

## Upload large file/dataset to data ingest website

The screenshot shows the Google Cloud Storage interface. On the left, a sidebar lists 'Cloud Storage', 'Buckets', 'Monitoring', and 'Settings'. The main area is titled 'Bucket details' for 'nyu-dataproc-hdfs-ingest'. It displays a table of objects with columns for Name, Size, Type, Created, and Storage class. Objects listed include 'Cat\_August\_2010-4.jpg', 'recording\_ids.parquet', 'results\_test\_lenskit.parquet', 'results\_train\_lenskit.parquet', 'results\_validation\_lenskit.parquet', and 'yelp\_academic\_dataset\_user.json'. A message at the bottom says 'gsutil URI copied to clipboard'.

## 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
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)=27700
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 Virtual memory (bytes)=2604171264
File Input Format Counters
Bytes Read=309
File Output Format Counters
Bytes Written=0
DistCp Counters
Bandwidth in Bbytes=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

SSH-in-browser

hy2228@nyu-edu:~\$ hadoop jar yelpUserRedo.jar YelpUser project\_redo\_yelpUser/yelp\_academic\_dataset\_user.json project\_redo\_yelpUser/output

2023-05-07 20:16:55,293 INFO Client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.24:8032

2023-05-07 20:16:55,490 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.24:10200

2023-05-07 20:16:55,660 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.

2023-05-07 20:16:55,674 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hy2228\_nyu\_edu/.staging/job\_1683077339950\_4557

2023-05-07 20:16:56,031 INFO mapreduce.JobSubmitter: number of splits:25

2023-05-07 20:16:56,302 INFO mapreduce.JobSubmitter: Submitting tokens for job: job\_1683077339950\_4557

2023-05-07 20:16:56,302 INFO mapreduce.JobSubmitter: Executing with tokens: []

2023-05-07 20:16:56,479 INFO conf.Configuration: resource-types.xml not found

2023-05-07 20:16:56,479 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.

2023-05-07 20:16:56,651 INFO impl.YarnClientImpl: Submitted application application\_1683077339950\_4557

2023-05-07 20:16:56,651 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application\_1683077339950\_4557/

2023-05-07 20:16:56,684 INFO mapreduce.Job: Running job: job\_1683077339950\_4557

2023-05-07 20:17:03,770 INFO mapreduce.Job: Job job\_1683077339950\_4557 running in uber mode : false

2023-05-07 20:17:03,771 INFO mapreduce.Job: map 0% reduce 0%

2023-05-07 20:17:11,854 INFO mapreduce.Job: map 12% reduce 0%

2023-05-07 20:17:12,860 INFO mapreduce.Job: map 20% reduce 0%

2023-05-07 20:17:13,865 INFO mapreduce.Job: map 44% reduce 0%

2023-05-07 20:17:14,873 INFO mapreduce.Job: map 52% reduce 0%

2023-05-07 20:17:15,879 INFO mapreduce.Job: map 76% reduce 0%

2023-05-07 20:17:16,884 INFO mapreduce.Job: map 84% reduce 0%

2023-05-07 20:17:17,889 INFO mapreduce.Job: map 96% reduce 0%

2023-05-07 20:17:18,994 INFO mapreduce.Job: map 100% reduce 0%

2023-05-07 20:17:32,961 INFO mapreduce.Job: map 100% reduce 100%

2023-05-07 20:17:33,974 INFO mapreduce.Job: Job job\_1683077339950\_4557 completed successfully

