Assignment 1: Hadoop MapReduce

Manish Kumar - 1009645840

1. The problem setup, input commands and the outputs are given in the Figure 1-3 below. The final result for number of lines is 58483.

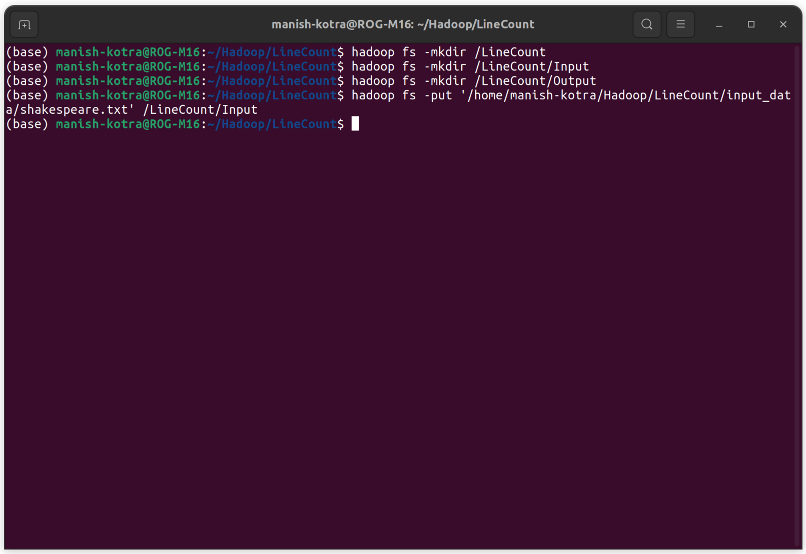

Figure 1. Initialize *Line Count* program

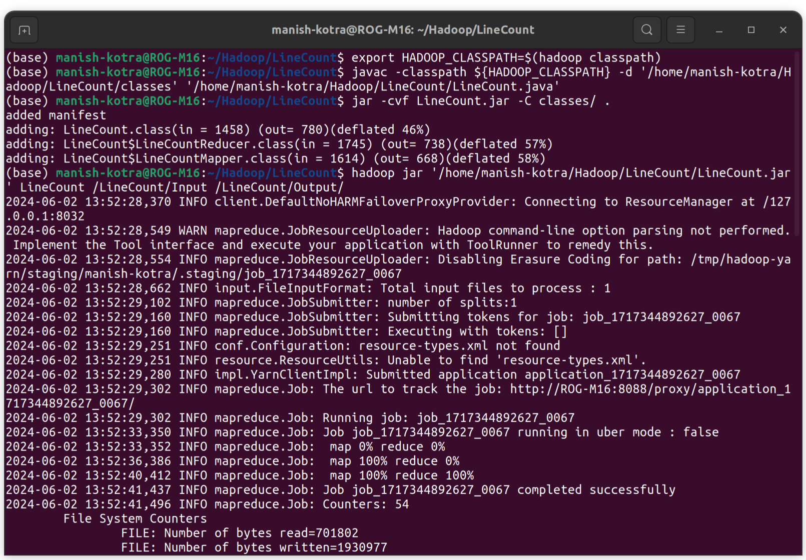

Figure 2. Run *Line Count* Program

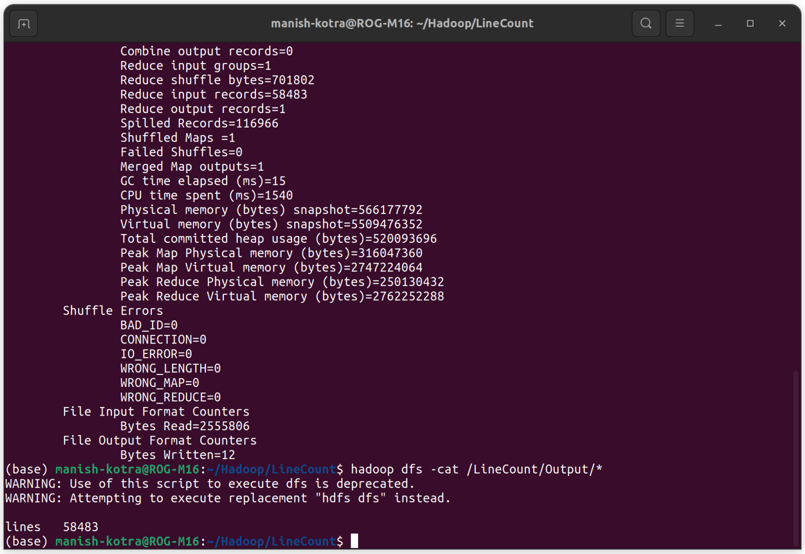

Figure 3. Output of *Line Count* program

2. The problem setup is given in Figure 4 below. The input commands and the final centroids for k-means algorithm for k=5 are shown in Figure 5 and 6, respectively. Similarly, input commands and the final centroids for k-means algorithm for k=8 are shown in Figure 7 and 8, respectively.

