accenture technology

Technology Services
Delivery Centers for
Technology in India

High performance. Delivered.

Learning and Knowledge Management

Collections : Lab Guide

Streaming Simulation using Amazon Kinesis Stream

&

Real Time Data Analytics using Amazon Kinesis Analytics

Developed & Tested

By

Karthigayen.Y

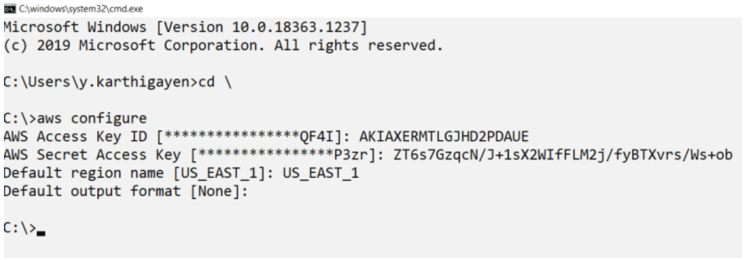
LKM, Accenture - ATCI

Context :

In this Lab we will be doing a Simualtion of Streaming Data from the Crime dataset using Amazon Kinesis Stream, And ingest that data into S3 and Amazon Kinesis Analytics , to perform basic Real Time Analysis of the Crime dataset.

Steps to simulate the streaming data from chicago_crime_dataset.csv file using Kinesis Stream

- Place the **chicago_crime_dataset.csv** file in a specific directory in C: drive on your windows machine.
    - Ex : C:\AWS_Dataset\chicago_crime_dataset.csv

- Download and install **AWS CLI V2** on windows and set the Environmental Variables as show below
    - https://docs.aws.amazon.com/cli/latest/userguide/install-cliv2-windows.html#cliv2-windows-install
    - 

```
Microsoft Windows [Version 10.0.18363.1237]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\Users\y.karthigayen>cd \

C:\>aws configure
AWS Access Key ID [****************QF4I]: AKIAXERMTLGJHD2PDAUE
AWS Secret Access Key [****************P3zr]: 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
Default region name [US_EAST_1]: US_EAST_1
Default output format [None]:

C:\>
```

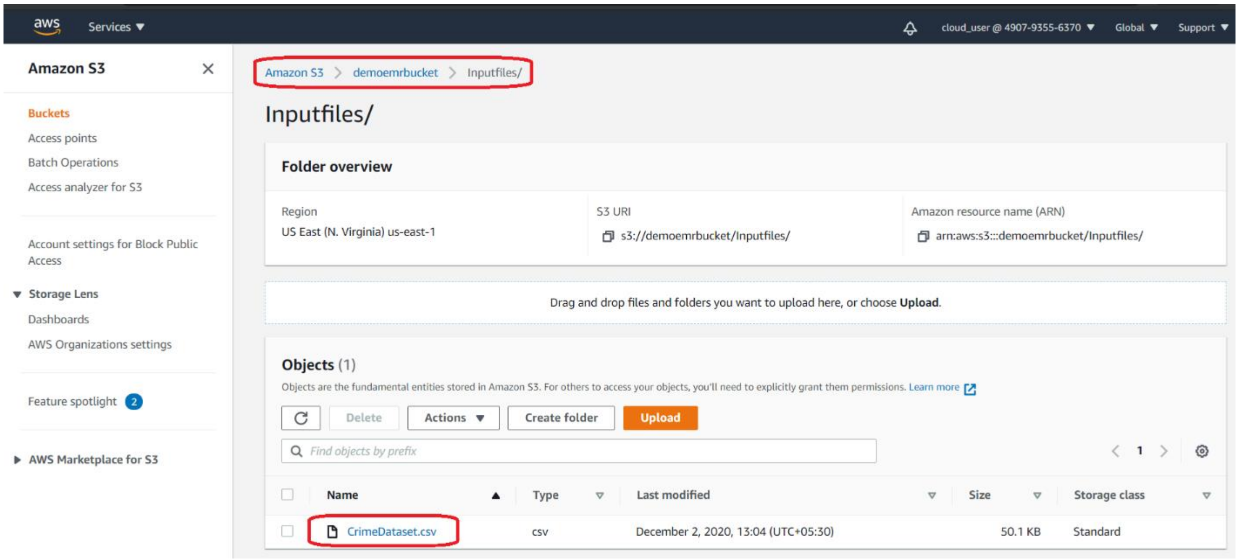
- Using the AWS console , create a Kinesis Data Stream named **“chicagoCrimeStream”.**

- Use the Maven Project shared with you, which has the below classes to read the Data from the specified location and provide us the simulation of streaming data and write the data into the stream which we had created.
    - **AwsKinesisClient.java**
    - **CrimeDataWriter.java**

Steps to read the Data from the Kinesis Stream and write it into a S3 bucket

- Using the AWS console, create a bucket named **“demoemrbucket”** on **S3 .**

- The below code will help us read the Data writen on to the **chicagoCrimeStream** and ingest the same into the S3 bucket which we had created
    - **CrimeDataS3Put.java**



Inputfiles/

Folder overview

Region	S3 URI	Amazon resource name (ARN)
US East (N. Virginia) us-east-1	s3://demoemrbucket/Inputfiles/	arn:aws:s3:::demoemrbucket/Inputfiles/

Drag and drop files and folders you want to upload here, or choose **Upload**.

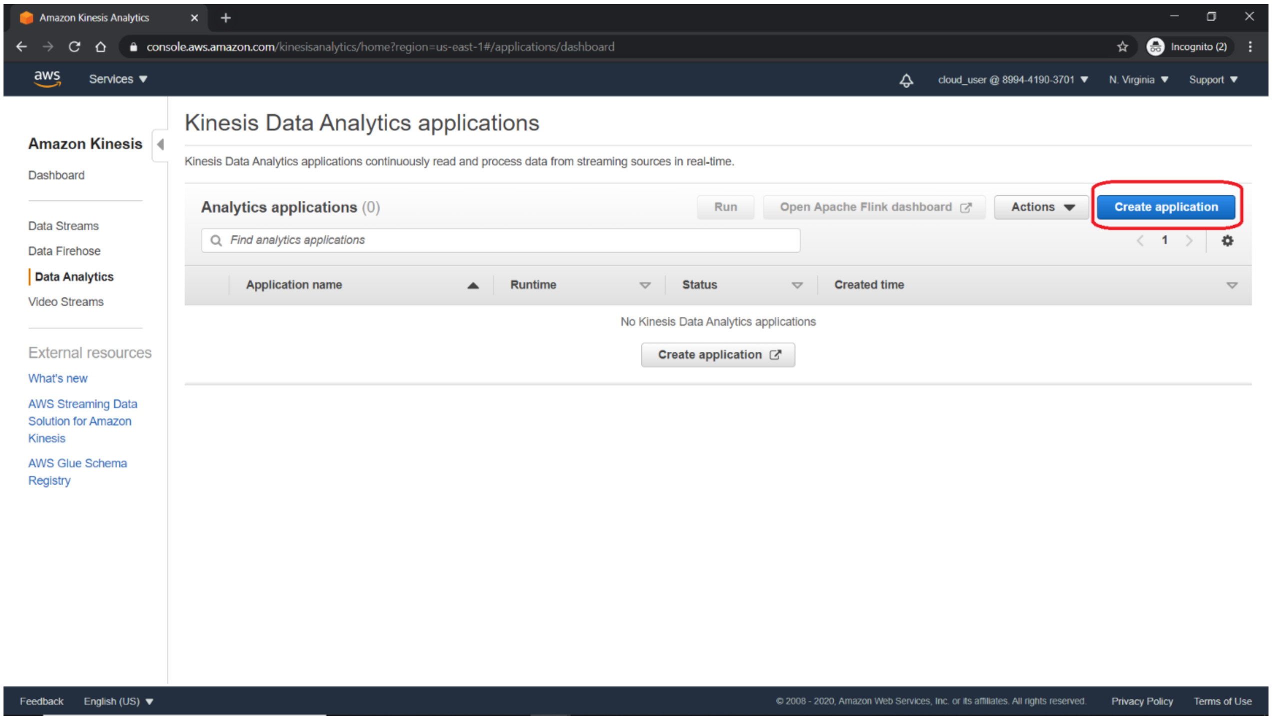
Objects (1)

