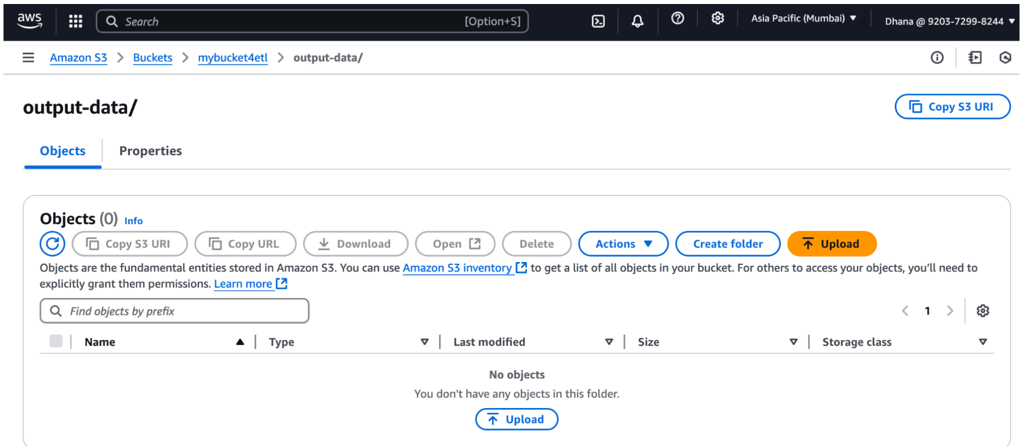
For real-time streaming, we are using **Kinesis Data Generator**.

First, we will create an S3 folder named **output-data**, as shown below. Next, we will create an IAM role with full access to **AWS Glue**, **Kinesis**, **Athena**, and **S3**.



output-data/

Objects (0) Info

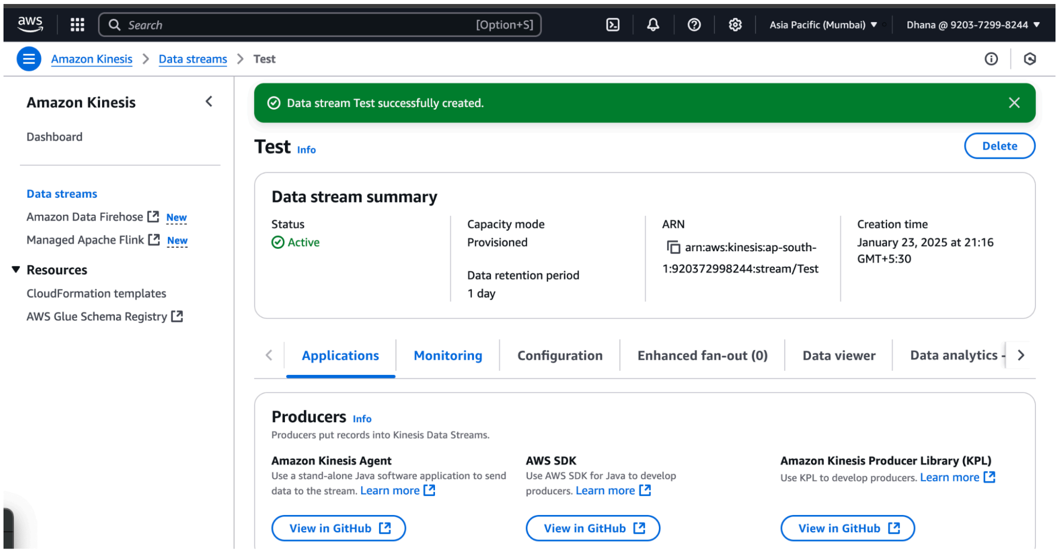
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name ▲	Type ▽	Last modified ▽	Size ▽	Storage class ▽

No objects

You don't have any objects in this folder.

We will create a **Kinesis Data Stream** for real-time data streaming.



Data stream Test successfully created.

