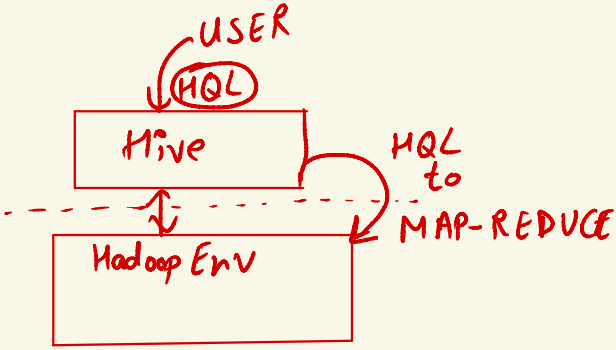
Challenges with Hadoop framework

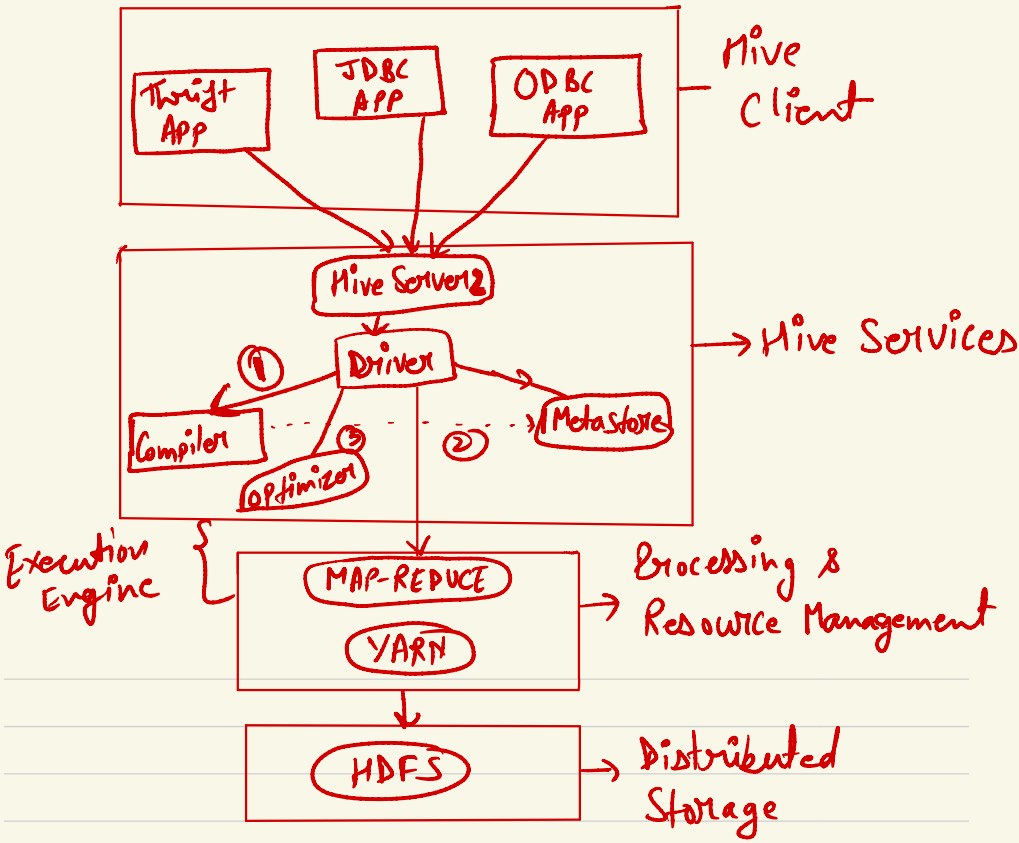
- We need to code a lot
- Difficult to adapt by Non-Tech or Analyst folks
- Difficult process for adhoc queries

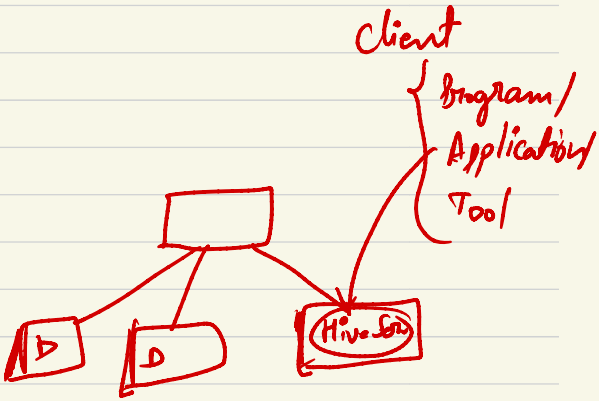


Hive

- It is a data warehousing service
- Build on top of Hadoop
- It uses HDFS file system
- It is SQL like framework and has its own query language known as HQL
- It converts HQL query into MAP-REDUCE code.
- It is used for heavy Analytical Queries
- Ad-hoc Queries

Hive Architecture





① Thrift Clients

The Hive server is based on Apache Thrift and that is why it can serve the requests from Thrift Clients.

