Part 4: Data Analysis using Hive

```
$ssh ds01@ds101

########
# HIVE #
########

### Create database and table for use
$echo $'CREATE DATABASE IF NOT EXISTS twmask;
USE twmask;
DROP TABLE IF EXISTS twmparquet;
CREATE EXTERNAL TABLE twmparquet (
  code string,
  name string,
  address string,
  tel string,
  adult INT,
  child INT,
  dt string
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS Parquet LOCATION '/dataset/twmpar';' > twmparquet.hsql

$hive -S -f twmparquet.hsql


### Q1: How many Drugstore were participating in distribution of
masks?
$time hive -S -e 'use twmask; select count(distinct code) from
twmparquet where substring(code,1,1)=5' 2>/dev/null
#6462


### Q2: How many adult masks were left in New Taipei City on April 30,
2020?
$time hive -S -e 'use twmask; select substring(dt,12,8) as t,
sum(adult), sum(child) from twmparquet where substring(dt,
1,10)="2020/04/30" and substring(code,3,2)="31" group by substring(dt,
12,8) order by t ' 2>/dev/null
#23:59:33        1096807        44323
$echo $((1096807+443230))
#1540037


### Q3: How many adult masks were left in Bade District, Taoyuan City
at 8am on May 12, 2020?
$time hive -S -e "use twmask; select substring(dt,12,8) as t,
sum(adult), sum(child) from twmparquet where substring(code,
3,4)="3208" and substring(dt,1,10)='2020/05/12' group by substring(dt,
```

```
12,8) order by t" 2>/dev/null
#08:02:33        119098 23843
$echo $((119098+23843))
#142941


### Q4: How many masks were distributed in Tainan City on May 3, 2020?
$time hive -S -e 'use twmask; select sum(diff_adult), sum(diff_child)
from (select code,max(adult)-min(adult) as diff_adult, max(child)-
min(child) as diff_child from twmparquet where substring(dt,
1,10)="2020/05/03" and substring(code,3,2) in ("05","21","41") group
by code order by code) a' 2>/dev/null
#27810   3789
$echo $((27810+3789))
#31599


#########
# LINUX #
#########

$ssh bigred@adm100

### Q1: How many Drugstore were participating in distribution of
masks?
$cat /opt/dataset/twmask/03-05-2020-23-49.csv | tr -s "" | cut -d',' -
f1 | grep -e "^5" | uniq > dstore.txt

$nano dstore.sh
$chmod +x dstore.sh
$./dstore.sh

#!/bin/bash
files=/opt/dataset/twmask/*.csv
for f in $files;
do
  cat $f | tr -s "" | cut -d',' -f1 | grep -e "^5" | uniq > $f.txt
done
ds=/opt/dataset/twmask/*.csv.txt
for d in $ds;
do
  comm -13 <(sort dstore.txt) <(sort $d) >> dstore.txt;
done

$cat dstore.txt | sort | uniq | wc -l
#6462


### Q2: How many adult masks were left in New Taipei City on April 30,
2020?
```

```
$cat /opt/dataset/twmask/30-04-2020-23-47.csv | grep '新北市' | cut -
d',' -f5 | python -c "import sys; print(sum(int(l) for l in
sys.stdin))"
$cat /opt/dataset/twmask/30-04-2020-23-47.csv | grep '新北市' | cut -
d',' -f6 | python -c "import sys; print(sum(int(l) for l in
sys.stdin))"
#adult:1096816
#child:443220
$echo $((1096816+443220))
#1540036

#same as
$cat /opt/dataset/twmask/30-04-2020-23-47.csv | grep -e "^..31" | cut
-d',' -f5 | python -c "import sys; print(sum(int(l) for l in
sys.stdin))"
$cat /opt/dataset/twmask/30-04-2020-23-47.csv | grep -e "^..31" | cut
-d',' -f6 | python -c "import sys; print(sum(int(l) for l in
sys.stdin))"

###something worth noticing
$cat /opt/dataset/twmask/01-05-2020-00-02.csv | grep '新北市' | cut -
d',' -f5 | python -c "import sys; print(sum(int(l) for l in
sys.stdin))"
$cat /opt/dataset/twmask/01-05-2020-00-02.csv | grep '新北市' | cut -
d',' -f6 | python -c "import sys; print(sum(int(l) for l in
sys.stdin))"
#adult:1096807
#child:443230
$echo $((1096807+443230))
#1540037


### Q3: How many adult masks were left in Bade District, Taoyuan City
at 8am on May 12, 2020?
### 08:04am
$cat /opt/dataset/twmask/12-05-2020-08-04.csv | grep '桃園市八德區' |
cut -d',' -f5 | python -c "import sys; print(sum(int(l) for l in
sys.stdin))"
$cat /opt/dataset/twmask/12-05-2020-08-04.csv | grep '桃園市八德區' |
cut -d',' -f6 | python -c "import sys; print(sum(int(l) for l in
sys.stdin))"
#adult:119098
#child:23843
$echo $((119098+23843))
#142941

#same as
$cat /opt/dataset/twmask/12-05-2020-08-04.csv | grep -e "^..3208" |
cut -d',' -f5 | python -c "import sys; print(sum(int(l) for l in
```

```
sys.stdin))"
$cat /opt/dataset/twmask/12-05-2020-08-04.csv | grep -e "^..3208" |
cut -d',' -f6 | python -c "import sys; print(sum(int(l) for l in
sys.stdin))"
```