ROBERT GORDON UNIVERSITY ABERDEEN

INFORMATICS INSTITUTE OF TECHNOLOGY

Informatics Institute of Technology
Department of Computing

Bsc(Hons) Artificial Intelligence and Data Science

Module: CM2606 Data Engineering
Module Coordinator: Dr.Kasun Jinasena

Coursework Report

Thareen Renuja
RGU ID – 2118809
IIT Student No. – 20211009

Introduction

In this project, I analyze the COVID-19 data and try to find patterns, trends, and risk factors for COVID-19 infection.

Dataset Selection

Dataset - A public data for COVID-19 research and development

[blogs/big-data/exploring-the-public-covid-19-data]

The dataset contains 10 tables.

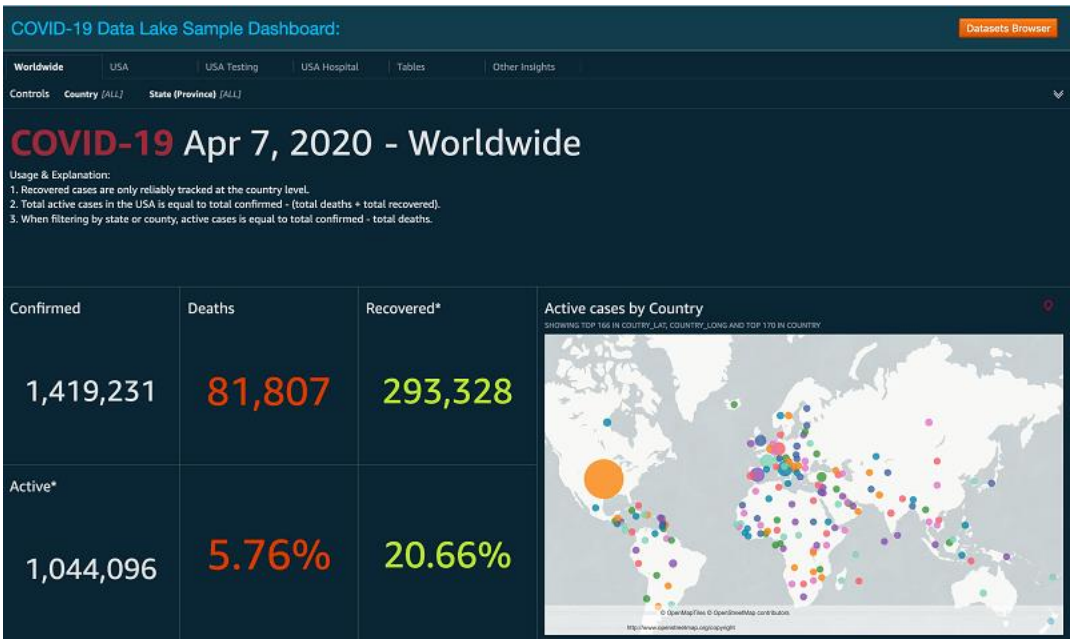
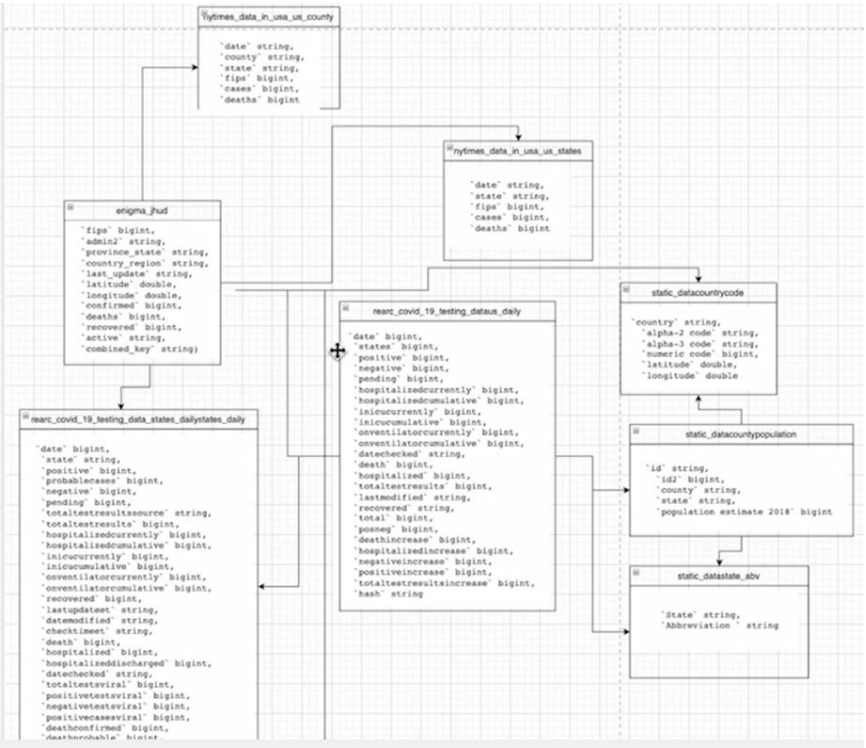


