

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:

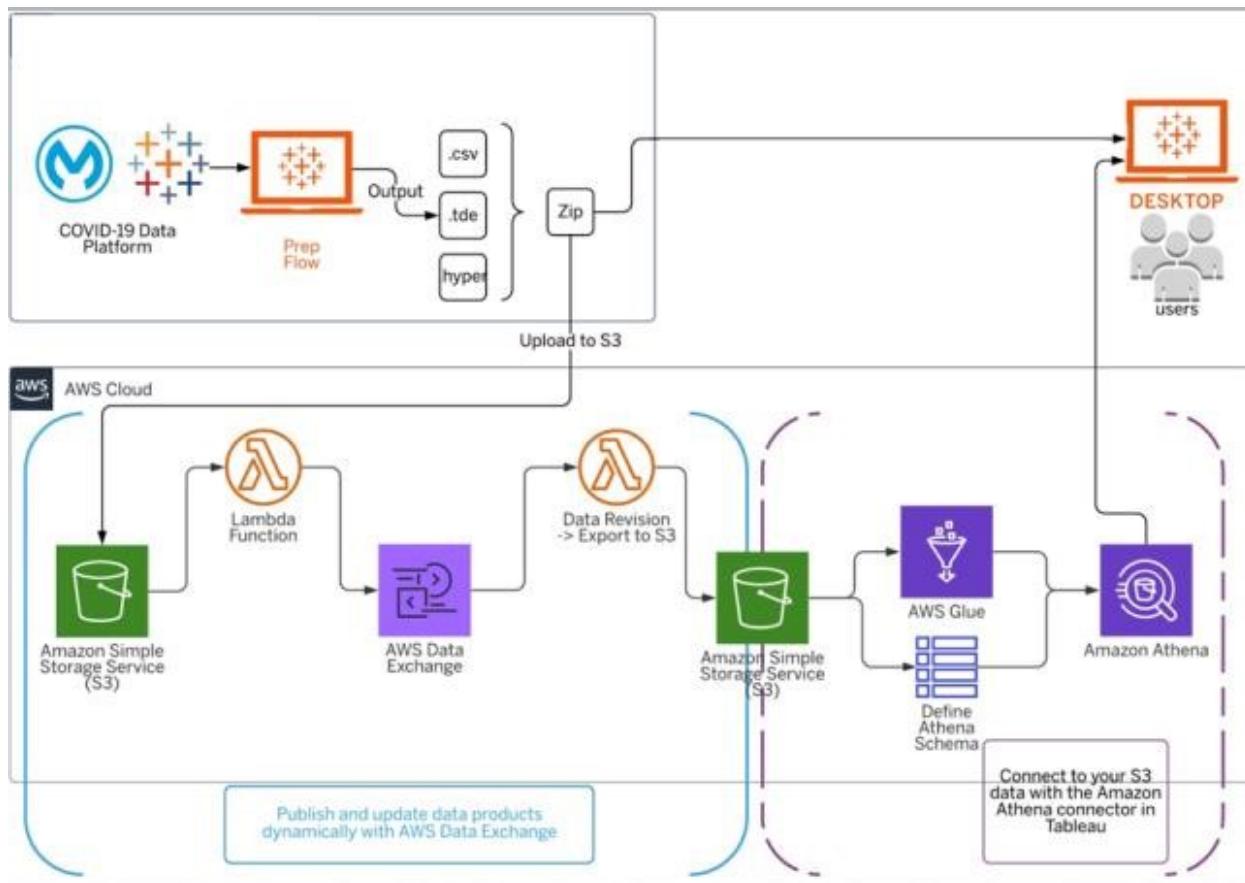
This project creates a centralized repository of trusted data from open source COVID-19 data providers using Amazon Web Services (AWS). Because of the ongoing COVID-19 pandemic our daily lives, people in every sector and country are accessing the data to stay informed and take appropriate measures in 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

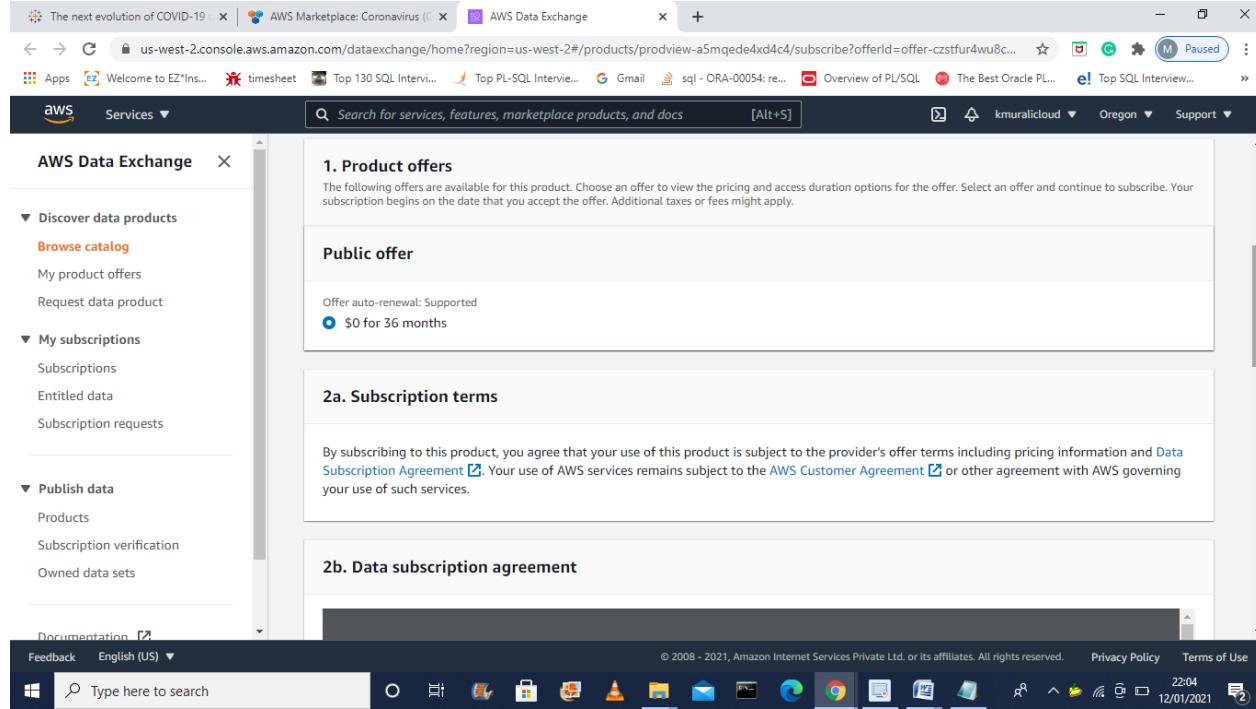
Click on Continue to Subscribe

The screenshot shows the AWS Marketplace interface. At the top, there are three tabs: 'The next evolution of COVID-19', 'AWS Marketplace: Coronavirus', and 'AWS Data Exchange'. The main content area is titled 'aws marketplace' and shows a product card for 'Coronavirus (COVID-19) Data Hub' provided by Tableau. The card includes a logo, a brief description, and an orange 'Continue to subscribe' button. Below the card, there are tabs for 'Product offers', 'Overview', 'Usage', and 'Support'. The 'Product offers' tab is selected, showing a 'Public offer' section with an auto-renewal option for \$0 for 36 months.

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

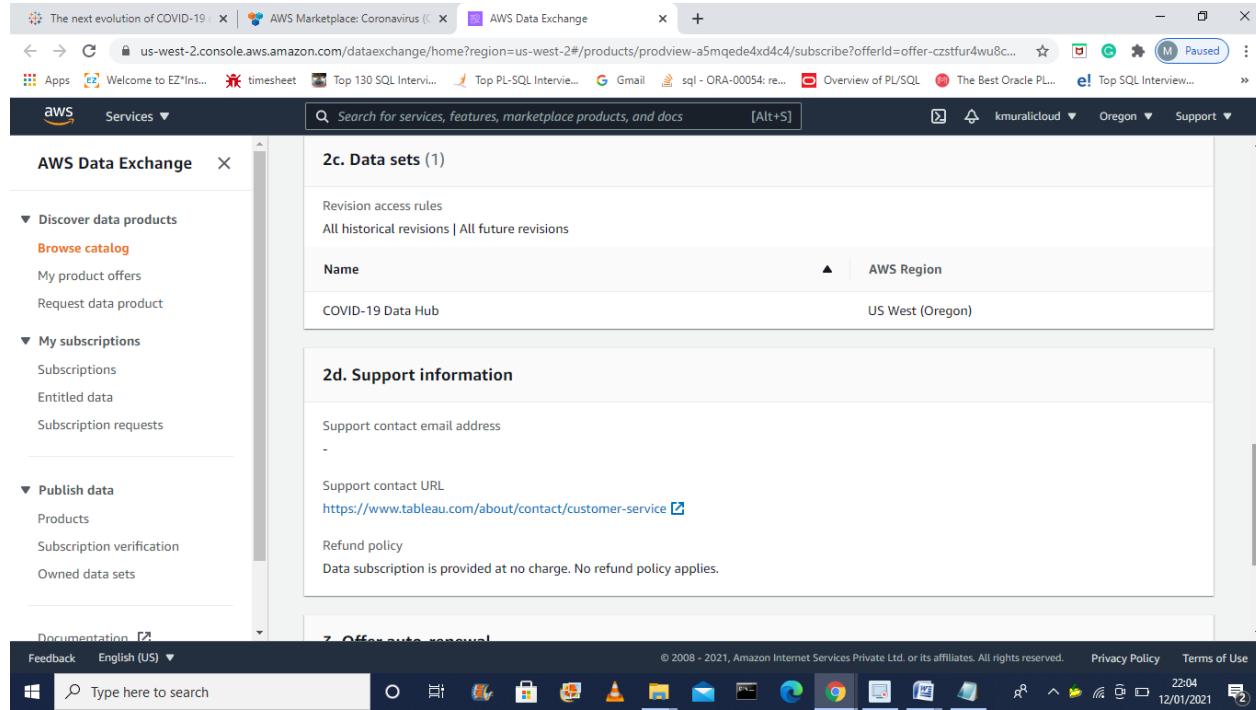
The screenshot shows the AWS Data Exchange service interface. At the top, there are three tabs: 'The next evolution of COVID-19', 'AWS Marketplace: Coronavirus', and 'AWS Data Exchange'. The main content area is titled 'AWS Data Exchange' and shows a product card for 'Coronavirus (COVID-19) Data Hub' provided by Tableau. The card includes a logo, a brief description, and a 'Free' status with a '36 month subscription available' note. Below the card, there is a 'Complete subscription' section with a pencil icon and a list of steps: 1. Choose a product offer and an offer option. 2. Review the subscription terms, Data Subscription Agreement (DSA), and refund policy. 3. Choose if you want to renew the offer when it expires. 4. Subscribe. The '1. Product offers' section is expanded, showing the same information as the previous screenshot.

The cost of Subscription is \$0 for 36 months



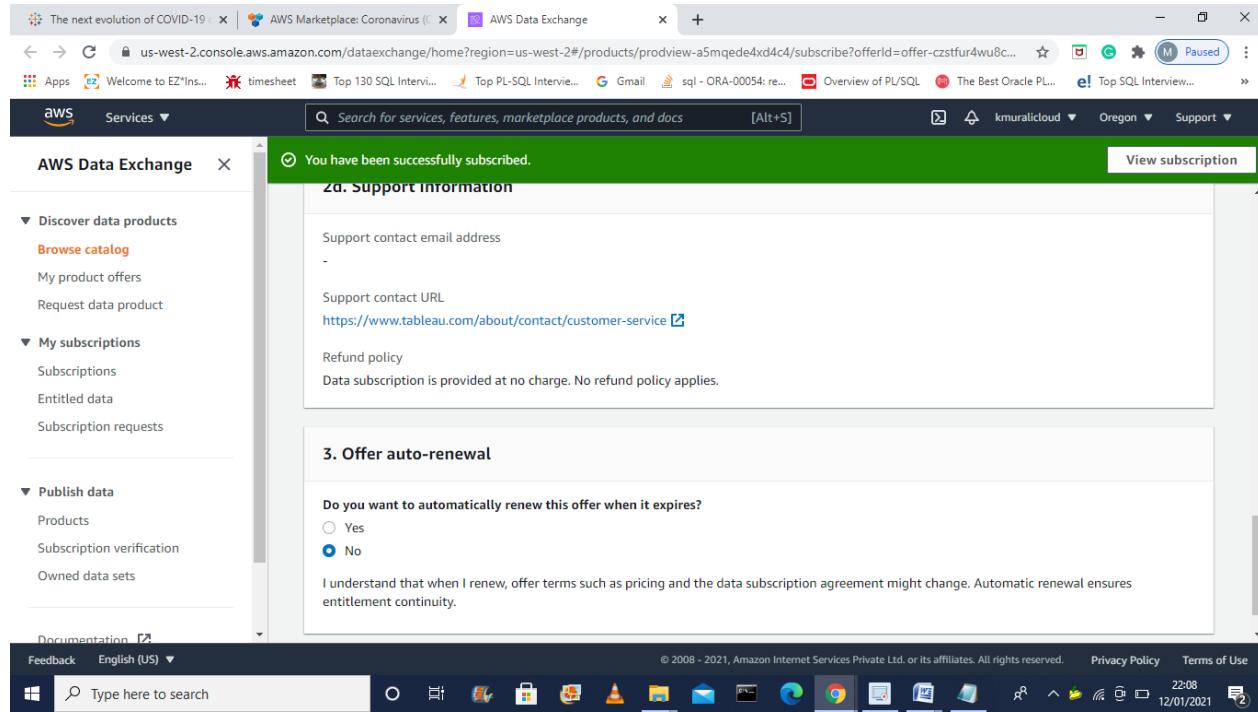
The screenshot shows the AWS Data Exchange interface. On the left, a sidebar lists categories like Discover data products, My subscriptions, and Publish data. The main content area displays a 'Public offer' for a product, highlighting a subscription duration of '\$0 for 36 months'. Below this, sections for 'Subscription terms' and 'Data subscription agreement' are visible. The bottom of the screen shows a Windows taskbar with various pinned icons.

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



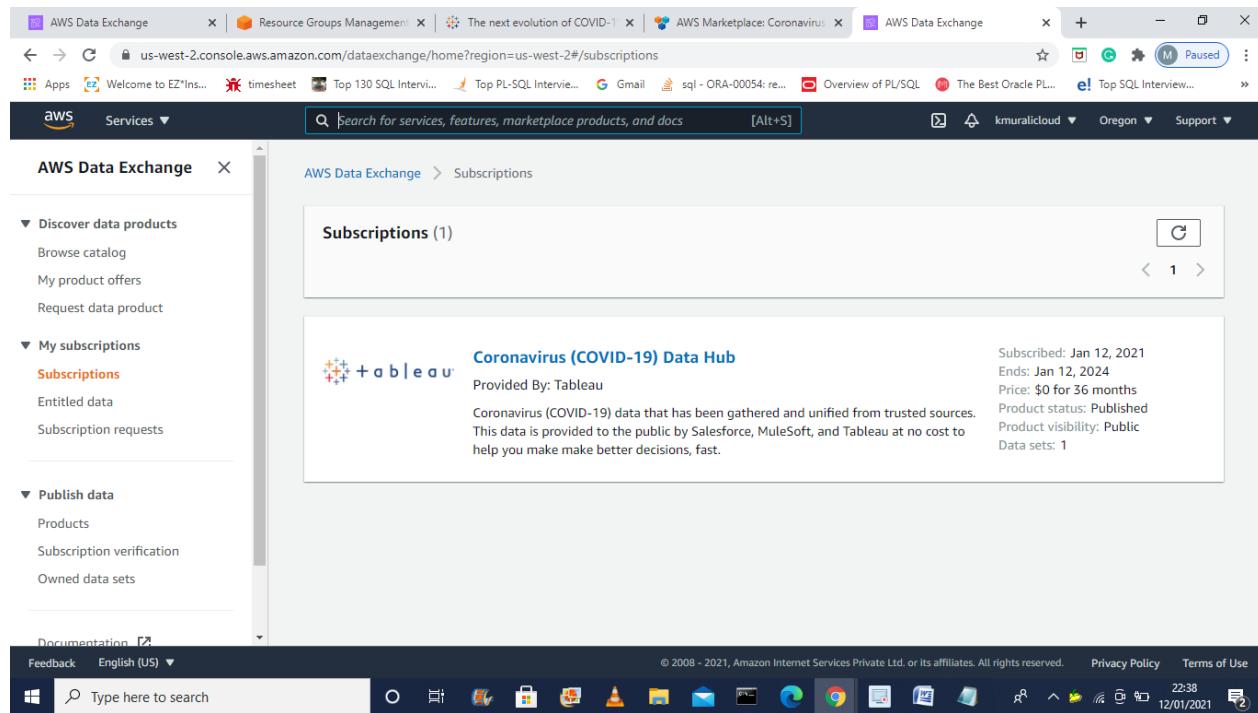
The screenshot shows the AWS Data Exchange interface, specifically the 'Data sets' section for the Covid-19 Data Hub. It lists revision access rules and shows a single dataset named 'COVID-19 Data Hub' located in the 'US West (Oregon)' region. The right side of the screen contains sections for 'Support information' and 'Offer terms and resources'. The bottom of the screen shows a Windows taskbar with various pinned icons.

Data set Subscribed Successfully



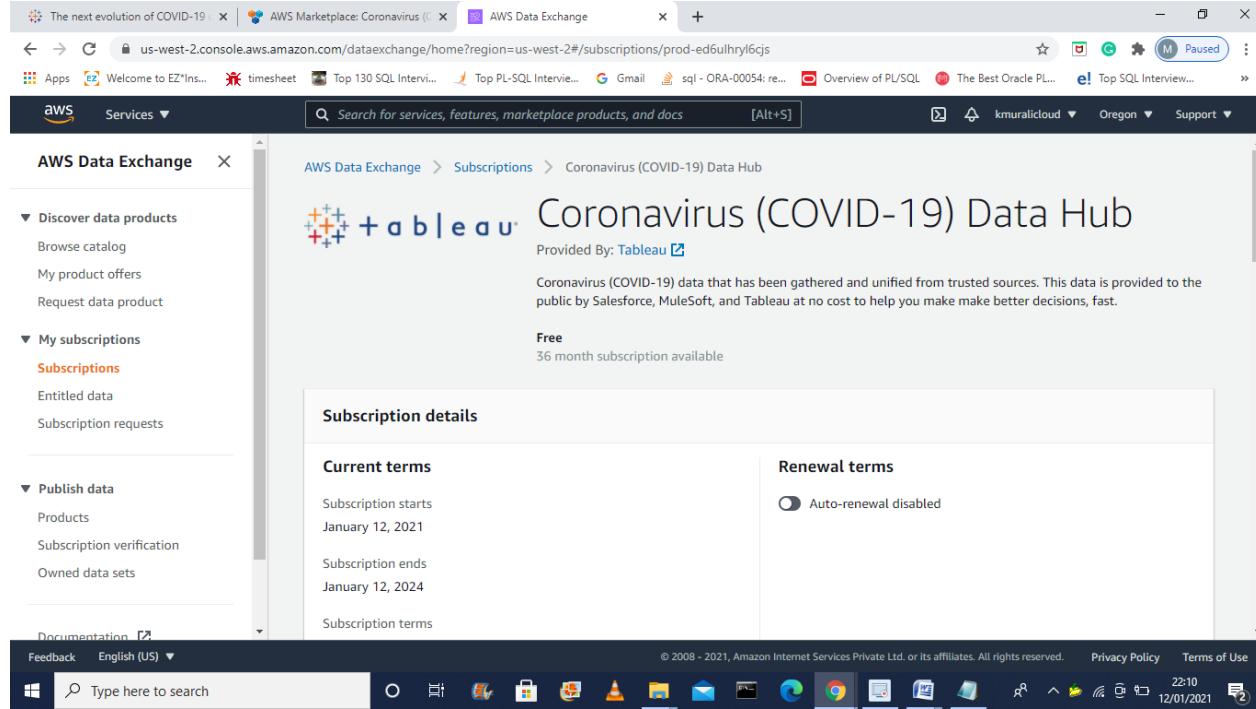
The screenshot shows a browser window with multiple tabs open. The active tab is titled "AWS Data Exchange". The main content area displays a green success message: "You have been successfully subscribed." Below this, there are two sections: "2d. Support information" and "3. Offer auto-renewal". In "Support information", it shows a support contact email address (redacted) and a URL (<https://www.tableau.com/about/contact/customer-service>). In "Offer auto-renewal", there is a question "Do you want to automatically renew this offer when it expires?" with two radio button options: "Yes" (unselected) and "No" (selected). A note below states: "I understand that when I renew, offer terms such as pricing and the data subscription agreement might change. Automatic renewal ensures entitlement continuity." The left sidebar of the AWS Data Exchange interface shows navigation links for "Discover data products", "My subscriptions", "Publish data", and "Documentation". The bottom of the screen shows a Windows taskbar with various pinned icons.

