**Cloud Project 2**

University: [San Jose State University (SJSU)](http://www.sjsu.edu/)

Course: [Cloud Technologies](http://info.sjsu.edu/web-dbgen/catalog/courses/CMPE281.html)

Professor: [Sanjay Garje](https://www.linkedin.com/in/sanjaygarje/)

Team: **TechnoSpartan**

|  |
| --- |
| Name |
| [Parvathi Pai](https://www.linkedin.com/in/parvathipai/) |
| [Shreya Ghotankar](https://www.linkedin.com/in/shreya-ghotankar/) |
| [Sania Gonsalves](https://www.linkedin.com/in/sania-gonsalves-8b961962/) |

Table of Contents

[**Project Idea** 3](#_Toc57710037)

[**Architecture Diagram** 3](#_Toc57710038)

[**Technologies and Services Used** 4](#_Toc57710039)

[**Features** 5](#_Toc57710040)

[**UI Screenshots** 5](#_Toc57710041)

[**AWS Configuration** 7](#_Toc57710042)

[**Twitter Data Resources:** 7](#_Toc57710043)

[**Databricks configurations** 9](#_Toc57710044)

[**COVID-19 Data Resources** 9](#_Toc57710045)

[**Front-end Deployment and Login Implementation** 13](#_Toc57710046)

[**CI/CD setup** 19](#_Toc57710047)

[**Future and Monetization** 19](#_Toc57710048)

[**References:** 20](#_Toc57710049)

# **Project Idea**

The severity of COVID-19 is not acknowledged by a lot of people which leads to incorrect information and decisions; to be regretted later. During this holiday season many are meeting and travelling across the world ignoring the consequences.

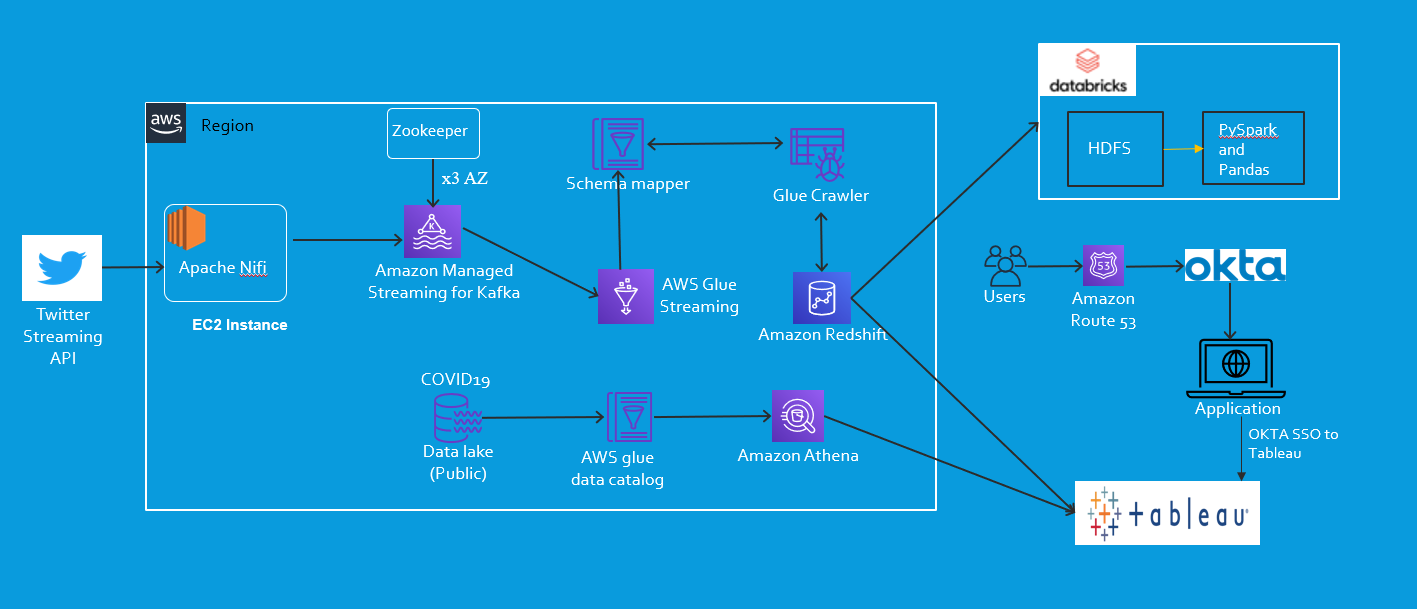
The data available today is humungous for any individual to manage and comprehend on their own.

Our project provides insights to people on the global rise of COVID-19 cases along with information on where people are meeting or planning to meetup. With our project we want to help people understand that COVID-19 should be taken seriously. People should be aware of the locations where the cases are high and for the same location they can see if people are meeting or not.

Our web application provides following visualizations –

* Data extracted from Twitter filtered over keywords like meetup, meet
* Real-time COVID-19 new cases, new deaths
* Predicted Deaths

# **Architecture Diagram**



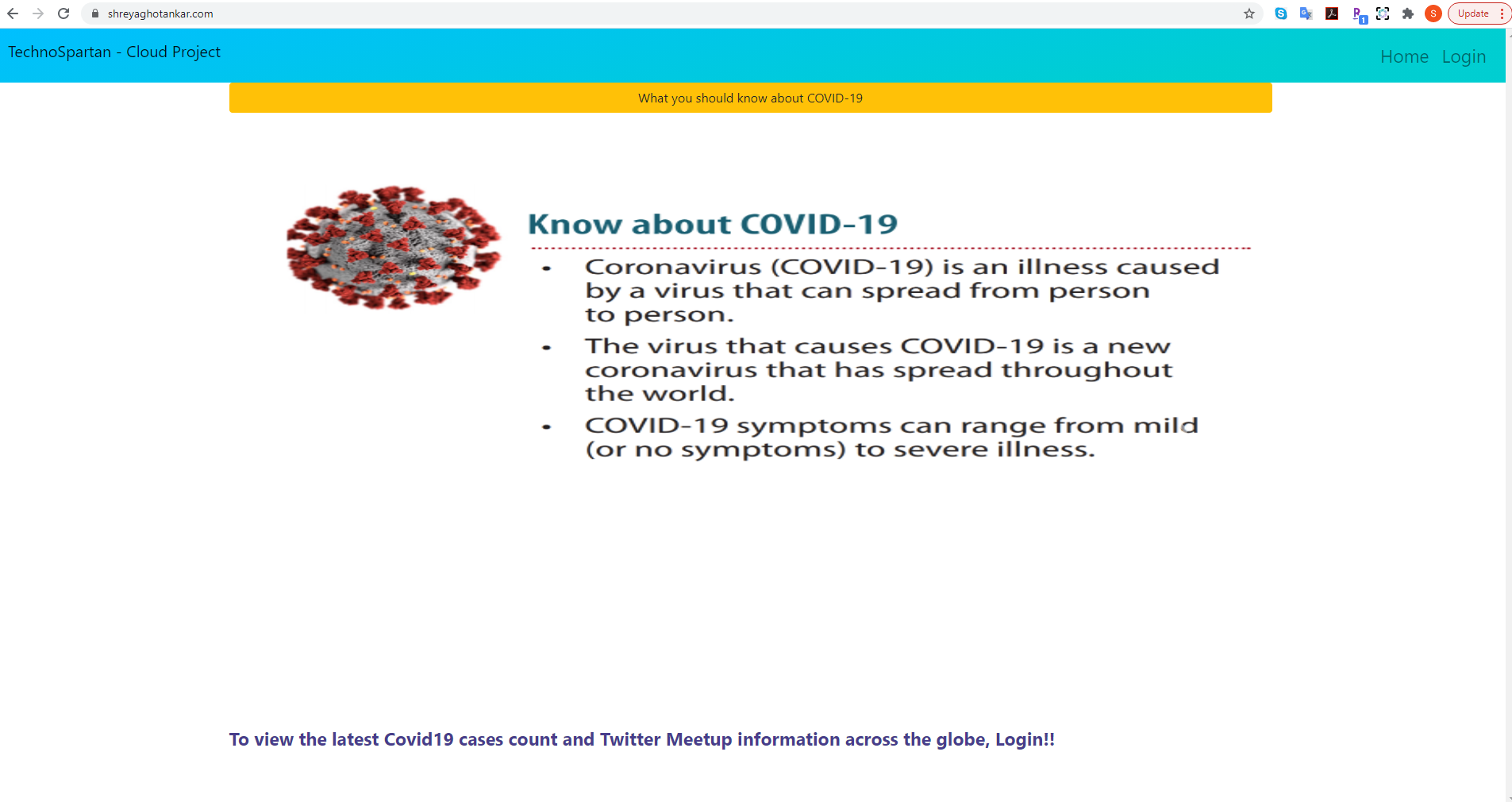
# **Technologies and Services Used**

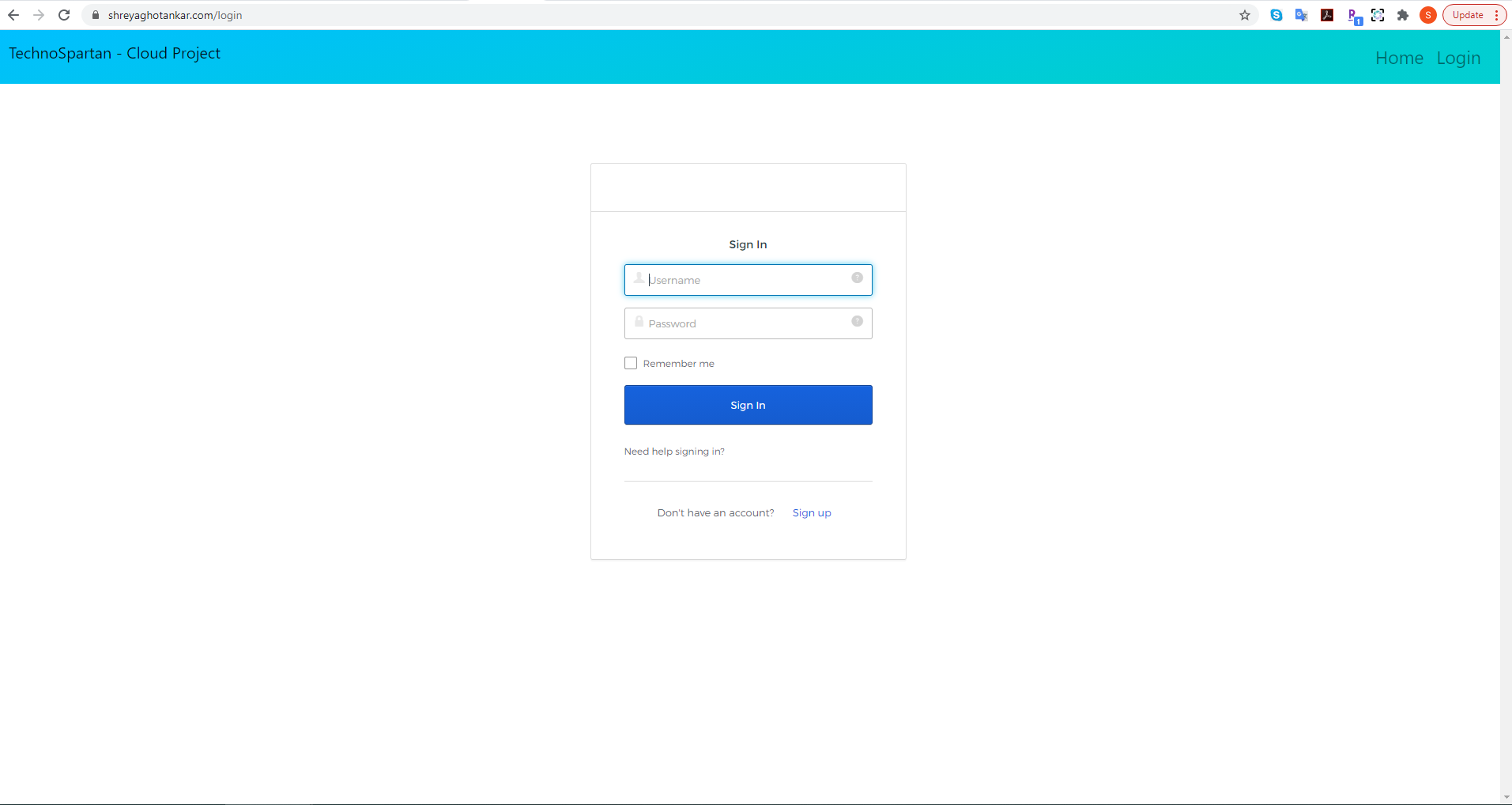
* For Web Application:
  + React JS
  + Amazon Route 53
  + Okta
* For COVID-19 Data:
* AWS S3 Data Lake
* AWS Glue Data Catalog
* Amazon Athena
* For Twitter Data:
* AWS EC2
* Kafka Stream – Amazon MSK
* Apache NIFI
* AWS Glue streaming
* AWS Glue Crawler
* Amazon Redshift
* For Data Visualization:
  + Tableau Desktop
  + Tableau Online Server (Trial-version)
* AWS CloudFormation
* AWS Amplify
* Data Analysis:
  + Databricks
  + HDFS
  + PySpark and Python

