

一个Linux上用python直接处理csv文件的方法

汪兴元

2015 年7 月21 日

摘要

我们常常需要处理体积很大的csv数据或日志文件。一般来讲，导入RDBMS来处理是常用的一种办法，但此办法也不完美，尤其是SQL这种描述型的语言。在表达算法的时候不如一般的程序设计语言的表达力强，某些应用场合虽然能够实现，但是难度太高，优化不易；批量处理大数据，通过在RDBMS上运行SQL效率并不高。

另一种办法就是使用Hadoop或Spark，或导入NoSQL中，再使用MapReduce。但是使用类似Hadoop之类的工具又可能太重，数据量不够运作一个Hadoop集群，还得承受其代价。

本文介绍了直接处理csv文件的一套办法，作为另一个数据处理的选项，示例代码用python，运行的操作系统为Linux，源程序及csv数据文件的编码采用utf8。在从原始数据到最终的目标结果，需要组合本文提到的各种算法，才能达到目的。通过管道-过滤器连接每一个算法，能够将整个处理算法分而治之，同时得到比较好的性能。

本文假设要处理的整个数据集无法一次装入机器的内存。

目录

1 校验数据	2
2 选择列	3
3 排序	4
4 转换ID字段	4
5 去重复值，整行distinct()	6
6 连接	7
6.1 内连接	7
6.2 左连接	8
6.3 右连接	8
6.4 全外连接	8
7 过滤数据	9
7.1 谓词及表示方法	9
7.2 谓词的连接关系	11
7.2.1 AND	11
7.2.2 OR	11
7.2.3 优先级	11

8 聚集	11
8.1 group by	11
8.2 having	11
8.3 max(), min(), avg()	11
9 补缺失值	11
10 转置	11

1 校验数据

csv文件可能存在错误。在正常情况下，csv文件不应存在错误。但是写入csv的过程并非事务型的，不象RDBMS那样有很好的ACID保证，故磁盘耗尽，进程异常退出、挂起等因素，直接导致csv文件的格式并非预期。

检查数据没有统一的方法，要视数据自身的特点来做检查。一般是检查一些数据正确所需的必要条件，但必要并不一定充分。一般的检查方法有：

1. 检查列数。如果列数不等于预期值，可以确定此行数据错误。
2. 检查每列数据的格式，比如，数字，日期，或其它满足指定格式的文本串。可以尝试将其解析为对应的类型，看能否成功；或用正则表达式验证。
3. 校验字段的值。前两项都对的情况下，可以检查数据值的取值范围。比如健在的人的年龄不能是负数，也不能是数百以上；历史数据中的记录时间不可能超过当前的日历时钟。

```
1 #!/usr/bin/python
2 # coding=utf-8
3 # validate_history_ai.py
4
5 import csv
6 import sys
7
8 if __name__ == '__main__':
9     infile = sys.stdin
10    outfile = sys.stdout
11
12    reader = csv.reader(infile, delimiter=',', quotechar='"',
13        quoting=csv.QUOTE_ALL, skipinitialspace=True)
14    writer = csv.writer(outfile, delimiter=',', quotechar='"',
15        quoting=csv.QUOTE_ALL, skipinitialspace=True)
16
17    for row in reader:
18        try:
19            # check if the number of columns is right.
20            if len(row) != 12
21                continue;
22            # check the format of each column...skipped.
23            # check the value range of each column...skipped.
24            writer.writerow(row)
25    except:
```

```
26         pass
27     finally:
28         pass
```

如果数据是已经通过有严格校验的系统中生成或导出的，原始数据本身无误，一般做以上三项检查可以查出我们能见到的全部错误。但是如果数据是人工填写或是有故障的机器生成的，这三项检查仍不能确保检查出全部错误。

对于错误的数据该如何处理，要视情况而定。有的业务场景，例如Web服务器的access log，价值不高；又如水温传感器输入的每分钟一次的历史数据，可以丢弃，之后使用插补法补缺；有的场景是不可以这样做的，如交易记录，需要重新导出此段数据或人工处理。样例代码中我们采用了丢弃的方法处理。

以上代码没有直接打开csv文件，而是从标准输入中读取文件，处理完毕之后再写到标准输出。这样的好处是便于通过管道过滤器连接多个处理进程，避免过高的耦合度。本文所有的处理程序都使用管道过滤器连接，不再赘述。

程序中对校验2.和3.未实现，可以练习下。对于机器导出的数据，做完1.可以去掉大部分错误。如果不打算实现2.和3.

