Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 7 – No. 12, January 2018 – www.caeaccess.org

Towards a Big Data Architectural Framework for Healthcare in Ghana

Edem Adjei
IPMC College of Technology
Ghana

Nana Kwame Gyamfi
Kumasi Technical University
Department of Computer Science

David Otoo-Arthur
Presbyterian Women's College of Edu, Ghana

ABSTRACT
Data is deemed as a gold mine if and only if it is analyzed and utilize. The healthcare system is one of the largest generators of data due to strict adherence to regulatory structure. Unfortunately, Big Data deployment in the healthcare industry has not catch up in Ghana and Africa at large. Big Data in healthcare has enormous benefits including the designing of Predictive models, analyzing disease patterns and tracking disease outbreaks, turning large data into actionable information, Evidence based health delivery through data analysis and Capture and analyze real time data from variety of locations. To achieve above mentioned potentials of Big Data, this thesis has taken a look at the structure of Big Data, which has led to the development of an architectural framework that will fit into the system Ghanaian healthcare system and how the variety of data will be handled and stored. A framework which will serve as a platform for data analytics in the Healthcare industry is also proposed. Finally, we propose a framework which will handle the new data generating devices used by health facilities that is the structured and unstructured data types.

Keywords
Healthcare System, framework, Big Data, structure and unstructured Data, architectural framework

1. INTRODUCTION
Access to healthcare must be basic [1]. During these processes of accessing healthcare, data is generated which is termed as the Medical Record of the patient and this medical record can grow over time depending on the frequency of a

patient at a health facility. This driven by record keeping, compliance and regulatory requirements and patient care [2]. With healthcare records still stored in Hard Copy and the population of Ghana will grow up to 27 million at the end of 2014 [3], this suggest it will be impossible to use these medium of storing health records in Hard copy form, therefore the need for rapid digitization of these large amount of data in the nearest future.

These healthcare data are important to the health facility, health practitioners and patient. The healthcare record shows the way the patient is responding to treatment and the progress being made [4]. This process of keeping health records also helps health practitioners in the case of referrals, It is also the only source of evidence for the health facility and or medical practitioner in the event of a lawsuit and fraud detection [2]. The records serve as prove of the processes the practitioner engaged in the delivery of healthcare to their respective client.

The varying nature of data types has made it imprudent for relational database management systems to store and manipulate these data types [5]. These data types require huge storage due to its exponential growth over time. Effectively dealing with these data types to maximise its benefits in healthcare means that, there should be live streaming of data, which can be accessed from various locations. These records should be also accessed in a timely fashion since it would involve peoples' life. The data would have the following characteristics, huge volume and scale, variety and heterogeneity of Data/Sources, speed and timeliness of information requirement, targeted services and solutions, data presentation, usability and interpretation and privacy, error handling and security [5].

One way of analysing digitised data can be through a database technology known as "Big Data" [5]. Big Data is defined by its basic characteristics which is termed as the 3V's and an addition of other "V" [2] that is Volume, Variety, Velocity and Veracity. Volume means large amount of data (size), variety means data comes in form heterogeneous data source and velocity meaning data arrives with high speed [6]. In the context of "big", means that the datasets grow so much that it becomes difficult to manage it using existing DBMS [5]. With the health sector being an essential part of every nations growth, it is important to put structures in place to deal with health-related issues and such a structure is a medical record system that can be accessed from any part of the country supporting a wide range of medical and healthcare function including decision support, diseases surveillance and population health management as also proposed by the Vice President of the republic of Ghana in December 2014. H.E. Amissah-Arthur (Vice President of Ghana) has also proposed in 2014 [7].

The Ministry of Health and Ghana Health Service has no electronic system for tracking record of data generated within a period of time. Even though it is difficult to ascertain the volume of data generated in Ghana, it can be established that, data generation is exponential. Korle-Bu Teaching Hospital has an average attendance of 1,360 patients in a day at the OPD alone. Korle-Bu Teaching Hospital also receives the most referrals than any hospital in Ghana according to a report by the Ministry of Finance [8]. In a less populated area such as the St. Martins De Porres Hospital at Agomanya [9] in the Eastern Region of Ghana has an average attendance of about 260 patients between 2008 and 2010. The total number of health facilities in Ghana as at the end of 2009 was 3217 [10]. Contrary to the situation in Ghana. A reports from the United States healthcare system stated that, medical record data collected reached 150 Exabytes and at the rate at which data is growing in the healthcare sector, it will soon reach zetabytes (1021 gigabyte) with Kaiser Permanent, a California based health network in Califonia alone is believed to have between 26.5 and 44 petabytes of data [6].

