CSC 555 - Assignment 5

Nisarg Patel

1)

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

(a) Treating the utility matrix as boolean, compute the Jaccard distance between each pair of users.

Solution :-
A and B have intersection of 4 and union of 8. Thus, Jaccard similarity SIM (A, B) = 4/8 and Jaccard distance DISTANCE (A, B) = 4/8.

A and C have intersection of 4 and union of 8. Thus, Jaccard similarity SIM (A, C) = 4/8 and Jaccard distance DISTANCE (A, C) = 4/8.

B and C have intersection of 4 and union of 8. Thus, Jaccard similarity SIM (B, C) = 4/8 and Jaccard distance DISTANCE (B, C) = 4/8.

(e) Normalize the matrix by subtracting from each nonblank entry the average value for its user.

Solution :-
Normalizing the matrix by subtracting from each nonblank entry with the average value for its user.

A [3/2	5/2	0	5/2	-3/2	0	1/2	-1/2]
B [0	5/4	9/4	5/4	-3/4	1/4	-3/4	0]
C [-1/4	0	-5/4	3/4	0	7/4	11/4	3/4]

After that we will compute cosine distance between each pair
Cosine(A,B) : 0.5465
Cosine(A,C) : 0.1634
Cosine(B,C) : (-0.3125)

b) **Content based filtering and collaborative filtering**
Content based filtering :- This system works with existing profile users. The profile has information about their taste and likes.

In this system when the user is new it gets idea by doing user's survey and then compare the positive rated item that are matching to his taste. Once that is done system will keep recommending based on user's positive ratings to that item and tags that are related to that item. For example, let's say that a particular user bought beer and he rated the beer 3/5. So then system will find some another beer that has 4/5 or 5/5 rating with different brand and then recommend that beer to user based on his previous purchase.

Collaborative filtering :- The collaborative filtering is used for finding users community that share appreciations. Let's say that two user that rated similar items so they might have similar taste. This is called so-called neighborhood. In this system user get recommendation for item if it is appreciated positively by their neighborhood.
It has two approaches

1) User based approach :- In this approach items are recommended based on an evaluation of items by its neighborhood.
2) Item based approach :- Refereeing that taste of users remain constant or slightly changed. Similar items build the neighborhood based on the user appreciation.

c)

To make matrix less sparse we can use collaborative filtering(CF) as mention above but it might not work all the time but in case of sparsity these technique might not work so we will try another method called **Baseline Prediction method**.

In **Baseline prediction method** we will average the rating value for that item that is rated by other user.

Knowledge base filtering is another method that we can use. In this method we only require domain specific knowledge and custom engineering.
It is product of User features and Item features. This method has major strength for existence of cold-start problems and problem with sparsity.

2)
Given the input data [(1pm, $10), (2pm, $15), (3pm, $15), (4pm, $20), (5pm, $10), (6pm, $20), (7pm, $30), (8pm, $25), (9pm, $25), (10pm, $30), (11pm, $30)].
a) What will the Hive (or Oracle/SQL) query "compute average price" return? (yes, this is as obvious as it seems, asked for comparison with following parts)
Solution: Hive or oracle/sql query for average price will be 20.9090.

b) What will a Storm query "compute average price per each 3 hour window" return? (tumbling, i.e., non-overlapping window of tuples, as many as you can fit).
Solution:
(1,10) (2,15) (3,15) (4,20) (5,10) (6,20) (7,30) (8,25) (9,25) (10,30) (11,30)

Now we will compute average price per each 3 hour window (non-overlapping window)

1	2	3
10	15	15

Avg price = 13.333

4	5	6
20	10	20

avg price = 16.666667

7	8	9
30	25	25

avg price = 26.666667

10	11	
30	30	

avg price = 20

c) What will a Storm query "compute average price per each 3 hour window" return? (sliding, i.e. overlapping window of tuples, moving the window forward 2 hours each time)

1	2	3
10	15	15

avg price = 13.333

3	4	5
15	20	10

avg price = 15

5	6	7
10	20	30

avg price = 20

7	8	9
30	25	25

avg price = 26.666667

9	10	11
25	30	30

Avg price = 28.33333

3)

a)

Installing spark on master and worker node. Adding slaves to master node in conf/slaves file and adding SPARK_HOME file to .bashrc file.

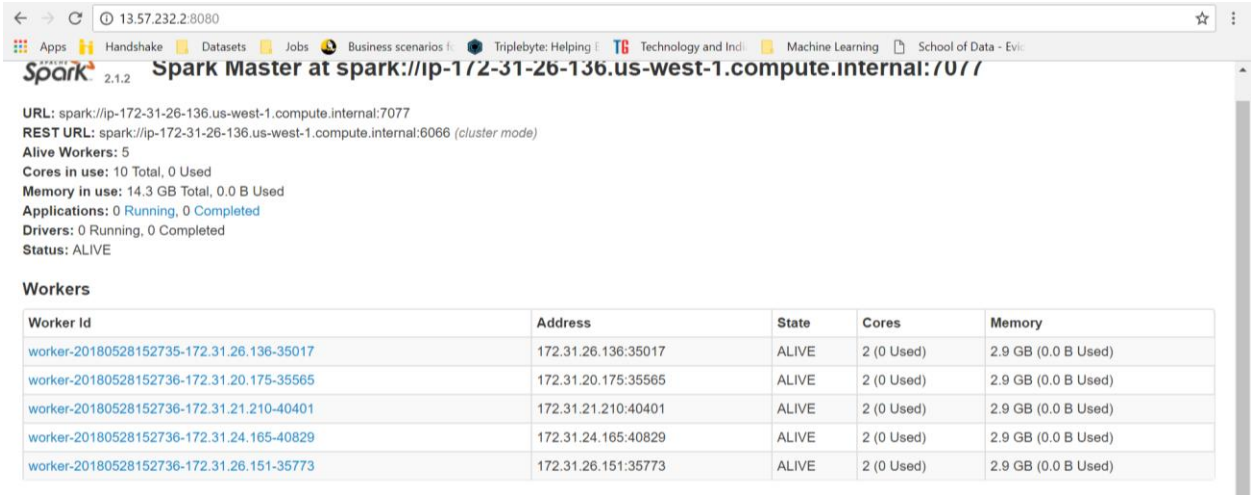
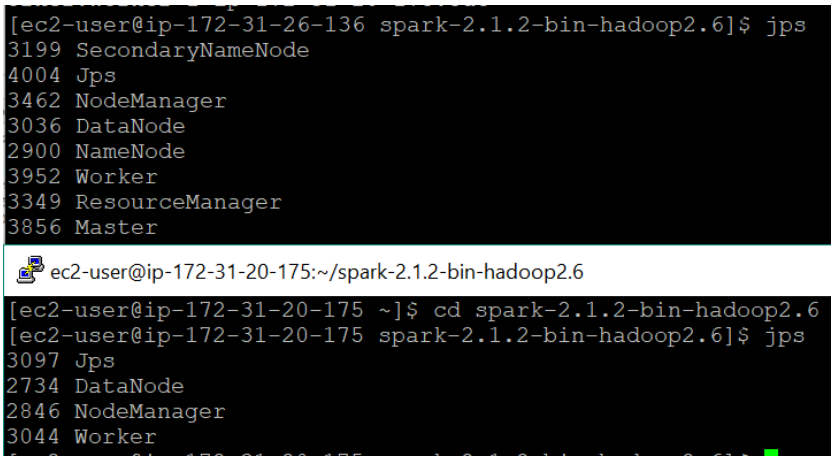


Fig : 5 node spark cluster



b)

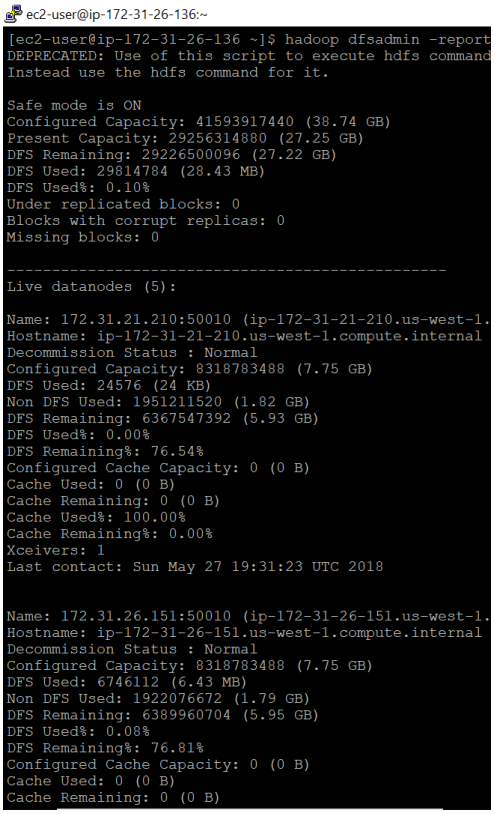
Resilient Distributed Dataset (RDD) is the fundamental unit of data in Apache Spark, which is a distributed collection of elements across cluster nodes and can perform parallel operations.

Spark operates on data in fault tolerant file systems like HDFS. In this, the data get replicated on one other node so that it can retrieve when a failure occurs.

4)

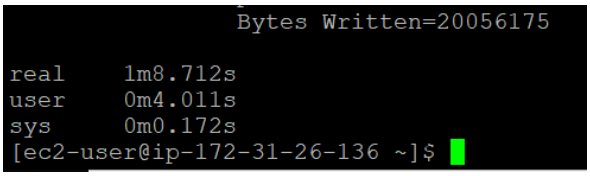
a)

Making current 4 node cluster to 5 node cluster.

