Big Data Characteristics, Value Chain and Challenges

Rabiul Islam Jony
Queensland University of Technology
7 PUBLICATIONS **52** CITATIONS

Rakibul Islam Rony
Universitat Politècnica de Catalunya
10 PUBLICATIONS **94** CITATIONS

Musfiqur Rahman
University of Liberal Arts Bangladesh (ULAB)
7 PUBLICATIONS **15** CITATIONS

Big Data Characteristics, Value Chain and Challenges

Rabiul Islam Jony[1], Rakibul Islam Rony[2]
Musfiqur Rahman[1], Abiduzzaman Rahat [1]
[1]University of Liberal Arts Bangladesh
[2]Primeasia University

***Abstract*—Recently the world is experiencing an deluge of data from different domains such as telecom, healthcare and supply chain systems. This growth of data has led to an explosion, coining the term Big Data. In addition to the growth in volume, Big Data also exhibits other unique characteristics, such as velocity and variety. This large volume, rapidly increasing and verities of data is becoming the key basis of completion, underpinning new waves of productivity growth, innovation and customer surplus. Big Data is about to offer tremendous insight to the organizations, but the traditional data analysis architecture is not capable to handle Big Data. Therefore, it calls for a sophisticated value chain and proper analytics to unearth the opportunity it holds. This research identifies the characteristics of Big Data and presents a sophisticated Big Data value chain as finding of this research. It also describes the typical challenges of Big Data, which are required to be solved. As a part of this research twenty experts from different industries and academies of Finland were interviewed.**

***Key words*—Big Data, Big Data characteristics, Big Data Value chain, Big Data Challenges.**

I. INTRODUCTION

In the year 2000, when the Sloan Digital Sky survey started their work, its telescope in New Mexico collected more data on its first few weeks than had been amassed in the entire history of astronomy. After one decade, now its archive contains around 140 terabytes of data. Another large synoptic Survey Telescope in Chile is predicted to collect the same quantity of data every five days by 2016 [1]. Wal-Mart, the retail giant, generates around 2.5 petabytes data of 1 million customers‟ transactions every hour [2]. Facebook, a social networking website stores 500+ terabytes of new data every day. Search engines like Google processes 20 petabytes of data every day [1]. All these examples show how much data the world contains and how rapidly the volume of data is growing. Until 2003, 5 exabytes of data were created by humans, and currently this amount of data is created in only two days [3]. The amount of data in the digital world reached 2.72 zettabytes in 2012 and is expected to double in every two years reaching 8 zettabytes by 2015 [4].Data is getting so large and complex, that it is becoming difficult to process using traditional data processing applications, and introducing big data.

The most popular definition of big data is defined by Gartner as “Big data is high-volume, high-velocity and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization” [5].

Big data is currently treated as a technology, which has been developed to handle large volumes of fast-changing and non-schematic data. Big data technology also provides companies, such as telecom operators, with an ideal platform for centralizing and storing and analyzing their structured, unstructured and semi-structured data. These yield major advantages in data analysis, knowledge discovery and new business opportunity identification.

According to McKinsey Global Institute (MGI) research, big data is the key basis of competition, underpinning new waves of productivity growth, innovation and customer surplus of the future market [6]. Therefore, organizations need to have a clear idea about the characteristics and value chain of Big Data. Big Data also comes with potential challenges, which needs to be attended well before starting Big Data initiatives.

In this paper, section II describes the characteristics of Big Data. Section III describes the value chain of Big Data, which is a finding of this research. In section IV the potential Big Data challenges are presented. Finally, the paper is concluded in section IV.

II. BIG DATA CHARACTERISTICS

The characteristics of big data are well defined in the definition by Gartner. The three Vs (volume, velocity and variety) can be considered as the main characteristics of big data. These characteristics are described below.

A. Volume

Data volume measures the amount of data available to an organization; the organization does not necessarily have to own all of it as long as it can access it [7]. The number of sources of data for an organization is growing. More data sources consisting large datasets increase the volume of data, which needs to be analyzed. As data volume increases, the value of different data records decreases in portion to age, type, richness and quality among the other factors [7].

