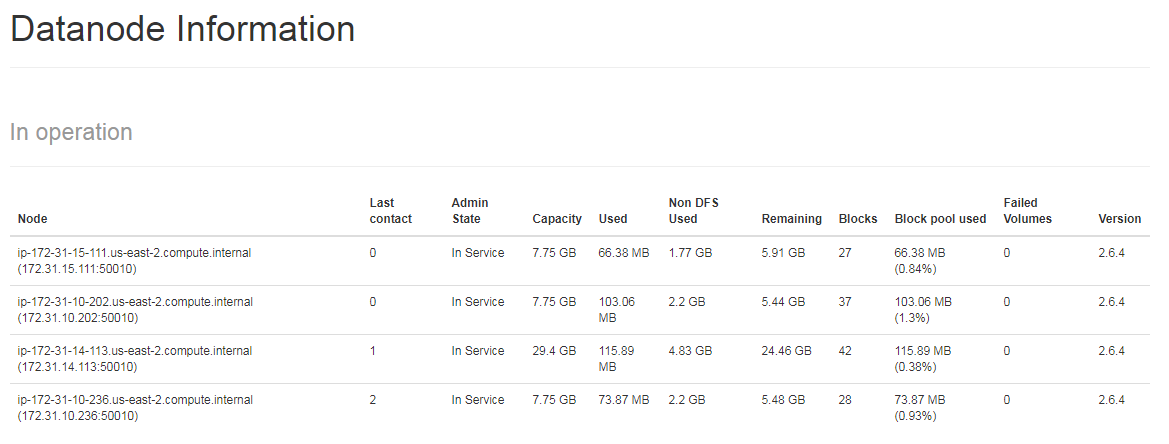
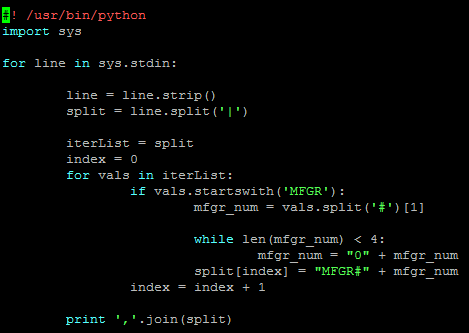
William Chirciu  
CSC 555 Project Phase 2

USED 4 NODE CLUSTER:



