The Intersection of Big Data and the Data Life Cycle: Impact on Data Management

Janet L. Kourik and Jiangping Wang

***Abstract*—Big data is an emerging approach to data processing and analytics that will demand advances in data management. Big data goes beyond traditional operational and data warehousing databases to address the rapid increase in both size and variety of data. Big data brings with it many opportunities to extend and enhance the value of data for business in many industries. This paper examines the dimensions of big data and opportunities that that big data offers. A big data life cycle is summarized. The four dimensions of data are then used to analyze the impact of big data on data management throughout the big data life cycle. Findings include that big data brings with it substantial challenges that impact every phase of the big data life cycle and will require substantial advances in data management.**

***Index Terms*—Big data, data management, data life cycle, analytics.**

I. Background

Data are growing at an exponential rate. The dramatic growth in both the volume and type of data is leading a sea change in data management that is encapsulated under the designation of big data. Big data offers exciting possibilities in many fields including business, manufacturing, health care, research, government, and scientific research [1]. Businesses that use the Internet and Web are poised to take advantage of new data sources in the form of clickstreams that can be used to tailor responses to the user. Big data is frequently used to customize the user experience by interests, searching habits, and page views. For example many shopping applications suggest additional products a customer might be interested in purchasing based on prior purchases by customers with similar demographics.

Big data as a paradigm shift crosses the boundaries of data management, database, information retrieval and knowledge management. Big data extends the existing fields of data management and analysis. As a result data management must adapt to accommodate significant challenges that existing data management techniques do not fully address.

Despite the frequent use of the phrase 'big data' there are many different definitions for the phrase. Nearly all definitions account for the enormous scale of data being considered. Consider for instance that google analyzes the clicks, links, and content on 1.5 trillion page views per day and delivers search results with personalized advertising in milliseconds. Businesses are struggling to manage the vast amounts of data.

Most definitions of big data go beyond scale to address other characteristics of the data that taken together pose complex challenges [2]. From an industry perspective, Gartner defines big data as "… high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation." The National Institute of Standards and Technology [3] Big Data Infrastructure Framework says big data "consists of extensive datasets – primarily in the characteristics of volume, variety, velocity, and/or variability – that require a scalable architecture for efficient storage, manipulation, and analysis." The NIST definition foreshadows the components and underlying technologies that challenge the management of big data. In fact, insight into the underlying technologies becomes more apparent as NIST defines the big data paradigm as "…the distribution of data systems across horizontally coupled, independent resources to achieve the scalability needed for the efficient processing of extensive datasets." De Mauro, Greco and Grimaldi suggest that big data is more than the information and technology and should also include both methods and impact [2], which we include under the dimension of value. The definition they propose brings all four elements into the forefront: "big data represents the information assets characterized by such high volume, velocity and variety to require specific technology and analytical methods for its transformation into value."

Other definitions of big data include similar Vs in varying quantities. For the purpose of this paper we adopt a 4 Vs approach that includes volume, velocity, variety, and veracity in the service of generating value.

Examples of big data focus on facts about customers, business, production, population, or weather for instance. Big data is distinctive in that it encompasses many data types. In addition to the traditional structured data associated with enterprise databases, big data includes semi-structured and unstructured data as well.

This paper examines the major dimensions of big data, and analyzes the big data environment for the inherent challenges facing data management. Further, each phase of the data lifecycle is analyzed to reveal the problems introduced by big data.

II. Primary Dimensions of Big Data

The primary dimensions of big data encompass many characteristics that can be succinctly classified as volume, velocity, variety, and veracity for the express purpose of generating value. These dimensions, as shown in Fig. 1, each reflect a variety of features and affect one another. The

Manuscript received August 12, 2017; revised November 20, 2017.

The authors are with Math & Computer Science Department, Webster University, St. Louis, Missouri, USA (e-mail: kourikjl@webster.edu, wang@webster.edu).

doi: 10.18178/ijke.2017.3.2.083

dimensions include some overlapping concerns and their interactions can produce challenges to managing big data.

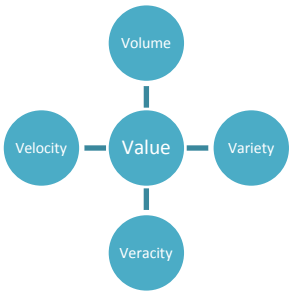

Fig. 1. Magnetization as a function of applied field.

