



## **ANALYSIS OF BANK'S UNSTRUCTURED DATA USING MAP REDUCE TECHNIQUE**

Jeny George<sup>1</sup>, A. Karthika<sup>1</sup>, Dr. Ulagamuthalvi V<sup>2</sup>

<sup>1</sup>Student, Faculty of Computing, Sathyabama University, Chennai, INDIA

<sup>2</sup>Asst. Professor, Faculty of Computing, Sathyabama University, Chennai, INDIA

**Abstract:** Data mining techniques have always been efficient and is still very much implementable, but Big Data concept is emerging. Data mining can process only structured data whereas Big Data deals with unstructured data. This system talks about tracking down and analysing the transactions of users who have multiple accounts in more than one bank. We particularly specify on transactions done in a specific or suspicious manner and generate a report on it using the hadoop tool. The multiple accounts and tracked down using the map reduce concept.

**Keywords:** Map reduce, Hadoop, Big data, Data mining, Unstructured data

### **I. Introduction**

Big data refers to extremely large data sets which makes it difficult for a typical database software tool to capture, store, handle and put into use. The reasons such an outbreak of Big Data is the increase of storage capacities, increase of processing power and abundant availability of data. It is capable of organizing information coming from multiple, heterogeneous, independent sources with complex and developing relationships and hence keeps growing. All of the world's data almost 90% of it has been generated over the last two years since the origin of Information Technology. Almost all of this of data have been generated from social networking sites (billions of unstructured data). There are many data types in Big Data like Banking, Insurance, Securities and Investment Services, Construction, Retail, Health care, Education etc. The large volume of available datasets allows us to apply more sophisticated models and get more accurate predictions about the future. Researchers have analysed influential and susceptible behaviour in social network using big data, similarly researchers have also done a thorough analysis of collective behaviour using swarm intelligence using large data sets concept [1][2]. Also papers and investigations on network analysis in social networks, social network and influences and also how twitter mood predicts the stock market etc have also been done which essentially proves that big data can be used for various analysis that can sound funny in the beginning but prove to be quite effective in the current scenario [3][4][5]. Also many algorithms have been introduced to mine big data throughout these years of the boom of data but none as such as the map reduce has been proven to be as effective as them. As you can see modern researchers also use this technique to mine large data sets [7]. No major development or analysis has been done on the banking sector which is what we have highlighted in our project. Big Data in action step wise: (i) Acquisition, (ii) Extraction, (iii) Integration, (iv) Analysis, (v) Interpretation, (iv) Decision.

In our country in most banks there are rules where people have to submit government proof when they transact money above a certain huge amount. Most people transact money on the margin line from various multiple accounts so as to avoid such submissions. This lot of people are who we analyse and separate in our project paper. It insists on three tier architecture: (a) Big Data implementation in multi system approach, (b) Application deployment-Banking, and (c) Extraction of useful information from unstructured data.

This paper is mainly implemented for banking domain. There will be two major departments: (1) Bank server for adding new clients and maintaining their accounts. During registration every user has to provide an ID proof to create an account in any bank. (2) Accounts monitoring server will monitor every user and their account status in different banks. This server will retrieve users who maintain and transact more than Rs. 50,000 / Annum in all 3 accounts in different banks using the same ID proof and thus map them and reduce those particular accounts.

### **II. Methodology**

The system design explains the flow of the control and how we have implemented the project:

#### **A. User Account Creation**

Initially, the user has to create his own account credentials and is only authenticated to access the network. Once the user creates as such, he is asked to sign in and request a particular service from the Service Provider. Based on it, the service will be processed and a response is send. All this information will be stored in the database of the Data Service Provider. In this system, we design an Interface to communicate with the server using the

programming Languages like Java/ .Net.If access is allowed by the server the user can access the articular services available to him/her by the service provider .

#### **B. Bank's Server**

Big Data Service Provider contains unstructured data abundant amounts of it in its database. Also, the Service Provider manages all the users details to validate them when they sign in into their account. These details will be stored in the database of the Service Provider. Also the Data Server redirects the users request to the Resource Assigning Module to process the particular request. The request of all the users in unison will be processed by the Resource Assigning Module. To connect to the client and with the other modules in the network, the server will establish a relation between them. This is why we are going to create the Interface as mentioned earlier. Also the Service Provider will send the user's request to the Resource Assigning Module in First in First out (FIFO) manner.

#### **C. Centralized server**

A sole sign-on protocol for the web, the CAS's purpose is to allow a user to access more than one application when the user provides their details (such as user id and password) only once. It also allows web applications to allow users access without gaining access to a user's secure detail, such as a password. It is also software package that implements this particular protocol.

#### **D. Mapping of User Account**

In this module we map or separate the users who have account in more than three banks with the same id proof and credentials. They are mapped by using Hadoop Map Reduce technique. The Hadoop tool basically implements the map-reduce algorithm to analyse the input of data.

