Analyzing COVID-19 Data with AWS Data Exchange, Amazon Athena, and Tableau July-2020 Team-8

Team 8:

S.No	Team Member Name
1	Anusha Pai G
2	Murali Karlapudi
3	Piyush Shukla
4	Puneet Dhiman

Version History:

Version	Date	Author	Key Changes
1.0	20-Nov-2020	Murali Karlapudi	Added Initial Document with problem statement
2.0	26-Nov-2020	Murali Karlapudi	Added Proposed Architecture & Solution Approach
3.0	30-Nov-2020	Anusha Pai G	Submitted changes for synopsis
4.0	25-Dec-2020	Murali Karlapudi	Added Effort & Cost Analysis submitted interim report
5.0	08-Jan-2021	Murali Karlapudi	Added project implementation and screenshots
6.0	14- Jan -2021	Murali Karlapudi	Added Tableau reports
7.0	23-Jan-2021	Murali Karlapudi	Roles& Responsibilities &Lessons learned &observations

Document References:

#	Title	Link
1.	The next evolution of COVID-19 data	https://www.tableau.com/about/blog/2020/5/next-evolution-covid-19-data
2.	Coronavirus (COVID-19) Data Hub	https://aws.amazon.com/marketplace/pp/prodview-a5mqede4xd4c4?qid=1589562921653&sr=0-5&ref_=srh_res_product_title
3.	AWS SDK for Python (Boto3)	https://aws.amazon.com/sdk-for-python/
4.	Using Athena with the JDBC Driver	https://docs.aws.amazon.com/athena/latest/ug/connect-with-jdbc.html
5.	64-bit Java for Windows	https://www.java.com/en/download
6.	Tableau Desktop	https://www.tableau.com/support/releases

Objective / Problem Statement:

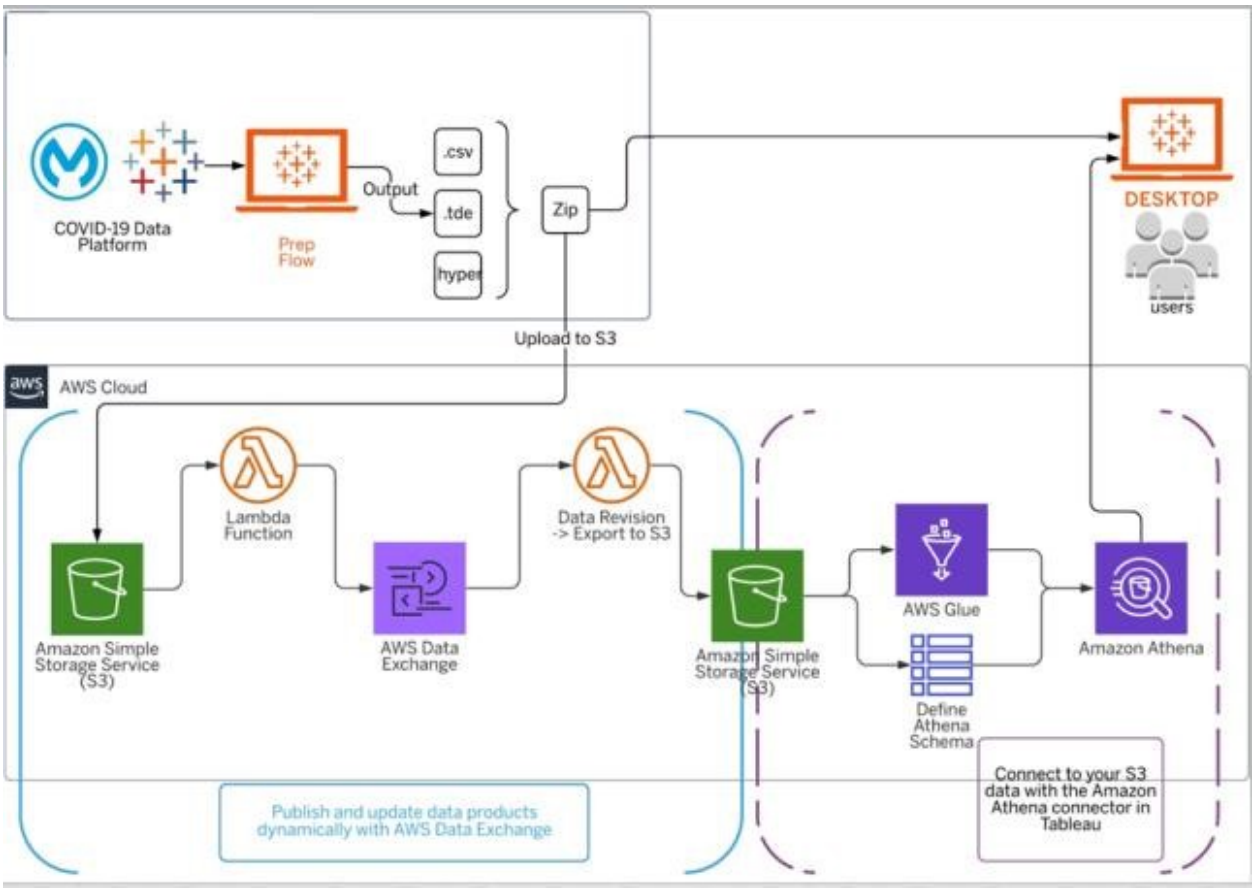
This project creates a centralized repository of trusted data from open source COVID-19 data providers using Amazon Web Services (AWS). Because of the ongoing COVID-19 pandemic our daily lives, people in every sector and country are accessing the data to stay informed and take appropriate measures in terms of keeping themselves safe. Organizations are using COVID-19 data to make critical decisions to continue with their BAU. Data is helping healthcare workers, researchers, scientists, public health officials, and first responders on the front lines as they care for patients, search for therapies, educate the public, influence policy, and communicate action plans.

In order to help everyone, visualize COVID-19 data confidently, Data from the AWS COVID-19 Data Lake is intended to be used for COVID-19 related research and development, and supports analytics in place with the data.

Proposed Solution:

This project aims to offer an accessible way for data professionals to build a dashboard that are updated with new data automatically. This is beneficial to make basic data visualizations without requiring any expertise in front end development. The architecture diagram shows how to ingest data into AWS Data Exchange, and dashboard it is using Tableau and Amazon Athena.

Architecture:



Analyzing and dashboarding using AWS Data Exchange, Amazon Athena, and Tableau.

Data Flow

- The COVID-19 Data Platform ingests and aggregates data across public sources and the COVID Tracking Project. The platform curates standardized data models for the analytical work.
- Data Prep converts the uploaded COVID-19 datasets into csv formats.
- These files are compressed in zip format and are uploaded to an Amazon S3 bucket.
- Amazon S3 triggers an Amazon CloudWatch event which invokes an AWS Lambda function to upload the data into AWS Data Exchange.
- Every time a new revision is published, AWS Data Exchange publishes a CloudWatch event sourced from AWS Data Exchange.

- A CloudWatch event rule triggers a Lambda function that creates an AWS Data Exchange job to export the revision's assets to a predefined S3 bucket.
- Amazon Athena that uses AWS Glue Data Catalog internally to fetch the data from S3.
- Amazon S3 bucket from Tableau Desktop to build dashboards using AWS Athena.

Implementation:

COVID-19 Data Platform

The next evolution of COVID-19 data:

Data is more critical than ever

As we begin emerging from the global COVID-19 crisis, businesses and communities alike are looking for ways to reopen safely. At Tableau we believe data will be a huge part of that journey. However, COVID-19 public data has been highly fragmented across different sources and difficult to make actionable. It's hard to know what data is reliable and most useful to aid the challenge of opening up our economy and returning to normal. Organizations across industries need a faster way to unlock data across applications and systems.

Introducing the COVID-19 Data Platform

Tableau introduced new data tools that will enable business leaders and developers to easily access public health data, diversity data, state and local policy guidance, mobility data, and much more. With these resources, your organization can better respond to time-sensitive, business-critical initiatives. Additionally, the COVID-19 Data Platform will empower IT teams to move faster by leveraging pre-built APIs and developer resources instead of manually building those from scratch.

How it works

The COVID-19 Data Platform ingests and aggregates data across public sources, including The New York Times, European Centre for Disease Prevention and Control, and the COVID Tracking Project. The platform then curates them into standardized data models that can be reliably used to make data-driven decisions, whether through visualizations or automated processes. With more data inputs, we enrich the data to include more information for deeper analysis. With this unification of data, we're making this data more reliable and trustworthy than ever before.

Utilizing technologies in Salesforce, MuleSoft, and Tableau, this crucial public data is made available as open APIs and within Tableau Public to be used in applications and new visualizations by anyone. The COVID-19 Data Platform is an open data service for developers and the Salesforce ecosystem. We now use it to power the Tableau COVID-19 Data Hub and the Work.com Command Center. By partnering with Akamai, we're able to serve API requests quickly and reliably anywhere in the world through their global platform. Organizations can also access the data through the **AWS Data Exchange**, further extending the value and collaboration to third-party developers.

Additionally, Tableau automated many of the data preparation processes using Tableau Prep, which allows us to publish more accurately and routinely every day. This means the data is in the best shape for analysis, and always up to date and ready for your important decisions.

How to access the data

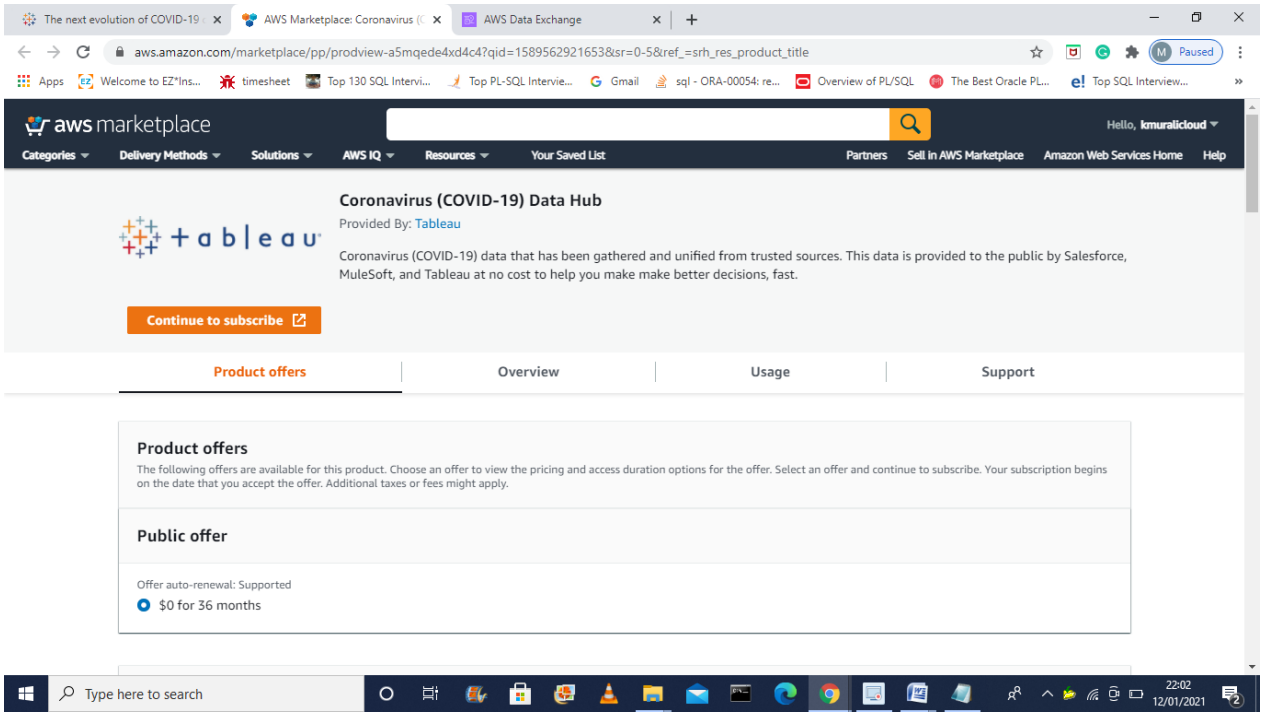
AWS Data Exchange:

The AWS Data Exchanges makes it easy to find, subscribe to, and use third-party data in the cloud. You can subscribe to COVID-19 activity data and more for free at the Coronavirus (COVID-19) Data Hub listing in the AWS Data Exchange.

The AWS Data Exchanges makes it easy to find, subscribe to, and use third-party data in the cloud. You can subscribe to COVID-19 activity data and more for free at the Coronavirus (COVID-19) Data Hub listing in the AWS Data Exchange.

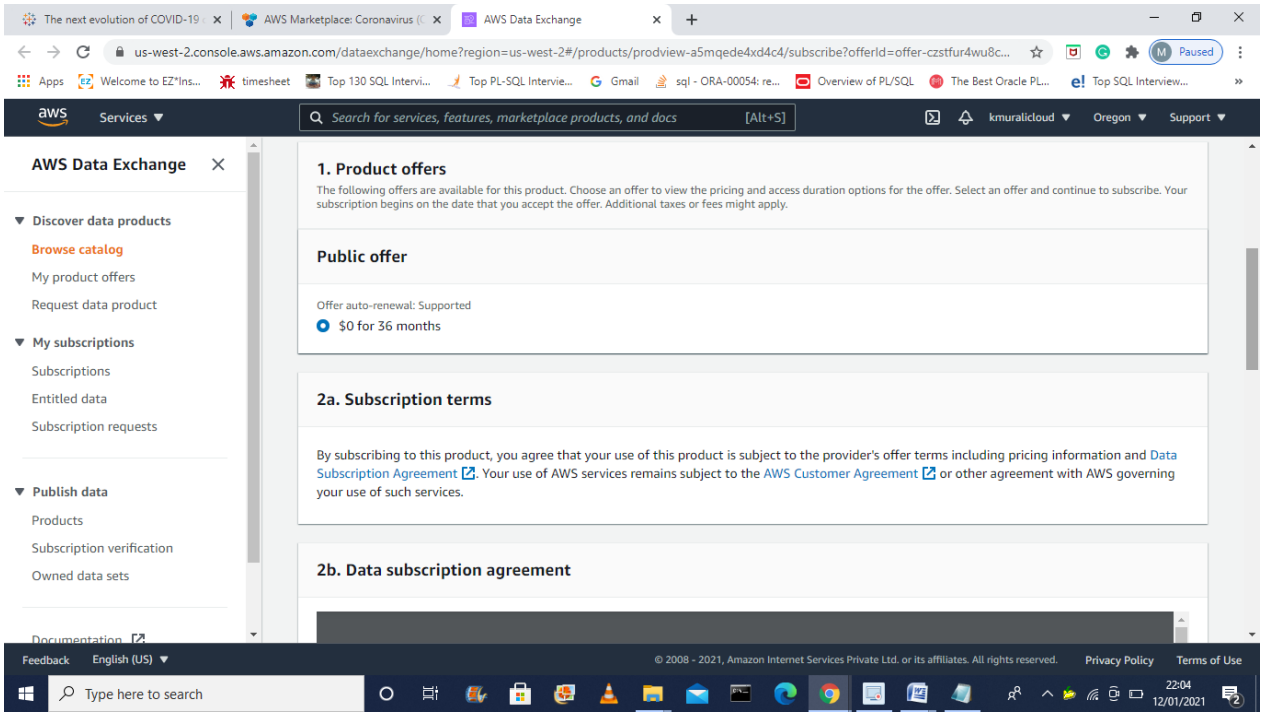
Subscribe the Coronavirus (COVID-19) Data Hub by cling the below link:
https://aws.amazon.com/marketplace/pp/prodview-a5mqede4xd4c4?qid=1589562921653&sr=0-5&ref_=srh_res_product_title

Click on Continue to Subscribe

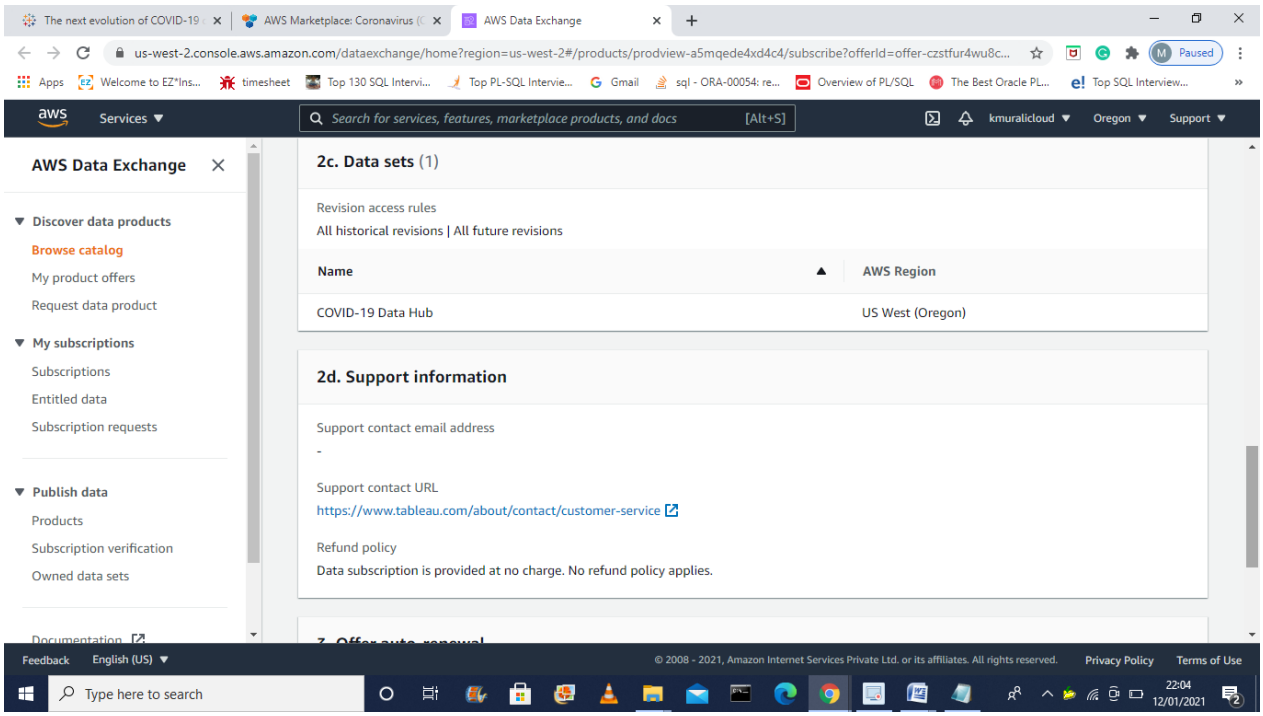


We can also browse catalog as mentioned below from the AWS market place

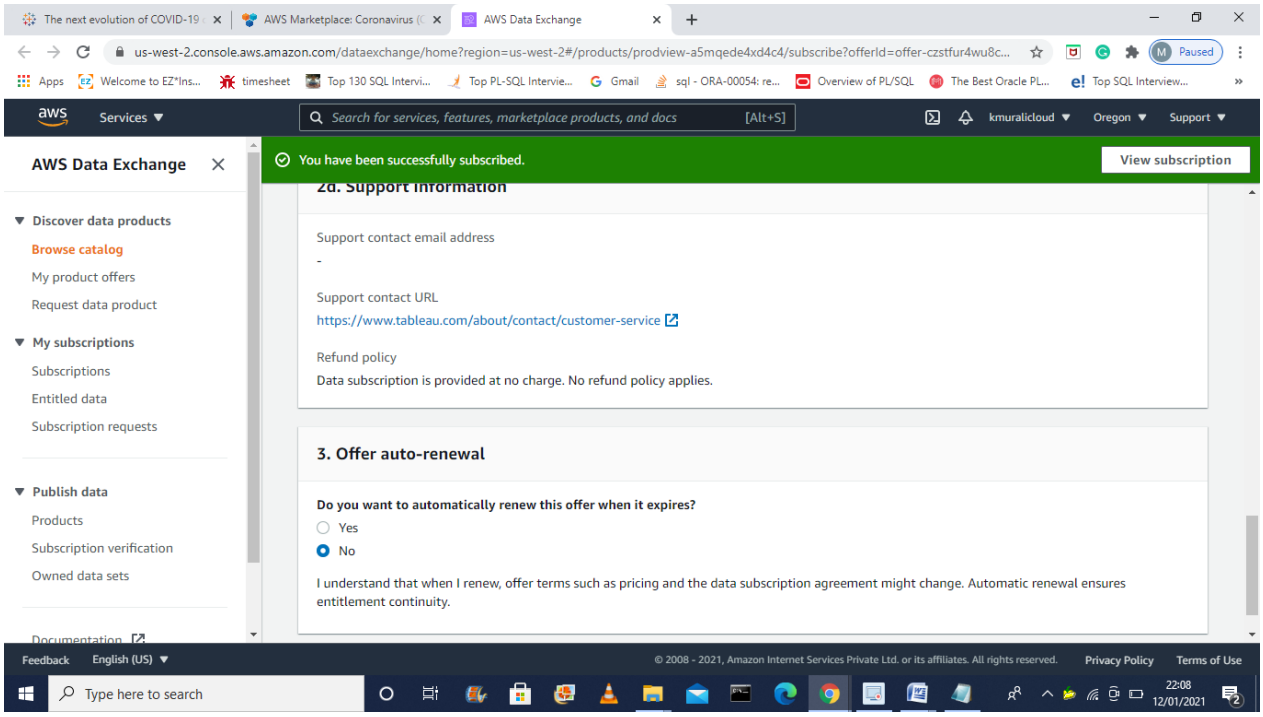
The cost of Subscription is $0 for 36 months



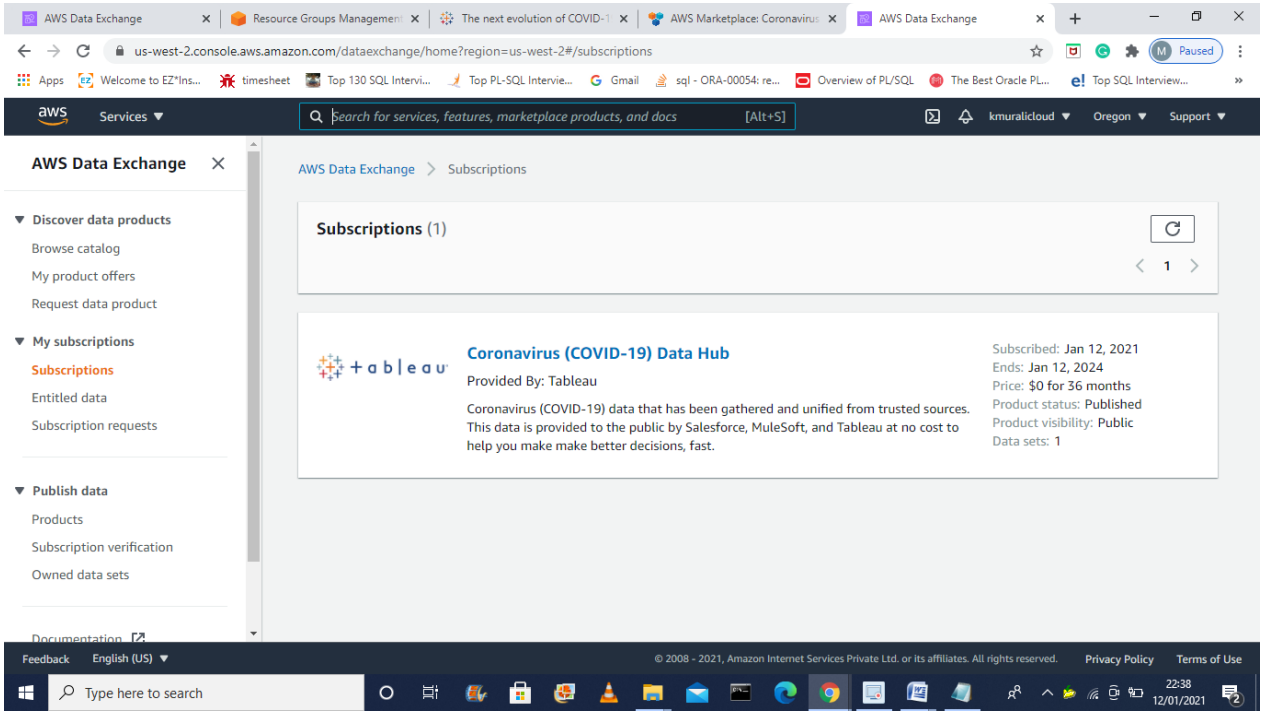
DataSets → Covid-19 Data Hub (US West Oregon) Region



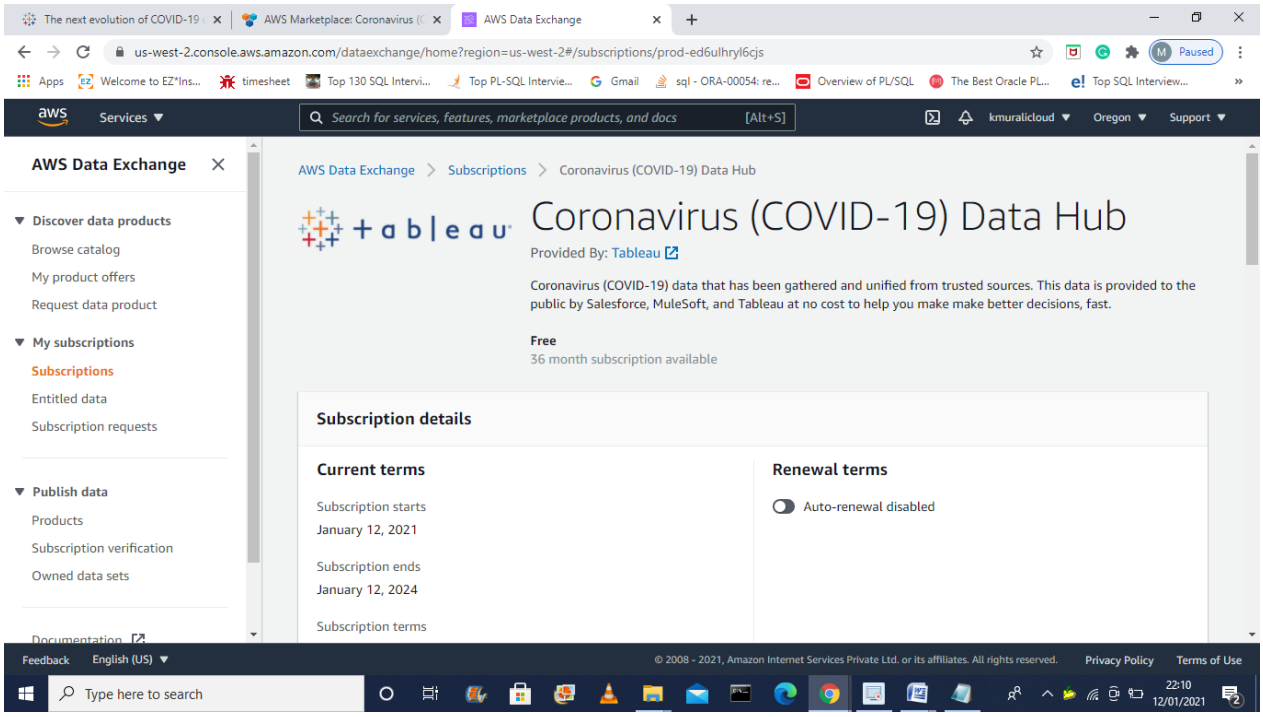
Data set Subscribed Successfully



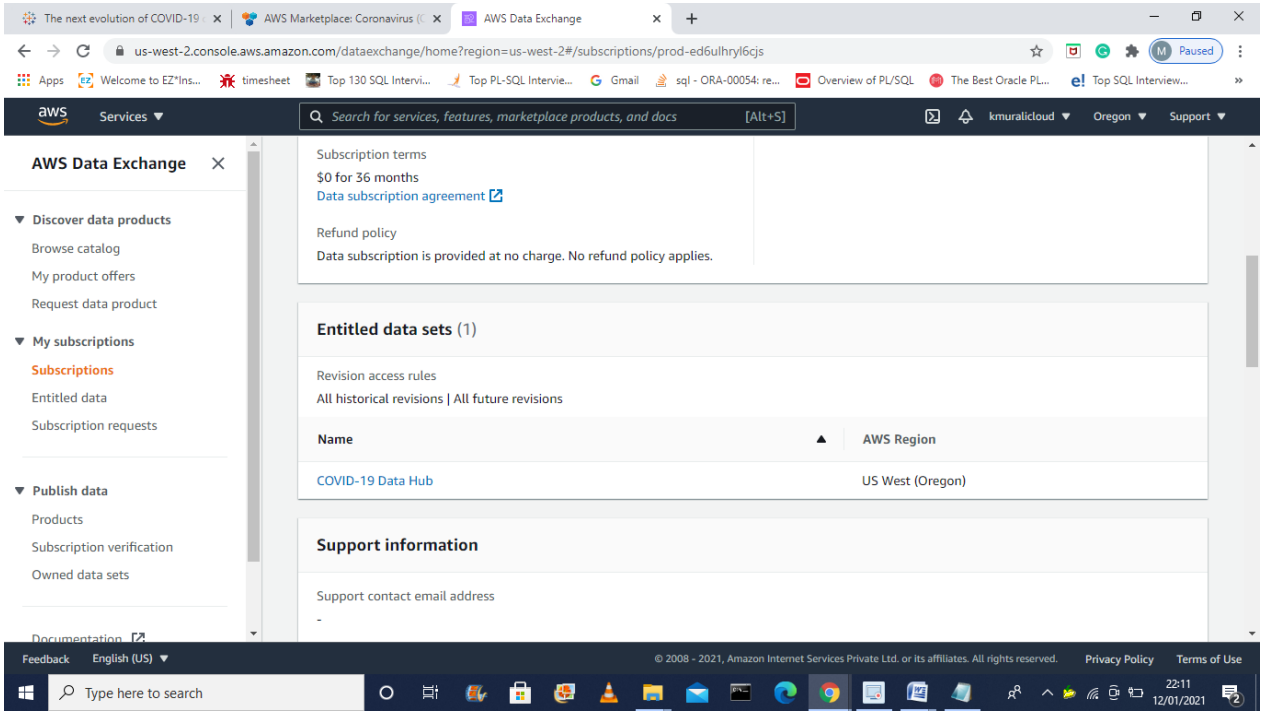
Go to My Subscriptions section:



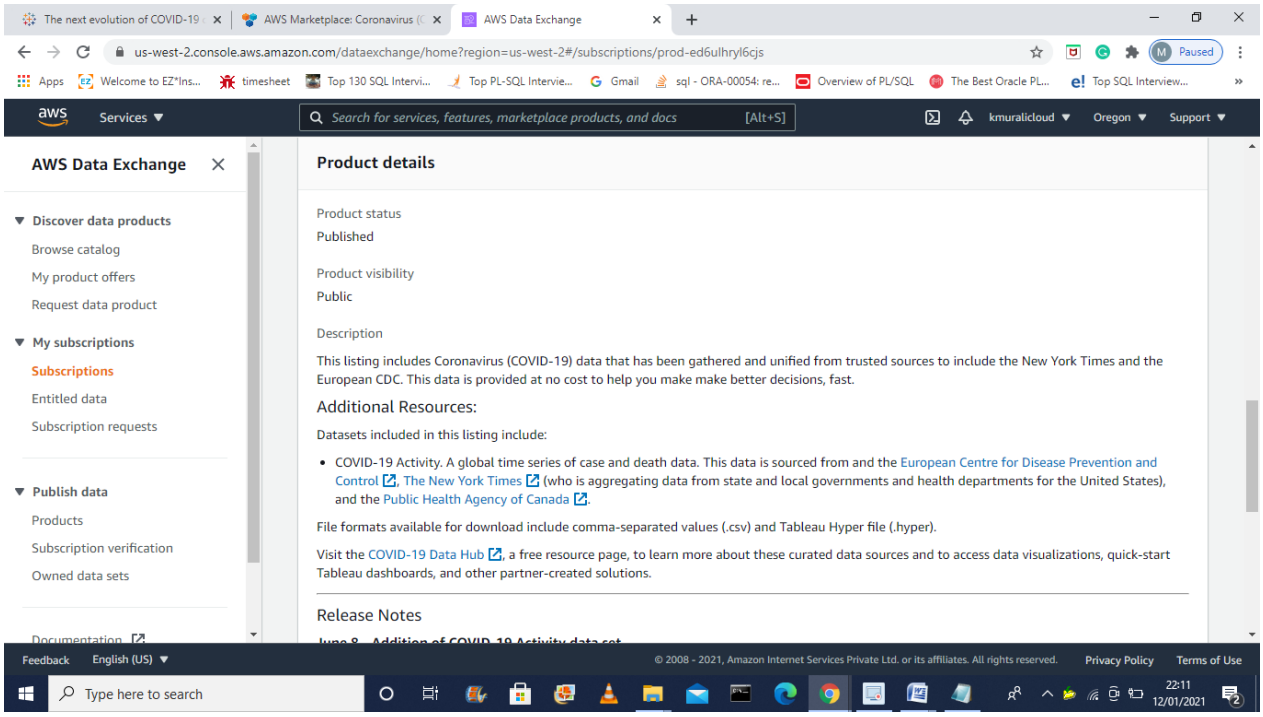
Subscription Details:



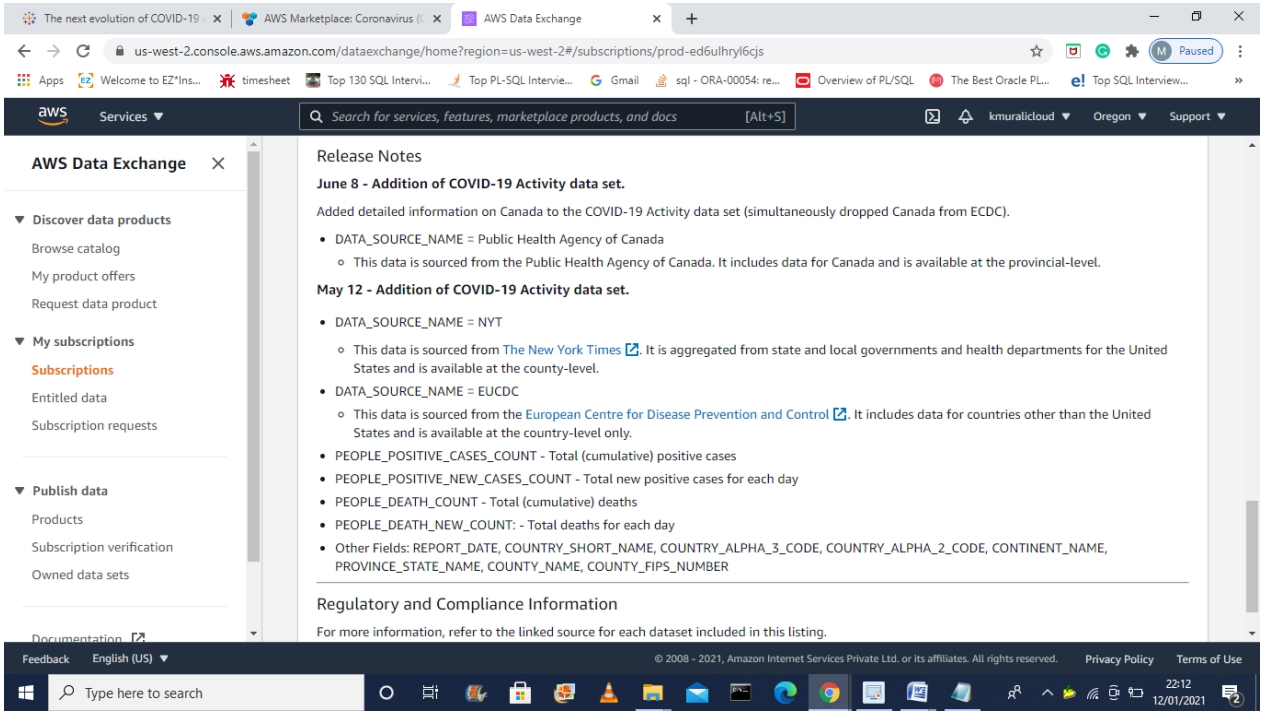
Subscription dataset:



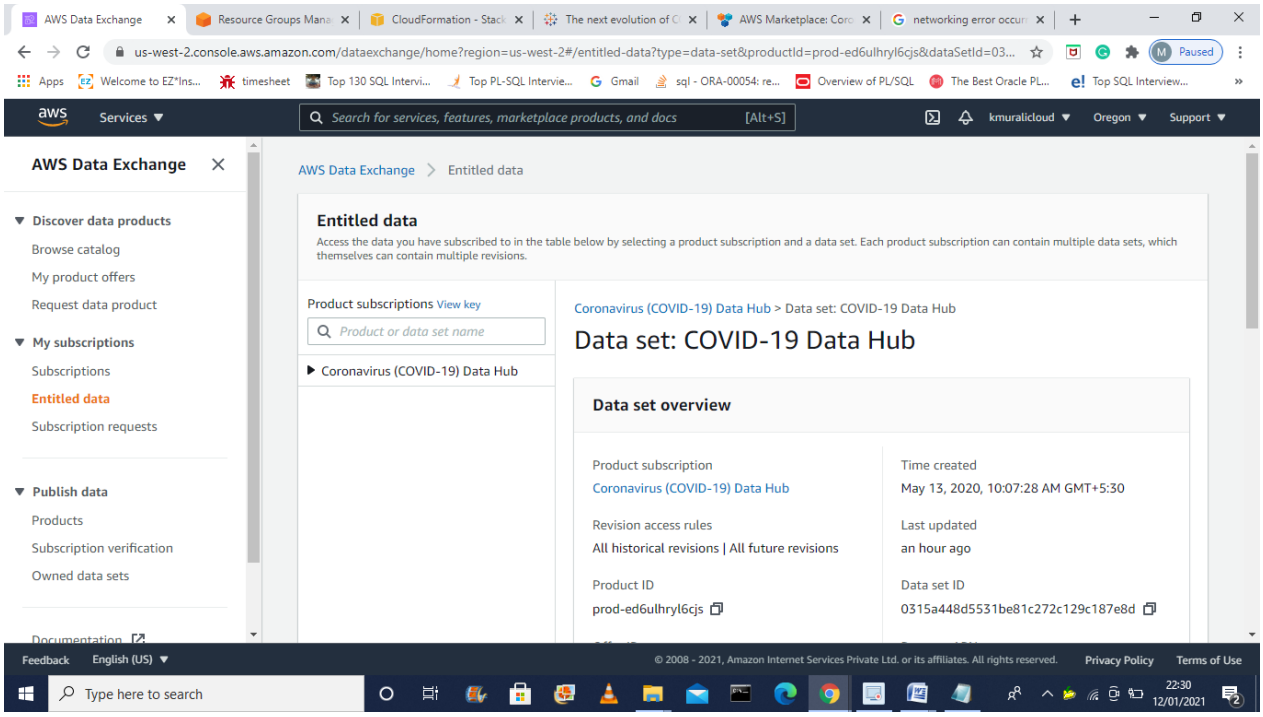
Subscribed product details:



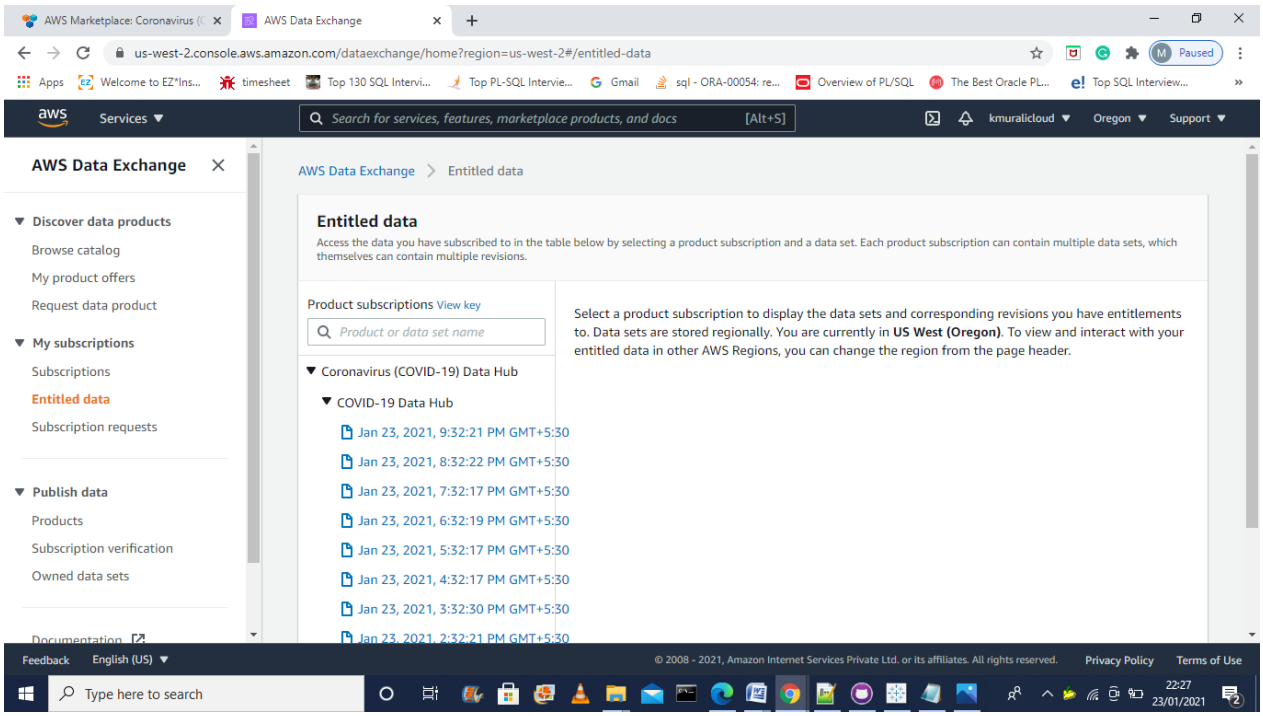
Data set fields description



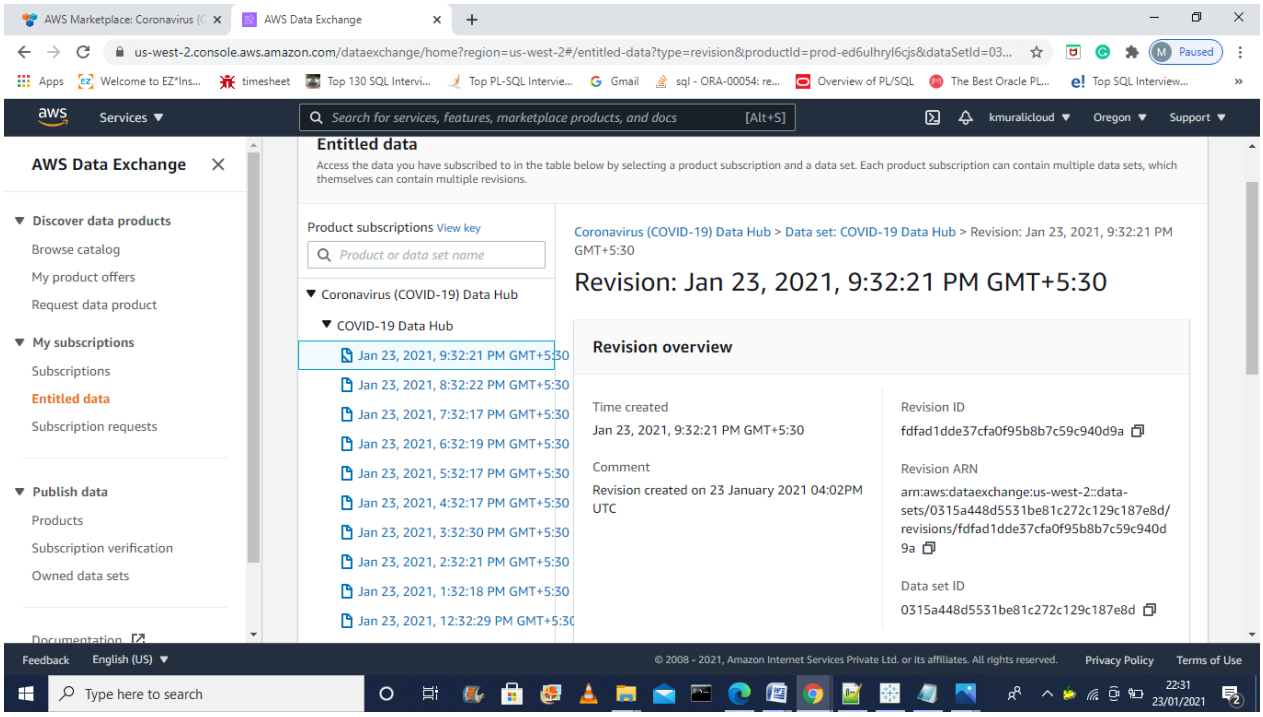
Entitled data



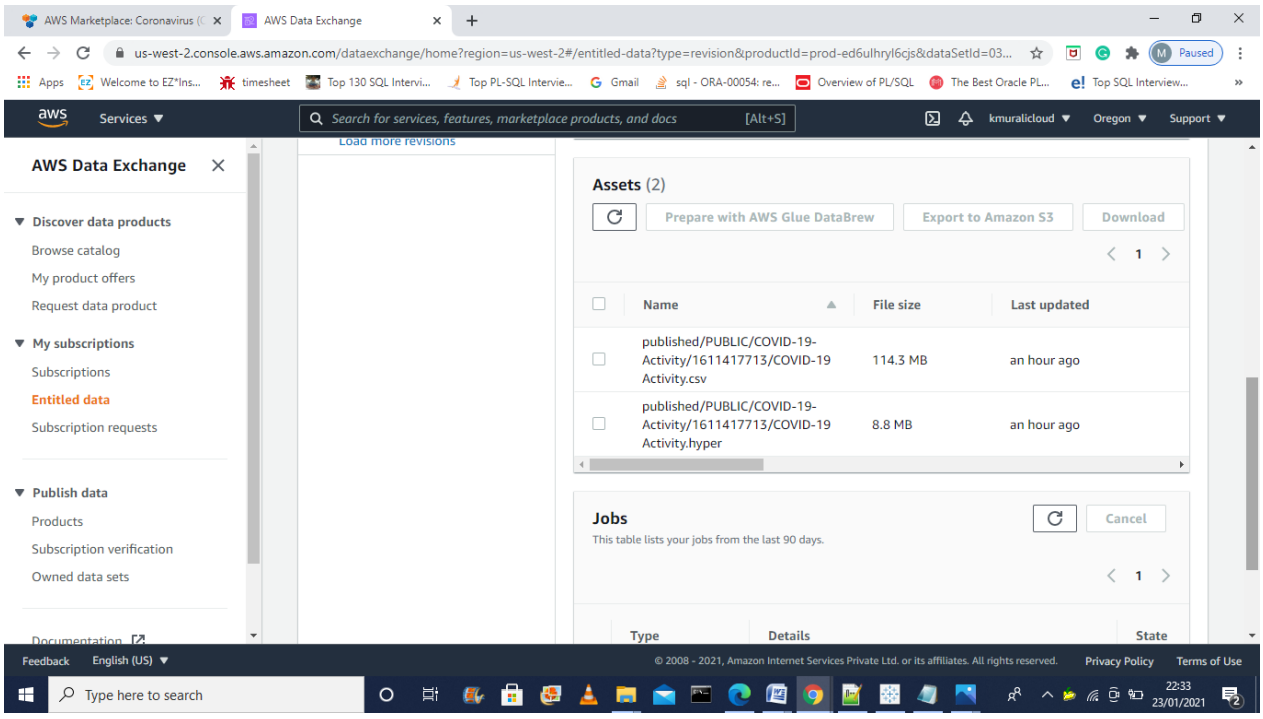
Entitled data , click on Latest Revision



Latest Revision Overview



File formats available for download include comma-separated values (.csv) and Tableau Hyper file (.hyper)



Cloud formation stack described in the next steps will create destination S3 bucket to store this csv/hyper file.

Find and acquire new data sets and retrieve new updates automatically using AWS Data Exchange

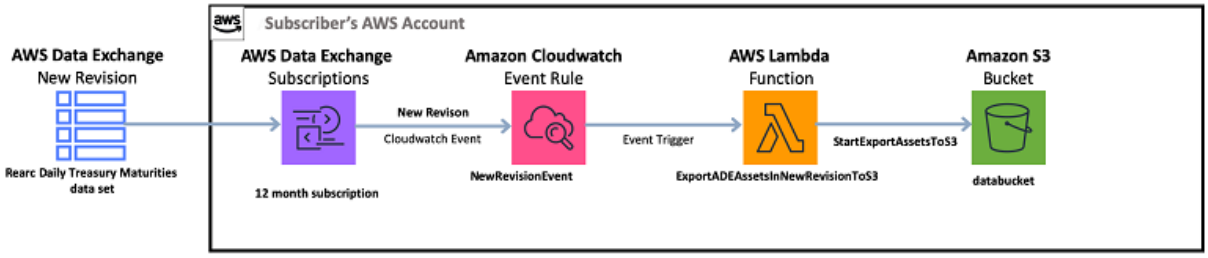
The solution has three steps:

1. Configure our prerequisites: an S3 bucket for our data and IAM permissions for using AWS Data Exchange.
2. Subscribe to a new data product in AWS Data Exchange, Subscribing to **Coronavirus (COVID-19) Data Hub**
3. Set up an automation using Amazon CloudWatch events to retrieve new revisions of subscribed data products in AWS Data Exchange automatically.

Automating the retrieval for new data set revisions

Providers update many products regularly by creating and publishing new revisions to the underlying data sets. For example, the **Coronavirus (COVID-19) Data Hub** data product is updated hourly. We want your analytics and visualizations to add these revisions to their insights easily. To do so, We need to set up an automation to retrieve the new files stored in newly published revisions.

The following diagram shows the workflow of this process.



Every time a new revision is published, AWS Data Exchange publishes a CloudWatch event sourced from aws.dataexchange. Using a Cloudwatch event rule to trigger a Lambda function, an AWS Data Exchange Job exports the revision's assets to a pre-defined S3 bucket. It is interesting to note that because AWS Data Exchange uses the asset name as a default S3 object key when exporting to Amazon S3, and since Covid-19 is publishing a new revision with the

same asset name every hour, this automation will always override the previous day's file with a new file, allowing you to always refer to the same S3 object, which will have the latest data.

An AWS CloudFormation template packages this automation. It contains all the necessary resources, including an S3 bucket to store the data, the Lambda function to export the data, its IAM role and policy, and the CloudWatch event rule to trigger the function. Packaging this automation in an AWS CloudFormation template makes it simple to repeat the automation for each data set you subscribe to. You can configure the template using the Data Set ID, which you can retrieve from the data set page that we have seen above.

we use a Lambda layer that extends the AWS Python SDK (boto3) that is built into the Lambda Python runtime by adding the AWS Data Exchange and AWS Marketplace Catalog API SDKs as of November 13, 2019.

Deploying the automation

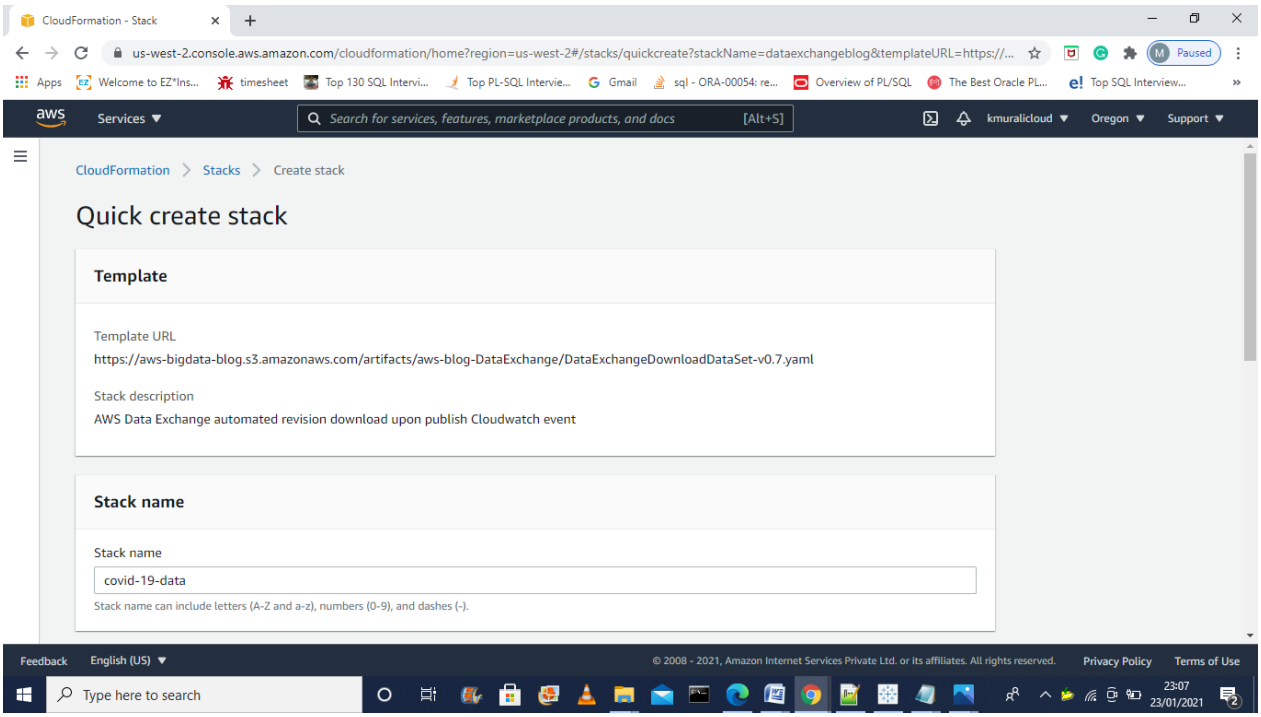
Before deploying the automation, make sure we are in the Region in which the data set is located. You can find this on the Subscription details page under Data sets. In our scenario we are under US West Oregon.

On the AWS CloudFormation console, choose Create Stack.

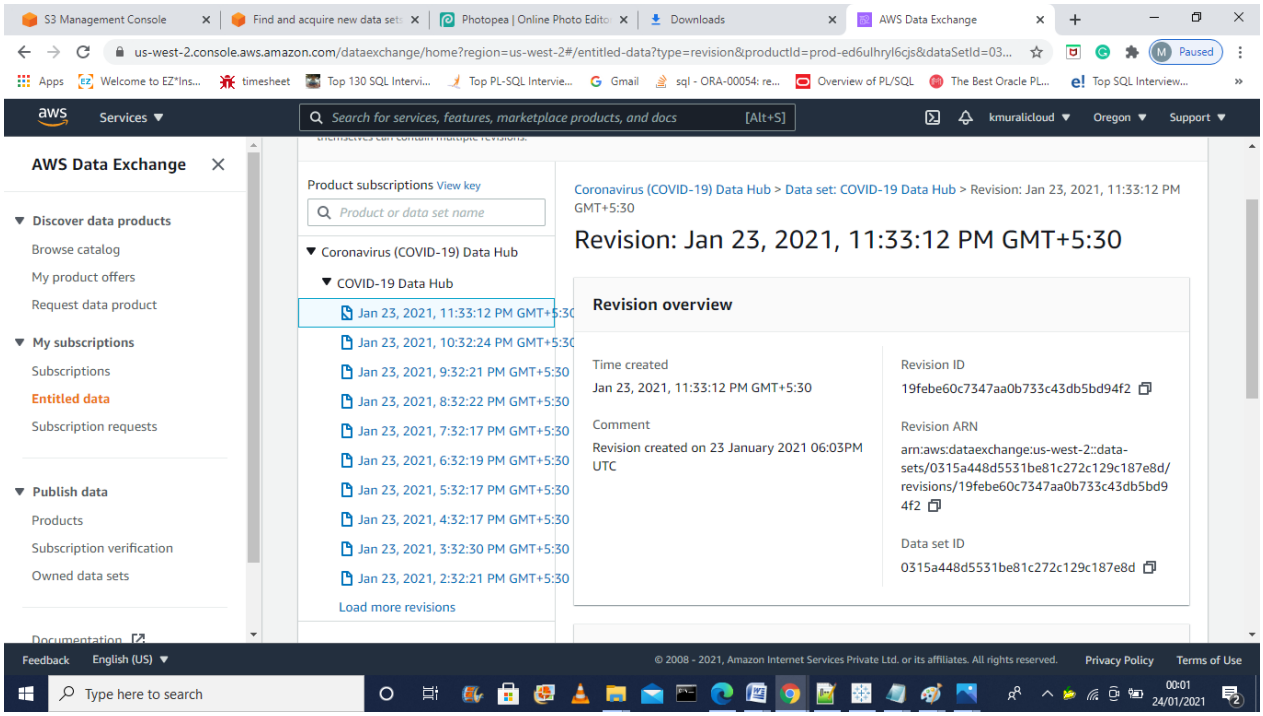
Template URL

https://aws-bigdata-blog.s3.amazonaws.com/artifacts/aws-blog-DataExchange/DataExchangeDownloadDataSet-v0.7.yaml

Provide the Stack name → covid-19-data



Provide the Data set ID



Data set ID

0315a448d5531be81c272c129c187e8d

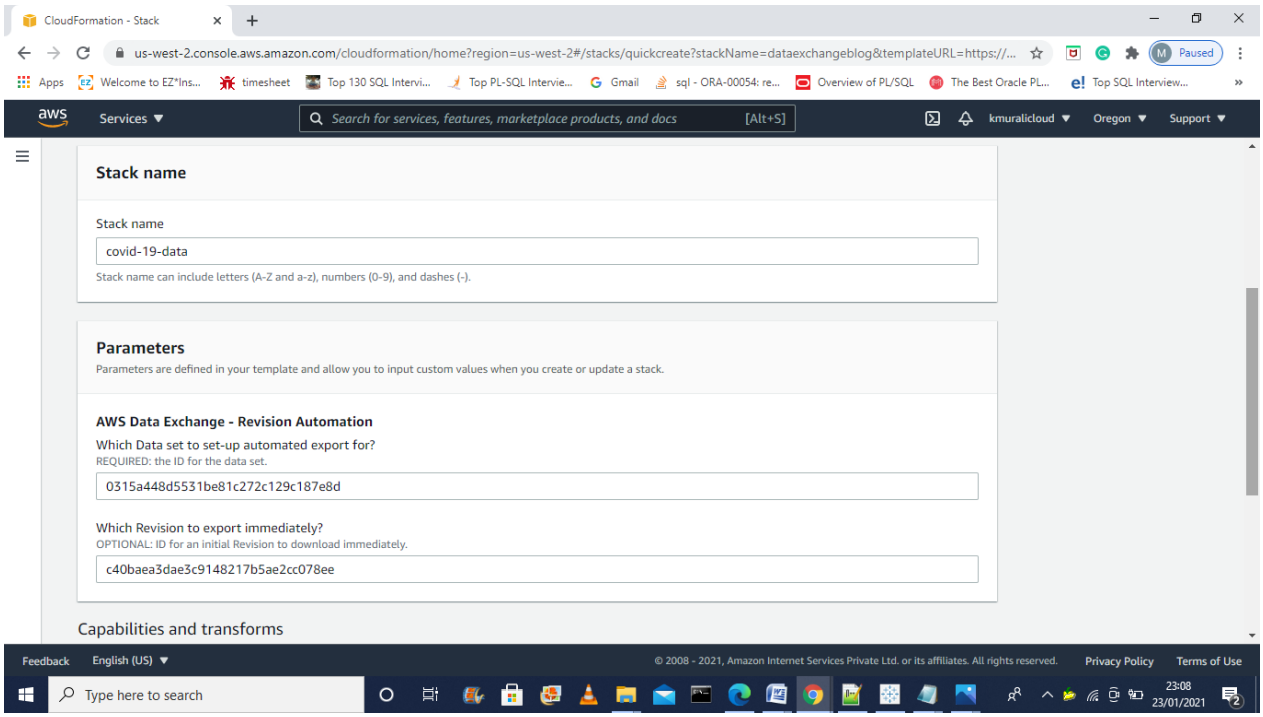
Copy the data set ID as shown in the above figure

Revision ID

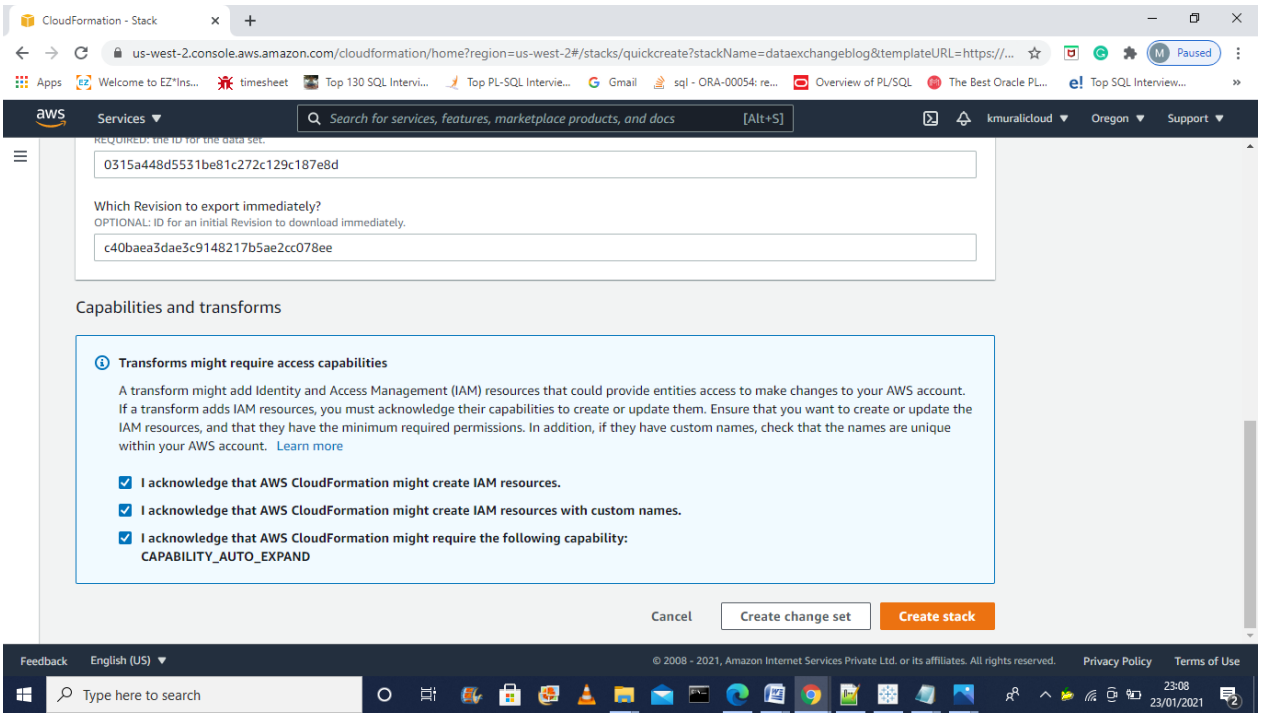
19febe60c7347aa0b733c43db5bd94f2

Copy the Revision ID from the above figure

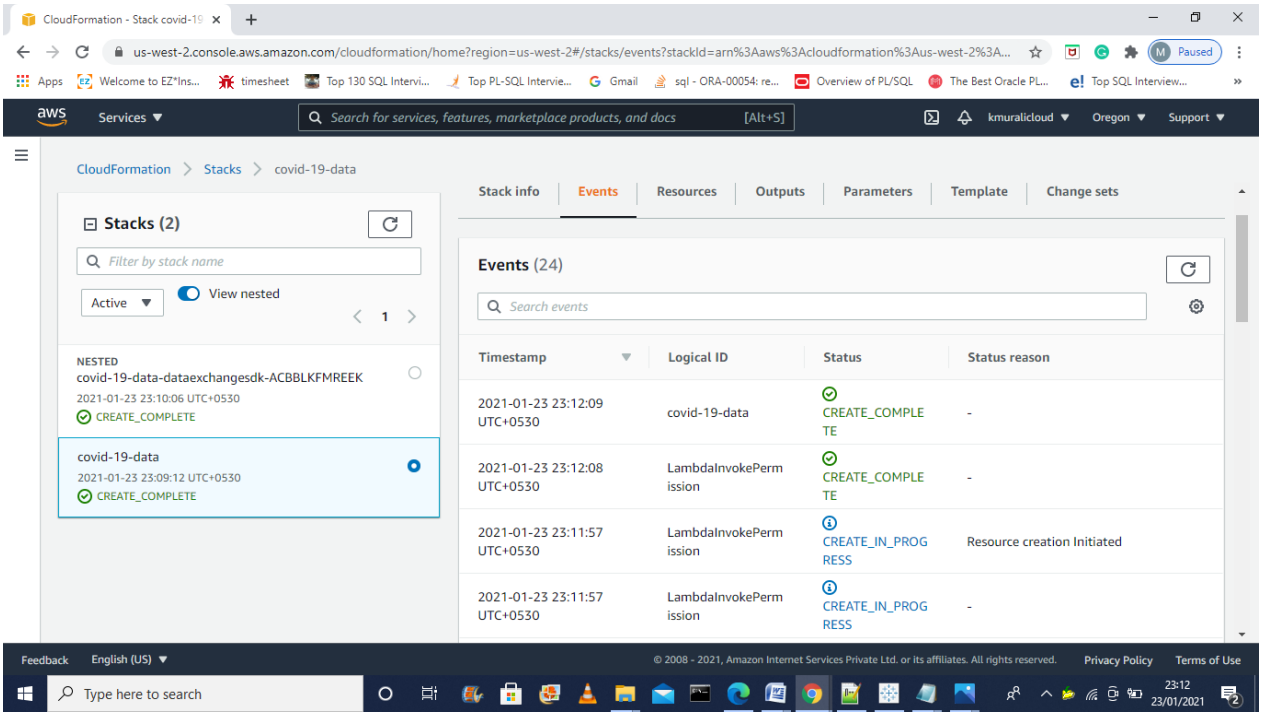
Enter both Data set ID, Revision ID in the parameters section as show below



