```
ip-10-0-41-79 login: vibhashekarnsgmail
Password:
Last login: Sun Apr 25 18:13:37 on pts/3187
[vibhashekarnsgmail@ip-10-0-41-79 ~]$ spark-shell
Setting default log level to "ERROR".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
21/04/25 18:14:34 WARN cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Attempted to request executors before the AM has registered!
21/04/25 18:14:34 WARN lineage.LineageWriter: Lineage directory /var/log/spark/lineage doesn't exist or is not writable. Lineage for this application will be disabled.
Spark context available as 'sc' (master = yarn, app id = application_1608530820093_34367).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.0-cdh6.3.2
      /_/

Using Scala version 2.11.12 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_144)
Type in expressions to have them evaluated.
Type :help for more information.

scala> import org.apache.spark.sql.SQLContext
import org.apache.spark.sql.SQLContext

scala>

scala> val sqlContext = new SQLContext(sc)
warning: there was one deprecation warning; re-run with -deprecation for details
sqlContext: org.apache.spark.sql.SQLContext = org.apache.spark.sql.SQLContext@72f5fb36

scala>

scala> val df = sqlContext.read.format("com.databricks.spark.csv").option("header", "true").option("inferSchema", "true").load("/user/vibhashekarnsgmail/Bank/Bank_Campaign_Data.csv")
21/04/25 18:14:51 WARN lineage.LineageWriter: Lineage directory /var/log/spark/lineage doesn't exist or is not writable. Lineage for this application will be disabled.
df: org.apache.spark.sql.DataFrame = [age: int, job: string ... 15 more fields]

scala>

scala> df.printSchema
root
 |-- age: integer (nullable = true)
 |-- job: string (nullable = true)
```

```
|-- marital: string (nullable = true)
|-- education: string (nullable = true)
|-- default: string (nullable = true)
|-- balance: integer (nullable = true)
|-- housing: string (nullable = true)
|-- loan: string (nullable = true)
|-- contact: string (nullable = true)
|-- day: integer (nullable = true)
|-- month: string (nullable = true)
|-- duration: integer (nullable = true)
|-- campaign: integer (nullable = true)
|-- pdays: integer (nullable = true)
|-- previous: integer (nullable = true)
|-- poutcome: string (nullable = true)
|-- y: string (nullable = true)


scala>

scala> df.show
+---+-----------+-------+---------+-------+-------+-------+----+-------+---+-----+--------+--------+-----+--------+--------+---+
|age|        job|marital|education|default|balance|housing|loan|contact|day|month|duration|campaign|pdays|previous|poutcome|  y|
+---+-----------+-------+---------+-------+-------+-------+----+-------+---+-----+--------+--------+-----+--------+--------+---+
| 58|  management|married| tertiary|     no|   2143|    yes|  no|unknown|  5|  may|     261|       1|   -1|       0| unknown| no|
| 44|  technician| single|secondary|     no|     29|    yes|  no|unknown|  5|  may|     151|       1|   -1|       0| unknown| no|
| 33|entrepreneur|married|secondary|     no|      2|    yes| yes|unknown|  5|  may|      76|       1|   -1|       0| unknown| no|
| 47| blue-collar|married|  unknown|     no|   1506|    yes|  no|unknown|  5|  may|      92|       1|   -1|       0| unknown| no|
| 33|     unknown| single|  unknown|     no|      1|     no|  no|unknown|  5|  may|     198|       1|   -1|       0| unknown| no|
| 35|  management|married| tertiary|     no|    231|    yes|  no|unknown|  5|  may|     139|       1|   -1|       0| unknown| no|
| 28|  management| single| tertiary|     no|    447|    yes| yes|unknown|  5|  may|     217|       1|   -1|       0| unknown| no|
| 42|entrepreneur|divorced| tertiary|    yes|      2|    yes|  no|unknown|  5|  may|     380|       1|   -1|       0| unknown| no|
| 58|     retired|married|  primary|     no|    121|    yes|  no|unknown|  5|  may|      50|       1|   -1|       0| unknown| no|
| 43|  technician| single|secondary|     no|    593|    yes|  no|unknown|  5|  may|      55|       1|   -1|       0| unknown| no|
| 41|      admin.|divorced|secondary|     no|    270|    yes|  no|unknown|  5|  may|     222|       1|   -1|       0| unknown| no|
| 29|      admin.| single|secondary|     no|    390|    yes|  no|unknown|  5|  may|     137|       1|   -1|       0| unknown| no|
| 53|  technician|married|secondary|     no|      6|    yes|  no|unknown|  5|  may|     517|       1|   -1|       0| unknown| no|
| 58|  technician|married|  unknown|     no|     71|    yes|  no|unknown|  5|  may|      71|       1|   -1|       0| unknown| no|
| 57|    services|married|secondary|     no|    162|    yes|  no|unknown|  5|  may|     174|       1|   -1|       0| unknown| no|
| 51|     retired|married|  primary|     no|    229|    yes|  no|unknown|  5|  may|     353|       1|   -1|       0| unknown| no|
| 45|      admin.| single|  unknown|     no|     13|    yes|  no|unknown|  5|  may|      98|       1|   -1|       0| unknown| no|
| 57| blue-collar|married|  primary|     no|     52|    yes|  no|unknown|  5|  may|      38|       1|   -1|       0| unknown| no|
| 60|     retired|married|  primary|     no|     60|    yes|  no|unknown|  5|  may|     219|       1|   -1|       0| unknown| no|
```

