03 Azure Data Factory

Azure Data Factory

Basic Level Questions (1-8)

1. What is Azure Data Factory and what are its primary use cases in data engineering?

What to look for: Understanding of ETL/ELT processes, data integration, cloud data movement, and orchestration capabilities.

2. Explain the main components of Azure Data Factory architecture.

What to look for: Pipelines, activities, datasets, linked services, triggers, and integration runtime concepts.

3. What is the difference between a Linked Service and a Dataset in Azure Data Factory?

What to look for: Linked Service as connection information, Dataset as data structure representation, and their relationship.

4. What are the different types of activities available in Azure Data Factory?

What to look for: Data movement, data transformation, and control activities with basic examples of each.

5. Explain what Integration Runtime is and the different types available.

What to look for: Azure IR, Self-hosted IR, and Azure-SSIS IR with their use cases and deployment scenarios.

6. What is a pipeline in Azure Data Factory and how do you create dependencies between activities?

What to look for: Pipeline as workflow container, activity dependencies, success/failure paths, and execution flow.

7. How do you monitor Azure Data Factory pipeline executions?

What to look for: Monitor hub, pipeline runs, activity runs, trigger runs, and basic alerting concepts.

8. What are the different ways to trigger a pipeline in Azure Data Factory?

What to look for: Schedule triggers, tumbling window triggers, event-based triggers, and manual triggers.

Intermediate Level Questions (9-17)

9. How would you implement incremental data loading using Azure Data Factory?

What to look for: Watermark patterns, lookup activities, conditional logic, and change data capture concepts.

10. Explain how you would handle error handling and retry logic in Azure Data Factory pipelines.

What to look for: Try-catch patterns, retry policies, failure paths, and notification mechanisms.

11. What is Data Flow in Azure Data Factory and when would you use it versus Copy Activity?

What to look for: Visual data transformation, complex transformations vs simple copy operations, performance considerations.

12. How would you implement dynamic pipelines that can process multiple files or datasets?

What to look for: Parameters, variables, ForEach loops, and metadata-driven approaches.

13. Explain how you would secure sensitive data and credentials in Azure Data Factory.

What to look for: Azure Key Vault integration, managed identity, service principals, and encryption.

14. How would you optimize the performance of a large data copy operation in Azure Data Factory?

What to look for: Parallel copy, data integration units (DIU), staging, and partitioning strategies.

15. What is the difference between Mapping Data Flow and Wrangling Data Flow?

What to look for: Code-free transformation vs Power Query integration, use cases, and performance characteristics.

16. How would you implement a data pipeline that processes files from multiple sources with different schemas?

What to look for: Schema drift handling, dynamic mapping, conditional transformations, and error handling.

17. Explain how you would set up CI/CD for Azure Data Factory using Azure DevOps or GitHub.

What to look for: ARM templates, branch strategies, deployment automation, and environment management.

Advanced/Difficult Level Questions (18-25)

18. Design a complex ETL solution that handles real-time and batch processing using Azure Data Factory.

What to look for: Hybrid architectures, event-driven patterns, lambda architecture, and service integration.

19. How would you implement a data lineage and governance solution with Azure Data Factory?

What to look for: Metadata capture, Azure Purview integration, data cataloging, and compliance tracking.

20. Explain how you would design a fault-tolerant, scalable data pipeline for processing TB-scale datasets.

What to look for: Distributed processing, checkpointing, recovery mechanisms, and performance optimization.

21. How would you implement a master data management (MDM) solution using Azure Data Factory?

What to look for: Data quality rules, deduplication logic, golden record creation, and data governance.

22. Design a solution for handling late-arriving data and out-of-order events in your data pipelines.

What to look for: Windowing strategies, state management, reprocessing logic, and data consistency.

23. How would you implement cost optimization strategies for Azure Data Factory in a large enterprise environment?

What to look for: DIU optimization, IR management, scheduling strategies, and resource utilization monitoring.

24. Explain how you would design a multi-tenant data pipeline architecture using Azure Data Factory.

What to look for: Tenant isolation, parameterization, security boundaries, and resource sharing strategies.

25. How would you implement a comprehensive data quality framework within Azure Data Factory pipelines?

What to look for: Validation rules, data profiling, quality metrics, alerting, and remediation workflows.

Technical Deep-Dive Scenarios

Scenario A: Performance Optimization

"Your pipeline copying 500GB of data is taking 6 hours. Walk me through your optimization approach."

Scenario B: Complex Transformation

"You need to join data from 5 different sources, apply business rules, and handle schema changes. Design the solution."

Scenario C: Error Recovery

"Your pipeline fails halfway through processing 1000 files. How do you implement resume capability?"

Scenario D: Compliance Requirements

"Design a pipeline that ensures GDPR compliance while processing customer data across multiple regions."

Hands-On Technical Questions

26. Expression and Function Usage

- "Write an ADF expression to extract the year and month from a filename like 'sales_2024_03_data.csv'"
- "How would you use the split() function to parse delimited data?"

27. JSON and REST API Integration

- "How would you handle pagination when calling a REST API in ADF?"
- "Write a pipeline to process nested JSON data with varying structures."

28. Custom Activities

- "When would you use a custom activity, and how would you implement one?"

- "Explain the difference between Azure Function Activity and Web Activity."

Follow-Up Questions:

- "How would you troubleshoot a pipeline that's intermittently failing?"
- "What's your approach to testing ADF pipelines before production deployment?"
- "How would you handle a requirement to process data in a specific order?"
- "Explain your strategy for managing pipeline versions and rollbacks."
- "How would you implement data archival policies using ADF?"

Real-World Integration Scenarios:

- Integration with Azure Synapse Analytics
- Connection to on-premises data sources
- Event-driven processing with Event Grid