

1. Importing RDBMS Data into HDFS

```
mohanram@ubuntu: ~  
File Edit View Search Terminal Tabs Help  
mohanram@ubuntu: ~/sqoop-1.4.7.bin_hadoop-2.6... x mohanram@ubuntu: ~ x mohanram@ubuntu: ~ x  
mohanram@ubuntu:~$ sqoop-import --connect jdbc:mysql://localhost/test --username root --password rootpasswordgiven --table Salaries  
Warning: /home/mohanram/sqoop-1.4.7.bin_hadoop-2.6.0/./hcatalog does not exist! HCatalog jobs will fail.  
Please set $HCAT_HOME to the root of your HCatalog installation.  
Warning: /home/mohanram/sqoop-1.4.7.bin_hadoop-2.6.0/./accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
Warning: /home/mohanram/sqoop-1.4.7.bin_hadoop-2.6.0/./zookeeper does not exist! Accumulo imports will fail.  
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.  
/usr/local/hadoop/hadoop-3.1.1/libexec/hadoop-functions.sh: line 2358: HADOOP_ORG.APACHE.SQOOP.SQOOP_USER: bad substitution  
/usr/local/hadoop/hadoop-3.1.1/libexec/hadoop-functions.sh: line 2453: HADOOP_ORG.APACHE.SQOOP.SQOOP_OPTS: bad substitution  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/usr/local/hadoop/hadoop-3.1.1/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/local/hbase/hbase-1.4.12/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]  
2020-01-20 06:13:38,299 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7  
2020-01-20 06:13:38,384 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.  
2020-01-20 06:13:38,763 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.  
2020-01-20 06:13:38,763 INFO tool.CodeGenTool: Beginning code generation  
2020-01-20 06:13:39,984 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Salaries` AS t LIMIT 1  
2020-01-20 06:13:40,335 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Salaries` AS t LIMIT 1  
2020-01-20 06:13:40,420 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop/hadoop-3.1.1  
Note: /tmp/sqoop-mohanram/compile/403de92b2d4882cc21dc92f9fd70586e/Salaries.java uses or overrides a deprecated API.  
Note: Recompile with -Xlint:deprecation for details.  
2020-01-20 06:13:58,685 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-mohanram/compile/403de92b2d4882cc21dc92f9fd70586e/Salaries.jar  
2020-01-20 06:13:58,987 WARN manager.MySQLManager: It looks like you are importing from mysql.  
2020-01-20 06:13:58,987 WARN manager.MySQLManager: This transfer can be faster! Use the --direct  
2020-01-20 06:13:58,988 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.  
2020-01-20 06:13:58,988 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)  
2020-01-20 06:13:59,025 INFO mapreduce.ImportJobBase: Beginning import of Salaries
```

```
mohanram@ubuntu: ~  
File Edit View Search Terminal Tabs Help  
mohanram@ubuntu: ~/sqoop-1.4.7.bin_hadoop-2.6... x mohanram@ubuntu: ~ x mohanram@ubuntu: ~ x  
2020-01-20 06:23:25,615 INFO mapreduce.Job: map 50% reduce 0%  
2020-01-20 06:23:26,626 INFO mapreduce.Job: map 100% reduce 0%  
2020-01-20 06:24:11,089 INFO mapreduce.Job: Job job_1579522838469_0001 completed successfully  
2020-01-20 06:24:11,694 INFO mapreduce.Job: Counters: 34  
File System Counters  
FILE: Number of bytes read=0  
FILE: Number of bytes written=894592  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0  
HDFS: Number of bytes read=401  
HDFS: Number of bytes written=2171  
HDFS: Number of read operations=24  
HDFS: Number of large read operations=0  
HDFS: Number of write operations=8  
Job Counters  
Failed map tasks=4  
Killed map tasks=2  
Launched map tasks=9  
Other local map tasks=9  
Total time spent by all maps in occupied slots (ms)=794075  
Total time spent by all reduces in occupied slots (ms)=0  
Total time spent by all map tasks (ms)=794075  
Total vcore-milliseconds taken by all map tasks=794075  
Total megabyte-milliseconds taken by all map tasks=813132800  
Map-Reduce Framework  
Map input records=100  
Map output records=100  
Input split bytes=401  
Spilled Records=0  
Failed Shuffles=0  
Merged Map outputs=0  
GC time elapsed (ms)=2074  
CPU time spent (ms)=25360  
Physical memory (bytes) snapshot=806187008
```

2. Exporting HDFS data to a RDBMS

```
mohanram@ubuntu: ~  
File Edit View Search Terminal Tabs Help  
mohanram@ubuntu: ~/sqoop-1.4.7.bin_hadoop-2.6... x mohanram@ubuntu: ~ x mohanram@ubuntu: ~ x  
mohanram@ubuntu:~$ sqoop export --connect jdbc:mysql://localhost/test --username root --password rootpasswordgiven --table salaries2 --export-dir saldata --input-fields-terminated-by ","  
Warning: /home/mohanram/sqoop-1.4.7.bin_hadoop-2.6.0/./hcatalog does not exist! HCatalog jobs will fail.  
