International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Efficient Query Handling On Big Data in Network Using Pattern Matching Algorithm: A Review

Prasadkumar Kale[1], Arti Mohanpurkar[2]

[1]Department of Computer Engineering, Pune University, Maharashtra
Dr. D Y School of Engineering and Technology, Lohegaon, Pune

[2]Assistant Professor Department of Computer Engineering, Pune University, Maharashtra
Dr. D Y School of Engineering and Technology, Lohegaon, Pune

Abstract: ***Manipulation of Queries on a large amount of data is done with various technologies. Query handling of Big data is a challenging task because of unstructured, real-time data. Efficient Query execution technique usage partition algorithm & pattern match algorithm for handling queries of the user. These techniques provide faster execution of such queries on the Big data in an efficient manner. Various authors propose query execution using DAG, ranges, unification & other methods. Range-aggregate queries are giving better performance with balance partition algorithm and map-reduce technique for handling large amount of queries with pattern matching technique.***

Keywords: Big data, Range-Aggregate query, Map-Reduce, Partitioning, Pattern matching

1. Introduction

Today Big data is the most demanded topic. The world is moving faster and the phrase becomes true 'World becomes a Village'. Every individual human wants to access network for staying connected with the world. These users may access a lot of data related to Geographical areas, political issues, neural network, health information and many more. There is another thing related to Big data is social sites and media. Social sites like Google for Gmail and most preferably for the search engine, Facebook, whatsApp are hit every day by billions of people around the world. These sites improve knowledge of human social networking, mathematicians, physicians and many more science fields by exchange of information in very small amount of time [1]. All these people search valuable information in just one click.

Big data processing is the main task. In this processing some frameworks are Hive, pig, Jaql like technologies play an important role described in [3] [4] [5] [6]. On the 6th Oct. 2014 Flip-kart announces an offer which is very cheap. Resulting in high sever processing is a very low small amount of time. According to Flip-kart there are billions of request hit within 30 min. For processing large amount of data and analyze that data various technologies are in use as mentioned above. The more fundamental challenge for Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. In many situations, the knowing extraction process which has to be very efficient and close to real time as storing all observed data is nearly unachievable. The unique data quantities require an effective data analysis and prediction platform to achieve fast response and real-time classification of such Big Data.

The main focus is on how data are analyzed, retrieved according to accuracy and an efficient manner. [2] Provided HACE theorem for categorizing the data into respective characteristic and discussed the data mining challenges. Now-a day's Map-Reduce frame wok is used for processing on OLAP and OLTP systems, which are updated periodically. Map-reduce technique [18] has one biggest characteristic, i.e. parallel execution. For the processing large amount of data HADOOP [19] [20] uses parallel processing techniques in which Map-Reduce technique is mostly used. This technique is easy to understand from the rest of the others. Cluster and Partition algorithms are used for processing on the big data. These things are effectively giving outputs, but not in satisfaction and their understanding level becomes more complex than others. Query mapping becomes more complicated with scientific databases. Mapping of queries of Big data web sources [17], presents a declarative meta - language for understanding the meaning of queries and map them into respective resources.

Most of query optimization algorithms [7] [8] are used graphs to analyze and operate effectively. The pattern matching algorithm is part of graph analysis. Distributed and live data can handle with this algorithm. The main importance of pattern matching algorithm is finding the patterns that are related to the outgoing or incoming data. Most time the DAG are used for query optimization. DAG is directed acyclic graph which does not have any cycle means better way a tree, so finding data will not end in Deadlock fashion. The pattern matching algorithm is mostly known to detect the attacks and prevent the attack, but here we are using it for finding the related queries.

2. Related Work

Feng Li [9] proposed a Map-Reduce Framework for supporting real-time OLAP system. The open source distributed key/value system; they called it as HBase and Stramed Map-Reduce as HStreaming for incremental updating. They proposed an R-store for Map-Reduce support on Real-time OLAP. They evaluate their performance results on the basis of TPC-H data.

Jiewen Huang [10] and colleagues introduce query optimization techniques based on distributed graph pattern

matching. They proposed two Frameworks that are from the linear and bushy plan is considered in System-R style dynamic programming algorithm and cycle detection algorithm for reduce intermediate result size. The computations reuse technique for eliminating redundant subqueries and traffic reduction.

