Estudo de Caso de Migração entre Banco de Dados Heterogêneos Utilizando Pentaho

Talita Pereira Rodrigues[1], Ricardo Bortolo Vieira[1]

[1]Faculdade Cidade Verde (FCV)
Maringá – PR – Brasil

talita_pr_@hotmail.com, professor.ricardovieira@gmail.com

Abstract. *This paper presents the process of data migration between heterogeneous databases through a case study. It was designed as tool of migration the Pentaho, using the ETL process to perform data transfer.*

Keywords: *Pentaho, Oracle, Database, SQL Server, ETL, Data Migration.*

Resumo. *Este artigo apresenta o processo de migração de dados entre banco de dados heterogêneos através de um estudo de caso. Foi definida como ferramenta de migração o Pentaho, utilizando o processo ETL para realizar a transferência dos dados.*

Palavras-chave: *Pentaho, Oracle, Banco de Dados, SQL Server, ETL, Migração de Dados.*

1. Introdução

As organizações estão sempre em busca de decisões corretas para alavancar o negócio e otimizar a metodologia de trabalho, constantemente em busca do melhor modelo é preciso ter opções para estas mudanças. Por essa razão, a gestão dos dados tem se tornado uma atividade de extrema importância dentro das organizações. Segundo Calazan e Raslan sistemas para banco de dados simplificam a tarefa de manter e recuperar uma grande quantidade de dados.

Para realizar uma migração entre banco de dados heterogêneos, é necessário estudar as documentações dos Sistemas Gerenciadores de Banco de Dados (SGBDs) e identificar as diferenças de sintaxe e funcionalidades, realizar testes de migração antes de realizar a mudança definitiva, com a finalidade de planejar o tempo gasto na janela de manutenção e espaço de armazenamento que pode mudar entre diferentes SGBDs.

Conforme pesquisas realizadas pela Bloor Research (2011), o cenário de migração dos dados está sendo aprimorado ao longo dos anos e indica que 62% dos projetos são entregues dentro do prazo e cumprindo o orçamento proposto. Esta evolução somente tornou-se possível com o advento das novas tecnologias e estudos dos processo

detalhados neste trabalho. A consequência é a redução dos riscos envolvidos.

Existem inúmeras ferramentas de mercado para auxiliar no processo de migração. A escolha da ferramenta para realizar a migração é de fundamental importância, veremos através de um estudo de caso, o uso do processo de Extração, Transformação e Carga (ETL) com a ferramenta *open-source Pentaho Comunity Data Integration*.

Este trabalho apresenta um breve referencial teórico sobre, linguagem SQL, Sistemas Gerenciadores de Banco de Dados abordando os dois alvos deste estudo SQLServer e Oracle, migração de dados, processo ETL e a ferramenta *Pentaho Comunity Data Integration*.

2. Fundamentação Teórica

2.1 SQL

SQL (*Structure Query Language*) é uma linguagem de definição e de manipulação de dados relacionais, desenvolvida nos laboratórios da IBM nos anos 70 e hoje padronizadas pelos comitês ISO/Ansi (GUIMARÃES, 2003).

A linguagem SQL foi desenvolvida para atender a todos os banco de dados relacionais e possibilitar que usuários possam conectar-se a qualquer banco usando a mesma base de aprendizagem.

A linguagem SQL é dividida nos seguintes componente:

Data Definition Language (DDL): permite a criação dos componentes do banco de dados, como tabelas, campos etc.

Principais comandos DDL:

- CREATE TABLE
- ALTER TABLE
- DROP TABLE

Data Manipulation Language (DML): permite a manipulação dos dados armazenados no banco de dados.

Comandos DML:

- INSERT

- DELETE
- UPDATE
- SELECT

Com o crescimento dos dados os fornecedores passaram a incorporar comandos procedurais (funções, procedimentos entre outros), estas linguagens se tornaram específicas de cada banco de dados.

