Creating Databases and Tables

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Understand managed, external and temporary tables

Learn how to insert data into tables from files and from other tables

Alter and drop tables

Get introduced to partitioning and bucketing

Storing Data in Hive

Storing Data in Hive

Data

The records in the table which holds the actual data

Metadata

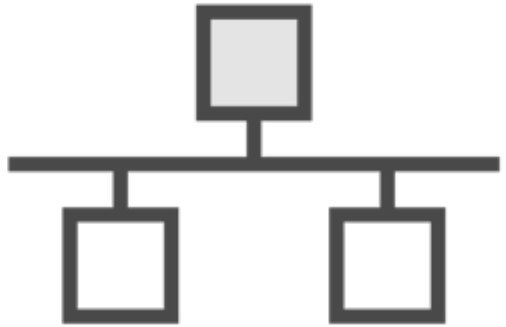
Information about the underlying data in the table

Data

Stored in HDFS, the reliable storage for data in Hadoop

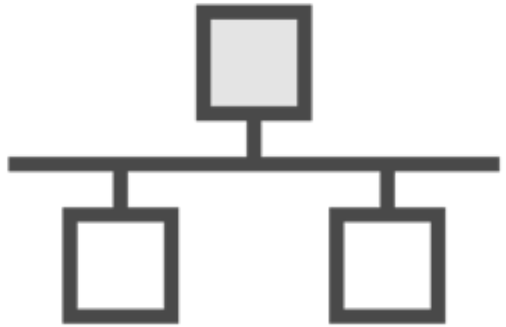
Files partitioned across multiple machines in the cluster

Data



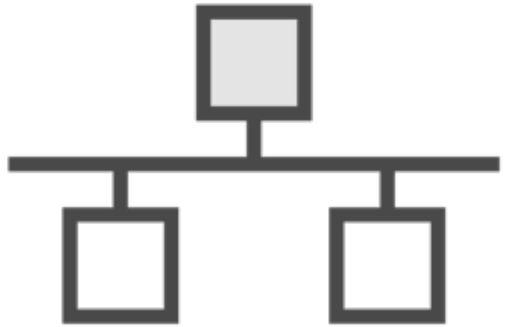
Stored in directories under Hive's warehouse directory

Data



hive.metastore.warehouse.dir
property in hive-site.xml

Data



Defaults to
/user/hive/warehouse

Metadata

Metastore, acts as a bridge between Hive and files in HDFS

A relational database with information on:

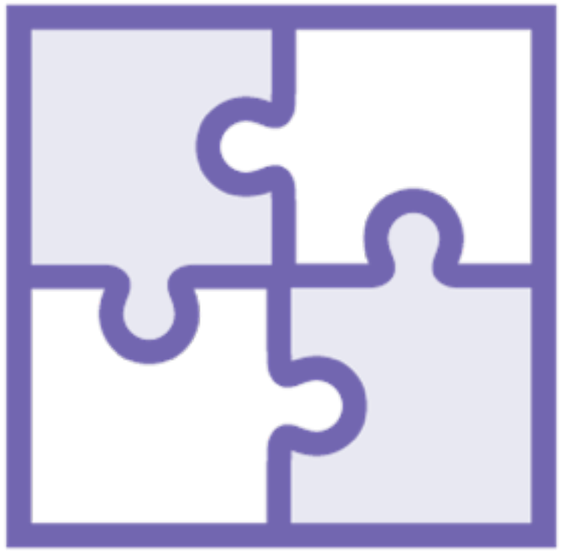
- databases, tables
- columns, owners, storage, serialization/ deserialization information
- user supplied metadata

