Master en
Big Data e Inteligencia Artificial

Module 5

Data Analytics

Lesson 0: Introduction to Data Warehouses

Juan Carlos Trujillo - Lucentia (jtrujillo@dlsi.ua.es)
Alejandro Maté - Lucentia (amate@dlsi.ua.es)

Universitat d'Alacant
Universidad de Alicante

1

Master en
Big Data e Inteligencia Artificial

Table of contents

Universitat d'Alacant
Universidad de Alicante

2

Introduction to Datawarehouses

- What are Data Warehouses and how do they arise?
- Evolution of Data Warehouse technologies
- Architecture and data flow in Data Warehouses
- Data Warehouse design techniques

3

Máster de Formación Permanente en BIG DATA e Inteligencia Artificial
Master en
Big Data e Inteligencia Artificial
khaos RESEARCH

What are Data Warehouses and how they arise?

- Development/evolution of data warehouse systems
  - From first file Management Systems (70's) to the current DataBase Management Systems (DBMS)
    - Efficient DBMS
    - Robust DBMS
    - Great variety of tools that faciliate their use
      - Servers
      - Back-end and Front-end tools

4

What are Data Warehouses and how they arise?

- Typical query on RDBMS
  - How many shoes did we sell last month?
- Highly competitive environments
  - Companies need to adopt strategy decisions
    - How many red shoes were sold in the northern zone, east and southern last month; compared to those sold in the same month last year?
    - What kind of client has been buying the model BMW 320i during the last 10 years?
      - Manager, Professor, operate worker, etc.?

Máster de Formación Permanente en BIG DATA e Inteligencia Artificial
Master en
Big Data e Inteligencia Artificial
khaos RESEARCH

What are Data Warehouses and how they arise?

- Req. 1. Vast volume of data (several years, clients, products, stores etc.)
  - Historical and normally from heterogeneous sources
- Req. 2. Have to be presented in a friendly and easy to use framework
  - Understanding strategic questions

What are Data Warehouses and how they arise?

- Are OLTP systems suitable for those decisions?
  - Problems
    - Historical data cannot be obtained from daily OLTP systems
    - Normally in different data sources
      - Suppliers, Clients, Components, Suppliers, components, faulty products, etc.
    - Managers cannot query those OLTP systems
    - And some more...

Máster de Formación Permanente en BIG DATA e Inteligencia Artificial
Master en
Big Data e Inteligencia Artificial
khaos RESEARCH

What are Data Warehouses and how they arise?

- Using OLTP systems would require:
  - Integrating data -> Time consuming (req. 1)
  - Could a Manager query those systems ? (req. 2)
  - **Impossible!!!!!**
- Let us go back in history and make a brief summary on the I.T. solutions used to manage historical data

Debido al volumen de los sistemas transaccionales es imposible hacer muchas queries, aparece el concepto de data warehouse

Introduction to Data Warehouses

- What are Data Warehouses and how do they arise?
- Evolution of Data Warehouse technologies
- Architecture and data flow in Data Warehouses
- Data Warehouse design techniques

9

Máster de Formación Permanente en BIG DATA e Inteligencia Artificial
Master en
Big Data e Inteligencia Artificial
khaos RESEARCH

Evolution of Data Warehouses technologies

- Data from legacy systems
  - 1970's we used huge mainframes (IBM)
    - Cobol, CICS, IMS, DB2, etc
  - 1980's platforms such as AS/400 y VAX/VMS
  - Nowadays, many business applications "run" on these systems
    - Gathering data and business rules for many years à difficult to take them to another system
    - Data is put in libraries where business applications will access
  - High cost for these business applications

10

Evolution of Data Warehouses technologies

- Information served at the desktop (1990's)
  - Reducing distance between final user and programmer
    - PC with spreadsheet, Analysis tools, etc.
  - Analysis tools to access data produced by legacy systems
  - Problem: data remain spread and are oriented to specific needs for certain final users
    - Partial solutions
    - No all users have the same "domain" on computers

Evolution of Data Warehouses technologies

- Decision Support Systems (DSS) and Executive Information Systems (EIS):
  - DSS: detailed information. Medium and low managers
  - EIS: consolidated information. High executives
    - More oriented to the Multidimensional view of data
  - They are similar and overlap functionalities
  - They are the Data warehouses predecessors
  - Still expensive and partial solutions for specific needs instead of a global solution

Evolution of Data Warehouses technologies

- Summary: common features for DSS and EIS:
  - Data described in standard business terms, instead of technical terms such as tuple, file or relational table.
    - Systems focused on non-technical users
  - Pre-processed data following business rule patterns
    - Benefits of purchased products in different stores
  - Consolidated and summary data views
    - Although they allow us to see data in detail, they rarely allow us to do it

Máster de Formación Permanente en BIG DATA e Inteligencia Artificial
Master en
Big Data e Inteligencia Artificial
khaos RESEARCH

Evolution of Data Warehouses technologies

- Data warehouse techniques and systems provide analytical tools provided by their forerunners
- Provide global solutions for an organization, instead of partial solutions
  - Data oriented to satisfy to the whole organization

