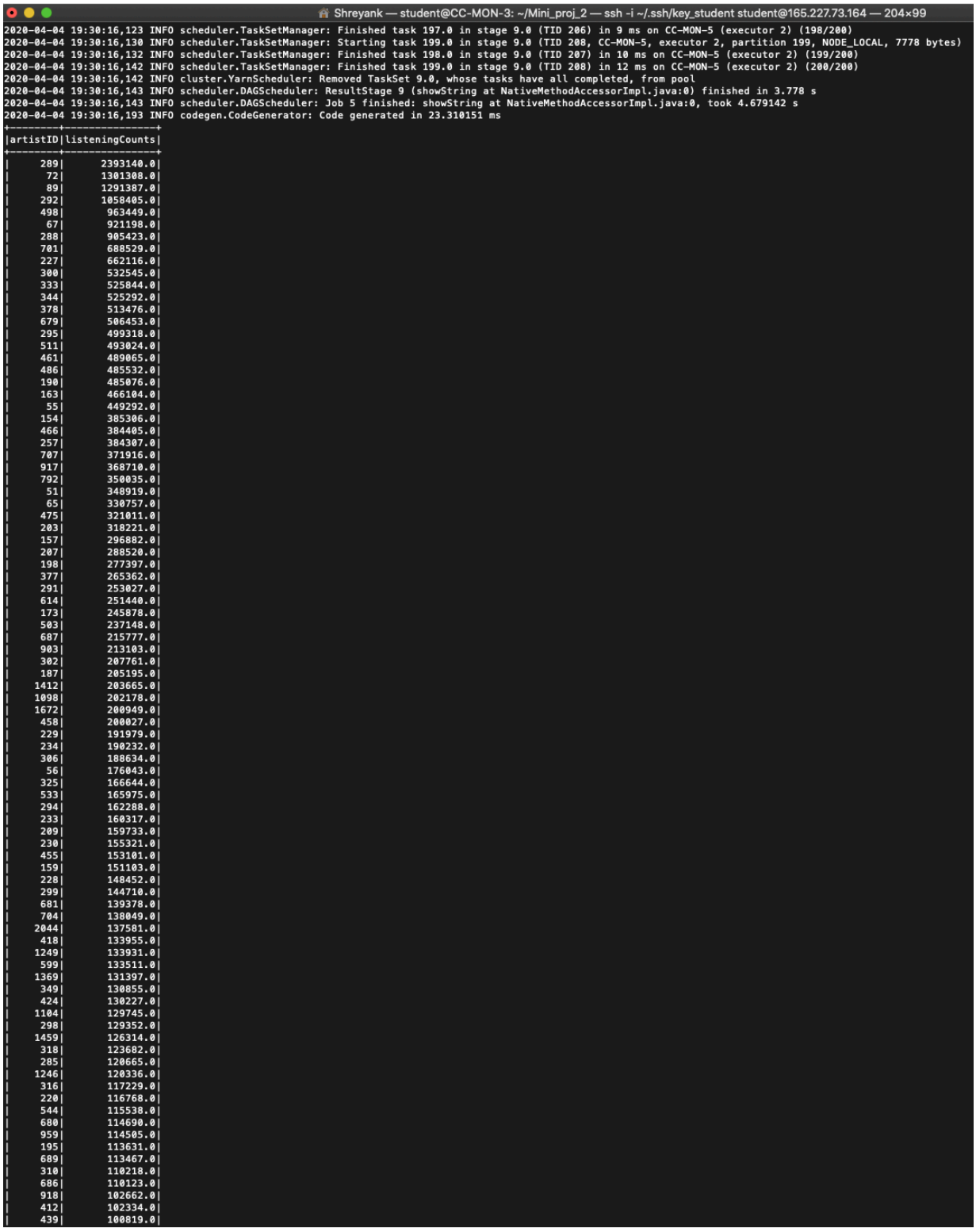
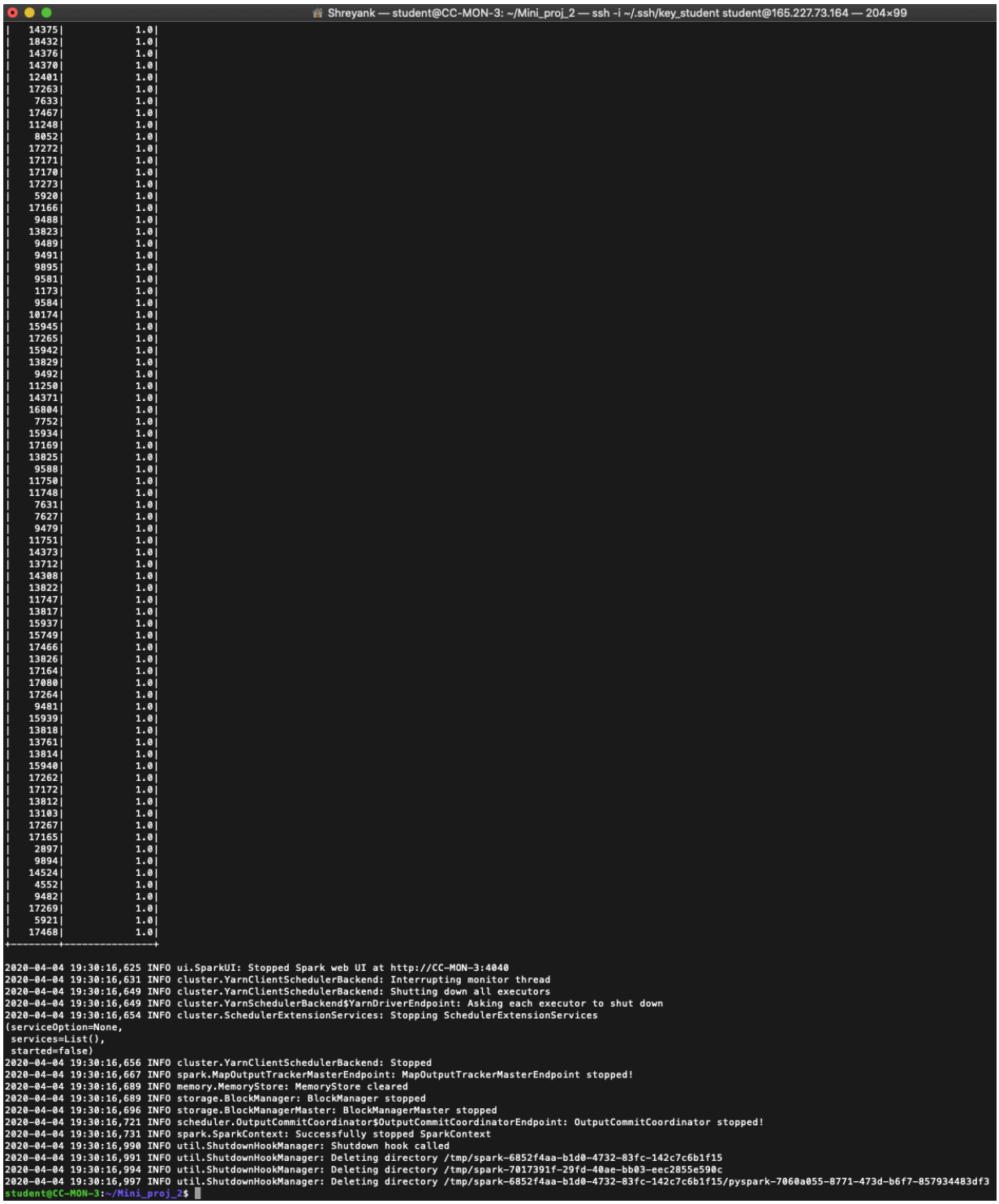
INFSCI 2750 - Mini-Project 2: Output Snapshot - Part 2 and Part 3