Test Info

Data stream summary

Status
Active

Capacity mode
Provisioned

Data retention period
1 day

ARN
arn:aws:kinesis:ap-south-1:920372998244:stream/Test

Creation time
January 23, 2025 at 21:16 GMT+5:30

Producers Info

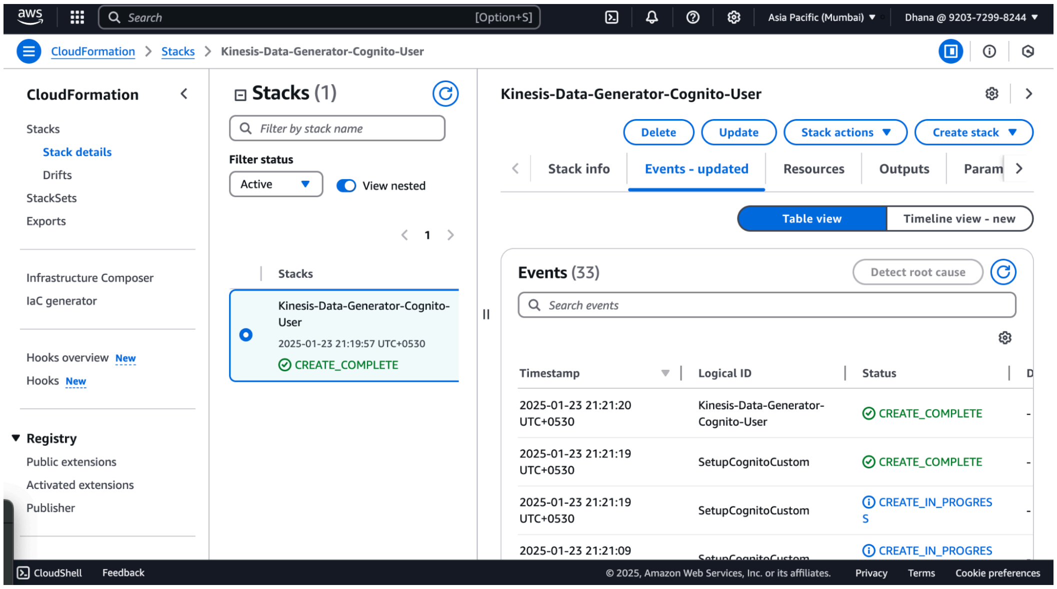
Producers put records into Kinesis Data Streams.

Amazon Kinesis Agent
Use a stand-alone Java software application to send data to the stream. Learn more

AWS SDK
Use AWS SDK for Java to develop producers. Learn more

Amazon Kinesis Producer Library (KPL)
Use KPL to develop producers. Learn more

Next, we will create a **Cognito user** for **Kinesis Data Generator** using a CloudFormation stack. You can use the following link to create it: https://awslabs.github.io/amazon-kinesis-data-generator/web/help.html



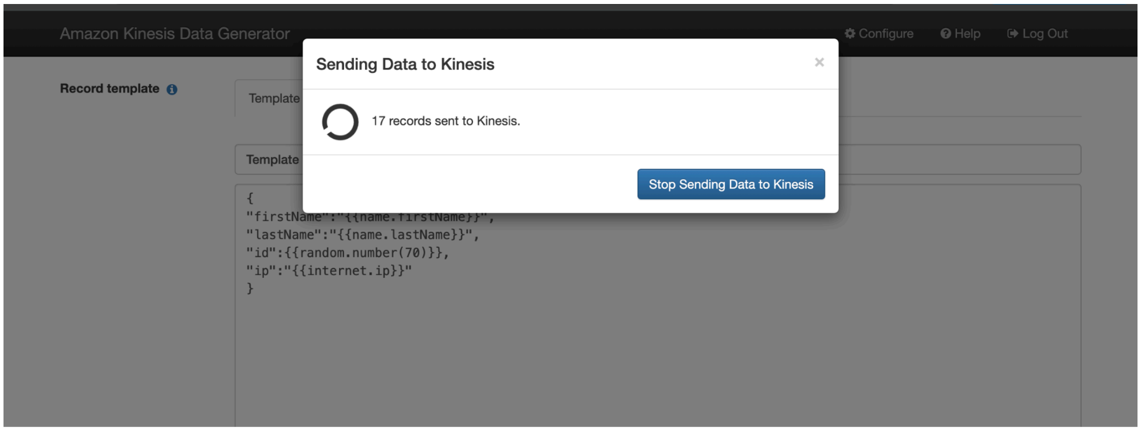
Kinesis-Data-Generator-Cognito-User
2025-01-23 21:19:57 UTC+0530
CREATE_COMPLETE

