Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer
- **Experience:** 3+ years
- **Location:** Bangalore
- **Work mode:** Office
- **Compensation:** ₹30+ LPA
- **Total Rounds:** 7
- **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
Phone Screen

1. Walk me through your background and past project experience.

2. Explain one project from your resume in layman's terms.

3. Describe your experience working with qualitative and quantitative methods.

Ankita Gulati

Shubh Goyal

Round 2
Algorithms & Problem-Solving

1. Find the minimum number of coins required to make up a given amount, given a set of coin denominations.

2. Implement a scalable Top-K words algorithm for Amazon product descriptions using a Count-Min Sketch.

3. What is the computational complexity of finding the most frequent word in a document?

4. Given a list of strings, describe how you would perform feature selection for document classification.

5. Given a dataset, how would you explain and calculate feature importance?

Ankita Gulati

Shubh Goyal

Round 3
Machine Learning & Statistics

1. Explain decision trees, random forests, and their differences.
2. Explain multivariate regression and logistic regression. What are the key parameters in logistic regression?
3. Explain PCA (Principal Component Analysis) and its applications.
4. Explain the assumptions and violations of K-means clustering.
5. Explain SVM (Support Vector Machines):
   • To a technical person (parameters, kernel functions).
   • To a non-technical person (intuition without math).
6. How would you design a classifier to predict the outcome of NFL games in real-time?
7. Describe different probability distributions and how you would model churn.
8. What is the difference between trees and random forests?

Ankita Gulati Shubh Goyal

Round 4
Data Modeling

1. Imagine you are hired by a hospital to analyze patient reports. Each report has:
• Symptoms + medical history (natural language text).
• Disease IDs (labels).
→ Propose a model to map symptoms to diseases.

2. You have 10 TBs of unstructured customer data. How would you extract valuable information from it? Walk through your approach in technical detail.

3. Based on a dataset we provide, what questions would you ask to prove that the programs generating this data add value to the company?

Ankita Gulati

Shubh Goyal

Round 5
Big Data & Scalability

1. How would you design a system to scale for billions of records?

2. How do you decide which big data tools (e.g., Spark, Hadoop, Kafka, Snowflake) should be used in different situations?

3. Explain how you would handle data ingestion, processing, and storage for large-scale distributed systems.

Ankita Gulati

Shubh Goyal

Round 6
Research Talk (Presentation)

Before the onsite interviews, candidates present a research talk to the group, typically covering:

1. A deep-dive into past projects/research work.

2. How your methods contribute to business value.

3. Key takeaways, challenges, and scalability considerations.

Ankita Gulati

Shubh Goyal

Round 7
HR

1. Why Salesforce?

2. Describe a time when you faced a major data engineering or ML pipeline challenge. How did you overcome it?

3. How do you keep yourself updated with the latest trends in data engineering/AI/ML?

4. Where do you see yourself contributing in the next 3–5 years at Salesforce?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal