2018 International Conference on Advances in Computing and Communication Engineering (ICACCE-2018)
Paris, France 22-23 June 2018

Analysis of Weather Prediction using Machine Learning & Big Data

Shubham Madan[1], Praveen Kumar[2], Seema Rawat[3], Tanupriya Choudhury[4]
Amity University Noida, Uttar Pradesh, India[1,2,3], UPES Dehradun[4]
shubhmadan20@gmail.com, pkumar3@amity.edu, srawat1@amity.edu,tanupriya@ddn.upes.ac.in

Abstract : ***The whole world is plagued by the dynamical clement and their facet, to cut back this facet effects up to some extent there are several techniques and algorithms through which we will predict the weather on the ready reference along with respective context of given information from past years example temperature, dew, humidity air pressure and wind direction, . When doing the analysis of existing data from past few years we inculcated the proposed scheme or techniques which have a tendency to conclude that, machine learning paradigm and permits us to research the given set of knowledge and extract the helpful information from the given dataset, thus so as to grasp the unsteady patterns of climatic conditions, a prognosticative model is also persuaded. During this paper or scheme, we have a tendency to explore progressive statistical linear regression and support vector machine techniques of machine learning that teams' constant kind information sets along and to prefigure the forecast or weather prediction. Under the proposed scheme we have a tendency to inculcate the augmented algorithmic rule that provides approximate and nearby results to forecast the climate for the next 5 days and at the end results are calculated on the idea of mathematical and statistical decision tree and conditions vide confusion matrix for more appropriate and accurate forecasting using Big Data.***

Keywords: L*inear regression, support vector machine, decision tree, confusion matrix, machine learning, big data.*

1. Introduction

Big Data contains tremendous and mammoth information in the organized, semi-organized and unstructured manner. That is the reason it is extremely hard to process, oversee and store to this kind of information. As of overdue extraordinary sorts of mechanism, techniques and procedures are there to deal with Big Data. Data mining[3] using machine learning is one of them which we have utilized as a part of this paper to oversee climate related information and predict the forecast and certain condition of future weather. Under this scheme we suggest that how to utilized the data mining and in order retrieval of data using machine learning in the expectation of climate and forecasting of the weather. Presently and now a days, we the people of India experiencing changing bad weather, pollution and their reactions. Typically in horticulture field, ranchers are confronting numerous issues because of surprising climate conditions. Climate anticipating is straightforwardly rely on the regular particles display noticeable all around like (O3) Ozone, Nitrogen (NO2) dioxide, (CO2) Carbon Dioxide, (SO2) Sulfur dioxide and so on. In this paper we have concentrated on particular area i.e. Delhi. To decrease these reactions up to some degree there are numerous strategies and calculations through which we can foresee the climate on the premise of given information. Data mining using machine learning procedure is utilized as a part of Weather expectation process. Climate is best natural[2] requirement in each period of our human life. So climate anticipating is going excessively utilized as a part of many fields like Food security calamities, Agriculture and science. In prior years we have no correct thought regarding climate conditions. So back then, we confronted numerous issues in sustenance administration process, industry and agribusiness[8] field. In any case, now in the period of progression we have numerous approaches to discover climate conditions. This is the explanation for applying information mining procedures to locate the climate conditions using Big Data and its Eco-System [6] along with machine learning techniques vide linear regression and support vector machine.

Data mining using machine learning are the way toward extracting important data from the extensive informational collection. The procedure of concentrate important data portrayed[6] as information revelation that can be connected on any extensive informational index. The primary data mining systems using machine learning are Classification, Clustering, Association and Regression. The distinctive Data digging methods utilized for taking care of climate changing and measuring issue. Climate measuring issue incorporate expectation[7] of temperature, rain, mist, winds, and storm and so forth. Climate sensors gather information consistently at numerous areas and assemble tremendous information. Climate anticipating is dependably a major test since it is difficult to foresee the condition of the air for the forthcoming future since atmosphere dataset is

