COVID19 Case Study

SCENARIO

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people who fall sick with COVID-19 will experience mild to moderate symptoms and recover without special treatment. Based on real data aggregated from different sources and countries, you are in charge to analyze this data and create a summary report.

To do this analysis we are going to use some Data at Scale common tools such as Hadoop, Pig, Hive and Impala.

Real COVID-19 data statistics is published every day (between 03:30 and 04:00 UTC) by the JHU (John Hopkins Universityi).

The data is published as a CSV file which contains more or less 3220 rows. • A second CSV file has been published earlier which contains lookup information such as the population of a country or a region.

For this case study, we will use COVID-19 data of May 13 2020 and lookup files into HDFS for further evaluation and processing, Clean, Normalize and filter the datasets using Pig Latin scripts, load the cleaned dataset into Hive and query it using Impala

```
%sh
##### TASK 01 ####
hdfs dfs -mkdir /COVID-19
```

FINISHED

Took 2 sec. Last updated by anonymous at May 18 2020, 9:43:11 PM.

```
%sh
#hdfs dfs -rm -r /COVID-LOOKUP
```

```
Deleted /COVID-LOOKUP
```

FINISHED

Took 2 sec. Last updated by anonymous at May 18 2020, 9:42:48 PM. (outdated)

```
%sh
hdfs dfs -ls /
```

FINISHED

```
Found 8 items
drwxr-xr-x   - root  supergroup          0 2020-05-18 18:43 /COVID-19
drwxrwxrwx   - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x   - hbase supergroup          0 2020-05-16 06:02 /hbase
drwxr-xr-x   - root  supergroup          0 2020-05-16 08:35 /piglab2
drwxr-xr-x   - solr  solr                0 2017-10-23 09:18 /solr
drwxrwxrwt   - hdfs  supergroup          0 2020-05-18 11:41 /tmp
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 09:17 /user
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 09:17 /var
```

Took 3 sec. Last updated by anonymous at May 18 2020, 9:43:16 PM.

```
%sh
hdfs dfs -put /home/cloudera/Downloads/05-13-2020.csv /COVID-19
```

FINISHED

Took 2 sec. Last updated by anonymous at May 18 2020, 9:43:21 PM.

```
%sh
hdfs dfs -put /home/cloudera/Downloads/UID_ISO_FIPS_LookUp_Table.csv /COVID-19
```

FINISHED

Took 2 sec. Last updated by anonymous at May 19 2020, 8:37:12 PM.

```
%sh
hdfs dfs -ls -h /COVID-19
```

FINISHED

```
Found 2 items
-rw-r--r--   1 root supergroup    327.4 K 2020-05-18 18:43 /COVID-19/05-13-2020.csv
-rw-r--r--   1 root supergroup    346.7 K 2020-05-19 17:37 /COVID-19/UID_ISO_FIPS_LookUp_Table.csv
```

Took 2 sec. Last updated by anonymous at May 19 2020, 8:37:18 PM.

```
%pig
--##### TASK 02 #####
-- list directories
fs -ls /
```

FINISHED

```
Found 8 items
drwxr-xr-x   - root  supergroup          0 2020-05-18 18:43 /COVID-19
drwxrwxrwx   - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x   - hbase supergroup          0 2020-05-16 06:02 /hbase
drwxr-xr-x   - root  supergroup          0 2020-05-16 08:35 /piglab2
drwxr-xr-x   - solr  solr                0 2017-10-23 09:18 /solr
drwxrwxrwt   - hdfs  supergroup          0 2020-05-18 11:41 /tmp
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 09:17 /user
drwxr-xr-x   - hdfs  supergroup          0 2017-10-23 09:17 /var
```

Took 0 sec. Last updated by anonymous at May 18 2020, 9:43:42 PM.

```
%pig
-- list items in COVID-19 directory
fs -ls -h /COVID-19
```

FINISHED

```
Found 2 items
-rw-r--r--   1 root supergroup    327.4 K 2020-05-18 18:43 /COVID-19/05-13-2020.csv
-rw-r--r--   1 root supergroup    346.7 K 2020-05-19 17:37 /COVID-19/UID_ISO_FIPS_LookUp_Table.csv
```

Took 0 sec. Last updated by anonymous at May 19 2020, 8:37:28 PM.

```
-- Load piggybank
register /usr/lib/pig/piggybank.jar;

define CSVLoader org.apache.pig.piggybank.storage.CSVLoader();

-- Load data
CovidRaw = Load '/COVID-19/05-13-2020.csv' using CSVLoader(',') as (FIPS:chararray, Admin2:chararray, Province_State:chararray,
    Country_Region:chararray, Last_Update:chararray,    Lat:double, Long_:double, Confirmed:int, Deaths:int, Recovered:int, Active:int,
    Combined_Key:chararray);

-- remove the header

Covid = Filter CovidRaw By not (FIPS == 'FIPS');
Covid5 = Limit Covid 5;
```

FINISHED

Took 0 sec. Last updated by anonymous at May 18 2020, 9:44:24 PM.

READY

```
%pig
-- Show 5 rows of data
DUMP Covid5;
```

```
(16001,Ada,Idaho,US,2020-05-14 03:32:28,43.4526575,-116.24155159999998,744,21,0,723,Ada, Idaho, US)
(19001,Adair,Iowa,US,2020-05-14 03:32:28,41.33075609,-94.47105874,4,0,0,4,Adair, Iowa, US)
(22001,Acadia,Louisiana,US,2020-05-14 03:32:28,30.2950649,-92.41419698,151,11,0,140,Acadia, Louisiana, US)
(45001,Abbeville,South Carolina,US,2020-05-14 03:32:28,34.22333378,-82.46170658,34,0,0,34,Abbeville, South Carolina, US)
(51001,Accomack,Virginia,US,2020-05-14 03:32:28,37.76707161,-75.63234615,545,7,0,538,Accomack, Virginia, US)
```

FINISHED

```
%pig
-- replace comma by dash
Covid_Replace_Comma = FOREACH Covid GENERATE Province_State, REPLACE(Country_Region,',',' ') as Country_Region, Last_Update, Lat, Long_, Confirmed, Deaths, Recovered,
    Active, REPLACE (Combined_Key,', ','-');
```

Took 0 sec. Last updated by anonymous at May 20 2020, 9:16:36 PM.

FINISHED

```
%pig
Covid_Replace_Comma5 = Limit Covid_Replace_Comma 5;
DUMP Covid_Replace_Comma5;
```

```
(Idaho,US,2020-05-14 03:32:28,43.4526575,-116.24155159999998,744,21,0,723,Ada-Idaho-US)
(Iowa,US,2020-05-14 03:32:28,41.33075609,-94.47105874,4,0,0,4,Adair-Iowa-US)
(Louisiana,US,2020-05-14 03:32:28,30.2950649,-92.41419698,151,11,0,140,Acadia-Louisiana-US)
(South Carolina,US,2020-05-14 03:32:28,34.22333378,-82.46170658,34,0,0,34,Abbeville-South Carolina-US)
(Virginia,US,2020-05-14 03:32:28,37.76707161,-75.63234615,545,7,0,538,Accomack-Virginia-US)
```

Took 43 sec. Last updated by anonymous at May 20 2020, 9:18:03 PM.

FINISHED

```
%pig
Covid_Clean_Output = FOREACH Covid GENERATE Province_State, Country_Region, Last_Update, Lat, Long_, Confirmed, Deaths, Recovered, Active;
```

Took 0 sec. Last updated by anonymous at May 20 2020, 9:18:54 PM.

```
%pig
-- Output the files in COVID-Clean directory
STORE Covid_Clean_Output INTO '/COVID-CLEAN/' USING PigStorage (',');
```

FINISHED

Took 17 sec. Last updated by anonymous at May 18 2020, 9:49:26 PM. (outdated)

```
%pig
-- Check that that file is saved
fs -ls -h /COVID-CLEAN/
```

FINISHED

```
Found 2 items
-rw-r--r--   1 root supergroup          0 2020-05-18 18:49 /COVID-CLEAN/_SUCCESS
-rw-r--r--   1 root supergroup    212.1 K 2020-05-18 18:49 /COVID-CLEAN/part-m-00000
```

Took 0 sec. Last updated by anonymous at May 18 2020, 9:49:36 PM.

```
%sh
#Output copy to local system
hdfs dfs -get '/COVID-CLEAN/' /home/cloudera/Documents
```

FINISHED

Took 3 sec. Last updated by anonymous at May 18 2020, 9:51:36 PM.