2.2 Sistemas Gerenciadores de Banco de Dados

2.2.1 Oracle

Oracle Corporation é uma empresa multinacional de tecnologia da informação especializada no desenvolvimento e comercialização de software e hardware para banco de dados, companhia fundada em 1977 por Larry Ellison, Bob Miner e Ed Oates quando viram a grande deficiência do mercado e o enorme potencial no negócio de banco de dados relacional. Está visão de mercado os tornou a maior empresa de software empresarial do mundo.

Com a evolução do SGBD desenvolvido pela companhia, em 1991 foi lançada a primeira versão da linguagem PL/SQL, *Procedural Language extention to SQL*, a partir de então era permitido executar a manipulação dos dados com uma maior complexidade, utilizando a funções e procedures. Está linguagem é sobretudo uma extensão do SQL nativo para o banco de dados Oracle. Desde então a empresa está aprimorando continuamente o PL/SQL para atender as necessidades de mercado.

Na Tabela 1 é apresentado um exemplo da criação de uma tabela no banco de dados Oracle com as suas definições para cada tipo de dado.

Tabela 1. Exemplo criação tabela em Oracle. (Fonte: A autora)

Oracle data type		
CREATE TABLE Tabela_01		
(Coluna_data	date	not null,
Coluna_inteiro	number	null,

```
Coluna_texto      long           null,
Coluna_varchar varchar2(10)      null)
```

2.2.2 Sql Server

O SQL Server é um banco de dados da empresa Microsoft, foi desenvolvido em parceria com a Sybase em 1988, no início era adicional do Windows NT, mais tarde foi aperfeiçoado e passou a ser vendido separadamente. A parceria entre a Sybase e Microsoft durou até 1994.

Conforme *Ranking* da DB-Engines, hoje o SQL Server é o terceiro banco mais utilizados entre as empresas.

Rank			DBMS	Database Model	Score		
Mar 2017	Feb 2017	Mar 2016			Mar 2017	Feb 2017	Mar 2016
1.	1.	1.	Oracle	Relational DBMS	1399.50	-4.33	-72.51
2.	2.	2.	MySQL	Relational DBMS	1376.07	-4.23	+28.36
3.	3.	3.	Microsoft SQL Server	Relational DBMS	1207.49	+4.04	+71.00
4.	4.	↑ 5.	PostgreSQL	Relational DBMS	357.64	+3.96	+58.01
5.	5.	↓ 4.	MongoDB	Document store	326.93	-8.57	+21.60
6.	6.	6.	DB2	Relational DBMS	184.91	-2.99	-3.02
7.	↑ 8.	7.	Microsoft Access	Relational DBMS	132.94	-0.45	-2.09
8.	↓ 7.	8.	Cassandra	Wide column store	129.19	-5.19	-1.14
9.	9.	↑ 10.	SQLite	Relational DBMS	116.19	+0.88	+10.42
10.	10.	↓ 9.	Redis	Key-value store	113.01	-1.03	+6.79

Figura 1. Raking Banco de Dados. (Fonte: DB-Engines, 2017)

A linguagem lançada para este SGBD é a Transact-SQL, também conhecida por T-SQL, é uma extensão ao padrão SQL-92 que trouxe melhorias de diversos recursos, como a manipulação de dados de forma eficiente, segura e exclusiva do SQL Server.

Na Tabela 2 é exibido o comando DML para criação de uma tabela em SQL Server, bem como a definição de seus tipos para cada campo.

Tabela 2. Exemplo criação tabela em SQLServer. (Fonte: A autora)

SQL Server data type

```
CREATE TABLE Tabela_01
(Coluna_data              datetime   not null,
Coluna_inteiro    int       null,
Coluna_texto      text      null,
Coluna_varchar            varchar(10)       null)
```

2.3 Migração de Dados

A migração entre bancos de dados heterogêneos desde o início da gestão da informação é uma questão com bastante relevância, visto sua importância e complexidade de execução.

