Big Data Course

Capstone Project   
Final Report

For students (instructor review required)

ⓒ2023 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung Innovation Campus, you must receive written consent from copyright holder.

|  |
| --- |
| Covid-19 Data Analysis |

16/08/24

Content

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. Simulated Scenario Description

2.2. Datasets Selection and Description

2.3. Data Ingestion Pipeline

2.4. Data Transformation Processing

2.5. Data Query and Insight

3. Results

3.1. Data Ingestion Scripts and Code

3.2. Data Transformation Scripts and Code

3.3. Description and Sample of Transformed Datasets

3.4. Data Visualization of Query Results

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Introduction

1.1. Background Information

My capstone project titled “Covid-19 Data Analysis” is a project that I have launched to leverage data engineering skills to process, analyze, and derive insights from COVID-19 data. The project typically includes tasks such as data collection from various sources, data cleaning, transformation, and storage in a structured format.

1.2. Motivation and Objective

The Covid-19 pandemic has had a profound impact on global well-being, dramatically changing the world. This is the motivation for me to launch this project. By analyzing the COVID-19 data, I aim to discover and gain a deeper understanding of many aspects of the pandemic. From this, I can propose solutions for the future.

1.3. Schedule and Milestones

* Thinking and choosing the topic
* Researching the topic, finding the information on the Internet
* Selecting the dataset
* Following the tutorial
* Learning, analyzing and understanding about it

2. Project Execution

2.1. Simulated Scenario Description

Imagine a situation where a national public health agency is facing a critical challenge: managing the overwhelming influx of COVID-19 data from hospitals, testing centers, and vaccination sites across the country. The agency needs to quickly process this data to track the spread of the virus, monitor vaccination efforts, and identify regions at risk of outbreaks. However, the data is coming in from various sources, in different formats, and with varying levels of quality.

2.2. Datasets Selection and Description

I chose the dataset from the AWS blog “A public data lake for analysis of COVID-19 data” because it’s published directly by Amazon, making it easy to find and understand. The data includes the following tables:

* Global Coronavirus (COVID-19) Data – Tracks confirmed COVID-19 cases in provinces, states, and countries across the world with a breakdown to the county level in the US.

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | Description | Source | Provider |
| enigma\_jhud | Confirmed COVID-19 cases | Johns Hopkins | Enigma |

* Coronavirus (COVID-19) Data in the United States – Tracks confirmed cases and deaths in the US by state and county.

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | Description | Source | Provider |
| nytimes\_states | Data on COVID-19 cases at US state level | NY Times | Rearc |
| nytimes\_counties | Data on COVID-19 cases at US county level |

* Coronavirus Disease (COVID-19) Testing Data – Tracks the number of people tested, pending tests, and positive and negative tests for COVID-19.

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | Description | Source | Provider |
| covid\_testing\_states\_daily | USA total test daily trend by state | COVID Tracking Project | Rearc |
| covid\_testing\_us\_daily | USA total test daily trend |
| covid\_testing\_us\_total | USA total tests |

* USA Hospital Beds – COVID-19 – Data on hospital beds and their utilization in the US.

|  |  |  |  |
| --- | --- | --- | --- |
| Table Name | Description | Source | Provider |
| hospital\_beds | Hospital beds and their utilization in the US | Definitive Healthcare | Rearc |

* Lookup tables to support visualizations.

|  |  |
| --- | --- |
| Table Name | Description |
| country\_codes | Lookup table for country codes |
| county\_populations | Lookup table for the population for each county based on recent census data |
| us\_state\_abbreviations | Lookup table for US state abbreviations |

- Data Modeling:

A diagram of a computer program

Description automatically generated

Diagram link:

https://app.diagrams.net/#W46a6b98eb81a0d1f%2F46A6B98EB81A0D1F!sab9d4031065f40d4971ea143c7d73fad#%7B%22pageId%22%3A%22R2lEEEUBdFMjLlhIrx00%22%7D

