UNIT -1: DATA ANALYTICS

Introduction to Data Analytics

- The term data analytics refers to the process of examining datasets to draw conclusions about the information they contain. Data analytic techniques enable you to take raw data and uncover patterns to extract valuable insights from it.
- Data Analytics refers to the techniques used to analyse data to enhance productivity and business gain. It is the discovery and communication of meaningful patterns in data. Data Analytics technique can help in finding the trends and metrics that would be used to optimize processes to increase the overall efficiency of a business or system.
- For example, manufacturing companies often record the runtime, downtime, and work queue for various machines and then analyse the data to better plan the workloads so the machines operate closer to peak capacity.
- The process involved in data analysis involves several different steps:

1. The first step is to determine the data requirements or how the data is grouped. Data may be separated by age, demographic, income, or gender. Data values may be numerical or be divided by category.
2. The second step in data analytics is the process of collecting it. This can be done through a variety of sources such as computers, online sources, cameras, environmental sources, or through personnel.
3. Once the data is collected, it must be organized so it can be analysed. Organization may take place on a spreadsheet or other form of software that can take statistical data.
4. The data is then cleaned up before analysis. This means it is scrubbed and checked to ensure there is no duplication or error, and that it is not incomplete. This step helps correct any errors before it goes on to a data analyst to be analysed.

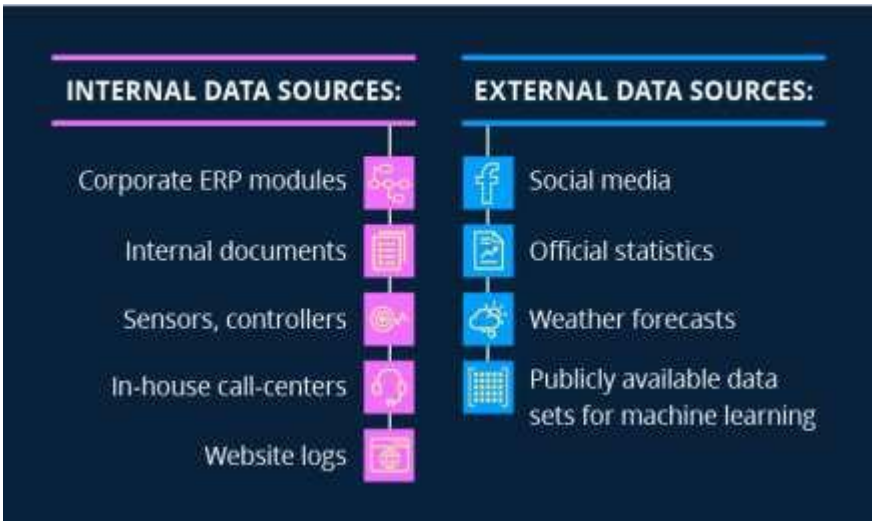
Types of Data Analytics

Data analytics is broken down into four basic types.

1. **Descriptive analytics** describes what has happened over a given period of time. Have the number of views gone up? Are sales stronger this month than last?
2. **Diagnostic analytics** focuses more on why something happened. This involves more diverse data inputs and a bit of hypothesizing. Did the weather affect beer sales? Did that latest marketing campaign impact sales?
3. **Predictive analytics** moves to what is likely going to happen in the near term. What happened to sales the last time we had a hot summer? How many weather models predict a hot summer this year?
4. **Prescriptive analytics** suggests a course of action. If the likelihood of a hot summer is measured as an average of these five weather models is above 58%, we should add an evening shift to the brewery and rent an additional tank to increase output.

Source of Data/ Big Data

- Data collection is the process of acquiring, collecting, extracting, and storing the voluminous amount of data which may be in the structured or unstructured form like text, video, audio, XML files, records, or other image files used in later stages of data analysis.



Sources of Data /Big Data

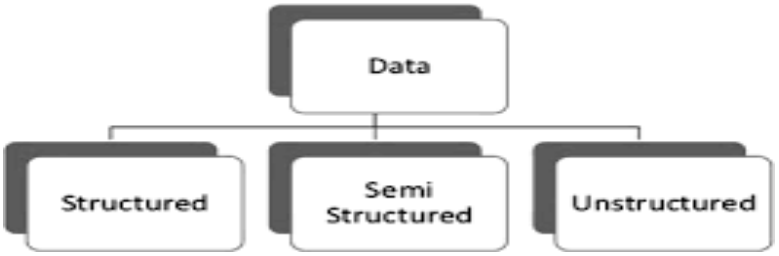
1. **Sensors/meters and activity records from electronic devices :** These kind of information is produced on real-time, the number and periodicity of observations of the observations will be variable, sometimes it will depend of a lap of time, on others of the occurrence of some event (per example a car passing by the vision angle of a camera) and in others will depend of manual manipulation (from an strict point of view it will be the same that the occurrence of an event)
2. **Social interactions:** The most common is the data produced in social networks. This kind of data implies qualitative and quantitative aspects which are of some interest to be measured. Quantitative aspects are easier to measure than qualitative aspects, first ones implies counting number of observations grouped by geographical or temporal characteristics, while the quality of the second ones mostly relies on the accuracy of the algorithms applied to extract the meaning of the contents which are commonly found as unstructured text written in natural language, examples of analysis that are made from this data are sentiment analysis, trend topics analysis, etc.
3. **Business transactions:** Data produced as a result of business activities can be recorded in structured or unstructured databases. When recorded on structured data bases the most common problem to analyze that information and get statistical indicators is the big volume of information and the periodicity of its production because sometimes these data is produced at a very fast pace, thousands of records can be produced in a second when big companies like supermarket chains are recording their sales.
4. **Electronic Files :** These refers to unstructured documents, statically or dynamically produced which are stored or published as electronic files, like Internet pages, videos, audios, PDF files, etc. They can have contents of special interest but are difficult to extract, different techniques could be used, like text mining, pattern recognition, and so

on. Quality of our measurements will mostly rely on the capacity to extract and correctly interpret all the representative information from those documents.

