CASE STUDY 1

1. **a)** To move the **ratings.csv** and **movies.csv** file into hadoop fs from local.
   - ➔ hadoop fs –put /movies.csv /
   - ➔ hadoop fs –put /ratings.csv /

```
acadgild@localhost:~
6102 ResourceManager
6231 Jps
5641 NameNode
4827 org.eclipse.equinox.launcher_1.4.0.v20161219-1356.jar
5739 DataNode
6204 NodeManager
[acadgild@localhost ~]$ hadoop fs -ls /
18/05/10 12:43:25 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 6 items
-rw-r--r--   1 acadgild supergroup    2283410 2018-05-09 18:16 /movies.csv
-rw-r--r--   1 acadgild supergroup  709550327 2018-05-09 18:17 /ratings.csv
drwxr-xr-x   - acadgild supergroup          0 2018-05-08 12:16 /resources
drwxr-xr-x   - acadgild supergroup          0 2018-02-02 12:49 /sqoopout111
drwxrwx---   - acadgild supergroup          0 2018-02-09 11:35 /tmp
drwxr-xr-x   - acadgild supergroup          0 2018-05-08 13:12 /user
[acadgild@localhost ~]$ ls
CaseStudy1.jar    Documents  eclipse-workspace  Music     source code  Videos
CaseStudyDataset  Downloads  flumeconf          Pictures  spooldir
Desktop           eclipse    install            Public    Templates
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop jar CaseStudy1.jar /movies.csv /ratings.csv caseO
utput1
18/05/10 12:44:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
18/05/10 12:44:13 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/05/10 12:44:16 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool
oolRunner to remedy this.
18/05/10 12:44:17 INFO input.FileInputFormat: Total input paths to process : 1
18/05/10 12:44:18 INFO input.FileInputFormat: Total input paths to process : 1
18/05/10 12:44:18 INFO mapreduce.JobSubmitter: number of splits:7
18/05/10 12:44:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1525936362839_0001
18/05/10 12:44:20 INFO impl.YarnClientImpl: Submitted application application_1525936362839_0001
18/05/10 12:44:20 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1525936362839_0001
18/05/10 12:44:20 INFO mapreduce.Job: Running job: job_1525936362839_0001
```

1. **b)** Report the hdfs block created by both the files (movies.csv and ratings.csv) on HDFS.
   - ➔ **Command used :** hadoop fsck /movies.csv
   - ➔ **Command used :** hadoop fsck /ratings.csv

➔ /movies.csv

```
acadgild@localhost:~
-rw-r--r--   1 acadgild supergroup  709550327 2018-05-09 18:17 /ratings.csv
drwxr-xr-x   - acadgild supergroup          0 2018-05-08 12:16 /resources
drwxr-xr-x   - acadgild supergroup          0 2018-02-02 12:49 /sqoopout111
drwxrwx---   - acadgild supergroup          0 2018-02-09 11:35 /tmp
drwxr-xr-x   - acadgild supergroup          0 2018-05-08 13:12 /user
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop fsck /movies.csv
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

18/05/10 19:00:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
 classes where applicable
Connecting to namenode via http://localhost:50070
FSCK started by acadgild (auth:SIMPLE) from /127.0.0.1 for path /movies.csv at Thu May 10 19:00:20 IST 2018
.Status: HEALTHY
 Total size:    2283410 B
 Total dirs:    0
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      1 (avg. block size 2283410 B)
 Minimally replicated blocks:   1 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          1
 Number of racks:               1
FSCK ended at Thu May 10 19:00:21 IST 2018 in 1094 milliseconds


The filesystem under path '/movies.csv' is HEALTHY
[acadgild@localhost ~]$
```

➔ /ratings.csv

```
acadgild@localhost:~
[acadgild@localhost ~]$ hadoop fsck /ratings.csv
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

18/05/10 19:00:50 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
 classes where applicable
Connecting to namenode via http://localhost:50070
FSCK started by acadgild (auth:SIMPLE) from /127.0.0.1 for path /ratings.csv at Thu May 10 19:00:53 IST 2018
.Status: HEALTHY
 Total size:    709550327 B
 Total dirs:    0
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      6 (avg. block size 118258387 B)
 Minimally replicated blocks:   6 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          1
 Number of racks:               1
FSCK ended at Thu May 10 19:00:53 IST 2018 in 7 milliseconds


The filesystem under path '/ratings.csv' is HEALTHY
[acadgild@localhost ~]$
```

2. Join the two tables using reduce side join and find out?
   a). the movies that user has rated?
   b). How many times a movie has been rated?
   c). the average rating given for a movie?

Sol: -> write a Mapper, reducer and a driver program.

->export the jar file and execute it on hadoop

->**command used** -> hadoop jar CaseStudy.jar /movies.csv /ratings.csv CaseOutput

```
acadgild@localhost:~
6102 ResourceManager
6231 Jps
5641 NameNode
4827 org.eclipse.equinox.launcher_1.4.0.v20161219-1356.jar
5739 DataNode
6204 NodeManager
[acadgild@localhost ~]$ hadoop fs -ls /
18/05/10 12:43:25 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 6 items
-rw-r--r--   1 acadgild supergroup    2283410 2018-05-09 18:16 /movies.csv
-rw-r--r--   1 acadgild supergroup  709550327 2018-05-09 18:17 /ratings.csv
drwxr-xr-x   - acadgild supergroup          0 2018-05-08 12:16 /resources
drwxr-xr-x   - acadgild supergroup          0 2018-02-02 12:49 /sqoopout111
drwxrwx---   - acadgild supergroup          0 2018-02-09 11:35 /tmp
drwxr-xr-x   - acadgild supergroup          0 2018-05-08 13:12 /user
[acadgild@localhost ~]$ ls
CaseStudy1.jar    Documents  eclipse-workspace  Music     source code  Videos
CaseStudyDataset  Downloads  flumeconf          Pictures  spooldir
Desktop           eclipse    install            Public    Templates
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ hadoop jar CaseStudy1.jar /movies.csv /ratings.csv caseO
utput1
18/05/10 12:44:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
18/05/10 12:44:13 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
18/05/10 12:44:16 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool
oolRunner to remedy this.
18/05/10 12:44:17 INFO input.FileInputFormat: Total input paths to process : 1
18/05/10 12:44:18 INFO input.FileInputFormat: Total input paths to process : 1
18/05/10 12:44:18 INFO mapreduce.JobSubmitter: number of splits:7
18/05/10 12:44:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1525936362839_0001
18/05/10 12:44:20 INFO impl.YarnClientImpl: Submitted application application_1525936362839_0001
18/05/10 12:44:20 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1525936362839_0001
18/05/10 12:44:20 INFO mapreduce.Job: Running job: job_1525936362839_0001
```

To check the output, command used:

➔ hadoop fs –cat caseOutput/part-r-0000 | head

```
acadgild@localhost:~
[acadgild@localhost ~]$ hadoop fs -cat caseOutput/part-r-00000 | head
18/05/10 18:54:06 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
Toy Story (1995)        66008   3.888157
GoldenEye (1995)        32534   3.431841
City Hall (1996)        4436    3.232304
Curdled (1996)  217     3.099078
"Comic  1       4.000000
Up in Smoke (1957)      3       3.666667
First Daughter (1999)   3       3.333333
"Flaw   14      3.714286
Battle of Los Angeles (2011)    44      2.522727
Jason Becker: Not Dead Yet (2012)       9       3.444444
cat: Unable to write to output stream.
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```