IJCTA, 10(9), 2017, pp. 129-144 ISSN: 0974-5572
© International Science Press

European Countries' Climatic Analysis and Forecasting

S. Dhamodharavadhani* **and R. Rathipriya****

ABSTRACT

Climate change concerns different sectors such as agriculture, forestry, urban and regional planning, nature conservation, water management, energy supply and tourism. Climate change prediction is one of the greatest natural resource challenges, today the world faces MapReduce is a programming model used for computation of large datasets. However, for the huge data, the standard Map Reduce algorithm is not able sufficiently towards satisfying the real application wants efficiently. This paper proposes a pre-processing algorithm support on MapReduce according to kind of huge climate data sets and also analyzing the data to forecasting the temperature. The main aim of this paper is to develop a technique based big data analytics for climate change problem using MapReduce, Seasonal Decomposition Model, Exponential Model and ARIMA Model.

Keywords: Climate data, MapReduce, Big data, Weather, Temperature, Climate variables, Seasonal Decomposition Model, Exponential Model, ARIMA Model

1. INTRODUCTION

The weather is an occurrence in the environment at any known time. The climate, in a fine intelligence, can be measured as the standard weather conditions, otherwise in an additional methodical precise way; it can be defined as the statistical explanation of the conditions of the mean and variability of applicable quantities more than an era of time . In a broader intelligence, the climate is the rank of the climate method which includes atmosphere, hydrosphere, cryosphere, surface lithosphere and biosphere. These fundamentals all decide the status and dynamics of the Earth's climate [1].

Weather is the random position on the environment about us, considered with temperature, wind, precipitation, clouds and other weather rudiments. Climate pass on to the standard weather and its unpredictability, more than certain time duration and a particular region. The climate is dissimilar from one region to another, depending on latitude, the coldness of the sea, vegetation, topography and other factors.

Table 1
Difference between Weather and Climate

Weather	*Climate*
It is an immediate full of atmosphere situation.	It is a criterion atmospheric state.
It can alter quickly, inside still not as much of the hour.	It maintains more than an era of 30 years, as defined by the World Climate Organization (WMO).
It prevails more than a small region.	It prevails more than a huge region.
It has just limited predictability.	It is approximately steady.
It depends primarily on the thickness (temperature and moisture) dissimilarity among one place and another place.	It depends on latitude, the coldness of the sea, vegetation, attendance or nonattendance of mountains, and other ecological issues.

* Research Scholar Department of Computer Science Periyar University Salem, India, *Email: vadhanimca2011@gmail.com*

** Assistant professor Department of Computer Science Periyar University Salem, India, *Email: rathipriyar@gmail.com*

Alike to weather, the climate unpredictability too happens in a dissimilar time range. The extensive word denotes of the climate of a place is extremely significant intended for our sympathetic as this decide a lot of issues which are helpful for human being livelihood. The following Table 1.1 is shown the difference between weather and climate.

1.1. Climatic Variables

Even though understanding the climate as hot or cold, infrequently perform to understand the climate is the effect of a slight stability among some elements, which include atmosphere, water systems, living organisms, topography.

These elements decide various features that preside over the climate. These factors are called climatic variables. Of these, the most major factors that are taken into account are rain, atmospheric pressure, the wind, humidity, and temperature. The general climatic variables are shown below.

1. *Temperature*: Temperature is the dimension of how hot and cold rather it. The most instant method, in which we can calculate this, is just by reacting it. It is exactly calculated by a thermometer. Temperature is input features in determining awake the weather and climate of an area. It presides over the air mass and causes wind.
2. *Atmospheric pressure*: It can be described as the energy per unit region wields besides a surface by the weight of atmospheric air. Atmospheric pressure depends on the mass of air. It controls the flow of air. Air moves from high pressure to low pressure. The force due to pressure difference causes movement of air or wind. This pressure difference is also called as the pressure gradient force. Movement of air from low to high pressure leads to the uplifting of air and causes the development of depression with clouds and rain.
3. *Rainfall*: Rainfall is the condensation of atmospheric water vapor into the earth's surface. Rainfall is a dangerous occurrence of our environment. It standardize the flow of water in the environment.
4. *Air density*: Air density is the mass per unit volume of air. It decreases with increasing altitude, temperature and humidity. Air density regulates the flow of air.

These climate variables are basic in climate data analysis research. In this paper to discuss the climate data format available in climate research, climate data analysis tools and methods and also the overview about climate model evaluation, which is useful for solving problems in climatologically research.

1.2. Temperature causes in Climate Change

All the parameters of the Earth's climate (the wind, rain, clouds, temperature…) are the result of energy transfer and transformations within the atmosphere at the Earth's surface and in the oceans. The temperature of the Earth results from a balance between energy coming into the Earth from the Sun (solar radiation). About half the solar radiation outstanding the Earth and its atmosphere is engrossed in the surface. The other half is engrossed by the atmosphere or replicate reverse into breathing space by clouds, little particle in the atmosphere, snow, ice and deserts by the Earth's exterior. Part of the energy absorbed at the Earth's surface is radiated back (or re-admitted) to the atmosphere and space in the form of heat energy. The temperature is considered to a measure of this heat energy. Part of it is reflected back to the Earth's surface by the atmosphere (the greenhouse effect) leading to a global average of around 14°C, well above the -19°C which would be felt without the natural greenhouse effect. Because the Earth is round and its position in the solar system, more solar energy is absorbed in the tropics creating temperature differences from the equator to the poles. Atmospheric and oceanic circulation contributes to reducing these differences by transporting heat from the tropics to the mid-latitudes and the Polar Regions. These equators to pole exchanges are the main driving force of the climate system.

The energy budget of the Earth can be changed, which in turn can affect the Earth's temperature. An increase in the greenhouse effect, feedbacks in the climate system, or other changes can modify the energy budget of the Earth. The high heat capacity of the oceans dampens the much higher temperature changes that would otherwise occur each day, each season and each year — both in coastal areas and often farther inland. Overall, the eruption of Mount Pinatubo caused quite a strong cooling of the global surface temperature (about 0.2°C) and in the troposphere (perhaps 0.4°C) from late 1991 to 1994[1].

The possibility that the Sun's energy output may have varied more appreciably in the past could explain the marked parallel between these changes and estimates of the Earth's surface temperature over much of the past four centuries. The global average surface air temperature is estimated to increase between 1.4°C and 5.8°C by 2100. Climate models cannot yet provide a detailed picture of regional climate change, but it is likely that nearly all land areas, particularly those at high latitudes in the winter season, will warm more rapidly than the global average. Most notable is the warming in the northern regions of North America, and northern and central Asia. This can be seen in the image to the right. In contrast, the warming is less than the expected global mean over South and Southeast Asia in summer and southern South America in winter. The surface temperature is likely to rise least in the North Atlantic and the circumpolar Southern Ocean.

MapReduce is an excellent model for distributed computing, introduced by Google in 2004. Each MapReduce job is composed of a certain number of map and reduce tasks. The MapReduce model for serving multiple jobs consists of a processor sharing queue for the map tasks and a multi-server queue for the reduce tasks [6].MapReduce is a key technology of using cloud computing to process a large amount of data. It is a parallel programming model and an associated implementation for processing and generating large datasets in a broad variety of real-world tasks. It is not only a programming model but also a task scheduling model. It is composing of two fundamental functions: map and reduce, defined by users. A map function is to handle a key/value pair to produce intermediate key/value pair. A reduce function is specified to combine all of the intermediate value with the same middle key [2]. MapReduce is typically used to perform distributed computing on clusters of computers. It abstracts the distributed computing from its complex details; such that programmers can handle large distributed system resources without any experience about a parallel or distributed system. Thereby, the effect originally achieved only by an expensive highperformance computer can be achieved by low-cost computing services. Some need to be adapted to take advantage of the efficiency of parallelization.

