DATA MANAGEMENT AND BIG DATA  
  
  
ALY6110, WINTER 2021  
MODULE 5 ASSIGNMENT

NAME: SHIVANI ADSAR

CLASS NUMBER & CLASS NAME: ALY6110 & DATA MANAGEMENT AND BIG DATA

CRN NUMBER: 22824

ASSIGNMENT NAME: WEEK 5 ASSIGNMENT 1

CONTACT: adsar.s@northeastern.edu

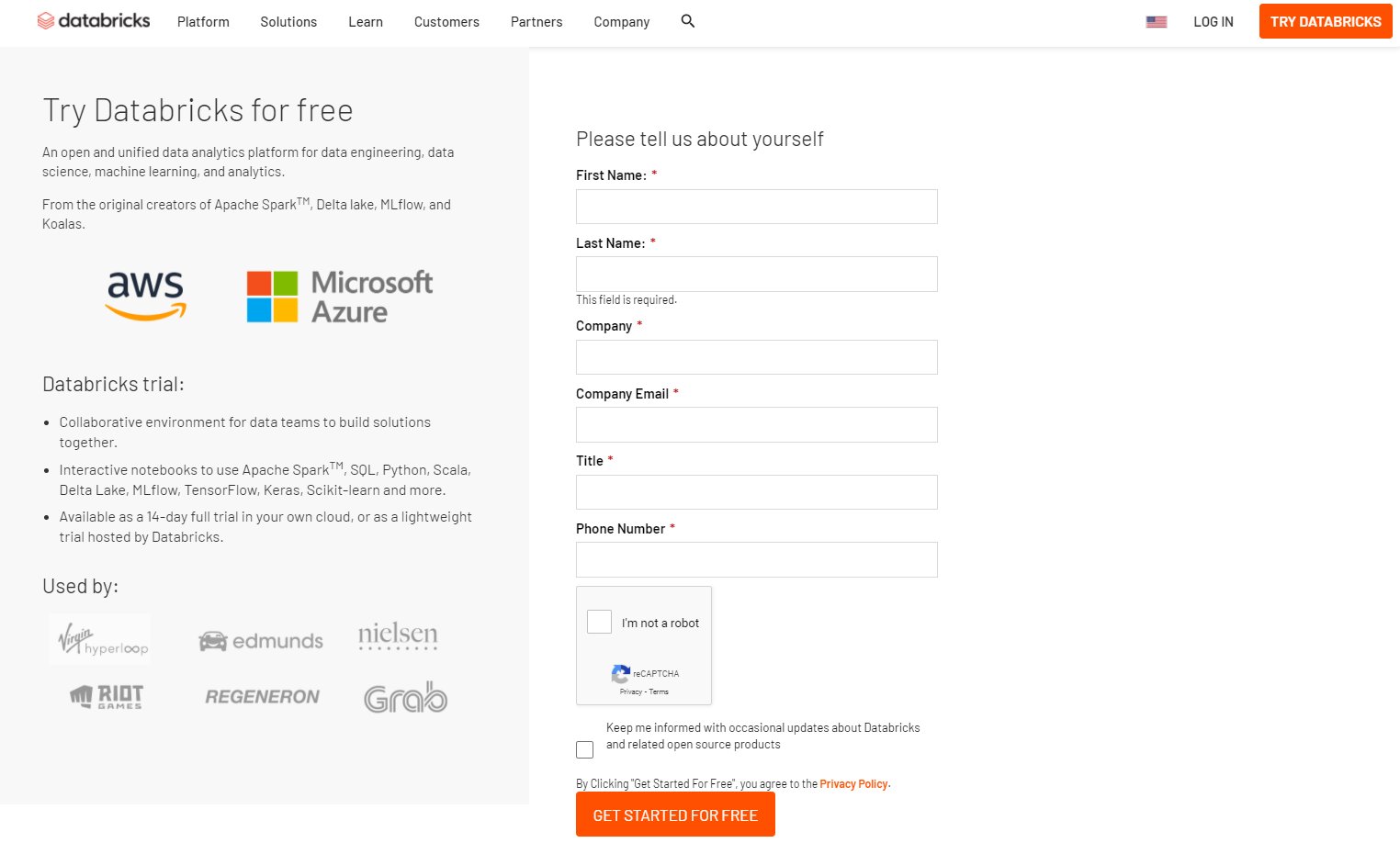
**Summary**

The assignment aims at understanding Microsoft Azure, and insights about the cloud storage and networking services. We were introduced to Azure tools for managing some of the services like controlling virtual machines, web applications and VPNs using PowerShell. This has helped in gaining tremendous knowledge about the cloud services.

**Content**

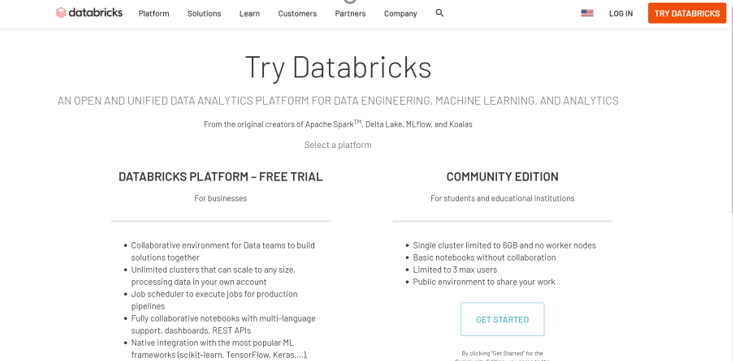
In this assignment, we have selected the second option of performing data analysis using Databricks.

We registered with Databricks using the <https://databricks.com/try-databricks> link.

