Khulna University Studies Special Issue ICSTEM4IR: 474-483, 2022

ORIGINAL ARTICLE

SCIENCE & ENGINEERING

LOCAL INVENTORY DEMAND FORECASTING OF E-COMMERCE WITH MAPREDUCE FRAMEWORK

Kazi Fardin Islam*, Mithun Rahman, and SK Alamgir Hossain

Computer Science and Engineering Discipline, Khulna University, Khulna 9208, Bangladesh

KUS: ICSTEM4IR-22/0082

Manuscript submitted: June 20, 2022 Accepted: September 27, 2022

Abstract

With the rapid expansion of e-commerce businesses shortest delivery time is a challenging task. As well as price and quality, delivery time is becoming an important key-factor in the race of business growth. To deliver the product to the buyer in the shortest time, the product should be at the local inventory at the time of purchasing. Which is a difficult task for e-commerce companies to store the right products at right time in local inventories. Extracting useful patterns and forecasting demand from purchase history is a tough job because of the huge volume and high variety nature of e-commerce data. The MapReduce algorithm of Big Data ecosystem can be used to pre-process the enormous amount of data and to summarize it into a suitable format which later can easily be used to build any forecasting model. A comparison of different classical time series forecasting methods with the pre-processing of MapReduce algorithm is focused on this work.

Keywords: Demand Forecasting, Inventory, Big Data, MapReduce, ARIMA

*Corresponding author: <abir1713@cseku.ac.bd>
DOI: https://doi.org/10.53808/KUS.2022.ICSTEM4IR.0082-se

Introduction

The area of e-commerce is increasing day by day, the opportunity for business growth is spectacular, nowadays consumers are very keen on online purchasing. Due to its delivery facilities and product varieties. For the increasing demand in competitive efficiency, inventory management prioritizes spotlight on ensuring accurate inventory quantities and automation of the decision-formulation action of a business. These flexibilities are considered to be a revolutionary development for accomplishing pace and exactness in e-commerce. An automated inventory control system offers accurate and precise business intuition that guide to make data-driven decisions to accomplish the particular profit goal (**Thiebaut**; **2019**). Demand forecasting of the products is one of those money-making insights. A predictive inventory management system processes a huge volume of past sales data and anticipates future demand for inventory which results in an improvement in lead times and cost overheads.

The fundamental concern is that customers need to have the option to buy the items they need, when they need them, on the channels they like. Neglecting to have the stock to satisfy this demand puts an e-commerce at the risk of missed sales opportunities (**Ajibola**; **2014**). For an e-commerce higher missed sales opportunity is a matter of great concern, as it pushes down the company's reputation among hundreds of competitors in the market. An efficient predictive system can suggest the right decisions about which item to stock, when to stock, and where to stock, which reduces the chance of missed sales opportunities. An e-commerce produces tons of data every day, which becomes tough work for conventional data processing systems to handle this huge-sized data, here comes the need for big data processing and analytics. Big data analytics permits businesses to gain different customer-related patterns and trends (**Anshari**; **2019; Aimee**; **2019**). Understanding these customer insights allows an e-commerce business to deliver the customers' wants at right time and at the right extent (**Randy**; **2021**). Big Data manages those datasets which cannot be perceived, stored, and processed by traditional approaches and technologies within a tolerable satisfactory time (**Li**; **2018**). In a survey from business executives of large companies conducted by NewVantage Partners in late 2020, 91.9% said they were accelerating their investments in big data and data analytics related AI initiatives, while 96% of companies reported successful outcomes in similar projects (**Randy**; **2007**). Inventory demand forecasting is one of those data-driven decision-making applications of big data analytics in the e-commerce business territory. We have implemented and tested different time series model upon the processed and summarized data to forecast inventory demand.

Data is arising as the world's most important resource for governments, organizations, and businesses to gain a competitive advantage. More than 300 million photos are uploaded per day on the world's largest social media Facebook, 510,000 comments are posted and 293,000 statuses updated in every minute (**Bernar**; **2018; Alicia**; **2017**). It is estimated that everyday world is generating 2.5 quintillion bytes of data from a variety of sources (**John**; **2018**). It was estimated that in

Khulna University Studies Special Issue (ICSTEM4IR): 474-483:2022
DOI: https://doi.org/10.53808/KUS.2022.ICSTEM4IR.0082-se

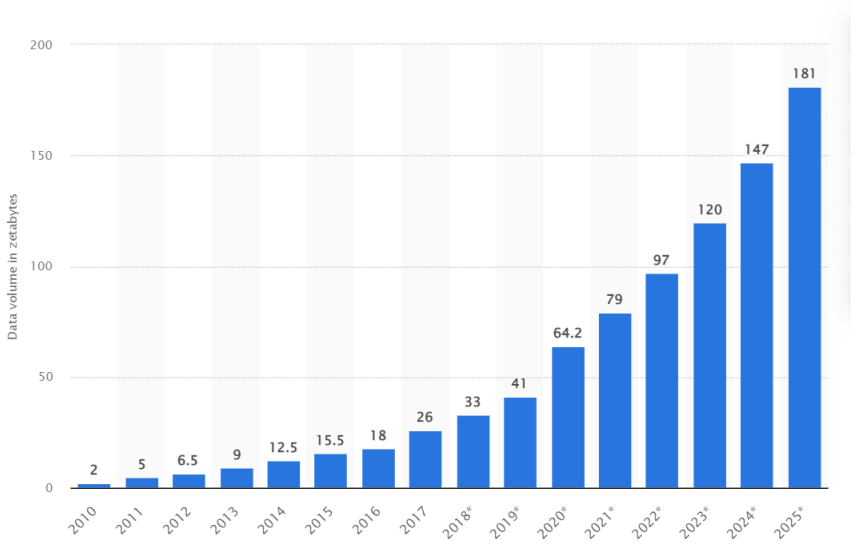

Figure 1: Amount of data consumed by 2025 in zettabytes **b5**

2020, the data size of the world will come to 40 Zeta bytes, but in 2020 the estimation got broken amazingly due to the COVID-19 pandemic making the new record of 64.2 Zeta bytes (**Arne**; **2021**). This should be affected on government, business and also individual scopes.

Related Works

Maobin Li et al. (**Li**; **2018**) studied Chinese e-commerce sales forecasting and experimented with individual and combined prediction model. They choose time series of various e-commerce products of Jingdong Company, China for predicting future sales. They analyzed the linearity and nonlinearity of ECS data and found their model adaptive to linear patterns of e-commerce retails and low performing to nonlinear patterns. They deployed ECS-ARIMA model on weekly sales data and then ECS-NARNN model on the same dataset. Found the ECS-NARNN model mapping better on the nonlinear patterns and ECS-ARIMA better in predicting linear elements of the time series.

The exploration job of Anupriya Jain et al. (**Jain**; **2020**) mainly focused on comparing two well-known forecasting algorithms namely Seasonal Auto-Regressive Integrated Moving Average(SARIMA) and Long Short-Term Memory Network(LSTM) for predicting how much a particular product is being sold. SARIMA is an expansion of the ARIMA model for fitting the regression model with seasonal time series. The authors considered intervals of twelve months as the seasonality value of SARIMA model. Although both algorithms performed incredibly well, the dataset size was only ten thousand. SARIMA model was slightly ahead of the LSTM model in the

context of long-term predictions.

Neha Verma et al.(**Verma**; **2017**) worked with an online shopping system with the assistance of Map-Reduce Framework. Hadoop Distributed File System (HDFS) was used here for storing data. Apriori technique was used for the finding pattern of sold products.

The Holt-Winters algorithm is used in the prediction of Walmart sales prediction in the works of Anita et al.(**Harsoor**; **2015**). They analyzed the sales forecasting using some Big Data applications. Because of being dealing with a huge amount of data they used HDFS and HIVE. For the sake of prediction, the Holt-Winters model with MapReduce showed its performance staggeringly. The main goal of their work was to find accurate information about investment and resource management. As the Holt-Winters model can't perform well with the random dataset, their model isn't useful to all types of datasets.

