CLOUD PROJECT ON DATA WAREHOUSE

RAHUL AGRAWAL (21200098)

INTRODUCTION

MOTIVATION

I've chosen to work on 'Cloud Data Warehouse for Fake News Dataset' as data warehousing is a system that is designed to help improve efficiency and retrieval speed on large-scale datasets. I wanted to understand the jargon associated with this process as it helps corporate decision-makers to derive insights and conclude on business decisions. Also, since I have already worked on MapReduce on previous practical sessions and gained some basic understanding earlier as a part of this curriculum, I wanted to try my hands on something different.

OBJECTIVE

The project aims on storing and organizing the dataset on a cloud warehouse (AWS RedShift) and a BI dashboard that can help in analyzing the dataset for future use cases. Further, the dataset must include data regarding the fake news which is increasing everyday and becoming common these days.

REASON

I found this interesting because we live in an era where fake WhatsApp forwards and Tweets may influence the minds of the uninitiated, tools and expertise must be put to practical use in not just preventing the spread of disinformation, but also in informing people about the news they consume. So, for this purpose I chose this dataset and aimed to build a system that can help in deriving some insights and analysis out of it.

The development of practical tools for users to acquire insight from the articles they consume, fact-checking websites, built-in plugins, and article parsers can be improved, made more accessible, and, most importantly, should raise awareness.

WHY CLOUD?

Implementing data warehouse on cloud rather than on a physical machine helps companies function better by focusing on what their business in preference to maintaining a room full of servers. There were several reasons why I chose AWS as my cloud service and hosted the application on it because –

1. Leading Cloud Provider: AWS is the most popular and widely accessible cloud platform that is available in almost every country and provides tons of services from warehousing to elastic computing as well as storage services.
2. Rapidly Scale Cloud Services: Scalability is a significant advantage of cloud computing. Unlike on-premises datacenters, where new gear must be purchased, public cloud services may swiftly scale up as needed. Your business may quickly scale back if the necessity for a particular resource (such as disk space or central processing units) reduces. This scale-up/scale-down methodology guarantees that companies have the resources they require at the time they require them.
3. Pay-As-You-Go Cost Savings: It is possible to save money by moving workloads to the cloud. You only pay for what you use with most cloud services. For businesses with set operation schedules, this is a significant advantage of cloud computing. Most providers let you automate the process of putting infrastructure up for a given task and then shutting it down once the job is done. As a result, you only pay for the hours the infrastructure is operational.

PROJECT'S SPECIFICATIONS AND REQUIREMENTS

SPECIFICATION

For this project, we need to identify a dataset that resembles the fake news features and then create a warehouse solution to store the same. Storing dataset in warehouse needs to be done precisely so that it does not contain any null values, this means data needs to be preprocessed. After successful store, we need to create a BI dashboard that will contain insightful charts which will help in deriving business decisions.

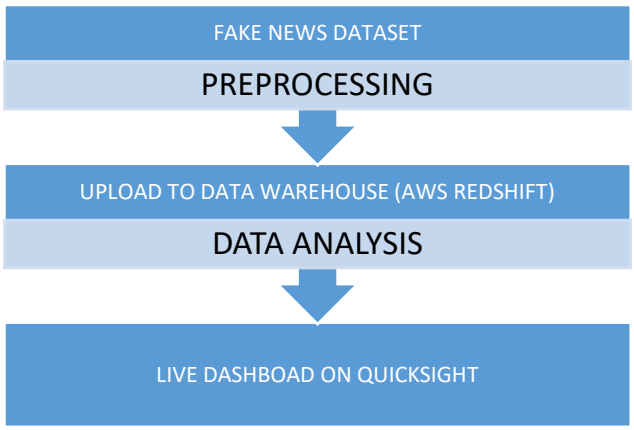
REQUIREMENT FOR APPLICATION

The requirements for this project are -

1. We will require a fake news dataset. Kaggle is one of the best-platform for sample datasets, I found one from here (https://www.kaggle.com/ruchi798/source-based-news-classification)
2. A cloud warehousing platform is needed for loading dataset and storing it for later use. AWS RedShift is one great solution which uses SQL to analyze structured and semi-structured data across data warehouses, operational databases, and data lakes, using AWS-designed hardware and machine learning to deliver the best price performance at any scale.
3. Lastly a BI tool is needed to generate a dashboard, for this AWS QuickSight can be a good solution as querying data from RedShift will be much easier and is very user friendly.

METHODOLOGY

SYSTEM ARCHITECTURE



METHODOLOGY

The following methodology was followed in order to implement the project:

1. Preprocessing of dataset:
   Here we cleaned the dataset for null values, as the number of null values were very less so the rows were removed from the dataset. There were line break in some column which would have caused issue while uploading data to warehouse, they were replaced by a space character. Headers were removed from the dataset and copied to a different file for later use.
2. Uploading of data:
   Data was then uploaded to AWS RedShift by creating a table inside the cluster. Here, we can perform data analysis using the queries of SQL.
3. Finally, for the live dashboard AWS QuickSight was used for analytics, data visualization and reporting. The dashboard was published and was shared to guest user whose credentials are shared in readme.txt file for verification.

DATA COLLECTION

For the purpose of this project, I downloaded the dataset present in Kaggle which contained almost 2096 rows and 12 columns. The type of news were 'bias', 'conspiracy', 'fake', 'bs', 'satire', 'hate', 'junksci', 'state'. The link of dataset is https://www.kaggle.com/ruchi798/source-based-news-classification.

DATA PREPROCESSING

For preprocessing the data, I used Anaconda Jupyter notebook (Python) and Numpy libraries to handle the null values. There are close to 2.5% values which are missing in the columns 'text_without_stopwords' and 'text' and close to 0.5% missing values in columns like 'title_without_stop_words', 'language' etc. We will be dropping these null values. After that removed the line break present in 'title' column and the headers of the csv file and pasted it to another csv 'headers.csv'. Lastly, I exported the data to different csv 'NewsDs.csv' for uploading it to AWS Redshift.

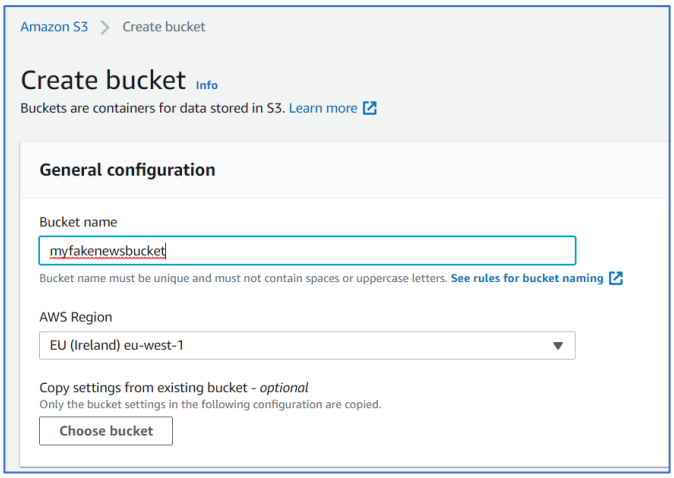
SYSTEM IMPLEMENTATION

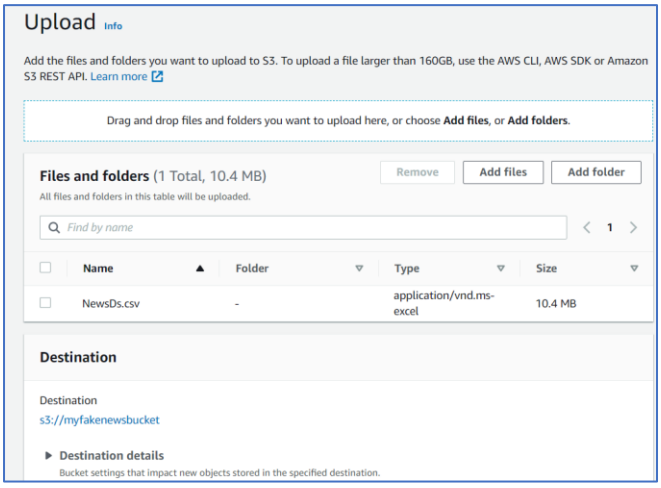
HOW DID I IMPLENT IT?

The following were used for implementation AWS S3 bucket for object storage, AWS RedShift for data warehousing and AWS QuickSight for BI Dashboard.

