Group-11
BDA Mini Project (CA-2)
Class : D20A

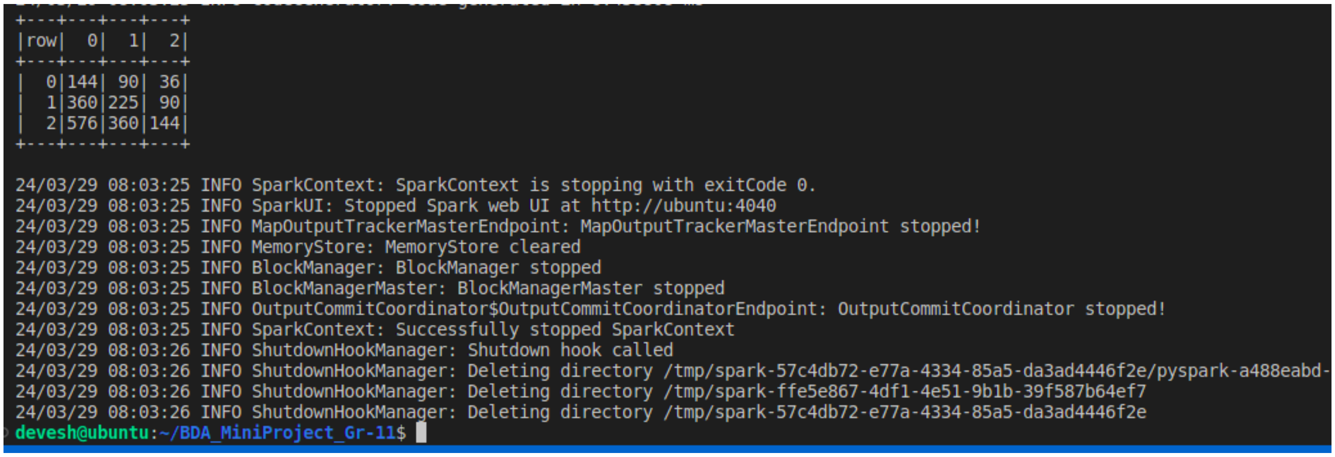
Roll No	Name
24	Soham Kayal
28	Mukul Kolpe
36	Sarvesh Limaye
56	Devesh Rahatekar

Matrix Multiplication :

```
from pyspark.sql import SparkSession
from pyspark.sql import functions as F

spark = SparkSession.builder \
    .appName("MatrixMultiplicationExample") \
    .getOrCreate()

matrix_A_data = [(1, 2, 3),
                 (4, 5, 6),
                 (7, 8, 9)]

matrix_B_data = [(9, 8, 7),
                 (6, 5, 4),
                 (3, 2, 1)]

matrix_A_df = spark.createDataFrame(matrix_A_data, ["A1", "A2", "A3"])
matrix_B_df = spark.createDataFrame(matrix_B_data, ["B1", "B2", "B3"])

result_matrix = matrix_A_df.crossJoin(matrix_B_df) \
    .withColumn("result", sum(F.col("A{0}".format(i + 1)) *
F.col("B{0}".format(j + 1))
                             for i in range(3) for j in range(3))) \
    .select("result") \
    .rdd.zipWithIndex() \
    .map(lambda x: (x[1] // 3, x[1] % 3, x[0][0])) \
    .toDF(["row", "col", "value"]) \
    .groupBy("row").pivot("col").agg(F.first("value"))
```

```
result_matrix.show()


spark.stop()
```

Output :