Xiaoqian Wang et al. (**Wang**; **2020**) proposed a framework for processing ultra-long time series. The distributed system calculates a weighted average of the local estimators which are delivered from the workers of the cluster, to decrease global loss function. They split the long series into several subseries to distribute the training task among slaves. Afterward, they compared the performance of distributed ARIMA and sole ARIMA model, they found that distributed ARIMA outperforms sole ARIMA in forecasting.

Leixiao Li at al. (**Li**; **2013**) proposed a Hadoop-based ARIMA model for weather forecasting. They concentrated on the issue of weather data mining and proposed a parallelized ARIMA model on Hadoop environment, that is scalable and easy to maintain. In that experiment, the historical weather data of past 10 years was used for building ARIMA model, and daily mean vapor pressure and daily mean humidity were predicted through different parameters. The system they offered is highly adaptive to handling huge meteorological data.

Mininath Bendre and Ramchandra Manthalkar (**Bendre**; **2019**) worked on investigating weather data of 100 years collected from 39 weather stations of Maharashtra, India. To analyze that ultra-long time series, they proposed a hybrid distributed forecasting model like the work of Xiaoqian Wang et al. (**Wang**; **2020**). They proposed predictive analytics based on the decomposition and classification on the modules. They visualized the decomposed series to identify trend, seasonality and other sophisticated components. The M-HM (MapReduce Hybrid Model) and neural network were also experimented in their extensive work (**Zaharia**; **2012; Parker**; **2012**). The M-HM model performed efficiently than M-ARIMA and M-KNN model.

Methodology

We have studied the weekly total quantity of a product in a certain region. We used MapReduce technique for processing the large dataset and applied different forecasting models on the summarized data. In the decomposition phase, the dominance of seasonality was noted. Hence, we tested the Seasonal ARIMA (SARIMA) along with the AR and ARIMA models.

Data Processing

The massive amount of raw data is collected from e-commerce's sales activity. In this study, we have analyzed the order history of a UK-based retailer. The dataset contains millions of entries of more than ten thousands unique products sold over three years time period. The features of datasets are Invoice No, Product Code, Product Description, Quantity, Purchase date, Unit price, Customer ID, and Location. As we are looking for the predictions of certain products during a certain time in a region, the Description and Unit price are not related that much.

Two features Invoice Date and Quantity of the specific Stock Code and Region are extracted through MapReduce algorithm. To protect the model from overfitting and for a balanced prediction model, the time period is resampled into weeks. This quantization along with the specific product query remarkably reduces the size of the dataset. The business queries are mainly deployed in the Mapper class because of its nature of mapping and transforming data. In the mapper class the instructions of preprocessing, noise and outlier handling are also stated.

The mapper class takes each entry as input, splits and converts it into key-value pairs, this pair type matches with the input type of the reducer class. In the MapReduce phase, the mapper function processes the selected key-value pairs. After that, it emits the processed selected key-value pairs for reducer. The Reducer function processes the certain values grouped by the same key, the Invoice Date and Quantity in our case. The preprocessing job follows the execution flow as described in Algorithm 1. As discussed earlier the Map output format should match the input format of the Reduce as follows:

Forecasting

Once the data is shaped into a suitable format, the forecasting model can be developed on that. In our work, we have experimented with different classical time series forecasting models. We primarily focused on developing an ARIMA (Autoregressive Integrated Moving Average) model analyzing the properties of the historical data of the selected product. Which is a statistical model that uses time-series data to predict future trends. ARIMA is an autoregressive model because it predicts future values based on past values. ARIMA gives forecasts based upon prior values in the series (Auto Regressive terms) and the errors made by previous predictions (Moving Average terms). This gives the model ability to immediately adjust for sudden changes in trends and helps in resulting in more accurate forecasts.

ARIMA is the combined Auto Regressive Moving Average model with an additional notion of integration. We applied different components of it in our data. It has three parameters. For the autoregressive part AR, p is the autoregressive term, for moving average part MA, q is the moving average part and d is the number of differences needed to make the time series stationary.