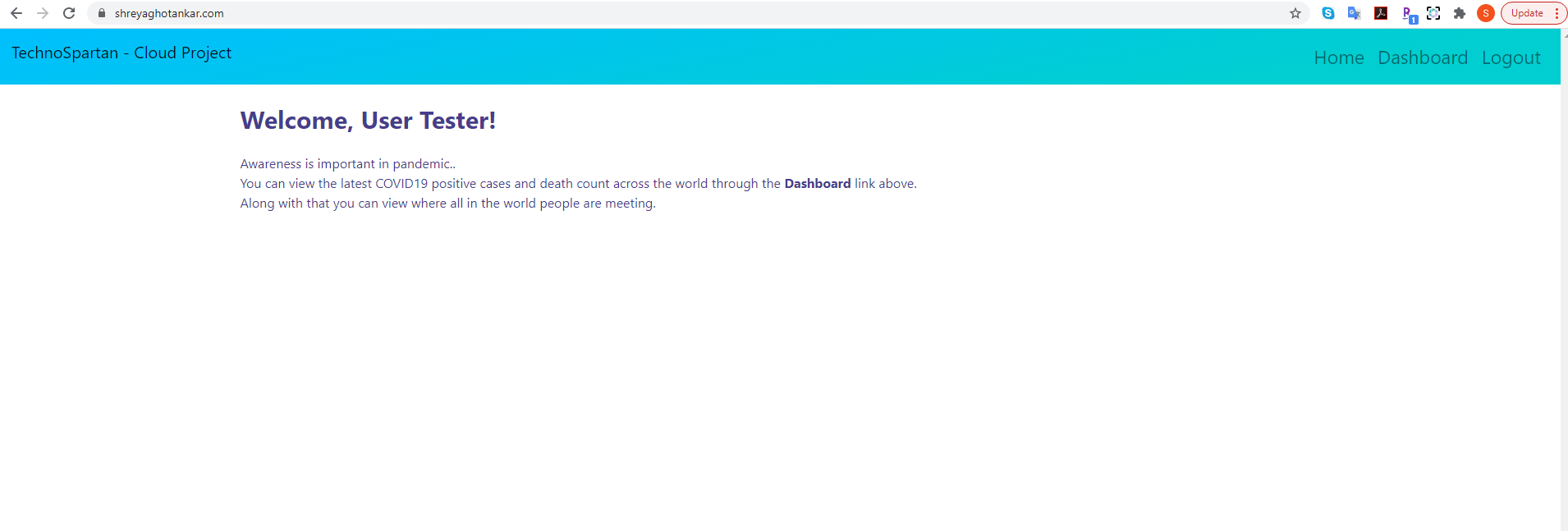
# **Features**

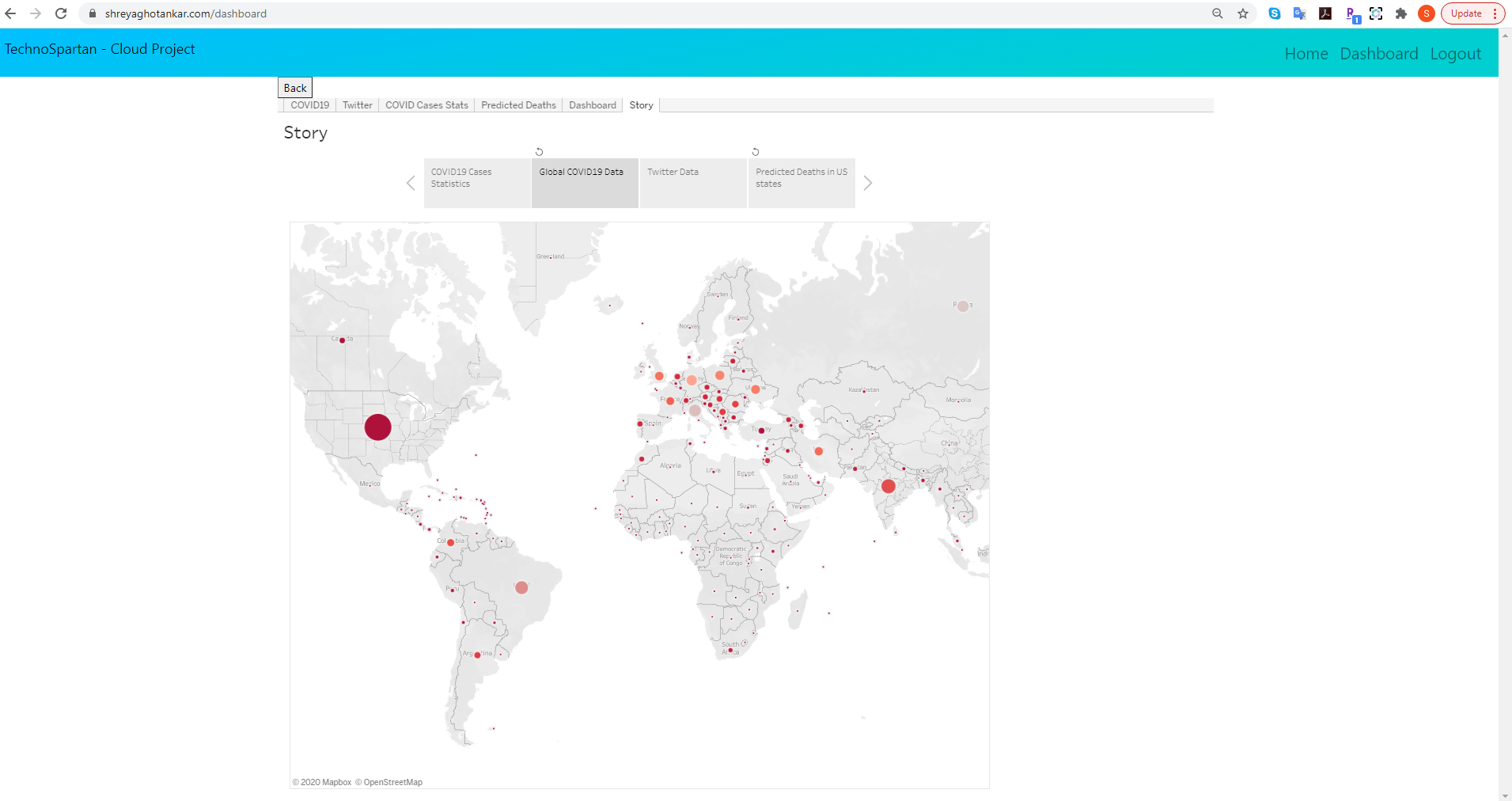
* Authentication and Authorization using OKTA.
* Single Sign-On to Tableau via OKTA.
* Real-time Data visualization on Tableau.
* Big data transformation, routing and visualization using Apache NIFI.
* Distributed coordination service using Zookeeper.
* Analytical data storage using Amazon Redshift
* Queuing mechanism using Amazon Managed Streaming for Apache Kafka.
* ETL jobs and data integration using AWS Glue
* Application has high availability and low latency.

# **UI Screenshots**









# **AWS Configuration**

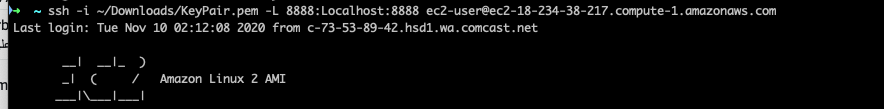
## **Twitter Data Resources:**

Steps: -

* Launch CloudFormation stack in us-east-1 the stack information is provided in the GitHub.Graphical user interface, text, application

  Description automatically generated

1. Login to the EC2 instance



1. Open the Nifi port from the EC2 to be projected in the local host

Graphical user interface, application, table, Excel

Description automatically generated

1. Create the connections between Glue and Redshift

Graphical user interface, application

Description automatically generated

1. Define Json Schema of tweets in Glue

Graphical user interface, application, Teams

Description automatically generated

1. Run the Glue crawler. The data will be populated in Redshift as shown in the figure.

Graphical user interface, text, application

Description automatically generated

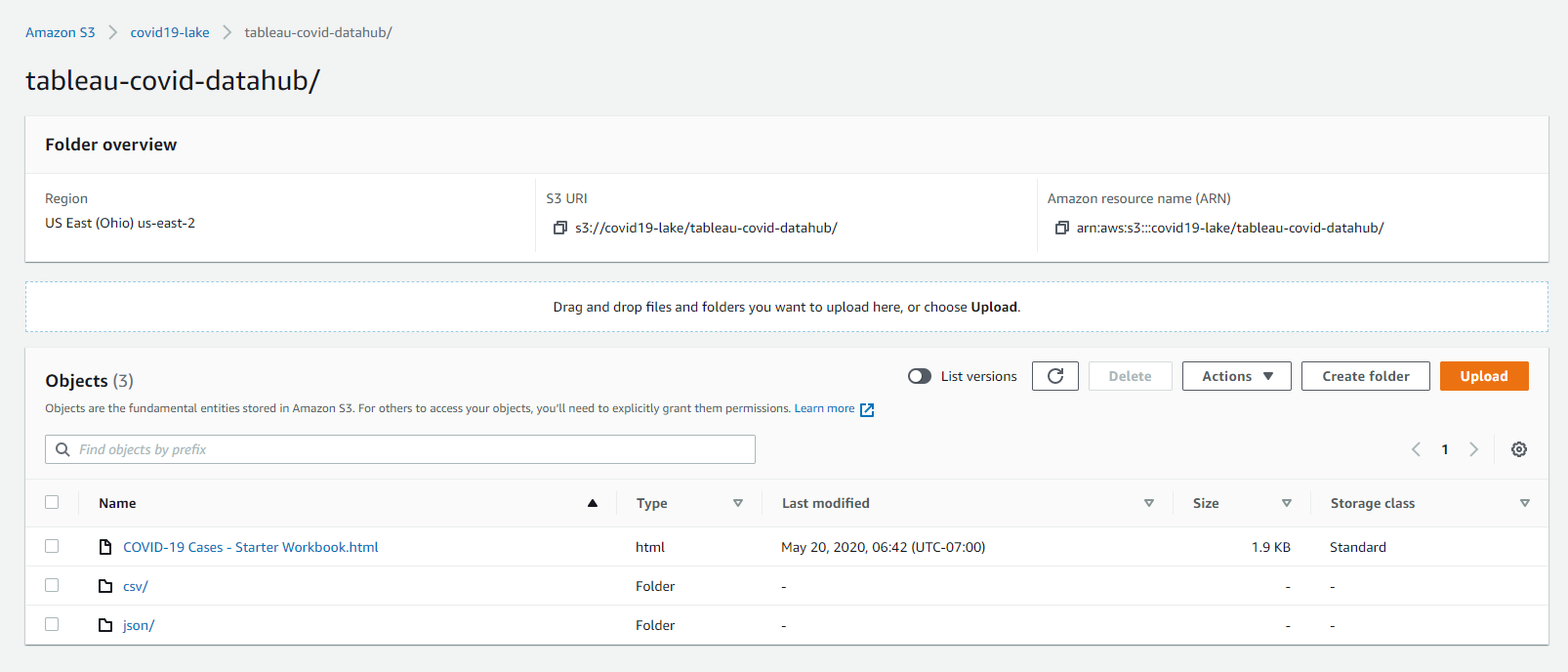
## **Databricks configurations**

1. Create an account on databricks
2. And use the link – for more analysis on visualization

<https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/8041837329188487/1492665670051231/419884982547132/latest.html>

