# [https://avatars2.githubusercontent.com/u/4156894?v=3&s=100](http://www.calstatela.edu/centers/hipic) CIS5560 Term Project Tutorial

#### Authors: Denis Akbari, Javier Gutierrez, Esdras Solorzano

#### Instructor: [Jongwook Woo](https://www.linkedin.com/in/jongwook-woo-7081a85)

#### Date: 05/22/2022

**Lab Tutorial**

Denis Akbari, Javier Gutierrez, Esdras Solorzano ([dakbari@calstatela.edu](mailto:dakbari@calstatela.edu), [esolorz9@calstatela.edu](mailto:esolorz9@calstatela.edu), [jgutie167@calstatela.edu](http://jgutie167@calstatela.edu))

05/22/2022

**Hive South Korea Covid-19 cases analysis using Hadoop Amazon Cluster**

**Objectives**

In this hands-on lab, you will learn how to:

* Get data manually from Kaggle using cURL and uploading it into Hadoop
* Create HDFS directories
* Create external Hive tables
* Downloading the file locally to your computer
* Tempo-spatial visualizations with Excel 3D Map to gain various insights while analyzing

**Platform Spec**

* Cluster Version: Hadoop 3.2.2-amzn-3.1
* CPU Speed: 2399.892 MHz
* # of nodes: 2
* Total Memory Size: 33G

Step 1: Upload data into Hadoop

This step is to get data manually from Kaggle into the Hadoop Amazon Cluster, create directories and putting the files into those directories:

1. Open a shell terminal and paste *ssh* command to connect to the Hadoop Amazon Cluster:

ssh username@54.218.99.3

1. Download data from Kaggle and upload data files to OneDrive
2. Copy cURL from OneDrive
3. Go to your terminal and load data into Hadoop by pasting cURL. Use this command to check if the file was uploaded:

ls -al

A picture containing graphical user interface

Description automatically generated

1. Make directory timeage in Hadoop:

hdfs dfs -mkdir tmp

hdfs dfs -mkdir tmp/timeage

1. Put file TimeAge.csv in created directory tmp/timeage:

hdfs dfs -put TimeAge.csv tmp/timeage

hdfs dfs -ls tmp/timeage



1. Make directory TimeGender in Hadoop:

hdfs dfs -mkdir tmp/timegender

1. Put file TimeGender.csv in directory tmp/timegender:

hdfs dfs -put TimeGender.csv tmp/timegender

hdfs dfs -ls tmp/timegender



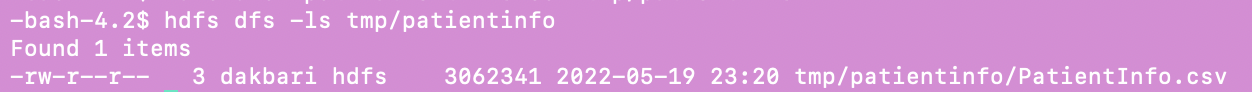
1. Make directory patientinfo:

hdfs dfs -mkdir tmp/patientinfo

1. Put file PatientInfo.csv in directory tmp/patientinfo:

hdfs dfs -put PatientInfo.csv tmp/patientinfo

hdfs dfs -ls tmp/patientinfo



Step 2: Create three Hive tables

This step is to create three external tables in Hive from data stored in HDFS. External tables preserve data in the original file format, while allowing Hive to have access to the data within the file. The Hive statements below create three new tables, named timeage, timegender and patientinfo,

1. Open **beeline CLI** (Command Line Shell Interface) that is equivalent to hive CLI environment as follows:

beeline

1. Delete the database if it already exists. Create your database and use that database for the tables you are going to create:

drop database yourDBName CASCADE;

create database yourDBName;

use yourDBName;

Text

Description automatically generated

1. Drop table If exists and create a timeage table:

DROP TABLE IF EXISTS TimeAge;

CREATE EXTERNAL TABLE IF NOT EXISTS timeage(Curdate STRING, Curtime INT, age STRING, confirmed INT, deceased INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION 'tmp/timeage' TBLPROPERTIES('skip.header.line.count'='1');

show tables;

Graphical user interface

Description automatically generated

1. Then select the first 10 rows to show:

select \* from TimeAge LIMIT 10;

A picture containing table

Description automatically generated

1. Use your database, drop table If exists and create a timegender table:

use yourDBName;