Shreyank Ranganath, shr74@pitt.edu
Kwesi Aguillera, kra40@pitt.edu
Aishwarya Jakka, aij12@pitt.edu

Part 2:
Total Listening Counts of Each Artist in Descending Order

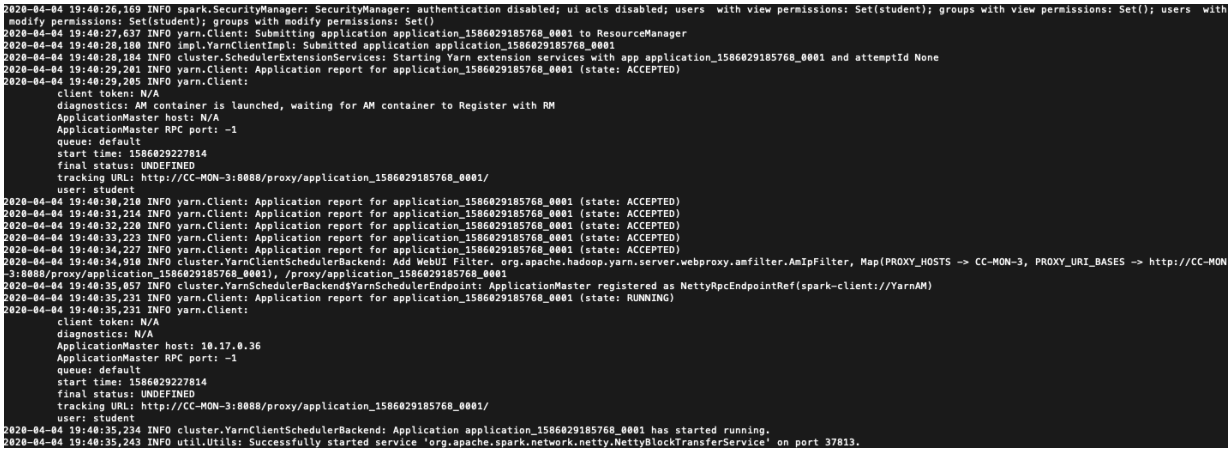


End of Output (snapshot)



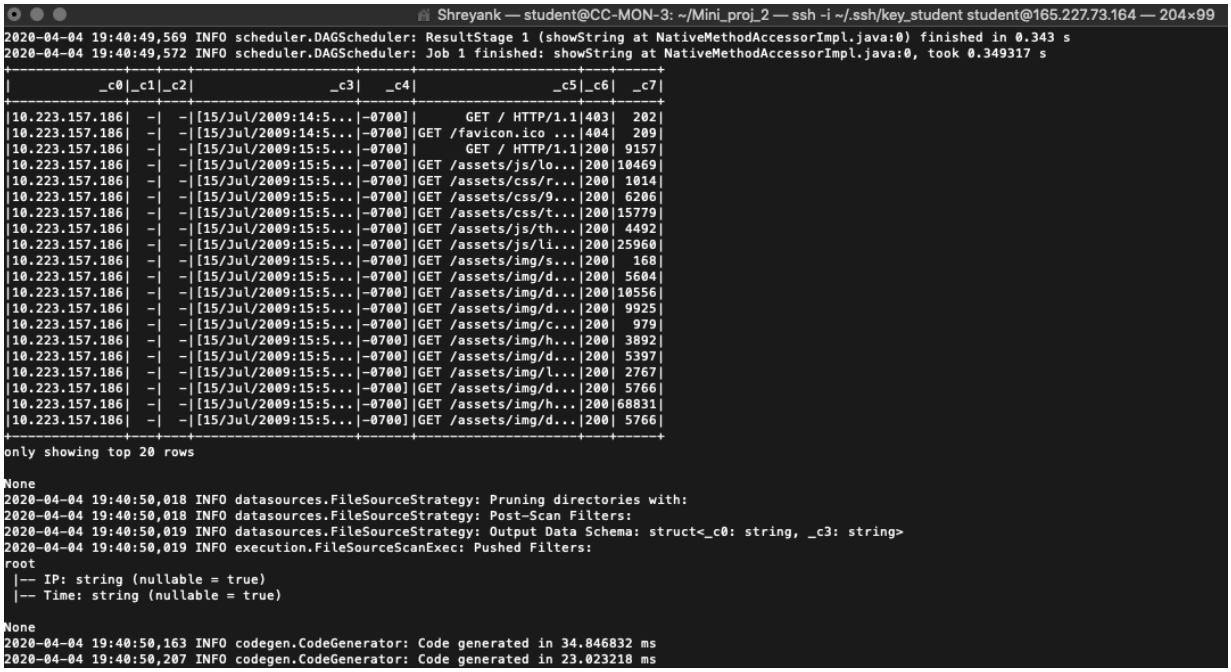
Part 3:
Snapshots of data at various stages of transformations

