Hive

Apache Hive is a data WH system that is built on top of hadoop.

→ data analysts
→ SQL developers

spark sql

Ease working on data
no Java or MR code

Facebook developed Hive ≈ 2010

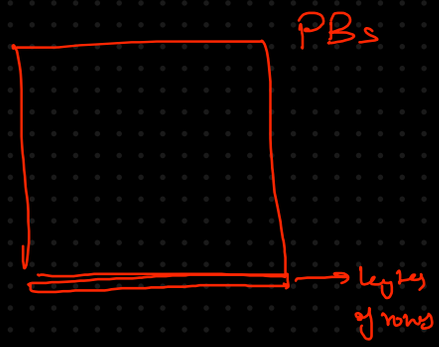
SQL like interface

HQL

PBs of data

Library =

librarian with a ipad / computer



How Hive makes data processing easier:

Without hive	With Hive
Java based MR code to process data	SQL like queries
Programmers	SQL users Analyst

Few Common Ques/misconceptions about Hive

Misconception

1. Hive is a database like SQL / Postgres SQL ?

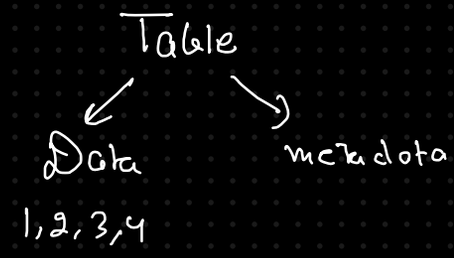
Hive is not a database but a data warehouse tool build on Hadoop.

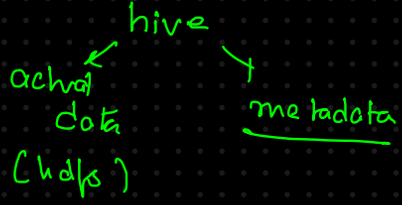
an interface to Query large datasets stored in hdfs

Feature	Traditional RDBMS (MySQL, PostgreSQL)	Apache Hive
Storage	Stores structured data in tables on disk	Uses **HDFS** for distributed storage
Query Execution	Uses in-memory processing	Uses **MapReduce, Tez, or Spark**
Transaction Type	OLTP (Row-level operations)	OLAP (Batch processing)

Feature	Traditional Databases (MySQL, PostgreSQL)	Apache Hive
Storage	Stores data in structured tables on disks	Stores data in **HDFS** (distributed storage)
Query Processing	Executes queries in-memory	Uses **MapReduce, Tez, or Spark** for distributed processing
Schema	Schema-on-Write (Data must fit schema)	Schema-on-Read (Can query raw files)
Transaction Type	OLTP (Transactional systems)	OLAP (Batch processing, analytics)
Speed	Fast for small queries, real-time updates	Optimized for large-scale batch processing

Hive is a metastore



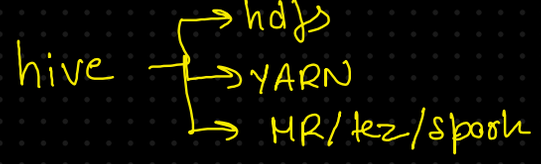


In big data world, any tech which supports data Querying not necessarily a dB

2. Hive is a replacement for Hadoop

built on top of hadoop & depends on hdfs, YARN

Query engine




```
Hive Session ID = 61b9b5ae-7609-498c-a06d-4fa12313a968
hive> set hive.execution.engine
    > ;
hive.execution.engine=tez
hive> set hive.execution.engine = mr
    > ;
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution eng
ine (i.e. spark, tez) or using Hive 1.X releases.
hive>
```

Why I teach things in depth ?

↳ Study nice or have doubts lifetime

end-to-end

MR ⟶ 3hrs

MR
Spark
Tez

3. Hive Queries run like normal SQL Queries DB

1st ⇒ abstraction of MR, Tez, Spark

~AI

select * from table ⟶ MR code

2nd distributed

4. Hive can perform Row level transactions like MySQL

⇒ update/delete

Hive as a technology
is designed for batch
Processing

Bazooka to kill an ant

→ it will not modify individual records, but create new partitions

5. Hive is always slower than Spark

Hive on MR ≃ slower

Hive on Spark/Tez ⇒ faster than MR

Hive is better for structured storage and
complex Queries

Use Case	Use Hive	Use Spark
Large-scale batch processing	✅ Yes	❌ No
Real-time analytics	❌ No	✅ Yes
ETL jobs on structured data	✅ Yes	❌ No
Streaming data processing	❌ No	✅ Yes

Hive only works with HDFS

s3, ADLS, GCS, hdfs

Feature	Hive Advantage
SQL-Like Interface	Allows non-programmers to work with big data.
Scalability	Handles **petabytes** of data.
Optimized Execution	Uses **MapReduce, Tez, or Spark** for distributed processing.
Storage Flexibility	Works with **HDFS, S3, Azure Blob, GCS**.
Schema Flexibility	Schema-on-read enables analyzing raw files.

Feature	Hive CLI	Beeline (Recommended)
Connection Type	Directly connects to Hive	Uses JDBC for **remote** execution
Security	No authentication	Supports authentication (LDAP, Kerberos)
Multiple Connections	Single session only	Supports multiple concurrent sessions
Performance	More resource-heavy	Optimized for **query performance**
Recommended?	❌ Deprecated	✅ Yes, for production use