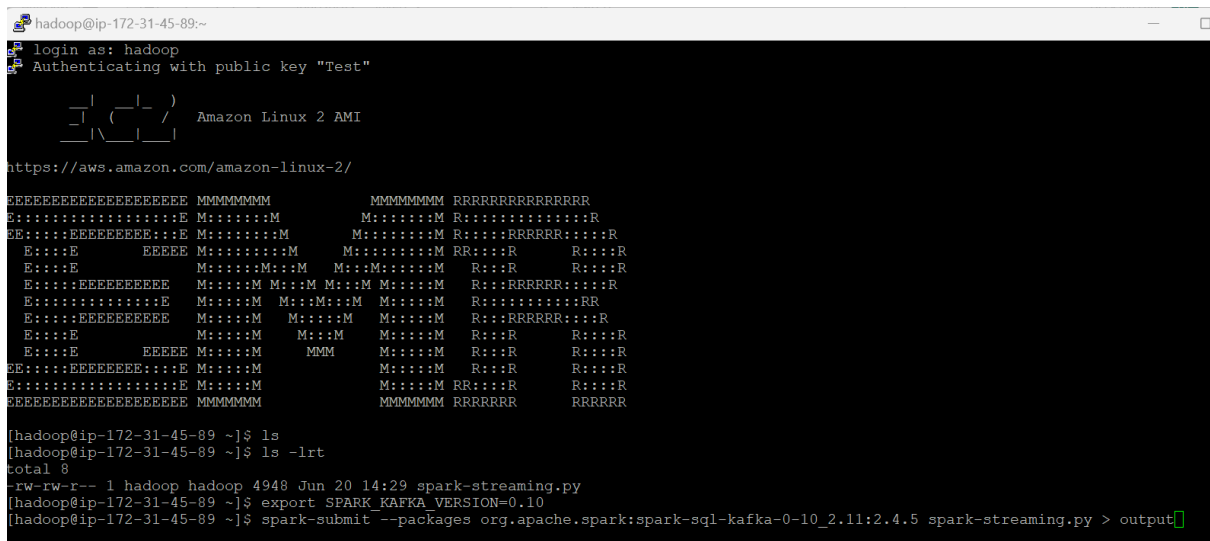


```
export SPARK_KAFKA_VERSION=0.10
```

```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark-streaming.py
> output
```



```

24/06/20 15:46:49 INFO hadoop@ip-172-31-45-89-
24/06/20 15:46:49 INFO YarnClient: Application report for application_1718898140041_0001 (state: ACCEPTED)
24/06/20 15:46:50 INFO Client: Application report for application_1718898140041_0001 (state: ACCEPTED)
24/06/20 15:46:51 INFO Client: Application report for application_1718898140041_0001 (state: ACCEPTED)
24/06/20 15:46:52 INFO Client: Application report for application_1718898140041_0001 (state: ACCEPTED)
24/06/20 15:46:53 INFO YarnClientSchedulerBackend: Add WebUI Filter. org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter, Map(PROXY_HOSTS -> ip-172-31-45-89.ec2.internal, PROXY_URI_BASES -> http://ip-172-31-45-89.ec2.internal:20888/proxy/application_1718898140041_0001), /proxy/application_1718898140041_0001
1
24/06/20 15:46:53 INFO Client: Application report for application_1718898140041_0001 (state: RUNNING)
24/06/20 15:46:53 INFO Client:
client token: N/A
diagnostics: N/A
ApplicationMaster host: 172.31.45.89
ApplicationMaster RPC port: -1
queue: default
start time: 1718898406259
final status: UNDEFINED
tracking URL: http://ip-172-31-45-89.ec2.internal:20888/proxy/application_1718898140041_0001/
user: hadoop
24/06/20 15:46:53 INFO YarnClientSchedulerBackend: Application application_1718898140041_0001 has started running.
24/06/20 15:46:53 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 36421.
24/06/20 15:46:53 INFO NettyBlockTransferService: Server created on ip-172-31-45-89.ec2.internal:36421
24/06/20 15:46:53 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
24/06/20 15:46:53 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, ip-172-31-45-89.ec2.internal, 36421, None)
24/06/20 15:46:53 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-45-89.ec2.internal:36421 with 1028.8 MB RAM, BlockManagerId(driver, ip-172-31-45-89.ec2.internal, 36421, None)
24/06/20 15:46:53 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, ip-172-31-45-89.ec2.internal, 36421, None)
24/06/20 15:46:53 INFO BlockManager: external shuffle service port = 7337
24/06/20 15:46:53 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, ip-172-31-45-89.ec2.internal, 36421, None)
24/06/20 15:46:53 INFO YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnRM)
24/06/20 15:46:53 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /metrics/json.
24/06/20 15:46:54 INFO EventLoggingListener: Logging events to hdfs://var/log/spark/apps/application_1718898140041_0001
24/06/20 15:46:54 INFO Utils: Using initial executors = 50, max of spark.dynamicAllocation.initialExecutors, spark.dynamicAllocation.minExecutors and spark.executor.instances
24/06/20 15:46:54 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
24/06/20 15:46:54 INFO SharedState: Loading hive config file: file:/etc/spark/conf/dist/hive-site.xml
24/06/20 15:46:55 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('hdfs://user/spark/warehouse').
24/06/20 15:46:55 INFO SharedState: Warehouse path is 'hdfs://user/spark/warehouse'.
24/06/20 15:46:55 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL.
24/06/20 15:46:55 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/json.
24/06/20 15:46:55 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution/json.
24/06/20 15:46:55 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution.json.
24/06/20 15:46:55 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /static/sql.
24/06/20 15:46:57 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint

```

```
hdfs dfs -ls
```

```
hadoop@ip-172-31-45-89:~$
login as: hadoop
Authenticating with public key "Test"
Last login: Thu Jun 20 15:47:48 2024

      _ _ _ _ _
     /   /   /   /
    /___/___/___/___/
   /___/___/___/___/
  /___/___/___/___/
 /___/___/___/___/
/___/___/___/___/

Amazon Linux 2 AMI

https://aws.amazon.com/amazon-linux-2/

EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M M::::::::M R:::::::::R
EE::::::::EEEEEEEE::E M::::::::M M::::::::M R::::RRRRR::::R
E::::E EEEEE M::::::::M M::::::::M RR::::R R::::R
E::::E M::::::::M M::::::::M M::::::::M R::::R R::::R
E::::EEEEEEEE M::::M M::M M::M M::M R::RRRRR::::R
E::::::::::::E M::::M M::M M::M M::M R::::::::RR
E::::EEEEEEEE M::::M M::M M::M M::M R::RRRRR::::R
E::::E M::::M M::M M::M M::M R::R R::R R::R
E::::E EEEEE M::::M M::M M::M R::R R::R R::R
EE::::::::EEEEEEEE::E M::::M M::M M::M R::R R::R
E::::::::::::E M::::M M::M M::M RR::::R R::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM MMMMMMM RRRRRRR RRRRRR

[hadoop@ip-172-31-45-89 ~]$ ls -lrt
total 16
-rw-rw-r-- 1 hadoop hadoop 4948 Jun 20 14:29 spark-streaming.py
-rw-rw-r-- 1 hadoop hadoop 4101 Jun 20 15:49 output
[hadoop@ip-172-31-45-89 ~]$
[hadoop@ip-172-31-45-89 ~]$ hdfs dfs -ls
Found 5 items
drwxr-xr-x - hadoop hadoop 0 2024-06-20 15:46 .sparkStaging
drwxr-xr-x - hadoop hadoop 0 2024-06-20 15:50 timeKPIvalue
drwxr-xr-x - hadoop hadoop 0 2024-06-20 15:47 timeKPIvalue_checkpoints
drwxr-xr-x - hadoop hadoop 0 2024-06-20 15:50 time_countryKPIvalue
drwxr-xr-x - hadoop hadoop 0 2024-06-20 15:47 time_countryKPIvalue_checkpoints
[hadoop@ip-172-31-45-89 ~]$
```

hdfs dfs -ls timeKPIvalue/

```
[hadoop@ip-172-31-45-89 ~]$
[hadoop@ip-172-31-45-89 ~]$
[hadoop@ip-172-31-45-89 ~]$ hdfs dfs -ls timeKPIvalue/
Found 8 items
drwxr-xr-x - hadoop hadoop 0 2024-06-20 15:51 timeKPIvalue/ _spark_metadata
-rw-r--r-- 1 hadoop hadoop 0 2024-06-20 15:47 timeKPIvalue/part-00000-397b4e6c-0152-4ac1-97e8-5a4d138a2480-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2024-06-20 15:51 timeKPIvalue/part-00000-40a8deb4-042d-4fa2-8d06-d66cbfab4cf4-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2024-06-20 15:49 timeKPIvalue/part-00000-51f72934-ff2f-484d-a5e9-5aef9421d219-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2024-06-20 15:48 timeKPIvalue/part-00000-92f635b7-170d-41cc-b9ea-eb6422efc23c-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2024-06-20 15:50 timeKPIvalue/part-00000-953a2e43-3ff3-474f-a05e-6de6aaa05f8d-c000.json
-rw-r--r-- 1 hadoop hadoop 200 2024-06-20 15:50 timeKPIvalue/part-00054-2c1efceb-58a2-466e-abfd-0483f0b85c47-c000.json
-rw-r--r-- 1 hadoop hadoop 203 2024-06-20 15:51 timeKPIvalue/part-00127-0d205589-c7ee-481b-b844-6440ddd0253e-c000.json
```