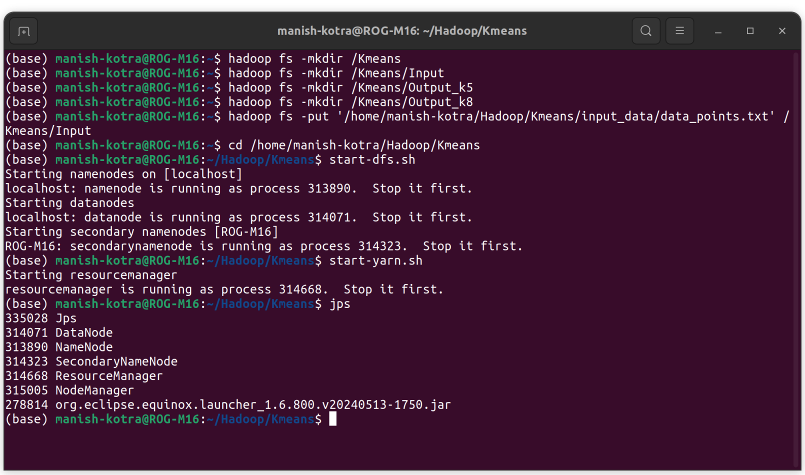

Figure 4. Initialize *KMeans Clustering* program

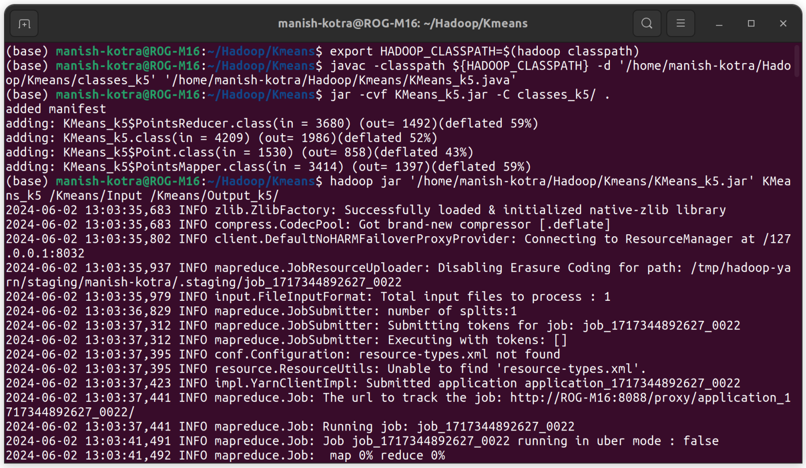

Figure 5. Run *KMeans Clustering* Program for k=5

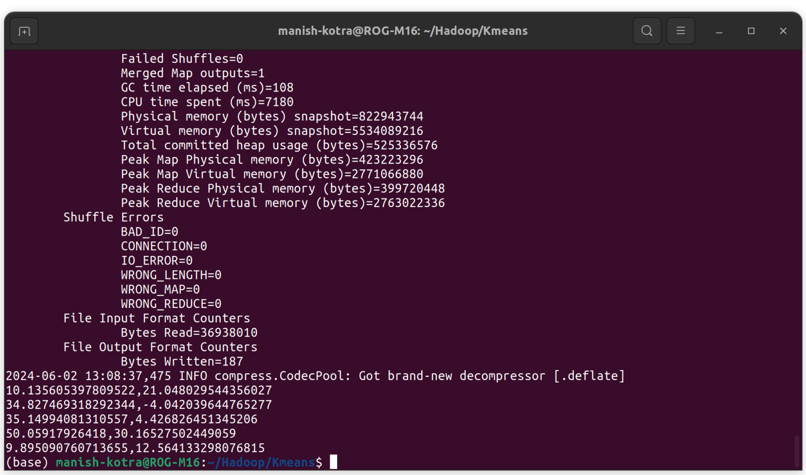

Figure 6. Output of *KMeans Clustering* program for k=5

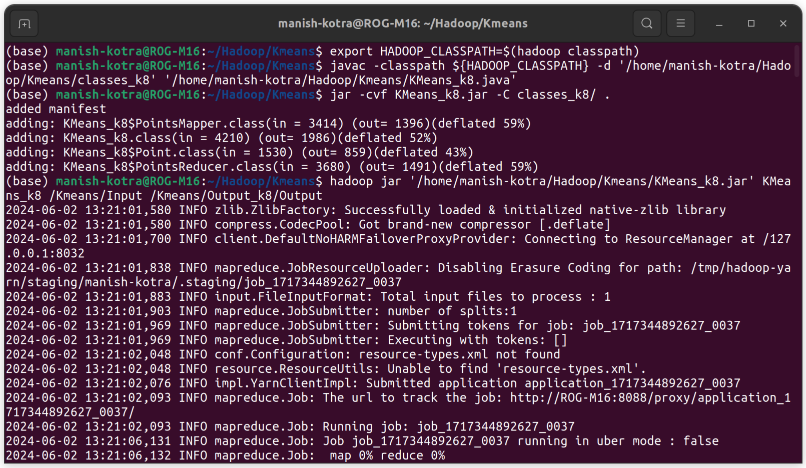

Figure 7. Run *KMeans Clustering* Program for k=8

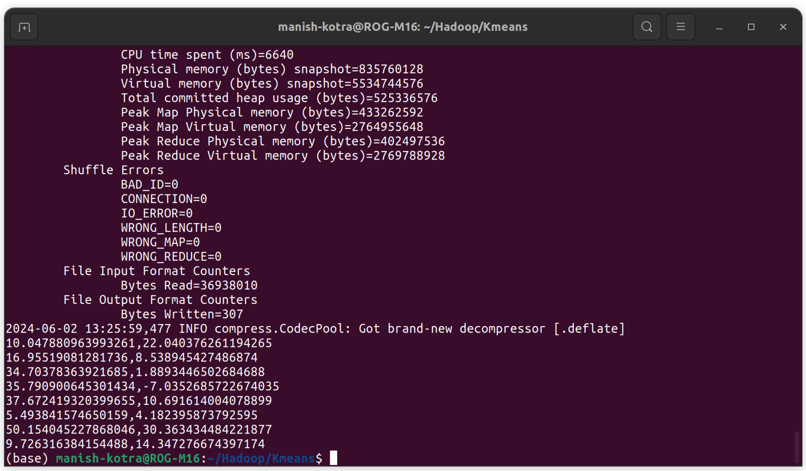

Figure 8. Output of *KMeans Clustering* program for k=8

3. Advantages of Using K-Means Clustering with MapReduce:

- Scalability: MapReduce allows K-Means to handle very large datasets by distributing the computation across multiple nodes, making it scalable to big data scenarios.
- Parallel Processing: The MapReduce framework inherently supports parallel processing, which can significantly speed up the clustering process.
- Efficiency: By breaking down the K-Means algorithm into map and reduce tasks, it can efficiently process large amounts of data in parallel, reducing the overall computation time.

Disadvantages of Using K-Means Clustering with MapReduce:

- Complexity: Implementing K-Means with MapReduce adds complexity to the setup and configuration, requiring knowledge of both the clustering algorithm and the MapReduce framework.
- Communication Overhead: There can be significant communication overhead between the map and reduce phases, which can impact performance, especially if the dataset is not large enough to justify the overhead.

4. Yes, the number of distance comparisons can be reduced by applying the Canopy Selection technique. This approach uses a cheap, approximate distance measure to group data points into overlapping subsets called canopies. Only the pairs of points within the same canopy are then considered for exact distance measurements, significantly reducing the number of expensive distance computations needed.

