

ML Incubator

**Data Engineering**

**Final Project**

Author:

Rakesh Sukumar

AWS Batch 2

Lab User ID: ugenpactaws26

Email: rakesh.sukumar@genpact.con

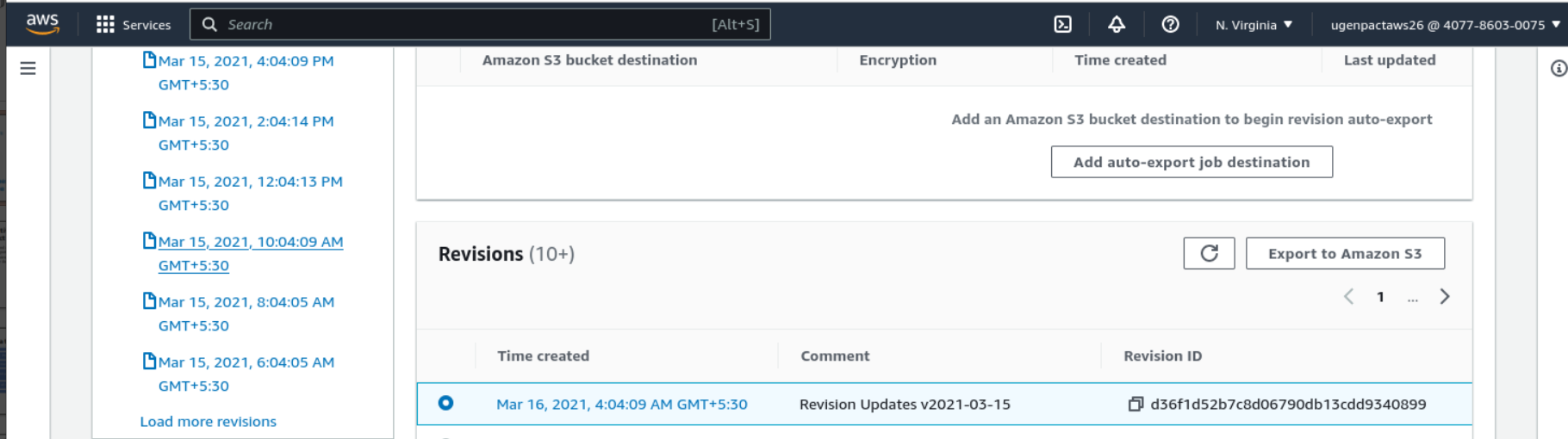
AWS Data Engineering Project

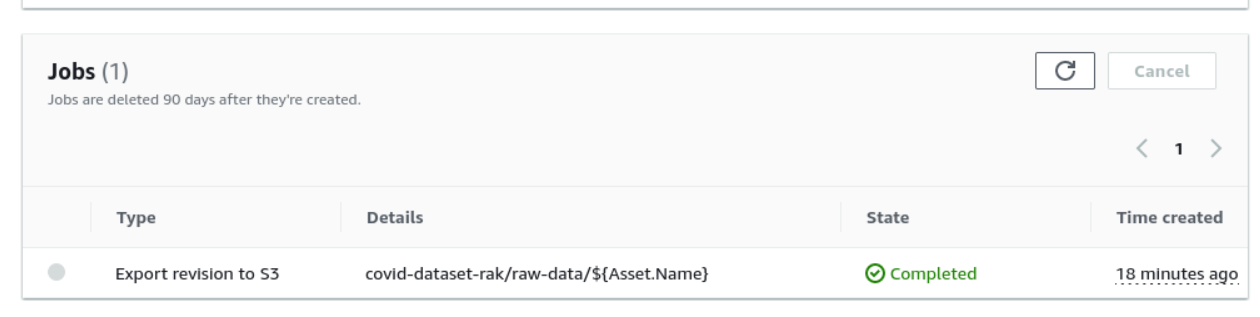
**Data Discovery , Processing , Loading Task**

**Task 1:**

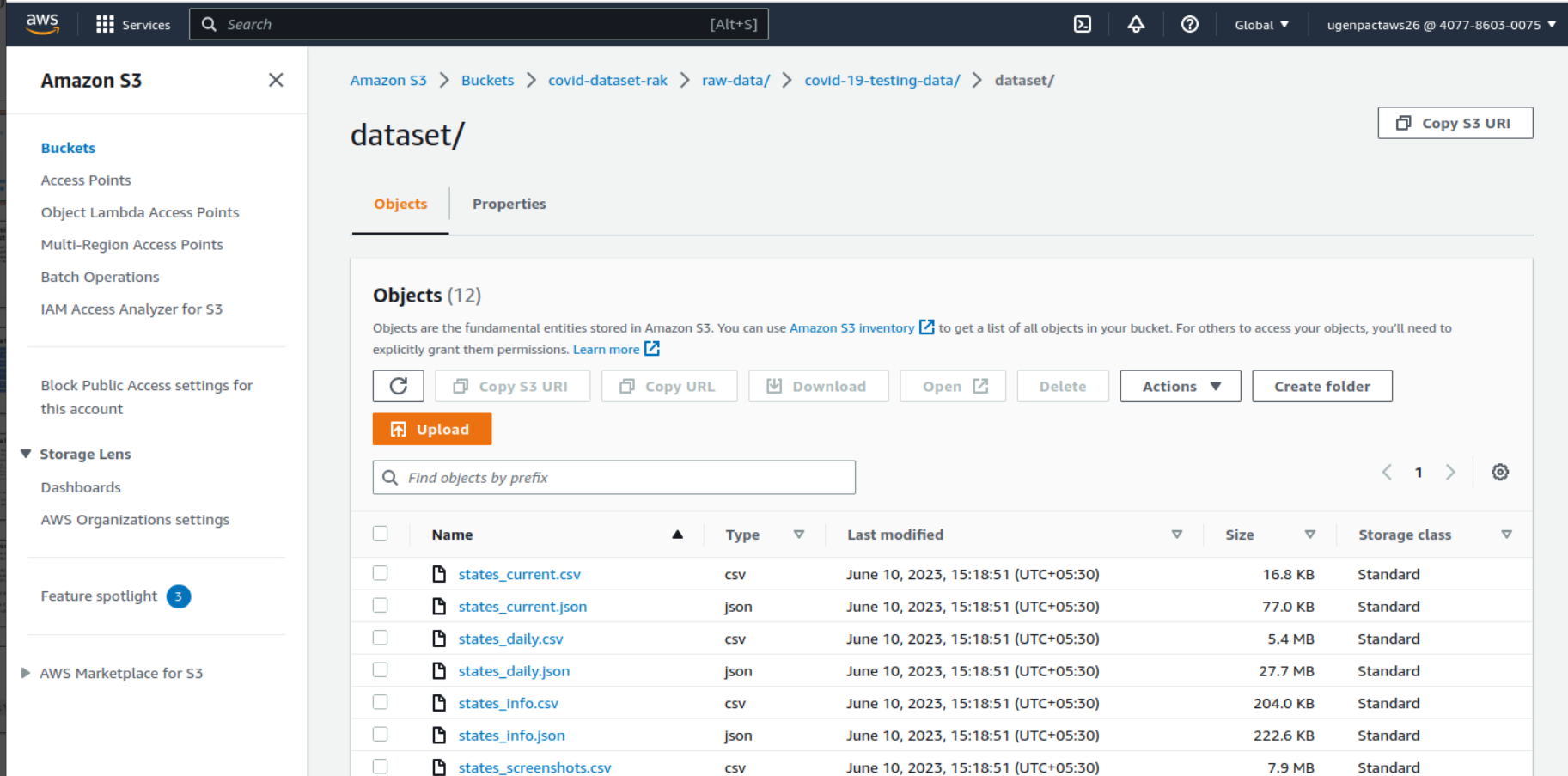
1. Load data to S3 bucket from AWS Data Exchange.