可以将列数从命令行上读取，成为更通用的子程序。这个作为练习。

2 选择列

原始数据中有可能只有部分列是所要关心的，其它的与要解决的问题无关，可以丢弃。因此要选择列，类似SQL语句中的SELECT所要办的事情。样例程序如下所示，这里我们只对第1, 3, 4, 5列感兴趣。

```
1  #!/usr/bin/python
2  # coding=utf-8
3  # select_history_ai.py
4
5  import csv
6  import sys
7
8  if __name__ == '__main__':
9      infile = sys.stdin
10     outfile = sys.stdout
11
12     reader = csv.reader(infile, delimiter=',', quotechar='"',
13         quoting=csv.QUOTE_ALL, skipinitialspace=True)
14     writer = csv.writer(outfile, delimiter=',', quotechar='"',
15         quoting=csv.QUOTE_ALL, skipinitialspace=True)
16
17     for row in reader:
18         outrow = []
19         outrow.append(row[0])
20         outrow.append(row[2])
21         outrow.append(row[3])
22         outrow.append(row[4])
23         writer.writerow(outrow)
```

可以将列下标从命令行输入，这样这个程序就成为一个通用的选择程序，而不是仅用于示例的场景。这个作为练习。

3 排序

我们对数据做去重、转置、分组、聚集，都需要先对数据排序。因此排序是非常重要的功能。不可或缺。

大数据的排序由于无法直接一次装入内存，故必须使用外部排序。如果能将数据切成能用内部排序的小块，排好序之后，再合并，那么我们就可以完成对大数据的排序。

我们打算从头实现一个大数据排序工具。利用Linux的工具`sort`来做排序。先将csv文件拆分。可以在导出或生成csv的时候，每达到某一固定尺寸就rotate一下，当前文件改名，再创建一个新文件当作当前文件。

假设文件已经切成能用内存装下的多个小文件。注意，使用切割工具切文件，必须从整行数据的位置断开，否则，切口处的那条数据被切坏了。

对每一个文件，执行`sort`。下面的两段代码是在Linux的终端上输入的命令，“\$”表示命令提示符。示例中有两个文件，这里只写了第一个文件的排序，另一个省略。

```
1 $ cat data/att-rec-utf8_part-1.txt \  
2 | python python-src/select_att_rec.py \  
3 | sort --field-separator=, --ignore-leading-blanks --stable \  
4 --key=1.2n,2 --key=2,3 >part-1.csv
```

此命令先用`select_att_rec.py`选择所要的列，再对选择的结果排序，之后将结果重定向到文件`part-1.csv`。所有小文件都排序好之后，再对排序后的文件做merge：

```
1 $ sort --field-separator=, --ignore-leading-blanks --stable \  
2 --key=1.2n,2 --key=2,3 -m part-*.csv > merge-sorted.csv
```

注意在做merge之前，必须先对每个文件排好序；merge时，使用的关键字、分隔符等参数必须与`sort`的时候所用的参数完全一致。

merge完之后，得到一个已经排好序的大文件。

4 转换ID字段

一般在集成两个系统的数据的时候，需要将两个系统的数据合并起来。但是两个系统的数据ID一般是相互独立分配的，因此，如果ID都是同一种类型，比如整数，几乎100%会发生冲突，即不相同的两个对象，其ID相同。即使用字符串之类的做ID，也难以不发生冲突。因此必须要建立两个系统间的对象的对应关系。

