Business Intelligence

Asterio K. Tanaka
http://www.uniriotec.br/~tanaka/SAIN
tanaka@uniriotec.br

Integração de Dados e ETL

Integração de Dados

- O termo “integração de dados” refere-se ao processo de combinar dados de diferentes fontes para prover uma única visão compreensível de todos os dados combinados.(*)
- É um problema antigo em sistemas de informação, em especial da área de banco de dados (anos 1980's)
- Pode ser abordado de diferentes formas, em diferentes níveis de abstração.
  - Desde a integração manual por usuários que interagem diretamente com todos os sistemas de informação relevantes
  - Até a integração física dos dados pela transferência de dados para um novo depósito de dados comum.
- Não seria um problema “se o mundo (de sistemas de informação) fosse perfeito...”

(*)Pentaho® Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration

Matt Casters, Roland Bouman, Jos van Dongen - Wiley 2010

Integração de Dados

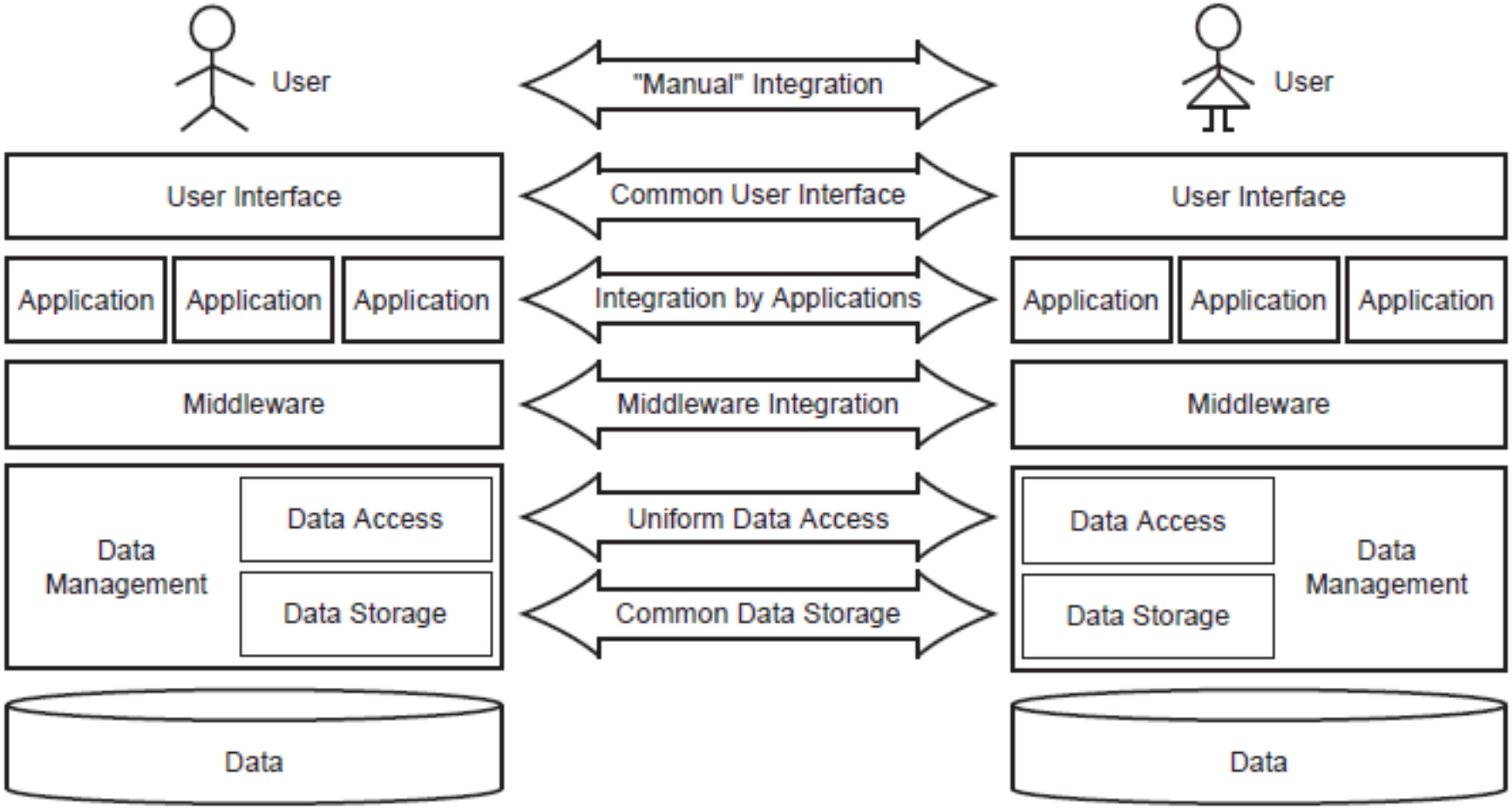


Fig. 1.1. General Integration Approaches on Different Architectural Levels

Data Integration: Problems, Approaches, and Perspectives
Patrick Ziegler and Klaus R. Dittrich - 2007

Definição de Data Warehouse

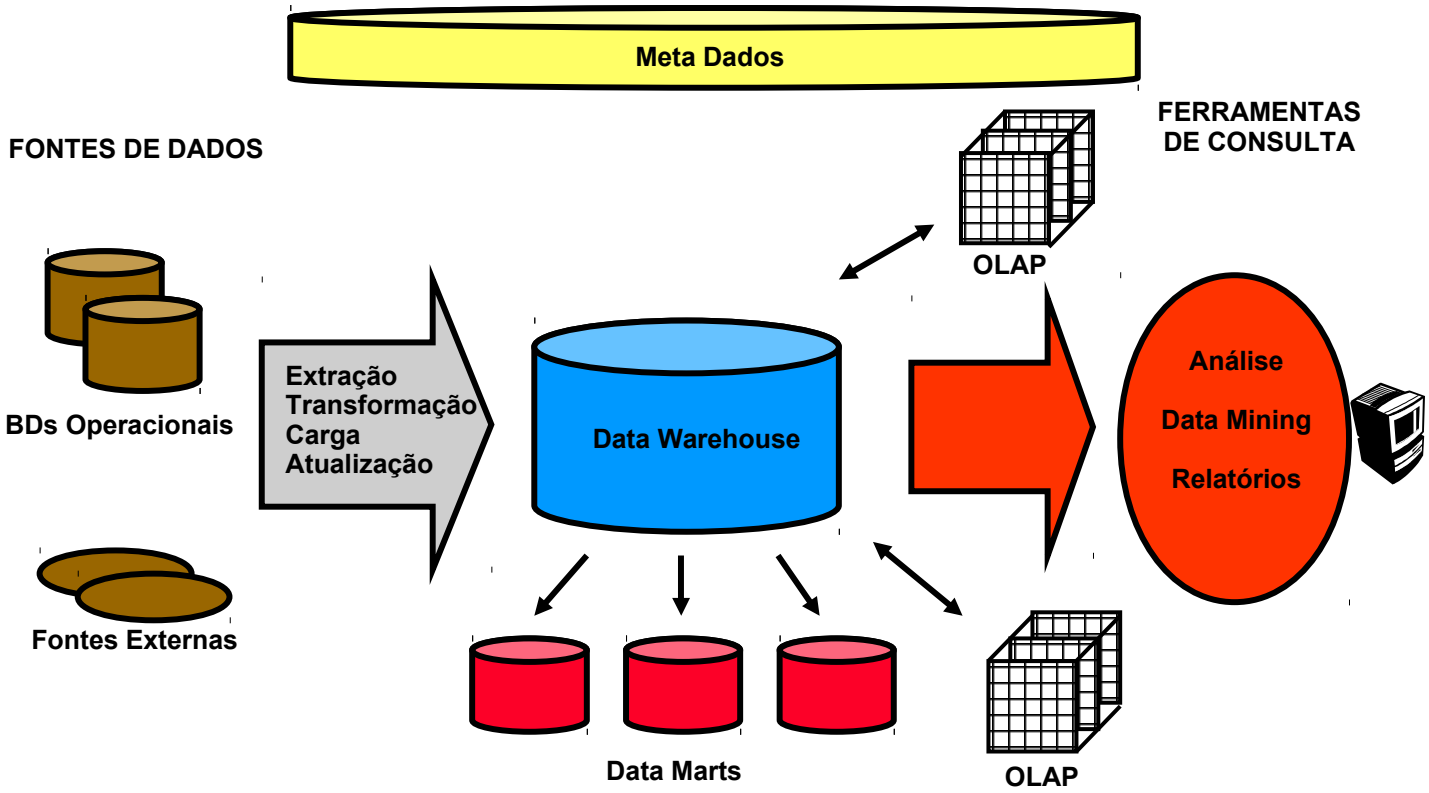
"*A Data Warehouse is a*
subject-oriented,
integrated,
time-variant,
non-volatile
collection of data in support
of management's decision-making process."