Practice Labs | vibhashekarnsgmail@ip-10-0-41 | slbdh-webconsole.corestack.io

```
| 45|      admin.|  single|  unknown|     no|     13|    yes|  no|unknown|  5|  may|      98|       1|   -1|       0| unknown| no|
| 57| blue-collar| married|  primary|     no|     52|    yes|  no|unknown|  5|  may|      38|       1|   -1|       0| unknown| no|
| 60|     retired| married|  primary|     no|     60|    yes|  no|unknown|  5|  may|     219|       1|   -1|       0| unknown| no|
| 33|    services| married|secondary|     no|      0|    yes|  no|unknown|  5|  may|      54|       1|   -1|       0| unknown| no|
+---+------------+--------+---------+-------+-------+-------+----+-------+---+-----+--------+--------+-----+--------+--------+---+
only showing top 20 rows


scala>

scala> df.select("age", "job", "marital", "education", "default", "balance", "housing", "loan", "contact", "day", "month", "duration", "campaign", "pdays", "previous", "poutcome", "y")
.show
+---+------------+--------+---------+-------+-------+-------+----+-------+---+-----+--------+--------+-----+--------+--------+---+
|age|         job| marital|education|default|balance|housing|loan|contact|day|month|duration|campaign|pdays|previous|poutcome|  y|
+---+------------+--------+---------+-------+-------+-------+----+-------+---+-----+--------+--------+-----+--------+--------+---+
| 58|  management| married| tertiary|     no|   2143|    yes|  no|unknown|  5|  may|     261|       1|   -1|       0| unknown| no|
| 44|  technician|  single|secondary|     no|     29|    yes|  no|unknown|  5|  may|     151|       1|   -1|       0| unknown| no|
| 33|entrepreneur| married|secondary|     no|      2|    yes| yes|unknown|  5|  may|      76|       1|   -1|       0| unknown| no|
| 47| blue-collar| married|  unknown|     no|   1506|    yes|  no|unknown|  5|  may|      92|       1|   -1|       0| unknown| no|
| 33|     unknown|  single|  unknown|     no|      1|     no|  no|unknown|  5|  may|     198|       1|   -1|       0| unknown| no|
| 35|  management| married| tertiary|     no|    231|    yes|  no|unknown|  5|  may|     139|       1|   -1|       0| unknown| no|
| 28|  management|  single| tertiary|     no|    447|    yes| yes|unknown|  5|  may|     217|       1|   -1|       0| unknown| no|
| 42|entrepreneur|divorced| tertiary|    yes|      2|    yes|  no|unknown|  5|  may|     380|       1|   -1|       0| unknown| no|
| 58|     retired| married|  primary|     no|    121|    yes|  no|unknown|  5|  may|      50|       1|   -1|       0| unknown| no|
| 43|  technician|  single|secondary|     no|    593|    yes|  no|unknown|  5|  may|      55|       1|   -1|       0| unknown| no|
| 41|      admin.|divorced|secondary|     no|    270|    yes|  no|unknown|  5|  may|     222|       1|   -1|       0| unknown| no|
| 29|      admin.|  single|secondary|     no|    390|    yes|  no|unknown|  5|  may|     137|       1|   -1|       0| unknown| no|
| 53|  technician| married|secondary|     no|      6|    yes|  no|unknown|  5|  may|     517|       1|   -1|       0| unknown| no|
| 58|  technician| married|  unknown|     no|     71|    yes|  no|unknown|  5|  may|      71|       1|   -1|       0| unknown| no|
| 57|    services| married|secondary|     no|    162|    yes|  no|unknown|  5|  may|     174|       1|   -1|       0| unknown| no|
| 51|     retired| married|  primary|     no|    229|    yes|  no|unknown|  5|  may|     353|       1|   -1|       0| unknown| no|
| 45|      admin.|  single|  unknown|     no|     13|    yes|  no|unknown|  5|  may|      98|       1|   -1|       0| unknown| no|
| 57| blue-collar| married|  primary|     no|     52|    yes|  no|unknown|  5|  may|      38|       1|   -1|       0| unknown| no|
| 60|     retired| married|  primary|     no|     60|    yes|  no|unknown|  5|  may|     219|       1|   -1|       0| unknown| no|
| 33|    services| married|secondary|     no|      0|    yes|  no|unknown|  5|  may|      54|       1|   -1|       0| unknown| no|
+---+------------+--------+---------+-------+-------+-------+----+-------+---+-----+--------+--------+-----+--------+--------+---+
only showing top 20 rows


scala>

scala> df.registerTempTable("spark_tbl_Marketing_Analysis")
```