Objects are the fundamental entities stored in Amazon S3. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

	Name	Type	Last modified	Size	Storage class
□	CrimeDataset.csv	csv	December 2, 2020, 13:04 (UTC+05:30)	50.1 KB	Standard

Steps to create a Kinesis Data Analytics Application and fetch the data from Kinesis Stream “chicagoCrimeStream” and do the Analysis.

Step 1 : Click on Create Application button


Amazon Kinesis Analytics
console.aws.amazon.com/kinesisanalytics/home?region=us-east-1#/applications/dashboard
Incognito (2)
aws
Services ▼
cloud_user @ 8994-4190-3701 ▼
N. Virginia ▼
Support ▼
Amazon Kinesis
Dashboard
Data Streams
Data Firehose
Data Analytics
Video Streams
External resources
What's new
AWS Streaming Data Solution for Amazon Kinesis
AWS Glue Schema Registry
Run
Open Apache Flink dashboard
Actions ▼
Create application
Find analytics applications
1
Feedback
English (US) ▼
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Kinesis Data Analytics applications

Kinesis Data Analytics applications continuously read and process data from streaming sources in real-time.

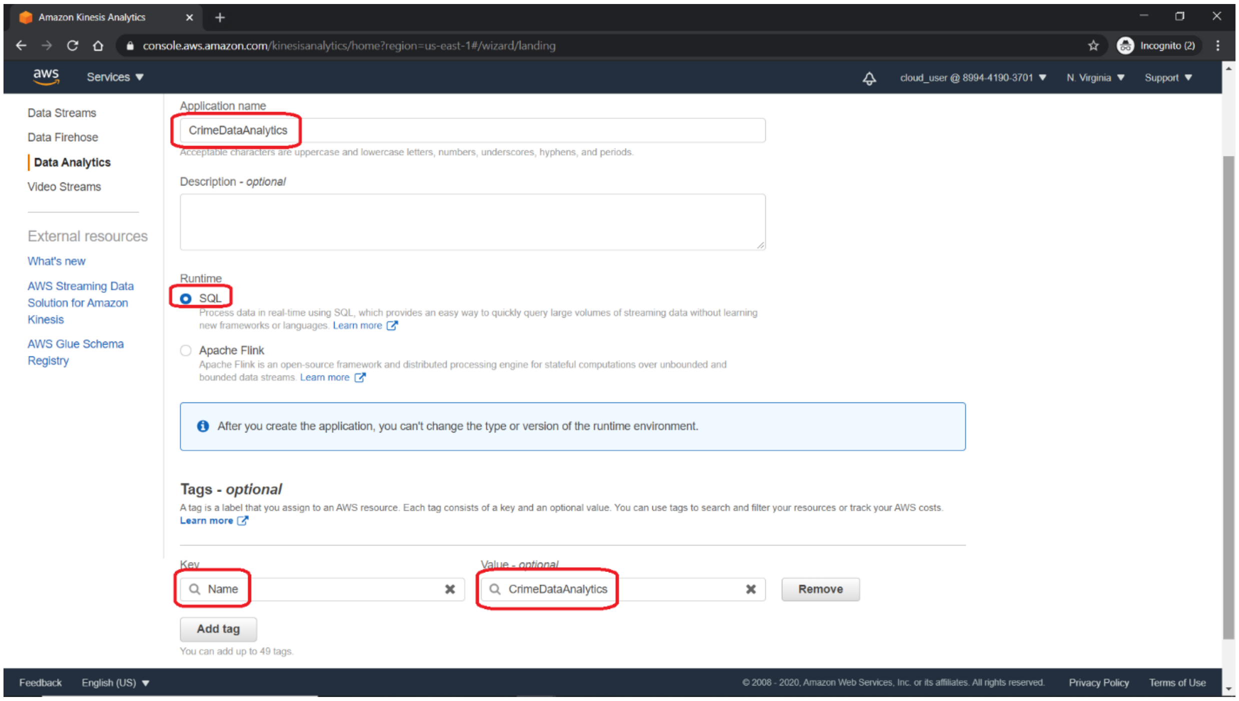
Analytics applications (0)

Application name	Runtime	Status	Created time

No Kinesis Data Analytics applications

Create application

Step 2 : Specify the name of the application as **CrimeDataAnalytics,** and the **SQL** as the environment to do the Analysis.



Application name

CrimeDataAnalytics

Acceptable characters are uppercase and lowercase letters, numbers, underscores, hyphens, and periods.

Description - *optional*

Runtime

● SQL

Process data in real-time using SQL, which provides an easy way to quickly query large volumes of streaming data without learning new frameworks or languages. Learn more

○ Apache Flink

Apache Flink is an open-source framework and distributed processing engine for stateful computations over unbounded and bounded data streams. Learn more

After you create the application, you can't change the type or version of the runtime environment.

