**20200624学习报告**

1813068郭冠琛

**学习内容**

* MapReduce核心功能：数据本地化，加快访问速度
* MapReduce

用于数据处理的编程模型

分为map阶段和reduce阶段（两个函数一次只处理一个键值对，被重复调用）

* map阶段

mapper接口，形参指定输入键、输入值、输出键、输出值类型

有自己的用于网络优化的类型

longwritable相当于长整型，intwritable为整型，text=string

map方法需要一个键和一个值，以及一个写入输出对象的collection（outputcollector），一个reporter

* reduce阶段

reducer接口形参同上

reduce输入必须与map输出类型匹配（在示例中map函数的输出对象，键为年份，值为年份中所有的气温，故reduce输入时值的参数为iterator）

* 运行阶段

jobconf对象指定作业参数，初始化时需要传入作业类对象（xxx.class）

setjobname

指定输入输出路径：FileInputFormat.addInputPath(jobconf对象，路径)

FileOutputFormat.setOutputPath(jobconf对象，路径)

setMapperClass和setReducerClass

setOutputKeyClass和setOutputValueClass（以上四个方法需要传入类对象）

* 分布式运行

jobtracker（总调度）tasktracker（执行）

map将输出写入本地硬盘，需要通过网络传输到reduce运行的节点进行合并

combiner：可以用于优化，减少map与reduce交互量（处理map输出结果），但系统不保证是否会在map结束后调用combiner或调用多少次

setCombinerClass

* hdfs

块：能处理的最小数据量

hadoop fs -copyFromLocal 本地文件路径 hdfs路径 将本地文件拷贝到hdfs（from->to相反）

hadoop fs -ls 显示文件目录下内容（第二列显示文件副本数）

hadoop distcp hdfs://namenode1/path hdfs://namenode2/path将名称节点1path目录复制到名称节点2的path下，用于在两个集群中传输数据（要求hdfs版本相同）

hadoop archive -archiveName 归档文件名 归档文件夹路径 归档文件存放位置 该指令生成一个.har归档文件，查看该文件中文件时需要使用归档文件路径/归档文件名/目录

* 调试作业

**需要解决的问题**

hdfs的java接口和数据流部分不太能理解

四五章内容不太能理解