Go to My Subscriptions section:



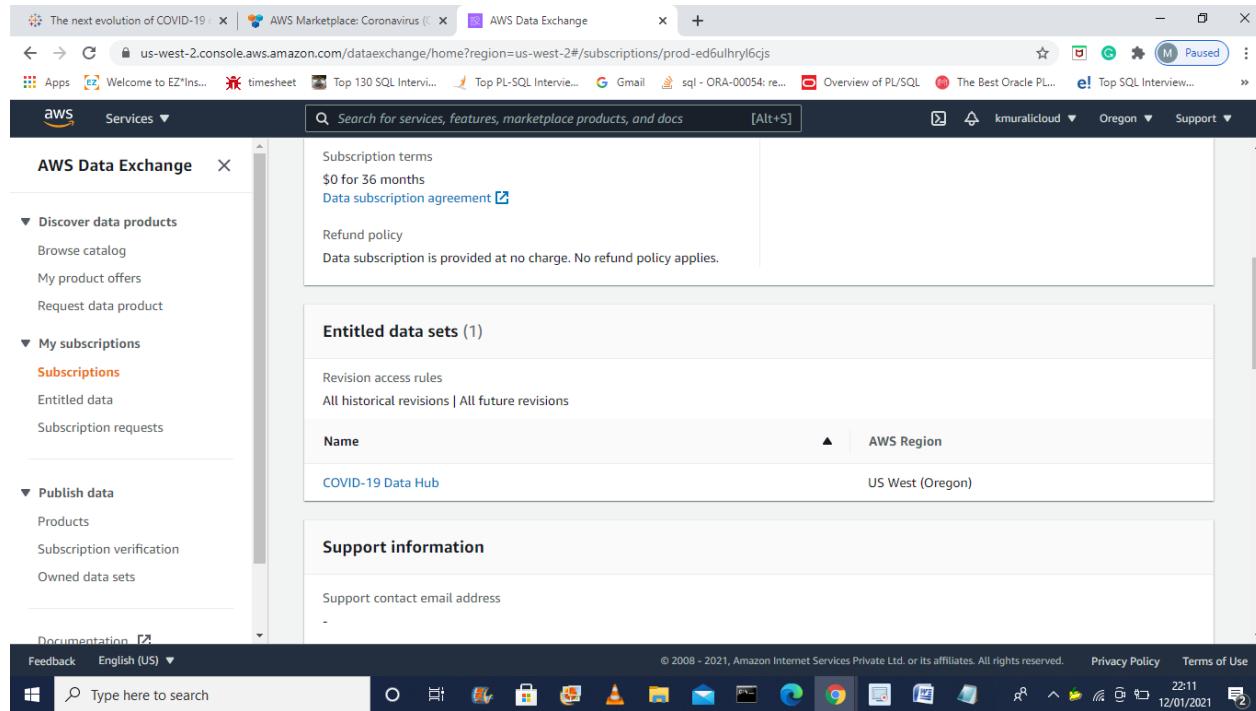
The screenshot shows a browser window with multiple tabs open. The active tab is titled "AWS Data Exchange". The main content area shows the "Subscriptions" section with a heading "Subscriptions (1)". Below this, there is a card for a subscription titled "Coronavirus (COVID-19) Data Hub". The card includes the provider "Provided By: Tableau", a description ("Coronavirus (COVID-19) data that has been gathered and unified from trusted sources. This data is provided to the public by Salesforce, MuleSoft, and Tableau at no cost to help you make better decisions, fast."), and details about the subscription: "Subscribed: Jan 12, 2021", "Ends: Jan 12, 2024", "Price: \$0 for 36 months", "Product status: Published", "Product visibility: Public", and "Data sets: 1". The left sidebar of the AWS Data Exchange interface shows navigation links for "Discover data products", "My subscriptions", "Publish data", and "Documentation". The bottom of the screen shows a Windows taskbar with various pinned icons.

Subscription Details:



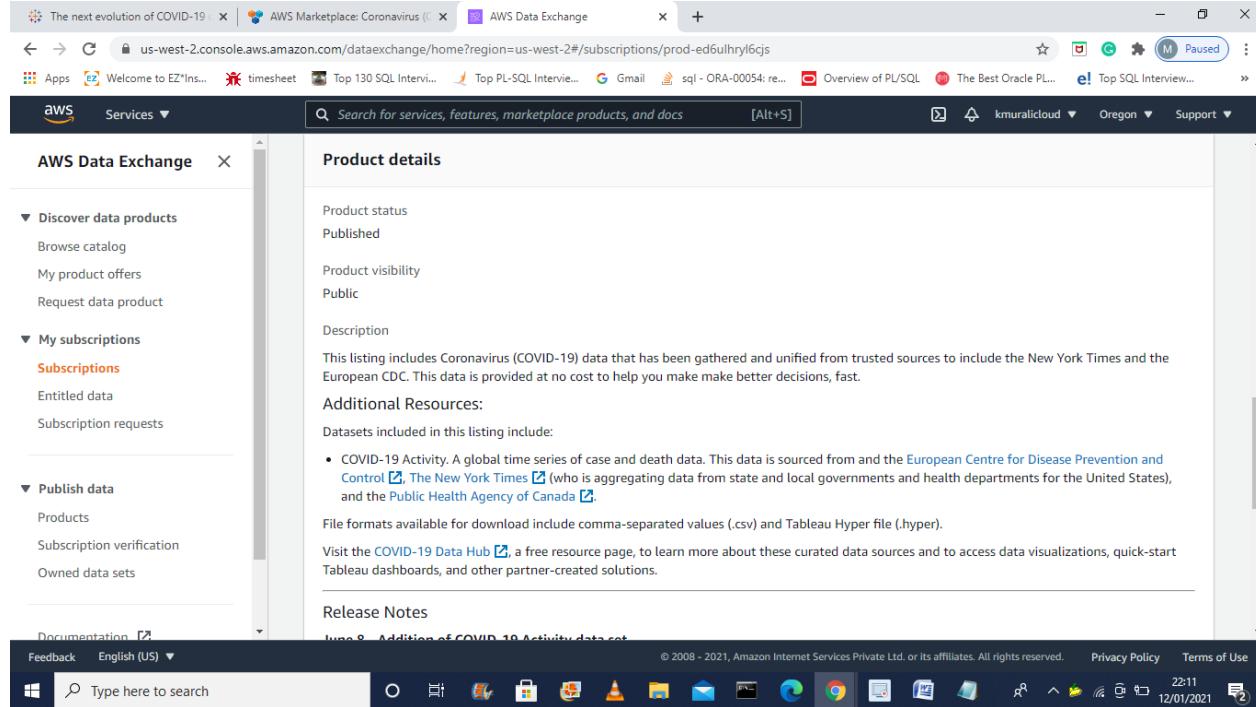
The screenshot shows the AWS Data Exchange interface. On the left, there's a sidebar with navigation links like 'Discover data products', 'My subscriptions', 'Subscriptions', 'Publish data', and 'Documentation'. The main content area displays the 'Coronavirus (COVID-19) Data Hub' by Tableau. It includes a brief description: 'Coronavirus (COVID-19) data that has been gathered and unified from trusted sources. This data is provided to the public by Salesforce, MuleSoft, and Tableau at no cost to help you make better decisions, fast.' Below this, it says 'Free' and '36 month subscription available'. A 'Subscription details' section is expanded, showing 'Current terms' (starts January 12, 2021, ends January 12, 2024) and 'Renewal terms' (auto-renewal disabled). At the bottom, there's a 'Subscription dataset' section with information about the dataset and its entitlements.

Subscription dataset:



This screenshot shows the same AWS Data Exchange interface as above, but focusing on the 'COVID-19 Data Hub' dataset. The 'Entitled data sets' section is expanded, showing 'Revision access rules' (All historical revisions | All future revisions) and a table for 'Name' (COVID-19 Data Hub) and 'AWS Region' (US West (Oregon)). The 'Support information' section is also visible at the bottom.

Subscribed product details:

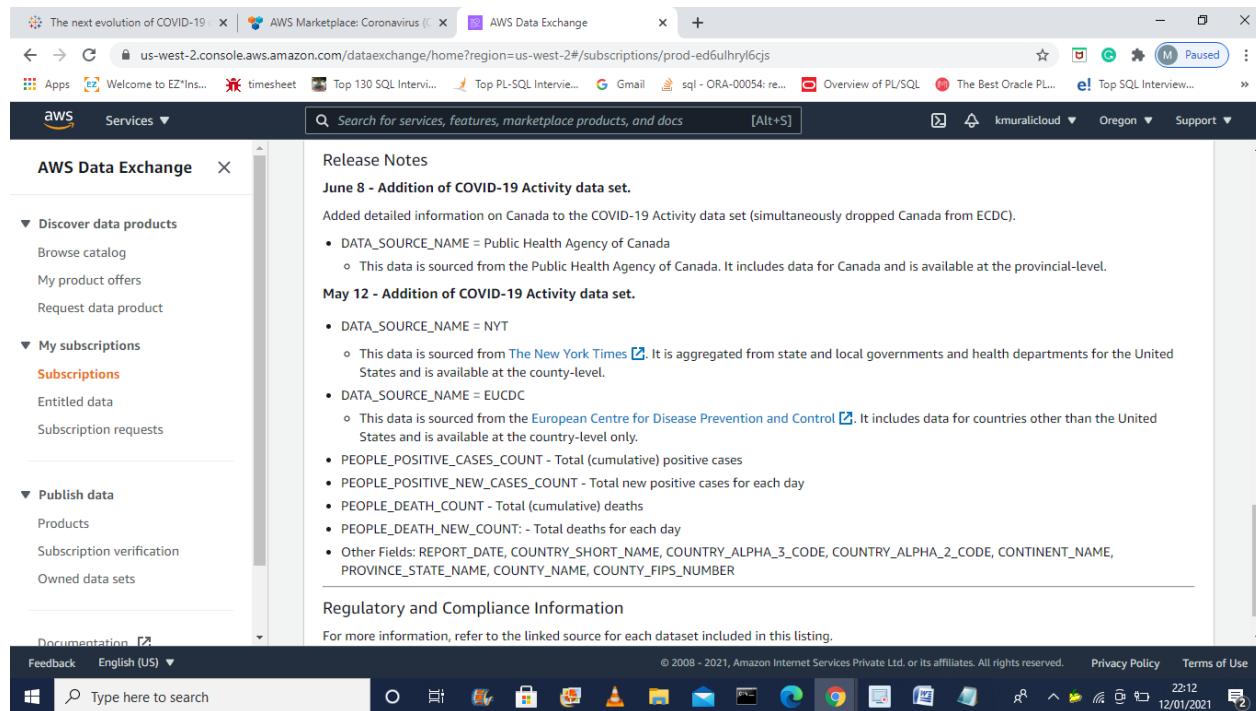


The screenshot shows the AWS Data Exchange product details page for the COVID-19 Activity dataset. The left sidebar lists categories like Discover data products, My subscriptions, and Publish data. The main content area displays the following information:

- Product details**
 - Product status: Published
 - Product visibility: Public
 - Description: This listing includes Coronavirus (COVID-19) data that has been gathered and unified from trusted sources to include the New York Times and the European CDC. This data is provided at no cost to help you make better decisions, fast.
 - Additional Resources:**
 - Datasets included in this listing include:
 - COVID-19 Activity: A global time series of case and death data. This data is sourced from and the European Centre for Disease Prevention and Control (who is aggregating data from state and local governments and health departments for the United States), and the Public Health Agency of Canada.
 - File formats available for download include comma-separated values (.csv) and Tableau Hyper file (.hyper).
 - Visit the COVID-19 Data Hub, a free resource page, to learn more about these curated data sources and to access data visualizations, quick-start Tableau dashboards, and other partner-created solutions.
- Release Notes**
 - June 8 - Addition of COVID-19 Activity data set

At the bottom, there's a footer with links to Feedback, English (US), Privacy Policy, Terms of Use, and a system tray showing the date and time (22:11, 12/01/2021).

Data set fields description



The screenshot shows the AWS Data Exchange product details page for the COVID-19 Activity dataset, specifically focusing on the Release Notes section. The left sidebar is identical to the previous screenshot. The main content area displays the following release notes:

- Release Notes**
 - June 8 - Addition of COVID-19 Activity data set.**
 - Added detailed information on Canada to the COVID-19 Activity data set (simultaneously dropped Canada from ECDC).
 - DATA_SOURCE_NAME = Public Health Agency of Canada
 - This data is sourced from the Public Health Agency of Canada. It includes data for Canada and is available at the provincial-level.
 - May 12 - Addition of COVID-19 Activity data set.**
 - DATA_SOURCE_NAME = NYT
 - This data is sourced from The New York Times. It is aggregated from state and local governments and health departments for the United States and is available at the county-level.
 - DATA_SOURCE_NAME = EUCDC
 - This data is sourced from the European Centre for Disease Prevention and Control. It includes data for countries other than the United States and is available at the country-level only.
 - PEOPLE_POSITIVE_CASES_COUNT - Total (cumulative) positive cases
 - PEOPLE_POSITIVE_NEW_CASES_COUNT - Total new positive cases for each day
 - PEOPLE_DEATH_COUNT - Total (cumulative) deaths
 - PEOPLE_DEATH_NEW_COUNT - Total deaths for each day
 - Other Fields: REPORT_DATE, COUNTRY_SHORT_NAME, COUNTRY_ALPHA_3_CODE, COUNTRY_ALPHA_2_CODE, CONTINENT_NAME, PROVINCE_STATE_NAME, COUNTY_NAME, COUNTY_FIPS_NUMBER
- Regulatory and Compliance Information**

For more information, refer to the linked source for each dataset included in this listing.

At the bottom, there's a footer with links to Feedback, English (US), Privacy Policy, Terms of Use, and a system tray showing the date and time (22:12, 12/01/2021).

Entitled data

The screenshot shows the AWS Data Exchange interface. On the left, a sidebar menu includes 'Discover data products', 'My subscriptions' (with 'Entitled data' selected), and 'Publish data'. The main content area is titled 'Entitled data' and shows a table for the 'Coronavirus (COVID-19) Data Hub'. The table includes columns for Product subscription, Time created, Revision access rules, Last updated, and Data set ID.

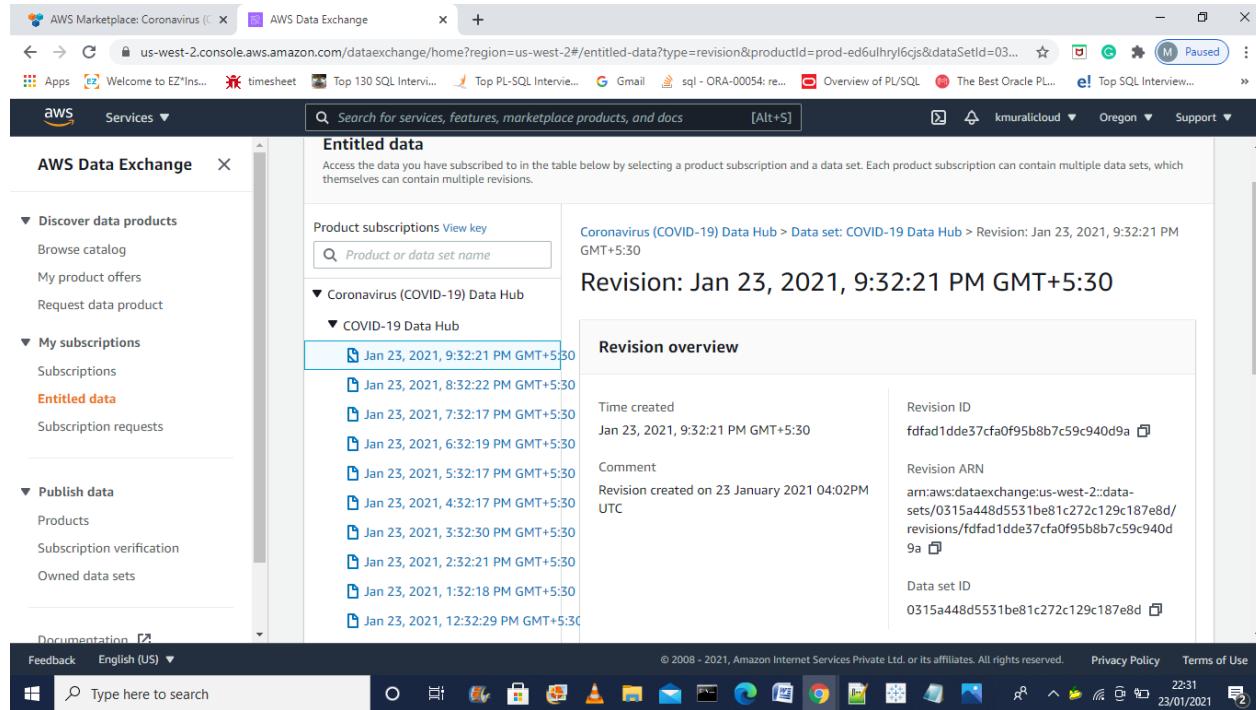
Data set overview	
Product subscription Coronavirus (COVID-19) Data Hub	Time created May 13, 2020, 10:07:28 AM GMT+5:30
Revision access rules All historical revisions All future revisions	Last updated an hour ago
Product ID prod-ed6ulhryl6cjs	Data set ID 0315a448d5531be81c272c129c187e8d

Entitled data , click on Latest Revision

The screenshot shows the same AWS Data Exchange interface as the previous one, but the main content area now displays a list of revisions for the 'COVID-19 Data Hub'. The revisions are listed in descending order of creation date, with the latest revision at the top.

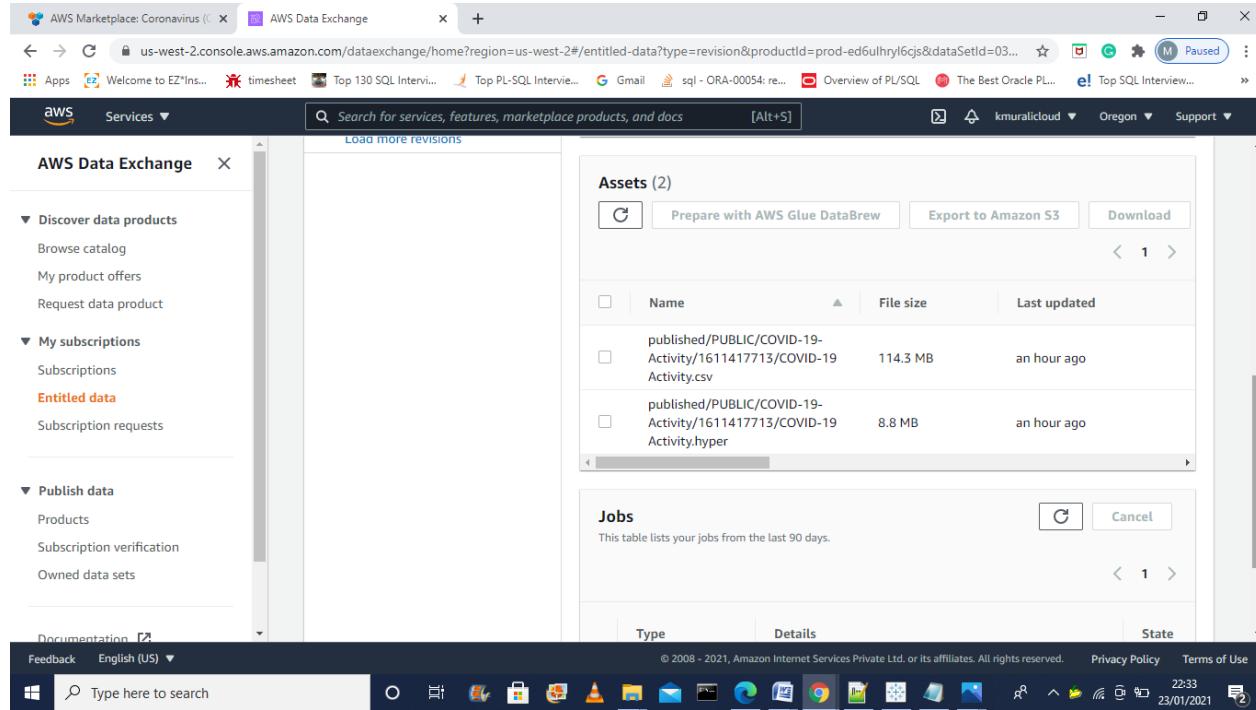
- Jan 23, 2021, 9:32:21 PM GMT+5:30
- Jan 23, 2021, 8:32:22 PM GMT+5:30
- Jan 23, 2021, 7:32:17 PM GMT+5:30
- Jan 23, 2021, 6:32:19 PM GMT+5:30
- Jan 23, 2021, 5:32:17 PM GMT+5:30
- Jan 23, 2021, 4:32:17 PM GMT+5:30
- Jan 23, 2021, 3:32:30 PM GMT+5:30
- Jan 23, 2021, 2:32:21 PM GMT+5:30

Latest Revision Overview



The screenshot shows the AWS Data Exchange service interface. On the left, there's a sidebar with navigation links like 'Discover data products', 'My subscriptions', and 'Publish data'. The main area is titled 'Entitled data' and shows a list of 'Product subscriptions'. A specific subscription for 'Coronavirus (COVID-19) Data Hub' is selected. Under this hub, the 'COVID-19 Data Hub' section is expanded, showing a list of revisions. The most recent revision is highlighted with a blue background and the timestamp 'Jan 23, 2021, 9:32:21 PM GMT+5:30'. Below this, the 'Revision overview' section provides details such as 'Time created' (Jan 23, 2021, 9:32:21 PM GMT+5:30), 'Comment' (Revision created on 23 January 2021 04:02PM UTC), 'Revision ID' (fdfad1dde37cfa0f95b8b7c59c940d9a), 'Revision ARN' (arn:aws:dataexchange:us-west-2::data-set/0315a448d5531be81c272c129c187e8d/revisions/fdfad1dde37cfa0f95b8b7c59c940d9a), and 'Data set ID' (0315a448d5531be81c272c129c187e8d).

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



This screenshot shows the AWS Data Exchange service interface again. The left sidebar is identical to the previous screenshot. The main area now displays the 'Assets' section, which lists two items: 'published/PUBLIC/COVID-19-Activity/1611417713/COVID-19Activity.csv' (114.3 MB, updated an hour ago) and 'published/PUBLIC/COVID-19-Activity/1611417713/COVID-19Activity.hyper' (8.8 MB, updated an hour ago). Below the assets, there's a 'Jobs' section with a note stating 'This table lists your jobs from the last 90 days.' The bottom of the screen shows the Windows taskbar with various pinned icons.

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

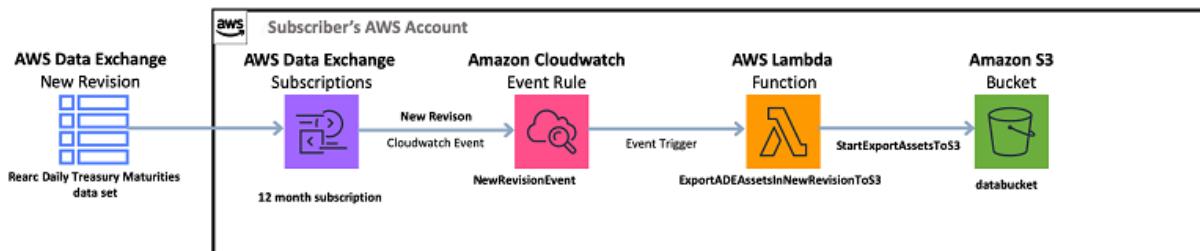
The solution has three steps:

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.