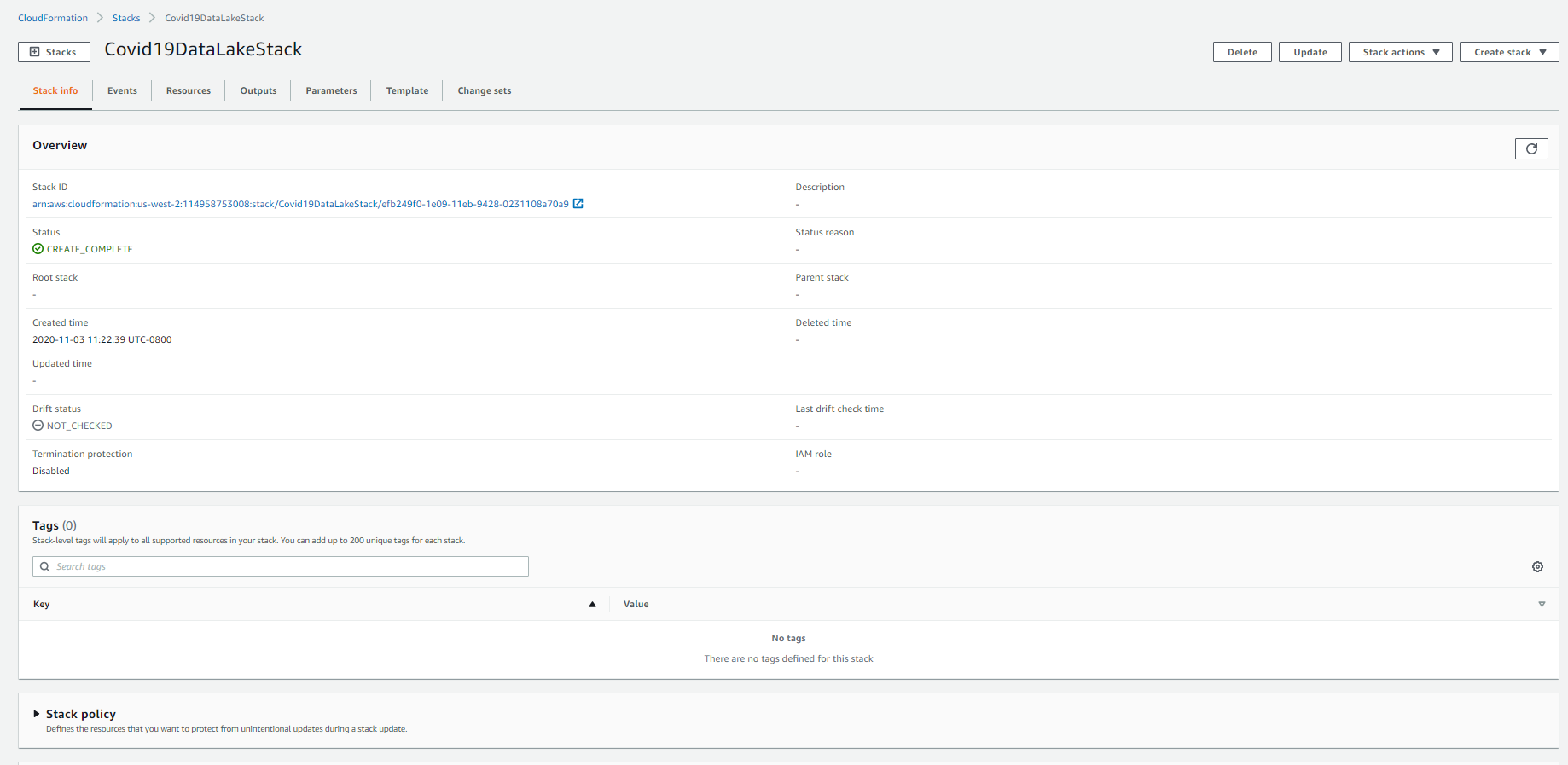
## **COVID-19 Data Resources**

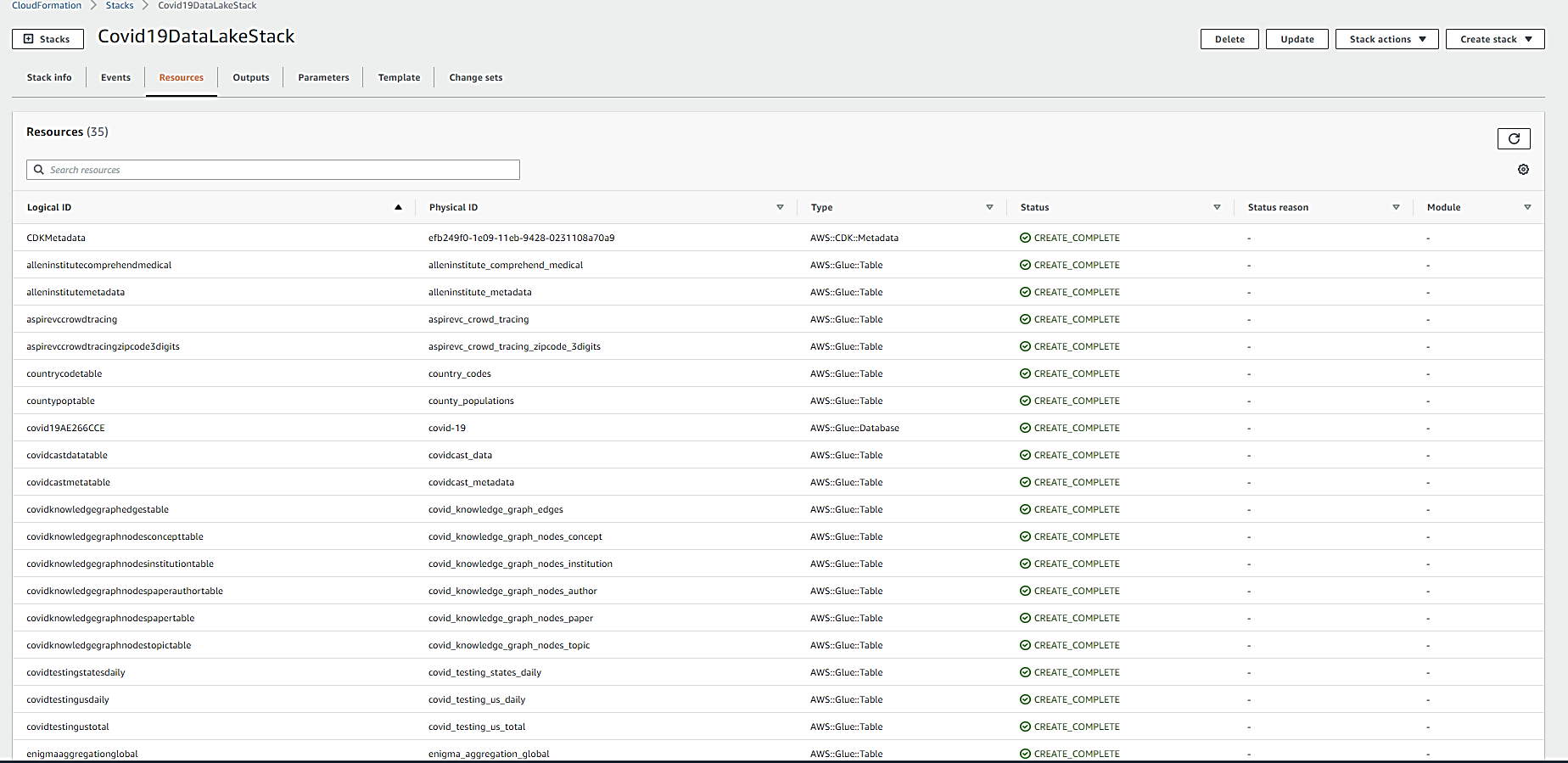
AWS S3 Data Lake

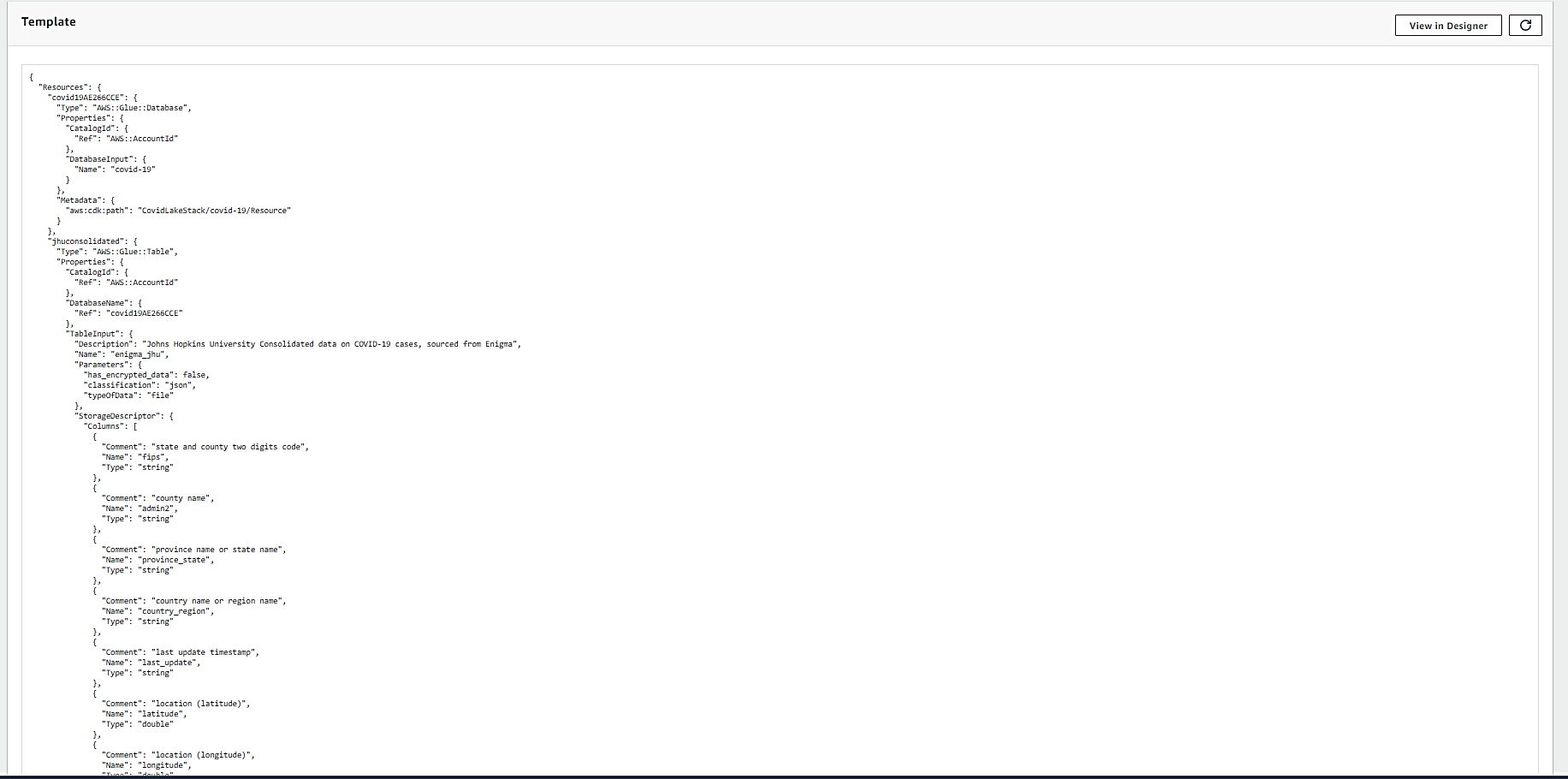


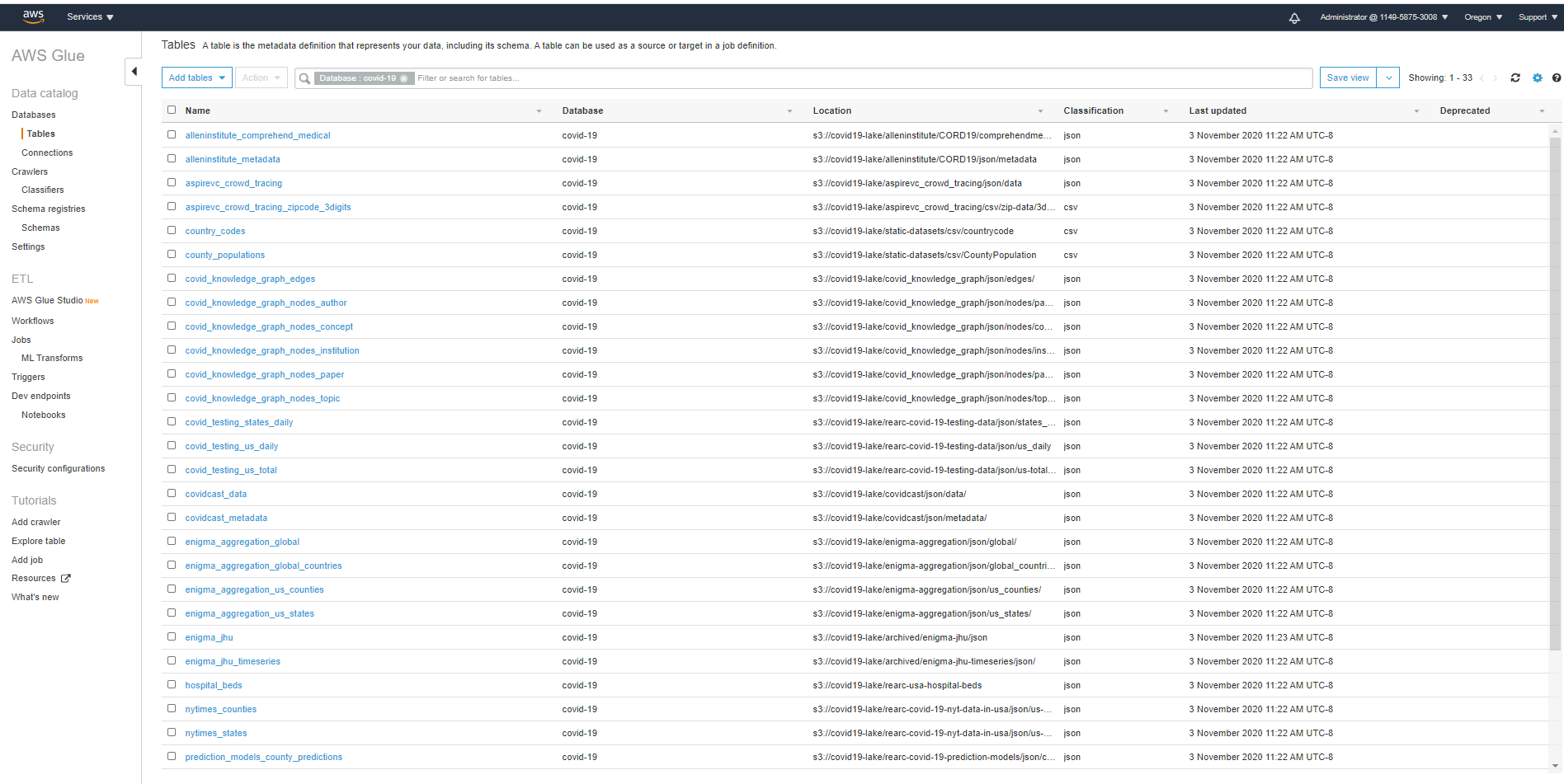
Use this CloudFormation [template](https://covid19-lake.s3.us-east-2.amazonaws.com/cfn/CovidLakeStack.template.json) to create AWS Glue Data Catalog and database to access the publicly available COVID-19 and query using Athena.