Kinesis-Data-Generator-Cognito-User

Events (33)

Timestamp	Logical ID	Status	D
2025-01-23 21:21:20 UTC+0530	Kinesis-Data-Generator-Cognito-User	CREATE_COMPLETE	-
2025-01-23 21:21:19 UTC+0530	SetupCognitoCustom	CREATE_COMPLETE	-
2025-01-23 21:21:19 UTC+0530	SetupCognitoCustom	CREATE_IN_PROGRESS	-
2025-01-23 21:21:09	SetupCognitoCustom	CREATE_IN_PROGRESS	

Once the stack creation is complete, log in to **Kinesis Data Generator using the link from Resources tab in CloudFormation** and start streaming data.
Use the following code for data generation:
{
"firstName":"{{name.firstName}}",
"lastName":"{{name.lastName}}",
"id":{{random.number(70)}},
"ip":"{{internet.ip}}"
}

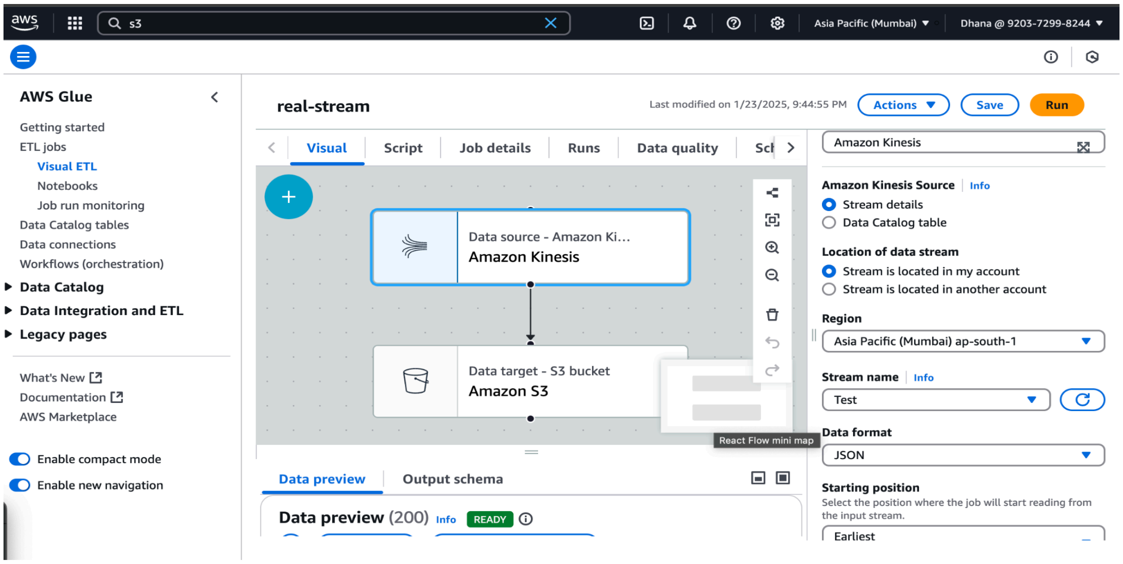


Sending Data to Kinesis

17 records sent to Kinesis.

```
{
"firstName":"{{name.firstName}}",
"lastName":"{{name.lastName}}",
"id":{{random.number(70)}},
"ip":"{{internet.ip}}"
}
```

Now, create a **Glue job** with the **source** set to the Kinesis Data Stream we created, and the **target** set to the S3 location.