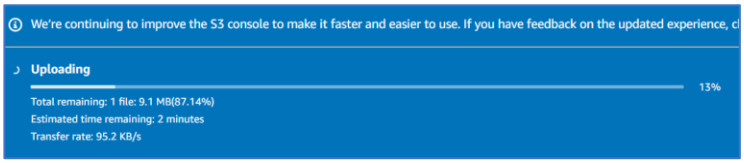
IMPLEMENTATION STEPS

1. The cleaned dataset was first loaded onto a S3 bucket called 'myfakenewsbucket'. This bucket will later be used my warehouse as an input file.

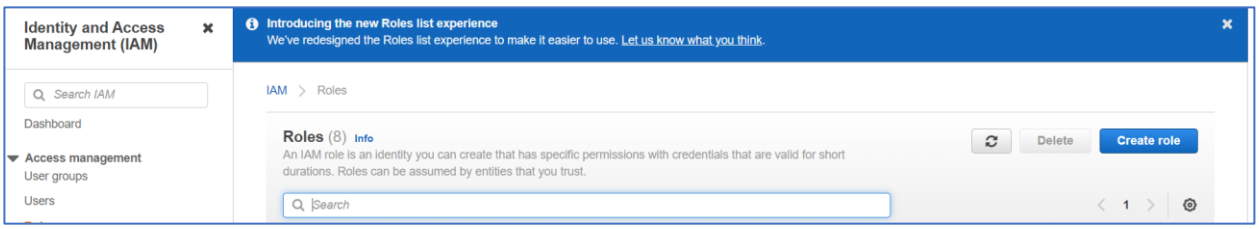


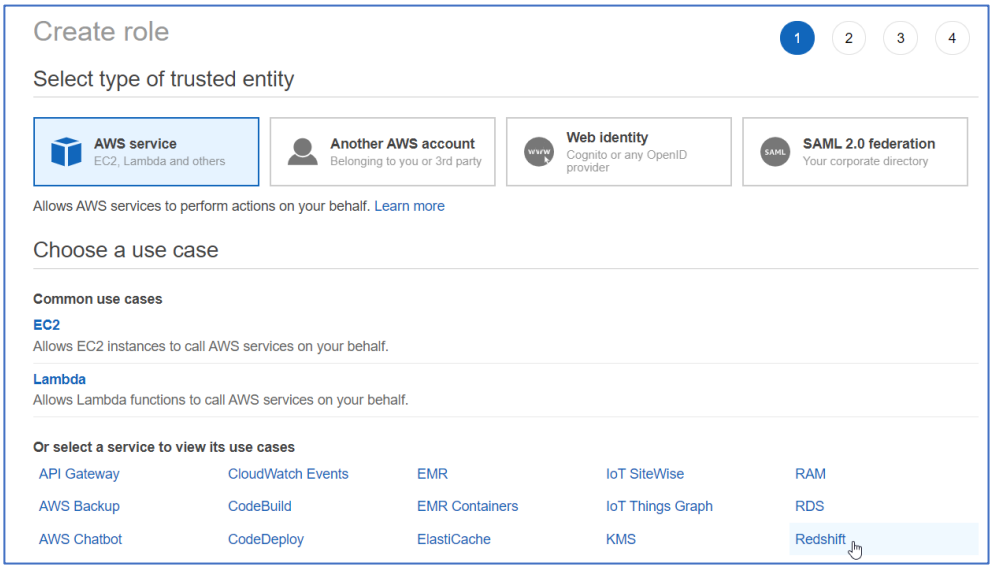


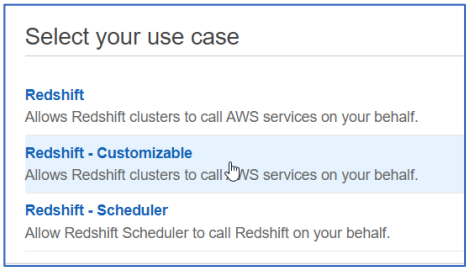
Wait untill data upload is complete.

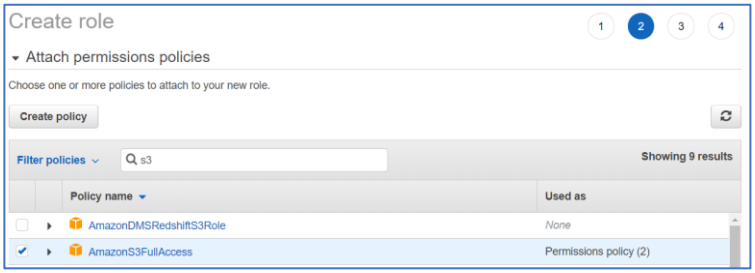


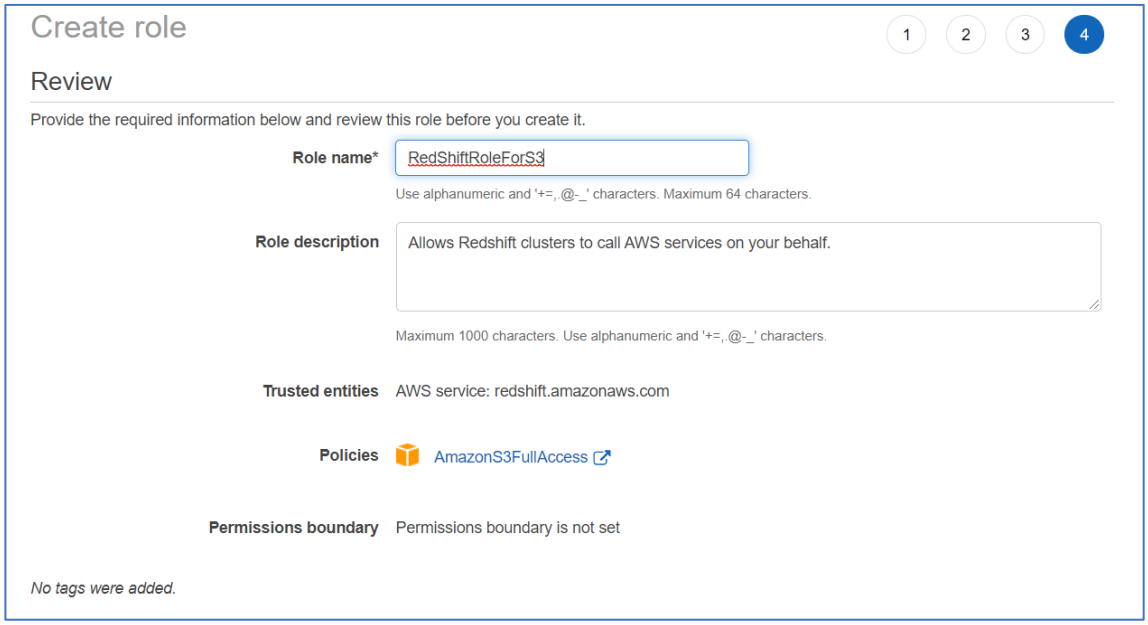
2. To access data from S3, RedShift will need an IAM role. Search IAM on AWS search bar and click on 'Create Role'. Click 'RedShift' after that and then click 'RedShift Customizable'. Click on Next. Search for 'AmazonS3FullAccess' in the search bar. Give a name for role on next screen like 'RedShiftRoleForS3'. Click Create Role.


Identity and Access Management (IAM)
Introducing the new Roles list experience
We've redesigned the Roles list experience to make it easier to use. Let us know what you think.
Search IAM
Dashboard
Access management
User groups
Users
IAM > Roles
Roles (8) Info
An IAM role is an identity you can create that has specific permissions with credentials that are valid for short durations. Roles can be assumed by entities that you trust.
Delete
Create role
Search

