Big Data Coursework 1

1.1 Coursework Question 1

1.1.1 Introduction

In this project, we are going to use weather data from NCDC to find descriptive statistics for different weather parameters . The data is collected from different weather stations in July 2007 and is processed daily, from 01/07 to 19/07. To calculate the statistics, no libraries are used, only the MapReduce framework.

MapReduce can be considered a programming paradigm capable of enabling massive scalability across a large number of servers in a Hadoop cluster. The MapReduce term refers to two separate jobs that are performed in the process. The first job is the mapping job. The mapping job takes a set of data as input and converts it into a set of key-value pairs. [1] The reduce job takes the output produced by the mapping job as input and combines the data into a smaller set. As implied by the name MapReduce, the map job is run before the reduce job. However, MapReduce has some limitations. The framework is not flexible. It has only one possible flow of execution. The framework can only offer solutions to problems that can be adapted to the MapReduce flow. In our case, since our aim was to extract some simple statistics from weather data, we managed to adapt to its flow. However, the amount of code written was high. This brings another limitation to the framework. the need for a lot of manual coding required even for simple common operations sunch as join, sort, filter etc. [2]

There are many columns in the dataset, but the only columns of interest for our project are the observation date, the wind speed, relative humidity, dew point temperature, and dry bulb temperature. By selecting only this data, the observations can be grouped by day and statistics like daily maximum, daily minimum daily average and variance can be calculated.

The pseudo-code for the map and reduce operations is shown below. For simplicity, for the reducer class only the mean function is shown. A separate reducer is used for every unique key or in this case date. For more information we could refer to the python code below.

```
class Mapper

    FOR EACH dataset row

        SET the key to the observation date

        SET the value as [dry bulb temperature, dew point temperature, humidity, wind spee
    d]

        OUTPUT the key/value pair

    END

class Reducer

    SET sum = 0

    SET n = 0

    FOR EACH key/value pair coming from the mapper

        SET n = n + 1
        SET sum = sum + value
    END
    SET mean = sum / n
```

1.1.2 Results

After applying the Map and Reducer to the weather data, a file with 19 rows of strings (one for each day) and a correlation matrix that describes the monthly correlation among, Humidity, Wind_Speed, and Dry_Bulb_Temp will be produced. Each row contains the following values separated by commas:

Day (YYYYMMDD), difference between Wind_Speed maximum and minimum, Humidity minimum, Dew_Point_Temp mean, and Dew_Point_Temp variance

e.g. 20070701, 31, 2, 54.4, 179.319630

From the results we can say that Relative Humidity is negatively correlated to Wind_Speed and Dew_Point_Temperature. On the last day 19/07 the humidity is higher than the other days meaning that it has rained. The difference between the maximum and minimum wind speed is lower, the mean of dew point temperature is higher, and its variance is lower than the other days. The range of wind speed difference except for a couple of days is between 30 and 40. The humidity minimum values are low except for the last day. The same thing stands for the dew point temperature. However, its variance is higher in the other days and lower in the last day.

1.2 Coursework Question 2

1.2.1 Introduction

The answer to this question is divided into two parts. In the first part we are going to perform a cluster analysis using only Apache Mahout. In the second part we are going to apply some data analysis to compare the performance of different clusters, plot the elbow graph and find the best number of k and distance measure.

The command-line tools of Apache Mahout are used to convert the dataset to a suitable file format for processing by Hadoop, create a TF-IDF feature matrix, and perform clustering with the K-Means algorithm.

K-Means Clustering

K-Means Clustering is an unsupervised algorithm. It groups an unlabelled set of data into different K clusters, where K is the number of clusters. Each of these clusters is associated with a centroid. The algorithm aims to minimize the sum of distances between the data points present in a cluster and their centroids. The K-Means clustering algorithm has the following steps:

1. Select the number of K's.
2. Select random number of K points as centroids.
3. Assign each remaining point to the nearest centroid.
4. Re-calculate centroids
5. Repeat steps 3 and 4 until there are no more changes.

Apache Mahout

Apache Mahout is an open-source project. It is mostly used for creating scalable machine learning algorithms. Mahout operates in addition to Hadoop framework, that allows the application of machine learning techniques using distributed computing. Mahout's algorithms include, clustering, recommendation mining, classification, etc.

Mahout Limitations

During this project there were no problems encountered with Mahout, probably because the british-corpus data was relatively small. However, according to different sources it is inefficient compared to other frameworks. It tends to run out of memory if the dataset is too large and some algorithms are not supported. Furthermore, there is a shortage of support. Other framework like MLlib offer faster computing time. [3] [4]

1.2.2 Results

To decide which is the best setting for K-Means clustering for this dataset, we are going to analyze the inter cluster, intra cluster density and density ratio extracted from the Mahout cluster dumps. The best setting is the one that produces clusters where the data points are similar and therefore densely packed. Furthermore, the best setting produces clusters that are different between them. So, in our case, the best distance measure has the lowest inter cluster density and the highest intra cluster and density ratio.

After computing the experiments, we can say that as the number of K increases, the average point distance for all distance measures decreases. The best distance measure is Manhattan because it has a higher intra cluster density and density ratio. Another good distance measure was the Euclidean distance measure, while Cosine performed the worst in all experiments. However, to produce more significant results, it would be necessary to extend the number of K's taken into consideration.

One way of finding the optimum number of clusters for the K-Means clustering algorithm is the Elbow Method. It is often used to determine the point of diminishing returns for selecting the K value. The Elbow method helps finding the best number of clusters, by using the WCSS values concept. WCSS stands for Within Cluster Sum of Squares and defines the total variations within a cluster. To find the optimal number of clusters, the elbow method follows four steps:

1. Run the cluster algorithm for different K values.
2. Calculate the sum of squares WCSS value for each K.
3. Plot a curve between sum of squares values and the number of clusters.
4. The sharp point of the bend is then considered as the optimal value for the number of clusters K.

For the cosine and euclidean distance measures the elbow point is not so clear, so to find the optimum number of clusters we might need to try a larger number of k. For the Manhattan distance measure, we could say that the optimum number of K is 8.

References

1. What is Apache MapReduce? IBM. Available at: https://www.ibm.com/topics/mapreduce (Accessed: December 20, 2022).
2. Kalavri, Vasiliki & Vlassov, Vladimir. (2013). MapReduce: Limitations, Optimizations and Open Issues. 1031-1038. 10.1109/TrustCom.2013.126.

3. Expert, E.R.P. (2022) Apache Mahout (workflow, features, Pros, and cons), ERP Information. Available at: https://www.erp-information.com/apache-mahout#Cons (Accessed: December 20, 2022).
4. Abid, E.B. (2022) Top 10 best machine learning tools (pros and cons), Cloud Infrastructure Services. Available at: https://cloudinfrastructureservices.co.uk/best-machine-learning-tools/ (Accessed: December 20, 2022).