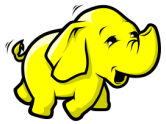


Apache Hive

Sunbeam Infotech

Hive scripts

- Hive scripts contains multiple Hive QL statements to executed sequentially (like .sql script).
- It includes Hive QL queries as well as configurations.
- Running using Hive CLI
  - terminal> hive -f /path/to/.hql
  - hive> SOURCE /path/to/.hql
- Running using Hive beeline
  - terminal> beeline -u ... -n ... -f /path/to/.hql
  - beeline> !run /path/to/.hql

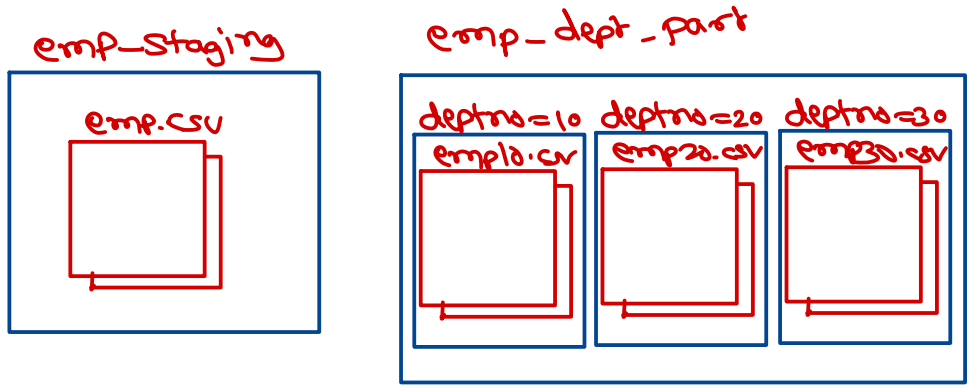
Managed Table vs External Tables

- CREATE TABLE statement.
- Located in HDFS warehouse directory
- Drop table operation drop table data (from HDFS) as well as table structure (from metastore).
- Loading data explicitly into the table (HDFS) after table creation.

- CREATE EXTERNAL TABLE statement.
- Located in HDFS directory.
- Drop table operation drop only table structure (from metastore). Data in HDFS is intact.
- When data is already present in HDFS and need to process it using HiveQL. Multiple tables (metadata) can refer to the same data files in HDFS.

