre são adicionados no campo do valor, e o ID irá

vincular a medição à informação da coleta.

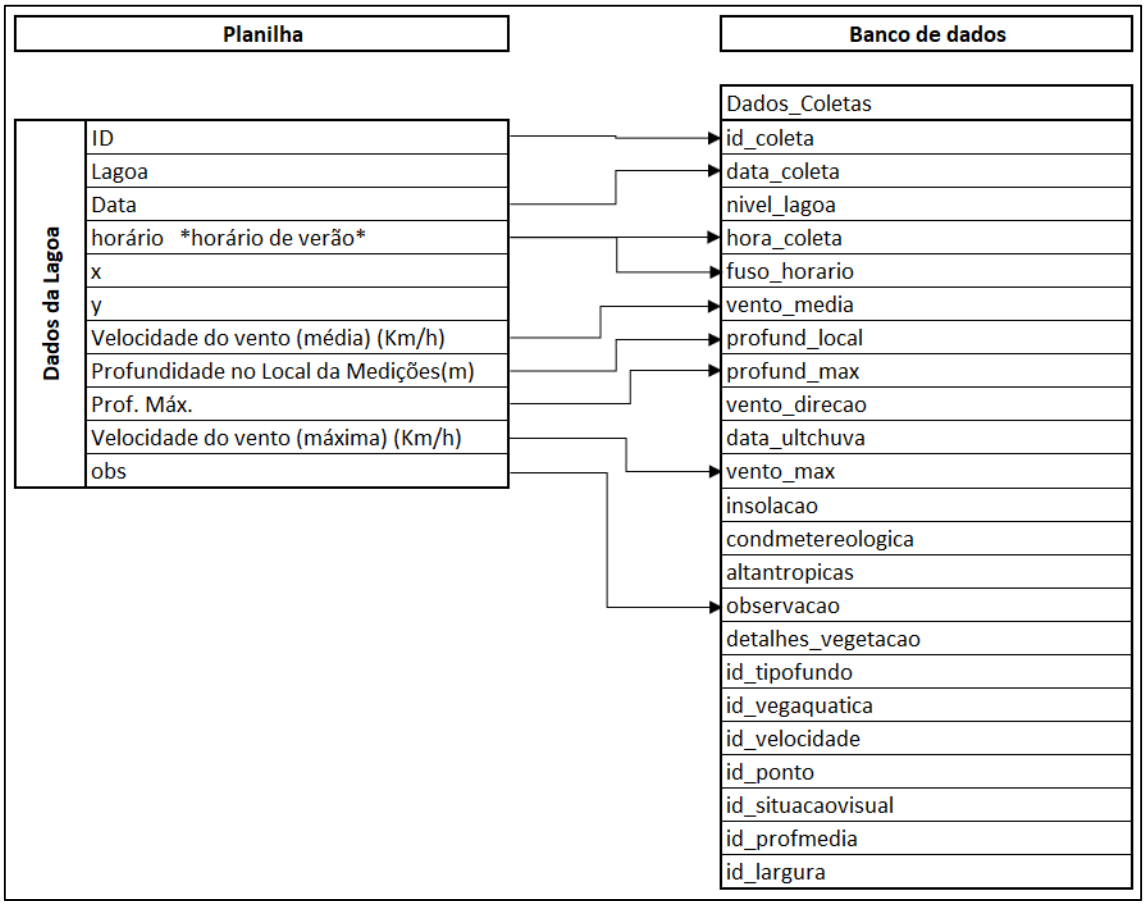


Figura 6. Mapeamento dos dados de entrada e saída das coletas

4.5 Aplicação do ETL

Após o mapeamento concluído, começaram os trabalhos com a ferramenta Pentaho Data Integration para a leitura do arquivo. As etapas de transformação dos dados foram separadas em etapas, cada uma abrangendo uma parte da transformação.

A primeira etapa foi a identificação dos pontos de coleta. Estes pontos serão comparados aos pontos já cadastrados no banco de dados a fim de identificar se estes já estão cadastrados. Em caso de já estar cadastrado, o sistema não irá carrega-lo novamente. Os dados restantes desta limpeza serão carregados no banco de dados ao final do projeto.

