Contemplation of Cash Inflow in ATM's using Big Data based on Hadoop Architecture

Mr. Saurabh Singh
Department of Computer Engineering
Army Institute of Technology, Savitribai Phule Pune University, (BE), India

Abstract— In today's world, one of the most challenging task for a bank is to maintain cash in their ATMs (Automated Teller Machines) so that they can easily serve their customers. Hence, we are attempting to build a model, which can predict the amount of cash that needs to be put into the ATM's. Cash Management and Optimization are crucial activity for any bank. Therefore, Bank needs to predict the cash inflow into the ATM's very accurately. For instance, cash demand is subject to follow the period of time, position of ATM's. In addition, the quantity of cash drawn may subject to factor such as Holiday pattern may follow weekly, monthly and annual cycles, Festivals that may cause dynamism in the overall requirement to avoid cash inflow cycle outburst using Open Source Technology.

Keywords: Regression, Time Series, Linear Regression, ARIMA

I. INTRODUCTION

An autoregressive integrated moving average (ARIMA) model is a generalization of an autoregressive moving average (ARMA) model. These models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting). AR part of a time series Y_t is that the observed value y(t) depends on linear combination of previous observed values up to a defined lag (denoted p), plus a random error term €(t) because partial dependence The AR(p) model can be expressed algebraically as:

$$yt = \varphi 1yt-1 + \varphi 2yt-2 + \dots\dots + \varphi pyt-p + \varepsilon t \quad (1.1)$$

MA part of a time series Y_t is that the observed value y(t) is a random error term plus linear combination of previous random error terms up to a defined lag (denoted q). The MA(q) model can be expressed algebraically as:

$$yt = \varepsilon t + \theta 1\varepsilon t-1 + \theta 2\varepsilon t-2 + \dots\dots + \theta q\varepsilon t-q \quad (1.2)$$

If we combine differencing with auto regression and a moving average model, we obtain a non-seasonal ARIMA model. Then equation can be written as:

$$\varphi 1B - \varphi 2B\,2 - \cdots \dots - \varphi pB\,p\,)\,(1-B)d\,yt = c + (1 - \theta 1B - \theta 1B\,2 - \cdots \dots - \theta 1B\,p\,) \quad (1.3)$$

€(t) AR(p) differences MA(q) Where B: Backward shift operator (shifting the data back one period)

Exponential smoothing is a technique that can be applied to time series data, either to produce smoothed data for presentation, or to make forecasts. Exponential smoothing assigns exponentially decreasing Exponential smoothing is commonly applied to financial market and economic data, but it can be used with any discrete set of repeated measurements. The raw data sequence is often represented by $\{x_t\}$, and the output of the exponential smoothing algorithm is commonly written as $\{s_t\}$, which may be regarded as a best estimate of what the next value of x will be.

We used ARIMA model in our project and used multiple regression also for forecasting the appropriate amount of ATM cash .because our data is much like which depends upon its past values so we have taken the past p values and we also considered it as a moving average model where forecasting depends on random error terms which follow a white noise process. We also considered it as an integrated model, which may be of order zero, one or two.

A series which is stationary after being differentiated n times is said to be integrated of order n. Therefore, for training and forecasting we have to remove the stationarity that is why we are using integrated model here hence by considering all our forecasting needs we can say that ARIMA model is best-suited model for our forecasting using time series analysis. We are not considering the data from different subjects like data of different atms hence we are using only the time series data, here we are not using cross sectional data. Our data is in a series of time period of intervals. We considered the all types of components like whether it is a trend, seasonal cyclical or random data. We are considering different patterns observed monthly, annually, daily, weekly, quarterly.

II. RELATED WORK

Numerous authors had done work to alleviate ATM cash problem. Ahmadreza Ghodratia et al. too present research on ATM cash management. They used genetic algorithm approach to determine refill amount of each ATM.

