International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

http://ijcjournal.org/

Comparative Study of Disk Resident and Column Oriented Memory Resident Technique for Healthcare Big Data Management Using Retrieval Time

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

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

[c]*Obafemi Awolowo University, Computer Science & Engineering Dept, Ile Ife, Nigeria*

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

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

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

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

Abstract

The rate at which information are being shared among people of diverse discipline, is continuously increasing the volume of data available for different forms of processing and storage. The channels for collecting data is increasing on daily basis; customers need to supply data to business owners; online social media keep on evolving; educational institutions are faced with keeping records of ever growing students' enrollment and keeping their records after graduating is now a challenge; health institutions keep on experiencing unprecedented growth in child birth on daily basis and the need to keep and maintain adequate health records is a necessity. This resultant data flood has called for the need to explore new cost effective storage options and analysis techniques in other to benefit from the dividends of Big Data. Some of the approaches involve investing more on hardware storage devices, some involve exploring other locations' facilities while some adopt improved software techniques. This paper is presenting some of the results obtained using software techniques. In this research, an improved column vector memory resident (in-memory) database management was employed to manage Big Data in which a comparative study of Disk and Memory resident Big Data mining from the study was shown.

Keywords: Big Data; columnar-oriented; caching; disk-I/O; disk-resident; memory-resident; speed-up.

---

* Corresponding author.

1. Introduction

Big data is a broad term used for describing data sets that are so large or complex that traditional data processing techniques and applications are inadequate to effectively process. It is also referred to as complex data that is difficult to process when using traditional database and software techniques [10]. Big data is the one that exceed the processing capacity of conventional database management systems [7]. Big data is a term that describes the large volume of data, both structured and unstructured that organizations generate on a day to day basis In the social media, Google is involved in processing Big Data that grew from a Terabyte (TB) of data a day in 2004 to processing 20 PetaByte (PB) in 2008 [2]. The characteristics of Big Data were said to encompass five V's: Volume, Velocity, Variety, Veracity, and Value. Volume relates to huge amounts of data, Velocity applies to the high pace at which new data is generated, Variety is the level of complexity of the data, Veracity measures the genuineness of the data, and Value evaluates the usefulness of the data [3]. From another perspective, data are seen as "Big Data", not because of it volume but for the inherent complexity, diversity and timeliness involved in the processing. It has been revealed that a lot of players in ICT world are constantly experimenting new ways of building applications and analytical tools to aid patients, physicians and healthcare stakeholders make better use of available healthcare information. It is expected that these innovators will employ interesting ideas for using big data so as to eventually reduce the cost of providing good health care services [8]. Figure 1 depicts a pictorial view of a Big data.

