****

**By**

**Sunil Raj Thota**

**ALY6110 - Data Management and Big Data**

**Class Number: 202135**

**Class Name: Spring 2021 CPS Quarter – Second Half**

**CRN: 80524**

**Week 4 Assignment 1 – Spark Basics in Databricks**

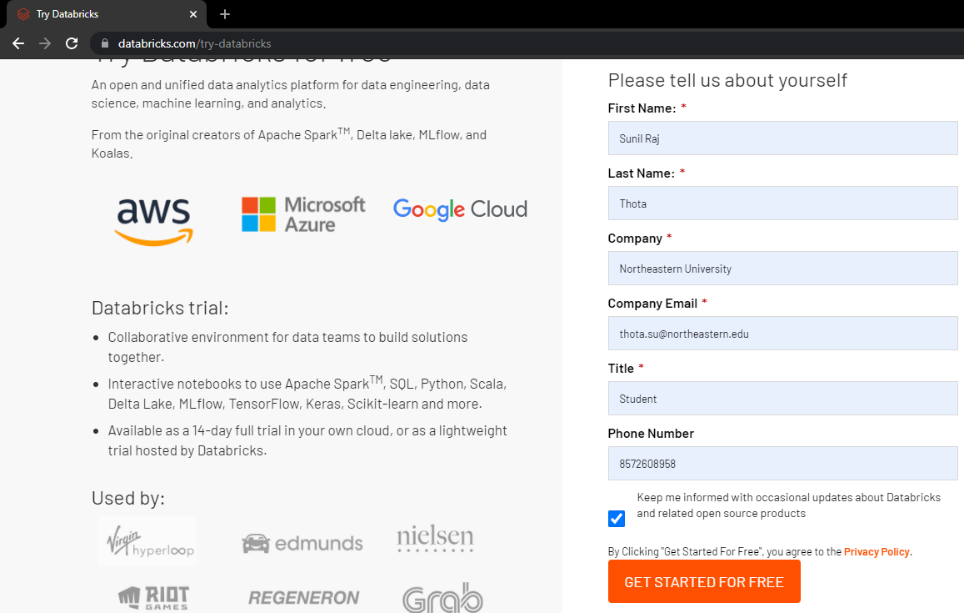
**thota.su@northeastern.edu**

**Summary:**

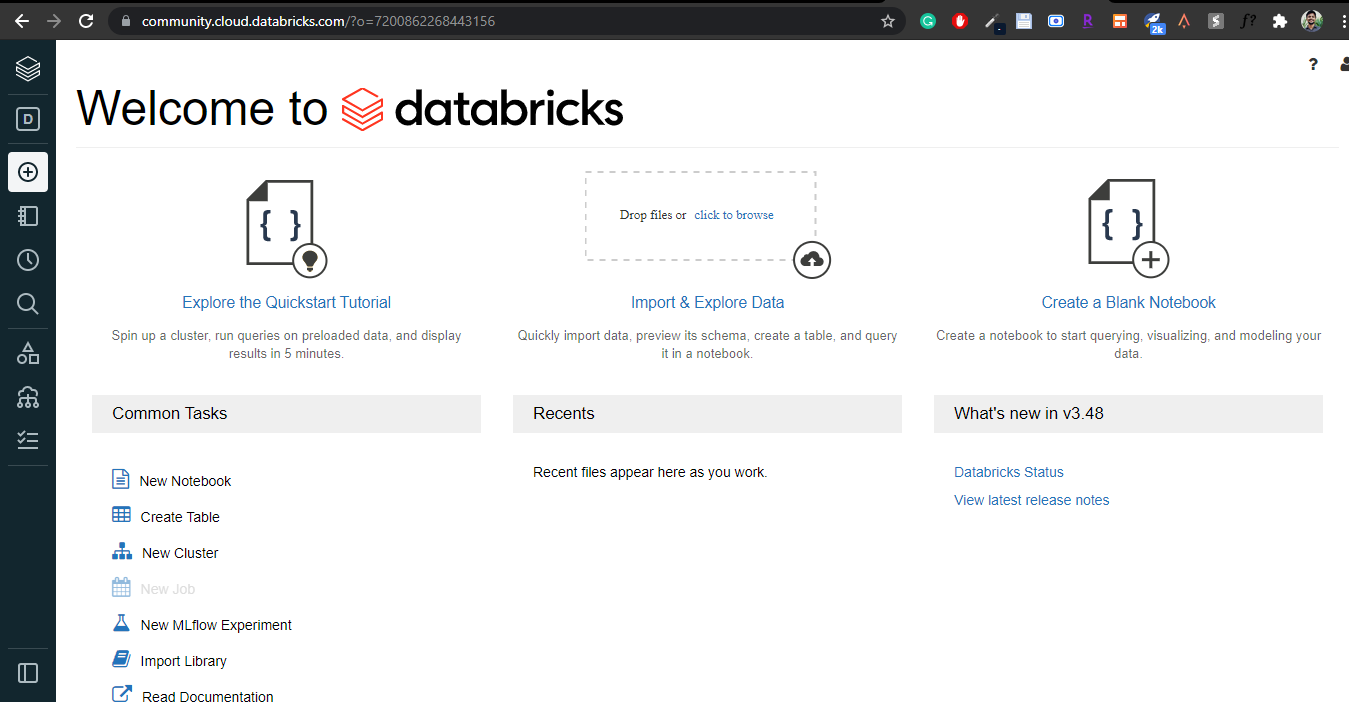
Every day the businesses are growing and producing large amounts of structured as well as non-structured data. This form of data is required for organizational teams of analysts especially data analysts to produce valuable insights from the following data. We'll go over how to use Spark within the Databricks platform in this assignment. It is a single Data Analytics platform that can be used for large-scale data engineering, data science, and SQL applications. It allows you to use Python or R to run Spark commands on a variety of machine learning libraries and datasets. Within Databricks, we'll be using Spark for this project.

