Avaliação da ferramenta Pentaho Comunity Data Integration Através de Estudos de Caso (TCC)[1]

Daniel Rodrigues Prigol[2]
Alexandre Tagliari Lazaretti[3]
Rafael Marisco Bertei[4]

RESUMO

A sociedade da informação e comunicação está imersa em uma elevada movimentação de dados que precisam ser tratados com maior atenção. Diante disto, pode-se notar que o acesso à informação cresce em grande escala, necessitando assim, que as empresas busquem recursos tecnológicos para garantir velocidade e desempenho nos seus serviços. Este artigo apresenta a descrição de três estudos de caso implementados com o objetivo de efetuar um processo de ETL de dados de diferentes extensões e da base de dados de armazenamento para uma base de dados final para cada estudo. O cenário desses estudos é baseado em problemáticas, em que há manipulação e tratativas de formatos de dados. Para a sua realização, foi utilizada a ferramenta Pentaho Comunity Data Integration, com o principal objetivo de analisar como a ferramenta se adapta aos estudos.

Palavras-chave: integração de dados; ferramentas ETL; migração massas de dados.

1 INTRODUÇÃO

Atualmente, os sistemas que cumprem o processo de extração, transformação e carga de dados, também conhecido como ETL (Extract, Transform and Load), tornam-se fundamentais para a estratégia de gerenciamento de dados de uma empresa, segundo relatório Enterprise ETL: Evolving And Indispensible To Your Data Management Strategy (DICAS PARA COMPUTADOR, 2016).

Integrar, tratar e interpretar dados cada vez mais tem se tornado essencial no cotidiano. Segundo empresa Invisual, “a ausência de uma fundação sólida, organizada e estruturada de dados empresariais atrapalha o desempenho, a transparência e a agilidade dos processos organizacionais, tanto do ponto de vista de gestão quanto de controle” (INVISUAL, 2016).

---

[1] Trabalho de Conclusão de Curso (TCC) apresentado ao Curso de Tecnologia em Sistemas para

[2] Graduando em Tecnologia em Sistemas para Internet pelo IFSUL campus Passo Fundo. E-mail: dr.prigol@gmail.com.
[3] Orientador, professor do IFSUL. Email: alexandre.lazzaretti@passofundo.ifsul.edu.br.
[4] Co-Orientador, professor do IFSUL. Email: rafael.bertei@passofundo.ifsul.edu.br.

Através da afirmação da empresa Invisual, sistemas englobando esta nomenclatura cada vez mais estão adquirindo seu respectivo espaço. Com eles, é possível cada vez mais criar integração de dados com embasamento tecnológico abrangente do ETL, seguindo estratégias que possibilitam a otimização do sistema, bem como sua autonomia e gerenciamento. Com isso, a escolha da ferramenta ETL, Pentaho Comunity Data Integration se deu devido a um levantamento de requisitos de ferramentas descritos neste trabalho na seção “Processo ETL”.

O presente trabalho apresenta um breve referencial teórico sobre integração e o processo ETL, a fim de expor a ligação entre estas duas nomenclaturas. O principal objetivo é analisar como a ferramenta se adapta a três estudos de caso descritos na seção de “Resultados”. Dessa maneira, foi possível realizar o processo de ETL entre bases de dados e arquivos de texto.

2 REFERENCIAL TEÓRICO

2.1 INTEGRAÇÃO DE BANCO DE DADOS

O processo de integração de dados é relevante já que dados oriundos de diferentes fontes podem ser manipulados de forma única dentro de um determinado domínio de aplicação (BARBOSA, 2001).

O principal foco da integração de dados segundo Kakugawa (2010), é a interação de dados de diferentes origens e destinos. Com a finalidade de tratá-los e gerar um resultado mais satisfatório diminuindo significativamente erros, perda e uso indevido de dados.

Quando se fala de integração de dados no que se refere a este trabalho pode-se destacar o processo ETL, sigla designada para dirimir extract, transform e load(extrair, transformar e carregar) (PENTAHO, 2016). Na próxima seção será explicado acerca deste processo.

2.1.1 PROCESSO ETL

Considera-se processo ETL, um conjunto de processos para trazer dados de sistemas para uma base de dados, providos não só de sistemas, mas também de websites, bases de e-mails e de redes sociais, arquivos de texto dos mais variados contextos e bases de dados pessoais (TANAKA, 2015).

O principal uso do processo ETL provém desde de um carregamento de uma base de dados, como também de uma diversidade de outros estudos de casos, sendo alguns deles capaz de gerar planilhas e modelos de base de dados, podendo prover retorno de dados para sistemas, com o intuito de resguardá-los e eliminar possíveis falhas que possam estar acontecendo (PENTAHO, 2016).

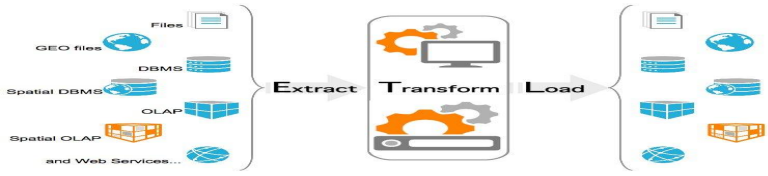
A terminologia ETL diz respeito a ferramentas que viabilizam a extração de um conjunto de dados oriundos de uma ou inúmeras origens, transformação e persistência de informações de um local para outro, ou seja, esse tipo de ferramenta pode ser utilizada em tarefas de modernização, readequação, validação e integração de massas de dados (KIMBALL; CASERTA, 2004).

Segundo Kimball e Caserta (2004), pode-se delimitar a abordagem ETL da seguinte forma:

- Extração (Extract): processamento necessário para conectar às fontes de dados, extraí-los e torná-los disponíveis para os passos subsequentes;
- Transformação (Transform): quaisquer funções aplicadas sobre os dados extraídos desde a extração das fontes até o carregamento nos alvos;
- Carregamento (Load): todo processamento requerido para carregar os dados no sistema alvo ou em um sistema simulado.

A Figura 1 ilustra as etapas do processo ETL, demonstrando as diferentes fontes que pode-se tratar.

Figura 1 - Ilustração do processo de ETL



Fonte: dbbest, 2016.

Para a construção de um sistema de ETL é preciso ter uma projeção bem definida da fonte de dados; das limitações destes dados; das linguagens utilizadas e suportadas; das ferramentas de ETL disponíveis e que atendam as necessidades;

das plataformas de Business Intelligence; dos inúmeros e diversos formatos de arquivo (TANAKA, 2015).