The data in this research is collected via survey, consisted of transaction date for year 2011-2012 of an Iranian bank Ayandeh (Ayandeh Bank of Iran). In this research authors on the basis of their exploration concluded that some of bank ATM need to be upload in 3 days while some of ATM should be upload with cash on daily basis. Saad M. Darwish also worked to improve the estimating correctness of ATM cash demand. In 2013 the approach which he used is the extension of ANN that is an Interval Type-2 Fuzzy Neural Network (IT2FNN). He used the simulated data of 25 ATM for his experiments the structure set of date he used consist of: everyday, weekly and monthly seasonality localized sudden changes (gazette/public holidays and festival effects) were used to imitate the customers' money withdrawal from ATMs that are categorized by different transaction volumes. The experiment showed that average forecast accuracy (per week) of the proposed technique is about 97.72% while the minimum forecast exactness is 94.15%. Abirami et al. 2014, used data mining approaches to deal with ATM cash prediction.

The key objectives of their work is to provide easy identification of ATM norm and to monitor the ATM usage (peak) time, so that ATM must be available when it is needed most. The data consist of 30 day transactions of a day for testing purpose and later it was extended to 30 ATMs

transaction carried in a month. The results in their work were predicted from the former data (past data) and track the appropriate solution which is needed. It happens that the types of transaction, which carried out frequently, are characterized and on the basis of that characterization, services will be provided to the customers which is better than the normal. Also peak hours were identified which demonstrate that at what day of a month a customer will use particular ATM most, so that more comfort can be provided in peak hours. Simutis et al. demonstrate an approach that based on artificial neural network (ANN) to forecast a daily cash need for every ATM in the bank network and they invented a procedure for cash upload of ATMs. They discussed existing solution for ATM cash flow prediction and observe bank network which comprises of 1225 ATMs. In their technique; they discussed the most important factors for ATMs maintenance such as cost of cash, cost of uploading and cost of daily services. Their work showed that in case of higher interest rate and minimum cost of money uploading in ATMs, their procedure reduces the ATMs maintenance costs up to 15-20%. However, they pointed out those further experimental studies that are necessary for the practical execution. Erol Genevois et al. highlighted the problem of ATM location and cash management in automated teller machines in 2015. In their research they discussed two problem which bank are, facing one is finding suitable location for ATM and other is cash management strategy. The author in this paper suggested new ways to plant new ATM according to its location. In addition, they also discussed in detailed regarding various technique opted by other researcher for optimal cash management.

III. PROPOSED SYSTEM

We are going to use multiple regression and ARIMA model for time series analysis to forecast and to solve the cash outflow problems in ATMS. First, we have the csv file of the ATMS data this data is an example of big data hence to analyse this data efficiently we used Hdfs using Hadoop architecture. It takes the data and divide it into different parts at different nodes .now this data is stored in hive metastore and for statistical analyses we used r to plot the scatter plot of data, which will predict the amount of cash that can be entered to resolve the cash outflow problem.

First, we trained our model on the basis of multiple regression including linear, quadratic and polynomial regression and after that for better forecasting, we used previous n values to predict the next outcome by using auto regressive model. After that we deduced that our data is moving around an average mean value hence we forecast the next value by the combination of previous values of variables and the previous errors and then we removed the stationarity of data therefore we used ARIMA model.

IV. PROBLEM DEFINITION

Problem Definition is "To manage the cash input into the ATM's by analyzing the varying trends subject to occasions such as holidays, festivals, fiscal burden, timings and various other aspects which may cause dynamism in the overall requirement to avoid cash inflow cycle outburst using open source technology.".

V. SYSTEM DESIGN

A. System Workflow

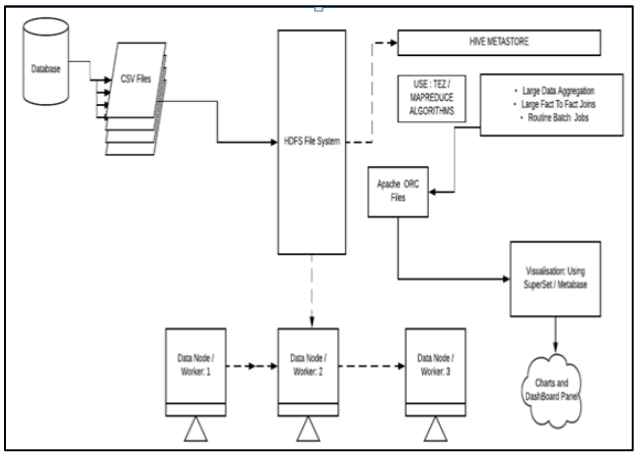

Fig. 1: Component Diagram

Fig.1 shows workflow of the whole system, which consists of Trained model and elastic search.

