Unit -3

A data warehouse is a large collection of business data used to help an organization make decisions. The concept of the data warehouse has existed since the 1980s, when it was developed to help transition data from merely powering operations to fueling decision support systems that reveal business intelligence. The large amount of data in data warehouses comes from different places such as internal applications such as marketing, sales, and finance; customer-facing apps; and external partner systems, among others.
On a technical level, a data warehouse periodically pulls data from those apps and systems; then, the data goes through formatting and import processes to match the data already in the warehouse. The data warehouse stores this processed data so it's ready for decision makers to access. How frequently data pulls occur, or how data is formatted, etc., will vary depending on the needs of the organization.
Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

Using Data Warehouse Information

There are decision support technologies that help utilize the data available in a data warehouse. These technologies help executives to use the warehouse quickly and effectively. They can gather data, analyze it, and take decisions based on the information present in the warehouse. The information gathered in a warehouse can be used in any of the following domains −

- **Tuning Production Strategies** − The product strategies can be well tuned by repositioning the products and managing the product portfolios by comparing the sales quarterly or yearly.
- **Customer Analysis** − Customer analysis is done by analyzing the customer's buying preferences, buying time, budget cycles, etc.
- **Operations Analysis** − Data warehousing also helps in customer relationship management, and making environmental corrections. The information also allows us to analyze business operations.

Integrating Heterogeneous Databases

To integrate heterogeneous databases, we have two approaches −

- Query-driven Approach
- Update-driven Approach

Query-Driven Approach

This is the traditional approach to integrate heterogeneous databases. This approach was used to build wrappers and integrators on top of multiple heterogeneous databases. These integrators are also known as mediators.

Process of Query-Driven Approach

- When a query is issued to a client side, a metadata dictionary translates the query into an appropriate form for individual heterogeneous sites involved.
- Now these queries are mapped and sent to the local query processor.
- The results from heterogeneous sites are integrated into a global answer set.

Disadvantages

- Query-driven approach needs complex integration and filtering processes.
- This approach is very inefficient.

- It is very expensive for frequent queries.
- This approach is also very expensive for queries that require aggregations.

Update-Driven Approach

This is an alternative to the traditional approach. Today's data warehouse systems follow update-driven approach rather than the traditional approach discussed earlier. In update-driven approach, the information from multiple heterogeneous sources are integrated in advance and are stored in a warehouse. This information is available for direct querying and analysis.

Advantages

This approach has the following advantages −

- This approach provides high performance.
- The data is copied, processed, integrated, annotated, summarized and restructured in semantic data store in advance.
- Query processing does not require an interface to process data at local sources.

Functions of Data Warehouse Tools and Utilities

The following are the functions of data warehouse tools and utilities −

- **Data Extraction** − Involves gathering data from multiple heterogeneous sources.
- **Data Cleaning** − Involves finding and correcting the errors in data.
- **Data Transformation** − Involves converting the data from legacy format to warehouse format.
- **Data Loading** − Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.
- **Refreshing** − Involves updating from data sources to warehouse.

Note − Data cleaning and data transformation are important steps in improving the quality of data and data mining results.

Some benefits of a data warehouse

Organizations that use a data warehouse to assist their analytics and business intelligence see a number of substantial benefits:

Better data — Adding data sources to a data warehouse enables organizations to ensure that they are collecting consistent and relevant data from that source. They don’t need to wonder whether the data will be accessible or inconsistent as it comes in to the system. This ensures higher data quality and data integrity for sound decision making.

Faster decisions — Data in a warehouse is in such consistent formats that it is ready to be analyzed. It also provides the analytical power and a more complete dataset to base decisions on hard facts. Therefore, decision makers no longer need to reply on hunches, incomplete data, or poor quality data and risk delivering slow and inaccurate results.

Delivers enhanced business intelligence

- By having access to information from various sources from a single platform, decision makers will no longer need to rely on limited data or their instinct. Additionally, data warehouses can effortlessly be applied to a business’s processes, for instance, market segmentation, sales, risk, inventory, and financial management.

- **Saves times**

- A data warehouse standardizes preserves, and stores data from distinct sources, aiding the consolidation and integration of all the data. Since critical data is available to all users, it allows them to make informed decisions on key aspects. In addition, executives can query the data themselves with little to no IT support, saving more time and money.

- **Enhances data quality and consistency**

- A data warehouse converts data from multiple sources into a consistent format. Since the data from across the organization is standardized, each department will produce results that are consistent. This will lead to more accurate data, which will become the basis for solid decisions.

- **Generates a high Return on Investment (ROI)**

- Companies experience higher revenues and cost savings than those that haven't invested in a data warehouse.

- **Provides competitive advantage**

- Data warehouses help get a holistic view of their current standing and evaluate opportunities and risks, thus providing companies with a competitive advantage.

- **Improves the decision-making process**

- Data warehousing provides better insights to decision makers by maintaining a cohesive database of current and historical data. By transforming data into purposeful information, decision makers can perform more functional, precise, and reliable analysis and create more useful reports with ease.

- **Enables organizations to forecast with confidence**

- Data professionals can analyze business data to make market forecasts, identify potential KPIs, and gauge predicated results, allowing key personnel to plan accordingly.

- **Streamlines the flow of information**

- Data warehousing facilitates the flow of information through a network connecting all related or non-related parties.

- **What a data warehouse is not**
- **1. It is not a database**
- It's easy to confuse a data warehouse with a database, since both concepts share some similarities. The primary difference, however, comes into effect when a business needs to perform analytics on a large data collection. Data warehouses are made to handle this type of task, while databases are not. Here's a comparison chart that tells the difference between the two:

	Database	**Data Warehouse**
What it is	Data collected for multiple transactional purposes. Optimized for read/write access.	Aggregated transactional data, transformed and stored for analytical purposes. Optimized for aggregation and retrieval of large data sets.
How it's used	Databases are made to quickly record and retrieve information.	Data warehouses store data from multiple databases, which makes it easier to analyze.
Types	Databases are used in data warehousing. However, the term usually refers to an online, transactional processing database. There are other types as	A data warehouse is an analytical database that layers on top of transactional databases to allow for analytics.

	well, including csv, html, and Excel spreadsheets used for database purposes.	

- **2. It is not a data lake**
- Although they both are built for business analytics purposes, the major difference between a data lake and a data warehouse is that a data lake stores all types of raw, structured, and unstructured data from all data sources in its native format until it is needed. By contrast, a data warehouse stores data in files or folders in a more organized fashion that is readily available for reporting and data analysis.

