Streaming Data Analytics with Amazon Kinesis Firehose and Redshift

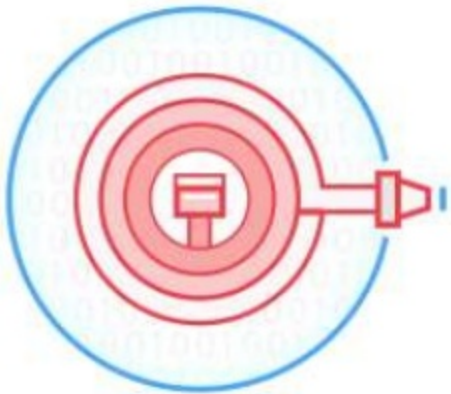
Ray Zhu, Sr. Product Manager, Amazon Kinesis

9/1/2016

Agenda

- Kinesis Firehose and Redshift
- Stream Data to Redshift
  - Step 1 Set Up Redshift DB and Table
  - Step 2 Create Firehose Delivery Stream
  - Step 3 Send Data to Firehose Delivery Stream
  - Step 4 Query and Analyze the Data from Redshift
  - Step 5 Monitor Streaming Data Pipeline

Kinesis Firehose



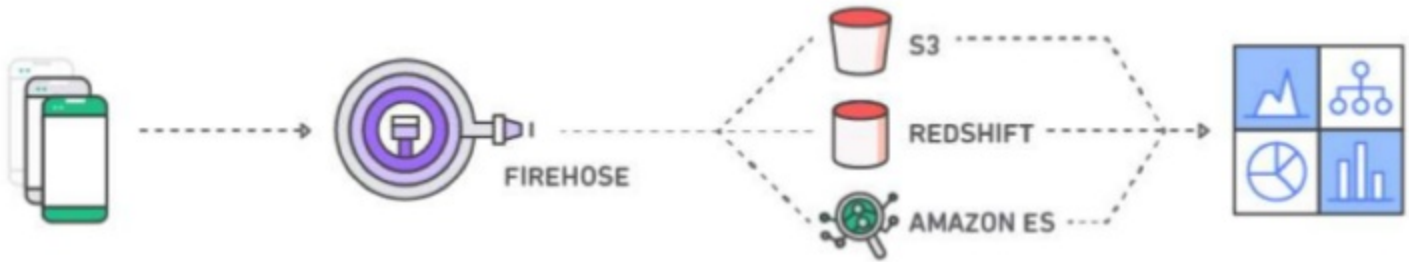
Load streaming data into Amazon S3, Amazon Redshift, and Amazon Elasticsearch Service

Amazon Redshift

Petabyte-scale data warehouse

Stream Data to Redshift

Data Flow Overview



Capture and submit streaming data to Firehose

Firehose loads streaming data continuously into Amazon S3, Redshift, or Elasticsearch Service

Analyze streaming data using your favorite analytical tools

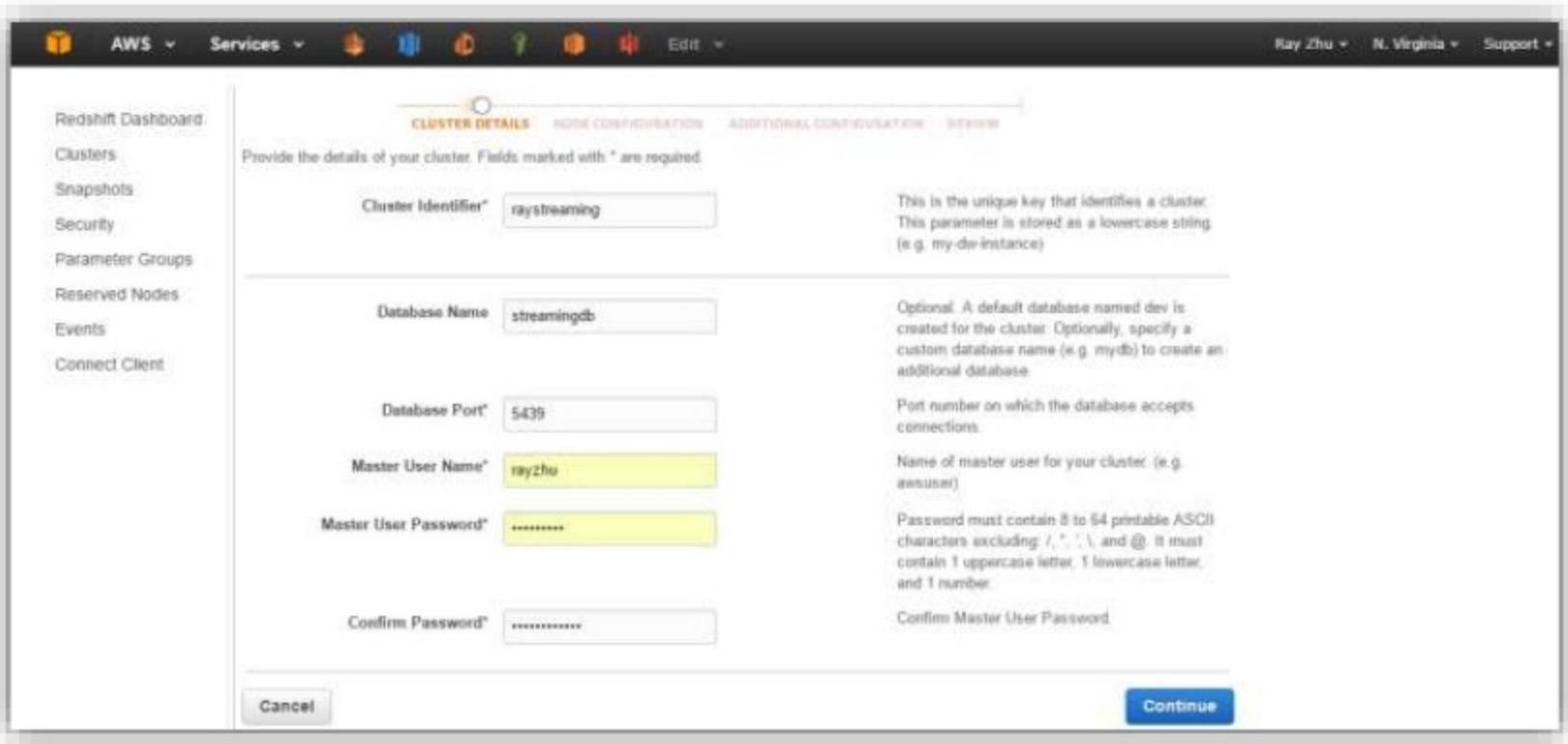
Zero administration: Capture and deliver streaming data into Amazon S3, Redshift, and Elasticsearch Service without writing an application or managing infrastructure.

Direct-to-data store integration: Batch, compress, and encrypt streaming data for delivery into data destinations in as little as 60 secs using simple configurations.

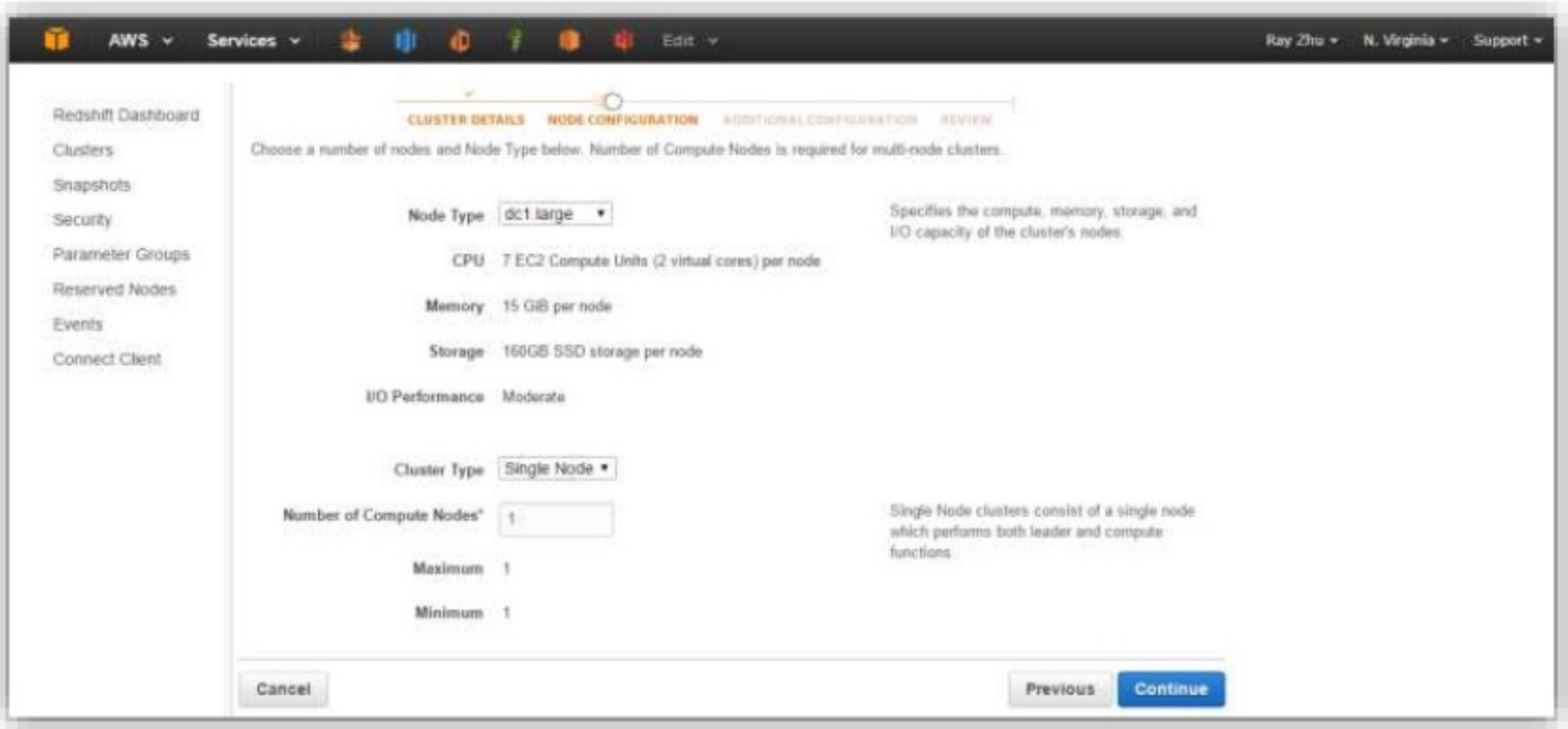
Seamless elasticity: Seamlessly scale to match data throughput without intervention.

