Business Intelligence and Data Analytics (IS 350)

Lecture 3

Business Intelligence and data warehousing

Dr. Abdalla sayed

2023 - 2024

What is Data warehouse (DW)

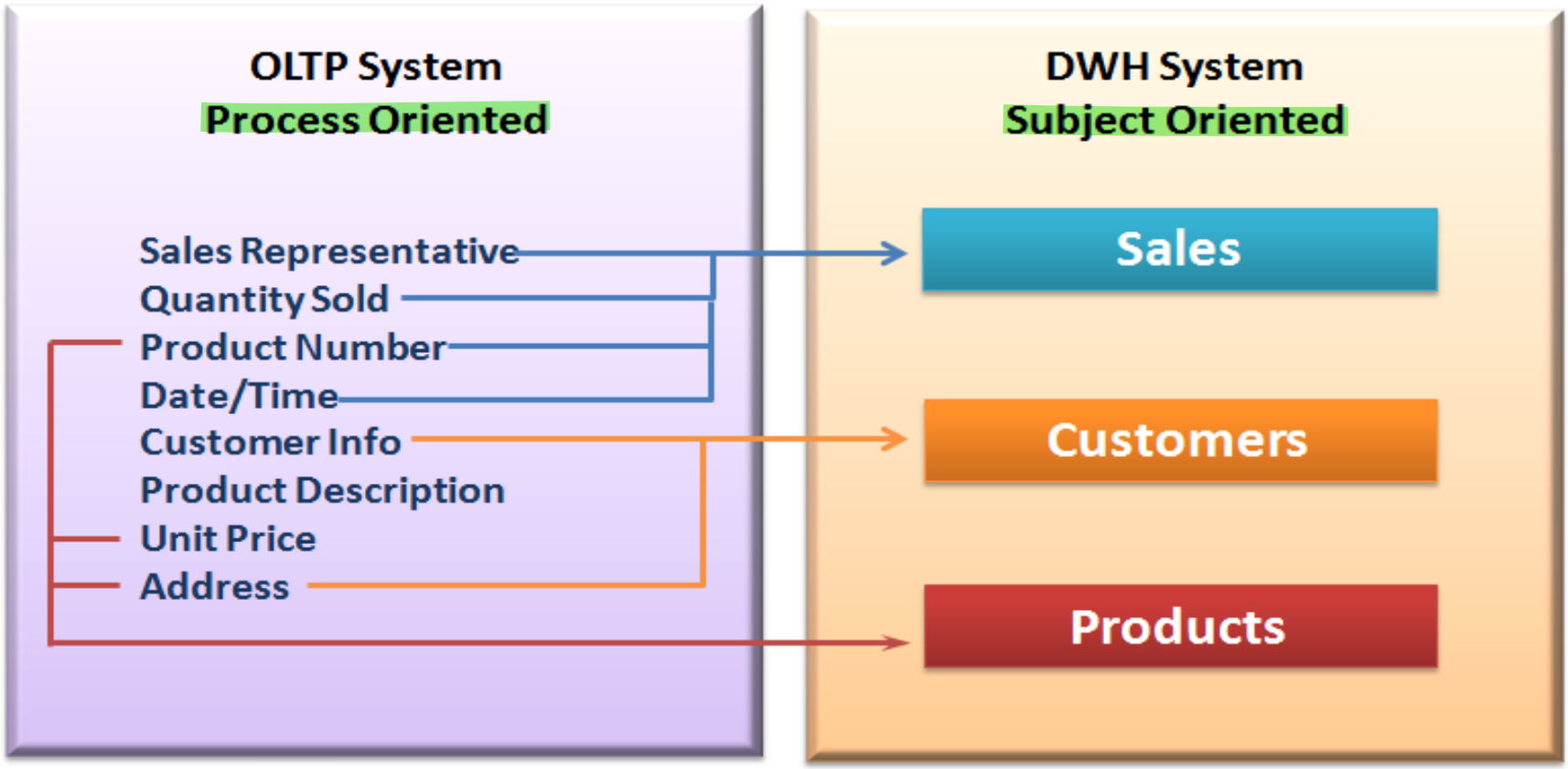
- **Data warehouse** is a pool of data produced to support decision making; it is also a repository of **current** and **historical** data of potential interest to managers throughout the organization.
- Data are usually structured to be available in a form ready for analytical processing activities (i.e., online analytical processing [OLAP], data mining, querying, reporting, and other decision support applications).