例如，商品销售网站和库管理系统如果是两套独立的系统，分别由不同的团队开发，比如库存管理系统是外购的，商品销售网站是自建的，如果要将两套系统的数据整起来合做分析，又没有办法修改系统达到统一，就需要做ID转换。

下面的程序从标准输入读入历史数据，从命令行读取基础数据文件名，加载基础数据到字典中，利用基础数据中的 $R(ID_{src}, ID_{dest})$ 关系，做了一个映射。逐行将原ID映射到目标ID的方法替换掉ID，实现ID转换的目的。

```
1 #! /usr/bin/python  
2 # coding=utf-8  
3 # transform_data_id.py  
4
```

```

5 import csv
6 import sys
7 import traceback
8
9 dataObjectIds = {}
10
11 if __name__ == '__main__':
12     infile = sys.stdin
13     outfile = sys.stdout
14     cfgfile = None
15     if(len(sys.argv) > 1) :
16         cfgfile = open(sys.argv[1], 'rb')
17
18     cfgreader = csv.reader(cfgfile, delimiter=',', quotechar='"',
19         quoting=csv.QUOTE_ALL, skipinitialspace=True)
20     reader = csv.reader(infile, delimiter=',', quotechar='"',
21         quoting=csv.QUOTE_ALL, skipinitialspace=True)
22     writer = csv.writer(outfile, delimiter=',', quotechar='"',
23         quoting=csv.QUOTE_ALL, skipinitialspace=True)
24
25     for row in cfgreader:
26         try:
27             key = int(row[0]), int(row[1])
28             value = int(row[2])
29             dataObjectIds[key] = value
30         except ValueError:
31             pass
32         finally:
33             pass
34
35     for row in reader:
36         try:
37             key = int(row[0]), int(row[2])
38             dataId = dataObjectIds.get(key, None)
39             converted = []
40             if dataId != None:
41                 converted = [dataId, row[1]]
42                 converted.extend(row[3:])
43                 writer.writerow(converted)
44         except:
45             print >> sys.stderr, row, traceback.format_exc()

```

一般来讲，大多数系统中的某一基础数据对象的ID是可以一次全部装入内存的。比如超市的商品ID，虽然可能有数十万条，但是对于现代计算机的内存来讲并不算多。

对于历史记录，其身的ID，或Primary Key的数量大，其总体积有可能单台机器的内存装不下。不过一般情况下，需要做转换的必要性往往不大，只对其通过Foreign Key引用的基础数据对象的ID有转换的价值。如果实在需要转换，使用6.1提到的内连接方法来做大表的连接，再选择所

要的字段，包括新ID，不选原ID字段即可。

这个方法除了可以用于转换ID的场合，还可以用于更换名称。比如，我要将公司员工的行为的分析结果公布，但是不想让人知道员工的真实姓名或工号；或者谁购买了伟哥，将要商品名换掉。

样例程序中的ID列号、多少列都是hardcode进代码的。可以将关系搞成动态的，从命令行读入。设 C_{i_1}, \dots, C_{i_m} 是从要转换的数据中选出来作为原ID的列， C_{j_1}, \dots, C_{j_n} 是要转成的目标ID列， $m, n \geq 1$ 。则转码可以视为从

$$\begin{aligned} transform : R_{src} &\rightarrow R_{dest} \\ ID_{src}(C_{i_1}, \dots, C_{i_m}) &\mapsto ID_{dest}(C_{j_1}, \dots, C_{j_n}) \end{aligned} \quad (1)$$

只需要在命令行上输入源ID的列号清单和目标ID的列号清单，在程序中，用迭代清单代替硬编码的取列语句即可。将hardcode转为动态，作为练习。

5 去重复值，整行distinct()

