

# 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.

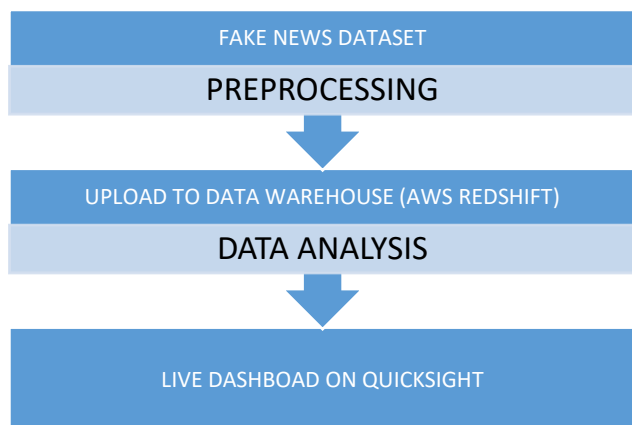
## REQUIREMENT FOR APPLICATION

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 IMPLMENT 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.

The image shows two side-by-side screenshots from the AWS S3 console. The left screenshot is titled 'Create bucket' and shows the 'General configuration' section. The 'Bucket name' field is filled with 'myfakenewsbucket', and the 'AWS Region' is set to 'EU (Ireland) eu-west-1'. The right screenshot is titled 'Upload' and shows a table of files and folders to be uploaded. The table has columns for Name, Folder, Type, and Size. A file named 'NewsDs.csv' is listed with a size of 10.4 MB. Below the table, the 'Destination' is set to 's3://myfakenewsbucket'.

Name	Folder	Type	Size
NewsDs.csv	-	application/vnd.ms-excel	10.4 MB

Wait untill data upload is complete.

The image shows a screenshot of the AWS S3 console's 'Uploading' progress bar. It indicates that the upload is 13% complete. The progress bar is blue with a white line showing the progress. Below the progress bar, it says 'Total remaining: 1 file: 9.1 MB(87.14%)', 'Estimated time remaining: 2 minutes', and 'Transfer rate: 95.2 KB/s'.

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)

Search IAM

Dashboard

Access management

User groups

Users

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.

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.

Search

Refresh

Delete

Create role

< 1 >

Settings

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

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

Refresh

Filter policies

Q s3

Showing 9 results

	Policy name	Used as
<input type="checkbox"/>	AmazonDMSRedshiftS3Role	None
<input checked="" type="checkbox"/>	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.

- 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.

The screenshot shows the 'Create cluster' page in the AWS Console. The breadcrumb navigation is 'Amazon Redshift > Clusters > Create cluster'. The page title is 'Create cluster' with an 'Info' link. Under 'Cluster configuration', the 'Cluster identifier' field is set to 'redshift-cluster-fake-news'. Below this, a message states: 'The identifier must be from 1-63 characters. Valid characters are a-z (lowercase only) and - (hyphen)'. Under 'What are you planning to use this cluster for?', the 'Free trial' option is selected. A note below states: 'When the free trial ends, delete your cluster to avoid incurring charges at on-demand rate for compute and storage. If you want to take a final snapshot of your cluster and store the snapshot on an S3, our on-demand rate applies.' At the bottom, the 'Calculated configuration summary' shows 'dc2.large | 1 node' and 'High performance with fixed local SSD storage'.

redshiftuser

Reduser123

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

The screenshot shows the 'Clusters (1)' page in the AWS Console. It features a search bar with the placeholder 'Filter clusters by property or value'. Below the search bar is a table with columns: 'Cluster', 'Cluster namespace', and 'Status'. There is one cluster listed: 'redshift-cluster-fake-ne...' with namespace '1f1f956f-af75-4447-99f8-0ead1aad82d' and status 'Modifying Creating'.

The screenshot shows the details page for the 'redshift-cluster-fake-news' cluster. It has buttons for 'Actions', 'Edit', 'Add partner integration', and 'Query data'. The 'General information' section contains the following details:

Property	Value
Cluster identifier	redshift-cluster-fake-news
Status	Available
Node type	dc2.large
Endpoint	redshift-cluster-fake-news.cczoov4l2hry.eu-...
Cluster namespace	1f1f956f-af75-4447-99f8-0ead1aad82d
Date created	December 03, 2021, 15:32 (UTC+00:00)
Number of nodes	1
JDBC URL	jdbc:redshift://redshift-cluster-fake-news.ccz...
Storage used	-
AQUA	Not available
ODBC URL	Driver={Amazon Redshift (x64)}; Server=reds...

