

Machine Learning and Data Mining

Last update: 08 October 2023

Academic Year 2023 – 2024
Alma Mater Studiorum · University of Bologna

Contents

1	Introduction	2
1.1	Data	2
1.1.1	Data sources	2
1.1.2	Software	2
1.1.3	Insight	2
2	Business Intelligence	4
2.1	Online Analytical Processing (Online Analytical Processing (OLAP))	4
2.1.1	Operators	4
2.2	Extraction, Transformation, Loading (Extraction, Transformation, Loading (ETL))	5
2.2.1	Extraction	5
2.2.2	Cleaning	5
2.2.3	Transformation	6
2.2.4	Loading	6
2.3	Data warehouse architectures	6
2.3.1	Single-layer architecture	7
2.3.2	Two-layer architecture	7
2.3.3	Three-layer architecture	7
2.4	Conceptual modeling	8
2.4.1	Aggregation operators	9
2.4.2	Logical design	9

Acronyms

BI Business Intelligence

DFM Dimensional Fact Model

DM Data Mart

DSS Decision Support System

DWH Data Warehouse

EIS Executive Information System

ERP Enterprise Resource Planning

ETL Extraction, Transformation, Loading

MIS Management Information System

OLAP Online Analytical Processing

OLTP Online Transaction Processing

1 Introduction

1.1 Data

Data Collection of raw values.

Data

Information Organized data (e.g. relationships, context, ...).

Information

Knowledge Understanding information.

Knowledge

1.1.1 Data sources

Transaction Business event that generates or modifies data in an information system (e.g. database).

Transaction

Signal Measure produced by a sensor.

Signal

External subjects

1.1.2 Software

Online Transaction Processing (OLTP) Class of programs to support transaction oriented applications and data storage. Suitable for real-time applications.

Online Transaction Processing

Enterprise Resource Planning (ERP) Integrated system to manage all the processes of a business. Uses a shared database for all applications. Suitable for real-time applications.

Enterprise Resource Planning

1.1.3 Insight

Decision can be classified as:

Structured Established and well understood situations. What is needed is known.

Structured decision

Unstructured Unplanned and unclear situations. What is needed for the decision is unknown.

Unstructured decision

Different levels of insight can be extracted by:

Management Information System (MIS) Standardized reporting system built on existing OLTP. Used for structured decisions.

Management Information System

Decision Support System (DSS) Analytical system to provide support for unstructured decisions.

Decision Support System

Executive Information System (EIS) Formulate high level decisions that impact the organization.

Executive Information System

Online Analytical Processing (OLAP) Grouped analysis of multidimensional data. Involves large amount of data.

Online Analytical Processing

Business Intelligence (BI) Applications, infrastructure, tools and best practices to analyze information. Business Intelligence

Big data Large and/or complex and/or fast changing collection of data that traditional DBMSs are unable to process. Big data

Structured e.g. relational tables.

Unstructured e.g. videos.

Semi-structured e.g. JSON.