Please set $HCAT_HOME to the root of your HCatalog installation.  
Warning: /home/mohanram/sqoop-1.4.7.bin_hadoop-2.6.0/./accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
Warning: /home/mohanram/sqoop-1.4.7.bin_hadoop-2.6.0/./zookeeper does not exist! Accumulo imports will fail.  
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.  
/usr/local/hadoop/hadoop-3.1.1/libexec/hadoop-functions.sh: line 2358: HADOOP_ORG.APACHE.SQOOP.SQOOP_USER: bad substitution  
/usr/local/hadoop/hadoop-3.1.1/libexec/hadoop-functions.sh: line 2453: HADOOP_ORG.APACHE.SQOOP.SQOOP_OPTS: bad substitution  
SLF4J: Class path contains multiple SLF4J bindings.  
SLF4J: Found binding in [jar:file:/usr/local/hadoop/hadoop-3.1.1/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: Found binding in [jar:file:/usr/local/hbase/hbase-1.4.12/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]  
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.  
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]  
2020-01-20 08:18:02,097 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7  
2020-01-20 08:18:02,178 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.  
2020-01-20 08:18:02,448 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.  
2020-01-20 08:18:02,454 INFO tool.CodeGenTool: Beginning code generation  
2020-01-20 08:18:03,235 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `salaries2` AS t LIMIT 1  
2020-01-20 08:18:03,307 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `salaries2` AS t LIMIT 1  
2020-01-20 08:18:03,320 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop/hadoop-3.1.1  
Note: /tmp/sqoop-mohanram/compile/112529c53f5ba930b07fcb97dd3128b6/salaries2.java uses or overrides a deprecated API.  
Note: Recompile with -Xlint:deprecation for details.  
2020-01-20 08:18:10,585 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-mohanram/compile/112529c53f5ba930b07fcb97dd3128b6/salaries2.jar  
2020-01-20 08:18:10,659 INFO mapreduce.ExportJobBase: Beginning export of salaries2  
2020-01-20 08:18:10,660 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address  