A screenshot of a computer

Description automatically generated2.3. Data Ingestion Pipeline

Pic 2.3.1. Data Pipeline

* Amazon S3: is an object storage service offering industry-leading scalability, data availability, security, and performance. It’s used for archiving raw data. Raw data includes CSV, JSON, images, videos, …
* Crawler: is a Amazon Glue feature. It’s used for crawling data. It automatically detects and creates metadata for data in S3. It also creates schemas for analytics services to understand and process data
* Amazon Athena: uses metadata generated by AWS Glue Crawler to understand the data structure
* AWS Glue: is a serverless tool developed for the purpose of ETL job
* Amazon Redshift: is a fully managed, petabyte-scale data warehouse service in the cloud

2.4. Data Transformation Processing

2.4.1. Crawling data from S3

- Step 1: We upload data from the source collected from the Internet to Amazon S3. Then we create crawlers for each table. The data after that you can see on AWS Athena

- Step 2: We analyze data using AWS Athena. The purpose of this step is to have an overview of the data

- Step 3: OLAP. We use this method to minimalize data, let us can perform complex data.

We have 4 tables.

**factCovid (Event table)**

* **Description: This is the main event table, archiving data about COVID-19.**
* **Columns:**
* FIPS: Unique identifiers for regions (linked to the dimRegion table)
* states: Name of states
* region
* confirmed: The number of confirmed cases
* death: The number of dead cases
* recovered: The number of recovered cases
* active: The number of cases that still be infected
* positive: The number of positive cases
* negative: The number of negative cases
* hospitalizedcurrently: The number of people being hospitalized currently
* hospitalized: The number of people hospitalized
* hospitalizeddischarged: The number of people discharged from the hospital
* **Relation**:
* Link to table dimRegion via column fips
* Link to table dimDate via column fips
* Link to table dimHospital via column fips

**2. dimRegion (Dimension table region)**

* **Description: Stores detailed information about the regions**
* **Columns**:
* fips: Unique identifiers for regions
* state: Name of states
* region
* lat: latitude
* lang: longtitude
* county
* state\_abb: Abbrevation of state
* **Relation:**
* Link to table factCovid via column fips

**3. dimHospital (Dimension table hospital)**

* **Description: Stores detailed information about hospitals**
* fips: Unique identifier for hospitals
* state: Name of states
* hos\_lat: Latitude of hospital.
* hos\_lang: Longtitude of hospital
* hq\_address:  Headquarters address
* hospital\_type: Type of hospital
* hospital\_name: Hospital name
* hq\_city: Headquarters city
* hq\_state: Headquarters state
* **Relation:**
* Link to table factCovid via column fips

**4. dimDate (Dimension table date)**

* **Description: Stores details about dates**.
* **Columns**:
* fips: Unique identifier for date.
* date
* month
* year
* is\_weekend: Is it a weekend or not (1 is a weekend, 0 is a weekday)
* **Relation:**
* Link to table factCovid via column fips

Schema diagram:

A diagram of a data flow

Description automatically generated

**Conclusion:**

* **The factCovid fact sheet is at the center of this data model, which stores key information about the COVID-19 epidemic situation**.
* **Dimension tables (dimRegion, dimHospital, dimDate) provide additional insights and help standardize the data in the event table**
* **Relationships between tables are established through primary and foreign key columns (fips), allowing for easy and efficient inter-table queries.**

- Step 4: ETL in Python

- Step 5: Save result to S3

- Step 6: Build tables on Redshift

- Step 7: Copy data on Redshift

- Step 8: Visualization on Power BI

3. Results

3.1. Data Ingestion Scripts and Code

**Crawler:**

A screenshot of a computer

Description automatically generated

Pic 3.1.1.

**Data on Athena:**

A screenshot of a computer

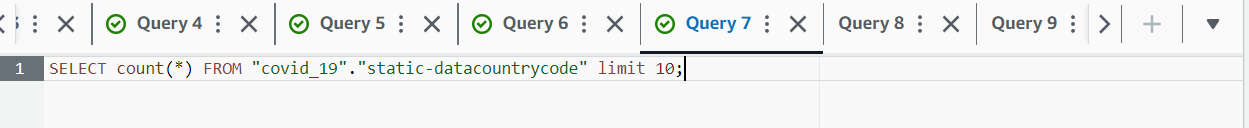
Description automatically generated

**Query results:**

A screenshot of a computer

Description automatically generated

**Counting static-datacountrycode**:



A screenshot of a computer

Description automatically generated

**The highest number of infected cases in table nytimes\_data\_in\_usa\_us\_county**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

