Hive 是什么？

- 数据仓库工具
- 将结构化数据转化为数据库表
- 将sql转化为mapreduce任务
- 适用于大批量的离线计算

Hive 特点

优点：

1、可扩展性,横向扩展，Hive 可以自由的扩展集群的规模

2、延展性，Hive 支持自定义函数，用户可以根据自己的需求来实现自己的函数

3、良好的容错性，可以保障即使有节点出现问题，SQL 语句仍可完成执行

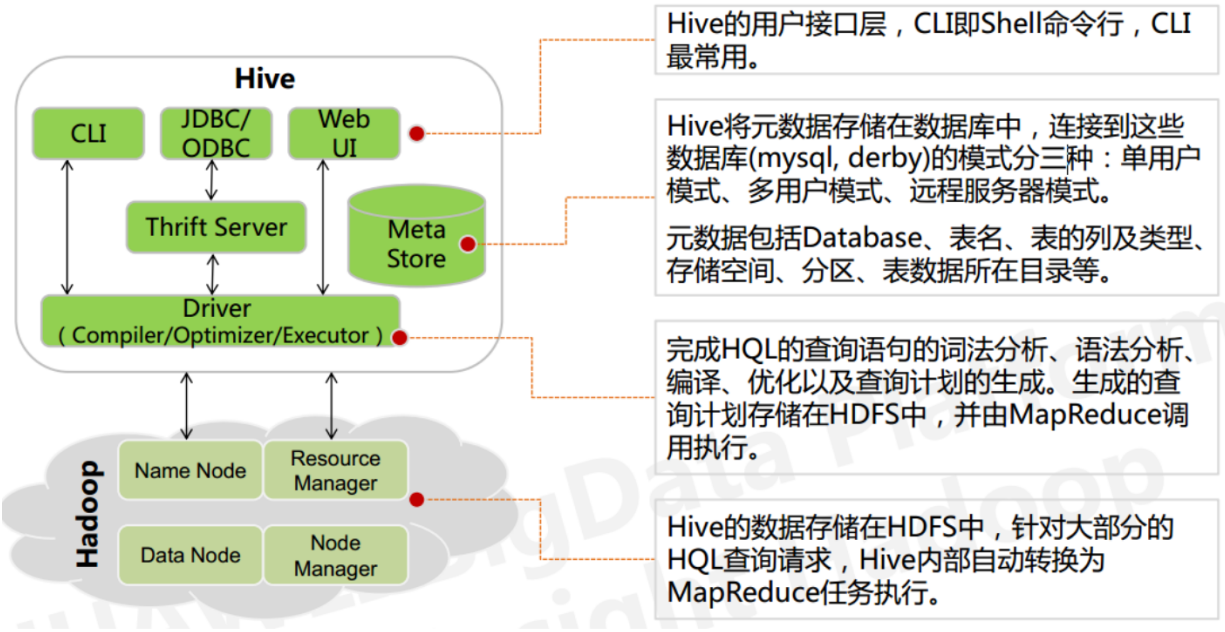
缺点：

1、Hive 不支持记录级别的增删改操作，但是用户可以通过查询生成新表或者将查询结果导入到文件中（当前选择的 hive-2.3.2 的版本支持记录级别的插入操作）

2、Hive 的查询延时很严重，因为 MapReduce Job 的启动过程消耗很长时间，所以不能用在交互查询系统中。

3、Hive 不支持事务（因为没有增删改，所以主要用来做 OLAP（联机分析处理），而不是 OLTP（联机事务处理），这就是数据处理的两大级别）。

Hive 架构



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

执行流程

HiveQL 通过命令行或者客户端提交，经过 Compiler 编译器，运用 MetaStore 中的元数据进行类型检测和语法分析，生成一个逻辑方案(Logical Plan)，然后通过的优化处理，产生 一个 MapReduce 任务。

Hive 的数据组织

Hive 的存储结构包括**数据库、表、视图、分区和表数据**等。数据库，表，分区等等都对 应 HDFS 上的一个目录。表数据对应 HDFS 对应目录下的文件。

Hive 中所有的数据都存储在 HDFS 中，没有专门的数据存储格式，因为 Hive 是**读模式（Schema On Read**），可支持 **TextFile，SequenceFile，RCFile 或者自定义格式**等。只需要在创建表的时候告诉 Hive 数据中的列分隔符和行分隔符，Hive 就可以解析数据

Hive 的默认列分隔符：控制符 **Ctrl + A，\x01**

Hive 的默认行分隔符：换行符 **\n**

数据模型

- **database：**在 HDFS 中表现为${hive.metastore.warehouse.dir}目录下一个文件夹
- **table（内部表）：**在 HDFS 中表现所属 database 目录下一个文件夹
- **external table（外部表/临时表）：**与 table 类似，不过其数据存放位置可以指定任意 HDFS 目录路径
- **partition**：在 HDFS 中表现为 table 目录下的子目录
- **bucket：**在 HDFS 中表现为同一个表目录或者分区目录下根据某个字段的值进行 hash 散 列之后的多个文件
- **view：**与传统数据库类似，**只读**，基于基本表创建

内部表、外部表、分区表和 Bucket 表的区别

内部表和外部表的区别

1. 删除内部表，删除表元数据和数据
2. 删除外部表，删除元数据，不删除数据
3. 内部表由 hive 进行管理，外部表由 hdfs 进行管理

内部表和外部表的使用选择

1. 如果数据的所有处理都在 Hive 中进行，那么倾向于 选择内部表，但是如果 Hive 和其他工具要针对相同的数据集进行处理，外部表更合适。
2. 使用外部表访问存储在 HDFS 上的初始数据，然后通过 Hive 转换数据并存到内部表中
3. 使用外部表的场景是针对一个数据集有多个不同的 Schema
4. 通过外部表和内部表的区别和使用选择的对比可以看出来，hive 其实仅仅只是对存储在 HDFS 上的数据提供了一种新的抽象。而不是管理存储在 HDFS 上的数据。所以不管创建内部 表还是外部表，都可以对 hive 表的数据存储目录中的数据进行增删操作。

分区表和分桶表的区别

1. Hive 数据表可以根据某些字段进行分区操作，细化数据管理，可以让部分查询更快。同 时表和分区也可以进一步被划分为 Buckets，分桶表的原理和 MapReduce 编程中的 HashPartitioner 的原理类似。
2. 分区和分桶都是细化数据管理，但是分区表是手动添加区分，由于 Hive 是读模式，所 以对添加进分区的数据不做模式校验，分桶表中的数据是按照某些分桶字段进行 hash 散列 形成的多个文件，所以数据的准确性也高很多

hive 与传统数据库的比较

参考文章：

> https://www.cnblogs.com/qingyunzong/p/8707885.html