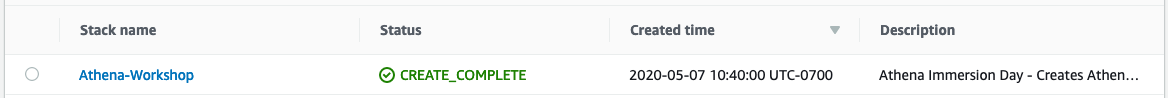
Using AWS CloudFormation Template

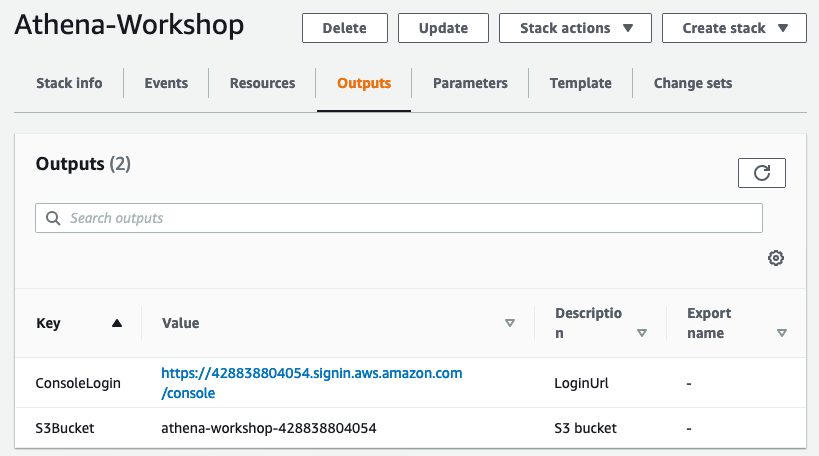
Before starting this Amazon Athena workshop, you need to create the required AWS resources. To do this, we provide AWS CloudFormation template to create a stack that contains the resources. When you create the stack, AWS creates a number of resources in your account.

[Click Here](https://console.aws.amazon.com/cloudformation/home?region=us-east-1#/stacks/new?stackName=Athena-Workshop&templateURL=https://aws-data-analytics-workshops.s3.amazonaws.com/athena-workshop/cloudformation/athena_immersion_day.yml) to launch the CloudFormation Stack.  
  
**NOTE:** This template is created for us-east-1 region (N. Virginia) and will not work in other regions. Don't change any of the default Database parameters in cloudformation.

The CloudFormation stack will roughly take 5 minutes to complete. The CloudFormation stack will create necessary IAM users, Athena Workgroups and Athena Named Queries for trying out the labs. Check the CloudFormation console and wait for the status **CREATE\_COMPLETE** as shown below:

[](https://athena-in-action.workshop.aws/images/athena_workshop_cft_complete.png)

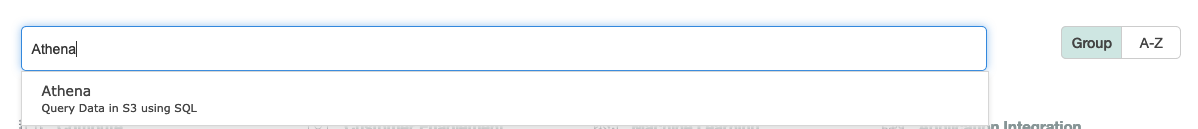
Click on the CloudFormation stack and go to outputs tab which will display the S3 bucket name created for this workshop:

[](https://athena-in-action.workshop.aws/images/athena_workshop_cft_output.png)

