

# Machine Learning and Data Mining

Last update: 14 October 2023

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

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Data . . . . .	2
1.1.1	Data sources . . . . .	2
1.1.2	Software . . . . .	2
1.1.3	Insight . . . . .	2
<b>2</b>	<b>Data warehouse</b>	<b>4</b>
2.1	Online Analytical Processing (OLAP) . . . . .	4
2.1.1	Operators . . . . .	4
2.2	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
<b>3</b>	<b>Data lake</b>	<b>11</b>
3.1	Traditional vs insight-driven data systems . . . . .	11
3.2	Data architecture evolution . . . . .	11
3.3	Components . . . . .	12
3.3.1	Data ingestion . . . . .	12
3.3.2	Storage . . . . .	12
3.3.3	Processing and analytics . . . . .	13
3.4	Architectures . . . . .	13
3.4.1	Lambda lake . . . . .	13
3.4.2	Kappa lake . . . . .	13
3.4.3	Delta lake . . . . .	14
3.5	Metadata . . . . .	14
<b>4</b>	<b>CRISP-DM</b>	<b>15</b>
4.1	Business understanding . . . . .	15
4.2	Data understanding . . . . .	15
4.3	Data preparation . . . . .	15
4.4	Modelling . . . . .	16
4.5	Evaluation . . . . .	16
4.6	Deployment . . . . .	16

# Acronyms

**BI** Business Intelligence

**CDC** Change Data Capture

**CRISP-DM** Cross Industry Standard Process for Data Mining

**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

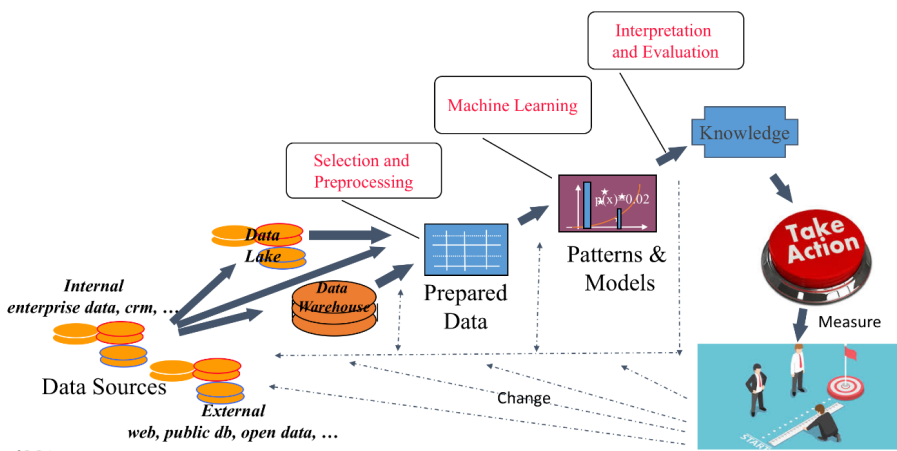


Figure 1.1: Data mining process

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

## 2 Data warehouse

**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.

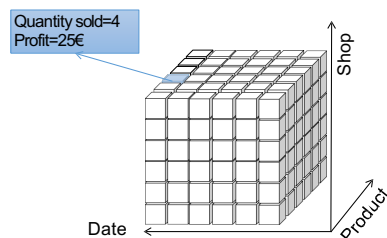
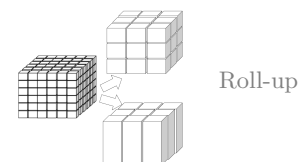


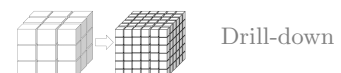
Figure 2.1: OLAP data cube

#### 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".

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.

The diagram illustrates the distributive property of multiplication. At the top, there are two separate groups of cubes. The left group consists of 3 columns, each with 4 orange cubes, totaling 12 orange cubes. The right group consists of 3 columns, each with 4 blue cubes, totaling 12 blue cubes. Below these, two yellow arrows point down to a single, larger array of cubes. This array is 3 columns wide and 8 cubes high. The cubes are arranged in alternating columns of orange and blue, with each column containing 4 cubes. This represents the sum of the two groups, 12 + 12 = 24, which is also equal to 3 times the sum of 4 and 4, or 3 times 8.

