STORAGE FOR AZURE:-

1. Blob(text file,videos,audio,CSV,MS Excel)
2. File(Access file from different location[ github ])
3. Table
4. Queue

TYPES OF TIER

HOT TIER (data that access frequently, higher storage cost but lower access cost)

COOL TIER (data that access infrequently, lower storage cost but higher access cost[30 days])

ARCHIVE TIER (data that access rarely, lower storage cost but higher access cost and higher retravel latency [180 days])

MyBDA

Youexcel

Mydatabase

aamna.database.windows.net

id name : aamnanazawan

AZURE DATA FACTORY(DAF)

Pipline

Activities

Datasets

Linked services

Integration runtime

Triger

Monitring

FOR DELETING MULTIPLE FILES(\*.CSV)

**PROCEDURE IN SQL:-[for deleting any ID]**

CREATE PROCEDURE DELETECUSTOMER @ID INT

AS DELETE FROM mahrukh WHERE ID=@ID

GO;

EXEC DELETECUSTOMER @ID=1

**PROCEDURE IN SQL:-[for inserting any ID]**

CREATE PROCEDURE INSERTDATA

@ID INT,

@Name VARCHAR(255),

@Salary INT

AS

BEGIN

INSERT INTO mahrukh (ID, Name, Salary)

VALUES (@ID, @Name, @Salary);

END;

GO

EXEC INSERTDATA

@ID = 1,

@Name = 'Aamna',

@Salary = 50000;

ANOTHER WAY TO INSERT DATA:-

CREATE PROCEDURE INSERTDATA1 @ID INT, @NAME VARCHAR(255), @SALARY INT

AS INSERT INTO mahrukh (ID, NAME, SALARY) VALUES (@ID, @Name, @Salary)

Synapse SQL have two consumption models:-

Dedicated(it charge more amount than serverless and per hour charge, openrowset does not apply on this) and **serverless**(its free)

Parquet file

OPENROWSET() is used to open file in azure aynapse

Code for open anyfile also file URL were taken from properties

SELECT \* FROM

    OPENROWSET(

        BULK 'https://youexcel.dfs.core.windows.net/aamna/mid.csv',

        FORMAT ='CSV',

        HEADER\_ROW = TRUE,

        PARSER\_VERSION = '2.0'

    ) AS [result]

(FOR TRANSFERING DATA FROM CSV TO SQL)

CREATE TABLE [dbo].[department]

(

Name nvarchar(218),

Age nvarchar(218),

City nvarchar(218)

)

WITH

(

DISTRIBUTION – ROUND\_ROBIN,

CLUSTERED COLUMNSTORE INDEX

--HEAP

)

GO

COPY INTO [dbo].[dept]

(Name 1, Age 2, City 3)

From 'https://youexcel.dfs.core.windows.net/aamna/mid.csv'

WITH

(

    FILE\_TYPE = 'CSV',

    MAXERRORS = 0,

    IDENTITY\_INSERT = 'OFF'

)

SELECT \* FROM [dbo].[dept]

Apache Spark pool(it works on parallel means to divide the work and do in less time also parquet file used in it) no of nodes used in it

We used this[df.printSchema()] instead of df.info()

QUERIES FOR APACHE SPARK NOTEBOOK USING PARQUET FILE

%%pyspark

df = spark.sql("SELECT \* FROM `nyctaxi`.`trip`")

display(df)

%%pyspark

df = spark.sql("""

  SELECT passenger\_count,

    SUM (trip\_distance) as SumTripDistance,

    AVG(trip\_distance) as AvgTripDistance

  FROM nyctaxi.trip

  WHERE trip\_distance>0 AND passenger\_count>0

  GROUP BY passenger\_count

  ORDER BY passenger\_count

""")

display(df)

df.write.saveAsTable("nyctaxi.passengercountstats")

creating sql table to csv and parquet file(only just write this command on notebook and then go to integrate button and set the pipeline and transfer data)

df = spark.sql("SELECT \* FROM nyctaxi.passengercountstats")

df1= df.repartition(1)

df1.write.mode("overwrite").csv("/nyctaxi/csvfile")

df.write.mode("overwrite").parquet("/nyctaxi/parquetfile")

External data-resources

External File format

If we write false in file format then it store all missing values as null.

Encoding UTF8|UTF16(Techniques for excel conversion)

FOR CREAING ANY EXTERNAL DATASOURCE

Use you

GO

CREATE MASTER KEY ENCRYPTION BY PASSWORD = 'Welcome1$Hello@'

CREATE DATABASE SCOPED CREDENTIAL demoCredential

WITH IDENTITY = '',

SECRET = 'sv=2022-11-02&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2024-08-15T18:11:47Z&st=2024-08-15T10:11:47Z&spr=https&sig=uH2dR81FraAvlDtnpWWrNz0LejI14YvnoCpp8vGIkHU%3D'(Shared access signature check mark all three of [all resources]-generate SAS token and then copy sas token)

GO

CREATE EXTERNAL DATA SOURCE demoDataSource WITH(

    LOCATION = 'https://youexcel.dfs.core.windows.net/', (Blob storage-select container-settings-endpoint-datalake storage-primary endpoint)

    CREDENTIAL = demoCredential

);

FOR CREAING ANY EXTERNAL FILEFORMAT(PARQUET)

CREATE EXTERNAL FILE FORMAT ParquetFileFormat

WITH

(

    FORMAT\_TYPE = PARQUET,

    DATA\_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'

)

FOR CREAING ANY EXTERNAL FILEFORMAT(CSV)

CREATE EXTERNAL FILE FORMAT CSVFileFormat

WITH

(

    FORMAT\_TYPE = DELIMITEDTEXT,

    DATA\_COMPRESSION = 'org.apache.hadoop.io.compress.GzipCodec',

    FORMAT\_OPTIONS (

        FIELD\_TERMINATOR = ',',

        STRING\_DELIMITER = '"',

        FIRST\_ROW = 1,

        USE\_TYPE\_DEFAULT = TRUE,

        ENCODING = 'UTF8'

    )

);

CETAS(CREATE EXTERNAL TABLE AS SELECT)

VIRTUAL MACHINE

\Users\student\Downloads\bdaclass\_key.pem [azureuser@20.2.220.10](mailto:azureuser@20.2.220.10)

Copy file path(\Users\student\Downloads)

Linux command

ssh -i C:\Users\student\Downloads\bdaclass\_key.pem azureuser@20.2.220.10

Ls(show the quantity of files)

Cd ..(one back folder)

mkdir [folder name](creating new folder)

rmdir [folder name](removing any folder)

cd(change directory)

rm [filename](removing any file)

touch one.txt (for creating any empty file)

nano one.txt(for writing anything in file)

cat one.txt(for reading any file)

pwd:

owner = u, group = g, other = o

drwxrwxr-x (d represent directory, rwx represent owner)

-rw-rw-r (r read, w write, owner can read and write and guest can only read)

* Represent file

Ls-l for specific file

Chmod u+x filename(for permission change, x means execute, u means users)

+ given permission

-Taking permission

Chmod o=rx filename(set read and execute permissions for others)

Read = 4 (chmod 4)

Write = 2 (chmod 2)

Execut = 1 (chmod 1)

For rwx = 7 (chmod 7)

Group = 2, other = 1

Bash sript

File extension is.sh

#!/bin/bash

Echo “hello world” (print Hello world)

For running script

./(file name)

For installing anything

Sudo apt-get update

Sudo apt-get install python3

Printing for loop

1 For i in {1..5}

Do

Date

Sleep 5

Echo “Print $i”

Done