Figure 1 indicates the volume of data stored in the world by year. It also predicts that the amount of data would be more than 40 zettabytes (10^{21}) by 2020 [8].

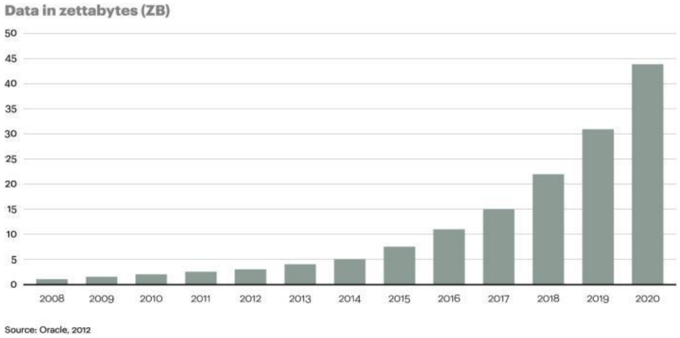

Figure 1: Data volume growth by year in zettabytes [8]

B. Velocity

Data velocity measures the speed of data creation, streaming and aggregation [7]. According to Svetlana Sicular from Gartner, velocity is the most misunderstood big data characteristic [9]. She describes that the data velocity is also about the rate changes, and about combining data sets that are coming with different speeds. She also argued that, the velocity of data also describes bursts of activities, rather than the usual steady tempo where velocity frequently equated to only real-time analytics [9].

C. Variety

Other than typical structured data, big data contains text, audio, images, videos, and many more unstructured and semi-structured data, which are available in many analog and digital formats. From analytics perspective, variety of data is the biggest challenge to effectively use it. Some researchers believe that, taming the data variety and volatility is the key of big data [2]. Data variety is also considered as a measure of the richness of the data presentation. Incomputable data formats, non-aligned data structures and inconsistent data semantics represents significant challenges that can lead to analytic sprawl [7].

Figure 2 shows the comparison between increment of unstructured, semi-structured data and structured data by years, which reflects the increment in variety of data.

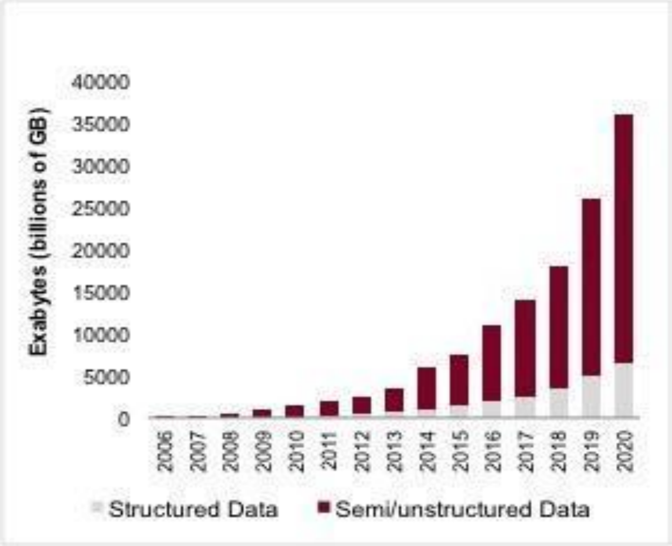

Figure 2: Growth of data variety by year

One of the big data vendors, IBM has coined additional V as the big data characteristics, which is veracity. By veracity, they address the inherent trustworthiness of the data. As big data will be used e.g. for decision making, it is important to make sure that the data can be trusted.

Some researchers mentioned „viability‟ and „value‟ as the fourth and the fifth characteristics leaving „veracity‟ out [10]. The characteristics of big data can also be described with HACE theorem. The theorem states that, "Big data starts with large-volume; **h**eterogeneous, **a**utonomous sources with distributed and decentralized control and seeks to explore **c**omplex and **e**volving relationships among data [11]. From the theorem the key characteristics are defined as:

1) Huge Data with Heterogeneous and Diverse Dimensionality: Here the „heterogeneous‟ feature refers to the different types of representations for the same individuals. The feature „diverse‟ reflects the variety of the features involved to represent each single observation.

