prominent Academy

Tech Mahindra

AZURE DATA FACTORY (ADF)

scenario-based interview questions

Project by Prominent Academy

Scenario:

- Your organization wants to integrate ADF with an Event Hub to process real-time streaming data.
- Questions:
- How would you set up ADF to process data from an Event Hub?
- What are the limitations of using ADF for real-time processing?
- How does ADF integrate with other Azure services like Stream Analytics for real-time use cases?

Scenario:
You have a large dataset stored in Azure Data Lake that needs to be processed by date partitions.
Questions:

- How would you design a pipeline to process data in partitions?
- What are the advantages of data partitioning in ADF?
- How would you use dynamic expressions to process each partition dynamically?

Scenario:
You have multiple pipelines, and one pipeline should only start after the successful completion of another.
Questions:

- How would you implement dependencies between pipelines in ADF?
- What are the pros and cons of using execute pipeline activity versus trigger chaining?
- How would you handle scenarios where one pipeline fails but others should continue?

Scenario:
You have a master pipeline that orchestrates the execution of multiple child pipelines. Some child pipelines are dependent on the output of others.
Questions:

- How would you design the master pipeline to handle dependencies between child pipelines?
- What are the differences between the "Wait" and "If Condition" activities in this context?
- How would you monitor and troubleshoot issues in a complex pipeline execution?

Scenario:
Your pipeline needs to be triggered whenever a file is uploaded to a specific container in Azure Blob Storage.
Questions:

- How would you set up an event-based trigger in ADF?
- What are the advantages and limitations of using event-based triggers?
- How would you handle scenarios where multiple files are uploaded simultaneously?

Scenario:
Your pipeline frequently encounters transient network issues when copying data from an on-premises database to Azure SQL Database.
Questions:

- How would you implement retry logic to handle transient errors?
- What settings in ADF activities allow for retries and delays?
- How would you monitor and alert on excessive retries in pipeline executions?

Scenario:
You are required to process only the files uploaded in the last 24 hours from Azure Blob Storage.
Questions:

- How would you filter files based on their upload timestamp?
- What expressions or functions would you use to calculate time-based conditions?
- How would you handle time zone differences when filtering files?

Scenario:
Your pipeline processes files daily from Azure Blob Storage and loads them into Azure SQL Database. Duplicate files occasionally appear in the storage container.
Questions:

- How would you design a pipeline to identify and skip duplicate files?
- How would you use metadata (e.g., file names or hashes) to track processed files?
- How would you recover if a duplicate file causes partial data corruption?

Scenario:
Your pipelines process a daily batch of files, and you need to ensure that if a pipeline fails, it can resume from the last successfully processed file.
Questions:

- How would you implement state management to track processed files?
- What role do control tables play in this scenario?
- How would you ensure idempotency in pipeline executions?

Scenario:
You need to process data from multiple Azure regions and consolidate it into a central Azure Data Lake in a cost-efficient manner.
Questions:

- How would you design a pipeline to handle geo-distributed data?
- What are the cost and performance considerations for cross-region data transfers?
- How can you use regional Integration Runtimes to optimize performance?

Scenario:
You are responsible for ensuring data quality before loading it into the destination system. This includes null checks, duplicate checks, and threshold-based validations.
Questions:

- How would you implement data quality checks in ADF?
- What role do Data Flow transformations like Filter, Aggregate, and Exists play in these checks?
- How would you handle rows that fail quality checks?

Scenario:
Your pipeline processes sensitive financial data that needs to be encrypted during transit and at rest.
Questions:

- How would you ensure end-to-end encryption for sensitive data in ADF?
- How can you use Azure Key Vault for managing credentials and encryption keys?
- What security best practices would you follow to secure data pipelines?

Scenario:
You need to process data stored in partitions (e.g., year/month/day folders), but only for specific time ranges based on runtime parameters.
Questions:

- How would you configure the Copy Activity or Data Flow to read specific partitions dynamically?
- What functions or expressions would you use to skip unnecessary partitions?
- How can you optimize pipeline performance when dealing with highly partitioned data?