AWS CloudFormation Stack for COVID-19 data lake



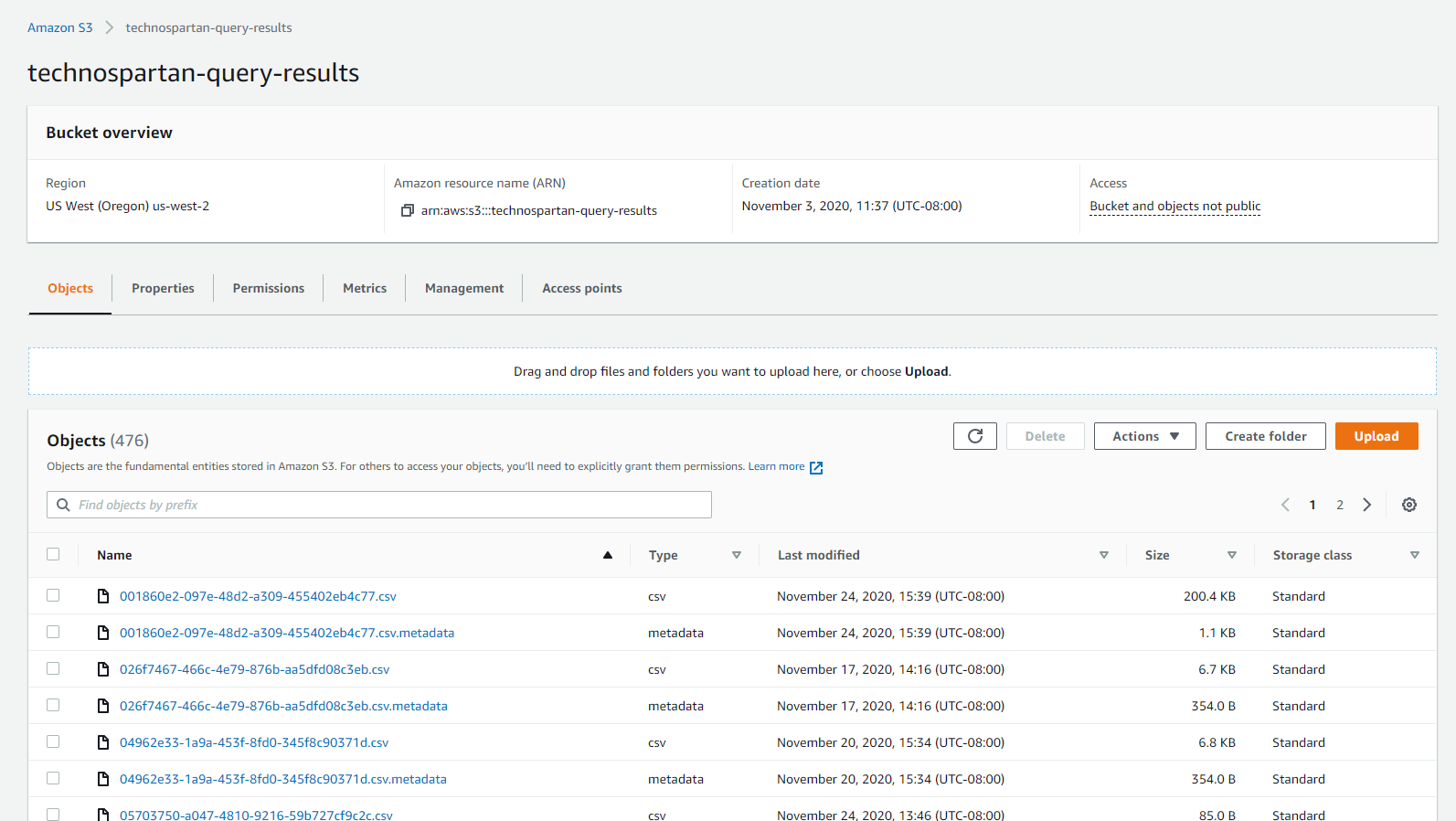




AWS Glue – Data Catalog

Amazon Athena



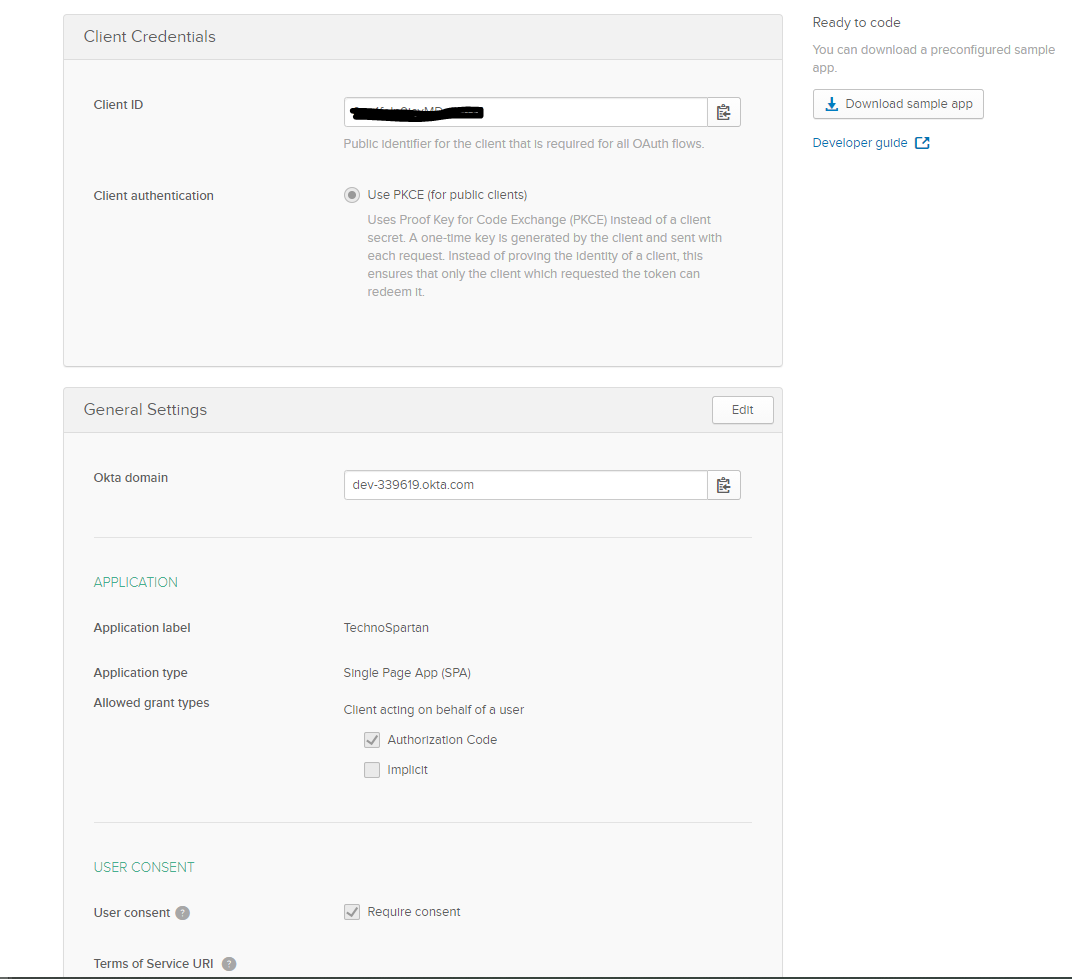
AWS S3 bucket for Athena query results

# **Front-end Deployment and Login Implementation**

This is a React JS application which displays visualization via Tableau. Following are the implementation details –

* Authenticate and Authorize Users

This has been achieved through OKTA as identity provider. We created application in okta org for our application and used the okta sign-in widget for sign-up and sign-in functionality.



