ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

Approximation of Rising Netbanking Operations Using Data Mining with Big Data under Cloud

R.Kabilan, Dr.N.Jayaveeran

Research Scholar, P.G & Research Department of Computer Science, Khadir Mohideen College, Adirampattinam, Thanjavur – District. India.

Associate Professor and Head, P.G & Research Department of Computer Science, Khadir Mohideen College, Adirampattinam, Thanjavur – District. India.

ABSTRACT: Banking industry, today is observing an Information Technology revolution. The shift towards internet banking is powered by the changing dynamics. Internet based technologies have changed the way we do banking. This paper analyses the net banking operations using data mining with big data under cloud computing. A data model is designed for prediction and approximation. The experimental results showed that this model is expandable, maintainable, and manages the massive data with high efficiency.

KEYWORDS: Hidden Markov Model, Baum Welch algorithm, K-means clustering, Cloud Computing

I. INTRODUCTION

The proposed data model analyses the past net banking operations and identifies the significant on the data. These identified patterns from the past data enable us to approximate the upcoming future consequences. To propose and build up such an exact data model a number of methods are reviewed and most promising approaches are collected. The model combines data mining techniques, big data and cloud computing. In addition, the cloud computing can offer stability, reliability and scalability when in service huge-scale application in virtual computing environment. Thus the proposed data model incorporates the parallel K-means clustering, Hidden Markov Model and Baum Welch algorithm

II. LITERATURE SURVEY

Many researches were released on Time series prediction, in this section we will discuss some of these work .In [2] Rohit Kumar Yadav, Ravi Khatri, proposed data model incorporates the Hidden Markov Model designed for prediction and for extraction of the weather condition observations the K-means clustering is used. In[9]Ahmad Tamimi, Mohammed Aldasht proposed an approach for Parallel Feature Extraction by Hidden Markov Model and Parallel K-Mean Clustering for Protein Sequences. In[12] Lev Brailovskiy and Dr.Maya Herman compare HMM with other models when used for prediction of financial time series. In[13] Shawn Hymel,Ihsan Akbar, Jeffrey H. Reed. Provedspeed benefits from parallelization are maximized when a large number of HMM states are used. In[1] Ran Jin, Chunhai Kou, Ruijuan Liu and Yefeng Li results provide research basis to better design a clustering partition algorithm in large data and high efficiency.In[14]Dr.Doreswamy and Ibrahim Gad provide a system that uses the historical weather data of a region and apply the MapReduce and Hadoop techniques to analysis these historical data.

III. DATA MINING WITH BIG DATA UNDER CLOUD

A.*DATAMINING*

The data mining techniques are offered to analyse the past data and prepare their experiences. This knowledge is used to recognize the similar type of data for classification, pattern extraction, building prediction and approximation. In this

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

proposed model the machine learning based classification, pattern extraction and approximation is considered in detail. In addition of that using the data mining technique an innovative approximation model is prepared using the hybrid technique of machine learning.

B.*BIG DATA*

To analyse and mine valuable information from the huge data has become a crucial problem as the traditional methods are hardly to achieve high scalability in data processing. Newly, MapReduce has come into view a major programming model in dealing with big data analytics. Apache Hadoop, is an open-source software platform implementation of MapReduce, has been extensively taken up by the society. Hadoop make easy the utilization of a large number of inexpensive commodity computers. In addition Hadoop provides support in dealing with faults which is particularly useful for long running jobs. In this proposed model MapReduce improves the clustering speed in large data and high efficiency.

C.*CLOUD COMPUTING*

Running Hadoop efficiently for big data needs clusters to be set up. Developments in the virtualization technology have significantly reduced the cost of setting up such clusters; however they still require major financial investments, certificate fees, and human interference in most cases. Cloud computing offers a gainful way of given that facilities for calculation and for processing of big data and also serves as a service model to support big data technologies. Various open source cloud computing frameworks such as Open Stack, Eucalyptus, OpenNebula and Apache CloudStack. We can set up platforms as a service (PaaS) such as Hadoop on top of this infrastructure for big data processing.

IV. PROPOSED DATA MODEL

The proposed data model is shown in figure. In this model the different components of predictive data model is established. These components are used for forming the complete system.