5. **Broadcastings:** Mainly referred to video and audio produced on real time, getting statistical data from the contents of this kind of electronic data by now is too complex and implies big computational and communications power, once solved the problems of converting "digital-analog" contents to "digital-data" contents we will have similar complications to process it like the ones that we can find on social interactions.

Classification of Data

Data classification is the process of organizing structured and unstructured data into defined categories that represent different types of data.

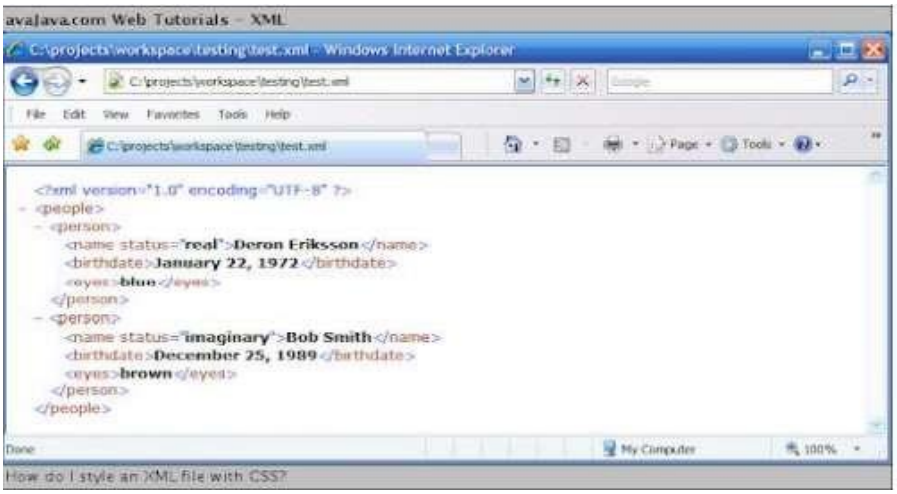


1. **Structured Data**: Structured data is information that has been formatted and transformed into a well-defined data model. Structured data is data with a high degree of organization, usually stored in some sort of spreadsheet. An example of structured data is shown: customer data of Your Model Car, using a spreadsheet (tabular form in rows and columns).

CUSTOMER

CUSTOMER_ID	LAST_NAME	FIRST_NAME	STREET	CITY	ZIP_CODE	COUNTRY
10302	Boucher	Leo	54, rue Royale	Nantes	44000	France
11244	Smith	Laurent	8489 Strong St	Las Vegas	83030	USA
11405	Han	James	636 St Kilda Road	Sydney	3004	Australia
11993	Mueller	Tomas	Berliner Weg 15	Tamm	71732	Germany
12111	Carter	Nataly	5 Tomahawk	Los Angeles	90006	USA
14121	Cortez	Nola	Av. Grande, 86	Madrid	28034	Spain
14400	Brown	Frank	165 S 7th St	Chester	33134	USA
14578	Wilson	Sarah	Seestreet #6101	Emory	1734	USA
14622	Jones	John	71 San Diego Ave	Arlington	69004	USA

2. **Semi Structured Data**: Semi-structured data is a type of data that has some consistent and definite characteristics, it does not confine into a rigid structure such as that needed for relational databases. Semi-structured is data which has some degree of organization in it. Example: Hypertext Mark-up language (HTML) files, JSON files.
   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 XML file.



3. **Unstructured Data**: Unstructured data is data with no pre-defined organizational form or specific format. Unstructured data is data that doesn't fit in a spreadsheet with rows and columns. Examples of unstructured data includes things like video, audio or image files, as well as log files, sensor or social media posts.

Properties	Structured data	Semi-structured data	Unstructured data
Technology	It is based on Relational database table	It is based on XML/RDF (Resource Description Framework).	It is based on character and binary data
Transaction management	Matured transaction and various concurrency techniques	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	Versioning over tuples,row,tables	Versioning over tuples or graph is possible	Versioned as a whole
Flexibility	It is schema dependent and less flexible	It is more flexible than structured data but less flexible than unstructured data	It is more flexible and there is absence of schema
Scalability	It is very difficult to scale DB schema	It's scaling is simpler than structured data	It is more scalable.
Robustness	Very robust	New technology, not very spread	—
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual queries are possible

Characteristics of Data/ Big Data

Gartner analyst Doug Laney listed the **3 'V's of Big Data – Variety, Velocity, and Volume.**

1. Variety

Variety of Big Data refers to structured, unstructured, and semi structured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more. Variety is one of the important characteristics of big data.

2. Velocity

Velocity essentially refers to the speed at which data is being created in real-time. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts.

3. Volume

Volume is one of the characteristics of big data. Big Data indicates huge 'volumes' of data that is being generated on a daily basis from various sources like social media platforms, business processes, machines, networks, human interactions, etc. Such a large amount of data are stored in data warehouses.

4. Veracity

Veracity is concerned with uncertainty or inaccuracy of the data. In many cases the data will be inaccurate hence filtering and selecting the data which is actually needed is a complicated task. A lot of statistical and analytical process has to go for data cleaning for choosing intrinsic data for decision making.