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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Efficient Analytical Architecture in Real-time Big Data for Remotely Sensing Application Using Hadoop Framework

S.J.Ohmpriya, Dr,Arokyraju, K.Ravikumar

P.G Student, Dept. of CSE, Rrase College of Engineering, Chennai, India[1]

Principal, Rrase College of Engineering, Chennai, India[2]

HOD, Dept. of CSE, Rrase College of Engineering, Chennai, India[3]

ABSTRACT*:* Big Data is the new experience curve in the new economy driven by data with high volume, velocity and variety. The real-time remote sensing Big Data seems at first, and extracting the useful information in an effective manner leads a system toward a massive computational challenges, such as to analyze, aggregate, and store, where data are remotely gathered. Keeping in view the over mentioned factors; there is a need for designing a system architecture that welcomes both real-time, as well as offline data processing. Therefore, in this paper, we discuss real-time Big Data analytical architecture for remote sensing satellite application. In Architecture contains three main units, such as 1) remote sensing Big Data acquisition unit (RSDU); 2) data processing unit (DPU); 3) data analysis decision unit (DADU). Firstly, RSDU gathers data from the satellite and transmit this data to the Base Station, where initial processing takes place. Later DPU plays an essential role in architecture for efficient processing of real-time Big Data by providing filtration, load balancing, and parallel processing. Then DADU is responsible for compilation, storage of the results, and generation of decision based on the results received from DPU. The architecture has the capability of dividing, load balancing, and parallel processing of only useful data. Thus, it results in efficiently analyzing real-time remote sensing Big Data using earth observatory system. Finally, a detailed analysis of remotely sensed earth observatory Big Data for land and sea area are provided using Hadoop.

KEYWORDS: Big Data, Filtration, Load Balancing, Hadoop

I. INTRODUCTION

BIG data analysis can discover trends of various social aspects and preferences of individual everyday behaviors. This provides a new opportunity to explore fundamental questions about the complex world. For example, to build an efficient investment strategy, they analyzed the massive behavioral data sets related to finance and yielded a profit of even 326 percent higher than that of a random investment strategy. Choi and Varian presented estimate sketches to forecast economic indicators, such as social unemployment, automobile sale, and even destinations for personal travelling. Currently, it is important to provide efficient methods and tools for big data analysis.

We give an application example of big data analysis. Distributed intrusion detection systems (DIDS) monitor and report anomaly activities or strange patterns on the network level. A DIDS detects anomalies via statistics information of summarizing traffic features from diverse sensors to improve false-alarm rates of detecting coordinated attacks. Such a scenario motivates a typical range-aggregatequeryproblem that summarizes aggregated features from all tuples within given queried ranges. Range-aggregatequeries are important tools in decision management, onlinesuggestion, trend estimation, and so on. It is a challengingproblem to quickly obtain range-aggregate queries results in big data environments. The big data involves a significant increase in data volumes, and the selected tuples maybe locate in different files or blocks. On the other hand, real-time systems aim to provide relevant results within seconds on massive data analysis.

DOI: 10.15680/IJIRCCE.2016. 0405224

II. RELATED WORK

Ahmed and Karypis 2012, Rezwan Ahmed, George Karypis, Algorithms for mining the evolution of conserved relational states in dynamic networks, Knowledge and Information Systems, December 2012 .They proposed the Dynamic networks that has been recognizedas a powerful abstraction to model and represent the temporalchanges and dynamic aspects of the data underlying manycomplex systems. Significant insights regarding the stable relationalpatterns among the entities can be gained by analyzingtemporal evolution of the complex entity relations. These can helpidentify the transitions from one conserved state to the next andmay provide evidence to the existence of external factors thatare responsible for changing the stable relational patterns inthese networks. But the Microarray gene-expression data and protein interaction data have their specific limitations, and as such, may not provide a measurement of all functional activities in a cell.

Alam et al. 2012, Md. HijbulAlam, JongWoo Ha, SangKeun Lee, Novel approaches to crawling important pages early, Knowledge and Information Systems, December 2012. They dictated that Web crawlers are essential to many Web applications, such as Web search engines, Web archives, and Web directories, which maintain Web pages in their local repositories. We studied that one of the most challenging issues for web search engines is finding high quality web pages or pages with high popularity for users and the problem of crawl scheduling that biases crawl ordering toward important pages.

