Apache HIVE

Image courtesy : https://en.wikipedia.org/wiki/Apache_Hive

Agenda

- **What is Apache HIVE**
- **Why HIVE**
- **How does HIVE fit into the Hadoop technology landscape**
- **Limitations of HIVE**

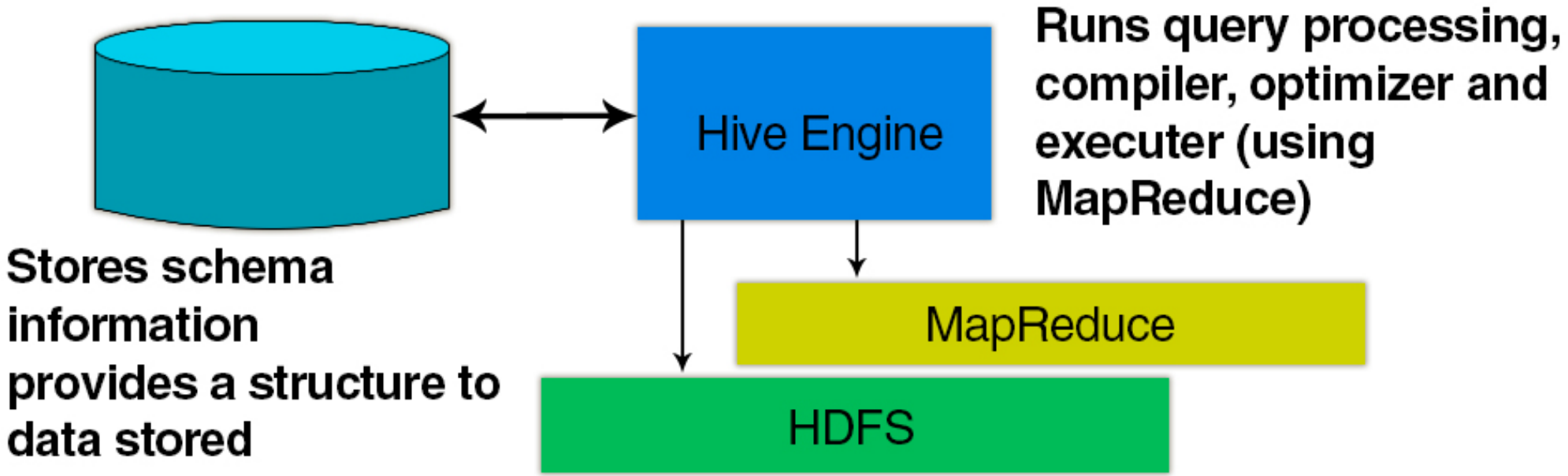
What is HIVE ?

- **HIVE is a query interface on top of Hadoop's native Map-Reduce**
- **HIVE is a data warehouse**
- **HIVE allows users to write SQL style queries in a native language known as Hive Query Language (HQL)**
- **HIVE execution engine converts the scripts written in HQL into JAR files (map reduce) to execute in the cluster**
- **HIVE reads data from HDFS**
- **Allows creation of tables to operate on structured data**
- **The table's schema information (table meta data) is saved in HIVE metastore which is borrowed from an RDBMS (Derby is default database)**
- **HIVE is not an RDBMS**

Why HIVE ?

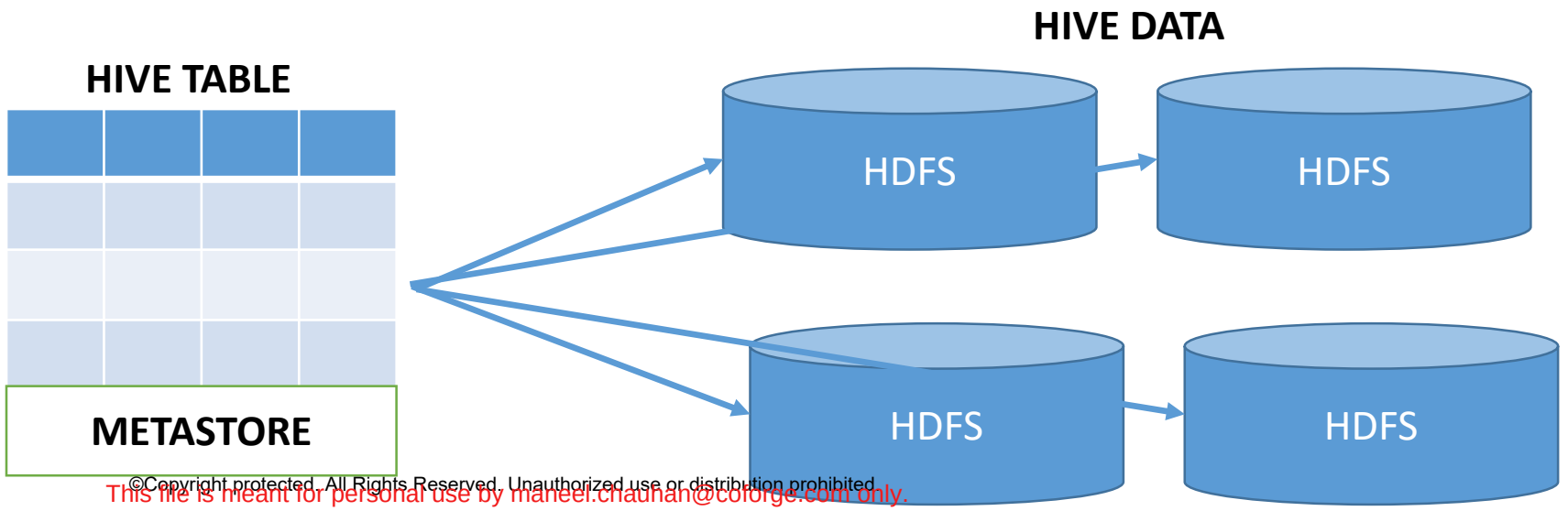
- **Hadoop is known for its Map-Reduce engine for parallelizing data processing operations using HDFS as its native file storage system**
- **Map Reduce does not provide user friendly libraries or interfaces to deal with unstructured data handling**
- **Very tight dependency of JAVA if one needs to use the Map-Reduce framework**
- **An operation like left inner join would need around 200-300 lines of code in JAVA Map-Reduce whereas in SQL it would just be a couple of lines of code**
- **Analysts from SQL experience of having come from RDBMS world and DW/BI world cannot program in JAVA in order to use**
- **To enable SQL developers to exploit the power of Hadoop, an abstraction interface was developed on top of native Map-Reduce**
- **This interface (engine) was called HIVE and was officially developed by Facebook and initial release was in the year 2010**

