International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

© Global Society of Scientific Research and Researchers

http://ijcjournal.org/

A Case for the Adoption of an In-Memory Based Technique for Healthcare Big Data Management

Famutimi R. F[a]*, Soriyan H. A.[b], Ibitoye A. O.[c], Famutimi T.I[d]

[a,c,d]*Computer Science & Info. Technology, Bowen University Iwo, Nigeria*

[b]*Computer Science & Engineering Dept., ObafemiAwolowo University, Ile-Ife, Nigeria*

[a]*Email: ranti.famutimi@bowenuniversity.edu.ng*

[b]*Email: hasoriyan@gmail.com*

[c]*Email: ibitoye_ayodeji@yahoo.com*

[d]*Email: ifeoluwatemitayo@gmail.com*

Abstract

In healthcare organizations, the amount of data that are generated daily are on the increase with every visit by patient. The generated data through vital signs' readings such as body temperature, pulse rate, respiratory rate, blood pressure, body weight among others are now accumulated into big data. Recently, the growth of data is averaged at about 35 percent annually. The implication is that the amount of storage needed to hold the data doubles within a period of three years. No doubt, if these data are processed and analyzed properly, it holds immense value in diagnosis and predictive medical conditions. However, the ever increasing volume of data has brought with it some big challenges. One of such is how healthcare organizations are going to store and access the vast amount of inherent information. In this paper, we discussed the need for storing medical Big Data in the main memory (In-Memory) as a way of addressing storage and access to information challenges of big data in health care delivery system. With current trends in technology advancement, there is an availability of storage systems with increased memory capacities. The storage of data in main memory can achieve a performance improvement of up to a factor of 100,000 or more. With this achievable performance, In-Memory Data Management proves to be a viable option.

Keywords: Big Data; in-memory; in-disk; columnar-oriented; caching; disk-I/O; OLAP.

1. Introduction

The large volume of data generated by healthcare organizations (and similar data intensive organizations) structured, unstructured, different varieties, high rate production, that are generated on a daily basis are referred to as Big data.

----

* Corresponding author.

It is also a broad term used for describing data sets that are either large or complex for the existing data processing applications tools to effectively process. [1]. Big data is the one that exceeds the processing capacity of conventional database management system [2]. While big data can be associated with it size (bytes), it can vary from one organization to another. This is so, because a particular big data size in an organization may not be big data in another organization, if there is an availability of the appropriate tools for processing it. In social media, Google is involved in processing Big Data that grew from 00 Terabyte (TB) of data a day in 2004 to processing 20 PetaByte (PB) in 2008 [3]. The characteristics of Big Data was said to encompass five V's: Volume, Velocity, Variety, Veracity, and Value. Volume pertains to vast amounts of data, Velocity applies to the high pace at which new data is generated, Variety pertains to the level of complexity of the data, Veracity measures the genuineness of the data, and Value evaluates how good the quality of the data is in reference to the intended results [4] .

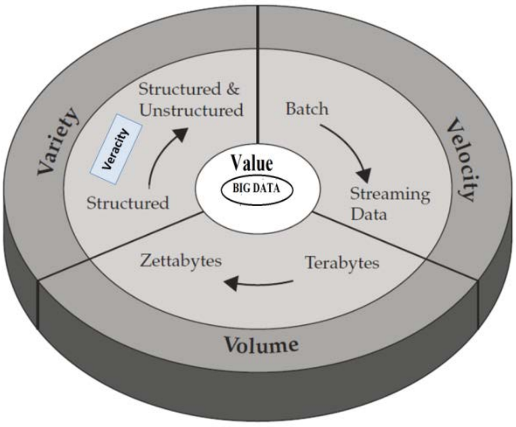

Figure 1.1: Pictorial view of Big data (Source: [5]), modified

It was estimated according to Berkeley researchers that the amount of data the world produced in the year 1999 was about 1.5 billion gigabytes. When the study was repeated in year 2003, the amount of data produced had doubled the previous amount, within a time space of three (3) years. In 2012, worldwide digital healthcare data was estimated to be equal to 500 petabytes and is expected to reach 25,000 petabytes in 2020 [6].

2. The relevance of big data

The size of data is not what matters, but what can be done with it. A normal size health data can be processed to obtain the treatment to be administered to a patient suffering from a particular sickness, the cost of treating an illness using the treatment chart, the availability of a particular drug in the pharmacy and so on. However, when you combine Big data with high powered and efficient analytical tools, you can accomplish broad task such as: predicting the root cause of an illness from an accumulated cases treated, a possible outbreak of an epidemic, a geographical spread of a particular disease and the possible period of the year a disease is likely to occur [7]. Also, analyzing treatment patterns by a physician can provide a better way of treating an ailment. A fraudulent transaction can be detected before it affects the organization, when using Big Data.

3. Challenges with big data

Big Data accommodates big volume with its complexities, the size of Big Data causes storage problem and makes it difficult when accessing data items with existing techniques, which leads to inaccurate analysis of data. In the critical application of Big Data such as the case of health domain, there is need for timely accessing of records. In addition, online analysis of data is essential for timely decision making to safe life and improve healthcare delivery. We live in a world that continuously make use of data intensive technologies. This will make the volume as well as the complexity of data been generated to continue to increase. The eventual result is the challenge of accessing and storage of this big medical data, which inevitably will require a well thought out solution. Such solution will resort to accurate predictions for management of big Data in many domains, detects fraudulent transactions in Online Analytical Processing (OLAP) and in particular improve healthcare delivery system through big data management systems' performance.

4. Proposed methodology

In this paper, we propose the use of single structure columnar In-Memory Data Management for addressing the accessing and storage challenges associated with Big Data. A special single structure columnar based model that references both columns and rows in a single structure is used to improve the materialization (extraction of record details) challenge. By this, a vector based dictionary encoding technique is employed on the memory resident columns. Through the advancement in technology, a positive change to the size of main memory availability with respect to time can be achieved. In previous years, the memory capacity of commodity servers has doubled every year culminating in servers with terabytes of main-memory [8,9]. Current technologies have revealed that data can be managed by database system using either an in-disk based techniques or in-memory based techniques. While the in-disk based techniques have been in use for quite a long time, the in-memory based technique is an evolving technique of data management, and it is still in its infancy stage with a promise of good success.

In an in-disk based data management, data reside mostly in the disk and it is moved to and from the memory as at when needed depending on the algorithm being used for memory management. An in-memory database system or main-memory database system is a breed of database management system that stores data entirely in main memory instead of keeping it on disk. In case of the normal Data Base Management System (DBMS) architecture known as in-disk, it is becoming more and more challenging to process the data and to produce analytical results in almost real time, which is Online Analytical Processing (OLAP). For in-disk databases, disk I/O operations are the main bottleneck, which result into very slow operations and can't be optimized beyond a limit being mechanical in nature [10]. In spite of the fact that traditional in-disk DBMS have introduced various measures to speed up its operations through different techniques, it has not brought the much desired result. As a result of advancement in technological innovations, the cost of main memory is drastically reducing, which makes storage of large amount of data (Big Data) in main memory a reality. The moment data is stored in the main memory, the speed of accessing and manipulation changes drastically. If data reside in the main memory, there will be no need for caching of data which on its own pose another logistics. Applications that require real-time processing will benefit immensely in the in-memory database technology approach.

With the noticeable advent of new applications and upcoming hardware improvements, tremendous changes are still expected to take place in enterprise software. The next generation database technologies will clearly deviate from traditional databases on disk to in-memory databases technology. It has been found out that the use of in-memory database technologies enables performance improvements by factors of up to 100,000 [11]. There are two types of In-Memory Data Management techniques: Row-oriented and Column-oriented (Columnar). In row-oriented storage, complete tuples (records) are stored in adjacent blocks of the memory. Columnar oriented storage technique stores complete columns in adjacent blocks. When all fields are not needed for processing, a row-oriented technique will still scan all the fields of a record before advancing to the next record for the selected field until the whole records are scanned. There is a delay being introduced while scanning unwanted fields. In case of columnar technique, as soon as the first required field (column) is obtained, all the adjacent successive storage locations contain the same column (field) of successive records. In columnar technique of data management, unwanted columns are not scanned, the moment the first item is obtained [11].

The most effective among the In-Memory Data Management techniques is the columnar- oriented one. Column-oriented storage, in contrast to row-oriented storage, is well suited for reading consecutive entries from a single column. This can be useful for aggregation and column scans. Effective In-Memory Data Management focuses on the technical details of in-memory columnar databases. Over the last years, in-memory databases and especially column-oriented in-memory databases grew in popularity and became widely researched. With modern hardware technologies characterized by increasing main memory capacities, In-Memory Data Management techniques is now a very viable option for managing Big Data in applications [12].

