hive实战1

大数据技术与应用专业

什么是Hive

- Hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供类SQL查询功能。
- 其本质是将SQL转换为MapReduce的任务进行运算，底层由HDFS来提供数据的存储，说白了hive可以理解为一个将SQL转换为MapReduce的任务的工具，甚至更进一步可以说hive就是一个MapReduce的客户端

为什么使用Hive

- 直接使用hadoop所面临的问题
  - 人员学习成本太高
  - 项目周期要求太短
  - MapReduce实现复杂查询逻辑开发难度太大

- 为什么要使用Hive
  - 操作接口采用类SQL语法，提供快速开发的能力。
  - 避免了去写MapReduce，减少开发人员的学习成本。
  - 功能扩展很方便。

Hive的特点

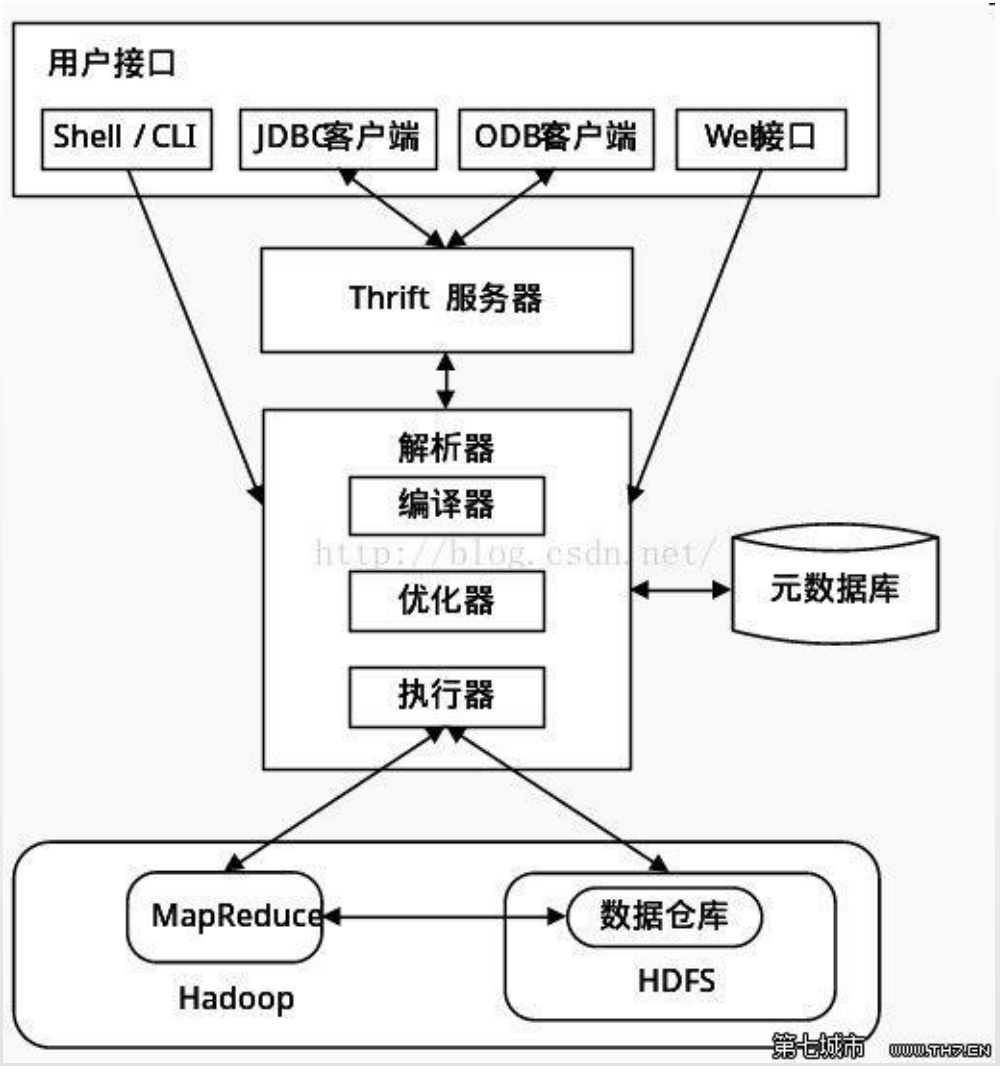
- 可扩展
  - Hive可以自由的扩展集群的规模，一般情况下不需要重启服务。
- 延展性
  - Hive支持用户自定义函数，用户可以根据自己的需求来实现自己的函数。
- 容错
  - 良好的容错性，节点出现问题SQL仍可完成执行。