We'll start by registering for Databricks and purchasing a community user subscription. Then, using Scala and Spark, we'll build a cluster and connect it to our notebook. For this project, we will be using a native dataset of ‘diamonds' from Databricks.

**Content:**

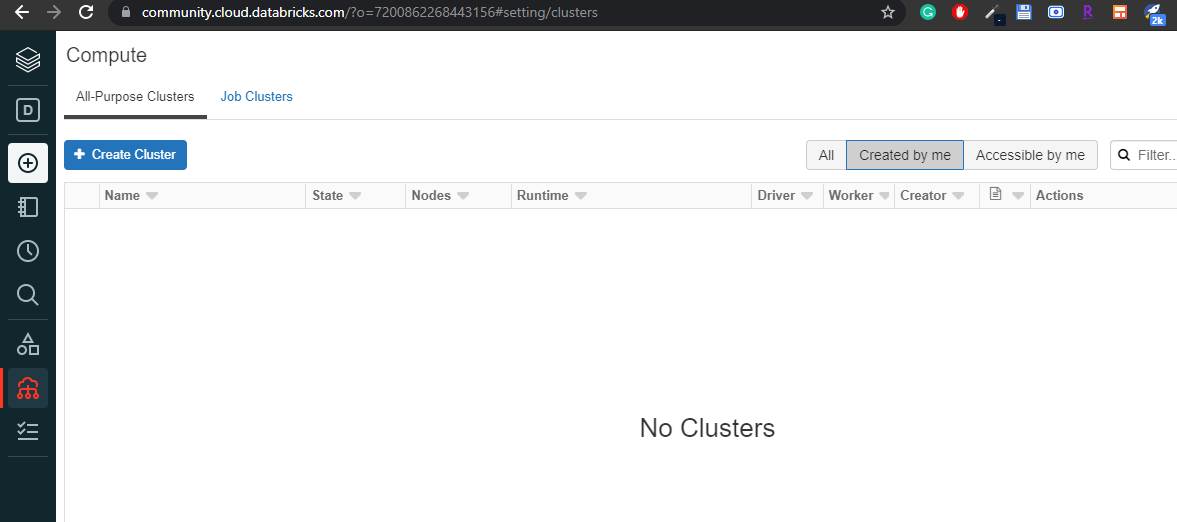


To begin, go to Databricks and register with a student id for free Community Edition access by following this link: <https://databricks.com/try-databricks>. Once the basic registration and setup are done, then we can log in to our account.

