

## Exercise 07

### Distributed Data Analytics

Syed Khalid Ahmed

Note : The codes are provided in the respective folders

### Exercise 1: Resilient Distributed Datasets

#### 1. rightOuterJoin and fullOuterJoin

```
hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7
[0.00545141392891...]
[0.00332375465950...]
[0.02905792648229...]
[0.0194026348357...]
[0.00781321173533...]
[0.0251776958206...]
[0.00607764788308...]
[0.05245233188799...]
hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.75 ./bin/spark-submit /home/khalid/Homework07/Exercise1/1.py
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
17/06/02 13:47:35 WARN Utils: Your hostname, ubuntu resolves to a loopback address: 127.0.0.1; using 192.168.142.141 instead (on interface eth0)
17/06/02 13:47:35 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
17/06/02 13:47:37 INFO SparkContext: Running Spark version 2.1.1
17/06/02 13:47:37 WARN SparkContext: Support for Java 7 is deprecated as of Spark 2.0.0
17/06/02 13:47:38 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/06/02 13:47:39 INFO SecurityManager: Changing view acls to: hduser
17/06/02 13:47:39 INFO SecurityManager: Changing modify acls to: hduser
17/06/02 13:47:39 INFO SecurityManager: Changing view acls groups to:
17/06/02 13:47:39 INFO SecurityManager: Changing modify acls groups to:
17/06/02 13:47:39 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(hduser); groups with view permissions: Set(); users with modify permissions: Set(hduser); groups with modify permissions: Set()
17/06/02 13:47:41 INFO Utils: Successfully started service 'sparkDriver' on port 34050.
17/06/02 13:47:41 INFO SparkEnv: Registering MapOutputTracker
17/06/02 13:47:41 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
17/06/02 13:47:41 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
17/06/02 13:47:41 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-acc20cc0-e72f-46b8-a295-3a7e04ad6a69
17/06/02 13:47:41 INFO MemoryStore: MemoryStore started with capacity 366.3 MB
17/06/02 13:47:42 INFO SparkEnv: Registering OutputCommitCoordinator
17/06/02 13:47:43 INFO Utils: Successfully started service 'SparkUI' on port 4040.
17/06/02 13:47:43 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.142.141:4040
17/06/02 13:47:44 INFO SparkContext: Added file file:/home/khalid/Homework07/Exercise1/1.py at file:/home/khalid/Homework07/Exercise1/1.py with timestamp 1496436464821
17/06/02 13:47:44 INFO Utils: Copying /home/khalid/Homework07/Exercise1/1.py to /tmp/spark-aa48ce7e-a331-49ef-be38-78aba9d8abf1/userfiles-c2133823-d6d0-47c5-9d0c-558e63a3dec5/1.py
17/06/02 13:47:45 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 35920.
17/06/02 13:47:45 INFO NettyBlockTransferService: Server created on 192.168.142.141:35920
17/06/02 13:47:45 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
17/06/02 13:47:45 INFO BlockManagerMaster: Registering block manager BlockManagerId(driver, 192.168.142.141, 35920, None)
17/06/02 13:47:45 INFO BlockManagerMasterEndpoint: Registering block manager 192.168.142.141:35920 with 366.3 MB RAM, BlockManagerId(driver, 192.168.142.141, 35920, None)
17/06/02 13:47:45 INFO BlockManager: Registered BlockManager BlockManagerId(driver, 192.168.142.141, 35920, None)
17/06/02 13:47:45 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.142.141, 35920, None)
17/06/02 13:47:46 INFO SharedState: Warehouse path is 'file:/home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7/spark-warehouse/'.

For right outer join : [('apache', (None, 1)), ('operation', (None, 1)), ('partition', (None, 1)), ('parallel', (None, 1)), ('lambda', (None, 1)), ('scala', (None, 1))]

For full outer join : [('python', (1, None)), ('apache', (None, 1)), ('rdd', (1, None)), ('spark', (1, None)), ('create', (1, None)), ('operation', (None, 1)), ('partition', (None, 1)), ('class', (1, None)), ('parallel', (None, 1)), ('context', (1, None)), ('lambda', (None, 1)), ('scala', (None, 1))]

hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.75
```