Tags - *optional*

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs. Learn more

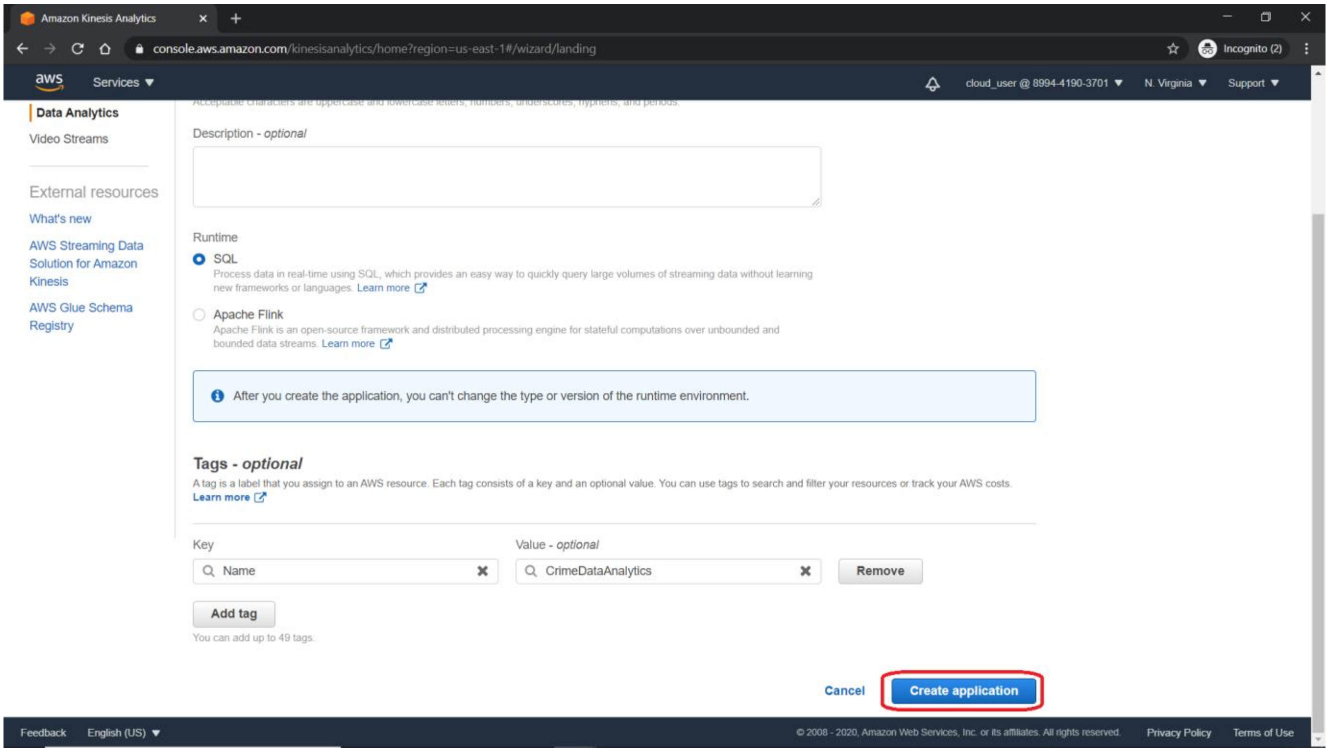
Key	Value - *optional*
Name	CrimeDataAnalytics

Add tag

You can add up to 49 tags.

Amazon Kinesis Analytics
console.aws.amazon.com/kinesisanalytics/home?region=us-east-1#/wizard/landing
Incognito (2)
aws
Services
cloud_user @ 8994-4190-3701
N. Virginia
Support
Data Streams
Data Firehose
Data Analytics
Video Streams
External resources
What's new
AWS Streaming Data Solution for Amazon Kinesis
AWS Glue Schema Registry
Remove
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Step 3 : Click on Create Application button



Description - *optional*

Runtime

● SQL

Process data in real-time using SQL, which provides an easy way to quickly query large volumes of streaming data without learning new frameworks or languages. Learn more

○ Apache Flink

Apache Flink is an open-source framework and distributed processing engine for stateful computations over unbounded and bounded data streams. Learn more

After you create the application, you can't change the type or version of the runtime environment.

Tags - *optional*

A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs. Learn more

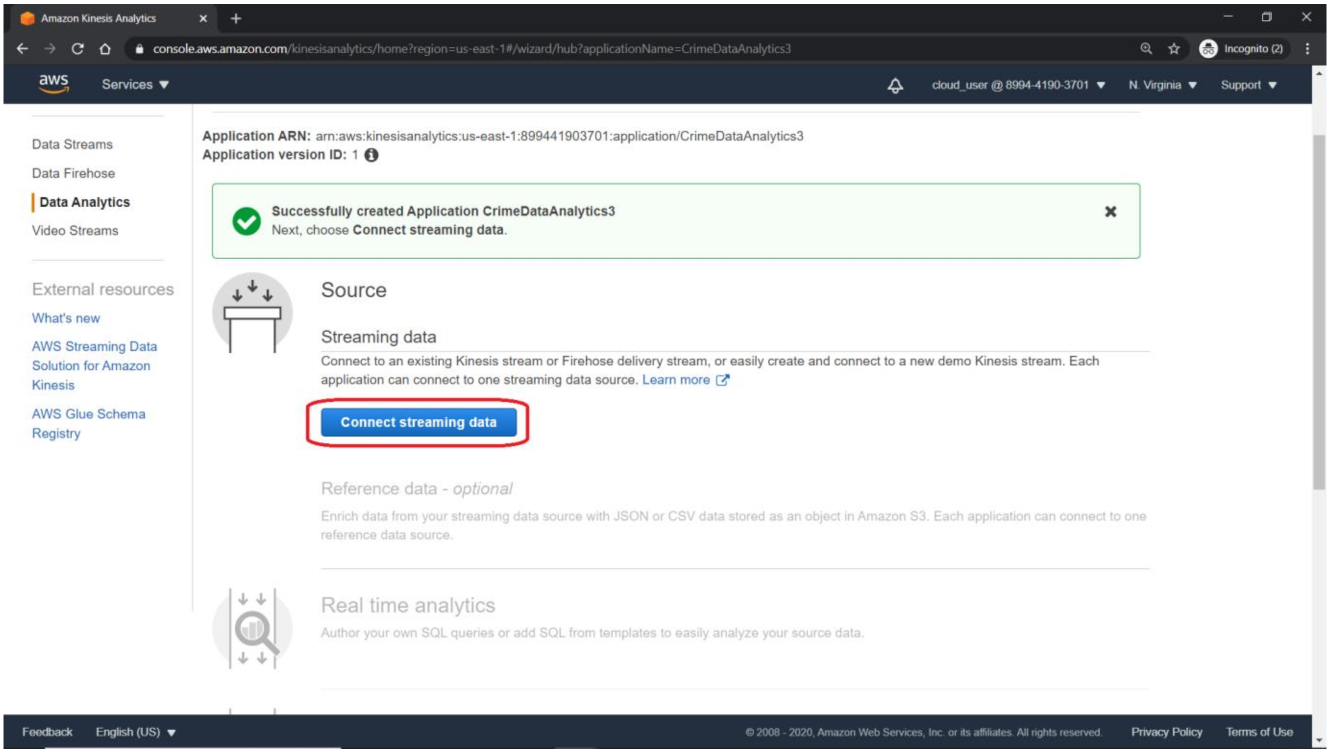
Key	Value - *optional*
Name	CrimeDataAnalytics

Add tag

You can add up to 49 tags.

Amazon Kinesis Analytics
console.aws.amazon.com/kinesisanalytics/home?region=us-east-1#/wizard/landing
Incognito (2)
aws
Services
cloud_user @ 8994-4190-3701
N. Virginia
Support
Data Analytics
Video Streams
External resources
What's new
AWS Streaming Data Solution for Amazon Kinesis
AWS Glue Schema Registry
Remove
Cancel
Create application
Feedback
English (US)
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Step 4 : Click on the Connect Streaming Data button and choose Kinesis Data Stream as the Source.