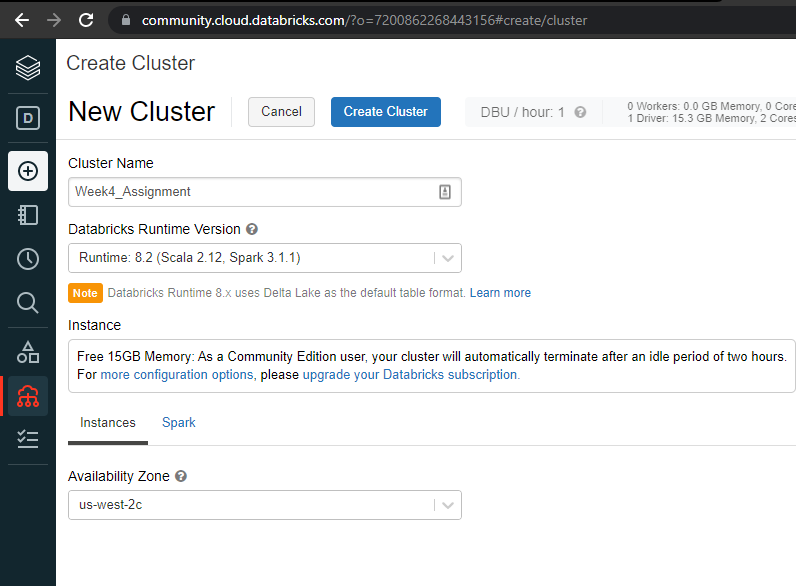


The Databricks Community Edition homepage is shown below. There is a catalog on the left-hand side that lists the many user functions available through the platform. Let’s begin with a Quickstart Guide as shown above.

To begin, click the Clusters button on the left-hand side to form a cluster. The page below will be shown.

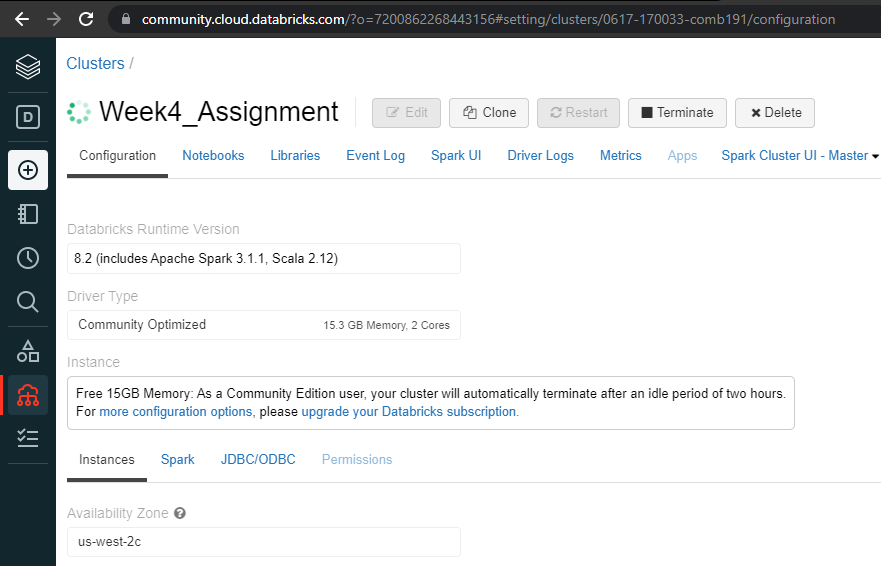


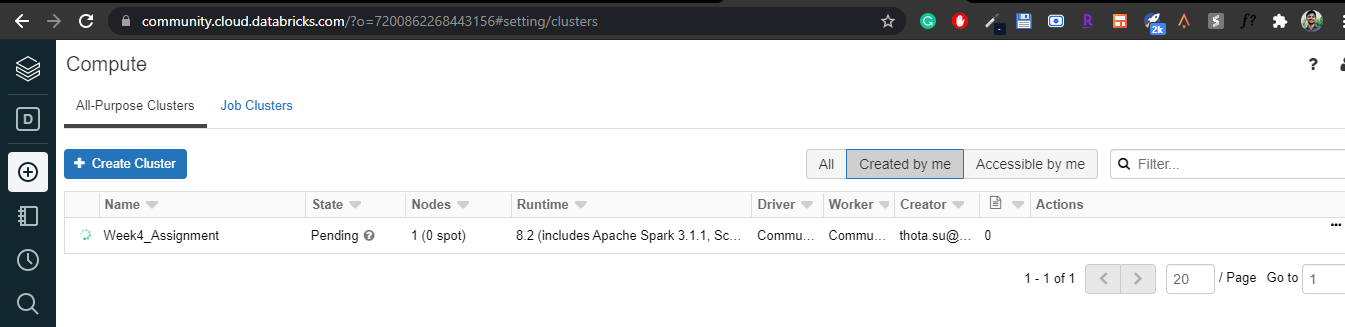
Click on the Create Cluster button and follow the below steps