2023-05-07 20:17:33,974 INFO mapreduce.Job: Counters: 55

File System Counters

FILE: Number of bytes read=149883619  
FILE: Number of bytes written=306197438  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0  
HDFS: Number of bytes read=3363455891  
HDFS: Number of bytes written=100196188  
HDFS: Number of read operations=80  
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=25  
Launched reduce tasks=1  
Rack-local map tasks=25  
Total time spent by all maps in occupied slots (ms)=909628  
Total time spent by all reduces in occupied slots (ms)=52796  
Total time spent by all map tasks (ms)=227407  
Total time spent by all reduce tasks (ms)=13199

print data in the output file

SSH-in-browser

UPLOAD FILE DOWNLOAD FILE

```

Ydhb8_W_MZVRaad3x0nWA,Grace,_3,2017,2,67,226,1,0,0
Ydwys_W5temHj6m1Vg7A,Emily,_15,2014,4,56,0,9,0,2
YdeTauahhoueTj5IVcwv,Tonya,_2,2016,1,0,31,11,5,4
YdeTo820gcUf01DLdzs0IQ,Cait,_1,2017,1,0,83,3,0,0
YdehdAGKan2DyPhpgCmg,Joe,_17,2014,3,22,2,8,0,1
Yde_wiV9xQxuKu9HZpKtCA,Jennifer,_23,2012,3,17,0,30,6,3
Ydeox4EPB_6cKaeEyyzgAQ,Cody,_3,2014,5,0,432,0,0,0
YdflomX45PjL_gYJLqMzG,Paul,_1,2013,1,0,16,38,10,0
Ydfm35r3c0n_7EHuJzB,Brynnie,_5,2013,2,0,2,3,0,0
YdgJzwt90c_hCp9kvCOEQ,Abigail,_1,2013,4,0,78,1,0,0
Ydrk04gHa8C7pff_Ge2W,Dimitrije,_3,2014,4,0,0,0,0,0
Ydfw05gvtpklcLoS613wQ,Lauren,_3,2015,5,0,0,0,0,0
Ydg_RM5vMSvvc0l6m-Tp4o,Alex,_3,2013,2,67,0,2,0,1
YdgQ0VgjdLmTVzHzUKqQ,C,_3,2009,2,67,2,4,1,1
YdgTwbpuY3_5MTC2RhDOFO,Tim,_8,2014,4,75,1,3,0,2
Ydg35r3mleCQq06x1_La,Hei,Itz,_3,2019,5,0,0,2,0,0
YdgxbgmtX8HNzHSBA82v2cg,Isa,_19,2008,3,95,0,8,4,5
YdgwfgvaWdMtsvOkpdyzvrg,Lauren,_8,2012,4,38,21,15,0,2
Ydhaseel_SKdmjW7Erqg9cA,Chris,_14,2012,3,71,1,8,0,0
Ydhiky9bUcl1A_vwccfrkr,Pete,_4,2014,5,0,1,2,0,0
Ydhjq014NJO4l3gUIYAKg,Julia,_7,2017,5,0,1,1,1,1
Ydhjq014NJO4l3gUIYAKg,Linda,_3,2013,2,38,15,1,1,0
Ydhjq014NJO4l3gUIYAKg,Linda,_3,2013,2,38,15,1,1,0
Ydhjq014NJO4l3gUIYAKg,Linda,_3,2013,2,38,15,1,1,0
Yd13DXdkzQtAbmQNgGX1A,Evan,_3,2012,2,67,0,6,1,0
Ydiw1EHVM4qjS5cBaXgw,Tanisha,_4,2018,5,0,0,0,0,0
YdixRdWgS_G00ux1_96Mng,Bill,_1,2013,2,0,0,0,1,0
Ydzi2np5_rc005WUzbyxgg,Maria,_299,2010,2,93,127,1132,269,490
Ydh15jgjIu61bw2D6ewk3A,Amanda,_1,2019,5,0,0,0,0,0
YdjM5y6ckz9uMoPjJw,Phil,_4,2014,4,0,1,4,1,0
Ydj5o7pzTN_tMwRYNj51A,Christina,_1,2021,5,0,8,0,0,0
Ydj1_g_oSBsX7ZJ5s8W1A,Alyssa,_2,2017,2,67,0,2,1,0
YdjMDTAZzrex9RYC7B12g,Jillian,_2,2014,5,0,330,1,1,1
YdjUYQoIfoAVwXwV8wyg,Chris,_1,2013,5,0,0,0,1,0
Ydjbb9hyrtwdw4DC1QwJ,Jarrin,_35,2009,4,11,721,35,10,19
YdkQDQ45_gS5qmEB99j,g,Sean,_1,2019,5,0,86,0,0,0
Ydk9Qb014NJO4l3gUIYAKg,Dawn,_1,2013,2,38,15,1,1,0
Ydk9Qb014NJO4l3gUIYAKg,Dawn,_1,2013,2,38,15,1,1,0
YdkkQc8KzGzRkgrzlyndW,Anna,_9,2017,4,0,196,6,0,0
Ydkk11z_w24UnrFVTXkQw,Matthew,_7,2011,2,88,337,9,6,2
Ydkkfa6_xw1Esj6pSpw75A,Andrew,_3,2012,4,67,1,0,0,0
YdkrpfxzDJeulkpLv8o82g,Gordon,_4,2013,5,0,9,1,0,0
YdkvXQJUeMbjz9820LkL1Q2,Andrea,_2,2014,5,0,0,0,0,0
Ydkw_71wodzKwHdLhOMAQ,Tom,_4,2016,4,0,0,0,0,0
Ydl9P20pxPRoPjxCpmOog,Kim,_1,2012,5,0,0,0,0,0
Ydlbh4sx4nuwJ47GvSNKw,James,_3,2015,2,0,0,1,4,0
YdlcnwRe53j0NK_FleoVpq,Chad,_13,2011,4,36,2,6,0,2
YdlhklOPqmKpVw_3u5QqKA,Nikki,_12,2020,4,67,0,0,0,0
YdltnuXyG7MKMpDUE_Jw,Naomi,_1,2021,1,0,0,1,0,0
YdltnuXyG7MKMpDUE_Jw,Naomi,_1,2021,1,0,0,1,0,0

