-

Big Data

K-means clustering with MapReduce

Table of Contents

Project Scope

This aim of this project is to implement the K-means algorithm into a series of MapReduce tasks and run them on a Hadoop cluster, for small dataset of 150 coordinates.

Introduction

This report describes the implementation of a K-means algorithm in Hadoop, following the MapReduce methodology. The algorithm implementation is run against a small test case of 150 coordinates, where the initial 3 centroid coordinates have been provided. The first section reviews the mathematics behind the k-means clustering algorithm. The second section details the advantages of running such algorithm in a MapReduce framework. The third section proposes a Hadoop implementation and records the performance runs with different MapReduce settings. This section also provides a *sklearn* implementation in *python*. This is an enabler for comparing the correctness of the Hadoop implementation. It also serves as a performance benchmark, for the proposed dataset.

Source Code

mapper_kmeans.py: The K-means mapper Python implementation.
reducer_kmeans.py: The K-means reducer Python implementation.
scikit-learn_kmeans.ipynb: the *sklearn* K-means implementation in Python.

Dataset

The dataset consists of two files: a *clusters.txt* and a *data.txt* file. The first file contains three records relating to the initial clusters. It lists the clusters id and the x/y coordinates. The second file contains 150 data points, listing their x/y coordinates.

K-means Clustering

The k-means clustering is the process of organising data into small number of collections (i.e. the clusters). Each data point should belong to one and only cluster. The K-means finds for each point the cluster it belongs to, by minimising the distance from the data point to the cluster centre [1]. In other words, the cluster with the nearest mean [2][3]. This means minimising the within-cluster sum of squares, as shown below:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Where:

X: this is an d-dimensional observation (from the set of d-dimensional observations (X1,X2,…,Xn), where n represents the number of observations.

μ_i: is the centre of points in S_i. S_i is a cluster belonging to the cluster list S= {S1,S2,…,Sk). K is the number of clusters.

The k-means belongs to the family of unsupervised machine learning algorithms. As noted in [4], k-means is a popular algorithm as it couples fast convergence with a low level of implementation complexity. The computational complexity is of order *O(nkt)*, where *n* is the number of observations, *k* the number of clusters and *t* the number of iterations required for a stable convergence. Therefore, parallelisation is well suited to reduce the time to completion of this algorithm. Each cluster can be generated in parallel (instead of sequentially). This makes the solution horizontally scalable. In the optimum case, where the number of parallel tasks (i.e. the number of iterations required to complete the k-means convergence) is equal to the number of available processes, the maximum time to complete the task will only be equal to the time required to complete the most time consumption task.

Parallelisation of K-means Clustering on Hadoop

With the availability of powerful and relatively cheap multi-processors on personal computer or server ringed infrastructures, maximum benefits can be obtained by rewriting the initial synchronous version of the K-means algorithm offered in [4] into parallel and distributable form. The approach proposed in [4] focuses on the parallelisation of the algorithm, this project will go one step beyond by proposing an implementation compatible with Hadoop (as it offers extra infrastructure benefits). The following two sections are dedicated to:

i) the high-level Hadoop infrastructure, its advantages in the context of a scalable and reliable solution
ii) the description of the implemented algorithm in Python.

MapReduce in the context of the Hadoop infrastructure (HDFS)

Apache Hadoop is an open source software framework ecosystem built for managing and processing very large data sets, with faults tolerance (e.g. hardware/network failure) built at its core [4]. The core Hadoop platform consist of:

- Hadoop Distributed File System (HDFS) - the data storage layer
- MapReduce programming model – the task processing layer
- A number of other software such as YARN (a resource management platform), Pig/Hive (a high-level query languages), Sparks (an open source processing engine built around speed, ease of use, and sophisticated analytics), etc. This report does not dive deeper into any of these powerful tools as this is not the core of the issue under analysis.

The Hadoop MapReduce framework enables end users to write MapReduce pieces of code that can be run as parallel tasks. It is supported by the infrastructure shown in Figure 1. In a nutshell, each node in Hadoop instance has a single *NameNode* running on the cluster (the secondary *NameNode* is down and only started when the main one is not running). The HDFS cluster can store very large file (gigabytes, petabytes, …, zettabytes files). Reliability is enforced by replicating across multiple hosts, and does not require a redundant array of expensive disks (RAID) storage on hosts. HDFS allows for data to be stored into files, which are internally split into blocks. These blocks are stored in a set of *DataNodes*. The *DataNode* (usually one per node in the cluster) manages storage attached to the nodes that they run on. The *NameNode* is mainly responsible for i) executing file system namespace operations, ii) determining the mapping of blocks to *DataNodes*, iii) and checking the *DataNode* internal state. The *NameNode* may decide to restart/stop or create *DataNodes* automatically based on its configuration file information. The *DataNodes* serve read and write requests from the file system's clients. The *DataNodes* also perform block creation, deletion, and replication following instructions from the *NameNode*.

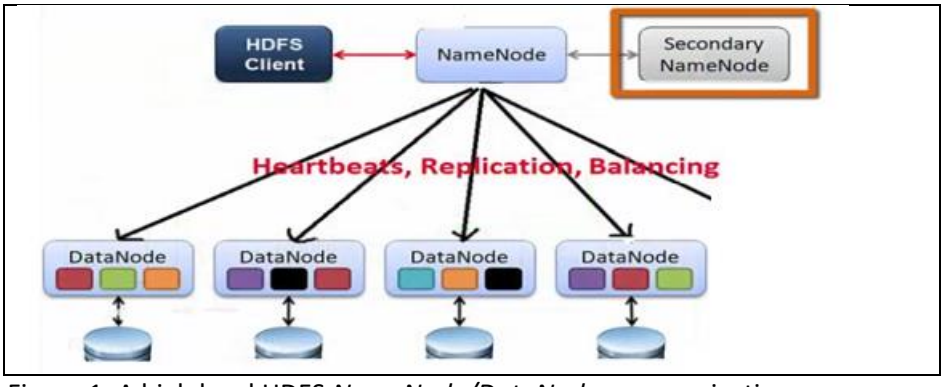


Figure 1: A high level HDFS *NameNode/DataNode* communication

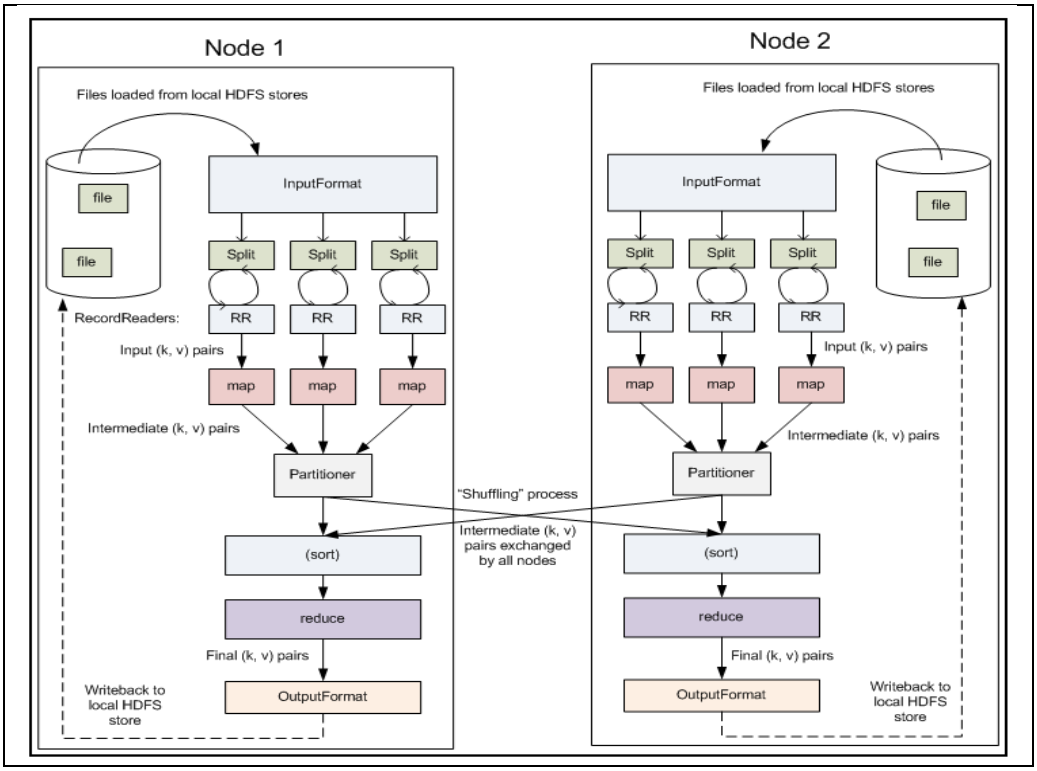


Figure 2: MapReduce in action

Figure 2 above, borrowed from [6], shows the data flow in a two nodes configuration, from the input to the output produced by the reducers. In our example, the input file is the *data.txt* file. The map task starts by reading the *clusters.txt* file to retrieve the initial clusters coordinates, then it reads each record contained in the *data.txt* file and calculates the *Euclidian* distance between each point and each cluster centroid. Each point in the *data.txt* file is associated to a *cluster id*, the key, and produce a value containing the x/y coordinates as well as a count flag for each record. Each map output emits an the output a <key,value> pair, e.g. <key=0, value=0.053139002741000294,0.0937351594239546,1>. There is no combiner in this implementation as there is no need for it on such small amount of data. Combiners, can help with reducing network traffic, by grouping the mappers output in advanced of the automatic shuffling/sort provided by the Hadoop infrastructure. However, as indicated in [5], there is an extra cost of running a combiner, in terms of execution time and maintenance. Therefore, performance tests should be carried out to establish whether it is worth adding an extra layer. This implementation will therefore only rely on the Hadoop shuffling/sort to distribute mapper tasks outputs to reducers. Each reducer takes the output of the shuffling/sort as an input and re-computes the new centroid, by averaging the x/y coordinates of each data point in each cluster. The average is calculated by adding all x coordinates (y coordinates), in a cluster, and divide by the number of points in the cluster. The operation is repeated for each cluster. The reducers output three new centroids (one for each of the initial cluster). Convergence is reached when the new centroids x/y coordinates do not change (or change below a user defined threshold), after a number of iterations. In the current implementation, the MapReduce is run 10 times. The detailed pseudo code of the MapReduce is presented in the next section.

Parallel K-means with MapReduce

The map and reduce pseudo code and implementation are detailed in the two tables below (pseudo code 1/2). The pseudo code has been extracted from [7]. The implementation only shows the most important functions. The entire code is available in the *mapper_kmeans.py* and *reducer_kmeans.py* files.

Input: Global variable *centers*, the offset *key*, the sample *value*
Output: <*key'*, *value'*> pair, where the *key'* is the index of the closest center point and *value'* is a string comprise of sample information

1. Construct the sample *instance* from *value*;
2. *minDis* = *Double.MAX_VALUE*;
3. *index* = -1;
4. For i=0 to *centers*.length do
   dis= *ComputeDist*(*instance*, *centers*[*i*]);
   If *dis* < *minDis* {
   minDis = *dis*;
   index = *i*;
   }
5. End For
6. Take *index* as *key'*;
7. Construct *value'* as a string comprise of the values of different dimensions;
8. output < *key'*, *value'* > pair;
9. End

```
#Calculate euclidian distance between two coordinates and return the distance
def get_distance_coords(x_coord, y_coord, center_x_coord, center_y_coord):
    dist = math.sqrt(math.pow(x_coord - center_x_coord,2) + math.pow(y_coord - center_y_coord,2))
    return dist

def get_nearest_cluster(x_coord, y_coord):
    #The cluster id is unknown originally, it is discovered by the code below and returned
    nearest_cluster_id = None
    #The closest_distance is set as a very far initial point on the grid
    #An initial distance of 1000 is more than enough in this case, as the points coordinates
    #are all between 0 and 1 in the current datasest.
    #This assumption may need to be reviewed with other datasets.
    #This could even be passed as a parameter...
    nearest_distance = 1000
    #For each cluster in the cluster list...
    for cluster in clusters:
        #Get the distance between the point coordinates and the cluster coordinates
        dist = get_distance_coords(x_coord, y_coord, cluster[1], cluster[2])
        #When the distance is less than the closest_distance, then
        #the closest_distance is reset to the current distance.
        #Else the process continues until a closer cluster is found.
        if dist < nearest_distance:
            nearest_cluster_id = cluster[0]
            nearest_distance = dist
    #Return the closest_cluster_id for point coordinates
    return nearest_cluster_id


#For each line in the standard input, the mapper produces a key/value pair output.
#It serves as input for the reducer. The key is the cluster_id.
#the value is a composite of the x and y point coordinates separated by a comma
for line in sys.stdin:
    #The line is empty, skip to the next line
    if len(line) == 0:
        continue
    #Get the point coordinates (x and y)
    coords = line.strip().split(" ")
    x_coord,y_coord = coords
    #Compute the nearest_cluster_id based on the point coordinates and
    #the in-memory cluster list (read from the CLUSTERS_FILENAME file)
    nearest_cluster_id = get_nearest_cluster(float(x_coord), float(y_coord))
    #Fabricate a key/value pair object ouput that serves as input for the reducer task.
    #Also add one at the end of the string in order to facilitate the counting of points
    #for a given cluster_id in the reducer.
    print ("%s\t%s" % (str(nearest_cluster_id),str(x_coord) + "," + str(y_coord) + "," + str(1) ))
```

Pseudo code 1 – The mapper pseudo code and implementation

Input: *key* is the index of the cluster, V is the list of the partial sums from different host
Output: $< key', value' >$ pair, where the *key*' is the index of the cluster, *value*' is a string representing the new center

1. Initialize one array record the sum of value of each dimensions of the samples contained in the same cluster, e.g. the samples in the list V;
2. Initialize a counter NUM as 0 to record the sum of sample number in the same cluster;
3. while(V.hasNext()){
   Construct the sample *instance* from V.next();
   Add the values of different dimensions of *instance* to the array
   NUM += *num*;
4. }
5. Divide the entries of the array by NUM to get the new center's coordinates;
6. Take *key* as *key*';
7. Construct *value*' as a string comprise of the *center*'s coordinates;
8. output $< key', value' >$ pair;
9. End

```python
import sys

#It is assumed cluster ids are always positive
current_cluster_id = -1
clusters = dict()

#For each line in the system input (i.e. the key/value pair) generated after shuffling/sorting and automatic shuffling.
#The key is the cluster_id, the value is a composite of the x and y point coordinates separated by a comma.
for line in sys.stdin:
    #The line is empty, skip to the next line
    if len(line) == 0:
        continue
    #Get the input data from the line
    data_mapped = line.strip().split("\t")
    #The data is not in the expected shape, then ignore and move to the next line
    if len(data_mapped) != 2:
        continue
    #Get the cluster_id and the sum of the x/y coordinates with the points count for a cluster
    cluster_id, count = data_mapped
    x_coord_sum, y_coord_sum, count = count.strip().split(",")

    #As the data has been shuffled/sorted,
    #the input data may contain more than one records with the same cluster_id.
    #We therefore need to reduce them here, prior to do the cluster new center calculation
    count = int(count)
    x_coord_sum = float(x_coord_sum)
    y_coord_sum = float(y_coord_sum)
    if current_cluster_id == cluster_id:
        count_total += count
        x_coord_total +=  x_coord_sum
        y_coord_total +=  y_coord_sum
    else:
        current_cluster_id = cluster_id
        count_total = count
        x_coord_total =  x_coord_sum
        y_coord_total =  y_coord_sum
    #This time the clustes map contains the final sum of the x/y coordinates and the final point count for a given cluster.
    clusters[current_cluster_id] = (x_coord_total, y_coord_total, count_total)

#Output the cluster_id, the average of x and y coordinates (new centroid) for each cluster
for key in clusters:
    x_coord_total, y_coord_total, count_total = clusters[key]
    print (str(key) + " " + str(x_coord_total/count_total) + " " + str(y_coord_total/count_total))
```

Pseudo code 2 – The reduce step

Result Generation

The *Hadoop* implementation and results

As shown in the *sklearn* implementation, only six iterations are required to attain convergence. Therefore, the Ma-Reduce task is also run manually and sequentially six times. As each task involves three reducers, three output files are produced containing for each one, one of the centroid coordinates. Having three reducers enable for maximum reducer parallelisation in this use case, as three centroid outputs are expected. It is necessary to merge them into one file before running the next job. This is achieved by running the *-getmerge* command line. By merging back under the *clusters.txt* initial file name, the merge replaces the content of the initial file with the new content. This has two advantages. First, it ensures there is only one version of the *clusters.txt* file at a time. Therefore, it simplifies both the command line argument list, and the python code (as the file name remains constant over several iterations). Second, it reduces the number of files (and therefore amount of data) stored on the server. Table 1 describes the steps involved. For the full explanation on how to run load the files on HDFS and run them, please c.f. Appendix B.