* To identify if people are meeting -

For determining this scenario, we have streamed data from Twitter using Kafka streaming service available in AWS and stored relevant data in Amazon Redshift.

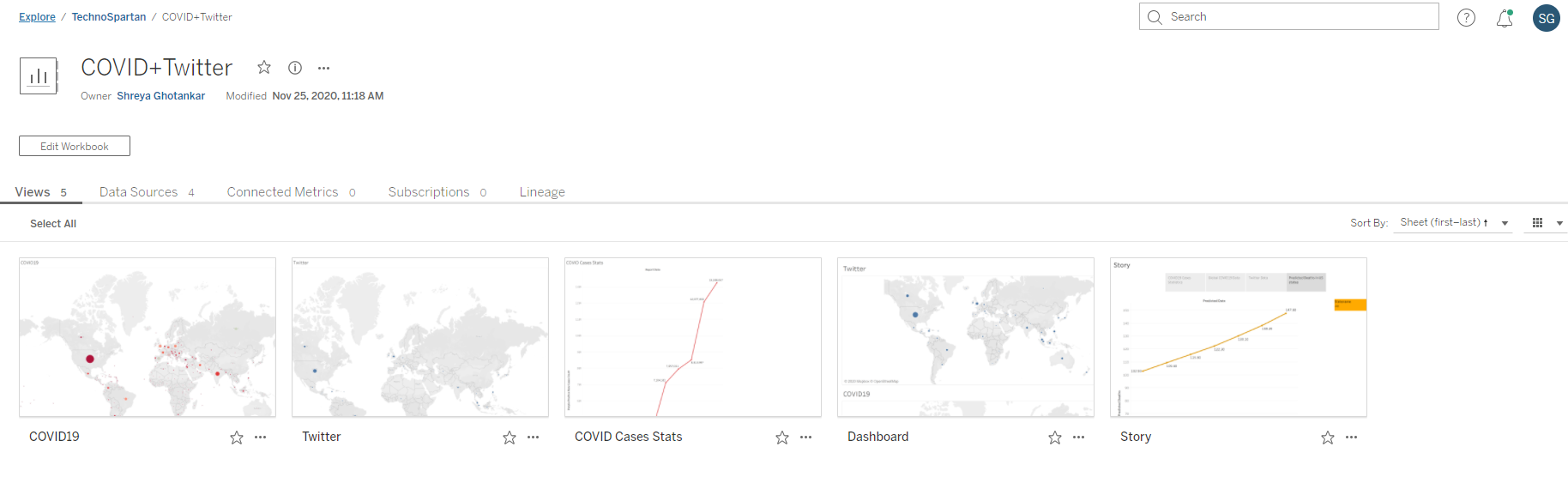
* COVID-19 Data:

To get the COVID-19 data that is valid, we have used public COVID-19 data lake made available by AWS. This data has been provided by multiple trusted sources like John Hopkins university, NYTimes, Tableau, REARC, AspireVC etc.

This data is then mapped to an AWS Glue Database and Amazon Athena is used to query this gigantic data.

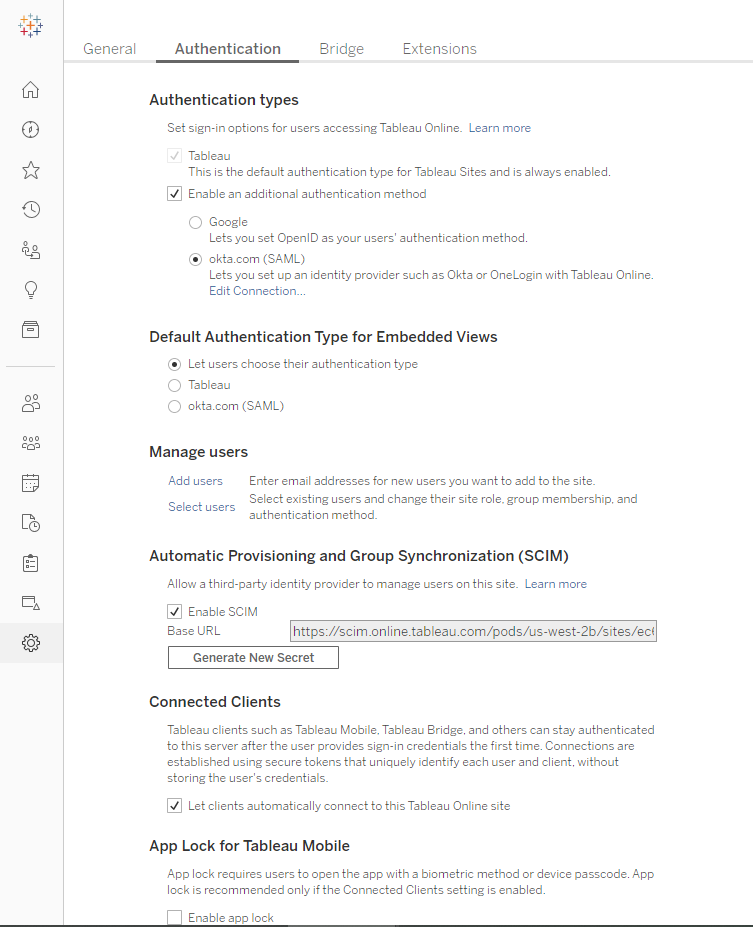
* Data Visualization:

To provide a global and country specific data to users we have utilized Tableau. Tableau provides map views and graphs to depict the data in a way that it is easier for user to understand. Tableau can connect to Amazon Redshift as well as to Amazon Athena where our data is residing. We have leveraged this feature and created our visualization components. Also, Tableau helps convert string location values to geographic locations quickly for use.

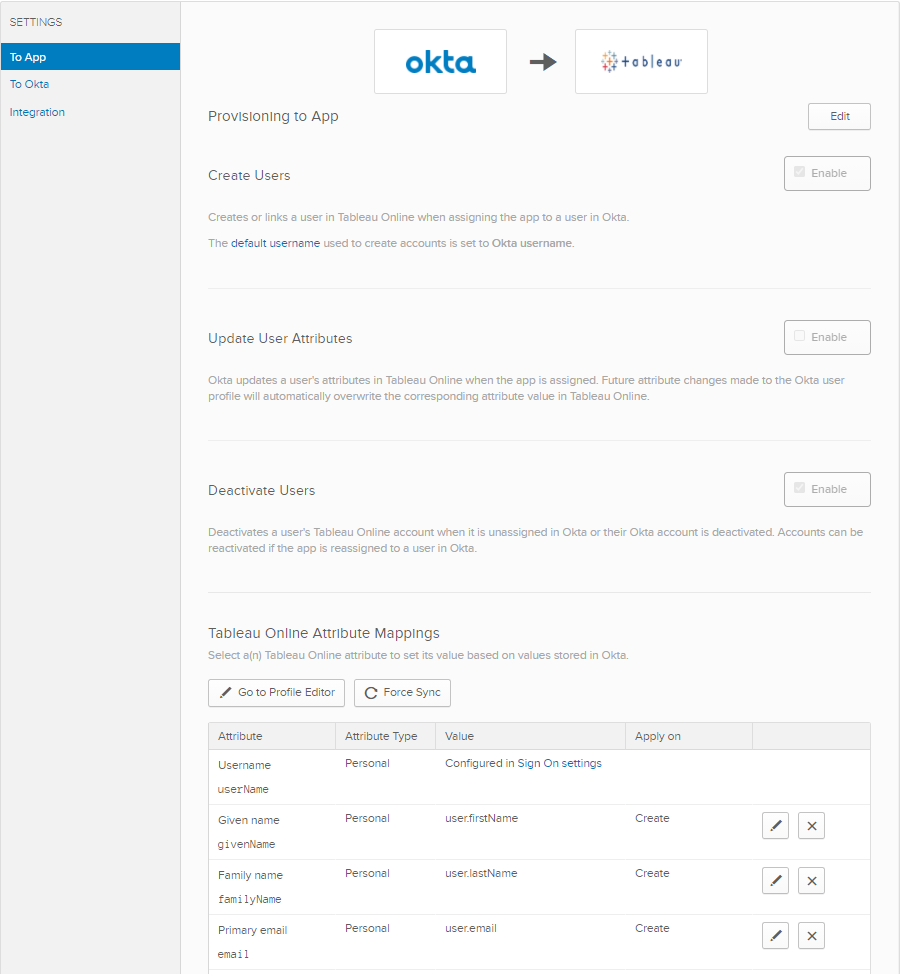


* To get the Tableau Viz available via our application, we have registered for a 14-day trial license of Tableau online where we have published our workbook.
* For user to view our Tableau Viz, they must have an account on our application. They do not require a Tableau license for this. We have implemented SSO to Tableau using Okta as authentication provider. Whenever a user sign-up on our application they get Tableau viewer role assigned to them which allows them to view the visualizations.