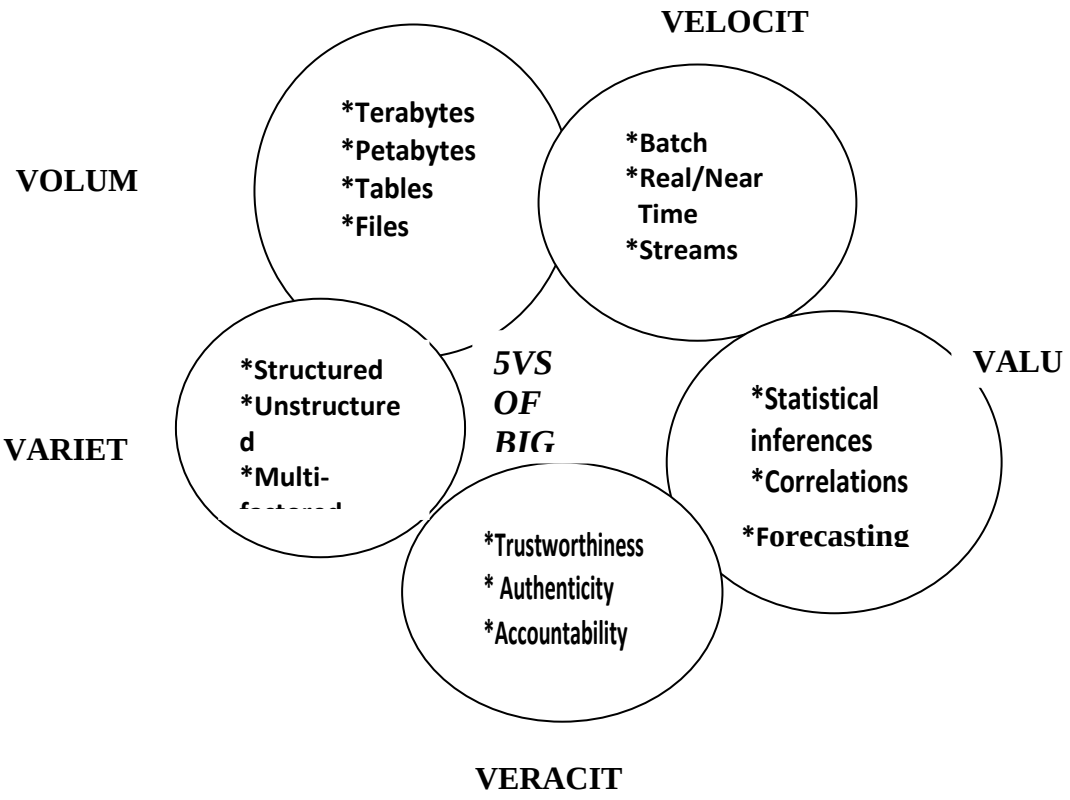

Figure 1: 5vs of big data: [3] modified

1.1 The place of big data in our contemporary society

According to [4] the place of Big data was expressly stated as comprising of the following: the incursion of Data into every industry and business function has now become an important factor of production; Big data creates value in several ways by creating transparency, enabling experimentation to discover needs, expose variability, improve performance, segmenting populations to customize actions, replacing and supporting human decision making with automated algorithms, innovating new business models, products, and services ;use of big data will become a key basis of competition and growth for individual firms through data driven decision making [1]. It is also expected that big data will open new avenues of production expansion and consumer satisfaction, greater gains for more sectors of the economy, high demand for needed skill that will to take the advantage of big data, access to data, industry structure.

Big Data, which is the information obtained from non-traditional sources ranging from blogs, social media, email, sensors, photographs to video footages, which are typically unstructured and voluminous has the potential of showing enterprises more insight into their customers, associates and business in general [14]. Big Data can provide answers to issues they never thought about. There are also more benefits to be derived when a business is viewed from a multidimensional view in addition to the traditional types of information they are familiar with. Big data, it is said to be similar to traditional data in all ways. It must be captured, stored, organized, analyzed. After analysis, the result must be integrated into the business' established processes so as to influence the operation of the business. The issue with Big Data is that because it comes from new sources of data types that were not previously extracted for insight, entrepreneurs are not used to collecting such information from the sources neither are they used to processing large volume of data both structured and unstructured. This then make room for missing many opportunities for gaining insight.

Today's information societies are characterized by huge storage of data, either from private domain or public domain. This suggests that any workable applicable must be able to upgrade datasets used for the domain of interest. Any organization that involves itself with gathering, analyzing, monitoring, filtering, searching, or organizing contents must address the issue of large-data, which is known as data-intensive processing [13].

1.2 Big data management approaches and challenges

The most common approach is the Disk-Resident (In-Disk) Database Management. In this system, also known as the traditional database management, it is the one that has its architecture on the industry standard Database Management Systems (DBMSs) which were the result of many decades of research and development activities in industry and academia. It is now a house hold name in all computing environments. In this systems data reside in the disk devices. DBMSs were used as pioneering systems for designing scalability and reliability techniques in many systems [9]. A traditional DBMS employs disk-based storage for the main part of the data as well as meta-data of the system. Main-memory is used for buffer-management and caching as well as for frequently accessed meta-data. Indexes, like the actual data, are mainly stored on disk. In these DBMS, transaction execution is designed to deal with disk input and output (I/O). This is the main bottleneck. In order to minimize the I/O bottleneck, it maintains a buffer pool of disk pages which currently reside in main