Na próxima etapa, houve a separação dos dados das coletas na planilha original. Os dados das coletas são características dos pontos de coleta da amostra (latitude e longitude), nome da lagoa, data e hora da coleta, dentre outros. Estes dados foram selecionados e exportados para uma planilha antes de ser importada ao banco, exatamente para homologar e verificar quais alterações irão afetar o resultado final. Nessa etapa, já foram verificadas se todas as informações estão preenchidas. Em caso de alguma não estar, irá ficar em branco o campo. Como os dados não possuem um identificador único, aqui também foi gerado um número sequencial e atribuído à cada linha importada.

Com a numeração sequencial já gerada, na próxima etapa fez-se a padronização das informações das coletas. Essa padronização faz com que os dados tenham uma mesma estrutura em comparação ao banco de dados que será carregado ao final da transformação. As principais padronizações feitas foram nos tipos de dados, data e hora, sendo esse último mais complexo por não apresentar uma estrutura fixa, conforme Figura 7.

Planilha	Banco de dados	
horário *horário de verão*	horário	horário de verão
*11:00	11:00:00	Sim
*10:30	10:30:00	Sim
10:00	10:00:00	Não
10:30*	10:30:00	Sim
09:30*	09:30:00	Sim
12:*	12:00:00	Sim
*15:30	15:30:00	Sim
*9:00	09:00:00	Sim
12:30	12:30:00	Não

Figura 7. Exemplo de transformação de horas

Como pode-se ver, a estrutura da hora não segue nenhum padrão. Há também um caractere especial que demonstra quando o período é horário de verão. Para essa padronização, foi necessária a análise do texto para identificar quais possuem o asterisco e gravar essa informação em outra coluna, quebra de hora e minutos conforme a estrutura padrão do banco de dados. No final, o resultado obtido foram uma estrutura igual ao utilizado nos bancos de dados, e possibilitando a importação futura.

No primeiro processo foi criado um identificador único para cada uma das coletas. Cada coleta possui mais de uma medição, e não necessariamente todas elas foram executadas em cada caso. São analisadas até 13 características em cada coleta. Cada medição possui um identificador único e este é vinculado com o identificador da coleta. O próximo passo então foi efetuar a limpeza dos dados que não foram coletados. Com essa limpeza, não foi necessário importar campos vazios, fazendo com que a importação fosse mais rápida e os recursos melhor utilizados. No total, haviam 1430 registros a serem importados. Após a limpeza, sobraram apenas 975, ou seja, 455 registros em branco que não há a necessidade de processar e registrar.

Na Figura 8, o PDI foi modelado para fazer a inclusão dos números sequenciais de identificação para ligar o número da coleta ao número da medição. Esse número necessitou ser incluído pois como se trata de tabelas diferentes para armazenamento dos dados de coletas e de medições, não havia uma opção de ligação entre os dados. Essa função é representada pela etapa 1, destacada na Figura 8.

Na etapa 2 da figura, foi elaborado o plano de padronização das horas como apresentado na Figura 8. Este plano foi necessário pois houveram muitas divergências entre os dados armazenados no campo das horas, onde este não possuía nada em comum

com os demais registros de hora. Outra dificuldade foi a indicação do horário de verão, que possuía somente um caractere padrão, mas em locais diferentes do registro. Um reconhecimento de caracteres teve de ser projetado para identificar estes caracteres especiais para a padronização.

