US-China Education Review B, March 2019, Vol. 9, No. 3, 90-97
doi: 10.17265/2161-6248/2019.03.002

Big Data and Knowledge Management: A Possible Course to Combine Them Together

Sam Hijazi
Texas Lutheran University, Texas, USA

Big data (BD) is the buzz phrase these days. Everyone is talking about its potential, its volume, its variety, and its velocity. Knowledge management (KM) has been around since the mid-1990s. The goals of KM have been to collect, store, categorize, mine, and process data into knowledge. The methods of knowledge acquisition varied from organizational culture to the next. Typical processes converted data into information through traditional databases and then applied business intelligence and data mining methodologies to extract knowledge. With the recent arrival of BD as a disruptive technology and the center of BD, this paper attempts to combine KM and BD fields together. Both areas could help each other tremendously. KM historically, when applied correctly, has helped managers to make decisions faster and better, prevented reinventing the wheel, preserved some talented processes through keeping track of best practices, and prompted innovation due to knowledge sharing and dissemination. BD deals with massive amount of data and does not require a traditional database to be effective. BD has its tools and requirement that can be enhanced through KM. The final aim of this paper is to recreate a model where both BD and KM coexist. The author hopes with a better understanding of both fields to develop a new course where the focus is a productive intersection of KM and BD. To keep up with changing times, this paper will bring the needed awareness of these fields for information systems and business students.

Keywords: big data, knowledge management, model, knowledge, value, class design, introductory, learning, business

Introduction and Problem Statement

The exponential increase in data size is well known point. O'Doherty (2012) discussed some interesting facts concerning big data (BD), data management, and data visualization. Considering that the article is an old one, it still brings some valid comparisons. The author stated that the volume of data created by the U.S. companies big enough fill 10,000 the size of the Library of Congress. A retailer who utilizes BD effectively could enhance its operating margin by more than 60%. Bad data costs the economy $600 billion every year. Another interesting statistic was that BD will cost businesses around $232 billion in 2016. Every minute, YouTube users upload 48 hours of videos resulting in a span of eight years of media to watch every day. The article predicted that by the year of 2015, 4.4 million information technology (IT) jobs globally would be needed to support BD. By 2020, we would create 35 zettabytes (10 ^ 21 Byte Approx.) of data. Finally, 1.9 million jobs related to BD would be created in the U.S. by the year of 2019.

Sam Hijazi, Dr., associate professor, Math, CS & IS Department, Texas Lutheran University.

Knowledge management (KM) has been around for over 20 years. In his website KM—knowledge management, David Skyrme (n.d.) answered the question: “Why manage knowledge?” The author stated, “Organizations are knowledge-intensive”. Knowledge is a valuable resource that provides meaning to their operations”. “If you look at the market value of a public company, it is typically 5-10 times greater than the assets (predominantly physical assets) recorded in its balance sheet”. KM could easily create a practical approach to education. According to Walter Smith (2012), the most current education models are abstract by nature. Education is meant to create learning, but does not show how learning works. The abstract style it creates is necessarily bad and has worked in the past, but definitely has not worked all the time. KM can create alternative education systems by providing the same learning opportunity to everyone. The author added that the learning process can be used at five levels of KM. These include building knowledge, applying knowledge, organizing knowledge, personalizing knowledge, and teaching knowledge. The most important part of using KM in education is that the learning will understand knowledge itself. “Learning becomes a dynamic, multi-dimensional, integrated, and interactive process, and knowledge is managed efficiently and effectively in school, college, and university, on the job, in our personal lives, and in the community”.

The above facts and findings do not require any additional proof in order for us to decide that we need to offer additional classes in BD within the information systems curriculum. Note that the output of most BD systems is knowledge instead of information. Understating the nature of knowledge and how to turn it into an action is critical. After searching the existing literature deeply, there was very little evidence to be found of how mixing these two fields benefit our students in the process of becoming effective decision makers upon their graduation. Knowledge is not a fad; rather it is the most valuable asset in our modern world and even in our past. Without it, we could not have preserved our civilization. The adage, “Knowledge is power” is completely true. Nations who have more knowledge in their fabric and economy are leading the scene around the world. It is a clear disadvantage to turn a blind eye to these important topics. As educators, we have the responsibility to find the topics that should be blended to create a “maximum effect” on the future of learners. If we do not react accordingly, students will have to invest in additional training, seminars, and online classes to catch up with their peers in a very competitive market. There is evidence that BD has been incorporated into the academic world but little has been done to link it to KM. This paper is at attempt to create that bridge.

