20/2/24

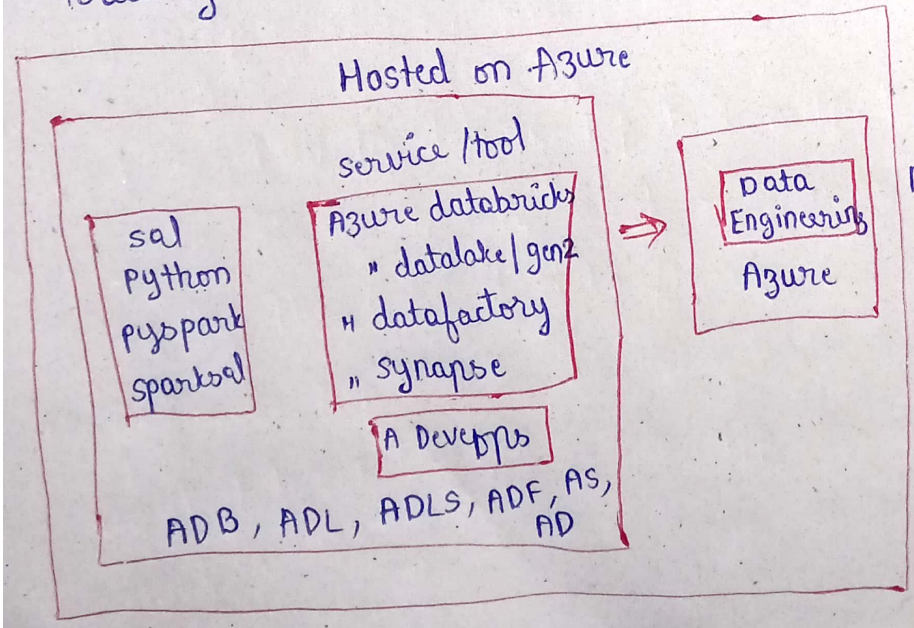
Azure Data Factory

* It is a cloud-based data integration service that allows you to create data-driven workflows in the cloud for orchestrating & automating data movement and data transformation.

* ADF doesnot store any data itself.
  (i) allows to create data-driven workflows for move-ment b/w supported data stores
  (ii) process data using compute services in other regions
  (iii) monitor & manage workflows - using UI mechas.

Uses / Use cases

1) supporting data migration
2) Getting data from client's server or online data to Azure Data Lake
3) for various data integration processes.
4) Integrating data from various ERP systems & loading it into Azure Synapse for reporting.

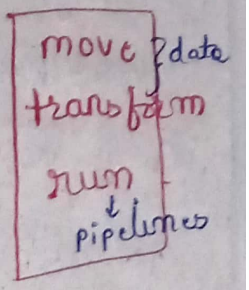


ADF :-
data driven pipeline
move, transform, run data in the pipeline of ADF

3 steps :-
1) Connect & Collect
2) Transform & Enrich
3) Publish

How Azure Data Factory work ?

* It allows to create data pipelines that move and transform data
* and then run pipelines in schedule (hourly, daily).
* Data that is consumed and produced by workflows is time-sliced data and we can schedule time



Steps :- Three steps

1) Connect and Collect 2) Transform and Enrich

3) Publish

① Connect & collect :-

* Connect to all required sources of data and processing such as saas services, file shares, FTP & web services.
* Use copy activity in data pipeline to move data from both on-premise and cloud source data stores to a centralization data store in the cloud for further analysis.

② transform & Enrich

* once data is present in centralized data store in cloud, it can be transformed using compute services such as HD Insight Hadoop, Spark, ADL Analytics and ML.

③ Publish

* Deliver transformed data from cloud to on-premise sources like SQL server or in cloud storage sources.

Data Migration activity with ADF

* By using ADF, data migration occurs b/w on-promise data store and cloud data store

Copy activity :- copies data from source data store to sink data store.