Após a análise destes requisitos apontados por Tanaka, foram projetados 3 estudos de casos, considerando possíveis problemas a serem enfrentados ao manipular dados. Esses estudos de caso estão descritos na seção de resultados.

A ferramenta Pentaho Comunity Data Integration(PDI) foi escolhida através de pesquisas baseadas em funcionalidades dentre ferramentas que realizam o processo ETL, adequando-a aos requisitos da tabela 1.

Tabela 1 – Análise da ferramenta

Número	Análise
1	**Suporte das plataformas**
2	**Possibilidade de integração com outras ferramentas e serviços**
3	**Carga de dados de diversas Origens**
4	**Serviço na nuvem**
5	**Possibilidade de customização de ferramenta quanto ao código (opensource)**
6	**Suporte de linguagem de programação**
7	**Execução das tarefas sem intervenção humana**

Fonte: Do Autor.

A tabela 1 lista um conjunto de possíveis funcionalidades que podem ser utilizadas na ferramenta PDI, que atendeu a todos os requisitos da tabela.

Na próxima seção será contextualizada a ferramenta e como ela se adequou aos requisitos listados na Tabela 1.

2.2 Pentaho Comunity Data Integration

Caracterizada como uma suíte de aplicativos Open Source para criação de Business Intelligence (BI), conhecida como Kettle ou PDI, a ferramenta Pentaho começou a ser construída em meados dos anos 2000, quando Matt Casters, um dos fundadores do Kettle Project, encontrou problemas com ferramentas de integração. Por isso, criou e trabalhou na elaboração de uma nova ferramenta que pudesse prover recursos de ETL (CASTERS; BOUMAN; DONGEN, 2010).

É importante levar em conta que a Suite Pentaho BI possui muitas funcionalidades, dentre as quais a principal é um mecanismo que provê soluções para as tarefas de ETL em projetos de integração.

Seus módulos mais importantes, segundo o site da tecnologia, são: Pentaho BI Plataform; Pentaho Data Integration – PDI - Kettle; Analysis View - Mondrian; Pentaho Reporting - Reporting; Weka - Data Mining (PENTAHO, 2016).

Como a maioria das ferramentas de integração, em relação aos recursos, possibilita a inserção de transformações e validações de dados baseadas em trechos de código escritos em linguagens de programação. No caso do Pentaho, destacam-se Java e JavaScript, a integração do workflow de ETL com sistemas externos, serviços de Cloud, entre outros (PENTAHO, 2016).

Logo, quanto aos drivers suportados para conexões de entrada e saída de dados, podem ser citados Oracle, MySQL, PostgreSQL, Sybase, Firebird SQL, Ingres, Borland Interbase, Oracle RDB, IBM Universe, SQLite, entre outros, além de drivers de conexão com a nuvem como SalesForce (PENTAHO, 2016).

3 RESULTADOS

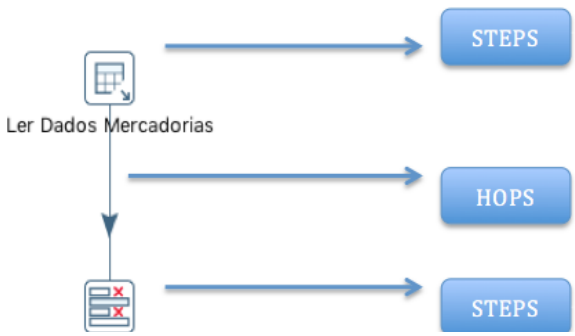
3.1 METODOLOGIA DO ESTUDO

Para avaliar a ferramenta PDI no processo ETL, faz-se necessário o conhecimento de algumas de suas funcionalidades. Dentre elas, destacam-se, neste trabalho, as seguintes: Step's, Hops e Transformação.

Step's são componentes utilizados para realização das tarefas a serem realizadas pelo usuário e suas respectivas transformações atendidas pela ferramenta, ou seja, são unidades mínimas de transformações.

Para ligar um step ao outro, são necessários segmentos que se denominam hops. A Figura 2 demonstra um exemplo de steps e hops.

Figura 2 - Exemplo da implementação de steps e hops



Fonte: Do autor.

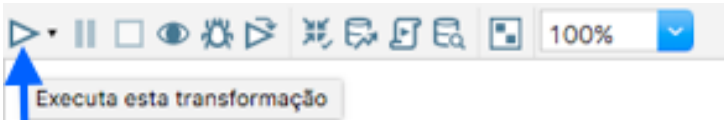
Transformações são entidades formadas pelos steps ligados através de hops, utilizados na manipulação do fluxo de dados em um workflow, que é um conjunto de transformações que pode compor um job. Job é uma entidade criada para execução de um processo.

Uma transformação pode ser criada através do menu FILE > Novo > TRANSFORMAÇÃO.

Para a realização deste trabalho, somente foram utilizadas as transformações, pois os job's estão mais relacionados à execução das transformações e não ao tratamento dos dados.

Por meio dessas definições sobre a ferramenta, foi iniciada uma nova transformação para cada estudo de caso. Para ser possível testar as transformações, basta clicar no símbolo convencional, localizado no canto superior esquerdo do menu do workflow, como demonstra a Figura 3.

Figura 3 - Executar a transformação



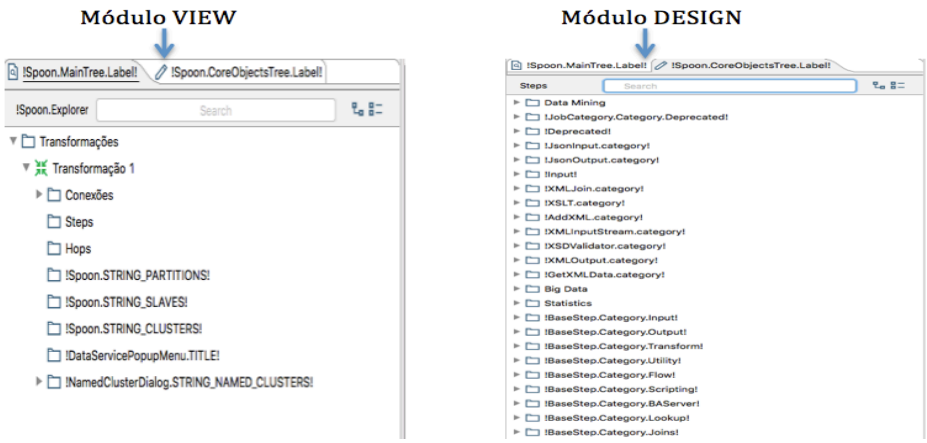
Fonte: Do autor.