Characterization of point pattern matching is done by the local descriptor called Line Graph spectral context. This work is done by Jun Tang [11] and his associates by doing an analysis of spectral methods and aiming to introduce a robust for positional jitter and outlier. Multiview spectral embedded technique is used for finding the similarities between descriptor by comparing their low dimensional embedding.

Kosaku Kimura [12] and fellows aimed to reduce the cost of data transmission between components that are processing nodes and interconnection service. Multi-query unification technique generates unified components for DFD. Unification methods are used nesting, clause assembly for collecting the queries and assemble into a single query for reduction of execution time. Results are calculated on the virtual DFD by applying two-stage unification on DSP using Esper and CDP using Hive. Better performance is of DSP using Esper.

For Big data analytics, i.e. high-level dataflow system an extensible and language independent framework m2r2 is described in Vasiliki Kalavri [13]. This prototype implementation is done on the Pig dataflow system and results handled automatically in catching, common subquery matching not only rewriting but also garbage collection. Evaluation is done using the TPC-H benchmark for pig and report reduction in query execution time by 65% on average.

Xiaochun Yun [14] proposed FastRAQ- big data query execution in a range-aggregate queries approach. A balanced partition algorithm is used first to divide big data into independent partitions, then local estimation sketch generated for each partition. FastRAQ gave result by summarizing local estimation from all partitions. The Linux platform is helpful for implementing FastRAQ and performance evaluated on billions of data records. According to the authors, FastRAQ can give good starting points for real-time big data. It solves the 1: n format range-aggregate query problem, but m:n formatted problem still outside there.

High performance computing (HPC) experienced explosive growth of data in recent days. Saba Sehrish [16] introducing MRAP (MapReduce with access patterns) techniques for demonstration of results with good percentage of throughput. MapReduce tool can be used for data analysis and reorganizing the HPC storage semantic and data-intensive systems. Running multiple MapReduce phase cause more overhead so authors provide data-centric scheduler to improve performance of MapReduce on HPC.

3. Big Data and Pattern Matching Algorithm

3.1. Big Data Characteristics

In traditional approach data is stored in tuples in the form of columns and rows. Big data characteristics are as follows:

- Volume - Data generated in large scale by machine and human interaction than a traditional data. For instance, Data generated in call centers, which is in terms of call recording, tagging of queries, request, complaints etc.
- Velocity – Social media data streams produce a large influx of opinions and relationships valuable to customer relationship management. That is like messages, photos on twitter or Facebook etc.
- Variety – Traditional databases use structured data, i.e. data schema and change slowly. In opposite of that nontraditional databases format exhibits dizzying rate of change.
- Complexity – Data management in big data is very complex task, when a large amount of data which is unstructured coming from various sources. This has to be linked, connected and correlated to grasp the information.

Big data also contain heterogeneous information, autonomous source and complex and evolving relationships. Heterogeneous mean data that is not in the same format because each enterprise, institutional and vendor has a different format as the copyright and other issues. Autonomous sources may generate the data as per the events are occurring in the system, for example, counting the job and completing various tasks in industries. The task may contain analysis of programs, testing of the programs. Humans are coming together because of their similarities with each other. These similarities may contain hobbies, biological relationships and mutual understanding of each other.

3.2. Big Data Challenges

Data access and computation on the related data for getting the related information in time. In such situation algorithms has to be very fast in terms of time complexity and other performance measures. In industry, there are a lot of data that need to be processed instantly so hardware increment required. Another method is putting data in-memory, but using a grid computing approach, where many machines are used to solve a problem. Both approaches allow organizations to explore huge data volumes and gain. Understanding the data takes a lot of time for getting the shape so that visualization may apply for it.

The value of data becomes jeopardized if data find and analyzed is unable to present at the right time when the customer want particular data. To deal with this problem companies need to have a data governance and data management in place to ensure data is clean. Plotting points on graph for analysis becomes difficult when dealing with extremely large amounts of data. One way to resolve this problem is cluster data into higher-level view where smaller groups become visible. In Data mining also have many challenges for the big data like Platform for computations, Data semantics and application knowledge with sharing, privacy and domain of data.