Type here to search | ENG | 23:51 | 25-04-2021

Practice Labs | vibhashekarnsgmail@ip-10-0-41 | slbdh-webconsole.corestack.io

```
warning: there was one deprecation warning; re-run with -deprecation for details

scala>

scala> sqlContext.sql("""select * from spark_tbl_Marketing_Analysis limit 10""").show();
+---+-----------+--------+---------+-------+-------+-------+----+-------+---+-----+--------+--------+-----+--------+--------+---+
|age|        job| marital|education|default|balance|housing|loan|contact|day|month|duration|campaign|pdays|previous|poutcome|  y|
+---+-----------+--------+---------+-------+-------+-------+----+-------+---+-----+--------+--------+-----+--------+--------+---+
| 58|  management| married| tertiary|     no|   2143|    yes|  no|unknown|  5|  may|     261|       1|   -1|       0| unknown| no|
| 44|  technician|  single|secondary|     no|     29|    yes|  no|unknown|  5|  may|     151|       1|   -1|       0| unknown| no|
| 33|entrepreneur| married|secondary|     no|      2|    yes| yes|unknown|  5|  may|      76|       1|   -1|       0| unknown| no|
| 47| blue-collar| married|  unknown|     no|   1506|    yes|  no|unknown|  5|  may|      92|       1|   -1|       0| unknown| no|
| 33|     unknown|  single|  unknown|     no|      1|     no|  no|unknown|  5|  may|     198|       1|   -1|       0| unknown| no|
| 35|  management| married| tertiary|     no|    231|    yes|  no|unknown|  5|  may|     139|       1|   -1|       0| unknown| no|
| 28|  management|  single| tertiary|     no|    447|    yes| yes|unknown|  5|  may|     217|       1|   -1|       0| unknown| no|
| 42|entrepreneur|divorced| tertiary|    yes|      2|    yes|  no|unknown|  5|  may|     380|       1|   -1|       0| unknown| no|
| 58|     retired| married|  primary|     no|    121|    yes|  no|unknown|  5|  may|      50|       1|   -1|       0| unknown| no|
| 43|  technician|  single|secondary|     no|    593|    yes|  no|unknown|  5|  may|      55|       1|   -1|       0| unknown| no|
+---+-----------+--------+---------+-------+-------+-------+----+-------+---+-----+--------+--------+-----+--------+--------+---+


scala>

scala> val tot_count = df.count()
tot_count: Long = 45211

scala>

scala> val reg_success = df.filter("y = 'yes'").count()
reg_success: Long = 5289

scala>

scala> val success_rate = tot_count.toFloat / reg_success
success_rate: Float = 8.548119

scala>

scala> val reg_fail = df.filter("y = 'no'").count()
reg_fail: Long = 39922

scala>
```