Step 1 Set Up Redshift DB and Table

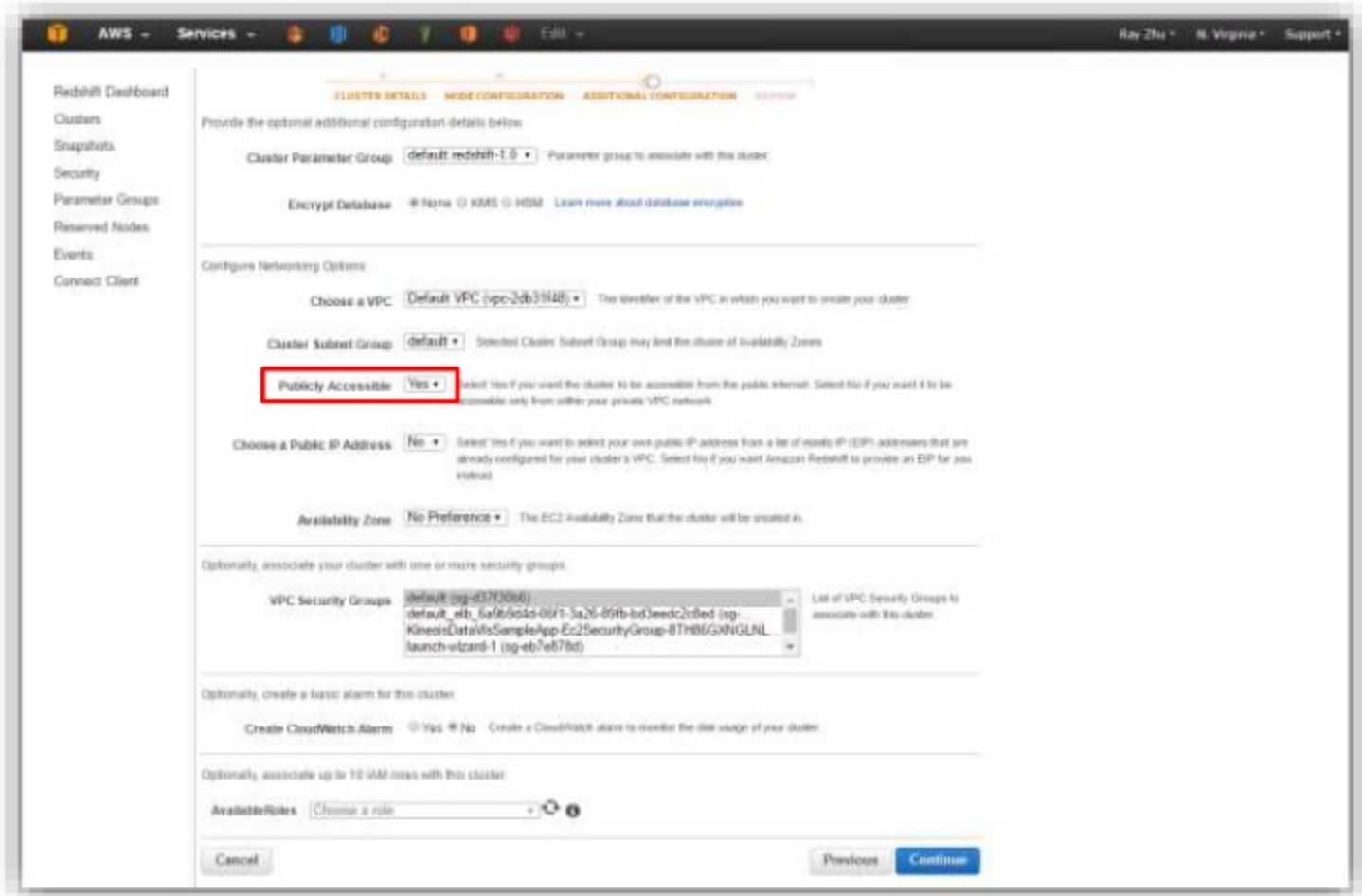
Cluster Details



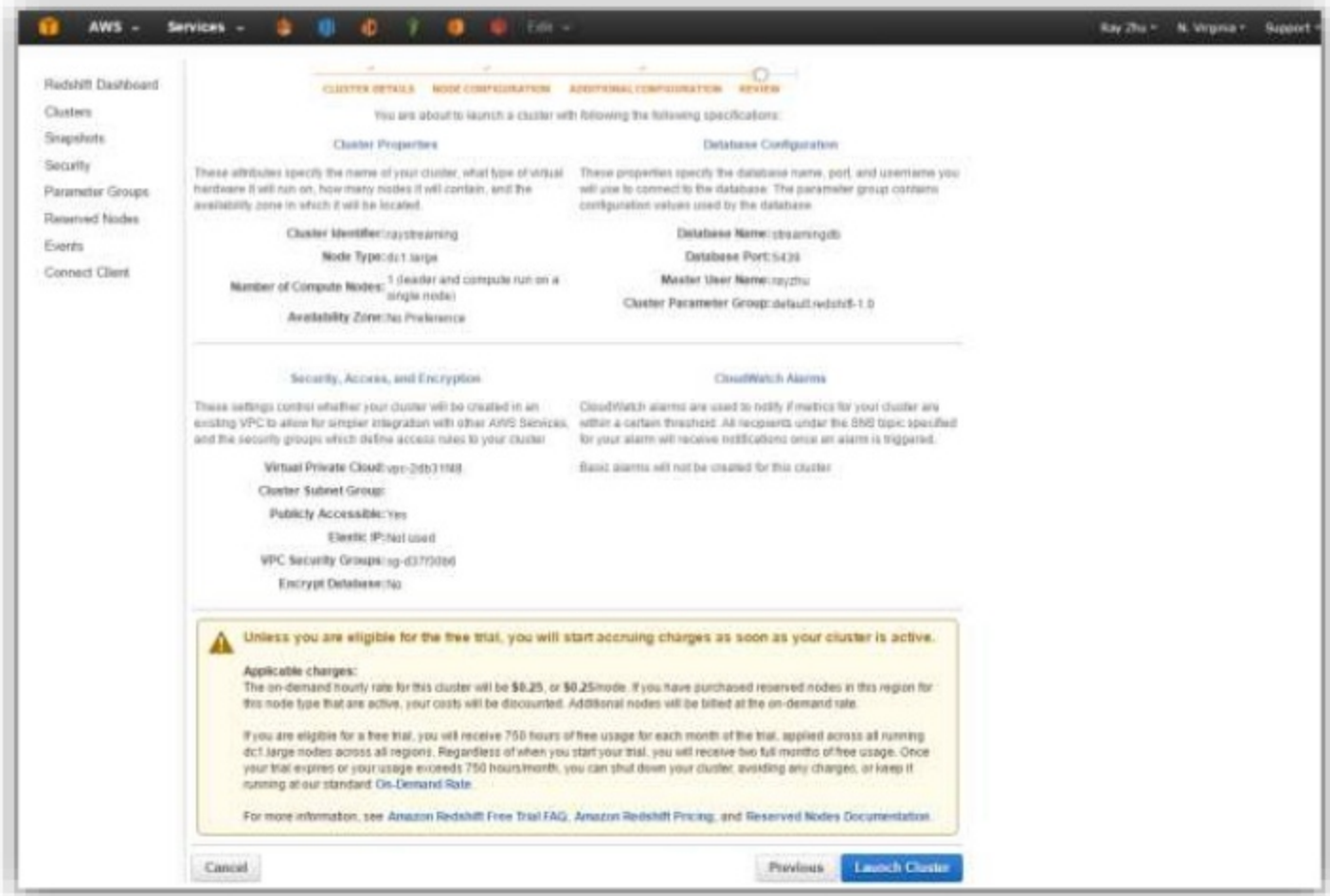
Node Configuration



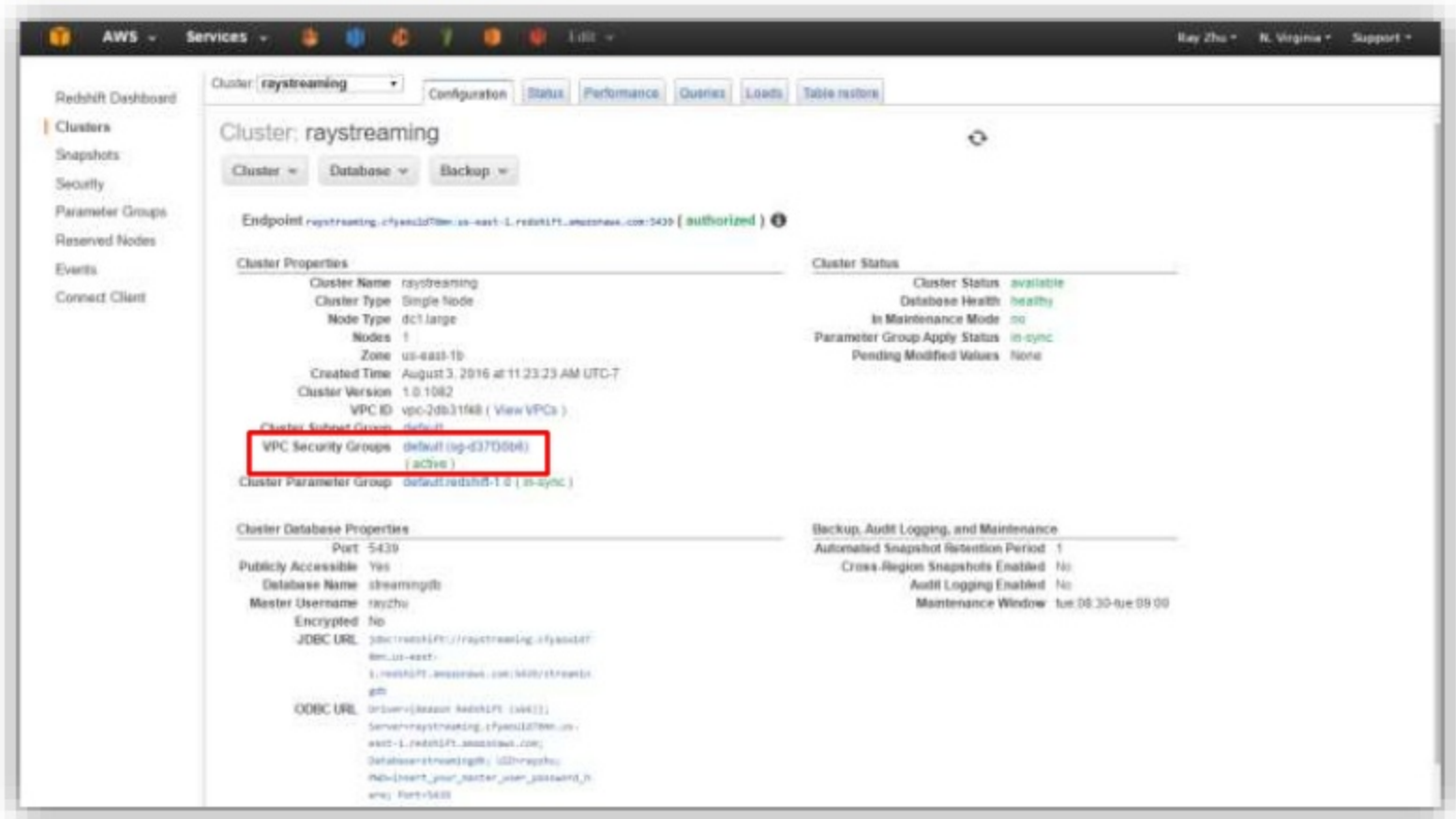
Additional Configuration



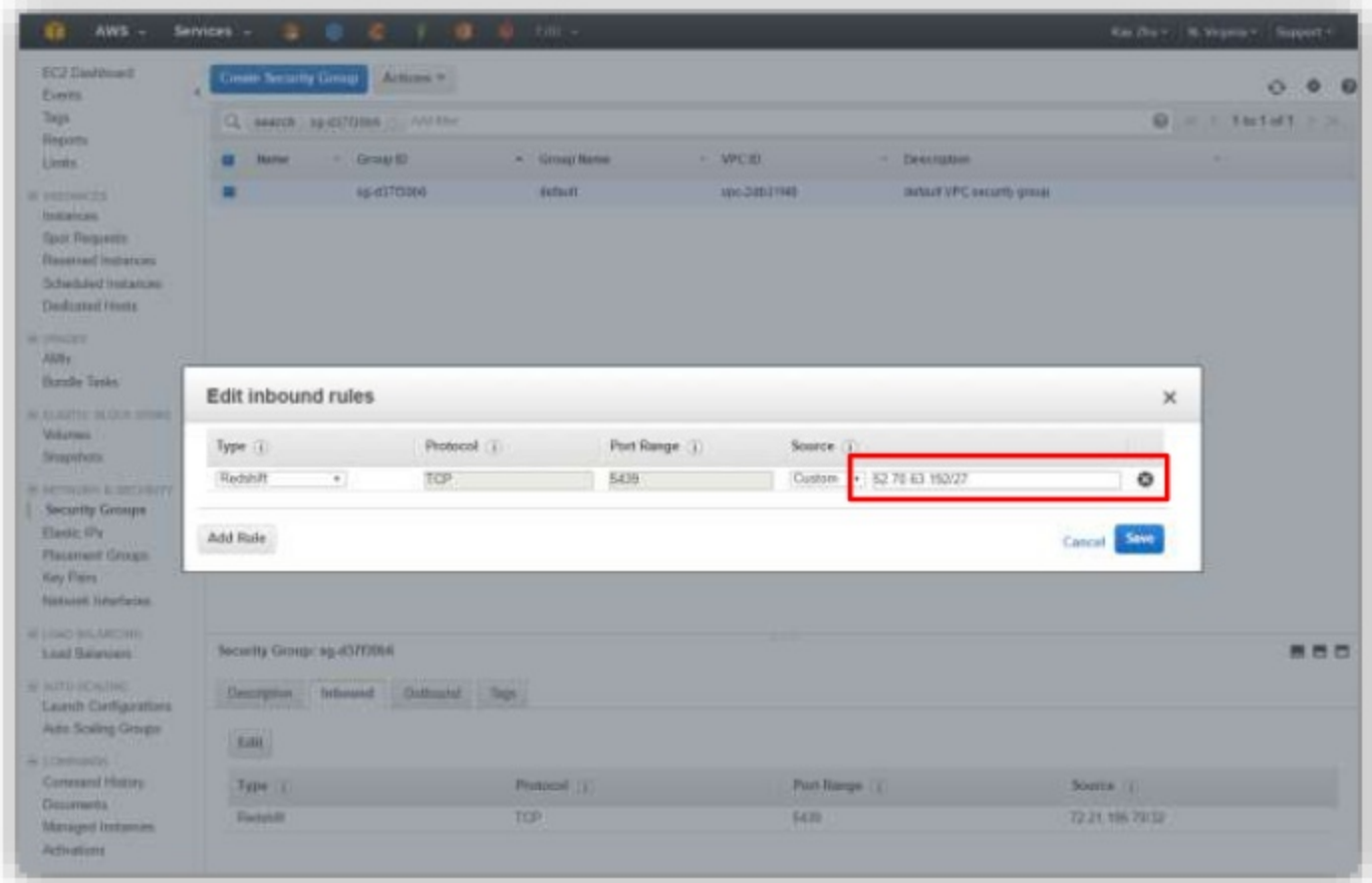
Review


AWS
Services
Ray Zhu
N. Virginia
Support
Redshift Dashboard
Clusters
Snapshots
Security
Parameter Groups
Reserved Nodes
Events
Connect Client
CLUSTER DETAILS
NODE CONFIGURATION
ADDITIONAL CONFIGURATION
REVIEW
You are about to launch a cluster with following the following specifications:
Cluster Properties
These attributes specify the name of your cluster, what type of virtual hardware it will run on, how many nodes it will contain, and the availability zone in which it will be located.
Cluster Identifier: raystreaming
Node Type: dc1.large
Number of Compute Nodes: 1 (leader and compute run on a single node)
Availability Zone: No Preference
Database Configuration
These properties specify the database name, port, and username you will use to connect to the database. The parameter group contains configuration values used by the database.
Database Name: streamingdb
Database Port: 5439
Master User Name: rayzhu
Cluster Parameter Group: default.redshift-1.0
Security, Access, and Encryption
These settings control whether your cluster will be created in an existing VPC to allow for simpler integration with other AWS Services, and the security groups which define access rules to your cluster.
Virtual Private Cloud: vpc-2db31f48
Cluster Subnet Group:
Publicly Accessible: Yes
Elastic IP: Not used
VPC Security Groups: sg-d37f30b6
Encrypt Database: No
CloudWatch Alarms
CloudWatch alarms are used to notify if metrics for your cluster are within a certain threshold. All recipients under the SNS topic specified for your alarm will receive notifications once an alarm is triggered.
Basic alarms will not be created for this cluster.
Unless you are eligible for the free trial, you will start accruing charges as soon as your cluster is active.
Applicable charges:
The on-demand hourly rate for this cluster will be $0.25, or $0.25/node. If you have purchased reserved nodes in this region for this node type that are active, your costs will be discounted. Additional nodes will be billed at the on-demand rate.
If you are eligible for a free trial, you will receive 750 hours of free usage for each month of the trial, applied across all running dc1.large nodes across all regions. Regardless of when you start your trial, you will receive two full months of free usage. Once your trial expires or your usage exceeds 750 hours/month, you can shut down your cluster, avoiding any charges, or keep it running at our standard On-Demand Rate.
For more information, see Amazon Redshift Free Trial FAQ, Amazon Redshift Pricing, and Reserved Nodes Documentation.
Cancel
Previous
Launch Cluster