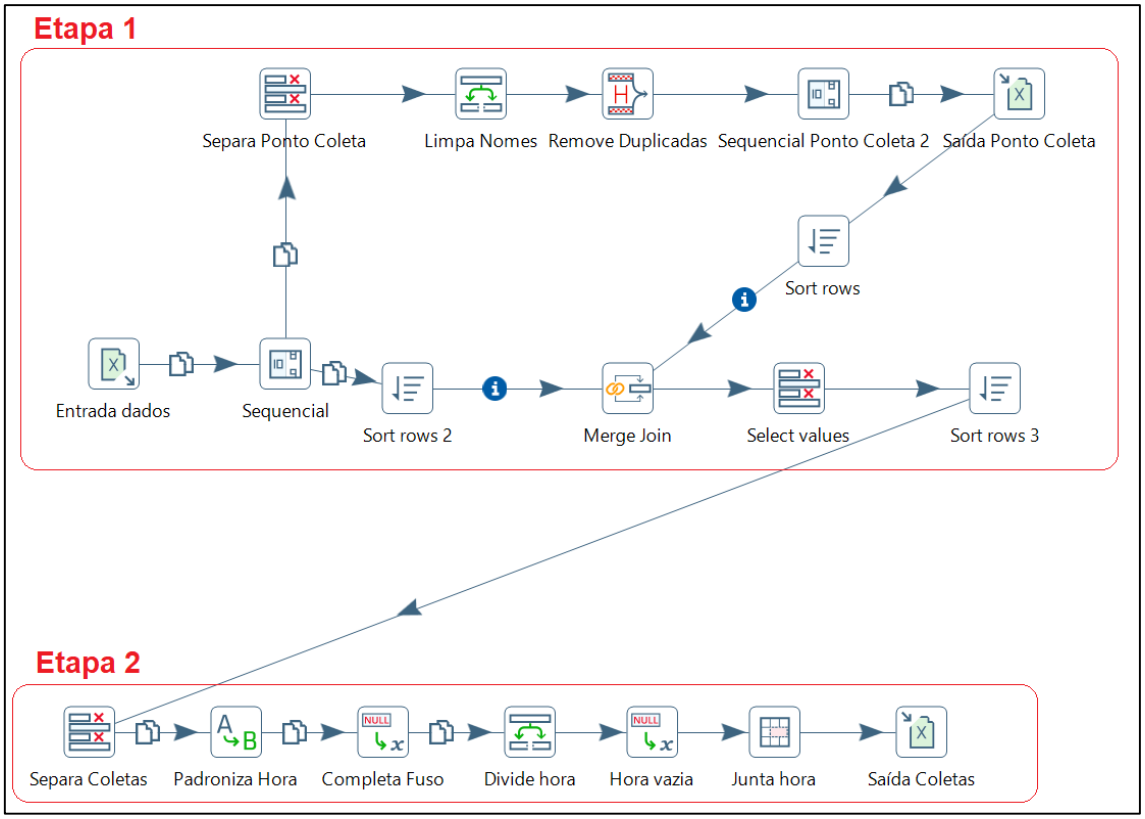


Figura 8 – Transformação dos dados das coletas

Cada um dos tipos de medição precisou ser analisado individualmente, como mostra a Figura 9, identificando no PDI qual é o tipo e código de identificação da medição, além dos valores documentados na planilha e o grau de precisão dos números decimais. Com isso, a estrutura do banco de dados teve uma otimização de performance pela opção de não armazenar dados vazios. A padronização dos dados, das medições, seguindo a mesma estrutura de casas decimais em todos os campos foi feita nessa etapa, para cada um dos tipos de medição.

Analisando os dados armazenados na planilha junto com os dados já existentes no banco de dados, foi identificado que diversas informações referentes às medições existentes no banco não estão sendo consideradas nessa planilha, como o a identificação do tipo da medição ser superficial ou de profundidade, tipos de coleta ser em arroio, rio, lagoa ou outro local. Essas informações devem ser analisadas e gravadas durante a transformação dos dados, para que, quando importados, não sejam armazenadas informações incoerentes ou vazias, fazendo com que os dados importados não fiquem completos e corretos. Essas informações possuem grande importância para análises dos tipos de medições e avaliar os locais de coleta e, para isso, tiveram que ser feitos cadastros

através da transformação, após a limpeza dos dados vazios. Essa função irá gravar as informações existentes e também os tipos de medição, assim como os locais de coleta.