A. Data Acquisition

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research question hypotheses, and evaluate outcomes. The data set contains information about users that are searching for the data and also marsh information. The page view static file contains data such as log of actions generated that were sent via up to filter the tosses request from an internal host. The filter is generating information like project name, page size of the request, and title of the page requested.

B. Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. There are a number of different tools and methods used for preprocessing, including:sampling, which selects a representative subset from a large population of datatransformation, which manipulates raw data to produce a single inputde-noising, which removes noise from data normalization, which organizes data for more efficient access andfeatureextraction, which pulls out specified data that is significant in some particular context.

C. Balanced Partitioning

A partitioning Framework for FARQ in which partitioning is help for balancing data chunks into respective partitions. These partitions hold data for increasing processing speed. According to large data record field partitioning algorithm is separating and analyzing that particular record. Also, it is assigned a record from large data tables to small data tables.The sampling of data is necessary for the analysis on the big data as the data is present in huge amount.Partitioning is the unit of big data utilizes load balancing and local aggregate queries.

D. Fast Result on Aggregate Query

The aggregation query is nothing but the aggregate functions used in the query like SQL, oracle, MySql and Sybase. There is Online Aggregate (OLA) that is used for improving the interactive behavior of database.For effective operations on database, batch mode is performing a key role.The user enters a query according to requirements. The query is come into users query processing block where the Querying Data algorithm (FARQ) is invoked. This is help for mapping queried data with histogram and go to respective partition for fetching respective data.

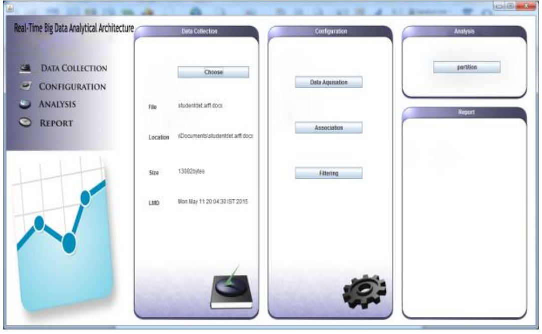
E. Performance Evaluation

After the project implementation of algorithms for processing of aggregate queries and applying an algorithm on them.The project has named node i.e. master node and data nodes (slave nodes).In this project first of all the partition of the data set is carried out which is uncontrolled as the Hadoop system does not control data. So by use of the balanced partition algorithm, data is controlled and chunks are created in the first stage of output. Then for mapping the data and utilizing the time histogram is generated. This is the second output of the project, which gives a benefit while matching contain with the user query .

III. RESULTS AND ANALYSIS

A detailed analysis of remotely sensed earth observatory Big Data for land and sea area are provided using Hadoop. Figure 1 shows the home page .

Fig 1 Main Window



File studentdet.arff.docx

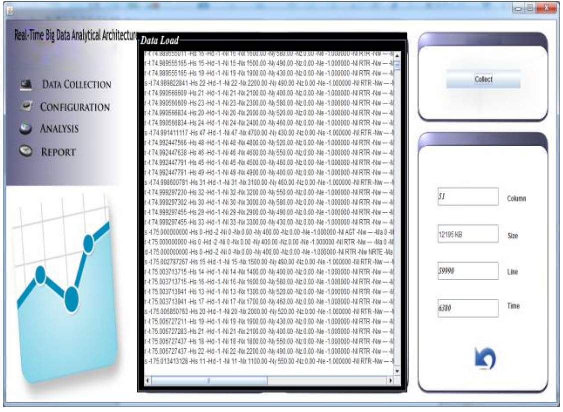
Location \Documents\studentdet.arff.docx

Size 13082bytes

LMD Mon May 11 20:04:30 IST 2015

A new data mining method that analyzes the time-persistent relations or states between the entities of the dynamic networks and captures all maximal non-redundant evolution paths of the stable relational states. Experimental results based on multiple datasets from real world applications show that the method is efficient and scalable.

