暑期实训学习报告

1813012　柳鑫

一．学习内容

① MapReduce 特性

② Hadoop 集群搭建

二．学习结果

① MapReduce 特性

1. 计数器

（1）MapReduce 包含的高级特性，计数器，数据集的排序和连接。

（2）作用：收集作业统计信息，质量控制或者应用级统计，辅助诊断系统故障。

（3）分组：MapReduce 任务计数器、文件系统计数器、fileinputformat 计数器、fileoutputformat 计数器、作业计数器。各组要么包含任务计数器，要么包含作业计数器。

（4）任务计数器，任务执行活成中采集任务相关信息，每个作业所有任务结果会被聚集起来，例如 map_input_records。

（5）任务计数器每次传输给 master 都是完成的传输，而非自上次传输之后的计数值，避免消息丢失引发错误，任务执行期间失败，相关计数器值会减小。

（6）作业计数器由 master 维护，无需网络间传输数据。

（7）Java 可以自定义计数器，如，数据不规范记录计数器。

2. 排序

（1）排序是 MapReduce 的核心计数，尽管应用本身可能不需要排序，但是仍可能使用 MapReduce 排序功能组织数据。

（2）部分排序、全排序、辅助排序。

3. 连接

（1）MapReduce 能执行大型数据间的连接操作，如果由 mapper 连接，则是 mapper 端连接，如果由 reducer 连接，则称为 reduce 端连接。

（2）map 端连接：map 函数执行执行连接，各个 map 输入数据必须先分区并且以特定方式排序。各个输入数据集会被划分成相同数量的分区，并且按照相同的连接键排序。同一键的所有记录均会放在同一分区之中。

（3）reduce 端连接：由于 reduce 端连接并不要求输入数据集符合特定结构，因为更为常用。因为需要经过 shuffle，所以效率会低一些。mapper 为各个记录标记源，并使用连接键作为 map 输出键，相同键的记录放在同一个 reducer 中。

② Hadoop 集群搭建

1. 首先检查每个节点是否已安装 java，并且配置环境变量以及免密操作 ssh。生成公钥私钥，然后将公钥拷贝到 authorized_keys 中。

2. 上传 hadoop 压缩包到 linux 系统并解压。

3. 在/etc/profile 或则~/.bash_profile 文件中配置 hadoop 的环境变量。

4. 分别在 hadoop-env.sh、mapred-env.sh、yarn-env.sh 中配置 JAVA_HOME。

5. 修改 core-site.xml 文件。

6. 修改 hdfs-site.xml 文件。

7. 修改 works 文件。

8. 将配置完整的 hadoop 拷贝到其他节点。

9. 启动集群：

（1）启动 journalnode 节点。

（2）格式化 namenode。

（3）开启 zookeeper 节点。

（4）开启 hdfs。