

# Applying Hadoop Framework for Demand Forecasting Application

Navya V K<sup>1</sup>; Nirmala Hiremani<sup>2</sup>

<sup>1</sup>PG Scholar, Dept. Of Computer Science and Engineering, VIAT, Muddenahalli;

<sup>2</sup>Assistant Professor, Dept. Of Computer Science and Engineering, VIAT, Muddenahalli

navya789@gmail.com

## ABSTRACT

*The retail business in India is expanding in leaps and bounds. With increasing competition, each retailer needs to correctly cope up with the impending demand. In any production environment it is important to get adequate volumes of orders that lead to the manufacture of products. In order to assess the demand of a product we need to observe the orders placed based on the dataset available which can be present or historical. In this paper we propose a time series based scheme to show the demand forecast for the forth coming years using Hadoop framework and forecasting models like ARIMA, Naive, Regression and Polynomial Regression. The values obtained from the above models forms the forecasted demand and can be pictorially shown with the help of drawing a line chart using JFreeChart.*

**Keywords:** Bigdata, Hadoop, HBase, Time Series Analysis.

## 1. INTRODUCTION

Demand forecasting [2] is used to forecast customer demand and it is the area of predictive analytics to understand consumer demand for goods and services. The demand enables the supplier to keep the right amount of stock on hand. If the demand is underestimated or overestimated it can cause a financial drain. Understanding demand keeps a company competitive in the market place. Understanding demand and prediction helps manufactures, suppliers and retailers [1]. To meet customer needs, we choose the appropriate forecasting models. Forecasting product demand is crucial to any supplier or retailer [10]. A forecast of future demand determines the quantities to be

produced, purchased and shipped. Demand forecasting helps in the basic operation from supplier's raw materials to furnished goods in customer's hands. In the case of most firms they cannot wait for demand to react to the customer orders.

In order to provide fast order cycle times the firms offer rapid delivery to their customers based on the forecast of future demand. If the demand is accurate forecasts can lead to efficient operations and customer service can be to a high level and if forecasts are inaccurate it may lead to inefficient and customer service might be done to a poor level. The efficiency and effectiveness of the logistics process improve the quality of demand forecast. Some of the key benefits

1. It is short time to market
2. Reduction in cost
3. Profit Margin can be increased
4. The accuracy of forecast can be increased
5. Supply chain can be improved

Here we propose a time series based scheme for the demand prediction using Big data platform [7] i.e. Hadoop framework. Here we collect sales data of many areas of a retail store. The data collected can be of large size. Since we use big data platform we can handle the data size more efficiently. The historical data collected is divided into datasets according to the areas and we place the data set to HDFS. Here we use four forecasting models like ARIMA, Naive, Regression analysis and Polynomial multivariate regression analysis.

The model building is done with the help of Map Reduce programming and the results can be obtained in HDFS. We also provide the provision of storing the results to HBase. The results can be shown pictorially with the help of line chart and the

graphical representation of the time series models helps us in the comparative study of the models.

## 2. TECHNIQUES AND TOOLS

### 2.1 TIME SERIES ANALYSIS

Time series [3] is a set of observations of well defined data items obtained through regular measurements over time. For example, measuring

the retail sales value of each month for the year comprises a time series. For our research we used the sales data collected time to time. Here we discuss mechanisms like Naive, ARIMA [13], Regression Analysis and Polynomial multivariate Regression analysis [13].

#### 2.1.1 Naive Forecasting Model

It is an estimating technique in which the last periods value are used as the periods forecast and no further adjustment is done. Usually naive forecast is used only to compare the forecasts generated by other techniques.

#### 2.1.2 Moving Average Model

It works on time series in which the value for a time period is the mean value of preceding and succeeding time periods. If given a set of observation it smoothes the peaks and troughs can be considered as an advantage.

#### 2.1.3 Regression Analysis

It deals with a single variable and data points are plotted to put on a straight line. The line is defined by its gradient or slope and point that intercept the x-axis. Assume x as independent variable and y as dependent variable and line can be represented as  $y = \text{intercept} + \text{slope} * x$

#### 2.1.4 Polynomial Multivariate Regression Analysis

It deals with a single variable and data points are plotted to put on a polynomial line. Assume x as independent variable and y as dependent variable then line can be represented as:

$$y = a_0 + a_1 * x + a_2 * x^2 + a_3 * x^3 + \dots + a_m * x^m$$

## 2.2 HADOOP

Hadoop [12] is a software framework which is open source and it is written in Java. It is used for distributed storage and can process even large data sets. The core parts of Hadoop are,

1. Hadoop Distributed File System (HDFS) as storage part and
2. MapReduce as the processing part.

Here, we use both HDFS and MapReduce for storage and building the forecast models.

