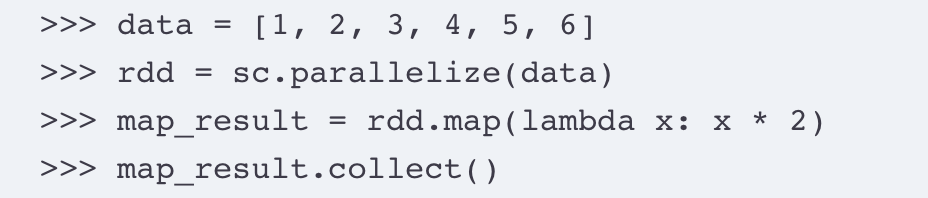
Assignment 4: Spark with Python

Name:

1. Execute the Transformation Operators: map (function), filter (function), and distinct (function) in the following, write down the commands and paste the output screenshots.



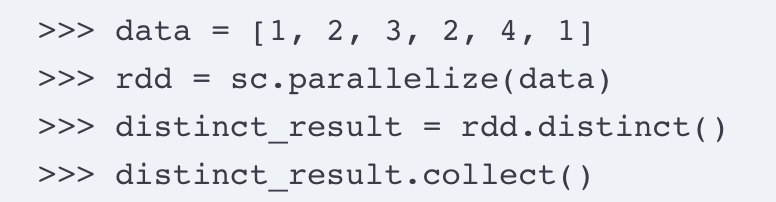
Text

Description automatically generated



Text

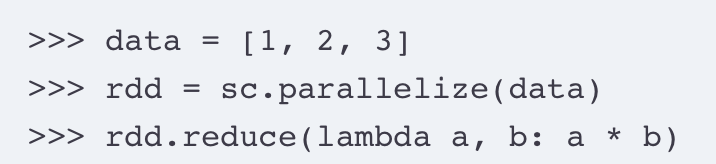
Description automatically generated



Text

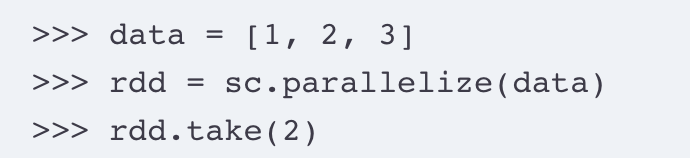
Description automatically generated

2. Execute the Action Operators: reduce ( ), take(n), and takeOrdered (n, key=func) in the following, write down the commands and paste the output screenshots.



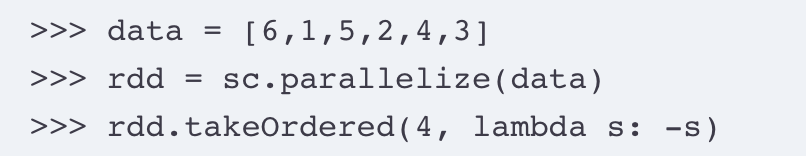
Text

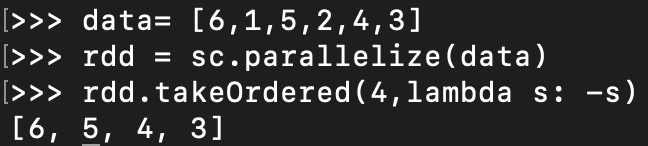
Description automatically generated



Text

Description automatically generated





3. Create a Python program that calculates the maximum temperature for each year using PySpark. The program should read the sample.txt data file from the HDFS input folder and place the output file in the output folder on HDFS. Execute the python program in PySpark, write down each steps of the commands, and paste the output screenshots.

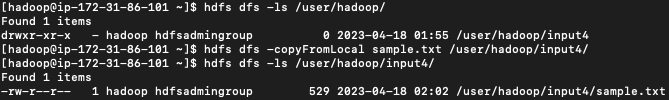
Create a directory for input file – input4

hdfs dfs -mkdir /user/Hadoop/input4/



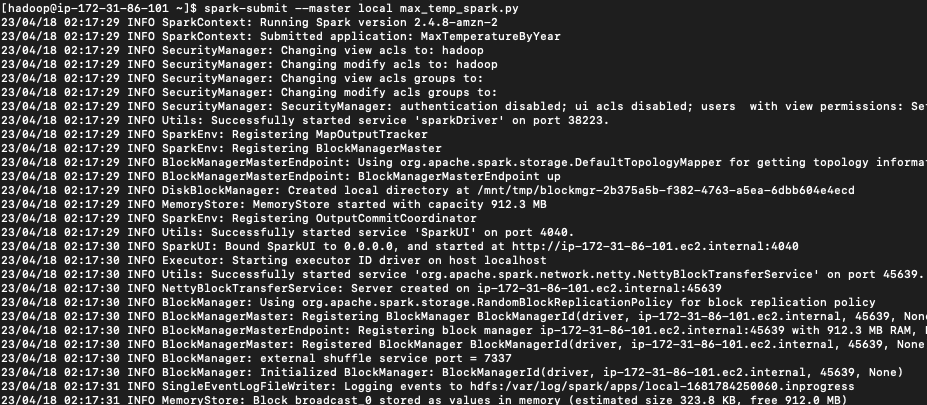
Copy the input file from local to input folder

hdfs dfs -copyFromLocal sample.txt /user/hadoop/input4/

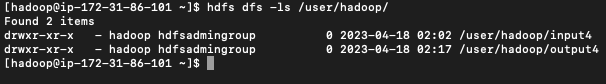


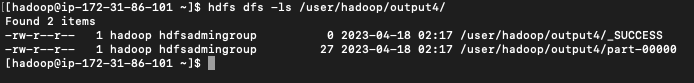
Run a max\_temp\_spark.py python file

Spark-submit –master local max\_temp\_spark.py



Output file – output4





Displaying output folder contents using -cat

Hdfs dfs -cat /user/Hadoop/output4/part-00000

Text

Description automatically generated