**学习日志**

姓名： 陈昭宇 学号：1812977 日期：6.23

**学习内容安排：**

上午：

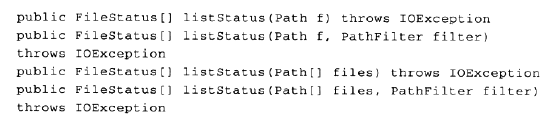
了解MapReduce以及Hadoop分布式文件系统（HDFS）

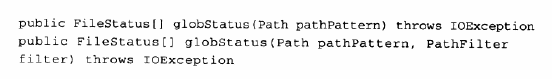
下午：

了解MapReduce以及Hadoop分布式文件系统（HDFS）

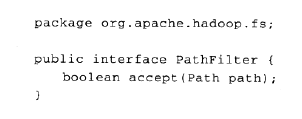
**学习反馈：**

已掌握知识：

1. map与reduce（map处理NCDC数据，提取需要的信息输出，然后通过mapreduce框架处理，然后再发送到reduce函数，reduce选择需要的输出）
2. map函数的实现（通过Mapper接口【泛型接口，指定输入键，输入值，输出键，输出值】）
3. LongWriteable 相当于Long,Text相当于String IntWritable 相当于Integer
4. Reduce通过Reducer接口实现
5. JobConf 对象指定了作业的各种参数。他授予你对整个作业的控制权
6. 新的API不兼容以前的API(以前偏向于接口，现在偏向于抽象类；放的包不一样；新的广泛使用context；新的API同时支持推拉的迭代；新的统一了配置，作业的控制的执行由Job类来负责)
7. Mapreduce的分布化
8. 数据流（用combiner 减少map与reduce之间的数据传输量）
9. Hadoop流（Ruby语言 -combiner 设置 -file Python版）
10. Hadoop管道C++接口 键和值是字节缓存，表示为STL的字符串。
11. 分布式文件系统：管理计算机网络存储HDFS，以流式数据访问模式存储超大文件（一次写入，多次读取 ；）：低延迟数据访问；大量的小文件；多用户写入，任意修改文件
12. HDFS的概念
    1. 块：代表一个磁盘能够读写的最小数据量；HDFS中默认为64MB【减少寻址开销】（小于一个块的内容不会独占一个）好处方便跨磁盘，利用抽象单元可以简化储存子系统fsck显示块的信息
    2. 名称节点与数据节点（名称节点管理者，数据节点工作者）
    3. 防止名称节点故障（复制，而写入NFS）
    4. 命令行接口 fs-copyFromLocal
    5. 文件系统（org.apache.hadoop.fs.FileSystem）
    6. Thrift API将hadoop文件
    7. 其他接口
13. Java接口
    1. 读取数据 java.net.URL打开数据流，从而从中读取数据
    2. SetURLStreamHandler-Factory
    3. 使用FileSystem API读取数据【get(Configuraction conf)throws IOException;get(URL uri, Configuration conf)throws IOException】
    4. 写入数据 public FSDataOutputStream create(Path f)throws IOExpection 重载方法 Progressable告知写入进度或者用append（）在一个已有文件中追加（public FSDataOutputStream append (Path f)throws IOExpection）
    5. 目录 public Boolean mkdirs(Path f)throws IOException
    6. 查询文件系统 FileSyetem getFileStatus（）方法，如果不存在会抛出异常
    7. 
    8. 文件格式



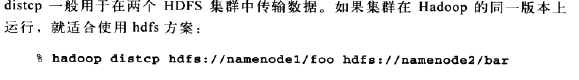
* 1. 可选PathFilter



* 1. 删除数据



1. 数据流
2. 通过distcp 进行并行复制



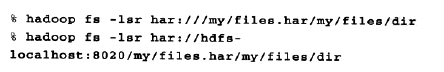
用-overwrite 和update会改变路径和目标路径的含义

1. Hadoop 归档文件



第一个选项为归档文件名称

递归

1. 

使用其他文件系统

未掌握知识：第一点没有领悟到Hadoop的妙处，就是一个人没有想过多文件系统的好处，第二个对于语法只是写，不知道具体的用的地方，很抽象。对于数据流没有办法理解

学习心得： Hadoop 这种基于并行结构的文件处理的软件，还没有明白，就我现在的理解，还是觉得很抽象。

。