Cloud Data Warehouse for Fake News Dataset

Cloud Computing Project Final Report

Rohit Dalvi - 20201672

A report submitted in part fulfilment of the Cloud Computing module of MSc Computer Science.

School of Computer Science and Informatics

University College Dublin

03 December 2021

Abstract

The digital age in which we live has given rise to a plethora of methods for creating and absorbing information and knowledge. Because of its ease of use, rapid transmission, and low cost, social media has become one of the greatest best channels for news generation and consumption. However, technology also allows "Fake News" to spread widely.

Fake news is news as well as stories that are intentionally designed to lie and mislead or mislead people who read. Traditionally, untruths were accepted as the norm, and media organizations were routinely regarded as untrusted. 'If you want to recognize any fact concerning politics, users must peruse at least three separate papers, compare at least three different versions with same actuality, and finally came to your own concluding,' it must have been stated. The speed at which fake news continues to spread now a day, however, is what tends to make it most concerning. And the role of innovations in the propagation of false, imprecise, or false facts has risen rapidly in the years since their discovery.

The origins, spread, as well as consequences of blatant propaganda are all complicated issues. In current history, countless researchers and organizations have begun to investigate fake news but instead suggested different solutions to the problem.

Table of Contents

1. Introduction

- Explain your motivation of choosing your project out of the two.
  I chose this project over other as it covered major concepts like data sourcing, extracting, cleaning, loading, and transforming. I have been using cloud services since many years which inclined my interest towards cloud-based platforms and tools. I was motivated by eliminating unwanted data in operational databases then loading into Data warehouse. Which is felt would be useful to analyse any informational meaningful metrics a data visualizing dashboard.

- Recalls the project objectives.
  We are required to create a cloud data warehouse solution to store and organize fake news data for this assignment. Choose one or more fake news sites for which we'd like to construct a cloud repository. Define the application for which we'd like to use the fake news data gathered. G go with a cloud data warehouse solution. Use a Hadoop network or something like validate the new data warehouse paradigm. Create a dashboard for your system that demonstrates how it can be utilized.

- Why in your view this is interesting?
  1. Cloud services speed up Big Data and business intelligence computing operations, enabling microservices-based software modernization, and are essential for an e-commerce software engine.
  2. It has the impact of streamlining and simplifying corporate applications, reducing time to market, and successfully addressing client requests. It's extremely impressive to be able to run such strong computing clusters of nodes in the cloud with such ease.

- Explain briefly how this useful in implementing it on Cloud Computing
  1. A Data Warehouse application is expensive, in fact the cost is probably proportional to the size of the DW Database (MetaData). If we know data reporting and analysis fundamentals then we can reduce the size of the problem and therefore the cost. Alternatively, you can include all the data available and the cost and delivery timescale which we put on individual applications/software’s
  2. Pulling data from multiple sources, transforming it into consumable formats, and storing it in a warehouse is vital for making sense of data. And with valuable data stored in warehouses, you can go beyond traditional analytics tools and query data with SQL to discover deep business insights.

2. Project's Specifications and Requirements

- Define the project full specifications
  This project requires the following components to run successfully:
  1. Data Source containing Fake news file (csv/txt/xml etc)
  2. Data cleansing tool/ Editor (Bash/Python/Excel)
  3. Data Management system which can hold our files (S3/SSMS/RDS)
  4. Extract-Transform-Load operation tool to fetch the source of data and extract them from there and load into the cluster's database. Then transform data accordingly as per the user's requirement, only relevant information must be taken to decrease the load over the servers. Load them to the tables created within the data management tools of the cluster.
  5. Perform any SQL DDL, DML commands if needed to polish the dataset. Create any subset of main dump data as in extracting top features can be done here.
  6. To visualize the data on any business intelligence platform we link the drivers of the cloud platform to this external application.

