

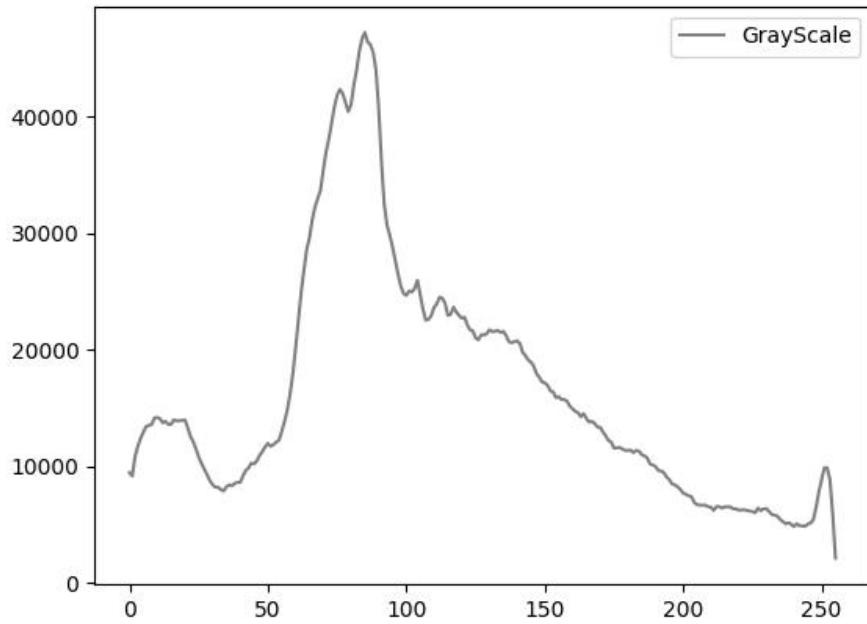
Distributed Data Analytics
Syed Khalid Ahmed
Marticulation number: 276970

Note : The Codes are provided in the respective folders

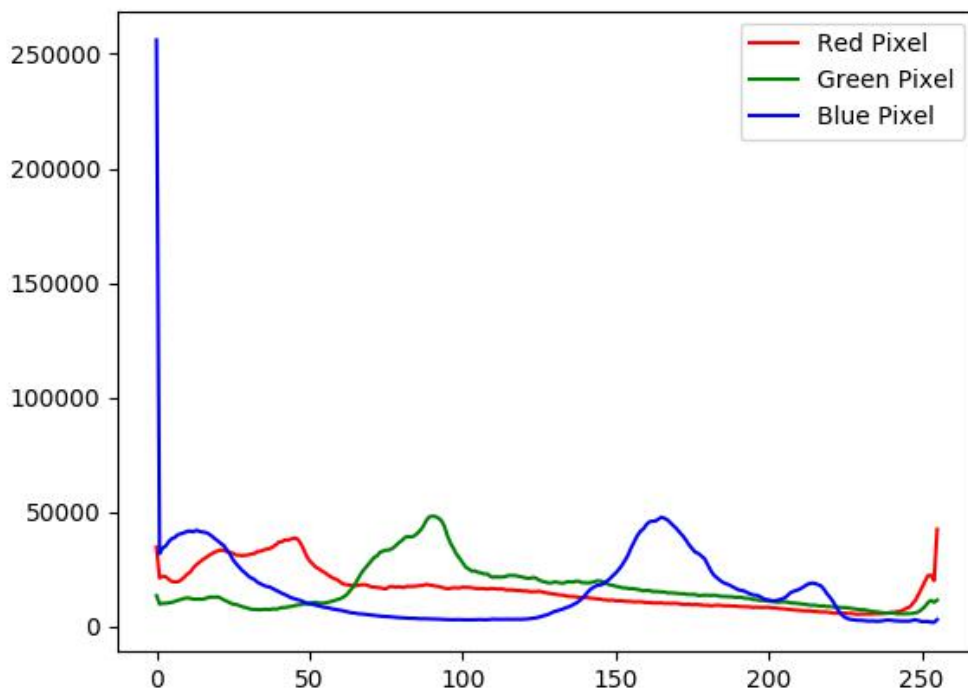
Exercise 01: Image Histogram using Aggregation

