Zen Technologies
Company Profile - Overview

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Senior Data Engineer
- **Experience:** 7 years
- **Location:** Mumbai
- **Work mode:** Hybrid
- **Compensation:** ₹35+ LPA
- **Total Rounds:** 6
- **Top Required Skills:**

1. SQL
2. PySpark / Python
3. Cloud Data Engineering
4. ETL / Data Modeling
5. Big Data & Streaming
6. System Design

Ankita Gulati

Shubh Goyal

Round 1
Technical

SQL Questions:

1. Write a query to fetch the 3rd highest salary per department, including ties. (Use DENSE_RANK in a CTE).
2. Given a table (id, date, sales_amount), calculate the 7-day rolling average sales. (Use ROWS BETWEEN 6 PRECEDING AND CURRENT ROW).
3. Write a recursive SQL query to print the employee → manager hierarchy chain.
4. How do you ensure NULL values do not affect aggregations like AVG()?

PySpark Coding:

1. Read a dataset from Azure Data Lake Storage, remove duplicates and nulls.
2. Given a transaction dataset, calculate total spend per user and rank them. .
3. Implement a PySpark job to join two large DataFrames efficiently. (Use broadcast() for smaller DataFrame).

Ankita Gulati

Shubh Goyal

Round 2
Spark, SQL & System Concepts

Spark / PySpark:

1. Explain the difference between RDD, DataFrame, and Dataset. When would you use each?
2. Explain narrow vs wide transformations with examples (e.g., map vs groupBy).
3. How do you handle data skew in Spark? (Techniques: key salting, repartitioning).
4. What happens in Spark DAG when you call an action?
5. Write a PySpark script to read streaming data from a folder and aggregate in mini-batches.

SparkSQL:

1. Given orders and customers, write a query to fetch customers who ordered above the average order amount.
2. Difference between LIMIT, ROW_NUMBER(), and Top N queries.
3. Explain partitioning and bucketing in SparkSQL.

Ankita Gulati Shubh Goyal

Architecture:

1. How would you design a data warehouse schema for financial transactions? (Fact vs Dimension, Star Schema)
2. How would you optimize a query running slow due to large dataset joins?

Ankita Gulati

Shubh Goyal

Round 3
Techno-Managerial

Azure Stack:

1. What are ADF pipelines, and how do you pass parameters across activities?
2. Difference between global parameters vs pipeline parameters in ADF.
3. How does Azure Synapse differ from Azure SQL DB?
4. How would you orchestrate a batch ETL pipeline with error handling and retries?
5. Explain how the Databricks-Snowflake connector works internally.

Scenario Design:

→ Design an ingestion pipeline to bring daily incremental customer data from an API into Azure Data Lake, process it with PySpark, and load into Synapse for reporting.

- Use ADF, Data Lake Gen2, Databricks.
- Cover incremental load strategies, partitioning, monitoring.

Ankita Gulati Shubh Goyal

Managerial Questions:

1. How do you estimate resources for a pipeline that ingests 1 TB/day?
2. If business demands near real-time, but infra costs are too high, how do you handle stakeholder conflict?

Ankita Gulati

Shubh Goyal

Round 4
Client Interview

SQL Questions:

1. Write a query to find customers who made transactions in all 12 months of a year.
2. Given (transaction_id, user_id, amount, transaction_date), detect consecutive duplicate transactions by the same user.
3. What are temporal tables in SQL Server? When would you use them?

ETL & Data Modeling:

1. Explain Batch vs Incremental Load. How do you implement an Upsert? (MERGE or staging tables).
2. Explain Slowly Changing Dimensions (SCD Types 1, 2, 3) with examples.
3. Compare Star Schema vs Snowflake Schema – which is better for OLAP?

Ankita Gulati

Shubh Goyal

Data Governance:

1. What is data masking? Provide an example in banking/healthcare.
2. How do you ensure auditability and lineage in ADF pipelines?

Round 5
Leadership / CTO Round

Project Deep Dive:

1. Walk me through a project where you optimized Spark jobs.
- Data volume handled.
- Optimizations applied (broadcast joins, caching, repartitioning).
- Measured impact (execution time, infra cost).

Azure & Cloud Design:

1. How would you design a high-availability data pipeline in Azure? (Retries, failover, SLA monitoring).
2. How do you ensure cost optimization in clusters? (Auto-scaling, spot instances, right-sizing executors).

Team Fitment:

1. How do you mentor junior engineers on SQL/Spark best practices?
2. Describe a time you handled a production issue at night. How did you manage it?

Thank You

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

Ankita Gulati

Shubh Goyal