B. System Modules

1) CSV files

CSV is a simple file format used to store tabular data, such as a spreadsheet or database. Files in the CSV format can be imported to and exported from programs that store data in tables, such as Microsoft Excel or OpenOffice Calc. CSV stands for "comma-separated values

2) HDFS

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project. Documents.

3) Hive

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis. Hive gives a SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. Traditional SQL queries must be implemented in the Map Reduce Java API to execute SQL applications and queries over distributed data. Hive provides the necessary SQL abstraction to integrate SQL-like queries (hiveql).

VI. USER INTERFACE

User will be provided with a web app interface to analyse and visualize the cash inflow and outflow of various ATM branches in different cities.

VII. INPUT DATASET

Input dataset includes bank holiday csv file and raw data csv file of atm transaction data which contains fields like

Atm_raw_data fields:

Id

Atm_id
Aggr_period_from
Aggr_period_till
Tot_wd
Sum_train_amt
Bank holiday fields:
Tdate
Wday
Holiday
Holiday_seq
Fest_religion

VIII. CONCLUSION AND FUTURE SCOPE

In this project we have highlighted the main aspect for undertaking this project. We have tried to explain, to our best possible, the main motivation for moving forward with such a project. We have tried to cover various works that have been done in this field of research through our literature survey.

We have tried our best to make the introduction as abstract and absolute as possible. The introduction comprises of all necessary details to make it easier to move forward with an understanding of the context but we have also ensured that it doesn't dive deep into the specifics of the topic, which might have resulted in increasing the ambiguity and making it more complex to interpret and process. We have tried to come up with a short and concise but elaborative and self-explaining problem statement for our project.

Based on whatever knowledge we gathered from literature survey of this research area we tried to come up with a comprehensive and complete understanding of the matter. We tried to gather each and every requirement for this project in the most unambiguous manner possible. We came up with the software requirement specification for this project which we hope will act as guideline or rulebook to be referred at various stages of project development.

We have decided the development model to be followed for the purpose of implementation of the project. As mentioned, we will be moving ahead with Extreme Programming Model for the software development life cycle of this project. We have formulated the entire system implementation to act as our guiding light for the development of this project. We have performed system design for this project by carefully taking into account each and every requirement of this project. We have performed both static and dynamic modelling for the purpose of complete design analysis of this project. For static modelling we have developed the class diagrams and component diagrams and for finishing with the dynamic modelling we have developed sequence diagram and activity diagram for this project.

We hope that with this comprehensive analysis of the project we would be able to develop our project without much hassles and complex issues.

REFERENCES

[1] Avril Coghlan. Little Book of R for Time Series. Available from http://a-little-book-of-r-for-timeseries readthedocs.org/en/latest/.
[2] http://www.dnaindia.com/money/1391225/report-atm-refilling-costs-up-as-withdrawal-limit-ishiked
[3] Joe Chong. Forecasting With MS Excel. McGraw-Hill; 1 Edition (January 11, 1993)
[4] Mirai Solutions GmbH. XLConnect: Excel Connector for R. Retrieved from http://cran.rproject.org/web/packages/XLConnect/index .html, http://mirai-solutions.com.
[5] Rob J Hyndman. fore cast: Forecasting functions for time series and linear models. Retrieved from http://cran.r-project.org/web/packages/forecast/index.html.
[6] Rob J Hyndman and George Athanasopoulos. Forecasting: Principles and practice. Available from http://otexts.com/fpp/.
[7] Ruey S. Tsay. An Introduction to analysis of financial data with R. John Wiley &Sons Inc., Hoboken, New Jersey.