Exporting data from AWS Data Exchange:





Datasets available in S3 bucket



2. Download datafiles and upload to a new folder with in one-table-per-folder format.

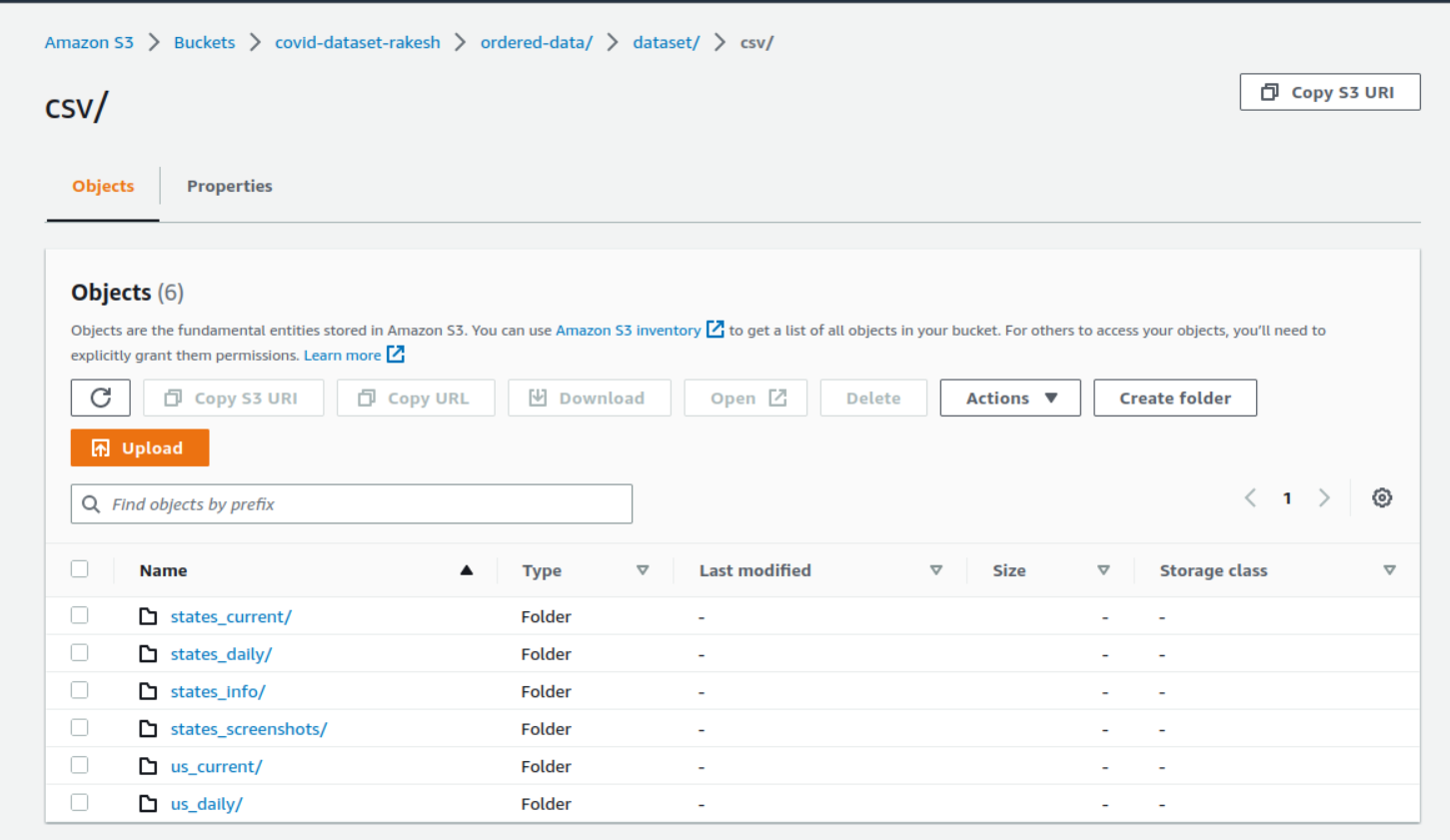
According to this [AWS re:POST post](https://repost.aws/knowledge-center/athena-empty-results), AWS Athena cannot retieve data if data for more than one tables are stored with the same s3 prefix. Hence I created separate folder for each table and upload the data to a new folder in the same bucket :

CSV Files:

* s3://covid-dataset-rakesh/ordered-data/dataset/csv/states\_current/states\_current.csv
* s3://covid-dataset-rakesh/ordered-data/dataset/csv/states\_daily/states\_daily.csv
* etc.

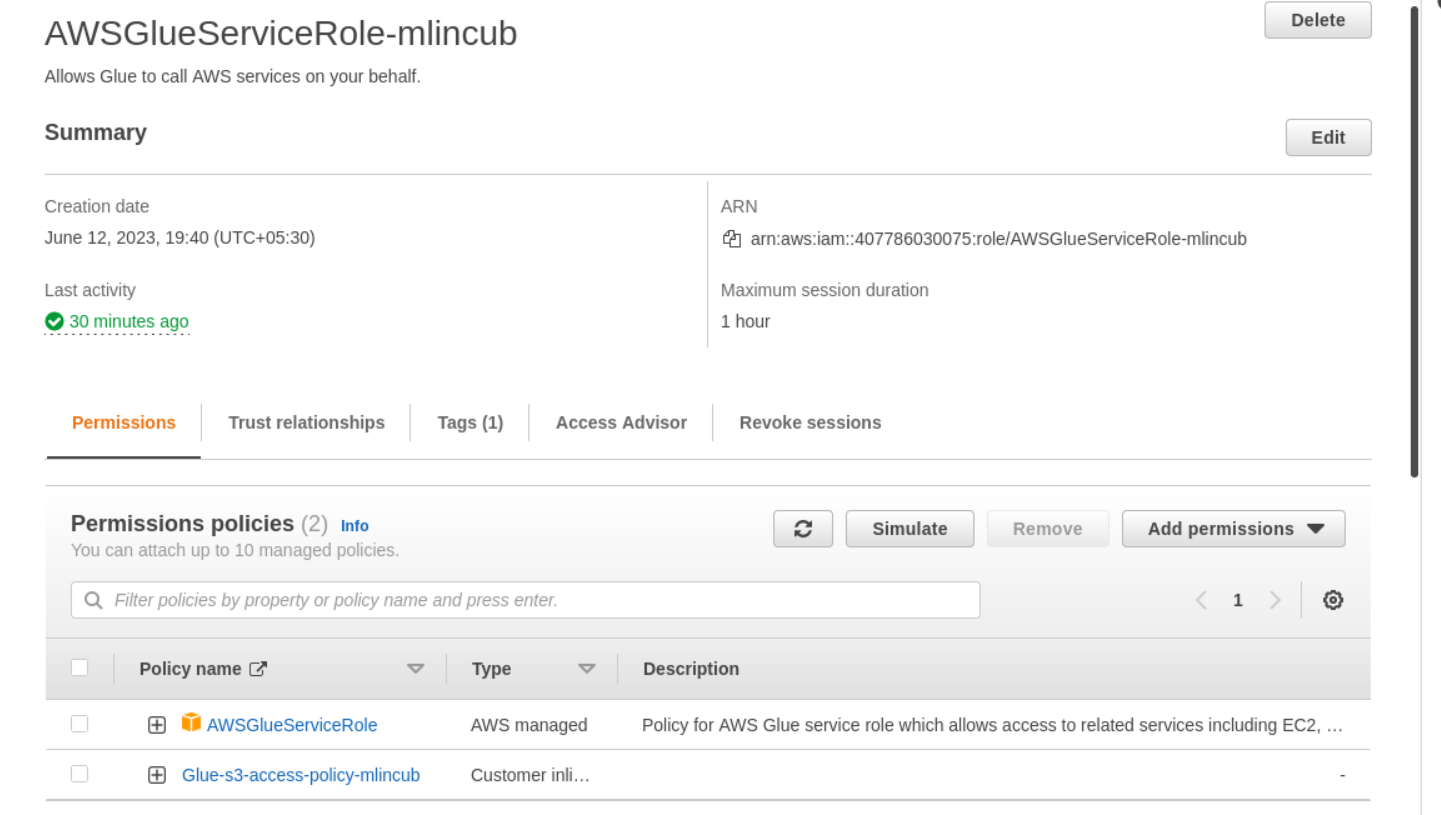
JSON Files:

* s3://covid-dataset-rakesh/ordered-data/dataset/json/states\_current/states\_current.json
* s3://covid-dataset-rakesh/ordered-data/dataset/json/states\_daily/states\_daily.json
* etc.



3. Open IAM and create a new for Glue to access S3

Create an IAM role with AWSGlueServiceRole policy and an inline policy with GetObject and PutObject access to the s3 bucket.



