Azure Data Factory

- ADF USE CASES
- DATAFLOW and its features

What is Azure Data Factory

- Azure Data Factory is Azure's cloud ETL service for scale-out serverless data integration and data transformation. You can also lift and shift existing SSIS packages to Azure and run them with full compatibility in ADF.
- It is the cloud-based ETL and data integration service that allows you to create data-driven workflows for orchestrating data movement and transforming data at scale.

Top Level Concepts in Data Level Factory

- Pipeline
- Activity
- Data Sets
- Linked Services
- Triggers

Pipeline

- A data factory might have one or more pipelines. A pipeline is a logical grouping of activities that performs a unit of work.

  For example, a pipeline can contain a group of activities that ingests data from an Azure blob, and then runs a Hive query on an HDInsight cluster to partition the data.

Activity

- Activities represent a processing step in a pipeline. For example, you might use a copy activity to copy data from one data store to another data store.

Linked Services and Data Sets

- Linked Services are much like connection strings, which define the connection information that's needed for Data Factory to connect to external resources.
- Datasets represent data structures within the data stores, which simply point to or reference the data you want to use in your activities.
- For example, an Azure Storage-linked service specifies a connection string to connect to the Azure Storage account. Additionally, an Azure blob dataset specifies the blob container and the folder that contains the data.

Triggers

- Triggers Determines when a pipeline execution needs to be kicked off. There are different types of triggers.

Pre-requisites

- Azure Subscription
- Azure Storage Account

Azure Data Factory

- Azure Portal UI
- Azure PowerShell (Install Azure PowerShell)
- .NET
- Python
- REST
- Resource Manager Template (Azure PowerShell Az Module)

Pipelines and Activities

- Pipeline is logical grouping of activities that together perform a task.
  Eg: Pipeline can have a set of activities that take data from ADLS and perform some transformation of data using U-SQL and load data in to SQL DB.

- Activities in a pipeline defines actions to perform on data.
  Eg. Copy Data activity can read data from one location of Blob storage and load it to Other location on Blob Storage.

Pipelines and Activities

Datasets identify data within different data stores, such as tables, files, folders and documents.

Linked Services and Datasets

- Linked services are used to connect Other resources with Azure Data Factory. Linked services are like connection strings resources to connect.
- Datasets are simply points or references the data, which you want to use in your activities as input or output

Triggers in Azure Data Factory

- Triggers - you can execute your pipeline.
- Triggers determine when a pipeline execution needs to be kicked off.
- Pipelines and triggers have a many-to-many relationship (except for the tumbling window trigger).
- Multiple triggers can kick-off a single pipeline, or a single trigger can kick off multiple pipelines.

Types of Triggers in Azure Data Factory

- Below are the Types of Triggers available In Azure Data Factory
  1. Schedule Trigger - A trigger that invokes a pipeline on a wall-clock schedule.
  2. Tumbling Window Trigger - A trigger that operates on a periodic interval, while also retaining state.
  3. Event-Based Trigger - A trigger that responds to an event.

Schedule Trigger in Azure Data Factory

- Schedule trigger runs pipelines on a wall-clock schedule. This trigger supports periodic and advanced calendar options.
  For example, the trigger supports intervals like “weekly” or “Monday” at 5:00 PM and Thursday agt 9:00 PM

  https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-schedule-trigger

Tumbling Window Trigger in ADF

- Tumbling window triggers are a type of trigger that fires at a periodic time interval from a specified start time, while retaining state.
- A tumbling window trigger has a one-to-one relationship with a pipeline and can only reference a singular pipeline.

Tumbling Window Trigger Dependency in ADF

- In order to build dependency chain and make sure that a trigger is executed only after successful execution of another trigger using Tumbling window Trigger Dependency Feature.

Note: A tumbling window trigger can depend on a maximum of two other triggers.

- You will be having access to Window Start Time and Window End Time values using below System Properties.

  trigger().outputs.windowStartTime

  trigger().outputs.windowEndTime

Integration Runtime in ADF

- The Integration Runtime (IR) is the compute infrastructure used by Azure Data Factory to provide the following data integration capabilities across different network requirements :
  - ➡ Data Flow: Execute a Data Flow in managed Azure compute environment
  - ➡ Data Movement: Copy data across data stores in public network and private network
  - ➡ Activity Dispatch: Dispatch and monitor transformation activities running on a variety of compute services
  - ➡ SSIS Package Execution: Execute SQL Server Integration Services (SSIS) packages in a managed Azure compute environment.

Types of Integration Runtimes

Data Factory offers three types Integration Runtime, and you should choose the type that best serve the data integration capabilities and network environment needs you are looking for. These three types are :