3.3. Big Data Analysis Technologies

Data analysis requires a lot of computing and complex time for result and understanding. Big data contain many techniques for analysis consists of a lot of computation which

is done using big data algorithms. There are some algorithms like cluster, map-reduce, data mining algorithms such as k-means, classification methods, support vector machine, apriori algorithm, EM, page rank, etc.. All these algorithms are effectively working according to their need.

For large growth of data, the data analytics uses advanced analytic techniques like predictive analytics, data mining, statistical analysis, complex SQL, data virtualization, and artificial intelligence. ADV (advanced data virtualization) is the best fit for the growing big data analytics. BI supports real time dashboard and key performance indicators (KPI) and sometimes OLAP cube for which in-memory databases will move. Text mining and text analytics give the unstructured data more efficiently. HDFS and Map-reduce is closely related by distributing parallel processing and combining the output. HADOOP uses a map-reduce technique for the analysis of data. Benefits of using map-reduce framework are 1) It will run a small amount of processes while data, analyzing, 2) simultaneously organize the migrated files.

3.4. Pattern Matching Algorithm

In analyzing the data patterns plays an important role in that each incoming data is analyzed. For a set of patterns for a set of objects in order to determine all possible matches method used is Rete Match Algorithm [15]. It maintains state information of objects which are matched and partially match until the object is present in the memory. There is another pattern matching algorithm also like exact pattern matching which usage searching of related patterns in giving text. Knuth-Morris-Pratt is another algorithm which is also on searching for patterns using Java techniques. RE pattern matching and grep algorithms are on regular expressions and they give more than one result for related pattern.

The Brute force exact pattern matching algorithm uses search techniques for finding the exact data. Applications of this algorithm are for web search engines, parsers, digital libraries, screen scraper. Other algorithms use DFA, grammar and regular expression for evaluation of patterns into the linear-time guarantee, no backup stream. The RE pattern matching algorithm gives multiple occurrences of patterns in text files.

4. System Architecture

The user will fire query according to his requirement of information on the database. This query is partitioned in various sub forms according to the words used in the query. This partition is useful for matching patterns of the indexes present on the database as quickly as possible. If the requested data is present in local database, then the query is executed and accordingly result is generated. This result will send to the user according to requested data. In case of data is not present on local server then it will send that query to main server or we can say that a Global one. This server also has an indexing which is generated by applying the partition algorithm on real-time data and scattering data according to the partitions.

These partitions hold an index mechanism which is based on the pattern. When a user query will enter its first analyzed and according to pattern it will retrieve related information as quickly as possible from related partition.

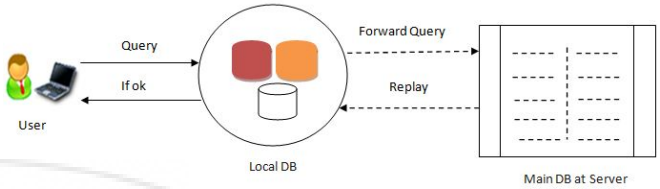

Figure 1: System Architecture

The data coming to database may be of any format so this data has to be formatted into a particular form and stored accordingly in the partition. The storage will done by using tries which are also used for storing the indexes.

5. Conclusion

In this paper, pattern matching technique is used for the retrieval of data. Partition algorithm is playing an important role for scattering of data according to data arriving on the sever. These partitions also contain an indexing system which is useful for analyzing the data. Query arrives at sever is partitioned into word. Pattern matching algorithms are used to process the relevant queries as quickly as possible. These idea willing to cover the data cube analysis and m:n problem of FastRAQ technique. The map-reduce framework with pattern matching technique gives better access than any other system for query analysis.