Amazon Kinesis Analytics
console.aws.amazon.com/kinesisanalytics/home?region=us-east-1#/wizard/hub?applicationName=CrimeDataAnalytics3
Incognito (2)
aws
Services ▼
cloud_user @ 8994-4190-3701 ▼
N. Virginia ▼
Support ▼
Data Streams
Data Firehose
Data Analytics
Video Streams
External resources
What's new
AWS Streaming Data Solution for Amazon Kinesis
AWS Glue Schema Registry
Connect streaming data
Feedback
English (US) ▼
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Application ARN: arn:aws:kinesisanalytics:us-east-1:899441903701:application/CrimeDataAnalytics3
Application version ID: 1

Successfully created Application CrimeDataAnalytics3
Next, choose **Connect streaming data**.

Source

Streaming data

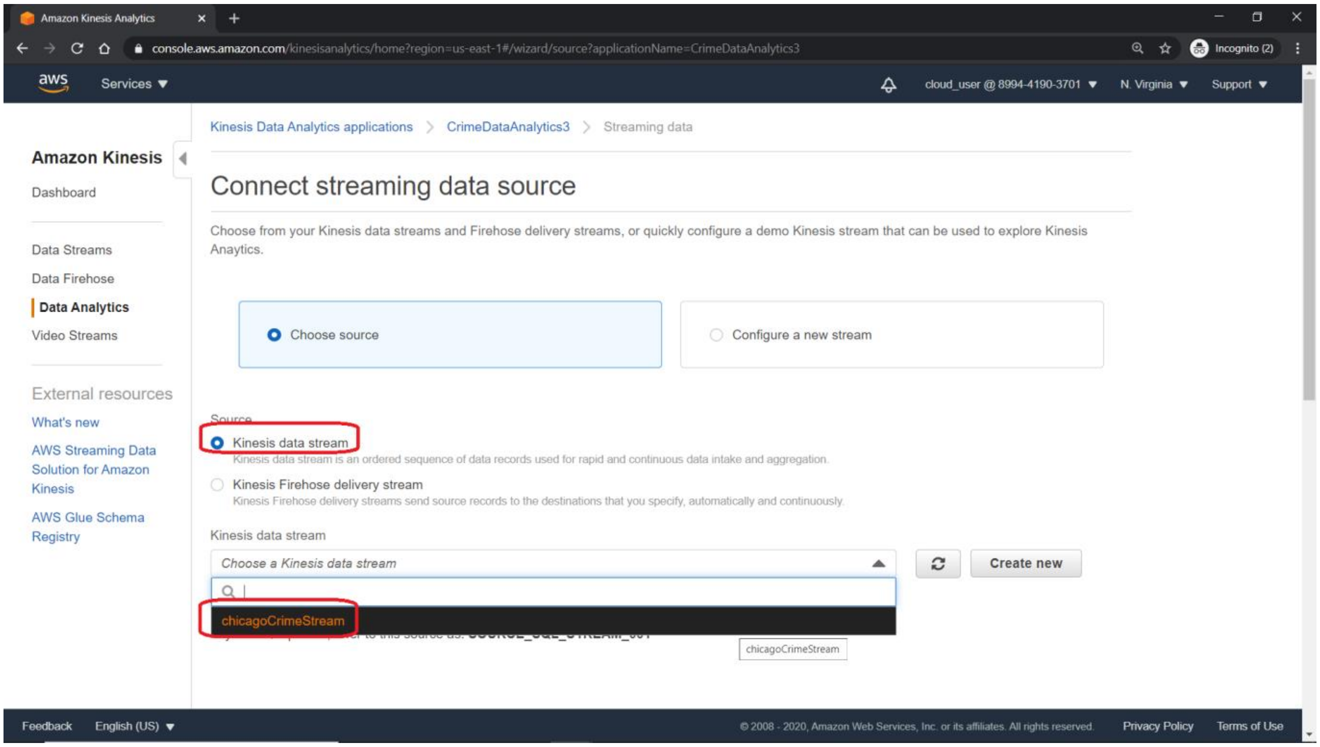
Connect to an existing Kinesis stream or Firehose delivery stream, or easily create and connect to a new demo Kinesis stream. Each application can connect to one streaming data source. Learn more

Reference data - *optional*

Enrich data from your streaming data source with JSON or CSV data stored as an object in Amazon S3. Each application can connect to one reference data source.

Real time analytics

Author your own SQL queries or add SQL from templates to easily analyze your source data.



Kinesis Data Analytics applications > CrimeDataAnalytics3 > Streaming data

Connect streaming data source

Choose from your Kinesis data streams and Firehose delivery streams, or quickly configure a demo Kinesis stream that can be used to explore Kinesis Anaytics.

● Choose source ○ Configure a new stream

Source

● Kinesis data stream
Kinesis data stream is an ordered sequence of data records used for rapid and continuous data intake and aggregation.

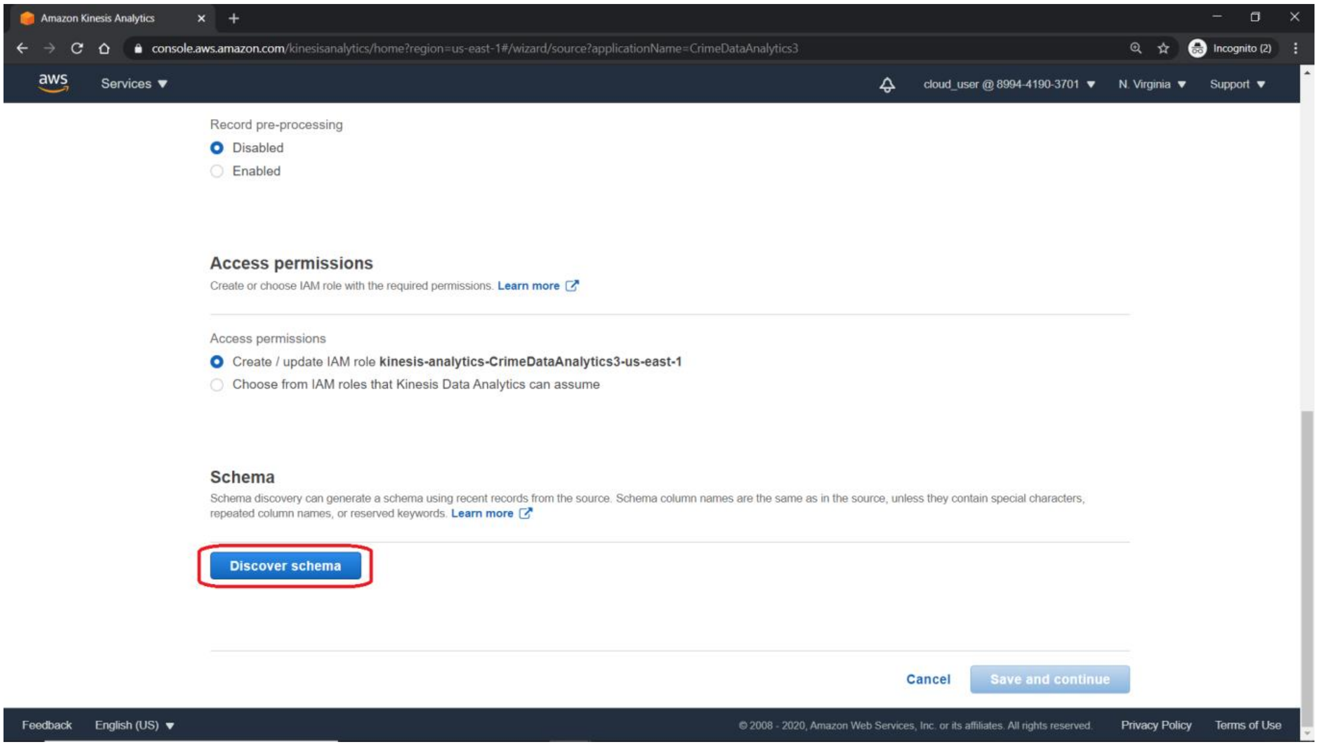
○ Kinesis Firehose delivery stream
Kinesis Firehose delivery streams send source records to the destinations that you specify, automatically and continuously.

