1.hadoop:

lucene（全文检索） -->Nutch(搜索引擎)

> GFS(谷歌文件系统) --->HDFS hadoop分布式文件系统
> MapReduce(数据的处理算法) -->MapReduce 分布式计算
> bigtable(列式数据库) --->HBase

主要包括:

1.分布式存储系统:HDFS

2.资源管理系统 :YARN -->负责集群资源的统一管理和调度

((hadoop2.0才有))

3.分布式计算框架: MapReduce

元数据:

对数据信息描述的数据(比如,某个资源在哪个块,起始地址是什么,大小为多少,)

namenode 持久化 有两个文件:fsimage,edits

SecondaryNameNode 合并fsimage,edits两个文件,加快启动速度

DN(dateNote) 和NN(namenode)保持心跳机制

副本放置策略:

1.找负载比较低的

2.放在不与第一个副本同一个地方

3.放在第二个副本同一个机架

hadoop作业流程:

1.客户向ResourceManager提交作业

2.ResourceManager的ApplicationManager通知一个NodeManager启动contrainer并在contrainer中启动ApplicationMaster负责这次作业

3.ApplicationMaster向ApplicationManager注册

4.ApplicationMaster向ResourceManager的ResourceSchedule轮询申请资源

5.申请到资源后,通知相应的NodeManager启动作业

6.NodeManager启动contrainer并执行相应的Map/Reduce Task

7.执行的Task向ApplicationMaster汇报作业情况

8.作业执行完成后,Application想ApplicationMaster注销作业

注:namenode: ResourceManager -> ResourceSchedule ,
ApplicationManager ;

datanode : NodeManager ; ApplicationManager ;
MapTask ;ReduceTask