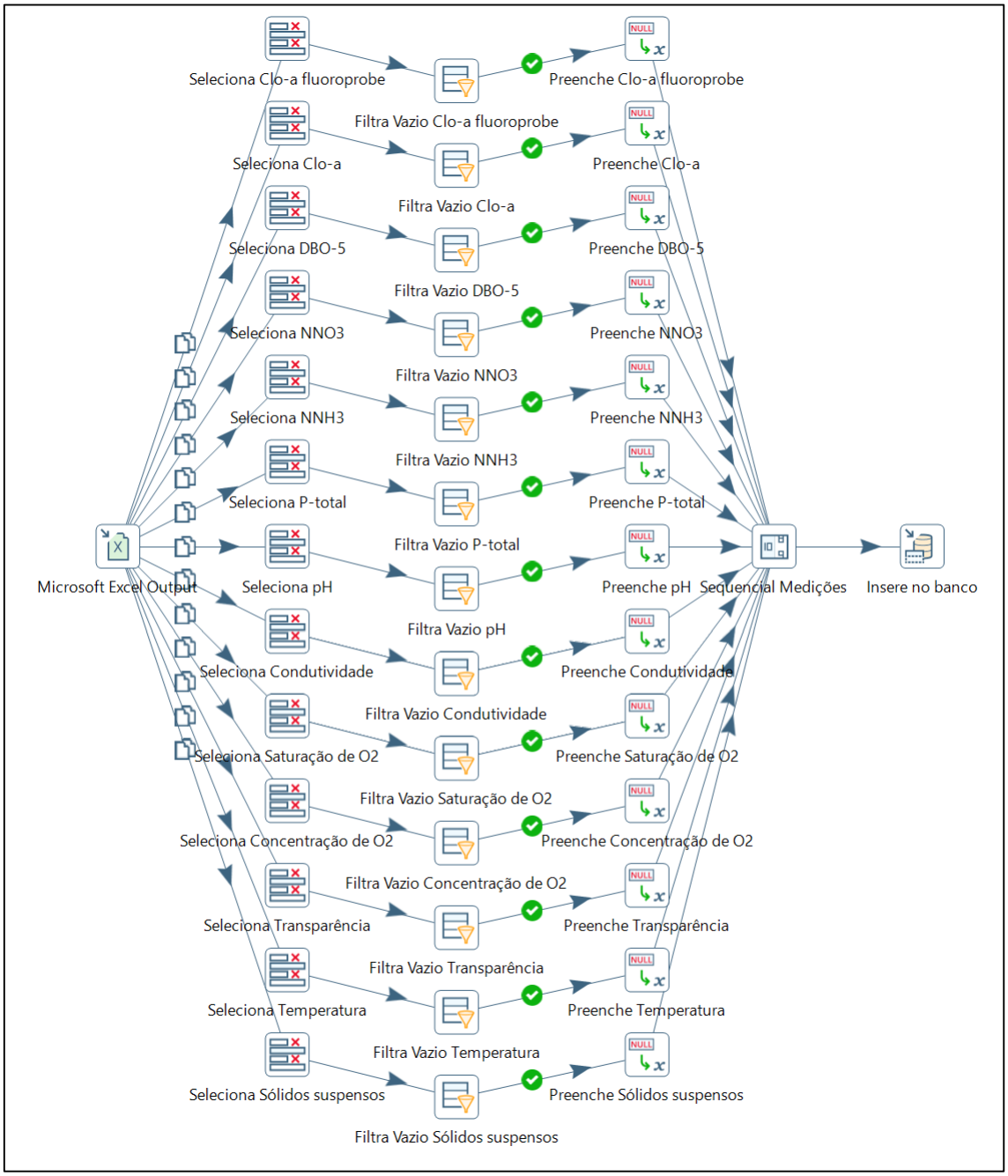


Figura 9 – Transformação dos dados das medições

Ao final destas etapas, as operações geraram arquivos de planilhas separadas para homologação dos dados. À cada execução, os arquivos eram substituídos e analisados novamente a fim de identificar as possíveis diferenças ocorridas nas mudanças dentro do PDI, que resultavam em menor tempo de resposta ou a correção de falhas que poderiam causar problemas de integridade dos dados.

Após homologados os dados de saída, o processo criado no Pentaho foi alterado para carregar os dados no banco de dados final, finalizando assim o processo de ETL.

O equipamento utilizado para a importação foi um notebook com processador i7-2670QM de 2,20 GHz, 6GB de memória RAM e Windows 10 64 bits. Cada teste foi registrado e cronometrado, e após 5 testes fazendo o mesmo processo, foi alcançada a marca de 3,8 segundos em média para todo o processamento. Nesse processamento, foi lido um arquivo de Excel de 110 linhas e 26 colunas. Ao final de todo o processo, foram gravados 975 registros de medições, dos 1430 existentes. Essa quantidade não gravada está vazia na planilha, não necessitando fazer o registro destas medições para melhorar a performance da importação e do banco de dados. Também foram gravados 32 registros de pontos de coletas diferentes. Estes pontos possuem uma informação de latitude e longitude, informação utilizada pelo sistema de georreferenciamento para mostrar no mapa o local exato da coleta.

4.6 Homologação dos dados

Após a importação dos dados para o banco de dados teste, os registros salvos foram verificados manualmente, utilizando a Tabela 2 para identificar as possíveis divergências que podem ocorrer durante o processo.

