Characteristics of Big Data – A Delphi study

by

© Raja Rajeshwari Sreenivasan

A thesis submitted to the

School of Graduate Studies

in partial fulfillment of the

requirements for the degree of

Master of Science in Management

Faculty of Business Administration

Memorial University of Newfoundland

December 2017

St. John’s Newfoundland

Abstract

Big Data has become increasingly important to organizations in various industries. Big Data is collected, stored and processed to obtain valuable insight and to support decision making and action. However, the question “what constitutes Big Data?” does not have a clear or singular answer. As Big Data is a developing phenomenon, concepts and constructs such as 3Vs, 4Vs, 5Vs, 6Vs, 7Vs and 9Vs have introduced by organizations and researchers. As a result, there is a lack of consensus on what constitutes the core characteristics of Big Data. Using four rounds of Delphi study, followed by a verification survey, this research aims to identify the core characteristics of Big Data as determined by practitioners and researchers and, among the key characteristics, which are deemed to be more important Though Volume, Velocity, and Variety are frequently mentioned as the 3Vs of Big Data, this study identifies Volume, Value, and Velocity to be the top three important characteristics of Big Data. Additionally, Visualization emerged as an important characteristic ranked much higher than Variety. We hope that the outcome of this study can contribute to better understand the key characteristics of Big Data.

Keywords: Big Data, Delphi method

Acknowledgement

I would like to express my sincere gratitude to my supervisor Dr. Jeffrey Parsons for his guidance and support to complete my Master's thesis.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1. Background and importance of the problem

In general, Big Data refers to "datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within a tolerable time" (Chen et al., 2014, p.173). The term "Big Data" is nebulous; it may mean different things for different organizations (De Mauro et al., 2015; Gandomi & Haider, 2015). Because of its rise in various fields, research on Big Data is quickly evolving; as it is a well-known area of developing research, concepts and constructs in Big Data are emerging. The lack of formal definition for the term "Big Data" gave rise to research with an aim to form a consensual definition.

Ronda-Pupo & Guerras-Martin (2012, p.163) suggest, "the level of consensus shown by a scientific community on a definition of a concept can be used as a measure of the progress of a discipline". When we consider Big Data, there is no single universally accepted definition; instead, there exist multiple definitions for Big Data proposed by various authors and stakeholders. The lack of a consensual definition means that many authors challenged or ignored previous definitions and proposed new ones. The existence of multiple and often contradicting definitions has also hampered our understanding of how Big Data is changing organizations and society.

Ward & Barker (2013) suggest that the lack of a formal definition has led research to evolve into multiple and inconsistent paths. They aimed to provide a consensual definition by combining common themes from existing works and previous definitions and created the following definition: "Big Data is a term describing the storage and analysis of large and complex data sets using a series of techniques, including but not limited to: NoSQL, MapReduce and machine learning" (Ward & Barker, 2013, p.4).

Dutcher (2014) questioned more than 40 leaders to define the phrase "Big Data" and found that, apart from three Vs (volume, velocity and variety) other characteristics such as tools used for analysis, insights gained, visualization, and processing methods were included in the Big Data definition. De Mauro et al. (2015) proposed a consensual definition for Big Data: "Big Data represents the information assets characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value" (De Mauro et al., 2015, p.103).

A mapping study conducted by Ylijoki and Porras (2016) analyzed the concept of Big Data and suggested that the characteristics such as value and velocity "do not define the characteristics of Big Data, but instead they reflect the usage of the data" (Ylijoki & Porras, 2016, p.77). Added to that, other aspects of Big Data such as technical, privacy, security and policy making are not characteristics of Big Data and Ylijoki & Porras suggested not to include them in the Big Data definition. To achieve clarity and coherence in Big Data definition, Ylijoki & Porras suggested that data and its usage should be considered separately.

Large sets of data (volume), which is considered to be one of the characteristics of Big Data, can be generated from various areas, including scientific research, health care, sensor data, business apps, and social media. The data generated are not only large in volume, but also varied in structure, generated at a rapid rate, and are difficult to handle by traditional database methods. Big Data in its raw form is not of much use, but when analyzed yields valuable results.

In 2001, Doug Laney, an analyst at the META Group, proposed the well-known 3Vs (volume, velocity, and variety) commonly associated with the term Big Data. The 3Vs were described as, "Volume: which means incoming data stream and cumulative volume of data; Velocity: which represents the pace data used to support interaction and generated by interactions; and Variety, which signifies the variety of incompatible and inconsistent data formats and data structures" (Gopinath et al., 2016, p.171).

Following that, IBM proposed 4Vs, (volume, velocity, variety and veracity). "Volume stands for scale of data; Velocity denotes to analyzing streaming data; Variety indicates different forms of data; Veracity implies uncertainty of data" (Gopinath et al., 2016, p.171). Added to that, Microsoft introduced two additional Vs, (variability and visibility). Variability refers to the complexity of data set. In comparison with "Variety" (or different data format), it means the number of variables in data sets; Visibility emphasizes that you need to have a full picture of data to make an informative decision (Gopinath et al., 2016). Well known as a developing area of research, the concepts and constructs in Big

Data are used every day by organizations and researchers. Not surprisingly, there is a disagreement about what constitutes the essential characteristics of Big Data.

The characteristics of Big Data mentioned in prior research articles are Volume, Velocity, and Variety (Laney, 2001); Veracity, Variability, Visualization, and Value (Mcnulty, 2014; Rijmenam, 2015); Viscosity, Virality, Volatility, Validity, and Viability (Biehn, 2013; NIST, 2014; B. Vorhies, 2014). Added to those characteristics, other characteristics have been introduced in blogs and web discussion. Using a Delphi research methodology, this research aims to identify experts' views on the core characteristics of Big Data.

1.2. Importance of this research

In general, research on Big Data is diverse and much importance is given to Big Data's definition, analytics, and developing tools for storing and analyzing the data. Less emphasis is given to the question "what are the characteristics of Big Data?" This research aims to identify core characteristics of Big Data from the perspective of Big Data practitioners and researchers. The questions addressed in this study are:

1. What are the characteristics of Big Data from the view of the Big Data practitioners and researchers?
2. What is the relative importance of these characteristics?

1.3. Thesis organization

The remainder of the thesis is organized as follows:

- Chapter 2 provides an overview of Big Data and its importance
- Chapter 3 provides a literature review of Vs in Big Data
- Chapter 4 presents the research methodology used in this research
- Chapter 5 discusses the finding of this research
- Chapter 6 summarizes the findings and their implications

Chapter 2: Scope of Big Data

2.1. Data growth

The actions performed in devices such as laptops, mobile phones, and tablets are converted into data by "digitization" and "datafication" (Ylijoki & Porras, 2016). Digitization is the process of converting text, images or sound into digital format (Parekh, 2001) and datafication is the process of connecting the digital data generated from devices and sensors to the internet (Mayer-Schnberger & Cukier, 2012). The amount of data resulting from the proliferation of devices and sensors is predicted to double every two years (IDC, 2014). In 2015, annually less than 10 Zettabytes of data were created and it is expected that the data growth will reach 44 Zettabytes per year in 2020 and 180 Zettabytes per year in 2025 (Press, 2016). The exponential increase in data generation is mainly because of the growth of the "web" and "internet of things (IoT)" (Uddin et al., 2014).

In the digital economy, data is vital for an organization as business stakeholders rely on data to derive valuable insights. The data collected in an organization can be classified as structured and unstructured.

- ***Structured data***: Data that is highly organized and strictly bound by the file and record structure. Structured data can be easily stored, updated and searched using standard algorithms. Structured data can be managed effectively by traditional

database systems. An example of structured data can be basic customer data which consists of name, addresses, and contact information.

- ***Unstructured data***: Data generated from a variety of sources and has varied structure, such as text documents, social media posts, emails, logs, videos, images, audio files, and sensor data, is referred to as unstructured data. Unstructured data are complex and do not fit inside a rigid data model like structured data. Unstructured data are usually voluminous in the range of petabyte (250 byte), exabyte (1018 byte), zettabyte (1021 byte), and yottabyte (1024 byte) and are generated at a high velocity so they cannot be handled or efficiently queried by a traditional database

Oracle has categorized Big Data into three types (Oracle, 2013)

- "Traditional enterprise data – customer information from CRM systems, transactional ERP data, web store transactions, and general ledger data.

- Machine-generated/sensor data – Call Detail Records ("CDR"), weblogs, smart meters, manufacturing sensors, equipment logs (often referred to as digital exhaust), and trading systems data.
- Social data – customer feedback streams, micro-blogging sites like Twitter, and social media platforms like Facebook."

2.1.1. Web-generated Big Data

Web 1.0 was static and mainly acted as an information provider. Businesses used Web 1.0 (Aghaei et al., 2012) as an advertisement platform and information provider. The users of Web 1.0 were passive, as they could only read the information present. Unlike Web 1.0, Web 2.0 was interactive (Prabhu, 2016) where users acted not only as an information receiver, but were also involved in content contribution and modification. The introduction of online social media like social networking services, forums, and blogs, enabled users to create, upload, and share content (Chen et al., 2014).

The tools of Web 2.0 are as follows (Prabhu, 2016):

- Social networks (Example: Facebook, LinkedIn, Twitter - where people can interact with friends and family, or use for professional purposes. People can post their photos, videos, comment, like, and share).
- Blogs - people share their thoughts on various issues using text, photos, and videos.
- Video sharing (YouTube) - people can upload, download, comment and share videos.
- Photo sharing (Example: Flickr: people can share and comment on photos).
- Wiki - users can edit the content.
- Podcast - digital media files (audio or video) downloaded from the internet through iPod or iPad- like devices).

- Virtual world - computer-based simulation environment.
- Content rating - any user reading the content on the website can rate the content.
- Widgets - stand-alone application in which users can turn their personal content into a dynamic web app which can then be shared on any website.

Web 3.0 opened up opportunities such as "personalization" and "recommendation" (Pattal, Li, & Zeng, 2009). The browsing behavior of the user, such as mostly viewed, recent search, likes and dislikes, browsing pattern, and mouse clicks, are stored and analyzed in order to make recommendations and personalization possible. Examples are: social network sites such as Facebook, Twitter, Pinterest, and LinkedIn; e-commerce sites like Amazon, and eBay; entertainment sites like Netflix and YouTube. The usage of internet around the world has tremendously increased; (Schultz, 2016) reported that internet usage has increased by 60 percent when compared with 2011. This increase is mainly because of the revolution in mobile technology. Initially, only computers were used to access the internet, but with the development and innovation of the mobile phone, the number of people having access to the internet drastically increased (Prabhu, 2016). In addition to the calling and texting options, mobile phones allow users to connect to the internet and provide options for utilizing the internet (send emails, read e-books, upload and download photos and videos, playing games online, video calling, and a number of apps for social networking, education, entertainment, and health) (Zhang, 2012). Lister (2017) reported that 80% of the time spent on social media is through mobile devices. In the case of YouTube, more than half of the views are from mobile devices (Donchev,

2017; Lister, 2017) and 80% of Twitter users are using mobiles to access the site (Aslam, 2017b). The Radicati Group, Inc. (Technology market research firm) conducted a research to put together statistics and forecasts for email and mobile email use from 2015 to 2019. They reported that over 205 billion emails are being sent and received per day and also predicted that by 2019 there will be an increase and reach over 246 billion emails per day (Team, 2015). Table 1 represents the statistical summary of data created by the users in the social media.

Schultz (2016) indicated that the amount of data created by users each day on the internet in 2016 was as follows:

- 6 billion daily Google Searches
- 500 million tweets sent each day
- More than 4 million hours of content uploaded to YouTube every day
- 3.6 billion Instagram likes each day
- 4.3 billion Facebook messages posted daily

Facebook (Lister, 2017; "The Top 20 Valuable Facebook Statistics," 2017)	• 22% of the world's total population uses Facebook • Every day, 100 million hours of video content are watched • 300 million photos are updated per day • Every 60 seconds, 510,000 comments are posted, and 293,000 statuses are updated, • Every second, five new profiles are created
YouTube (Donchev, 2017)	• YouTube has 1,300,000,000 users • Every minute, 300 hours of video are uploaded • Every day, almost 5 billion videos are watched • YouTube has 30 million visitors per day
LinkedIn (Aslam, 2017b)	• LinkedIn has 467 million users • On a weekly basis, 3 million users share content • Over 19.7 million slide share presentations have been uploaded
Twitter (Aslam, 2017c)	• Twitter has 317 million users • 500 million tweets are tweeted per day
Instagram (Aslam, 2017a)	• 95 million photos are uploaded per day • 40 billion photos are shared per day • So far more than 40 billion photos have been uploaded to Instagram
Snapchat (Lister, 2017)	• 400 million snaps are shared on Snapchat per day, • 9,000 photos are shared every second

Table 1: Statistical summary of data created by users

2.1.2. Internet of things-based Big Data

The definition of internet of things (loT) given by (Xu, He, & Li, 2014, p.2233) states that, "IoT can be considered as a global network infrastructure composed of numerous connected devices that rely on sensory, communication, networking, and information processing technologies". The two foundational technologies for IoT are RFID (radio frequency identification) and WSNs (wireless sensor networks). Other technologies and devices used are barcodes, smartphones, social networks, GPS, and cloud computing (Xu, He, & Li, 2014). Some examples of IoT products as mentioned by (Rose and Eldridge, 2015) are internet-enabled appliances, home automation components, energy management devices, wearable fitness and health monitoring devices, network-enabled medical devices, networked vehicles, intelligent traffic systems, and sensors embedded in roads and bridges. For example the intelligent transportation system of a big city generates 144 Gigabytes of data every day (Ma et al., 2012), the Ford GT (super car) generates up to 100 GB of data per hour (Kanellos, 2016), and it has been estimated that for every hour 25 gigabytes of data is generated from the connected cars to the cloud (Quartz, 2015).

2.2. Importance of Big Data

Big Data, because of its volume and variety of data structure, is hard to store, retrieve, and analyze using traditional relational database management methods. Challenges associated with Big Data are "capturing data, curation, storage, searching, sharing, transfer, analysis, presentation" (Kaur et al., 2016, p.55). Big Data in its raw form is not of much use, but when analyzed, can yield valuable results. In recent years, Big Data has gained interest and application in various fields such as government, healthcare, retail sector, agriculture, research, online and social media, telecom industry, and banking. The increasing interest around Big Data is mainly because it provides valuable insight to organizations to make judgments, recommendations and decisions (Chen et al., 2014). Following are some example Big Data applications.

Healthcare

In health care, the sensors used in home monitoring devices for chronically ill patients collect the vital details of the patient and make it available to the physician for analysis, thereby reducing the need for visits to the hospital and admittance (Dijcks, 2012). Big Data analytics in health care is used to provide personalized health care and also acts as a valuable resource to control and prevent the spreading of disease (Pai, 2017). The clinical data available from Electronic health records (EHRs) and the data available on social media sites are analyzed and used for public health surveillance. For example, in diabetes surveillance data from EHR (such as laboratory report, physical exam result,

medication used) and social media data are used to identify the population at high risk for diabetes and also in disease management (Eggleston & Weitzman, 2014).

