A Weather Prediction Model with Big Data

Ning Yang
Seidenberg School of CSIS
Pace University
New York, NY
Email: ny00685p@pace.edu

Lewis Westfall
Seidenberg School of CSIS
Pace University
New York, NY
Email:lw19277w@pace.edu

Ms. Preeti Dalvi
Seidenberg School of CSIS
Pace University
New York, NY
Email:pd21567n@pace.edu

Abstract—Weather prediction is the application of technology to predict the weather for a given location based on historical or current data as applicable. It has been getting a lot of attention due to the unexpected changes that have been occurring. A very large amount of weather data is recorded because of the number of sites, the number of data elements and because the observations are recorded hourly. Using traditional methods to analyze the data has become very resource dependent. This amount of data makes weather analysis and prediction an ideal candidate for Big Data technology. This project focuses on predicting the target location's barometric pressure by using the data recorded from sites around it and uses a novel regression model, which combines several important factors of weather data to enhance the prediction rate.
Key words: Weather Prediction; Big Data; Regression Model; Weather Pressure.

1. Introduction

Weather prediction is the application of science and technology to predict the conditions of the atmosphere for a given location and time by collecting quantitative data about the current state of the atmosphere at a given place [1]. The prediction of climate has proven to be very important and useful for it is always relating to the judgment of government to help protect the growth of agricultural crops, the warning of hurricanes, floods and so on. However, the conditions' analysis is usually combined with large volume of data, which made it a significant challenge and a good candidate for Big data technology.

MapReduce is one of the best-known Big Data methods for turning raw data into useful information. It is a method for taking large data sets and performing computations on it across multiple computers, in parallel. It serve as a model for how to program, and is often used to refer to the actual implementation of this model [2]. Thus, it is used as general model in many technologies such as Apache Hadoop which is perhaps the most influential and established tool for analyzing big data.

The Hadoop with MapReduce has been the leading open source framework for many years while the Apache Spark [3] has become the lingua franca of big data analytics for many organizations because of its fast data processing and ease of development. The Apache Spark combined with R platform is used for our work due to its real-time streamed data analyze and generality.

The weather data is always collected by the weather station directly. However, can we get the weather data quickly if the sensors at a weather station are not functioning for some reason? The weather information in an area usually affect each other. Thus, we make a hypothesis that the weather conditions can be calculated for a site from weather data collected from the stations around the target site using a model constructed from the data analysis.

The goal in our project is to predict the target locations weather information by a new model using the weather data sets from the cities around the destination. We will analyze the data and create a model for the prediction. A novel regression line model, combined with standard distance is supplied to describe the information of the test. We will also do a comparison between the new and original models to show the improvements of our work.

The remainder of this paper is organized as follows:

- Section two introduces recent research based on the weather prediction with different methods and models.
- Section three introduces the backgrounds of this research.
- Section four introduces the methods and models we use.
- The rest of the sections introduce our experiments on this project, conclusion, our enhancement and future works.

2. Literature Review

Weather prediction is a hot topic. Much related research has been done with different models and methods. This section covers a number of papers in this field, which were

published recently. A detailed analysis is supplied for each study to show their advantages and improvements.

In [4], the author introduces a novel Train Delay Prediction System whose function is to provide an integrated and holistic view of operational performance and enable high levels of rail operations efficiency of the Italian railway network. The weather data is working as the exogenous sources combined with historical data of train movements to build this reliable and robust data-driven model. Four different methods such as RFI systems, Random Forest algorithm, kerneled version of RLS and Extreme learning Machine are supplied to make the comparisons to prove the novel RFI with weather data perform up to twice better than the current state-of-the-art-methodologies. Meanwhile, more data sources would be add to the system as the conditions to improve the new system in the future.

Khalid and Mazlina supply a Big Data Prediction Framework [5] for temperature based on MapReduce to improve the accurate prediction rate of weather. This new framework gives faster processing of data due to its highly parallelization and distributed characteristics. The experiment environment is based on a cluster with three PCs (one as name Node while the others work as data Node). Their result shows that the additional of more systems to the distributed network gives faster processing of the data processing and the combination of big data technology and commercial industry has the potential to greatly enhance the weather forecast's efficiency.