Em um projeto de implantação ou desenvolvimento de softwares, é comum existir a necessidade de migrar dados de sistemas legados para o novo sistema ou até mesmo por necessidade de mudança de fornecedores do sistema gerenciador de banco de dados. No ponto de vista do projeto a migração de dados, por si só, não agrega qualquer valor ao negócio e talvez por isso não é dada a devida atenção. Contudo pode proporcionar o rápido acesso a informação quando necessária uma tomada de decisão estratégica. Segundo Sarkis (2001) normalmente, as aplicações existentes nestes sistemas legados antigos, foram geradas a partir das necessidades em resultados operacionais imediatos da corporação, estando todos os dados armazenados de forma detalhada não tendo desta forma uma documentação histórica capaz de atender uma análise de visão à longo prazo dos negócios da empresa e nem a possibilidade de identificar padrões e tendências, que são feitos de maneira mais eficiente, quando os dados estão sumarizados.

De acordo com os dados de pesquisa realizada pela Bloor Research (2007) com a participação de 700 empresas da Forbes Global 2000, 84% dos projetos de migração de dados, falham ou excedem o tempo/orçamento previstos. Está pesquisa reforça a atenção que deve ser dada, no processo de migração para que todo esforço transcorra no cronograma previsto.

2.4. ETL

ETL, vem do inglês *Extract Transform Load*, ou seja, Extração Transformação Carga. O

ETL permite lidar com todo o segmento de extração de dados independente da fonte, transformando estes dados conforme estabelecimento da regra de negócio imposta e por fim a carga dos dados em um Data Warehouse ou outro banco de dados dentro de uma organização. Sendo que para determinar um processo ETL tem-se como obrigatoriedade apenas a extração e carga dos dados a sua transformação não se faz necessária em todas as situações.

Segundo Tanaka (2015) os principais passos do processo ETL são:

- Extração: processamento necessário para conectar às fontes de dados, extrair os dados e torná-los disponíveis para os passos subsequentes;
- Transformação: quaisquer funções aplicadas sobre os dados extraídos desde a extração das fontes até o carregamento nos alvos;
- Carregamento: todo o processamento requerido para carregar os dados no sistema alvo.

A Figura 2 expõe os passos do processo ETL, de suas diferentes fontes de dados e opções de exportação das transformações.

Figura 2. Modelo de processo ETL. (Fonte: Spatialytics:GeoKettle, 2017)

No lado esquerdo podemos observar os dados de origem, na maioria dos casos, provenientes de Bancos de Dados (BD) ou então de arquivos podendo ter diferentes formatos. Os dados são obtidos destas fontes através de rotinas de extração que podem

fornecer informações iguais ou modificadas. Seguidamente, esses dados são replicados para a *Data Staging Area* (DSA) onde pode ou não ocorrer a transformação e/ou limpeza dos dados antes de serem carregados para o armazenamento de destino.

De acordo com Kimball e Caserta (2004) no *Data Warehouse* (DW), os dados normalmente utilizados estão localizados em BD multidimensionais. É importante que se tenha consciência que as alterações nos dados não afetam as fontes originais, mas sim, os dados no momento da extração para o repositório do DW. Mais ainda, que os ajustes são modelados de acordo com as necessidades do modelo de DW, atendendo assim as restrições que são necessárias para este modelo.

2.5. *Pentaho Data Integration*

De acordo com Natálio (2011) o *Pentaho Data Integration*, também conhecido por *Kettle* (*Kettle Extraction, Transport, Transformation and Loading Environment*), é o conjunto de ferramentas *open-source* responsável pelos processos de ETL da *Pentaho Business Intelligence Suite*. A sua característica principal é ser baseado em modelos (guardados sob a forma de metadados) que representam as transformações e os fluxos de dados que ocorrem num determinado processo de ETL.

Embora as ferramentas ETL sejam amplamente utilizadas em ambiente de Data Warehouse, o PDI também pode ser empregada para outros fins, como:

- Migração de dados entre aplicações ou banco de dados.
- Exportação de dados de SGBD para arquivos.
- Carregamento de grandes volumes de dados em SGBD's.
- Limpeza de dados.
- Integração entre aplicações.

