

INFSCI 2750 - Mini-Project 2: Output Snapshot - Part 2 and Part 3

Shreyank Ranganath, shr74@pitt.edu

Kwesi Aguilera, kra40@pitt.edu

Aishwarya Jakka, aij12@pitt.edu

Part 2:

Total Listening Counts of Each Artist in Descending Order

```
Shreyank — student@CC-MON-3: ~/Mini_proj_2 — ssh -i ~/.ssh/key_student student@165.227.73.164 — 204x99
2020-04-04 19:30:16,123 INFO scheduler.TaskSetManager: Finished task 197.0 in stage 9.0 (TID 206) in 9 ms on CC-MON-5 (executor 2) (198/200)
2020-04-04 19:30:16,130 INFO scheduler.TaskSetManager: Starting task 199.0 in stage 9.0 (TID 208, CC-MON-5, executor 2, partition 199, NODE_LOCAL, 7778 bytes)
2020-04-04 19:30:16,132 INFO scheduler.TaskSetManager: Finished task 198.0 in stage 9.0 (TID 207) in 10 ms on CC-MON-5 (executor 2) (199/200)
2020-04-04 19:30:16,142 INFO scheduler.TaskSetManager: Finished task 199.0 in stage 9.0 (TID 208) in 12 ms on CC-MON-5 (executor 2) (200/200)
2020-04-04 19:30:16,142 INFO cluster.YarnScheduler: Removed TaskSet 9.0, whose tasks have all completed, from pool
2020-04-04 19:30:16,143 INFO scheduler.DAGScheduler: ResultStage 9 (showString at NativeMethodAccessorImpl.java:0) finished in 3.778 s
2020-04-04 19:30:16,143 INFO scheduler.DAGScheduler: Job 5 finished: showString at NativeMethodAccessorImpl.java:0, took 4.679142 s
2020-04-04 19:30:16,193 INFO codegen.CodeGenerator: Code generated in 23.318151 ms

+-----+
|artistID|listeningCounts|
+-----+
|289|2393140.0|
|72|1301308.0|
|89|1291387.0|
|292|1058405.0|
|498|963449.0|
|67|921198.0|
|288|905423.0|
|701|688529.0|
|227|662116.0|
|300|532545.0|
|333|525844.0|
|344|525292.0|
|378|513476.0|
|679|506453.0|
|295|499318.0|
|511|493024.0|
|461|489065.0|
|486|485532.0|
|190|485076.0|
|163|466104.0|
|55|449292.0|
|154|385306.0|
|466|384405.0|
|257|384307.0|
|707|371916.0|
|917|368710.0|
|792|350035.0|
|51|348919.0|
|65|330757.0|
|475|321011.0|
|203|310221.0|
|157|296882.0|
|207|288520.0|
|198|277397.0|
|377|265362.0|
|291|253027.0|
|614|251440.0|
|173|245878.0|
|503|237148.0|
|687|215777.0|
|903|213103.0|
|302|207761.0|
|187|205195.0|
|1412|200665.0|
|1098|200278.0|
|1672|200049.0|
|458|200027.0|
|229|191979.0|
|234|190232.0|
|306|188634.0|
|56|176043.0|
|325|166644.0|
|533|165975.0|
|294|162288.0|
|233|160317.0|
|209|159733.0|
|230|155321.0|
|455|153101.0|
|159|151103.0|
|228|148452.0|
|299|144710.0|
|681|139378.0|
|704|138049.0|
|2044|137581.0|
|418|133955.0|
|1249|133931.0|
|599|133511.0|
|1369|131307.0|
|349|130855.0|
|424|130227.0|
|1104|129745.0|
|298|129352.0|
|1459|126314.0|
|318|123682.0|
|285|120665.0|
|1246|120336.0|
|316|117229.0|
|220|116768.0|
|544|115538.0|
|680|114690.0|
|959|114505.0|
|195|113631.0|
|689|113467.0|
|310|110218.0|
|686|110123.0|
|918|102662.0|
|412|102334.0|
|439|100819.0|
```

