■ Data Engineering Interview Preparation Guide ■

Covers concepts, real-world problems, and curated interview questions across 10 companies

■ Step-by-Step Prep Roadmap
■ Real-Time Project Scenarios
■ Scenarios Based Notes
■ Interview experiences
■ Real Interview Experiences & Solutions
■ 100+ Curated Interview Questions

■ Real Interview Experiences from Top Companies

TCS

Theoretical Questions:

– What is the architecture of a data lake?
– Explain the role of ETL in data engineering.
– What is schema evolution in big data systems?
– Difference between batch and stream processing.
– How does partitioning improve query performance?

Coding Questions:

– Write a PySpark code to remove nulls and duplicates from a DataFrame.
– Transform nested JSON into tabular format using Python.
– Calculate running average using window function in Spark.

Real-Time Scenario:

– Build an incremental data pipeline to load daily sales from flat files to a Delta table.

Infosys

Theoretical Questions:

– Explain distributed computing in context of Spark.
– How does Kafka handle message durability and delivery?
– What are benefits of using Parquet over CSV?
– Difference between Spark SQL and Hive.
– Explain fault tolerance in data pipelines.

Coding Questions:

– Write a Kafka consumer that listens to real-time stock updates.
– Load and aggregate data using Spark SQL.
– Read a large file in chunks using Python.

Real-Time Scenario:

– Build a customer churn detection pipeline using historical activity data.

Capgemini

Theoretical Questions:

– How do you monitor and optimize Airflow DAGs?

– What is a Slowly Changing Dimension (SCD)?

– Difference between OLAP and OLTP.

– What are Z-order and compaction in Delta Lake?

– How do watermarking and event-time windows work in streaming?

Coding Questions:

– Write Airflow DAG to load data daily to a warehouse.

– Merge two datasets on customer_id using PySpark.

– Generate MD5 hash for data integrity check in Python.

Real-Time Scenario:

– Design a pipeline to cleanse and aggregate IoT sensor data every hour.

Wipro

Theoretical Questions:

– What is the CAP theorem and its implication on NoSQL?

– Explain how data skew can affect Spark jobs.

– What is CDC (Change Data Capture)?

– Difference between row-level and file-level operations in Delta Lake.

– When to use broadcast joins?

Coding Questions:

– Identify data skew using PySpark DataFrame APIs.

– Capture changed rows using timestamp in SQL.

– Write a lambda function to filter nested dict data.

Real-Time Scenario:

– Implement real-time CDC pipeline from PostgreSQL to S3 via Kafka.

Deloitte

Theoretical Questions:

– Explain orchestration vs workflow in data pipelines.

– How do you handle schema mismatch in streaming data?

– What is Lakehouse architecture?

– Difference between append and overwrite in Spark writes.

– When to use Snowflake vs BigQuery?

Coding Questions:

– Query Snowflake using Python connector.

– Join multiple tables using Spark SQL.

– Create partitioned external table in Hive.

Real-Time Scenario:

– Create end-to-end pipeline to track invoice payments with SLA alerts.

Amazon

Theoretical Questions:

– How does Kinesis differ from Kafka?

– What is eventual consistency and where is it used?

– Describe data encryption at rest and in transit.

– What is the difference between EMR and Glue?

– How do you handle retries in data pipeline failures?

Coding Questions:

– Write Python script to encrypt a file using AES.

– Ingest streaming data to Redshift using Glue.

– Create checkpointing in Spark Structured Streaming.

Real-Time Scenario:

– Build product recommendation engine using S3 + EMR + SageMaker.

Google

Theoretical Questions:

– What is BigQuery and how is it priced?

– Explain the difference between DataProc and Dataflow.

– How is sharding implemented in Bigtable?

– What are some common anti-patterns in big data queries?

– How do materialized views help with performance?

Coding Questions:

– Create a scheduled query in BigQuery.

– Write Python code to interact with Google Cloud Storage.

– Export BigQuery results as CSV using Python.

Real-Time Scenario:

– Build unified analytics pipeline using Pub/Sub, Dataflow, and BigQuery.

Microsoft

Theoretical Questions:

– Difference between Azure Synapse and Azure Data Factory.

– Explain PolyBase and external tables in Synapse.

– What are sink transformations in ADF?

– What is Delta Lake and how is it implemented on Azure?

– How to handle large-scale data ingestion securely?

Coding Questions:

– Create Azure Data Factory pipeline to load JSON to SQL DB.

– Transform data using Dataflow in ADF.

– Query Delta table in Synapse using Spark SQL.

Real-Time Scenario:

– Pipeline to sync ERP system data with Azure Data Lake for analytics.

EY

Theoretical Questions:

– What is the importance of metadata management?

– Explain lineage tracking in data engineering.

– Difference between mutable and immutable data stores.

– What is data modeling and its types?

– How to ensure GDPR compliance in data pipelines?

Coding Questions:

– Extract schema from Parquet file using PyArrow.

– Log job progress and metrics using Python logger.

– Write Python function to validate data against schema.

Real-Time Scenario:

– Build secure data pipeline with audit logs and field-level encryption.

JP Morgan

Theoretical Questions:

– What is a DAG in workflow orchestration?

– Explain real-time fraud detection system.

– How to handle backpressure in streaming systems?

– Difference between micro-batching and event processing.

– What is lambda architecture?

Coding Questions:

– Create DAG with conditional branching in Airflow.

– Set up alert on SLA miss in Airflow using email operator.

– Detect and deduplicate records in Kafka stream.

Real-Time Scenario:

– Build fraud detection engine using Kafka + Spark + Cassandra.