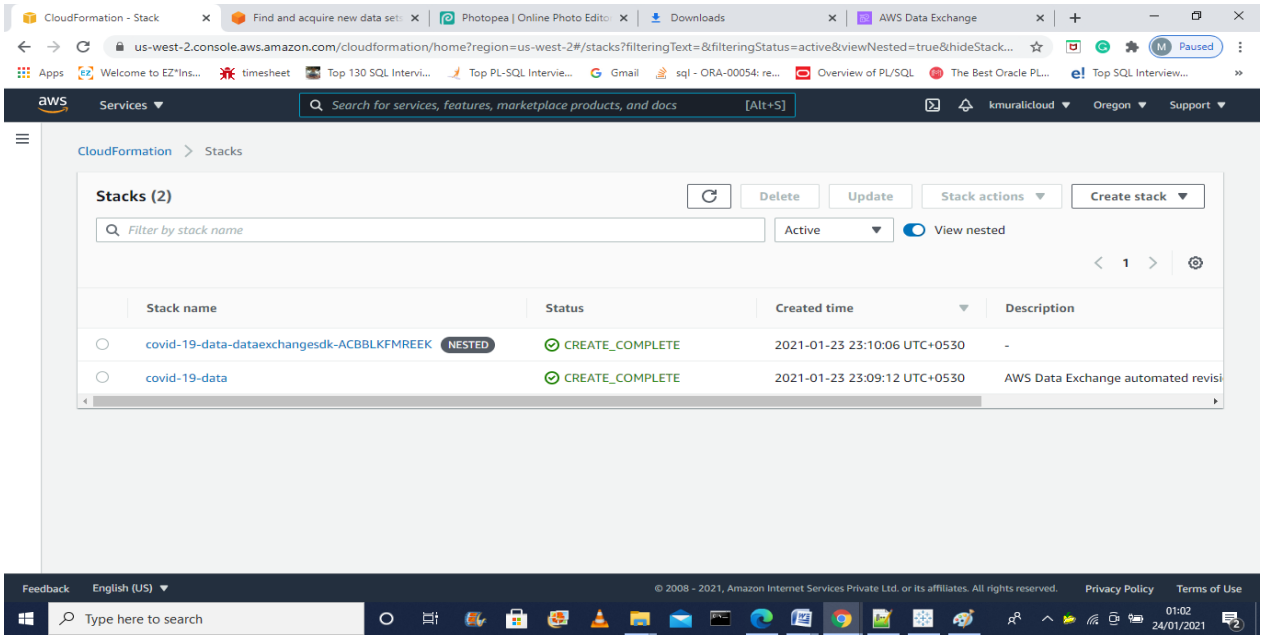
Create Stack



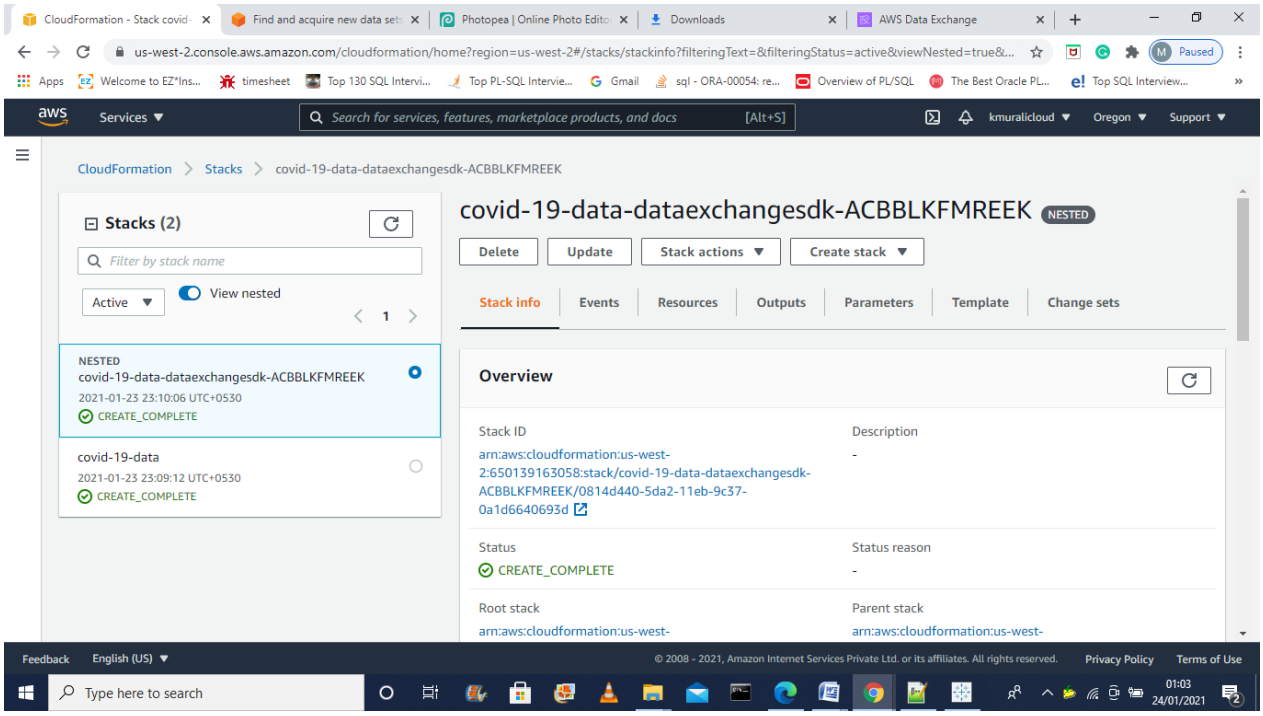
Stack created successfully



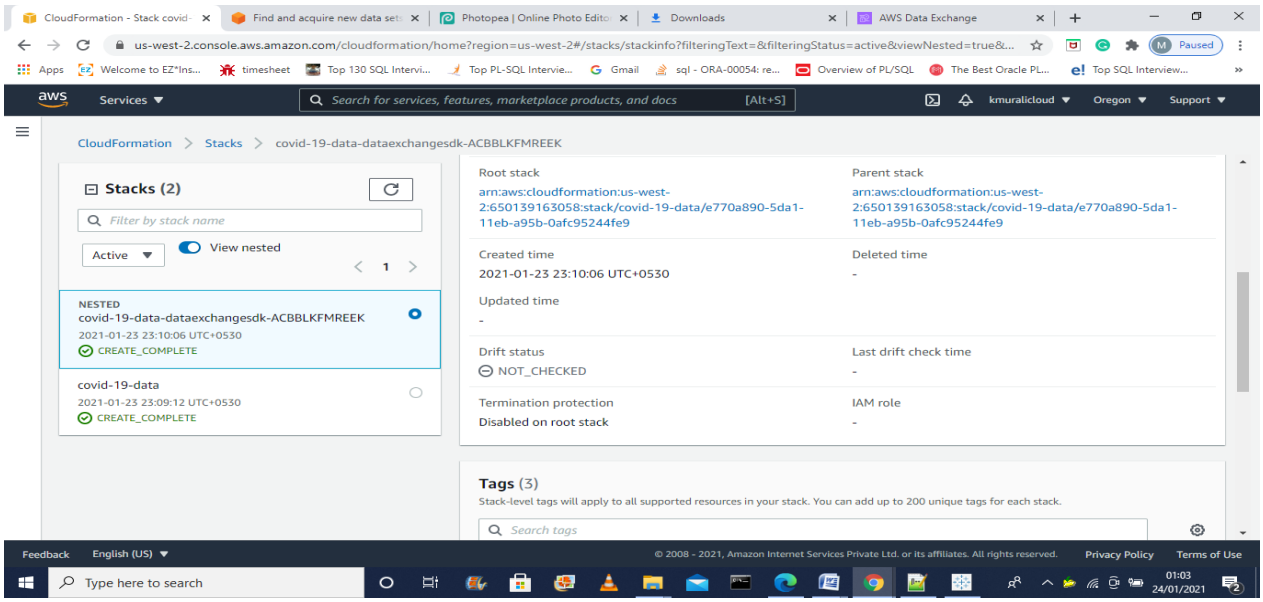
Created Stacks details in below few set of slides: covid-19-data, covid-19-data-dataexchangesdk-ACBBLKFMREEK as shown below



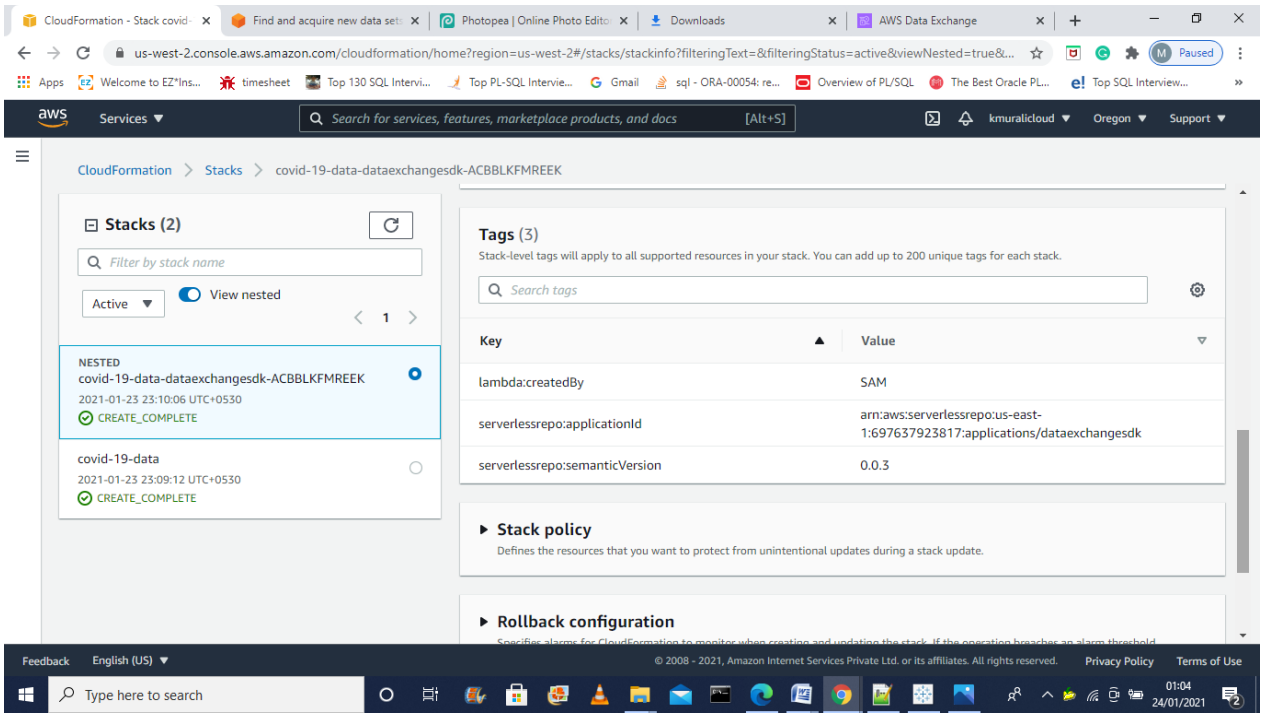
Covid-19-data-dataexhangesdk details



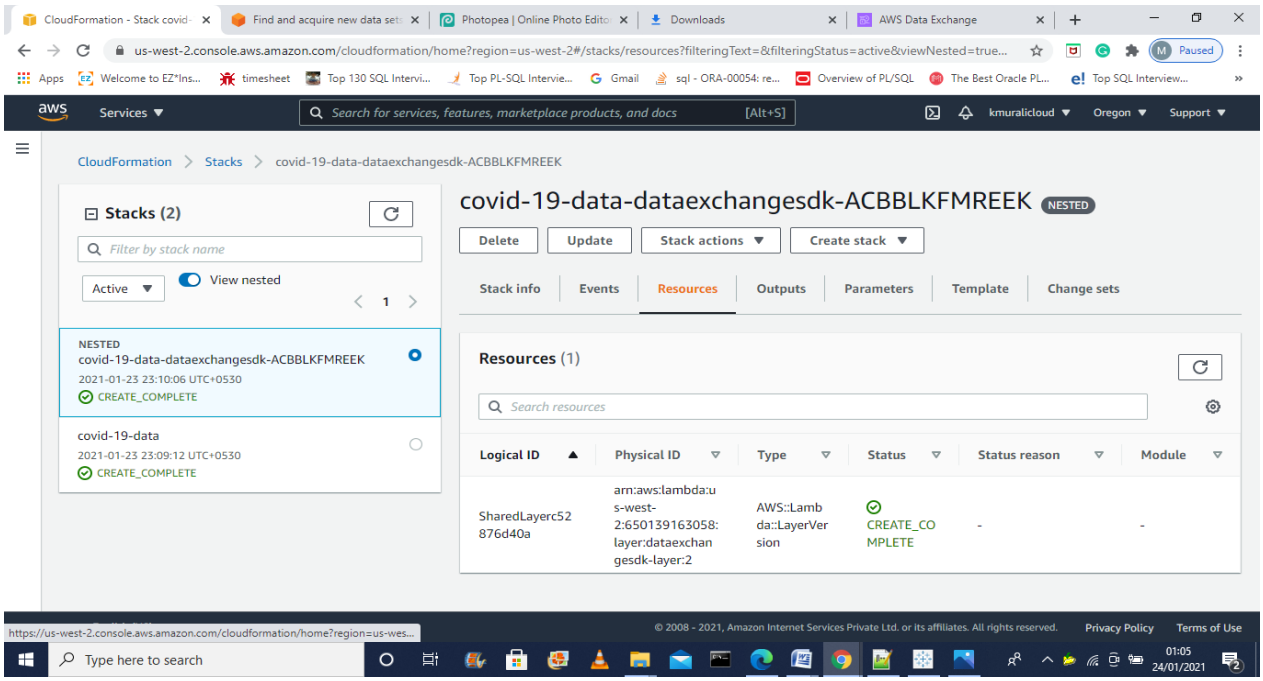
Covid-19-data-dataexhangesdk details



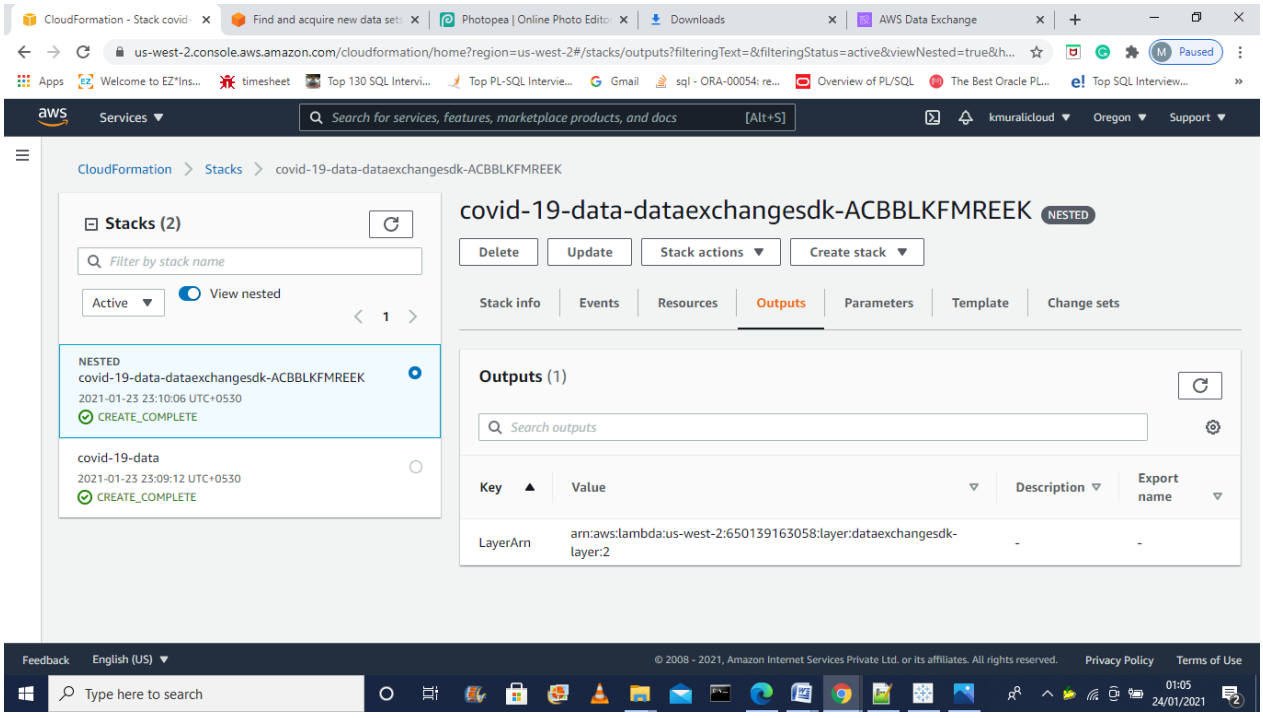
Covid-19-data-dataexhangesdk Tags



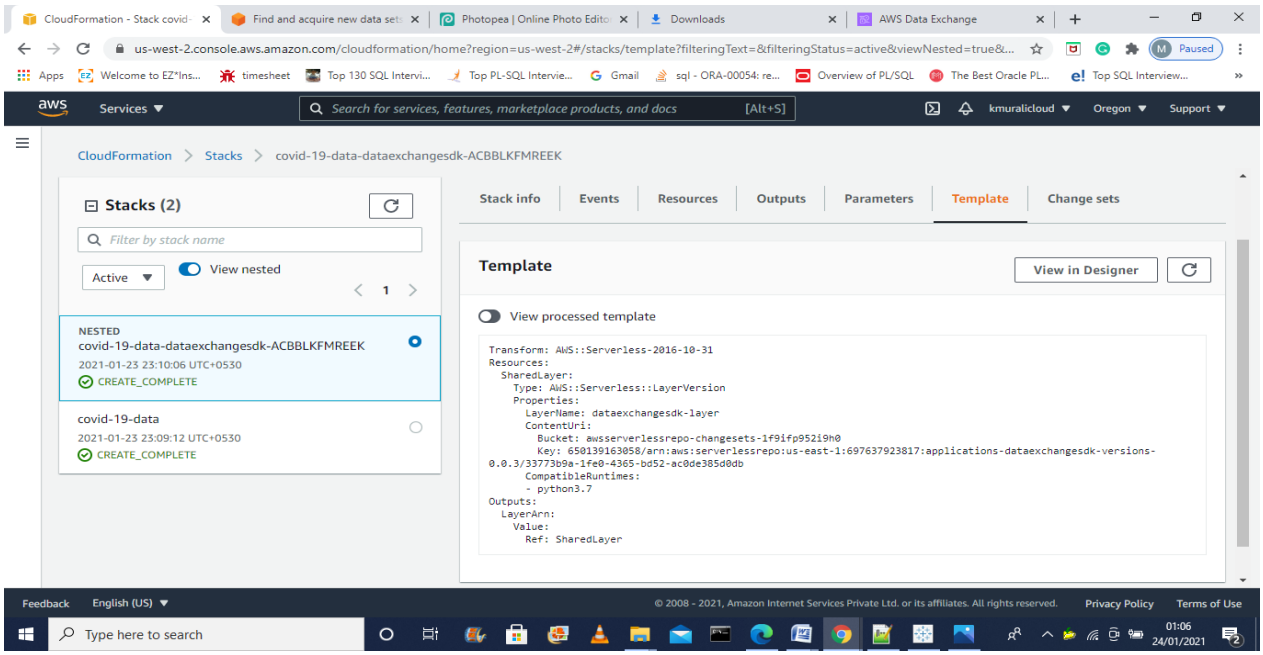
Resources of the covid-19-data-dataexchangesdk



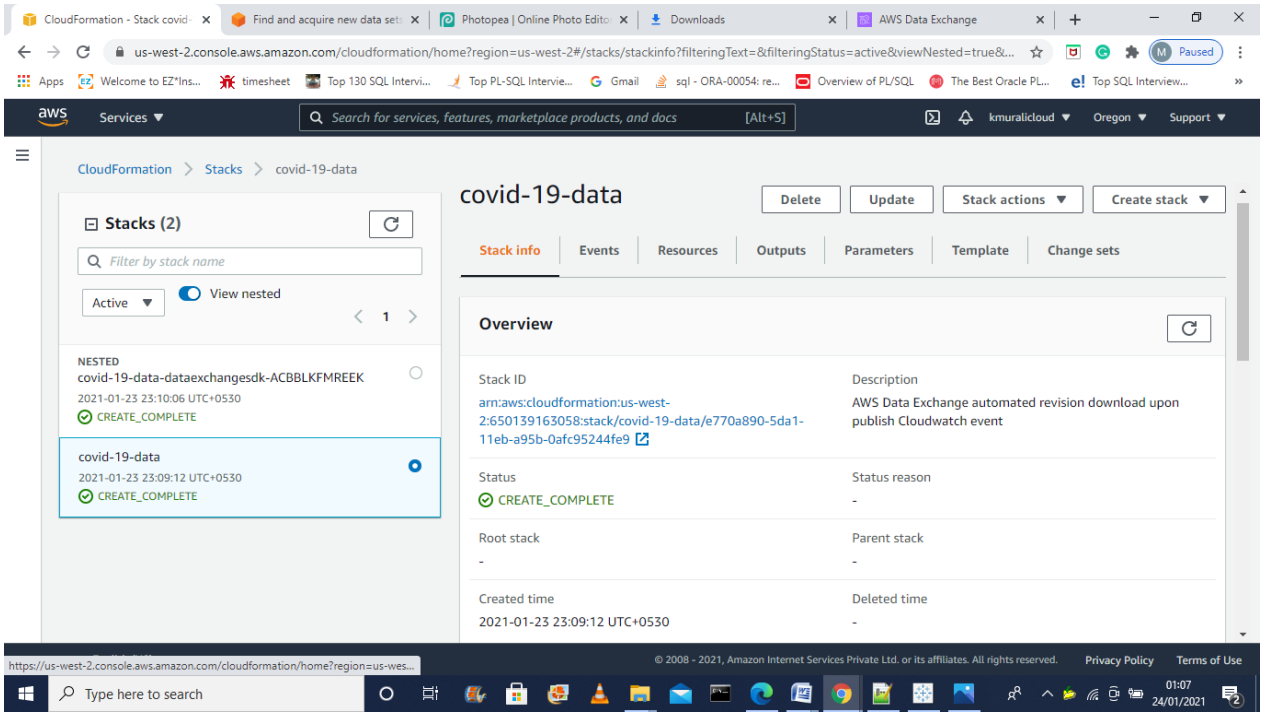
Outputs of the stack



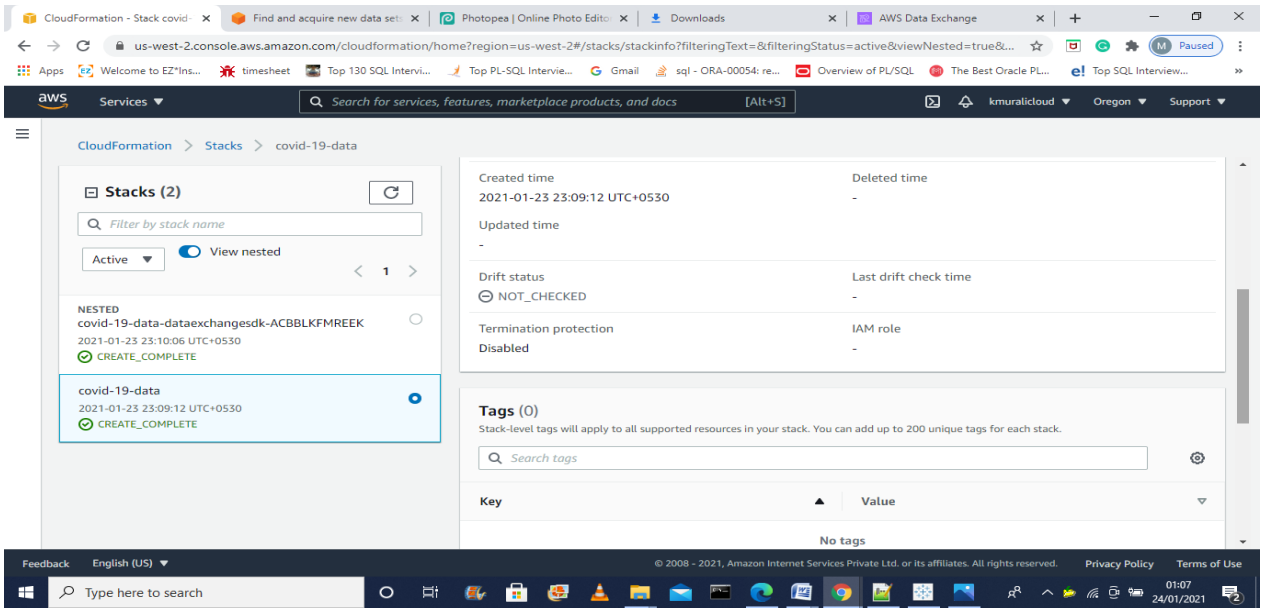
template of the covid-19-data-dataexchangesdk



