Hive vs HBase

初接触Hadoop技术的朋友肯定会对它体系下寄生的个个开源项目糊涂了，我敢保证Hive,Pig,HBase这些开源技术会把你搞的有些糊涂，不要紧糊涂的不止你一个，如某个菜鸟的帖子的疑问，when to use Hbase and when to use Hive？....请教了^_^没关系这里我帮大家理清每个技术的原理和思路。

Pig

一种操作hadoop的轻量级脚本语言，最初又雅虎公司推出，不过现在正在走下坡路了。当初雅虎自己慢慢退出pig的维护之后将它开源贡献到开源社区由所有爱好者来维护。不过现在还是有些公司在用，不过我认为与其使用pig不如使用hive。：）

Piq是一种数据流语言，用来快速轻松的处理巨大的数据。

Pig包含两个部分：Pig Interface,Pig Latin。

Pig可以非常方便的处理HDFS和HBase的数据，和Hive一样,Pig可以非常高效的处理其需要做的，通过直接操作Pig查询可以节省大量的劳动和时间。当你想在你的数据上做一些转换，并且不想编写MapReduce jobs就可以用Pig.

Hive

不想用程序语言开发MapReduce的朋友比如DB们，熟悉SQL的朋友可以使用Hive开离线的进行数据处理与分析工作。

注意Hive现在适合在离线下进行数据的操作，就是说不适合在挂在真实的生产环境中进行实时的在线查询或操作，因为一个字"慢"。相反

起源于FaceBook,Hive在Hadoop中扮演数据仓库的角色。建立在Hadoop集群的最顶层，对存储在Hadoop群上的数据提供类SQL的接口进行操作。你可以用HiveQL进行select,join,等等操作。

如果你有数据仓库的需求并且你擅长写SQL并且不想写MapReduce jobs就可以用Hive代替。

HBase

HBase作为面向列的数据库运行在HDFS之上，HDFS缺乏随即读写操作，HBase正是为此而出现。HBase以Google BigTable为蓝本，以键值对的形式存储。项目的目标就是快速在主机内数十亿行数据中定位所需的数据并访问它。

HBase是一个数据库，一个NoSql的数据库，像其他数据库一样提供随即读写功能，Hadoop不能满足实时需要，HBase正可以满足。如果你需要实时访问一些数据，就把它存入HBase。

你可以用Hadoop作为静态数据仓库，HBase作为数据存储，放那些进行一些操作会改变的数据。

Pig VS Hive

Hive更适合于数据仓库的任务，Hive主要用于静态的结构以及需要经常分析的工作。Hive与SQL相似促使 其成为Hadoop与其他BI工具结合的理想交集。

Pig赋予开发人员在大数据集领域更多的灵活性，并允许开发简洁的脚本用于转换数据流以便嵌入到较大的 应用程序。

Pig相比Hive相对轻量，它主要的优势是相比于直接使用Hadoop Java APIs可大幅削减代码量。正因为如此，Pig仍然是吸引大量的软件开发人员。

Hive和Pig都可以与HBase组合使用，Hive和Pig还为HBase提供了高层语言支持，使得在HBase上进行数据统计处理变的非常简单

Hive VS HBase

Hive是建立在Hadoop之上为了减少MapReduce jobs编写工作的批处理系统，HBase是为了支持弥补Hadoop对实时操作的缺陷的项目 。

想象你在操作RMDB数据库，如果是全表扫描，就用Hive+Hadoop,如果是索引访问，就用HBase+Hadoop 。

Hive query就是MapReduce jobs可以从5分钟到数小时不止，HBase是非常高效的，肯定比Hive高效的多。