FIAP ON

BUSINESS INTELLIGENCE (BI)

ETL
- ADQUIRINDO DADOS PARA A TOMADA DE DECISÃO

FERNANDO BARBOSA LIMA

4

LISTA DE FIGURAS

SUMÁRIO

4 ETL – ADQUIRINDO DADOS PARA A TOMADA DE DECISÃO

4.1 Relembrando conceitos

Aprendemos que Business Intelligence é um termo abrangente, formado por um conjunto de itens que, juntos, permitem o acesso e a análise de informações para melhorar e otimizar decisões e desempenho em uma corporação.

As tomadas de decisão são realizadas sobre informações geradas a partir de dados disponíveis, obtidos em fontes internas ou externas à corporação. Segundo Inmon (2005), um Data Warehouse, que é um conjunto de dados de apoio às decisões gerenciais, integrado, não volátil, variável em relação ao tempo e baseado em assuntos. Integrado, pois os dados são coletados a partir de uma variedade de fontes distintas e fundidos em um todo coerente.

Em uma solução de DW ou Data Mart, a Staging Area é a área de trabalho que recebe os dados das fontes internas e externas para que eles sejam trabalhados, antes de disponibilizados para a consulta dos tomadores de decisões. A utilização de Data Marts, em função do tempo de implementação muito menor e o retorno de investimento mais rápido, é a proposta de construção de DW mais aceita no mercado atualmente.

Vale lembrar que Data Marts são subconjuntos de dados corporativos, geralmente focados em assuntos especiais e de valor para um departamento da corporação, unidade corporativa ou conjunto de usuários. Na metodologia de estruturação Bottom-Up, proposta por Kimball, um DM pode ser considerado um pequeno DW ou um subconjunto de um DW.

Um Data Mart é definido pelo escopo funcional que atende e não pelo seu tamanho.

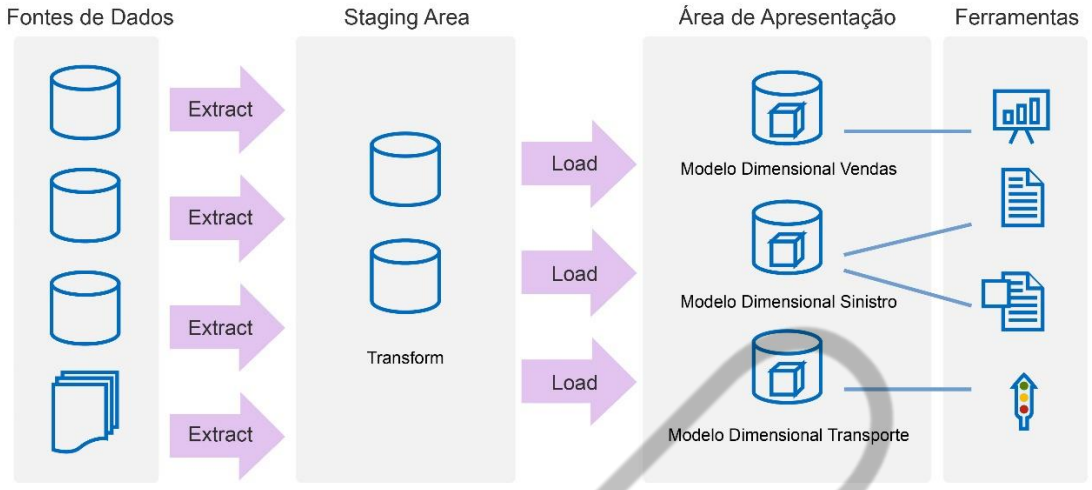

Figura 4.1 – Componentes do ETL
Fonte: Elaborado pelo autor (2018)

4.2 Extract - Transform - Load (ETL)

Extract Transform Load (Extração Transformação Carga) – ETL é o processo de extração, transformação e carga dos dados oriundos de fontes diversas, em modelos dimensionais no DW, para que os usuários finais possam realizar consultas e tomar decisões.

- **Extração**: obtenção dos dados brutos, de fontes internas e externas para a Staging Area (SA) e para o ODS (quando existente).
- **Transformação**: limpeza dos dados extraídos, aplicação de regras de negócio e validação dos dados na SA como preparação para carga.
- **Carga**: inserção dos dados transformados na SA em bases de dados do DW para serem acessados por ferramentas de consulta, geradores de relatórios, dashboards e cubos OLAP.

O sistema ETL é considerado por Kimball como a principal fundação de um Data Warehouse. Um ETL, devidamente projetado, extrai dados diversos, impõe qualidade e padrões de consistência, conforme os dados de modo que fontes distintas possam ser usadas em conjunto e, finalmente, carregam os dados em um formato pronto para análise e tomada de decisões.

Embora seja um processo do DW que não é visível para os usuários finais, segundo Kimball, o ETL é muito importante, pois consome facilmente 70% dos recursos necessários para a implementação e manutenção de um Data Warehouse típico, pois um ETL deve ser capaz de acessar diferentes bases de dados e ler diversos formatos de arquivos, utilizados por toda a organização.

