哈尔滨工业大学

**<<大数据分析>>**

**实验报告一**

**(2022年度春季学期)**

|  |  |
| --- | --- |
| **姓名：** | **李金宣** |
| **学号：** | **1181910201** |
| **学院：** | **大数据** |
| **教师：** | **张开旗 杨东华** |

实验一 数据预处理

## 实验目的

掌握数据预处理的步骤和方法，包括数据抽样、数据过滤、数据标准化和归一化、数据清洗。理解数据预处理各个步骤在大数据环境下的实现方式。

## 实验环境

操作系统：Windows、Linux虚拟机(vmware ubuntu 20.04)

框架：伪分布式Hadoop环境

编程语言：Java

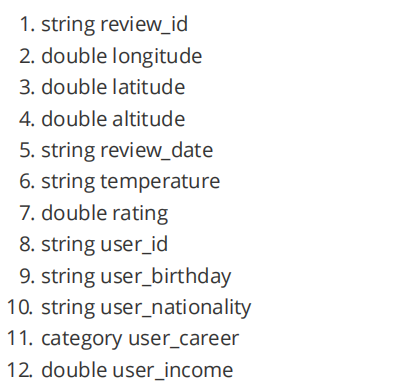
## 三、实验过程及结果

3.1 数据抽样

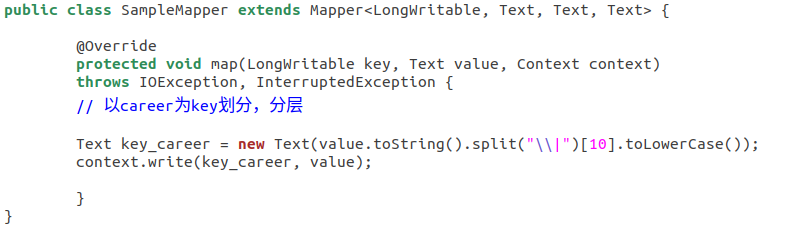
抽样的MapReduce过程：

输入文件路径/data.txt，输出文件路径/D\_Sample

data.txt的数据类型



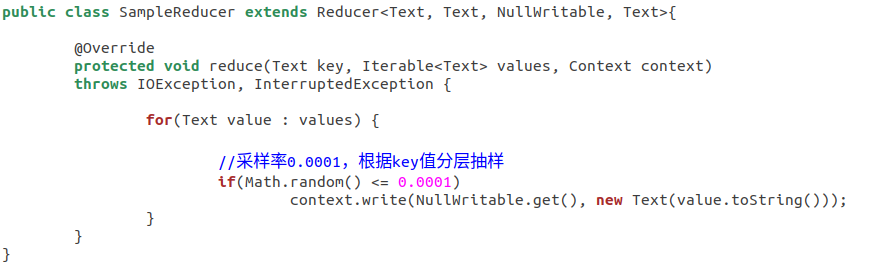
Mapper：



分层抽样所使用的属性为user\_career属性，对于传入Mapper的data.txt源数据集文件中的每一条，将其第11条属性career转化为text类型，并作为Mapper的key值传递给Reducer，将每条数据本身作为value值传递给Reducer。

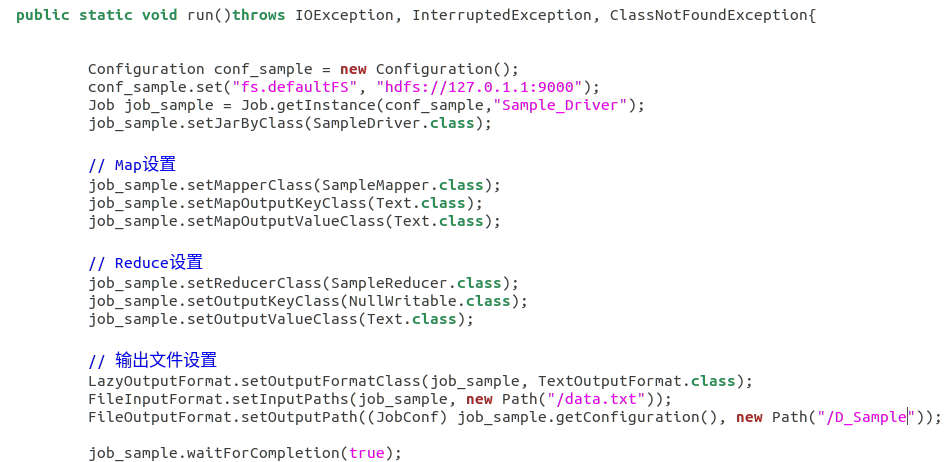
Reducer：

每个career(Reducer过程的key)都对应自己的value值，实际相当于把value值按其career属性分层，在Reducer过程中分布式地对每个career的层进行抽样，采样率为0.0001，通过随机数的方式采样，将采样后数据写入/D\_Sample文件。



