interative output of etl process is in

\profiling_code\zack_ly1339\review_pricechange\interative_shell_mode.md

With some explaination.

My NetId: ly1339

ETL Job files:

1. step0_1_peel.py
2. step2_peel.py

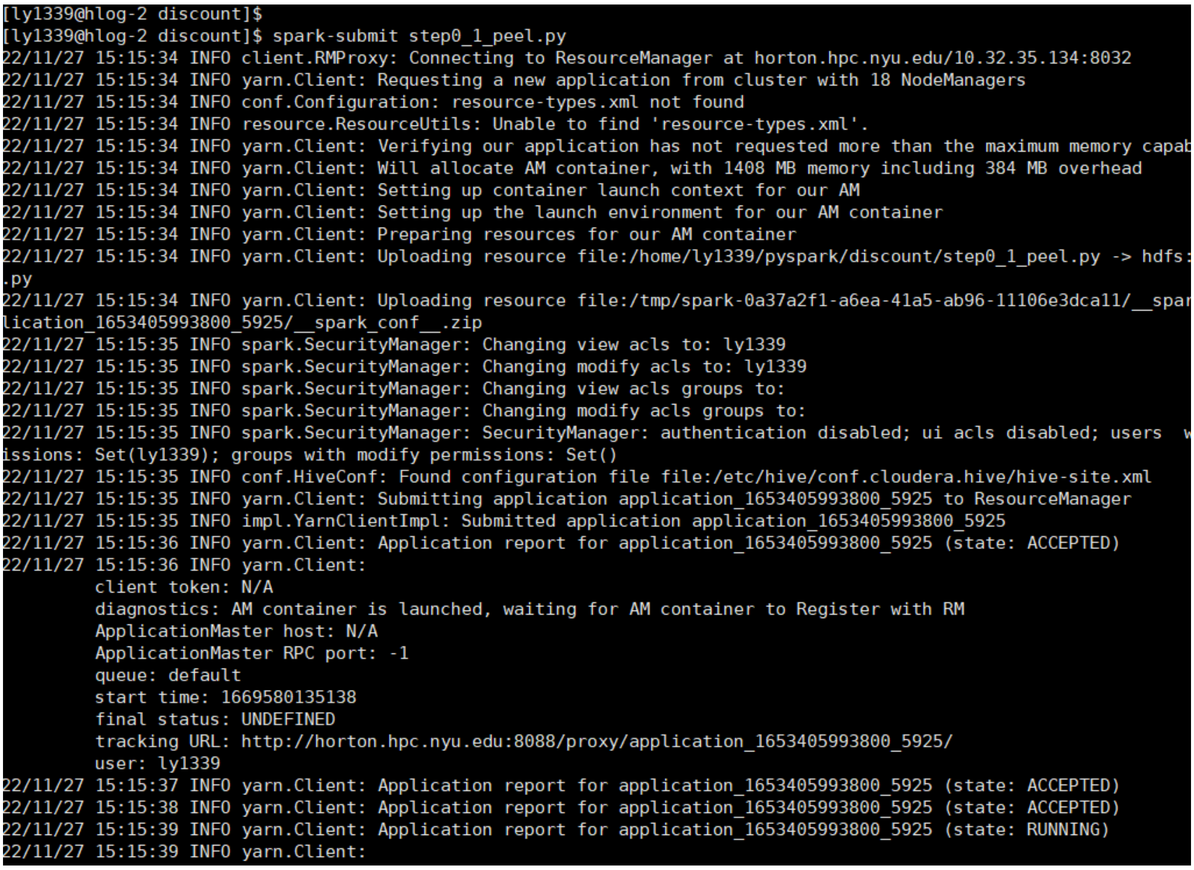
Run the job by:

```
spark-submit xxx.py
```