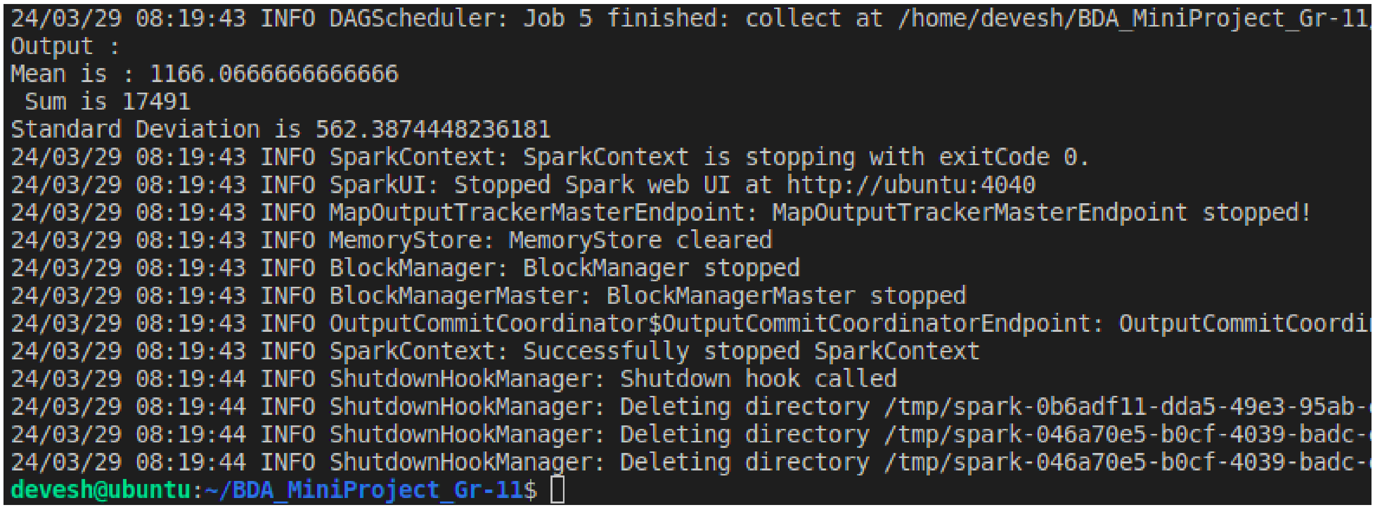


```
+---+---+---+---+
|row|  0|  1|  2|
+---+---+---+---+
|  0|144| 90| 36|
|  1|360|225| 90|
|  2|576|360|144|
+---+---+---+---+

24/03/29 08:03:25 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/03/29 08:03:25 INFO SparkUI: Stopped Spark web UI at http://ubuntu:4040
24/03/29 08:03:25 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/03/29 08:03:25 INFO MemoryStore: MemoryStore cleared
24/03/29 08:03:25 INFO BlockManager: BlockManager stopped
24/03/29 08:03:25 INFO BlockManagerMaster: BlockManagerMaster stopped
24/03/29 08:03:25 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/03/29 08:03:25 INFO SparkContext: Successfully stopped SparkContext
24/03/29 08:03:26 INFO ShutdownHookManager: Shutdown hook called
24/03/29 08:03:26 INFO ShutdownHookManager: Deleting directory /tmp/spark-57c4db72-e77a-4334-85a5-da3ad4446f2e/pyspark-a488eabd-
24/03/29 08:03:26 INFO ShutdownHookManager: Deleting directory /tmp/spark-ffe5e867-4df1-4e51-9b1b-39f587b64ef7
24/03/29 08:03:26 INFO ShutdownHookManager: Deleting directory /tmp/spark-57c4db72-e77a-4334-85a5-da3ad4446f2e
devesh@ubuntu:~/BDA_MiniProject_Gr-11$
```

Aggregations (Mean, Sum, Standard Deviation) :

