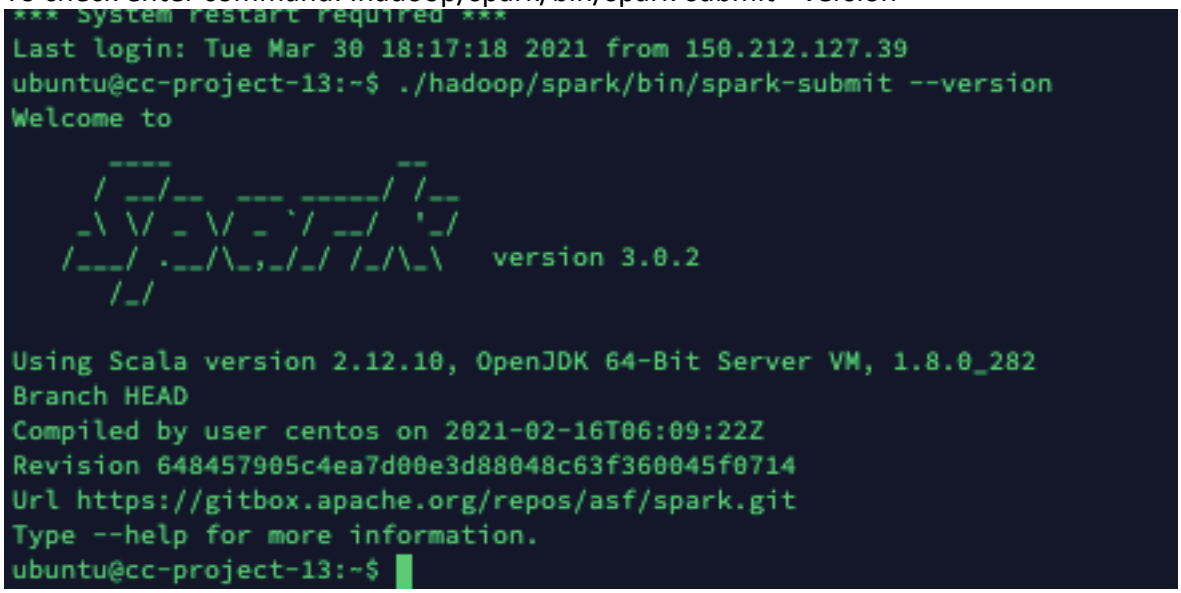
Mini Project 2

Group:
Sweta Rawal (swr22@pitt.edu)
Tazeen Khan (tfk12@pitt.edu)
Neha Shah (nes95@pitt.edu)

The zip contains two folders
Program 1:
Artist.java (jar is /home/ubuntu/target folder in cc-project-13)
Program 2:
p2.py and dependencies3.zip for running program inside the vm cluster
Folder jupyternotebook contains spark program for running it on juypter along with save model and html of the output

Part 1: Setting up spark
Installed version 3.0.2 in instance cc-project-13
To check enter command: .hadoop/spark/bin/spark-submit --version



Part 2: Developing Spark programs

1. To run the program use below command (Since the output is very big I have the command to store the output in consoleoutfile.txt file.)-

   ./hadoop/spark/bin/spark-submit --class artists.Artists --master yarn target/artists-1.0-SNAPSHOT.jar > /home/ubuntu/consoleoutfile.txt 2>&1

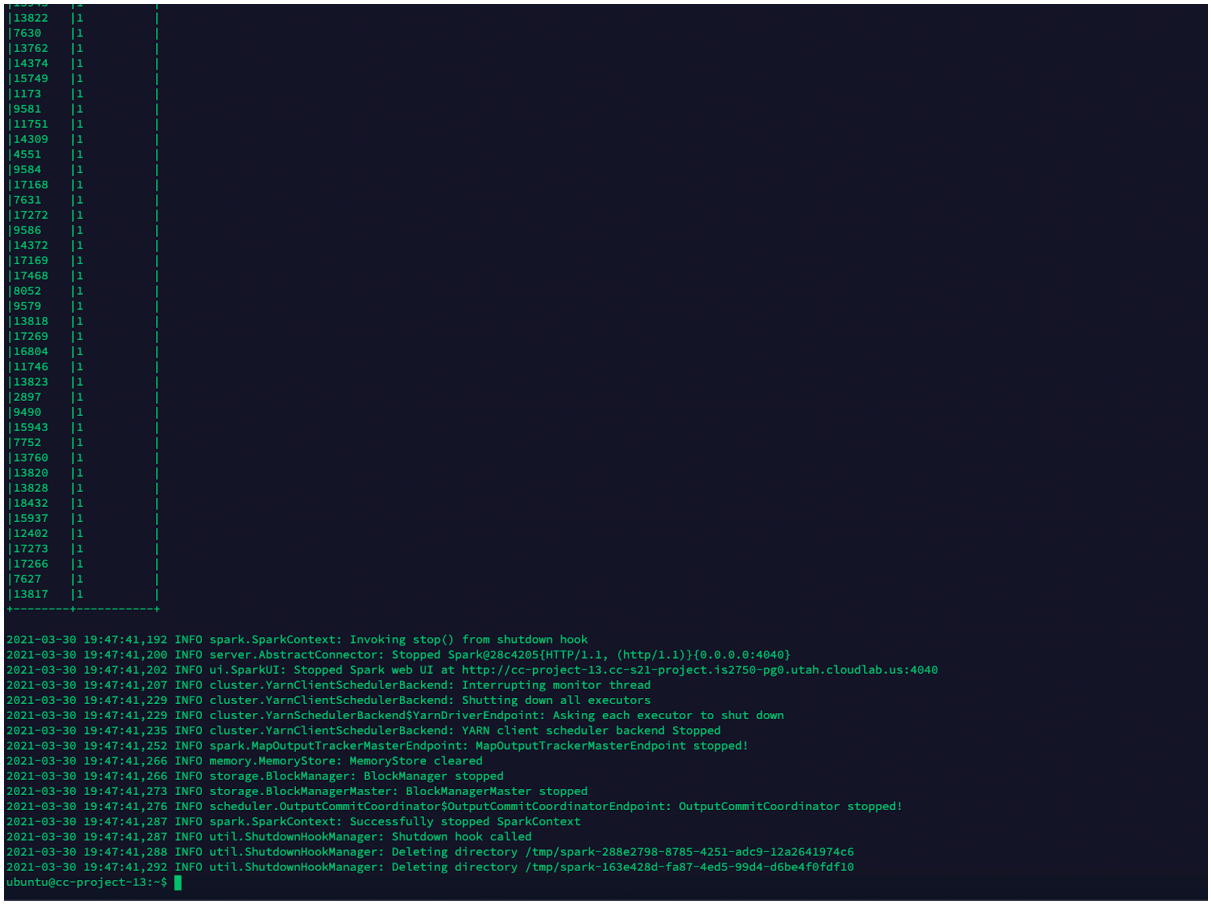
2. To see the output on the command line run below-