The distance metric used for canopy clustering should be one that is computationally inexpensive and able to provide a rough estimate of similarity. A commonly used metric is the inverted index, commonly used in information retrieval systems, are very efficient in high dimensions and can find elements near the query by examining only a small fraction of a data set.

5. Yes, it is possible to apply Canopy Selection on MapReduce. Here's a high-level implementation approach:

- Map Phase: Each data point is assigned to multiple mappers based on a cheap distance metric. Each mapper processes a subset of data points and identifies potential canopies by assigning points to canopies if they fall within a loose threshold distance from a randomly selected center point. The mapper outputs canopy assignments for each data point.
- Reduce Phase: Reducers collect canopy assignments from mapper and refine the canopies by collecting all the points belonging to the same canopy. Reducers then finalize the canopies, outputting the data points and their associated canopies.

6. Yes, it is possible to combine Canopy Selection with K-Means on MapReduce. The process involves the first two steps given in answer to question 5 above. After applying these steps of canopy selection the following steps of K-means algorithms is used to refine clusters and centroids:

- K-Means Initialization (Map Phase): Within each canopy, mappers initialize K-Means by selecting initial centroids. Each mapper processes points in its canopy, computing initial distances to centroids.
- K-Means Iteration (Map and Reduce Phases): Map Phase: Mappers assign points to the nearest centroid within each canopy. Reduce Phase: Reducers recompute the centroids based on the assigned points. These steps are iterated until convergence.