Geralmente, essa não é uma tarefa simples, pois muitas fontes de dados podem ter dificuldades particulares para acesso.

A manutenção é composta por atividades auxiliares na evolução e no gerenciamento do Data Warehouse. Temos tarefas específicas para gerenciamento de jobs, performance, planos de backup, verificação de itens de segurança e compliance. Em relação à compliance, após a aprovação do ato Sarbanes-Oxley, as organizações foram obrigadas a garantir a veracidade do que elas relatam e fornecer provas de que os números relatados são precisos, completos e não foram adulterados.

Em casos de auditoria, isso implica que a gestão do ETL deve possuir cópia das fontes de dados, controles de segurança de acesso aos dados, detalhes do fluxo de extração, transformação e carga dos dados e, principalmente, algoritmos de transformação muito bem documentados.

4.3 Staging Area e o ETL

Kimball sugere a seguinte analogia: imagine um restaurante: os clientes do restaurante são os usuários finais do DW e a comida é um dado. A comida é oferecida aos clientes na sala de jantar e deve ser servida, exatamente como o cliente espera, limpa, organizada e representada de forma que cada item que compõe o pedido possa ser facilmente identificado e consumido.

Antes da comida ser servida, ela é preparada na cozinha sob a supervisão de um experiente chef. Na cozinha, os ingredientes são selecionados, limpos, fatiados, integrados, cozidos e preparados para apresentação.

A cozinha é uma área de trabalho fora dos limites para os clientes do restaurante, o acesso à cozinha, onde a comida ainda está em preparo, prejudica a melhor experiência gastronômica do cliente.

Em um restaurante renomado, se um cliente solicita informações sobre a preparação de alimentos, o chef deve sair da cozinha para explicar o processo de preparação de alimentos ao cliente na sala de jantar, um ambiente seguro e limpo, onde o cliente está confortável.

A Staging Area (AS) é a cozinha do Data Warehouse. É um ambiente acessível apenas para profissionais experientes que desempenham a função de integradores de dados. É um ambiente fora dos limites para os usuários finais, onde os dados são colocados depois que são extraídos dos sistemas de fontes, são limpos, manipulados e preparados.

Um ETL útil para os usuários finais deve obter, preparar e levar os dados da cozinha, até a área de apresentação de um Data Warehouse.

4.4 A área de apresentação

Como pudemos perceber, o propósito de um ETL é obter dados das fontes de origem, transformá-los e carregá-los nas tabelas modeladas dimensionalmente, para serem acessados diretamente pelas ferramentas de consulta, geradores de relatórios, geradores de dashboards e cubos OLAP.

Os dados na área de apresentação são o que os usuários finais realmente veem. Data Marts são componentes importantes da área de apresentação, pois cada DM é formado por um conjunto de tabelas dimensionais e fato, que juntas comportam dados de um determinado processo comercial.

4.5 O ETL na linha do tempo

O ETL tornou-se realidade nos anos 1970, quando as organizações começaram a usar vários repositórios de documentos e bancos de dados para armazenar diferentes tipos de informações comerciais. A necessidade de integrar dados espalhados por essas fontes cresceu rapidamente. O ETL tornou-se o método padrão para tirar dados de fontes diferentes e transformá-los antes de carregá-los.

Em meados dos anos 1990, os Data Warehouses ganharam espaço. Grandes armazéns de dados surgiram para fornecer acesso integrado a dados de vários sistemas, processados em mainframes, minicomputadores, computadores pessoais e planilhas.

Ao longo do tempo, o número de formatos de dados, fontes e sistemas expandiu-se tremendamente. Extrair, transformar, carregar agora é apenas um dos vários métodos que as organizações usam para coletar, importar e processar dados.

No início, antes que as ferramentas ETL existissem, a única maneira de integrar dados de fontes diferentes eram programas codificados em linguagens como COBOL, RPG e PL/SQL. Apesar de serem propensos a erros, lentos para desenvolver, pesados para processar logs e difíceis para manter, estudos apontam que aproximadamente 45% de todo o processo de ETL hoje, ainda é feito por códigos feitos manualmente ou através de geradores em diversas linguagens de programação.

Os geradores de código também apresentam limitações, já que eles trabalham com apenas um conjunto limitado de bancos de dados e, muitas vezes, são exclusivos para um deles.

Em contrapartida, a geração mais atual de ferramentas ETL inclui um mecanismo de propósito geral, que executa a transformação e um conjunto de metadados que armazenam a lógica de transformação.

Como as ferramentas ETL baseadas em um motor de execução são independentes de armazéns de dados de origem e de destino, elas são mais versáteis do que gerar códigos manualmente ou a partir de ferramentas. O Pentaho Data Integration - PDI é um ótimo exemplo de ferramenta ETL atual.

As ferramentas atuais não podem deixar de lado o Big Data e a Suíte Pentaho que fornecem ferramentas para extrair, preparar, integrar dados e realizar análises visuais. De Hadoop e Spark (NoSQL) e Bancos de Dados relacionais, o Pentaho fornece soluções para trabalhar com Big Data e para gerar importantes insights.