(W. Inmon)

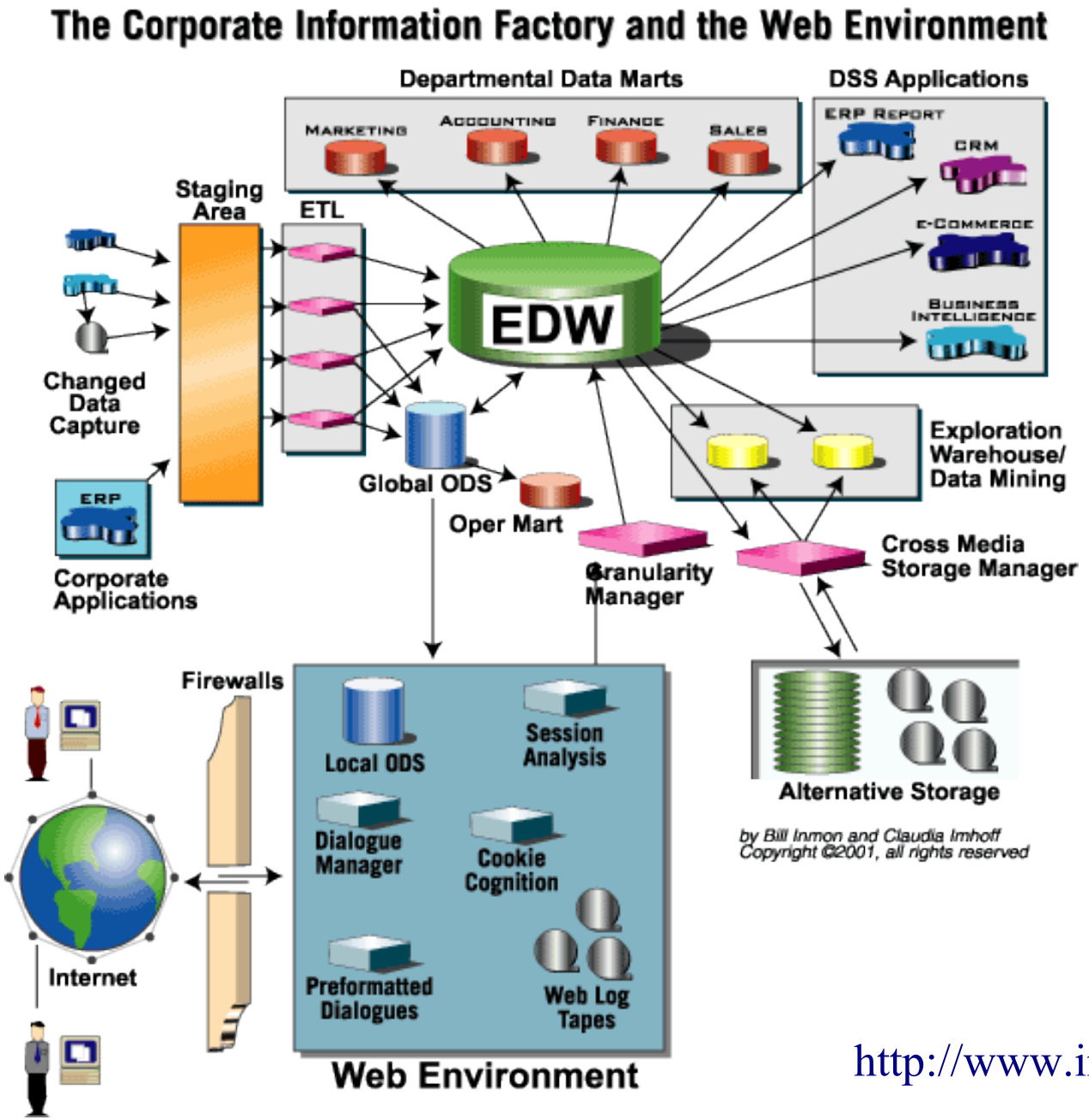
Data Warehouse – Integrated

- Construído integrando fontes de dados múltiplas e heterogêneas.
  - bancos de dados relacionais, arquivos comuns, registros de transações on line, etc.
- Limpeza de dados e técnicas de integração de dados são aplicadas.
  - Assegura consistência em convenções de nomes, estruturas de codificação, medidas de atributos, etc., entre diferentes fontes de dados.
