zeta

Data Engineering Interview Questions

Ankita Gulati

Shubh Goyal

Job Details

- **Position:** Data Engineer II
- **Experience:** 5+ years
- **Location:** Bangalore / Mumbai
- **Work mode:** Office
- **Compensation:** ₹30-40LPA
- **Total Rounds:** 3
- **Top Required Skills:**

1. SQL
2. PySpark
3. ETL Development
4. AWS
5. Data Modeling
6. Data Warehousing
7. System Design / Architecture

Ankita Gulati

Shubh Goyal

Round 1
SQL & Database Concepts

SQL

1. What are the different types of joins in SQL? Briefly explain each with an example.
2. What is the difference between UNION, UNION ALL, and UNION DISTINCT? Which one would you prefer for very large datasets, and why?
3. Suppose you have an employee table and a manager table. You want a combined list of unique IDs from both. Since managers are also employees, their IDs may already exist in the employee table, but some may not.
   - How would you write a query using UNION ALL to ensure no duplicates?
4. From a performance perspective, which is more efficient: UNION ALL with DISTINCT, UNION ALL with window functions or plain UNION ?

Ankita Gulati Shubh Goyal

5. Explain the difference between normalization and denormalization. When would you prefer which ?
6. Give an example of a table in First Normal Form (1NF) and then transform it into Second Normal Form (2NF) with explanation.

SQL - Scenarios

7. Given an Employee table with columns id (primary key), name, department, and manager_id (which references id and can be NULL if the employee has no manager, with the condition that no employee manages themselves), write a SQL query to display each employee along with the name of their manager.
8. You have an orders table with: order_id, customer_id, order_date, amount, reason.
Write a query to calculate the 7-day rolling average of amount for each reason.

Round 2
Data Management Concepts

Data Quality & CDC

1. How does a data validation framework like Great Expectations fit into ensuring data quality in an OLTP → OLAP pipeline?
2. Schema changes in OLTP often require application code updates. How can the impact of schema changes be minimized?
3. In an OLTP system, how can triggers be used to capture every insert, update, or delete into a history table?
4. What would be the schema for a Card table and its corresponding CardHistory table? How does each column support change tracking?
5. Why do we include valid_from and valid_to columns in history tables when the latest state can be derived from entry order? What additional value do they provide?

Ankita Gulati

Shubh Goyal

6. In a history-tracking design, how would a trigger: Insert a new version of a record, and Update the valid_to of the previous version?

7. How would you extend the same history-tracking mechanism to related tables like Account and Limits, which are linked to Card?

8. If the client expects a consolidated CardExtract table with card, account, and limit details plus the exact change date: How would you design this table?

9. Given CardHistory, AccountHistory, and LimitHistory, how would you build the CardExtract target model in OLAP so that it captures changes from all three entities?

10. Can you demonstrate (SQL or pseudocode) how to join history tables to generate a daily extract of changes across cards, accounts, and limits?

11. If only the LimitHistory changes (e.g., spending limit adjustment), how do you ensure the CardExtract still reflects this change?

Ankita Gulati

Shubh Goyal

12. On a given day, how would you detect which records across history tables actually changed, so only those are extracted?

13. If millions of cards exist but only a few thousand change daily, how would you optimize the extract to avoid recomputing all history?

14. What is the best way to identify the specific set of impacted cards that should be processed for the extract?

Ankita Gulati

Shubh Goyal

Round 3
Architecture &System Design

Data Architecture & TechnologyDesign

1. You mentioned Airflow, Spark, and Glue for transformations. How do you decide which one to use? Provide scenarios where Spark, Glue, or Airflow would be the best fit.
2. A bank wants Zeta's platform to run on Azure (no AWS). How would you design the architecture to be cloud-agnostic? How would you decouple transformation, storage, and orchestration from cloud infra? Would you use open-source abstractions (Spark, dbt, Airflow) and storage adapters (S3, ADLS, GCS)? How do you ensure business logic stays the same while only infra plugins change?

Data Integration & Schema Management

3. A bank already has a working Zeta pipeline, warehouse, and reporting. Now they want to migrate product data in different formats and have acquired a smaller bank with incomplete datasets (e.g., missing fraud/risk data). How would you design ingestion and transformation to ensure correctness, quality, and business continuity for BI/fraud monitoring, while handling multi-source integration and missing data?
4. In schema mapping, how do you map source fields to Zeta's canonical model? Provide an example (source field + datatype → Zeta field + transformation). What kinds of mismatches/transformations do you anticipate?
5. If Zeta's schema defines a field as mandatory but the source provides it as optional, how would you handle it (defaults, validation errors, NULLs)?
6. If the source provides extra fields beyond Zeta's schema (e.g., 120 fields vs 100 in Zeta), how would you handle them (ignore, store separately, extend schema)?

Ankita Gulati

Shubh Goyal

Pipeline Design & Platform Enchancement

7. Walk through your pipeline layers — ingestion, raw lake, schema mapping, validation, staging, transformation, curated, BI/reporting. Which design considerations (correctness, integrity, observability, scalability, compliance, maintainability) are you addressing at each stage?
8. Zeta wants a self-serve onboarding model (BYOD – bring your own data). What should be self-service (connectors, schema upload, field mapping, DQ rules, scheduling)? How do you differentiate between building a platform vs a service in this context?

Ankita Gulati

Shubh Goyal

Thank You

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

Ankita Gulati

Shubh Goyal