Output:

For Grayscale:



For RGB:



Exercise 02: Decision Tree on Iris dataset

Outputs:

I have run the code with different hyper-parameters. The results are included in the snapshots below:

1) Impurity = Entropy , Depth = 3

```
khalid@ubuntu: ~/Downloads/spark-2.1.1-bin-hadoop2.7
khalid@ubuntu:~/Downloads/spark-2.1.1-bin-hadoop2.7$ ./bin/spark-submit /home/khalid/Documents/Homework08/Exercise02.py
DecisionTreeModel classifier of depth 3 with 7 nodes
If (feature 3 <= 0.6)
  Predict: 1.0
Else (feature 3 > 0.6)
  If (feature 3 <= 1.7)
    If (feature 2 <= 4.9)
      Predict: 2.0
    Else (feature 2 > 4.9)
      Predict: 3.0
  Else (feature 3 > 1.7)
    Predict: 3.0

Precision of Setosa 1.0
Precision of Versicolor 1.0
Precision of Virginica 0.833333333333

Recall of Setosa 1.0
Recall of Versicolor 0.857142857143
Recall of Virginica 1.0

F-1 Score /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7/python/lib/pyspark.zip/pyspark/mllib/evaluation.py:262: UserWarning: Deprecated in 2.0.0. Use accuracy.
0.96

Confusion Matrix
[[ 13.  0.  0.]
 [  0.  6.  1.]
 [  0.  0.  5.]]

khalid@ubuntu:~/Downloads/spark-2.1.1-bin-hadoop2.7$
```

2) Impurity = Entropy , Depth = 4

```
khalid@ubuntu: ~/Downloads/spark-2.1.1-bin-hadoop2.7
khalid@ubuntu:~/Downloads/spark-2.1.1-bin-hadoop2.7$ ./bin/spark-submit /home/khalid/Documents/Homework08/Exercise02.py
DecisionTreeModel classifier of depth 4 with 15 nodes
If (feature 3 <= 0.5)
  Predict: 1.0
Else (feature 3 > 0.5)
  If (feature 3 <= 1.7)
    If (feature 2 <= 4.9)
      If (feature 0 <= 4.9)
        Predict: 2.0
      Else (feature 0 > 4.9)
        Predict: 2.0
    Else (feature 2 > 4.9)
      If (feature 0 <= 6.1)
        Predict: 2.0
      Else (feature 0 > 6.1)
        Predict: 3.0
  Else (feature 3 > 1.7)
    If (feature 2 <= 4.9)
      If (feature 1 <= 3.0)
        Predict: 3.0
      Else (feature 1 > 3.0)
        Predict: 2.0
    Else (feature 2 > 4.9)
      Predict: 3.0

Precision of Setosa 1.0
Precision of Versicolor 0.846153846154
Precision of Virginica 0.875

Recall of Setosa 0.9
Recall of Versicolor 0.916666666667
Recall of Virginica 0.875

F-1 Score /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7/python/lib/pyspark.zip/pyspark/mllib/evaluation.py:262: UserWarning: Deprecated in 2.0.0. Use accuracy.
0.9

Confusion Matrix
[[ 9.  1.  0.]
 [ 0. 11.  1.]]

khalid@ubuntu:~/Downloads/spark-2.1.1-bin-hadoop2.7$
```

3) Impurity = Gini , Depth = 3

```
khalid@ubuntu: ~/Downloads/spark-2.1.1-bin-hadoop2.7
khalid@ubuntu:~/Downloads/spark-2.1.1-bin-hadoop2.7$ ./bin/spark-submit /home/khalid/Documents/Homework08/Exercise02.py
DecisionTreeModel classifier of depth 3 with 9 nodes
  If (feature 3 <= 0.5)
    Predict: 1.0
  Else (feature 3 > 0.5)
    If (feature 3 <= 1.6)
      If (feature 2 <= 4.9)
        Predict: 2.0
      Else (feature 2 > 4.9)
        Predict: 3.0
    Else (feature 3 > 1.6)
      If (feature 2 <= 5.0)
        Predict: 3.0
      Else (feature 2 > 5.0)
        Predict: 3.0

Precision of Setosa 1.0
Precision of Versicolor 1.0
Precision of Virginica 1.0

Recall of Setosa 1.0
Recall of Versicolor 1.0
Recall of Virginica 1.0

F-1 Score /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7/python/lib/pyspark.zip/pyspark/mllib/evaluation.py:262: UserWarning: Deprecated in 2.0.0. Use accuracy.
1.0

Confusion Matrix
[[ 6.  0.  0.]
 [ 0.  7.  0.]
 [ 0.  0.  9.]]

khalid@ubuntu:~/Downloads/spark-2.1.1-bin-hadoop2.7$
```

