Scientific Research Publishing

Journal of Computer and Communications, 2017, 5, 40-48
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Big Data for Organizations: A Review

Pwint Phyu Khine[1], Wang Zhao Shun[1,2]

[1]School of Information and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing, China
[2]Beijing Key Laboratory of Knowledge Engineering for Material Science, Beijing, China
Email: pwintphyukhinecs@gmail.com, zhswang@aliyun.com, zhswang@sohu.com

How to cite this paper: Khine, P.P. and Shun, W.Z. (2017) Big Data for Organizations: A Review. *Journal of Computer and Communications*, **5**, 40-48.
https://doi.org/10.4236/jcc.2017.53005

Received: January 18, 2017
Accepted: March 10, 2017
Published: March 13, 2017

Abstract

Big data challenges current information technologies (IT landscape) while promising a more competitive and efficient contributions to business organizations. What big data can contribute to is what organizations have been wanted for a long time ago. This paper presents the nature of big data and how organizations can advance their systems with big data technologies. By improving the efficiency and effectiveness of organizations, people can benefit the can take advantages of a more convenient life contributed by Information Technology.

Keywords

Big Data, Big Data Models, Organization, Information System

1. Introduction

Business organizations have been using big data to improve their competitive advantages. According to McKinsey [1], organizations which can fully apply big data get competitive advantages over its competitors. Facebook users uploads hundreds of Terabytes of data each day and these social media data are used for developing more advanced analysis which aim is to take more value from user data. Search Engines like Google and Yahoo are already monetizing by associating appropriate ads based on user queries (*i.e.* Google use big data to give the right ads to the right user in a split seconds). In applying information systems to improve their organization system, most government organization left behind compared to the business organizations [2]. Meanwhile some government already take initiative to get the advantages of big data. E.g. Obama's government announced investment of more than $200 million for Big Data R & D in Scientific Foundations in 2012 [3]. Today, people are living in the data age where data become oxygen to people as organizations are producing data more than they

can handle leading to big data era.

This paper is sectioned as follows: Section II of this paper describes Big Data Definitions, Big Data Differences and Sources within data, Big Data characteristics, and databases and ELT process of big data. Section IV is mainly concerned with the relationship between big data information systems and organizations, how big data system should be implemented and big data core techniques for organizations. Section IV is the conclusion of the paper.

2. Big Data for Organizations

The nature of Big Data can be expressed by studying the big data definition, the data hierarchy and sources of big data, and its prominent characteristics, databases and processes. According to [1], there are five organization domains for big data to create value and based on the size of the potential including health care, manufacturing, public sector administration, retail and global personal location data. There are many potential organizations which require big data solution such as scientific discovery (e.g. astronomical organizations, weather predictions) with huge amount of data.

2.1. Big Data Definition

Big data refers to the world of digital data which becomes enormous to be handled by traditional data handling techniques. Big data is defined in here as "a large volume of digital data which require different kinds of velocity based on the requirements of the application domains which has a wide variety of data types and sources for the implementation of the big data project depending on the nature of the organization."

Big data can be further categorized into Big Data Science and Big data framework [4]. Big data science is "the study of techniques covering the acquisition, conditioning, and evaluation of big data", whereas big data frameworks are "software libraries along with their associated algorithms that enable distributed processing and analysis of big data problems across clusters of computer units." It is also stated that "an instantiation of one or more big data frameworks is known as big data infrastructure."

2.2. Big Data Differences and Sources within Data

According to the basic data hierarchy as described in **Table 1**, different levels of computer systems have been emerged based on the nature of the application domains and organizations to extract the required value of the data (required hierarchy level data). Big data, instead, try to get value from since the "data" steps by applying big data theories and techniques regardless of types and level of information systems.

Based on the movement of data, data can be classified into "data in motion" and "data at rest". "Data in motion" means data which have not stored in a storage medium *i.e.* moving data such as streaming data comes from IoTs devices. They need to control almost in real time and need interactive controlling. "Data

Table 1. Hierarchy of data.

