Armazéns de Dados

Departamento de Engenharia Informática (DEI/ISEP)
Paulo Oliveira
pjo@isep.ipp.pt

Adaptado do Original de:
Fátima Rodrigues (DEI/ISEP)

1

Data Warehouse Architectures

2

Bibliography

- Mastering Data Warehouse Design: Relational and Dimensional Techniques
  Claudia Imhoff, Nicholas Galemmo, Jonathan G. Geger
  Wiley, 2003

  Chapters 1, 13

- The Data Warehouse Lifecycle Toolkit: Experts Methods for Designing, Developing, and Deploying Data Warehouses
  Ralph Kimball, Laura Reeves, Margy Ross, Warren Thornthwaite
  Willey, 1998

  Chapters 8, 9, 10

- Modern Database Management
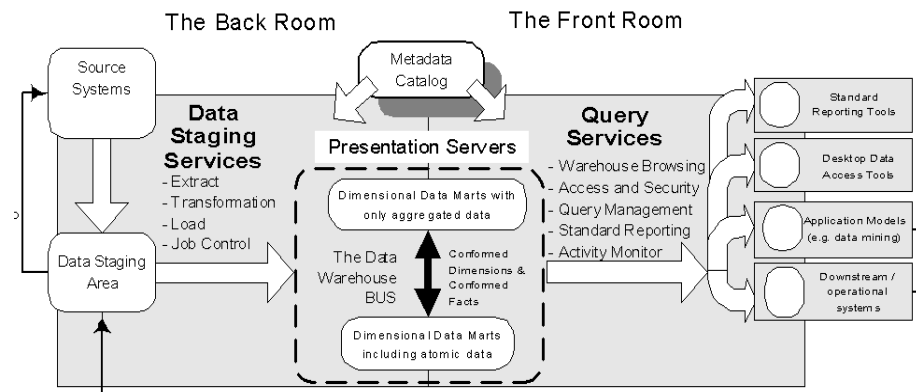 J.Hoffer, M.Prescott, H. Topi
  Prentice Hall, 2008

  Chapter 11

3

3

Data Warehouse Bus Architecture (Ralph Kimball)

4

DW BUS Architecture

The Back Room   The Front Room

Source Systems

Metadata Catalog

Data Staging Services
- Extract
- Transformation
- Load
- Job Control

Data Staging Area

Presentation Servers

Dimensional Data Marts with only aggregated data

The Data Warehouse BUS

Conformed Dimensions & Conformed Facts

Dimensional Data Marts including atomic data

Query Services
- Warehouse Browsing
- Access and Security
- Query Management
- Standard Reporting
- Activity Monitor

Standard Reporting Tools

Desktop Data Access Tools

Application Models (e.g. data mining)

Downstream / operational systems

- DW is built on a series of incremental data marts
  - "**Bottom-up**" or **incremental methodology**
- Has two major types of components: **services** and **data stores**

5

5

DW BUS Architecture

- Is divided into two groups of components and processes
  - **Back-room (data acquisition)**
    - Part responsible for gathering and preparing the data
    - Where data acquisition and data staging processes take place
  - **Front-room (data access)**
    - Part responsible for delivering data to business users

- Flow of data from source systems to user desktop is supported by the **metadata catalog**

- Includes two types of data marts in the data presentation area
  - **Atomic data marts**
  - **Aggregated data marts**

6

6

3

Back-Room

- Where the **data staging process** takes place

- **Engine room** of the DW

- **Primary concern**:
  - Getting the right data, with the appropriate transformations, at the right time, and load it into the DW

Source Systems

- Are the **obvious sources** of interesting business data

- Other high-value sources may be **external to the business**
  - Demographic customer information, target customer lists, and competitive sales data

- Data storage types are dictated by the source system
  - Many older legacy systems are standard mainframe data storage facilities: IMS, IDMS, VSAM, and DB2 are common

- **Flat file** is one often standard source for the DW

- Understanding their nature **is critical** for creating the back-room architecture

Data Staging Area

- Is both a **storage area** and **set of processes** commonly referred as **Extraction, Transformation and Loading** (ETL), not seen by end-users