如果在内存中，用set来装数据，自动去重。但是我们的数据量大，不能装入内存，因此直接用set不行。如果对每个输入值，都去扫描一遍记录，如果共有 n 个记录，则对第1个记录，要扫描第2, \dots , n 共 $n-1$ 个记录；对第2个记录，要扫描第3, \dots , n 共 $n-2$ 个记录，对第 $n-1$ 个记录要扫描第 n 个记录共1个记录。于是一共要扫描文件 $n-1$ 次，转储到文件 $n-1$ 次，其中最后一次是结果，中间 $n-2$ 次是临时文件。这个性能相当差。共扫描记录的个次是

$$(n-1) + \dots + 1 = \frac{1}{2}n(n-1). \quad (2)$$

csv文件每行的长度不固定，因此，定位记录的位置很难，往往扫描的记录数是 n^2 。

如果将数据先排序，则相同的两条记录会相邻，成为一组，于是，我们只需要比较相邻的两条数据是否相同，如果相同就丢弃一条，如果不同就输出这一条。示例代码如下，假设数据事先已经排好序。

```
1 #!/bin/env python
2 # coding=utf-8
3 # remove_duplicate.py
4
5 import csv
6 import sys
7
8 currentRow = None
9 def duplicate(row):
10     global currentRow
11
12     result = True
13     if(currentRow == None):
14         result = False
15     else:
16         for i, j in zip(row, currentRow):
17             if i != j:
18                 result = False
19             break
20     currentRow = row
```

```

21     return result
22
23 if __name__ == '__main__':
24     infile = sys.stdin
25     outfile = sys.stdout
26     if(len(sys.argv) > 1) :
27         infile = open(sys.argv[1], 'rb')
28
29     reader = csv.reader(infile, delimiter=',', quotechar='"',
30         quoting=csv.QUOTE_ALL, skipinitialspace=True)
31     writer = csv.writer(outfile, delimiter=',', quotechar='"',
32         quoting=csv.QUOTE_ALL, skipinitialspace=True)
33
34     for row in reader:
35         if(not duplicate(row)):
36             writer.writerow(row)

```

6 连接

数据仅代表同一个表或关系的csv是不够的。两个表csv连接跟写SQL查询一样常见。如果是历史记录表跟基础数据表的连接，一般基础数据很小，历史数据大，所以可以采用 4 所述的方法。即把历史表的Key作为dict的Key，基础表的Key作为dict的Value，逐行转换。

如果是内连接，则对于每行历史记录，必须找到对应的基础表中的记录才算连接上，否则要丢弃；而对于左连接，找不到对应的记录，则用空值填充。

对于右连接，和全外连接，由于历史数据很大，不能直接装入内存，这办法不适用。

接下来描述不将表装入内存的连接方法。

6.1 内连接

先将两个表的csv按连接的关键字排序，升序或降序都行，但必须都是升序或都是降序。

再逐行连接。不妨假设都按升序排好了序。设两个表分别为 R_1 , R_2 ，当前从这两张表取的数据行的行 r_1 , r_2 。

1. 先从 R_1 中取第一行为 r_1 ，取其连接的Key，设为 K_1 ，取 R_2 中的第一行为 r_2 ，取其Key， K_2 ，进行比较。如果 $K_1 < K_2$ ，由于 R_2 为升序， r_2 及 R_2 余下的行的Key均不小于 K_2 ，因此， r_1 与 r_2 及 R_2 余下任一行都连接不上。于是再取 R_1 中的下一行为 r_1 。
2. 如果 $K_1 > K_2$ ，由于 R_1 升序， r_1 及 R_1 余下的行的Key均不小于 K_1 ，因此， r_2 与 r_1 及 R_1 余下任一行都连接不上。于是再取 R_2 中的下一行为 r_2 。
3. 如果 $K_1 = K_2$ ，则连接成功，输出连接后的行。再取 R_2 的下一行为 r_2 。重复以上步骤，直至所有的数据处理完毕。

此算法的python实现如下。

```

1 #! /bin/env python
2 # coding=utf-8
3 # remove_duplicate.py
4
5 import csv

