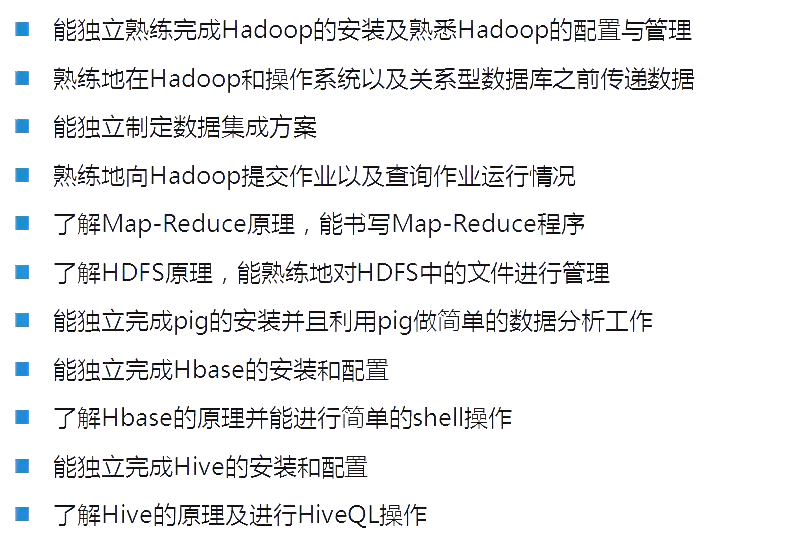
Hadoop笔记

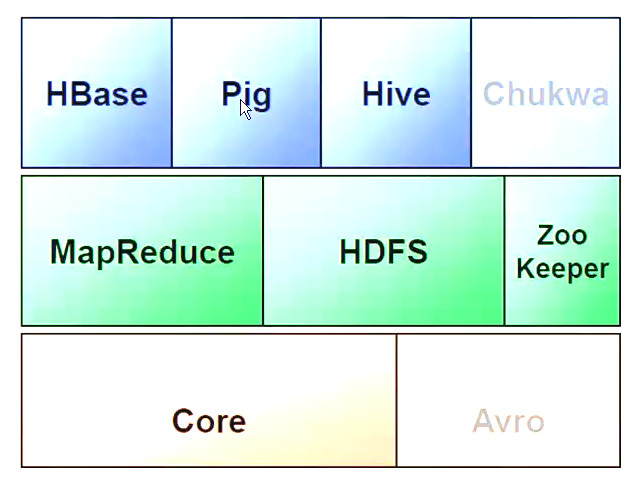
# Hadoop的技能



# Hadoop的特点

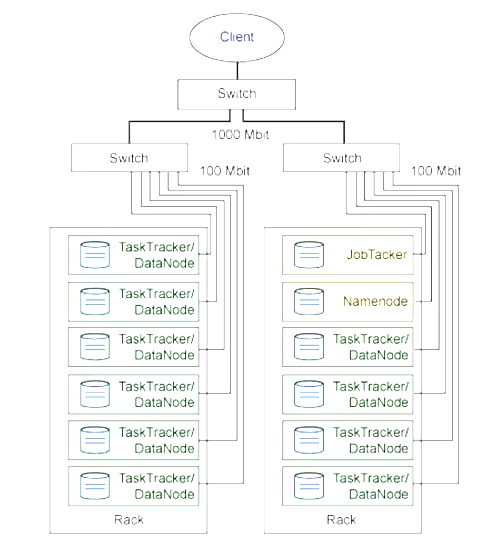
扩容能力、成本低、高效率、可靠性

# Hadoop家族



# Hadoop的架构

## Hadoop集群的物理分布

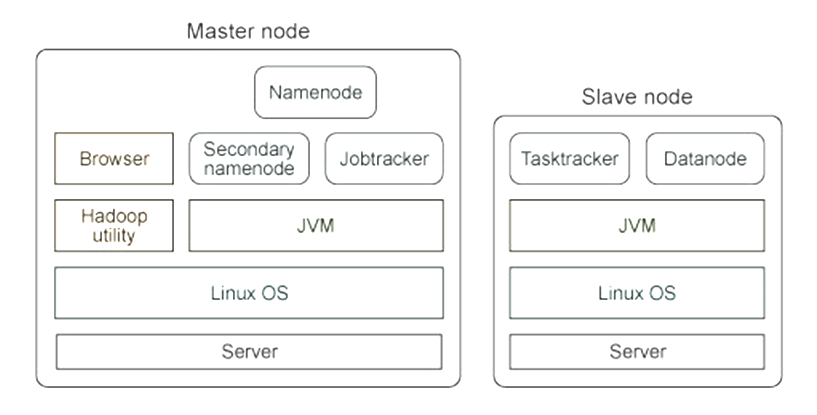


Master：NameNode、SecondaryNameNode、JobTracker

Slave：TaskTracker、DataNode

Master不是唯一的

## 单结点物理结构



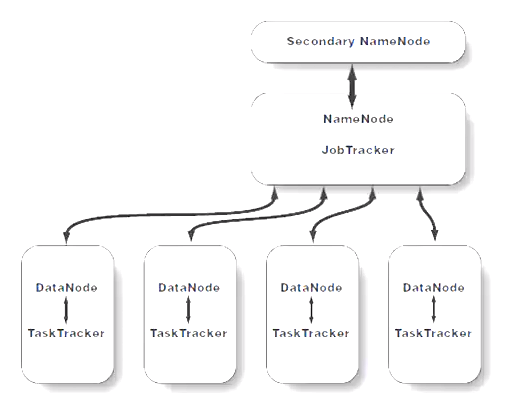
# MapReduce的架构

主从结构

主节点，一个：JobTracker

从节点，多个：TaskTracker

## JobTracker



用于处理作业的后台程序

接收客户提交的计算任务

把任务分配给TaskTracker执行

决定有哪些文件参与处理，然后切割task并分配节点

监控tasktracker，重启失败的task

每个集群只有一个JobTracker，位于master节点

## TaskTracker

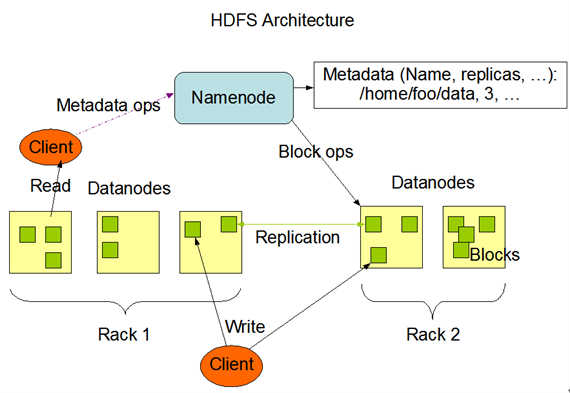
位于slave节点，与datanode结合

管理各自节点的task（JobTracker分配）

每个节点只有一个tasktracker，但一个tasktracker可以启动多个jvm，用于并行执行map或reduce任务