   ./hadoop/spark/bin/spark-submit --class artists.Artists --master yarn target/artists-1.0-SNAPSHOT.jar

```
To see these additional updates run: apt list --upgradable


*** System restart required ***
Last login: Tue Mar 30 18:17:22 2021 from 150.212.127.39
ubuntu@cc-project-13:~$ ./hadoop/spark/bin/spark-submit --class artists.Artists --master yarn target/artists-1.0-SNAPSHOT.jar
```

```
   - Reduce system reboots and improve kernel security. Activate at:
     https://ubuntu.com/livepatch

24 packages can be updated.
0 of these updates are security updates.
To see these additional updates run: apt list --upgradable


*** System restart required ***
Last login: Tue Mar 30 18:17:22 2021 from 150.212.127.39
ubuntu@cc-project-13:~$ ./hadoop/spark/bin/spark-submit --class artists.Artists --master yarn target/artists-1.0-SNAPSHOT.jar
2021-03-30 19:47:02,130 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2021-03-30 19:47:02,357 INFO spark.SparkContext: Running Spark version 3.0.2
2021-03-30 19:47:02,446 INFO resource.ResourceUtils: ==============================================================
2021-03-30 19:47:02,448 INFO resource.ResourceUtils: Resources for spark.driver:

2021-03-30 19:47:02,448 INFO resource.ResourceUtils: ==============================================================
2021-03-30 19:47:02,449 INFO spark.SparkContext: Submitted application: total_count
2021-03-30 19:47:02,512 INFO spark.SecurityManager: Changing view acls to: ubuntu
2021-03-30 19:47:02,512 INFO spark.SecurityManager: Changing modify acls to: ubuntu
2021-03-30 19:47:02,512 INFO spark.SecurityManager: Changing view acls groups to:
2021-03-30 19:47:02,512 INFO spark.SecurityManager: Changing modify acls groups to:
2021-03-30 19:47:02,512 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(ubuntu); groups with vie
 users  with modify permissions: Set(ubuntu); groups with modify permissions: Set()
2021-03-30 19:47:02,752 INFO util.Utils: Successfully started service 'sparkDriver' on port 34379.
2021-03-30 19:47:02,784 INFO spark.SparkEnv: Registering MapOutputTracker
2021-03-30 19:47:02,817 INFO spark.SparkEnv: Registering BlockManagerMaster
2021-03-30 19:47:02,835 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
2021-03-30 19:47:02,836 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
2021-03-30 19:47:02,872 INFO spark.SparkEnv: Registering BlockManagerMasterHeartbeat
2021-03-30 19:47:02,885 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-f932f2d0-5fb9-47db-bb52-77acbb8be5c1
2021-03-30 19:47:02,910 INFO memory.MemoryStore: MemoryStore started with capacity 366.3 MiB
2021-03-30 19:47:02,963 INFO spark.SparkEnv: Registering OutputCommitCoordinator
```

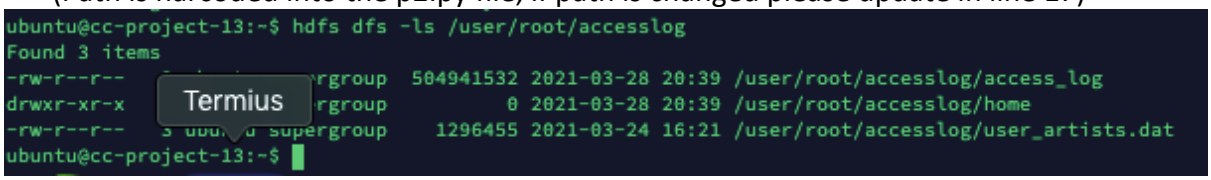
```
2021-03-30 19:47:40,929 INFO scheduler.TaskSetManager: Finished task 70.0 in stage 18.0 (TID 599) in 10 ms on cc-project-14.cc-s21-project.is2750-pg0.utah.cloudlab.us (exe
2021-03-30 19:47:40,937 INFO scheduler.TaskSetManager: Starting task 72.0 in stage 18.0 (TID 601, cc-project-14.cc-s21-project.is2750-pg0.utah.cloudlab.us, executor 2, par
AL, 7336 bytes)
2021-03-30 19:47:40,938 INFO scheduler.TaskSetManager: Finished task 71.0 in stage 18.0 (TID 600) in 8 ms on cc-project-14.cc-s21-project.is2750-pg0.utah.cloudlab.us (exec
2021-03-30 19:47:40,946 INFO scheduler.TaskSetManager: Finished task 72.0 in stage 18.0 (TID 601) in 9 ms on cc-project-14.cc-s21-project.is2750-pg0.utah.cloudlab.us (exec
2021-03-30 19:47:40,946 INFO cluster.YarnScheduler: Removed TaskSet 18.0, whose tasks have all completed, from pool
2021-03-30 19:47:40,947 INFO scheduler.DAGScheduler: ResultStage 18 (show at Artists.java:21) finished in 0.660 s
2021-03-30 19:47:40,947 INFO scheduler.DAGScheduler: Job 7 is finished. Cancelling potential speculative or zombie tasks for this job
2021-03-30 19:47:40,947 INFO cluster.YarnScheduler: Killing all running tasks in stage 18: Stage finished
2021-03-30 19:47:40,948 INFO scheduler.DAGScheduler: Job 7 finished: show at Artists.java:21, took 0.673690 s
2021-03-30 19:47:40,980 INFO codegen.CodeGenerator: Code generated in 14.877918 ms
+--------+-----------+
|artistID|sum(weight)|
+--------+-----------+
|289     |2393140    |
|72      |1301308    |
|89      |1291387    |
|292     |1058405    |
|498     |963449     |
|67      |921198     |
|288     |905423     |
|701     |688529     |
|227     |662116     |
|300     |532545     |
|333     |525844     |
|344     |525292     |
|378     |513476     |
|679     |506453     |
|295     |499318     |
|511     |493024     |
|461     |489065     |
|486     |485532     |
|190     |485076     |
|163     |466104     |
|55      |449292     |
|154     |385306     |
|466     |384405     |
|257     |384307     |
|707     |371916     |
|917     |368710     |
|792     |350035     |
|51      |348919     |
|65      |330757     |
|475     |321011     |
|203     |318221     |
|157     |296882     |
|207     |288520     |
|198     |277397     |
|377     |265362     |
|291     |253027     |
|614     |251440     |
|173     |245878     |
|503     |237148     |
|687     |215777     |
|903     |213103     |
|302     |207761     |
|187     |205195     |
|1412    |203665     |
```