The Benefits of KM

Laal (2010) stated that KM has witnessed an increase in its popularity in the last decade. The author explored the concern whether or not KM is a fad. The findings indicate strongly that KM is not a fad and it is here to stay, mainly because our economy is based mostly on intellectual capital, another way to say knowledge. KM is recommended for all organizations since it helps in creating, capturing, sharing, and leveraging knowledge for all decision makers.

In another helpful article, David Skyrme (n.d.) discussed the benefits of KM. The author stated that we all know that organizations are “knowledge-intensive”. Knowledge is the most vital resource to compete with others. However, organizations do not manage their knowledge the way they manage their finances. The author reviewed 15 years of experience in the knowledge domain and divided his discussion into three main categories. These are:

1. Benefits from efficient processing of information and knowledge. This category discussed:

(a) Quicker access to information;

(b) Less redundancy and duplication;

(c) More time for professional to focus on more important issues;

(d) Knowing the source of knowledge and knowing who does what;

(e) Improved quality of information and knowledge;

(f) Access to current knowledge and thinking.

2. If the above benefits were established, this should lead us to the second category which is the internal benefits to the organization. This category discussed:

(a) Avoiding worst practices and sharing the best ones;

(b) Speeding up the time to market new products or services;

(c) Avoiding reinventing the wheels which lead to cost reduction;

(d) Capturing valuable knowledge before experts retire or move to other organizations;

(e) Reducing time to process information which result into faster problem solving and cost reduction.

3. Just like there were some internal benefits, there are benefits to stakeholders, especially customers. This category includes:

(a) Improving customer retention and satisfaction;

(b) Faster problem solving;

(c) Being consistent with all customers regardless of their geographical location;

(d) Acquiring more insight from the customers which improve the quality of the products or services;

(e) Better value for the cost;

(f) Improved reputation in the market.

The Benefits of BD and Its Vs

If you read any article about BD, more likely you are going to be exposed to the three main Vs of BD. These are volume, variety, and velocity. For the purpose of expanding the KM model, the paper will cover these factors and examine if there any other additional ones. Firican (2017) discussed these characteristics to understand the nature, advantages, and challenges of BD.

Volume

This is the most common attribute of BD, knowing that 90% of the existing data were created in the past two years. Also, Firican (2007) stated some staggering data where every minute, people upload 300 hours of video to YouTube. In 2016, appropriately 1.1 trillion pictures were taken, and this number most likely to rise by 9% in 2017. With the number of mobile devices, it is not surprising to see that the amount of data passing through global mobile traffic added up to 6.2 Exabytes per month. Exabyte is equal to 10 ^ 18 approximately.

Velocity

This characteristic is refers to the speed of generating, producing, refreshing, and streaming data. Velocity means data are accessed in real time and little time is wasted to access it.

Variety

This attribute means the nature of data itself. Most data are not structured as you would see in traditional databases. Data are mostly semi-structured or unstructured. In addition to multi-media data types, Firican mentioned click, sensor, and machine as a few examples.

Firican did not stop at the traditional three Vs, rather he stated that there are seven other Vs that should be

considered. These include: variability, veracity, validity, vulnerability, volatility, visualization, and value. For the purpose of the paper, it is important to expose the reader to these characteristics in order to understand their impact on the recent thinking concerning BD.

Variability

This attribute has to do with data types and sources. Variability refers also to the uneven speed it takes to load the data in the database engine.

Veracity

It is the classic GIGO, garbage in, and garbage out. This is considered one of the most serious V factor, knowing dirty data could erase the value of BD and the cost associated with it.

Validity

Validity is similar to veracity. According to Firican, 60% of data scientists spend their time cleansing the data to get ready for analysis. It is required to have a policy to assure that we have quality and consistent data.