Demo

Explore the warehouse directory where Hive data is stored

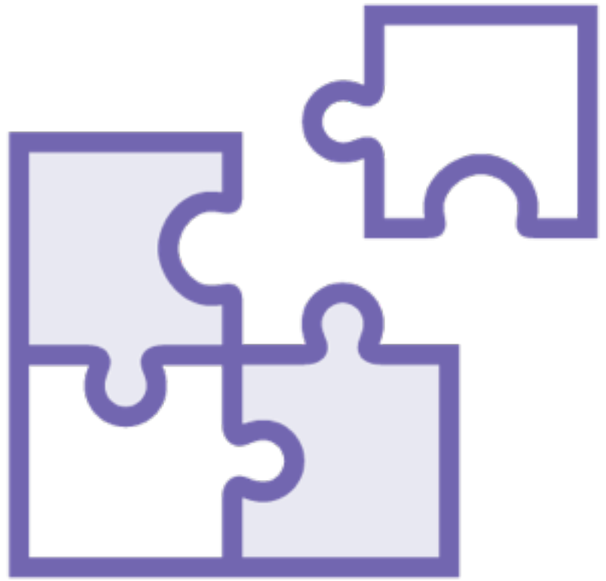
Tables in Hive

Hive Tables



Managed

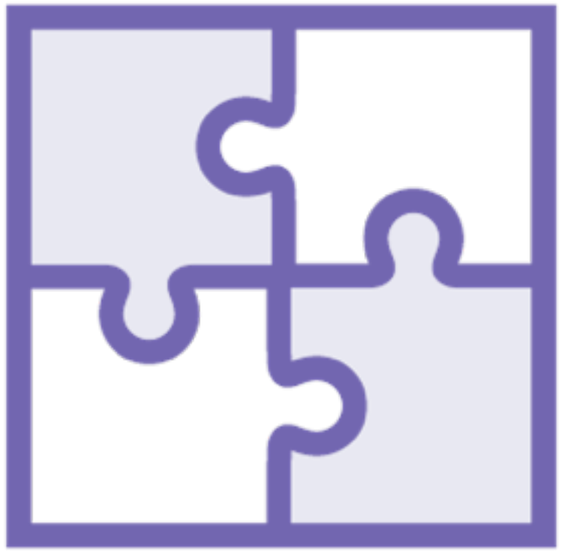
Data managed by Hive and stored in the warehouse directory