- **3. It is not a data mart**
- Data warehouses are also sometimes confused with data marts. But data warehouses are generally much bigger and contain a greater variety of data, while data marts are limited in their application.
- Data marts are often subsets of a warehouse, designed to easily deliver specific data to a specific user, for a specific application. In the simplest terms, data marts can be thought of as single-subject, while data warehouses cover multiple subjects.

Characteristics of Data Warehouse Design

The following are the main characteristics of data warehousing design development and best practices:

Theme-Focused

A data warehouse design uses a particular theme. It provides information concerning a subject rather than a business's operations. These themes can be related to sales, advertising, marketing, and more.

Instead of focusing on the business operations or transactions, data warehousing emphasizes on business intelligence (BI) that is, displaying and analyzing data for decision-making. It also offers a straightforward and succinct interpretation of the particular theme by eliminating data that may not be useful for decision-makers.

Unified

A data warehouse design unifies and integrates all analogous data from different databases in a collectively acceptable way using data modeling. It incorporates data from diverse sources such as relational and non-relational databases, flat files, mainframe, cloud-based systems, etc. Besides, a data warehouse must maintain consistent nomenclature, layout, and coding to facilitate effective data analysis.

Time Variance

Unlike other operational systems, data warehouse stores data collected over an extensive time horizon. The data gathered is identified with specific time duration and provides insights from the past perspective. Moreover, when data is entered into the warehouse, it cannot be restructured or altered.

Non-volatility

Another important characteristic is non-volatility which means that the preceding data is not removed when new data is loaded to the data warehouse. Moreover, data is only readable and can be intermittently refreshed to deliver a complete and updated picture to the user.

Types of Data Warehouse Architecture

A data warehouse architecture defines the arrangement of data and the storing structure. As the data must be organized and cleansed to be valuable, a modern data warehouse architecture centres on identifying the most effective technique of extracting information from raw data in the staging area and converting it into a simple consumable structure using a dimensional model that delivers valuable business intelligence.

When designing a company's data warehouse, there are three main types of architecture to take into consideration.

Single-tier architecture

A single-tier data warehouse architecture centers on producing a dense set of data and reducing the volume of data deposited. Although it is beneficial for eliminating redundancies, this architecture is not suitable for businesses with complex data requirements and numerous data streams. This is where 2-tier and 3-tier architecture of data warehouse comes in as they both deal with more complex data streams.

Two-tier architecture

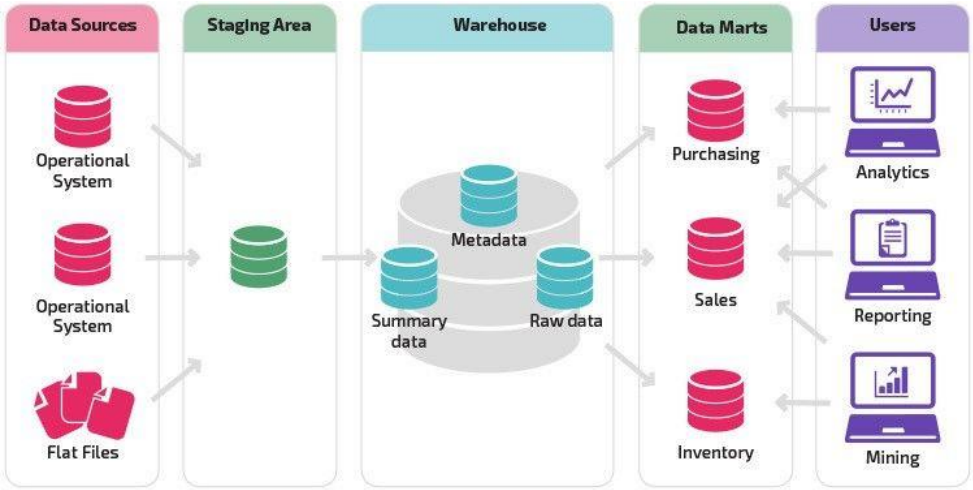
This architecture splits the tangible data sources from the warehouse itself. Although it is more efficient at data storage and organization, the two-tier architecture is not scalable. Moreover, it only supports a nominal number of users.

Three-tier architecture

This is the most common type of modern data warehouse architecture as it produces a well-organized data flow from raw information to valuable insights.

The bottom tier typically comprises of the databank server that creates an abstraction layer on data from numerous sources, like transactional databanks utilized for front-end uses.

The middle tier includes an Online Analytical Processing (OLAP) server. From a user's perspective, this level alters the data into an arrangement that is more suitable for analysis and multifaceted probing. Since it includes OLAP server pre-built in the architecture, we can also call it the OLAP focused data warehouse.



The third and the topmost tier is the client level which includes the tools and Application Programming Interface (API) used for high-level data analysis, inquiring, and reporting. However, barely people also include the 4-tier architecture of data warehouse but it is often not considered as integral as other three types of data warehouse architecture.

These are the different types of data warehouse architecture in data mining. Now let's learn about the elements of a data warehouse (DWH) architecture and how they help build and scale a data warehouse in detail.

Main Components of Data Warehouse Architecture

Now that we have discussed the three data warehouse architectures, let's look at the main constituents of a data warehouse.

A data warehouse design mainly consists of six key components.

1. Data Warehouse Database

The central component of a data warehousing architecture is a databank that stocks all enterprise data and makes it manageable for reporting. Obviously, this means you need to choose which kind of database you'll use to store data in your warehouse.

The following are the four database types that you can use:

- **Typical relational databases** which are the row-centered databases you perhaps use on an everyday basis. For example, Microsoft SQL Server, SAP, Oracle, and IBM DB2.
- **Analytics databases** are precisely developed for data storage to sustain and manage analytics. For example, Teradata and Greenplum.
- **Data warehouse applications** which aren't exactly a kind of storage databases, but several dealers now offer applications that offer software for data management as well as hardware for storing data. For example, SAP Hana, Oracle Exadata, and IBM Netezza.
- **Cloud-based databases** which can be hosted and retrieved on the cloud so that you don't have to procure any hardware to set up your data warehouse. For example, Amazon Redshift, Microsoft Azure SQL, and Google BigQuery.

2. Extraction, Transformation, and Loading Tools (ETL)

ETL tools are central to a data warehouse architecture. These tools help with extracting data from different sources, transforming it into a suitable arrangement, and loading it into a data warehouse.

The ETL tool you choose will determine:

- The time expended in data extraction
- Approaches to extracting data
- Kind of transformations applied and the simplicity to do so
- Business rule definition for data validation and cleansing to improve end-product analytics
- Filling mislaid data
- Outlining information distribution from the fundamental depository to your BI applications

3. Metadata

Metadata describes the data warehouse and offers a framework for data. It helps in constructing, preserving, handling and making use of the data warehouse.

