Sales data analysis using Hadoop and forecasting based on time-series ARIMA Model

Md. Rajib Hossen*, Prof Dr.M M A Hashem†, Rubayet Rahman Rongon‡
Dept of Computer Science and Engineering
KUET,Khulna-9203
*Email: rajibcse2k10@gmail.com
†Email: mma_hashem@hotmail.com
‡Email: rubayetrongon@gmail.com

***Abstract*—Day-by-day data is growing faster and making crucial business decision based on data becoming difficult at the same time. We need to process large volumes of data to get business insights. Hadoop is one of the most used Big Data Analysis tools. In this paper, With the help of this distributed big data processing tool we overcame the problem of extracting relevant information from a huge data dump. We experimented with sales data of 1year. We feed the data to Hadoop File system and extracted time-series data from that stored data using Hadoop computing engine. To forecast from our learning data, we used time-series ARIMA Model. We fit our distilled time series data into a forecasting model, ARIMA. Then compared this model with other exponential equation i.e holt winter model. Based on the better model, we mathematically derived a forecasting equation. Using that equation, we can predict the future sales comparatively better than other models.**

***Keywords*—Big Data, Hadoop, MapReduce, Data Mining, Forecasting,ARIMA Model**

I. Introduction

Big Data is a broad term in the sense that it deals with data that meet 3 V properties. These are velocity, volume and variety. In details, we say that data must be in high volume, and increasing as fast as possible and final properties is called variety that means various types of data like audio, video, pure text, photo, gps data, sensor data etc. There are lots of tools and systems for processing big data nowadays but among all these methods Hadoop is one of the oldest big data tools. It has a distributed file system which is called Hadoop Distributed File Systems (HDFS). Hadoop also comes with a processing engine called MapReduce

A. Hadoop Architecture

Hadoop come's from Google's MapReduce and Google File System(GFS) [7]. It is a Java-based framework that can process huge data sets in a parallel and distributed environment. It is part of the Apache project sponsored by the Apache Software Foundation. It is mainly used to process terabytes of data in a cluster of nodes [6]. Its distributed system enables faster data transfers among nodes. It replicates data between nodes, so the failure of any nodes is not an issue [6]. The best thing of Hadoop is that it can run on commodity hardware's and it is horizontally scalable. Node failure is common in big data clusters and failure of one node in a multi-node cluster is tolerable and economical.

Figure 1 shows the overall system of Hadoop environment.

There is only one name node in an HDFS cluster. It works as a master node that store file system metadata and fs images. There are many Data Nodes which stores data. HDFS creates a file system namespace and allows users data to be kept in files. There is block size for file and each file is split into blocks and these blocks are kept in the data nodes. Name node defines the mappings of blocks to data nodes [7]. HDFS is designed in a way that it can store large files reliably in multiple machines in clusters.

1) Name Node: The master name node manages entire file systems by keeping the index of data location, control access to files, manage replication etc. Files and directories metadata are kept in name node and it manages operations like reading, writing, replicating. Name Node doesn't contain any data.

2) Data Node: Data node store data assigned by name node. At first data files are split into several blocks. Data nodes execute the read-write requests given by clients. Data nodes also execute commands of creating, deleting and replicating blocks of file instructed by name node.

B. Map-Reduce Programming Framework

In 2004 Google first introduced MapReduce, a framework for processing large datasets. MapReduce is scalable, fault-tolerant and efficient programming techniques for processing a large amount of data. [11] Discuss the MapReduce techniques, comparisons between other data analysis languages and implementation in Hadoop in details. There is a map function that generates key/value pair and reduce function that merges all intermediate value associated with the same key.