Inline policy JSON:

{

"Version": "2012-10-17",

"Statement": [

{

"Sid": "VisualEditor0",

"Effect": "Allow",

"Action": [

"s3:PutObject",

"s3:GetObject"

],

"Resource": "arn:aws:s3:::covid-dataset-rakesh/\*"

}

]

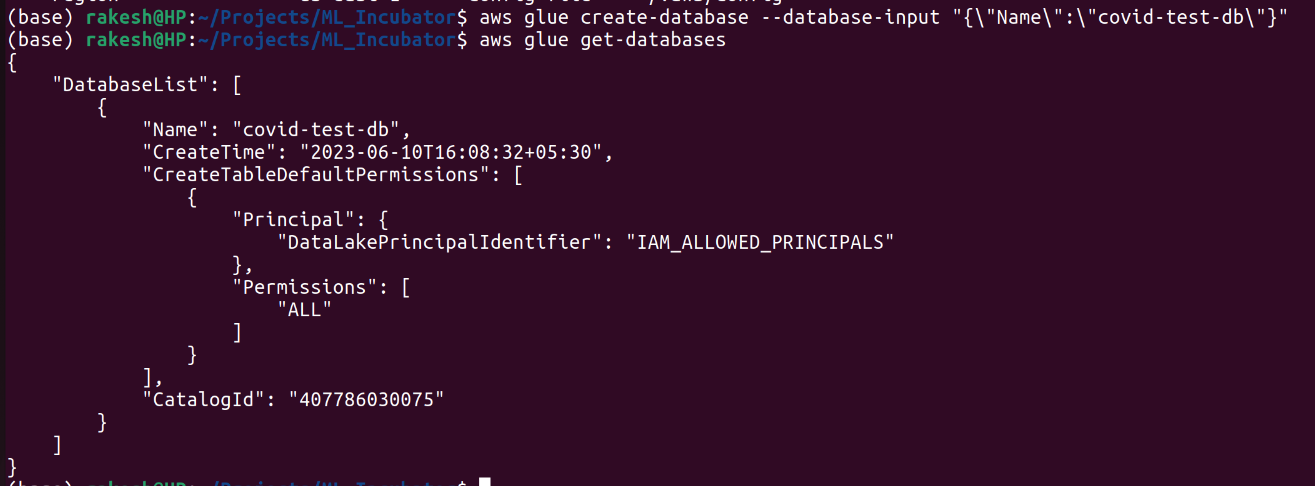
}

4. Create a Glue database from CLI

aws glue create-database --database-input "{\"Name\":\"covid-test-db\"}"

Use following command to view the database created:

aws glue get-databases



5. Create Crawler for csv files

aws glue create-crawler \

--name covid-s3-crawler-csv \

--role AWSGlueServiceRole-mlincub \

--database-name covid-test-db \

--targets "{\"S3Targets\": [\

{\

\"Path\": \"s3://covid-dataset-rakesh/ordered-data/dataset\",\

\"Exclusions\": [\"\*\*.json\"] \

}\

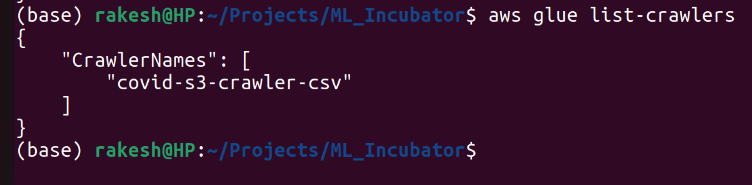
]}" \

--recrawl-policy "{\"RecrawlBehavior\": \"CRAWL\_NEW\_FOLDERS\_ONLY\"}" \

--tags "{\"KeyName\": \"project\", \"Value\": \"mlincub\"}"

Check the crawler:

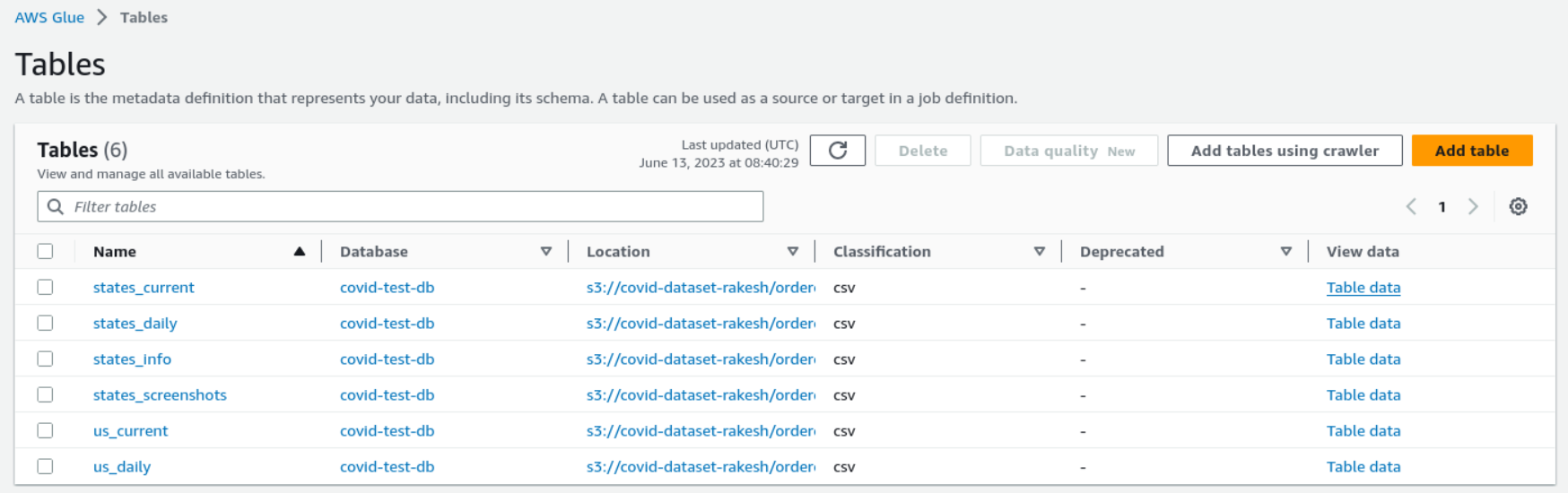
aws glue list-crawlers



Run the crawler:

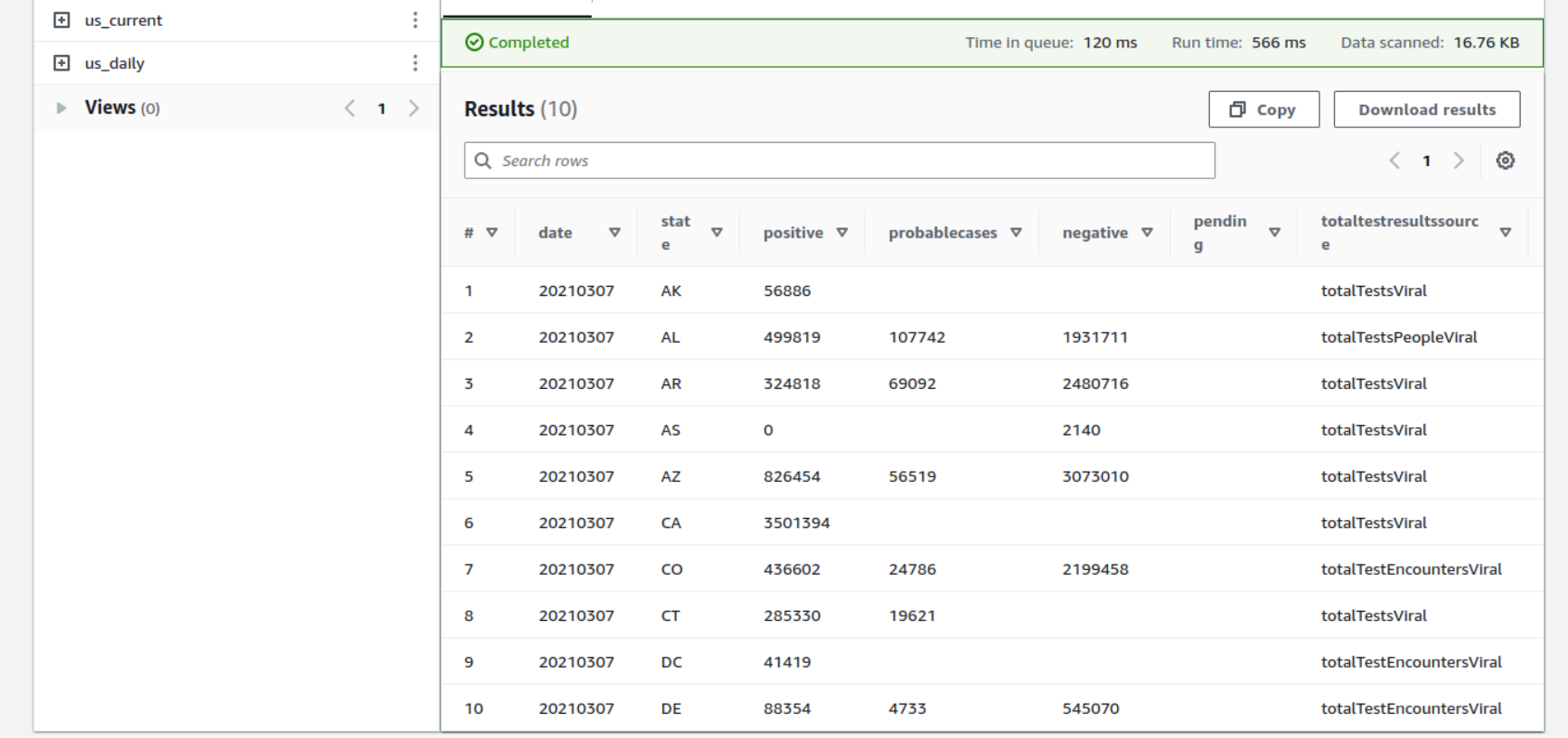
aws glue start-crawler --name covid-s3-crawler-csv

CSV Tables created



Data Queried from Athena

SELECT \* FROM "AwsDataCatalog"."covid-test-db"."states\_current" limit 10;