In this paper, develop the algorithm in the MapReduce framework. This paper is organized as follows. Related work is reviewed in section 2. In section 3, introduce the MapReduce programming model. In section 4, to describe temperature causes in climate change. In section 5, describe a methodology for large-scale climate data using MapReduce. In section 5, conduct the experiment to evaluate large-scale climate data. Finally, conclude the paper in section 6.

2. BACKGROUND

More sophisticated systems and techniques for climate change and analysis of large meteorological datasets are available in the literature table as shown below.

Author	*Contribution*
Fang, V. S. Sheng, XueZhi Wen, and	This paper proposes an improved MK-means algorithm (MK-means) based on MapReduce according to characteristics of large climate datasets. The results show that our K-means algorithm deployed in the large-scale climate data processing system is feasible and efficient. Next, we will further optimize the algorithm and integrate the system with other parallel and distributed algorithms into the system to meet the challenge of Big Data.
B¨ose et al (2010)	Implemented several incremental data mining algorithms including Naive Bayes and PCA to deal with large datasets,

(*contd...Table*)

Author	*Contribution*
Chao et al (2011)	Proposed a parallel Co -means algorithm based on MapReduce, which basically distributes the clustering load over a given number of processors
Li et al (2011)	Ensemble learning method-bagging to overcome the instability and sensitivity to outliers in clustering on large Datasets. There has been work on developing algorithms and approximation algorithms that fit into the MapReduce
Zhao et al (2009)	Fast parallel -means clustering algorithm based on the MapReduce framework; however, their approach does not consider the characteristics of large climate datasets and cannot achieve good results.
Sudhakaran & Chue Hong (2011)	Evaluate a scientific application from the environmental sciences for its suitability to use the MapReduce framework. Consider cccgistemp — a Python reimplementation of the original NASA GISS model for estimating global temperature change — which takes land and ocean temperature records from different sites, removes duplicate records, and adjusts for urbanization effects before calculating the 12 months running mean global temperature. The application consists of several stages, each displaying differing characteristics, and three stages have been ported to use Hadoop with the mrjob library.
Hansen & Lebedeff (1987)	Error estimates are based in part on studies of how accurately the actual station distributions are able to reproduce temperature change in a global data set produced by a three-dimensional general circulation model with realistic variability. Find that meaningful global temperature change can be obtained for the past century, despite the fact that the climate stations are confined mainly to continental and island locations. The results indicate a global warming of about 0.5°–0.7°C in the past century, with the warming of similar magnitude in both hemispheres; the northern hemisphere result is similar to that found by several other investigators
Githeko et al. (2000)	Estimated that with the global temperature rise from 1ºC-3.5ºC by 2100, the likelihood of much vector-borne diseases will also increase due to environmental changes.
Hales et al. (2002)	According to the result of climate change six billion people will be at risk of contracting dengue fever, compared with 3.5 billion people if the climate does not change.
Hay et al. (2006)	Used models to predict that approximately 260-320 million more people will be affected globally by malaria by 2080 as the changing climate creates new transmission zones.
Yusa et al (2015)	Furthermore, the adverse health impacts due to warming temperatures and droughts have also been documented across the United States of America (USA) and Canada in regard to respiratory problems from dust and water-borne diseases.
Hsiang et al (2013)	The tertiary effects of climate change on health are a result of many factors: the psycho-social distress from disasters, displacement of habitation and livelihood, resource, water and food shortages, and related conflict.
Parmeshwari, P. Sabnis, Chaitali A. Laulkar et al (2014)	MapReduce is a widely used data-parallel programming model for large-scale data analysis. The framework is shown to be scalable to thousand of computing nodes and reliable on commodity clusters. MapReduce provides simple programming interfaces with two functions: map and reduce. Moreover, MapReduce offers other benefits, including load balancing, high scalability, and fault tolerance. The challenge escalates when we consider that data are dynamically and continuously produced, from different geographical locations. Recently, many researchers tend to implement and deploy data-intensive or computation-intensive algorithms on MapReduce parallel computing framework for high processing efficiency
J. Dean, and S. Ghemawat, et al. (2008)	Job Tracker is the only master control, which can run on any computer in the cluster for scheduling and managing other Task Trackers, allocating Map task and Reduce task to free Task Trackers for parallel running and monitoring the condition of the tasks. There can be more than one Task Tracker. Task Tracker is in charge of the implementation of the tasks.
J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling, et al	Climate data are dramatically increasing in volume and complexity since users of these data in the scientific community and the public are rapidly increasing. Faced with such large-scale climate data, high-efficient computing power (more than a trillion times) is urgently

(*contd...Table*)

Author	*Contribution*
(2011).	required. Therefore, establishing a cloud computing the weather information processing system is very important and significant [1]
J. Dean and S. Ghemawat, et al (2004).	MapReduce is a key technology of using cloud computing to process a large amount of data. It is a parallel programming model and an associated implementation for processing and generating large datasets in a broad variety of the real world. tasks proposed by Google. It is not only a programming model but also a task scheduling model. It is composed of two fundamental functions: Map and Reduce, defined by users. A Map function is to handle a key/value pair to produce intermediate key/value pair. A Reduce function is specified to combine all of the intermediate value with the same middle key [2]
Parag N. Kolhe 1 Mr. Rahul M. Ugale 2 Miss Shraddha V Shingne 3, et al (2015).	Today the volume of data has been enormously increasing as a result of advances in data generation, collection and storage technologies. The effect of climate prediction on society, business, agriculture and almost all aspects of human life, force the scientist to give proper attention to the matter. Weather is a continuous, data intensive, multidimensional, dynamic process that makes weather forecasting a formidable challenge. The prediction of rare events is a pressing scientific problem. We primarily concentrate on identifying weather patterns in the long term while consistent with global climate change on weather patterns, identify rare/outlying patterns that coincide with rare events data mining. This paper proposes an adaptive clustering pattern detection method for the detection of rare patterns in climate change using data mining techniques which use a k-means algorithm where an open number of states as clusters to accommodate the dynamic temporarily of data
Ramya M G, Chetan Balaji, Girish L, et al (2015).	Query Analysis module is processing phase includes two parts of data reading/analyzing and the establishment of forecast result. MapReduce is an execution engine suitable for large data processing and can significantly improve the response speed for returning query results. In the second part, we implement prediction function for establishing forecast data through k-means cluster algorithm. This architecture has an ability of mass storage of climate data, efficient query and analysis, climate change prediction. We proposed a prediction technique with high accuracy.
T V Rajini kanth1, V V SSS Balaram2 and N. Rajasekhar, et al. (2014)	The temperature in terms of min or max or mean irrespective of it is increasing gradually and is found through k-means cluster analysis. The predictions can be done using the linear regression line equations that are found in an effective manner. The future scope of this is it can be extended to any huge data sets with various attributes /parameters for effective analysis and accurate prediction.

However, only some revisions on dealing with the large-scale meteorological data using MapReduce have been reported. In this paper, present a preprocessing technique using MapReduce for very large meteorological data and to analysis the climate data.