Configure VPC Security Group



Configure VPC Security Group

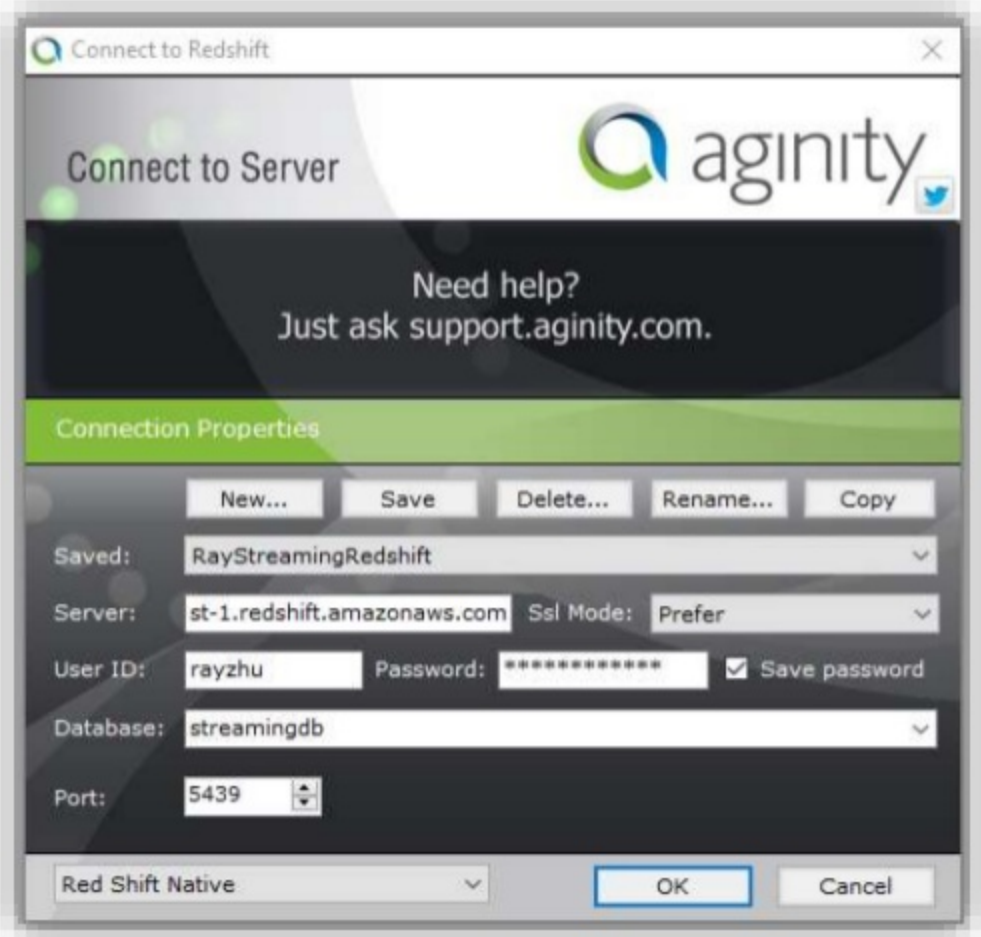


US East (N. Virginia)
52.70.63.192/27

US West (Oregon)
52.89.255.224/27

EU (Ireland)
52.19.239.192/27

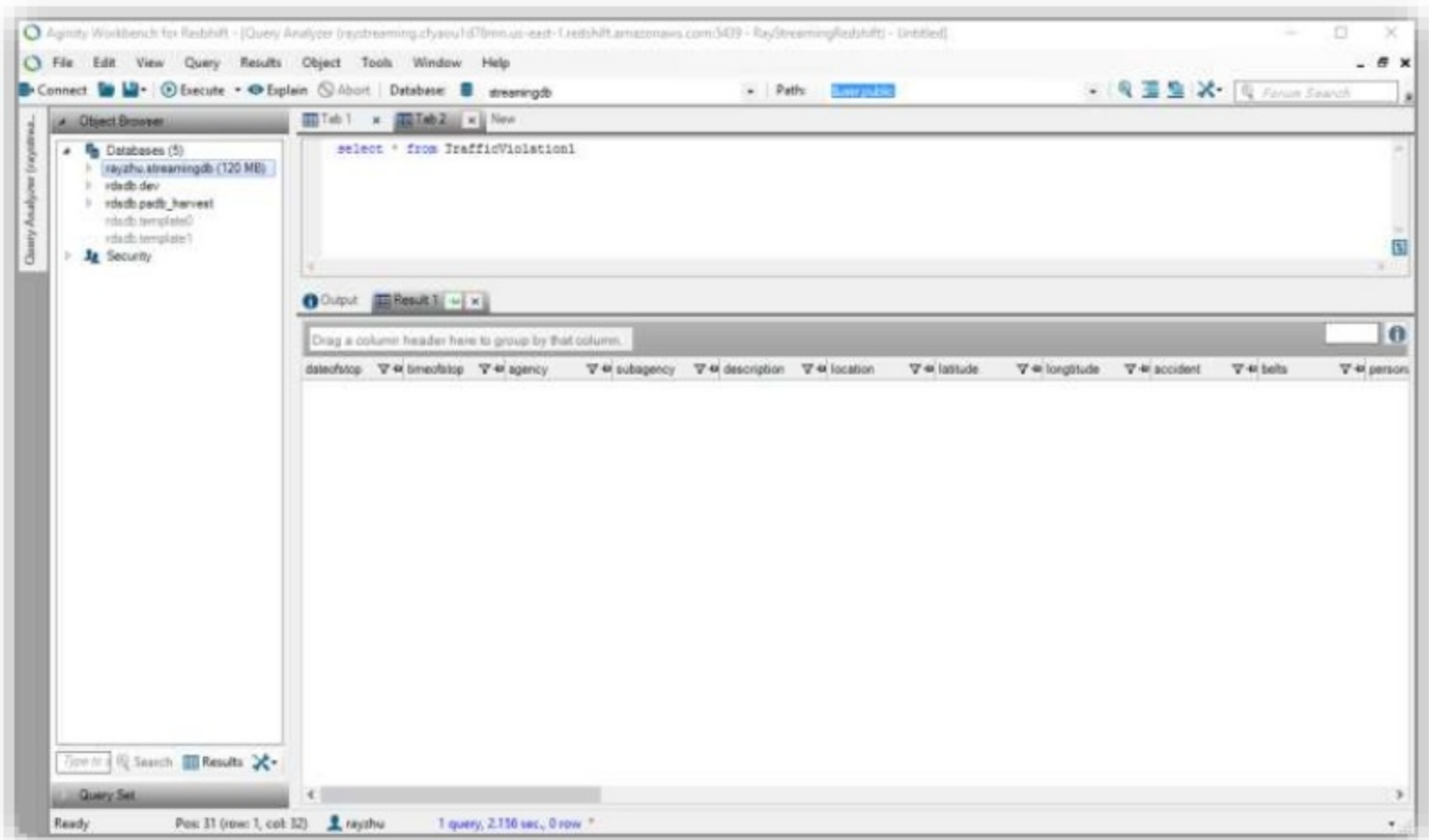
Connect to Redshift DB



Create a Redshift Table

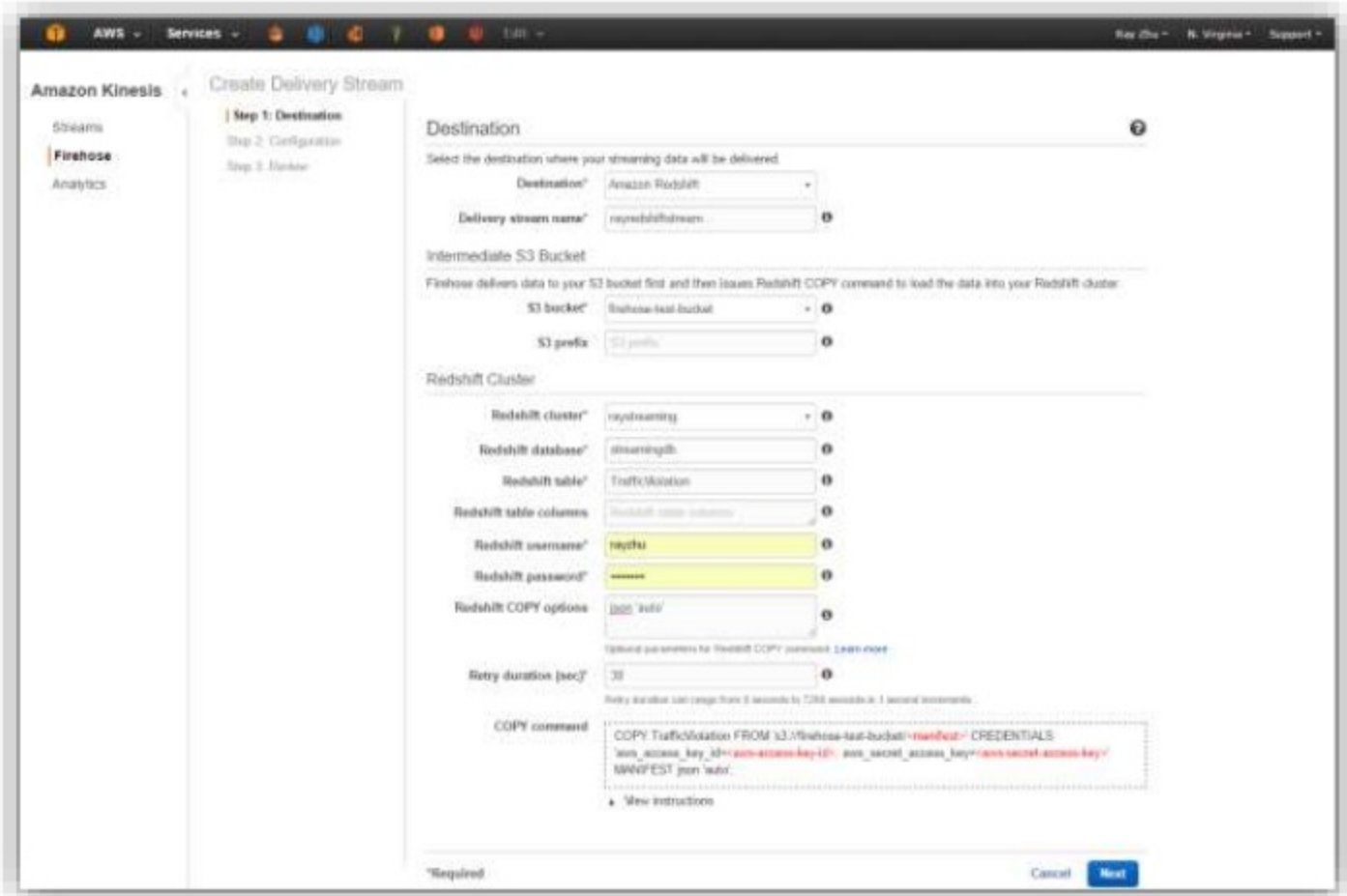
```
create table TrafficViolation(
dateofstop date,
timeofstop timestamp,
agency varchar(100),
subagency varchar(100),
description varchar(300),
location varchar(100),
latitude varchar(100),
longtitude varchar(100),
accident varchar(100),
belts varchar(100),
personalinjury varchar(100),
propertydamage varchar(100),
fatal varchar(100),
commlicense varchar(100),
hazmat varchar(100),
commvehicle varchar(100),
alcohol varchar(100),
workzone varchar(100),
state varchar(100),
veichletype varchar(100),
year varchar(100),
make varchar(100),
model varchar(100),
color varchar(100),
violation varchar(100),
type varchar(100),
charge varchar(100),
article varchar(100),
contributed varchar(100),
race varchar(100),
gender varchar(100),
drivercity varchar(100),
driverstate varchar(100),
dlstate varchar(100),
arresttype varchar(100),
geolocation varchar(100));
```

