Integration of Big Data & Cloud Computing To Detect Black Money Rotation with Range – Aggregate Queries

K. Kedharewsari [#1], V. Maria Anu [*2], V. Rajalakshmi [#3]

[#] Department of Information Technology, Sathyabama University, Chennai, Tamil Nadu, India

[1] kedha.k@gmail.com

[2] Mariaanu18@gmail.com

[3] rajalakshmi.bala03@gmail.com

***Abstract*— the big data is difficult to be analyzed due to the presence and characteristics of huge amount of data. Hadoop technology plays a key role in analyzing the large scale data. The aggregate queries are executed on more columns concurrently and it is difficult for huge amount of data. This paper is proposing the method in which the fast RAQ is dividing the big data in to autonomous partitions by means of a balanced partition algorithm and later for each partition a local assessment sketch is generated. By the arrival of the range-aggregate query demand the fast RAQ gets the result in a direct manner by shortening local estimate from all partition and then the cooperative results are provided. Thus in fast RAQ technique three tier Architecture is insisted and they are of 1.Extracting the helpful information's from Unstructured Data, 2.Implementation of the big data in Multi system Approach, 3.Application Deployment – Insurance/ Banking. This paper is implement for the banking domain process and two major departments are involved in this process and they are 1.To maintain the accounts and for adding new clients the Bank Server is used. To create account in any bank the user have to give their ID proof at the time of registration.2.Account Monitoring Server is used for monitoring every users accounts in various banks and this server is used for retrieving the users who are maintaining and transacting more than Rs 50,000 per annum in various bank accounts by using the similar ID proof is identified by Map Reduce technique. The Online Aggregation is a smart sampling-based method that is performed to provide response to aggregation query by an approximation to the last outcome, with the self-assurance interval which is becoming tighter eventually. It is built into a Map-Reduce-based cloud scheme for analytics of the big data that allows the user to save the money by means of killing the calculation early and to observe the query progress when the enough accuracy is achieved.**

Keyword- Map Reduce, Partition Algorithm, Fast RAQ, Range-aggregate query, Tracking Black Money

I. INTRODUCTION

The analysis of big data is used for discovering the preference of everyday's individual behavior and trends of different social aspects. A new chance to explore the primary queries about the composite world is provided here. For illustration, the author pries analyzed the enormous behavior of data sets that are connected to money and an income of even 326 percent which is superior to an arbitrary investment policy is yielded to build the well-organized investment strategy. The approximation sketch to forecast the economic indicators like automobile sale, even destination for private travelling, social unemployment are presented by Varian and Choi. At present it is very necessary to provide resourceful method and tool for big data study. An application of big data analysis is given. The big data is defined as the breaking of huge quantity of the data in to minor parts for enhanced understanding. Big data is produced since every individual is using the commercial, internet and social access, Business sites and educational sites for accepting. Every individual likes to connect to his/her friends, colleagues, family by the means of internet. Now a day's knowledge of technology and new updates are done with the help of internet that is creating life more attractive and simple. Big data is developed in all the things and here the logic is performed that another way to earn money in business. The universes is moving rapid and the appearance is authentic world is turning to town. Every person needs a system to mingle among the world. The information's are documented by the clients like political issues, wellbeing data, Topographical regions, neural system and more. Media and the social locales is the alternative thing that is familiar by the big data. The social locales like Google with Gmail and whatsapp face book and preferably for the web index are consistently success because of several numbers of users as wide and far as possible. Learning the human long series interpersonal communication, doctors, mathematicians and numerous additional science fields by deal of data in a little determination of time [1] is enhanced by the locales. Each person wants the information by a solitary click. Preparing the enormous information is the basic errand. The few structures that are transformed are pig, Hive, Jaql like innovation assuming imperative part depicted in this. The flip kart submits an offer which is tremendously shabby on sixth Oct 2014.Bringing the elevated disjoin preparation is the short determination of time. Billions of appeals strike within 30 min according to the flip kart. For transforming and