Spark - yarn - diagnostic message



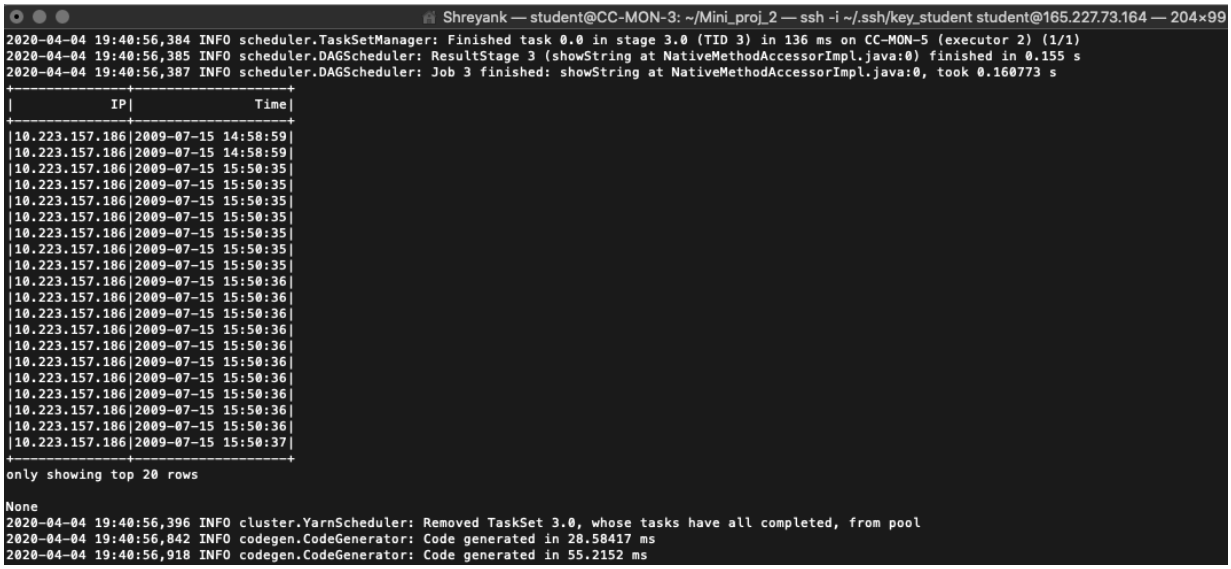
Data transformation outputs:

Read data frame from HDFS

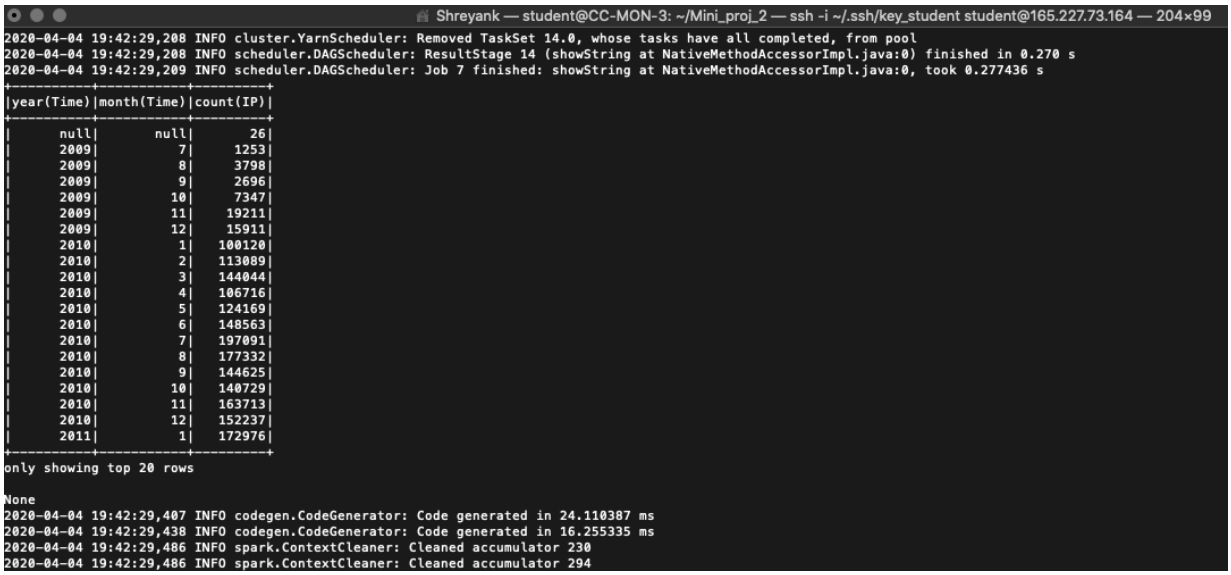


Select only columns 0 and 3

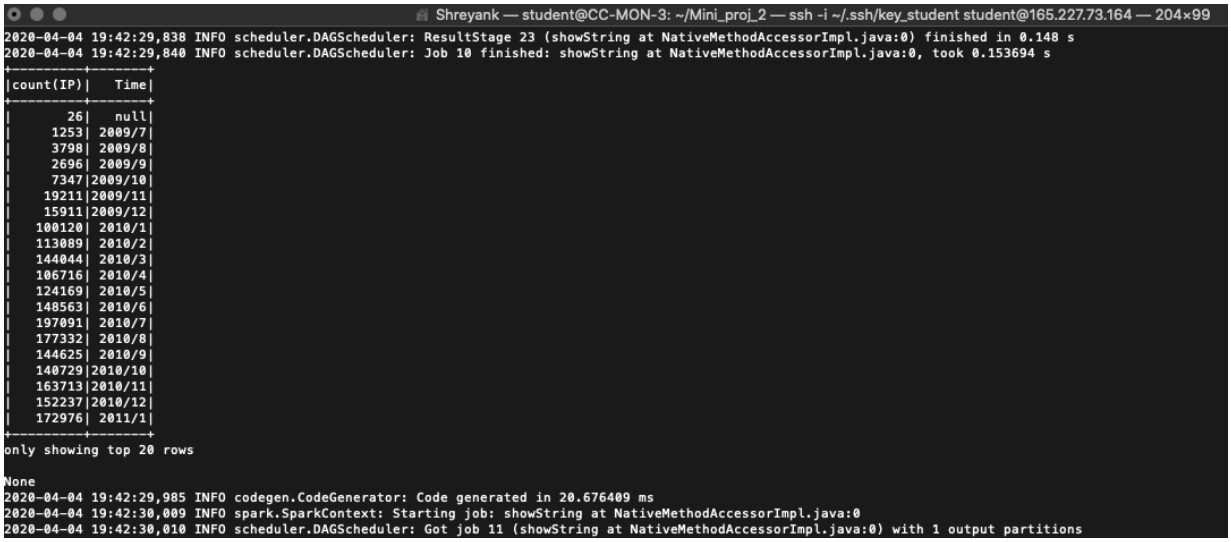
Remove '[' from Time column and convert times to datetime format



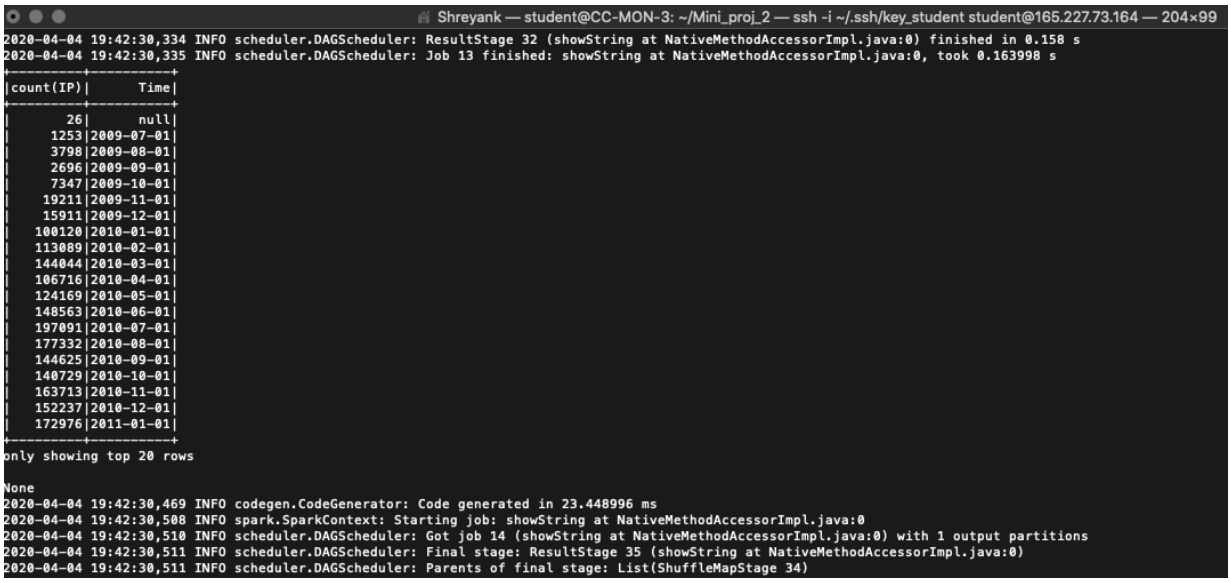
Get counts of IPs and Group by month of year.