Data warehouse characteristics

- A data warehouse is a **subject-oriented**, **integrated**, **time-variant**, **nonvolatile** collection of data in support of management's decision-making process (Inmon, 2005).
- **subject-oriented:**
  - Data are organized by detailed subject, such as sales, products, or customers, containing only information relevant for decision support. Subject orientation enables users to determine not only how their business is performing, but also why.
  - A data warehouse differs from an operational database in that most operational databases have a product orientation and are tuned to handle transactions that update the database. Subject orientation provides a more comprehensive view of the organization.

Data warehouse characteristics

- **subject-oriented:**



Data warehouse characteristics

- **Integrated:**
  - Integration is closely related to subject orientation. Data warehouses must place data from different **sources** into a **consistent** format.
  - To do so, they must deal with naming conflicts and discrepancies among units of measure. A data warehouse is presumed to be totally integrated.

Data warehouse characteristics

- **Time variant:**
  - A warehouse maintains historical data. The data do not necessarily provide current status (except in real-time systems). They detect trends, deviations, and long-term relationships for forecasting and comparisons, leading to decision making.
  - Every data warehouse has a temporal quality. Time is the one important dimension that all data warehouses must support.
  - Data for analysis from multiple sources contain multiple time points (e.g., daily, weekly, monthly views).

Data warehouse characteristics

- **Nonvolatile :**
  - After data are entered into a data warehouse, users cannot change or update the data. Obsolete data are discarded, and changes are recorded as new data.
  - Data is loaded as snapshots : when changes occurs , a new snapshot is created .
  - The result : Historical record of data is kept in the data warehouse.

Data warehouse characteristics

- **Some additional characteristics may include the following :**
  - **Web based**: Data warehouses are typically designed to provide an efficient computing environment for web application.
  - **Relational/multidimensional :**A data warehouse uses either a relational structure.
  - **Client/server** : A data warehouse uses the client/server architecture to provide easy access for end users.
  - **Real time** : Newer data warehouses provide real-time, or active, data-access and analysis capabilities
  - **Metadata** : A data warehouse contains metadata (data about data) about how the data are organized and how to effectively use them.

Data warehousing

- Whereas data warehouse is a repository of data, **data warehousing** is literally the entire process.
- **Data warehousing** is a discipline that results in applications that provide decision support capability, allows ready access to business information, and creates business insight.

Data warehousing main types

- The three main types of data warehouses are :
  - Data marts (DMs)
  - operational data stores (ODS)
  - enterprise data warehouses (EDW)

Data Mart

- Whereas a data warehouse combines databases across an entire enterprise, a data mart (DM) is usually smaller and focuses on a particular subject or department.
- A DM is a subset of a data warehouse, typically consisting of a single subject area (e.g., marketing, operations).
- A DM can be either **dependent** or **independent.**

Data Mart

- **A dependent data mart** is a subset that is created directly from the data warehouse. It has the advantages of using a consistent data model and providing quality data. Dependent DMs support the concept of a single enterprise-wide data model, but the data warehouse must be constructed first.
- A dependent DM ensures that the end user is viewing the same version of the data that is accessed by all other data warehouse users. The high cost of data warehouses limits their use to large companies.

Data Mart

- Many firms use a lower-cost, scaled-down version of a data warehouse referred to as an **independent** DM subset.
- **independent data mart** is a small warehouse designed for a strategic business unit or a department, but its source is not an **EDW**.

operational data stores (ODS)

- **Operational data store** (ODS) provides a fairly recent form of customer information file.
- This type of database is often used as an interim **staging** area for a data warehouse.
- Unlike the static contents of a data warehouse, the contents of an ODS are updated throughout the course of business operations.
- An ODS is used for short-term decisions involving mission-critical applications rather than for the medium- and long-term decisions associated with an **EDW**.

operational data stores (ODS)

- An ODS is similar to short-term memory in that it stores only very recent information. In comparison, a data warehouse is like long-term memory because it stores permanent information.
- An ODS consolidates data from multiple source systems and provides a near–real-time, integrated view of volatile, current data.
- The exchange, transfer, and load (**ETL**) processes for an ODS are identical to those for a data warehouse.