Database configurations

Change admin user password

Rotate encryption keys

Edit

Database name

dev

Port

5439

Admin user name

redshiftuser

Parameter group

Defines database parameter and query queues for all the databases.

default.redshift-1.0

SSH ingestion setting (cluster public key)

ssh-rsa AAAAB3NzaC1yc2EAAAADAQAB...

Encryption

Disabled

AWS KMS key ID

-

Audit logging

Disabled

Network and security settings

Edit

Virtual private cloud (VPC)

vpc-809e20f9

Subnet

default

Endpoint URL

-

Availability Zone

eu-west-1a

Enhanced VPC routing

Disabled

VPC security group

Specify which instances and devices can connect to the cluster.

sg-caa3239c

Publicly accessible

Allow instances and devices outside the VPC to connect to your database through the cluster endpoint.

Disabled

- 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.

aws

Services

Search for services, features, blogs, docs, and more

[Alt+S]

Ireland

Rahul @ 8437-9296-4210

Capacity Reservations

Images

AMIs

Elastic Block Store

Volumes

Snapshots

Lifecycle Manager

Network & Security

Security Groups

Security Groups (1/1)

Filter security groups

search: sg-caa3239c

Clear filters

	Name	Security group ID	Security group name	VPC ID	Description	Owner
<input checked="" type="checkbox"/>	-	sg-caa3239c	default	vpc-809e20f9	default VPC security gr...	843792964210

Inbound rules

Security group rule ID

Type

Protocol

Port range

Source

Description - optional

sgr-011a25cbe0b31a910	Redshift	TCP	5439	Custom	0.0.0.0/0	
-----------------------	----------	-----	------	--------	-----------	--

Add rule

Cancel

Preview changes

Save rules

- 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.

Actions

Edit

Add partner

Manage cluster

Resize

Reboot

Relocate

Pause

Delete

Defer maintenance

Configure AQUA

Modify publicly accessible setting

Edit publicly accessible

Publicly accessible

Allow instances and devices outside the VPC to connect to your database through the cluster endpoint.

Disable

Enable

Elastic IP address

Specify the Elastic IP address used to connect to the cluster.

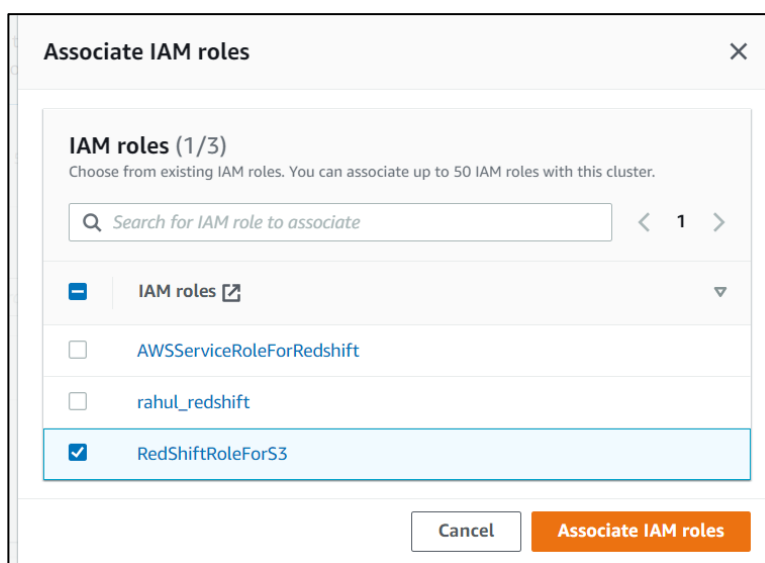
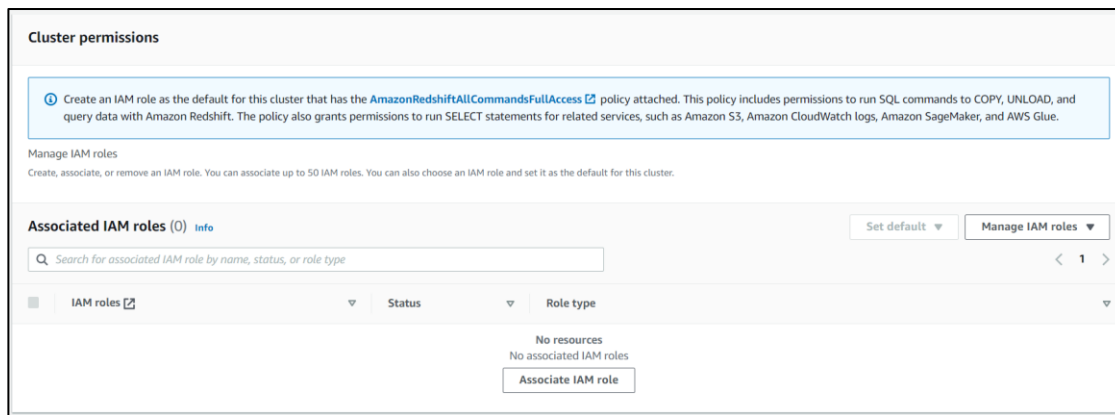
52.208.61.100