Action	**Command Line**
Run the 1st iteration	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,**clusters.txt** -D mapred.map.tasks=2 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output **iteration1**
Merge	hadoop fs -getmerge **iteration1 clusters.txt**
Run the 2nd iteration	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py, **clusters.txt** -D mapred.map.tasks=2 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output **iteration2**
Merge	hadoop fs -getmerge **iteration2 clusters.txt**
Run the 3rd iteration	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py, **clusters.txt** -D mapred.map.tasks=2 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output **iteration3**
Merge	hadoop fs -getmerge **iteration3 clusters.txt**
Run the 4th iteration	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py, **clusters.txt** -D mapred.map.tasks=2 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output- **iteration4**
Merge	hadoop fs -getmerge **iteration4 clusters.txt**
Run the 5th iteration	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py, **clusters.txt** -D mapred.map.tasks=2 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output **iteration5**
Merge	hadoop fs -getmerge **iteration5 clusters.txt**
Run the 6th iteration	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py, **clusters.txt** -D mapred.map.tasks=2 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output **iteration6**
Merge	hadoop fs -getmerge **iteration6 clusters.txt**

Table 1 – Sequential running of the MapReduce tasks

Table 2 shows the result of after each iteration and the final result obtained in the last merged. The results are very close to the one obtained in *sklearn,* as shown in the next section.

```
[fmare001@dsm1 ~]$ hadoop fs -ls iteration1
Found 4 items
-rw-r--r--   3 fmare001 hadoop          0 2017-02-09 18:10 iteration1/_SUCCESS
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:10 iteration1/part-00000
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:10 iteration1/part-00001
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:10 iteration1/part-00002
[fmare001@dsm1 ~]$
```

```
[fmare001@dsm1 ~]$ hadoop fs -ls iteration2
Found 4 items
-rw-r--r--   3 fmare001 hadoop          0 2017-02-09 18:12 iteration2/_SUCCESS
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:12 iteration2/part-00000
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:12 iteration2/part-00001
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:12 iteration2/part-00002
[fmare001@dsm1 ~]$
```

```
[fmare001@dsm1 ~]$ hadoop fs -ls iteration3
Found 4 items
-rw-r--r--   3 fmare001 hadoop          0 2017-02-09 18:14 iteration3/_SUCCESS
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:14 iteration3/part-00000
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:14 iteration3/part-00001
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:14 iteration3/part-00002
[fmare001@dsm1 ~]$
```

```
[fmare001@dsm1 ~]$ hadoop fs -ls iteration4
Found 4 items
-rw-r--r--   3 fmare001 hadoop          0 2017-02-09 18:15 iteration4/_SUCCESS
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:15 iteration4/part-00000
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:15 iteration4/part-00001
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:15 iteration4/part-00002
```

```
[fmare001@dsm1 ~]$ hadoop fs -ls iteration5
Found 4 items
-rw-r--r--   3 fmare001 hadoop          0 2017-02-09 18:15 iteration5/_SUCCESS
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:15 iteration5/part-00000
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:15 iteration5/part-00001
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:15 iteration5/part-00002
[fmare001@dsm1 ~]$
```

```
[fmare001@dsm1 ~]$ hadoop fs -ls iteration6
Found 4 items
-rw-r--r--   3 fmare001 hadoop          0 2017-02-09 18:16 iteration6/_SUCCESS
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:16 iteration6/part-00000
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:16 iteration6/part-00001
-rw-r--r--   3 fmare001 hadoop         33 2017-02-09 18:16 iteration6/part-00002
```

clusters - Notepad

File Edit Format View Help

```
2 0.612260390886 0.653529296302
0 0.200302217487 0.186358139854
1 0.842339504006 0.690283942841
```

Table2 - List of output produced after each iteration and the final result.

The *sklearn* implementation and results

The full code is available in *scikit-learn_kmeans.ipynb* file. The main functions are produced in *code block 1* below. As shown in the second section of *code block 1*, the new centroid have been generated for 1,2,….,10 iterations. The aim was to verify the results with the proposed Hadoop implementation. Two functions are used:

i) *produce_centroids ()*: it generates new centroids given the cluster number, and a defined number of iterations.

ii) *k_means_graph()*: it is used to generate a graphic representation of the centroid and the area surrounding each centroid and its points.

Note: To respect schema compatibility between the *data.txt* and the *clusters.txt* files in *sklearn,* a new cluster file named *clusters_sklearn.txt* has been created. It is a copy of the original *clusters.txt* file, from which the cluster id column has been removed.

In [2]:
```
#Generate the cluster graph
def k_means_graph(centroids):
    #Step size of the mesh.
    h = .02
    x_min, x_max = coords[:, 0].min() - 0.25, coords[:, 0].max() + 0.25
    y_min, y_max = coords[:, 1].min() - 0.25, coords[:, 1].max() + 0.25
    xx, yy = np.meshgrid(np.arange(x_min, x_max, h), np.arange(y_min, y_max, h))

    # Obtain labels for each point in mesh.
    Z = kmeans.predict(np.c_[xx.ravel(), yy.ravel()])
    # Put the result into a color plot
    Z = Z.reshape(xx.shape)

    plt.imshow(Z,  interpolation='nearest',
                   extent=(xx.min(), xx.max(), yy.min(), yy.max()),
                   cmap=plt.cm.Paired,
                   aspect='auto', origin='lower')


    plt.scatter(coords[:, 0], coords[:, 1],  color='b', )
    plt.scatter(centroids[:, 0], centroids[:, 1], marker='x',  color='w', s=50, linewidths=2)

    #plt.scatter(centroids[:, 0], centroids[:, 1])
    plt.title('K-means clustering on the coordinates dataset (k=3)\n'
          'Centroids are marked with white cross')
    plt.xlim(x_min, x_max)
    plt.ylim(y_min, y_max)
    plt.show()

#Produce the centroids for a given cluster number, an intial array of clusters and a number of iterations
def produce_centroids(p_n_clusters, p_init, p_max_iter):
    kmeans = KMeans(n_clusters=p_n_clusters, init=p_init, max_iter=p_max_iter).fit(coords)
    #Three clusters are being used
    centroids = kmeans.cluster_centers_
    print("The cluster centers for clusters= " +  str (p_n_clusters) + " and max_iter= " + str (p_max_iter))
    print(centroids)
    return kmeans, centroids

print("Shared function success")
```

In [10]:
```
CLUSTERS_FILENAME = "clusters_sklearn.txt"
DATA_FILENAME = "data.txt"

#Load the initial clusters into an array
initial_custers = np.loadtxt(CLUSTERS_FILENAME, dtype=np.float)
#Load the data into an array
coords = np.loadtxt(DATA_FILENAME, dtype=np.float)
#Generate the centroids for different iteration
kmeans, centroids = produce_centroids(3,initial_custers,1)
kmeans, centroids = produce_centroids(3,initial_custers,2)
kmeans, centroids = produce_centroids(3,initial_custers,3)
kmeans, centroids = produce_centroids(3,initial_custers,4)
kmeans, centroids = produce_centroids(3,initial_custers,5)
kmeans, centroids = produce_centroids(3,initial_custers,6)
kmeans, centroids = produce_centroids(3,initial_custers,7)
kmeans, centroids = produce_centroids(3,initial_custers,8)
kmeans, centroids = produce_centroids(3,initial_custers,9)
kmeans, centroids = produce_centroids(3,initial_custers,10)
k_means_graph(centroids)

print("Main body success")
```

Code block 1 – The k-means implementation using the *sklearn* python library

The result obtained as shown in Figure 3 below. The convergence is obtained after 6 iterations. This is a small number as the dataset comprises only 3 clusters and 150 data points. The final centroid coordinates are shown in the green box. The below map represents the final centroids (white cross), the area around each centroid and the points contained in each of the area.

```
The cluster centers for clusters= 3 and max_iter= 1
[[ 0.20198948  0.18973645]
 [ 0.96172426  0.8243381 ]
 [ 0.66127243  0.65832504]]
The cluster centers for clusters= 3 and max_iter= 2
[[ 0.20030222  0.18635814]
 [ 0.896381    0.82655273]
 [ 0.64757489  0.64807195]]
The cluster centers for clusters= 3 and max_iter= 3
[[ 0.20030222  0.18635814]
 [ 0.88061428  0.80540784]
 [ 0.64379772  0.64645514]]
The cluster centers for clusters= 3 and max_iter= 4
[[ 0.20030222  0.18635814]
 [ 0.84418909  0.75423336]
 [ 0.62868944  0.64218098]]
The cluster centers for clusters= 3 and max_iter= 5
[[ 0.20030222  0.18635814]
 [ 0.84595376  0.7101183 ]
 [ 0.61714045  0.64887742]]
The cluster centers for clusters= 3 and max_iter= 6
[[ 0.20030222  0.18635814]
 [ 0.8423395   0.69028394]
 [ 0.61226039  0.6535293 ]]
The cluster centers for clusters= 3 and max_iter= 7
[[ 0.20030222  0.18635814]
 [ 0.8423395   0.69028394]
 [ 0.61226039  0.6535293 ]]
The cluster centers for clusters= 3 and max_iter= 8
[[ 0.20030222  0.18635814]
 [ 0.8423395   0.69028394]
 [ 0.61226039  0.6535293 ]]
The cluster centers for clusters= 3 and max_iter= 9
[[ 0.20030222  0.18635814]
 [ 0.8423395   0.69028394]
 [ 0.61226039  0.6535293 ]]
The cluster centers for clusters= 3 and max_iter= 10
[[ 0.20030222  0.18635814]
 [ 0.8423395   0.69028394]
 [ 0.61226039  0.6535293 ]]
```

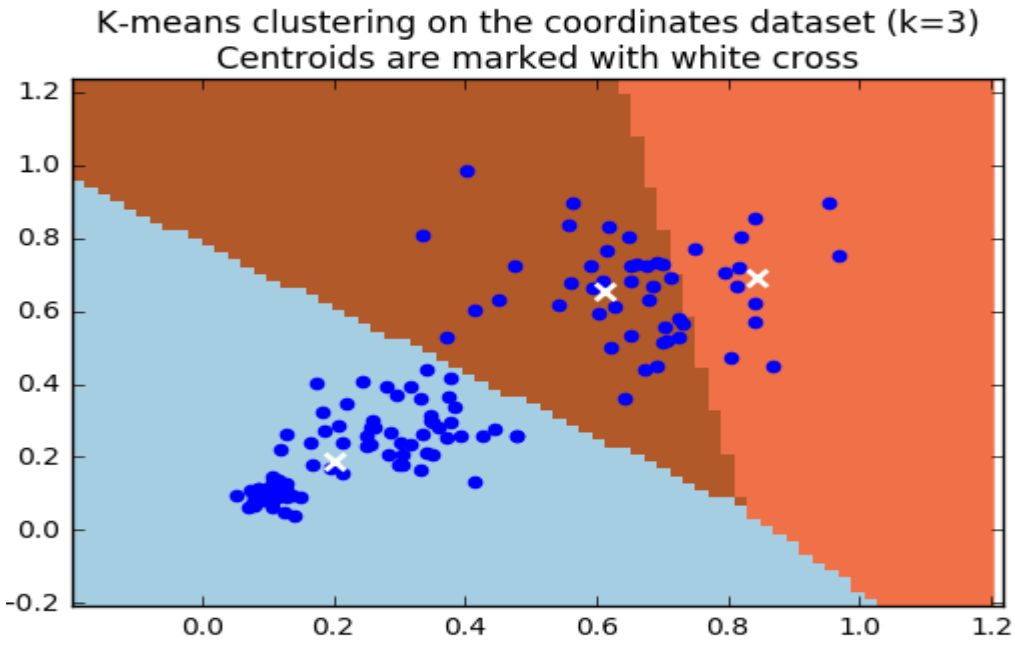


Figure 3 – The k-means results and converge level

Implementations Performance Comparison

First, the experiment shows that both the Hadoop MapReduce output and the Python code returns the same centroids after the optimal six iterations. Second, Figure 4, derived from the figures present in Appendix C, shows that the shortest time to completion is 23 seconds for MapReduce tasks. It is reached with a Hadoop MapReduce configuration set-up with 2 mappers and 3 reducers. It is interesting to note that, for a constant number of reducer (whether 1 or 3), increasing the number of mappers does not necessary reduce the time to completion. This is shown in the case with the mapping number is set to 50 (M=50_R=1) or 100 ((M=100_R=1), for a constant number of reducer set to 1. The picture is even clearer when the number of reducers is set to 3. The increase of mappers always increase the time to completion. This could be explained by extra network overheads produced by the Hadoop framework. Third, it is interesting to note that for such a small data size (150 records), the time to completion in Hadoop (green rectangle) is 23 times slower than running the same task in Python using the *sklearn* library (yellow rectangle). This underlines the fact that Hadoop is certainly efficient for very large datasets. For very small ones, this is not the right tool, as it is counterproductive.

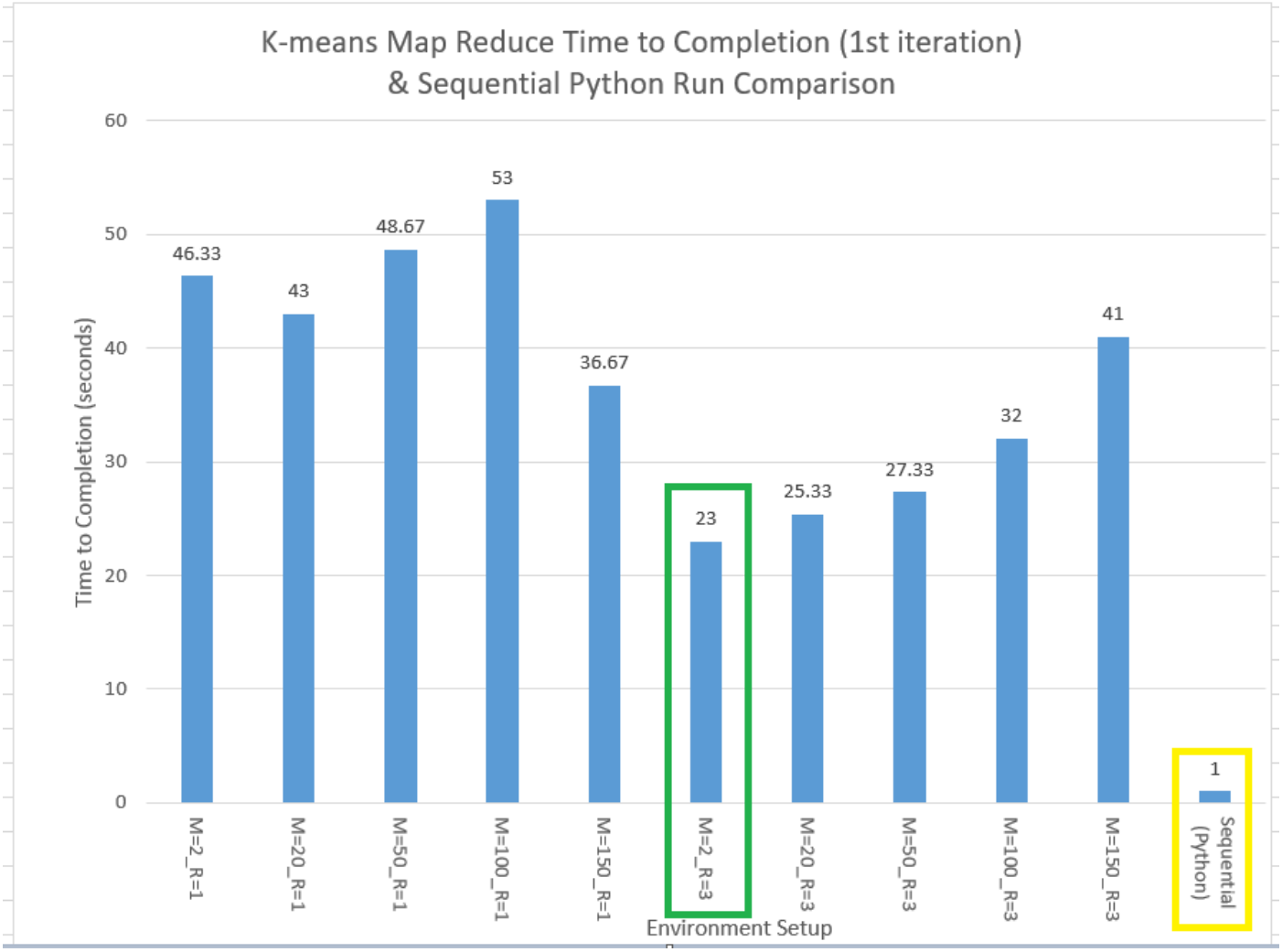


Figure 4 – Performance Test Results. M=2 means 2 mappers and R=1 means 1 reducers

Bibliography

[1] K-means clustering [Online]. Available At: http://onmyphd.com/?p=k-means.clustering [Accessed: 09-February-2017]
[2] *k*-means clustering [Online]. Available At: https://en.wikipedia.org/wiki/K-means_clustering [Accessed: 09-February-2017]
[3] MacQueen J., *SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS* [Online], Available At: http://www.stat.ucla.edu/~macqueen/PP07.pdf [Accessed: 09-February-2017]
[4] Kerdprasop K., Kerdprasop N., *Parallelization of K-Means Clustering on Multi-Core Processors* [Online], Available At: http://www.wseas.us/e-library/conferences/2010/Japan/ACS/ACS-74.pdf [Accessed: 09-February-2017]
[5] Lam C., *Hadoop in Action* [Online], Available At: http://www.chinastor.org/upload/2013-11/13111115436557.pdf, p98, [Accessed: 09-February-2017]
[6] *Module 4: MapReduce* [Online], Available At: https://developer.yahoo.com/hadoop/tutorial/module4.html, [Accessed: 09-February-2017]
[7] Zhao W., Huifang M., Qing H., *Parallel K-Means Clustering Based on MapReduce*[Online], Available At: http://www.cs.ucsb.edu/~veronika/MAE/parallelkmeansmapreduce_zhao.pdf, [Accessed: 09-February-2017]