Dr.Asha and Shobha [6] describe a data mining study of agricultural meteorological patterns collected from the meteorological center of Bengaluru district. K-means and Hierarchical clustering techniques are used to exact patterns and obtain results, which play a crucial role in the decision making for sustainable agriculture. The result analysis is clustered by the types of crops such as mango, grapes, potatoes, and so on for each one has different conditions for analysis. The results show that the cluster techniques are effective to predict the information of weather details and the Hierarchical algorithm performs better than K-means.

In [7], the authors introduce a project that aims to forecast the chances of rainfall by using predictive analysis in Hadoop. This model captures relationships among many factors in the data to assign a score or weight pattern for future rainfall prediction by using historical data. The process is in an efficient manner for the large volume of data can be well processed by the big data techniques. The main method for the analysis is classify the weather type by using Naive Bayes with the weather data attribute of humidity. The system design and the plot of mean, maximum and precipitation parameter of humidity are supplied to improve that the more weather information can be efficiently predicted by using Naive Bayes in Hadoop Framework.

Po-Chen Chen and Mladen Kezunovic demonstrate a way to utilize historical weather data and climate change projections in a large (macro) geographical area to predict future electric load patterns in a relatively small (micro) geographical area in their paper [8]. The impact of temperature rising is based on the load while the deviations of the result is large depending on the changing data. Both the data and model are from Coupled Model Intercomparison Project 5. The future and historical peak load consumption are both supplied to show that the novel framework is proposed and its efficiency is higher than most models, which are also aimed at this research for it shows larger numbers for the temperature's increase.

Vincent and Katherine describe a method for lossy compression of weather data [9] by representing the data as a sparse and adaptive subset. This output is used for solving an optimization problem for the minimal loss of information. The series methods are combined with Numerical Weather Prediction (NWP) to support those users who require substantially smaller data sets in exchange for some loss of information. Mean squared error (MSE) and peak-signal to noise ratio (PSNR) are used to judge the performance of various algorithms while the long running Genetic Algorithm (GA) gives the highest PSNR and the least loss of information. They enhance the result that the large data sets can be reduced to the size of an email attachment and the loss of information can be minimized by use of adaptive sampling.

In [10], the author proposes to use a geostatistical interpolation technique called Kriging to create short term weather predictions from scattered weather observations derived from surveillance data. This method can accurately capture the spatio-temporal distribution of the temperature and wind data, which allows obtaining high-quality local, short-term weather predictions and providing at the same time a measure of the uncertainty associated with the prediction. For their research, the UK-ST framework has been particularized with the trend models and can be used to predict the spatio-temporal variograms focusing on temperature and wind speeds. Many methods are supplied while the cross-validation is supplied to prove that wind and temperature models generated using this technique can accurately capture the spatio-temporal distribution of these weather variables.

3. Background

Our project combines Apache Spark and a linear regression model. This section will introduce the basic information of Big Data and Regression model we use in our research. 3.1. Big Data

Big data is high volume, velocity and variety data set module [11]. It is a term , which describes the large volume of data, both structured and unstructured, which inundates a business on a day-to-day basis. However, it's not the amount of data, it is what the organizations does with the data and the related analysis based on that data. Big data

can be analyzed for insights, which can lead to better decisions and strategic business moves.

Big data originals model is 3V model as follows:

- Volume: Big data implies enormous volumes of data. The volume of big data to be analyzed is massive.
- Variety: The variety means that the data sets have many different types of data, including structured and unstructured.
- Velocity: It refers to the speed at which new data is generated and the speed at which data moves around.

Further research on big data has exposed two new characteristics and these have been added into the 3V model. The new characteristics are Veracity (Trustworthiness of data) and Value (It is well and good to have access to big data, but if it doesn't have value, it is useless) to make the 5V model now (shown in figure.1).