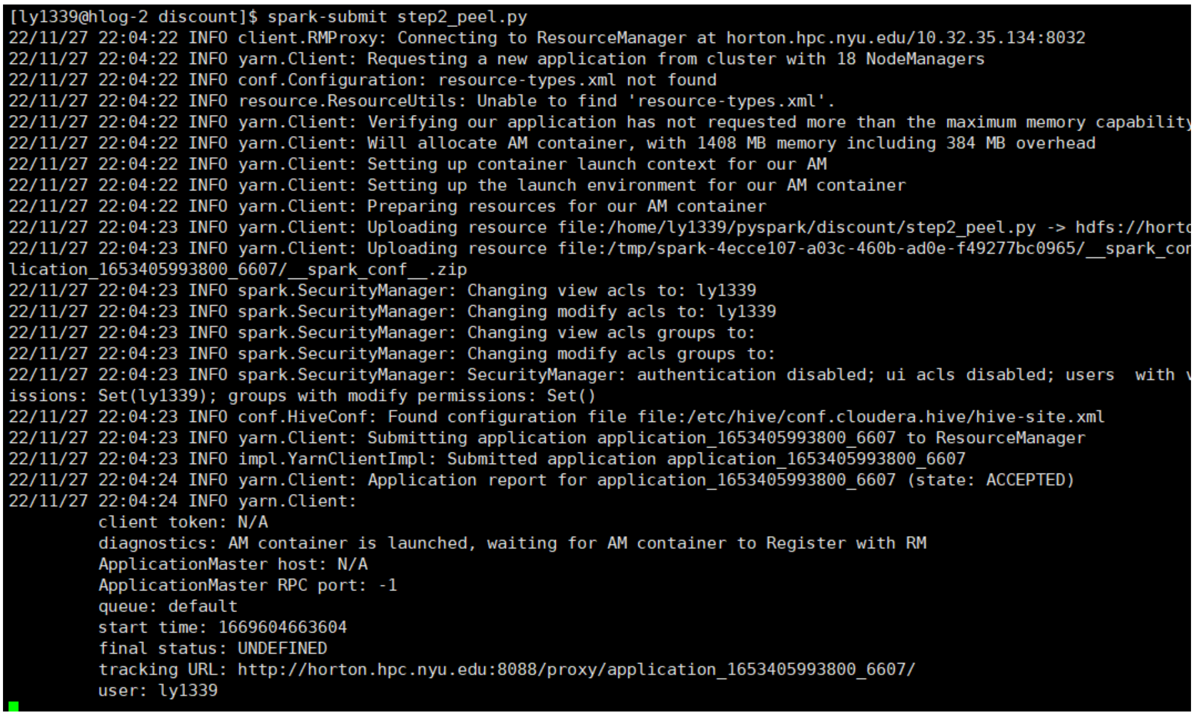
You should be able to specify my hdfs folder by changeing

```
hdfs_path = "./project/discount/"
```