Fig 2: Data Acquisition



51 Column

12195 KB Size

59990 Line

6380 Time

We propose a set of crawling algorithms for effective and efficient crawl ordering by prioritizing important pages with the well-known Page Rank as the importance metric. In order to score URLs, the proposed algorithms utilize various features, including partial link structure, inter-host links, page titles, and topic relevance. We conduct a large-scale experiment using publicly available data sets to examine the effect of each feature on crawl ordering and evaluate the performance of many algorithms. The experimental results verify the efficacy of our schemes. In particular, compared with the representative Rank Mass crawler, the FPR-title-host algorithm reduces computational overhead by a factor as great as three in running time while improving effectiveness by 5 % in cumulative Page Rank.

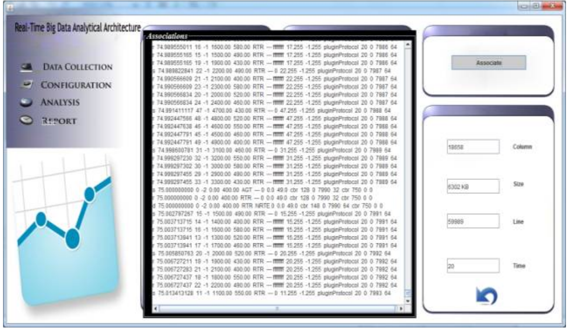
ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

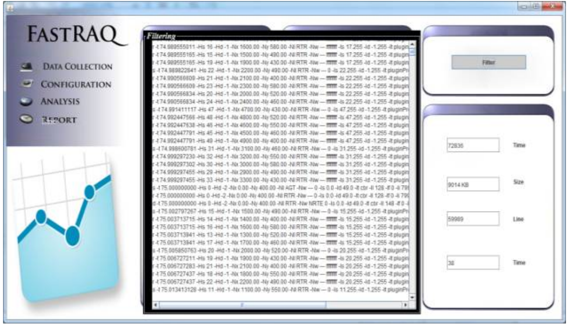
Fig 3: Data Association



```
74.989555011 16 -1 1600.00 580.00 RTR — ffffffff 17.255 -1.255 pluginProtocol 20 0 7986 64
r 74.989555165 15 -1 1500.00 490.00 RTR — ffffffff 17.255 -1.255 pluginProtocol 20 0 7986 64
r 74.989555165 19 -1 1900.00 430.00 RTR — ffffffff 17.255 -1.255 pluginProtocol 20 0 7986 64
s 74.989822841 22 -1 2200.00 490.00 RTR — 0 22.255 -1.255 pluginProtocol 20 0 7987 64
r 74.990566609 21 -1 2100.00 400.00 RTR — ffffffff 22.255 -1.255 pluginProtocol 20 0 7987 64
r 74.990566609 23 -1 2300.00 580.00 RTR — ffffffff 22.255 -1.255 pluginProtocol 20 0 7987 64
r 74.990566834 20 -1 2000.00 520.00 RTR — ffffffff 22.255 -1.255 pluginProtocol 20 0 7987 64
r 74.990566834 24 -1 2400.00 460.00 RTR — ffffffff 22.255 -1.255 pluginProtocol 20 0 7987 64
s 74.991411117 47 -1 4700.00 430.00 RTR — 0 47.255 -1.255 pluginProtocol 20 0 7988 64
r 74.992447566 48 -1 4800.00 520.00 RTR — ffffffff 47.255 -1.255 pluginProtocol 20 0 7988 64
r 74.992447638 46 -1 4600.00 550.00 RTR — ffffffff 47.255 -1.255 pluginProtocol 20 0 7988 64
r 74.992447791 45 -1 4500.00 460.00 RTR — ffffffff 47.255 -1.255 pluginProtocol 20 0 7988 64
r 74.992447791 49 -1 4900.00 400.00 RTR — ffffffff 47.255 -1.255 pluginProtocol 20 0 7988 64
s 74.998600781 31 -1 3100.00 460.00 RTR — 0 31.255 -1.255 pluginProtocol 20 0 7989 64
r 74.999297230 32 -1 3200.00 550.00 RTR — ffffffff 31.255 -1.255 pluginProtocol 20 0 7989 64
r 74.999297302 30 -1 3000.00 580.00 RTR — ffffffff 31.255 -1.255 pluginProtocol 20 0 7989 64
r 74.999297455 29 -1 2900.00 490.00 RTR — ffffffff 31.255 -1.255 pluginProtocol 20 0 7989 64
r 74.999297455 33 -1 3300.00 430.00 RTR — ffffffff 31.255 -1.255 pluginProtocol 20 0 7989 64
s 75.000000000 0 -2 0.00 400.00 AGT — 0 0.0 49.0 cbr 128 0 7990 32 cbr 750 0 0
r 75.000000000 0 -2 0.00 400.00 RTR — 0 0.0 49.0 cbr 128 0 7990 32 cbr 750 0 0
d 75.000000000 0 -2 0.00 400.00 RTR NRTE 0 0.0 49.0 cbr 148 0 7990 64 cbr 750 0 0
s 75.002797267 15 -1 1500.00 490.00 RTR — 0 15.255 -1.255 pluginProtocol 20 0 7991 64
r 75.003713715 14 -1 1400.00 400.00 RTR — ffffffff 15.255 -1.255 pluginProtocol 20 0 7991 64
r 75.003713715 16 -1 1600.00 580.00 RTR — ffffffff 15.255 -1.255 pluginProtocol 20 0 7991 64
r 75.003713941 13 -1 1300.00 520.00 RTR — ffffffff 15.255 -1.255 pluginProtocol 20 0 7991 64
r 75.003713941 17 -1 1700.00 460.00 RTR — ffffffff 15.255 -1.255 pluginProtocol 20 0 7991 64
s 75.005850763 20 -1 2000.00 520.00 RTR — 0 20.255 -1.255 pluginProtocol 20 0 7992 64
r 75.006727211 19 -1 1900.00 430.00 RTR — ffffffff 20.255 -1.255 pluginProtocol 20 0 7992 64
r 75.006727283 21 -1 2100.00 400.00 RTR — ffffffff 20.255 -1.255 pluginProtocol 20 0 7992 64
r 75.006727437 18 -1 1800.00 550.00 RTR — ffffffff 20.255 -1.255 pluginProtocol 20 0 7992 64
r 75.006727437 22 -1 2200.00 490.00 RTR — ffffffff 20.255 -1.255 pluginProtocol 20 0 7992 64
s 75.013413128 11 -1 1100.00 550.00 RTR — 0 11.255 -1.255 pluginProtocol 20 0 7993 64
```