O software PDI, mais especificamente o módulo Spoon, forneceu suporte ao processo ETL através de uma interface gráfica, permitindo implementar as etapas do processo de maneira prática. As principais funções do PDI são chamadas de transformação (*transformation*) e tarefa *(job*). (OLIVA, 2015).

Para criar um processo ETL, devemos conhecer as funções oferecidas pela aplicação, é destacado neste estudo as seguintes: Steps, Hops e Transformação.

- Steps: São componentes utilizados para execução das tarefas definidas pelo

usuário em uma transformação.

- Hops: Componentes de ligação entre os steps que definem a direção da execução da transformação.
- Transformação: Entidade formada pelos Steps ligados através de hops, utilizados na manipulação do fluxo de dados em um workflow.

A Figura 3 traz um exemplo da representação gráfica das etapas do processo ETL com o uso do PDI.

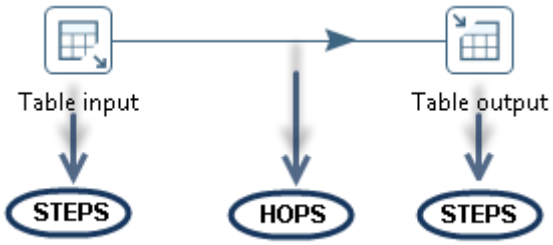

Figura 3. Exemplo etapas do processo ETL. (Fonte: A autora)

3. Estudo de Caso

Foi realizado um estudo com a proposta de migrar os dados, partindo de um banco de dados SQLServer para o Oracle utilizando a ferramenta *Pentaho Data Integration*. Neste processo será considerado somente a migração dos dados, visto que as estrutura de tabelas e campos serão criados por um software proprietário que será chamada de Ferramenta 1. Esta ferramenta possui uma funcionalidade, como o PDI para migração dos dados, que será apresentada no decorrer do estudo de caso comparando sua performance em relação ao Pentaho.

A Figura 4 representa o processo de execução do trabalho.

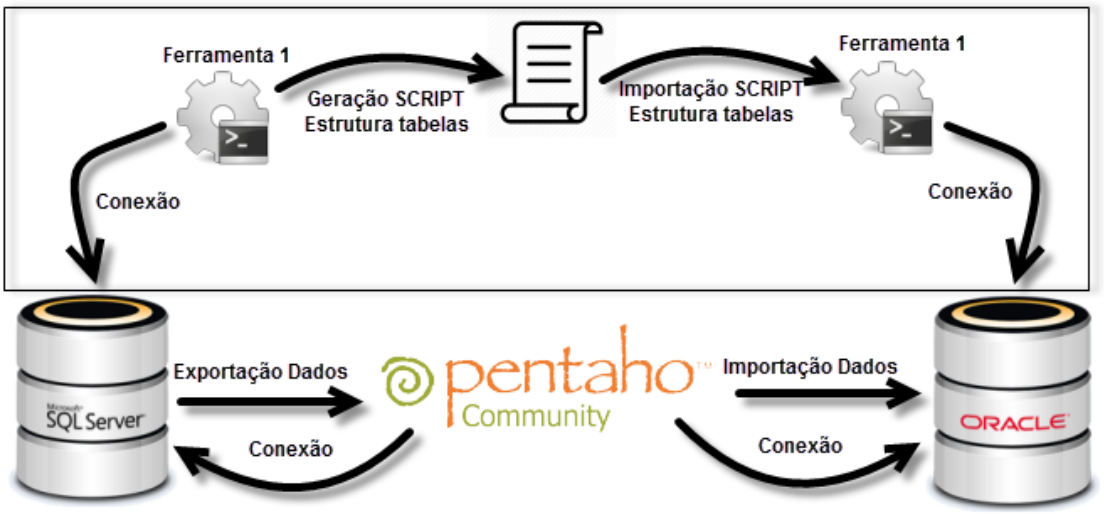

Figura 4. Fluxo de trabalho. (Fonte: A autora)

No quadrante em destaque na Figura 4 conectou-se no banco de dados SQLServer com a Ferramenta 1 para extrair o script de criação das tabelas e campos do sistema, conectou-se na mesma aplicação com o banco de dados Oracle para importar o script gerado, com esta importação o banco de dados de destino estará preparado para receber a migração dos dados.

