**December, 2018**



**Immersion Day**

Serverless Data Lake with AWS

# Table of Contents

[Table of Contents 2](#_Toc1030540)

[Revision History 3](#_Toc1030541)

[Disclaimer 3](#_Toc1030542)

[Lab Pre-Requisites 4](#_Toc1030543)

[Expected Costs 4](#_Toc1030544)

[Lab 3.1 – [Optional] Data Visualization with AWS & 3PP Tools 5](#_Toc1030545)

[Lab 3.1.1 Querying and Visualising Data with Athena & Quicksight 5](#_Toc1030546)

[Additional Lab Prerequisites for this Optional Lab 5](#_Toc1030547)

[Lab 3.1.2 [Optional] Querying Dataset with JDBC Endpoints using SQL Workbench 11](#_Toc1030548)

[Additional Lab Prerequisites fort this Optional Lab 12](#_Toc1030549)

[Summary & Next Steps 16](#_Toc1030550)

[Deleting Lab Resources 17](#_Toc1030551)

[References 18](#_Toc1030552)

# Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| Revision | Date | By | Change Log |
| PA6 | 03-Dec-2018 | @akirmak | First 3 labs reer reviewed by @baya |
| PA7 | 13-Dec-2018 | @akirmak | Initial draft, ready for internal joint review. |
| PA8 | 14-Dec-2018 | @akirmak | Joint reviewed by @halilb and @serdarn |
| PA9 | 16-Dec-2018 | @akirmak | Joint review feedback incorporated. Fixed an error in Lab 2.4 instructions. Shortened Lab 3.1 |

# Disclaimer

The work is provided “as is” without warranties or conditions of any kind, either express or implied, including warranties.

# Lab Pre-Requisites

In order to complete this immersion day lab, you need to complete Labs 1 and 2. If you haven’t performed Lab 1 & 2, and would like to work on a different dataset, you can still follow the instructions in this lab, but you may need to change steps slightly (e.g. S3 path, IAM role etc.)

# Expected Costs

* You are responsible for the cost of the AWS services used while running this lab. As of the date of publication, the baseline cost for running this solution as-is should be around:
  + Quicksight: 9 $/month per user

# Lab 3.1 – [Optional] Data Visualization with AWS & 3PP Tools

This is an optional lab, meaning you could do it yourself, or watch the instructor doing.

## Lab 3.1.1 Querying and Visualising Data with Athena & Quicksight

This is a very short lab. You already used Athena in the previous labs.

You will work with AWS services Athena & Quicksight do reporting & visualization on your data.

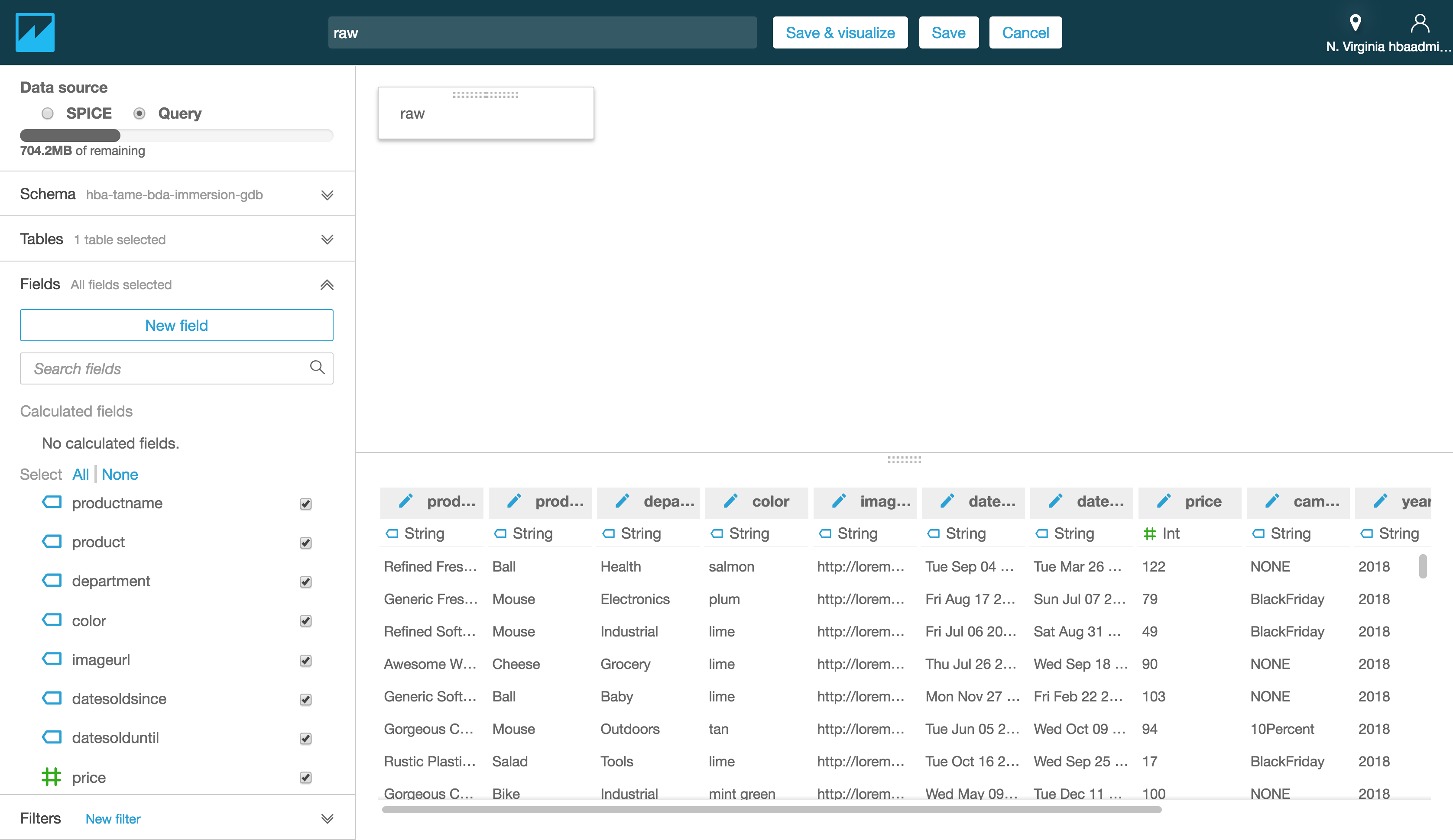
### Additional Lab Prerequisites for this Optional Lab

* **New User:** If you want to do this lab, You need to create a new user for Amazon Quicksight service.
* **Additional Costs:** You can create a free trial account (if you are eligible), or you can create a standar edition user and use the service for a month, which will cost you around 12 USD (billed monthly).
* **User Creation Instructions:** We won’t provide detailed instructions to create Quicksight user. It’s simple. Check 3 minute instruction video here: <https://aws.amazon.com/tr/quicksight/getting-started/>

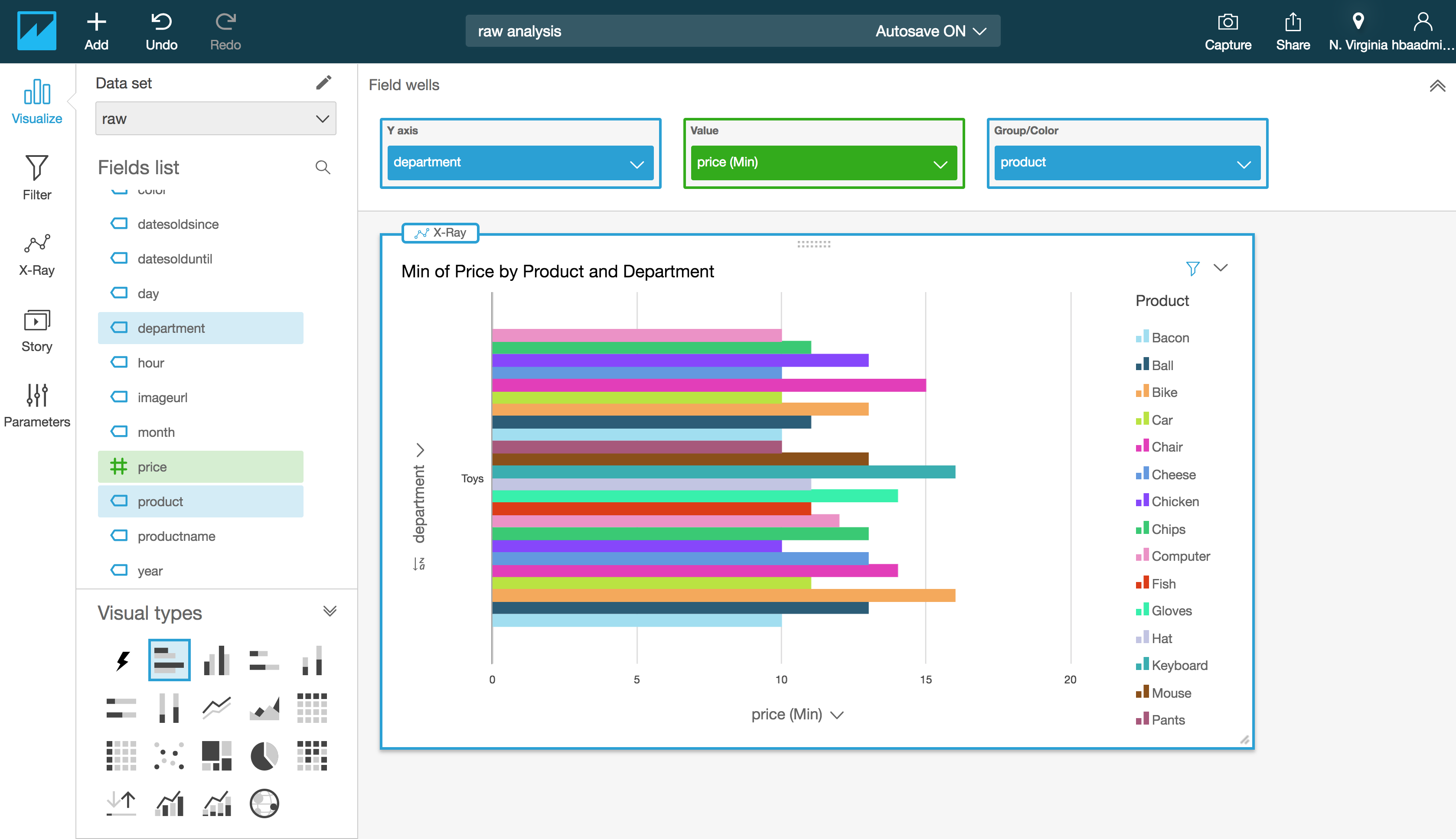
1. Open the Quicksight service. Select create dataset from new Data source. Select “Athena” as data source. Select the DB: “<your-initials>-tame-bda-immersion-gdb” DB. Then select the table: “Raw” that contains the simulated product catalogue data.