Table Name	Description	Source
nytimes_states	Data on COVID-19 cases at US state level	NY Times
nytimes_counties	Data on COVID-19 cases at US county level	NY Times
enigma_jhu	Confirmed COVID-19 case	Johns Hopkins

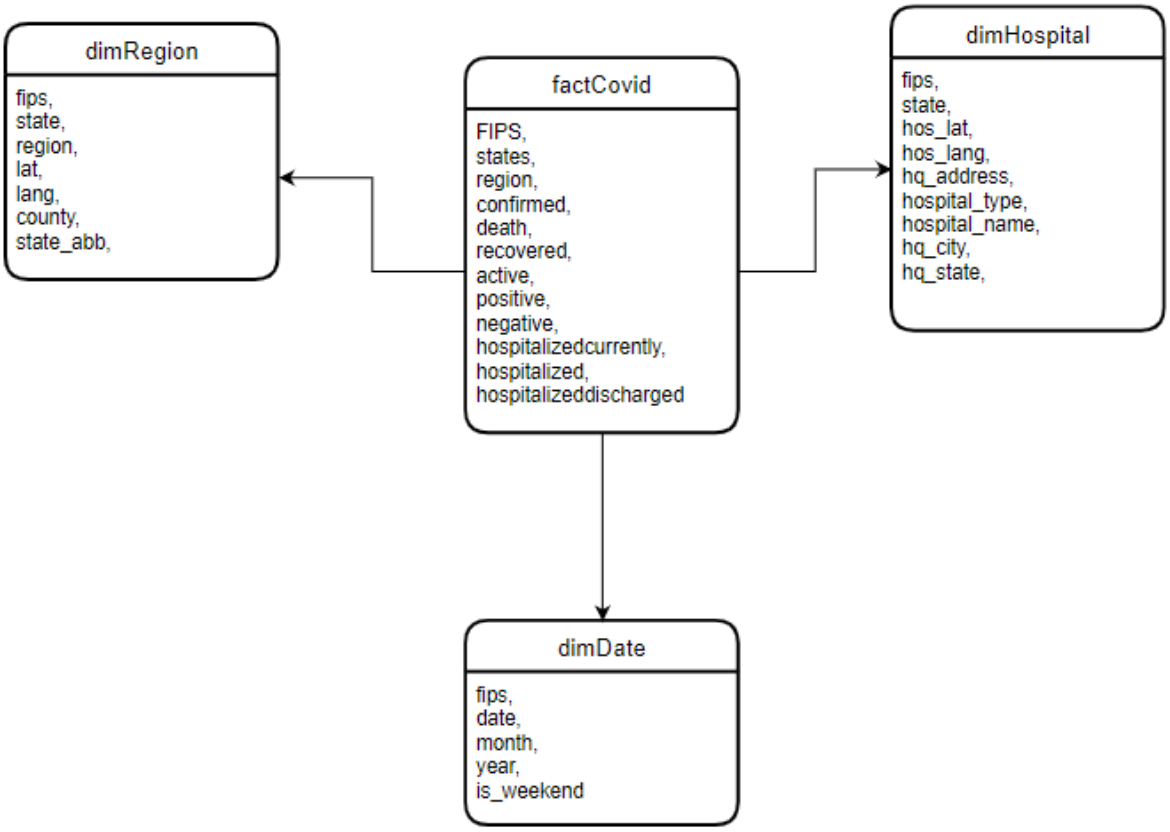
covid_testing_states_daily	USA total test daily trend by state	COVID Tracking Project
covid_testing_us_daily	USA total test daily trend	COVID Tracking Project
covid_testing_us_total	USA total tests	COVID Tracking Project
hospital_beds	Hospital beds and their utilization in the US	Definitive Healthcare
country_codes	Lookup table for country codes	Allen Institute for AI
county_populations	Lookup table for the population for each county based on recent census data	Allen Institute for AI
us_state_abbreviations	Lookup table for US state abbreviations	Allen Institute for AI

I download this dataset and manually upload to my AWS s3 bucket

Relational data model (Based on Primary key – ‘fips’)



Dimension model with 4 tables



3 dimenstion tables - (dimRegion, dimHospital , dimDate)

1 Fact table - (factCovid)

Insight Generation Mechanism

This step involves ingesting data from data sources into data lake in Amazon S3. I use AWS Glue for data ingestion.