Foi realizado um levantamento para identificar as tabelas com maior quantidade de registros, na Tabela 3 é apresentada a lista de tabelas a serem migradas, bem como a quantidade de registro e espaço ocupado em disco por estas entidades.

Tabela 3. Listagem das Tabelas usados na Migração. (Fonte: A autora)

Table Name	# Records	Data (KB)
SAM_GUIA_EVENTOS	16.067.985	12.775.616
ANS_TISMONITORAMENTO_GUIA_PROC	35.794.795	5.220.288
FIS_REGAUX03	15.711.357	4.687.416
SAM_GUIA_EVENTOS_GLOSA	11.741.812	4.277.016
FIS_REGAUX04	14.068.359	3.041.808
ANS_TISMONITORAMENTO_GUIA	5.167.290	2.760.776
SAM_GUIA	2.918.048	2.440.776
FIS_REGAUXPRODUTO	40.455.744	2.242.928

FIS_REGAUX06	6.956.883	1.749.720
SFN_CONTAB_LANC	9.031.717	1.331.344
SAM_GUIA_PROCEDIMENTO	12.949.047	1.290.312
TMP_PRECOS	656.803	1.287.008
SAM_PRONTUARIO	10.636.212	1.201.664
SFN_FATURA	2.198.722	1.183.624
SFN_FATURA_LANC	8.153.636	1.088.528
SFN_FATURA_LANC_MOD	2.827.319	1.076.984
SFN_ROTINAARQUIVO_DOC	3.275.213	1.073.592
SFN_DOCUMENTO	1.660.025	1.023.816
SAM_GUIA_EVENTOS_NEGACAO	2.493.978	789.344
SAM_PEG_OCORRENCIA	1.882.689	607.248
SAM_AUTORIZ_EVENTOGERADO	1.182.250	586.536

Para migração destas tabelas foi realizado o mapeamento via Pentaho. Inicialmente foi estabelecida uma nova transformação e as conexões para os dois banco de dados. A partir do levantamento de tabelas desenvolvidos os processos ETL na interface gráfica do PDI o Spoon.