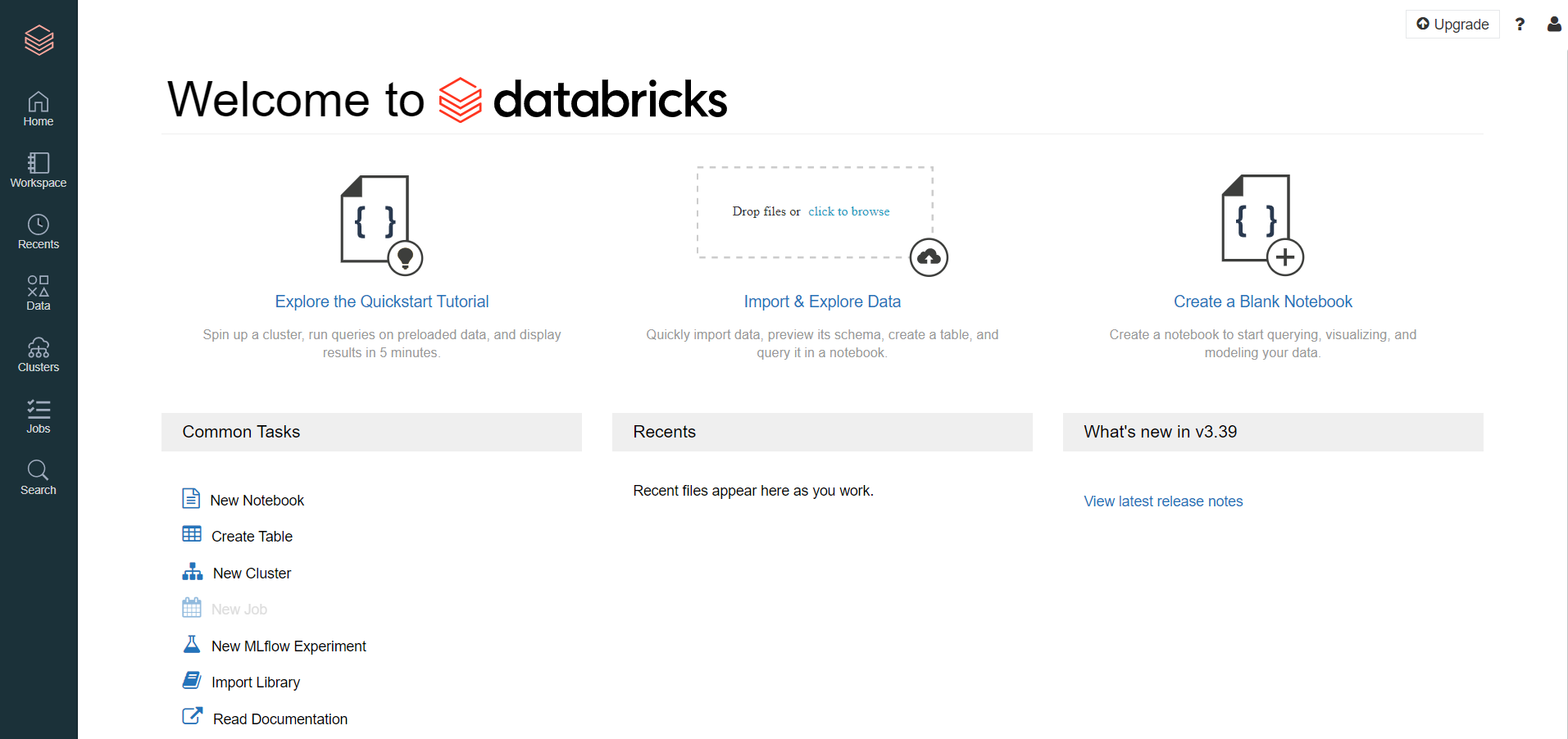


Databricks provides a free Community edition for students and educational institutions. We have used this Community Edition to work on Databricks.

We start by clicking on the “Get Started” button as shown next to the Community Edition.

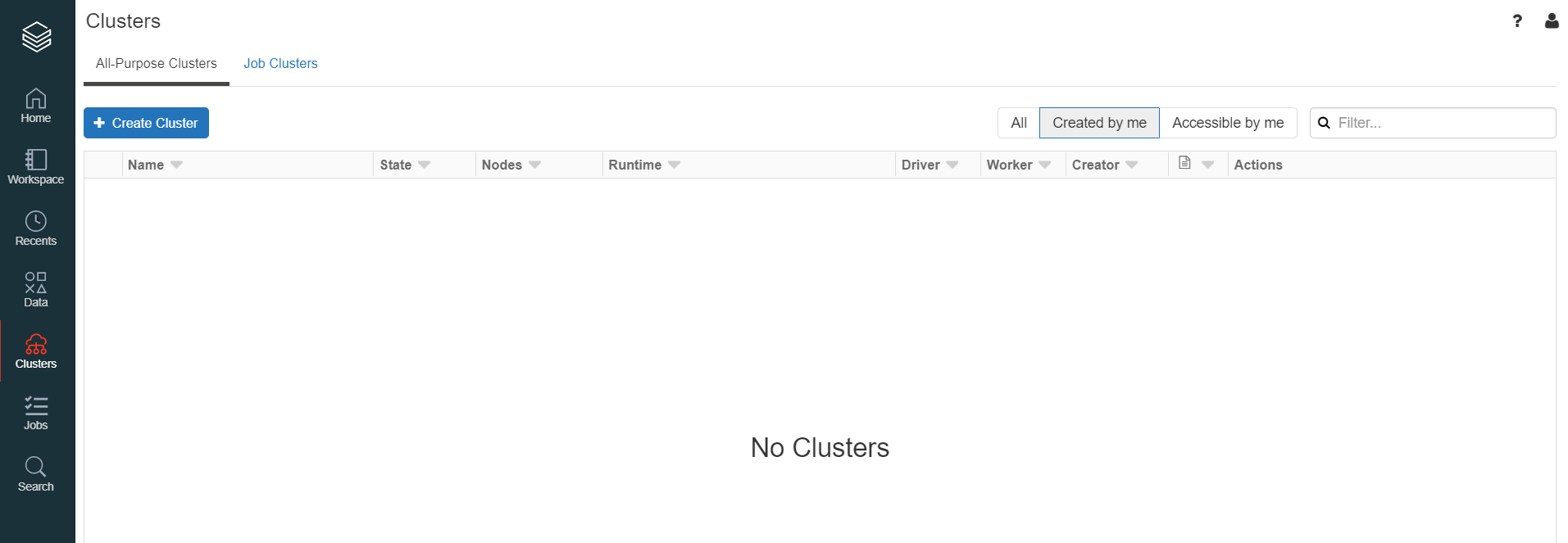


After we logged into Databricks, we have used the QuickStart Tutorial for easier navigation.