**Map()** procedure: It performs filtering and sorting(such as sorting students by their grades, sorting account holders using their unique customer Id

**Reduce()** procedure: It performs a summary operation(segregating account holders based on their transaction amount

MapReduce system arranges the process by marshalling the distributed servers, parallelly handling multiple kinds, administrating all connections and data transfer between various parts of the system.In this paper, Map() procedure sort out the users who used the same ID for more than three bank accounts and Reduce() procedure segregates the users who made transaction more than Rs.50,000-Rs.1,00,000 per annum in all the three bank accounts.

#### **E. Account Transaction Review**

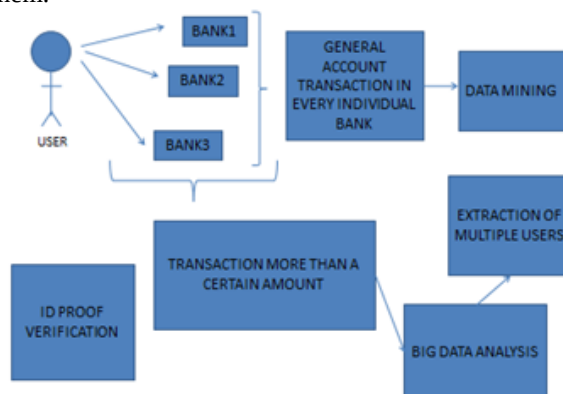
In this module we are getting information about the users who have accounts in three different banks and we also filter the transaction done by them. Finally we review the information transacted by the user that is, if they transact more than Rs.50,000-Rs.1,00,00/ANNUM in all the three accounts. We create a report based on their transactions in a day, month or a year depending on the threshold specified.

#### **F. Tracking of suspicious users**

By extracting the vital information from the unstructured data for transaction of bank service and tax service we analyse the ratio through our application, and produce the best output for both transaction and unpaid tax amount.

### **III. Architecture Diagram**

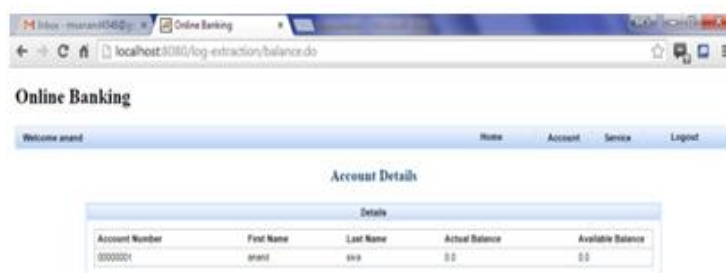
The diagram here shows the flow of the system in a gist. A user who has multiple accounts is tracked down using a unique id generated every time he signs in his account using the same address id. From this we generate his transactions and analyse them.



### **IV. Result and Discussions**

This is an example of a particular users transactions. So similarly when we search for a particular users transaction or log in to the admin of the bank and search for a specific user, based on the threshold specified the whole set of his transactions is generated. So in the tool when you type for a user it is checked if he has multiple

accounts and if so it is map and reduced and a report of his transactions is created. With this report we can analyse the amount of money hes transacted in a month.



## V. Conclusion

We have therefore found that big data can prove to be very useful not only in social media sites and networks but also in banking applications, medical sectors etc. We have successfully implemented the system and have been succesful in generating reports of users with multiple accounts and their transactions. Similar work can be done in the future for medical, insurance applications as well where we analyse common diseases or frauds respectively.

## References

- [1] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks," *Science*, vol. 337, pp. 337-341, 2012.
- [2] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," *Knowledge and Information Systems*, vol. 33, no. 3, pp. 523-547, Dec. 2012.
- [3] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," *J. Computational Science*, vol. 2, no. 1, pp. 1-8, 2011.
- [4] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences," *Science*, vol. 323, pp. 892-895, 2009.
- [5] D. Centola, "The Spread of Behavior in an Online Social Network Experiment," *Science*, vol. 329, pp. 1194-1197, 2010.
- [6] E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects," *Nature*, vol. 489, pp. 49-51, 2012.
- [7] C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, and K. Olukotun, "Map-Reduce for Machine Learning on Multicore," *Proc. 20th Ann. Conf. Neural Information Processing Systems (NIPS '06)*, pp. 281-288, 2006.
- [8] E.Y. Chang, H. Bai, and K. Zhu, "Parallel Algorithms for Mining Large-Scale Rich-Media Data," *Proc. 17th ACM Int'l Conf. Multimedia, (MM '09)*, pp. 917-918, 2009.
- [9] Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, "Efficient Algorithms for Influence Maximization in Social Networks," *Knowledge and Information Systems*, vol. 33, no. 3, pp. 577-601, Dec. 2012.