Government

Governments use Big Data analytics to improve their services to the country/citizens by addressing challenges in the country such as "health care, terrorism, job creation and natural disaster" (Kim et al., 2014, p.78). One of the applications of Big Data is the concept of smart cities, which "utilize multiple technologies to improve the performance of health, transportation, energy, education and water services leading to higher levels of comfort of their citizens" (Al Nuaimi et al., 2015, p.1).

Agriculture

Precision farming, as described by (Herring, 2001), refers to the use of information and technology-based systems for within-field management of crops. Using Global Positioning System (GPS) and Geographical Information Systems (GIS), farmers receive information about the soil condition, plant growth, sunlight, amount of pesticides and insecticides to use, and water requirements. Sensors are used in agricultural field to gather data such as temperature, water composition, and soil fertility. Data gathered by sensors is analyzed to discover optimal environmental conditions to promote plant growth (Sravanthi & Reddy, 2015).

Retail and Telecom

Companies in the retail sector analyze their customer's data to provide a personalized shopping environment for individual users (Singh & Singla, 2015). Telecom industries store and analyze customer service logs, recorded call details, and customer emails to understand and improve the customer service and satisfaction (Mukherjee & Ravi, 2016).

Detection purpose

Huge volume of customer transaction data is collected by banks and other financial institutions. Fraud detection can be made by analyzing customer transaction data. By doing so, the safety of both the customers and the bank are ensured (Mukherjee & Ravi, 2016). High-frequency sensor data generated from engines and machines of cars, wind mills, and pipelines are stored and analyzed to understand the performance and also to diagnose any problems (Su, 2013).

User generated data

User-generated content in the form of reviews, feedbacks, and ratings are used by both businesses and customers. Businesses utilize user-generated content to understand their customer satisfaction and business growth. Customers read the reviews and feedback of the product posted by the product users before making purchases. In everyday life, reviews and comments are used for various purposes, including planning a trip, choosing a restaurant, and finding a medical doctor. One such extended application of user-generated content gave rise to a novel approach called citizen science. Citizen science

refers to “participation of lay citizens in design, funding, data collection, analysis, report or dissemination of scientific research” (Del Savio et al., 2016, p.1). Citizen science projects have been conducted in areas like astronomy, botany, zoology, environmental science, ecology, entomology, ornithology, phenology, seismology, herpetology, computer science, art history, cetology, climatology, health, wild life, economy, humanities, sociology, geography, marine, and metrology (“List of citizen science projects,” n.d.).

Recommendation to users

Big Data analytics is used extensively by various industries to provide recommendation to users as follows:

- Web log data of customers are stored and analyzed to understand shopping behavior (Kalota, 2015). By doing so, a personalized shopping environment can be given to individual customers. Examples include eBay and Amazon. These companies analyze the buying behavior of individual customers and provide recommendations/suggestions of products that the customer might be interested in. This is also how recommendations are suggested to the viewers on Netflix and YouTube.
- In the case of food industry, Big Data is used to track the quality of the food, and also to provide recommendations for the individual user. Big Data analytics are used in companies such as Starbucks, Dominos, and Subway, where the

customer's data are analyzed and personalized offers are made to the customers (Mukherjee & Ravi, 2016).

- Time and location data are collected and analyzed to provide timely suggestions to the users (Su, 2013). For example, if the user is on a road trip, the details about a nearby gas station, restaurant, coffee shop and other relevant information is suggested based on the time and the geographical location of the user.
- Social network data such as the data from Facebook, Twitter, and LinkedIn are analyzed to understand the user's interests and the recommendations are suggested based on that. For example, if a Facebook user is interested in sports and recreation, groups related to sports and recreations and related feeds are suggested to that particular user. In the case of LinkedIn, job suggestions are generated to the individual user based on their personal interest (Su, 2013).

As discussed above Big Data has valuable application in various industries. The International Data Corporation (IDC) estimated that organizations using Big Data – such as Government (Federal/Central), Professional services, Telecommunications, and Retail – are estimated to generate revenue of more than $10 billion in 2019 (IDC, 2016). The same report also predicted faster revenue growth in industries such as Banking, Healthcare, Utilities and Resource industries. According to Kisker (2015), Big Data is like crude oil - it needs filtering and refining to unlock its value and make it usable. So in order to analyze Big Data and to implement suitable Big Data analytics it is important to understand Big Data characteristics.

Chapter 3: Proposed Characteristics of Big Data

3.1. Vs in Big Data

In 2001, Doug Laney, an analyst in META Group proposed 3Vs of Big Data (Laney, 2001). Laney highlighted the data explosion as the result of E-commerce, and put forth the data management challenges along three dimensions: "Volume, Velocity and Variety." These three characteristics are widely accepted and have been reported in both academic and practitioner literatures. Other potential characteristics of Big Data are "Veracity, Variability, Visualization, and Value," which have been consolidated from various sources (McNulty, 2014; Rijmenam, 2013; Uddin et al., 2014). According to (Gandomi & Haider, 2015), other organizations which introduced additional Vs to Big Data are as follows: IBM coined "Veracity", Oracle introduced "Value" and SAS introduced "Variability". In addition to the above mentioned widely used seven Vs, other Vs have been reported (Neil Biehn; NIST, 2015; Vorhies, 2013) as characteristics of Big Data: namely, "Viscosity, Virality, Volatility, Validity, and Viability." The characteristic of Big Data is an evolving research area and new characteristics are being introduced by academic researchers and organizations.

The V's grouped in literature are as follows:

- **3Vs - Volume, Velocity, Variety** (Laney, 2001)
- **4Vs - Volume, Velocity, Variety, Value** (Chen et al., 2014; Thiyagarajan & Venkatachalapathy, 2014)
- **5Vs - Volume, Velocity, Variety, Value, Veracity** (Anuradha & Ishwarappa, 2015; Hadi et al., 2014; Knilans, 2014; Rowe, 2016)
- **6Vs - Volume, Velocity, Variety, Veracity, Validity, Volatility** (Yassin, 2014)
- **7Vs - Volume, Velocity, Variety, Veracity, Value, Validity, Volatility** (Uddin et al., 2014)
- **9Vs - Volume, Velocity, Variety, Veracity, Value, Variability, Validity, Volatility, Visualization** (Owais & Hussein, 2016)

Next, I examine how each of these characteristics has been defined and used.

3.1.1. Volume

Volume refers to the size of the data; it is one of the characteristics which differentiates data from Big Data and is the reason why traditional relational database management systems cannot deal with Big Data (W. Vorhies, 2014). The data generated every day are increasing exponentially (Shukla et al., 2015). It is estimated that 90% of the data we have today was created in the past two years (Geczy, 2014). For example:

- Every hour Wal-Mart generates 2.5 petabytes of customer transaction data (Zaslavsky et al., 2012)

- In a span of 30 minutes, a single jet engine generates 10TB of data (Dijcks, 2012);
- Each year the sensors in airplanes generate 2.5 billion terabytes of data and self-driving cars generate 2 petabytes of data (Rijmenam, 2013);
- Sensors used in oil wells generate 10 exabytes of data annually;
- The Square Kilometer Array telescope generates 1 Exabyte of data daily;
- Every day by storing 16 petabytes of data, UPS tracks and captures 16.3 million packages and responds to 39.5 million tracking requests (W. Vorhies, 2014).

It has been estimated that by 2020, 40 zettabytes of data will be created, and that will represent 300 times the amount of data that was generated in 2005 (Mobertz, 2013). The definition for volume given by Afshin Goodarzi, chief analyst at 1010data Inc. is, "volume is the number of rows, or the number of events that are happening in the real world that are getting captured in some way, a row at a time" (Rowe, 2016, p.29). This huge volume of data is generated from various areas (Zaslavsky et al., 2012), including meteorology, genomics, physics, simulations, biology, environmental science, sensors collecting real time information (Su et al., 2016) web data (customer web behavior), text data (email, news, documents) (Mukherjee & Ravi, 2016) smart phones, e-commerce (Thiyagarajan & Venkatachalapathy, 2014) social networks (Facebook, LinkedIn, Instagram), web logs, satellite images, broadcast audio streams, banking transactions, web page contents, scans of government documents, GPS, financial market data (Pai, 2017), military surveillance, natural disaster and resource management, private sector, retail, RFID, call detail recording, genomics, biogeochemical, life science advanced health care system and

insurance, and stock exchanges (Uddin et al., 2014). Moreover, what is considered to be Big Data today may not be the Big Data tomorrow (Gandomi & Haider, 2015; Zaslavsky et al., 2012).

3.1.2. Velocity

As defined in Al Nuaimi et al., (2015), velocity is the speed at which the data is generated, stored, analyzed and processed. In the case of Big Data, the data is generated at a very high velocity, and thus it is challenging to analyze. There are several examples of high velocity data generation: in each session of the New York stock exchange, around one terabyte of trade information is captured (Mobertz, 2013). On average, more than 40,000 search queries are processed by Google every second (Firican, 2017). Air quality monitoring sensors sense and send data every minute, and smart meters are used to get a detailed level of minute to minute energy consumption of each device (Gary Barnett, 2014). Telescopes used for research purposes like astronomy, genome sequences generated in biological research, weather forecasting satellites, IoT connected with home appliances, and smart cities, all generate data in a very high velocity.

Advancement in the devices, such as smart phones and sensors, enhances the rapid generation of data (Gandomi & Haider, 2015). Data are generated at a high rate from mobile phones through mobile apps provide valuable information to the seller, such as the demographical location of the customer and buying patterns. Thus, by analyzing this information, the seller can increase the sales value by targeting the right customer with

the right product. One such example is the ads that pop up when a user uses a sophisticated content website, such as Amazon or Netflix. Ads such as these are selected and displayed for that particular user by capturing, analyzing and storing users' web behavior (W. Vorhies, 2014). The data generating at this high velocity turns into voluminous data. Data from social media like Facebook and Twitter are generated at a great velocity at every second (Dijcks, 2012).

3.1.3. Variety

Variety refers to the different data structures which are generated (De Mauro et al., 2015). In the case of traditional data, the data generated are structured, whereas in Big Data, the data generated are often semi-structured or unstructured, like images, audio, video, transactions, and log data (Owais & Hussein, 2016). For example, the advancement in internet enabled social media, such as Facebook, Twitter, and Instagram to generate data in the forms of pictures (JPEG, GIF, 3D), videos, audio, and increased the usage of hash tags.

Sensors are used in research and industry for various purposes, and they generate different varieties of data such as temperature, air quality, water quality, traffic speed, and goods movement. Examples of unstructured data are images generated from satellites for weather forecasting or surveillance purposes, data generated for scientific research purposes such as tide/wave movements, earth rotation patterns etc., videos generated for security, and traffic purposes (Hurwitz et al., 2013). For example, car sensors generate a variety of data such as speed, temperature of the engine, and fuel level. Sensors in

smartphones such as Global Positioning System (GPS) generate data about the geolocation (Davis, 2015). Sensors used in the agricultural field generate a variety of data such as water levels in the soil, fertility of the soil, the level of pesticides and insecticides required, as well as plant growth pattern data (Herring, 2001). One application of IoT is wearable devices, where sensors generate a variety of data about a person such as sleeping patterns, heartbeat, walking speed, food intake, water consumed, and calories burned. In the case of e-commerce, the actions performed by the customer on the website, such as time spent, browsing pattern and purchase history are all stored as data for analysis purposes.

3.1.4. Veracity

Gandomi & Haider (2015) note that IBM coined "veracity" as the fourth V of Big Data. (Uddin et al., 2014, p.3) define veracity as "truthfulness of data." If the data is not trustworthy, analyzing such data will not yield valuable results. The quality of the data determines the accuracy of the data analysis. For example, sensors are used in various fields (IoT, air quality monitoring, weather forecast, astronomy, and for various other purposes) (Anuradha & Ishwarappa, 2015). If the data generated from these sensors are not truthful, then analyzing the data will not generate a reliable result for decision making.

Veracity refers to the biases, noise, and abnormality in data (Owais & Hussein, 2016; Yassin, 2014). Big Data are collected, stored and analyzed because they provide valuable insight to organizations for making judgments, recommendations, and decision-making

(Chen et al., 2014). In Traditional Business Intelligence solutions, data is often acquired from one or more business systems. The data extracted is highly structured and comes from a reliable source system. The data acquired is cleansed and integrated in to a repository such as a Data warehouse or Data mart. In case of Big Data, the data is acquired from a variety of sources and it can be both structured and unstructured.

Human generated data (emails, videos, photos, text messages etc.) and machine generated data (log files, sensor data etc.) are all included for analysis. Consider the scenario where organizations need to aggregate structured data like their customers' information from business systems with unstructured data like customer emails, tweets etc. There can be threats to data veracity as the external sources of data can be unverified, or the data itself like a tweet or email can be incomplete or misleading. Organizations must take into account the level of uncertainty of the data sources used for analysis during their decision making process.

For instance, consider the case where a business launched a new product and would like to have a real-time analysis of how the product was received by the customers. Sentiment analysis can be used to collect and analyze customer opinions about a particular product or service. The data collected is mostly unstructured data like tweets, comments and posts gathered from various social media platforms. Even though such real-time analysis can be a key factor for success, the variety of data sources and the velocity at which social media data is generated leads to a certain level of uncertainty and incompleteness in data.

The veracity issues like uncertainty, incompleteness and trustworthiness of the data must be addressed before making any decisions based on the data analysis

3.1.5. Value

As mentioned by (Gandomi & Haider, 2015), Oracle introduced value as a defining attribute of Big Data. Big Data is stored and analyzed for a purpose, and that purpose depends on the organization and the kind of data collected. For example, data from the continuous monitoring of a person in an intensive care unit is tracked and analyzed for life-saving purposes, whereas data generated from a wearable device can monitor walking speeds, and sleeping patterns which can be analyzed for a healthy lifestyle. Data are stored and analyzed for their intended purposes. Anuradha & Ishwarappa (2015) refer to the value of Big Data as the extent to which the data collected, after analysis, can contribute to the intended purpose. As mentioned by (Kaisler et al., 2013), data value is the measure of usefulness of the data in decision making. In addition, (Wu et al., 2016, p.7) indicated that value answers the question, "Does the data contain any valuable information for my business needs?" Anuradha & Ishwarappa (2015) consider value to be the most important aspect of Big Data, because if the Big Data collected does not hold value, then it is considered to be useless. Thiyagarajan & Venkatachalapathy (2014) described value as the desired outcome of Big Data processing.