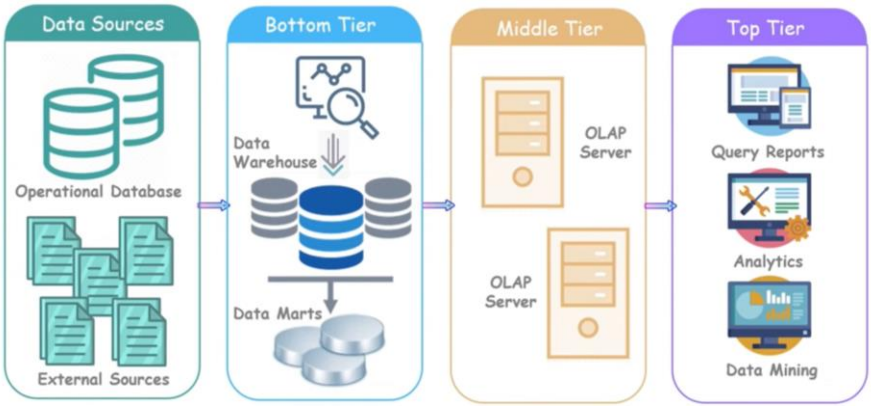
- What are the requirements of such application you would like to implement?
  There are few implementations I had thought of adding to this application.
  1. Taking the current file used for this project we can publish the analytics found and compare this with the prediction of such fake news circulating across calculated from this result.
  2. Also, it would be beneficial to purge this information in s3 buckets and keep on uploading snapshots of updated data, this will keep track of variance in fake reported news in the upcoming year.
  3. Data management and retention policy can be set as per the company's decision due to which targets such as innocent parties can be identified and saved from facing such false accusations and news generated by democratic votes.
  4. Most of the tweets/articles/newspapers containing fake news contain harsh words which can be identified and stored in an inline data file which can be run as per the cloud scheduling tool. This will detect any such false information and sort them into separate container.
  5. There are multiple methods to pre-process the data and the perform ETL to save them using merge-join-lookup-sort transformations in Talend, SSIS, AWS Glue etc.
  6. Not only fake news but such an application can be used in various fields where websites data are handled by multiple users. One can implement a feature where we bookmark the latest comment, post made. Geographically recognize actives happening actively for any event.

3. Methodology & System Architecture

- What is the methodology followed to develop and implement the project?

The plan was to first choose a decent fake news data set and download it. After it got approved check for any columns which might not pass through the data load query. The free tier version of redshift has been used with single node, dc2.large. To store and access the data file S3 has been used as a data storage tool. Once the cluster was up and running, I set the inbound rules and applied read only access to redshift and full access to S3. Loading data in the query window took some time. After making successful connection with tableau using ODBC 64bit driver I implemented a dashboard containing graphical visualizations which is hosted live on a tableau server.

- What is the system architecture you would like to implement and deliver?
    1. Users initially launch a set of nodes and provision them, after which they upload data and carry out an analysis. Part of a broader Amazon Web Services (AWS) ecosystem, the Redshift data warehousing service offers various features.
    2. My system architecture looks somewhat like this. It covers all the layers mentioned which we can say backend to frontend implementation.

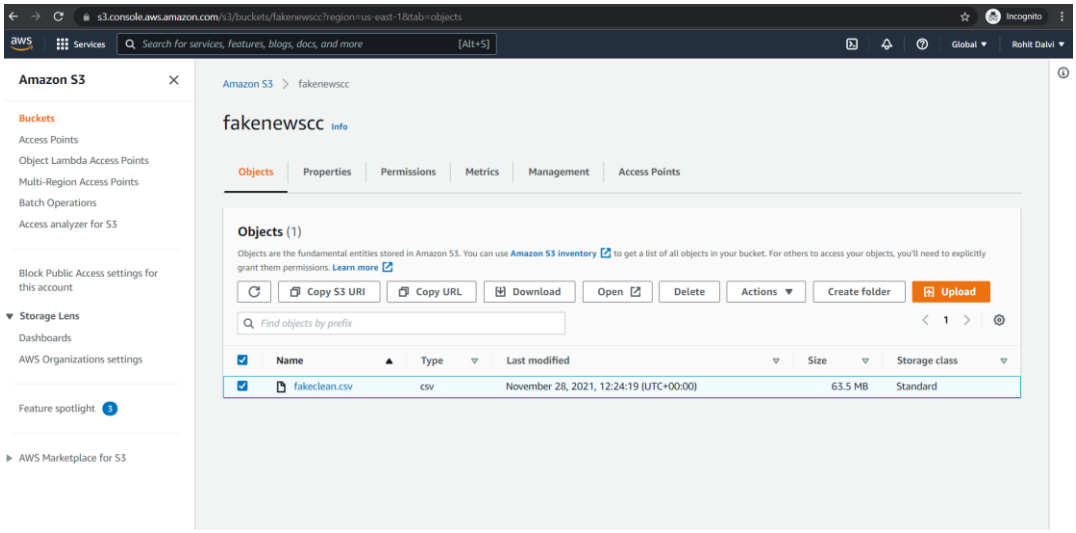


4. Implementation Details

- How did you implement it (language, environment, …)?
    1. I implemented the cloud warehousing solution using AWS Redshift. Created a cloud repository where I could extract my cleaned csv data set. Initially I found the source file which contained fake news that had to be cleaned. I have done my data cleansing using bash commands.