5. Conclusion

In healthcare big data sources, there are repetitions of data. A technique that uses a memory based compression such as dictionary encoding, augmented with appropriate algorithm for storage and retrieval of information, as well as appreciable reduction in input and output overhead, (single structure columnar in-memory database management) has been proposed for healthcare big data management. Many advantages offered by In-Memory data management and the upcoming advancement in memory technologies have been outlined. It is also expected that with the continuous improvements in main memory technology and high speed networks, the future Enterprise Solutions will embrace fully, the use of many versions of In-Memory based data Management for Big Data in healthcare systems .

6. Recommendations

The study focuses on the storage of text data. Since big data consists of all varieties of data, an attempt can be made to study the effect of other data formats when using this proposed methodology.

References

[1]. M. Khalilian, N. Mustapha, N. Sulaiman (2016). ”Data stream clustering by divide and conquer approach based on vector model". Journal of Big Data. (2016) [On-line]. 3:1. Available: https://journalofbigdata.springeropen.com/ [Jan 31, 2017].

[2]. Gartner (2011). “Pattern-Based Strategy: Getting Value from Big Data”. Gartner Group press release. July 2011. [On-line]. Available: thttp://www.gartner.com/it/page.jsp?id=1731916 [Feb 10, 2017].

[3]. J. Dean, and S. Ghemawat (2008). “MapReduce: Simplified data processing on large clusters”. Communications of the ACM, 51(1):107–113, 2008 [On-line]. Available: https://research.google.com/archive/mapreduce-osdi04.pdf [July 26, 2017]

[4]. Y. Demchenko, C. Ngo, and P. Membrey (2013). “Architecture Framework and Components for the Big Data Ecosystem.System and Network Engineering (SNE)”. publication.Universiteit van Amsterdam [On-line]. Available : www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf. [July 26, 2017]

[5]. M.Grobelnik (2012). “Big Data – Growing torrent”. Stavanger, May 8, 2012. Jozef Stefan Institute Ljubljana, Slovenia. [On-line]. Available: https://www.planet-data.eu/sites/default/files/presentations/Big_Data_Tutorial_part4.pdf . [July 26, 2017]

[6]. J. Sun, and C.K. Reddy (2013). “Big Data Analytics for HealthCare” .Tutorial presentation at the SIAM InternationalConference on Data Mining, Austin, TX, 2013. [On-line] . Available: https://www.siam.org/meetings/sdm13/sun.pdf [July 26, 2017]

[7]. P. Groves, B.,Kayyali, D. Knott, S. Kuiken (2013) “The ‘big data’revolution in healthcare”. Center for US Health System ReformBusiness Technology Office Publication. [On-line]. Available: www.pharmatalents.es/assets/files/Big_Data_Revolution.pdf [June 5, 2017]

[8]. P.A. Bernstein, C.W. Reid, and S. Das (2011). “Hyder - A TransactionalRecord Manager for Shared Flash”. In: CIDR. 2011, pp. 9–20.5th Biennial Conference on Innovative Data Systems Research (CIDR '11) January 9-12, 2011, Asilomar, California, USA. [On-line]. Available : http://cidrdb.org/cidr2011/Papers/CIDR11_Paper2.pdf [July 11, 2017]

[9]. H. Zhang, G. Chen, B. Chin, and K. Tan (2015) “In-Memory Big Data Management and Processing” . IEEE Transactions On Knowledge and Data Engineering, vol. 27, no. 7 [On-line] . Available: ieeexplore.ieee.org/document/7097722. [July 10, 2017]

[10]. M.K. Gupta, V. Verma, M.S. Verma (2013).”In-Memory Database Systems - A Paradigm Shift”. International Journal oFEf Engineering Trends and Technology (IJETT) – Volume 6 Number 6- Dec 2013 ISSN: 2231-5381.:333 [On-line]. Available: http://www.ijettjournal.org[Feb, 11, 2017].

[11]. H.Plattner (2015), “The Inner Mechanics of In-Memory Databases”. Hasso Plattner Institute of IT Systems Engineering, Universitat Potsdam [On-line]. Available: https://open.hpi.de/courses/imdb2015 [July 25, 2017]

[12]. A. Kemper, T. Neumann, F. Funke, V. Leis and H. Mühe (2012) “Hyper: Adapting columnar main-memory data management for transactional and query processing”. IEEE Data Eng. Bull., 35(1):46–51, 2012 [On-line]. Available:. http://sites.computer.org/debull/A12mar/p46.pdf