4) Impurity = Gini , Depth = 4

```
khalid@ubuntu: ~/Downloads/spark-2.1.1-bin-hadoop2.7
khalid@ubuntu:~/Downloads/spark-2.1.1-bin-hadoop2.7$ ./bin/spark-submit /home/khalid/Documents/Homework08/Exercise02.py
DecisionTreeModel classifier of depth 4 with 15 nodes
  If (feature 3 <= 0.5)
    Predict: 1.0
  Else (feature 3 > 0.5)
    If (feature 2 <= 4.9)
      If (feature 3 <= 1.6)
        If (feature 1 <= 3.4)
          Predict: 2.0
        Else (feature 1 > 3.4)
          Predict: 1.0
      Else (feature 3 > 1.6)
        If (feature 1 <= 3.0)
          Predict: 3.0
        Else (feature 1 > 3.0)
          Predict: 2.0
    Else (feature 2 > 4.9)
      If (feature 3 <= 1.6)
        If (feature 0 <= 6.0)
          Predict: 2.0
        Else (feature 0 > 6.0)
          Predict: 3.0
      Else (feature 3 > 1.6)
        Predict: 3.0

Precision of Setosa 1.0
Precision of Versicolor 1.0
Precision of Virginica 0.90909090909091

Recall of Setosa 1.0
Recall of Versicolor 0.875
Recall of Virginica 1.0

F-1 Score /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7/python/lib/pyspark.zip/pyspark/mllib/evaluation.py:262: UserWarning: Deprecated in 2.0.0. Use accuracy.
0.965517241379

Confusion Matrix
[[ 11.  0.  0.]
 [ 0.  7.  1.]
 [ 0.  0. 10.]]
```

Exercise 03: Naive Bayes on Spam Dataset

Output:

```
khalid@ubuntu: ~/Downloads/spark-2.1.1-bin-hadoop2.7
khalid@ubuntu:~/Downloads/spark-2.1.1-bin-hadoop2.7$ ./bin/spark-submit /home/khalid/Documents/Homework08/Spark-Histogram.py
khalid@ubuntu:~/Downloads/spark-2.1.1-bin-hadoop2.7$ ./bin/spark-submit /home/khalid/Documents/Homework08/Exercise03.py

F1 metric = 0.929586

khalid@ubuntu:~/Downloads/spark-2.1.1-bin-hadoop2.7$
```

