* Apache hive:

→ The Apache hive data warehouse software facilitates reading, writing and managing large datasets residing in distributed storage and queried using SQL syntax.

→ Built on top of Apache hadoop

→ provides following features:

↳ tools to enable easy access to data via SQL, thus enabling data warehousing tasks such extract / transform / load (ETL), reporting & data analysis.

↳ Allows programmers to plug in custom Mappers & Reducers.

↳ A mechanism to impose structure on a variety of data formats.

↳ Access to files stored either directly in Apache HDFS or in other data storage systems such as Apache HBase.

↳ Query execution via Apache Tez, Apache Spark, or Map Reduce.

→ What is hive?

↳ Hive is a data warehousing infrastructure based on Apache hadoop.

↳ Hadoop provides massive scale out and fault tolerance capabilities for data storage and processing on commodity hardware.

↳ Hive is designed to enable easy data summarization, ad-hoc querying and analysis of large volume of data

↳ It provides SQL which enables users to do ad-hoc querying, summarization and data analysis easily.

↳ At the same time, Hive's SQL gives users multiple places to integrate their own functionality to do custom analysis, such as User Defined functions (UDFs).

→ What is Hive NOT? x

↳ Hive is not designed for online transaction processing. (OLTP)

↳ It is best used for traditional data warehousing task.
Such as, Online Analytic processing (OLAP).

* Apache Hive:

→ Apache hive is a Data warehousing package built on top of hadoop. and it is used for data analysis.

→ Hive is targeted towards users who are comfortable with SQL.

→ It is similar to SQL and called HiveQL.

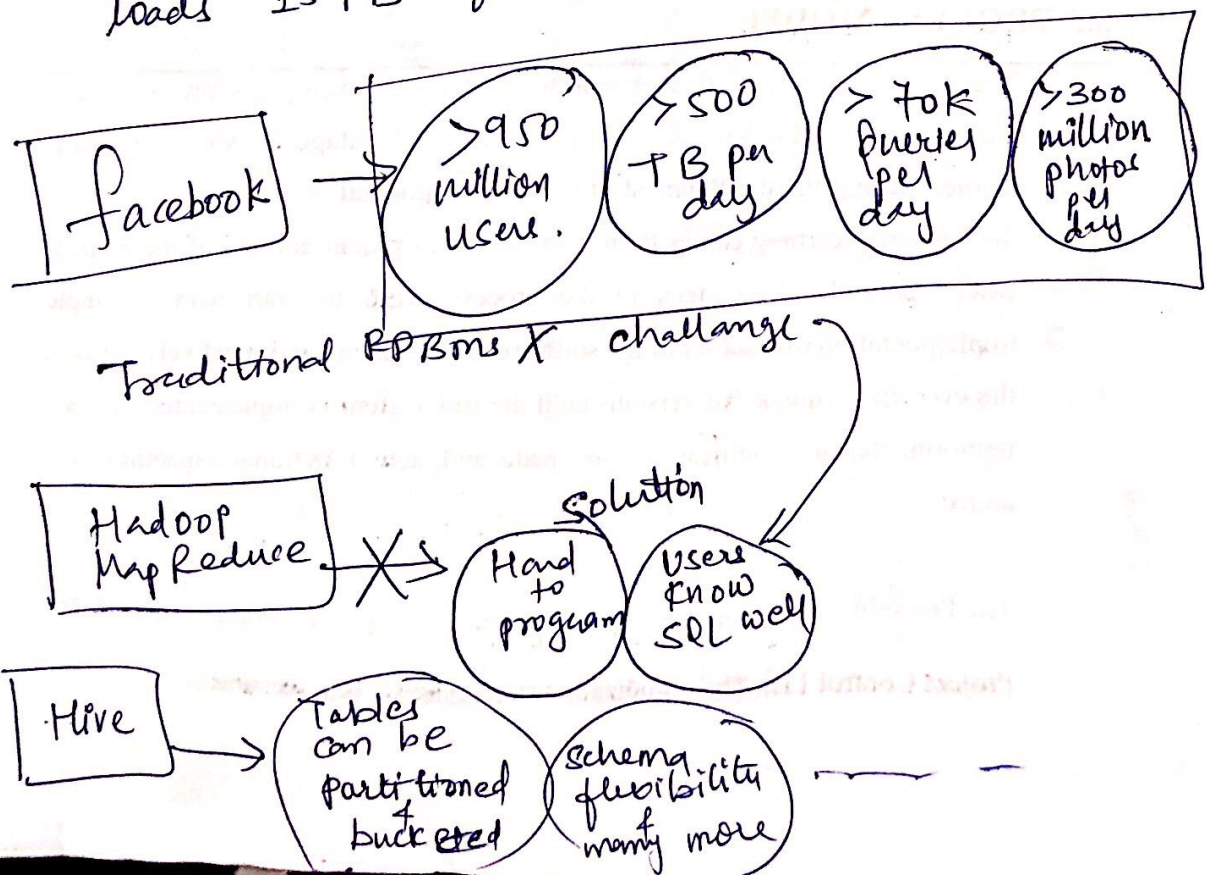
→ Apache hive is used to abstract complexity of hadoop.