2) Autonomous Sources with Distributed and Decentralized Control: „Autonomous‟ feature describes the ability of each data sources to generate and collect information without any centralized control.

3) Complex and Evolving Relationships: With volume of data the complexity and the correlations among them increases.

In summary, based on the characteristics described above, big data can be defined as large volume, high velocity and verities of data, which is complex to process with traditional applications, but able to bring new business opportunities to the industries by enhanced insight generation

III. Big Data Value Chain

Few decades ago, Michale E. Porter first introduced the concept of value chain, where he explained a value chain as a series of activities that create and build value as it progresses [12]. Finally, these activities culminated in total value, which the organizations then deliver to its customer [13]. In 1988, R. L. Ackoff first specified data value chain [14], which was a hierarchy based on filtration, reduction, and transformation, showing how data lead to information, knowledge and finally to wisdom. He presented Data-Information-Knowledge-Wisdom hierarchy as a pyramid which produces a series of opposing terms including misinformation, error, ignorance and stupidity when inverted [15]. Ackoff has fitted wisdom on the top of the hierarchy pyramid followed by knowledge, information and then the data or the raw data.

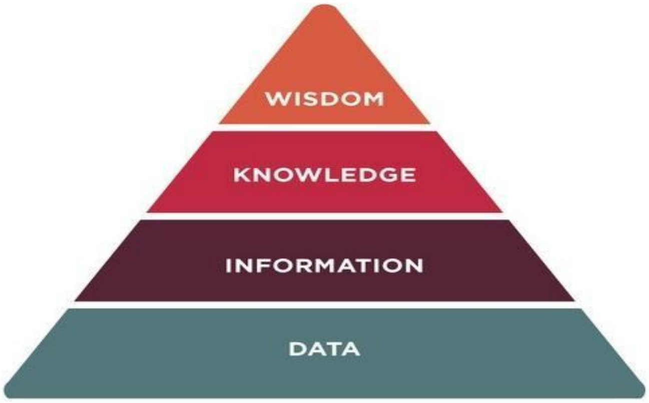

Figure 3: The Data-Information-Knowledge-Wisdom hierarchy pyramid

Table 1 below describes the four components of Data-Information-Knowledge-Wisdom hierarchy

Category	Description
Data	Data is raw. It simply exists and has no significance beyond its existence and it does not have any meaning of itself. Data can also be defined as Computerized representation of models and attributes of real or simulated entities.
Information	Information is the data that has been given meaning by way of relational connection. Information can also be defined as the data that represents the results of the computational process such as statistical analysis, providing answers to questions, such as „who‟, „what‟, „where‟, and „when‟.
Knowledge	Knowledge is the appropriate collection of the information calculated out of raw data and its intent has to be useful. Knowledge might also be defined as the data that represents the results of a computer-simulated cognitive process, such as perception, learning, and reasoning. Knowledge is the application of data and information which provides the answers to „how‟ questions.
Wisdom	Wisdom represents the ability to see the long-term consequences of any act and evaluate them relatively to the ideal of total control. Wisdom is a non-deterministic and non-probabilistic process that answers questions like „what needs to be done and why‟. Wisdom can also be defined as the process by which the outcome can be judged.

Table 1 Data-Information-Knowledge-Wisdom hierarchy

The Big Data value chain in this research is divided into three stages, naming Data sources, Preprocessing and storing, and Processing and Visualization, where each step increases value.

A. Data sources, types and accessibility

Data source is the first stage of the Big Data value chain. The data types and accessibility are also included in the sources tag because these also define the value. This step can be divided into three sub-divisions naming availability, amount and accessibility. These define the value of the data sources.

In Figure 4 below the Big Data value chain is presented. Where we can see that, difficult to easy is mentioned in this step, which means if the data from the sources is easily accessible, it has higher value.

This stage of the value chain lies under the data section of the Data-Information-Knowledge-Wisdom pyramid. For ease of graphics design, the pyramid is drawn horizontally in the value chain.

B. Preprocessing and Storing

This stage of the value chain brings the information out of the data, and belongs to the information part of the pyramid.

