20180321_hive 和 HBASE的区别

先放结论：**Hbase和Hive在大数据架构中处在不同位置，Hbase主要解决实时数据查询问题，Hive主要解决数据处理和计算问题，一般是配合使用。**

一、区别：

1. Hbase： Hadoop database 的简称，也就是基于Hadoop数据库，是一种NoSQL数据库，主要适用于海量明细数据（十亿、百亿）的随机实时查询，如日志明细、交易清单、轨迹行为等。
2. Hive：Hive是Hadoop数据仓库，**严格来说，不是数据库，主要是让开发人员能够通过SQL来计算和处理HDFS上的结构化数据，适用于离线的批量数据计算。**

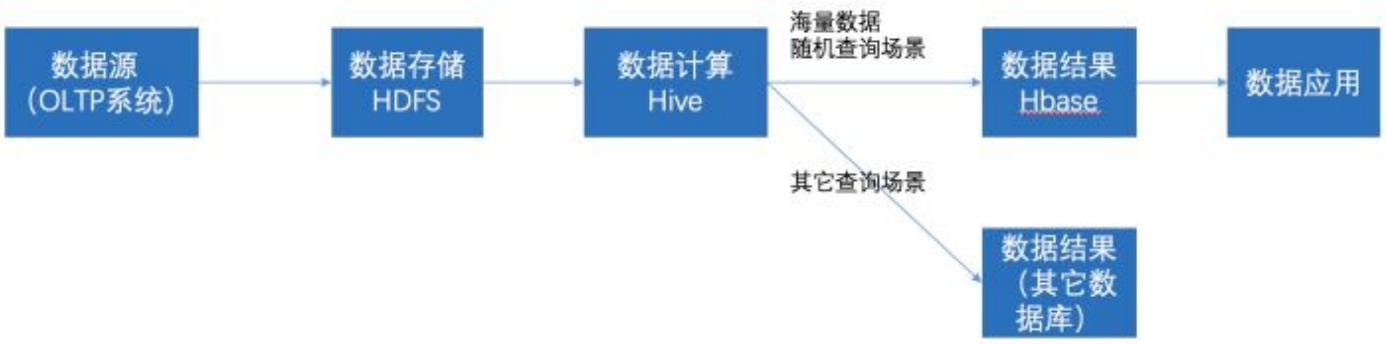
- 通过元数据来描述Hdfs上的结构化文本数据，通俗点来说，就是定义一张表来描述HDFS上的结构化文本，包括各列数据名称，数据类型是什么等，方便我们处理数据，当前很多SQL ON Hadoop的计算引擎均用的是hive的元数据，如Spark SQL、Impala等；
- 基于第一点，通过SQL来处理和计算HDFS的数据，Hive会将SQL翻译为Mapreduce来处理数据；

HBASE-数据库

二、关系

在大数据架构中，Hive和HBase是协作关系，数据流一般如下图：

1. 通过ETL工具将数据源抽取到HDFS存储；
2. 通过Hive清洗、处理和计算原始数据；
3. HIve清洗处理后的结果，如果是面向海量数据随机查询场景的可存入Hbase
4. 数据应用从HBase查询数据；



---

1. **两者分别是什么?**

   Apache Hive是一个构建在Hadoop基础设施之上的数据仓库。通过Hive可以使用HQL语言查询存放在HDFS上的数据。HQL是一种类SQL语言，这种语言最终被转化为Map/Reduce. 虽然Hive提供了SQL查询功能，但是**Hive不能够进行交互查询–因为它只能够在Haoop上批量的执行Hadoop。**

Apache HBase是一种Key/Value系统，它运行在HDFS之上。和Hive不一样，Hbase的能够在它的数据库上实时运行，而不是运行MapReduce任务。Hive被分区为表格，表格又被进一步分割为列簇。列簇必须使用schema定义，列簇将某一类型列集合起来（列不要求schema定义）。例如，“message”列簇可能包含：“to”, ”from” “date”, “subject”, 和”body”. 每一个 key/value对在Hbase中被定义为一个cell，每一个key由row-key，列簇、列和时间戳。在Hbase中，行是key/value映射的集合，这个映射通过row-key来唯一标识。Hbase利用Hadoop的基础设施，可以利用通用的设备进行水平的扩展。

1. 两者的特点

   Hive帮助熟悉SQL的人运行MapReduce任务。因为它是JDBC兼容的，同时，它也能够和现存的SQL工具整合在一起。运行Hive查询会花费很长时间，因为它会默认遍历表中所有的数据。虽然有这样的缺点，一次遍历的数据量可以通过Hive的分区机制来控制。分区允许在数据集上运行过滤查询，这些数据集存储在不同的文件夹内，查询的时候只遍历指定文件夹（分区）中的数据。这种机制可以用来，例如，只处理在某一个时间范围内的文件，只要这些文件名中包括了时间格式。

HBase通过存储key/value来工作。它支持四种主要的操作：增加或者更新行，查看一个范围内的cell，获取指定的行，删除指定的行、列或者是列的版本。版本信息用来获取历史数据（每一行的历史数据可以被删除，然后通过Hbase compactions就可以释放出空间）。虽然HBase包括表格，但是schema仅仅被表格和列簇所要求，列不需要schema。Hbase的表格包括增加/计数功能。

1. 限制

   Hive目前不支持更新操作。另外，由于hive在hadoop上运行批量操作，它需要花费很长的时间，通常是几分钟到几个小时才可以获取到查询的结果。Hive必须提供预先定义好的schema将文件和目录映射到列，并且Hive与ACID不兼容。

HBase查询是通过特定的语言来编写的，这种语言需要重新学习。类SQL的功能可以通过Apache Phonenix实现，但这是以必须提供schema为代价的。另外，**Hbase也并不是兼容所有的ACID特性，虽然它支持某些特性。最后但不是最重要的–为了运行Hbase，Zookeeper是必须的，zookeeper是一个用来进行分布式协调的服务，这些服务包括配置服务，维护元信息和命名空间服务。**

1. 应用场景

Hive适合用来对一段时间内的数据进行分析查询，例如，用来计算趋势或者网站的日志。Hive不应该用来进行实时的查询。因为它需要很长时间才可以返回结果。

Hbase非常适合用来进行大数据的实时查询。Facebook用Hbase进行消息和实时的分析。它也可以用来统计Facebook的连接数。

1. 总结

Hive和Hbase是两种基于Hadoop的不同技术–Hive是一种类SQL的引擎，并且运行MapReduce任务，Hbase是一种在Hadoop之上的NoSQL 的Key/vale数据库。当然，这两种工具是可以同时使用的。就像用Google来搜索，用FaceBook进行社交一样，Hive可以用来进行统计查询，HBase可以用来进行实时查询，数据也可以从Hive写到Hbase，设置再从Hbase写回Hive。