Kinesis data stream

Choose a Kinesis data stream

chicagoCrimeStream

Step 5 : Click on the Discover Schema button to Automatically detect the schema for Streaming Data


Amazon Kinesis Analytics
console.aws.amazon.com/kinesisanalytics/home?region=us-east-1#/wizard/source?applicationName=CrimeDataAnalytics3
Incognito (2)
aws
Services ▼
cloud_user @ 8994-4190-3701 ▼
N. Virginia ▼
Support ▼
Feedback
English (US) ▼
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

Record pre-processing

● Disabled

○ Enabled

Access permissions

Create or choose IAM role with the required permissions. Learn more

Access permissions

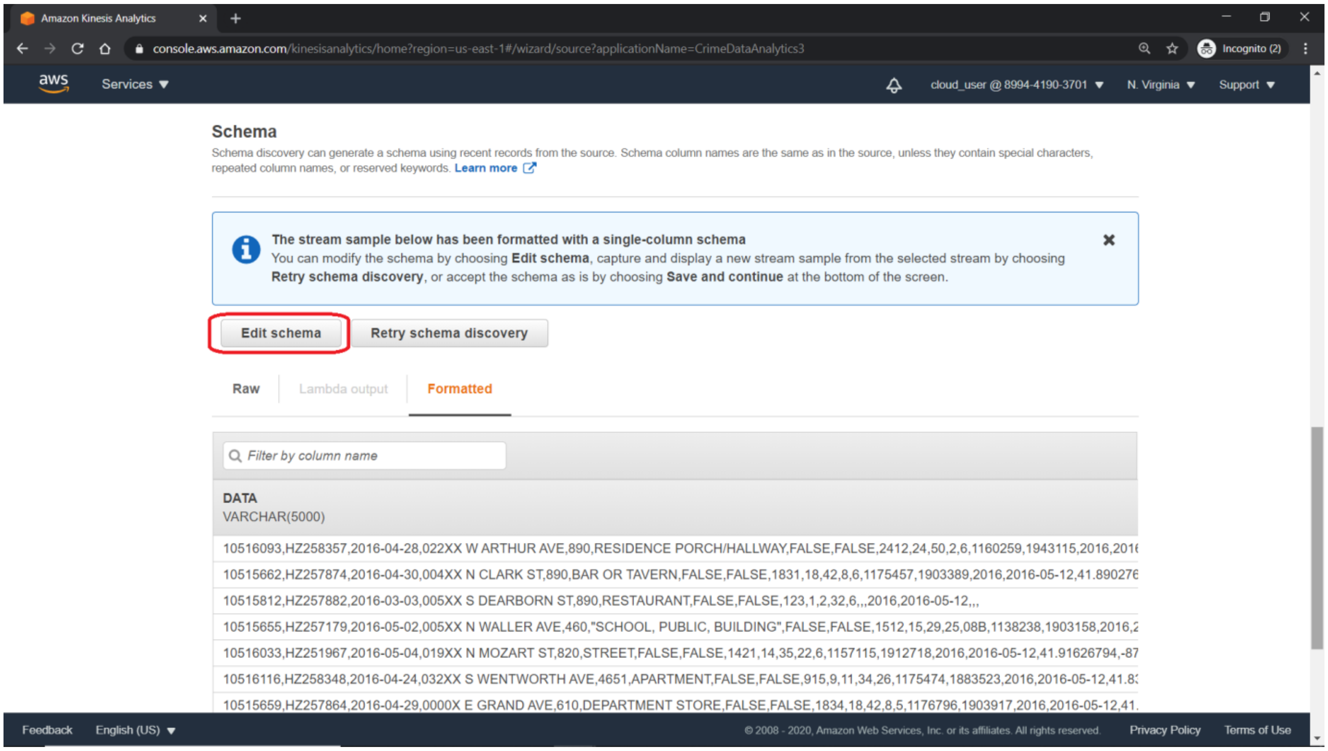
● Create / update IAM role **kinesis-analytics-CrimeDataAnalytics3-us-east-1**

○ Choose from IAM roles that Kinesis Data Analytics can assume

Schema

Schema discovery can generate a schema using recent records from the source. Schema column names are the same as in the source, unless they contain special characters, repeated column names, or reserved keywords. Learn more

Step 6 : If the schema is not as expected , then click on Edit Schema button and edit the schema as required.


Amazon Kinesis Analytics
console.aws.amazon.com/kinesisanalytics/home?region=us-east-1#/wizard/source?applicationName=CrimeDataAnalytics3
Incognito (2)
aws
Services ▼
cloud_user @ 8994-4190-3701 ▼
N. Virginia ▼
Support ▼
Feedback
English (US) ▼
© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use

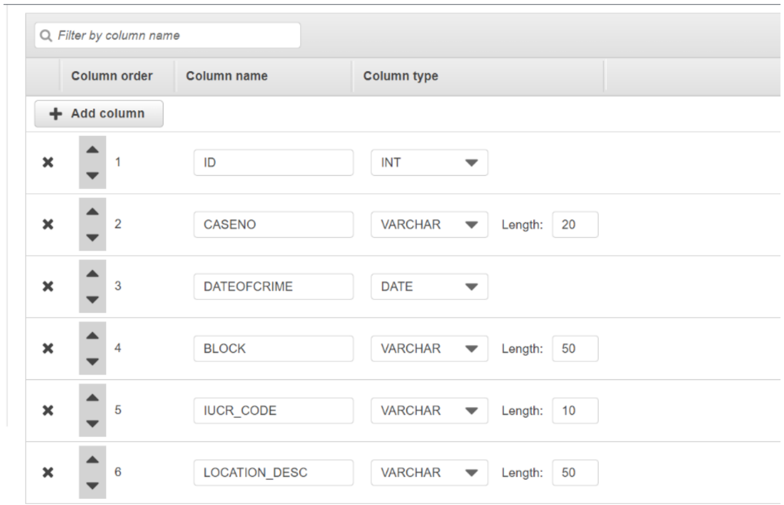
Schema

Schema discovery can generate a schema using recent records from the source. Schema column names are the same as in the source, unless they contain special characters, repeated column names, or reserved keywords. Learn more

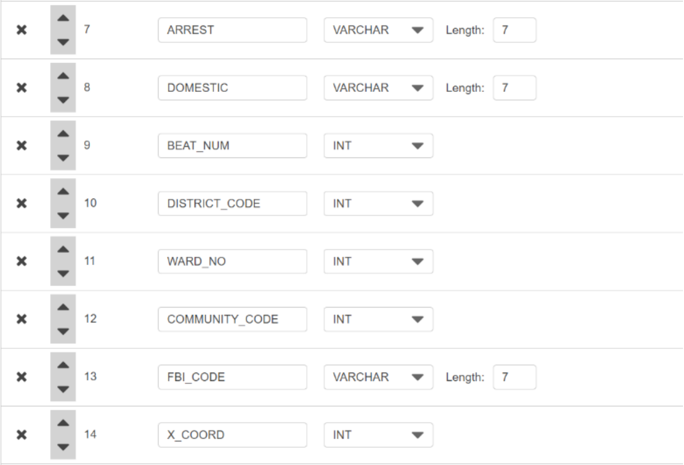
The stream sample below has been formatted with a single-column schema
You can modify the schema by choosing **Edit schema**, capture and display a new stream sample from the selected stream by choosing **Retry schema discovery**, or accept the schema as is by choosing **Save and continue** at the bottom of the screen.

