International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 03 | Mar -2017 www.irjet.net p-ISSN: 2395-0072

FAST RANGE AGGREGATE QUERIES FOR BIG DATA ANALYSIS

M.R.ABHIMAN RAAM[1], V.S.ARAVINDAKSHAN[2], P.A.HARISH[3] , S.KARTHIK[4]

[1] *Department of information technology, Valliammai Engineering College, Tamilnadu, India*
[2] *Department of information technology, Valliammai Engineering College, Tamilnadu, India*
[3] *Department of information technology, Valliammai Engineering College, Tamilnadu, India*
[4]*Department of information technology, Valliammai Engineering College, Tamilnadu, India*

Abstract - *In big data environment, Aggregate Queries are important tools in finding individual persons behavior, trends and various activities in the real world. The Aggregation is applied to employ aggregate function on all the tuples within a specified range. Existing approaches for this method is not enough to provide fast results for large datasets such as in banks, financial institutions, etc. It is important to provide effective methods and tools for big data analysis. In the proposed system, Fast RAQ divides big data into different partitions using Partitioning Algorithm and generates a local value for each individual partition. When a Query request is given, this algorithm obtains the result directly by grouping the local estimates from all tuples and provides a collective results. This system applies Fast RAQ for Banking Domain. The banking datasets are divided into multiple tuples and stored in different sets of the database across different places. This proposed method tracks multiple accounts maintained in different banks of same user and their transaction details. This helps in finding out tax violators using their unique id.*

Key Words: **Hadoop, Bigdata, Cloud, Fast RAQ, Balanced partitioning**

1.INTRODUCTION

Big data analysis is generally used to explore the hidden patterns from the large datasets. This provides a new approach to discover the solutions for the various difficulties in the real world. It is vital to provide a cost-effective and time saving methods and tools for the analysis in big data environment.

The main aim of the project is to identify the tax violators in the banking sector. To track the transaction of users in multiple banks and monitor them using Fast Range Aggregate Queries with Balance Partitioning algorithm. The result is obtained by summarizing local estimates from all the partitions and provides a collective results.

The transaction details of different banks are taken and stored in the cloud. The datasets are partitioned into different tuples before uploading it into the cloud storage. The algorithm partitions the datasets according to its attributes, interests, etc.

In this project, the time taken to process the given query is enormously reduced. It provides an efficient and accurate results for the large datasets. The Hadoop and map reduce concept is used for storage and processing of large datasets.

2. LITERATURE SURVEY

Xiaochun Yun, et al.. [1] This paper describes about the implementation of low cost and fast approach technique for getting accurate results in big data analysis using queries.

Zhiqiang Zhang, et al.. [2] proposed Hadoop online Aggregation in the distributed environment. The random sampling and sample size estimation are analyzed. This two sample values are calculated according to (1)user calculated sample value, (2)system calculated sample value. It also ensures that approximate aggregation results are produced.

A. Munar, et al.. [3] It is based on the highly scalable and fault tolerance map reduce model for the use of large scale database. It uses various big data analytics to handle systems with different requirement specification. This paper mainly focuses on providing good performance even when there is an enormous increase in the database.

Y. Shi, et al.. [4] Cloud Based system for Online Aggregation which provides progressive approximate aggregate answers for both single table and multiple joined tables.

3. PROPOSED SYSTEM

In the Proposed System, the data sets are divided into different partitions using the partitioning algorithm. Then a sample value is obtained from each individual partitions and the analyzes made on the datasets is obtained. When the query arrives another cost factor involved in the analyzing of big data are cost of network synchronization and the scanning of files in every transaction while passing the range aggregate queries meanwhile in our proposed system since our query is fast range aggregated it passes through every tuple, counter values from the aggregated columns and the sample values from the rows are calculated the cost of network synchronization of files and the scanning of files can be reduced. Which it leads to it produce of accurate result.

4. TECHNOLOGIES USED

4.1 Big Data Analysis

The process of analyzing large amount of datasets to explore the hidden patterns, facts, etc.

4.2 Hadoop

Hadoop is used for the purpose of storing and analyzing large datasets in the distributed environment.

4.3 My SQL Query Browser

It is used to store the datasets. It act as a backend in the project.

