DATA ENGINEERING 101
Azure Data Factory

Concepts to get started

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Factory

A cloud-based data integration service that allows creation, scheduling, and orchestration of data workflows.

Building ETL processes, data migration, data transformation.

Shwetank Singh
GritSetGrow - GSGLearn.com

Pipeline

A logical grouping of activities that together perform a task. Pipelines allow activities to be linked together.

ETL pipelines, data migration pipelines, orchestrating data workflows.

Activity

Represents a single step in a pipeline. Types include data movement, data transformation, and control activities.

Copy activity for data movement, data transformation using Azure Data Lake Analytics, Databricks, or custom scripts.

Linked Service

Defines the connection information needed for Data Factory to connect to external resources.

Connecting to Azure Blob Storage, SQL Database, Cosmos DB, REST APIs.

Dataset

Represents the data structure within the linked data stores. Used in activities for reading or writing data.

CSV files in Blob Storage, tables in SQL Database, documents in Cosmos DB.

Mapping Data Flow

A visually designed data transformation feature in ADF that allows for code-free data transformations.

Data aggregation, data joins, data filtering, and transformation logic applied to incoming data streams.

Wrangling Data Flow

A feature that allows users to perform data transformation using the Power Query Editor within ADF.

Data wrangling, cleaning, reshaping, and enrichment using Power Query interface.

Shwetank Singh
GritSetGrow - GSGLearn.com

Control Flow

Orchestrates how activities are executed within a pipeline. Includes conditional, looping, and branching constructs.

If activities, ForEach loops, Until loops, Switch case activities.

Triggers

Used to schedule or trigger pipelines based on time or events.

Schedule-based triggers, tumbling window triggers, event-based triggers.

Integration Runtime (IR)

The compute infrastructure used by ADF to perform data movement and transformation.

Azure Integration Runtime, Self-hosted Integration Runtime, and Azure-SSIS Integration Runtime.

Data Movement

The process of moving data from source to destination. ADF supports various connectors and protocols.

Copy data from on-premises SQL Server to Azure SQL Database, transfer files between Blob Storage and Data Lake.

Parameterization

Allows the creation of dynamic, reusable pipelines. Parameters can be passed to pipelines, datasets, and linked services.

Passing different file names or table names to a single pipeline.

Shwetank Singh
GritSetGrow - GSGLearn.com

Monitoring

ADF provides a monitoring dashboard to view the status of pipeline runs, debug runs, and triggers.

Monitoring pipeline execution success/failure, reviewing activity logs, and diagnosing issues.

Git Integration

Allows integration with Git repositories for version control of ADF pipelines and assets.

Using GitHub or Azure Repos for managing ADF assets, promoting code changes across environments.

Debugging

ADF offers debugging capabilities to test pipeline activities before publishing.

Running pipelines in debug mode, testing activities with test data.

Data Flow Debug

Specific to Mapping Data Flows, allows for debugging of data transformation logic.

Viewing row-level data as it passes through transformation steps, testing transformations with sample data.

Global Parameters

Reusable parameters available across pipelines in a Data Factory instance.

Defining a global connection string or file path that can be used by multiple pipelines.

Expression Language

ADF's expression language allows dynamic content and logic within activities and datasets.

String concatenation, conditional expressions, date functions, and array manipulation.

Azure Key Vault Integration

Allows secure storage and retrieval of secrets, keys, and certificates used by ADF.

Securely connecting to databases using secrets stored in Key Vault, accessing API keys securely.

SSIS Integration

ADF allows running SQL Server Integration Services (SSIS) packages in the cloud with full compatibility.

Migrating on-premises SSIS packages to run in ADF using Azure-SSIS Integration Runtime.

Data Preview

Provides a snapshot view of the data in the dataset before execution, specific to Mapping Data Flows.

Previewing data after applying a transformation to validate the results.

Pipeline Parameters

Allows defining parameters within a pipeline to make it more flexible and reusable.

Creating a parameterized pipeline for copying data between different storage accounts or databases.

ADF SDKs and APIs

Provides programmatic access to create, manage, and monitor ADF resources using various languages.

Automating pipeline deployment using Azure SDKs for .NET, Python, or REST API.

Power Query

A low-code/no-code feature that allows data transformation within ADF using a familiar Power Query interface.

Data wrangling, simple transformations, data cleanup, and enrichment tasks.

If Condition

Allows conditional logic within a pipeline to execute activities based on a Boolean expression.

Execute activity A if condition is true, otherwise execute activity B.

Shwetank Singh
GritSetGrow - GSGLearn.com

ForEach

Executes a set of activities in a loop over a collection of items.

Loop through a list of files and perform a copy operation for each file.

Shwetank Singh
GritSetGrow - GSGLearn.com

Until

Executes a set of activities in a loop until a specified condition is met.

Retry an operation until a success condition is met, such as waiting for a file to arrive in storage.

Wait

Pauses pipeline execution for a specified period.

Wait for 10 minutes before proceeding to the next activity.

Execute Pipeline

Invokes another pipeline from within a pipeline.

Modularize complex workflows by calling sub-pipelines.

Switch

Allows branching logic based on a value. Similar to a switch-case statement in programming.

Execute different activities based on the value of a dataset field or parameter.

Set Variable

Sets the value of a variable that can be used later in the pipeline.

Set a variable with the current timestamp for logging purposes.

Shwetank Singh
GritSetGrow - GSGLearn.com

Append Variable

Appends values to an array variable.

Collect and store file names processed during pipeline execution.

Copy Data

Moves data from a source to a destination. Supports over 90 data connectors.

Copy data from SQL Server to Azure Data Lake Storage, transfer files from SFTP to Blob Storage.

Lookup

Retrieves data from a dataset and makes it available for subsequent activities.

Look up a value in a SQL table to use in a conditional activity.

Get Metadata

Retrieves metadata information (e.g., file size, last modified date) from data in a dataset.

Check the size of a file before processing it.

Mapping Data Flow

Provides a code-free experience for transforming data at scale.

Join, filter, aggregate, and perform complex transformations on data in a visually designed data flow.

Wrangling Data Flow

Allows for data preparation and transformation using Power Query.

Data cleanup, normalization, and transformation using Power Query language.

Shwetank Singh
GritSetGrow - GSGLearn.com

Data Flow Debug

Enables testing and debugging of Mapping Data Flows with real data.

Test and validate transformations by inspecting row-level data during the design phase.

Shwetank Singh
GritSetGrow - GSGLearn.com

Databricks Notebook

Executes an Azure Databricks notebook as part of a pipeline.

Run complex data transformations, machine learning models, or custom logic in Databricks.

Azure HDInsight

Runs jobs on an Azure HDInsight cluster, including Hive, Pig, Spark, etc.

Process big data using Hadoop or Spark, run Hive queries, or Pig scripts.

Azure Machine Learning

Invokes an Azure Machine Learning pipeline endpoint to run ML models.

Trigger a machine learning model training or scoring pipeline as part of a data processing workflow.

Shwetank Singh
GritSetGrow - GSGLearn.com

Stored Procedure

Executes a stored procedure in a relational database.

Run a stored procedure in Azure SQL Database or SQL Server to perform complex database operations.

Shwetank Singh
GritSetGrow - GSGLearn.com

Web Activity

Makes a call to a REST endpoint.

Trigger a web service, pass data between ADF and external systems via REST APIs.

Azure Function

Executes an Azure Function as part of a pipeline.

Run serverless code, perform lightweight transformations, or trigger external processes.

Execute SSIS Package

Executes an SSIS package stored in Azure SQL Database, Azure File Storage, or SSISDB.

Migrate and run existing SSIS ETL packages in the cloud using Azure-SSIS Integration Runtime.

Custom Activity

Runs custom code on an Azure Batch Service.

Execute complex transformations or logic that isn't supported by out-of-the-box activities using custom code.

Aggregate Data Flow

Summarizes data by performing aggregation operations.

Calculate sum, average, count, min, max, etc., on a dataset.

Conditional Split

Data Flow

Routes data rows to different streams based on a condition.

Filter data into different outputs based on conditional logic.

Shwetank Singh
GritSetGrow - GSGLearn.com

Derived Column

Data Flow

Adds or replaces columns in the dataset with derived values.

Create new columns by applying expressions or transformations on existing columns.

Shwetank Singh
GritSetGrow - GSGLearn.com

Filter **Data Flow**

Filters rows in the dataset based on a Boolean expression.

Remove unwanted rows from the dataset based on a condition.

Shwetank Singh
GritSetGrow - GSGLearn.com

Join Data Flow

Combines two datasets based on a common key or condition.

Perform inner, outer, left, or right joins on datasets.

Lookup Data Flow

Performs a lookup of data in another dataset and returns the matching rows.

Enrich data by adding columns from another dataset based on a key.

Shwetank Singh
GritSetGrow - GSGLearn.com

Select **Data Flow**

Chooses which columns to include or exclude in the output.

Remove unnecessary columns from the output.

Sort Data Flow

Orders the rows in a dataset based on specified columns.

Sort data before performing further operations like windowing.

Union Data Flow

Combines rows from two or more datasets into a single dataset.

Merge datasets with the same schema.

Shwetank Singh
GritSetGrow - GSGLearn.com

Window Data Flow

Performs window functions, such as ranking or calculating moving averages, based on a specified window.

Calculate running totals, moving averages, or rank rows within partitions.

Pivot **Data Flow**

Converts rows into columns based on an aggregation function.

Transform categorical data into a more suitable format for analysis.

Unpivot **Data Flow**

Converts columns into rows to normalize the data.

Flatten a wide table into a long table for easier processing.

Exists Data Flow

Filters rows based on the existence of corresponding rows in another dataset.

Keep only the rows that have matching entries in another dataset.

Thank You