

```
\ \ :\ \ : / \ \ :\ \ / : / \ \ :\ \ \ \ \ \ :\ \ / : / \ \ :\ \ / : /  
\\ \ :\ :\ / \ \ :\ :\ / \ \ :\ :\ / \ \ :\ :\ / \ \ :\ :\ /  
\\_ \ / \ \ :\ :\ / \ \ :\ :\ / \ \ :\ :\ /  
\\_ \ / \ \ :\ :\ / \ \ :\ :\ /  
  
*****  
https://wikis.nyu.edu/display/NYUHPC/Clusters+-+Dumbo  
*****  
[[sk8325@login-2-1 ~]$ hdfs dfs -mkdir project  
[[sk8325@login-2-1 ~]$ hdfs dfs -put /scratch/sk8325/  
crime.csv pluto.csv subway.csv  
[[sk8325@login-2-1 ~]$ hdfs dfs -put /scratch/sk8325/crime.csv project  
[[sk8325@login-2-1 ~]$ hdfs dfs -put /scratch/sk8325/subway.csv project  
[[sk8325@login-2-1 ~]$ hdfs dfs -put /scratch/sk8325/pluto.csv project  
[[sk8325@login-2-1 ~]$ hdfs dfs -ls project/  
Found 3 items  
-rw-r--r--+ 3 sk8325 users 2142626367 2020-04-10 16:07 project/crime.csv  
-rw-r--r--+ 3 sk8325 users 296575240 2020-04-10 16:07 project/pluto.csv  
-rw-r--r--+ 3 sk8325 users 64187 2020-04-10 16:07 project/subway.csv  
[[sk8325@login-2-1 ~]$  
[[sk8325@login-2-1 ~]$  
[[sk8325@login-2-1 ~]$ hdfs dfs -setfacl -R -m default:user:rss638:rwx project  
[[sk8325@login-2-1 ~]$ hdfs dfs -setfacl -R -m user:rss638:rwx project  
[[sk8325@login-2-1 ~]$ hdfs dfs -getfacl project  
# file: project  
# owner: sk8325  
# group: users  
user::rwx  
user:hive:rwx  
user:impala:rwx  
user:rss638:rwx  
group::r-x  
mask::rwx  
other::r-x  
default:user::rwx  
default:user:hive:rwx  
default:user:impala:rwx  
default:user:rss638:rwx  
default:group::r-x  
default:mask::rwx  
default:other::r-x  
  
[[sk8325@login-2-1 ~]$ spark2-submit --class App project/PropertyValueAnalytic-assembly-1.0.0.jar project
```

```
ApplicationMaster host: 10.0.255.236
ApplicationMaster RPC port: 0
queue: root.users.sk8325
start time: 1586550509349
final status: UNDEFINED
tracking URL: http://babar.es.its.nyu.edu:8088/proxy/application_1579315286937_40634/
user: sk8325
20/04/10 16:28:32 INFO cluster.YarnClientSchedulerBackend: Application application_1579315286937_40634 has started running.
20/04/10 16:28:32 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 44801.
20/04/10 16:28:32 INFO netty.NettyBlockTransferService: Server created on login-2-1.local:44801
20/04/10 16:28:32 INFO storage.BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
20/04/10 16:28:32 INFO storage.BlockManagerMaster: Registering BlockManager BlockManagerId(driver, login-2-1.local, 44801, None)
20/04/10 16:28:32 INFO storage.BlockManagerMasterEndpoint: Registering block manager login-2-1.local:44801 with 366.3 MB RAM, BlockManagerId(driver, login-2-1.local, 44801, None)
20/04/10 16:28:32 INFO storage.BlockManagerMaster: Registered BlockManager BlockManagerId(driver, login-2-1.local, 44801, None)
20/04/10 16:28:32 INFO storage.BlockManager: external shuffle service port = 7337
20/04/10 16:28:32 INFO storage.BlockManager: Initialized BlockManager: BlockManagerId(driver, login-2-1.local, 44801, None)
20/04/10 16:28:32 INFO cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnAM)
20/04/10 16:28:32 INFO ui.JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /metrics/json.