与jobtracker交互，执行JobTracker分配的计算任务

**# HDFS体系结构**



**HDFS的架构**

主从结构

**## NameNode**

HDFS的守护程序

对内存和IO进行集中管理

接收用户

提供名称查询服务，是一个jetty服务器。

保存的metadata信息（文件owership和permission、跟踪文件如何被分割成文件块，文件块又被哪些节点存储）

NameNode是单点

**metadata**存储到**fsimage**



## DataNode:

保存块

HDFS数据块读写到本地文件系统

启动DN线程时向NN汇报block信息

向NN发送心跳保持联系

Block的副本放置策略：

第一个副本，放置在本地机架的节点上

第二个副本，第二个副本放在同一机架的另一个节点上

第三个副本，放置在不同机架的节点上

Block默认64MB，不够默认仍占用一个block的命名空间，但物理上并不会。

Block大小和副本数由client端上传文件时hdfs设置，其中副本数可以变更，block不可以上传后变更。

## 数据损坏

DN读取block的时候，会计算checksum

## SecondaryNameNode（SNN）

它不是NN的热备份

它可以作为冷备份：

将本地保存的fsimage导入

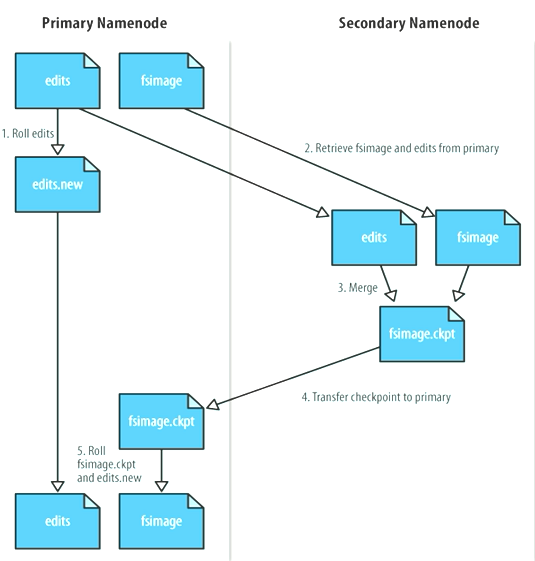
修改cluster所有DN的NameNode地址

修改所有client端NameNode地址

或者修改SNN IP为原NN IP

它的工作是帮助NN合并edits，减少NN启动时间

## SecondaryNameNode

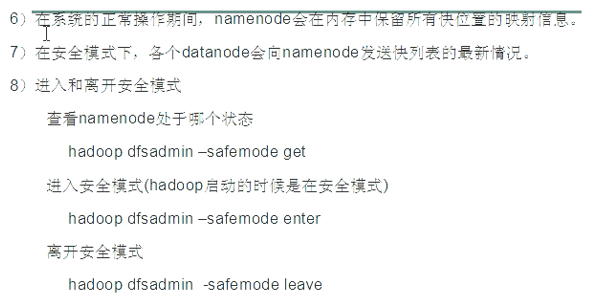