## 2. Count 's' using map-reduce

```
hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7
hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7 x hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7 x khalid@ubuntu: ~/Downloads
17/06/02 13:47:45 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 192.168.142.141, 35920, None)
17/06/02 13:47:45 INFO BlockManagerMasterEndpoint: Registering block manager 192.168.142.141:35920 with 366.3 MB RAM, BlockManagerId(driver, 192.168.142.141, 35920, None)
17/06/02 13:47:45 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 192.168.142.141, 35920, None)
17/06/02 13:47:45 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.142.141, 35920, None)
17/06/02 13:47:46 INFO SharedState: Warehouse path is 'file:/home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7/spark-warehouse/'.

For right outer join : (('apache', (None, 1)), ('operation', (None, 1)), ('partition', (None, 1)), ('parallel', (None, 1)), ('lambda', (None, 1)), ('scala', (None, 1)))

For full outer join : (('python', (1, None)), ('apache', (None, 1)), ('rdd', (1, None)), ('spark', (1, None)), ('create', (1, None)), ('operation', (None, 1)), ('partition', (None, 1)), ('class', (1, None)), ('parallel', (None, 1)), ('context', (1, None)), ('lambda', (None, 1)), ('scala', (None, 1)))

hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7.5 ./bin/spark-submit /home/khalid/Homework07/Exercise1/2.py
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
17/06/02 13:49:21 WARN Utils: Your hostname, ubuntu resolves to a loopback address: 127.0.0.1; using 192.168.142.141 instead (on interface eth0)
17/06/02 13:49:21 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
17/06/02 13:49:23 WARN SparkContext: Running spark version 2.1.1
17/06/02 13:49:23 WARN SparkContext: Support for Java 7 is deprecated as of Spark 2.0.0
17/06/02 13:49:24 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/06/02 13:49:26 INFO SecurityManager: Changing view acls to: hduser
17/06/02 13:49:26 INFO SecurityManager: Changing modify acls to: hduser
17/06/02 13:49:26 INFO SecurityManager: Changing view acls groups to:
17/06/02 13:49:26 INFO SecurityManager: Changing modify acls groups to:
17/06/02 13:49:26 INFO SecurityManager: SecurityManager: authentication disabled; ut acls disabled; users with view permissions: Set(hduser); groups with view permissions: Set(); users with modify permissions: Set(hduser); groups with modify permissions: Set()
17/06/02 13:49:27 INFO Utils: Successfully started service 'sparkDriver' on port 37366.
17/06/02 13:49:28 INFO SparkEnv: Registering MapOutputTracker
17/06/02 13:49:28 INFO SparkEnv: Registering BlockManagerMaster
17/06/02 13:49:28 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
17/06/02 13:49:28 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-a7884e57-8528-4c29-b4d7-fff2f803e3b
17/06/02 13:49:28 INFO MemoryStore: MemoryStore started with capacity 366.3 MB
17/06/02 13:49:28 INFO SparkEnv: Registering OutputCommitCoordinator
17/06/02 13:49:30 INFO Utils: Successfully started service 'SparkUI' on port 4040.
17/06/02 13:49:30 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.142.141:4040
17/06/02 13:49:31 INFO SparkContext: Added file file:/home/khalid/Homework07/Exercise1/2.py at file:/home/khalid/Homework07/Exercise1/2.py with timestamp 1496436571466
17/06/02 13:49:31 INFO Utils: Copying /home/khalid/Homework07/Exercise1/2.py to /tmp/spark-b29c6d49-fab1-4be7-92c9-54e7c5a6c3b/userfiles-522f48ad-7f8e-499a-84bf-0fca92f81fa/2.py
17/06/02 13:49:31 INFO Executor: Starting executor ID driver on host localhost
17/06/02 13:49:31 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 46305.
17/06/02 13:49:31 INFO NettyBlockTransferService: Server created on 192.168.142.141:46305
17/06/02 13:49:31 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
17/06/02 13:49:32 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 192.168.142.141, 46305, None)
17/06/02 13:49:32 INFO BlockManagerMasterEndpoint: Registering block manager 192.168.142.141:46305 with 366.3 MB RAM, BlockManagerId(driver, 192.168.142.141, 46305, None)
17/06/02 13:49:32 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 192.168.142.141, 46305, None)
17/06/02 13:49:32 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.142.141, 46305, None)
17/06/02 13:49:33 INFO SharedState: Warehouse path is 'file:/home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7/spark-warehouse/'.

Total Count -> (('s', 4)]

hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7.5
```