Value increases with the capability of colleting, loading, and preparing the data. There are different kinds of data types in big data, and capability of reading all types of data increases value. The data preparing capability also increases the value. Typically data needs to be stored in this phase, but if real-time analysis is required, data might be stored after the actual analysis.

The preprocessing step of the value chain reflects the ETL (Extract-Transform-Load) process. The ETL process is described below.

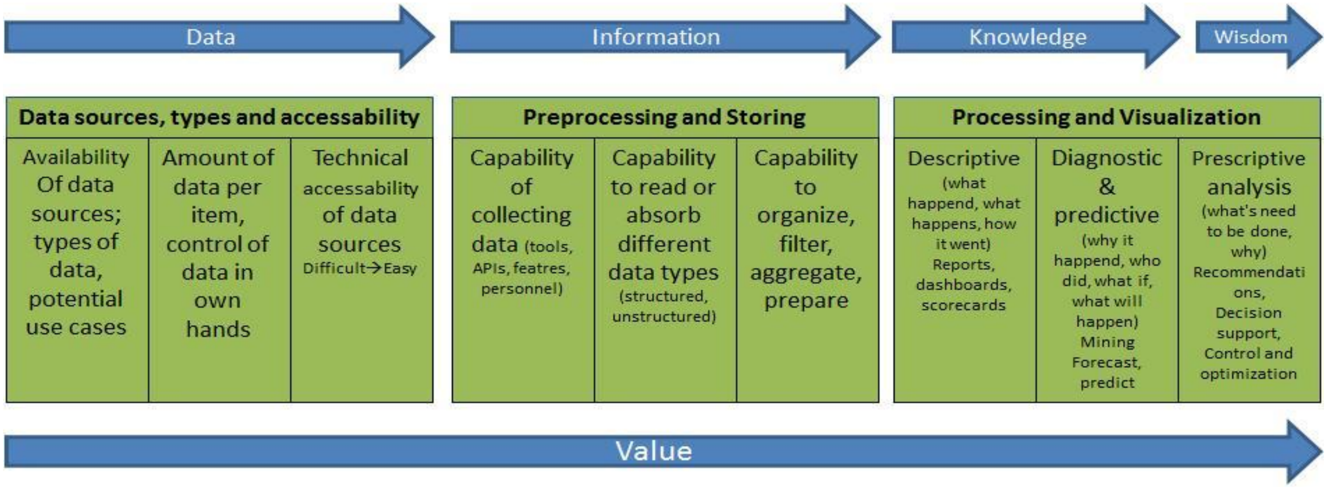
Data
Information
Knowledge
Wisdom
Data sources, types and accessability
Availability Of data sources; types of data, potential use cases
Amount of data per item, control of data in own hands
Technical accessability of data sources Difficult→Easy
Preprocessing and Storing
Capability of collecting data (tools, APIs, featres, personnel)
Capability to read or absorb different data types (structured, unstructured)
Capability to organize, filter, aggregate, prepare
Processing and Visualization
Descriptive (what happend, what happens, how it went) Reports, dashboards, scorecards
Diagnostic & predictive (why it happend, who did, what if, what will happen) Mining Forecast, predict
Prescriptive analysis (what's need to be done, why) Recommendations, Decision support, Control and optimization
Value

Figure 4: Big Data Value Chain

- **Extract** is the first step of ETL process which covers the data extraction from the source system and makes it accessible for further processing. The goal of this step is to retrieve required data from all the sources with little resources, and not to affect the process in terms of performance, response time negatively. Data extraction can be performed in several ways like update notification, incremental extraction and full extraction.
- **Transform** step cleans the data, which is important to ensure the quality of the data. When the data is cleaned then transform step applies a set or rules to transform the data from source to target. This includes several tasks, such as translating coded values, encoding free-from values, sorting, joining the data from multiple sources, aggregation and splitting according to the application requirements.
- **Load** phase loads the transformed data into the end target. Depending on the requirements of the applications, this process varies widely. Typically the target of the load phase is the databases or data warehouses. During the load step it is also necessary to ensure that the load is performed correctly with the minimal resources usage.