Hierarchy	Description
Data	Any piece of raw information that is unprocessed e.g. name, quality, sound, image, etc.
Information	Data is processed into a useful form that become information. e.g. employee information (data about employee)
Knowledge	Information is advanced by adding more contents from human experts that become knowledge (e.g. Pension data about employee)
Business Insight	Information is extracted and used in a way that help improve the business processes. (e.g. predicting the trends of customer buying patterns based on current information)

at rest" are data that can be retrieved from the storage systems such as data from warehouses, RDBMS (Relational Database Management Systems) databases, File systems e.g. HDFS (Hadoop Distributed File Systems), etc.

Traditional "Bringing data to perform operations" style is not suitable in voluminous big data because it will definitely waste the huge amount of computational power. Therefore, big data adopts the style of "Operations go where data exist" to reduce computational costs which is done by using the already well-established distributed and parallel computing technology [5]. Big data is also different from traditional data paradigm. Traditional data warehouses approaches map data into predefined schema and used "Schema-on-write" approach. But when big data handle data, there is no predefined schema. Instead, the required schema definition is retrieved from data itself. Therefore, big data approach can be considered as "Schema-on-Read" approach.

In information age with the proliferation of data in every corner of the world, the sources of big data can be difficult to differentiate. Big data sourced in the proliferation of social media, IoTs, traditional operation systems and people involvement. The sources of big data stated in [4] are IoTs (Internet of Things) such as sensor, social networks such as Twitter, open data permitted to be used by government or some business organizations (e.g. twitter data) and crowd sourcing which encourage people to provide and enter data especially for massive scale projects (e.g. census data). The popularity, major changes or new emergence of different organizations will create the new sources of big data. E.g. in the past, data from social media organizations such as Facebook, twitter, are not predicted to become a big data source. Currently, data from mobile phones handled by telecommunication companies, and IoTs for different scientific researches become important big data sources. In future, transportation vehicles with Machine-to-Machine communication (data for automobile manufacturing firms), and data from Smart city with many interconnected IoT devices will become the big data sources because of their involvement in people daily life.

2.3. Characteristics of Big Data

The most prominent features of big data are characterized as Vs. The first three

Vs of Big data are Volume for huge data amount, Variety for different types of data, and Velocity for different data rate required by different kinds of systems [6].

Volume: When the scale of the data surpass the traditional store or technique, these volume of data can be generally labeled as the big data volume. Based on the types of organization, the amount of data volume can vary from one place to another from gigabytes, terabytes, petabytes, etc. [1]. Volume is the original characteristic for the emergence of big data.

Variety: Include structured data defined in specific type and structure (e.g. string, numeric, etc. data types which can be found in most RDBMS databases), semi-structured data which has no specific type but have some defined structure (e.g. XML tags, location data), unstructured data with no structure (e.g. audio, voice, etc.) which their structures have to be discovered yet [7], and multi-structured data which include all these structured, semi-structured and unstructured features [7] [8]. Variety comes from the complexity of data from different information systems of target organization.

Velocity: Velocity means the rate of data required by the application systems based on the target organization domain. The velocity of big data can be considered in increasing order as batch, near real-time, real-time and stream [7]. The bigger data volume, the more challenges will likely velocity face. Velocity the one of the most difficult characteristics in big data to handle [8].

As more and more organizations are trying to use big data, big data Vs characteristics become to appear one after another such as value, veracity and validity. Value mean that data retrieved from big data must support the objective of the target organization and should create a surplus value for the organization [7]. Veracity should address confidentiality in data availablefor providing required data integrity and security. Validity means that the data must come from valid source and clean because these big data will be analyzed and the results will be applied in business operations of the target organization.

Another V of data is "Viability" or Volatility of data. Viability means the time data need to survive *i.e.* in a way, the data life time regardless of the systems. Based on viability, data in the organizations can be classified as data with unlimited lifetime and data with limited lifetime. These data also need to be retrieved and used in a point of time. Viability is also the reason the volume challenge occurs in organizations.

