07 Data Ingestion & Integration

Interview Questions

Azure Blob Storage, Azure Data Factory Integration & JDBC Connections

Basic Level Questions (1-10)

1. What is Azure Blob Storage and what are its primary use cases in data engineering? *Expected Answer: Object storage service for unstructured data, used for data lakes, backup, archival, and as a source/sink for data pipelines.*

2. Explain the different storage tiers available in Azure Blob Storage. *Expected Answer: Hot (frequently accessed), Cool (infrequently accessed, stored for 30+ days), Archive (rarely accessed, stored for 180+ days).*

3. What is Azure Data Factory and how does it fit into the data engineering ecosystem? *Expected Answer: Cloud-based ETL/ELT service for data integration, orchestration, and movement between various sources and destinations.*

4. What are the main components of an Azure Data Factory pipeline? *Expected Answer: Activities, datasets, linked services, triggers, integration runtime, and pipelines.*

5. What is JDBC and why is it important for data engineers? *Expected Answer: Java Database Connectivity - API for connecting and executing queries against databases, enables database vendor independence.*

6. How do you authenticate to Azure Blob Storage from an application? *Expected Answer: Connection strings, access keys, SAS tokens, Azure AD authentication, managed identities.*

7. What file formats are commonly stored in Azure Blob Storage for data engineering workloads? *Expected Answer: Parquet, CSV, JSON, Avro, ORC, Delta Lake format.*

8. What is a linked service in Azure Data Factory? *Expected Answer: Connection information to external resources like databases, storage accounts, or services.*

9. Explain the difference between a dataset and a linked service in ADF. *Expected Answer: Linked service defines connection, dataset defines data structure and location within that connection.*

10. What are the common JDBC driver types and when would you use each? *Expected Answer: Type 1 (JDBC-ODBC bridge), Type 2 (Native-API), Type 3 (Network Protocol), Type 4 (Thin driver - most common for modern applications).*

Intermediate Level Questions (11-20)

11. How would you implement incremental data loading from a SQL database to Azure Blob Storage using ADF? *Expected Answer: Use watermark approach with high watermark activity, lookup activity, and conditional logic to process only new/changed records.*

12. Explain the concept of Integration Runtime in Azure Data Factory and its types. *Expected Answer: Compute infrastructure for data movement and transformation. Types: Azure IR (cloud), Self-hosted IR (on-premises/private networks), Azure-SSIS IR (SSIS packages).*

13. How do you handle schema evolution when ingesting data into Azure Blob Storage? *Expected Answer: Use schema-on-read approach, implement versioning, use flexible formats like JSON or Parquet with schema registry.*

14. What are the best practices for organizing data in Azure Blob Storage for analytics workloads? *Expected Answer: Partition by date/time, use proper folder structure, implement naming conventions, consider file sizes (avoid small files).*

15. How would you implement error handling and retry logic in an ADF pipeline? *Expected Answer: Use try-catch activities, configure retry policies, implement dead letter queues, set up monitoring and* *alerting.*

16. Explain connection pooling in JDBC and why it's important for data engineering applications. *Expected Answer: Reuses database connections to improve performance and reduce overhead. Important for high-throughput data processing applications.*

17. How do you secure JDBC connections in a production environment? *Expected Answer: Use SSL/TLS encryption, implement proper authentication, use connection pooling with timeout settings, store credentials securely.*

18. What are the different copy activity settings in ADF for optimizing data transfer performance? *Expected Answer: Parallel copies, degree of parallelism, data integration units (DIU), compression settings, fault tolerance configurations.*

19. How would you implement data validation in an ADF pipeline when ingesting from JDBC sources? *Expected Answer: Use data flow activities for validation, implement row count checks, schema validation, data quality rules, and conditional activities for failure handling.*

20. Explain the difference between Azure Blob Storage hierarchical namespace enabled vs disabled. *Expected Answer: With hierarchical namespace (ADLS Gen2), you get file system semantics with directories, better performance for analytics, POSIX permissions.*

Advanced/Difficult Level Questions (21-30)

21. Design a fault-tolerant data ingestion architecture that handles failures gracefully when moving data from multiple JDBC sources to Azure Blob Storage via ADF. *Expected Answer: Implement circuit breaker patterns, dead letter queues, checkpoint mechanisms, idempotent operations, monitoring with automatic recovery procedures.*

22. How would you optimize ADF pipeline performance when dealing with large-scale JDBC data extraction (10TB+ daily)? *Expected Answer: Implement parallel processing, partition data extraction, use incremental loading, optimize JDBC fetch size, implement data compression, use multiple integration runtimes.*

23. Explain how you would implement a CDC (Change Data Capture) solution using ADF to sync data from on-premises SQL Server to Azure Blob Storage. *Expected Answer: Use SQL Server CDC features or triggers, implement watermark tracking, design delta detection logic, handle deletes with soft delete patterns or separate tracking.*

24. How do you handle transactional consistency when ingesting data from multiple related JDBC sources into Azure Blob Storage? *Expected Answer: Implement distributed transaction patterns, use staging areas, implement compensation patterns, consider eventual consistency models with reconciliation processes.*

25. Design a solution for real-time data streaming from JDBC sources to Azure Blob Storage while maintaining exactly-once delivery semantics. *Expected Answer: Use Apache Kafka with JDBC connectors, implement idempotent operations, use Azure Event Hubs with checkpointing, design deduplication mechanisms.*

26. How would you implement a metadata-driven ADF framework that can dynamically generate pipelines for JDBC to Blob Storage ingestion? *Expected Answer: Create configuration tables, use parameterized pipelines, implement pipeline templates, use ADF REST APIs for dynamic pipeline creation, implement configuration-driven execution.*

27. Explain how you would handle data encryption at rest and in transit when moving sensitive data from JDBC sources to Azure Blob Storage. *Expected Answer: Use TLS for JDBC connections, implement customer-managed keys for Blob Storage, use Azure Key Vault for secrets management, implement column-level encryption where needed.*

28. How do you implement proper partitioning strategies in Azure Blob Storage to optimize query performance for different access patterns? *Expected Answer: Analyze query patterns, implement appropriate partitioning schemes (date, geography, customer), consider partition elimination, balance partition size, implement partition pruning strategies.*

29. Design a monitoring and alerting solution for a complex ADF pipeline that ingests data from 50+ JDBC sources to Azure Blob Storage with SLA requirements. *Expected Answer: Implement comprehensive logging, use Azure Monitor with custom metrics, set up proactive alerting, create dashboards for operational visibility, implement automated recovery procedures.*

30. How would you implement a data lineage and governance solution for data flowing from JDBC sources through ADF to Azure Blob Storage? *Expected Answer: Use Azure Purview for data catalog, implement custom lineage tracking, create data classification policies, implement automated data quality checks, maintain metadata repositories.*