**Hive South Korea Covid-19 cases analysis**

**using Hadoop Amazon Cluster**

Authors: Denis Akbari, Javier Gutierrez, Esdras Solorzano

Department of Information Systems, California State University Los Angeles

CIS4560-01 Introduction to Big Data

([dakbari@calstatela.ed](http://dakbari@calstatela.edu)u, [jgutie167@calstatela.edu](mailto:jgutie167@calstatela.edu), [esolorz9@calstatela.edu](mailto:esolorz9@calstatela.edu))

**Abstract:** This paper will explain the process used for Hive South Korea Covid-19 cases analysis using Hadoop Amazon Cluster.

The project's main focus is to understand handling big data files and data cleaning using Hadoop Amazon Cluster.

Analysis of this data will be executed with the help of Microsoft Excel, which will be depicting of 3D Map visuals.

1. **Introduction**

This project uses Hadoop and Beeline to process the South Korea Covid-19 cases dataset. The dataset we will be using will be age, gender, provinces, status, etc.

The dataset was chosen for this data because of the recent pandemic. Approximately 10,000 South Koreans have been infected by COVID-19. The KCDC (Korea Centers for Disease Control & Prevention) releases COVID-19 information rapidly and transparently. The data was extracted from research from the KCDC and local governments. Based on KCDC and local government reports, we create a structured dataset. We analyze and visualize the data using data mining and visualization techniques. The dataset is unique as we’re looking into the data of cases of Covid-19 in South Korea. It's relevant because we need a dataset that can be sent into HDFS and Beeline and finally into a visual representation. We created a visualization of what age groups were infected the most and the number of people who died in those age groups.

1. **Related Work**

Kaggle is a website known for its datasets. Within also exists other datasets, but in this project, the dataset specifically chosen is the South Korea Covid-19 cases other datasets based on Covid-19 cases also exist, varying from cases from different countries.

One dataset that is a little similar to this project is titled Covid-19 Explained through Visualizations.[1] This dataset focuses on other countries and their Covid-19 cases rise and fall and demonstrates using only visuals of graphs and charts but tries to predict the cases and fatalities for the month of April 2020 by using the data of the three months before April. Some visuals only focus on one country, those countries being the United States of America, India, Italy, Spain, Chine, and South Korea, and its rise of cases and death counts while they also have other visuals where they have the top 10 countries with their rise in cases. Just like in our project we’ll be analyzing our tempo spacial visual.

Another dataset similar to this project is titled Covid-19 - Analysis, Visualization & Comparisons.[2] This dataset also focuses on Covid-19 cases in the rise and falls using bars, maps, and bubble plot points all coming from around the world and using all that data into one visual. Just like our project, we’ll be analyzing only the data set in South Korea and the tempo spacial visual.

1. **Specifications**

Each file consists of data such as the Patient's information, gender, location, dates, etc. Data collected for the datasets are based on the KCDC. The dataset is 10.47GB. Table 1 shows the files and size of the files of the dataset.

*Table 1 Data Specification*

|  |  |
| --- | --- |
| Data Set | Size (10.47GB) |
| TimeAge.csv | 27.11KB |
| Region.csv | 19.08KB |
| Case.csv | 11.71KB |
| cord\_19\_embeddings\_2021-05-31.csv | 8.84GB |
| Metadata.csv | 1.57GB |
| PatientInfo.csv | 488.86KB |
| Policy.csv | 5.71KB |
| SearchTrend.csv | 71.72KB |
| SeoulFloating.csv | 49.68MB |
| Time.csv | 6.6KB |
| TimeGender.csv | 6.97KB |
| TimeProvince.csv | 92.03KB |
| Weather.csv | 1.5MB |

The table below shows the specifications for the Oracle cluster used in the Hadoop specification in the project.

*Table 2 H/W Specification*

|  |  |
| --- | --- |
| Cluster Version | Hadoop 3.2.2-amzn-3.1 |
| Cluster Number of Nodes | 3 |
| Memory Size | 33GB |
| CPU Speed | 2399.892MHz |

1. **Implementation Flowchart**

The raw dataset used for this project was downloaded from a trusted source. The process is shown in the flowchart below (Figure 1). Thirteen .csv files are uploaded to HDFS. Beeline is used to create the tables, clean data, sort out the info, and export the results. Once an output file has been downloaded, we can open the file in an Excel 3D map.

Diagram

Description automatically generated

*Figure 1 – Implementation Flowchart*

1. **Data Cleaning**

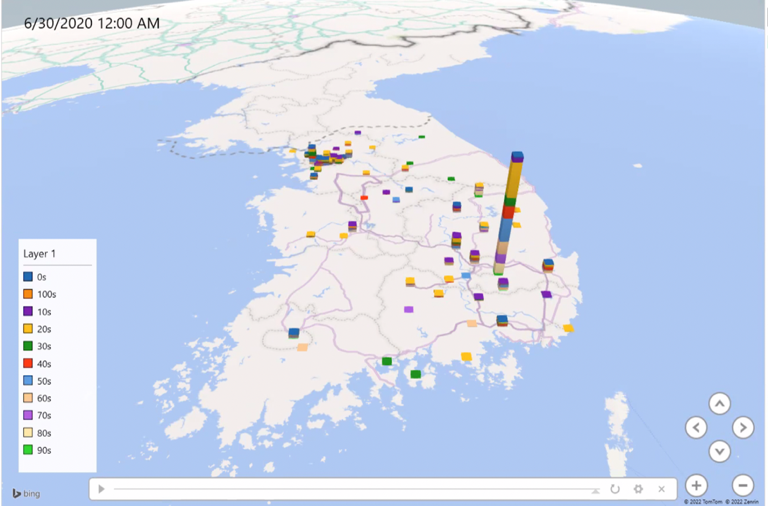
Raw files were uploaded and stored into HDFS and then loaded onto Beeline. All files carry different information that was separately cleaned and then exported through execution for results. The data found on Kaggle was mostly cleaned data already. Therefore, the data cleaning step of thye process was not a long one.

1. **Analysis and Visualization**

Once data cleaning was finished, files were extracted into Excel. We used different interactive maps in Excel in order to show how Covid affected different cities and age ranges as well as confirmed cases and confirmed cases marked as deceased. The Excel 3D map helped with tempo-spatial visualizations and four visualizations were done to gain more insight on the data while analyzing.

**6.1 3D Map Confirmed Cases Tempo-Spatial Visualization**

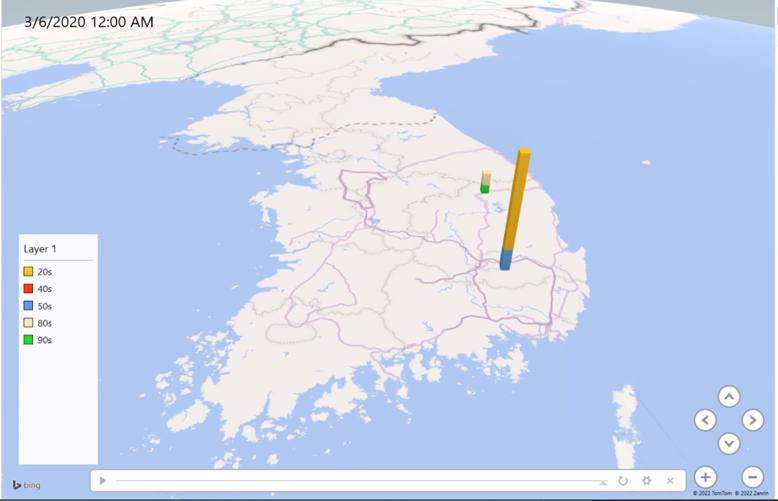
From this map (Figure 2), we can see the number of people in their age group who had confirmed cases of Covid-19 in specific cities[[1]](#footnote-2) in South Korea. In the Excel 3D Map, *Location* was set to ‘city’, *Category* was set to ‘age’, and *Time* was set to ‘confirmed\_date’. This map tells us the number of people in their age group who had confirmed cases of Covid-19 in South Korea.



*Figure 2 – Confirmed Covid Cases by City*

**6.2 3D Map Top 5 Confirmed Cases by Age Ranges**

From this map (Figure 3), we can see the top 5 age ranges with the most confirmed Covid-19 cases in South Korea. We can conlude that the age ranges 20s, 40s, 50s, 80s, and 90s are the top 5.

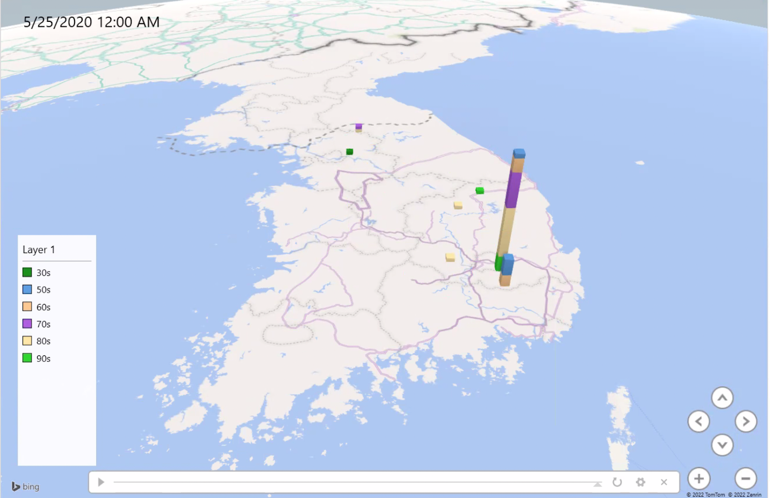


*Figure 3 – Confirmed Covid Age Ranges*

**6.3 3D Map Deceased Cases Visualization**

In this Excel 3D Map (Figure 4), we have the *Location* set to ‘city,’ *Category* set to ‘age,’ and *Time* set to ‘deceased\_date.’