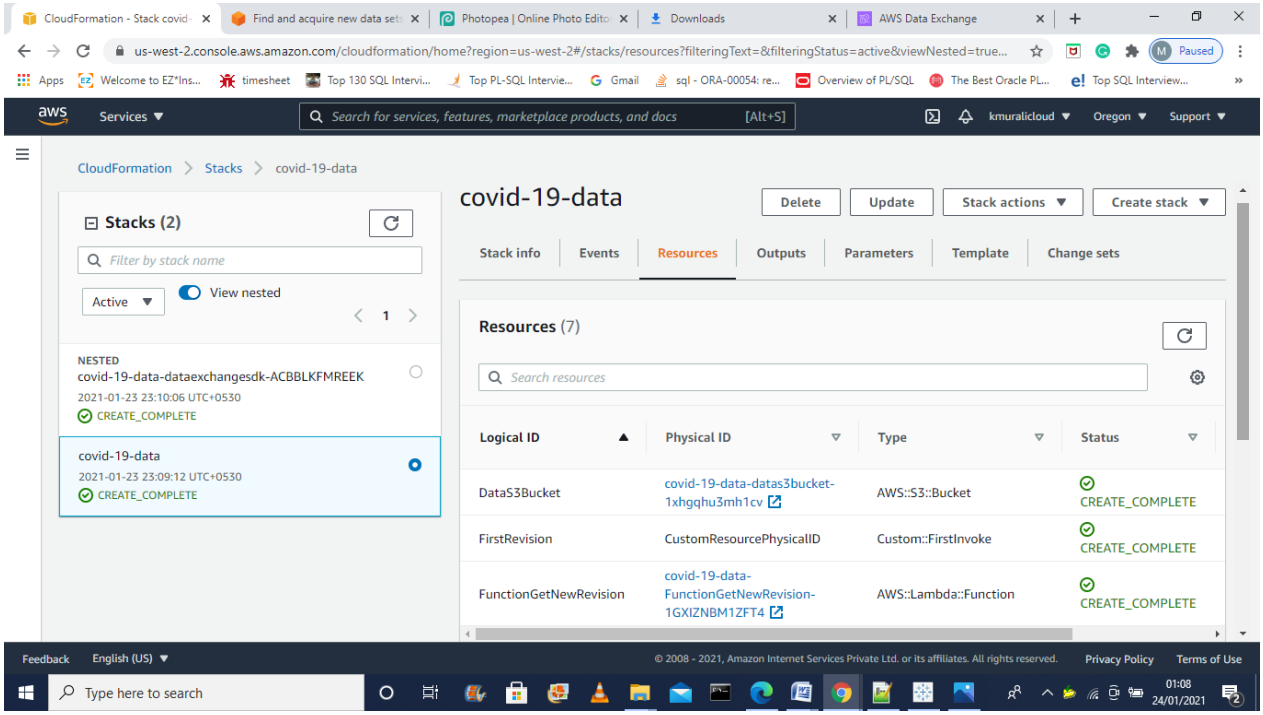
Covid-19-data summary



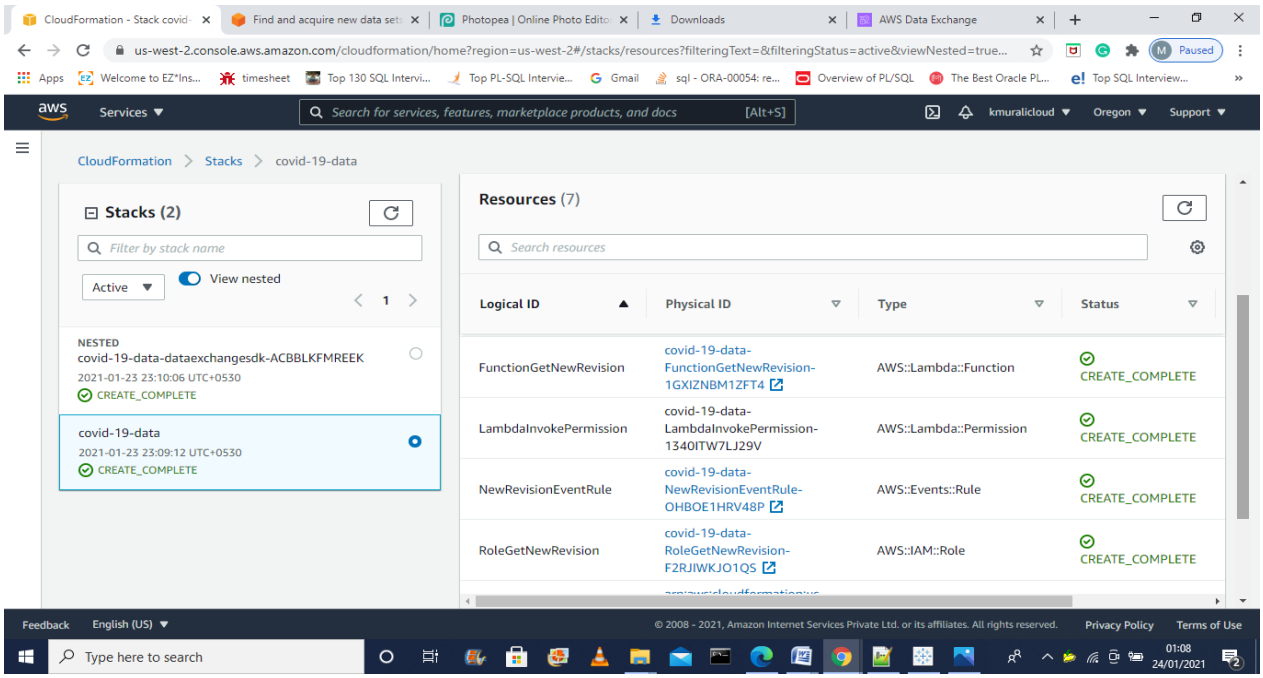
covid-19-data stack summary



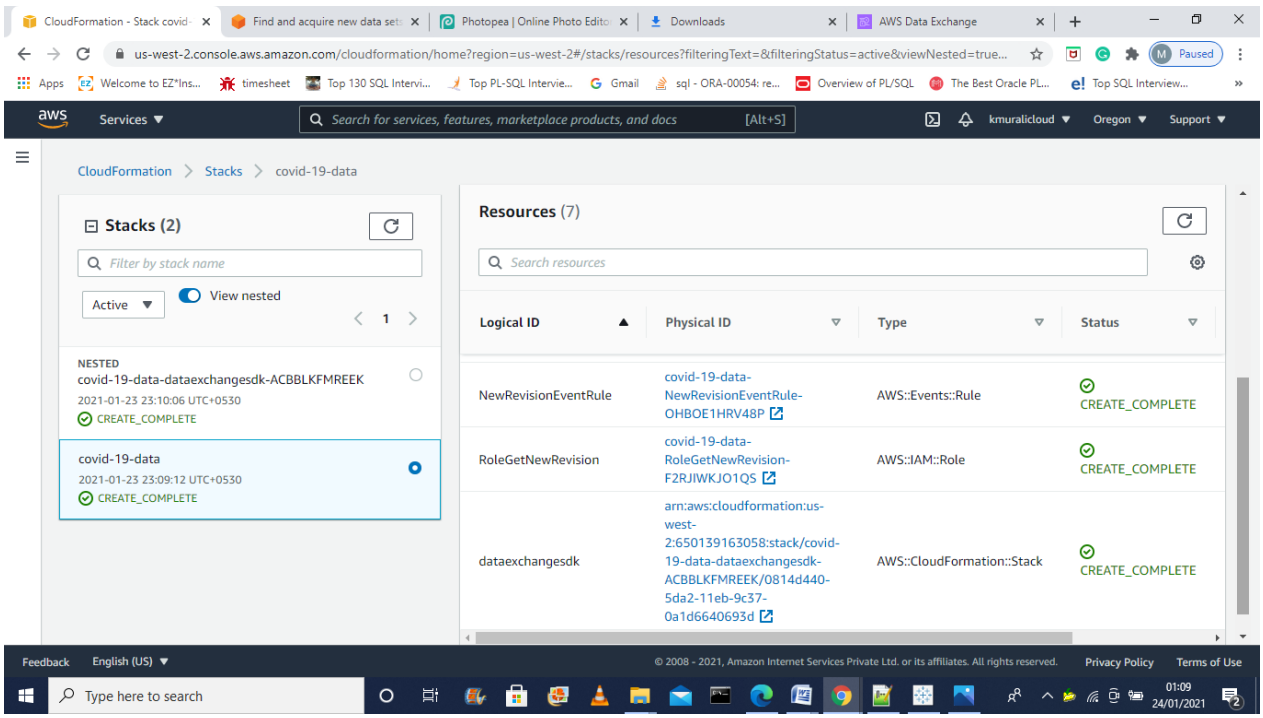
Resources used in covid-19-data stack



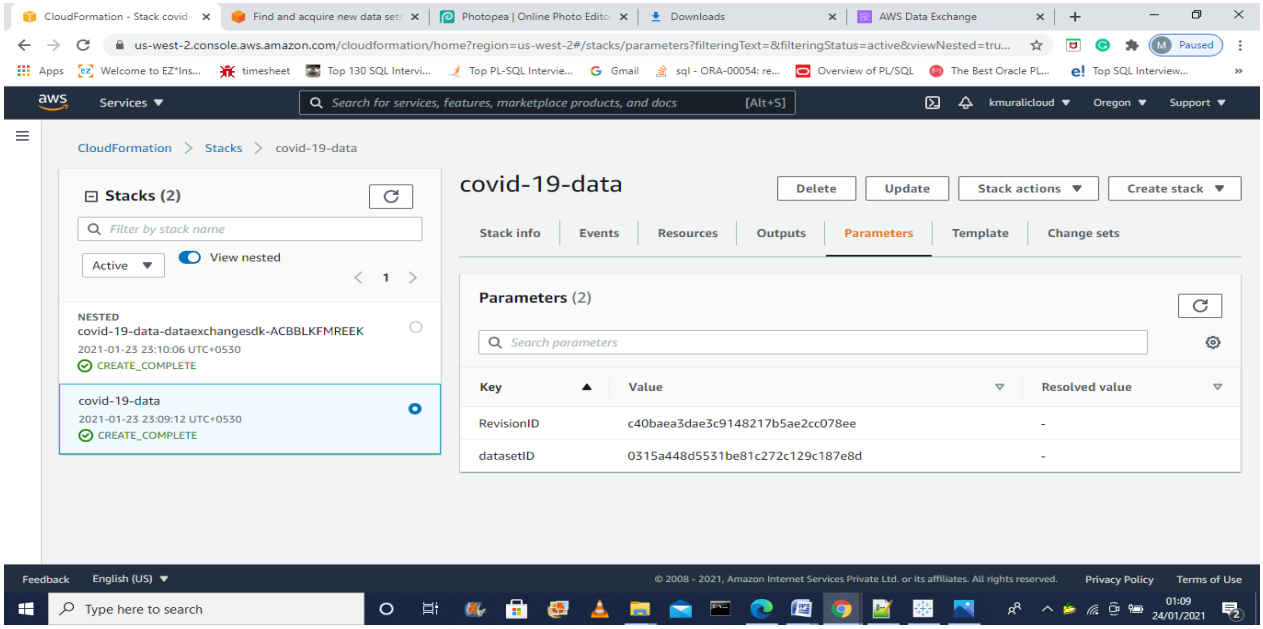
Resources used in covid-19-data stack



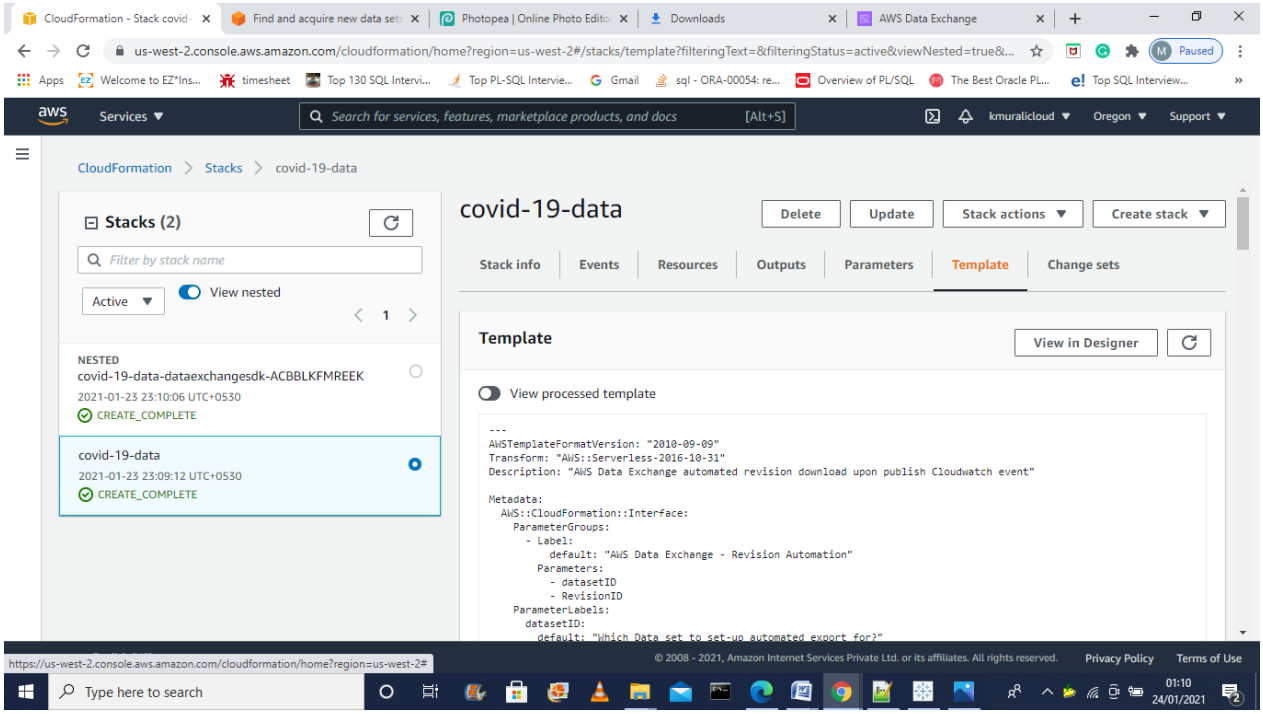
Covid-19-data stack summary



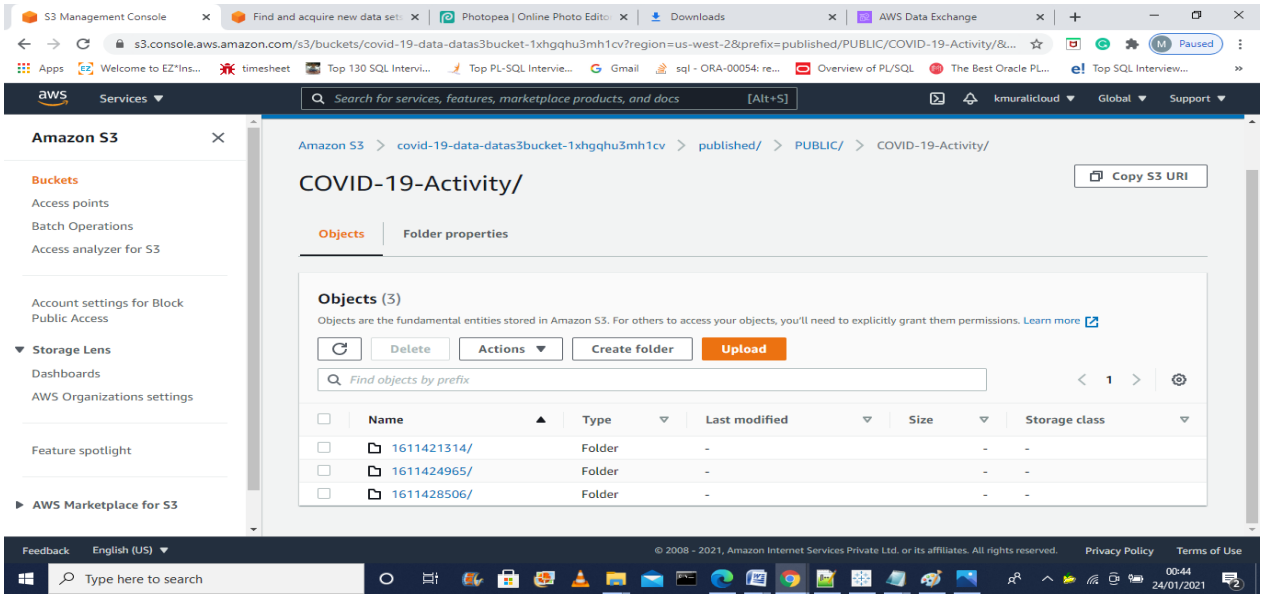
Parameters required for this stack



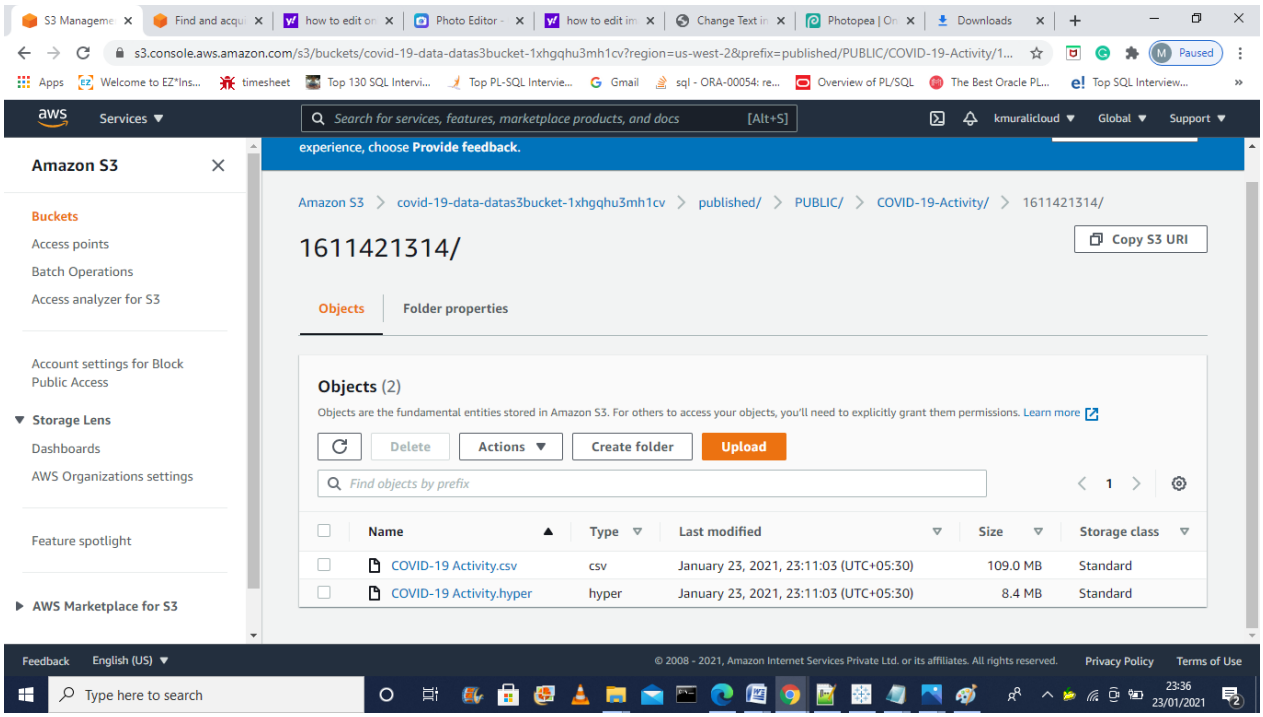
Template details



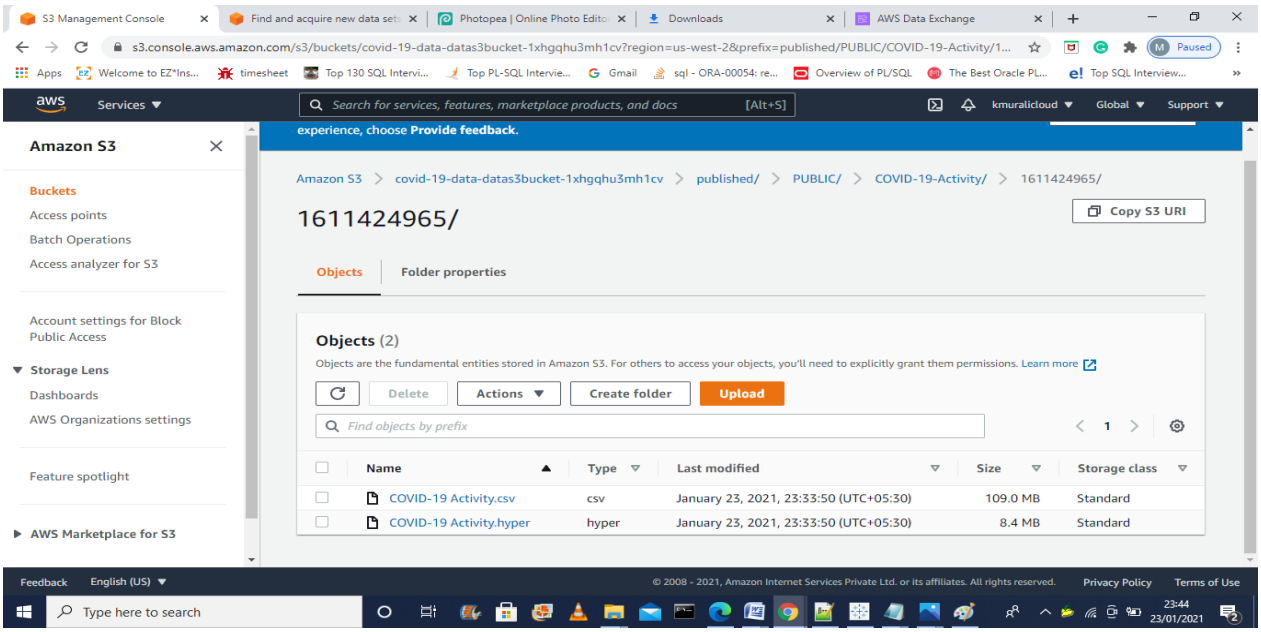
Now AWS Lambda exporting and creating new data sets as show in the below figure for each specific revision of the aws exchange data



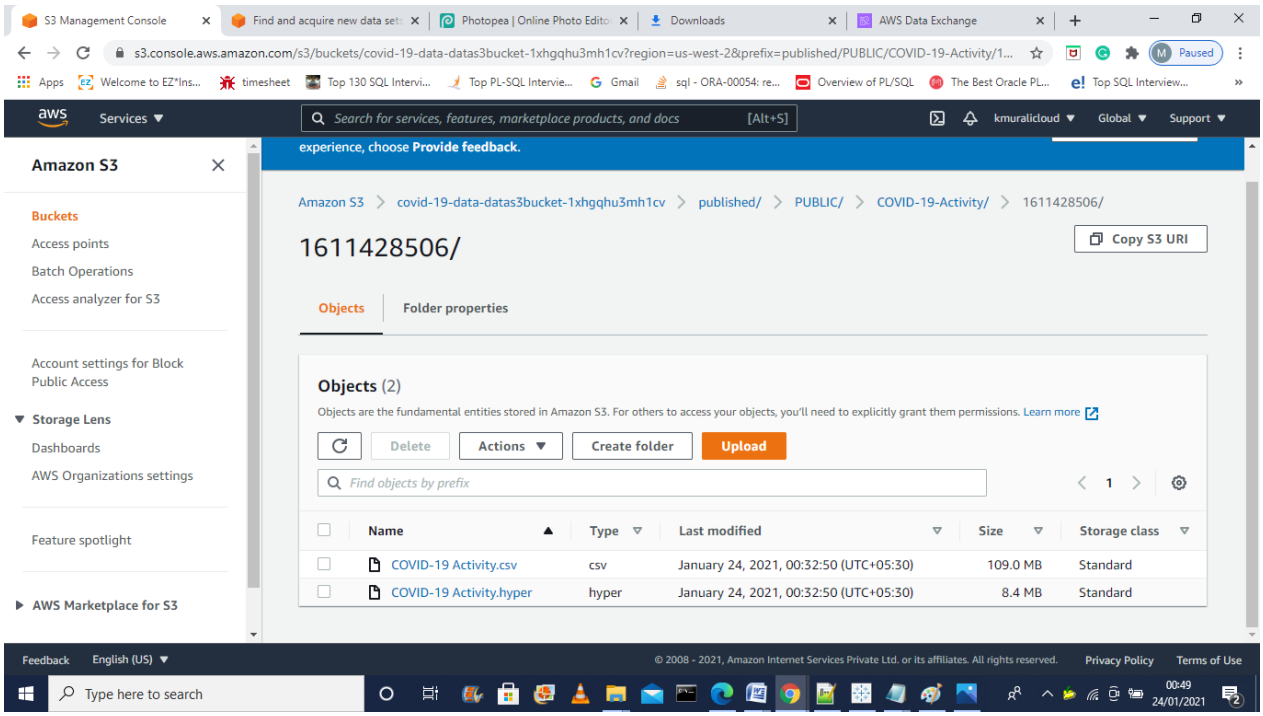
S3 bucket folder created for revision1611421314



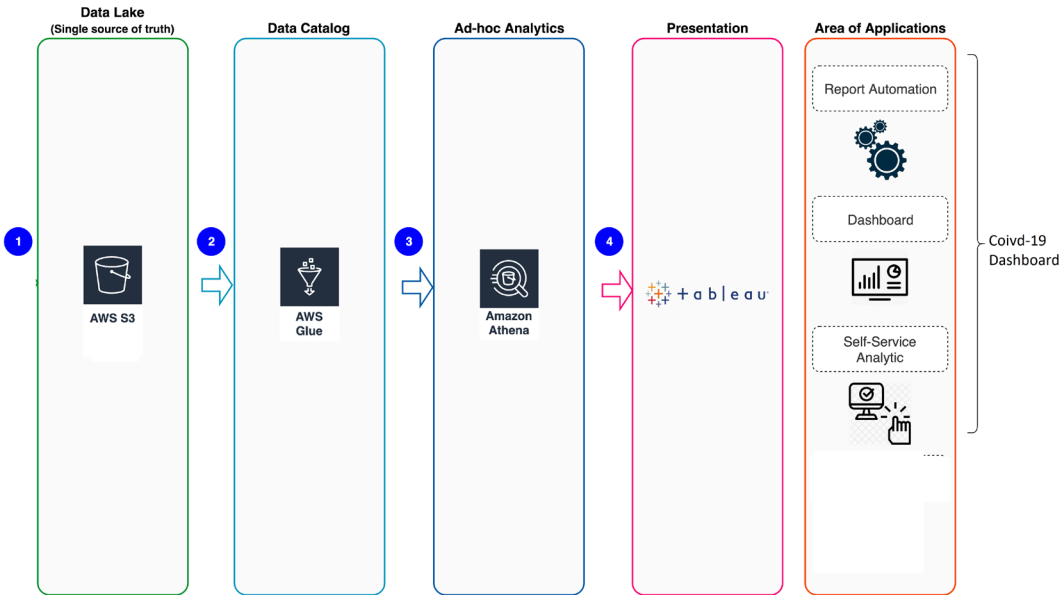
S3 bucket folder created for revision1611424965



For every hour 33min AWS Lambda extract data from aws exchange and place under s3 bucket



Connect to Amazon S3 bucket from Tableau Desktop to build dashboards using Amazon Athena that uses AWS Glue Data Catalog internally to fetch the data from S3.



As illustrated in the preceding diagram, this is a big data processing in this model:

1. Upload the Covid-19 data into the S3 bucket using Lambda function from covid-19 data hub.

2. Setup Glue data catalog to create Glue table.

3. Athena runs the query to create target table with Glue data catalog.

4. Tableau will be configured, connect to Athena, to retrieve data, and show the analytic figure on Tableau dashboard.

AWS Glue introduction

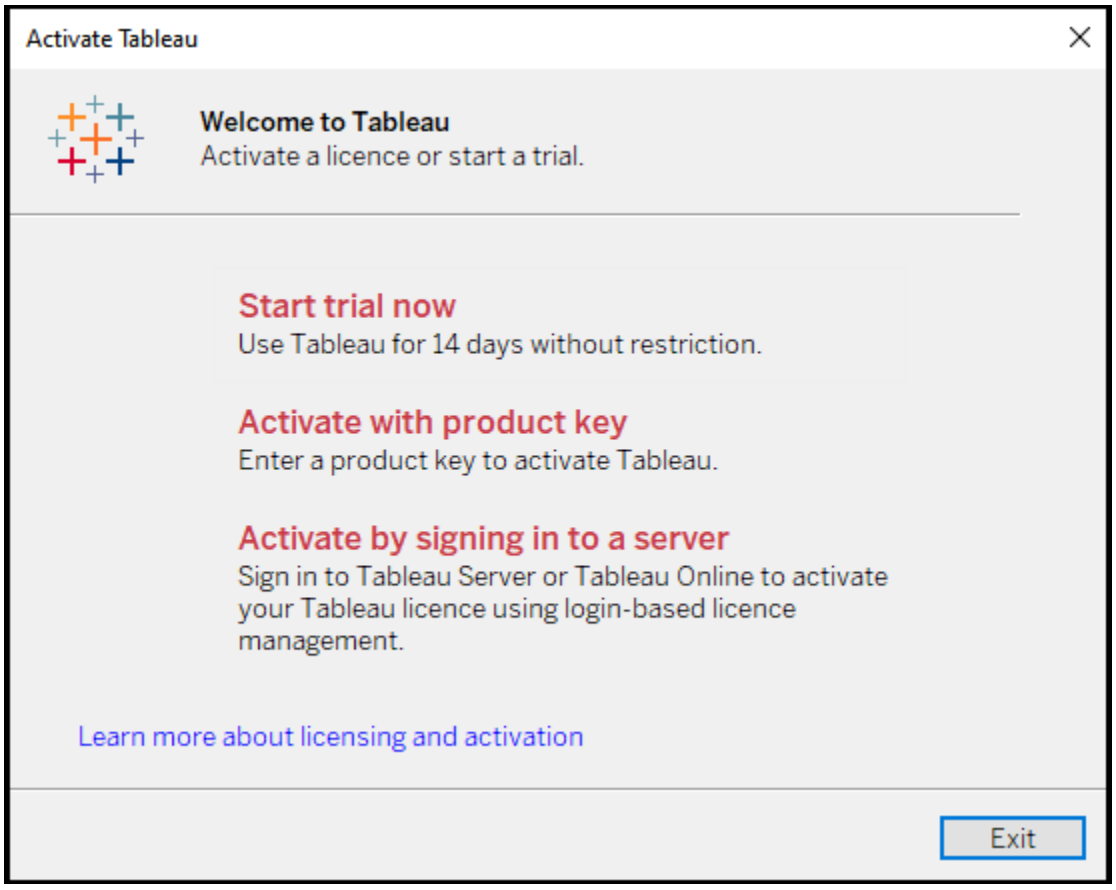
AWS Glue is a fully managed data catalog and ETL (extract, transform, and load) service that simplifies and automates the difficult and time-consuming tasks of data discovery, conversion, and job scheduling. AWS Glue crawls our data sources and constructs a data catalog using pre-built classifiers for popular data formats and data types, including CSV, Apache Parquet, JSON, and more. It is significantly reducing the time and effort that it takes to derive business insights quickly from an Amazon S3 data lake by discovering the structure and form of your data. Also automatically crawls your Amazon S3 data, identifies data formats, and then suggests schemas for use with other AWS analytic services.

Amazon Athena introduction

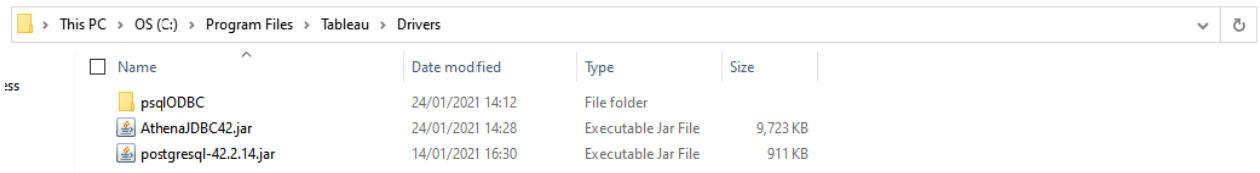
Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run. Athena is easy to use. Simply point to your data in Amazon S3, define the schema, and start querying using standard SQL. Most results are delivered within seconds. With Athena, there's no need for complex ETL jobs to prepare your data for analysis. This makes it easy for anyone with SQL skills to quickly analyze large-scale datasets. Athena is out-of-the-box integrated with AWS Glue Data Catalog, allowing you to create a unified metadata repository across various services, crawl data sources to discover schemas and populate your Catalog with new and modified table and partition definitions, and maintain schema versioning. You can also use Glue's fully-managed ETL capabilities to transform data or convert it into columnar formats to optimize cost and improve performance.

Step 0 - Prerequisites

1. Sign-in a AWS account, and make sure you have select **Oregon** region
2. Make sure account we have permission to create IAM role for following services: **S3, Glue, Athena**
3. Make sure we have created the **Access key** and **Secret access key** that have **Athena** fully permission to connect to Tableau
4. Download **Tableau Desktop** on our laptop.
   Click below link to download
   https://www.tableau.com/support/releases
   Note that download the latest version (2020.4 for this example)
   Make sure that we have to use our 14-day trail



5. Setup AWS Athena Driver for Tableau Desktop
   If Java is not already installed on windows pc, download and install the latest Java version from https://www.java.com/en/download.
   Download the JDBC driver (.jar file) from the Amazon Athena User Guide on Amazon's website.
   https://docs.aws.amazon.com/athena/latest/ug/connect-with-jdbc.html
   For Windows, Move the downloaded .jar file to C:\Program Files\Tableau\Drivers.



Step 1 - AWS environment setup

login to AWS console

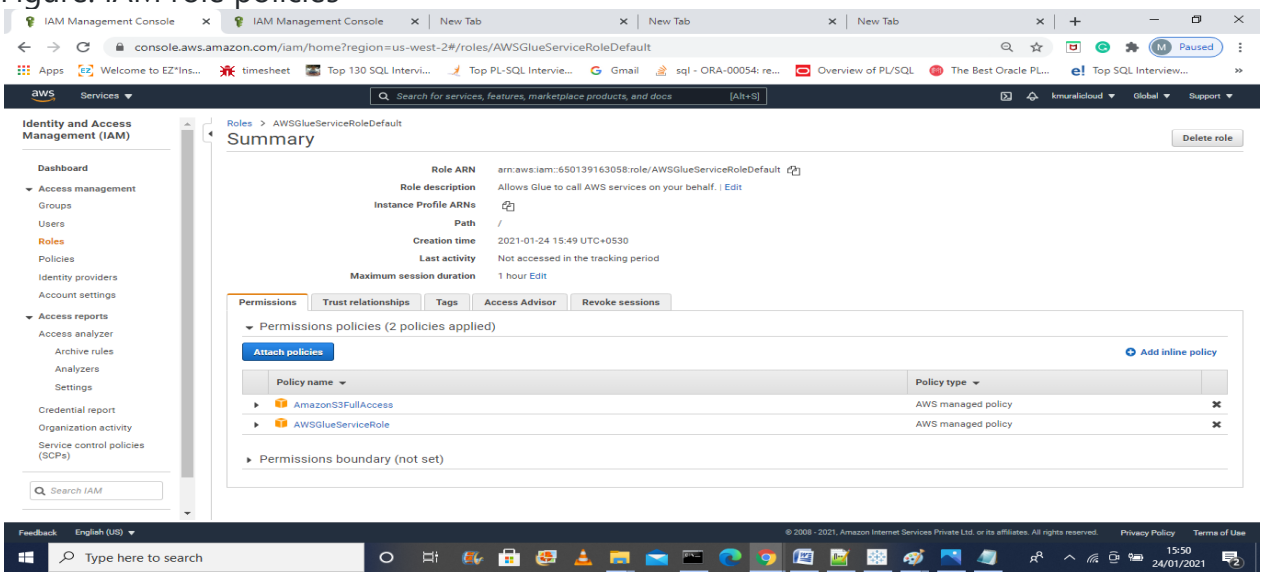
Create Access key and Secret access key on AWS

- To create a new secret access key for your root account, use the security credentials page. Expand the **Access Keys** section, and then click **Create New Root Key**.
- To create a new secret access key for an IAM user, open the IAM console. Click **Users** in the **Details** pane, click the appropriate IAM user, and then click **Create Access Key** on the **Security Credentials** tab.
- Download the newly created credentials (**csv file**), when prompted to do so in the key creation wizard

Setp 2 - Create IAM roles for Glue service

- On the service menu, click IAM.
- In the navigation pane, choose Roles.
- Click Create role.
- For role type, choose AWS Service, find and choose Glue, and choose Next: Permissions.
- On the Attach permissions policy page, search and choose AmazonS3FullAccess, AWSGlueServiceRole, and choose Next: Tags then click Next: Review.
- On the Review page, enter the following detail:
  Role name: AWSGlueServiceRoleDefault
- Click Create role.
- Choose Roles page, select the role AWSGlueServiceDefault you just created.
- Now confirm you have policies as below figure

- Figure: IAM role policies



We successfully create the role that allow AWS Glue get access to S3.

Setp 3 - Create S3 bucket for data lake and staging

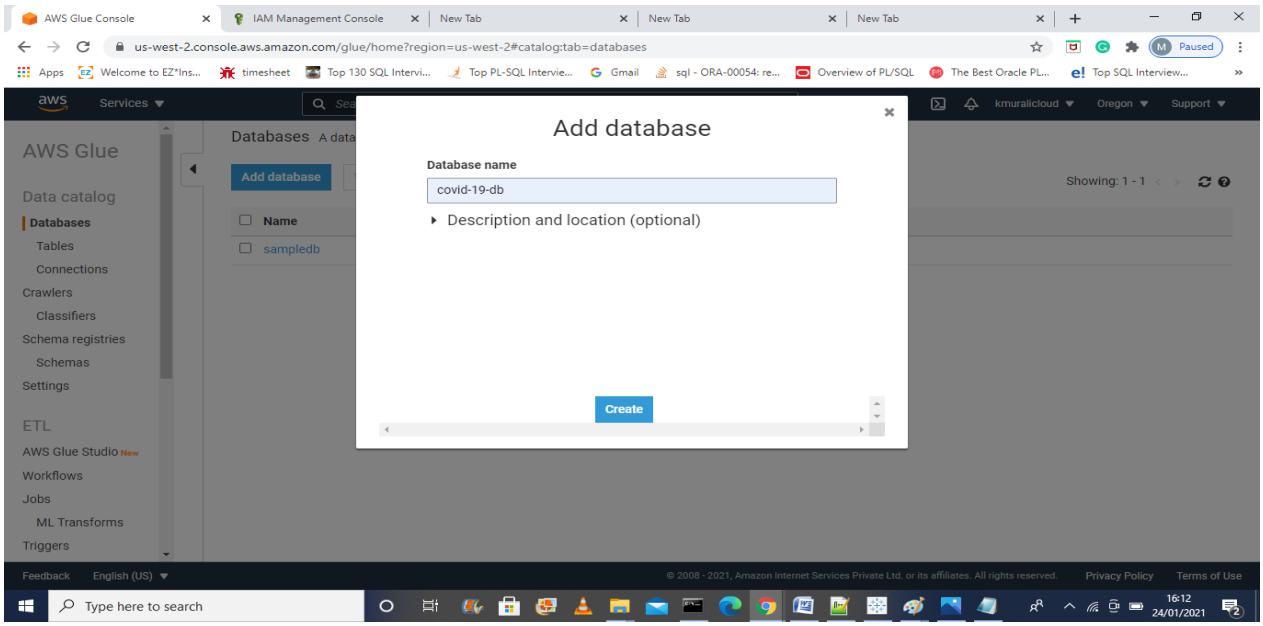
In our Scenario AWS Lambda will create S3 bucket as mentioned in the above screenshots.

Step 4 - Setup AWS Glue data catalog

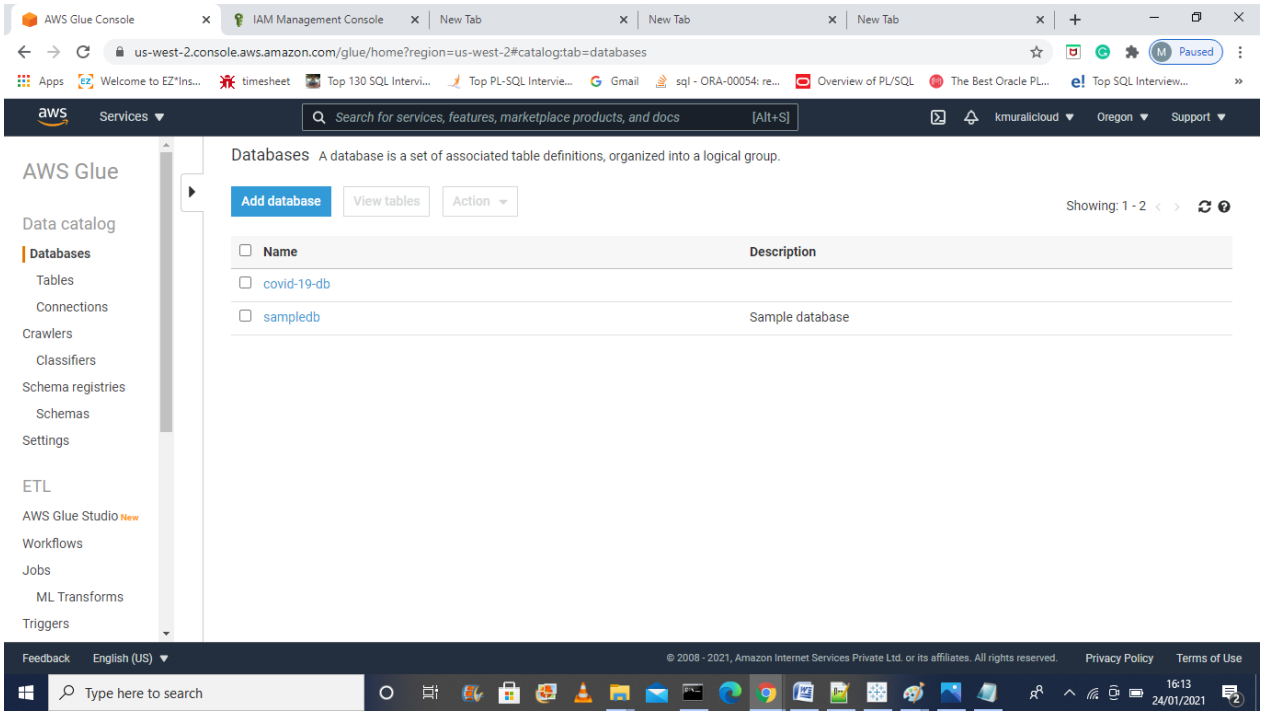
Create database, tables, crawlers, in Glue Data Catalog

- On the **Services** menu, click **AWS Glue**.
- In the console, choose **Add database**. In the **Database name**, type **covid-19-db**, and choose **Create**.

