Data Science Project Report

1. Principal Investigator

Member 1: Kishan Kathiriya, N03362005, kathirik1@hawkmail.newpaltz.edu
Member 2: Harsh Parikh, N03370552, parikhh1@hawkmail.newpaltz.edu
Member 3: Kishan Malaviya, N03348838, malaviyk1@hawkmail.newpaltz.edu

Task	Member 1	Member 2	Member 3	Total
Introduction	33%	33%	34%	100%
Background	33%	34%	33%	100%
Implementation	34%	33%	33%	100%
Experiment Results and Discussion	34%	33%	33%	100%
Conclusion	33%	34%	33%	100%
Other contribution and explain	33%	33%	34%	100%

2. Title of Project

Clustering Data for Centre for Medicaid and Medicare Services

3. Mentoring

Professor **Min Chen**, Department of Computer Science, SUNY-New Paltz

chenm@newpaltz.edu

4. Introduction

4.1 Project Motivation

- Implementation of MapReduce for handling large data set

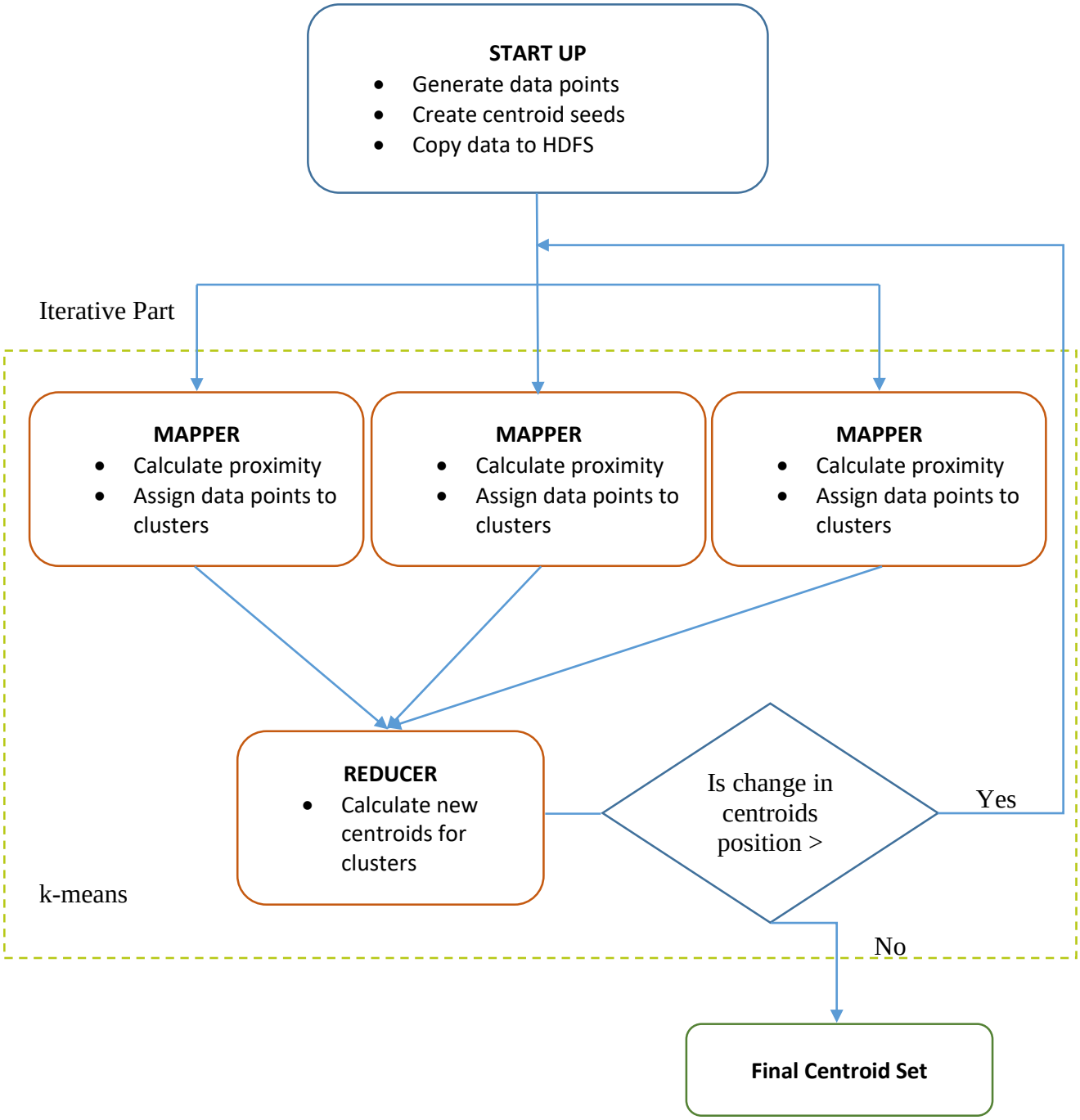
4.2 Aims and Objectives

- To classify large amount of data provided by Centre for Medical and Medicaid Services (CMS) using k-means algorithm implementing MapReduce Framework component of Hadoop.

5. Background/History of the Study

- CMS offers researchers and other health care professionals a broad range of quantitative information on our programs, from estimates of future Medicare and Medicaid spending to enrolment, spending, and claims data, and a broad range of consumer research to help its partners and staff. CMS also conducts demonstration projects to explore alternative policies of health care coverage and delivery. These demonstration project typically cover a limited timeframe, geographic area, and scope of coverage.

6. Approach and Implementation



Pre Process:

- Very first step of these project is Pre Process; In which we will put our sample dataset file to the MapReduce System.
- Sample Dataset file consists of 28 columns and 9287877 records among which we will only use 8 columns listed below for clustering.
    - LINE_SRVC_CNT,
    - BENE_UNIQUE_CNT,
    - BENE_DAY_SRVC_CNT,
    - AVERAGE_MEDICARE_ALLOWED_AMT,
    - STDEV_MEDICARE_ALLOWED_AMT,
    - AVERAGE_SUBMITTED_CHRG_AMT,
    - STDEV_SUBMITTED_CHRG_AMT,
    - AVERAGE_MEDICARE_PAYMENT_AMT,
