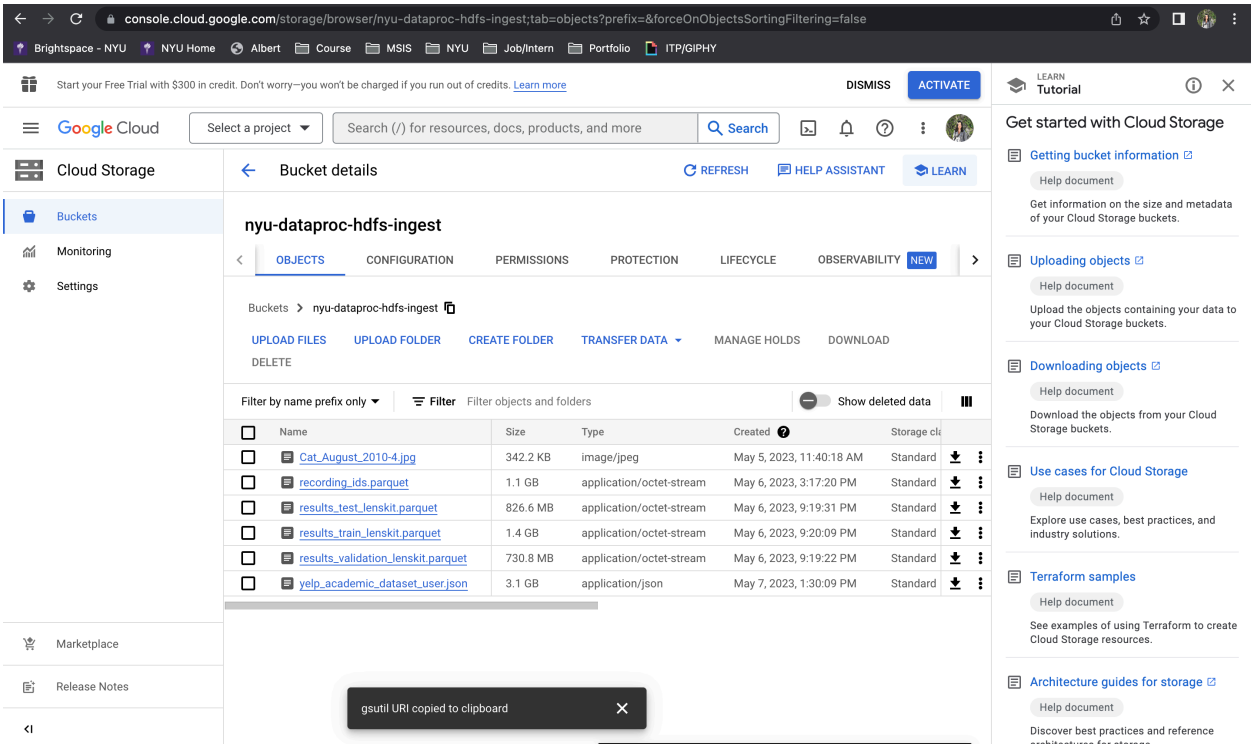
Upload large file/dataset to data ingest website



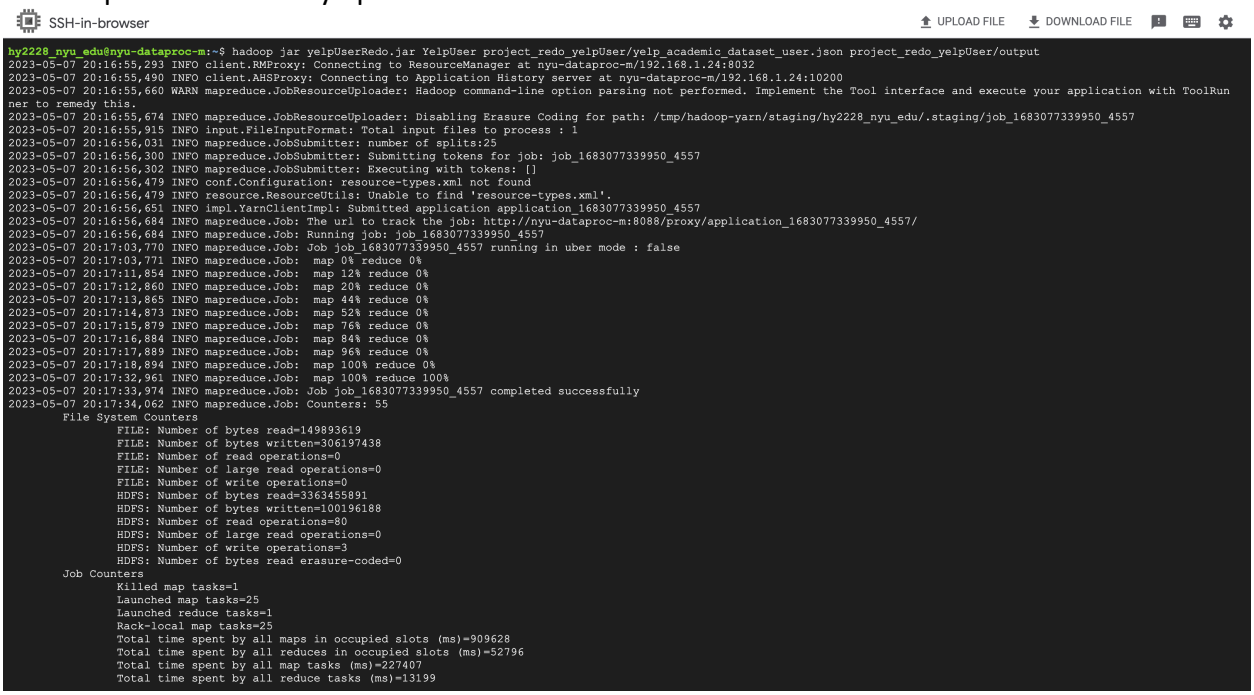
Copy large file from data ingest website to Hadoop

```
hy2228_nyu_edu@nyu-dataproc-m:~$ hadoop fs -ls project_redo_yelpUser
hy2228_nyu_edu@nyu-dataproc-m:~$ hadoop distcp gs://nyu-dataproc-hdfs-ingest/yelp_academic_dataset_user.json /user/hy2228_nyu_edu/project_redo_yelpUser
```

SSH-in-browser UPLOAD FILE DOWNLOAD FILE

```
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                GS: Number of bytes read=3363329011
                GS: Number of bytes written=0
                GS: Number of read operations=410563
                GS: Number of large read operations=0
                GS: Number of write operations=0
                HDFS: Number of bytes read=454
                HDFS: Number of bytes written=3363329011
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=6
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=36798
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=36798
                Total vcore-milliseconds taken by all map tasks=36798
                Total megabyte-milliseconds taken by all map tasks=37681152
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=145