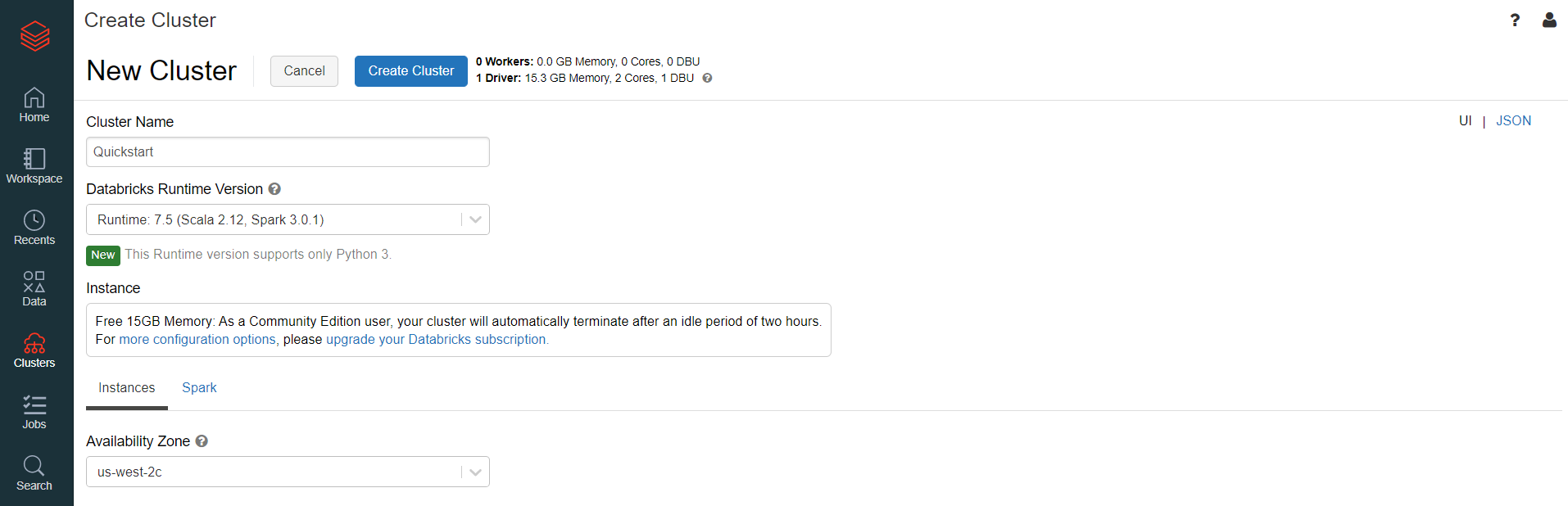


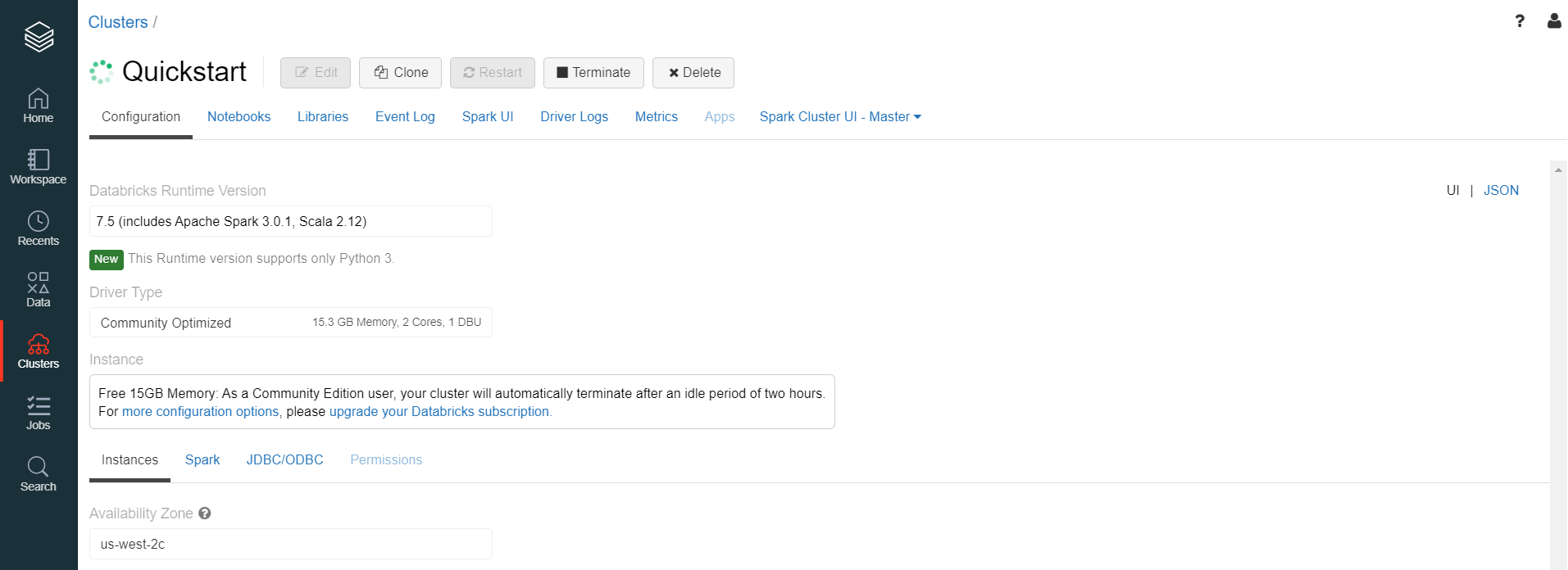
Now, we open the cluster in a new window by right clicking the “Clusters” icon on the left panel and then opening it into a new tab.

The new tab will open up a Clusters page, as seen below:

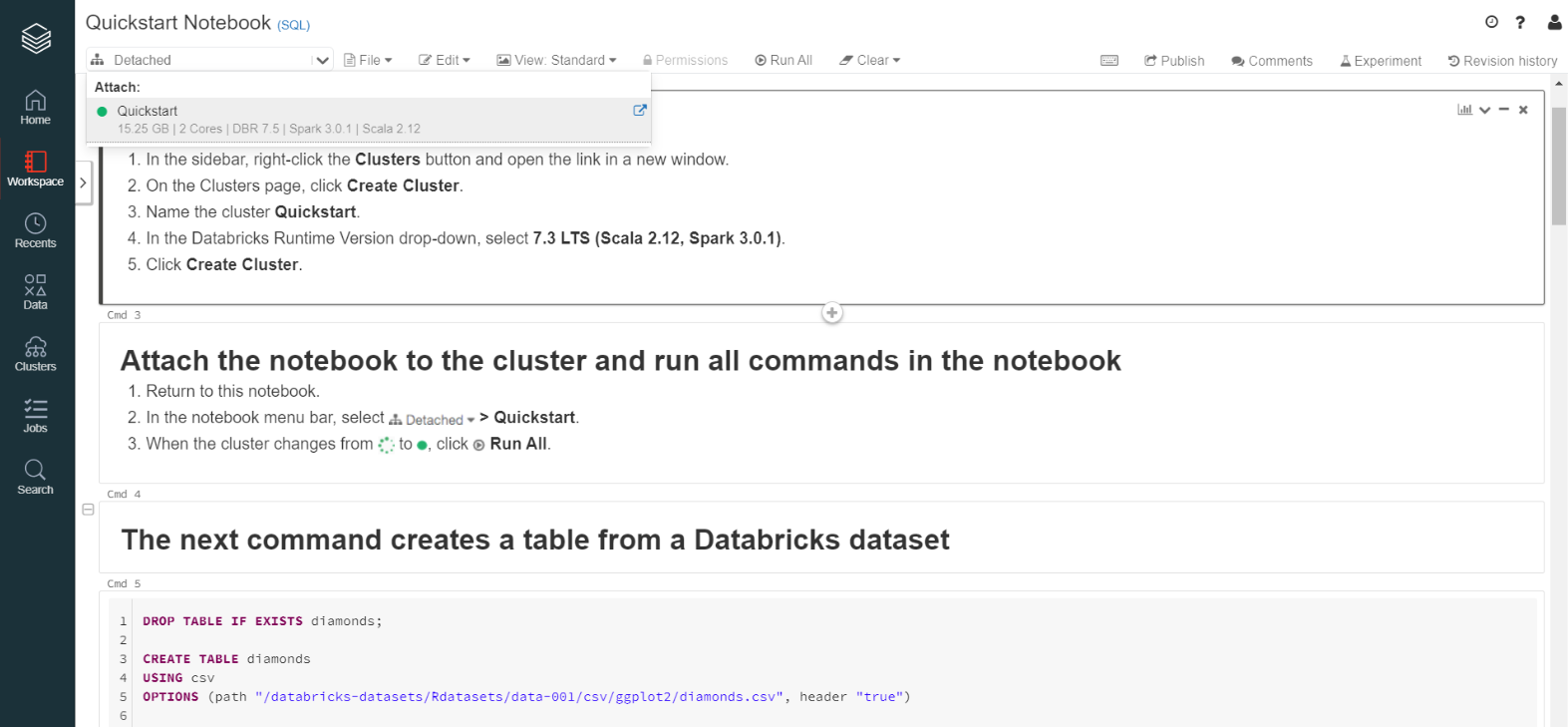


We will create a new cluster by clicking on “Create Cluster”, then, add ‘Quickstart’ as the cluster name and select the Databricks Runtime Version as Runtime: 7.5 (Scala 2.12, Spark 3.0.1)

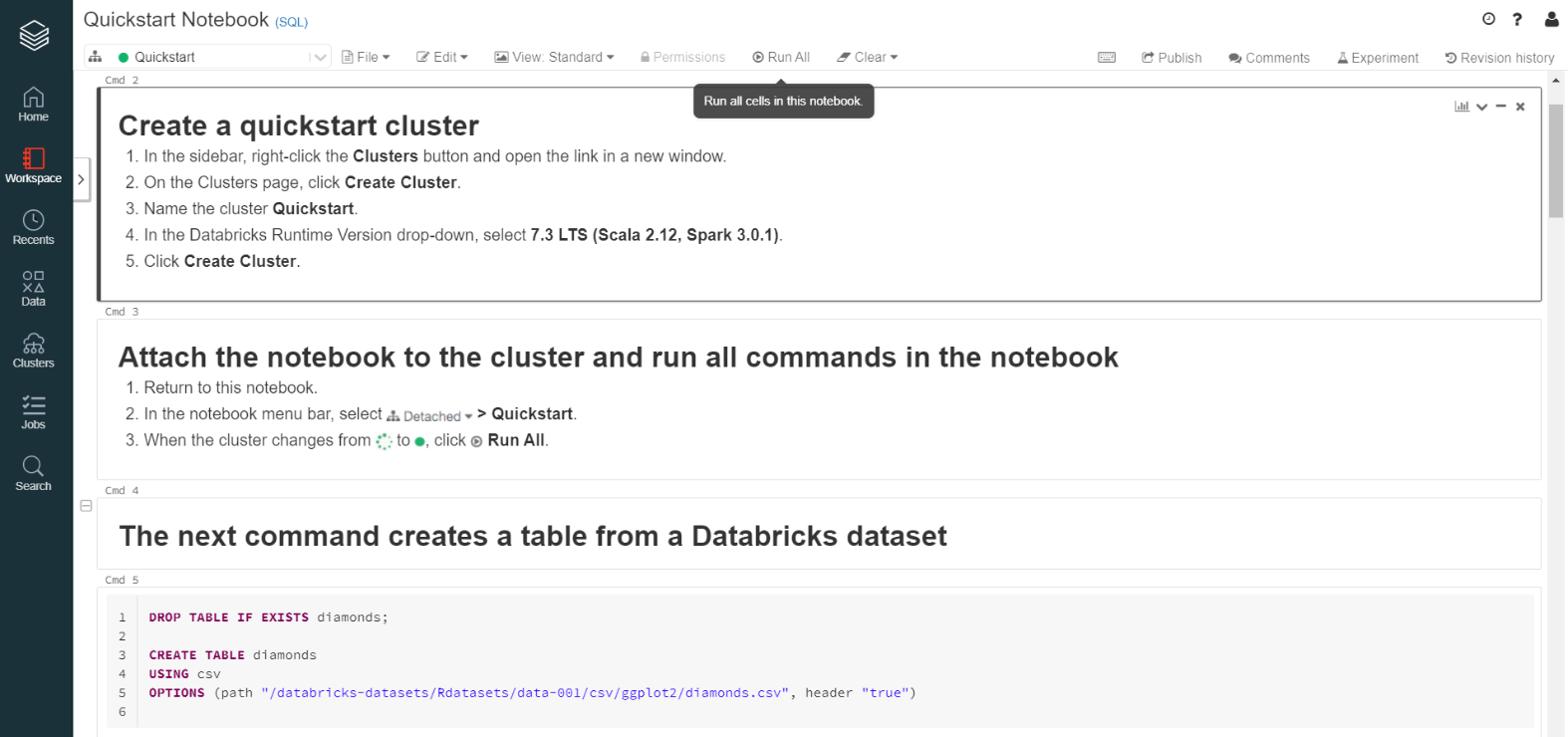


When we click on “Create Cluster”, we see the transitioning page as below:  