3.2. Data Transformation Scripts and Code

**Connect to Athena:**

A screen shot of a computer

Description automatically generated

A black screen with white text

Description automatically generated

**Create a function to load query results from Amazon Athena into a pandas DataFrame**

A computer screen shot of a program

Description automatically generated

**Create a query response from Athena (here is table enigma\_jhud):**

A screen shot of a computer code

Description automatically generated

**Query successfully returned code 200:**

A screenshot of a computer program

Description automatically generated

**Load the data into the dataFrame and check:**

A screenshot of a computer

Description automatically generated

**Do the same with the "nytimes\_data\_in\_usa\_us\_county", "nytimes\_data\_in\_usa\_us\_states", "static\_datacountrycode" tables,...  Note with the "static\_datastate\_abv" table:A screenshot of a computer

Description automatically generated**

**After exporting the data from the dataframe, we see that row 0 must be the name of the column, but the result returned is a data. To convert row 0 to a column name, we use the pandas library to edit:**

A screenshot of a computer

Description automatically generated

**After saving the data in the dataFrame, we create the factCovid, dimHostpital, dimDate, dimRegion tables:**A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**The dimDate table has an unnormalized date attribute:**

**A screenshot of a computer

Description automatically generated**

**Results:**

A black rectangular object with a black border

Description automatically generated

**Copy the data to s3 (in the output folder):**

* We use the StringIO module in python's io library:

A screenshot of a computer program

Description automatically generated

* The above line of code saves the data of the factCovid dataFrame into a file factCovid.csv the "output" folder
* Similar to dimDate, dimHospital, dimRegion:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Create a data frame on Redshift:**

A black screen with red and blue text

Description automatically generated

A computer screen with a black background

Description automatically generated

A screenshot of a computer

Description automatically generated

A black screen with a black background

Description automatically generated

**Create a cluster on redshift (configure IAMROLE, access pubblic mode, add TCP ports) and connect via redshift\_connector library:A screen shot of a computer

Description automatically generated**

**Next, create a cursor and push the database frame up:**

**A black rectangular object with blue lines

Description automatically generated**

A black screen with a white text

Description automatically generated

A black rectangular object with white text

Description automatically generated

A black rectangle with blue lines

Description automatically generated  **After having the frame on Redshift, we copy the data from the output file (including factCovid.csv, dimDate.csv,... files) to the database on Redshift:**

A screenshot of a computer

Description automatically generated

3.3. Description and Sample of Transformed Datasets

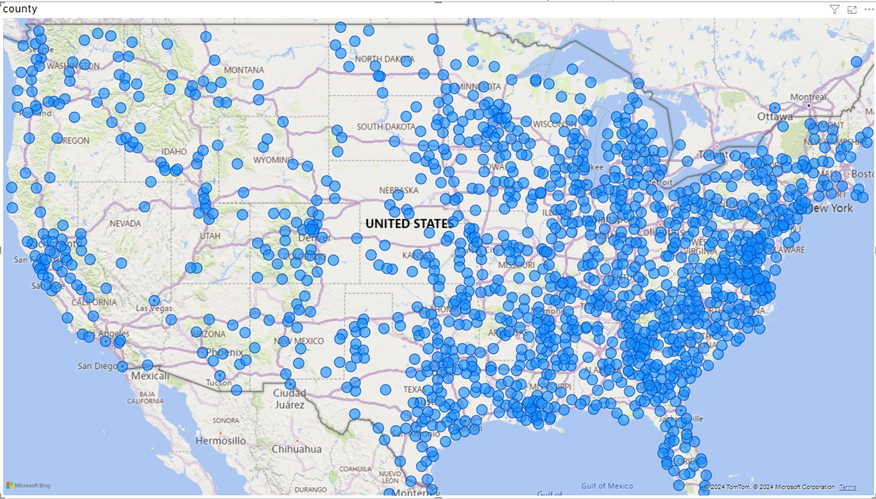
Data after transformation process:

