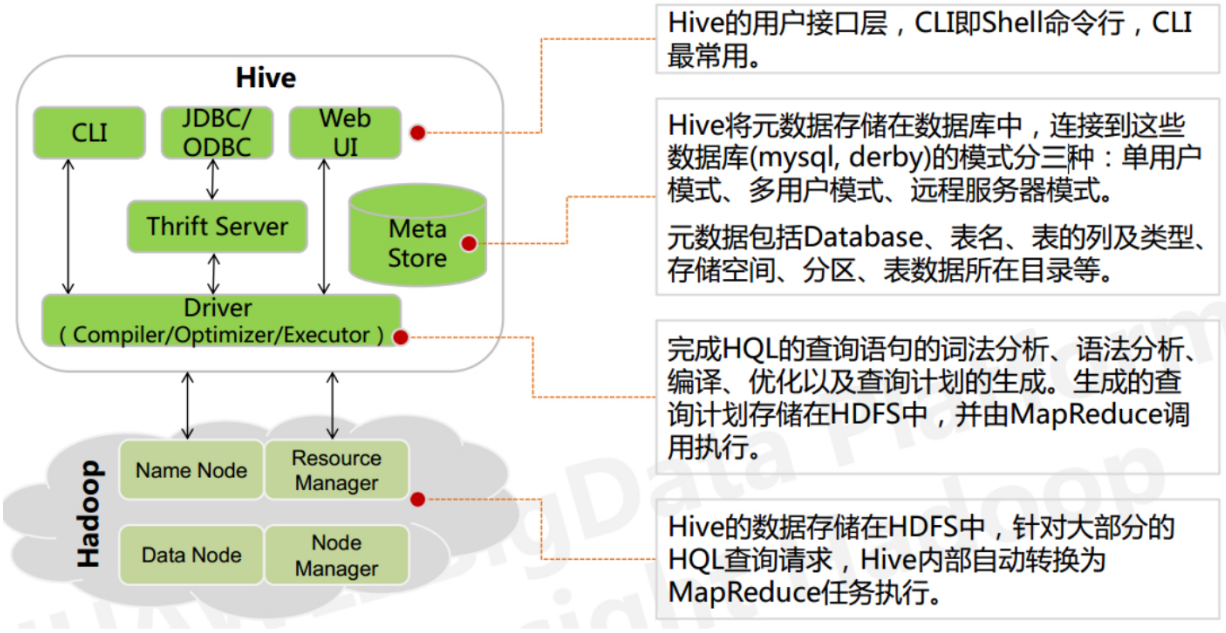
hive 是构建于Hadoop之上的数据仓库，通过将hdfs上的结构化数据映射为数据库表实现对大批量数据的查询操作，底层存储依赖于HDFS，SQL查询依赖于MapReduce，因此hive**不适用于少量数据的插入和实时性查询**，**不支持更新、删除操作**，通常被**用于大批量数据导入与大批量离线查询。**

架构



hive 的架构可以理解为了处理**将SQL语句转化为MapReduce任务的工具**。

Thrift Server

Facebook 开发的一个软件框架，可以用来进行可扩展且跨语言的服务的开发， Hive 集成了该服务，能让不同的编程语言调用 Hive 的接口。

Driver

完成 HQL 查询语句从词法分析，语法分析，编译，优化，以及生成逻辑执行 计划的生成。生成的逻辑执行计划存储在 HDFS 中，并随后由 MapReduce 调用执行,Hive 的核心是驱动引擎， 驱动引擎由四部分组成：

(1) 解释器：解释器的作用是将 HiveSQL 语句转换为抽象语法树（ AST ）

(2) 编译器：编译器是将语法树编译为逻辑执行计划

(3) 优化器：优化器是对逻辑执行计划进行优化

(4) 执行器：执行器是调用底层的运行框架执行逻辑执行计划

Meta Store

元数据，存储在 Hive 中的数据的描述信息。

Hive 中的元数据通常包括：**表的名字，表的列和分区及其属性，表的属性（内部表和 外部表），表的数据所在目录**

Metastore 默认存在自带的 Derby 数据库中。缺点就是不适合多用户操作，并且数据存 储目录不固定。数据库跟着 Hive 走，极度不方便管理

解决方案：通常存我们自己创建的 MySQL 库（本地 或 远程）**Hive 和 MySQL 之间通过 MetaStore 服务交互**