Após, é preciso clicar em "Launch" (botão do meio na parte inferior), para que a transformação seja executada.

A ferramenta é dotada de dois módulos, view e design. Em view é possível obter tudo o que foi utilizado ou carregado no projeto, por exemplo: steps, hops e as conexões utilizadas para as integrações. Este módulo atua como uma espécie de recuperador de processo. Além disso, é possível criar os componentes ou alterar os que já foram criados no projeto.

No módulo design, é possível carregar para o workflow todos os steps a serem utilizados no projeto. A Figura 4 ilustra um exemplo dos módulos view e design.

Figura 4 - Ilustração dos módulos



Fonte: Do Autor.

Ao observar a Figura 4, nota-se que, através do campo de pesquisa, é possível procurar por elementos, de acordo com o módulo em que está associado. Essas pesquisas podem ser feitas tanto pelo nome que pode ser dado ao elemento, quanto pelo nome do elemento, por exemplo: conexões, steps, hops ou o nome dado à escolha do usuário, como “Selecionar dados de Vendas” ou Ler Dados de Pessoas etc.

Com o intuito de extrair e carregar dados, respectivamente na base de dados de saída e de entrada, foi necessário conectá-las com a ferramenta. Isso foi possível devido às bases de dados já se encontrarem previamente criadas no sgbd.

Para realização desta tarefa, clica-se na aba View e depois, com o botão direito, em Conexões. Dessa forma, abre uma caixa de diálogo que permite clicar em Novo.

A configuração da conexão varia de acordo com o sgbd em que se está trabalhando.

As conexões podem ser realizadas sem a necessidade de instalação de drives para reconhecimento dos bancos de dados.

A ferramenta dispõe de criação de conexões simultâneas, ou seja, pode-se conectar com diversas bases de dados de múltiplos sgbd’s.

Uma vez criada, essas conexões podem ser utilizadas em uma única ou ainda em diversas transformações, conforme a necessidade do usuário.

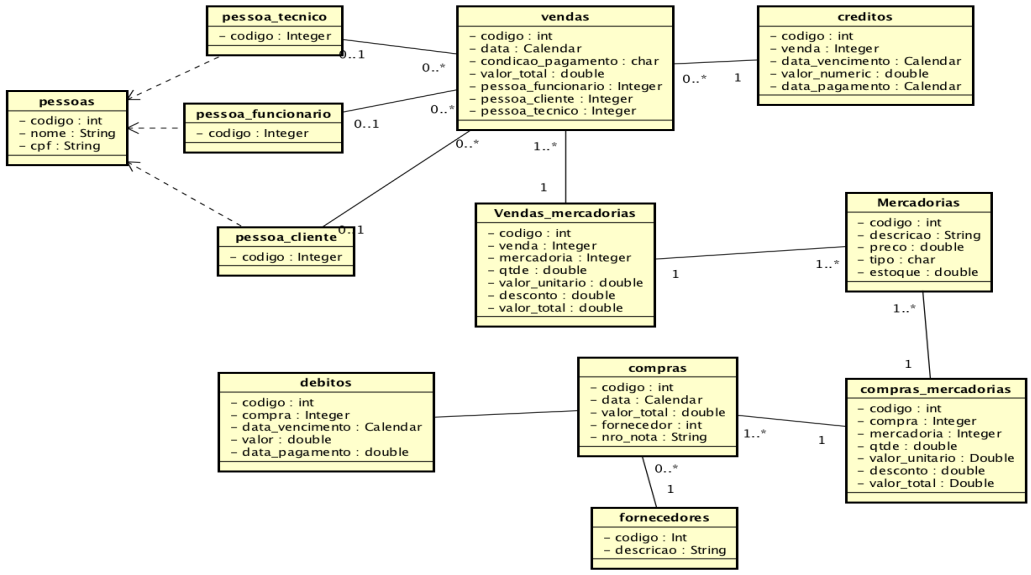
Para este trabalho considera-se que um arquivo estruturado possui dados que se apresentam de forma organizada, seguindo um padrão em suas colunas e um

arquivo semiestruturado possui dados que não seguem uma ordem se apresentando desorganizadamente. Na próxima seção esta descrito o primeiro estudo de caso.

3.2 Descrição do estudo de caso 1

O estudo de caso 1 está relacionado com a integração de dois bancos de dados com o mesmo esquema. O sgbd utilizado é o postgresql. O objetivo deste estudo de caso é avaliar como a ferramenta se comporta em relação a este tipo de integração. O modelo conceitual dos bancos de dados é descrito na Figura 5.

Figura 5 - Diagrama de classes do banco de saída dos dados



Fonte: Do autor.

O diagrama da Figura 5 refere-se a uma base remota que será integrada a uma base de dados centralizada. Ele reflete um sistema de compra, venda e manutenção, no qual armazenam-se pessoas que podem ser clientes, ou funcionários e/ou técnicos. Também possui o controle de compras e vendas.

Uma das possibilidades a ser avaliada neste estudo de caso é o tratamento da integridade referencial, ou seja, da relação entre chaves primárias e chaves estrangeiras.

3.3 Resultado do estudo de caso 1

Para avaliar o uso da ferramenta PDI, optou-se por implementar a integração entre vendas e mercadorias, podendo detalhar todos os requisitos levantados na problemática do estudo de caso.

Primeiramente, é necessário criar uma nova transformação e estar configurada a conexão da ferramenta com o banco de dados. Após realizadas as devidas configurações, no módulo design, deve-se buscar através do campo de pesquisa ou através do menu pelo step “Table Input” que retorna os dados do banco de saída que deseja manipular.

Para utilizar o step, basta clicar nele e arrastá-lo para o workflow. Para este step foi dado o nome de “Ler dados vendas” e, após, foi conectado com a base de dados de saída. Para isso, deve-se clicar no botão “!BaseStepDialog.WizardConnectionButton.Label”, localizado logo abaixo do nome do step.

Esse botão, redireciona para a configuração de uma conexão, sendo que, deve-se colocar um nome para esta conexão, selecionar o nome do sgbd, após realizadas estas configurações, deve-se seguir os próximos passos da configuração.