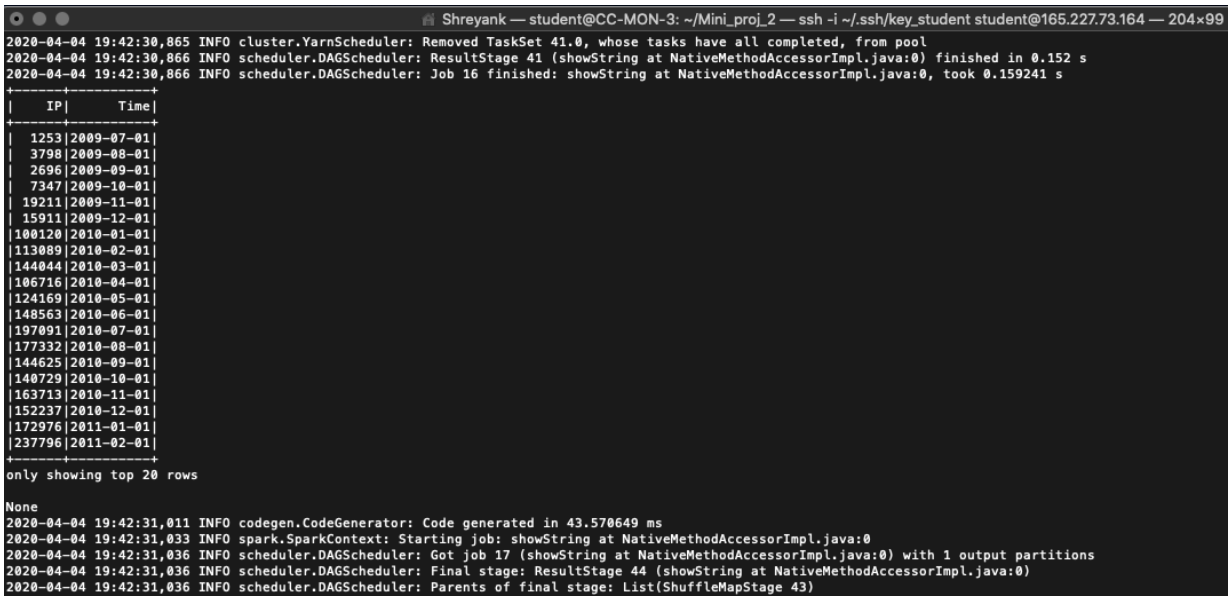
Concatenate month and year columns



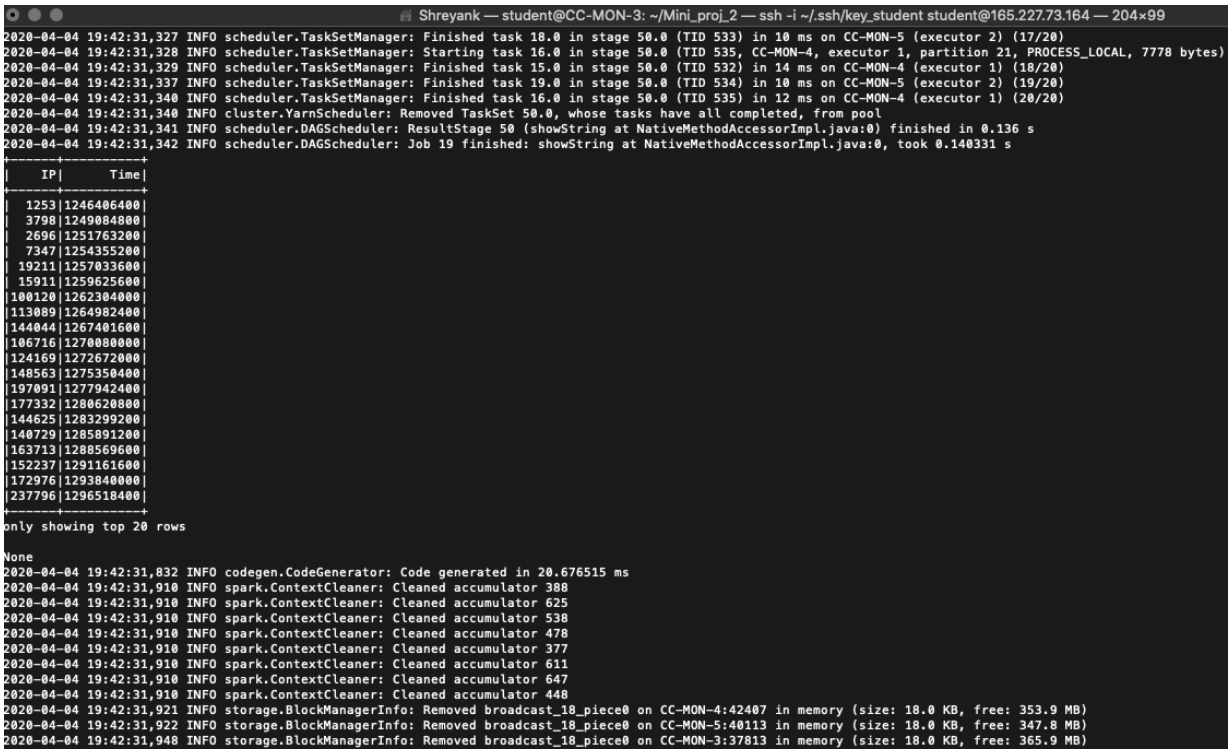
Convert Time column to timestamp format



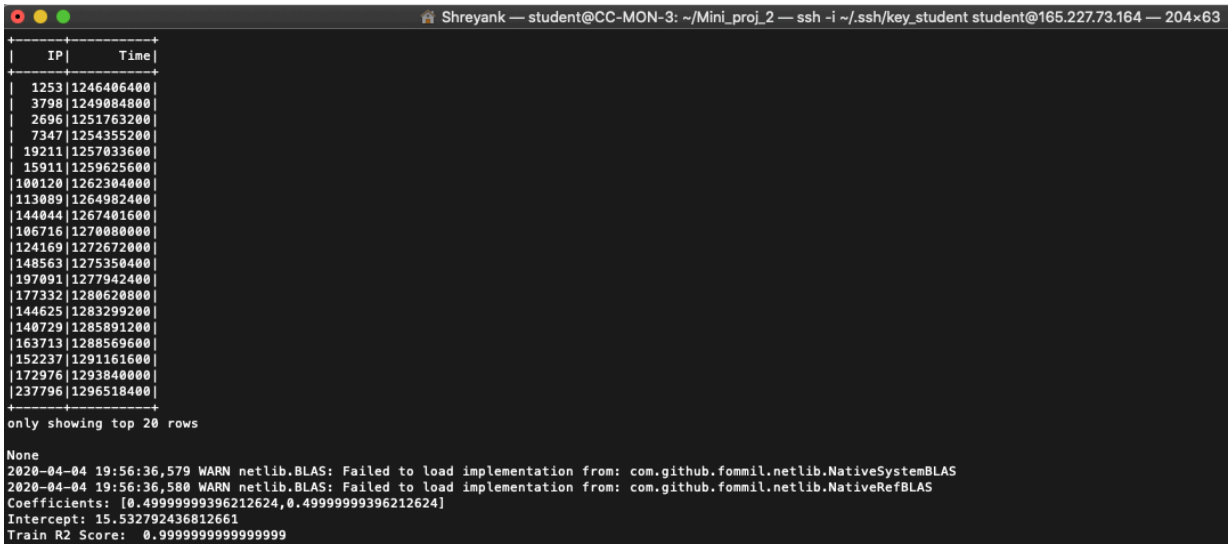
Remove null values



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



Linear Regression model coefficients, intercept and Train R2 Score



spark.stop()

```
2020-04-04 19:44:26,518 INFO spark.SparkContext: Invoking stop() from shutdown hook
2020-04-04 19:44:26,529 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 19:44:26,559 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$
```