3. MAPREDUCE OVERVIEW

MapReduce is a framework for processing highly distributable problems across huge datasets using a large number of computers (nodes). The map and reduce functions of Map-Reduce are both defined with respect to data structured in <key, value> pairs. As said before, MapReduce is developed by Google. Its libraries have been written in many programming languages, such as Java, Python, and C++ [13–16]. It is mainly used to process large-scale (TB-level) data files. MapReduce is not only a simplified programming model but an efficient distributed scheduling model. Programming is very simple in such a cloud computing environment. The treatment of clusters is handled by the platform, including the reliability and scalability [17].Application developers only need to focus on the application itself. Map and Reduce are the two basic computing units of the MapReduce model. Massive data is cut into unrelated blocks by Map program and scheduled to lots of computers to process, achieving distributed computing.

Then the results from these computers are summarized and outputted by Reduce program. In MapReduce, massive data are processed in parallel. Data is initially partitioned across the nodes (computers) of a cluster and stored in a distributed file system (DFS). Data is represented as (key, value) pairs. The computation of the two functions is expressed formally as follows [5]: The program is used to calculate the annual maximum temperature [18]. A Map function is used to extract all the years and the temperatures (key/value pairs) appeared in the text, and these pairs are sent to an intermediate temporary space specified by MapReduce. Through intermediate processing by the Map function, the key/value pairs are grouped according to the key, so that each year is followed by a list of temperatures. Then, a Reduce function is only to find the maximum number through a whole list. The result is the annual maximum temperature. The intermediate results of each step of the execution process of MapReduce, including Map and Reduce phases, which both use all nodes in the cluster. Between the Map and Reduce phases, there is an intermediate phase, which concatenates the intermediate results with the same key into a list. The list will be used by the Reduce function to output the maximum temperature of a certain year [2].

MapReduce is a programming paradigm used for computation of large datasets. A standard MapReduce process computes terabytes or even peta bytes of data on interconnected systems forming a cluster of nodes. MapReduce implementation splits the huge data into chunks that are independently fed to the nodes so the number and size of each chunk of data are dependent on the number of nodes connected to the network. The programmer designs a Map function that uses a (key, value) pair for computation. The Map function results in the creation of another set of data in the form of (key, value) pair which is known as the intermediate data set.

The program also designs a Reduce function that combines value elements of the (key, value) paired intermediate data set having the same intermediate key. [10] Map and Reduce steps are separate and distinct and complete freedom is given to the programmer to design them. Each of the Map and Reduce steps is performed in parallel on pairs of (Key, value) data members. Thereby the program is segmented into two distinct and well-defined stages namely Map and Reduce. The Map stage involves execution of a function on a given data set in the form of (key, value) and generates the intermediate data set. The generated intermediate data set is then organized for the implementation of the Reduce operation. Data transfer takes place between the Map and Reduce functions. The Reduce function compiles all the data sets bearing the particular key and this process is repeated for all the various key values [19]. The final output produced by the Reduce call is also a dataset of (key, value) pairs. An important thing to note is that the execution of the Reduce function is possible only after the Mapping process is complete [17]. Each MapReduce Framework has a solo Job Tracker and multiple task trackers. Each node connected to the network has the right to behave as a slave Task Tracker. The issues like division of data to various nodes, task scheduling, node failures, task failure management, communication of nodes, monitoring the task progress are all taken care by the master node. The data used as input and output data is stored in the file-system [16] [20]

The Google’s MapReduce programming model is shown in Figure 1. To further understand the MapReduce programming model, the pseudo code of program based on MapReduce as shown below

4. METHODOLOGY

The proposed methodology consists of following steps

- Removing Missing Values
- Map Reduce-based framework for pre-processing climate data
- Seasonal Decomposition Model
- Exponential Model
- ARIMA Model

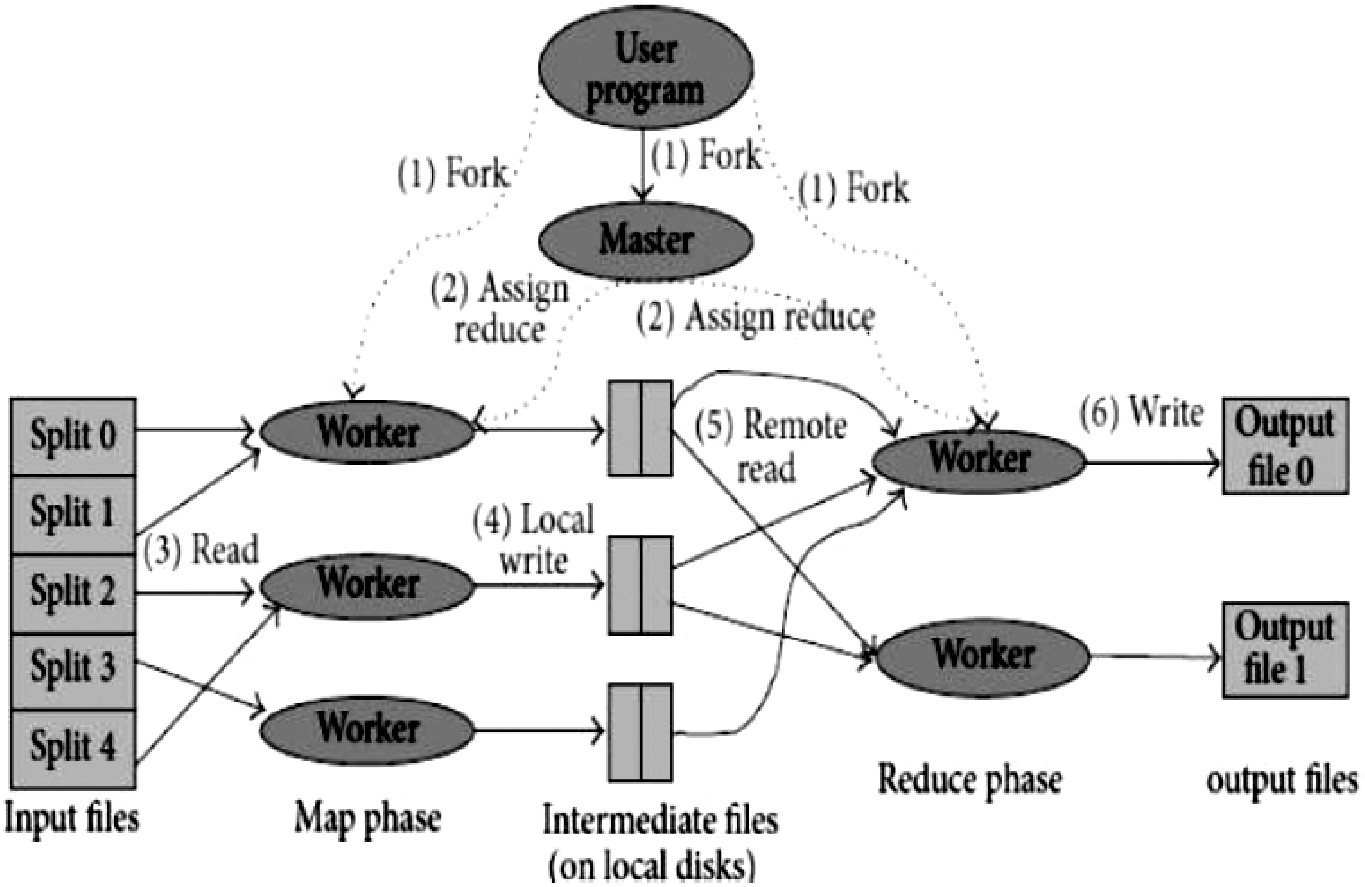


Figure 1: Google's MapReduce programming model [20]

4.1. Removing Missing Values

Many data sets contains one or more missing values. It is convenient to code missing values as NAN (Not a Number) to preserve the structure of data sets crosswise multiple variables and observations. These dataset file format (MISSING VALUE CODE = -9999). Before mapping year-wise data to remove the missing value - 9999 using the matlab function as shown below [4].

```
Removing Missing
Vaules function preprocess
foldername='H:\d';
fname=fullfile(foldername);
ds=datastore(fname, 'ReadVariableNames',false,'delimiter','\t')

for i=1:length(ds.Files)
   fname=ds.Files(i);
   ss=load(char(fname));
   id1=find (~(ss(:,4)==-9999));
   ss1=ss(id1,:);
   dlmwrite(char(fname),ss1,'delimiter','\t','newline','pc');
```

4.2. Map Reduce-based framework for pre-processing climate data

To further understand the MapReduce programming model, the agenda based on MapReduce is shown in Algorithm 1. The program is used to calculate the Max (maximum temperature) [18]. A Map function is used to extract all the years and station (key/value pairs) appeared in the text, and these pairs are sent to an intermediate temporary space specified by MapReduce. Through intermediate processing by the Map function, the key/value pairs are grouped according to the key, so that every year is followed by a list of the station. Then, a Reduce function is only to find the years through a whole list of stations. The result is the each and every year are having a list of all stations with temperature value [20]. Figure 1.4 shows the intermediate results of each step of the execution process of MapReduce, including Map and Reduce phases, which equally use all nodes in the cluster. Between the Map and Reduce phases, there is an intermediate phase, which concatenates the intermediate results with the same key into a list. The list will be use by the Reduce function to output the year wise data of a certain station [5] [14] [15]

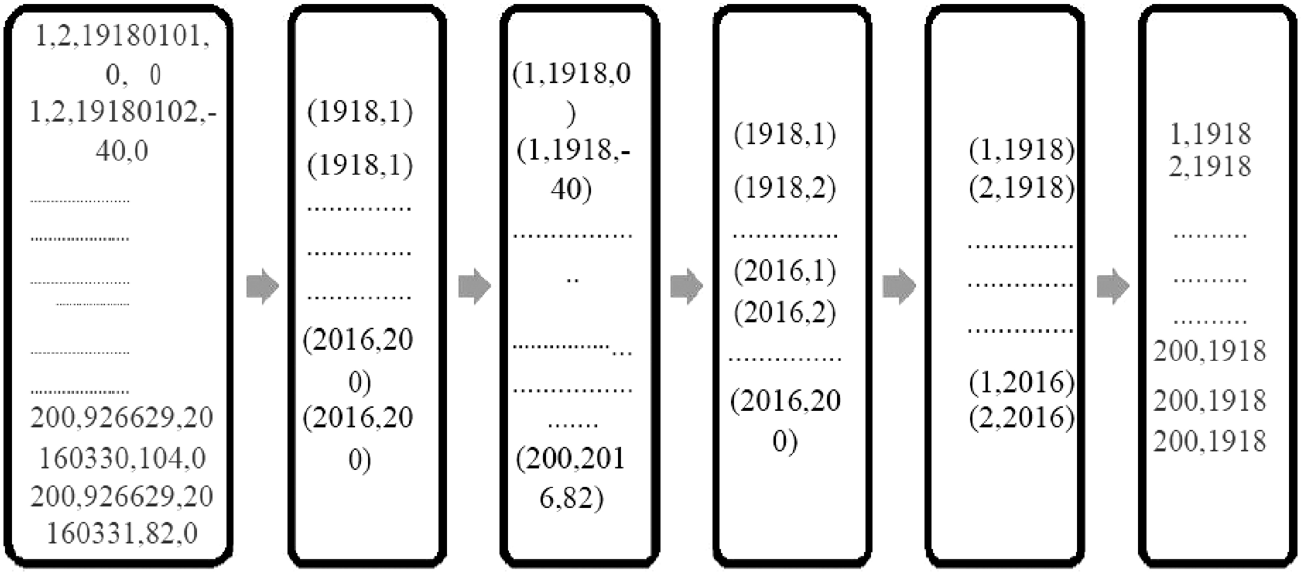


Figure 2: An example of the execution process of MapReduce, including the intermediate results of each step

Algorithm 4

```
map(String input key, String input value): //input key: Year //input value: ?Station?sid
for each year and Station id s in input value:
EmitIntermediate( , );
reduce(String output key, Interator intermediate values):
// output key: year
// intermediate values: a list of year wise
```

4.3. Seasonal Decomposition Model

A seasonal time series consists of a trend component, a seasonal component and an irregular component. Decomposing the time series means separating the time series into these three components: that is, estimating these three components. To estimate the trend component and seasonal component of a seasonal time series that can be described using an additive model. This function estimates the trend, seasonal, and irregular components of a time series that can be described using an additive model [21].

The function “decompose()” returns a list object as its result, where the estimates of the seasonal component, trend component and irregular component are stored in named elements of that list objects, called “seasonal”, “trend”, and “random” respectively.

Time series data can exhibit a huge variety of patterns and it is helpful to categorize some of the patterns and behaviours that can be seen in time series.

It is also sometimes useful to try to split a time series into several components, each representing one of the underlying pattern categories. Often this is done to help understand the time series better, but it can also be used to improve forecasts.

Trend: A trend exists when there is a long-term increase or decrease in the data. It does not have to be linear. Sometimes we will refer to a trend “changing direction” when it might go from an increasing trend to a decreasing trend [21].

Seasonal: A seasonal pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week). Seasonality is always of a fixed and known period.

Random: The random component could be analyzed for such things as mean absolute size, or mean squared size (variance), or possibly even for whether the component is actually random or might be modeled with an ARIMA model [21].

4.4. Exponential Model

If you have a time series that can be described using an additive model with a constant level and no seasonality, you can use simple exponential smoothing to make short-term forecasts. The simple exponential smoothing method provides a way of estimating the level at the current time point. Smoothing is controlled by the parameter alpha; for the estimate of the level at the current time point. The value of alpha; lies between 0 and 1. Values of alpha that are close to 0 means that little weight is placed on the most recent observations when making forecasts of future values [21].

4.5. ARIMA Model

Exponential smoothing methods are useful for making forecasts, and make no assumptions about the correlations between successive values of the time series. However, if you want to make prediction intervals for forecasts made using exponential smoothing methods, the prediction intervals require that the forecast errors are uncorrelated and are normally distributed with mean zero and constant variance [21].

While exponential smoothing methods do not make any assumptions about correlations between successive values of the time series, in some cases you can make a better predictive model by taking correlations of the data into account. Autoregressive Integrated Moving Average (ARIMA) models include an explicit statistical model of the irregular component of a time series, which allows for nonzero autocorrelations in the irregular component [21].

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1. Overview of Dataset

The dataset is collected from “European climate assessment & dataset” (eca&d), the file created on: 14-04-2016. These data can be used freely for non-commercial research provided that the following source is acknowledged: klein tank, a.m.g. and co-authors, 2002.. Data and metadata available at http://www.ecad.eu.

	Data Description
Missing Value Code	-9999
STAID	Station identifier
SOUID	Source identifier
DATE	Date YYYYMMDD
TX	Maximum temperature in 0.1 &#176; C
Q_TX	Quality code for TX (0='valid'; 1='suspect'; 9='missing').