Ghana's population is expected to grow up to 27 million [3] in

2014, comparatively to Kaiser Permanent of California health network, data for healthcare in Ghana is expected to be between 79.5 and 132 petabytes. This therefore emphasises the necessity of getting Big Data system to manage the flow of data in the heath sector.

Data are evolving; exponentially growing and with varying types; it is inefficient or almost impossible to use Relational Databases Management Systems to handle this variety of data streaming into the system. Relational Databases Management system can easily handle homogenous data but not all types heterogeneous data in the health care sector which includes ultra scan data, video, audio, pictures, location data, simulations Magnetic Resonance Imaging (MRI) and so on [6]

These data types to be stored and managed has huge sizes. This means that, it will occupy huge space ranging from the size of petabyte (1015) to yottabyte (1024). The maximum available space for SQL 2014 for Business Intelligence, Enterprise, Standard and Web applications is 524 petabytes (524x1015) [11], this means that, with the exponential growth of data generated in the medical sector on daily basis, it will be difficult and store data in this type of RDBMS.

The problems identified in the medical record keeping system of Health Facilities are

- Huge data is generated daily at healthcare facilities daily with storage and quick access to patient history very difficult. (this can be evident as stated in the introduction)
- No definitive way of analysing huge data generated over a period of time, making data collected almost unimportant and irrelevant.
- The evolution of different types of data that are to be stored in the medical field has changed and cannot be handled by RDBMS.

The purpose for this research is therefore to develop an architectural framework for keeping Healthcare data of Health facilities in Ghana via the Big Data technology.

2. RELATED WORK

By definition, Wullianallur *et al*, explains big data in healthcare as "an electronic health data set so large and complex that they are difficult or impossible to manage with traditional software and hardware nor can easily be managed with traditional or common data management tools and methods. [13]" According to Wullianallur *et al*,, the healthcare industry is being historically noted for the generation of very large amount of data which are stored in hard copy form, driven by the practice of record keeping and compliance to regulatory requirements. The digitization of the healthcare delivery reduces cost in areas like clinical decision support, population health management and disease surveillance which needs big data to manage; these data generated is huge with reports from US healthcare system alone reaching 150 Exabyte in 2011. This data is expected to grow between zettabyte and yottabytes [13].

An example is the Kaiser Permanete, a health network based in California with more than 9 million members is between 26.5 and 44 petabytes rich in data from their Electronic Health Record system which includes images and annotations accumulated over time with newer forms of Big data such as 3D imaging, genomics and biometric sensors also fuelling the exponential growth of data in healthcare [13]. Other places where Big Data has being successfully deployed to for Healthcare delivery is the North York General Hospital in Toronto Canada, Columbia University Medical Centre, Rizzoli Orthopaedic Institute in Bologna, Italy, Brigham and Women's Hospital and many more. Africa and for that matter Ghana has not harnessed the potentials of Big Data analytics.

3. BIG DATA ARCHITECTURAL FRAMEWORK FOR HEALTHCARE IN GHANA

The section looks at the process involved in developing proposed architecture for big data in healthcare in Ghana. This will involve ontological designs that depict the stages or states of the architectural framework for the Big Data in healthcare. The generic model for dealing with healthcare data as depicted in Fig. 1 deals with the data sources to where data transformation, modelling and analysis are done.

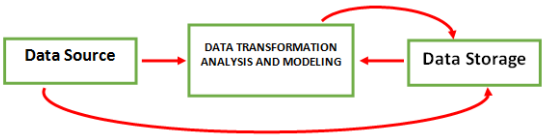

Fig. 1 A generalized model from the processes of healthcare delivery in Ghana

a. Data Sources: Data generated in healthcare may be either internal or external as illustrated in fig 3.2. The Ministry of Health in Ghana categorizes healthcare facilities into many groups according to the size, the services they deliver and ownership. Even though these are islands of health facilities, there Ministry of Health and the Ghana Health Service serve as a regulatory body for these healthcare facilities. These facilities includes teaching, regional and district hospitals and clinics which is government owned, health training institution and research centers, faith based mission facilities and traditional and alternate health providers and Private Hospitals, Clinics and maternity Homes.