Another forecasting model we experimented with is SARIMA, which is Seasonal Autoregressive Integrated Moving Average. SARIMA is an extension of ARIMA that is used for analyzing

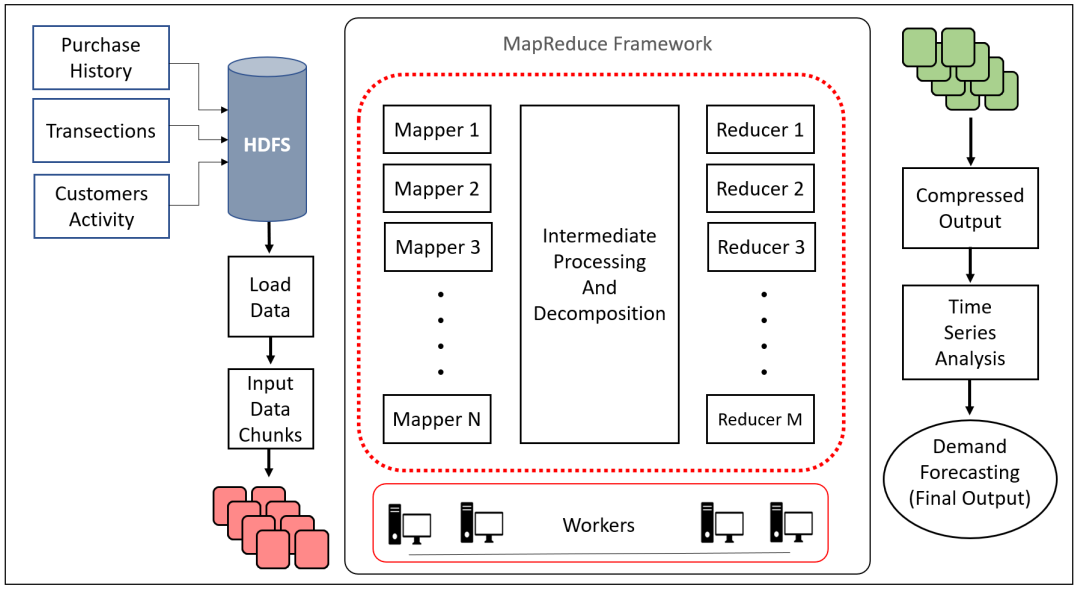

Figure 2: Proposed Method.

Algorithm 1 Data Preprocessing Algoritm

1: Map Class
2: Input: ($key \leftarrow name\ of\ the\ input, value \leftarrow value\ of\ the\ input$)
3: Output: (key, value)
4: Mapper(key, value)
5: **if** (($P : P \in Product\ Attribute\ \&\ P = Desired\_Product$) and ($A : A \in Area\ Attribute\ \&\ A = Desired\_Area$)) **then**
6: $key \leftarrow date$
7: **end if**
8: Intermediate: (key, value)
9: Reduce Class
10: Input: ($key \leftarrow name\ of\ the\ mapped\ data, values \leftarrow list\ of\ map\ data$)
11: Output: (key, value)
12: Reducer(key, value)
13: **while** values.hasNext() **do**
14: $sum \leftarrow sum + getnext(values)$
15: **end while**
16: Output: (key, value) # return the final output

the time series that have both trend and seasonality. This model is expressed as SARIMA(p, d, q)(P, D, Q)[s]. Where P is the seasonal autoregressive order, Q is the moving average order and s is the seasonal cycle. As we resampled the e-commerce data in weeks, so the value of s will be 52 in our case.
The first task of the ARIMA forecasting method is transforming the time series into a stationary one, the value of d is used for differencing the series d times to make it stationary. To identify the order of autoregressive term p, the assist of PACF (Partial Autocorrelation Factor) plot is needed. PACF can be defined as the correlation between the time series and its lag, by eliminating the contributions from the intermediate lags. This factor is the direct correlation between a lag and the series. To determine the order of AR, the number of lags that crosses the significance boundary in the PACF plot is considered as the initial value of p. The order of MA is also determined using ACF plot in the same way of determination of q. After figuring out the appropriate values of p,d,q the ARIMA model is framed in that configuration and best model is evaluated by residual analysis.