End of Output (snapshot)

```
Shreyank — student@CC-MON-3: ~/Mini_proj_2 — ssh -i ~/.ssh/key_student student@165.227.73.164 — 204x99
| 14375| 1.0|
| 18432| 1.0|
| 14376| 1.0|
| 14370| 1.0|
| 12401| 1.0|
| 17263| 1.0|
| 7633| 1.0|
| 17467| 1.0|
| 11248| 1.0|
| 8052| 1.0|
| 17272| 1.0|
| 17171| 1.0|
| 17170| 1.0|
| 17273| 1.0|
| 5920| 1.0|
| 17166| 1.0|
| 9488| 1.0|
| 13823| 1.0|
| 9489| 1.0|
| 9491| 1.0|
| 9895| 1.0|
| 9581| 1.0|
| 1173| 1.0|
| 9584| 1.0|
| 10174| 1.0|
| 15945| 1.0|
| 17265| 1.0|
| 15942| 1.0|
| 13829| 1.0|
| 9492| 1.0|
| 11250| 1.0|
| 14371| 1.0|
| 16804| 1.0|
| 7752| 1.0|
| 15934| 1.0|
| 17169| 1.0|
| 13825| 1.0|
| 9580| 1.0|
| 11750| 1.0|
| 11748| 1.0|
| 7631| 1.0|
| 7627| 1.0|
| 9479| 1.0|
| 11751| 1.0|
| 14373| 1.0|
| 13712| 1.0|
| 14308| 1.0|
| 13822| 1.0|
| 11747| 1.0|
| 13817| 1.0|
| 15937| 1.0|
| 15749| 1.0|
| 17466| 1.0|
| 13826| 1.0|
| 17164| 1.0|
| 17080| 1.0|
| 17264| 1.0|
| 9481| 1.0|
| 15939| 1.0|
| 13818| 1.0|
| 13761| 1.0|
| 13814| 1.0|
| 15940| 1.0|
| 17262| 1.0|
| 17172| 1.0|
| 13812| 1.0|
| 13103| 1.0|
| 17267| 1.0|
| 17165| 1.0|
| 2897| 1.0|
| 9894| 1.0|
| 14524| 1.0|
| 4552| 1.0|
| 9482| 1.0|
| 17269| 1.0|
| 5921| 1.0|
| 17468| 1.0|
|-----|-----|
2020-04-04 19:30:16,625 INFO ui.SparkUI: Stopped Spark web UI at http://CC-MON-3:4040
2020-04-04 19:30:16,631 INFO cluster.YarnClientSchedulerBackend: Interrupting monitor thread
2020-04-04 19:30:16,649 INFO cluster.YarnClientSchedulerBackend: Shutting down all executors
2020-04-04 19:30:16,649 INFO cluster.YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
2020-04-04 19:30:16,654 INFO cluster.SchedulerExtensionServices: Stopping SchedulerExtensionServices
(serviceOption=None,
services=List(),
started=false)
2020-04-04 19:30:16,656 INFO cluster.YarnClientSchedulerBackend: Stopped
2020-04-04 19:30:16,667 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
2020-04-04 19:30:16,689 INFO memory.MemoryStore: MemoryStore cleared
2020-04-04 19:30:16,689 INFO storage.BlockManager: BlockManager stopped
2020-04-04 19:30:16,696 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
2020-04-04 19:30:16,721 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
2020-04-04 19:30:16,731 INFO spark.SparkContext: Successfully stopped SparkContext
2020-04-04 19:30:16,990 INFO util.ShutdownHookManager: Shutdown hook called
2020-04-04 19:30:16,991 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-6852f4aa-b1d0-4732-83fc-142c7c6b1f15
2020-04-04 19:30:16,994 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-7017301f-29fd-40ae-bb03-ec2855e590c
2020-04-04 19:30:16,997 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-6852f4aa-b1d0-4732-83fc-142c7c6b1f15/pyspark-7060a055-8771-473d-b6f7-857934483df3
student@CC-MON-3:~/Mini_proj_2$
```

Part 3:

Snapshots of data at various stages of transformations

Spark - yarn - diagnostic message

```
2020-04-04 19:40:26,169 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acts disabled; users with view permissions: Set(student); groups with view permissions: Set(); users With
modify permissions: Set(student); groups with modify permissions: Set()
2020-04-04 19:40:27,637 INFO yarn.Client: Submitting application application_1586029185768_0001 to ResourceManager
2020-04-04 19:40:29,180 INFO impl.YarnClientImpl: Submitted application application_1586029185768_0001
2020-04-04 19:40:29,186 INFO cluster.SchedulerExtensionServices: Starting Yarn extension services with app application_1586029185768_0001 and attemptId None
2020-04-04 19:40:29,201 INFO yarn.Client: Application report for application_1586029185768_0001 (state: ACCEPTED)
2020-04-04 19:40:29,205 INFO yarn.Client:
  client token: N/A
  diagnostics: AM container is launched, waiting for AM container to register with RM
  ApplicationMaster host: N/A
  ApplicationMaster RPC port: -1
  queue: default
  start time: 1586029227814
  final status: UNDEFINED
  tracking URL: http://CC-MON-3:8088/proxy/application_1586029185768_0001/
  user: student
2020-04-04 19:40:30,210 INFO yarn.Client: Application report for application_1586029185768_0001 (state: ACCEPTED)
2020-04-04 19:40:31,214 INFO yarn.Client: Application report for application_1586029185768_0001 (state: ACCEPTED)
2020-04-04 19:40:32,220 INFO yarn.Client: Application report for application_1586029185768_0001 (state: ACCEPTED)
2020-04-04 19:40:33,223 INFO yarn.Client: Application report for application_1586029185768_0001 (state: ACCEPTED)
2020-04-04 19:40:34,227 INFO yarn.Client: Application report for application_1586029185768_0001 (state: ACCEPTED)
2020-04-04 19:40:34,910 INFO cluster.YarnClientSchedulerBackend: Add WebUI Filter. org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter, Map(PROXY_HOSTS -> CC-MON-3, PROXY_URI_BASES -> http://CC-MON-3:8088/proxy/application_1586029185768_0001), /proxy/application_1586029185768_0001
2020-04-04 19:40:35,057 INFO cluster.YarnSchedulerBackendsYarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnAM)
2020-04-04 19:40:35,233 INFO yarn.Client: Application report for application_1586029185768_0001 (state: RUNNING)
2020-04-04 19:40:35,231 INFO yarn.Client:
  client token: N/A
  diagnostics: N/A
  ApplicationMaster host: 10.17.0.36
  ApplicationMaster RPC port: -1
  queue: default
  start time: 1586030227814
  final status: UNDEFINED
  tracking URL: http://CC-MON-3:8088/proxy/application_1586029185768_0001/
  user: student
2020-04-04 19:40:35,234 INFO cluster.YarnClientSchedulerBackend: Application application_1586029185768_0001 has started running.
2020-04-04 19:40:35,249 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 37013.
```

Data transformation outputs:

Read data frame from HDFS

```
Shreyank — student@CC-MON-3: ~/Mini_proj_2 — ssh -i ~/.ssh/key_student student@165.227.73.164 — 204x99
2020-04-04 19:40:49,569 INFO scheduler.DAGScheduler: ResultStage 1 (showString at NativeMethodAccessorImpl.java:0) finished in 0.343 s
2020-04-04 19:40:49,572 INFO scheduler.DAGScheduler: Job 1 finished: showString at NativeMethodAccessorImpl.java:0, took 0.349317 s
```

	_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

```
None
2020-04-04 19:40:50,018 INFO datasources.FileSourceStrategy: Pruning directories with:
2020-04-04 19:40:50,018 INFO datasources.FileSourceStrategy: Post-Scan Filters:
2020-04-04 19:40:50,019 INFO datasources.FileSourceStrategy: Output Data Schema: struct<_c0: string, _c3: string>
2020-04-04 19:40:50,019 INFO execution.FileSourceScanExec: Pushed Filters:
root
|-- IP: string (nullable = true)
|-- Time: string (nullable = true)
None
2020-04-04 19:40:50,163 INFO codegen.CodeGenerator: Code generated in 34.846832 ms
2020-04-04 19:40:50,207 INFO codegen.CodeGenerator: Code generated in 23.023218 ms
```