It can be characterized into two types:

- **Technical Metadata**, which comprises information that can be used by developers and managers when executing warehouse development and administration tasks.

- **Business Metadata**, which comprises information that offers easily understandable standpoint of the data stored in the warehouse.

Metadata plays an important role for the businesses as well as the technical teams to understand the data present in the warehouse and to convert it into information.

4. Data Warehouse Access Tools

A data warehouse uses a database or group of databases as a foundation. Corporate users generally cannot work with databases directly. This is why they use the assisstance of several tools. Some of these tools include:

- **Query and reporting tools**, which help users produce corporate reports for analysis that can be in the form of spreadsheets, calculations, or interactive visuals.
- **Application development tools**, which help create tailored reports and present them in interpretations intended for particular reporting purposes.
- **Data mining tools**, which systematize the procedure of identifying arrays and links in huge quantities of data using cutting-edge statistical modeling methods.
- **OLAP tools**, which help construct a multi-dimensional data warehouse and allow analysis of enterprise data from numerous viewpoints.

5. Data Warehouse Bus

It defines the data flow within a data warehousing bus architecture and includes a data mart. A data mart is an access level used to transfer data to the users. It is used for partitioning data which is produced for the particular user group.

6. Reporting Layer

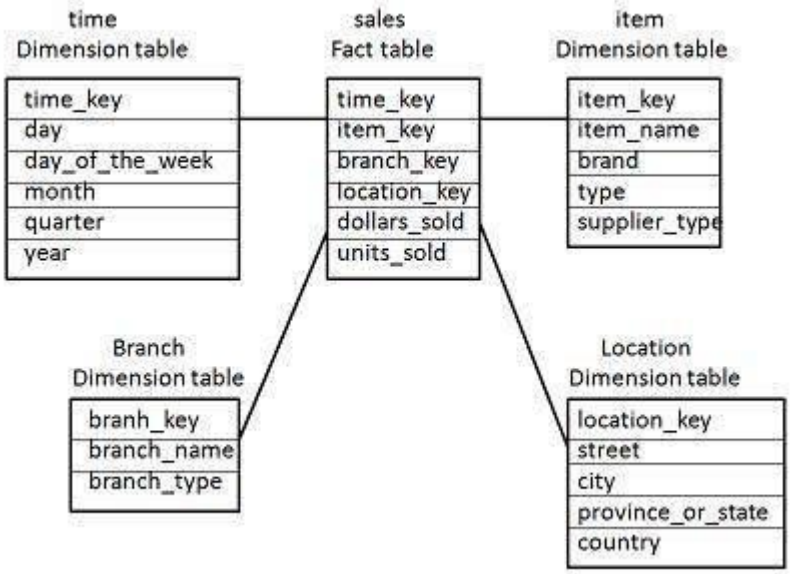
The reporting layer in the data warehouse allows the end-users to access the BI interface or BI database architecture. The purpose of this layer is to act as a dashboard for data visualization, create reports, and take out any required information.

Data Warehousing - Schemas

Schema is a logical description of the entire database. It includes the name and description of records of all record types including all associated data-items and aggregates. Much like a database, a data warehouse also requires to maintain a schema. A database uses relational model, while a data warehouse uses Star, Snowflake, and Fact Constellation schema. In this chapter, we will discuss the schemas used in a data warehouse.

Star Schema

- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.

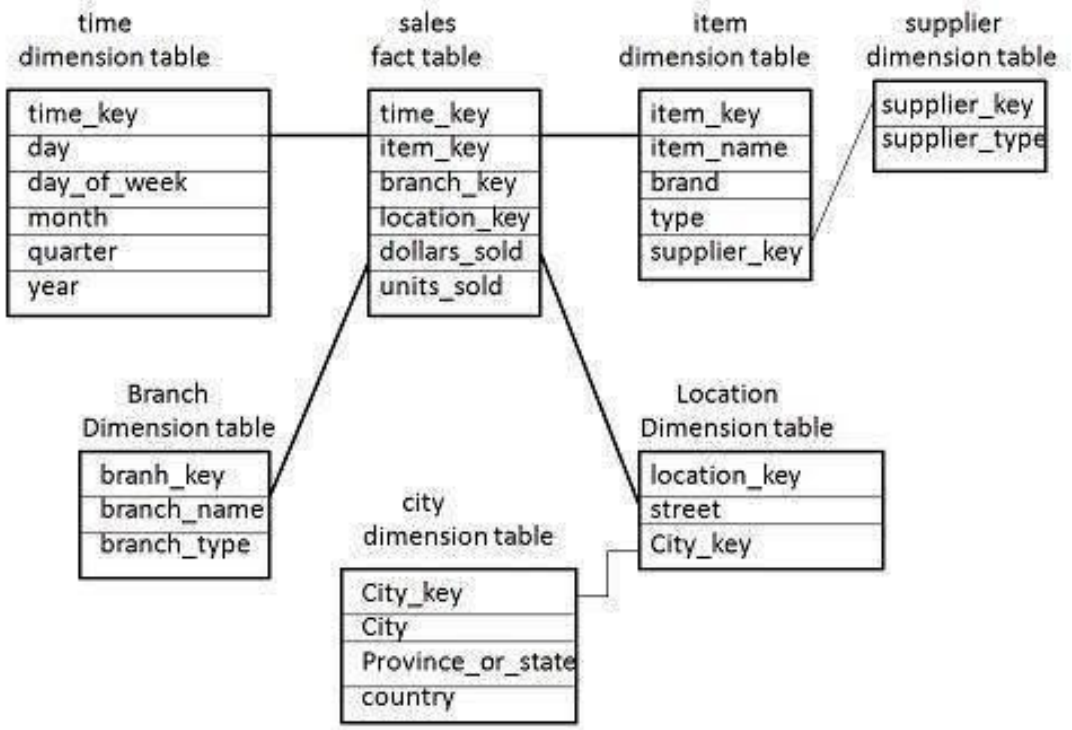


- There is a fact table at the centre. It contains the keys to each of four dimensions.
- The fact table also contains the attributes, namely dollars sold and units sold.

Note − Each dimension has only one dimension table and each table holds a set of attributes. For example, the location dimension table contains the attribute set {location_key, street, city, province_or_state,country}. This constraint may cause data redundancy. For example, "Vancouver" and "Victoria" both the cities are in the Canadian province of British Columbia. The entries for such cities may cause data redundancy along the attributes province_or_state and country.

Snowflake Schema

- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.