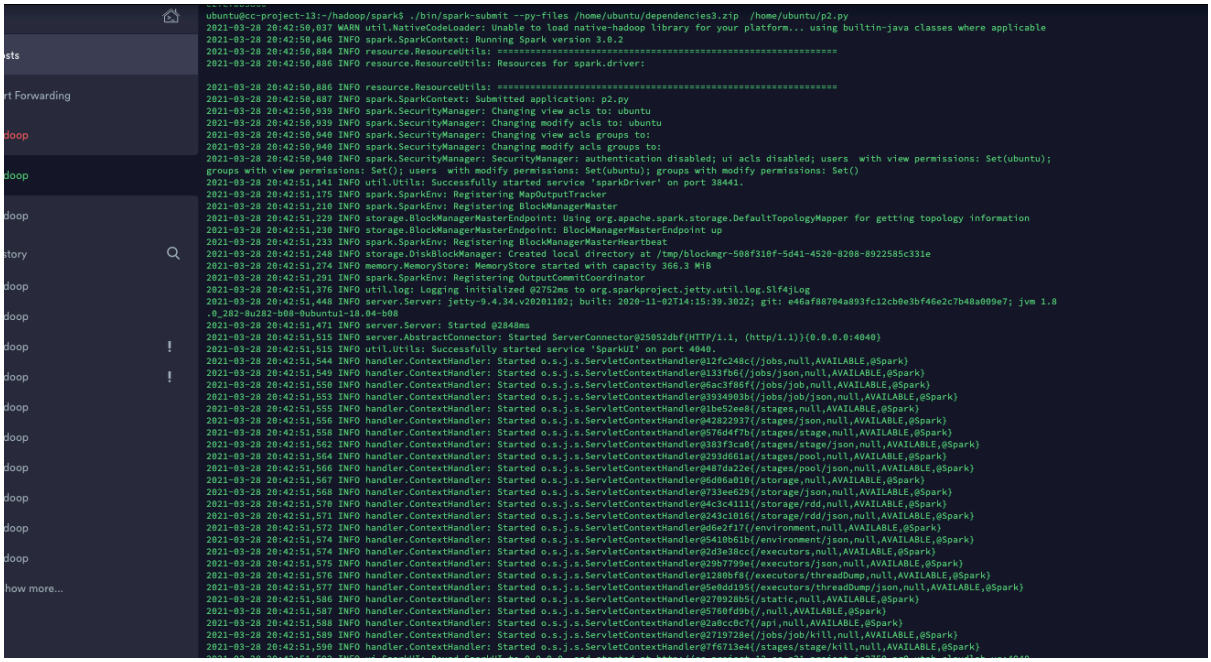
To see full output check the pre generated output in consoleoutfile.txt file or run again.

Part 3: Developing Spark programs 2

1. Access_Log file is placed at hdfs://user/root/accesslog/access_log
   (Path is harcoded into the p2.py file, if path is changed please update in line 17)



2. To run the program external dependencies3.zip is required which is placed at /home/ubuntu
3. Execute the following command to run the program:
   a. Cd Hadoop
   b. Cd spark
   c. ./bin/spark-submit --py-files /home/ubuntu/dependencies3.zip /home/ubuntu/p2.py



```
ubuntu@cc-project-13:~/hadoop/spark$ ./bin/spark-submit --py-files /home/ubuntu/dependencies3.zip  /home/ubuntu/p2.py
2021-03-28 20:42:50,037 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2021-03-28 20:42:50,846 INFO spark.SparkContext: Running Spark version 3.0.2
2021-03-28 20:42:50,884 INFO resource.ResourceUtils: ==============================================================
2021-03-28 20:42:50,886 INFO resource.ResourceUtils: Resources for spark.driver:

2021-03-28 20:42:50,886 INFO resource.ResourceUtils: ==============================================================
2021-03-28 20:42:50,887 INFO spark.SparkContext: Submitted application: p2.py
2021-03-28 20:42:50,939 INFO spark.SecurityManager: Changing view acls to: ubuntu
2021-03-28 20:42:50,939 INFO spark.SecurityManager: Changing modify acls to: ubuntu
2021-03-28 20:42:50,940 INFO spark.SecurityManager: Changing view acls groups to:
2021-03-28 20:42:50,940 INFO spark.SecurityManager: Changing modify acls groups to:
2021-03-28 20:42:50,940 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(ubuntu);
groups with view permissions: Set(); users  with modify permissions: Set(ubuntu); groups with modify permissions: Set()
2021-03-28 20:42:51,141 INFO util.Utils: Successfully started service 'sparkDriver' on port 38441.
2021-03-28 20:42:51,175 INFO spark.SparkEnv: Registering MapOutputTracker
2021-03-28 20:42:51,210 INFO spark.SparkEnv: Registering BlockManagerMaster
2021-03-28 20:42:51,229 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
2021-03-28 20:42:51,230 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
2021-03-28 20:42:51,233 INFO spark.SparkEnv: Registering BlockManagerMasterHeartbeat
2021-03-28 20:42:51,248 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-508f310f-5d41-4520-8208-8922585c331e
2021-03-28 20:42:51,274 INFO memory.MemoryStore: MemoryStore started with capacity 366.3 MiB
2021-03-28 20:42:51,291 INFO spark.SparkEnv: Registering OutputCommitCoordinator
2021-03-28 20:42:51,376 INFO util.log: Logging initialized @2752ms to org.sparkproject.jetty.util.log.Slf4jLog
2021-03-28 20:42:51,440 INFO server.Server: jetty-9.4.34.v20201102; built: 2020-11-02T14:15:39.302Z; git: e46af88704a893fc12cb0e3bf46e2c7b48a009e7; jvm 1.8
.0_282-8u282-b08-0ubuntu1~18.04-b08
2021-03-28 20:42:51,471 INFO server.Server: Started @2848ms
2021-03-28 20:42:51,515 INFO server.AbstractConnector: Started ServerConnector@25052dbf{HTTP/1.1, (http/1.1)}{0.0.0.0:4040}
2021-03-28 20:42:51,515 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
2021-03-28 20:42:51,544 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@12fc248c{/jobs,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,549 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@133fb6{/jobs/json,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,550 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@6ac3f86f{/jobs/job,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,553 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@3934903b{/jobs/job/json,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,555 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@1be52ee8{/stages,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,556 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@42822937{/stages/json,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,558 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@576d4f7b{/stages/stage,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,562 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@383f3ca0{/stages/stage/json,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,564 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@293d661a{/stages/pool,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,566 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@407da22e{/stages/pool/json,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,567 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@6d06a010{/storage,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,568 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@733ee629{/storage/json,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,570 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@4c3c4111{/storage/rdd,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,571 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@243c1016{/storage/rdd/json,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,572 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@d6e2f17{/environment,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,574 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@5410b61b{/environment/json,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,574 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@2d3e38cc{/executors,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,575 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@29b7799e{/executors/json,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,576 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@1280bf8{/executors/threadDump,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,577 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@5e0dd195{/executors/threadDump/json,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,586 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@270928b5{/static,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,587 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@5760fd9b{/,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,588 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@2a8cc0c7{/api,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,589 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@2719728e{/jobs/job/kill,null,AVAILABLE,@Spark}
2021-03-28 20:42:51,590 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7f6713e4{/stages/stage/kill,null,AVAILABLE,@Spark}
```