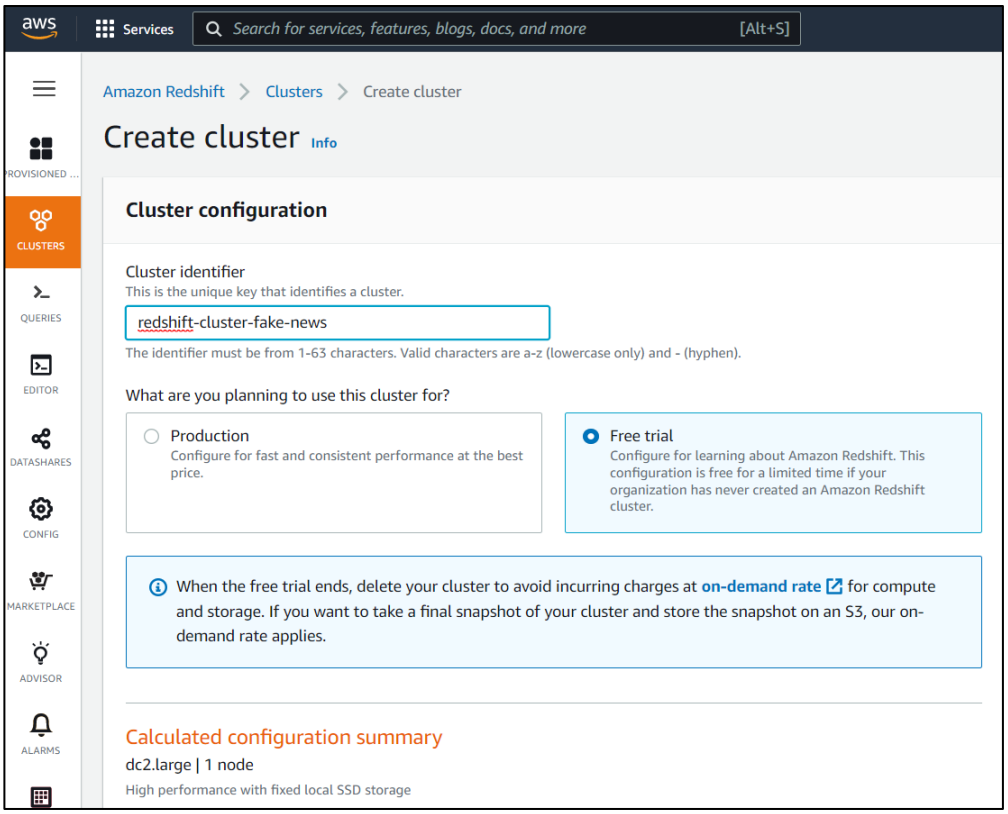

Create role
1 2 3 4
Select type of trusted entity
AWS service
EC2, Lambda and others
Another AWS account
Belonging to you or 3rd party
Web identity
Cognito or any OpenID provider
SAML 2.0 federation
Your corporate directory
Allows AWS services to perform actions on your behalf. Learn more
Choose a use case
Common use cases
EC2
Allows EC2 instances to call AWS services on your behalf.
Lambda
Allows Lambda functions to call AWS services on your behalf.
Or select a service to view its use cases
API Gateway
CloudWatch Events
EMR
IoT SiteWise
RAM
AWS Backup
CodeBuild
EMR Containers
IoT Things Graph
RDS
AWS Chatbot
CodeDeploy
ElastiCache
KMS
Redshift


Select your use case
Redshift
Allows Redshift clusters to call AWS services on your behalf.
Redshift - Customizable
Allows Redshift clusters to call AWS services on your behalf.
Redshift - Scheduler
Allow Redshift Scheduler to call Redshift on your behalf.


Create role
1 2 3 4
Attach permissions policies
Choose one or more policies to attach to your new role.
Create policy
Filter policies
s3
Showing 9 results
Policy name
Used as
AmazonDMSRedshiftS3Role
None
AmazonS3FullAccess
Permissions policy (2)


Create role
1 2 3 4
Review
Provide the required information below and review this role before you create it.
Role name*
RedShiftRoleForS3
Use alphanumeric and '+=,.@-_' characters. Maximum 64 characters.
Role description
Allows Redshift clusters to call AWS services on your behalf.
Maximum 1000 characters. Use alphanumeric and '+=,.@-_' characters.
Trusted entities
AWS service: redshift.amazonaws.com
Policies
AmazonS3FullAccess
Permissions boundary
Permissions boundary is not set
No tags were added.

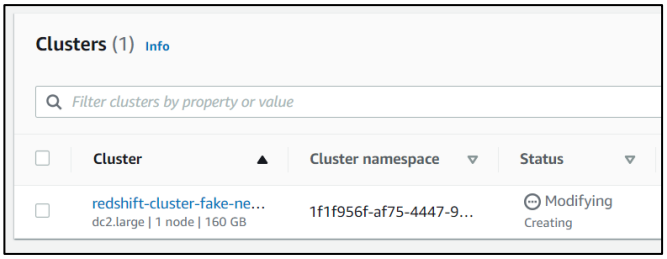
3. Then search RedShift in AWS Console. Click on ‘Create Cluster’ then give a name to cluster ‘redshift-cluster-fake-news’. Select Free Trial and select username and password then click on create.

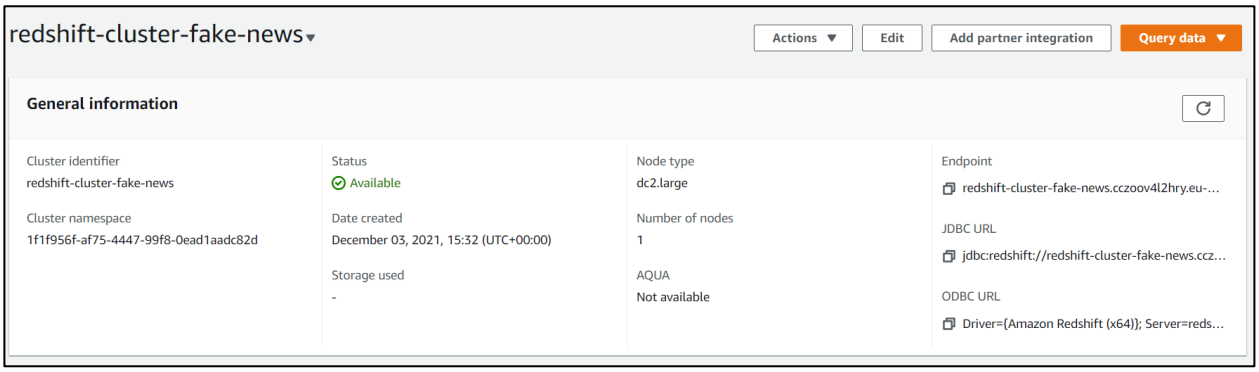


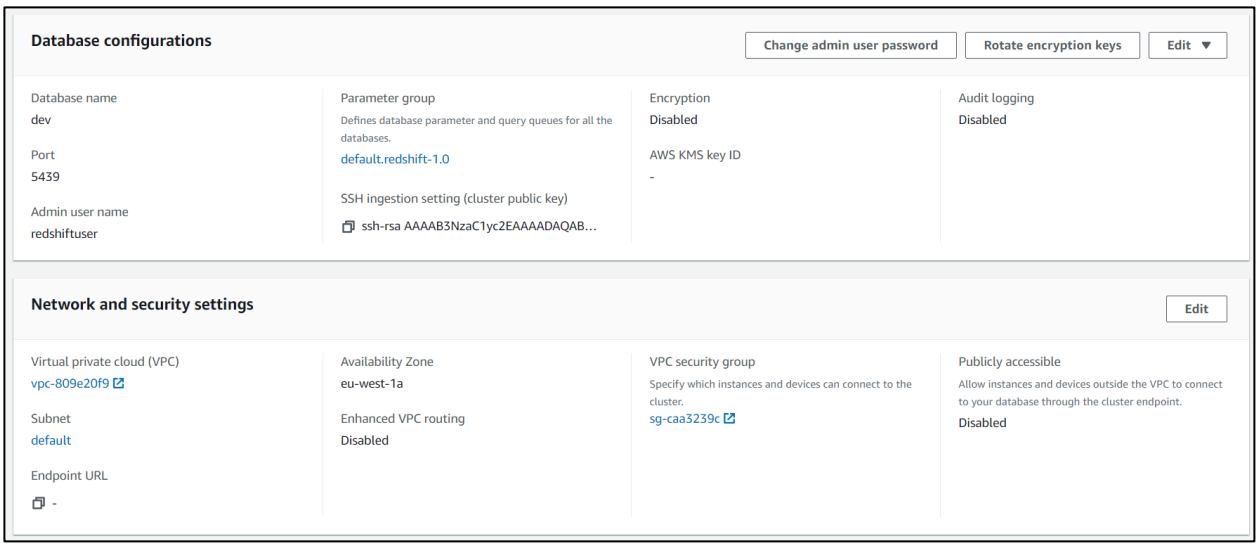
redshiftuser

Reduser123

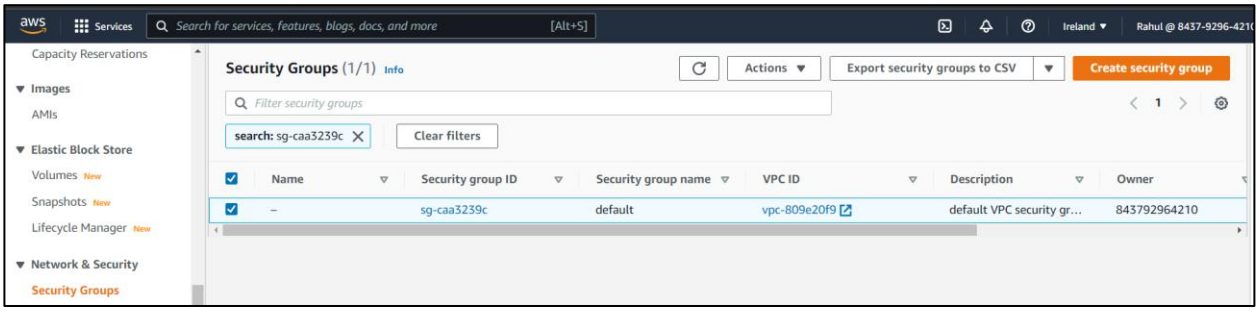
4. Then cluster will go to modifying state then then it will be available.