20/04/10 16:28:32 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@747f6c5a{/metrics/json,null,AVAILABLE,@Spark}
20/04/10 16:28:33 INFO scheduler.EventLoggingListener: Logging events to hdfs://dumbo/user/spark/spark2ApplicationHistory/application_1579315286937_40634
20/04/10 16:28:33 INFO spark.SparkContext: Registered listener com.cloudera.spark.lineage.NavigatorAppListener
20/04/10 16:28:35 INFO cluster.YarnSchedulerBackend$YarnDriverEndpoint: Registered executor NettyRpcEndpointRef(spark-client://Executor) (10.0.255.236:52864) with ID 2
20/04/10 16:28:35 INFO storage.BlockManagerMasterEndpoint: Registering block manager compute-2-12.local:43640 with 366.3 MB RAM, BlockManagerId(2, compute-2-12.local, 43640, None)
20/04/10 16:28:35 INFO cluster.YarnSchedulerBackend$YarnDriverEndpoint: Registered executor NettyRpcEndpointRef(spark-client://Executor) (10.0.255.239:59810) with ID 1
20/04/10 16:28:35 INFO storage.BlockManagerMasterEndpoint: Registering block manager compute-1-3.local:44142 with 366.3 MB RAM, BlockManagerId(1, compute-1-3.local, 44142, None)
20/04/10 16:28:35 INFO cluster.YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.8
Crime Cleaning took: 328092 msecs
Subway Cleaning took: 1296 msecs
Pluto Cleaning took: 188765 msecs
Crime Profiling took: 252888 msecs
Subway Profiling took: 8269 msecs
Pluto Profiling took: 44886 msecs
Total Application time: 835538 msecs
Cleaning and Profiling Done!
[sk8325@login-2-1 ~]$
```

```
[sk8325@login-2-1 ~]$ hdfs dfs -ls project/output
Found 2 items
drwxr-xr-x+ - sk8325 users          0 2020-04-10 16:34 project/output/cleaned
drwxr-xr-x+ - sk8325 users          0 2020-04-10 16:41 project/output/profile
[[sk8325@login-2-1 ~]$ hdfs dfs -ls project/output/cleaned
Found 3 items
drwxr-xr-x+ - sk8325 users          0 2020-04-10 16:34 project/output/cleaned/crime
drwxr-xr-x+ - sk8325 users          0 2020-04-10 16:37 project/output/cleaned/pluto
drwxr-xr-x+ - sk8325 users          0 2020-04-10 16:34 project/output/cleaned/subway
[[sk8325@login-2-1 ~]$ hdfs dfs -ls project/output/cleaned/crime
Found 17 items
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:34 project/output/cleaned/crime/_SUCCESS
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:30 project/output/cleaned/crime/part-00000
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:30 project/output/cleaned/crime/part-00001
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:31 project/output/cleaned/crime/part-00002
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:31 project/output/cleaned/crime/part-00003
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:30 project/output/cleaned/crime/part-00004
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:32 project/output/cleaned/crime/part-00005
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:32 project/output/cleaned/crime/part-00006
-rw-r--r--+ 3 sk8325 users      52175511 2020-04-10 16:29 project/output/cleaned/crime/part-00007
-rw-r--r--+ 3 sk8325 users      17073084 2020-04-10 16:29 project/output/cleaned/crime/part-00008
-rw-r--r--+ 3 sk8325 users      32377552 2020-04-10 16:32 project/output/cleaned/crime/part-00009
-rw-r--r--+ 3 sk8325 users      32294217 2020-04-10 16:33 project/output/cleaned/crime/part-00010
-rw-r--r--+ 3 sk8325 users      32320062 2020-04-10 16:30 project/output/cleaned/crime/part-00011
-rw-r--r--+ 3 sk8325 users      32328675 2020-04-10 16:33 project/output/cleaned/crime/part-00012
-rw-r--r--+ 3 sk8325 users      32328873 2020-04-10 16:33 project/output/cleaned/crime/part-00013
-rw-r--r--+ 3 sk8325 users      32324170 2020-04-10 16:31 project/output/cleaned/crime/part-00014
-rw-r--r--+ 3 sk8325 users      31211639 2020-04-10 16:33 project/output/cleaned/crime/part-00015
[[sk8325@login-2-1 ~]$ hdfs dfs -ls project/output/cleaned/subway
Found 3 items