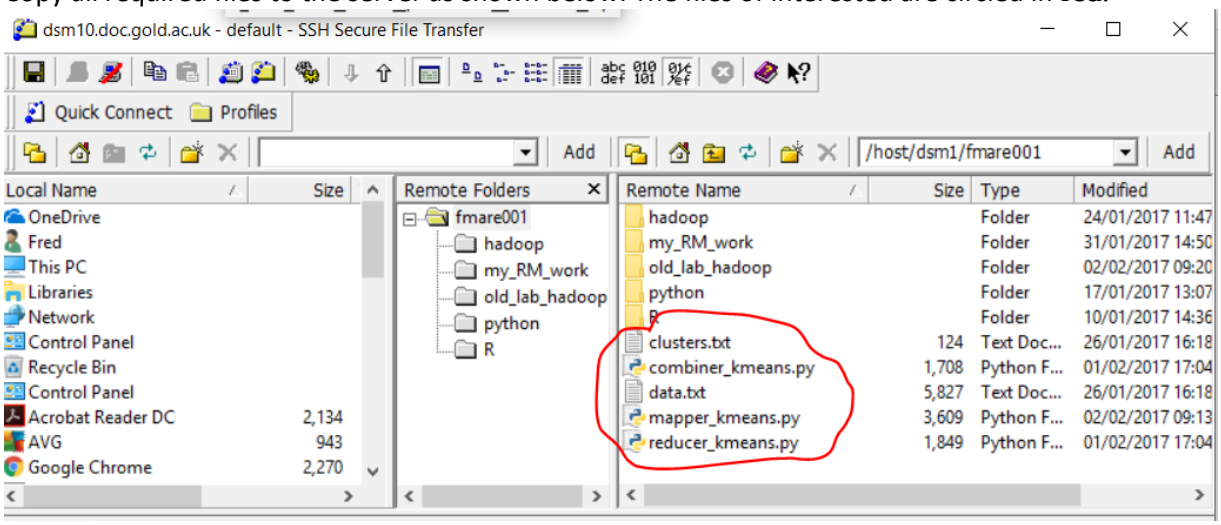
Appendix

Appendix A – How to run the code locally on the windows 10 box

1. Python 3.x should be installed on the box
2. The *data.txt*, *cluster.txt*, *mapper_kmeans.py* and *reducer_kmeans.py* file should be stored in the same directory.
3. Open a command line and cd to the directory where the mapper/ reducer files live
4. Run the following cmd in the window "python mapper_kmeans.py < data.txt | sort | reducer_kmeans.py"

Appendix B – How to run the code on a Hadoop server cluster

1. Copy all required files to the server as shown below. The files of interested are circled in red.



2. Log into the server and cd to the location where the files have been stored (here the user directory)
3. From this location, give the python scripts the executable mode (permission) bit, by executing a command such as :
   chmod +x kmeans.py mapper_kmeans.py reducer_kmeans.py
4. Then copy all files to Hadoop file system, by running the following commands:
   hadoop fs -copyFromLocal clusters.txt
   hadoop fs -copyFromLocal data.txt
5. Check the code is running without any errors as a python command line:
   cat data.txt | ./mapper_kmeans.py | sort | ./reducer_kmeans.py
6. Run the MapReduce job on the server. For example:
   hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=2 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output iteration1

Appendix C – Hadoop MapReduce Performance Test Summary

The below two tables present the time performance summary results of running a MapReduce task with different numbers of mappers and reducers. The details of the output are presented in Appendix D.

Map task #	Reduce task #	Cmd Line	Run Id	Start Time	End Time	Duration (seconds)	Average Duration (seconds)
2	1	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_1 The three runs log successfully in the vanilla_1, vanilla_2 and vanilla_3 output. Only the first command line showing vanilla_1 is mentioned above.	1	17/02/07 09:05:56	17/02/07 09:06:37	42	46.33
			2	17/02/07 09:09:13	17/02/07 09:09:59	46	
			3	17/02/07 09:16:00	17/02/07 09:16:51	51	
20	1	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt **-D mapred.map.tasks=20** -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_4	4	17/02/07 10:01:46	17/02/07 10:02:25	39	43
			5	17/02/07 10:07:02	17/02/07 10:07:48	49	
			6	17/02/07 10:11:17	17/02/07 10:11:52	41	
50	1	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt **-D mapr ed.map.tasks=50** -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_7	7	17/02/07 10:18:39	17/02/07 10:19:29	50	48.67
			8	17/02/07 10:22:20	17/02/07 10:23:07	47	
			9	17/02/07 10:26:01	17/02/07 10:26:50	49	
100	1	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt **-D mapred.map.tasks=100** -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_10	10	17/02/07 10:29:29	17/02/07 10:29:58	89	53
			11	17/02/07 10:32:31	17/02/07 10:33:02	31	
			12	17/02/07 10:35:03	17/02/07 10:35:36	39	
150	1	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt **-D mapred.map.tasks=150** -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_13	13	17/02/07 10:38:05	17/02/07 10:38:44	39	36.67
			14	17/02/07 10:42:35	17/02/07 10:43:14	39	
			15	17/02/07 10:44:54	17/02/07 10:45:33	32	
2	3	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt **-D mapred.map.tasks=2 -D mapred.reduce.tasks=3** -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_16	16	17/02/07 10:52:16	17/02/07 10:52:38	16	23
			17	17/02/07 10:55:51	17/02/07 10:56:17	26	
			18	17/02/07 10:58:32	17/02/07 10:58:59	27	
20	3	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt **-D mapred.map.tasks=20 -D mapred.reduce.tasks=3** -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_19	19	17/02/07 11:02:44	17/02/07 11:03:11	27	25.33
			20	17/02/07 11:49:53	17/02/07 11:50:18	25	
			21	17/02/07 11:51:51	17/02/07 11:52:15	24	

50	3	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -**D mapred.map.tasks=50 -D mapred.reduce.tasks=3** -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_22	22	17/02/07 11:58:27	17/02/07 11:58:53	26	27.33
			23	17/02/07 12:01:47	17/02/07 12:02:20	27	
			24	17/02/07 12:04:04	17/02/07 12:04:33	29	
100	3	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -**D mapred.map.tasks=100 -D mapred.reduce.tasks=3** -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_25	25	17/02/07 12:07:20	17/02/07 12:07:54	34	32
			26	17/02/07 12:10:45	17/02/07 12:11:18	33	
			27	17/02/07 12:12:35	17/02/07 12:13:04	29	
150	3	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -**D mapred.map.tasks=150 -D mapred.reduce.tasks=3** -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_28	28	17/02/07 12:15:55	17/02/07 12:16:35	40	41
			29	17/02/07 12:26:16	17/02/07 12:26:56	40	
			30	17/02/07 12:28:03	17/02/07 12:28:46	43	

Appendix D – Hadoop MapReduce Runs Output

Run Id	Command Input and Out
1	[fmare001@dsm6 ~]$ hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt - mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_1 packageJobJar: [/tmp/hadoop-unjar7265035614362852728/] [] /tmp/streamjob2069450753511166672.jar tmpDir=null 17/02/07 09:05:56 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 09:05:57 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 09:05:59 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 09:06:00 INFO mapreduce.JobSubmitter: number of splits:2 17/02/07 09:06:00 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0259 17/02/07 09:06:01 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0259 17/02/07 09:06:01 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0259/ 17/02/07 09:06:01 INFO mapreduce.Job: Running job: job_1484631414223_0259 17/02/07 09:06:10 INFO mapreduce.Job: Job job_1484631414223_0259 running in uber mode : false 17/02/07 09:06:10 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 09:06:22 INFO mapreduce.Job: map 50% reduce 0% 17/02/07 09:06:23 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 09:06:36 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 09:06:37 INFO mapreduce.Job: Job job_1484631414223_0259 completed successfully 17/02/07 09:06:37 INFO mapreduce.Job: Counters: 49 File System Counters FILE: Number of bytes read=6734 FILE: Number of bytes written=340521 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=7192 HDFS: Number of bytes written=96 HDFS: Number of read operations=9 HDFS: Number of large read operations=0 HDFS: Number of write operations=2 Job Counters Launched map tasks=2 Launched reduce tasks=1 Rack-local map tasks=2 Total time spent by all maps in occupied slots (ms)=19449 Total time spent by all reduces in occupied slots (ms)=11605 Total time spent by all map tasks (ms)=19449 Total time spent by all reduce tasks (ms)=11605 Total vcore-seconds taken by all map tasks=19449 Total vcore-seconds taken by all reduce tasks=11605 Total megabyte-seconds taken by all map tasks=19915776 Total megabyte-seconds taken by all reduce tasks=11883520 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428

	Map output materialized bytes=6740 Input split bytes=182 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=6740 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =2 Failed Shuffles=0 Merged Map outputs=2 GC time elapsed (ms)=189 CPU time spent (ms)=2800 Physical memory (bytes) snapshot=701222912 Virtual memory (bytes) snapshot=3018395648 Total committed heap usage (bytes)=603979776 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=7010 File Output Format Counters Bytes Written=96 17/02/07 09:06:37 INFO streaming.StreamJob: Output directory: vanilla_1
2	[fmare001@dsm6 ~]$ hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -mapper mapper_kmea ns.py -reducer reducer_kmeans.py -input data.txt -output vanilla_2 packageJobJar: [/tmp/hadoop-unjar9013099818303454/] [] /tmp/streamjob435237820333828711.jar tmpDir=null 17/02/07 09:09:13 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 09:09:14 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 09:09:16 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 09:09:17 INFO mapreduce.JobSubmitter: number of splits:2 17/02/07 09:09:17 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0260 17/02/07 09:09:18 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0260 17/02/07 09:09:18 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0260/ 17/02/07 09:09:18 INFO mapreduce.Job: Running job: job_1484631414223_0260 17/02/07 09:09:35 INFO mapreduce.Job: Job job_1484631414223_0260 running in uber mode : false 17/02/07 09:09:35 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 09:09:48 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 09:09:58 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 09:09:59 INFO mapreduce.Job: Job job_1484631414223_0260 completed successfully 17/02/07 09:09:59 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6734 FILE: Number of bytes written=340518 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=7192 HDFS: Number of bytes written=96 HDFS: Number of read operations=9 HDFS: Number of large read operations=0 HDFS: Number of write operations=2 Job Counters Launched map tasks=2 Launched reduce tasks=1 Data-local map tasks=1 Rack-local map tasks=1 Total time spent by all maps in occupied slots (ms)=20733 Total time spent by all reduces in occupied slots (ms)=8360 Total time spent by all map tasks (ms)=20733 Total time spent by all reduce tasks (ms)=8360 Total vcore-seconds taken by all map tasks=20733 Total vcore-seconds taken by all reduce tasks=8360 Total megabyte-seconds taken by all map tasks=21230592 Total megabyte-seconds taken by all reduce tasks=8560640 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428

	Map output materialized bytes=6740 Input split bytes=182 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=6740 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =2 Failed Shuffles=0 Merged Map outputs=2 GC time elapsed (ms)=173 CPU time spent (ms)=2940 Physical memory (bytes) snapshot=708911104 Virtual memory (bytes) snapshot=3005394944 Total committed heap usage (bytes)=603979776 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=7010 File Output Format Counters Bytes Written=96 17/02/07 09:09:59 INFO streaming.StreamJob: Output directory: vanilla_2
3	[fmare001@dsm6 ~]$ hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -mapper mapper_kmea ns.py -reducer reducer_kmeans.py -input data.txt -output vanilla_3 packageJobJar: [/tmp/hadoop-unjar9107472755937574834/] [] /tmp/streamjob1282018103941524506.jar tmpDir=null 17/02/07 09:16:00 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 09:16:01 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 09:16:04 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 09:16:04 INFO mapreduce.JobSubmitter: number of splits:2 17/02/07 09:16:05 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0261 17/02/07 09:16:06 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0261 17/02/07 09:16:06 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0261/ 17/02/07 09:16:06 INFO mapreduce.Job: Running job: job_1484631414223_0261 17/02/07 09:16:23 INFO mapreduce.Job: Job job_1484631414223_0261 running in uber mode : false 17/02/07 09:16:23 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 09:16:34 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 09:16:49 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 09:16:50 INFO mapreduce.Job: Job job_1484631414223_0261 completed successfully 17/02/07 09:16:51 INFO mapreduce.Job: Counters: 49 File System Counters FILE: Number of bytes read=6734 FILE: Number of bytes written=340521 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=7192 HDFS: Number of bytes written=96 HDFS: Number of read operations=9 HDFS: Number of large read operations=0 HDFS: Number of write operations=2 Job Counters Launched map tasks=2 Launched reduce tasks=1 Rack-local map tasks=2 Total time spent by all maps in occupied slots (ms)=18411 Total time spent by all reduces in occupied slots (ms)=11893 Total time spent by all map tasks (ms)=18411 Total time spent by all reduce tasks (ms)=11893 Total vcore-seconds taken by all map tasks=18411 Total vcore-seconds taken by all reduce tasks=11893 Total megabyte-seconds taken by all map tasks=18852864 Total megabyte-seconds taken by all reduce tasks=12178432 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=6740

	Input split bytes=182 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=6740 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =2 Failed Shuffles=0 Merged Map outputs=2 GC time elapsed (ms)=223 CPU time spent (ms)=2950 Physical memory (bytes) snapshot=701702144 Virtual memory (bytes) snapshot=3005566976 Total committed heap usage (bytes)=603979776 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=7010 File Output Format Counters Bytes Written=96 17/02/07 09:16:51 INFO streaming.StreamJob: Output directory: vanilla_3
4	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=20 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_4 17/02/07 10:01:46 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps packageJobJar: [/tmp/hadoop-unjar5551257645863748404/] [] /tmp/streamjob8474597247880314960.jar tmpDir=null 17/02/07 10:01:48 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:01:49 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:01:56 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:01:56 INFO mapreduce.JobSubmitter: number of splits:20 17/02/07 10:01:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0262 17/02/07 10:01:58 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0262 17/02/07 10:01:58 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0262/ 17/02/07 10:01:58 INFO mapreduce.Job: Running job: job_1484631414223_0262 17/02/07 10:02:08 INFO mapreduce.Job: Job job_1484631414223_0262 running in uber mode : false 17/02/07 10:02:08 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 10:02:14 INFO mapreduce.Job: map 5% reduce 0% 17/02/07 10:02:16 INFO mapreduce.Job: map 15% reduce 0% 17/02/07 10:02:18 INFO mapreduce.Job: map 25% reduce 0% 17/02/07 10:02:20 INFO mapreduce.Job: map 55% reduce 0% 17/02/07 10:02:21 INFO mapreduce.Job: map 85% reduce 0% 17/02/07 10:02:22 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 10:02:23 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 10:02:24 INFO mapreduce.Job: Job job_1484631414223_0262 completed successfully 17/02/07 10:02:25 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6734 FILE: Number of bytes written=2303272 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=57049 HDFS: Number of bytes written=96 HDFS: Number of read operations=63 HDFS: Number of large read operations=0 HDFS: Number of write operations=2 Job Counters Launched map tasks=20 Launched reduce tasks=1 Data-local map tasks=1 Rack-local map tasks=19 Total time spent by all maps in occupied slots (ms)=161935 Total time spent by all reduces in occupied slots (ms)=6789 Total time spent by all map tasks (ms)=161935 Total time spent by all reduce tasks (ms)=6789 Total vcore-seconds taken by all map tasks=161935 Total vcore-seconds taken by all reduce tasks=6789

	Total megabyte-seconds taken by all map tasks=165821440 Total megabyte-seconds taken by all reduce tasks=6951936 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=6848 Input split bytes=1820 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=6848 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =20 Failed Shuffles=0 Merged Map outputs=20 GC time elapsed (ms)=1579 CPU time spent (ms)=14960 Physical memory (bytes) snapshot=5490307072 Virtual memory (bytes) snapshot=21064613888 Total committed heap usage (bytes)=4227858432 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=55229 File Output Format Counters Bytes Written=96 17/02/07 10:02:25 INFO streaming.StreamJob: Output directory: vanilla_4
5	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=20 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_5 17/02/07 10:07:02 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps packageJobJar: [/tmp/hadoop-unjar6152309119000503431/] [] /tmp/streamjob6218087641027449131.jar tmpDir=null 17/02/07 10:07:05 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:07:05 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:07:08 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:07:08 INFO mapreduce.JobSubmitter: number of splits:20 17/02/07 10:07:09 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0263 17/02/07 10:07:10 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0263 17/02/07 10:07:10 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0263/ 17/02/07 10:07:10 INFO mapreduce.Job: Running job: job_1484631414223_0263 17/02/07 10:07:26 INFO mapreduce.Job: Job job_1484631414223_0263 running in uber mode : false 17/02/07 10:07:26 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 10:07:36 INFO mapreduce.Job: map 15% reduce 0% 17/02/07 10:07:37 INFO mapreduce.Job: map 20% reduce 0% 17/02/07 10:07:38 INFO mapreduce.Job: map 35% reduce 0% 17/02/07 10:07:39 INFO mapreduce.Job: map 80% reduce 0% 17/02/07 10:07:40 INFO mapreduce.Job: map 95% reduce 0% 17/02/07 10:07:41 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 10:07:47 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 10:07:48 INFO mapreduce.Job: Job job_1484631414223_0263 completed successfully 17/02/07 10:07:48 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6734 FILE: Number of bytes written=2303272 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=57049 HDFS: Number of bytes written=96 HDFS: Number of read operations=63 HDFS: Number of large read operations=0 HDFS: Number of write operations=2 Job Counters Launched map tasks=20 Launched reduce tasks=1 Data-local map tasks=6