The data used in our experiments has 5 attributes Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment.

S. No	*Column Name*	*Description*
1	STAID	Station identifier
2	SOUID	Source identifier
3	DATE	Date YYYYMMDD
4	TX	Maximum temperature in 0.1 &#176; C
5	Q_TX	Quality code for TX (0='valid'; 1='suspect'; 9='missing').

Dataset	*File Name*	*Capacity*	*Matrix*	*Type*
1	TX_STAID000001	1,262 KB	619*5	1918 year daily dataset
2	TX_STAID000002	1,262 KB	619*5	1919 year daily datasetz
3	TX_STAID000003	1,262 KB	619*5	1920 year daily dataset
4	TX_STAID000004	1,262 KB	619*5	1921 year daily dataset
	…………………………………..………………………………………………………			
	…………………………………..………………………………………………………			
11216	TX_STAID011216	1,262 KB	619*5	2016 year daily dataset

5.2. Experimental Results

After preprocessing the dataset to get the year wise data the following are shown the input and output of preprocessing technique.

Input: data of each automatic station

File 1: data of automatic station 1

File 2: data of automatic station 2

...

File ?: data[of automatic] station ?

Output: key, Value pair, key is the intermediate value of the clustering, while the value is intermediate value associated with the same key the sample distance is calculated based on Value; the minimum distance is repeatedly calculated based on the method of the center of mass;

To evaluate the performance of these algorithm for climate datasets, each year represents as a cluster. Calculate the mean value of each group to find the maximum value of the given year [13]. In these experiments is given daily maximum temperature value of the year 1918 to march 2016. The proposed algorithm is used to find the Max (Maximum Temperature), the result shows the following figure 3. The x axis represents as

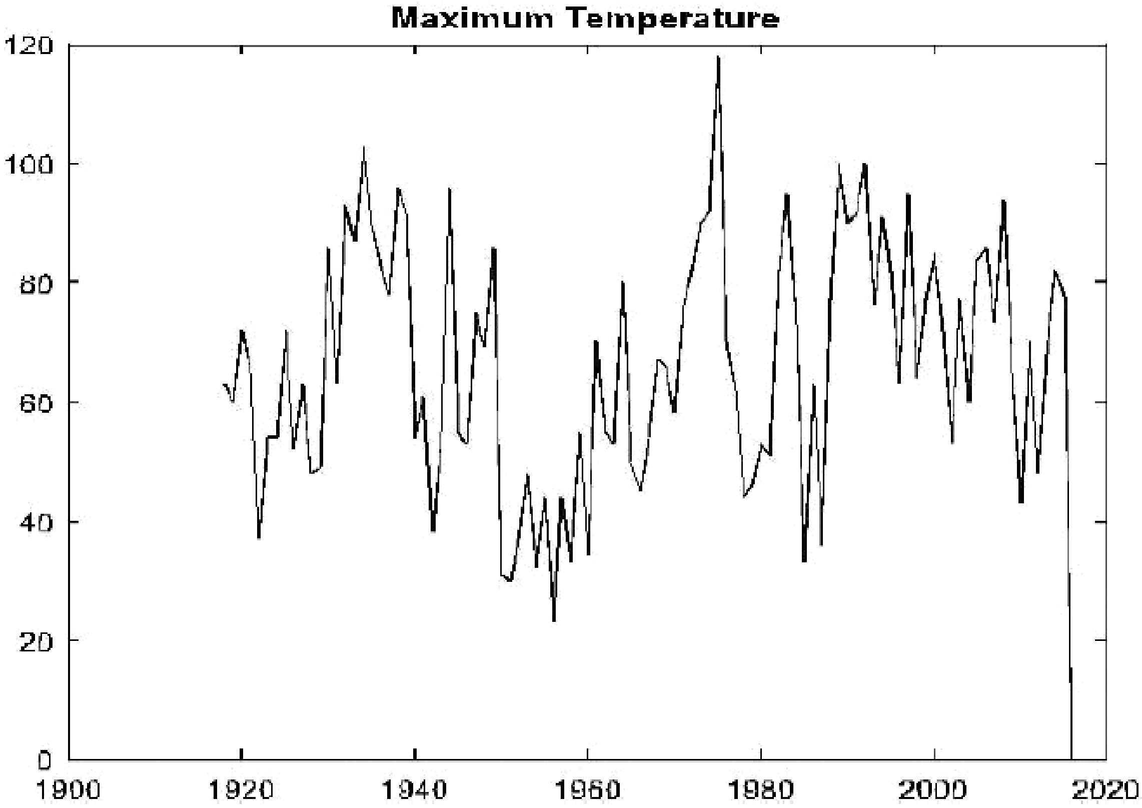


Figure 3: Maximum Temperature

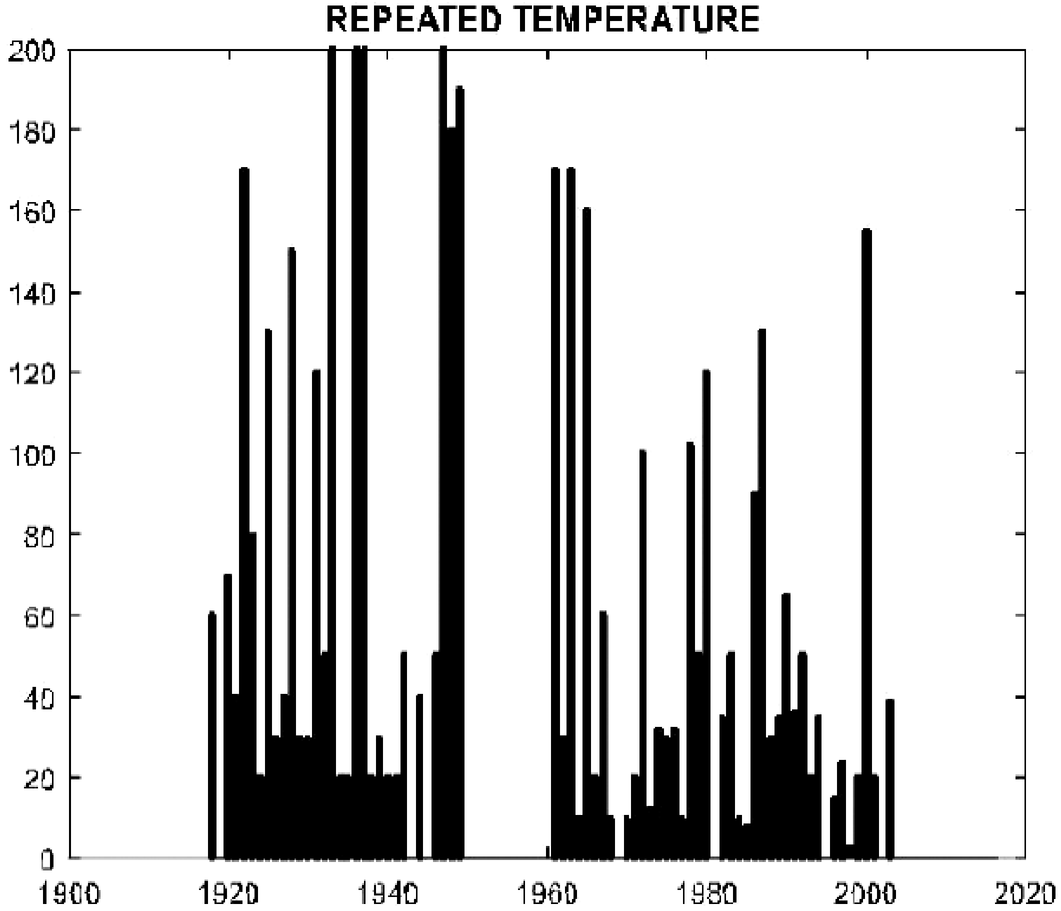


Figure 4: Repeated Temperature

year and y axis represent as the maximum value of temperature. A plot is shown the Max (Maximum Temperature) in each year. In 1970 as highest Max (Maximum Temperature), 1950 as smallest Max (Maximum Temperature) value.