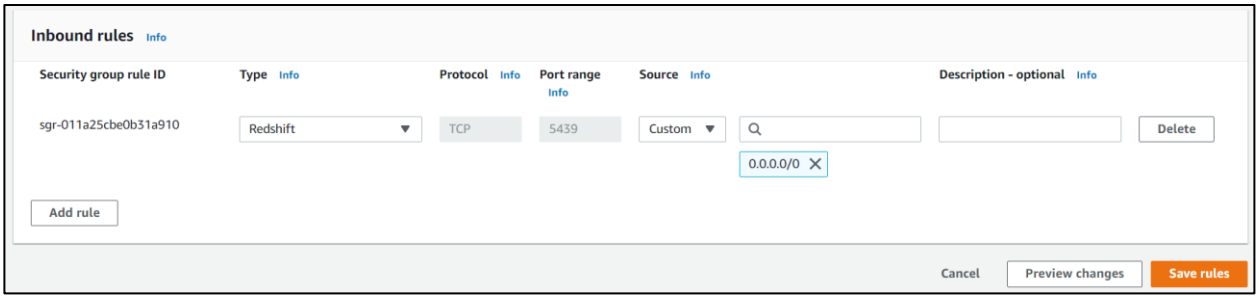




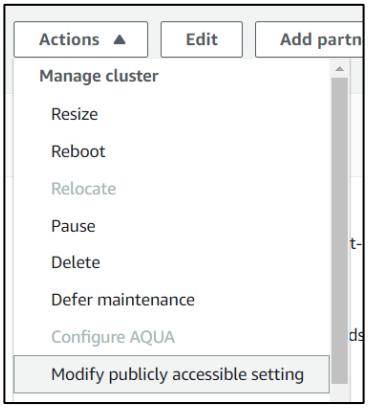


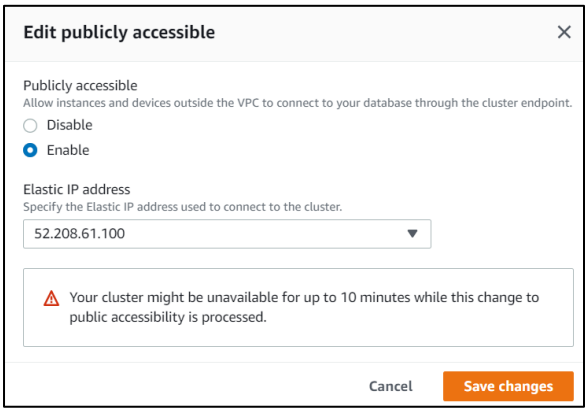
5. Then go down to ‘Network and Security settings’ and open the link in ‘VPC security group’. Change the inbound rule of security group as provided below. Click on save rule.



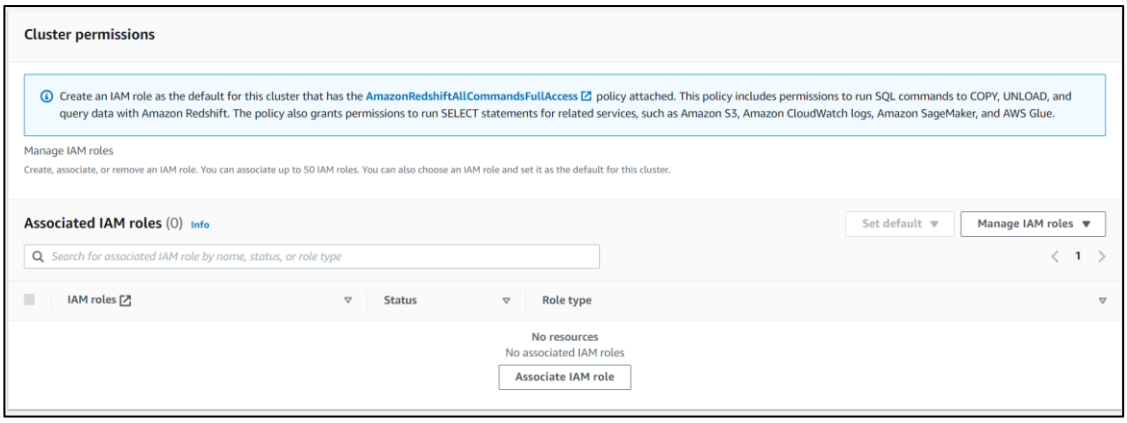


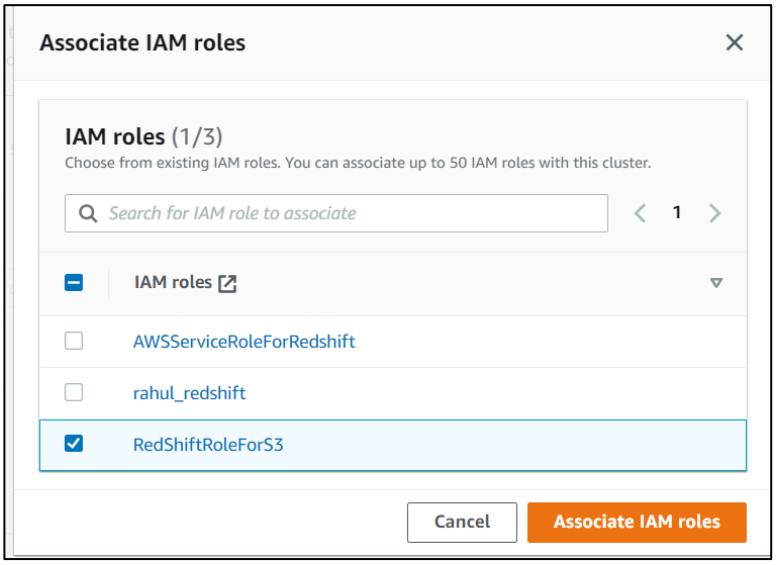
6. Go back to Redshift cluster (‘redshift-cluster-fake-news’) and click on ‘Actions -> Modify publicly accessible setting’. A new pop up will open select ‘Enable’ and choose IP address from drop down. Click save changes.



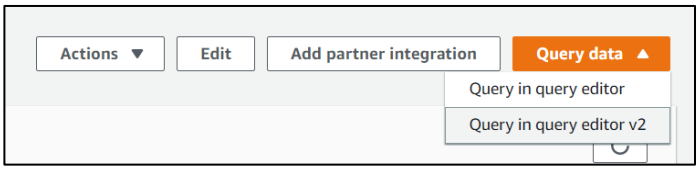


7. Now attach the IAM role created earlier to the RedShift cluster. Navigate to ‘Cluster permission’ and then click on ‘Associate IAM Role’. After that select the role created earlier in this process.