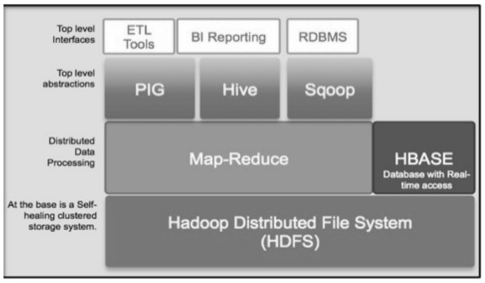
Top level Interfaces
ETL Tools
BI Reporting
RDBMS
Top level abstractions
PIG
Hive
Sqoop
Distributed Data Processing
Map-Reduce
HBASE
Database with Real-time access
At the base is a Self-healing clustered storage system.
Hadoop Distributed File System (HDFS)

Fig. 1. Hadoop architecture at high-level view

1) Map Step: The master node receives the input, split them into smaller sub-module and distributes them to worker nodes. A worker node may repeat this steps again and could make a multi-level tree structure. The worker nodes process the data and send the processed data back to the master node. Map takes one pair of data with a type in one data domain, and returns a list of pairs in a different domain:

Map (k1, v1) →list (K2, v2)

2) Reduce Step: The master nodes collect all the mapped data of the sub-problems and combine them to form the output the solution to the problem it was trying to solve. The reduce function then applied in parallel to each group which in turn produces a collection of values.

Reduce (K2, list (v2)) → list (v3)

C. Forecasting & ARIMA Model

[16] In statistics and econometric, and in time series analysis, there is a popular model called Auto Regressive integrated moving average (ARIMA) model. To better understand the data or to give a prediction, these time series models are used. When data show evidence of non-stationary then these model can be used to remove non-stationary.

ARIMA Models:

- Autoregressive (AR) process: Series current value depend on its previous values.
- Moving Average(MA) process: The current deviation from mean depends on previous deviations.
- Both combined form ARMA process
- Another part is named Integrated. So it becomes ARIMA model.

ARIMA is also known as Box-Jenkins approach. Details explanation of arima model has given in this website [18]. All the variations and forms of ARIMA model are explained here.

Some equations of ARIMA (p,d,q) models:

$$ARIMA(2,0,1) \quad Y_t = a_1Y_{t-1} + a_2Y_{t-2} + b_1\epsilon_{t-1} \tag{1}$$

$$ARIMA(3,0,1) \quad Y_t = a_1Y_{t-1} + a_2Y_{t-2} + a_3Y_{t-3} + b_1\epsilon_{t-1} \tag{2}$$

D. Box Jenkins Approach

[17] Box-Jenkins model applies to ARMA or ARIMA model to find the best fit of the time series data. The Box-jenkins provides a flowchart for ARIMA model estimating.

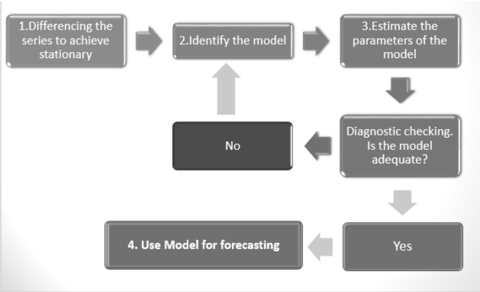
1.Differencing the series to achieve stationary
2.Identify the model
3.Estimate the parameters of the model
Diagnostic checking. Is the model adequate?
No
Yes
4. Use Model for forecasting

Fig. 2. Box-Jenkins Approach

1) Step 1: Stationarity: Dicky fuller test is used to check the stationary of data. Stationarity means mean and variance of data will be remain same for all time period.

Dicky Fuller Test

- p value has to be less than 0.05 or 5%
- if p value greater than 0.05 or 5% you accept the null hypothesis you conclude that time series has a unit root
- In that case I will first difference the series before proceding with analysis.

2) Step 2: Identification: Identification of order of p, q and d where p, q and d is for AR process, MA process and integration respectively. If the data is non-stationary then we must have to make it stationary. To do that we need to take differentiate of the data. The order to differentiate to obtain stationary is d.

After making the series stationary, we have to approximate the value of p and q. For this we use auto-correlation function (ACF) and Partial Auto-correlation Function(PACF). We can examine the ACF and PACF curve to help identify the proper number of lagged by y(AR) and ϵ (MA) terms.

