Data Engineer Interview Preparation ? 73 Strings

SQL ? Theory & Hands-On

Q1. Find the second highest salary from an `employees` table.

SELECT MAX(salary) AS second_highest_salary

FROM employees

WHERE salary < (SELECT MAX(salary) FROM employees);

Q2. Optimize this query:

Original: SELECT * FROM orders WHERE customer_id = 42 AND order_date > '2024-01-01';

Optimized: CREATE INDEX idx_customer_order ON orders(customer_id, order_date);

SELECT order_id, total_amount FROM orders WHERE customer_id = 42 AND order_date > '2024-01-01';

Q3. Find the month with the highest total revenue.

SELECT DATE_TRUNC('month', sale_date) AS month, SUM(amount) AS total_revenue

FROM sales

GROUP BY month

ORDER BY total_revenue DESC

LIMIT 1;

Q4. JOIN types explained.

- INNER JOIN: Only matching rows.

- LEFT JOIN: All left rows, with matched right.

- FULL OUTER JOIN: All rows from both tables.

ETL & Data Pipelines

Q5. Describe a data pipeline you've built.

Ingested HRMS data, normalized using Python, stored in S3, processed using Spark. Monitored via

Datadog.

Q6. Handling schema changes in ETL?

Schema evolution, validation scripts, metadata versioning, and alerting on mismatches.

Spark, PySpark, and Databricks

Q7. Optimize a PySpark job.

- Use .cache(), reduce shuffling, prefer built-ins, use broadcast joins.

Q8. PySpark hands-on example.

df1.join(df2, 'user_id').filter(df2.amount > 1000).select(df1.name, df2.amount).show()

Q9. What are Delta tables?

Delta tables are ACID-compliant tables with time travel, schema enforcement, and fast performance.

Python for Data Engineering

Q10. Python for automation.

Used to ingest API data, transform and validate it using pandas, and store into S3.

Q11. Flatten a nested dictionary.

```
def flatten_dict(d, parent_key='', sep='_'):
  # function to flatten nested dicts
  ...
```

Cloud & Monitoring (AWS, Datadog)

Q12. Cost optimization on AWS.

- Instance right-sizing, scheduled shutdowns, lifecycle rules on S3, consolidated logging.

Q13. Pipeline health monitoring.

- Use Datadog, Grafana for metrics. Alerts via Sentry & OpsGenie.

System Design & Domain-Specific

Q14. Structuring investor reports.

Extract using Python libs (Tabula, pandas), normalize to schema, store in Delta tables.

Q15. Real-time portfolio monitoring.

Use APIs + Kafka + Spark Streaming + Delta Lake + Redis + React dashboards.

Behavioral / Project Experience

Q16. Migration of 100+ clients.

Used blue-green deployment, feature flags, validation scripts, and careful monitoring.

Q17. Why 73 Strings?

Excited about data, AI, and finance convergence. Series B support and deep tech problem-solving is appealing.

What to Expect on Intervue.io

Overview

Live coding in SQL/Python, real-time problem solving, behavioral Qs.

Prepare on LeetCode, review pipelines, and practice design Qs.