Select only columns 0 and 3

```
Shreyank — student@CC-MON-3: ~/Mini_proj_2 — ssh -i ~/.ssh/key_student student@165.227.73.164 — 204x99
2020-04-04 19:40:55,803 INFO storage.BlockManagerInfo: Added rdd_16_0 in memory on CC-MON-5:40113 (size: 5.9 MB, free: 360.3 MB)
2020-04-04 19:40:55,907 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 5514 ms on CC-MON-5 (executor 2) (1/1)
2020-04-04 19:40:55,908 INFO cluster.YarnScheduler: Removed TaskSet 2.0, whose tasks have all completed, from pool
2020-04-04 19:40:55,909 INFO scheduler.DAGScheduler: ResultStage 2 (showString at NativeMethodAccessorImpl.java:0) finished in 5.566 s
2020-04-04 19:40:55,909 INFO scheduler.DAGScheduler: Job 2 finished: showString at NativeMethodAccessorImpl.java:0, took 5.579907 s

+-----+-----+
|      IP|      Time|
+-----+-----+
|10.223.157.186|[15/Jul/2009:14:5...|
|10.223.157.186|[15/Jul/2009:14:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
|10.223.157.186|[15/Jul/2009:15:5...|
+-----+-----+

only showing top 20 rows

None
['IP', 'Time']
2020-04-04 19:40:56,185 INFO codegen.CodeGenerator: Code generated in 46.969244 ms
2020-04-04 19:40:56,226 INFO spark.SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
2020-04-04 19:40:56,227 INFO scheduler.DAGScheduler: Got job 3 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions
```

Remove '[' from Time column and convert times to datetime format

```
Shreyank — student@CC-MON-3: ~/Mini_proj_2 — ssh -i ~/.ssh/key_student student@165.227.73.164 — 204x99
2020-04-04 19:40:56,384 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 136 ms on CC-MON-5 (executor 2) (1/1)
2020-04-04 19:40:56,385 INFO scheduler.DAGScheduler: ResultStage 3 (showString at NativeMethodAccessorImpl.java:0) finished in 0.155 s
2020-04-04 19:40:56,387 INFO scheduler.DAGScheduler: Job 3 finished: showString at NativeMethodAccessorImpl.java:0, took 0.160773 s

+-----+-----+
|      IP|      Time|
+-----+-----+
|10.223.157.186|2009-07-15 14:58:59|
|10.223.157.186|2009-07-15 14:58:59|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:35|
|10.223.157.186|2009-07-15 15:50:37|
+-----+-----+

only showing top 20 rows

None
2020-04-04 19:40:56,396 INFO cluster.YarnScheduler: Removed TaskSet 3.0, whose tasks have all completed, from pool
2020-04-04 19:40:56,842 INFO codegen.CodeGenerator: Code generated in 28.58417 ms
2020-04-04 19:40:56,918 INFO codegen.CodeGenerator: Code generated in 55.2152 ms
```

Get counts of IPs and Group by month of year.

```
Shreyank — student@CC-MON-3: ~/Mini_proj_2 — ssh -i ~/.ssh/key_student student@165.227.73.164 — 204x99
2020-04-04 19:42:29,208 INFO cluster.YarnScheduler: Removed TaskSet 14.0, whose tasks have all completed, from pool
2020-04-04 19:42:29,208 INFO scheduler.DAGScheduler: ResultStage 14 (showString at NativeMethodAccessorImpl.java:0) finished in 0.270 s
2020-04-04 19:42:29,209 INFO scheduler.DAGScheduler: Job 7 finished: showString at NativeMethodAccessorImpl.java:0, took 0.277436 s

+-----+-----+-----+
|year(Time)|month(Time)|count(IP)|
+-----+-----+-----+
|null|null|26|
|2009|7|1253|
|2009|8|3798|
|2009|9|2696|
|2009|10|7347|
|2009|11|19211|
|2009|12|15911|
|2010|1|100120|
|2010|2|113089|
|2010|3|144044|
|2010|4|106716|
|2010|5|124169|
|2010|6|148563|
|2010|7|197091|
|2010|8|177332|
|2010|9|144625|
|2010|10|140729|
|2010|11|163713|
|2010|12|152237|
|2011|1|172976|
+-----+-----+-----+

only showing top 20 rows

None
2020-04-04 19:42:29,407 INFO codegen.CodeGenerator: Code generated in 24.110387 ms
2020-04-04 19:42:29,438 INFO codegen.CodeGenerator: Code generated in 16.255335 ms
2020-04-04 19:42:29,486 INFO spark.ContextCleaner: Cleaned accumulator 230
2020-04-04 19:42:29,486 INFO spark.ContextCleaner: Cleaned accumulator 294
```