```

```

6 import sys
7
8 if __name__ == '__main__':
9     leftTableIn = sys.stdin
10    rightTableIn = open(sys.argv[1], 'rb')
11    outfile = sys.stdout
12
13    leftReader = csv.reader(leftTableIn, delimiter=',',
14        quotechar='"', quoting=csv.QUOTE_ALL, skipinitialspace=True)
15    rightReader = csv.reader(rightTableIn, delimiter=',',
16        quotechar='"', quoting=csv.QUOTE_ALL, skipinitialspace=True)
17    writer = csv.writer(outfile, delimiter=',', quotechar='"',
18        quoting=csv.QUOTE_ALL, skipinitialspace=True)
19
20    try:
21        r1 = leftReader.__next__()
22        r2 = rightReader.__next__()
23        while True:
24            key1 = (r1[0], r1[2])
25            key2 = (r1[0], r1[1])
26            if key1 < key2:
27                r1 = leftReader.__next__()
28            elif key1 > key2:
29                r2 = rightReader.__next__()
30            else:
31                row = []
32                row.extend(r1)
33                row.append(r2[2])
34                writer.writerow(row)
35    except StopIteration:
36        pass;

```

此算法利用数据集的顺序，避免了重复扫描csv，只需one pass即可完成连接；缺点是必须先排序。

可以考虑将程序写成通用的，在命令行上指定连接谓词。这个可以作为练习。

6.2 左连接

同前述内连接，但是步情况1.连接不上 r_2 的时候，将 r_1 补齐空列之后输出。

6.3 右连接

同前述左连接，但是将 R_1 、 R_2 的位置互换即可。

6.4 全外连接

可以有两种办法可选：

1. R_1 、 R_2 先做一遍左连接， R_1 、 R_2 再做一遍右连接，合并两个连接的结果，最后去掉重复值即可。注意要保持左连接和右连接的列顺序是相同的。
2. R_1 、 R_2 先做左连接，再将结果与 R_2 做一个类似右连接的操作，区别于右连接的是：对于有匹配的行，直接丢弃；对于无匹配的行，填写空值后输出。

办法1.简单，但需要的步骤多，每个步骤都要用到文件；办法2.稍复杂，但步骤少可以只写一次文件，采用管道直接连接中间结果。

从性能上考虑，办法2. 较优。

7 过滤数据

筛选满足条件的数据是信息检索中必不可少的部分。将用户的Search Criteria转成过滤数据的Predicate。Search Criteria可能包含多个条件，为了简化问题，需要将每个条件做成一个Predicate，只做一项测试，多个条件需要多个Predicate，并且每个Predicate 之间可能是AND，也可能是OR，并且有优先级顺序要求。

7.1 谓词及表示方法

谓词是一个返回Boolean值的表达式，用于测试对象的性质是否为真，或对象之间的关系是否为真。

例如测试订单是否用银行卡支付；帐户余额是否大于100块； a 同学的考试分数是否比 b 同学的高，等等。

```
1 for row in input:
2     if predicate.test(row):
3         output.writerow(row)
```

如果只测试其中的某列或几列，也可以如此：

```
1 for row in input:
2     if predicate.test(tuple([row[0], row[2]])):
3         output.writerow(row)
```

经验不多的开发者在写过滤条件时，往往不注意将各谓词分解为基本的形式，并且与其它过滤条件，可能还与过滤数据无关的问题，合并一起解决。这样一来，大片的其它逻辑与谓词测试的内容混合在一起，修改起来要同时考虑多个问题。

这样在问题不太复杂的时候问题不大，而且可能性能比分离关注点来考虑更简单。在解决一个较大的问题的时候，往往复杂到做不下去。

例如，使用的“Predicate”并不返回Boolean值，而是直接返回满足测试条件的数据，并且做了处理，列数，字段的含义都可能已经变了，因为恰好下一步也需要这个数据，这是常见的问题之一。例如，

```
1 for row in input:
2     result = predicate.test(row):
3     if result != None:
4         output.writerow(result)
```