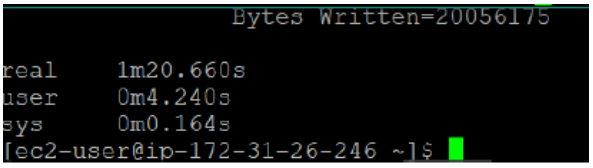


b)

time take by 5 node cluster.



time taken by 4 node cluster.



c)

After shutting down one of the node time taken by 4 node cluster increases compare to 5 node cluster.

It didn't generate any error.

```
        Map-Reduce Framework
                Map input records=5284546
                Map output records=18562366
                Map output bytes=279356680
                Map output materialized bytes=26902454
                Input split bytes=210
                Combine input records=20053191
                Combine output records=2673165
                Reduce input groups=1040390
                Reduce shuffle bytes=26902454
                Reduce input records=1182340
                Reduce output records=1040390
                Spilled Records=3855505
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=623
                CPU time spent (ms)=42140
                Physical memory (bytes) snapshot=769323008
                Virtual memory (bytes) snapshot=2975682560
                Total committed heap usage (bytes)=559415296
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=231153099
        File Output Format Counters
                Bytes Written=20056175

real    1m42.019s
user    0m3.880s
sys     0m0.210s
[ec2-user@ip-172-31-26-136 ~]$
```

d)

Removing <property> tag from core-site.xml configuration file.

```
  GNU nano 2.5.3             File: core-site.xml                    Modi

  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>

<name>fs.defaultFS</name>
  <value>hdfs://172.31.26.136/</value>
</property>
</configuration>
```

Showing error when we start hadoop again after changing configuration file.

```
[ec2-user@ip-172-31-26-246 hadoop-2.6.4]$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
[Fatal Error] core-site.xml:24:3: The element type "configuration" must be termi
nated by the matching end-tag "</configuration>".
18/05/27 23:46:06 FATAL conf.Configuration: error parsing conf core-site.xml
org.xml.sax.SAXParseException; systemId: file:/home/ec2-user/hadoop-2.6.4/etc/ha
doop/core-site.xml; lineNumber: 24; columnNumber: 3; The element type "configura
tion" must be terminated by the matching end-tag "</configuration>".
```

5)

- bin/mahout splitDataset --input movielens/ratings.csv --output ml_dataset -- trainingPercentage 0.9 --probePercentage 0.1 --tempDir dataset/tmp

```
18/05/28 03:08:24 INFO Job:  map 100% reduce 0%
18/05/28 03:08:26 INFO Job: Job job_1527475426499_0003 completed successfully
18/05/28 03:08:26 INFO Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=107086
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=21770226
                HDFS: Number of bytes written=1159137
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=59046
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=59046
                Total vcore-milliseconds taken by all map tasks=59046
                Total megabyte-milliseconds taken by all map tasks=60463104
        Map-Reduce Framework
                Map input records=1000209
                Map output records=100344
                Input split bytes=142
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=803
                CPU time spent (ms)=22450
                Physical memory (bytes) snapshot=174280704
                Virtual memory (bytes) snapshot=1000816640
                Total committed heap usage (bytes)=105381888
        File Input Format Counters
                Bytes Read=21770084
        File Output Format Counters
                Bytes Written=1159137
18/05/28 03:08:26 INFO MahoutDriver: Program took 420108 ms (Minutes: 7.0018)
[ec2-user@ip-172-31-26-136 apache-mahout-distribution-0.11.2]$
```

```
[ec2-user@ip-172-31-26-136 ~]$ hadoop fs -ls ml_dataset/probeSet
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2018-05-28 03:08 ml_dataset/probeSet/_SUCCESS
-rw-r--r--   2 ec2-user supergroup    1159137 2018-05-28 03:08 ml_dataset/probeSet/part-m-00000
[ec2-user@ip-172-31-26-136 ~]$ hadoop fs -ls ml_dataset/trainingSet
Found 2 items
-rw-r--r--   2 ec2-user supergroup          0 2018-05-28 03:06 ml_dataset/trainingSet/_SUCCESS
-rw-r--r--   2 ec2-user supergroup   10394319 2018-05-28 03:06 ml_dataset/trainingSet/part-m-00000
[ec2-user@ip-172-31-26-136 ~]$ hadoop fs -ls movielens/ratings.csv
-rw-r--r--   2 ec2-user supergroup   11553456 2018-05-28 02:56 movielens/ratings.csv
[ec2-user@ip-172-31-26-136 ~]$
```

