Business intelligence and decision support using distinct mapreduce with access patterns (DMRAP) in big data analytics

Article *in* Journal of Advanced Research in Dynamical and Control Systems · June 2017

CITATIONS
0

READS
269

4 authors, including:

Javed Parvez Shaik
Vels University
9 PUBLICATIONS **130** CITATIONS
SEE PROFILE

Senthil Kumar Janahan
Lovely Professional University
11 PUBLICATIONS **110** CITATIONS
SEE PROFILE

Arun Sahayadhas
Vels Institute of Science Technology and Advanced Studies
34 PUBLICATIONS **1,257** CITATIONS
SEE PROFILE

Business Intelligence and Decision Support Using Distinct MapReduce with Access Patterns (DMRAP) in Big Data Analytics

Shaik Javed Parvez, Assistant Professor, School of Engineering, Vels University, Pallavaram, Chennai, India.

Senthil Kumar Janahan, Assistant Professor, School of Engineering, Vels University, Pallavaram, Chennai, India.

Dr.S. Arun, Associate Professor, School of Engineering, Vels University, Pallavaram, Chennai, India.

Dr.R. Anandan, Assistant Professor, School of Engineering, Vels University, Pallavaram, Chennai, India.

Abstract--- In modern business environment, data is playing a vital role in taking quick decisions. The growth of data is uncertain and huge over period of time and also handling such huge amount of unstructured and structured data is difficult for business analytics, for quick decision making we need efficient tools that are capable of handling and processing Bigdata for analytical purposes. Traditional available systems such as data mining techniques, association rules and other mining methodologies are insufficient to store, handle and process the spontaneous amount of data generated day by day. Even though Hadoop, MapReduce and MapReduce Access Patterns (MRAP) involved in increasing the productivity and fault tolerance, improving the performance is complicated task for handling the vast amount of data. So, we propose a new methodology Distinct MapReduce with Access Patterns (DMRAP) in association with Hadoop MapReduce and business intelligence tools that supports in taking quick decisions and are capable of handling and processing Big data. The methodology used here is unique and more tangible, scalable, reduces cost and time for wide variety of data analytics, real time operations in order to predict and forecast the future business needs.

Keywords--- Big Data, DSS, Hadoop MapReduce, Business Intelligence.

I. Introduction

Business Intelligence is an umbrella term up to expectation composed concerning many tools, technologies and applications for collecting, analyzing and giving access to data to help the users make efficient and effective intelligent business decisions. It is a system that gathers data and combines, analyses to present information concerning business for supporting intelligent decision making for business in a better way.

Information is developing at a gigantic speed making it hard to deal with such huge measure of information (exabytes). The main trouble in taking care of such extensive measure of information is on account of that the volume is expanding quickly in contrast with the computing resources[1].There is a tremendous issue to store and deal with the vast volume of data, in addition there is also issue to break down and extricate meaningful information from it. There are also a few methods to deal with collecting, storing, processing, and analyzing big data. Managing such vast amount of unstructured and structured data for business analytics and quick decision making we need efficient tools that are capable of handling and processing Big data for analytical purposes.

Big Data

2.5 quintillion bytes of information are created day to day. Information sources-Sensors used to gather atmospheric data, post to networking media sites, computerized imaging and recordings, buy transaction records and GPS signs of mobile phone to give some examples. We have come into the period of "Big Data." Just as the individuals producing this information continuously, it can be still broken down progressively by high performance computing networks, along these lines making it possible for enhanced basic decision making. The International Data Corporation (IDC) trusts associations that are the best ready to settle on progressing business choices utilizing Big Data [2]. Huge Data indicate data sets whose volume and/or Variety are exceeding the capacity of frequently utilized computing devices to capture, manage and process the information in an average elapsed by time which is important for business. The difficulty can be identified with data collection, search, storage, analytics, visualization, sharing and so forth. It implies information which is too enormous, too quick, or too hard for existing devices to process. Here, "Too enormous" indicates that associations progressively should be able to manage petabyte scale collection of information that is derived from click streams, sensors, transaction histories or elsewhere. "Too quick" indicates that is information huge, as well as it should be handled quickly for instance, to perform fraud detection at

a point of sale or determine what ad to be shown to a user on a webpage. "Too hard" indicates information that will not fit into a present processing tool or that needs some kind of study that existing tools can't provide. Enormous Data is reported by the following 4 Vs:

1. Volume – Is the huge amount of information generated each and every second that are bigger than what the routine relational database frameworks can adjust to.
2. Velocity –Isthe recurrence at which new information is generated, collected and shared.
3. Variety –Isthe rapidly increasing different kind of information (from online networking feeds to budgetary information, from sensor information to photographs, from voice recordings to video capture) that no longer positions into perfect, easy to consume structures.
4. Veracity –Is what which refers to the numerous sources and kinds of information both structured and unstructured. Nowadays information comes as photographs, messages, PDFs, monitoring devices, audio and so on. This variety of unstructured information makes it difficult for mining, capacity and analyzing the information.

In the recent 35 years, data management standards, for example, declarative querying, cost-based optimization and logical independencehave driven to a multi-billion dollar industry. More critically, these specialized progress have empowered the first round of business intelligence applications and established the framework for analysing and managing Big Data today[2]. More than 500 million tweets are posted by Twitter every day [8].Over 766 million active users per day in 2014 is reported by Weibo (Weibo, 2015). The greater use of social media such as Weibo, YouTube and Twitter has put up nearly 90% of the total data available today[10].Many challenges that businesses face today is expected to be handled by Big data analytics [9].

MapReduce

The Hadoop fundamentally has two principle segments that are MapReduce programming model and Hadoop Distributed File System (HDFS) [3].

Hadoop is a distributed system written in Java which consolidates highlights like those of the MapReduce programming paradigm and Google file System [4].MapReduce is a critical innovation which was proposed by Google. MapReduce is a rearranged programming model and is a main part of Hadoop for parallel processing of limitless amount of information. It calms software engineers from the weight of parallelization issues while permitting them to openly focus on application advancement[4]. HDFS and MapReduce are adaptable and fault tolerant model that conceal all difficulties for Big Data analytics. MapReduce engine which uses TaskTracker and JobTracker that handle checking and execution of jobs. Hadoop is the most notable and open source execution of MapReduce programming model [5].

MapReduce is a program and a technique for distributed computing in light of java. The MapReduce includes two essential tasks to be specific Map and Reduce as shown in the fig(1). In Map phase, a set of data is taken and converted to anotherst of data ,where the tuples(key/value pairs) are considered as individual elements which are broken down. In reduce phase, the output from a mappase is taken as an input which in turn combines those data tuples into a more smaller set of tuples. As the sequence of the name MapReduce infers, the reduce task is always performed after the map job.

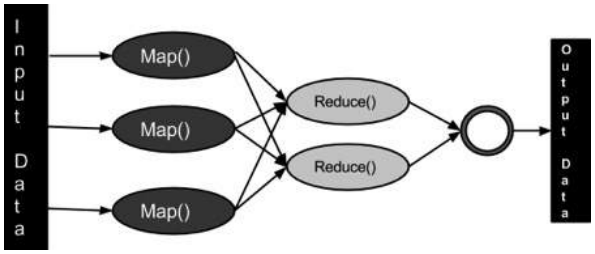

Figure 1: MapReduce

In a MapReduce program, the Map stage reads information in contiguous chunk, and produces information as (key, value) sets for the reduce stage. The reduce stage then joins all of the values with the same keys to yield the result [6].

Business Intelligence

The way toward collecting, arranging and analyzing business data and transforming it into useful and significant data is mostly referred to as Business Analytics or Business Intelligence. With business intelligence, organizations

have more better understanding into their yielding new opportunities, association, competitive advantages, rectifications to existing systems or procedures and more.

Business knowledge (BI) is the capacity of an organization to make meaningful utilization of information it gathers over the span of time. Its everyday business operations [11].

The BI could play a critical part in enhancing organizational performance by recognizing new opportunities, highlighting potential dangers, uncovering new business experiences and upgrading decision making forms among numerous different advantages [13][19].

The advance of computing and web innovations have encouraged gathering of huge volume of diverse data from various sources on a continuous basis posturing new difficulties and opportunities for business intelligence. This data include both structures and unstructured, unpredictable and basic data. For instance, Wal-Mart can deal with more than 1 million transactions for each hour [12].

Business insight incorporates a scope of areas, for example, customer intelligence, competitor intelligence, product intelligence, strategic intelligence, market intelligence, business counter intelligence and technological intelligence [11].