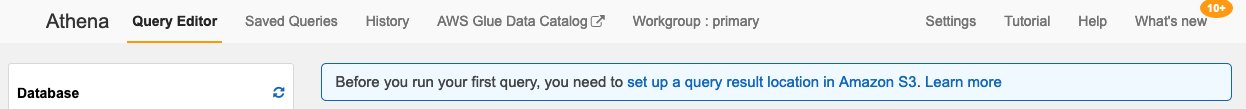
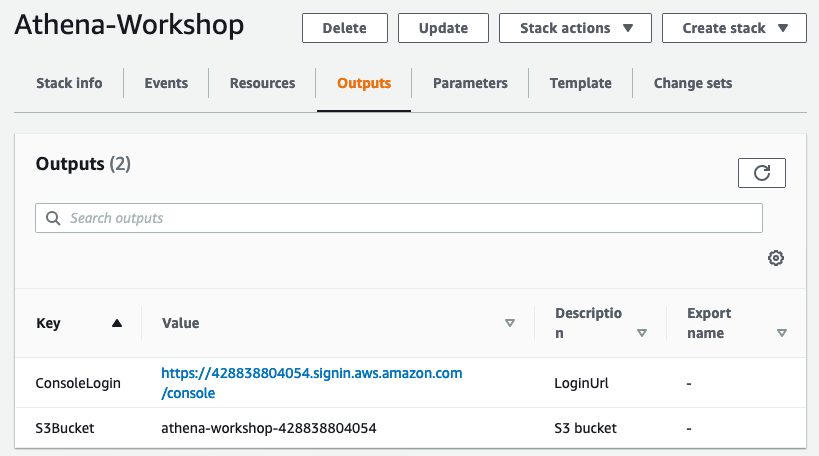
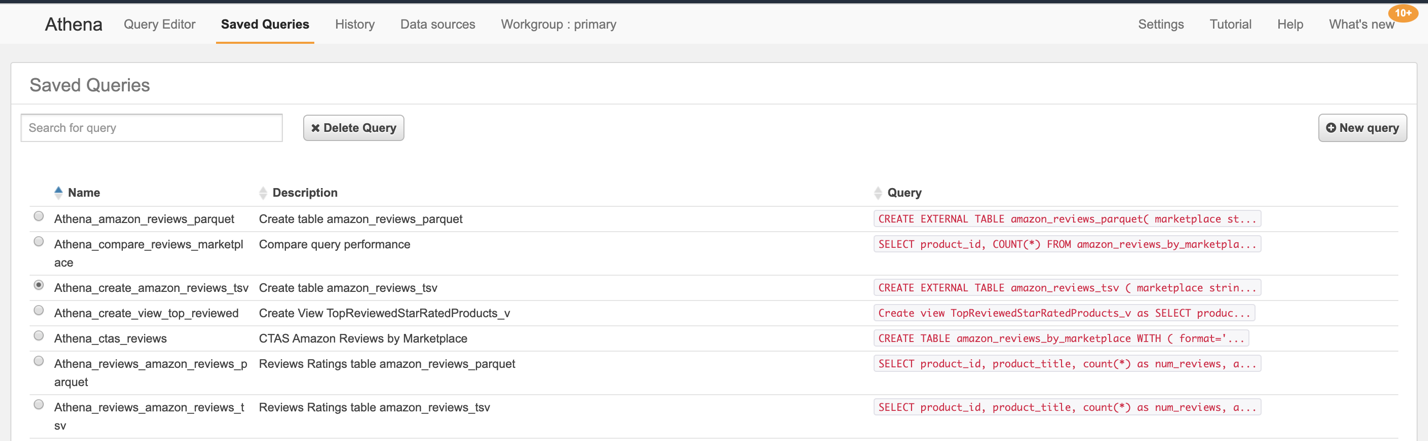
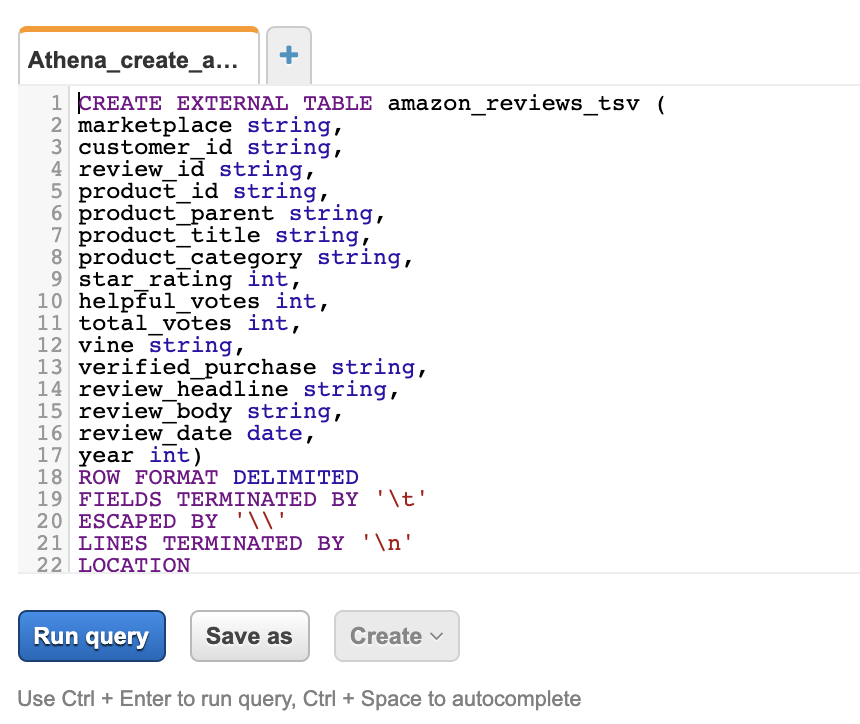
## View Datasets

To view source dataset in S3, access below URL  
  
[Amazon Product Reviews Dataset](https://console.aws.amazon.com/s3/home?region=us-east-1&bucket=amazon-reviews-pds)  
  
Notice that the tsv folder has multiple files compressed using gzip. Also notice that file size varies from 12 MB to 2.6 GB. Parquet folder has sub-folders on product category and going down one level, you would notice that files are compressed using snappy. File size is more uniform.  
  
[Flight Delay Dataset](https://console.aws.amazon.com/s3/home?region=us-east-1&bucket=athena-examples/flight/)  
  
Navigate to flight folder and check the data under csv and parquet folder.