Volume is the dimension most evident in the big data phrase itself. Big data continues to grow faster and faster. At any given time approximately 90% of the world's data was created in the preceding two years [4]. This has been and will continue to be true over time.

Presently growth in data volume translates into 2.5 quintillion bytes (Exabytes) of data created every day. In a world of accustomed to measuring data in gigabytes (GB) higher measures of data size are now in common use. Retail consumers can purchase hard drives in terabytes (TB) that are equivalent to 1000 Gb. Data warehouses have reached the size of petabytes (PB) that are equivalent to 1,000,000 GB. It is estimated that 40 zettabytes (43 trillion gigabytes) of data will be created by 2020.

Velocity is a dimension related to the speed at which data flows. Unlike traditional business transactions, big data encompasses streaming and continuous data. In many cases the data streams are generated by automated systems and surpass rates of data production in most enterprise systems.

The dimension of velocity also involves two other characteristics – analysis and volatility. Implicit in velocity is the need to rapidly analyze the data to respond in similar timeframes. In many cases the flow of data exceeds the rate at which data can be stored and analyzed suggesting the volatility, or transient nature, of the data. Volatility concerns include for instance how long the data is valid and how long should it be stored. It is easy to see how the dimension of velocity interacts with other dimensions of volume and veracity.

Variety is the dimension that addresses the heterogeneous nature of big data. This dimension reflects the broad range of data types and data representation that may be involved in big data processing. Unlike transaction data which is structured and discrete in nature, big data involves semi-structured and unstructured data. Network and communication data may be considered semi-structured information as they encapsulate the unstructured data of text within network communication structures such as packets. Unstructured data includes text, images, and other non-transactional data. Video, audio, click stream and sensor data introduce streaming and continuous data into the dimension of variety along with different data representations. The variety mixture of above data formats and types are more resulted from interaction between people and machines, hence data presents more on its multi-structured nature. Additional processing of semi-structured and unstructured data is needed to develop metadata and prepare the data for analysis that can be meaningful and generate value.

The heterogeneous nature of the variety dimension also refers to the need to integrate data from dissimilar sources for further processing and analysis. For example internal operational data may benefit from alignment with customer behavior, social media and external market data. In each case the data type and representation of the data is likely to differ.

Veracity is a dimension of big data that incorporates several facets such as uncertainty, accuracy, and validity of data. Many sources of big data may have poor data quality or are inherently uncertain. For example, customer sentiments while valuable are subjective and not measured on precise scale. Further, when data is more precise, the accuracy and validity of data may be questioned based its source. Timeliness and freshness of data also have an impact on the accuracy and validity and by extension data quality.

Value is the core of big data that addresses the cost/benefit proposition. Big data must answer business questions or provide for the discovery of meaningful and actionable data in order to contribute value to business. Unlike typical operational and data warehouse data, discovery of value is less direct.

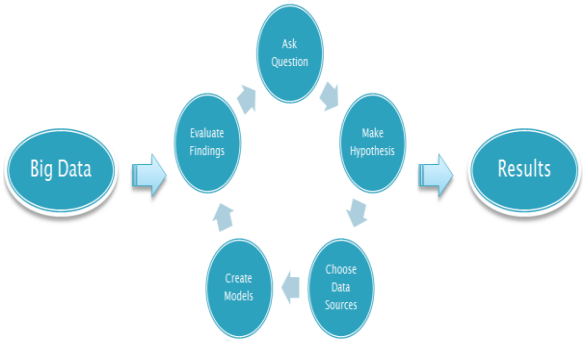

Fig. 2. Big data value discovery process.

Discovering value in big data involves iterative exploration via queries, models, analytical algorithms, and visualizations, as shown in Fig. 2. For example, big data value can be revealed through discovering consumer preference or sentiment to make a relevant offer onsite as well as in time. Machine learning and many other data mining techniques are a critical element of value discovery on large volume datasets because many types of big data are sparse in value. That is, the data may be valuable but with low value density when compared to the volume of data analyzed [5].

III. Challenges and Impact on Data Management

A traditional view of the data management life cycle varies but typically includes six or seven phases. One classic model is POSMAD, an acronym that represents the information resource life cycle as six fundamental phases of an information resource life cycle – plan, obtain, store and share, maintain, apply, and dispose [6]. From an industry point of view, another model includes roughly seven phases: data capture, data maintenance, data synthesis, data usage, data publication, archive and purging [7]. The phases in both models closely align on most essential concepts.