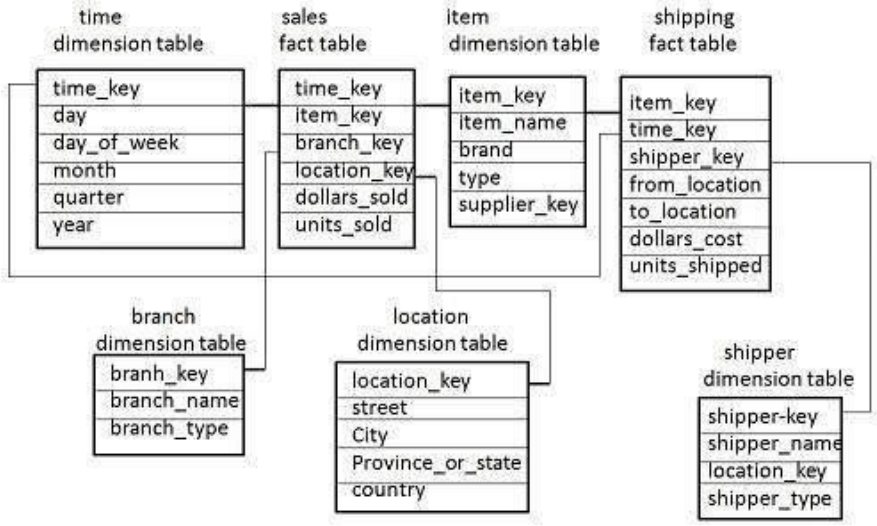


- Now the item dimension table contains the attributes item_key, item_name, type, brand, and supplier-key.
- The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier_key and supplier_type.

Note − Due to normalization in the Snowflake schema, the redundancy is reduced and therefore, it becomes easy to maintain and the save storage space.

Fact Constellation Schema

- A fact constellation has multiple fact tables. It is also known as galaxy schema.
- The following diagram shows two fact tables, namely sales and shipping.



- The sales fact table is same as that in the star schema.
- The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.
- The shipping fact table also contains two measures, namely dollars sold and units sold.
- It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.

Schema Definition

Multidimensional schema is defined using Data Mining Query Language (DMQL). The two primitives, cube definition and dimension definition, can be used for defining the data warehouses and data marts.

Syntax for Cube Definition
define cube < cube_name > [< dimension-list > }: < measure_list >

Syntax for Dimension Definition
define dimension < dimension_name > as (< attribute_or_dimension_list >)

Star Schema Definition

The star schema that we have discussed can be defined using Data Mining Query Language (DMQL) as follows −

define cube sales star [time, item, branch, location]:

dollars sold = sum(sales in dollars), units sold = count(*)

define dimension time as (time key, day, day of week, month, quarter, year)
define dimension item as (item key, item name, brand, type, supplier type)
define dimension branch as (branch key, branch name, branch type)
define dimension location as (location key, street, city, province or state, country)

Snowflake Schema Definition

Snowflake schema can be defined using DMQL as follows −

define cube sales snowflake [time, item, branch, location]:

dollars sold = sum(sales in dollars), units sold = count(*)

define dimension time as (time key, day, day of week, month, quarter, year)
define dimension item as (item key, item name, brand, type, supplier (supplier key, supplier type))
define dimension branch as (branch key, branch name, branch type)
define dimension location as (location key, street, city (city key, city, province or state, country))

Fact Constellation Schema Definition

Fact constellation schema can be defined using DMQL as follows −

define cube sales [time, item, branch, location]:

dollars sold = sum(sales in dollars), units sold = count(*)

define dimension time as (time key, day, day of week, month, quarter, year)
define dimension item as (item key, item name, brand, type, supplier type)
define dimension branch as (branch key, branch name, branch type)
define dimension location as (location key, street, city, province or state,country)
define cube shipping [time, item, shipper, from location, to location]:

dollars cost = sum(cost in dollars), units shipped = count(*)

define dimension time as time in cube sales
define dimension item as item in cube sales
define dimension shipper as (shipper key, shipper name, location as location in cube sales, shipper type)
define dimension from location as location in cube sales
define dimension to location as location in cube sales

What is Data?

The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or

Introduction

Big data is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large datasets. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing has greatly expanded in recent years.

In this article, we will talk about big data on a fundamental level and define common concepts you might come across while researching the subject. We will also take a high-level look at some of the processes and technologies currently being used in this space.

What Is Big Data?

An exact definition of "big data" is difficult to nail down because projects, vendors, practitioners, and business professionals use it quite differently. With that in mind, generally speaking, **big data** is:

- large datasets
- the category of computing strategies and technologies that are used to handle large datasets

In this context, "large dataset" means a dataset too large to reasonably process or store with traditional tooling or on a single computer. This means that the common scale of big datasets is constantly shifting and may vary significantly from organization to organization.

Why Are Big Data Systems Different?

The basic requirements for working with big data are the same as the requirements for working with datasets of any size. However, the massive scale, the speed of ingesting and processing, and the characteristics of the data that must be dealt with at each stage of the process present significant new challenges when designing solutions. The goal of most big data systems is to surface insights and connections from large volumes of heterogeneous data that would not be possible using conventional methods.

In 2001, Gartner's Doug Laney first presented what became known as the "three Vs of big data" to describe some of the characteristics that make big data different from other data processing:

Volume

The sheer scale of the information processed helps define big data systems. These datasets can be orders of magnitude larger than traditional datasets, which demands more thought at each stage of the processing and storage life cycle.

Often, because the work requirements exceed the capabilities of a single computer, this becomes a challenge of pooling, allocating, and coordinating resources from groups of computers. Cluster management and algorithms capable of breaking tasks into smaller pieces become increasingly important.

Velocity

Another way in which big data differs significantly from other data systems is the speed that information moves through the system. Data is frequently flowing into the system from multiple sources and is often expected to be processed in real time to gain insights and update the current understanding of the system.

This focus on near instant feedback has driven many big data practitioners away from a batch-oriented approach and closer to a real-time streaming system. Data is constantly being added, massaged, processed, and analyzed in order to keep up with the influx of new information and to surface valuable information early when it is most relevant. These ideas require robust systems with highly available components to guard against failures along the data pipeline.

Variety

Big data problems are often unique because of the wide range of both the sources being processed and their relative quality.

Data can be ingested from internal systems like application and server logs, from social media feeds and other external APIs, from physical device sensors, and from other providers. Big data seeks to handle potentially useful data regardless of where it's coming from by consolidating all information into a single system.

The formats and types of media can vary significantly as well. Rich media like images, video files, and audio recordings are ingested alongside text files, structured logs, etc. While more traditional data processing systems might expect data to enter the pipeline already labeled, formatted, and organized, big data systems usually accept and store data closer to its raw state. Ideally, any transformations or changes to the raw data will happen in memory at the time of processing.