2020-01-20 08:18:10,924 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
2020-01-20 08:18:11,005 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
```

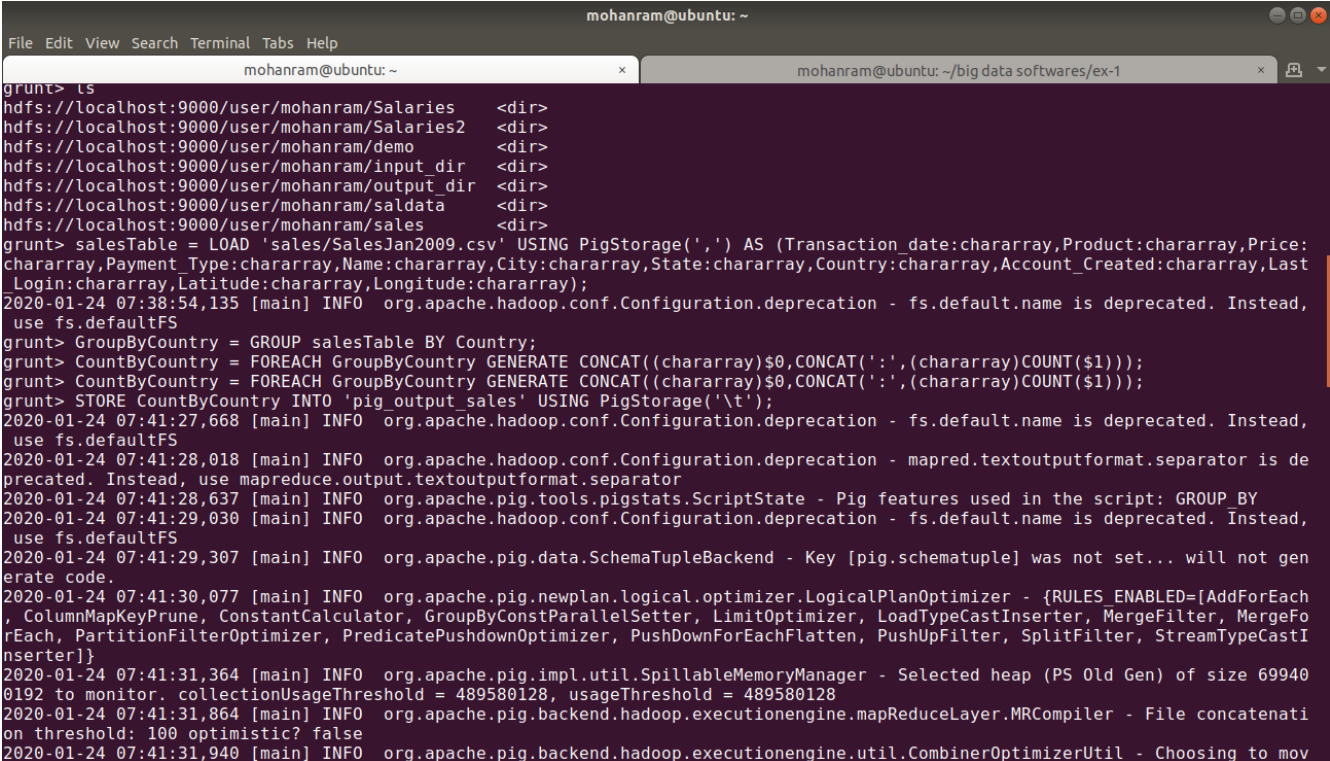
```
mohanram@ubuntu: ~  
File Edit View Search Terminal Tabs Help  
mohanram@ubuntu: ~/sqoop-1.4.7.bin_hadoop-2.6... x mohanram@ubuntu: ~ x mohanram@ubuntu: ~ x  
M | 28 | 708172 | 957939 | 39 |  
M | 62 | 97199 | 961997 | 40 |  
M | 3 | 126290 | 968574 | 41 |  
F | 65 | 110065 | 949548 | 42 |  
M | 51 | 503425 | 956028 | 43 |  
F | 31 | 431725 | 906856 | 44 |  
F | 20 | 720585 | 978506 | 45 |  
M | 23 | 140663 | 921686 | 46 |  
M | 15 | 527577 | 990315 | 47 |  
F | 23 | 238218 | 993288 | 48 |  
M | 21 | 534215 | 988942 | 49 |  
F | 12 | 755232 | 942419 | 50 |  
M | 75 | 739523 | 944506 | 51 |  
M | 33 | 213444 | 946035 | 52 |  
F | 55 | 833711 | 903963 | 53 |  
M | 60 | 118339 | 923977 | 54 |  
M | 17 | 45287 | 952692 | 55 |  
M | 11 | 302586 | 953043 | 56 |  
M | 50 | 209967 | 970516 | 57 |  
F | 23 | 909330 | 909390 | 58 |  
F | 10 | 191823 | 997805 | 59 |  
M | 95 | 556758 | 989742 | 60 |  
F | 25 | 451777 | 918317 | 61 |  
F | 26 | 971680 | 990316 | 62 |  
F | 59 | 25452 | 927802 | 63 |  
M | 83 | 98892 | 946476 | 64 |  
M | 52 | 165582 | 970657 | 65 |  
M | 70 | 934726 | 913378 | 66 |  
M | 34 | 219518 | 948818 | 67 |  
M | 93 | 615393 | 927830 | 68 |  
F | 20 | 163065 | 950594 | 69 |  
F | 18 | 139147 | 979649 | 70 |  
M | 21 | 906327 | 931648 | 71 |  
M | 35 | 231792 | 904677 | 72 |  
F | 100 | 616469 | 949969 | 73 |
```

3. Understanding a MapReduce Jog and Running a MapReduce Job

```
mohanram@ubuntu: ~/big data softwares/ex-1
File Edit View Search Terminal Help
mohanram@ubuntu:~/big data softwares/ex-1$ hadoop jar wc.jar WordCount input_dir output_dir
2020-01-23 23:58:07,237 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
a classes where applicable
2020-01-23 23:58:08,988 INFO client.RMPProxy: Connecting to ResourceManager at /0.0.0.0:8032
2020-01-23 23:58:10,304 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Too
l interface and execute your application with ToolRunner to remedy this.