dissecting the huge information's several advances that are specified previously are used. Majority of basic test that are performed by the massive Information request is that to examine the outgoing volumes of data and to provide concentration on learning for upcoming activities or helpful data. In frequent situation, the known extraction process that should be remarkably well-organized and near to continuous as putting away all the information that is watched is almost impossible. The information amount that is outstanding requires the forecast step to accomplish to rapid reaction and victorious information examination and even categorization of enormous Information. The main center is that the method on how the information is examined and recovered by a resourceful way. Classifying the data into character trademark and also examining the data mining difficulties is done by Gave HACE hypothesis. Currently a day's for master access on OLTP plus OLAP frameworks the Guide reduces sketch work is utilized, which are upgraded occasionally [2].Parallel execution is the guide decrease method that has one of the greatest trademarks. To handle the huge amount of information the HADOOP technique will use the parallel handling events in which Guide reduce technique is usually utilized. This is a straight forward system from the remaining of others. For transforming the enormous information the Allotment calculations and Bunch are used. These stuffs are providing the yields but not in completion and also their knowledge level is more mind boggling than others. Question map is more puzzled with scientific database. Mapping the queries of Big Data network sources provides an illustrative Meta - dialect for comprehending the significance of queries and guides them into every person. The utilized charts are superior part of query development to work and examine efficiently. The portion of chart investigation [3] is a pattern matching calculation. This calculations deal with live and appropriate information. Finding the examples which are recognized with the approaching information or friend is the main consequence of outline matching calculation.

II. RELATED WORK

Online aggregation and the data sampling are the two fields that are generated as a proposed work in our paper. The Online aggregation method was initially proposed in [4] involving the "group by" aggregations that is focused on the single-table queries. The focus in [4] is improved by the effort in [5] by giving the huge-sample and deterministic assurance time computing method in case of multi table and single-table and queries. The estimate and the query dispensation algorithms are considered in the circumstance of joins above multi-tables [7, 6 and 8] for OLA. In [6] ripple joins known to be the family of link algorithms is presented. The hash joins and the conventional mass nested-loops are the basis of Ripple joins that are developed to reduce the time. The effort in [8] is extending the unique ripple joins that is used in speeding up the junction by parallelizing the sampling and query processing. If the memory overflows and the accurate data sharing are unknown then statistical guarantee cannot be provided. The effort in [7] is making the query approximation and maintaining the probabilistic self-assurance bounds even if the input is not fitting in the memory. In order to run the uncertainty processing concurrently the shrink phase is added and the approaches which combine the ripple joins algorithms and traditional sort-merge joins with the combine phase to inform the result. In the scattered environment the online aggregation is complete in [9] the scattered hash table network. All the above effort is in the circumstance of conventional databases. In the background of cloud computing the online aggregations investigate is renewed at present and a few studies are based on the process of Map Reduce [10, 12]. Hadoop Online Prototype (HOP) [10] is pipelining the MapReduce technique of Hadoop technique that allows the posterior operators to consume the production of sign operators previous to the originator operator to complete. The unique snapshots of the MapReduce jobs at the data in need intervals can be provided by HOP and scaling up the snapshots with the job development is supporting the OLA without any such confidence limits of the query estimate. To implement the OLA over the Map Reduce [11] the Bayesian structure based approach is used. During the estimate processing this approach consider the connection among aggregate rate of every mass and the processing time where the each blocks processing time and scheduling time is taken into account as the observed data. The approach is focusing on the solitary table aggregate query which is having one MapReduce job and not taking into account of the aggregate query over joined numerous tables, which are having the numerous MapReduce jobs. The data sampling is necessary for the online handling of the aggregate queries and in the DBMS field lots of job has been done. Row-level sampling [12] plus block level sampling or page level samplings [13] are the sampling of two levels component in the obtainable sampling techniques. True uniform-randomness that is the origin of several approximate algorithms is provided by the Row-level sampling. Since data is forever clustered by lots of pages or blocks the row level sampling will be very costly. When generating statistics the block level sampling is further proficient but it is flat to errors. The force of block-level sampling resting on statistic estimation for distinct-value estimation and histogram and the equivalent statistical estimators with block-level samplings proposals is analyzed in the work [13]. The page level sampling method and the row level sampling technique is combined by a block level sampling system which is projected in [11].Every work is carried in the ground of single-site DBMS. In the field of dispersed DBMS, the efficiency and accuracy of diverse sampling method for query amount opinion in the equivalent DBMS is compared in the work [14], by the use of simple random sampling and stratified random sampling with the component of page-level and row-level [15,16]. In the online aggregation of spread environment [16, 8] the stratified sampling is