Architectural overview

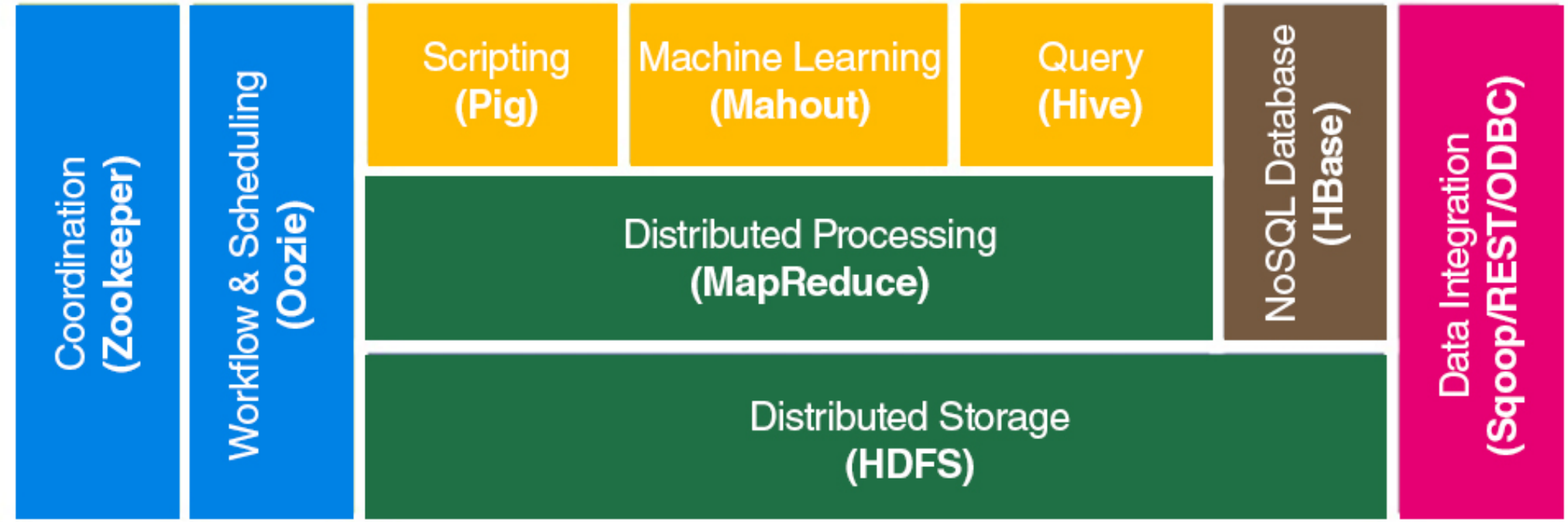


Working of HIVE

- **Hive allows a way to project a table structure on the data in HDFS (structured data in HDFS)**
- **The table meta data is saved separately from the data**
- **In reality, we do not actually load the data into the place where HIVE tables are created**
- **HIVE table information (table meta data is saved in meta store)**



How does HIVE fit into the Hadoop ecosystem



Things HIVE cannot do efficiently

- **Ad hoc real time queries**
- **OLTP (Online Line Transaction Processing)**
- **No ACID support (ACID support is limited)**
- **Not suited for frequent updates and inserts (inserts and updates are allowed in recent releases of HIVE)**
- **Not recommended for small data sets**
- **Not meant for unstructured data analysis**

DEMO