6. Create crawler for JSON files and Run the crawler:

Creating json using the below command created a single table with the name json\_json and considered individual table folders as partitions. Hence I added the below crawler configuration parameter to create-crawler command which created 6 tables as expected.

aws glue create-crawler \

--name covid-s3-crawler-json \

--role AWSGlueServiceRole-mlincub \

--database-name covid-test-db \

--targets "{\"S3Targets\": [\

{\

\"Path\": \"s3://covid-dataset-rakesh/ordered-data/dataset\",\

\"Exclusions\": [\"\*\*.csv\"] \

}\

]}" \

--table-prefix json\_ \

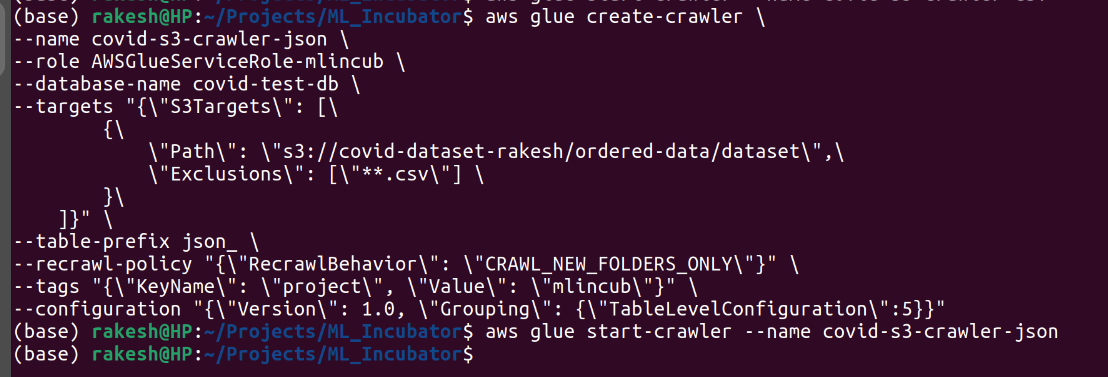
--recrawl-policy "{\"RecrawlBehavior\": \"CRAWL\_NEW\_FOLDERS\_ONLY\"}" \

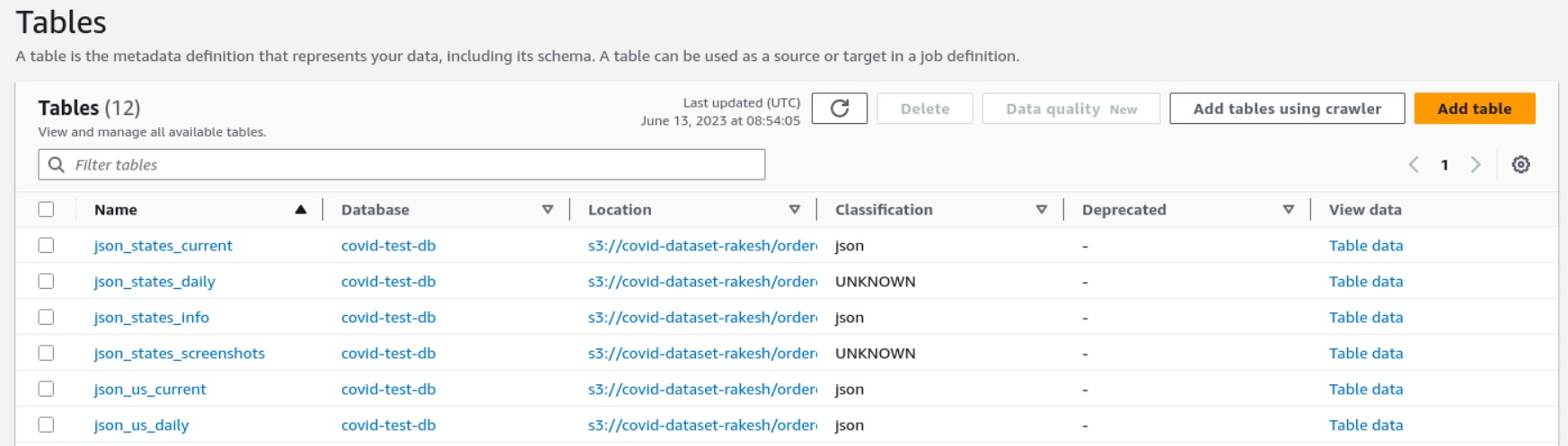
--tags "{\"KeyName\": \"project\", \"Value\": \"mlincub\"}" \

--configuration "{\"Version\": 1.0, \"Grouping\": {\"TableLevelConfiguration\":5}}"

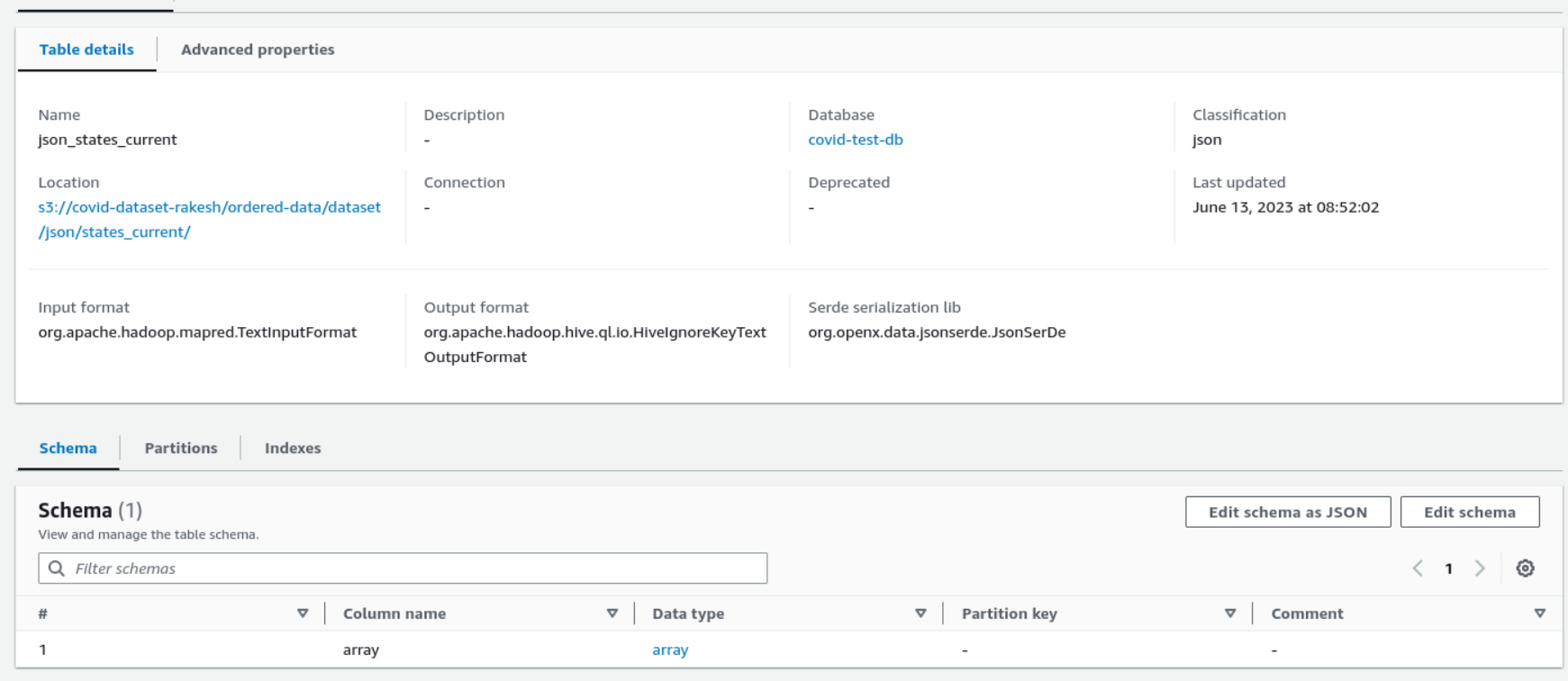
# Run the crawler for json files

aws glue start-crawler --name covid-s3-crawler-json





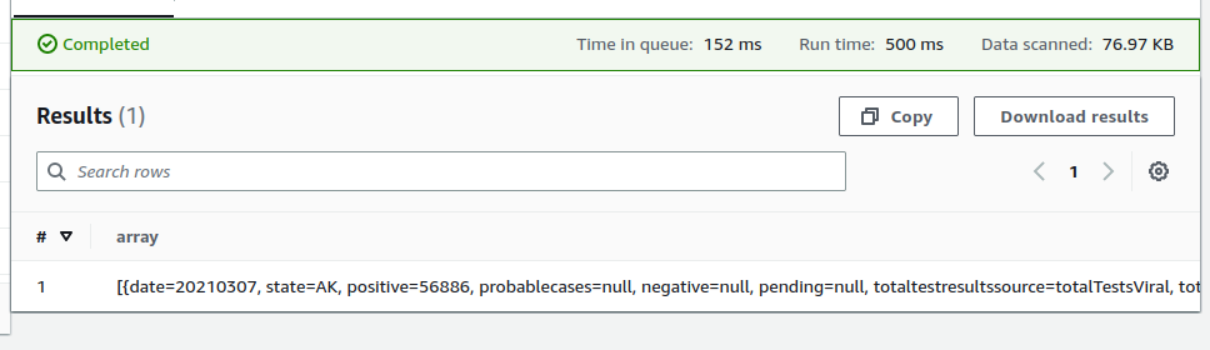
The json crawler could not identify the file types for all input files. Also, the schema identified has one column of type array.



