what is databricks and why databricks?

Databricks is a company which was founded by the creators of apache spark.

behind the scenes databricks uses apache spark to process massive data efficiently.

challenges with the traditional data platforms

- multiple tools (ingestion, ETL, orchestration, governance)
it is very hard to integrate so many tools.

- complex infrastructure - need to manage clusters manually, very difficult to scale and maintain.

- slow and inefficient pipelines - your traditional etl pipelines are slow

- Lack of collaboration - teams end up working in isolation. Sharing the notebooks/code is painfull.

- unreliable data lakes -

Data lake -> holds your huge data (Data Science/ ML team)
Data Warehouse -> performance / reliability (BI / reporting teams)

Data Lake -> Data Warehouse

there are 2 systems that need to be maintained

extra ETL activity required

same copy at different locations

No ACID transactions

Schema mismatches can break the pipeline.

No versioning

Audit history

- Vendor Locking

ingestion - Lakeflow connect

ETL - Lakeflow Declarative pipelines (DLT)

orchestration - Lakeflow Jobs

governance - Unity Catalog

Data Lake house - Delta Lake

Delta Lake = Data Lake + Data Warehouse
Data Lake - Amazon S3/ ADLS gen2 / Google Cloud Storage

ingestion - Lakeflow connect

ETL - Lakeflow Declarative pipelines (DLT)

orchestration - Lakeflow Jobs

Data Engineer / Data Scientist / Data Analyst
Unity Catalog - Governance / Metadata
Delta Lake
azure/aws/gcp - Datalake

while learning how we will practise

community edition - it has very limited options.

Unity catalog

Databricks has replaced community edition with a free edition.

this free edition is really better than the community edition.

azure databricks

outlook account

30 days free trial (azure account) $200 (17000 INR)

they upgrade to a paid version

$200 (for 30 days..)

if you have to create a azure databricks account

azure free trial account valid for 30 days.

databricks free edition account

to create a azure free account

free edition

azure databricks

portal.azure.com

login into azure portal

to use databricks we need to create a databricks workspace

step 1 - resource group in azure

<project>-<env>-<region>-rg

databricks-training-centralindia-rg

once the resource group is created.

search for databricks

workspace name - <project>-<env>-<region>-dbws

training-centralindia-dbws

when you use databricks then how is your cost determined

3 node cluster

the cost for these 3 vms + service cost

infra cost (azure)

software cost (Databricks) DBU

30 day free trial for azure (Infra free for 30 days)

Databricks cost

when we create a cluster in databricks in backend it will run VM's on azure.

so in which resource group these backend resources should reside

I want to run some spark code

I should be writing the code in a notebook

dbfs (databricks file system)

==========

High level architecture of Databricks
==========

https://learn.microsoft.com/en-us/azure/databricks/getting-started/overview

control plane (databricks cloud account)
==========
- Metadata and ACL
configurations (cluster config, notebook config, job info)

web url through which workspace is accessible - https://adb-1578161282373452.12.azuredatabricks.net/?o=1578161282373452&l=en

workspace id - 1578161282373452

data plane / compute plane (customer cloud account)
==========
client data
classic compute cluster

This cluster is a classic compute cluster.

Client cloud account.

==========

A organization can have a single databricks account.

and will most likely have multiple databricks workspaces.

whats the need for more than one workspace?

for different environments (dev, stg, prod)
for different BU's (sales, finance, IT)

Tables
=======

A table has 2 parts

Data + Metadata

Data will be stored in a files (in your own cloud account)
Metadata - this will be kept in a Metastore

for example - Hive Metastore (Legacy)

Each workspace will get its own hive metastore (legacy feature)

It has a 2 level namespace

schema/database

Schema.tablename

retaildb - customer, orders

retaildb.orders

what are the issues with hive metastore

- Hive metastore offers basic metadata management but lacks, fine grained access controls.

it lacks build in data governance features like
- data lineage
- auditing
- access controls

these limitations makes it challenging to manage security and compliance.

Moreover hive metastore can deal only with structured data.

tables in the hive metastore do not benefit from the full set of security and governance features provided by the unity catalog. such as built in auditing, lineage and access control.

Now in the latest versions of databricks, the workspace it is unity catalog enabled.

Now whenever we create a databricks workspace we will get

an automatically provisined unity catalog metastore..

a catalog named after your workspace.

Metastore

Catalogs

Schemas / Databases

Tables

to go to the accounts console -

https://accounts.azuredatabricks.net/

the account admin

for each region there will be just one metastore

centralindia

your first workspace in central india

one metastore in central india will be created.

later if you create 4 more workspaces in centralindia

all of these 4 workspaces will be attached with this metastore.

if you create a new workspace in another region us-east

then one more metastore will be created

why this metastore - it will store metadata + security related configs

can one workspace have more than one metastore?

No it cannot have.

can one metastore be associated with multiple workspaces?

It can be if its in the same region.

The best practise is to just have 1 metastore in one region and then attact this metastore to all the workspaces in that region.

Metastore

Catalogs (will be same as your workspace name)

Schemas / Databases

Tables

whenever you create a new workspace

it will give a new catalog with the name of workspace

in this default catalog (default, information_schema)

you cannot shared the managed resource group among different databricks workspaces.