Results and Discussions

It is a laborious task to deal with a huge amount of data. First of all, we reduced the huge dataset to a small size of the dataset to make it suitable for analysis. For the sake of reducing the dataset, we took advantage of Hadoop-based Map Reduced framework. Time series problems differ from other types of machine learning problems. We modeled the time series problem as supervised learning. We conducted three types of time series models namely Auto Regression(AR), Auto-Regressive Integrated Moving Average(ARIMA), and Seasonal Auto-Regressive Integrated Moving Average(SARIMA).

Descriptive Analysis

Because of being supervised learning, we smoothed the data followed by decomposing the data for checking trends, seasonality, and stationarity. We used the MapReduce model for preprocessing our data and transforming it into intermediate input. From our decompostion a clear upward trend was noted which suggests that the quantity of sold products is rising day by day. Also, a clear seasonality was noted.

Forecasting Models

The main purpose of the three models was to forecast the number of products sold per week. From the Partial Auto Correlation Function(PACF) , we got p = 1. With help of the Ade fuller test, we came to know that we no need to differentiate data one time so we got d = 1. From the Auto Correlation Function(ACF), we got q = 4.So, we conducted AR(1), and ARIMA(1,1,4).Also We conducted SARIMA(2,1,0)(1,1,0)[52].

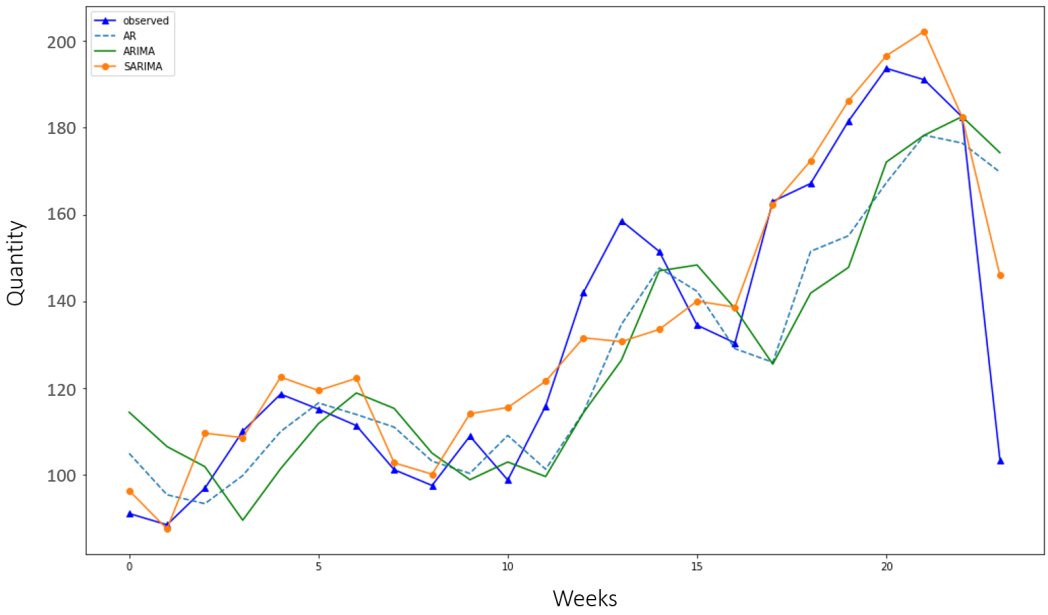

Figure 3: Prediction comparison of different models.

Table 1: Error Rate Table

Model	MAE	RMSE	MAPE
AR	18.13	23.51	15.23
ARIMA	14.63	20.42	14.47
SARIMA	13.91	16.34	10.09

SARIMA showed comparatively better performance. AR model had 18.13, 23.51 and 15.23 respectively Mean Absolue Error(MAE), Root Mean Squared Error(RMSE) and Mean Absolute Percentage Error(MAPE) while ARIMA model had 14.63, 20.42 and 14.47. Surprisingly, SARIMA had 13.91, 16.34 and 10.09 respectively Mean Absolue Error(MAE), Root Mean Squared Error(RMSE) and Mean Absolute Percentage Error(MAPE).

We have shown AR, ARIMA and SARIMA model line graph in one single line graph in Figure 3. Those graph lines are also suggest that SARIMA model is better than any other models.