Your cluster might be unavailable for up to 10 minutes while this change to public accessibility is processed.

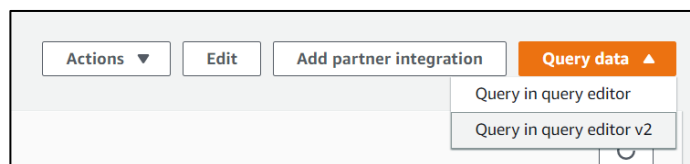
Cancel

Save changes

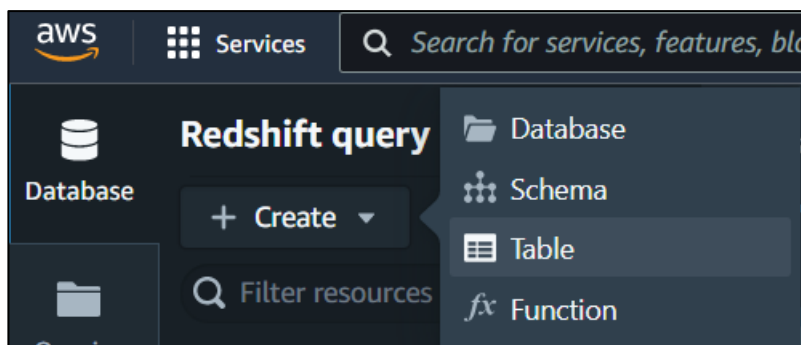
7. Now attach the IAM role created earlier to the RedShift cluster. Navigate to 'Cluster permission' and then click on 'Associate IAM Role'. After that select the role created earlier in this process.



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'.

**Create table**

Schema: public Table name: fakenews\_warehouse Load from CSV + Add field Column options

Column name Data type Encoding

Column name	Data type	Encoding
sno	VARCHAR	No selection
author	VARCHAR	No selection
published	VARCHAR	No selection
title	VARCHAR	No selection
text	VARCHAR	No selection
language	VARCHAR	No selection
site_url	VARCHAR	No selection
main_img_url	VARCHAR	No selection
type	VARCHAR	No selection
label	VARCHAR	No selection
title_without_stop...	VARCHAR	No selection
text_without_stop...	VARCHAR	No selection
hasImage	VARCHAR	No selection

Default value: Custom, Empty string, NULL, No default value (selected)

Automatically increment: Enable (disabled)

Not NULL: Enable (disabled)

Size: 65565

Keys: Primary key, Unique key (disabled)