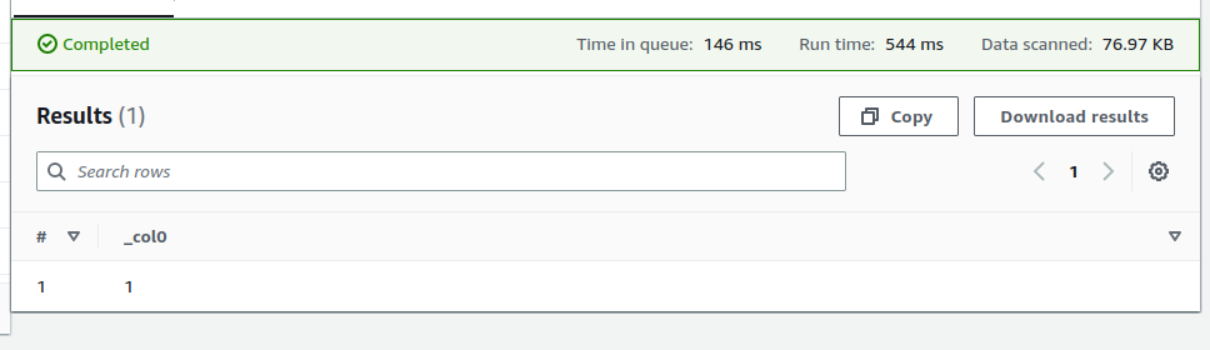
Athena could not retrieve data for tables with unknown classification above.

Athena retrived all data as a single record for tables with classification “json”.

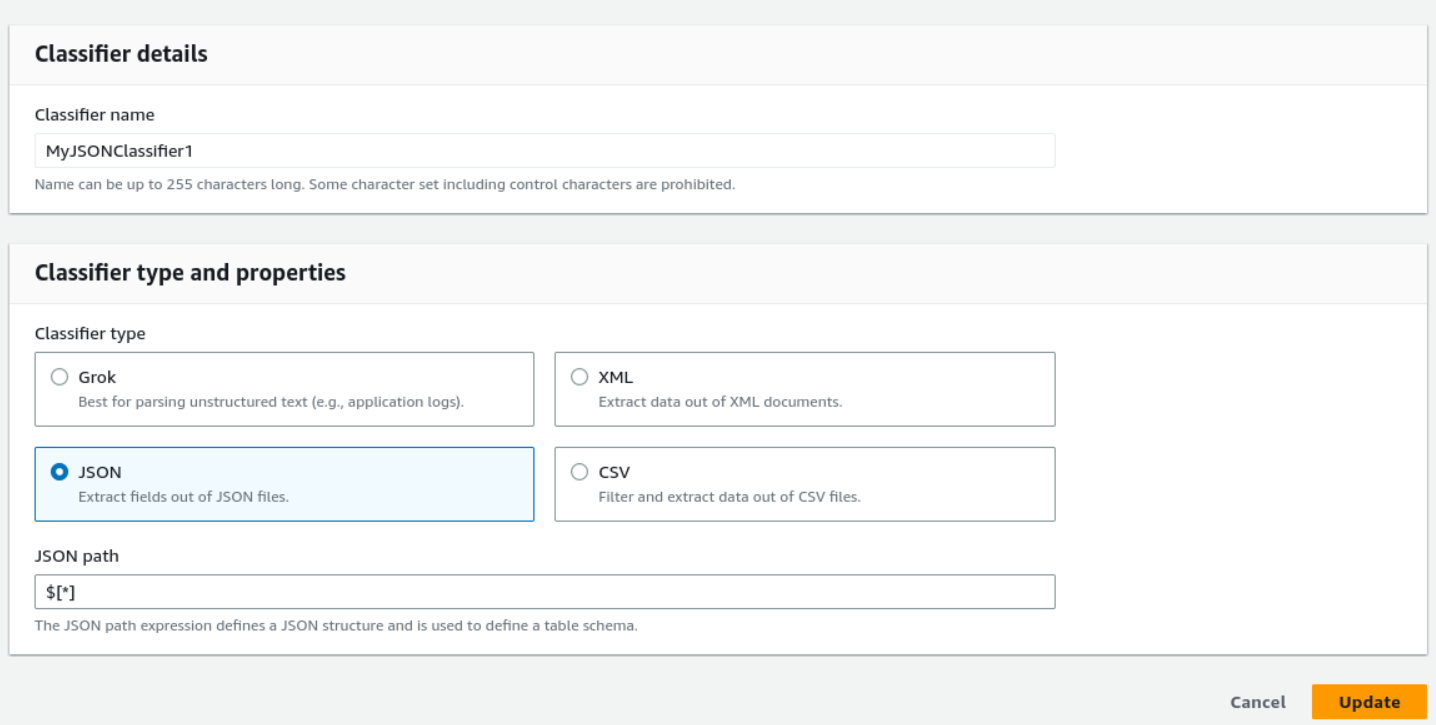
Athena Query: SELECT \* FROM "AwsDataCatalog"."covid-test-db"."json\_states\_current" limit 10;



Athena Query: SELECT COUNT(\*) FROM "AwsDataCatalog"."covid-test-db"."json\_states\_current";



7. Glue crawler could identify the schema correctly with a custom json classifier.



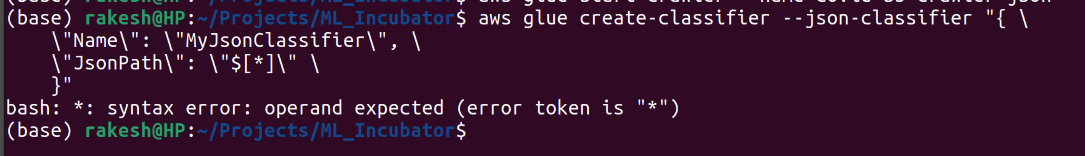
Note: I created the above classifier from Console and the CLI command to create the classifier was failing with some strange error. I couldn’t find any useful reference to fix this issue. Also, I felt the creation of custom json classifier from CLI is not well documented.

aws glue create-classifier --json-classifier "{ \

\"Name\": \"MyJsonClassifier1\", \

\"JsonPath\": \"$[\*]\" \

}"



I create a separate database is Glue to test the custom classifier.

aws glue create-database --database-input "{\"Name\":\"covid-test-db-new\"}"