#### 2.2.1 HDFS (Hadoop Distributed File System)

Hadoop has a storage part which is fault tolerant. It has the capability of storing large amount of information and without losing data it can survive the failure of important parts of the storage infrastructure. File's contents are stored inside data node and it is having large sizes of blocks. (e.g.:- 64MB) and name node stores the Meta information and information of the blocks.

#### 2.2.2 MapReduce

It is a programming model and is basically used to process and generate datasets of large size. It uses a parallel distributed algorithm on cluster. When a job is given to MapReduce framework it divides it into Map tasks and it is given to different nodes for running. Only a part of input is dealt by the Map task and after processing those intermediate key-value pairs are given to the Reduce function. Based on the specific key the Reduce function will merge the process and output is generated as per the client requirements. A MapReduce program has a Map () procedure for filtering and sorting and Reduce () procedure for summary operation.

#### 2.2.3 HBase

It is written in Java which is an open source distributed database. HBase is non-relational. It runs on top of HDFS and is a part of Hadoop. HBase Schema can have several tables and each table is a set of column families. HTable is a separate file used to store the column family. Row keywords, column keywords and timestamp form the index of the table. A series of data lines can be stored in HBase table. Each line has line keywords, optional timestamp and some columns that possibly contain data.

## 3. METHODOLOGY

We are implementing demand forecasting application using big data platform Hadoop. We utilize HDFS, MapReduce and HBase. We are using Hadoop framework because it is open source and we can handle large data sets and processing of the data is done fast using MapReduce. Forecasting approaches are of 3 types like Qualitative, Quantitative and Time Series approach [5]. In this

work we deal with time series approach where the future values are related to its past values. The forecasting models used in this work are ARIMA, Naive forecasting model, Regression model and

Polynomial multivariate Regression model. Our work proposes a design that can be used for demand forecasting application, as shown in Fig. 1.

