Hadoop安装流程:

1. 准备N台虚拟机(关闭防火墙，设置静态IP，设置主机名称等)
   1. 查看/etc/udev/rules.d/70-persistent-net.rules,复制address码
   2. 修改/etc/sysconfig/network-scripts/ifcfg-eth0中的HWADDR，并修改其IPADDR为制定ip
   3. 修改/etc/sysconfig/network中的主机名称
2. 安装JDK
3. 配置环境变量
4. 安装hadoop
5. 配置环境变量
6. 配置hadoop集群
7. 单点启动
8. 配置ssh
9. 集群启动并测试集群
   1. 在hadoop的etc/hadoop文件夹下新建slaves文件，输入datanode的节点地址或主机名
10. 集群时间同步（主机）
    1. 检查ntp是否安装
    2. 修改ntp配置(/etc/ntp.conf)
       1. 修改1：授权相同网段上的机器都可以查询及同步这台机器上的时间，即将#restrict 192.168.1.0 mask 255.255.255.0 nomodify notrap的#去掉
       2. 修改2：集群在局域网中，不使用互联网上的时间，即注释掉联网服务器
       3. 添加3：当节点断网时，依然可以采用本地时间为其他服务器作为标准。添加

server 127.127.1.0

fudge 127.127.1.0 stratum

* 1. 修改/etc/sysconfig/ntpd，让硬件时间一起同步，添加SYNC\_HWCLOCK=YES
  2. 重新启动ntpd
  3. 设置ntpd开机启动，chkconfig ntpd on

1. 集群时间同步（其他机器）
   1. 每十分钟同步一次(同步任务crontab -e)
      1. \*/10 \* \* \* \* /usr/sbin/ntpdate hostname

Xsync脚本:

#! /bin/bash

pcount=$#

if((pcount==0));then

echo no args

exit

fi

#2 获取文件名称

p1=$1

fname=`basename $p1`

echo fname=$fname

#3 获取文件路径

fdir=`cd -P $(dirname $p1); pwd`

echo fdir=$fdir

#4 获取当前用户名

user=`whoami`

for((host=1;host<4;host++));do

echo ------------------------enpcszhnnt0$host----------------------

rsync -r $fdir/$fname $user@enpcszhnnt0$host:$fdir/$fname

done

hdfs部分指令:

1. 刷新节点：hdfs dfsadmin -refreshNodes
2. 集群间数据拷贝：hdfs distcp source dest
3. 将小文件合并成har文件：hadoop archive -archiveName input.har -p source dest
4. 快照：hdfs dfsadmin -allowSnapshot 路径

hdfs dfsadmin -disallowSnapshot 路径

hdfs dfs -createSnapshot 路径

hdfs lsSnapshottableDir

hdfs snapshotDiff <snapshotDir> <from> <to>

hdfs dfs -deleteSnapshot <path> <snapshotname>

hdfs dfs -cp <snapshotName> <Dir>

MapReduce介绍:

包含进程——MapTask，ReduceTask，MrAppMaster

Map阶段：

1. 读数据，并按行处理
2. 通过分隔符，划分单词
3. 生成（k，v）键值对
4. 将所有KV键值对中的单词，按照单词首字母分成两个分区溢写到磁盘里

Reduce阶段(依赖于Map阶段的结果，并行)：