The emergence of "big data" technology and analytics

Bernice Purcell
Holy Family University

ABSTRACT

The Internet has made new sources of vast amount of data available to business executives. Big data is comprised of datasets too large to be handled by traditional database systems. To remain competitive business executives need to adopt the new technologies and techniques emerging due to big data.

Big data includes structured data, semistructured and unstructured data. Structured data are those data formatted for use in a database management system. Semistructured and unstructured data include all types of unformatted data including multimedia and social media content. Big data are also provided by myriad hardware objects, including sensors and actuators embedded in physical objects, which are termed the Internet of Things.

Data storage techniques used for big data include multiple clustered network-attached storage (NAS) and object-based storage. Clustered NAS employs storage devices attached to a network. Groups of storage devices attached to different networks are then clustered together. Object-based storage systems distribute sets of objects over a distributed storage system.

Hadoop, used to process unstructured and semistructured big data, uses the map-reduce paradigm to locate all relevant data then select only the data directly answering the query. NoSQL, MongoDB, and TerraStore process structured big data. NoSQL data is characterized by being basically available, soft state (changeable), and eventually consistent. MongoDB and TerraStore are both NoSQL-related products used for document-oriented applications.

The advent of the age of big data poses opportunities and challenges for businesses. Previously unavailable forms of data can now be saved, retrieved, and processed. However, changes to hardware, software, and data processing techniques are necessary to employ this new paradigm.

Keywords: big data, scale-out network attached storage, data analytics, Hadoop, NoSQL

BIG DATA IMPACTS BUSINESS ENTERPRISES

Data are generated in a growing number of ways. Use of traditional transactional databases has been supplemented by multimedia content, social media, and myriad types of sensors (Manyika et al., 2011). Advances in information technology allow users to capture, communicate, aggregate, store and analyze enormous pools of data, known as "big data" (Manyika et al., 2011). However, the new data collection methodologies pose a dilemma for businesses that have depended upon database technology to store and process data.

"Big data" derives its name from the fact that the datasets are large enough that typical database systems are unable to capture, save, and analyze these datasets (Manyika et al., 2011). The actual size of big data varies by business sector, software tools available in the sector, and average dataset sizes within the sector (Manyika et al., 2011). Best estimates of size range from a few dozen terabytes to many petabytes (Manyiak et al., 2011).

In order to benefit from big data, new storage technologies and analysis methods need to be adopted. Business executives must determine the new technologies and methodologies best suited to their information needs. Business executives ignoring the growing field of big data will eventually become non-competitive.

TYPES AND SOURCES OF BIG DATA

Executives need to be cognizant of the types of data they need to deal with. There are three main types of data, regardless of whether or not a company is using big data – unstructured data, structured data, and semistructured data. Unstructured data are data in the format in which they were collected; no formatting is used (Coronel, Morris, & Rob, 2013). Some examples of unstructured data are PDF's, e-mails, and documents (Baltzan, 2012). Structured data are formatted to allow storage, use, and generation of information (Coronel, Morris, & Rob, 2013). Traditional transactional databases store structured data (Manyika et al., 2011). Semistructured data have been processed to some extent (Coronel, Morris, & Rob, 2013). XML or HTML-tagged text are examples of semistructured data (Manyika et al., 2011). Business executives with traditional database management systems need to broaden their data horizons to include collection, storage, and processing of unstructed and semistructured data

Data collection of unstructured and semistructured data is done through several internet-based technologies. Chui, Löffler, and Roberts (2010) describe sensors providing big data as being part of the Internet of Things. The Internet of Things is described as sensors and actuators that are embedded in physical objects that provide data through wired and wireless networks (Chui, Löffler, & Roberts, 2010). Some industries that are creating and using big data are those that have recently begun digitization of their data content; these industries include entertainment, healthcare, life sciences, video surveillance, transportation, logistics, retail, utilities, and telecommunications (Chui, Löffler, & Roberts, 2010). Devices generating data in these

industries include IPTV cameras, GPS transceiver, RFID tag readers, smart meters, and cell phones (Chui, Löffler, & Roberts, 2010).

BIG DATA STORAGE TECHNOLOGIES

The ability to store massive amounts of data is a necessity for business executives to use big data. Two major means of storing big data are clustered network-attached storage (NAS), also called scale-out NAS, and object-based storage systems (Sliwa, 2011). Without a change to data storage technology, executives will not be able to collect big data.

Scale-out NAS is built upon a traditional NAS system. NAS is a storage device that is based on a computer with no keyboard or mouse; this computer only serves as a device to retrieve data for users (White, 2011). To support the demands of big data, several NAS devices are connected, or clustered, and each NAS device can search through devices attached to the other NAS devices.

As indicated in Figure 1 (Appendix), each NAS is attached to several storage devices, which the NAS is able to search. In turn this "NAS pod" is connected by a switch to another "NAS pod" which does the same function. Because the pods are connected through the switch, both pods can be searched for data by any client. Clients may be directly connected on a local network, a VPN, or somewhere on the cloud attached through a network.