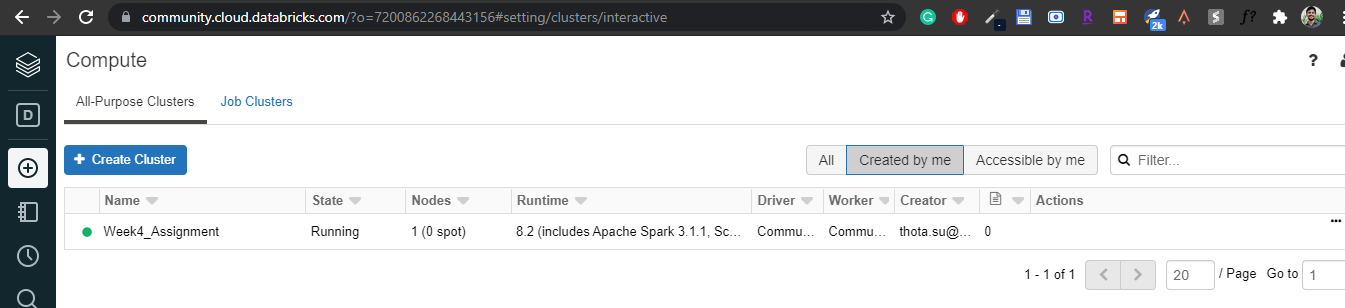


The Cluster can be renamed to ‘Week4\_Assignment.' Then choose Databricks Runtime Version 8.2. (Scala 2.12, Spark 3.1.1). And Availability Zone as ‘us-west-2c’. To create a cluster, click the ‘Create Cluster button. In Databricks, a cluster is a collection of processing resources and configuration settings used to conduct data engineering, data analytics, data science, and machine learning projects.

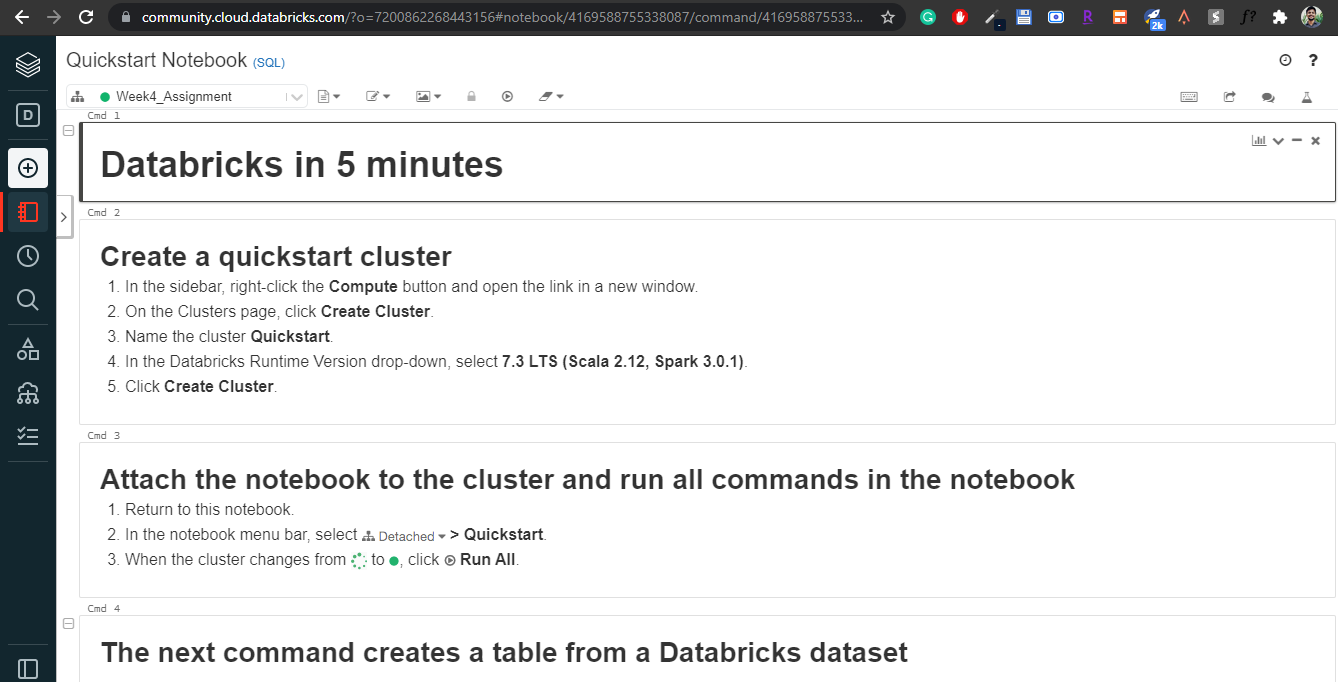
After that you will see a screen as shown below:







