Assignment-11

Name-Ajay Chaudhary
Batch-Data Engineering(Batch-1)

SparkSession

Creating a SparkSession instance would be the first statement you would write to the program with RDD, DataFrame and Dataset. SparkSession will be created using SparkSession.builder() builder pattern.

```
#Create SparkSession
import org.apache.spark.sql.SparkSession
val spark:SparkSession = SparkSession.builder()
      .master("local[1]")
      .appName("PySpark_example")
      .getOrCreate()
```

RDD creation(Resilient Distributed Datasets)

Two ways to create RDD-

- sparkContext.parallelize()
- sparkContext.textFile()

sparkContext.parallelize()

```
[14]:
# Import SparkSession
from pyspark.sql import SparkSession


# Create SparkSession
spark = SparkSession.builder \
      .master("local[1]") \
      .appName("SparkByExamples.com") \
      .getOrCreate()
dataList = [("Java", 20000), ("Python", 100000), ("Scala", 3000)]
rdd=spark.sparkContext.parallelize(dataList)
result=rdd.collect()
print("RDD Contents:",result)
```

```
RDD Contents: [('Java', 20000), ('Python', 100000), ('Scala', 3000)]
```

sparkContext.textFile()

[16]:

```
rdd2 = spark.sparkContext.textFile("/Users/ajaychaudhary/Downloads/test.txt")
print(rdd2.collect())
```

```
['one 1', 'eleven 11']
```

Spark.read.csv

```
import pyspark
```

[3]:

```
from pyspark.sql import SparkSession
```

[4]:

```
spark=SparkSession.builder.appName("Pyspark").getOrCreate()
```

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(ne
wLevel).
24/02/05 15:07:14 WARN NativeCodeLoader: Unable to load native-hadoop library for yo
ur platform... using builtin-java classes where applicable
```

[5]:

```
spark
```

[5]:

SparkSession - in-memory

SparkContext

Spark UI

Version	`v3.5.0`
Master	`local[*]`
AppName	`Pyspark`

[6]:

```
df=spark.read.csv("/Users/ajaychaudhary/Downloads/Marks_data.csv")
df
```

[6]:

```
DataFrame[_c0: string, _c1: string, _c2: string, _c3: string]
```

```
df.show()
```

```
+-----+--------+--------+----+
| _c0|     _c1|     _c2|_c3|
+-----+--------+--------+----+
|Name|M1 Score|M2 Score|age|
|Alex|      62|      80| 20|
|Brad|      45|      56| 19|
|Joey|      85|      98| 21|
|NULL|      54|      79| 20|
|abhi|    NULL|    NULL| 20|
+-----+--------+--------+----+
```

[14]:

I have run the same commands for creating a rdd on databricks -