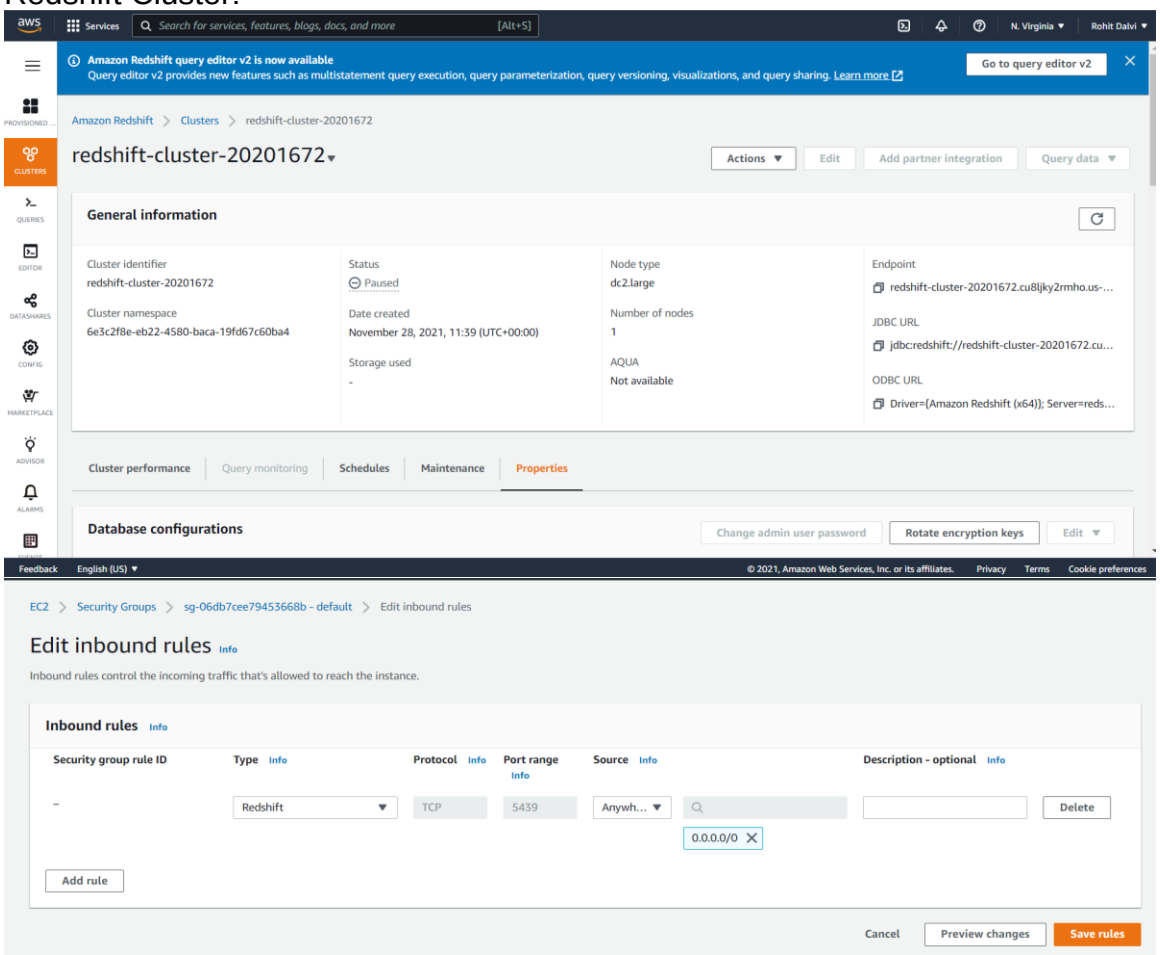
```
#!/bin/bash

cp fake.csv fakeclean.csv
while grep -q '"[^"][^"]*|.*"' fakeclean.csv ;do
    sed -i 's/\("[^"][^"]*\)|\(.*"\)/\1;\2/' fakeclean.csv
done
```