	Rack-local map tasks=14 Total time spent by all maps in occupied slots (ms)=173256 Total time spent by all reduces in occupied slots (ms)=9112 Total time spent by all map tasks (ms)=173256 Total time spent by all reduce tasks (ms)=9112 Total vcore-seconds taken by all map tasks=173256 Total vcore-seconds taken by all reduce tasks=9112 Total megabyte-seconds taken by all map tasks=177414144 Total megabyte-seconds taken by all reduce tasks=9330688 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=6848 Input split bytes=1820 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=6848 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =20 Failed Shuffles=0 Merged Map outputs=20 GC time elapsed (ms)=1472 CPU time spent (ms)=15470 Physical memory (bytes) snapshot=5509173248 Virtual memory (bytes) snapshot=21022781440 Total committed heap usage (bytes)=4227858432 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=55229 File Output Format Counters Bytes Written=96 17/02/07 10:07:48 INFO streaming.StreamJob: Output directory: vanilla_5
6	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=20 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_6 17/02/07 10:11:17 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps packageJobJar: [/tmp/hadoop-unjar7436709825213920148/] [] /tmp/streamjob7409175664356983014.jar tmpDir=null 17/02/07 10:11:19 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:11:20 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:11:21 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:11:22 INFO mapreduce.JobSubmitter: number of splits:20 17/02/07 10:11:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0264 17/02/07 10:11:23 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0264 17/02/07 10:11:23 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0264/ 17/02/07 10:11:23 INFO mapreduce.Job: Running job: job_1484631414223_0264 17/02/07 10:11:33 INFO mapreduce.Job: Job job_1484631414223_0264 running in uber mode : false 17/02/07 10:11:33 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 10:11:40 INFO mapreduce.Job: map 5% reduce 0% 17/02/07 10:11:41 INFO mapreduce.Job: map 10% reduce 0% 17/02/07 10:11:45 INFO mapreduce.Job: map 30% reduce 0% 17/02/07 10:11:46 INFO mapreduce.Job: map 75% reduce 0% 17/02/07 10:11:47 INFO mapreduce.Job: map 95% reduce 0% 17/02/07 10:11:48 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 10:11:51 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 10:11:52 INFO mapreduce.Job: Job job_1484631414223_0264 completed successfully 17/02/07 10:11:52 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6734 FILE: Number of bytes written=2303272 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=57049 HDFS: Number of bytes written=96

	HDFS: Number of read operations=63 HDFS: Number of large read operations=0 HDFS: Number of write operations=2 Job Counters Launched map tasks=20 Launched reduce tasks=1 Data-local map tasks=1 Rack-local map tasks=19 Total time spent by all maps in occupied slots (ms)=170472 Total time spent by all reduces in occupied slots (ms)=8347 Total time spent by all map tasks (ms)=170472 Total time spent by all reduce tasks (ms)=8347 Total vcore-seconds taken by all map tasks=170472 Total vcore-seconds taken by all reduce tasks=8347 Total megabyte-seconds taken by all map tasks=174563328 Total megabyte-seconds taken by all reduce tasks=8547328 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=6848 Input split bytes=1820 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=6848 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =20 Failed Shuffles=0 Merged Map outputs=20 GC time elapsed (ms)=1320 CPU time spent (ms)=15320 Physical memory (bytes) snapshot=5509873664 Virtual memory (bytes) snapshot=21048688640 Total committed heap usage (bytes)=4227858432 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=55229 File Output Format Counters Bytes Written=96 17/02/07 10:11:52 INFO streaming.StreamJob: Output directory: vanilla_6
7	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapr ed.map.tasks=50 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_7 17/02/07 10:18:39 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps packageJobJar: [/tmp/hadoop-unjar4873364327632448227/] [] /tmp/streamjob3633061171693395048.jar tmpDir=null 17/02/07 10:18:41 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:18:42 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:18:45 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:18:45 INFO mapreduce.JobSubmitter: number of splits:51 17/02/07 10:18:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0265 17/02/07 10:18:46 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0265 17/02/07 10:18:46 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0265/ 17/02/07 10:18:47 INFO mapreduce.Job: Running job: job_1484631414223_0265 17/02/07 10:19:01 INFO mapreduce.Job: Job job_1484631414223_0265 running in uber mode : false 17/02/07 10:19:01 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 10:19:10 INFO mapreduce.Job: map 2% reduce 0% 17/02/07 10:19:11 INFO mapreduce.Job: map 6% reduce 0% 17/02/07 10:19:12 INFO mapreduce.Job: map 12% reduce 0% 17/02/07 10:19:13 INFO mapreduce.Job: map 24% reduce 0% 17/02/07 10:19:14 INFO mapreduce.Job: map 39% reduce 0% 17/02/07 10:19:15 INFO mapreduce.Job: map 63% reduce 0% 17/02/07 10:19:16 INFO mapreduce.Job: map 73% reduce 0% 17/02/07 10:19:17 INFO mapreduce.Job: map 75% reduce 0% 17/02/07 10:19:18 INFO mapreduce.Job: map 80% reduce 0% 17/02/07 10:19:19 INFO mapreduce.Job: map 84% reduce 0% 17/02/07 10:19:20 INFO mapreduce.Job: map 88% reduce 0%

	17/02/07 10:19:22 INFO mapreduce.Job: map 92% reduce 0% 17/02/07 10:19:23 INFO mapreduce.Job: map 94% reduce 0% 17/02/07 10:19:24 INFO mapreduce.Job: map 98% reduce 0% 17/02/07 10:19:25 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 10:19:26 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 10:19:28 INFO mapreduce.Job: Job job_1484631414223_0265 completed successfully 17/02/07 10:19:29 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6734 FILE: Number of bytes written=5683574 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=140133 HDFS: Number of bytes written=96 HDFS: Number of read operations=156 HDFS: Number of large read operations=0 HDFS: Number of write operations=2 Job Counters Launched map tasks=51 Launched reduce tasks=1 Data-local map tasks=13 Rack-local map tasks=38 Total time spent by all maps in occupied slots (ms)=428333 Total time spent by all reduces in occupied slots (ms)=12572 Total time spent by all map tasks (ms)=428333 Total time spent by all reduce tasks (ms)=12572 Total vcore-seconds taken by all map tasks=428333 Total vcore-seconds taken by all reduce tasks=12572 Total megabyte-seconds taken by all map tasks=438612992 Total megabyte-seconds taken by all reduce tasks=12873728 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=7034 Input split bytes=4641 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=7034 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =51 Failed Shuffles=0 Merged Map outputs=51 GC time elapsed (ms)=3692 CPU time spent (ms)=37550 Physical memory (bytes) snapshot=13782253568 Virtual memory (bytes) snapshot=52130512896 Total committed heap usage (bytes)=10468982784 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=135492 File Output Format Counters Bytes Written=96 17/02/07 10:19:29 INFO streaming.StreamJob: Output directory: vanilla_7
8	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=50 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_8 17/02/07 10:22:20 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps packageJobJar: [/tmp/hadoop-unjar1872207383042615174/] [] /tmp/streamjob2246427828326528476.jar tmpDir=null 17/02/07 10:22:23 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:22:24 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:22:25 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:22:26 INFO mapreduce.JobSubmitter: number of splits:51 17/02/07 10:22:27 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0266

```
17/02/07 10:22:27 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0266
17/02/07 10:22:28 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0266/
17/02/07 10:22:28 INFO mapreduce.Job: Running job: job_1484631414223_0266
17/02/07 10:22:42 INFO mapreduce.Job: Job job_1484631414223_0266 running in uber mode : false
17/02/07 10:22:42 INFO mapreduce.Job:  map 0% reduce 0%
17/02/07 10:22:51 INFO mapreduce.Job:  map 2% reduce 0%
17/02/07 10:22:52 INFO mapreduce.Job:  map 8% reduce 0%
17/02/07 10:22:53 INFO mapreduce.Job:  map 22% reduce 0%
17/02/07 10:22:54 INFO mapreduce.Job:  map 29% reduce 0%
17/02/07 10:22:55 INFO mapreduce.Job:  map 39% reduce 0%
17/02/07 10:22:56 INFO mapreduce.Job:  map 59% reduce 0%
17/02/07 10:22:57 INFO mapreduce.Job:  map 65% reduce 0%
17/02/07 10:22:58 INFO mapreduce.Job:  map 78% reduce 0%
17/02/07 10:22:59 INFO mapreduce.Job:  map 80% reduce 0%
17/02/07 10:23:01 INFO mapreduce.Job:  map 84% reduce 0%
17/02/07 10:23:02 INFO mapreduce.Job:  map 92% reduce 0%
17/02/07 10:23:03 INFO mapreduce.Job:  map 94% reduce 0%
17/02/07 10:23:04 INFO mapreduce.Job:  map 100% reduce 0%
17/02/07 10:23:05 INFO mapreduce.Job:  map 100% reduce 100%
17/02/07 10:23:07 INFO mapreduce.Job: Job job_1484631414223_0266 completed successfully
17/02/07 10:23:07 INFO mapreduce.Job: Counters: 50
    File System Counters
        FILE: Number of bytes read=6734
        FILE: Number of bytes written=5683574
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=140133
        HDFS: Number of bytes written=96
        HDFS: Number of read operations=156
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
    Job Counters
        Launched map tasks=51
        Launched reduce tasks=1
        Data-local map tasks=15
        Rack-local map tasks=36
        Total time spent by all maps in occupied slots (ms)=438559
        Total time spent by all reduces in occupied slots (ms)=11152
        Total time spent by all map tasks (ms)=438559
        Total time spent by all reduce tasks (ms)=11152
        Total vcore-seconds taken by all map tasks=438559
        Total vcore-seconds taken by all reduce tasks=11152
        Total megabyte-seconds taken by all map tasks=449084416
        Total megabyte-seconds taken by all reduce tasks=11419648
    MapReduce Framework
        Map input records=150
        Map output records=150
        Map output bytes=6428
        Map output materialized bytes=7034
        Input split bytes=4641
        Combine input records=0
        Combine output records=0
        Reduce input groups=3
        Reduce shuffle bytes=7034
        Reduce input records=150
        Reduce output records=3
        Spilled Records=300
        Shuffled Maps =51
        Failed Shuffles=0
        Merged Map outputs=51
        GC time elapsed (ms)=3590
        CPU time spent (ms)=37410
        Physical memory (bytes) snapshot=13738942464
        Virtual memory (bytes) snapshot=52116664320
        Total committed heap usage (bytes)=10468982784
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=135492
```

	File Output Format Counters Bytes Written=96 17/02/07 10:23:07 INFO streaming.StreamJob: Output directory: vanilla_8
9	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=50 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_9 17/02/07 10:26:01 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps packageJobJar: [/tmp/hadoop-unjar507229172017285589/] [] /tmp/streamjob2853278273328228968.jar tmpDir=null 17/02/07 10:26:04 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:26:05 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:26:07 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:26:07 INFO mapreduce.JobSubmitter: number of splits:51 17/02/07 10:26:08 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0267 17/02/07 10:26:09 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0267 17/02/07 10:26:09 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0267/ 17/02/07 10:26:09 INFO mapreduce.Job: Running job: job_1484631414223_0267 17/02/07 10:26:23 INFO mapreduce.Job: Job job_1484631414223_0267 running in uber mode : false 17/02/07 10:26:23 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 10:26:32 INFO mapreduce.Job: map 4% reduce 0% 17/02/07 10:26:33 INFO mapreduce.Job: map 6% reduce 0% 17/02/07 10:26:34 INFO mapreduce.Job: map 20% reduce 0% 17/02/07 10:26:35 INFO mapreduce.Job: map 25% reduce 0% 17/02/07 10:26:36 INFO mapreduce.Job: map 47% reduce 0% 17/02/07 10:26:37 INFO mapreduce.Job: map 65% reduce 0% 17/02/07 10:26:38 INFO mapreduce.Job: map 73% reduce 0% 17/02/07 10:26:39 INFO mapreduce.Job: map 78% reduce 0% 17/02/07 10:26:40 INFO mapreduce.Job: map 82% reduce 0% 17/02/07 10:26:41 INFO mapreduce.Job: map 84% reduce 0% 17/02/07 10:26:43 INFO mapreduce.Job: map 86% reduce 0% 17/02/07 10:26:44 INFO mapreduce.Job: map 88% reduce 0% 17/02/07 10:26:45 INFO mapreduce.Job: map 98% reduce 0% 17/02/07 10:26:46 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 10:26:48 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 10:26:49 INFO mapreduce.Job: Job job_1484631414223_0267 completed successfully 17/02/07 10:26:50 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6734 FILE: Number of bytes written=5683574 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=140133 HDFS: Number of bytes written=96 HDFS: Number of read operations=156 HDFS: Number of large read operations=0 HDFS: Number of write operations=2 Job Counters Launched map tasks=51 Launched reduce tasks=1 Data-local map tasks=15 Rack-local map tasks=36 Total time spent by all maps in occupied slots (ms)=435549 Total time spent by all reduces in occupied slots (ms)=12498 Total time spent by all map tasks (ms)=435549 Total time spent by all reduce tasks (ms)=12498 Total vcore-seconds taken by all map tasks=435549 Total vcore-seconds taken by all reduce tasks=12498 Total megabyte-seconds taken by all map tasks=446002176 Total megabyte-seconds taken by all reduce tasks=12797952 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=7034 Input split bytes=4641 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=7034 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =51 Failed Shuffles=0

	Merged Map outputs=51 GC time elapsed (ms)=3829 CPU time spent (ms)=36040 Physical memory (bytes) snapshot=13749518336 Virtual memory (bytes) snapshot=52077764608 Total committed heap usage (bytes)=10468982784 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=135492 File Output Format Counters Bytes Written=96 17/02/07 10:26:50 INFO streaming.StreamJob: Output directory: vanilla_9
10	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=100 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.tx t -output vanilla_10 17/02/07 10:29:29 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps packageJobJar: [/tmp/hadoop-unjar6086947376192107096/] [] /tmp/streamjob8548337041755276841.jar tmpDir=null 17/02/07 10:29:30 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:29:30 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:29:31 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:29:31 INFO mapreduce.JobSubmitter: number of splits:101 17/02/07 10:29:32 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0268 17/02/07 10:29:32 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0268 17/02/07 10:29:32 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0268/ 17/02/07 10:29:32 INFO mapreduce.Job: Running job: job_1484631414223_0268 17/02/07 10:29:37 INFO mapreduce.Job: Job job_1484631414223_0268 running in uber mode : false 17/02/07 10:29:37 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 10:29:41 INFO mapreduce.Job: map 1% reduce 0% 17/02/07 10:29:42 INFO mapreduce.Job: map 2% reduce 0% 17/02/07 10:29:44 INFO mapreduce.Job: map 9% reduce 0% 17/02/07 10:29:45 INFO mapreduce.Job: map 11% reduce 0% 17/02/07 10:29:46 INFO mapreduce.Job: map 19% reduce 0% 17/02/07 10:29:47 INFO mapreduce.Job: map 24% reduce 0% 17/02/07 10:29:48 INFO mapreduce.Job: map 31% reduce 0% 17/02/07 10:29:49 INFO mapreduce.Job: map 38% reduce 0% 17/02/07 10:29:50 INFO mapreduce.Job: map 44% reduce 0% 17/02/07 10:29:51 INFO mapreduce.Job: map 55% reduce 0% 17/02/07 10:29:52 INFO mapreduce.Job: map 66% reduce 0% 17/02/07 10:29:53 INFO mapreduce.Job: map 77% reduce 0% 17/02/07 10:29:54 INFO mapreduce.Job: map 90% reduce 0% 17/02/07 10:29:55 INFO mapreduce.Job: map 99% reduce 0% 17/02/07 10:29:56 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 10:29:58 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 10:29:58 INFO mapreduce.Job: Job job_1484631414223_0268 completed successfully 17/02/07 10:29:58 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6734 FILE: Number of bytes written=11135879 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=278118 HDFS: Number of bytes written=96 HDFS: Number of read operations=306 HDFS: Number of large read operations=0 HDFS: Number of write operations=2 Job Counters Launched map tasks=101 Launched reduce tasks=1 Data-local map tasks=10 Rack-local map tasks=91 Total time spent by all maps in occupied slots (ms)=752166 Total time spent by all reduces in occupied slots (ms)=11406 Total time spent by all map tasks (ms)=752166 Total time spent by all reduce tasks (ms)=11406 Total vcore-seconds taken by all map tasks=752166 Total vcore-seconds taken by all reduce tasks=11406 Total megabyte-seconds taken by all map tasks=770217984