DATA VARCHAR(5000)
10516093,HZ258357,2016-04-28,022XX W ARTHUR AVE,890,RESIDENCE PORCH/HALLWAY,FALSE,FALSE,2412,24,50,2,6,1160259,1943115,2016,2016
10515662,HZ257874,2016-04-30,004XX N CLARK ST,890,BAR OR TAVERN,FALSE,FALSE,1831,18,42,8,6,1175457,1903389,2016,2016-05-12,41.890276
10515812,HZ257882,2016-03-03,005XX S DEARBORN ST,890,RESTAURANT,FALSE,FALSE,123,1,2,32,6,,,2016,2016-05-12,,,
10515655,HZ257179,2016-05-02,005XX N WALLER AVE,460,"SCHOOL, PUBLIC, BUILDING",FALSE,FALSE,1512,15,29,25,08B,1138238,1903158,2016,2
10516033,HZ251967,2016-05-04,019XX N MOZART ST,820,STREET,FALSE,FALSE,1421,14,35,22,6,1157115,1912718,2016,2016-05-12,41.91626794,-87
10516116,HZ258348,2016-04-24,032XX S WENTWORTH AVE,4651,APARTMENT,FALSE,FALSE,915,9,11,34,26,1175474,1883523,2016,2016-05-12,41.83
10515659,HZ257864,2016-04-29,0000X E GRAND AVE,610,DEPARTMENT STORE,FALSE,FALSE,1834,18,42,8,5,1176796,1903917,2016,2016-05-12,41.

Step 7 : Add the below mentioned Columns for Crime Dataset.

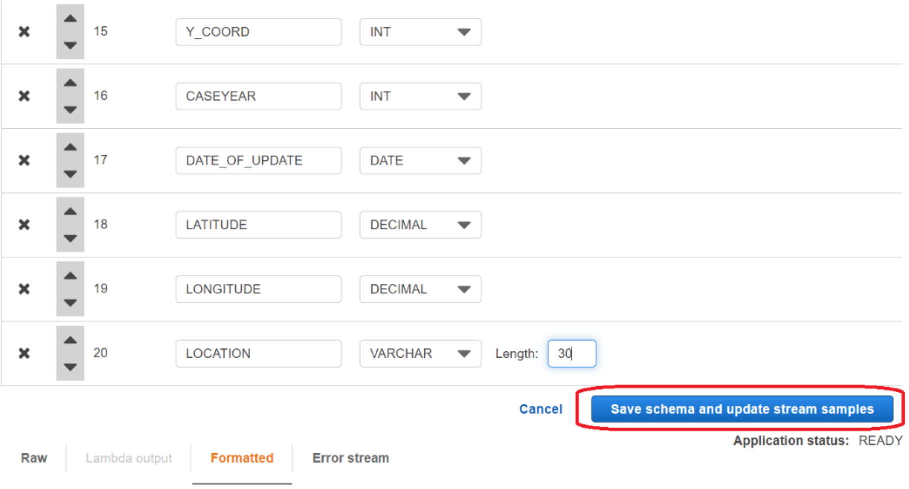


Column order	Column name	Column type	Length
1	ID	INT	
2	CASENO	VARCHAR	20
3	DATEOFCRIME	DATE	
4	BLOCK	VARCHAR	50
5	IUCR_CODE	VARCHAR	10
6	LOCATION_DESC	VARCHAR	50



Column order	Column name	Column type	Length
7	ARREST	VARCHAR	7
8	DOMESTIC	VARCHAR	7
9	BEAT_NUM	INT	
10	DISTRICT_CODE	INT	
11	WARD_NO	INT	
12	COMMUNITY_CODE	INT	
13	FBI_CODE	VARCHAR	7
14	X_COORD	INT	

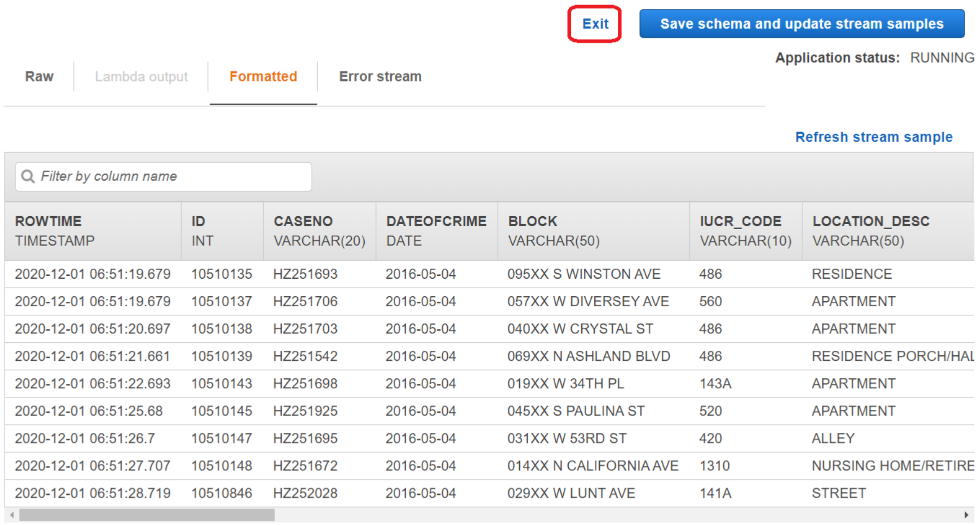
Step 8 : Once all the columns are add , click on Save Schema button



Column order	Column name	Column type	Length
15	Y_COORD	INT	
16	CASEYEAR	INT	
17	DATE_OF_UPDATE	DATE	
18	LATITUDE	DECIMAL	
19	LONGITUDE	DECIMAL	
20	LOCATION	VARCHAR	30