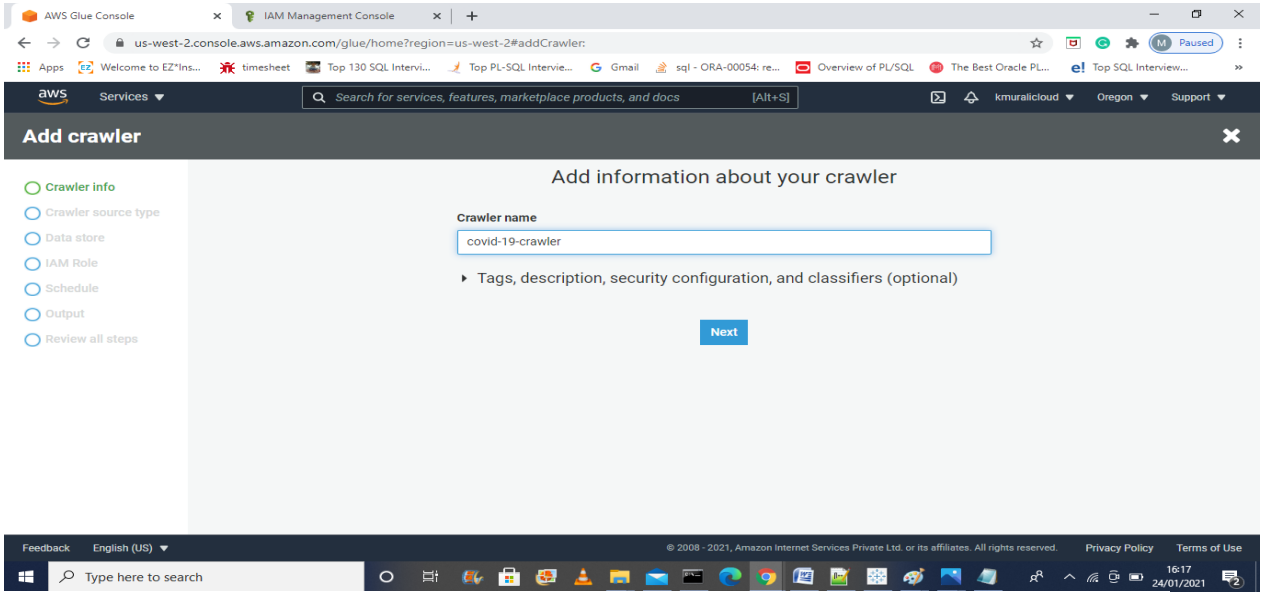
Create Database



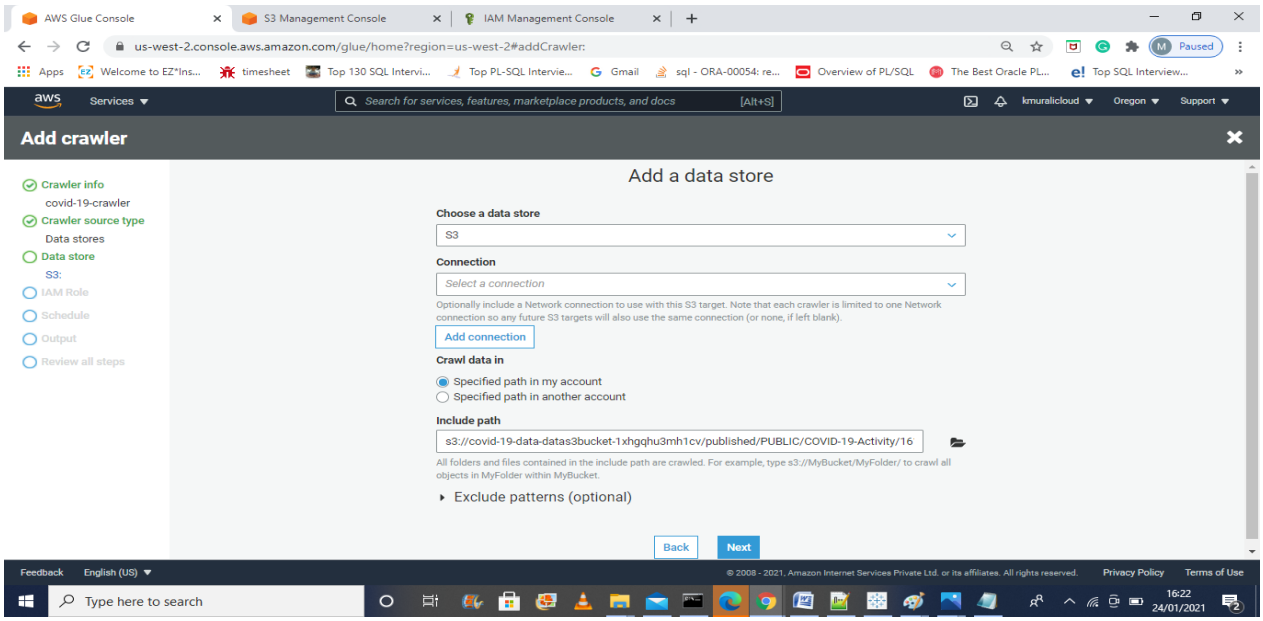
Database added as shown below



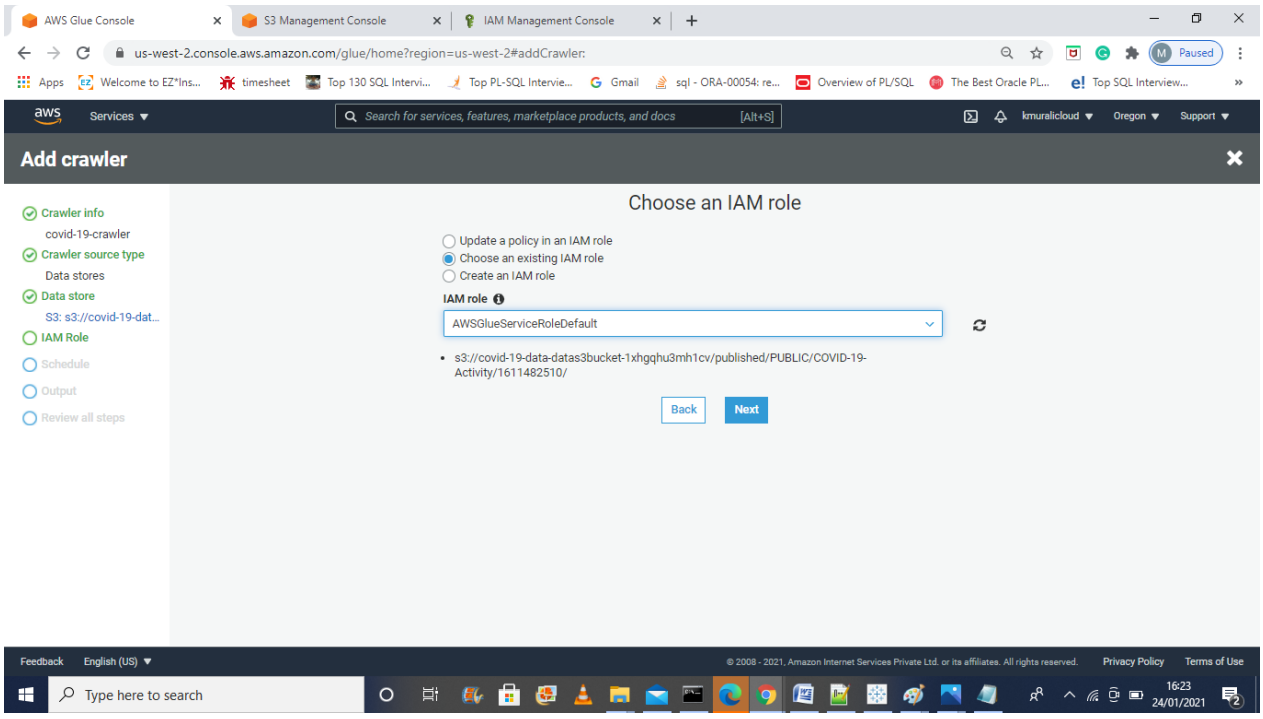
- Choose **Crawlers** in the navigation pane, choose **Add crawler**. Enter the Crawler name **basics-crawler**, and choose **Next**.



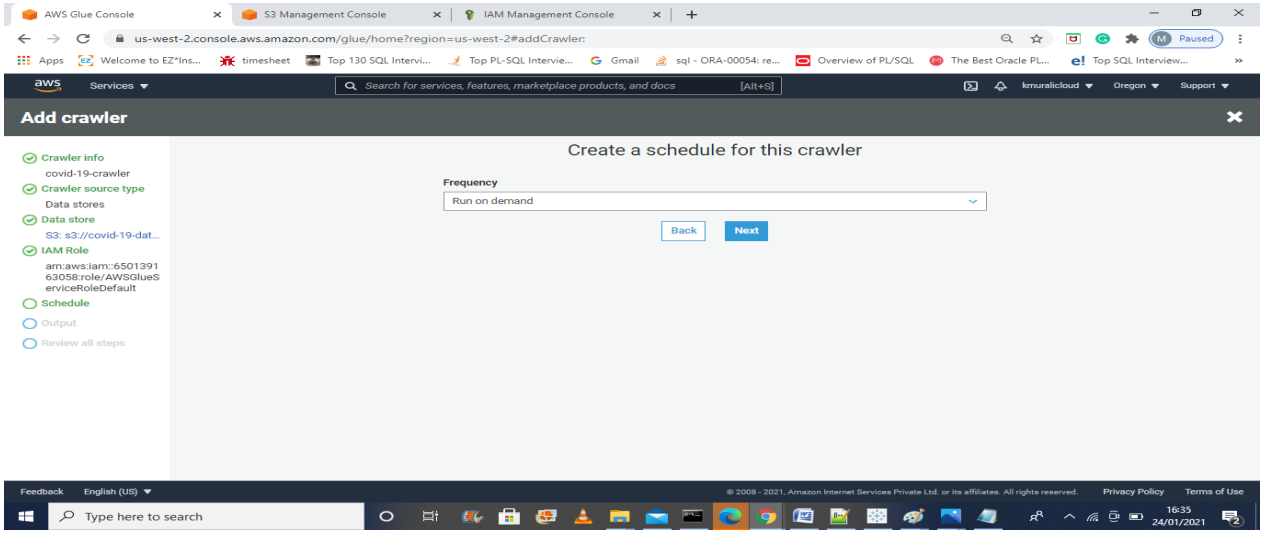
- On the **Add a data store** page, choose **S3** as data store.
- Select **Specified path in my account**.
- Select **Lambda created** folder in the bucket, and choose **Next**.



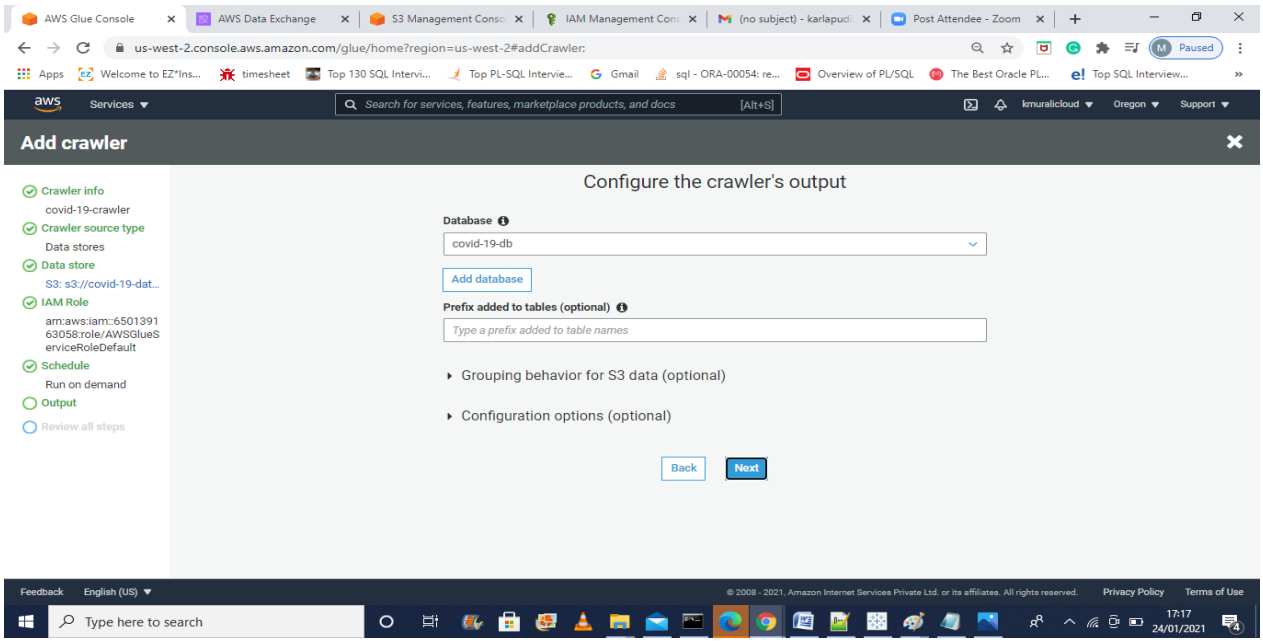
- On **Add another data store** page, choose **No**, and choose **Next**.
- Select **Choose an existing IAM role**, and choose throle **AWSGlueServiceRoleDefault** you just created in the drop-down list, and choose **Next**



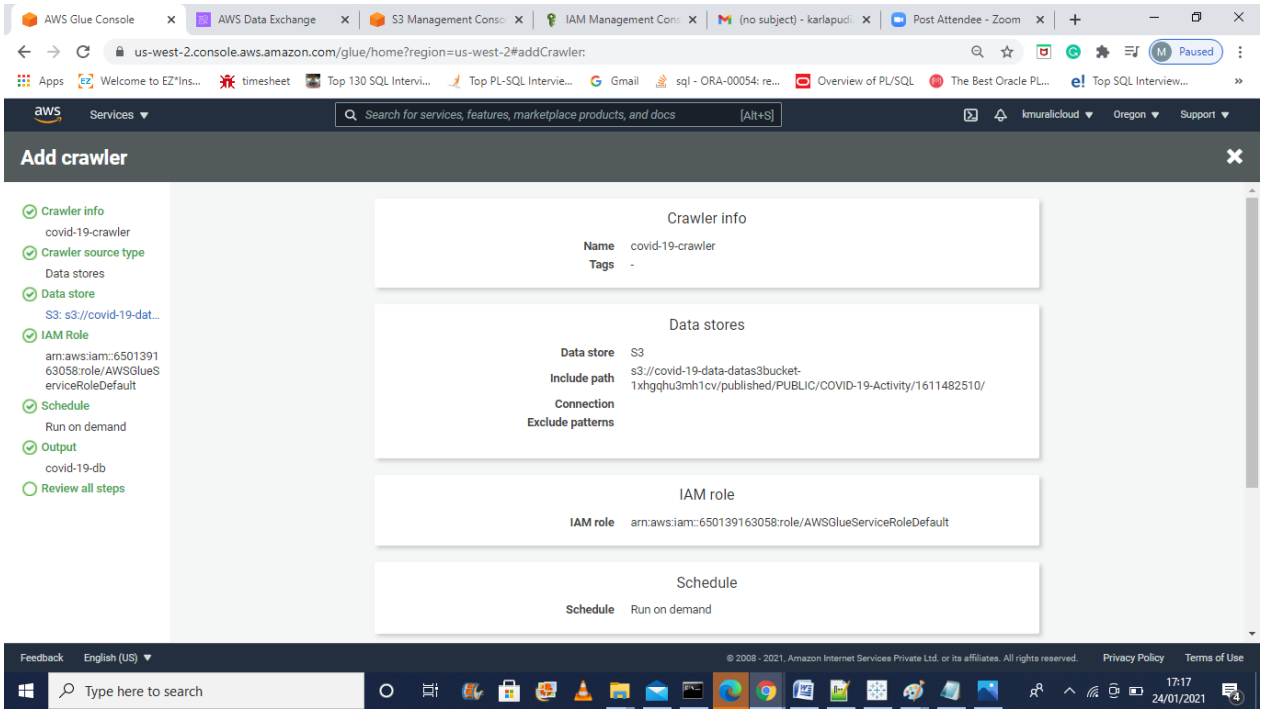
- For **Frequency**, choose **Run on demand**, and choose **Next**.



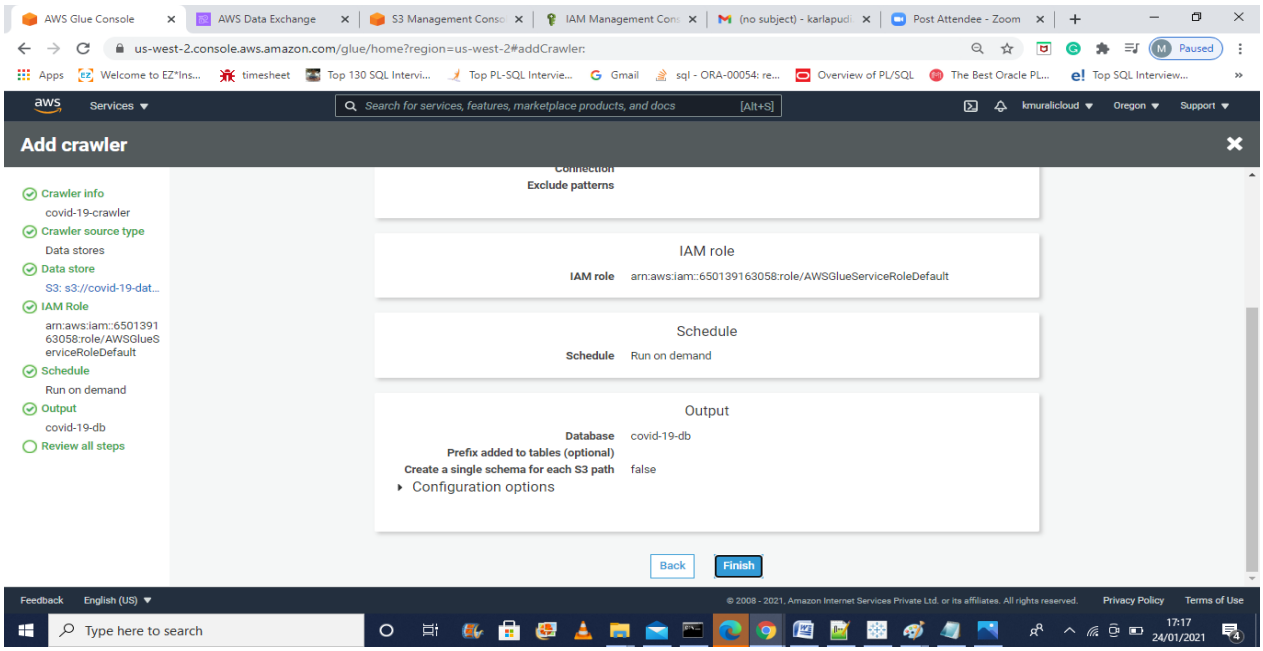
- For **Database**, choose **covid-19-data**, and choose **Next**.



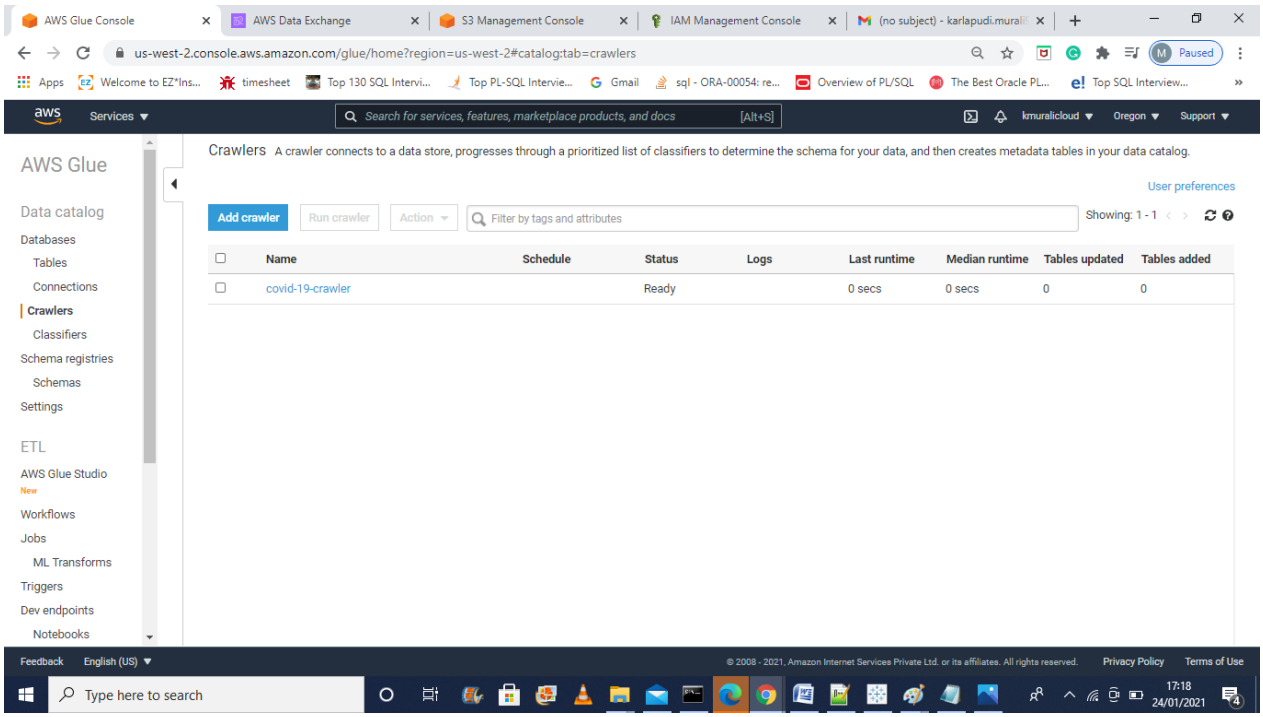
- Review the steps, and choose Finish.



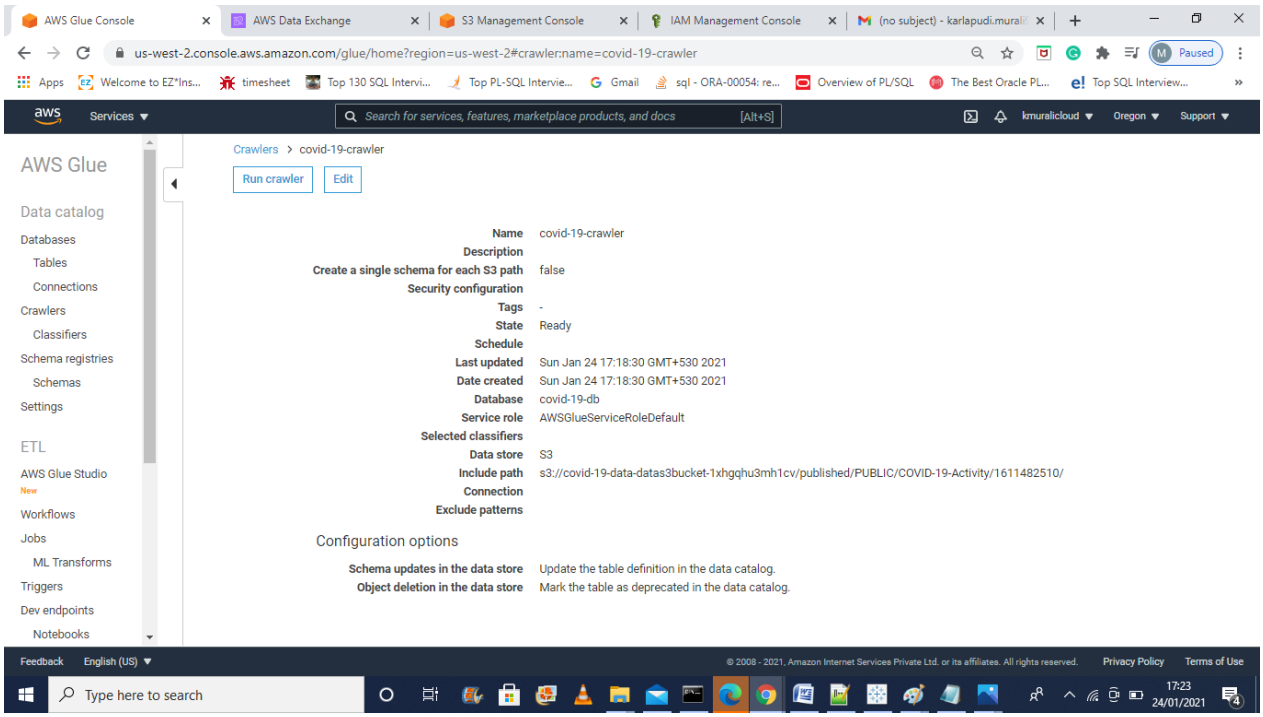
Crawler creation summary page, click on finish to create



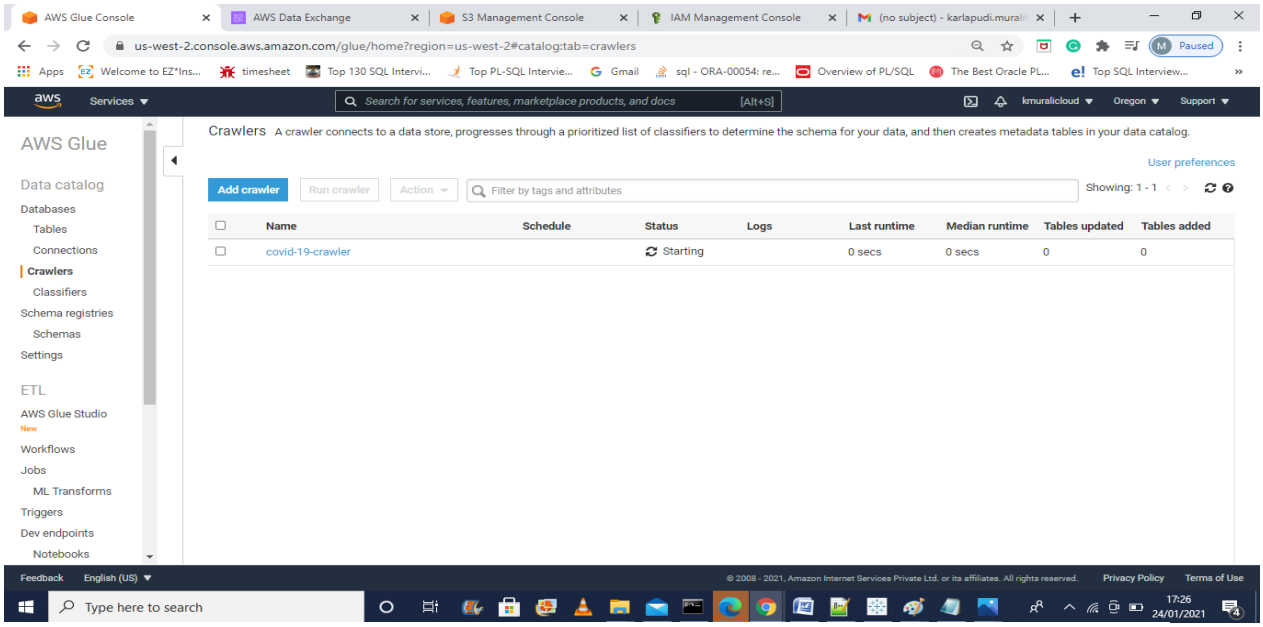
Covid-19 Crawler created



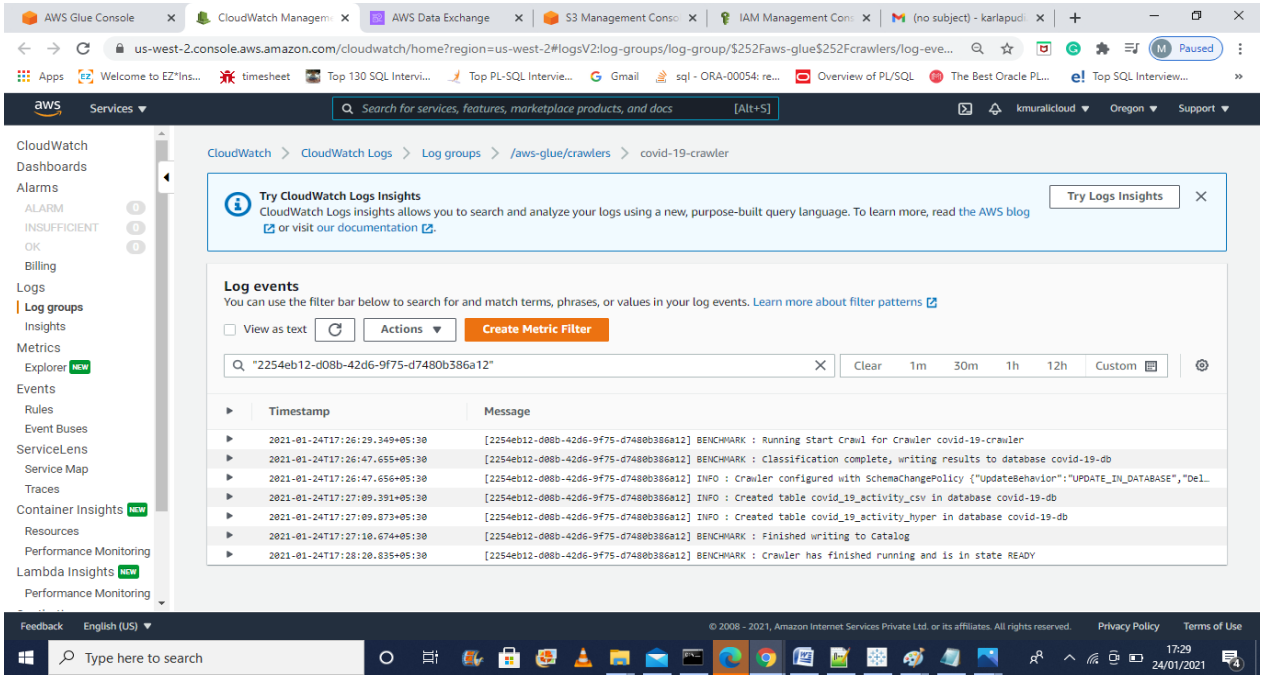
Covid-19 Crawler properties:



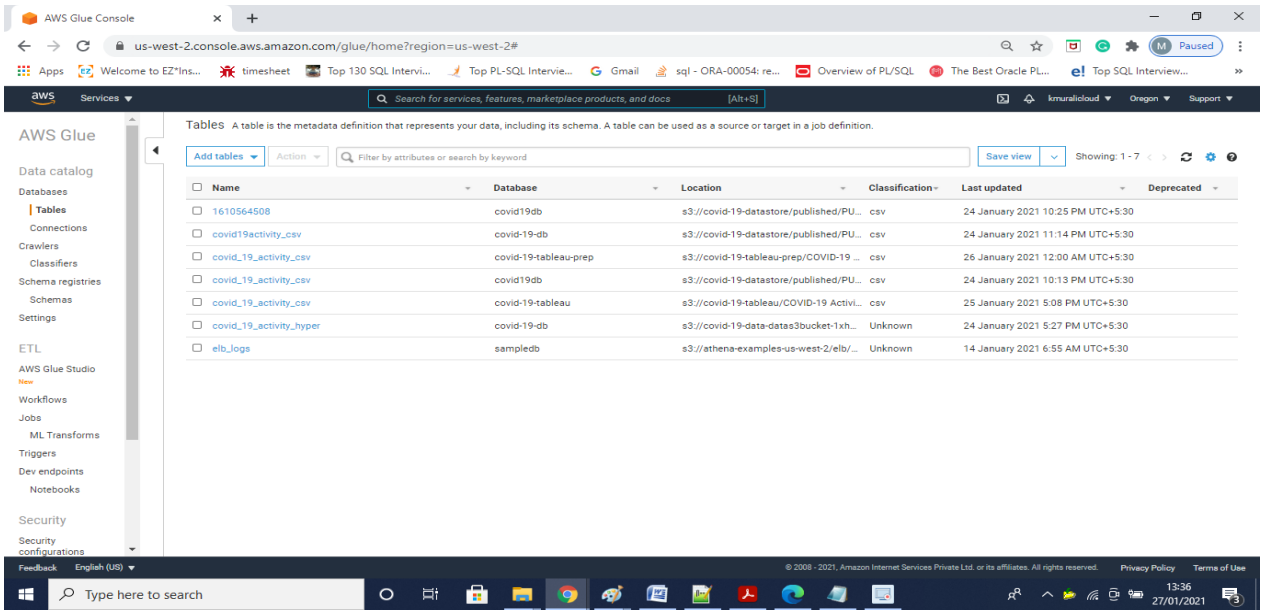
- The crawler is ready to run. Choose **Run it now.**
  Now the **basics-crawler** is crawling the data in basics folder in S3 bucket.



