Name: Bas, Vince Justine C. Lab No. 1

Git Repo/ Colab Link: https://colab.research.google.com/drive/1LshEvcXwwPUaYztNJv6TnviWI7Ixac4o

Date: 1/27/25

Objective

Clearly state the objective of the lab. What is the purpose of this activity? What are you trying to achieve or learn?

To familiarize myself with the concept and use of Apache Spark, an indispensable tool used for quickly and efficiently scanning and manipulating large amounts of data for organizations to make strategic decisions.

Introduction

Provide background information relevant to the lab. Briefly explain the concepts, tools, or technologies involved. Include any theoretical concepts or algorithms being used.

The tools we utilized was a software called Spark, it is a data processing program used to quickly and effectively scan and/or manipulate large amounts of data that will be used to guide strategic decisions an organization makes.

Methodology

Describe the steps you followed in the lab. Include algorithms, data flow, or specific Spark transformations/actions if applicable. Use numbered or bullet points for clarity.

* **Installed and set up PySpark**
* **Configured PySpark to run on the terminal properly**
* **Ran the pyspark command, invoking PySpark**
* **Creating a spark app**
* **Creating an array of values**
* **Performing different transformations on those values**

Results and Analysis

Present the outcomes of your experiment. Use tables, graphs, or screenshots of results as needed. Provide a detailed analysis of the results, highlighting key findings.

Challenges and Solutions

Document any issues encountered during the lab and how you resolved them.

Challenge: Setting up my local system to support and run PySpark.

Solution: Utilizing the internet and its resources to troubleshoot and eventually fix the problem.

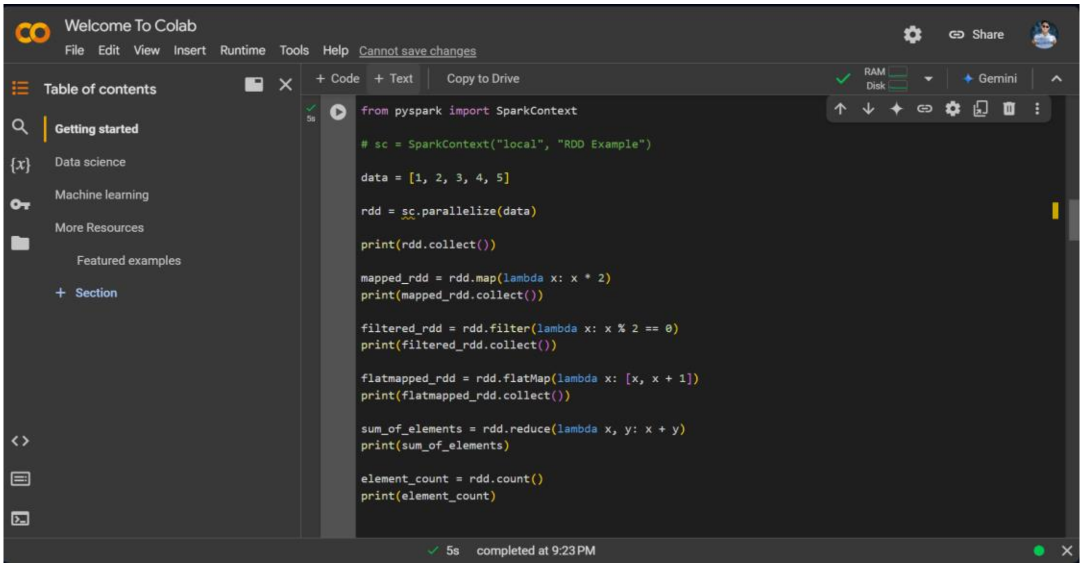
Conclusion

Summarize the key takeaways from the lab. Did you meet the objectives? What did you learn about the tools, techniques, or concepts?

Spark is a data processing tool indispensable to data analysts in order to efficiently and effectively scan and manipulate large amounts of data that will prove useful in allowing organizations to make good decisions based on the interpretation of the data.

Documentation:

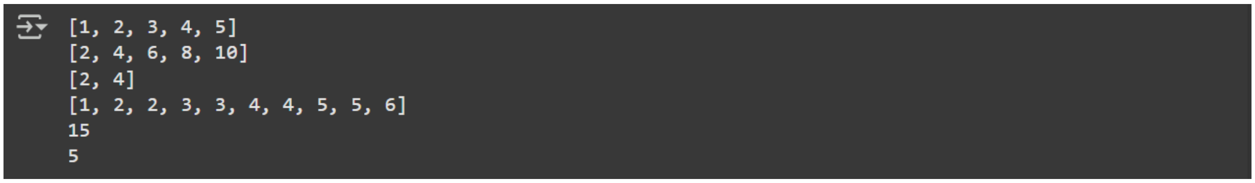
A screenshot of the initialization of an rdd, and the transformations I have performed upon the data:



```python
from pyspark import SparkContext

# sc = SparkContext("local", "RDD Example")

data = [1, 2, 3, 4, 5]

rdd = sc.parallelize(data)

print(rdd.collect())

mapped_rdd = rdd.map(lambda x: x * 2)
print(mapped_rdd.collect())

filtered_rdd = rdd.filter(lambda x: x % 2 == 0)
print(filtered_rdd.collect())

flatmapped_rdd = rdd.flatMap(lambda x: [x, x + 1])
print(flatmapped_rdd.collect())

sum_of_elements = rdd.reduce(lambda x, y: x + y)
print(sum_of_elements)

element_count = rdd.count()
print(element_count)
```

✓ 5s completed at 9:23 PM

The results I got from running the cell in google colab containing all the various transformations and values afterwards:



```
[1, 2, 3, 4, 5]
[2, 4, 6, 8, 10]
[2, 4]
[1, 2, 2, 3, 3, 4, 4, 5, 5, 6]
15
5
```