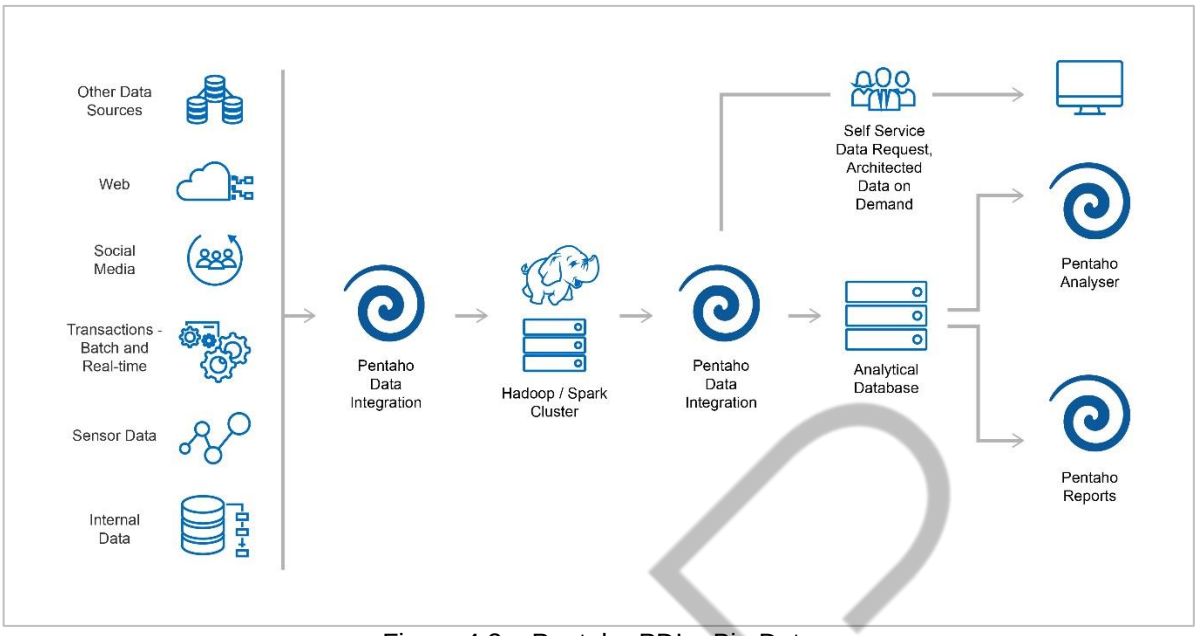

Figura 4.2 – Pentaho PDI e Big Data
Fonte: http://www.pentaho.com (2018)

4.6 Ferramentas ETL ou Código?

Como em qualquer outra situação, existem vantagens e desvantagens em ambas as abordagens. Não podemos desconsiderar que o percentual comentado de soluções ETL codificadas à mão contabiliza os projetos herdados do período em que não existiam ou ainda estavam em desenvolvimento as principais ferramentas de ETL atuais. Abaixo apresentamos as principais vantagens de cada abordagem.

4.6.1 ETLs baseados em ferramentas

Desenvolvimento do ETL se torna mais simples, mais rápido e mais barato.

Pessoas técnicas com habilidades comerciais, mas sem grandes capacidades de programação, conseguem utilizar as ferramentas ETL de forma eficaz.

As ferramentas de ETL mais modernas são autoexplicativas e possuem mecanismos de documentação simplificados. Diferencial significativo para manutenções, principalmente no mercado de TI atual, com equipes sofrendo alta rotatividade.

Utilizando ferramentas, a manutenção de um ETL é mais fácil de ser feita, comparando-se à manutenção de código. Ferramentas utilizam interfaces visuais que

criam ETLs através design gráficos, facilitando a interpretação, a documentação e o entendimento de quem fará a manutenção.

Um ETL pode ser reaproveitado graficamente dentro de outros ETLs ou utilizado como um template.

Processos complexos, com vários fluxos e dependências, portanto, com várias atividades de ETL, são facilmente orquestrados criando-se jobs de forma gráfica.

Ferramentas ETL atuais possuem repositórios de metadados integrados que podem sincronizar metadados de sistemas de origem, com metadados de bancos de dados de destino e de outras ferramentas de BI.

A maioria das ferramentas ETL gera automaticamente metadados em cada etapa do processo e impõe uma forma de trabalho consistente, baseada em metadados, que todos os desenvolvedores do ETL devem seguir.

A maioria das ferramentas ETL possui capacidades internas que ajudam a documentação, aportam facilidade de criação e de mudanças de gerenciamento, lidam com dependências e com tratamento e logs de erros.

As ferramentas ETL possuem diversos conectores para as mais diversas fontes de dados e destinos de carga e são capazes de executar todos os tipos de transformações de dados complexas.

As ferramentas ETL geralmente oferecem funcionalidades de criptografia e compressão de dados.

A maioria das ferramentas ETL oferece um bom desempenho, mesmo para volumes de dados muito grandes.

Ferramentas de ETL disponibilizam meios para trabalhar a qualidade dos dados através de componentes que são capazes de executar algoritmos complexos.