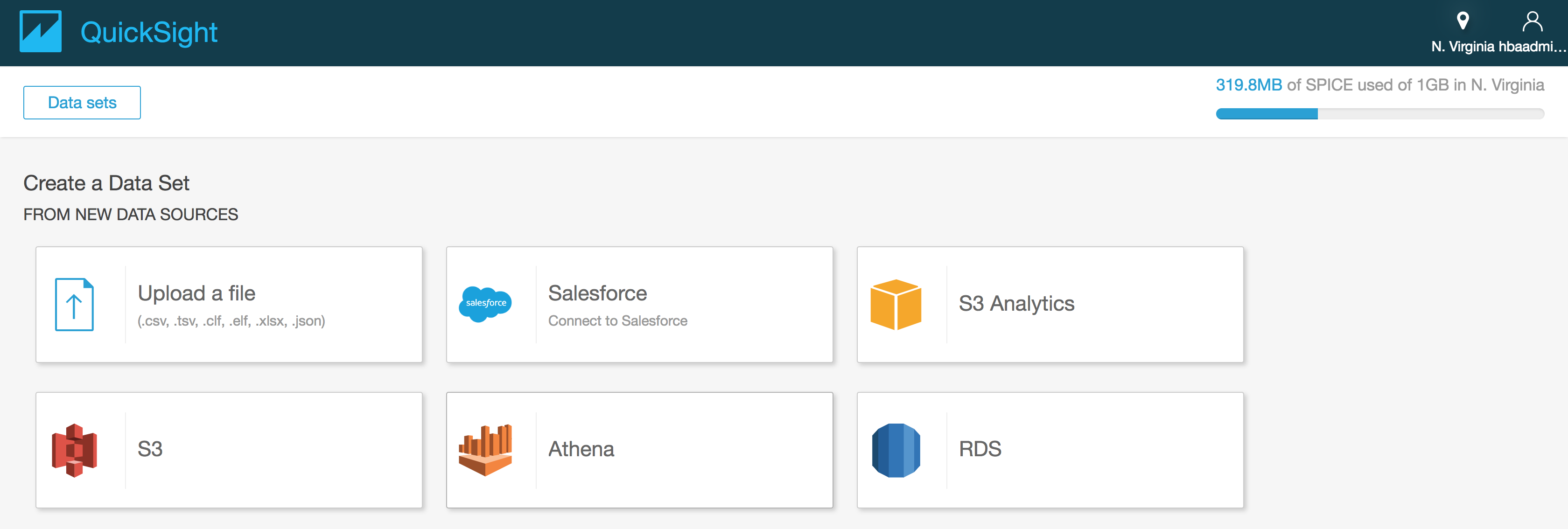
1. Select edit/preview data

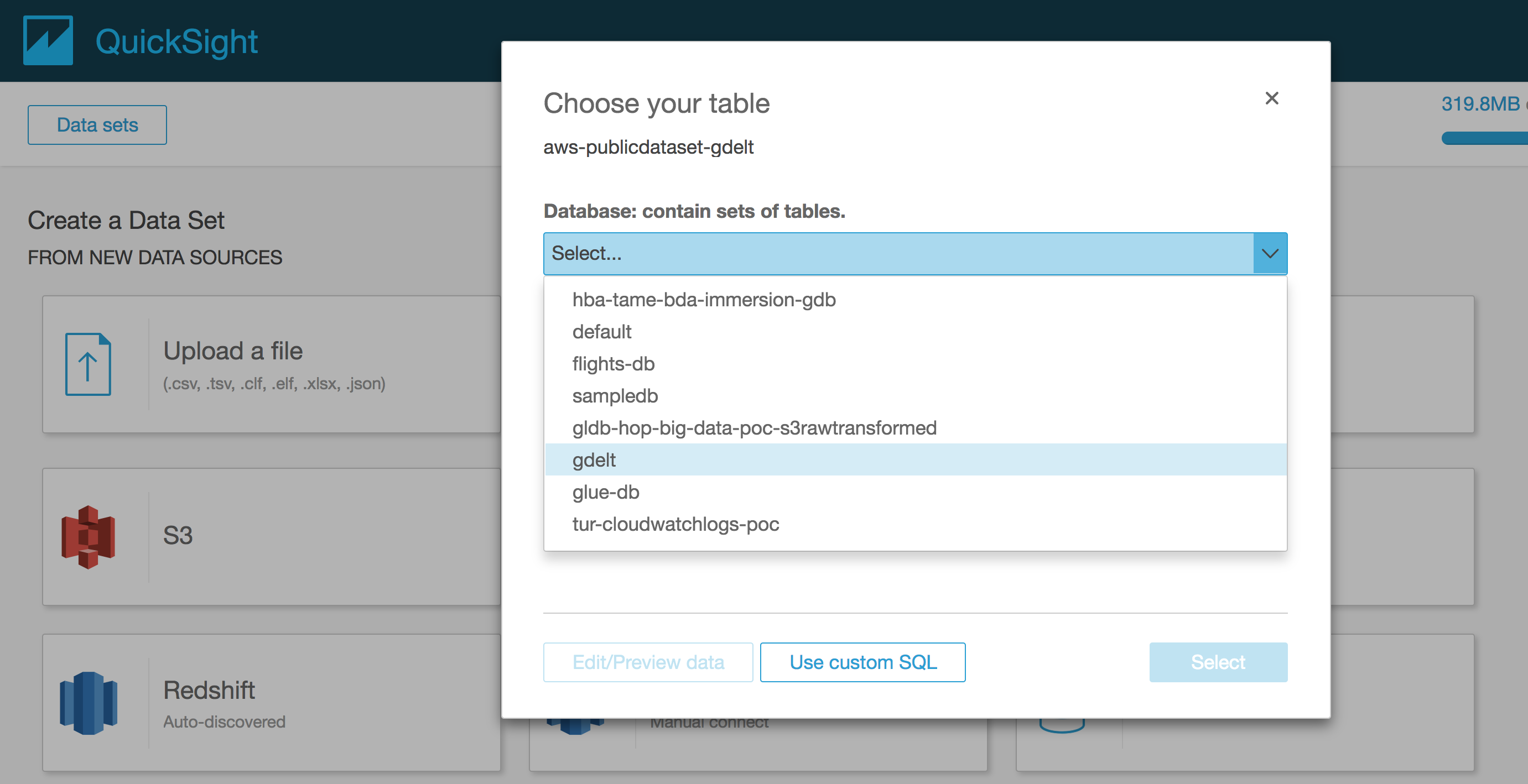


1. Select save & visualize.
2. Select Horizontal Bar Chart as visual type.
3. Select column called “department” and drag and drop it to Y axis field.
4. Select column called “price” and drag and drop it to Value field.
5. Select Group/Color field, and select “product”
6. The diagram will show product prices by department as diagram below:
7. Try creating your own diagram for product catalogue data.