② JDBC Clients

Hive allows for the Java applications to connect to it using JDBC driver.

③ ODBC Clients Hive allows ODBC based applications (open database Connectivity) to connect to Hive.

④ Hive Server2 ↓ It is the successor of Hive Server1. It enables clients to submit queries for execution. Designed to support JDBC, ODBC & Thrift Connection

→ Hive Server1 doesn't handle concurrent requests from more than one client which is resolved in Hive Server2

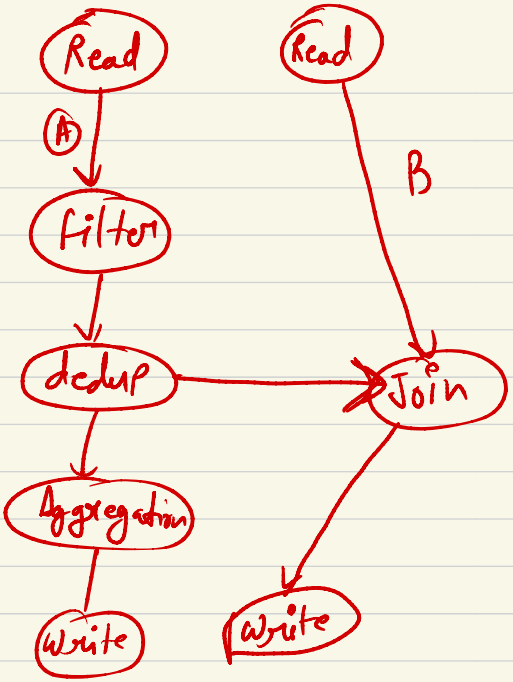
⑤ Hive Driver → It receives HQL statement from users and creates the session handles for the query and sends the query to the compiler.

⑥ Hive Compiler → It parses the Query. It performs type-checking on the different Query

blocks and query expressions by using the metadata stored in the metastore and generates an Execution plan.

The execution plan created by the compiler is known as DAG (Directed Acyclic Graph).

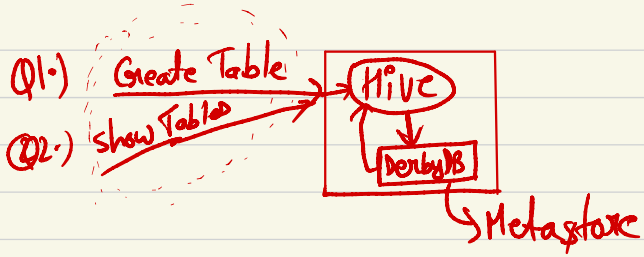
a.) Read employee Data
b.) filter data
c.) Dedup check
d.) Aggregation
e.) write



⑦ Optimizer → It performs the transformation on the execution plan and split the task to improve efficiency.

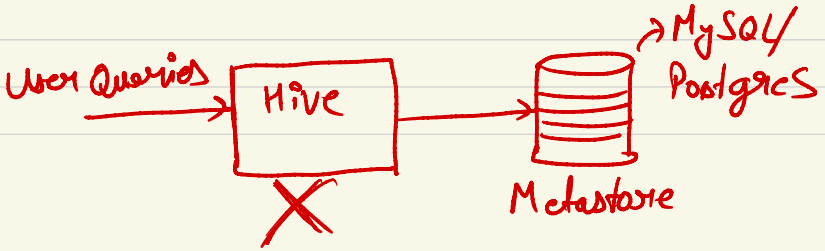
⑧ Metastore → It is a central repository which stores the metadata information about the structure of tables, partitions, including columns and data type.

It also stores the information of serializer & deserializer, required for the read/write operation.



* In case of DerbyDB/Internal DB if Hive is removed then metadata will also be lost.

* It doesn't support concurrent request (DerbyDB)



⑨ Execution Engine → After the compilation & optimization steps, executes the execution plan created by compiler in order of their dependencies using Hadoop Map-Reduce.

---

x) Working of Hive

① executeQuery → The user interface calls the execute interface to the driver.

② getPlan → The driver accepts the query, creates a session handle for the query, and passes the query to the compiler for generating execution plan.

③ getMetaData → The compiler sends the metadata request to the metastore.

④ sendMetaData → The metadata store will send the metadata to the compiler.

Now compiler generates DAG.

⑤ sendPlan → The compiler then sends the generated execution plan to the driver.

⑥ executePlan → Driver sends the execution plan to the execution engine for executing the plan.

⑦ Submit job to Map-Reduce: Based on the query plan map-reduce task will be created to get the final result

⑧ sendResult → Driver will collect results and send it to Client/ Interface.