Other Characteristics

Various individuals and organizations have suggested expanding the original three Vs, though these proposals have tended to describe challenges rather than qualities of big data. Some common additions are:

- **Veracity**: The variety of sources and the complexity of the processing can lead to challenges in evaluating the quality of the data (and consequently, the quality of the resulting analysis)
- **Variability**: Variation in the data leads to wide variation in quality. Additional resources may be needed to identify, process, or filter low quality data to make it more useful.
- **Value**: The ultimate challenge of big data is delivering value. Sometimes, the systems and processes in place are complex enough that using the data and extracting actual value can become difficult.

Types of Big Data

Following are the types of Big Data:

1. **Structured**
2. **Unstructured**
3. **Semi-structured**

Structured

Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes.

Do you know? 10^{21} bytes equal to ***1 zettabyte*** or ***one billion terabytes*** forms ***a zettabyte***.

Looking at these figures one can easily understand why the name Big Data is given and imagine the challenges involved in its storage and processing.

Do you know? Data stored in a relational database management system is one example of a '**structured**' data.

Examples of Structured Data

An 'Employee' table in a database is an example of Structured Data

Employee_ID	Employee_Name	Gender	Department	Salary_In_lacs
2365	Rajesh Kulkarni	Male	Finance	650000

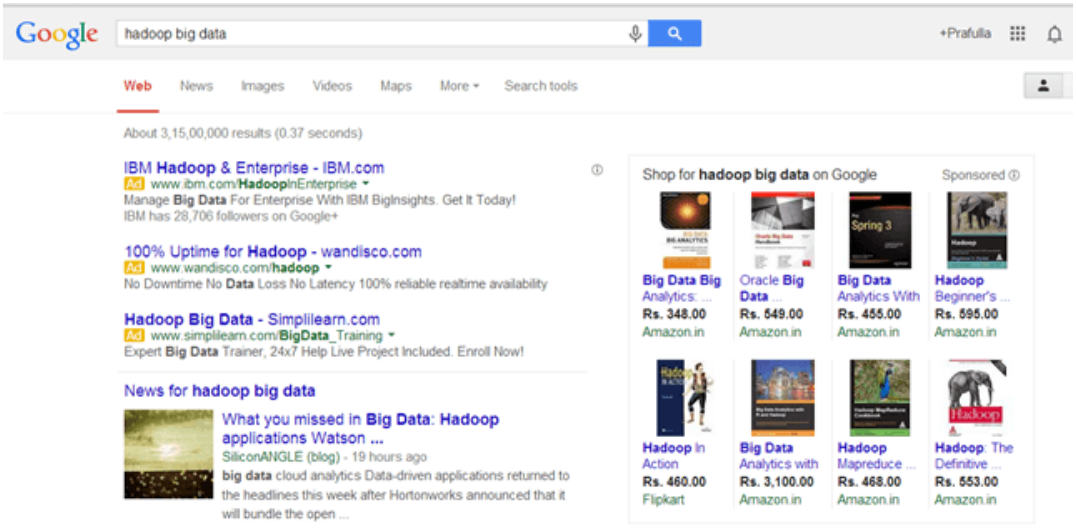
3398	Pratibha Joshi	Female	Admin	650000
7465	Shushil Roy	Male	Admin	500000
7500	Shubhojit Das	Male	Finance	500000
7699	Priya Sane	Female	Finance	550000

Unstructured

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.

Examples of Un-structured Data

The output returned by 'Google Search'



Semi-structured

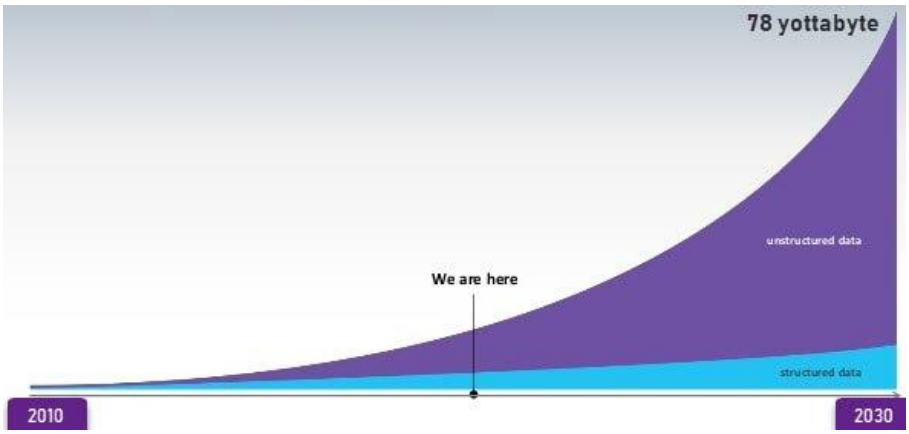
Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS. Example of semi-structured data is a data represented in an XML file.

Examples of Semi-structured Data

Personal data stored in an XML file-

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

Data Growth over the years



Please note that web application data, which is unstructured, consists of log files, transaction history files etc. OLTP systems are built to work with structured data wherein data is stored in relations (tables).

Introduction to Data Mining

"Drowning in Data yet Starving for Knowledge"???

"Computers have promised us a fountain of wisdom but delivered a flood of data"
William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus

"Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?"
T. S. Eliot

History of Data Mining

In the **1990s**, the term "Data Mining" was introduced, but data mining is the evolution of a sector with an extensive history.

Early techniques of identifying patterns in data include Bayes theorem (**1700s**), and the evolution of regression (**1800s**). The generation and growing power of computer science have boosted data collection, storage, and manipulation as data sets have broad in size and complexity level. Explicit hands-on data investigation has progressively been improved with indirect, automatic data processing, and other computer science discoveries such as neural networks, clustering, genetic algorithms (**1950s**), decision trees (**1960s**), and supporting vector machines (**1990s**).

Data mining origins are traced back to three family lines: Classical statistics, Artificial intelligence, and Machine learning.

Classical statistics:

Statistics are the basis of most technology on which data mining is built, such as regression analysis, standard deviation, standard distribution, standard variance, discriminatory analysis, cluster analysis, and confidence intervals. All of these are used to analyze data and data connection.

Artificial Intelligence:

AI or Artificial intelligence is based on heuristics as opposed to statistics. It tries to apply human- thought like processing to statistical problems. A specific AI concept was adopted by some high-end commercial products, such as query optimization modules for **Relational Database Management System(RDBMS)**.

Machine Learning:

Machine learning is a combination of statistics and AI. It might be considered as an evolution of AI because it mixes AI heuristics with complex statistical analysis. Machine learning tries to enable computer programs to know about the data they are studying so that programs make a distinct decision based on the characteristics of the data examined. It uses statistics for basic concepts and adding more AI heuristics and algorithms to accomplish its target.