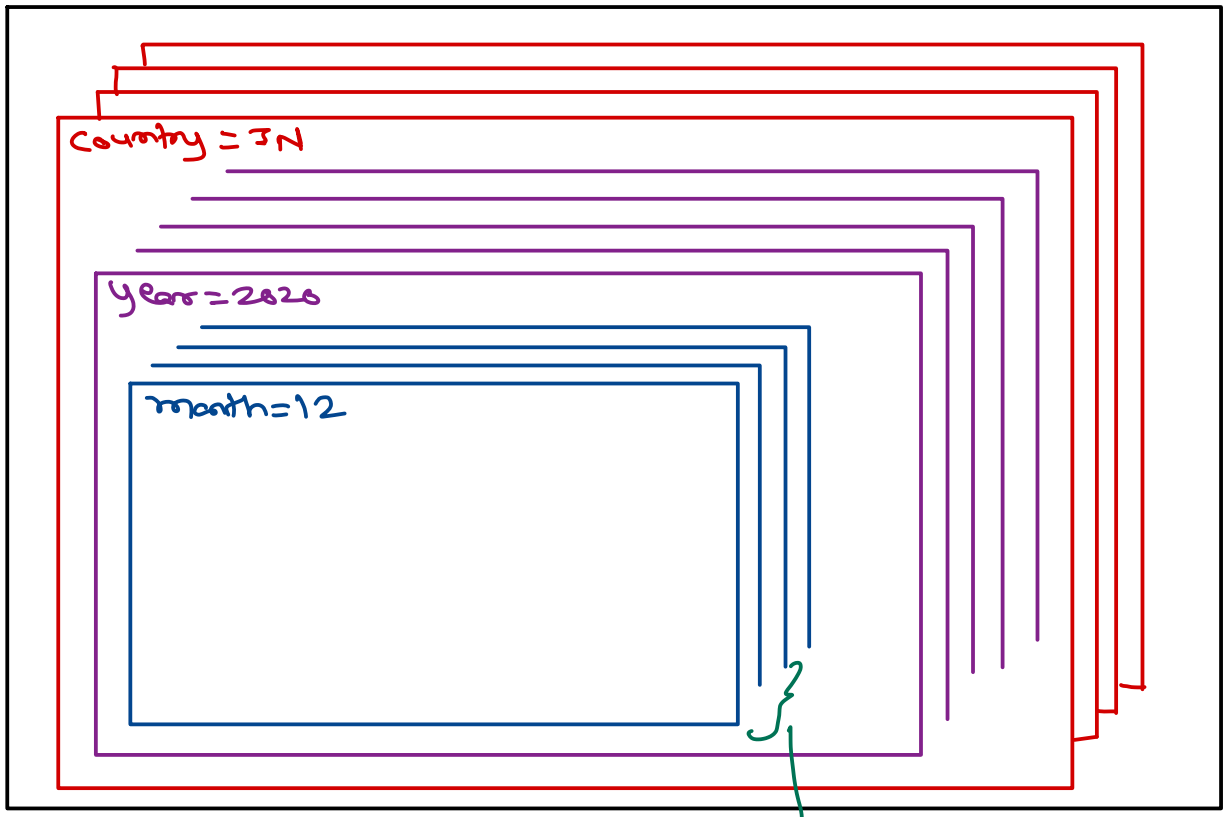
Partitioning

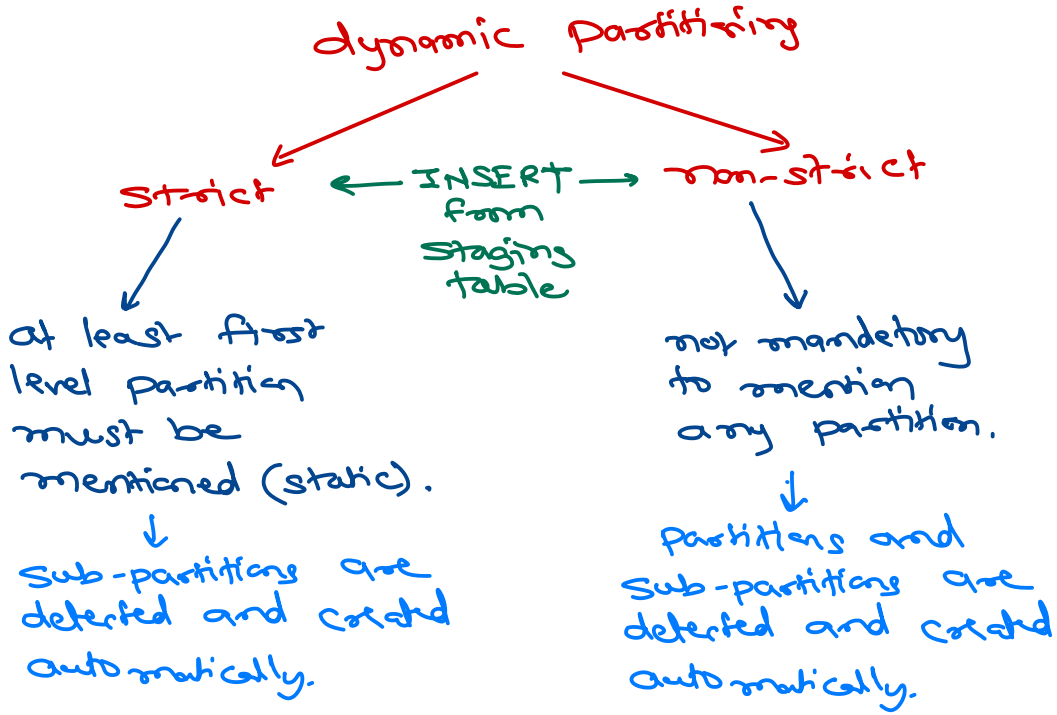
- Data is divided into multiple sub-directories under HDFS (table location) based on value of one or more columns.
- When query is fired for given value of column, only respective sub-directories data will be processed. This significantly improves performance.
- Examples:
  - Emp partitioned dept-wise
  - Emp partitioned job-wise
  - Emp partitioned dept-job-wise
- Static partitioning
  - Data is ingested partition-wise.
  - Very fast operation.
- Dynamic partitioning
  - Data is ingested in staging table.
  - Data is loaded partition-wise into main table using MR.