Tabela 2 – Resumo de conferência dos dados

Análise Quantitativa	Planilha	Importado	Situação
Quantidade de coletas	110	110	OK
Quantidade de medições	975	975	OK
Média das medições	35,2933	35,2933	OK
Maior medição	769,6 (Clo-a)	769,6 (Clo-a)	OK
Menor medição	0,0013 (sólidos)	0,0013 (sólidos)	OK
Comparação aleatória			
Quantidade de medições de Condutividade	79	79	OK
Quantidade de coleta no mês de Janeiro de 2015	25	25	OK
Quantidade de medições de PH maior que 10	4	4	OK
Quantidade de coletas feitas antes das 12:00	48	48	OK

Fonte: Autor (2018)

As comparações foram feitas entre a planilha original e o banco de dados. Nessas comparações, foi levado em consideração a integridade dos dados importados, com maior foco no comparativo dos números e a confiança nos valores depositados. Após isso, foram analisados os dados que foram incluídos, com as contagens de medições importadas, a fim de que toda a base de dados seja carregada de forma a suprir a necessidade do projeto.

Cada coleta pode possuir até 13 informações de diferentes tipos de medição. Baseando-se nos dados apresentados até agora, foi encontrada uma média de 8,86 medições por coleta. Para facilitar o cálculo, foi utilizada a quantidade de 9 medições em média para cada registro de coleta. Em 5 simulações foram cronometrados os tempos para o cadastro de uma coleta com 9 medições. Ao final, a média desses cadastros deu um total de 1 minuto e 48 segundos do início até a conclusão do cadastro.

Com esse número em mente, chegamos à um total de 3 horas e 42 minutos para registrar todas as 110 coletas e as 975 medições. Além do maior tempo para digitação e

retrabalho dos dados, podemos citar também a possibilidade de erro nas digitações e o retrabalho para a conferência das quantidades digitadas. A possibilidade de inserir dados errados é um dos fatores que prejudica o processo de migração de dados de uma base para outra. Este tempo foi estimado de um usuário com conhecimento do portal. Para pessoas com menos conhecimento, este número pode ser muito maior.

Ao final da importação dos dados, já existindo informações antigas e adicionando as novas, o portal LACOS Web já possui diversos dados que podem ser analisados por ferramentas de BI, consultas e acompanhamentos dos dados. O conhecimento gerado ao final do projeto mostra que podemos estruturar as decisões tomadas, baseando-as em informações conhecidas. Com o comparativo dos valores descritos neste capítulo, podemos notar que é válido o investimento do tempo em uma migração de dados, para não necessitar a digitação manual dos dados existente.

A Figura 10 mostra o layout simples das coletas. Cada uma destas pode ser visualizada separadamente, mostrando todas as medições armazenadas, com os dados identificados e padronizados para o padrão do portal.