What is NOT data mining?

Data Mining, noun: ***"Torturing data until it confesses ... and if you torture it enough, it will confess to anything"***

Jeff Jonas, IBM

"An Unethical Econometric practice of massaging and manipulating the data to obtain the desired results"

W.S. Brown "Introducing Econometrics"

"A buzz word for what used to be known as DBMS reports"

An Anonymous Data Mining Skeptic

What is data mining?

"The non trivial extraction of implicit, previously unknown, and potentially useful information from data"

William J Frawley, Gregory Piatetsky-Shapiro and Christopher J Matheus

- Data mining finds valuable information hidden in large volumes of data.
- Data mining is the analysis of data and the use of software techniques for finding patterns and regularities in sets of data.
- The computer is responsible for finding the patterns by identifying the underlying rules and features in the data.
- It is possible to "strike gold" in unexpected places as the data mining software extracts patterns not previously discernible or so obvious that no-one has noticed them before.
- Mining analogy:
    - large volumes of data are sifted in an attempt to find something worthwhile.
    - in a mining operation large amounts of low grade materials are sifted through in order to find something of value.

Data Mining - an interdisciplinary field

- Databases
- Statistics
- High Performance Computing
- Machine Learning
- Visualization
- Mathematics

Related Technologies

Machine Learning vs. Data Mining

- Large Data sets in Data Mining
- Efficiency of Algorithms is important
- Scalability of Algorithms is important
- Real World Data
- Lots of Missing Values
- Pre-existing data - not user generated

- Data not static - prone to updates
- Efficient methods for data retrieval available for use
- Domain Knowledge in the form of integrity constraints available.

Data Mining vs. DBMS

- Example DBMS Reports
  - Last months sales for each service type
  - Sales per service grouped by customer sex or age bracket
  - List of customers who lapsed their policy
- Questions answered using Data Mining
  - What characteristics do customers that lapse their policy have in common and how do they differ from customers who renew their policy?
  - Which motor insurance policy holders would be potential customers for my House Content Insurance policy?

Data Warehouse

- Data Warehouse: centralized data repository which can be queried for business benefit.
- Data Warehousing makes it possible to
  - extract archived operational data
  - overcome inconsistencies between different legacy data formats
  - integrate data throughout an enterprise, regardless of location, format, or communication requirements
  - incorporate additional or expert information

On-line Analytical Processing (OLAP)

- Multi-Dimensional Data Model (Data Cube)
- Operations:
  - Roll-up
  - Drill-down
  - Slice and dice
  - Rotate

Statistical Analysis

- Ill-suited for Nominal and Structured Data Types
- Completely data driven - incorporation of domain knowledge not possible
- Interpretation of results is difficult and daunting
- Requires expert user guidance

Data Mining Goals

Classification

- DM system learns from examples or the data how to partition or classify the data i.e. it formulates classification rules
- Example - customer database in a bank
  - Question - Is a new customer applying for a loan a good investment or not?
  - Typical rule formulated:

    if STATUS = married and INCOME > 10000 and HOUSE_OWNER = yes then INVESTMENT_TYPE = good

Association

- Rules that associate one attribute of a relation to another
- Set oriented approaches are the most efficient means of discovering such rules
- Example - supermarket database
    - 72% of all the records that contain items A and B also contain item C
    - the specific percentage of occurrences, 72 is the confidence factor of the rule

Sequence/Temporal

- Sequential pattern functions analyze collections of related records and detect frequently occurring patterns over a period of time
- Difference between sequence rules and other rules is the temporal factor
- Example - retailers database can be used to discover the set of purchases that frequently precedes the purchase of a microwave oven

Database management systems (DBMS), Online Analytical Processing (OLAP) and Data Mining

Area	**DBMS**	**OLAP**	**Data Mining**
Task	Extraction of detailed and summary data	Summaries, trends and forecasts	Knowledge discovery of hidden patterns and insights
Type of result	Information	Analysis	Insight and Prediction
Method	Deduction (Ask the question, verify with data)	Multidimensional data modeling, Aggregation, Statistics	Induction (Build the model, apply it to new data, get the result)
Example question	Who purchased mutual funds in the last 3 years?	What is the average income of mutual fund buyers by region by year?	Who will buy a mutual fund in the next 6 months and why?

Stages of the data mining process

- Data pre-processing
    - Heterogeneity resolution
    - Data cleansing
    - Data transformation
    - Data reduction
    - Discretization and generating concept hierarchies
- Creating a data model: applying Data Mining tools to extract knowledge from data
- Testing the model: the performance of the model (e.g. accuracy, completeness) is tested on independent data (not used to create the model)
- Interpretation and evaluation: the user bias can direct DM tools to areas of interest
    - Attributes of interest in databases
    - Goal of discovery
    - Domain knowledge
    - Prior knowledge or belief about the domain

Techniques

Set oriented database methods

- Statistics: can be used in several data mining stages
  - data cleansing: removal of erroneous or irrelevant data
  - EDA (exploratory data analysis): frequency counts, histograms etc.
  - data selection and sampling: reduce the scale of computation
  - attribute re-definition
  - Data analysis - measures of association and relationships between attributes, interestingness of rules, classification etc.
- Visualization: enhances EDA, makes patterns more visible
- Clustering (Cluster Analysis)
  - Clustering and segmentation is basically partitioning the database so that each partition or group is similar according to some criteria or metric
  - Clustering according to similarity is a concept which appears in many disciplines, e.g. clustering of molecules in chemistry
  - Data mining applications make use of clustering according to similarity, e.g. segmenting a client/customer base
  - It provides sub-groups of a population for further analysis or action - very important when dealing with very large databases

Knowledge Representation Methods

- Neural Networks
  - a trained neural network can be thought of as an "expert" in the category of information it has been given to analyze
  - provides projections given new situations of interest and answers "what if" questions
  - problems include:
    - the resulting network is viewed as a black box
    - no explanation of the results is given i.e. difficult for the user to interpret the results
    - difficult to incorporate user intervention
    - slow to train due to their iterative nature
- Decision trees
  - used to represent knowledge
  - built using a training set of data and can then be used to classify new objects
  - problems are:
    - opaque structure - difficult to understand
    - missing data can cause performance problems
    - they become cumbersome for large data sets
- Rules
  - probably the most common form of representation
  - tend to be simple and intuitive
  - unstructured and less rigid
  - problems are:
    - difficult to maintain
    - inadequate to represent many types of knowledge
  - Example format: if X then Y

Data Mining Applications

- Credit Assessment
- Stock Market Prediction
- Fault Diagnosis in Production Systems
- Medical Discovery

- Fraud Detection
- Hazard Forecasting
- Buying Trends Analysis
- Organizational Restructuring
- Target Mailing
- Knowledge Acquisition
- Scientific Discovery
- Semantics based Performance Enhancement of DBMS

- **Verification model:**
- It takes a **hypothesis from the user** and **tests the validity** of it against the data. The user has to come across the facts about the data using a number of techniques such as queries, multi-dimensional analysis and visualization to the guide to the exploration of the data being inspected.
- **Discovery model:**
- Knowledge discovery is the **concept of analyzing large amount of data** and **gathering out relevant information** leading to the knowledge discovery process for extracting meaningful rules, patterns and models from data.

