Naya College

Cloud Data Engineering

Apache Spark

Lab1: PySpark Word Count

Objective:
In this exercise, you will get hands-on experience with PySpark, a big data processing framework. Your primary objective is to understand the basics of distributed data processing using Resilient Distributed Datasets (RDDs) in Spark. By the end of this exercise, you will have developed a simple application that counts the frequency of each word in a given text.

Background:
Word count is a classic example in the big data realm, often used to demonstrate the fundamental concepts of distributed data processing. While seemingly simple, this exercise effectively illustrates the power and efficiency of distributed computing.

Description:
You will work on a PySpark application using the PyCharm IDE. The tasks involved include:

1. **Setting up the Workspace:**
    - Create a folder named **exercises_one** and a Python script titled **my_first_app.py** within this folder using PyCharm.
2. **Data Definition:**
    - Define a multi-line text string in the Python script:

```
text = """This is example of spark application in Python
Python is very common development language and it also one of Spark supported languages
The library of Spark in Python called PySpark
In this example you will implements word count application using PySpark
Good luck!!""".split("\n")
```

3. **Distributed Data Processing:**
    - Convert the text string into an RDD, which distributes the data across the Spark cluster.
    - Split the text into individual words.
    - Convert these words into key-value pairs, with each word being a key and the value being 1.
    - Aggregate these pairs by their key (word) to get the total count for each unique word.
4. **Output:**
    - Display the count of each word in the console.
5. **Cleanup:**
    - Properly stop the Spark context to ensure efficient resource usage.

Expected Outcome:
Once you run the PySpark application, you should see a list of words from the given text and

their respective counts. This will give you insights into the most frequently occurring words in the text.

Skills Developed:
Upon completion, you will have practiced:

- Setting up and using Spark Context.
- Creating and manipulating RDDs.
- Applying transformations and actions on RDDs.
- Understanding key-value pair operations in Spark.

Tools/Software Needed:

- PyCharm IDE
- PySpark