memory. The scenario is such that: when a transaction requires a page which does not currently reside in main-memory, the page is fetched from disk resulting in a delay in the order of tenth of milliseconds [6]. During this suspension, traditional disk-based database systems will try to use the available CPU time to make progression other transactions which are not waiting for I/O at the moment. This requires interleaved execution of transactions and a concurrency control mechanism which ensures compliance with a defined isolation level [12]. Main memory is that portion generally referred to as 'RAM' or internal memory of the system. It is the storage portion that data must reside before the CPU can process it. This is unlike the Disk storage that is not directly accessible by the CPU, hence the need to move it to the main memory before any further processing can take place. Main memory is a volatile storage made of dynamic Ram (DRAM). It losses it information when electric or the power source is withdrawn. In-spite of this, the available of DRAM with vast amount of storage and lowering cost of purchase have made it a viable choice for an In-Memory database management. In the emerging In-Memory database management, the main memory is now taking over the assignment of Disk in traditional database management, and the DISK is now being considered as a backing up storage. Data are transmitted between Disk and main Memory in pages, where a page consist of many disk sectors. A buffer is usually used to facilitate the movement of data between the disk and the main memory due to speed differential of the two devices. Buffer is also a form of bridge between the Disk and main Memory. When the contents of main Memory do not need refreshing as long as there is power supply, it is referred to as Static Random Access (SRAM). However when the contents are periodically refreshed, even with the presence of power supply, it is termed as Dynamic Random Access Memory (DRAM) [11]. The dynamic Random Access memory can be further divided into Synchronous Dynamic Random Access Memory (SDRAM), which is a faster variation of DRAM and Mobile Random Access Memory (MRAM). The focus of this research is the use of DRAM type of main Memory to store data.

2. Materials and methods

According to [5], if a column vector based dictionary encoding technique is employed to store data in the main Memory, there will be a significant improvement in the time taken to retrieve record items stored when compared with the common method of storing data in the Disk. The approach used in this paper is based on this technique. We therefore implemented a Main Memory-Resident database management system using a column vector management technique for the fields of a health care records' database (Made in Nigeria Primary and Hospital Information System - MINPHIS). This information system has been indicating processing challenges due to the size of its database. We benchmarked it with a Microsoft SQL traditional database management system on the same database with the same platform and architecture. Enquiries for retrieving record items consisting of different data sets of records were carried out on the database using each approach (Disk storage approach) and (main Memory storage approach) so as to compute the time taken for each. Different retrieval enquires were carried out and the average time taken were recorded for ten query evocations (runs) on each data set retrieved. Enquiries with one-field such as: select pat-id from patient-history; two fields such as: select pat-id, surname from patient-history; three fields such as: select pat-id, surname, other-name from patient-history; four fields such as: select pat-id, surname, other-name, date-of-discharge from patient-history; five fields such as: select pat-id, surname, other-name, date-of-discharge, state-of-discharge from patient-history; six fields such as: select pat-id, surname, other-name, date-of-discharge, state-of-discharge, condition-on-discharge from

patient-history; seven fields such as: select pat-id, surname, other-name, date-of-discharge, state-of-discharge, condition-on-discharge, diagnosis from patient-history. The improvement on the time taken to retrieve each dataset (Speed Up) was also computed

3. Results

The tabular and the graphical interpretations of the results were shown below. Table 1 shows the result obtained from disk resident database with 1000 records being retrieved for many runs, table 2 shows the result of main-memory resident database, table 3 compares the two results, table 4 shows the speed up obtained with main memory resident. Figure 1 is a graphical representation showing retrieval time versus number of fields, and figure 2 is showing speed-up value versus number of fields retrieved when main memory and disk database are compared.

Table 1: Disk resident database retrieval of 1000 records in milliseconds

Number of Runs	One-Field	Two-Fields	Three-Fields	Four-Fields	Five-Fields
1	546	827	905	983	1092
2	562	920	905	1092	1154
3	561	811	921	1030	1092
4	561	796	920	1014	1101
5	562	842	905	998	1155
6	561	827	889	983	1139
7	561	795	905	998	1108
8	562	811	920	982	1139
9	562	811	921	998	1154
10	562	811	905	983	1076
AVERAGE	560	825	910	1006	1121

Table 2: Memory resident database retrieval of 1000 records in milliseconds

Number of Runs	One-Field	Two-Fields	Three-Fields	Four-Fields	Five-Fields
1	281	468	577	640	718
2	265	468	608	687	718
3	265	468	593	655	749
4	249	484	593	671	718
5	265	484	593	639	827
6	296	453	593	686	749
7	296	468	639	671	811
8	310	453	578	640	827
9	327	546	577	639	718
10	281	458	593	639	717
AVERAGE	284	475	594	657	755

Table 3: Main-memory resident/ disk resident compard

fields	retrieval time for main memory DB	retrieval time for disk database
1	284	560
2	475	825
3	594	910
4	657	1060
5	755	1121