18658 Column
6302 KB Size
59989 Line
20 Time

Fig 4: Data Filtering



```
r -t 74.989555011 -Hs 16 -Hd -1 -Nx 1600.00 -Ny 580.00 -Nl RTR -Nw — ffffffff -Is 17.255 -Id -1.255 -It plugin
r -t 74.989555165 -Hs 15 -Hd -1 -Nx 1500.00 -Ny 490.00 -Nl RTR -Nw — ffffffff -Is 17.255 -Id -1.255 -It plugin
r -t 74.989555165 -Hs 19 -Hd -1 -Nx 1900.00 -Ny 430.00 -Nl RTR -Nw — ffffffff -Is 17.255 -Id -1.255 -It plugin
s -t 74.989822841 -Hs 22 -Hd -1 -Nx 2200.00 -Ny 490.00 -Nl RTR -Nw — 0 -Is 22.255 -Id -1.255 -It pluginPr
r -t 74.990566609 -Hs 21 -Hd -1 -Nx 2100.00 -Ny 400.00 -Nl RTR -Nw — ffffffff -Is 22.255 -Id -1.255 -It plugin
r -t 74.990566609 -Hs 23 -Hd -1 -Nx 2300.00 -Ny 580.00 -Nl RTR -Nw — ffffffff -Is 22.255 -Id -1.255 -It plugin
r -t 74.990566834 -Hs 20 -Hd -1 -Nx 2000.00 -Ny 520.00 -Nl RTR -Nw — ffffffff -Is 22.255 -Id -1.255 -It plugin
r -t 74.990566834 -Hs 24 -Hd -1 -Nx 2400.00 -Ny 460.00 -Nl RTR -Nw — ffffffff -Is 22.255 -Id -1.255 -It plugin
s -t 74.991411117 -Hs 47 -Hd -1 -Nx 4700.00 -Ny 430.00 -Nl RTR -Nw — 0 -Is 47.255 -Id -1.255 -It pluginPr
r -t 74.992447566 -Hs 48 -Hd -1 -Nx 4800.00 -Ny 520.00 -Nl RTR -Nw — ffffffff -Is 47.255 -Id -1.255 -It plugin
r -t 74.992447638 -Hs 46 -Hd -1 -Nx 4600.00 -Ny 550.00 -Nl RTR -Nw — ffffffff -Is 47.255 -Id -1.255 -It plugin
r -t 74.992447791 -Hs 45 -Hd -1 -Nx 4500.00 -Ny 460.00 -Nl RTR -Nw — ffffffff -Is 47.255 -Id -1.255 -It plugin
r -t 74.992447791 -Hs 49 -Hd -1 -Nx 4900.00 -Ny 400.00 -Nl RTR -Nw — ffffffff -Is 47.255 -Id -1.255 -It plugin
s -t 74.998600781 -Hs 31 -Hd -1 -Nx 3100.00 -Ny 460.00 -Nl RTR -Nw — 0 -Is 31.255 -Id -1.255 -It pluginPr
r -t 74.999297230 -Hs 32 -Hd -1 -Nx 3200.00 -Ny 550.00 -Nl RTR -Nw — ffffffff -Is 31.255 -Id -1.255 -It plugin
r -t 74.999297302 -Hs 30 -Hd -1 -Nx 3000.00 -Ny 580.00 -Nl RTR -Nw — ffffffff -Is 31.255 -Id -1.255 -It plugin