The health record of patience will be generated internally at the healthcare facilties. This includes the vitals and biodata of the patient in a digital format could be termed as a Electronic Health Record (EHR). The other form of health data is derived from Clinical Decision Support Systems (CDSS) that assist health professionals and physicians with decision making. In Ghana, the Most of the health facilities are hooked onto the District Health Information Management Systems (DHIMS 2) [13] an Information Management System to capture service delivery points at the hospitals and merged at the district level and transmitted to the region and finally at the national level. Information Management System to In a Health Summit (May, 2014) it was agreed to hook the teaching hospitals, private sector and quasi- government health institutions onto the DHIMS 2.

The External sources may be from Governments as a regulatory body for healthcare delivery in Ghana, Laboratories, and Pharmacies which do not reside in the health facility. With the NHIA as a body which deliver health insurance services, there will be data from the authority which seek to check validity, fraud, claims payment and so on. There are also of healthcare devices which has embedded sensors to stream location data to healthcare facilities. Social media for the healthcare will also stream huge sums of data which needs to be analysed for decision making.

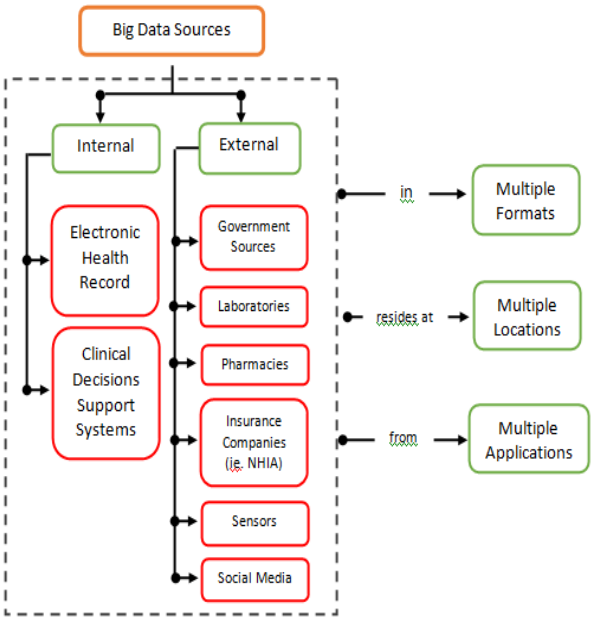

Fig. 3.2 Big Data sources and immerging Data trends in healthcare

b. Big Data Tools

Hadoop Distributed File System (HDFS): Hadoop Distributed File System enables the underlying storage for Hadoop cluster. It divides the data into smaller parts and distributes it across the various servers and or node. HDFS handles fault tolerance using data replication by creating replicas of each data block on different datanode for the purpose of reliability and availability [22].

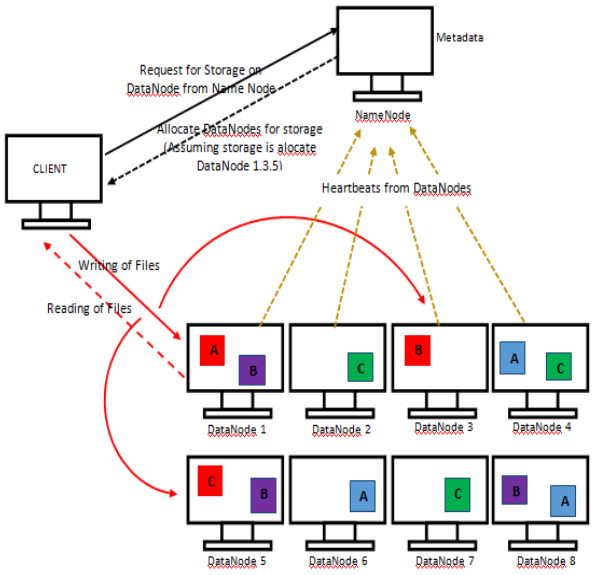

Fig. 3 HDFS architecture where the files are Replicated onto other DataNodes

MapReduce: MapReduce provides the interface for distribution of sub-tasks and the gathering of outputs. When task are executed, Map Reduce tracks the processing of each server or node. MapReduce was introduced to provide easy-to-use programming model that features fault tolerance, automatic parallelization, scalability and data locality-based optimizations according to Lui *et al* [15] .He further explains that main concept behind the MapReduce algorithms is to split work to different computing units to achieve parallelism.