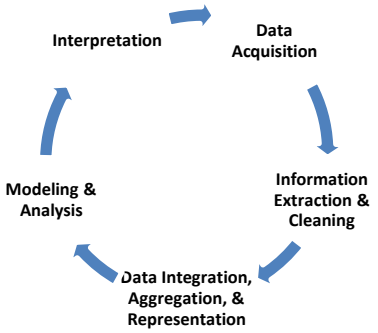

Fig. 3. Big data life cycle.

Big data life cycle itself in Fig. 3, closely resembles the traditional data management life cycle with five major phases including [8]:

- Data acquisition
- Information extraction and cleaning
- Data integration, aggregation and representation
- Modeling and analysis
- Interpretation

Using the lenses of the dimensions, we examine each phase of the data life cycle to distinguish where big data will impact data management.

A. *Data Acquisition Phase*

During the acquisition phase, the dimensions of volume, velocity, variety, and veracity all expose challenges in data management. The volume of data often exceeds our current capabilities to process and store. Machines and devices generate annually 1000 Exabyte of data with projections of increasing volume by 20 times in the coming decade [9]. The bandwidth of communications places constraints on the transmission of large volumes of data. New techniques are needed to efficiently transfer and receive vast volumes of data.

Velocity points to efficiency demands that have increased given the speed of data and timely preparation. Near real time efficiencies are particularly important in sensor networks and bio-medical applications. New data network and transmission systems are needed to support the demands of big data [10].

The data environment has grown in complexity crossing the boundaries of data sources, content format, and data stores. Data sources now include social media, machine generated data, sensor measurements, public datasets, curated datasets, multimedia, and the internet-of-things (IoT) in addition to transaction and operational data. Content formats include structured, semi-structured, and unstructured data. A wide range of techniques are needed to support column-oriented, document-oriented, graph and key-value pair data stores [11].

Variety dimension brings with it dissimilar data sources, formats, and stores that challenge the data management task of cleaning and transforming the data. For example, health care data typically involves structured data from providers and insurance companies but it also needs to integrate unstructured data such as x-rays and other imaging tests. In the veracity dimension, accuracy in healthcare data is critical as errors may have profound consequences for patients.

B. *Information Extraction & Cleaning Phase*

Volume and velocity data dimensions point to the scale, access methods, and speed needed in storage for support of information extraction and cleaning. Hard drives may reasonably store up to four petabytes of data yet face high latency in read-time [11]. Solid state memory supports up to two terabytes of data with faster data access and transfer. Cloud storage can provide the scalability needed but fees accrue for storage and access. When data volume and speeds exceed existing capabilities data reduction strategies such as sampling and compression techniques may be needed yet are problematic to apply when the data of value may not yet be known.

Variety and veracity data dimensions reveal several challenges to data management as big data is notably messy. Unstructured data and continuous data make it difficult to extract let alone clean data. Extraction needs to pull the required data and apply a structured format. Such a process will vary significantly by the nature of the data and become particularly challenging when many heterogeneous sources of data are involved. Extracting tags and features from multimedia such as image, video, and audio, is still a growing area of research. Traditional extract, transform, and load (ETL) processes are insufficient for big data.

Many of the sources of data have inherent reliability issues ranging from the subjective nature of social media data to noise associated with cyber-physical data such as the sensors or the IoT. Veracity is further diminished when the freshness of data from external sources may be unknown or lacking.

The value of big data will be diminished without improvements and success in the extraction and cleaning of big data.

C. *Data Integration, Aggregation & Representation Phase*

Once more the dimensions of volume, velocity, variety, and veracity all pose challenges to data management during the integration, aggregation and representation phase.

Master Data Management (MDM) attempts to guide organizations in creating and maintaining a single point of reference for authoritative data in an enterprise. Typically MDM relies on metadata and a relational database. Big data challenges traditional MDM through the data dimensions of variety, volume, velocity and veracity. A broader range of data sources and data types at enormous volume and speed are not readily handled by MDM architectures yet metadata and MDM are critical to big data [12].

Integration of data from various internal and external sources is far more difficult in the heterogeneous environment of big data. Historically, the unassuming task of entity identification across structured databases has been a challenge. Such a task becomes much more difficult as the variety of data and sources increases.

Scalability and performance are barriers to timely processing of big data. New algorithms, storage, file systems and processing capabilities are needed. Cleaning and aggregation would benefit from improvements in CPU and GPUs, parallel processing, and distributed processing. As systems become distributed additional concerns about consistency, availability, and partition tolerance (of the

CAP theorem) need to be addressed.