Enterprise data warehouses (EDW)

- An **enterprise data warehouse (EDW)** is a large-scale data warehouse that is used across the enterprise for decision support.
- The large-scale nature of an EDW provides integration of data from many sources into a standard format for effective BI and decision support applications.
- EDWs are used to provide data for many types of decision support systems (DSS), including customer relationship management (CRM), supply chain management (SCM), business performance management (BPM), business activity monitoring, product life cycle management, revenue management, and sometimes even knowledge management systems.

Metadata

- Metadata are data about data
- Metadata describe the structure of and some meaning about data, thereby contributing to their effective or ineffective use.
- Patterns are another way to view metadata. According to the pattern view, we can differentiate between syntactic metadata (i.e., data describing the syntax of data), structural metadata (i.e., data describing the structure of the data), and semantic metadata (i.e., data describing the meaning of the data in a specific domain).

Data warehouses process

- Many organizations need to create data warehouses—massive data stores of time series data for decision support.
- Data are imported from various external and internal resources and are **cleansed** and **organized** in a manner consistent with the organization's needs.
- After the data are populated in the data warehouse, DMs can be loaded for a specific area or department.
- Alternatively, DMs can be created first, as needed, and then integrated into an EDW.

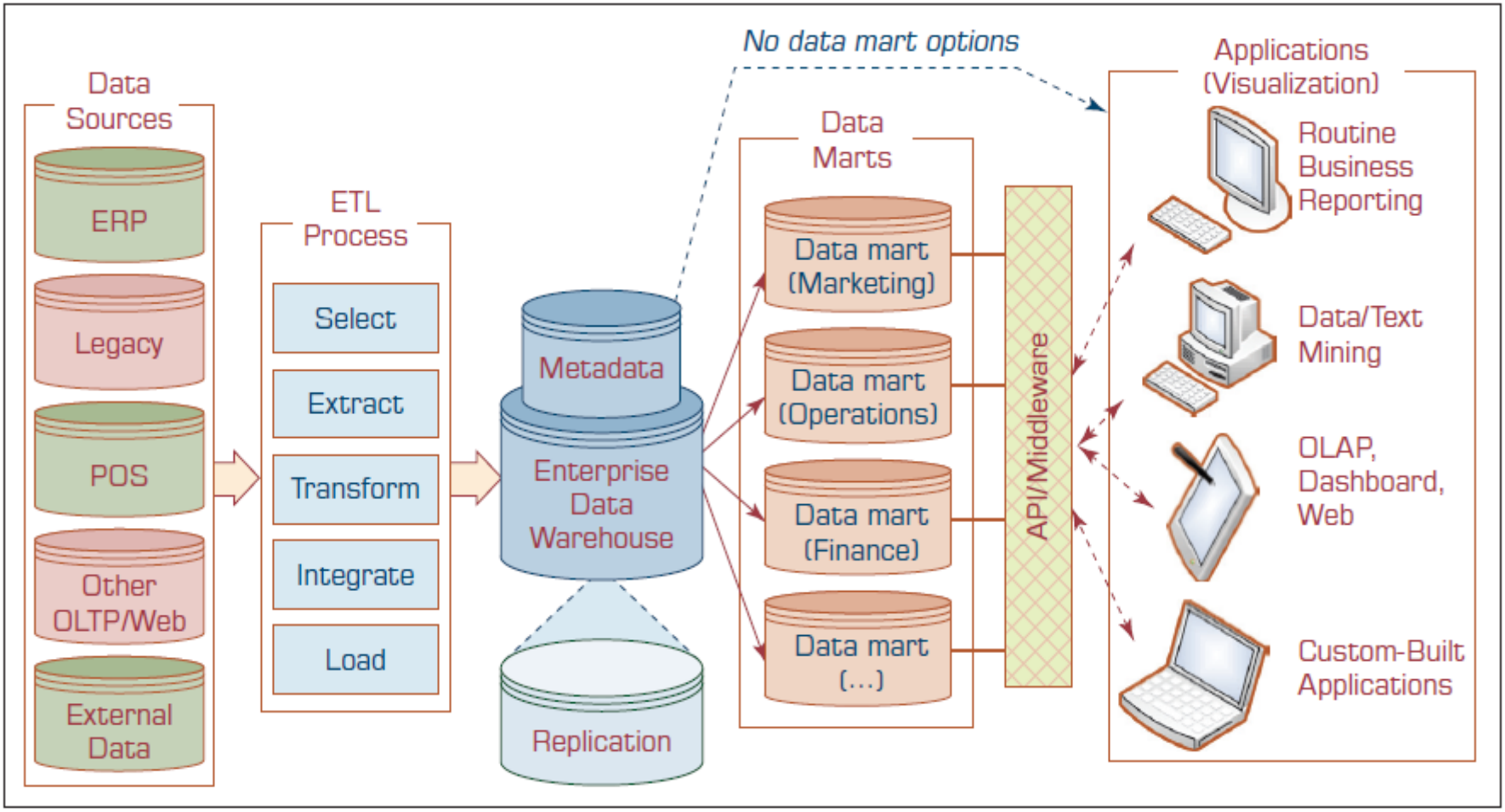
Data warehouses components

- **Data sources:** Data are sourced from multiple independent operational "legacy" systems and possibly from external data providers. Data may also come from an OLTP or enterprise resource planning (ERP) system. Web data in the form of Web logs may also feed to a data warehouse.