3) ACF Curve: Auto-correlation is a correlation coefficient.It's not correlation between two different variables rather correlation between two values of same variable at time X_i and X_{i+k}. Correlation with lag1, lag2, lag3 etc. The ACF represents the degree of persistence over respective lags of a variable.

$$\rho_k = Y_k/Y_0 = co - variance at lag k/variance \tag{3}$$

$$\rho_k = \frac{E[(y_t - \mu)(y_{t-k} - \mu)]^2}{E[(y_t - \mu)^2]} \tag{4}$$

4) PACF curve: Partial regression coefficient - is the partial regression coefficient,θ_{kk} in the k_th order auto regression. Normally, partial correlation means correlation between two variables is the amount of correlation between them which is not determined by their mutual correlations with set of other variables.

$$Y_t = \theta_{k_1}Y_{t-1} + \theta_{k_2}Y_{t-2} + ... + \theta_{k_k}Y_{t-k} + \epsilon_t \tag{5}$$

TABLE I. PROPERTIES OF ACF AND PACF CURVE OF AR,MA AND ARMA PROCESS

Process	MA(q)	AR(p)	ARMA(p,q)
Auto Correlation Function	cuts off	Infinite,Tails Off	Infinite,Tails off,dampening exponentially
Partial Auto Correlation Funciton	Infinite, dampening exponentially	Cuts off	Infinite,Tails off,dampening exponentially

Partial correlation measures the degree of association between two random variables, with the effect of a set of controlling random variables removed.

Table I list the properties curve.

II. RELATED WORKS

Nandimath et al. [6] proposed a distributed way for storing data in small pieces. They found that these data can't be processed in an existing centralized data processing manner. They also faced the efficiency of a time factor. So, with the use of parallel architecture, these large companies solve the problems of extracting relevant information from a huge data dump. They used one of the best open source tools, Hadoop to solve the data processing problems.

Garcia and Wang [8] worked on Querying Large Web Public RDF Datasets on Amazon Cloud Using Hadoop and Open Source Parsers. They wrote about the easiness of using Map-Reduce on large Dataset with Big Data Technology. Nowadays, Large public datasets are available and also can be found on the Amazon Web Service (AWS) Cloud. In particular, Web Data Commons has extracted and posted RDF Quads from the Common Crawl Corpus found on AWS which comprises over five billion web pages of the Internet. But they found infancy technology to process a large amount of data. As, within the last couple of years, AWS and EMR have facilitated processing of big files with parallelization along with a distributed file systems. There existed RDF technology and methodology and these were available commercially and open source. They proposed and presented advantages of Elastic Map Reduce (EMR) process. They also used and analyzed RDF tools against large dataset especially in a distributed environment which is relatively new.

Mukherjee et al. [9] compared a widely used clustered file system: VERITAS Cluster File System (SF-CFS) with Hadoop Distributed File System (HDFS) with the help of popular MapReduce benchmarks like Terasort, DFS-IO, and Gridmix on top of Apache Hadoop. In their experiments, VxCFS performance was same as Hadoop and sometimes it outperformed in many cases. This way, now enterprises can solve their problem of big data analytics need, with a traditional and existing shared storage without migrating them to different storage model in their data centers.

Time series data with Hadoop is not new. [10] uses Hadoop to implement time series data for stock analysis. Various works were done by ARIMA model. Lots of research conducted and still many researchers and business-analytics using ARIMA model to predict future sales. In business intelligence, it has great a impact as it is helps to take a critical decision, understanding market trend etc. chen [1], use ARIMA model to predict Chinese automobile demand. The authors tries to predict the demand of automobiles based on monthly data of automobile sales by Chinese Association of Automobile Manufacturers (CAAM) in January 2001 - June 2011 The author also evaluate the performance of various models.