PART 1: DATA TRANSFORMATION

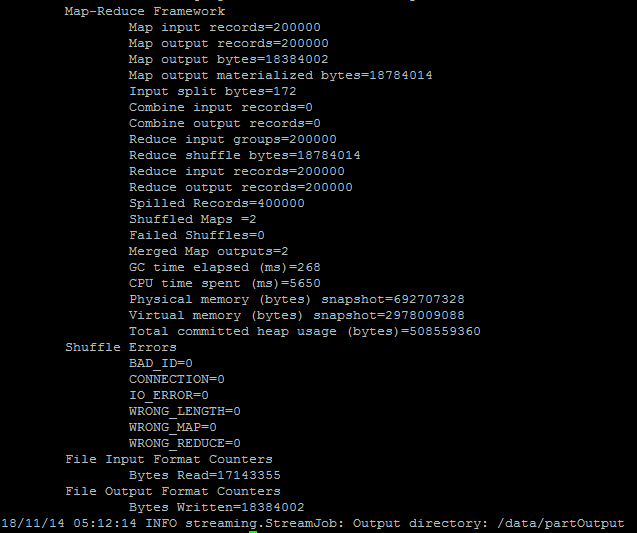
**Hadoop Streaming**

**Python Code:  
**

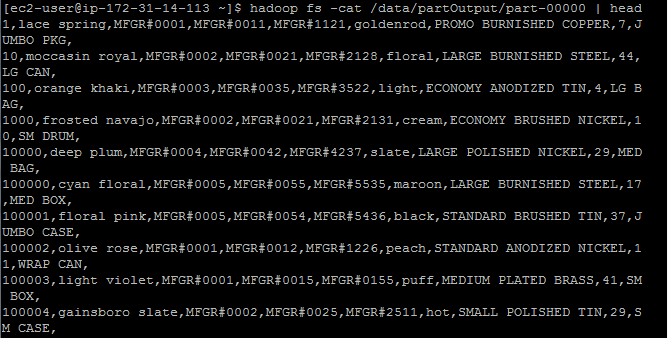
**Command:**

****

**After Running Job:**

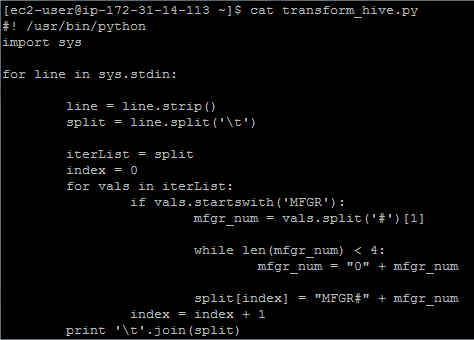
****

**Some Output:**

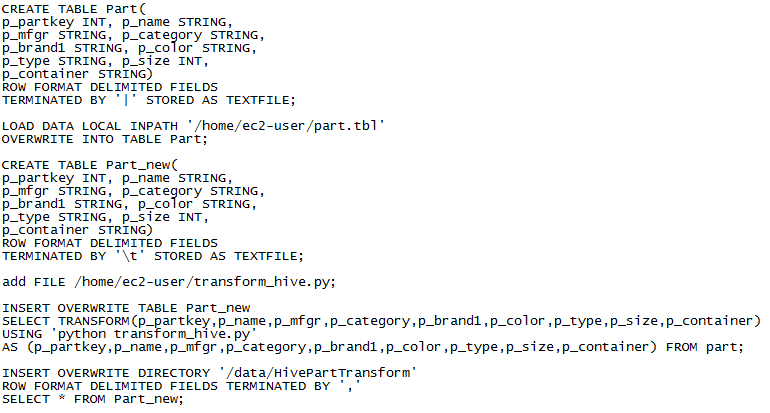
****

**HIVE**

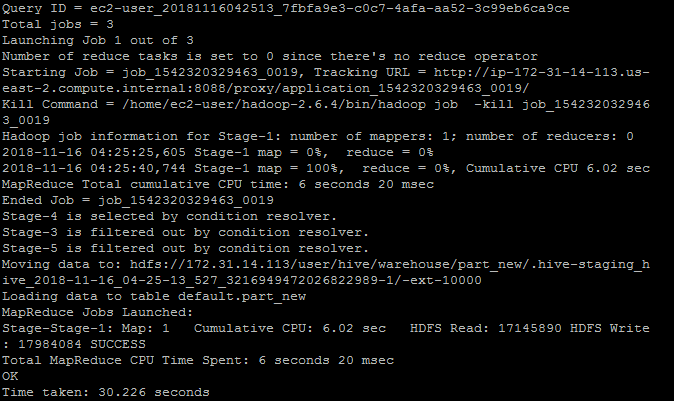
**Python Code:**

****

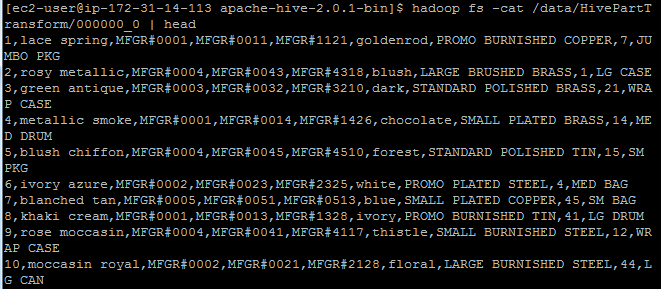
**Commands:**

****

**After running job:**

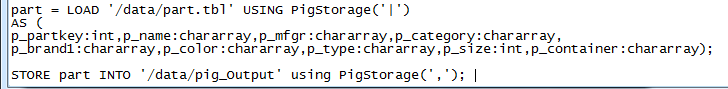
****

**Output:**

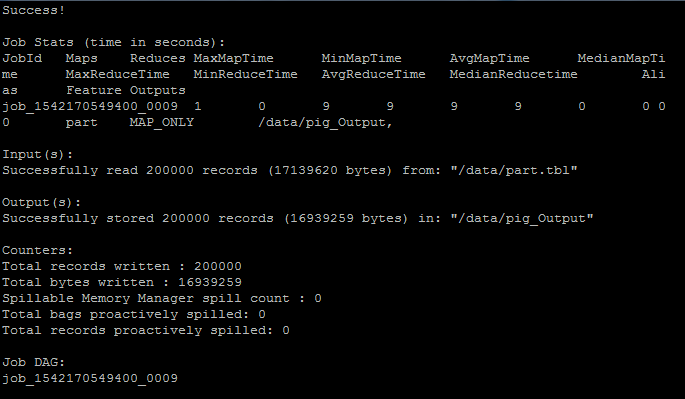
****

**PIG**

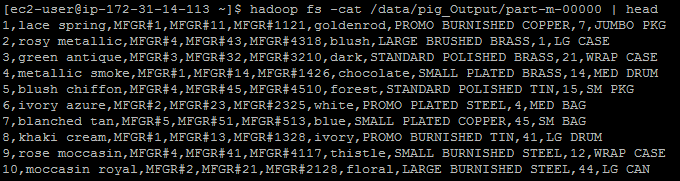
**Commands:**

****

**After running job:**

****

**Output:**

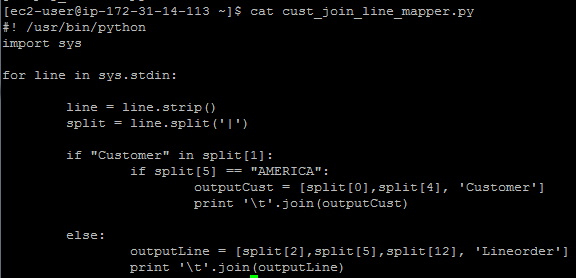
****

PART 2: Queries

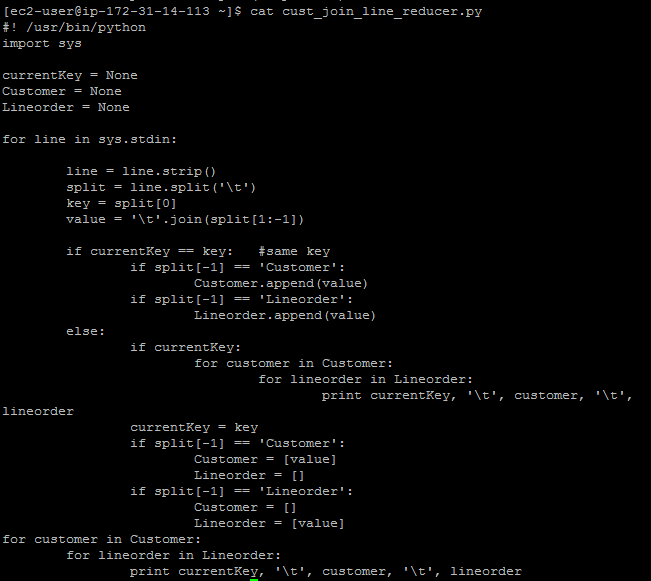