978-1-5386-4485-0/18/$31.00 ©2018 IEEE

capricious and again day to day changes as indicated by worldwide atmosphere changes in context to past scenarios. The information utilized is from the INDIA METEOROLOGICAL DEPARTMENT (IMD), the arrangement of dataset bolsters a rich arrangement of meteorological components, which are great contender for investigation with huge information since it is semi-organized and record situated. The term Big Data came around 2005, which implies datasets that are tremendous, moreover high in collection and speed, which makes them difficult to process using ordinary devices and frameworks. Huge information made colossal[4] business and social open doors in each field, empowering the revelation of beforehand shrouded designs and the advancement of new bits of knowledge to decide, running from web hunt to content proposal and computational scenarios. The term Big Data is presently utilized wherever in our everyday life and it is a present innovation and furthermore going to manage the world in future and has risen on the grounds that individuals and diverse organizations makes expanding utilization of information concentrated advancements. Huge information sizes are right now extending from a Terabyte to Zettabyte in a solitary informational collection. Like the physical universe, the advanced universe is huge. As per look into led by IDC, from 2005 to 2020, the advanced universe will develop from 130 Exabytes to 40,000 Exabyte's, or 40 trillion gigabytes. From now, the advanced universe will about twofold at regular intervals until 2020. As expressed by IBM, with machine-to-machine (M2M) correspondences, on the web/portable informal[10] communities and unavoidable handheld gadgets it makes 2.5 quintillion bytes of information in every day.

Attributes of Big data– Big Data has copious attributes detailed by n V's qualities. Collection of V's characteristics of the Big Data were gathered from numerous scientist's productions to have Nine V's characteristics (9V's attributes). These 9V's qualities are:

- **Veracity**: Enormous Data veracity alludes to the inclinations, commotion, and irregularity in information..
- **Variety**: Organized, semi-organized, and unstructured information other than content and more information composes have risen, for example, record, log, sound, and half and half information.
- **Velocity**: The developed or made data at a speedier pace than some time recently, in which the distinctive mediums of Big Data increment the yield matter.
- **Volume**: the measure of information is known as volume of information, where the measure of information keeps on detonating.
- **Validity**: the correct data or information that is exact for the utilize plan. Most probably, authorized data is used for deciding for legitimate choices.
- **Variability**: the data streams may be greatly incompatible with discontinuous peaks, frequent and occasion triggered peak data burdens can be trying to oversee, especially with the inclusion of unstructured data.
- **Volatility**: When maintenance period lapses, we can without much of a stretch crush it.
- **Visualization**: implies complex charts that can incorporate a few factors of information while as yet staying justifiable and lucid
- **Value** It has a low-esteem thickness because of extricating an incentive from monstrous information. Helpful information should be separated from any information write and from a colossal measure of information.

2. Related Work

Related works included a wide range of and fascinating systems to attempt to perform climate figures. While a lot of current determining innovation includes reenactments in light of material science and differential conditions, numerous new methodologies from computerized reasoning utilized essentially machine learning strategies, generally neural systems while others used models which had a probabilistic approach, for example, Bayesian systems. From 3 papers on climatic expectation from machine learning we inspected, 2 of 3 utilized neural systems while one utilized help vector machines. The most noticeable machine learning model is using neural systems for determining climate on account of the capacity to catch the indirect conditions of previous climate patterns and approaching climate setting, dissimilar to the straight relapse and practical relapse models that we utilized. This gives the upside of not accepting basic direct conditions of all highlights over our models. Approaches using neural systems , one [3] utilized a mixture demonstrate that utilized neural systems to show the material science behind climate estimating while the other [4] connected adapting all the more specifically to anticipating climate conditions. Likewise, the approach utilizing bolster vector machines [6] additionally connected the classifier straightforwardly for climate forecast yet was more restricted in scope than the neural system approaches. Different methodologies for climate guaging included utilizing Bayesian systems. One intriguing model [2] utilized Bayesian systems to model and make climate expectations however utilized a machine learning calculation to locate the most ideal Bayesian systems and parameters which was computationally costly due to the substantial measure of various conditions yet performed extremely well. Another approach [1] concentrated on a more particular instance of anticipating extreme climate for a particular topographical area which

Table 1: Sample data showing the 5 features.

Number	Name	Value
1	Classification	Clear
2	Maximum	Temperature (F) 57
3	Minimum	Temperature (F) 33
4	Mean Humidity	Humidity 43
5	Mean Pressure	Atmospheric Pressure in 30.13

restricted the requirement for calibrating Bayesian system conditions however was constrained in scope.