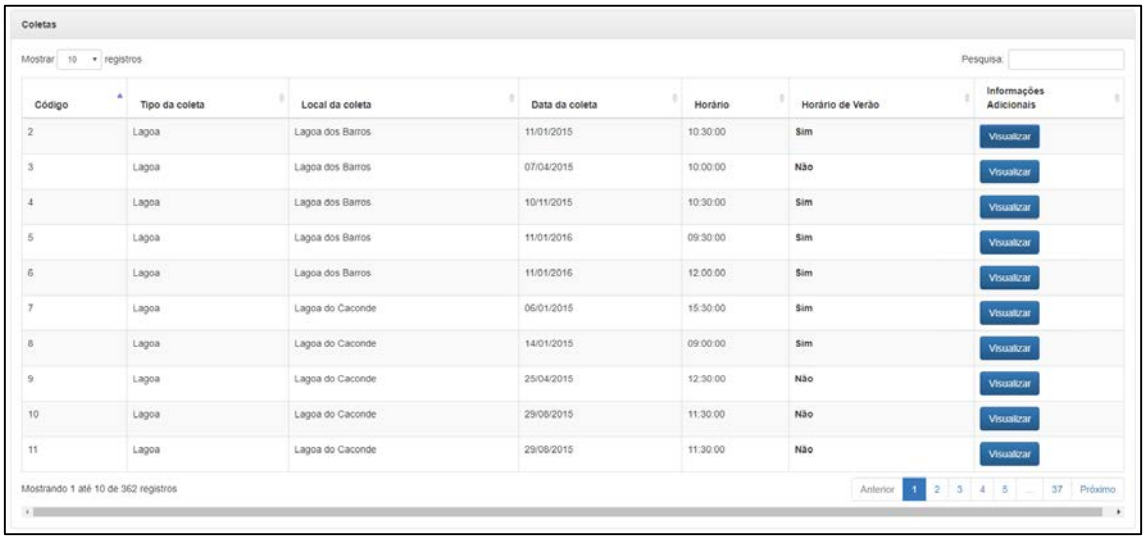
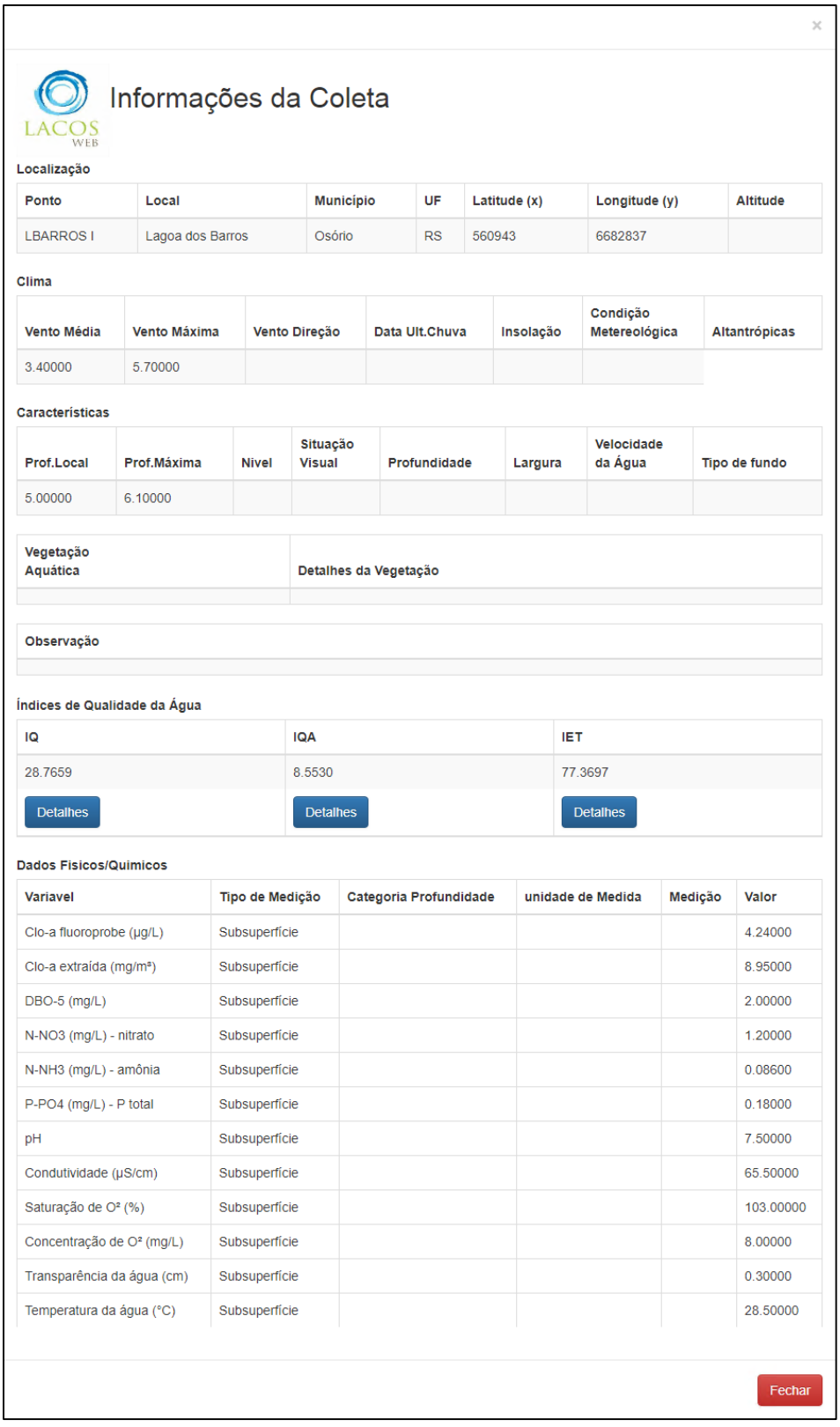


Figura 10. Tela de consulta das medições

Nessa tela, são apresentadas as coletas feitas e armazenadas, juntamente com algumas informações básicas das coletas, como data e hora do registro, os tipos de coletas e o local da coleta. Caso o usuário queira ver as informações mais detalhadas daquela coleta, a opção “Visualizar” no campo de informações adicionais irá expandir a visualização das coletas para uma mais detalhada, somente da coleta selecionada.

Na Figura 11 é apresentada a visão que o usuário que acessa o portal possui, com todas as medições que existem registradas para a coleta selecionada. As coletas possuem dados únicos, como informações de municípios e localização, além do índices de qualidade da água. Nesta visão também são apresentados os valores das medições importados ou cadastrados, com suas respectivas informações e identificações.