Partitioning

select ... from orders
where country='IN' and year=2020 and month in (10,11,12)





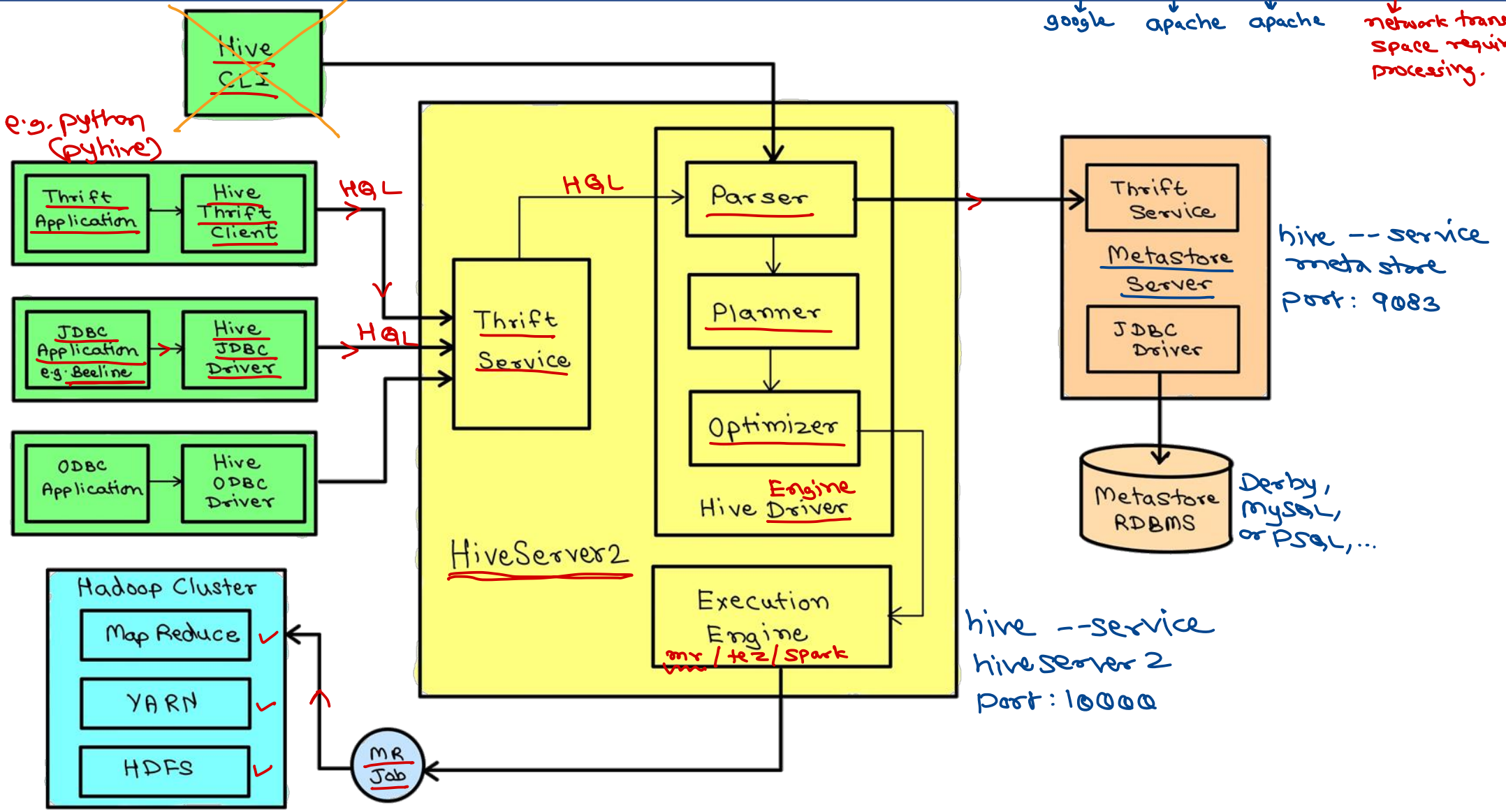
Hive functions

- Hive have many built-in functions.
  - Single Row Functions
    - Row → Function → Row
    - e.g. LENGTH(), CONCAT(), ROUND(), ...
  - Group/Aggregate Functions
    - Rows → Function → Row
    - e.g. SUM(), AVG(), COUNT(), ...
  - Table generation Functions
    - Row → Function → Rows
    - e.g. EXPLODE(), ...
- Hive function help is available in Hive documentation:
  - https://cwiki.apache.org/confluence/display/Hive/LanguageManual+UDF
- Hive UDF can be written in Java/Python.

Hive Architecture

Hive = metadata + data + processing
metadata: table struct, RDBMS
data: HDFS, Warehouse
processing: HDFS api & Map Reduce

IPC over network/sockets
- Text (plain / xml/json) - Compatible but slow process
- Binary (protobuf / avro / thrift) - efficient execution
  - protobuf: google
  - avro: apache
  - thrift: apache
  - efficient execution: network transfer, space required, processing.



Thank you!

Nilesh Ghule <nilesh@sunbeaminfo.com>