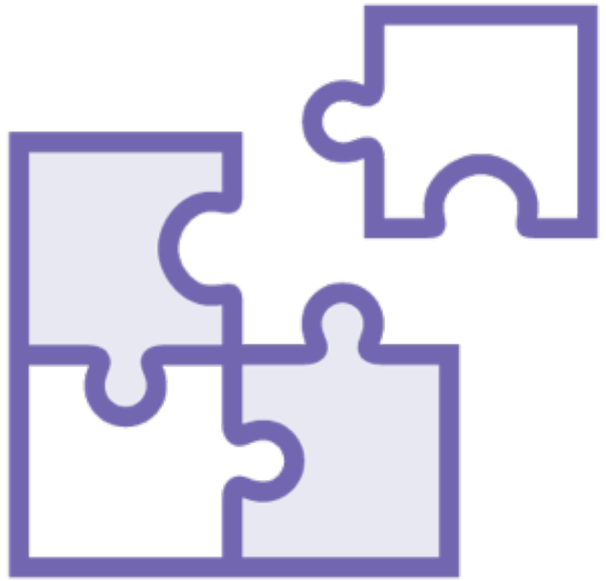
External

Data not fully managed by Hive and exists outside the warehouse directory

Hive Tables



Managed



External

The metadata for both is in the metastore

Managed Tables

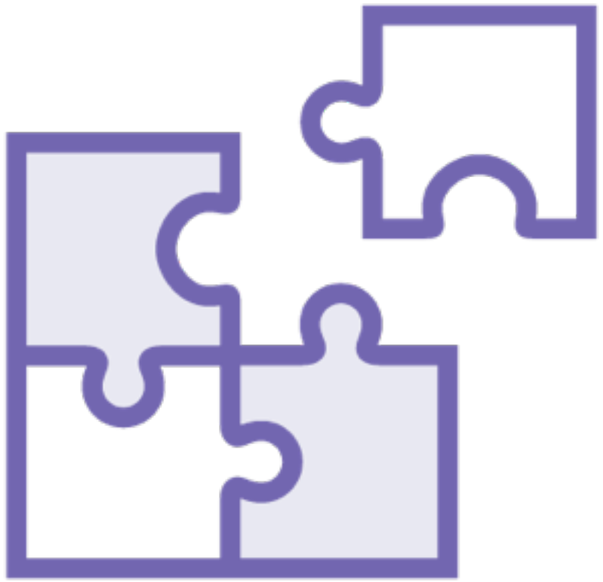
All tables so far have been managed tables

Hive owns the files and directories

These can be modified by other technologies

Deleting a managed table deletes both data and metadata

External Tables



Share the underlying data across other technologies

Hadoop, Pig, HBase all of these may access and edit those files

Deleting an external table deletes only the metadata

Demo

Create an external table in Hive

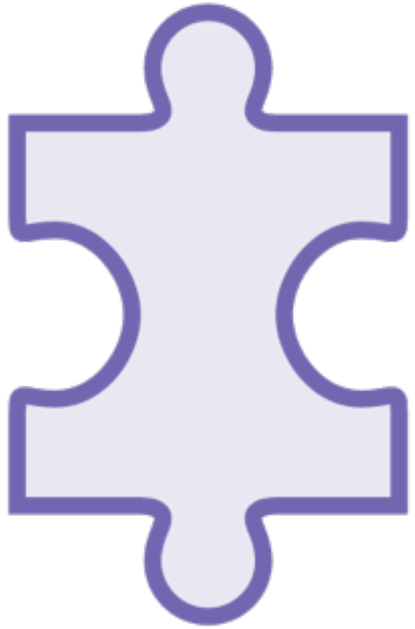
Delete a external table to see how it is different from a managed table

Demo

Explore the options available while creating a table

Temporary Tables

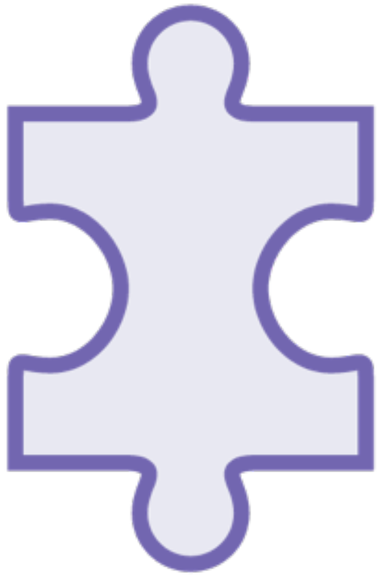
Temporary Tables



Temporary

Tables created within a Hive session, deleted when the session ends

Temporary Tables



Store temporary data

Tables of the same name can be created by different users

Do not support partitions and indexes

Can have the same name as a permanent table

Demo

Temporary tables in Hive

Inserting Data into Hive Tables

Inserting Data

Demo

Load data into tables

- from files

- from other existing tables

Load multiple tables from a single table

Delete data from tables

Deleting and Updating Data in Hive

Deleting and Updating Data

Hive tables do not support row level deletes and updates by default

It is possible to get ACID compliant* Hive tables by setting up special properties in hive-site.xml