real-stream

Last modified on 1/23/2025, 9:44:55 PM

Amazon Kinesis

Amazon Kinesis Source Info

- ● Stream details
- ○ Data Catalog table

Location of data stream

- ● Stream is located in my account
- ○ Stream is located in another account

Region

Asia Pacific (Mumbai) ap-south-1

Stream name Info

Test

Data format

JSON

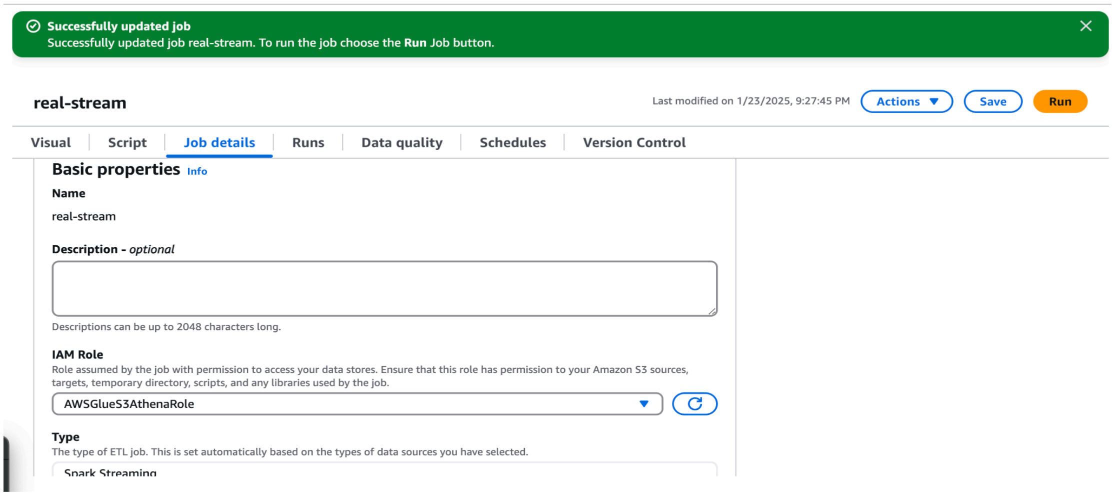
Starting position

Select the position where the job will start reading from the input stream.

Earliest

Data preview (200)

Edit job Details and **Run** the job.



Successfully updated job

Successfully updated job real-stream. To run the job choose the **Run** Job button.

real-stream

Last modified on 1/23/2025, 9:27:45 PM

Basic properties

Name

real-stream

Description - *optional*

Descriptions can be up to 2048 characters long.

IAM Role

Role assumed by the job with permission to access your data stores. Ensure that this role has permission to your Amazon S3 sources, targets, temporary directory, scripts, and any libraries used by the job.