Create a Redshift Table

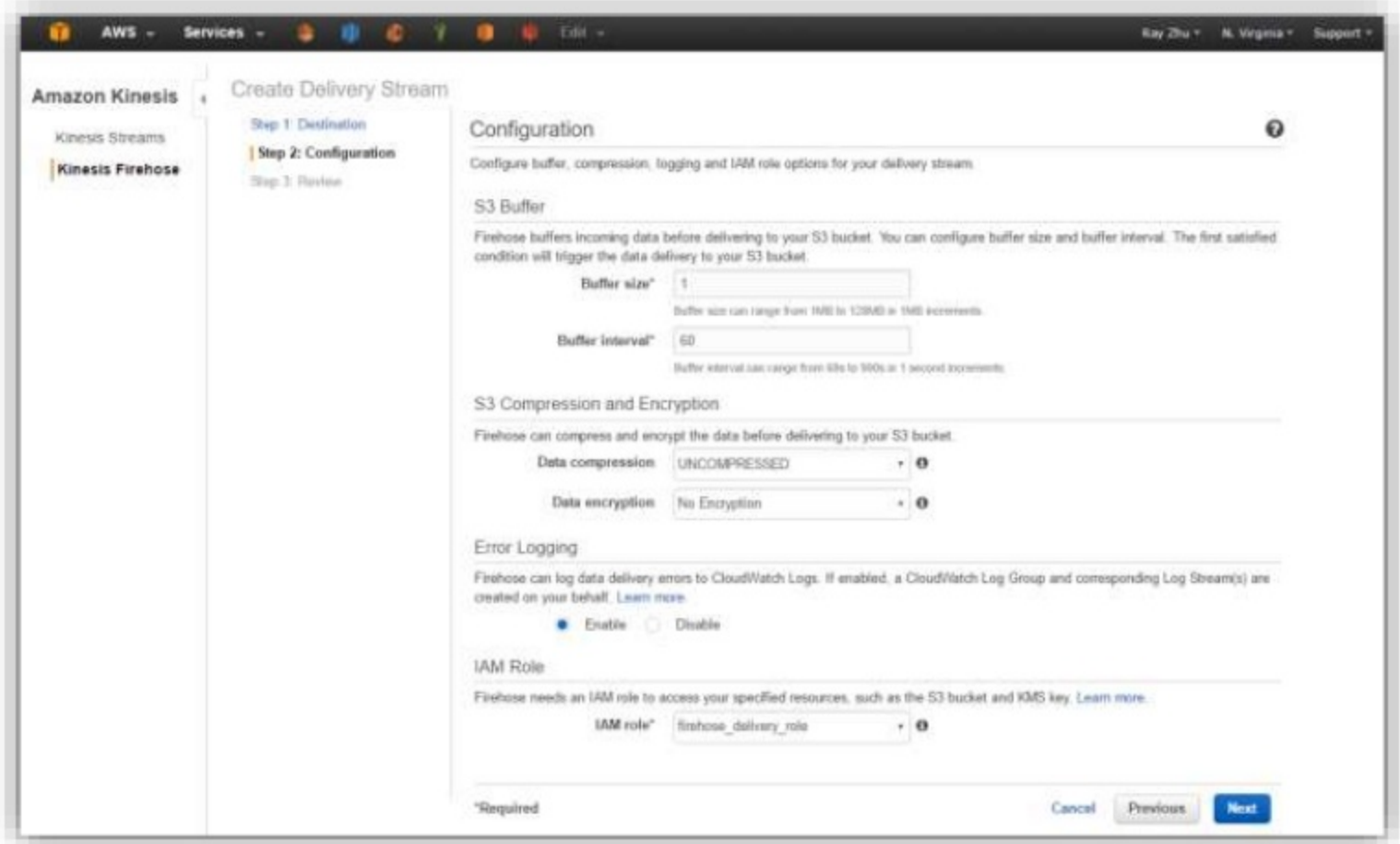


Step 2 Set Up Firehose Delivery Stream

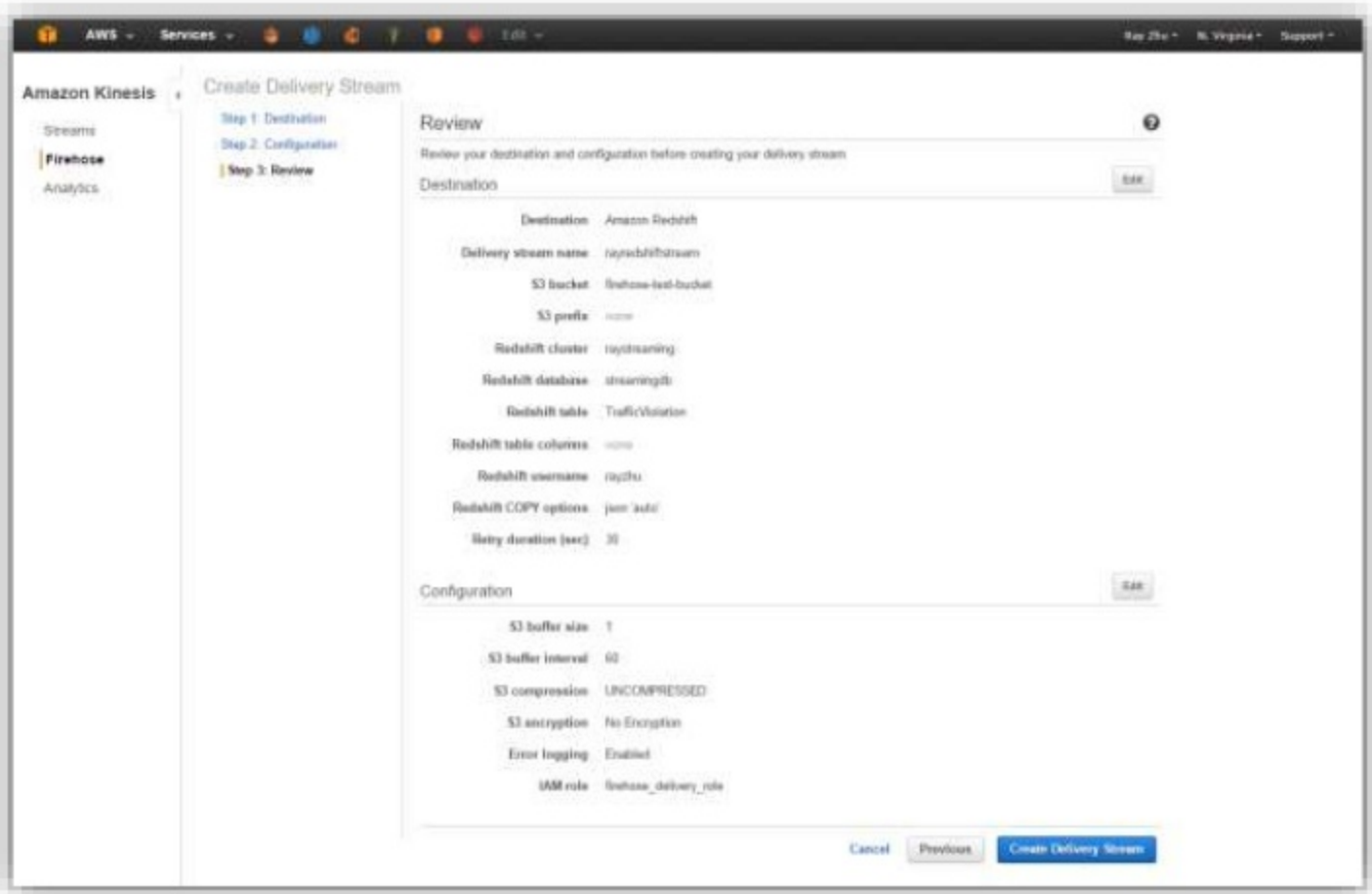
Destination



Configuration



Review


AWS
Services
Ray Zhu
N. Virginia
Support
Amazon Kinesis
Streams
Firehose
Analytics
Create Delivery Stream
Step 1: Destination
Step 2: Configuration
Step 3: Review
Review
Review your destination and configuration before creating your delivery stream.
Destination
Edit
Destination Amazon Redshift
Delivery stream name raytestfirehosestream
S3 bucket firehose-test-bucket
S3 prefix none
Redshift cluster raystreaming
Redshift database streamingdb
Redshift table TrafficViolation
Redshift table columns none
Redshift username raydu
Redshift COPY options json 'auto'
Retry duration (sec) 30
Configuration
Edit
S3 buffer size 1
S3 buffer interval 60
S3 compression UNCOMPRESSED
S3 encryption No Encryption
Error logging Enabled
IAM role firehose_delivery_role
Cancel
Previous
Create Delivery Stream

Step 3 Send Data to Firehose Delivery Stream

Sample Data

US Government Open Data: https://catalog.data.gov/dataset/traffic-violations-56dda