In object-based storage systems, users deal not with files but with sets of objects which are distributed over several devices (Wang, Brandt, Miller, & Long, 2004). Object-based storage systems provide high capacity and throughput as well as reliability and scalability, which are all needed for big data storage (Wang, Brandt, Miller, & Long, 2004). It is the layout of the objects themselves is what provides the efficiency of the storage and searching, rather than the configuration of the storage system as in scale-out NAS.

BIG DATA ANALYTICS

Storing big data is only part of the picture. Special techniques are needed to analyze big data. Executives need to become familiar with the big data methodologies, adopt the technology appropriate for their business, and ensure that employees develop skill with the technology.

Data storage techniques differ depending on whether the data are unstructured or structured. Unstructured and semistructured data can be analyzed using software like Hadoop. Users analyzing structured big data can use software such as NoSQL, MongoDB, and TerraStore.

Hadoop is based on a programming paradigm called MapReduce, as discussed in Google's 2004 paper on Hadoop (Eaton, Deroos, Deutsch, Lapis, & Zikopoulos, 2012). The name MapReduce comes from the two distinct tasks that the Hadoop program will perform using key-value pairs when a query is made (Eaton, Deroos, Deutsch, Lapis, & Zikopoulos, 2012). The mapping task is given a piece of data known as a key to search on, finds relevant values based on this key, and converts the key and values into another dataset query (Eaton, Deroos,

Deutsch, Lapis, & Zikopoulos, 2012). The reducing task takes the final resultant output (the key and value combinations) from the mapping and reduces the output into a small dataset which answers the query (Eaton, Deroos, Deutsch, Lapis, & Zikopoulos, 2012). Hadoop works well in a scale-out NAS environment. The mapping task will search all possible datasets for the data being queried. Due to the size of the environment, this will produce a huge dataset for the output. The reduce task will analyze the dataset output from mapping and check that only data the directly answers the query is returned. For example, if the user queries the system for the highest sales amount for each of four sales people, the map task will search the system for all sales for the four sales people, and the reduce task will limit the output to the highest sales amount for each sales person. Researchers from Techaisle found that 73% of businesses in their study preferred using Hadoop because of its capability to process large volumes of big data (Business & Finance Week editors, 2012).

Due to the volume of data stored, structured data can also be considered big data depending upon how it is stored (scale-out NAS or object-based storage). There are several different software options commonly used to analyze structured big data. NoSQL, which can mean either 'no SQL' or 'not only SQL,' is characterized by data that is Basically Available, Soft state, and Eventually consistent (BASE), rather than the traditional database data characteristics of Atomicity, Consistency, Isolation, and Durability (ACID) (Oracle, 2011). Data analyzed using NoSQL, therefore, is at times in a state of transition and may not be directly available; the data is in flux rather than set as in traditional database environments. MongoDb and TerraStore are both NoSQL-related products that are used for "document-oriented applications" such as storage and searching of whole invoices rather than the individual data fields from the invoice (Sasirekha, 2011).

THE IMPORTANCE OF BIG DATA TO THE BUSINESS WORLD

The importance of big data to business executives is derived from the data collected. Previously, executives relied solely on structured data collected and stored in a traditional database. Data collected from social media and the Internet of Things provides unstructured data that is constantly updated (Chui, Löffler, & Roberts, 2010). Analysis of these data will provide new information for executives that will enable them to maintain a competitive stance in their business environment. Thirty-four percent of business executives currently using business intelligence plan to employ big data analytics (Business & Finance Week editors, 2012).

Manyika et al. (2011) propose five major contributions big data can make to businesses: 1) transparency creation, 2) performance improvement, 3) population segmentation, 4) decision making support, and 5) innovative business models, products, and services. Creating data transparency within a business enables data to be shared more easily among departments. For example, data from research and development, engineering, and manufacturing units within a business can be integrated to enable concurrent product engineering, reducing time to market and improving quality (Manyika et al., 2011). Big data can provide more accurate and detailed

performance data in real-time or near real-time, allowing managers to analyze performance variability and understand causes of the variability (Manyika et al., 2011). While market segmentation has been used for years, big data can provide highly specific segmentations enabling production of tailored products and services (Manyika et al., 2011). Increasingly sophisticated analytics can be employed using big data to support decision makers in minimizing risks and finding new insights, thus improving the decision making process (Manyika et al., 2011). New products, services, and even business models can emerge from analysis of big data (Manyika et al., 2011). One example is use of real-time location-based data enabling property and causality insurance adjusters to price policies based on where and how people drive (Manyika et al., 2011).