## HDFS权限

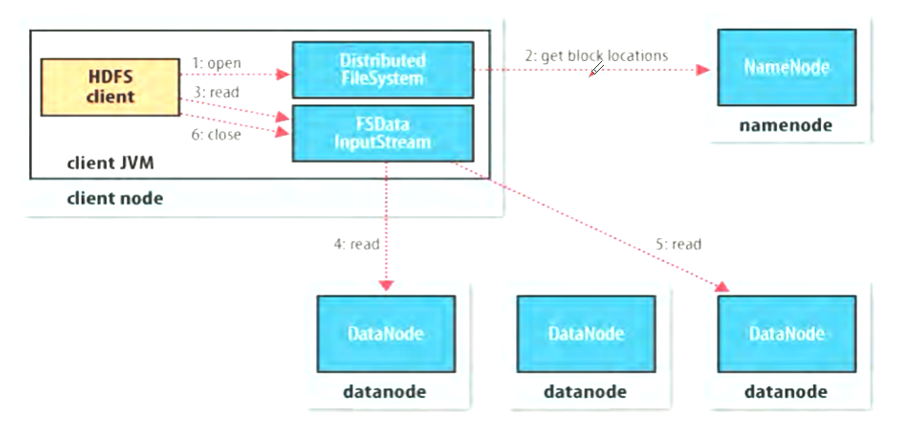
rwx权限

## 安全模式

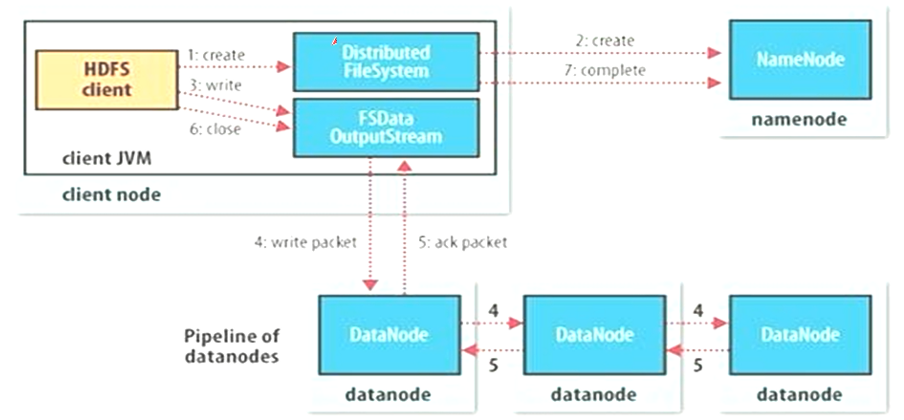




## HDFS文件读取



## HDFS文件写入



## HDFS开发常用命令

列出hdfs下的文件

dfs –ls

列出HDFS目录下某个文档中的文件

dfs –ls in (HDFS下名为in的文档中的文件)

创建文件夹

hadoop fs –mkdir /usr/example

上传文件

hadoop fs –put test test1 (目录下的test1文件上传到HDFS上并重命名为test)

将HDFS中的文件复制到本地系统中

hadoop dfs -get in getin  (将HDFS中的in文件复制到本地系统并命名为getin)

删除HDFS下的文档

hadoop dfs -rmr out

查看HDFS下的某个文件

hadoop dfs -cat in/\*

管理与更新

hadoop dfsadmin -report

安全模式

退出 hadoop dfsadmin -safemode leave

进入 hadoop dfsadmin -safemode enter

# 安装和配置

虚拟机

ubuntu 14.04

jdk1.8.0\_111

hadoop-1.1.2.tar.gz

## 伪分布式安装

**关闭防火墙** （主机和虚拟机ufw disable）

**修改ip （桥接模式 静态IP）**

**修改hostname**

**设置ssh自动登录 （虚拟机ubuntu安装openssh）**

**安装jdk**

**安装hadoop**

## vmware的三种网络方式

Vmnet0 桥接