09/30/2014,23:51:00,MCP,"1st district, Rockville",DRIVER FAILURE TO STOP AT STEADY CIRCULAR RED SIGNAL,PARK RD AT HUNGERFORD DR,,,No,No,No,No,No,No,No,No,No,No,MD,02 - Automobile,2014,FORD,MUSTANG,BLACK,Citation,21-202(h1),Transportation Article,No,BLACK,M,ROCKVILLE,MD,MD,A - Marked Patrol,

03/31/2015,23:59:00,MCP,"2nd district, Bethesda",HEADLIGHTS (*),CONNECTICUT AT METROPOLITAN AVE,,,No,No,No,No,No,No,No,No,No,No,MD,02 - Automobile,2003,HONDA,2S,BLUE,ESERO,55*,,No,HISPANIC,M,SILVER SPRING,MD,MD,A - Marked Patrol,

09/30/2014,23:30:00,MCP,"5th district, Germantown",FAILURE TO DISPLAY TWO LIGHTED FRONT LAMPS WHEN REQUIRED,OBSERVATION @ RIDGE ROAD,,,No,No,No,No,No,No,No,No,No,No,MD,02 - Automobile,2009,TOYOTA,CAMRY,RED,Warning,22-226(a),Transportation Article,No,BLACK,F,GERMANTOWN,MD,MD,A - Marked Patrol,

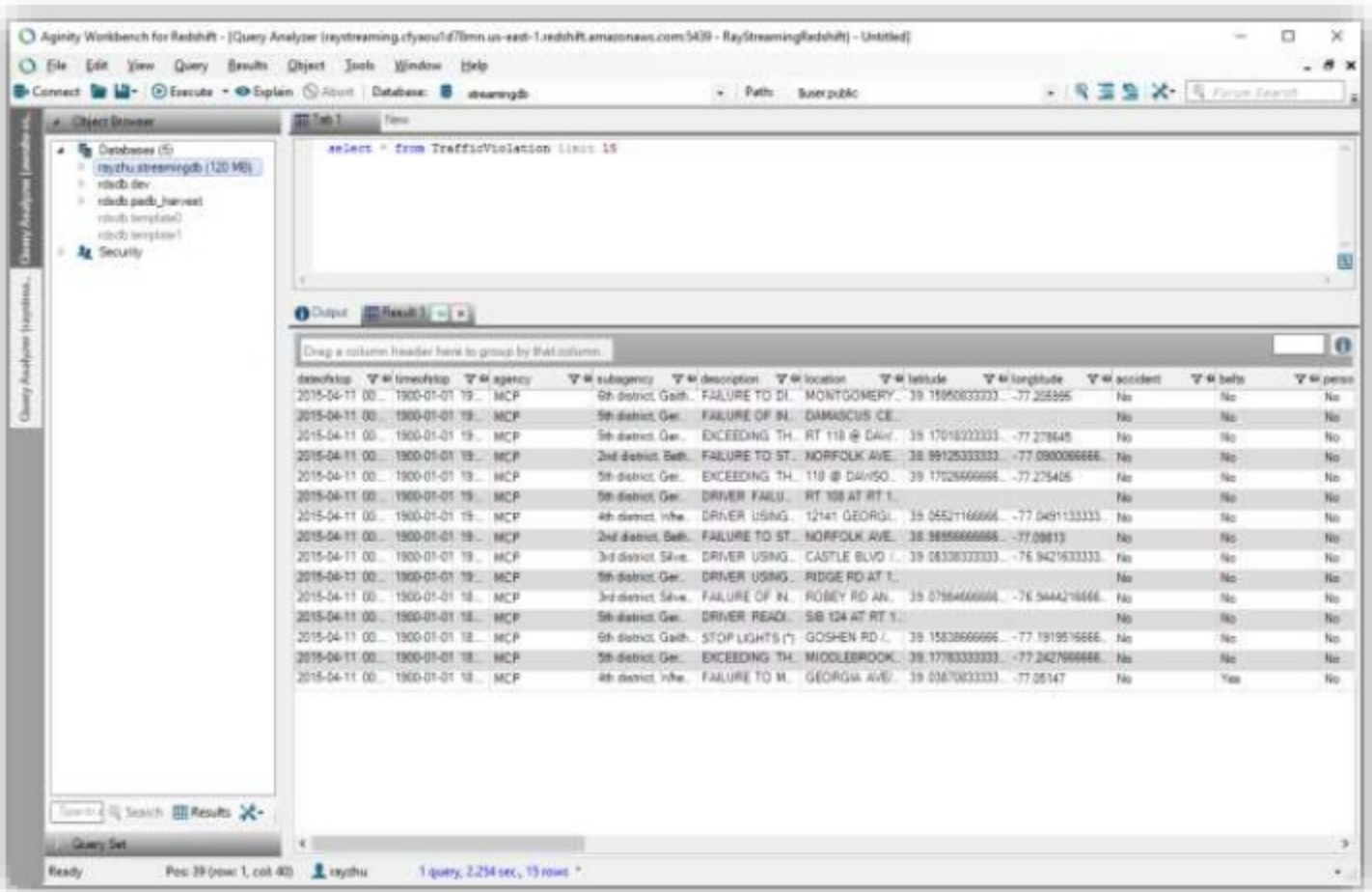
03/31/2015,23:59:00,MCP,"5th district, Germantown",DRIVER FAILURE TO STOP AT STOP SIGN LINE,W/B PLYERS MILL RD AT METROPOLITAN AVE,39.0338233333333,-77.07544,No,No,No,No,No,No,No,No,No,No,MD,02 - Automobile,2007,ACURA,MDX,BLACK,Warning,21-707(a),Transportation Article,No,WHITE,F,KENSINGTON,MD,MD,A - Marked Patrol,"(39.0338233333333, -77.07544)"