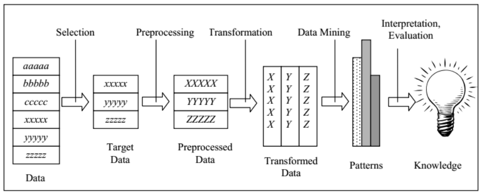

Fig. 3. The Data Mining Process

salehian [2], used ARIMA model for forecasting thermal rating of transmission lines. As they collect information from real-time tension monitoring systems, it helps to create more elaborate models. These models help to approximate dynamically, track thermal rating patterns and behavioral changes. These new models can be used to generate forecasts of future rating patterns. [12] Use ARIMA model for forecasting customer short-term load.They also compare their transfer function ARIMA model with the universe ARIMA model. [13] Here's they forecast dependent Internet traffic with Adjusted ARIMA Model for modeling. They also provide a comparison between their ARIMA Model with traditional ARIMA Model including performance improvement. Peiyuan et al. [3] This paper proposes a stochastic wind power model based on an Autoregressive Integrated Moving Average (ARIMA) process. The model takes into account the nonstationarity and also limitation of stochastic wind power generation. The model is constructed based on wind power measurement of one year from the Nysted offshore wind farm in Denmark. Banerjee [5], Here they predict the stock market using ARIMA model. They collected data on the monthly closing stock indices of Sensex for six years (2007-2012) of India and based on the data they developed an appropriate model which help them to forecast future unobserved values of Indian stock market indices.

III. METHODOLOGY & IMPLEMENTATION

Tools and techniques of data science are not limited and there is not a hard and fast rule to use some particular tools and techniques. Hadoop is one of the ecosystems that comes with lots of individual components and help to process large data sets. So, in our work, we followed the basic data mining working procedure to process data.

A. Working Procedure

Data Mining is a broad term. In this paper, we mainly refereed data mining steps to the algorithmic steps it recommends which also known as knowledge discovery in databases (KDD) process. This entire process, as originally envisioned by [15] is shown in Fig: 3

The first three steps in 3 involve preparing the data for mining. Data should be collected from a large and diverse set of sources, any necessary pre-processing must then be performed, and finally, the data should be transformed to apply for the data mining algorithms that are applied in the data mining step. In our work, we did not need the preprocessing step. We selected the data and transformed that using map-reduce. On the fourth step, we used ARIMA model to generate forecasting pattern like graph and equation on which we used interpretation and evaluation to generate our desired knowledge. Our workflow can be described as given in Figure 4

There are many attributes presented that we don't need in our data. So before working with them, we just need to remove unnecessary attributes. Feature selection is important as it may lead to unexpected results. Also, Feature selection helps to avoid over-fitting and also to improve model performance and to provide faster and more cost-effective models. We worked with two attributes here that are date and sell amount.

Transformation converts formats of data from source to destination data system. Our transformation can be divided as:

- Mapping-maps date to sales column of each row of the data-sets.
- Reduction-Reduce to single date from multiple date entry and aggregate sales on each date sales.

We fit our whole reduced data to the ARIMA model. Before that, we compared this model with exponential model i.e halt-winter model and found that ARIMA is better. So, we fit our data on ARIMA model in order to extract pattern. Using different test and parameters we found our final mathematical equation that can efficiently predict the future sales.

Evolution of pattern is basically a decision that gives us a common knowledge on our data. After every step processing our aim is to give a knowledge that can be helpful in our scenario on which data we are working on. In our work, from the ARIMA model what we got a mathematical equation that can be used to calculate future sales, from that we can say what the seller will know about the sales and could be prepared for that can definitely help him.

We divide our whole work into several steps. We work with one step at a time. After completing a step, we go to the next steps. We can describe our whole implementation procedure in figure 5

B. Hadoop Implementation

We install hadoop in pseudo-distribution mode. There are several modes of installtion of hadoop such as single-mode, pseudo-distribution mode and cluster mode.

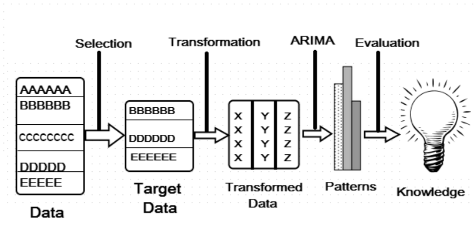

Fig. 4. Work flow of the whole process

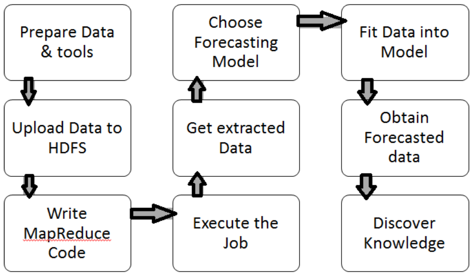

Fig. 5. Implementation procedure of our work

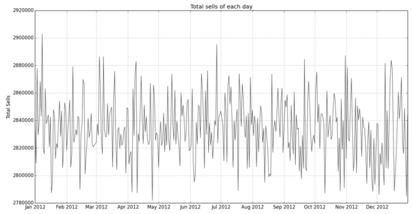

Fig. 6. Plot of Time vs Sales Data

1) Environment: We worked in Linux operating system as Hadoop is developed and tested in this environment.

Operating system: Ubuntu 14.10 32 bit

Hadoop: version 2.6.0

Programming Language: Python 2.7.3 (used for MapReduce), R(for ARIMA model)

2) Data sets: We choose time series data of a shop. We consider oneyear data. Datasets contain date, time, a name of the item, name of the store, the price of the item and payment system.

The meta data of our data sets is given below.

Date—Time—-Item Name—-Store Name——Price——Payment System

As the data is already structured we didn't need any preprocessing. After the initial step, we uploaded the data from our local machine to HDFS. When the data uploading is done, we focus to transform the data. We use MapReduce, Hadoop computing engine to process our data. Hadoop support many languages for writing MapReduce code. Mainly As Hadoop runs on JVM platform, MapReduce code is written in Java. There is Hadoop Streaming service that allows python as a MapReduce programming language. After installing required tools, uploading data to hdfs and writing MapReduce code, we start processing the data. We executed the job in Hadoop. Hadoop framework will distribute the data, apply computation on the data and will return reduced data as a format we provided.

Our Hadoop part is done. we will now fit the data into a forecasting model.

C. ARIMA Steps

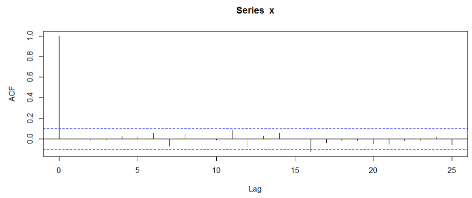

Fig. 7. ACF Curve of our Dataset

1) Stationarity check: From the data we found that the above data is strongly stationary. We also can run a test on the data. The test is called dicky- fuller test.

Below this line a boxed environment is used

```
Dickey-Fuller = -19.091, Lag order = 0, p-value = 0.01 alternative hypothesis: stationary Warning message: In adf.test(z, alternative = ”stationary”, k = 0) : p-value smaller than printed p-value
```

2) Identification step: The identification step is to estimate the value of p and q. The first step is acf curve.

From the curve 7 we saw that acf curve is dampening exponentially.

The pacf curve from our dataset given below:

So from above the curve we can estimate the value of p and q. p = 1, d = 0, q = 1

So the estimated equation of ARIMA (1,0,1) model is:

$$Y_t = a_1 Y_{t-1} + b_1 \epsilon_{t-1} \tag{6}$$

D. Estimation

Estimation is selection process model to better predict. One feature of R language is used. R has a function name auto.arima(). Its job is to predict ARIMA model from the raw data set. We use it and compare with our model. We also test some extra model to identify the appropriate model for our data.

So heres the estimated model. We choose model which have lower AIC and BIC.

AIC Akaike Information Criteria

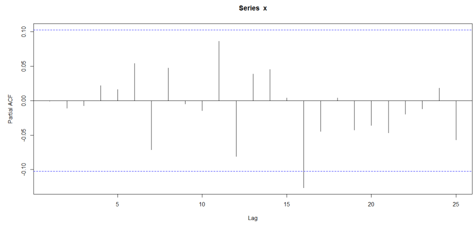

Fig. 8. PACF Curve of our Dataset

BIC - Bayesian Information Criteria.