## Exercise 2: Dataframe API and Spark SQL

```
hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7
hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7 x hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7 x khalid@ubuntu: ~/Downloads
17/06/02 13:54:49 INFO Utils: Successfully started service 'SparkUI' on port 4040.
17/06/02 13:54:49 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.142.141:4040
17/06/02 13:54:50 INFO SparkContext: Added file file:/home/khalid/Homework07/Exercise2/FinalDataset.py at file:/home/khalid/Homework07/Exercise2/FinalDataset.py with timestamp 1496436899914
17/06/02 13:54:50 INFO SparkContext: Copying /home/khalid/Homework07/Exercise2/FinalDataset.py to /tmp/spark-B9166325-c8a8-4e75-8891-4c7349d9fb40/userfiles-c9f23639-bd97-487d-904e-b59cc170497e/FinalDataset.py
17/06/02 13:54:51 INFO Executor: Starting executor ID driver on host localhost
17/06/02 13:54:51 INFO NettyBlockTransferService: Server created on 192.168.142.141:35808.
17/06/02 13:54:51 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
17/06/02 13:54:51 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 192.168.142.141, 35808, None)
17/06/02 13:54:51 INFO BlockManagerMasterEndpoint: Registering block manager 192.168.142.141:35808 with 366.3 MB RAM, BlockManagerId(driver, 192.168.142.141, 35808, None)
17/06/02 13:54:51 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 192.168.142.141, 35808, None)
17/06/02 13:54:51 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.142.141, 35808, None)
17/06/02 13:54:52 INFO SharedState: Warehouse path is 'file:/home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7/spark-warehouse/'.

Snapshot of the final dataset
-----+-----+-----+-----+-----+-----+
course|dob|first_name|last_name|points|s_id|
-----+-----+-----+-----+-----+-----+
Humanities and Art|October 10, 1983|Alan|Zoe|10|1|
Computer Science|September 20, 1980|Martin|Genberg|20|2|
Graphic Design|June 12, 1982|Athur|Watson|20|3|
Graphic Design|April 5, 1987|Anabelle|Sanberg|12|4|
Psychology|November 1, 1978|Kira|Schomer|11|5|
Business|17 February 1981|Christian|Klrian|10|6|
Machine Learning|1 January 1984|Barbara|Ballard|14|7|
Deep Learning|January 15, 1978|John|--|10|8|
Machine Learning|26 December 1989|Marcus|Carson|20|9|
Physics|30 December 1987|Martal|Brooks|11|10|
Data Analytics|June 12, 1975|Holly|Schwartz|12|11|
Computer Science|July 2, 1985|April|Black|11|12|
Computer Science|July 22, 1980|Irene|Bradley|13|13|
Psychology|7 February 1986|Mark|Weber|12|14|
Informatics|May 10, 1987|Rosal|Norran|9|15|
Business|August 10, 1984|Martin|Steele|7|16|
Machine Learning|16 December 1990|Collin|Martinez|9|17|
Data Analytics|unknown|Bridget|Twain|6|18|
Business|7 March 1980|Darlene|Hills|20|19|
Data Analytics|June 2, 1985|Zachary|--|10|20|

None

The histogram distribution is :
([s, 10, 15, 20], [4, 12, 4])

hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7.5
```

## Exercise 3: Pipeline API and ML workflows

The pipeline API allows us to chain multiple functions and give the output to a Machine Learning Algorithm for learning. Spark ML is a relatively new concept which allows us to do efficient machine learning.

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

17/06/02 13:57:21 INFO SparkContext: Running spark version 2.1.1
17/06/02 13:57:21 WARN SparkContext: Support for Java 7 is deprecated as of Spark 2.0.0
17/06/02 13:57:22 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/06/02 13:57:23 INFO SecurityManager: Changing view acls to: hduser
17/06/02 13:57:23 INFO SecurityManager: Changing view acls groups to:
17/06/02 13:57:23 INFO SecurityManager: Changing modify acls to: hduser
17/06/02 13:57:23 INFO SecurityManager: authentication disabled; ut acls disabled; users with view permissions: Set(hduser); groups with view permissions: Set(); users with modify permissions: Set(hduser); groups with modify permissions: Set()
17/06/02 13:57:24 INFO Utils: Successfully started service 'sparkDriver' on port 45159.
17/06/02 13:57:25 INFO SparkEnv: Registering MapOutputTracker
17/06/02 13:57:25 INFO SparkEnv: Registering BlockManagerMaster
17/06/02 13:57:25 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
17/06/02 13:57:25 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
17/06/02 13:57:25 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-c11d481c-fc57-4836-b853-343f000a2e39
17/06/02 13:57:25 INFO MemoryStore: MemoryStore started with capacity 366.3 MB
17/06/02 13:57:26 INFO SparkEnv: Registering OutputCommitCoordinator
17/06/02 13:57:27 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.142.141:4040
17/06/02 13:57:28 INFO SparkContext: Added file file:/home/khalid/Homework07/Exercise3/TF-IDF.py at file:/home/khalid/Homework07/Exercise3/TF-IDF.py with timestamp 1496437048814
17/06/02 13:57:29 INFO Executor: Copying /home/khalid/Homework07/Exercise3/TF-IDF.py to /tmp/spark-ec394490-da29-4a45-b85f-a79ca9079516/userFiles-13b2a7d1-f6c5-4564-89e2-5e7c782d73ab/TF-IDF.py
17/06/02 13:57:29 INFO Executor: Starting executor ID driver on host localhost
17/06/02 13:57:29 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 41154.
17/06/02 13:57:29 INFO NettyBlockTransferService: Server created on 192.168.142.141:41154
17/06/02 13:57:29 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
17/06/02 13:57:29 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 192.168.142.141, 41154, None)
17/06/02 13:57:29 INFO BlockManagerMasterEndpoint: Registering block manager 192.168.142.141:41154 with 366.3 MB RAM, BlockManagerId(driver, 192.168.142.141, 41154, None)
17/06/02 13:57:29 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 192.168.142.141, 41154, None)
17/06/02 13:57:29 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.142.141, 41154, None)
17/06/02 13:57:30 INFO SharedState: Warehouse path is 'file:/home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7/spark-warehouse/'.