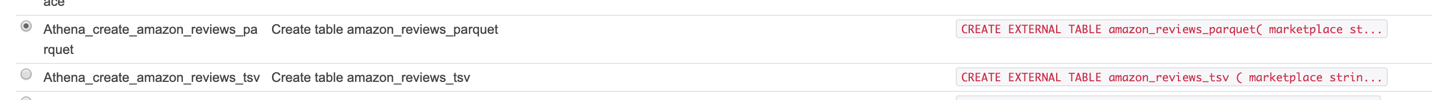
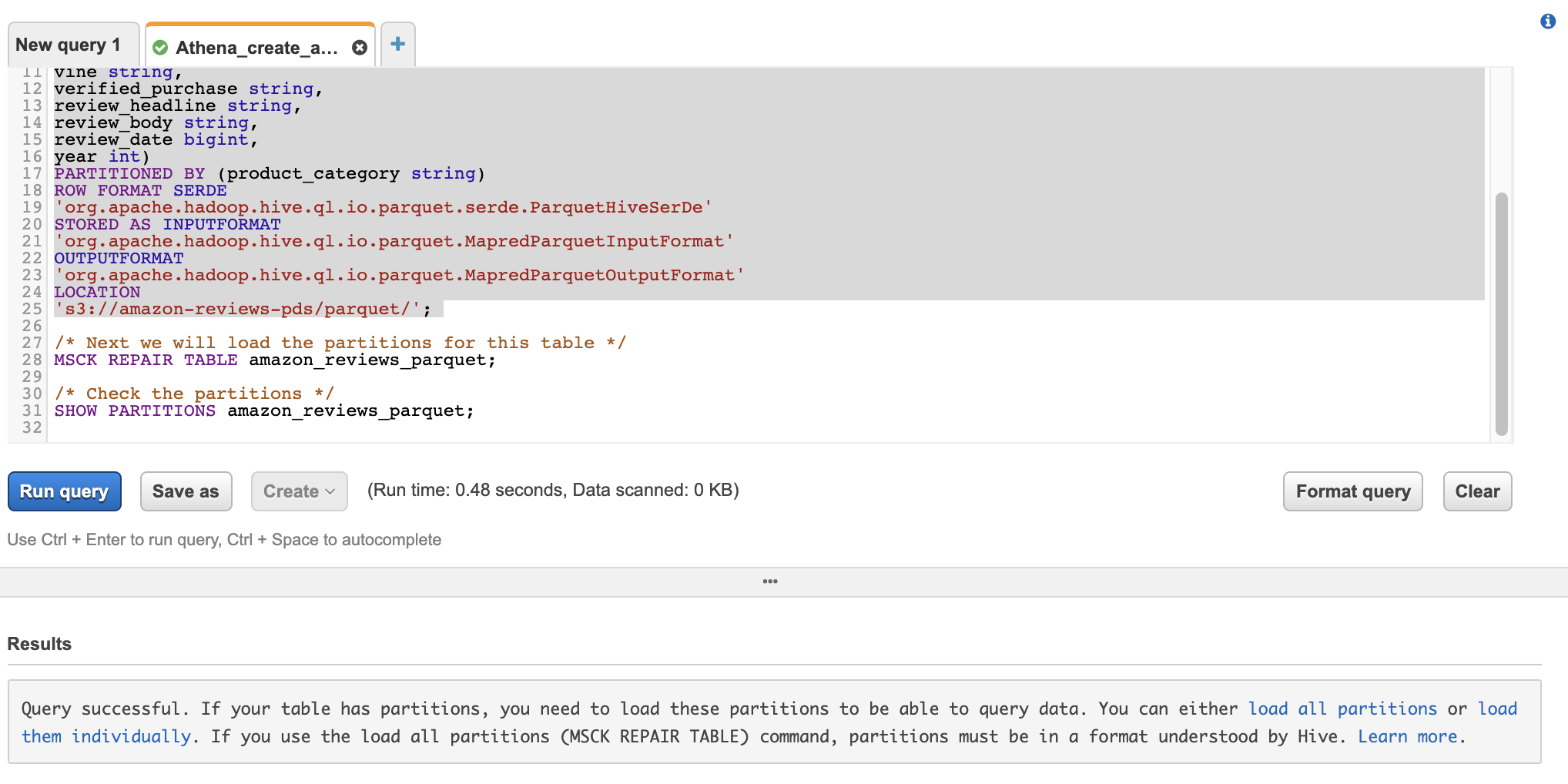
Athena Interface - Create Tables and Run Queries

From the services menu type Athena and go to the console.[](https://athena-in-action.workshop.aws/images/services_athena.png)

Once on the Athena console click, “Get Started” and then on “Set up a query result location in Amazon S3” and enter the S3 bucket name from Cloudformation output. S3 url in Athena requires a "/" at the end. Make sure to enter **"/"** at the end of the bucket location for the query results.