Vulnerability

Security is an issue with small or BD. Firican referred to the hacking in May 2016 which resulted in the stealing of information from 167 million LinkedIn accounts and 360 million passwords and emails from MySpace users.

Volatility

This in reference to the freshness of the data and how long it stays relevant and useful. As a result of velocity and volume of BD, management must consider its volatility. Firican added that data must be related to your business needs and functions.

Visualization

Firican stated that limitation of memory and poor scalability and response time could be a challenge when visualizing massive amount of data. Traditional graphs would not work for billion pieces of data, therefore, other graphic methods, such as data clustering, sunbursts, parallel coordinates, circular network diagrams, cone tree, or sunburst should be considered.

Value

This attribute is considered by many as the most important one. It makes sense to say with business value, every other V is a waste of time. Firican emphasized values, such as understating of our customers, creating targets, optimizing processes, or in general, improving business performance. Extracting value from BD cannot be attained without a valid strategy

Discussion

In 2003, the author of this paper presented a model for knowledge creation. Later, the model was modified to emphasize action as the final output of any knowledge creation project. Without action, knowledge, no matter how costly it is, will be useless. The model is organic by nature and adjustable to the changes in the IT and the business world.

The Link Between KM and BD

Lamont (2012) stated, “A goal of KM over the years has been the ability to integrate information from

multiple perspectives to provide the insights required for valid decision-making". The article emphasized that the job of KM is not only to learn about our organizations, but also to transform them. The article stated that regardless of our measure of success, customer stratification, successful development, robust security, or profit, to excel in the "Knowledge Age", organizations and people must mature through the different stages of knowledge to transform their surroundings. Lamont (2017) discussed the need for KM programs to hire data scientists. There is a clear evidence that data science is becoming critical to all fields by providing opportunities for better employment regardless of the stage of their careers.

As we notice from this model (see Figure 1), it was created for handling knowledge creation/management in a small data world. From the model above (Hijazi, 2006), it is clear to the observer that the data processor is a Database Management System (DBMS). The use of the process is still valid for BD, except the engine must be updated. Here comes Hadoop to provide a major and timely addition and not a replacement. DBMS will continue to stay with us. Many businesses depend on their DBMS and could not imagine replacing it.