	Total megabyte-seconds taken by all reduce tasks=11679744 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=7334 Input split bytes=9191 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=7334 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =101 Failed Shuffles=0 Merged Map outputs=101 GC time elapsed (ms)=6813 CPU time spent (ms)=69380 Physical memory (bytes) snapshot=27102998528 Virtual memory (bytes) snapshot=102157180928 Total committed heap usage (bytes)=20535312384 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=268927 File Output Format Counters Bytes Written=96 17/02/07 10:29:58 INFO streaming.StreamJob: Output directory: vanilla_10
11	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=100 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.tx t -output vanilla_11 17/02/07 10:32:31 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps packageJobJar: [/tmp/hadoop-unjar4965630914206129108/] [] /tmp/streamjob6225881589167188160.jar tmpDir=null 17/02/07 10:32:32 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:32:32 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:32:34 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:32:34 INFO mapreduce.JobSubmitter: number of splits:101 17/02/07 10:32:34 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0269 17/02/07 10:32:34 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0269 17/02/07 10:32:34 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0269/ 17/02/07 10:32:34 INFO mapreduce.Job: Running job: job_1484631414223_0269 17/02/07 10:32:42 INFO mapreduce.Job: Job job_1484631414223_0269 running in uber mode : false 17/02/07 10:32:42 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 10:32:46 INFO mapreduce.Job: map 1% reduce 0% 17/02/07 10:32:47 INFO mapreduce.Job: map 2% reduce 0% 17/02/07 10:32:49 INFO mapreduce.Job: map 4% reduce 0% 17/02/07 10:32:50 INFO mapreduce.Job: map 10% reduce 0% 17/02/07 10:32:51 INFO mapreduce.Job: map 13% reduce 0% 17/02/07 10:32:52 INFO mapreduce.Job: map 22% reduce 0% 17/02/07 10:32:53 INFO mapreduce.Job: map 30% reduce 0% 17/02/07 10:32:54 INFO mapreduce.Job: map 37% reduce 0% 17/02/07 10:32:55 INFO mapreduce.Job: map 43% reduce 0% 17/02/07 10:32:56 INFO mapreduce.Job: map 53% reduce 0% 17/02/07 10:32:57 INFO mapreduce.Job: map 68% reduce 0% 17/02/07 10:32:58 INFO mapreduce.Job: map 79% reduce 0% 17/02/07 10:32:59 INFO mapreduce.Job: map 90% reduce 0% 17/02/07 10:33:00 INFO mapreduce.Job: map 98% reduce 0% 17/02/07 10:33:01 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 10:33:02 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 10:33:02 INFO mapreduce.Job: Job job_1484631414223_0269 completed successfully 17/02/07 10:33:02 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6734 FILE: Number of bytes written=11135879 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=278118

	HDFS: Number of bytes written=96 HDFS: Number of read operations=306 HDFS: Number of large read operations=0 HDFS: Number of write operations=2 Job Counters Launched map tasks=101 Launched reduce tasks=1 Data-local map tasks=30 Rack-local map tasks=71 Total time spent by all maps in occupied slots (ms)=761400 Total time spent by all reduces in occupied slots (ms)=8972 Total time spent by all map tasks (ms)=761400 Total time spent by all reduce tasks (ms)=8972 Total vcore-seconds taken by all map tasks=761400 Total vcore-seconds taken by all reduce tasks=8972 Total megabyte-seconds taken by all map tasks=779673600 Total megabyte-seconds taken by all reduce tasks=9187328 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=7334 Input split bytes=9191 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=7334 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =101 Failed Shuffles=0 Merged Map outputs=101 GC time elapsed (ms)=7206 CPU time spent (ms)=68560 Physical memory (bytes) snapshot=27022860288 Virtual memory (bytes) snapshot=102161883136 Total committed heap usage (bytes)=20535312384 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=268927 File Output Format Counters Bytes Written=96 17/02/07 10:33:02 INFO streaming.StreamJob: Output directory: vanilla_11
12	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=100 -mapper mapper_kmeans.py -reducer reducer_kmeans.py - input data.tx t -output vanilla_12 17/02/07 10:35:03 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps packageJobJar: [/tmp/hadoop-unjar9096292514579411096/] [] /tmp/streamjob2410403676479534023.jar tmpDir=null 17/02/07 10:35:04 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:35:04 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:35:05 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:35:06 INFO mapreduce.JobSubmitter: number of splits:101 17/02/07 10:35:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0270 17/02/07 10:35:06 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0270 17/02/07 10:35:06 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0270/ 17/02/07 10:35:06 INFO mapreduce.Job: Running job: job_1484631414223_0270 17/02/07 10:35:14 INFO mapreduce.Job: Job job_1484631414223_0270 running in uber mode : false 17/02/07 10:35:14 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 10:35:19 INFO mapreduce.Job: map 1% reduce 0% 17/02/07 10:35:20 INFO mapreduce.Job: map 2% reduce 0% 17/02/07 10:35:21 INFO mapreduce.Job: map 3% reduce 0% 17/02/07 10:35:23 INFO mapreduce.Job: map 10% reduce 0% 17/02/07 10:35:24 INFO mapreduce.Job: map 15% reduce 0% 17/02/07 10:35:25 INFO mapreduce.Job: map 22% reduce 0% 17/02/07 10:35:26 INFO mapreduce.Job: map 27% reduce 0% 17/02/07 10:35:27 INFO mapreduce.Job: map 35% reduce 0% 17/02/07 10:35:28 INFO mapreduce.Job: map 39% reduce 0%

	17/02/07 10:35:29 INFO mapreduce.Job: map 50% reduce 0% 17/02/07 10:35:30 INFO mapreduce.Job: map 59% reduce 0% 17/02/07 10:35:31 INFO mapreduce.Job: map 72% reduce 0% 17/02/07 10:35:32 INFO mapreduce.Job: map 87% reduce 0% 17/02/07 10:35:33 INFO mapreduce.Job: map 96% reduce 0% 17/02/07 10:35:34 INFO mapreduce.Job: map 99% reduce 0% 17/02/07 10:35:35 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 10:35:36 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 10:35:36 INFO mapreduce.Job: Job job_1484631414223_0270 completed successfully 17/02/07 10:35:36 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6734 FILE: Number of bytes written=11135879 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=278118 HDFS: Number of bytes written=96 HDFS: Number of read operations=306 HDFS: Number of large read operations=0 HDFS: Number of write operations=2 Job Counters Launched map tasks=101 Launched reduce tasks=1 Data-local map tasks=32 Rack-local map tasks=69 Total time spent by all maps in occupied slots (ms)=757624 Total time spent by all reduces in occupied slots (ms)=10297 Total time spent by all map tasks (ms)=757624 Total time spent by all reduce tasks (ms)=10297 Total vcore-seconds taken by all map tasks=757624 Total vcore-seconds taken by all reduce tasks=10297 Total megabyte-seconds taken by all map tasks=775806976 Total megabyte-seconds taken by all reduce tasks=10544128 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=7334 Input split bytes=9191 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=7334 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =101 Failed Shuffles=0 Merged Map outputs=101 GC time elapsed (ms)=7152 CPU time spent (ms)=69650 Physical memory (bytes) snapshot=27085701120 Virtual memory (bytes) snapshot=102239137792 Total committed heap usage (bytes)=20535312384 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=268927 File Output Format Counters Bytes Written=96 17/02/07 10:35:36 INFO streaming.StreamJob: Output directory: vanilla_12
13	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=150 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_13 17/02/07 10:38:05 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps packageJobJar: [/tmp/hadoop-unjar1854590614601538885/] [] /tmp/streamjob1765774783824820053.jar tmpDir=null 17/02/07 10:38:06 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:38:06 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:38:07 INFO mapred.FileInputFormat: Total input paths to process : 1

```
17/02/07 10:38:08 INFO mapreduce.JobSubmitter: number of splits:154
17/02/07 10:38:08 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0271
17/02/07 10:38:08 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0271
17/02/07 10:38:08 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0271/
17/02/07 10:38:08 INFO mapreduce.Job: Running job: job_1484631414223_0271
17/02/07 10:38:15 INFO mapreduce.Job: Job job_1484631414223_0271 running in uber mode : false
17/02/07 10:38:15 INFO mapreduce.Job:  map 0% reduce 0%
17/02/07 10:38:22 INFO mapreduce.Job:  map 1% reduce 0%
17/02/07 10:38:24 INFO mapreduce.Job:  map 2% reduce 0%
17/02/07 10:38:25 INFO mapreduce.Job:  map 4% reduce 0%
17/02/07 10:38:26 INFO mapreduce.Job:  map 8% reduce 0%
17/02/07 10:38:27 INFO mapreduce.Job:  map 16% reduce 0%
17/02/07 10:38:28 INFO mapreduce.Job:  map 19% reduce 0%
17/02/07 10:38:29 INFO mapreduce.Job:  map 25% reduce 0%
17/02/07 10:38:30 INFO mapreduce.Job:  map 29% reduce 0%
17/02/07 10:38:31 INFO mapreduce.Job:  map 36% reduce 0%
17/02/07 10:38:32 INFO mapreduce.Job:  map 42% reduce 0%
17/02/07 10:38:33 INFO mapreduce.Job:  map 46% reduce 0%
17/02/07 10:38:34 INFO mapreduce.Job:  map 52% reduce 0%
17/02/07 10:38:35 INFO mapreduce.Job:  map 58% reduce 0%
17/02/07 10:38:36 INFO mapreduce.Job:  map 62% reduce 0%
17/02/07 10:38:37 INFO mapreduce.Job:  map 68% reduce 0%
17/02/07 10:38:38 INFO mapreduce.Job:  map 74% reduce 0%
17/02/07 10:38:39 INFO mapreduce.Job:  map 79% reduce 0%
17/02/07 10:38:40 INFO mapreduce.Job:  map 87% reduce 0%
17/02/07 10:38:41 INFO mapreduce.Job:  map 93% reduce 0%
17/02/07 10:38:42 INFO mapreduce.Job:  map 99% reduce 29%
17/02/07 10:38:43 INFO mapreduce.Job:  map 100% reduce 29%
17/02/07 10:38:44 INFO mapreduce.Job:  map 100% reduce 100%
17/02/07 10:38:44 INFO mapreduce.Job: Job job_1484631414223_0271 completed successfully
17/02/07 10:38:44 INFO mapreduce.Job: Counters: 50
    File System Counters
        FILE: Number of bytes read=6734
        FILE: Number of bytes written=16915264
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=423398
        HDFS: Number of bytes written=96
        HDFS: Number of read operations=465
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
    Job Counters
        Launched map tasks=154
        Launched reduce tasks=1
        Data-local map tasks=46
        Rack-local map tasks=108
        Total time spent by all maps in occupied slots (ms)=1216460
        Total time spent by all reduces in occupied slots (ms)=15590
        Total time spent by all map tasks (ms)=1216460
        Total time spent by all reduce tasks (ms)=15590
        Total vcore-seconds taken by all map tasks=1216460
        Total vcore-seconds taken by all reduce tasks=15590
        Total megabyte-seconds taken by all map tasks=1245655040
        Total megabyte-seconds taken by all reduce tasks=15964160
    MapReduce Framework
        Map input records=150
        Map output records=150
        Map output bytes=6428
        Map output materialized bytes=7652
        Input split bytes=14014
        Combine input records=0
        Combine output records=0
        Reduce input groups=3
        Reduce shuffle bytes=7652
        Reduce input records=150
        Reduce output records=3
        Spilled Records=300
        Shuffled Maps =154
        Failed Shuffles=0
        Merged Map outputs=154
        GC time elapsed (ms)=10722
        CPU time spent (ms)=108850
        Physical memory (bytes) snapshot=41232044032
        Virtual memory (bytes) snapshot=155279478784
```

	Total committed heap usage (bytes)=31205621760 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=409384 File Output Format Counters Bytes Written=96 17/02/07 10:38:44 INFO streaming.StreamJob: Output directory: vanilla_13
14	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=150 -mapper mapper_kmeans.py -reducer reducer_kmeans.py - input data.tx t -output vanilla_14 17/02/07 10:42:35 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps packageJobJar: [/tmp/hadoop-unjar2617583311978981774/] [] /tmp/streamjob2688950315052375749.jar tmpDir=null 17/02/07 10:42:36 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:42:36 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:42:37 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:42:38 INFO mapreduce.JobSubmitter: number of splits:154 17/02/07 10:42:38 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0274 17/02/07 10:42:38 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0274 17/02/07 10:42:38 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0274/ 17/02/07 10:42:38 INFO mapreduce.Job: Running job: job_1484631414223_0274 17/02/07 10:42:46 INFO mapreduce.Job: Job job_1484631414223_0274 running in uber mode : false 17/02/07 10:42:46 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 10:42:51 INFO mapreduce.Job: map 1% reduce 0% 17/02/07 10:42:53 INFO mapreduce.Job: map 3% reduce 0% 17/02/07 10:42:54 INFO mapreduce.Job: map 6% reduce 0% 17/02/07 10:42:55 INFO mapreduce.Job: map 10% reduce 0% 17/02/07 10:42:56 INFO mapreduce.Job: map 13% reduce 0% 17/02/07 10:42:57 INFO mapreduce.Job: map 17% reduce 0% 17/02/07 10:42:58 INFO mapreduce.Job: map 20% reduce 0% 17/02/07 10:42:59 INFO mapreduce.Job: map 27% reduce 0% 17/02/07 10:43:00 INFO mapreduce.Job: map 32% reduce 0% 17/02/07 10:43:01 INFO mapreduce.Job: map 39% reduce 0% 17/02/07 10:43:02 INFO mapreduce.Job: map 43% reduce 0% 17/02/07 10:43:03 INFO mapreduce.Job: map 52% reduce 0% 17/02/07 10:43:04 INFO mapreduce.Job: map 56% reduce 0% 17/02/07 10:43:05 INFO mapreduce.Job: map 60% reduce 0% 17/02/07 10:43:06 INFO mapreduce.Job: map 66% reduce 0% 17/02/07 10:43:07 INFO mapreduce.Job: map 71% reduce 0% 17/02/07 10:43:08 INFO mapreduce.Job: map 75% reduce 23% 17/02/07 10:43:09 INFO mapreduce.Job: map 81% reduce 23% 17/02/07 10:43:10 INFO mapreduce.Job: map 86% reduce 23% 17/02/07 10:43:11 INFO mapreduce.Job: map 96% reduce 27% 17/02/07 10:43:12 INFO mapreduce.Job: map 100% reduce 27% 17/02/07 10:43:13 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 10:43:14 INFO mapreduce.Job: Job job_1484631414223_0274 completed successfully 17/02/07 10:43:14 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6734 FILE: Number of bytes written=16915264 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=423398 HDFS: Number of bytes written=96 HDFS: Number of read operations=465 HDFS: Number of large read operations=0 HDFS: Number of write operations=2 Job Counters Launched map tasks=154 Launched reduce tasks=1 Data-local map tasks=46 Rack-local map tasks=108 Total time spent by all maps in occupied slots (ms)=1214005 Total time spent by all reduces in occupied slots (ms)=16754 Total time spent by all map tasks (ms)=1214005 Total time spent by all reduce tasks (ms)=16754 Total vcore-seconds taken by all map tasks=1214005 Total vcore-seconds taken by all reduce tasks=16754

	Total megabyte-seconds taken by all map tasks=1243141120 Total megabyte-seconds taken by all reduce tasks=17156096 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=7652 Input split bytes=14014 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=7652 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =154 Failed Shuffles=0 Merged Map outputs=154 GC time elapsed (ms)=10694 CPU time spent (ms)=108590 Physical memory (bytes) snapshot=41290731520 Virtual memory (bytes) snapshot=155411431424 Total committed heap usage (bytes)=31205621760 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=409384 File Output Format Counters Bytes Written=96 17/02/07 10:43:14 INFO streaming.StreamJob: Output directory: vanilla_14
15	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=150 -mapper mapper_kmeans.py -reducer reducer_kmeans.py - input data.tx t -output vanilla_15 17/02/07 10:44:54 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps packageJobJar: [/tmp/hadoop-unjar7719427521441820447] [] /tmp/streamjob3150122714179887281.jar tmpDir=null 17/02/07 10:44:55 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:44:56 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:44:57 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:44:57 INFO mapreduce.JobSubmitter: number of splits:154 17/02/07 10:44:58 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0275 17/02/07 10:44:58 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0275 17/02/07 10:44:58 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0275/ 17/02/07 10:44:58 INFO mapreduce.Job: Running job: job_1484631414223_0275 17/02/07 10:45:05 INFO mapreduce.Job: Job job_1484631414223_0275 running in uber mode : false 17/02/07 10:45:05 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 10:45:09 INFO mapreduce.Job: map 1% reduce 0% 17/02/07 10:45:12 INFO mapreduce.Job: map 2% reduce 0% 17/02/07 10:45:14 INFO mapreduce.Job: map 6% reduce 0% 17/02/07 10:45:15 INFO mapreduce.Job: map 10% reduce 0% 17/02/07 10:45:16 INFO mapreduce.Job: map 16% reduce 0% 17/02/07 10:45:17 INFO mapreduce.Job: map 21% reduce 0% 17/02/07 10:45:18 INFO mapreduce.Job: map 23% reduce 0% 17/02/07 10:45:19 INFO mapreduce.Job: map 29% reduce 0% 17/02/07 10:45:20 INFO mapreduce.Job: map 34% reduce 0% 17/02/07 10:45:21 INFO mapreduce.Job: map 40% reduce 0% 17/02/07 10:45:22 INFO mapreduce.Job: map 45% reduce 0% 17/02/07 10:45:23 INFO mapreduce.Job: map 53% reduce 0% 17/02/07 10:45:24 INFO mapreduce.Job: map 56% reduce 0% 17/02/07 10:45:25 INFO mapreduce.Job: map 62% reduce 0% 17/02/07 10:45:26 INFO mapreduce.Job: map 67% reduce 0% 17/02/07 10:45:27 INFO mapreduce.Job: map 73% reduce 23% 17/02/07 10:45:28 INFO mapreduce.Job: map 80% reduce 23% 17/02/07 10:45:29 INFO mapreduce.Job: map 84% reduce 23% 17/02/07 10:45:30 INFO mapreduce.Job: map 94% reduce 29% 17/02/07 10:45:31 INFO mapreduce.Job: map 99% reduce 29% 17/02/07 10:45:32 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 10:45:33 INFO mapreduce.Job: Job job_1484631414223_0275 completed successfully 17/02/07 10:45:33 INFO mapreduce.Job: Counters: 50 File System Counters