References

[1] Wei Tan, M. Brian Blake & Iman Saleh, Schahram Dustdar, "Social-Network-Sourced Big Data Analytics", IEEE Internet Computing, September/October 2013.
[2] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding "Data Mining with Big Data", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, January 2014
[3] F. Gates, O. Natkovich, S. Chopra, P. Kamath, S. M. Narayana-murthy, C. Olston, B. Reed, S. Srinivasan, and U. Srivastava, "Building a high-level dataflow system on top of map-reduce: the pig experience," Proc. VLDB Endow., vol. 2, no. 2, pp. 1414–1425, Aug. 2009.
[4] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. "Pig latin: a not-so-foreign language for data processing. In SIGMOD" , pages 1099–1110, 2008.
[5] Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu,P. Wyckoff, and R. Murthy, "Hive: a warehousing solution over a map-reduce framework," Proc. VLDB Endow., vol. 2, no. 2, pp. 1626–1629,Aug. 2009.
[6] K. S. Beyer, V. Ercegovac, R. Gemulla, A. Balmin, M. Y. Eltabakh, C.C. Kanne, F. Ozcan, and E. J. Shekita, "Jaql: A scripting language for large scale semistructured data analysis." PVLDB, vol. 4, no. 12, pp. 1272–1283, 2011.
[7] W. Hong and M. Stonebraker. "Optimization of parallel query execution plans in xprs", PDIS '91

[8] R. S. G. Lanzelotte, P. Valduriez, and M. Zait. "On the effectiveness of optimization search strategies for parallel execution spaces", In VLDB, pages 493–504, 1993.
[9] Feng Li, M. Tamer Ozsu, Gang Chen and Beng Chin Ooi," R-Store: A Scalable Distributed System for Supporting Real-time Analytics", ", IEEE ICDE Conference 2014.
[10] Jiwen Huang, Kartik Venkatraman, Daniel J. Abadi,"Query Optimization of Distributed Pattern Matching", IEEE ICDE Conference, 2014.
[11] Jun Tang, Ling Shao, Simon Jones, "Point Pattern Matching Based on Line Graph Spectral Context and Descriptor Embedding".
[12] Kosaku Kimura, Yoshihide Nomura, Hidetoshi Kurihara, Koji Yamamoto and Rieko Yamamoto,"Multi-Query Unification for Generating Efficient Big Data Processing Components from a DFD", IEEE Sixth International Conference on Cloud Computing,2013.
[13] Vasiliki Kalavri, Hui Shang, Vladimir Vlassov," m2r2: A Framework for Results Materialization and Reuse in High-Level Dataflow Systems for Big Data", IEEE 16th conference on ICCSE,2013.
[14] Xiaochun Yun, Guangjun Wu, Guangyan Zhang, Keqin Li, and Shupeng Wang, " FastRAQ: A Fast Approach to Range-Aggregate Queries in Big Data Environments", IEEE Transactions On Cloud Computing, Vol. 6, No. 1, January 2014.
[15] Charles L. Forgy ,"Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem", Artificial Intelligence ,1982.
[16] Saba Sehrish, Grant Mackey, Pengju Shang, Jun Wang,"Supporting HPC Analytics Applications with Access Patterns Using Data Restructuring and Data-Centric Scheduling Techniques in MapReduce", IEEE Transactions on Parallel and Distributed Systems, Vol. 24, No. 1, January 2013.
[17] Hasan M. Jamil, "Mapping Abstract Queries to Big Data Web Resources for On-the-fly Data Integration and Information Retrieval", IEEE ICDE Workshops 2014.
[18] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," Commun. ACM, vol. 51, no. 1, pp. 107–113, Jan. 2008.
[19] Bina Kotiyal, Ankit Kumar, Bhaskar Pant, RH Goudar, "Big Data: Mining of Log File through Hadoop".
[20] Y. Bu, B. Howe, M. Balazinska, and M. D. Ernst, "Haloop: eficient iterative data processing on large clusters," Proc. VLDB Endow., vol. 3, no. 1-2, pp. 285–296, Sep. 2010.

Author Profile

Prasadkumar Kale Research Scholar Dr. D.Y.Patil School of Engineering & Technology, Pune, Savitribai Phule Pune University. He received B.E. in Computer Engineering from Computer Engineering, Department of ICOER, Wagholi, Pune from Pune University. Currently He is perusing M.E. in Computer Network from Dr. D. Y. Patil School of Engineering & Technology, Pune, University of Pune.

Prof. Arti Mohanpurkar received the B.E. and M.E degrees in Computer Science Engineering, and pursuing Ph.D. Now she is HOD of Computer Engineering Department, Dr. D. Y. Patil School of Engineering & Technology, Savitribai Phule Pune University, India