The ETL process framework is shown in the Figure 5 below.

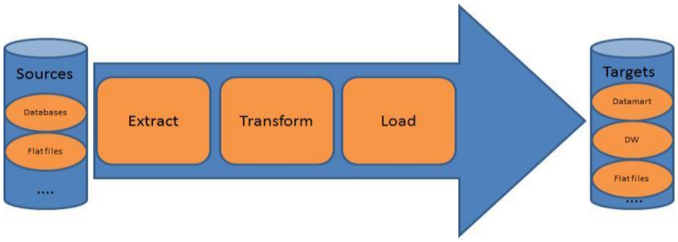

Figure 5: Typical ETL process framework

C. *Processing and Visualization*

This stage of the value chain creates the highest value and can also be called as „Analytics and Visualization‟. It lays into both knowledge and wisdom parts of the pyramid. Descriptive analysis works on past and present results and answers questions, such as what happened, what happens, and how it went. On the other hand, diagnostic and predictive analysis investigate the results and answer questions, such as why it happened, who did and what is going to happen. Both the processes increase value and bring knowledge. Prescriptive analysis works on future and includes questions, such as what is needed to be done and why, also brings wisdom. Wisdom has the highest value in the value chain [16].

The big data value chain shows the process of converting data into information, knowledge, and finally to wisdom. It also shows that every stage plays an important mediator role to enable information and generates insights from data.

IV. Big Data Challenges

Big data also has some significant challenges, some of them are mentioned below:

- **Storage**: The first and foremost challenge of big data is the storing. Traditional data warehouses are not typically made for it. Organizations that are trying to adopt big data strategy need to build a new warehouse, which is capable of storing big data for them.
- **Complexity**: The three dimensions of big data, namely volume, velocity, and variety make it more complex and challenging to analyze than the other traditional data.
- **Management**: Big data management systems available in the current market are not able to satisfy

the needs of it, thus a re-construction of the information framework is needed. Re-organizing the data in this re-constructed frame work is another big challenge.

- **Preprocessing**: Finding out the right data from big amount of data which also have verities in it, is typically challenging. Big data preprocessing requires collection capability, statistical analysis, and integration capability of large amount of data. Traditional extract-transform- load (ETL) tools are not able to fulfill these requirements.

- **Analytics**: Big data analytic is a highly mathematics intensive analytic modeling exercise which requires proper tools and skilled people. Big data also requires highly capable tools for data visualization, because traditional tools are typically made for small amount of data.

- **Utilization gap**: Christine Moorman stated that the biggest challenge regarding big data is the Utilization gap [17]. When asked to report the percentage of project in which their companies are using marketing analytics that are available, CMOs report a dismal of only 30% usage rate [17]. In [18], the hardest challenge of big data is mentioned as, taking the insights generated from the analytics and utilizing them to change the way business operates.

- **Lack of skilled people**: As big data is a new concept and requires newer technologies; there is a lack of skilled people for it. According to Gartner, big data demand will reach around 4.4 million jobs globally by 2015, with two third of these positions remaining unfilled [19].

- **Privacy**: In several research studies, privacy concern has defined as the biggest barrier for big data [7], [20]. When it comes to the customer personal data and how it is used, people generally don‟t like surprises. The study [20] shows, how the location data and the social media data are hampering users‟ privacy, and the users are not concerned about it. The social media data is being one big topic about the users‟ privacy issue in recent days but the user location data being as a privacy issue has not got that much attention yet [20].

- **Security**: Big data security management is also one challenging task. Traditional security mechanisms, which are tailored to secure the small-scale data, are inadequate for it.

- **Real-time analysis**: Big data requires real-time analysis, which is sometimes challenging. Real-time analysis requires high-velocity streaming analysis of big amount of data and typical data analysis tools are incapable of doing so.

- **Additional challenges**: There are more additional challenges regarding big data, such as transportation of data, dynamic design requirement, and scaling.