In the map phase, Sara *et al* explains that, input data is split into independent sub-programs and distributes it from the master node to the workers node for parallel processing in smaller units and back to the master node [16]. In the reduce phase, the master node takes the results to all the sub problems and put them together to form an output.

Fig 4. The MapReduce Process

HBase: HBase is a column-Oriented database management system that sits on top of HDFS. It is well suited for sparse data sets, which are common in many big data use cases. Unlike relational database systems, HBase does not support a structured query language like SQL; in fact, HBase isn't a relational data store at all. HBase applications are written in Java much like a typical MapReduce application. HBase does support writing applications in Avro, REST, and Thrift. An HBase system comprises a set of tables. Each table contains rows and columns, much like a traditional database. Each table must have an element defined as a Primary Key, and all access attempts to HBase tables must use this Primary Key" [17](IBM).

Hive: Hive is a runtime Hadoop support architecture that leverages Structure Query Languages (SQL) with the Hadoop platform. It permits SQL programmers to develop Hive Query Language (HQL) statements akin to typical SQL statements. Hive provides tools to enable easy data extract/transform/load (ETL), a mechanism to impose structure on a variety of data formats, access to files stored either directly in Apache HDFS or in other data storage systems such as Apache HBase and query execution via MapReduce. [18]

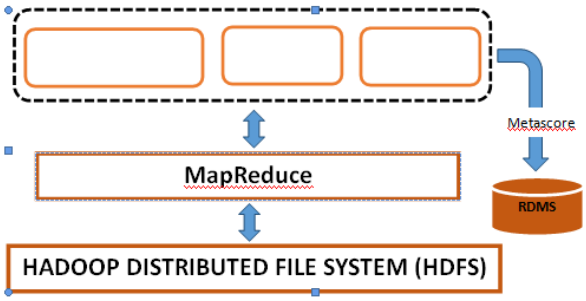

Fig 3.6.1 Hive Architecture

Due to the fact most of the healthcare facilities have digitalised their operations, the data generated are mainly from query languages. It is therefore important to have a means of processing data of such nature which is mostly generated from internally and Hive does the job of processing these query languages. Hive is built on Hadoop and uses map reduce for execution and HDFs for storage to query and manage unstructured data as if it were structured. as shows in fig. 5 Even though Hive is for query languages, it is not designed for Online Transaction Protocols and real time queries.