Table 4: Retrieval times with speed up

fields	main memory resident database	disk resident database	speed up
1	284	560	2
2	475	825	1.7
3	594	910	1.5
4	657	1060	1.6
5	755	1121	1.5

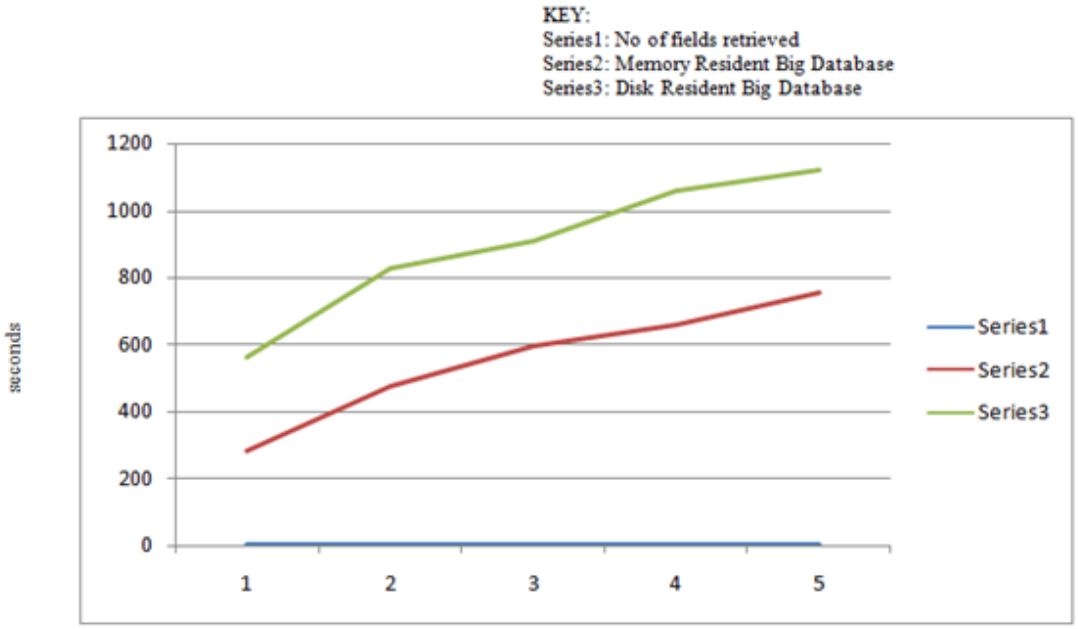

Figure 1: Graphical representation showing retrieval time versus number of fields

