Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 3+ years
- **Location:** Gurgaon
- **Work mode:** Hybrid
- **Compensation:** ₹22–27LPA
- **Total Rounds:** 5
- **Top Required Skills:**

1. SQL
2. Spark
3. ETL Development & Data Lake Design
4. Cloud Systems
5. Performance Tuning

Ankita Gulati

Shubh Goyal

Round 1
Programming&DWH Basics

Python

1. Write a Python program that generates the squares of numbers from 1 to n using a generator function with the yield keyword.
2. Write a Python program to swap two numbers without using a third variable.

Data Warehouse & Cloud

1. How is Snowflake different from other traditional data warehouses or modern data lakes? Discuss its architecture, compute-storage separation, micro-partitioning, cloning, time-travel features, and cost structure.

Ankita Gulati

Shubh Goyal

System Design

1. Design a relational database schema for the Airtel Wynk music app. Include entities for users, artists, songs, playlists, subscriptions, and play events. Clearly describe the relationships (one-to-many, many-to-many) between these entities.

Ankita Gulati

Shubh Goyal

Round 2
SQL Scenarios

Sessionization Problem (Hive + Spark)

1. You are given a clickstream dataset stored in Hive with the schema:

click_time,user_id
2018-01-01 11:00:00,u1
2018-01-01 12:10:00,u1
2018-01-01 13:00:00,u1
2018-01-01 13:50:00,u1
2018-01-01 14:40:00,u1
2018-01-01 15:30:00,u1
2018-01-01 16:20:00,u1
2018-01-01 16:50:00,u1
2018-01-01 11:00:00,u2
2018-01-02 11:00:00,u2

A session is defined as follows:

- It expires if there is inactivity greater than 1 hour.
- It remains active for a maximum of 2 hours from the start.

Using Spark, enrich the dataset by assigning a session ID for each row based on these rules.

SQL Procedure

1. You are given a table customer_transactions with columns:
id (unique transaction ID)
customer_name (name of the customer)
transaction_time (timestamp of the transaction)
transaction_amount (transaction value)
Write a MySQL stored procedure that returns the customer names for whom every pair of consecutive transactions occurred exactly 10 seconds apart.

Input:

id	customer_name	transaction_time	transaction_amount
1	Lillian Nelson	2017-01-01 10:10:15	10
2	Susan Moore	2017-01-01 11:11:11	20
3	Kian Lawrence	2017-01-01 12:12:12	10 4
Lillian Nelson	2017-01-01 10:10:20	30 5
Lillian Nelson	2017-01-01 10:10:30	40
6	Susan Moore	2017-01-01 11:11:21	50

Expected Output:

customer_name

Susan Moore

Ankita Gulati

Shubh Goyal

Unpivot Problem (Scala / Spark)

1. Given the following table:

```
EmpId | Name | Location1 | Location2 | Location3
------+--------+-----------+-----------+----------
1     | Gaurav | Pune      | Bangalore | Hyderabad
2     | Risabh | Mumbai    | Bangalore | Pune
```

Transform it into the following format using Scala (Spark):

```
EmpId | Name | Location
------+--------+----------
1     | Gaurav | Pune
1     | Gaurav | Bangalore
1     | Gaurav | Hyderabad
2     | Risabh | Mumbai
2     | Risabh | Pune
```

Explain how you will handle null or missing values.

Ankita Gulati Shubh Goyal

Round 3
System Design & Coding

Spark Coding & Data Lake Operations

1. Show how to read data from a JDBC source in Spark, specifying options for URL, table, user credentials, and partitioning for parallel reads.
2. Write Spark code to read a JSON file and convert it into a DataFrame. Compare using schema inference with defining an explicit schema. Show how to handle multi-line JSON.
3. Given a nested JSON dataset, flatten it into a flat schema using Spark SQL functions such as explode and column dot-notation.
4. Explain how to implement incremental data loads from a data warehouse into a data lake using Spark. Discuss strategies such as tracking the maximum updated_at field, handling change-data-capture, and using MERGE with Delta Lake or Apache Hudi.
5. Write Spark code to write a large dataset into Amazon S3 partitioned by day. Explain how you would control the number of output files, avoid generating too many small files, and use appropriate output committers.

Ankita Gulati

Shubh Goyal

Round 4

System Design-WynkPopularity Models

1. Design the clickstream events for the Wynk music app. Examples include song_play, artist_follow, playlist_add, playlist_play, hellotune_set, search, and browse. Define the required fields for each event, such as user ID, device, app version, timestamps, content IDs, play duration, location, network type, and session ID
2. Propose a data model for calculating artist popularity. The score should be based on the number of song plays, the number of times an artist is followed, and the number of times a hellotune is set for their songs.
3. Propose a data model for calculating playlist popularity. The score should be based on the popularity of the songs and artists in the playlist. Suggest a formula that incorporates weighted averages and recency decay.
4. Describe how the clickstream data will flow through data lake layers (raw → refined → semantic) and how aggregates (daily, hourly) will be computed. Discuss whether the pipelines should be batch, streaming, or hybrid.

Ankita Gulati

Shubh Goyal

Round 5
PerformanceTuningonEMR Spark

1.Scenario:

- Input Data Size: 2 TB
- Cluster Size: 14 TB
- Input File Partitions: 8000
- Cluster: EMR with Spark 3.x and Spark 2.4.7
- Total Memory: 14,336 GB
- Output: Write into S3, partitioned by day
- Target Shuffle Partitions: 1500

Questions:

- Propose Spark configuration parameters for this workload, including executor memory, driver memory, memory overhead, number of cores per executor, minimum and maximum executors, and shuffle partitions. Justify each parameter choice.
- How would you handle data skew in this workload? Discuss strategies such as salting, repartitioning, and skew join optimization.
- What improvements in Spark 3.x (such as Adaptive Query Execution) can help optimize this workload compared to Spark 2.4.7?

Ankita Gulati

Shubh Goyal

Thank You

Best of luck with your upcoming interviews — you've got this!

Ankita Gulati

Shubh Goyal