Azure Data Lake Gen 2

==========

Azure Storage account

so the only change we did to create this is that we checkedin the option of datalake gen2 when

creating the storage account -> under the advanced section

the icon is different for the containers when we talk about blob and datalake gen2

1. when we are creating a datalake storage it supports hierarchy

when talking about the blob storage there is a flat hierarchy.

2. when talking about big data analytics workloads then the datalake gen 2 storage account is

more performant than the raw blob storage account.

Low cost

storage - hot, cold, archive

high availability

Continuation on Azure Data Lake Gen

==========

ACL - access control list

https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

Access tiers -

1. hot - access is faster

2. cold/cool - access will be a little slower

3. archive - that we want a backup

archive is available at the blob level

by default the access tier is hot

cool -> hot (if we try to change the access tier from cool to hot in first 30 days then an early

deletion fee is charged)

archive -> hot/cool (if we try to change the access tier from archive to hot/cool before 180 days

then an early deletion fee is charged)

rehydration (a process that is followed when we move a blob from archive to hot/cool) it takes

several hours

standard/high

storage cost

==========

hot access tier - highest

cool - medium

archive - lowest

retrieval cost

==========

hot - least

cool - medium

archive - most

Azure storage pricing

https://azure.microsoft.com/en-in/pricing/details/storage/blobs/

Lifecycle management of storage accounts

==========

Manually changing the access tiers is not a good solution and is not practically possible

always.

Cost of blob storage

==========

1. volume of data stored per month

2. quantity and types of operations performed

3. Data transfer cost

4. Data redundancy option that are selected

Compare

Hadoop - HDFS

AWS - s3

Azure - Data Lake Gen2

GCP - Google Storage

Azure Databricks - session 1

==========

what is databricks and why databricks?

Databricks - a company created by creators of apache spark

when we talk about open source version of spark

=> infrastructure management

=> software installation

=> upgrade

=> lack of user interface

=> managing the security

Apache spark based unified analytics platform optimized for cloud.

Pricing of Databricks -

Infrastructure charges

Software charges

Azure + Databricks

Microsoft Azure is provding us the hardware

software charges are taken by Databricks

4 machines - 4 cores and 8 GB RAM

.5$ per hour (2$ per hour)

.5 DBU per machine (2 DBU)

.5$ * 2 = 1$

is databricks offering only given on azure cloud?

azure

aws

gcp

databricks is managed 1st party service on azure

transparent

security

unified billing

Azure Databricks - session 2

==========

cluster - driver node and a bunch of worker nodes

Notebook - your code will be written here

3 options when creating a cluster

==========

1. all purpose cluster (interactive purpose)

2. job cluster (during the job run the cluster will be created and once the job finishes it terminates)

3. pool

All purpose

==========

created manually

persistent

suitable for interactive workloads

shared among many users

expensive to run

Job Cluster

==========

created by the jobs

terminated at the end of the job

suitable for automated workloads

isolated for a job

cheaper in terms of cost

cluster modes

==========

1. single node - the driver will be acting like worker also

suitable for single user

2. standard - we can have multiple workers

suitable for single user

3. high concurency

Both the single node and standard clusters are suitable for single user

1. fault in one user job might impact other user job

2. long running jobs might hamper interactive jobs..

single node and standard - sql, python, scala , R

high concurrency - scala is not supported.

spark offering is with azure synapse

databricks benefits

1. it is more optimized

2. databricks we will find latest versions

3. in synapse you have support for .net

memory optimized - machine learning workloads

compute optimized - streaming workloads

storage optimized - high disk through put

general purpose -

gpu accelerated - deep learning

Azure Databricks - session 3

==========

Create a notebook and write some sample spark code

%md (for documentation)

% is a magic command

the role of magic commands is to write various codes in same notebook

scala

python

sql

R

auxillary magic commands

%fs

Azure Databricks - session 4

==========

DBFS (databricks file system)

dbfs is wrapper on the

azure blob, adls gen2

architecture of databricks

==========

control plane - in databricks subscription

databrics UX, cluster Manager, DBFS, Cluster metadata

data plane - in your azure subscription

vnet, nsg, azure blob storage

Session 5

==========

how to create a free Azure account

azure.microsoft.com

some services are free for 12 months

+

some other 40 services are free forever

+

we get $200 free credits which we can utilize for 30 days

portal.azure.com

Subscription - billing unit

every service is treated like a resource..