    - » Por exemplo, Preço de Hotel: moeda, taxas, inclui café da manhã, etc.
  - Quando os dados são carregados no Data Warehouse, são convertidos para o padrão adotado.

Arquitetura Genérica de um Data Warehouse



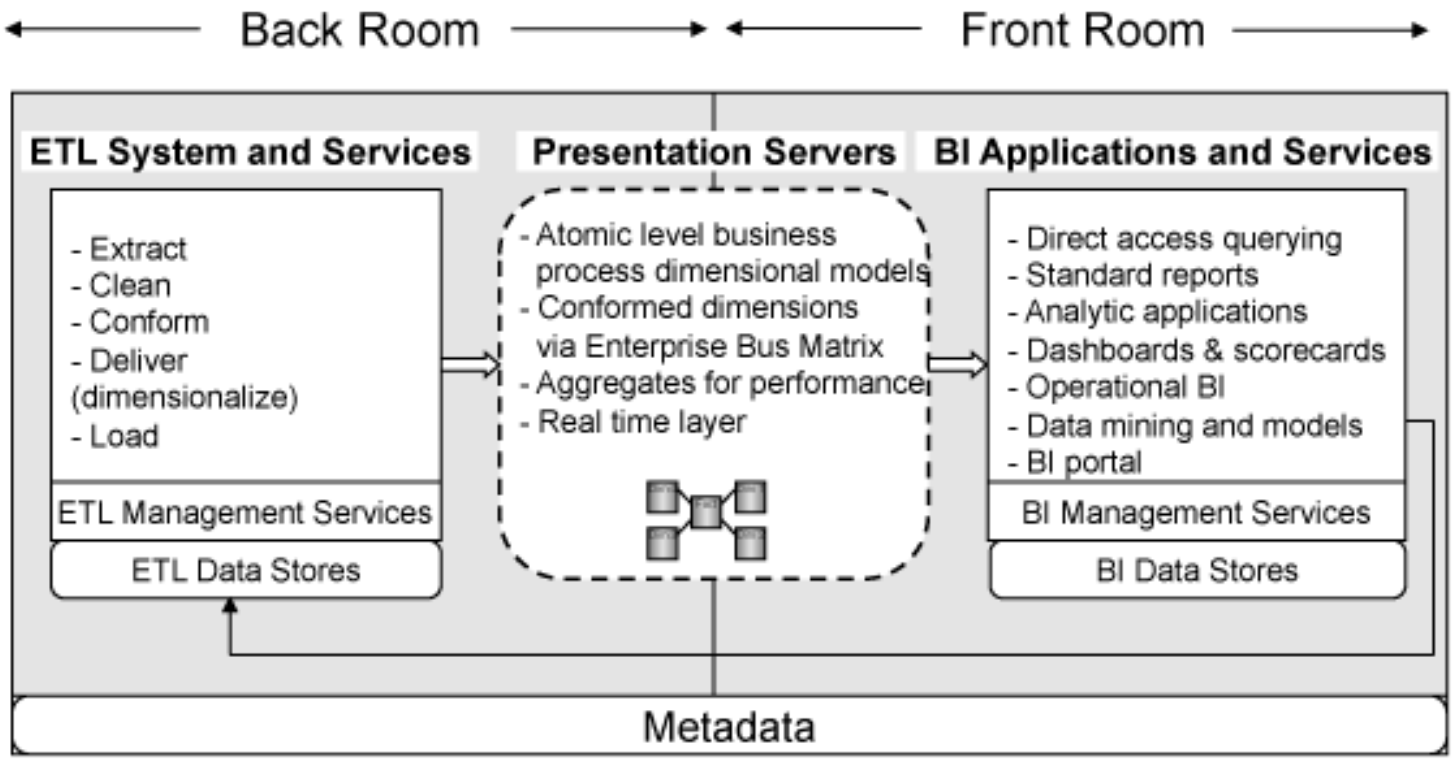
An Overview of Data Warehousing and OLAP Technology
Surajit Chaudhuri, Umeshwar Dayal SIGMOD Record 1997