Deleting and Updating Data

There is a lot of fine print around exact support for ACID

Not covered in this course

Partitioning and Bucketing of Tables

Partitioning and Bucketing

Partitioning

Bucketing

Splits data into smaller, manageable parts

Partitioning and Bucketing

Partitioning

Bucketing

Enables performance optimizations

Partitioning

Data may be naturally split into logical units

Partitioning

Each of these units will be stored in a different directory

WA

OR

CA

CT

NY

GA

Partitioning

State specific queries will run only on data in one directory

WA

OR

CA

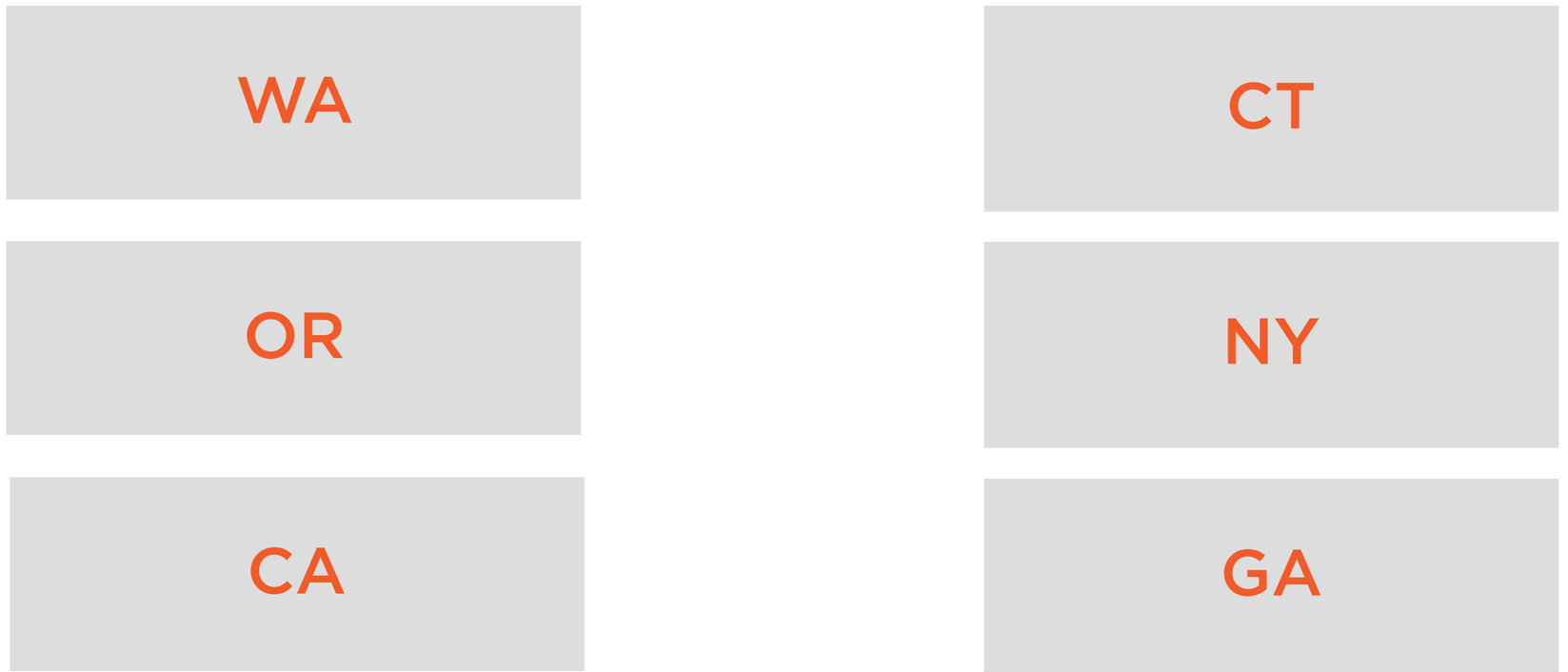
CT

NY

GA

Partitioning

Splits may not of the same size



Partitioning and Bucketing

Partitioning

Bucketing

Bucketing

Size of each split should be the same

Bucketing

Hash of a column value - address, name, timestamp anything

Bucketing

Each bucket is a separate file

Bucketing

Makes sampling and joining data more efficient

Partitioning and Bucketing

Not covered in this course

Summary

Understood managed, external and temporary tables

Learnt how to insert data into tables from files and from other tables

Explored the alter and drop table commands

Got an overview of partitioning and bucketing of Hive tables