Type here to search | ENG | 23:51 | 25-04-2021

```
scala> val fail_rate = tot_count.toFloat / reg_fail
fail_rate: Float = 1.1324834

scala>

scala> val age_stat = sqlContext.sql("select max(age), min(age), avg(age) from spark_tbl_Marketing_Analysis")
age_stat: org.apache.spark.sql.DataFrame = [max(age): int, min(age): int ... 1 more field]

scala> age_stat.show()
+--------+--------+------------------+
|max(age)|min(age)|         avg(age)|
+--------+--------+------------------+
|      95|      18|40.93621021432837|
+--------+--------+------------------+


scala>

scala> val cust_qual = sqlContext.sql("select percentile_approx(balance, .5), avg(balance) from spark_tbl_Marketing_Analysis")
cust_qual: org.apache.spark.sql.DataFrame = [percentile_approx(balance, CAST(0.5 AS DOUBLE), 10000): int, avg(balance): double]

scala> cust_qual.show()
+-----------------------------------------------------+------------------+
|percentile_approx(balance, CAST(0.5 AS DOUBLE), 10000)|      avg(balance)|
+-----------------------------------------------------+------------------+
|                                                  448|1362.2720576850766|
+-----------------------------------------------------+------------------+


scala>

scala> val age_mktg = sqlContext.sql("select y, avg(age) from spark_tbl_Marketing_Analysis group by y")
age_mktg: org.apache.spark.sql.DataFrame = [y: string, avg(age): double]

scala> age_mktg.show()
+---+------------------+
|  y|          avg(age)|
+---+------------------+
| no| 40.83898602274435|
|yes|41.670069956513515|
+---+------------------+
```

Practice Labs | vibhashekarnsgmail@ip-10-0-41 | slbdh-webconsole.corestack.io

```
scala>

scala> val marital_mktg = sqlContext.sql("select marital, y, count(marital) from spark_tbl_Marketing_Analysis group by marital, y order by y")
marital_mktg: org.apache.spark.sql.DataFrame = [marital: string, y: string ... 1 more field]

scala> marital_mktg.show()
+--------+---+--------------+
| marital|  y|count(marital)|
+--------+---+--------------+
|divorced| no|          4585|
|  single| no|         10878|
| married| no|         24459|
|divorced|yes|           622|
|  single|yes|          1912|
| married|yes|          2755|
+--------+---+--------------+


scala>

scala> val age_mktg_subs = sqlContext.sql("select marital, y, count(marital),avg(age) from spark_tbl_Marketing_Analysis group by marital, y order by y")
age_mktg_subs: org.apache.spark.sql.DataFrame = [marital: string, y: string ... 2 more fields]

scala> age_mktg_subs.show()
+--------+---+--------------+------------------+
| marital|  y|count(marital)|          avg(age)|
+--------+---+--------------+------------------+
|divorced| no|          4585| 45.31297709923664|
|  single| no|         10878| 33.96258503401361|
| married| no|         24459| 43.05854695613067|
|  single|yes|          1912| 32.22907949790795|
|divorced|yes|           622|49.247588424437296|
| married|yes|          2755| 46.51143375680581|
+--------+---+--------------+------------------+


scala>

scala> val age_camp = sqlContext.sql("select y, age, count(y) from spark_tbl_Marketing_Analysis group by y, age order by y, age")
age_camp: org.apache.spark.sql.DataFrame = [y: string, age: int ... 1 more field]

scala> age_camp.show()
```

Type here to search | ENG | 23:52 | 25-04-2021

Practice Labs | vibhashekarnsgmail@ip-10-0-41 | slbdh-webconsole.corestack.io

```
|  single| no|         10878| 33.96258503401361|
| married| no|         24459| 43.05854695613067|
|  single|yes|          1912| 32.22907949790795|
|divorced|yes|           622|49.247588424437296|
| married|yes|          2755| 46.51143375680581|
+--------+---+--------------+------------------+


scala>

scala> val age_camp = sqlContext.sql("select y, age, count(y) from spark_tbl_Marketing_Analysis group by y, age order by y, age")
age_camp: org.apache.spark.sql.DataFrame = [y: string, age: int ... 1 more field]

scala> age_camp.show()
+---+---+--------+
|  y|age|count(y)|
+---+---+--------+
| no| 18|       5|
| no| 19|      24|
| no| 20|      35|
| no| 21|      57|
| no| 22|      89|
| no| 23|     158|
| no| 24|     234|
| no| 25|     414|
| no| 26|     671|
| no| 27|     768|
| no| 28|     876|
| no| 29|    1014|
| no| 30|    1540|
| no| 31|    1790|
| no| 32|    1864|
| no| 33|    1762|
| no| 34|    1732|
| no| 35|    1685|
| no| 36|    1611|
| no| 37|    1526|
+---+---+--------+
only showing top 20 rows


scala>
```

Type here to search | ENG | 23:53 | 25-04-2021