Informações da Coleta

Localização

Ponto	Local	Município	UF	Latitude (x)	Longitude (y)	Altitude
LBARROS I	Lagoa dos Barros	Osório	RS	560943	6682837	

Clima

Vento Média	Vento Máxima	Vento Direção	Data Ult.Chuva	Insolação	Condição Metereológica	Altantrópicas
3.40000	5.70000					

Características

Prof.Local	Prof.Máxima	Nivel	Situação Visual	Profundidade	Largura	Velocidade da Água	Tipo de fundo
5.00000	6.10000						

Vegetação Aquática	Detalhes da Vegetação

Observação

Índices de Qualidade da Água

IQ	IQA	IET
28.7659	8.5530	77.3697
Detalhes	Detalhes	Detalhes

Dados Físicos/Químicos

Variavel	Tipo de Medição	Categoria Profundidade	unidade de Medida	Medição	Valor
Clo-a fluoroprobe (µg/L)	Subsuperfície				4.24000
Clo-a extraída (mg/m³)	Subsuperfície				8.95000
DBO-5 (mg/L)	Subsuperfície				2.00000
N-NO3 (mg/L) - nitrato	Subsuperfície				1.20000
N-NH3 (mg/L) - amônia	Subsuperfície				0.08600
P-PO4 (mg/L) - P total	Subsuperfície				0.18000
pH	Subsuperfície				7.50000
Condutividade (µS/cm)	Subsuperfície				65.50000
Saturação de O² (%)	Subsuperfície				103.00000
Concentração de O² (mg/L)	Subsuperfície				8.00000
Transparência da água (cm)	Subsuperfície				0.30000
Temperatura da água (°C)	Subsuperfície				28.50000

Fechar

Figura 11. Visão detalhada da coleta e medições

5. Considerações Finais

Após o estudo e os testes das ferramentas, pode-se identificar grandes possibilidades de utilização das mesmas para diversas soluções do dia a dia das empresas. Existe também a possibilidade da utilização destas soluções em pesquisas acadêmicas, como ocorreu no objeto deste estudo. Além de todo o conhecimento adquirido no estudo e das possibilidades de utilização das ferramentas analisadas em outros projetos, devemos destacar também a utilização da solução para o meio acadêmico, enriquecendo a base de conhecimento dos dados biológicos e químicos das lagoas já examinadas. Com essa base unificada e os dados já importados, abre-se uma grande possibilidade para análise das informações e tomadas de decisões úteis para a melhora do meio ambiente em que vivemos.

Com a análise dos dados pode-se identificar padrões de ocorrências de danos ecológicos, influências que o ambiente pode causar nos recursos hídricos e o acompanhamento futuro da região, podendo ser monitorados rios e lagos a fim de constatar alterações consideráveis no ecossistema da região, identificar o que causou a transformação e armazenar as informações de maneira segura para futuras pesquisas.

Estes dados, assim como o próprio portal desenvolvido para o projeto auxiliam na análise das informações e apresentam, de maneira organizada e simples, as características do ambiente estudado. Os dados podem ser utilizados no meio acadêmico para fins didáticos e servir como exemplo de projetos a serem feitos no futuro, visando conhecer cada vez mais o nosso meio ambiente.

Referências

ABREU, Fábio Silva Gomes da Gama e. (2008). Desmistificando o conceito de ETL. Revista de Sistemas de Informação n°. 02. Macaé.

ANTONIOLLI, Solange. (2016). Projeto da área administrativa do sistema LACOS Web. Caxias do Sul.

ASPIN, Adam. (2012). SQL Server 2012 Data Integration recipes: Solutions for integration services and other ETL tools.

BANSAL, Srividya; KAGEMANN, Sebastian. (2014). Semantic Extract-Transform-Load framework for Big Data Integration. Tempe, Arizona.

BARKER, Adam; HEMERT, Jano van. (2007). Scientific Workflow: A survey and research directions. Edimburgo, Reino Unido.

BARROS, Adriano Soares de; CAMPOS, Fernando Celso de (2006). Uma discussão sobre a aplicação de processo de KDD e técnicas de mineração de dados na indústria automobilística. Bauru.

BERNSTEIN, Philip A.; HAAS, Laura M. (2008). Information integration in the enterprise.

BIGOLIN, Nara Martini. (2015). Padronização de processos: BI e KDD. Frederico Westphalen.

