

# Chapter 6 – Data WareHouse Part I

INFO401

B.Wakim

# Outline

- I. Definition and characteristics of a data warehouse
- II. Architecture of a data warehouse
- III. Lifecycle of a data warehouse
  - 1. Analysis
  - 2. Design (Dimensional Modeling)
  - 3. Import data (ETL)
  - 4. Install front-end tools
  - 5. Test and deploy
- IV. Usage of a data warehouse
- V. Advantages of the warehousing approach
- VI. Operational DB Systems (OLTP) vs. Data Warehouse (OLAP)
- VII. Data Warehouse vs. Data Mining



# I- Definition

- It is a system used for reporting and data analysis: A central repository of integrated data from one or more sources.
- It is a “*Queryable source of data in the enterprise*”
- They store current and historical data and are used for creating trending reports for senior management reporting such as annual comparisons.
- *Examples:*
  - Sales
  - Purchasing
  - Supplier Promotions: National, Region, Store Level





# Definition

- A database filled with large volumes of cross-indexed historical business information that users can access with various query tools.
- The warehouse usually resides on its own server and is separate from the transaction-processing or “run-the-business” systems.



# Purpose of a data warehouse

- Provides an architecture for the flow of data from operational systems to decision support systems
  - DW involves a many record analysis, during which all data has to be loaded
- Used to discover trends and patterns
  - Present opportunities
  - Identify problems





# Example: Sales Data Warehouse Example (1)

- **Requirements Gathering**
- This phase involved the following steps:
  - Collect some business questions the users want an answer for.
  - Gather details/requirements from the business users.
  - Get user sign off on the business questions.
- **Business Questions:**
  - Can we profile our “best subscribers” to pull lists of “like” non-subscribers that we could touch in some way
  - Who exists in the marketplace and have we touched them?
  - Can we build a loyalty model based on a subscriber’s payment history?



# Example: Sales Data Warehouse Example(2)

## Sales Agent Details

Sales Agent	Source Name	Phone	Address	St	City	ZIP	# Sales	Units Sold	\$ Sold	Discount Cost	Premium Cost
AD CONTRACT START	ULTIMATE TAN OF SMITHFIELD	(757) 365-9400	13412 BENNS CHURCH BLVD	VA	SMITHFIELD	23430	1	260	\$30.99	\$37.91	\$5.00
AD CONTRACT START	CHOREY & ASSOCIATE	(757) 539-7451	330 W CONSTANCE RD # 100	VA	SUFFOLK	23434	1	260	\$30.99	\$37.91	\$5.00
AD CONTRACT START	CHOREY AND ASSOCIATE	(757) 539-7454	804 W WASHINGTON ST	VA	SUFFOLK	23434	1	260	\$30.99	\$37.91	\$5.00
AD CONTRACT START	VIRGINIA STAGE CO	0	254 GRANBY ST	VA	NORFOLK	23510	1	260	\$30.99	\$37.91	\$5.00
AD CONTRACT START	SPINE & ORTHAPEDIC CTR, PC	0	6160 KEMPSVILLE CIR # 303A	VA	NORFOLK	23502	1	260	\$30.99	\$37.91	\$5.00
AD CONTRACT START	COUNTRYWIDE HOME LOAN	0	3000 WOODLAWN DR	VA	SUFFOLK	23434	1	260	\$30.99	\$37.91	\$5.00
AD CONTRACT START	WHITE, E.D.	0	730 10TH ST	VA	VIRGINIA BEACH	23451	1	260	\$30.99	\$37.91	\$5.00
AD CONTRACT START	PERMANENT COATING SOLUTIONS IN	(757) 539-4366	434 N MAIN ST # D	VA	SUFFOLK	23434	1	260	\$30.99	\$37.91	\$5.00
							8	2,080	\$247.94	\$303.26	\$40.00



# Example: Sales Data Warehouse Example(3)

Solicitor Sales					
Sales Type	Sales Channel	Sales Agent	Number Of Sales	Cost Per Unit	Retention
Carrier Sales					
	Carrier Sales				
		Carrier Sales	2	\$0.00	50.0%
DM Sales					
	DM Sales				
		DM Sales	124	\$0.13	13.7%
Non-Solicited					
	Other				
		2005 THISISHAMPRDS FREE2WKSAMP	8	\$0.00	12.5%
		AD CONTRACT START	8	\$0.17	62.5%
		ALLCONNECT	1,914	\$0.04	43.5%
		CAN'T AFFORD DM "2 WKS FREE"	7	\$0.10	71.4%
		COLLECTIONS TEAM	15	\$0.14	73.3%
		COOLSAVINGS.COM	9,415	\$21.55	7.1%
		CUSTOMER SERVICE PROMO STARTS	1,934	\$0.15	73.7%
		DATA ENTRY STARTS	49	\$0.14	59.2%