Application status: READY

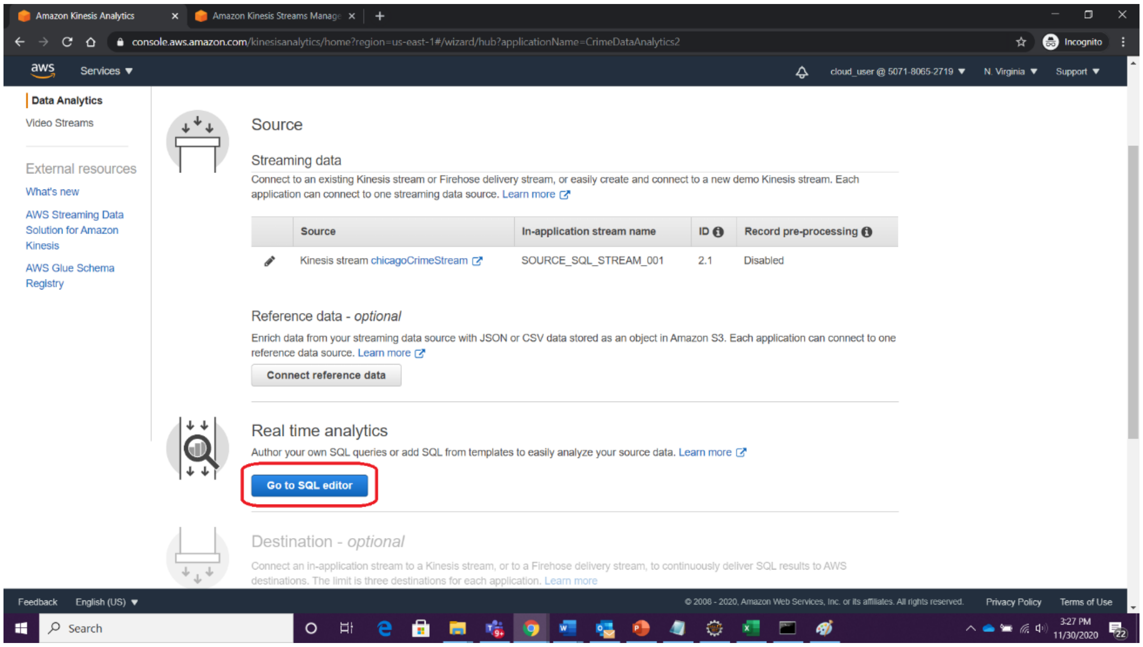
Step 9 : Now click on Exit(Done) link to go to Analytics page.



Application status: RUNNING

ROWTIME TIMESTAMP	ID INT	CASENO VARCHAR(20)	DATEOFCRIME DATE	BLOCK VARCHAR(50)	IUCR_CODE VARCHAR(10)	LOCATION_DESC VARCHAR(50)
2020-12-01 06:51:19.679	10510135	HZ251693	2016-05-04	095XX S WINSTON AVE	486	RESIDENCE
2020-12-01 06:51:19.679	10510137	HZ251706	2016-05-04	057XX W DIVERSEY AVE	560	APARTMENT
2020-12-01 06:51:20.697	10510138	HZ251703	2016-05-04	040XX W CRYSTAL ST	486	APARTMENT
2020-12-01 06:51:21.661	10510139	HZ251542	2016-05-04	069XX N ASHLAND BLVD	486	RESIDENCE PORCH/HAL
2020-12-01 06:51:22.693	10510143	HZ251698	2016-05-04	019XX W 34TH PL	143A	APARTMENT
2020-12-01 06:51:25.68	10510145	HZ251925	2016-05-04	045XX S PAULINA ST	520	APARTMENT
2020-12-01 06:51:26.7	10510147	HZ251695	2016-05-04	031XX W 53RD ST	420	ALLEY
2020-12-01 06:51:27.707	10510148	HZ251672	2016-05-04	014XX N CALIFORNIA AVE	1310	NURSING HOME/RETIRE
2020-12-01 06:51:28.719	10510846	HZ252028	2016-05-04	029XX W LUNT AVE	141A	STREET

Step 10 : Click on Go to SQL editor , to create the Analytical Application code



console.aws.amazon.com/kinesisanalytics/home?region=us-east-1#/wizard/hub?applicationName=CrimeDataAnalytics2

Source

Streaming data

Connect to an existing Kinesis stream or Firehose delivery stream, or easily create and connect to a new demo Kinesis stream. Each application can connect to one streaming data source. Learn more

Source	In-application stream name	ID	Record pre-processing
Kinesis stream chicagoCrimeStream	SOURCE_SQL_STREAM_001	2.1	Disabled

Reference data - *optional*

Enrich data from your streaming data source with JSON or CSV data stored as an object in Amazon S3. Each application can connect to one reference data source. Learn more

Real time analytics

Author your own SQL queries or add SQL from templates to easily analyze your source data. Learn more

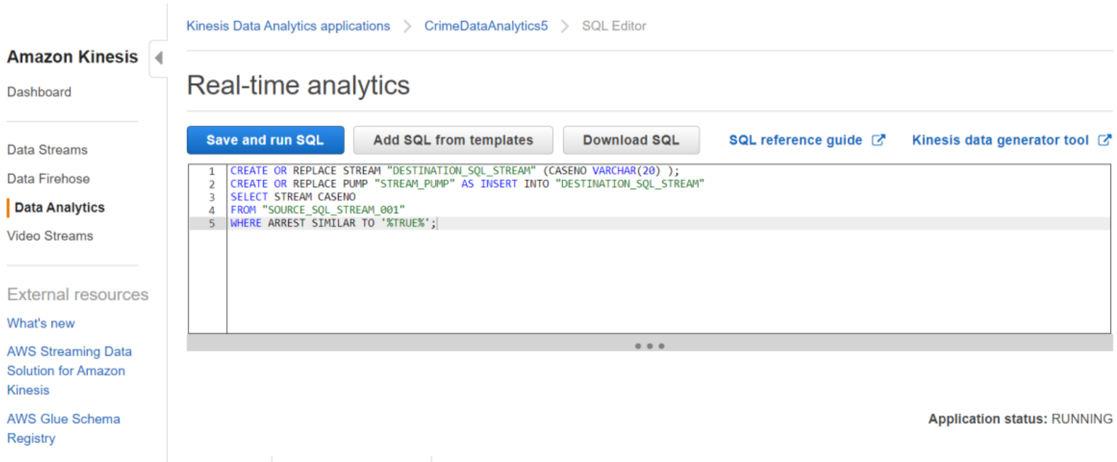
Destination - *optional*

Connect an in-application stream to a Kinesis stream, or to a Firehose delivery stream, to continuously deliver SQL results to AWS destinations. The limit is three destinations for each application. Learn more

© 2008 - 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Steps to Write Analytical Queries on Kinesis- Analytics:

- Create a Destination Stream
- Create a Stream Pump to push the resultant data into the Destination stream
- With the Analytical query triggered on the Source Data Stream



```
CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (CASENO VARCHAR(20) );
CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"
SELECT STREAM CASENO
FROM "SOURCE_SQL_STREAM_001"
WHERE ARREST SIMILAR TO '%TRUE%';
```

Application status: RUNNING

Use Case 1 : To list all the Case numbers where Arrest has happened

CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (CASENO VARCHAR(20));