2020-01-23 23:58:10,370 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/mohanra
m/.staging/job_1579851660161_0001
2020-01-23 23:58:10,852 INFO input.FileInputFormat: Total input files to process : 1
2020-01-23 23:58:11,131 INFO mapreduce.JobSubmitter: number of splits:1
2020-01-23 23:58:11,703 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. In
stead, use yarn.system-metrics-publisher.enabled
2020-01-23 23:58:12,405 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1579851660161_0001
2020-01-23 23:58:12,408 INFO mapreduce.JobSubmitter: Executing with tokens: []
2020-01-23 23:58:13,238 INFO conf.Configuration: resource-types.xml not found
2020-01-23 23:58:13,239 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2020-01-23 23:58:14,249 INFO impl.YarnClientImpl: Submitted application application_1579851660161_0001
2020-01-23 23:58:14,421 INFO mapreduce.Job: The url to track the job: http://ubuntu:8088/proxy/application_1579851660161_0001/
2020-01-23 23:58:14,423 INFO mapreduce.Job: Running job: job_1579851660161_0001
2020-01-23 23:58:34,066 INFO mapreduce.Job: Job job_1579851660161_0001 running in uber mode : false
2020-01-23 23:58:34,076 INFO mapreduce.Job: map 0% reduce 0%
2020-01-23 23:58:45,740 INFO mapreduce.Job: Task Id : attempt_1579851660161_0001_m_000000_0, Status : FAILED
[2020-01-23 23:58:43.841]Container [pid=6186,containerID=container_1579851660161_0001_01_000002] is running 350059008B beyond t
he 'VIRTUAL' memory limit. Current usage: 80.3 MB of 1 GB physical memory used; 2.4 GB of 2.1 GB virtual memory used. Killing c
ontainer.
Dump of the process-tree for container 1579851660161_0001_01_000002 :
|- PID PPID PGRPID SESSID CMD_NAME USER_MODE_TIME(MILLIS) SYSTEM_TIME(MILLIS) VMEM_USAGE(BYTES) RSSMEM_USAGE(PAGES) FUL
L_CMD_LINE
|- 6186 6184 6186 6186 (bash) 0 2 10153984 698 /bin/bash -c /usr/lib/jvm/java-8-openjdk-amd64/bin/java -Djava.net.prefe
rIPv4Stack=true -Dhadoop.metrics.log.level=WARN -Xmx820m -Djava.io.tmpdir=/tmp/hadoop-mohanram/nm-local-dir/usercache/mohanra
m/appcache/application_1579851660161_0001/container_1579851660161_0001_01_000002/tmp -Dlog4j.configuration=container-log4j.prop
erties -Dyarn.app.container.log.dir=/usr/local/hadoop/hadoop-3.1.1/logs/userlogs/application_1579851660161_0001/container_15798
51660161_0001_01_000002 -Dyarn.app.container.log.filesize=0 -Dhadoop.root.logger=INFO,CLA -Dhadoop.root.logfile=syslog org.apac
he.hadoop.mapred.YarnChild 192.168.61.142 41327 attempt_1579851660161_0001_m_000000_0 2 1>/usr/local/hadoop/hadoop-3.1.1/logs/u
serlogs/application_1579851660161_0001/container_1579851660161_0001_01_000002/stdout 2>/usr/local/hadoop/hadoop-3.1.1/logs/user
logs/application_1579851660161_0001/container_1579851660161_0001_01_000002/stderr
```

```
mohanram@ubuntu:~/big data softwares/ex-1$ hadoop fs -ls
2020-01-24 00:06:13,924 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
a classes where applicable
Found 5 items
drwxr-xr-x - mohanram supergroup 0 2020-01-20 08:16 Salaries
drwxr-xr-x - mohanram supergroup 0 2020-01-20 07:03 Salaries2
drwxr-xr-x - mohanram supergroup 0 2020-01-24 00:04 input_dir
drwxr-xr-x - mohanram supergroup 0 2020-01-24 00:05 output_dir
drwxr-xr-x - mohanram supergroup 0 2020-01-20 08:17 saldata
mohanram@ubuntu:~/big data softwares/ex-1$ hadoop fs -ls output_dir
2020-01-24 00:06:21,807 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
a classes where applicable
Found 2 items
-rw-r--r-- 1 mohanram supergroup 0 2020-01-24 00:05 output_dir/_SUCCESS