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=240
                CPU time spent (ms)=27700
                Physical memory (bytes) snapshot=471797760
                Virtual memory (bytes) snapshot=2604171264
                Total committed heap usage (bytes)=584056832
                Peak Map Physical memory (bytes)=471797760
                Peak Map Virtual memory (bytes)=2604171264
        File Input Format Counters
                Bytes Read=309
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bandwidth in Btyes=101919060
                Bytes Copied=3363329011
                Bytes Expected=3363329011
                Files Copied=1
hy2228_nyu_edu@nyu-dataproc-m:~$ hadoop fs -ls project_redo_yelpUser
Found 1 items
-rw-r--r--   1 hy2228_nyu_edu hy2228_nyu_edu 3363329011 2023-05-07 19:13 project_redo_yelpUser/yelp_academic_dataset_user.json
hy2228_nyu_edu@nyu-dataproc-m:~$
```

Run MapReduce Job for yelp user dataset



print data in the output file

SSH-in-browser   UPLOAD FILE   DOWNLOAD FILE

```
Yddh8b_W_MZVRa0d3x0nWA,Grace,3,2017,2.67,226,1,0,0
Yddwys-W5temHj6mlVqL7A,Emily,15,2014,4.56,0,9,0,2
Yde7auaYhHouAeTJaIVcvw,Tonya,2,2016,1.0,31,11,5,4
