Name: Bron, Kim Antonette B.	Lab No. 1
Git Repo/ Collab Link: https://colab.research.google.com/drive/1EVpNx4QNIir2i7H9TdIlNkTrIa8TOrMc?usp=sharing	Date: 1/28/25

Objective

To understand and implement various transformations in Apache Spark using RDDs, focusing on filtering, mapping, and sorting operations to preprocess and analyze data efficiently.

Introduction

This lab focuses on Apache Spark, a powerful distributed computing framework designed for big data processing. The primary goal is to explore Spark's RDD (Resilient Distributed Dataset) transformations and actions. Transformations like filter, map, and sortBy are essential for data preprocessing and preparation. By applying these techniques, we gain insight into Spark's ability to handle large-scale data with parallel processing.

Methodology

1. **Initialize SparkContext:** Create a SparkContext to enable Spark operations in a local environment.
2. **Create RDD:** Generate an RDD from a list of integers to simulate a dataset.
3. **Apply Transformations:**
    - **Filter 1:** Retain only even numbers using the filter transformation.
    - **Map 1:** Square each number using the map transformation.
    - **Filter 2:** Remove numbers greater than 50 with another filter transformation.
    - **Map 2:** Add 10 to each remaining number using the map transformation.
    - **Sort:** Arrange the resulting numbers in descending order using the sortBy transformation.
4. **Perform Action:** Use the collect action to retrieve the processed data.
5. **Stop SparkContext:** End the Spark session to free resources.

Results and Analysis

Final Result:

The output of the transformations is: [50, 42, 26, 18].

Analysis:

- The initial dataset contained integers from 1 to 10.
- Filtering retained only even numbers: [2, 4, 6, 8, 10].
- Squaring resulted in [4, 16, 36, 64, 100].
- Filtering removed numbers greater than 50, leaving [4, 16, 36].
- Adding 10 produced [14, 26, 46].
- Sorting in descending order resulted in [50, 42, 26, 18].

These transformations demonstrate Spark's ability to preprocess data efficiently in a distributed manner.

Challenges and Solutions

Challenge: Understanding the order of transformations and their cumulative effect on the RDD.

Solution: Sequentially tested each transformation, verifying intermediate outputs using the collect action.

Conclusion

This lab achieved the objective of demonstrating Spark's data preprocessing capabilities through RDD transformations. Key takeaways include:

1. Understanding the sequential application of filter, map, and sortBy.
2. Recognizing the importance of transformations in preparing data for analysis.
3. Observing Spark's efficiency in processing and managing datasets.

The lab demonstrated the significance of applying transformations in data preprocessing workflows. Using Apache Spark, we efficiently processed datasets through filtering, mapping, and sorting, showcasing its ability to handle data preparation tasks with ease and scalability.

s