TASK 03

READY

```
%sh
##### TASK 03 #####
#Check file properly loaded
hdfs dfs -ls /COVID-19/
```

```
Found 2 items
-rw-r--r--   1 root supergroup     335286 2020-05-18 18:43 /COVID-19/05-13-2020.csv
-rw-r--r--   1 root supergroup     355025 2020-05-19 17:37 /COVID-19/UID_ISO_FIPS_LookUp_Table.csv
```

READY

```
%pig
-- Load piggybank
register /usr/lib/pig/piggybank.jar;

define CSVLoader org.apache.pig.piggybank.storage.CSVLoader();

-- Load data
LookUp_Table_Raw = Load '/COVID-19/UID_ISO_FIPS_LookUp_Table.csv' using CSVLoader(',') as (UID:int, iso2:chararray, iso3:chararray, code3:chararray, FIPS:chararray,
    Admin2:chararray, Province_State:chararray, Country_Region:chararray, Lat:double, Long_:double, Combined_Key:chararray, Population:int);

-- remove the header

LookUp_Table = Filter LookUp_Table_Raw By not (FIPS == 'FIPS');
```

FINISHED

```
%pig
-- replace comma by dash
LookUp_Table_dash = FOREACH LookUp_Table GENERATE UID, iso2, iso3, Province_State, REPLACE(Country_Region,',',' ') as Country_Region, Lat, Long_, Population,
    Combined_Key, REPLACE (Combined_Key,',','-');

LookUp_Table_dash_only = FILTER LookUp_Table_dash by (Combined_Key matches '.*-.*');
LookUp_Table_dash_only_5 = limit LookUp_Table_dash_only 5;
```

Took 1 sec. Last updated by anonymous at May 20 2020, 9:11:14 PM.

```
%pig
-- Check some rows where commas were replaced by dash
DUMP LookUp_Table_dash_only_5;
```

FINISHED

```
(72407,ES,ESP,Castilla - La Mancha,Spain,39.2796,-3.0977,2034877,Castilla - La Mancha, Spain,Castilla - La Mancha- Spain)
(84002105,US,USA,Alaska,US,58.29307365,-135.6424424,2148,Hoonah-Angoon, Alaska, US,Hoonah-Angoon- Alaska- US)
(84070020,US,USA,Utah,US,41.27116049,-111.9145117,272337,Weber-Morgan, Utah, US,Weber-Morgan- Utah- US)
(624,GW,GNB,,Guinea-Bissau,11.8037,-15.1804,1967998,Guinea-Bissau,Guinea-Bissau)
(626,TL,TLS,,Timor-Leste,-8.874217,125.727539,1318442,Timor-Leste,Timor-Leste)
```

Took 44 sec. Last updated by anonymous at May 18 2020, 9:53:40 PM.

```
%pig
-- Only keep data with UID with 3 digits
LookUp_Table_UID3 = FILTER LookUp_Table_dash by (UID <1000);

LookUp_Table_UID3_desc = ORDER LookUp_Table_UID3 by UID DESC;
LookUp_Table_UID3_desc_5 = LIMIT LookUp_Table_UID3_desc 5;
```

FINISHED

Took 0 sec. Last updated by anonymous at May 18 2020, 9:53:45 PM.

```
%pig
DUMP LookUp_Table_UID3_desc_5;
```

FINISHED

```
(894,ZM,ZMB,,Zambia,-13.133897,27.849332,18383956,Zambia,Zambia)
(887,YE,YEM,,Yemen,15.552727,48.516388,29825968,Yemen,Yemen)
(862,VE,VEN,,Venezuela,6.4238,-66.5897,28435943,Venezuela,Venezuela)
(860,UZ,UZB,,Uzbekistan,41.377491,64.585262,33469199,Uzbekistan,Uzbekistan)
(858,UY,URY,,Uruguay,-32.5228,-55.7658,3473727,Uruguay,Uruguay)
```

Took 1 min 24 sec. Last updated by anonymous at May 18 2020, 9:55:12 PM.

```
%pig
LookUp_Table_Clean = FOREACH LookUp_Table_UID3 GENERATE UID, iso2, iso3, Province_State, Country_Region, Lat, Long_, Population;
```

FINISHED

Took 0 sec. Last updated by anonymous at May 18 2020, 9:55:22 PM.

