DS Questions

1. **Explain Amazon Web service in detail**

   Amazon web service is an online platform that provides scalable and cost-effective cloud computing solutions.

   AWS is a broadly adopted cloud platform that offers several on-demand operations like compute power, database storage, content delivery, etc., to help corporates scale and grow.

   Applications of AWS

   AWS enables businesses to build a number of sophisticated applications. Organizations of every industry and of every size, can run every imaginable use case on AWS. Here are some of the most common applications of AWS:-

   1. Storage and Backup

   One of the reasons why many businesses use AWS is because it offers multiple types of storage to choose from and is easily accessible as well. It can be used for storage and file indexing as well as to run critical business applications.

   2. Websites

   Businesses can host their websites on the AWS cloud, similar to other web applications.

   3. Gaming

   There is a lot of computing power needed to run gaming applications. AWS makes it easier to provide the best online gaming experience to gamers across the world.

   4. Mobile, Web and Social Applications

   A feature that separates AWS from other cloud services is its capability to launch and scale mobile, e-commerce, and SaaS applications. API-driven code on AWS can enable companies to build uncompromisingly scalable applications without requiring any OS and other systems.

Companies using AWS

Whether it’s technology giants, startups, government, food manufacturers or retail organizations, there are so many companies across the world using AWS to develop, deploy and host applications. According to Amazon, the number of active AWS users exceeds 1,000,000. Here is a list of companies using AWS:

- Netflix
- Intuit
- Coinbase
- Finra
- Johnson & Johnson
- Capital One
- Adobe
- Airbnb
- AOL

Advantages of AWS Services

The power of AWS services lies in the fact that it enables businesses to reach the marketplaces with little initial investment. Here are some advantages of AWS services:

1. Security

There is a false misconception that data stored in a public cloud is not secure. On the contrary, not only does AWS offer security tools that are cheaper than other alternatives, but it is one of the most secure, extensive, and reliable cloud platforms.

2. Global Availability

AWS has 80 Availability Zones across 25 geographic regions global data centers.

3**. Scalability and Flexibility**

AWS offers unlimited flexibility and scalability on demand. This enables organizations to plan their infrastructure roadmap on a subscription basis without full commitment.

AWS Services

Amazon has many services for cloud applications. Let us list down a few key services of the AWS ecosystem and a brief description of how developers use them in their business.

Amazon has a list of services:

- Compute service
- Storage
- Database
- Networking and delivery of content
- Security tools

2. **Difference between unstructed and semi structured**
3. **Difference between unstructed and and structure data**

Difference Between Structured, Semi-structured, and Unstructured Data

Parameters	Structured Data	Semi-Structured Data	Unstructured Data
Data Structure	The information and data have a predefined organization.	The contained data and information have organizational properties- but are different from predefined structured data.	There is no predefined organization for the available data and information in the system or database.
Technology Used	Structured Data words on the basis of relational database tables.	Semi-Structured Data works on the basis of Relational Data Framework (RDF) or XML.	Unstructured data works on the basis of binary data and the available characters.
Flexibility	The data depends a lot on the schema. Thus, there is less flexibility.	The data is comparatively less flexible than unstructured data but way more flexible than the structured data.	Schema is totally absent. Thus, it is the most flexible of all.
Management of Transaction	It has a mature type of transaction. Also, there are various techniques of concurrency.	It adapts the transaction from DBMS. It is not of mature type.	It consists of no management of transaction or concurrency.
Management of Version	It is possible to version over tables, rows, and tuples.	It is possible to version over graphs or tuples.	It is possible to version the data as a whole.
Robustness	Structured data is very robust in nature.	Semi-Structured Data is a fairly new technology. Thus, it is not very robust in nature.	–
Scalability	Scaling a database schema is very difficult. Thus, a structured database offers lower scalability.	Scaling a Semi-Structured type of data is comparatively much more feasible.	An unstructured data type is the most scalable in nature.
Performance of Query	A structured type of query makes complex joining possible.	Semi-structured queries over various nodes (anonymous) are most definitely possible.	Unstructured data only allows textual types of queries.

4. **Explain data cleaning and data extraction in detail**

   What is data cleaning?

   Data cleaning is the process of ensuring data is correct, consistent and usable. You can clean data by identifying errors or corruptions, correcting or deleting them, or manually processing data as needed to prevent the same errors from occurring.

   Most aspects of data cleaning can be done through the use of software tools, but a portion of it must be done manually. Although this can make data cleaning an overwhelming task, it is an essential part of managing company data.

   What are the benefits of data cleaning?

   There are many benefits to having clean data:

   1. It removes major errors and inconsistencies that are inevitable when multiple sources of data are being pulled into one dataset.
   2. Using tools to clean up data will make everyone on your team more efficient as you'll be able to quickly get what you need from the data available to you.
   3. Fewer errors mean happier customers and fewer frustrated employees.
   4. It allows you to map different data functions, and better understand what your data is intended to do, and learn where it is coming from.

   What is Data Extraction?

   Data extraction is the process of collecting or retrieving disparate types of data from a variety of sources, many of which may be poorly organized or completely unstructured. Data extraction makes it possible to consolidate, process, and refine data so that it can be stored in a centralized location in order to be transformed. These locations may be on-site, cloud-based, or a hybrid of the two.

5. **Explain data collection in detail**

   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.

   In the process of big data analysis, “Data collection” is the initial step before starting to analyze the patterns or useful information in data. The data which is to be analyzed must be collected from different valid sources.

   The data which is collected is known as raw data which is not useful now but on cleaning the impure and utilizing that data for further analysis forms information, the information obtained is known as “knowledge”.

   Data collection starts with asking some questions such as what type of data is to be collected and what is the source of collection.