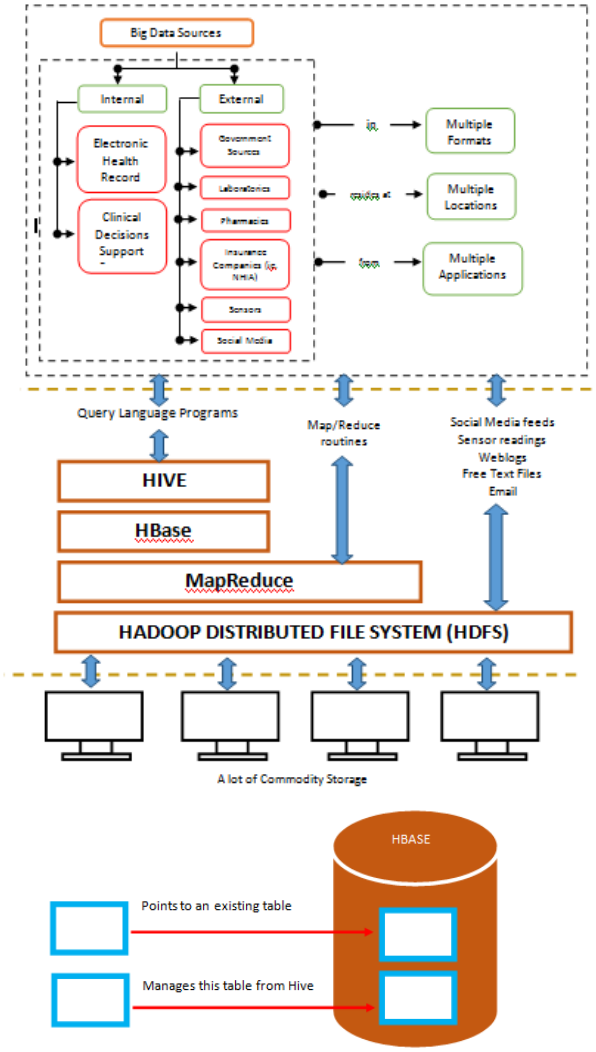

Fig 3.6.1 Hive used on HBase

Hive is normally used on HBase. This is possible because Hive can use tables that already exist in HBase or manage its own ones. Both categories of tables in all instance resides in the HBase.

c. Dealing with data in healthcare

The proposed architecture has the upper layer dealing with the various data sources. It is therefore important to set requirements which will underpin the processes of managing data. The requirement in the upper layer of the architecture is based on the characteristics of Big Data in healthcare. The characteristics as already established is volume, variety, velocity, and veracity. As proposed by Marcus Maier In his MSc thesis, I will integrate his requirement made with the upper layer of my architecture to define the functional and non-functional requirements of the architecture in Healthcare.

The functional requirement can be termed as a specific behaviour of a part of a system while the non-functional requirements system deals the criteria to determine the operations or behaviour of a system. Even though these requirement are supposed to be measurable, it may not be feasible so far as an architecture is concerned.

Requirements for handling growing data volume in healthcare: The goal of the proposed architecture is to handle volume. Handling the growing volume of data in healthcare, there should be an adherence to an ability to store the data. The functional requirement pertaining volume for Big Data in healthcare is Data Storage. The rationale behind this is to be able to store data and data results after analysis. Due to the fact that, healthcare data has different formats and its exponential growth as seen in the proposed architecture (Fig 3.7), it is therefore important to develop non-functional requirement which is scalability and performance relating to the growth of data and timeliness (analysing and getting results directly when data flows)

Requirements handling for data variety in healthcare: As indicated in the proposed architecture, the sources and types of data varies. The architectures proposes the heterogeneity of the sources of data as internal or external and the varying data types. In handing all these varying data types from varying data sources, the main functional requirement will be how the Data will be integrated (a global schema that will that will be able to integrate the span of varying data types and data sources) and data extraction (defines the sources of healthcare data system to acquire data from as defined in the proposed architecture – Fig.3.7 and the format of this data). The non-functional requirement in this domain is Multi-Structured data due to data types and sources of healthcare data.

Requirements handling for data velocity in healthcare: The goal of this requirement of the proposed architecture is to tackle the idea of stream processing. This real time data can be from sources such as Click stream and interactive data from social networks, health plan websites and smart phone applications, Machine to machine data (readings from remote sensors, measuring instruments and other vital signs devices). The challenge is the speed of the incoming. The major requirement in velocity in big data is analysing streaming of Data and the processing the streaming data.

Requirement handling for data veracity: Data is deemed useful if a meaning can be derived from it. Veracity has to deal with the trustworthiness of data and this is very important especially big data analytics important in the health sector. The first requirement is to be sieve between trustworthy and untrustworthy data. This functional requirement has two main sub functional requirement improve Data quality to decrease Uncertainty and tracking trustworthiness within the voluminous data streaming in for healthcare analytics.

These requirements will be handled by the second and third layer of the. The next chapter will now look at how some of these data will be implemented using the Hadoop technology based on the upper layer of the architecture.

4. CONCLUSION

Harnessing the Big Data paradigm in healthcare delivery in Ghana will serve a lot of purposes including;

- Predictive models.
- Analysing disease patterns and tracking disease outbreaks.
- Turning large data into actionable information.
- Evidence based health delivery through data analysis.
- Capture and analyse real time data from variety of locations.

Big Data has come to improve data analytics and deploying it in healthcare will massively improve the healthcare delivery system in Ghana.

5. REFERENCES

[1] World Health Organization, "http://www.who.int/universal_health_coverage/en/," 8 August 2015. [Online]. Available: http://www.who.int.

[2] R. Wullianallur, "Data Mining in Health Care. In health informatics. Improving efficiency and productivity, Edited by Kudyba S. Taylor and Francis," 2014.

[3] Ghana Statistical Service, "National Population Projection 2010 to 2014," Ghana Statistical Service, Accra, 2014.

