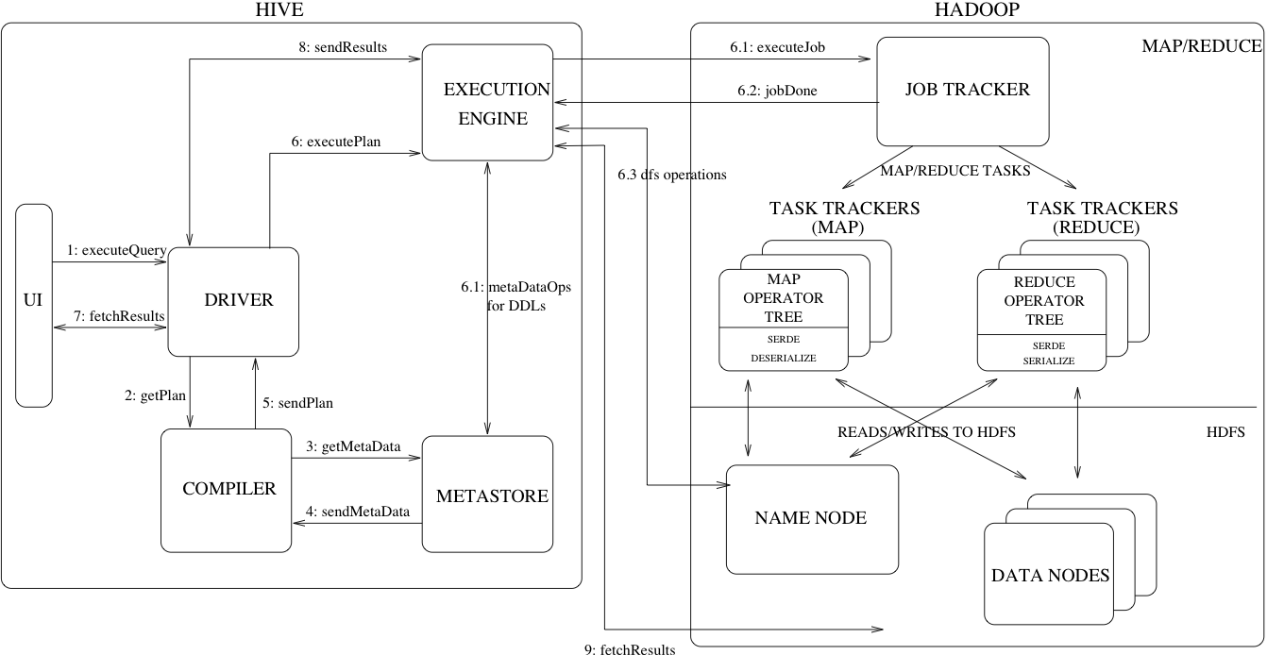
Hive is a data warehousing infrastructure based on Apache Hadoop



* **Databases**: Namespaces function to avoid naming conflicts for tables, views, partitions, columns, and so on.
* **Tables**: Homogeneous units of data which have the same schema.
* **Partitions** 分区: Each Table can have one or more partition Keys which determines how the data is stored. Partitions—apart from being storage units—also allow the user to efficiently identify the rows that satisfy a specified criteria.
* **Buckets** 分桶: Data in each partition may in turn be divided into Buckets based on the value of a hash function of some column of the Table. 【不常用】
* **Metastore** is a central repository of metadata for Hive tables and partitions in a relational database.

**Language Manual**

1. 表结构

LOCATION：表示文件所在位置，俗称EXTERNAL外表，如：HDFS、S3等。

LOAD DATA：数据加载动作；默认追加，带OVERWRITE表示覆盖。

INDEX：has been **removed** in version 3.0

1. 表查询

ORDER BY：in order to impose total order of all results, there has to be one reducer to sort the final output.

SORT BY：sort the rows before feeding the rows to a reducer.

DISTRIBUTE BY：All rows with the same Distribute By columns will go to the same reducer.

CLUSTER BY：is a short-cut for both Distribute By and Sort By.

1. 表优化

EXPLAIN

HiveServer2 has its own CLI called **Beeline**.

1. 碎文件

concatenate 命令只支持 RCFILE 和 ORC 文件类型。

1. 数据倾斜