   Most of the data collected are of two types known as “qualitative data“ which is a group of non-numerical data such as words, sentences mostly focus on behavior and actions of the group and another one is “quantitative data” which is in numerical forms and can be calculated using different scientific tools and sampling data.

6. **Explain eda in detail**

   Exploratory Data Analysis (EDA)

   Exploratory Data Analysis or EDA is used to take insights from the data. Data Scientists and Analysts try to find different patterns, relations, and anomalies in the data using some statistical graphs and other visualization techniques. Following things are part of EDA :

   1. Get maximum insights from a data set
   2. Uncover underlying structure
   3. Extract important variables from the dataset
   4. Detect outliers and anomalies (if any)

5. Test underlying assumptions
6. Determine the optimal factor settings
7. Getting a better understanding of data
8. Identifying various data patterns
9. Getting a better understanding of the problem statement

The main purpose of EDA is to detect any errors, outliers as well as to understand different patterns in the data. It allows Analysts to understand the data better before making any assumptions.

The outcomes of EDA help businesses to know their customers, expand their business and take decisions accordingly.

The basic uses of the Data Visualization technique are as follows:

- It is a powerful technique to explore the data with **presentable** and **interpretable** results.
- In the **data mining process**, it acts as a primary step in the pre-processing portion.
- It supports the **data cleaning process** by finding incorrect data and corrupted or missing values.
- It also helps to **construct and select variables**, which means we have to determine which variable to include and discard in the analysis.
- In the process of **Data Reduction**, it also plays a crucial role while combining the categories.

7. **Define data science and explain types of data**

   Data Science : is an interdisciplinary field that focuses on extracting knowledge from data sets which are typically huge in amount. The field encompasses analysis, preparing data for analysis, and presenting findings to inform high-level decisions in an organization. As such, it incorporates skills from computer science, mathematics, statics, information visualization, graphic, and business.

1. Quantitative data

Quantitative data seems to be the easiest to explain. It answers key questions such as “how many, “how much” and “how often”.

Quantitative data can be expressed as a number or can be quantified. Simply put, it can be measured by numerical variables.

Quantitative data are easily amenable to statistical manipulation and can be represented by a wide variety of statistical types of graphs and charts such as line, bar graph, scatter plot, and etc.

Examples of quantitative data:

- Scores on tests and exams e.g. 85, 67, 90 and etc.
- The weight of a person or a subject.
- Your shoe size.
- The temperature in a room.

There are 2 general types of quantitative data: discrete data and continuous data

2. Qualitative data

Qualitative data can’t be expressed as a number and can’t be measured. Qualitative data consist of words, pictures, and symbols, not numbers.

Qualitative data is also called categorical data because the information can be sorted by category, not by number.

Qualitative data can answer questions such as “how this has happened” or and “why this has happened”.

Examples of qualitative data:

- Colors e.g. the color of the sea
- Your favorite holiday destination such as Hawaii, New Zealand and etc.
- Names as John, Patricia
- Ethnicity such as American Indian, Asian, etc.

There are 2 general types of qualitative data: nominal data and ordinal data.

3. Nominal data

Nominal data is used just for labeling variables, without any type of quantitative value. The name ‘nominal’ comes from the Latin word “nomen” which means ‘name’.

The nominal data just name a thing without applying it to order. Actually, the nominal data could just be called “labels.”

Examples of Nominal Data:

- Gender (Women, Men)
- Hair color (Blonde, Brown, Brunette, Red, etc.)
- Marital status (Married, Single, Widowed)
- Ethnicity (Hispanic, Asian)

As you see from the examples there is no intrinsic ordering to the variables.

Eye color is a nominal variable having a few categories (Blue, Green, Brown) and there is no way to order these categories from highest to lowest.

4. Ordinal data

Ordinal data shows where a number is in order. This is the crucial difference from nominal types of data.

Ordinal data is data which is placed into some kind of order by their position on a scale. Ordinal data may indicate superiority.

However, you cannot do arithmetic with ordinal numbers because they only show sequence.

Ordinal variables are considered as “in between” qualitative and quantitative variables.

In other words, the ordinal data is qualitative data for which the values are ordered.

In comparison with nominal data, the second one is qualitative data for which the values cannot be placed in an ordered.

We can also assign numbers to ordinal data to show their relative position. But we cannot do math with those numbers. For example: “first, second, third…etc.”

Examples of Ordinal Data:

- The first, second and third person in a competition.

- Letter grades: A, B, C, and etc.
- When a company asks a customer to rate the sales experience on a scale of 1-10.
- Economic status: low, medium and high.

5. Discrete data

Discrete data is a count that involves only integers. The discrete values cannot be subdivided into parts.

For example, the number of children in a class is discrete data. You can count whole individuals. You can’t count 1.5 kids.

To put in other words, discrete data can take only certain values. The data variables cannot be divided into smaller parts.

It has a limited number of possible values e.g. days of the month.

Examples of discrete data:

- The number of students in a class.
- The number of workers in a company.
- The number of home runs in a baseball game.
- The number of test questions you answered correctly

6. Continuous data

Continuous data is information that could be meaningfully divided into finer levels. It can be measured on a scale or continuum and can have almost any numeric value.

For example, you can measure your height at very precise scales — meters, centimeters, millimeters and etc.

You can record continuous data at so many different measurements – width, temperature, time, and etc. This is where the key difference from discrete types of data lies.

The continuous variables can take any value between two numbers. For example, between 50 and 72 inches, there are literally millions of possible heights: 52.04762 inches, 69.948376 inches and etc.

A good great rule for defining if a data is continuous or discrete is that if the point of measurement can be reduced in half and still make sense, the data is continuous.

Examples of continuous data:

- The amount of time required to complete a project.
- The height of children.
- The square footage of a two-bedroom house.
- The speed of cars.