healthy worn. In the circumstance of random sampling [10, 9], online aggregation in the cloud is assumed by the accessible job of OLA over Map Reduce, and no particular sampling techniques are projected.

A. Problem Definition

The range-aggregate issue in big data environment is the issue in this paper. To analyze a huge quantity of data there is no technique followed. The data occurrence in each part is not properly utilized. To process a huge quantity of data Hadoop tool is used.

III. FAST RAQ

A new estimated reasonable approach that produces accurate estimations rapidly for range-aggregate query in big data environments is known to be the Fast RAQ which is our proposed approach. For Fast RAQ the ad-hoc range aggregate queries have the Oð N PBÞ time complication and Oð1Þ time difficulty for data updates. FastRAQ will have Oð1Þ time complication for range collective queries when proportion of edge-bucket cardinality (h0) is little. It is believed that the Fast RAQ process is providing a better initial point for mounting real-time answering system for big data analysis method.

A. *Black Money Detection Using Map Reduce Technique*

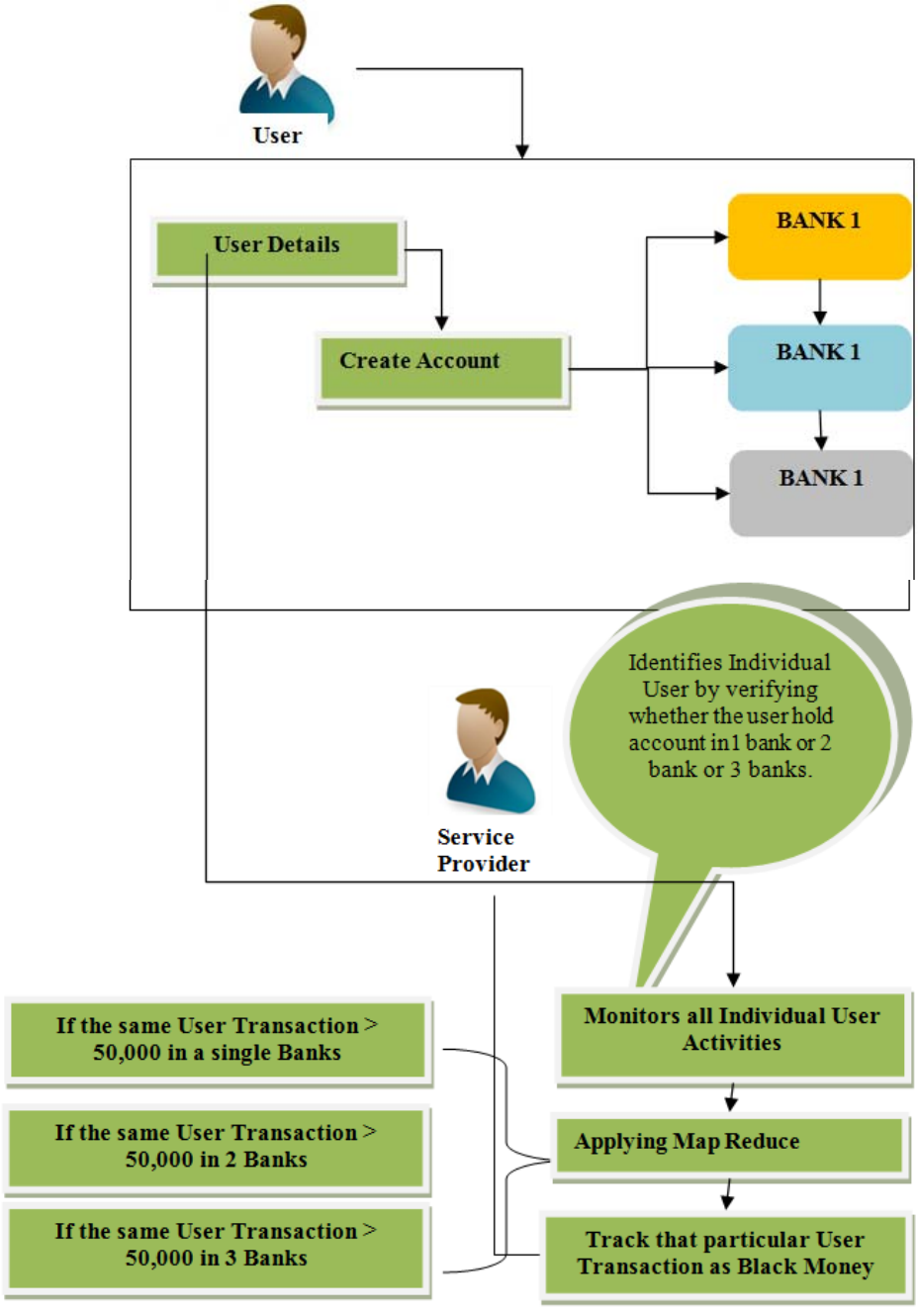


Figure 1 Black Money Detection Using Map Reduce Technique

B. Creating Account in a Bank

The user has to generate an account through similar ID proof in three dissimilar banks. So that the three various banks which is having the similar ID proof with the similar user details allows the client to use the network. The user need to login to their account once their account is been created and then have to request the job from service provider. The Service provider will develop the User request Job and provides a response to them which is based on the User request. The Database of the Data Service supplier stores all the user details. Using the programming languages like Java/ .Net the User Interface Frame can communicate with the Data Server throughout Network code and this is designed in this paper. As a result of transferring the request to Server Provider, the server authenticates the User and can admission the requested data if they are genuine by the Server.

C. Server Provider

