16/2/24

Data Lake Storage Gen2 is created

ADLS => Azure data Lake storage GEN-2 => Service name in Azure

Note :-

* If we want to store data without performing analysis on data, set Hierarchical Namespace option to Disabled.
* You can also use blob storage to archive rarely used data or to store website assets such as images and Media

Datalake Storage Gen2

* If we are performing analytics on data, set up the storage acc as an Azure Data Lake Storage Gen2 account by setting the Hierarchical Namespace option to Enabled.
* Both ADLS Gen 2 is integrated into Azure Storage platform, appln can use either the Blob APIs or Azure Data Lake Storage Gen2 file system APIs to access data.

Creating ADLS account

Search for storage account

↓

account name :- adls gen2 storage1049

↓

Central India

⇓

Standard

⇓

GRS

⇓

Go to Advanced

⇓

Enable Hierarchical Namespace

⇓

Review + Create

⇓

Deployment is in progress

⇓

Deployed successfully

⇓

Go to resource

⇓

Left side Configuration

⇓

Enable Anonymous Blob Access

⇓

Refresh

⇓

Left side ⇒ Container

⇓

Create container

⇓

Upload file

⇓

Left side ⇒ Access key

⇓

Copy Connection string

⇓

Go to storage Explorer

⇓

Connect to Azure Resource

⇓

storage account or Service

⇓

Connection string

⇓

Enter Connection Info

⇓

Name to display

paste URL

⇓

Connect

It displays new connection is added

— x —

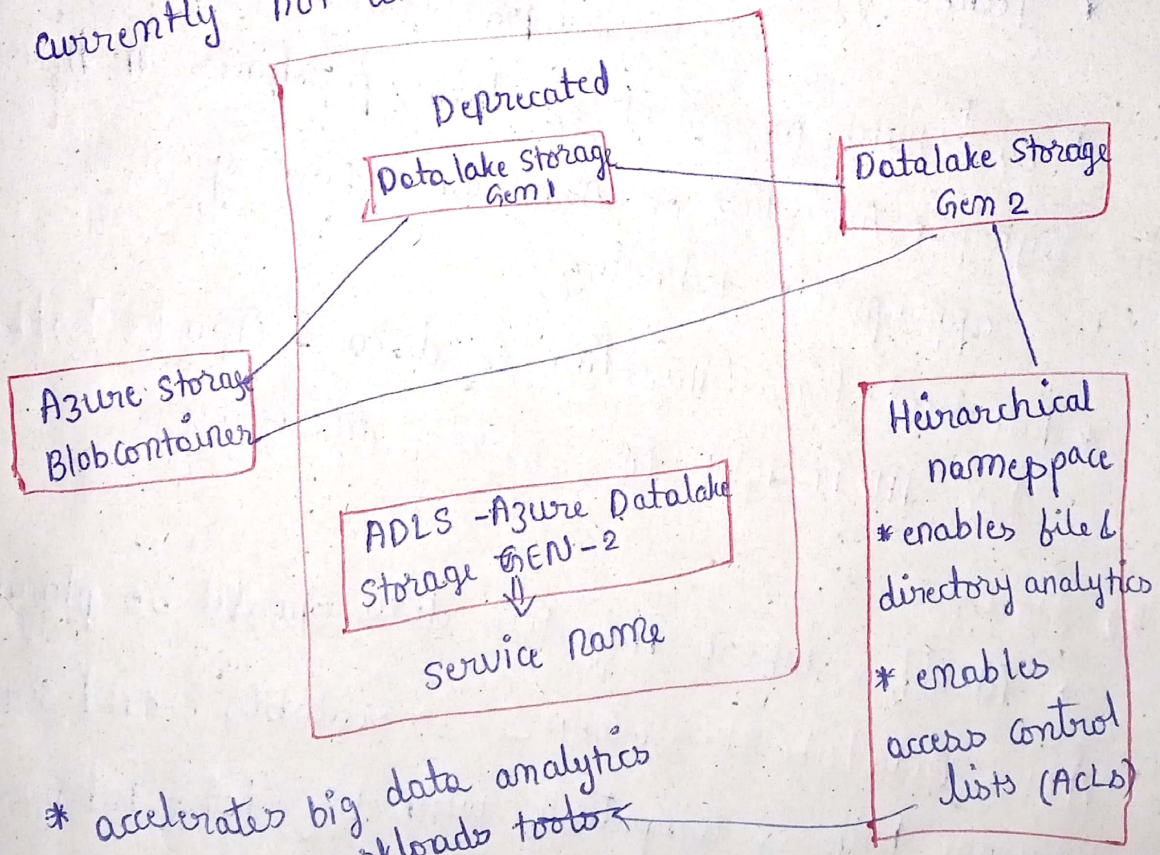
ADLS

* In Azure Blob storage, we can store large amounts of unstructured ("object") data in a flat namespace within a blob container.

* Azure Data Lake Storage Gen 2 ⇒ builds on blob storage and optimizes I/O of high-volume data by using hierarchical namespace that organizes blob data into ~~directories~~ directories and stores meta data about each directory.

* Hierarchical namespace ⇒ keep the data organized which yields better storage and retrieval performance for analytical use and lowers cost of analysis.

* Azure Data Lake Storage Gen 1 ⇒ deprecated and currently not used.



* accelerates big data analytics workloads tools

* Complemented by Datalake Storage Gen-2

Understand the stages for processing big data

* Data lakes have fundamental role in wide range of big data architectures.
* These architectures involve creation of
  (i) enterprise data warehouse
  (ii) Advanced analytics against big data
  (iii) real-time analytical soln

Four stages

1) Ingest :-
   * This phase identifies technology & processes that are used to acquire source data
   * This data comes from logs, files and other unstructured data in data lake
   * Technology used → depends on freq of data transferred.
   * Ex :- For batch movement of data, pipelines in Azure Synapse Analytics or Azure Data Factory are appropriate technology
   * For real-time ingestion of data - Apache Kafka for HDInsight (or) Stream Analytics

2) Store :-
   * identifies where ingested data should be placed
   - Azure Data Lake Storage Gen 2 ⇒ scalable and secure storage soln ⇒ Compatible with big data technologies

3) Prep and train :-

* identifies technologies that are used → to perform data preparation, model training & scoring for ML solns.

Common Technologies ⇒ Azure Synapse Analytics, Azure Databricks, Azure HDInsight, Azure ML.

4) Model & Serve :-

* involves technologies that will present data to users

* Visualization tools - MS power BI or analytical data stores such as Azure Synapse Analytics.

* Combination of multiple technologies will be used depending on business requirements.