Entitled data sets (1)	
Revision access rules	
All historical revisions All future revisions	
Name	AWS Region
COVID-19 Data Hub	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

CloudFormation - Stack

CloudFormation > Stacks > Create stack

Quick create stack

Template

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

Stack description
 AWS Data Exchange automated revision download upon publish Cloudwatch event

Stack name

Stack name
 Stack name can include letters (A-Z and a-z), numbers (0-9), and dashes (-).

Provide the Data set ID

Revision overview	
Time created	Jan 23, 2021, 11:33:12 PM GMT+5:30
Comment	Revision created on 23 January 2021 06:03PM UTC
Revision ID	19febe60c7347aa0b733c43db5bd94f2
Revision ARN	arn:aws:dataexchange:us-west-2::datasets/0315a448d5531be81c272c129c187e8d/ revisions/19febe60c7347aa0b733c43db5bd94f2
Data set ID	0315a448d5531be81c272c129c187e8d

Data set ID

0315a448d5531be81c272c129c187e8d 

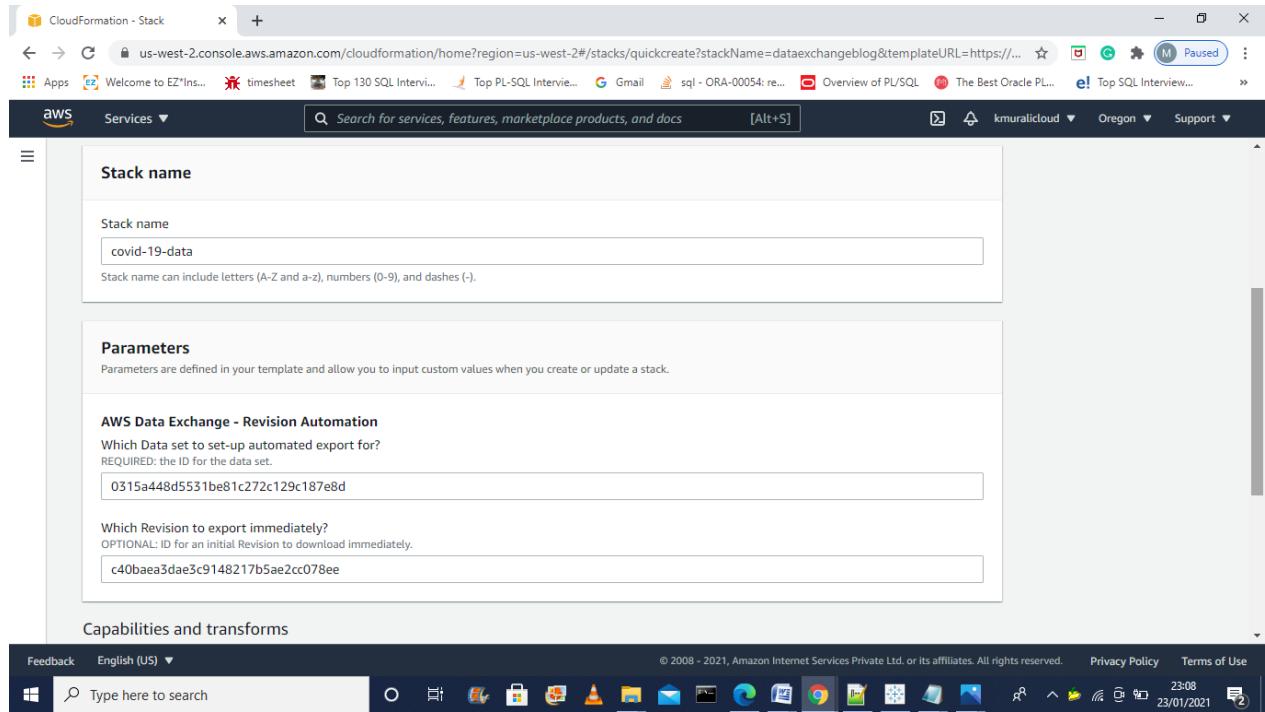
Copy the data set ID as shown in the above figure

Revision ID

19febe60c7347aa0b733c43db5bd94f2 

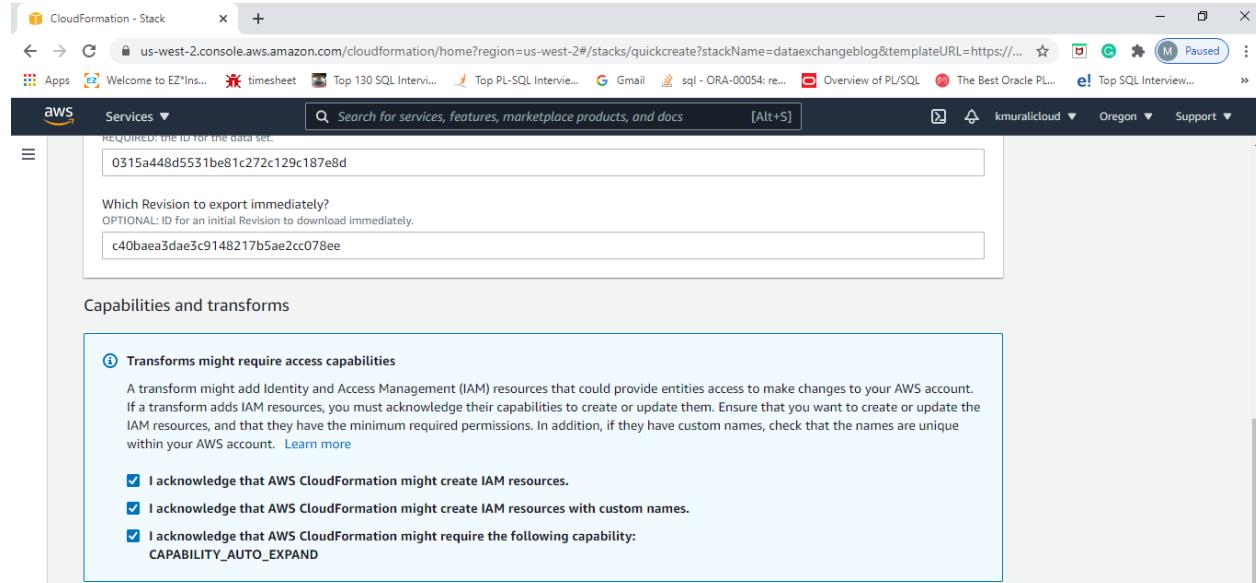
Copy the Revision ID from the above figure

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



The screenshot shows the AWS CloudFormation console interface. In the 'Stack name' section, the stack name 'covid-19-data' is entered. In the 'Parameters' section, two fields are filled: 'Which Data set to set-up automated export for?' with value '0315a448d5531be81c272c129c187e8d' and 'Which Revision to export immediately?' with value 'c40baea3dae3c9148217b5ae2cc078ee'. The bottom of the screen shows the Windows taskbar with various pinned icons.

Create Stack



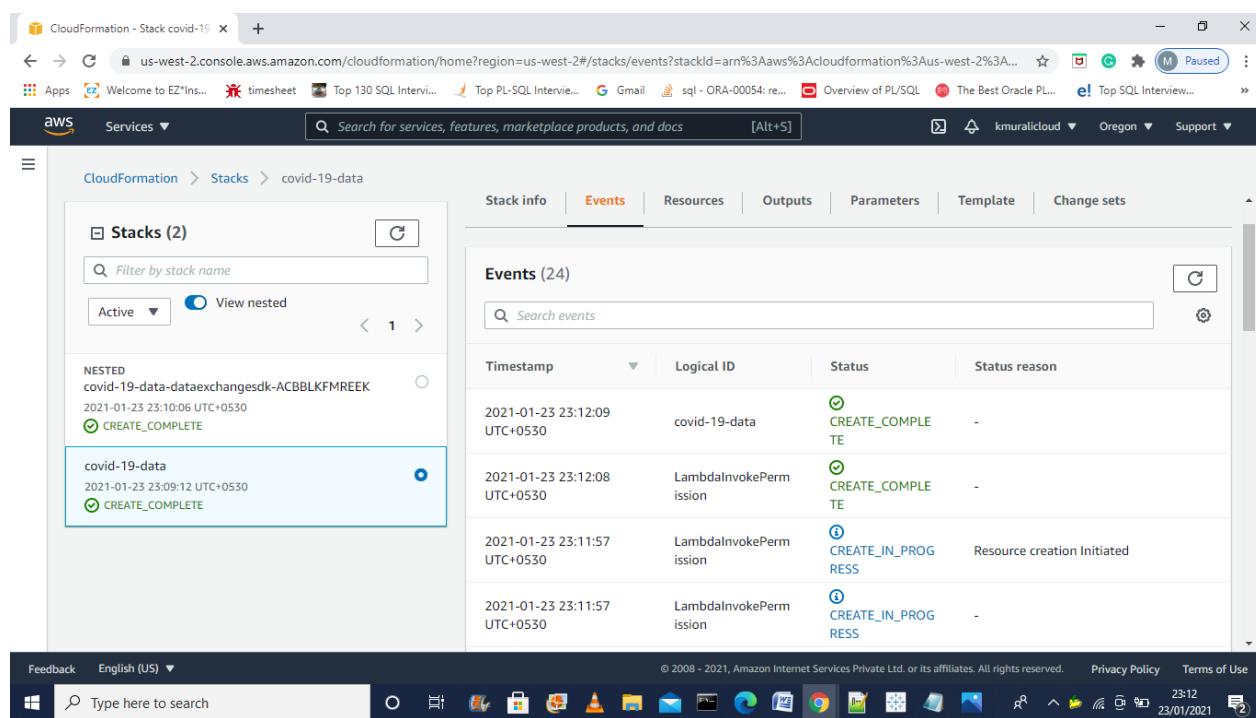
The screenshot shows the 'Create Stack' wizard in the AWS CloudFormation console. The current step is 'Transforms might require access capabilities'. It contains a note about IAM resources and three checked checkboxes:

- I acknowledge that AWS CloudFormation might create IAM resources.
- I acknowledge that AWS CloudFormation might create IAM resources with custom names.
- I acknowledge that AWS CloudFormation might require the following capability: CAPABILITY_AUTO_EXPAND

At the bottom are 'Cancel', 'Create change set', and 'Create stack' buttons.

Feedback English (US) ▾ © 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. Privacy Policy Terms of Use 23:08 23/01/2021

Stack created successfully



The screenshot shows the 'Stacks' page in the AWS CloudFormation console. The 'Events' tab is selected. It displays a table of events for the 'covid-19-data' stack:

Timestamp	Logical ID	Status	Status reason
2021-01-23 23:10:06 UTC+0530	covid-19-data	CREATE_COMPLETE	-
2021-01-23 23:12:08 UTC+0530	LambdaInvokePermission	CREATE_COMPLETE	-
2021-01-23 23:11:57 UTC+0530	LambdaInvokePermission	CREATE_IN_PROGRESS	Resource creation Initiated
2021-01-23 23:11:57 UTC+0530	LambdaInvokePermission	CREATE_IN_PROGRESS	-

Feedback English (US) ▾ © 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. Privacy Policy Terms of Use 23:12 23/01/2021

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

Stack name	Status	Created time	Description
covid-19-data-dataexchangesdk-ACBBLKFMREEK	CREATE_COMPLETE	2021-01-23 23:10:06 UTC+0530	-
covid-19-data	CREATE_COMPLETE	2021-01-23 23:09:12 UTC+0530	AWS Data Exchange automated revision

Covid-19-data-dataexchangesdk details

covid-19-data-dataexchangesdk-ACBBLKFMREEK (NESTED)

Overview

Stack ID	arn:aws:cloudformation:us-west-2:650139163058:stack/covid-19-data-dataexchangesdk-ACBBLKFMREEK/0814d440-5da2-11eb-9c37-0a1d6640693d	Description
Status	CREATE_COMPLETE	Status reason
Root stack	arn:aws:cloudformation:us-west-	Parent stack

Covid-19-data-dataexchangesdk details

CloudFormation - Stack covid - Find and acquire new data set... | Photopia | Online Photo Edit... | Downloads | AWS Data Exchange | Paused

CloudFormation Services Search for services, features, marketplace products, and docs [Alt+S]

CloudFormation > Stacks > covid-19-data-dataexchangesdk-ACBBLKFMREEK

Stacks (2)

- NESTED** covid-19-data-dataexchangesdk-ACBBLKFMREEK
 - Created time: 2021-01-23 23:10:06 UTC+0530
 - Status: CREATE_COMPLETE
- covid-19-data
 - Created time: 2021-01-23 23:09:12 UTC+0530
 - Status: CREATE_COMPLETE

Root stack
arn:aws:cloudformation:us-west-2:650139163058:stack/covid-19-data/e770a890-5da1-11eb-a95b-0afc95244fe9

Parent stack
arn:aws:cloudformation:us-west-2:650139163058:stack/covid-19-data/e770a890-5da1-11eb-a95b-0afc95244fe9

Created time 2021-01-23 23:10:06 UTC+0530 **Deleted time** -

Updated time -

Drift status NOT_CHECKED **Last drift check time** -

Termination protection Disabled on root stack **IAM role** -

Tags (3)
Stack-level tags will apply to all supported resources in your stack. You can add up to 200 unique tags for each stack.

Feedback English (US) ▾ Type here to search © 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. Privacy Policy Terms of Use 01:03 24/01/2021

Covid-19-data-dataexchangesdk Tags

CloudFormation - Stack covid - Find and acquire new data set... | Photopia | Online Photo Edit... | Downloads | AWS Data Exchange | Paused

CloudFormation Services Search for services, features, marketplace products, and docs [Alt+S]

CloudFormation > Stacks > covid-19-data-dataexchangesdk-ACBBLKFMREEK

Stacks (2)

- NESTED** covid-19-data-dataexchangesdk-ACBBLKFMREEK
 - Created time: 2021-01-23 23:10:06 UTC+0530
 - Status: CREATE_COMPLETE
- covid-19-data
 - Created time: 2021-01-23 23:09:12 UTC+0530
 - Status: CREATE_COMPLETE

Tags (3)
Stack-level tags will apply to all supported resources in your stack. You can add up to 200 unique tags for each stack.

Key	Value
lambda:createdBy	SAM
serverlessrepo:applicationId	arn:aws:serverlessrepo:us-east-1:697637923817:applications/dataexchangesdk
serverlessrepo:semanticVersion	0.0.3

▶ Stack policy
Defines the resources that you want to protect from unintentional updates during a stack update.

▶ Rollback configuration
Specifies alarms for CloudFormation to monitor when creating and updating the stack. If the operation breaches an alarm threshold...

Feedback English (US) ▾ Type here to search © 2008 - 2021, Amazon Internet Services Private Ltd. or its affiliates. All rights reserved. Privacy Policy Terms of Use 01:04 24/01/2021

Resources of the covid-19-data-dataexchangesdk

Logical ID	Physical ID	Type	Status	Status reason	Module
SharedLayer52	arn:aws:lambda:us-west-2:650139163058:layer:dataexchangesdk-layer:2	AWS::Lambda::LayerVersion	CREATE_COMPLETE	-	-

Outputs of the stack

Key	Value	Description	Export name
LayerArn	arn:aws:lambda:us-west-2:650139163058:layer:dataexchangesdk-layer:2	-	-

template of the covid-19-data-dataexchangesdk

The screenshot shows the AWS CloudFormation console with the URL <https://us-west-2.console.aws.amazon.com/cloudformation/home?region=us-west-2#/stacks/template?filteringText=&filteringStatus=active&viewNested=true&...>. The page displays a CloudFormation stack named "covid-19-data-dataexchangesdk-ACBBLKFMREEK". The "Template" tab is selected, showing the AWS CloudFormation JSON template. The template details a "SharedLayer" layer version with properties like LayerName: "dataexchangesdk-layer", ContentUri: "awsserverlessrepo-changesets-1f91fp95219h0", and CompatibleRuntimes: "python3.7". The "Outputs" section shows a single output named "LayerArn" with a value of "Ref: SharedLayer". The browser's status bar indicates the URL is <https://us-west-2.console.aws.amazon.com/cloudformation/home?region=us-west-2#/stacks/template?filteringText=&filteringStatus=active&viewNested=true&...>.

```

Transform: AWS::Serverless-2016-10-31
Resources:
  SharedLayer:
    Type: AWS::Serverless::LayerVersion
    Properties:
      LayerName: dataexchangesdk-layer
      ContentUri:
        Bucket: awsserverlessrepo-changesets-1f91fp95219h0
        Key: 650139163058/arn:aws:serverlessrepo:us-east-1:697637923817:applications-dataexchangesdk-versions-0.0.3/3377309a-1fe0-4365-bd52-ac0de385d0bd
      CompatibleRuntimes:
        - python3.7
    Outputs:
      LayerArn:
        Value:
          Ref: SharedLayer
  
```

Covid-19-data summary

The screenshot shows the AWS CloudFormation console with the URL <https://us-west-2.console.aws.amazon.com/cloudformation/home?region=us-west-2#/stacks/stackinfo?filteringText=&filteringStatus=active&viewNested=true&...>. The page displays a CloudFormation stack named "covid-19-data". The "Stack info" tab is selected, showing the "Overview" section. The stack ID is "arn:aws:cloudformation:us-west-2:650139163058:stack/covid-19-data/e770a890-5da1-11eb-a95b-0acf95244fe9". The status is "CREATE_COMPLETE". Other details include the root stack being the same, created time as 2021-01-23 23:09:12 UTC+0530, and deleted time as "-". The browser's status bar indicates the URL is <https://us-west-2.console.aws.amazon.com/cloudformation/home?region=us-west-2#/stacks/stackinfo?filteringText=&filteringStatus=active&viewNested=true&...>.

Stack ID	Description
arn:aws:cloudformation:us-west-2:650139163058:stack/covid-19-data/e770a890-5da1-11eb-a95b-0acf95244fe9	AWS Data Exchange automated revision download upon publish Cloudwatch event

Status	Status reason
CREATE_COMPLETE	-

Root stack	Parent stack
-	-

Created time	Deleted time
2021-01-23 23:09:12 UTC+0530	-

covid-19-data stack summary

The screenshot shows the AWS CloudFormation console with the following details for the 'covid-19-data' stack:

- Stacks (2)**
- Created time:** 2021-01-23 23:09:12 UTC+0530
- Deleted time:** -
- Updated time:** -
- Drift status:** NOT_CHECKED
- Termination protection:** Disabled
- IAM role:** -
- Tags (0):** Stack-level tags will apply to all supported resources in your stack. You can add up to 200 unique tags for each stack.

Resources used in covid-19-data stack

The screenshot shows the AWS CloudFormation console with the following details for the 'covid-19-data' stack:

- Stacks (2)**
- Resources (7)**
- Logical ID** | **Physical ID** | **Type** | **Status**
- DataS3Bucket | covid-19-data-datas3bucket-1xhgqhu3mh1cv | AWS::S3::Bucket | CREATE_COMPLETE
- FirstRevision | CustomResourcePhysicalID | Custom::FirstInvoke | CREATE_COMPLETE
- FunctionGetNewRevision | covid-19-data-FunctionGetNewRevision-1GXIZNBM1ZFT4 | AWS::Lambda::Function | CREATE_COMPLETE

Resources used in covid-19-data stack

The screenshot shows the AWS CloudFormation console with the 'CloudFormation > Stacks > covid-19-data' path selected. On the left, the 'Stacks (2)' section lists two stacks: 'covid-19-data-dataexchangesdk-ACBBLKFMRREK' (nested) and 'covid-19-data' (main). The main stack is currently selected. On the right, the 'Resources (7)' section displays seven resources with their logical IDs, physical IDs, types, and statuses:

Logical ID	Physical ID	Type	Status
FunctionGetNewRevision	covid-19-data-FunctionGetNewRevision-1GXIZNBM1ZFT4	AWS::Lambda::Function	CREATE_COMPLETE
LambdaInvokePermission	covid-19-data-LambdaInvokePermission-1340ITW7LJ29V	AWS::Lambda::Permission	CREATE_COMPLETE
NewRevisionEventRule	covid-19-data-NewRevisionEventRule-OHBOE1HRV4BP	AWS::Events::Rule	CREATE_COMPLETE
RoleGetNewRevision	covid-19-data-RoleGetNewRevision-F2RJIWKJ01QS	AWS::IAM::Role	CREATE_COMPLETE
arn:aws:cloudformation:us-west-2:650139163058:stack/covid-19-data-dataexchangesdk-ACBBLKFMRREK/0814d440-5da2-11eb-9c37-0a1d6640693d		AWS::CloudFormation::Stack	CREATE_COMPLETE
arn:aws:lambda:arn:aws:lambda:us-west-2:1340ITW7LJ29V:lambdafunction:covid-19-data-FunctionGetNewRevision-1GXIZNBM1ZFT4		AWS::Lambda::Function	CREATE_FAILED

Covid-19-data stack summary

This screenshot is identical to the one above, showing the AWS CloudFormation console with the 'CloudFormation > Stacks > covid-19-data' path selected. The 'Stacks (2)' section shows the same two stacks, and the 'Resources (7)' section shows the same seven resources with their details and statuses.

Parameters required for this stack

The screenshot shows the AWS CloudFormation console with the 'Parameters' tab selected for the 'covid-19-data' stack. The 'Parameters' section lists two entries:

Key	Value	Resolved value
RevisionID	c40baea3dae3c9148217b5ae2cc078ee	-
datasetID	0315a448d5531be81c272c129c187e8d	-

Template details

The screenshot shows the AWS CloudFormation console with the 'Template' tab selected for the 'covid-19-data' stack. The template code is displayed in a large text area:

```

AWSTemplateFormatVersion: "2010-09-09"
Transform: "AWS::Serverless-2016-10-31"
Description: "AWS Data Exchange automated revision download upon publish Cloudwatch event"

Metadata:
  AWS::CloudFormation::Interface:
    ParameterGroups:
      - Label:
          default: "AWS Data Exchange - Revision Automation"
        Parameters:
          - datasetID
          - RevisionID
    ParameterLabels:
      datasetID:
        default: "Which Data set to set-up automated export for?"

