Microsoft

Data Engineer Interview Questions (0-3 years) 25 LPA

SQL + DataModeling

1. Write a query to identify users who placed more than 2 orders in a week but skipped one specific weekday.
2. How would you handle Slowly Changing Dimensions (Type 2) in Azure Synapse?
3. What partitioning and indexing strategies would you use for a billion-row fact table?
4. Explain the difference between CUBE, ROLLUP, and GROUPING SETS.
5. How do you optimize a query that's performing poorly?
6. When would you choose star schema vs snowflake schema in data modeling?
7. What are window functions? Give an example use case with ROW_NUMBER, RANK, or LAG/LEAD.
8. How would you detect duplicate records in a table and remove them efficiently?
9. Explain normalization vs denormalization with examples.
10. How do you handle late-arriving dimensions in ETL?

Python + DSA

1. **Given a stream of events (user_id, timestamp), detect fraud patterns based on time gaps.**
2. **Implement a string compression function that supports UTF-8 multi-byte characters.**
3. **Design an LRU cache using files instead of memory, tracking read/write costs.**
4. **Explain time complexity of different Python data structures (dict, set, list).**
5. **How do you handle memory management in Python?**
6. **Implement a function to check if a string is a valid palindrome, ignoring punctuation.**
7. **How would you implement a producer-consumer system in Python?**
8. **Solve: Find the k-th largest element in an unsorted list .**
9. **Implement a custom exception class and show how you'd use it in error handling.**
10. **Explain the difference between deep copy vs shallow copy in Python.**

PySpark + System Design

1. **Design a real-time recommendation system with <1s latency for millions of users.**
2. **How would you optimize joins in PySpark (broadcast vs shuffle)?**
3. **Explain when you would use caching vs persistence in PySpark.**
4. **Describe how you'd build a data pipeline with schema evolution (Bronze → Silver → Gold).**
5. **How do you debug and optimize a slow PySpark job?**
6. **What is predicate pushdown and why does it matter?**
7. **Explain the difference between repartition and coalesce.**
8. **How do you monitor PySpark jobs in production?**
9. **When would you use bucketing in Spark SQL?**
10. **Explain fault tolerance in Spark (lineage + DAG).**
11. **How would you design a CDC (Change Data Capture) pipeline in Spark with exactly-once guarantees?**
12. **Compare batch vs streaming in Spark. When would you choose each?**