Huge amount of data are stored in the Big Data Service Provider data storage. The service provider will preserve all the information's of the user to authenticate when the users are logging in to their respected account. The Database of the Service Provider stores the user information. To process the requested Job of User the Data attendant will forward the User requested job to the supply conveying unit. The resource assigning module will process all the request of the users. The communication with the additional modules of the Network and the Client is done by establishing a connection between them by means of the Data Server. A User Interface outline should be created for this purpose. The User Job demand is sended to the Resource Assign unit by Cloud Service source in a Fist in First out (FIFO) manner.

D. Map Reducing Technique

This module explains about the users who are having account in additional three banks and can be mapped by using the Hadoop map reducing method. The data about the clients who gave extra three accounts in the bank is obtained in this module and also the transaction done by the user can be filtered and we can evaluate the information that is transacted by the user to their user throughout for manual or online.

E. Tracking Black Money

The needed information is extracted from the shapeless data for the tax service and bank service transaction. Hence the most excellent output for both the transaction and unpaid tax amount is produced and the ratio throughout our application is analyzed.

F. Algorithm

Partition Algorithm

Input: Details (D), Vector Set

Output: Partition identifier PID

1. Detail has to resolve into different column families.
2. Calculate Class Identifier (CID) with value ranges

Get Segregation vector VSG from Vs with CID

3. VSG =< CID ; V range > Set target for Partition identifier,
4. P I D =< CID; random [1; V pi * Vrange] > ; Build
5. Sample in partitioning P I D ;
6. counters P I D counter P I D + 1;
7. Add PID Add PID + N;
8. sample PID add a; b ;c ;range =counter PID ;
9. SID Hash (PID; counter PID);
10. Send Detail to partition PID; return PID;

FARQ

Input: P d ;

P d: Select add other Col Nam where

z_i 1 < Col Nam < z_i 2 ,z_j 1 < Col Nam < z_j 2

Output: RA;

R A: result set of aggregate query.

