Recap

- MapReduce counters
- Performance tuning in MapReduce jobs
- MapReduce job chaining
- Pig
- SQL Commands

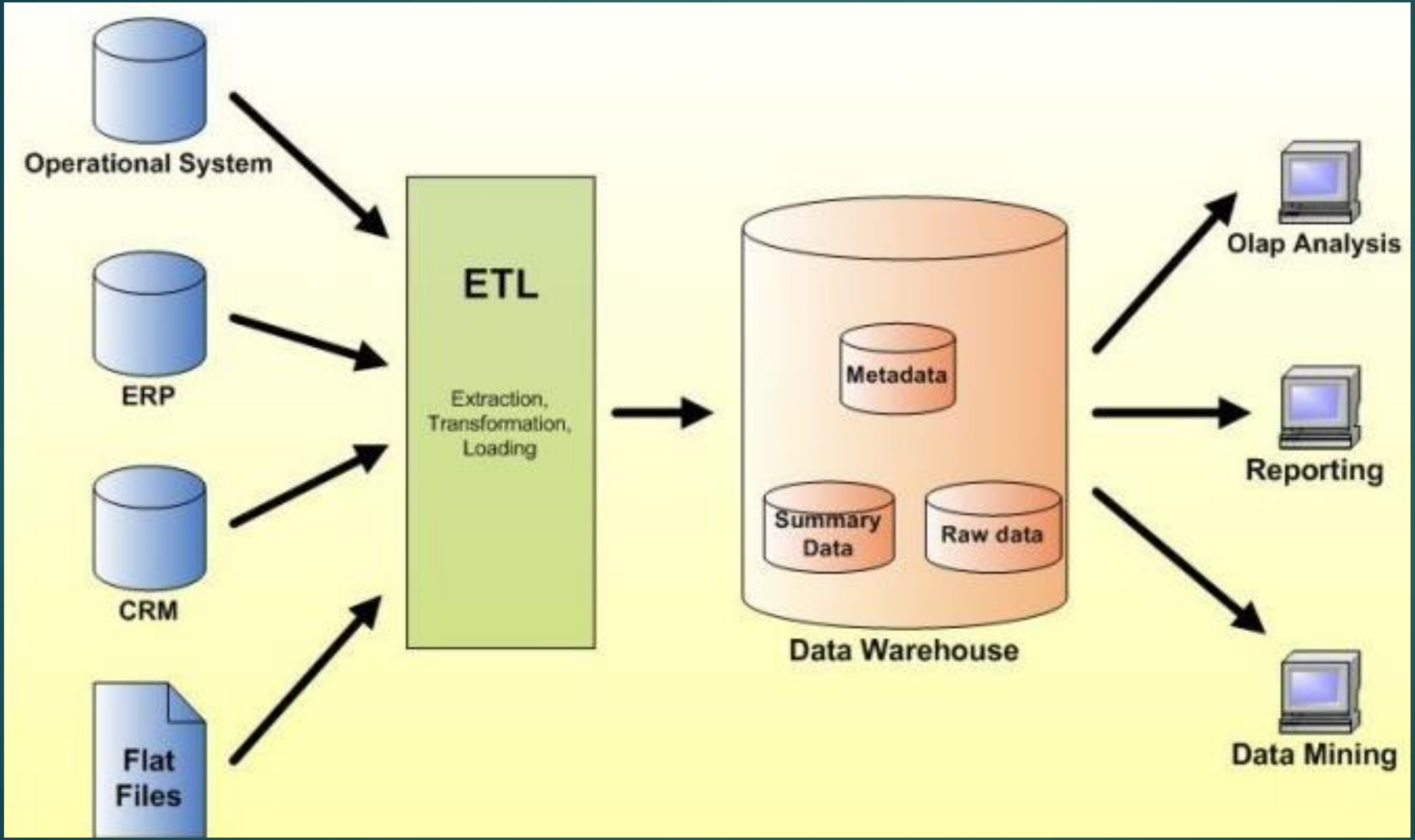
Agenda for today

- Hive
- Impala

Hive

- A data warehousing tool developed by Facebook and then contributed as Apache project
- Reliable batch processing on very large; structured dataset
- SQL on Hadoop

Typical data warehouse



Basic data warehouse cycle

- Lets try out such a very basic cycle using hive
- Please refer to first section of exercise document for more details