Para selecionar os dados deste step, deve-se buscar por um novo step de nome “Select”, no campo de pesquisa do módulo design. Deve-se arrastar este step para o workflow e conectar o step anterior neste step, através do hop. Para isso, deve-se ir com o *mouse* sobre o step anterior e clicar no quarto ícone, após isso será habilitada uma seta quando se clica sobre o step a ser conectado.

Após a ligação entre os steps, configura-se o step “Select”. Para ser possível obter os dados da tabela desejada, basta clicar no botão “Get Select”, o qual vai retornar todas as colunas da tabela selecionada, podendo-se selecionar ou excluir a coluna desejada. Após realizadas as configurações, deve-se clicar em “OK”.

Para inserir na nova base de dados, deve-se buscar por “Output” na barra de pesquisa. Deve-se arrastar o step para o workflow, conectar o step anterior através do hop e configurá-lo. Para acessar as configurações, basta seguir os mesmos passos realizados anteriormente.

Dentro das configurações deste step, foi alterado o nome para “Inserir dados vendas”, devendo-se configurar uma conexão da mesma forma que foi o step “Table Input”. No passo seguinte, colocou-se o nome do banco de entrada de dados. Nesse

caso, foi nomeado como “PC2_entrada. O restante das configurações foram as mesmas utilizadas anteriormente.

Foi inserido o nome da tabela que receberá os dados no campo “TargetTable”. Para que sejam transferidos os dados, acionou-se o botão “SQL”, localizado no canto inferior direito, para que fosse criada uma tabela no banco de dados de entrada.

Se tudo foi configurado corretamente, abre uma tela informando um sql de criação de uma nova tabela, deve-se clicar em “Execute”.

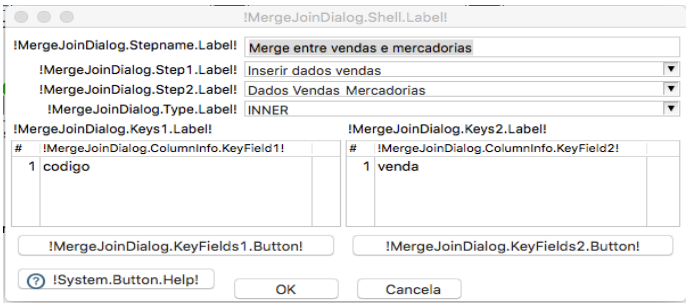
Após realizada a configuração do step clicou-se em “OK”. Para verificar se os dados irão ser inseridos, é preciso executar a transformação. Assim, a ferramenta retorna um console na parte inferior do workflow, demonstrando se ocorreu sucesso ou não. Caso tenha ocorrido sucesso, neste console é demonstrado todo o processo que foi realizado de entrada e saída de dados e aparece um ícone de certo(√) em cada step que está dentro do workflow.

Após realizada a transformação dos dados da venda, o mesmo processo foi adaptado para os itens da venda. Depois de configurados os steps para os itens da venda, deve-se clicar em executar novamente.

O banco de dados de saída possui uma ligação de muitos para muitos entre essas tabelas e uma tabela de ligação chamada de “vendas_mercadorias”. Foi realizado o mesmo processo das transformações anteriores para retornar aos dados desta tabela. Para ser possível tratar os dados e realizar um merge entre as tabelas, é necessário buscar pelo step chamado de “MergeJoin. A seguir, arrastou-se o step para o workflow realizar a conexão entre o step de output de “vendas” e o step de output de “venda_mercadorias” e configurar da mesma forma que os outros steps.

A seguir, alterou-se o nome para “Merge entre vendas e mercadorias”, selecionou-se o nome do step de vendas e o nome do step de vendas_mercadorias. Posteriormente, selecionaram-se as chaves a serem mergeadas entre os steps, como ilustrado na Figura 6.

Figura 6 - Merge entre vendas e vendas_mercadorias



Fonte: Do autor.

Depois, realizou-se o mesmo processo e configurou-se um step de merge entre mercadorias e vendas_mercadorias. Para buscar os dados e realizar o merge entre as chaves e a ferramenta e poder tratar a integridade referencial, selecionaram-se e inseriram-se os dados com o step “Dimension Update”. Buscou-se por este step, arrastou-se o step para o workflow, configurou-se o step com o nome “Dimensão de vendas”. A conexão com o banco de entrada de dados e o nome para a nova tabela ficou “dimensão_vendas”.

Então, ligou-se código com código e configurou-se a nova chave como “codigo_id”. Com a aba “FieldsTab” selecionada, clicou-se no botão “GetFields”, localizado no canto inferior, para obter os dados que vêm do merge.

Para enviar os dados para o banco, clicou-se no botão de SQL, localizado no canto inferior direito, que criou uma tabela para recebimento dos dados. Abriu uma caixa de diálogo com o sql a ser inserido, então clicou-se em “Executar” e após em “OK” do step.

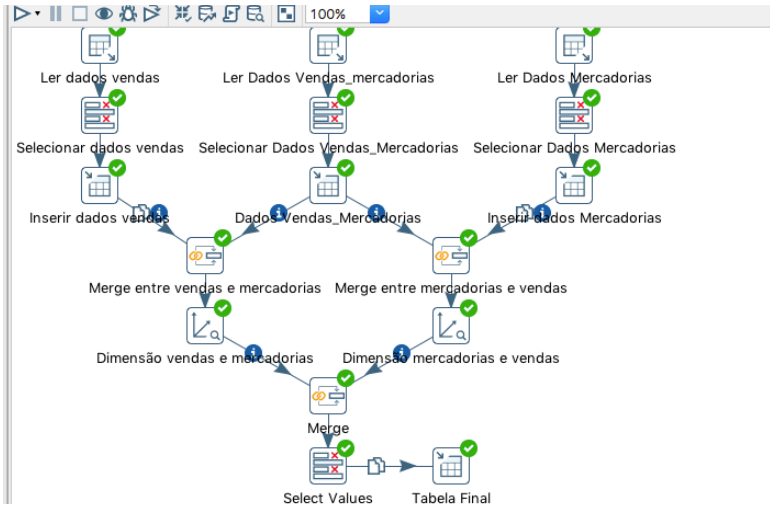
Para realizar o mesmo processo para o step de mercadorias a workflow, foi necessário realizar mais um merge, utilizando como chaves os códigos que foram enviados dos step’s anteriores. Através do step “select values”, selecionaram-se os dados desejados para a nova tabela.

Optou-se por inserir os dados finais, com o step “Output”, pois a integridade estava sendo controlada nos steps “Dimension”.