Implementam recursos de auditoria e tracking, facilitando o entendimento sobre a fonte de onde o dado foi extraído, que transformações ele sofreu e para onde o dado foi carregado.

Ferramentas tornam a segurança mais modular, aplicando regras por etapas do processo papéis (extração, transformação, cargas, agendamento, entre outros).

4.6.2 ETLs baseados em programas

Programas possuem total controle sobre o processo, não existe qualquer tipo de componente caixa preta ou desconhecimento de como uma determinada parte do ETL funciona. Todo o processo é conhecido, podendo ser alterado, evoluído e customizado da forma que o programador desejar, desde que respeitando os limites da linguagem de programação adotada.

A escolha da linguagem de programação é feita de acordo com a governança de TI da organização, que usualmente aproveita a plataforma de hardware/software vigente e o maior conhecimento dos profissionais internos, para escolha.

A codificação interna dispensa contratos de suporte e manutenção com fabricantes de ferramentas. Também não possui o risco de um fabricante descontinuar ou alterar radicalmente um produto. Por outro lado, evoluções tecnológicas obtidas sem esforço em novas versões de ferramentas, não são uma realidade neste cenário.

4.7 Pentaho

4.7.1 A Suíte Pentaho

O Pentaho é uma solução open source desenvolvida em Java, composta por um conjunto de ferramentas visuais para análise de dados. Capaz de combinar a integração de dados com o processamento analítico, o Pentaho acelera significativamente o processo de obtenção, transformação, carga, predição e análise de informações.

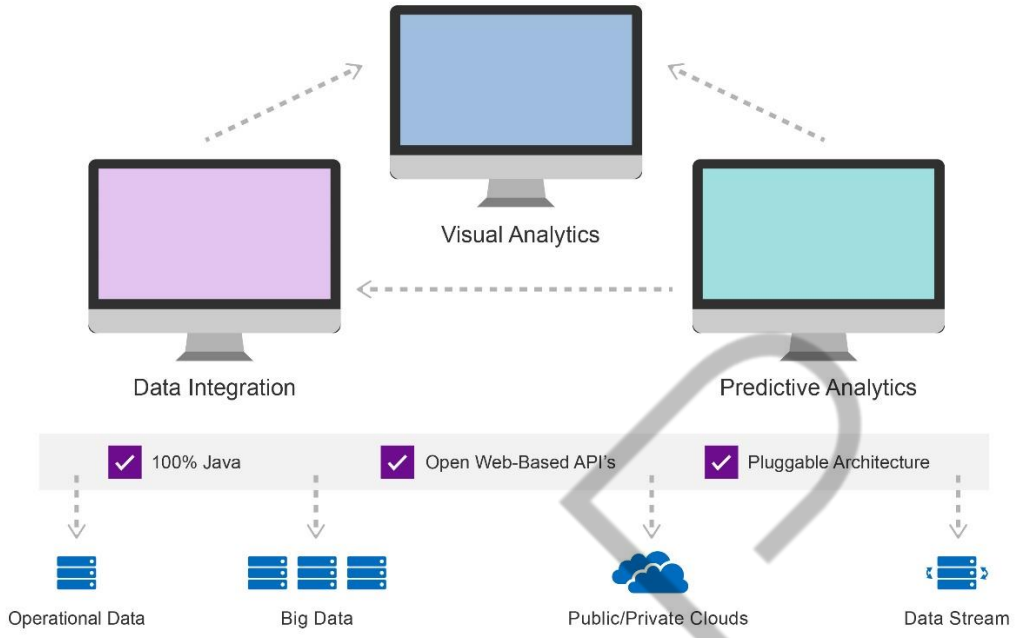

Figura 4.3 – Pentaho
Fonte: http://www.pentaho.com (2018)

O Pentaho é customizável e possui uma interface amigável que ajuda os usuários a visualizar os dados, e permite que eles tomem decisões comerciais inteligentes. O Pentaho é projetado para assegurar que cada membro de sua equipe, desenvolvedores ou usuários conhecedores do negócio, possam facilmente converter dados em valor;

A suíte é composta por um conjunto de ferramentas e funcionalidades que podem ser usadas para os seguintes propósitos:

- Integrações de dados.
- Interface of Things.
- Business Analytics.
- Big Data.
- Cloud Analytics.
- Ad Hoc Analysis.
- On-line Analytical Processing (OLAP).
- Predictive Analysis.

- Ad Hoc Reporting.
- Performance Measurements.