-rw-r--r-- 1 mohanram supergroup 36 2020-01-24 00:05 output_dir/part-r-00000
mohanram@ubuntu:~/big data softwares/ex-1$ hadoop fs -cat output_dir/part-r-00000
2020-01-24 00:06:40,486 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
a classes where applicable
a 5
bear 2
cat 2
dog 3
human 3
is 5
mohanram@ubuntu:~/big data softwares/ex-1$
```

4. Understanding Pig, Getting Started with Pig and Exploring Data with Pig



```
mohanram@ubuntu: ~  
File Edit View Search Terminal Tabs Help  
mohanram@ubuntu: ~ x mohanram@ubuntu: ~/big data softwares/ex-1 x  
grunt> ls  
hdfs://localhost:9000/user/mohanram/Salaries <dir>  
hdfs://localhost:9000/user/mohanram/Salaries2 <dir>  
hdfs://localhost:9000/user/mohanram/demo <dir>  
hdfs://localhost:9000/user/mohanram/input_dir <dir>  
hdfs://localhost:9000/user/mohanram/output_dir <dir>  
hdfs://localhost:9000/user/mohanram/saldata <dir>  
hdfs://localhost:9000/user/mohanram/sales <dir>  
grunt> salesTable = LOAD 'sales/SalesJan2009.csv' USING PigStorage(',') AS (Transaction_date:chararray,Product:chararray,Price:  
chararray,Payment_Type:chararray,Name:chararray,City:chararray,State:chararray,County:chararray,Account_Created:chararray,Last  
_Login:chararray,Latitude:chararray,Longitude:chararray);  
2020-01-24 07:38:54,135 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead,  
use fs.defaultFS  
grunt> GroupByCountry = GROUP salesTable BY Country;  
grunt> CountByCountry = FOREACH GroupByCountry GENERATE CONCAT((chararray)$0,CONCAT(':', (chararray)COUNT($1)));  
grunt> CountByCountry = FOREACH GroupByCountry GENERATE CONCAT((chararray)$0,CONCAT(':', (chararray)COUNT($1)));  
grunt> STORE CountByCountry INTO 'pig_output_sales' USING PigStorage('\t');  
2020-01-24 07:41:27,668 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead,  
use fs.defaultFS  
2020-01-24 07:41:28,018 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is de  
precated. Instead, use mapreduce.output.textoutputformat.separator  
2020-01-24 07:41:28,637 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP BY  
2020-01-24 07:41:29,030 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead,  
use fs.defaultFS  
2020-01-24 07:41:29,307 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not gen  
erate code.  
2020-01-24 07:41:30,077 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach  
, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeFo  
rEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastI  
nserter]}  
2020-01-24 07:41:31,364 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (PS Old Gen) of size 69940  
0192 to monitor. collectionUsageThreshold = 489580128, usageThreshold = 489580128  
2020-01-24 07:41:31,864 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenati  
on threshold: 100 optimistic? false  
2020-01-24 07:41:31,940 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.CombinerOptimizerUtil - Choosing to mov
```

In the below picture, I have used salary data in pig to give the output.

M,66,3732775,4411009
M,70,5359148,6953356
M,35,3115920,4717758
F,27,6784037,4194575
F,64,4781851,6894186
M,62,2775487,5455827
M,72,3301035,5552327
F,26,5562198,6299705
F,29,2933377,4723247
F,38,3809586,6999096
M,71,8452403,5536268
F,73,8637132,6169476
F,56,471633,4085208
M,68,3966201,4294712
M,67,329293,4362304
F,69,6042108,4745612
M,64,4373440,6433782
F,57,4313351,5435795
M,39,6605330,6817221
M,70,3751198,6915101
F,35,1043714,5872953
M,26,9549144,5317057
F,76,4423550,5882178
F,50,3782888,5571300
M,58,5511036,6253977