Yde7o82OgcUF0iDL8zs0IQ,Cait,1,2017,1.0,83,3,0,0
YdeHd6AGKAm2DYpHpqPcmg,Joe,17,2014,3.22,2,8,0,1
Yde_wHV9xQuKn9HZ5pkCtA,Jennifer,23,2012,3.17,0,30,6,3
Ydeosx4EQ_6cKaeEvyGbAQ,Cody,3,2014,5.0,432,0,0,0
Ydf0mX45ZaNn-Gyx-k3-ZQ,Ann,30,2009,4.16,0,32,3,10
YdfTLVKn3kcN-0j7EHuUgQ,Bryanne,5,2019,3.2,0,3,0,0
YdfVJzwt9Oc_hC8pkvGOPQ,Abigail,1,2019,4.0,78,1,0,0
YdfWo4gEHa8C7pff_Gs2Hw,Dimitrios J.,3,2014,4.0,0,0,0,0
YdfwO5gvtpkLcLLoS6L3wQ,Lauren,3,2015,5.0,0,0,0,0
Ydg-RM5vMSvVcoi6m-tPdQ,Alex,3,2013,2.67,0,2,0,1
YdgBOkVqjDLmTyVzHzUKaQ,C.,3,2008,2.67,2,4,1,1
YdgFwPbuY3_5MTc2RhbOFQ,Tim,8,2014,4.75,1,3,0,2
YdgV35r9mleGCQq06x1_lA,Hey Itz,3,2019,5.0,0,2,0,0
YdgXbgmTX8hNzHSBA82vCg,Isa,19,2008,3.95,0,8,4,5
YdgwfgvaWdMtsvOkgDYzvg,Lauren,8,2012,4.38,21,15,0,2
YdhAseO_SKdmjW7ErqQ9cA,Chris,14,2012,3.71,1,8,0,0
YdhIk9yDUclAf_vwccPkrg,Pete,4,2014,5.0,1,2,0,0
YdhIqQJ_4NJO4L3qUIYAxg,Julia,7,2017,5.0,1,1,1,1
YdhJA0x-2WmcLagSRyMl2Q,Linda,3,2013,2.33,15,1,1,0
Ydi9l5cq53pSKLUgrSb1Zg,Vania,2,2020,5.0,8,2,1,1
YdiSDXdkzg0tabmQWqGYIA,Evan,3,2012,2.67,0,6,1,0
YdiWiEHVM4qcjS5cBaXGww,Tanisha,4,2018,5.0,0,0,0,0
YdiXRdWqS_-G00xJ_96Mhg,Bill,1,2013,2.0,0,0,1,0
YdiZnp5_rc00SWUZybeXgg,Maria,299,2010,2.93,127,1132,269,490
Ydih5JgJiU6Lbw2D6ewk3A,Amanda,1,2019,5.0,0,0,0,0
