Logo

Description automatically generated

**ALY 6110: Data Management and Big Data**

**Introduction to Sparks (Databricks)**

**CRN: 70633**

**Week 4: Assignment**

**NUID: 002938209**

**Email: jain.abhin@northeastern.edu**

**Submitted by: Abhinav Jain**

**Date: 10/20/2022**

**Submitted to: Prof. Valeriy Shevchenko**

**Option 2**

**Summary**

In this assignment, we are setting up the platform on the Data bricks community edition for the implementation of the project for a better understanding to handle big data in the databricks environment. As per the requirement, the tutorial from the data bricks community platform will give the hands-on to execute the queries on the diamond dataset. In this assignment, we are going to solve queries by using SQL.

By choosing option 2, using the Databricks server to form the cluster by using the community version and will use a notebook while doing the tutorial we are following steps to understand the basics of the databricks server. If we look at apache spark is an open source, which is distributed processing system to analyze the query in a fast environment. In 2017 Apache become the best and the most popular big data distributed processing framework with 365,000 meetups. Whereas, Apache start in 2009 as a part of UC Berkeley’s AMPlab with the involvement of a few researchers and students.

**Content:**

Step 1: The Login page opens after you click the link and enter your personal information. signing up for the data bricks server.link: [https://databricks.com/try-databricks Links to an external site.](https://databricks.com/try-databricks%20Links%20to%20an%20external%20site.).

Graphical user interface, text, application, email

Description automatically generated

Step 2: In this below step we can see as per the instruction we can perform the Guide for QuickStart Tutorial where we need to spin up a cluster, for running the queries where we have data already on the data bricks platform.

Graphical user interface, application

Description automatically generated

Step 3: In this step we can see that Quick start Notebook is displayed on the databricks platform where we must follow the instruction. First of all, we have to create the cluster from the sidebar, by right click and pressing the compute button to open the new window.

Graphical user interface, text, application, email

Description automatically generated

Step 3 Created the cluster which has 15 GB space and 2 cores with 1DBU which is data bricks unit processing per hour. It will provide the space in the server to analyze the data in databricks. The following screenshot will tell about learning about the databricks in 5 minutes. There is a quickstart cluster in which we have to open the sidebar with create cluster.

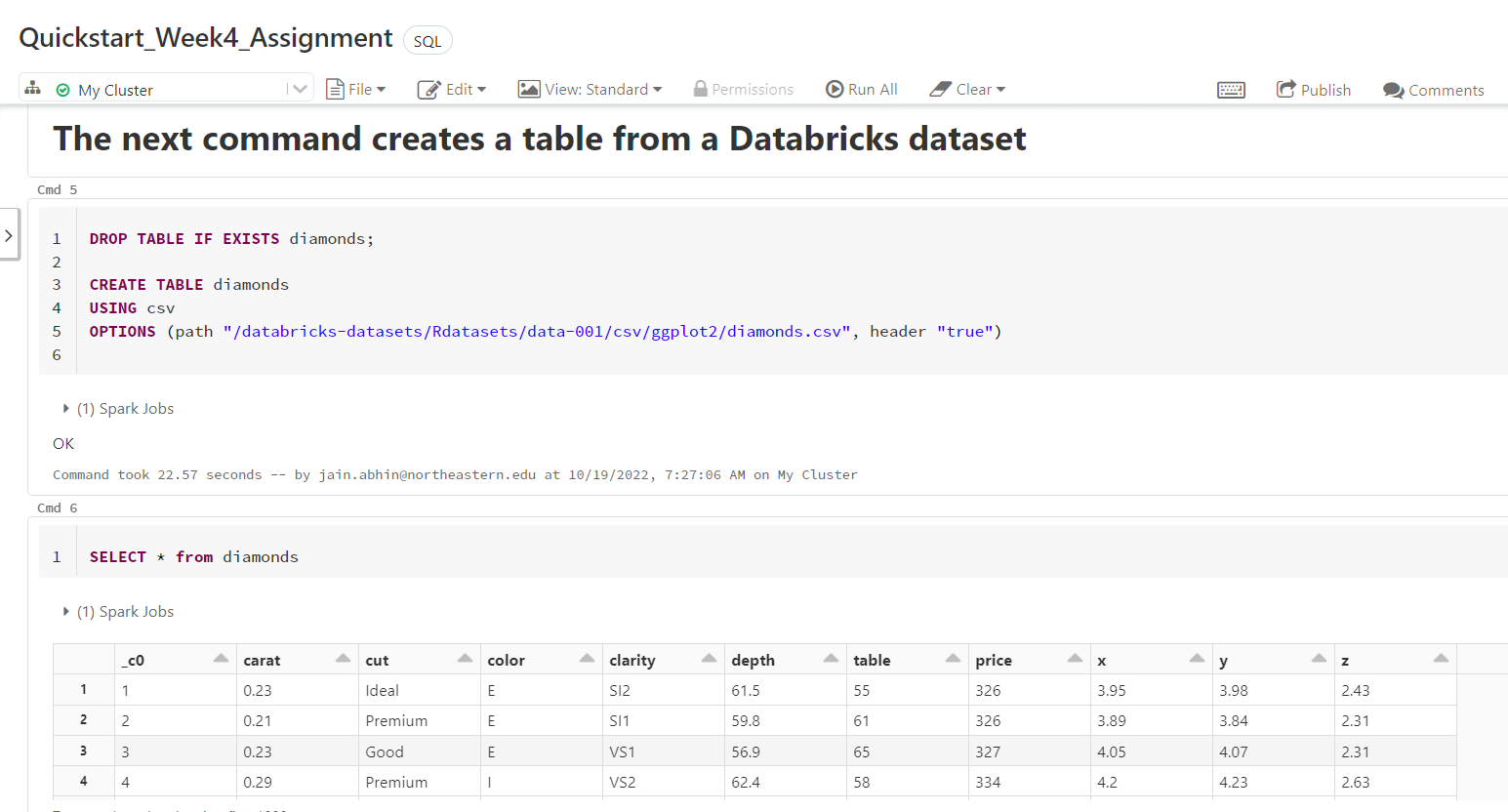
Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

