Indroduction to Data

➢ Data

Data is a raw form of knowledge and, on its own, doesn't carry any significance or purpose. In other words, you have to interpret data for it to have meaning. Data can be simple—and may even seem useless until it is analyzed, organized, and interpreted.

▪ Differentiate between data and information

- Data is a collection of facts, while information puts those facts into context.
- While data is raw and unorganized, information is organized.
- Data points are individual and sometimes unrelated. Information maps out that data to provide a big-picture view of how it all fits together.
- Data, on its own, is meaningless. When it's analyzed and interpreted, it becomes meaningful information.
- Data does not depend on information; however, information depends on data.
- Data typically comes in the form of graphs, numbers, figures, or statistics. Information is typically presented through words, language, thoughts, and ideas.

➢ Big Data

Big data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

- Types of big data

- Structured
- Unstructured
- Semi-structured

Differentiate between structured Unstructured Semi-structured

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.

➢ Quantitative data and Qualitative data

- **Quantitative data :** Quantitative data are used when a researcher is trying to quantify a problem, or address the "what" or "how many" aspects of a research question. It is data that can either be counted or compared on a numeric scale. For example, it could be the number of first year students at Macalester, or the ratings on a scale of 1-4 of the quality of food served at Cafe Mac. This data are usually gathered using instruments, such as a questionnaire which includes a ratings scale or a thermometer to collect weather data. Statistical analysis software, such as SPSS, is often used to analyze quantitative data.

- **Qualitative data:** Qualitative data describes qualities or characteristics. It is collected using questionnaires, interviews, or observation, and frequently appears in narrative form. For example, it could be notes taken during a focus group on the quality of the food at Cafe Mac, or responses from an open-ended questionnaire. Qualitative data may be difficult to precisely measure and analyze. The data may be in the form of descriptive words that can be examined for patterns or meaning, sometimes through the use of coding. Coding allows the researcher to categorize qualitative data to identify themes that correspond with the research questions and to perform quantitative analysis.

➢ Different V's in big data

Big data is a collection of data from many different sources and is often describe by five characteristics: volume, value, variety, velocity, and veracity.

- **Volume:** the size and amounts of big data that companies manage and analyse

- **Value:** the most important "V" from the perspective of the business, the value of big data usually comes from insight discovery and pattern recognition that lead to more effective operations, stronger customer relationships and other clear and quantifiable business benefits
- **Variety:** the diversity and range of different data types, including unstructured data, semi-structured data and raw data
- **Velocity:** the speed at which companies receive, store and manage data – e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time
- **Veracity:** the "truth" or accuracy of data and information assets, which often determines executive-level confidence

➢ Tools used in big data

- Apache Spark
- Apache Hadoop
- Google cloud platform
- MongoDB
- Sisense
- RapidMiner

➢ Different types of data

The data is classified into four categories:

- Nominal data
- Ordinal data
- Discrete data
- Continuous data

Nominal data: Nominal Data is used to label variables without any order or quantitative value. The color of hair can be considered nominal data, as one color can't be compared with another color.

Ordinal data: Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale. These data are used for observation like customer satisfaction, happiness, etc., but we can't do any arithmetical tasks on them.

Discrete data: The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers. The total number of students in a class is an example of discrete data. These data can't be broken into decimal or fraction values.

Continuous data: Continuous data are in the form of fractional numbers. It can be the version of an android phone, the height of a person, the length of an object, etc. Continuous data represents information that can be divided into smaller levels. The continuous variable can take any value within a range.