-----
content|features|
-----
Bout to check my ...|(20,2,3,4,7,8,9,...|
Yeeeee four my l...|(20,0,1,2,4,5,6,...|
Its been a long t...|(20,0,1,4,5,6,7,...|
And I really don...|(20,0,1,5,6,9,10,...|
Last break is ove...|(20,1,4,5,6,9,11,...|
Winhighschool tre...|(20,0,1,3,4,6,7,...|
The other Credit ...|(20,0,1,2,4,6,8,...|
Jantne hold just...|(20,0,1,2,4,5,6,...|
RevHumisdom: You...|(20,0,3,5,6,7,8,...|
ralphmarston: whe...|(20,0,2,5,6,7,8,...|
RT @USER_a7dacf42...|(20,2,3,6,7,8,10,...|
I kno they mean w...|(20,1,3,4,5,6,...|
@USER_19251f5f LO...|(20,0,1,2,4,8,9,...|
@USER_e99c8bbe hm...|(20,3,4,10,13,14,...|
@PeterwildeM id ...|(20,0,3,4,6,7,8,...|
Now I'm in love w...|(20,0,1,3,5,7,6,...|
Shlt that crazy m...|(20,0,3,4,5,6,7,...|
Lol funny I don't...|(20,2,4,5,6,7,8,...|
@USER_fc39199 RT...|(20,0,2,4,7,8,10,...|
Won't be watching...|(20,1,4,7,8,9,10,...|
-----
```

## Exercise 4: Word2Vec

Word2Vec gives the vector representation of a string.

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

17/06/02 13:59:18 INFO SparkContext: Running spark version 2.1.1
17/06/02 13:59:18 WARN SparkContext: Support for Java 7 is deprecated as of Spark 2.0.0
17/06/02 13:59:20 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/06/02 13:59:21 INFO SecurityManager: Changing view acls to: hduser
17/06/02 13:59:21 INFO SecurityManager: Changing view acls groups to:
17/06/02 13:59:21 INFO SecurityManager: Changing modify acls to: hduser
17/06/02 13:59:21 INFO SecurityManager: authentication disabled; ut acls disabled; users with view permissions: Set(hduser); groups with view permissions: Set(); users with modify permissions: Set(hduser); groups with modify permissions: Set()
17/06/02 13:59:22 INFO Utils: Successfully started service 'sparkDriver' on port 35061.
17/06/02 13:59:22 INFO SparkEnv: Registering MapOutputTracker
17/06/02 13:59:22 INFO SparkEnv: Registering BlockManagerMaster
17/06/02 13:59:22 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
17/06/02 13:59:22 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
17/06/02 13:59:23 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-7b166bd-5031-4b56-8571-a4b13f4dbbce
17/06/02 13:59:23 INFO MemoryStore: MemoryStore started with capacity 366.3 MB
17/06/02 13:59:23 INFO SparkEnv: Registering OutputCommitCoordinator
17/06/02 13:59:25 INFO Utils: Successfully started service 'SparkUI' on port 4040.
17/06/02 13:59:25 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.142.141:4040
17/06/02 13:59:26 INFO SparkContext: Added file file:/home/khalid/Homework07/Exercise4/word2Vec.py at file:/home/khalid/Homework07/Exercise4/word2Vec.py with timestamp 1496437168186
17/06/02 13:59:26 INFO Executor: Copying /home/khalid/Homework07/Exercise4/word2Vec.py to /tmp/spark-cf3640f6-25c7-4988-b19b-f743f3b97112/userFiles-4b88de7a-492b-48e2-a180-f4ebae2e1ab/word2Vec.py
17/06/02 13:59:26 INFO Executor: Starting executor ID driver on host localhost
17/06/02 13:59:26 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 36719.
17/06/02 13:59:26 INFO NettyBlockTransferService: Server created on 192.168.142.141:36719
17/06/02 13:59:26 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
17/06/02 13:59:26 INFO BlockManagerMaster: Registering block manager 192.168.142.141:36719 with 366.3 MB RAM, BlockManagerId(driver, 192.168.142.141, 36719, None)
17/06/02 13:59:26 INFO BlockManagerMasterEndpoint: Registering BlockManager BlockManagerId(driver, 192.168.142.141, 36719, None)
17/06/02 13:59:26 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.142.141, 36719, None)
17/06/02 13:59:28 INFO SharedState: Warehouse path is 'file:/home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7/spark-warehouse/'.

-----
result|
-----
[[-0.0054715927047...|
[-0.0049304981075...|
[-0.0046137770930...|
[0.00676818879452...|
[-0.0066674195429...|
[0.00556259759176...|
[0.01400073827244...|
[0.01640053901176...|
[0.02322069885364...|
[-0.0284605090957...|
[-0.0124416192993...|
[-0.0105149815691...|
[0.00545143392891...|
[0.00332375465950...|
[0.02905792648229...|
[-0.0194026348357...|
[0.0078132173533...|
[0.02517776958206...|
[0.00807764780308...|
[0.024523188709...|
-----
```