5. ARCHITECTURE DIAGRAM

The user's banking data is partitioned into multiple Tuples and stored in different sets of Database. Fast RAQ first divides big data into independent partitions with a balanced partitioning algorithm, and then generates a local value for each individual partition. When a Query request is given, this algorithm obtains the result directly by grouping the local estimates from all tuples and provides a collective results. A Map Reduce usually splits the input data-set into independent chunks which are processed by the map tasks in parallel. This sorts the outputs of the maps, which are then input to the reduce tasks. A particular value is fixed in order to filter required data from the whole bunch of database.

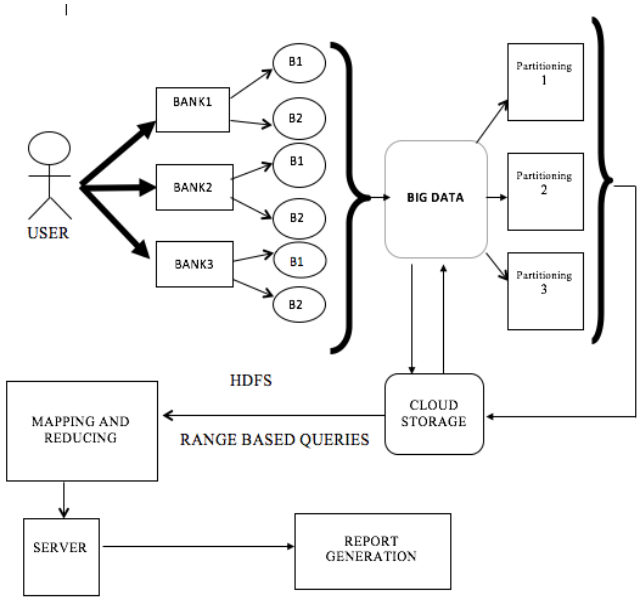

Fig -1: Architecture Diagram

This results in finding out tax violators.

5.1 User Registration

First the user wants to create an account using first name, last name, document id. After creating the account an unique account number and password is generated for user. All the user details are stored in remote server.

5.2 Remote Server

Remote server contain information about the user. Also the service provider will maintain the user information. It provides authentication when a user tries to login to the account. The user information will be stored in remote server.

5.3 Cloud Setup

The banking transaction of different banks from different locations are uploaded to the cloud storage.

5.4 Big Data Setup

Everyday several millions of transactions can take place in the banking sector. It is a challenging process to provide an efficient way to explore the big data. The transaction details may contain a single day transaction or weekly, monthly transaction.

5.5 Data Split-up/Partitioning

The large datasets are split into small individual datasets using the algorithm to analyze the datasets more effectively.

5.6 Map Reduce Framework

Map Reduce Frame work is a part of Hadoop. This module takes input data and converts it into a set of data, where individual elements are broken down into tuples. In our project the query is given as input and the request is given to job tracker. Task tracker process the request in all tuples and users with accounts more than in 3 Banks then those people is the input of Map task. The output from a map task as input to reduce task and combines those data tuples into a smaller set of tuples.

5.7 Unique Id's are tracked

The output from the map reduce frame work consist of the users who have account in three banks. These are tracked by the unique id. In this module we will track the users with deposited more than Rs.50000 per Annum in all three banks. This helps to find out the tax violator.