Send Data

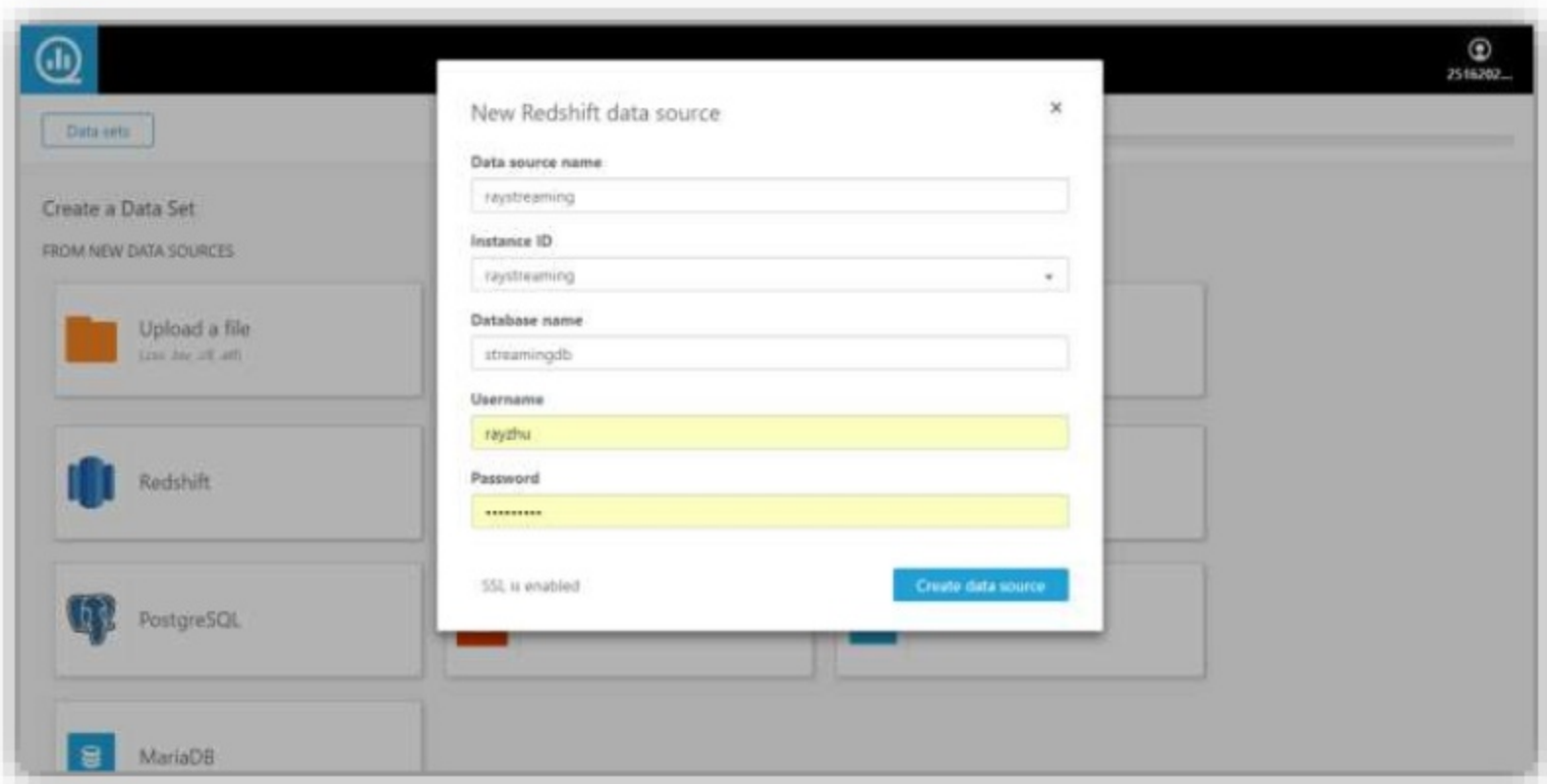
```
PutRecordRequest putRecordRequest = new PutRecordRequest();
putRecordRequest.setDeliveryStreamName(deliveryStreamName);

String data = line + "\n";
Record record = createRecord(data);
putRecordRequest.setRecord(record);

FirehoseClient.putRecord(putRecordRequest);
```