FINISHED

```
%pig
-- Output the file in COVID-Clean directory
STORE LookUp_Table_Clean INTO '/COVID-LOOKUP/' USING PigStorage (',');
```

Took 16 sec. Last updated by anonymous at May 18 2020, 9:55:46 PM.

FINISHED

```
%sh
#Output copy to local system
hdfs dfs -get '/COVID-LOOKUP/' /home/cloudera/Documents
```

Took 2 sec. Last updated by anonymous at May 18 2020, 9:56:01 PM.

FINISHED

```
%sh
#Check files in COVID-LOOKUP FOLDER
hdfs dfs -ls /COVID-LOOKUP/
```

```
Found 2 items
-rw-r--r--   1 root supergroup          0 2020-05-18 18:55 /COVID-LOOKUP/_SUCCESS
-rw-r--r--   1 root supergroup      10438 2020-05-18 18:55 /COVID-LOOKUP/part-m-00000
```

Took 2 sec. Last updated by anonymous at May 18 2020, 9:56:05 PM.

FINISHED

```
%sh
##### TASK 04 #####
hdfs dfs -ls /COVID-CLEAN/
```

```
Found 2 items
-rw-r--r--   1 root supergroup          0 2020-05-18 18:49 /COVID-CLEAN/_SUCCESS
-rw-r--r--   1 root supergroup     217207 2020-05-18 18:49 /COVID-CLEAN/part-m-00000
```

Took 2 sec. Last updated by anonymous at May 18 2020, 9:56:10 PM.

FINISHED

```
%hive
-- Create Database
CREATE DATABASE Covid
```

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at May 18 2020, 10:00:50 PM.

FINISHED

```
%hive
-- Use Covid Database
use Covid
```

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at May 18 2020, 10:00:52 PM.

FINISHED

```
%hive
-- Use to drop table as needed
DROP TABLE Covid
```

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at May 18 2020, 10:00:55 PM.

```
%hive
--Create Hive non-managed table for COIVD

create external table Covid(
    Province_State string,
    Country_Region string,
    Last_Update string,
    Lat double,
    Long_ double,
    Confirmed int,
    Deaths int,
    Recovered int,
    Active int)
    ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE
    Location '/COVID-CLEAN/'
    TBLPROPERTIES ("skip.header.line.count" = "0")
```

FINISHED

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at May 18 2020, 10:01:22 PM.

```
%hive
select * from Covid limit 5
```

FINISHED

settings

covid.province_state	covid.country_region	covid.last_update	covid.lat	covid.long_	covid.confirmed	covid.deaths	covid.recovered	covid.active
South Carolina	US	2020-05-14 03:32:28	34.22333378	-82.46170658	34	0	0	34
Louisiana	US	2020-05-14 03:32:28	30.2950649	-92.41419698	151	11	0	140
Virginia	US	2020-05-14 03:32:28	37.76707161	-75.63234615	545	7	0	538
Idaho	US	2020-05-14 03:32:28	43.4526575	-116.24155159999998	744	21	0	723
Iowa	US	2020-05-14 03:32:28	41.33075609	-94.47105874	4	0	0	4

Took 0 sec. Last updated by anonymous at May 18 2020, 10:01:26 PM.

```
-- Count number of rows
select Count(*) from Covid
```

FINISHED

settings

_c0
3239

Took 19 sec. Last updated by anonymous at May 18 2020, 10:01:53 PM.

```
%hive
--Display table properties and size
show create table Covid
```

FINISHED

settings

createtab_stmt
'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION
'hdfs://quickstart.cloudera:8020/COVID-CLEAN'
TBLPROPERTIES (
'COLUMN_STATS_ACCURATE'='false',
'numFiles'='1',
'numRows'='-1',
'rawDataSize'='-1',

Took 0 sec. Last updated by anonymous at May 18 2020, 10:03:16 PM.