Driver：

设置输入文件路径/data.txt，输出文件路径/D\_Sample



以下的2、3、4过程皆在一轮Mapreduce中完成

3.2 数据过滤

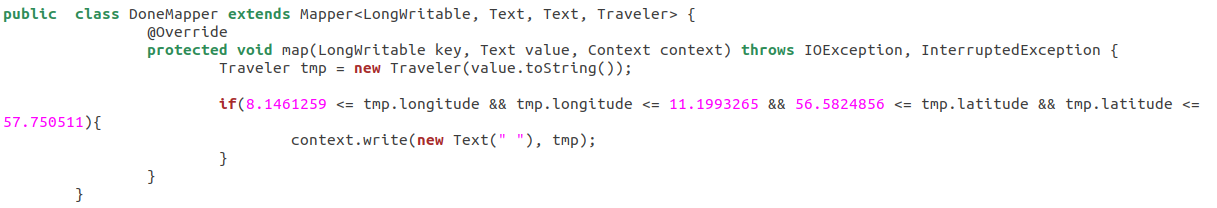
输入文件路径/D\_Sample/part-r-00000，输出文件路径/D\_Filter

定义一个Traveler类：



Taveler类中的私有变量其数据类型及名称与源数据格式相同。在构造方法Traveler()中将输入文件中每行的字符串类型，转化为Traveler类中各种与源数据格式相同的变量类型，方便之后处理，而对于rating与income的缺失值，暂时以默认值-1.0进行填充，更进一步的填充将在后面的数据清洗中进行。

Mapper：



Mapper中的数据过滤，将输入文件的每一行转化为Traveler类之后，按其longitude和latitude值进行数据过滤，在奇异值边界以内的value才可被选择传递到Reduer中，不需要分类，故传递到Reducer的key值为空，直接传递即可，而传入到Reducer的value值为是字符串类型被转化成Taveler类型的结果。

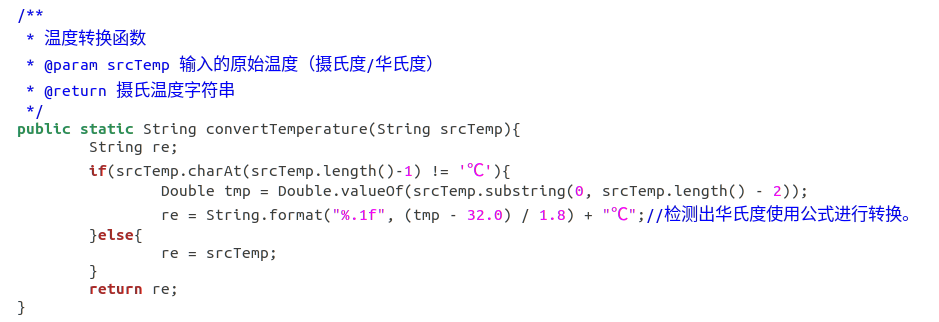
3.3 数据格式转换与归一化

3.4 数据清洗（缺失值填充）

格式转换：

定义一个convert类，两个方法将日期与温度进行格式转换，日期转换为2018-03- 21形式，温度转化为摄氏度形式，在上节3.2中提到的Traveler类构造方法Traveler()中,顺便调用了convert类的两个方法，将类型转换后传递给Redycer每一条数据的日期与温度都转换成了统一的格式，省去了一次类型转换的MapReduce过程。