### Bonus : MongoDB Assignment

### Bonus 1:

1.

```
db.student.find({}, {"name.last":1, "course_gpas":{$slice:[5,5]}, "_id":0}).pretty()
```

```

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

{"course_gpas": [
  3.03,
  2.66,
  3.77,
  3.38,
  3.97
]},
{"name": {
  "last": "Martinez"
},
"course_gpas": [
  3.55,
  2.81,
  3.99,
  2.71,
  3.13
]},
{"name": {
  "last": "Ford"
},
"course_gpas": [
  2.22,
  2.23,
  3.6,
  3.85,
  2.09
]},
{"name": {
  "last": "Washington"
},
"course_gpas": [
  2.34,
  3.05,
  3.57
]},
{"name": {
  "last": "Steele"
},
"course_gpas": [
  2.63,
  3.61,
  2.86,
  3.18,
  2.77
]}
}
type "tt" for more
>

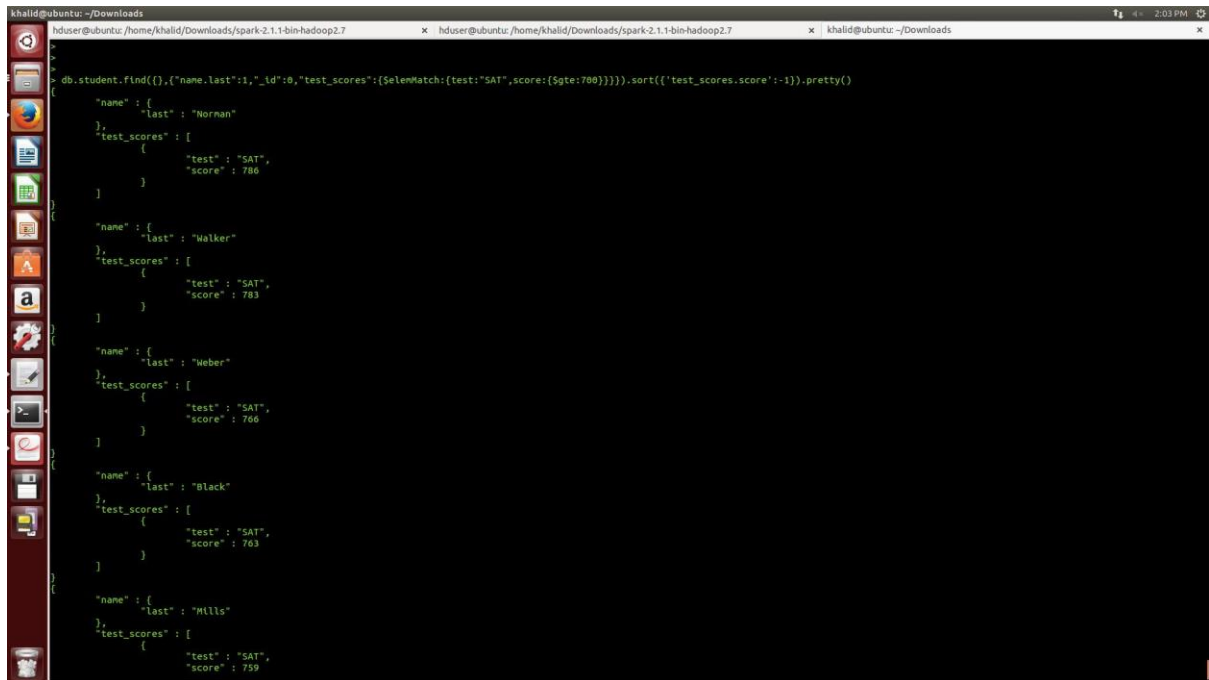
```

(Next Page)



2.

```
db.student.find({},{"name.last":1,"_id":0,"test_scores":{"$elemMatch:{test:"SAT",score:{$gte:700}}}}).  
sort({'test_scores.score':-1}).pretty()
```

