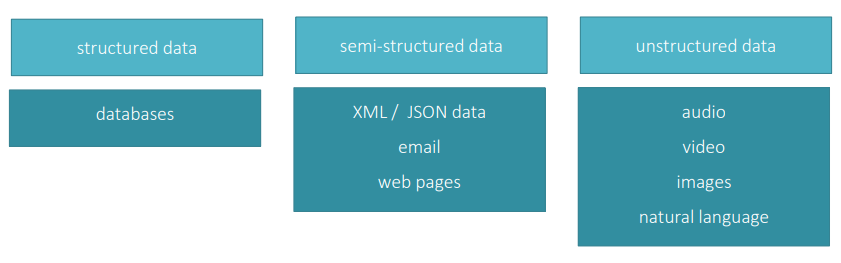
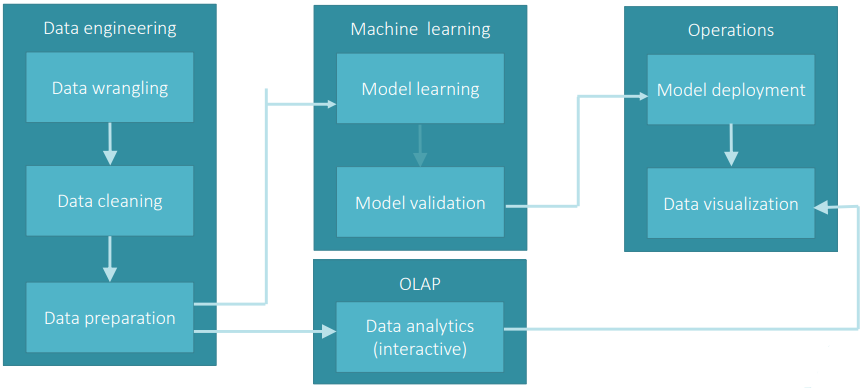
Data Mining

**Data**



**Typical Pipeline**

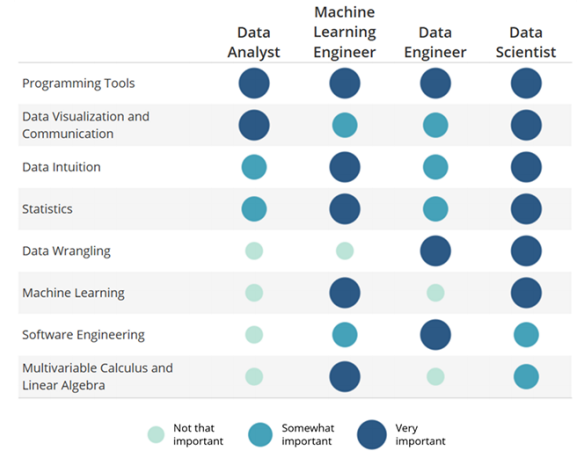


**Data Science** is the field that comprises of everything that’s related to data cleansing, preparation, and analysis.

**Big Data** is something that can be used to analyze insights which can lead to better decision and strategic business moves.

**Data Analytics** envolves automating insights into a certain dataset as well as supposes the usage of queries anda data aggregation procedures.

**Skills**



**Data mining users** – Scientists

* Flat files, simple databases, etc
* Study of a specific problema, just once
* Statistical methods; classification, clustering, etc

**Data warehousing & OLAP platforms** – Companies

* Large database schemas
* Periodic information, monitoring and trends analysis
* Reports and interactive data analysis (dashboards) for domain experts

**Multidimensional data modeling** is a graphical formalismo for modeling analytic applications and decision support systems. Representation of business events in data cubes. The basic concepts are facts, measures and dimensions; other concepts include hierarchies, descriptive attributes, cross-dimension attributes, multiple and optical arcs or convergence.

**Dimensions** are a fact property with a finite domain describing one of the coordinates to analyze. Represent the perspectives to which na organization wants to analyze data. Generally, represent answers to questions such as: when, where, what, who. Usually, at least one of the dimensions should represent time.

**Concept hierarchies** are relationships between dimension attributes. Directed trees composed by attributes describing a dimension and whose arcs model one-to-many associations between dimension attributes. Determine how events can be aggregated in decision-making process.

