**学习心得**

18183023 谭智元 6.26

**今天学习的内容：**

1.MapReduce应用开发：

**MapReduce执行的基本步骤：**

1）Map端工作：

1.1 读取要操作的文件–这步会将文件的内容格式化成键值对的形式，键为每一行的起始位置偏移，值为每一行的内容。

1.2 调用map进行处理–在这步使用自定义的Mapper类来实现自己的逻辑，输入的数据为1.1格式化的键值对，输入的数据也是键值对的形式。

1.3 对map的处理结果进行分区–map处理完毕之后可以根据自己的业务需求来对键值对进行分区处理，比如，将类型不同的结果保存在不同的文件中等。这里设置几个分区，后面就会有对应的几个Reducer来处理相应分区中的内容。

1.4 分区之后，对每个分区的数据进行排序，分组–排序按照从小到大进行排列，排序完毕之后，会将键值对中，key相同的选项 的value进行合并。如，所有的键值对中，可能存在

hello 1

hello 1

key都是hello，进行合并之后变成

hello 2

可以根据自己的业务需求对排序和合并的处理进行干涉和实现。

1.5 归约（combiner）–简单的说就是在map端进行一次reduce处理，但是和真正的reduce处理不同之处在于：combiner只能处理本地数据，不能跨网络处理。通过map端的combiner处理可以减少输出的数据，因为数据都是通过网络传输的，其目的是为了减轻网络传输的压力和后边reduce的工作量。并不能取代reduce。

2）Reduce端工作：

2.1 通过网络将数据copy到各个reduce。

2.2 调用reduce进行处理–reduce接收的数据是整个map端处理完毕之后的键值对，输出的也是键值对的集合，是最终的结果。

2.3 将结果输出到hdfs文件系统的路径中。

**开发流程：**

1.本地程序开发测试：从海量数据中抽取小部分数据到本地，使用IDE等工具进行开发，并在测试数据集上进行程序运行、逻辑等测试

2.集群环境运行：本地测试通过后，就可以将代码运行在海量数据中，这时候90%的小bug已经得到修复了，剩余要解决的就是生产环境中的问题了

3.调优：程序能够在集群中运行起来并不代表成功，通过一些集群、程序调优方式可以让你的代码跑的更好、更快

2.Hadoop的集群：

进行了理论学习，发现步骤在原来的单机环境的步骤上多出了以下几个步骤：

1）修改HOSTS配置：在/etc/hosts中添加集群所有主机的 IP 和HostName的对应关系，IP 与主机名之间使用一个 TAB 键分隔，多个主机名之间同样使用 TAB 键进行分隔。

2）禁用SELinux（root权限）：为了避免由于SELinux安全问题导致安装失败，暂时先将其关闭，以后再根据需要进行开启。使用的命令如下：

setenforce 0（将 SELinux 的状态临时设置为 Permissive 模式（立即生效））

3）安装Zookeeper（感觉非必须）：由于两个NameNode需要使用Zookeeper来实现故障时主备自动切换，所以在这里需要先提前安装并启动Zookeeper。（如果不使用两个NameNode：Active和Standby是否可以不需要，存疑）

4）将认证文件复制到其他主机，并设置认证文件访问权限

5）修改配置：修改core-site.xml、hdfs-site.xml、mapred-site.xml、yarn-site.xml、workers、hadoop-env.sh文件（要修改比单机环境多了许多）。配置文件全部修改完成之后，将修改后的配置分发到集群的其他节点上。

6）启动Hadoop集群：

a)启动JournalNode：

首先，需要逐个启动三个JournalNode节点，使用jps命令查看是否启动成功。

b)格式化Active NameNode：

第一次启动HDFS需要先进行格式化 NameNode，在 NN1上执行格式化命令。

c)启动Active NameNode ：

格式化完成之后，在 NN1上执行如下命令：

hdfs --daemon start namenode

d)同步数据到Standby NameNode：

在 NN2上执行如下命令，将nn1的元数据信息同步到NN2：

hdfs namenode -bootstrapStandby

e)启动其余节点：

start-dfs.sh

f)启动ZKFC ：

在NN1 上执行如下操作，先在ZooKeeper上创建一个znode用来存储自动故障转移相关数据，再启动 ZKFC 守护进程。

hdfs zkfc -formatZK

hdfs --daemon start zkfc

g)启动YARN

在NN1上执行如下命令启动 Yarn 。

**反思：**

1. 在浏览大量资料时发现，其进行java代码编写的时候似乎都是使用IDE进行开发，并不是使用终端开发。在浏览之后发现，确实有相关插件（基于Eclipse）来实现IDE的MapReduce开发。但是这又引出我的疑问：是否需要在虚拟机上装载eclipse以便于开发？还是直接基于本地Eclipse引入插件包进行开发后将.jar包导回虚拟机进行使用？如果是这样那么需要在本地机上再重新配置Hadoop环境才能使用？基于以上种种疑问，并不敢贸然进行环境的搭建，需要等待疑问由老师解决之后再进行进一步操作。

2.由于没有通知互信使得Hadoop集群的搭建没有在线上如期合作，之后将进行尝试