Examples of the value of Big Data in different organizations are: in the banking sector, Big Data analysis is used for fraud detection, and illegal trading; in the US Food and Drug Administration (FDA), Big Data is used to detect and understand illness and

diseases caused by food; in retail and wholesale, trading Big Data is used for inventory analysis, and to understand customers' shopping patterns; in the transportation industry, Big Data provides value by route planning and traffic control (Sivakumar, 2015). (Wamba et al., 2015, p.240) define value as "the extent to which Big Data generates economically worthy insights and/or benefits through extraction and transformation." International Data Corporation (IDC) estimated revenue opportunity for the industries in 2019 are: discrete manufacturing ($22.8 billion), banking ($22.1 billion), process manufacturing ($16.4 billion) and more than $10 billion in retail, telecommunication, federal government, and professional services.

3.1.6. Variability

SAS introduced the term variability as an additional dimension of Big Data which represents the variation in the rate of data flow (Gandomi & Haider, 2015). Owais & Hussein (2016) define variability as the inconsistency in the data flow rate. For example, consider social media where the rate of the data flow is highly inconsistent. In the case of Facebook and Twitter, people generate data almost every minute in the form of pictures, videos, and audio, but the rate of data generated is not the same all the time. The data generation rate peaks during some events like an election, natural disaster, etc. Microsoft referred to variability as the "complexity in the data set" (Gopinath et al., 2016, p.171), and "number of variables in the data set" (Wu et al., 2016, p.5). McNulty (2014) refers to variability as the change in the meaning of the data. DeVan (2016) differentiated the term variability from variety with the following example: a coffee shop may have different

blends of coffee and that is known as variety, but if the same blend of coffee tastes different every day, that is referred as variability. This means data might have a different meaning depending on the context of the usage. User-generated data in the form of comments and feedback are analyzed by different industries to understand customer satisfaction and business growth. In cases like this, the word may carry a different meaning based on the place of usage; such characteristics are known as variability. In the case of social networking, sites like Facebook, Twitter, and blogs, where people express their thoughts through written comments (example: opinion about the election), careful Big Data analysis is required because of the variable nature of the meanings in words.

3.1.7. Validity

Validity emphasizes how to obtain the required data by avoiding biases (Wu et al., 2016). Validity means the data is correct and accurate for the intended use (Owais & Hussein, 2016). According to Uddin et al. (2014), Validity refers to correctness and accuracy of data with respect to the intended usage. They suggest that the data with no veracity issues may not valid for a particular application if not understood correctly. The data set may be complete but does it tell the user what it purports to and is it valid for the current business context. Data quality not only depends on the completeness but also on the business environment and the business purpose it is supposed to serve. Only the data that conform to the business requirements can be considered valid. In health care, huge volume of sensor-generated data is generated by remote monitoring or continuous monitoring of patients. If the data generated are not valid, the analysis will lead to

erroneous results and making decisions based on those results would put the patient's life at risk.

Storing and analyzing Big Data is a complex procedure, as Big Data storage requires special hardware, and analysis requires Big Data analytics tools. To perform Big Data analysis, a company requires financial investment. IoT is widely used for environmental research like air quality monitoring, water quality monitoring, weather prediction, and natural disaster prediction. If the data used for analysis is biased, the results obtained will not be valid and making decisions or judgments based on the invalid result might lead to disaster.

3.1.8. Volatility

Volatility in Big Data refers to "how long data is valid and how long should it be stored" (Cartledge, 2016; Maheshwari, 2015; Normandeau, 2013; Yassin, 2014)**.** Volatility in Big Data also considers "at what point the data is no longer relevant to the current analysis" (Klarity, 2015). Big Data is stored and analyzed for various purposes like research, entertainment, healthcare, business etc. In case of the stock market and telecom industries, Big Data collected in recent time (current day/day before) is found to be more useful (Joseph, 2012). In online transactions in e-commerce, customer purchase histories are usually stored for the duration of one year (Owais & Hussein, 2016). In cases of data collected from sensors for research purposes like genomics, and astronomy, data are stored and analyzed for a long time and the volatility of the Big Data depends on different organizations (Hurwitz et al., 2013). Some organizations will only store the most recent

customer data in their business systems. This ensures faster retrieval for analysis. Another option is to process data "on the fly" and discard irrelevant information. Volatility of Big Data will also depend on the following:

- Volume - How much data is to be stored and what is the cost of storage?
- Value - How long does the data remain relevant and produce value?
- Data processing - Does the data need to be repeatedly processed?

3.1.9. Viability

The primary purpose of storing and analyzing Big Data in organizations is to gain insight from the data. Big Data are analyzed to uncover relationships among the variables used in the analysis. Viability refers to the process of "careful selection of attributes and factors that are most likely to predict outcomes that matter most to businesses" (Biehn, 2013, p.8). It had been believed by data scientists that "5% of the attributes in the data are responsible for 95% of the benefits" ("The Viability Of Big Data - Infographic"). A Big Data project consumes resources and time, so before investing in such a project, it is imperative to establish its viability and feasibility. This will reduce the risk of incurring large costs without generating greater benefits. Determining viability of data involves looking at the multidimensional data and identifying a set of attributes that would predict the outcome that is valuable for the business. For example, it can start with a simple hypothesis like the effect of customers' age and credit limit on his/her propensity to buy a certain product. Then a sample set of data is extracted, and statistically analyzed to identify any significant correlation between the chosen variables and the outcome. If

viability is established, further investment can be done to collect and refine the data source.

3.1.10. Virality

Different definitions proposed for the characteristic "virality" in Big Data are as follows:

- "Virality is a measure of the spread rate of data across the network" (Vermeend Willem, 2013, p.17).
- "Virality describes how quickly information gets dispersed across people to people (P2P) networks. Virality measures how quickly data is spread and shared to each unique node" (Wang, 2012, p.6).
- Virality refers to "rate at which the data spreads; how often it is picked up and repeated by other users or events" (Vorhies, 2014, p.24).

Going viral refers to your company's content (audio, video, comments or blogs) getting liked, shared, retweeted, reposted, commented on, etc. It is considered as an important factor for the growth of many start-ups and online businesses. By having their content (audio, video, posts, comments, Tweets) shared and liked by other users, they will be visible to more people, and it is a means by which they can acquire new customers at a lower cost.

3.1.11. Viscosity

Viscosity in Big Data is defined as follows:

- "Viscosity measures the resistance to flow in the volume of data. This resistance can come from different data sources, friction from integration flow rates, and from processing required to turn the data into insight" (Wang, 2012, p.6).
- Viscosity describes "the latency or lag time in the data, relative to the event being described" (Vorhies, 2014, p.23).
- Viscosity also refers to "the slowness in navigating the data, for example, by the variety of sources, the data flow speeds or the complexity of the required processing" (Vermeend & Ossyren, 2017, p.14).

Since data can originate from a variety of data sources, highly viscous data would require a lot of effort for transformation, interpretation, and integration. Highly viscous data would increase the cost of analysis and effort to extract valuable information. For example, many enterprises find it challenging to monitor and extract data from social media (tweets, Facebook comments, and posts) and integrate into the traditional data analysis process due to the nature and structure of data. These enterprises miss out on valuable insights, since the data is highly time sensitive and a delay in data processing diminishes its value.

3.1.12. Visualization

One of the biggest challenges of Big Data is how to make sense of Big Data once they are collected, processed and stored. The sheer volume and complexity of the data are so overwhelming that many organizations struggle to gain any competitive advantage by analyzing it. Big Data must be presented in a way that is easily comprehensible for heterogeneous audiences. Visualization is the process of "making all that vast amount of data comprehensible in a manner that is easy to understand and read" (Owais & Hussein, 2016; Rijmenam, 2015). Owais & Hussein (2016) refers visualization as a way to explore and understand data, in the same way that the human brain processes information. It enables decision makers to understand the meaning of different data values and to identify relationships and patterns easily. One major advantage of visualization is that it enables exploratory analysis that is accessible, even to decision makers, without a specific background in statistics. This allows critical data analysis to be a part of the everyday decision-making process.

In business, Big Data visualization is used to understand a "complete view of the customers" (Wang et al., 2015, p.35). In the modern economy, businesses would require a complete picture of their customers to stay competitive and improve customer retention. This requires that they understand how each customer is transacting with their company, how they are finding out about available products and services, and how are they interacting with them. Businesses must also track customer sentiments like what they are saying about their company, and what they are saying about the competition.

Visualization is an essential tool for integrating all of these analytical components and providing the big picture to the business user.

3.2. Summary of Vs

Laney (2001) used 3Vs (Volume, Velocity and Variety) to describe the technical challenges for existing data base management technologies and how to address them. Initially 3Vs "Volume, Velocity and Variety" were considered as the three main dimensions of Big Data. Over the years, Big Data technologies have evolved rapidly and many new definitions of Big Data are introduced from different perspectives. Several new characteristics were added to Big Data such as IBM added "Veracity", Oracle introduced "Value" and SAS introduced "Variability" (Gandomi & Haider, 2015). The Big Data definition has evolved to include not only the core data characteristics but also the infrastructure for management and the value that can be derived from them. There is a difference of opinion between the literature definition and practitioner's perspective.

Some argue that Big Data definitions should include only the 3Vs, as they describe the core characteristics and consider "value" as a derived property and suggested not to include it in Big Data definition (Grimes, 2013). Big Data practitioners argue that organizations could incur huge loss if they invest in Big Data infrastructure without understanding the value that can be derived from them and its usefulness to the organization (Dutcher, 2014). So any discussion of Big Data must not only include the 3Vs but also include the tools used and insights or value that can be derived from the

data. This Delphi study is aimed at identifying the core characteristics of Big Data from the perspective of Big Data practitioners and researchers.

The 12 V's that are included in the study were collected from existing literature and are widely considered as the core characteristics of Big Data. When we consider the V's that define the core characteristics of Big Data, not all of the characteristics are orthogonal, nor do they all define independent aspects of Big Data. Many of the characteristics of Big Data are related to each other. In particular:

- ***Velocity and Volume***: Velocity is the speed at which the data is generated, stored, analyzed and processed. An increase in velocity will result in increase in the volume of data.
- ***Velocity, Variety and Veracity***: Veracity refers to truthfulness or trustworthiness of data. An increase in the velocity and the variety of data (structured, semi-structured and unstructured) will typically lead to a higher level of uncertainty and, hence, reduce the trustworthiness of the data.
- ***Visualization:*** The velocity of data generation and variety of data with complex relationships make it challenging to develop meaningful visualization.
- ***Value:*** The eventual value derived will depend on Volume, Velocity, Variety and Veracity of the data. It will also depend on how quickly data can be analyzed and provide meaningful visualization for business stake holders.

Chapter 4: A Delphi Study of Big Data Characteristics

4.1. Delphi research methodology

The Delphi method "is an iterative process, normally three to four rounds, involving a series of questionnaires, each building on the results of the previous one" (Somerville, 2008, p.1). The main aim of Delphi is to solicit the opinion of experts (participants) (Hilbert, 2016) and to get a reliable consensus (Okoli & Pawlowski, 2004). The Delphi method dates back to the 1950s, when the RAND Corporation used this technique to obtain a reliable consensus from the group of experts about a military defense project (Bourgeois et al., 2006; Dalkey, 1963; Habibi et al., 2014). Linstone & Turoff (1975) defined Delphi as, "a method for structuring a group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem". The Delphi method uses iterative questions with a controlled feedback technique to solicit the opinion of the experts. The Delphi method is used in various fields (Schmidt, 1997, p.764) such as "public administration, medicine, technology diffusion, management, social work, education, operations management and information systems." According to Okoli & Pawlowski (2004), the Delphi method is a popular tool in information systems to identify and prioritize the issue for decision making. The Delphi method is considered to be an attractive and flexible method for graduate students (Skulmoski et al., 2007).

Four characteristics of the Delphi method are (i) anonymity (ii) iteration (iii) controlled feedback (iv) statistical aggregation (Habibi et al., 2014; Skulmoski et al., 2007; Somerville, 2008).

Anonymity: Some of the shortcomings of traditional group-based techniques are that the group members who have high self-confidence can dominate the other member in the group and also there is a case for coercion to conform to a certain viewpoint. The participant's anonymity in Delphi overcomes the above shortcomings (Habibi et al., 2014; Hsu & Sandford, 2007)

Iteration: In the controlled feedback process, a well-documented summary of the prior iteration results are distributed to each participant. This provides participants an opportunity to review their previous response and provide additional insights. Multiple iterations are expected to elicit more insightful opinions from the participants and minimize the effects of noise (Hsu & Sandford, 2007). The number of iterations required to attain consensus among the group is different for different studies (Hsu & Sandford, 2007; Keeney, Hasson, & McKenna, 2001). In general, three to five rounds of iteration are used (Hsu & Sandford, 2007; Somerville, 2008).

Controlled feedback: Feedback given to the participants in each round provides an opportunity to know about the perspective of their fellow participants (Skulmoski et al., 2007). By looking at the feedback, participants can review or change their opinion. Controlled feedback is used in Delphi to reduce the "effect of noise" (Hsu & Sandford, 2007).

Statistical aggregation: Statistical aggregation of a group's response "allows for quantitative analysis and interpretation of data" (Skulmoski et al., 2007, p.2). This provides an opportunity to represent the options given by each participant involved in the study (Hsu & Sandford, 2007).

4.2. Reasons for using Delphi for this study

The Delphi method can be used when there is an incomplete knowledge about the existing problem (Skulmoski et al., 2007). In case of this research, the number of proposed characteristics of Big Data is growing, as new characteristics are introduced from different perspectives as mentioned in Chapter 3.

Gartner analyst (Laney, 2001) used 3Vs (volume, velocity, and variety) to describe Big Data as a phenomenon and how it challenges traditional data management principles. Added to that he also discussed about the need for modern approaches to data management. Over the years the tools and technologies surrounding Big Data storage and management have evolved and organizations' outlook of Big Data has also changed. Organizations no longer consider Big Data as a data management problem, but as an opportunity to gain competitive advantage and drive businesses forward.

This can be noted in the updated definition by Gartner in 2012, "Big Data is high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" (Sicular, 2013, p,3). Along with the 3Vs in the original definition, the updated Big Data definition

included the technologies used for information processing and the value or insights gained from Big Data.

Swoyer (2012) presented the case for a fourth V: value. He says it's not enough to focus on the three different dimensions of the data alone and we need to focus on what can be done with the data, and how it can drive businesses.

Dutcher (2014) asked more than 40 thought leaders from various industries to define the phrase "Big Data" and found that there were multiple definitions for the term. The definitions varied based on the industry that the person belonged, but the common theme was the definitions were beyond the 3Vs and included other aspects like deriving insights, value, visualization, decision making, pattern recognition etc.

Therefore we can infer that from a practitioner or organizational perspective, the term Big Data not only includes the 3Vs but also the tools, technologies, analytics and insights that are derived from the data.

But there is another school of thought that is against adding other characteristics to the 3Vs of Big Data. Ylijoki & Porras (2016) in their mapping study suggested that separating Big Data from intended usage will clarify the definition and help us understand the characteristics of Big Data better. They concluded that even though analysis and data usage are essential elements they must not be included in defining the characteristics of Big Data.