Bonus : Aggregation in MongoDB

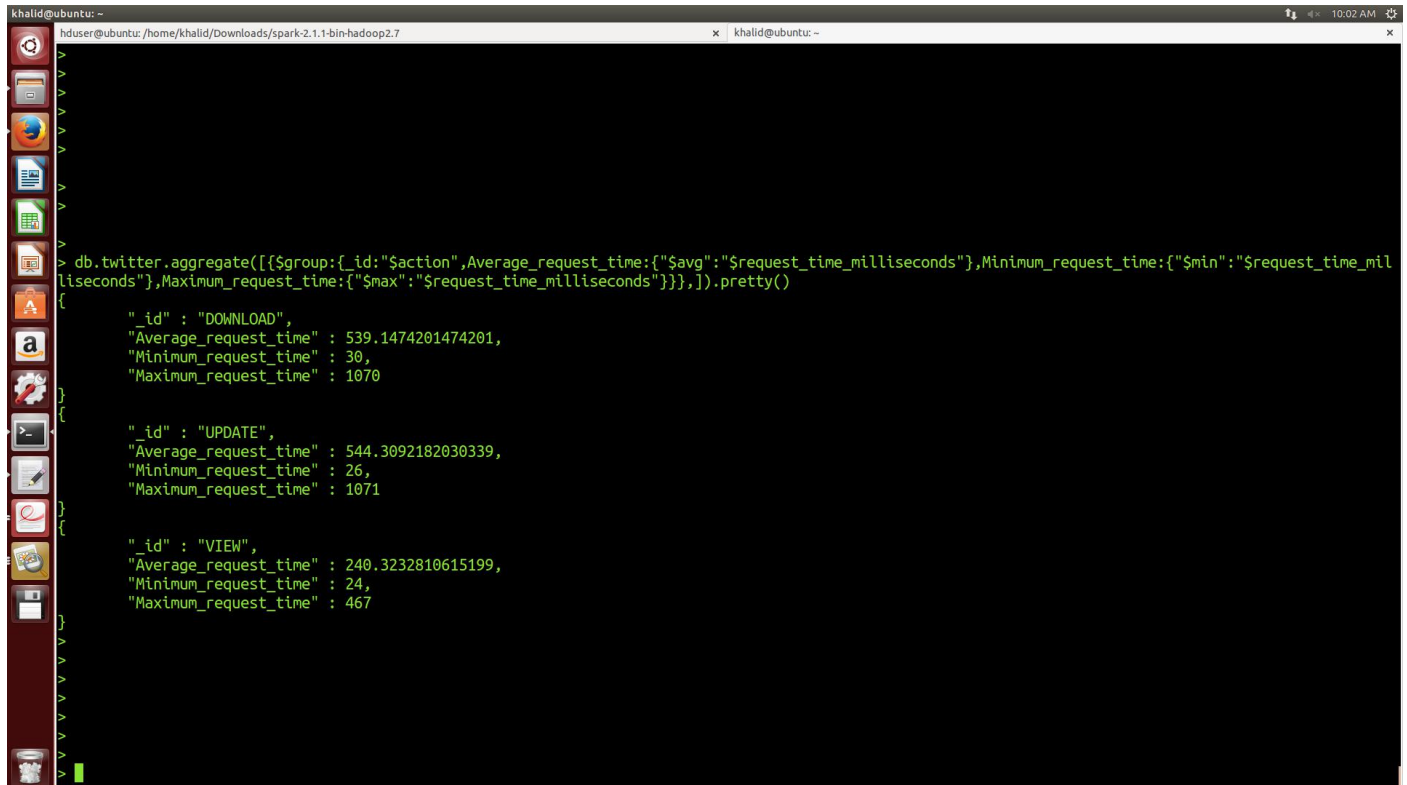
1) `db.twitter.aggregate([{$group:{_id:"$action", count:{$sum:1}}])`

```
khalid@ubuntu: ~
hduuser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7
khalid@ubuntu: ~

> db.twitter.aggregate([{$group:{_id:"$action", count:{$sum:1}}])
{ "_id" : "DOWNLOAD", "count" : 814 }
{ "_id" : "UPDATE", "count" : 857 }
{ "_id" : "VIEW", "count" : 829 }
```