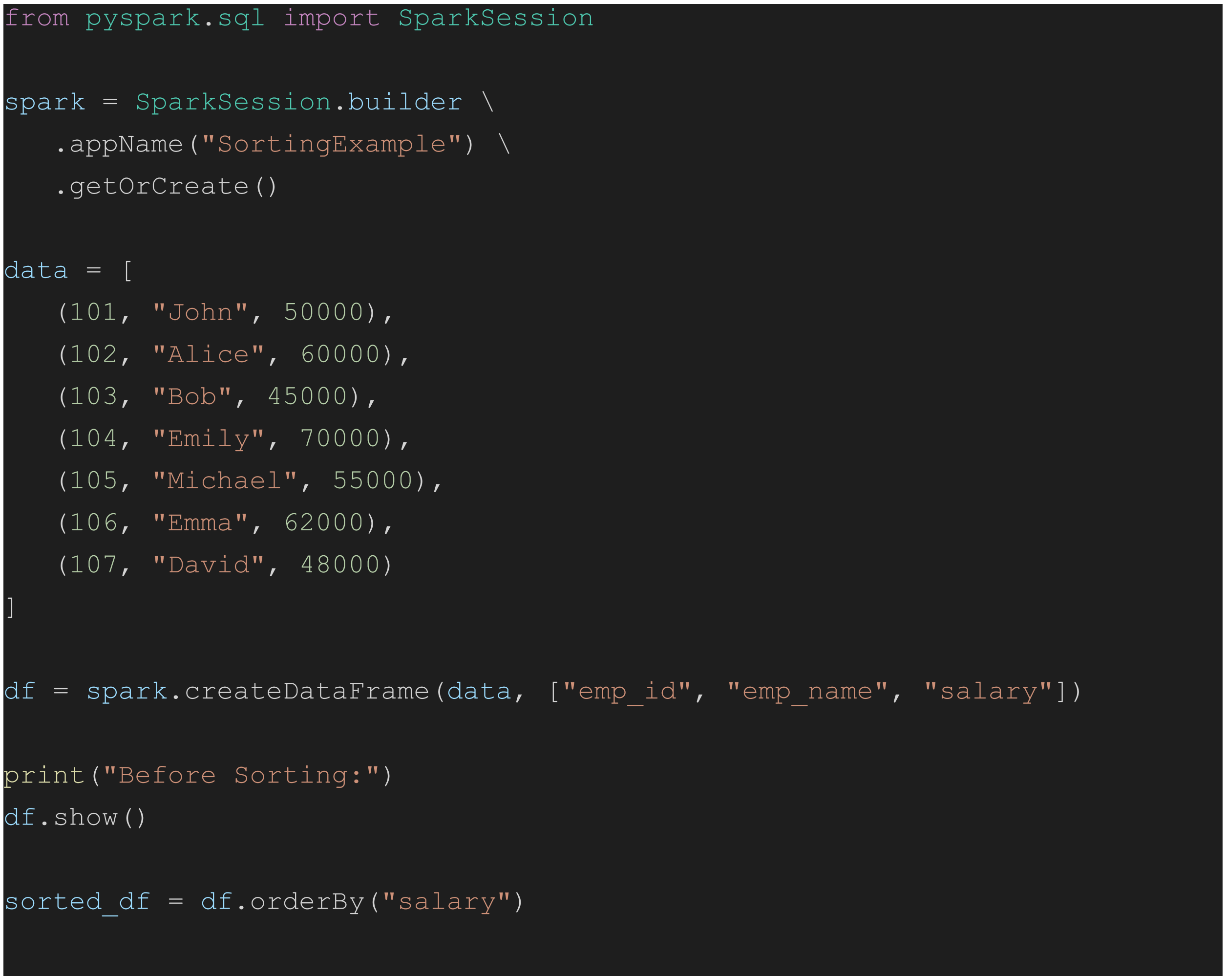
```
from pyspark.sql import SparkSession
from pyspark.sql import functions as F

spark = SparkSession.builder \
    .appName("Aggregations") \
    .getOrCreate()

data = [(225,), (346,), (518,), (687,), (823,), (944,), (1056,),
        (1223,), (1375,), (1442,), (1565,), (1678,), (1790,), (1876,),
        (1943,)]

df = spark.createDataFrame(data, ["measurement"])

mean = df.agg(F.mean("measurement")).collect()[0][0]
sum_val = df.agg(F.sum("measurement")).collect()[0][0]
std_dev = df.agg(F.stddev("measurement")).collect()[0][0]
print("Output : ")
print(f"Mean is : {mean}\n Sum is {sum_val} \nStandard Deviation is
{std_dev}" )
spark.stop()
```