上面的代码中，`result == row`未必能成立。光看这段代码，天知道`predicate.test(row)`里面做了什么。

做Predicate的时候要把握以下原则：

1. Predicate只测试数据是否满足所指定的条件，满足返回True，不满足返回False。不能更改输入的数据。
2. 不做与指定的测试无关的任何其它工作，例如，提取数据，计算平均值之类。如果所需要做的其它工作与测试结果有关，应交给Predicate的使用者来完成；如果无关，则更应分开来做，既不能在Predicate里做，又不能在Predicate的使用者里面做。
3. Predicate不能有状态。其返回值仅取决于输入的条件，包括数据和测试条件。重复测试同一条数据，得到的结果应相同。如果使用lambda或闭包，其中捕获的参数不应发生变化。

下面对这3条举例说明。

原则1. 可能出于下一步的数据处理方便的目的，在当前Predicate中对输入数据作调整，容易实现，又比较方便。这样做，从名义上，Predicate只做测试，而不作修改，而事实上却做修改，容易误导代码的读者；而且关注点不分离，引起复杂度上升，例如，如果下阶段增加了一个Predicate需要没有修改的数据，这样就有矛盾了：如果去掉原Predicate对数据的更改，则后面的处理不能正常运行，也要改。

原则2. 例如，有人做了一个比较大小的Predicate，用于扫描csv文件是否为升序，通过比较第 n 条和第 $n+1$ 条数据的大小。因此，可以顺便用来统计csv文件的行数。假设数据满足升序条件，对每一条数据都做了扫描。

如果csv文件总共有 N 条数据，Predicate将会被调用 $N-1$ 次。反过来，如果Predicate被调用了 $N-1$ 次，那么csv文件的行数应是 N 行。

假设文件既不满足升序，又不满足降序，那么在中间某一位置 k ，Predicate返回False，测试终止，这时候Predicate统计的行数只能是前面有 $k+1$ 条数据满足条件，这个与行数不符。

另有一排序功能需要用到比较的Predicate，于是该Predicate直接被用于排序，但排序会用到多少次比较与数据的条数的关系不是这样的，于是这个行数统计毫无意义。

原则3. 例如比较大小。如果 $a > b$ ，第一次测试`greaterPred.test(a, b) == True`应能成立，而且无论做多少次测试，只要 a, b 保持原值，`greaterPred.test(a, b) == True`必须都成立。

如果使用闭包，捕获的参数可能是引用一个闭包外的对象。由于这个对象不是常量，且有并发访问，这个对象的状态可能会被修改，引起闭包的状态也随之变更。这样一来，闭包的Test Criteria就变了，使用同样的数据做测试，返回值可能会与改变前不同。

谓词分两种类型的：Unary和Binary。前者只需要输入一条数据，后者需要两条。其它类型的可以通过这两类组合运算，或者将数据整合得到。

```
1 def LE(a, b):
2     return a <= b
3
4 pred = lambda x : LE(x, 10)
5
6 for row in inputFile:
7     if pred(row):
8         writer.writerow(row)
```

测试一条数据中的某一属性是否为真，可以用Unary。某些情况看似需要两个数据输入的，也可以用Unary，但是使用lambda或闭包。例如，测试一个数是否大于某个常数，需要比较两个数，但是常数可以使用lambda和一个LE()函数来捕获，这样，得到的lambda对象只需要输入一个数。而对于检查顺序是否升序这种情况，要比较的两个对象没有一个常量，因此直接用二元的LE()函数。

```
1 def isAscending(l):
2     n = None
3     for i in l:
```

```
4     if n is not None:
5         if not LE(n, i):
6             return False
7     n = i
8     return True
```

对于Predicate的接口形式，只要使用起来方便即可。如果过滤数据的时候用到了很多个Predicate，每个Predicate最好按类型统一接口，这样可以用一致的方法处理；这些Predicate最好组合成一个Predicate，能一次做全部的测试，以便调用者在使用时，不必关注Predicate之间的关系细节，只需关注它满足或不满足条件。

7.2 谓词的连接关系

7.2.1 AND

7.2.2 OR

7.2.3 优先级

8 聚集

8.1 group by

8.2 having

8.3 max(), min(), avg()

9 补缺失值

10 转置