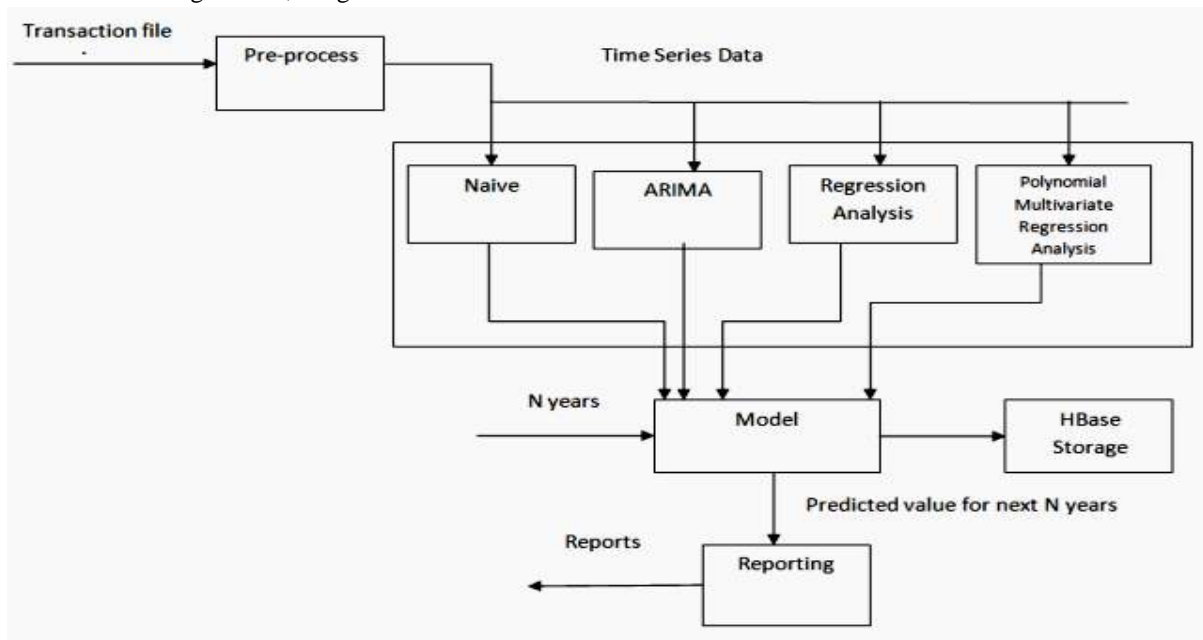


Fig1: Demand forecasting application

### 3.1 BASIC OPERATIONS

#### 3.1.1 Pre-process the transaction file

In this module the historical data is collected from the various stores of an enterprise and it is put in the HDFS so that it can be used in further stages. The transaction file that we are dealing with can be of huge size and we utilize the big data platform. As the Hadoop framework provides us with HDFS through which we are able to store a large sized file which can be of size in PB. We are forecasting the future demand based on this historical data that we have collected. Once it is stored at the HDFS we have to load this observation and give as input to our model. The data that we have collected should be free from errors and in correct format. So the initial preprocessing is being done. The data that we have collected is a time series data of N years and it is given as input to the model.

#### 3.1.2 Building the Model

Once we have the observations with us the next step is to build the model. Here we build the models like ARIMA, Naive, Regression Analysis and Polynomial Multivariate Regression Analysis.

#### 3.1.3 Prediction made using the Model

In this module we will see how we use the models to predict the future demand. The observation loaded in the

HDFS is given to the models as input. We write the code for Map-Reduce and within the map function we implement the models by providing the necessary datasets. After the map function it is the responsibility of the reduce function to collect the output and place in a secured HBase. Storing of the predicted demand helps in the further stages. We take the four models so that it helps us in comparing of the future demand.

#### 3.1.4 Secured HBase Storage

After the prediction is done the predicted demand is stored in the HBase. Now the HBase offers security. The Reduce function of the Map-Reduce is responsible for collecting the predicted values and those values can be sensitive. So such values are placed in a secured HBase Storage.

#### 3.1.5 Forecasted demand and Reporting

Once the forecasted demand is available in the HBase it can be retrieved and it is shown to the respective authorized people so that it becomes helpful in their supply chain. The prediction of N years is obtained here. The values of the predicted demand can be seen using the log file and this module also provides the necessary reports. The values are loaded and the report for the models based on the predicted values is drawn so that user gets an idea of the predicted demand clearly. We show the comparison of four models by drawing the chart which

helps in clearly identifying the forecasted demand and becomes helpful in the supply chain of products.

#### 4. EXPERIMENT RESULTS

For our experiments we have utilized the sales record of a retail store having 20 branches in different areas. Here, we utilize the data collected from 2005-2014 of 20 different areas and we forecast the demand for the next 3 years 2015-2017. The data was collected and initially it was divided into dataset of different areas using Dataset Creator. The code was written in Java and executed in Net Beans platform. Since we are using the big data platform we use HDFS, MapReduce and HBase. Then based on the area in which we want to forecast we have loaded that area sales dataset to HDFS for storage. Here we can an option of forecasting the demand of 4-5 or more than that areas at the same time. Load the desired areas to the HDFS and the model building, processing is done together for all the areas and results can be obtained.

Here we try to build 4 models which help in the demand forecast. The 4 models used here are ARIMA, Naive forecasting, Regression and Polynomial Regression Model. We utilize the Open forecast package for model building. Open Forecast is a general purpose package where forecasting models are written in Java so that we can apply it to any data series. We implement ARIMA using MovingAverageModel, Naive using NaiveForecastingModel, Regression using Regression Model and Polynomial multivariate Regression using Polynomial Regression Model. The map reduce code run as the background process which reads the data from the HDFS and the Map() does the model building and processing and the forecast values are collected at the Reduce(). The results are also loaded to HDFS which can be read from HDFS and displayed through a form in Net Beans. The results of all the 4 models can be displayed.

We can store the results in HBase table so that we can utilize the values for future use. It is implemented as results is read from the HDFS and loaded to HBase table. HBase provides a secured way of storage which is given by the Hadoop Framework.

The forecasted demand values given by the 4 models can be shown pictorially with the help of Line chart. We draw the Line chart with the help of JFreeChart. JFreeChart is an open source library available for Java that allows users to easily generate graphs and charts. It is particularly effective for when a user needs to regenerate graphs that change on a frequent basis. The Line chart of 4 models clearly gives us an idea about the forecast and it also helps us in the comparative study. In chart we plot the year 2015-2017 (years to be forecasted) in X-axis and demand forecast in Y-axis. We create the chart by calling the function createchart() which in turn call the function

createchartdataset(). We read the results from HDFS and call the function createchartdataset() and load the results from ARIMA, Naive, Regression and Polynomial model to it. By drawing the line chart we can pictorially get an idea of the forecast made by the models and we can do a comparative study also.

Fig.2 shows the line chart of demand forecast using 4 forecasting models.

