# Individuell uppgift 4, Sammanfattning

**Yasir Riyadh Jabbar (TIDAA KTH)**

**Kurslitteratur: Kapitel 11, 12**

## CHAPTER 11 Data Warehouses and Data Mining

### 11.1 Data Warehousing and Data Mining

Data Warehousing combines data that collected from many sources into one big database. It used for data analysis, decision making (called online analytical processing (**OLAP**)). It has many features:

1) Data items are linked together
2) Data is static (no updated/deleted)
3) Combined data that comes from different sources is made consistent
4) Stored data has specific time period

Data mining is the process of finding deviations, patterns, and correlations within large data sets to predict outcomes. By using statistics principles, it can be discovered new info that cannot simply be recovered in normal ways.

### 11.2 Operational DB

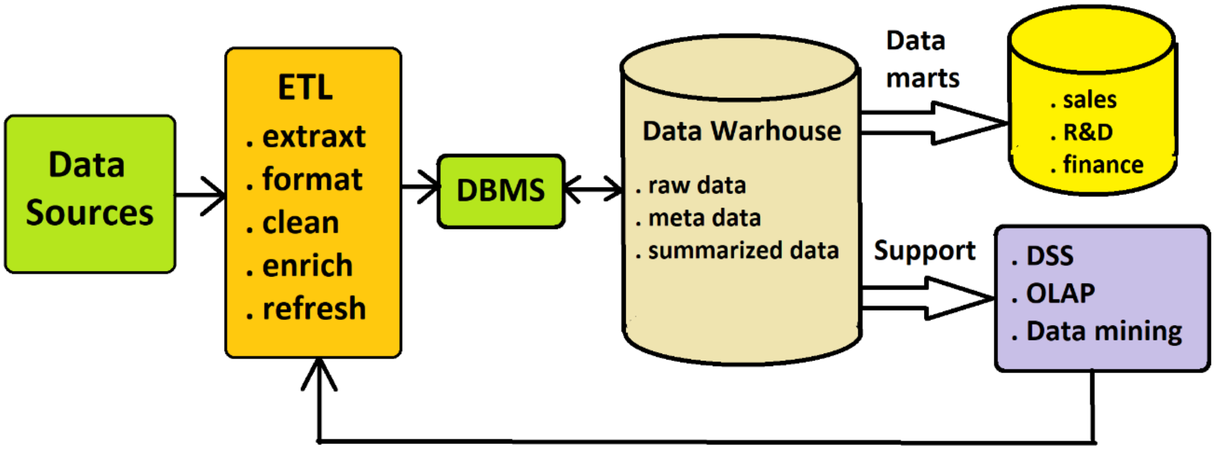
Operational DB is the source of info for data warehouse. Operational DBMS is called Online Transactions Processing (**OLTP**) and is used to deal with dynamic data in real-time.

| Data Warehouse | Operational DB |
|---|---|
| Support OLAP | Support OLTP |
| Concerned with historical data | Concerned with current data |
| No updated/deleted (just added) | Data is updated |
| Used for analysis of business | Used for real-time business |
| Access many rows/table | Access one row/table |
| Subject-oriented | Process-oriented |
| Perform fast retrievals of huge data | Perform inserts and updates of small data |

### 11.3 Architecture

Data sources are some operational DBs, independent files, environmental and financial data. Before putting data into data model for the warehouse and loaded, it must be consolidated into unified single repository by extraction, transformation, and loading (ETL) to ensure data integrity, validity, and quality. The next stage is DBMS to support data warehouse that has stores Metadata and DB components. Data warehouse is then used to:

- support requests for OLAP
- offer info for decision support systems (DSS)
- support data mining tools to determine new info



## 11.4 Developing

Two approaches for designing data warehouse:

1) **Top-Down**: Here we describe data warehouse as subject-oriented, time-variant, non-volatile. Also, integrated data repository is validated and stored normalized. The benefit here is developing new data mart is easy but cost of implementing is high.

2) **Bottom-Up**: Data mart here is formed first to necessary reporting and analytical abilities for business subjects. Here documents can be produced quickly and data warehouse could be extended.