Concatenate month and year columns

```
Shreyank — student@CC-MON-3: ~/Mini_proj_2 — ssh -i ~/.ssh/key_student student@165.227.73.164 — 204x99
2020-04-04 19:42:29,838 INFO scheduler.DAGScheduler: ResultStage 23 (showString at NativeMethodAccessorImpl.java:0) finished in 0.148 s
2020-04-04 19:42:29,840 INFO scheduler.DAGScheduler: Job 10 finished: showString at NativeMethodAccessorImpl.java:0, took 0.153694 s

+-----+-----+
|count(IP)|Time|
+-----+-----+
|26|null|
|1253|2009/7|
|3798|2009/8|
|2696|2009/9|
|7347|2009/10|
|19211|2009/11|
|15911|2009/12|
|100120|2010/1|
|113089|2010/2|
|144044|2010/3|
|106716|2010/4|
|124169|2010/5|
|148563|2010/6|
|197091|2010/7|
|177332|2010/8|
|144625|2010/9|
|140729|2010/10|
|163713|2010/11|
|152237|2010/12|
|172976|2011/1|
+-----+-----+

only showing top 20 rows

None
2020-04-04 19:42:29,985 INFO codegen.CodeGenerator: Code generated in 20.676409 ms
2020-04-04 19:42:30,009 INFO spark.SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
2020-04-04 19:42:30,010 INFO scheduler.DAGScheduler: Got job 11 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions
```

Convert Time column to timestamp format

```
Shreyank — student@CC-MON-3: ~/Mini_proj_2 — ssh -i ~/.ssh/key_student student@165.227.73.164 — 204x99
2020-04-04 19:42:30,334 INFO scheduler.DAGScheduler: ResultStage 32 (showString at NativeMethodAccessorImpl.java:0) finished in 0.158 s
2020-04-04 19:42:30,335 INFO scheduler.DAGScheduler: Job 13 finished: showString at NativeMethodAccessorImpl.java:0, took 0.163998 s

+-----+-----+
| count(IP) | Time |
+-----+-----+
| 26 | null |
| 1253 | 2009-07-01 |
| 3798 | 2009-08-01 |
| 2696 | 2009-09-01 |
| 7347 | 2009-10-01 |
| 19211 | 2009-11-01 |
| 15911 | 2009-12-01 |
| 100120 | 2010-01-01 |
| 113089 | 2010-02-01 |
| 144044 | 2010-03-01 |
| 106716 | 2010-04-01 |
| 124169 | 2010-05-01 |
| 148563 | 2010-06-01 |
| 197091 | 2010-07-01 |
| 177332 | 2010-08-01 |
| 144625 | 2010-09-01 |
| 140729 | 2010-10-01 |
| 163713 | 2010-11-01 |
| 152237 | 2010-12-01 |
| 172976 | 2011-01-01 |
+-----+-----+
only showing top 20 rows

None
2020-04-04 19:42:30,469 INFO codegen.CodeGenerator: Code generated in 23.448996 ms
2020-04-04 19:42:30,508 INFO spark.SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
2020-04-04 19:42:30,510 INFO scheduler.DAGScheduler: Got job 14 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions
2020-04-04 19:42:30,511 INFO scheduler.DAGScheduler: Final stage: ResultStage 35 (showString at NativeMethodAccessorImpl.java:0)
2020-04-04 19:42:30,511 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 34)
```

