学习日志

姓名：辛昊洋　　学号：1813090　　日期：6.26

学习内容安排：（以 6.15 为例）

上午：

MapReduce 特性

下午：

Hadoop 集群搭建

学习反馈：

已掌握知识：

特性：

1、计数器

1.1、MapReduce 包含的高级特性，计数器，数据集的排序和连接

1.2、计数器作用，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

1.3、计数器分组 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

1.4、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

1.5、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

1.6、作业计数器由 master 维护，无需网络间传输数据

1.7、java 可以自定义计数器，如，数据不规范记录计数器

2、排序

2.1、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

2.2、部分排序、全排序、辅助排序

3、连接

3.1、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

3.2、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

3.3、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

未掌握知识：
集群搭建按照网上的操作出现了一些问题还未解决。