→ This ~~can~~ allows traditional Map/Reduce programmers to plug in their custom mappers and reducers.

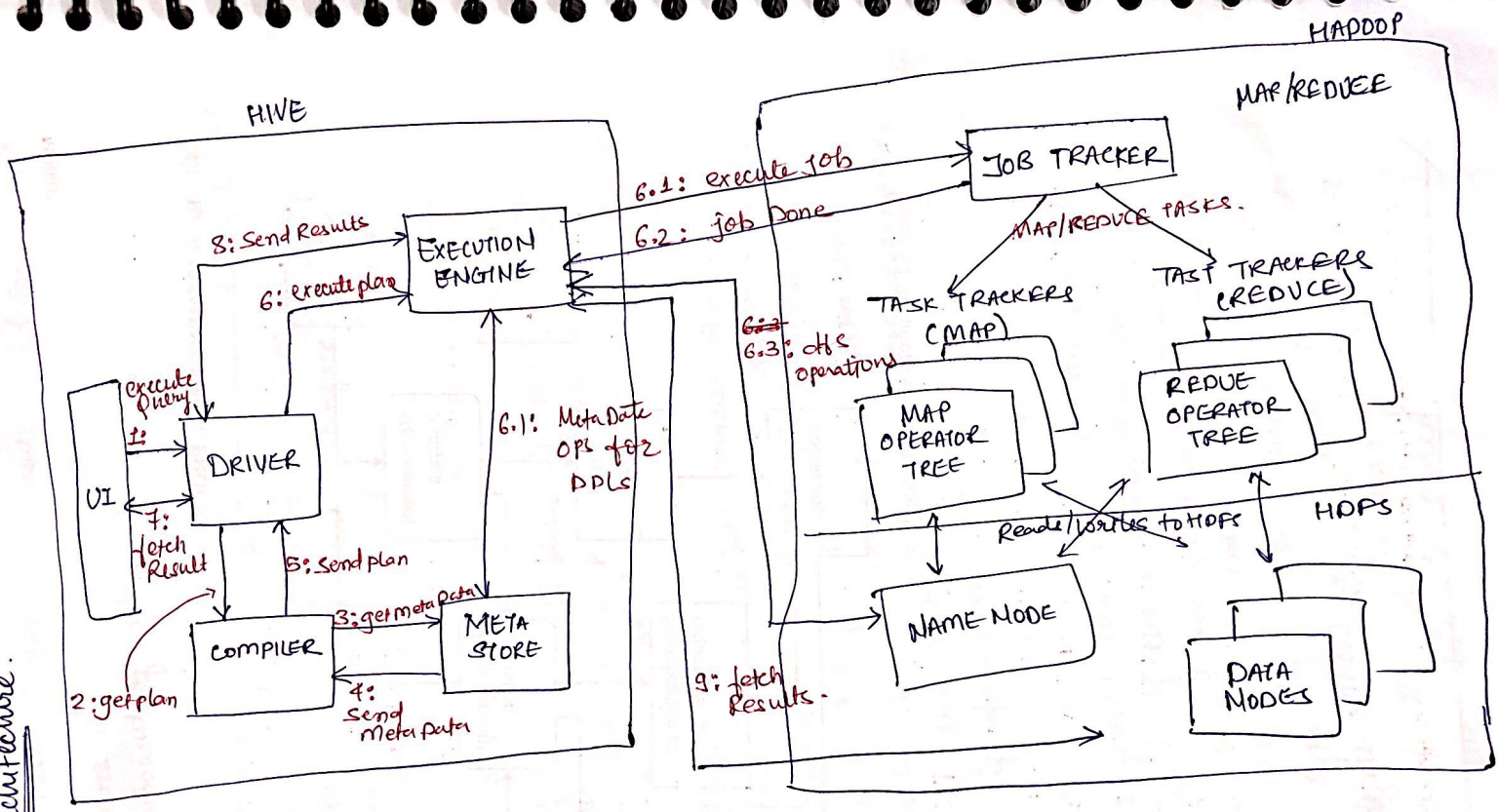
→ There is no need to learn java.

→ Hive, an open source peta-byte scale data warehousing framework based on hadoop, was developed by the Data infrastructure Team at facebook.

→ Hive-hadoop cluster at facebook stores more than 2PB of raw data and regularly loads 15 PB of data on daily basis.



Hive Architecture :



⟹ figure shows major components of Hive & its interactions with Hadoop.

→ Main component of Hive are:

① UI:
- The user interface for users to submit queries & other operations to the system.
- Through command line interface or a web based GUI

② Driver:
- The component which receives the queries.
- This component implements the notion of session handles & provides execute and fetch APIs modeled on JDBC/ODBC interface.