Big data analytic can help organizations to better process big data for enhancing consumer loyalty, managing risk of supply chain, creating competitive intelligence, giving business ongoing insights to help settle on important choices and optimizing pricing if correctly used [14][15][16][17].

As per an examination, a retailer that can utilize big data appropriately has the potential capacity to expand 60% of working edges by acquiring market share over its opponents and exploiting the detailed buyer information [18].

II. Methodology

To deal with Big data, Range of techniques and methods have been developed and technologies to determine what is needed and what is fit in processing Big data and satisfying the requirements. We need considerable processing power for complex data, reduced data pre-processing techniques, scalable, distributed and fault tolerant data processing capabilities. The following part describes the need for DMRAP and how it differs from other MapReduce access patterns in handling data.

MRAP Types

MapReduce Access Patterns (MRAP) is an ideal combination of data access semantic and the programming framework in which it has been demonstrated that throughput improvement and input-output data performances over large datasetsare immensely complex. MapReduce program fig 2.1(a) describes filtering original datasets, JobTracker enables TaskTracker to accept jobs and instructs the NameNode to locate data in HDFS which is to be processed.There are many MapReduce Patterns in fig 2.1(b) which are available as such Flat data structures, Foreign/Recursive key Joins, Hash Joins, Distribute Computation, Chain MapReduce phases, Simplify Reduce, for understanding the data processing and analysis.

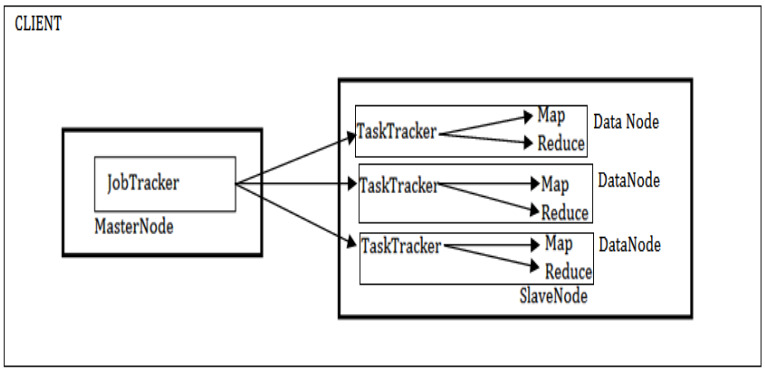

Figure 2.1(a): Filtering Original Datasets

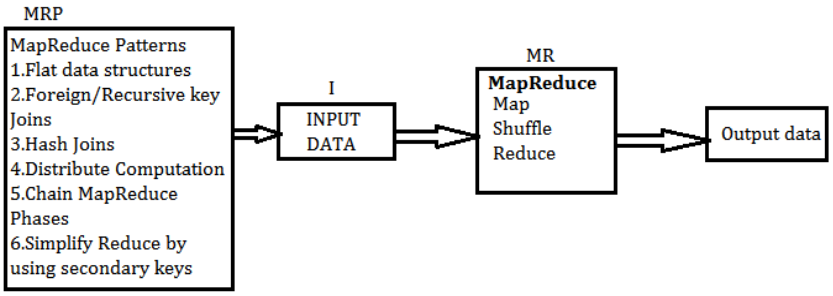

Figure 2.1(b): MapReduce Patterns

MapReduce program consists of two stages for any patterns, Stage one is about to Map (M) and stage two is about to Reduce (R). For two different datasets we can consider patterns one for smaller contiguous regions and another one for smaller non-contiguous regions. To, handle these two patterns MRAP has two components such as MRAP API and MRAP data restructuring. To reduce pre-processing techniques and optimising the data layouts MRAP uses pre-written programs for every access patterns. Evaluating the MRAP framework with different Data Access Patterns needs traditional MapReduce technologies. Increase in nodes results in extension of pre-processing techniques and reduces the fault tolerance. To overcome these drawbacks there is an essential need for improvised Access Patterns where it minimises the number of MapReduce phase.

DMRAP

Introducing Distinct MapReduce Access Patterns (DMRAP) which differs from old MapReduce Access Patterns techniques and uses a capacity scheduler that are capable of managing and assigning resources using a Resource Manager (RM), Application Master (AM) for each application interacts with resource manager and a Node Manager (NM) for managing user processes for node. The DMRAP architecture is described in the Fig (2.2).