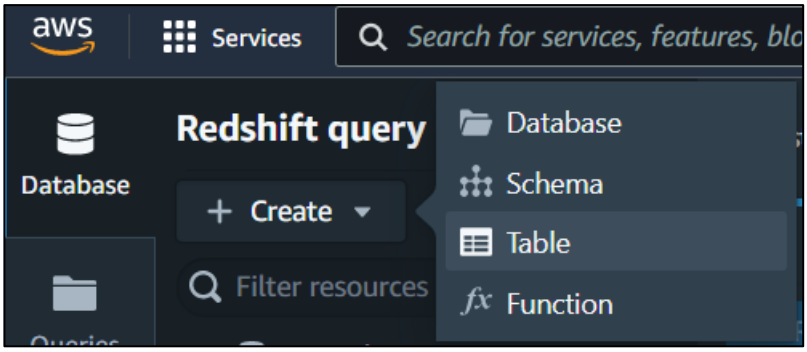




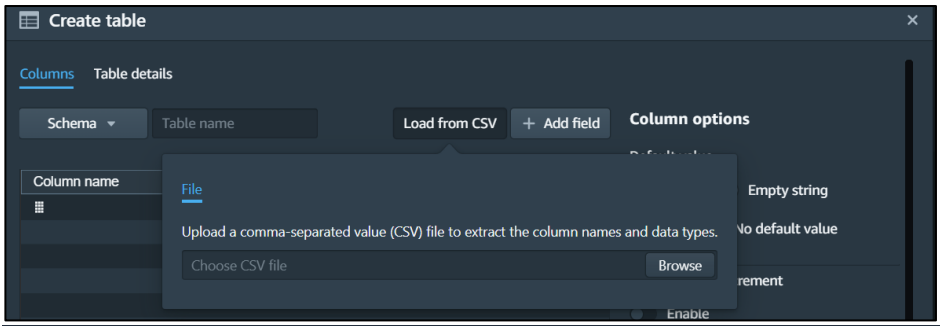
8. So till now we have uploaded data to S3, updated security setting of RedShift, and attached IAM Role. Now, Click on ‘Query Data -> Query in query editor v2’.

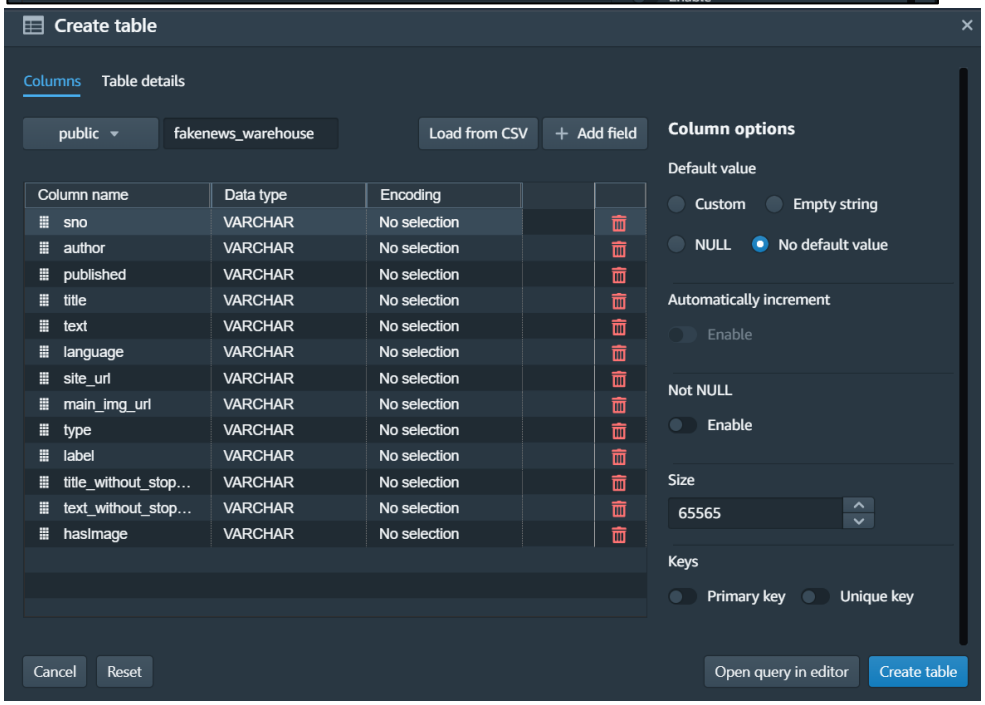


9. Then Click on Create then select ‘Table’. We will first create a table before uploading the data.

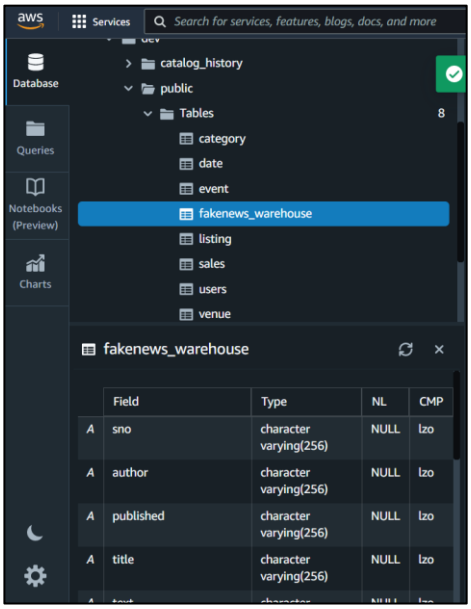


10. Click ‘Load from CSV’ and then choose ‘header.csv’ file from local system. This will load the headers and update the dialog box, give table name like **‘fakenews_warehouse’** then click on ‘Create Table’.

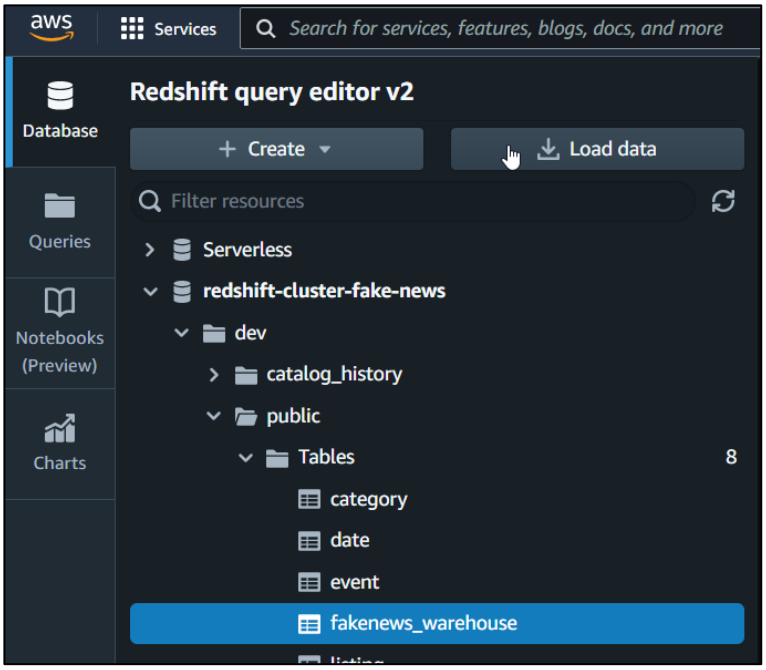


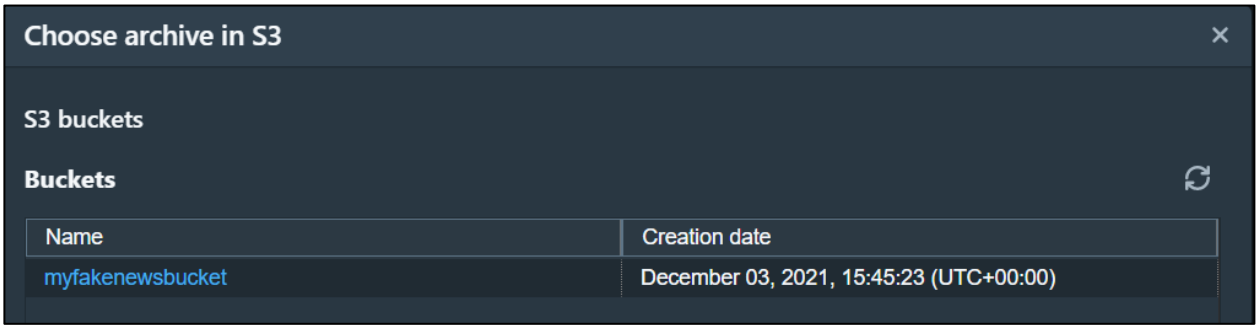


11. Check once again if table is created under public -> Tables -> fakenews_warehouse. Clicking on table will generate the schema below.