1.1 Hadoop

Hadoop is generally utilized as a part of enormous information apps, e.g., spam separating, organize looking, clickstream investigation, and social suggestion. Few illustrative cases are underneath. As proclaimed, Hadoop is run by Yahoo in many servers for helping items in administration at 4 server farms, e.g. searching and spam separating, and so on. At introduce, the greatest Hadoop bunch has around four thousand hubs, yet the quantity of hubs will be expanded to around ten thousand with the arrival of Hadoop 2.0. Around the same time, Facebook reported that their Hadoop bunch can process 100 PB information, which developed by 0.5 PB for every day as in November 2012. Some outstanding offices that utilization Hadoop to lead appropriated calculation are recorded in [13]. What's more, numerous organizations give Hadoop business execution as well as help, including Cloudera, IBM, MapR, EMC, and Oracle. As indicated by the Gartner Research, Bigdata Analytics is a slanting subject in 2014 [14]. Hadoop is an open system generally utilized for Bigdata Analytics. MapReduce is a programming worldview related with the Hadoop.

2 Literature Survey

A. Adamu Galadima portrays a short take a gander at the Arduino microcontroller and some of its applications and how it can be utilized as a part of learning. Arduino is an open source microcontroller utilized as a part of electronic prototyping. Arduino equipment and its segments might be taken a gander at. Programming and the Environment that Arduino keeps running on are both taken a gander at as well. A few applications will be taken as illustrations that can help make learning Arduino additionally fascinating. This can be utilized as a noteworthy method to urge understudies and others to take in more about gadgets and programming.

B. Jeffrey Cohen display information parallel calculations for advanced factual systems, with an emphasis on thickness strategies. At last, he responds on database framework includes that empower deft outline and adaptable calculation improvement utilizing both SQL and Map Reduce interfaces over an assortment of capacity instruments.

C. Brian Dolan display the outline rationality, methods and experience giving MAD examination to one of the world's biggest promoting systems at Fox Audience Network, utilizing the Green plum parallel database framework. We depict database plan approachs that help the light-footed working style of examiners in these settings.

D. R. P. Singh clarify why a cloud-based arrangement is required, depict our model usage, and investigate some case applications we have executed that show individual information[11] proprietorship, control, and examination. He address these issues by outlining and executing a cloud-based engineering that furnishes buyers with quick access and fine-grained control over their utilization information, and also the capacity To break down this information with calculations of their picking, including outsider applications that investigate that information in a protection saving style.

E. Jeffrey Dean depicts the essential programming model and gives a few cases. Many ware machines are run using Map and Reduce: numerous terabyte of data is formed due to Map Reduce calculations on a huge number [7] of machines

F. Panagiotis D. Diamantoulakis finds the usage of the Data Analytics in field like Smart Grid for management of energy dynamically. There is a 2 way flow among suppliers and consumers of power and data for optimizing power for economic efficiency, security, tenability. DEM or dynamic energy management is promoted by this infrastructure for consumers and producers of micro energy . Reduction of cost of power by user participation is an important part.

G. L. Aniello investigates the possibility of a structure utilizing various information sources to enhance assurance capacities of CIs. Difficulties and openings are examined along three fundamental research bearings: I) utilization of particular and heterogeneous information sources, ii) checking with versatile

granularity, and iii) assault demonstrating and runtime mix of various information examination procedures.

4. Proposed Work

The most outrageous temperature, slightest temperature, mean clamminess, mean barometrical weight, and atmosphere gathering for consistently from year 1996 to 2017 for Delhi, India were gained from Weather Department website. [10][11] Primitively, there are 9 atmosphere orders: clear, scattered fogs, to some degree shady, generally shady, dimness, overcast, rain, tempest, and snow. Since an extensive parcel of these requests are practically identical and some are meagrely populated, these were diminished to four atmosphere groupings by joining scattered fogs and not entirely shady into sensibly shady; generally shady, foggy, and shady into extraordinarily shady; and rain, tempest, and precipitation instead of snow. Past years data were used to set up the counts, and the latest years data acted like test set and the alluded data for 1st month using the table 1 depicted parameters.