Output :



```
24/03/29 08:19:43 INFO DAGScheduler: Job 5 finished: collect at /home/devesh/BDA_MiniProject_Gr-11/
Output :
Mean is : 1166.0666666666666
 Sum is 17491
Standard Deviation is 562.3874448236181
24/03/29 08:19:43 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/03/29 08:19:43 INFO SparkUI: Stopped Spark web UI at http://ubuntu:4040
24/03/29 08:19:43 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/03/29 08:19:43 INFO MemoryStore: MemoryStore cleared
24/03/29 08:19:43 INFO BlockManager: BlockManager stopped
24/03/29 08:19:43 INFO BlockManagerMaster: BlockManagerMaster stopped
24/03/29 08:19:43 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordi
24/03/29 08:19:43 INFO SparkContext: Successfully stopped SparkContext
24/03/29 08:19:44 INFO ShutdownHookManager: Shutdown hook called
24/03/29 08:19:44 INFO ShutdownHookManager: Deleting directory /tmp/spark-0b6adf11-dda5-49e3-95ab-
24/03/29 08:19:44 INFO ShutdownHookManager: Deleting directory /tmp/spark-046a70e5-b0cf-4039-badc-
24/03/29 08:19:44 INFO ShutdownHookManager: Deleting directory /tmp/spark-046a70e5-b0cf-4039-badc-
devesh@ubuntu:~/BDA_MiniProject_Gr-11$
```

Sorting :

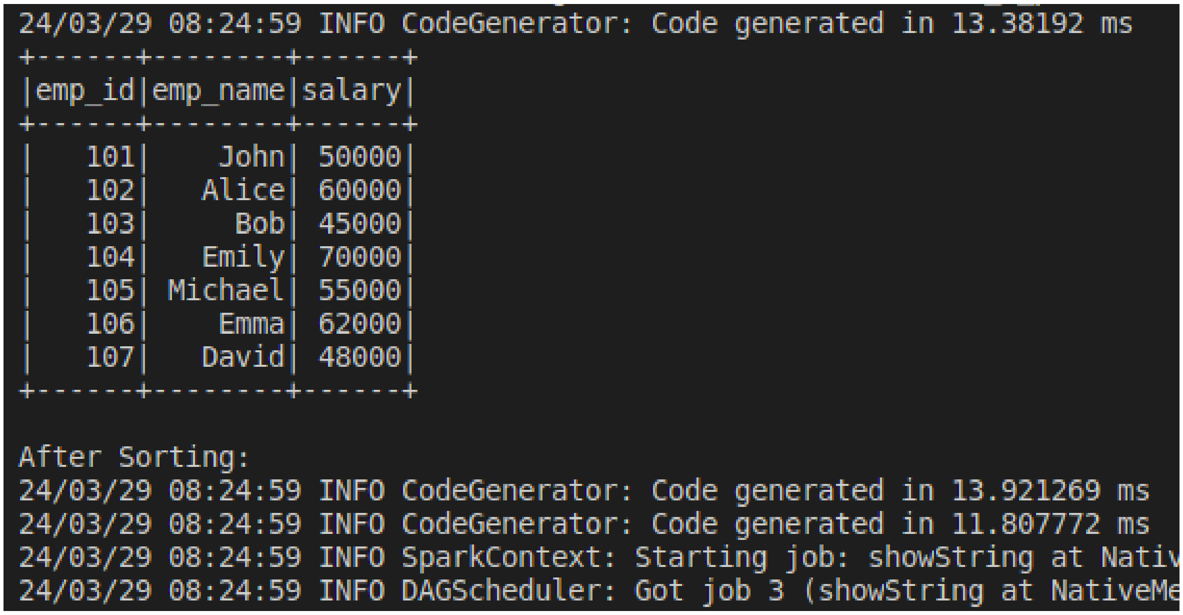


```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
   .appName("SortingExample") \
   .getOrCreate()

data = [
   (101, "John", 50000),
   (102, "Alice", 60000),
   (103, "Bob", 45000),
   (104, "Emily", 70000),
   (105, "Michael", 55000),
   (106, "Emma", 62000),
   (107, "David", 48000)
]

df = spark.createDataFrame(data, ["emp_id", "emp_name", "salary"])

print("Before Sorting:")
df.show()

sorted_df = df.orderBy("salary")
```

```
print("After Sorting:")
sorted_df.show()

spark.stop()
```