Seth Grimes, an analytics strategy consultant with Alta Plana, argues that the "3V's (volume, volume, velocity and variety) do a fine job of defining Big Data. Don't be misled by variability, veracity, validity and value" (Grimes, 2013, p.1). According to Seth Grimes

- "Viability isn't a Big Data property. It's a quality that you determine via Big Data analytics.
- Variability and veracity are similarly analytics-derived qualities that relate more to data uses than to the data itself.
- Variability, veracity, validity and value aren't intrinsic, definitional Big Data properties. They are not absolutes. By contrast, they reflect the uses you intend for your data."

From the above discussion it is clear that there is a lack of consensus about the core characteristics of Big Data. Thus by using Delphi method this research aims to identify the core characteristics of Big Data from the perspective of Big Data practitioners and researchers.

4.3. Delphi methodology used in this research

This research using the Delphi method aims to solicit the opinion of participants about the characteristics of Big Data, and aims to achieve consensus in four rounds of iterative survey (Delphi). Following the four rounds of Delphi, additional survey was conducted to collect feedback about the level of agreement with the result obtained from the Delphi study. The Delphi method used in this research was framed based on the guidelines in Skulmoski et al. (2007). Figure 1 depicts the Delphi process used in this research.

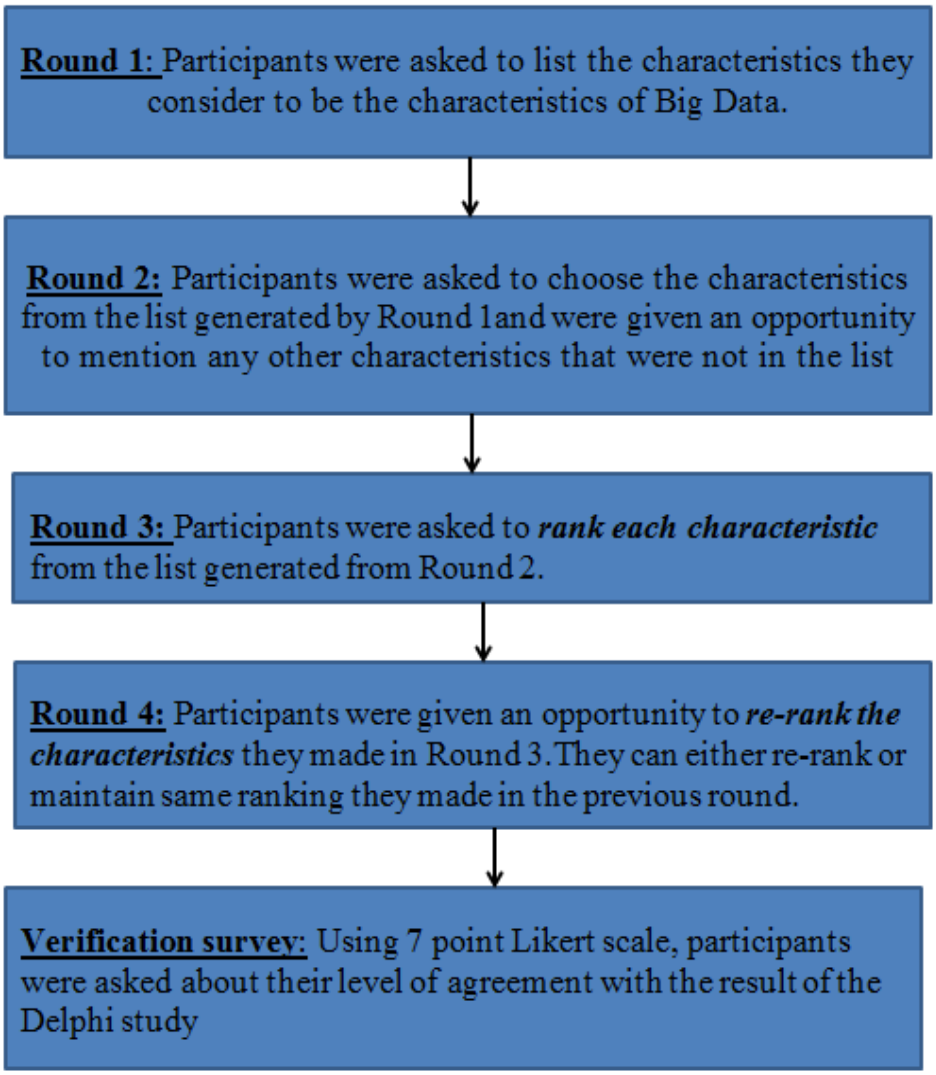


Figure 1: Delphi process used in this research

An overview of the four rounds used in this research is as follows:

Round 1: Participants were asked to list the characteristics they consider to be the characteristics of Big Data.

Round 2: Participants were requested to ***choose the characteristics*** of Big Data from the list generated from Round 1. Additionally, participants were given an opportunity to mention any other characteristics that were not presented in the list.

Round 3: Participants were asked to ***rank each characteristic*** from the list generated from Round 2.

Round 4: Participants were given an opportunity to ***re-rank the characteristics*** they made in Round 3. Participants can either re-rank or maintain same ranking they made in the previous round.

4.4. Participant selection

The quality or the accuracy of the result from the Delphi study is directly related to the quality of the participants involved in the study. Therefore, in a Delphi study, experts are generally used as participants. Participant selection is considered to be a vital part of the Delphi method (Hsu & Sandford, 2007; Skulmoski et al., 2007). This research is aimed at identifying the core characteristics of Big Data, and the participants for this research are Big Data practitioners and researchers. Purposive sampling (Cooper & Blumberg, 2011) was used to select the participants (Big Data practitioners and Big Data researchers). As defined by (Dalkey, 1963, p.11) “purposive sampling is a type of non- probability

sampling that is most effective when one needs to study a certain cultural domain with knowledgeable experts within."

One of the important steps in purposive sampling is to "define the qualities the informant(s) should or should not have" (Tongco et al., 2007, p.149). Burtch (2013), defines Big Data professionals as "individuals who can apply sophisticated quantitative skills to data describing transactions, interactions or other behaviors of people to derive insights and prescribe actions" (Linda Burtch, 2013, p.4).

In this research, participants are categorized as Big Data practitioners and researchers as follows

- Big Data practitioner - uses Big Data tools and technologies in an organization
- Big Data researcher - does academic research related to Big Data

In the case of Big Data practitioner, based on internet search, twenty jobs related to Big Data in organizations, and technical skills required to perform those jobs were identified. It was found that the job title and job description were not uniform across various organization. So instead of using job titles, we decided to recruit participants based on their skills set related to Big Data.

To participate in the Delphi study Big Data practitioners were required to have one or more of the following skill sets:

- Knowledge and experience with Big Data technologies such as Hadoop, Spark, Scala, Pig, Hive, Sqoop, Flume, HBase, and MapReduce
- Knowledge of predictive analytics techniques (e.g. predictive modeling, statistical programming, machine learning, data mining, data visualization)
- Experience and proficiency in skills relevant for Big Data (e.g. Java, Scala, Python, Perl, C++, SQL, Hive-QL, R, Scikit-Learn, Mahout, Matlab)
- Possess a strong foundation in Databases, Data modeling techniques, and ETL implementations

Additionally, if any participants choose "none of the above skill" option but still considered himself/herself as a Big Data practitioner, a comment box was provided to describe their skill set to participate in the survey.

To qualify as a Big Data researcher, the participant should have done academic research related to Big Data.

Guidelines for participant recruitment for this research was framed based on the criteria mentioned by (Skulmoski et al., 2007, p.1).

- "Knowledge and experience with the issues under investigation
- Capacity and willingness to participate
- Sufficient time to participate in iterative online survey"

The capacity of the participants to participate in the survey was analyzed from the qualifications mentioned by participants in the survey. The willingness of the participant to take part in the study was obtained from the participants as consent and provide their email address. Since the Delphi method is based on an iterative survey, the consent form (see Appendix) explained to participants the iterations and time required for each iteration before they start.

This study deals with humans, so the proposal was approved by the Interdisciplinary Committee on Ethics in Human Research (ICEHR). The following are explanations about each round of Delphi methodology used in this research.

4.5. Round1

Participant selection, a vital component in Delphi, was made in this round. The participants were categorized as Big Data practitioner or Big Data researcher (detailed explanation for categorization is in [4.4]). In Round 1, the survey link carrying consent form and survey questions (attached in Appendix) was posted in the AISWORLD listserve, Linked In, Research Gate, and Twitter. The consent form provided participants a detailed description about the research, role of participants, and time required to complete the survey. After providing consent, participants could start the survey. Participants' email addresses were collected in Round 1 to send the next round of survey questions. The reason for collecting email addresses was explained to the participants in the consent form. The Round 1 survey question (see Appendix) was open-ended question, and a comment box was used in the survey to solicit the opinion of the participants.

Participant's responses for the open-ended questions provided an opportunity for the researcher to understand more about participants view on Big Data characteristics. The deadline to complete the survey was mentioned to the participants, and responses received after the deadline were not considered.

4.6. Round 2

The Round 2 survey was framed based on the result obtained from Round 1. Unlike Round 1 where the survey link was posted publicly allowing anyone to participate, the Round 2 survey link was sent to the participants who participated in Round 1. This was done by sending email invitations from Survey Monkey using the email addresses collected form participants in Round 1. Round 2 aimed to obtain a consolidated list of Big Data characteristics from the participants. Round 2 survey questions (attached in Appendix) asked participants to select the characteristics of Big Data from the list obtained from literature and from Round 1. In addition, this round provided an opportunity for the participants to add new characteristics of Big Data that were not mentioned in the given list. The deadline to complete the survey was mentioned to the participants in the consent form and responses received after the deadline were not considered.

4.7. Round 3

The question for Round 3 was framed based on the results obtained from Round 2, and participants were given feedback about the results of Round 2. Similar to Round 2, in Round 3 participants who responded to Round 2 received a survey link to participate in Round 3 through an email invitation from Survey Monkey. Round 3 aimed to obtain a ranking for each characteristic of Big Data obtained from Round 2. The Round 3 survey question asked participants to rank each characteristic of Big Data in the list obtained from the result of Round 2. The deadline to complete the survey was mentioned to the participants, and response received after the deadline was not considered.

4.8. Round 4

Survey question for Round 4 was framed based on the results obtained from Round 3, and participants were given feedback about the results of Round 3. Similar to Round 3, in Round 4, participants who responded to Round 3 received a survey link to participate in Round 4 through an email invitation from Survey Monkey. Round 4's objective was to attain a consensus among the group about the characteristics of Big Data. The Round 4 survey question provided an opportunity for participants to re-rank each characteristic of Big Data, based on the group ranking obtained in Round 3. The participants looking at the feedback (Round 3) had an opportunity to re-rank the characteristics made earlier, or to stick with their earlier ranking. The deadline to complete the survey was mentioned to the participants, and responses received after the deadline were not considered.

4.9. Round continuation

In this research, if consensus was not reached in four rounds, the rounds would have continued until the consensus is reached. The consensus here represents no change in the group mean ranking of the characteristics of Big Data.

4.10. Verification survey

Upon reaching consensus among the group of experts about the characteristics of Big Data, a survey was conducted to collect feedback about the level of agreement with the result obtained from Delphi study. The survey link carrying the consent form and survey questions was posted in the AISWORLD listserve, LinkedIn, Research Gate, and Twitter. The consent form provided participants a detailed description about the research, role of participants, and time required to complete the survey. After providing consent, participants were able to start the survey. In this survey, participants were shown the results of the Delphi study (Round 4) and were asked to share their feedback about the result. Likert scaling technique, on the scale of 1-7 was used to collect participant's level of agreement with the result of Delphi study. The scaling used were: 1 (Extremely disagree), 2 (Mostly disagree), 3 (Somewhat disagree), 4 (Neither agree nor disagree), 5 (Somewhat agree), 6 (Mostly agree), and 7 (Extremely agree). Added to that participants were also asked to justify the reason for their level of agreement with the Delphi result and a comment box was used to collect the answer from the participants.

Chapter-5: Results

5.1. Round 1

The survey link for Round 1 was posted on the AISWORLD listserve, LinkedIn, Twitter, and Facebook. The survey questions used in this research are provided in the Appendix. Detailed descriptions about the research, research method (iterative survey), time required to participate in the survey, benefits, confidentiality, and anonymity were explained to the participants in the consent form. By agreeing to the consent form, participants were able to access the survey questions for Round 1. Email addresses were collected from the participants to send the Round 2 (and subsequent rounds) survey questions. Participants were categorized as Big Data practitioners or Big Data researchers based on the criteria mentioned in Section 4.4. The open-ended question used in Round 1 was "***As a Big Data practitioner/ Big Data researcher what do you consider to be the characteristics of Big Data? List as many characteristics as you can.***" The answer to this question was collected in a comment box.

5.1.1. Result of Round 1

Round 1 was posted in AISWORLD listserve, LinkedIn, Twitter, and Facebook for a span of 8 weeks. 11 participants (3 Big Data practitioners and 8 Big Data researchers) responded to the Round 1 survey and submitted their email address to contact for Round 2. Using the comment box, participants provided a brief description about Big Data characteristics. Based on the response received for the question to categorize themselves

as a Big Data practitioner or reserchers, all the 11 participants were accepted to participate in the Delphi study. The objective of this round was to collect set of Big Data characteristics that could be used for the next round. About 20% of the participants mentioned the 3Vs (volume, velocity and variety) and rest of the participants provided other characteristics such as: sense making, deriving insights, prediction and statistics, large data sets from a variety of resources, visualization technique, and analyzing many data points together. From the various characteritics gathered from the participants, we can infer that their prepective of Big Data included not only the 3Vs but also the analytical process involved and insigths that can be derived from the data.

5.2. Round 2

Based on the response received from Round 1 and literature review (mentioned in Chapter 3), 12 Big Data characteristics were identified. To provide a simplified description for each characteristic to the participants, we reviewed the literature (mentioned in Chapter 3) and framed a one line definition for all 12 characteristics as shown in Table 2. The characteristics listed in the table are in alphabetical order.

Characteristics	Definition
Validity	appropriateness of the data for its intended use
Value	identifying what is valuable and then transforming and extracting that data for analysis
Variability	variation in the data flow rates
Variety	structural heterogeneity in a data set
Velocity	rate at which data are generated and the speed at which it should be analyzed and acted upon
Veracity	accuracy of the data
Viability	selection of attributes and factors that is most likely to predict outcomes that matter most to businesses
Virality	rate at which the data spreads; how often it is picked up and repeated by other users or events
Viscosity	latency or lag time in the data relative to the event being described
Visualization	transferring immense scale of data into something easily comprehended and actionable
Volatility	tendency for data structures to change over time
Volume	magnitude or size of data

Table 2: Definition used in Round 2

In this round, participants were asked to choose the characteristics that they consider to be the characteristics of Big Data from the given list. Also, participants were given an opportunity to mention Big Data characteristics that were not mentioned in the list. The question used in this round was "***What do you consider to be the important characteristics of Big Data? Please select as many characteristics from the following list as you feel are appropriate***." To provide an opportunity for the participants to come up with other Big Data characteristics the following question was used "***Please list any other important characteristics of Big Data not mentioned in the list above***" and the comment box was used to collect the answer. The Round 2 survey link was sent to the 11 participants (3 Big Data practitioners and 8 Big Data researchers) who participated in Round 1. The survey link was sent in the form of an email invitation, using the email addresses collected from the 11 participants in Round 1. Participants were given two weeks to respond to the survey.