A terminal window on a Linux system (Ubuntu) showing the execution of a MongoDB query. The query filters for students with an SAT score of 700 or higher and sorts them by score in descending order. The results are displayed in a pretty-printed JSON format. The terminal window has a dark background and a light-colored text. The query is entered at the prompt, and the results are shown below it. The results are a list of five objects, each representing a student with their name, last name, and test scores.

```
khalid@ubuntu: ~/Downloads  
hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7 x hduser@ubuntu: /home/khalid/Downloads/spark-2.1.1-bin-hadoop2.7 x khalid@ubuntu: ~/Downloads  
db.student.find({},{"name.last":1,"_id":0,"test_scores":{"$elemMatch:{test:"SAT",score:{$gte:700}}}}).sort({'test_scores.score':-1}).pretty()  
{  
  "name" : {  
    "last" : "Norman"  
  },  
  "test_scores" : [  
    {  
      "test" : "SAT",  
      "score" : 786  
    }  
  ]  
},  
{  
  "name" : {  
    "last" : "Walker"  
  },  
  "test_scores" : [  
    {  
      "test" : "SAT",  
      "score" : 783  
    }  
  ]  
},  
{  
  "name" : {  
    "last" : "Heber"  
  },  
  "test_scores" : [  
    {  
      "test" : "SAT",  
      "score" : 766  
    }  
  ]  
},  
{  
  "name" : {  
    "last" : "Black"  
  },  
  "test_scores" : [  
    {  
      "test" : "SAT",  
      "score" : 763  
    }  
  ]  
},  
{  
  "name" : {  
    "last" : "Hills"  
  },  
  "test_scores" : [  
    {  
      "test" : "SAT",  
      "score" : 759  
    }  
  ]  
}
```

(Next Page)

3.

```
db.student.update({},{$set:{"evaluations":[]}},{upsert:false,multi:true})
```

```
remarks = [{"eval_comment":"This student is very clever"}, {"eval_comment":"This student always  
submits exercises on time"}]
```



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

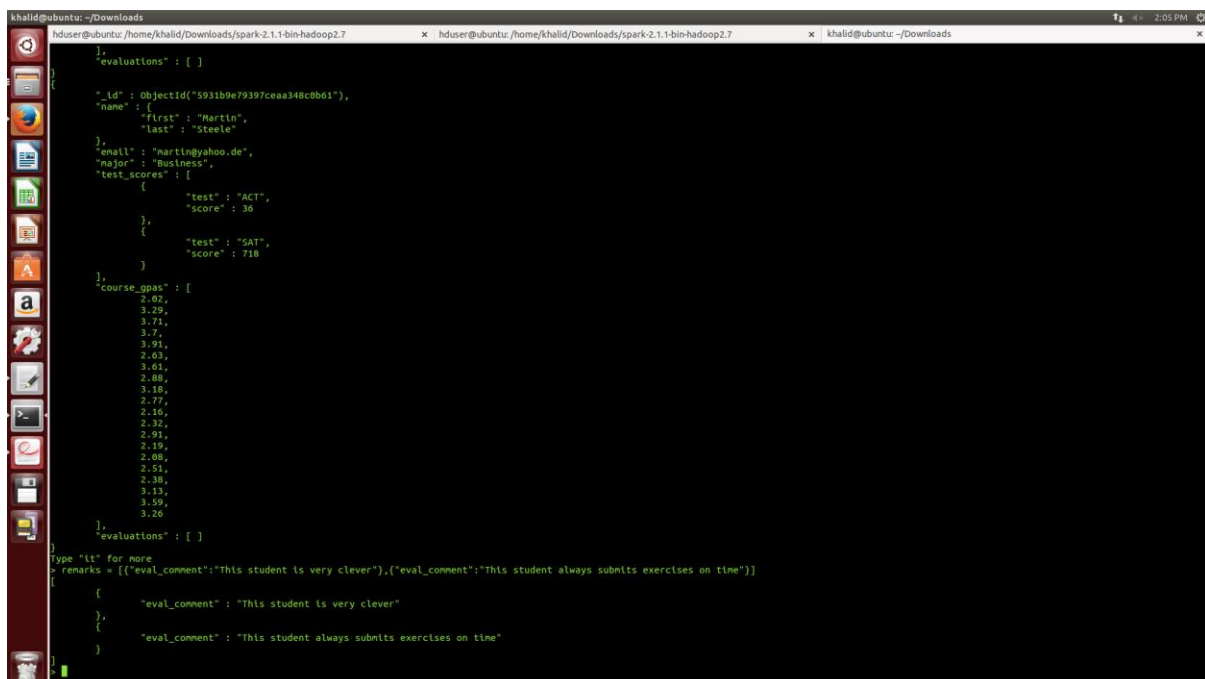
3.6,
2.15,
3.39,
3.16,
2.92,
2.96,
3.41,
3.19,
3.5

},
"evaluations" : [ ]
}

{
  "_id" : ObjectId("5931b9e79397ceaa348c0b60"),
  "name" : {
    "first" : "Virginia",
    "last" : "Washington"
  },
  "email" : "virginiag@mail.de",
  "major" : "Graphic Design",
  "test_scores" : [
    {
      "test" : "ACT",
      "score" : 21
    },
    {
      "test" : "SAT",
      "score" : 424
    }
  ],
  "course_gpas" : [
    2.13,
    3.62,
    3.88,
    3.05,
    3.64,
    2.34,
    3.05,
    3.57
  ],
  "evaluations" : [ ]
}