Após finalizadas as configurações, foi executada novamente a transformação e os dados foram transferidos para as devidas tabelas corretamente.

A Figura 7 ilustra como a ferramenta se apresenta após executar a workflow.

Figura 7 – Workflow da ferramenta após a execução

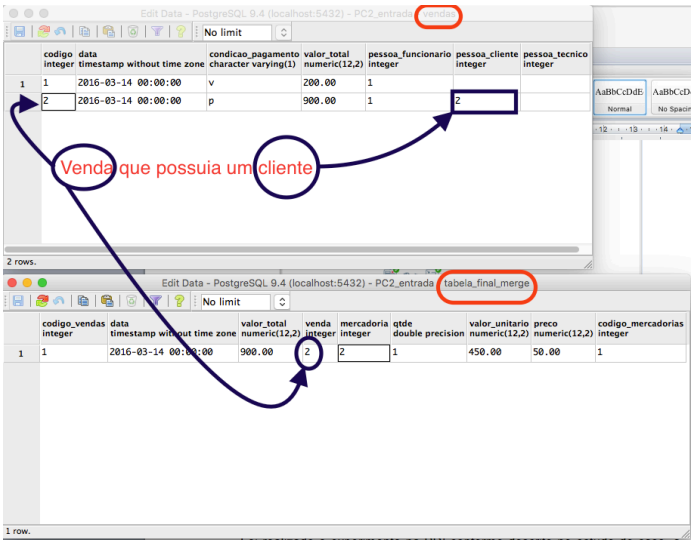


Fonte: Do autor.

Foi realizado um controle para que a ferramenta armazenasse as vendas somente quando tivesse dados de cliente. Para isso, foi configurada a dimensão entre vendas e mercadorias.

Como ilustrado na Figura 8, somente foi armazenada na “Tabela Final” a venda que possuía cliente.

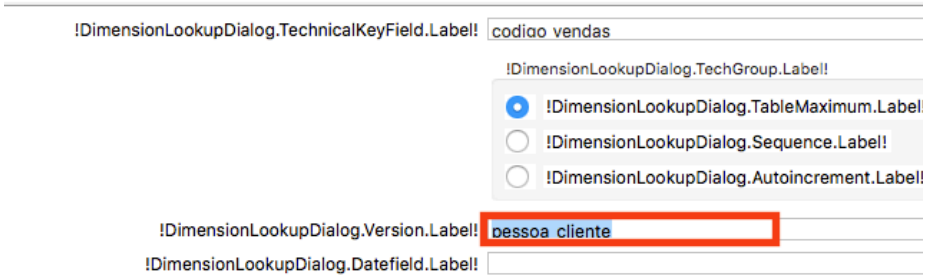
Figura 8 - Tabelas do banco de dados



Fonte: Do autor.

Para que a ferramenta obtivesse esse nível de abstração de merge, foi realizado o controle no step “Dimensão vendas e mercadorias” e alterada a versão para filtrar por pessoa, como ilustrado na Figura 9.

Figura 9 - Controle do merge



Fonte: Do autor.

Para as outras tabelas, foram realizadas as mesmas configurações e tratativas.

3.4 Estudo de caso 2

O estudo de caso 2 está relacionado com a integração de um arquivo de texto estruturado com um banco de dados. O sgbd utilizado é o PostgreSQL. O objetivo deste estudo de caso é avaliar como a ferramenta se comporta em relação à integração de arquivos de texto estruturados. Um exemplo para este tipo de arquivo é ilustrado na Figura 10.

Figura 10 - Exemplo de arquivo estruturado com separadores

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	24505420	01/01/12	0	974.2			22.1	21.3	22.8	0	96.9	1.6	44	3		0	0
2	24505420	01/01/12	1	973.5			21.8	21.3	22.5	0	96.4	1.7	40	2.6		0	0
3	24505420	01/01/12	2	973			21.2	20.8	21.9	0	97	1.6	50	2.4		0	0
4	24505420	01/01/12	3	972.5			20.9	20.4	21.5	0	98.2	1.6	61	2.4		0	0
5	24505420	01/01/12	4	972.2			20	18.9	21	0	98.1	1.4	47	2.3		0	0
6	24505420	01/01/12	5	972.8			19.2	18.2	20.4	0	99.1	1.1	39	2		0	0
7	24505420	01/01/12	6	973.4			17.8	16.7	19.5	1	98.1	0.5	77	1.4		0	0
8	24505420	01/01/12	7	974			18.3	16.8	20.8	99	95.9	0.9	68	1.4		0	0
9	24505420	01/01/12	8	974.5			22	20.4	23.6	297	87.4	0.6	39	1.5		0	0
10	24505420	01/01/12	9	974.9			23.2	21.7	24.6	514	83.5	1	294	2.7		0	0
11	24505420	01/01/12	10	975.1			25.5	24.1	27	710	73.2	1.3	277	3		0	0
12	24505420	01/01/12	11	975.1			27.4	26.4	28.4	867	65.4	2.8	219	5.7		0	0
13	24505420	01/01/12	12	974.9			28.4	27.3	29.6	970	57.2	3.4	215	5.8		0	0
14	24505420	01/01/12	13	974.2			29.2	28.4	30.3	1007	53.5	3.3	237	5.9		0	0
15	24505420	01/01/12	14	973.6			30.3	29.4	31.3	990	49.6	3.4	216	6.2		0	0
16	24505420	01/01/12	15	973			31	29.9	32	917	45.9	3.7	205	7.4		0	0
17	24505420	01/01/12	16	972.4			31.5	30.9	32.4	767	41.2	3.9	239	7.7		0	0
18	24505420	01/01/12	17	971.9			31.2	29.9	32.1	549	41.9	3.9	229	6.7		0	0

Fonte: Do autor.

Uma das possibilidades a ser avaliada é o tratamento do cabeçalho das colunas e seus respectivos dados. A ilustração da Figura 10 refere-se há uma base

remota que será integrada a uma base de dados centralizada. A base remota reflete um arquivo de texto que se apresenta organizado em forma de colunas, sem cabeçalho.

3.5 Resultado do Estudo de Caso 2

Para estudar o uso da ferramenta, optou-se por criar todas as colunas do arquivo na ferramenta. Primeiramente foi necessário criar uma nova transformação e configurar a conexão da ferramenta com o banco de dados.