5.2.1. Result of Round 2

At the end of two weeks, out of 11 participants contacted for Round 2, 4 participants responded back to the survey. The 4 participants were composed of 3 Big Data researchers and 1 Big Data practitioner. The responses received for Round 2 from the 4 participants are shown in Table 3 and the characteristics in the table are listed in alphabetical order.

Characteristics	Number of participants selecting this characteristic
Validity	3
Value	3
Variability	0
Variety	2
Velocity	2
Veracity	2
Viability	2
Virality	0
Viscosity	0
Visualization	4
Volatility	0
Volume	4

Table 3: Participant response for each characteristic in Round 2

In Round 2, the participants were requested to choose the core characteristics of Big Data based on the list of 12 characteristics provided and add any other characteristic that they believe to be the core characteristics of Big Data.

Volume, Visualization, Validity and Value were chosen 75% or more of the participants, and 50% of the participants choose Variety, Velocity, Veracity and Viability.

Additionally, new characteristics such as Density and Usefulness were introduced by two of the participants.

The definition for the characteristics "Density" as given by the participant is as follows "the ratio of meaningful information to data points. In 'small data', like a customer database, each record has meaning of its own. 'Big Data' is characterized by combining lots of data points together to find a small number of insights. For example, millions of transactions might be analyzed to generate a much smaller number of product recommendations". Though "*Usefulness*" is closely related to "*Value*", since it was provided by the participant as an additional characteristic it was included in the next rounds of the study.

The 3Vs (Volume, Velocity and Variety) have been traditionally described as the three dimensions of Big Data, in Round 2 of our study, we found that Volume, Visualization, Validity and Value were the most popular choice among the participants. Visualization, Validity and Value seems to be preferred over Variety and Velocity.

5.3. Round 3

The 4 participants (3 Big Data researchers and 1 Big Data practitioner) who responded to Round 2 received the survey link for Round 3 in the form of an email invitation. The analysis of the Round 2 results was used in the framing of the question for Round 3. The characteristics "visualization, volume, validity, value, variety, velocity, veracity, viability, density, and usefulness" were used in framing questions for Round 3. The

participants received the feedback from Round 2 in the form of bar graphs, as shown in Figure 2.

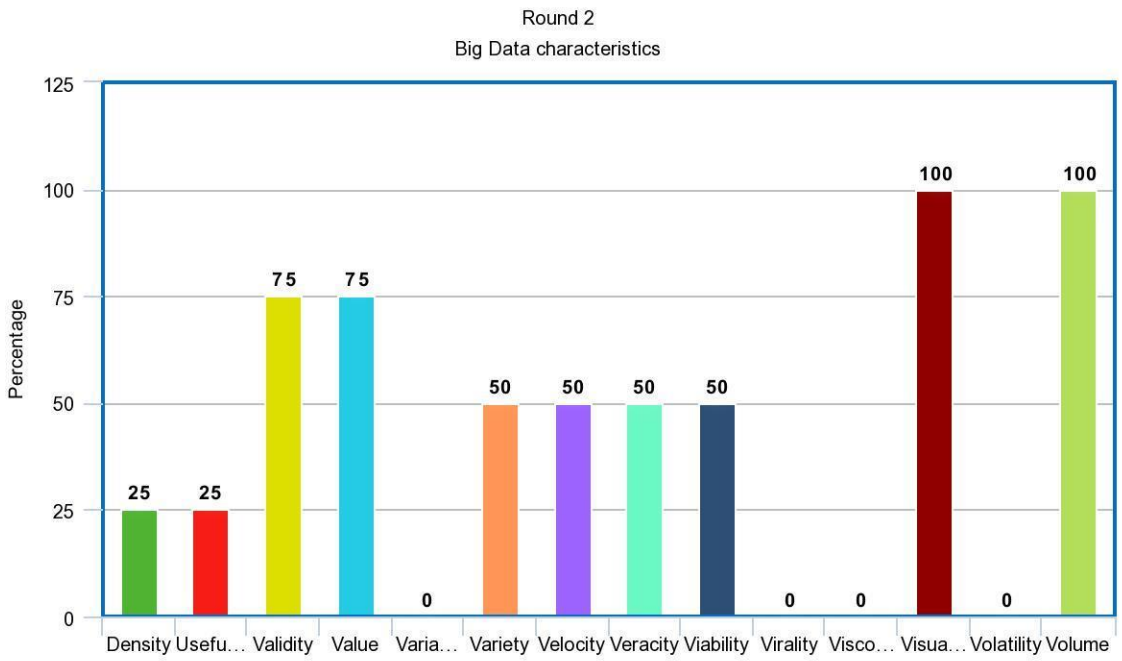


Figure 2: Feedback given to participants in Round 3

The aim of giving feedback is to give an opportunity for the participant to know about the perspective of fellow participants. This round obtained rankings for the individual characteristics such as "visualization, volume, validity, value, variety, velocity, veracity, viability, density, and usefulness." The question used in Round 3 was "***The following characteristics are listed in order based on response from Round 2. Please provide your individual rank for each characteristic on a scale of 1 (highest) – 10 (lowest)***." The characteristics are listed based on the group mean ranking response of Round 2 [Figure 3].

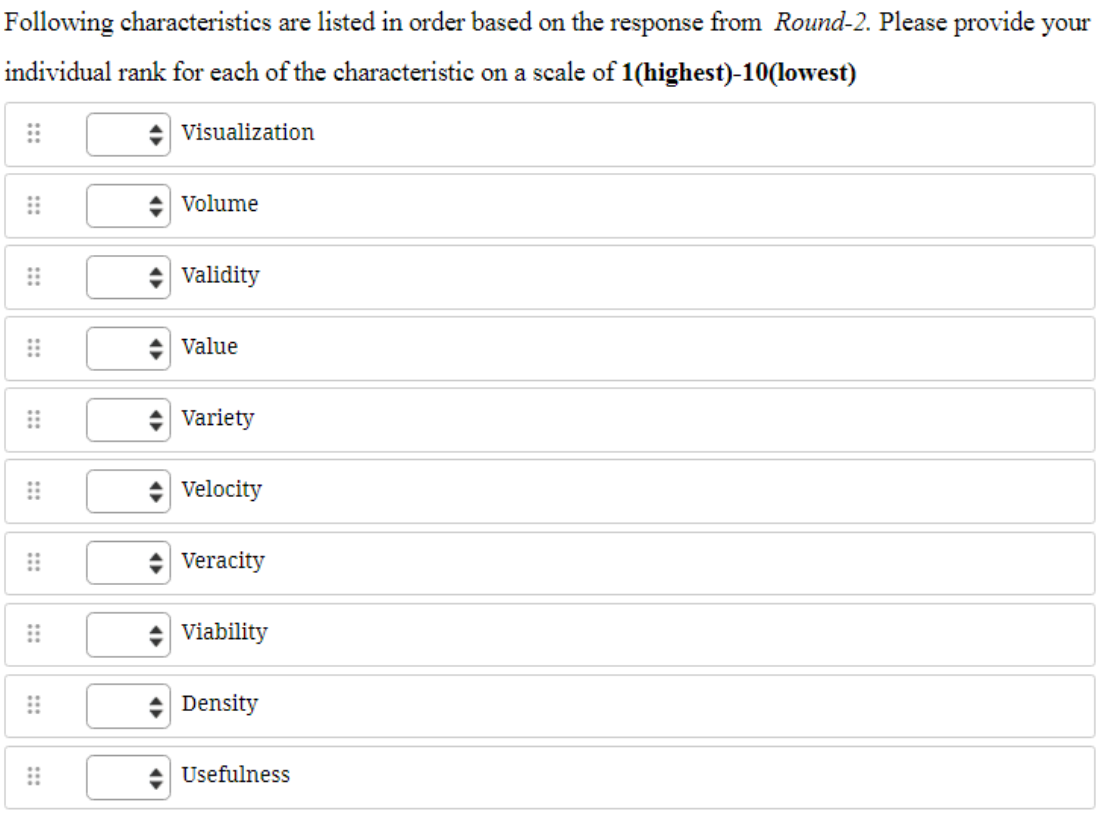


Figure 3: Ranking question used in Round 3

5.3.1. Result of Round 3

Participants in Round 3 were given two weeks to respond back to the survey, and the closing date was mentioned to the participants in the email survey link. All 4 participants (3 Big Data researchers and 1 Big Data practitioner) who received the survey link for Round 3 replied to the survey. Their responses are shown in Table 4.

Characteristics of Big Data	Rank given for each characteristics by four participants 1(highest) to 10(lowest)				Mean Ranking
	Participant-1	Participant-2	Participant-3	Participant-4	
Visualization	6	8	1	4	4.75
Volume	2	1	2	7	3
Validity	7	5	4	3	4.75
Value	5	4	5	2	4
Variety	8	3	6	8	6.25
Velocity	3	2	7		4
Veracity	9	6	9		8
Viability	10	7	10	5	8
Density	1	10	8	6	6.25
Usefulness	4	9	3	1	4.25

Table 4: Participants response for individual characteristics in Round 3

In Round 3, we requested participants to rank each of the characteristics, the rationale being that it would elicit better insights from the participants since they would have to bring in their experience to evaluate each characteristic and rank them.

Volume, Value, Velocity, Usefulness and Visualization were the Top 5 ranked characteristics in Round 3 based on mean ranking.

Interestingly, Velocity which was not such a popular choice in Round 2 (only 50% of participants choose Velocity) was ranked much higher in Round 3. When the participants were asked to rank the characteristics, Velocity seems to have much higher importance than Visualization, Validity and Veracity.

Visualization, Value and Usefulness emerged as some of the important characteristics along with Volume and Velocity.

Characteristics	Mean ranking
Volume	3
Value	4
Velocity	4
Usefulness	4.25
Visualization	4.75
Validity	4.75
Variety	6.25
Density	6.25
Veracity	8
Viability	8

Table 5: Mean ranking response received from Round 3

5.4. Round 4

The response received from Round 3 was analyzed, and mean ranking was calculated for each characteristic. The 4 participants (3 Big Data researcher and 1 Big Data practitioner) who replied to Round 3 received the link to Round 4 through email invitation. Participants were given two weeks to respond to the survey. The analyzed result of Round 3 was given to participants in the form of feedback as shown in Figure 4. Feedback to the participants provided an opportunity for them to understand what happened in the previous round, and about the perspective of the other participants in the same group. Round 4 aimed at gaining consensus among the group of participants.

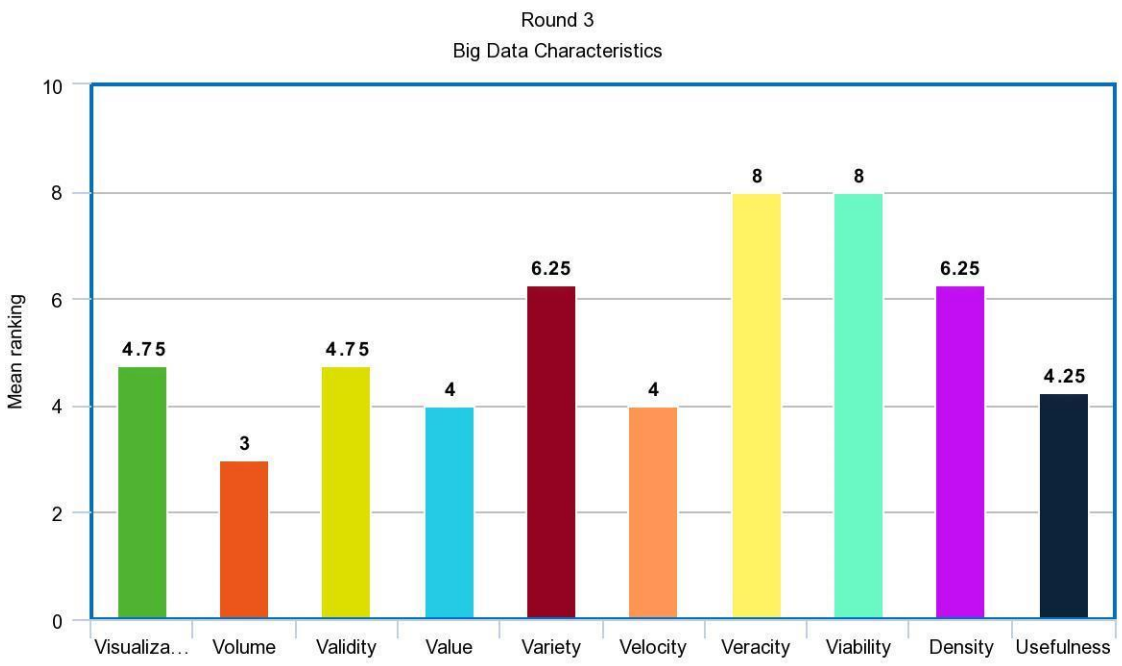


Figure 4: Feedback given to participants in Round 4

A comparison of group ranking and the participant rank for each characteristic was shown to the participants in the Round 4 survey link. Round 4 gave a chance to the participants to re-rank the characteristics based on the feedback (result from Round 3). The table used in the Round 4 survey link was different for each of the four participants. The table used in Round 4, listed both the group ranking and individual ranking, to provide an opportunity for the participants to understand the perspective of the fellow participants. An example is shown in Table 6.

Characteristics	**Group ranking**	**Your ranking**
Volume	1	2
Value	2	5
Velocity	2	3
Usefulness	4	4
Visualization	5	6
Validity	5	7
Variety	7	8
Density	7	1
Veracity	9	9
Viability	9	10

Table 6: Group ranking and individual ranking used in Round 4

In this round, the participants could either re-rank or stick with the ranking made in the previous round.

5.4.1. Result of Round 4

Participants were given two weeks to respond to the survey. All 4 participants (3 Big Data researchers and 1 Big Data practitioner) who received the survey link for Round 4 responded, and all of the participants elected to stay with their ranking in the previous iteration. Thus, a consensus among the participants was achieved for the characteristics of Big Data. Thus, the core characteristics of Big Data and the order of their importance identified from the Delphi study is:

1. Volume
2. Value
3. Velocity
4. Usefulness
5. Visualization
6. Validity
7. Variety
8. Density
9. Veracity
10. Viability

5.5. Verification survey

In general, following a Delphi study, to verify or generalize the Delphi results, interviews or surveys are used. The objective of the verification survey used in this study was to collect opinions from a wider group of audience about the results of the Delphi study. The verification survey was conducted upon reaching consensus among the group of experts about the characteristics of Big Data (which included ranking and re-ranking). In the verification survey, we asked the participants to agree/ disagree with the results.