A screenshot of a computer

Description automatically generated

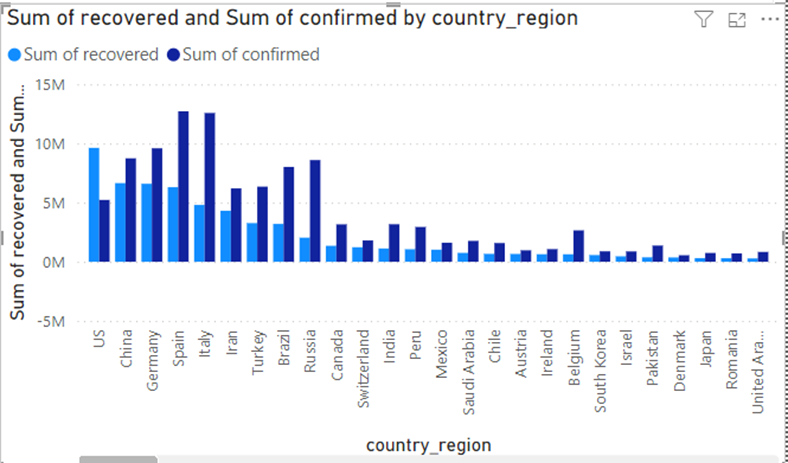
3.4. Data Visualization of Query Results

* Map of distribution of COVID-19 cases across the United States

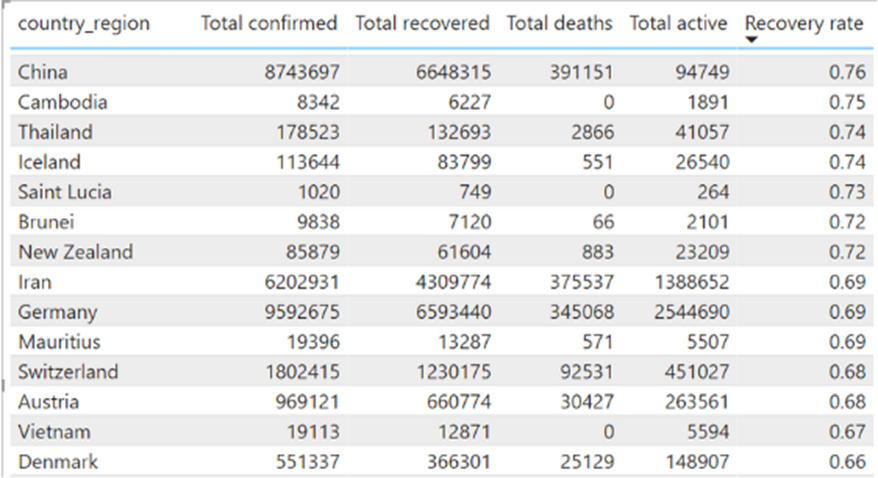


Based on the distribution of COVID-19 cases across the United States, it is evident that the majority of cases are concentrated more heavily in the eastern part of the country. This pattern can be attributed to several factors, including higher population density, greater urbanization, and more significant travel and economic activities in the East. Major metropolitan areas such as New York City, Washington D.C., and Boston have large populations and serve as hubs for international travel, which likely contributed to the initial spread and continued transmission of the virus. Additionally, the close proximity of states in the eastern region may have facilitated the inter-state spread of COVID-19, leading to higher case numbers compared to the western part of the country.

* Column chart showing total COVID-19 infections of countries around the world over time from January 19 to March 7

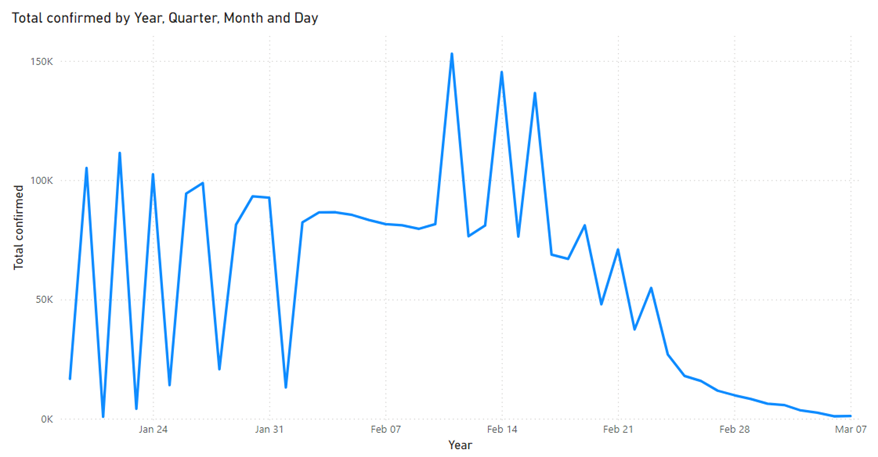


* The table shows the recovery rate of COVID-19 cases in the world over time from January 19 to March 7