Ydj0M5yV6ckz9wuOMoPJJw,Phil,4,2014,4.0,1,4,1,0
YdjEo7pzTN-tMwyRYNJ5iA,Christina,1,2021,5.0,8,0,0,0
YdjI_gE_oSBsX7ZJ5s8WLA,Alyssa,2,2017,2.67,0,2,1,0
YdjMD7LAZzreX9RYc7bI2g,Jillian,2,2014,5.0,330,1,1,1
YdjUOYQo1FoXAVwxWVs8vg,Chris,1,2013,5.0,0,0,1,0
YdjbbE9hyrtwdwo4DClQJw,Jarrin,35,2009,4.11,721,35,10,19
Ydk0DQ45-gS5qMhEB98i_g,Sean,1,2019,5.0,86,0,0,0
Ydk90LpzdiEpYO19HXxDDg,David,3,2013,3.33,1,1,0,0
YdkMUu0Bt1fpmH8sAX34yQ,A,7,2021,3.86,0,1,0,2
YdkaQc8cKGzJRkgariyn4w,Anna,9,2017,4.0,196,6,0,0
Ydki1Lq_w24UurFVTXbkQw,Matthew,7,2011,2.88,337,9,6,2
Ydkifa6_xw1Esj8pSFw75A,Andrew,3,2012,4.67,1,0,0,0
YdkrpfzDDJeUkPLv8Oa8Zg,Gordon,4,2013,5.0,9,1,0,0
YdkvXQUeMbqjz982oLkL2Q,Andrea,2,2014,5.0,0,0,0,0
Ydkw_7lwodzAOwHdLhOMAQ,Tom,4,2016,4.0,0,0,0,0
Ydl9P20pxEPRoPJxCPMOOg,Kim,1,2012,5.0,0,0,0,0
YdlBh4sx4nuwRJ47Gv5NXw,James,3,2015,2.0,0,1,4,0
YdlGnwRe53qoNK_P1eoVpQ,Chad,13,2011,4.36,2,6,0,2
YdlHkIOPqMkPVw_3u5QqkA,Nikki,12,2020,4.67,0,0,0,0
YdlTnJXyGw7wMKNpDUE_Jw,Naomi,1,2021,1.0,0,1,0,0
Ydle-OafBVgiya2_vV72rA,Jared,3,2012,3.67,20,0,0,0
```

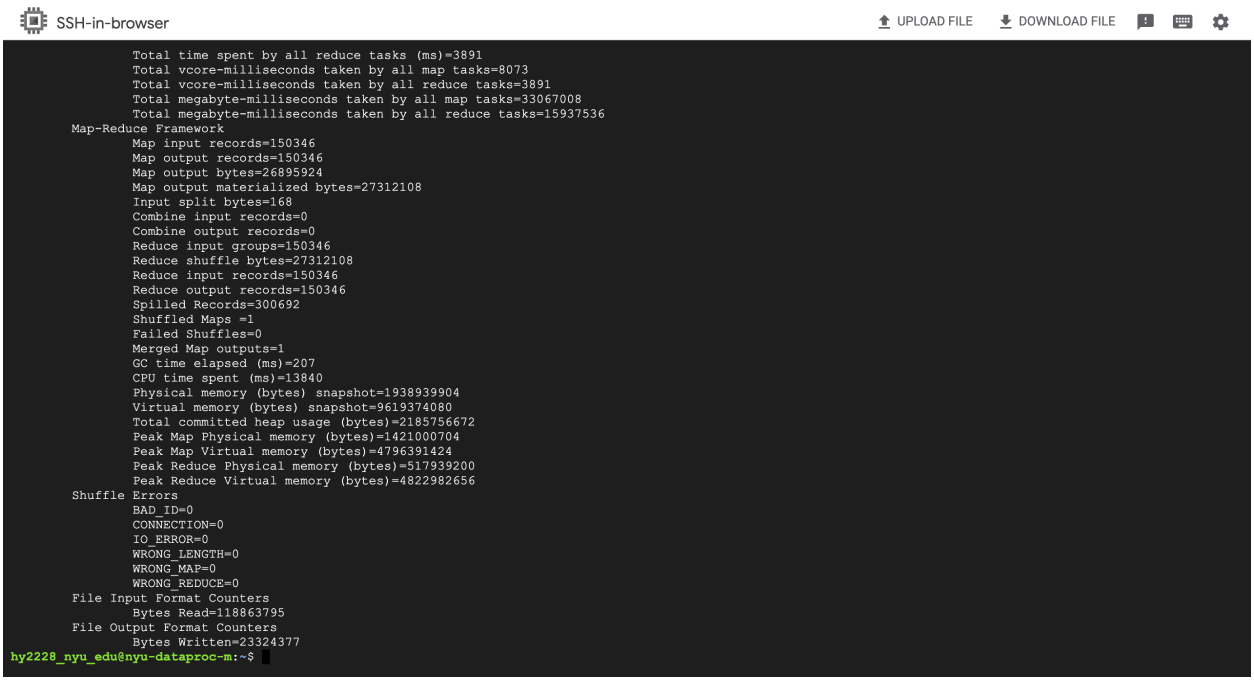
get the file from hadoop

```
hy2228_nyu_edu@nyu-dataproc-m:~$ hadoop fs -get project_redo_yelpUser/output/part-r-00000
hy2228_nyu_edu@nyu-dataproc-m:~$ ls