2.4. Database Systems and Extract-Load-Transform (ELT) in Big Data

Traditional RDBMS with ACID properties–(Atomicity, Consistency, Isolation and Durability) is only intended for structured data cannot handled all V's requirements of big data and cannot provide horizontal scalability, availability and performance [9]. Therefore, NoSQL (not only SQL) databases are need to use based on the domains of the organizations such as Mongo DB, Couch DB for documentation databases, Neo4j for graph databases, HBase columnar database

for sparse data, etc. NoSQL database use the BASE properties (Basically Available, Soft state, Eventual consistency). Because big data are based on parallel computing and distributed technology, CAP (Consistency, Availability, and Partition) theorem will affect in big data technologies [10].

Data warehouses and data marts store valid and cleaned data by the process of ETL (Extract-Transform-Load). Preprocessed, highly summarized and integrated (Transformed) data are loaded into the data warehouses for further usage [11]. Because of heterogeneous sources of big data, traditional transformation process will charge a huge computational burden. Therefore, big data first "Load" all the data, and then transform only the required data based on need of the systems in the organizations. The process can change into Extract-Load-Transform. As a result, new idea like "Data Lake" also emerged which try to store all data generated by organizations and has overpower of data warehouses and data mart although there are critics for becoming a "data swamp" [12].

3. Big Data in Organizations and Information Systems

Many different kind of organizations are now applying and implementing big data in various types of information systems based on their organizational needs. Information systems emerged according to the requirements of the organizations which are based on what organizations do, how they do and organizational goals. According to Mintzberg, five different kinds of organization are classified based on the organization's structure, shape and management as (1) Entrepreneurial structure—a small startup firm, (2) Machine Bureaucracy—medium sized manufacturing firm with definite structure, (3) Divisionalized bureaucracy—a multi-national organization which produces different kinds of products controlled by the central headquarter, (4) Professional bureaucracy—an organization relys on the efficiency of individuals such as law firms, universities, etc., and (5) Adhocracy such as consulting firm. Different kinds of information systems are required based on the work the target organization does.

Information systems required by the organization and the nature of problems within them reflects the types of organizational structure. Systems are structured procedures for the regulations of the organization limited by organization boundary. These boundary express the relationship between systems and its environment (organization). Information systems collects and redistribute data within internal operations of the organization and organization environment using the three basic simplest procedures-inputting data, performing processing and outputting the information. Among the organization and systems are "Business processes" which are logically related tasks with formal rules to accomplish a specific work which need to coordinate throughout the organization hierarchy [2]. These organizational theories are always true regardless of old or new evolving data methodologies.

3.1. Relationship between Organization and Information Systems

The relationship between organization and information systems are called socio-

technical effects. This socio-technical model suggests that all these components-organizational structure, people, job tasks and Information Technology (IT)-must be changed simultaneously to achieve the objective of the target organization and information systems [2]. Sometimes, these changes can result in changing business goals, relationship with people, and business processes for target organization, blur the organizational boundaries and cause the flattening of the organization [1] [2]. Big data transforms traditional siloed information systems in the organizations into digital nervous systems with information in and out of relating organizational systems. Organization resistance to change is need to be considered in every implementation of Information systems. The most common reason for failure of large projects is not the failure of the technology, but organizational and political resistance to change [2]. Big data projects need to avoid these kind of mistake and implement based on not only from information system perspective but also from organizational perspective.

3.2. Implementing Big Data Systems in Organizations

The work [13] provide a layered view of big data system. To make the complexity of big data system simpler, the big data system can be decomposed into a layered structure according to a conceptual hierarchy. The layers are "Infrastructure Layer" with raw ICT resources, "Computing Layer" which encapsulating various data tools into a middleware layer that runs over raw ICT resources, and "Application layer" which exploits the interface provided by the programming models to implement various data analysis functions to develop various field related applications in different organizations.

Different scholars are considering the system development life cycle of big data system project. Based on IBM's three-phases to build big data projects, the work in [4] proposed a holistic view for implementing the big data projects.

Phase 1. Planning: Involves Global Strategy Elaboration where the main idea is that the most important thing to consider is not technology but business objectives.