Figure 3: An example of the implementation of MapReduce, including the intermediate results of each step

In the given dataset each and every year as repeated maximum temperature value. The proposed algorithm is used to find the repeated value of each year and also to find the count of repeated temperature value of the station [11] [12]. The given figure 4 is shown the repeated temperature value of each year in the 20th-century x-axis represents the year and y-axis represent the value of maximum temperature. In figure 5 shown the counts of repeated temperature value are shown the figure. Here x-axis represents the year and y-axis represent as values of station id.

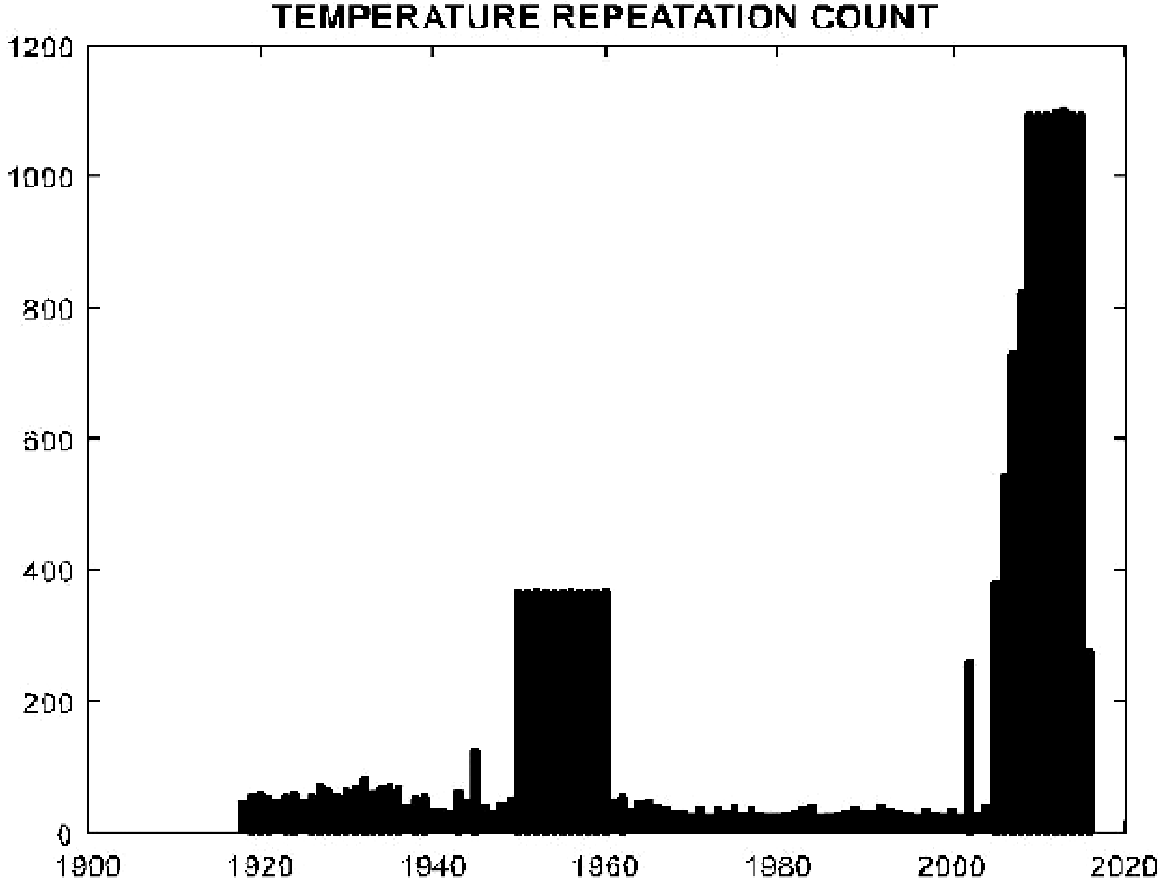


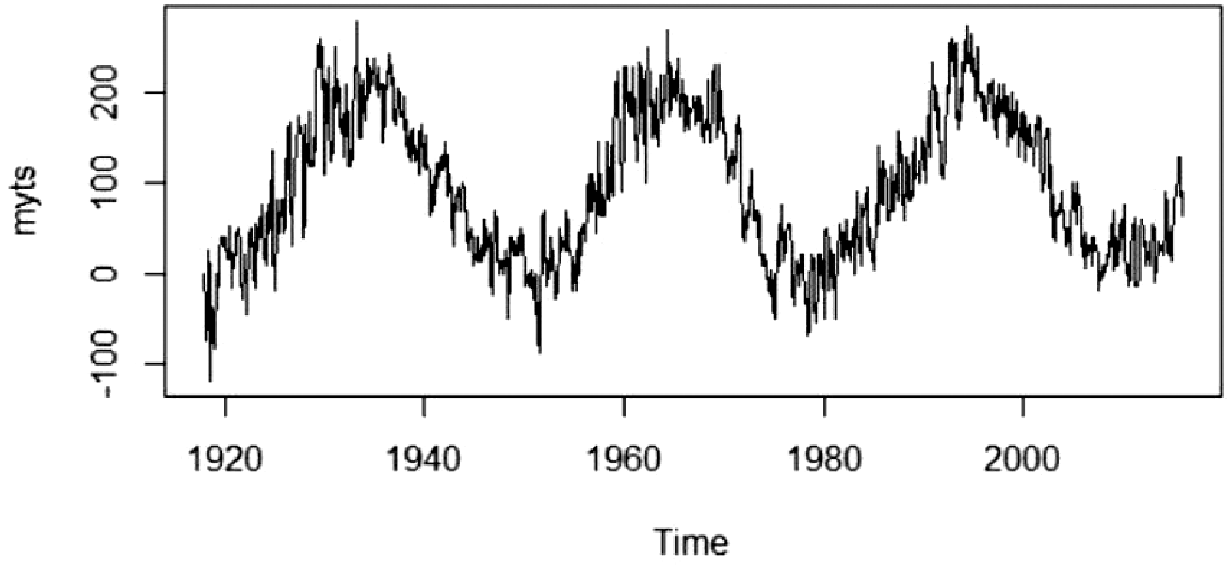
Figure 5: Temperature Repetition Count

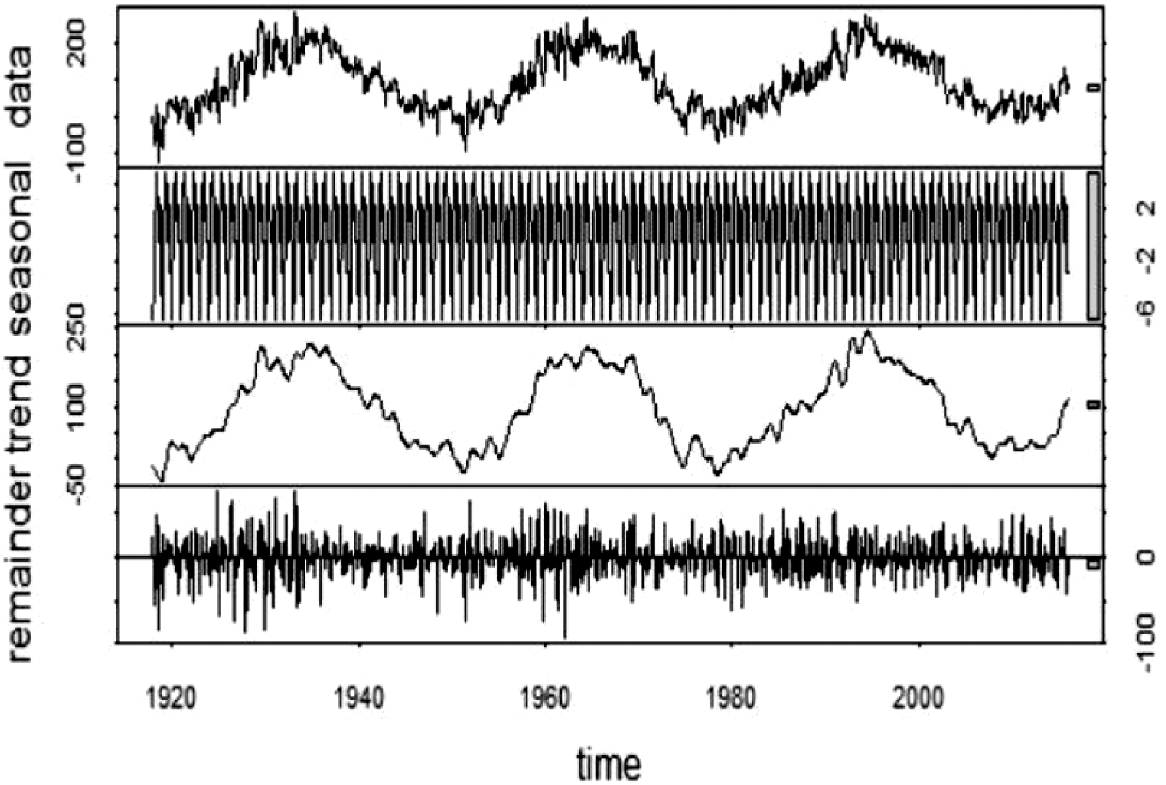
The below analysis the year wise data from 1918 to 2016 in 11216 stations in European counties to find the maximum temperature, repeated temperature and temperature repetition count of each and every year in given stations.