FINISHED

```
%sh
# Size of table on HDFS
hdfs dfs -ls '/COVID-CLEAN/' #to match statement above
hdfs dfs -ls -h '/COVID-CLEAN/'
```

```
Found 2 items
-rw-r--r--   1 root supergroup          0 2020-05-18 18:49 /COVID-CLEAN/_SUCCESS
-rw-r--r--   1 root supergroup     217207 2020-05-18 18:49 /COVID-CLEAN/part-m-00000
Found 2 items
-rw-r--r--   1 root supergroup          0 2020-05-18 18:49 /COVID-CLEAN/_SUCCESS
-rw-r--r--   1 root supergroup    212.1 K 2020-05-18 18:49 /COVID-CLEAN/part-m-00000
```

Took 4 sec. Last updated by anonymous at May 18 2020, 10:03:31 PM.

FINISHED

```
%hive
DROP TABLE lookup
```

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at May 18 2020, 10:03:31 PM.

```
%hive
--TASK 05 - Create Hive non-managed table for lookup
create external table lookup(
    UID int,
    iso2 string,
    iso3 string,
    Province_State string,
    Country_Region string,
    Lat double,
    Long_ double,
    Population double)
    ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    Location '/COVID-LOOKUP/'
    TBLPROPERTIES ("skip.header.line.count" = "0")
```

FINISHED

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at May 18 2020, 10:03:35 PM.

```
%hive
select * from lookup limit 5
```

FINISHED

settings

lookup.uid	lookup.iso2	lookup.iso3	lookup.province_state	lookup.country_region	lookup.lat	lookup.long_	lookup.population
4	AF	AFG		Afghanistan	33.93911	67.709953	38928341
8	AL	ALB		Albania	41.1533	20.1683	2877800
12	DZ	DZA		Algeria	28.0339	1.6596	43851043
20	AD	AND		Andorra	42.5063	1.5218	77265
24	AO	AGO		Angola	-11.2027	17.8739	32866268

Took 0 sec. Last updated by anonymous at May 18 2020, 10:03:39 PM.

```
%hive
-- Count number of rows
select Count(*) from lookup
```

FINISHED

settings

_c0
217

Took 20 sec. Last updated by anonymous at May 18 2020, 10:04:11 PM.

```
%hive
--Display table properties and size
show create table lookup
```

FINISHED

settings

createtab_stmt
TBLPROPERTIES (
'COLUMN_STATS_ACCURATE'='false',
'numFiles'='1',
'numRows'='-1',
'rawDataSize'='-1',
'skip.header.line.count'='0',
'totalSize'='10438',
'transient_lastDdlTime'='1589853815')

Took 0 sec. Last updated by anonymous at May 18 2020, 10:05:17 PM.

```
%sh
# Size of table on HDFS
hdfs dfs -ls '/COVID-LOOKUP/' #to match statement above
hdfs dfs -ls -h '/COVID-LOOKUP/'
```

FINISHED

```
Found 2 items
-rw-r--r--   1 root supergroup          0 2020-05-18 11:41 /COVID-LOOKUP/_SUCCESS
-rw-r--r--   1 root supergroup      10438 2020-05-18 11:41 /COVID-LOOKUP/part-m-00000
Found 2 items
-rw-r--r--   1 root supergroup          0 2020-05-18 11:41 /COVID-LOOKUP/_SUCCESS
-rw-r--r--   1 root supergroup     10.2 K 2020-05-18 11:41 /COVID-LOOKUP/part-m-00000
```

Took 4 sec. Last updated by anonymous at May 18 2020, 3:19:45 PM.

```
%hive
--TASK 06 -
DROP TABLE covid_part
```

FINISHED

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at May 18 2020, 3:19:49 PM. (outdated)

```
%hive
Create Hive managed table for COIVD


create table covid_part(
    Province_State string,
    Last_Update string,
    Lat double,
    Long_ double,
    Confirmed int,
    Deaths int,
    Recovered int,
    Active int)
    partitioned by (Country_Region string) CLUSTERED BY (Province_State) SORTED BY (Confirmed) into 4 buckets
    ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE
    TBLPROPERTIES ("skip.header.line.count" = "0")
```