Anaylitics Structured decision driven by data. Anaylitics

Data mining Discovery process for unstructured decisions. Data mining

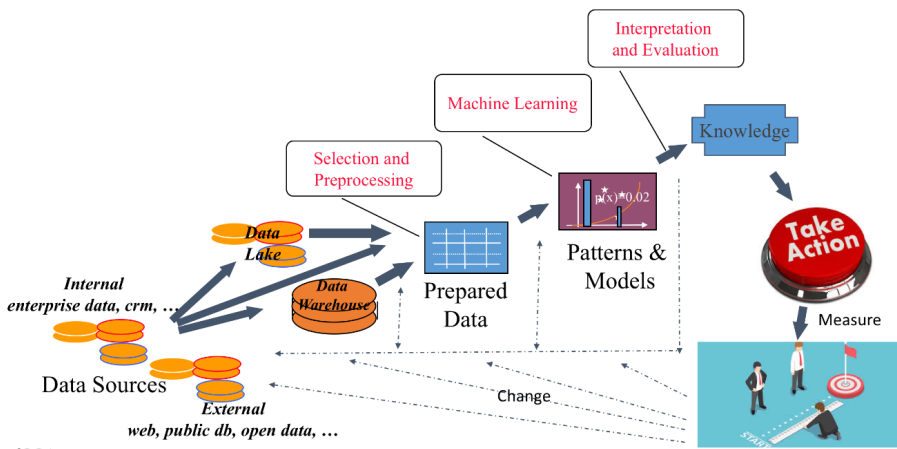


Figure 1.1: Data mining process

Machine learning Learning models and algorithms that allow to extract patterns from data. Machine learning

2 Business Intelligence

Business Intelligence Transform raw data into information. Deliver the right information to the right people at the right time through the right channel. Business Intelligence

Data Warehouse (DWH) Optimized repository that stores information for decision making processes. DWHs are a specific type of DSS. Data Warehouse

Features:

- Subject-oriented: focused on enterprise specific concepts.
- Integrates data from different sources and provides an unified view.
- Non-volatile storage with change tracking.

Data Mart (DM) Subset of the primary DWH with information relevant to a specific business area. Data Mart

2.1 Online Analytical Processing (OLAP)

OLAP analyses Able to interactively navigate the information in a data warehouse. Allows to visualize different levels of aggregation. Online Analytical Processing (OLAP)

OLAP session Navigation path created by the operations that a user applied.

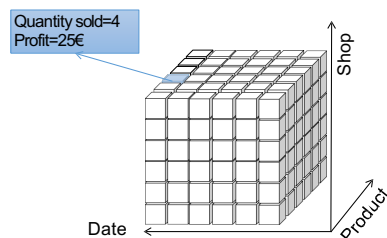
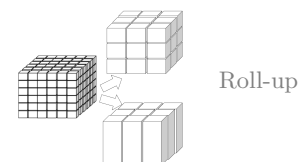


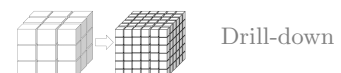
Figure 2.1: OLAP data cube

2.1.1 Operators

Roll-up Increases the level of aggregation (i.e. GROUP BY in SQL). Some details are collapsed together.



Drill-down Reduces the level of aggregation. Some details are reintroduced.



The diagram illustrates the process of dividing a large cube into smaller units. On the left is a large cube composed of a 10x10x10 grid of smaller cubes. An arrow points from this large cube to a single 10x10 slice, which is labeled "slice". Another arrow points from the slice to a single 1x10 die, which is labeled "dice". This visualizes how a large volume is partitioned into smaller, measurable units.

The dice operator reduces the number of data being analyzed (i.e. LIMIT in SQL).

Changes the layout of the data, to analyze it from a different viewpoint.

Drill-through

Order ID	Order Date	Ship Date	Ship Mode	Customer Name	Segment	City	State	Country
17-1219-128300	12/19/2017	12/20/2017	Same Day	Gunga Industries	Corporate	Mumbai	India	France
18-1214-128297	12/14/2018	12/20/2018	Second Class	Sara Garcia	Corporate	Houston	Texas	France
18-1214-128301	12/14/2018	12/20/2018	Standard Class	Vanessa	Corporate	Washington	District of Columbia	Germany
18-1214-128300	12/14/2018	12/20/2018	Standard Class	Vanessa	Corporate	Houston	Texas	Germany
18-1214-128302	12/14/2018	12/20/2018	Standard Class	Vanessa	Corporate	Houston	Texas	Germany
18-1214-128303	12/14/2018	12/20/2018	Standard Class	Vanessa Garcia	Corporate	Waco	Texas	Germany
18-1214-128302	12/14/2018	12/20/2018	Standard Class	Vanessa Garcia	Corporate	Waco	Texas	Germany
18-1214-128302	12/14/2018	12/20/2018	Standard Class	Vanessa Garcia	Corporate	Waco	Texas	Germany
18-1214-128302	12/14/2018	12/20/2018	Standard Class	Vanessa Garcia	Corporate	Waco	Texas	Germany

The ETL process extracts, integrates and cleans operational data that will be loaded into a data warehouse.