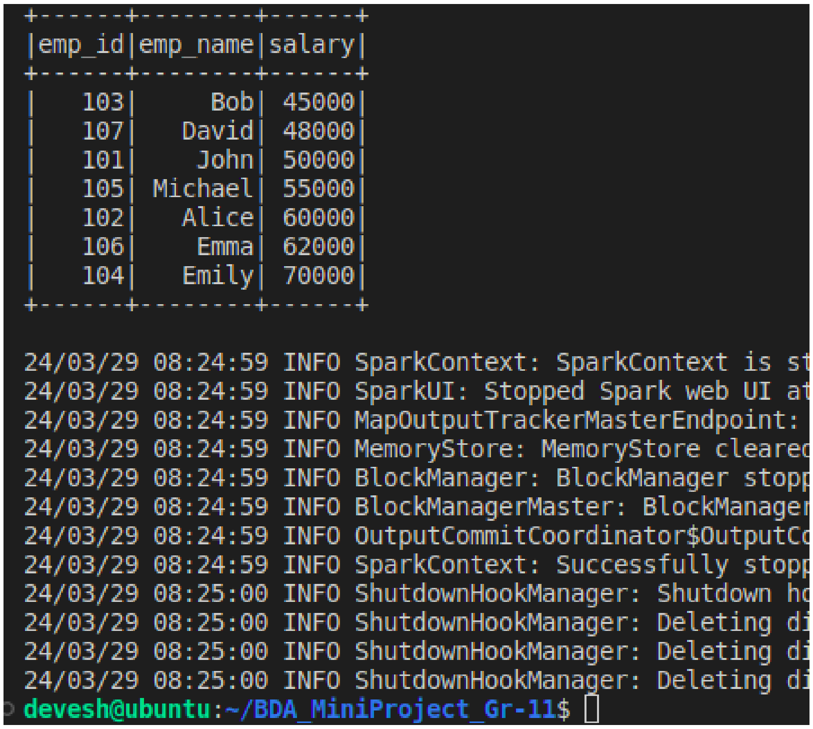
Output :

Before Sorting :



```
24/03/29 08:24:59 INFO CodeGenerator: Code generated in 13.38192 ms
+------+--------+------+
|emp_id|emp_name|salary|
+------+--------+------+
|   101|    John| 50000|
|   102|   Alice| 60000|
|   103|     Bob| 45000|
|   104|   Emily| 70000|
|   105| Michael| 55000|
|   106|    Emma| 62000|
|   107|   David| 48000|
+------+--------+------+

After Sorting:
24/03/29 08:24:59 INFO CodeGenerator: Code generated in 13.921269 ms
24/03/29 08:24:59 INFO CodeGenerator: Code generated in 11.807772 ms
24/03/29 08:24:59 INFO SparkContext: Starting job: showString at Nativ
24/03/29 08:24:59 INFO DAGScheduler: Got job 3 (showString at NativeMe
```

After Sorting :



