

# University of Massachusetts Boston

## MSIS685 – Big data Analytics – Spring 2021

### Programming Assignment I: using Spark/Pyspark to analyze a bank dataset

Submitted by - Alisha Warke

1.

BD\_prog\_Assgn1\_Alisha (Python)

My Cluster

```
1 # File location and type
2 file_location = "/FileStore/shared_uploads/a.warke001@umb.edu/bank_full.csv"
3 file_type = "csv"
4
5 # CSV options
6 infer_schema = "true"
7 first_row_is_header = "true"
8 delimiter = ";"
9
10 # The applied options are for CSV files. For other file types, these will be ignored.
11 dataf = spark.read.format(file_type) \
12     .option("inferSchema", infer_schema) \
13     .option("header", first_row_is_header) \
14     .option("sep", delimiter) \
15     .load(file_location)
16
17 display(dataf)
```

(3) Spark Jobs

dataf: pyspark.sql.dataframe.DataFrame = [age: integer, job: string ... 15 more fields]

	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
1	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
2	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
3	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
4	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
5	1	no	no	unknown	5	may	198	1	-1	0	unknown	no
6	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no

Showing the first 1000 rows.

Command took 2.07 seconds -- by a.warke001@umb.edu at 4/26/2021, 4:15:16 AM on My Cluster

Cmd 3

```
1 # Create a view or table
```

2.

BD\_prog\_Assgn1\_Alisha (Python)

My Cluster

```
1 dataf.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)
```

Command took 0.03 seconds -- by a.warke001@umb.edu at 4/26/2021, 4:18:35 AM on My Cluster

3.

BD\_prog\_Assgn1\_Alisha (Python)

My Cluster File Edit View: Standard Permissions Run All Clear Publish Comments Experiment Revision history

Command took 0.03 seconds -- by a.warke001@umb.edu at 4/26/2021, 4:18:35 AM on My Cluster

Cmd 8

```
1 # count number of each type of job
2 dataf.groupby("job").count().show()
```

(2) Spark Jobs

job	count
management	9458
retired	2264
unknown	288
self-employed	1579
student	938
blue-collar	9732
entrepreneur	1487
admin.	5171
technician	7597
services	4154
housemaid	1240
unemployed	1303

Command took 1.03 seconds -- by a.warke001@umb.edu at 4/26/2021, 4:18:53 AM on My Cluster

4.

BD\_prog\_Assgn1\_Alisha (Python)

My Cluster File Edit View: Standard Permissions Run All Clear Publish Comments Experiment Revision history

Command took 0.95 seconds -- by a.warke001@umb.edu at 4/26/2021, 6:30:58 AM on My Cluster

Cmd 9

```
1 #count people who age is above 30
2 dataf.filter(dataf.age > 30).count()
```

(2) Spark Jobs

Out[55]: 38181

Command took 0.95 seconds -- by a.warke001@umb.edu at 4/26/2021, 6:30:58 AM on My Cluster

Cmd 10

5.

BD\_prog\_Assgn1\_Alisha (Python)

My Cluster File Edit View: Standard Permissions Run All Clear Publish Comments Experiment Revision history

Command took 0.95 seconds -- by a.warke001@umb.edu at 4/26/2021, 6:30:58 AM on My Cluster

Cmd 10

```
1 %sql
2
3 SELECT bank_full_csv.education, count(bank_full_csv.education) from bank_full_csv where bank_full_csv.y=="yes" group by bank_full_csv.education
```

(2) Spark Jobs

education

education	count	percentage
unknown	5171	5%
tertiary	1487	11%
secondary	4154	46%
primary	1240	38%

Plot Options...

Command took 0.69 seconds -- by a.warke001@umb.edu at 4/26/2021, 6:32:13 AM on My Cluster

Cmd 11

6.

BD\_prog\_Assgn1\_Alisha (Python)

My Cluster File Edit View: Standard Permissions Run All Clear Publish Comments Experiment Revision history

Cmd 11

```

1 #count number of people who have a personal loan and who dont
2 dataf.groupby('loan').count().show()
3 #count total loans
4 dataf.select("loan").count()

```

▶ (4) Spark Jobs

```

+-----+
| loan|count|
+-----+
| no|37967|
| yes| 7244|
+-----+

```

Out[51]: 45211

Command took 1.64 seconds -- by a.warke001@umb.edu at 4/26/2021, 6:10:22 AM on My Cluster

7.