```

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

Amazon S3 > covid-19-data-datas3bucket-1xhgqhu3mh1cv > published/ > PUBLIC/ > COVID-19-Activity/

COVID-19-Activity/

Objects (3)

Name	Type	Last modified	Size	Storage class
1611421314/	Folder	-	-	-
1611424965/	Folder	-	-	-
1611428506/	Folder	-	-	-

S3 bucket folder created for revision1611421314

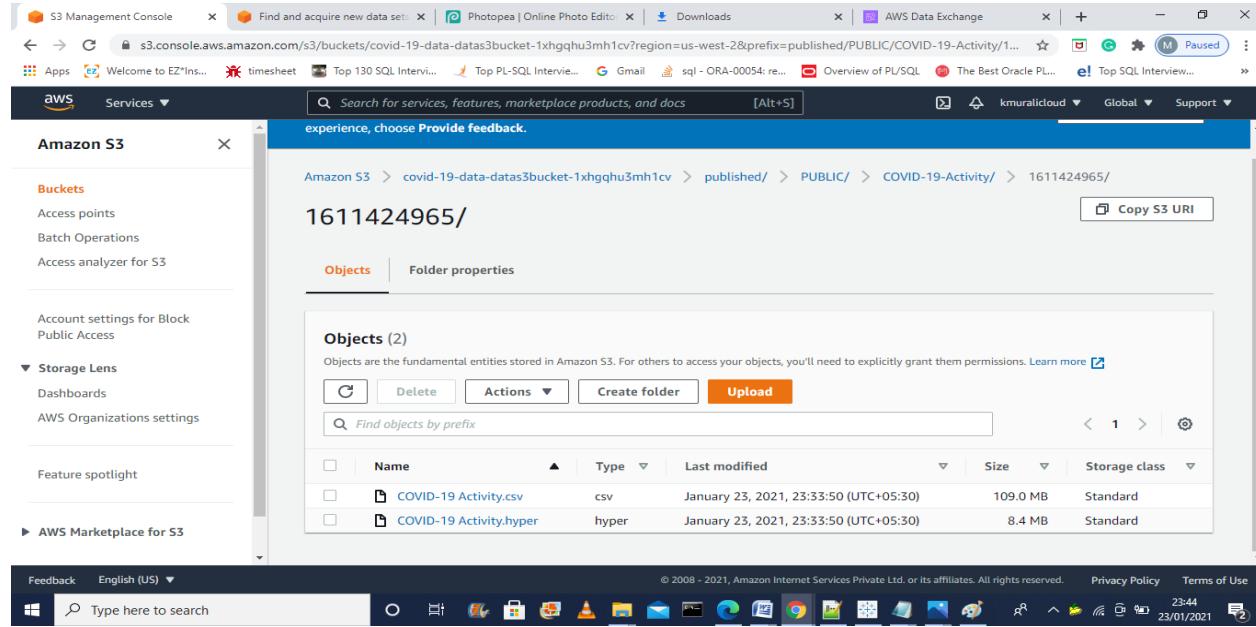
Amazon S3 > covid-19-data-datas3bucket-1xhgqhu3mh1cv > published/ > PUBLIC/ > COVID-19-Activity/ > 1611421314/

1611421314/

Objects (2)

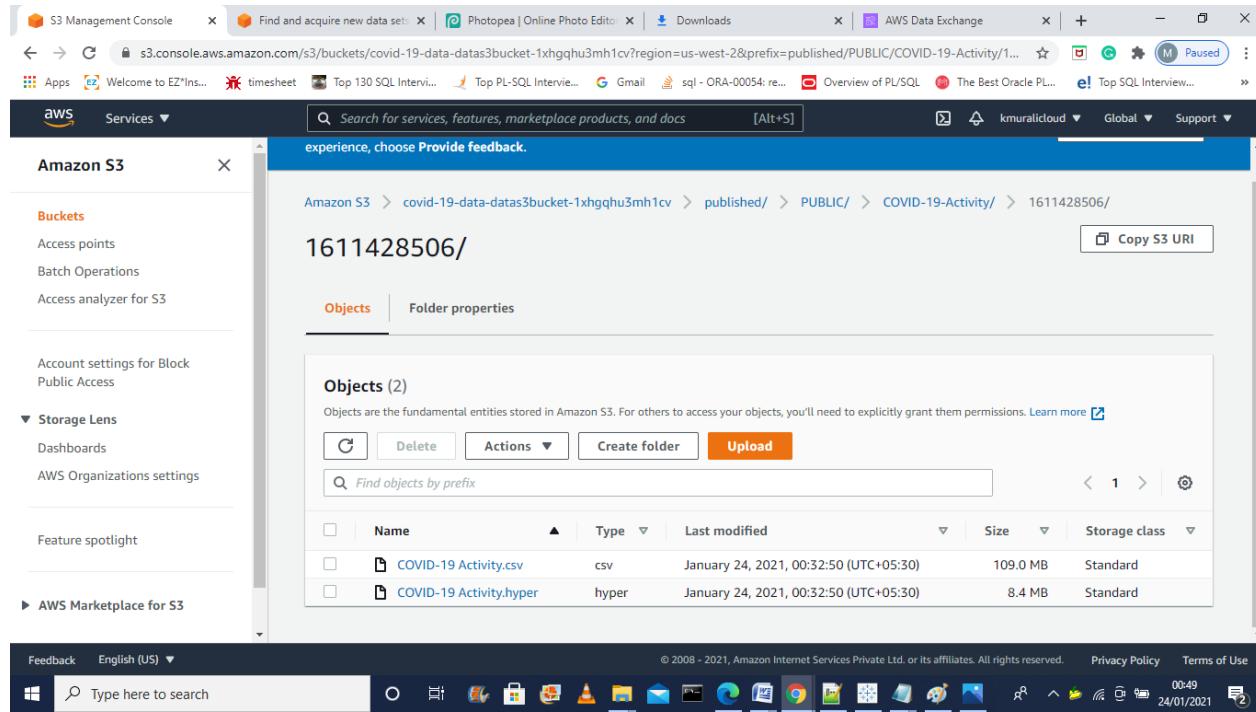
Name	Type	Last modified	Size	Storage class
COVID-19 Activity.csv	csv	January 23, 2021, 23:11:03 (UTC+05:30)	109.0 MB	Standard
COVID-19 Activity.hyper	hyper	January 23, 2021, 23:11:03 (UTC+05:30)	8.4 MB	Standard

S3 bucket folder created for revision1611424965



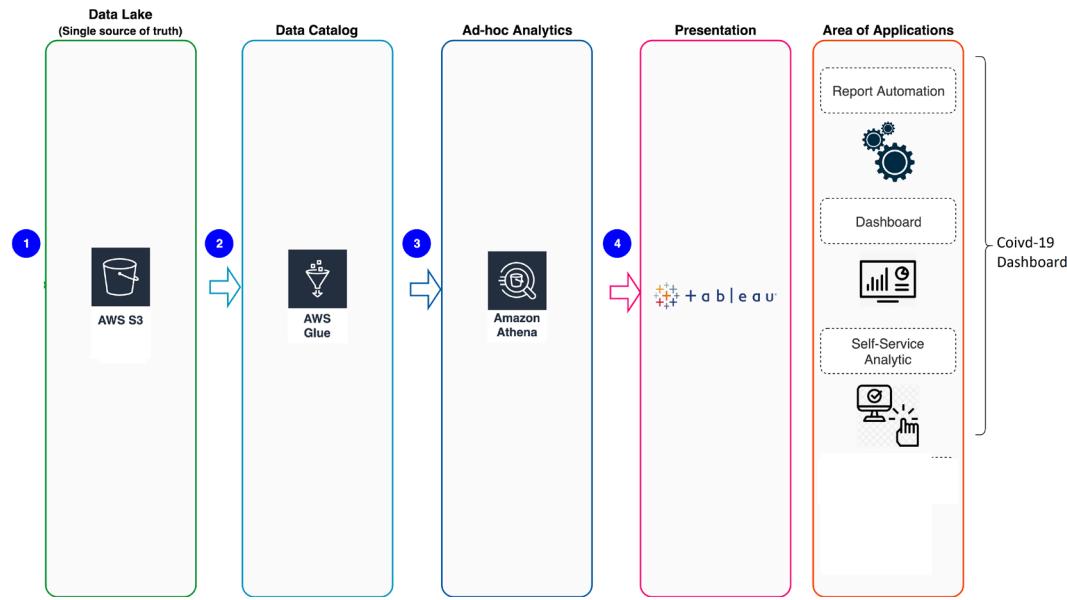
The screenshot shows the AWS S3 Management Console interface. On the left, the navigation pane includes 'Buckets', 'Storage Lens', and 'AWS Marketplace for S3'. The main content area displays a folder named '1611424965/'. Underneath it, there are two objects: 'COVID-19 Activity.csv' (CSV file, 109.0 MB) and 'COVID-19 Activity.hyper' (HyperText file, 8.4 MB), both last modified on January 23, 2021.

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



The screenshot shows the AWS S3 Management Console interface. The navigation pane is identical to the previous one. The main content area displays a folder named '1611428506/'. Underneath it, there are two objects: 'COVID-19 Activity.csv' (CSV file, 109.0 MB) and 'COVID-19 Activity.hyper' (HyperText file, 8.4 MB), both last modified on January 24, 2021.

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

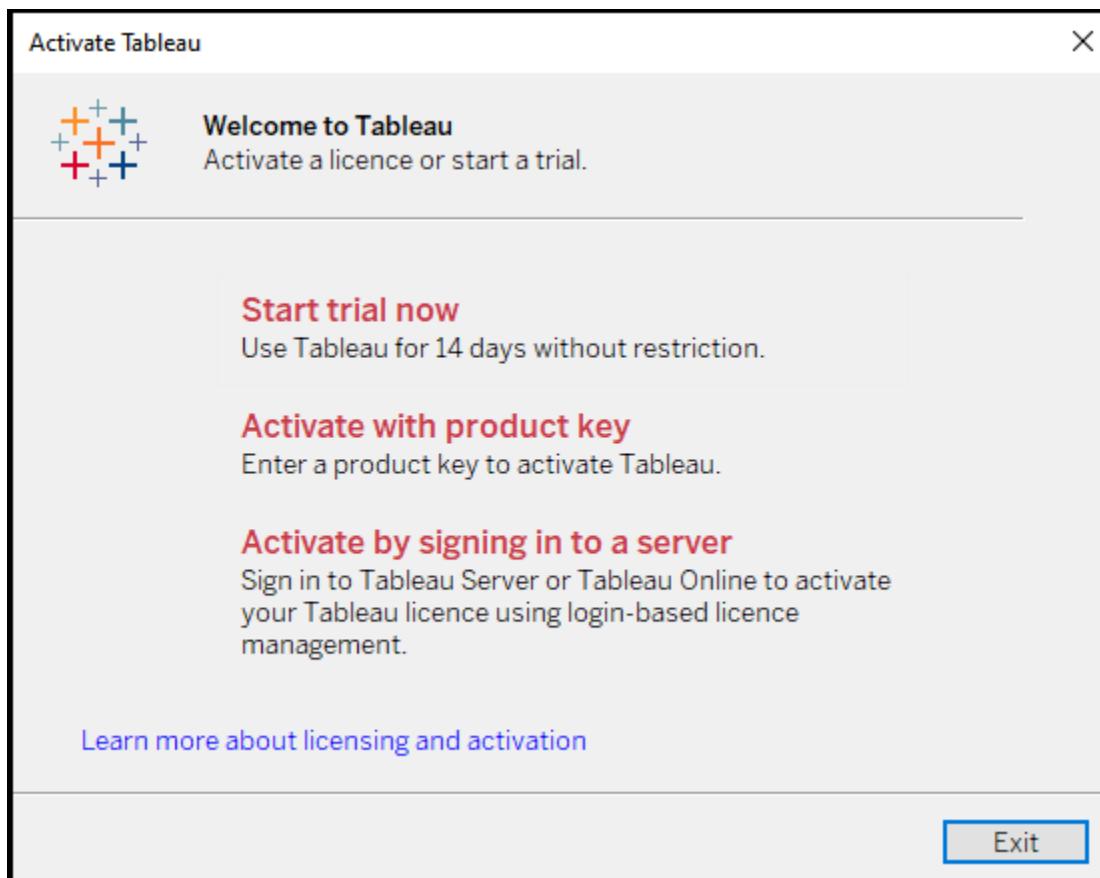
AWS Glue is a fully managed data catalog and ETL (extract, transform, and load) service that simplifies and automates the difficult and time-consuming tasks of data discovery, conversion, and job scheduling. AWS Glue crawls our data sources and constructs a data catalog using pre-built classifiers for 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 trial



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.

This PC > OS (C:) > Program Files > Tableau > Drivers					
	Name	Date modified	Type	Size	
ISS	psqlODBC	24/01/2021 14:12	File folder		
	AthenaJDBC42.jar	24/01/2021 14:28	Executable Jar File	9,723 KB	
	postgresql-42.2.14.jar	14/01/2021 16:30	Executable Jar File	911 KB	

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

Step 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

Policy name	Policy type
AmazonS3FullAccess	AWS managed policy
AWSGlueServiceRole	AWS managed policy

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

Step 3 - Create S3 bucket for data lake and staging

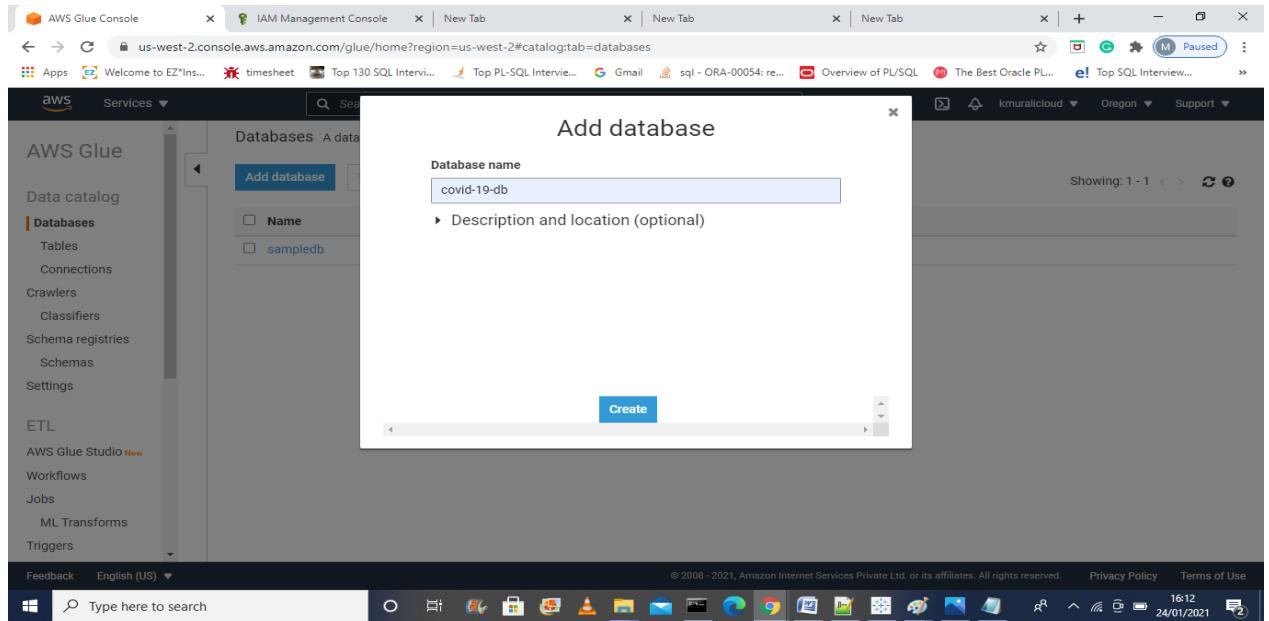
In our Scenario AWS Lambda will create S3 bucket as mentioned in the above screenshots.

Step 4 - Setup AWS Glue data catalog

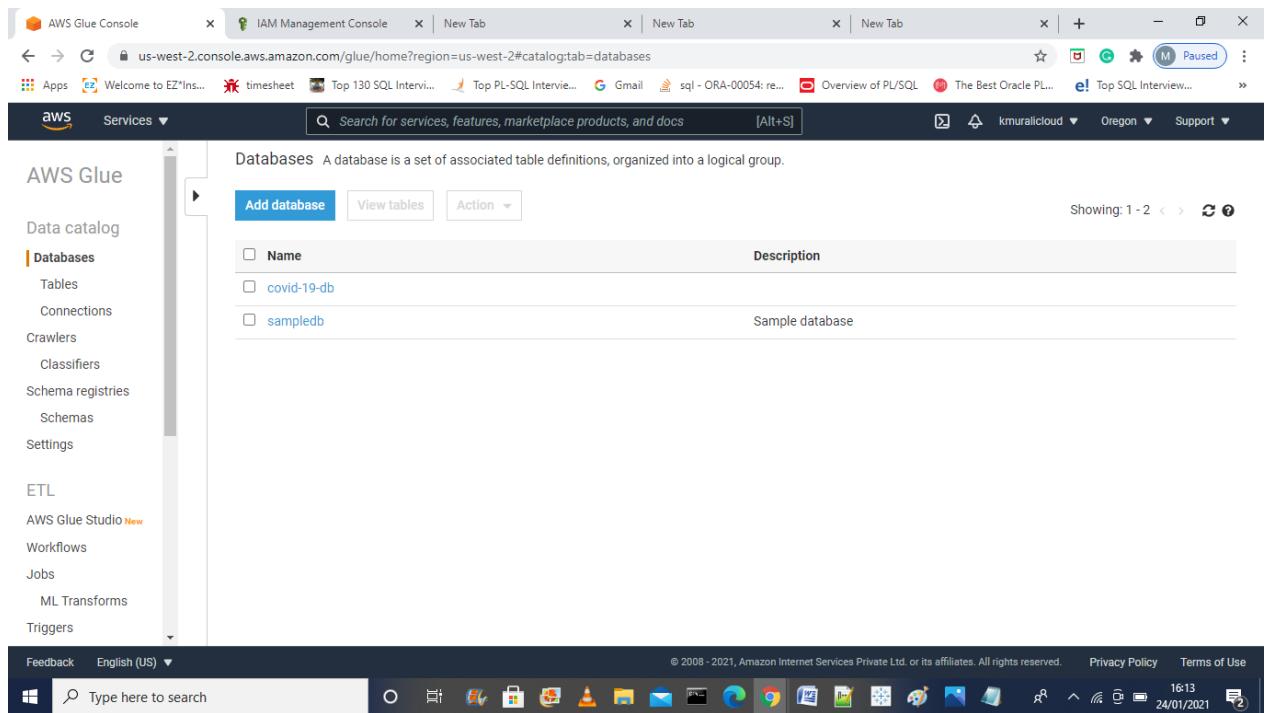
Create database, tables, crawlers, in Glue Data Catalog

