**Clustering MLlib Implementation**

Step1. Launching session.

* Please look at Step 4 and Step 5 in the document “Setup\_Sandbox” to launch a session

Step2. Setup SPARK HOME

* Please setup the SPARK HOME variable (see Step 4 in the document “configuration\_for\_spark”)

Step3. Copy the code.

* Create a directory for Clustering code file.

**mkdir clustering**

* Copy code file “clustering.txt” from resource page to sandbox (Step 6 in Setup\_Sandbox)
* Convert txt to py

**mv clustering.txt clustering.py**

Step4. Input file.

* Create a Directory in HDFS for input file.

**hadoop fs -mkdir -p /cluster/input**

* Copy the input file named “stocks.txt” available on the resource page to sandbox (step 6 in Setup\_Sandbox) and upload it to HDFS

**hadoop fs -put stocks.txt /cluster/input**

Step5. Create a directory for output directories.

**hadoop fs -mkdir /cluster/output**

Step6.Executing code

* Navigate to “clustering” directory and run the below command.

**spark-submit clustring.py /cluster/input /cluster/output/low /cluster/output/high**

Step7. Checking output.

* Output will be printed on screen and also will be written into the HDFS file system creating two new directories under /cluster/output.
* Directory /cluster/output/low contains stock labels which are in low density cluster.
* Directory /cluster/output/high contains stock labels which are in high density cluster.
* Sample outputs:





