Zeppelin-Spark Ass…

```
%md
Mirul Patel (N01489347)
```

READY

```
%md
KEY STEPS:

1)I have read the data from HDFS
2)Perfromed filtration and aggregation on data
3)created new dataframe and stored the results of aggregation and then save/write that file into HDFS
4)created a temporary views
5)performed SQL queries like JOIN and various aggregation on that views for desire results
```

READY

```
%spark2

val ws = spark.read.option("header","true").option("inferSchema", "true").csv("/tmp/worldsales.csv")
```

FINISHED

```
ws: org.apache.spark.sql.DataFrame = [Id: int, Region: string ... 13 more fields]
```

Took 0 sec. Last updated by anonymous at October 10 2022, 11:05:49 PM. (outdated)

```
%spark2
ws.printSchema()
```

FINISHED

```
root
 |-- Id: integer (nullable = true)
 |-- Region: string (nullable = true)
 |-- Country: string (nullable = true)
 |-- Item_Type: string (nullable = true)
 |-- Sales_Channel: string (nullable = true)
 |-- Order_Priority: string (nullable = true)
 |-- Order_Date: string (nullable = true)
 |-- Order_ID: integer (nullable = true)
 |-- Ship_Date: string (nullable = true)
 |-- Units_Sold: integer (nullable = true)
 |-- Unit_Price: double (nullable = true)
 |-- Unit_Cost: double (nullable = true)
 |-- Total_Revenue: double (nullable = true)
 |-- Total_Cost: double (nullable = true)
 |-- Total_Profit: double (nullable = true)
```

Took 1 sec. Last updated by anonymous at October 10 2022, 11:05:55 PM.

```
%spark2
ws.head(5)
```

FINISHED

```
res6: Array[org.apache.spark.sql.Row] = Array([1,Middle East and North Africa,Libya,Cosmetics,Offline,
M,10/18/2014,686800706,10/31/2014,8446,437.2,263.33,3692591.2,2224085.18,1468506.02], [2,North Americ
```

```
a,Canada,Vegetables,Online,M,11/7/2011,185941302,12/8/2011,3018,154.06,90.93,464953.08,274426.74,19052
6.34], [3,Middle East and North Africa,Libya,Baby Food,Offline,C,10/31/2016,246222341,12/9/2016,1517,2
55.28,159.42,387259.76,241840.14,145419.62], [4,Asia,Japan,Cereal,Offline,C,4/10/2010,161442649,5/12/2
010,3322,205.7,117.11,683335.4,389039.42,294295.98], [5,Sub-Saharan Africa,Chad,Fruits,Offline,H,8/16/
2011,645713555,8/31/2011,9845,9.33,6.92,91853.85,68127.4,23726.45])
```

Took 0 sec. Last updated by anonymous at October 10 2022, 2:56:08 PM.

```
%spark2
val wsfilter = ws.filter(ws("Units_Sold") > 8000 and ws("Unit_Cost") > 500).toDF("Id",
"Region","Country","Item_Type","Sales_Channel","Order_Priority","Order_Date","Order_ID","Ship_Date","U
    ,"Total_Revenue","Total_Cost","Total_Profit")
```

FINISHED

```
wsfilter: org.apache.spark.sql.DataFrame = [Id: int, Region: string ... 13 more fields]
```

Took 1 sec. Last updated by anonymous at October 11 2022, 10:14:53 AM.

```
%spark2
wsfilter.collect()
```

FINISHED

```
res26: Array[org.apache.spark.sql.Row] = Array([20,Sub-Saharan Africa,Senegal,Household,Offline,L,8/2
7/2012,247802054,9/8/2012,8989,668.27,502.54,6007079.03,4517332.06,1489746.97], [37,Sub-Saharan Afric
a,Swaziland,Office Supplies,Offline,H,10/3/2013,405785882,10/22/2013,9915,651.21,524.96,6456747.15,520
4978.4,1251768.75])
```

Took 0 sec. Last updated by anonymous at October 11 2022, 10:15:12 AM.

```
%spark2
val grouppbyresults = ws.groupBy("Region").count().toDF("Region","Count")
```

FINISHED

```
grouppbyresults: org.apache.spark.sql.DataFrame = [Region: string, Count: bigint]
```

Took 0 sec. Last updated by anonymous at October 11 2022, 10:15:23 AM.

```
%spark2
grouppbyresults.collect()
```

FINISHED

```
res27: Array[org.apache.spark.sql.Row] = Array([Middle East and North Africa,6], [Australia and Oceani
a,2], [Europe,12], [Sub-Saharan Africa,15], [Central America and the Caribbean,6], [North America,3],
 [Asia,5])
```

Took 1 sec. Last updated by anonymous at October 11 2022, 10:15:33 AM.

```
%spark2
grouppbyresults.repartition(1).write.mode("overwrite").option("header","true").csv("/tmp/subset.csv")
```

FINISHED

Took 1 sec. Last updated by anonymous at October 10 2022, 11:19:15 PM.

```
%spark2
grouppbyresults.createOrReplaceTempView("Regionview")
```

FINISHED

Took 0 sec. Last updated by anonymous at October 10 2022, 11:27:27 PM.

```
%spark2
```

FINISHED

```
ws.createOrReplaceTempView("Salesview")
```

Took 0 sec. Last updated by anonymous at October 11 2022, 10:29:13 AM.

```
%spark2.sql
SELECT * FROM Regionview
```

FINISHED

Region
Middle East and North Africa
Australia and Oceania
Europe
Sub-Saharan Africa
Central America and the Caribbean
North America
Asia

Took 1 sec. Last updated by anonymous at October 10 2022, 11:46:12 PM. (outdated)

```
%spark2.sql
SELECT Region, SUM(Units_sold) as Total_Unit_sale FROM Salesview
GROUP BY Region
```

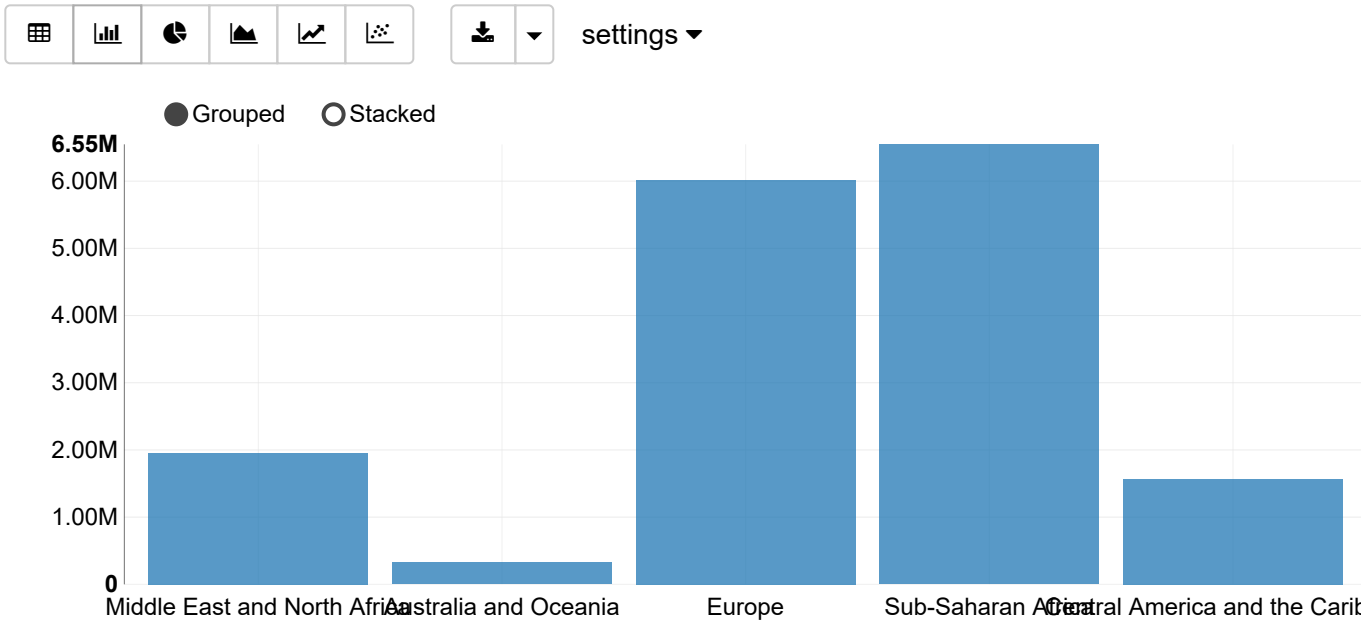
FINISHED

Region	Total_Unit_
Middle East and North Africa	24273
Australia and Oceania	5207
Europe	67424
Sub-Saharan Africa	81582
Central America and the Caribbean	30380
North America	10607
Asia	24189

Took 1 sec. Last updated by anonymous at October 11 2022, 10:29:26 AM.

```
%spark2.sql
SELECT Region, SUM(total_profit) as TOTAL_PROFIT
FROM Salesview
GROUP BY Region
```