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

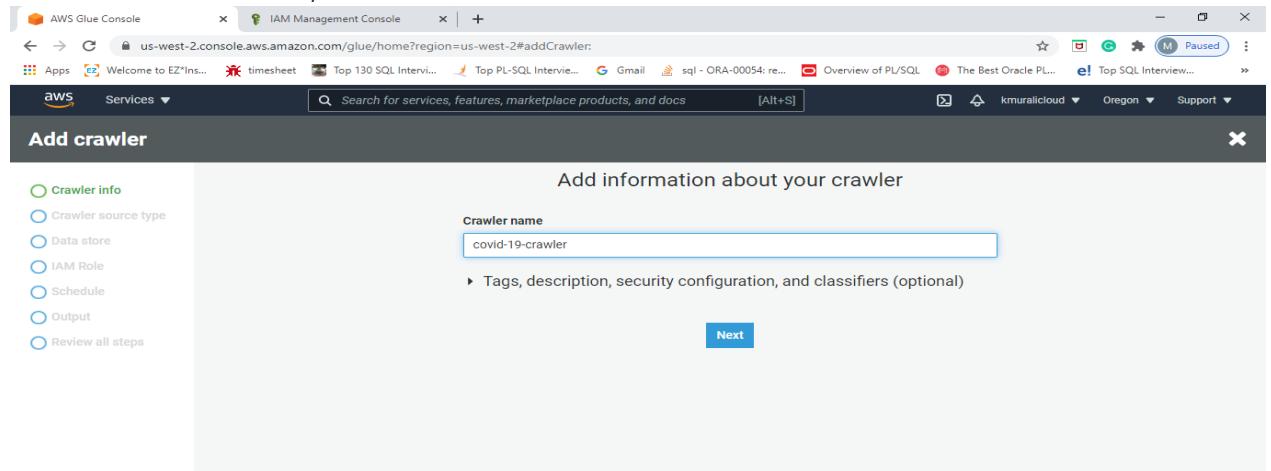
Create Database



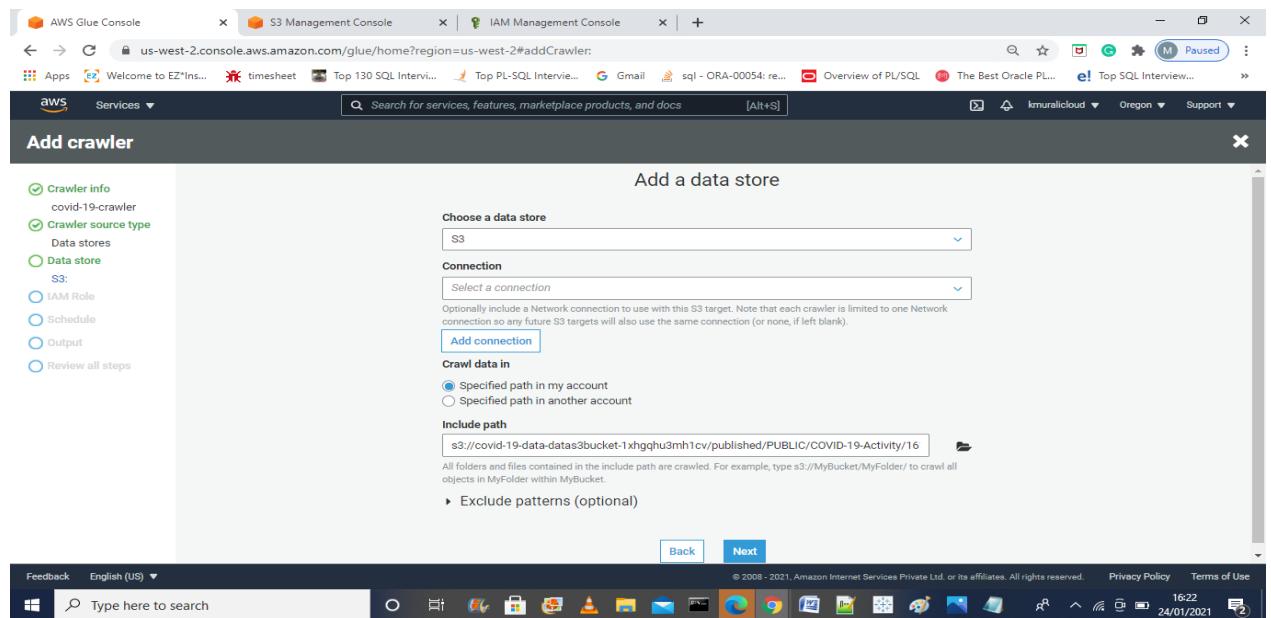
Database added as shown below



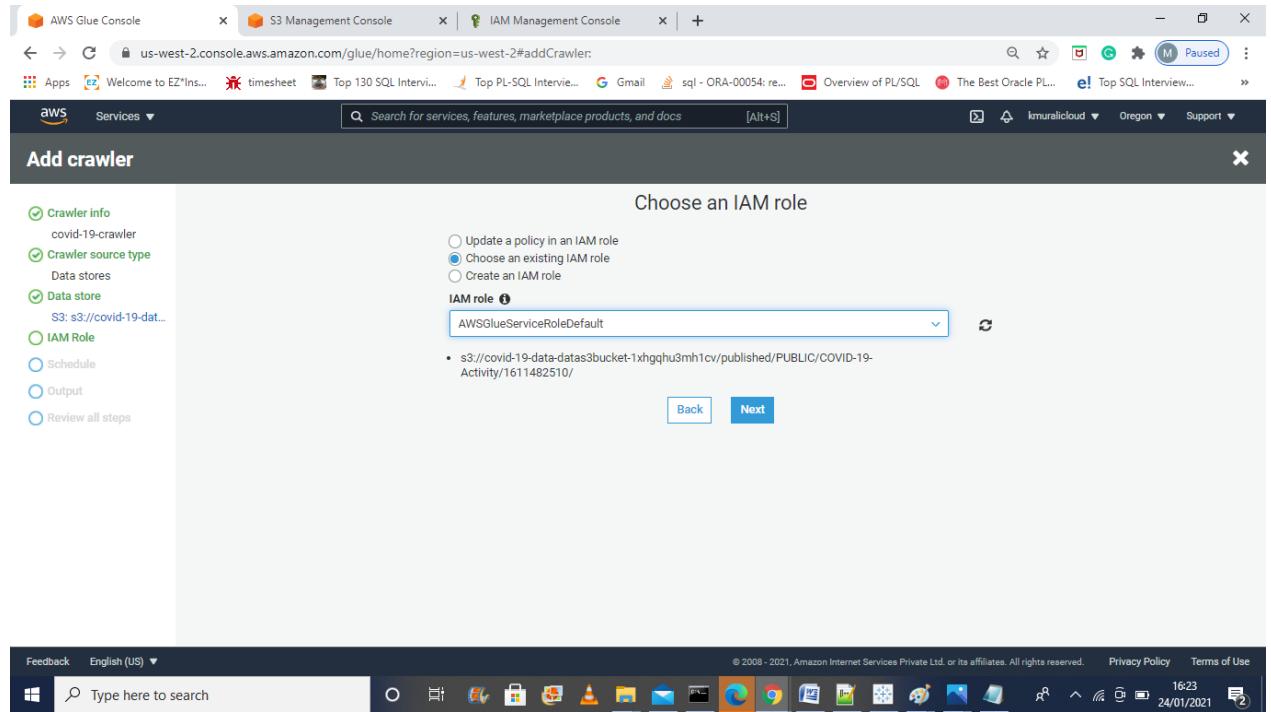
- Choose **Crawlers** in the navigation pane, choose **Add crawler**. Enter the Crawler name **basics-crawler**, and choose **Next**.



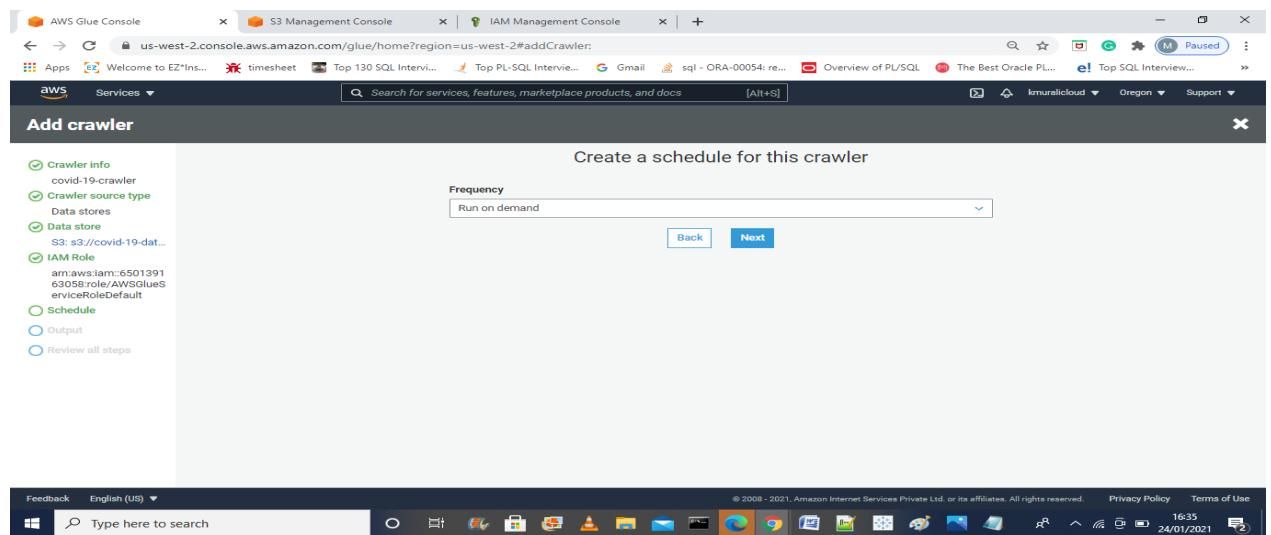
- On the **Add a data store** page, choose **S3** as data store.
- Select **Specified path in my account**.
- Select **Lambda created** folder in the bucket, and choose **Next**.



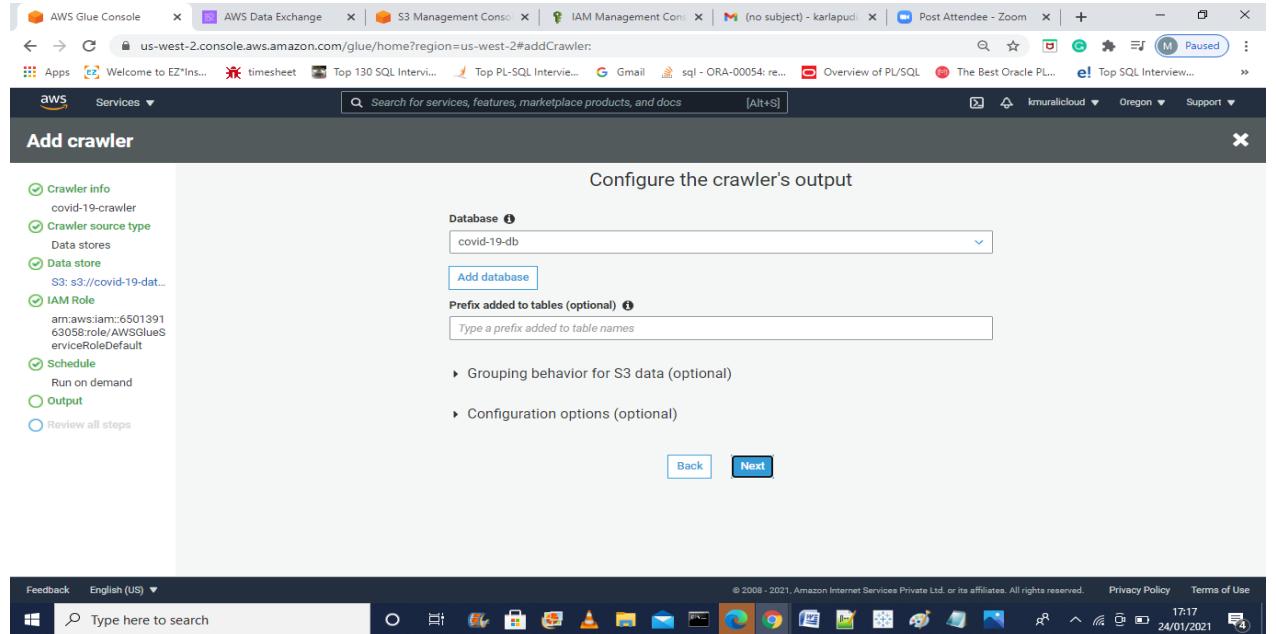
- On **Add another data store** page, choose **No**, and choose **Next**.
- Select **Choose an existing IAM role**, and choose the role **AWSGlueServiceRoleDefault** you just created in the drop-down list, and choose **Next**



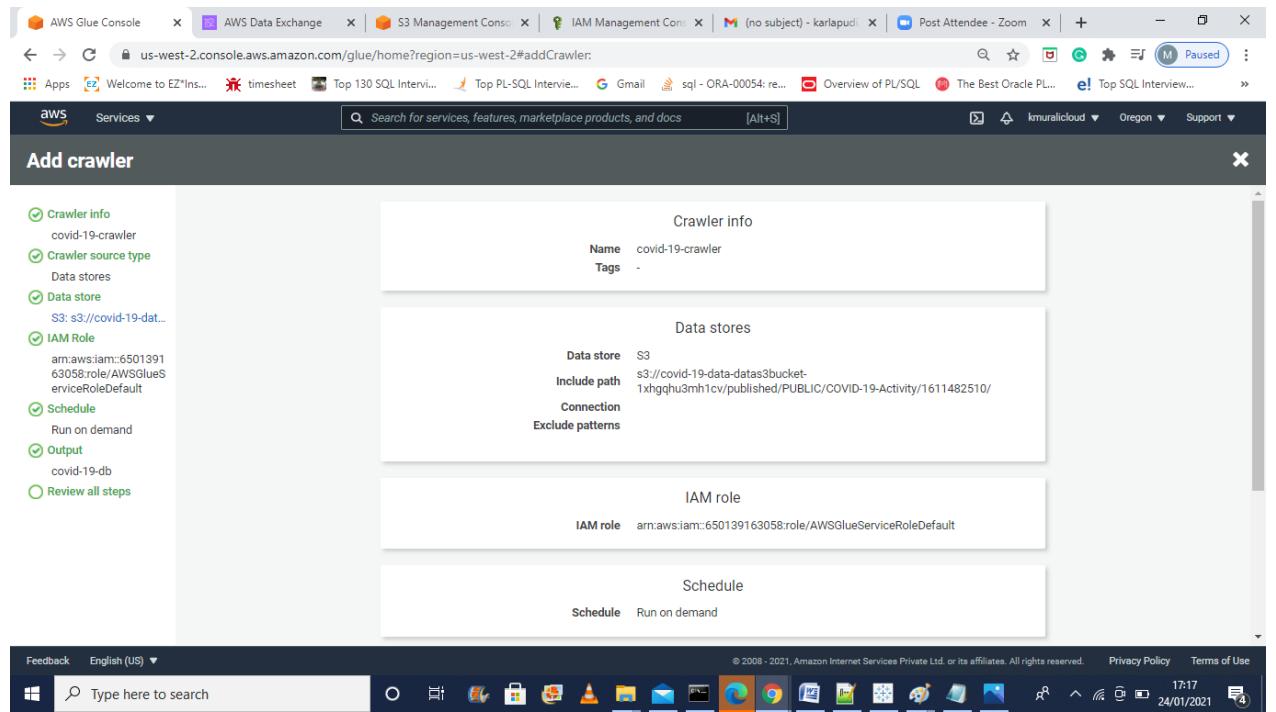
- For **Frequency**, choose **Run on demand**, and choose **Next**.



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



- Review the steps, and choose Finish.



Crawler creation summary page, click on finish to create

Add crawler

Connection

- Crawler info: covid-19-crawler
- Crawler source type: Data stores
- Data store: S3: s3://covid-19-dat...
- IAM Role: arn:aws:iam::650139163058:role/AWSGlueServiceRoleDefault
- Schedule: Run on demand
- Output: covid-19-db

IAM role: arn:aws:iam::650139163058:role/AWSGlueServiceRoleDefault

Schedule: Run on demand

Output

Database: covid-19-db

Prefix added to tables (optional): covid-19-

Create a single schema for each S3 path: false

Configuration options

Back **Finish**

Covid-19 Crawler created

AWS Glue

Crawlers A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.

Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
covid-19-crawler		Ready		0 secs	0 secs	0	0

Action

User preferences

Showing: 1 - 1

Covid-19 Crawler properties:

The screenshot shows the AWS Glue Console interface. On the left, there's a sidebar with various options like Data catalog, Databases, Connections, Crawlers, Classifiers, Schema registries, Schemas, Settings, ETL, AWS Glue Studio, Workflows, Jobs, ML Transforms, Triggers, Dev endpoints, and Notebooks. The 'Crawlers' section is currently selected. In the main content area, it says 'Crawlers > covid-19-crawler'. There are two buttons: 'Run crawler' (highlighted in blue) and 'Edit'. Below this, the crawler details are listed:

- Name: covid-19-crawler
- Description: Create a single schema for each S3 path: false
- Security configuration: Tags: -, State: Ready, Schedule: Last updated: Sun Jan 24 17:18:30 GMT+530 2021, Date created: Sun Jan 24 17:18:30 GMT+530 2021, Database: covid-19-db, Service role: AWSGlueServiceRoleDefault
- Selected classifiers: Data store: S3, Include path: s3://covid-19-data-datas3bucket-1xhgqhu3mh1cv/published/PUBLIC/COVID-19-Activity/1611482510/, Connection: Exclude patterns:
- Configuration options: Schema updates in the data store: Update the table definition in the data catalog, Object deletion in the data store: Mark the table as deprecated in the data catalog.

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

The screenshot shows the AWS Glue Console interface, similar to the previous one but with a different view. The sidebar is the same. In the main content area, it says 'Crawlers A crawler connects to a data store, progresses through a prioritized list of classifiers to determine the schema for your data, and then creates metadata tables in your data catalog.' Below this, there's a table with the following data:

Name	Schedule	Status	Logs	Last runtime	Median runtime	Tables updated	Tables added
covid-19-crawler		Starting		0 secs	0 secs	0	0

Crawler cloudwatch logs:

The screenshot shows the AWS CloudWatch Logs Insights interface. The URL is [us-west-2.console.aws.amazon.com/cloudwatch/home?region=us-west-2#logsV2:log-groups/log-group/\\$252Faws-glue\\$252Fcrawlers/log-eve...](https://us-west-2.console.aws.amazon.com/cloudwatch/home?region=us-west-2#logsV2:log-groups/log-group/$252Faws-glue$252Fcrawlers/log-eve...). The left sidebar shows navigation for CloudWatch services like Dashboards, Alarms, Metrics, and Log groups. The main content area displays log events for the crawler 'covid-19-crawler'. A search bar at the top has the query '2254eb12-d08b-42d6-9f75-d7480b386a12'. Below the search bar are buttons for 'View as text' (unchecked), 'Actions' (dropdown), and 'Create Metric Filter' (button). The log events table has columns for 'Timestamp' and 'Message'. The messages show the crawler starting, writing results to database covid-19-db, creating tables covid_19_activity_csv and covid_19_activity_hyper, and finishing running.

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

The screenshot shows the AWS Glue Studio interface. The URL is us-west-2.console.aws.amazon.com/glue/home?region=us-west-2#. The left sidebar shows navigation for AWS Glue services like Databases, Crawlers, Schema registries, and ETL. The main content area shows a table titled 'Tables' with metadata for various tables. The table includes columns for Name, Database, Location, Classification, Last updated, and Deprecated. Tables listed include 1610564508, covid19activity_csv, covid_19_activity_csv, covid_19_activity_csv, covid_19_activity_csv, covid_19_activity_tableau, and elb_logs.

We can get the table information such as S3 location

Name	Description	Database	Classification	Location	Connection	Deprecated	Last updated	Input format	Output format	Serde serialization lib	Serde parameters
covid_19_activity_csv		covid-19-tableau-prep	csv	s3://covid-19-tableau-prep/COVID-19 Activity Csv/		No	Tue Jan 26 00:00:39 GMT+530 2021	org.apache.hadoop.mapred.TextInputFormat	org.apache.hadoop.hive.io.HiveIgnoreKeyTextOutputFormat	org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	field.delim ,

Table properties	CrawlerSchemaSerializerVersion	recordCount	averageRecordSize	CrawlerSchemaDeserializerVersion					
skip.header.line.count	1	sizeKey	4232872	objectCount	1	UPDATED_BY_CRAWLER	covid-19-tableau-prep		
compressionType	none	columnsOrdered	true	areColumnsQuoted	false	delimiter	,	typeOfData	file

Schema	Column name	Data type	Partition key	Comment
1	daterep	string		

we can also get the table schema

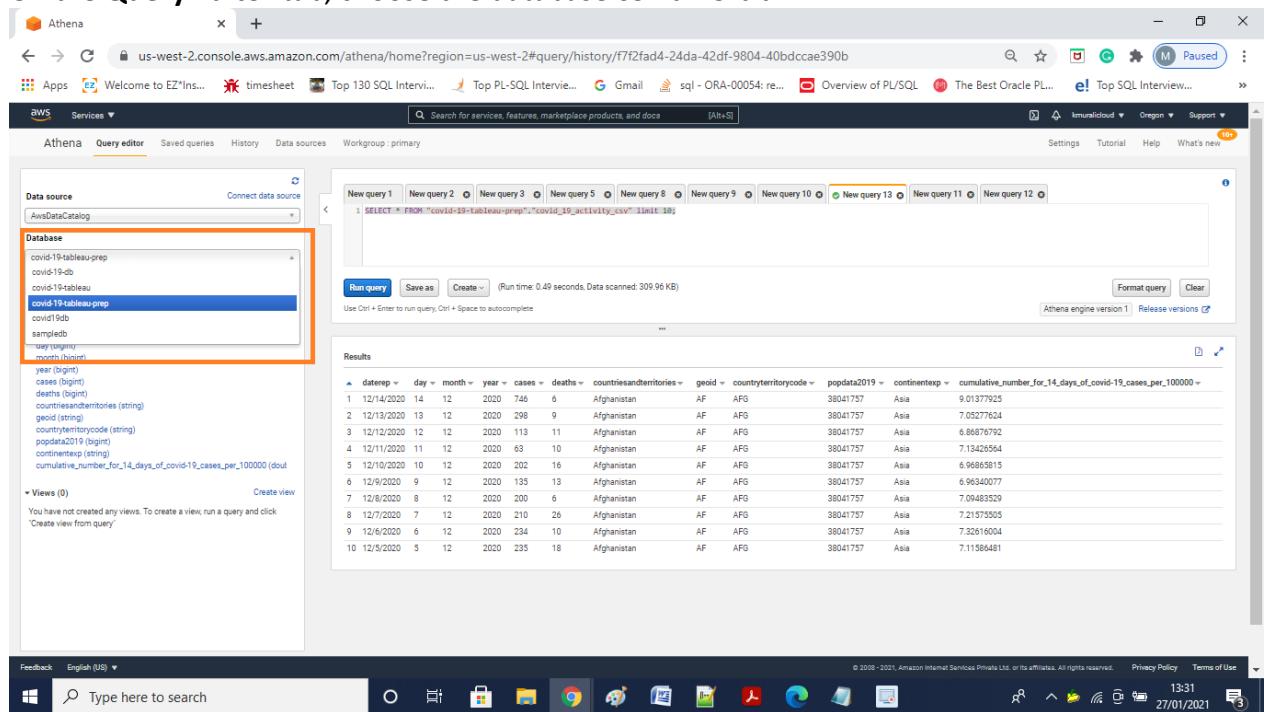
Column name	Data type	Partition key	Comment
1	string		
2	bigint		
3	bigint		
4	bigint		
5	bigint		
6	bigint		
7	string		
8	string		
9	string		
10	bigint		
11	string		
12	double		
19_cases_per_100000	double		

Now we successfully 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**.