 YelpUserProfile.jar   pagerankInputs.txt      yelpBusiness.txt                 yelpUserCheck.jar                      yelp_academic_dataset_user_inputs.json
 maxTempRedo.jar       part-r-00000           'yelpJoinedDataFL(woHeader).txt'  yelpUserDataClean.jar                  yelpuserInputs.json
 merged_output.txt     project_output          yelpJoinedDataFL.txt             yelpUserProfileCheck.jar
 merged_output2.txt    smallWeather1.txt       yelpReview.txt                   yelpUserRedo.jar
 pageRank20.jar        temperatureInputs.txt   yelpUser.txt                     yelp_academic_dataset_user.json.csupload
```

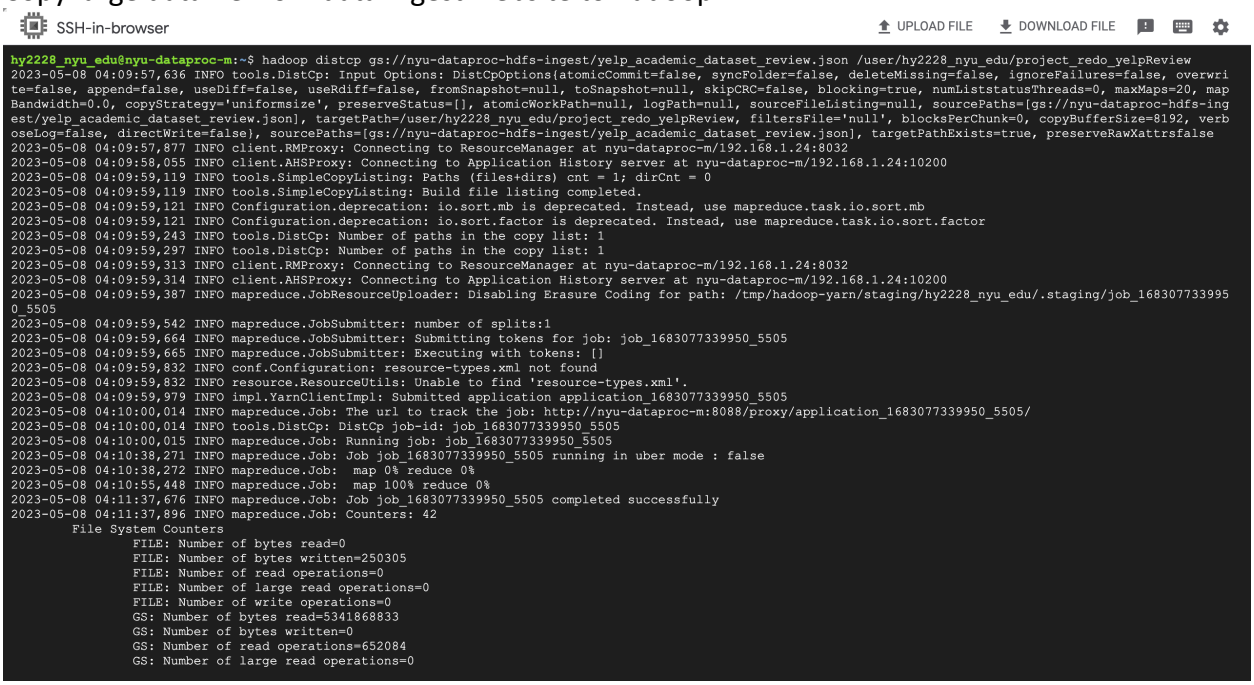
Run MR for business dataset

SSH-in-browser   UPLOAD FILE   DOWNLOAD FILE

```
hy2228_nyu_edu@nyu-dataproc-m:~$ hadoop jar yelpBusinessRedo.jar YelpBusiness project_redo_yelpBusiness/yelp_academic_dataset_business.json project_redo_yelpBus
iness/output
2023-05-08 03:15:07,236 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.24:8032
2023-05-08 03:15:07,421 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.24:10200
2023-05-08 03:15:07,703 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your appl
ication with ToolRunner to remedy this.