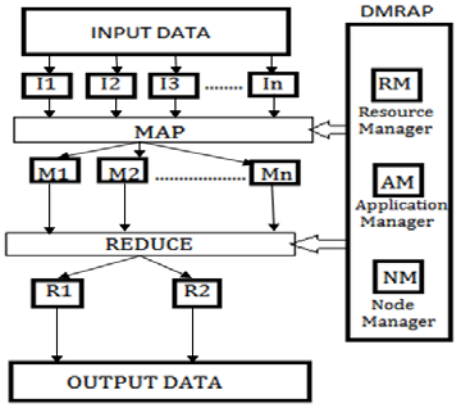

Figure 2.2: DMRAP Architecture

The purpose of DMRAP is to combine Multiple Data Access Patterns and uses MapReduce program are of at least one in which a Map (M) and a Reduce (R) with inputs and outputs defined in form of Key and Value pairs. This approach guarantees each Map task with merge set of operations to reduce the multiple occurrence of a datasets.

In DMRAP we provide set of three Managers to access patterns continue with the Resource Manager (RM) analysis the dataset in one phases i.e. a Map phase. In the first part, there is a combination inputs made from different patterns matching similar inputs from I1 to In and the count of Key and Value pairs are identified. In the second phase, Application Manager (AM) analyses the Key and Value pairs and constructs the Map with specified node.

Now the Node Manager(NM) analyses the node and configures the Reduce (R) in a new format where MapReduce and MRAP logically operates the applications such as standard HDFS and restructuring data formulated from different smaller items. The data appropriate to a specified business needs could be easily generated and the

result of DMRAP is increased when compared to the traditional MapReduce and MRAP implementations requires only one phase of input-output. The total number of bytes accessed by MRAP and MapReduce implementations i.e. read or written on an average of 50% more when compared to DMRAP implementations. The overall performance is improved by 28% in Business analytics methodologies.

III. Results and Discussions

To make meaningful use of data in the course of day to day business operations, Solutions to be made on structured and internal data that potentially leads to decision making and creates new opportunities for Business Intelligence. The exploratory nature of study was taken in this paper by analysing the supermarket products, items that influences frequent purchasing behaviour. Identifying the customer needs and creation of new marketing opportunities and sales opportunities covering Big data analytics on Business Intelligence.

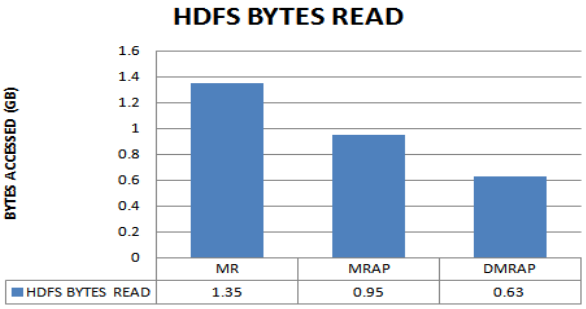

Figure 3.1: DMRAP Implementation of Read Mapping

In a supermarket we analysed large set of purchase bills of the customers, we identified some of the items were associated with other items such as bread, milk, bun, jam, butter, tea, sugar, cheese etc together considered as input data to the DMRAP system and performed operations among 30GB to 60GB data in each slot input and reduced 15 Map tasks of 4GB regions and depicts 1GB Map task as a worst case scenario. The mapped items for raw data leads to 0 results read map. But in MapReduce we found 1.35 GB data mitigate the read/write issue and does not work with HDFS shown in the Fig (3.1).

MRAP also significantly outperforms almost nearing 1GB (0.95GB) of data mitigate the read/write issue. But DMRAP uses 0.63GB of data mitigate for read/write issue. So this is almost 50% reduced MR and increases fault tolerance and scalability of data.