Phase 2. Implementation: This stages are divided into 1) data collecting from major big data sources, 2) data preprocessing by data cleaning for valid data, integrating different data types and sources, transformation (mapping data elements from source to destination systems and reducing data into a smaller structure (sometimes data discretization as a part of it), 3) smart data analysis *i.e.* using advanced analytics to extract value from a huge set of data, apply advanced algorithms to perform complex analytics on either structured or unstructured data, 4) representation and visualization for guiding the analysis process and presenting the results in a meaningful way.

Phase 3. Post implementation: This phase involves 1) actionable and timely insight extraction stage based on the nature of organization and the value that organization is seeking which decide whether the success and failure of big data project, 2) Evaluation stage evaluates a Big data project, it is stated that diverse data inputs, their quality, and expected results are required to consider.

Based on this big data project life cycle, organization can develop their own big data projects. The best way to implement big data projects is to use both technologies that are before and after big data. E.g. use both Hadoop and warehouse because they implement each other. US government considers "all contents as data" when implementing big data projects. In digital era, data has the power to change the world and need careful implementation.

3.3. Big Data Core Techniques for Organizations

There are generally two types of processing in big data-batch processing and real-time processing based on the domain nature of the organization. The fundamental of big data technology is based on MapReduce Model [14] by Google for processing batch work load of their user data. It is based on scale out model of the commodity servers. Later, real-time processing models such as twitter's Storm, Yahoo's S4, etc. become appear because of the near-real time, real time and stream processing requirements of organizations.

The core of MapReduce model is the power of "divide and conquer method" by distributing the jobs on the clusters of commodity servers with two steps (Map and Reduce) [14]. Jobs are divided and distributed over the clusters, and the completed jobs (intermediate results) from Map phases are sent to the reduce phase to perform required operations. In a way, In the MapReduce paradigm, the Map function performs filtering and sorting and Reduce function carries out grouping and aggregation operations. There are many implementations of MapReduce algorithm which are in open source or proprietary. Among the open source frameworks, the most prominent one is "Hadoop" with two main components—"MapReduce Engine" and "Hadoop Distributed File System (HDFS)"—In the HDFS cluster, files are broken into blocks that are stored in the DataNodes. NameNode maintains meta-data of these file blocks and keeps tracks of operations of Data Node [7]. MapReduce provide scalability by distributed execution and reliability by reassigning the failed jobs [9]. Other than MapReduce Engine and HDFS, Hadoop has a wide variety of ecosystem such as Hive for warehouses, Pig for query, YARN for resource management, Sqoop for data transfer, Zookeeper for coordination, etc. and many others. Hadoop ecosystem will continue to grow as new big data systems appeared according to the need of the different organizations.

Organizations with interactive nature and high response time require real-time processing.

Although MapReduce is dominant batch processing model, real-time processing models are still competing with each other, each with their own competitive advantages.

"Storm" is a prominent big data technology for Real-time processing. The famous user of storm is Twitter. Different from MapReduce, Storm use a topology which is a graph of spouts and bolts that are connected with stream grouping. Storm consume data streams which are unbounded sequences of tuples, splits the consumed streams, and processes these split data streams. The pro-

cessed data stream is again consumed and this process is repeated until the operation is halted by user. Spout performsas a source of streams in a topology, and Bolt consumes streams and produce new streams, as they execute in parallel [15].

There are other real-time processing tools for Big Data such as Yahoo's S4 (Simple Scalable Streaming System) which is based on the combination of actor models and MapReduce model. S4 works with Processing Elements (PEs) that consume the keyed data events. Messages are transmitted between PEs in the form of data events. Each PE's state is inaccessible to other PEs and event emission and consumption is the only mode of interaction between PEs. Processing Nodes (PN) are the logical hosts of PEs which are responsible for listening to the events, executing operating on the incoming events, dispatching events with the assistance of the communication layer, and emitting output events [16]. There is no specific winner in stream processing models, and organizations can use appropriate data models that are consistent with their works.