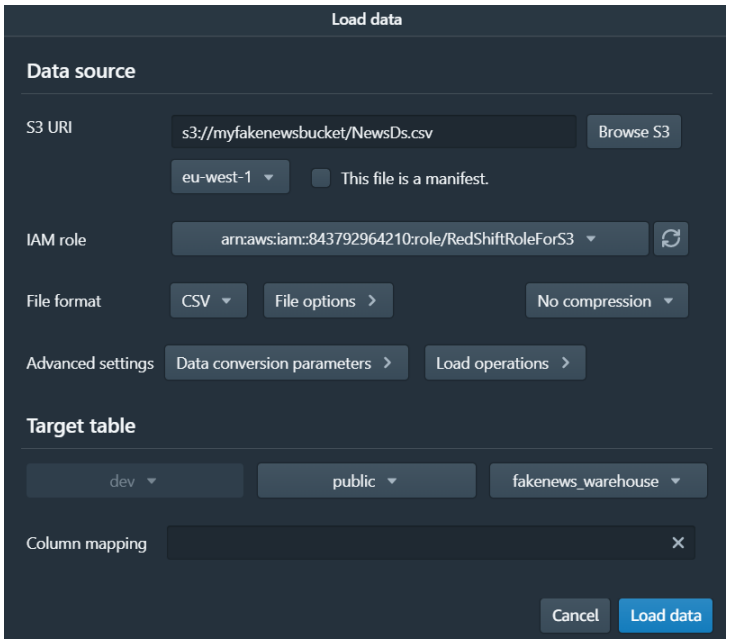


12. Now select the table and then click on ‘Load Data’. Select the bucket and file from S3.

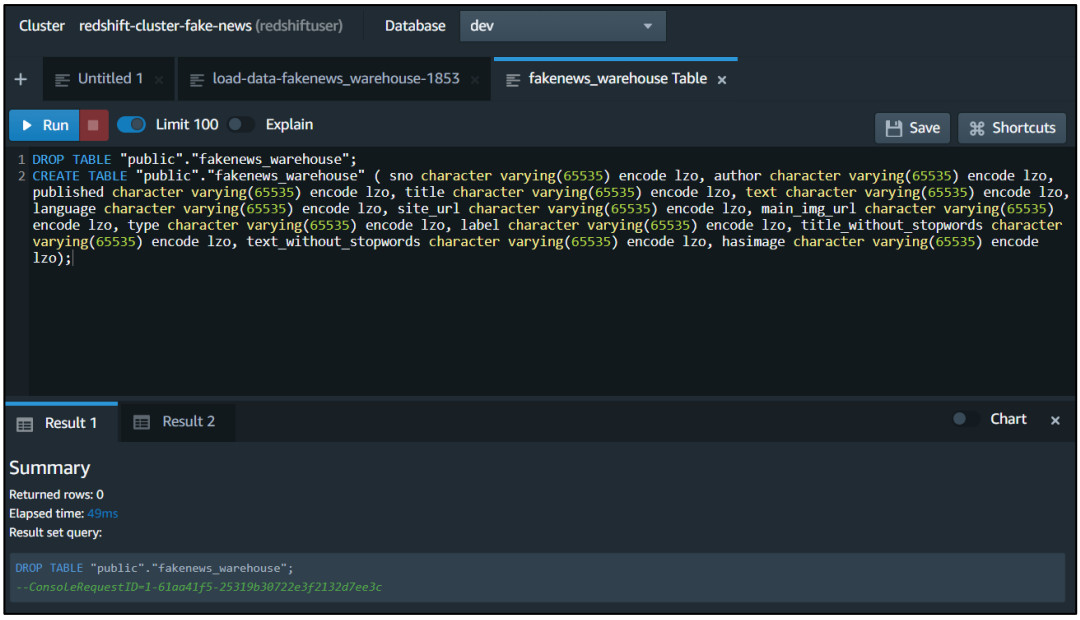


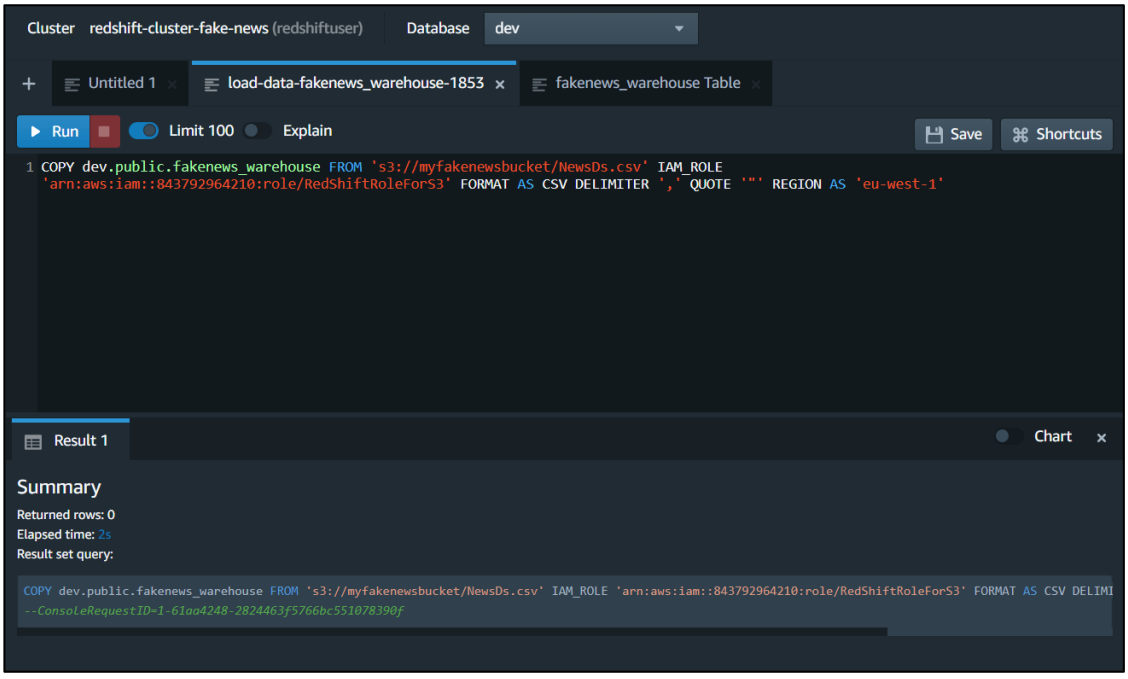


13. After selecting the file, attach the IAM role (RedShiftRoleForS3) and choose target table i.e. **public -> fakenews_warehouse.** Click on Load Data.

