Hive 基本概念

1、Hive 简介

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

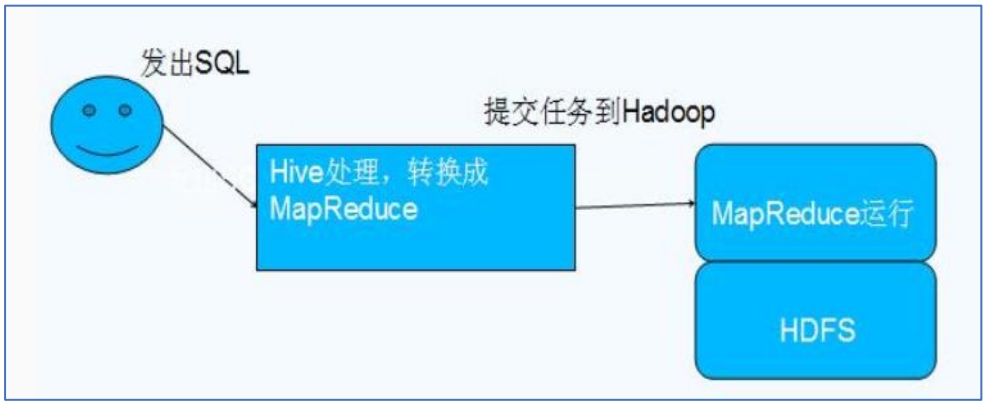
  Hive 支持用户自定义函数，用户可以根据自己的需求来实现自己的函数。

- 容错

  良好的容错性，节点出现问题 SQL 仍可完成执行。

2、Hive 与 Hadoop 的关系

Hive 利用 HDFS 存储数据，利用 MapReduce 查询分析数据



3、Hive 与传统数据库对比

hive 用于海量数据的离线数据分析

	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

总结：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