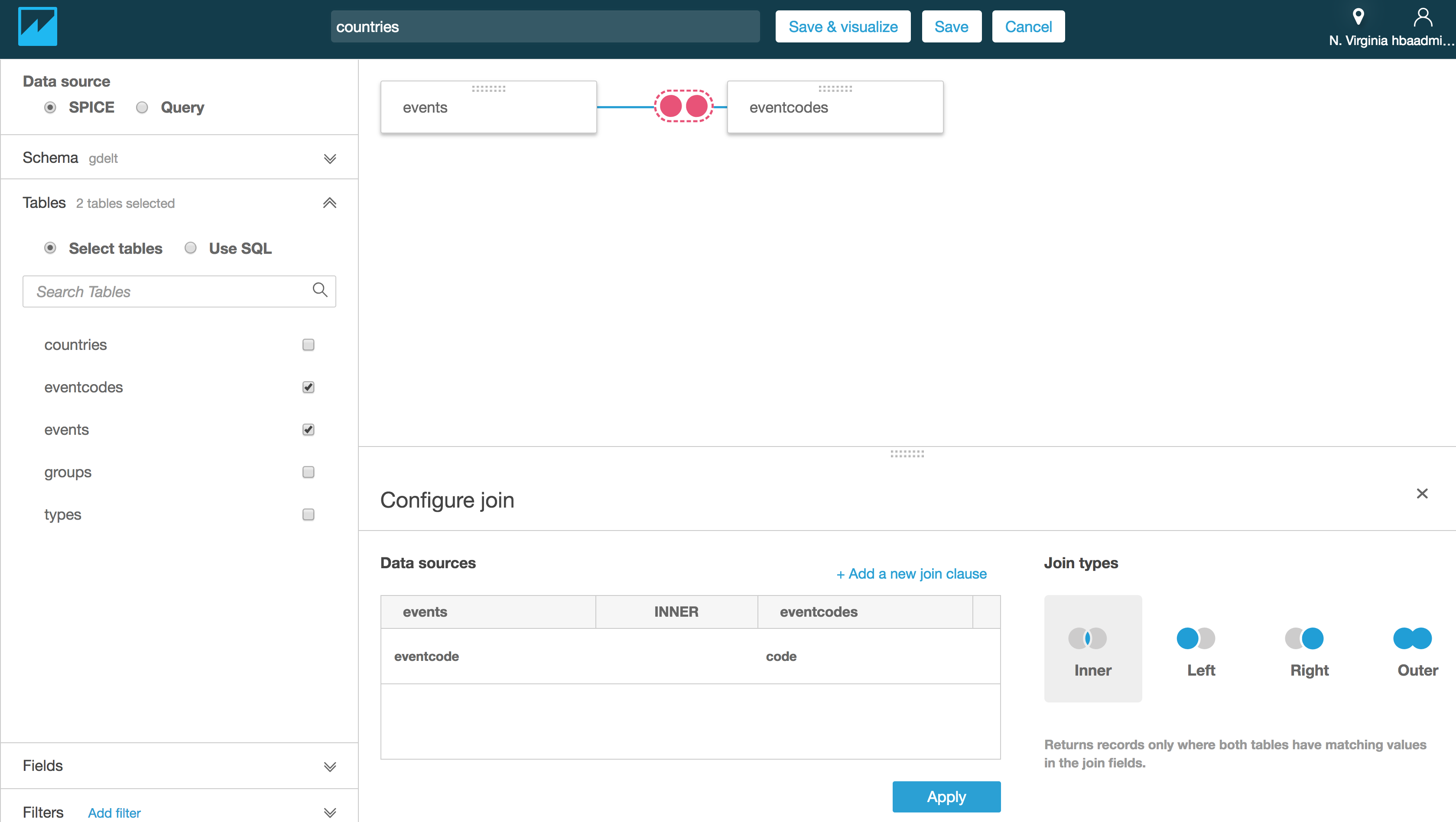


1. Now, let’s work with GDELT data. Click the Athena logo top left to go back to dashboard. Click “Managa Data” and “New Data Set” to create a data set again. Select Athena as Data Source, and name your data source “gdelt”. Select the DB called “gdelt” this time, and choose the “events” table, and then click “Edit/Preview data”.





1. Instead of 1 table we want to join 2 tables. Select “Add data” from the menu and select the “eventcodes” table. You will be presented with options on how to JOIN these tables. Select INNER join (this option returns only where both tables have matching values in the join fields).



1. At this point click “Cancel” in the top menu and restart the process by clicking “New Data Set”. Select Athena as Data Source, and name your data source “gdelt” (again). Instead of picking a table from the dropdown, select “Use custom SQL”, and enter the Athena query from the last lab for “Count Obama/Putin events per category”:

SELECT eventcode,

gdelt.eventcodes.description,

nb\_events

FROM (SELECT gdelt.events.eventcode,

COUNT(gdelt.events.globaleventid) AS nb\_events

FROM gdelt.events

WHERE actor1name='BARACK OBAMA'and actor2name='VLADIMIR PUTIN'

GROUP BY gdelt.events.eventcode

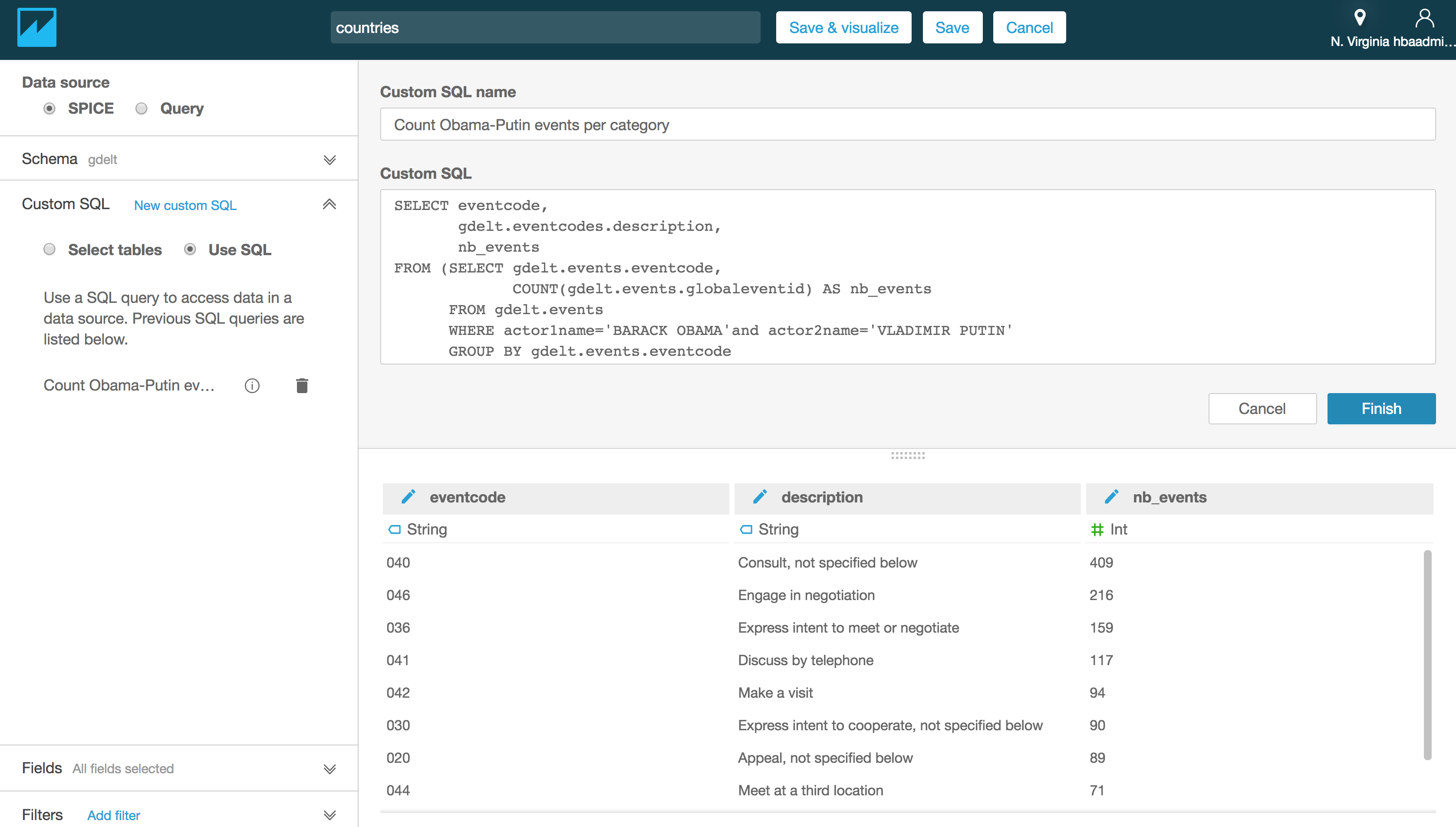
ORDER BY nb\_events DESC)