FINISHED

```
Query executed successfully. Affected rows : -1
```

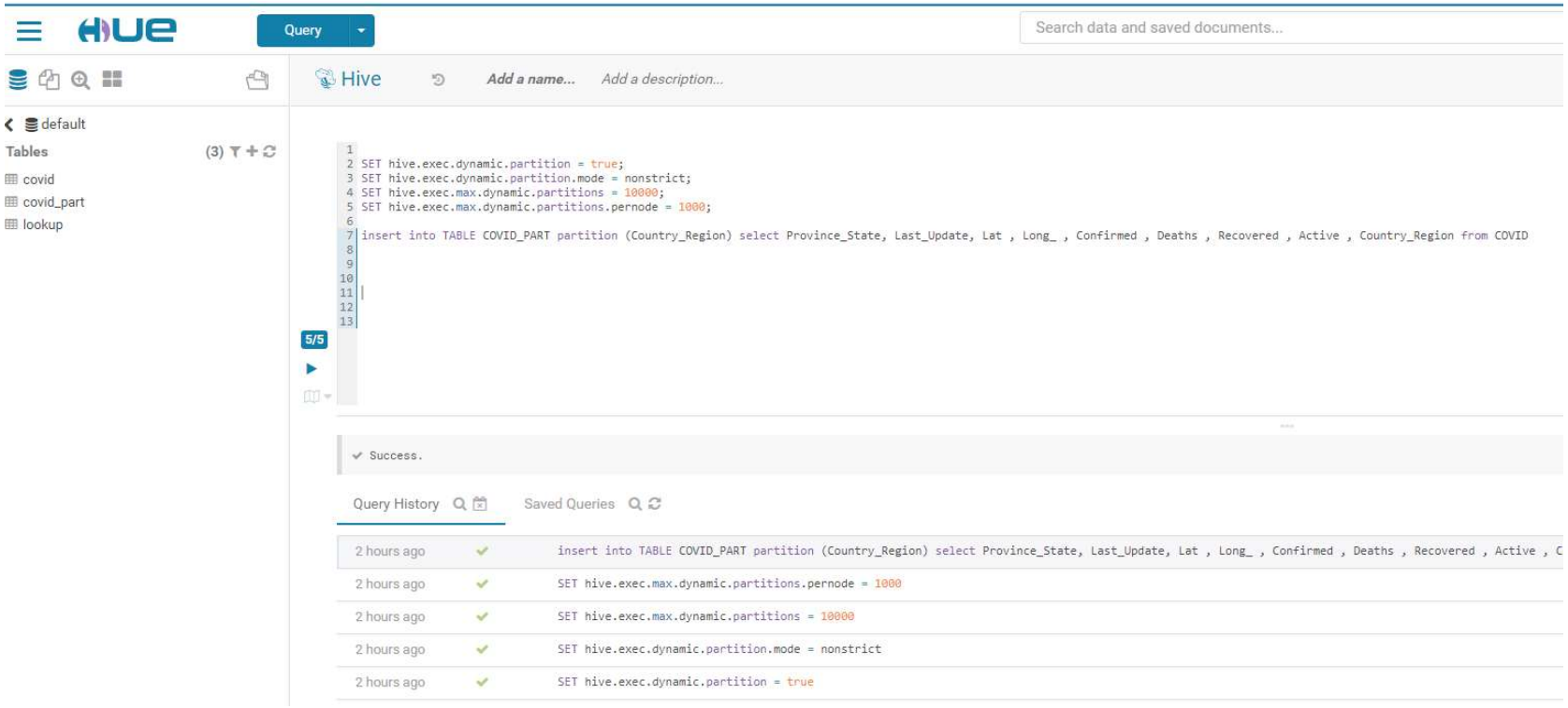
Took 0 sec. Last updated by anonymous at May 18 2020, 10:05:34 PM. (outdated)

```
%hive
show partitions COVID_PART
```

FINISHED

settings

partition



- **Comments:** We had to run the code in Hue as it would not generate any output in zeppelin.

```
%hive
-- Need to run this statement in HUE
SET hive.exec.dynamic.partition = true;
SET hive.exec.dynamic.partition.mode = nonstrict;
SET hive.exec.max.dynamic.partitions = 10000;
SET hive.exec.max.dynamic.partitions.pernode = 1000;

insert into TABLE COVID_PART partition (Country_Region) select Province_State, Last_Update, Lat , Long_ , Confirmed , Deaths , Recovered , Active ,
    Country_Region from Covid.COVID
```

FINISHED

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at May 18 2020, 10:05:40 PM.

```
%hive
show partitions covid.COVID_PART
```