Evolution of Data Warehouses technologies

- The Data Warehouse (DW)
  - Systems that store historical data to be used by Decision Support Systems
  - Basically, they are query systems focused on extracting knowledge from the stored historical data
  - The data analysis -> On-Line Analytical Processing (OLAP)
  - Using the multidimensional modeling (cubes, hypercubes, etc)

Máster de Formación Permanente en BIG DATA e Inteligencia Artificial
Master en
Big Data e Inteligencia Artificial
khaos RESEARCH

Evolution of Data Warehouses technologies

- Definition by W. Inmon (the "father", 1992)

"A subject-oriented, integrated, time-variant, and non-volatile collection of data used in support of management's decisions"

LUCENTIA UNIVERSITY
ucentia
Universitat d'Alacant
Universidad de Alicante

Evolution of Data Warehouses technologies

- Subject oriented
  - The data warehouse is organized by “data subjects” that are relevant to the organization.
    - Subjects: Sales, Purchases, shipments, etc.
    - Context of analysis: clients, suppliers, products, etc…
  - Multidimensional Modeling (First approach)
    - Facts -> Activities of a high interest
    - Dimensions -> context of analysis

Evolution of Data Warehouses technologies

- Integrated
  - Data integrated from different data sources to provide a comprehensive view
- Time variant
  - Historical data: related to a time period and incremented periodically
- Non-volatile
  - Data are not updated or erased by users. New data is alwayes added.

Evolution of Data Warehouses technologies

	OLTP	DW/OLAP
User	Clerk, IT Professional	Knowledge Worker
Function	Day-to-day Operations	Decision Support
Database Design	Application-oriented (E-R based)	Subject-oriented (Star, Snowflake)
Data	Current, Isolated	Historical, Consolidated
View	Detailed, Flat Relational	Summarized, Multidimensional
Usage	Structured, Repetitive	Ad-Hoc
Unit of Work	Short, Simple Transaction	Complex Query
Access	Read / Write	Read Mostly
# Records Accessed	Tens	Millions
# Users	"Thousands"	"Hundreds"
Database Size	100 MB-GB	100 GB-TB
Performance Metric	Transaction Throughput	Query Throughput Response

Introduction to Data Warehouses

- What are Data Warehouses and how do they arise?
- Evolution of Data Warehouse technologies
- Architecture and data flow in Data Warehouses
- Data Warehouse design techniques

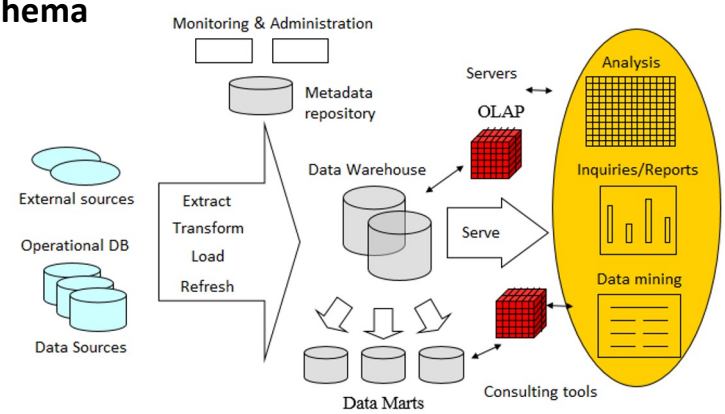
Architecture and data flow in Data Warehouses

- There are different variants of the architecture of DWs according to the needs of the business
- Some of them:
  - **Traditional architecture**: Designed for analysis of large amounts of structured data
  - **Real-time**: When decisions require data as soon as they are generated
  - **For "Big Data"**: When the information to be treated includes unstructured information (social networks!)

Máster de Formación Permanente en BIG DATA e Inteligencia Artificial
Master en
Big Data e Inteligencia Artificial
khaos RESEARCH

Architecture and data flow in Data Warehouses

- **General schema**



Universitat d'Alacant
Universidad de Alicante

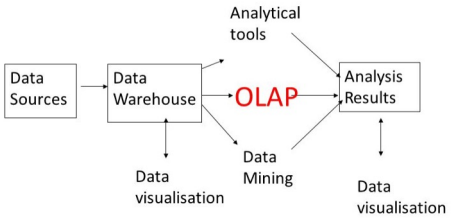
Architecture and data flow in Data Warehouses

- Three layer architecture
  - Repository server or Data Warehouse database
    - Almost always a Relational DBMS
  - OLAP Servers
    - Relational OLAP (ROLAP)
      - Extend relational DBMS to allow MD operations
    - Multidimensional OLAP (MOLAP)
      - Directly implements the MD model in vectors

Máster de Formación Permanente en BIG DATA e Inteligencia Artificial
Master en
Big Data e Inteligencia Artificial
khaos RESEARCH

Architecture and data flow in Data Warehouses

- Three Layer Architecture (II)
  - Customers -> Tools
    - Reports and consultations
    - OLAP (On-Line Analytical Processing)
    - Data Mining