r -t 74.999297455 -Hs 29 -Hd -1 -Nx 2900.00 -Ny 490.00 -Nl RTR -Nw — ffffffff -Is 31.255 -Id -1.255 -It plugin
r -t 74.999297455 -Hs 33 -Hd -1 -Nx 3300.00 -Ny 430.00 -Nl RTR -Nw — ffffffff -Is 31.255 -Id -1.255 -It plugin
s -t 75.000000000 -Hs 0 -Hd -2 -Nx 0.00 -Ny 400.00 -Nl AGT -Nw — 0 -Is 0.0 -Id 49.0 -It cbr -Il 128 -If 0 -Ii 79
r -t 75.000000000 -Hs 0 -Hd -2 -Nx 0.00 -Ny 400.00 -Nl RTR -Nw — 0 -Is 0.0 -Id 49.0 -It cbr -Il 128 -If 0 -Ii 79
d -t 75.000000000 -Hs 0 -Hd -2 -Nx 0.00 -Ny 400.00 -Nl RTR -Nw NRTE 0 -Is 0.0 -Id 49.0 -It cbr -Il 148 -If 0 -
s -t 75.002797267 -Hs 15 -Hd -1 -Nx 1500.00 -Ny 490.00 -Nl RTR -Nw — 0 -Is 15.255 -Id -1.255 -It pluginPr
r -t 75.003713715 -Hs 14 -Hd -1 -Nx 1400.00 -Ny 400.00 -Nl RTR -Nw — ffffffff -Is 15.255 -Id -1.255 -It plugin
r -t 75.003713715 -Hs 16 -Hd -1 -Nx 1600.00 -Ny 580.00 -Nl RTR -Nw — ffffffff -Is 15.255 -Id -1.255 -It plugin
r -t 75.003713941 -Hs 13 -Hd -1 -Nx 1300.00 -Ny 520.00 -Nl RTR -Nw — ffffffff -Is 15.255 -Id -1.255 -It plugin
r -t 75.003713941 -Hs 17 -Hd -1 -Nx 1700.00 -Ny 460.00 -Nl RTR -Nw — ffffffff -Is 15.255 -Id -1.255 -It plugin
s -t 75.005850763 -Hs 20 -Hd -1 -Nx 2000.00 -Ny 520.00 -Nl RTR -Nw — 0 -Is 20.255 -Id -1.255 -It pluginPr
r -t 75.006727211 -Hs 19 -Hd -1 -Nx 1900.00 -Ny 430.00 -Nl RTR -Nw — ffffffff -Is 20.255 -Id -1.255 -It plugin
r -t 75.006727283 -Hs 21 -Hd -1 -Nx 2100.00 -Ny 400.00 -Nl RTR -Nw — ffffffff -Is 20.255 -Id -1.255 -It plugin
r -t 75.006727437 -Hs 18 -Hd -1 -Nx 1800.00 -Ny 550.00 -Nl RTR -Nw — ffffffff -Is 20.255 -Id -1.255 -It plugin
r -t 75.006727437 -Hs 22 -Hd -1 -Nx 2200.00 -Ny 490.00 -Nl RTR -Nw — ffffffff -Is 20.255 -Id -1.255 -It plugin
s -t 75.013413128 -Hs 11 -Hd -1 -Nx 1100.00 -Ny 550.00 -Nl RTR -Nw — 0 -Is 11.255 -Id -1.255 -It pluginPr
```