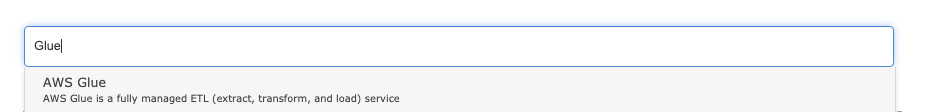
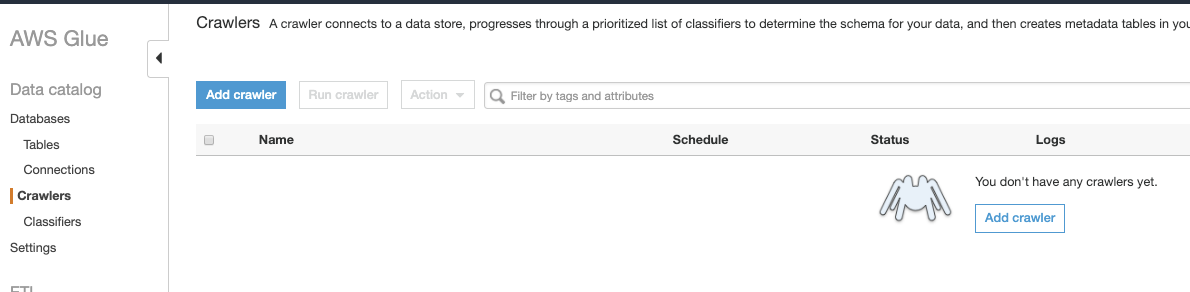
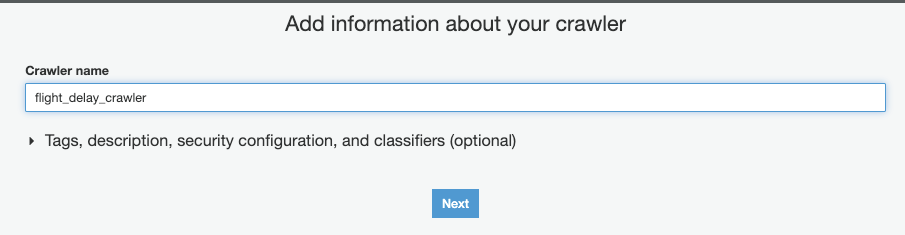
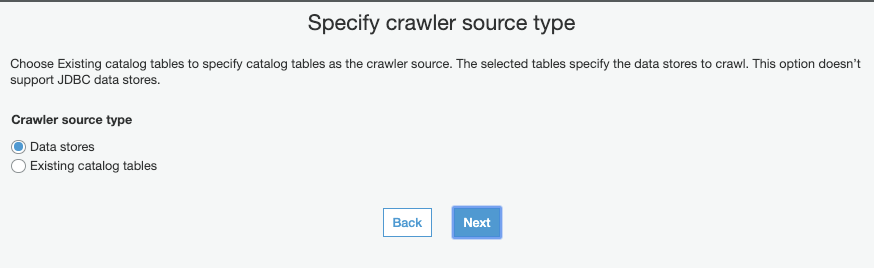
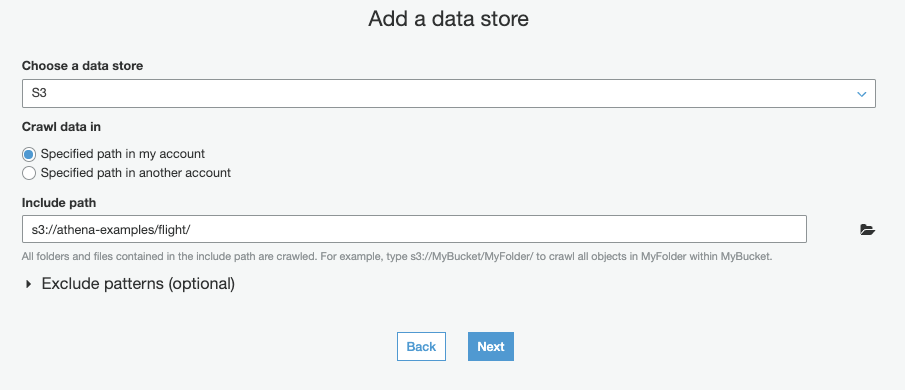
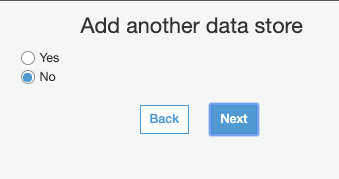
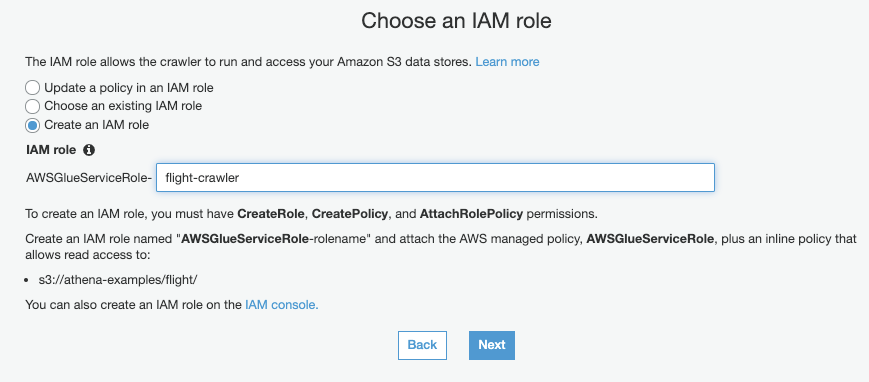
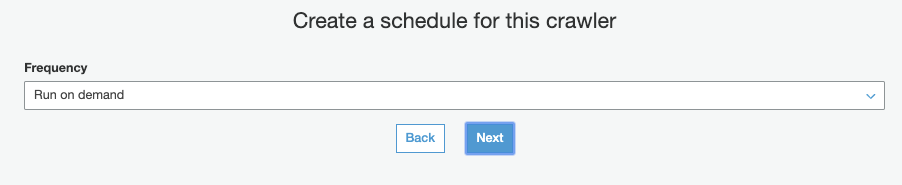
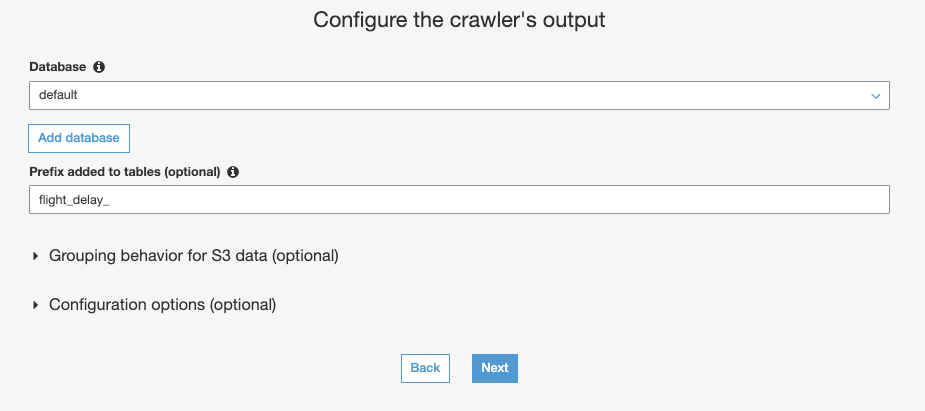
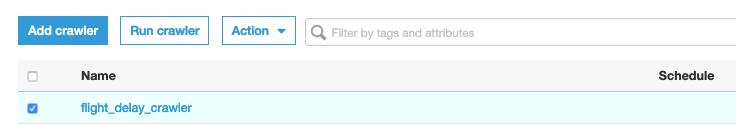
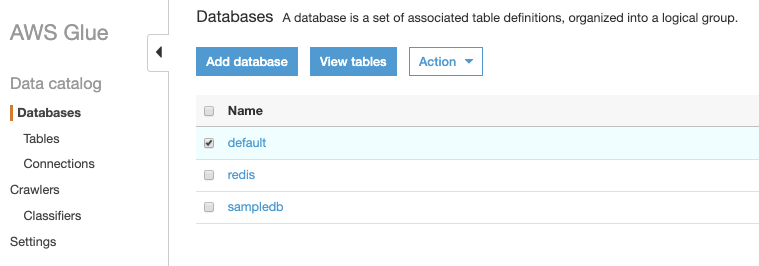
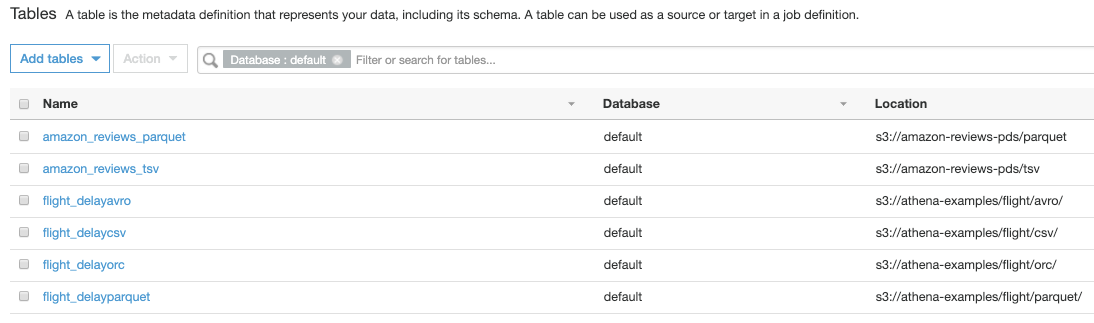
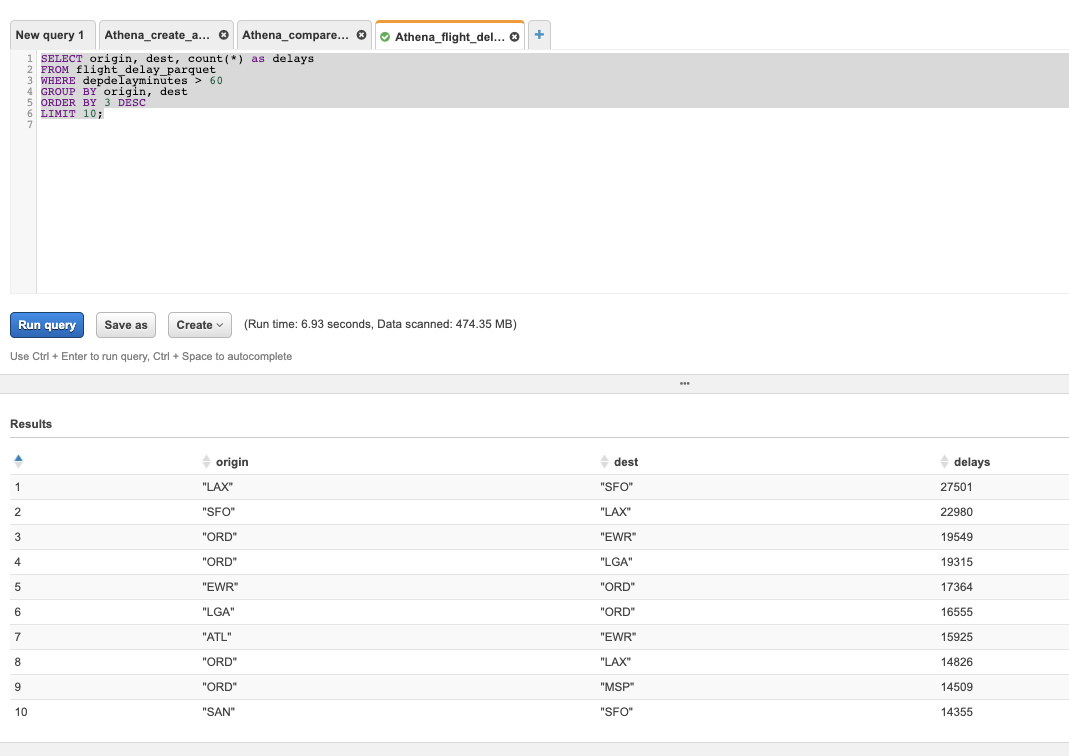
[](https://athena-in-action.workshop.aws/images/athena_save_qry_result.png)[](https://athena-in-action.workshop.aws/images/athena_workshop_cft_output.png)Click on Saved Queries and Select **Athena\_create\_amazon\_reviews\_tsv**[](https://athena-in-action.workshop.aws/images/athena_basics_saved_qry.png)[](https://athena-in-action.workshop.aws/images/athena_saved_qry1.png)Click on Run query to create the table.  
Click on Saved Queries and Select **Athena\_create\_amazon\_reviews\_parquet** and select the table create query and run the the query. Make sure to select one query at a