**Hadoop Streaming**

***Joining Customer and Lineorder Tables***

**Mapper:**

****

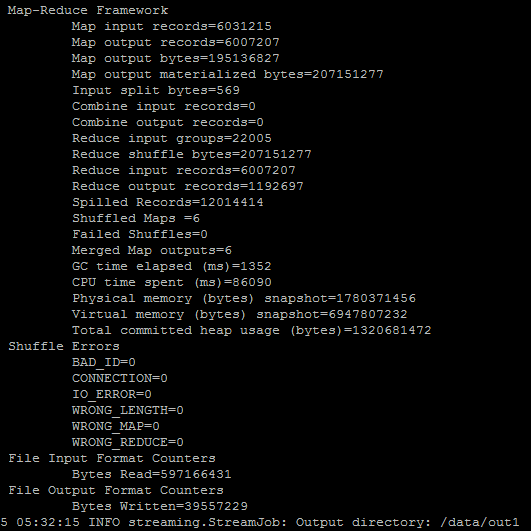
**Reducer:**

****

**Command Line:**

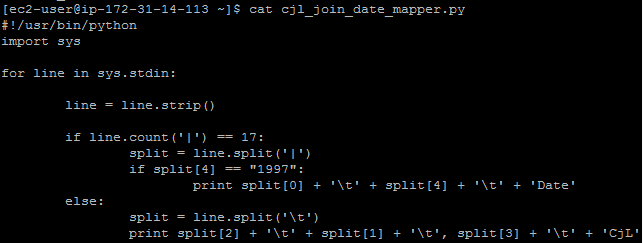
****

**After running job:**

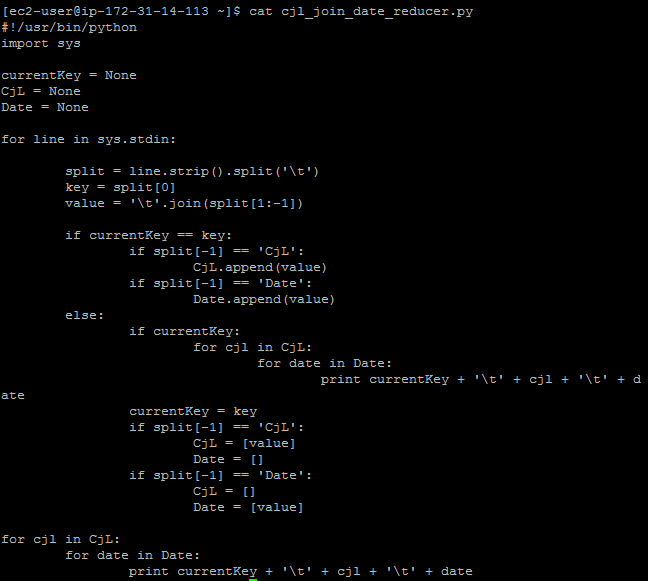
****

***Joining CjL and dwdate tables***

**Mapper:**

****

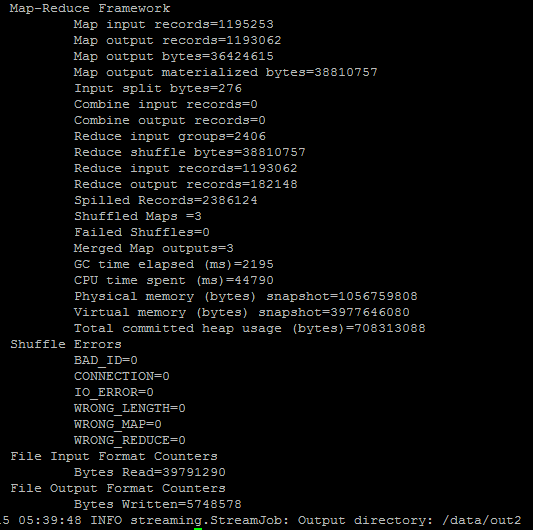
**Reducer:**

****

**Command Line:**

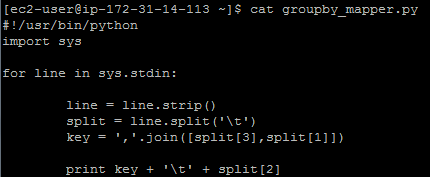
