Communications of the IIMA

Volume 13 | Issue 2 Article 1

2013

An Architecture for Big Data Analytics

Joseph O. Chan
Roosevely University

An Architecture for Big Data Analytics

Joseph O. Chan
Roosevelt University, USA
jchan@roosevelt.edu

ABSTRACT

Big Data is the new experience curve in the new economy driven by data with high volume, velocity, variety, and veracity. They come from various sources that include the Internet, mobile devices, social media, geospatial devices, sensors, and other machine-generated data. Unlocking the value of Big Data allows business to better sense and respond to the environment, and is becoming a key to creating competitive advantages in a complex and rapidly changing market. Government is also taking notice of the Big Data phenomenon and has created initiatives to exploit Big Data in many areas such as science and engineering, healthcare and national security. Traditional data processing and analysis of structured data using RDBMS and data warehousing no longer satisfy the challenges of Big Data. Technology trends for Big Data embrace open source software, commodity servers, and massively parallel-distributed processing platforms. Analytics is at the core of exploiting values from Big Data to create consumable insights for business and government. This paper presents architecture for Big Data Analytics and explores Big Data technologies that include NoSQL databases, Hadoop Distributed File System and MapReduce.

Keywords: Big data, big data analytics, NoSQL, Hadoop, distributed file system, MapReduce

INTRODUCTION

Technology has played the roles of enabler and driver in the evolution of the economies spanning the eras characterized by agriculture, manufacturing, service, and knowledge assets. Alongside the change of the economies, technology has evolved across the mainframe computer, the PC, client-server computing, the Internet, cloud computing, mobile computing and social networking. Big Data emerges as the latest stage of the evolution that combines three trends in technology, which Minelli, Chambers, and Dhiraj (2013) described as the three perfect storms: computing, data and convergence. The computing storm results from the exponential growth of processing power as predicted by Moore's Law, mobile computing, social network, and cloud computing. The data storm results from the accessibility of data with high volume, velocity and variety. The convergence storm results from the availability of open-source technology and commodity hardware. Big Data from the technology standpoint are datasets that require beyond the currently available technological capacity. From the business standpoint, they represent a new strategy of creating actionable business insights enabling organizations to sense and respond in a rapidly changing environment.

The impact of Big Data is felt across many sectors and industries. King and Rosenbush (2013) reported that Sears Holding Corporation and Wal-Mart Stores, Inc., use Big Data database for marketing efforts, and Chevron Corporation uses them to process seismic data in the search for new reserves of oil and gas. Winslow (2013) reported the use of Big Data to collect data on the care of hundreds of thousands of cancer patients and use it to help guide treatment of other patients across the healthcare system. Mahrt and Scharkow (2013) described the value of Big Data in digital media research, where the *data rush* through social media promises new insights about consumers' behavior. The impact of real-time social media was felt when the Dow was down over a hundred and forty points in early trading on April 23, 2013 after the Associated Press reported the false tweet about an attack at the White House (Associated Press, 2013). Government has taken notice of Big Data as well and has announced a Big Data Research and Development initiative to improve the ability to extract knowledge and insights from large and complex collections of digital data to help solve some of the Nation's most pressing challenges spanning across concerns in science and engineering, healthcare and national security (Executive Office of the President, 2012). Harbert (2013) postulated that Big Data helps to create big career opportunities that include data scientists, data architects, data visualizers, data change agents, data engineers and operators. Chief executive roles, such as chief data officer and chief analytics officer, also emerge as companies recognize Big Data as an important corporate asset. Analytics that creates consumable business insights is at the heart of exploiting Big Data for business benefits. Traditional technologies for structured data using SQL-based RDBMS and data warehousing are not suitable for Big Data with high volume, velocity and variety. This paper presents the architectural components for Big Data analytics and explores the paradigm shift in Big Data technology towards NoSQL databases, open source software, cheap commodity hardware, and massively parallel computing platforms.

WHAT IS BIG DATA

According to Manuika et al. (2011), Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. The concept of Big Data is therefore relative to the storage and processing capability of prevalent technology of the time. The exponential growth of transistor density was predicted by Intel co-founder Gordon Moore (1965, 1975). Moore's Law states that the number of transistors on a chip doubles about every two years, resulting in the rapidly rising processing power and declining hardware costs. As described in Jacobs (2009), the 1980 US Census data would be considered as Big Data in the 1980s, where the IBM 3850 Mass Storage System with a capacity of 102.2GB was the monster storage device of its day (da Cruz, 1982). This would certainly not be considered as Big Data today where a personal computer can afford such capacity. In today's environment, the size of datasets that may be considered as Big Data could range from terabytes (10^{12} bytes), petabytes (10^{15} bytes), to exabytes (10^{18} bytes), depending on the industry, how data is used and other characteristics. The Gartner Group (Gartner, Inc., 2011) characterized Big Data by the three V's: volume, velocity and variety. Other characteristics such as veracity and value have been added to the definition by other researchers.