2023-05-08 03:15:07,716 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hy2228_nyu_edu/.staging/job_168307733995
0_5407
2023-05-08 03:15:07,942 INFO input.FileInputFormat: Total input files to process : 1
2023-05-08 03:15:08,010 INFO mapreduce.JobSubmitter: number of splits:1
2023-05-08 03:15:08,175 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1683077339950_5407
2023-05-08 03:15:08,176 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-05-08 03:15:08,353 INFO conf.Configuration: resource-types.xml not found
2023-05-08 03:15:08,353 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-05-08 03:15:08,543 INFO impl.YarnClientImpl: Submitted application application_1683077339950_5407
2023-05-08 03:15:08,579 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1683077339950_5407/
2023-05-08 03:15:08,580 INFO mapreduce.Job: Running job: job_1683077339950_5407
2023-05-08 03:15:26,729 INFO mapreduce.Job: Job job_1683077339950_5407 running in uber mode : false
2023-05-08 03:15:26,730 INFO mapreduce.Job:  map 0% reduce 0%
2023-05-08 03:15:37,835 INFO mapreduce.Job:  map 100% reduce 0%
2023-05-08 03:15:43,868 INFO mapreduce.Job:  map 100% reduce 100%
2023-05-08 03:15:44,881 INFO mapreduce.Job: Job job_1683077339950_5407 completed successfully
2023-05-08 03:15:44,977 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=27312108
                FILE: Number of bytes written=55117349
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=118863963
                HDFS: Number of bytes written=23324377
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=32292
                Total time spent by all reduces in occupied slots (ms)=15564
                Total time spent by all map tasks (ms)=8073
```



Copy large datafile from data ingest website to hadoop



```
                GS: Number of bytes read=5341868833
                GS: Number of bytes written=0
                GS: Number of read operations=652084
                GS: Number of large read operations=0
                GS: Number of write operations=0
                HDFS: Number of bytes read=456
                HDFS: Number of bytes written=5341868833
                HDFS: Number of read operations=12
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=6
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=55547
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=55547
                Total vcore-milliseconds taken by all map tasks=55547
                Total megabyte-milliseconds taken by all map tasks=56880128
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=143
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=237
                CPU time spent (ms)=40960
                Physical memory (bytes) snapshot=486567936
                Virtual memory (bytes) snapshot=2587996160
                Total committed heap usage (bytes)=594542592
                Peak Map Physical memory (bytes)=486567936
                Peak Map Virtual memory (bytes)=2604707840
        File Input Format Counters
                Bytes Read=313
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bandwidth in Btyes=102728246
                Bytes Copied=5341868833
                Bytes Expected=5341868833
                Files Copied=1
```

MR job

SSH-in-browser UPLOAD FILE DOWNLOAD FILE

```
hy2228_nyu_edu@nyu-dataproc-m:~$ hadoop jar yelpReviewRedo.jar YelpReview project_redo_yelpReview/yelp_academic_dataset_review.json project_redo_yelpReview/outp
ut
2023-05-08 04:30:28,915 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.24:8032
2023-05-08 04:30:29,120 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.24:10200
2023-05-08 04:30:29,301 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your appl
ication with ToolRunner to remedy this.