Pipeline Design

Data Source: COVID-19 research data

Data Lake: Amazon S3 is a popular data lake storage option provided by AWS. It allows me to store large amounts of data in a simple and scalable manner, making it suitable for big data and data lake architectures.

Ingestion Flow: This step involves ingesting data from data sources into data lake in Amazon S3. I used AWS Glue for data ingestion.

Data Storage Option: Once data is ingested into data lake, I can choose different storage options in Amazon S3, such as object storage, file storage, or block storage. I choose file storage

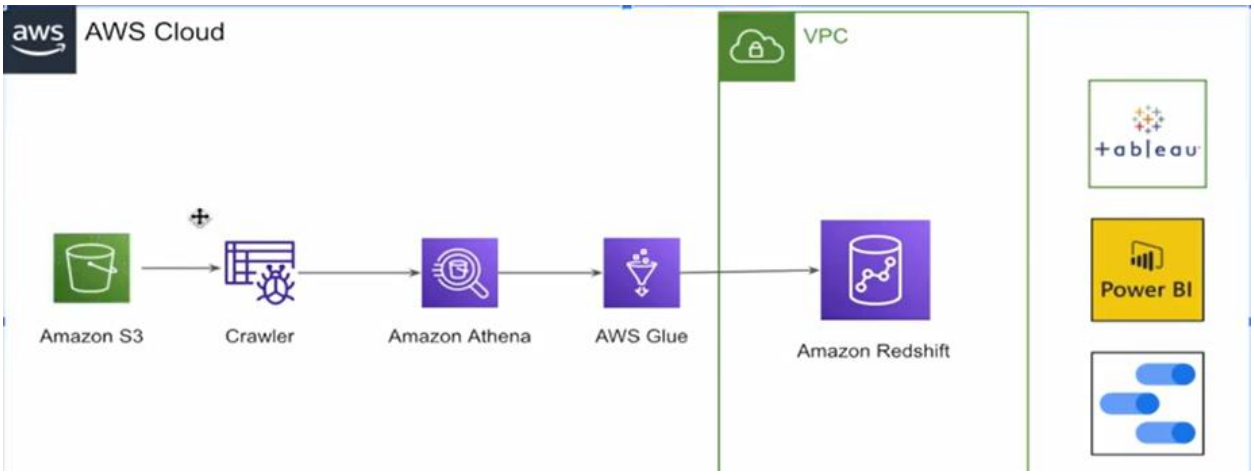
ETL Flow: Extract, Transform, Load (ETL) is a common data processing workflow in data engineering. AWS Glue is a managed ETL service that allows us to create, schedule, and run ETL jobs to transform data stored in data lake or other sources. I use Python code with AWS Glue to define my ETL jobs. python code -

Athena: Amazon Athena is a serverless query service that allows us to run SQL-like queries directly on our data stored in Amazon S3. I can use Athena to analyze, query, and visualize data without having to move it to another data store.

Redshift: Amazon Redshift is a fully managed data warehouse service that allows me to analyze large amounts of data at scale. I use AWS Glue to transform and load data from your data lake to Redshift for further analysis and reporting.

An orchestration mechanism

Overall, AWS Step Functions is a powerful and flexible orchestration mechanism that can help me to automate data ingestion pipeline, making it easier to manage and maintain complex data workflows in a serverless manner.



Discussions & Conclusion

The use of AWS Glue for data ingestion and AWS Step Functions for workflow orchestration offers several benefits in terms of data processing efficiency, scalability, and reliability. By leveraging AWS Glue's crawlers and jobs, data can be automatically extracted, transformed, and cataloged from various sources, simplifying the data ingestion process. AWS Glue also provides features for data transformation, data enrichment, and data aggregation.

References

https://youtu.be/GmR2migiu6A

https://youtu.be/zNj6XlbP2nA

https://youtu.be/gFWu-SSzRzc

https://aws.amazon.com/blogs/big-data/exploring-the-public-aws-covid-19-data-lake/