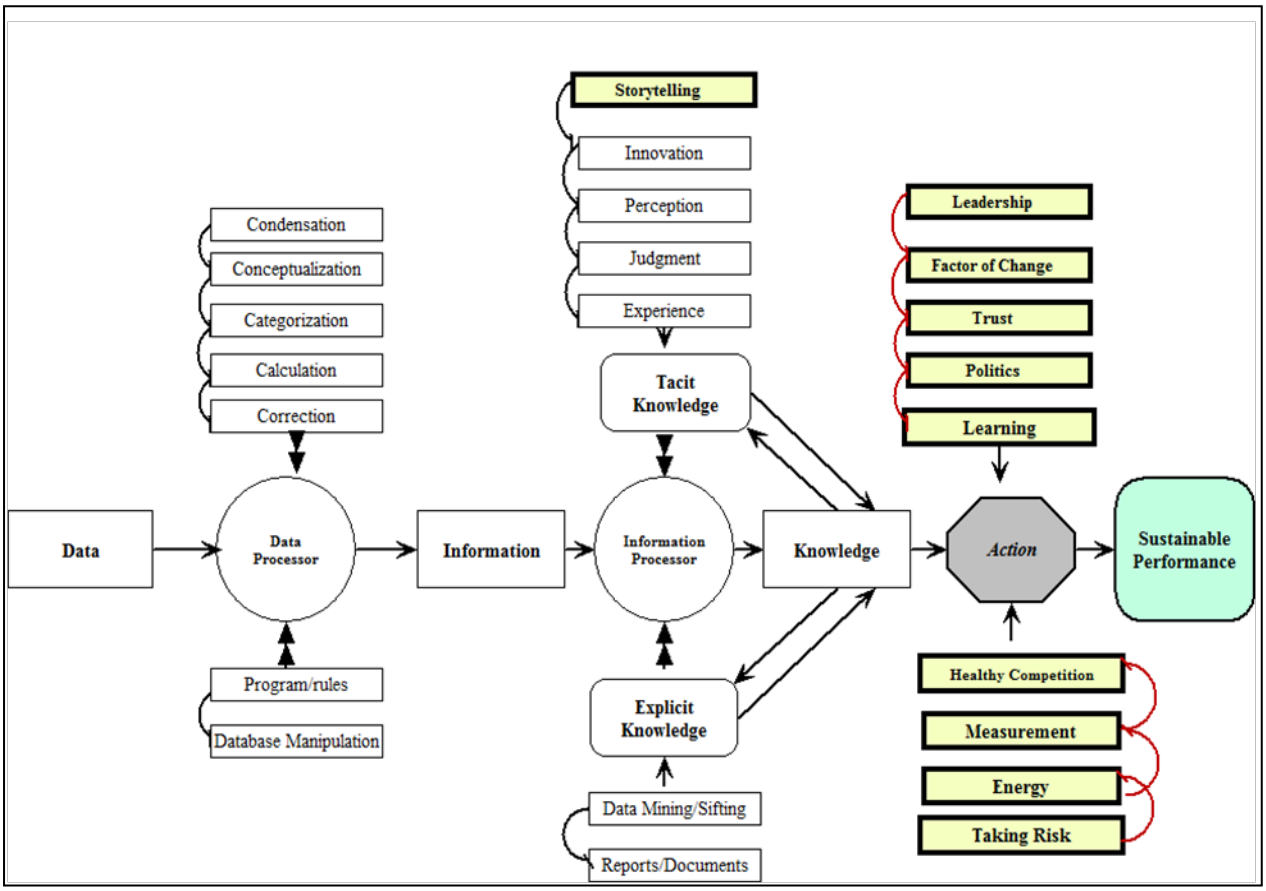

Figure 1. The early model.

Tacit knowledge will not change, it still resides mostly in people heads. Their intuition, experiences, judgment, innovation, perception, and many other important factors will continue to be a huge input to the process of knowledge creation. What will change is the technology and size of the tools that support explicit knowledge. It should be clear now that for BD to be successful, we need to add other designed and developed technologies. In addition to the relationship among data, information and knowledge, databases, queries, and

reporting, the classes will at least need to learn most technologies pertaining to BD. These include MapReduce, Hadoop, and Hive, as they are discussed below. The list also added other known and important technologies of BD for additional knowledge.

Rodrigues (2012) interviewed Dr. Kaur about the 10 emerging technologies for BD. The discussion ironically has covered most technologies that have become stranded. These included:

MapReduce. This is a programming pattern that allows for scalable execution for thousands of servers or even clusters of servers. The found tasks in MapReduce are the map task where a dataset is modified into pair values or records and the Reduce where a group of outputs from the map are clustered resulting in a number of records.

Hadoop. It the most common implementation of MapReduce and can work with multiple data sources. One clear and useful application of Hadoop is handling large and constantly changing data, such as those found in weather forecasting or social-media.

Hive. This technology is similar to structured query language (SQL) syntax. It uses business intelligence (BI) to query Hadoop clusters. It gives a developer a similar feeling to a conventional data store which results in the widespread use of Hadoop. Hive was developed by Facebook, but later became open-source.

PIG. PIG's function is similar to Hive, however, it uses a Perl-like language to query data stored in a Hadoop cluster. Similar to Hive, it was developed by a private developer, Yahoo, but later ended up as open-source.

WibiData. This tool combines Hadoop with a web analytics capability. It works with HBase as the database layer on top of Hadoop. It provides websites the ability to work with their user data in order to respond to the user's choice in real-time. It also gives a user personized contents, recommendations, and decision making help.

PLATFORA. This technology adds a friendly face to Hadoop. Hadoop requires intensive training and PLATFORA adds an abstract layer to organize and simplify the access to datasets stored in Hadoop.

Storage technologies. With the tremendous growth of data, there is a need to find different techniques for storing volumes of data. Data compression and visualization are the reasons associated with BD.

SkyTree. Rodrigues added that SkyTree is an analytics platform and machine learning platform in the area of BD. SkyTree handles volumes of data associated with machine learning where conventional tools would be able to do the job.