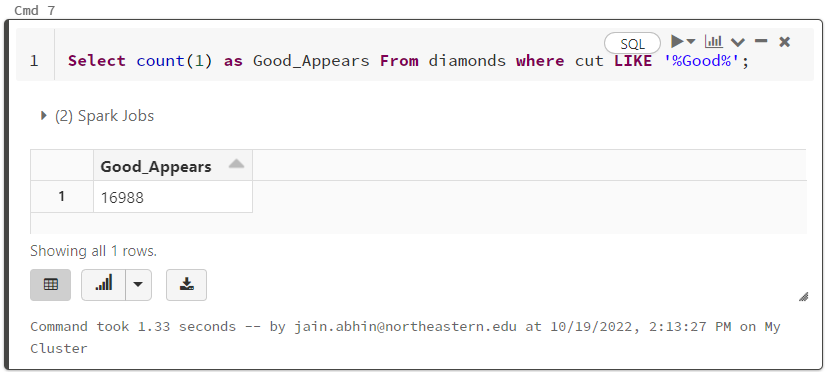
Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

Step 4: In the below screenshot we can find that by creating the table of diamonds dataset for the existing dataset from the databricks. We have 11 variables in the diamond dataset with their values by using those values we will try to find the answers of the following questions. In this selected diamond dataset.

***Question 1****: How many times is the word "Good" appears in the Diamonds dataset?*

Answer 1: In this query, we used where with like constraints we found the total number of words “GOOD” appeared when running the query on the diamond dataset 16988 times



***Question 2****: How many diamond's with colors "J' is in the diamonds dataset?*

Answer 2: In this step, we found the colors with “J” in the diamond dataset where the result for this count of color ‘J’ are 2808 in number

Graphical user interface, text, application, email

Description automatically generated

**Comments :Pros and Cons**

**Pros while performing the task**

* It provides an advanced analytics environment to analyze a huge amount of data with various tools.
* It is easy to use while implementing the data on the data bricks
* Fast processing speed for running the command.
* We can use multilanguage while performing the task on the same window. Although, I used SQL in this assignment.
* Apache is the most powerful tool to handle big data.
* Easy access to big data through API and databases.

**Cons while performing the task**

* The process of optimizing the process is not automatic, we have to run it by clicking on the run command.
* Maintaining the file in Apache spark is complex. It doesn’t provide its own file management system. It takes the services from Hadoop and Cloud servers
* The difficulty for the multi-user environment developer.
* It supports a limited number of large files than a small number of small files.

Initially, difficult to understand the environment as this was the new tool used to handle the big data while easy to create the cluster with limited space for the dataset. For that increase the space as per the project requirement. To resolve the issues initially it’s difficult to analyze the error and resolve it. But now its error becomes familiar with time.

**Conclusion Summary**

In conclusion, the most interesting part of this assignment is the databricks provide easy access which helps in understanding the big data in a short time and provide a quick solution to the query with the support of various programming language. In this, we have used SQL to perform the task, on the other hand, we can use python to provide better programming to handle big data.

Databricks server will be used to create the cluster using the community version, and we'll be utilizing a notebook as we work through the lesson to help us grasp the fundamentals of the databricks server. Apache Spark is a distributed processing system that is open source and designed to evaluate queries quickly. Apache spark is used as one of the open sources, to meet the fastest computing environment for well-designed architectures to meet the requirement of huge calculations. Highly used to capture a large amount of data load.

It provides better insights for the real-time system to get the analysis from the social media while building the large data analysis for managing the huge data in the data analytics environment. Machine learning gives the successful implementation with big data in various fields like fraud detection, healthcare, the manufacturing sector, tracking accidents for the insurance sectors, and tracking email. A spark is a tool that we can see from the outside. By using this tool, the performance of the various project increased, and get quick outcomes while implementing the projects.

**References:**

[1] Big Data

Schmarzo

<https://aws.amazon.com/big-data/what-is-spark/>

[2] The Future Of Big Data Analytics & Data Science: 5 Trends Of Tomorrow

<https://www.montecarlodata.com/blog-the-future-of-big-data-analytics-and-data-science/#:~:text=In%20the%20future%2C%20big%20data,informed%20decisions%20and%20increased%20competitiveness.>

[3] Building the Future with Databricks!

Syal

<https://medium.datadriveninvestor.com/building-the-future-with-databricks-f18eb702fba9>