LUCENTIA UA
ucentia
Universitat d'Alacant
Universidad de Alicante

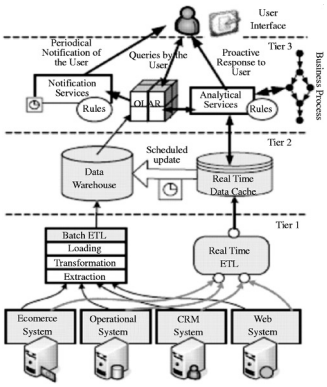
Architecture and data flow in Data Warehouses

- Why Separate Data Warehouse?
  - Performance
    - Complex OLAP queries -> Server deceleration
    - Methods of implementation, access, etc. Different
  - Functionality
    - Data not available -> Historical
    - Consolidated data (aggregates, sums, summaries, etc.) from different sources
    - Quality of data
    - Different sources -> different representations, etc ...

Máster de Formación Permanente en BIG DATA e Inteligencia Artificial
Master en
Big Data e Inteligencia Artificial
khaos RESEARCH

Architecture and data flow in Data Warehouses

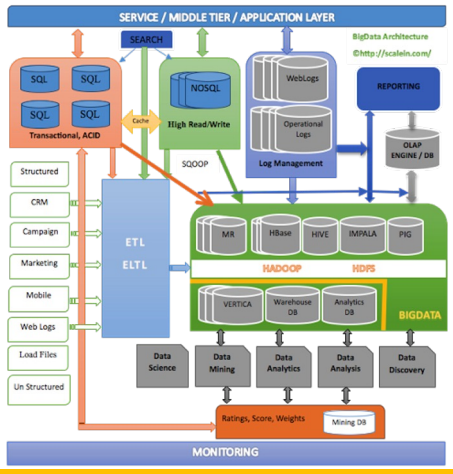
- Traditional architecture: more in detail in the following section
- In real time:



Source: Nguyen Manh *et al.* (2005)

Architecture and data flow in Data Warehouses

- Typical BigData Architecture: Characteristics:
  - Storage of different types of data
    - Semi-structured (Marketing / campaigns / mobile / web logs)
    - Structured
    - Log files
  - Loading data from different databases (MySQL, Oracle, PostgreSQL, MongoDB, etc.)
  - Data mining
  - Analytical
  - Data warehouses for reporting
  - Batch Analysis (Hadoop)
  - Web caching
  - Search



Introduction to Data Warehouses

- What are Data Warehouses and how do they arise?
- Evolution of Data Warehouse technologies
- Architecture and data flow in Data Warehouses
- Data Warehouse design techniques

Data Warehouse design techniques

- Data Warehousing techniques
  - Efficiently integrate database technologies with data analysis technologies
    - Databases: A DBMS that supports the DW repository
    - Data Analysis: Tools that allow us to accomplish an easy data analysis
      - The more extended ones: OLAP tools
        - Multidimensional Analysis based on the multi model

Máster de Formación Permanente en BIG DATA e Inteligencia Artificial
Master en
Big Data e Inteligencia Artificial
khaos RESEARCH

Data Warehouse design techniques

- Example: show product **sales** grouped by sold **products**, **stores** where they were sold and **time**

Sales			Product.Grou = "Grocery"			
			Food		Drink	
			Refrig.	Fresh	soft	Alcohol
Store. Region = "Comunidad Valenciana"	Alicante	Albatera	100	200	300	400
		Elche	500	600	700	800
	Valencia	Sagunto	900	1000	1100	1200
		Cullera	1300	1400	1500	1600

Universitat d'Alacant
Universidad de Alicante

Data Warehouse design techniques

- Advantages for companies
  - Decisions supported by reliable data
  - Profitability of investments
  - Increasing competitiveness in hostile environments
  - Friendly frameworks -> managers analyze data by themselves
  - At last managers can understand computers -> We have achieved it !!!
- Disadvantages
  - Underestimate the required resources to populate the DA from the operational data
  - Bad planning of the needed effort to achieve a good DW design
  - Never ended -> continuous increasing of ad-hoc requirements
  - ¡¡¡¡ **Be careful** !!!! The more data we have -> The more security we need

Data Warehouse design techniques

- First Methodological approach for data warehouse design
  - Top-down
    - Design and modeling the DW based on user's requirements.
    - Gather data to populate the DW from data sources.
    - Design ETL processes
    - It is normally the most used approach
    - Only applicable to very simple projects

Data Warehouse design techniques

- Bottom-up
  - Design and modeling the DW based on the data already existing in the operational data sources of the Enterprise.
  - Design ETL processes
  - The final user analysis has to be based on the design instead of guiding the whole design process
- Hybrid (combining both approaches)

33

Module 5

Data Analytics

Lesson 0: Introduction to Data Warehouses

Juan Carlos Trujillo - Lucentia (jtrujillo@dlsi.ua.es)
Alejandro Maté - Lucentia (amate@dlsi.ua.es)

34