Introduction To Data

Assignment 1

1. Difference between data and Information

- Data is a raw and unorganized fact that is required to be processed to make it meaningful whereas Information is a set of data that is processed in a meaningful way according to the given requirement.
- Data does not have any specific purpose whereas Information carries a meaning that has been assigned by interpreting data.
- Data alone has no significance while Information is significant by itself.
- Data never depends on Information while Information is dependent on Data.
- Data measured in bits and bytes, on the other hand, Information is measured in meaningful units like time, quantity, etc.
- Data can be structured, tabular data, graph, data tree whereas Information is language, ideas, and thoughts based on the given data.

2. How data is useful to us ?

Data = Knowledge. Good data provides indisputable evidence, while anecdotal evidence, assumptions, or abstract observation might lead to wasted resources due to taking action based on an incorrect conclusion.

Most Compelling Benefits of Data and Analytics

1. Customer Acquisition and Retention
2. Focused and Targeted Promotions
3. Potential Risks Identification
4. Innovate
5. Complex Supplier Networks
6. Cost optimization
7. Improve Efficiency.

3. What is Big data

Big data refers to the large, diverse sets of information that grow at ever-increasing rates. It encompasses the volume of information, the velocity or speed at which it is created and collected, and the variety or scope of the data points being covered.

4. Differences between Structured, Semi-structured and Unstructured data:

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

5. What is the difference between quantitative and qualitative data

Qualitative data	Quantitative data
Can't be measured.	Can be quantified and is measurable.
Can be quantified and is measurable.	The data is expressed as numbers and values.
The data describes qualities or characteristics.	The data is statistical and structured.
The data is nonstatistical and unstructured.	The data answers the questions "how much," "how many," or "how often"

Qualitative data	Quantitative data
The data can be collected using questionnaires, interviews, focus groups, or observation.	The data can be collected through instruments, tests, experiments, surveys, market reports, and metrics.
Examples include a person's name, hair color, and occupation.	Examples include age, height, and the number of visitors a website gets per day.

6. The 5 V's of Big Data

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

7. Most Popular Big Data Analytics Tools

- Apache Hadoop
- Apache Spark
- Flink
- Apache Storm
- Apache Cassandra
- MongoDB
- Kafka
- Tableau
- RapidMiner
- R Programming

8. Different Types of Data

Structured data

Structured data adheres to a pre-defined data model. This model describes how data is recorded, and it defines the attributes and provides information about the data type (e.g. name, date, number) and restrictions on their values (e.g. number of characters). This level of organisation means that data can be entered, stored, queried, or analysed by machines.

Structured data includes:

- names
- dates
- phone numbers

Unstructured and semi-structured data

Unlike structured data, unstructured data requires human interpretation. Consider a block of text. Computers can read each word, or sentence, but they can't (yet) determine the meaning or tone of the text without human intervention. As you'll discover later in the course, data scientists are trying to solve this problem with machine learning and other types of artificial intelligence.

Other examples of unstructured data include:

- images (human- and machine-generated)
- video files
- audio files
- social-media posts