The aim of the verification survey was to solicit feedback about the mean ranking of Big Data characteristics obtained from the Delphi study and to validate the result obtained. The survey link was posted on the AISWORLD listserve, LinkedIn, Twitter, and Facebook. A detailed description about the research, time required to participate in the survey, benefits, confidentiality, and anonymity were explained to the participants in the consent form. After consenting to participate in the study, participants gained access to the survey questions.

In this round, participants were given a bar graph representing the result of the Big Data characteristics, and were asked to rate their level of agreement in the 7 point Likert scale, and a comment box was also used to obtain the reason for their level of agreement. Figure 5 presents the bar graph given to the participants in the verification survey. The question used in this survey was "***On the scale of 1-7, how much do you agree with the results of the figure above?" the response options were:*** *1 (Extremely disagree), 2*

(Mostly disagree), 3 (Somewhat disagree), 4 (Neither agree nor disagree), 5 (Somewhat agree), 6 (Mostly agree), and 7 (Extremely agree).

The following figure represents the result of survey response (mean ranking of individual characteristics of Big Data) received from my study using "Delphi" method.

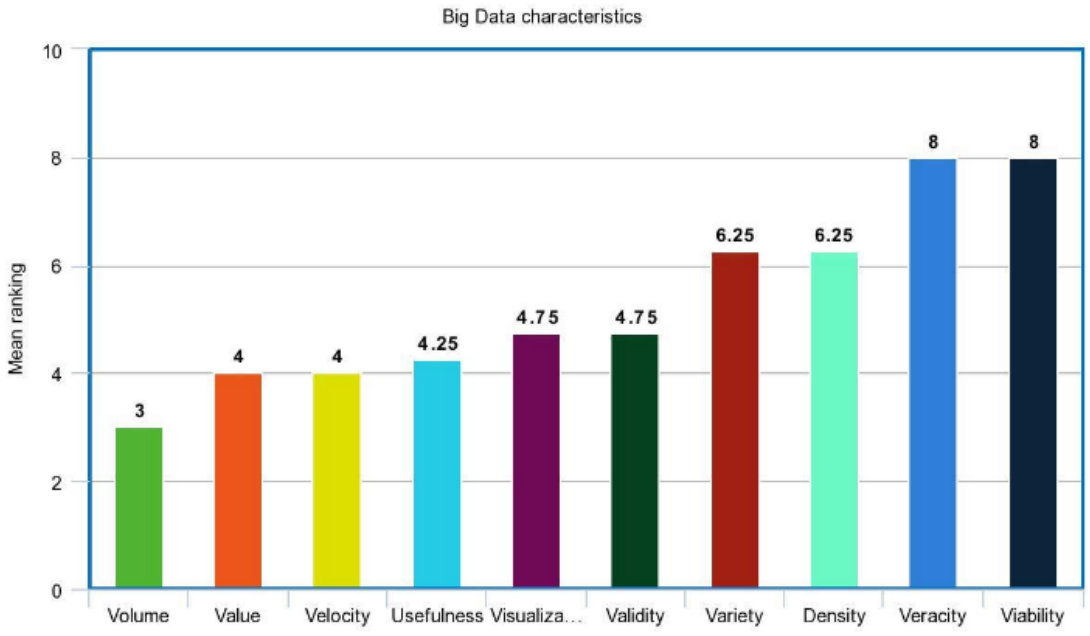


Figure 5: Mean ranking of the characteristics from Delphi study

5.5.1. Verification survey result

The survey was open for 45 days. Eighteen people responded to the survey and responses from all the eighteen participants were included in this study. Unlike the Delphi survey, this survey had no specific criteria for the Big Data practitioners and researchers to participate in the survey. The participants were broadly categorized as Big Data

practitioners and Big Data researchers. Among the 18 participants, there were 8 practitioners and 10 researchers. Figure 6 summarizes the responses received.

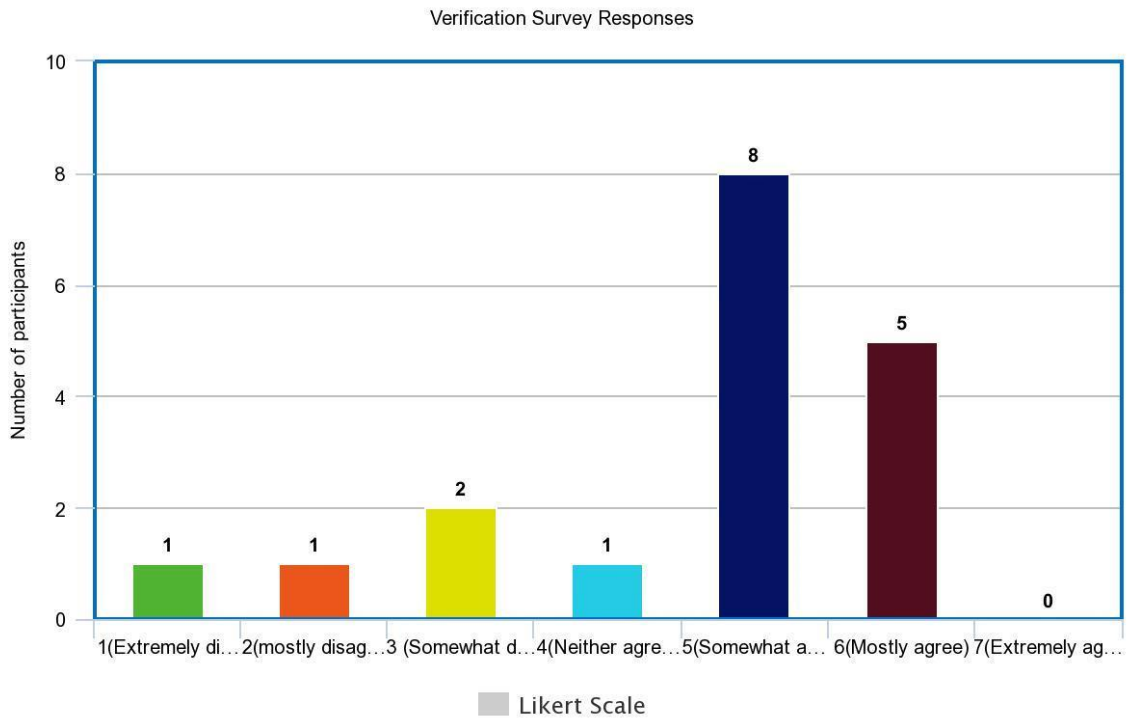


Figure 6: Response received from verification survey

The reason for having the verification survey was to collect the opinion from a wider group of audience about the results of the Delphi study. The weighted average calculated for the participant's level of agreement with the result of Delphi study is 5.

About 70% of the participants agreed to the results of the Delphi study. Additionally, the participants had the option to provide their comment about the ranking of Big Data

characteristics and based on the comments received, about 15% of the participants believe that the characteristic Variety and Veracity should be ranked much higher.

Among Big data practitioners, 75% somewhat agree and 13% mostly agree to the results of the Delphi study. Among Big data researchers 20% somewhat agree and 40% mostly agree to the results of the Delphi study. Though participants were categorized into two groups as Big Data practitioners and researchers because of the low number of participants it was not possible to differentiate between practitioners' view and researchers' view of Big Data.

Chapter 6

Summary

Over the years the tools and technologies surrounding Big Data storage and management have evolved and organizations' outlooks of Big Data have changed. Advances in computing technology have made it possible to manage huge volumes of data without requiring expensive server or super computers. Open source platform like Hadoop with its MapReduce programming model can process large amount of data in a cost and time effective manner using commodity machines. Cloud computing and Software as a service (SaaS) business models have made Big Data analytics accessible for even small business. Big Data is no longer considered as a data management problem, but as an opportunity or platform that can be exploited to drive businesses forward.

One of the main reasons many Big Data initiatives fail is not because of the volume of data, but what the organization tries to achieve after gathering and storing the data. Instead of looking at the input (data), business should start looking at the output (value) there are trying to derive and then decide if they have a Big Data problem.

Value seems to be the important factor that would decide the outcome of Big Data investment and Value is dependent on data Validity. When we consider Big Data 80% of the data is unstructured and hence there is a certain degree of uncertainty and inaccuracy attached to it. Data quality standards are still being developed and not yet standardized. Also the quality attributes are contextual and vary depending on the industry.

Through four rounds of Delphi study and verification survey this research aimed to identify the core characteristics of Big Data. It is found that along with the 3 Vs, the participants also considered Value, Usefulness, Validity and Visualization to be the core characteristics of Big Data. Value, Usefulness and Validity were ranked among the Top 5 characteristics of Big Data based on Group Mean ranking.

In Round 1, when participants were asked an open-ended question to list the characteristics of Big Data, the responses received were volume, velocity, variety, visualization and privacy issues. However, in Round 2, when they were given a list and asked to select the characteristics that they consider to be Big Data characteristics, additional characteristics such as validity, value, veracity, and viability were chosen.

Also, when they were given an opportunity to mention characteristics that were not in the list, two new characteristics – Density and Usefulness – were mentioned by the participants. Note that, when the participants were exposed to the suggestions by others in Round 2, they tended to select more characteristics. This indicates the view of Big Data is still evolving, and it is not limited to the 3 Vs.

Based on the results of my study, I conclude that any discussion of Big Data should have a holistic view and should not only include the 3Vs but also the data analytical processes that were used to gain insights value and the data quality and data governance process followed collect and store the data.

According to a study conducted by IDG (International Data Group), 78% of the enterprises have acknowledged that data strategy, collection and analysis have the potential to completely change the way they conduct business (Jones, 2017). Big Data is changing how business are run traditionally and many organization are moving to a data-driven decision making model. There is a necessity to understand the Big Data much beyond the scope of the 3Vs. The people who are using the data must understand how the data is accumulated and stored and how data fits in their business work flow to achieve their goals.

Limitation and Possible extension

One limitation of this research is the number of participants. Initially, this research was started with the idea to categorize the participants into two groups as Big Data practitioners or researchers, and to conduct the iterative survey for the two groups separately. The aim was to collect the core characteristics of Big Data from the perspective of Big Data practitioners and Big Data researchers independently, and then to examine the similarities and differences between the two groups. This is a Delphi study and new participants could not be introduced in the later stages. Based on this, the study continued with the responses from the four participants. Because of the low number of responses, the participants were not categorized into two groups. Instead they were considered as one group and the study aimed to achieve consensus within the group.

In future, this research can be extended with large number of participants from different Big Data industries. Categorizing the participants based on their area of working (banking, healthcare, retail industry, etc.) and conducting a Delphi study might help in identifying characteristics of Big Data from the perspective of different industries.

Bibliography

Aghaei, S., Nemaakhstbh, M. A., & Farsani, H. K. (2012). Evolution of the world wide web: From WEB 1.0 TO WEB 4.0. *International Journal of Web & Semantic Technology*, *3*(1), 1–10.

Al Nuaimi, E., Al Neyadi, H., Mohamed, N., & Al-Jaroodi, J. (2015). Applications of big data to smart cities. *Journal of Internet Services and Applications*, *6*(1), 25.

Anuradha, J., & Ishwarappa. (2015). A brief introduction on Big data 5Vs characteristics and Hadoop Technology. *Procedia Computer Science*, *48*, 319–324.

Aslam, S. (2017). Instagram by the Numbers: Stats, Demographics & Fun Facts. Retrieved from https://www.omnicoreagency.com/instagram-statistics/

Aslam, S. (2017). Linkedin by the Numbers: Stats, Demographics & Fun Facts. Retrieved from https://www.omnicoreagency.com/linkedin-statistics/

Aslam, S. (2017). Twitter by the Numbers: Stats, Demographics & Fun Facts. Retrieved from https://www.omnicoreagency.com/twitter-statistics/

Biehn, N. (2013). The missing Vs in big data: viability and value. *Wired. Innovation Insights Community*, *5*.

Bourgeois, J., Pugmire, L., Stevenson, K., Swanson, N., & Swanson, B. (2006). The Delphi method: A qualitative means to a better future. *URL: Http://www. Freequality. org/documents/knowledge/Delphimethod. Pdf (Citirano 2. 11. 2011)*.

Cartledge, C. (2016). How Many Vs are there in Big Data?, 1–4. Retrieved from http://clc-ent.com/TBDE/Docs/vs.pdf

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209.

Chen, M., Mao, S., Zhang, Y., & Leung, V. C. M. (2014). Big Data, 59–79. https://doi.org/10.1007/978-3-319-06245-7

Da Xu, L., He, W., & Li, S. (2014). Internet of things in industries: A survey. *IEEE Transactions on Industrial Informatics*, *10*(4), 2233–2243.

Dalkey, N. & O. H. (1963). An experimental application of the Delphi method to the use of experts. *Management Science*, *9(3)*.

Davis, B. (2015). The 7 pillars of Big Data. *Petroleum Review*, (January), 34–36.

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *AIP Conference Proceedings*, *1644*(2015), 97–104.

Del Savio, L., Prainsack, B., & Buyx, A. (2016). Crowdsourcing the Human Gut. Is crowdsourcing also "citizen science"? *Journal of Science Communication*, *15*(3), 1–16.

DeVan, A. (2016). The 7 V's of Big Data. Retrieved from https://www.impactradius.com/blog/7-vs-big-data/

Dijcks, J. (2012). Oracle: Big data for the enterprise. *Oracle White Paper*.

Donald R. Cooper, Boris Blumberg, P. S. S. (2011). *Business Research Methods*.

Donchev, D. (2017). 36 Mind Blowing YouTube Facts, Figures and Statistics – 2017. Retrieved from https://fortunelords.com/youtube-statistics/

Dutcher, J. (2014). What is Big Data? Retrieved from https://datascience.berkeley.edu/what-is-big-data/

Eggleston, E. M., & Weitzman, E. R. (2014). Innovative uses of Electronic Health Records and social media for public health surveillance. *Currrent Diabetes Report*, *14(3)*.

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

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

Gary Barnett. (2014). Managing Data in the Internet of Things. Retrieved from http://hotdesks.org/docs/Managing-Data-in-the-Internet-of-Things.pdf

Geczy, P. (2014). Big data characteristics. *The Macrotheme Review*, *3*(6), 94–104.

Gopinath, V., Krishna Rao, S., Yallmmanda, C., & Prakash, K. P. (2016). The Journey of Big Data : 3 V ' s to 3 2 V ' s. *IJRCCT*, *5(3)*, 170–175.