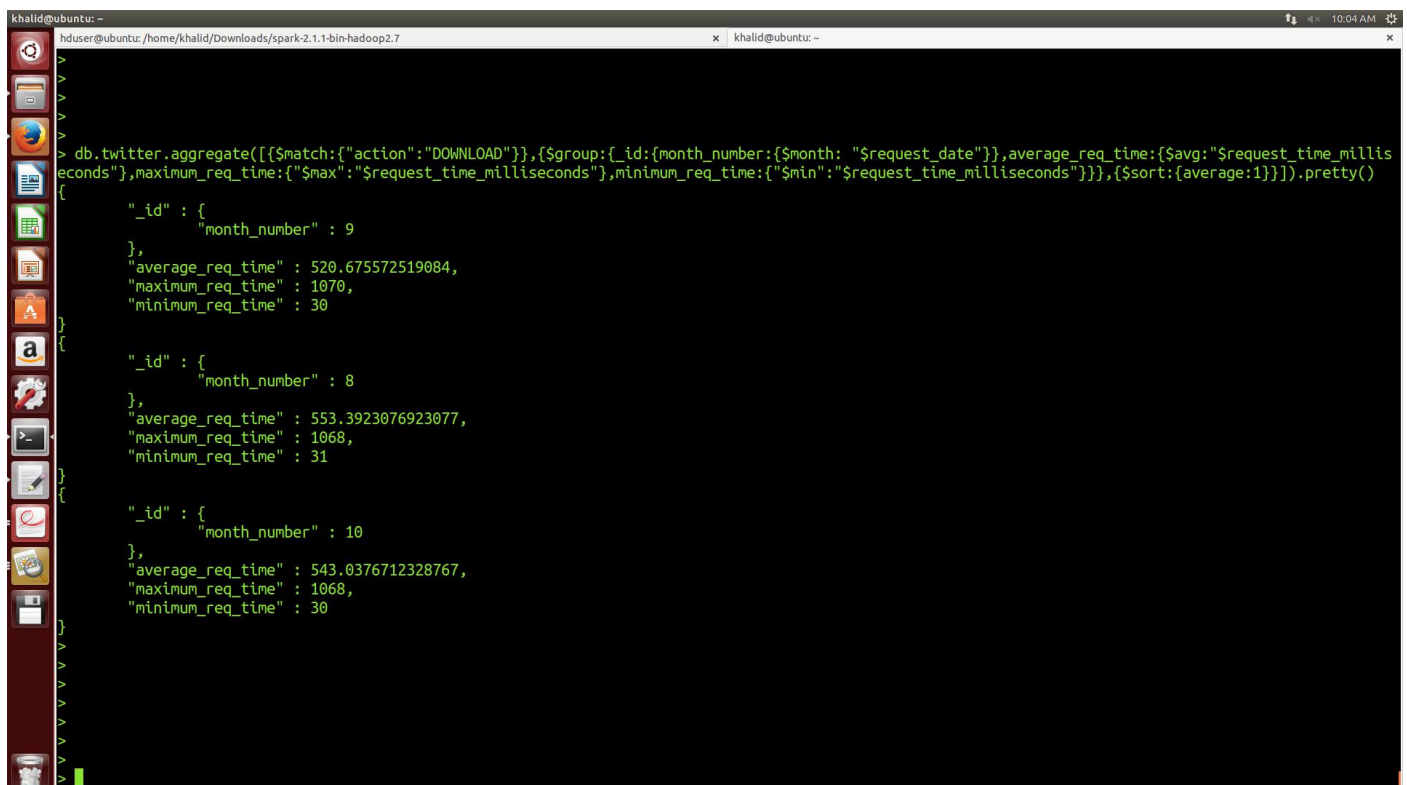
2)

```
db.twitter.aggregate([{$group:{_id:"$action",Average_request_time:{"$avg":"$request_time_milliseconds"},Minimum_request_time:{"$min":"$request_time_milliseconds"},Maximum_request_time:{"$max":"$request_time_milliseconds"}},$sort:{_id:1}}],1).pretty()
```



```
khalid@ubuntu: ~  
hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7  
x khalid@ubuntu: ~  
> db.twitter.aggregate([{$group:{_id:"$action",Average_request_time:{"$avg":"$request_time_milliseconds"},Minimum_request_time:{"$min":"$request_time_milliseconds"},Maximum_request_time:{"$max":"$request_time_milliseconds"}},$sort:{_id:1}}],1).pretty()  
{  
  "_id" : "DOWNLOAD",  
  "Average_request_time" : 539.1474201474201,  
  "Minimum_request_time" : 30,  
  "Maximum_request_time" : 1070  
},  
{  
  "_id" : "UPDATE",  
  "Average_request_time" : 544.3092182030339,  
  "Minimum_request_time" : 26,  
  "Maximum_request_time" : 1071  
},  
{  
  "_id" : "VIEW",  
  "Average_request_time" : 240.3232810615199,  
  "Minimum_request_time" : 24,  
  "Maximum_request_time" : 467  
}
```

