Recap:

Hive - Data warehouse of Hadoop

Hive execution engines -> Mapreduce, spark, Tez

Metadata + data => Hive

HiveQL

Managed table vs External table

Partitioning vs Bucketing

File formats

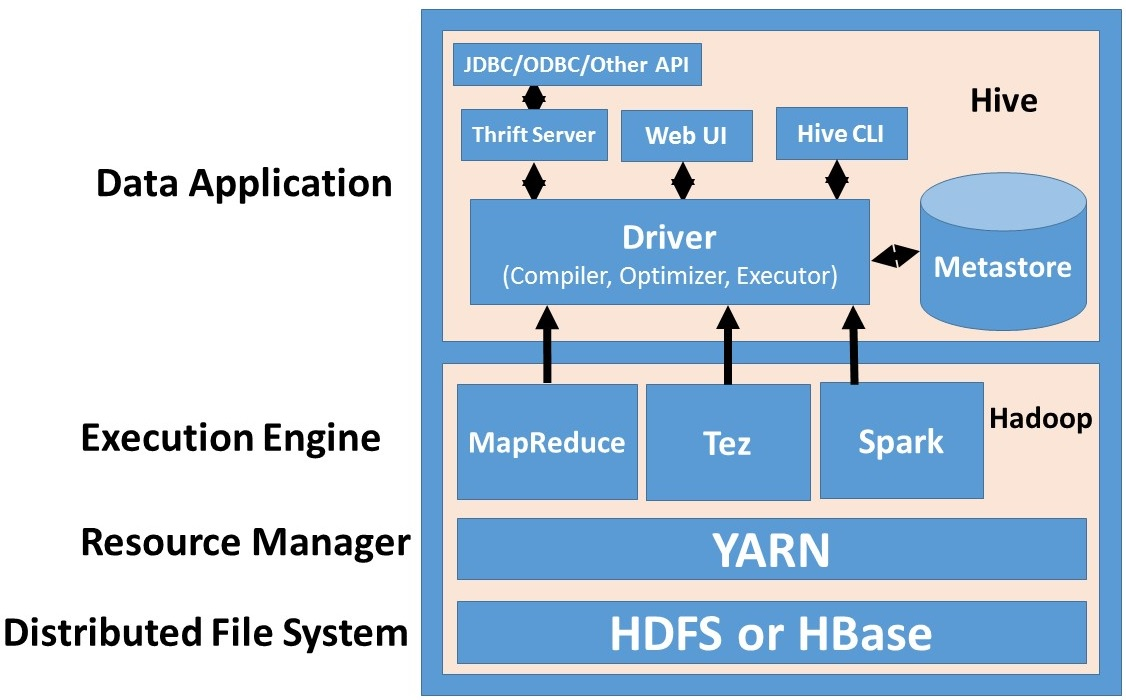
Pyspark => partitions

Repartition vs coalesce

Dataframe writer mode --->

Different ways to load data into Hive

1. Insert
2. Load data local inpath
3. Load data inpath (from existing HDFS directory) --it moves the data from HDFs to Hive HDFS location
4. SQOOP -- Bulk data transfer tool
5. ETL tools
6. Pyspark → DF --write -> hive table



======================================

Spark context is an entry point of spark job

Spark context → sc → only one context runs in a single JVM

Spark session → spark -> it allows multiple context to run it in a single JVM (java virtual machine )

Temp view vs global temp view

Single session vs cross-session