Microsoft Data Engineer Interview Guide

Round 1: Big Data, PySpark & Data Warehousing

Focus Areas:

- PySpark internals and performance tuning
- Hadoop ecosystem knowledge
- Dimensional modeling in data warehousing

Topics Covered:

PySpark & Performance Optimization:

- DAGs: Jobs → Stages → Tasks — understand the Spark UI layout.
- Narrow vs. Wide transformations (e.g., map, filter vs. join, groupBy).
- Broadcast join vs. Shuffle join — when to use each and the memory implications.
- Reading .explain() output and using the Spark UI for debugging.

Sample Questions:

- Explain how Spark groups transformations into stages. What causes a stage boundary?
- When would you choose a broadcast join over a shuffle join? Any memory risks?
- What's the difference between narrow and wide transformations?
- How would you diagnose a slow PySpark job?

Hadoop Ecosystem:

- HDFS architecture: Block size, replication factor, and role of NameNode.
- YARN: Resource management and containerization.
- MapReduce vs. Spark — tradeoffs and use cases.

Sample Questions:

- How does HDFS handle fault tolerance?
- Compare Spark and MapReduce for iterative workloads.
- What happens if the NameNode goes down?
- How is resource allocation handled in YARN?

Data Warehousing & Dimensional Modeling:

- Fact vs. Dimension tables.
- Star vs. Snowflake schemas.
- Partitioning and indexing for performance.
- Slowly Changing Dimensions (SCD Type 2) strategy.

Sample Questions:

- How would you model SCD Type 2 in a data warehouse?
- When would you choose a Snowflake schema over a Star schema?
- What's the role of surrogate keys in dimensional modeling?
- How do partitions improve query performance in fact tables?

Round 2: Data Structures & Algorithms (DSA)

Focus Areas:

- Problem-solving, edge case handling, and code efficiency.
- Clean and optimal Python coding.

Problem 1: Dutch National Flag Problem

- Three-pointer technique for in-place partitioning.
- Used when dealing with fixed categories (like 0s, 1s, and 2s).

Problem 2: Interval Merging

- Merge overlapping intervals with sorting or timeline methods.
- Optimization: Use bucket sort or event counters for large-scale merging.

Sample Questions:

- Solve the Dutch National Flag problem in one pass. How would you handle it?
- Given a list of intervals, merge the overlaps. How do you optimize it?
- What's the time and space complexity of both solutions?
- How would you test these functions with edge cases?

Round 3: System Design & Dimensional Modeling

Focus Areas:

- Real-world pipeline architecture.
- Tool choices, scalability, and fault tolerance.

Pipeline Design:

- Batch vs. Streaming ingestion.
- Kafka, Spark Streaming, Redshift, Snowflake, and S3 as building blocks.
- Parquet vs. Avro: format tradeoffs.
- Query optimizations: filter pushdown, partition pruning, caching.

Dimensional Modeling Design:

- Surrogate keys, SCD handling.
- Partition strategies for fact tables.
- Designing scalable star schema models.

Sample Questions:

- Design a data pipeline to ingest and process clickstream data in near real-time.
- What storage format would you choose for analytics-heavy workloads and why?
- How would you manage schema evolution in your data lake?
- Design a data model to track orders, payments, and shipping — handle changes in customer address.

Round 4: Managerial + Optimization Round

Focus Areas:

- System-level thinking and performance tuning
- CI/CD for data pipelines
- Behavioral insights and leadership fit

Technical Deep Dive:

- Explain trade-offs in your previous architecture decisions.
- Identify and resolve data skew and long-running job issues.

CI/CD for Data Engineering:

- Tools: Airflow, Jenkins, GitHub Actions, Databricks Jobs
- Strategies: deployment pipelines, testing, rollback, data validation.

Behavioral and Leadership Principles:

- Adapting to new tools.
- Managing production outages.
- Working across teams under tight deadlines.

Sample Questions:

- Tell me about a time when a Spark job failed in production. How did you fix it?
- How do you set up CI/CD for a PySpark ETL workflow?
- Describe a time when you had to learn a new tool quickly to complete a task.
- How do you ensure data quality and validation in a fast-moving team?

Glassdoor Microsoft Review –

https://www.glassdoor.co.in/Reviews/Microsoft-Reviews-E1651.htm

Microsoft Careers –

https://careers.microsoft.com/v2/global/en/home.html

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