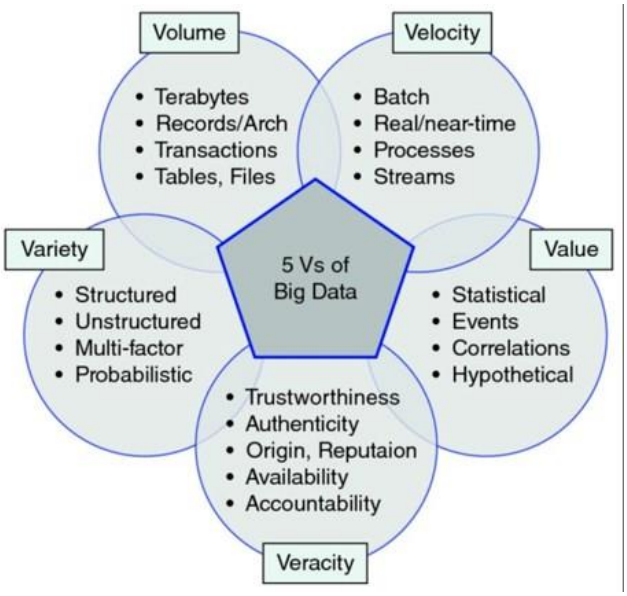


Figure 1. Big Data's 5V Model

3.2. Apache Spark

Apache Spark [12] is a fast and general engine for largescale data processing which is known for its speed, ease of use and generality and can be run on many platforms such as Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3. It contains five components (shown in figure.2):

- Apache Spark Core: The foundation for parallel and distributed processing of large data sets. It provides In-Memory computing and referencing data sets in external storage systems.
- Spark SQL: The component on top of Spark Core which introduces a new data abstraction called Schema RDD. RDD: Resilient Distributed Data sets.
- Spark Streaming: It ingests data in mini-batches and performs RDD transformations on those minibatches of data.
- MLib(Machine learning Library)is a low-level machine learning library that can be called from Scala, Python and Java programming languages.
- GraphX: It is an Apache Sparks API for graphs and graph-parallel computation. This component supports multiple use cases like social network analysis, recommendation and fraud detection. Other graph databases can also be used but they require several systems to create the entire computation pipeline.

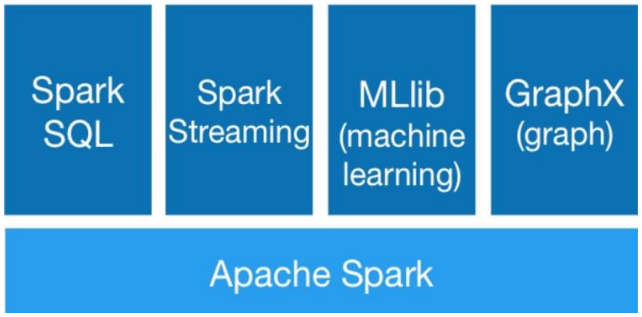


Figure 2. Apache Spark's Ecosystem

Resilient Distributed Data sets (RDD) [13] is the fundamental data structure of Spark. It is an immutable distributed collection of objects. Each data set in RDD is divided into logical partitions, which may be computed on different nodes of the cluster. RDDs can be described as:

- Immutable distributed collection of objects.
- A read-only, logically partitioned collection of record.
- Many types of classes are supported by RDD such as python, java and so on
- Parallelizing an existing collection or Referencing a data set in an external storage system are the two ways to create an RDD.

RDD supports in-memory processing computation. This means, it stores the state of memory as an object across the jobs and the object is shareable between those jobs. Data sharing in memory is 10 to 100 times faster than network and disk, which is usually used by MapReduce. Thus, the Spark's processing speed is much faster than MapReduce, which spends more than 90 percent of the time doing HDFS read-write operations.

3.3. Regression model

Regression analysis [14] is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables.

A function of the independent variables called the regression function is to be estimated. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the prediction of the regression function using a probability distribution. The function of linear regression can be described as:

- Given a variable y and a number of variables X_1, ..., X_p that may be related to y, linear regression analysis can be applied to quantify the strength of the relationship between y and the X_j, to assess which X_j may have no relationship with y at all, and to identify which subsets of the X_j contain redundant information about y (shown in formula 1).

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\cdots+\beta_p x_p+\epsilon \quad (1)$$

4. Methodology and Data set

Our project aims at predict the barometric station pressure at the target, dependent, location from data recorded at the independent weather stations which are around the target. In order to improve the prediction rate, we include the distance of each independent weather station to the dependent weather station to the model. This section will describe the information in the data sets and the model we built.

4.1. Information of data sets

The data set we used is from the National Oceanic and Atmospheric Administration (NOAA). It contains weather information for 2016 from eight different cities in New York State; Binghamton, Buffalo, Cattaraugus, Dansville, Penn Yan, Rochester, Syracuse, and Wellsville (shown in figure.3). Our target location is the Penn Yan weather station, which is marked with red hexagon. This data set has 20 different attributes, which are recorded hourly. After our analysis, we choose five attributes (HourlyDryBulbTemp, HourlyWetBulbTemp, HourlyDewPointTemp, HoulyRelativeHumidity and HourlyWindSpeed) as the parameter for our regression model analysis.