Corporate Information Factory de Inmon Versão 2013

http://www.inmoncif.com/library/cif/

Data WareHouse (Kimball 2013)



http://www.kimballgroup.com/data-warehouse-and-business-intelligence-resources/kimball-core-concepts/

http://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/technical-dw-bi-system-architecture/

Arquitetura de BI com Pentaho

http://www.pentaho.com/
http://community.pentaho.com/

Data Integration & ETL

- **Independentemente da arquitetura de DW ou de BI, há um processo comum de integração de dados chamado de ETL (Extract, Transform, Load).**
- **Data Integration e ETL são termos usados indistintamente no mercado de BI, embora ETL seja apenas um possível cenário de integração de dados.**
- **Na plataforma Pentaho, a ferramenta ETL, originalmente denominada K.E.T.T.L.E. (Kettle Extration, Transformation, Transportation and Loading Environment), foi rebatizada como Pentaho Data Integration (PDI).**

Pentaho® Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration

Matt Casters, Roland Bouman, Jos van Dongen

Wiley 2010

Data Integration & ETL

- **Embora seja uma atividade de "back room" que não é visível para os usuários finais (e patrocinadores), ETL facilmente consome 70% dos recursos (pessoas, tempo, dinheiro) necessários para a implementação e manutenção de um data warehouse típico.**
- **Algumas frases de efeito sobre ETL:**
  - **"The Extract-Transform-Load (ETL) system is the foundation of the data warehouse."**
  - **"The ETL system makes or breaks the data warehouse"**
  - **"ETL is both a simple and a complicated subject."**

The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data

Ralph Kimball, Joe Caserta

Wiley 2006

Processo de ETL

- Definição original de ETL:
  - O conjunto de processos para trazer dados de sistemas OLTP para um data warehouse.
- Mas nas soluções modernas de ETL
  - Os dados provêm não só de sistemas OLTP mas de websites, arquivos “flat”, bases de dados de e-mails e de redes sociais, planilhas e bases de dados pessoais.
  - O uso de ETL não é só para carregar um único data warehouse, mas pode ter muitos outros casos de uso, como carregar data marts, gerar planilhas e modelos de data mining, e até mesmo retornar dados de volta para sistemas OLTP.
- Passos principais do processo
  - **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.

Processo de ETL

- Qual a melhor maneira de projetar e construir um sistema de ETL? “Depende”
  - das fontes de dados;
  - das limitações dos dados;
  - das linguagens de script;
  - das ferramentas de ETL disponíveis;
  - das habilidades do pessoal envolvido (TI e negócio);
  - da plataforma de BI;
  - Etc.

Requisitos de ETL

- Requisitos e restrições a considerar antes de projetar um sistema de ETL
  - Necessidades do negócio (KPIs)
  - Conformidade legal dos dados
  - Qualidade dos dados
  - Segurança
  - Integração de dados e sistemas
  - Latência dos dados
  - Archiving and Lineage
  - BI Delivery Interfaces
  - Available Skills
  - Legacy licenses

Quatro macroprocessos de ETL

- Embora seja conhecido pela sigla ETL, são quatro os macroprocessos, com 34 subsistemas, segundo Kimball:
  - **Extracting**. Gathering raw data from the source systems and usually writing it to disk in the ETL environment before any significant restructuring of the data takes place. → 3 subsistemas
  - **Cleaning and conforming**. Sending source data through a series of processing steps in the ETL system to improve the quality of the data received from the source, and merging data from two or more sources to create and enforce conformed dimensions and conformed metrics. → 5 subsistemas
  - **Delivering**. Physically structuring and loading the data into the presentation server's target dimensional models. → 13 subsistemas
  - **Managing**. Managing the related systems and processes of the ETL environment in a coherent manner. → 13 subsistemas

The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data

Ralph Kimball, Joe Caserta

Wiley 2006

Os 34 subsistemas de ETL

- **Extracting: Getting Data into the Data Warehouse**
  1. Data Profiling
  2. Change Data Capture System
  3. Extract System