Grimes, S. (2013). Big Data: Avoid "Wanna V" Confusion. Retrieved from https://www.informationweek.com/big-data/big-data-analytics/big-data-avoid-wanna-v-confusion/d/d-id/1111077?page_number=1

Habibi, A., Sarafrazi, A., & Izadyar, S. (2014). Delphi technique theoretical framework in qualitative research. *The International Journal of Engineering and Science*, *3(4)*, 8–13.

Habibi, A., Sarafrazi, A., & Izadyar, S. (2014). Delphi Technique Theoretical Framework in Qualitative Research. *The International Journal Of Engineering And Science*, 2319–1813. https://doi.org/10.1016/S0169-2070(99)00018-7

Hadi, H. J., Shnain, A. H., Hadishaheed, S., & Ahmad, A. H. (2014). Big Data and five V's characteristics. *Proceedings of IRF International Conference*, (1), 29–36.

Herring, D. (2001). Precision farming. Retrieved from https://earthobservatory.nasa.gov/Features/PrecisionFarming/

Hilbert, M. (2016). Big data for development: A review of promises and challenges. *Development Policy Review*, *34*(1), 135–174.

Hsu, C. C., & Sandford, B. A. (2007). Practical assessment. *Research & Evaluation, The Ohio*.

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big data for dummies*. *John Wiley & Sons*. Retrieved from http://www.dummies.com/programming/big-data/engineering/how-to-ensure-the-validity-veracity-and-volatility-of-big-data/

IDC. (2014). The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things. Retrieved from https://www.emc.com/leadership/digital-universe/2014iview/digital-universe-of-opportunities-vernon-turner.htm

Jones, M. C. (2017). It's Time for Small Business to Embrace Big Data.

Joseph, J. (2012). Taming Data Variety and Volatility is Key for Big Data Analytics. Retrieved from http://www.lavastorm.com/blog/2012/11/14/taming-data-variety-and-volatility-is-key-for-big-data-analytics/

Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big data: Issues and challenges moving forward. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (pp. 995–1004). IEEE.

Kalota, F. (2015). Applications of Big Data in Education. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, *9*(5), 1602–1607.

Kanellos, M. (2016). 152,000 smart devices every minute in 2025: IDC outlines the future of smart things. Retrieved from https://www.forbes.com/sites/michaelkanellos/2016/03/03/152000-smart-devices-

every-minute-in-2025-idc-outlines-the-future-of-smart-things/#60d765254b63

Kaur, I., Kaur, N., Tanisha, Gurmeen, & Deepi. (2016). Big Data Management: Characteristics, Challenges and Solutions. *International Journal of Computer Science And Technology*.

Keeney, S., Hasson, F., & McKenna, H. P. (2001). A critical review of the Delphi technique as a research methodology for nursing. *International Journal of Nursing Studies*, *38*, 195–200.

Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big-Data Applications in the Government Sector. *Communications of the ACM*, *57*(3), 78–85.

Klarity. (2015). Dimensions of Big Data. Retrieved from http://www.klarity-analytics.com/2015/07/27/dimensions-of-big-data/

Knilans, E. (2014). The 5 V ' s of Big Data. *Avnet*, *7*, 3–6. Retrieved from http://ats.avnet.com/na/en-us/news/Pages/The-5-Vs-of-Big-Data.aspx

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity and Variety. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Linda Burtch. (2013). Salaries for Big Data Professionals- The Burtch works study.

Linstone, H. A., & Turoff, M. (1975). *Delphi Method: Techniques and Applications* (Vol. 29). Addison-Wesley Reading, MA.

List of citizen science projects. (n.d.). Retrieved from https://en.wikipedia.org/wiki/List_of_citizen_science_projects

Lister, M. (2017). 40 Essential Social Media Marketing Statistics for 2017. Retrieved from http://www.wordstream.com/blog/ws/2017/01/05/social-media-marketing-statistics

Ma, Y., Rao, J., Hu, W., Meng, X., Han, X., Zhang, Y., … Liu, C. (2012). An efficient index for massive IOT data in cloud environment. In *In Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 2129–2133).

Maheshwari, R. (2015). 3 V's or 7 V's - What's the Value of Big Data? Retrieved from https://www.linkedin.com/pulse/3-vs-7-whats-value-big-data-rajiv-maheshwari

Mayer-Schnberger, V., & Cukier, K. (2012). Big data: A revolution that will transform how we live, work, and think.

McNulty, E. (2014). Understanding big data: the seven V's. *Dataconomy. Com*.

Michael, K., & Miller, K. (2013). Big data: New opportunities and new challenges [guest editors' introduction]. *Computer*, *46*(6), 22–24.

Mobertz, L. (2013). The Four V's of Big Data [INFOGRAPHIC]. Retrieved from https://blog.dashburst.com/infographic/big-data-volume-variety-velocity/

Mukherjee, S., & Ravi, S. (2016). Big Data – Concepts, Applications, Challenges and Future Scope. *International Journal of Advanced Research in Computer and Communication Engineering*, *5*(2).

NIST. (2014). Big data interoperability framework: volume 1, definitions. *NIST Special Publication, Information Technology Laboratory, Gaithersburg*, *23*.

Normandeau, K. (2013). Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity. Retrieved from http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, *42*(1), 15–29.

Oracle. (2013). *An Oracle White Paper Oracle: Big Data for the Enterprise*. Retrieved from http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf

Owais, S. S., & Hussein, N. S. (2016). Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data. *International Journal of Advanced Computer Science and Applications*, *7*(3), 254–258.

Pai, T. (2017). Big Data New Challenges, Tools and Techniques.

Parekh, H. (2001). Digitization: An overview of iisues.

Pattal, M. M. I., Li, Y., & Zeng, J. (2009). Web 3.0: A real personal web! More opportunities and more threats. In *NGMAST 2009 - 3rd International Conference on Next Generation Mobile Applications, Services and Technologies* (pp. 125–128). IEEE.

Prabhu, D. (2016). Application of web 2.0 and web 3.0: an overview, *2*(1), 54–62.

Press, G. (2016). IoT Mid-Year Update From IDC And Other Research Firms. Retrieved from https://www.forbes.com/sites/gilpress/2016/08/05/iot-mid-year-update-from-idc-and-other-research-firms/#11397a7b55c5

Quartz. (2015). Connected cars will send 25 gigabytes of data to the cloud every hour. Retrieved from https://qz.com/344466/connected-cars-will-send-25-gigabytes-of-data-to-the-cloud-every-hour/

Rijmenam, M. van. (2013). Why The 3V's Are Not Sufficient To Describe Big Data', Datafloq.

Ronda-Pupo, G. A., & Guerras-Martin, L. A. (2012). Dynamics of the evolution of the strategy concept 1962–2008: A co-word analysis. *Strategic Management Journal, 33(2),* 162–188.

Rose, K., Eldridge, S., & Chapin, L. (2015). The Internet of Things (IoT): An Overview--Understanding the Issues and Challenges of a More Connected World. *Internet Society*.

Rowe, S. Del. (2016). Beyond the three V's of Big data.

Schmidt, R. C. (1997). Managing Delphi surveys using nonparametric statistical techniques. *Decision Sciences, 28*(3), 763–774.

Schultz, J. (2016). How Much Data is Created on the Internet Each Day? Retrieved from https://www.gwava.com/blog/internet-data-created-daily

Shukla, S., Kukade, V., & Mujawar, S. (2015). Big Data : Concept , Handling and Challenges : An Overview. *International Journal of Computer Applications, 114*(11).

Sicular, S. (2013). Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s.

Singh, J., & Singla, V. (2015). Big Data: Tools and Technologies in Big Data. *International Journal of Computer Applications*, *112*(15).

Sivakumar. (2015). How Top 10 Industries Use Big Data Applications. Retrieved from http://www.datascienceassn.org/content/how-top-10-industries-use-big-data-applications

Skulmoski, G. J., Hartman, F. T., & Krahan, J. (2007). The Delphi method for graduate research. *Journal of Information Technology Education*, *6*.

Somerville, J. A. (2008). *Effective Use of the Delphi Process in Research: Its Characteristics, Strengths and Limitations*.

Sravanthi, K., & Reddy, T. S. (2015). Applications of Big data in Various Fields. *International Journal of Computer Science and Information Technologies (IJCSIT)*.

Su, X. (2013). Introduction to Big Data. Retrieved from http://aitel.hist.no/fag/big/lessons/lesson2.pdf

Su, X., Pattnaik, K., & Prasad Mishra, B. S. (2016). Introduction to Big Data Analysis, 1–20. Retrieved from http://link.springer.com/10.1007/978-3-319-27520-8_1

Swoyer, S. (2012). Big Data - Why the 3Vs Just Don't Make Sense.

Team, R. (2015). Email Statistics Report, 2015-2019. The Radicati Group. Retrieved from http://www.radicati.com/wp/wp-content/uploads/2015/02/Email-Statistics-Report-2015-2019-Executive-Summary.pdf

The Top 20 Valuable Facebook Statistics. (2017). Retrieved from https://zephoria.com/top-15-valuable-facebook-statistics/

The Viability Of Big Data - Infographic. (n.d.). Retrieved from https://datafloq.com/read/viability-of-big-data-infographic/418

Thiyagarajan, V. S., & Venkatachalapathy, K. (2014). Isolating Values From Big Data With The Help Of Four V's. *International Journal of Research in Engineering and Technology*, *4*(1), 132–135.

Tongco, M. D. C. (2007). Purposive sampling as a tool for informant selection. *Ethnobotany Research and Applications*, *5*, 147--158.

Uddin, M. F., Gupta, N., & Khan, M. A. (2014). Seven V's of Big Data understanding Big Data to extract value. In *Proceedings of 2014 Zone 1 Conference of the American Society for Engineering Education (ASEE Zone 1)*.

Understanding the 7 V's of Big Data. (2015). Retrieved from http://www.optimusinfo.com/blog/understanding-the-7-vs-of-big-data/

Vermeend, W., & Ossyren, A. (2017). *Revolution Big Data An exploration of the profound impact*. Business Web Solutions BV. Retrieved from http://www.bigdatawereld.nl/9

Vermeend Willem. (2013). *The impact of the internet: How the internet is changing the way we think, learn, work, do business and make money*. Retrieved from http://www.ebusinessbook.nl/152.php

Vorhies, B. (2014). How Many "V"s in Big Data--The Characteristics that Define Big Data. *Data Science Central*.

Vorhies, W. (2014). How many V's in big data? The characteristics that define big data. Data Science Central.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How "big data" can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, *165*, 234–246.

Wang, L., Wang, G., & Alexander, C. A. (2015). Big Data and Visualization: Methods, Challenges and Technology Progress. *Digital Technologies*, *1*(1), 33–38. https://doi.org/10.12691/dt-1-1-7

Wang, R. (2012). Monday's musings: Beyond the three V's of big data-Viscosity and virality. *A Software Insider's Point of View*.

Ward, J. S., & Barker, A. (2013). Undefined by data: a survey of big data definitions. *arXiv Preprint arXiv:1309.5821*.

Wu, C., Buyya, R., & Ramamohanarao, K. (2016). Big Data analytics= machine learning+ cloud computing. *arXiv Preprint arXiv:1601.03115*.

Yassin, A. T. (2014). Analyzing 6Vs of big data using system dynamics. *The 2nd Scientific Conference of the College of Science*.

Ylijoki, O., & Porras, J. (2016). Perspectives to definition of big data: a mapping study and discussion. *Journal of Innovation Management*, *4*(1), 69–91.

Zaslavsky, A., Perera, C., & Georgakopoulos, D. (2012). Sensing as a service and big data. *arXiv Preprint arXiv:1301.0159*.

Zhang, G. (2012). The impact of mobile technology on people's lives. Retrieved from https://dmsp.digital.eca.ed.ac.uk/blog/literaryhighstreet2012/2012/04/27/the-impact-of-mobile-technology-on-peoples-live/

Appendix

Consent form

You are invited to take part in a research project entitled "*Characteristics of Big Data*"

This form is part of the process of informed consent. It should give you a basic idea of what the research is about and what your participation will involve. It also describes your right to withdraw from the study. In order to decide whether you wish to participate in this research study, you should understand enough about its risks and benefits to be able to make an informed decision. This is the informed consent process. Take time to read this carefully and to understand the information given to you. Please contact the researcher, Raja Rajeshwari Sreenivasan (email address: rrs347@mun.ca), if you have any questions about the study or would like more information before you consent.

It is entirely up to you to decide whether to take part in this research. If you choose not to take part in this research or if you decide to withdraw from the research once it has started, there will be no negative consequences for you, now or in the future.

Introduction:
My name is Raja Rajeshwari Sreenivasan and I am Master's student in the Faculty of Business Administration at Memorial University of Newfoundland. As part of my Master's thesis I am conducting research under the supervision of Dr. Jeffrey Parsons.

Purpose of study:
The purpose of this research is to identify the characteristics of Big Data. In general, Big Data refers to a large set of data that is impossible to manage and process using traditional data processing tools. Huge volumes of data are generated in many areas, including scientific research, health care, sensor data, business applications and social media. The data generated are not only huge in volume, but also varied in structure, and generated at a rapid rate. Well known as a developing area of research, the concepts and constructs in Big Data are used every day by organizations and researchers. Not surprisingly there is a disagreement about what constitute the essential characteristics of Big Data. Using the "Delphi method" (research method), this research aims to identify expert views on the core characteristics of Big Data.

What you will do in this study:
This study will be conducted in the form of iterative survey. The main aim of this research is to solicit the opinion of participants about the characteristics of Big Data. This research will be conducted in two steps. Step 1: Participants will be asked to list the characteristics of Big Data. Step 2: Participants will be asked to rank and re-rank their opinion until no change in the group mean re-ranking is achieved.

Length of time:
This study will follow an iterative survey. Iterations will continue until no change in the group mean re-ranking is achieved. I anticipate there to be approximately three iterations. Each iteration will take 5-10 minutes. You have 2 weeks to respond to each iteration of the survey and the deadline will be clearly mentioned in all the iterations. Response received after deadline will not be considered because the question for next iteration is framed based on the response received from the previous iteration.

Withdrawal from the study:
If you do not respond to the survey you will not be included in the next iteration of the survey.

- Once when you click "Do not agree" in page-3 of the survey, the survey ends.
- At the end of the survey, only when you click "Done" button the response is submitted.
- At any time during the survey, you can click "Exit" and quit the survey
- Possible benefits:
- At the end of each round, the researcher will analyze the (group response received) result and share with the participants. Thus, the participants will get to know about their fellow participants responses in each round. This will give the participants an opportunity to learn what their peers perceive to be the core characteristics of Big Data

Possible risks:
There is no potential risk involved in this research.

Confidentiality:
The ethical duty of confidentiality includes safeguarding participant's identities, personal information, and data from unauthorized access, use, or disclosure. The data from this research will be analyzed in my thesis and may be published in academic conferences and academic journals; however, the data will be reported in aggregate form, so that it will not be possible to identify individual participants.