Remove null values

```
Shreyank — student@CC-MON-3: ~/Mini_proj_2 — ssh -i ~/.ssh/key_student student@165.227.73.164 — 204x99
2020-04-04 19:42:30,865 INFO cluster.YarnScheduler: Removed TaskSet 41.0, whose tasks have all completed, from pool
2020-04-04 19:42:30,866 INFO scheduler.DAGScheduler: ResultStage 41 (showString at NativeMethodAccessorImpl.java:0) finished in 0.152 s
2020-04-04 19:42:30,866 INFO scheduler.DAGScheduler: Job 16 finished: showString at NativeMethodAccessorImpl.java:0, took 0.159241 s

+-----+-----+
| IP | Time |
+-----+-----+
| 1253 | 2009-07-01 |
| 3798 | 2009-08-01 |
| 2696 | 2009-09-01 |
| 7347 | 2009-10-01 |
| 19211 | 2009-11-01 |
| 15911 | 2009-12-01 |
| 100120 | 2010-01-01 |
| 113089 | 2010-02-01 |
| 144044 | 2010-03-01 |
| 106716 | 2010-04-01 |
| 124169 | 2010-05-01 |
| 148563 | 2010-06-01 |
| 197091 | 2010-07-01 |
| 177332 | 2010-08-01 |
| 144625 | 2010-09-01 |
| 140729 | 2010-10-01 |
| 163713 | 2010-11-01 |
| 152237 | 2010-12-01 |
| 172976 | 2011-01-01 |
| 237796 | 2011-02-01 |
+-----+-----+
only showing top 20 rows

None
2020-04-04 19:42:31,011 INFO codegen.CodeGenerator: Code generated in 43.570649 ms
2020-04-04 19:42:31,033 INFO spark.SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
2020-04-04 19:42:31,036 INFO scheduler.DAGScheduler: Got job 17 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions
2020-04-04 19:42:31,036 INFO scheduler.DAGScheduler: Final stage: ResultStage 44 (showString at NativeMethodAccessorImpl.java:0)
2020-04-04 19:42:31,036 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 43)
```

Convert timestamp to unix_timestamp - (to ordinal function not available in PySpark)

```
Shreyank — student@CC-MON-3: ~/Mini_proj_2 — ssh -i ~/.ssh/key_student student@165.227.73.164 — 204x99
2020-04-04 19:42:31,327 INFO scheduler.TaskSetManager: Finished task 18.0 in stage 50.0 (TID 533) in 10 ms on CC-MON-5 (executor 2) (17/20)
2020-04-04 19:42:31,328 INFO scheduler.TaskSetManager: Starting task 16.0 in stage 50.0 (TID 535, CC-MON-4, executor 1, partition 21, PROCESS_LOCAL, 7778 bytes)
2020-04-04 19:42:31,329 INFO scheduler.TaskSetManager: Finished task 15.0 in stage 50.0 (TID 532) in 14 ms on CC-MON-4 (executor 1) (18/20)
2020-04-04 19:42:31,337 INFO scheduler.TaskSetManager: Finished task 19.0 in stage 50.0 (TID 534) in 10 ms on CC-MON-5 (executor 2) (19/20)
2020-04-04 19:42:31,340 INFO cluster.YarnScheduler: Removed TaskSet 50.0, whose tasks have all completed, from pool
2020-04-04 19:42:31,341 INFO scheduler.DAGScheduler: ResultStage 50 (showString at NativeMethodAccessorImpl.java:0) finished in 0.136 s
2020-04-04 19:42:31,342 INFO scheduler.DAGScheduler: Job 19 finished: showString at NativeMethodAccessorImpl.java:0, took 0.140331 s

+----+-----+
| IP | Time |
+----+-----+
| 1253 | 1246406400 |
| 3798 | 1249084800 |
| 2696 | 1251763200 |
| 7347 | 1254355200 |
| 19211 | 1257033600 |
| 15911 | 1259625600 |
| 100120 | 1262304000 |
| 113089 | 1264982400 |
| 144044 | 1267401600 |
| 106716 | 1270080000 |
| 124169 | 1272672000 |
| 148563 | 1275350400 |
| 197091 | 1277942400 |
| 177332 | 1280620800 |
| 144625 | 1283299200 |
| 140729 | 1285891200 |
| 163713 | 1288569600 |
| 152237 | 1291161600 |
| 172976 | 1293840000 |
| 237796 | 1296518400 |
+----+-----+
only showing top 20 rows

None
2020-04-04 19:42:31,832 INFO codegen.CodeGenerator: Code generated in 20.676515 ms
2020-04-04 19:42:31,910 INFO spark.ContextCleaner: Cleaned accumulator 388
2020-04-04 19:42:31,910 INFO spark.ContextCleaner: Cleaned accumulator 625
2020-04-04 19:42:31,910 INFO spark.ContextCleaner: Cleaned accumulator 538
2020-04-04 19:42:31,910 INFO spark.ContextCleaner: Cleaned accumulator 478
2020-04-04 19:42:31,910 INFO spark.ContextCleaner: Cleaned accumulator 377
2020-04-04 19:42:31,910 INFO spark.ContextCleaner: Cleaned accumulator 611
2020-04-04 19:42:31,910 INFO spark.ContextCleaner: Cleaned accumulator 647
2020-04-04 19:42:31,910 INFO spark.ContextCleaner: Cleaned accumulator 448
2020-04-04 19:42:31,921 INFO storage.BlockManagerInfo: Removed broadcast_18_piece0 on CC-MON-4:42407 in memory (size: 18.0 KB, free: 353.9 MB)
2020-04-04 19:42:31,922 INFO storage.BlockManagerInfo: Removed broadcast_18_piece0 on CC-MON-5:40113 in memory (size: 18.0 KB, free: 347.8 MB)
2020-04-04 19:42:31,948 INFO storage.BlockManagerInfo: Removed broadcast_18_piece0 on CC-MON-3:37813 in memory (size: 18.0 KB, free: 365.9 MB)
```