- **Everything between** the source system and the DW presentation server

- Where much of the **data transformation** takes place and much of the **added value** of the DW is created
  - **Cleaning the data**
    - Correcting misspellings, resolving domain conflicts, dealing with missing values, or parsing into standard formats
  - **Integrating data** from multiple sources
  - **De-duplicating data**
  - **Assigning surrogate DW keys**
  - **...**

9

9

Back-Room Services

- **Tools and techniques** employed in the data staging process
  - Also known as *data staging services*

- **Service** is an elementary function or task, that can be as simple as:
  - Creating a table in a database
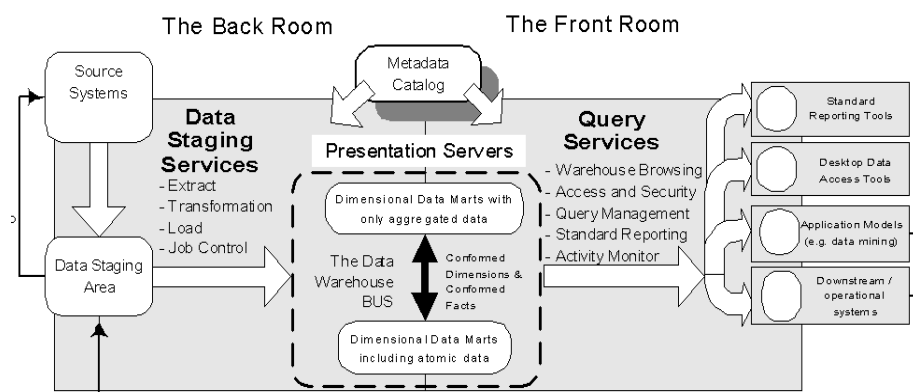 - Copying data from one table to another

10

10

Back-Room Services

- Extract services
  - **Pulling the data from the source system(s)**
  - Largest effort in the DW project, especially if the source systems are decades-old or mainframe-based

- Data transformation services
  - Acts performed on the data to **convert it into something presentable to users and valuable to the business**

- Data loading services
  - Set of services responsible by **loading the data into the DW**

- Job control services
  - Captures metadata regarding the **progress and statistics of job execution**

11

11

DW BUS Architecture

The Back Room   The Front Room

Source Systems

Metadata Catalog

Data Staging Services
- Extract
- Transformation
- Load
- Job Control

Data Staging Area

Presentation Servers

Dimensional Data Marts with only aggregated data

The Data Warehouse BUS

Conformed Dimensions & Conformed Facts

Dimensional Data Marts including atomic data

Query Services
- Warehouse Browsing
- Access and Security
- Query Management
- Standard Reporting
- Activity Monitor

Standard Reporting Tools

Desktop Data Access Tools

Application Models (e.g. data mining)

Downstream / operational systems

- DW is built on a series of incremental data marts
  - "**Bottom-up**" or **incremental methodology**

- Has two major types of components: **services** and **data stores**

12

12

6

Presentation Server

- Where the **data is stored** for direct querying by end-users, OLAP tools, reporting systems and other applications

- Is a **series of integrated data marts**
  - Data mart presents the data from a single business process

- Data in the queryable presentation server of the DW must be:
  - **Dimensional**
  - **Atomic** (to unpredictable ad-hoc user queries)

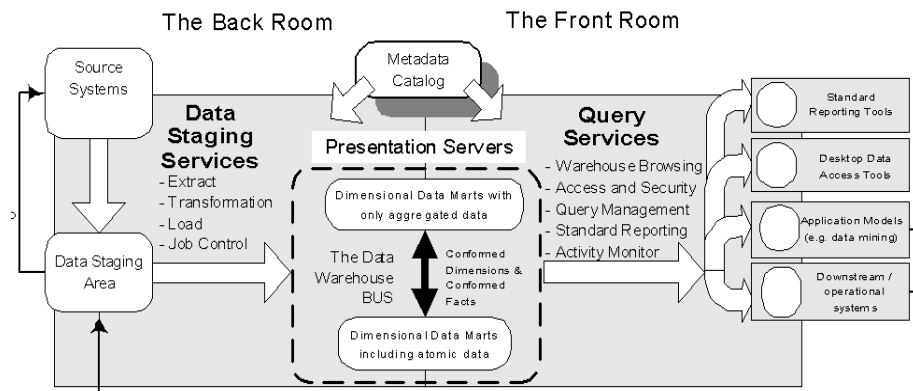- All data marts must be built using *common/ shared dimensions*