[4] IBM, "IBM," [Online]. Available: http://www-1.ibm.com/software/data/bigdata/industry-healthcare.html. [Accessed November 2014].

[5] Microsoft, "Microsoft," [Online]. Available: https://msdn.microsoft.com/en-us/library/cc645993.aspx. [Accessed 10 February 2015].

[6] T. Firat and A. John Keane, "Big Data Framework," IEEE Computer Society, pp. 1494-1499, 2013.

[7] S. Pandey and V. Tokekar, "Prominence of MapReduce in Big Data Processing," IEEE Computer Society, pp. 555-560, 2014.

[8] Ghana Broadcasting Corporation , "GBC," [Online]. Available: http://gbcghana.com/1.1922464. [Accessed 22 January 2015].

[9] Minstry of Finance, Ghana, "PIPELINE PROJECTS," Ministry of Finance, Ghana, Accra, 2012.

[10] St. Martin De Porres Hospital, "2010 ANNUAL REPORT," National Catholic Health Services , Agomanya, ER, 2010.

[11] Ghana Health Service , "2010 GHS Facts and Figures," Ghana Health Service, Accra, 2010.

[12] R. Wullianallur and V. Raghupathi, "Big Data analytics in healthcare: promise and potential," Health Information Science and Systems , 2014.

[13] /www.ispor.org, "www.ispor.org," [Online]. Available: http://www.ispor.org/meetings/montreal0614/presentations/africa-forum.pdf. [Accessed 7 June 2015].

[14] N. Patel, P. Narenda, M. Hasan, S. Parth and P. Mayur, "Improving HDFS write performance using efficient relica placement," IEEE, pp. 36-39, 2014.

[15] R. Liu, L. Qicheng, L. Feng, L. Mei and L. J. Lee, "Big Data Architecture for IT Incident Management," IEEE, pp. 424-429, 2014.

[16] D. R. Sara, V. Lopez, M. Jose and F. Herrera, "On the use of MapReduce for imbalanced big data using Random forest," Elsevier - Information Science, pp. 112-137, 2014.

[17] IBM, "Infoshere," 23 June 2015. [Online]. Available: https://www-01.ibm.com/support/knowledgecenter/SSPT3X_3.0.0/com.ibm.swg.im.infosphere.biginsights.jaql.doc/doc/c0057749.html .

[18] apache.org, "Hive," 23 June 2015. [Online]. Available: https://hive.apache.org/.

[19] Y. Demchenko, C. d. Laat and P. Membrey, "Defining Architecture Component of Big Data Ecosystem," IEEE, pp. 104-112, 2014.

[20] G. Katarina, M. Hayes, W. A. Higashino, A. L'Heureux and M. Capretz, "Challenges for MapReduce in Big Data," IEEE Computer Society, pp. 182-188, 2014.

[21] R. Lu, H. Zhu, X. Liu, J. Liu and J. Shao, "Towards efficient and privacy-preserving Computing in Big Data," IEEE Network, pp. 46-50, 2014.

[22] O. Mehmet and B. Mustafa, "Big Data Challeneges in Information Engineeering Curriculum," IEEE, 2014.

[23] P. Zadrozny and R. Kodali, Big Data Analytics Using Splung, Berkley CA, USA: Apress, 2013.

[24] McKinsey GLobal Institue, "Big data: The next frontier," 2011.

[25] T. H. Davenport, "At the Big Data Crossroads:turning towards a smarter travel experience," Amadeus IT Group, 2013.

[26] B. Vorhies, "Data Magnum," [Online]. Available: http://data-magnum.com/how-many-vs-in-big-data-the-characteristics-that-define-big-data. [Accessed 10 May 2015].

[27] R. Lui, Q. Li and J. Lee, "Big Data Architecture for IT Incident Management," IEEE, pp. 424-429, 2014.

[28] M. Wigan and R. Clark, "BIg Data's Big Unintended Consequences," IEEE Computer Society , pp. 46-53, 2013.

[29] apache.org, "cwiki.apache.org," 2013 JUne 2015. [Online]. Available: https://cwiki.apache.org/confluence/display/Hive/Home;jsessionid=19F58B7B013C7CC06A3E5FE23C477B2D .