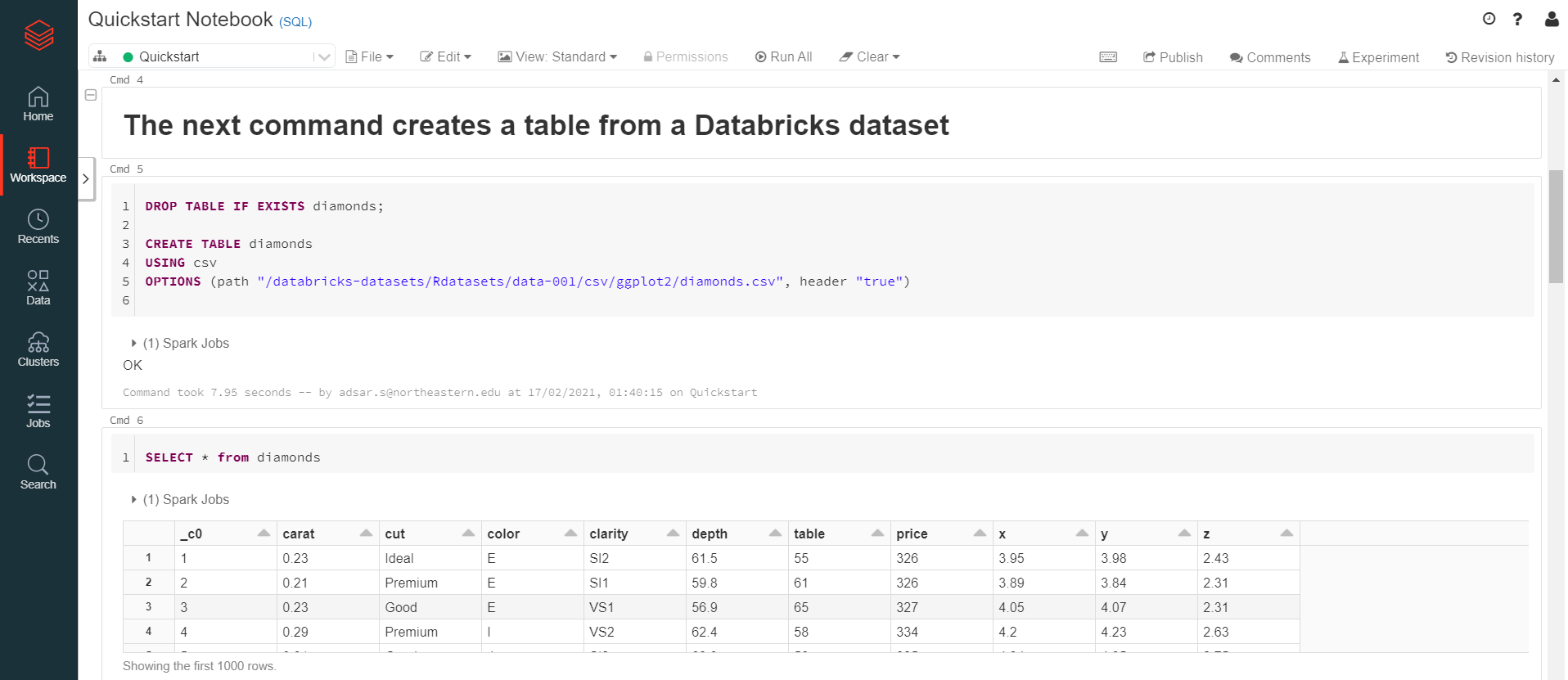
Further, we will need to associate the cluster with notebook, so we return back to the tutorial and click on the Workspace icon, followed by Detached icon and Quickstart Notebook.



When we see the cluster change from the green search symbol to a green dot, we will click on “Run All”.



After the execution completes, we observe that, a table called “Diamonds” is created using queries as below:



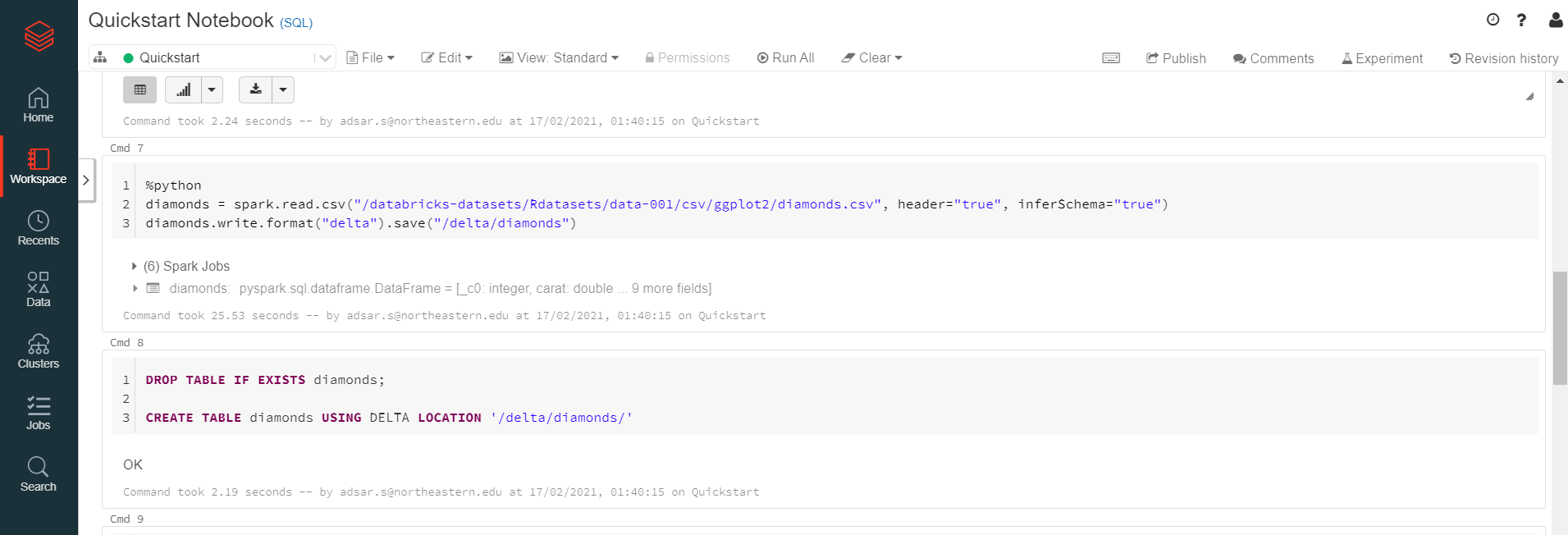
In order to select all the records present in the “Diamonds” table, the SELECT \* query has been used.

Further, we import the csv for the diamonds dataset into spark and write the output of the data.

