Instruction for Data Profiling

In this instruction, I will introduce how I use Map/Reduce to get the result we want. Since the raw data contains too many useless data, that' s the reason why we need to deal with the data before we do the data analysis. In this case, to get the daily population number, we need two rounds of Map/Reduce to achieve it because the raw data is too complicated.

1. Sum all turnstiles data up.

C/A	Unit	SCP	Station	Line Name	Division	Date	Time	Description	Entries	Exits
A002	R051	02-00-00	59 ST	NQR456W	BMT	09/19/2020	00:00:00	REGULAR	7460236	2537220
A002	R051	02-00-00	59 ST	NQR456W	BMT	09/19/2020	04:00:00	REGULAR	7460241	2537220
A002	R051	02-00-00	59 ST	NQR456W	BMT	09/19/2020	08:00:00	REGULAR	7460260	2537239
A002	R051	02-00-00	59 ST	NQR456W	BMT	09/19/2020	12:00:00	REGULAR	7460289	2537289
A002	R051	02-00-00	59 ST	NQR456W	BMT	09/19/2020	16:00:00	REGULAR	7460377	2537315
A002	R051	02-00-00	59 ST	NQR456W	BMT	09/19/2020	20:00:00	REGULAR	7460531	2537338
A002	R051	02-00-00	59 ST	NQR456W	BMT	09/18/2020	00:00:00	REGULAR	7459724	2537009
A002	R051	02-00-00	59 ST	NQR456W	BMT	09/18/2020	04:00:00	REGULAR	7459730	2537010
A002	R051	02-00-00	59 ST	NQR456W	BMT	09/18/2020	08:00:00	REGULAR	7459746	2537056
A002	R051	02-00-00	59 ST	NQR456W	BMT	09/18/2020	12:00:00	REGULAR	7459800	2537160
A002	R051	02-00-00	59 ST	NQR456W	BMT	09/18/2020	16:00:00	REGULAR	7459956	2537190
A002	R051	02-00-00	59 ST	NQR456W	BMT	09/18/2020	20:00:00	REGULAR	7460157	2537215

As the screenshot above shows,

C/A and Unit represent the specific one station, we can easily use the attribute Station as the unique key.

Line Name, Division, and Description are all useless so we will remove all three in the map/reduce process.

SCP represents the turnstile number, which means one station can have several SCP values.

For each turnstile, it will store the data every four hours. We can only use the earliest one in a day to get the initial population number. We can remove those unnecessary data during Map process.

After that, what we have right now include Station, Date, SCP, Entries, and Exits. Since there is more than one turnstile in the same station, we need to sum all those SCP up to get the total number of a day. We can do it during the Reduce process. That' s the first round Map/Reduce.

2. Get the daily number for both entries and exits.

After the first round Map/Reduce, what we have now are four attributes, which contain Station, Date, accumulated_Entries, accumulated_Exits. The data looks like the screenshot below.

```
[hz2169@login-1-1 task]$ head populationSum.csv
1 AV,01/01/2020,456452912,445300773
1 AV,01/02/2020,456459021,445308361
1 AV,01/03/2020,456474713,445326424
1 AV,01/04/2020,456491239,445345342
1 AV,01/05/2020,456498116,445353107
1 AV,01/06/2020,456503137,445358889
1 AV,01/07/2020,456519848,445377816
1 AV,01/08/2020,456536905,445397517
1 AV,01/09/2020,456554573,445418080
1 AV,01/10/2020,456572607,445438554
[hz2169@login-1-1 task]$ tail populationSum.csv
ZEREGA AV,11/11/2020,2661472,1702301
ZEREGA AV,11/12/2020,2662416,1703425
ZEREGA AV,11/13/2020,2663479,1704715
ZEREGA AV,11/14/2020,2664537,1705928
ZEREGA AV,11/15/2020,2665203,1706702
ZEREGA AV,11/16/2020,2665653,1707295
ZEREGA AV,11/17/2020,2666749,1708558
ZEREGA AV,11/18/2020,2667909,1709843
ZEREGA AV,11/19/2020,2669053,1711116
ZEREGA AV,11/20/2020,2670149,1712405
```

In order to get the daily number of both Entries and Exits. What we need to do is calculate the difference between two constant days. That' s why we need the second round Map/Reduce process because we need to first sum all turnstiles data up to get the total number in the first round Map/Reduce process and the data produced from round one will be the input file of round two.

3. Result

After two round Map/Reduce, the result looks like as below.

result_2020

Station	Date	Daily_Entry	Daily_Exit
1 AV	01/01/2020	6109	7588
1 AV	01/02/2020	15692	18063
1 AV	01/03/2020	16526	18918
1 AV	01/04/2020	6877	7765
1 AV	01/05/2020	5021	5782
1 AV	01/06/2020	16711	18927
1 AV	01/07/2020	17057	19701
1 AV	01/08/2020	17668	20563
1 AV	01/09/2020	18034	20474
1 AV	01/10/2020	18710	21325
1 AV	01/11/2020	8015	8974
1 AV	01/12/2020	5329	6222
1 AV	01/13/2020	16637	19261
1 AV	01/14/2020	17533	20327
1 AV	01/15/2020	17898	20211