	FILE: Number of bytes read=6734 FILE: Number of bytes written=16915264 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=423398 HDFS: Number of bytes written=96 HDFS: Number of read operations=465 HDFS: Number of large read operations=0 HDFS: Number of write operations=2 Job Counters Launched map tasks=154 Launched reduce tasks=1 Data-local map tasks=46 Rack-local map tasks=108 Total time spent by all maps in occupied slots (ms)=1204461 Total time spent by all reduces in occupied slots (ms)=16235 Total time spent by all map tasks (ms)=1204461 Total time spent by all reduce tasks (ms)=16235 Total vcore-seconds taken by all map tasks=1204461 Total vcore-seconds taken by all reduce tasks=16235 Total megabyte-seconds taken by all map tasks=1233368064 Total megabyte-seconds taken by all reduce tasks=16624640 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=7652 Input split bytes=14014 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=7652 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =154 Failed Shuffles=0 Merged Map outputs=154 GC time elapsed (ms)=10042 CPU time spent (ms)=106620 Physical memory (bytes) snapshot=41271472128 Virtual memory (bytes) snapshot=155324547072 Total committed heap usage (bytes)=31205621760 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=409384 File Output Format Counters Bytes Written=96 17/02/07 10:45:33 INFO streaming.StreamJob: Output directory: vanilla_15
16	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=2 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_16 17/02/07 10:52:16 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 10:52:16 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar5468235695749898444/] [] /tmp/streamjob6386798246551011486.jar tmpDir=null 17/02/07 10:52:17 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:52:17 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:52:18 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:52:19 INFO mapreduce.JobSubmitter: number of splits:2 17/02/07 10:52:19 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0279 17/02/07 10:52:19 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0279 17/02/07 10:52:19 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0279/ 17/02/07 10:52:19 INFO mapreduce.Job: Running job: job_1484631414223_0279 17/02/07 10:52:25 INFO mapreduce.Job: Job job_1484631414223_0279 running in uber mode : false 17/02/07 10:52:25 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 10:52:29 INFO mapreduce.Job: map 50% reduce 0% 17/02/07 10:52:31 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 10:52:37 INFO mapreduce.Job: map 100% reduce 67%

	17/02/07 10:52:38 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 10:52:38 INFO mapreduce.Job: Job job_1484631414223_0279 completed successfully 17/02/07 10:52:38 INFO mapreduce.Job: Counters: 49 File System Counters FILE: Number of bytes read=6746 FILE: Number of bytes written=558446 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=7192 HDFS: Number of bytes written=96 HDFS: Number of read operations=15 HDFS: Number of large read operations=0 HDFS: Number of write operations=6 Job Counters Launched map tasks=2 Launched reduce tasks=3 Rack-local map tasks=2 Total time spent by all maps in occupied slots (ms)=6145 Total time spent by all reduces in occupied slots (ms)=13989 Total time spent by all map tasks (ms)=6145 Total time spent by all reduce tasks (ms)=13989 Total vcore-seconds taken by all map tasks=6145 Total vcore-seconds taken by all reduce tasks=13989 Total megabyte-seconds taken by all map tasks=6292480 Total megabyte-seconds taken by all reduce tasks=14324736 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=6764 Input split bytes=182 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=6764 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =6 Failed Shuffles=0 Merged Map outputs=6 GC time elapsed (ms)=281 CPU time spent (ms)=5600 Physical memory (bytes) snapshot=1041047552 Virtual memory (bytes) snapshot=4993167360 Total committed heap usage (bytes)=1006632960 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=7010 File Output Format Counters Bytes Written=96 17/02/07 10:52:38 INFO streaming.StreamJob: Output directory: f vanilla_16
17	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=2 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer_ kmeans.py -input data.txt -output vanilla_17 17/02/07 10:55:51 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 10:55:51 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar3194829060474950188/] [] /tmp/streamjob1104822289088609755.jar tmpDir=null 17/02/07 10:55:52 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:55:52 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:55:53 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:55:53 INFO mapreduce.JobSubmitter: number of splits:2 17/02/07 10:55:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0281 17/02/07 10:55:54 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0281 17/02/07 10:55:54 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0281/ 17/02/07 10:55:54 INFO mapreduce.Job: Running job: job_1484631414223_0281 17/02/07 10:56:02 INFO mapreduce.Job: Job job_1484631414223_0281 running in uber mode : false

	17/02/07 10:56:02 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 10:56:08 INFO mapreduce.Job: map 50% reduce 0% 17/02/07 10:56:09 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 10:56:15 INFO mapreduce.Job: map 100% reduce 33% 17/02/07 10:56:16 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 10:56:16 INFO mapreduce.Job: Job job_1484631414223_0281 completed successfully 17/02/07 10:56:17 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6746 FILE: Number of bytes written=558441 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=7192 HDFS: Number of bytes written=96 HDFS: Number of read operations=15 HDFS: Number of large read operations=0 HDFS: Number of write operations=6 Job Counters Launched map tasks=2 Launched reduce tasks=3 Data-local map tasks=1 Rack-local map tasks=1 Total time spent by all maps in occupied slots (ms)=8955 Total time spent by all reduces in occupied slots (ms)=14094 Total time spent by all map tasks (ms)=8955 Total time spent by all reduce tasks (ms)=14094 Total vcore-seconds taken by all map tasks=8955 Total vcore-seconds taken by all reduce tasks=14094 Total megabyte-seconds taken by all map tasks=9169920 Total megabyte-seconds taken by all reduce tasks=14432256 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=6764 Input split bytes=182 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=6764 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =6 Failed Shuffles=0 Merged Map outputs=6 GC time elapsed (ms)=463 CPU time spent (ms)=5660 Physical memory (bytes) snapshot=1041649664 Virtual memory (bytes) snapshot=5025718272 Total committed heap usage (bytes)=1006632960 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=7010 File Output Format Counters Bytes Written=96 17/02/07 10:56:17 INFO streaming.StreamJob: Output directory: vanilla_17
18	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=2 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer_ kmeans.py -input data.txt -output vanilla_18 17/02/07 10:58:32 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 10:58:32 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar4348829281393311363/] [] /tmp/streamjob6290854078675142857.jar tmpDir=null 17/02/07 10:58:33 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:58:34 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 10:58:35 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 10:58:35 INFO mapreduce.JobSubmitter: number of splits:2

	17/02/07 10:58:35 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0282 17/02/07 10:58:36 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0282 17/02/07 10:58:36 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0282/ 17/02/07 10:58:36 INFO mapreduce.Job: Running job: job_1484631414223_0282 17/02/07 10:58:44 INFO mapreduce.Job: Job job_1484631414223_0282 running in uber mode : false 17/02/07 10:58:44 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 10:58:50 INFO mapreduce.Job: map 50% reduce 0% 17/02/07 10:58:52 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 10:58:56 INFO mapreduce.Job: map 100% reduce 33% 17/02/07 10:58:58 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 10:58:59 INFO mapreduce.Job: Job job_1484631414223_0282 completed successfully 17/02/07 10:58:59 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6746 FILE: Number of bytes written=558441 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=7192 HDFS: Number of bytes written=96 HDFS: Number of read operations=15 HDFS: Number of large read operations=0 HDFS: Number of write operations=6 Job Counters Launched map tasks=2 Launched reduce tasks=3 Data-local map tasks=1 Rack-local map tasks=1 Total time spent by all maps in occupied slots (ms)=9188 Total time spent by all reduces in occupied slots (ms)=12021 Total time spent by all map tasks (ms)=9188 Total time spent by all reduce tasks (ms)=12021 Total vcore-seconds taken by all map tasks=9188 Total vcore-seconds taken by all reduce tasks=12021 Total megabyte-seconds taken by all map tasks=9408512 Total megabyte-seconds taken by all reduce tasks=12309504 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=6764 Input split bytes=182 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=6764 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =6 Failed Shuffles=0 Merged Map outputs=6 GC time elapsed (ms)=298 CPU time spent (ms)=4980 Physical memory (bytes) snapshot=1041178624 Virtual memory (bytes) snapshot=5008785408 Total committed heap usage (bytes)=1006632960 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=7010 File Output Format Counters Bytes Written=96 17/02/07 10:58:59 INFO streaming.StreamJob: Output directory: vanilla_18
19	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=20 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer _kmeans.py -input data.txt -output vanilla_19 17/02/07 11:02:44 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 11:02:44 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces

```
packageJobJar: [/tmp/hadoop-unjar6007957234395849530/] [] /tmp/streamjob7283961400827419783.jar tmpDir=null
17/02/07 11:02:45 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032
17/02/07 11:02:46 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032
17/02/07 11:02:47 INFO mapred.FileInputFormat: Total input paths to process : 1
17/02/07 11:02:47 INFO mapreduce.JobSubmitter: number of splits:20
17/02/07 11:02:47 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0284
17/02/07 11:02:48 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0284
17/02/07 11:02:48 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0284/
17/02/07 11:02:48 INFO mapreduce.Job: Running job: job_1484631414223_0284
17/02/07 11:02:57 INFO mapreduce.Job: Job job_1484631414223_0284 running in uber mode : false
17/02/07 11:02:57 INFO mapreduce.Job:  map 0% reduce 0%
17/02/07 11:03:02 INFO mapreduce.Job:  map 5% reduce 0%
17/02/07 11:03:03 INFO mapreduce.Job:  map 10% reduce 0%
17/02/07 11:03:04 INFO mapreduce.Job:  map 25% reduce 0%
17/02/07 11:03:05 INFO mapreduce.Job:  map 60% reduce 0%
17/02/07 11:03:06 INFO mapreduce.Job:  map 90% reduce 0%
17/02/07 11:03:07 INFO mapreduce.Job:  map 100% reduce 0%
17/02/07 11:03:10 INFO mapreduce.Job:  map 100% reduce 100%
17/02/07 11:03:11 INFO mapreduce.Job: Job job_1484631414223_0284 completed successfully
17/02/07 11:03:11 INFO mapreduce.Job: Counters: 50
    File System Counters
        FILE: Number of bytes read=6746
        FILE: Number of bytes written=2521662
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=57049
        HDFS: Number of bytes written=96
        HDFS: Number of read operations=69
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=6
    Job Counters
        Launched map tasks=20
        Launched reduce tasks=3
        Data-local map tasks=6
        Rack-local map tasks=14
        Total time spent by all maps in occupied slots (ms)=102710
        Total time spent by all reduces in occupied slots (ms)=13776
        Total time spent by all map tasks (ms)=102710
        Total time spent by all reduce tasks (ms)=13776
        Total vcore-seconds taken by all map tasks=102710
        Total vcore-seconds taken by all reduce tasks=13776
        Total megabyte-seconds taken by all map tasks=105175040
        Total megabyte-seconds taken by all reduce tasks=14106624
    MapReduce Framework
        Map input records=150
        Map output records=150
        Map output bytes=6428
        Map output materialized bytes=7088
        Input split bytes=1820
        Combine input records=0
        Combine output records=0
        Reduce input groups=3
        Reduce shuffle bytes=7088
        Reduce input records=150
        Reduce output records=3
        Spilled Records=300
        Shuffled Maps =60
        Failed Shuffles=0
        Merged Map outputs=60
        GC time elapsed (ms)=1705
        CPU time spent (ms)=17720
        Physical memory (bytes) snapshot=5841137664
        Virtual memory (bytes) snapshot=23057002496
        Total committed heap usage (bytes)=4630511616
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=55229
    File Output Format Counters
```

	Bytes Written=96 17/02/07 11:03:11 INFO streaming.StreamJob: Output directory: vanilla_19
20	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=20 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer _kmeans.py -input data.txt -output vanilla_20 17/02/07 11:49:53 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 11:49:53 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar1572058830967147068/] [] /tmp/streamjob7271727448159438712.jar tmpDir=null 17/02/07 11:49:54 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 11:49:55 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 11:49:56 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 11:49:56 INFO mapreduce.JobSubmitter: number of splits:20 17/02/07 11:49:56 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0294 17/02/07 11:49:57 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0294 17/02/07 11:49:57 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0294/ 17/02/07 11:49:57 INFO mapreduce.Job: Running job: job_1484631414223_0294 17/02/07 11:50:04 INFO mapreduce.Job: Job job_1484631414223_0294 running in uber mode : false 17/02/07 11:50:04 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 11:50:09 INFO mapreduce.Job: map 5% reduce 0% 17/02/07 11:50:10 INFO mapreduce.Job: map 10% reduce 0% 17/02/07 11:50:11 INFO mapreduce.Job: map 20% reduce 0% 17/02/07 11:50:12 INFO mapreduce.Job: map 75% reduce 0% 17/02/07 11:50:13 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 11:50:16 INFO mapreduce.Job: map 100% reduce 67% 17/02/07 11:50:18 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 11:50:18 INFO mapreduce.Job: Job job_1484631414223_0294 completed successfully 17/02/07 11:50:18 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6746 FILE: Number of bytes written=2521662 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=57049 HDFS: Number of bytes written=96 HDFS: Number of read operations=69 HDFS: Number of large read operations=0 HDFS: Number of write operations=6 Job Counters Launched map tasks=20 Launched reduce tasks=3 Data-local map tasks=6 Rack-local map tasks=14 Total time spent by all maps in occupied slots (ms)=101140 Total time spent by all reduces in occupied slots (ms)=16430 Total time spent by all map tasks (ms)=101140 Total time spent by all reduce tasks (ms)=16430 Total vcore-seconds taken by all map tasks=101140 Total vcore-seconds taken by all reduce tasks=16430 Total megabyte-seconds taken by all map tasks=103567360 Total megabyte-seconds taken by all reduce tasks=16824320 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=7088 Input split bytes=1820 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=7088 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =60 Failed Shuffles=0 Merged Map outputs=60 GC time elapsed (ms)=1304 CPU time spent (ms)=17560 Physical memory (bytes) snapshot=5829816320 Virtual memory (bytes) snapshot=23082905600 Total committed heap usage (bytes)=4630511616 Shuffle Errors BAD_ID=0

	CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=55229 File Output Format Counters Bytes Written=96 17/02/07 11:50:18 INFO streaming.StreamJob: Output directory: vanilla_20
21	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=20 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer _kmeans.py -input data.txt -output vanilla_21 17/02/07 11:51:51 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 11:51:51 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar2200670503082847981/] [] /tmp/streamjob675806489607298077.jar tmpDir=null 17/02/07 11:51:52 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 11:51:52 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 11:51:53 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 11:51:54 INFO mapreduce.JobSubmitter: number of splits:20 17/02/07 11:51:54 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0296 17/02/07 11:51:54 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0296 17/02/07 11:51:54 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0296/ 17/02/07 11:51:54 INFO mapreduce.Job: Running job: job_1484631414223_0296 17/02/07 11:52:01 INFO mapreduce.Job: Job job_1484631414223_0296 running in uber mode : false 17/02/07 11:52:01 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 11:52:06 INFO mapreduce.Job: map 10% reduce 0% 17/02/07 11:52:07 INFO mapreduce.Job: map 15% reduce 0% 17/02/07 11:52:08 INFO mapreduce.Job: map 60% reduce 0% 17/02/07 11:52:09 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 11:52:14 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 11:52:15 INFO mapreduce.Job: Job job_1484631414223_0296 completed successfully 17/02/07 11:52:15 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6746 FILE: Number of bytes written=2521639 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=57049 HDFS: Number of bytes written=96 HDFS: Number of read operations=69 HDFS: Number of large read operations=0 HDFS: Number of write operations=6 Job Counters Launched map tasks=20 Launched reduce tasks=3 Data-local map tasks=6 Rack-local map tasks=14 Total time spent by all maps in occupied slots (ms)=99916 Total time spent by all reduces in occupied slots (ms)=13710 Total time spent by all map tasks (ms)=99916 Total time spent by all reduce tasks (ms)=13710 Total vcore-seconds taken by all map tasks=99916 Total vcore-seconds taken by all reduce tasks=13710 Total megabyte-seconds taken by all map tasks=102313984 Total megabyte-seconds taken by all reduce tasks=14039040 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=7088 Input split bytes=1820 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=7088 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =60 Failed Shuffles=0 Merged Map outputs=60 GC time elapsed (ms)=1267