The veracity dimension of data uncovers some substantial problems with data quality in big data. Research shows that the scale of data alone multiplies preexisting data quality problems in direct proportion to size. Further, the complexity and interaction of data may increase data quality problems geometrically or exponentially [12]. Manually created data errors are difficult to correct by automated processes. This suggests that social media and crowdsourced data will continue to pose significant data quality challenges.

The value of big data hinges on continuing research into tools and techniques that address data quality from data integration and aggregation.

D. Modeling & Analysis Phase

The dimensions of volume, velocity, variety, veracity all impact data modeling and analysis and in turn the value that can be derived from big data. Big data does not fit into the traditional model of relational tables with well-structured rows and columns due to its variety data types.

Volume, velocity and variety dimensions require data modeling for big data to go beyond traditional relational database design. Relational databases are less able to handle diverse data formats and data types or scale. There are many emerging models that are essential to handling big data such as NoSQL databases, use of key-value pairs, and schema-on-read. NoSQL works with unstructured data storing documents for instance and separates data storage from data management [1]. The storage portion of NoSQL often uses key-value pairs to enable more efficient access across data types while avoiding conversion into a fixed schema.

Other emerging databases are column-oriented. Traditional relational database design uses a schema-on-write approach. Big data is benefiting from new schema-on-read approach that reduces delays by writing data first and organizing it later. The schema-on-read offers flexibility to store data in many formats and ease of reuse. To take advantage of both traditional data models and emerging models for big data, it may be necessary to integrate both in a unified platform for big data analytics [13].

Analytics and visualizations are computationally demanding. Efforts to improve the efficiency of computation include the use of in-memory databases and distributed systems. Both approaches reduce I/O overhead and execution time and scale well. Both techniques exact new data management issues. For instance, costs may be greater and security is a greater challenge. Examples of vendor-specific in-memory database limitations include limited data types, use of attribute constraints and computations as well as some recovery techniques such as mirroring and snapshots (SQL Server). Distributed systems may help availability and performance at the expense of consistency. New approaches to large-scale parallel programming such as MapReduce are needed to improve the performance of NoSQL databases. A drawback is that partitioning for MapReduce may introduce errors in analytical processes [12].

The variety dimension builds on predictive analytics and is reflected in the growth in types of analysis such as text, audio, video, social media, and real-time analytics.

Many traditional statistical techniques are not appropriate for the complexity of big data or are computationally very demanding. Traditional methods that are appropriate for big data analysis include cluster, factor, correlation and regression analysis as well as machine learning and data mining techniques. More complex methods include Bloom Filter, hashing, indexing, trie-tree, and parallel computing [10].

Another area of research includes large-scale network analysis that models explicit and implicit interactions [14]. Network analysis works well with some visualization techniques.

E. Interpretation Phase

All big data dimensions drive value that dominates the interpretation phase. Industry expectations for value from big data often emphasize outcomes related to customers. Optimization of operations and manufacturing are expected to a similar degree as financial and risk management. An interesting expectation is that big data will offer new business models [10].

How do the expectations compare to actual value creation? Big data research finds that value is created via the following categories: supporting transparency; discovering needs exceptions, variability and improving performance; segmenting populations to customize actions; supporting or replacing human decision making with automated algorithms; and innovating new business models, products, and services [15].

Interpretation still relies heavily on human evaluation and judgement in the design of analytical procedures as well as their results [16]. Developing a concurrent culture of data-driven decision making is a critical part of the big data challenge. Eventually, the value of big data lies on extracting useful and meaningful knowledge from the data that can solve business problems.

IV. Conclusion

Using the four dimensions of big data, we examined the phases of the big data life cycle. Each dimension was used to reveal the impact of big data on the life cycle phase and data management in general. It is now clear that the dimensions of big data have an impact on data management in every phase of the big data life cycle, as shown in Fig. 4.

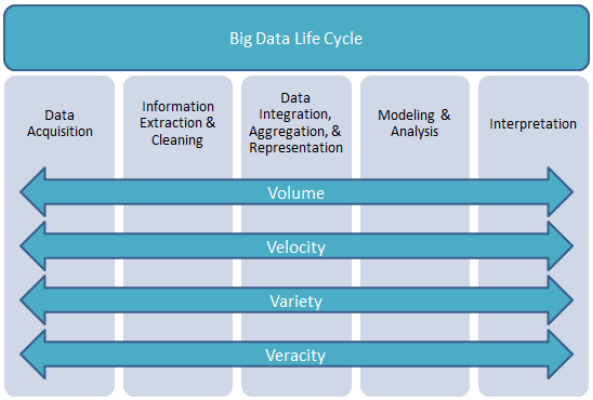

Fig. 4. Interaction of big data life cycle and dimensions.