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:34 project/output/cleaned/subway/_SUCCESS
-rw-r--r--+ 3 sk8325 users      13772 2020-04-10 16:34 project/output/cleaned/subway/part-00000
-rw-r--r--+ 3 sk8325 users      14372 2020-04-10 16:34 project/output/cleaned/subway/part-00001
[[sk8325@login-2-1 ~]$ hdfs dfs -ls project/output/cleaned/pluto
Found 4 items
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/cleaned/pluto/_SUCCESS
-rw-r--r--+ 3 sk8325 users      39569366 2020-04-10 16:36 project/output/cleaned/pluto/part-00000
-rw-r--r--+ 3 sk8325 users      39201505 2020-04-10 16:36 project/output/cleaned/pluto/part-00001
-rw-r--r--+ 3 sk8325 users      8144107 2020-04-10 16:37 project/output/cleaned/pluto/part-00002
[[sk8325@login-2-1 ~]$ hdfs dfs -ls project/output/profile
Found 3 items
drwxr-xr-x+ - sk8325 users          0 2020-04-10 16:41 project/output/profile/crime
drwxr-xr-x+ - sk8325 users          0 2020-04-10 16:42 project/output/profile/pluto
```

```
[sk8325@login-2-1 ~]$ hdfs dfs -ls project/output/profile/crime
Found 9 items
drwxr-xr-x+  - sk8325 users          0 2020-04-10 16:38 project/output/profile/crime/boroughs
drwxr-xr-x+  - sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/count
drwxr-xr-x+  - sk8325 users          0 2020-04-10 16:38 project/output/profile/crime/dates
drwxr-xr-x+  - sk8325 users          0 2020-04-10 16:39 project/output/profile/crime/levels
drwxr-xr-x+  - sk8325 users          0 2020-04-10 16:39 project/output/profile/crime/offense_descs
drwxr-xr-x+  - sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount
drwxr-xr-x+  - sk8325 users          0 2020-04-10 16:40 project/output/profile/crime/suspect_age
drwxr-xr-x+  - sk8325 users          0 2020-04-10 16:40 project/output/profile/crime/suspect_race
drwxr-xr-x+  - sk8325 users          0 2020-04-10 16:41 project/output/profile/crime/suspect_sex
[[sk8325@login-2-1 ~]$ hdfs dfs -ls project/output/profile/crime/originalCount
Found 17 items
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/_SUCCESS
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00000
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00001
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00002
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00003
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00004
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00005
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00006
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00007
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00008
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00009
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00010
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00011
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00012
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00013
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00014
-rw-r--r--+ 3 sk8325 users          16 2020-04-10 16:37 project/output/profile/crime/originalCount/part-00015
[[sk8325@login-2-1 ~]$ hdfs dfs -cat project/output/profile/crime/originalCount/part-00015
Count : 6847945
[[sk8325@login-2-1 ~]$ hdfs dfs -cat project/output/profile/crime/originalCount/part-00014
[[sk8325@login-2-1 ~]$ hdfs dfs -ls project/output/profile/crime/count
Found 17 items
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/count/_SUCCESS
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/count/part-00000
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/count/part-00001
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/count/part-00002
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/count/part-00003
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/count/part-00004
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/count/part-00005
-rw-r--r--+ 3 sk8325 users          0 2020-04-10 16:37 project/output/profile/crime/count/part-00006
```