****

**After running job:**

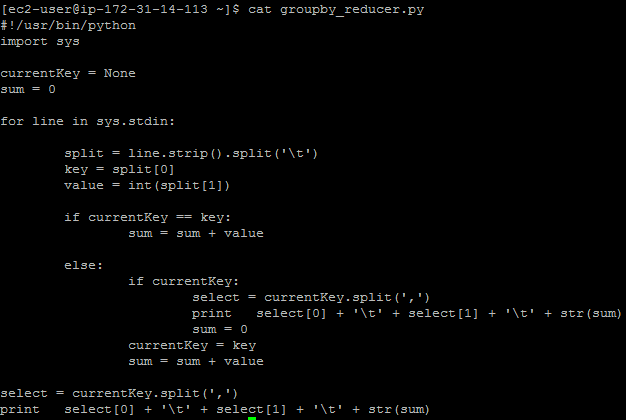
****

***Group By***

**Mapper:**

****

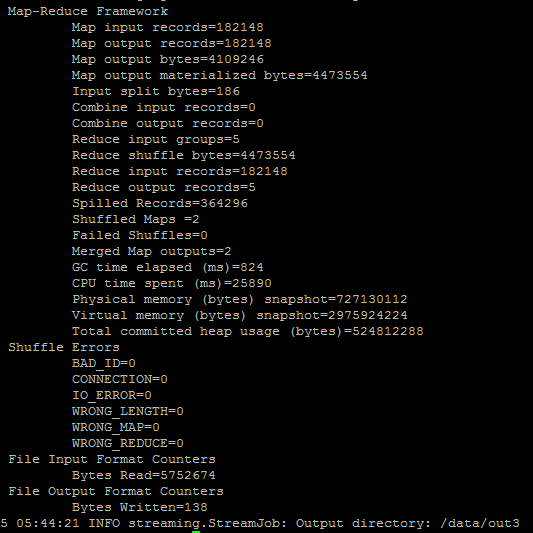
**Reducer:**

****

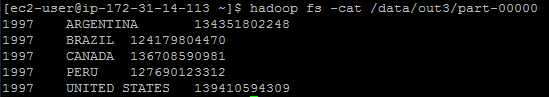
**Command Line:**

****

**After running job:**

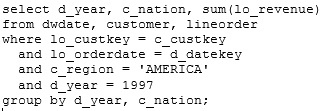
****

**Final Output:**

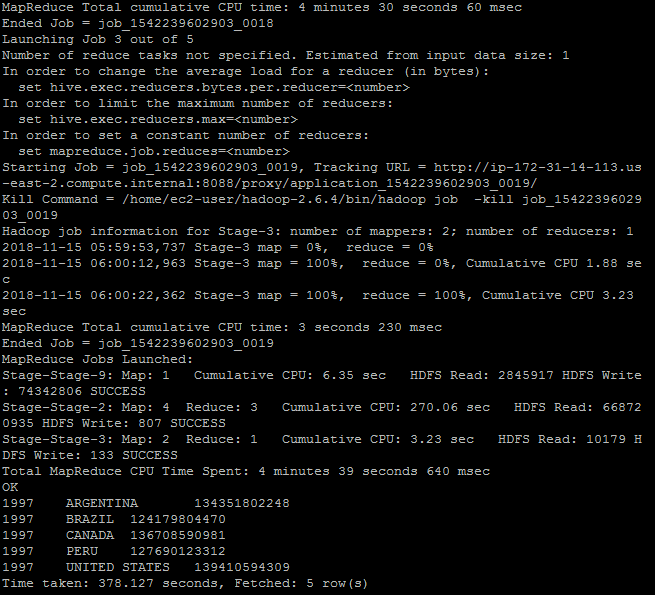