FINISHED

settings

partition
country_region= Sint Eustatius and Saba
country_region=Afghanistan
country_region=Albania
country_region=Algeria
country_region=Andorra
country_region=Angola
country_region=Antigua and Barbuda
country_region=Argentina

Took 0 sec. Last updated by anonymous at May 18 2020, 10:07:17 PM.

```
%hive
-- Count number of rows
select Count(*) from covid_part
```

FINISHED

settings

_c0
3239

Took 20 sec. Last updated by anonymous at May 18 2020, 10:07:42 PM.

```
%hive
--##### TASK 07 #####
drop view v_lookup
```

FINISHED

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at May 18 2020, 10:07:49 PM. (outdated)

```
%hive
CREATE VIEW v_lookup AS select max(uid) as uid,max(iso2) as iso2,max(iso3) as iso3,max(province_state) as province_state,country_region, max(lat) as lat, max(long_) as long_, sum(population) as population from lookup group by country_region
```

FINISHED

```
Query executed successfully. Affected rows : -1
```

Took 0 sec. Last updated by anonymous at May 18 2020, 10:07:52 PM.

```
%hive
select * from v_lookup limit 5
```

FINISHED

settings

v_lookup.uid	v_lookup.iso2	v_lookup.iso3	v_lookup.province_state	v_lookup.country_region	v_lookup.lat	v_lookup.long_	v_lookup.population
535	BQ	BES	Bonaire	Sint Eustatius and Saba	null	12.1784	-68.2385
4	AF	AFG		Afghanistan	33.93911	67.709953	38928341
8	AL	ALB		Albania	41.1533	20.1683	2877800
12	DZ	DZA		Algeria	28.0339	1.6596	43851043
20	AD	AND		Andorra	42.5063	1.5218	77265

Took 18 sec. Last updated by anonymous at May 18 2020, 10:08:14 PM.

```
%hive
-- Number of rows
select Count(1) from v_lookup
```

READY

settings

_c0
186

```
%sh
 hdfs dfs -ls /user/hive/warehouse/
```

READY

```
Found 2 items
drwxrwxrwx   - hive supergroup          0 2020-05-18 19:05 /user/hive/warehouse/covid.db
drwxrwxrwx   - hive supergroup          0 2020-05-18 12:26 /user/hive/warehouse/covid_part
```

```
%impala
-- ##### TASK 08 ######

invalidate metadata
```

FINISHED

```
Query executed successfully. Affected rows : -1
```

Took 5 sec. Last updated by anonymous at May 18 2020, 10:09:17 PM. (outdated)

```
%impala
Show tables in Covid
```

FINISHED

settings

name
covid
covid_part
lookup
v_lookup

Took 0 sec. Last updated by anonymous at May 18 2020, 10:09:19 PM.

```
%impala
select SUM(CONFIRMED) as total_CONFIRMED, SUM(DEATHS) as total_deaths, SUM(ACTIVE) as total_active from Covid.Covid
```

FINISHED

settings

total_confirmed	total_deaths	total_active
4336021	307934	2515075

Took 10 sec. Last updated by anonymous at May 18 2020, 10:09:35 PM.

```
%impala

-- Report top 10 confirmed with % of total confirmed

select A.COUNTRY_REGION , A.TOTAL_CONFIRMED_Region, A.TOTAL_CONFIRMED_Region/B.TOTAL_CONFIRMED*100  as Percentage

FROM
(select COUNTRY_REGION , SUM(CONFIRMED) AS TOTAL_CONFIRMED_Region, SUM(DEATHS) AS TOTAL_DEATHS_Region, SUM(ACTIVE) AS TOTAL_ACTIVE_Region from
    covid.covid GROUP BY COUNTRY_REGION) A,
(SELECT SUM(CONFIRMED) AS TOTAL_CONFIRMED, SUM(DEATHS) AS TOTAL_DEATHS, SUM(ACTIVE) AS TOTAL_ACTIVE  from covid.covid) B

ORDER BY  Percentage desc nulls last
```

FINISHED

settings

country_region	total_confirmed_region	percentage
US	1390406	32.06640373743577
Russia	242271	5.587403751042719
United Kingdom	230985	5.327119033786968
Spain	228691	5.274213385959155
Italy	222104	5.122299915060374
Brazil	190137	4.385057175691723
France	178184	4.109389691608966
Germany	174098	4.015155830656724