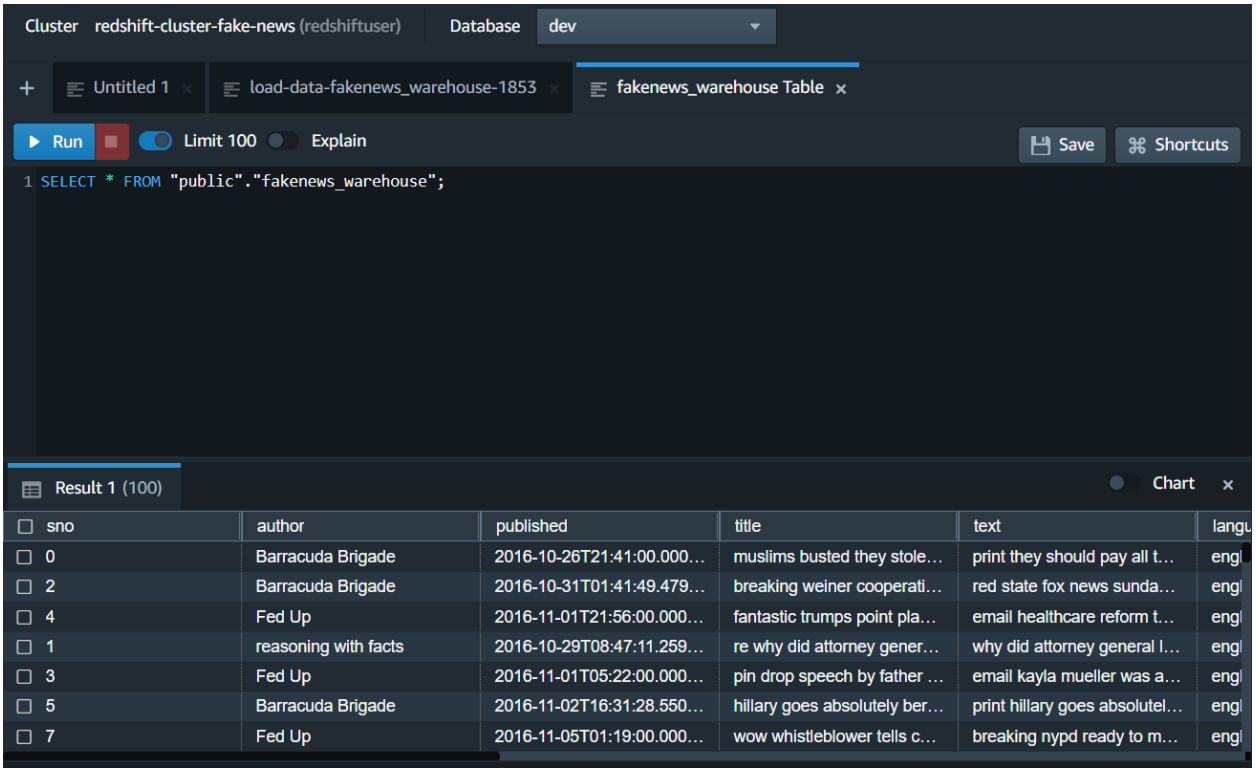


14. This may throw error as the columns are not designed to handle large number of characters. So we will drop the table and create table again with these two commands. To perform this task, right click on table and click on 'Show Table Definition'. We will moodify that by the following.
    DROP TABLE "public"."fakenews_warehouse";
    CREATE TABLE "public"."fakenews_warehouse" (sno character varying(65535) encode lzo, author character varying(65535) encode lzo, published character varying(65535) encode lzo, title character varying(65535) encode lzo, text character varying(65535) encode lzo, language character varying(65535) encode lzo, site_url character varying(65535) encode lzo, main_img_url character varying(65535) encode lzo, type character varying(65535) encode lzo, label character varying(65535) encode lzo, title_without_stopwords character varying(65535) encode lzo, text_without_stopwords character varying(65535) encode lzo, hasimage character varying(65535) encode lzo);
    Run the queries and you can see the summary. Then Run the query of loading data from S3, this time it will be successful.



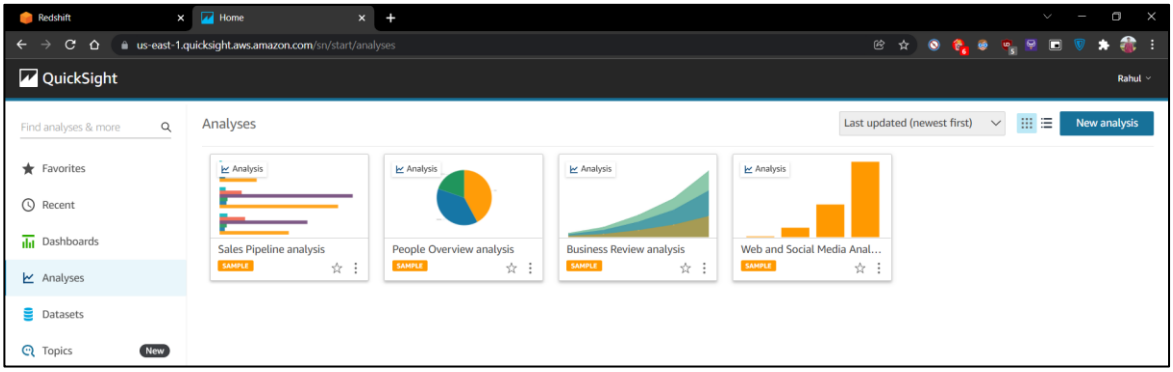


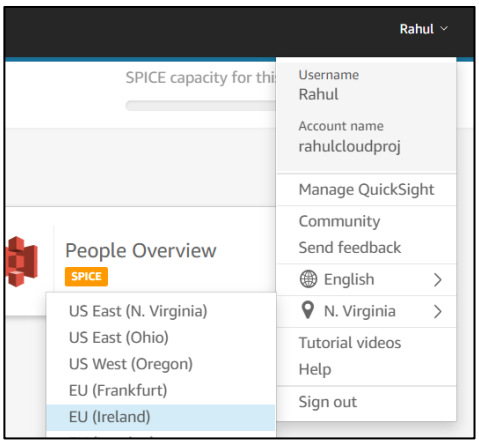
15. We can check the data using the query **Select * from “public”.”fakenews_earehouse”**

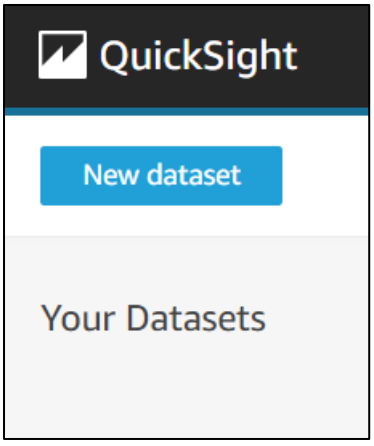


sno	author	published	title	text	langu
0	Barracuda Brigade	2016-10-26T21:41:00.000…	muslims busted they stole…	print they should pay all t…	engl
2	Barracuda Brigade	2016-10-31T01:41:49.479…	breaking weiner cooperati…	red state fox news sunda…	engl
4	Fed Up	2016-11-01T21:56:00.000…	fantastic trumps point pla…	email healthcare reform t…	engl
1	reasoning with facts	2016-10-29T08:47:11.259…	re why did attorney gener…	why did attorney general l…	engl
3	Fed Up	2016-11-01T05:22:00.000…	pin drop speech by father …	email kayla mueller was a…	engl
5	Barracuda Brigade	2016-11-02T16:31:28.550…	hillary goes absolutely ber…	print hillary goes absolutel…	engl
7	Fed Up	2016-11-05T01:19:00.000…	wow whistleblower tells c…	breaking nypd ready to m…	engl

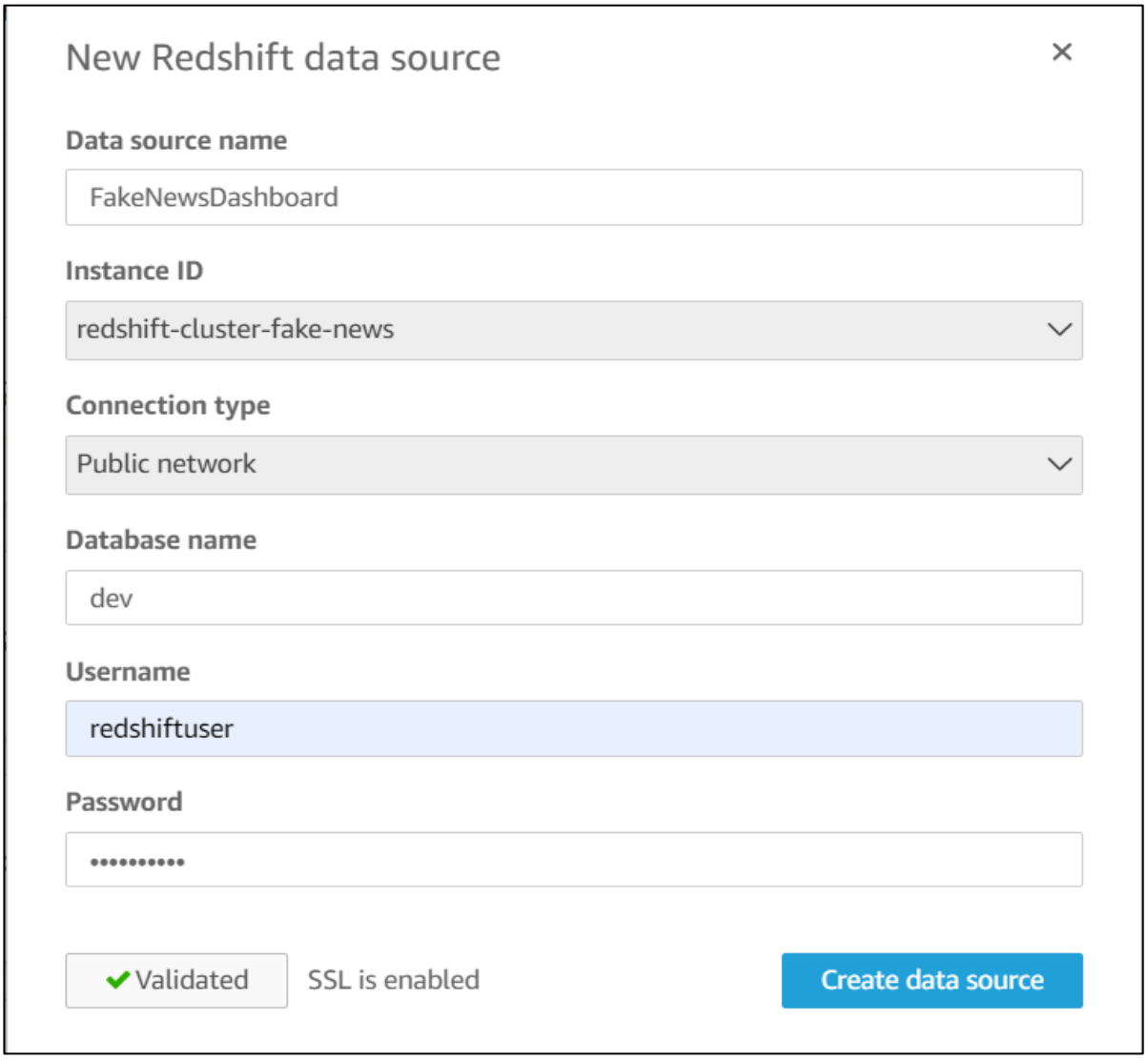
16. So loading data to warehouse is complete. Now search QuickSight on AWS Console. This will open a console like this. Change the region where RedShift cluster is hosted. In this scenario we hosted it in Ireland, change region to Ireland and then click on “New Analysis -> New Dataset”.