{
  "_id" : ObjectId("5931b9e79397ceaa348c0b61"),
  "name" : {
    "first" : "Martin",
    "last" : "Steele"
  },
  "email" : "martinyahoo.de",
  "major" : "Business",
  "test_scores" : [
    {
      "test" : "ACT",
      "score" : 36
    },
    {
      "test" : "SAT",
      "score" : 718
    }
  ],
  "course_gpas" : [
    2.02,
    3.29,
    3.71,
    3.7,
    3.91,
    2.63,
    3.01,
    2.88,
    3.18,
    2.77,
    2.16,
    2.32,
    2.91,
    2.19,
    2.08,
    2.51,
    2.38,
    3.13,
    3.59,
    3.26
  ],
  "evaluations" : [ ]
}

type "it" for more
> remarks = [{"eval_comment":"This student is very clever"}, {"eval_comment":"This student always submits exercises on time"}]
{
  {
    "eval_comment" : "This student is very clever"
  },
  {
    "eval_comment" : "This student always submits exercises on time"
  }
}
```



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

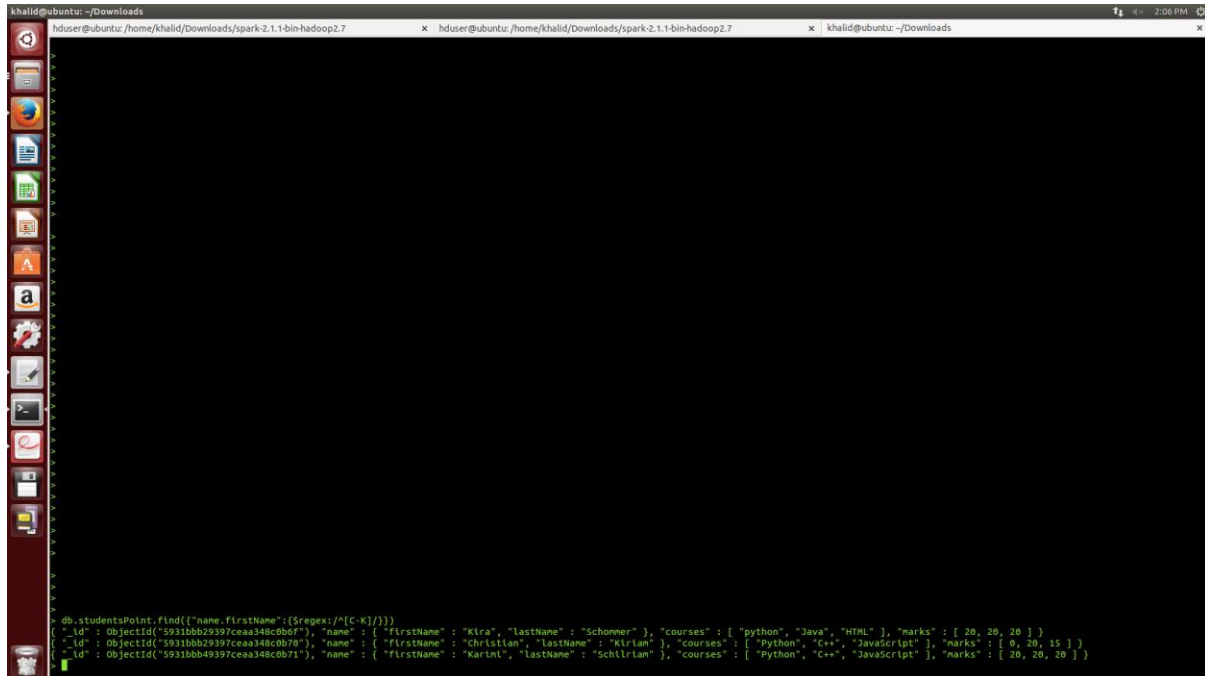
{
  "_id" : ObjectId("5931b9e79397ceaa348c0b61"),
  "name" : {
    "first" : "Martin",
    "last" : "Steele"
  },
  "email" : "martinyahoo.de",
  "major" : "Business",
  "test_scores" : [
    {
      "test" : "ACT",
      "score" : 36
    },
    {
      "test" : "SAT",
      "score" : 718
    }
  ],
  "course_gpas" : [
    2.02,
    3.29,
    3.71,
    3.7,
    3.91,
    2.63,
    3.01,
    2.88,
    3.18,
    2.77,
    2.16,
    2.32,
    2.91,
    2.19,
    2.08,
    2.51,
    2.38,
    3.13,
    3.59,
    3.26
  ],
  "evaluations" : [ ]
}

type "it" for more
> remarks = [{"eval_comment":"This student is very clever"}, {"eval_comment":"This student always submits exercises on time"}]
{
  {
    "eval_comment" : "This student is very clever"
  },
  {
    "eval_comment" : "This student always submits exercises on time"
  }
}
```

