sample

March 3, 2025

[ ]:
```python
# Import necessary modules from PySpark
from pyspark import SparkContext
from pyspark.sql import SparkSession

# Option A: Using SparkContext to work with RDDs
def test_rdd():
    # Create a SparkContext (runs in local mode)
    sc = SparkContext("local", "SimpleRDDTest")
    # Create an RDD from a list
    numbers = sc.parallelize([10, 20, 30, 40, 50])
    # Compute the sum of the RDD elements
    total = numbers.reduce(lambda a, b: a + b)
    print("Sum of numbers in the RDD:", total)
    # Stop the SparkContext
    sc.stop()

# Option B: Using SparkSession to work with DataFrames
def test_dataframe():
    # Create a SparkSession
    spark = SparkSession.builder.appName("SimpleDataFrameTest").getOrCreate()
    # Create a simple DataFrame from a list of tuples
    data = [("Alice", 34), ("Bob", 45), ("Cathy", 29)]
    df = spark.createDataFrame(data, ["Name", "Age"])
    # Show the DataFrame
    df.show()
    # Stop the SparkSession
    spark.stop()

if __name__ == "__main__":
    print("Testing Spark with RDDs:")
    test_rdd()
    print("\nTesting Spark with DataFrames:")
    test_dataframe()
```