

Name: Rohan Vinayak Chaudhari

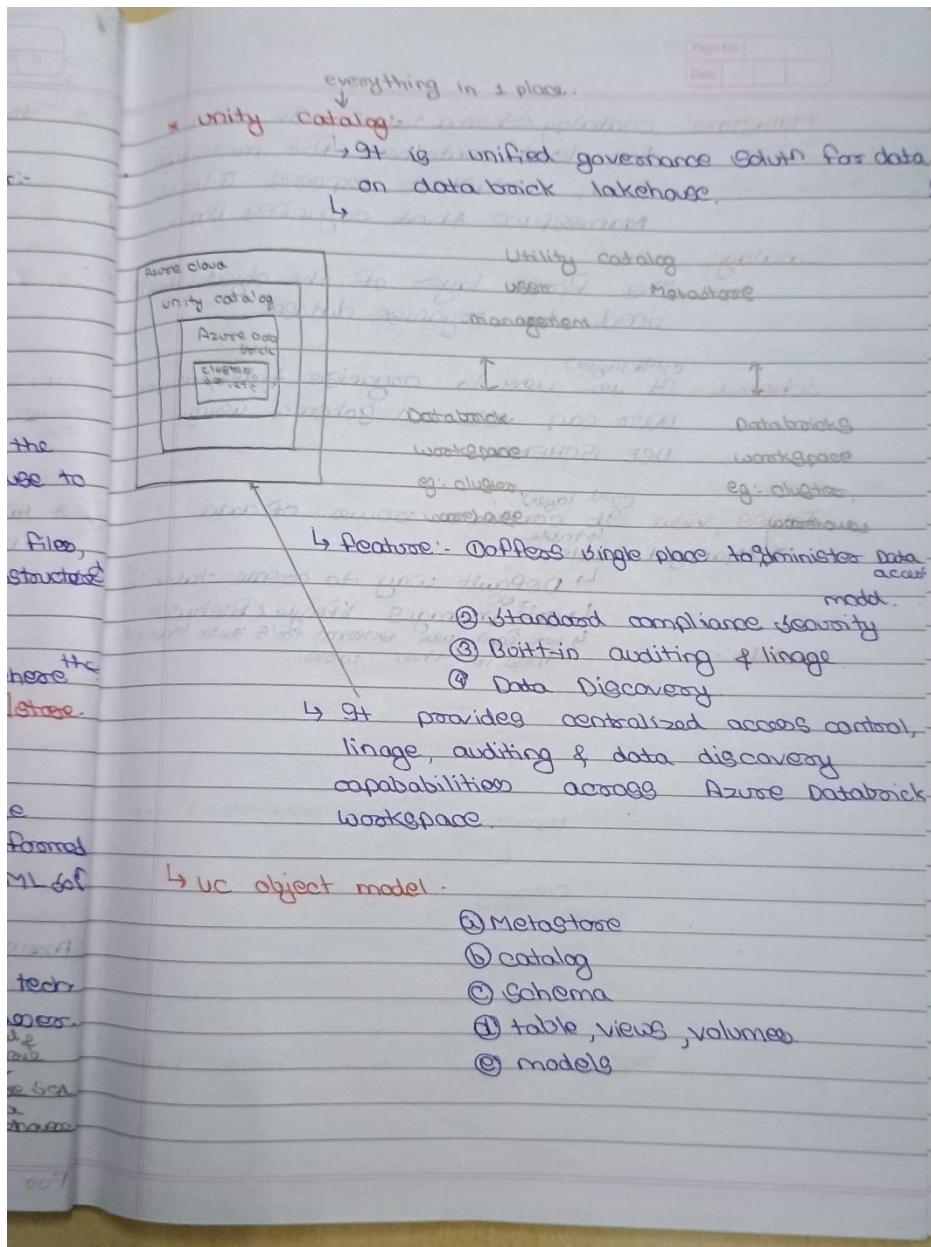
Batch: Data Engineering

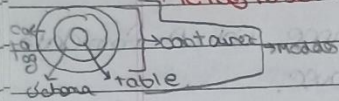
Date: 19/02/2024

Topic: Azure DataBricks

Solution:

1. Azure Databricks:





Metastore:- catalog schema table:- It is a top level container for metadata. Each metastore exposes 3 level namespace that organises data.

catalog:-

The first layer of the object hierarchy used to organise dataset.

schema:- ^{2nd layer} It is used to organise table, view, user can see all schema using `USE SCHEMA`.

Tables, view:- ^{3rd layer} It contains rows of data.

* **Managed table**

↳ Default way to create table in UC

↳ ~~UC~~ manages lifecycle & file by UC

↳ we can't use external tools to manipulate files in this tables.

external table

Metastore

can connect with

Azure Storage Acct, ADLS gen

Catalog

organise

schema

Tables

Views

Volume

Models

Functions

File layout - Delta, CSV, JSON, Parquet, ORC, text

External table:-

- ↳ They are the table whose data lifecycle & file layouts are not managed by UC.
- ↳ It is used for large amount of existing data.
- ↳ When you drop external table UC does not delete the underlying data.

* Job Scheduling:-

- ↳ It is used when we want to run our job (notebook) at a specific time.