**20200626学习报告**

1813068郭冠琛

**学习内容**

* 计数器

内置计数器：被相关任务维护并定期发送状态到tasktracker，然后发送到jobtracker，这样就可以在全局范围内汇总。（实际由jobtracker维护，不需要通过网络发送）

自定义：java枚举型

getcounters（）返回所有计数器

findcounters（）返回特定计数器

* 排序

记录到达reducer之前按键排序，但值不排序

patitioner保证键相同的数据由同一个reducer处理（reducer通过分区按键成组）

* 联接

多个数据集

map端联接：到达map前即合并 要求输入数据集必须按相同大小分区且按相同顺序排序

reduce端联接：mapper给每条记录打上源标签且以联接键作为map输出键

* Hadoop集群安装

Hadoop可以运行在任意一个具有JVM的平台

多机器需要映射节点到机架上，放置MapReduce时将优先选择机架内传输

* 名称节点在内存中保存整个命名空间文件和块的元数据，需要高内存
* 运行hdfs需要指明名称节点，使用fs.default.name指定主机URL和端口（默认8020）
* Hadoop后台程序会运行一个为后台程序间通信的RPC服务器和一个HTTP服务器
* Hadoop基准测试

查看所有用于测试的jar

testDFSIO：测试作为并行读写文件的便捷方法

RandomWriter生成随机数据，以每个节点十个map且每个map生成键值长度不同的10GB二进制数据的方式运行MapReduce作业。

sort排序

sortvalidator检验排序前与排序后的数据来确认排序是否正确

mrbench多次运行一个小作业

nnbench名称节点压力测试

gridmix模拟现实集群作业负荷的集群测试

**需要解决的问题**

排序方式

次要数据的存储

Hadoop集群配置

hadoop后台服务器

由于缺乏实践以及部分前置知识，Hadoop预习效果较差，需要更深入的学习与理解。