## Bonus 2: Regular Expressions

1.

```
db.studentsPoint.find({"name.firstName":{"regex:/^[C-K]/}})
```

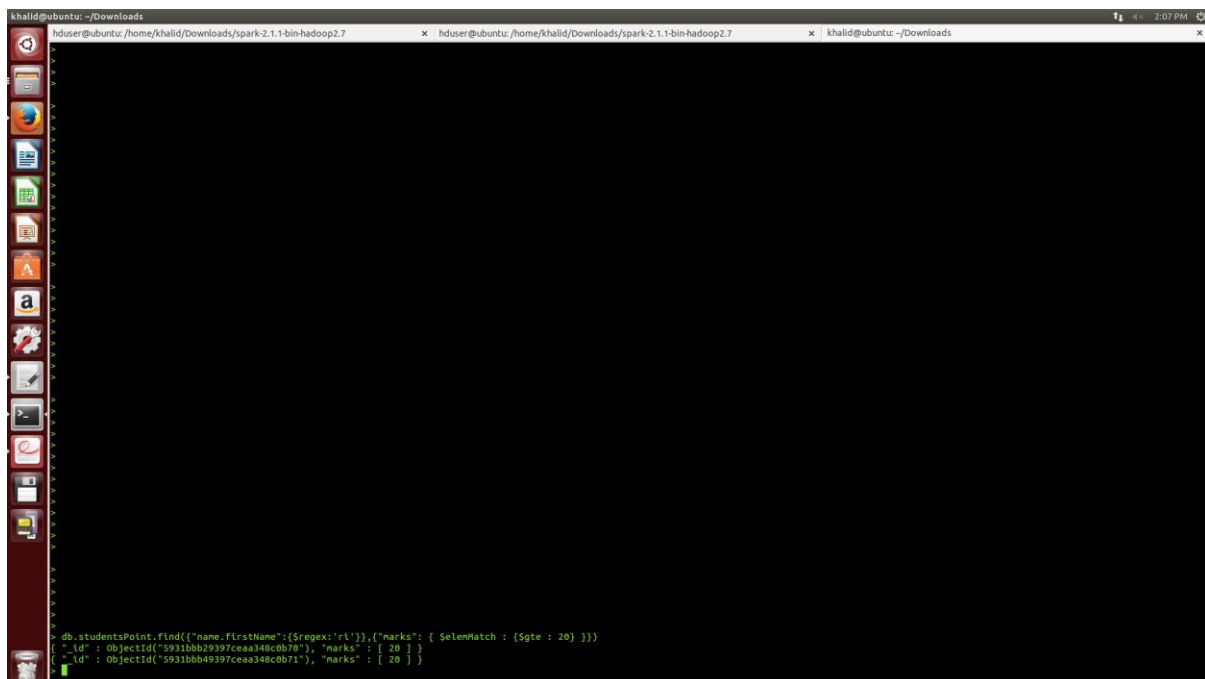


A terminal window on an Ubuntu system showing the execution of a MongoDB query. The query filters students by first name using a regular expression. The output displays three documents with their IDs, names, courses, and marks.

```
db.studentsPoint.find({"name.firstName":{"regex:/^[C-K]/}})
{"_id" : ObjectId("5931bbb29397ceaa348c0b0f"), "name" : { "firstName" : "Kira", "lastName" : "Schonner" }, "courses" : [ "python", "Java", "HTML" ], "marks" : [ 20, 20, 20 ] }
{"_id" : ObjectId("5931bbb29397ceaa348c0b70"), "name" : { "firstName" : "Christian", "lastName" : "Kirlan" }, "courses" : [ "Python", "C++", "JavaScript" ], "marks" : [ 0, 20, 15 ] }
{"_id" : ObjectId("5931bbb49397ceaa348c0b71"), "name" : { "firstName" : "Karlnt", "lastName" : "Schllrlan" }, "courses" : [ "Python", "C++", "JavaScript" ], "marks" : [ 20, 20, 20 ] }
```

3.

```
db.studentsPoint.find({"name.firstName":{"regex:'ri'"}},{"marks": { $elemMatch : { $gte : 20 } }})
```



A terminal window on an Ubuntu system showing the execution of a MongoDB query. The query filters students by first name containing 'ri' and marks greater than or equal to 20. The output displays two documents with their IDs, names, and marks.

```
db.studentsPoint.find({"name.firstName":{"regex:'ri'"}},{"marks": { $elemMatch : { $gte : 20 } }})
{"_id" : ObjectId("5931bbb29397ceaa348c0b70"), "marks" : [ 20 ] }
{"_id" : ObjectId("5931bbb49397ceaa348c0b71"), "marks" : [ 20 ] }
```