	CPU time spent (ms)=17720 Physical memory (bytes) snapshot=5839773696 Virtual memory (bytes) snapshot=23070302208 Total committed heap usage (bytes)=4630511616 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=55229 File Output Format Counters Bytes Written=96 17/02/07 11:52:15 INFO streaming.StreamJob: Output directory: vanilla_21
22	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=50 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_22 17/02/07 11:58:27 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 11:58:27 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar8318037824710400720/] [] /tmp/streamjob4361609476029949222.jar tmpDir=null 17/02/07 11:58:28 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 11:58:28 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 11:58:29 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 11:58:29 INFO mapreduce.JobSubmitter: number of splits:51 17/02/07 11:58:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0299 17/02/07 11:58:30 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0299 17/02/07 11:58:30 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0299/ 17/02/07 11:58:30 INFO mapreduce.Job: Running job: job_1484631414223_0299 17/02/07 11:58:37 INFO mapreduce.Job: Job job_1484631414223_0299 running in uber mode : false 17/02/07 11:58:37 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 11:58:42 INFO mapreduce.Job: map 2% reduce 0% 17/02/07 11:58:43 INFO mapreduce.Job: map 4% reduce 0% 17/02/07 11:58:44 INFO mapreduce.Job: map 6% reduce 0% 17/02/07 11:58:45 INFO mapreduce.Job: map 20% reduce 0% 17/02/07 11:58:46 INFO mapreduce.Job: map 27% reduce 0% 17/02/07 11:58:47 INFO mapreduce.Job: map 47% reduce 0% 17/02/07 11:58:48 INFO mapreduce.Job: map 59% reduce 0% 17/02/07 11:58:49 INFO mapreduce.Job: map 90% reduce 0% 17/02/07 11:58:50 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 11:58:51 INFO mapreduce.Job: map 100% reduce 33% 17/02/07 11:58:53 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 11:58:53 INFO mapreduce.Job: Job job_1484631414223_0299 completed successfully 17/02/07 11:58:53 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6746 FILE: Number of bytes written=5902770 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=140133 HDFS: Number of bytes written=96 HDFS: Number of read operations=162 HDFS: Number of large read operations=0 HDFS: Number of write operations=6 Job Counters Launched map tasks=51 Launched reduce tasks=3 Data-local map tasks=16 Rack-local map tasks=35 Total time spent by all maps in occupied slots (ms)=329279 Total time spent by all reduces in occupied slots (ms)=16235 Total time spent by all map tasks (ms)=329279 Total time spent by all reduce tasks (ms)=16235 Total vcore-seconds taken by all map tasks=329279 Total vcore-seconds taken by all reduce tasks=16235 Total megabyte-seconds taken by all map tasks=337181696 Total megabyte-seconds taken by all reduce tasks=16624640 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=7646 Input split bytes=4641

	Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=7646 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =153 Failed Shuffles=0 Merged Map outputs=153 GC time elapsed (ms)=3729 CPU time spent (ms)=38550 Physical memory (bytes) snapshot=14134317056 Virtual memory (bytes) snapshot=54185570304 Total committed heap usage (bytes)=10871635968 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=135492 File Output Format Counters Bytes Written=96 17/02/07 11:58:53 INFO streaming.StreamJob: Output directory: vanilla_22
23	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=50 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_22 17/02/07 12:01:47 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 12:01:47 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar8897203717535713254/] [] /tmp/streamjob2348708624847746241.jar tmpDir=null 17/02/07 12:01:48 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:01:48 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:01:48 ERROR streaming.StreamJob: Error Launching job : Output directory hdfs://dsm1:9000/user/fmare001/vanilla_22 already exists Streaming Command Failed! [fmare001@dsm6 ~]$ hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=50 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer _kmeans.py -input data.txt -output vanilla_23 17/02/07 12:01:53 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 12:01:53 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar2987494775902704926/] [] /tmp/streamjob506852496107395173.jar tmpDir=null 17/02/07 12:01:54 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:01:54 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:01:55 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 12:01:56 INFO mapreduce.JobSubmitter: number of splits:51 17/02/07 12:01:56 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0300 17/02/07 12:01:56 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0300 17/02/07 12:01:56 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0300/ 17/02/07 12:01:56 INFO mapreduce.Job: Running job: job_1484631414223_0300 17/02/07 12:02:04 INFO mapreduce.Job: Job job_1484631414223_0300 running in uber mode : false 17/02/07 12:02:04 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 12:02:10 INFO mapreduce.Job: map 4% reduce 0% 17/02/07 12:02:11 INFO mapreduce.Job: map 6% reduce 0% 17/02/07 12:02:12 INFO mapreduce.Job: map 16% reduce 0% 17/02/07 12:02:13 INFO mapreduce.Job: map 29% reduce 0% 17/02/07 12:02:14 INFO mapreduce.Job: map 41% reduce 0% 17/02/07 12:02:15 INFO mapreduce.Job: map 57% reduce 0% 17/02/07 12:02:16 INFO mapreduce.Job: map 84% reduce 0% 17/02/07 12:02:17 INFO mapreduce.Job: map 96% reduce 0% 17/02/07 12:02:18 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 12:02:19 INFO mapreduce.Job: map 100% reduce 67% 17/02/07 12:02:20 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 12:02:20 INFO mapreduce.Job: Job job_1484631414223_0300 completed successfully 17/02/07 12:02:20 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6746 FILE: Number of bytes written=5902716 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=140133

	HDFS: Number of bytes written=96 HDFS: Number of read operations=162 HDFS: Number of large read operations=0 HDFS: Number of write operations=6 Job Counters Launched map tasks=51 Launched reduce tasks=3 Data-local map tasks=16 Rack-local map tasks=35 Total time spent by all maps in occupied slots (ms)=328591 Total time spent by all reduces in occupied slots (ms)=16184 Total time spent by all map tasks (ms)=328591 Total time spent by all reduce tasks (ms)=16184 Total vcore-seconds taken by all map tasks=328591 Total vcore-seconds taken by all reduce tasks=16184 Total megabyte-seconds taken by all map tasks=336477184 Total megabyte-seconds taken by all reduce tasks=16572416 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=7646 Input split bytes=4641 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=7646 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =153 Failed Shuffles=0 Merged Map outputs=153 GC time elapsed (ms)=3923 CPU time spent (ms)=38930 Physical memory (bytes) snapshot=14036598784 Virtual memory (bytes) snapshot=54085140480 Total committed heap usage (bytes)=10871635968 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=135492 File Output Format Counters Bytes Written=96 17/02/07 12:02:20 INFO streaming.StreamJob: Output directory: vanilla_23
24	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=50 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer _kmeans.py -input data.txt -output vanilla_24 17/02/07 12:04:04 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 12:04:04 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar3406654966286405136/] [] /tmp/streamjob7640639391458261279.jar tmpDir=null 17/02/07 12:04:05 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:04:05 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:04:07 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 12:04:07 INFO mapreduce.JobSubmitter: number of splits:51 17/02/07 12:04:08 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0301 17/02/07 12:04:08 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0301 17/02/07 12:04:08 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0301/ 17/02/07 12:04:08 INFO mapreduce.Job: Running job: job_1484631414223_0301 17/02/07 12:04:15 INFO mapreduce.Job: Job job_1484631414223_0301 running in uber mode : false 17/02/07 12:04:15 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 12:04:20 INFO mapreduce.Job: map 2% reduce 0% 17/02/07 12:04:21 INFO mapreduce.Job: map 4% reduce 0% 17/02/07 12:04:22 INFO mapreduce.Job: map 6% reduce 0% 17/02/07 12:04:23 INFO mapreduce.Job: map 18% reduce 0% 17/02/07 12:04:24 INFO mapreduce.Job: map 33% reduce 0% 17/02/07 12:04:25 INFO mapreduce.Job: map 49% reduce 0% 17/02/07 12:04:26 INFO mapreduce.Job: map 65% reduce 0% 17/02/07 12:04:27 INFO mapreduce.Job: map 90% reduce 0%

	17/02/07 12:04:28 INFO mapreduce.Job: map 96% reduce 0% 17/02/07 12:04:29 INFO mapreduce.Job: map 98% reduce 0% 17/02/07 12:04:30 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 12:04:31 INFO mapreduce.Job: map 100% reduce 67% 17/02/07 12:04:32 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 12:04:33 INFO mapreduce.Job: Job job_1484631414223_0301 completed successfully 17/02/07 12:04:33 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6746 FILE: Number of bytes written=5902770 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=140133 HDFS: Number of bytes written=96 HDFS: Number of read operations=162 HDFS: Number of large read operations=0 HDFS: Number of write operations=6 Job Counters Launched map tasks=51 Launched reduce tasks=3 Data-local map tasks=17 Rack-local map tasks=35 Total time spent by all maps in occupied slots (ms)=331556 Total time spent by all reduces in occupied slots (ms)=20335 Total time spent by all map tasks (ms)=331556 Total time spent by all reduce tasks (ms)=20335 Total vcore-seconds taken by all map tasks=331556 Total vcore-seconds taken by all reduce tasks=20335 Total megabyte-seconds taken by all map tasks=339513344 Total megabyte-seconds taken by all reduce tasks=20823040 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=7646 Input split bytes=4641 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=7646 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =153 Failed Shuffles=0 Merged Map outputs=153 GC time elapsed (ms)=3892 CPU time spent (ms)=39460 Physical memory (bytes) snapshot=14094266368 Virtual memory (bytes) snapshot=54125899776 Total committed heap usage (bytes)=10871635968 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=135492 File Output Format Counters Bytes Written=96 17/02/07 12:04:33 INFO streaming.StreamJob: Output directory: vanilla_24
25	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=100 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py - reducer reduce r_kmeans.py -input data.txt -output vanilla_25 17/02/07 12:07:20 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 12:07:20 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar5502499174451860924/] [] /tmp/streamjob7173527383409939920.jar tmpDir=null 17/02/07 12:07:21 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:07:22 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:07:23 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 12:07:23 INFO mapreduce.JobSubmitter: number of splits:101

```
17/02/07 12:07:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0302
17/02/07 12:07:24 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0302
17/02/07 12:07:24 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0302/
17/02/07 12:07:24 INFO mapreduce.Job: Running job: job_1484631414223_0302
17/02/07 12:07:32 INFO mapreduce.Job: Job job_1484631414223_0302 running in uber mode : false
17/02/07 12:07:32 INFO mapreduce.Job:  map 0% reduce 0%
17/02/07 12:07:37 INFO mapreduce.Job:  map 1% reduce 0%
17/02/07 12:07:38 INFO mapreduce.Job:  map 2% reduce 0%
17/02/07 12:07:40 INFO mapreduce.Job:  map 9% reduce 0%
17/02/07 12:07:41 INFO mapreduce.Job:  map 13% reduce 0%
17/02/07 12:07:42 INFO mapreduce.Job:  map 19% reduce 0%
17/02/07 12:07:43 INFO mapreduce.Job:  map 24% reduce 0%
17/02/07 12:07:44 INFO mapreduce.Job:  map 35% reduce 0%
17/02/07 12:07:45 INFO mapreduce.Job:  map 42% reduce 0%
17/02/07 12:07:46 INFO mapreduce.Job:  map 50% reduce 0%
17/02/07 12:07:47 INFO mapreduce.Job:  map 60% reduce 0%
17/02/07 12:07:48 INFO mapreduce.Job:  map 71% reduce 0%
17/02/07 12:07:49 INFO mapreduce.Job:  map 84% reduce 0%
17/02/07 12:07:50 INFO mapreduce.Job:  map 93% reduce 0%
17/02/07 12:07:51 INFO mapreduce.Job:  map 100% reduce 0%
17/02/07 12:07:52 INFO mapreduce.Job:  map 100% reduce 33%
17/02/07 12:07:53 INFO mapreduce.Job:  map 100% reduce 100%
17/02/07 12:07:54 INFO mapreduce.Job: Job job_1484631414223_0302 completed successfully
17/02/07 12:07:54 INFO mapreduce.Job: Counters: 50
    File System Counters
        FILE: Number of bytes read=6746
        FILE: Number of bytes written=11356275
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=278118
        HDFS: Number of bytes written=96
        HDFS: Number of read operations=312
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=6
    Job Counters
        Launched map tasks=101
        Launched reduce tasks=3
        Data-local map tasks=29
        Rack-local map tasks=72
        Total time spent by all maps in occupied slots (ms)=769709
        Total time spent by all reduces in occupied slots (ms)=30217
        Total time spent by all map tasks (ms)=769709
        Total time spent by all reduce tasks (ms)=30217
        Total vcore-seconds taken by all map tasks=769709
        Total vcore-seconds taken by all reduce tasks=30217
        Total megabyte-seconds taken by all map tasks=788182016
        Total megabyte-seconds taken by all reduce tasks=30942208
    MapReduce Framework
        Map input records=150
        Map output records=150
        Map output bytes=6428
        Map output materialized bytes=8546
        Input split bytes=9191
        Combine input records=0
        Combine output records=0
        Reduce input groups=3
        Reduce shuffle bytes=8546
        Reduce input records=150
        Reduce output records=3
        Spilled Records=300
        Shuffled Maps =303
        Failed Shuffles=0
        Merged Map outputs=303
        GC time elapsed (ms)=7354
        CPU time spent (ms)=74760
        Physical memory (bytes) snapshot=27429158912
        Virtual memory (bytes) snapshot=104184516608
        Total committed heap usage (bytes)=20937965568
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
```

	WRONG_REDUCE=0 File Input Format Counters Bytes Read=268927 File Output Format Counters Bytes Written=96 17/02/07 12:07:54 INFO streaming.StreamJob: Output directory: vanilla_25
26	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=100 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py - reducer reduce r_kmeans.py -input data.txt -output vanilla_26 17/02/07 12:10:45 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 12:10:45 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar6753919729467273032/] [] /tmp/streamjob3867047809174245884.jar tmpDir=null 17/02/07 12:10:46 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:10:47 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:10:48 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 12:10:48 INFO mapreduce.JobSubmitter: number of splits:101 17/02/07 12:10:48 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0303 17/02/07 12:10:49 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0303 17/02/07 12:10:49 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0303/ 17/02/07 12:10:49 INFO mapreduce.Job: Running job: job_1484631414223_0303 17/02/07 12:10:56 INFO mapreduce.Job: Job job_1484631414223_0303 running in uber mode : false 17/02/07 12:10:56 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 12:11:02 INFO mapreduce.Job: map 1% reduce 0% 17/02/07 12:11:03 INFO mapreduce.Job: map 2% reduce 0% 17/02/07 12:11:04 INFO mapreduce.Job: map 4% reduce 0% 17/02/07 12:11:05 INFO mapreduce.Job: map 10% reduce 0% 17/02/07 12:11:06 INFO mapreduce.Job: map 18% reduce 0% 17/02/07 12:11:07 INFO mapreduce.Job: map 22% reduce 0% 17/02/07 12:11:08 INFO mapreduce.Job: map 29% reduce 0% 17/02/07 12:11:09 INFO mapreduce.Job: map 36% reduce 0% 17/02/07 12:11:10 INFO mapreduce.Job: map 44% reduce 0% 17/02/07 12:11:11 INFO mapreduce.Job: map 53% reduce 0% 17/02/07 12:11:12 INFO mapreduce.Job: map 62% reduce 0% 17/02/07 12:11:13 INFO mapreduce.Job: map 75% reduce 0% 17/02/07 12:11:14 INFO mapreduce.Job: map 88% reduce 0% 17/02/07 12:11:15 INFO mapreduce.Job: map 94% reduce 0% 17/02/07 12:11:16 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 12:11:17 INFO mapreduce.Job: map 100% reduce 33% 17/02/07 12:11:18 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 12:11:18 INFO mapreduce.Job: Job job_1484631414223_0303 completed successfully 17/02/07 12:11:18 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6746 FILE: Number of bytes written=11356275 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=278118 HDFS: Number of bytes written=96 HDFS: Number of read operations=312 HDFS: Number of large read operations=0 HDFS: Number of write operations=6 Job Counters Launched map tasks=101 Launched reduce tasks=3 Data-local map tasks=32 Rack-local map tasks=69 Total time spent by all maps in occupied slots (ms)=767377 Total time spent by all reduces in occupied slots (ms)=29240 Total time spent by all map tasks (ms)=767377 Total time spent by all reduce tasks (ms)=29240 Total vcore-seconds taken by all map tasks=767377 Total vcore-seconds taken by all reduce tasks=29240 Total megabyte-seconds taken by all map tasks=785794048 Total megabyte-seconds taken by all reduce tasks=29941760 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=8546 Input split bytes=9191 Combine input records=0 Combine output records=0 Reduce input groups=3