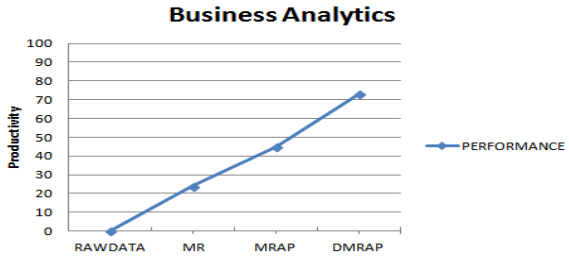

Figure 3.2: Overall Performance of DMRAP with Other Access Patterns

The above diagram illustrates the overall performance of DMRAP with other access patterns in order to increase the productivity that is the backbone of every business in the modern trends.

IV. Conclusion

Handling huge amount of unstructured and structured data is difficult for business analytics, for quick decision making. MapReduce and MapReduce Access Patterns (MRAP) involved in improving the productivity but lags in storage issues and speed due to the demand increased for Big data solutions. We introduced a new methodology Distinct MapReduce with Access Patterns (DMRAP) that is capable of handling and processing Big data in three stages, and also increases the productivity and performance of data processing regardless of the amount of data collected from different locations are of datasets. The final results of DMRAP is observed around 50% of improvement in processing data and 28% of increase in performance and productivity.

References

[1] Katal, A., Wazid, M. and Goudar, R.H. Big data: issues, challenges, tools and good practices. In *Sixth International Conference on Contemporary Computing (IC3)*, 2013, 404-409.

[2] Bansal, S. and Rana, D.A. Transitioning from Relational Databases to Big Data. *International Journal of Advanced Research in Computer Science and Software Engineering* **4** (1) (2014).

[3] Uzunkaya, C., Ensari, T. and Kavurucu, Y. Hadoop ecosystem and its analysis on tweets. *Procedia-Social and Behavioral Sciences*, 2015, 1890-1897.

[4] Kulkarni, A.P. and Khandewal, M. Survey on Hadoop and Introduction to YARN. *International Journal of Emerging Technology and Advanced Engineering* **4** (5) (2014) 82-87.

[5] Ghazi, M.R. and Gangodkar, D. Hadoop, MapReduce and HDFS: A Developers Perspective. *Procedia Computer Science*, 2015, 45-50.

[6] Sehrish, S., Mackey, G., Wang, J. and Bent, J. Mrap: A novel mapreduce-based framework to support HPC analytics applications with access patterns. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, 2010, 107-118.

[7] Ram, J., Zhang, C. and Koronios, A. The Implications of Big Data Analytics on Business Intelligence: A Qualitative Study in China. *Procedia Computer Science*, 2016, 221-226.

[8] Internetlivestats.com. Twitter Statistics. Retrieved from. http://www.internetlivestats.com/twitterstatistics/ (2015).

[9] Marín-Ortega, P.M., Dmitriyev, V., Abilov, M. and Gómez, J.M. ELTA: New Approach in Designing Business Intelligence Solutions in Era of Big Data. *Procedia Technology*, 2014, 667-674.

[10] Sharma, S., Tim, U.S., Wong, J., Gadia, S. and Sharma, S. A brief review on leading big data models. *Data Science Journal* **13** (2014) 138-157.

[11] Kimble, C. and Milolidakis, G. Big data and business intelligence: Debunking the myths. *Global Business and Organizational Excellence* **35** (1) (2015) 23-34.

[12] The Economist. Data, data everywhere. Available at http://www.economist.com/node/15557443 [Accessed 16th 2015].

[13] Selene Xia, B. and Gong, P. Review of business intelligence through data analysis. *Benchmarking: An International Journal* **21** (2) (2014) 300-311.

[14] Davenport, H.T. How strategists use big data to support internal business decisions, discovery and production. *Strategy & Leadership* **42** (4) (2014) 45-50.

[15] Erevelles, S., Fukawa, N. and Swayne, L. Big Data consumer analytics and the transformation of marketing. *Journal of Business Research* **69** (2) (2016) 897-904.

[16] Narayanan, V. Using big-data analytics to manage data deluge and unlock real-time business insights. *The Journal of Equipment Lease Financing* **32** (2) (2014).

[17] Wang, L. and Alexander, C.A. Big data driven supply chain management and business administration. *American Journal of Economics and Business Administration* 7 (2) (2015) 60.

[18] Tankard, C. Cultural issues in security and privacy. *Network security* (2012) 5-8.

[19] Kowalczyk, D.W.I.M. and Buxmann, P. Big Data and information processing in organizational decision processes. *Business & Information Systems Engineering* **6** (5) (2014) 267-278.