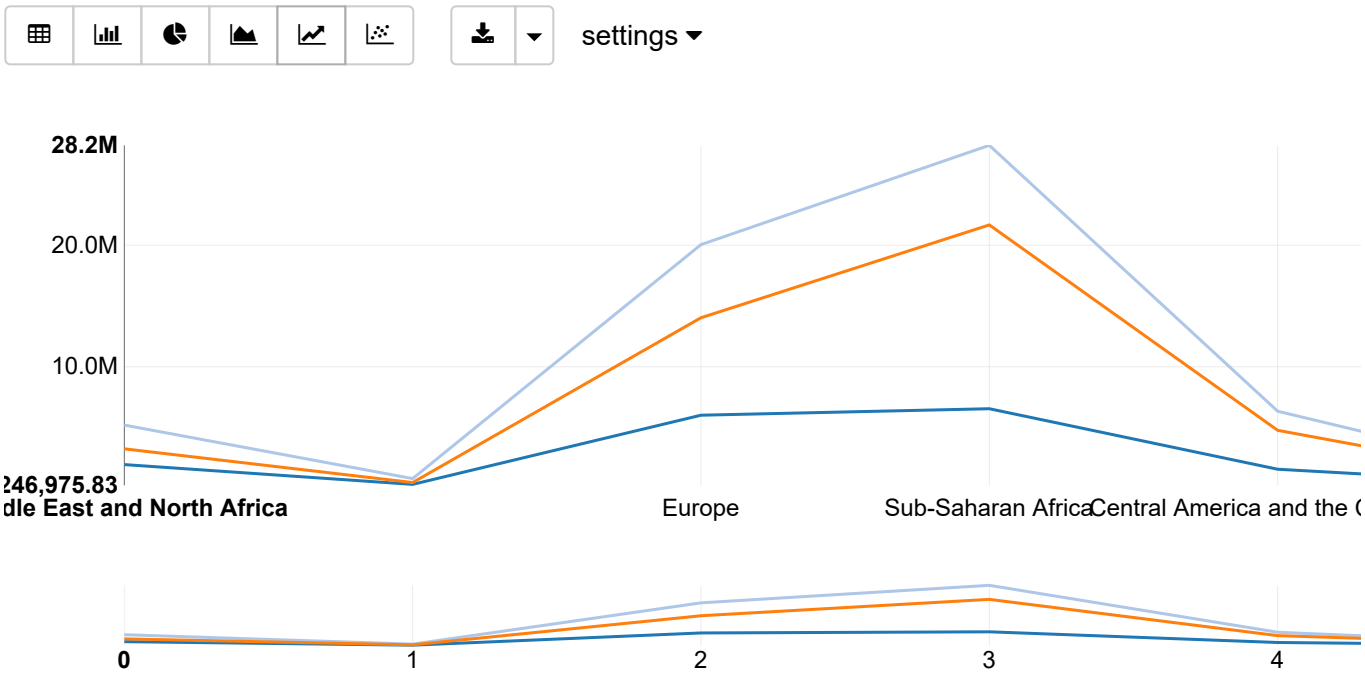
FINISHED



Took 1 sec. Last updated by anonymous at October 11 2022, 10:30:03 AM. (outdated)

```
%spark2.sql
SELECT Region, SUM(total_profit) as PROFIT, SUM(total_revenue) as REVENUE, SUM(total_cost) as COST
FROM Salesview
GROUP BY Region
```

FINISHED



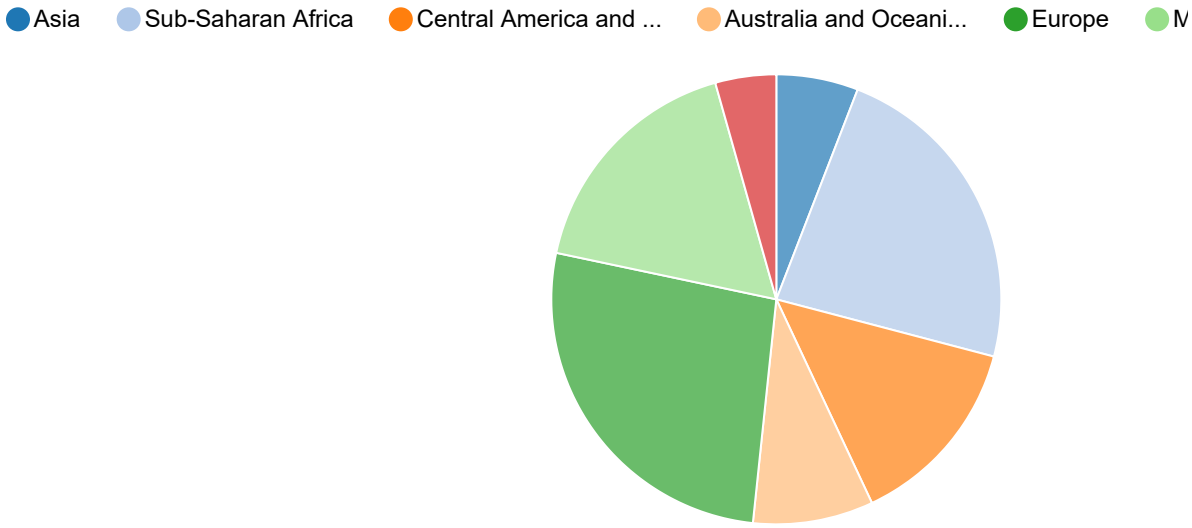
Took 2 sec. Last updated by anonymous at October 11 2022, 10:30:51 AM. (outdated)

```
%spark2.sql
```

FINISHED

```
SELECT s.Region, AVG(s.total_profit) as Average_Profit, r.Count
FROM Salesview s
JOIN Regionview r
ON r.Region=s.Region
```

settings



Took 2 sec. Last updated by anonymous at October 11 2022, 10:59:07 AM. (outdated)

READY

```
%md
Based on above query results:
-Europe is the most profitable region
And if company/customer want to open stores based on profit then the first choice should be Europe, fo
and Central America and the Caribbean
```