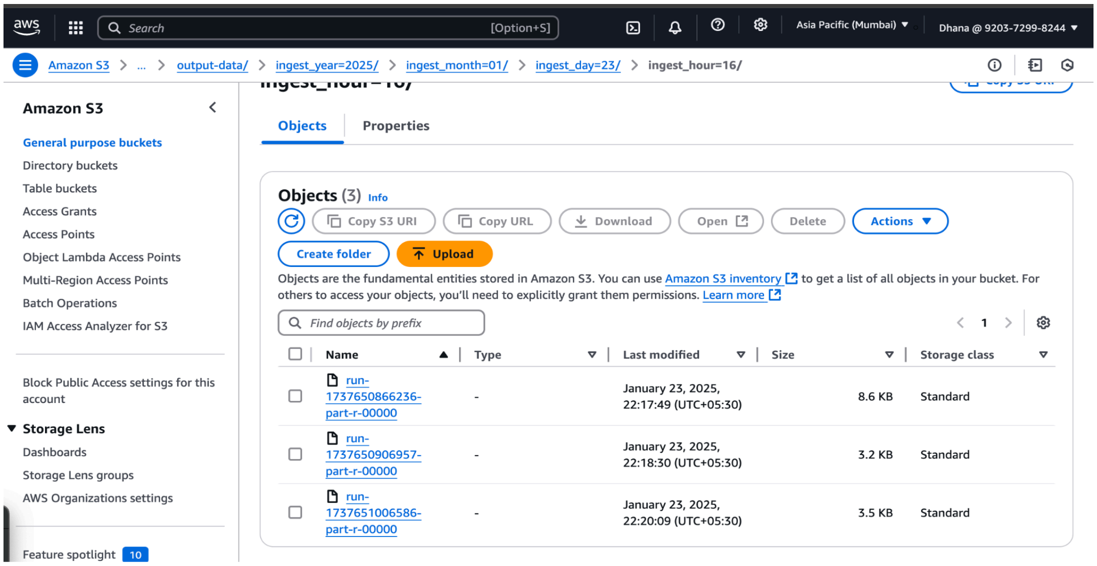
AWSGlueS3AthenaRole

Type

The type of ETL job. This is set automatically based on the types of data sources you have selected.

Spark Streaming

Once job is completed we should see the files placed in S3 from streaming data.

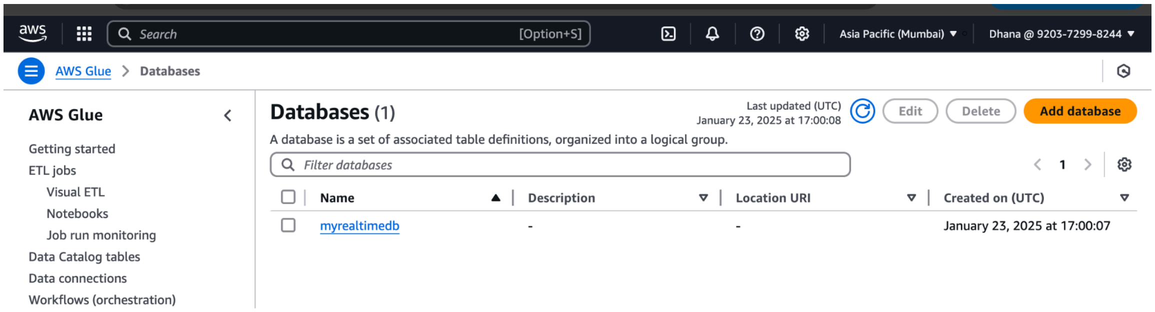


Objects (3)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

	Name	Type	Last modified	Size	Storage class
□	run-1737650866236-part-r-00000	-	January 23, 2025, 22:17:49 (UTC+05:30)	8.6 KB	Standard
□	run-1737650906957-part-r-00000	-	January 23, 2025, 22:18:30 (UTC+05:30)	3.2 KB	Standard
□	run-1737651006586-part-r-00000	-	January 23, 2025, 22:20:09 (UTC+05:30)	3.5 KB	Standard

Next, we will create a **database** in **Glue** and then **create Crawler** and **run the crawler for table creation.**

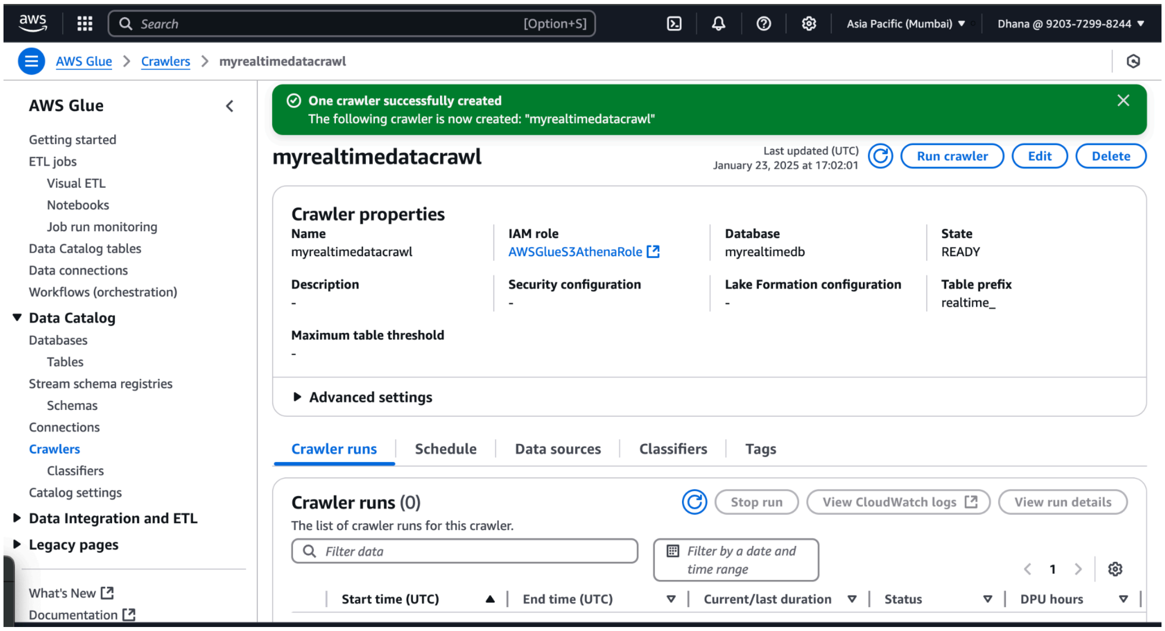


Databases (1)

Last updated (UTC)
January 23, 2025 at 17:00:08

A database is a set of associated table definitions, organized into a logical group.

	Name	Description	Location URI	Created on (UTC)
□	myrealtimedb	-	-	January 23, 2025 at 17:00:07



One crawler successfully created
The following crawler is now created: "myrealtimedatacrawl"

myrealtimedatacrawl

Last updated (UTC)
January 23, 2025 at 17:02:01

Crawler properties

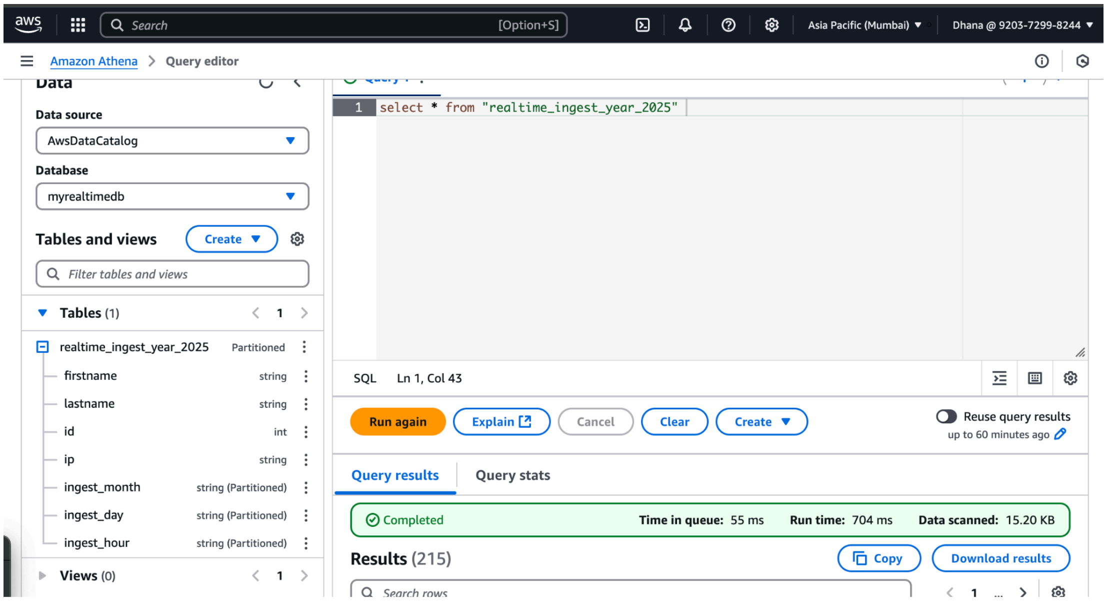
Name	IAM role	Database	State
myrealtimedatacrawl	AWSGlueS3AthenaRole	myrealtimedb	READY
Description	**Security configuration**	**Lake Formation configuration**	**Table prefix**
-	-	-	realtime_
Maximum table threshold			
-			

Crawler runs (0)

The list of crawler runs for this crawler.