- **Cleaning and conforming: Improving Data Quality Culture and Processes**
  4. Data Cleansing System
  5. Error Event Schema
  6. Audit Dimension Assembler
  7. Deduplication System
  8. Conforming System

Os 34 subsistemas de ETL

- **Delivering: Prepare for Presentation**
  9. Slowly Changing Dimension Manager
  10. Surrogate Key Generator
  11. Hierarchy Manager
  12. Special Dimensions Manager
  13. Fact Table Builders
  14. Surrogate Key Pipeline
  15. Multivalued Dimension Bridge Table Builder
  16. Late Arriving Data Handler
  17. Dimension Manager System
  18. Fact Provider System
  19. Aggregate Builder
  20. OLAP Cube Builder
  21. Data Propagation Manager

Os 34 subsistemas de ETL

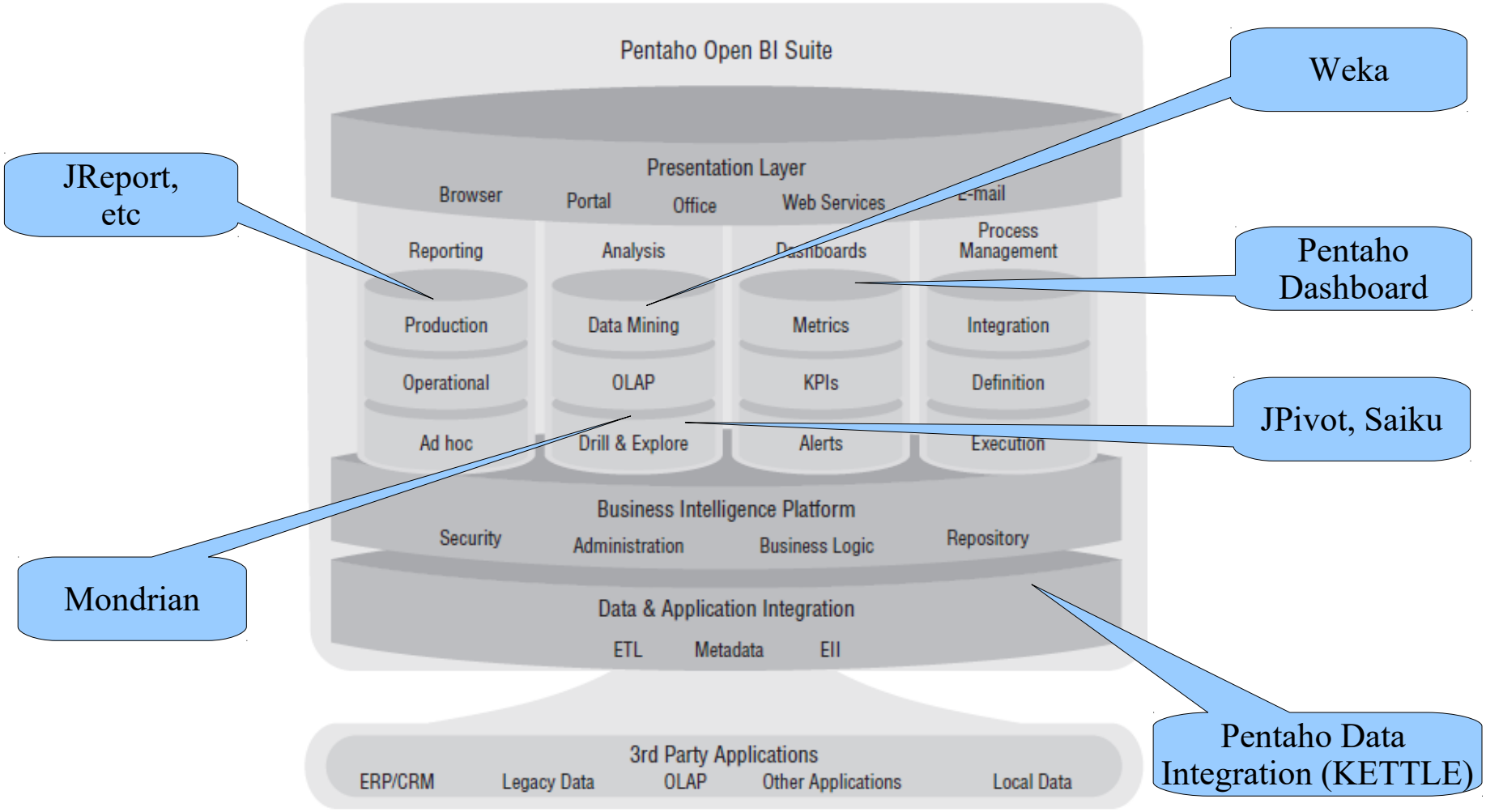
- **Managing the ETL environment**
  22. Job Scheduler
  23. Backup System
  24. Recovery and Restart System
  25. Version Control System
  26. Version Migration System
  27. Workflow Monitor
  28. Sorting System
  29. Lineage and Dependency Analyzer
  30. Problem Escalation System
  31. Parallelizing/Pipelining System
  32. Security System
  33. Compliance Manager
  34. Metadata Repository Manager