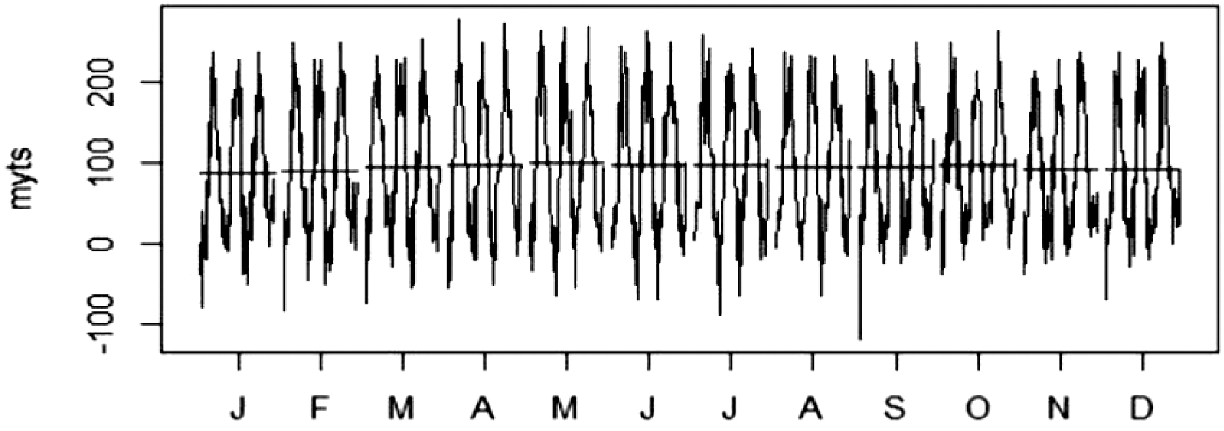
Seasonal Decomposition Model: In the seasonal decomposition model first part to creating a time series as shown the below figure.

The plot above shows maximum temperature is 280 and minimum temperature is - 120 in station id 1. X axis represents the value of the temperature and Y axis represents the year.

The plot above shows the original time series (top), the estimated trend constituent (second from top), the estimated seasonal constituent (third from top), and the predictable uneven constituent (bottom). See


myts
-100
0
100
200
1920
1940
1960
1980
2000
Time

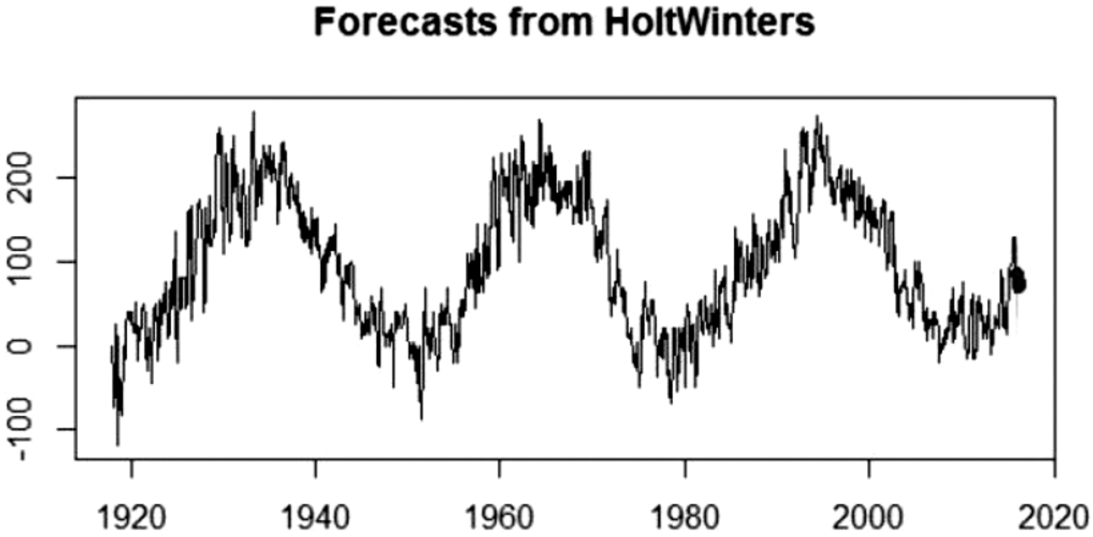

data
seasonal
trend
remainder
-100
200
2
-2
-6
-50
100
250
0
-100
1920
1940
1960
1980
2000
time


myts
-100
0
100
200
J
F
M
A
M
J
J
A
S
O
N
D

that the estimated trend constituent shows a small decrease from about 24 in 1947 to about 22 in 1948, followed by a stable increase from then on to about 27 in 1959.

