Hive Pig 实习报告

薛飞跃 1700017831

实验步骤

1. 按照指导安装 Hive Pig 并调试
2. 分别在 Hive 和 Pig 中导入数据，并完成相应的查询

实验过程

Hive：

1. 建表，并导入数据。

```
0: jdbc:hive2://localhost:10000> create table covid(Day string,Country string,Province string,Lat float,Long float,Confirmed int,Recovered int,Deaths int )
. . . . . .> ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
. . . . . .> WITH SERDEPROPERTIES (
. . . . . .>     "separatorChar" = ",",
. . . . . .>     "quoteChar" = "\"",
. . . . . .>     "escapeChar" = "\\"
. . . . . .> )
. . . . . .> tblproperties("skip.header.line.count"="1") ;
No rows affected (0.809 seconds)
0: jdbc:hive2://localhost:10000> load data local inpath '/usr/local/hive/apache-hive-2.3.6-bin/warehouse/time-series-19-covid-combined_csv.csv' overwrite into table covid;
No rows affected (1.988 seconds)
0: jdbc:hive2://localhost:10000> show table
table       table_name
0: jdbc:hive2://localhost:10000> show tables;
+-----------+
| tab_name  |
+-----------+
| covid     |
+-----------+
1 row selected (0.398 seconds)
```

2. 查询病毒在全球、中国湖北、中国大陆其它省份的病死率。

```
0: jdbc:hive2://localhost:10000> select sum(deaths)/sum(confirmed) from covid;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different exe
+---------------------+
|         _c0         |
+---------------------+
| 0.0564521118355406  |
+---------------------+
1 row selected (4.87 seconds)
0: jdbc:hive2://localhost:10000> select sum(deaths)/sum(confirmed) from covid where country='China' and province='Hubei';
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different exe
+----------------------+
|         _c0          |
+----------------------+
| 0.04206659913628231  |
+----------------------+
1 row selected (5.267 seconds)
0: jdbc:hive2://localhost:10000> select sum(deaths)/sum(confirmed) from covid where country='China' and province<>'Hubei';
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different exe
+----------------------+
|         _c0          |
+----------------------+
| 0.007507953842939335  |
+----------------------+
1 row selected (3.752 seconds)
```

3. 查询确诊人数超过 5000 的国家/地区中病死率前三的国家/地区。

```
0: jdbc:hive2://localhost:10000> select country,sum(deaths)/sum(confirmed) as deathrate  from covid
. . . . . .> group by country having sum(confirmed)>5000 order by deathrate desc limit 3;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a
+-------------+---------------------+
|   country   |      deathrate      |
+-------------+---------------------+
| Algeria     | 0.11279048656499636  |
| Italy       | 0.11262467597732054  |
| San Marino  | 0.101905311778291    |
+-------------+---------------------+
```

4.查询确诊人数超过 5000 的国家/地区中治愈率前三的国家/地区。

```
0: jdbc:hive2://localhost:10000> select country,sum(recovered)/sum(confirmed) as recoverrate  from covid
. . . . . .> group by country having sum(confirmed)>5000 order by recoverrate desc limit 3;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a di
+-------------------+---------------------+
|      country      |     recoverrate     |
+-------------------+---------------------+
| China             | 0.6397352747649925  |
| Bahrain           | 0.4862359231031737  |
| Diamond Princess  | 0.4051197398245787  |
+-------------------+---------------------+
```

Pig：

1. 加载数据。

```
grunt> covid = load '/usr/local/hive/apache-hive-2.3.6-bin/time-series-19-covid-combined_csv.csv'
using PigStorage(',') as (day:chararray,country:chararray,province:chararray,lat:float,long:float,
confirmed:int,recovered:int,deaths:int);
```

2. 查询病毒在全球的病死率。

```
temp = group covid all;
```

```
grunt> temp2 = foreach temp generate (float)SUM(covid.deaths)/SUM(covid.confirmed);
```

3. 查询病毒在中国湖北的病死率。

```
grunt> temp = filter covid by province == 'Hubei';
```

```
grunt> temp2 = group temp all;
```

```
(all,0.0420666)
grunt> temp3 = foreach temp2 generate group,(float)SUM(temp.$7)/SUM(temp.$5);
```

4. 查询病毒在中国大陆其它省份的病死率。

```
(all,0.007507954)
grunt> temp = filter covid by country=='China' and province != 'Hubei';
```

temp2 和 temp3 和第 3 步相同。

5. 按国家分组，计算每个国家的病死率和确诊人数，保留确诊人数超过 5000 的国家，按病死率降序排列，选择前三名。

```
grunt> temp = group covid by country;
```

```
grunt> temp1 = foreach temp generate group,(float)SUM(covid.$7)/SUM(covid.$5),SUM(covid.$5);
```

```
grunt> temp2 = filter temp1 by $2 >5000;
```

```
grunt> temp3 = order temp2 by $1 desc;
```

```
(Algeria,0.11279049,22032)
(Italy,0.112624675,2817704)
grunt> temp4 = limit temp3 3;
```

6. 查询确诊人数 5000 以上的治愈率前三的国家/地区。

```
grunt> temp1 = foreach temp generate group,(float)SUM(covid.$6)/SUM(covid.$5);
```

```
(China,0.6397353,5179408)
(Bahrain,0.48623592,17582)
grunt> dump temp4;
```

心得和总结：

1. 互相验证，pig 和 hive 查询的结果都正确且相同。
2. csv 是逗号分隔符文件，而每一列存储的字符可能本身包含逗号，如果仅仅用逗号分隔来读取文件，可能会导致一些问题。
3. hive 直接套用 sql 的语法，pig 的语法要特别注意一些细节，如等号前面必须有空格，

sum，count 等符号注意大小写，建表时，Date 是数据类型的保留字符，尽量避开用 Date 作为列名。