**Assignment 6**

**SEIS 736 Big Data Architecture**

Please complete your assignment and save it as a PDF or Word document then submit it electronically in the Assignments section of Canvas. If you submit multiple assignments, only the latest submission will be graded.

**Section 1 of 2: Reading invoice data in Spark**

In this homework assignment, you will be leveraging Databricks Community Edition and Spark to answer some questions regarding an invoice dataset. You may leverage Python or Scala, but you are expected to write code that completes the following tasks.

Databricks CE URL: <https://community.cloud.databricks.com/>

1. To get started, create a spark cluster in the Databricks console. Once your cluster is up and running, take a screenshot and post it below.
2. Read the invoice CSV into a resilient distributed dataset (RDD) using the code below. Collect the first five rows and print them. Take a screenshot of both the code and printed output and include it here.

invoice\_rdd = sc.textFile("/databricks-datasets/online\_retail/data-001/data.csv")

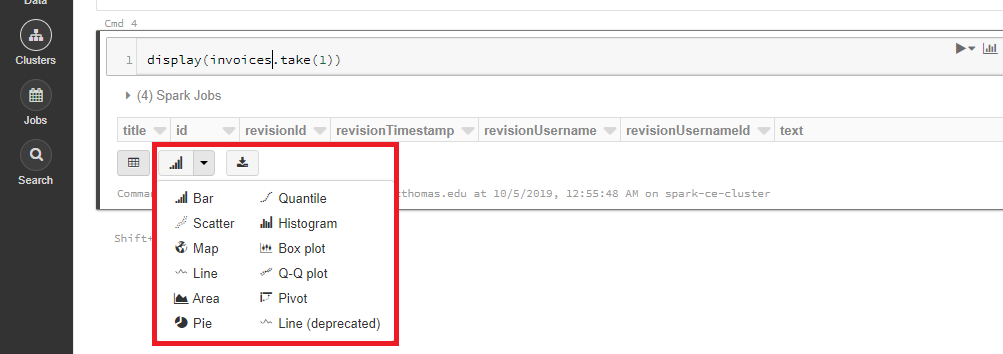
print(invoice\_rdd.take(5))

**Section 2 of 2: Answer the following questions regarding invoice data**

For each question below, please:

* Use **map** and **reduce** functions to answer the question.
* Provide the snippet of Spark code that you used to answer the question.
* Include a screenshot of your notebook that includes both the code and the printed answer.

1. Which customer in the dataset has spent the most on products? The quantity multiplied by the unit price will give you the total dollar amount spent per invoice line.
2. What is the product description for the best selling product in the dataset? We will define "Best Selling" as the product with the highest quantity sold.
3. How much has each country spent on products? The output should have two columns, one being the country and the other being the gross dollar amount spent across all products. Sort the output by the dollar amount, descending. Print the entire output, showing a gross dollar amount for each country.
4. What is the highest-grossing day in the dataset? Again, use quantity multiplied by unit price to get the revenue per line.
5. Finally, try out one of Databrick's visualizations. Note that you will need to convert back to a DataFrame in order to visualize the data (hint: look at rdd.toDF()). Create an appropriate DataFrame for visualization and call display on it.

Take a screenshot of your code and the resulting visualization. You can find available visualizations by expanding this icon at the bottom of a cell: