Name: Sarvesh Patil
Class: D20A
Roll No: 52

Mini project CA2

Write MapReduce/Spark Program to perform

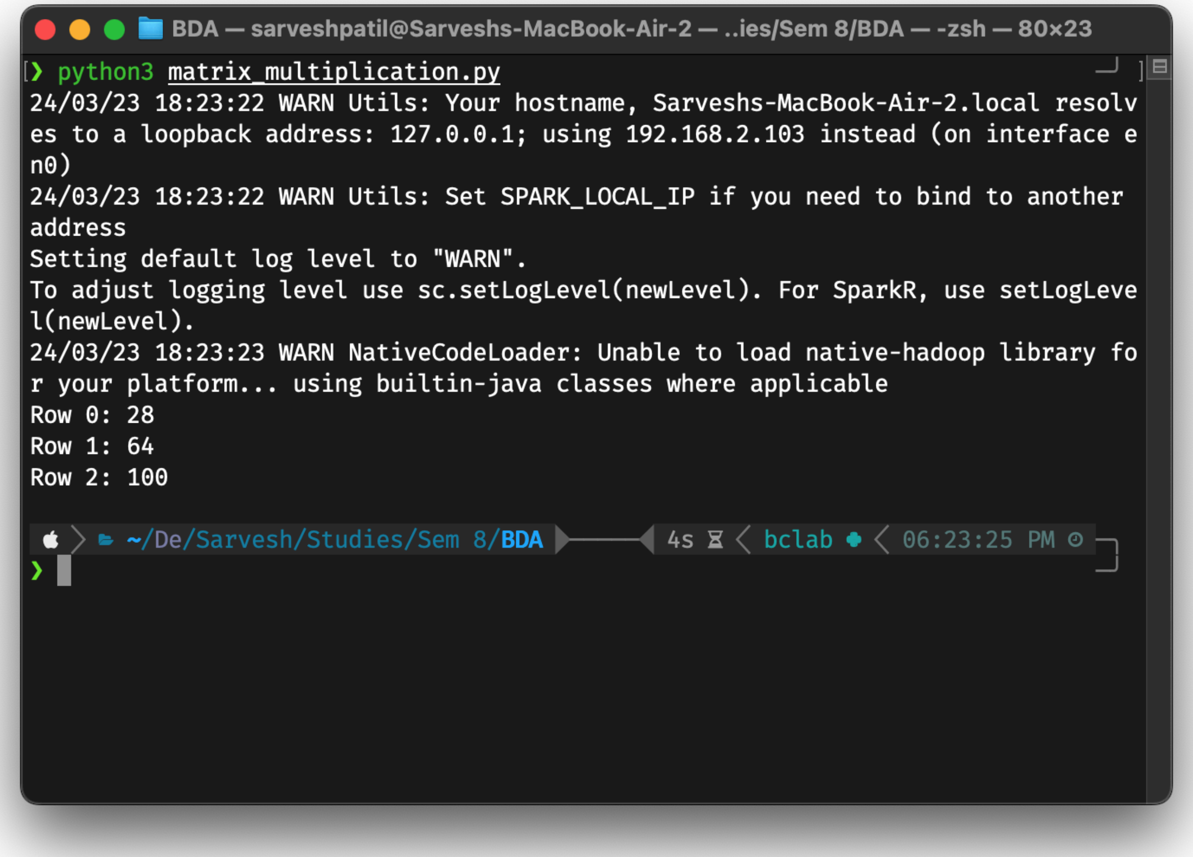
1. **Matrix Vector Multiplication**

Code:

```
from pyspark import SparkContext, SparkConf

# Initialize SparkContext
conf = SparkConf().setAppName("MatrixVectorMultiplication")
sc = SparkContext(conf=conf)

# Input matrix and vector
matrix = [
   (0, [1, 2, 3]),
   (1, [4, 5, 6]),
   (2, [7, 8, 9])
]
vector = [2, 4, 6]

# Broadcast the vector to all nodes in the cluster
broadcast_vector = sc.broadcast(vector)

# Perform matrix-vector multiplication using MapReduce
result = sc.parallelize(matrix) \
           .map(lambda row: (row[0], sum([row[1][i] * broadcast_vector.value[i] for i in range(len(row[1]))]))) \
   .collect()

# Print the result
for row_id, value in sorted(result):
   print(f"Row {row_id}: {value}")

# Stop SparkContext
```

```
sc.stop()
```

Output:



BDA — sarveshpatil@Sarveshs-MacBook-Air-2 — ..ies/Sem 8/BDA — -zsh — 80×23

```
❯ python3 matrix_multiplication.py
24/03/23 18:23:22 WARN Utils: Your hostname, Sarveshs-MacBook-Air-2.local resolv
es to a loopback address: 127.0.0.1; using 192.168.2.103 instead (on interface e
n0)
24/03/23 18:23:22 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
24/03/23 18:23:23 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
Row 0: 28
Row 1: 64
Row 2: 100

 ~/De/Sarvesh/Studies/Sem 8/BDA ──────── 4s ⌛ bclab ● 06:23:25 PM
❯
```

Name: Sarvesh Patil
Class: D20A
Roll No: 52

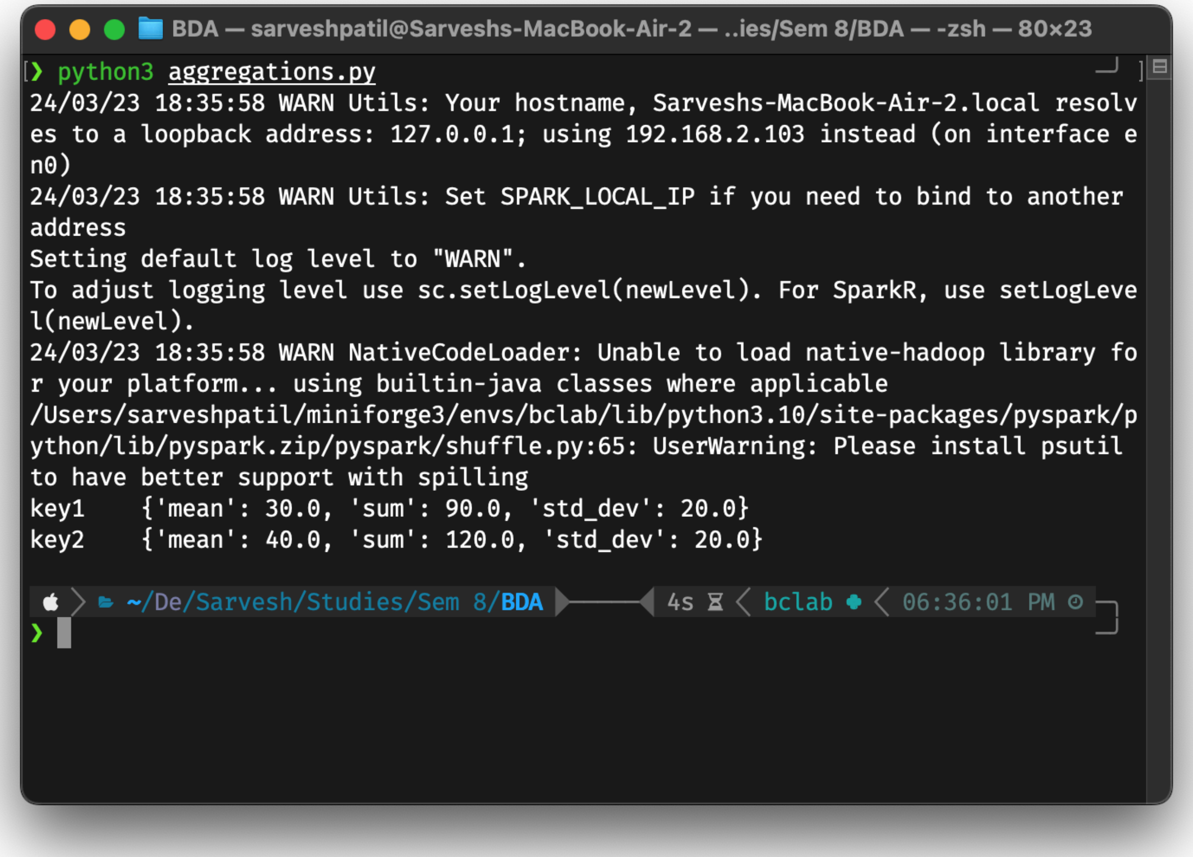
2. Aggregations - Mean, Sum, Std Deviation

Code:

```
from pyspark import SparkContext
from math import sqrt

# Dummy input data
input_data = [
    'key1\t10',
    'key2\t20',
    'key1\t30',
    'key2\t40',
    'key1\t50',
    'key2\t60',
]

def map_func(line):
    key, value = line.split('\t')
    return key, float(value)

def reduce_func(data):
    values = [x for x in data]
    mean_val = sum(values) / len(values)
    sum_val = sum(values)
     std_dev_val = sqrt(sum((x - mean_val)**2 for x in values) / (len(values) - 1)) if len(values) > 1 else 0
    return {
        'mean': mean_val,
        'sum': sum_val,
        'std_dev': std_dev_val
    }

if __name__ == '__main__':
    sc = SparkContext('local', 'AggregationSpark')
    lines = sc.parallelize(input_data)
```

```
mapped = lines.map(map_func)
grouped = mapped.groupByKey()
result = grouped.mapValues(list).mapValues(reduce_func)
output = result.collect()
for key, value in output:
    print(f'{key}\t{value}')
sc.stop()
```

Output:


BDA — sarveshpatil@Sarveshs-MacBook-Air-2 — ..ies/Sem 8/BDA — -zsh — 80×23

```
❯ python3 aggregations.py
24/03/23 18:35:58 WARN Utils: Your hostname, Sarveshs-MacBook-Air-2.local resolv
es to a loopback address: 127.0.0.1; using 192.168.2.103 instead (on interface e
n0)
24/03/23 18:35:58 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
24/03/23 18:35:58 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
/Users/sarveshpatil/miniforge3/envs/bclab/lib/python3.10/site-packages/pyspark/p
ython/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil
to have better support with spilling
key1    {'mean': 30.0, 'sum': 90.0, 'std_dev': 20.0}
key2    {'mean': 40.0, 'sum': 120.0, 'std_dev': 20.0}

  ~/De/Sarvesh/Studies/Sem 8/BDA ──────── 4s ⌛ bclab  06:36:01 PM
❯
```

Name: Sarvesh Patil
Class: D20A
Roll No: 52

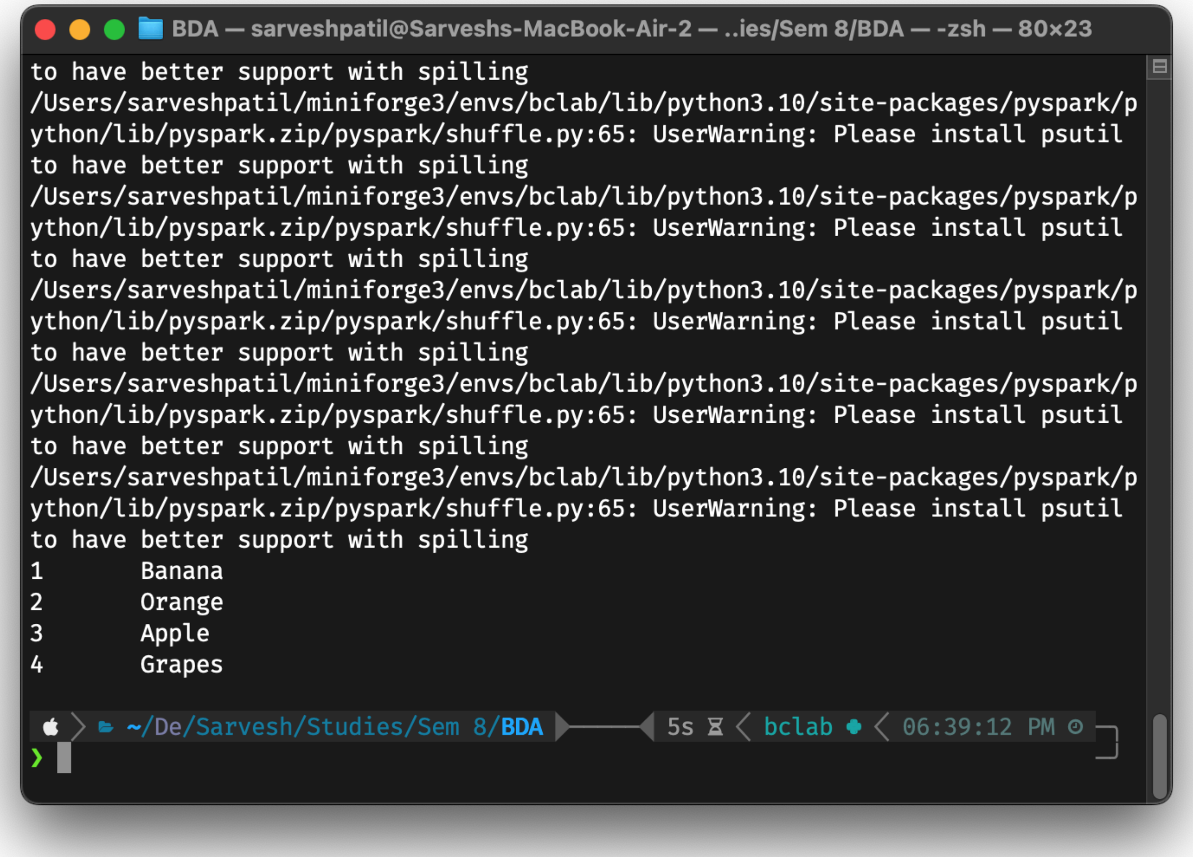
3. Sort the data

Code:

```
from pyspark.sql import SparkSession

# Create a Spark session
spark = SparkSession.builder \
    .appName("SortData") \
    .getOrCreate()

# Define dummy input data
dummy_data = [
    "3\tApple",
    "1\tBanana",
    "2\tOrange",
    "4\tGrapes"
]

# Create RDD from dummy data
data_rdd = spark.sparkContext.parallelize(dummy_data)

# Sort the data
sorted_data = data_rdd.sortBy(lambda x: x.split('\t')[0])

# Collect and print the sorted data
sorted_results = sorted_data.collect()
for result in sorted_results:
    print(result)

# Stop the Spark session
spark.stop()
```

Name: Sarvesh Patil
Class: D20A
Roll No: 52

Output:


BDA — sarveshpatil@Sarveshs-MacBook-Air-2 — ..ies/Sem 8/BDA — -zsh — 80×23

```
to have better support with spilling
/Users/sarveshpatil/miniforge3/envs/bclab/lib/python3.10/site-packages/pyspark/p
ython/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil
to have better support with spilling
/Users/sarveshpatil/miniforge3/envs/bclab/lib/python3.10/site-packages/pyspark/p
ython/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil
to have better support with spilling
/Users/sarveshpatil/miniforge3/envs/bclab/lib/python3.10/site-packages/pyspark/p
ython/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil
to have better support with spilling
/Users/sarveshpatil/miniforge3/envs/bclab/lib/python3.10/site-packages/pyspark/p
ython/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil
to have better support with spilling
/Users/sarveshpatil/miniforge3/envs/bclab/lib/python3.10/site-packages/pyspark/p
ython/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil
to have better support with spilling
1       Banana
2       Orange
3       Apple
4       Grapes

 ~/De/Sarvesh/Studies/Sem 8/BDA ────── 5s ⧗  bclab  06:39:12 PM
❯
```