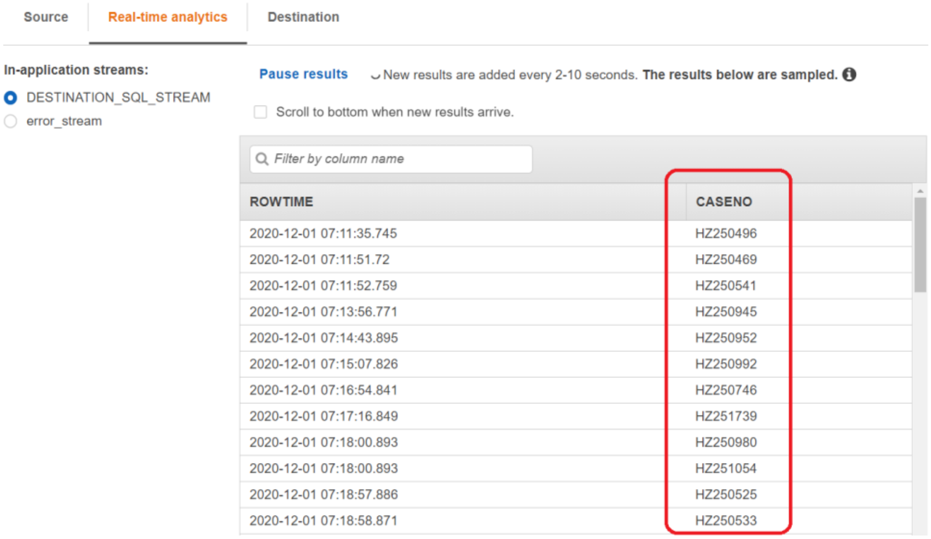
CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM CASENO

FROM "SOURCE_SQL_STREAM_001"

WHERE ARREST SIMILAR TO '%TRUE%';

Destination Stream Output :



ROWTIME	CASENO
2020-12-01 07:11:35.745	HZ250496
2020-12-01 07:11:51.72	HZ250469
2020-12-01 07:11:52.759	HZ250541
2020-12-01 07:13:56.771	HZ250945
2020-12-01 07:14:43.895	HZ250952
2020-12-01 07:15:07.826	HZ250992
2020-12-01 07:16:54.841	HZ250746
2020-12-01 07:17:16.849	HZ251739
2020-12-01 07:18:00.893	HZ250980
2020-12-01 07:18:00.893	HZ251054
2020-12-01 07:18:57.886	HZ250525
2020-12-01 07:18:58.871	HZ250533

Use Case 2 : To check if the Case-ID is unique :

CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (ID INT, ID_count INTEGER);

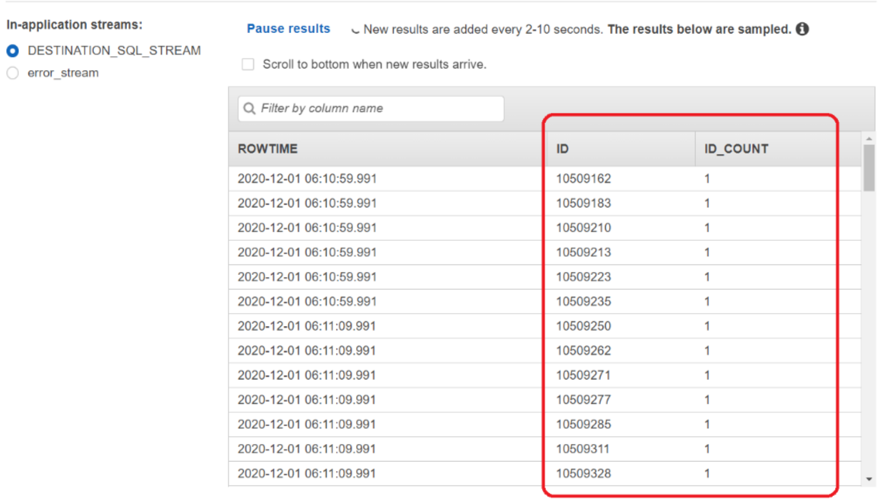
CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM ID, COUNT(*) AS ID_count

FROM "SOURCE_SQL_STREAM_001"

GROUP BY ID, FLOOR(("SOURCE_SQL_STREAM_001".ROWTIME - TIMESTAMP '2020-12-01 11:30:00') SECOND / 10 TO SECOND);

Output :


In-application streams:
● DESTINATION_SQL_STREAM
○ error_stream
Pause results
New results are added every 2-10 seconds. The results below are sampled.
□ Scroll to bottom when new results arrive.
Filter by column name

ROWTIME	ID	ID_COUNT
2020-12-01 06:10:59.991	10509162	1
2020-12-01 06:10:59.991	10509183	1
2020-12-01 06:10:59.991	10509210	1
2020-12-01 06:10:59.991	10509213	1
2020-12-01 06:10:59.991	10509223	1
2020-12-01 06:10:59.991	10509235	1
2020-12-01 06:11:09.991	10509250	1
2020-12-01 06:11:09.991	10509262	1
2020-12-01 06:11:09.991	10509271	1
2020-12-01 06:11:09.991	10509277	1
2020-12-01 06:11:09.991	10509285	1
2020-12-01 06:11:09.991	10509311	1
2020-12-01 06:11:09.991	10509328	1

Use Case 3 : Create report on total number of crime cases on each day from crimes dataset

CREATE OR REPLACE STREAM "DESTINATION_SQL_STREAM" (DATEOFCRIME DATE, COUNT_OF_CASES INT);

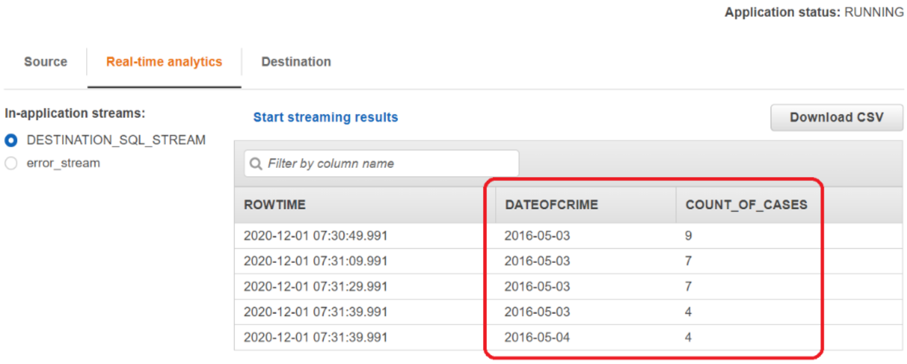
CREATE OR REPLACE PUMP "STREAM_PUMP" AS INSERT INTO "DESTINATION_SQL_STREAM"

SELECT STREAM DATEOFCRIME, COUNT(CASENO) AS COUNT_OF_CASES

FROM "SOURCE_SQL_STREAM_001"

GROUP BY DATEOFCRIME, FLOOR(("SOURCE_SQL_STREAM_001".ROWTIME - TIMESTAMP '2020-12-01 11:30:00') SECOND / 10 TO SECOND);

Output : Note - the aggregation is for specific duration (2 to 10 Sec), hence the groups may repeat.



ROWTIME	DATEOFCRIME	COUNT_OF_CASES
2020-12-01 07:30:49.991	2016-05-03	9
2020-12-01 07:31:09.991	2016-05-03	7
2020-12-01 07:31:29.991	2016-05-03	7
2020-12-01 07:31:39.991	2016-05-03	4
2020-12-01 07:31:39.991	2016-05-04	4