## Drill-through

Order ID	Order Date	Ship Date	Ship Mode	Customer Name	Segment	City	State	Country
17-0129-128000	15/01/2017	17/01/2017	Same Day	Gunga Sengupta	Corporate	Mumbai	India-France	France
15-0124-026907	17/01/2017	17/01/2017	Second Day	Sara Garcia	Corporate	Orlando	India-France	France
15-0124-026908	16/01/2017	16/01/2017	First Day	Vanessa	Corporate	Orlando	India-France	France
15-0124-046800	16/01/2017	16/01/2017	Standard Class	Vanessa	Corporate	Huygheboon	Germany-Japan	Germany
15-0124-046801	16/01/2017	16/01/2017	Standard Class	Vanessa	Corporate	Huygheboon	Germany-Japan	Germany
15-0124-046802	17/01/2017	17/01/2017	Standard Class	Vanessa	Corporate	Waller (Dach)	North-West-Germany	Germany
15-0124-046803	17/01/2017	17/01/2017	Standard Class	Vanessa	Corporate	Waller (Dach)	North-West-Germany	Germany
15-0124-046804	17/01/2017	17/01/2017	Standard Class	Vanessa	Corporate	Waller (Dach)	North-West-Germany	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:

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

### 2.2.3 Transformation

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

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

### 2.2.4 Loading

Adding data into a data warehouse:

<b>Refresh</b>	The entire DWH is rewritten.	Refresh loading
<b>Update</b>	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:

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

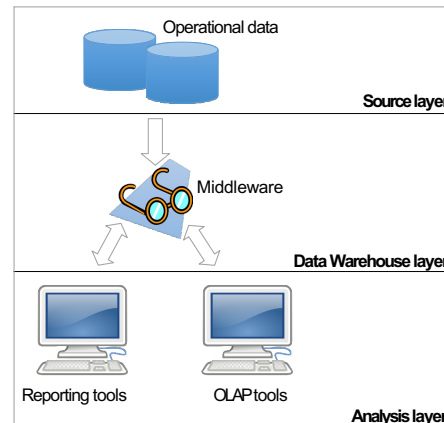
---

<sup>1</sup>[https://en.wikipedia.org/wiki/Database\\_normalization](https://en.wikipedia.org/wiki/Database_normalization)



### 2.3.1 Single-layer architecture

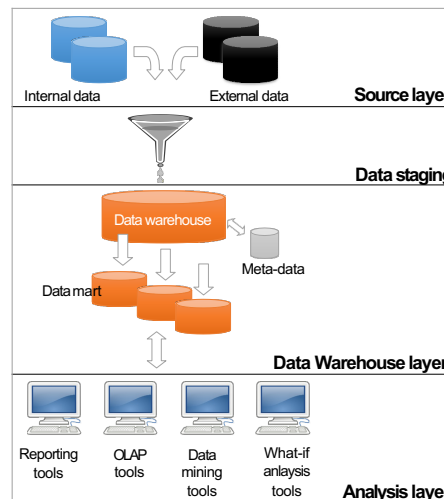
- Minimizes the amount of data stored (i.e. no redundances).
- The source layer is the only physical layer (i.e. no separation).
- A middleware provides the DWH features.



Single-layer architecture

### 2.3.2 Two-layer architecture

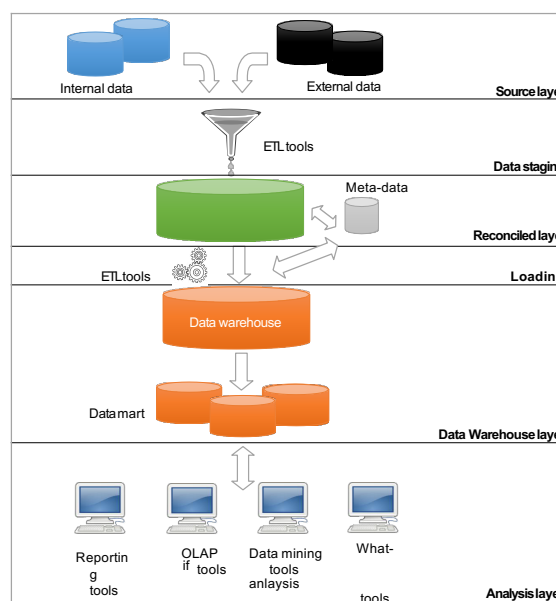
- Source data (source layer) are physically separated from the DWH (data warehouse layer).
- A staging layer applies ETL procedures before populating the DWH.
- The DWH is a centralized repository from which data marts can be created. Metadata repositories store information on sources, staging and data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

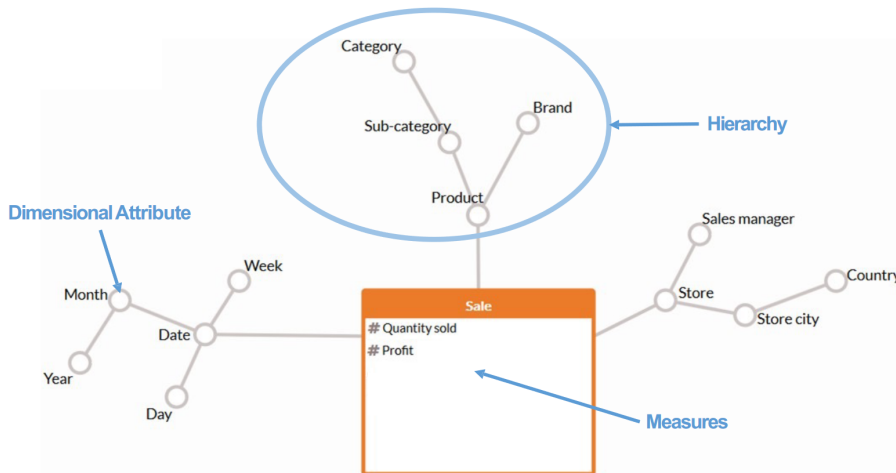


Figure 2.2: Example of DFM

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

SUM      SUM

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

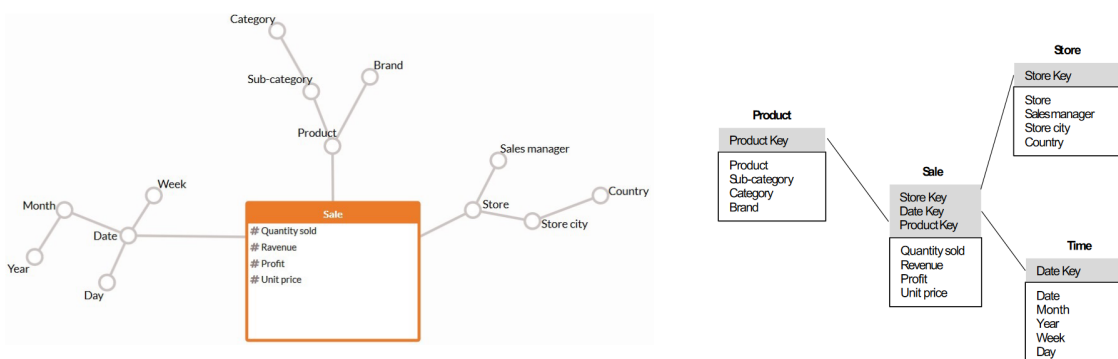


Figure 2.4: Example of star schema

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

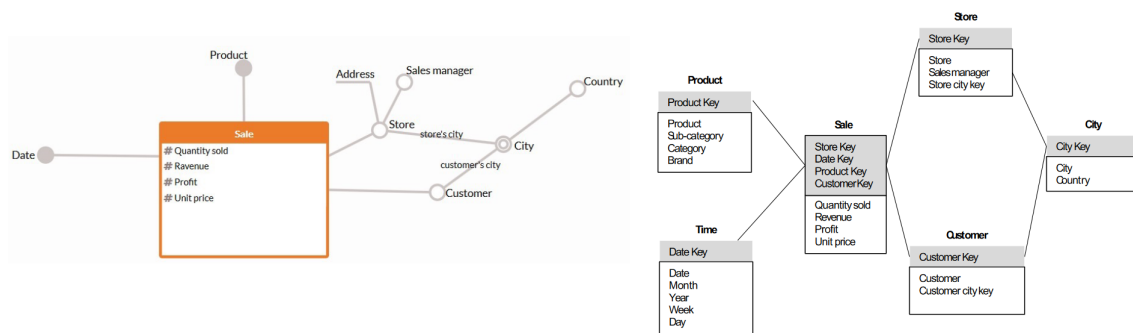


Figure 2.5: Example of snowflake schema

## 3 Data lake

**Dark data** Acquired and stored data that are never used for decision-making processes. Dark data

**Data lake** Repository to store raw (unstructured) data. It has the following features: Data lake

- Does not enforce a schema on write.
- Allows flexible access and applies schemas on read.
- Single source of truth.
- Low cost and scalable.

**Storage** Stored data can be classified as:

**Hot** A low volume of highly requested data that require low latency. More expensive HW/SW. Hot storage

**Cold** A large amount of data that does not have latency requirements. Less expensive. Cold storage

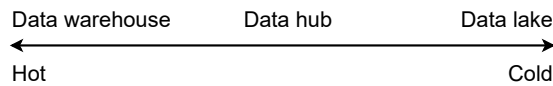


Figure 3.1: Data storage technologies

### 3.1 Traditional vs insight-driven data systems

	Traditional (data warehouse)	Insight-driven (data lake)
<b>Sources</b>	Structured data	Structured, semi-structured and un-structured data
<b>Storage</b>	Limited ingestion and storage capability	Virtually unlimited ingestion and storage capability
<b>Schema</b>	Schema designed upfront	Schema not fixed
<b>Transformations</b>	ETL upfront	Transformations on query
<b>Analytics</b>	SQL, BI tools, full-text search	Traditional methods, self-service BI, big data, machine learning, ...
<b>Price</b>	High storage cost	Low storage cost
<b>Performance</b>	Fast queries	Scalability/speed/cost tradeoffs
<b>Quality</b>	High data quality	Depends on the use case

### 3.2 Data architecture evolution

**Traditional data warehouse** (i.e. in-house data warehouse)

- Structured data with predefined schemas.
- High setup and maintenance cost. Not scalable.

Traditional data warehouse

- Relational high-quality data.
- Slow data ingestion.

### Modern cloud data warehouse

Modern cloud data warehouse

- Structured and semi-structured data.
- Low setup and maintenance cost. Scalable and easier disaster recovery.
- Relational high-quality data and mixed data.
- Fast data ingestion if supported.

### On-premise big data (i.e. in-house data lake)

On-premise big data

- Any type of data with schemas on read.
- High setup and maintenance cost.
- Fast data ingestion.

### Cloud data lake

Cloud data lake

- Any type of data with schemas on read.
- Low setup and maintenance cost. Scalable and easier disaster recovery.
- Fast data ingestion.

## 3.3 Components

### 3.3.1 Data ingestion

Data ingestion

**Workload migration** Inserting all the data from an existing source.

**Incremental ingestion** Inserting changes since the last ingestion.

**Streaming ingestion** Continuously inserting data.

**Change Data Capture (CDC)** Mechanism to detect changes and insert the new data into the data lake (possibly in real-time).

Change Data Capture (CDC)

### 3.3.2 Storage

**Raw** Immutable data useful for disaster recovery.

Raw storage

**Optimized** Optimized raw data for faster query.

Optimized storage

**Analytics** Ready to use data.

Analytics storage

### Columnar storage

- Homogenous data are stores contiguously.
- Speeds up methods that process entire columns (i.e. all the values of a feature).
- Insertion becomes slower.

**Data catalog** Methods to add descriptive metadata to a data lake. This is useful to prevent an unorganized data lake (data swamp).

### 3.3.3 Processing and analytics

Processing and analytics

**Interactive analytics** Interactive queries to large volumes of data. The results are stored back in the data lake.

**Big data analytics** Data aggregations and transformations.

**Real-time analytics** Streaming analysis.

## 3.4 Architectures

### 3.4.1 Lambda lake

Lambda lake

**Batch layer** Receives and stores the data. Prepares the batch views for the serving layer.

**Serving layer** Indexes batch views for faster queries.

**Speed layer** Receives the data and prepares real-time views. The views are also stored in the serving layer.

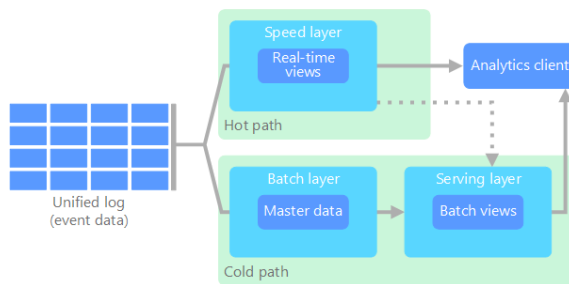


Figure 3.2: Lambda lake architecture

### 3.4.2 Kappa lake

The data are stored in a long-term store. Computations only happen in the speed layer (avoids lambda lake redundancy between batch layer and speed layer).

Kappa lake

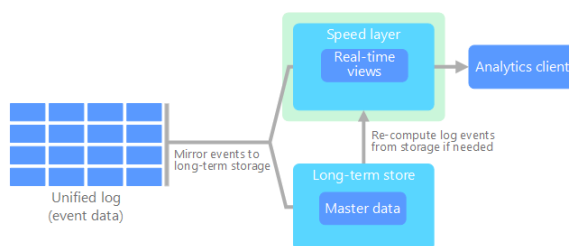


Figure 3.3: Kappa lake architecture

### 3.4.3 Delta lake

Framework that adds features on top of an existing data lake.

Delta lake

- ACID transactions
- Scalable metadata handling
- Data versioning
- Unified batch and streaming
- Schema enforcement

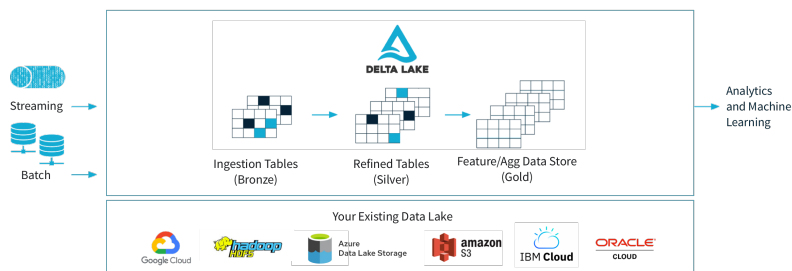


Figure 3.4: Delta lake architecture

## 3.5 Metadata

Metadata are used to organize a data lake. Useful metadata are:

Metadata

**Source** Origin of the data.

**Schema** Structure of the data.

**Format** File format or encoding.

**Quality metrics** (e.g. percentage of missing values).

**Lifecycle** Retention policies and archiving rules.

**Ownership**

**Lineage** History of applied transformations or dependencies.

**Access control**

**Classification** Sensitivity level of the data.

**Usage information** Record of who accessed the data and how it is used.



## 4 CRISP-DM

**Cross Industry Standard Process for Data Mining** Standardized process for data mining. CRISP-DM

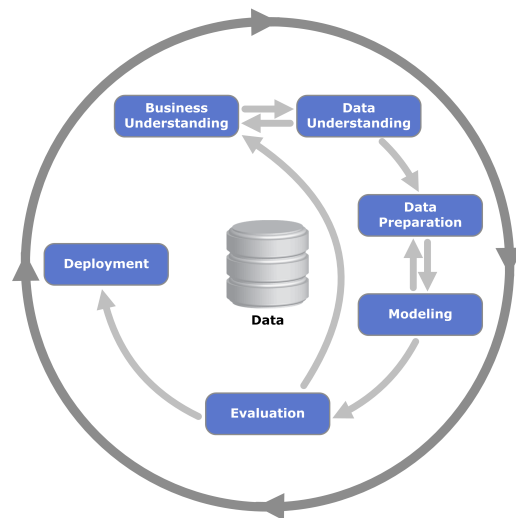


Figure 4.1: CRISP-DM workflow

### 4.1 Business understanding

- Determine the objective and the success criteria.
- Feasibility study.
- Produce a plan.

Business understanding

### 4.2 Data understanding

- Determine the available (raw) data.
- Determine the cost of the data.
- Collect, describe, explore and verify data.

Data understanding

### 4.3 Data preparation

- Data cleaning.
- Data transformations.

Data preparation

## 4.4 Modelling

- Select modelling technique.
- Build/train the model.

Modelling

## 4.5 Evaluation

- Evaluate results.
- Review process.

Evaluation

## 4.6 Deployment

- Plan deployment.
- Plan monitoring and maintenance.
- Final report and review.

Deployment