time and run it.

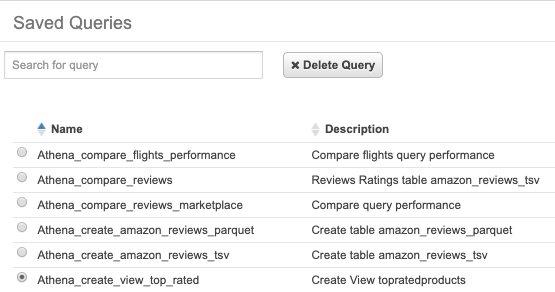
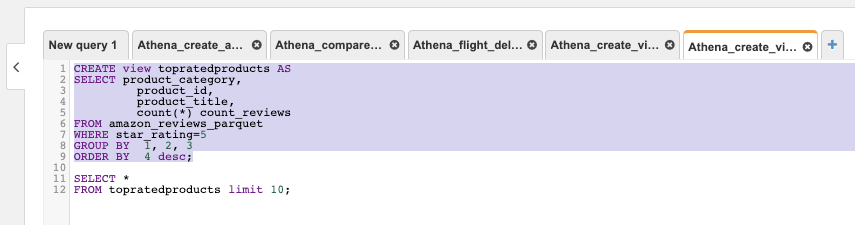
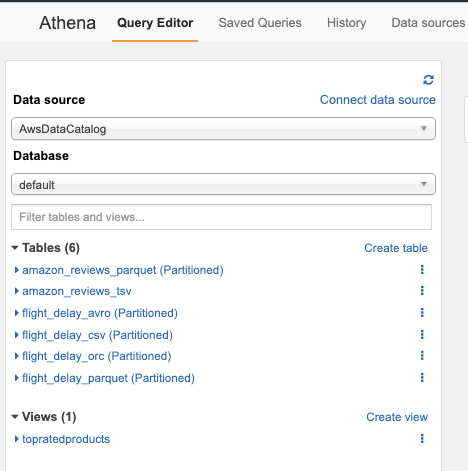
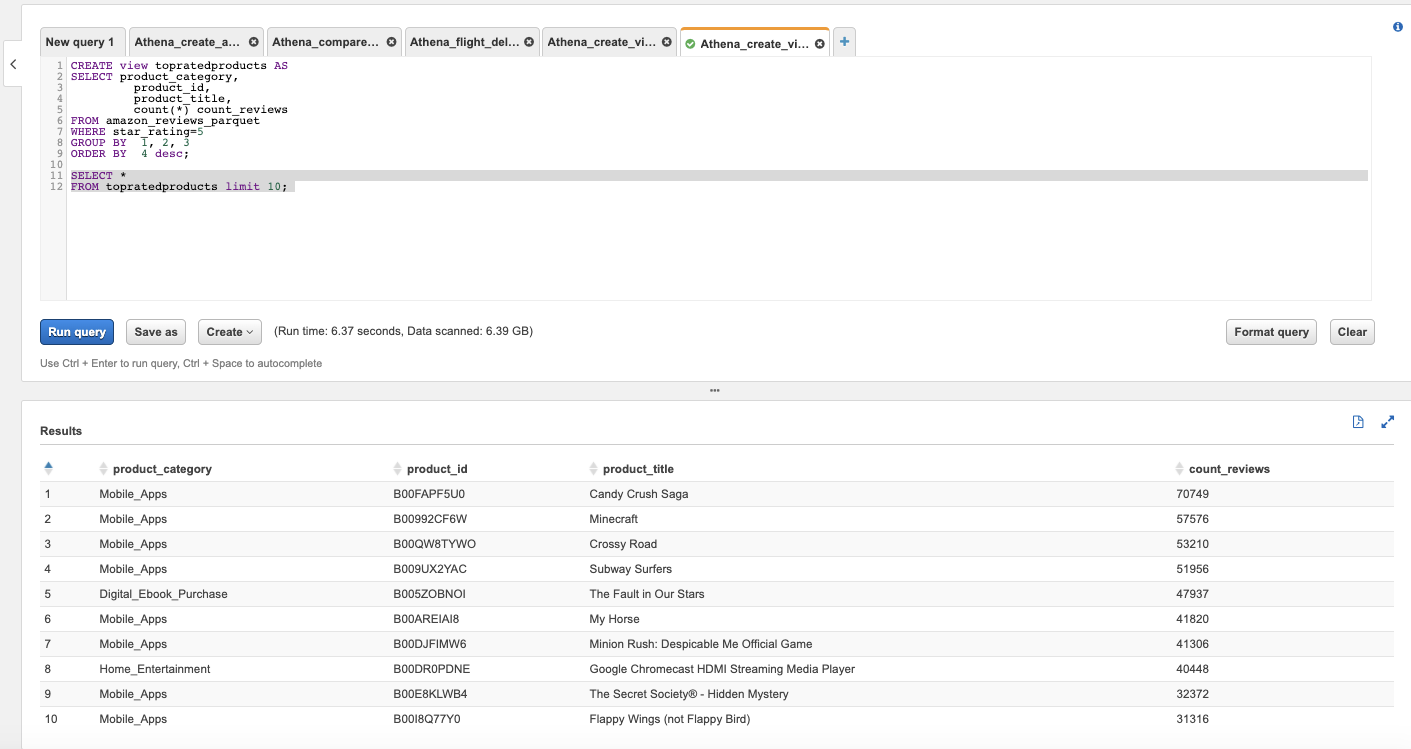
[](https://athena-in-action.workshop.aws/images/athena_saved_qry2.png)[](https://athena-in-action.workshop.aws/images/athena_saved_qry3.png)Once the query completes it will display a message to add partitions.  
Run the next query to add partitions.  
Next query will display the partitions.  
  
Now that we have the tables created we will run some queries to test the performance between the tables:  
Open Saved Queries and select **Athena\_compare\_reviews**  
Run the queries in order and see the difference in amount of data scanned and time taken:

|  |  |  |  |
| --- | --- | --- | --- |
| **Query** | **Table** | **Time Taken** | **Data Scanned** |
| Top Ten Products By Avg Reviews | amazon\_reviews\_tsv | 2 min 1 sec | 32.23 GB |
| Top Ten Products By Avg Reviews | amazon\_reviews\_parquet | 7.88 sec | 6.34 GB |
| Top Ten Products By Avg Reviews  in Mobile Apps category | amazon\_reviews\_tsv | 1 minute 32 seconds | 32.21 GB |
| Top Ten Products By Avg Reviews  in Mobile Apps category | amazon\_reviews\_parquet | 3.94 seconds | 57.01 MB |