These are used for selecting best models. The small the value, the better the model.

So the final equation of the model after estimation is given below

$$Y_t = -1.4065 * Y_{t-1} - 0.9712 * Y_{t-2} + 1.4089 * \epsilon_{t-1} + 0.9530 * \epsilon_{t-2} \tag{7}$$

Where Y_{t-1} is the previous data of t time and ϵ is the error term.

E. Goodness of FIT

For the final task, we tested goodness of fit of the models. The goodness of fit tells how much error occurs in forecasting. There are many systems for error estimation such as mean absolute error, mean absolute percent error, mean square error, root mean square error.

There are other techniques for error measurement. The below table list out the estimation models errors.

From the table III we see that ARIMA(2,0,2) has least error comparing with others.

IV. Experimental Result

We divide the experimental result in two section. First hadoop part and second ARIMA model part.

A. Hadoop Output

Hadoop has many packages including it. We use only Map Reduce and HDFS.

Input data size: 250MB

Output/Reduced Data Size: 8KB

Time to extract Data: 2.17sec

B. Forecasting Output

The outcome of the all above work is to predict future sales. So we get an equation to predict the future sales in business and now it will help to take a business decision. The equation for predicting the future value from the data is given in equation 7

The figure 9 shows the forecasting data where the x-axis is the day and y-axis is the sales of each day. Here the first 365 days is the original data. Then the forecasted data. The forecast is somewhat close to the original. The gray range is confidence.

If we plot the forecasted data along with raw data then the scenario will be like this:

TABLE II. Estimation of Various ARIMA Model

	A1	A2	B1	B2	AIC
Auto.arima()	-1.4065	-0.9712	1.4089	0.9530	8360.94
ARIMA(1,0,1)	-0.7641			0.7547	8360.64
ARIMA(0,0,2)			-0.0007	-0.0106	8360.68
ARIMA(2,0,1)	-0.5114	0.0019	0.5109		8362.72
ARIMA(1,0,2)	-0.7737		0.7741	0.013	8362.57

TABLE III. Goodness of Fit of different Models

	ME	RMSE	MAE	MPE	MAPE
Arma(2,0,2)	-15.65645	22416.8	17778.59	-0.006796	0.627134
Arima(1,0,1)	6.863218	22539.6	17837.49	-0.006076	0.6291727
Arima(1,0,2)	1.182253	22537.6	17836.35	0.0062751	0.6291311
Arima(2,0,1)	0.914088	22542.1	17844.02	0.0062870	0.6294027

C. Knowledge

The average of the first 10 days of New Year of forecasted value is 2834134.8
The average of the middle 10 days of forecasted sales is 2862568.6
The average of the last 10 days of forecasted sales is 2833646

From our result, we found that sales will be more in middle ten (10) days than first and last 10 days. If sales amount is high we can conclude that sales will increase on those days. So an important decision here is that the businessman should be prepared in the middle ten days to fulfill the heavy loads of sales.

V. Future Work

In this work, we only predict sales of each day based on previous year sales. To predict more efficiently, we need other parameters also like seasoning, occasional sales. In future, we will consider more attributes to forecast sales. We are planning to make a multi-node Hadoop cluster. Also, there are many processing engines available for processing big data. We will use Apache Spark instead of MapReduce in Hadoop ecosystem. We will also implement other forecasting models as well as Hadoop Machine Learning components named Mahout. We are planning to analyze and forecast stock markets, traffic load by day, user behavior in particular situation based on previous action etc.

VI. Conclusion

Through this studies, Our main objective was to familiar with data science and its tools and techniques. We used pseudo-distribution mode of Hadoop. We process and reduce data Using Hadoop. After reducing the data, we used ARIMA model to forecast future data. At first, we go in ARIMA process to estimate the model. After estimating the model we used R to make an ARIMA model from the data. Then we compare the both models. We accept the model which has lowest AIC, BIC. AIC determines the best fit of the data. Finally, we take the goodness of fit test. Calculate various error of the models and we find that our accepted model has least error comparing to others. Then from the model with least error, we forecast sales data.