In analyzing the COVID-19 recovery rates, it is notable that China has the highest recovery rate at 0.76. This suggests that China has implemented effective treatment and containment measures, leading to a significant number of recoveries. Similarly, Vietnam's recovery rate is also relatively high at 0.67. What is particularly remarkable about Vietnam's case is that there have been no recorded deaths, highlighting the country's successful efforts in managing the pandemic and providing effective healthcare. These high recovery rates reflect the strong response and healthcare strategies implemented by both countries.

* Line chart showing the world's number of covid-19 infections over time from January 19 to March 7



The line graph shows a significant spike in global COVID-19 cases from early February 2021 to mid-February 2021. This sharp increase could be attributed to several factors, including the emergence of new, more transmissible variants of the virus during that period, which spread rapidly across multiple countries. Additionally, the holiday season and winter months likely contributed to higher transmission rates due to increased indoor gatherings and travel. The surge in cases during this time highlights the challenges in controlling the pandemic and the importance of timely public health interventions to mitigate the spread of COVID-19.

4. Projected Impact

4.1. Accomplishments and Benefits

1. **Community and Stakeholder Engagement**:
   * Collaborated with public health officials, researchers, and other stakeholders to ensure the data analysis addressed critical questions and needs.
   * Published findings and data analyses in reports, academic journals, and public dashboards, promoting transparency and informed public discourse.

**Benefits:**

1. **Enhanced Decision-Making**:
   * Enabled data-driven decision-making, leading to more effective and timely responses to the pandemic.
   * Supported the allocation of medical resources, such as ICU beds and ventilators, based on data-driven forecasts of demand.
2. **Improved Public Health Outcomes**:
   * Helped reduce the spread of COVID-19 by identifying hotspots, predicting outbreak trends, and informing targeted interventions.
   * Supported vaccination efforts by analyzing distribution patterns and identifying areas with low vaccination coverage.
3. **Increased Transparency and Trust**:
   * Provided the public and stakeholders with clear, reliable data, fostering trust in public health measures and government actions.
   * Empowered citizens with information to make informed decisions about their health and safety.
4. **Resource Optimization**:
   * Optimized the use of resources by forecasting future needs, thus avoiding both shortages and waste, and ensuring that resources reached the areas most in need.
   * Streamlined data operations, reducing the time and cost associated with manual data processing and analysis.
5. **Long-Term Impact on Data Infrastructure**:
   * Established a robust data infrastructure that can be repurposed for future public health emergencies or other large-scale data analysis projects.
   * Enhanced the overall data capabilities of public health institutions, leaving a lasting impact on their ability to respond to future crises.

These accomplishments and benefits underscore the importance of data engineering in managing a global health crisis like the COVID-19 pandemic, demonstrating how data-driven approaches can significantly improve outcomes and save lives.

4.2. Future Improvements

**Future Improvements:**

1. **Enhanced Data Integration**:
   * **Incorporation of Additional Data Sources**: Integrate more diverse data sources, such as social media, mobility data, economic indicators, and environmental factors, to provide a more comprehensive view of the pandemic's impact.
   * **Cross-Border Data Collaboration**: Establish partnerships with international organizations to facilitate data sharing and integration across borders, enabling a more global perspective on the pandemic.
2. **Real-Time Data Processing**:
   * **Improved Real-Time Capabilities**: Enhance the data pipeline to support real-time or near-real-time data processing, allowing for immediate analysis and response to emerging trends.
   * **Automated Anomaly Detection**: Implement automated systems to detect anomalies or sudden changes in data (e.g., spikes in cases or vaccine distribution issues) in real-time, triggering alerts for quick intervention.

5. Team Member Review and Comment