Q1: Difference between data vs information?

Data	Information
Data is unorganized and unrefined facts	Information comprises processed, organized data presented in a meaningful context
Data is an individual unit that contains raw materials which do not carry any specific meaning.	Information is a group of data that collectively carries a logical meaning.
Data doesn't depend on information.	Information depends on data.
Raw data alone is insufficient for decision making	Information is sufficient for decision making
An example of data is a student's test score	The average score of a class is the information derived from the given data.

Q2: How data is useful for 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.

Q3: what is big data?

Big data refers to data that is so large, fast or complex that it's difficult or impossible to process using traditional methods. The act of accessing and storing large amounts of information for analytics has been around for a long time.

Q4: differentiate between structured data, unstructured data and semi-structured data?

Big Data includes huge volume, high velocity, and extensible variety of data. These are 3 types: Structured data, Semi-structured data, and Unstructured data.

1. Structured data –
   Structured data is data whose elements are addressable for effective analysis. It has been organized into a formatted repository that is typically a database. It concerns all data which can be stored in database SQL in a table with rows and columns. They have relational keys and can easily be mapped into pre-designed fields. Today, those data

are most processed in the development and simplest way to manage information. Example: Relational data.

2. Semi-Structured data –

   Semi-structured data is information that does not reside in a relational database but that has some organizational properties that make it easier to analyze. With some processes, you can store them in the relation database (it could be very hard for some kind of semi-structured data), but Semi-structured exist to ease space. Example: XML data.

3. Unstructured data –

   Unstructured data is a data which is not organized in a predefined manner or does not have a predefined data model, thus it is not a good fit for a mainstream relational database. So for Unstructured data, there are alternative platforms for storing and managing, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications. Example: Word, PDF, Text, Media logs.

Q5: What are qualitative and quantitative data?

Quantitative data refers to any information that can be quantified, counted or measured, and given a numerical value. Qualitative data is descriptive in nature, expressed in terms of language rather than numerical values. Quantitative research is based on numeric data.

Q6: What is the 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.

Q7: Name some popular tools used in big data?

Some popular tools are:

Apache Spark.

Apache Hadoop.

Apache Flink.

Google Cloud Platform.

MongoDB.

Sisense.

RapidMiner.

Q8: what are different types of data ? Explain.

Data :-1: quantitative data

2: qualitative data

1:quantitative data:

Qualitative data is data that can be felt or described.

These are two types;

1: Discrete:-

Discrete data is quantitative data that has fixed numerical values incapable of breaking down into smaller parts. An example of discrete data would be the number of children a person has.

2: Continuous:-

Continuous data refers to data that can be measured. This data has values that are not fixed and have an infinite number of possible values. These measurements can also be broken down into smaller individual parts. Some examples of continuous data would include: The height or weight of a person.

These are two types:-

1:Interval:

Interval data, also called an integer, is defined as a data type which is measured along a scale, in which each point is placed at equal distance from one another. Interval data always appears in the form of numbers or numerical values where the distance between the two points is standardized and equal.

2:ratio:

Ratio data is a form of quantitative (numeric) data. It measures variables on a continuous scale, with an equal distance between adjacent values. While it shares these features with interval data (another type of quantitative data), a distinguishing property of ratio data is that it has a 'true zero.

Qualitative data:-

Qualitative data is information that cannot be counted, measured or easily expressed using numbers. It is collected from text, audio and images and shared through data visualization tools, such as word clouds, concept maps, graph databases, timelines and infographics.

These are two types:-

1:Nominal:

Nominal data is a type of qualitative data which groups variables into categories. You can think of these categories as nouns or labels; they are purely descriptive, they don't have any quantitative or numeric value, and the various categories cannot be placed into any kind of meaningful order or hierarchy.

2:Ordinal:

Ordinal data is a categorical, statistical data type where the variables have natural, ordered categories and the distances between the categories are not known. These data exist on an ordinal scale.

Q9: How is the statistical significance of an insight assessed?

Statistical significance is often calculated with statistical hypothesis testing, which tests the validity of a hypothesis by figuring out the probability that your results have happened by chance.

10: what is mean?

Mean is the average of the given numbers and is calculated by dividing the sum of given numbers by the total number of numbers.

Mean = sum of all observation/ total no of observation

11: what is the meaning of standard deviation?

A standard deviation (or σ) is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out.

12: what is correlation?

in statistical terms we use correlation to denote association between two quantitative variables. We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other.

13: what is the meaning of covariance?

Covariance is a statistical tool investors use to measure the relationship between the movement of two asset prices. A positive covariance means asset prices are moving in the same general direction. A negative covariance means asset prices are moving in opposite directions.

14: where is inferential statistics used?

Inferential statistics are often used to compare the differences between the treatment groups. Inferential statistics use measurements from the sample of subjects in the experiment to compare the treatment groups and make generalizations about the larger population of subjects.

15: what is one sample t-test?

The one-sample t-test is used when we want to know whether our sample comes from a particular population but we do not have full population information available to us. For instance, we may want to know if a particular sample of college students is similar to or different from college students in general.

16: what is the relationship between standard deviation and standard variance?

Standard deviation is the square root of the variance and is expressed in the same units as the data set.

17: what is one- way ANOVA test?

One-Way ANOVA ("analysis of variance") compares the means of two or more independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. One-Way ANOVA is a parametric test. This test is also known as: One-Factor ANOVA.