This work method contains three algorithms such as Clustering, Hidden Markov Models and Baum-Welch algorithm were parallelized. In the following sections detailed description of each algorithm will be discussed.

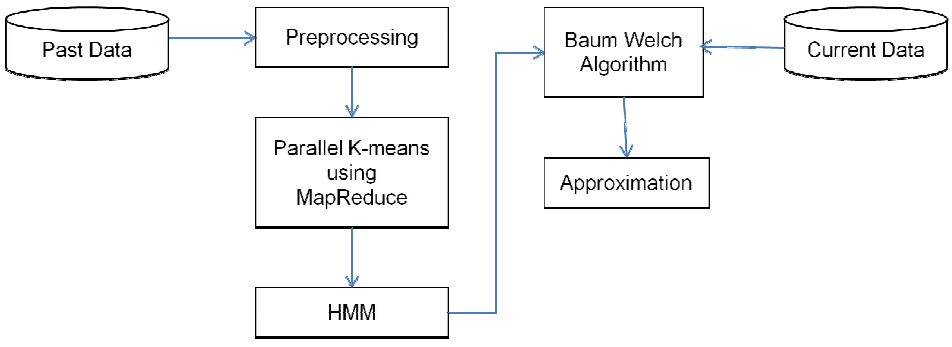


Figure 1. Proposed Model

Past Data: It means collected data about past events and circumstances pertaining to a particular subject. It is the digital information outlining activity, conditions and trends in an organization's past. In this model Data is collected from Bank wise volumes in ECS/NEFT/RTGS/Mobile Transactions, according to the Reserve Bank of India open data.

Pre-processing: The pre-processing is a method by which the data is refined, distorted and cleaned for improving the quality of the past data on which the data model is prepared for approximation. Cleaning and filtering of the data might

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

be clearly passed out with respect to the data mining algorithm. The data has fields like bank, outward debit transactions, received inward debit transactions and amount.

A.*PARALLEL K-MEANS USING MAPREDUCE*

Parallel K-means algorithm requires a kind of MapReduce job.

The Algorithm flowchart is shown in figure. Transaction record is filtered which is input to the tool. From the filtered records, transaction values are extracted. After that each transaction values are aggregated and their minimum, maximum and average is calculated.

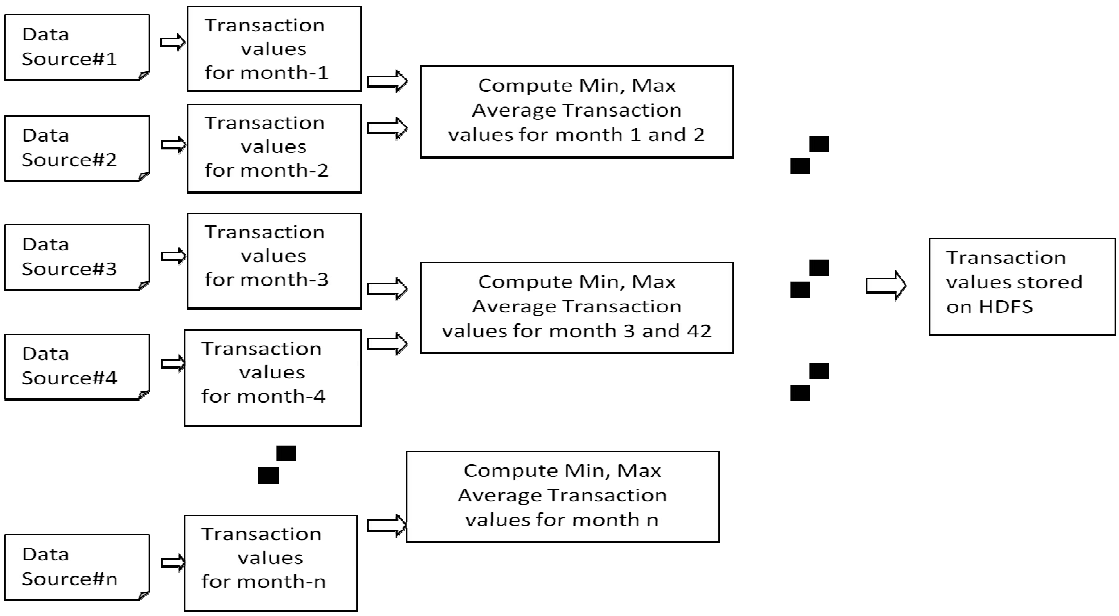


Figure 2.k-means parallel process based on map reduce

B.*HIDDEN MARKOV MODEL*

Hidden Markov Models (HMM) is a powerful machine learningmodel. HMM's main usage has been in solving classification and pattern recognition problems. In recent days, attempts have been made to use HMM for prediction ingeneral and prediction of time series in particular. However, this is not straightforward. To overcome the challenges in predicting time serieswith HMM some hybrid approaches have been applied.

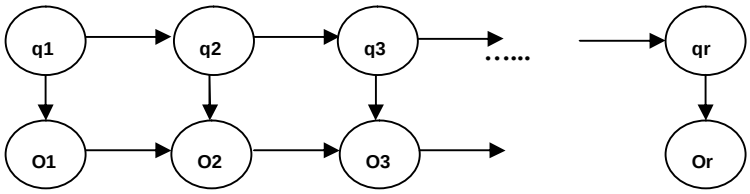


Figure 3. Hidden Markov Model

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

In the Hidden Markov Model, we have a series of Observations O = (o1, o2… oT). Eachobservation has M possible values (oi = v1, v2 …vM). The observations are directlycontrolled by a set of hidden variables which we cannot see X = (q1, q2… qT), where each hidden variable has N possible values (qi = s1, s2 …sN).

The main property of Hidden Markov Model is the Markov property, which is that the hidden variable qi is only depended to its previous hiddenvariable$q$i-1. All hidden variables before qi-1 have no effect in qi.
We use a set of parameters λ = {A, B, Π} to represent a Hidden Markov Model where:

Transition Matrix A = {ai,j} s.t.ai,j = P(qt = sj | qt-1 = si)In this transition matrix the transaction values and next transaction values are organized on the basis of the matrix.
Observation Matrix B = {bi(vk)} s.t. bi(vk) = P(ot = vk | qt = si)

It is the matrix on which the transaction values stored on HDFS is organized in order to grow the observation matrix of the Hidden Markov Model. Moreover the data objects class labels are recognized here as the states of the events. These states are the usual events which are necessary to predict.

Initialization Vector Π = {πi} s.t. πi = P(q1 = si)

The hidden Markov model is responsible to accepting these two matrixes as input and producing the learned model for prediction.

These algorithms parallelizedby distribute all the sequences to the available process. Let P be the number of processor. Each processor must handle O/P sequences using the preceding algorithm.

Current Data: The present value computation allowed for all of the data to be significant because it all was comparable on the same scale. In order to predict the upcoming transaction value the system required to take input the pattern, based on the observation and transitional patterns the system generate the next possible pattern.

C.*BAUM-WELCH ALGORITHM*

This algorithm is used for unsupervised training of a Hidden MarkovModel (HMM). It is an iterative algorithm which reviews the HMM model parameters in each iteration until a local maximum is reached. In each iteration, it computes the forward variables and backward variables.

The forward variables are denoted αt (i) where 0≤t≤T−1 and 0≤i≤N−1 for an observation sequence of length T and N hidden states. The backward variables βt (t) are defined similarly.

In each iteration, the Baum-Welch algorithm uses the forward and backward variables to revise the HMM model parameters, namely, the transition probability matrix, the emission probability matrix and the initial probability matrix.

These algorithms parallelized by distribute all the sequences to the available process.

Approximation: It is the result obtained from the proposed model. To foresee with HMM we will exploit our understanding of the past data. The rational is if somewhat happened in the past that is some pattern appeared, than it will with high probability appear in the future. Therefore we will look for similar to pattern and within this patter most possible study that will be used to analyse our approximation results.

V. EXPERIMENT AND ANALYSIS OF RESULTS

Dataset that used to test this model contains Bank wise Volumes in ECS/NEFT/RTGS/Mobile Transactions, according to the Reserve Bank of India open data from October 2015 to October 2016.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

A.*SET UP THE EXPERIMENTAL ENVIRONMENT*

By building the Hadoop cloud computing platform to validate the effectiveness of the model. The hardware includes five PC machines, one machine as a master node, loading scheduling and real time monitoring of task, the remaining four machines as aslave node, loading distributed processing of task. Each node has 16 Intel(R) Xeon(R) CPU E5-2670 (2.6GHz) cores, 64GB RAM and 250GB storage capacity. All the nodes were interconnected by a 1 gigabit Ethernet switch. Hadoop version 2.7.0 was used to setup the cluster.

This model can be evaluated based on accuracy and speed up. For each number of nodes parameter we did three experiments.
Communication time (T_{comm}) - Time that its jobs use receiving and sending messages.
Calculation time (T_{calc})- Time spent manipulating calculation rather than communication or idle.
Total execution time (T) - Itcan be well-defined in two methods: (i) The total of calculation, communication and idle times on an arbitrary processor. (ii) The total of these times done all processors divided by the number of processors P.
Efficiency- Fraction of time those processors employ doing helpful job. It can be defined as $E_{relative} = T_1/PT_p$

Nodes	1	2	4	6	8	10
Average Communication Time	0	0.12	0.9	0.067	0.12	0.51
Average Calculation Time	4636	1003	1112	994	1063	1089
Average idle Time	0	90.2	54.06	44.73	39.47	37.21
Total execution Time	4636	1093.32	1166.96	1038.8	1102.59	1126.72
Average Real Time	4636	1119	1273	1083	1097	1095
Speedup	1	2.71	3.22	3.86	3.91	3.77
Accuracy (%)	99.2	98.32	98.33	98.22	98.74	98.14

Table 1.Experiment Results

Table1 and Figure4 show the average results in details for these experiments. All averages time appears in the table were measured in seconds.

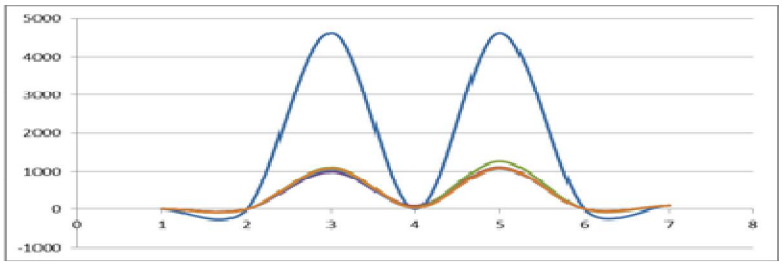


Figure.4Experiment Results

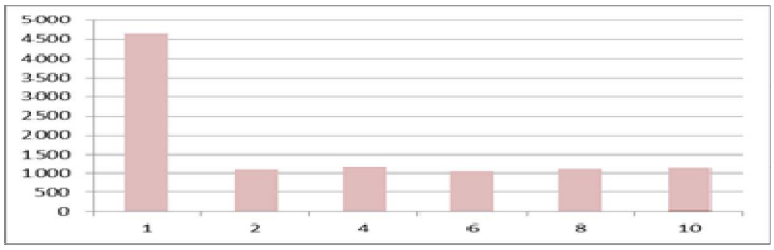
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017



Figur5. Total Execution time

Figure.5 presents Total execution time. Figure.6 presents the distribution of execution time between communication, calculation, and idle time. We note that if the data size is small, the communication time is almost zero. The idle time is also very low compared to computation time, this occur since our problem contains a big parallel independent portion.

