hadoop版本

1、Apache

官方版本

2、Cloudera（CDH）

使用下载最多的版本，稳定，有商业支持，在Apache的基础上打上了一些patch。推荐使用。

3、HDP（Hortonworks Data Platform）

Hortonworks公司发行版本

hadoop各个节点介绍：

NameNode：这是Hadoop的守护进程（是JVM的进程）。负责记录文件是如何分割成数据块，以及这些数据块分别存储到哪些数据节点上。对内存进行集中管理。NameNode在整个hadoop中只有一个。一旦NameNode服务器宕机，整个系统将无法运行。

DataNode：集群中的每个从服务器都运行一个DataNode后台服务。这个后台程序负责将HDFS数据块写到本地的文件系统。

SecondaryNameNode：用来监控HDFS状态的辅助后台程序。如保存NameNode的快照。

JobTracker：用户连接应用程序和hadoop。每一个hadoop集群中只有一个。它一般运行在Master节点上。

TaskTracker：负责与DataTracker进行结合。

hadoop的三大核心：

1、HDFS：Hadoop Distributed File System，分布式文件系统。

2、MapReduce：并行计算框架。

3、TARN：Yet Another Resource Negotiator，资源管理调度系统。

谷歌云计算平台技术架构

分布式文件系统Google Distributed File System

并行数据处理MapReduce

分布式锁Chubby

结构化数据表BigTable

GFS -- HDFS

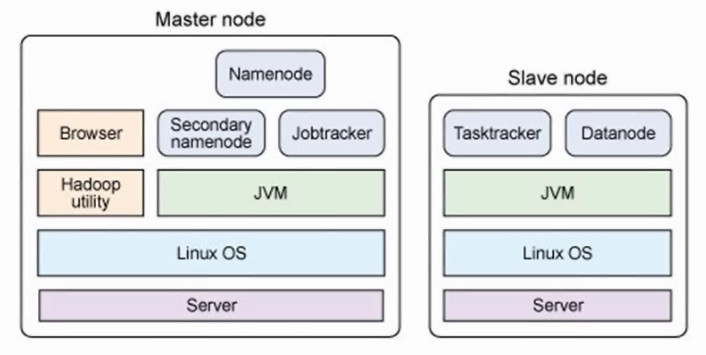
BigTable-Hbase

Chubby

谷歌云计算应用

MapReduce

hadoop的集群架构



NameNode的工作原理！！！

hdfs适用于一次写入多次查询 的情况。不支持并发写的情况，小文件不适合。

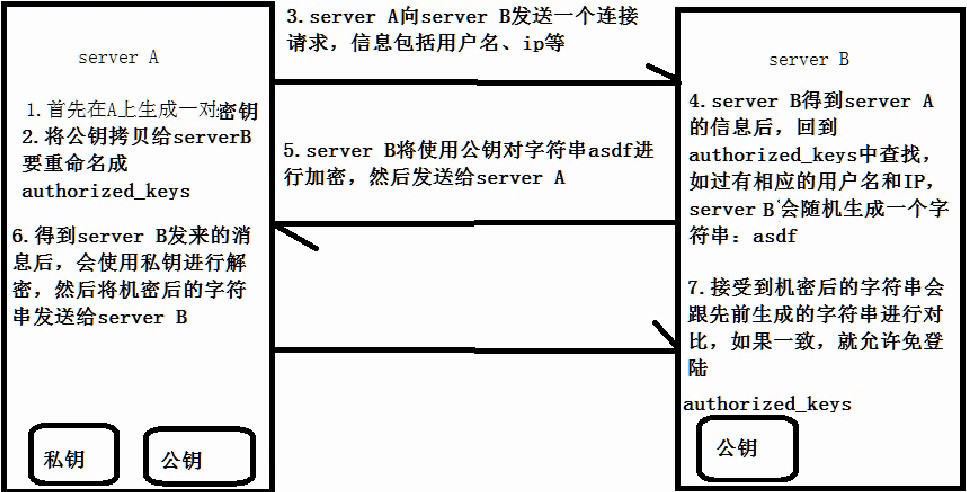
hadoop的linux下的命令行操作。

这些命令号和linux的命令行几乎一样，只不过是在前面加入上了“hadoop fs”。

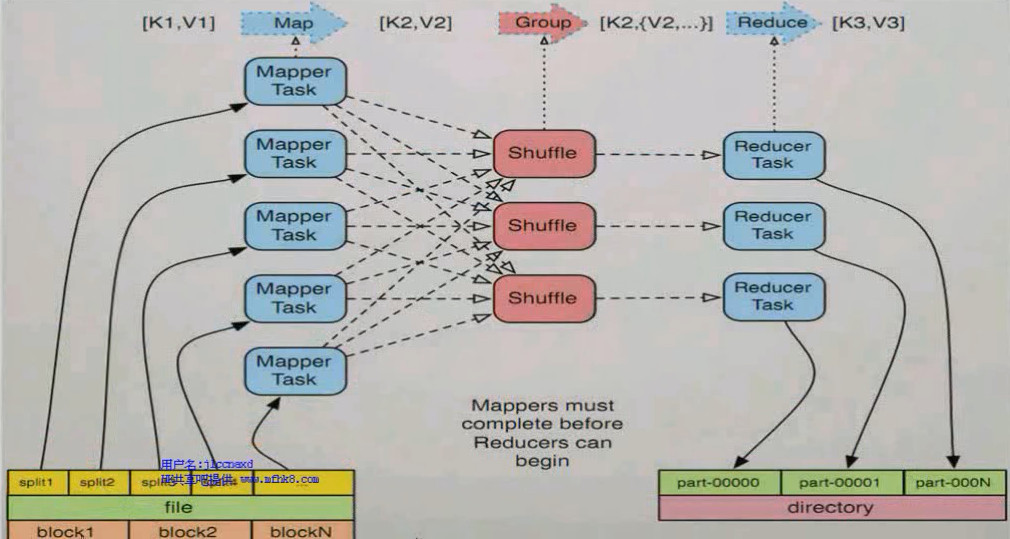
hadoop fs -ls hafs://192.168.58.88:9000/相当于

hadoop fs -ls /

ssh登入过程分析：



MapReduce原理



◆执行步骤：

1. map任务处理

1.1 读取输入文件内容，解析成key、value对。对输入文件的每一行，解析成key、value对。每一个键值对调用一次map函数。

1.2 写自己的逻辑，对输入的key、value处理，转换成新的key、value输出。

1.3 对输出的key、value进行分区。

1.4 对不同分区的数据，按照key进行排序、分组。相同key的value放到一个集合中。

1.5 (可选)分组后的数据进行归约。

2.reduce任务处理

2.1 对多个map任务的输出，按照不同的分区，通过网络copy到不同的reduce节点。

2.2 对多个map任务的输出进行合并、排序。写reduce函数自己的逻辑，对输入的key、value处理，转换成新的key、value输出。

2.3 把reduce的输出保存到文件中。