Cancel Reset Open query in editor Create table

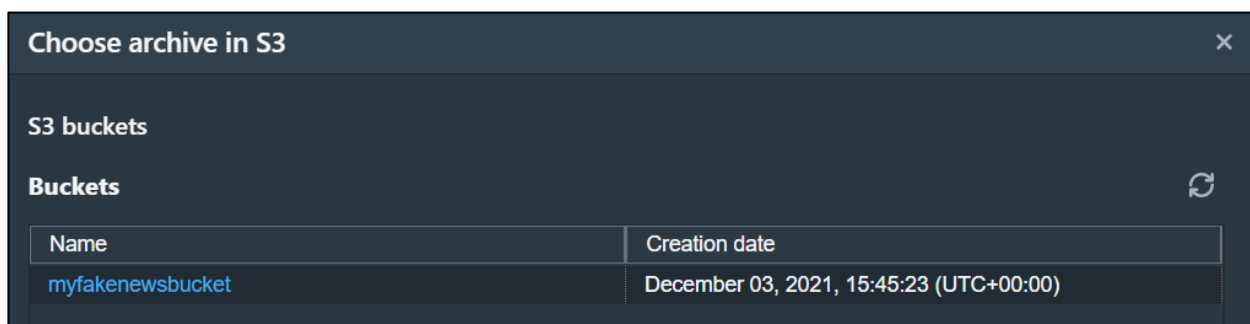
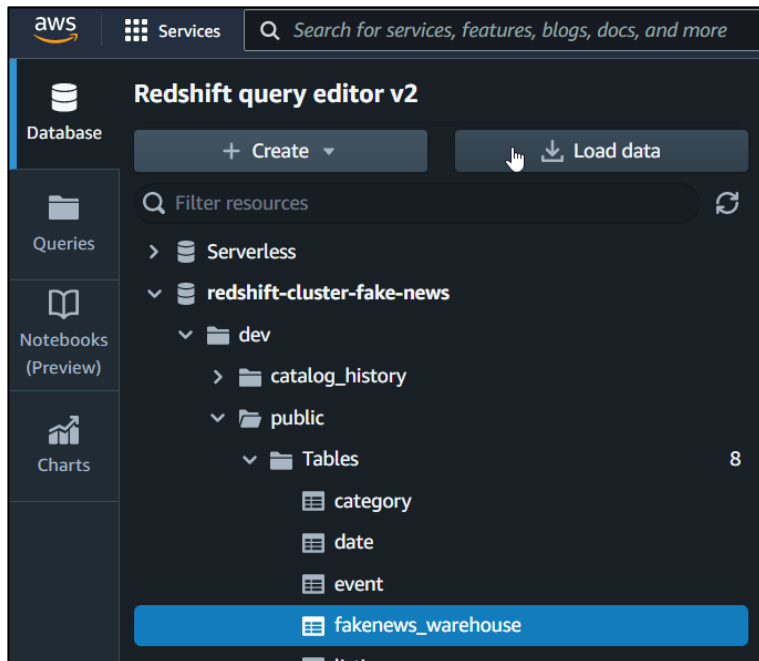
11. Check once again if table is created under public -> Tables -> fakenews\_warehouse. Clicking on table will generate the schema below.

**fakenews\_warehouse**

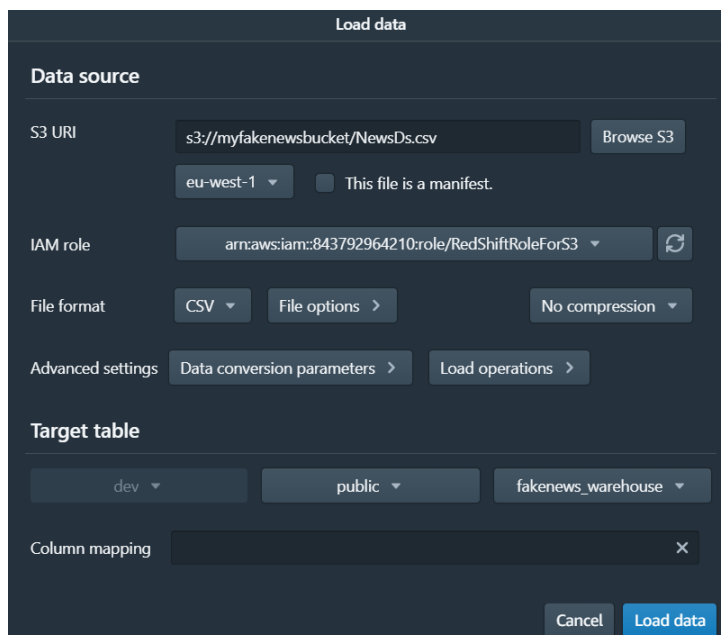
Field	Type	NL	CMP
A sno	character varying(256)	NULL	lzo
A author	character varying(256)	NULL	lzo
A published	character varying(256)	NULL	lzo
A title	character varying(256)	NULL	lzo