An organization, before starting big data projects needs to make new policies to mitigate these challenges, and select tools which are truly capable for it. Otherwise, the project acquires a large possibility to be failed in the middle of the process which will cause the organization financial loss**.**

V. Conclusion

In this paper the basic characteristics of Big Data is described and then a summarized definition of Big Data is presented. This paper also presents a Big Data value chain as a result of this research. The value chain defines how the value of data increase as the process continues. The challenges of Big Data this paper presents are based on internet research and literature study. The challenges may also vary from organization to organization depending on their resources and data sizes. As the field of Big Data is still in a great state of change and development, the possibilities for research in this area are extensive and the interest on the findings will certainly rise if the adoption of big data rises as expected.

VI. Bibliography

[1] The Economist, "Data, data everywhere," *The Economist*, February 2010.

[2] Infosys, "Big Data: Challenges and Opportunities," 2013.

[3] Eric Schmidt. (2010, August) techcrunch. [Online]. http://techcrunch.com/2010/08/04/schmidt-data/

[4] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*, San Diego, CA, 2013, pp. 42 - 47.

[5] Mark A. Beyer and Douglas Laney, "The Importance of 'Big Data': A Definition," Gartner, Analysis Report G00235055, 2012.

[6] McKinsey & Company, "Big Data: The next frontier for innovation, competition, and productivity," 2011.

[7] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big Data: Issues and Challenges Moving Forward," in *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, Wailea, Maui, HI, 2013, pp. 995-1004.

[8] ATKearney. (2013, January) Big Data and the Creative Destruction of Today's Busnsiness Model. [Online]. https://www.atkearney.com/strategic-it/ideas- insights/article/-/asset_publisher/LCcgOeS4t85g/content/big-data-and-the-creative-destruction-of-today-s-business-models/10192

[9] S. Sicular, "Gartner's Big Data definition consists of three parts," *Forbes*, March 2013.

[10] Neil Biehn. (2013) WIRED. [Online].

http://www.wired.com/insights/2013/05/the-missing-vs- in-big-data-viability-and-value/

[11] X. Wu, X. Zhu, G.Q. Wu, and W. Ding, "Data Mining with Big Data. Knowledge and Data Engineering," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 1, pp. 97-107, June 2013.

[12] M.E. Porter, *The Competitive Advantage: Creating and Sustaining Superior Performance*. New tork, USA: Free press, 1985.

[13] H. Miller and P. Mork, "From Data to Decisions: A Value Chain for Big Data," *IT Professional*, vol. 15, no. 1, pp. 57-59, February 2013.

[14] R. Ackloff, "From Data to Wisdom," *Journal of Applied Systems Analysis*, vol. 16, pp. 3-9, 1989.

[15] Jay H. Bernstein, "The Data-Information-Knowledge-Wisdom Hierarchy and its Antithesis," *North American Symposium on Knowledge Organization*, pp. 65-75, 2011.

[16] Andrew Stein, "Big Data and Analytics, The Analytics Value Chain – Part 3," *Steinvox*, October 2012.

[17] C. Moorman, "The Utilization Gap: Big Data's Biggest Challenge," *Forbes*, March 2013.

[18] McGuire. (2013, March) Youtube. [Online]. https://www.youtube.com/watch?v=Sc5FFY-IVDQ

[19] Gartner, "Gartner Reveals Top Predictions for IT Organizations and Users for 2013 and Beyond," 2012.

[20] M. Smith, C. Szongott, B. Henne, and G. Von Voigt, "Big data privacy issues in public social media," in *Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on*, Campione d'Italia, 2012, pp. 1-6.

[21] Olaf Acker, Adrian Blockus, and Florian Pötscher, "Benefiting from big data: A new approach for the telecom industry," Booz & Company, Analysis Report 2013.

[22] Abdelghani Bellaachia, *Data preprocessing*, 1st ed. Washington, USA, 2011.

[23] D Tanasa and Antipolis Sophia, "Advanced data preprocessing for intersites Web usage mining," *Intelligent Systems, IEEE*, vol. 19, no. 2, pp. 59 - 65, March 2004.

[24] Fazel Famili, Wei-min Shen, Richard Weber, and Evangelos Simoudis, "Data Preprocessing and Intelligent Data Analysis," *Intelligent Data Analysis*, vol. 1, no. 1, pp. 3-23, 1997.