Conclusions

Demand forecasting is the artery of businesses nowadays especially the E-Commerce business. Since it can limit deals drop. We understood from our examination SARIMA model performs

better for seasonal data. As we showed that SARIMA and ARIMA models perform really in the case of time series forecasting. Our plan is work with Holt-Winter Method forecasting. Also, Our tentative arrangement is to work with Long Short-Term Memory(LSTM) for forecast purposes and assess the exhibitions of LSTM and time series models. An efficient demand prediction model will help an ecommerce to take appropriate decisions in product storing and to deliver the products in the shortest time.

References

Thiebaut, R. (2019). AI Revolution: How data can identify and shape consumer behavior in ecommerce. In Entrepreneurship and Development in the 21st Century. Emerald Publishing Limited.

Ajibola, A. S., & Goosen, L. (2014). Missed opportunities in mobile E-commerce usability. Int. J. Sci. Eng. Res.(IJSER), 5(6), 954-956.

Harsoor, A. S., & Patil, A. (2015). Forecast of sales of Walmart store using big data applications. International Journal of Research in Engineering and Technology, 4(6), 51-59

Anshari, M., Almunawar, M. N., Lim, S. A., & Al-Mudimigh, A. (2019). Customer relationship management and big data enabled: Personalization & customization of services. Applied Computing and Informatics, 15(2), 94-101.

Arne Holst. (2021). Amount of data created, consumed, and stored 2010-2025. https://www.statista.com/statistics/871513/worldwide-data-created/

Bernard Marr.(2018). How Much Data Do We Create Every Day, The Mind-Blowing Stats Everyone Should Read. https://www.forbes.com /sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=2c464e6160ba

RandyBean.(2021). NewVantage Partners Releases 2021 Big Data and AI Executive Survey. https://www.businesswire.com/news/home/20210104005022/en/NewVantage-Partners-Releases-2021-Big-Data-and-AI-Executive-Survey

Alicia Tan. (2017). How Coca-Cola uses data to supercharge its superbrand status. https://www.adma.com.au/resources/how-coca-cola-uses-data-to-supercharge-its-superbrand-status

John Kopanakis. (2018). 5 Real-World Examples of How Brands are Using Big Data Analytics. https://www.mentionlytics.com/blog/5-real-world-examples-of-how-brands-are-using-big-data-analytics/

Dr. Andreas Stephan Huber et al. (2014). Big data: Potentials from a risk management perspective.

Aimee Manning. (2019). How Big Data is Improving Inventory Management. https://www.bigcommerce.com/blog/data-inventory-management/

Li, M., Ji, S., & Liu, G. (2018). Forecasting of Chinese E-commerce sales: an empirical comparison of ARIMA, nonlinear autoregressive neural network, and a combined ARIMA-NARNN model. Mathematical Problems in Engineering, 2018.

Verma, N., & Singh, J. (2017). An intelligent approach to big data analytics for sustainable retail environment using Apriori-MapReduce framework. Industrial Management & Data Systems.

Wang, X., Kang, Y., Hyndman, R. J., & Li, F. (2020). Distributed ARIMA models for ultra-long time series. arXiv preprint arXiv:2007.09577.

Jain, A., Karthikeyan, V., Sahana, B., Shambhavi, B. R., Sindhu, K., & Balaji, S. (2020, November). Demand Forecasting for E-Commerce Platforms. In 2020 IEEE International Conference for Innovation in Technology (INOCON) (pp. 1-4). IEEE.

Li, L., Ma, Z., Liu, L., & Fan, Y. (2013). Hadoop-based ARIMA algorithm and its application in weather forecast. International Journal of Database Theory and Application, 6(5), 119-132.

Bendre, M., & Manthalkar, R. (2019). Time series decomposition and predictive analytics using MapReduce framework. Expert Systems with Applications, 116, 108-120.

M. Zaharia et al., ''Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing,'' in Proc. 9th USENIX Conf. Netw.Syst. Design Implement. (NSDI), 2012, p. 2

C. Parker, ''Unexpected challenges in large scale machine learning,'' in Proc. 1st Int. Workshop Big Data, Streams Heterogeneous Source Mining Algorithms, Syst., Programm. Models Appl. (BigMine), 2012, pp. 1–6.