链滴

HBase & Hive

作者：moloee

原文链接：https://ld246.com/article/1517731324933

来源网站：链滴

1. Hive中的表是纯逻辑表，就只是表的定义等，即表的元数据。Hive本身不存储数据，它完全依赖DFS和MapReduce。这样就可以将结构化的数据文件映射为为一张数据库表，并提供完整的SQL查功能，并将SQL语句最终转换为MapReduce任务进行运行。 而HBase表是物理表，适合存放非结构的数据。

2. Hive是基于MapReduce来处理数据,而MapReduce处理数据是基于行的模式；HBase处理数据基于列的而不是基于行的模式，适合海量数据的随机访问。

3. HBase的表是疏松的存储的，因此用户可以给行定义各种不同的列；而Hive表是稠密型，即定义少列，每一行有存储固定列数的数据。

4. Hive使用Hadoop来分析处理数据，而Hadoop系统是批处理系统，因此不能保证处理的低迟延题；而HBase是近实时系统，支持实时查询。

5. Hive不提供row-level的更新，它适用于大量append-only数据集（如日志）的批任务处理。而于HBase的查询，支持和row-level的更新。

6. Hive提供完整的SQL实现，通常被用来做一些基于历史数据的挖掘、分析。而HBase不适用与有jon，多级索引，表关系复杂的应用场景。

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；

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；

uploading...

是的，根据google论文来的，类似的系统还有Cassandra。Google当年设计bigtable的原因在于公内部各业务线需求差异太大，无论从查询性能还是存储schema等，导致没有办法搞一个大招解决所部门的需求。后来还是很吊的Jeffrey一票人设计出来的bigtable。早期google的web页面就存在bigtble里。HBase根据论文，社区的一帮人搞出来的。现在主要的contributor应该是Cloudera和Hortoworks的人。HBase本质上是一个database，可以认为它是一个很大的hashmap。你可以看到HBas很多核心的机制在于它的compaction和split，以及WAL，ragion管理等。而它可以秒级返回，得益hash的设计、bloom filter、memory cache等，但这绝对不是它设计的初衷，只能说是一个考虑点者优化。另外，本质上讲，把Hive和HBase放到一起对比是毫无理由的，这两个系统根本就是完全不的东西，设计目的、架构、生态中的位置都是完全不同的。希望这个回答令你满意。：）

非常感谢详细的回复。我是这么理解的，hbase的目标是解决海量数据的随机查询，key-value、compaction、split、wal、region、memory cache等是围绕这个目标而采用的技术手段。另外，hive和hbase是完全不同的东西我是认同的，在文中也由相关的表述。谢谢，一起讨论！

其实真正为解决adhoc查询的系统是你提到的impala（虽然它现在半死不死）。database的核心是储，访问只是附属品。Anyway，你怎么认为这个系统，你开心就好，我有时候会比较钻牛角尖，勿。

没事，探讨而已，不同思想碰撞一下。impala适合olap多维分析的adhoc场景，但高并发能力不行，base适合单表的清单数据高并发基于某个key的查询，当然现在kylin的OLAP分析底层也是基于hbas来做。