Como ensinar/aprender ETL?

- **Ver em detalhes os 34 subsistemas de ETL? No way!**
- **Aprender fazendo**
  - ETL é um processo essencialmente PRÁTICO.
  - Laboratório de Pentaho Data Integration, a.k.a. KETTLE
  - Uma das ferramentas integradas na plataforma Pentaho BI
  - Projeto open source encampado pela Pentaho em 2006 (desenvolvido por Matt Casters desde 2001)

http://community.pentaho.com/projects/data-integration/

Arquitetura da Plataforma Pentaho BI



Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL
Roland Bouman, Jos van Dongen - Wiley, 2009

Funcionalidades do Pentaho Data Integration

- **Integração de dados de diversas fontes**
  - Bancos de dados
  - Planilhas eletrônicas
  - Arquivos texto, CSV, XML, Big Data
- **Processo de ETL para carga de dados em um Data Warehouse ou Data Mart**
  - Extração de dados de diferentes fontes e formatos
    - » Validação e descarte de dados de acordo com regras e padrões
  - Transformação dos dados de acordo com requisitos técnicos e do negócio
    - » Conversão dos tipos de dados, filtragem de dados e sumarizações
  - Carga dos dados transformados em uma base de dados (DW/DM)
    - » Reescrita dos dados e adição de novas informações

Funcionalidades do Pentaho Data Integration

- **Atividades de Extração**
  - Captura dos dados
    - » Leitura a partir de diversas fontes
    - » Identificação de mudanças desde a última extração.
  - Staging
    - » Armazenamento temporário dos dados.

Funcionalidades do Pentaho Data Integration

- **Atividades de Transformação**
  - Validação dos dados
    - » Verificação se os dados estão corretos e precisos.
    - » Filtragem de dados inválidos.
  - Limpeza dos dados
    - » Correção de dados inválidos.
  - Decodificação
    - » Conversão de atributos (numéricos, categóricos) para adequação a um padrão ou regra.
  - Agregação
  - Geração e gerenciamento de chaves
    - » Dimensões identificadas por chaves substitutas (“surrogates”).