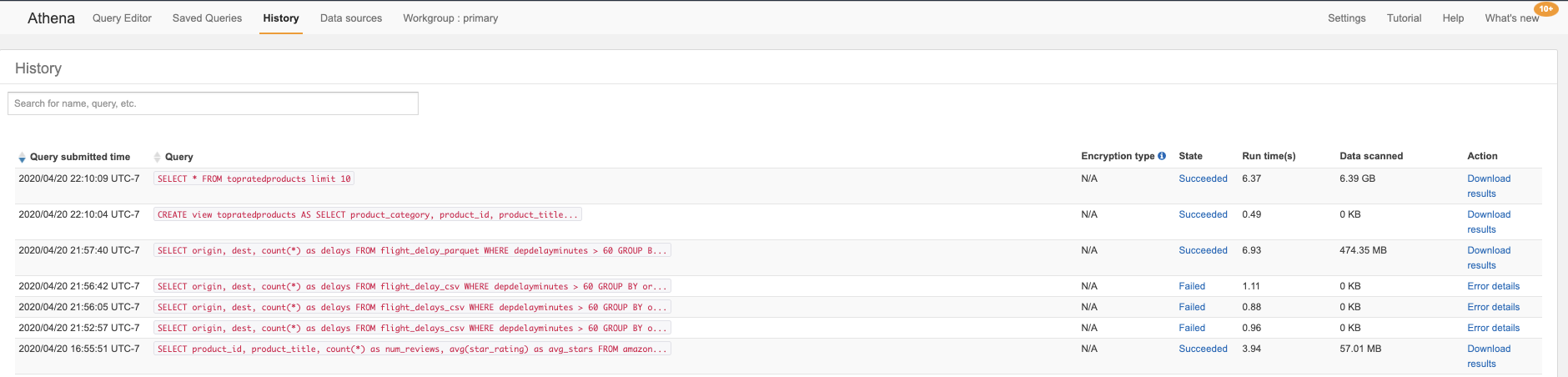
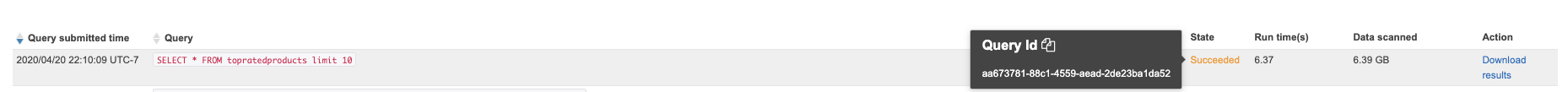
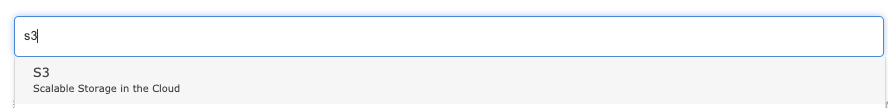
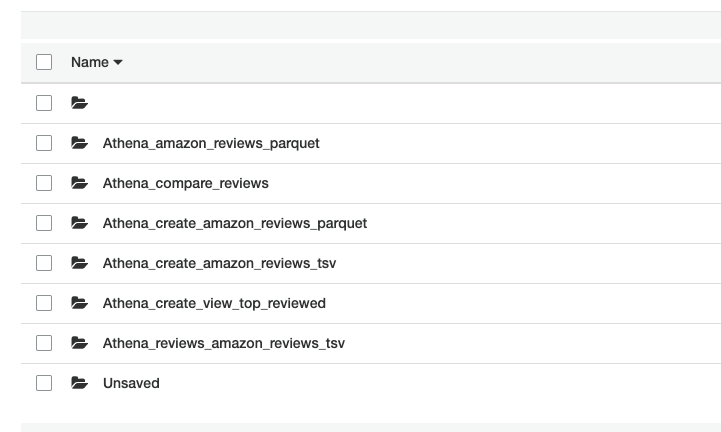
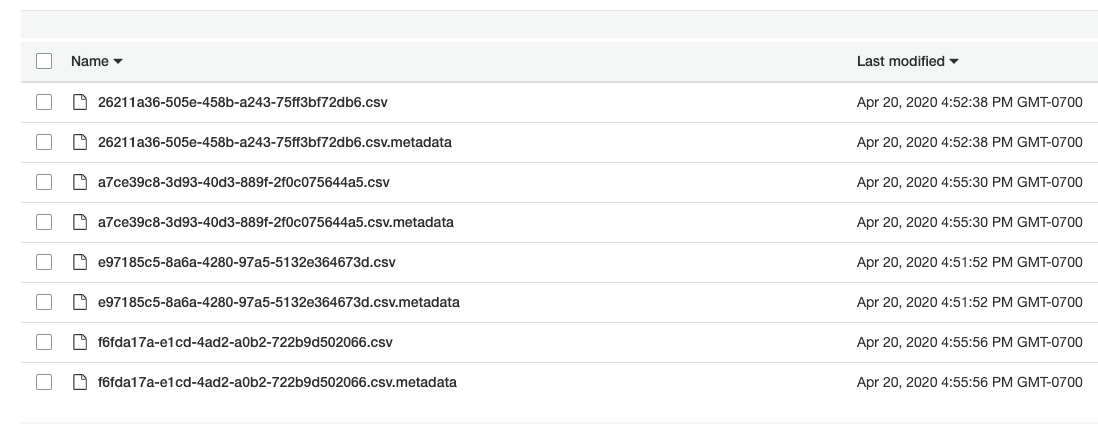
Create Tables with Glue

In this lab we will use Glue Crawlers to crawl the dataset for Flight Delay and then use the tables created by Glue Crawlers to query using Athena. Goto Services and type Glue. Click on AWS Glue.[](https://athena-in-action.workshop.aws/images/services_glue.png)On the Glue console click on Crawlers and then Add Crawler[](https://athena-in-action.workshop.aws/images/glue_1.png)[](https://athena-in-action.workshop.aws/images/glue_2.png)[](https://athena-in-action.workshop.aws/images/glue_3.png)Enter Path: **s3://athena-examples/flight/**[](https://athena-in-action.workshop.aws/images/glue_4.png)[](https://athena-in-action.workshop.aws/images/glue_5.png)[](https://athena-in-action.workshop.aws/images/glue_6.png)[](https://athena-in-action.workshop.aws/images/glue_7.png)database: default  
Prefix: **flight\_delay\_**[](https://athena-in-action.workshop.aws/images/glue_8.png)Click on Next and then Finish. Once the crawler is created select the crawler and choose **Run crawler**. Click on that and the crawler will start running.[](https://athena-in-action.workshop.aws/images/glue_9.png)Once the crawler completes crawling you should see 4 tables added.[](https://athena-in-action.workshop.aws/images/glue_10.png)From the Glue Console Click on Databases → Default → “Tables in default”[](https://athena-in-action.workshop.aws/images/glue_11.png)You should now see 4 more tables added to the database.[](https://athena-in-action.workshop.aws/images/glue_12.png)We will now go back to Athena Console and then click on Saved Queries and select **Athena\_flight\_delay\_60**[](https://athena-in-action.workshop.aws/images/glue_13.png)

Athena Create View

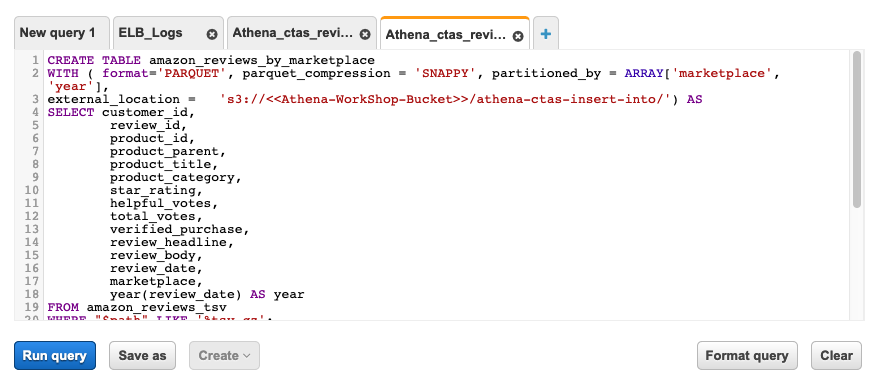
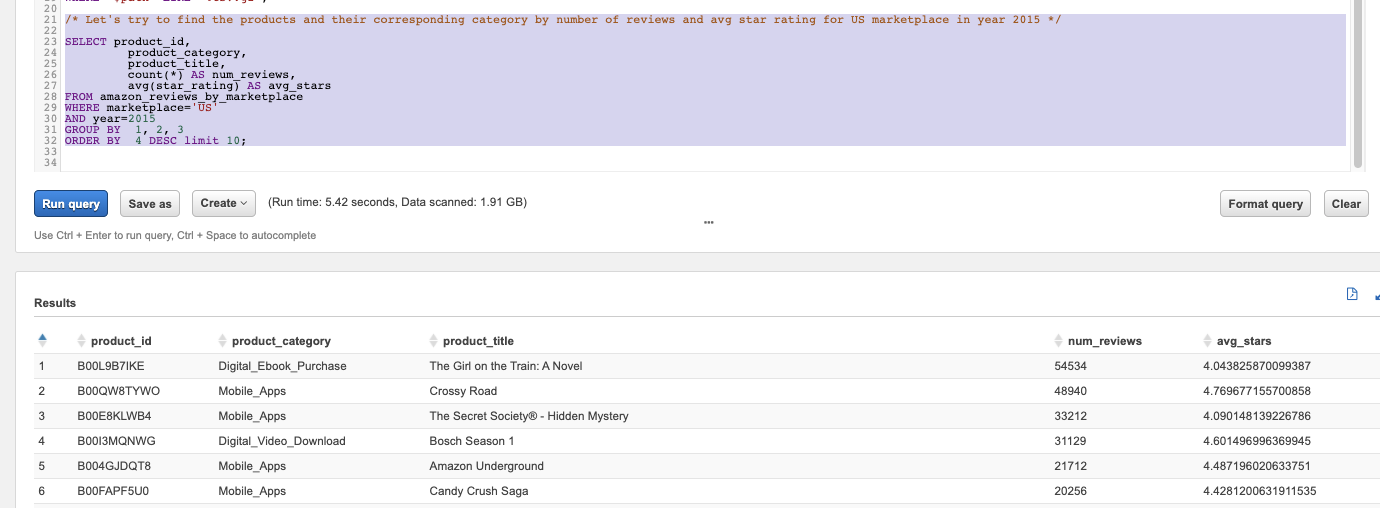
A view in Amazon Athena is a logical, not a physical table. The query that defines a view runs each time the view is referenced in a query. You can create a view from a SELECT query and then reference this view in future queries. For more information, see [CREATE VIEW](https://docs.aws.amazon.com/athena/latest/ug/create-view.html).  
  
Click on Saved Queries under Athena Console and the open **Athena\_create\_view\_top\_rated** query.[](https://athena-in-action.workshop.aws/images/athena_view_1.png)[](https://athena-in-action.workshop.aws/images/athena_view_2.png)  
Run the first query by highlighting which will create the view called **topratedproducts**. Once the view is created it will appear under Athena Views.[](https://athena-in-action.workshop.aws/images/athena_view_3.png)Select the second query and see the top 10 products by rating.[](https://athena-in-action.workshop.aws/images/athena_view_4.png)

## Athena Query History and Query Results

To view the history of queries click on **History**. You will be able to see all the queries submitted in this workgroup, their State, Run Time and Data scanned. You can also click on Download results to download as csv data.[](https://athena-in-action.workshop.aws/images/athena_qry_res_1.png)To see the Query id click on the state - Succeeded or Failed on one of the queries.[](https://athena-in-action.workshop.aws/images/athena_qry_res_5.png)Next Open the S3 console by typing S3 in the Services[](https://athena-in-action.workshop.aws/images/athena_qry_res_2.png)Once in the S3 console browse to the bucket starting with **athena-workshop-<account-id>** You should see something like this. There is a prefix for each Saved Query and Unsaved folder which stores the results of unsaved queries executed from Athena console. Browse one the the prefix **Athena\_compare\_reviews** and you would see the results stored with a query id.  
  
**Note:** If you don't see Unsaved folder run one of the previous queries in a new Query tab in Athena and check the S3 bucket again.[](https://athena-in-action.workshop.aws/images/athena_qry_res_3.png)[](https://athena-in-action.workshop.aws/images/athena_qry_res_4.png)

## ETL using Athena CTAS

Most of the times the raw data coming into the data lake or tables is in csv or text format. These formats are not optimal for querying with Athena and other engines. It is advisable to convert the data into columnar formats like parquet. In this lab we will using Athena’s Create Table As Select ( CTAS ) query to transform a table from tsv format to parquet and store with compression and partitioning.  
Open Athena → Saved Queries and click on **Athena\_ctas\_reviews**  
**Note:** Change the name of **<<Athena-WorkShop-Bucket>>** with the bucket name from CloudFormation outputs.[](https://athena-in-action.workshop.aws/images/athena_ctas_1.png)

It will take few minutes for this query to run. Once the query is completed, run the next query to see the results.[](https://athena-in-action.workshop.aws/images/athena_ctas_2.png)[](https://athena-in-action.workshop.aws/images/athena_ctas_3.png)