[30] apache.org, "Hbase," 23 June 2015. [Online]. Available: http://hbase.apache.org/ .

[31] IBM, "Bigsql," 23 June 2015. [Online]. Available: http://www.ibm.com/developerworks/library/bd-bigsql/.

[32] teradata.com, "Big Data Reference Architecture," 23 June 2015. [Online]. Available: http://thinkbig.teradata.com/leading_big_data_technologies/big-data-reference-architecture/.

[33] bigdatawg.nist.gov, "bigdatawg.nist.gov," 23 June 2015. [Online]. Available: http://bigdatawg.nist.gov/_uploadfiles/M0055_v1_7606723276.pdf.

[34] X. Li, F. Zhang and Yongliang, "Resreach on Big Data Architecture, Key Technologies and its Measures," IEEE, pp. 1-4, 2013.

[35] E. Dede and e. al, "A processing pipeline for cassandra Datasets based on Hadoop streaming," IEEE International Congress om Big Data, pp. 168-175, 2014.

[36] Y. Xiaomeng, L. Fangming, L. Jiangchuan and H. Jin, "Building a Network Highway for Big Data: Architecture and Challenges," IEEE Network, pp. 5-12, 2014.

[37] C. Kai-Sang, R. K. MacKinnon and F. Jiang, "Reeducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data," IEEE internation Congress on Big Data, pp. 351-322, 2014.

[38] X. Cheng, C. Hu, Y. Li, W. Lin and H. Zuo, "Data Evolution Analysis of Virtual DataSpace for Managing the Big Data Cycle," IEEE, pp. 2054-2063, 2013.

[39] T.-W. Kim, K.-H. Park, S.-H. Yi and H.-C. Kim, "A Big Data Framwork for u-Healthcare Systems Utilizing Vital Signs," IEEE, pp. 494-497, 2014.

[40] J. Qian, P. Lv, X. Yue, C. Liu and Z. Jing, "Hierarchical attribute reduction algorithm for Big data using MapReduce," ELVEVIER, pp. 1-14, 2014.

[41] L. Gu, D. Zeng, L. Peng and S. Guo, "Cost minimization for Big Data Processing in Geo-Distribution Data Centers," IEEE, pp. 1-10, 2013.

[42] T. Gavrilova and M. Gladkova, "Big Data Structuring: The role of Visaul Models and Ontologies," ELSEVIER, pp. 336-343, 2014.

[43] S. delRio, V. Lopez, J. M. Benitez and F. Herrera, "On the use of MapReduce for imbalance big data using Random Forest," ELSEVIER, pp. 112-137, 2014.

[44] M. Wigan and R. Clarke, "Big Data's Big Unintended Consequences," IEEE, pp. 46-56, 2013.

[45] M. C.Okur and M. Buyukkececi, "Big Data challeneges in Information Engineering Curriculum," IEEE, 2014.

[46] A.Castigilione, M. Gribaudo, M. Iacono and F. Palmieri, "Exploiting mean field to model performance of Big data architectures," ELSEVIER, 2014.

[47] N. Sun, J. Morris, J. Xu, X. Zhu and M. Xie, "iCARE: A framework for big data-based banking customer analytics," IBM Coperation , 2014.

[48] T. White, "Comparism with other System - RDBMS," in Hadoop, The definitive Guide, Califinia , O'Reilly Media Inc., 2011, p. 5.

[49] M. Herland, T. Khosgoftaar and R. Wald, "A review of data mining using big data in health informatics," Journal of Big Data, 2014.

[50] S. O. Fadiya, S. Saydam and V. V. Zira, "Advancing big data for humanitarian needs," Elsevier, pp. 88-95, 2014.

[51] K. Kambatla, G. Kollias, V. Kumar and A. Grama, "Trends in Big Data Analytics," Elsevier, pp. 2561-1573, 2014.

[52] K. Ghani, K. Zheng, J. Wei and C. Friedman, "Harnessing Bog Data for Health Care and Research: Are Urologist Ready," Sciencedirect, pp. 975-977, 2014.

[53] Institute for Health Technology Transformation, "Transforming Healthcare through Big Data," 2013.

[54] B. Vorhies, "Data Magnum," [Online]. Available: http://data-magnum.com/how-many-vs-in-big-data-the-characteristics-that-define-big-data. [Accessed 10 May 2015].