Data Marts in Presentation Server

- **Atomic data marts**
  - Hold multidimensional data at the **lowest detail level**

- **Aggregated data marts**
  - Hold multidimensional data that have been **aggregated**
  - Improve **query performance**
  - Loaded from the **data staging area** or from the **atomic data marts**

- All star schema-based data marts **may or not reside within the same database instance**

- Collection of star schemas which share dimensions and facts is the basis of the **DW Bus Architecture**

DW BUS Architecture

The Back Room | The Front Room

- Metadata Catalog
- Source Systems
- **Data Staging Services**
  - Extract
  - Transformation
  - Load
  - Job Control
- Data Staging Area
- **Presentation Servers**
  - Dimensional Data Marts with only aggregated data
  - The Data Warehouse BUS
  - Conformed Dimensions & Conformed Facts
  - Dimensional Data Marts including atomic data
- **Query Services**
  - Warehouse Browsing
  - Access and Security
  - Query Management
  - Standard Reporting
  - Activity Monitor
- Standard Reporting Tools
- Desktop Data Access Tools
- Application Models (e.g. data mining)
- Downstream / operational systems

- DW is built on a series of incremental data marts
  - "**Bottom-up**" or **incremental methodology**

- Has two major types of components: **services** and **data stores**

15

15

Front-Room

- **Public face of the DW**
  - It's what the business users see and work with day-to-day
- **Data access services are between the users and the data**, hiding some of the complexities and helping them to find what they are looking for

16

16

8

Front-Room Services

- Warehouse browsing
  - Takes advantage of the metadata catalog to support the users in their efforts to **find and access the data they need**
- Access and security services
  - Facilitate a **user's connection to the DW**
- Activity monitoring services
  - Capture information about the **use of the DW**
- Query management services
  - Set of capabilities that manage the exchange between the **query formulation**, the **execution of the query** on the database, and the **return of the result set** to the desktop
- Standard reporting services
  - **Ability to create fixed-format reports** that have limited user interaction and regular execution schedules

17

17

Corporate Information Factory (Bill Inmon)

18

CIF Architecture
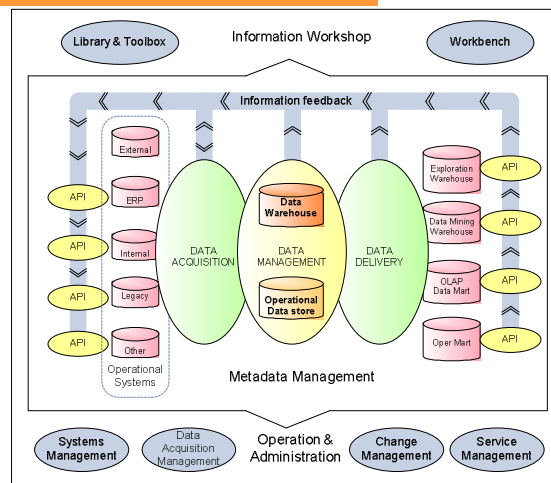
Logical architecture with **major databases** and **processes** to effectively and efficiently move data from source systems to business users
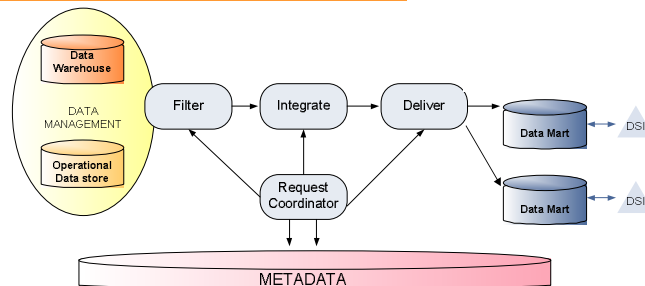
Data Acquisition
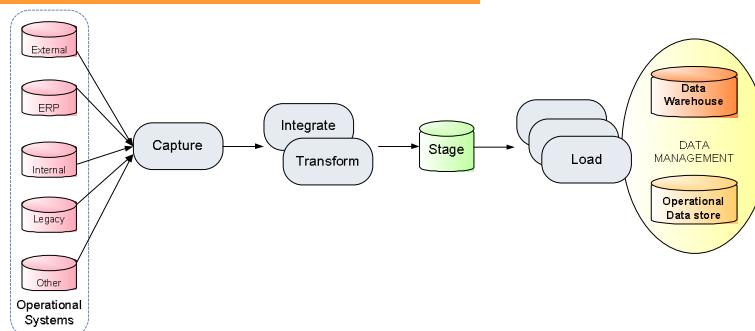
- Set of processes and programs that extract data from the operational systems to the **DW** and to the **Operational Data Store (ODS)**