Data Mining

- Data mining is the process of **extracting the useful information**, which is stored in the large database.
- It is a powerful tool, which is useful for organizations to retrieve the useful information from available data warehouses.
- Data mining can be applied to relational databases, object-oriented databases, data warehouses, structured-unstructured databases, etc.
- Data mining is used in numerous areas like banking, insurance companies, pharmaceutical companies etc.

Patterns in Data Mining

1. Association

The items or objects in relational databases, transactional databases or any other information repositories are considered, while finding **associations or correlations.**

2. Classification

- The goal of classification is to construct a model with the help of historical data that can accurately predict the value.
- It maps the data into the predefined groups or classes and searches for the new patterns.

For example:
To predict weather on a particular day will be categorized into - sunny, rainy, or cloudy.

3. Regression

- Regression creates predictive models. Regression analysis is used to make predictions based on existing data by applying formulas.
- Regression is very useful for finding (or predicting) the information on the basis of previously known information.

4. Cluster analysis

- It is a process of portioning a set of data into a set of meaningful subclass, called as cluster.
- It is used to place the data elements into the related groups without advanced knowledge of the group definitions.

5. Forecasting
Forecasting is concerned with the discovery of knowledge or information patterns in data that can lead to reasonable predictions about the future.

Technologies used in data mining

Several techniques used in the development of data mining methods. Some of them are mentioned below:
1. Statistics:

It uses the mathematical analysis to express representations, model and summarize empirical data or real world observations.

Statistical analysis involves the collection of methods, applicable to large amount of data to conclude and report the trend.

2. Machine learning

Arthur Samuel defined machine learning as a field of study that gives computers the ability to learn without being programmed.

When the new data is entered in the computer, algorithms help the data to grow or change due to machine learning.

In machine learning, an algorithm is constructed to predict the data from the available database (Predictive analysis).

It is related to computational statistics.

The four types of machine learning are:

1. Supervised learning

It is based on the classification.

It is also called as inductive learning. In this method, the desired outputs are included in the training dataset.

2. Unsupervised learning
unsupervised learning is based on clustering. Clusters are formed on the basis of similarity measures and desired outputs are not included in the training dataset.

3. Semi-supervised learning
Semi-supervised learning includes some desired outputs to the training dataset to generate the appropriate functions. This method generally avoids the large number of labeled examples (i.e. desired outputs) .

4. Active learning

Active learning is a powerful approach in analyzing the data efficiently.

The algorithm is designed in such a way that, the desired output should be decided by the algorithm itself (the user plays important role in this type).

3. Information retrieval

Information deals with uncertain representations of the semantics of objects (text, images). For example: Finding relevant information from a large document.

4. Database systems and data warehouse

Databases are used for the purpose of recording the data as well as data warehousing.

Online Transactional Processing (OLTP) uses databases for day to day transaction purpose.

To remove the redundant data and save the storage space, data is normalized and stored in the form of tables.

Entity-Relational modeling techniques are used for relational database management system design.

Data warehouses are used to store historical data which helps to take strategical decision for business.

It is used for online analytical processing (OALP), which helps to analyze the data.

5. Decision support system

Decision support system is a category of information system. It is very useful in decision making for organizations.

It is an interactive software based system which helps decision makers to extract useful information from the data, documents to make the decision.

KDD and Data mining

The process of discovering knowledge in data and application of data mining techniques are referred to as Knowledge Discovery in Databases (KDD).

KDD consists of various application domains such as artificial intelligence, pattern recognition, machine learning and data visualization.

The main goal of KDD is to extract knowledge from large databases with the help of data mining methods.

The different steps of KDD are as given below:

1. Data cleaning:
In this step, noise and irrelevant data are removed from the database.

2. Data integration:
In this step, the heterogeneous data sources are merged into a single data source.

3. Data selection:
In this step, the data which is relevant to the analysis process gets retrieved from the database.

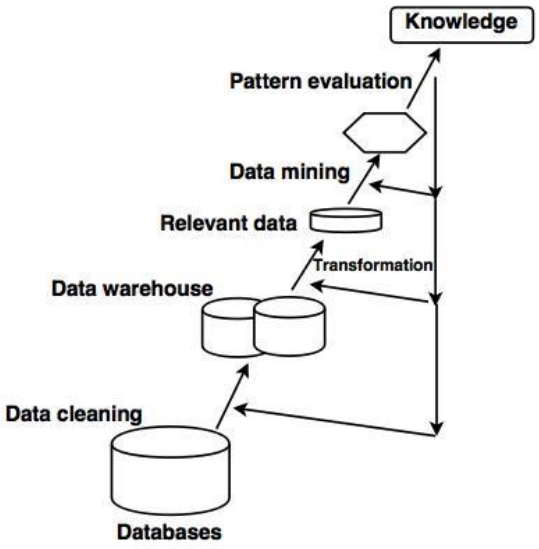
4. Data transformation:
In this step, the selected data is transformed in such forms which are suitable for data mining.

5. Data mining:

In this step, the various techniques are applied to extract the data patterns.

6. Pattern evaluation:
In this step, the different data patterns are evaluated.

7. Knowledge representation:
This is the final step of KDD, which represents the knowledge.



Knowledge Discovery in Databases (KDD)

KDD vs Datamining

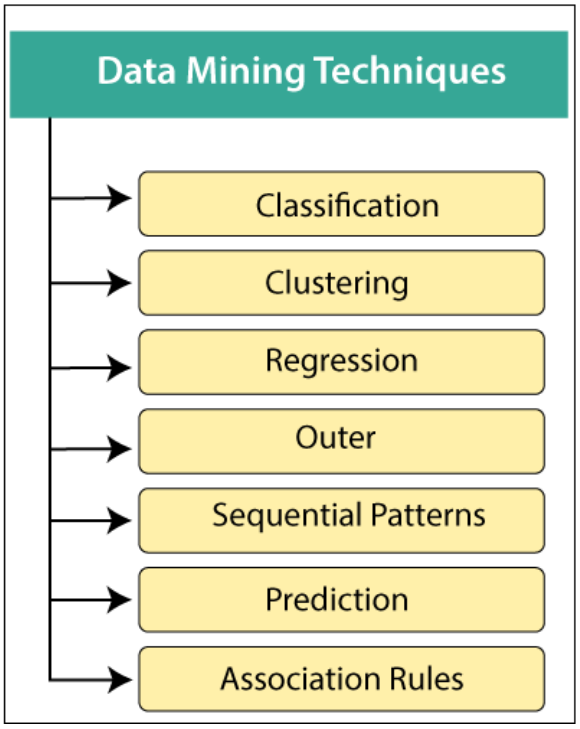
KDD	Data Mining
It is a field of computer science, that helps to extract useful and previously undiscovered knowledge from the large database by using various tools and theories.	**Data mining is one of the important steps in the KDD process.** It includes suitable algorithm based on the objective of the KDD process to identify the patterns from the database.

