Name: DHWANI GUPTA

Enrollment Number: EBEON0223761985

Introduction to Data Theory Questions

1.Differentiate between data and information.

- The difference between the date and information is shown below in the table: -

Definition	**DATA**	**INFORMATION**
	Data is collection of **facts** that has been translated into a form that computer can process. Date can be a Number, Observations, Words, Measurements.	When data are processed, organized, structured, or presented in a given context so as to make them useful, they are called "information "**DATA** presented in useful manner"
Usage	Raw data alone is insufficient for decision making and Data doesn't depend on information. **Organized data** is presented in useful manner to get information.	Information is **sufficient** for decision making. Information **depends** on data.
Example: Global Temperature or Sales Experience	The history of temperature **readings** all over the world for the **past 100 years** is data. Customer feedback is a form of data.	If this **data is organized** and **analyzed** to find that global temperature is rising, then that is information. After Analyzing the data, the observations and conclusions are information like Site, Sales person review.

2.How data is useful for us?

Data = Knowledge. Data Enables Better Decision Making. Good Data Will Also Give You the Justification and Evidence You Need to Back Up These Decisions So That You Can Feel Confident Explaining Your Reasoning Going Forward. Without Solid Data, You're Much More Likely to Make Mistakes and Reach Incorrect Conclusions.

DATA ⟶ INFORMATION ⟶ INSIGHTS

Data Is Used to Predicted Future, Solving Problem and Making Strategies.

OLA Case Study

Ola Used Data to Improve Their Customer Satisfaction.

Ola Basically Has Millions and Billions of Bytes of Data Consisting of Their Drivers, Customers and Also Has Information About Every Single Trip That Takes Place on Its Platform. So, They Basically Have Insights of Time & Day of Booking, Pick-Up & Drop-Up Locations, And Much More. However, Containing So Much Data with Them Ola's Business- Deeply Rooted in Data Science. Data Science Helps Ola to Understand the Kind of Aura A Traveler Prefers in The Cab.

Eg:- Infotainment Preference, Playlist Preference Etc.

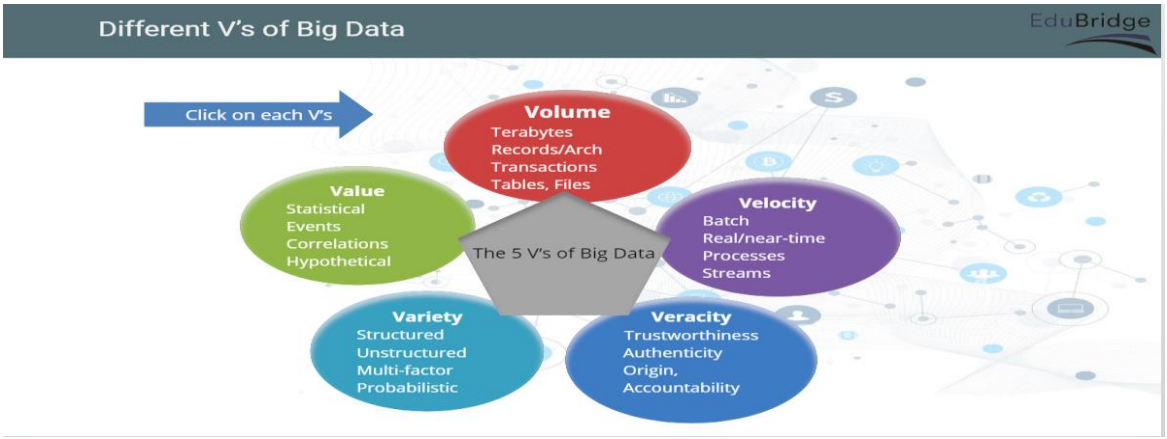
Ola Has Collected Data from Approximately "**Nine Lakh**" Cabs on The Road Which Helps Them to Create an Accurate Real-Time Scenario of Traffic Around the Roads. The Information of The Routes Around the Road and They Use This Data to Create Unique **ETA (Estimated Time Arrival)** Models. **It Suggest Less Congested Routes to Magnify the Commuting Experience.**

OTTs Platforms

OTTs platforms also used data to marks a presence in the market. Now a days OTTs are giving tuff competition to the cinemas and TV. According to a report released by TechJury, the OTT streaming industry has shown a hike of around 99% between April 2019-April 2020. It is predicted that by 2027, the Over-the-Top streaming industry will be worth 184.27 billion USD.

3.What is big data?

Big data refers to large and complex datasets that are beyond the processing capabilities of traditional data processing tools and techniques. It refers to the immense volume, velocity, and variety of data that organizations collect and analyze to gain valuable insights and make informed decisions.



4.Differentiate between structured, semi-structured and unstructured data?

