

Experiment 5

Aim: Data Analysis with Spark: Use Apache Spark to analyze large datasets by loading them into Spark RDDs (Resilient Distributed Datasets) and performing operations such as filtering, mapping, and aggregation.

Directory:

```
C:\Users\rawat\Documents\7 SEMESTER\Big Data Analytics\Lab\Experiment 5>spark-submit Spark_Data_Analysis.py
24/10/18 15:28:37 INFO SparkContext: Running Spark version 3.5.2
24/10/18 15:28:37 INFO SparkContext: OS info Windows 11, 10.0, amd64
24/10/18 15:28:37 INFO SparkContext: Java version 1.8.0_421
24/10/18 15:28:37 INFO ResourceUtils: =====
24/10/18 15:28:37 INFO ResourceUtils: No custom resources configured for spark.driver.
24/10/18 15:28:37 INFO ResourceUtils: =====
24/10/18 15:28:37 INFO SparkContext: Submitted application: Iris Data Analysis
24/10/18 15:28:37 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
24/10/18 15:28:37 INFO ResourceProfile: Limiting resource is cpu
24/10/18 15:28:37 INFO ResourceProfileManager: Added ResourceProfile id: 0
24/10/18 15:28:38 INFO SecurityManager: Changing view acls to: rawat
24/10/18 15:28:38 INFO SecurityManager: Changing modify acls to: rawat
24/10/18 15:28:38 INFO SecurityManager: Changing view acls groups to:
24/10/18 15:28:38 INFO SecurityManager: Changing modify acls groups to:
24/10/18 15:28:38 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: rawat; groups with view permissions: EMPTY; users with modify permissions: rawat; groups with modify permissions: EMPTY
24/10/18 15:28:38 INFO Utils: Successfully started service 'sparkDriver' on port 51851.
```

Fig 1: Running Spark in Directory

Displaying Dataset:

First few rows of the encoded Iris dataset:

```
24/10/18 15:28:42 INFO SparkContext: Starting job: runJob at PythonRDD.scala:181
24/10/18 15:28:42 INFO DAGScheduler: Got job 1 (runJob at PythonRDD.scala:181) with 1 output partitions
24/10/18 15:28:42 INFO DAGScheduler: Final stage: ResultStage 1 (runJob at PythonRDD.scala:181)
24/10/18 15:28:42 INFO DAGScheduler: Parents of final stage: List()
24/10/18 15:28:42 INFO DAGScheduler: Missing parents: List()
24/10/18 15:28:42 INFO DAGScheduler: Submitting ResultStage 1 (PythonRDD[3] at RDD at PythonRDD.scala:53)
```

Fig 2: Displaying First Rows Messages

```
24/10/18 15:28:42 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
24/10/18 15:28:42 INFO DAGScheduler: Job 1 finished: runJob at PythonRDD.scala:181, took 0.715150 s
[[['5.1', '3.5', '1.4', '0.2', '0'], ['4.9', '3.0', '1.4', '0.2', '0'], ['4.7', '3.2', '1.3', '0.2', '0'], ['4.6', '3.1', '1.5', '0.2', '0'], ['5.0', '3.6', '1.4', '0.2', '0']]]
```

Fig 3: Displaying First Rows

Filtering and Aggregation in Dataset:

```
===== Average Sepal Length =====
24/10/18 15:28:42 INFO SparkContext: Starting job: collect at C:\Users\rawat\Documents\7 SEMESTER\Big Data Analytics\Lab\Experiment 5\Spark_Data_Analysis.py:38
24/10/18 15:28:42 INFO DAGScheduler: Registering RDD 5 (aggregateByKey at C:\Users\rawat\Documents\7 SEMESTER\Big Data Analytics\Lab\Experiment 5\Spark_Data_Analysis.py:27) as input to shuffle 0
```

Fig 4: Displaying Aggregation Messages

```
24/10/18 15:28:45 INFO
7057 s
0: 5.005999999999999
1: 5.935999999999999
2: 6.587999999999998
```

Fig 5: Aggregation by Sepal Length

Mapping Species and Aggregation in Dataset:

```
===== Average Values for All Columns by Species =====
```

Fig 6: Average Values for Species Column Message

```
2245 s
0 - Sepal Length: 5.005999999999999, Sepal Width: 3.4180000000000006, Petal Length: 1.464, Petal Width: 0.24399999999999999
1 - Sepal Length: 5.935999999999999, Sepal Width: 2.77, Petal Length: 4.26, Petal Width: 1.3259999999999998
2 - Sepal Length: 6.587999999999998, Sepal Width: 2.9739999999999998, Petal Length: 5.552, Petal Width: 2.026
```

Fig 7: Average Values for Species Column

Counting Occurrence of Species in Dataset:

```
===== Count of Instances per Species =====
```

Fig 8: Count of Column Message

```
18245 s
0: 50
1: 50
2: 50
24/10/18 15:28:51 INFO SparkContext: SparkContext is stopping with exitCode 0.
```

Fig 9: Count of Column Unique Values in Species