Para trabalhar com arquivos ".CSV[5]", a ferramenta já possui um step específico que deve ser colocado no workflow e configurado. Nas configurações, alterou-se o nome para "CSV". Para que a ferramenta buscasse o arquivo, clicou-se no botão "Navegar", localizado no canto superior direito. Neste caso, o delimitador de espaço entre os dados foi o ponto-e-vírgula (;), inserido nesta configuração.

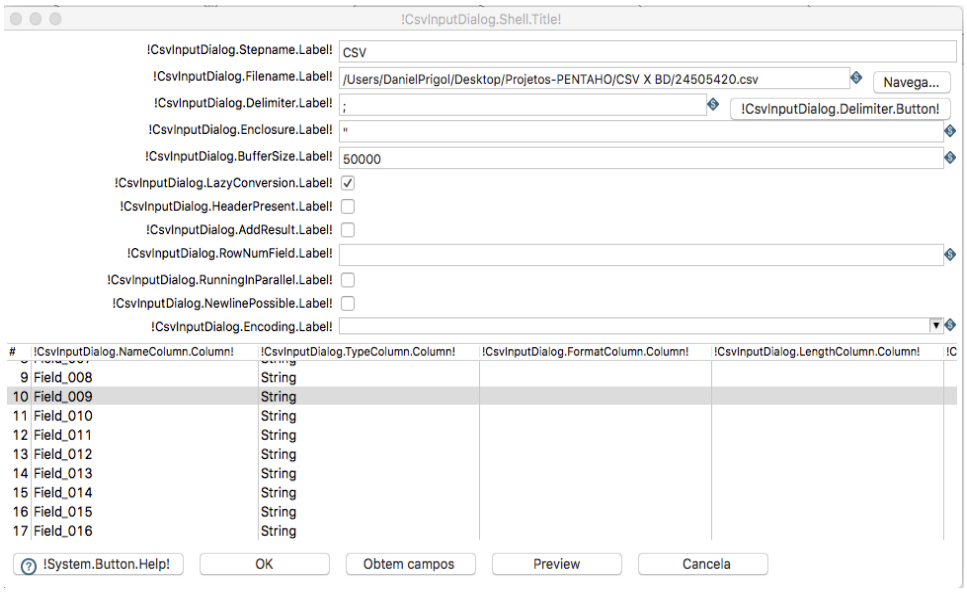
Na parte inferior da configuração, foi possível buscar pelas colunas do arquivo de texto. Como o arquivo não possuía cabeçalho, a ferramenta nomeou automaticamente cada coluna como "Field" e um número em ordem cronológica.

Por exemplo, neste caso, o arquivo de texto possuía 17 colunas separadas por ponto-e-vírgula (;), a ferramenta buscou por estas colunas e, como não havia cabeçalho para nomeá-las, a ferramenta nomeou automaticamente como field_000 text, field_001 text, field_002 text, sucessivamente até chegar ao field_016 text.

É possível alterar estas colunas para um nome mais amigável. No caso deste estudo, não foi alterado o nome que a ferramenta ordenou. Após isso, finalizaram-se as configurações clicando em OK. As configurações podem ser visualizadas na Figura 11.

[5] CSV Comma-separated values – É um formato de arquivo de texto.

Figura 11 – Configuração do step CSV estruturado



Fonte: Do autor.

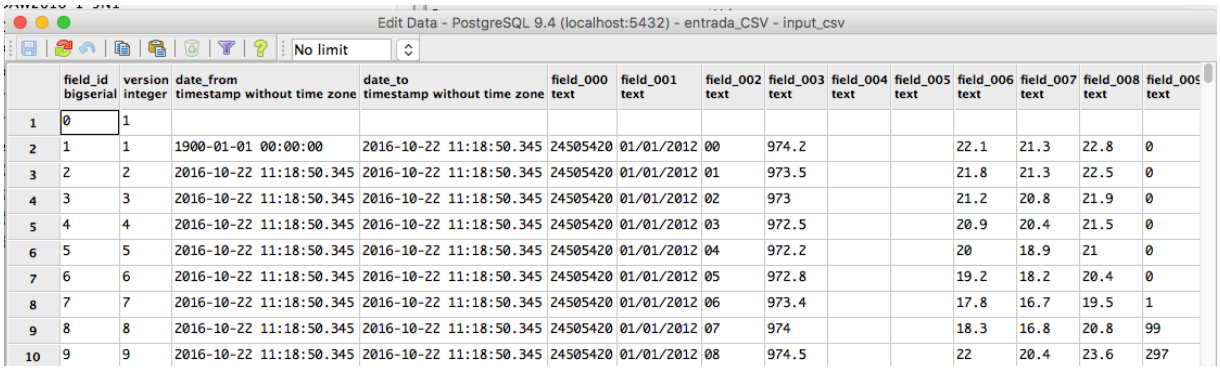
Com o step do arquivo CSV configurado, buscou-se por um step que realiza o insert ou update chamado de “Dimension Update”. Ao configurá-lo, foi conectado o step CSV ao step de output, lembrando que, caso não exista a tabela, ela deve ser criada na ferramenta. A workflow se apresenta como ilustrada na Figura 12.

Figura 12 – Workflow após as configurações e ligações

Fonte: Do autor.

Após executada o workflow, os dados foram transferidos, para uma nova tabela sem perdas, tratando a integridade referencial, cuidando-se para que os arquivos não fossem replicados. A Figura 13 ilustra como ficou a tabela do banco de dados. Nela é possível notar que obtiveram-se um id e uma versão, que foram inseridos através da ferramenta para fins de controle de integridade.

Figura 13 – Resultado após a execução da transformação



Edit Data - PostgreSQL 9.4 (localhost:5432) - entrada_CSV - input_csv

No limit

	field_id bigserial	version integer	date_from timestamp without time zone	date_to timestamp without time zone	field_000 text	field_001 text	field_002 text	field_003 text	field_004 text	field_005 text	field_006 text	field_007 text	field_008 text	field_009 text
1	0	1												
2	1	1	1900-01-01 00:00:00	2016-10-22 11:18:50.345	24505420	01/01/2012	00	974.2			22.1	21.3	22.8	0
3	2	2	2016-10-22 11:18:50.345	2016-10-22 11:18:50.345	24505420	01/01/2012	01	973.5			21.8	21.3	22.5	0
4	3	3	2016-10-22 11:18:50.345	2016-10-22 11:18:50.345	24505420	01/01/2012	02	973			21.2	20.8	21.9	0
5	4	4	2016-10-22 11:18:50.345	2016-10-22 11:18:50.345	24505420	01/01/2012	03	972.5			20.9	20.4	21.5	0
6	5	5	2016-10-22 11:18:50.345	2016-10-22 11:18:50.345	24505420	01/01/2012	04	972.2			20	18.9	21	0
7	6	6	2016-10-22 11:18:50.345	2016-10-22 11:18:50.345	24505420	01/01/2012	05	972.8			19.2	18.2	20.4	0
8	7	7	2016-10-22 11:18:50.345	2016-10-22 11:18:50.345	24505420	01/01/2012	06	973.4			17.8	16.7	19.5	1
9	8	8	2016-10-22 11:18:50.345	2016-10-22 11:18:50.345	24505420	01/01/2012	07	974			18.3	16.8	20.8	99
10	9	9	2016-10-22 11:18:50.345	2016-10-22 11:18:50.345	24505420	01/01/2012	08	974.5			22	20.4	23.6	297