Number	Name	Value
1	Classification	Clear
2	Maximum	Temperature (F) 57
3	Minimum	Temperature (F) 33
4	Mean Humidity	Humidity 43
5	Mean Pressure	Atmospheric Pressure in 30.13

Table 1 : Parameters for Regression and Classification

The essential count which is used was straight backslide, that tries to suspect the temperature which is high and low as an immediate blend of the attributes. Since straight backslide can't be used with gathering data, this computation did not use the atmosphere course of action of consistently [13]. As needs be, just 8 attributes are utilized: the best temperature, minimum temperature, mean moistness, and mean climatic weight for each of the past two days. In this way, i-th join of consistent days, x (I) ∈ R9 is a 9 dimension component, where $x_0 = 1$ is portrayed as the square term. Let y (I) ∈ R 14 imply the 14-dimensional vector that contains these sums for the I-th match of progressive days utilizing direct relapse and further utilizing help vector machine arrangement limit the blunder work utilizing:

$$\frac{1}{2} w^T w - v\rho + \frac{1}{N}\sum_{i=1}^{N} \xi_i$$

subject to the constraints:

$$y_i\left(w^T \phi(x_i) + b\right) \geq \rho - \xi_i, \xi_i \geq 0, i = 1,\ldots,N \text{ and } \rho \geq 0$$

For this type of SVM the error function is:

$$\frac{1}{2} w^T w + C\sum_{i=1}^{N} \xi_i + C\sum_{i=1}^{N} \xi^*{}_i$$

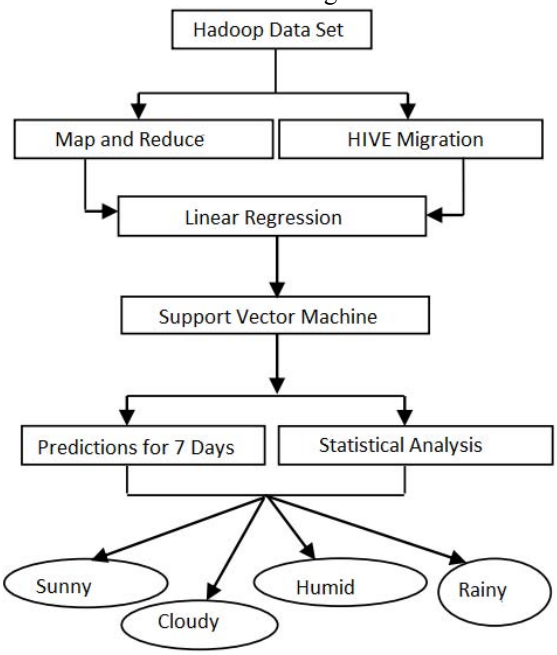
which we minimize subject to:

$$w^T \phi(x_i) + b - y_i \leq \varepsilon + \xi_i^*$$

$$y_i - w^T \phi(x_i) - b_i \leq \varepsilon + \xi_i$$

$$\xi_i, \xi_i^* \geq 0, i = 1,\ldots,N$$

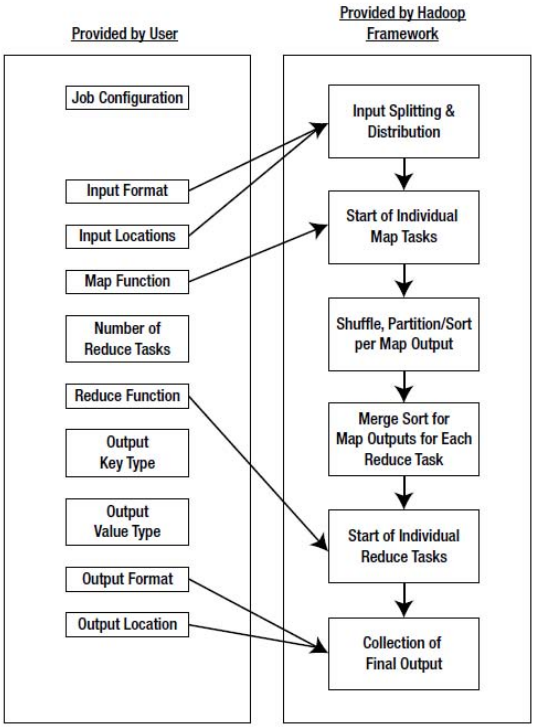
The work flow model of the proposed scheme is as under vide Figure 1 :



4. Design and Implementation

To gain the desired goals and results in proposed scheme the probalastic scenarios i.e. linear regression and SVM have be used via Big Data MapReduce. The below steps depicts the workflow and implementation of proposed scheme.

Step 1. Map Reduce using Big Data (Hadoop)



Step 2 Linear Regression :

In totality, dataset was obtained from 6100 to 7800 (records from meteorological department is obtained for regression) for at slightest seven attributes are regressed by which waning combinations were calibrated under this scheme. In erstwhile expression, of the 7800 rows forming the data cluster is formerly selected for use in this study below depicts the linear regression model.

The righteousness of fit [7] character for the model calibrations are obtainable in below equation, and the calibrated coefficients are shown in table 4. However presents standard error (S_e) calculated as:-

$$S_e = \sqrt{\frac{1}{n-m}\sum(y-\hat{y})^2}$$

where n is the number of observations,
m is the number of coefficients or exponents being calibrated,
y is the observed discharge (from the PeakFQ output), and
$\hat{y}$ is the predicted output calibrated by the regression tool.

Standard deviation (S_y) is calculated as

$$, = \sqrt{\frac{1}{n-1}\sum(y-\bar{y})^2},$$

where $\bar{y}$ is the mean of the discharges for the return period (T).

Explained variance (R^2) is calculated as

$$R^2 = \frac{1}{n^2 \cdot S_e^2 \cdot S_x^2}\left[\sum(y-\hat{y})\cdot(y-\bar{y})\right]^2$$

where

$$S_x = \sqrt{\frac{1}{n-1}\sum(\hat{y}-\bar{\hat{y}})^2},$$

in which $\bar{\hat{y}}$ is the mean of the predicted discharges for the return period

Step 3 Support Vector Machine :

1: Input: $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$, C, ϵ
2: $\mathcal{W} \leftarrow \emptyset$
3: **repeat**
4: $(\mathbf{w}, \xi) \leftarrow \text{argmin}_{\mathbf{w},\xi\geq 0} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\xi$
s.t. $\forall(\mathbf{c}^+,\mathbf{c}^-)\in\mathcal{W}$: $\frac{1}{m}\mathbf{w}^T\sum_{i=1}^{n}(c_i^+ - c_i^-)\mathbf{x}_i \geq \frac{1}{2m}\sum_{i=1}^{n}(c_i^+ + c_i^-) - \xi$
5: sort S by decreasing $\mathbf{w}^T\mathbf{x}_i$
6: $\mathbf{c}^+ \leftarrow 0; \mathbf{c}^- \leftarrow 0$
7: $n_r \leftarrow$ number of examples with $y_i = r$
8: **for** $r = 2, \ldots, R$ **do**
9: $i \leftarrow 1; j \leftarrow 1; a \leftarrow 0; b \leftarrow 0$
10: **while** $i \leq n$ **do**
11: **if** $y_i = r$ **then**
12: **while** $(j \leq n) \wedge (\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j < 1)$ **do**
13: **if** $y_j < r$ **then**
14: $b++; c_j^- \leftarrow c_j^- + (n_r - a + 1)$
15: **end if**
16: $j++$
17: **end while**
18: $a++; c_i^+ \leftarrow c_i^+ + b$
19: **end if**
20: $i++$
21: **end while**
22: **end for**
23: $\mathcal{W} \leftarrow \mathcal{W} \cup \{(\mathbf{c}^+, \mathbf{c}^-)\}$
24: **until** $\frac{1}{2m}\sum_{i=1}^{n}(c_i^+ + c_i^-) - \frac{1}{m}\sum_{i=1}^{n}(c_i^+ - c_i^-)(w^T\mathbf{x}_i) \leq \xi + \epsilon$
25: return$(\mathbf{w}, \xi)$

Results

DAY	DESCRIPTION	HIGH / LOW	PRECIP	WIND	HUMIDITY
WED 31-Jan	Fog Early / Clearing Late	--11°	0%	W 7 km/h	78%
THU 01-Feb	Sunny	26°10°	0%	WNW 12 km/h	57%
FRI 02-Feb	Mostly Sunny	25°11°	0%	WNW 9 km/h	66%
SAT 03-Feb	Sunny	25°9°	0%	NE 8 km/h	64%
SUN 04-Feb	Sunny	23°8°	0%	NW 15 km/h	55%
MON 05-Feb	Mostly Sunny	22°10°	0%	WNW 15 km/h	51%
TUE 06-Feb	Cloudy	23°12°	0%	WNW 16 km/h	48%
WED 07-Feb	AM Clouds / PM Sun	23°11°	0%	W 15 km/h	48%
THU 08-Feb	Partly Cloudy	24°11°	10%	WNW 12 km/h	64%
FRI 09-Feb	Partly Cloudy	23°10°	10%	WNW 11 km/h	70%
SAT 10-Feb	Mostly Sunny	23°10°	20%	WSW 9 km/h	75%
SUN 11-Feb	Mostly Sunny	24°10°	20%	WSW 9 km/h	72%
MON 12-Feb	Partly Cloudy	24°11°	0%	W 10 km/h	70%
TUE 13-Feb	Mostly Sunny	25°12°	0%	WNW 12 km/h	66%
WED 14-Feb	Partly Cloudy	26°14°	0%	WNW 13 km/h	57%