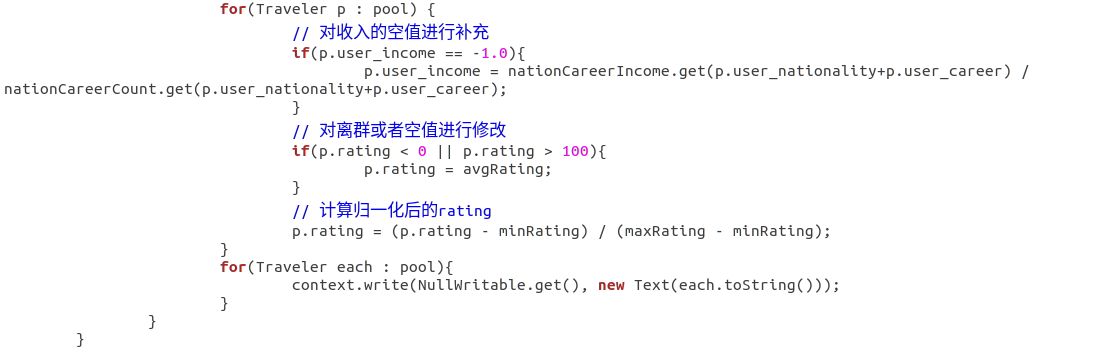
归一化：

在Reducer过程中，先设置一个Traveler类型的动态数列pool，将每一个传递到Reducer的value值都存入其中，之后可以遍历这个pool中的每一个元素再进行清洗并写入输出文件，省去了为了求最大最小值等参数的Mapreduce后，需要再进行一轮清洗数据的一次Mapreduce过程。

首先遍历每一个Reducer接受的value值，求出需要归一化与填充的属性rating的最大值，最小值，总和，平均数等参数，求出需要填充的属性income的上述参数

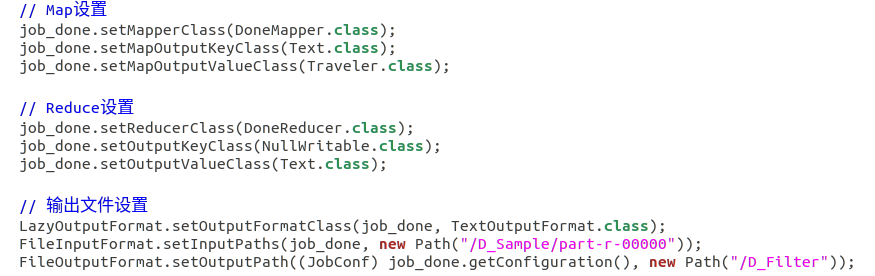


然后遍历之前的pool动态数组，其中的元素正是之前Reducer接受的所有value，这样使用pool遍历省去了一轮MapReduce过程，将计算出的rating与income的平均值作为填充值填入缺失的属性值中，并利用公式，将rating进行归一化后填入，最后将处理后的数据写到输出文件中，至此格式转换，归一化，数据填充皆已完成。