```

## 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 yelpJoineduserDataFL (woHeader).txt yelpUserDatasetClean.jar yelpuserinputs.json
merged_output.txt project_output yelpJoineduserDataEL.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_yelpBusiness/output
2023-05-08 03:15:07,236 INFO mapred.YARNProxy: Connecting to ResourceManager at nyu-dataprocm-/192.168.1.24:8032
2023-05-08 03:15:07,421 INFO client.AHSProxy: Connecting to Application History server at nyu-dataprocm-/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 application 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-dataprocm-: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

```

SSH-in-browser

UPLOAD FILE DOWNLOAD FILE ⚙️ 📈⚙️

```
Total time spent by all reduce tasks (ms)=3891
Total vcore-milliseconds taken by all map tasks=8073
Total vcore-milliseconds taken by all reduce tasks=3891
Total megabyte-milliseconds taken by all map tasks=33067008
Total megabyte-milliseconds taken by all reduce tasks=15937536

Map-Reduce Framework
Map input records=150346
Map output records=150346
Map output bytes=26895924
Map output materialized bytes=27312108
Input split bytes=168
Combine input records=0
Combine output records=0
Reduce input groups=150346
Reduce shuffle bytes=27312108
Reduce input records=150346
Reduce output records=150346
Spilled Records=300692
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=207
CPU time spent (ms)=13840
Physical memory (bytes) snapshot=1938939904
Virtual memory (bytes) snapshot=9619374080
Total committed heap usage (bytes)=2185756672
Peak Map Physical memory (bytes)=1421000704
Peak Map Virtual memory (bytes)=4796391424
Peak Reduce Physical memory (bytes)=517939200
Peak Reduce Virtual memory (bytes)=4822982656

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=118863795
File Output Format Counters
Bytes Written=23324377
hy2228_nyu_edu@nyu-dataproc-m:~$
```

## Copy large datafile from data ingest website to hadoop

SSH-in-browser

UPLOAD FILE DOWNLOAD FILE ⚙️ 📈⚙️