****

**HIVE**

**Command Line:**

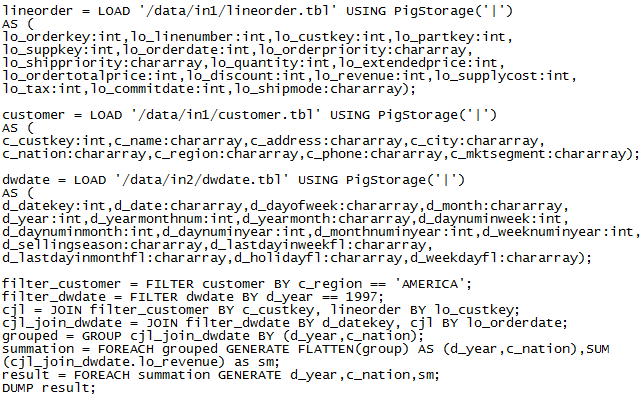
****

**After running job:**

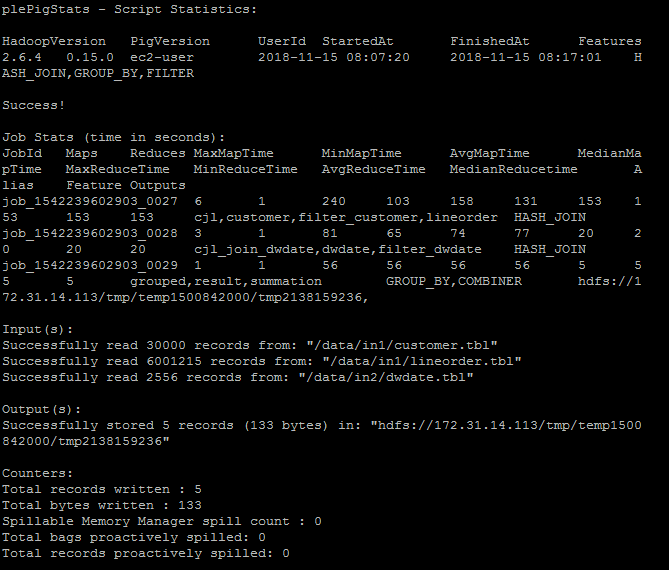
****

**PIG**

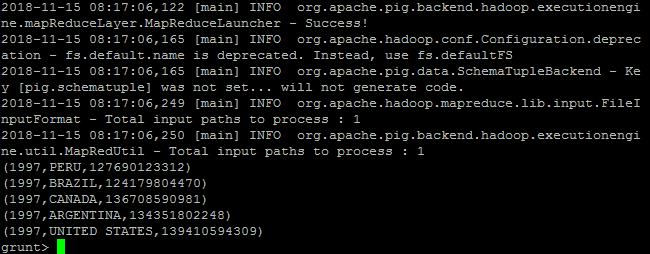
**Command Line:**

****

**After running job:**

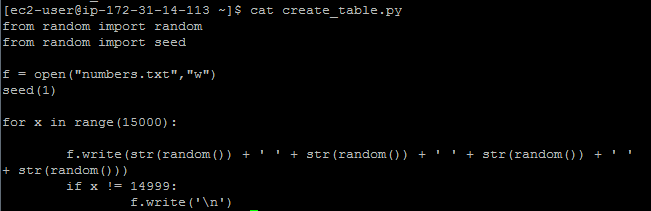
****

**Output:**

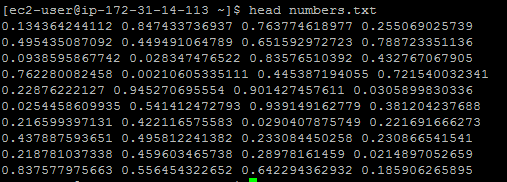
****

PART 3: Clustering

**Creating Numeric File:**

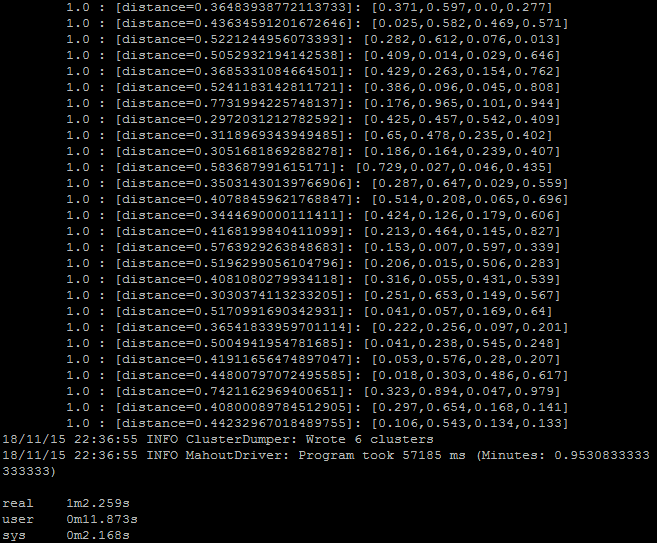