Create json crawler with custom classifier

aws glue create-crawler \

--name covid-s3-crawler-json-cc \

--role AWSGlueServiceRole-mlincub \

--database-name covid-test-db-new \

--targets "{\"S3Targets\": [\

{\

\"Path\": \"s3://covid-dataset-rakesh/ordered-data/dataset\",\

\"Exclusions\": [\"\*\*.csv\"] \

}\

]}" \

--table-prefix json\_ \

--recrawl-policy "{\"RecrawlBehavior\": \"CRAWL\_NEW\_FOLDERS\_ONLY\"}" \

--tags "{\"KeyName\": \"project\", \"Value\": \"mlincub\"}" \

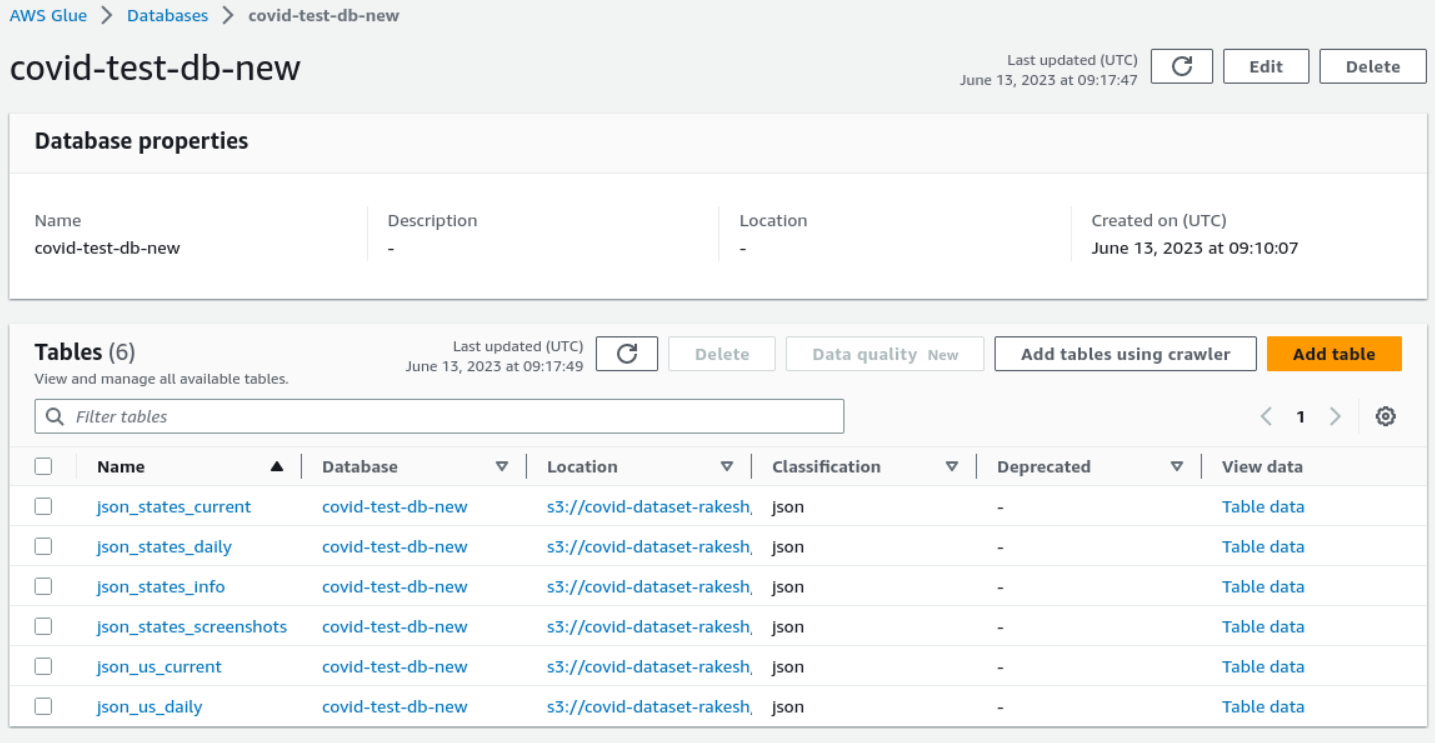
--configuration "{\"Version\": 1.0, \"Grouping\": {\"TableLevelConfiguration\":5}}" \

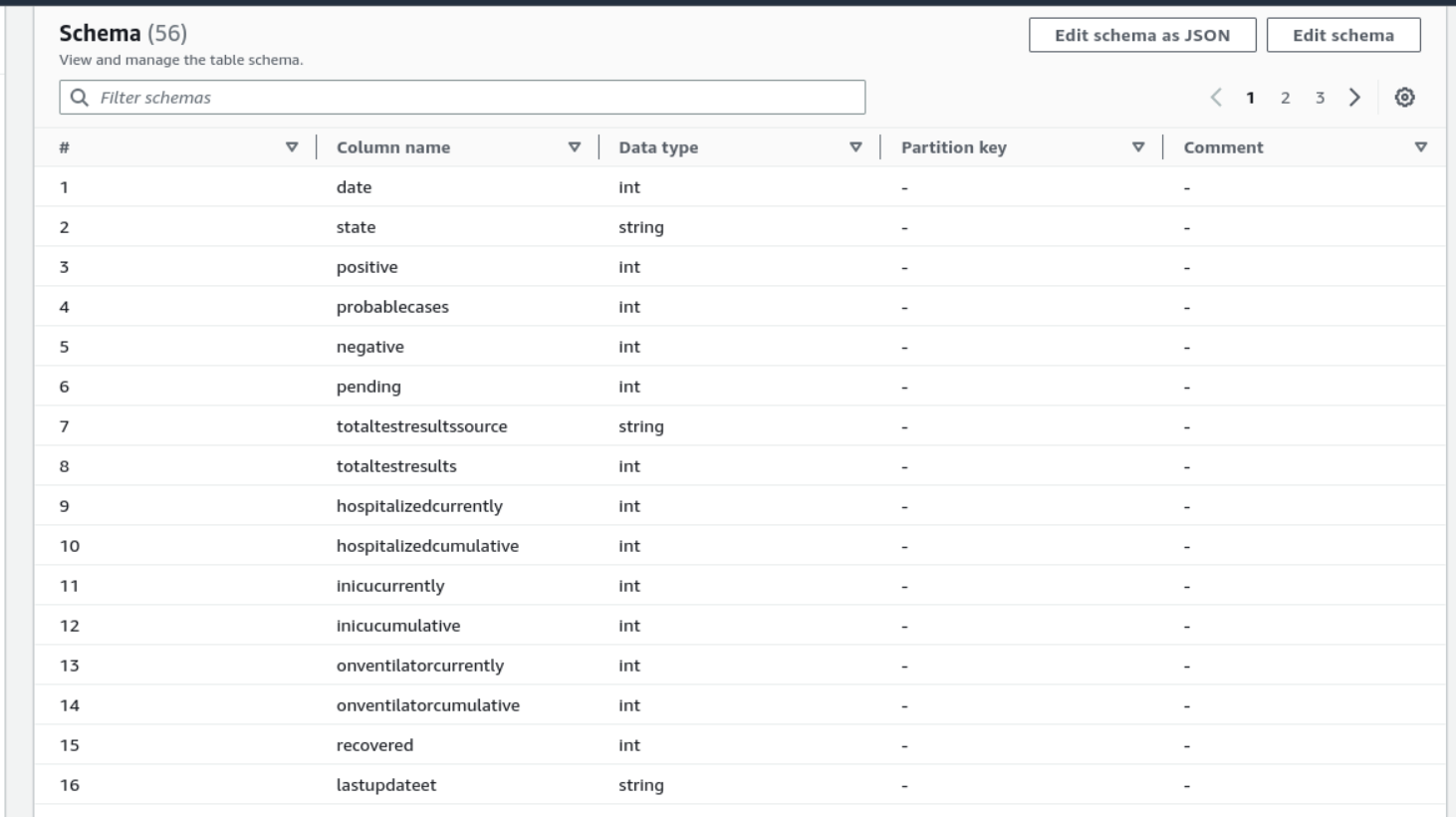
--classifiers MyJSONClassifier1

Run the crawler

aws glue start-crawler --name covid-s3-crawler-json-cc

The Glue crawler is able to classify the files create schemas correctly.

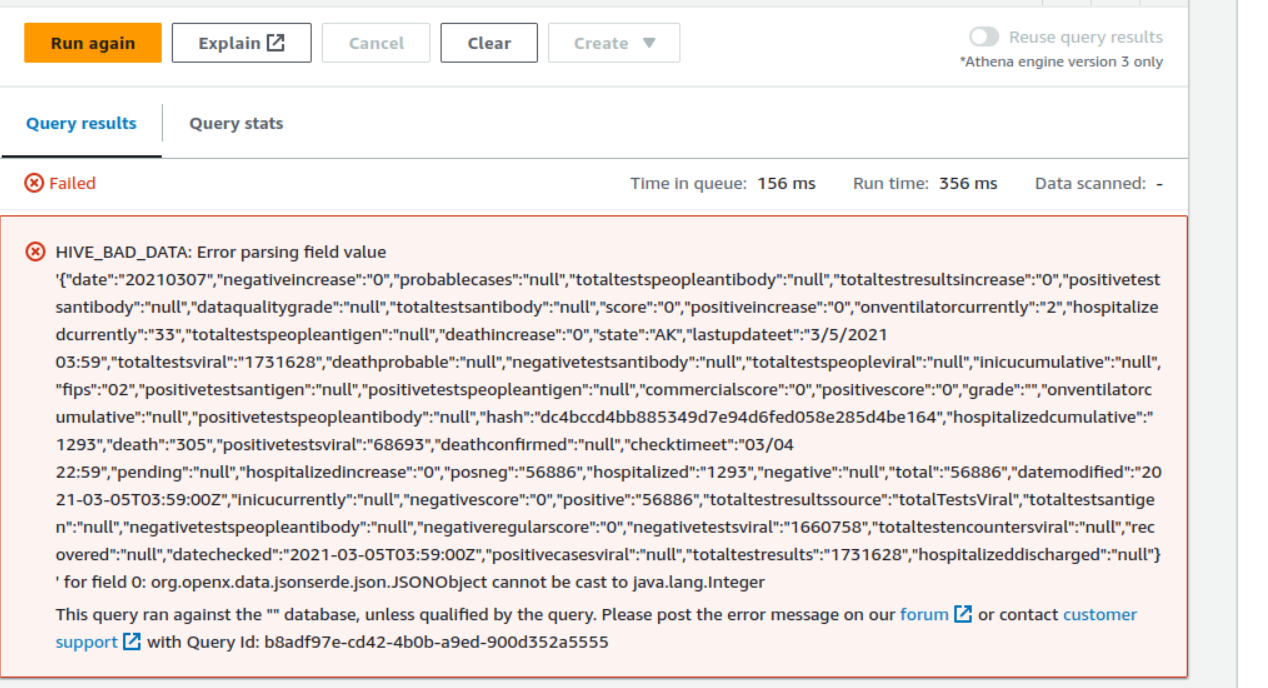




However, we are not able to query the data from Athena still.