The screenshot shows the AWS Athena Query Editor interface. The left sidebar displays the 'Data source' dropdown set to 'AwsDataCatalog' and the 'Database' dropdown set to 'covid19-tableau-prep'. The main area shows a query history with one entry:

```
1: SELECT * FROM "covid-19-tableau-prep"."covid_19_activity_csv" limit 10;
```

The results pane displays the output of the query, which is a table of COVID-19 activity data for Afghanistan. The columns are: datenp, day, month, year, cases, deaths, countriesandterritories, geoid, countryterritorycode, popdata2019, continentexp, and cumulative_number_for_14_days_of_covid-19_cases_per_100000. The data is as follows:

datenp	day	month	year	cases	deaths	countriesandterritories	geoid	countryterritorycode	popdata2019	continentexp	cumulative_number_for_14_days_of_covid-19_cases_per_100000	
1	12/14/2020	14	12	2020	746	6	Afghanistan	AF	AFG	38041757	Asia	9.0377925
2	12/13/2020	13	12	2020	298	9	Afghanistan	AF	AFG	38041757	Asia	7.05277624
3	12/12/2020	12	12	2020	113	11	Afghanistan	AF	AFG	38041757	Asia	6.88786792
4	12/11/2020	11	12	2020	63	10	Afghanistan	AF	AFG	38041757	Asia	7.13426564
5	12/10/2020	10	12	2020	202	16	Afghanistan	AF	AFG	38041757	Asia	7.98865815
6	12/9/2020	9	12	2020	135	13	Afghanistan	AF	AFG	38041757	Asia	9.9340077
7	12/8/2020	8	12	2020	200	6	Afghanistan	AF	AFG	38041757	Asia	7.09483529
8	12/7/2020	7	12	2020	210	26	Afghanistan	AF	AFG	38041757	Asia	7.21575305
9	12/6/2020	6	12	2020	234	10	Afghanistan	AF	AFG	38041757	Asia	7.32616004
10	12/5/2020	5	12	2020	235	18	Afghanistan	AF	AFG	38041757	Asia	7.11586481

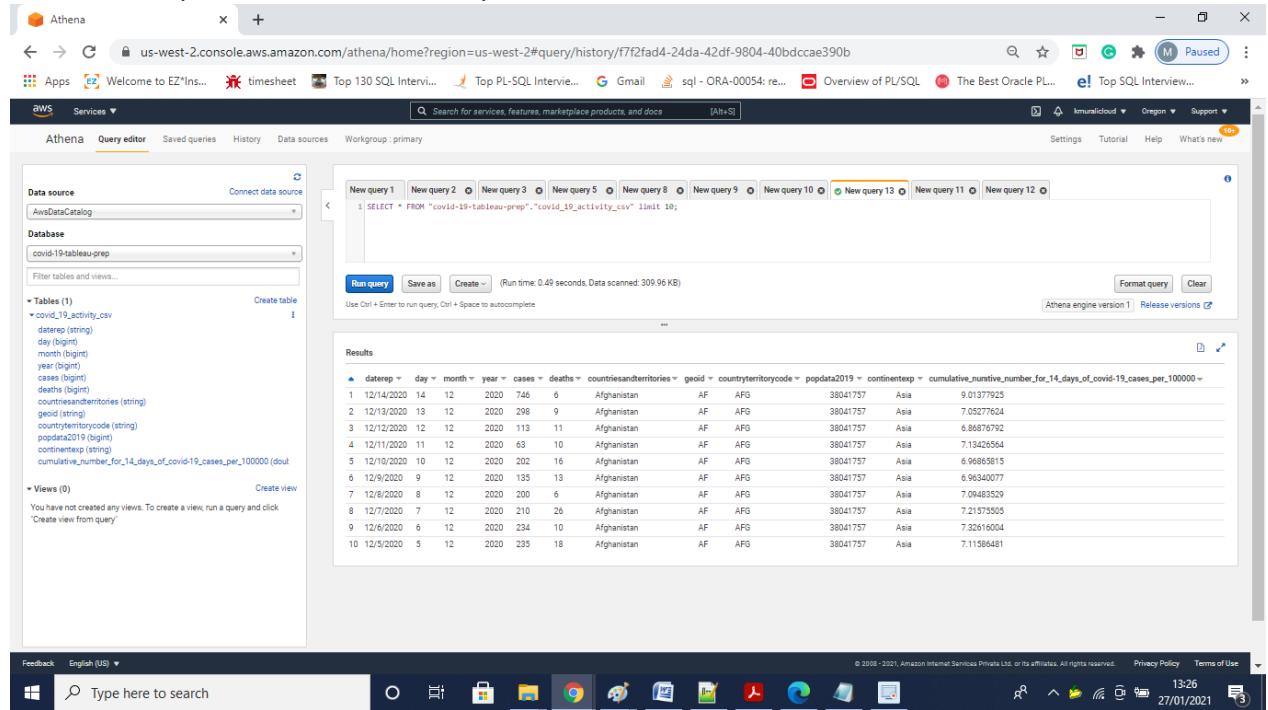
dateRep	day	month	year	cases	deaths	countriesAndTerritories	geoid	countryTerritoryCode	popData2019	continentExp	cumulative_number_for_14_days_of_covid_19_cases_per_100000
12/14/2020	14	12	2020	746	6	Afghanistan	AF	AFG	38041757	Asia	9.01377925
12/13/2020	13	12	2020	298	9	Afghanistan	AF	AFG	38041757	Asia	7.05277624
12/12/2020	12	12	2020	113	11	Afghanistan	AF	AFG	38041757	Asia	6.88678792
12/11/2020	11	12	2020	63	10	Afghanistan	AF	AFG	38041757	Asia	7.13426564
12/10/2020	10	12	2020	202	16	Afghanistan	AF	AFG	38041757	Asia	6.96865815
12/9/2020	9	12	2020	135	13	Afghanistan	AF	AFG	38041757	Asia	6.93402077
12/8/2020	8	12	2020	200	6	Afghanistan	AF	AFG	38041757	Asia	7.09483529
12/7/2020	7	12	2020	210	26	Afghanistan	AF	AFG	38041757	Asia	7.21575505
12/6/2020	6	12	2020	234	10	Afghanistan	AF	AFG	38041757	Asia	7.32616004
12/5/2020	5	12	2020	235	18	Afghanistan	AF	AFG	38041757	Asia	7.11586481

- Query the data, paste below standard SQL in the blank:

```
SELECT * FROM "covid-19-tableau-prep"."covid_19_activity_csv" limit 10;
```

dateRep	day	month	year	cases	deaths	countriesAndTerritories	geoid	countryTerritoryCode	popData2019	continentExp	cumulative_number_for_14_days_of_covid_19_cases_per_100000
12/14/2020	14	12	2020	746	6	Afghanistan	AF	AFG	38041757	Asia	9.01377925
12/13/2020	13	12	2020	298	9	Afghanistan	AF	AFG	38041757	Asia	7.05277624
12/12/2020	12	12	2020	113	11	Afghanistan	AF	AFG	38041757	Asia	6.88678792
12/11/2020	11	12	2020	63	10	Afghanistan	AF	AFG	38041757	Asia	7.13426564
12/10/2020	10	12	2020	202	16	Afghanistan	AF	AFG	38041757	Asia	6.96865815
12/9/2020	9	12	2020	135	13	Afghanistan	AF	AFG	38041757	Asia	6.93402077
12/8/2020	8	12	2020	200	6	Afghanistan	AF	AFG	38041757	Asia	7.09483529
12/7/2020	7	12	2020	210	26	Afghanistan	AF	AFG	38041757	Asia	7.21575505
12/6/2020	6	12	2020	234	10	Afghanistan	AF	AFG	38041757	Asia	7.32616004
12/5/2020	5	12	2020	235	18	Afghanistan	AF	AFG	38041757	Asia	7.11586481

- You can also preview the table to explore the data



The screenshot shows the AWS Athena Query Editor interface. On the left, the sidebar displays the data source (AthenaDataCatalog) and database (covid19-tableau-prep). Under 'Tables (1)', there is a single table named 'covid_19_activity_csv'. The main area shows a query editor with the following SQL code:

```
SELECT * FROM "covid-19-tableau-prep"."covid_19_activity_csv" limit 10;
```

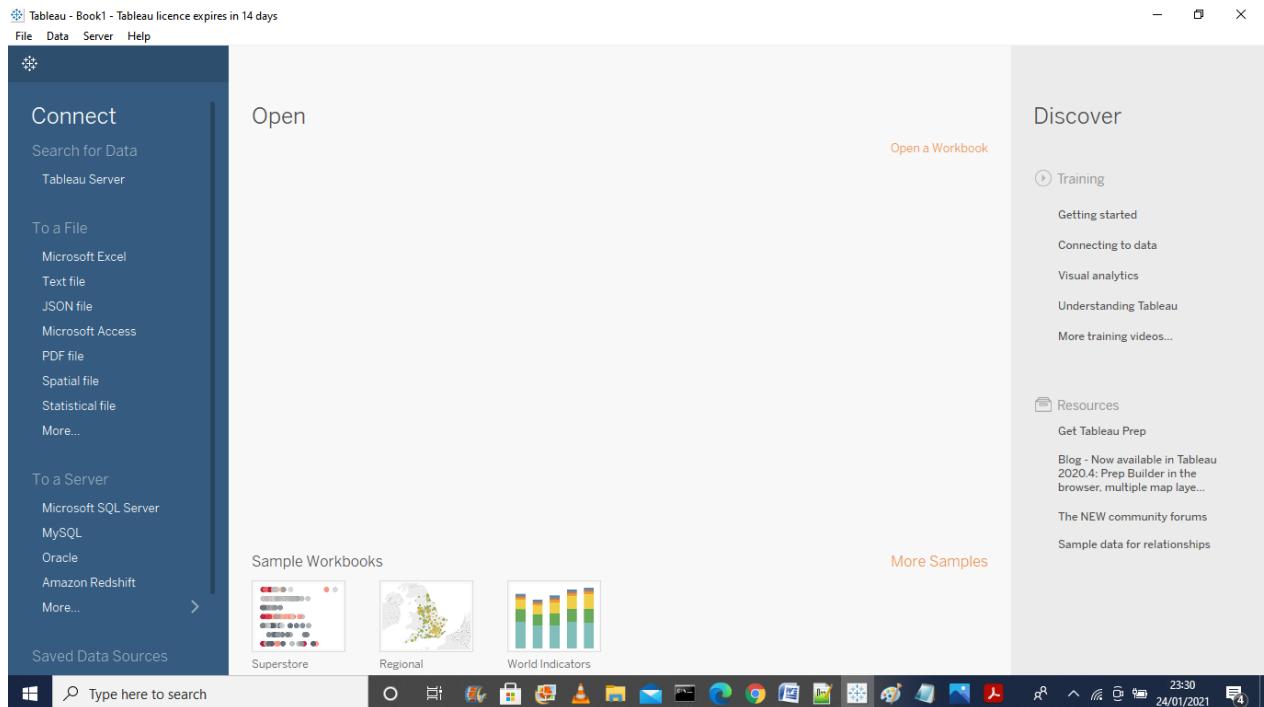
Below the query, the results are displayed in a table format. The columns are: daterep, day, month, year, cases, deaths, countriesandterritories, geoid, countryterritorycode, popdata2019, continentexp, and cumulative_nunlive_number_for_14_days_of_covid-19_cases_per_100000. The data for 10 rows is as follows:

daterep	day	month	year	cases	deaths	countriesandterritories	geoid	countryterritorycode	popdata2019	continentexp	cumulative_nunlive_number_for_14_days_of_covid-19_cases_per_100000	
1	12/14/2020	14	12	2020	746	6	Afghanistan	AF	AFG	38041757	Asia	9.01377925
2	12/13/2020	13	12	2020	298	9	Afghanistan	AF	AFG	38041757	Asia	7.05277624
3	12/12/2020	12	12	2020	113	11	Afghanistan	AF	AFG	38041757	Asia	6.88676792
4	12/11/2020	11	12	2020	63	10	Afghanistan	AF	AFG	38041757	Asia	7.13426564
5	12/10/2020	10	12	2020	206	16	Afghanistan	AF	AFG	38041757	Asia	6.98685815
6	12/9/2020	9	12	2020	135	13	Afghanistan	AF	AFG	38041757	Asia	6.96340077
7	12/8/2020	8	12	2020	200	6	Afghanistan	AF	AFG	38041757	Asia	7.00482529
8	12/7/2020	7	12	2020	210	26	Afghanistan	AF	AFG	38041757	Asia	7.21575505
9	12/6/2020	6	12	2020	234	10	Afghanistan	AF	AFG	38041757	Asia	7.32616004
10	12/5/2020	5	12	2020	235	18	Afghanistan	AF	AFG	38041757	Asia	7.11586481

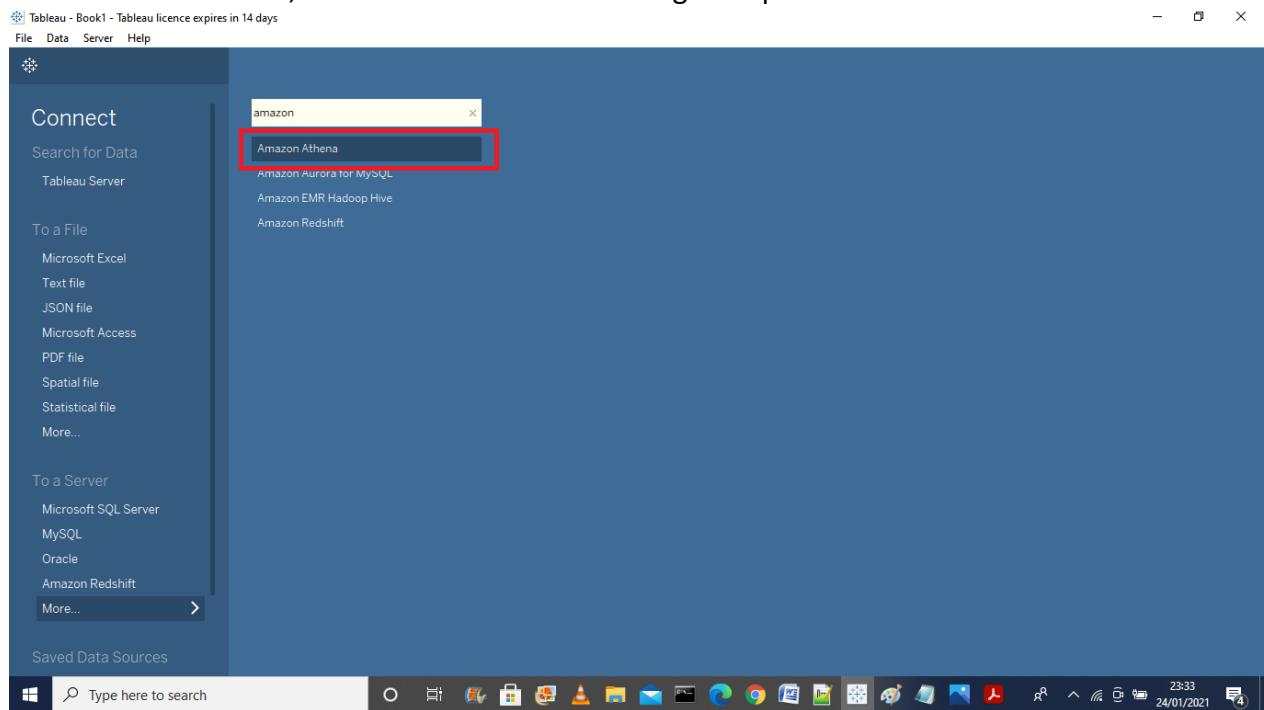
Step 6 - Setup Tableau desktop connection to Athena

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

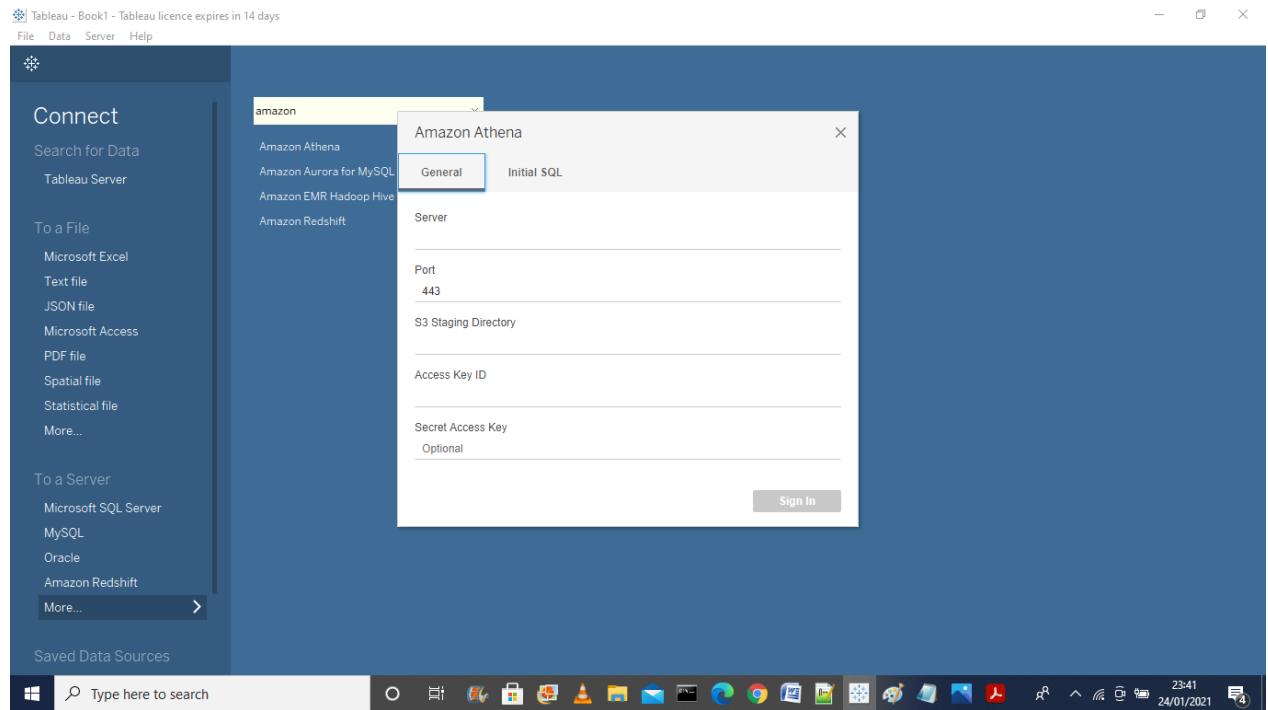


Tableau & Athena Connection Properties

Amazon Athena

General **Initial SQL**

Server
athena.us-west-2.amazonaws.com

Port
 443

S3 Staging Directory
 s3://covid-19-anthe-result/

Access Key ID
 AKIAIIATJUK367F6N6KA

Secret Access Key
 Optional

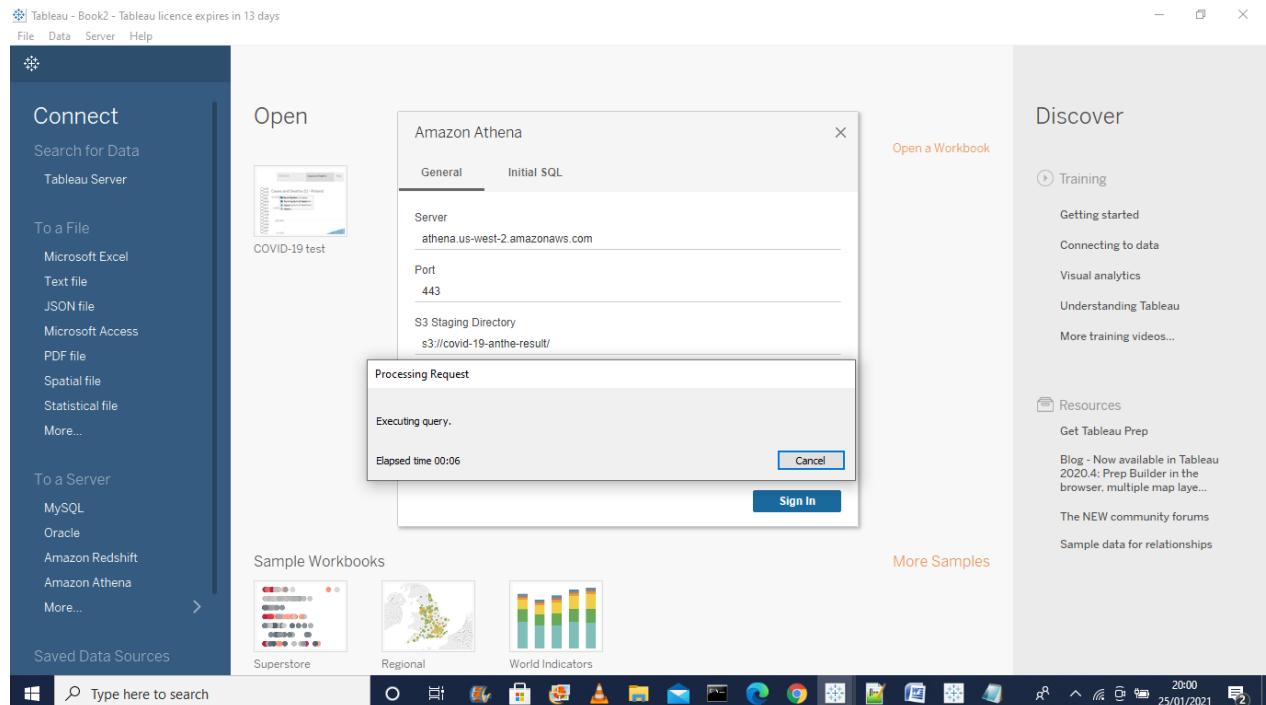
Sign In

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

The screenshot shows the AWS Athena Settings dialog box overlaid on the main Athena Query editor interface. In the dialog box, the 'Query result location' field is set to 's3://covid-19-anthe-result/'. The background shows the Query editor with a table of COVID-19 data.

people_positive_cases_count	county_name	province_state_name	report_date	continent_name	data_source_name	people_death_new_count	county_j
1	19	Hart	Kentucky	2020-05-16	America	New York Times	0
2	19	Hart	Kentucky	2020-05-17	America	New York Times	0
3	19	Hart	Kentucky	2020-05-18	America	New York Times	0
4	20	Hart	Kentucky	2020-05-19	America	New York Times	0
5	20	Hart	Kentucky	2020-05-20	America	New York Times	0
6	20	Hart	Kentucky	2020-05-21	America	New York Times	0
7	21	Hart	Kentucky	2020-05-22	America	New York Times	0

- 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

Tableau - Book1 - Tableau licence expires in 13 days

File Data Server Window Help

Connections Add

athena.us-w...azonaws.com Amazon Athena

Catalogue AwsDataCatalog

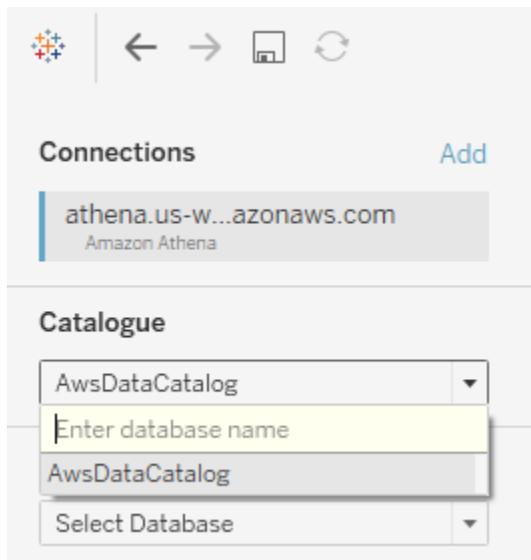
Database Select Database

AwsDataCatalog

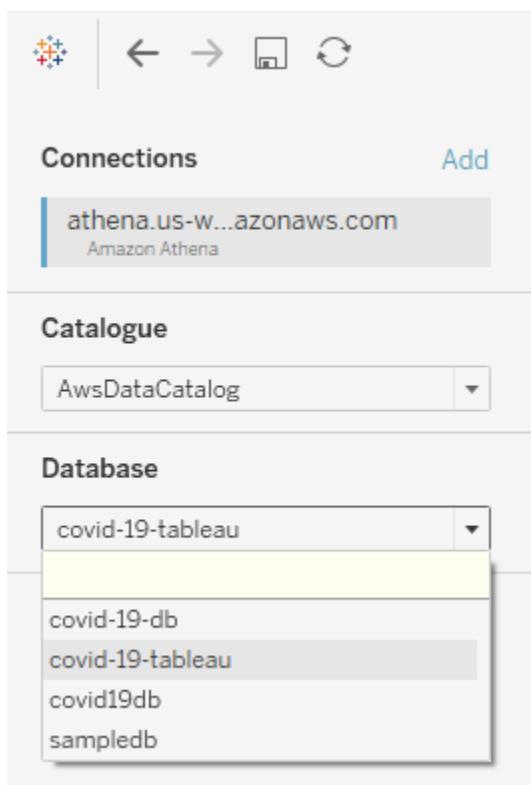
Drag tables here

Sort fields Data source order Show aliases Show hidden fields rows

- Type here to search
- 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

The screenshot shows the Tableau Data Source interface. On the left, there are sections for 'Connections', 'Catalogue', 'Database', and 'Table'. The 'Table' section is highlighted with a red border and contains a list of tables, with 'covid_19_activity_csv' being the selected item.