JOIN gdelt.eventcodes ON eventcode = gdelt.eventcodes.code

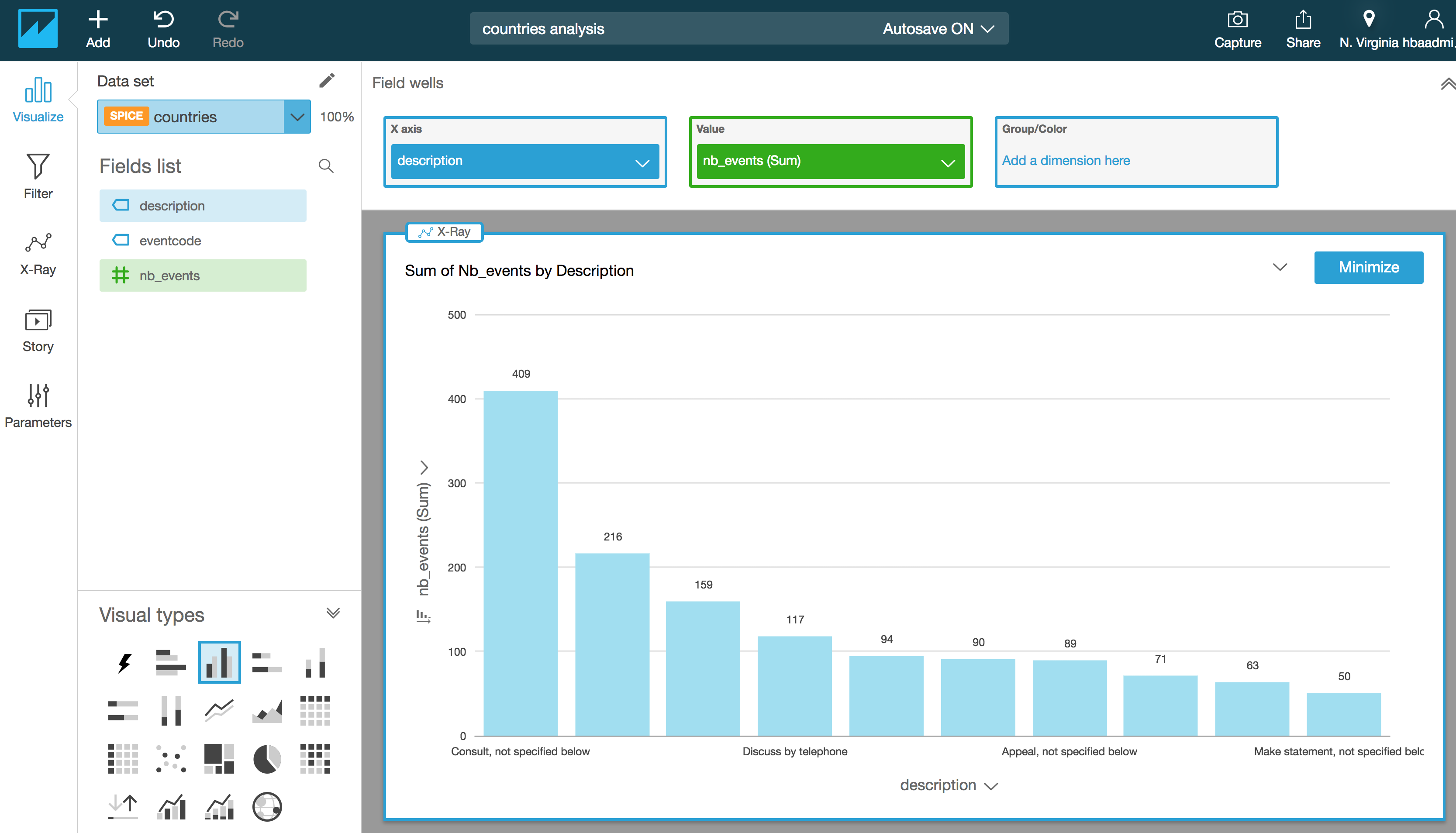
WHERE nb\_events >= 50

ORDER BY nb\_events DESC

1. Click “Confirm Query”. Then select SPICE as data source and click “Visualize”. Note: SPICE is Amazon QuickSight's in-memory optimized calculation engine, designed specifically for fast, ad hoc data visualization. SPICE stores your data in a system architected for high availability, where it is saved until you choose to delete it. You can improve the performance of database data sets by importing the data into SPICE instead of using a direct query to the database.



1. Try creating the graph below:
   1. Visual type: vertical bar chart
   2. X axis = description
   3. Value = nb\_events (Sum)



## Lab 3.1.2 [Optional] Querying Dataset with JDBC Endpoints using SQL Workbench

Many customers use their existing BI tools for reporting and Visualization. You can connect your existing tools to your data lake with typical JDBC connections.

This is an optional lab, meaning you could do it yourself, or watch the instructor doing.

