**STEP for ETL Development**

1.Create Azure Account: Azure free trial.

2.Storage Account>>Create Storage Account>> Free trial + create new resource group+ name + India region+ standard performance >>next >> hierarchical name space (data lake gen2) >>review >>create

3.Data Storage>> container >>name >>create >> click on the created container>>add directory>>name>> create raw data and transformed data directory or folder.

4.Data Factories >>create data factory >>add existing resource group+name+region>next>>next>>review>>create>>click on created pipeline>> launch studio >>author>>pipeline>>activities>>move and transform>>drag and drop>>copy data>>click on the component>>it has general,source,sink tab.

General

Provide name of the file or data

Source

a.Upload csv files to git hub

b.Open the raw file and copy the url

Source tab in ADF>>new>>http>>delimited text>>linked services(creating link from ADF source to git hub)>>new>>name+base url(paste the raw data git url)+authentication type(anonymous)>>create>>mark first row as header+none>>ok>>preview data(check everything looks good or not)

Sink Tab

Sink>>new>>azure data lake>>continue>>delimited text>>name>>linked services>>new>> subscription(free account) + storage account>>create>>filepath>>browse>>click on the storage account created by you>>raw folder>>add file name with extention + import schema(none)>>ok

Validate>>debug>>if everything looks good then publish

I

5.Azure data brick>>create>>subscription(free trial) + Resource Group+workspace name + region+pricing(premium) >> review + create

Click to the databrick created above>>go to resources >>launch workspace>> compute

a.policy:unrestricted

b.single node(multinode machines scaling horizontally)

c.memory core 14 gb 4 core

d.create compute

new>>notebook>>give it a name

What is done here?

Create connection from azure data bricks to our azure data storage to easily access the data

Steps to connect

a.Go to databrick and connect to the dedicated notebook

b.Search for app registration in azure portal

new registration>>name>>register>>copy client and tenant id

certificate & secret>>client secret >>new client secret key>>name>>add>>copy secret key value

c.Go to data brick notebook: configure and mount(provided in the code)

d.give permission to access to azure app data from data lake.The above is the configuration the

below is to establish connection

Go to your container>> IAM >>add role assignment >> storage blob >>contributer

>>next>>select member(give app name)>>next >>review assign

6. Read and Write file

a.In jupyternotebook we need to create spark session object but in databricks it is not required.It is implicitly created.

b.Go to jupyternotebook

7.Create Synapse Analytics

Synapse aanalytic>>create>>free trial+workspace name+India region+data lake accout name+filesystem>>review and create

Click to the synapse created >>open synapse studio>>data>>lake database>>give it a name>>click table>>click +Table >>from data lake >>give table name(table name of transformed data) >> linked services(select default one) >>input file folder(select the file:part file) >>continue >> infer column names >. Continue >>check everything about the table, if it looks good then validate and publish.

**STEPS for Dashboarding**

1.Create separate container for parquet under storage account

2.Create new pipeline to convert CSV to parquet

Synapse>>integrate>> pipeline>>move and transform>copy data >>name it

General Tab

Name: Give a proper name

Source Tab

+>>data lake storage gen>>delimited text>>name>>linked services >> select existing link>>add all the csv files from raw data

Sink Tab

+>>datalake storage gen2>>parquet>>name>>linked service(existing)>>select folder ad gove file name with extension

3.Debug pipeline: do it for each csv files

4.Create new lake spark database

Synapse>>data>>+>>lake database>>name it>>create new spark table>>name+same linked service>>point to the parquet file>ok>>continue>>create

Create for all parquet files

5. Query spark table using serverless sql

Synapse>>”your created workspace”>>copy serverless sql endpoint

**OPEN POWER BI**

Get data>>azure sql database>>server:paste the above url + database + Direct query >>using Microsoft account login >> load all the tables.