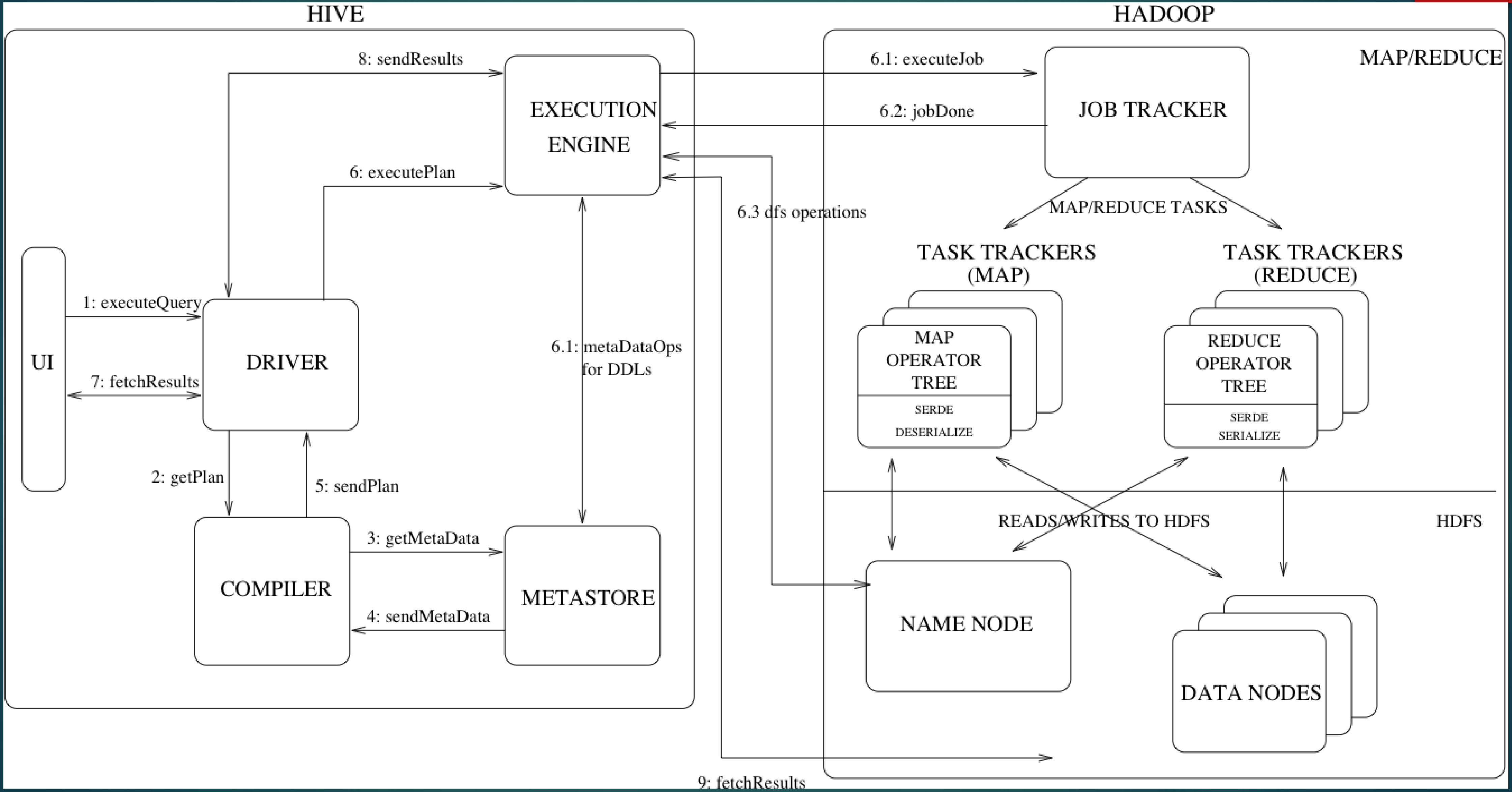
Impala

- Ad-hoc querying tool developed by Cloudera and contributed as Apache project
- Massively Parallel Processing(MPP) architecture on Hadoop
- Provides most of the SQL functionalities
- Uses Hive metastore to store table metadata

References

- Hive reference book http://shop.oreilly.com/product/0636920023555.do
- Impala command reference https://www.cloudera.com/documentation/enterprise/5-9-x/topics/impala_langref_sql.html
- Teradata MPP architecture https://www.tutorialspoint.com/teradata/teradata_architecture.htm
- By Davod - Own work, using File:Apache Hive logo.jpg as base., Apache License 2.0, https://commons.wikimedia.org/w/index.php?curid=44338923
- Hive DML commands https://cwiki.apache.org/confluence/display/Hive/LanguageManual+Select

Hive architecture


HIVE
HADOOP
MAP/REDUCE
8: sendResults
EXECUTION ENGINE
6.1: executeJob
6.2: jobDone
JOB TRACKER
6: executePlan
6.3 dfs operations
MAP/REDUCE TASKS
TASK TRACKERS (MAP)
TASK TRACKERS (REDUCE)
UI
1: executeQuery
DRIVER
7: fetchResults
6.1: metaDataOps for DDLs
MAP OPERATOR TREE
SERDE DESERIALIZE
REDUCE OPERATOR TREE
SERDE SERIALIZE
2: getPlan
5: sendPlan
READS/WRITES TO HDFS
HDFS
3: getMetaData
COMPILER
METASTORE
4: sendMetaData
NAME NODE
DATA NODES
9: fetchResults