黑客派

整体了解HADOOP框架及一些开源项目

作者：unhappydepig

原文链接：https://hacpai.com/article/1464779414782

来源网站：黑客派

Hadoop框架中，有很多优秀的工具，帮助我们解决工作中的问题。

<h2 id="toc_h2_0">Hadoop的位置</h2>

从上图可以看出，越往右，实时性越高，越往上，涉及到算法等越多。
越往上，越往右就越火……

<h2 id="toc_h2_1">Hadoop框架中一些简介</h2>

<h2 id="toc_h2_2">HDFS</h2>
HDFS，（Hadoop Distributed File System） hadoop分布式文件系统。在Google开源有关DF的论文后，由一位大牛开发而成。HDFS的建立在集群之上，适合PB级大量数据的存储，扩展性强，错性高。它也是Hadoop集群的基础，大部分内容都存在了HDFS上。

<h2 id="toc_h2_3">MapReduce</h2>
MapReduce，是Hadoop中的计算框架，由两部分构成。Map操作以及Reduce操作。MapRed ce，会生成计算的任务，分配到各个节点上，执行计算。这样就避免了移动集群上面的数据。而且其部，也有容错的功能。在计算过程中，某个节点宕掉之后，会有策略进行应对。Hadoop集群，上层一些工具，比如Hive或者Pig等，都会转换为基本的MapReduce任务来执行。

<h2 id="toc_h2_4">HBase</h2>
HBase源自谷歌的BigTable。HBase是面向列存储的数据库，性能高，扩展性强，可靠性高。HB se的内容，存储在HDFS上，当然它也可以使用其他的文件系统，如S3等。HBase作为一个顶级项目使用频率很高。如：我们可以用来存储，爬虫爬来的网页的信息等。具体的HBase的概念请见后续详说明。延迟较低。

<h2 id="toc_h2_5">Hive</h2>
Hive，是一个查询的工具，在HBase中，对于SQL的支持不太好。而Hive解决了这一类的问题。sql形式操作hbase，更爽一些。Hive编写的一些sql语句，其实最后也还是会变成MapReduce程序当然这种查询，不能与关系型数据库mysql等比较，hive查询时，是秒级或分钟级的，时间比较长。

<h2 id="toc_h2_6">Sqoop</h2>
Sqoop，也是一个很神奇的数据同步工具。在关系型数据库中，我们会遇到一种情景，将Oracle据导入到Mysql，或者将Mysql数据，导入到Oracle。那其实Sqoop也是类似的功能。sqoop可以将O acle，Mysql等关系型数据库中的数据，导入到HBase，HDFS上，当然也可以从HDFS或HBase导入Mysql或Oracle上。

<h2 id="toc_h2_7">Flume</h2>
Flume，是日志收集工具，是分布式的，可靠的，容错的，可以定制的。应用场景如：100台服

器，需要监测各个服务器的运行情况，这时可以用flume将各个服务器的日志，收集过来。Flume也两个版本。Flume OG 和Flume NG。现在基本都用NG了。

<h2 id="toc_h2_8">Impala</h2>
Impala是Cloudera公司主导开发的新型查询系统，它提供SQL语义，能查询存储在Hadoop的HFS和HBase中的PB级大数据。已有的Hive系统虽然也提供了SQL语义，但由于Hive底层执行使用的MapReduce引擎，仍然是一个批处理过程，难以满足查询的交互性。相比之下，Impala的最大特点是最大卖点就是它的快速。Imapa可以和Phoenix，Spark Sql联系起来了解一下。

<h2 id="toc_h2_9">Spark</h2>
Spark是一个内存计算的框架。目前一个大的趋势。MapReduce会有很大的IO操作，而Spark是内存中计算。速度是Hadoop的10倍（官网上这样说的）。Spark是目前一个趋势，是需要了解的。

<h2 id="toc_h2_10">Zookeeper</h2>
Zookeeper，动物管理员。Zookeeper叫分布式协作服务。作用主要是，统一命名，状态同步集群管理，配置同步。Zookeeper在HBase，以及Hadoop2.x中，都有用到。

<h2 id="toc_h2_11">Mahout</h2>
数据挖掘算法库，里面内置了大量的算法。可以用来做预测、分类、聚类等。工具很强大，但是术要求能力较高。

<h2 id="toc_h2_12">Pig</h2>
和Hive类似。具体区别自己搜搜。Pig可以构建数据仓库。可用来对数据仓库中数据，进行查询析。Pig也有自己的查询语法，很不幸，不是sql形式，Pig Latin。

<h2 id="toc_h2_13">Ambari</h2>
Ambari是一个管理平台。可以对集群进行统一的部署。也是很方便的。