## 1 修改本地hosts文件

web端查看HDFS文件系统 <http://hadoop100:50070>

不能查看，需要修改本地hosts文件，在如下路径：C:\Windows\System32\drivers\etc

修改本地hosts文件，添加192.168.60.100 hadoop100

保存没有权限，另存为，然后替换即可

## 2 格式化namenode

**格式化NameNode**（第一次启动时格式化，以后就不要总格式化）

[atguigu@hadoop101 hadoop-2.7.2]$ hadoop namenode -format

注意：格式化NameNode，会产生新的集群id,导致NameNode和DataNode的集群id不一致，集群找不到已往数据。所以，格式NameNode时，一定要先删除hadoop-2.7.2/data数据和log日志，然后再格式化NameNode。

## 3 xshell鼠标右键修改

点击 xshell 中的工具，选项，鼠标和键盘在相应的位置做修改；

## 4 集群注意事项

（1）hdfs-site.xml 中 50070 是dfs.http.address NameNode web管理端口，因此可以在浏览器中输入 <http://hadoop102:50070访问，注意hadoop102是namenode所在的节点；50090>是备份名称节点的http协议访问地址与端口，因此可以使用备份节点加上端口查看；

（2）在查看wordcount执行结果的时候，也需要在namenode所在的节点执行

**hadoop fs -cat /wc/part-r-00000；**

（3）日志的开启以及收集功能的配置都在RM所在节点配置，然后执行脚本分发，保证每个节点的配置文件都是一样的，然后重新启用yarn节点，在日志所在节点执行；

1. data/tmp/dfs/ 目录下分成不同的文件夹，存放不同节点数据信息；

（5）hadoop fs -D dfs.replication=2 -put /home/atguigu/hello.txt /

// 上传文件的同时指定副本数

（6）hdfs dfsadmin -rollEdits // Rolls the edit log 滚动编辑日志

（7）文件归档：hadoop archive -archiveName test.har -p /hello test\* hi\* /

hadoop fs -get har:///test2.har/hi1 hadoop fs -ls har:///test2.har

（8）MR细节：移动计算而不是移动数据;

Mapreduce程序输入文本格式为utf-8无pom；

排序的过程只会根据key排序；

分区，快排，合并，溢写；

一个reducetask负责一个分区，每个分区单独有序，整体没有顺序；reducer拿一组数据【封装迭代器相同key的为一组】 调用一次reduce；

shuffle很大程度上就是为了排序，需求不涉及排序，不需要reduce，不要shuffle；combine只有在溢写之前调用；

## 5 hive语句注意点

（1）collect\_set(col), collect\_list(col) 可以作用与group以后将每个组的所有结果放在集合里面；

（2）hive错误的详细提示，在hive.log文件中，可以查看；

（3）select age,count(\*) count\_age from people group by age having count\_age>10，hiving语句可以使用count\_age，说明having在select之前执行，MySQL底层有优化器，将count\_age改为count(\*)，在严格模式不能正确执行；

## 6 flume注意点

会不会丢数据？看使用的异步source类型如果没有缓存机制，如果channel满了，可能会丢数据。