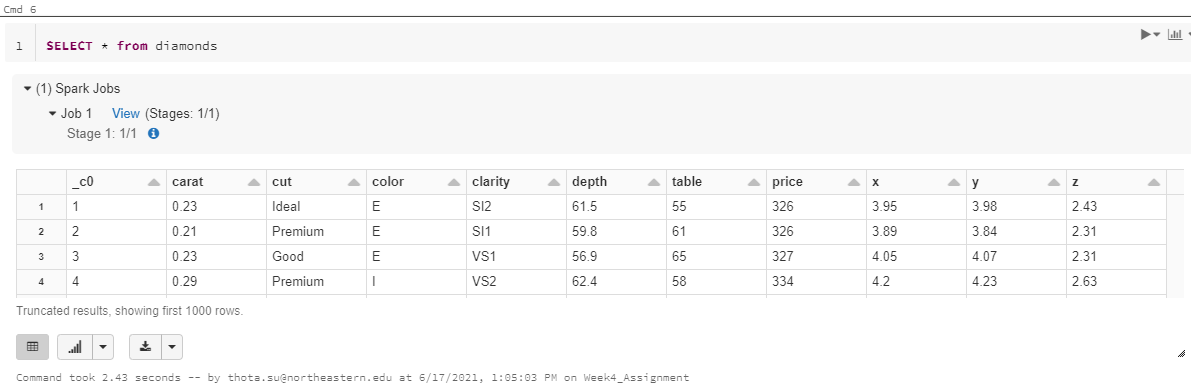
By selecting the Workspace button, we can now return to our notebook. As illustrated below, we must select a cluster from the top panel. As illustrated above, there is a green rotating circle at the start. We'll wait for the green dot to appear, as illustrated below.



Our necessary resources have now been added to our notebook. The Quickstart notebook includes code that has already been developed. We'll run each code one at a time, or we can click the Run button at the top and wait for it to finish, or we may press Ctrl + Enter to run a cell.



The CREATE TABLE command is used to create a table named ‘diamonds' using the CSV file type from the specified file path. The dataset will be stored in a table called diamonds by this command. To see the table's head, we can use the following command.



SELECT \* diamonds from the table

This \* operator will extract all of the data frame's columns.



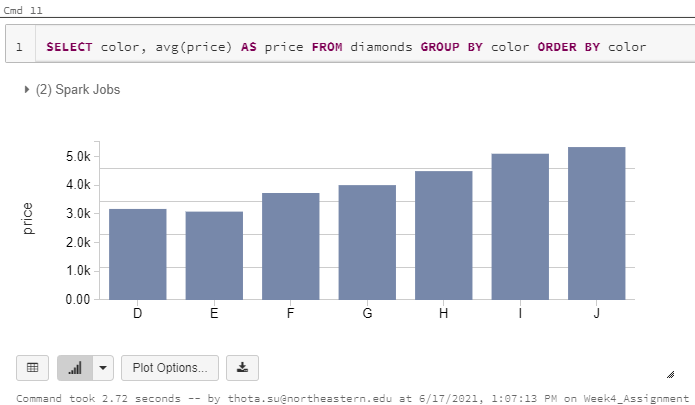
Using the Python programming language, we can perform the identical dataset retrieval procedure. We'll utilize the ‘%python' keyword to do so, and we'll be able to run Python scripts.