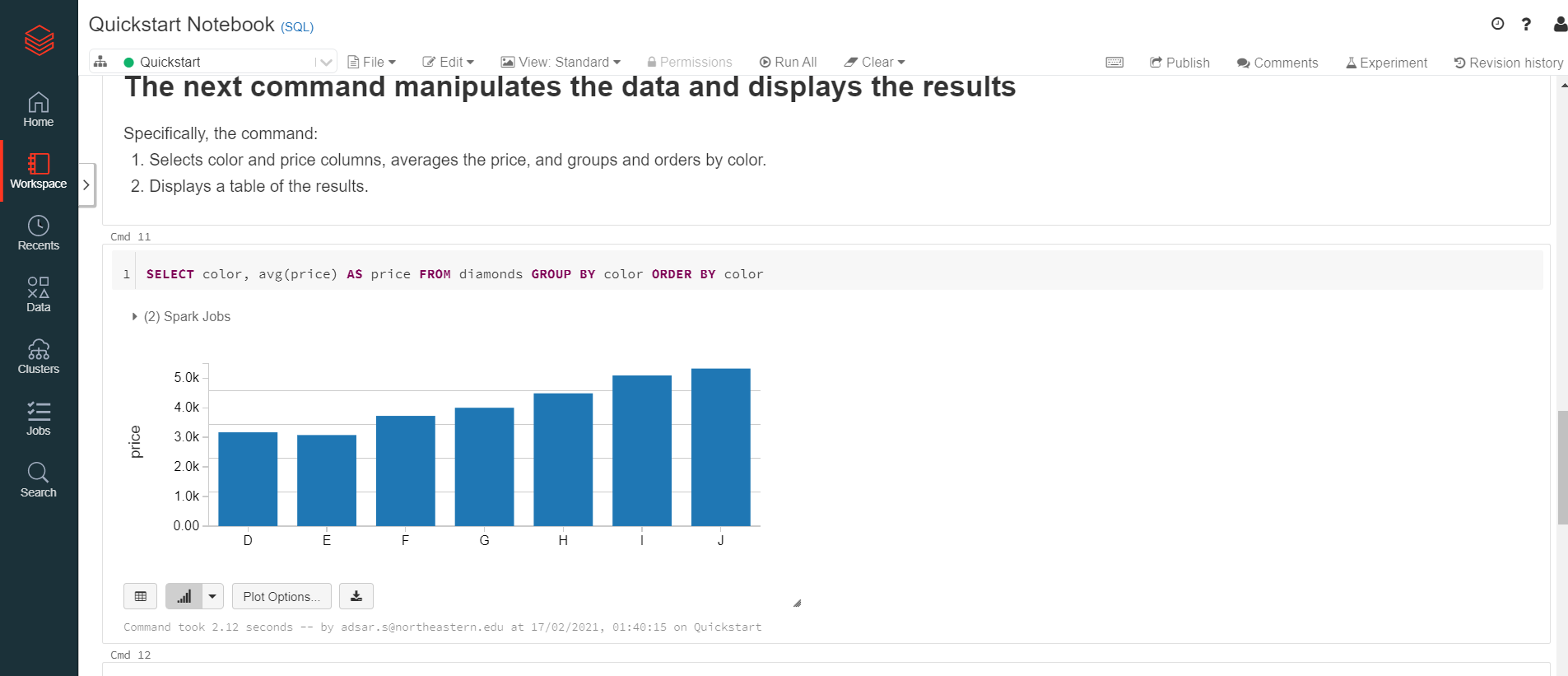


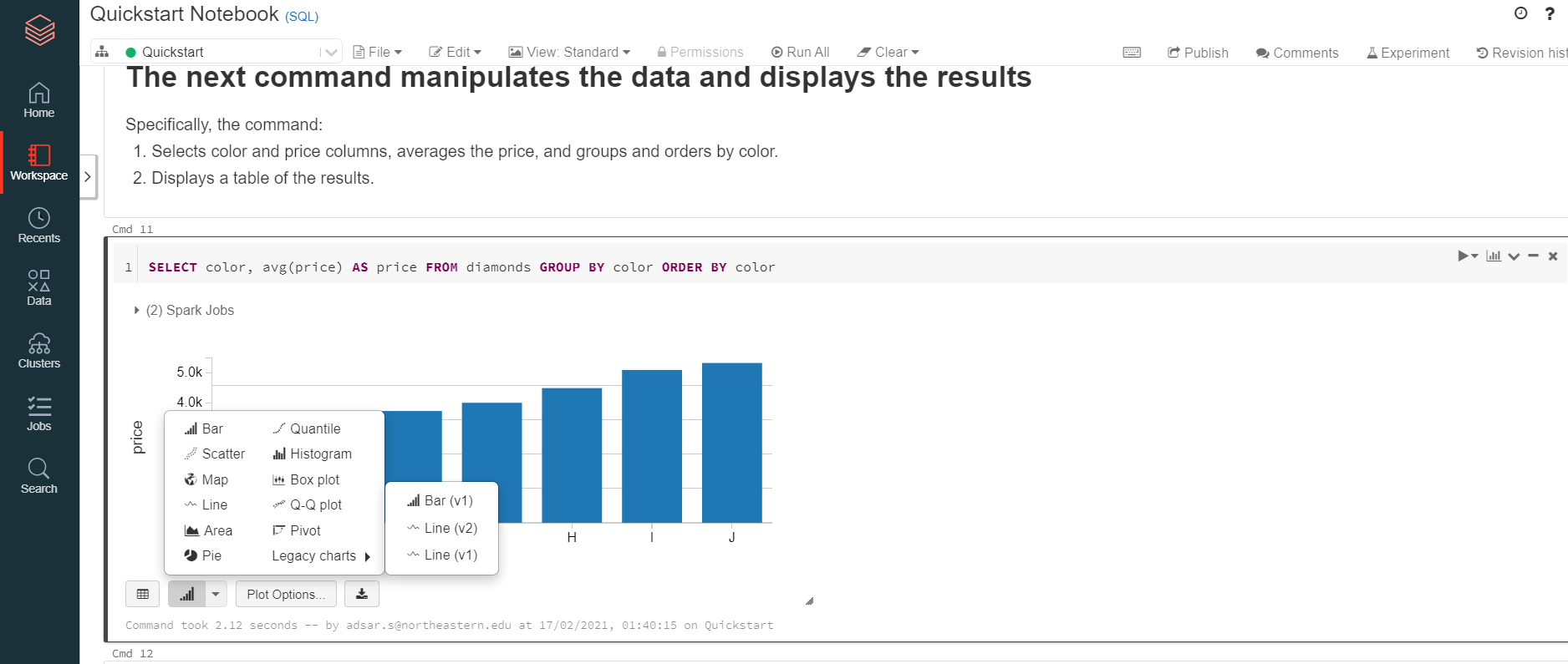
Moreover, the Databricks tool seems effective as we can view the output within seconds of execution.

Since, we have created a recent data import into a Delta location, we have used this data and dropped the previous dataset.