BD in the cloud. Rodrigues concluded the meeting with Dr. Kuar by stating that everything mentioned above is amiable in the cloud. Vendors are offering Hadoop clusters to meet business needs and to be scaled to their demand. BD and cloud computing are intertwined where cloud computing gives the chance to all companies to join the bandwagon of BD.

The new mode. As a result of the modification, the new model is ready to handle the new components of BD (see Figure 2).

The model still keeps all the helpful and productive steps, we have learned from knowledge creation/management. The model also keeps all the intangible factors, such as leadership, factor of change, trust, politics, and meta-learning as deter-minal factors for any application of technology to succeed. The end result is sustainable performance where success alone is not enough. Success needs to be evaluated, recharged, and ready to deal with all business obstacles that prevent it from being achieved.

The Importance of the Study

KM has been around for some time. Organizations have gained greater understanding of the value of knowledge as a major asset to their survival. BD has burst into the scene with a call for a change in the way we capture, cleanse, process, update, and sort through unimaginable volume of data a few years ago. This study attempts to show the impact of BD as an inescapable phenomenon and to link it to the wealth of managing knowledge. Knowledge is the outcome of both a BD project and a KM program. Why not combine them together? This study modified an earlier model for KM, but added the components pertaining to BD. The hope is to develop a class where both topics will be introduced together to students.

