Crack Any Data Engineering Interview with These 30 Questions  
=====================================================  
  
  
  
  
1. Write an SQL query to find the second highest salary from an employee table.  
  
2. How do you handle NULL values in SQL joins?  
  
3. Write a Python script to read a CSV file and load it into a DataFrame.  
  
4. How do you handle exceptions in Python using try-except blocks?  
  
5. In PySpark, how would you perform a join operation between two large DataFrames efficiently?  
  
6. Write a PySpark code to find the top 3 customers with the highest revenue per region.  
  
7. What is the difference between partitioning and bucketing in PySpark?  
  
8. How do you implement Slowly Changing Dimensions (SCD) in a data warehouse?  
  
9. Explain the concept of star schema and snowflake schema in data modeling.  
  
10. How would you design a fact table for an e-commerce platform?  
  
11. How do you build an ETL pipeline using Azure Data Factory?  
  
12. What are the different types of triggers in ADF and when to use them?  
  
13. Explain the architecture of Azure Databricks and its integration with Delta Lake.  
  
14. Write a PySpark code to process streaming data from Event Hub in Databricks.  
  
15. How do you optimize query performance in Azure Synapse Analytics?  
  
16. How would you design a data warehouse for a retail business using Synapse?  
  
17. What are the best practices for securing data in Azure Data Lake Storage?  
  
18. How do you manage access control and secrets using Azure Key Vault?  
  
19. Write a PySpark script to load data from ADLS into a Delta table.  
  
20. How do you implement data lineage and governance in Microsoft Purview?  
  
21. Build a real-time analytics pipeline using Event Hub, Stream Analytics, and Synapse.  
  
22. How would you handle late-arriving data in a batch ETL pipeline?  
  
23. Write an SQL query to calculate the customer churn rate over the last 6 months.  
  
24. How do you implement incremental data loading in ADF pipelines?  
  
25. Write a Python script to validate data quality and detect anomalies.  
  
26. How do you perform schema evolution in Delta Lake?  
  
27. How would you design a data pipeline to handle both batch and streaming data?  
  
28. Write a PySpark code to perform window functions for ranking sales data.  
  
29. How do you optimize storage and query performance in a Synapse dedicated pool?  
  
30. Build an end-to-end data pipeline that ingests data from multiple sources, transforms it in Databricks, and loads it into Synapse for reporting.