- Drag and drop the table covid19activity_csv

The screenshot shows the Tableau Data Source interface with the 'covid_19_activity_csv' table selected. The main pane displays the table schema:

Daterep	Day	Month	Year	Cases	Deaths	Abc	Geoid	Abc	Abc	Abc
covid_19_activity_	covid_19_a...	covid_19_activi...	covid_19_ac...	covid_19_acti...						

Below the schema, there are buttons for 'Update Now' and 'Automatically Update'. The bottom of the screen shows the Windows taskbar with various application icons.

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

The screenshot shows the Tableau Data Source interface. On the left, the 'Connections' pane shows 'athena.us-w...azonaws.com' (Amazon Athena) selected. The 'Catalogue' pane shows 'AwsDataCatalog' and 'Database' set to 'covid-19-tableau'. The 'Table' pane shows 'covid_19_activity_csv' selected. The main area displays a preview of the data with columns: #, covid_19_act..., #, covid_19_act..., #, covid_19_act..., Abc, covid_19_act..., Abc, covid_19_act..., #, covid_19_act..., Abc, covid_19_act..., #, covid_19_act.... The data preview shows rows for Afghanistan from 2020 with various values for cases, deaths, and geoid. A message at the bottom says 'Need more data? Drag tables here to relate them. Learn more'.

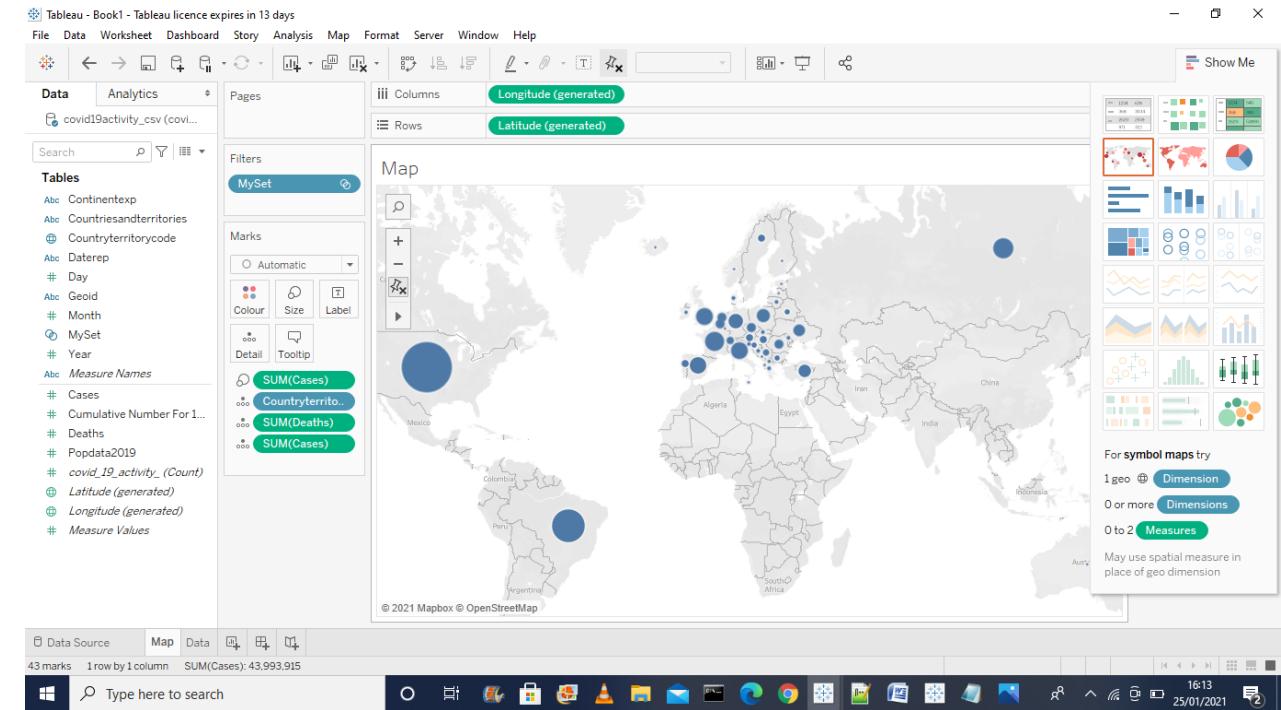
- Create Set countryTerritorycode as shown below for the codes

The screenshot shows the 'Create Set' dialog box in Tableau. The 'Name' field is set to 'MySet'. The 'Select from list' option is selected. A list of countries is shown in a dropdown: UGA, UKR, URY, USA, UZB, VAT, VCT, VEN, VGB, VIR, VNM. Below the list are buttons for 'All' and 'None'. The 'Summary' section shows the selected field as '[Countryterritorycode]'. The 'OK' button is visible at the bottom right of the dialog.

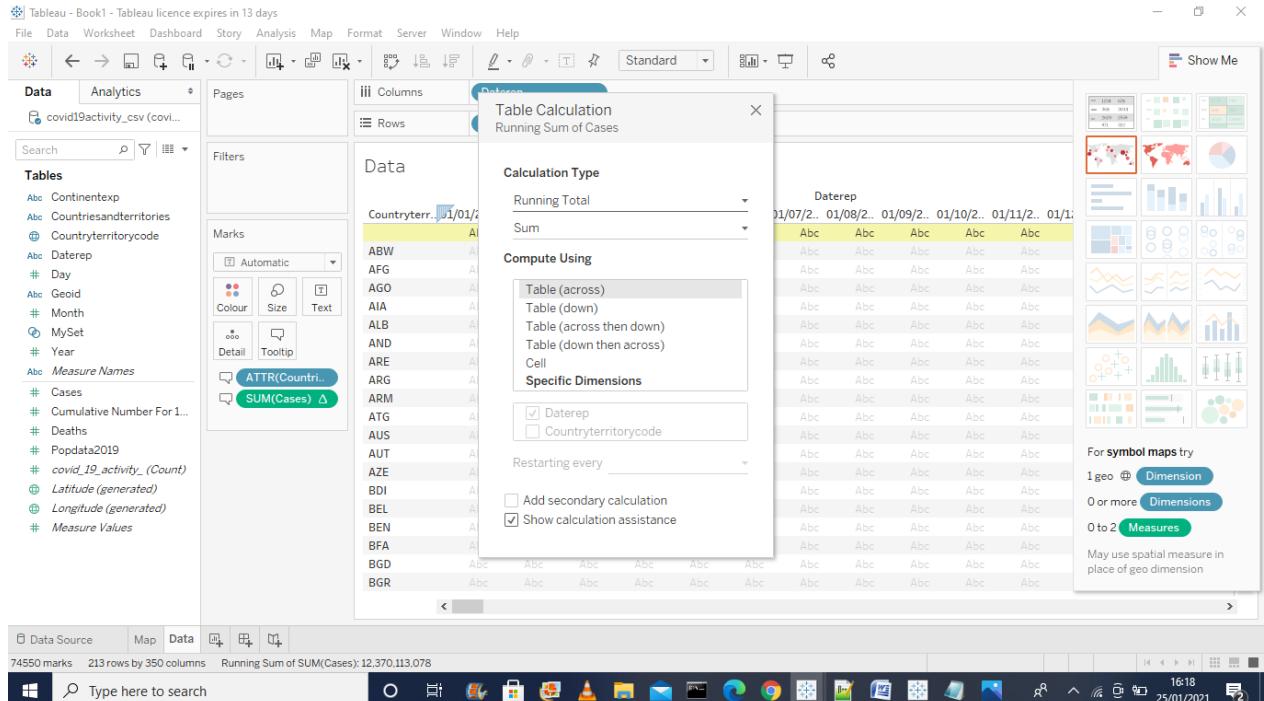
- Myset created for the below countryterritorycode code
 ALB,AUS,AUT,BEL,BGR,BIH,BLR,BRA,CHE,CZE,DEU,DNK,ESP,EST,FIN,FRA,GBR,GRC,HRV,

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



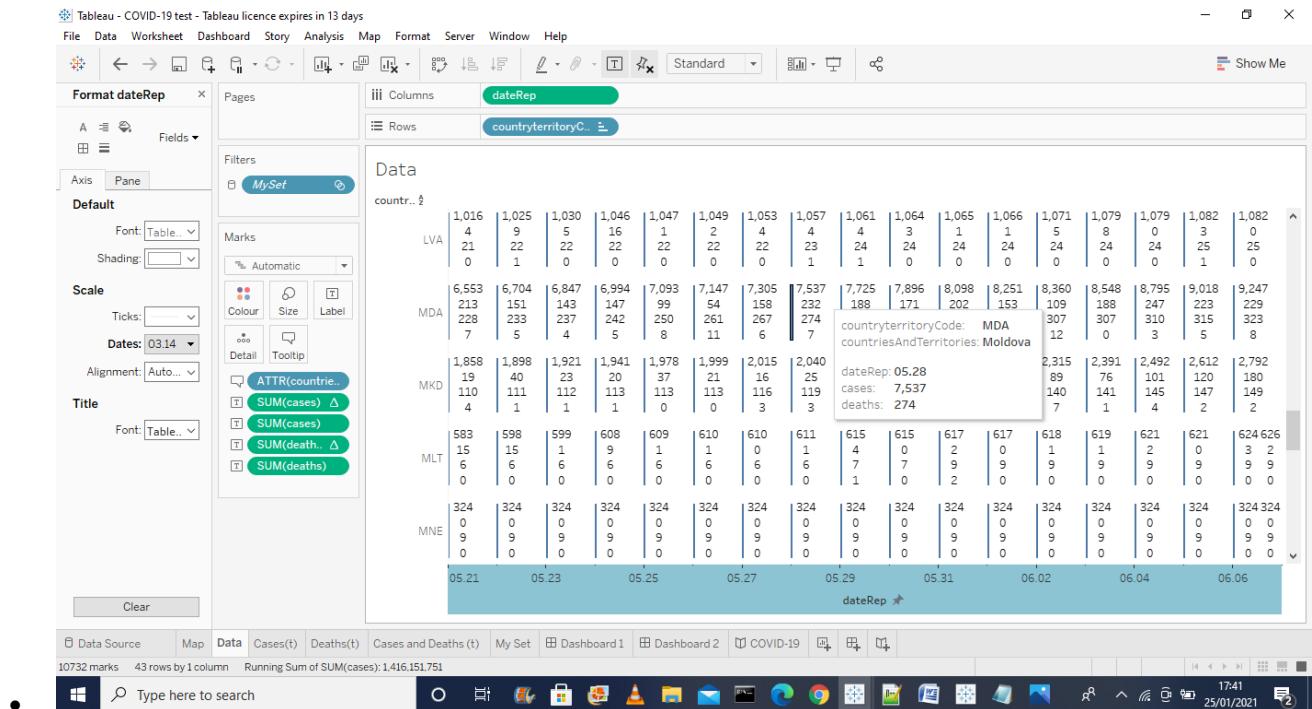
- Running total dashboard creation



The screenshot shows the Tableau interface with the following details:

- Tableau - Book1 - Tableau licence expires in 13 days**
- Data** pane: Shows a table named "covid19activity_csv" with columns: DateRep, CountryTerritoryCode, and SUM(Cases).
- Table Calculation** pane: A "Running Sum of Cases" calculation is being created under "Data". The "Calculation Type" is set to "Running Total" and "Sum". The "Compute Using" dropdown is expanded, showing "Table (across)" and "Table (down)". The "Specific Dimensions" dropdown is also expanded, showing "Daterep" and "Countryterritorycode".
- Dashboard**: A preview of a dashboard is shown, featuring a grid of data points for various countries over time.
- Bottom Navigation**: Includes "Data Source", "Map", "Data", "Cases(t)", "Deaths(t)", "Cases and Deaths (t)", "My Set", "Dashboard 1", "Dashboard 2", and "COVID-19".

- Data Dashboard creation

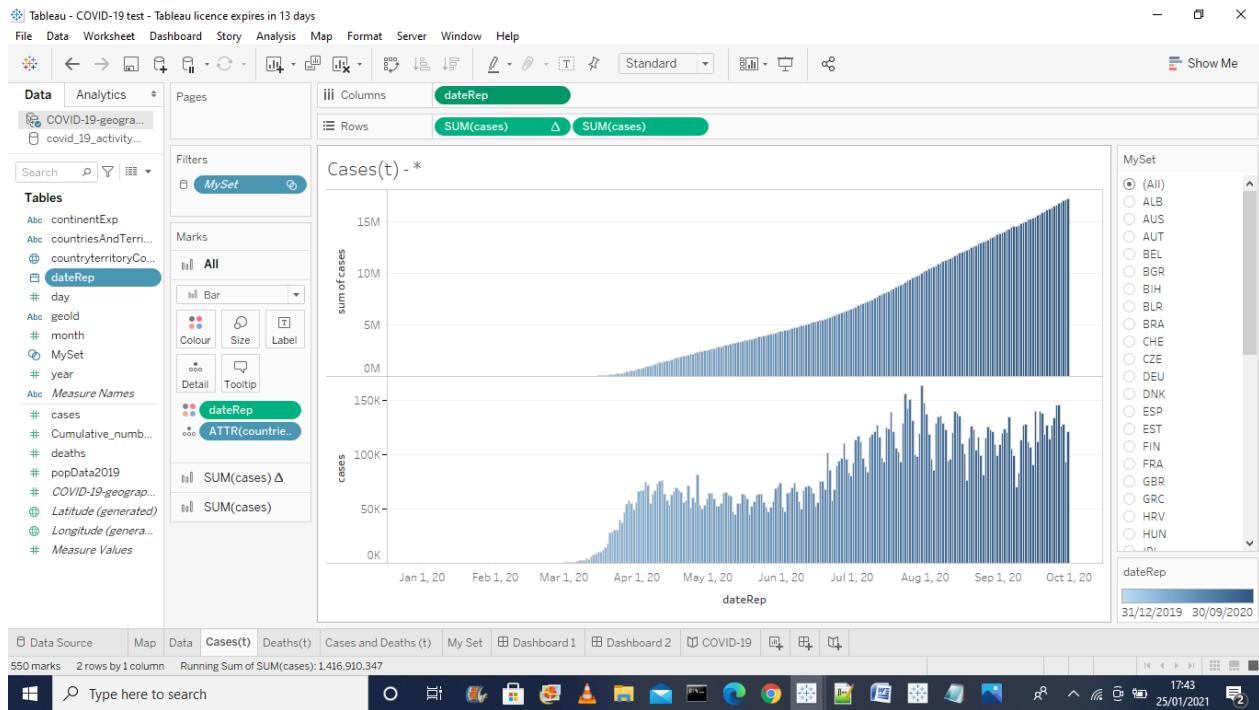


The screenshot shows the Tableau interface with the following details:

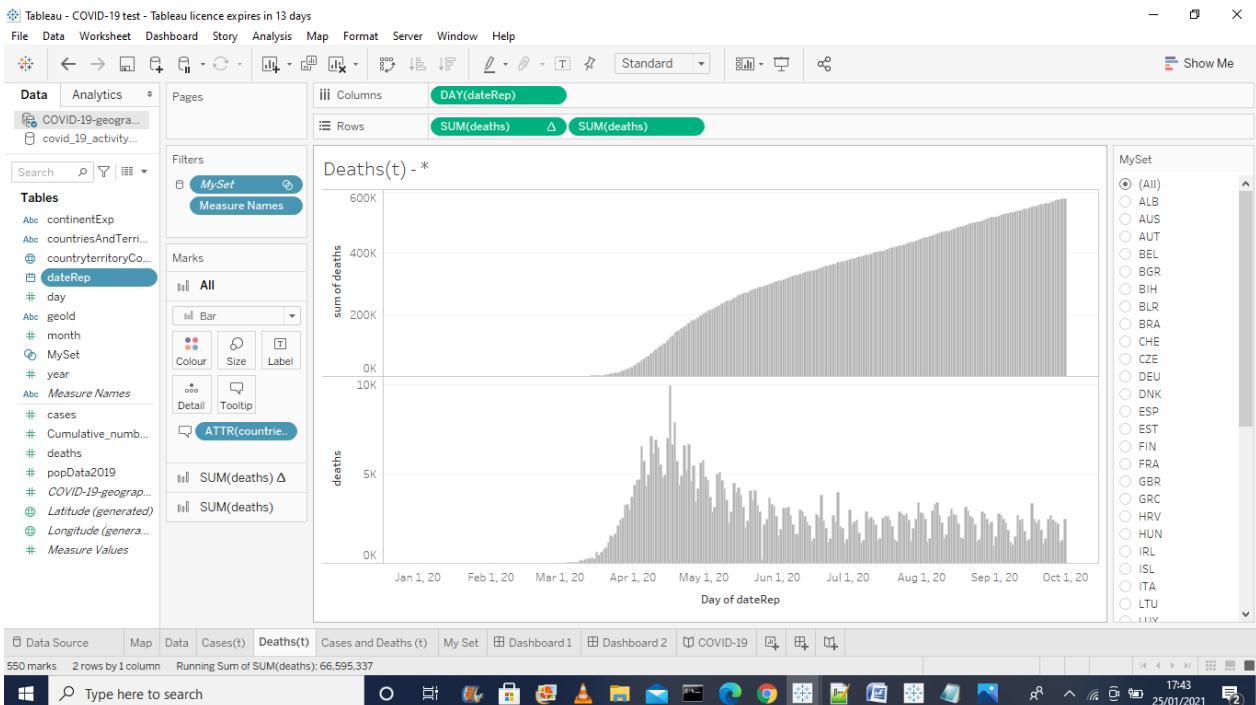
- Tableau - COVID-19 test - Tableau licence expires in 13 days**
- Format dateRep** pane: Settings for the "dateRep" field, including "Font: Table..", "Shading: []", "Scale: Ticks: []", "Dates: 03.14", "Alignment: Auto..", and "Title: Font: Table..".
- Data** pane: A large grid of data for various countries and territories. The grid includes columns for dateRep, countryterritoryCode, and countryname, along with numerous numerical values for cases and deaths.
- Bottom Navigation**: Includes "Data Source", "Map", "Data", "Cases(t)", "Deaths(t)", "Cases and Deaths (t)", "My Set", "Dashboard 1", "Dashboard 2", and "COVID-19".