CONCLUSION

Big data poses opportunities and challenges for businesses. Previously untapped sources of data are able to be stored and processed. Unstructured data previously available, such as invoice data, can be stored in a new, more convenient and meaningful format, and can employ text searching techniques. Data analytics will supplant the use of only structured queries of relational database management system. Benefits of big data use to business executives include enhanced data sharing through transparency, improved performance though analysis, augmented market segmentation, increased decision support through advanced analytics, and greater ability to innovate products, services and business models. Business owners need to follow trends in big data carefully to make the decision that fits their businesses.

REFERENCES

Baltzan, P. (2012). *Business driven information systems*, (3rd ed.). New York: McGraw-Hill.

Business & Finance Week Editors. (2012, 12 May). "Data analytics: 34 percent of mid-market businesses using business intelligence are planning to adopt big data analytics; Lack of expertise among SMBs is main barrier." *Business & Finance Week.* Retrieved from http://search.proquest.com.proxy1.ncu.edu/docview/1010520318?accountid=28180

Chui, M., Löffler, M., & Roberts, R. (2010, March). "The Internet of things." *McKinsey Quarterly*. Retrieved from https://www.mckinseyquarterly.com /The_Internet_of_Things_2538

Coronel, C., Morris, S., & Rob, P. (2013). *Database Systems: Design, Implementation, and Management*, (10th Ed.). Boston: Cengage Learning.

Eaton, Deroos, Deutsch, Lapis, & Zikopoulos. (2012). *Understanding big data: Analytics for enterprise class Hadoop and streaming data.* New York: McGraw-Hill.

Manyika, J., Chui,. M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011, June). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. Retrieved from http://www.mckinsey.com/Insights /MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_ innovation

Oracle. (2011, September). Oracle NoSQL Database. Redwood Shores, CA: Oracle Corporation. Retrieved from http://www.oracle.com/webapps/dialogue/ns/dlgwelcome.jsp?p_ext=Y&p_dlg_id=11739928&src=7328025&Act=24&sckw=WWMK11054205MPP001.GCM.8318.1020

Sasirekha, R. (2011). *NoSQL, the database for the cloud*. New York: Tata Consultancy Service. Retrieved from http://www.tcs.com/SiteCollectionDocuments/White%20Papers/Consulting_Whitepaper_No-SQL-Database-For-The-Cloud_04_2011.pdf

Sliwa, C. (2011, June 16). Scale-out NAS, object storage, cloud gateways replacing traditional NAS. Retrieved from http://searchstorage.techtarget.com/feature/Scale-out-NAS-object-storage-cloud-gateways-replacing-file-storage

Wang, Brandt, Miller, & Long. (2004, April). OBFS: A file system for object-based storage devices. *Design* (2004), 283 – 300. Retrieved from http://www.mendeley.com/research/obfs-a-file-system-for-objectbased-storage-devices/

White, C. (2011). *Data Communications and Computer Networks: A business user's approach*, (6th ed.). Boston: Cengage Learning.

APPENDIX

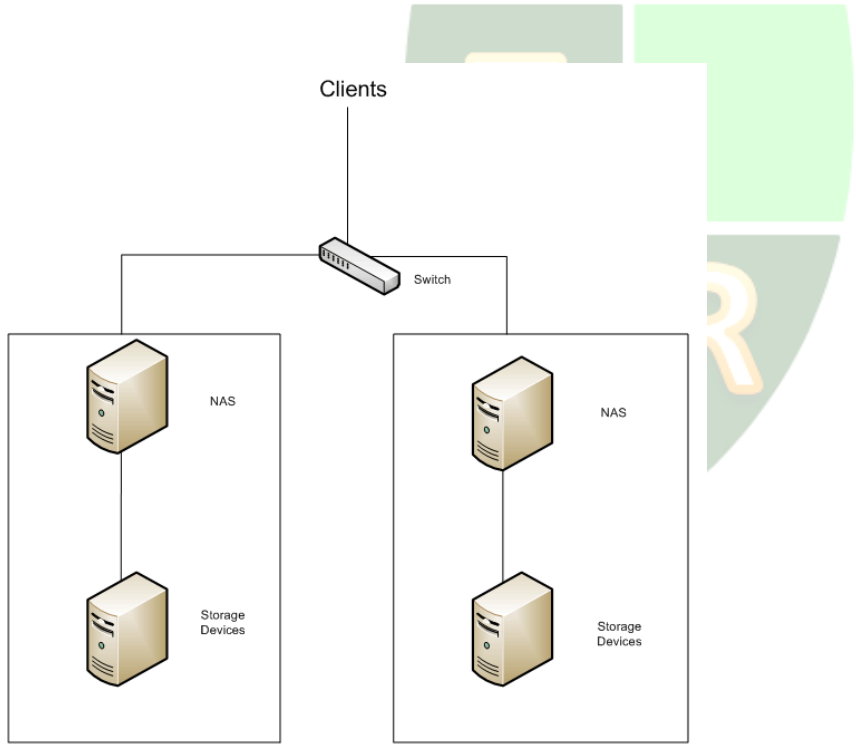

Figure 1