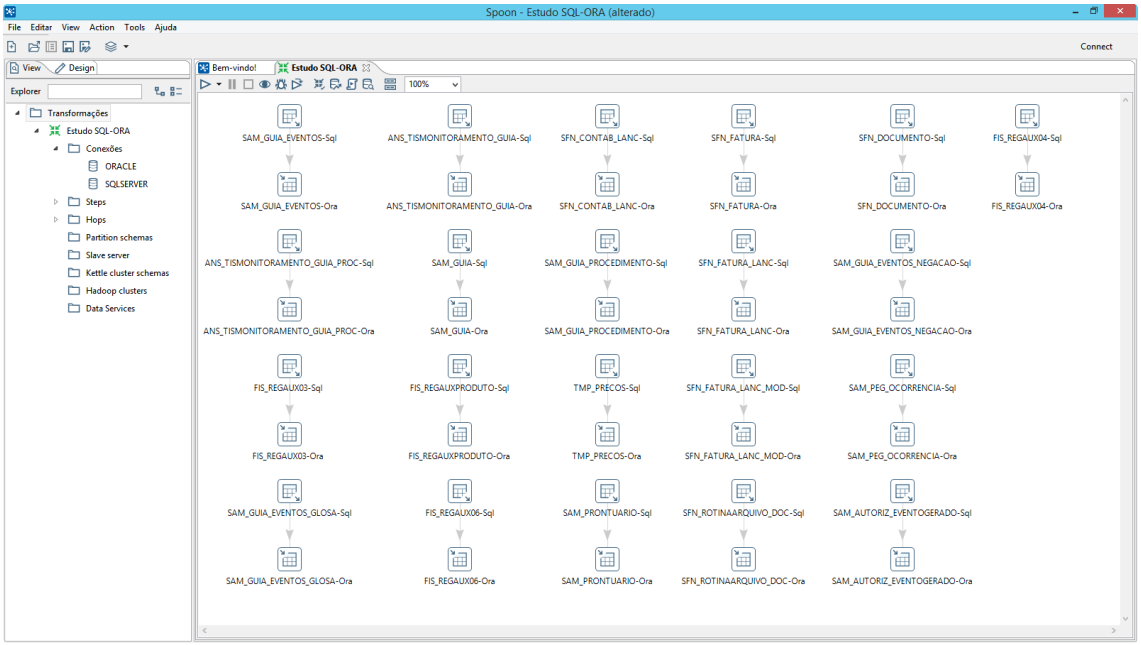

Figura 5. Transformação. (Fonte: A autora)

Como pode-se observar na Figura 5 do lado esquerdo as conexões criadas, e no lado direito o mapeamento de todas as tabelas que serão migradas. Para construção das mudanças foram utilizados os componentes *Table input* para conexão no banco SQLServer e *Table output* conectado no Oracle não será necessário realizar nenhuma transformação dos dados na migração. Portanto na extração dos dados executou-se o comando SELECT * FROM ***TABELA*** para coletar os registros, e na inserção no banco de dados de destino parametrizado o *commit* a cada cem mil registros.

O processo de migração ocorrerá de forma única sem concorrência de hardware durante a transferência dos dados, com isso poderá ser observado a capacidade de movimentação dos dados de cada uma das tabelas. Não será levado em consideração o esquema de banco de dados nesta migração, visto que o objetivo principal é migrar as tabelas com maior quantidade de registros.

4. Análise dos Dados

Para elucidar o ganho no uso do Pentaho na migração dos dados, é preciso evidenciou-se a migração das mesmas tabelas utilizando a Ferramenta 1 que foi citada anteriormente, na Tabela 4 compara-se do tempo gasto em cada software.

Tabela 4. Relação Tabelas x Tempo. (Fonte: A autora)

Table Name	PDI	Ferramenta 1
SAM_GUIA_EVENTOS	00:42:38	04:50:30
ANS_TISMONITORAMENTO_GUIA_PROC	00:17:15	10:02:25
FIS_REGAUX03	00:17:24	04:40:31
SAM_GUIA_EVENTOS_GLOSA	00:30:21	03:29:17
FIS_REGAUX04	00:18:22	04:13:59
ANS_TISMONITORAMENTO_GUIA	00:23:45	01:45:08
SAM_GUIA	00:27:26	01:12:05
FIS_REGAUXPRODUTO	00:13:16	11:33:56
FIS_REGAUX06	00:23:54	02:28:59
SFN_CONTAB_LANC	00:08:21	02:53:48
SAM_GUIA_PROCEDIMENTO	00:06:51	04:02:57
TMP_PRECOS	00:06:19	00:18:39
SAM_PRONTUARIO	00:21:39	03:01:42
SFN_FATURA	00:14:15	01:01:56
SFN_FATURA_LANC	00:13:44	02:28:30
SFN_FATURA_LANC_MOD	00:04:47	01:10:16
SFN_ROTINAARQUIVO_DOC	00:03:19	01:24:43
SFN_DOCUMENTO	00:05:15	00:46:48
SAM_GUIA_EVENTOS_NEGACAO	00:04:59	01:05:38
SAM_PEG_OCORRENCIA	00:03:22	00:52:51
SAM_AUTORIZ_EVENTOGERADO	00:03:43	00:33:10

Fundamentada com os dados descritos na Tabela 4, foi obtido o tempo médio de migração de duas horas vinte e oito minutos e trinta segundos para transferência dos registros utilizando a Ferramenta 1, enquanto com o Pentaho foi alcançado o tempo médio de treze minutos e quarenta e quatro segundos.