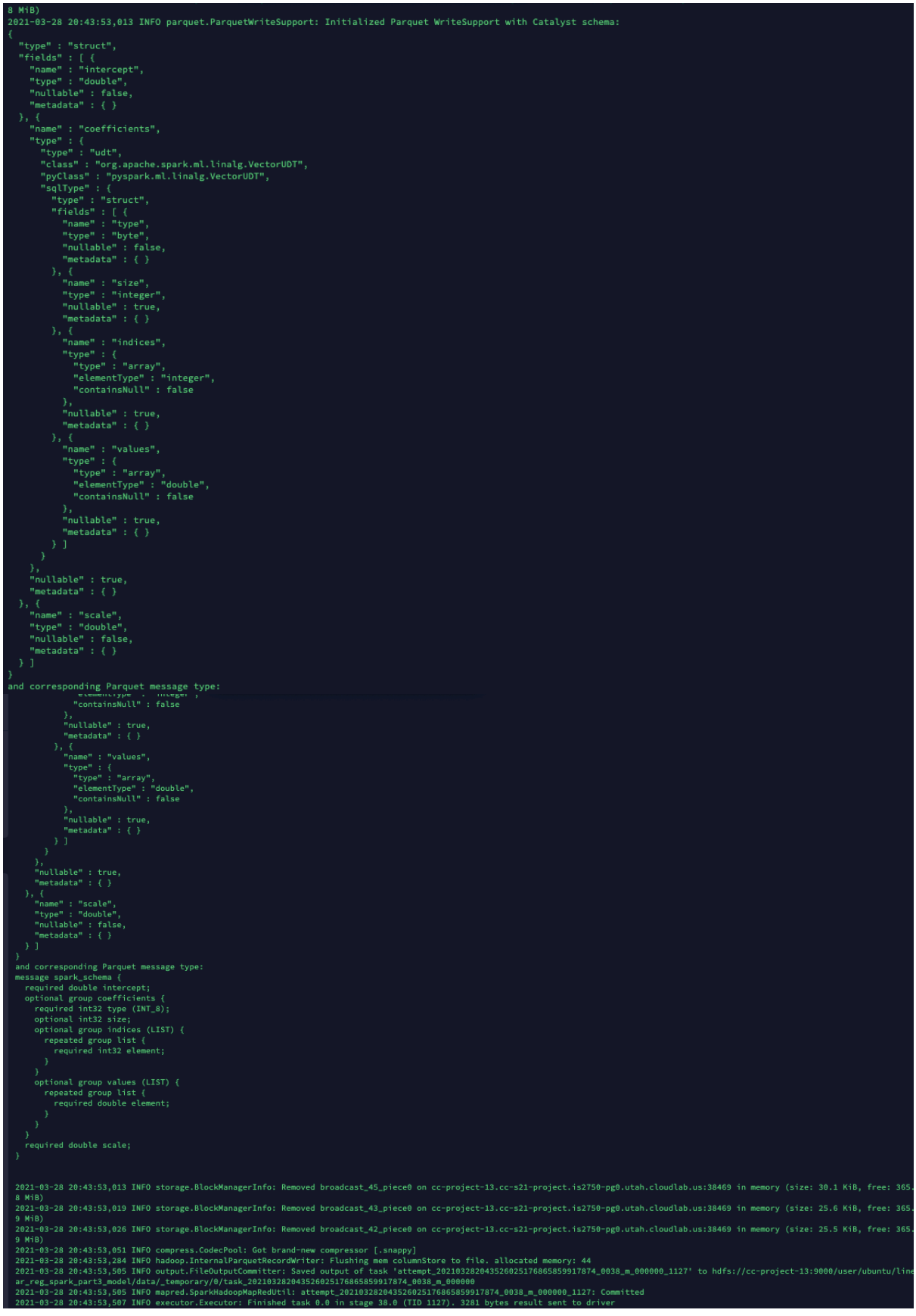
```
2021-03-28 20:43:01,130 INFO codegen.CodeGenerator: Code generated in 16.357893 ms
2021-03-28 20:43:01,153 INFO executor.Executor: Finished task 0.0 in stage 2.0 (TID 5). 2403 bytes result sent to driver
2021-03-28 20:43:01,154 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 5) in 60 ms on cc-project-13.cc-s21-project.is2750-pg0.utah.clou
dlab.us (executor driver) (1/1)
2021-03-28 20:43:01,154 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
2021-03-28 20:43:01,155 INFO scheduler.DAGScheduler: ResultStage 2 (showString at NativeMethodAccessorImpl.java:0) finished in 0.078 s
2021-03-28 20:43:01,155 INFO scheduler.DAGScheduler: Job 2 is finished. Cancelling potential speculative or zombie tasks for this job
2021-03-28 20:43:01,155 INFO scheduler.TaskSchedulerImpl: Killing all running tasks in stage 2: Stage finished
2021-03-28 20:43:01,155 INFO scheduler.DAGScheduler: Job 2 finished: showString at NativeMethodAccessorImpl.java:0, took 0.082848 s
2021-03-28 20:43:01,184 INFO codegen.CodeGenerator: Code generated in 20.345887 ms
+-------------+---+---+--------------------+------+--------------------+---+-----+
|          _c0|_c1|_c2|                 _c3|   _c4|                 _c5|_c6|  _c7|
+-------------+---+---+--------------------+------+--------------------+---+-----+
|10.223.157.186|  -|  -|[15/Jul/2009:14:5...|-0700]|      GET / HTTP/1.1|403|  202|
|10.223.157.186|  -|  -|[15/Jul/2009:14:5...|-0700]|GET /favicon.ico ...|404|  209|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|      GET / HTTP/1.1|200| 9157|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/js/lo...|200|10469|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/css/r...|200| 1014|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/css/9...|200| 6206|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/css/t...|200|15779|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/js/th...|200| 4492|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/js/li...|200|25960|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/img/s...|200|  168|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/img/d...|200| 5604|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/img/d...|200|10556|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/img/d...|200| 9925|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/img/c...|200|  979|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/img/h...|200| 3892|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/img/d...|200| 5397|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/img/l...|200| 2767|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/img/d...|200| 5766|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/img/h...|200|68831|
|10.223.157.186|  -|  -|[15/Jul/2009:15:5...|-0700]|GET /assets/img/d...|200| 5766|
+-------------+---+---+--------------------+------+--------------------+---+-----+
only showing top 20 rows

2021-03-28 20:43:01,249 INFO datasources.FileSourceStrategy: Pushed Filters:
2021-03-28 20:43:01,249 INFO datasources.FileSourceStrategy: Post-Scan Filters:
2021-03-28 20:43:01,249 INFO datasources.FileSourceStrategy: Output Data Schema: struct<_c0: string, _c3: string>
2021-03-28 20:43:01,285 INFO codegen.CodeGenerator: Code generated in 17.556135 ms
2021-03-28 20:43:01,290 INFO memory.MemoryStore: Block broadcast_6 stored as values in memory (estimated size 307.7 KiB, free 365.0 MiB)
2021-03-28 20:43:01,299 INFO memory.MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated size 27.9 KiB, free 364.9 MiB)
2021-03-28 20:43:01,300 INFO storage.BlockManagerInfo: Added broadcast_6_piece0 in memory on cc-project-13.cc-s21-project.is2750-pg0.utah.cloudlab.us:38469
 (size: 27.9 KiB, free: 366.2 MiB)
2021-03-28 20:43:01,301 INFO spark.SparkContext: Created broadcast 6 from showString at NativeMethodAccessorImpl.java:0
2021-03-28 20:43:01,302 INFO execution.FileSourceScanExec: Planning scan with bin packing, max size: 127283959 bytes, open cost is considered as scanning 4
194304 bytes.
2021-03-28 20:43:01,310 INFO spark.SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
2021-03-28 20:43:01,311 INFO scheduler.DAGScheduler: Got job 3 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions
2021-03-28 20:43:01,311 INFO scheduler.DAGScheduler: Final stage: ResultStage 3 (showString at NativeMethodAccessorImpl.java:0)
2021-03-28 20:43:01,311 INFO scheduler.DAGScheduler: Parents of final stage: List()
2021-03-28 20:43:01,311 INFO scheduler.DAGScheduler: Missing parents: List()
2021-03-28 20:43:01,312 INFO scheduler.DAGScheduler: Submitting ResultStage 3 (MapPartitionsRDD[16] at showString at NativeMethodAccessorImpl.java:0), whic
h has no missing parents
2021-03-28 20:43:01,315 INFO memory.MemoryStore: Block broadcast_7 stored as values in memory (estimated size 13.6 KiB, free 364.9 MiB)
2021-03-28 20:43:01,317 INFO memory.MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 6.7 KiB, free 364.9 MiB)
2021-03-28 20:43:01,318 INFO storage.BlockManagerInfo: Added broadcast_7_piece0 in memory on cc-project-13.cc-s21-project.is2750-pg0.utah.cloudlab.us:38469
 (size: 6.7 KiB, free: 366.2 MiB)
2021-03-28 20:43:01,318 INFO spark.SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:1223
```

```
2021-03-28 20:43:01,348 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
2021-03-28 20:43:01,348 INFO scheduler.DAGScheduler: ResultStage 3 (showString at NativeMethodAccessorImpl.java:0) finished in 0.035 s
2021-03-28 20:43:01,349 INFO scheduler.DAGScheduler: Job 3 is finished. Cancelling potential speculative or zombie tasks for this job
2021-03-28 20:43:01,349 INFO scheduler.TaskSchedulerImpl: Killing all running tasks in stage 3: Stage finished
2021-03-28 20:43:01,349 INFO scheduler.DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:0, took 0.039124 s
2021-03-28 20:43:01,362 INFO codegen.CodeGenerator: Code generated in 9.701367 ms
+--------------+--------------------+
|           _c0|                Time|
+--------------+--------------------+
|10.223.157.186|15/Jul/2009:14:58:59|
|10.223.157.186|15/Jul/2009:14:58:59|
|10.223.157.186|15/Jul/2009:15:50:35|
|10.223.157.186|15/Jul/2009:15:50:35|
|10.223.157.186|15/Jul/2009:15:50:35|
|10.223.157.186|15/Jul/2009:15:50:35|
|10.223.157.186|15/Jul/2009:15:50:35|
|10.223.157.186|15/Jul/2009:15:50:35|
|10.223.157.186|15/Jul/2009:15:50:35|
|10.223.157.186|15/Jul/2009:15:50:36|
|10.223.157.186|15/Jul/2009:15:50:36|
|10.223.157.186|15/Jul/2009:15:50:36|
|10.223.157.186|15/Jul/2009:15:50:36|
|10.223.157.186|15/Jul/2009:15:50:36|
|10.223.157.186|15/Jul/2009:15:50:36|
|10.223.157.186|15/Jul/2009:15:50:36|
|10.223.157.186|15/Jul/2009:15:50:36|
|10.223.157.186|15/Jul/2009:15:50:36|
|10.223.157.186|15/Jul/2009:15:50:37|
+--------------+--------------------+
only showing top 20 rows

2021-03-28 20:43:01,433 INFO datasources.FileSourceStrategy: Pushed Filters:
2021-03-28 20:43:01,433 INFO datasources.FileSourceStrategy: Post-Scan Filters:
2021-03-28 20:43:01,433 INFO datasources.FileSourceStrategy: Output Data Schema: struct<_c0: string, _c3: string>
2021-03-28 20:43:01,468 INFO codegen.CodeGenerator: Code generated in 20.751431 ms
2021-03-28 20:43:01,472 INFO memory.MemoryStore: Block broadcast_8 stored as values in memory (estimated size 307.7 KiB, free 364.6 MiB)
2021-03-28 20:43:01,482 INFO memory.MemoryStore: Block broadcast_8_piece0 stored as bytes in memory (estimated size 27.9 KiB, free 364.6 MiB)
2021-03-28 20:43:01,482 INFO storage.BlockManagerInfo: Added broadcast_8_piece0 in memory on cc-project-13.cc-s21-project.is2750-pg0.utah.cloudlab.us:38469
 (size: 27.9 KiB, free: 366.1 MiB)
2021-03-28 20:43:01,483 INFO spark.SparkContext: Created broadcast 8 from showString at NativeMethodAccessorImpl.java:0
2021-03-28 20:43:01,484 INFO execution.FileSourceScanExec: Planning scan with bin packing, max size: 127283959 bytes, open cost is considered as scanning 4
194304 bytes.
2021-03-28 20:43:01,493 INFO spark.SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
2021-03-28 20:43:01,494 INFO scheduler.DAGScheduler: Got job 4 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions
2021-03-28 20:43:01,494 INFO scheduler.DAGScheduler: Final stage: ResultStage 4 (showString at NativeMethodAccessorImpl.java:0)
2021-03-28 20:43:01,494 INFO scheduler.DAGScheduler: Parents of final stage: List()
```

```
udlab.us (executor driver) (1/4)