each resource has to be in a resource group..

subscription -> Resource groups ->

Subscription

pay as you go/Free

Resource group

Customer 360/VM-rg

Resources

vm, storage...

customer360-stg-rg

customer360-prd-rg

subscription

resource group

resource

Resource Manager

api/sdk portal powershell CLI/Bash

Json Template

Resource Manager

Additional Information

1. each resource can be a part of only one resource group.

2. resource groups cant be nested..

3. resources can be moved between the resource groups..

4. resource groups have their own location and that means the metadata for the resource group

is in that location.

5. resources in the resource groups can reside in different locations.

Azure Databricks - session 7

==========

DBFS - Databricks file system

object store - blob, datalake gen2

databricks - dbfs

DBFS

==========

Object Store

DBFS is a distributed file system mounted into a databricks workspace.

It is basically an abstraction on top of your scalable object storage.

By default you do not see an option to browse your DBFS file system..

How to get the DBFS browse facility on the browser

DButils

========

provides the utility functions

dbutils.help()

dbutils.fs.help()

dbutils.fs.help('cp')

File system utility

==========

cp, head, rm, mkdirs, ls, mv

dbutils.fs.ls('/FileStore')

dbutils.fs.ls('dbfs:/FileStore')

Azure Databricks - session 8

==========

DBFS

dbutils

file system utility

data utility

==========

summarize(df)

notebook utility

==========

exit

run

5 notebooks

wrapper notebook

Azure Databricks - session 9

==========

File system utility

data utility

notebook utility

widgets utility

==========

combobox, dropdown, multiselect, text

dbutils.widgets.combobox(name = 'orderstatus', defaultValue = 'CLOSED', choices =

['CLOSED','COMPLETE','PROCESSING'], label = 'ORDER STATUS')

combobox - you can either select a value from the existing ones or you can type in your value

dropdown - you can only select from existing values

multiselect - you can select more than one value

text - free flowing text

get() - is to get the value of the widget

remove('<name of the widget>')

removeAll() - to remove all the widgets

Azure Databricks - session 10

==========

how to pass parameters to a notebook from another notebook

wrapper notebook which will execute the child notebook and this wrapper notebook also send some

parameters to the child notebook

Mount Point

==========

Blob storage -

storage account name - ttstorageaccount100

container name - inputdatasets

DBFS

/mnt/retaildb

to mount

dbutils.fs.mount(source, mountpoint, extraconfigs)

dbutils.fs.mount(

source = 'wasbs://inputdatasets@ttstorageaccount100.blob.core.windows.net',

mount_point = '/mnt/retaildb',

extra_configs={'fs.azure.account.key.ttstorageaccount100.blob.core.windows.net':'SuTUyxyYr/ooc0gF

Abqwq3cRt5ApE3sAbCMBSLL8trA/jBt0BEiPRYliwgCCPDSBNY4zvC1eHLBj+AStdIk73A=='}

)

1. Account Key

2. SAS key - shared access signature

3. Service Principle

dbutils.fs.unmount('/mnt/retaildb')

dbutils.fs.mounts() - it tells all the available mount points along with the source storage path

1. how to pass the parameter from one notebook to another

2. how to create a mount point

Azure Databricks - session 11

==========

databricks workspace

adls gen2 storage account

Databricks CLI

step 1 - Download and install python 3

step 2 - make sure pip is installed

curl https://bootstrap.pypa.io/get-pip.py -o get-pip.py

python3 get-pip.py

step 3 - pip install databricks-cli

databricks --help

databricks configure --help

databricks configure --token

we need to provide the databricks workspace url

url: https://adb-7847222293192977.17.azuredatabricks.net/

token: dapi0fc8d093461ad9e89c35e4290a87a16b

upload the customers.csv file in raw folder inside FileStore

databricks fs cat dbfs:/FileStore/raw/customers.csv

cat .databrickscfg