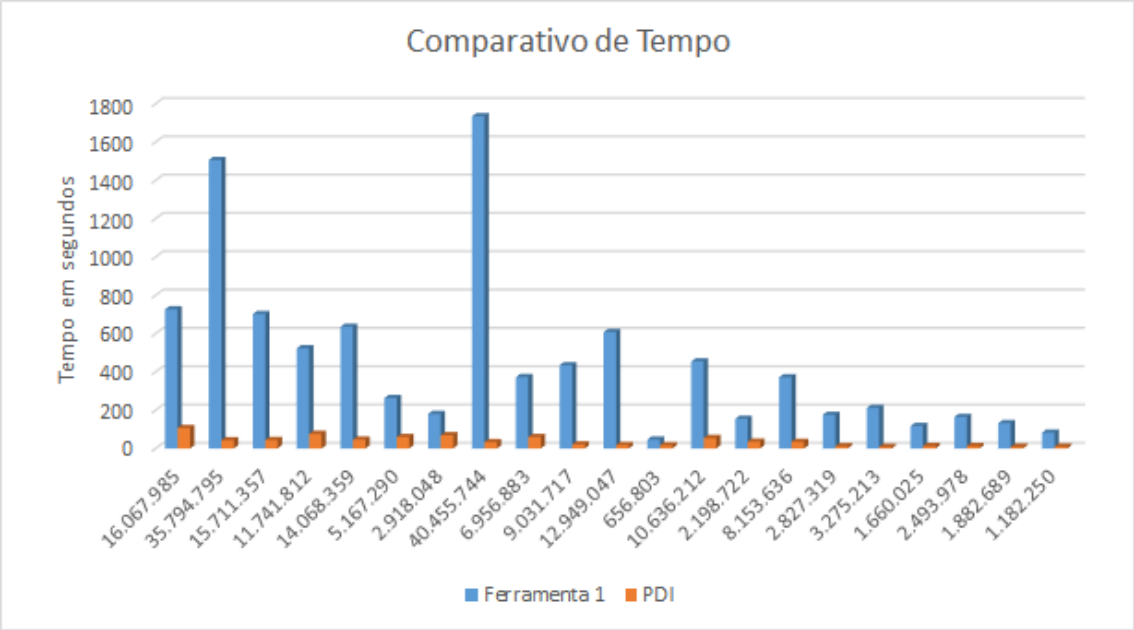

Figura 6. Comparativo de tempo migração. (Fonte: A autora)

A Figura 6 ilustra que o principal ganho no uso do PDI para migração dos dados é o tempo, porém é necessário sempre levar em consideração que para realizar a migração dos dados é preciso fazer o levantamento das tabelas, e o mapeamento das mesmas via Pentaho. No entanto com a interface gráfica fornecida Spoon é possível ter uma experiência intuitiva.

Outras vantagens no uso do *Pentaho Data Integration* é a viabilidade visto que não há custos para aquisição nem licenciamento, tornado-o uma alternativa excelente para o negócio, pois não irá exigir gastos para implementação. Por ser um aplicativo *open-source* permite a customização do *software* podendo ser personalizado caso necessidade.

5. Conclusões

Baseado nos resultados obtidos a ferramenta PDI mostrou-se extremamente eficiente no estudo de caso proposto, para construção das movimentações dos dados optou-se em utilizar alguns dos muitos recursos oferecidos pela ferramenta, tornando por vezes complexo o entendimento e configurações para atender as premissas da atividade.

É notável que uma série de fatores podem interferir na performance do Pentaho, fatores estes como, tráfego de rede, capacidade de processamento da máquina onde está sendo executado a migração, bem como dos servidores de banco de dados de origem e

destino dos dados.

Dentro do contexto estudado, seria ideal se houvesse uma metodologia, onde pudesse atingir todas as fases que compõem a atividade de migração dos dados, garantindo assim a qualidade do processo como um todo, pois apenas a migração dos dados não garante a integridade dos mesmo, fica a cargo de teste posteriores para realizar as validações necessárias.