	Reduce shuffle bytes=8546 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =303 Failed Shuffles=0 Merged Map outputs=303 GC time elapsed (ms)=7527 CPU time spent (ms)=73330 Physical memory (bytes) snapshot=27380195328 Virtual memory (bytes) snapshot=104127143936 Total committed heap usage (bytes)=20937965568 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=268927 File Output Format Counters Bytes Written=96 17/02/07 12:11:18 INFO streaming.StreamJob: Output directory: vanilla_26
27	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=100 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reduce r_kmeans.py -input data.txt -output vanilla_27 17/02/07 12:12:35 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 12:12:35 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar2668334762407567922/] [] /tmp/streamjob5803400854930229552.jar tmpDir=null 17/02/07 12:12:36 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:12:36 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:12:37 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 12:12:37 INFO mapreduce.JobSubmitter: number of splits:101 17/02/07 12:12:38 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0304 17/02/07 12:12:38 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0304 17/02/07 12:12:38 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0304/ 17/02/07 12:12:38 INFO mapreduce.Job: Running job: job_1484631414223_0304 17/02/07 12:12:43 INFO mapreduce.Job: Job job_1484631414223_0304 running in uber mode : false 17/02/07 12:12:43 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 12:12:47 INFO mapreduce.Job: map 1% reduce 0% 17/02/07 12:12:48 INFO mapreduce.Job: map 2% reduce 0% 17/02/07 12:12:50 INFO mapreduce.Job: map 6% reduce 0% 17/02/07 12:12:51 INFO mapreduce.Job: map 11% reduce 0% 17/02/07 12:12:52 INFO mapreduce.Job: map 16% reduce 0% 17/02/07 12:12:53 INFO mapreduce.Job: map 22% reduce 0% 17/02/07 12:12:54 INFO mapreduce.Job: map 30% reduce 0% 17/02/07 12:12:55 INFO mapreduce.Job: map 37% reduce 0% 17/02/07 12:12:56 INFO mapreduce.Job: map 44% reduce 0% 17/02/07 12:12:57 INFO mapreduce.Job: map 50% reduce 0% 17/02/07 12:12:58 INFO mapreduce.Job: map 62% reduce 0% 17/02/07 12:12:59 INFO mapreduce.Job: map 74% reduce 0% 17/02/07 12:13:00 INFO mapreduce.Job: map 89% reduce 0% 17/02/07 12:13:01 INFO mapreduce.Job: map 94% reduce 0% 17/02/07 12:13:02 INFO mapreduce.Job: map 100% reduce 0% 17/02/07 12:13:03 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 12:13:04 INFO mapreduce.Job: Job job_1484631414223_0304 completed successfully 17/02/07 12:13:04 INFO mapreduce.Job: Counters: 51 File System Counters FILE: Number of bytes read=6746 FILE: Number of bytes written=11356275 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=278118 HDFS: Number of bytes written=96 HDFS: Number of read operations=312 HDFS: Number of large read operations=0 HDFS: Number of write operations=6 Job Counters Killed map tasks=1 Launched map tasks=102 Launched reduce tasks=3 Data-local map tasks=11

	Rack-local map tasks=91 Total time spent by all maps in occupied slots (ms)=771655 Total time spent by all reduces in occupied slots (ms)=27541 Total time spent by all map tasks (ms)=771655 Total time spent by all reduce tasks (ms)=27541 Total vcore-seconds taken by all map tasks=771655 Total vcore-seconds taken by all reduce tasks=27541 Total megabyte-seconds taken by all map tasks=790174720 Total megabyte-seconds taken by all reduce tasks=28201984 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=8546 Input split bytes=9191 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=8546 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =303 Failed Shuffles=0 Merged Map outputs=303 GC time elapsed (ms)=6959 CPU time spent (ms)=73100 Physical memory (bytes) snapshot=27491721216 Virtual memory (bytes) snapshot=104289509376 Total committed heap usage (bytes)=20937965568 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=268927 File Output Format Counters Bytes Written=96 17/02/07 12:13:04 INFO streaming.StreamJob: Output directory: vanilla_27
28	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=150 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py -reducer reducer_kmeans.py -input data.txt -output vanilla_28 17/02/07 12:15:55 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 12:15:55 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar4271610616294651757/] [] /tmp/streamjob4392419037645778311.jar tmpDir=null 17/02/07 12:15:56 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:15:56 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:15:57 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 12:15:58 INFO mapreduce.JobSubmitter: number of splits:154 17/02/07 12:15:58 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0305 17/02/07 12:15:58 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0305 17/02/07 12:15:58 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0305/ 17/02/07 12:15:58 INFO mapreduce.Job: Running job: job_1484631414223_0305 17/02/07 12:16:04 INFO mapreduce.Job: Job job_1484631414223_0305 running in uber mode : false 17/02/07 12:16:04 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 12:16:10 INFO mapreduce.Job: map 1% reduce 0% 17/02/07 12:16:13 INFO mapreduce.Job: map 3% reduce 0% 17/02/07 12:16:14 INFO mapreduce.Job: map 6% reduce 0% 17/02/07 12:16:15 INFO mapreduce.Job: map 12% reduce 0% 17/02/07 12:16:16 INFO mapreduce.Job: map 16% reduce 0% 17/02/07 12:16:17 INFO mapreduce.Job: map 19% reduce 0% 17/02/07 12:16:18 INFO mapreduce.Job: map 24% reduce 0% 17/02/07 12:16:19 INFO mapreduce.Job: map 30% reduce 0% 17/02/07 12:16:20 INFO mapreduce.Job: map 34% reduce 0% 17/02/07 12:16:21 INFO mapreduce.Job: map 38% reduce 0% 17/02/07 12:16:22 INFO mapreduce.Job: map 45% reduce 0% 17/02/07 12:16:23 INFO mapreduce.Job: map 53% reduce 0% 17/02/07 12:16:24 INFO mapreduce.Job: map 58% reduce 0% 17/02/07 12:16:25 INFO mapreduce.Job: map 62% reduce 0% 17/02/07 12:16:26 INFO mapreduce.Job: map 66% reduce 0% 17/02/07 12:16:27 INFO mapreduce.Job: map 72% reduce 7% 17/02/07 12:16:28 INFO mapreduce.Job: map 78% reduce 7%

	17/02/07 12:16:29 INFO mapreduce.Job: map 82% reduce 24% 17/02/07 12:16:30 INFO mapreduce.Job: map 90% reduce 26% 17/02/07 12:16:31 INFO mapreduce.Job: map 96% reduce 26% 17/02/07 12:16:32 INFO mapreduce.Job: map 98% reduce 30% 17/02/07 12:16:33 INFO mapreduce.Job: map 100% reduce 32% 17/02/07 12:16:34 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 12:16:35 INFO mapreduce.Job: Job job_1484631414223_0305 completed successfully 17/02/07 12:16:35 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6746 FILE: Number of bytes written=17136985 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=423398 HDFS: Number of bytes written=96 HDFS: Number of read operations=471 HDFS: Number of large read operations=0 HDFS: Number of write operations=6 Job Counters Launched map tasks=154 Launched reduce tasks=3 Data-local map tasks=49 Rack-local map tasks=105 Total time spent by all maps in occupied slots (ms)=1202301 Total time spent by all reduces in occupied slots (ms)=53753 Total time spent by all map tasks (ms)=1202301 Total time spent by all reduce tasks (ms)=53753 Total vcore-seconds taken by all map tasks=1202301 Total vcore-seconds taken by all reduce tasks=53753 Total megabyte-seconds taken by all map tasks=1231156224 Total megabyte-seconds taken by all reduce tasks=55043072 MapReduce Framework Map input records=150 Map output records=150 Map output bytes=6428 Map output materialized bytes=9500 Input split bytes=14014 Combine input records=0 Combine output records=0 Reduce input groups=3 Reduce shuffle bytes=9500 Reduce input records=150 Reduce output records=3 Spilled Records=300 Shuffled Maps =462 Failed Shuffles=0 Merged Map outputs=462 GC time elapsed (ms)=11347 CPU time spent (ms)=110900 Physical memory (bytes) snapshot=41548230656 Virtual memory (bytes) snapshot=157248761856 Total committed heap usage (bytes)=31608274944 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=409384 File Output Format Counters Bytes Written=96 17/02/07 12:16:35 INFO streaming.StreamJob: Output directory: vanilla_28
29	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=150 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py - reducer reduce r_kmeans.py -input data.txt -output vanilla_29 17/02/07 12:26:15 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 12:26:15 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar6041964666497112654/] [] /tmp/streamjob457219947676695399.jar tmpDir=null 17/02/07 12:26:16 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:26:16 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:26:17 INFO mapred.FileInputFormat: Total input paths to process : 1

```
17/02/07 12:26:17 INFO mapreduce.JobSubmitter: number of splits:154
17/02/07 12:26:17 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0306
17/02/07 12:26:18 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0306
17/02/07 12:26:18 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0306/
17/02/07 12:26:18 INFO mapreduce.Job: Running job: job_1484631414223_0306
17/02/07 12:26:25 INFO mapreduce.Job: Job job_1484631414223_0306 running in uber mode : false
17/02/07 12:26:25 INFO mapreduce.Job:  map 0% reduce 0%
17/02/07 12:26:31 INFO mapreduce.Job:  map 1% reduce 0%
17/02/07 12:26:32 INFO mapreduce.Job:  map 2% reduce 0%
17/02/07 12:26:33 INFO mapreduce.Job:  map 4% reduce 0%
17/02/07 12:26:34 INFO mapreduce.Job:  map 8% reduce 0%
17/02/07 12:26:35 INFO mapreduce.Job:  map 11% reduce 0%
17/02/07 12:26:36 INFO mapreduce.Job:  map 16% reduce 0%
17/02/07 12:26:37 INFO mapreduce.Job:  map 19% reduce 0%
17/02/07 12:26:38 INFO mapreduce.Job:  map 25% reduce 0%
17/02/07 12:26:39 INFO mapreduce.Job:  map 31% reduce 0%
17/02/07 12:26:40 INFO mapreduce.Job:  map 39% reduce 0%
17/02/07 12:26:41 INFO mapreduce.Job:  map 41% reduce 0%
17/02/07 12:26:42 INFO mapreduce.Job:  map 49% reduce 0%
17/02/07 12:26:43 INFO mapreduce.Job:  map 55% reduce 0%
17/02/07 12:26:44 INFO mapreduce.Job:  map 58% reduce 0%
17/02/07 12:26:45 INFO mapreduce.Job:  map 61% reduce 0%
17/02/07 12:26:46 INFO mapreduce.Job:  map 68% reduce 0%
17/02/07 12:26:47 INFO mapreduce.Job:  map 72% reduce 0%
17/02/07 12:26:48 INFO mapreduce.Job:  map 79% reduce 0%
17/02/07 12:26:49 INFO mapreduce.Job:  map 86% reduce 9%
17/02/07 12:26:50 INFO mapreduce.Job:  map 91% reduce 27%
17/02/07 12:26:51 INFO mapreduce.Job:  map 97% reduce 27%
17/02/07 12:26:52 INFO mapreduce.Job:  map 100% reduce 29%
17/02/07 12:26:53 INFO mapreduce.Job:  map 100% reduce 78%
17/02/07 12:26:54 INFO mapreduce.Job:  map 100% reduce 100%
17/02/07 12:26:55 INFO mapreduce.Job: Job job_1484631414223_0306 completed successfully
17/02/07 12:26:56 INFO mapreduce.Job: Counters: 50
    File System Counters
        FILE: Number of bytes read=6746
        FILE: Number of bytes written=17136828
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=423398
        HDFS: Number of bytes written=96
        HDFS: Number of read operations=471
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=6
    Job Counters
        Launched map tasks=154
        Launched reduce tasks=3
        Data-local map tasks=48
        Rack-local map tasks=106
        Total time spent by all maps in occupied slots (ms)=1217266
        Total time spent by all reduces in occupied slots (ms)=51800
        Total time spent by all map tasks (ms)=1217266
        Total time spent by all reduce tasks (ms)=51800
        Total vcore-seconds taken by all map tasks=1217266
        Total vcore-seconds taken by all reduce tasks=51800
        Total megabyte-seconds taken by all map tasks=1246480384
        Total megabyte-seconds taken by all reduce tasks=53043200
    MapReduce Framework
        Map input records=150
        Map output records=150
        Map output bytes=6428
        Map output materialized bytes=9500
        Input split bytes=14014
        Combine input records=0
        Combine output records=0
        Reduce input groups=3
        Reduce shuffle bytes=9500
        Reduce input records=150
        Reduce output records=3
        Spilled Records=300
        Shuffled Maps =462
        Failed Shuffles=0
        Merged Map outputs=462
        GC time elapsed (ms)=11145
        CPU time spent (ms)=112130
```

	Physical memory (bytes) snapshot=41539645440 Virtual memory (bytes) snapshot=157433729024 Total committed heap usage (bytes)=31608274944 Shuffle Errors BAD_ID=0 CONNECTION=0 IO_ERROR=0 WRONG_LENGTH=0 WRONG_MAP=0 WRONG_REDUCE=0 File Input Format Counters Bytes Read=409384 File Output Format Counters Bytes Written=96 17/02/07 12:26:56 INFO streaming.StreamJob: Output directory: vanilla_29
30	hadoop jar /usr/local/hadoop-2.6.0/share/hadoop/tools/lib/hadoop-streaming-2.6.0.jar -files mapper_kmeans.py,reducer_kmeans.py,clusters.txt -D mapred.map.tasks=150 -D mapred.reduce.tasks=3 -mapper mapper_kmeans.py - reducer reduce r_kmeans.py -input data.txt -output vanilla_30 17/02/07 12:28:03 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps 17/02/07 12:28:03 INFO Configuration.deprecation: mapred.reduce.tasks is deprecated. Instead, use mapreduce.job.reduces packageJobJar: [/tmp/hadoop-unjar5590870676266488099/] [] /tmp/streamjob5862505121619372268.jar tmpDir=null 17/02/07 12:28:04 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:28:05 INFO client.RMProxy: Connecting to ResourceManager at dsm1/172.16.100.155:8032 17/02/07 12:28:06 INFO mapred.FileInputFormat: Total input paths to process : 1 17/02/07 12:28:06 INFO mapreduce.JobSubmitter: number of splits:154 17/02/07 12:28:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1484631414223_0308 17/02/07 12:28:07 INFO impl.YarnClientImpl: Submitted application application_1484631414223_0308 17/02/07 12:28:07 INFO mapreduce.Job: The url to track the job: http://dsm1:8088/proxy/application_1484631414223_0308/ 17/02/07 12:28:07 INFO mapreduce.Job: Running job: job_1484631414223_0308 17/02/07 12:28:14 INFO mapreduce.Job: Job job_1484631414223_0308 running in uber mode : false 17/02/07 12:28:14 INFO mapreduce.Job: map 0% reduce 0% 17/02/07 12:28:19 INFO mapreduce.Job: map 1% reduce 0% 17/02/07 12:28:22 INFO mapreduce.Job: map 4% reduce 0% 17/02/07 12:28:23 INFO mapreduce.Job: map 7% reduce 0% 17/02/07 12:28:24 INFO mapreduce.Job: map 12% reduce 0% 17/02/07 12:28:25 INFO mapreduce.Job: map 14% reduce 0% 17/02/07 12:28:26 INFO mapreduce.Job: map 19% reduce 0% 17/02/07 12:28:27 INFO mapreduce.Job: map 21% reduce 0% 17/02/07 12:28:28 INFO mapreduce.Job: map 29% reduce 0% 17/02/07 12:28:29 INFO mapreduce.Job: map 36% reduce 0% 17/02/07 12:28:30 INFO mapreduce.Job: map 42% reduce 0% 17/02/07 12:28:31 INFO mapreduce.Job: map 47% reduce 0% 17/02/07 12:28:32 INFO mapreduce.Job: map 53% reduce 0% 17/02/07 12:28:33 INFO mapreduce.Job: map 58% reduce 0% 17/02/07 12:28:34 INFO mapreduce.Job: map 62% reduce 0% 17/02/07 12:28:35 INFO mapreduce.Job: map 66% reduce 0% 17/02/07 12:28:36 INFO mapreduce.Job: map 73% reduce 0% 17/02/07 12:28:37 INFO mapreduce.Job: map 78% reduce 8% 17/02/07 12:28:38 INFO mapreduce.Job: map 84% reduce 8% 17/02/07 12:28:39 INFO mapreduce.Job: map 90% reduce 8% 17/02/07 12:28:40 INFO mapreduce.Job: map 97% reduce 20% 17/02/07 12:28:41 INFO mapreduce.Job: map 99% reduce 31% 17/02/07 12:28:42 INFO mapreduce.Job: map 100% reduce 31% 17/02/07 12:28:43 INFO mapreduce.Job: map 100% reduce 77% 17/02/07 12:28:44 INFO mapreduce.Job: map 100% reduce 100% 17/02/07 12:28:45 INFO mapreduce.Job: Job job_1484631414223_0308 completed successfully 17/02/07 12:28:46 INFO mapreduce.Job: Counters: 50 File System Counters FILE: Number of bytes read=6746 FILE: Number of bytes written=17136985 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=423398 HDFS: Number of bytes written=96 HDFS: Number of read operations=471 HDFS: Number of large read operations=0 HDFS: Number of write operations=6 Job Counters Launched map tasks=154 Launched reduce tasks=3 Data-local map tasks=46 Rack-local map tasks=108 Total time spent by all maps in occupied slots (ms)=1221882

```
                Total time spent by all reduces in occupied slots (ms)=49480
                Total time spent by all map tasks (ms)=1221882
                Total time spent by all reduce tasks (ms)=49480
                Total vcore-seconds taken by all map tasks=1221882
                Total vcore-seconds taken by all reduce tasks=49480
                Total megabyte-seconds taken by all map tasks=1251207168
                Total megabyte-seconds taken by all reduce tasks=50667520
        MapReduce Framework
                Map input records=150
                Map output records=150
                Map output bytes=6428
                Map output materialized bytes=9500
                Input split bytes=14014
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=9500
                Reduce input records=150
                Reduce output records=3
                Spilled Records=300
                Shuffled Maps =462
                Failed Shuffles=0
                Merged Map outputs=462
                GC time elapsed (ms)=10672
                CPU time spent (ms)=112340
                Physical memory (bytes) snapshot=41598382080
                Virtual memory (bytes) snapshot=157333233664
                Total committed heap usage (bytes)=31608274944
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=409384
        File Output Format Counters
                Bytes Written=96
17/02/07 12:28:46 INFO streaming.StreamJob: Output directory: vanilla_30
```