Scenario:
You are tasked with processing unstructured data like log files or free-form text stored in Azure Blob Storage.
Questions:

- How would you handle unstructured data in ADF?
- What external tools (e.g., Databricks, Cognitive Services) can you integrate with ADF for parsing or extracting insights?
- How would you transform this data into a structured format for downstream processing?

Scenario:
You have a pipeline with multiple parallel activities, and one of the activities fails intermittently due to source system issues.
Questions:

- How would you implement exception handling for individual activities in ADF?
- How can you ensure that the pipeline continues processing unaffected branches?
- What strategies would you use to retry or log failed activities?

Scenario:
Your pipeline needs to process files dynamically based on folder structure and file patterns in Azure Data Lake.
Questions:

- How would you use wildcard file paths in ADF to process specific files?
- How can you create folders dynamically based on runtime parameters?
- What are the challenges of managing large numbers of folders and files, and how would you address them?

Scenario:
Your team needs to collaborate with another team to build pipelines that share dependencies and datasets.
Questions:

- How would you manage shared resources (e.g., Linked Services, Datasets) across teams?
- What strategies would you use to avoid conflicts in pipeline development?
- How can Git integration help streamline collaboration between teams?

Scenario:
You need to notify stakeholders immediately when a pipeline or activity fails, including error details.
Questions:

- How would you implement real-time error notifications using Azure Monitor or Logic Apps?
- How can you configure email or SMS alerts for pipeline failures?
- What are the key metrics and logs to monitor for proactive issue detection?

Scenario:
You are part of a large organization with multiple teams working on separate ADF projects. Central governance is required for Linked Services, triggers, and naming conventions.
Questions:

- How would you implement centralized governance for ADF projects?
- How can Azure Policy or Resource Manager templates enforce naming conventions?
- What strategies would you use to manage shared Linked Services across teams?

Scenario:
You need to load data from multiple sources into corresponding tables in a destination, with dynamic schema mapping.
Questions:

- How would you configure dynamic sink mapping in a Copy Activity?
- How can parameterization help in automating schema mapping?
- What challenges might you encounter when handling mismatched schemas?

Scenario:
You need to implement data retention policies for your pipelines, ensuring that data older than a certain period is deleted or archived.
Questions:

- How would you automate data retention policies in ADF?
- What role does the Delete Activity play in this process?
- How can you monitor and validate the successful execution of retention policies?

Scenario:
You need to process semi-structured data (e.g., JSON files with varying schemas) stored in Azure Blob Storage.
Questions:

- How would you handle schema variability while processing semi-structured data in ADF?
- What transformations would you use in Mapping Data Flows to parse JSON data?
- How can you flatten hierarchical data structures for downstream consumption?

Scenario:
Your pipeline processes files in batches based on their upload time, dynamically creating batches for every 24-hour period.
Questions:

- How would you design a pipeline to identify and process dynamic file batches?
- How can you use metadata from Azure Blob Storage to determine batch boundaries?
- What challenges might arise in handling late-arriving files, and how would you address them?

Scenario:
Your pipelines need to adapt dynamically based on metadata, such as file names, schema definitions, or transformation rules stored in a database.
Questions:

- How would you design a metadata-driven pipeline in ADF?
- How can Lookup and ForEach Activities be used to retrieve and apply metadata?
- What are the advantages of a metadata-driven approach in large-scale ETL processes?

prominent Academy

Yesterday, one of our students faced a tough reality during an interview with Tech Mahindra, one of the top MNCs.

💡 Here's how we ensure success:
✅ Detailed training on ADF modules, from pipelines to debugging.
✅ Real-world project experience and mock interviews.
✅ One-on-one sessions to strengthen weak areas.
✅ Unlimited interview calls until you land your dream job!

📞 **Connect with us at**

+91 98604 38743

and get the guidance you need to land your dream job!

#SQL #DataEngineering #DataScience #DataAnalytics #CareerSupport