Azure

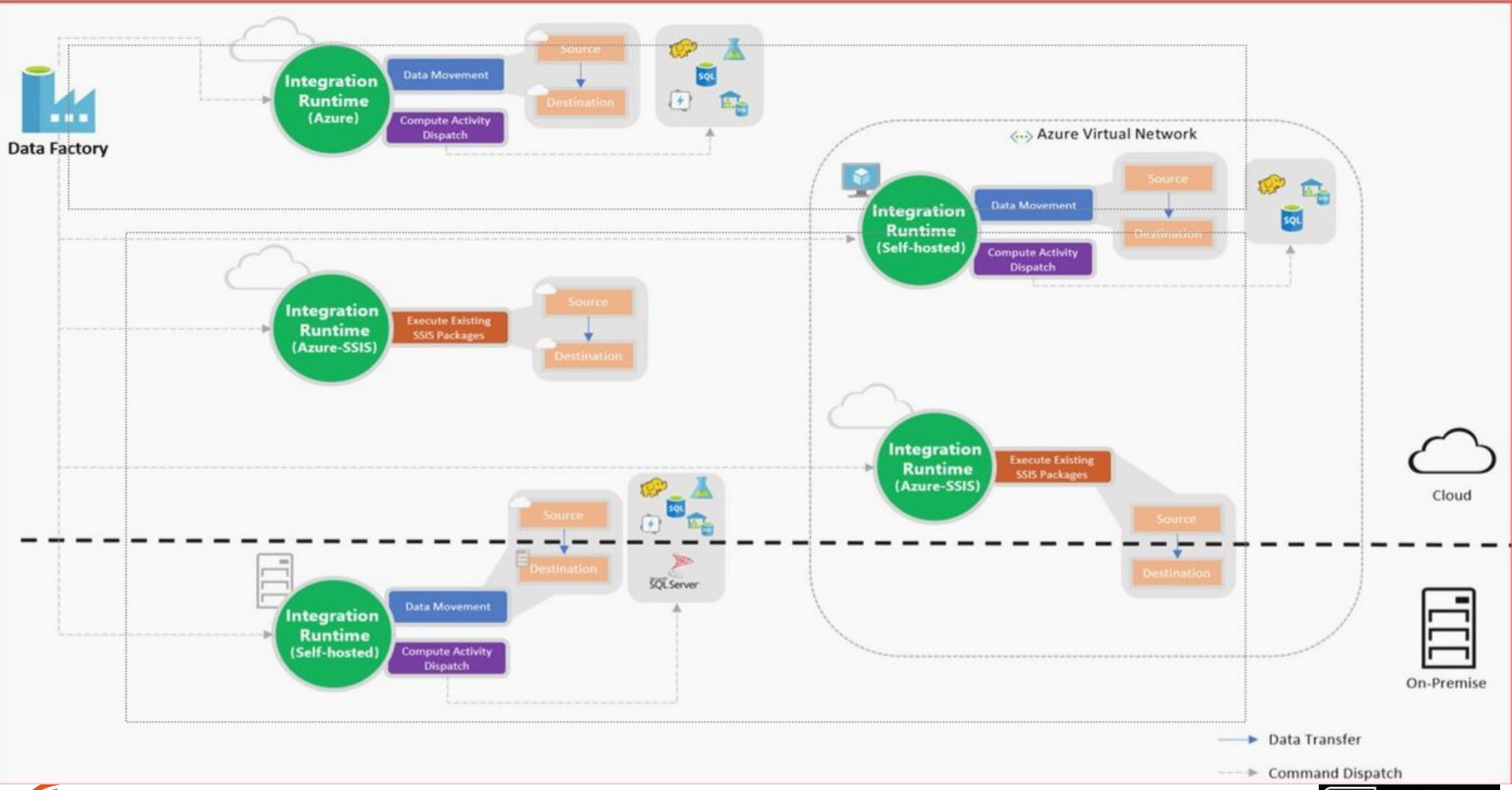
Self-hosted

Azure SSIS

IR type	Public network	Private network
Azure	Data Flow Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

Self-Hosted Integration runtime

- Self-Hosted Integration runtime is capable
  - Performing data movement activities between cloud data stores and in private network
  - Running transform activities against compute resources in on-premises or Azure virtual network
- Self-hosted IR needs to be installed on an on-premise machine or a virtual machine inside a private network. Currently, its supports running the self-hosted IR on a Windows operating system.


Data Factory
Integration Runtime (Azure)
Data Movement
Compute Activity Dispatch
Source
Destination
Azure Virtual Network
Integration Runtime (Self-hosted)
Data Movement
Compute Activity Dispatch
Source
Destination
Integration Runtime (Azure-SSIS)
Execute Existing SSIS Packages
Source
Destination
Integration Runtime (Azure-SSIS)
Execute Existing SSIS Packages
Source
Destination
Cloud
SQL Server
Integration Runtime (Self-hosted)
Data Movement
Compute Activity Dispatch
Source
Destination
On-Premise
Data Transfer
Command Dispatch

Data flows in ADF

- Data flows feature in Azure data factory will allow you to develop graphical data transformation logic that can be executed as activities in ADF pipelines
- Your Data flow will execute on your own Azure databricks cluster for scaled out data processing using spark.
- ADF internally handles all code translation, spark optimization and execution of transformation.

Filter Transformation

- The Filter transforms allows row filtering based upon a condition. The output stream includes all rows that match the filtering condition. The filter transformation is similar to a WHERE clause in SQL.

JOIN Transformation

- Use the join transformation to combine data from two sources or streams in a mapping data flow. The output stream will include all columns from both sources match based on a join condition.

Conditional Split Transformation

- The conditional split transformation routes data rows to different streams based on matching conditions. The conditional split transformation is similar to a CASE decision structure in a programming language.

Derived Column Transformation

- Use the derived column transformation to generate new columns in your data flow or to modify existing fields.

Exists Transformation in Data Flow

- The exists transformation is a row filtering transformation that checks whether your data exists in another source or stream.The output stream includes all rows in the left stream that either exist or don’t exist in the right stream. The exists transformation is similar to SQL WHERE EXISTS and SQL WHERE DOES NOT EXISTS.

Lookup Transformation in Data Flow

- A lookup transformation is similar to a left outer join. All rows from the primary stream will exist in the output stream with additional columns from the lookup stream.

Select Transformation in Data Flow

- Use the select transformation to rename, drop, or reorder columns. This transformation doesn't alter row data, but chooses which columns are propagated downstream.

Thanks!

Contact us:
training@apps2fusion.com
+44 207 101 9262
+ 1 212 404 1735

www.apps2fusion.com