Volume

Das and Kumar (2013, p. 153) described that "it took from the dawn of civilization to 2003 to create 5 Exabytes (10^{18} bytes) of information; we now create the same volume in just two days." It is projected that digital data will grow to 8 Zettabytes (10^{21} bytes) by 2015, which is equivalent to 18 million Libraries of Congress. It further described that unstructured data from text, document, image, video, etc. will account for 90% of all data created in the next decade. Big Data comes from different sources that may include transactional enterprise data, machine-generated data, and social data (Oracle, 2012). Enterprise transactional data comes from a variety of applications that include enterprise applications such as ERP, SCM and CRM systems, Web transactions, e-business and m-business transactions. Sathi (2012) described that a communications service provider with 100 million customers would occupy five petabytes of data if stored for 100 days. IBM (2013) provided examples of volume that include turning 12 terabytes of Tweets created each day into improved product sentiment analysis, and converting 350 billion annual meter readings to better predict power consumption.

Machine-generated data contributes to Big Data. They may come from sensors such as environmental and manufacturing sensors, smart meters, smart cards, scanning equipment, and machine-to-machine electronic tenders. Maugh (2009) indicated that 70 millions of CT scans were performed a year, which would generate about 0.1 GB of data per day with 0.5 MB per CT scan image (Strickland, 2004). Aylor (2001) described the launch of the first high-speed satellite seismic data transmission from a vessel off Brazil to the Houston processing center, and indicated that the WesternGeco Patriot seismic vessel using an IBM SPII supercomputer generated seismic data on an average of 93 GB/day. Evans (2012) discussed the explosion of the Big Data set as 40 billion new devices are connected to the Internet in the next few years. A contributing factor to this explosion is the machine-to-machine (M2M) data from remote devices being used more broadly throughout enterprises. Crosby (2008) described the many applications of M2M that include utility companies harvesting energy products using remote sensors to detect important parameters at oil drill sites, traffic control using sensors to monitor traffic volume and speed, telemedicine where patients wear special monitors that gather information about the way their heart is working, and business for tracking inventory and security. A recent development of M2M is the use of sensors by companies to gather real-time information on how teams of employees work and interact (Silverman 2013).

Social data comes from a variety of social media such as Twitter and Facebook. Terdiman (2012) reported that Twitter has hit half a billion tweets a day. André, Bernstein, and Luther (2012) indicated that millions of people read billions of tweets every day. Tam (2013) also reported that Facebook active monthly users have increased to over one billion. These statistics suggest that terabytes of data would be generated daily and petabytes of data would be created, collected, read, interpreted, and responded over a short period of time.

Velocity

The definition of Big Data goes beyond the dimension of volume; it includes the types and frequency of data that are disruptive to traditional database management tools. Minelli et al.

(2013) described velocity as the speed at which data is created, accumulated, ingested, and processed. Sathi (2012) described velocity in terms of throughput and latency. Harris remarked:

> Note that it's not all about size, and that's especially true for the financial markets, where even many years of time series tick data does not come close to the data volumes processed by the likes of Google and Facebook. But what the financial markets might lack in data size, it makes up for in complexity and frequency (2012, para 3).

Social media data streams, such as Twitter data, produce a large influx of data at high frequency, which ensures large volumes, over 8 TB per day (Oracle, 2012). Real-time analysis and response to such data are characteristic of Big Data management in many business situations. IBM (2013) provided examples of velocity that include scrutinizing 5 million trade events created each day to identify potential fraud, and analyzing 500 million daily call detail records in real-time to predict customer churn faster. Maddox (2012), indicated a "10 millisecond in the spot demonstrates just how fast technology is able to provide feedback to its advertisers about online audience behavior" (para. 8). Marketers use real-time social media, Web browsing and transactional data to tailor real-time responses to target individuals and segments. Companies monitor their internal operations and environments in real-time and generate real-time responses. Big Data with high velocity has created opportunities and requirements for organizations to increase the capability of real-time sense and response.