Athena Query:

SELECT \* FROM "AwsDataCatalog"."covid-test-db-new"."json\_states\_current" limit 10;



**Task 2:**

1. Pull the latest Glue docker image:

docker pull amazon/aws-glue-libs:glue\_libs\_4.0.0\_image\_01

2. Set AWS\_PROFILE and JUPYTER\_WORKSPACE\_LOCATION environment variables:

export AWS\_PROFILE=default

export AWS\_DEFAULT\_REGION=us-east-1

export JUPYTER\_WORKSPACE\_LOCATION=/home/rakesh/Projects/ML\_Incubator/glue\_workspace

3. Run the Glue docker image

docker run -it -v ~/.aws:/home/glue\_user/.aws \

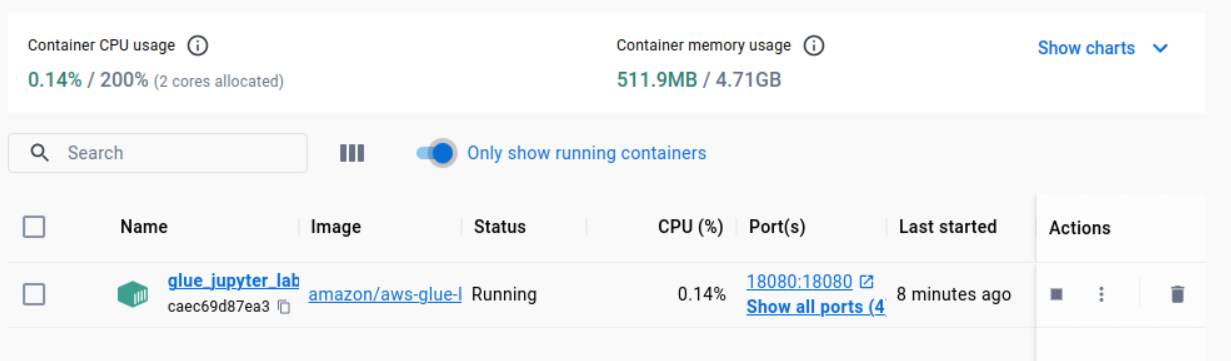
-v $JUPYTER\_WORKSPACE\_LOCATION:/home/glue\_user/workspace/jupyter\_workspace/ \

-e AWS\_PROFILE=$AWS\_PROFILE -e DISABLE\_SSL=true --rm -p 4040:4040 -p 18080:18080 \

-p 8998:8998 -p 8888:8888 \

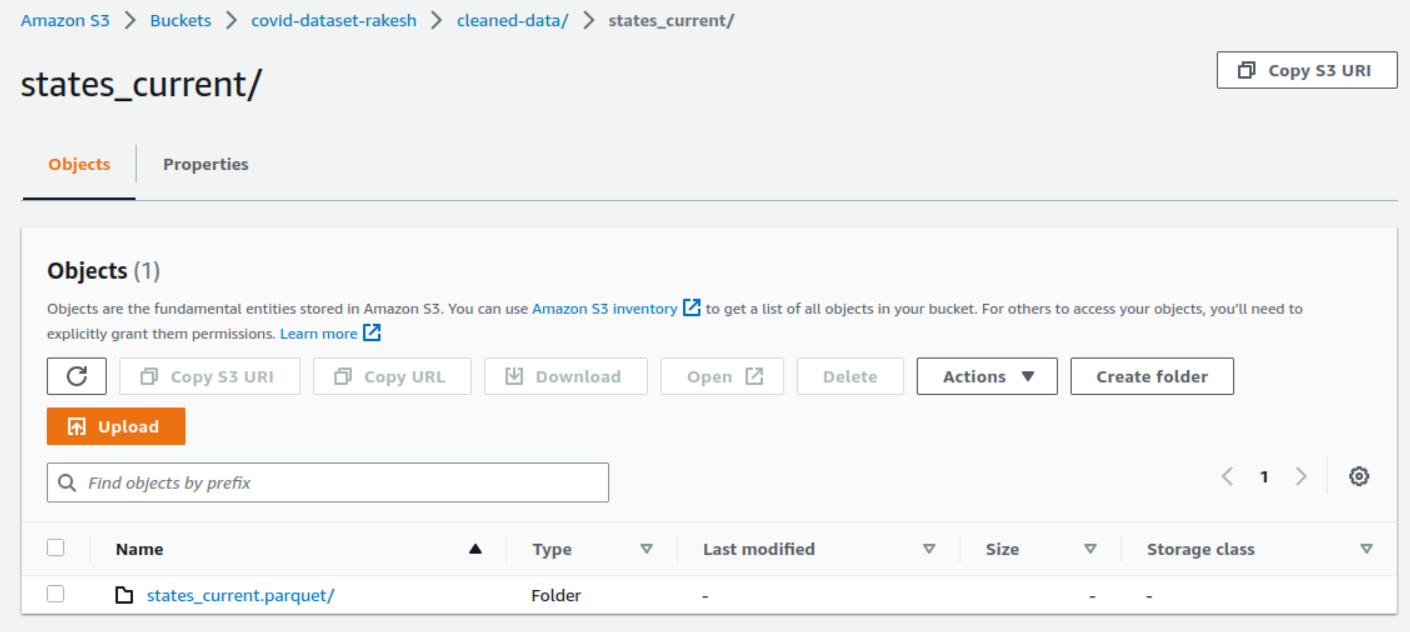
--name glue\_jupyter\_lab amazon/aws-glue-libs:glue\_libs\_4.0.0\_image\_01 \

/home/glue\_user/jupyter/jupyter\_start.sh



4. Open http://127.0.0.1:8888/lab and choose Glue Spark Local (PySpark) under Notebook.

Develop a jupyter notebook to read csv files from s3, correct the table schema and save the data back to s3 in parquet format.



5. Create a new database and crawlers for parquet files

Create a new database:

aws glue create-database --database-input "{\"Name\":\"covid\_parquet\_db\"}"

Create crawler:

aws glue create-crawler \

--name covid-s3-crawler-parquet \

--role AWSGlueServiceRole-mlincub \

--database-name covid\_parquet\_db \

--targets "{\"S3Targets\": [\

{\

\"Path\": \"s3://covid-dataset-rakesh/cleaned-data/\"\

}\

]}" \

--table-prefix par\_ \

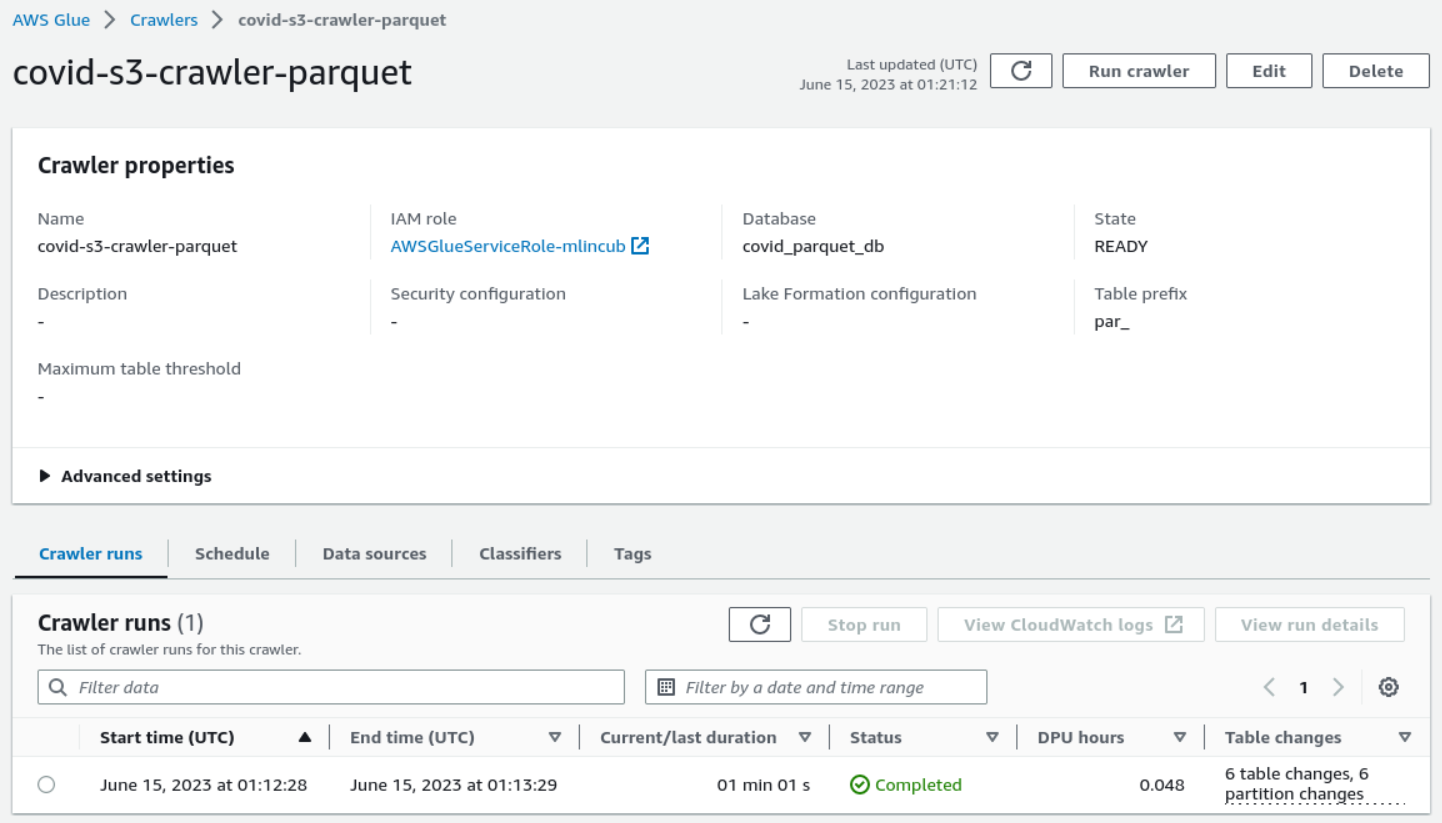
--recrawl-policy "{\"RecrawlBehavior\": \"CRAWL\_NEW\_FOLDERS\_ONLY\"}" \