- Perform the **cleaning**, **integration** and **transformation** of the data into an enterprise format

Data Delivery

- Process that moves data from the DW or ODS into **data marts** and **oper marts**
- Like in the acquisition layer, data is manipulated as it is moved
- Origin is the **DW** or **ODS**, which already contains high quality integrated data that **conforms to the enterprise business rules**

DW vs. ODS

- **DW** – A **subject-oriented**, **integrated**, **time variant** and **non-volatile** collection of data used in strategic decision making [Inmon and Hackathorn, 1994]

- **ODS**
  - Data is **fully integrated** like in a DW
  - **Data is current**
  - **Data is volatile or updatable** (no history is retained)
  - Data is **usually entirely detailed**
  - Source of **near real-time** and **accurate data** – accessible from anywhere in the corporation

Data Marts vs. Oper Marts

- **Data marts** are customized and/or aggregated subsets of data derived from the DW
  - Where the analytical activities take place

- Data in each data mart is usually **tailored for specific analytical requirements** of a business unit or function
  - Product profitability analysis, sales analyses, …

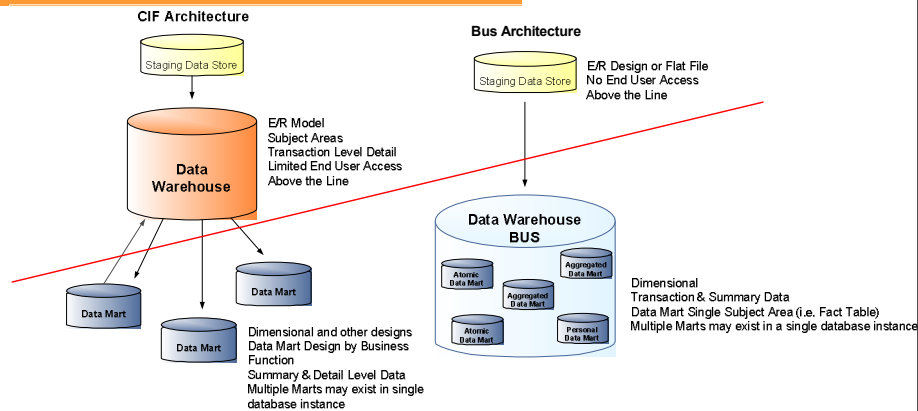- **Oper marts** are derived from the **ODS** and used to provide the business community with **dimensional access to current operational data**

23

23

Bus Architecture vs CIF Architecture

24

Comparing CIF and BUS Architectures

Access is usually not allowed above the red line
- Back room is completely off-limits to users in the BUS architecture
- Direct access to the CIF DW is discouraged

25

CIF vs. BUS: Differences

- **No physical repository equivalent** to the CIF DW in the BUS architecture

- BUS DW is the collection of **atomic and aggregate data marts**

- BUS architecture data marts **(star schemas)** are significantly different from the design of the CIF DW **(relational schema)**

- Various data marts schemas "conform" through **common dimensions** in the BUS architecture

- In the BUS architecture **all components are dimensional**, **except the data staging area**