6. FLOW DIAGRAM

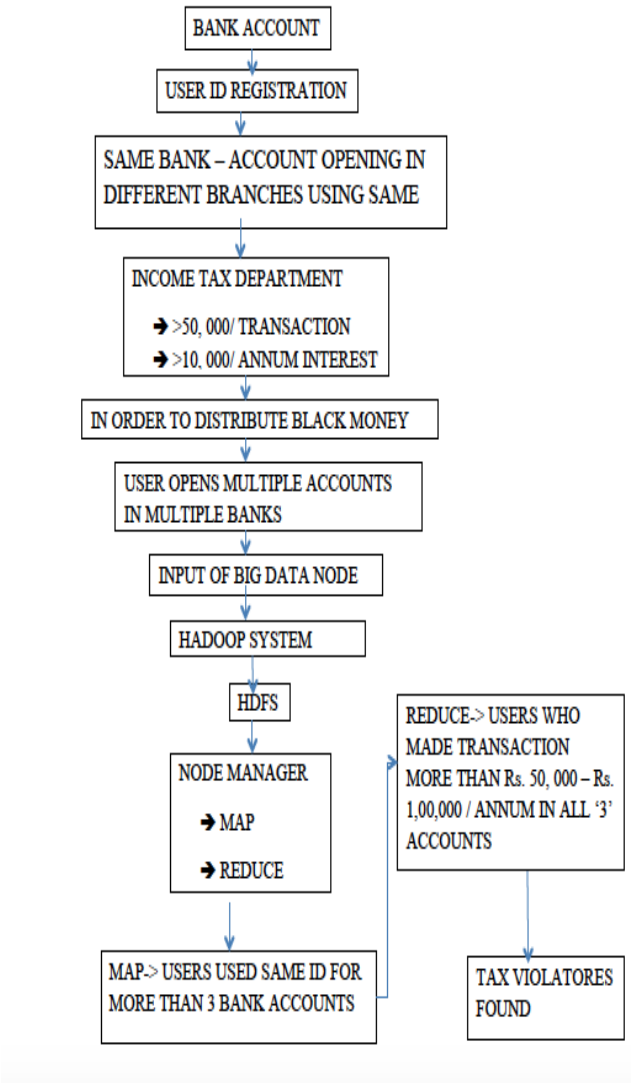

Fig -1: Flow Diagram

7. CONCLUSIONS

A new approximate answering approach which provides accurate results for large datasets is explained. The main aim of the project is to identify the tax violators in the banking sector. To track the transaction of users in multiple banks and monitor them using Aggregate Queries with Partitioning algorithm. The result is obtained by grouping the local values from all the different partitions and collective results are provided. The transaction details of different banks are taken and stored in the cloud. The datasets are partitioned into different tuples before uploading it into the cloud storage. The algorithm partitions the datasets according to its attributes, interests, etc. The improvement in this paper includes the analyses of the large datasets efficiently by reducing the cost to update, change queries, etc.

REFERENCES

[1] Xiaochun Yun, Guangjun Wu, Guangyan Zhang, Keqin Li, and Shupeng Wang,"A Fast Approach to Range-Aggregate Queries in Big Data Environments", IEEE,vol. 3, no. 2,pg no 206-218,June 2015

[2] Zhiqiang Zhang, Jianghua Hu,!Xiaoqin Xie,"An Online Approximate Aggregation Query Processing Method Based on Hadoop", IEEE, pg no 117-122,2016

[3] A. Munar, E. Chiner, I. Sales, "A Big Data Financial Information Management Architecture for Global Banking", IEEE, pg no 385-388,2014

[4] Yantao Gan, X. Meng, Y.Shi, "A Cloud-Based System for Online Aggregation", June 2013

[5] C.-T. Ho, R. Agrawal, N. Megiddo, and R. Srikant,, "Range queries in OLAP data cubes," ACM SIGMOD Rec., vol. 26, no. 2, pp. 73–88, 1997.

[6] Konstantin Andreev ∗,Harald Rac¨ ke ,"Balanced graph partitioning", 2004

[7] P. Mika and G. Tummarello, "Web semantics in the clouds," IEEE Intell. Syst., 23, no. 5, pp. 82–87, Sep./Oct. 2008.

[8] T. Preis, H. S. Moat, and E. H. Stanley, "Quantifying trading behavior in financial markets using Google trends," Sci. Rep.,vol. 3, p. 1684, 2013.

[9] H. Choi and H. Varian, "Predicting the present with Google trends," Econ. Rec., vol. 88, no. s1, pp. 2–9, 2012.

[10] W. Liang, H. Wang, and M. E. Orlowska, "Range queries in dynamic OLAP data cubes," Data Knowl. Eng., vol. 34, no. 1, pp. 21–38, Jul. 2000.

[11] G. Mishne, J. Dalton, Z. Li, A. Sharma, and J. Lin, "Fast data in the era of big data: Twitter's real-time related query suggestion architecture," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2013, pp. 1147–1158.

[12] J. M. Hellerstein, P. J. Haas, and H. J. Wang, "Online aggregation," ACM SIGMOD Rec., vol. 26, no. 2, 1997, pp. 171–182.

[13] P. J. Haas and J. M. Hellerstein, "Ripple joins for online aggregation," in ACMSIGMOD Rec., vol. 28, no. 2, pp. 287–298, 1999.

[14] E. Zeitler and T. Risch, "Massive scale-out of expensive continuous queries," Proc. VLDB Endowment, vol. 4, no. 11, pp. 1181–1188, 2011.

[15] N.Pansare, V.Borkar, C. Jermaine, and T. Condie , "Online aggregation for large Map Reduce jobs," Proc. VLDB Endowment, vol. 4, no. 11, pp. 1135–1145, 2011.