BD\_prog\_Assgn1\_Alisha (Python)

My Cluster File Edit View: Standard Permissions Run All Clear Publish Comments Experiment Revision history

Cmd 12

```

1 #count number of people who are married, single and divorced
2 dataf.groupby('marital').count().show()

```

▶ (2) Spark Jobs

```

+-----+
| marital|count|
+-----+
| divorced| 5207|
| married|27214|
| single|12799|
+-----+

```

Command took 1.24 seconds -- by a.warke001@umb.edu at 4/26/2021, 4:21:01 AM on My Cluster

8.

BD\_prog\_Assgn1\_Alisha (Python)

My Cluster File Edit View: Standard Permissions Run All Clear Publish Comments Experiment Revision history

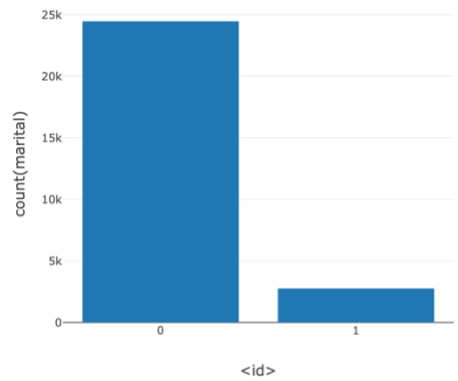
Cmd 13

```

1 %sql
2 /* count 0 =number of married clients who have not subscribed a term deposit */
3 /* count 1 =number of married clients who have subscribed a term deposit */
4
5
6 SELECT count(bank_full_csv.marital) from bank_full_csv where bank_full_csv.marital=="married" group by bank_full_csv.y