2021-03-28 20:43:09,254 INFO executor.Executor: Finished task 3.0 in stage 8.0 (TID 16). 3764 bytes result sent to driver
2021-03-28 20:43:09,255 INFO executor.Executor: Finished task 1.0 in stage 8.0 (TID 14). 3754 bytes result sent to driver
2021-03-28 20:43:09,256 INFO scheduler.TaskSetManager: Finished task 3.0 in stage 8.0 (TID 16) in 30 ms on cc-project-13.cc-s21-project.is2750-pg0.utah.c
udlab.us (executor driver) (2/4)
2021-03-28 20:43:09,257 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 8.0 (TID 14) in 32 ms on cc-project-13.cc-s21-project.is2750-pg0.utah.c
udlab.us (executor driver) (3/4)
2021-03-28 20:43:09,262 INFO executor.Executor: Finished task 0.0 in stage 8.0 (TID 13). 3785 bytes result sent to driver
2021-03-28 20:43:09,263 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 8.0 (TID 13) in 39 ms on cc-project-13.cc-s21-project.is2750-pg0.utah.c
udlab.us (executor driver) (4/4)
2021-03-28 20:43:09,263 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 8.0, whose tasks have all completed, from pool
2021-03-28 20:43:09,263 INFO scheduler.DAGScheduler: ResultStage 8 (showString at NativeMethodAccessorImpl.java:0) finished in 0.045 s
2021-03-28 20:43:09,264 INFO scheduler.DAGScheduler: Job 6 is finished. Cancelling potential speculative or zombie tasks for this job
2021-03-28 20:43:09,264 INFO scheduler.TaskSchedulerImpl: Killing all running tasks in stage 8: Stage finished
2021-03-28 20:43:09,265 INFO scheduler.DAGScheduler: Job 6 finished: showString at NativeMethodAccessorImpl.java:0, took 0.053355 s
2021-03-28 20:43:09,275 INFO codegen.CodeGenerator: Code generated in 7.431512 ms
+----------+-----+
|      year|count|
+----------+-----+
| 0.1258848| 2051|
|0.12814848| 7885|
| 0.1249776|  115|
|0.12559104|   32|
|0.12852864| 4203|
|0.12962592| 3238|
|0.12963456| 4868|
|  0.130464| 6501|
|0.13179456| 9857|
|0.12712896| 4564|
|0.12819168| 4848|
|0.12957408| 3536|
|0.13099968| 8722|
|0.13109472| 8993|
|0.13138848| 5632|
|0.13215744| 7051|
|0.12794976| 7376|
|0.12923712| 4842|
| 0.1266624| 2222|
|0.12712032| 6862|
+----------+-----+
only showing top 20 rows

2021-03-28 20:43:09,565 INFO datasources.FileSourceStrategy: Pushed Filters:
2021-03-28 20:43:09,565 INFO datasources.FileSourceStrategy: Post-Scan Filters:
2021-03-28 20:43:09,566 INFO datasources.FileSourceStrategy: Output Data Schema: struct<_c3: string>
2021-03-28 20:43:09,642 INFO codegen.CodeGenerator: Code generated in 38.393172 ms
2021-03-28 20:43:09,652 INFO memory.MemoryStore: Block broadcast_14 stored as values in memory (estimated size 307.7 KiB, free 364.8 MiB)
2021-03-28 20:43:09,660 INFO memory.MemoryStore: Block broadcast_14_piece0 stored as bytes in memory (estimated size 27.9 KiB, free 364.8 MiB)
2021-03-28 20:43:09,661 INFO storage.BlockManagerInfo: Added broadcast_14_piece0 in memory on cc-project-13.cc-s21-project.is2750-pg0.utah.cloudlab.us:38
9 (size: 27.9 KiB, free: 366.1 MiB)
2021-03-28 20:43:09,662 INFO spark.SparkContext: Created broadcast 14 from showString at NativeMethodAccessorImpl.java:0
2021-03-28 20:43:09,663 INFO execution.FileSourceScanExec: Planning scan with bin packing, max size: 127283959 bytes, open cost is considered as scanning
194304 bytes.
2021-03-28 20:43:09,715 INFO spark.SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
2021-03-28 20:43:09,720 INFO scheduler.DAGScheduler: Registering RDD 31 (showString at NativeMethodAccessorImpl.java:0) as input to shuffle 1
2021-03-28 20:43:09,721 INFO scheduler.DAGScheduler: Got job 7 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions
2021-03-28 20:43:09,721 INFO storage.BlockManagerInfo: Removed broadcast_12_piece0 on cc-project-13.cc-s21-project.is2750-pg0.utah.cloudlab.us:38469 in m
ory (size: 16.5 KiB, free: 366.1 MiB)
2021-03-28 20:43:09,721 INFO scheduler.DAGScheduler: Final stage: ResultStage 10 (showString at NativeMethodAccessorImpl.java:0)
```

```
2021-03-28 20:43:15,361 INFO executor.Executor: Finished task 1.0 in stage 12.0 (TID 23). 3824 bytes result sent to driver
2021-03-28 20:43:15,361 INFO executor.Executor: Finished task 3.0 in stage 12.0 (TID 25). 3840 bytes result sent to driver
2021-03-28 20:43:15,362 INFO executor.Executor: Finished task 2.0 in stage 12.0 (TID 24). 3692 bytes result sent to driver