Regardless of batch or real-time, there are many open source and proprietary software framework for big data. Open source big data framework are Hadoop, EPCC (High Performance Computing Cluster), etc. [7]. Many other proprietary big data tools such as IBM BigInsight, Accumulo, Microsoft Azure, etc. has been successfully used in many business areas of different organizations. Now, big data tools and libraries are available in other languages such as Python, R, etc. for many different kinds of specific organizations.

4. Conclusion

Big data is a very wide and multi-disciplinary field which requires the collaboration from different research areas and organizations from various sources. Big data may change the traditional ETL process into Extract-Load-Transform (ELT) process as big data give more advantages in moving algorithms near where the data exist. Like other information systems, the success of big data projects depend on organizational resistance to change. Organizational structure, people, tasks and information technologies need to change simultaneously to get the desired results. Based on the layered view of the big data [13], big data projects can implement with step-by-step roadmap [4]. Big data sources will vary based on the past, present and future of the organizations and information systems. Big data have power to change the landscape of organization and information systems because of its different unique nature from traditional paradigms. Using big data technologies can make organizations get overall advantage with better efficiency and effectiveness. The future of big data will be the digital nervous systems for organization where every possible systems need to consider the big data as a must have technology. Data age is coming now.

Acknowledgements

I want to express my gratitude for my supervisor Professor Wang Zhao Shun for encouraging and giving suggestions for improving my paper.

References

[1] Manyika, J., *et al*. (2011) Big Data: The Next Frontier for Innovation, Competition, and Productivity. San Francisco, McKinsey Global Institute, CA, USA.

[2] Laudon, K.C. and Laudon, J.P. (2012) Management Information Systems: Managing the Digital Firm. 13th Edition, Pearson Education, US.

[3] House, W. (2012) Fact Sheet: Big Data across the Federal Government.

[4] Mousanif, H., Sabah, H., Douiji, Y. and Sayad, Y.O. (2014) From Big Data to Big Projects: A Step-by-Step Roadmap. *International Conference on Future Internet of Things and Cloud*, 373-378

[5] Oracle Enterprise Architecture White Paper (March 2016) An Enterprise Architect's Guide to Big Data: Reference Architecture Overview.

[6] Laney, D. (2001) 3D Data Management: Controlling Data Volume, Velocity and Variety, Gartner Report.

[7] Sagiroglu, S. and Sinanc, D. (2013) Big Data: A Review. *International Conference on Collaboration Technologies and Systems* (*CTS*), 42-47.

[8] de Roos, D., Zikopoulos, P.C., Melnyk, R.B., Brown, B. and Coss, R. (2012) Hadoop for Dummies. John Wiley & Sons, Inc., Hoboken, New Jersey, US.

[9] Grolinger, K., Hayes, M., Higashino, W.A., L'Heureux, A., Allison, D.S. and Capretz1, M.A.M. (2014) Challenges of MapReduce in Big Data, IEEE 10th World Congress on Services, 182-189.

[10] Hurwitz, J.S., Nugent, A., Halper, F. and Kaufman, M. (2012) Big Data for Dummies, 1st Edition, John Wiley & Sons, Inc, Hoboken, New Jersey, US.

[11] Han, J., Kamber, M. and Pei, J. (2006) Data Mining: Concepts and Techniques. 3rd Edition, Elsevier (Singapore).

[12] Data Lake. https://en.m.wikipedia.org/wiki/Data_lake

[13] Hu, H., Wen, Y.G., Chua, T.-S. and Li, X.L. (2014) Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access*, **2**, 652-687. https://doi.org/10.1109/ACCESS.2014.2332453

[14] Dean, J. and Ghemawat, S. (2008) MapReduce: Simplified Data Processing on Large Clusters. *Commun ACM*, 107-113. https://doi.org/10.1145/1327452.1327492

[15] Storm Project. http://storm.apache.org/releases/2.0.0-SNAPSHOT/Concepts.html

[16] Neumeyer, L., Robbins, B., Nair, A. and Kesari, A. (2010) S4: Distributed Stream Computing Platform. 2010 *IEEE International Conference on Data Mining Workshops* (*ICDMW*). https://doi.org/10.1109/ICDMW.2010.172