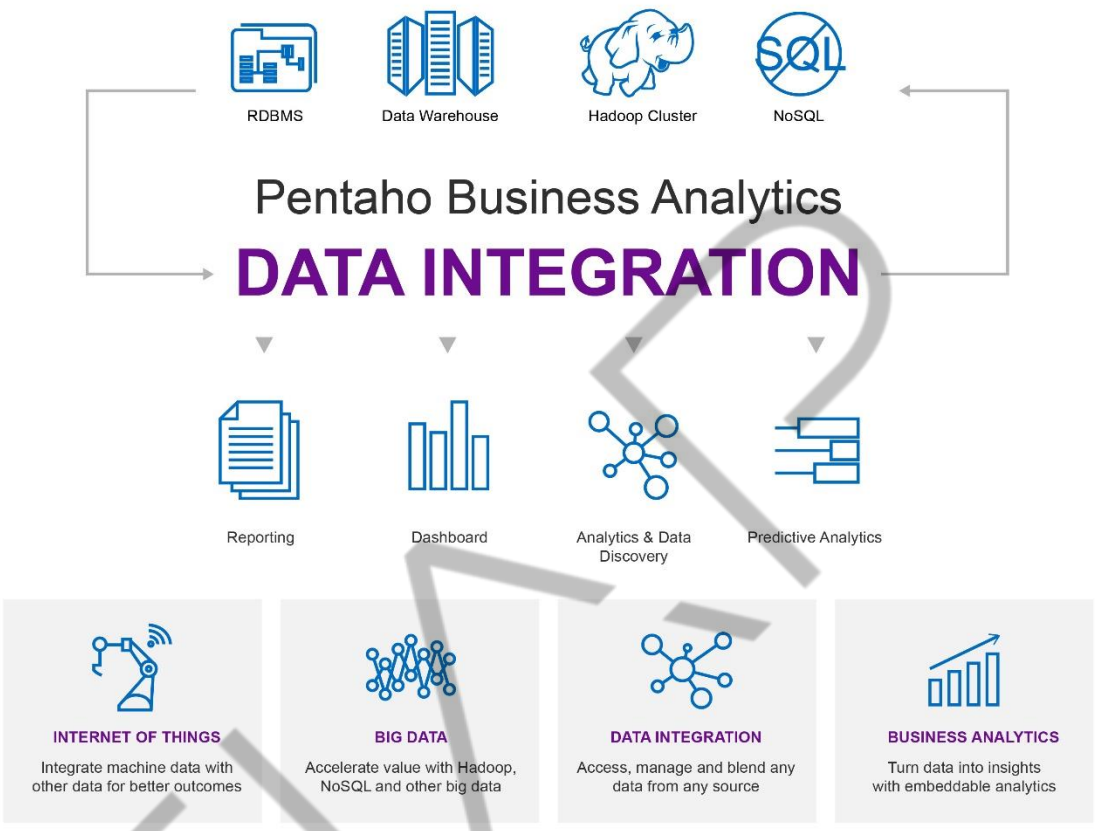

Figura 4.4 – Pentaho B.A.
Fonte: http://www.pentaho.com (2018)

Com uma série de ferramentas avançadas de análise, o Pentaho foi projetado para se encaixar bem com as plataformas móveis, permitindo que os usuários realizem análises em tablets e smartphones.

4.7.2 Produtos Pentaho

Os produtos Pentaho são usados para acessar, integrar, manipular, visualizar e analisar dados provenientes das mais diversas fontes, tais como: flat files, bancos de dados relacionais, Hadoop, bancos de dados NoSQL, bancos de dados analíticos, fluxos de mídia social e da nuvem, mesmo que não se tenha experiência de codificação. Já programadores podem usar uma extensa API disponibilizada pelo

Pentaho para personalizar relatórios, consultas, transformações e para estender a funcionalidades existentes.

Os produtos incluem componentes baseados na web e ferramentas de design.

4.7.3 Componentes baseados na Web

Os componentes baseados na Web são utilizados para compartilhar soluções de business intelligence, analises de dados, relatórios e painéis integrados.

Esses componentes incluem:

- **Pentaho User Console (PUC)**: ambiente de design para acessos de análise, relatórios iterativos e designer de dashboards. O console também oferece recursos de administração do sistema para configurar o servidor Pentaho.
- **Analyzer**: capaz de criar gráficos geográficos, gráficos de dispersão, mapas de calor e multi-gráficos. Com o analisyzer o usuário pode filtrar dados, adicionar parâmetros de consulta, configurar links de detalhamento, aplicar formatação condicional e gerar hiperlinks;
- **Interactive Reports**: interface gráfica usada para criar relatórios operacionais simples e on-demand, sem depender de programadores. O usuário de negócio pode criar e adicionar rapidamente elementos a um relatório e formatá-los da maneira que quiser.
- **Dashboard Designer**: ferramenta usada para escolher modelos de layout, temas e conteúdo para criar dashboards (painéis) visualmente atrativos que ajudem os tomadores de decisão a obter conhecimento crítico rapidamente. O usuário que constrói o dashboard pode combinar uma grande variedade de conteúdo, incluindo relatórios interativos, visualizações do analisador e conteúdo colaborativo.
- **CTools**: framework para a criação de painéis usando tecnologias da web, como JavaScript, CSS e HTML. Na elaboração, o usuário pode facilmente criar painéis dinâmicos para usuários explorar e entender grandes

quantidades de dados usando uma variedade de gráficos, tabelas e outros componentes.

- **Data Source Wizard**: define fontes de dados que contém as informações desejadas e orienta o usuário a criar relatórios de análise, através da criação dos primeiros modelos relacionais ou multidimensionais.

Data Source Model Editor: ajuda os analistas a aprimorar modelos de dados relacionais e multidimensionais graficamente. É possível arrastar os campos para os locais apropriados, misturar e combinar campos de diferentes tabelas, adicionar campos a mais ou remover um campo.