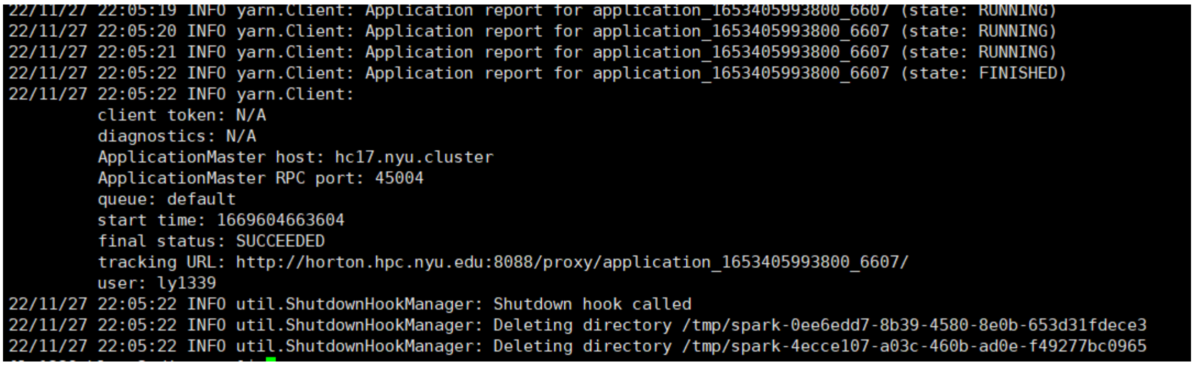
line from code.



```
[ly1339@hlog-2 discount]$
[ly1339@hlog-2 discount]$ spark-submit step0_1_peel.py
22/11/27 15:15:34 INFO client.RMProxy: Connecting to ResourceManager at horton.hpc.nyu.edu/10.32.35.134:8032
22/11/27 15:15:34 INFO yarn.Client: Requesting a new application from cluster with 18 NodeManagers
22/11/27 15:15:34 INFO conf.Configuration: resource-types.xml not found
22/11/27 15:15:34 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/11/27 15:15:34 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capab
22/11/27 15:15:34 INFO yarn.Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead
22/11/27 15:15:34 INFO yarn.Client: Setting up container launch context for our AM
22/11/27 15:15:34 INFO yarn.Client: Setting up the launch environment for our AM container
22/11/27 15:15:34 INFO yarn.Client: Preparing resources for our AM container
22/11/27 15:15:34 INFO yarn.Client: Uploading resource file:/home/ly1339/pyspark/discount/step0_1_peel.py -> hdfs:
.py
22/11/27 15:15:34 INFO yarn.Client: Uploading resource file:/tmp/spark-0a37a2f1-a6ea-41a5-ab96-11106e3dca11/__spar
lication_1653405993800_5925/__spark_conf__.zip
22/11/27 15:15:35 INFO spark.SecurityManager: Changing view acls to: ly1339
22/11/27 15:15:35 INFO spark.SecurityManager: Changing modify acls to: ly1339
22/11/27 15:15:35 INFO spark.SecurityManager: Changing view acls groups to:
22/11/27 15:15:35 INFO spark.SecurityManager: Changing modify acls groups to:
22/11/27 15:15:35 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  w
issions: Set(ly1339); groups with modify permissions: Set()
22/11/27 15:15:35 INFO conf.HiveConf: Found configuration file file:/etc/hive/conf.cloudera.hive/hive-site.xml
22/11/27 15:15:35 INFO yarn.Client: Submitting application application_1653405993800_5925 to ResourceManager
22/11/27 15:15:35 INFO impl.YarnClientImpl: Submitted application application_1653405993800_5925
22/11/27 15:15:36 INFO yarn.Client: Application report for application_1653405993800_5925 (state: ACCEPTED)
22/11/27 15:15:36 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: 1669580135138
         final status: UNDEFINED
         tracking URL: http://horton.hpc.nyu.edu:8088/proxy/application_1653405993800_5925/
         user: ly1339
22/11/27 15:15:37 INFO yarn.Client: Application report for application_1653405993800_5925 (state: ACCEPTED)
22/11/27 15:15:38 INFO yarn.Client: Application report for application_1653405993800_5925 (state: ACCEPTED)
22/11/27 15:15:39 INFO yarn.Client: Application report for application_1653405993800_5925 (state: RUNNING)
22/11/27 15:15:39 INFO yarn.Client:
```



```
[ly1339@hlog-2 discount]$ spark-submit step2_peel.py
22/11/27 22:04:22 INFO client.RMProxy: Connecting to ResourceManager at horton.hpc.nyu.edu/10.32.35.134:8032
22/11/27 22:04:22 INFO yarn.Client: Requesting a new application from cluster with 18 NodeManagers
22/11/27 22:04:22 INFO conf.Configuration: resource-types.xml not found
22/11/27 22:04:22 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
22/11/27 22:04:22 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability
22/11/27 22:04:22 INFO yarn.Client: Will allocate AM container, with 1408 MB memory including 384 MB overhead
22/11/27 22:04:22 INFO yarn.Client: Setting up container launch context for our AM
22/11/27 22:04:22 INFO yarn.Client: Setting up the launch environment for our AM container
22/11/27 22:04:22 INFO yarn.Client: Preparing resources for our AM container
22/11/27 22:04:23 INFO yarn.Client: Uploading resource file:/home/ly1339/pyspark/discount/step2_peel.py -> hdfs://hortc
22/11/27 22:04:23 INFO yarn.Client: Uploading resource file:/tmp/spark-4ecce107-a03c-460b-ad0e-f49277bc0965/__spark_cor
lication_1653405993800_6607/__spark_conf__.zip
22/11/27 22:04:23 INFO spark.SecurityManager: Changing view acls to: ly1339
22/11/27 22:04:23 INFO spark.SecurityManager: Changing modify acls to: ly1339
22/11/27 22:04:23 INFO spark.SecurityManager: Changing view acls groups to:
22/11/27 22:04:23 INFO spark.SecurityManager: Changing modify acls groups to:
22/11/27 22:04:23 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with v
issions: Set(ly1339); groups with modify permissions: Set()
22/11/27 22:04:23 INFO conf.HiveConf: Found configuration file file:/etc/hive/conf.cloudera.hive/hive-site.xml
22/11/27 22:04:23 INFO yarn.Client: Submitting application application_1653405993800_6607 to ResourceManager
22/11/27 22:04:23 INFO impl.YarnClientImpl: Submitted application application_1653405993800_6607
22/11/27 22:04:24 INFO yarn.Client: Application report for application_1653405993800_6607 (state: ACCEPTED)
22/11/27 22:04:24 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: 1669604663604
         final status: UNDEFINED
         tracking URL: http://horton.hpc.nyu.edu:8088/proxy/application_1653405993800_6607/
         user: ly1339
```



```
22/11/27 22:05:19 INFO yarn.Client: Application report for application_1653405993800_6607 (state: RUNNING)
22/11/27 22:05:20 INFO yarn.Client: Application report for application_1653405993800_6607 (state: RUNNING)
22/11/27 22:05:21 INFO yarn.Client: Application report for application_1653405993800_6607 (state: RUNNING)
22/11/27 22:05:22 INFO yarn.Client: Application report for application_1653405993800_6607 (state: FINISHED)
22/11/27 22:05:22 INFO yarn.Client:
         client token: N/A
         diagnostics: N/A
         ApplicationMaster host: hc17.nyu.cluster
         ApplicationMaster RPC port: 45004
         queue: default
         start time: 1669604663604
         final status: SUCCEEDED
         tracking URL: http://horton.hpc.nyu.edu:8088/proxy/application_1653405993800_6607/
         user: ly1339
22/11/27 22:05:22 INFO util.ShutdownHookManager: Shutdown hook called
22/11/27 22:05:22 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-0ee6edd7-8b39-4580-8e0b-653d31fdece3
22/11/27 22:05:22 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-4ecce107-a03c-460b-ad0e-f49277bc0965
```