- time bin/mahout parallelALS --input ml_dataset/trainingSet/ --output als/out --tempDir als/tmp --numFeatures 20 --numIterations 3 --lambda 0.065

```
        Map-Reduce Framework
                Map input records=3693
                Map output records=3693
                Input split bytes=133
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=45
                CPU time spent (ms)=3340
                Physical memory (bytes) snapshot=179945472
                Virtual memory (bytes) snapshot=992460800
                Total committed heap usage (bytes)=91750400
        File Input Format Counters
                Bytes Read=8229392
        File Output Format Counters
                Bytes Written=648819
18/05/28 15:41:50 INFO MahoutDriver: Program took 166853 ms (Minutes: 2.780883333333333)

real    2m51.974s
user    0m11.078s
sys     0m2.832s
```

Time taken = 2 minutes 51 second

- bin/mahout evaluateFactorization --input ml_dataset/probeSet/ --output als/rmse/ --userFeatures als/out/U/ --itemFeatures als/out/M/ --tempDir als/tmp

```
18/05/28 15:45:12 INFO Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=107305
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1159269
                HDFS: Number of bytes written=96
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=4671
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4671
                Total vcore-milliseconds taken by all map tasks=4671
                Total megabyte-milliseconds taken by all map tasks=4783104
        Map-Reduce Framework
                Map input records=100344
                Map output records=0
                Input split bytes=132
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=35
                CPU time spent (ms)=1180
                Physical memory (bytes) snapshot=172167168
                Virtual memory (bytes) snapshot=995819520
                Total committed heap usage (bytes)=105381888
        File Input Format Counters
                Bytes Read=1159137
        File Output Format Counters
                Bytes Written=96
18/05/28 15:45:12 INFO MahoutDriver: Program took 16077 ms (Minutes: 0.26795)
[ec2-user@ip-172-31-26-136 apache-mahout-distribution-0.11.2]$
```

- RMSE value

```
                Bytes Written=1621416
18/05/28 15:53:10 INFO MahoutDriver: Program took 15619 ms (Minutes: 0.26033333333333336)
[ec2-user@ip-172-31-26-136 apache-mahout-distribution-0.11.2]$ hadoop fs -ls als/rmse1/rmse.txt
-rw-r--r--   2 ec2-user supergroup         17 2018-05-28 15:53 als/rmse1/rmse.txt
[ec2-user@ip-172-31-26-136 apache-mahout-distribution-0.11.2]$ hadoop fs -cat als/rmse1/rmse.txt
0.8879398701928 92[ec2-user@ip-172-31-26-136 apache-mahout-distribution-0.11.2]$
```

RMSE value = 0.8879

- bin/mahout recommendfactorized --input als/out/userRatings/ --output recommendations/ --userFeatures als/out/U/ --itemFeatures als/out/M/ --numRecommendations 6 --maxRating 5

$HADOOP_HOME/bin/hadoop fs -cat recommendations/part-m-00000 | head

```
[ec2-user@ip-172-31-26-136 apache-mahout-distribution-0.11.2]$ hadoop fs -cat recommendations1/part-m-00000 | head
1       [572:5.0,1198:4.5111012,1226:4.451013,2762:4.428101,318:4.4273634,356:4.4218764]
2       [572:4.9995685,527:4.4384665,2762:4.2802663,953:4.235685,2028:4.2116976,1961:4.1653876]
3       [572:5.0,2332:4.7771373,2834:4.6750913,297:4.5984774,2762:4.5095468,2028:4.4464846]
4       [858:5.0,2360:5.0,3823:5.0,3925:5.0,1263:5.0,1221:5.0]
5       [1420:4.5614343,1423:4.4033694,2362:4.296197,2483:4.2709947,3608:4.245638,2360:4.242081]
6       [572:5.0,2101:5.0,2197:4.8744535,3585:4.752102,3853:4.6765885,2175:4.579724]
7       [1198:4.7477117,1262:4.684964,260:4.661096,572:4.651185,1283:4.631059,1272:4.629972]
8       [318:4.6216855,2905:4.607397,1218:4.5965147,1148:4.5765195,2494:4.561526,50:4.558065]
9       [260:4.3284574,2905:4.3178415,1196:4.3084817,1198:4.29593,296:4.2230816,858:4.212307]
10      [572:5.0,2197:4.8247514,2565:4.504516,3147:4.4493046,1907:4.4007344,1207:4.3625064]
cat: Unable to write to output stream.
[ec2-user@ip-172-31-26-136 apache-mahout-distribution-0.11.2]$
```

- Top movie recommendation by movie ID for user 4,5,6

4 [858:5.0,2360:5.0,3823:5.0,3925:5.0,1263:5.0,1221:5.0]

5 [1420:4.561, 1423:4.403, 2362:4.296, 2483:4.270, 3608:4.245, 2360:4.24]

6 [572:5.0,2101:5.0,2197:4.8744535,3585:4.752102,3853:4.6765885,2175:4.579724]