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 modify that by the following.

```
DROP TABLE "public"."fakenews_warehouse";
```

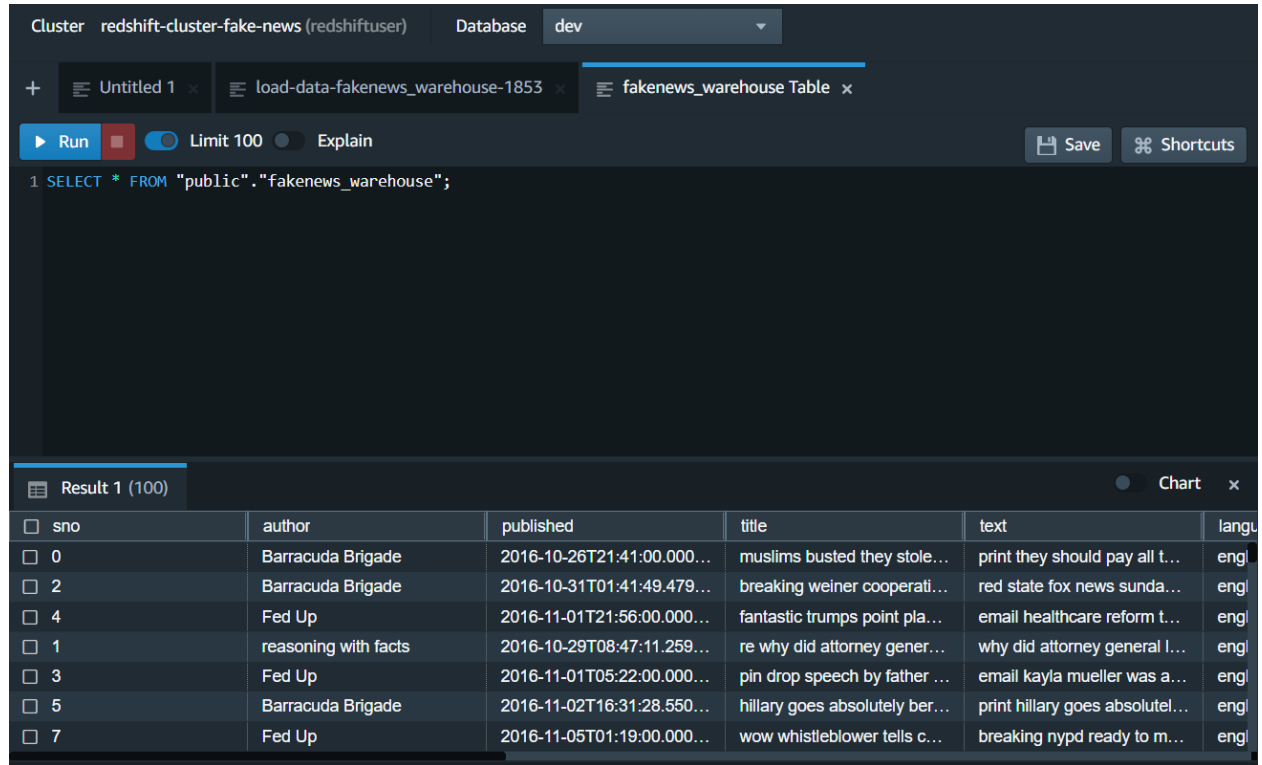
```
CREATE TABLE "public"."fakenews_warehouse" ( sno character varying(65535) encode lzo, author character varying(65535) encode lzo, published character varying(65535) encode lzo, title character varying(65535) encode lzo, text character varying(65535) encode lzo, language character varying(65535) encode 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.

The screenshot shows the AWS Redshift console interface. At the top, the cluster is 'redshift-cluster-fake-news (redshiftuser)' and the database is 'dev'. The query editor shows two SQL commands: 1. DROP TABLE "public"."fakenews\_warehouse"; 2. 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); The 'Run' button is highlighted. Below the query editor, the 'Result 1' tab is selected, showing a 'Summary' section with 'Returned rows: 0', 'Elapsed time: 49ms', and 'Result set query: DROP TABLE "public"."fakenews\_warehouse";'. A console log at the bottom shows the request ID: --ConsoleRequestID=1-61aa41f5-25319b30722e3f2132d7ee3c.

The screenshot shows the AWS Redshift console interface. At the top, the cluster is 'redshift-cluster-fake-news (redshiftuser)' and the database is 'dev'. The query editor shows a single SQL command: 1. COPY dev.public.fakenews\_warehouse FROM 's3://myfakenewsbucket/NewsDs.csv' IAM\_ROLE 