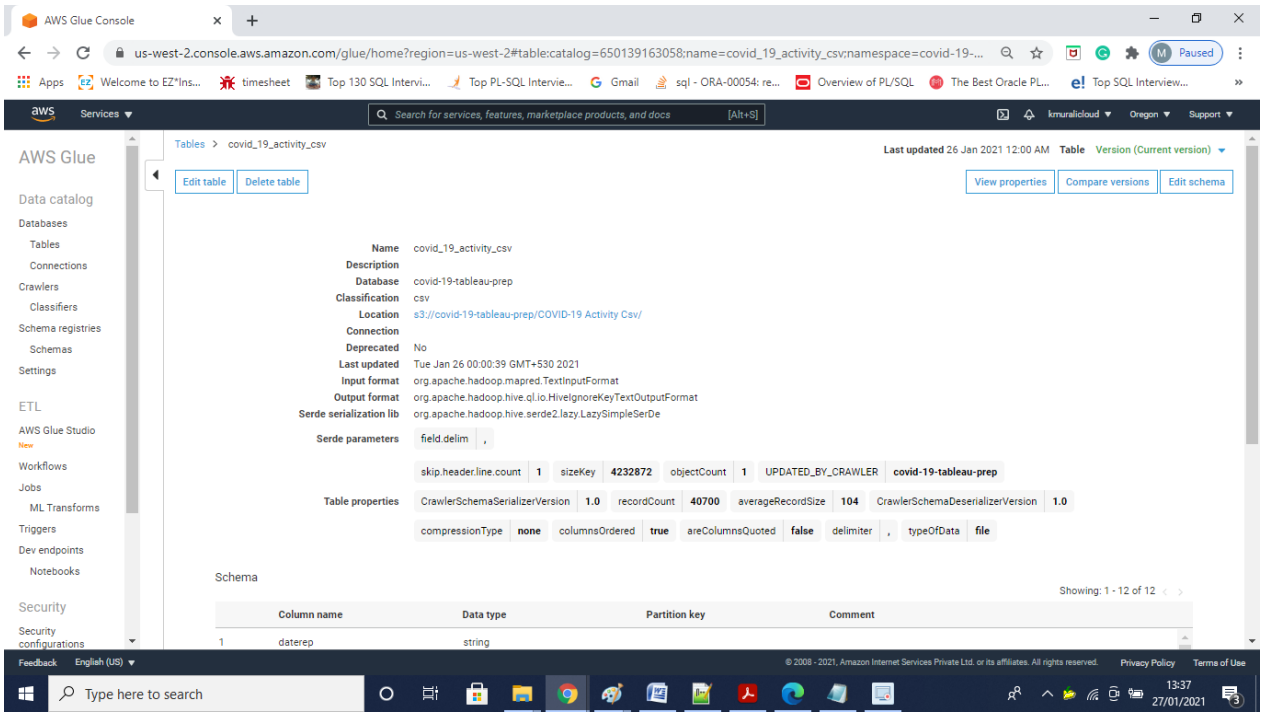
Crawler cloudwatch logs:



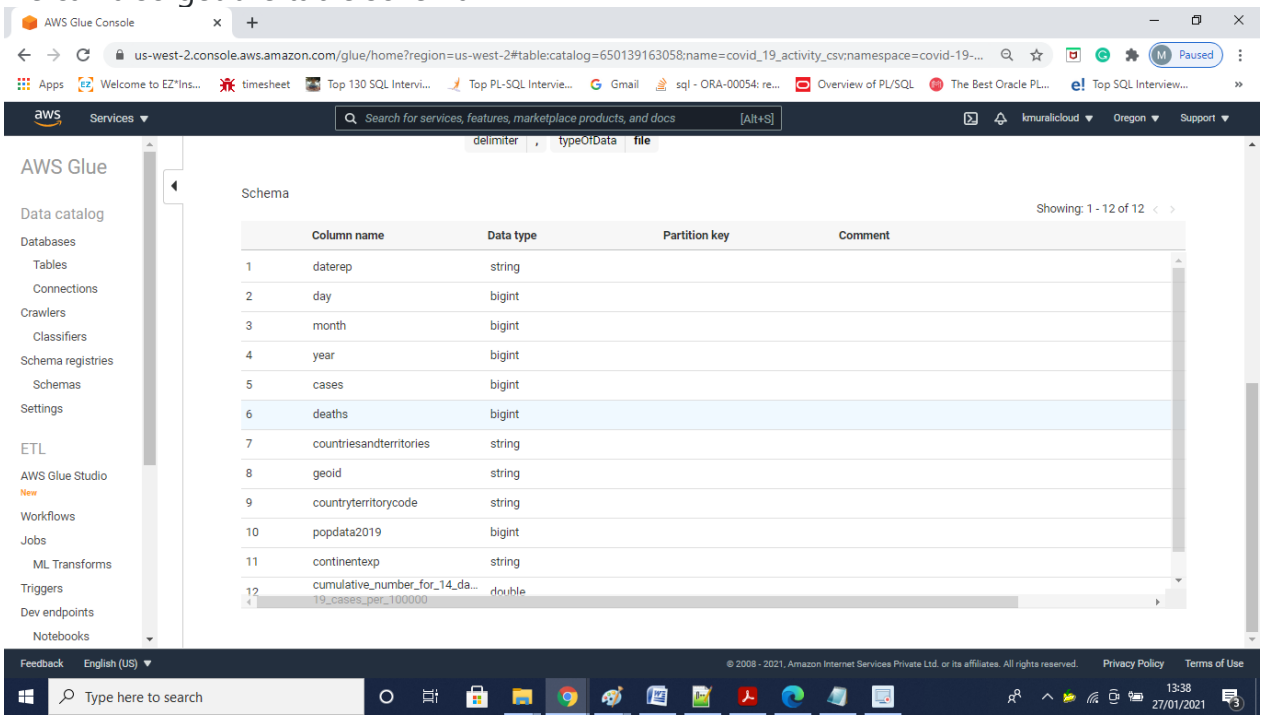
- When the crawler has finished, two table has been added. Choose **Tables** in the left navigation pane, and then choose **basics** to confirmed.



We can get the table information such as S3 location



we can also get the table schema

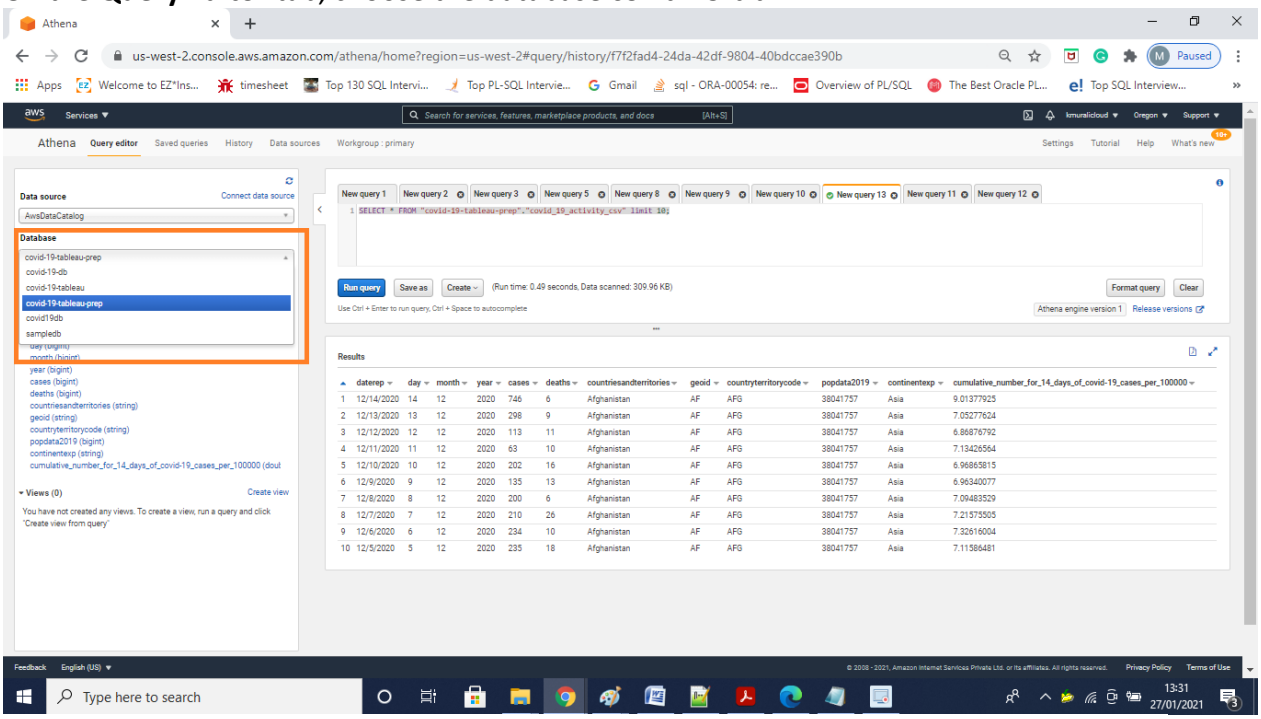


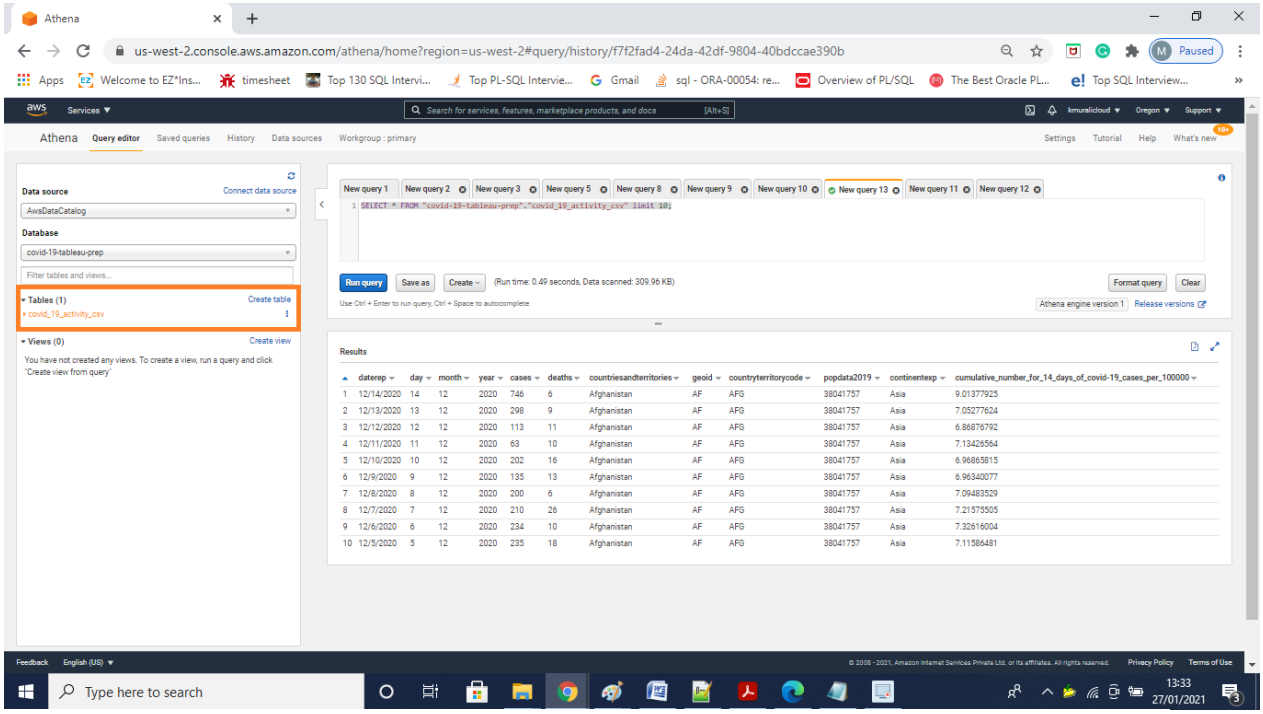
Now we successfully to setup AWS Glue data catalog and create Glue table with covid-19 data

Step 5 - Ad Hoc query in with AWS Athena

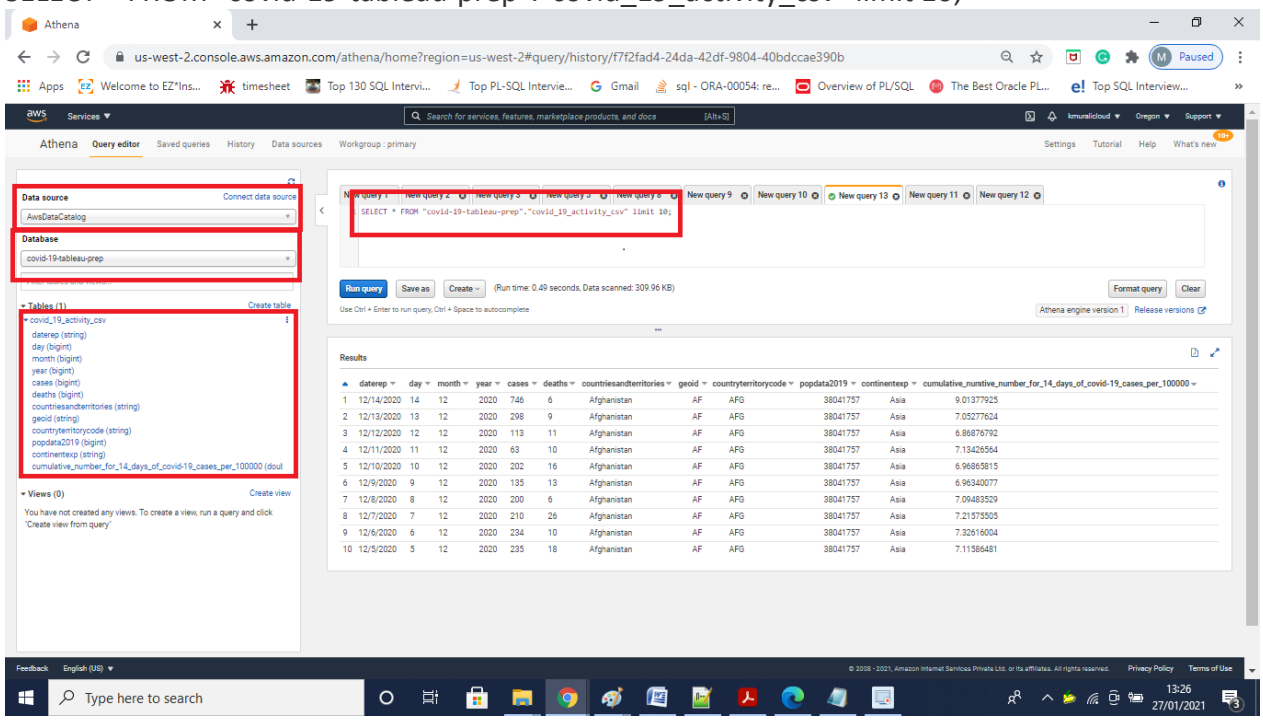
Athena can query the data in an easy way with Glue Data Catalog

- On the **Services** menu, click **Athena**.
- On the **Query Editor** tab, choose the database **covid-19-db**.

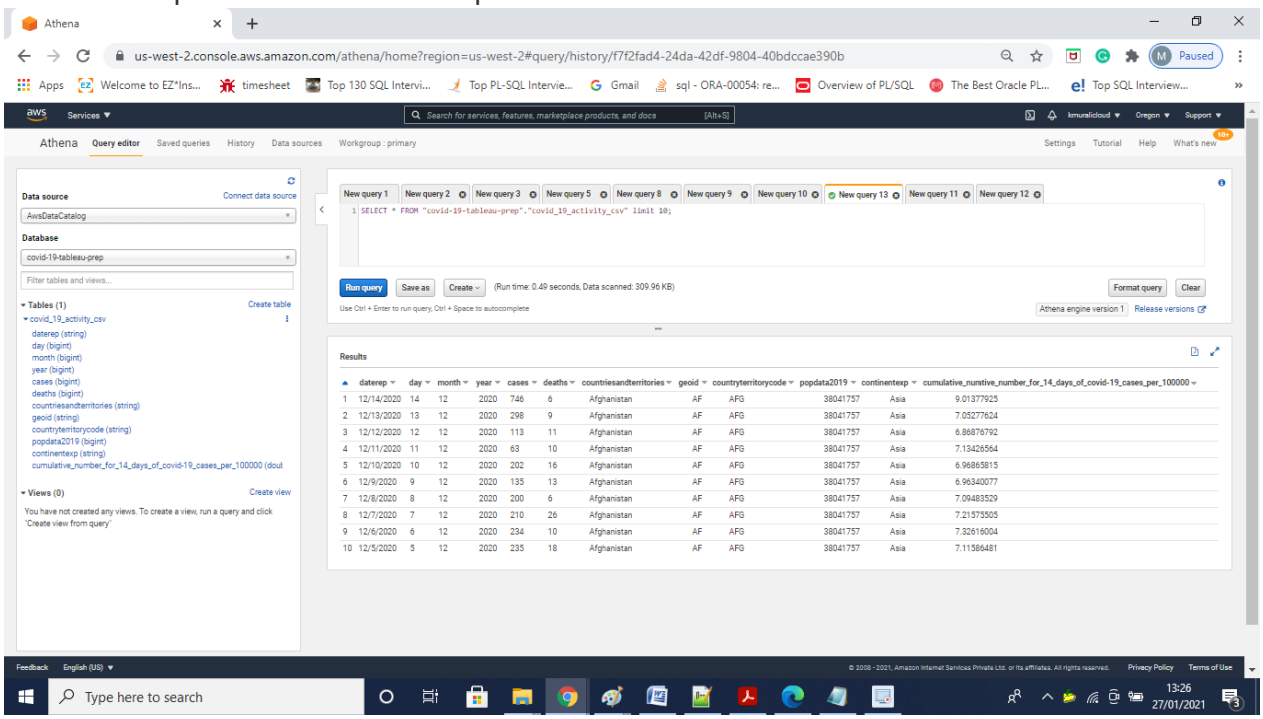




- Query the data, paste below standard SQL in the blank:
SELECT * FROM "covid-19-tableau-prep"."covid_19_activity_csv" limit 10;



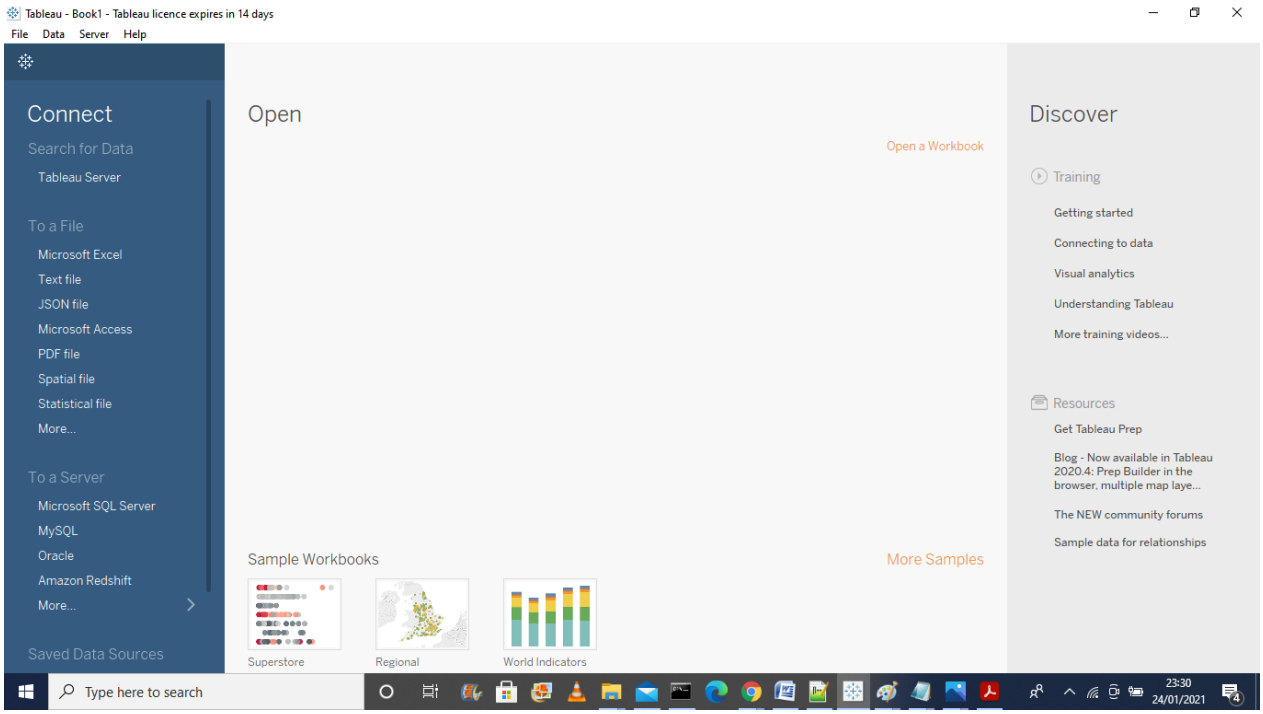
- You can also preview the table to explore the data



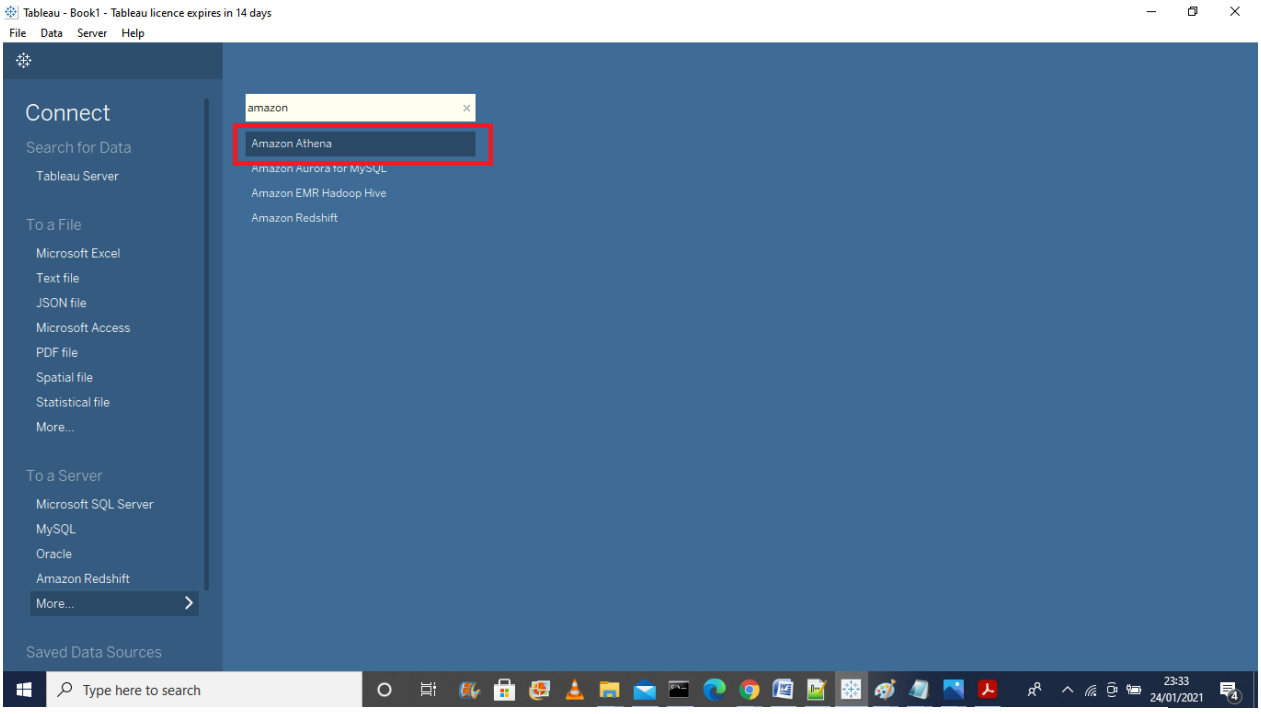
Step 6 - Setup Tableau desktop connection to Athena

The following steps will show you how to use Tableau to create the views with Athena table.

- First we need to download Tableau Desktop on your laptop.
- In this step assume we have installed Tableau Desktop.
- Open Tableau Desktop you will see this screen



- To connect to Athena, click **Amazon Athena** in navigation pane left side



- Enter **"athena.ap-southeast-1.amazonaws.com"** in **Server**
- Enter port for 443

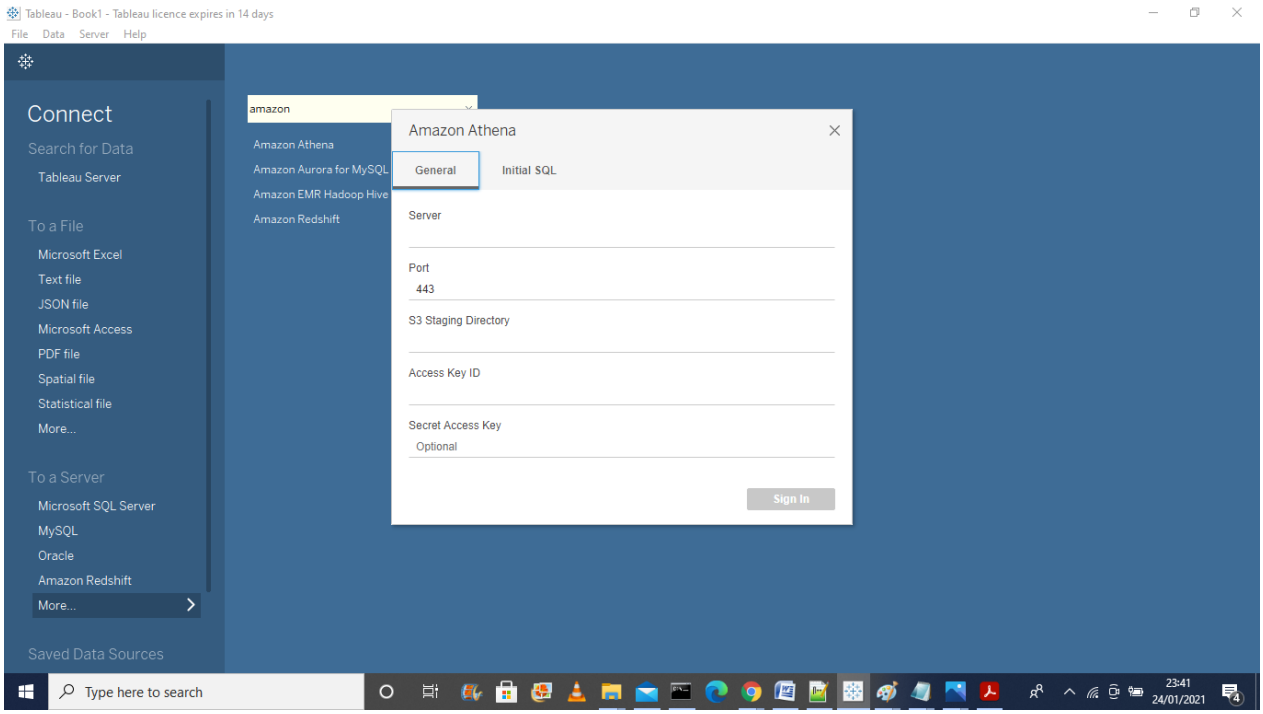
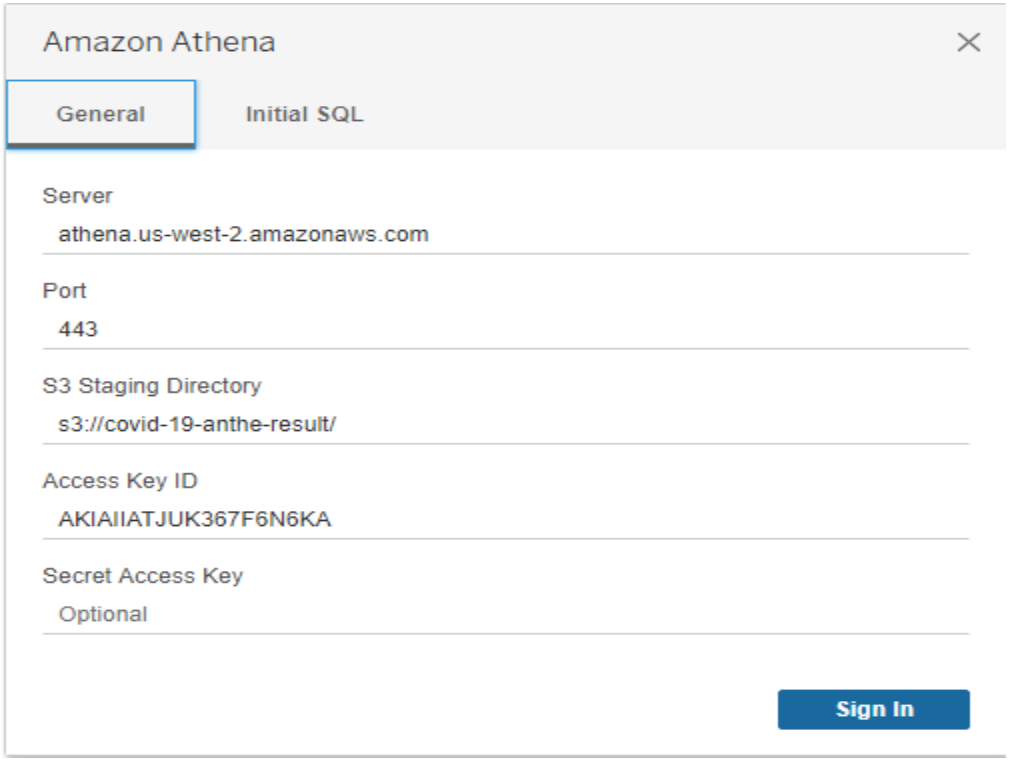
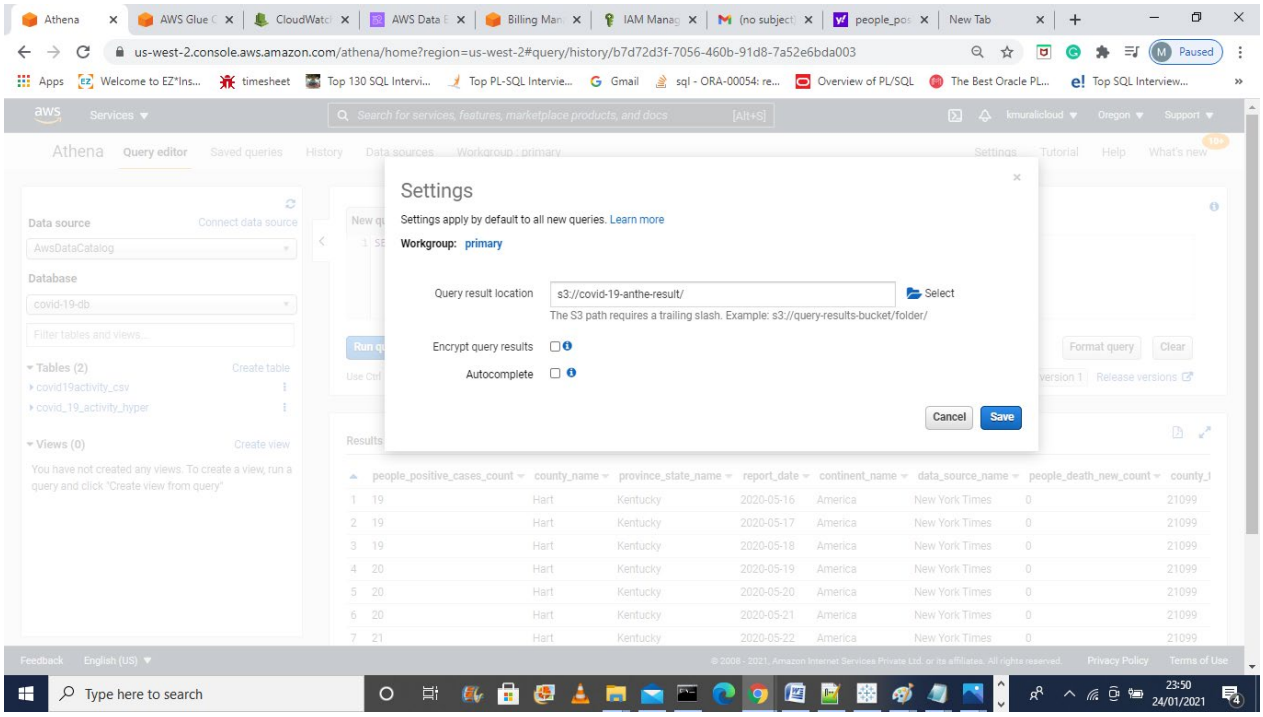