****

**First 10 lines in Numeric File:**

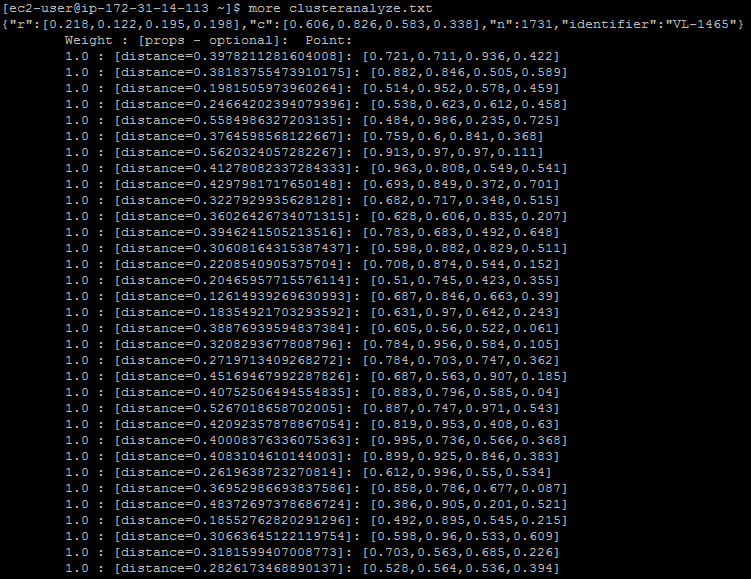
****

**Part A : Using ‘numbers.txt’ as input data**

**After running** time mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job

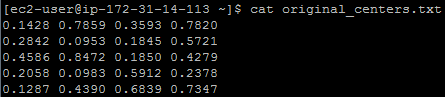
****

**After running** mahout clusterdump --input output/clusters-1-final --pointsDir output/clusteredPoints --output clusteranalyze.txt

