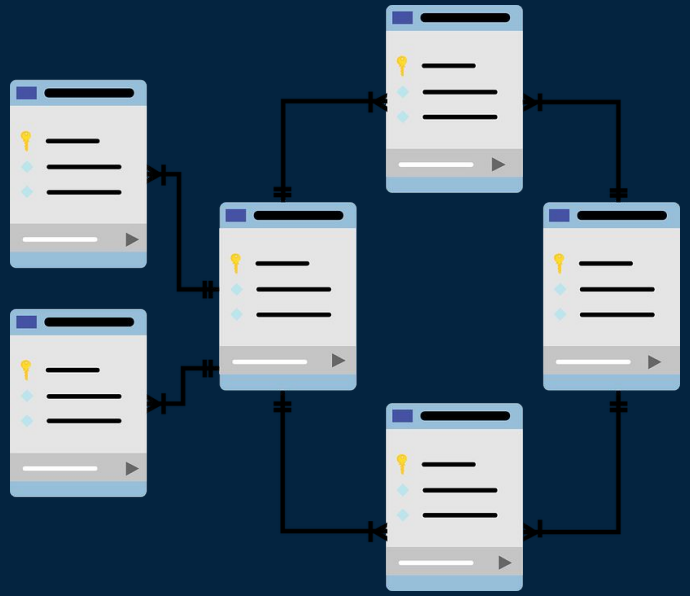
What are Relational Entities in Databricks?

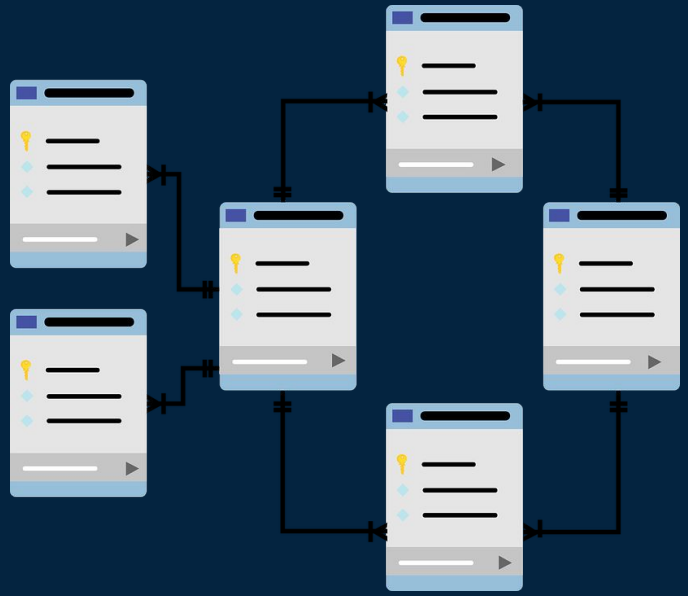


Learning Objectives

- Describe relational entities in Databricks
- Explain the workspace - storage layer relationship
- Introduce the LOCATION argument