We can now edit data and run SQL queries according to our needs. We'll try to figure out how much diamonds cost based on their color. We can accomplish that by using the command below.

*SELECT color, avg(price) AS price FROM diamonds*

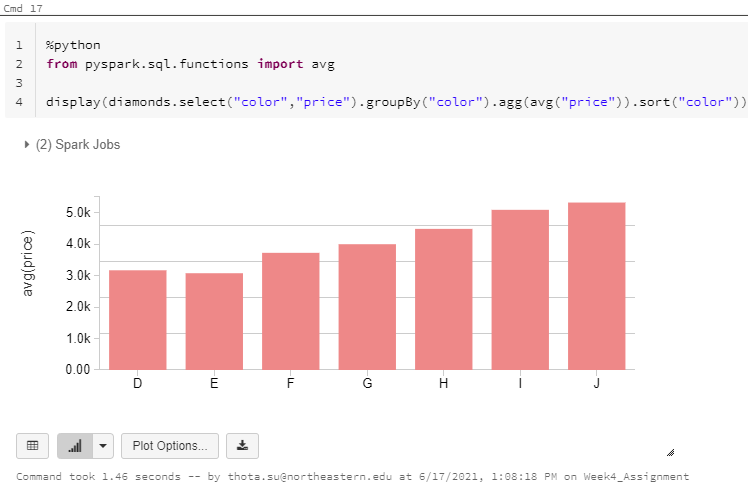
*GROUP BY color*

*ORDER BY color*



We take the table's color and average price, group it by color, then order it by color. Databricks offers a function that allows you to create graphs without having to perform any additional commands. Any table that is obtained by running a query can be displayed as a graph. This will generate a bar chart based on the pricing of the diamond hue.

Other plot options, such as line, area, and scatter plot charts, are available by selecting the chart button below. The following code shows how to achieve the same thing in Python. To get the same result, we employ the magical ‘% python' operator and perform the following command.



**Answers:**

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

The SQL command below is used to answer this question.

*SELECT COUNT(\*) FROM diamonds*

*WHERE cut LIKE 'Good' OR cut LIKE 'Very Good'*



**We can see 16988 as a result. The number of times the term "Good" appears in the data. This includes the occurrence in the word 'Very Good.' Because there are two values with the word ‘Good’ in them, we used a where clause for each of them, and the final result is shown above, with the total result being (Very Good -> 12082 + Good -> 4906), which equals 16988.**

1. **How many diamonds with the color "J' are in the Diamonds dataset?**

The SQL command below is used to answer this question.

*SELECT COUNT(color) FROM diamonds*

*WHERE color LIKE ‘J’*



It's clear that the Where condition is being used to check the total number of counts of diamonds with the color ‘J’, we discovered that there are only 2808 total count counts with ‘J’ as their color.

**Comments:**

**Pros:**

Databricks is a fantastic platform for anyone learning how to utilize Spark. Databricks is simple to use because it has a Spark setup for Python, R, and other machine learning libraries built-in. You don't have to manually configure Spark and set it up on the system, which is a time-consuming operation. Being a part of the big data community, I was always looking forward to putting my SQL and Python skills to the test. The introduction to Databricks allowed me to learn not just SQL, but also Scala, Python, and R.

**Cons:**

Only 15 GB of initial storage is available with Databricks. Any data that is larger than this will necessitate the purchase of additional storage. One issue I've run into is why, even though the instance is not larger, the cluster takes so long to start. A license is necessary for business use, as well as additional storage. It is, nevertheless, an excellent platform for learning and getting started for individuals.

**Conclusion:**

Databricks can be used for any type of analysis, as well as machine learning and data science research. It comes pre-configured for Spark, eliminating the need to do it manually. It also enables the execution of SQL queries on data. It is a cloud platform that can be used for any big data project. It is, however, a proprietary tool that requires a paid subscription.

It gives you 15 GB of storage space to import and save your files at first. Any data that is larger than that will necessitate the purchase of additional storage. It's a fantastic cloud service since it allows you to pick and choose resources and versions of Spark and Scala-based on your needs.

**References:**

[1] Databricks. Try Databricks for free. Retrieved from <https://databricks.com/try-databricks>