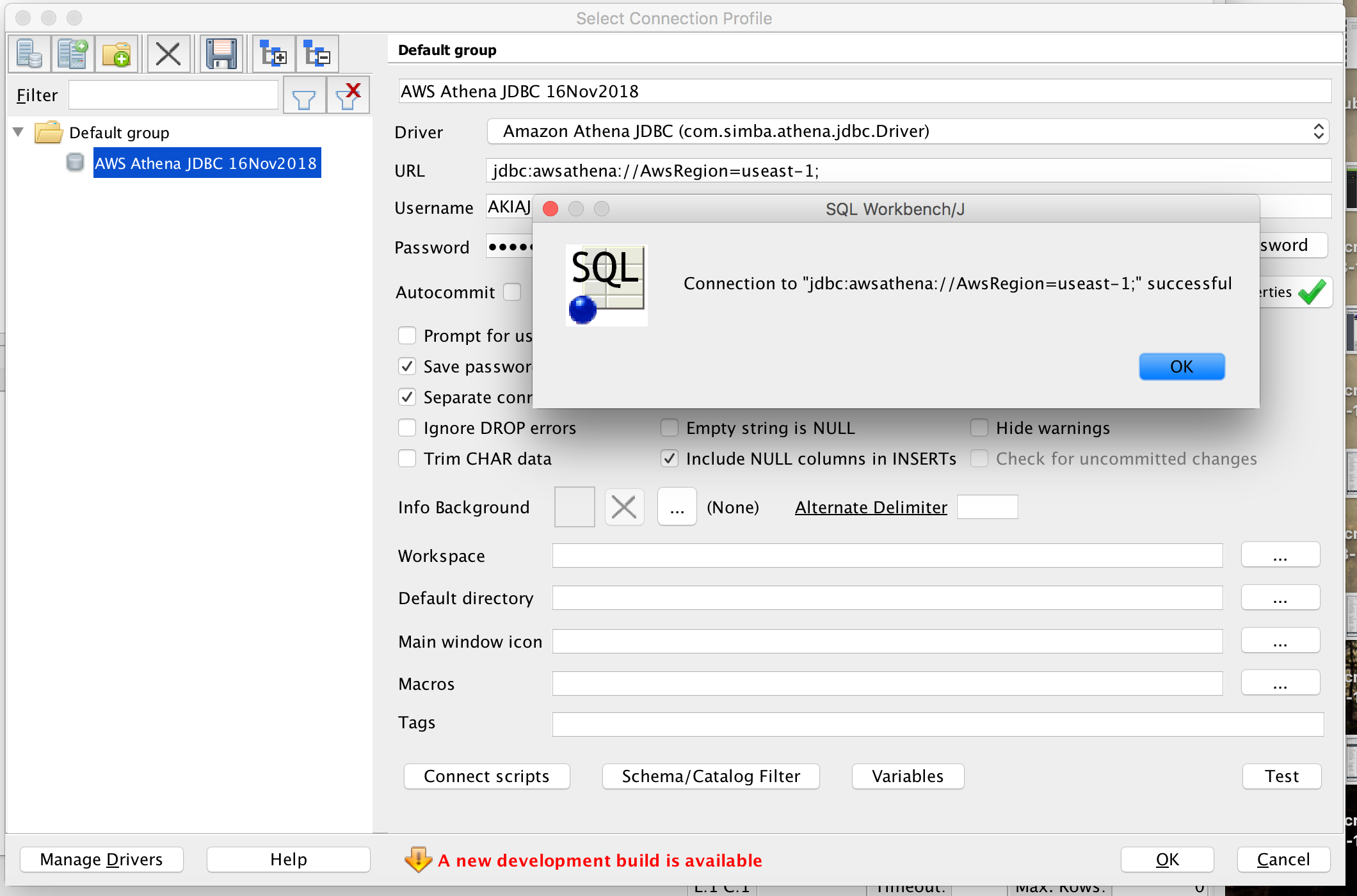
You will work with a 3rd Party SQL Client tool to connect to a data source using JDBC. The example below uses SQL Workbench, because there are detailed instructions on how to connect it to Athena service via JDBC.