CASTEJON, Emiliano F.; MONTES, Antonio. (2001). Aplicação do método KDD para a detecção de eventos anômalos no tráfego de rede. São José dos Campos.

CASTERS, Matt; BOUMAN, Roland; DONGEN, Jos van. (2010). Pentaho Kettle solutions: Building open source ETL solutions with Pentaho Data Integration. Indianápolis, Indiana.

COPELLI, Ederson Luis. (2016). Proposta de módulo de avaliação da qualidade da água para o portal LACOS Web. Caxias do Sul.

DEGAN, Joyce Otsuka Côrtes. (2005). Integração de dados corporativos: uma proposta de arquitetura baseada em serviços de dados. Campinas.

ECKERSON, Wayne; WHITE, Colin. (2003). Evaluating ETL and Data Integration plataforms. Seattle, Washington.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory, SMYTH, Padhraic. (1996). From data minig to knowledge discovery in databases.

FERRANDIN, Mauri. (2002). Integrando bancos de dados heterogêneos através do padrão XML. Florianópolis.

FERREIRA, João; MIRANDA, Miguel; ABELHA, António; MACHADO, José. (2010). O processo ETL em sistemas Data Warehouse. Braga, Portugal.

FERREIRA, Lucas de Conto. (2016). Proposta de georreferenciamento das coletas para o LACOS Web. Caxias do Sul.

FERREIRA, Rick Miranda. (2014). Solução de ETL desenvolvida para o cliente CETIC.br - Centro de Estudos sobre as Tecnologias da Informática e da Comunicação. Rio de Janeiro.

FLOREA, Alexandra Maria Ioana; DIACONITA, Vlad; BOLOGA, Ramona. (2015). Data integration approaches using ETL. Bucareste, Romênia.

IGNOATTO, Maicon Luiz (2016). LACOS Web: Projeto de banco de dados para ecologia das lagoas. Caxias do Sul.

JUNIOR, Jair Sampaio Soares; QUINTELLA, Rogério Hermida. (2005). Descoberta de conhecimento em bases de dados públicas: uma proposta de estruturação metodológica. Rio de Janeiro.

JÚNIOR, Libório de Oliveira; MATOS, Simone Nasser. (2010). O descobrimento do conhecimento em base de dados com a técnica da mineração de dados e o apoio do sistema Help Desk como suporte de decisões na organização. Ponta Grossa.

KIMBAL, Ralph; CASERTA, Joe. (2004). The data warehouse ETL toolkit. Indianapolis, Indiana.

KOTOPOULIS, Alex. (2014). Best practices for real-time data warehousing. Redwood Shores, Califórnia.

LANZER, Rosane Maria. (2005). Lagoas Costeiras: Patrimônio ambiental do Rio Grande do Sul. Pelotas.

LANZER, Rosane Maria; RAMOS, Bernardo Villanueva de Castro; MARCHETT, Cassiano Alves. (2013). Impactos ambientais do turismo em lagoas costeiras do Rio Grande do Sul. Rio de Janeiro.

LINS, Matheus Icaro Agra; JUNIOR, José da Silva Duda; CUNHA, Mônica Ximenes Carneiro da. (2016). Terceirização de sistemas de informação no setor público: Uma revisão sistemática da literatura. Florianópolis.

PAIVA, Simone Bastos; ARAGÃO, Paulo Ortiz Rocha de; PEREIRA, Sandra Leandro. (2005). Gestão de conhecimento em uma organização baseada em conhecimento: uma abordagem qualitativa.

PENEDO, Janaina R.; CAPRA, Eliane P. (2012). Mineração de dados na descoberta do padrão de usuários de um sistema de educação à distância. Rio de Janeiro.

PERINI, Vinícius Lahm. (2017). Integração de ferramentas de administração e segurança BYOD. Caxias do Sul.

PRASS, Fernando Sarturi. (2012). KDD - Uma visão geral do processo.

PULVIRENTI, Adrán Sergio; ROLDÁN, María Carina. (2011). Pentaho Data Integration 4 cookbook: Over 70 recipes to solve ETL problems using Pentaho Kettle. Editora PACKT Books. Birmingham, Reino Unido.

SALES, Anderson Alex de Souza; GOMES, Georgia Regina Rodrigues. (2014). Aplicação do processo de descoberta de conhecimento em bases de dados num processo seletivo de educação a distância. Engenharia: Múltiplos Saberes e atuações, Juiz de Fora - MG.