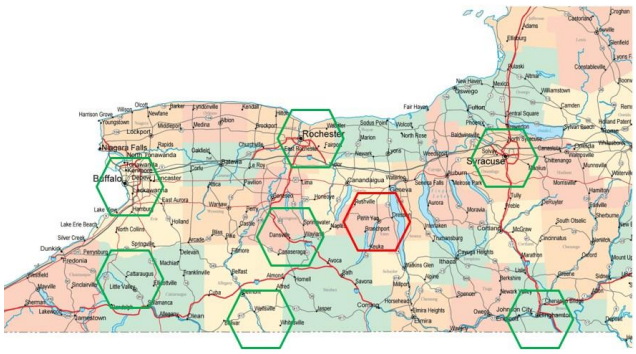


Figure 3. The information of weather stations

4.2. Model Design

For our project, we set three steps to get the final model:

1) For the hypothesis, the Pen Yann station's weather sensors is not functioning, thus we need to collect the required data from the other seven weather stations. We generated the regression model for each weather station, the relation is between the station pressure and the five attribute we choose, taking Binghamton's data set as an example, the information of its coefficients of model is shown in figure 3. The resulting equation is:

```
Call:
lm(formula = BING$HOURLYStationPressure ~ BING$HOURLYDRYBULBTEMPF +
    BING$HOURLYWETBULBTEMPF + BING$HOURLYDewPointTempF + BING$HOURLYRelativeHumic
 +
    BING$HOURLYWindSpeed)

Residuals:
     Min       1Q   Median       3Q      Max
-0.78057 -0.11890  0.01184  0.13376  0.60844

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   28.8104659  0.0428175 672.866  < 2e-16 ***
BING$HOURLYDRYBULBTEMPF       -0.0108434  0.0013486  -8.041 9.62e-16 ***
BING$HOURLYWETBULBTEMPF        0.0107605  0.0019360   5.558 2.77e-08 ***
BING$HOURLYDewPointTempF       0.0004340  0.0014922   0.291    0.771
BING$HOURLYRelativeHumidity -0.0049781  0.0004550 -10.942  < 2e-16 ***
BING$HOURLYWindSpeed        -0.0179267  0.0003798 -47.204  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1957 on 14404 degrees of freedom
  (658 observations deleted due to missingness)
Multiple R-squared:  0.1849,    Adjusted R-squared:  0.1846
F-statistic: 653.5 on 5 and 14404 DF,  p-value: < 2.2e-16
```

Figure 4. Coefficients of Binghamton

$$y_1=28.8104659-0.108434x_1+0.0107605x_2+0.000434x_3-0.0049781x_4-0.0179267x_5 \quad (2)$$

2) We use different cities' regression model to test each city's pressure. Then we combined them together with the target location's pressure attributes to build the new data sets that are used to calculate target location's pressure.
3) For the new data set, we calculate the distance for each cities using its latitudes and longitude, and set the $\log_{100}$ (distance) combined with pressure from each cities as the parameter to build the new model. The coefficients are shown in figure 4.

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     20.068208   0.755231  26.572  < 2e-16 ***
pressure$ROHOURLYStationPressure    0.288432   0.009255  31.166  < 2e-16 ***
pressure$SYRHOURLYStationPressure  -0.050704   0.008910  -5.691 1.29e-08 ***
pressure$WELLHOURLYStationPressure  0.051953   0.010410   4.991 6.10e-07 ***
pressure$BINGHOURLYStationPressure  0.021460   0.010612   2.022   0.0432 *
pressure$BUFFHOURLYStationPressure  0.037451   0.008736   4.287 1.83e-05 ***
pressure$CATAHOURLYStationPressure  0.065200   0.011431   5.704 1.20e-08 ***
pressure$DANSHOURLYStationPressure -0.099830   0.011532  -8.656  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5. Coefficients of the new model

The resulting model is:

$$\alpha = 20.06821 + 0.24991y_1/\log_{100}(d_1) - 0.04926y_2/\log_{100}(d_2) + 0.05130y_3/\log_{100}(d_3) + 0.02154y_4/\log_{100}(d_4) + 0.04327y_5/\log_{100}(d_5) + 0.06970y_6/\log_{100}(d_6) - 0.07519y_7/\log_{100}(d_7) \quad (3)$$

5. Experiment results and analysis

In this section, we will introduce the characteristic analysis of this model and the prediction result what we got.