Name: Sarvesh Patil
Class: D20A
Roll No: 52

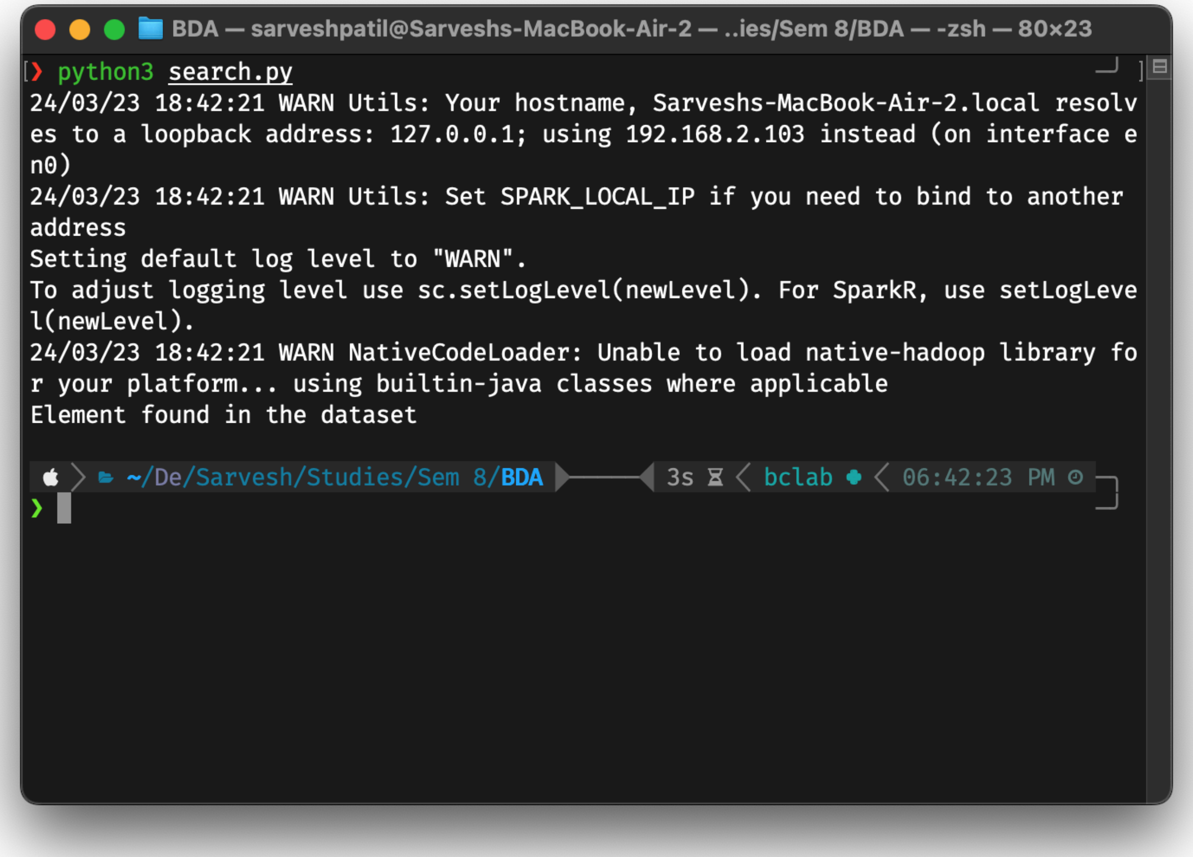
4. Search a data element

Code:

```
from pyspark import SparkContext, SparkConf

# Create a Spark context
conf = SparkConf().setAppName("SearchElement").setMaster("local")
sc = SparkContext(conf=conf)

# Define the data to be searched
data = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

# Parallelize the data into RDD (Resilient Distributed Dataset)
rdd = sc.parallelize(data)

# Define the search function
def search_element(element):
    return element == 5  # Change the search element as needed

# Map function to search for the element in the dataset
result = rdd.map(search_element)

# Collect the results
search_result = result.collect()

# Print the search result
if True in search_result:
    print("Element found in the dataset")
else:
    print("Element not found in the dataset")

# Stop the Spark context
sc.stop()
```

Name: Sarvesh Patil
Class: D20A
Roll No: 52

Output:



BDA — sarveshpatil@Sarveshs-MacBook-Air-2 — ..ies/Sem 8/BDA — -zsh — 80×23

```
❯ python3 search.py
24/03/23 18:42:21 WARN Utils: Your hostname, Sarveshs-MacBook-Air-2.local resolv
es to a loopback address: 127.0.0.1; using 192.168.2.103 instead (on interface e
n0)
24/03/23 18:42:21 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
24/03/23 18:42:21 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
Element found in the dataset

 ~/De/Sarvesh/Studies/Sem 8/BDA ─────── 3s ⌛ bclab ● 06:42:23 PM
❯
```