Primary event – indicates the occurrence of a fact and determines the finest aggregation gain.

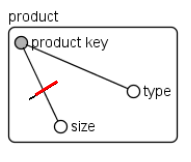
Secondary events – aggregation of primary events to sumarize a set of primary events.

Attributes of a dimension can be organized in a partial order, forming a lattice structure.

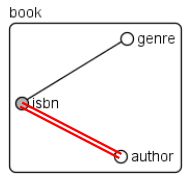
**Cross-dimension attribute** is a dimensional or descriptive atribute whose value is determined by the combination of two or more dimenssion attributes, possibly belonging to different hierarchies.

**Optional and multiple arcs**

Optional



Multiple



**Facts**

Events occurring in the enterprise world that are relevant for decision making process. Dynamic events, represented in the data source. Highly repeated events are good candidates.

Event facts – Model events in real-world.

Snapshot facts – Model the state of a process at a given time instant.

**Measures**

Numeric properties describing quantitative aspects of interest for analysis. Types:

* Additive – Values can be combined along any dimension
* Semi-additive – May occur when the fact is of type snapshot
* Non-additive measures – Values cannot be combined along any direction

**The process of multidimensional data modeling**

1. Understand and analyze business needs
2. Choose the gain of the business process
3. Design the dimensions that will apply to each fact table record
4. Choose the measures that will populate each fact table record

**Data Warehouse** for the whole organization. Collects all the information about subjects spanning the entire organization. Implementation cycle measured in months or years.

Issues:

* It’s often assumed that the only dynamic componentes in a cube are the events
* Dimensions are assumed to be static
* Today for yesterday – all events are referred to the current configuration
* Yesterday for today – all events are referred to some past configuration
* Historical truth – requires redundancy

**Data Mart** for part of na organization. Collects a subject of corporate data that is of value to specific users. Implementation cycle measured in weeks.

**Star Schema**

Facts table

* Compused by numeric values and FKs
* Large number of records

Dimension tables

* Not normalized
* Large number of attributes
* Few records relatively to facts table

**Snow-flake schema**

Some dimensions are normalized – avoids redundancy, but the structure is more complex and query operations are less eficiente than for star schema

**Constellation schema**

Several facts tables sharing common dimensions

**Types of slowly changing dimensions**

* Type 0 – no effort is made to reflect the changes in the dimension. All attributes preserver original values.
* Type 1 – changed data overwrites old entries
* Type 2 – the whole history is stored in the database
* Type 3 – only previous values of a dimension are written into the database
* Type 4 – store current data in the dimension table and older versions in historical tables

**Degenerate dimensions** – dimension that consists of a single indentifier. Easier to implemente and yields better performance

**Hierarchies**

* Parent-child hierarchies
* Unbalenced hierarchies
* Non-covering hierarchies

**Non-strict hierarchies** – many-to-many relationships cause problems with respect to aggregations. In ROLAP include na association table.

**Metadata in a data warehouse**

* defines the warehouse objects
* Acts as a directory to help locating the contentes of a data warehouse
* Describes the mapping of data from operational environment to data warehouse environment
* Is used in extraction, cleaning, transformation, querying, reporting and loading tools and processes

Categories:

* Business metadata – includes the data ownership information, business definition, and changing policies
* Technical metadata – includes database system names, table and column names and sizes, data types and allowed values, as well as structural information such as primary and foreign key attributes and indices
* Operational metada – includes currency of data and data lineage. Currency of data means whether the data is actived, archived, or purged. Lineage of data means the history of data migrated and transformation applied on it

**Extraction, transformation and loading (ETL)**

ETL/data preprocessing helps to improve the quality of data.

Main tasks:

* Data cleaning
  + Filling missing values
  + Smoothing noisy data
  + Identifying or removing outliers
  + Resolving inconsistencies
* Data integration
  + From multiple databases, data cubes, documents or flat files
    - Schema integration – finding attributes identifying same concepts in different sources; metadata and documentation are helpful
    - Object mapping – use of different names or values for the same object (entity)
  + Some redundancy can be detected by correlation analysis
* Data transformation