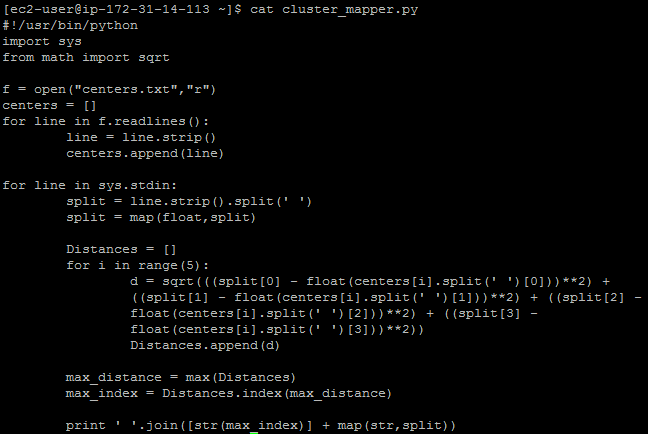


**Part B**

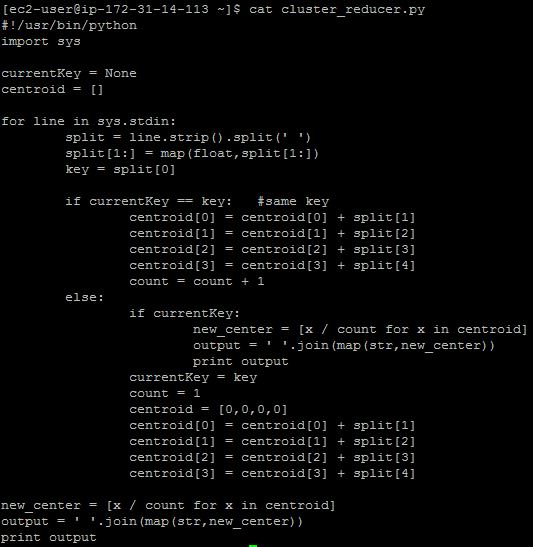
**Original random centers:**

****

**Mapper: Distances calculated using Euclidean distance**

****

**Reducer:**

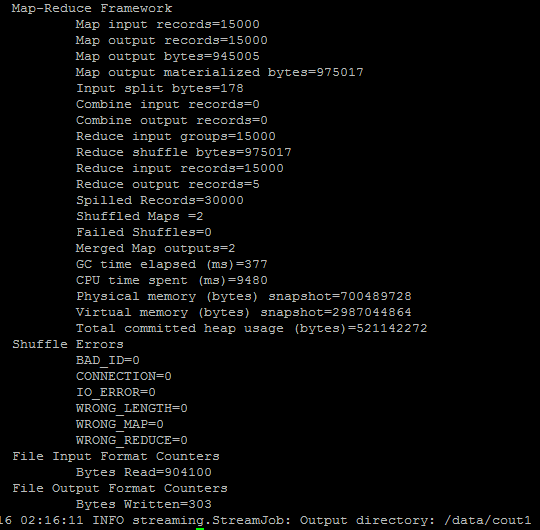
****

**1st Iteration**

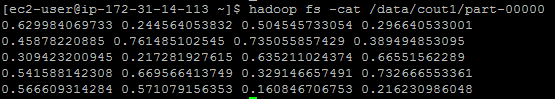
**Command Line:**

****

**After running job:**

****

**New cluster centers for next iteration:**

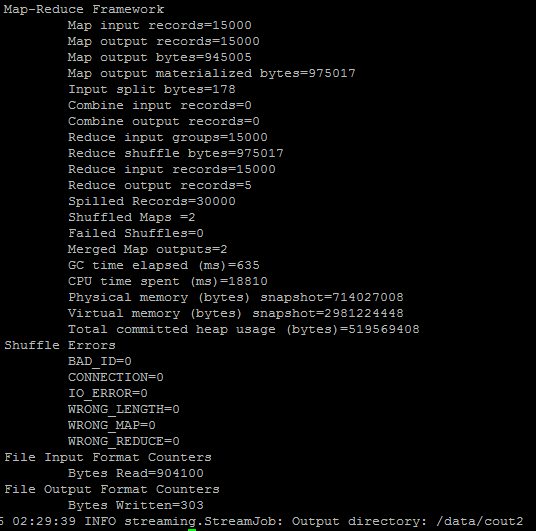
****

**2nd Iteration**

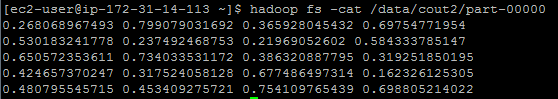
**Command line:**

****

**After running job:**

****

**New cluster centers for next iteration:**

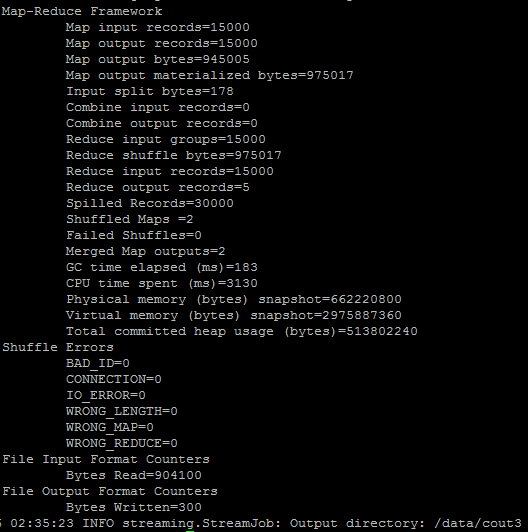
****

**3rd Iteration**

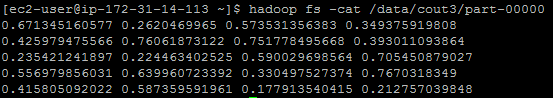
**Command line:**

****

**After running job:**

****

**Final cluster centers:**

****