Fragma Data Systems Data Engineer Interview Guide – Experienced 3+

Round 1: HR Screening

This round focuses on assessing communication, motivation, and background.

Sample Questions:

1. Walk me through your resume.
2. Why are you interested in this role and Fragma Data Systems?
3. Describe a challenging project you worked on.
4. What technologies are you most comfortable with?
5. Are you open to learning new tools and technologies?
6. What is your notice period, and are you interviewing elsewhere?
7. What are your salary expectations?
8. How do you keep yourself updated with new data engineering trends?
9. What motivates you to work in data engineering?

Round 2: Technical Assessment (Take-Home Coding)

This round involves solving real-world data engineering tasks using Python, SQL, and Spark.

Example Tasks and Questions:

1. Implement a PySpark job to read CSV data, perform joins, and store output as partitioned Parquet.
2. Write a SQL query to find the top 5 products by sales per region.
3. Optimize a query fetching customer data with a rolling 6-month sales sum.
4. Develop a Python script to clean data by removing duplicates and handling missing values.
5. Explain how you would design a partition strategy for a large dataset in HDFS.
6. Write a function to detect schema evolution issues in Parquet files.
7. Implement a Kafka consumer that writes streaming data into a database.
8. Write a PySpark code snippet to filter rows with a specific condition.
9. Create a SQL query to identify customers with purchases above a dynamic threshold.
10. Explain steps to optimize data read performance from cloud storage (S3 or Azure Blob).

Round 3: Technical Interview (Spark, SQL, and Performance Tuning)

This round tests knowledge of core Spark concepts, PySpark, and SQL optimization.

Sample Questions:

1. How does Spark execute a job (explain the DAG and stages)?
2. Describe the difference between Spark RDDs, DataFrames, and Datasets.
3. Explain the benefits of using DataFrames over RDDs.
4. What are narrow and wide transformations in Spark? Give examples.
5. How do you reduce shuffle operations in Spark?
6. Explain the use of broadcast variables in PySpark.
7. How does lazy evaluation work in Spark?
8. Write a SQL query to find employees earning the second-highest salary.
9. How do you handle out-of-memory errors in Spark jobs?
10. Describe your approach to handling skewed data.

Round 4: Advanced Technical Interview (Kafka, Data Pipelines, and Partitioning)

This round covers real-time data systems, Kafka, and pipeline design.

Sample Questions:

1. Explain the architecture of Kafka and its core components.
2. How does Kafka ensure message durability and reliability?
3. What is the role of Zookeeper in Kafka?
4. How do you monitor consumer lag in Kafka, and how can you reduce it?
5. Describe your approach to managing offsets in Kafka.
6. Explain the difference between repartition() and coalesce() in Spark.
7. How do you optimize partitioning when dealing with large datasets?
8. What are the different delivery semantics in Kafka (at least-once, at-most-once, exactly-once)?
9. How would you design a data pipeline to handle late-arriving data?
10. Discuss approaches for fault-tolerant data ingestion in real-time systems.

Round 5: System Design and Techno-Managerial Round

This round evaluates designing data systems and decision-making skills.

Scenario Question:

Design a scalable system for processing real-time sales data from multiple stores, storing it for analytics, and generating reports.

Key Questions and Discussion Points:

1. What data storage would you use for real-time analytics? Why?
2. Explain your choice of streaming framework (Kafka, Spark Streaming, etc.).
3. How would you design the schema for transactional data storage?
4. Describe your approach to managing data deduplication.
5. How would you handle fault tolerance in your pipeline?
6. What optimizations would you apply for partitioning strategies?
7. How would you handle schema evolution in a real-time data system?
8. Describe your monitoring strategy for this pipeline.
9. Discuss scalability challenges and solutions for growing data volumes.
10. How would you incorporate data security and access control?

Conclusion

This interview process reflects a balanced mix of coding, theoretical knowledge, and system design. Preparing thoroughly across these topics enhances success chances.

Glassdoor Fragma Data Systems Review –

https://www.glassdoor.co.in/Reviews/Fragma-Data-Systems-Reviews-E2376618.htm

Fragma Data Systems Careers –

https://fragmadata.com/

Subscribe to my YouTube Channel for Free Data Engineering Content –

https://www.youtube.com/@shubhamwadekar27

Connect with me here –

https://bento.me/shubhamwadekar

Checkout more Interview Preparation Material on –

https://topmate.io/shubham_wadekar