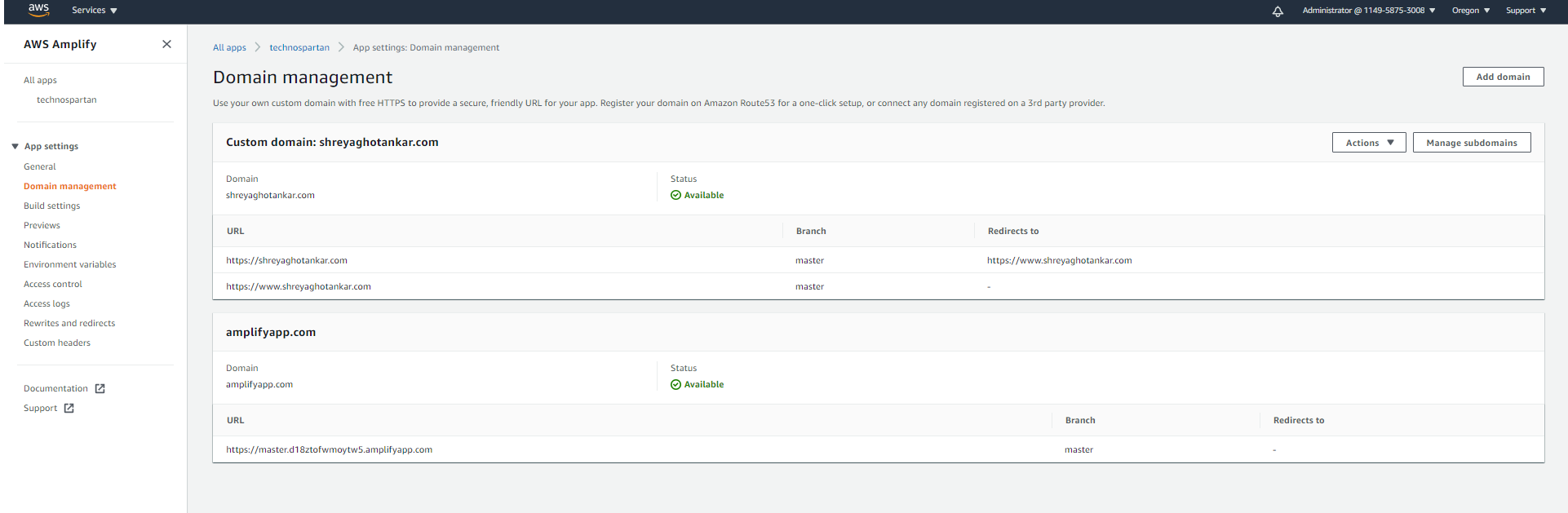
Enabled Okta Authentication for Tableau



* Created a Tableau application in Okta and linked the users that come in through TechnoSpartan application to it.
* Enabled System for Cross-domain Identity Management (SCIM) feature to allow users created in Okta to be created as viewers in Tableau.

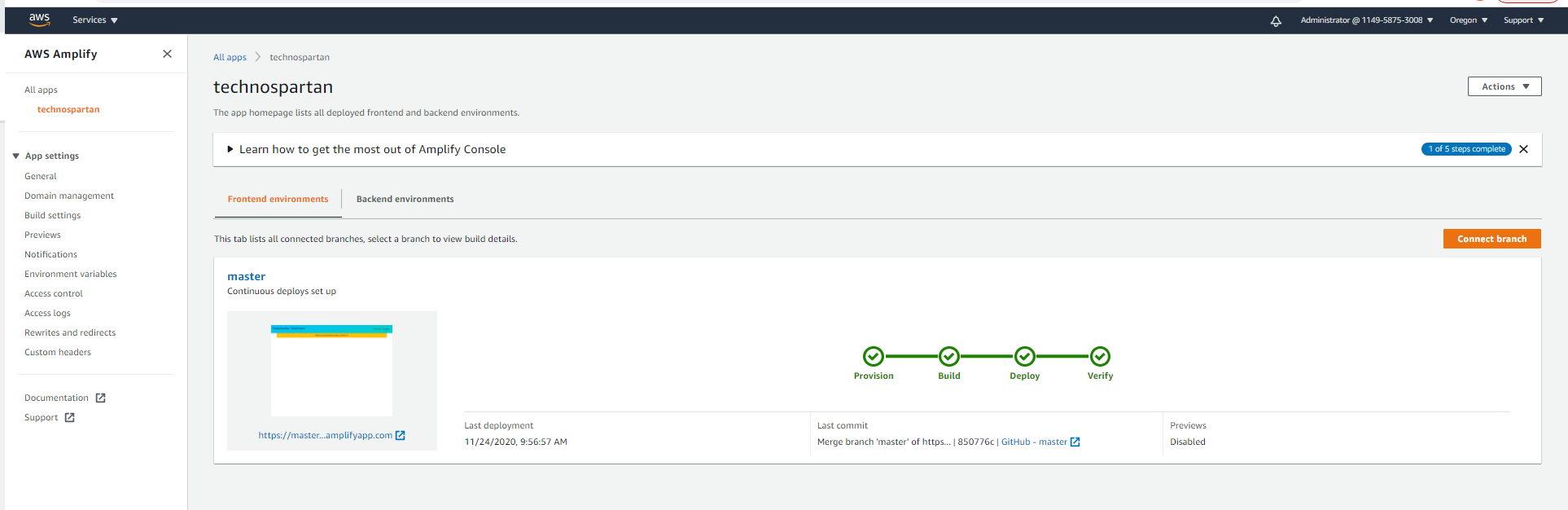


* Route 53 – Registered domain and configured it in AWS Amplify.



# **CI/CD setup**

We are using AWS Amplify as a CI/CD setup. We have connected our project GitHub master repo to AWS Amplify which on every commit or merge pull request event builds and deploy the code.



# **Future and Monetization**

* There is not much social media integration with COVID-19 data.
* Use the API in messaging platforms and alert the users to the COVID-19 rate in the meetup locations.
* Add extensions to personal and professional communication and social media applications such as Slack, Teams, Facebook messenger etc. to alert on the COVID-19 risk in a mentioned location based on analytical data.
* Use NLP to extract the real time tweet locations and intent from the social media text.
* With funding, we could use other AWS technologies and improve the efficiency of big data pipeline.

# **References:**

* <https://aws.amazon.com/blogs/big-data/a-public-data-lake-for-analysis-of-covid-19-data/>
* <https://www.cdc.gov/coronavirus/2019-ncov/downloads/2019-ncov-factsheet.pdf>
* <https://console.aws.amazon.com/s3/buckets/covid19-lake?region=us-east-2&tab=objects>
* <https://aws.amazon.com/blogs/big-data/stream-twitter-data-into-amazon-redshift-using-amazon-msk-and-aws-glue-streaming-etl/>
* <https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/tweet-object>
* <http://docs.tweepy.org/en/v3.4.0/streaming_how_to.html>
* <https://developer.okta.com/code/react/okta_react_sign-in_widget/>
* <https://learning.oreilly.com/library/view/learning-spark-2nd/9781492050032/>