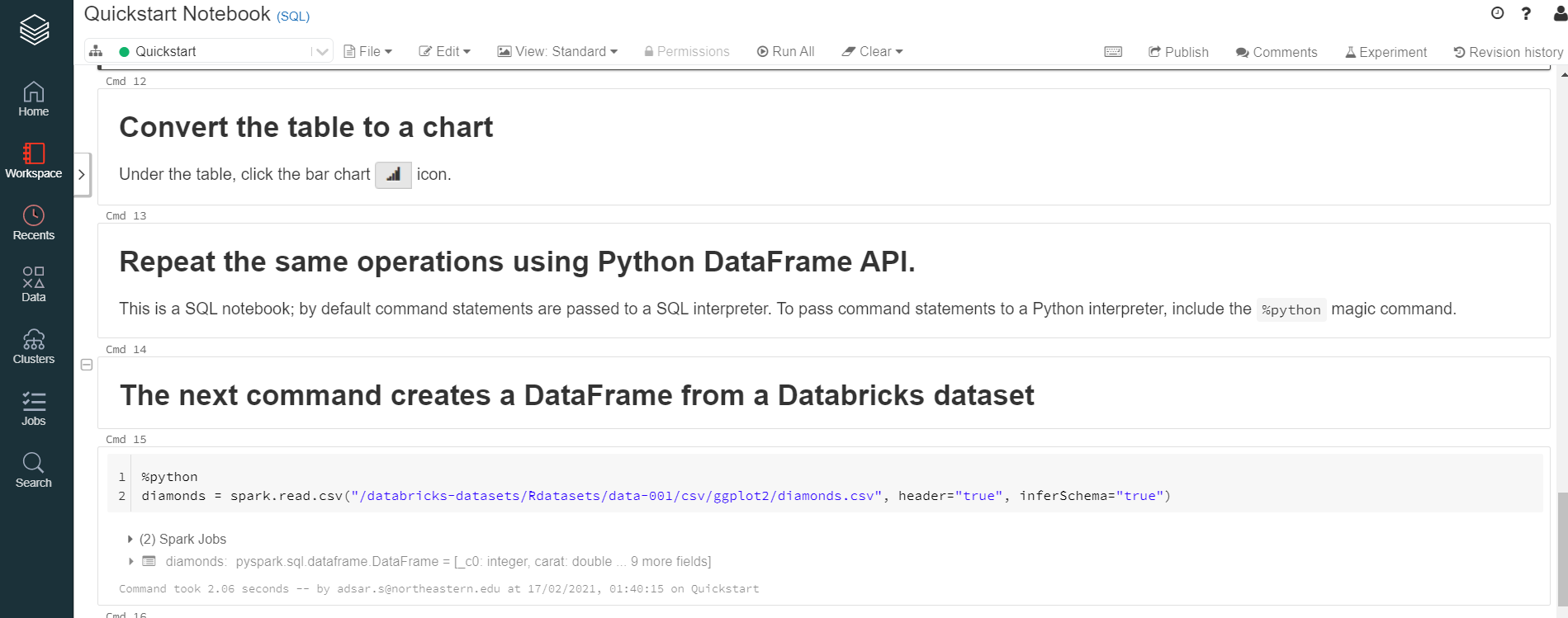


We can create data visualizations with Databricks by manipulating the data. In the below query, we have calculated the average price of diamonds and then grouped the results by a color.

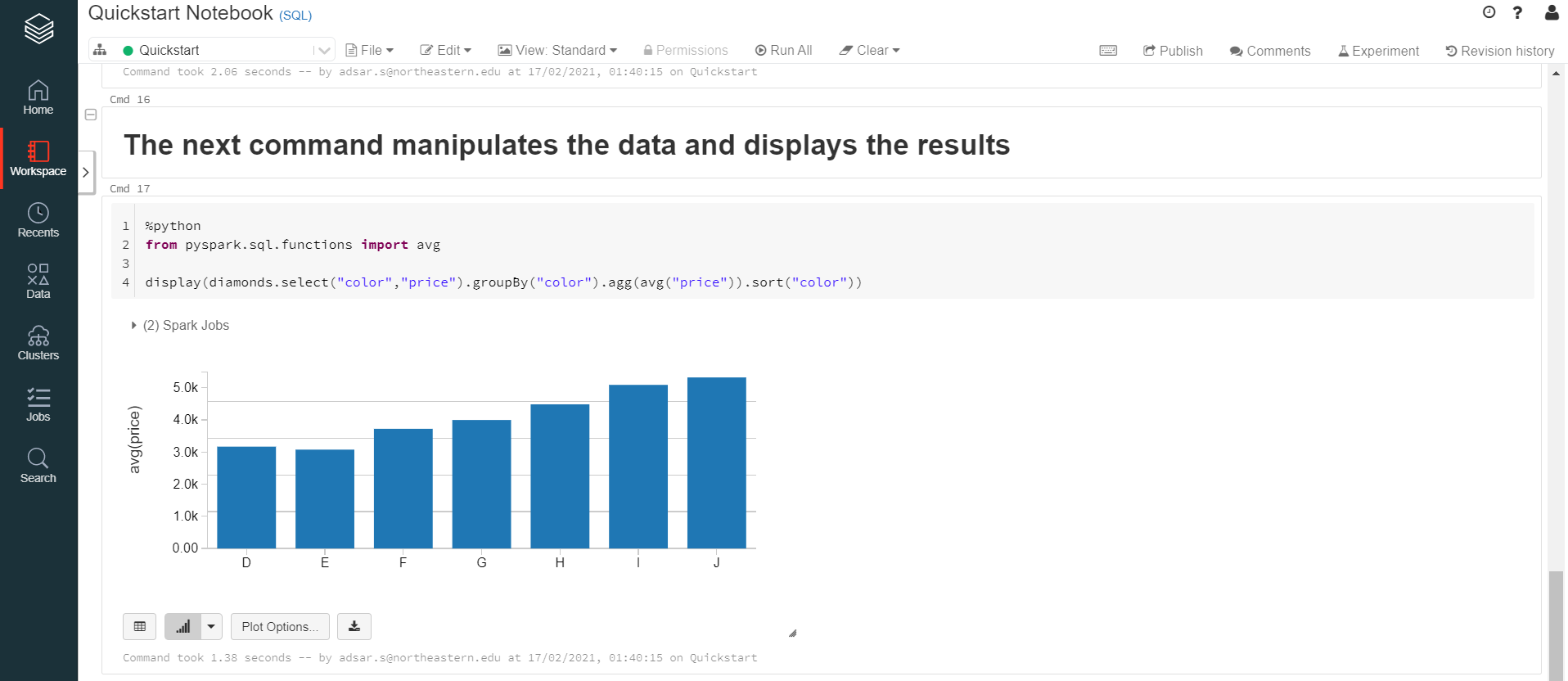


We can select any visualization based on the required data:  


We can create similar queries and results with python data frames.



