Nicolas Escobar (escobarn)  
Harsh Reddy (hagandav)

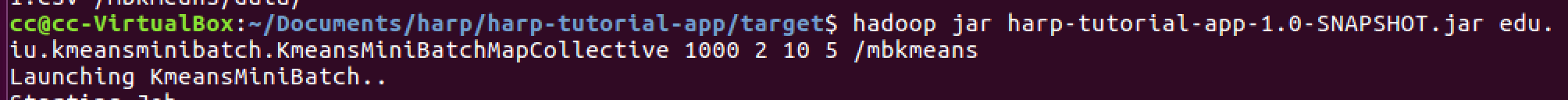
Cloud Computing - Project 8 Report  
Bonus Credits

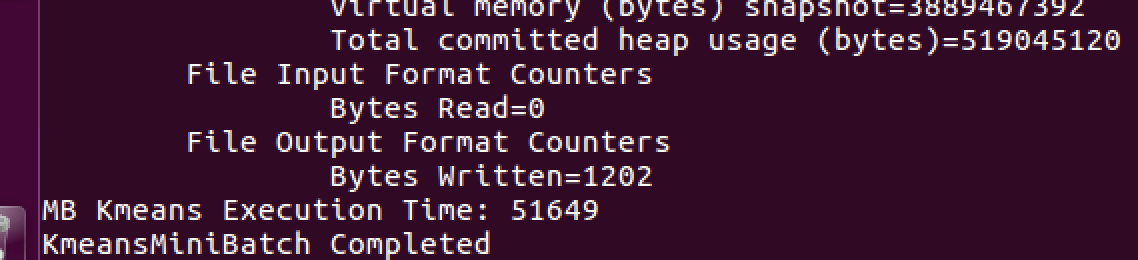
1. **Perform experiments on various (small, medium, large, etc) datasets**

We executed the following experiments:

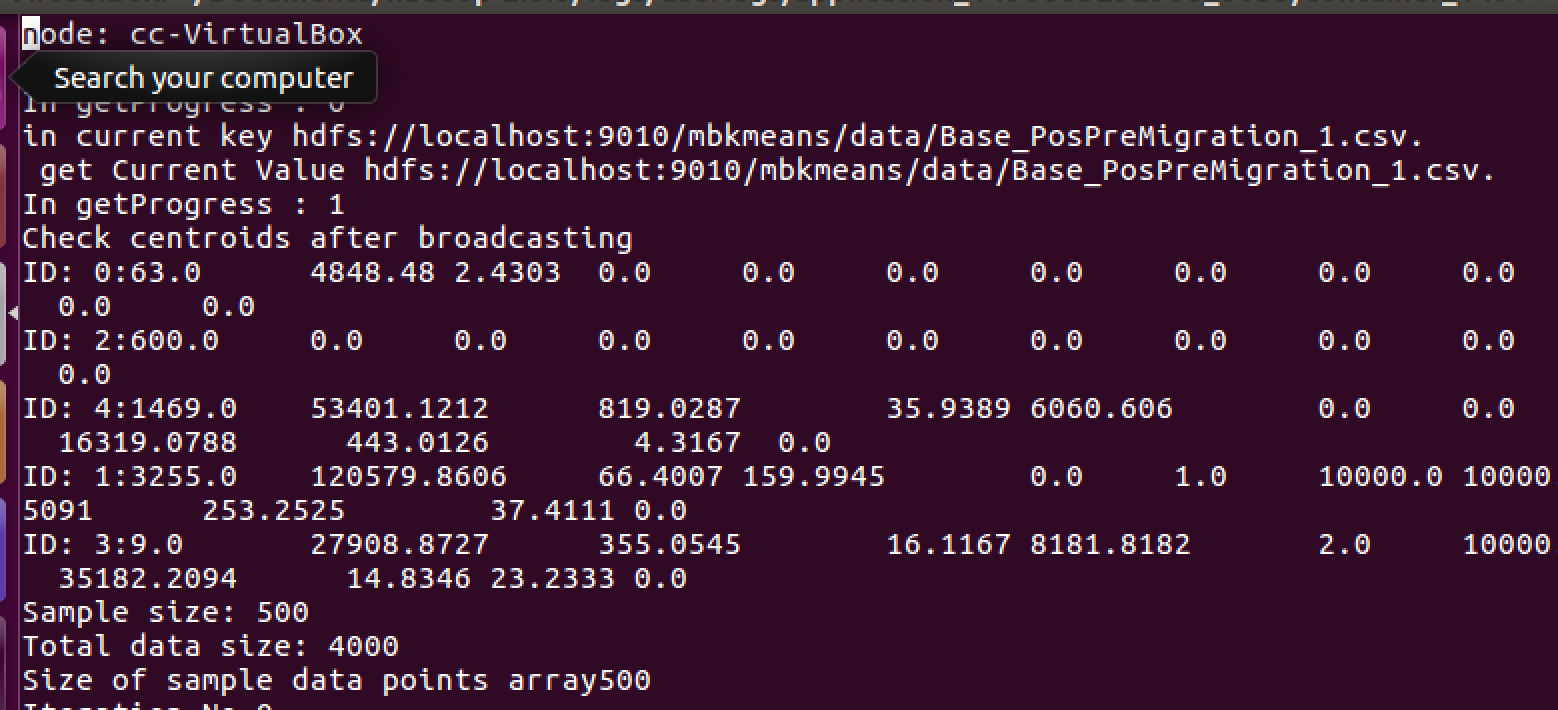
**Experiment #1 (Small data set, small batch size)**

* Input data: two files with 4,000 rows each
* Infrastructure: Local VM
* Mappers: 2
* Iterations: 10
* Centroids: 5
* Batch size: 1,000
* Execution time: 51,649 ms.





Log file execution sample:

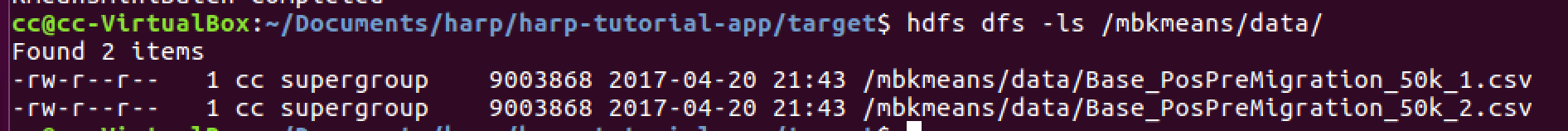


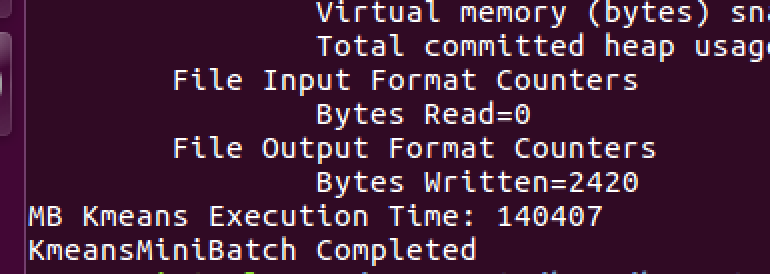
**Experiment #2 (Large data set, small batch size)**

* Input data: two files with over 50,000 rows each
* Infrastructure: Local VM
* Mappers: 2
* Iterations: 10
* Centroids: 5
* Batch size: 5,000
* Execution time: 51,426 ms.

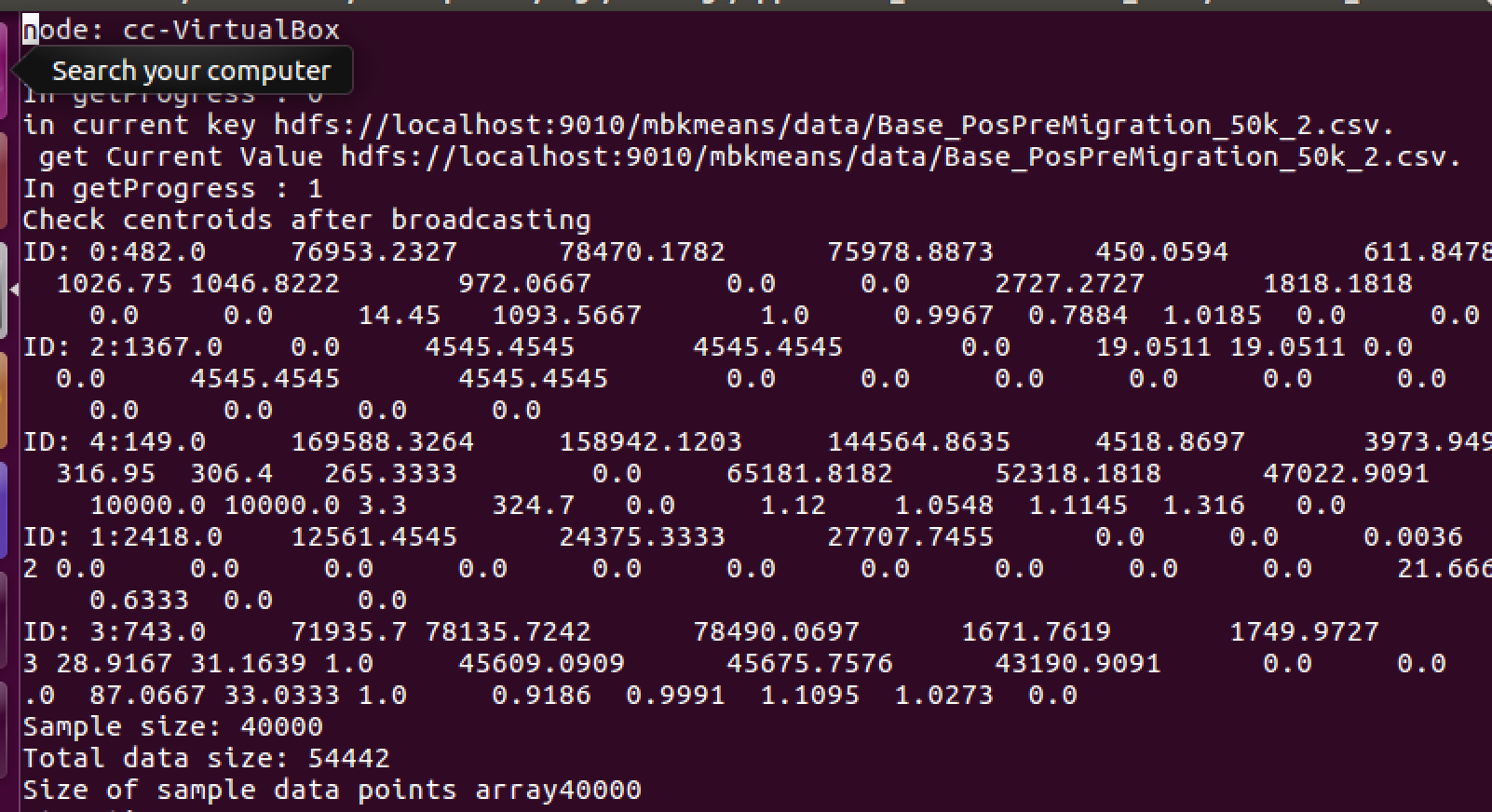
**Experiment #3 (Large data set, big batch size)**

* Input data: two files with over 50,000 rows each
* Infrastructure: Local VM
* Mappers: 2
* Iterations: 10
* Centroids: 5
* Batch size: 80,000
* Execution time: 140,407 ms.





Log file execution sample:

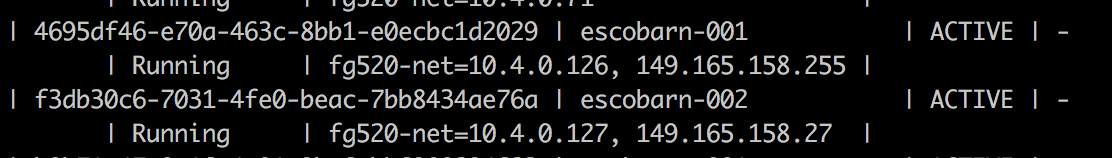


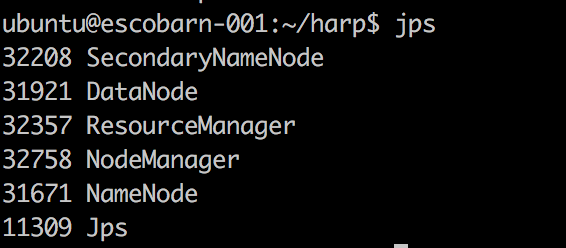
1. **Test your algorithm on at least 2 nodes on FutureSystem.**

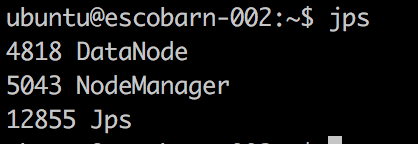
We configured two nodes on FutureSystem which are:

* escobarn-001 🡪 149.165.158.255
* escobarn-002 🡪 149.165.158.27

These nodes are running hadoop-2.6.5 with java 1.8 (openjdk 1.8.0\_111) and harp. We followed the instructions from <https://dsc-spidal.github.io/harp/docs/getting-started-cluster/> to configure and execute the cluster.







We have tested our algorithm against a large data set on the FutureSystems nodes and obtained the following results:

* Input data: two files with over 50,000 rows each
* Infrastructure: Future Systems Nodes running Ubuntu
* Mappers: 2
* Iterations: 10
* Centroids: 5
* Batch size: 80,000
* Execution time: 81,608 ms.

Screenshots of execution logs can be found below:

Sample execution logs