1. Request P d must be provided to all partitions.
2. for each part i in partitions do
3. Evaluate cardinality of range

z $_i$ 1 < Col N am i < z i 2 from the histogram and let calculate CE i be i th dimension estimator.

4. Estimate cardinality of range z_j 1 < Col Nam j < z_j 2 from histogram and let CE j be j th dimension evaluator.

5. combine estimators CE i and CE j by the logical operator, and evaluate combined cardinality CE combine .

6. Count i h (C E combine) h - function of cardinality evaluator.

7. Calculate samples for Agg Col and let sample i be sample,

8. ADD i count I sample i : //ADD i - result of local range-aggregate query.

9. end

10.set almost answer of aggregate Query RA .

11.Let R A P N

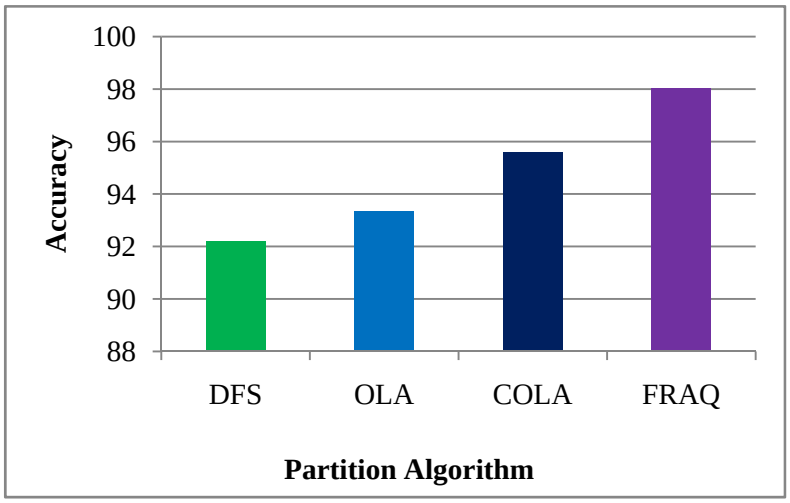
i =1 ADD i ,

12.where N- number of partitions. Return RA

IV. EXPERIMENTAL RESULT

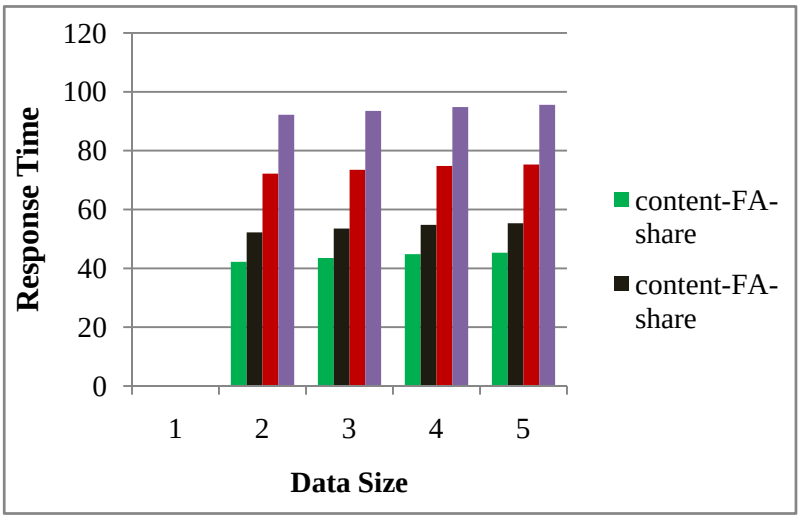
The performance of our proposed approach is identified by an experimental result which is conducted with the following requirements. CPU G2020, Windows 7, processor speed of 2.90 GHz and Intel Pentium are the subsequent configurations are used to execute our projected methods.

A. Accuracy Measurement



The above Figure 2 is presenting the accuracy of the proposed technique FRAQ which is better than other existing technique. The proposed technique is providing more accuracy over the existing issues.

B. Response Time Computation



The above figure 3 is presenting the Response Time of the proposed method Hadoop which is better than other existing technique. The proposed method is providing more response time over the existing issues.