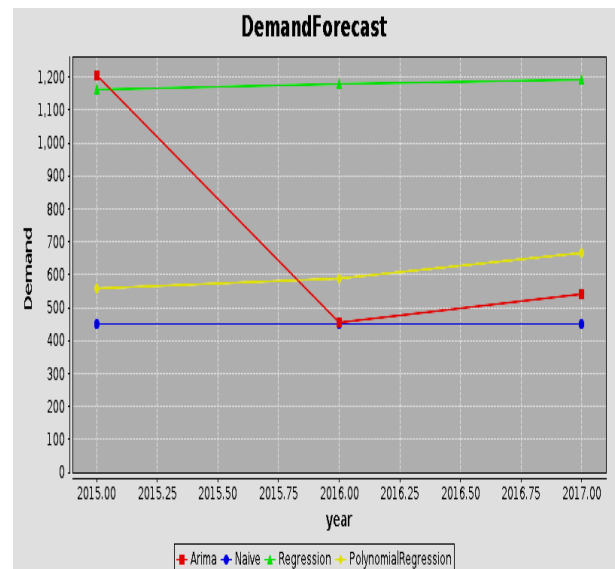


Fig.2: Line chart for demand forecast

#### 5. CONCLUSION

Our work is intended provide an effective forecasting solution to a retailer who is having stores in 20 areas. The solution provides enables the retailer to foresee the demand for product in next 3 years based on the previous sales. This helps retailer to make business decisions such as product price, marketing. It may also help in promotions to develop and plan. If forecasts are made accurate it may help retail business owners to extend profits. The open source technology like Hadoop is used in implementation that reduces the cost of implementation and can deal with large data sets. In addition to that the various time series based forecasting models helps in achieving more accuracy and efficiency.

#### 6. REFERENCES

- [1]. Vijay Gabale, Ashutosh Dhekne, "Demand Forecasting in the Indian Retail Industry Applied Economics".
- [2]. <http://www.statsoft.com/Textbook/Demand-Forecasting>
- [3]. Kumara M.P.T.R, Fernando W.M.S, Perera J.M.C.U, Philips C.H.C, Department of Computer Science and Engineering, University of Moratuwa, "Time Series Prediction Algorithms: Literature Review".

- [4]. Vaida Pilinkien, "Market Demand Forecasting Models and their Elements in the Context of Competitive Market".
- [5]. <http://www.ipredict.it/ForecastingMethods.aspx>
- [6]. Prashant Shrivastava, S.Pandiraj and Dr. J.Jagadeesan, "Big Data analytics in forecasting lakes levels", IJAIEM, vol.3, March 2014.
- [7]. <https://Infosys.uni-saarland.de/publications/BigDataTutorial.pdf>
- [8]. O'ReillyMedia (2013) "Disruptive Possibilities: How Big Data changes Everything".
- [9]. <http://smallbusiness.chron.com/explain-forecasting-retail-37966.html>
- [10]. <http://www.oracle.com/us/products/applications/retail/supply-chain-planning/demandforecasting/overview/retail-demand-forecasting-overview-1561396.html>
- [11]. M.Jayashree, "Data Mining: Exploring Bigdata using Hadoop and MapReduce", IJESR, vol.4, January 2013.
- [12]. Tom White, Hadoop The Definitive Guide, O'Reilly Publication.
- [13]. Priyanka Sinha, "Multivariate Polynomial Regression in Data Mining: Methodology, Problems and Solutions", IJSER, vol. 4, December-2013
- [14]. International Journal from Science & Engineering Research Support Society (SERSC): "Hadoop-based ARIMA Algorithm and its Application in Weather Forecast", vol.6, November 2011.
- [15]. Hastie, T., Tibshirani, R., & Friedman, J. H., "The elements of statistical learning: Data mining, inference, and prediction", New York: Springer, 2001.
- [16]. Weiss and Indurkha, "Predictive data mining: A practical guide", New York: Morgan-Kaufman, 1997.
- [17] D. Jiang et al., "The Performance of MapReduce: An In-depth Study". PVLDB, pp.472-483, 2010.
- [18] J. Lin et al., Full-Text Indexing for Optimizing Selection Operations in Large-Scale Data Analytics MapReduce Workshop, 2011.
- [19]. Vignesh Prajapati, "Big Data Analytics with R and Hadoop", PACKT Publishing, 2013.
- [20]. Kay-Yut Chen, Leslie R. Fine and Bernardo A. Huberman, "Predicting the Future", Journal on Information Systems Frontiers, vol. 5, pp. 47-61, 2003.