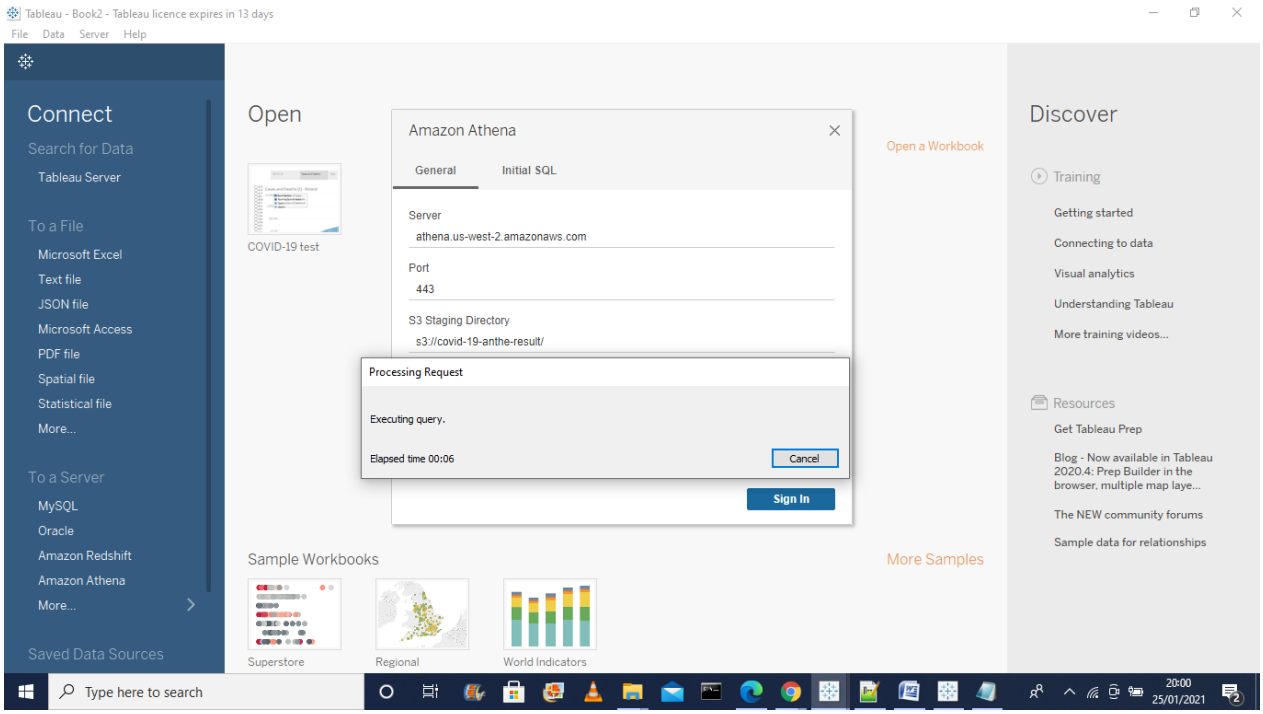
Tableau & Athena Connection Properties



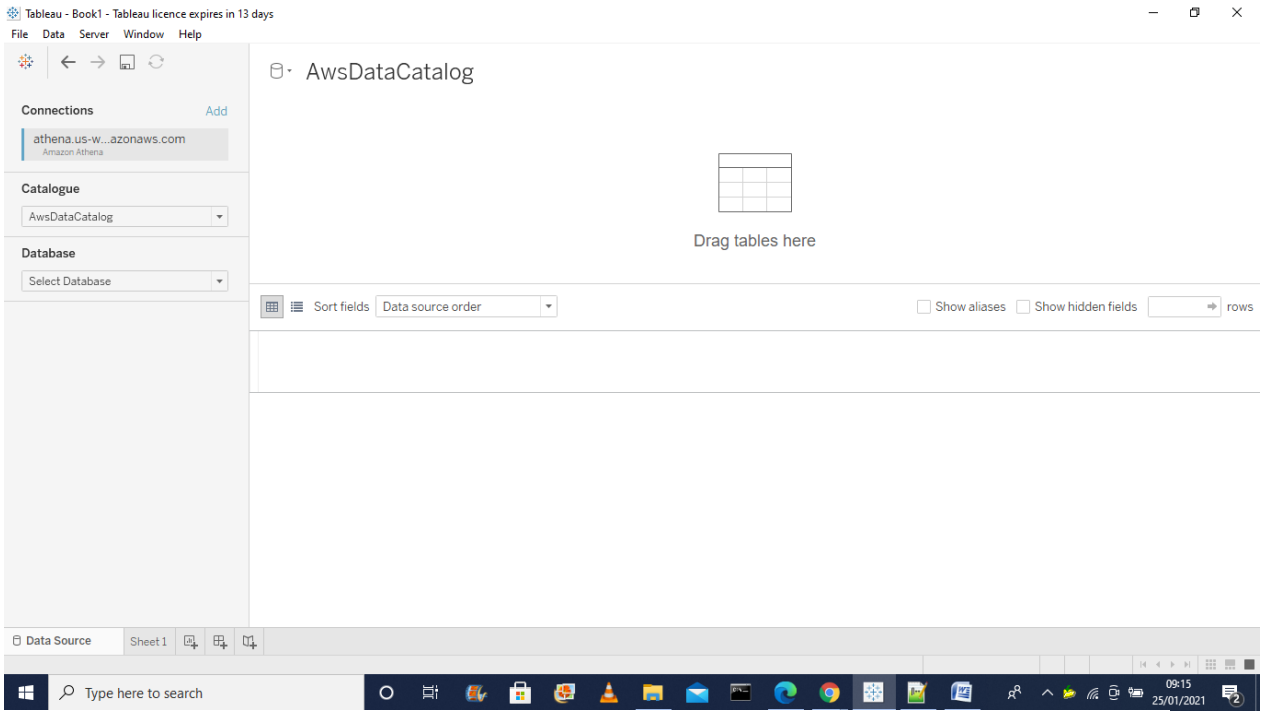
- Enter Staging Directory for your Athena query result S3 bucket
  Go to Athena console and click Settings to get the staging directory path



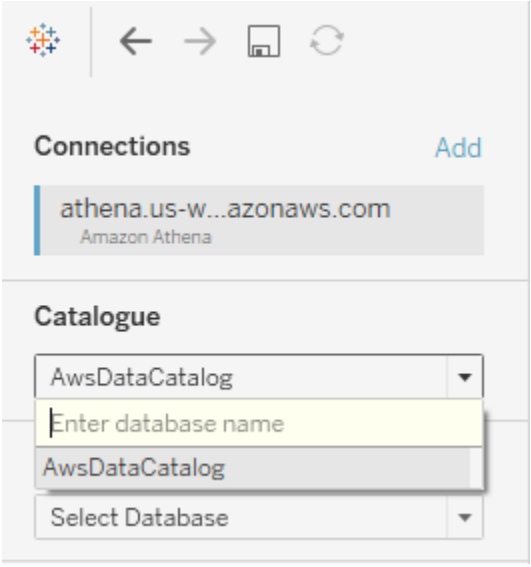
- Enter **Access Key ID** and **Secret Access Key** which you have created on AWS (**you can view these two items in credential csv**) then click **sign in**



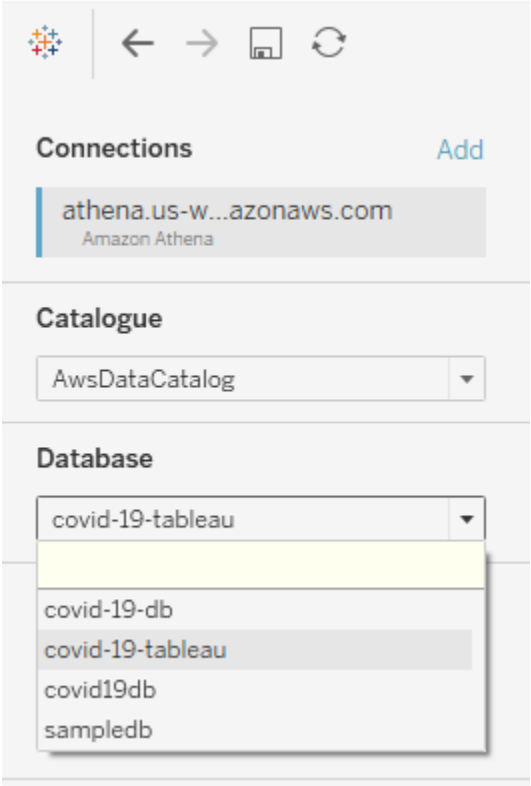
Connected to tableau



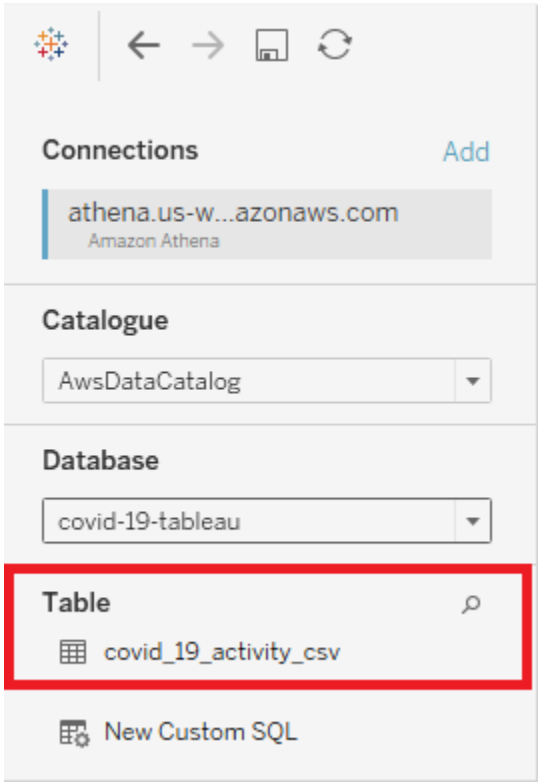
- Select Catalogue as AwsDataCatalog from the below dropdown value.



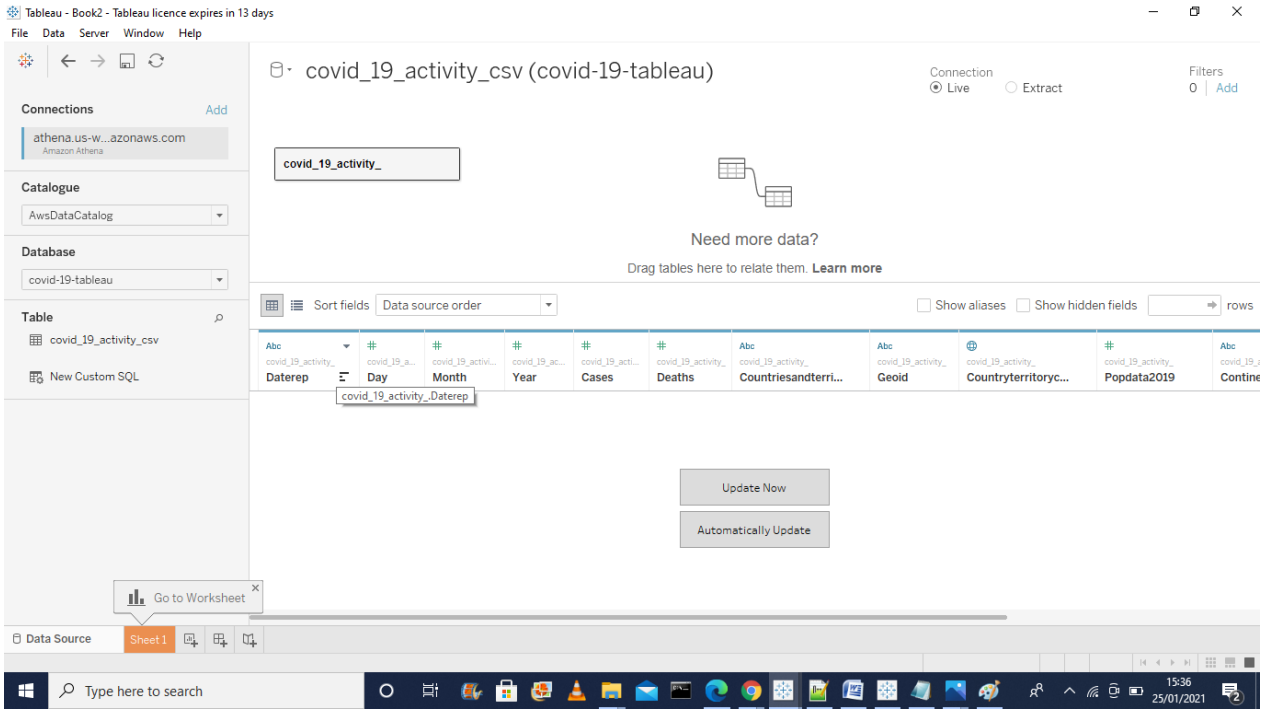
- Select the Database what we created in the AWS Anthena as “covid-19-tableau”



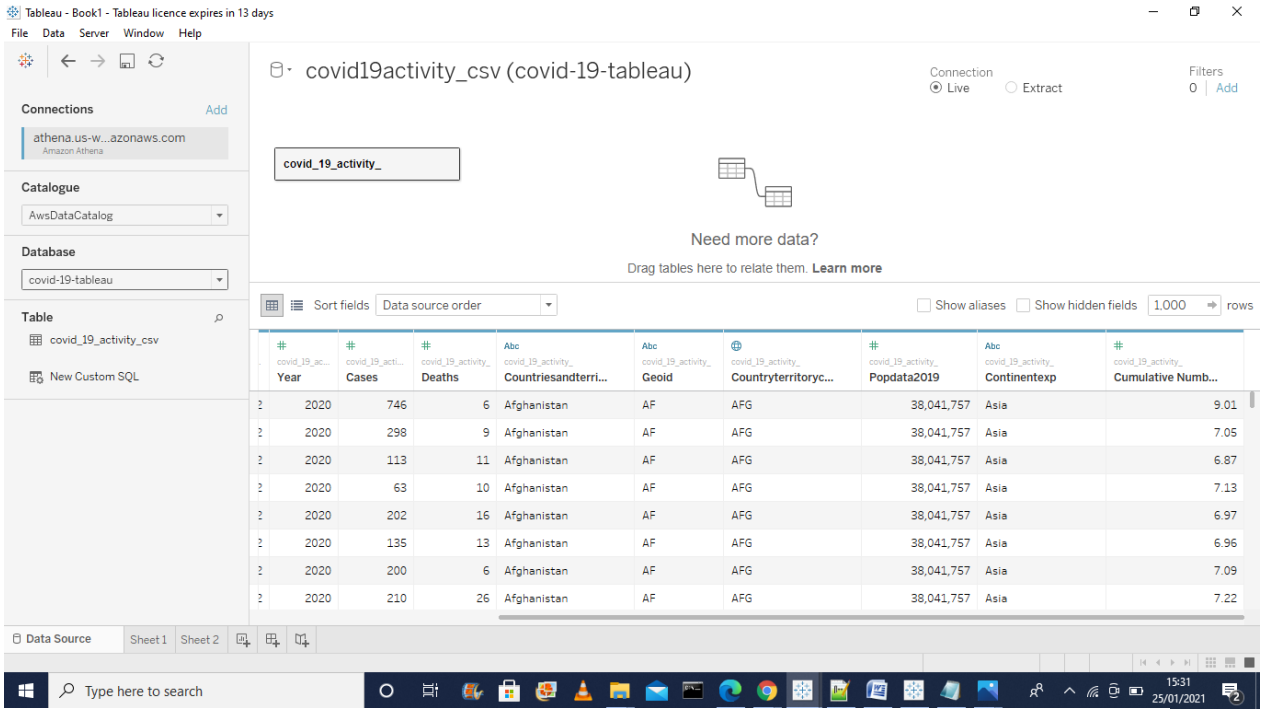
- Tables created in the Anthe will populate automatically as show in the below figure



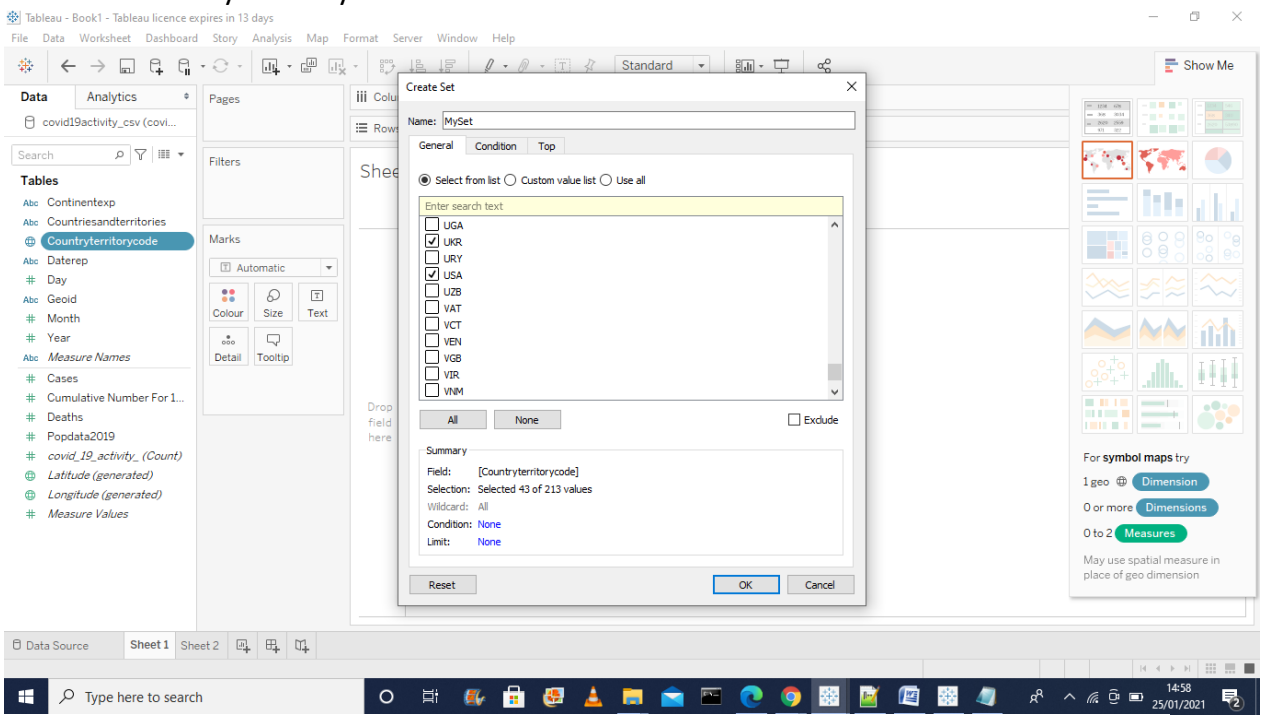
- Drag and drop the table covid19activity_csv



- Click on Update Now to load data in to the Tableau



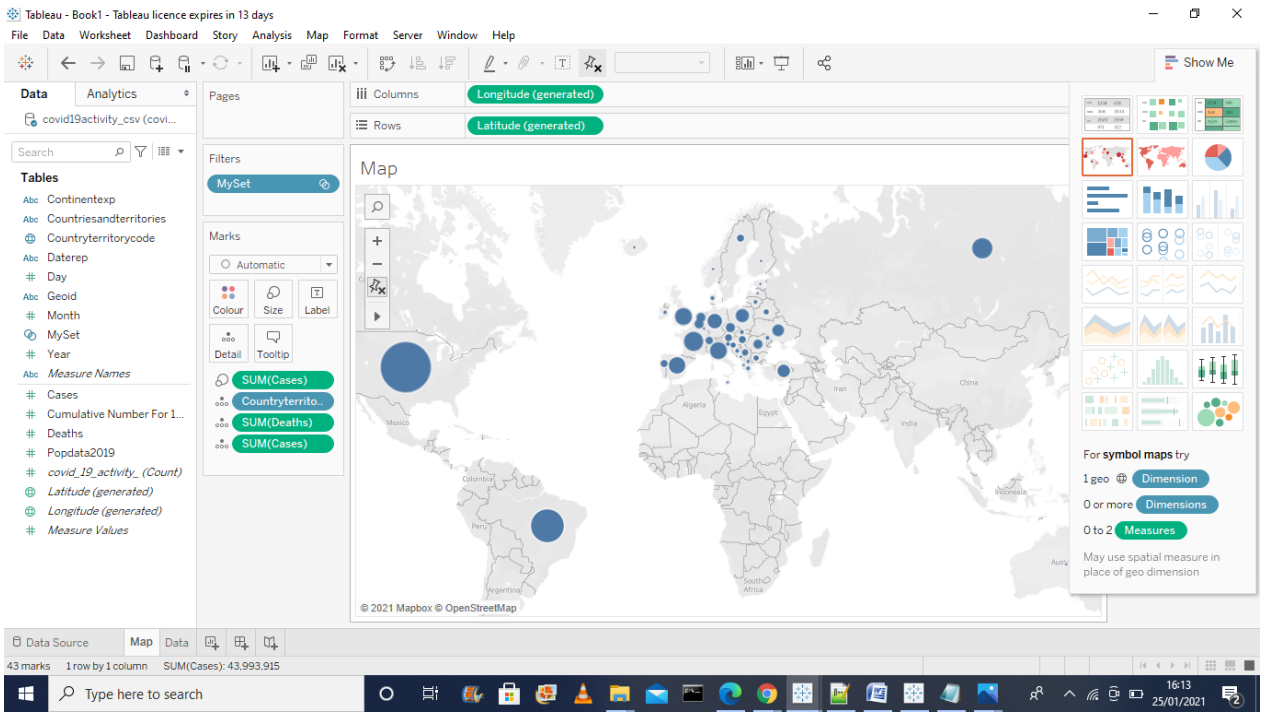
- Create Set counryTerritorycode as shown below for the codes



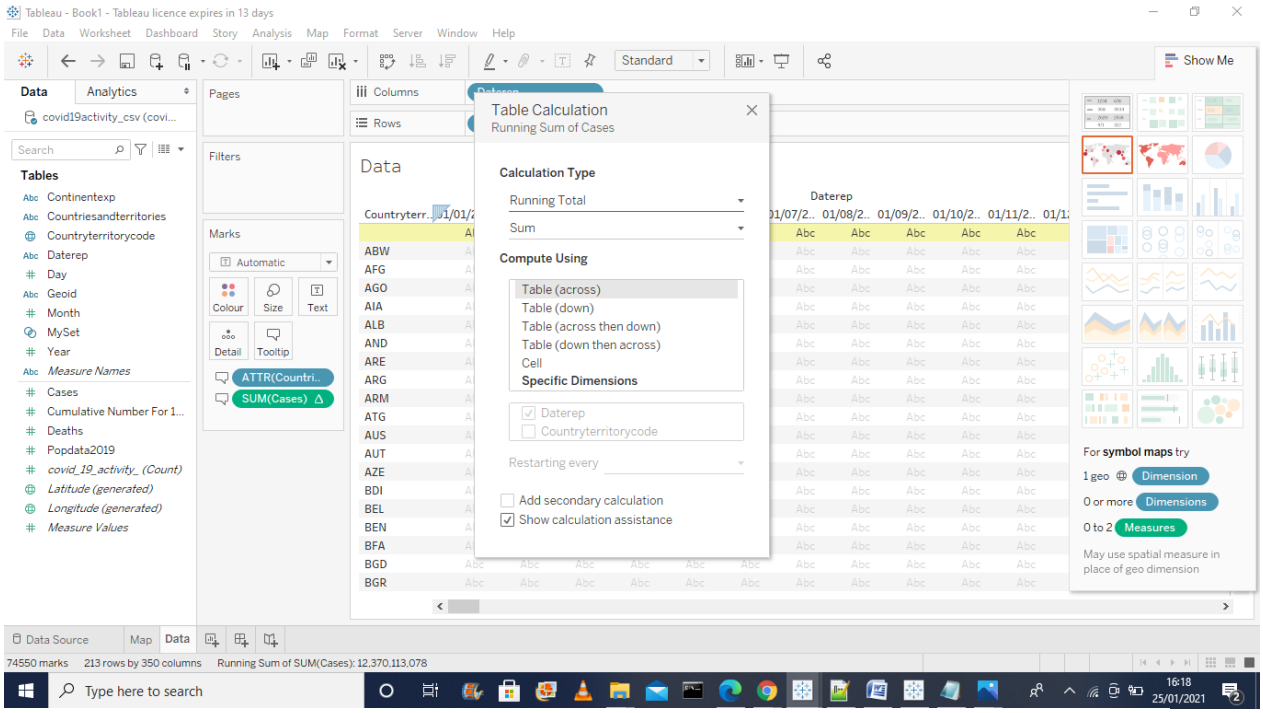
- Myset created for the below countryterritorycode code ALB,AUS,AUT,BEL,BGR,BIH,BLR,BRA,CHE,CZE,DEU,DNK,ESP,EST,FIN,FRA,GBR,GRC,HRV,