3) db.twitter.aggregate([{\$match:{"action":"DOWNLOAD"}},{\$group:{_id:{month_number:{\$month: "\$request_date"}},average_req_time:{\$avg:"\$request_time_milliseconds"},maximum_req_time:{"\$max":"\$request_time_milliseconds"},minimum_req_time:{"\$min":"\$request_time_milliseconds"}},{\$sort:{average:1}}],1).pretty()



```
khalid@ubuntu: ~  
hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7  
x khalid@ubuntu: ~  
> db.twitter.aggregate([{$match:{"action":"DOWNLOAD"}},{$group:{_id:{month_number:{$month: "$request_date"}},average_req_time:{$avg:"$request_time_milliseconds"},maximum_req_time:{"$max":"$request_time_milliseconds"},minimum_req_time:{"$min":"$request_time_milliseconds"}},{$sort:{average:1}}],1).pretty()  
{  
  "_id" : {  
    "month_number" : 9  
  },  
  "average_req_time" : 520.675572519084,  
  "maximum_req_time" : 1070,  
  "minimum_req_time" : 30  
},  
{  
  "_id" : {  
    "month_number" : 8  
  },  
  "average_req_time" : 553.3923076923077,  
  "maximum_req_time" : 1068,  
  "minimum_req_time" : 31  
},  
{  
  "_id" : {  
    "month_number" : 10  
  },  
  "average_req_time" : 543.0376712328767,  
  "maximum_req_time" : 1068,  
  "minimum_req_time" : 30  
}
```

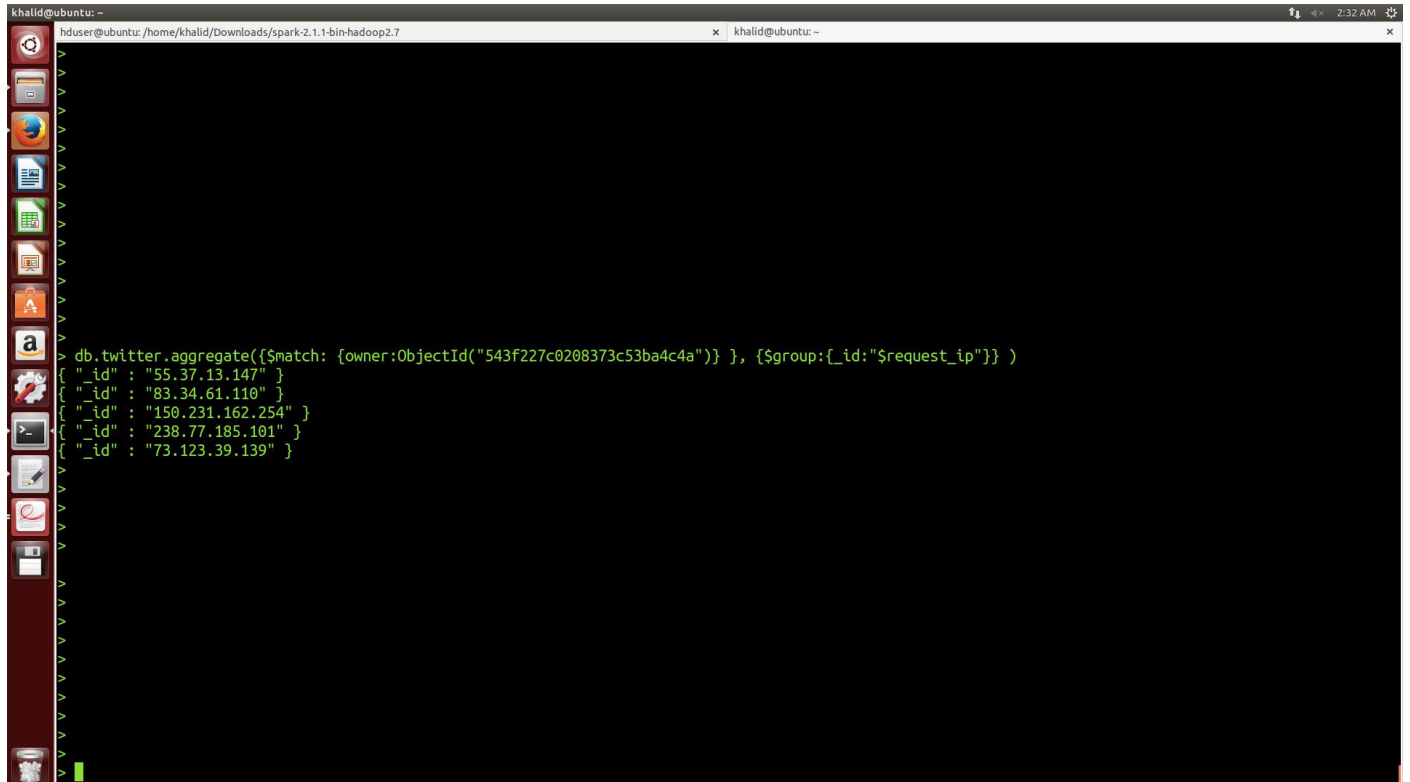
4) `db.twitter.aggregate({$match: {$and: [{action:"UPDATE"} , {owner:ObjectId("543f227c0208373c53ba4c4a")}}] }, {$group:{_id:"$action", Number_of_occurence:{$sum: 1}}})`