DROP TABLE IF EXISTS TimeGender;

CREATE EXTERNAL TABLE IF NOT EXISTS timegender (Curdate STRING, Curtime INT, sex STRING, confirmed INT, deceased INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION 'tmp/timegender' TBLPROPERTIES('skip.header.line.count'='1');

show tables;

Graphical user interface

Description automatically generated

1. Then select the first 10 rows to show:

select \* from TimeGender LIMIT 10;

A picture containing table

Description automatically generated

1. Use your database, drop table If exists and create a patientinfo table:

use yourDBName;

DROP TABLE IF EXISTS PatientInfo;

CREATE EXTERNAL TABLE IF NOT EXISTS patientinfo (patientId INT, sex STRING, age INT, country STRING, province STRING, city STRING, infection\_case STRING, infected\_by INT, contact\_number INT, symptom\_onset\_date STRING, confirmed\_date STRING, released\_date STRING, deceased\_date STRING, state STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION 'tmp/patientinfo' TBLPROPERTIES('skip.header.line.count'='1');

show tables;

A picture containing graphical user interface

Description automatically generated

1. Then select the first 10 rows to show:

select \* from PatientInfo LIMIT 10;

Graphical user interface

Description automatically generated

1. Open another terminal with git bash in order to import the output file using your lab computer. You have to download the file to your lab computer. For example your output file at the Amazon Cluster server is located at /home/username/PatientInfo.csv and remotely copied to the file “000000\_0.csv”:

scp username@54.218.99.3:/home/username/PatientInfo.csv 000000\_0.csv

Step 3: Tempo-Spatial Visualizations

This step is to make tempo-spatial visualizations to gain various insights while analyzing the Covid-19 cases in South Korea.

1. To visualize location type of results on map, convert csv file to excel and click on 3D map button under insert tab:

Graphical user interface, application, table, Excel

Description automatically generated

1. Have **Location** set to ‘city’, **Category** set to ‘age’, **Time** set to ‘confirmed\_date’. This map tells us the number of people in their age group who had confirmed cases of Covid-19 in certain cities in South Korea.

A screenshot of a computer

Description automatically generated with medium confidence

1. Keep this graph and under **Time** click to **Filters** and in **Count of Age** select the range of top 5. This graph tells us the top 5 age ranges with the most confirmed Covid-19 cases in South Korea:

Map

Description automatically generated

1. Have **Location** set to ‘city’, **Category** set to ‘age’, **Time** set to ‘deceased\_date’. This map tells us the number of people in their age range that have passed away from Covid-19 in certain cities in South Korea:

Map

Description automatically generated

1. Keep this graph and under **Time** click to **Filters** and in **Count of Age** select the range of top 2. This graph tells us the top 2 age ranges with the most deceased Covid-19 cases in South Korea:

A screenshot of a map

Description automatically generated with medium confidence

References

* 1. Data Source:

[1] Data Science for COVID-19. Retrieved from <https://www.kaggle.com/kimjihoo/coronavirusdataset?fbclid=IwAR1V1oLbBY3gako_2s27chpybGrF8l5eH7NWyj685Ol9BKNyqJ0Ym-YCThw>

[2]COVID-19 Open Research Dataset Challenge. Retrieved from <https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge?select=metadata.csv>

* 1. Github: <https://github.com/dakbari2/Group5Project>
  2. References:

[1] COVID-19 Explained through Visualizations Retrieved from <https://www.kaggle.com/code/anshuls235/covid19-explained-through-visualizations>

[2] Eemerging COVID-19 success story: South Korea learned their lessons of MERS. Retrieved from <https://ourworldindata.org/covid-exemplar-south-korea>

[3] South Korea: Coronavirus Pandemic Country Profile Retrieved from <https://ourworldindata.org/coronavirus/country/south-korea>

[4] Understanding South Korea’s Response to the COVID-19 Outbreak: A Real-Time Aanalysis. Retreieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7766828/>

[5] Coronavirus Disease-19: The First 7,755 Cases in the Republic of Korea. Retreieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7104685/>

[6] How One Country is Beating Covid Despite 600,000 New Cases a Day. Retreieved from <https://www.bloomberg.com/news/articles/2022-03-17/how-south-korea-is-beating-covid-despite-600-000-new-cases-a-day>