

Data Engineering Interview Questions



Ankita Gulati

Shubh Goyal



Job Details

- **Position:** Senior Data Engineer
- **Experience:** 5+ years
- **Location:** Hyderabad
- **Work mode:** Hybrid
- **Compensation:** ₹20–23LPA
- **Total Rounds:** 2
- **Top Required Skills:**
 1. SQL
 2. PySpark
 3. ETL Development
 4. AWS

Round 1

Python & SQL

Python Coding & Logic

1. Write Python code to split -

full name = "Saumya Shukla Massmutual"
into first, middle, and last names.

2. Write a Python program where -

input list = [a, b, c, d]
output_list = [a, bb, ccc, ddd].

3. Extract only string values from the given list:

list1 = [10, 20, "Jessa", 12, "Emma"].

4. If you want to identify an integer in a string using regex, which pattern would you use?

SQL & Joins

5. Write a query to find duplicate records, first identifying them and then deleting them.

6. Given two tables:

- Table A = [1, 1, 1, 1]

- Table B = [1, 1, 1, 1]

Explain the results of INNER JOIN, LEFT JOIN, RIGHT JOIN, and FULL JOIN between these tables.

Data Modeling & Keys

7. What are fact and dimension tables?

8. Explain the difference between a primary key and a unique key.

9. What is a foreign key?

10. Suppose your fact table has arrived but the dimension table has not yet arrived. What would you do?

11. Which table do you load first – fact table or dimension table?

Round 2

API, Cloud & Data Engineer

Rest API & Authorization

1. While making a POST API call, what would the grant type value be?
2. If you made a POST request to obtain a token, what will the payload contain?
3. Which HTTP methods are you using in your project?
4. Explain the different HTTP methods.
5. You are receiving status code 401. What does this mean?
6. What type of authorization have you used to get the token?
7. Regarding the REST API in your project, can you walk me through what you worked on? Specifically, did you handle tasks like extracting data from the source, applying transformations, and loading it into the final landing zone?

Ankita Gulati

Shubh Goyal

Data Pipelines & Schema Handling

8. In AWS Glue Crawler, how can you customize it to handle specific data formats and structures? What are the implications of customizing the classification and schema evolution process?
9. Since AWS Glue Crawler uses schema inference to extract metadata, how does it handle a case where a single column contains multiple or inconsistent data types?
10. Suppose you have multiple data source models that need to be integrated. How would you ensure data consistency across all sources during integration?
11. Suppose your data sources are stored in S3 with multiple file formats, and your ETL components include AWS Glue, Airflow, or RPSS, with Snowflake as the data warehouse. How would you design an end-to-end data pipeline for this setup?

12. The customer wants the pipeline to be designed in such a way that any addition or removal of columns (schema changes) should be automatically handled without updating the code. How would you design the pipeline to support dynamic schema handling?

13. You have multiple pipelines. If there are frequent changes in some component of the pipeline, what approach would you follow to handle such changes effectively?

EMR & Cluster Sizing

14. Let's say a customer comes to you and says: "I have around 1 terabyte of data to process. How much memory should I allocate to the EMR cluster for efficient processing?"

How would you estimate or decide the right memory configuration for this workload?

Thank You

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