2023-05-08 04:30:29,387 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hy2228_nyu_edu/.staging/job_168307733995
0_5548
2023-05-08 04:30:29,635 INFO input.FileInputFormat: Total input files to process : 1
2023-05-08 04:30:29,737 INFO mapreduce.JobSubmitter: number of splits:40
2023-05-08 04:30:29,923 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1683077339950_5548
2023-05-08 04:30:29,925 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-05-08 04:30:30,109 INFO conf.Configuration: resource-types.xml not found
2023-05-08 04:30:30,110 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-05-08 04:30:30,295 INFO impl.YarnClientImpl: Submitted application application_1683077339950_5548
2023-05-08 04:30:30,332 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1683077339950_5548/
2023-05-08 04:30:30,332 INFO mapreduce.Job: Running job: job_1683077339950_5548
2023-05-08 04:30:38,441 INFO mapreduce.Job: Job job_1683077339950_5548 running in uber mode : false
2023-05-08 04:30:38,443 INFO mapreduce.Job:  map 0% reduce 0%
2023-05-08 04:30:48,539 INFO mapreduce.Job:  map 8% reduce 0%
2023-05-08 04:30:49,545 INFO mapreduce.Job:  map 20% reduce 0%
2023-05-08 04:30:50,550 INFO mapreduce.Job:  map 22% reduce 0%
2023-05-08 04:30:51,556 INFO mapreduce.Job:  map 28% reduce 0%
2023-05-08 04:30:52,562 INFO mapreduce.Job:  map 38% reduce 0%
2023-05-08 04:30:53,570 INFO mapreduce.Job:  map 40% reduce 0%
2023-05-08 04:30:54,578 INFO mapreduce.Job:  map 70% reduce 0%
2023-05-08 04:30:55,584 INFO mapreduce.Job:  map 85% reduce 0%
2023-05-08 04:30:56,589 INFO mapreduce.Job:  map 95% reduce 0%
2023-05-08 04:30:57,594 INFO mapreduce.Job:  map 100% reduce 0%
2023-05-08 04:31:15,688 INFO mapreduce.Job:  map 100% reduce 86%
2023-05-08 04:31:18,702 INFO mapreduce.Job:  map 100% reduce 100%
2023-05-08 04:31:20,721 INFO mapreduce.Job: Job job_1683077339950_5548 completed successfully
2023-05-08 04:31:20,824 INFO mapreduce.Job: Counters: 56
        File System Counters
                FILE: Number of bytes read=685047446
                FILE: Number of bytes written=1380205016
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=5342035137
                HDFS: Number of bytes written=510290440
                HDFS: Number of read operations=125
```

```
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Killed map tasks=1
                Launched map tasks=40
                Launched reduce tasks=1
                Data-local map tasks=6
                Rack-local map tasks=34
                Total time spent by all maps in occupied slots (ms)=1984144
                Total time spent by all reduces in occupied slots (ms)=75768
                Total time spent by all map tasks (ms)=496036
                Total time spent by all reduce tasks (ms)=18942
                Total vcore-milliseconds taken by all map tasks=496036
                Total vcore-milliseconds taken by all reduce tasks=18942
                Total megabyte-milliseconds taken by all map tasks=2031763456
                Total megabyte-milliseconds taken by all reduce tasks=77586432
        Map-Reduce Framework
                Map input records=6990280
                Map output records=6990280
                Map output bytes=671066880
                Map output materialized bytes=685047680
                Input split bytes=6560
                Combine input records=0
                Combine output records=0
                Reduce input groups=6990280
                Reduce shuffle bytes=685047680
                Reduce input records=6990280
                Reduce output records=6990280
                Spilled Records=13980560
                Shuffled Maps =40
                Failed Shuffles=0
                Merged Map outputs=40
                GC time elapsed (ms)=14986