--tags "{\"KeyName\": \"project\", \"Value\": \"mlincub\"}" \

--configuration "{\"Version\": 1.0, \"Grouping\": {\"TableLevelConfiguration\":3}}"

Run the crawler:

aws glue start-crawler --name covid-s3-crawler-parquet



Query the data from Athena

-- SELECT top 3 months with highest average covid positive cases

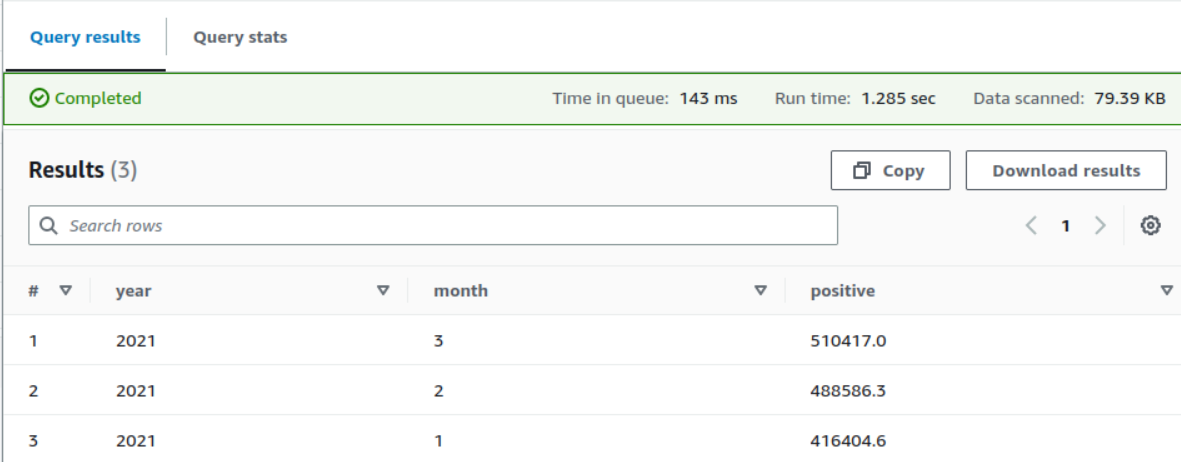
SELECT YEAR(date) AS year, MONTH(date) AS month, ROUND(AVG(positive), 1) AS positive

FROM "AwsDataCatalog"."covid\_parquet\_db"."par\_states\_daily"

GROUP BY YEAR(date), MONTH(date)

ORDER BY positive DESC

LIMIT 3;



-- Select top 5 days with highest test positivity rate (by states)

SELECT A.state, B.name, date, positiveTestsViral, totalTestsViral,

(CAST(positiveTestsViral AS REAL)/CAST(totalTestsViral AS REAL)) AS positivityRate

FROM "AwsDataCatalog"."covid\_parquet\_db"."par\_states\_daily" AS A

LEFT JOIN (SELECT DISTINCT state, name

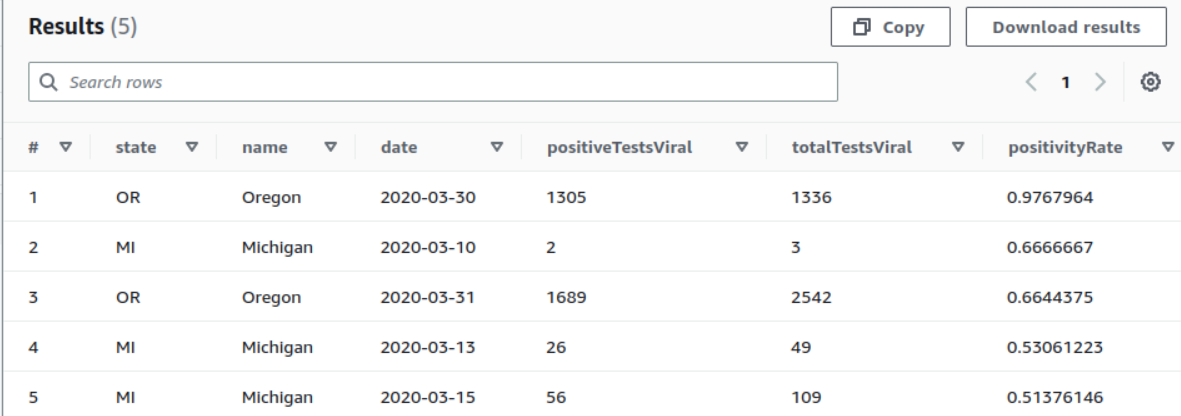
FROM "AwsDataCatalog"."covid\_parquet\_db"."par\_states\_info") AS B

ON A.state = B.state

WHERE totalTestsViral > 0

ORDER BY positivityRate DESC

LIMIT 5;



# Select top 5 states with highest monthly fatality rates. Consider only states with at-least 100 death in the month:

SELECT A.state, YEAR(date) AS year, MONTH(date) AS month,

SUM(positiveIncrease) AS positive, SUM(deathIncrease) AS death,

(CAST(SUM(deathIncrease) AS REAL)/CAST(SUM(positiveIncrease) AS REAL)) AS fatalityRate

FROM "AwsDataCatalog"."covid\_parquet\_db"."par\_states\_daily" AS A

LEFT JOIN (SELECT DISTINCT state, name

FROM "AwsDataCatalog"."covid\_parquet\_db"."par\_states\_info") AS B

ON A.state = B.state

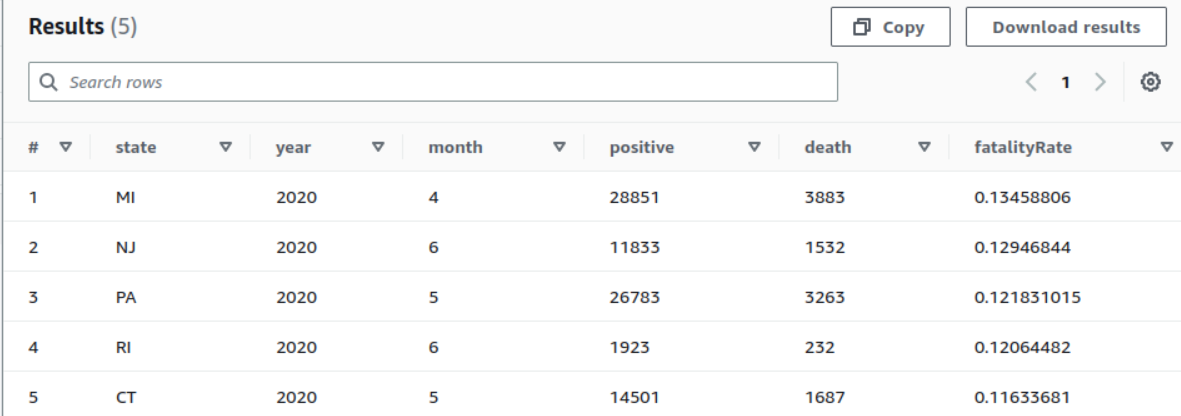
GROUP BY A.state, YEAR(date), MONTH(date)

HAVING SUM(positiveIncrease) > 0

AND SUM(deathIncrease) > 100

ORDER BY fatalityRate DESC

LIMIT 5;



Git History Screenshot