- Cases Dashboard alone

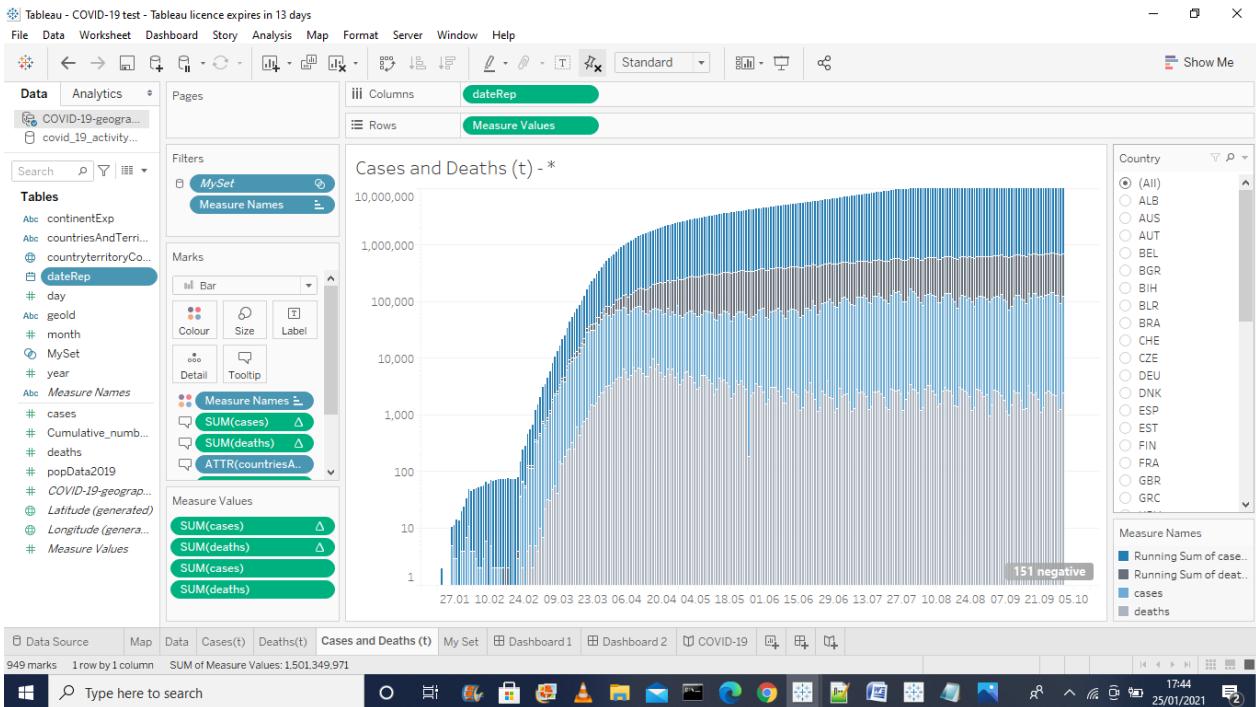
-



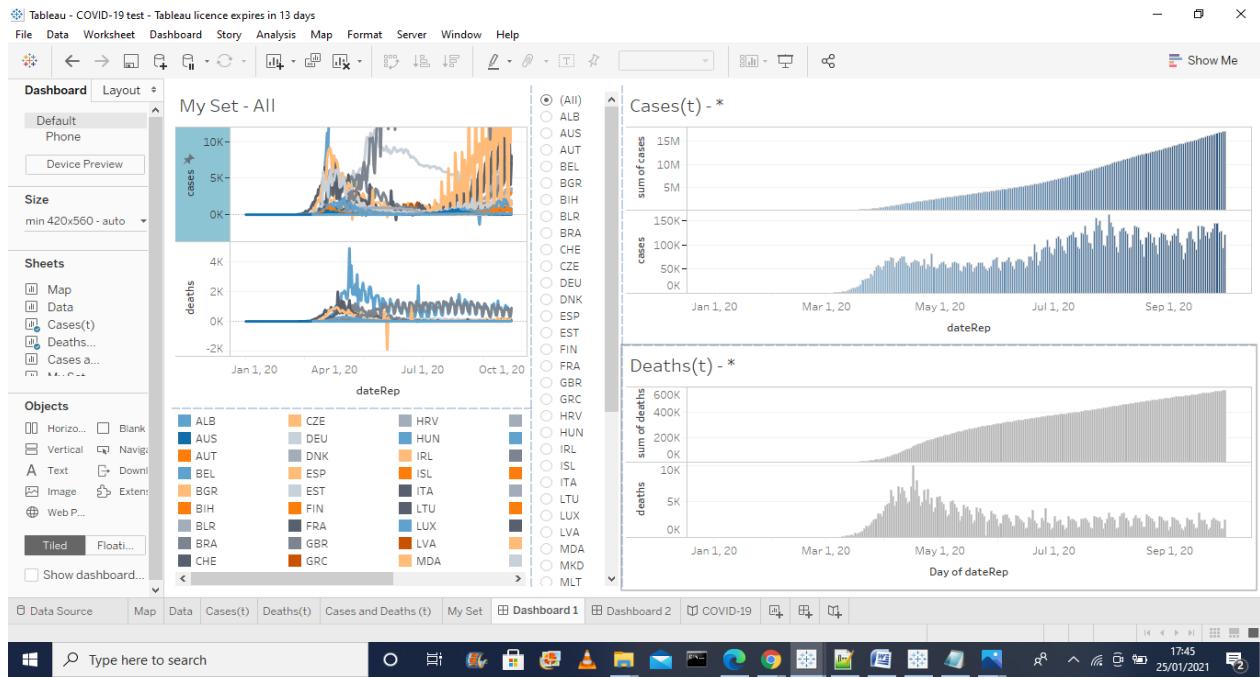
- Death Dashboard



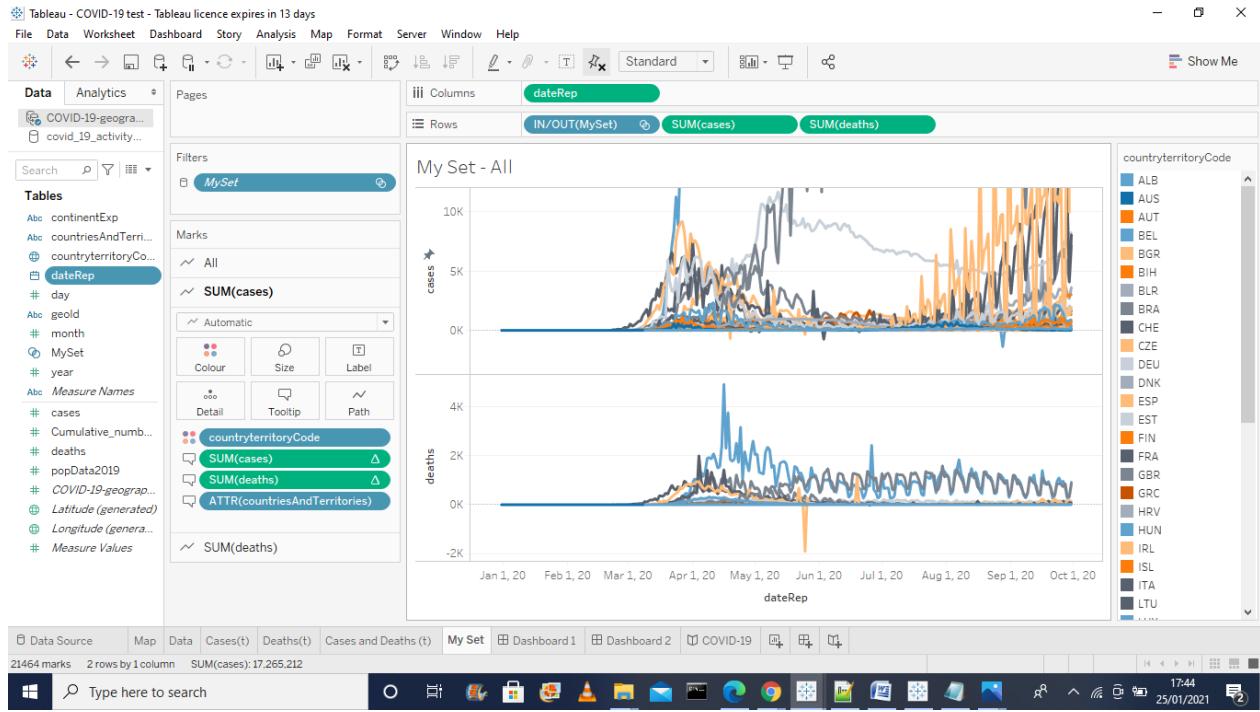
Cases & Death Dashboard



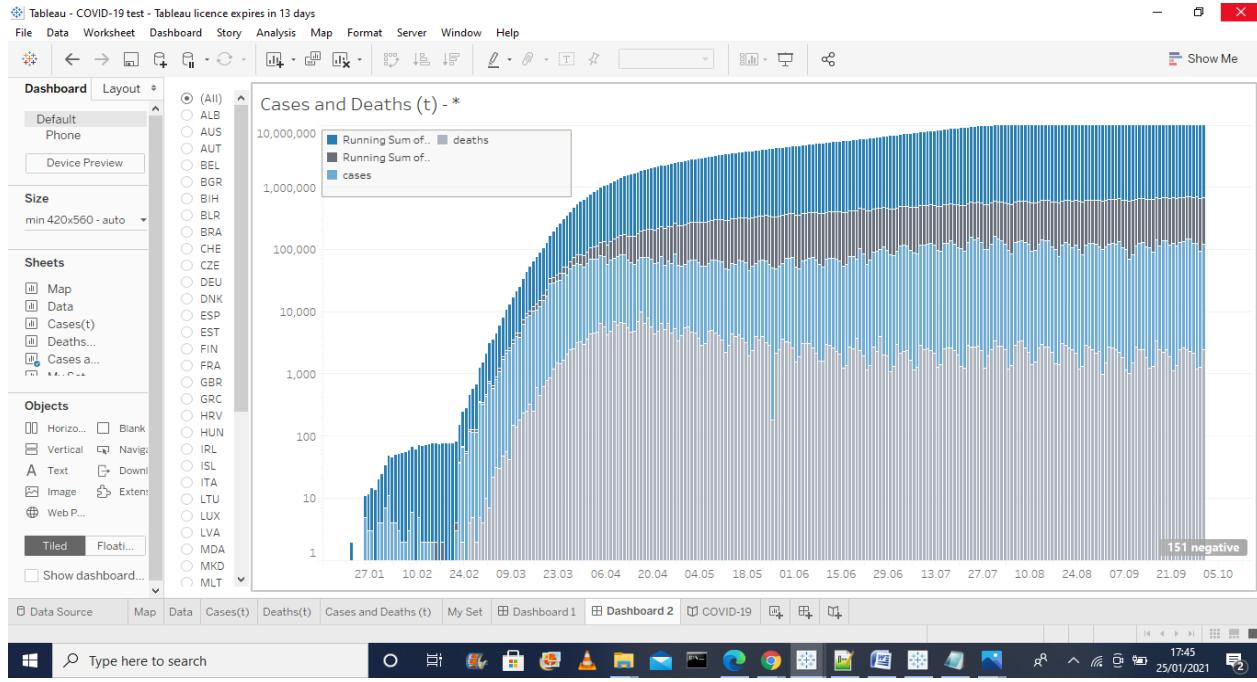
Consolidated Dashboard -1



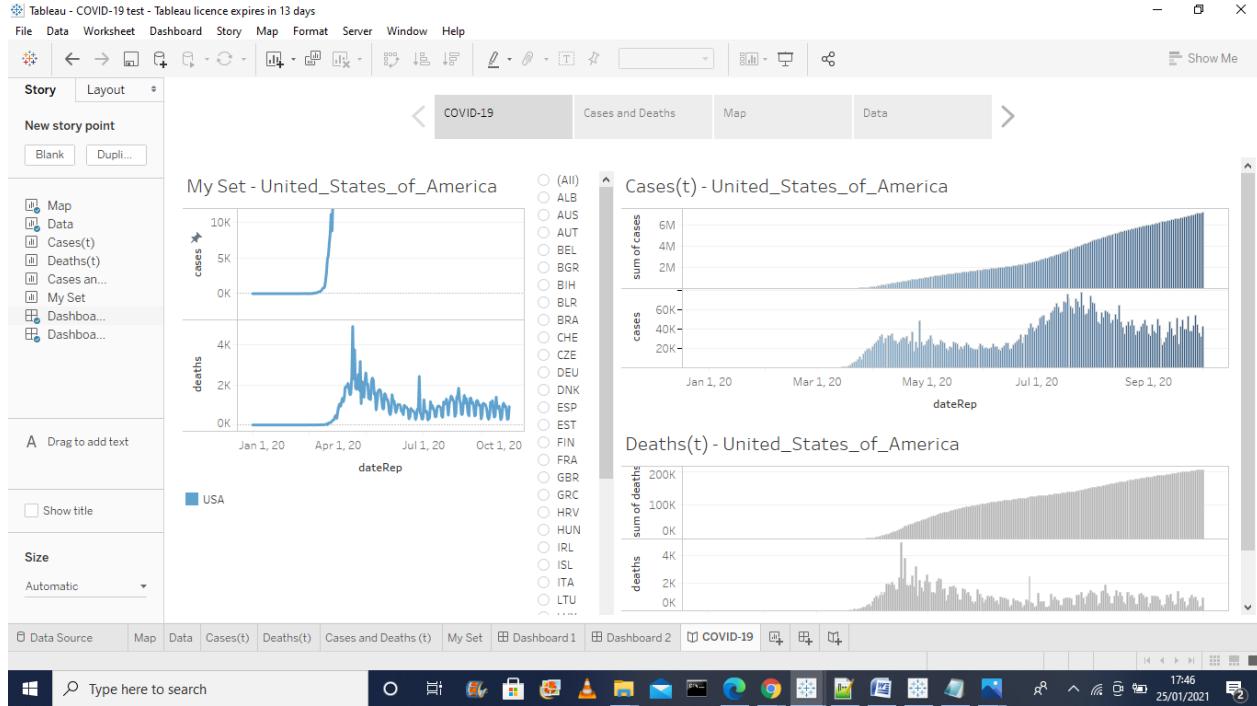
My Set Dashboard:



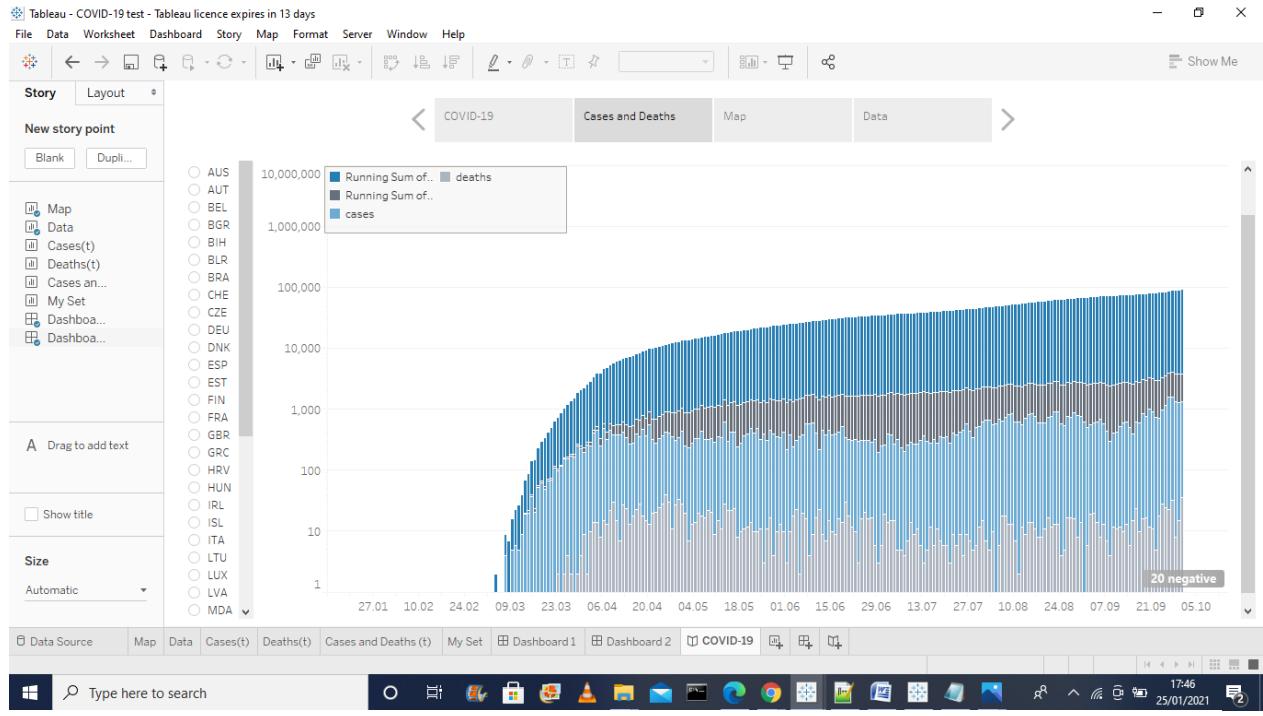
Cases & Deaths Dashboard2



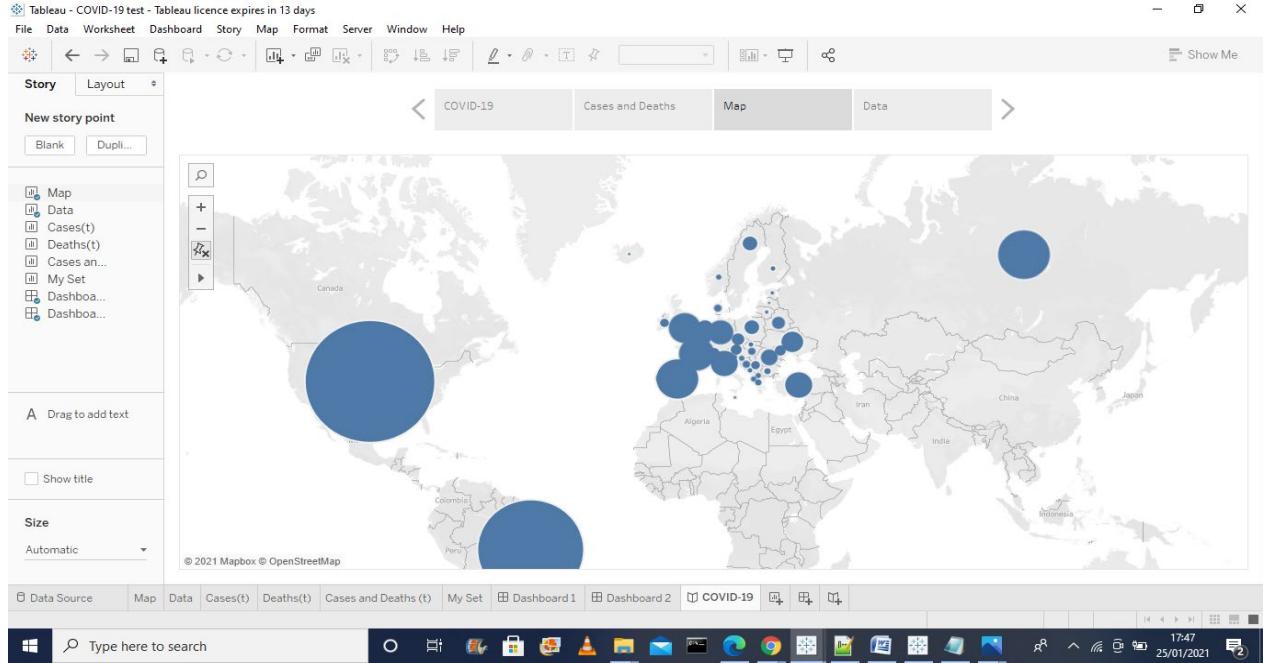
Covid-19 Dashboard for US:



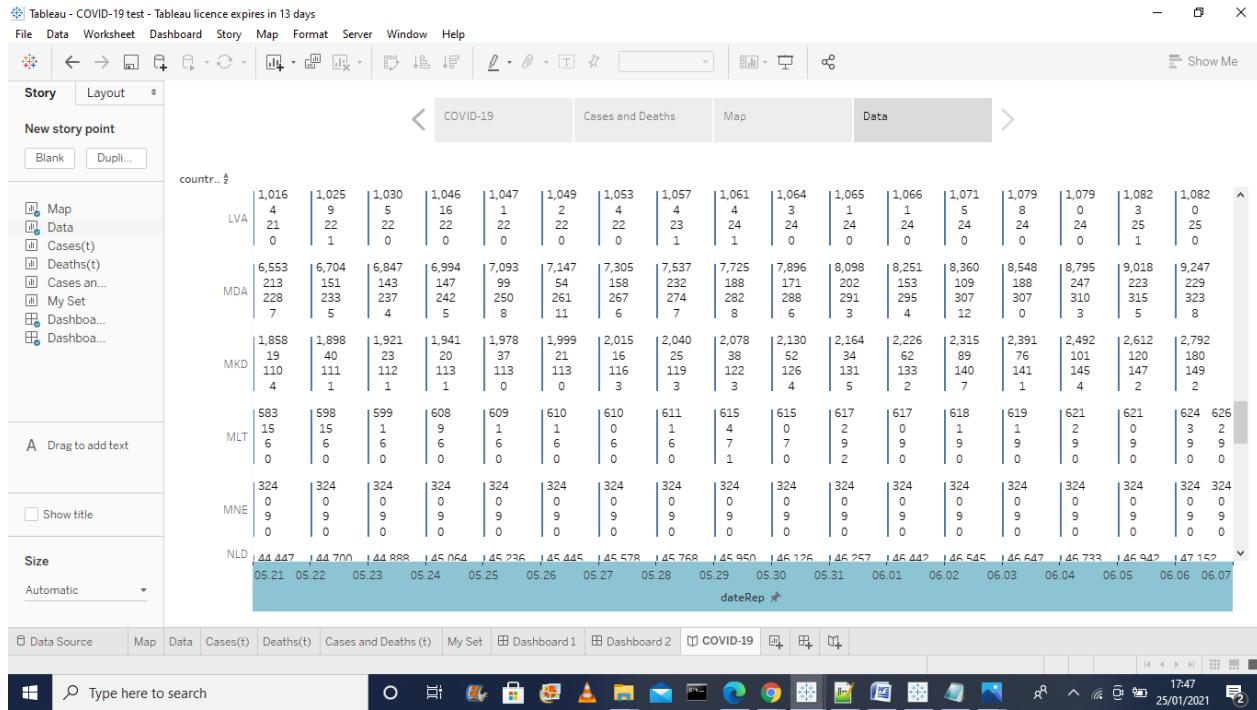
Covid-19 Dashboard for Cases & Deaths



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 analysis.

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. Without 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-datasource.
- 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 trial 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 \text{ 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:

Service	AWS Free Tier usage limit	Current usage	Forecasted usage	MTD actual usage %	MTD forecasted usage %
Amazon Simple Storage Service	2,000 Put, Copy, Post or List Requests of Amazon S3	1,055 Requests	1,258 Requests	52.75%	62.89%
Amazon Simple Storage Service	20,000 Get Requests of Amazon S3	1,524 Requests	1,817 Requests	7.62%	9.09%
Amazon Simple Storage Service	5 GB of Amazon S3 standard storage	0 GB-Mo	0 GB-Mo	4.94%	5.89%
AWS Key Management Service	20,000 free requests per month for AWS Key Management Service	145 Requests	173 Requests	0.73%	0.86%
AWS Glue	\$0 for AWS Glue Data Catalog requests under the free tier	959 Request	1,143 Request	0.10%	0.11%
AWS Lambda	400,000 seconds of compute time per month for AWS Lambda	72 seconds	86 seconds	0.02%	0.02%
AWS Lambda	1,000,000 free requests per month for AWS Lambda	58 Requests	69 Requests	0.01%	0.01%
Amazon Simple Queue Service	1,000,000 Requests of Amazon Simple Queue Service	17 Requests	20 Requests	0.00%	0.00%
AmazonCloudWatch	5 GB of Log Data Ingestion for Amazon Cloudwatch	0 GB	0 GB	0.00%	0.00%
Amazon Simple Notification Service	1,000,000 Requests for Amazon Simple Notification Service (USW2)	7 Requests	8 Requests	0.00%	0.00%

Billing & Cost Management Dashboard for this project implementation

Spend Summary

Welcome to the AWS Billing & Cost Management console. Your last month, month-to-date, and month-end forecasted costs appear below.

Current month-to-date balance for January 2021, the exchange rate for the Payment Currency is estimated.
 0.32 USD which converts to **23.58 INR** at today's exchange rate of 73.683338

Month-to-Date Spend by Service

The chart below shows the proportion of costs spent for each service you use.

Service	Cost
Glue	\$0.32
Athena	\$0.00
CloudWatch	\$0.00
DataTransfer	\$0.00
Other Services	\$0.00
Tax	\$0.05
Total	\$0.32

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.