Step 4 Query and Analyze the Data from Redshift

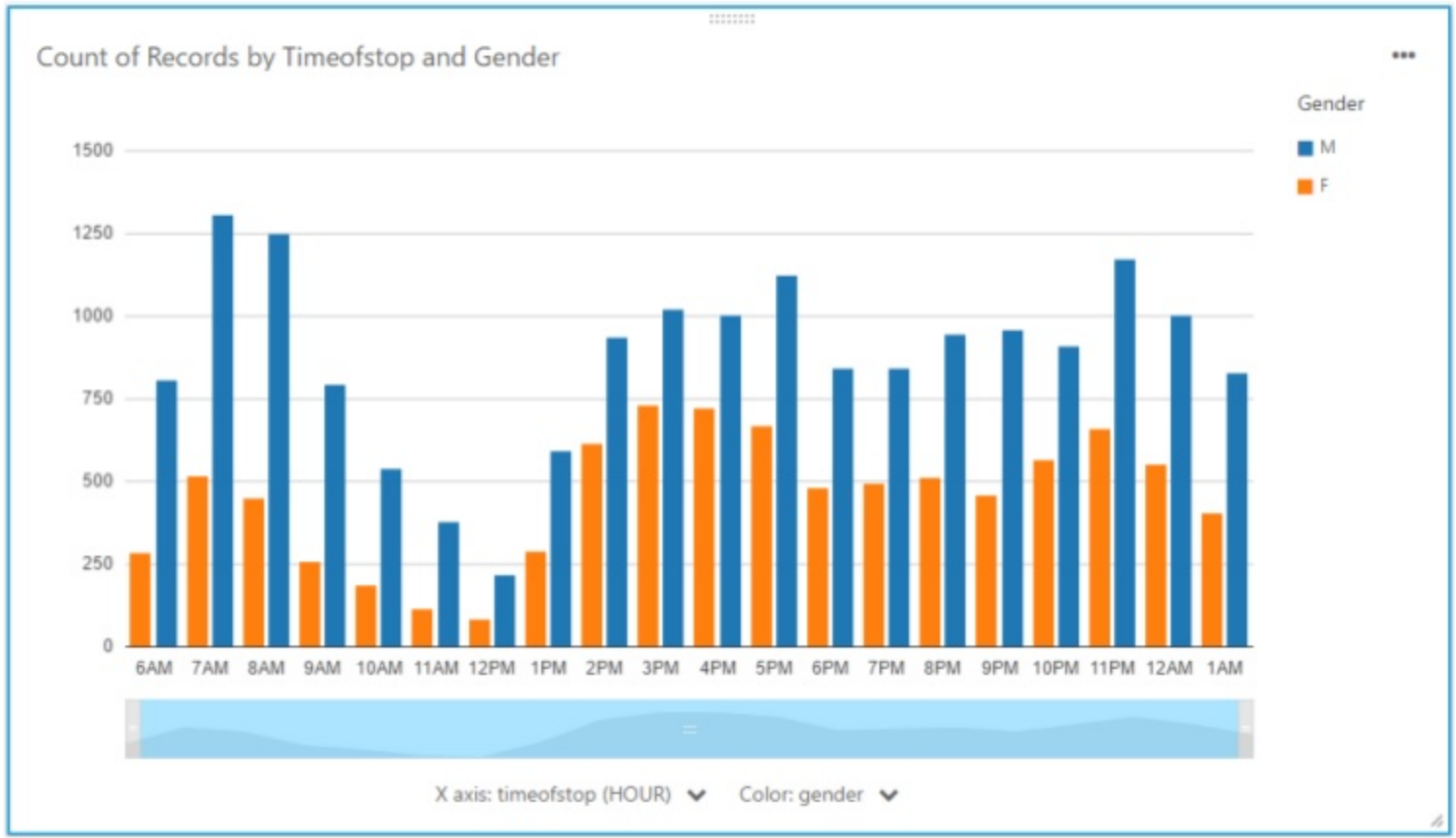
Query Data


Aginity Workbench for Redshift - [Query Analyzer (raystreaming.cfyaou1d78mn.us-east-1.redshift.amazonaws.com:5439 - RayStreamingRedshift) - Untitled]
File Edit View Query Results Object Tools Window Help
Connect Execute Explain Abort Database: streamingdb Path: $user,public
Object Browser
Databases (5)
raythu.streamingdb (120 MB)
rdsdb.dev
rdsdb.padb_harvest
rdsdb.template0
rdsdb.template1
Security
Tab 1
select * from TrafficViolation limit 15
Output
Result 1
Drag a column header here to group by that column.
datevioltop timevioltop agency subagency description location latitude longitude accident belts perso
2015-04-11 00... 1900-01-01 19... MCP 6th district, Gaith... FAILURE TO DI... MONTGOMERY... 39.1695833333... -77.205995 No No No
2015-04-11 00... 1900-01-01 19... MCP 5th district, Ger... FAILURE OF IN... DAMASCUS CE... No No No
2015-04-11 00... 1900-01-01 19... MCP 5th district, Ger... EXCEEDING TH... RT 118 @ DAV... 39.1701833333... -77.278645 No No No
2015-04-11 00... 1900-01-01 19... MCP 2nd district, Beth... FAILURE TO ST... NORFOLK AVE... 38.9912533333... -77.0900666666... No No No
2015-04-11 00... 1900-01-01 19... MCP 5th district, Ger... EXCEEDING TH... 118 @ DAWSO... 39.1702666666... -77.275405 No No No
2015-04-11 00... 1900-01-01 19... MCP 5th district, Ger... DRIVER FAILU... RT 108 AT RT 1... No No No
2015-04-11 00... 1900-01-01 19... MCP 4th district, Whe... DRIVER USING... 12141 GEORGI... 39.0552116666... -77.0491133333... No No No
2015-04-11 00... 1900-01-01 19... MCP 2nd district, Beth... FAILURE TO ST... NORFOLK AVE... 38.9895666666... -77.09813 No No No
2015-04-11 00... 1900-01-01 19... MCP 3rd district, Silve... DRIVER USING... CASTLE BLVD /... 39.0838333333... -76.9421633333... No No No
2015-04-11 00... 1900-01-01 19... MCP 5th district, Ger... DRIVER USING... RIDGE RD AT 1... No No No
2015-04-11 00... 1900-01-01 18... MCP 3rd district, Silve... FAILURE OF IN... ROBEY RD AN... 39.0766466666... -76.9444216666... No No No
2015-04-11 00... 1900-01-01 18... MCP 5th district, Ger... DRIVER READI... SB 124 AT RT 1... No No No
2015-04-11 00... 1900-01-01 18... MCP 6th district, Gaith... STOP LIGHTS (*) GOSHEN RD /... 39.1583866666... -77.1919516666... No No No
2015-04-11 00... 1900-01-01 18... MCP 5th district, Ger... EXCEEDING TH... MIDDLEBROOK... 39.1778333333... -77.2427666666... No No No
2015-04-11 00... 1900-01-01 18... MCP 4th district, Whe... FAILURE TO M... GEORGIA AVE... 39.0367083333... -77.0514? No Yes No
Search Results
Query Set
Ready Pos: 39 (row: 1, col: 40) raythu 1 query, 2.254 sec, 15 rows

Connect via QuickSight

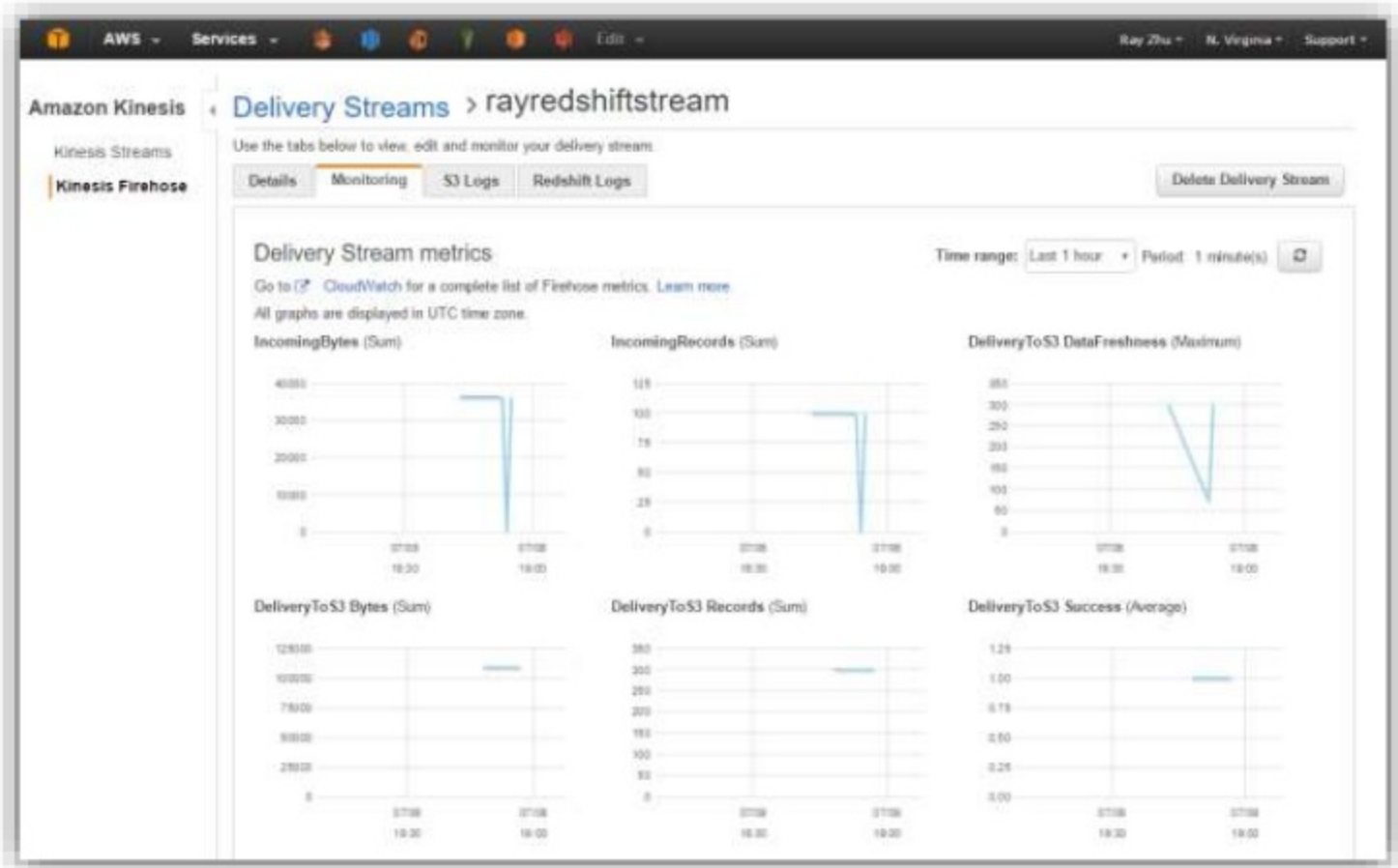


Visualize Data

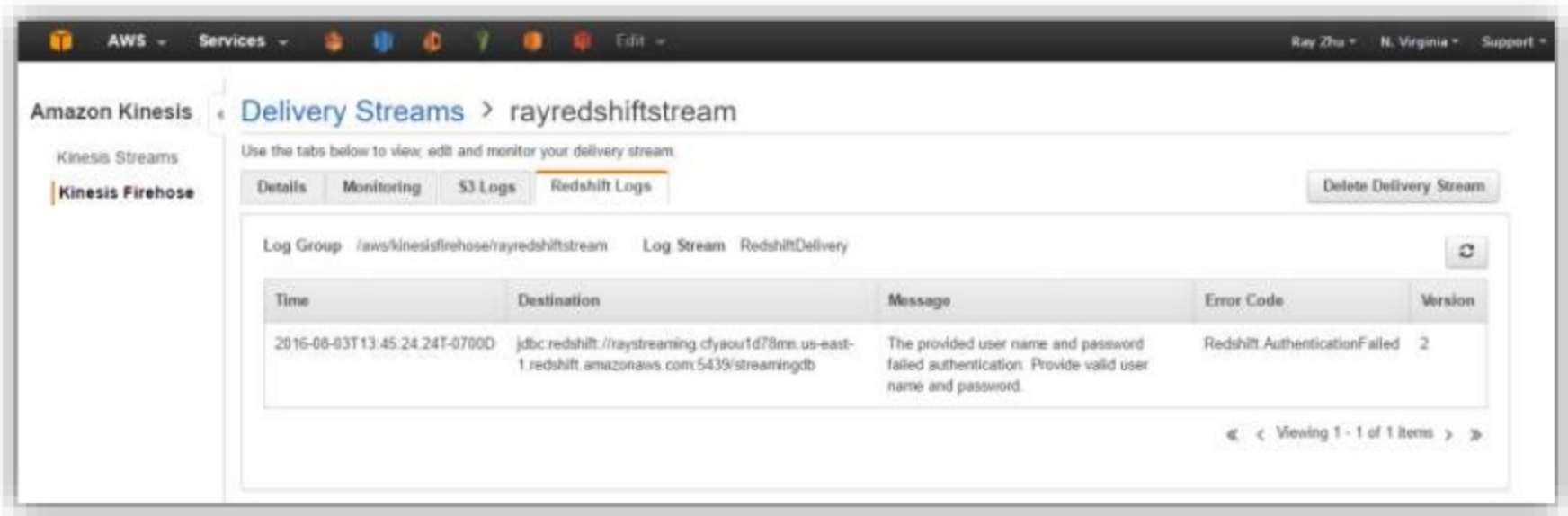


Step 5 Monitor Streaming Data Pipeline

Monitor with CloudWatch Metrics



Monitor with CloudWatch Logs



Q & A

Thank you!