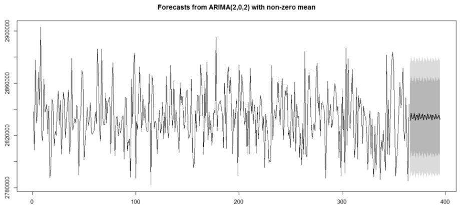

Fig. 9. Forecast from ARIMA(2,0,2) with non-zero mean

References

[1] Daoping Chen,*Chinese automobile demand prediction based on ARIMA model*,2011,volume 4,pages 2197-2201,Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on

[2] A. Salehian,*ARIMA time series modeling for forecasting thermal rating of transmission lines*,2003,volume 3,pages 875-879,Transmission and Distribution Conference and Exposition, 2003 IEEE PES.

[3] Peiyuan Chen and Pedersen, T. and Bak-Jensen, B. and Zhe Chen, *ARIMA-Based Time Series Model of Stochastic Wind Power Generation*,2010, volume 25 number 2,Power Systems, IEEE Transactions on

[4] Kamalpreet Singh and Ravinder Kaur, *Hadoop: Addressing Challenges of Big Data*, 2014,number 2 pages 686-689,Advance Computing Conference (IACC), 2014 IEEE International Transaction on.

[5] D.Banerjee. *Forecasting of Indian stock market using time-series ARIMA model*,pages 131-135,Business and Information Management (ICBIM), 2014 2nd International Conference on

[6] Nandimath et al. *Big data analysis using Apache Hadoop*,2013 pages 700-703,Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on

[7] Aditya B. et al *Addressing Big Data Problem Using Hadoop and Map Reduce*,year 2012,page 06-08, Addressing Big Data Problem Using Hadoop and Map Reduce NIRMA UNIVERSITY INTERNATIONAL CONFERENCE ON ENGINEERING

[8] Garcia et al. *Analysis of Big Data Technologies and Method - Query Large Web Public RDF Datasets on Amazon Cloud Using Hadoop and Open Source Parsers*, year 2013, pages 244-251, Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on

[9] Mukherjee et al. *Shared disk big data analytics with Apache Hadoop*, year 2012,pages 1-6, High Performance Computing (HiPC), 2012 19th International Conference on

[10] Y. Xie et al. *Implementation of time series data clustering based on SVD for stock data analysis on hadoop platform*, 2014,pages 2007-2010, 2014 9th IEEE Conference on Industrial Electronics and Applications

[11] S. Pandey and V. Tokekar, *Prominence of MapReduce in Big Data Processing*,2014,pages 555-560, Communication Systems and Network Technologies (CSNT), 2014 Fourth International Conference on

[12] M. Y. Cho, J. C. Hwang and C. S. Chen, *Customer short term load forecasting by using ARIMA transfer function model* Energy Management and Power Delivery, 1995. Proceedings of EMPD '95., 1995 International Conference on, 1995, pp. 317-322 vol.1.

[13] H. M. A. El Hag and S. M. Sharif, *An adjusted ARIMA model for internet traffic* AFRICON 2007, Windhoek, 2007, pp. 1-6.

[14] P. Zhang, X. Wu, X. Wang and S. Bi, *Short-term load forecasting based on big data technologies* in CSEE Journal of Power and Energy Systems, vol. 1, no. 3, pp. 59-67, Sept. 2015.

[15] Fayyad, Usama and Piatetsky-Shapiro, Gregory and Smyth, Padhraic, *Knowledge Discovery and Data Mining: Towards a Unifying Framework*, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, series-kDD'96, 1996, Portland, Oregon, numpages-7, pp. 82–88, AAAI Press

[16] *Auto Regressive Integrated Moving Average*, http://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average accessed: 2015-04-26

[17] *BOX-Jenkins Approach*, http://en.wikipedia.org/wiki/Box-Jenkins, accessed: 2015-04-26

[18] *Introduction to ARIMA Model*, http://people.duke.edu/~rnau/411arim.htm Accessed: 2015-04-26