Os componentes web do Pentaho 8 podem ser executados em qualquer um dos servidores de aplicação abaixo:

- JBoss EAP 7.x com Oracle Java 8.x.
- Tomcat 8.0.x com Oracle Java 8.x.

O Pentaho foi desenvolvido para se integrar com sistemas de autenticação de segurança de terceiros:

- Active Directory.
- CAS.
- Integrated Microsoft Windows Authentication.
- LDAP.
- RDBMS.

4.7.4 Ferramentas de design

Use ferramentas de design Pentaho para desenvolver e refinar como seus valores de dados são relatados, modelados, transformados e armazenados.

Essas ferramentas incluem:

- **Pentaho Data Integration (PDI):** provê funcionalidades de Extração, Transformação e Carregamento (ETL), como esta ferramenta é o nosso foco, falaremos mais sobre ela adiante.

- **Report Designer**: usado para gerar relatórios detalhados usando praticamente qualquer fonte de dados; permite que usuários de negócios criem relatórios altamente detalhados e com alta qualidade de impressão, com base em dados adequadamente preparados.
- **Aggregation Designer**: fornece uma interface simples que permite que a criação de tabelas agregadas para melhorar o desempenho de cubos OLAP do Pentaho Analysis (Mondrian).
- **Metadata Editor**: é capaz de criar modelos de metadados que mapeiam a estrutura física do seu banco de dados para modelos de negócio lógico. Simplifica muito a criação de relatórios por usuários finais.
- **Schema Workbench**: permite a criação de modelos multidimensionais (Mondrian). Os analistas podem criar modelos Mondrian graficamente ou defini-los, usando arquivos XML estruturados manualmente.

4.7.5 Pentaho Data Integration (PDI)

O Pentaho Data Integration (PDI) é uma ferramenta visual que fornece os recursos para Extract, Transform e Load (ETL) com o intuito de eliminar a codificação e a complexidade, permitindo ao usuário com conhecimentos básicos, criar as tarefas de captura, limpeza, transformação e armazenamento de dados, usando um processo uniforme, consistente e fácil de usar.

4.7.5.1 Facilidade de uso

- Disponibiliza biblioteca rica de componentes para acessar, preparar, tratar e gravar dados de fontes diversas.
- Possui diversos tipos de conexões facilmente configuráveis, com bancos de dados relacionais, não relacionais, arquivos, serviços, entre outros.
- Possibilita a criação de ETLs através de designer gráfico, de forma intuitiva, arrastando e soltando componentes.

- Possui exibições visuais para modelagem e visualização de dados em tempo real, durante os processos de extração, transformação e carga dos dados.
- Fornece recursos de orquestração para determinar fluxos e combinar transformações, incluindo funcionalidades de teste que disparam notificações e alertas durante a execução do trabalho.
- Possui suporte robusto, sem necessidade de codificação, para distribuições Hadoop, Spark, NoSQL e bancos de dados analíticos.
- Integra modelos analíticos avançados em R, Python e Weka para operacionalizar modelos preditivos.
- Trabalha com templates de integração de dados dinâmicos que economizam tempo e reduzem riscos de erros, quando reutilizados.

CONCLUSÃO

O objetivo deste conteúdo foi abordar o essencial sobre ETL e Pentaho. Existem diversos pontos a serem explorados, como o dimensionamento de ambientes para grandes cargas, outras implementações de modificação lenta, a criação de job complexos, o consumo de serviços pelo PDI, entre outros pontos. Portanto, sugerimos a leitura dos diversos conteúdos adicionais que são abordados nos livros e URLs de referência deste capítulo.

REFERÊNCIAS

BOUMAN, Roland; VAN DONGEN, Jos. **Pentaho solutions**: Business Intelligence and Data Warehousing with Pentaho and MYSQL. Wiley, 2009.

CASTERS, Matt; BOUMAN, Roland; VAN DONGEN, Jos. **Pentaho Kettle solutions:** building open source ETL solutions with Pentaho Data Integration. John Wiley & Sons, 2010.

GOODMAN, Nicholas. **Pentaho Data Integration:** Scaling out large data volume processing in the cloud or on premise. Bayon Technologies, 2009.

GONÇALVES, Marcio. **Extração de dados para Data Warehouse**. Rio de Janeiro/RJ: Axcell Books, 2003.

HITACHI. **Pentaho Data Integration**. 2017. Disponível em: <https://help.pentaho.com/Documentation/8.0/Products/Data_Integration>. Acesso em: 10 fev. 2018.

_____. **PDI Job Tutorial.** 2017. Disponível em: <https://help.pentaho.com/Documentation/8.0/Setup/Evaluation/Tutorials/0C0/030>. Acesso em: 10 fev. 2018.

_____. **Work with Jobs**. 2017. Disponível em: <https://help.pentaho.com/Documentation/8.0/Products/Data_Integration/Data_Integration_Perspective/040>. Acesso em: 10 fev. 2018.

_____. **Pentaho Data Integration Steps.** 2017. Disponível em: <https://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+Steps>. Acesso em: 10 fev. 2018.