- The difference between the Structured, Semi Structured and Unstructured Data is shown below in the table: -

Parameters	STRUCTURED DATA	SEMI STRUCTURED DATA	UNSTRUCTURED DATA
Definition	Structured data is essentially data that conforms to a data model, has a well-defined structure, and follows a consistent order. It should also be easily accessible and/or usable by people or computer programs.	Semi-structured data combine unstructured and structured data because it contains elements of both.	Unstructured data is essentially everything else. Unstructured data has an internal structure but is not structured via predefined data models or schema. It may be textual or non-textual, and human- or machine-generated.
Advantages	Easy to access and store, indexed feature, data mining becomes easier, easy to update and delete anything	Flexibility, Portability, Ease of use	More flexibility, offers more insights, quickly accumulated
Disadvantages	Limited usage, Limited storage options, Increased complexity, Reduced flexibility, Increased costs	More challenging to query, more challenging to analyze, less reliable	Time-consuming and expensive, Difficult to analyze, requires specific tools, Hard to store
Tools/Examples	Excel files or SQL databases	HTML code, graphs and tables, e-mails, XML documents	Word documents, email messages, PowerPoint presentations, survey responses, transcripts of call center interactions and posts from blogs and social media sites.

5.What are quantitative data and qualitative data?

- The difference between **quantitative data and qualitative data** is shown below in the table: -

Parameters	Quantitative Data	Qualitative Data
Definition	Quantitative data is data that can be counted or measured in numerical values. The two main types of quantitative data are discrete data and continuous data. Height in feet, age in years, and weight in pounds are examples of quantitative data. Qualitative data is descriptive data that is not expressed numerically.	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.
Purpose	Answer "How many/much?" question	Answer "Why?" question
Data type	Number/Statistical result	Observation, Symbol, Word etc.
Approach	Measure and test	Observe and interpret
Analysis	Statistical analysis	Grouping of common data /non-statistical analysis

6.What are the different V's in big data?

There are 5 different V's in big data.

I. **Volume**: The name "big data" implies that it can store a large amount of data. The volume of data determines whether it is classified as big data or not. As a result, volume is crucial. Volume refers to the amount/quantity of data created, such as Every hour, Walmart customers' transactions generate approximately 2.5 petabytes of data for the company.

II. **Velocity**: Velocity refers to the rate at which data moves, such as how Facebook users send an average of 31.25 million messages and watch 2.77 million videos every minute of every day over the internet. Speed is always a deciding factor in any sport. To meet user demand, how quickly and spontaneously data is generated and processed will determine the data's true potential. In big data, the flow of data will be massive and continuous.

III. **Variety**: The term "variety" refers to the various types of data that can be created, such as structured, semi-structured, and unstructured data. Unstructured data includes things like sending emails with attachments through Gmail, as well as posting comments with external links. Pictures, audio clips, and video clips are all examples of unstructured data.

IV. **Veracity**: Veracity is the quality or trustworthiness of the data. Just how accurate is all this data? A lot of data and a big variety of data with fast access are not enough. The data must have quality and produce credible results that enable right action when it comes to end of life decision making.

V. **Value**: Value in Big Data Refers to the worth of the data being extracted. Having endless amounts of data is one thing, but unless it can be turned into value it is useless. While there is a clear link between data and insights, this does not always mean there is value in Big Data.

7.Name some popular tools used in big data.

There are several popular tools used in big data processing and analysis. Here are some of them:

i. **Hadoop**: Hadoop is an open-source framework that enables distributed processing of large datasets across clusters of computers. It includes the Hadoop Distributed File System (HDFS) for storage and the MapReduce programming model for processing.

ii. **Apache Spark**: Apache Spark is a fast and general-purpose cluster computing system that provides in-memory processing capabilities. It supports a wide range of data processing tasks, including batch processing, streaming, machine learning, and graph processing.

iii. **Apache Kafka**: Apache Kafka is a distributed streaming platform used for building real-time data pipelines and streaming applications. It is designed to handle high-throughput, fault-tolerant, and scalable data streaming.

iv. **Apache Flink**: Apache Flink is an open-source stream processing framework with powerful event-time processing and state management capabilities. It supports both batch and stream processing, and it's known for its low-latency and high-throughput processing.

v. **Apache Hive**: Apache Hive is a data warehousing and SQL-like query language for big data built on top of Hadoop. It provides a mechanism to query and analyze large datasets stored in Hadoop's HDFS.

vi. **Apache HBase**: Apache HBase is a distributed, scalable, and consistent NoSQL database built on top of Hadoop. It is designed for random, real-time read/write access to large datasets.

vii. **Apache Cassandra**: Apache Cassandra is a highly scalable and distributed NoSQL database that can handle massive amounts of structured and unstructured data across many commodity servers. It offers high availability and fault tolerance.

viii. **Apache Storm**: Apache Storm is a distributed real-time stream processing system. It is used for processing high-velocity, real-time data streams and provides reliable and fault-tolerant stream processing capabilities.

ix. **Elasticsearch**: Elasticsearch is a real-time distributed search and analytics engine. It is often used for log analysis, full-text search, and real-time data analysis.

x. **Splunk**: Splunk is a popular software platform used for searching, monitoring, and analyzing machine-generated big data, including logs, events, and other types of unstructured data.