Variety

Data source has dramatically increased from traditional transactional processing to others that include Internet data (e.g., clickstream and social media), research data (e.g., surveys and industry reports), location data (e.g., mobile device data and geospatial data), images (e.g., surveillance and satellites), supply chain data (e.g., EDI, vendor catalogs), and devices data (e.g., sensors and RFID devices), (Minelli et al., 2013). Big Data concerns structured, semi-structured and unstructured data. Structured data are data with a formal structure of data models such as fixed fields in a record or a file, or columns and rows in a relational table. Unstructured data are data with no identifiable formal structure of data models. They include unstructured text in documents, emails and blogs, PDF files, audio, video, images, click streams and Web contents. Semi-structured data type does not fit into a formal structure of data models, but it contains tags that separate semantic elements, which includes the capability to enforce hierarchies within the data (Minelli et al., 2013). Data variety contributes to the complexity of capturing, storing, processing and performing analytics of Big Data. IBM (2013) provided examples of variety that include monitoring hundreds of live video feeds from surveillance cameras to target points of interest, and exploiting the 80% data growth in images, video and documents to improve customer satisfaction.

Veracity

As Big Data comes from different sources outside the control of the firm, accuracy and completeness of data become a concern. Sathi (2012) described veracity as representing both the credibility of the data source as well as the suitability of data for the target audience. Establishing

trust in Bid Data presents a huge challenge as the variety and number of sources grows (IBM 2013).

Value

Oracle (2013a) addressed the commercial value that any new source and forms of data can add to the business, and to what extent they can be predicted so that the ROI can be calculated and projected budget acquired. The value dimension adds the business perspective to Big Data in addition to technical factors in volume, velocity, variety and veracity. Assessing the value of Big Data in its alignment of business objectives is essential in its implementation.

AN ACHITECTURE FOR BIG DATA

In the economy fueled by the Internet and globalization, and amplified by social networks and mobile computing, Big Data is becoming an enterprise concern. The capability to capture, process and analyze Big Data can provide tremendous competitive advantage by increasing the firm's capability to respond to the dynamic market conditions and customer needs. In the following, a client-server architecture for Big Data is presented (Figure 1).

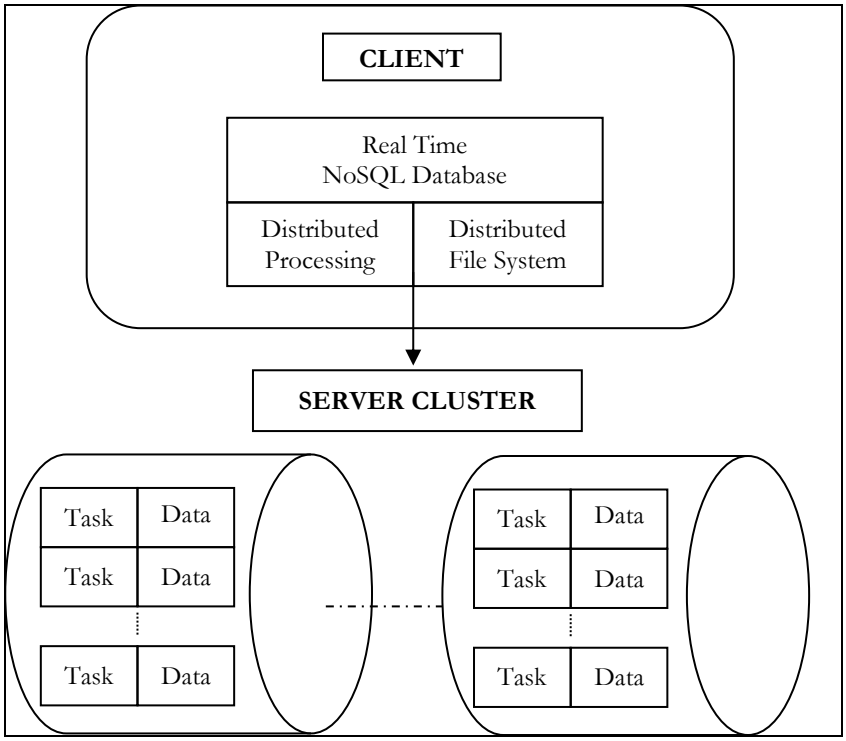

Figure 1: A Conceptual Cluster Architecture for Big Data.

The Client Level Architecture

The client level architecture consists of NoSQL databases, distributed file systems and a distributed processing framework. NoSQL is commonly interpreted as "Not Only SQL." NoSQL

databases are non-relational, non-SQL based, and store data in key-value pairs, which work well with unrelated data (Minelli, 2013). NoSQL databases provide distributed, highly scalable data storage for Big Data. Yen (2009) presented a *NoSQL taxonomy that includes* key-value-cache, key-value-store, eventually-consistent key-value-store, ordered-key-value-store, data-structures server, tuple-store, object database, document store, and wide columnar store. NoSQL (n.d.) listed 150 examples of NoSQL databases by different categories. A popular example of NoSQL database is Apache Hbase. According to Apache (2013b), Apache Hbase is an open-source, distributed, versioned, column-oriented store that provides random, real-time read/write access to Big Data. Oracle (2013b) described the Oracle NoSQL Database as a distributed key-value database designed to provide **highly reliable, scalable and available** data storage across a configurable set of systems that function as storage nodes.