Linear Regression model coefficients, intercept and Train R2 Score

```
Shreyank — student@CC-MON-3: ~/Mini_proj_2 — ssh -i ~/.ssh/key_student student@165.227.73.164 — 204x63

+----+-----+
| IP | Time |
+----+-----+
| 1253 | 1246406400 |
| 3798 | 1249084800 |
| 2696 | 1251763200 |
| 7347 | 1254355200 |
| 19211 | 1257033600 |
| 15911 | 1259625600 |
| 100120 | 1262304000 |
| 113089 | 1264982400 |
| 144044 | 1267401600 |
| 106716 | 1270080000 |
| 124169 | 1272672000 |
| 148563 | 1275350400 |
| 197091 | 1277942400 |
| 177332 | 1280620800 |
| 144625 | 1283299200 |
| 140729 | 1285891200 |
| 163713 | 1288569600 |
| 152237 | 1291161600 |
| 172976 | 1293840000 |
| 237796 | 1296518400 |
+----+-----+
only showing top 20 rows

None
2020-04-04 19:56:36,579 WARN netlib.BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS
2020-04-04 19:56:36,580 WARN netlib.BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
Coefficients: [0.499999999396212624,0.499999999396212624]
Intercept: 15.532792436812661
Train R2 Score: 0.9999999999999999
```

spark.stop()


```
2020-04-04 19:44:26,510 INFO spark.SparkContext: Invoking stop() from shutdown hook
2020-04-04 19:44:26,520 INFO server.AbstractConnector: Stopped Spark@690a3f90(HTTP/1.1,[http/1.1]){0.0.0.0:4040}
2020-04-04 19:44:26,533 INFO ui.SparkUI: Stopped Spark web UI at http://CC-MON-3:4040
2020-04-04 19:44:26,540 INFO cluster.YarnClientSchedulerBackend: Interrupting monitor thread
2020-04-04 19:44:26,550 INFO cluster.YarnClientSchedulerBackend: Shutting down all executors
2020-04-04 19:44:26,560 INFO cluster.YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
2020-04-04 19:44:26,567 INFO cluster.SchedulerExtensionServices: Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
2020-04-04 19:44:26,569 INFO cluster.YarnClientSchedulerBackend: Stopped
2020-04-04 19:44:26,585 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
2020-04-04 19:44:26,597 INFO memory.MemoryStore: MemoryStore cleared
2020-04-04 19:44:26,598 INFO storage.BlockManager: BlockManager stopped
2020-04-04 19:44:26,599 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
2020-04-04 19:44:26,602 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
2020-04-04 19:44:26,612 INFO spark.SparkContext: Successfully stopped SparkContext
2020-04-04 19:44:26,613 INFO util.ShutdownHookManager: Shutdown hook called
2020-04-04 19:44:26,617 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-5ef561ca-d40f-47a7-b303-88d6727342cd/pyspark-66fed44b-39fb-4eb0-bc2c-235bb94e9d4d
2020-04-04 19:44:26,626 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-8ba4aee2-79e9-4045-8079-786812a660e2
2020-04-04 19:44:26,636 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-5ef561ca-d40f-47a7-b303-88d6727342cd
student@CC-MON-3:~/Mini_proj_2$
```