                CPU time spent (ms)=420450
                Physical memory (bytes) snapshot=59018932224
                Virtual memory (bytes) snapshot=196897505280
                Total committed heap usage (bytes)=58265174016
                Peak Map Physical memory (bytes)=1463873536
                Peak Map Virtual memory (bytes)=4830834688
                Peak Reduce Physical memory (bytes)=2015522816
                Peak Reduce Virtual memory (bytes)=4814077952
        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=5342028577
        File Output Format Counters
                Bytes Written=510290440
hy2228_nyu_edu@nyu-dataproc-m:~$
```

Create external tables in Hive

SSH-in-browser UPLOAD FILE DOWNLOAD FILE

```
hy2228_nyu_edu@nyu-dataproc-m:~$ beeline -u jdbc:hive2://localhost:10000
Connecting to jdbc:hive2://localhost:10000
Connected to: Apache Hive (version 3.1.2)
Driver: Hive JDBC (version 3.1.2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.2 by Apache Hive
0: jdbc:hive2://localhost:10000> set hive.execution.engine=mr;
No rows affected (0.048 seconds)
0: jdbc:hive2://localhost:10000> set hive.fetch.task.conversion=minimal;
No rows affected (0.004 seconds)
0: jdbc:hive2://localhost:10000> use hy2228_nyu_edu;
No rows affected (0.075 seconds)
0: jdbc:hive2://localhost:10000> show tables;
+-----------------+
|    tab_name     |
+-----------------+
| w1              |
| w3              |
| yelpbusiness    |
| yelpjoineddata  |
| yelpreview      |
| yelpuser        |
+-----------------+
6 rows selected (0.129 seconds)
0: jdbc:hive2://localhost:10000> create external table yelpUserRedo (user_id string, name string, review_count bigint, yelping_since string, average_stars float, friends_count bigi
nt, useful bigint, funny bigint, cool bigint)
. . . . . . . . . . . . . . . .> row format delimited fields terminated by ','
. . . . . . . . . . . . . . . .> location '/user/hy2228_nyu_edu/project_redo_yelpUser/output';
No rows affected (0.162 seconds)
0: jdbc:hive2://localhost:10000> show tables;
+-----------------+
|    tab_name     |
+-----------------+
| w1              |
| w3              |
| yelpbusiness    |
| yelpjoineddata  |
| yelpreview      |
| yelpuser        |
| yelpuserredo    |
+-----------------+
7 rows selected (0.051 seconds)
0: jdbc:hive2://localhost:10000> create external table yelpBusinessRedo (business_id string, name string, city string, state string, postal_code string, latitude float, longitude f
loat, stars float, review_count bigint, is_open int, categories string)
. . . . . . . . . . . . . . . .> row format delimited fields terminated by ','
. . . . . . . . . . . . . . . .> location '/user/hy2228_nyu_edu/project_redo_yelpBusiness/output/';
No rows affected (0.13 seconds)
0: jdbc:hive2://localhost:10000> create external table yelpReviewRedo (review_id string, user_id string, business_id string, stars int)
. . . . . . . . . . . . . . . .> row format delimited fields terminated by ','
```

```
. . . . . .> location '/user/hy2228_nyu_edu/project_redo_yelpBusiness/output/';
No rows affected (0.13 seconds)
0: jdbc:hive2://localhost:10000> create external table yelpReviewRedo (review_id string, user_id string, business_id string, stars int)
. . . . . .> row format delimited fields terminated by ','
. . . . . .> location '/user/hy2228_nyu_edu/project_redo_yelpReview/output';
No rows affected (0.1 seconds)
0: jdbc:hive2://localhost:10000> show tables;
+------------------+
|     tab_name     |
+------------------+
| w1               |
| w3               |
| yelpbusiness     |
| yelpbusinessredo |
| yelpjoineddata   |
| yelpreview       |
| yelpreviewredo   |
| yelpuser         |
| yelpuserredo     |
+------------------+
9 rows selected (0.053 seconds)
0: jdbc:hive2://localhost:10000>
```