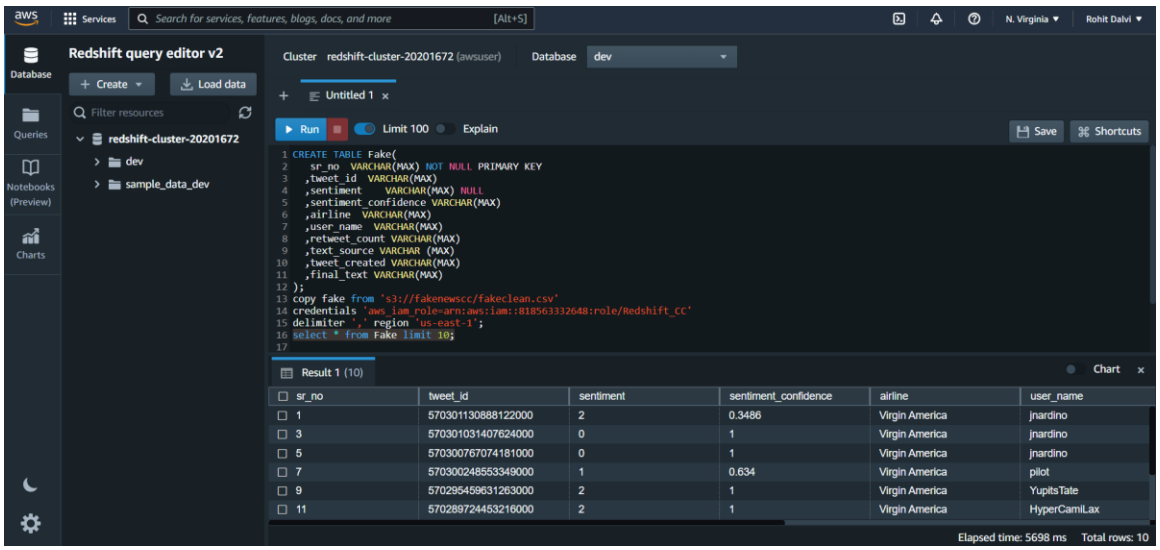


2. This cleaned data was then loaded into AWS S3 bucket which was linked to redshift cluster. The S3 URI was synched with redshift by assigning correct role and policies to gain access over these two platform services.

Redshift Cluster:



- How much have you implemented (functions, operations, etc.)?
    1. Data cleaning is done with the help of bash commands.
    2. Data storage in S3 bucket.
    3. Creating appropriate IAM Role for the S3 as well as Redshift policy roles.
    4. Setting the inbound rules inside EC2 security groups.
    5. Making the created redshift cluster publicly accessible.
    6. Creation of SQL table using redshift query editor and dumping data from s3 bucket.
    7. Installing Amazon Redshift ODBC DSN for configuring tableau with data source (Redshift).
    8. Built a reporting dashboard (Application) which contains graph and charts for visualizing data. Also hosted the same on tableau public keeping a single clickable link for user's access.



- Evaluation.
    1. A thorough testing has been done over the whole process since beginning. The dataset has been run through strict conditions where all columns and rows were parsed through queries.
    2. Exact location and variable data types were checked for best results. Column transformations were taken care of to identify accurate characteristics of source columns.
    3. End-point was tested for its network connectivity using {telnet endpoint}.
    4. Apart from this visually graphically finding insights over the data made evaluation much better on tableau.

- Explain how the system should be installed and used?
    1. The final application is published online. Thus, by clicking on the url one can access and even download the .twb file which can be loaded into the computer and evaluated to see the pages and final creation of the dashboard.
    2. To view the metrics, it can be seen in blocks on which if hovered tooltip is also visible to display the counts and relevant data shown.

3. The Cluster is available in AWS redshift which keeps the data live but if application has already extracted data dump, then it can be accessed anytime anywhere.

Click on the below link to access the Tableau dashboard.

https://public.tableau.com/views/FakeNews_16385413864100/FakeNews_Dashboard?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

5. Conclusion

- State clearly how much have you achieved in developing and implementing the project.
    1. Identified source of fake news and created a cloud repository.
    2. Chosen S3 as its stores data as it provides object storage through a web service interface.
    3. Created a data warehouse using Redshift cluster by keeping all policies, rules, synchronization in mind.
    4. Connected database to visualizing tool and built an application/dashboard which is hosted online and can be accessible at one click.

- Explain the challenges you had to overcome.
    1. There were multiple challenges been faced during the implementation such as cleaning of dataset like removing ‘,’ from the columns to ensure correct separation of fields by delimiter.
    2. Then Making it publicly accessible and what rules to apply in security group took much time. Last but not the least, creation of dashboard with relevant meaningful features was challenging.
- What did you learn from this project?
    1. Due to characteristics of Cloud computing like elasticity, scalability and deployment time, reliability, and reduced costs, the DW in the cloud have great potential.
    2. For a data warehousing system to be able to utilize the capabilities of the cloud it will have to be both highly parallel and distributed, while complying with many requirements discussed in this paper.
    3. This makes distributing and parallelizing less of an issue.
    4. Security issues will likely always be involved in the decision of moving a data warehousing systems or data marts into the cloud.