- **Data extraction and transformation:** Data are extracted and properly transformed using custom-written or commercial software called **ETL**.
- **Data loading:** Data are loaded into a staging area, where they are transformed and cleansed. The data are then ready to load into the data warehouse and/or DMs.
- **Comprehensive database:** Essentially, this is the EDW to support all decision analysis by providing relevant summarized and detailed information originating from many different sources.

Data warehouses components

- **Metadata**: Metadata are maintained so that they can be assessed by IT personnel and users. Metadata include software programs about data and rules for organizing data summaries that are easy to index and search, especially with Web tools.
- **Middleware tools**: Middleware tools enable access to the data warehouse. Power users such as analysts may write their own SQL queries. Others may employ a managed query environment, such as Business Objects, to access data. There are many front-end applications that business users can use to interact with data stored in the data repositories, including data mining, OLAP, reporting tools, and data visualization tools.

Data warehouses components



ETL

- **Data integration** comprises three major processes that, when correctly implemented, permit data to be accessed and made accessible to an array of ETL and analysis tools and the data warehousing environment: data access (i.e., the ability to access and extract data from any data source), data federation (i.e., the integration of business views across multiple data stores), and change capture (based on the identification, capture, and delivery of the changes made to enterprise data sources).
- A major purpose of a data warehouse is to integrate data from multiple systems.