### Additional Lab Prerequisites fort this Optional Lab

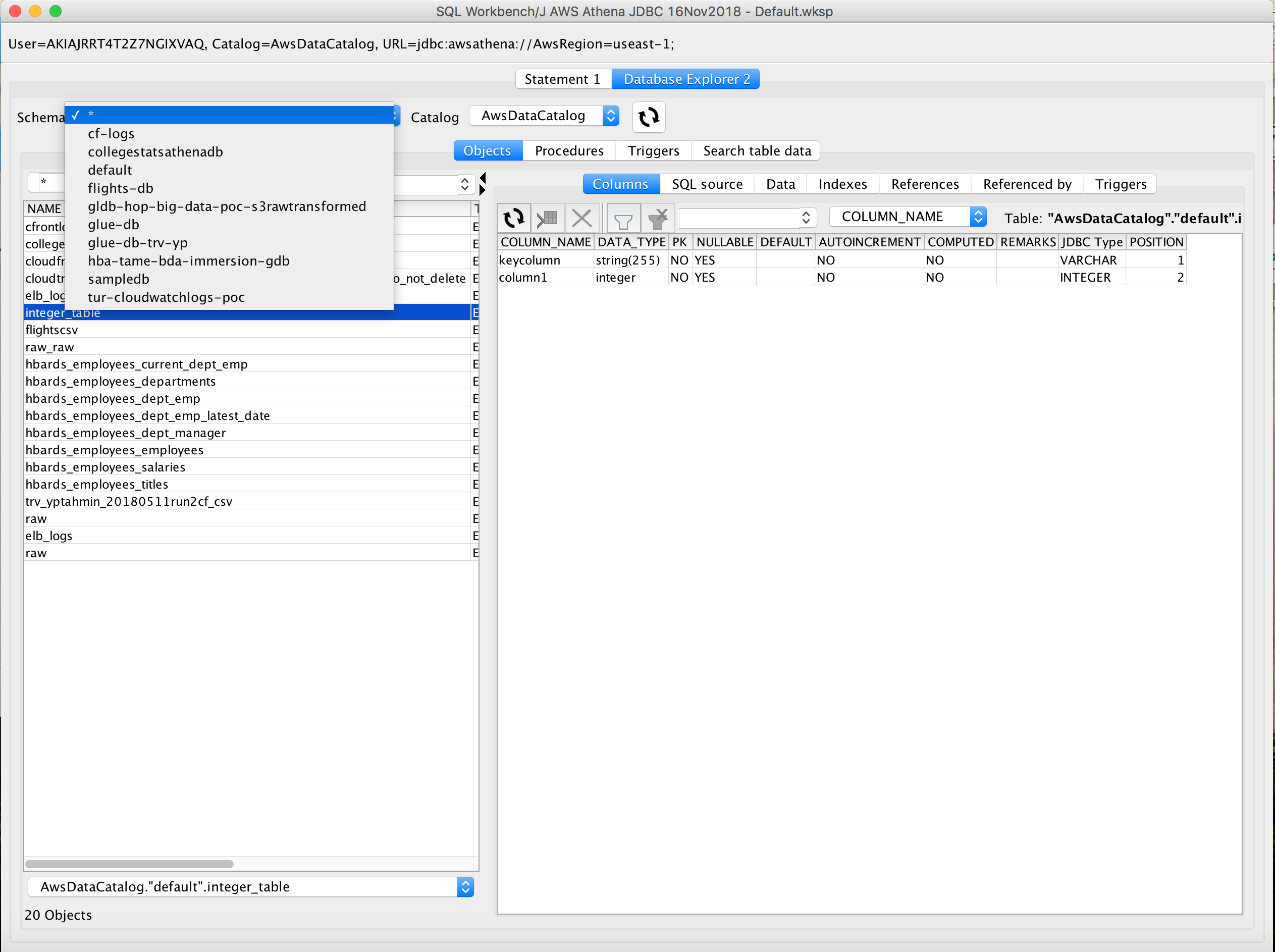
1. **IAM Access Keys for your user:** You’ll need an Access key and secret key that has Access to Athena service.
2. **SQL Workbench Tool:** 
   1. Installation: If you haven’t installed SQL Workbench prior to this lab, install it now. Follow the instructions here: <https://docs.aws.amazon.com/athena/latest/ug/connect-with-jdbc.html> or directly here: <https://s3.amazonaws.com/athena-downloads/drivers/JDBC/SimbaAthenaJDBC_2.0.6/docs/Simba+Athena+JDBC+Driver+Install+and+Configuration+Guide.pdf>
   2. Configuration: According to the guide, enter the following information:
      1. select File > Manage Drivers
         1. In the Manage Drivers dialog box, specify the following values in the fields:
         2. Name: Athena JDBC Driver
         3. Version: AthenaJDBC42.jar
         4. Classname com.simba.athena.jdbc.Driver
         5. Sample URL: jdbc:awsathena://AwsRegion=us-east-1;
         6. Username The access key provided by your AWS account.
         7. Password The secret key provided by your AWS account
         8. Click Extended Properties,
         9. set the S3OutputLocation property to an S3 bucket and folder you have read/write access to (i.e.
      2. You can now use the Simba Athena JDBC Driver in SQL Workbench to query and view data.

To explore Data with SQL Workbench

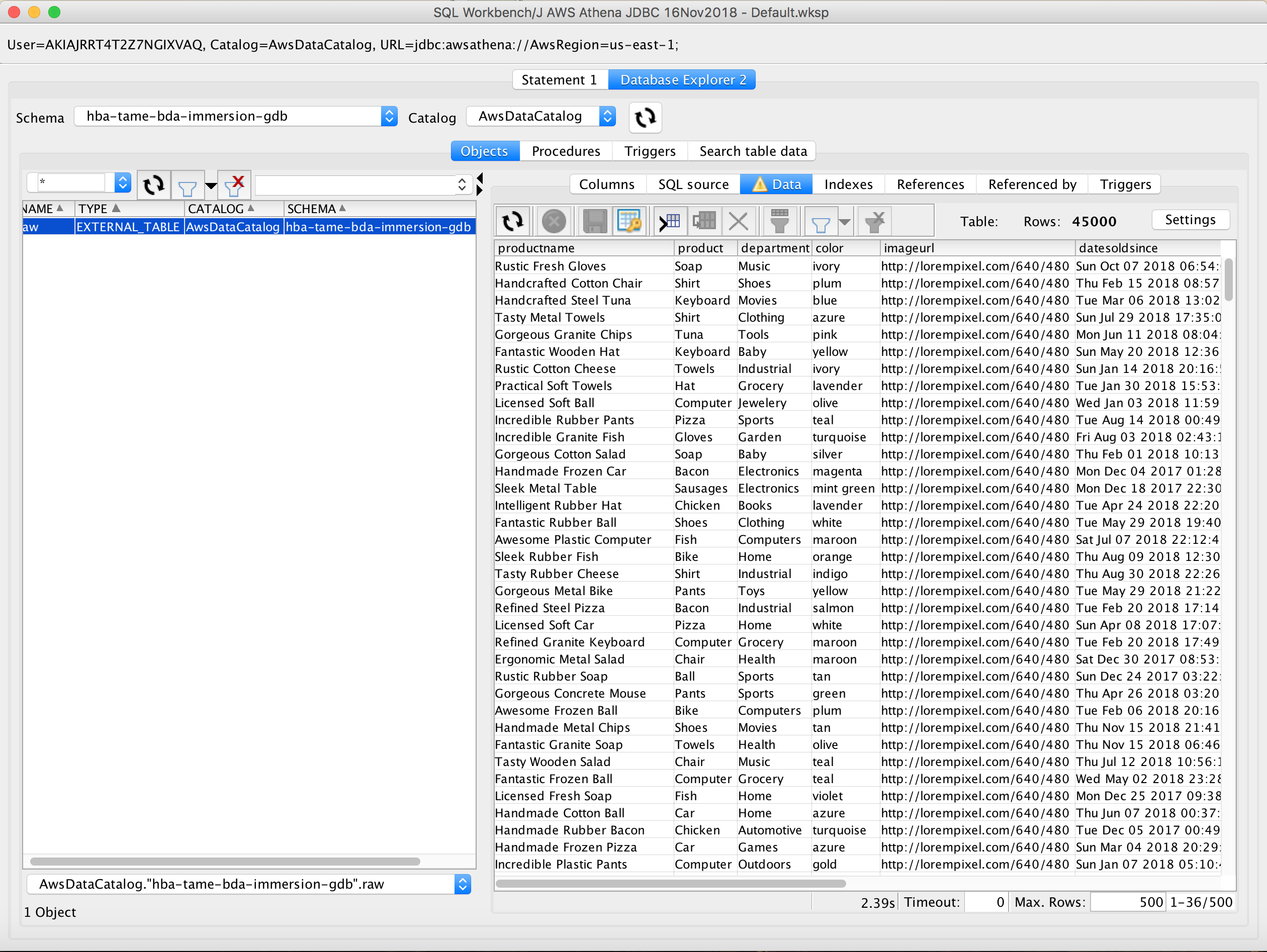
1. Open SQL Workbench and connect to the configured JDBC Endpoint