hdfs dfs -ls time_countryKPIvalue/

```
[hadoop@ip-172-31-45-89 ~]$
[hadoop@ip-172-31-45-89 ~]$
[hadoop@ip-172-31-45-89 ~]$ hdfs dfs -ls time_countryKPIvalue/
Found 14 items
drwxr-xr-x - hadoop hadoop 0 2024-06-20 15:52 time_countryKPIvalue/ _spark_metadata
-rw-r--r-- 1 hadoop hadoop 0 2024-06-20 15:50 time_countryKPIvalue/part-00000-21c9bbcc-a0c0-4a62-b71f-8f3df5221957-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2024-06-20 15:51 time_countryKPIvalue/part-00000-32c2c091-430c-48f8-b50f-e6dc20efe082-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2024-06-20 15:49 time_countryKPIvalue/part-00000-33564d53-15b3-4943-94d6-df32a4fd927c-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2024-06-20 15:47 time_countryKPIvalue/part-00000-519052f8-5a5f-4e0e-b330-aaea062dc373-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2024-06-20 15:48 time_countryKPIvalue/part-00000-5aa43399-0719-4abe-93e7-bf0e7aab04ea-c000.json
-rw-r--r-- 1 hadoop hadoop 0 2024-06-20 15:52 time_countryKPIvalue/part-00000-df092061-3083-4de1-913b-cd98fe1a708e-c000.json
-rw-r--r-- 1 hadoop hadoop 169 2024-06-20 15:52 time_countryKPIvalue/part-00013-7a0a2587-274e-4f95-924b-e5178e8044a1-c000.json
-rw-r--r-- 1 hadoop hadoop 160 2024-06-20 15:52 time_countryKPIvalue/part-00014-5e714d00-f5bf-498e-b240-080e694c50f1-c000.json
-rw-r--r-- 1 hadoop hadoop 158 2024-06-20 15:50 time_countryKPIvalue/part-00024-c50660fd-b61d-46dc-a5c8-20c41c7ed47b-c000.json
-rw-r--r-- 1 hadoop hadoop 159 2024-06-20 15:52 time_countryKPIvalue/part-00054-0045d5b1-4c81-481a-ba82-a3e8f265a46d-c000.json
-rw-r--r-- 1 hadoop hadoop 169 2024-06-20 15:50 time_countryKPIvalue/part-00097-33882eaa-2571-4a12-b631-8ac1961339a3-c000.json
-rw-r--r-- 1 hadoop hadoop 160 2024-06-20 15:51 time_countryKPIvalue/part-00103-df6d42f3-2110-4bef-9ba0-18cbef07649e-c000.json
-rw-r--r-- 1 hadoop hadoop 186 2024-06-20 15:51 time_countryKPIvalue/part-00192-9c553172-253b-4e0f-bb42-7abfa2abce81-c000.json
```

hdfs dfs -get timeKPIvalue/* /home/hadoop/timeKPIvalue/

hdfs dfs -get time_countryKPIvalue/* /home/hadoop/time_countryKPIvalue/

```
[hadoop@ip-172-31-45-89 ~]$
[hadoop@ip-172-31-45-89 ~]$
[hadoop@ip-172-31-45-89 ~]$
[hadoop@ip-172-31-45-89 ~]$
[hadoop@ip-172-31-45-89 ~]$ hdfs dfs -get timeKPIvalue/* /home/hadoop/timeKPIvalue/
[hadoop@ip-172-31-45-89 ~]$ hdfs dfs -get time_countryKPIvalue/* /home/hadoop/time_countryKPIvalue/
[hadoop@ip-172-31-45-89 ~]$
[hadoop@ip-172-31-45-89 ~]$
[hadoop@ip-172-31-45-89 ~]$
```