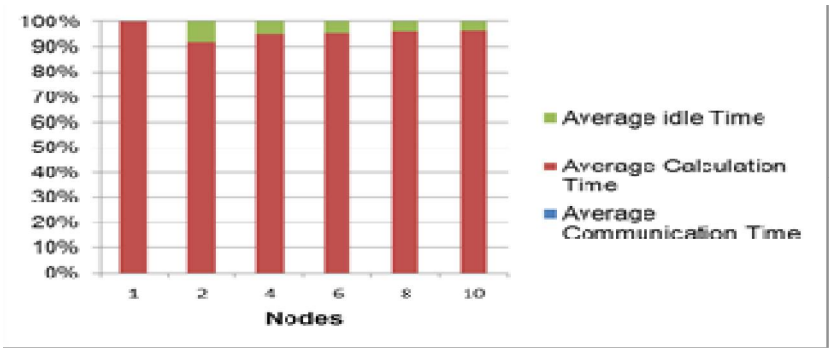


Figure6. Distribution of execution time

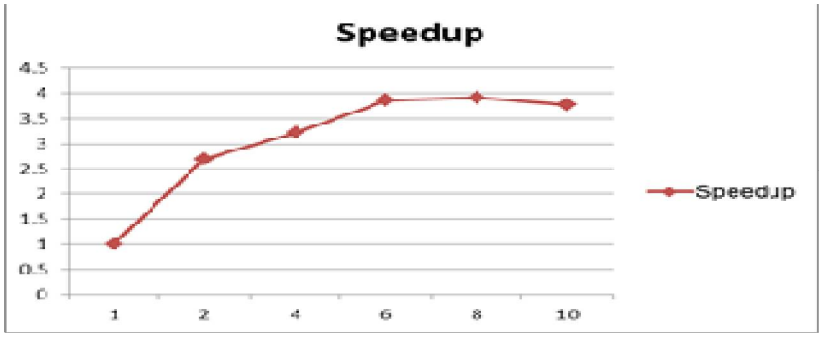


Figure 7. Speedup curve

Speedup attempts to assess the capability of the parallelism to improve the execution time. In this model the speedup obtained by dividing the average real time when having one node minus on average real time for parallel nodes. Figure 7 presents the speedup curve.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 3, March 2017

VI.CONCLUSION AND FUTURE WORK

This paper analyses the net banking operations using data mining with big data under cloud computing and provides a model that combines and parallelize k means algorithm, Hidden Markov Model and Baum Welch algorithm through the deployment of Hadoop. The future scope of this is it can be extended to any enormous data sets with different parameters / attributes for effective analysis, approximation and accurate prediction.

REFERENCES

1. Ran Jin, Chunhai Kou, Ruijuan Liu and Yefeng Li,' Efficient parallel spectral clustering algorithm design for large data sets under cloud computing Environment Journal of Cloud Computing' Advances, Systems and Applications 2013 Springer.
2. Rohit Kumar Yadav, Ravi Khatri, 'A Weather Forecasting Model using the Data Mining Technique, International Journal of Computer Applications (0975 – 8887) Volume 139 – No.14, April 2016.
3. Jianbin Cui, 'Parallelizing K-means with Hadoop/Mahout for Big Data Analytics' Thesis.
4. Katrina Leigh LaCurts, 'Application Workload Prediction and Placement in Cloud Computing Systems' Thesis.
5. Haiyan Song, Leixiao Li2 and YuhongFan,'Applied research on data mining platform for weather forecast based on cloud storage' 1 October 2014, www.cmnt.lv.
6.SamrudhiTabhane, Prof. R.A.Fadnavis,' Large data computing using Clustering algorithms based on Hadoop' International Journal of Engineering Research and General Science Volume 3, Issue 2, March-April, 2015, ISSN 2091-2730.
7. ChalabiBaya,'Implementation of a solution Cloud Computing withMapReducemodel'High Performance Computing Symposium 2013 (HPCS 2013) IOP Publishing.
8. DengZhenrong, Deng Xing1, Zhang Chuan1, Xu Liang1 and Huang Wenming, K-Means Algorithm Based on Parallel Improving and Applying' The Open Cybernetics &Systemics Journal, 2015.
9.Ahmad Tamimi, Mohammed Aldasht, 'Parallel Feature Extraction by Hidden Markov Model and Parallel K-Mean Clustering for Protein Sequences' acadamia.edu.
10. https://en.wikipedia.org/wiki/Hidden_Markov_model
11. https://en.wikipedia.org/wiki/Baum%E2%80%93Welch_algorithm
12. Lev Brailovskiy and Dr.Maya Herman.'Prediction of Financial Time Series Using Hidden Markov Models' ASE bigdata/social com / cybersecurity conference, Stanford University, May 27-31, 2014.
13.Shawn Hymel, Ihsan Akbar, Jeffrey H. Reed.' Parallel implementation of hidden markov models for wireless applications' sdr11 Technical conference and product eposition.
14. Dr.Doreswamy and Ibrahim Gad BIG Data Techniques: HADOOP and map reduce for weather forecasting International Journal of Latest Trends in Engineering and Technology Special Issue SACAIM 2016, pp. 194-199.