## 11.5 Data Models

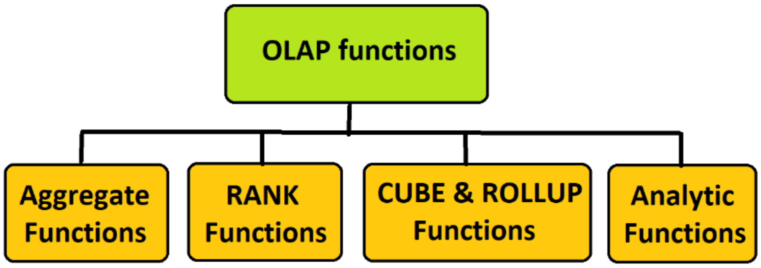
Data warehouse must use consistent model. Each model has 3 main components: DB Server, OLAP server, and Front-end tool.

- Relational **ROLAP** is used for big data that stored in relation tables. There are 3 types of ROLAP schemas: star, snowflake, and columnar data form.
- When it uses limited data volumes and stored in multidimensional array, it is called **MOLAP**.
- Hybrid **HOLAP** uses both ROLAP (to store detailed data) and MOLAP (to store aggregated data).

| Parameter | ROLAP | MOLAP | HOLAP |
|---|---|---|---|
| Processing time | very slow | fast | fast |
| Storage space | large | medium | small |
| Latency | Low | high | medium |
| Query response time | slow | fast | medium |

## 11.6 Data Warehouse Queries and SQL

The OLAP functions can be classified as:



1) Aggregate Functions: SUM, MAX, MIN, COUNT, and AVG
2) RANK Functions: RANK (), DENSE_RANK ( )
3) CUBE and ROLLUP Functions: GROUP BY CUBE, GROUP BY ROLLUP
4) Analytic Functions: CORR (correlation), REGR (regression), CUME_DIST (cumulative distribution)

## 11.7 Views and View Materialization

View allows DB models to be customized (presents some calculated data, or gives a summary info). Queries and views in data warehouse environment is considered complex and query modification may result some delay in response time. Another method is to materialize views (it is customized object that contains the results of a query), precomputing, then storing them for later use (also indexes can be used to reduce delay).

## 11.8 Data Warehouse Vendors

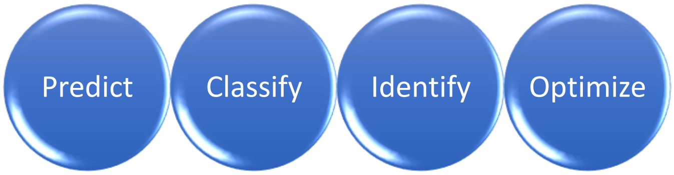
Special software and hardware platforms are required for data warehouse (ex. Oracle, IBM Netezza, Microsoft, and SAP). Data warehouse appliances are integrated systems that installed and optimized and include complete package of servers, OS, and DBMS.

## 11.9 Data Mining

Data warehouse is used as data source for data mining and include summarized data and raw data (taken from operational DB). Data mining uses statistics and AI techniques (machine learning) to determine new info from big data. It requires knowledge of environment of the object being determined. Data format must be flat file (single record rather than normalized relational tables) and contains values either numerical or categorical.

## 11.10 Purpose of Data Mining

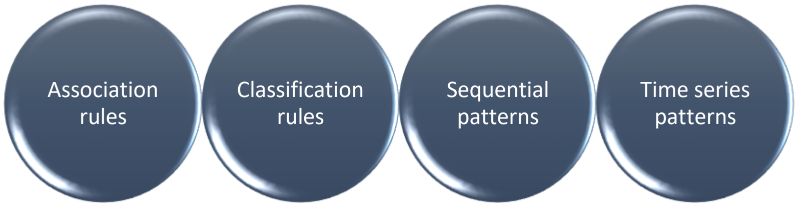
Generally, data mining Provides knowledge to achieve:



1) Predict future behavior by studying data from multiple sources and expect some features.
2) Classify items in categories for better understanding relationships and connections between items.
3) Identify existence of activity by studying characteristics of prior events to determine the possible new activities.
4) Optimize the use of resources for best results and maximize productivity and efficiency.

