**MapReduce分析与优化**

**(50 / 50分)**

MapReduce程序效率的瓶颈在于两点：  
1.计算机性能

CPU、内存、磁盘、网络

2.I/O操作优化

数据倾斜

Map和Reduce数设置不合理

Map运行时间太长，导致Reduce等待过久

小文件过多

大量的不可分块的超大文件

Spill次数过多

Merge次数过多

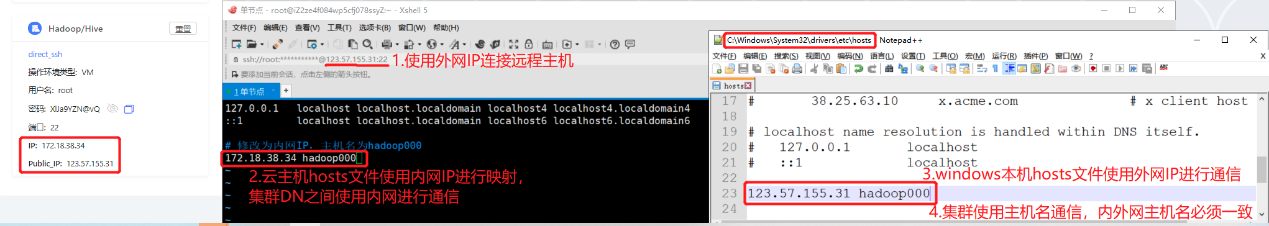
可参考方法：

减少溢写（spill）次数：通过调整mapreduce.task.io.sort.mb及mapreduce.map.sort.spill.percent参数值（内存缓冲区大小为100MB，达到80%就开始溢出到本地磁盘文件中），增大触发spill的内存上限，减少spill次数，从而减少磁盘I/O次数。

减少合并（merge）次数：通过调整mapreduce.task.io.sort.factor参数，增大merge时同时打开的文件数目，减少merge的次数，从而缩短MapReduce处理时间。

**1.本地开发工具连接Hadoop集群**

**1.本机映射名为hadoop000，云主机Hadoop/Hive的hosts文件中IP需要修改为内网IP，需要修改为实际内网IP（服务器地址有内外网之分），本地eclipse所在主机需要使用外网IP设置映射；**  
Linux/Mac系统下文件地址：/etc/hosts  
Windows系统下文件地址：C:\Windows\System32\drivers\etc\hosts



**2.云主机搭建的Hadoop集群，集群之间通过内网通信，本地eclipse开发工具需要使用域名进行访问。**

Hadoop配置文件是以内网IP作为机器间通信的IP。在这种情况下,我们能够访问到namenode机器，namenode会给我们数据所在机器的IP地址供我们访问数据传输服务，但是当写数据的时候，NameNode和DataNode是通过内网通信的，返回的是datanode内网的IP,我们无法根据该IP访问datanode服务器。将默认的通过IP访问，改为通过域名方式访问。

// 使用hdfs的fs功能，客户端就会访问core-site.xml配置文件

// 设置客户端访问datanode使用hostname来进行访问

**conf**.set("dfs.client.use.datanode.hostname", "true");

// 设置core-site.xml中的属性fs.defaultFS和属性值，注意主机名必须和设置的hosts主机名一致

**conf**.set("fs.defaultFS","hdfs://hadoop000:9000");

**MapReduce分析与优化**

**(50 / 50 分)**

**项目背景**

目前我国是世界上最大的移动通信市场，消费者购买力也保持上升趋势，其市场潜力巨大。本项目针对某商城手机销售数据，采用Hadoop技术进行分析预测手机市场各品牌形象力，从而为消费者和运营商提供客观可靠的技术性参考。

**数据说明**

项目数据来源于某商城手机销售日志数据，由品牌、手机名称、价格、销量、好评率、购买链接六个属性组成，数据间以逗号进行分割。

# 数据路径：/root/data/mobile.txt

品牌,手机名称,价格,销量,好评率,购买链接

Apple,商品名称：AppleiPhone 12,6799,301534,93,http://item.jd.com/100009077475.html

vivo,商品名称：vivoiQOO Neo3,2998,322076,94,http://item.jd.com/100012820028.html

**需求说明**

针对手机日志数据，对同一品牌手机进行相关分析，计算同一品牌：

* 手机总销售量；（同一品牌手机价格\*销量）
* 总销售额；（同一品牌手机总销售额）
* 平均好评率。（同一品牌手机平均好评率=同一品牌手机总好评率/同一品牌手机总数，结果取整数）

**示例数据**

21KE 8193 1558867 95

360 11166 2222034 92

考核条件如下 :

1.

将数据/root/data/mobile.txt上传至HDFS根目录下

(10/ 10分)

操作环境：

Hadoop- hive

2.

根据步骤说明，编写程序，结果保存至HDFS文件系统/mobile目录下（结果使用默认文件名，不做改动）

(20/ 20分)

操作环境：

Hadoop- hive

3.

将HDFS上结果文件保存至本地/root/data目录下

(20/ 20分)

操作环境：

Hadoop- hive、Hadoop- hive