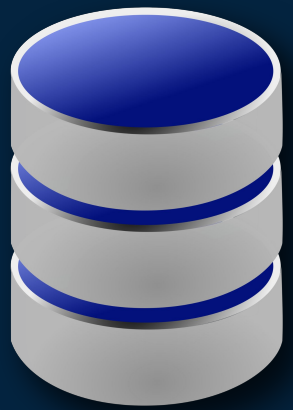
Relational Entities in Databricks

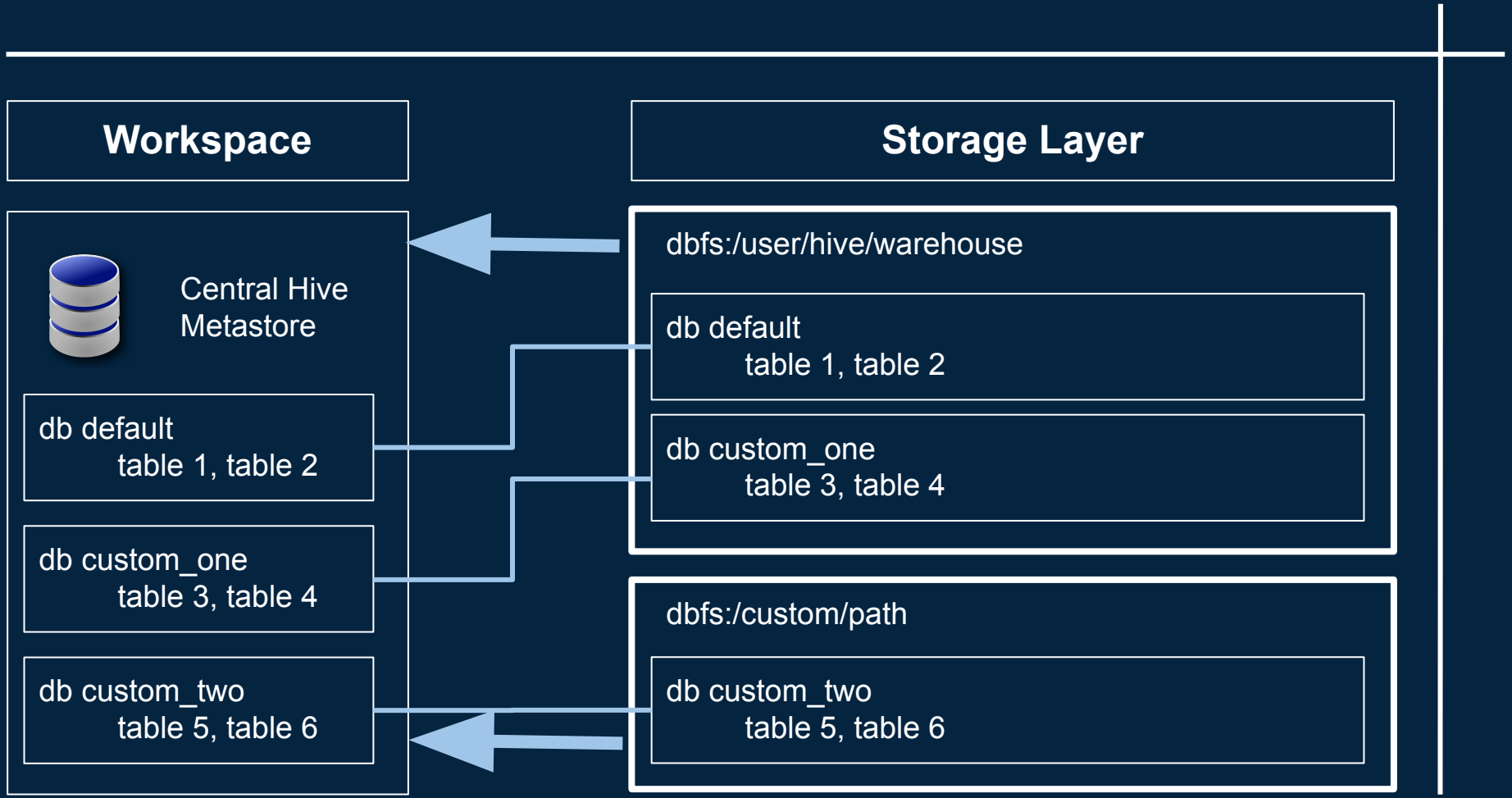
- Databases
- Tables
- Views
- CTEs



Databases

- Databases in Databricks are essentially schemas in Hive metastore
- They are essentially repositories of metadata regarding other relational entities, such as tables
- CREATE DATABASE db_name
- CREATE SCHEMA db_name




Workspace
Storage Layer
Central Hive Metastore
db default
table 1, table 2
db custom_one
table 3, table 4
db custom_two
table 5, table 6
dbfs:/user/hive/warehouse
db default
table 1, table 2
db custom_one
table 3, table 4
dbfs:/custom/path
db custom_two
table 5, table 6

The LOCATION argument

- All metadata are always stored in the central Hive metastore
- By default the actual data get stored at the path dbfs:/user/hive/warehouse
- With the LOCATION keyword it is possible to specify a storage location different than dfbs:/user/hive/warehouse
- CREATE SCHEMA db_custom_two
  LOCATION ‘dbfs:/custom/path/db_custom_two.db‘

Tables

Managed Tables

- Created under the database directory, in the actual storage layer
- Databricks manages both the metadata and the data
- Dropping the table means deleting both metadata and data
- CREATE TABLE table_name

External (Unmanaged) Tables

- Created outside of the database directory
- Databricks manages only the metadata
- Dropping the table means deleting only the metadata
- CREATE TABLE table_name LOCATION dbfs:/custom/path

Views

- Views in Databricks are essentially the logical result of a query. They do not persist data physically.
- (Stored) Views
- Temporary Views
- Global Temporary Views

Views

(Stored) Views

- Traditional Views persisted objects in Databricks Hive Metastore

Temporary Views

- They last exclusively for the duration of a Spark Session
- End when a new notebook is opened, detached or reattached from a cluster, or if a cluster is restarted

Global Temporary Views

- They last until their cluster runs
- They are added to a cluster temp db called global_temp

Common Table Expressions (CTEs)

- A CTE is a logical result of a query.
- It defines a temporary result set that can be referenced multiple times within the scope of a SQL statement.
- A CTE lasts only for the running of the SQL statement that contains it.