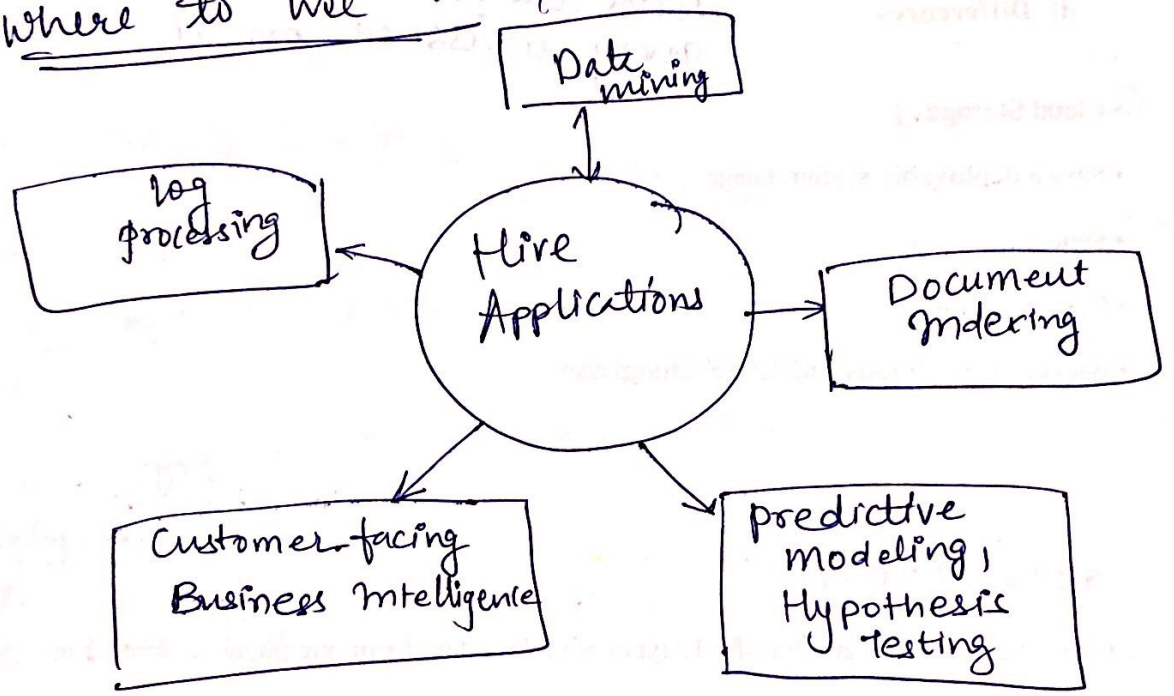
③ Compiler:
- The component that parses query, does semantic analysis on different query blocks & query expressions & eventually generates an execution plan with the help of the table and partition metadata looked up from the metastore.

④ Metastore:
- The component that stores all the structure information of the various tables and partitions ~~metadata looked up from the metastore~~. in the warehouse including column and column type information, the serializes & deserializes necessary to read and write data and the corresponding HDFS files where the data is stored.

(5) Execution Engine:

- The component which executes the execution plan created by the compiler.
- The plan is a DAG of stages.
- The execution engine manages the dependencies between these diff^n stages of the plan & executes these stages on the appropriate system component.

* Where to use hive?



→ If the data loaded and the schema does not match, then it is rejected

ratings → This is called as schema on write

- 5,000,000 → which means when we are writing the data at that time schema is enforced.

onal data → Hive supports schema on read, which means data is checked with the schema when any query is issued on it.

* Difference between Hive & Traditional Database:

HIVE	RDBMS
→ Hive enforces schema on read i.e schema does not Verify loading data.	→ RDBMS enforces schema on write i.e schema verify loading data, else rejected.
→ Hive is based on the notion of write once, Read many times.	→ RDBMS is designed for Read & write many times.
→ Hive data size is petabytes	→ maximum data size is Terabytes.
→ Hive doesn't support OLTP but it support OLAP.	→ only OLTP.
→ suited for static data analysis. (non real time data). ex. text file.	→ best suited for dynamic data analysis (real time data). ex. data from the sensors & web feeds.
→ Record level update is not possible in hive.	→ Record level Updates, Insertions, deletes & transactions are possible.
→ Hive is very easily scalable at low cost	→ RDBMS is not scalable to low cost.
→ Hive resembles a traditional DB by supporting SQL but it is not a database.	→ It is a database.

→ No support for indexes because data is always scanned.	→ supports indexes, it is very important for performance.
→ focus on only analytics	→ focus on analytics or online device connected to the network.
→ Distributed processing done via map/reduce.	→ Distributed processing varies by vendor. (company or person)
→ Scales up to hundred of nodes.	→ Scales to beyond 20 nodes.