'arn:aws:iam::843792964210:role/RedShiftRoleForS3' FORMAT AS CSV DELIMITER ',' QUOTE '''' REGION AS 'eu-west-1'. The 'Run' button is highlighted. Below the query editor, the 'Result 1' tab is selected, showing a 'Summary' section with 'Returned rows: 0', 'Elapsed time: 2s', and 'Result set query: COPY dev.public.fakenews\_warehouse FROM 's3://myfakenewsbucket/NewsDs.csv' IAM\_ROLE 'arn:aws:iam::843792964210:role/RedShiftRoleForS3' FORMAT AS CSV DELIMI'. A console log at the bottom shows the request ID: --ConsoleRequestID=1-61aa4248-2824463f5766bc551078390f.

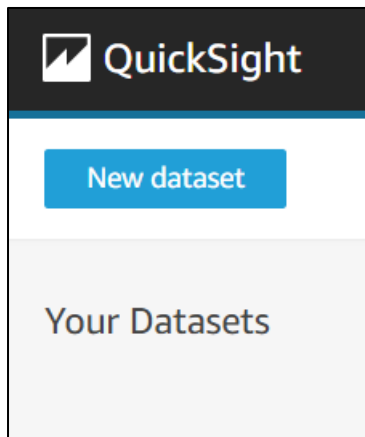
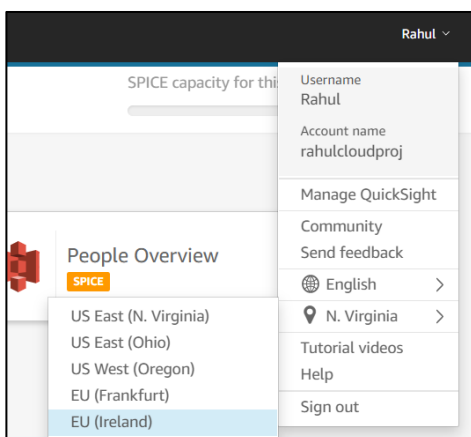
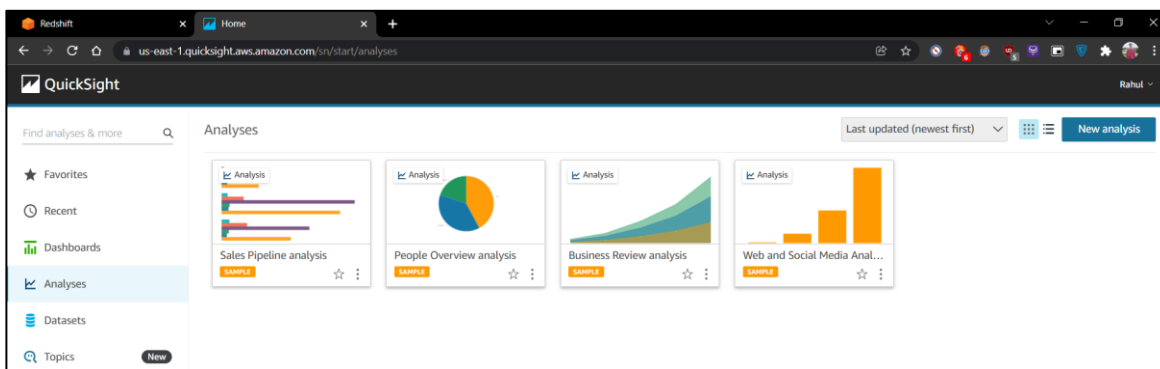
15. We can check the data using the query **Select \* from “public.”fakenews\_earhouse”**



The screenshot shows the AWS Redshift console interface. At the top, the cluster is 'redshift-cluster-fake-news (redshiftuser)' and the database is 'dev'. The query editor shows the SQL query: `1 SELECT * FROM "public"."fakenews_warehouse";`. Below the query, the 'Result 1 (100)' tab is active, displaying a table with 6 columns: sno, author, published, title, text, and language. The table contains 8 rows of data.

sno	author	published	title	text	language
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 unda...	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 I...	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 credentials of redshift user and click on “Validate” and then “Create Data Source”.