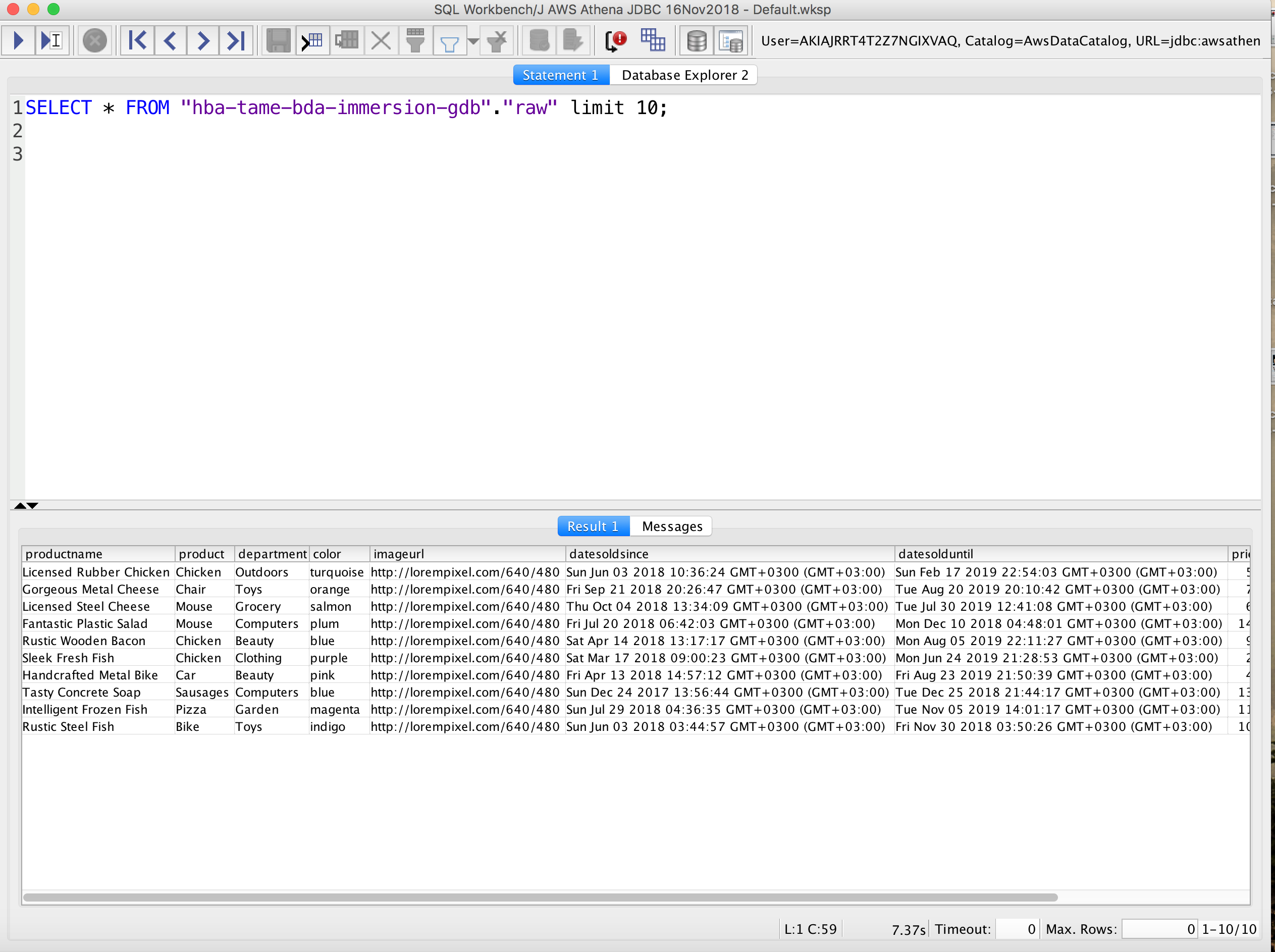
1. Select Database Explorer from the GUI
2. Select schema: <yourinitials>-tame-bda-immersion-gdb



1. Select preview data.



1. Or alternatively, enter an SQL directly.



1. Note that the SQL statement you entered is running against a set of JSON files stored in S3. You can use your own reporting and visualization tools to connect to Athena with this method.

# Summary & Next Steps

Congratulations. You have successfully created visualizations of the data ingested in Lab 1 and Lab 2 without servers.

So, what’s next? One option is use Machine Learning services to create insights from your data.

At AWS, we’ve seen a ton of progress in machine learning during the past 12 months, with customers using Amazon SageMaker – a fully-managed service which has put ML into the hands of tens of thousands of developers and data scientists – to [find fraud](https://aws.amazon.com/solutions/case-studies/Intuit/?hp=tile&story=intuit" \t "_blank), [predict pitches](https://aws.amazon.com/statcastai/?hp=tile&story=mlb" \t "_blank), and [tune engines](https://aws.amazon.com/solutions/case-studies/formula-one/?hp=tile&story=f1" \t "_blank). As of writing (November 2018), AWS have added nearly 100 new features and capabilities since we introduced SageMaker at re:Invent 2017 last year, with the vast majority based on customer feedback. Several major new announcements have been made at ReInvent 2018 for Amazon SageMaker.

# Deleting Lab Resources

Make sure you terminate the resources below to avoid bills.

* Quicksight
  + Delete your user.
* JDBC Connection
  + Delete the security key you created for JDBC connection.

# References

1. Website, AWS Data Lake, <https://aws.amazon.com/tr/big-data/datalakes-and-analytics/what-is-a-data-lake/>
2. Whitepaper, “Lambda Architecture for Batch and Stream Processing, AWS, October 2018, <https://d0.awsstatic.com/whitepapers/lambda-architecure-on-for-batch-aws.pdf>