5. Conclusion and Future Work.

Both machine learning algorithms using hadoop lead realistic perfection were outflanked by proficient climate or weather determining directions or forecasting, in spite of the fact that the error in their execution diminished altogether for later days approx next 5 days, demonstrating that over longer timeframes, our models may beat proficient ones. Direct relapse turned out to be a low inclination, high change display while useful relapse ended up being to be a high predisposition, low difference demonstrate.

Results are intrinsically a high and accurate as demonstrated as it is steady for exceptions and forecasting, so one approach to enhance the straight relapse show is by accumulation of more information using linear regression and SVM. Showing that the decision of model was efficient and effective that its expectations can be enhanced by promote accumulation of information under the proposed scheme. For future scope the same can be incorporated over apache spark for concurrent prediction of weather whereas the same can be compare with the results obtained from sensors.

References

[1] Abramson, Bruce, et al. ”Hailfinder: A Bayesian system for forecasting severe weather.”International Journal of Forecasting12.1 (1996): 57-71.

[2] Cofno, Antonio S., et al. ”Bayesian networks for probabilistic weather prediction.”15th Eureopean Conference on Artificial Intelligence (ECAI). 2002.

[3] Krasnopolsky, Vladimir M., and Michael S. FoxRabinovitz. ”Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction.”Neural Networks19.2 (2006): 122-134.

[4] Lai, Loi Lei, et al. ”Intelligent weather forecast.”Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on. Vol. 7. IEEE, 2004.

[5] Ng, Andrew. ”CS229 Lecture Notes Supervised Learning” 2016.

[6] Radhika, Y., and M. Shashi. ”Atmospheric temperature prediction using support vector machines.”International Journal of Computer Theory and Engineering1.1 (2009): 55.

[7] ”Stanford, CA” in Weather Underground, The Weather Company, 2016. [Online]. Available: https://www.wunderground.com/us/ca/paloalto/zmw: 94305.1.99999. Accessed: Nov 20, 2016.

[8] Gupta, Subham Kumar, Seema Rawat, and Praveen Kumar. "A novel based security architecture of cloud computing." In Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2014 3rd International Conference on, pp. 1-6. IEEE, 2014.

[9] Saini, Parag, Tanupriya Choudhury, Praveen Kumar, and Seema Rawat. "Proposal and implementation of a novel scheme for image and emotion recognition using Hadoop." In Smart Technologies For Smart Nation (SmartTechCon), 2017 International Conference On, pp. 1358-1363. IEEE, 2017.

[10] Weather.com,http://www.weather.com

[11] Kadambari, Sanchita, Seema Rawat, and Praveen Kumar. "A Comprehensive Study on Big Data and Its Future Opportunities." In Proceedings of the 2014 Fourth International Conference on Advanced Computing & Communication Technologies, pp. 277-281. IEEE Computer Society, 2014.

[12] Gupta, Subham Kumar, Seema Rawat, and Praveen Kumar. "A novel based security architecture of cloud computing." In Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions), 2014 3rd International Conference on, pp. 1-6. IEEE, 2014.

[13] Wiki (2013). Applications and organizations using hadoop. http://wiki.apache.org/hadoop/PoweredBy

[14] Gartner Research Cycle 2014, http://www.gartner .com

[15] K. Morton, M. Balazinska and D. Grossman, “Paratimer: a progress indicator for MapReduce DAGs”, In Proceedings of the 2010 international conference on Management of data, 2010, pp.507–518.

[16] Lu, Wei, et al. “Efficient processing of k nearest neighbor joins using MapReduce”, Proceedings of the VLDB Endowment, Vol. 5, NO. 10, 2012, pp. 1016-1027.

[17] J. Dean and S. Ghemawat, “Mapreduce: simplied data processing on large clusters”, in OSDI 2004: Proceedings of 6th Symposium on Operating System Design and Implementation.