Unity catalog (Metadata / Governance)

Metastore

catalog

schema

table, view, volumes, functions, model

3 level namespace

=========

Lets see how to create a table

There are 3 types of tables

- Managed table
- external table
- foreign table

Managed table (data + metadata is managed by databricks)

the metadata will be stored in the metastore

where will your data be stored

I am creating a table and then inserting records

where the data will be stored.

Metastore - one metastore for a workspace

catalogs - location for default catalog

schemas / databases

tables

whenever you create a workspace - but there is no storage location associated with the metastore.

storage account name - dbstorage2af6o74djoavs
container name - unity-catalog-storage

abfss://unity-catalog-storage@dbstorage2af6o74djoavs.dfs.core.windows.net/15781616282373452/__unitystorage/catalogs/2408b6f9-bbea-4b3a-8e0a-3b8b22712862

storage account name - databricksdemo101sa

container name - unitycatalog

folder - catalog

I want databricks to access this storage

go to azure portal and say "access connector for azure databricks"

after your access connector is created.

we need to give access to the storage account.

our connector has the access to the storage account.

I have to create a external location

storage account name - databricksdemo101sa
container name - unitycatalog

folder - catalog

external location - its a cloud path
storage credentials

to enable fine grained access controls

cloud location
storage credential - mycredential1

myexternallocation1

========

metastore - location

catalog - location

schema - location

table

training-centralus-dbws

abfss://unitycatalog@databricksdemo101sa.dfs.core.windows.net/metastore

storage cred: adls_raw_storage_cred

we just defined a location at metastore level.

we want to also define a location at catalog level.

how to create a external location using code

we will use the existing credentials

CREATE EXTERNAL LOCATION adls_raw_storage_catalog_ext_loc URL

'abfss://unitycatalog@databricksdemo101sa.dfs.core.windows.net/catalog'
WITH (CREDENTIAL adls_raw_storage_cred)
COMMENT 'external location catalog level for raw sales data in adls'

CREATE EXTERNAL LOCATION adls_raw_storage_schema_ext_loc URL
'abfss://unitycatalog@databricksdemo101sa.dfs.core.windows.net/schema'
WITH (CREDENTIAL adls_raw_storage_cred)
COMMENT 'external location schema level for raw sales data in adls'

Managed table vs External table
==========

```
CREATE TABLE sales (
    sale_id INT,
    customer_name STRING,
    product STRING,
    quantity INT,
    amount DOUBLE
);
```

the managed table by default will be a delta table

```
CREATE TABLE sales (
    sale_id INT,
    customer_name STRING,
    product STRING,
    quantity INT,
    amount DOUBLE
)
location '<path>';
```

in old days when hive metastore was the default one..

if you drop a managed table (both your data and metadata get dropped)

with external tables (only metadata was dropped and data files were not touched)

unity catalog

managed table when we drop them (the metadata is lost, and data has a retention of 7 days)

till 7 days you can use the UNDROP TO undrop the table

external table (only metadata is dropped data is not touched)

Managed tables
==========

Fully governed and optimized by the unity catalog

Databricks recommends using managed tables to take advantage of:

- Reduced storage and compute costs
- faster query performance
- automatic table maintainance and optimization
- secure access for non databricks clients via open API's
- Supports Delta Lake and Iceberg formats
- Automatic upgrades to the latest platform features.

why optimized - because managed tables learn from your read and write patterns.

Unity catalog supports the UNDROP TABLE command to recover dropped managed tables for 7 days.

EXTERNAL TABLES
==========

```
CREATE TABLE sales (
    sale_id INT,
    customer_name STRING,
    product STRING,
    quantity INT,
    amount DOUBLE
)
location '<path>';
```

unity catalog governs data access permissions but does not manage data lifecycle, optimizations, storage location or layout

when you drop an external table, the data files are not deleted.

when you create an external table, you can either register an existing directory of data files as a table or
provide a path to create a new directory.

Databricks recommends using external tables for the following use cases:

- you need to register a table backed by existing data that is not compatible with unity catalog manage tables.

- you also require direct access to the data from non databricks clients.

unity catalog privileges are not enforced when users access data files from external systems.

DELTA
CSV
JSON
AVRO
PARQUET
ORC
TEXT

```
CREATE EXTERNAL LOCATION adls_storage_external_data_loc URL
'abfss://unitycatalog@databricksdemo101sa.dfs.core.windows.net/external'
WITH (CREDENTIAL adls_raw_storage_cred)
COMMENT 'external location for handling external data in adls'
```

```
CREATE TABLE sales (
    sale_id INT,
    customer_name STRING,
    product STRING,
    quantity INT,
    amount DOUBLE
)
location
'abfss://unitycatalog@databricksdemo101sa.dfs.core.windows.net/external';
```

```
INSERT INTO sales VALUES
```

```
(1, 'Alice', 'Laptop', 1, 1200.00),
(2, 'Bob', 'Tablet', 2, 800.00),
(3, 'Charlie', 'Phone', 1, 500.00);


CREATE EXTERNAL LOCATION adls_storage_external_sales_data_loc URL
'abfss://unitycatalog@databricksdemo101sa.dfs.core.windows.net/external_1/sal
es'
WITH (CREDENTIAL adls_raw_storage_cred)
COMMENT 'external location for handling external sales data in adls'


%sql
CREATE TABLE sales_ext (
    sale_id INT,
    customer_name STRING,
    product STRING,
    quantity INT,
    amount DOUBLE
)
USING CSV
location
'abfss://unitycatalog@databricksdemo101sa.dfs.core.windows.net/external_1/sal
es';
```