The next layers consist of the distributed file system that is scalable and can handle a large volume of data, and a distributed processing framework that distributes computations over large server clusters. Tantisiriroj, Patil, and Gibson (2008) described the Internet services file systems to include Google file system, Amazon Simple Storage Service and the open-source Hadoop distributed files system. A popular platform is the Apache Hadoop. According to Apache (2013a), Apache Hadoop is a framework for distributed processing of large data sets across clusters of computers, and is designed to scale up from a few servers to thousands of machines, each offering local computation and storage. The two critical components for Hadoop are the Hadoop distributed file system (HDFS) and MapReduce (Minelli, et al., 2013). HDFS is the storage system and distributes data files over large server clusters and provides high-throughput access to large data sets. MapReduce is the distributed processing framework for parallel processing of large data sets. It distributes computing jobs to each server in the cluster and collects the results.

The Server Level Architecture

The server level architecture for Big Data consists of parallel computing platforms that can handle the associated volume and speed. Minelli et al. (2013) described three prominent parallel computing options: clusters or grids, massively parallel processing (MPP), and high performance computing (HPC). According to Buyya, Yeo, Venugopal, Broberg and Brandic (2009), clusters or grids are types of parallel and distributed systems, where a cluster consists of a collection of inter-connected stand-alone computers working together as a single integrated computing resource, and a grid enables the sharing, selection, and aggregation of geographically distributed autonomous resources dynamically at runtime.

A commonly used architecture for Hadoop consists of client machines and clusters of loosely coupled commodity servers that serve as the HDFS distributed data storage and MapReduce distributed data processing. Hedlund (2011) described the three major categories of machine roles in a Hadoop deployment that consist of Client machines, Master nodes and Slave nodes. The role of the Client machine is to load data into the cluster, submit MapReduce jobs, and retrieve results of the job when it is finished (Hedlund 2011). There are two types of Master nodes, the HDFS nodes and the MapReduce nodes. The HDFS nodes consist of Name Nodes, which keep the directory of all files in the HDFS file system. Client applications submit jobs to MapReduce nodes, which consist of Job Trackers that assign MapReduce tasks to slave nodes.

The Job Tracker consults with the Name Node to determine the location of the Data Node where the data resides, and assigns the task to the Task Tracker that resides in the same node, which can execute the task. While HDFS is a distributed file system that is well suited for the storage of large files, it does not provide fast individual record lookups; whereas, HBase, built on top of HDFS provides fast record lookups and updates (Apache 2013c). Apache HBase provides random, real-time read/write access to Big Data (Apache 2013b). Lipcon (2012) described that HDFS was originally designed for high-latency high-throughput batch analytic systems like MapReduce, and that HBase improved its suitability for real-time systems low-latency performance. In this architecture, the Hadoop HDFS provides a fault-tolerant and scalable distributed data storage for Big Data, Hadoop MapReduce provides the fault-tolerant distributed processing over large data sets across the Hadoop cluster, and HBase provides the real-time random access to Big Data. Figure 2 illustrates the Hadoop architecture using HBase, HDFS and MapReduce.