Start time (UTC)	End time (UTC)	Current/last duration	Status	DPU hours

And then Use **Athena** to query the data

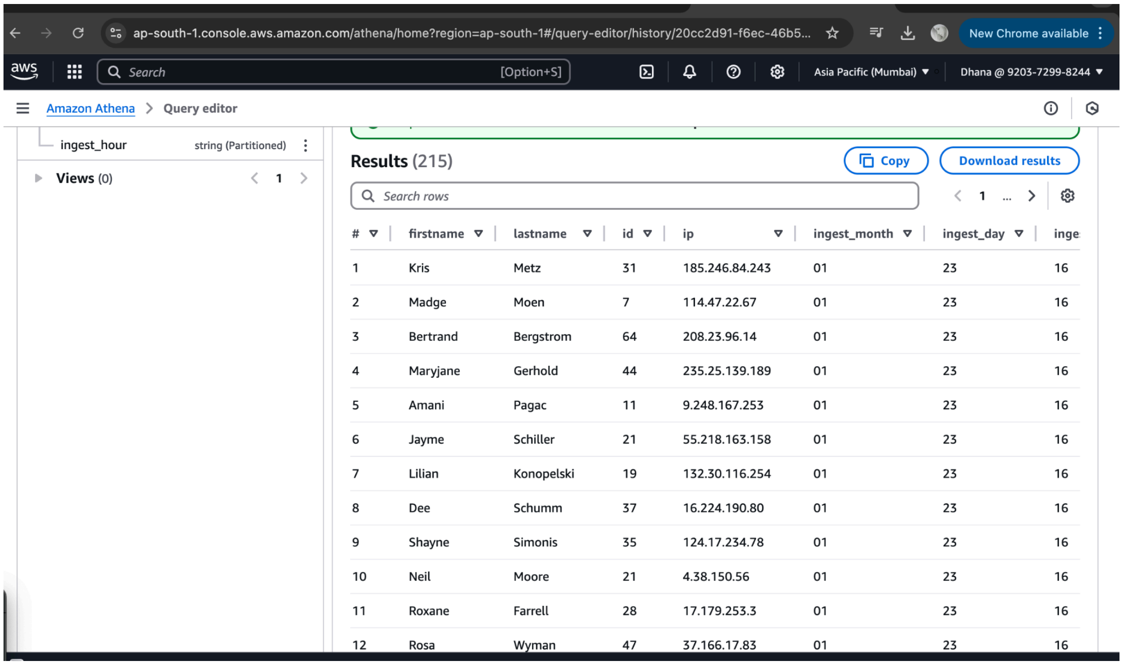


```
select * from "realtime_ingest_year_2025"
```

realtime_ingest_year_2025	Partitioned
firstname	string
lastname	string
id	int
ip	string
ingest_month	string (Partitioned)
ingest_day	string (Partitioned)
ingest_hour	string (Partitioned)

Completed **Time in queue:** 55 ms **Run time:** 704 ms **Data scanned:** 15.20 KB

Results (215)



Results (215)

#	firstname	lastname	id	ip	ingest_month	ingest_day	inge
1	Kris	Metz	31	185.246.84.243	01	23	16
2	Madge	Moen	7	114.47.22.67	01	23	16
3	Bertrand	Bergstrom	64	208.23.96.14	01	23	16
4	Maryjane	Gerhold	44	235.25.139.189	01	23	16
5	Amani	Pagac	11	9.248.167.253	01	23	16
6	Jayme	Schiller	21	55.218.163.158	01	23	16
7	Lilian	Konopelski	19	132.30.116.254	01	23	16
8	Dee	Schumm	37	16.224.190.80	01	23	16
9	Shayne	Simonis	35	124.17.234.78	01	23	16
10	Neil	Moore	21	4.38.150.56	01	23	16
11	Roxane	Farrell	28	17.179.253.3	01	23	16
12	Rosa	Wyman	47	37.166.17.83	01	23	16