2021-03-28 20:43:15,363 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 12.0 (TID 23) in 23 ms on cc-project-13.cc-s21-project.is2750-pg0.utah.cl
oudlab.us (executor driver) (1/4)
2021-03-28 20:43:15,363 INFO scheduler.TaskSetManager: Finished task 3.0 in stage 12.0 (TID 25) in 23 ms on cc-project-13.cc-s21-project.is2750-pg0.utah.cl
oudlab.us (executor driver) (2/4)
2021-03-28 20:43:15,364 INFO scheduler.TaskSetManager: Finished task 2.0 in stage 12.0 (TID 24) in 23 ms on cc-project-13.cc-s21-project.is2750-pg0.utah.cl
oudlab.us (executor driver) (3/4)
2021-03-28 20:43:15,364 INFO executor.Executor: Finished task 0.0 in stage 12.0 (TID 22). 3856 bytes result sent to driver
2021-03-28 20:43:15,365 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 12.0 (TID 22) in 26 ms on cc-project-13.cc-s21-project.is2750-pg0.utah.cl
oudlab.us (executor driver) (4/4)
2021-03-28 20:43:15,365 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 12.0, whose tasks have all completed, from pool
2021-03-28 20:43:15,366 INFO scheduler.DAGScheduler: ResultStage 12 (showString at NativeMethodAccessorImpl.java:0) finished in 0.033 s
2021-03-28 20:43:15,366 INFO scheduler.DAGScheduler: Job 8 is finished. Cancelling potential speculative or zombie tasks for this job
2021-03-28 20:43:15,366 INFO scheduler.TaskSchedulerImpl: Killing all running tasks in stage 12: Stage finished
2021-03-28 20:43:15,366 INFO scheduler.DAGScheduler: Job 8 finished: showString at NativeMethodAccessorImpl.java:0, took 0.036463 s
2021-03-28 20:43:15,380 INFO codegen.CodeGenerator: Code generated in 11.157409 ms
+----------+-----+------------+
|      year|count|       input|
+----------+-----+------------+
| 0.1258848| 2051| [0.1258848]|
|0.12814848| 7885|[0.12814848]|
| 0.1249776|  115| [0.1249776]|
|0.12559104|   32|[0.12559104]|
|0.12852864| 4203|[0.12852864]|
|0.12962592| 3238|[0.12962592]|
|0.12963456| 4868|[0.12963456]|
|  0.130464| 6501|  [0.130464]|
|0.13179456| 9857|[0.13179456]|
|0.12712896| 4564|[0.12712896]|
|0.12819168| 4848|[0.12819168]|
|0.12957408| 3536|[0.12957408]|
|0.13099968| 8722|[0.13099968]|
|0.13109472| 8993|[0.13109472]|
|0.13138848| 5632|[0.13138848]|
|0.13215744| 7051|[0.13215744]|
|0.12794976| 7376|[0.12794976]|
|0.12923712| 4842|[0.12923712]|
| 0.1266624| 2222| [0.1266624]|
|0.12712032| 6862|[0.12712032]|
+----------+-----+------------+
only showing top 20 rows

2021-03-28 20:43:15,522 INFO datasources.FileSourceStrategy: Pushed Filters:
2021-03-28 20:43:15,523 INFO datasources.FileSourceStrategy: Post-Scan Filters:
2021-03-28 20:43:15,523 INFO datasources.FileSourceStrategy: Output Data Schema: struct<_c3: string>
2021-03-28 20:43:15,587 INFO codegen.CodeGenerator: Code generated in 40.303887 ms
2021-03-28 20:43:15,597 INFO memory.MemoryStore: Block broadcast_18 stored as values in memory (estimated size 307.7 KiB, free 364.4 MiB)
2021-03-28 20:43:15,605 INFO memory.MemoryStore: Block broadcast_18_piece0 stored as bytes in memory (estimated size 27.9 KiB, free 364.4 MiB)
2021-03-28 20:43:15,606 INFO storage.BlockManagerInfo: Added broadcast_18_piece0 in memory on cc-project-13.cc-s21-project.is2750-pg0.utah.cloudlab.us:3846
9 (size: 27.9 KiB, free: 366.1 MiB)
2021-03-28 20:43:15,607 INFO spark.SparkContext: Created broadcast 18 from rdd at Predictor.scala:80
2021-03-28 20:43:15,607 INFO execution.FileSourceScanExec: Planning scan with bin packing, max size: 127283959 bytes, open cost is considered as scanning 4
194304 bytes.
2021-03-28 20:43:15,626 INFO util.Instrumentation: [51ad6109] Stage class: LinearRegression
2021-03-28 20:43:15,627 INFO util.Instrumentation: [51ad6109] Stage uid: LinearRegression_8cc66242dc17
2021-03-28 20:43:15,651 INFO datasources.FileSourceStrategy: Pushed Filters:
2021-03-28 20:43:15,651 INFO datasources.FileSourceStrategy: Post-Scan Filters:
```

```
2021-03-28 20:43:52,441 INFO executor.Executor: Finished task 0.0 in stage 35.0 (TID 1116). 3855 bytes result sent to driver
2021-03-28 20:43:52,441 INFO executor.Executor: Finished task 8.0 in stage 35.0 (TID 1124). 3855 bytes result sent to driver