The volume dimension heavily impacts scalability and

performance. The velocity dimension affects transmission and performance. The variety dimension raises issues with interoperability, openness and consistency. The veracity dimension highlights many concerns with data quality. The data management issues raised by volume, velocity, variety, and veracity have direct impact on the ability to generate value from big data. As more distributed computing becomes part of big data, availability, reliability and fault tolerance become critical concerns. In addition, security, risk management, and costs are issues that cross all dimensions of big data. It is also evident that people with skills in big data, data science and analytics are essential to the success of big data applications [16].

From analyzing the dimensions of big data it is clear that there are many limitations in data management that must be addressed through research and development in a variety of areas. Advances in data management are necessary in all phases of the big data life cycle so the business and analytical users can rely on richer yet high quality data for them to navigate through various data and information, verify hypothesis, analyze patterns, and perform knowledgeable data-driven decisions making. The value of big data can only be reached from insightful understanding of manageable and actionable data.

REFERENCES

[1] C. L. P. Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information Sciences*, vol. 275, pp. 314-347, 2014.

[2] A. De Mauro, M. Greco, and M. Grimaldi, "What is big data? A consensual definition and a review of key research topics," in *Proc. 4th International Conference on Integrated Information (IC-ININFO 2014).*

[3] National Institute of Standards and Technology (NIST). DRAFT NIST Big Data Interoperability Framework: vol. 7, Standards Roadmap, NIST.

[4] P. Heller and D. Piziak, *An Enterprise Architecture White Paper – An Enterprise Architect's Guide to Big Data*, 2015.

[5] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, pp. 137-144, 2015.

[6] D. McGilvray, *Executing Data Quality Projects*, Amsterdam: Morgan Kaufmann Publishers, 2008.

[7] M. Chisholm, "7 phases of a data life cycle," *Information Management*, p. 4, 2015.

[8] H. V. Jagadis, J. Gehreke, A. Labrinidis, Y. Papakonstantinoue, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big data and its technical challenges," *Communications of the ACM*, vol. 57, no. 7, pp. 86-94. 2014.

[9] S. Yin and O. Kaynak, "Big data for modern industry: Challenges and trends," in *Proc. of the IEEE*, 2015, pp. 143-146.

[10] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Network Applications*, vol. 19, pp. 171-209, 2014.

[11] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems*, vol. 47, pp. 98-115. 2015.

[12] D. Becker, B. McMullen, and T. D. King, "Big data, big data quality problem," in *Proc. IEEE International Conf. on Big Data, Santa Clara, California*, 2015, pp. 2644-2652.

[13] J. P. Dijcks and M. Gubar, "Integrating SQL and hadoop," *Business Intelligence Journal*, vol. 19, no. 2. pp. 49-55, 2014.

[14] P. B. Goes, "Big data and IS research," *MIS Quarterly*, vol. 38, no. 3, p. 6, 2014.

[15] S. F. Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How 'big data' can make big impact: findings from a systematic review and a longitudinal case study," *International Journal of Production Economics*, vol. 165, pp. 234-246, 2015.

[16] A. McAfee and E. Brynjolfsson, "Big data: The management revolution," *Harvard Business Review*, pp. 59-68, 2012.

Jiangping Wang is an associate professor of computer science at Webster University. He has a B.A. from Chongqing University, China, has a M.S. from the University of Leeds, United Kingdom and a Ph.D from the Missouri University of Science and Technology, Rolla, Missouri, USA. Wang's areas of teaching include database design, database applications, data warehousing, web databases, database in web services, and distributed application development. His areas of research include database management systems, decision support systems, business intelligence, e-commerce data processing, and software project managemen.

Janet L. Kourik is a professor in the Mathematics and Computer Science Department of Webster University in St. Louis, Missouri, US. She has a B.S. and C.S. from Webster University, an M.A. from Webster University and a Ph.D from Nova Southeastern University. Kourik's areas of teaching include database concepts and applications, information systems, operating systems, and distributed systems. Her areas of research include databases and analytics, agile methods, informatics, and computer science education.