## 11.11 Types of Knowledge Discovered

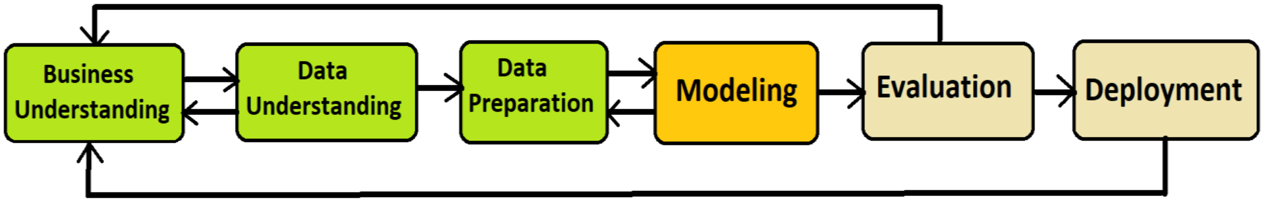
Knowledge can be characterized in different ways:



- Association rules: have form **{x} ⇒ {y}**. Two measures here, **support** (percentage of transactions in data set that contain all items in both sides) and **confidence** (is measure of how rule proves to be true)
- Classification rules: using **training set** method to put the item (whose class is currently unknown) in correct category.
- Sequential patterns: the prediction of object behavior in one transaction in another transaction.
- Time series patterns: sequence of events (all same type) that can be examined to determine patterns and sequences.

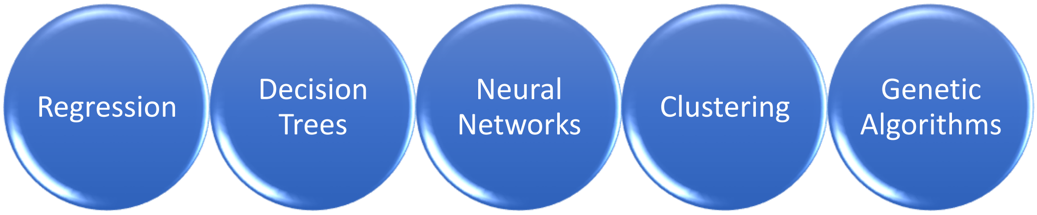
## 11.12 Models and Methods Used

Generally, the process steps of data mining model are:



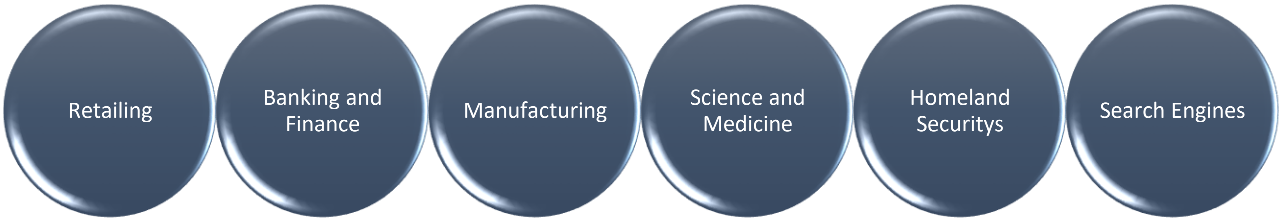
- Business Understanding: requirements, objectives of the project and problem formulation
- Data Understanding: data is examined (charts and graphs) to detect any noticeable subsets
- Data Preparation: constructing the input data, cleaning and reformatting
- Modeling: selecting the appropriate model, identifying dependent variable (target) and independent variables (predictor)
- Evaluation: tested, validated, and evaluated to check it achieves the specified objectives
- Deployment: developed model putting into use

Every model is created by different algorithms and depends on required output.



- Regression: statistical method for predicting dependent variable given independent variables. Also, can use curve-fitting method to find the equation which fits observed variables
- Decision Trees: used for establishing classification system based on common variables or for making prediction algorithm for target object.
- Neural Networks: it is non-linear models that use learning process and adapt their learning to new info by testing other samples
- Clustering: partitions set of objects based on features then aggregates them according to likenesses (disjoint or overlapping). The aim here to implement specific join algorithm
- Genetic Algorithms: It is advanced data classification operates as adaptive search algorithm and classify best optimum solution between several candidate solutions (individuals). It executes iteratively combination, mutation, selection, then encoding process to develop successive generation of another models (such as decision trees or neural networks layers).