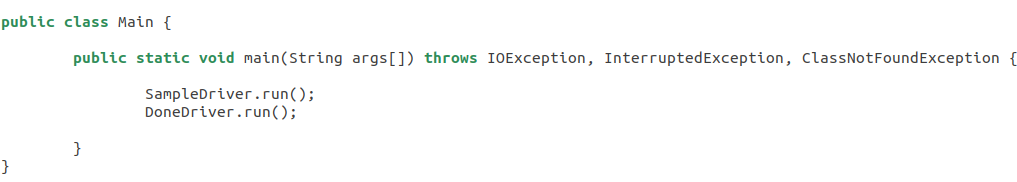


Driver

输入文件路径/D\_Sample/part-r-00000，输出文件路径/D\_Filter



Main函数运行

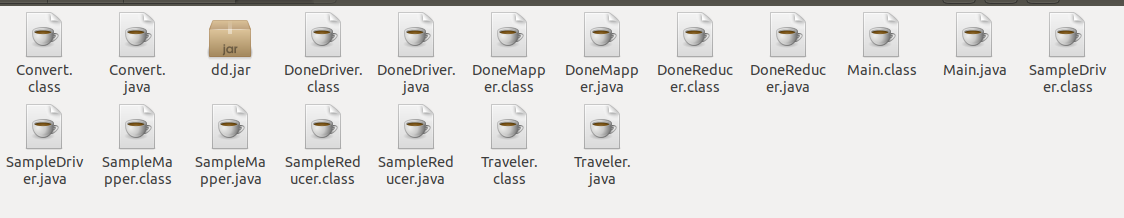


3.5 运行结果

javac \*.java

jar cvfe dd.jar Main \*

将java文件编译后打包为jar程序包



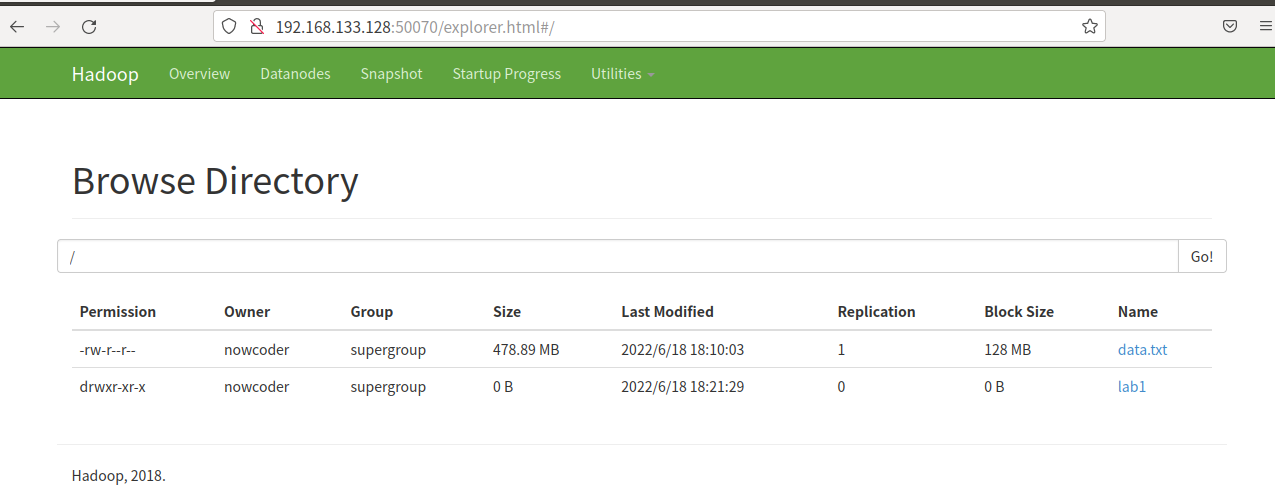
hdfs namenode -format 格式化NameNode

hadoop-daemon.sh start namenode 启动 NameNode

hadoop-daemon.sh start datanode启动 DataNode

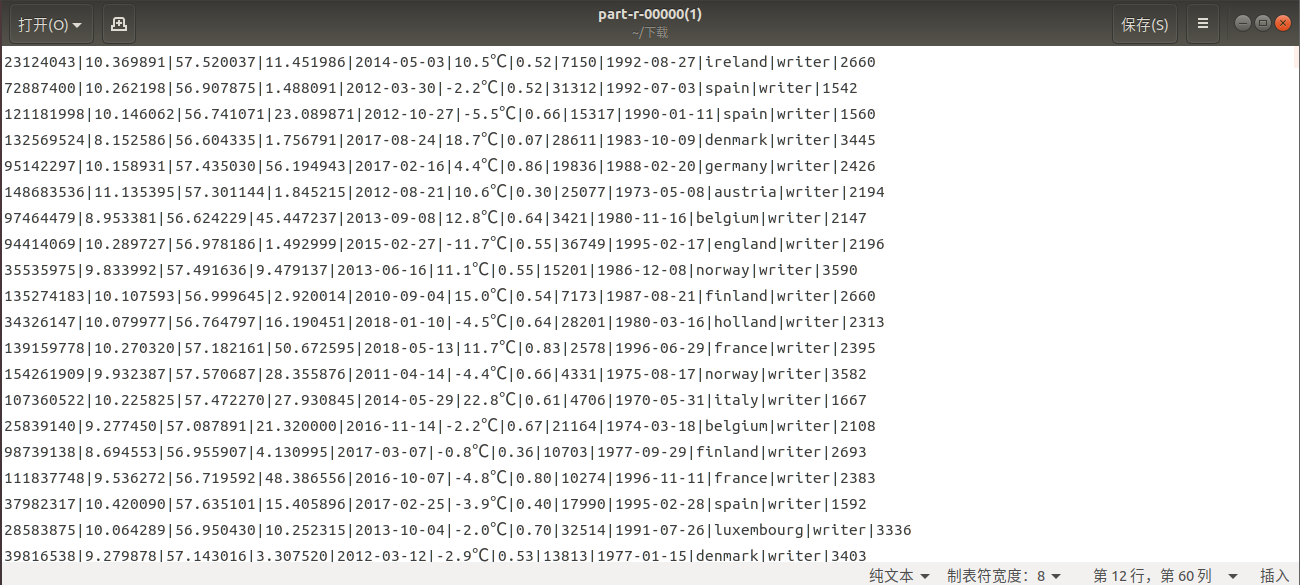
hdfs dfs -put /home/nowcoder/hadoop/hadoop-2.7.7/lab1/data.txt /

将本地文件放到HDFS上



hadoop jar dd.jar Main运行

从HDFS下载的运行后得到的结果文件



## 四、实验心得

实验环境hadoop在Linux虚拟机的初次配置花了很长时间，遇到过很多问题，用了很多天解决。

对MapReduce编程初次使用很不熟悉，用了很长时间了学习相关编程知识。

为降低Mapreduce轮数想了很多方法，最后采用了如实验报告中多种方法的综合，如Traveler类直接转化，使用动态数组存储Reducer的值减少Mapreduce等，成功将Mapreduce轮数降低为2轮。