```
hy2228_nyu_edu@nyu-dataproc-m:~$ hadoop distcp gs://nyu-dataproc-hdfs-ingest/yelp_academic_dataset_review.json /user/hy2228_nyu_edu/project_redo_yelpReview
2023-05-08 04:09:57,636 INFO client.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false, append=false, useBiff=false, useRdiff=false, fromSnapshot=null, skipCRC=false, blocking=true, numListStatusThreads=0, maxMaps=20, mapBandwidth=0.0, copyStrategy='uniformsize', preserveStatus=[], atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[gs://nyu-dataproc-hdfs-ingest/yelp_academic_dataset_review.json], targetPath=/user/hy2228_nyu_edu/project_redo_yelpReview, filtersFile='null', blocksPerChunk=0, copyBufferSize=8192, verb=oLog=false, directWrite=false}, sourcePaths=[gs://nyu-dataproc-hdfs-ingest/yelp_academic_dataset_review.json], targetPathExists=true, preserveRawXattrs=false
2023-05-08 04:09:57,877 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.24:10203
2023-05-08 04:09:59,119 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.24:10200
2023-05-08 04:09:59,119 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
2023-05-08 04:09:59,119 INFO tools.SimpleCopyListing: Build file listing completed.
2023-05-08 04:09:59,121 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
2023-05-08 04:09:59,121 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
2023-05-08 04:09:59,243 INFO tools.DistCp: Number of paths in the copy list: 1
2023-05-08 04:09:59,297 INFO tools.DistCp: Number of paths in the copy list: 1
2023-05-08 04:09:59,313 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.24:10203
2023-05-08 04:09:59,314 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.24:10200
2023-05-08 04:09:59,387 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hy2228_nyu_edu/.staging/job_168307733995
0_5505
2023-05-08 04:09:59,542 INFO mapreduce.JobSubmitter: number of splits:1
2023-05-08 04:09:59,664 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1683077339950_5505
2023-05-08 04:09:59,665 INFO mapreduce.JobSubmitter: Executing with tokens: []
2023-05-08 04:09:59,832 INFO conf.Configuration: resource-types.xml not found
2023-05-08 04:09:59,832 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2023-05-08 04:09:59,979 INFO impl.YarnClientImpl: Submitted application application_1683077339950_5505
2023-05-08 04:10:00,014 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1683077339950_5505/
2023-05-08 04:10:00,014 INFO tools.DistCp: DistCp job-id: job_1683077339950_5505
2023-05-08 04:10:00,015 INFO mapreduce.Job: Running job: job_1683077339950_5505
2023-05-08 04:10:38,271 INFO mapreduce.Job: Job job_1683077339950_5505 running in uber mode : false
2023-05-08 04:10:38,272 INFO mapreduce.Job: map 0% reduce 0%
2023-05-08 04:10:55,448 INFO mapreduce.Job: map 100% reduce 0%
2023-05-08 04:11:37,676 INFO mapreduce.Job: Job job_1683077339950_5505 completed successfully
2023-05-08 04:11:37,896 INFO mapreduce.Job: Counters: 42
File System Counters
FILE: Number of bytes read=0
FILE: Number of bytes written=250305
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
GS: Number of bytes read=534186833
GS: Number of bytes written=0
GS: Number of read operations=652084
GS: Number of large read operations=0
```

```

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 Bytes=102728246
Bytes Copied=5341868833
Bytes Expected=5341868833
Files Copied=1

```

## MR job

 SSH-in-browser

▲ UPLOAD FILE ▾ DOWNLOAD FILE ✖️ 📁 ⚙️ ⚙️

```

hy2228_nyu_edu@nyu-dataproj-m:~$ hadoop jar yelpReviewRedo.jar YelpReview project_redo_yelpReview/yelp_academic_dataset_review.json project_redo_yelpReview/output
2023-05-08 04:30:28,915 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproj-m/192.168.1.24:8032
2023-05-08 04:30:29,120 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproj-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 application 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_1683077339950_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-dataproj-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)=420450
  CPU time spent (ms)=14986
  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-dataprof-m:~$ 

```

## Create external tables in Hive

SSH-in-browser

```

hy2228_nyu_edu@nyu-dataprof-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 bigint, 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 float, 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 | 
| yelpjoinedata | 
| yelpreview | 
| yelpreviewredo | 
| yelpuser | 
| yelpuserredo | 
+-----+-----+
9 rows selected (0.053 seconds)
0: jdbc:hive2://localhost:10000>
```