A terminal window titled 'khalid@ubuntu: ~' showing a MongoDB aggregation command and its output. The command filters for 'UPDATE' actions by a specific owner and groups them by action, summing the occurrences. The output shows that the 'UPDATE' action occurred 4 times.

```
khalid@ubuntu: ~  
hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7  
> db.twitter.aggregate({$match: {$and: [{action:"UPDATE"} , {owner:ObjectId("543f227c0208373c53ba4c4a")}}] }, {$group:{_id:"$action", Number_of_occurence:{$sum: 1}}})  
{ "_id" : "UPDATE", "Number_of_occurence" : 4 }
```

5) `db.twitter.aggregate({$match: {owner:ObjectId("543f227c0208373c53ba4c4a")}}, {$group:{_id:"$request_ip"}})`



A terminal window titled 'khalid@ubuntu: ~' showing a MongoDB aggregation command and its output. The command filters for a specific owner and groups the results by request IP. The output is an array of five objects, each containing a request IP.

```
khalid@ubuntu: ~  
hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7  
> db.twitter.aggregate({$match: {owner:ObjectId("543f227c0208373c53ba4c4a")}}, {$group:{_id:"$request_ip"}} )  
{ "_id" : "55.37.13.147" }  
{ "_id" : "83.34.61.110" }  
{ "_id" : "150.231.162.254" }  
{ "_id" : "238.77.185.101" }  
{ "_id" : "73.123.39.139" }
```

6) db.twitter.aggregate({\$match: {request_method:"SET"}}, {\$group:{_id:"\$request_ip", Total_Sum:{\$sum: 1}}})

```
khalid@ubuntu: ~
hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7
x khalid@ubuntu: ~

>
>
>
> db.twitter.aggregate({$match: {request_method:"SET"}}, {$group:{_id:"$request_ip", Total_Sum:{$sum: 1}}})
{ "_id" : "57.162.89.255", "Total_Sum" : 1 }
{ "_id" : "251.165.16.151", "Total_Sum" : 1 }
{ "_id" : "138.169.128.244", "Total_Sum" : 1 }
{ "_id" : "188.215.78.3", "Total_Sum" : 1 }
{ "_id" : "217.61.247.57", "Total_Sum" : 1 }
{ "_id" : "253.219.92.200", "Total_Sum" : 1 }
{ "_id" : "10.157.106.21", "Total_Sum" : 1 }
{ "_id" : "115.75.58.181", "Total_Sum" : 1 }
{ "_id" : "119.217.2.215", "Total_Sum" : 1 }
{ "_id" : "58.130.203.25", "Total_Sum" : 1 }
{ "_id" : "44.82.196.95", "Total_Sum" : 1 }
{ "_id" : "33.46.81.134", "Total_Sum" : 1 }
{ "_id" : "115.50.100.112", "Total_Sum" : 1 }
{ "_id" : "53.105.137.16", "Total_Sum" : 3 }
{ "_id" : "168.186.115.23", "Total_Sum" : 3 }
{ "_id" : "158.193.173.115", "Total_Sum" : 2 }
{ "_id" : "69.157.62.35", "Total_Sum" : 1 }
{ "_id" : "168.247.235.56", "Total_Sum" : 1 }
{ "_id" : "153.203.41.247", "Total_Sum" : 2 }
{ "_id" : "176.104.7.66", "Total_Sum" : 1 }
Type "it" for more
>
>
>
>
>
```