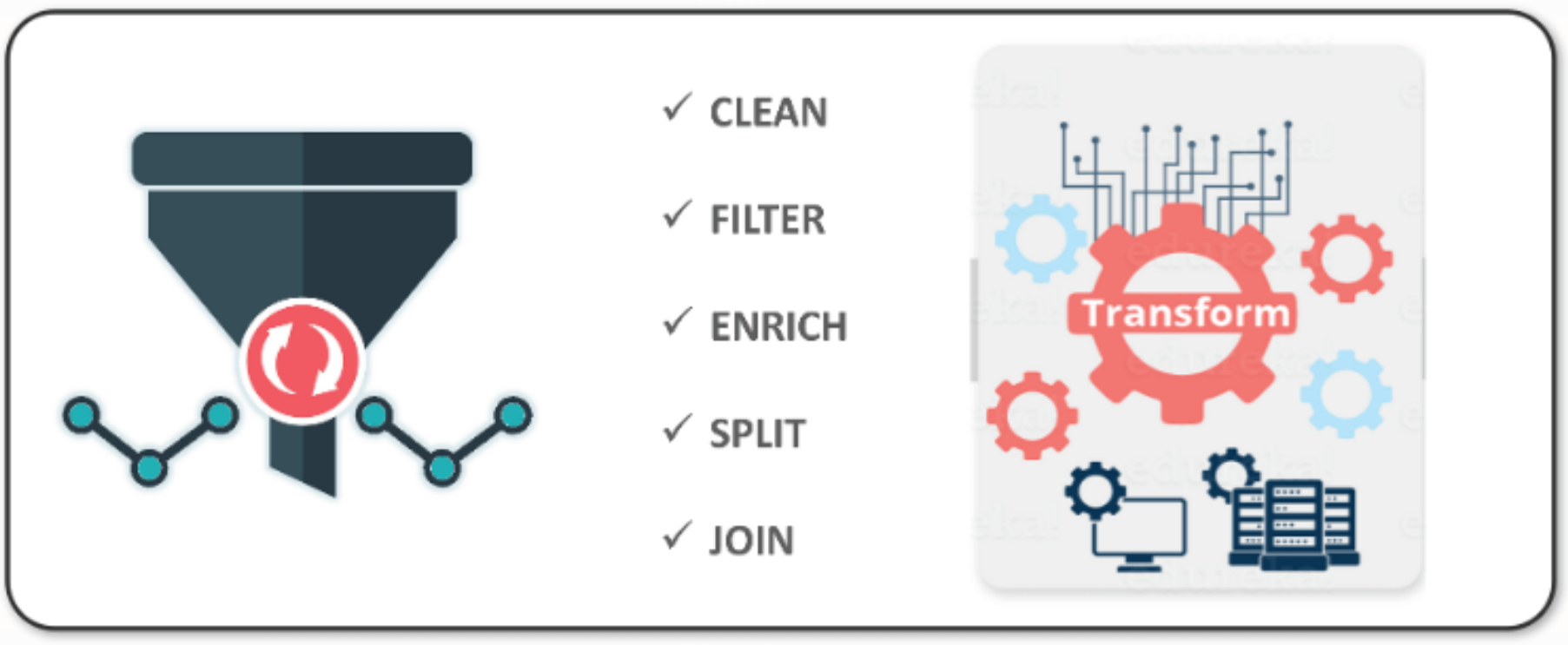
ETL

- **Extraction, Transformation, and Load**
  - The **heart** of the technical side of the data warehousing process is extraction, transformation, and load (ETL).
  - The ETL process is an **integral** component in any data-centric project.
  - **Extraction** : is the process of extracting (reading) data from various homogenous and heterogeneous data sources.
  - **Transformation** : converting the extracted data from its previous form into the form in which it needs to be so that it can be placed into a data warehouse or simply another database.

ETL

- **Extraction, Transformation, and Load**

Transformation: In transformation, entire data is analyzed and various functions are applied on it in order to load the data to the target database in a cleaned and general format.



