INSAID
INTERNATIONAL SCHOOL OF AI AND DATA SCIENCE

DATA ENGINEER INTERVIEW QUESTIONS

1. Explain Data Engineering.

Data engineering is a term used in big data. It focuses on the application of data collection and research. The data generated from various sources are just raw data. Data engineering helps to convert this raw data into useful information.

2. What is Data Modelling?

Data modeling is the method of documenting complex software design as a diagram so that anyone can easily understand. It is a conceptual representation of data objects that are associated between various data objects and the rules.

3. Distinguish between structured and unstructured data

Following is a difference between structured and unstructured data:

Parameter	Structured Data	Unstructured Data
Storage	DBMS	Unmanaged file structures
Standard	ADO.net, ODBC, and SQL	STMP, XML, CSV, and SMS
Integration Tool	ELT (Extract, Transform, Load)	Manual data entry or batch processing that includes codes

scaling	Schema scaling is difficult	Scaling is very easy.

4. List important fields or languages used by data engineer
Here are a few fields or languages used by data engineer:

- Probability as well as linear algebra
- Machine learning
- Trend analysis and regression
- Hive QL and SQL databases

5. What is Big Data?
It is a large amount of structured and unstructured data that cannot be easily processed by traditional data storage methods. Data engineers are using Hadoop to manage big data.

6. How to see the database structure in MySQL?
In order to see database structure in MySQL, you can use

DESCRIBE command. Syntax of this command is DESCRIBE Table name;.

7. Compare Relational and Non-Relational Databases
A relational database is one where data is stored in the form of a table. Each table has a schema, which is the columns and types a record is required to have. Each schema must have at least one primary key that uniquely identifies that record. In other words, there are no duplicate rows in your database. Moreover, each table can be related to other tables using foreign keys.

One important aspect of relational databases is that a change in a schema must be applied to all records. This can sometimes cause breakages and big headaches during migrations. Non-relational databases tackle things in a different way. They

are inherently schema-less, which means that records can be saved with different schemas and with a different, nested structure. Records can still have primary keys, but a change in the schema is done on an entry-by-entry basis.

You would need to perform a speed comparison test based on the type of function being performed. You can choose INSERT, UPDATE, DELETE, or another function. Schema design, indices, the number of aggregations, and the number of records will also affect this analysis, so you'll need to test thoroughly. You'll learn more about how to do this later on.

Databases also differ in scalability. A non-relational database may be less of a headache to distribute. That's because a collection of related records can be easily stored on a particular node. On the other hand, relational databases require more thought and usually make use of a master-slave system.

8. How to speed Up SQL Queries?

Speed depends on various factors, but is mostly affected by how many of each of the following are present:

- Joins
- Aggregations
- Traversals
- Records

The greater the number of joins, the higher the complexity and the larger the number of traversals in tables. Multiple joins are quite expensive to perform on several thousands of records involving several tables because the database also needs to cache the intermediate result! At this point, you might start to think about how to increase your memory size.

Speed is also affected by whether or not there are indices present in the database. Indices are extremely important and allow you to quickly search through a table and find a match for some column specified in the query.

Indices sort the records at the cost of higher insert time, as well as some storage. Multiple columns can be combined to create a single index. For example, the columns date and price might be combined because your query depends on both conditions.

9. List various types of design schemas in Data Modelling

There are mainly two types of schemas in data modeling: 1) Star schema and 2) Snowflake schema.

Star Schema or Star Join Schema is the simplest type of Data Warehouse schema. It is known as a star schema because its structure is like a star. In the Star schema, the center of the star may have one fact table and multiple associated dimension tables. This schema is used for querying large data sets.

A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions. It is so-called as snowflake because its diagram looks like a Snowflake. The dimension tables are normalized, that splits data into additional tables.

10. Explain the features of Hadoop

Important features of Hadoop are:

- It is an open-source framework that is available freeware.
- Hadoop is compatible with the many types of hardware and easy to access new hardware within a specific node.
- Hadoop supports faster-distributed processing of data.
- It stores the data in the cluster, which is independent of the rest of the operations.
- Hadoop allows creating 3 replicas for each block with different nodes.

11. How to deploy a big data solution?

Follow the following steps in order to deploy a big data solution.

1) Integrate data using data sources like RDBMS, SAP, MySQL, Salesforce

2) Store data extracted data in either NoSQL database or HDFS.

3) Deploy big data solutions using processing frameworks like Pig, Spark, and MapReduce.

12. What does SerDe mean in Hive? List components available in Hive data model.

SerDe is a short name for Serializer or Deserializer. In Hive, SerDe allows you to read data from table to and write to a specific field in any format you want.

There are the following components in the Hive data model:

- Tables
- Partitions
- Buckets

13. Explain the use of Hive in the Hadoop ecosystem.

Hive provides an interface to manage data stored in the Hadoop ecosystem. Hive is used for mapping and working with HBase tables. Hive queries are converted into MapReduce jobs in order to hide the complexity associated with creating and running MapReduce jobs.

14. What is Metastore in Hive?

It stores schema as well as the Hive table location.

Hive table defines, mappings, and metadata that are stored in Metastore. This can be stored in RDBMS supported by JPOX.

15. Explain how data analytics and big data can increase company revenue?

Following are the ways how data analytics and big data can increase company revenue:

- Use data efficiently to make sure that business growth.
- Increase customer value.
- Turning analytical to improve staffing levels forecasts.
- Cutting down the production cost of the organizations