Sendo assim, é possível compreender que a migração de dados não é, por si só, uma tecnologia, mas sim uma tarefa especializada que precisa ser suportada por uma variedade de ferramentas e técnicas, devendo ser amplamente planejada para se obter sucesso.

6. Trabalhos Futuros

Como proposta para trabalho futuro, pode-se analisar aplicação deste estudo com a importação das tabelas paralelo migrando mais de uma tabela ao mesmo tempo, podendo ter execução simultânea de múltiplos processos em diferentes fluxos de dados no mesmo *job.*

Pode-se ainda explorar a funcionalidade *Slave Server* disponível na ferramenta Pentaho onde permite o processamento em máquinas distribuídas, aumentando a escalabilidade podendo eventualmente distribuir a carga de dados entre vários servidores.

7. Referências

BLOOR Research, 2007. Data Migration in the Global 2000. Relatório Técnico. United Kigdom. Disponível em <http://www.bloorresearch.com/research/white-paper/data-migration/>. Acesso em: 12 mar. 2017.

BLOOR Research, 2011. Data Migration. Relatório Técnico. United Kigdom. Disponível em <http://www.bloorresearch.com/research/white-paper/data-migration-white-paper/>. Acesso em: 11 abr. 2017.

CALAZANS, Angélica T. S.; RASLAN Daniela A. Data Warehouse: conceitos e aplicações. Brasília: Gestão de TI, Universitas, 2014. Disponível em: <https://www.publicacoes.uniceub.br/gti/article/view/2612/2400>. Acesso em: 08 abr. 2017.

DB-ENGINES, 2017. Disponível em: <http://db-engines.com/en/ranking> Acesso em: 20 mar. 2017.

GUIMARÃES, Célio Cardoso, 1942. Fundamentos de bancos de dados: modelagem,

projeto e linguagem SQL. Campinas, SP: Editora da Unicamp, 2003.

JACOBS, Diogo Rafael; CARVALHO, J. V. DDL, Lidando com as diferenças das instruções SQL nos diferentes SGBD's. Novo Hamburgo: Centro Universitário Feevale. Disponível em: < http://www.sirc.unifra.br/artigos2006/SIRC-Artigo25.pdf >. Acesso em: 01 ago. 2016.

KIMBALL, R e CASERTA, Joe. The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. John Wiley & Sons, 2004

NATALIO, João T. R. Mendes. Desenvolvimento de uma Framework de Business Performance Management. Portugal: Departamento de Informática, Universidade de Lisboa, 2011. Disponível em: <http://repositorio.ul.pt/bitstream/10451/9151/1/ulfc104470_tm_Joao_Natalio.pdf> Acesso em: 17 mar. 2017

OLIVA, Samuel Zanferdini. Ambiente de Data Warehousing para integração de dados de saúde pública em âmbito de gestão regional. Ribeirão Preto: Bioengenharia, Universidade de São Paulo, 2015. Disponível em: <http://www.teses.usp.br/teses/disponiveis/82/82131/tde-28032016-144310/>. Acesso em: 10 mar. 2017.

OLIVEIRA, Celso Henrique Pedroso. SQL: Curso Prático. São Paulo, SP: Editora Novatec, 2002.

SARKIS, Laura C. Data Warehouse: O Processo de Migração de Dados. Florianópolis: Ciência da Computação, Universidade Federal de Santa Catarina, 2001. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/80047/227423.pdf?sequence=1>. Acesso em: 08 abr. 2017

SPATIALYTICS: GeoKettle, 2017. Disponível em: <http://www.spatialytics.org/projects/geokettle>. Acesso em 14 mar. 2017.

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: 08 mar. 2017.

TEOREY, Tobey J. et al. Projeto e modelagem de banco de dados. Tradução de Daniel Vieira. Rio de Janeiro: Elsevier, 2014.