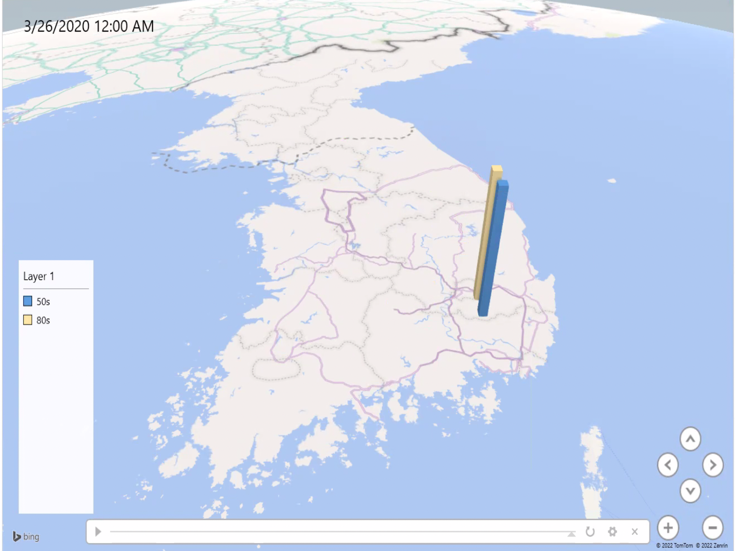
This map shows the number of people in their age range that have passed away from Covid-19. Those age ranges are go from the 30s - 90s.



*Figure 4 – Confirmed Deceased by Age*

**6.4 3D Map Top 2 Deceased Cases’ Age Ranges**

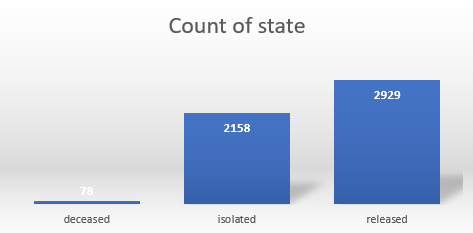
From this map (Figure 5), we see the top 2 age ranges with the most deceased Covid-19 cases in South Korea. From our previous knowldege, although the age range of 20s was between the most confirmed cases that did not mean that it was in the top 2 of deceased cases. Indeed, the age ranges of 50s and 80s are the top 2 age ranges with the most deceased Covid-19 cases in South Korea. With this information we can confirm that age plays a big role in the number of confirmed deceased cases.



*Figure 5 – Top 2 Deceased by Age Ranges*

**6.5 2D Chart Deceased, Isolated, and Released**

From this chart (Figure 6), we can see the total amount of people who were marked as deceased, isolated, and released after being a confirmed case. As we see from the chart the number of deceased is very low compared to isolated and released. This is great information for the KCDC because we believe that the isolation of these Covid-19 patients correlates with the low number of the deceased and the high number of released.



*Figure 6 – Deceased, Isolated, and Released*

1. **Conclusion**

Finally, with all the information and data provided above we can conclude the following.

1. The capital of South Korea (Seoul) had the most confirmed cases.
2. The top 5 confirmed cases by age range were 20’s, 40’s, 50’s, 80’s and 90’s.
3. The number of people in their age range that have passed away we can see that it goes from the 30’s - 90’s.
4. The top 2 deceased cases by age range were the 50’s and 80’s. From this we understand how age impacts deceased cases.
5. The number of deceased, isolated and released patients after being a confirmed case.

From the data we were able to extract, we were able to provide an interactive tool for analysis and data manipulation. Further work can be done to provide more information and create more data driven conclusions about Covid-19 in South Korea.

For more information, tutorials and code visit project’s GitHub link[[2]](#footnote-3).

**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>

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

[8] 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. From the map, we can also see that the capital of South Korea (Seoul) has the most confirmed cases and the age range it affected the most was the 20’s. [↑](#footnote-ref-2)
2. GitHub Link: <https://github.com/dakbari2/Group5Project> [↑](#footnote-ref-3)