## 11.13 Applications of Data Mining



1) Retailing: to allow retailer to offer personalized shopping skill to clients to build client loyalty, and to identify the clients how are more likely to make purchase in response to the Ads
2) Banking and Finance: building a form to determine whether credit will be offered to new applicants, and to determine whether claims are fraudulent
3) Manufacturing: data mining is used to determine the best use of resources, to maximize efficiency and lowest cost for products, and to improve the design by checking product defects data and feedback from clients
4) Science and Medicine: data mining is used to determine effectiveness of treatments: to analyze effects of drugs, to find relationships, in astronomy, for weather prediction, and in bioinformatics
5) Homeland Security: to Identify individual terrorists and track activities
6) Search Engines: to identify and classify resources for user requests by examining big data on Internet

## 11.14 Data Mining Software Providers

Data mining (DM) software is offered either as:

- standalone product
- part of a DB
- part of data warehouse management system

The most popular venders:

1) SAS offers DM software includes tools for process (from data preparation to scoring) and several algorithms to build predicting and classifying models.
2) ODM (Oracle) offers various advanced DM algorithms for irregularity detection, regression, association rules, classification, and segmentation
3) IBM’s SPSS offers auto data preparation, systematic functions, algorithms and interactive data visualization.
4) SAP offers DM tools for various applications such as social network analysis.

# CHAPTER 12 Big Data and NoSQL

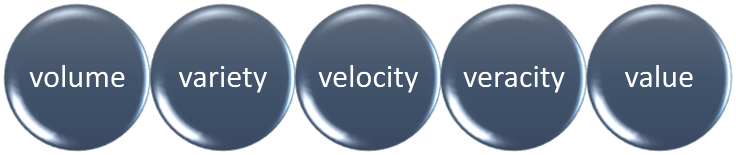
## 12.1 Big data technologies

Big data is just massive data sets which cannot be treated with conventional DB technology (ex. Facebook stores over 500 TB data and analyzes it every day). There are special technologies that deal with big data:

- Hadoop: framework that store and process efficiently big data sets from GB to PB of data. The backbone of Hadoop is Hadoop distributed file system (**HDFS**) and Map reduce parallel programming model (**MR**). It also used data warehouse system **Hive** and **HiveQL** to query and analyze data that stored in HDFS
- NoSQL: (Not only SQL): Unlike relational DB, NoSQL is not limited by fixed schema model. Has many good features as flexible schemas, horizontal scaling, and fast queries.
- NewSQL: to obtain more scalable solutions, it provides distributed DB architectures with cloud computing for applications that deal with massive data

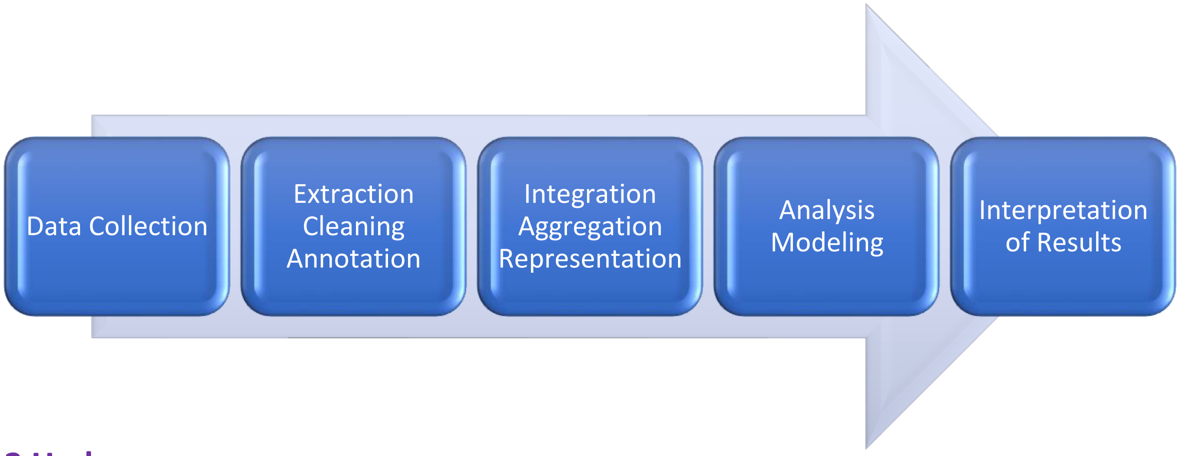
## 12.2 Defining Big Data

Big data can be defined as (5 Vs):



- Volume: big amounts of collected data from different sources (sensors, monitors, social media, and log files)
- Variety: different forms of generated data (text, video, audio, unstructured data)
- Velocity: different arrived data rates make problems to capture and analyze data. this leads to use suitable data sampling techniques for real-time analysis and streaming data
- Veracity: accuracy of analyzed data (correctness, validity, trustworthiness)
- Value: the best outcomes (knowledge, insight, competitive edge, discovery) that can be obtained from analysis of big data to businesses fields and research labs.