```

▶ (2) Spark Jobs

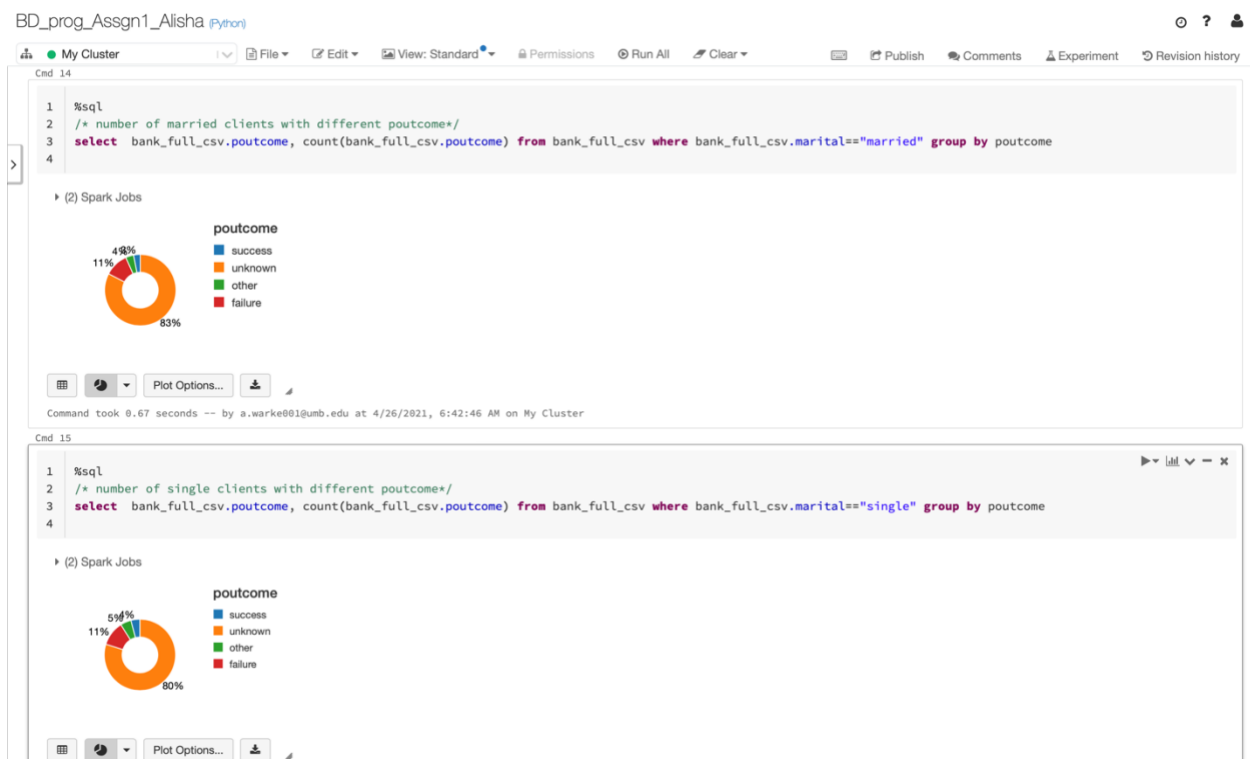


count(marital)

<id>

Command took 0.58 seconds -- by a.warke001@umb.edu at 4/26/2021, 6:36:52 AM on My Cluster

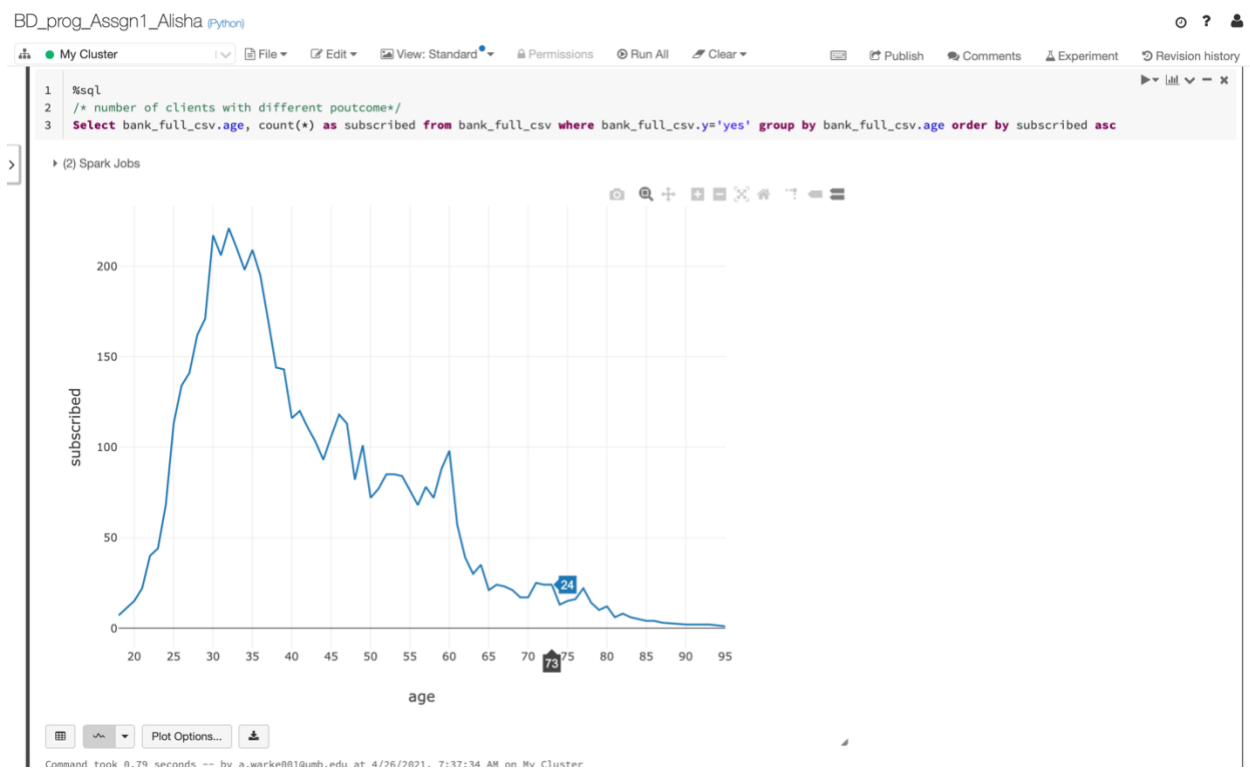
9.



10.



11.



12.



=====\*