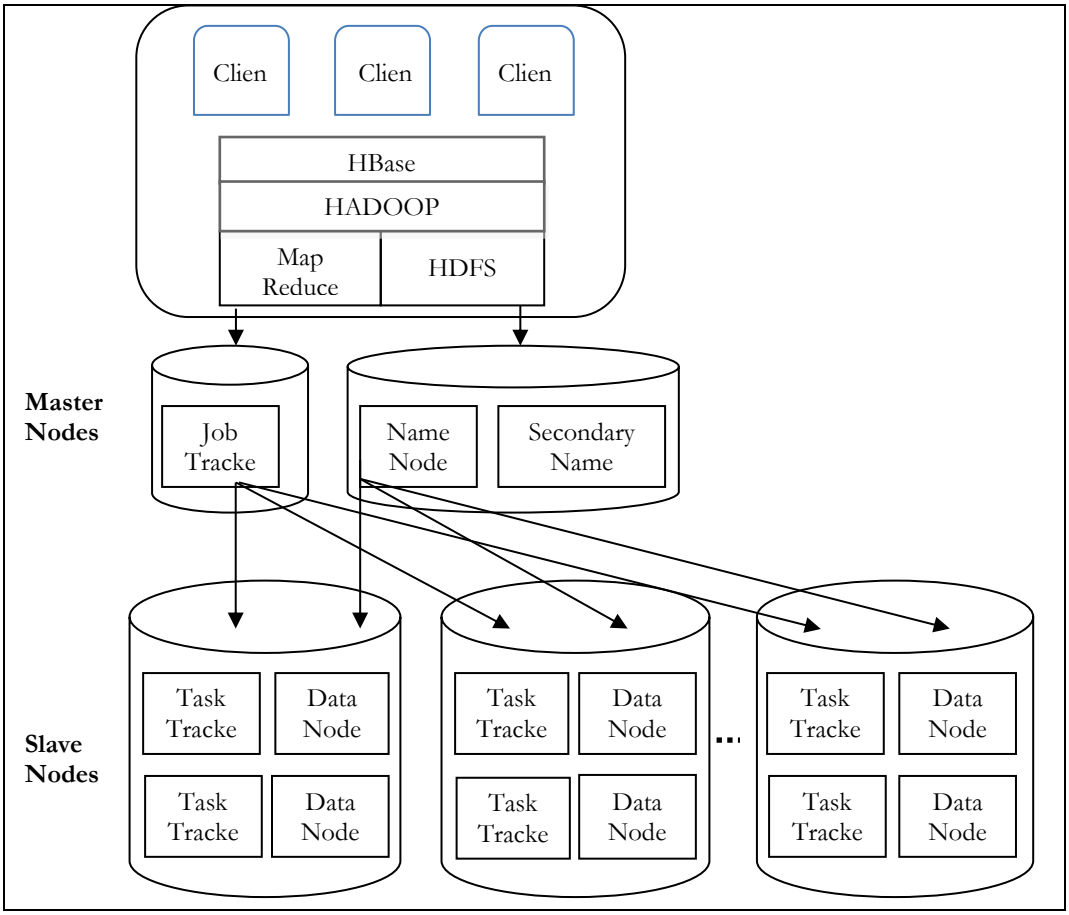

Figure 2: HBase/Hadoop Cluster Architecture for Big Data.

Kim, Raman, Liu, Lee, and August (2010) pointed out that while clusters of commodity servers are the most popular form of large-scale parallel computers, they might not be suitable for highly inter-node dependent general-purpose applications programs. It proposed a runtime system distributed software Multi-threaded Transactional memory (DSMTX) to address inter-node communication costs. Another option for parallel computing platform is MPP (Massively

Parallel Processing). Minelli et al. (2013) described MPP as combining storage, memory, and computation to create a platform. While nodes in a cluster network are mainly independent, the nodes in MPP are tightly interconnected by dedicated high-speed networks, allowing the high-speed collaboration between processors. Schulmeister (2005) described the Taradata MPP architecture as a collection of SMP (symmetric multi-processing) nodes, each consisting of dual processors and memory, connected by BYNET Interconnect.

BIG DATA ANALYTICS

Chen, Chiang, and Storey (2012) provided a classification of business intelligence and analytics (BI&A) into three categories. BI&A 1.0 is characterized by DBMS-based and structured content. It utilizes traditional analytic tools via data warehousing, ETL, OLAP and data mining. BI&A 2.0 is characterized by Web-based and unstructured content. It utilizes tools in information retrieval, opinion mining, question answering, Web analytics, social media analytics, social network analysis, and spatial-temporal analysis. BI&A 3.0 is characterized by mobile and sensor-based content. It utilizes tools in location-awareness analysis, person-centered analysis, context-relevant analysis, and mobile visualization and HCI. BI&A 2.0 and 3.0 would require a platform that can handle the huge volume, velocity and variety of data. The Big Data analytics architecture described below utilizes the massively parallel, distributed storage and processing framework as provided by Hadoop HDFS and MapReduce.

As opposed to some belief that Big Data has pronounced the obsolescence of data warehousing, it remains a viable technology for Big Data analytics of huge volume of structured data. Furthermore, there is synergy between data warehousing and the Hapdoop type Big Data architecture. Unstructured data from sensors, M2M devices, social media and Web applications can be stored in Hadoop and be MapReduced later for meaningful insight (Sathi, 2012). MapReduced data can then be integrated with the data warehouse for further analytic processing. Conversely, data warehouse can be a data source for complex Hadoop jobs, simultaneously leveraging the massively parallel capabilities of two systems (Awadallah & Graham, 2011). Real-time location data from GPS or smartphones can be combined with historic data from the data warehouse to provide real-time insight for marketers to promote products targeted to the individual customer based on real-time location data and customer profile. Figure 3 illustrates an architecture for Big Data analytics.