Extraction,
Transformation,
Loading (ETL)

Extracted operational data can be:

Strucured data

Unstructured data

Static The entirety of the operational data are extracted to populate the data warehouse for the first time.

Static extraction

Incremental extraction

Operational data may contain:

Missing data

Wrong values (e.g. 30th of February)

5

Typos

Methods to clean and increase the quality of the data are:

Dictionary-based techniques	Lookup tables to substitute abbreviations, synonyms or typos. Applicable if the domain is known and limited.	Dictionary-based cleaning
Approximate merging	Merging data that do not have a common key.	Approximate merging
Approximate join	Use non-key attributes to join two tables (e.g. using the name and surname instead of a unique identifier).	
Similarity approach	Use similarity functions (e.g. edit distance) to merge multiple instances of the same information (e.g. typo in customer surname).	
Ad-hoc algorithms		Ad-hoc algorithms

2.2.3 Transformation

Data are transformed to respect the format of the data warehouse:

Conversion	Modifications of types and formats (e.g. date format)	Conversion
Enrichment	Creating new information by using existing attributes (e.g. compute profit from receipts and expenses)	Enrichment
Separation and concatenation	Denormalization of the data: introduces redundances (i.e. breaks normal form ¹) to speed up operations.	Separation and concatenation

2.2.4 Loading

Adding data into a data warehouse:

Refresh	The entire DWH is rewritten.	Refresh loading
Update	Only the changes are added to the DWH. Old data are not modified.	Update loading

2.3 Data warehouse architectures

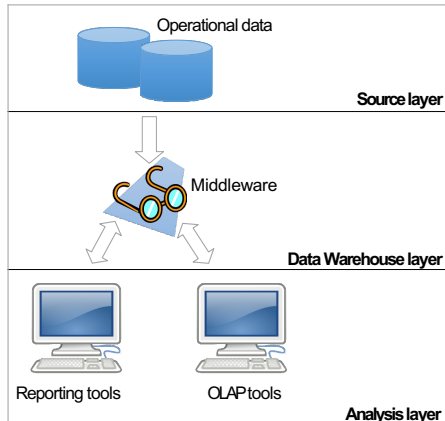
The architecture of a data warehouse should meet the following requirements:

Separation	Separate the analytical and transactional workflows.
Scalability	Hardware and software should be easily upgradable.
Extensibility	Capability to host new applications and technologies without the need to redesign the system.
Security	Access control.
Administrability	Easily manageable.

¹https://en.wikipedia.org/wiki/Database_normalization

2.3.1 Single-layer architecture

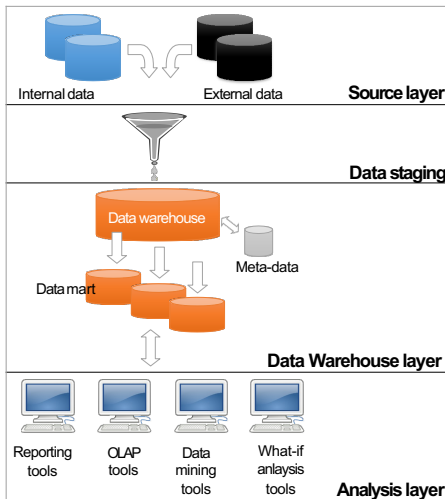
- Minimizes the amount of data stored (i.e. no redundancies).
- The source layer is the only physical layer (i.e. no separation).
- A middleware provides the DWH features.