Took 0 sec. Last updated by anonymous at May 18 2020, 10:09:42 PM.

```
%impala
-- Report top 10 deaths with % of total deaths

select A.COUNTRY_REGION , A.TOTAL_Deaths_Region, A.TOTAL_Deaths_Region/B.TOTAL_Deaths*100  as Percentage

FROM
(select COUNTRY_REGION , SUM(CONFIRMED) AS TOTAL_CONFIRMED_Region, SUM(DEATHS) AS TOTAL_DEATHS_Region, SUM(ACTIVE) AS TOTAL_ACTIVE_Region from
    covid.covid GROUP BY COUNTRY_REGION) A,
(SELECT SUM(CONFIRMED) AS TOTAL_CONFIRMED, SUM(DEATHS) AS TOTAL_DEATHS, SUM(ACTIVE) AS TOTAL_ACTIVE  from covid.covid) B

ORDER BY  Percentage desc nulls last
```

FINISHED

settings

country_region	total_deaths_region	percentage
US	84119	27.317217325790594
United Kingdom	33264	10.802314781738943
Italy	31106	10.101515259763456
Spain	27104	8.801886118453954
France	27077	8.793118005806438
Brazil	13240	4.299622646411244
Korea	10991	3.5692713373645004
Belgium	8843	2.871719264517721

Took 1 sec. Last updated by anonymous at May 18 2020, 10:09:54 PM.

```
%impala
-- Report top 10 actives with % of total actives

select A.COUNTRY_REGION , A.TOTAL_Active_Region, A.TOTAL_Active_Region/B.TOTAL_Active*100  as Percentage

FROM
(select COUNTRY_REGION , SUM(CONFIRMED) AS TOTAL_CONFIRMED_Region, SUM(DEATHS) AS TOTAL_DEATHS_Region, SUM(ACTIVE) AS TOTAL_ACTIVE_Region from
    covid.covid GROUP BY COUNTRY_REGION) A,
(SELECT SUM(CONFIRMED) AS TOTAL_CONFIRMED, SUM(DEATHS) AS TOTAL_DEATHS, SUM(ACTIVE) AS TOTAL_ACTIVE  from covid.covid) B

ORDER BY  Percentage desc nulls last
```

FINISHED

settings

country_region	total_active_region	percentage
US	1067859	42.458336232517915
United Kingdom	196689	7.820402970090355
Russia	192056	7.636193751677385
Brazil	98473	3.9153106766199812
France	92321	3.670705644960886
Italy	78457	3.1194695983221177
Spain	60764	2.4159915708279076
Peru	49813	1.9805771199666014

Took 0 sec. Last updated by anonymous at May 18 2020, 10:11:20 PM.

```
%impala
-- Death rate by population for top 10

SELECT T.COUNTRY_REGION, T.TOTAL_DEATHS,V_LOOKUP.TOTAL_POPULATION ,(T.TOTAL_DEATHS/V_LOOKUP.TOTAL_POPULATION)* 100 DEATHS_PERCENTAGE

FROM
(SELECT COUNTRY_REGION , SUM(CONFIRMED) AS TOTAL_CONFIRMED, SUM(DEATHS) AS TOTAL_DEATHS, SUM(ACTIVE) AS TOTAL_ACTIVE from covid.covid GROUP BY COUNTRY_REGION) T,
(SELECT COUNTRY_REGION, SUM(POPULATION) AS TOTAL_POPULATION FROM COVID.LOOKUP GROUP BY COUNTRY_REGION) V_LOOKUP

WHERE T.COUNTRY_REGION = V_LOOKUP.COUNTRY_REGION
ORDER BY  DEATHS_PERCENTAGE DESC NULLS LAST
LIMIT 10
```

FINISHED

settings

country_region	total_deaths	total_population	deaths_percentage
San Marino	41	33938	0.12080853320761388
Belgium	8843	11589616	0.07630106122584217
Andorra	49	77265	0.06341810651653401
Spain	27104	46754783	0.0579705396130274
Italy	31106	60461828	0.051447336325987365
United Kingdom	33264	68225159	0.0487562073691906
France	27077	68136441	0.03973938116315762
Sweden	3460	10099270	0.034259901953309496

Took 1 sec. Last updated by anonymous at May 20 2020, 9:21:18 PM.