_____. **Schedule a Transformation or Job.** 2017. Disponível em: <https://help.pentaho.com/Documentation/8.0/Products/Data_Integration/Schedule_Perspective>. Acesso em: 10 fev. 2018.

_____. **Work with Pentaho Repositories**. 2017. Disponível em: <https://help.pentaho.com/Documentation/8.0/Products/Data_Integration/Repositori>. Acesso em: 10 fev. 2018.

_____. **Navigate the PDI Client (Spoon)**. 2017. Disponível em: <https://help.pentaho.com/Documentation/8.0/Products/Data_Integration/PDI_Client>. Acesso em: 10 fev. 2018.

_____. **Components Reference. 2017.** Disponível em: <https://help.pentaho.com/Documentation/8.0/Setup/Components_Reference>. Acesso em: 10 fev. 2018.

_____. **Basic Concepts of PDI: Transformations, Jobs and Hops.** 2017. Disponível em:

<https://help.pentaho.com/Documentation/8.0/Products/Data_Integration/Data_Integration_Perspective/010>. Acesso em: 10 fev. 2018.

_____. **Basic Concepts of PDI: Transformations, Jobs and Hops.** 2017. Disponível em: <https://help.pentaho.com/Documentation/8.0/Products/Data_Integration/Data_Integration_Perspective/010>. Acesso em: 10 fev. 2018.

INMON, William H. **Building the data warehouse**. John Wiley & Sons, 2005.

KIMBALL, Ralph; ROSS, Margy. **The data warehouse toolkit:** the complete guide to dimensional modeling. John Wiley & Sons, 2011.

PULVIRENTI, Adrián Sergio. **Pentaho data integration 4 cookbook**. Packt Publishing Ltd, 2011.

RALPH, Kimball; MARGY, Ross. **The data Warehouse ETL Toolkit**. Wiley, 2008.

SINGH, Harry S. **Data warehouse:** conceitos, tecnologias, implementação e gerenciamento. Tradução Mônica Rosemberg. Makron Books: São Paulo, 2001.

GLOSSÁRIO

Active Directory	Implementação de serviço de diretório no protocolo LDAP que armazena informações sobre objetos em rede de computadores e disponibiliza essas informações a usuários e administradores desta rede.
Business Analytics	A análise de negócios (BA) refere-se às habilidades, tecnologias, práticas para a exploração iterativa contínua e investigação do desempenho das empresas passadas para obter informações e impulsionar o planejamento de negócios. BA concentra-se no desenvolvimento de novos conhecimentos e na compreensão do desempenho do negócio com base em dados e métodos estatísticos.
CAS	Content Addressable Storage, também conhecido como armazenamento associativo ou abreviado como CAS, é um mecanismo para armazenar informações que podem ser recuperadas com base em seu conteúdo, não o seu local de armazenamento. É normalmente utilizado para o armazenamento de alta velocidade e recuperação de conteúdo fixo, como documentos armazenados para cumprimento da regulamentação do governo.
COBOL	Sigla de COmmon Business Oriented Language - Linguagem Comum Orientada para os Negócios é uma linguagem de programação orientada para o processamento de banco de dados comerciais.
CSV	Os arquivos Comma-separated values, também conhecido como CSV, são arquivos de texto de formato regulamentado pelo RFC 4180, que faz uma ordenação de bytes ou um formato de terminador de linha. É comumente usado em softwares offices, tais como o Microsoft Excel e o LibreOffice Calc.
Data Mart	Subconjuntos de dados corporativos, geralmente focados em assuntos especiais e de valor para um departamento da corporação, unidade corporativa ou conjunto de usuários. Um data mart é definido pelo escopo funcional que atende e não pelo seu tamanho. Geralmente é considerado como subconjunto de um Data Warehouse

DW	Data Warehouse é um conjunto de dados de apoio às decisões gerenciais, integrado, não volátil, variável em relação ao tempo e baseado em assuntos.
ETL	Extract Transform Load (Extração Transformação Carga) é o processo de extração, transformação e carga dos dados, oriundos de fontes diversas em modelos dimensionais no DW, para que os usuários finais possam realizar consultas e tomar decisões
Hadoop	Plataforma de software em Java de computação distribuída voltada para clusters e processamento de grandes volumes de dados, com atenção a tolerância a falhas. Foi inspirada no MapReduce e no GoogleFS (GFS).
Interactive Reports	Interface gráfica usada para criar relatórios operacionais simples e on-demand, sem depender de programadores.
JBoss	Servidor de aplicação de código fonte aberto baseado na plataforma JEE e implementado completamente na linguagem de programação Java. Em virtude disso, ele pode ser usado em qualquer Sistema Operacional que suporte a referida linguagem
Kimball	O Prof. Dr. (PhD) Ralph Kimball é um dos precursores dos conceitos de data warehouse e sistemas para análise de dados transacionais. Desde 1982 vem desenvolvendo pesquisas e conceitos que hoje são utilizados em diversas ferramentas de software para data warehouse. Ele é conhecido por suas convicções de que o data warehouse deve ser desenhado de forma compreensível e rápida. Sua metodologia, conhecida como modelagem dimensional, é frequentemente usada para permitir o compartilhamento de dimensões.
LDAP	Lightweight Directory Access Protocol, ou LDAP, é um protocolo de aplicação aberto, livre de fornecedor e padrão de indústria para acessar e manter serviços de informação de diretório distribuído sobre uma rede de Protocolo da Internet (IP).