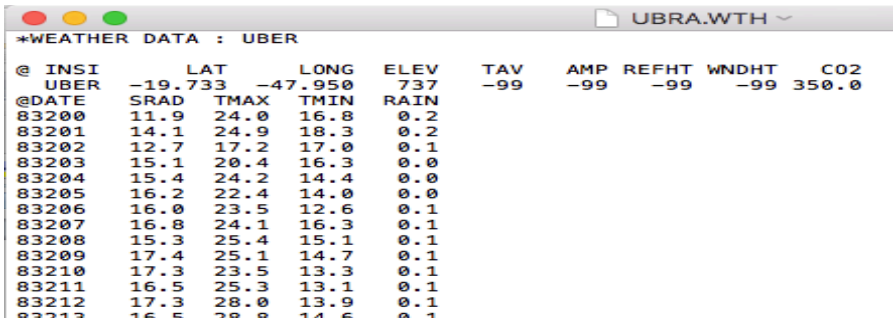
Fonte: Do autor.

3.6 Descrição do Estudo de caso 3

O estudo de caso 3 está relacionado com a integração de um arquivo de texto semiestruturado, de extensão ".WTH[6]", com um banco de dados implementado no sgbd PostgreSQL.

Um exemplo para este tipo de arquivo foi ilustrado na Figura 14.

Figura 14 - Exemplo de arquivo semiestruturado



UBRA.WTH

```
*WEATHER DATA : UBER

@ INSI      LAT      LONG   ELEV    TAV    AMP REFHT WNDHT    CO2
  UBER   -19.733   -47.950    737    -99    -99    -99    -99  350.0
@DATE    SRAD   TMAX   TMIN   RAIN
83200    11.9   24.0   16.8    0.2
83201    14.1   24.9   18.3    0.2
83202    12.7   17.2   17.0    0.1
83203    15.1   20.4   16.3    0.0
83204    15.4   24.2   14.4    0.0
83205    16.2   22.4   14.0    0.0
83206    16.0   23.5   12.6    0.1
83207    16.8   24.1   16.3    0.1
83208    15.3   25.4   15.1    0.1
83209    17.4   25.1   14.7    0.1
83210    17.3   23.5   13.3    0.1
83211    16.5   25.3   13.1    0.1
83212    17.3   28.0   13.9    0.1
```

Fonte: Do autor.

Para esse estudo avaliou-se os mesmos critérios do estudo de caso 2.

3.7 Resultado do estudo de caso 3

Para estudar o uso da ferramenta, também optou-se por criar todas as colunas das tabelas na ferramenta.

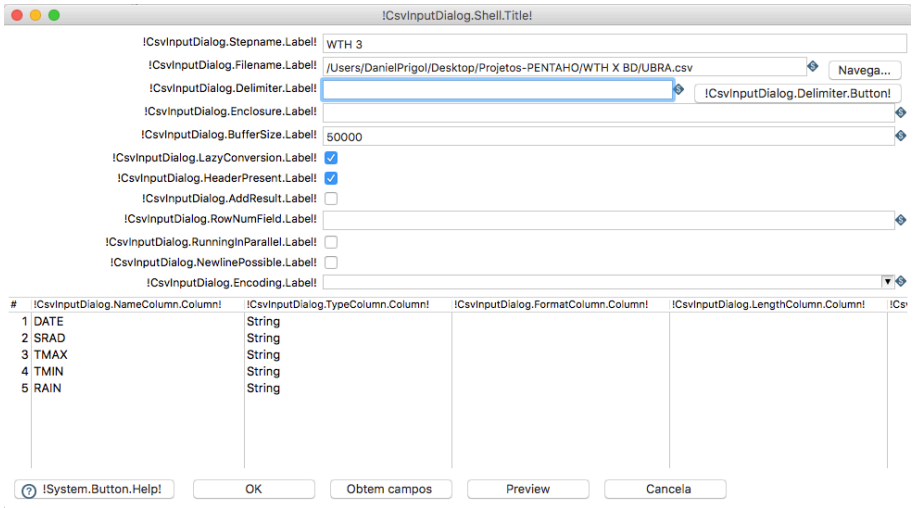
[6] WTH – É um formato de arquivo de texto comum no armazenamento de dados temporais.

Primeiramente, foi necessário criar uma nova transformação e configurar a conexão da ferramenta com o banco de dados. Devido a questões de reconhecimento de arquivo, a ferramenta não apresentou um step que identifique arquivos de textos desta extensão. Com isso, foi alterado o tipo de extensão para “.CSV” visando o reconhecimento do arquivo de texto juntamente com a ferramenta. Para recuperar o arquivo, foi utilizado o mesmo tratamento do estudo de caso 2.

Neste caso, o campo que preenche o delimitador foi deixado em branco e selecionado o campo header, pois o arquivo não possuía separador e possuía cabeçalho.

E, por fim, com o objetivo de retornar os dados, acionou-se o botão “Obtem Campos”. Como havia cabeçalho, a ferramenta nomeou automaticamente cada coluna buscando cada valor de cabeçalho. Pôde-se alterar estas colunas para o nome que achar devidamente necessário. Essas configurações são ilustradas na Figura 15.

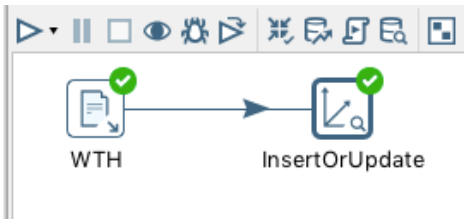
Figura 15 – Configuração do step CSV semiestruturado



Fonte: Do autor.