26

CIF vs. BUS: Similarities

Both architectures:

- Have a **separate staging area**, **metadata management**, **data acquisition** and **data delivery** processes

- Power of information resides in the **atomic data**, which embed all available information dimensionality

- Existence of **dimensional data marts**
    - **Aggregate data mart** in BUS architecture is **usually the same** as the **data mart** in the CIF architecture
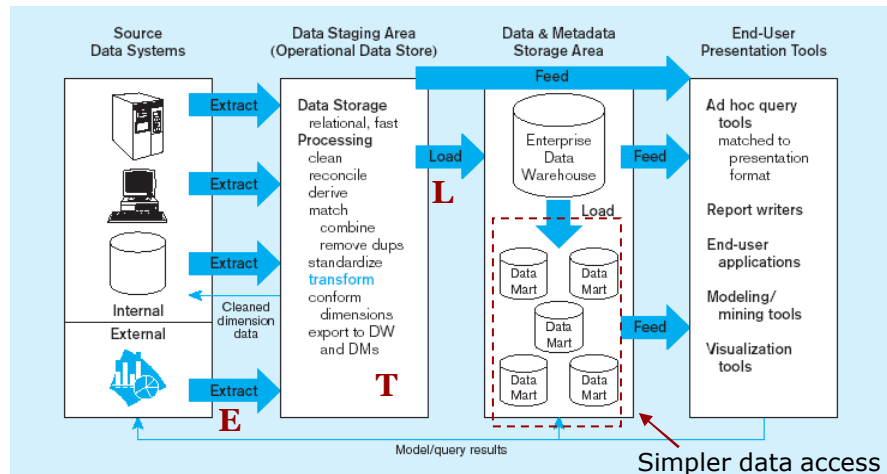
Kimball DW Architecture
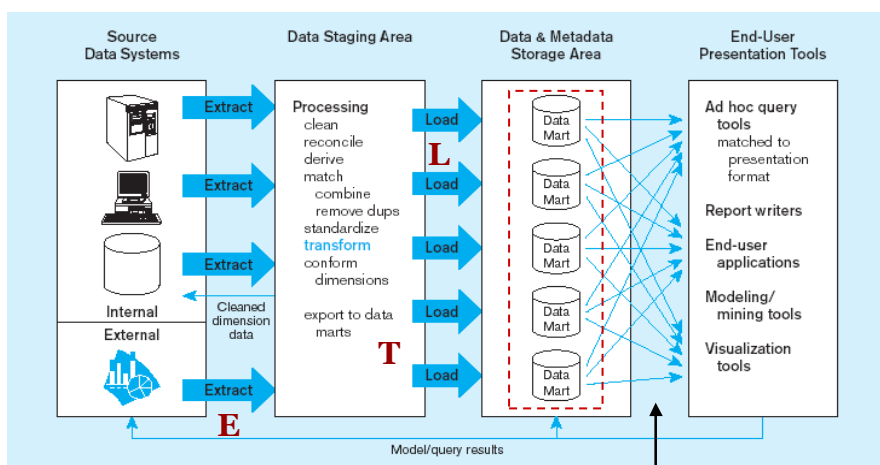
➔ Separate ETL for each **independent** data mart

➔ No single consolidated DW from the beginning

Inmon DW Architecture

➔ Single ETL for **Enterprise Data Warehouse**

➔ Data marts loaded from Enterprise DW

Kimball Approach

- **Most common approach**

- Begins with **a single data mart** and **others are added over time** for more subject areas
  - Will require an overall **integration plan**

- Relatively **inexpensive** and **easy** to start to implement
  - Can be used as a **proof of concept** for DW

- Separate ETL process is developed for each data mart, which yields **costly redundant data and processing efforts**

- Can perpetuate the **"silos of information" problem**

- Key is to a have an **overall plan for integrating the different data marts**

Inmon Approach

- **Comprehensive DW is built initially**

- Data marts are built using **aggregate subsets of the data in the DW**

- Like all complex projects, it is **expensive**, **time consuming**, and **prone to failure**

- When successful, it results in an **integrated and scalable DW**

31

31