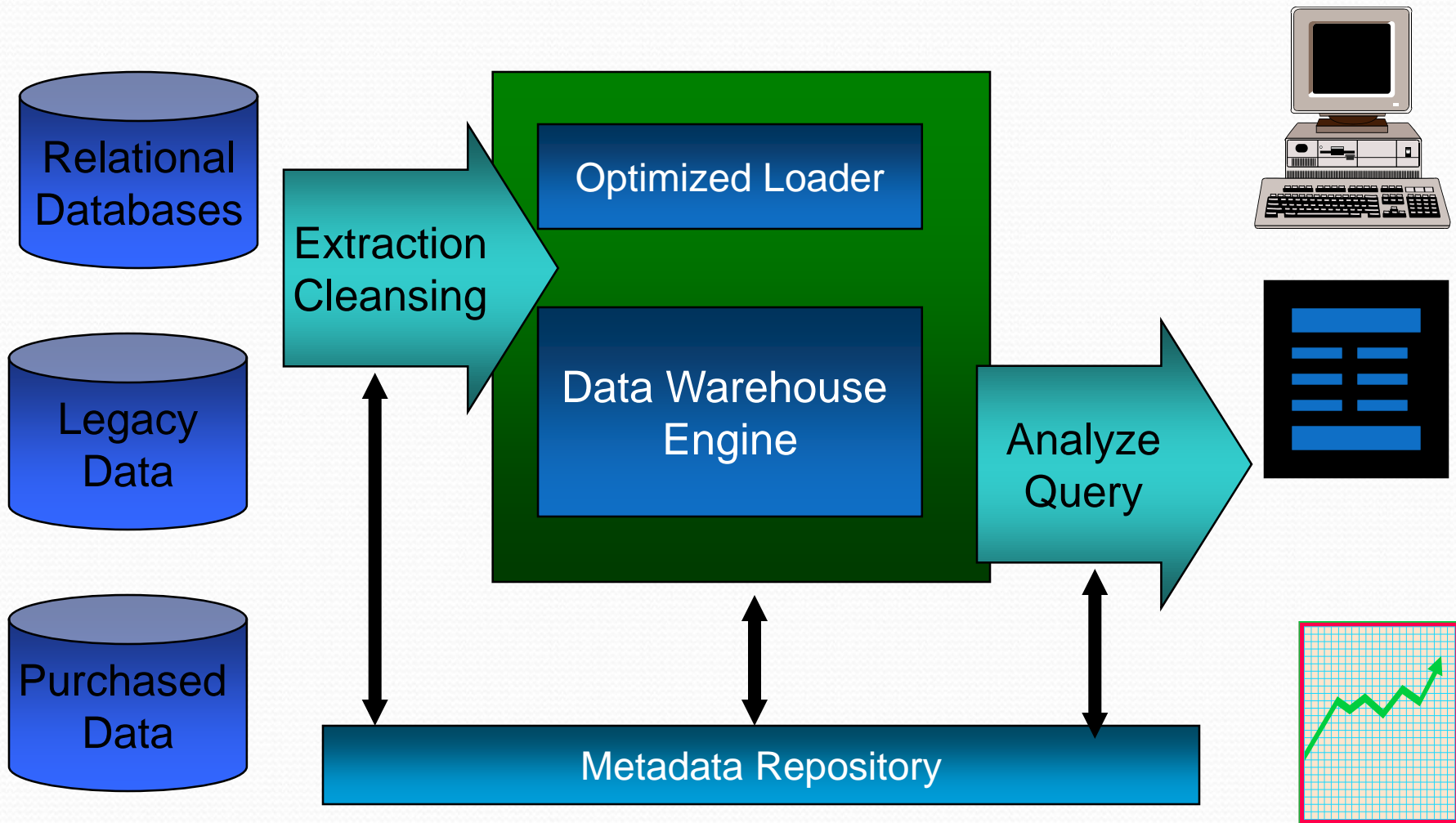


# Characteristics of a DW

- **Subject-Oriented:** A data warehouse can be used to analyze a particular subject area. For example, "sales" can be a particular subject.
- **Integrated:** A data warehouse integrates data from multiple data sources.
  - For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.
- **Time-Variant:** Historical data is kept in a data warehouse.
  - For example, one can retrieve data from 3 months, 6 months, 12 months, or even older data from a data warehouse.
  - This contrasts with a transactions system, where often only the most recent data is kept. For example, a transaction system may hold the most recent address of a customer, where a data warehouse can hold all addresses associated with a customer.
- **Non-volatile:** Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.