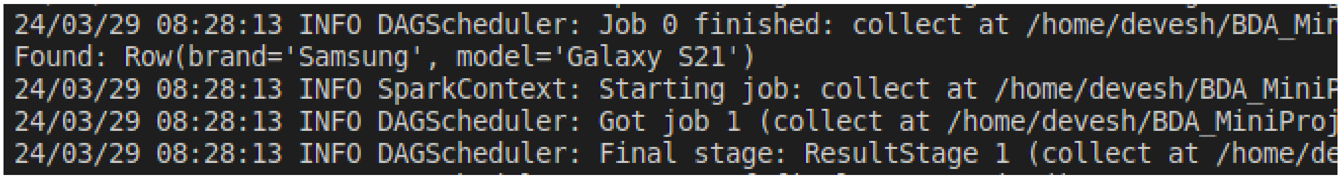
```
+------+--------+------+
|emp_id|emp_name|salary|
+------+--------+------+
|   103|     Bob| 45000|
|   107|   David| 48000|
|   101|    John| 50000|
|   105| Michael| 55000|
|   102|   Alice| 60000|
|   106|    Emma| 62000|
|   104|   Emily| 70000|
+------+--------+------+

24/03/29 08:24:59 INFO SparkContext: SparkContext is st
24/03/29 08:24:59 INFO SparkUI: Stopped Spark web UI at
24/03/29 08:24:59 INFO MapOutputTrackerMasterEndpoint:
24/03/29 08:24:59 INFO MemoryStore: MemoryStore cleared
24/03/29 08:24:59 INFO BlockManager: BlockManager stopp
24/03/29 08:24:59 INFO BlockManagerMaster: BlockManager
24/03/29 08:24:59 INFO OutputCommitCoordinator$OutputCo
24/03/29 08:24:59 INFO SparkContext: Successfully stopp
24/03/29 08:25:00 INFO ShutdownHookManager: Shutdown ho
24/03/29 08:25:00 INFO ShutdownHookManager: Deleting di
24/03/29 08:25:00 INFO ShutdownHookManager: Deleting di
24/03/29 08:25:00 INFO ShutdownHookManager: Deleting di
devesh@ubuntu:~/BDA_MiniProject_Gr-11$
```

Searching a data Element :

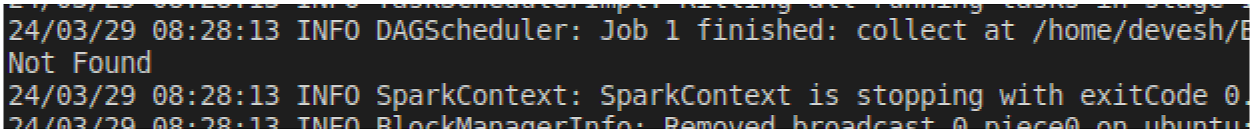
Code :

```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
   .appName("Search") \
   .getOrCreate()

data = [("Apple", "iPhone 13"), ("Samsung", "Galaxy S21"), ("Google",
"Pixel 6"),
       ("Huawei", "Mate 40"), ("Xiaomi", "Mi 11"), ("OnePlus", "9 Pro")]

df = spark.createDataFrame(data, ["brand", "model"])

search_result = df.filter(df.brand == "Samsung").collect()
if search_result:
   print("Found:", search_result[0])
else:
   print("Not Found")

search_result = df.filter(df.brand == "Sony").collect()
if search_result:
   print("Found:", search_result[0])
else:
   print("Not Found")

spark.stop()
```

Output :



```
24/03/29 08:28:13 INFO DAGScheduler: Job 0 finished: collect at /home/devesh/BDA_Min
Found: Row(brand='Samsung', model='Galaxy S21')
24/03/29 08:28:13 INFO SparkContext: Starting job: collect at /home/devesh/BDA_MiniP
24/03/29 08:28:13 INFO DAGScheduler: Got job 1 (collect at /home/devesh/BDA_MiniProj
24/03/29 08:28:13 INFO DAGScheduler: Final stage: ResultStage 1 (collect at /home/de
```