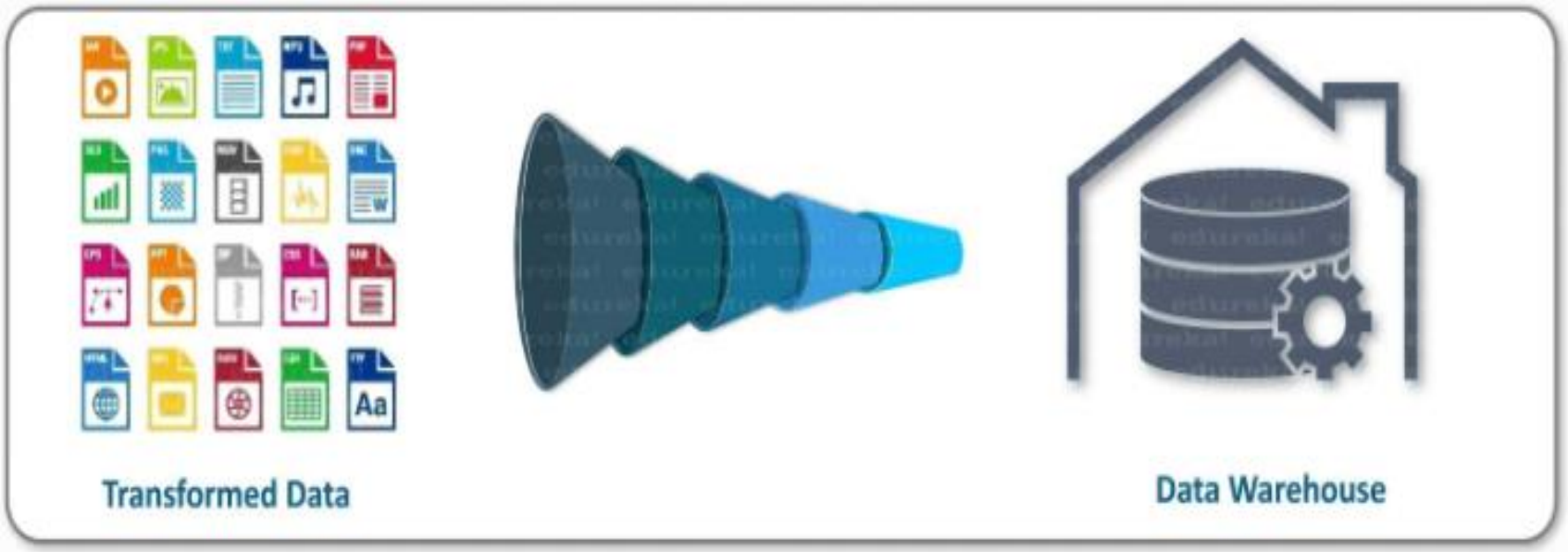
ETL

- **Extraction, Transformation, and Load**
  - **Transformation** occurs by using rules or lookup tables or by combining the data with other data.
  - **Load** : putting the data into the data warehouse.
  - The three database functions are integrated into one tool to pull data out of one or more databases and place them into another, consolidated database or a data warehouse.

ETL

- **Extraction, Transformation, and Load**

Load: Loading is the process of loading the processed data to a target data repository using minimal resources.



ETL tools


IBM Infosphere Information Server
Fast track your information
edureka!
talend
6.4
Talend Open Studio
for Data Integration
ORACLE
DATA INTEGRATOR
sas
Data Integration
informatica
POWERCENTER
Microsoft
SQL Server
INTEGRATION SERVICES
SAP
Business Objects

Dimensional modeling

- Like enterprise relationship (ER) modeling, **dimensional modeling** is a logical design technique**.**
- Dimensional modeling is much better suited for business intelligence (BI) applications and data warehousing (DW).
- It depicts business processes throughout an enterprise and organizes that data and its structure in a logical way.
- The purpose of dimensional modeling is to enable BI reporting, query, and analysis.
- The key concepts in dimensional modeling are **facts**, **dimensions**, and **attributes**. There are different types of facts, depending on whether they can be added together.

Dimensional modeling

- Dimensions can have different hierarchies, and have attributes that define the who, what, where, and why of the dimensional model. The grain, or level of **granularity**, is another key concept with dimensional modeling, as it determines the level of detail.
- Facts, dimensions, and attributes can be organized in several ways, called **schemas**. The choice of schema depends on variables such as the **type** of reporting that the model needs to facilitate and the type of BI tool being used.
- Building a dimensional model includes additional puzzle pieces such as calendar and time dimensions, and more complicated pieces such as degenerative dimensions and consolidated fact tables.

Dimensional modeling

- **Facts**
  - A **fact** is a measurement of a business activity, such as a business event or transaction, and is generally numeric. Examples of facts are sales, expenses, and inventory levels. Numeric measurements may include counts, dollar amounts, percentages, or ratios.
  - A **fact** is a collection of related data consisting of **measures**.
  - In a data warehouse, facts are implemented in the core tables in which all of the numeric data is stored.
  - Facts can be **aggregated** or **derived**. For example, you can sum up the total revenue or calculate the profitability of a set of sales transactions.