### New Redshift data source ×

**Data source name**

**Instance ID**

redshift-cluster-fake-news ▼

**Connection type**

Public network ▼

**Database name**

**Username**

**Password**

✓ Validated

SSL is enabled

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.

### Choose your table ×

FakeNewsDashboard

Schema: contain sets of tables.

public ▼

Tables: contain the data you can visualize.

☐ category

☐ date

☐ event

☒ fakenews\_warehouse

### Finish dataset creation ×

Table: fakenews\_warehouse

Estimated table si... 36MB

Data source: FakeNewsDashboard

Schema: public

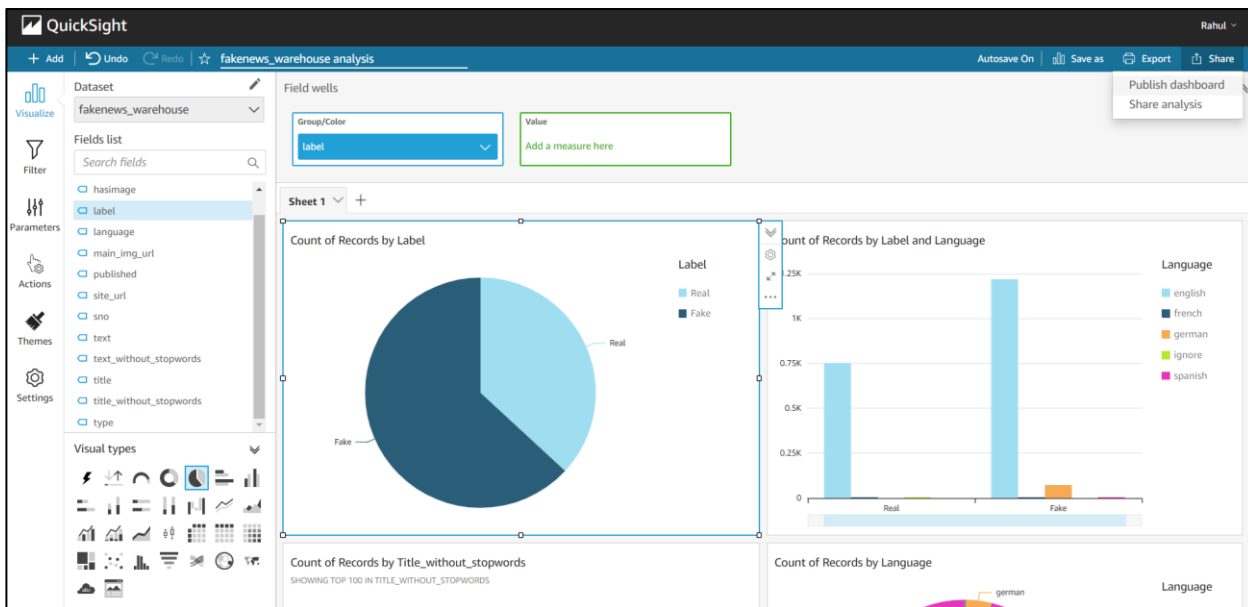
☐ Import to SPICE for quicker analytics ✗ Not enough SPICE capacity

☒ Directly query your data

Edit/Preview data

Visualize

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”.



The screenshot shows the 'Publish a dashboard' dialog box in Amazon QuickSight. The dialog box has a title bar with a close button. It contains two radio buttons: 'Publish new dashboard as' (selected) and 'Replace an existing dashboard'. Below the 'Publish new dashboard as' radio button is a text input field containing the text 'FakeNewsDashboard\_21200098'. Below the 'Replace an existing dashboard' radio button is a dropdown menu. At the bottom left, there is a link 'Advanced publish options' with a chevron icon. At the bottom right, there is a blue button labeled 'Publish dashboard'.

## 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.