Structured data are captured through various data sources including OLTP systems, legacy systems and external systems. It goes through the ETL process from the source systems to the target data warehouse. Traditional business intelligence (BI) batched analytical processing tools such as online analytical processing (OLAP), data mining, and query and reporting, can be used to create the business intelligence to enhance business operations and decision processes. Unstructured and semi-structured Big Data sources can be of a wide variety that includes data from clickstreams, social media, machine-to-machine, mobile device, sensors, documents and reports, Web logs, call records, scientific research, satellites, and geospatial devices. They are loaded into the Hadoop Distributed File System cluster. Hadoop MapReduce provides the fault-tolerant distributed processing framework across the Hadoop cluster, where batched analytics

can be performed. Actionable insight resulting from Hadoop MapReduce analytics and business intelligence analytics can be consumed by operational and analytical applications.

While Hadoop is highly scalable and can perform massively parallel computing for Big Data, it is a batch system with high latency, and would not be suitable for processing of real-time events. Minelli et al. (2013) described geospatial intelligence as using data about space and time to improve the quality of predictive analysis. For example, real-time recommendations of places of interest can be based on the real-time location from smartphone usage. This real-time information can be combined with batched analytics to improve the quality of the predictions. Other examples of real-time analytic applications include real-time trending of social media data, real-time Web click stream analysis, algorithmic trading, and real-time M2M analysis. Real-time NoSQL databases such as HBase can be used in conjunction with Hadoop to provide real-time read/write of Hadoop data. Emerging technologies for Big Data real-time analytics include technologies for collection and aggregation of real-time data for Hadoop, in-memory analytic systems, and real-time analytics applications for processing of data stored in Hadoop. Real-time insight created by real-time analytics can be consumed by real-time operations and decision processes.