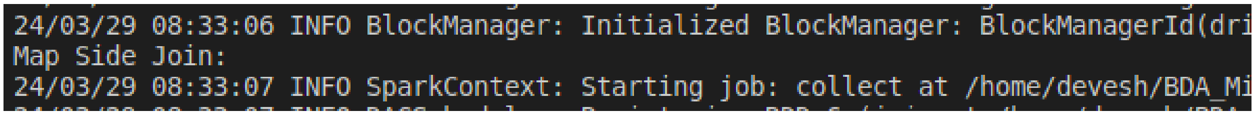


```
24/03/29 08:28:13 INFO DAGScheduler: Job 1 finished: collect at /home/devesh/E
Not Found
24/03/29 08:28:13 INFO SparkContext: SparkContext is stopping with exitCode 0.
24/03/29 08:28:13 INFO BlockManagerInfo: Removed broadcast 0 piece0 on ubuntu-
```

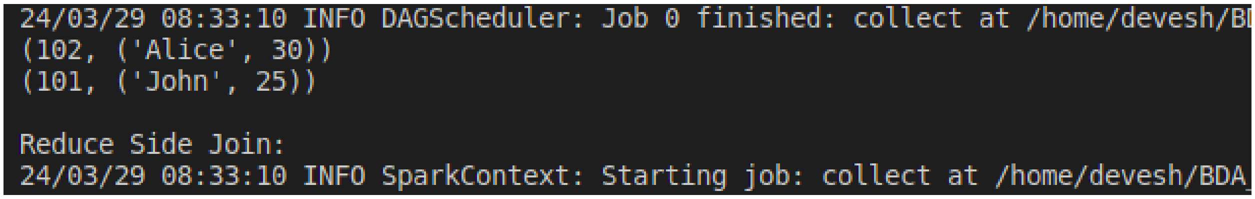
Joins - Map side and Reduce side :
Code :

```python
from pyspark import SparkContext

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

left_data = sc.parallelize([(101, "John"), (102, "Alice"), (103, "Bob"), (104, "Emily")])
right_data = sc.parallelize([(101, 25), (102, 30), (105, 28)])

map_join = left_data.join(right_data)

reduce_join = left_data.union(right_data).reduceByKey(lambda x, y: (x, y))

print("Map Side Join:")
for result in map_join.collect():
    print(result)

print("\nReduce Side Join:")
for result in reduce_join.collect():
    print(result)

sc.stop()
```

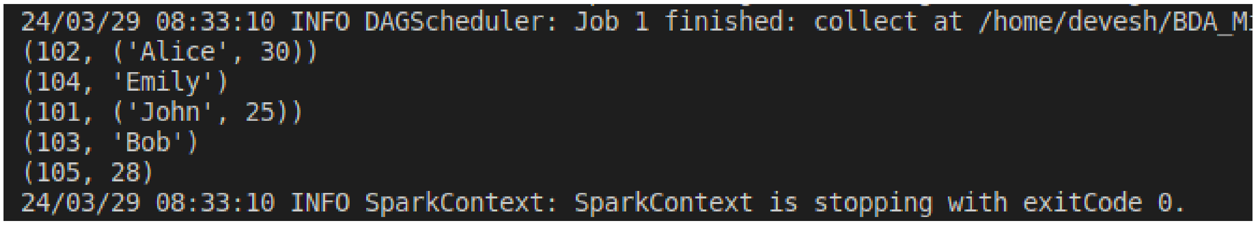
Output :



```
24/03/29 08:33:06 INFO BlockManager: Initialized BlockManager: BlockManagerId(dri
Map Side Join:
24/03/29 08:33:07 INFO SparkContext: Starting job: collect at /home/devesh/BDA_Mi
```



```
24/03/29 08:33:10 INFO DAGScheduler: Job 0 finished: collect at /home/devesh/BD
(102, ('Alice', 30))
(101, ('John', 25))

Reduce Side Join:
24/03/29 08:33:10 INFO SparkContext: Starting job: collect at /home/devesh/BDA_
```



```
24/03/29 08:33:10 INFO DAGScheduler: Job 1 finished: collect at /home/devesh/BDA_Mi
(102, ('Alice', 30))
(104, 'Emily')
(101, ('John', 25))
(103, 'Bob')
(105, 28)
24/03/29 08:33:10 INFO SparkContext: SparkContext is stopping with exitCode 0.
```