    - STDEV_MEDICARE_PAYME
- Preprocess.java file will get the data file and extract 8 columns from that data and create equal chunks on HDFS.

Min-Max:

- In order to determine the cost effectiveness we must sort the data.
- Minmax.java file will get the minimum and maximum values and sort the data

Standardization:

- These data consists of values ranging from Zero to Billions.
- Therefore to classify clusters we must have data in specific range.
- The standardmap.java file will get the data in the range of 0 to 1 on the calculated minmax using the following formula :

$$\frac{Value-Min}{Max-Min}$$

K- means Algorithm:

- We need to assign cluster to each rows.
- k_means_medicare.java file implements the k-means algorithm on the calculated standardized minmax; assigning cluster to each row. This process is iterative till it meets the exit criteria i.e. Δ which is calculated after every assignments.

Map Reduce:

- A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system.
- It runs with two Jobs called Map task and Reduce task
- Map: Mapper class will calculate proximities of data and assign data points to clusters
- Reduce: Reduce class will calculate new centroids for clusters in order to determine centroid position is > Δ or not.

7. Experiment Results and Discussion

- With the data set available (around ~3 GB), we tried to cluster the data and classified the data into 3 clusters.

8. Conclusion

- The output of the program gives us the clustering of the data.
- To execute the program on CMS data without using MapReduce can take double or triple the time. Implementing MapReduce decreased the time to 5-10 minutes.
- Advantages:
    - Fraud Detection
    - Customers can know based on the classification that which treatment comes in the budget at particular location.
    - Based on the clusters we can extract best providers in an area.

9. References

- https://www.cms.gov/Research-Statistics-Data-and-Systems/Research-Statistics-Data-and-Systems.html
- http://hadoop.apache.org/
- http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-single-node-cluster/