skyme

浅谈Hive vs. HBase

2年前 1217

对于刚接触大数据的用户来说，要想区分Hive与HBase是有一定难度的。本文将尝试从其各自的定义、特点、限制、应用场景等角度来进行分析，以作抛砖引玉之用。

Hive是什么？

Apache Hive是一个构建于Hadoop(分布式系统基础架构)顶层的数据仓库，注意这里不是数据库。Hive可以看作是用户编程接口，它本身不存储和计算数据；它依赖于HDFS(Hadoop分布式文件系统)和MapReduce(一种编程模型，映射与化简；用于大数据并行运算)。其对HDFS的操作类似于SQL—名为HQL，它提供了丰富的SQL

查询方式来分析存储在HDFS中的数据；HQL经过编译转为MapReduce作业后通过自己的SQL 去查询分析需要的内容；这样一来，即使不熟悉MapReduce 的用户也可以很方便地利用SQL 语言查询、汇总、分析数据。而MapReduce开发人员可以把己写的mapper 和reducer 作为插件来支持Hive 做更复杂的数据分析。

HBase是什么？

Apache HBase是运行于HDFS顶层的NoSQL(=Not Only SQL，泛指非关系型的数据库)数据库系统。区别于Hive，HBase具备随即读写功能，是一种面向列的数据库。HBase以表的形式存储数据，表由行和列组成，列划分为若干个列簇(row family)。例如：一个消息列簇包含了发送者、接受者、发送日期、消息标题以及消息内容。每一对键值在HBase会被定义为一个Cell，其中，键由row-key(行键)，列簇，列，时间戳构

成。而在HBase中每一行代表由行键标识的键值映射组合。Hbase目标主要依靠横向扩展，通过不断增加廉价的商用服务器，来增加计算和存储能力。

特性

遵从JDBC的Hive不但可以让具SQL知识的用户来间接执行MapReduce作业，同时里面也整合了目前基于SQL的操作工具。不过，由于默认的数据读取是全表遍历的，其时间的耗费也不可避免地相对较大。尽管如此，不尽相同的Hive分区方法，其遍历读取的数据量也是能够有所限制的。Hive分区允许对存储在独立文件上的数据进行筛选查询，返回的是筛选后的数据。例如针对日期的日志文件访问，前提是该类文件的文件名包含日期信息。

HBase以键值对的形式储存数据。其包含了4种主要的数据操作方式:

添加或更新数据行

扫描获取某范围内的cells

为某一具体数据行返回对应的cells

从数据表中删除数据行/列，或列的描述信息

列信息可用于获取数据变动前的取值（透过HBase压缩策略可以删除列信息历史记录来释放存储空间）。

限制

Hive不支持常规的SQL更新语句，如：数据插入，更新，删除。因为其对数据的操作是针对整个数据表的。同时该特点也使得数据查询用时以数分钟甚至数小时来进行计算。此外，其MapReduce转换过程必须遵从预定义的转换规则。

HBase的数据查询是有一套属于自己类似SQL的操作语言的，这个需要一定的学习来掌握。此外，要运行HBase，ZooKeeper是需要配备的。ZooKeeper是一个针对大型分布式系统的可靠协调系统，提供的功能包括：配置维护、名字服务、分布式同步、组服务等。

应用举例

Hive适用于网络日志等数据量大、静态的数据查询。例如：用户消费行为记录，网站访问足迹等。但是不适用于联机实时在线查询的场合。

HBase能在大数据联机实时查询场合大展身手。例如：Fackbook就利用其对用户间的传送的消息进行联机实时分析。

小结

Hive与HBase两者是基于Hadoop上不同的技术。Hive是一种能执行MapReduce作业的类SQL编程接口，Hbase是一种非关系型的数据库结构。结合这两者自身的特点，互相结合使用或许能收到相得益彰的效果。例如：利用Hive处理静态离线数据，利用HBase进行联机实时查询，而后对两者间的结果集进行整合归并，从而使得数据完整且永葆青春，为进一步的商业分析提供良好支持。

作者

skyme

TA的文章

Cloudera Manager简介

web日志中的频繁访问日志挖掘

scala实现单例模式

相关文章

浅谈开源大数据平台的演变

浅谈开源大数据平台的演变

夏梦竹谈**Hive vs. HBase**的区别

浅谈**zookeeper**的在**hbase**集群中的作用

Hadoop生态上几个技术的关系与区别：**hive**、**pig**、**hbase** 关系与区别