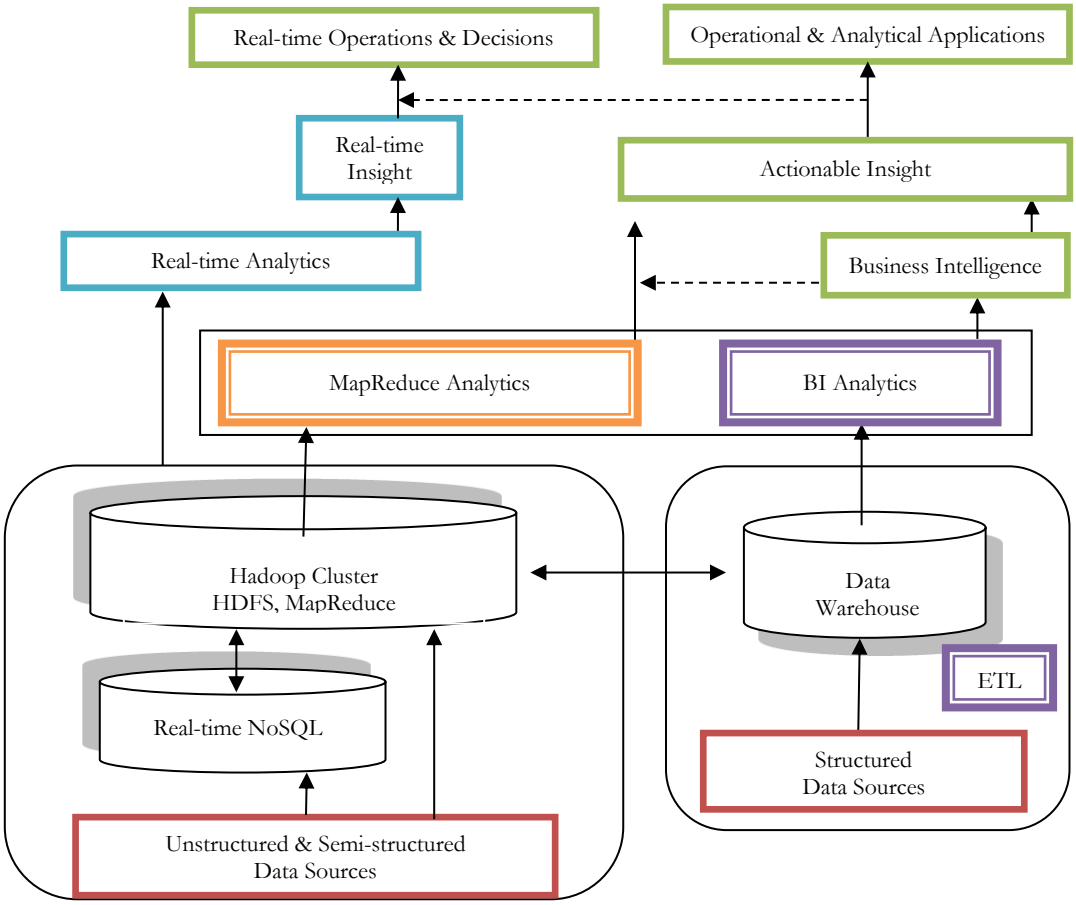

Figure 3: An Architecture for Big Data Analytics

CONCLUSIONS

The arrival of Big Data in society has prompted business and government to take actions to exploit its value and application. This paper described the characteristics of Big Data and presented an architecture for Big Data analytics. Big Data technology deviates from traditional data management SQL-based RDBMS approaches as it deals with data with high volume, velocity and variety. The new paradigm moves towards NoSQL databases, massively parallel and scalable computing platforms, open-source software, and commodity servers. This paper discussed an architecture using real-time NoSQL databases, the Hadoop HDFS distributed data storage and MapReduce distributed data processing over a cluster of commodity servers. It discussed running batched and real-time analytics on the Hadoop platform. Its bidirectional relationship with traditional data warehouse and data mining analytics platform was described. The practical significance of the architecture is to provide a blueprint for organizations moving towards the implementation of Big Data in their enterprise. The capability for organizations to collect and process Big Data about individuals or groups causes various privacy concerns. Further study could address the ethical issues that may arise from Big Data and the measures that society can take to mitigate such concerns.

REFERENCES

André, P., Bernstein, M. S., & Luther, K. (2012). Who gives a tweet? Evaluating microblog content value. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW12, 2012)*, 471-474. doi: http://dx.doi.org/10.1145/2145204.2145277

Apache Software Foundation. (2013a). *Welcome to Apache Hadoop*. Retrieved from http://hadoop.apache.org/

Apache Software Foundation. (2013b). *Welcome to Apache HBase*. Retrieved from http://hbase.apache.org/

Apache Software Foundation. (2013c). *Architecture overview: What is the difference between HBase and Hadoop/HDFS?* Retrieved from http://hbase.apache.org/book/architecture.html#arch.overview

Associated Press. (2013, April 23). Hackers compromise AP Twitter account. *The Wall Street Journal*, U.S. ed.

Awadallah, A., & Graham, D. (2011). *Hadoop and the data warehouse: When to use which.* Dayton, OH: Teradata Corporation. Retrieved from http://www.teradata.com/white-papers/Hadoop-and-the-Data-Warehouse-When-to-Use-Which/

Aylor, W. K. (2001). Geology & geophysics: Seismic milestone achieved with high-speed satellite data movement. *Offshore, 61*(8). Retrieved from http://www.offshore-mag.com/articles/print/volume-61/issue-8/news/geology-geophysics-seismic-milestone-achieved-with-high-speed-satellite-data-movement.html

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems, 25(6)*, 599-616. doi: 10.1016/j.future.2008.12.001

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*(4), 1165-1188.

Crosby, T. (2008, April 24). How machine-to-machine communication works. *HowStuffWorks.* Retrieved from http://computer.howstuffworks.com/m2m-communication.htm

da Cruz, F. (1982). The IBM 3850 mass storage system. *Columbia University Computing history: A chronology of computing at Columbia University.* Retrieved from http://www.columbia.edu/cu/computinghistory/mss.html

Das, T. K., & Kumar, P. M. (2013). Big data analytics: A framework for unstructured data analysis. *International Journal of Engineering and Technology, 5*(1), 153-156. Retrieved from http://www.enggjournals.com/ijet/docs/IJET13-05-01-081.pdf

Evans, B. (2012, November 6). Big data set to explode as 40 billion new devices connect to internet. *Forbes*. Retrieved from http://www.forbes.com/sites/oracle/2012/11/06/big-data-set-to-explode-as-40-billion-new-devices-connect-to-internet/

Executive Office of the President, Office of Science and Technology Policy. (2012, March 29). *Obama administration unveils "big data" initiative: Announces $200 million in new R&D investments* [Press release]. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf

Gartner, Inc. (2011, June 27). *Gartner says solving 'big data' challenge involves more than just managing volumes of data* [Press release]. Retrieved from http://www.gartner.com/newsroom/id/1731916

Harbert, T. (2012, September 20). Big data big jobs? *Computerworld*. Retrieved from http://www.computerworld.com/s/article/9231445/Big_data_big_jobs_?pageNumber=1

Harris, P. (2012). *Big data: The other side of low latency* [web log]. Retrieved from http://low-latency.com/blog/big-data-other-side-low-latency

Hedlund, B. (2011). *Understanding Hadoop clusters and the network*. Retrieved from http://bradhedlund.com/2011/09/10/understanding-hadoop-clusters-and-the-network/

IBM. (2013). *Bringing big data to the enterprise*. Retrieved from: http://www-01.ibm.com/software/data/bigdata/

Jacobs, A. (2009). The pathologies of dig data. *Communications of the ACM, 52*(8), 36-44. doi: 10.1145/1536616.1536632

Kim, H., Raman, A., Liu, F., Lee, J., & August, D. I. (2010). Scalable speculative parallelization on commodity clusters. *Proceedings of the 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO '43)*, 3-14. doi: 10.1109/MICRO.2010.19

King, R., & Rosenbush, S. (2013, March 13). Big data broadens its range. *The Wall Street Journal, business technology ed*.