The process steps of using big data can be shown in following figure:

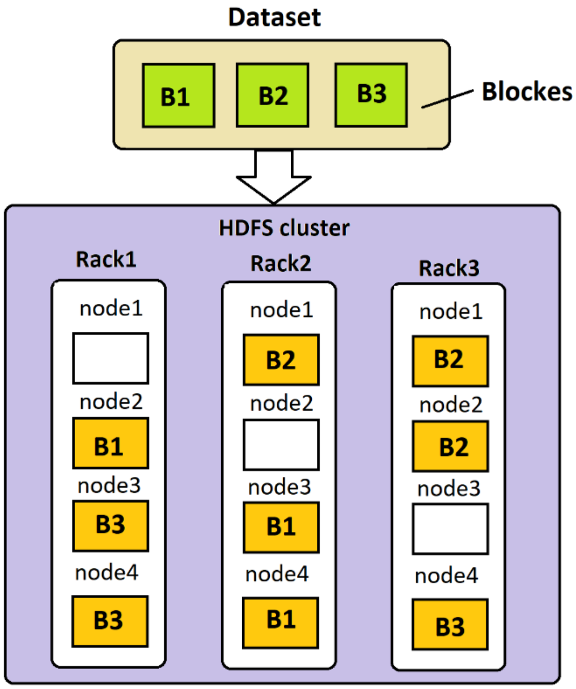


## 12.3 Hadoop

It is framework stores big data in distributed systems and parallelly processes it. The main components are HDFS and MR parallel programming model. It contains combination of structured, semi-structured, and unstructured info.

### HDFS

Data in DFS is stored in server and can be accessed and processed to share info and files among users on network (in controlled and authorized way). It can represent big data that is not store in single machine by distributing it across network. HDFS divides file into blocks (64 MB each) then of each block replicates copies and distributes the blocks across several computers. Each HDFS cluster is made of nodes (single computer) and racks (30-40 nodes).



### MapReduce

MapReduce parallel programming model is used to parallelize operations of file reading (from blocks) to perform data computation then merges the results. Initially, data is converted to **key-value** pairs.

- map step filters and convert data into suitable form
- reduce step then makes some calculation or aggregation over data

### Hive and HiveQL

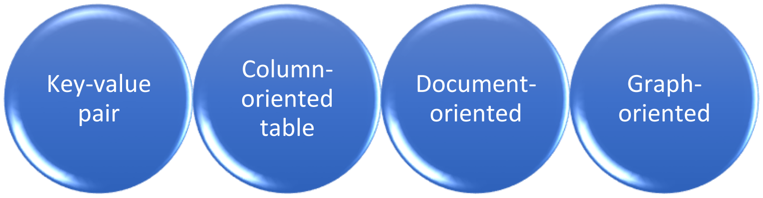
Hive is a system provides DB query functions for access and manipulation data in Hadoop. it is lacking functionalities updates, transactions, and indexes but it maps Hadoop data to table structures. HiveQL is query language that supports queries over Hive table structures, summarization and analysis of data.

## 12.4 NoSQL

NoSQL is not limited by fixed schema model. It has many good features as flexible schemas, horizontal scaling, and fast real-time processing queries.

- efficient handling of big data (read/write) in real-time with low latency
- can handle millions of users (horizontal scaling)
- can handle structured, semi-structured, and unstructured data
- support parallel processing abilities

### Types of NoSQL systems



Example: HBase column-oriented is non-relational DB and data is stored in individual columns and indexed by unique row key.

## 12.5 NewSQL

NewSQL is new distributed DB architectures with cloud computing. It obtains more scalable solutions, and provides distributed DB architectures with cloud computing for applications that deal with massive data that satisfy ACID properties.