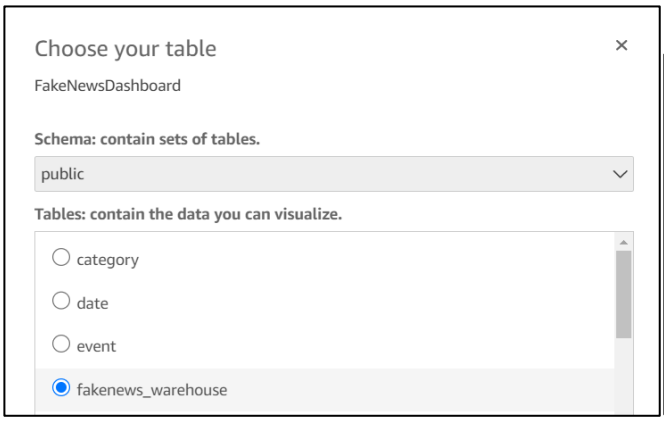


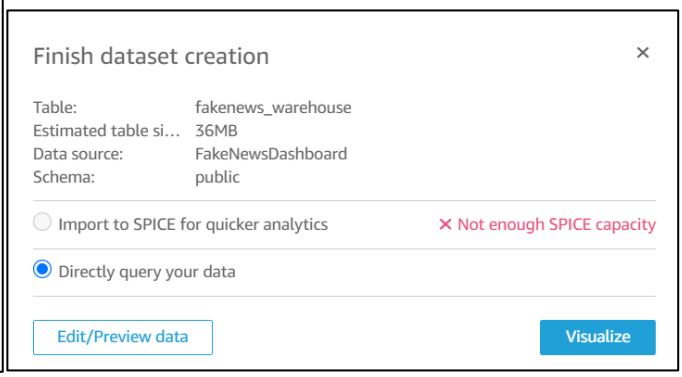


17. Provide a name to data source “FakeNewsDashboard” and select instance Id “redshift-cluster-fake-news” and then connection type as “Public”. Database name as “dev”, give credentails of redshift user and click on “Validate” and then “Create Data Source”.

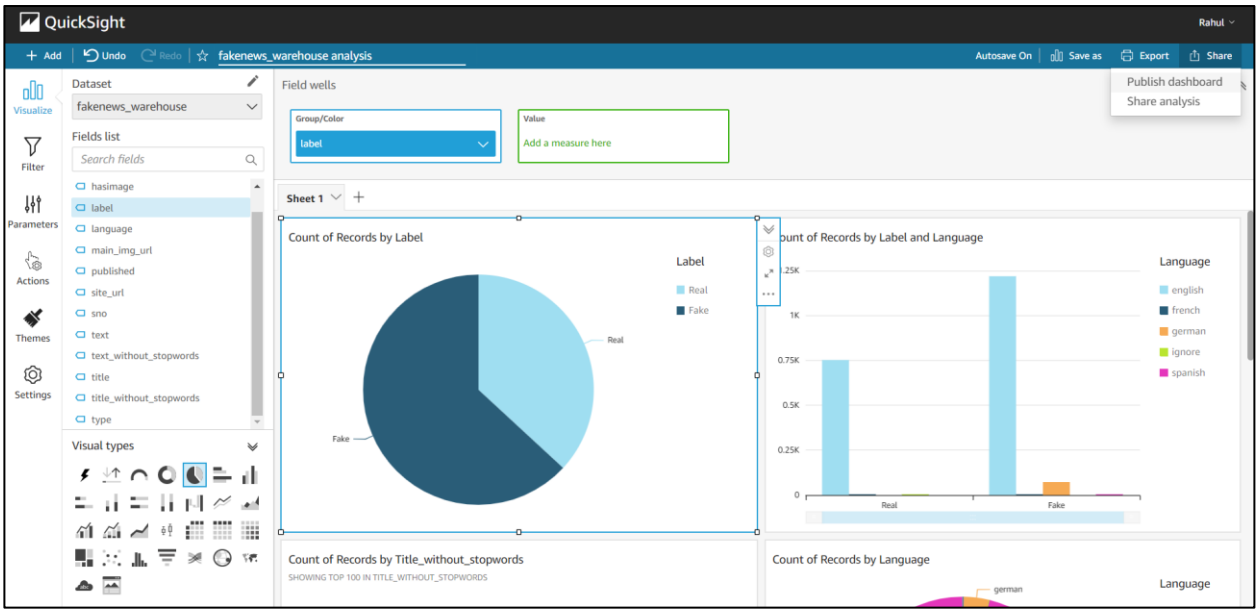


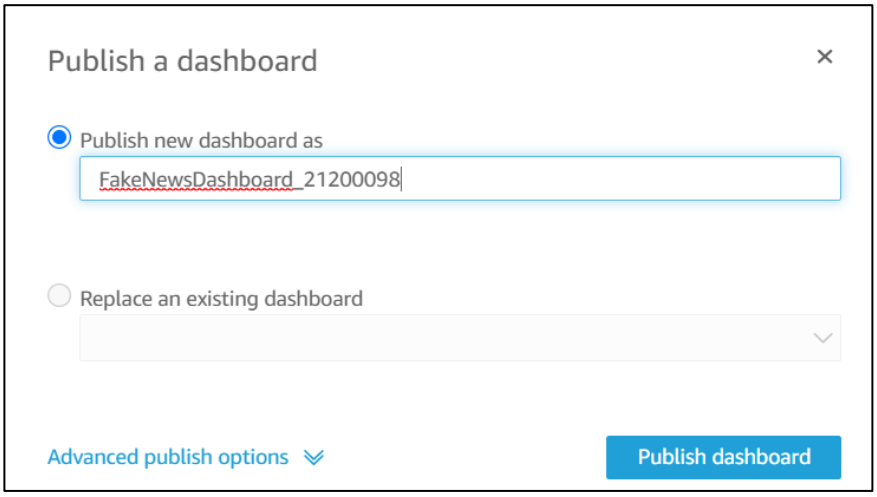
18. Choose the table to be visualized (fakenews_warehouse) and then to finish dataset creation select “Directly query your data”. Click “Visualize” to finish.





19. Select columns to create pie charts and bar graphs. After that click on “Share” and then to “Publish dashboard”. Publish the dashboard by giving a name “FakeNewsDashboard_21200098”.





CONCLUSION

I am delighted to say that I’m able to complete the project and as per stated requirements I preprocessed the data, hosted it in a cloud warehouse and created a BI dashboard for same.

https://eu-west-1.quicksight.aws.amazon.com/sn/dashboards/74e87c69-fe5d-44f6-984a-395817b72aa3

CHALLENGES I FACED

There were many challenges that I faced while completing the project. Few of them were when I was unable to load data to csv due to line break present in data, I had to preprocess the data again and then reupload it. Even after uploading it I figured out that size of column needed to be increased in order to fit the data in database. Later on I was struggling while connecting the RedShift to QuickSight due to security issue, I figured out later those inbound rules needed to be changed and RedShift access was to be added.

LEARNINGS

I would like to thank the faculty of Cloud Computing for giving me the opportunity to work on implementing a Cloud Data Warehouse Solution for Fake News, as it was a vast learning experience for me. Starting from understanding the use of data sets, necessity of cleaning the data as per the requirement in hand before starting the actual analysis to hosting a data warehouse on the cloud and creating a live dashboard to share the analysis the entire journey has been knowledge gaining. Also, this gave me some firsthand experience of getting stuck at various points and producing innovative solutions within a given period.