Name: Sarvesh Patil
Class: D20A
Roll No: 52

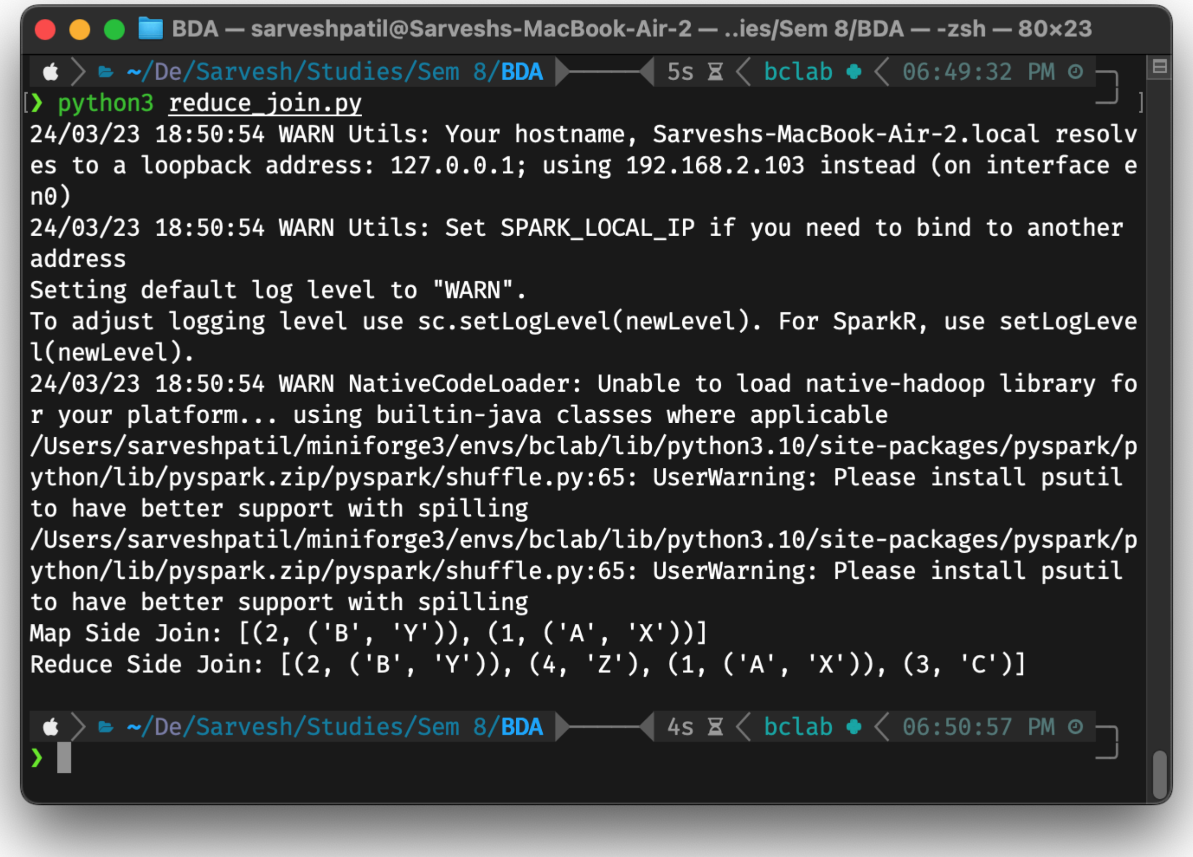
5. Joins - Map Side and Reduce Side

Code:

```
# Using Spark for Joins - Map Side and Reduce Side
from pyspark import SparkContext

# Initialize SparkContext
sc = SparkContext("local", "Joins")

# Create RDDs for left and right datasets
left_data = sc.parallelize([(1, "A"), (2, "B"), (3, "C")])
right_data = sc.parallelize([(1, "X"), (2, "Y"), (4, "Z")])

# Perform map-side join
map_join = left_data.join(right_data)

# Perform reduce-side join
reduce_join = left_data.union(right_data).reduceByKey(lambda x, y: (x, y))

# Print the results
print("Map Side Join:", map_join.collect())
print("Reduce Side Join:", reduce_join.collect())

# Stop SparkContext
sc.stop()
```

Name: Sarvesh Patil
Class: D20A
Roll No: 52

Output:


BDA — sarveshpatil@Sarveshs-MacBook-Air-2 — ..ies/Sem 8/BDA — -zsh — 80×23

```
 ~/De/Sarvesh/Studies/Sem 8/BDA ──── 5s ⧗ bclab ● 06:49:32 PM
❯ python3 reduce_join.py
24/03/23 18:50:54 WARN Utils: Your hostname, Sarveshs-MacBook-Air-2.local resolv
es to a loopback address: 127.0.0.1; using 192.168.2.103 instead (on interface e
n0)
24/03/23 18:50:54 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
24/03/23 18:50:54 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
/Users/sarveshpatil/miniforge3/envs/bclab/lib/python3.10/site-packages/pyspark/p
ython/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil
to have better support with spilling
/Users/sarveshpatil/miniforge3/envs/bclab/lib/python3.10/site-packages/pyspark/p
ython/lib/pyspark.zip/pyspark/shuffle.py:65: UserWarning: Please install psutil
to have better support with spilling
Map Side Join: [(2, ('B', 'Y')), (1, ('A', 'X'))]
Reduce Side Join: [(2, ('B', 'Y')), (4, 'Z'), (1, ('A', 'X')), (3, 'C')]

 ~/De/Sarvesh/Studies/Sem 8/BDA ──── 4s ⧗ bclab ● 06:50:57 PM
❯ 
```