V. CONCLUSION

Big data is known to be the uncertain, real time and unstructured data that are present in an enormous amount. Even there are different technologies existing in today's world querying on such data is a quiet challenging task. The exact pattern matching method and Balance partition method, proposed in this paper are useful for managing theses queries. The balance partition technique is used for dividing the big data into division at first and then it stores in particular partition. The indexing is provided in the partitions which are used through an accurate pattern matching method for successful managing of queries. Also the paper is implemented on the crown of Hadoop technologies which sustain the java language.

REFERENCES

[1] Arti Mohan purkar, Prasad kumar Kale "Efficient Query Handling On Big Data in Network Using Pattern Matching Algorithm: A Review" International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064.
[2] Prasadkumar Kale1 and Arti Mohanpurkar "Big Data Analysis using Partition Technique" International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2871-2875.
[3] P.Ravinder Rao S.V.Sridhar V.RamaKrishna "An Optimistic Approach for Query Construction and Execution in Cloud Computing Environment" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013 ISSN: 2277 128X
[4] P. J. Haas and J. M. Heller stein. June 1999 "Ripple joins for online aggregation". In SIGMOD 1999 Conference Proceedings, pages 287–298.
[5] T.Condie, N.Conway, P.Alvaro. June 2010 "Online aggregation and continuous query support in map reduce". In SIGMOD 2010 Conference Proceedings.
[6] P. J. Haas. August 1997. "Large-sample and deterministic confidence intervals for online aggregation". In SSDBM 1997 Conference Proceedings.
[7] J.M.Hellerstein, P.J.Haas, and H.J.Wang. May 1997. "Online aggregation". In SIGMOD 1997 Conference Proceedings.
[8] C. Jermaine, A. Dobra, S. Arumugam. June 2005. "A disk-based join with probabilistic guarantees". In SIGMOD 2005 Conference Proceedings.
[9] N. Pansare, V. R. Borkar, C. Jermaine, and T. Condie. August 2011. "Online aggregation for large map reduces jobs". In VLDB 2011 Conference Proceedings.
[10] S. Chaudhuri, G. Das, and U. Srivastava. June 2004. "Effective use of block-level sampling in statistics estimation". In SIGMOD 2004 Conference.
[11] F. Olken and D. Rotem. August 1989. "Random sampling from b+ trees". In VLDB 1989 Conference Proceedings.
[12] Wikipedia Page Traffic Statistics. Available at http://aws.amazon.com/datasets/2596.
[13] S. Blanas, J. M. Patel, V. Ercegovac, J. Rao, E. J. Shekita, and Y. Tian. June 2010. "A comparison of join algorithms for log processing in map reduces". In SIGMOD 2010 Conference Proceedings.
[14] F. Olken and D. Rotem. April 1990. "Random sampling from database files: A survey". In SSDBM 1990 Conference Proceedings.
[15] S. Seshadri and J. F. Naughton. March 1992. "Sampling issues in parallel database systems". In EDBT 1992 Conference Proceedings.
[16] S. Wu, S. Jiang, B. C. Ooi, and K. L. Tan. August 2009. "Distributed online aggregation". In VLDB 2009 Conference Proceedings.

Ms. K.Kedharewsari obtained her B.Tech degree from Sathyabama University. She is persuing M.Tech degree in the Department of Information technology from Sathyabama University

V. Maria Anu obtained her B.Tech degree from University of Madras. Then she obtained her M.E degree in Computer Science from Sathyabama University .She is currently working toward the PhD degree in the Department of Computer Science, Sathyabama University. Her research interests include data mining, RFID Data Management. She is Assistant Professor at Department of Information Technology, Sathyabama University.

Ms.V.Rajalakshmi received B.E degree from University of Madras in 2002, M.E degree from Sathyabama University in 2005. She is persuing Ph.D in Sathyabama University in the area of privacy preservation in data mining. She is in the teaching profession for the past 12 years and currently doing research. She is interested in developing computational algorithms, solving problems using neural networks, information security etc. She has published 14 research papers in various journals.