72836 Time
9014 KB Size
59989 Line
38 Time

IV. CONCLUSION

In this paper ,we presents a new data mining method that analyzes the time-persistent relations or states between the entities of the dynamic networks and captures all maximal non-redundant evolution paths of the stable relational states. Experimental results based on multiple datasets from real world applications show that the method is efficient and scalable, we propose a set of crawling algorithms for effective and efficient crawl ordering by prioritizing important pages with the well-known and this proposed algorithms utilize various features, including partial link structure, inter-host links, page titles, and topic relevance. We also conduct a large scale randomized experiment involving millions of subjects on the popular social networking websitewhich is used to identify influential and susceptible in large social networks of consumers. wealso propose FastRAQ—a new approximateanswering approach that acquires accurate estimationsquickly for range-aggregate queries in big data environments.FastRAQ has Oð1Þ time complexity for data updatesand O(N/(P*B)) time complexity for ad-hoc range-aggregatequeries. If the ratio of edge-bucket cardinality (h0) is small enough, FastRAQ even has Oð1Þ time complexity for rangeaggregatequeries.We believe that FastRAQ provides a good starting pointfor developing real-time answering methods for big dataanalysis.

DOI: 10.15680/IJIRCCE.2016. 0405224

V. FUTURE ENHANCEMENT

There are also some interesting directions for ourfuture work. First, FastRAQ can solve the 1:n format range-aggregatequeries problem, i.e., there is one aggregation columnand n index columns in a record. We plan to investigatehow our solution can be extended to the case of m:n formatproblem, i.e., there are m aggregation columns and n indexcolumns in a same record. Second, FastRAQ is now runningin homogeneous environments. We will further explore howFastRAQ can be applied in heterogeneous context or even asa tool to boost the performance of data analysis in DBaas.

REFERENCES

[1] P. Mika and G. Tummarello, “Web semantics in the clouds,” IEEEIntell. Syst., vol. 23, no. 5, pp. 82–87, Sep./Oct. 2008.

[2] T. Preis, H. S. Moat, and E. H. Stanley, “Quantifying tradingbehavior in financial markets using Google trends,” Sci. Rep.,vol. 3, p. 1684, 2013.

[3] H. Choi and H. Varian, “Predicting the present with Googletrends,” Econ.Rec., vol. 88, no. s1, pp. 2–9, 2012.

[4] C.-T. Ho, R. Agrawal, N. Megiddo, and R. Srikant,, “Rangequeries in OLAP data cubes,” ACM SIGMOD Rec., vol. 26, no. 2,pp. 73–88, 1997.

[5] G. Mishne, J. Dalton, Z. Li, A. Sharma, and J. Lin, “Fast data in theera of big data: Twitter’s real-time related query suggestionarchitecture,” in Proc. ACM SIGMOD Int. Conf. Manage. Data,2013, pp. 1147–1158.

[6] W. Liang, H. Wang, and M. E. Orlowska, “Range queries indynamic OLAP data cubes,” Data Knowl. Eng., vol. 34, no. 1,pp. 21–38, Jul. 2000.