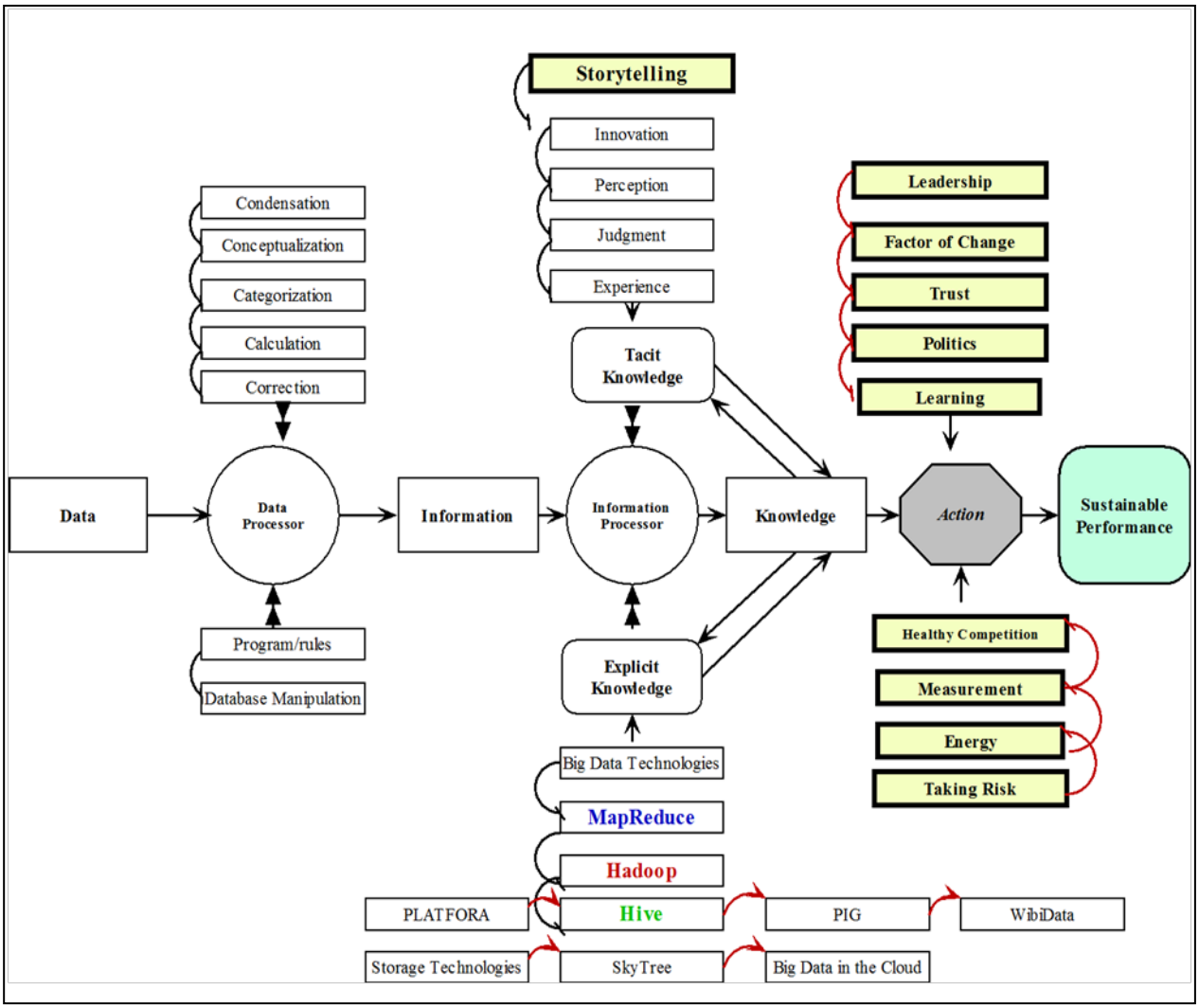

Figure 2. The new model.

Conclusion

There is no question that BD is a permanent and increasing phenomenon. As educators, we need to respond to changes in the business world. After reviewing keys areas in the field of KM and BD fields, it led to the modification of an early KM model to include those components that will generate explicit knowledge from massive amount of data. The model acknowledged the need for the addition of BD technologies.

However, it left all the earlier factors emphasized by a KM program, especially the ones that guarantee an action and sustainable performance at the end. In addition to the alteration of the module, this research should give the reader a good exposure to both fields where key concepts are included in the model to develop a new class. The study shows an alarming rate of increase in the volume of data. However, this will generate an opportunity to all concerned parties that data regardless of its nature—structured, semi-structured, or unstructured—will be used to increase our knowledge repository. Students in the field of business and information systems must know the value of both fields and more importantly how combine them in order to combine their strengths.

References

Firican, G. (2017). *The 10 Vs of big data*. Retrieved March 3, 2017, from https://upside.tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx

Hijazi, S., & Kelly, L. (2003). *Knowledge creation in higher education institutions: A conceptual model.* Retrieved March 12, 2017, from http://fits.depauw.edu/ascue/Proceedings/2003/p78.pdf

Knowledge Management Presentation. (2006, August). *Conference by the Society for applied learning technology (SALT).* Retrieved March 12, 2017, from http://www.salt.org/docs/wprogram06.pdf

Knowledge management as a model for education. (2012). Retrieved May 23, 2017, from https://evolllution.com/opinions/knowledge-management-as-a-model-for-education/

Laal, M. (2010). *Knowledge management in higher education.* Retrieved May 4, 2012, from http://ac.els-cdn.com/S1877050910004655/1-s2.0-S1877050910004655-main.pdf?_tid=63b6b32e-2186-11e7-9e96-00000aacb35e&acdnat=1492224854_bb01d2d056353a2ecc5f51d779781aa0

Lamont, J. (2012). *Big data has big implications for knowledge management.* Retrieved April 23, 2017, from http://www.kmworld.com/Articles/Editorial/Features/Big-data-has-big-implications-for-knowledge-management-81440.aspx

Lamont, J. (2017). *KM education: Data science takes the lead*. Retrieved April 4, 2017, from http://www.kmworld.com/Articles/Editorial/Features/KM-education-Data-science-takes-the- lead-117210.aspx

O'Doherty, P. (2012). *20 shocking fact and figures about "big data"*. Retrieved April 4, 2017, from https://www.espatial.com/articles/20-shocking-facts-and-figures-about-big-data

Rodrigues, T. (2012). *10 emerging technologies for "big data"*. Retrieved April 3, 2017, from http://www.techrepublic.com/blog/big-data-analytics/10-emerging-technologies-for-big-data/

Skyrme, D. (n.d.). *Why manage knowledge?* Retrieved April 4, 2017, from https://www.skyrme.com/kmbasics/whykm.htm

What is knowledge management. (n.d.). Retrieved April 3, 2017, from http://www.kminstitute.org/content/what-knowledge-management