Hive到底是什么

写在前面的话
，学Hive这么久了，发现目前国内还没有一本完整的介绍Hive的书籍，而且互联网上面的资料很乱，于是我决定写一些关于《Hive的那些事》
序列文章，分享给大家。我会在接下来的时间整理有关Hive的资料，如果对Hive的东西感兴趣，请关注本博客。https://www.iteblog.com/archives/tag/hive-technology/

Hive最初是应Facebook每天产生的海量新兴社会网络数据进行管理和机器学习的需求而产生和发展的。那么，到底什么是Hive，我们先看看Hive官网Wiki是如何介绍Hive的(https://cwiki.apache.org/confluence/display/Hive/Home)：

The Apache Hive™ data warehouse software facilitates querying and managing large datasets residing in distributed storage. Built on top of Apache Hadoop™, it provides:
(1)、Tools to enable easy data extract/transform/load (ETL)
(2)、A mechanism to impose structure on a variety of data formats
(3)、Access to files stored either directly in Apache HDFS™ or in other data storage systems such as Apache HBase™
(4)、Query execution via MapReduce

上面英文的大致意思是：Apache
Hive数据仓库软件提供对存储在分布式中的大型数据集的查询和管理，它本身是建立在Apache Hadoop之上，主要提供以下功能：它提供了一系列的工具，可用来对数据进行提取/转化/加载（ETL）；是一种可以存储、查询和分析存储在HDFS（或者HBase）中的大规模数据的机制；查询是通过MapReduce来完成的（并不是所有的查询都需要MapReduce来完成，比如select * from XXX就不需要；在Hive0.11对类似select a,b from
XXX的查询通
过配置也可以不通过MapRe
duce来完成，具体怎么配置请参见本博客《Hive：简单查询不启用Mapreduce job而启用Fetch task》）。
从上面的定义我们可以了解到，Hive是一种建立在Hadoop文件系统上的数据仓库架构，并对存储在HDFS中的数据进行分析和管理；那么，我们如何来分析和管理那些数据呢？
Hive定义了一种类似SQL的查询语言，被称为HQL，对于熟悉SQL的用户可以直接利用Hive来查询数据。同时，这个语言也允许熟悉 MapReduce 开发者们开发自定义的mappers和reducers来处理内建的mappers和reducers无法完成的复杂的分析工作。Hive可以允许用户编写自己定义的函数UDF，来在查询中使用。Hive中有3种UDF：User Defined Functions（UDF）、User Defined Aggregation Functions（UDAF）、User Defined Table Generating
Functions（UDTF）。
今天，Hive已经是一个成功的Apache项目，很多组织把它用作一个通用的、可伸缩的数据处理平台。
当然，Hive和传统的关系型数据库有很大的区别，Hive将外部的任务解析成一个MapReduce可执行计划，而启动MapReduce是一个高延迟的一件事，每次提交任务和执行任务都需要消耗很

多时间，这也就决定Hive只能处理一些高延迟的应用（如果你想处理低延迟的应用，你可以去考虑一下Hbase）。同时，由于设计的目标不一样，Hive目前还不支持事务；不能对表数据进行修改（不能更新、删除、插入；只能通过文件追加数据、重新导入数据）；不能对列建立索引（但是Hive支持索引的建立，但是不能提高Hive的查询速度。如果你想提高Hive的查询速度，请学习Hive的分区、桶的应用）。