Lipcon, T. (2012). *HBase and HDFS: Past, present, future* [Slideshow]. Retrieved from http://www.slideshare.net/cloudera/1-todd-lipcon-past-present-futurepdf

Maddox, K. (2012, July 16). Turn ad inspired by 'Mad Men': Cloud computing company debuts spot that mimics show during season finale. *B2B.* Retrieved from http://www.btobonline.com/apps/pbcs.dll/article?AID=/20120716/ADVERTISING01/307169932/0/SEARCH

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media, 57*(1), 20-33. doi: 10.1080/08838151.2012.761700

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. New York, NY: McKinsey Global Institute.

Maugh, T. H., II. (2009, December 15). Overuse of CT scans will lead to new cancer deaths, a study shows. *Los Angeles Times*. Retrieved from http://articles.latimes.com/2009/dec/15/science/la-sci-ct-scans15-2009dec15

Minelli, M., Chambers, M., & Dhiraj, A. (2013). *Big data, big analytics: Emerging business intelligence and analytic trends for today's businesses.* Hoboken, NJ: John Wiley & Sons, Inc.

Moore, G. E. (1965, April). Cramming more components onto integrated circuits. *Electronics, 114-117*

Moore, G. E. (1975). Progress in digital integrated electronics. *1975 International Electron Devices Meeting, 21*, 11-13.

NoSQL. (n.d.). NoSQL definition. *NoSQL Archive*. Retrieved from http://nosql-database.org/

Oracle. (2012). *Oracle: Big data for the enterprise* [White paper]. Retrieved from http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf

Oracle. (2013a). *Information management and big data: A reference architecture* [White paper]. http://www.oracle.com/technetwork/topics/entarch/articles/info-mgmt-big-data-ref-arch-1902853.pdf

Oracle. (2013b). *Oracle NoSQL database*. Retrieved from http://www.oracle.com/technetwork/products/nosqldb/overview/index.html

Sathi, A. (2012). *Big data analytics: Disruptive technologies for changing the game.* Boise, ID: MC Press.

Schulmeister, B. (2005). *Teradata: Architecture, technology, scalability, performance and vision for active enterprise data warehousing* [PowerPoint slides]. Retrieved from http://www.informatik.hu-berlin.de/forschung/gebiete/wbi/teaching/archive/sose05/dbs2/slides/teradata_060628.pdf

Silverman, E. (2013, March 7). Tracking sensors invade the workplace: Devices on workers, furniture offer clues for boosting productivity. *The Wall Street Journal.* Retrieved from http://online.wsj.com/article/SB10001424127887324034804578344303429080678.html

Strickland, N. H. (2004). Multidetector CT: What do we do with all the images generated? *The British Journal of Radiology, 77*(1), S14-S19. Retrieved from http://bjr.birjournals.org/content/77/suppl_1/S14.long . doi: 10.1259/bjr/95034282.

Tam, D. (2013, January 30). Facebook by the numbers: 1.06 billion monthly active users. *CNET News.* Retrieved from http://news.cnet.com/8301-1023_3-57566550-93/facebook-by-the-numbers-1.06-billion-monthly-active-users/

Tantisiriroj, W., Patil, S., & Gibson, G. (2008). *Data intensive file systems for internet services: A rose by any other name*. Pittsburgh, PA: Parallel Data Laboratory, Carnegie Mellon University. Retrieved from http://www.pdl.cs.cmu.edu/PDL-FTP/PDSI/CMU-PDL-08-114.pdf

Terdiman, D. (2012, October 26). Report: Twitter hits half a billion tweets a day. *CNET News*. Retrieved from http://news.cnet.com/8301-1023_3-57541566-93/report-twitter-hits-half-a-billion-tweets-a-day/

Yen, S. (2009, October 1). *NoSQL is a horseless carriage*. Retrieved from http://dl.getdropbox.com/u/2075876/nosql-steve-yen.pdf

Winslow, R. (2013, March 26). Big data for cancer care. *The Wall Street Journal*. Retrieved from http://online.wsj.com/article/SB10001424127887323466204578384732911187000.html

This Page Was Left Blank Intentionally.