No of fields retrieved

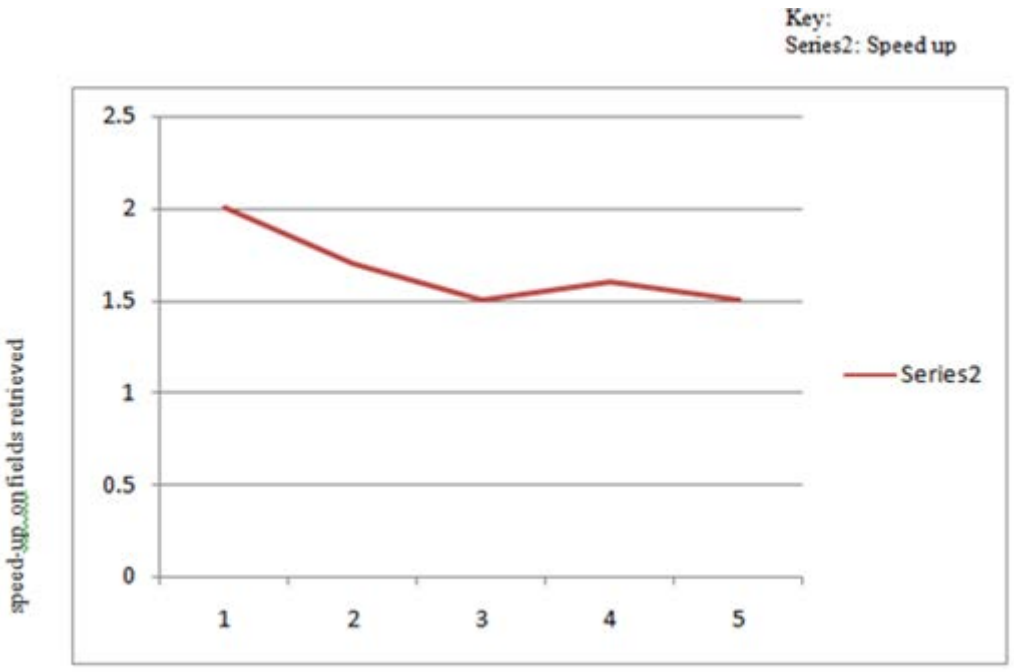

Figure 2: Graphical representation showing speed-up value versus number of fields retrieved

No of fields retrieved from big data

4. Conclusion

In this study a new method based on column vector memory resident (in-memory) database management was employed to manage health care big data complexities. The study used column vector approach because it is a software based technique that can be used to manage database resident in the main memory. Also, a health care database records that has exhibited the characteristics of big data was used for the implementation of our prototype. From our experiment, it was shown that for all categories of data retrieval from the Big Data, the memory resident database provided an average of about a speed up of 2, meaning up to 200% improvement on the time taken to retrieve thousands of data items when compared with the disk resident database..

5. Recommendation

The study used only retrieval time to evaluate the proposed technique for managing healthcare related big data; other parameters can also be used to evaluate the technique in this study.

References

[1]. Brynjolfsson, E., Hitt, L., and Kim, H. (2011). “Strength in Numbers: How Does Data-Driven Decision making Affect Firm Performance?” ICIS 2011 Proceedings. 13. https://aisel.aisnet.org/icis2011/proceedings/economicvalueIS/13.

[2]. Dean, J. and Ghemawat, S.(2008).”MapReduce: Simplified data processing on large clusters”.

Communications of the ACM, 51(1):107–113.

[3]. Demchenko, Y., Ngo, C,.Membrey, P.(2013). "Architecture Framework and Components for the Big Data Ecosystem". System and Network Engineering (SNE) publication.Universiteit van amsterdam

[4]. Eijnatten, J.V., Toine, P. and Verheul, J. (2013). "Big Data for Global History, The Transformative Promise of Digital Humanities". BMGN - Low Countries Historical Review , Volume 128-4 (2013) ,| pp. 55-77

[5]. Famutimi, R. F, Soriyan, H. A, Ibitoye, A. O., and Famutimi, T.I. (2017). "A Case for the Adoption of an In-Memory Based Technique for Healthcare Big Data Management". International Journal of Computer (IJC). 27(1):141-145.

[6]. Gajakosh, S and Takalikar, M.(2013). "Multitenant Software as a Service: Application Development Approach". International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-3 Number-3 Issue-11.

[7]. Gartner (2011). "Pattern-Based Strategy: Getting Value from Big Data". Gartner Group press release. July 2011. Available at http://www.gartner.com/it/page.jsp?id=1731916

[8]. Groves, P.,Kayyali, B., Knott, D.,Kuiken, S.,(2013) "The 'big data'revolution in healthcare". Center for US Health System Reform Business Technology Office Publication.

[9]. Hellerstein, J.M., Stonebraker, M. and Hamilton,J.(2007). "Architecture of a database management System". Foundations and Trends in Databases. Vol. 1, No. 2 (2007) 141–259

[10]. Khalilian, M.,Mustapha, N., Sulaiman, N. (2016). "Data stream clustering by divide and conquer approach based on vector model". Journal of Big Data. (2016) 3:1

[11]. Lankhorst, M.H.R., Ketelaars, B.W.S.M. and. Wolters R. A. M. (2005). "Low-cost and nanoscale non-volatile memory concept for future silicon chips". Nature Materials 4, 347 - 352 (2005).doi:10.1038/nmat1350.

[12]. Larson, P. (2013) "Evolving the Architecture of SQL Server for Modern Hardware".IMDM 2013.

[13]. Lin, J. and Dyer, C. (2010). "Data-Intensive Text Processing with MapReduce. Manuscript of a book in the Morgan & Claypool Synthesis". University of Maryland, College publication Park.https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf. Accessed December 22, 2015.

[14].Oracle (2014). "Integrate for Insight". Oracle Big Data strategy guide, http://www.oracle.com/us/technologies /big-data/big-data-strategy-guide-1536569.pdf accessed December 22, 2017.