* Azure supports various data stores such as source or sink data stores like
  1) Azure Blob Storage
  2) Azure Cosmos DB
  3) Azure data lake store
  4) Oracle
  5) Cassandra

* ADF supports transformation activities
  1) Hive
  2) MapReduce
  3) Spark

  } can be added to pipelines either individually or chained with other activities

* For visualization → Power BI, Azure Synapse

Key Components of ADF

⇓

Work together to define input & output data, process-ing events and schedule and resources required to execute the desired data flow.

1) Datasets represent data structures within data stores

input dataset → represents i/p for an activity in pipeline

output " → " o/p " " " " "

Ex:- Azure Blob dataset ⇒ specifies blob container & folder in Azure blob storage from which pipeline should read data

i/p

SQL table dataset ⇒ specifies table to which o/p data is written by activity

o/p

2) A pipeline is group of activities

* they are used to group activities into unit that together performs task
* DF can have one (or) more pipelines

Ex:- pipeline → group of activities → 1) ingest data from Azure Blob
2) runs a Hive query on HDInsight cluster to partition data

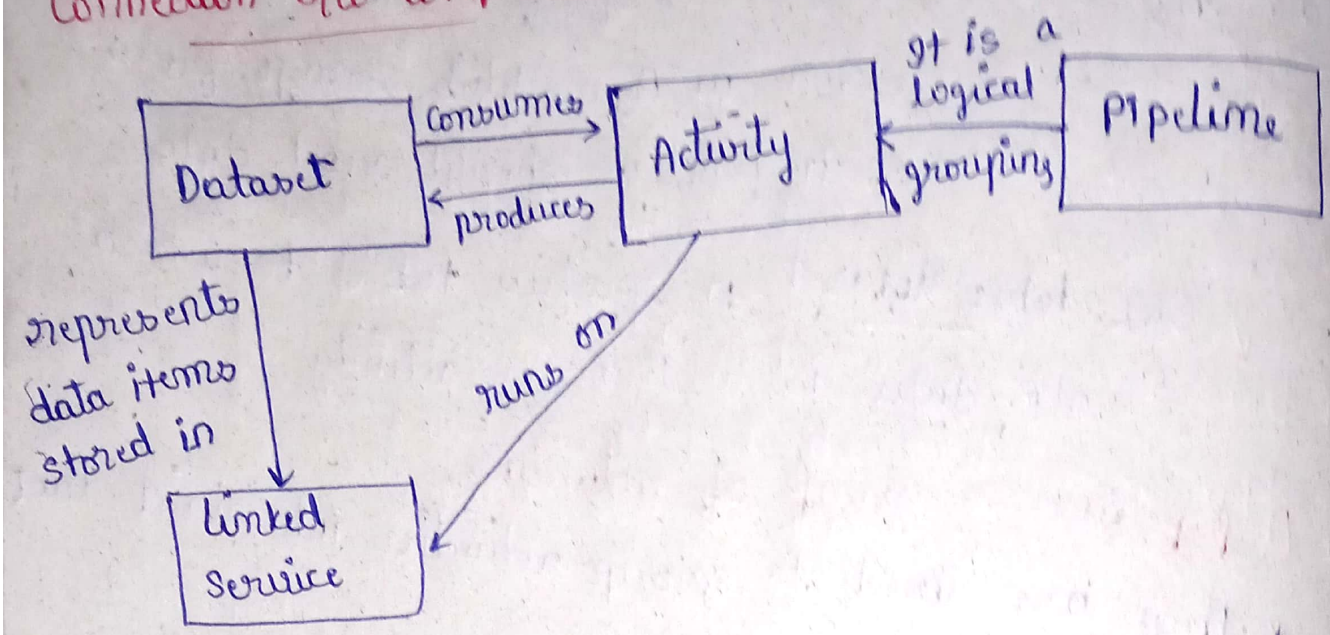
3) Activities define actions to perform on data

* ADF supports two types of activities
  1) Data movement
  2) Data transformation

4) Linked services define information needed for ADF to connect to external resources.

* Example :- Azure storage linked service specifies a connection string to connect to Azure storage account.

Connection b/w Components



Tools / APIs to create data pipelines in ADF

1) Azure portal
2) Visual Studio
3) Powershell
4) .NET API
5) REST API
6) Azure Resource Manager Template.