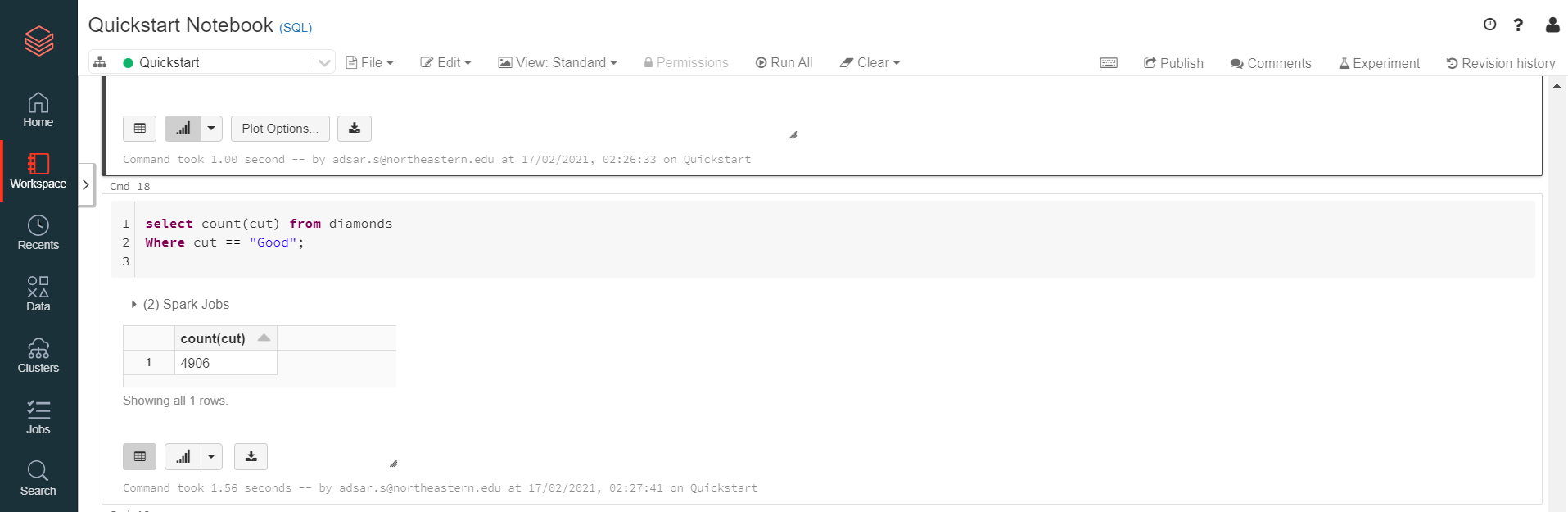
We have used python language for coding, this code selects the required library, and then, the diamonds are displayed based on the color and average price.



**Answers**

1. How many times is the word "Good" appears in the Diamonds dataset?

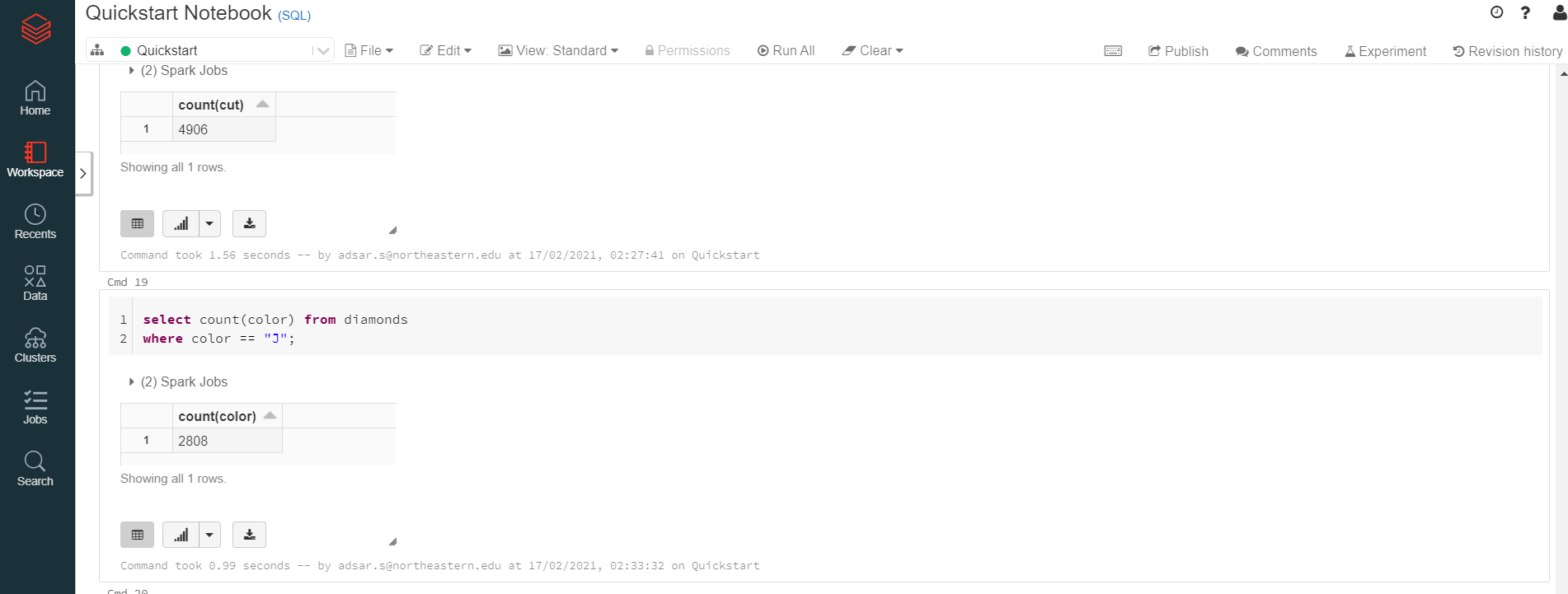
In order to view the number of times “Good” appears in the Diamonds data, we will use the WHERE clause to filter the data.



It can be seen that the word “Good” appears 4096 times in the diamonds data.

1. How many diamond's with colors "J' is in the diamonds dataset?

We have selected count for the color field in the diamonds data, and then filtered using the WHERE clause for the color “J”. It seems there are 2808 records that contain the color “J”.



**Comments**

Advantages

* The Databricks interactive platform is fast and effective for huge datasets and therefore yields higher performance
* The Community Edition is available free for students which helps in gaining practical experience without purchasing the licensed version.
* The platform provides training on Apache Spark for beginners which helps in gaining knowledge about the platform.
* The tool is user interactive and simple, so a beginner can interact easily.
* Moreover, It is easy to detect errors faced within the platform and provides good debugging facility.

Disadvantages

* The platform does not seem to have enough dashboarding tools to perform data visuals and dashboards to display relevant metrics.
* The Databricks platform seems costly, and less affordable for regular usage.
* Also, the platform does not support any integration with Git.

**Conclusion**

The Databricks is a cloud based interactive platform for managing and implementing big data analytics. We have worked on performing data analysis using SQL and Spark on the Diamonds dataset.

The platform seemed interactive for developing clusters and implementing SQL on them, for faster execution. Moreover, the platform supports Delta Lakes that help in faster execution of queries with Apache Spark. Moreover, the platform provides some tutorials that help in gaining information about the platform for easier navigation. The platform has the ability is manage huge volumes of data as it is scalable, and overall is reliable platform to perform big data analysis.

**References**

[1] Making big data easy. (2020, April 29). Retrieved February 17, 2021, from <https://databricks.com/blog/2014/07/14/databricks-cloud-making-big-data-easy.html>

[2] Spark sql: Manipulating structured data using apache spark. (2020, April 29). Retrieved February 17, 2021, from https://databricks.com/blog/2014/03/26/spark-sql-manipulating-structured-data-using-spark-2.html