Hive架构

基本组成

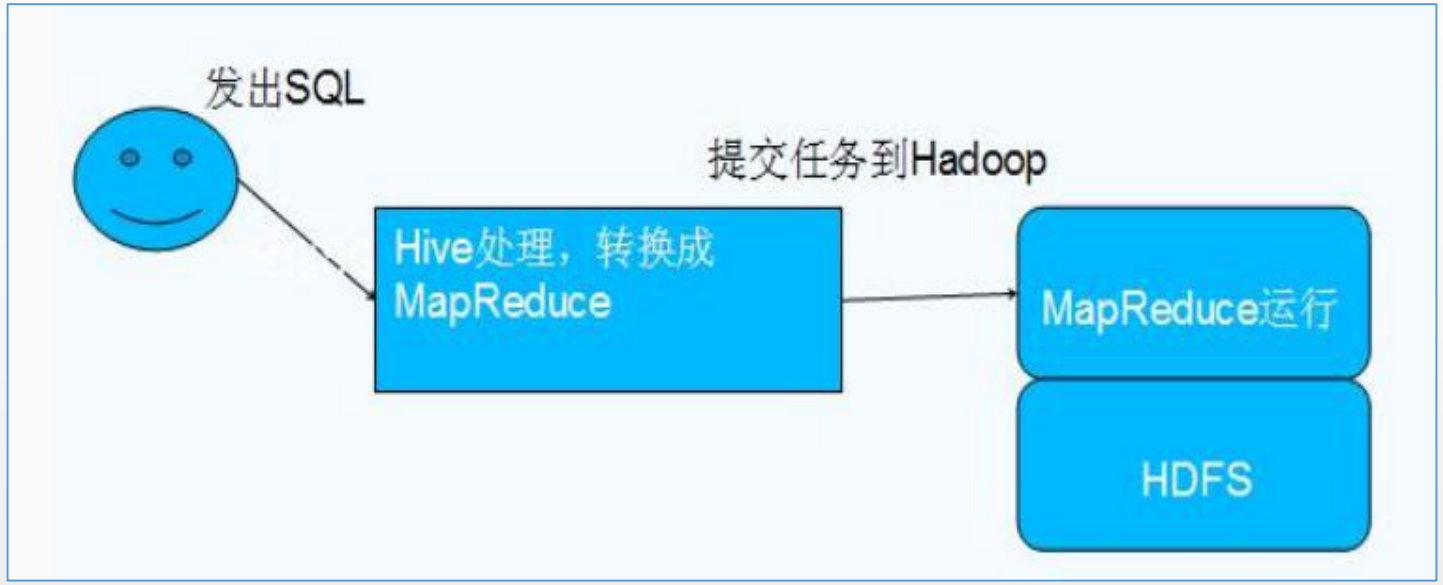
用户接口：包括CLI、JDBC/ODBC、WebGUI。其中，CLI(command line interface)为shell命令行；JDBC/ODBC是Hive的JAVA实现，与传统数据库JDBC类似；WebGUI是通过浏览器访问Hive。
元数据存储：通常是存储在关系数据库如mysql/derby中。Hive 将元数据存储在数据库中。Hive 中的元数据包括表的名字，表的列和分区及其属性，表的属性（是否为外部表等），表的数据所在目录等。

解释器、编译器、优化器、执行器:完成HQL 查询语句从词法分析、语法分析、编译、优化以及查询计划的生成。生成的查询计划存储在HDFS 中，并在随后有MapReduce 调用执行。



Hive与Hadoop的关系

- Hive利用HDFS存储数据，利用MapReduce查询分析数据



Hive与传统数据库对比

- hive用于海量数据的离线数据分析
- hive具有sql数据库的外表，但应用场景完全不同，hive只适合用来做批量数据统计分析

	Hive	RDBMS
查询语言	HQL	SQL
数据存储	HDFS	Raw Device or Local FS
执行	MapReduce	Excutor
执行延迟	高	低
处理数据规模	大	小
索引	0.8版本后加入位图索引	有复杂的索引

Hive的数据存储

- 1、Hive中所有的数据都存储在 HDFS 中，没有专门的数据存储格式（可支持Text，SequenceFile，ParquetFile，ORC格式RCFILE等）
  - SequenceFile是hadoop中的一种文件格式：
  - 文件内容是以序列化的kv对象来组织的

- 2、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。

- 3、Hive 中包含以下数据模型：DB、Table，External Table，Partition，Bucket。
  - db：在hdfs中表现为${hive.metastore.warehouse.dir}目录下一个文件夹
  - table：在hdfs中表现所属db目录下一个文件夹
  - external table：与table类似，不过其数据存放位置可以在任意指定路径
  - 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下根据hash散列之后的多个文件