Vmnet1 主机模式

Vmnet8 NAT模式

Vmware虚拟机三种网络模式详解

*地址：http://blog.csdn.net/noob\_f/article/details/51099040*

桥接模式Vmnet0

什么是桥接模式？桥接模式就是将主机网卡与虚拟机虚拟的网卡利用虚拟网桥进行通信。在桥接的作用下，类似于把物理主机虚拟为一个交换机，所有桥接设置的虚拟机连接到这个交换机的一个接口上，物理主机也同样插在这个交换机当中，所以所有桥接下的网卡与网卡都是交换模式的，相互可以访问而不干扰。在桥接模式下，虚拟机ip地址需要与主机在同一个网段，如果需要联网，则网关与DNS需要与主机网卡一致。

适用： 桥接模式配置简单，但如果你的网络环境是ip资源很缺少或对ip管理比较严格的话，那桥接模式就不太适用了。

NAT地址转换模式（Vmnet8）

如果你的网络ip资源紧缺，但是你又希望你的虚拟机能够联网，这时候NAT模式是最好的选择。NAT模式借助虚拟NAT设备和虚拟DHCP服务器，使得虚拟机可以联网。

主机模式（VMnet1）

Host-Only模式其实就是NAT模式去除了虚拟NAT设备，然后使用VMware Network Adapter VMnet1虚拟网卡连接VMnet1虚拟交换机来与虚拟机通信的，Host-Only模式将虚拟机与外网隔开，使得虚拟机成为一个独立的系统，只与主机相互通讯。

其他

关闭虚拟机防火墙 ufw disable

ubuntu重启网卡

先关闭接口:sudo ifconfig eth0 down

然后打开:sudo ifconfig eth0 up

###

Vmware虚拟机-Ubuntu14.04

IP地址：192.168.3.117

账户Saber密码1992

账户root密码root