Dimensional modeling

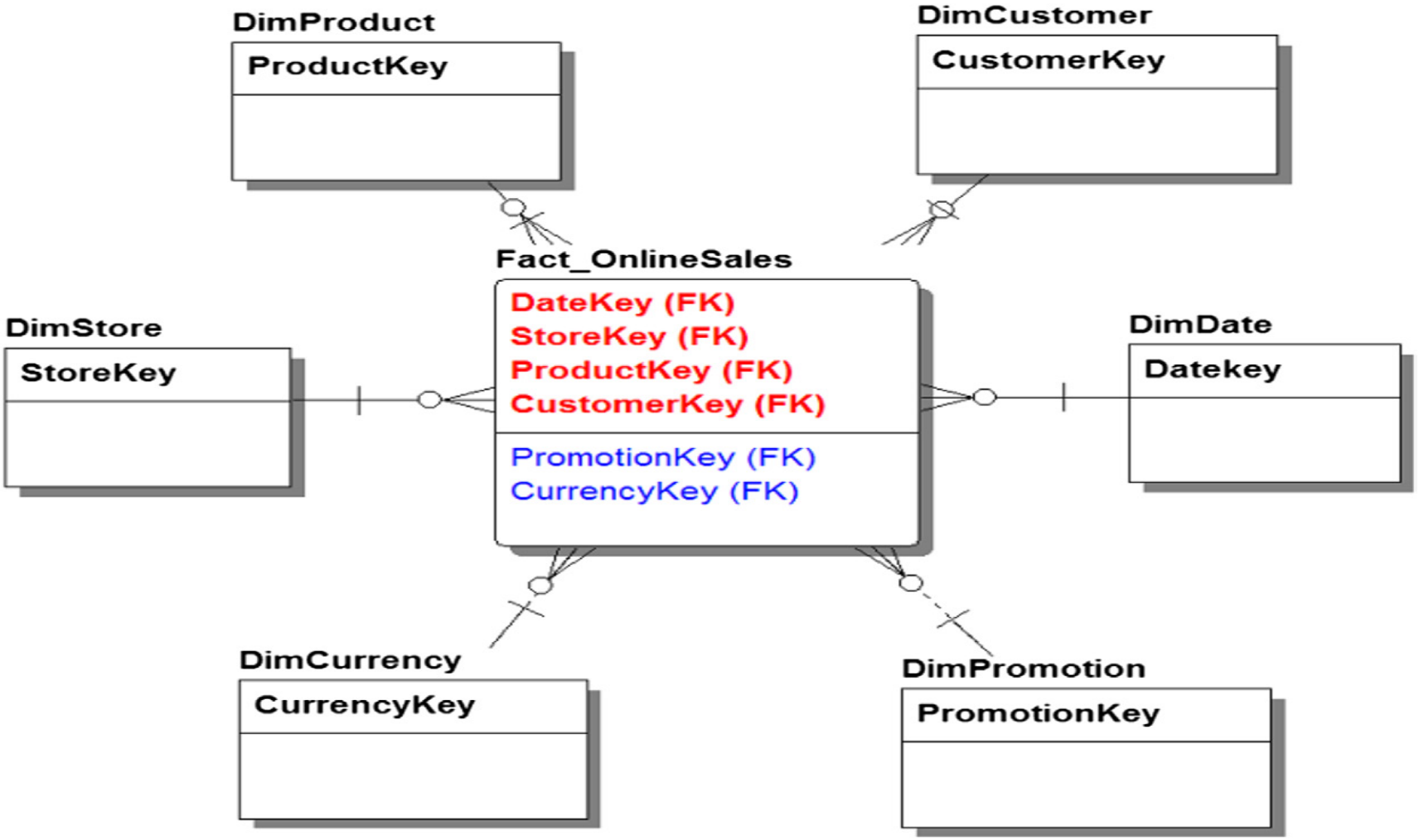
- **Facts**
  - A **fact** is a measurement of Facts provide the measurements of how well or how poorly the business is performing.
  - A fact is also referred to as organizational performance measure.
  - Fact contains **redundancy**.
  - **Ninety-percent** of the data in a dimensional model is typically located in the fact tables.
  - The key dimensional modeling design concerns when working with the data in fact tables are how to minimize and standardize it and make it consistent.

Dimensional modeling

- **Facts**
  - Fact tables are composed of two types of columns: **keys and measures.**
  - The first, the key column, consists of a group **of foreign keys (FK)** that point to the primary keys of dimensional tables that are associated with this fact table to enable business analysis.
  - The relationships between fact tables and the dimensions are one-to-many

Dimensional modeling

- **Facts**



Dimensional modeling

- **Facts**
  - The second type of column in a fact table is the actual **measures** of the business activity such as the **sales revenue and order quantity**.
  - Every measurement has a grain, which is the level of detail in the measurement of an event such as a unit of measure, currency used, or ending daily balance of an account.
  - For example SalesQuantity, SalesAmount, ReturnAmount, ReturnQuantity, DiscountAmount, DiscountQuantity, and TotalCost that apply to a customer for a product purchased at a specific time.
  - All of these measures are related to the business event (the sale) that the fact represents and they have a level of granularity related to that event.

Dimensional modeling

- **Facts**

.

Measures in fact tables

Dimensional modeling

- **Dimensions**
  - A dimension is an entity that establishes the business context for the measures (facts) used by an enterprise.
  - Dimensions define the who, what, where, and why of the dimensional model, and group similar attributes into a category or subject area. Examples of dimensions are product, geography, customers, employees, and time.
  - Whereas facts are **numeric**, dimensions are **descriptive** in nature (although some of those descriptions, such as a product list price, may be numeric).

Dimensional modeling

- **Dimensions**