Single-layer architecture

2.3.2 Two-layer architecture

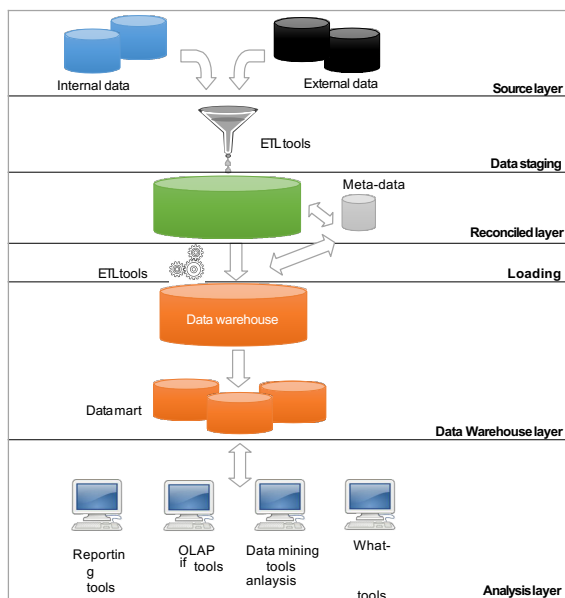
- Source data (source layer) are physically separated from the DWH (data warehouse layer).
- A staging layer applies ETL procedures before populating the DWH.
- The DWH is a centralized repository from which data marts can be created. Metadata repositories store information on sources, staging and data marts schematics.



Two-layer architecture

2.3.3 Three-layer architecture

- A reconciled layer enhances the cleaned data coming from the staging step by adding enterprise-level details (i.e. adds more redundancy before populating the DWH).



Three-layer architecture

2.4 Conceptual modeling

Dimensional Fact Model (DFM) Conceptual model to support the design of data marts.

The main concepts are:

Fact Concept relevant to decision-making processes (e.g. sales).

Measure Numerical property to describe a fact (e.g. profit).

Dimension Property of a fact with a finite domain (e.g. date).

Dimensional attribute Property of a dimension (e.g. month).

Hierarchy A tree where the root is a dimension and nodes are dimensional attributes (e.g. date \rightarrow month).

Primary event Occurrence of a fact. It is described by a tuple with a value for each dimension and each measure.

Secondary event Aggregation of primary events. Measures of primary events are aggregated if they have the same (preselected) dimensional attributes.

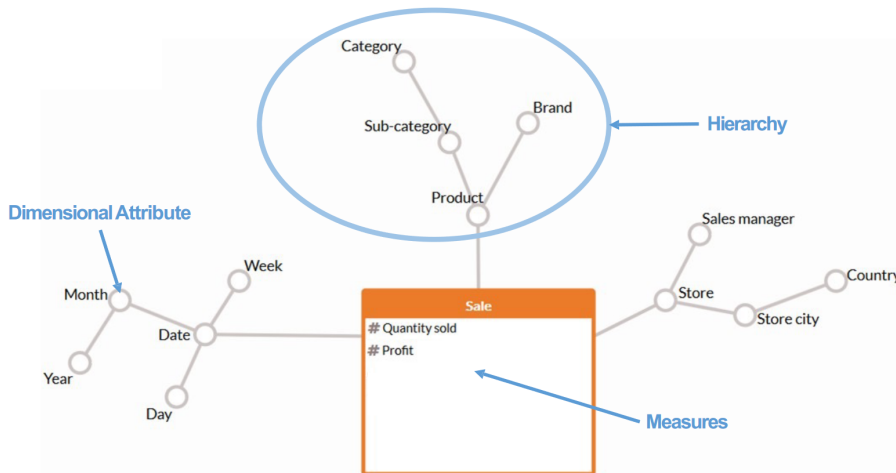


Figure 2.2: Example of DFM

Primary events				
Date	Store	Product	Qty sold	Profit
01/03/15	Central store	Milk	20	60
01/03/15	Central store	Coke	25	50
02/03/15	Central store	Bread	40	70
10/03/15	Central store	Wine	15	150

Secondary event				
Month	Store	Category	Qty sold	Profit
March 2015	Central store	Food and Beverages	100	330

Figure 2.3: Example of primary and secondary events

2.4.1 Aggregation operators

Measures can be classified as:

- Flow measures** Evaluated cumulatively with respect to a time interval (e.g. quantity sold). Flow measures
- Level measures** Evaluated at a particular time (e.g. number of products in inventory). Level measures
- Unit measures** Evaluated at a particular time but expressed in relative terms (e.g. unit price). Unit measures

Aggregation operators can be classified as:

- Distributive** Able to calculate aggregates from partial aggregates (e.g. SUM, MIN, MAX). Distributive operators
- Algebraic** Requires a finite number of support measures to compute the result (e.g. AVG). Algebraic operators
- Holistic** Requires an infinite number of support measures to compute the result (e.g. RANK). Holistic operators
- Additivity** A measure is additive along a dimension if an aggregation operator can be applied. Additive measure

	Temporal hierarchies	Non-temporal hierarchies
Flow measures	SUM, AVG, MIN, MAX	SUM, AVG, MIN, MAX
Level measures	AVG, MIN, MAX	SUM, AVG, MIN, MAX
Unit measures	AVG, MIN, MAX	AVG, MIN, MAX

Table 2.1: Allowed operators for each measure type

2.4.2 Logical design

Defining the data structures (e.g. tables and relationships) according to a conceptual model. There are mainly two strategies:

- Star schema** A fact table that contains all the measures and linked to dimensional tables. Star schema

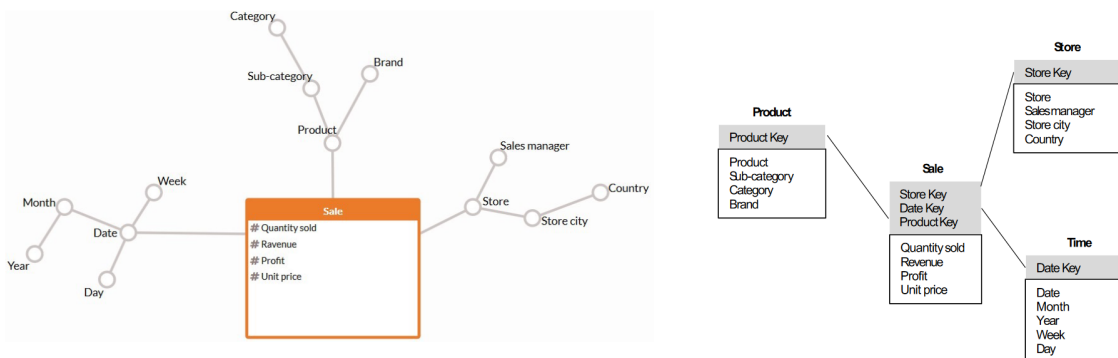


Figure 2.4: Example of star schema

- Snowflake schema** A star schema variant with partially normalized dimension tables. Snowflake schema

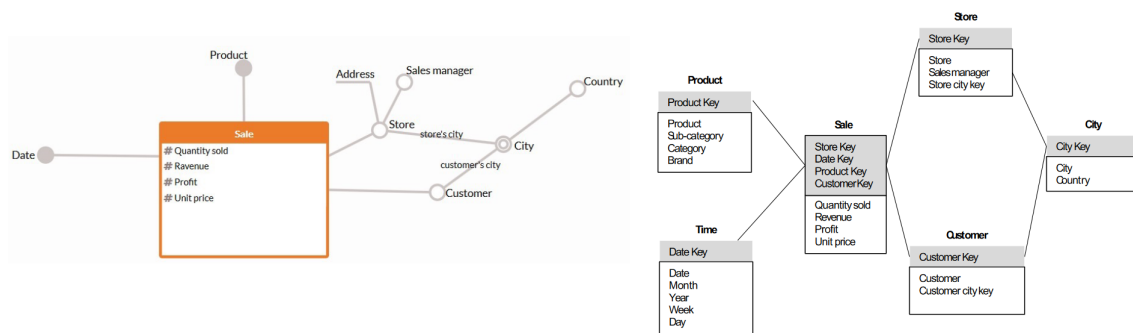


Figure 2.5: Example of snowflake schema