HUN,IRL,ISL,ITA,LTU,LUX,LVA,MDA,MKD,MLT,MNE,NLD,NOR,POL,PRT,ROU,RUS,SRB,SVK, SVN,SWE,TUR,UKR,USA

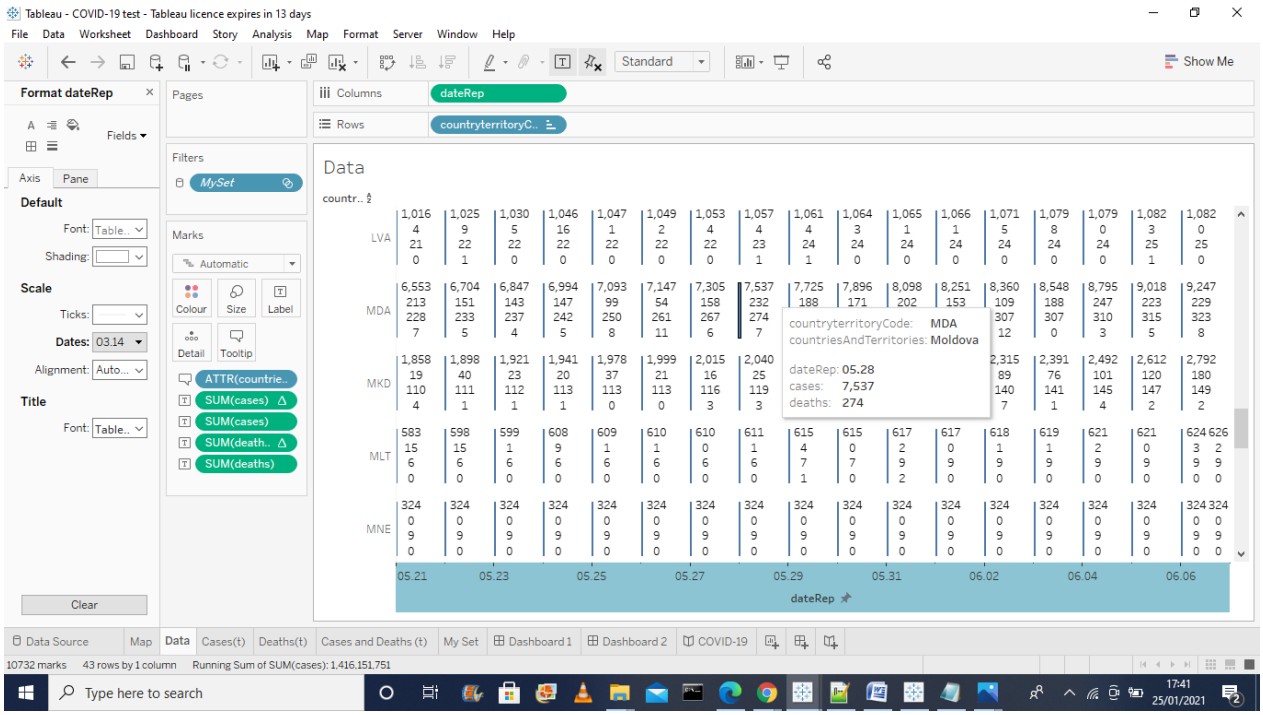
Created Map Dashboard as shown below

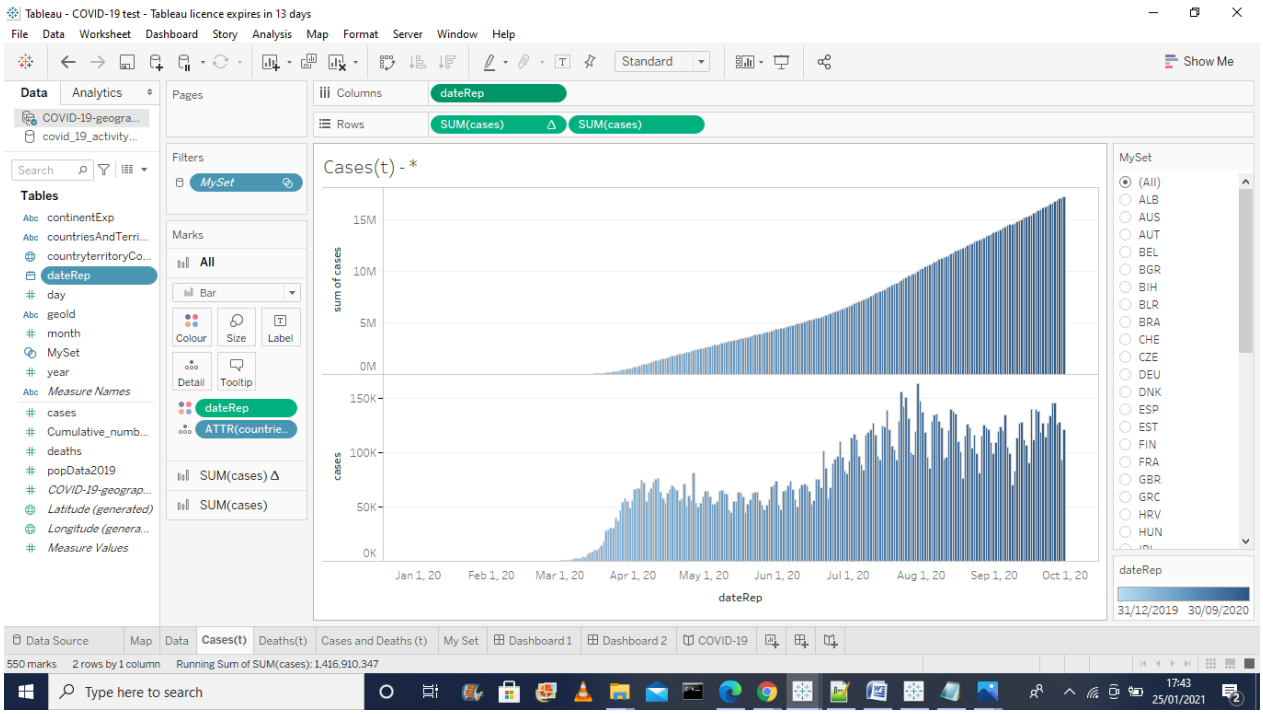
- 

- Running total dashboard creation

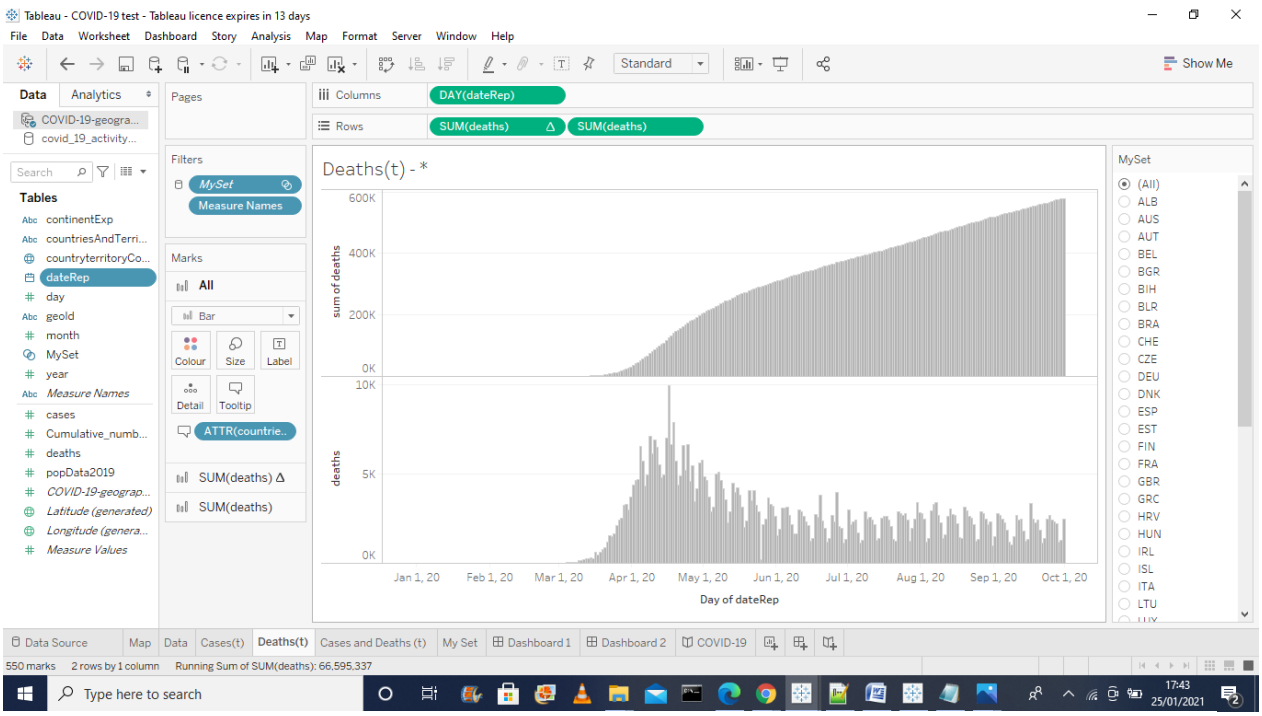


- Data Dashboard creation

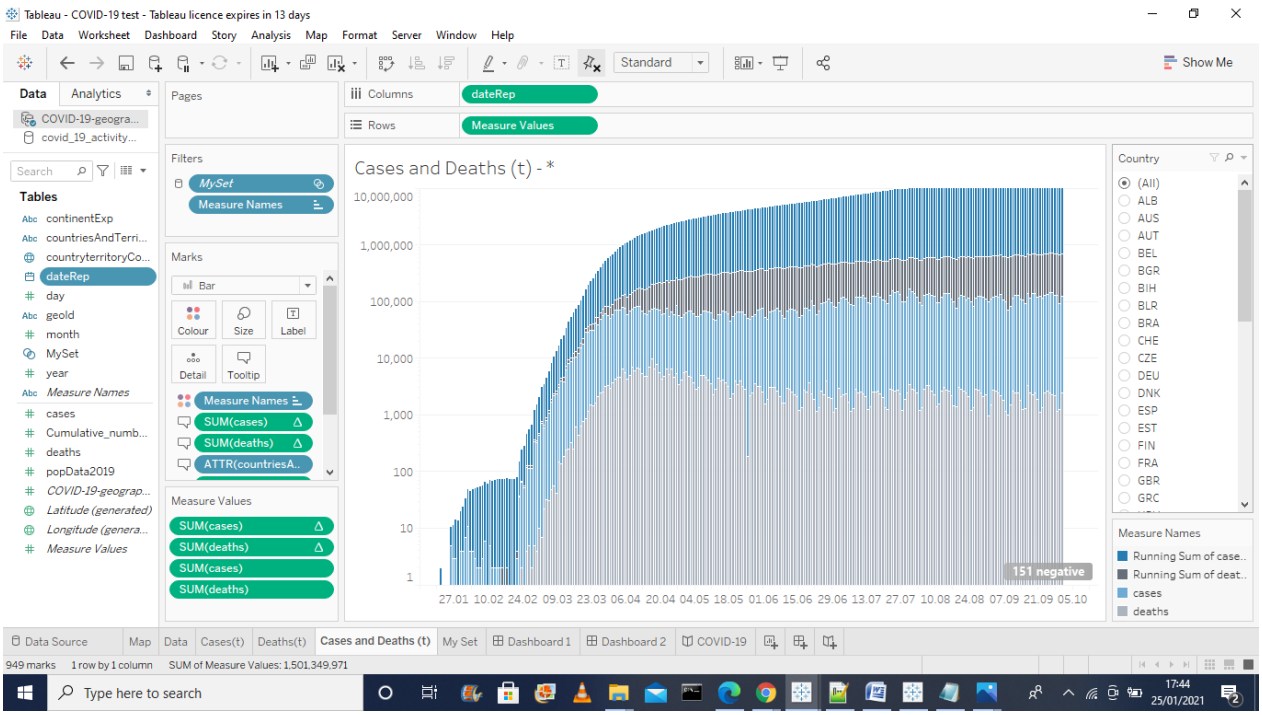


- Cases Dashboard alone
-


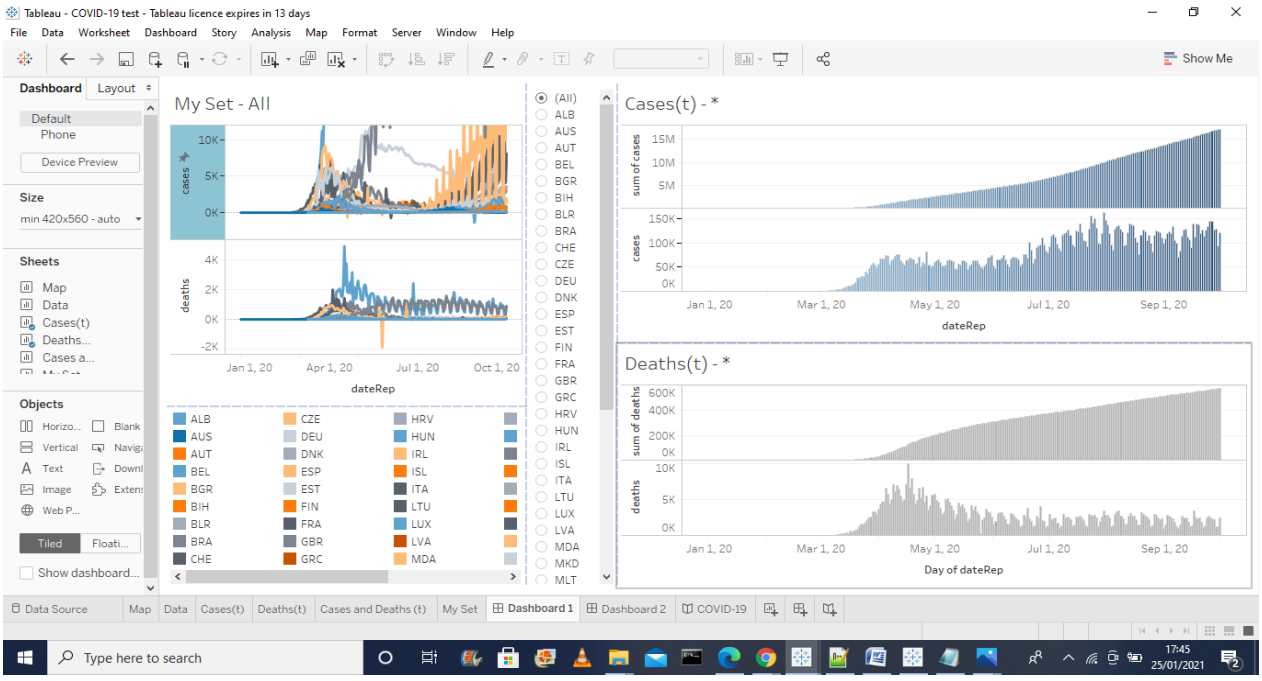
-
- Death Dashboard



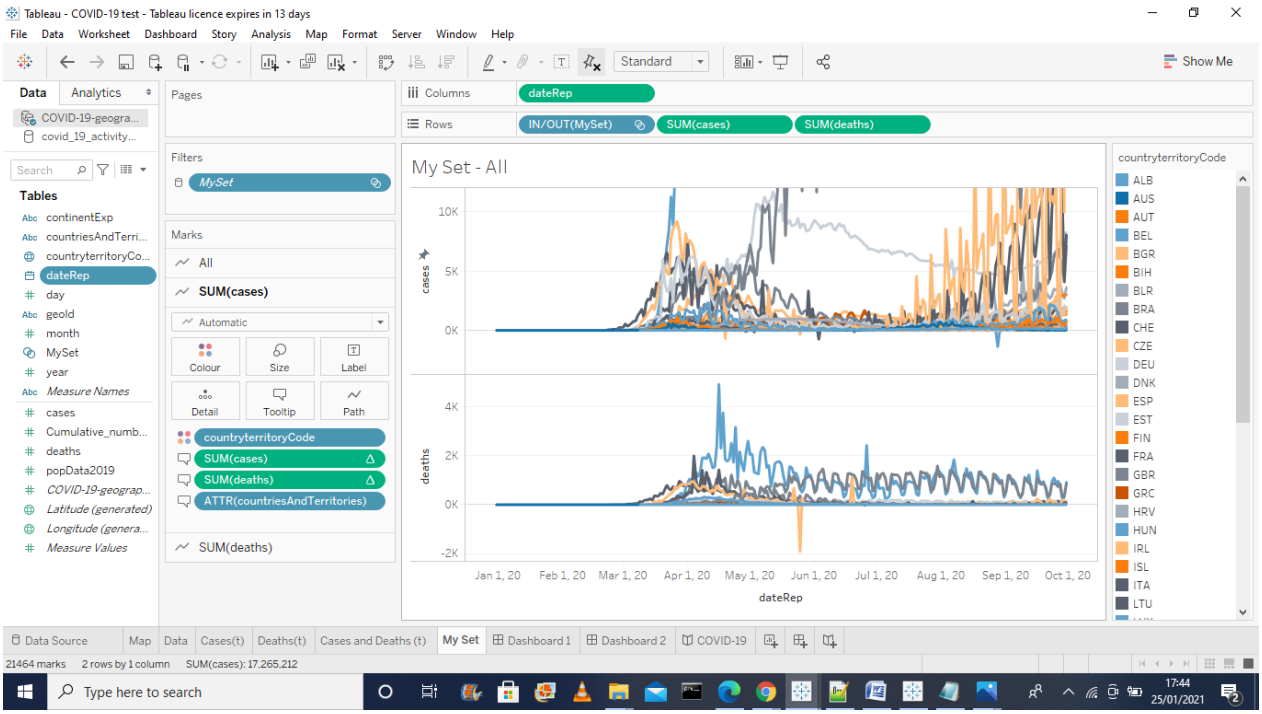
Cases & Death Dashboard



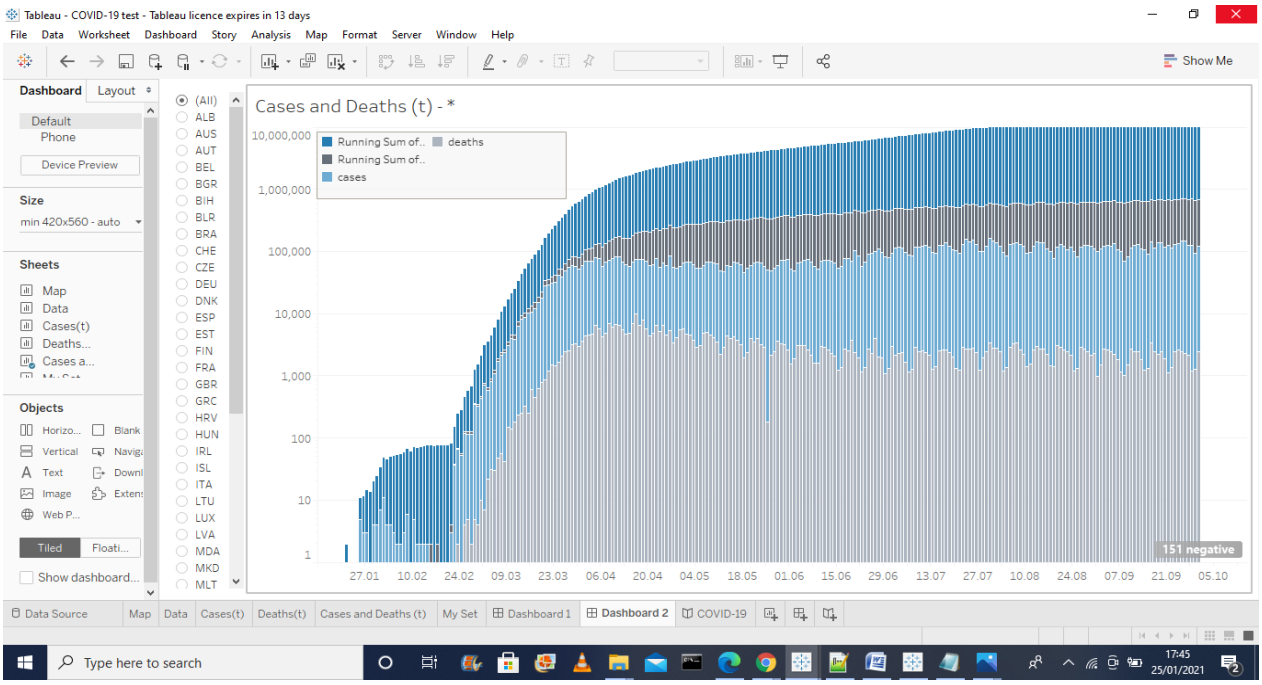
Consolidated Dashboard -1



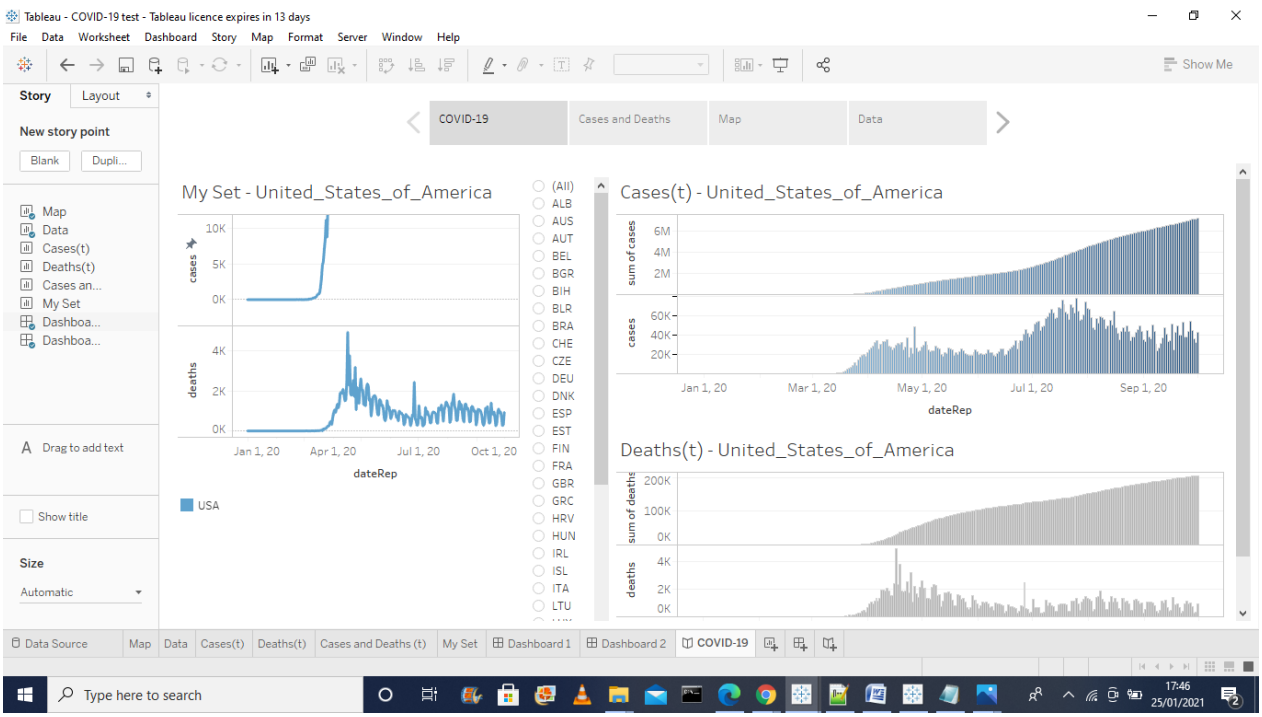
My Set Dashboard:



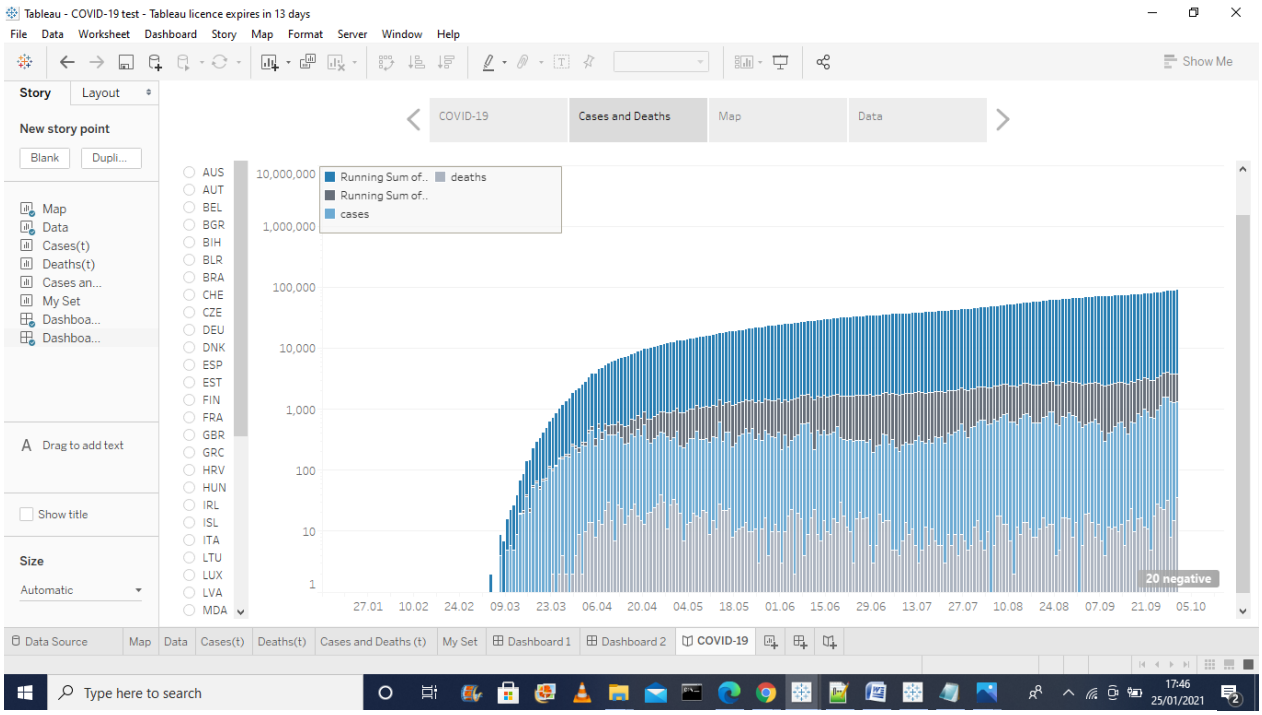
Cases & Deaths Dashboard2



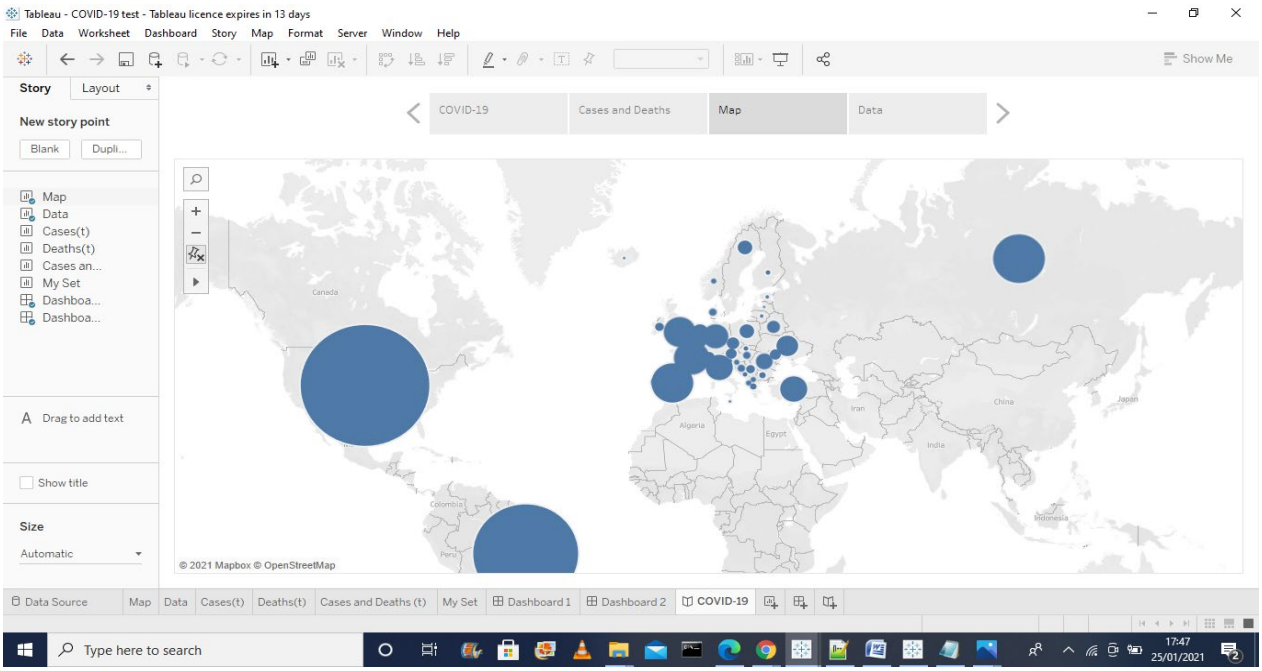
Covid-19 Dashboard for US:



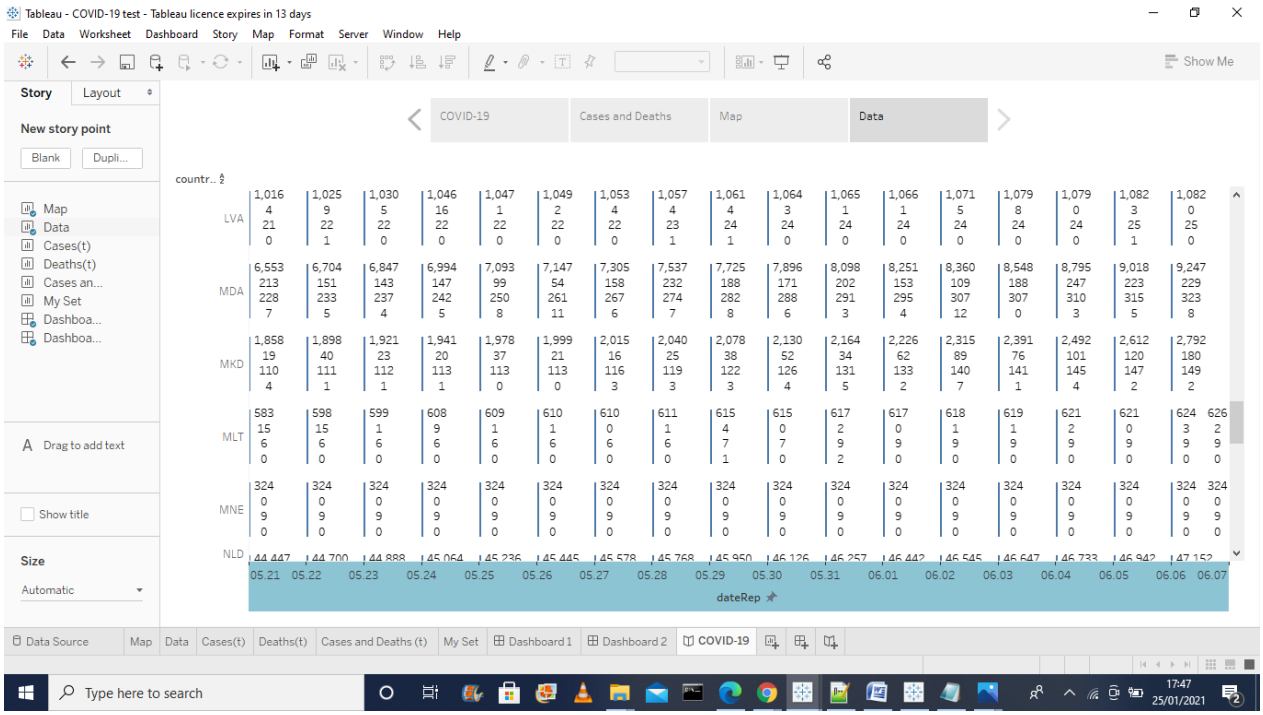
Covid-19 Dashboard for Cases & Deaths



Covid-19 Map Dashboard in a single view:



Covid-19 Data Dashboard in a Single view:



Role of Cloud:

Analyzing and dashboarding COVID-19 datasets using AWS Data Exchange, Amazon Glue, Amazon Glue Data Catalog, Amazon Athena, and Tableau. This approach is suggested for customers who are already using Amazon Athena or have just started using AWS and would like to build a dashboard using Tableau. This portal has to be available with minimal downtime and what better so hosting the portal online will have better availability. Moreover, as a part of AWS initiative consolidated Data sets is available using AWS Data Exchange service which can be utilized to project/predict real time graphs and analysi.

COVID-19 public data has been highly fragmented across different sources and difficult to make actionable. It's hard to know what data is reliable and most useful to aid the challenge of opening up our economy and returning to normal. Organizations across industries need a faster way to unlock data across applications and systems.To help everyone visualize COVID-19 data confidently and responsibly, Amazon Web Services (AWS) and its partners Salesforce, Tableau, and MuleSoft to create a centralized repository of trusted data from open source COVID-19 data providers.The AWS Data Exchanges makes it easy to find, subscribe to, and use third-party data in the cloud. We can subscribe to COVID-19 activity data and more for free at the Coronavirus (COVID-19) Data Hub listing in the AWS Data Exchange. By using various services of AWS like AWS Data Exchange, Amazon CloudWatch,Amazon CloudWatch Events,AWS Lambda,Amazon S3 ,AWS Glue,AWS Glue Data Catalog, Amazon Athena we can work with large covid-19 data set more flexible and make ease to build analytics by using Tableau Desktop, Tableau Public. Wihtout AWS cloud marketplace it is very difficult to get all different type of sources of data at one place to build Analytics.

Assumptions & Risks:

- Relying on the data which is made available by the AWS.
- Data consolidation for COVID-19 is a tedious task and we must rely on the various data providers to pull the data. Authenticity of the data can be a risk.

Lessons Learned & Observations:

- Zero records returned on query against table created by crawler. we need to select one level above the actual dataset in crawler s3 path selection.

  We just created a table by selecting the "Automatically (AWS Glue crawler)" create table option in Athena. My database and table show up in the pane on the left of the Athena query screen, and the table details page in glue shows a record count of 2416664. However, when I try to query the table in Athena (even if I select the "Preview table" option), the Results panel shows the column headers but under that it says "Zero rows returned".
- if any data type is date format in data source crawler identified as string ,corresponding graphs on top string not properly rendering.
- While copying JDBC Tableau athena driver need to restart the system.
- For Date datatype athena table creating as string datatype by crawler if we change the data type to Date corresponding athena table result for that column data is null instead of original content.
- If possible we thought to use Tableau Public to plot graphs as tableau public is not connecting athena directly we used Tableau Desktop.
- For graphs building we also considered various data sources not only the covid-19-datasoruce.
- dataexchangesdk bith boto3 implementation understand with python programming.
- Few of the IAM roles we missed out initial while implementation time we corrected it.
- Antena working if the folder having only one data file.
- Glue not identifying few data files like hyper,xls etc, if we process the dataset schema is not creating(table is not creating).
- String to date format conversion not working properly. May be we need to use views to convert string to date.
- We relied on the data provided by tableau for cleansing/cleaning/indigestion.
- Tableau is the licensed one we used 14 days trail for our simulation
- Tableau Public, a free software from Tableau that allows anyone to create interactive data visualization for the web. tableau not having connectors to connect anthena.

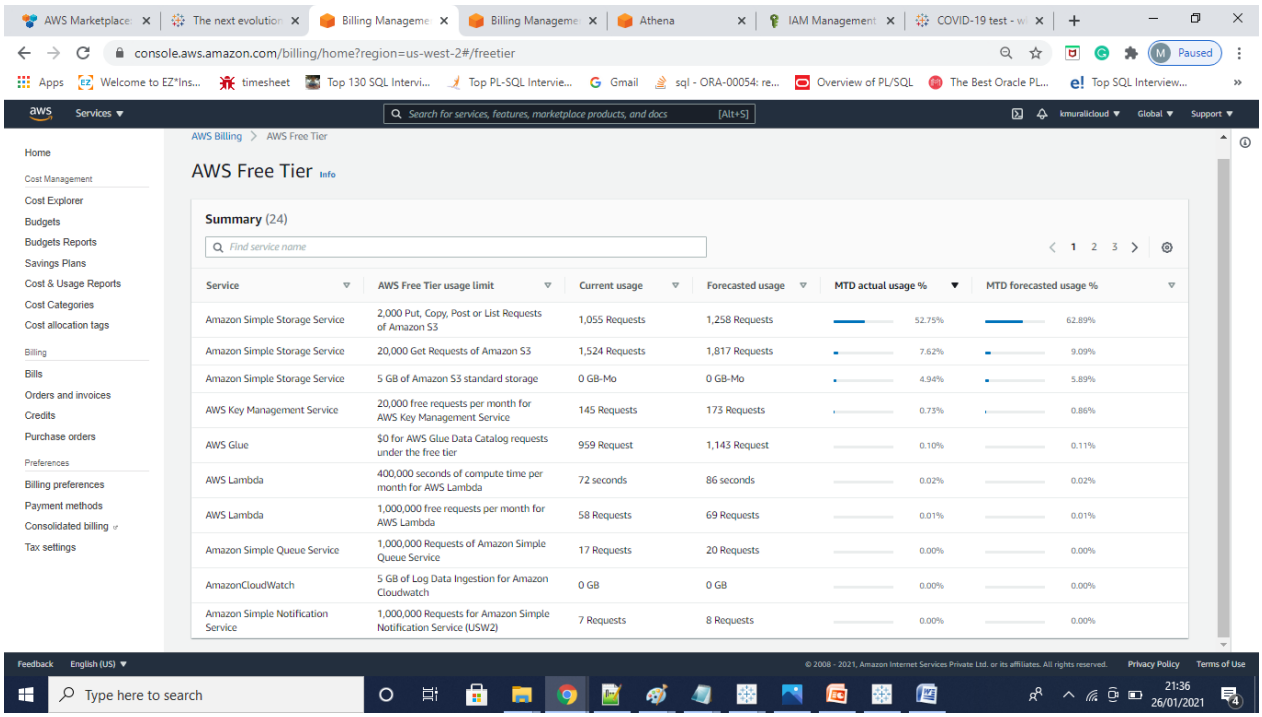
Effort & Cost Analysis:

- **Tableau Desktop**, We used 14 days trail period to plot the Dashboards
- **AWS Data Exchange**
  https://aws.amazon.com/marketplace/pp/prodview-a5mqede4xd4c4?qid=1589562921653&sr=0-5&ref_=srh_res_product_title#overview
  $0 for 36 months
- **AWS Glue**
  Region: Oregon (us-west-2)
  $0.27
- **Amazon Athena**
  Region: Oregon (us-west-2)
  $0
- **Amazon Simple Storage Service (S3)**
  Region: Oregon (us-west-2)
  Used free tier
  $0
- **AWS Lambda**
  Region: Oregon (us-west-2)
  Lambda Function - Include Free Tier
  $0
- **Tableau Public**
  Tableau Public Desktop is a free software
  $0
- **Total Cost**
  $0.32 =23.58 INR

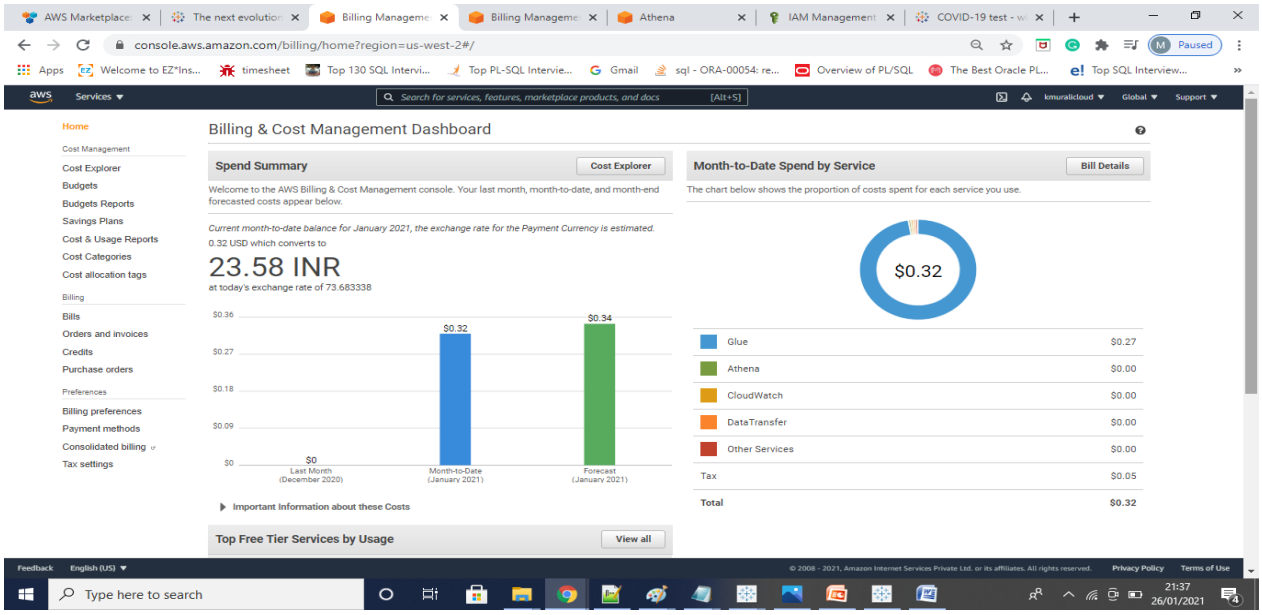
  We estimated $7.19/monthly before project implementation , (7.19/30 days*2=$0.47), for the two days implementation the total cost for all the services is around $0.32

  Estimated → $0.47
  Actual → $0.32

Consumption of AWS Free Tier usage for this project:



Billing & Cost Management Dashboard for this project implementation



Conclusion:

1. **Real world impact** - Provide real time analysis of the COVID-19 Data in order to make mission critical decisions.
2. **Any innovation** in the approach for implementing the project - AWS Data Exchange service is not being covered in our curriculum so with this project we will gain working knowledge of AWS Data Exchange also. Also, Tableau is market leader in visualization; we are trying to showcase Tableau also.
3. **Used Components (TAGS)**: Amazon Athena, Amazon CloudWatch, Amazon CloudWatch Events, Amazon S3, Analytics, AWS Data Exchange, AWS Glue, AWS Glue Data Catalog, AWS Lambda, Tableau, Tableau Public, Lambda functions are the primary skills which we will be achieving using this project.

Team Roles and Responsibilities:

Anusha:

- Involved in the Architecture creation for cloud watch events, lambda, AWS Data Exchange and S3 bucket.
- Cloud Watch configuration to trigger the event for the data pull.
- Worked on the Lambda function to be used to pull the revision data from the AWS Data Exchange to S3 by using AWS Python SDK with boto3
- Cloud formation stack creation. Worked on the yaml file creation
- Creation of the S3 bucket by using automated fashion.
- Involved in the documentation.

Murali :

- Proposed the Architecture for the problem statement by leveraging the various architectural references provided by AWS Documentation.
- Designed the correct set of services required for the project and also did various services integrations.
- Configured the Tableau for data visualization.
- Created various Tableau Dashboards.
- Worked on the Tableau Amazon Athena connector Integrations.
- Worked on the project cost estimation.
- Worked on the major part of solution design documentation.

Piysuh:

- Involved in the Architecture creation for Aws Glue Data Catalog, AWS Glue, Aws Athena.
- Defining Crawler
- Worked on the Various Athena queries.
- Created required Databases for the AWS Glue.
- Involved in the Project cost estimations.
- Analyzed the data source data properties to create glue schema definition.

Puneet:

- Proposed the capstone project idea.
- Analyzed the various data sets required for the project implementation.
- Creation of various IAM roles required for the project.
- Involved in the Architecture creation for Aws Data Exchange, Covid-19 Data platform.
- Involved in Lambda function implementation with dataexchangesdk with anusha.
- Involved in few tableau dashboard design with murali.

Acknowledgement:

The success and final outcome of this project required a lot of guidance and assistance from many people and we are extremely privileged to have got this all along with the completion of our project.

All that we have done is only due to such supervision and assistance from **Nirmallya Mukherjee** Lecturers and All other cloud mentors who shared their knowledge to make this project successfully and we would not forget to thank them. Thanks, GreatLearning for providing such a great opportunity to learn and became cloud masters --From PGPCCJULY_2020_Team-8

Special Thanks:

We would like to recognize **M. Sai Kalyana Chakravarthy, Tech Lead, Infosys, Chennai**, for the invaluable assistance that provided in Tableau dashboard implementation.