Funcionalidades do Pentaho Data Integration

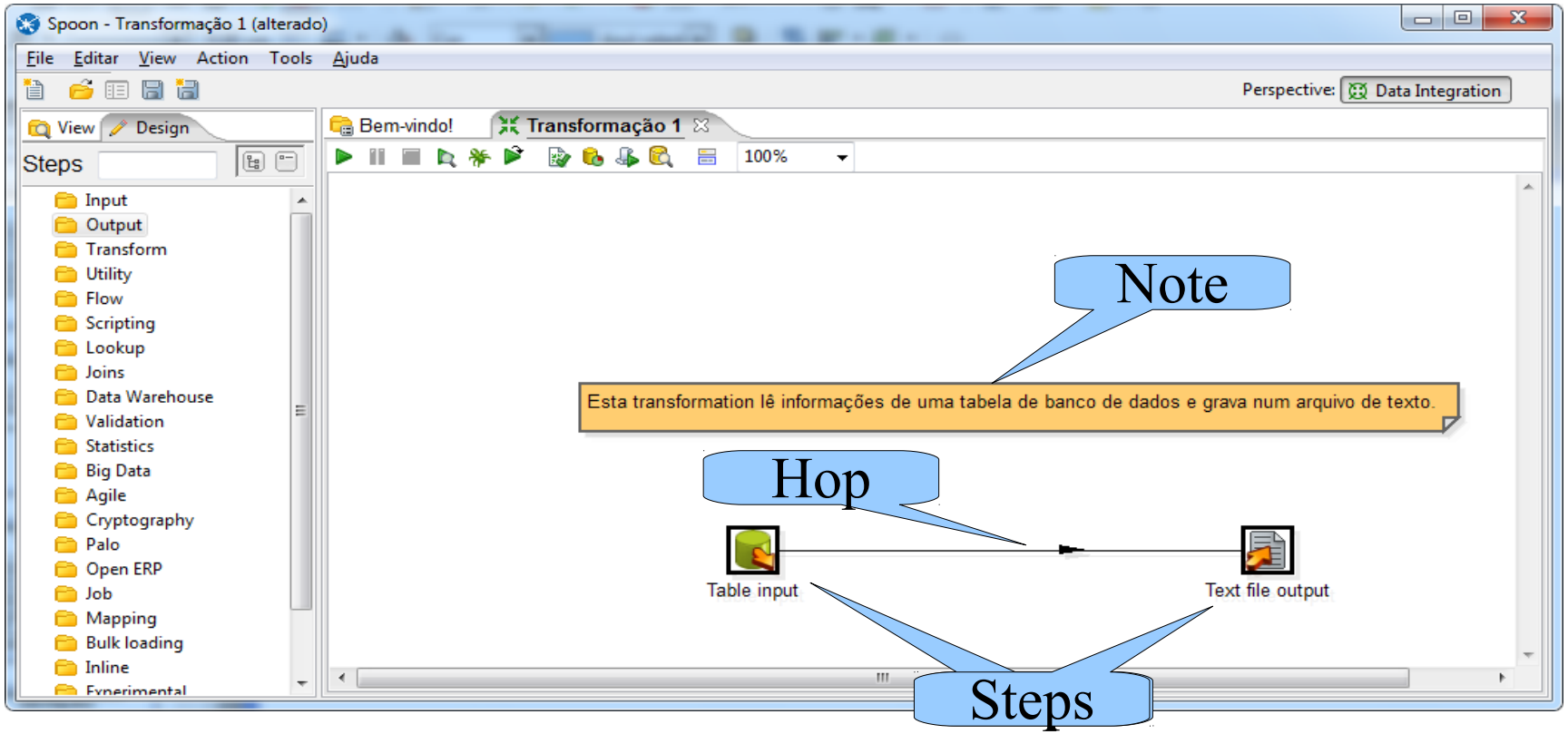
- **Atividades de Carregamento**
  - Carregamento e manutenção das tabelas de dimensões
    - » Adição e atualização de linhas das tabelas de dimensões.
  - Carregamento das tabelas de fatos
    - » Adição de linhas à tabela de fatos.
    - » Atualização de atributos de status
- **Mais do que carga de DW, o PDI vem sendo usado para outputs diversos, por exemplo:**
  - para geração de relatórios
  - para carga de bancos de dados “convencionais”
  - como entrada para diversas aplicações (vide caso NSA)

Componentes do PDI
- Transformations -

- A transformação (***transformation***) é o carro-chefe de uma solução de ETL. Ela lida com a manipulação de linhas ou dados no significado mais amplo possível da sigla ETL. Consiste em um ou mais passos (***steps***) que realizam trabalhos básicos de ETL, como ler dados de arquivos, filtrar linhas, limpar dados, ou carregar dados em um banco de dados.
- Os ***steps*** são ligados por saltos (***hops***), que definem um canal unidirecional que permite aos dados fluírem entre os ***steps*** que estão ligados. No PDI, a unidade de dados é a linha, e um fluxo de dados é o movimento de linhas de ***step*** em ***step***.
- Além de ***steps*** e ***hops***, ***transformations*** também podem conter ***notes***. São pequenas caixas com notas que podem ser colocadas em qualquer lugar, com um texto arbitrário para documentar a ***transformation***.

Componentes do PDI
- Transformations -

- ***Transformations, Steps, Hops, Notes***



Interface gráfica do PDI: spoon

Laboratório de Introdução ao Pentaho Data Integration

Vide Tarefa no Moodle

Referências