Mondrian	Servidor de processamento analítico on-line (OLAP) de código aberto escrito em JAVA, o Mondrian responde a consultas com rapidez suficiente para permitir uma exploração interativa dos dados, mesmo que eles tenham milhões de registros, ocupando vários gigabytes. Ele traz análise multidimensional para as massas, permitindo que os usuários examinem dados de negócios através de perfuração e tabulação de informações cruzadas.
ODS	Os bancos de dados operacionais são bancos de dados normalizados, desenvolvidos em algumas soluções de BI, para atender a necessidades analíticas sobre processos específicos em uma empresa.
OLAP	É capacidade de manipular e analisar um grande volume de dados através de múltiplas perspectivas e assim monitorar os fatos e indicadores mais relevantes da organização, por meio de painéis de controle e relatórios executivos desenvolvidos para facilitar a visualização, o entendimento dos fatos e a tomada de decisões.
OLTP	Processamento de transações on-line (OLTP) descreve a forma como os dados são processados por um sistema informatizado. Sistemas OLTP armazenam seus dados de forma normalizada e geralmente, processam enormes quantidades de operações CRUD, realizadas pelo usuário final.
PDI	Pentaho Data Integration: também conhecido como Kettle, é uma ferramenta de código aberto para extração, transformação e carga (ETL) de dados
Pentaho	Software de código aberto para inteligência empresarial, desenvolvido em Java. A solução cobre as áreas de ETL (Extrac, Transform and Load), reporting, OLAP e mineração de dados (data-mining). Realiza análises de big data, trabalha nativamente com bancos de dados NoSQL e Hadoop, além de poder processar dados de forma distribuída nativamente em cluster, pode rodar sobre o Hadoop em cluster alcançando velocidades extremamente rápidas.

PL/SQL	Acrónimo para a expressão inglesa Procedural Language/Structured Query Language. É uma extensão da linguagem padrão SQL para o SGBD Oracle da Oracle Corporation. Linguagem procedural da Oracle que estende a linguagem SQL. Permite que a manipulação de dados seja incluída em unidades de programas.
Python	Linguagem de programação de alto nível, interpretada, de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte. Foi lançada por Guido van Rossum em 1991. Possui um modelo de desenvolvimento comunitário, aberto e gerenciado pela organização sem fins lucrativos Python Software Foundation.
R	Linguagem e um ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos. Criada originalmente por Ross Ihaka e por Robert Gentleman no departamento de Estatística da universidade de Auckland, Nova Zelândia, e foi desenvolvido em um esforço colaborativo de pessoas em vários locais do mundo.
RDBMS	Sistema de gerenciamento de banco de dados (SGBD) baseado no modelo relacional inventado por Edgar F. Codd, do Laboratório de Pesquisa San Jose da IBM . A maioria dos bancos de dados em uso generalizado baseia-se no modelo de banco de dados relacional.
RPG	Report Program Generator é uma linguagem de programação através da qual se especificam os campos a partir dos quais deveriam ser obtidos os dados para gerar relatórios impressos. Foi criada pela IBM em 1959 e comercializada a partir de 1961 visando facilitar o desenvolvimento de programas.
Sarbanes-Oxley	Também conhecida como SOx, a lei foi sancionada em 2002 pelo Congresso dos Estados Unidos para proteger investidores e demais stakeholders dos erros das escriturações contábeis e práticas fraudulentas. A lei surgiu como resposta a uma série de escândalos financeiros que atingiram empresas como Xerox, Enron, Tyco, WorldCom etc.

Spark	Apache Spark™ é um mecanismo rápido e geral para o processamento de dados em grande escala.
Spoon	É a transformação gráfica e o designer de trabalho associado ao conjunto Pentaho Data Integration - também conhecido como o projeto Kettle.
Staging Area	SA é um ambiente acessível apenas para profissionais experientes que desempenham a função de integradores de dados. É um ambiente fora dos limites para os usuários finais, onde os dados são colocados depois que são extraídos dos sistemas de fontes, são limpos, manipulados e preparados.
Tomcat	O software Apache Tomcat é uma implementação de código aberto do Java Servlet, JavaServer Pages, Java Expression Language e Java WebSocket. Essas linguagens são desenvolvidas no Java Community Process.
Weka	O pacote de software Weka (Waikato Environment for Knowledge Analysis) começou a ser escrito em 1993, usando Java, na Universidade de Waikato, Nova Zelândia sendo adquirido posteriormente por uma empresa no final de 2006. O Weka tem como objectivo agregar algoritmos provenientes de diferentes abordagens/paradigmas na subárea da inteligência artificial dedicada ao estudo de aprendizagem de máquina. O Weka encontra-se licenciado ao abrigo da General Public License sendo, portanto, possível estudar e alterar o código fonte do software.