Com o step do arquivo CSV configurado, é preciso buscar por outro step nomeado “Dimension Update”, com as mesmas configurações utilizadas no estudo de caso 2. Ao configurá-lo e conectá-lo, a workflow se apresenta conforme ilustrado na Figura 16.

Figura 16 – Workflow após as configurações e ligações



Fonte: Do autor.

Após executada a workflow, todos os dados foram transferidos para uma nova tabela sem perdas ou falhas.

A Figura 17 ilustra como ficou a tabela do banco de dados. Conforme pode-se observar, foram armazenados um id e uma versão para fins de controle de integridade e o nomeou-se cada coluna conforme o cabeçalho do arquivo.

Figura 17 – Base de dados após a execução da transformação

	field_id bigserial	version integer	date_from timestamp without time zone	date_to timestamp without time zone	srad text	DATE text	tmax text	tmin text	rain text
1	0	1							
2	1	1	1900-01-01 00:00:00	2016-10-23 23:43:07.059	11.9	83200	24.0	16.8	0.2
3	2	1	1900-01-01 00:00:00	2016-10-23 23:43:07.059	14.1	83201	24.9	18.3	0.6
4	3	1	1900-01-01 00:00:00	2016-10-23 23:43:07.059	12.7	83202	17.2	17.0	2.2
5	4	1	1900-01-01 00:00:00	2016-10-23 23:43:07.059	15.1	83203	20.4	16.3	1.6
6	5	1	1900-01-01 00:00:00	2016-10-23 23:43:07.059	16.2	83204	22.4	14.0	0.7
7	6	1	1900-01-01 00:00:00	2016-10-23 23:43:07.059	16.0	83205	23.5	12.6	0.9
8	7	1	1900-01-01 00:00:00	2016-10-23 23:43:07.059	16.8	83206	24.1	16.3	1.8
9	8	1	1900-01-01 00:00:00	2016-10-23 23:43:07.059	15.3	83207	25.4	15.1	1.3
10	9	1	1900-01-01 00:00:00	2016-10-23 23:43:07.059	17.4	83208	25.1	14.7	2.4
11	10	1	1900-01-01 00:00:00	2016-10-23 23:43:07.059	18.2	83209	25.3	13.3	2.8

Fonte: Do autor.

4 CONSIDERAÇÕES FINAIS

A ferramenta PDI demonstrou-se eficiente nos tratamentos dos estudos de casos. Com seus recursos foi possível realizar o processo de ETL de uma maneira fácil e rápida, atendendo aos pré-requisitos levantados em cada estudo, como tratamento da integridade referencial, carga e versionamento de dados, controlando a atualização ou inserção na base.

Para o desenvolvimento das integrações fez-se necessário o uso de alguns dos muitos recursos que a ferramenta apresenta, tornando inicialmente, um tanto

quanto, complexo a compreensão e configuração de cada um, no entanto, foi necessário estudar quais os recursos eram mais adequados no tratamento dos estudos de casos.

Os resultados alcançados com os estudos realizados foram de acordo com os objetivos delimitados, exceto pelo fato da ferramenta não ter reconhecido um tipo de extensão de arquivo de texto. Como uma tarefa futura pretende-se explorar a ferramenta a ponto de utilizar linguagens de programação, tratamentos de máscara de dados e automatização das transformações, se aprofundando na criação de Jobs, que não foram utilizados neste trabalho.

ABSTRACT

The information and communication society is immersed in a high data movement that needs to be treated with more attention. Given this, it can be noted that access to information grows on a large scale, thus requiring companies to seek technological resources to ensure speed and performance in their services. This article presents the description of three case studies implemented with the objective of carrying out an ETL process of data from different extensions and from the storage database to a final database for each study. The scenario of these studies is based on problems, in which there is manipulation and treatment of data formats. For its realization, the Pentaho Community Data Integration tool was used, with the main purpose of analyzing how the tool adapts.

Keywords: integration of data; ETL tools; mass data migration.

REFERÊNCIAS

BARBOSA, Alvaro C. P. Middleware para Integração de Dados Heterogêneos Baseado em Composição de Frameworks Disponivel em: <ftp://ftp.inf.puc-rio.br/pub/docs/theses/01_PhD_barbosa.pdf> Acesso em: 12 out. 2016.

CASTERS M., BOUMAN R., DONGEN J. V.,Pentaho Kettle Solutions – Building Open Source ETL Solutions with Pentaho Data Integration 1ª ed. Indianápolis: Wiley, 2010.

DBBEST TECHNOLOGIES - Extract-Transform-Load (ETL) Technologies – Part 1.Disponivel em: https://www.dbbest.com/blog/extract-transform-load-etl-technologies-part-1/. Acesso em: 12 set. 2016.

DICAS PARA COMPUTADOR, Forrester Research - solução Informática Powercenter com a tecnologia de ETL. Disponível em: <http://www.dicasparacomputador.com/forrester-research-solucao-informatica-powercenter-como-tecnologia-etl#ixzz4BSWOE7PD>. Acesso em: 12 set. 2016.

INVISUAL. Disponível em: <http://www.invisual.com.br/integracao-etl.php>. Acesso em: 12 set. 2016.

KAKUGAWA , Fernando Ryoji. Integração de Bancos de Dados Heterogêneos Utilizando Grades Computacionais. Disponível em: <http://www.teses.usp.br/teses/disponiveis/3/3141/tde-07012011-145400/publico/Dissertacao_Fernando_Ryoji_Kakugawa.pdf >. Acesso em: 06 set 2016.

KIMBALL, R.;CASERTA, J. *The Data Warehouse ETL Toolkit*: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. EUA: Wiley Publishing, Inc. 2004.

NETO, Trajano C. M. Avaliação das ferramentas etl open-source talend e kettle para projetos de data warehouse em empresas de pequeno porte Disponível em: <http://www.ambientelivre.com.br/downloads/doc_download/87-tcc-ferramentas-de-etl-open-source-talend-e-kettle.html>. Acesso em: 14 set. 2016.

PENTAHO, Kettle. Project. Disponivel em: <http://community.pentaho.com/projects/data-integration/>. Acesso em: 12 set. 2016.

TANAKA, Asterio. Tópicos Avançados de Banco de Dados (Business Intelligence) - Integração de Dados e ETL. Disponível em: <http://www.uniriotec.br/~tanaka/SAIN/03-ETL-2015.1.pdf>. Acesso em: 31 out. 2016.