## II- Architecture of a DW





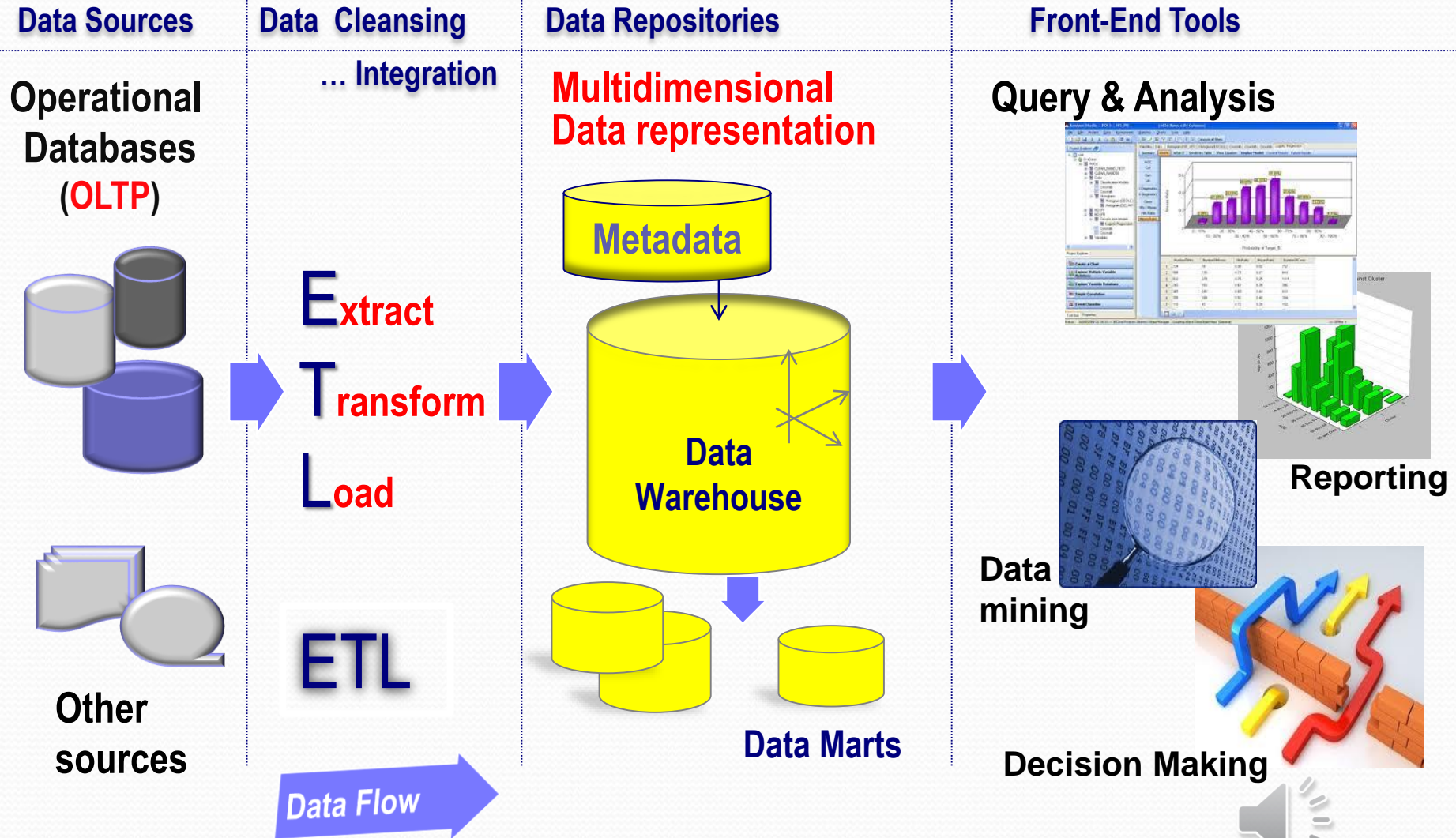
# III- Lifecycle

- **Analysis**
  - Business requirement gathering - Meet business user's requirement.
  - Identify the OLTP Systems (Oracle, MS-SQL, etc.) / Databases
  - Initial data gathering - Identify main subjects
- **Designing Data warehouse**
  - Data Modeling and design (star schema)
  - Implementing Data warehouse – choosing platform, creating data warehouse
- **Importing data:** Loading & testing - ETL phase
  - Designing and development of ETL jobs / verification
  - Run ETL jobs - Data load into Data warehouse
  - Verification & Testing
- **Install front-end tools:**
  - Requirement gathering for BI reports (Front End)
  - Building (or purchasing) a front-end reporting tool.





# DW Multi-Tiered Architecture



# Key Terms

- Data warehouse
- Dimensional model
- Normalized model
- Relational database
- OLAP (Online Analytical Processing)
- OLTP (Online Transaction Processing)
- ETL (extraction, transformation, load)
- Business Intelligence (BI) application





## III.1- Analysis

- Identify:
  - Target Questions
  - Data needs
  - Timeliness of data
  - Granularity
- Create an enterprise-level data dictionary
- Dimensional analysis: Identify facts and dimensions