2021-03-28 20:43:52,442 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 35.0 (TID 1116) in 24 ms on cc-project-13.cc-s21-project.is2750-pg0.utah.
cloudlab.us (executor driver) (6/9)
2021-03-28 20:43:52,442 INFO scheduler.TaskSetManager: Finished task 8.0 in stage 35.0 (TID 1124) in 6 ms on cc-project-13.cc-s21-project.is2750-pg0.utah.c
loudlab.us (executor driver) (7/9)
2021-03-28 20:43:52,442 INFO executor.Executor: Finished task 7.0 in stage 35.0 (TID 1123). 3855 bytes result sent to driver
2021-03-28 20:43:52,443 INFO executor.Executor: Finished task 6.0 in stage 35.0 (TID 1122). 3953 bytes result sent to driver
2021-03-28 20:43:52,443 INFO scheduler.TaskSetManager: Finished task 7.0 in stage 35.0 (TID 1123) in 10 ms on cc-project-13.cc-s21-project.is2750-pg0.utah.
cloudlab.us (executor driver) (8/9)
2021-03-28 20:43:52,443 INFO scheduler.TaskSetManager: Finished task 6.0 in stage 35.0 (TID 1122) in 13 ms on cc-project-13.cc-s21-project.is2750-pg0.utah.
cloudlab.us (executor driver) (9/9)
2021-03-28 20:43:52,443 INFO scheduler.TaskSchedulerImpl: Removed TaskSet 35.0, whose tasks have all completed, from pool
2021-03-28 20:43:52,443 INFO scheduler.DAGScheduler: ResultStage 35 (showString at NativeMethodAccessorImpl.java:0) finished in 0.029 s
2021-03-28 20:43:52,443 INFO scheduler.DAGScheduler: Job 17 is finished. Cancelling potential speculative or zombie tasks for this job
2021-03-28 20:43:52,443 INFO scheduler.TaskSchedulerImpl: Killing all running tasks in stage 35: Stage finished
2021-03-28 20:43:52,444 INFO scheduler.DAGScheduler: Job 17 finished: showString at NativeMethodAccessorImpl.java:0, took 0.031195 s
2021-03-28 20:43:52,452 INFO codegen.CodeGenerator: Code generated in 7.517669 ms
+----------+-----+------------+------------------+
|      year|count|       input|        prediction|
+----------+-----+------------+------------------+
| 0.1258848| 2051| [0.1258848]|2206.8133906202274|
| 0.1266624| 2222| [0.1266624]|3119.6718135910924|
| 0.1302048|10496| [0.1302048]| 7278.249073791638|
|0.12797568| 8927|[0.12797568]| 4661.388261275191|
|0.13038624| 5853|[0.13038624]| 7491.249372484803|
|0.13101696| 5085|[0.13101696]| 8231.678982227837|
|0.12664512| 6546|[0.12664512]| 3099.386070858396|
| 0.1271376| 3256| [0.1271376]|3677.5297387399187|
| 0.1301184| 5622| [0.1301184]| 7176.820360128186|
|0.13042944| 4668|[0.13042944]| 7541.963729316543|
| 0.1248912|   34| [0.1248912]| 1040.383183490805|
|0.13021344| 7708|[0.13021344]| 7288.391945157986|
|0.13004064| 8814|[0.13004064]| 7085.534517831111|
|0.12794112| 4848|[0.12794112]| 4620.816775809799|
| 0.1258416|  662| [0.1258416]| 2156.099033788516|
| 0.1281312| 7004| [0.1281312]| 4843.959945869341|
|0.12888288| 4756|[0.12888288]| 5726.389754741191|
|0.12807936| 7000|[0.12807936]|4783.1027176713105|
|0.12877056| 3512|[0.12877056]| 5594.532426978723|
|0.13005792| 5772|[0.13005792]| 7105.820260563778|
+----------+-----+------------+------------------+
only showing top 20 rows

2021-03-28 20:43:52,508 INFO Configuration.deprecation: mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputformat.outputdir
2021-03-28 20:43:52,511 INFO io.HadoopMapRedCommitProtocol: Using output committer class org.apache.hadoop.mapred.FileOutputCommitter
2021-03-28 20:43:52,513 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
2021-03-28 20:43:52,513 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup f
ailures: false
```



The model is saved with name linear_reg_spark_part3_model
To check the saved model execute: hdfs dfs -ls linear_reg_spark_part3_model

```
ubuntu@cc-project-13:~$ hdfs dfs -ls linear_reg_spark_part3_model
Found 2 items
drwxr-xr-x   - ubuntu supergroup          0 2021-03-28 20:46 linear_reg_spark_part3_model/data
drwxr-xr-x   - ubuntu supergroup          0 2021-03-28 20:46 linear_reg_spark_part3_model/metadata
```

execute: hdfs dfs -ls linear_reg_spark_part3_model/data

```
ubuntu@cc-project-13:~$ hdfs dfs -ls linear_reg_spark_part3_model/data/
Found 2 items
-rw-r--r--   3 ubuntu supergroup          0 2021-03-28 20:46 linear_reg_spark_part3_model/data/_SUCCESS
-rw-r--r--   3 ubuntu supergroup       1943 2021-03-28 20:46 linear_reg_spark_part3_model/data/part-00000-3bcc953f-7374-4e09-95fa-c8c2084df42a-c000.snappy.
parquet
ubuntu@cc-project-13:~$
```