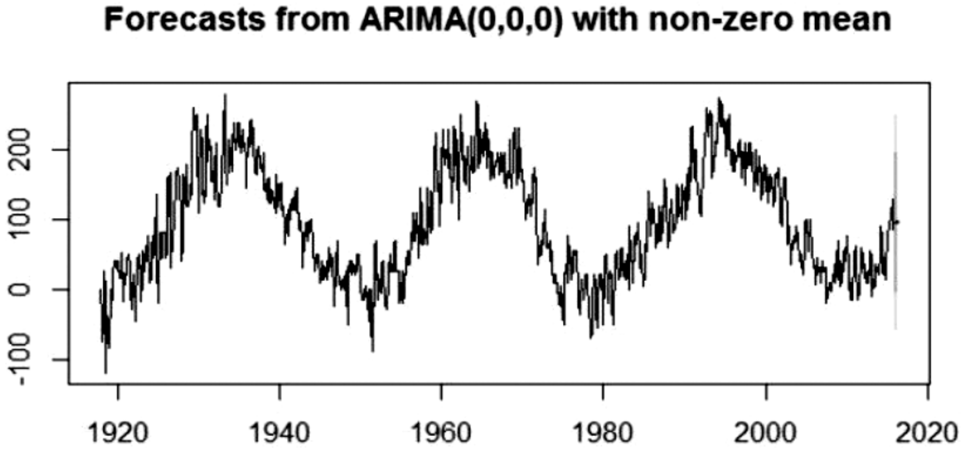
You can see from the plot to present is approximately constant level (the mean stays constant at about 25 inches). The random fluctuations in the time series appear to be approximately stable in size over time, so it is probably suitable to explain the data using an additive representation. Thus, we can make forecasts using simple exponential smoothing.

Exponential Model



Training set	
ME	-1.081313e-11
RMSE	79.87873
MAE	69.38719
MPE	-Inf
MAPE	Inf
MASE	3.584189
ACF1	0.9466095

ARIMA Model



Period	*Point Forecast*	*Lo 80*	*Hi 80*
Jan 2016	85.12572	50.341134	119.9103
Feb 2016	76.81480	34.028576	119.6010
Mar 2016	71.84778	22.297443	121.3981
Apr 2016	78.10817	22.577044	133.6393
May 2016	77.72317	16.763505	138.6828
Jun 2016	78.32690	12.354373	144.2994
July 2016	82.98903	12.330713	153.6473
Aug 2016	85.59730	10.518915	160.6757
Sep 2016	83.20554	3.928285	162.4828
Oct 2016	84.72331	1.434832	168.0118
Nov 2016	80.77931	-6.358713	167.9173
Dec 2016	89.11719	-1.729338	179.9637
Jan 2017	85.93045	-9.572033	181.4329
Feb 2017	77.61953	-21.318824	176.5579
Mar 2017	72.65251	-29.626021	174.9310

These model are used to forecasting the maximum temperature in station 1 in given dataset. In the dataset are having the temperature values from 1918 to 2015 .Here to forecast the temperature value Jan 2016 to Mar 2017 are shown below the table and plotting figure.

6. CONCLUSION

In this paper, the proposed MapReduce algorithms are used to pre-processing the huge dataset. After pre-processing the dataset first part we separate the whole dataset into clusters. Here to take the year as cluster tag. Based on the year wise data to evaluate the experimental result and analysis. To calculate the mean value of the cluster to find the max (maximum temperature) and min (maximum temperature). Then to group the repeated temperature value of each year and counts the repetition of temperature value of each station id. The results show that MapReduce has more computing ability, capability and scalability. ARIMA model is efficient to forecasting the time series analysis. Further, optimize the algorithm and integrate the system to meet the challenge of Big Data.

REFERENCES

[1] J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling, "Climate data challenges in the 21st century," Science, vol. 331, no. 6018, pp. 700– 702, 2011.

[2] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," in Proceedings of the 6th Symposium on Operating System Design and Implementation (OSDI '04), pp. 1–6, San Francisco, Calif, USA, December 2004.

[3] X. Wu, V. Kumar, Q. J. Ross et al., "Top 10 algorithms in data mining," Knowledge and Information Systems, vol. 14, no. 1, pp. 1–37, 2008.

[4] W. Zhao, H. Ma, and Q. He, "Parallel K-means clustering based on MapReduce," Cloud Computing, Springer, vol. 5931, pp. 674– 679, 2009.

[5] R. Vernica, M. J. Carey, and C. Li, "Efficient parallel set-similarity joins using MapReduce," in Proceedings of the International Conference on Management of Data (SIGMOD '10), pp. 495–506, Indianapolis, Ind, USA, June 2010.

[6] L. Chao, Y. Yan, and R. Tonny, "A parallel Cop-Kmeans clustering algorithm based on MapReduce framework," Advances in Intelligent and Soft Computing, vol. 123, pp. 93–102, 2011.

[7] A. Ene, S. Im, and B. Moseley, "Fast clustering using MapReduce," in Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11), pp. 681–689, August 2011.

[8] H.-G. Li, G.-Q. Wu, X.-G. Hu, J. Zhang, A. Li, and X. Wu, "Kmeans clustering with bagging and MapReduce," in Proceedings of the 44th Hawaii International Conference on System Sciences (HICSS '10), pp. 1– 8, January 2011.

[9] J.-H. Bose,¨ A. Andrzejak, and M. Hogqvist,¨ “Beyond online aggregation: parallel and incremental data mining with online mapreduce,” in Proceedings of the Workshop on Massive Data Analytics on the Cloud (MDAC ’10), pp. 1–6, April 2010.

[10] C. T. Chu, S. K. Kim, Y. A. Lin et al., “Map reduce for machine learning on multicore,” in Advances in Neural Information Processing Systems 19, pp. 281–288, 2006.

[11] F. Chierichetti, R. Kumar, and A. Tomkins, “Max-cover in mapreduce,” in Proceedings of the 19th International World Wide Web Conference (WWW ’10), pp. 231–240, April 2010.

[12] A. Clifton and K. J. Lundquist, “Data clustering reveals climate impacts on local wind phenomena,” Journal of Applied Meteo rology and Climatology, vol. 51, pp. 1547–1557, 2012.

[13] G. Amit and D. Sara, “A survey on cloud computing,” Tech. Rep. CS 508, University of British Columbia, 2009.

[14] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, “Evaluating MapReduce for multi-core and multiprocessor systems,” in Proceedings of the 13th IEEE International Symposium on High Performance Computer Architecture (HPCA ’07), pp. 13–24, Phoenix, Ariz, USA, February 2007.

[15] R. Lammel,¨ “Google’s MapReduce programming model - Revis-ited,” Science of Computer Programming, vol. 68, no. 3, pp. 208– 237, 2007.

[16] B. He, W. Fang, Q. Luo, N. K. Govindaraju, and T. Wang, “Mars: a MapReduce framework on graphics processors,” in Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques (PACT ’08), pp. 260– 269, ACM, New York, NY,USA, October 2008.

[17] J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.

[18] T. White, Hadoop: The Definitive Guide, O’Reilly Media, Inc., 2009.

[19] X. Xu, J. Jager,¨ and H.-P. Kriegel, “A fast parallel clustering algorithm for large spatial databases,” Data Mining and Knowledge Discovery, vol. 3, no. 3, pp. 263–290, 1999.

[20] Wei Fang, V. S. Sheng, XueZhi Wen, and Wubin Pan, “Meteorological Data Analysis Using MapReduce”, The Scientific World Journal , 2014.

[21] http://a-little-book-of-r-for-timeseries.readthedocs.io/en/latest/src/timeserie