Data Mining Techniques

Data mining includes the utilization of refined data analysis tools to find previously unknown, valid patterns and relationships in huge data sets. These tools can incorporate statistical models, machine learning techniques, and mathematical algorithms, such as neural networks or decision trees. Thus, data mining incorporates analysis and prediction.

Depending on various methods and technologies from the intersection of machine learning, database management, and statistics, professionals in data mining have devoted their careers to better understanding how to process and make conclusions from the huge amount of data, but what are the methods they use to make it happen?

In recent data mining projects, various major data mining techniques have been developed and used, including association, classification, clustering, prediction, sequential patterns, and regression.



1. Classification:

This technique is used to obtain important and relevant information about data and metadata. This data mining technique helps to classify data in different classes.

Data mining techniques can be classified by different criteria, as follows:

i. **Classification of Data mining frameworks as per the type of data sources mined:**
This classification is as per the type of data handled. For example, multimedia, spatial data, text data, time-series data, World Wide Web, and so on..

ii. **Classification of data mining frameworks as per the database involved:**
This classification based on the data model involved. For example. Object-oriented database, transactional database, relational database, and so on..

iii. **Classification of data mining frameworks as per the kind of knowledge discovered:**
This classification depends on the types of knowledge discovered or data mining functionalities. For example, discrimination, classification, clustering, characterization, etc. some frameworks tend to be extensive frameworks offering a few data mining functionalities together..

iv. **Classification of data mining frameworks according to data mining techniques used:**
This classification is as per the data analysis approach utilized, such as neural networks, machine learning, genetic algorithms, visualization, statistics, data warehouse-oriented or database-oriented, etc.
The classification can also take into account, the level of user interaction involved in the data mining procedure, such as query-driven systems, autonomous systems, or interactive exploratory systems.

2. Clustering:

Clustering is a division of information into groups of connected objects. Describing the data by a few clusters mainly loses certain confine details, but accomplishes improvement. It models data by its clusters. Data modelling puts clustering from a historical point of view rooted in statistics, mathematics, and numerical analysis. From a machine learning point of view, clusters relate to hidden patterns, the search for clusters is unsupervised learning, and the subsequent framework represents a data concept. From a practical point of view,

clustering plays an extraordinary job in data mining applications. For example, scientific data exploration, text mining, information retrieval, spatial database applications, CRM, Web analysis, computational biology, medical diagnostics, and much more.

In other words, we can say that Clustering analysis is a data mining technique to identify similar data. This technique helps to recognize the differences and similarities between the data. Clustering is very similar to the classification, but it involves grouping chunks of data together based on their similarities.

3. Regression:

Regression analysis is the data mining process is used to identify and analyze the relationship between variables because of the presence of the other factor. It is used to define the probability of the specific variable. Regression, primarily a form of planning and modeling. For example, we might use it to project certain costs, depending on other factors such as availability, consumer demand, and competition. Primarily it gives the exact relationship between two or more variables in the given data set.

4. Association Rules:

This data mining technique helps to discover a link between two or more items. It finds a hidden pattern in the data set.
Association rules are if-then statements that support to show the probability of interactions between data items within large data sets in different types of databases. Association rule mining has several applications and is commonly used to help sales correlations in data or medical data sets.
The way the algorithm works is that you have various data, For example, a list of grocery items that you have been buying for the last six months. It calculates a percentage of items being purchased together.
These are three major measurements technique:

- **Lift:**
  This measurement technique measures the accuracy of the confidence over how often item B is purchased.
  (Confidence) / (item B)/ (Entire dataset)
- **Support:**
  This measurement technique measures how often multiple items are purchased and compared it to the overall dataset.
  (Item A + Item B) / (Entire dataset)
- **Confidence:**
  This measurement technique measures how often item B is purchased when item A is purchased as well.
  (Item A + Item B)/ (Item A)

5. Outer detection:

This type of data mining technique relates to the observation of data items in the data set, which do not match an expected pattern or expected behavior. This technique may be used in various domains like intrusion, detection, fraud detection, etc. It is also known as Outlier Analysis or Outilier mining. The outlier is a data point that diverges too much from the rest of the dataset. The majority of the real-world datasets have an outlier. Outlier detection plays a significant role in the data mining field. Outlier detection is valuable in numerous fields like network interruption identification, credit or debit card fraud detection, detecting outlying in wireless sensor network data, etc.

6. Sequential Patterns:

The sequential pattern is a data mining technique specialized for **evaluating sequential data** to discover sequential patterns. It comprises of finding interesting subsequences in a set of sequences, where the stake of a sequence can be measured in terms of different criteria like length, occurrence frequency, etc.

In other words, this technique of data mining helps to discover or recognize similar patterns in transaction data over some time.

7. Prediction:

Prediction used a combination of other data mining techniques such as trends, clustering, classification, etc. It analyzes past events or instances in the right sequence to predict a future event.

Benefits of Data Mining:

- Data mining technique helps companies to get knowledge-based information.
- Data mining helps organizations to make the profitable adjustments in operation and production.
- The data mining is a cost-effective and efficient solution compared to other statistical data applications.
- Data mining helps with the decision-making process.
- Facilitates automated prediction of trends and behaviors as well as automated discovery of hidden patterns.
- It can be implemented in new systems as well as existing platforms
- It is the speedy process which makes it easy for the users to analyze huge amount of data in less time.

Data Mining Tools

Following are 2 popular Data Mining Tools widely used in Industry

R-language:

R language is an open source tool for statistical computing and graphics. R has a wide variety of statistical, classical statistical tests, time-series analysis, classification and graphical techniques. It offers effective data handing and storage facility.

Oracle Data Mining:

Oracle Data Mining popularly knowns as ODM is a module of the Oracle Advanced Analytics Database. This Data mining tool allows data analysts to generate detailed insights and makes predictions. It helps predict customer behavior, develops customer profiles, identifies cross-selling opportunities.