5.1. Model check

The quality of a regression model is usually judged from many aspects. We analyze our model from three ways to check its efficiency. The figure.6 shows the normal Quantile Quantile plot (Q Q) of this model which can reflect the distribution of the predict pressure's probability distributions against the original data sets. We can find that the linearity of points reflects that most of the data fits this model well and is normally distributed.

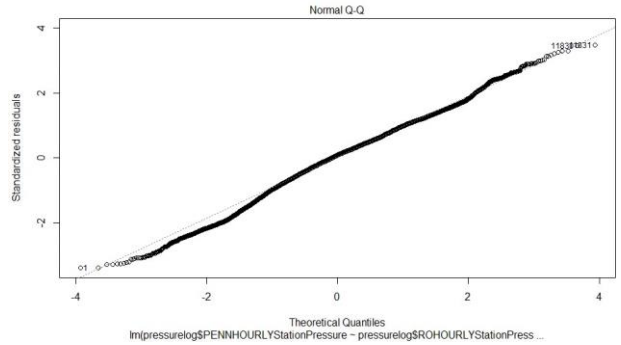


Figure 6. The normal QQ of model

The figure.7 shows the information of residuals vs fitted for this model. we can find that the model mostly meet the liner model assumption and the residuals are not too far away from 0, the abstract standardized values are less than 0.5 to make sure the assumption's linearity and homoscedasticity.

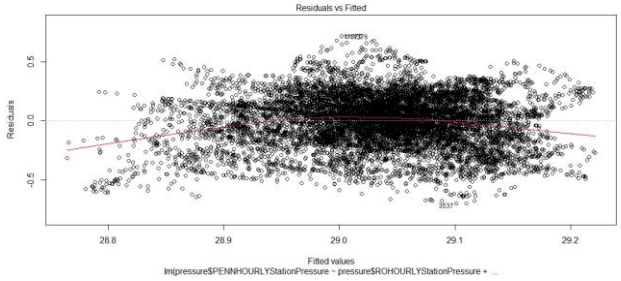


Figure 7. The residuals vs fitted

5.2. Result analysis

We use the model to get the prediction station pressure while the result cannot be matched exactly, the figure.8 shows the range of the differences between the calculated values and the actual values.

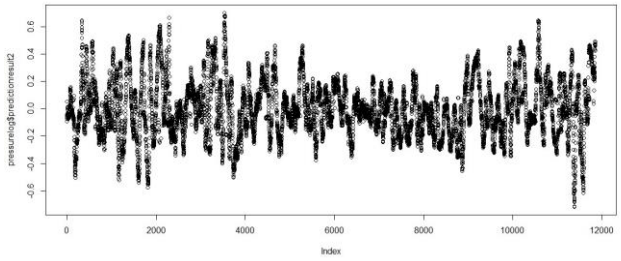


Figure 8. The range analysis of the result

We first selected the error limit of 0.25 as our criteria for having an accurate result. Figure.9 shows the distribution of the information of detection and we achieved a 78% accuracy rate.

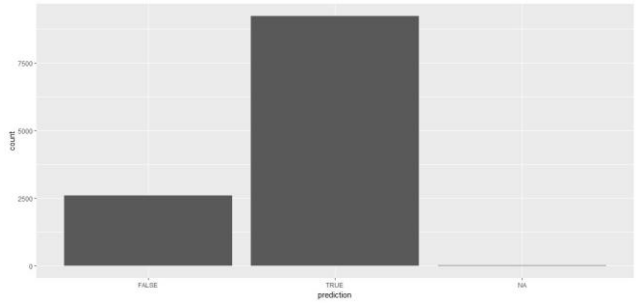


Figure 9. Prediction Rate

In order to get a more realistic error limit we did a K-Means clustering on the range of differences shown in Figure.10 We found that we had two clusters with centers at 0.09 and 0.32. We set the error limit to 0.32 to improve

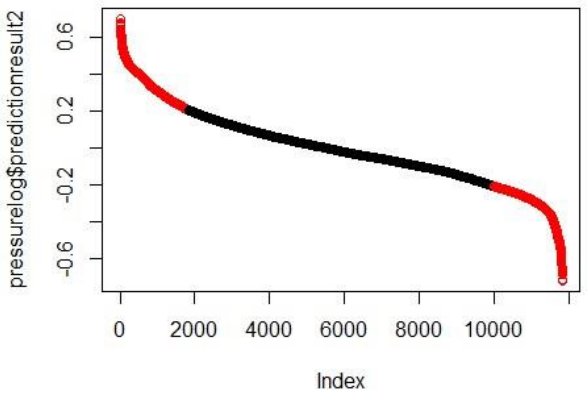
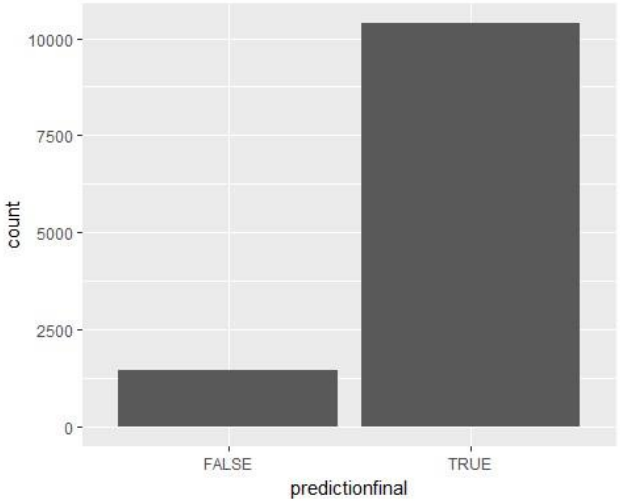


Figure 10. The cluster by using k-means

the accuracy rate. The distributed information is shown in Figure.11 and the prediction rate is about 88.78%.



distribution of Truth table

6. Conclusion and Future Work

From the result and model analysis, we can prove that our model reflects the target location's weather information with the prediction rate of 88.78%. Using big data technology, we enhanced the efficiency of dealing with large amount of weather data. However, the weather information prediction is only a basic use of these techniques. There are also many other weather feature predictions to explore. For our future work, we will continue the related researches in two ways. We will build models to predict other weather metrics, such as rain, fog and so on, to see if the regression model can make good predictions on other conditions.

[8] Po-Chen Chen; Mladen Kezunovic,"Load consumption prediction utilizing historical weather data and climate change projections", 2017 19th International Conference on Intelligent System Application to Power Systems (ISAP),Year: 2017,Pages: 1 - 6.

[9] Vincent van Leijen; Marijn Boone; Katherine Horgan,"Making numerical weather predictions portable compression of weather data for use in radar propagation modeling", 2017 USNC-URSI Radio Science Meeting (Joint with AP-S Symposium) ,Year: 2017 ,Pages: 1 - 2.

[10] R. Dalmau, M. Perez-Batlle, and X. Prats. "Estimation and prediction of weather variables from surveillance data using spatio-temporal Kriging". 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC). Year: 2017, Pages 1-8.

[11] Li, Jingui, Xuelian Lin, Xiaolong Cui, and Yue Ye., "Improving the shuffle of hadoop mapreduce". Cloud Computing Technology and Science (CloudCom), 2013 IEEE 5th International Conference on, vol.
1, pp. 266-273. IEEE, 2013

[12] https://www.tutorialspoint.com/apache spark/apache spark rdd.html

[13] https://en.wikipedia.org/wiki/Regression analysis Figure 11. The

Meanwhile, we will also use the historical data in one place to find the rules to predict future weather information.

References

[1] https://en.wikipedia.org/wiki/Weather forecasting

[2] https://en.wikipedia.org/wiki/MapReduce

[3] R.D. Schneider, Hadoop for Dummies Special Edition, John Wiley and Sons Canada, 978-1-118-25051-8, 2012

[4] Luca Oneto; Emanuele Fumeo; Giorgio Clerico; Renzo Canepa; Federico Papa; Carlo Dambra; Nadia Mazzino; Davide Anguita,"Advanced Analytics for Train Delay Prediction Systems by Including Exogenous Weather Data ",2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)

[5] Khalid Adam Ismail; Mazlina Abdul Majid; Jasni Mohamed Zain; Noor Akma Abu Bakar,"Big Data prediction framework for weather Temperature based on MapReduce algorithm", 2016 IEEE Conference on Open Systems (ICOS) Year: 2016

[6] N. Shobha; T. Asha, "Monitoring weather based meteorological data: Clustering approach for analysis " .2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA),Year: 2017,Pages: 75 - 81

[7] Sunil Navadia; Pintukumar Yadav; Jobin Thomas; Shakila Shaikh, "Weather prediction: A novel approach for measuring and analyzing weather data ",2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC),Year: 2017,Pages: 414 - 417.