Anonymity:
Anonymity refers to protecting participants identifying characteristics, such as name or description of physical appearance. This research is based on Delphi research methodology, and anonymity of the participants is one of the key features of Delphi method. No information about the participants will be made available to anyone participating in the study. Anonymity will be maintained both throughout and after the study. Throughout the study, anonymity will be well maintained to ensure that the participants are free to share their opinion about the characteristics of Big Data. Only the number of participants in each round will be made available to participants in the group.

The email address collected will be used only to share the analyzed survey response and to invite the participant for the next iteration of the survey .

Storage of Data:
Only the personnel involved in this study will have access to the data. The ID log will be kept on a separate password-protected computer from the participant data. For additional security, the data-file itself will also be password protected. The data will be retained for the required duration of 5 years and then securely disposed. The on-line survey company, Survey Monkey hosting this survey is located in the United States. The US Patriot Act allows authorities to access the records of internet service providers. Therefore, anonymity and confidentiality cannot be guaranteed. If you choose to participate in this survey, you understand that your responses to the survey questions will be stored and may be accessed in the US. The security and privacy policy for the web survey company can be found at security and privacy.

Reporting of Results:
The final result will be published in a thesis and may be published in academic conference and academic journals. The thesis will be publicly available at the QEII library of Memorial University of Newfoundland. However, the data will be reported in aggregate form, so that it will not be possible to identify individual participants.

Sharing of Results with Participants:
This research is based on Delphi research methodology, wherein feedback to the participants is essential. At the end of each round, participants will get feedback in the form of a bar graph representing mean ranking of each characteristics. After analyzing the group response (in the form of feedback), the participants will be given a chance to rank, and re-rank their opinion. In the iterative survey, questions will be framed only based on the previous round response from the participants.

Questions:
You are welcome to ask questions at any time before, during, or after your participation in this research. If you would like more information about this study, please contact Raja Rajeshwari Sreenivasan by rrs347@mun.ca or Dr.Jeffrey Parsons by jeffreyp@mun.ca.

The proposal for this research has been reviewed by the Interdisciplinary Committee on Ethics in Human Research and found to be in compliance with Memorial University's ethics policy. If you have ethical concerns about the research, such as the way you have been treated or your rights as a participant, you may contact the Chairperson of the ICEHR at icehr@mun.ca or by telephone at 709-864-2861.

Consent:

By completing this survey you agree that:

- You have read the information about the research.
- You have been advised that you may ask questions about this study and receive answers prior to continuing.
- You are satisfied that any questions you had have been addressed.
- You understand what the study is about and what you will be doing.
- You understand that you are free to withdraw participation from the study by closing your browser window or navigating away from this page, without having to give a reason and that doing so will not affect you now or in the future.
- You understand that if you choose to withdraw, you are free to do so by not participating in the next iteration. The data collected from previous iteration will still be used in the research.
- In all the iterations you are free to omit questions that you do not wish to answer.

By consenting to this online survey, you do not give up your legal rights and do not release the researchers from their professional responsibilities. Please retain a copy of this consent information for your records.

I am aware of the research procedure and I agree to participate in the survey

- o I agree
- o I do not agree

Round 1

Thank you for your time and interest to participate in the survey "Characteristics of Big Data".

Please enter your email address

How would you categorize yourself?

- Big Data practitioner - uses Big Data tools and technologies in an organization
- Big Data researcher - does academic research related to Big Data

Big Data practitioner

Please choose options that best describe your skill set

- Knowledge and experience with Big Data technologies such as Hadoop, Spark, Scala, Pig, Hive, Sqoop, Flume, HBase, and MapReduce
- Knowledge of predictive analytics techniques (e.g. predictive modeling, statistical programming, machine learning, data mining, data visualization)
- Experience and proficiency in skills relevant for Big Data (e.g. Java, Scala, Python, Perl, C++,SQL, Hive-QL, R, Scikit-Learn, Mahout, MATLAB)
- Possess strong foundation in Databases, Data modelling techniques and ETL implementations
- Proficient in querying and analyzing very large scale structured and unstructured data sources
- None of the above

If you choose "none of the above" please describe briefly your other skill that would qualify you as a Big Data practitioner

Please briefly describe your work with Big Data

Please enter your years of experience working as a Big Data practitioner

As a Big Data practitioner, what do you consider to be the characteristics of Big Data? List as many characteristics as you can

Big Data researcher

Please enter your area of research in Big Data

Please enter your number of publications related to Big Data

As a Big Data researcher, what do you consider to be the characteristics of Big Data? List as many characteristics as you can.

Thank you for your time and consideration. You have reached the end of Round 1. We will get back to you with the result of Round 1 and survey link to Round 2 shortly.

Round 2

Thank you for your participation in Round 1 of my survey. Your response is valuable in making consensus about characteristics of Big Data. Your Round 2 survey starts below.

What do you consider to be the important characteristics of Big Data? Please select as many characteristics from the following list as you feel are appropriate.

- ***Validity***: appropriateness of the data for its intended use.
- ***Value:*** identifying what is valuable and then transforming and extracting that data for analysis.
- ***Variability***: variation in the data flow rates.
- ***Variety:*** structural heterogeneity in a data set.
- ***Velocity***: rate at which data are generated and the speed at which it should be analyzed and acted upon.
- ***Veracity***: accuracy of the data.
- ***Viability***: selection of attributes and factors that is most likely to predict outcomes that matter most to businesses
- ***Virality:*** rate at which the data spreads; how often it is picked up and repeated by other users or events.
- ***Viscosity***: latency or lag time in the data relative to the event being described.
- ***Visualization:*** transferring immense scale of data into something easily comprehended and actionable.
- ***Volatility:*** tendency for data structures to change over time.
- ***Volume:*** magnitude or size of data.

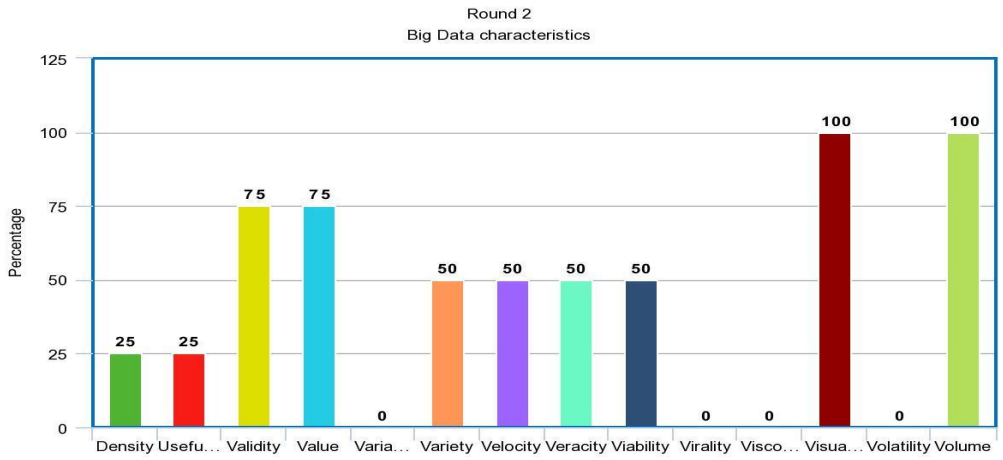
Please list any other important characteristics of Big Data not mentioned in the list above?

Thank you for your time and consideration. You have reached the end of Round 2. We will get back to you with the result of Round 2 and survey link to Round 3 shortly.

Round 3

Thank you for your participation in Round 2 of my survey. Your response is valuable in making consensus about characteristics of Big Data. Your Round 3 survey starts below.

The following figure represents the result of survey response received from Round 2.



Following characteristics are listed in order based on the response from Round2. Please provide your individual rank for each of the characteristic on a scale of 1(highest)-10(lowest)

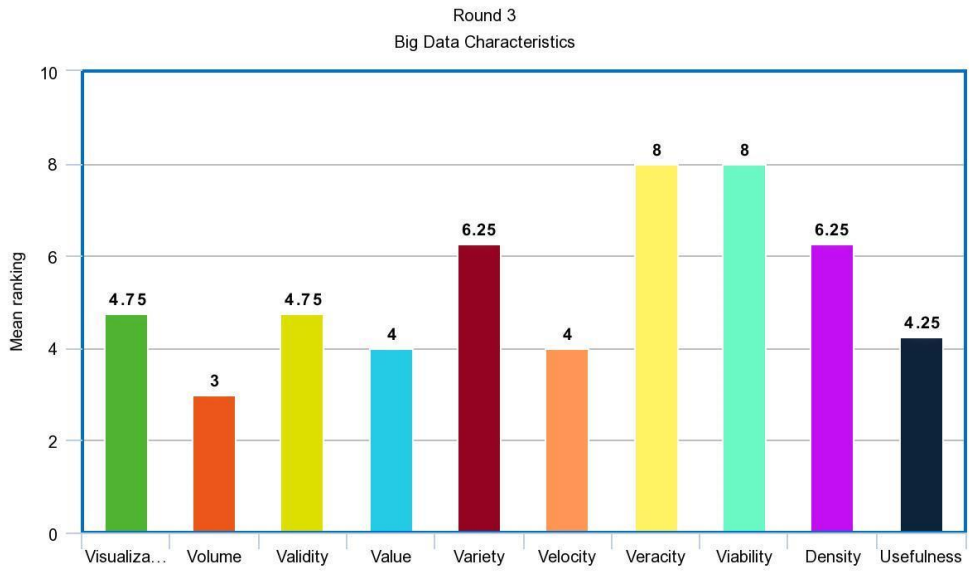
- Visualization
- Volume
- Validity
- Value
- Variety
- Velocity
- Veracity
- Viability
- Density
- Usefulness

Thank you for your time and consideration. You have reached the end of Round 3. We will get back to you with the result of Round 3 and survey link to Round 4 shortly.

Sample Question: Round 4

Thank you for your participation in Round 3 of my survey. Your response is valuable in making consensus about characteristics of Big Data. Your Round 4 survey starts below

The following figure represents the result of survey response (mean ranking) received from Round 3.



Characteristics	Group Ranking	Your Ranking
Volume	1	2
Value	2	5
Velocity	2	7
Usefulness	4	3
Validity	5	4
Visualization	5	1
Variety	7	6
Density	7	8
Veracity	9	9
Viability	9	10

For this Round, you may either re-rank the characteristics or stick with the ranking you made in previous Round. Please choose the option below.

- Re-rank the characteristics
- Stick with ranking made in previous round

If you would like to stick with the ranking made in previous Round (Round-3) please type "YES" in the box below

Thank you for your participation

Verification Survey

You are invited to take part in a research project entitled "Characteristics of Big Data"

This form is part of the process of informed consent. It should give you a basic idea of what the research is about and what your participation will involve. It also describes your right to withdraw from the study. In order to decide whether you wish to participate in this research study, you should understand enough about its risks and benefits to be able to make an informed decision. This is the informed consent process. Take time to read this carefully and to understand the information given to you. Please contact the researcher, Raja Rajeshwari Sreenivasan (email address: rrs347@mun.ca), if you have any questions about the study or would like more information before you consent.

It is entirely up to you to decide whether to take part in this research. If you choose not to take part in this research or if you decide to withdraw from the research once it has started, there will be no negative consequences for you, now or in the future.

Introduction:
My name is Raja Rajeshwari Sreenivasan and I am Master's student in the Faculty of Business Administration at Memorial University of Newfoundland. As part of my Master's thesis I am conducting research under the supervision of Dr. Jeffrey Parsons.

Purpose of study:
The purpose of this research is to identify the characteristics of Big Data. Well known as a developing area of research, the concepts and constructs in Big Data are used every day by organizations and researchers. Not surprisingly, there is a disagreement about what constitute the essential characteristics of Big Data. *In my previous study, using the Delphi method I have gathered expert views on the core characteristics of Big Data and created mean ranking for the characteristics. This survey aims to extend my Delphi study by soliciting opinion/feedback on the "mean ranking" of Big Data characteristics. This will help identify if there is any concurrence between expert's view on Big Data characteristics.*

What you will do in this study:
This study will be conducted in the form of survey. Participants will be asked to share their opinion/feedback on the result of my previous study (expert views on the core characteristics of Big Data).

Length of time:

This survey will take about 5-10 minutes.

Withdrawal from the study:

Once when you click "Do not agree" in page-3 of the survey, the survey ends.

At any time during the survey, you can click exit and quit.

At the end of the survey, only when you click "Done" button the response is submitted.

Possible benefits:

This will give the participants an opportunity to learn more about core characteristics of Big Data.

Possible risks:

There is no potential risk involved in this research.

Confidentiality:

The ethical duty of confidentiality includes safeguarding participant's identities, personal information, and data from unauthorized access, use, or disclosure. The data from this research will be analyzed in my thesis and may be published in academic conferences and academic journals; however, the data will be reported in aggregate form, so that it will not be possible to identify individual participants.

Anonymity:

Anonymity refers to protecting participants identifying characteristics, such as name or description of physical appearance. No information about the participants will be made available to anyone participating in the study. Anonymity will be maintained both throughout and after the study.

Storage of Data:

Only the personnel involved in this study will have access to the data. The ID log will be kept on a separate password-protected computer from the participant data. For additional security, the data-file itself will also be password protected. The data will be retained for the required duration of 5 years and then securely disposed. The on-line survey company, Survey Monkey hosting this survey is located in the United States. The US Patriot Act allows authorities to access the records of internet service providers. Therefore, anonymity and confidentiality cannot be guaranteed. If you choose to participate in this survey, you understand that your responses to the survey questions will be stored and may be accessed in the US. The security and privacy policy for the web survey company can be found at security and privacy.

Reporting of Results:
The final result will be published in a thesis and may be published in academic conference and academic journals. The thesis will be publicly available at the QEII library of Memorial University of Newfoundland. However, the data will be reported in aggregate form, so that it will not be possible to identify individual participants.

Sharing of Results with Participants:
If the participants are interested to know about the result of the study they can submit their email address in the survey

Questions:
You are welcome to ask questions at any time before, during, or after your participation in this research. If you would like more information about this study, please contact Raja Rajeshwari Sreenivasan by rrs347@mun.ca or Dr.Jeffrey Parsons by jeffreyp@mun.ca. The proposal for this research has been reviewed by the Interdisciplinary Committee on Ethics in Human Research and found to be in compliance with Memorial University's ethics policy. If you have ethical concerns about the research, such as the way you have been treated or your rights as a participant

Consent:

By completing this survey you agree that:

- You have read the information about the research.
- You have been advised that you may ask questions about this study and receive answers prior to continuing.
- You are satisfied that any questions you had have been addressed.
- You understand what the study is about and what you will be doing.
- You understand that you are free to withdraw participation from the study by closing your browser window or navigating away from this page, without having to give a reason and that doing so will not affect you now or in the future.
- You understand that if you choose to withdraw, you are free to do so by not participating in the next iteration. The data collected from previous iteration will still be used in the research.
- In all the iterations you are free to omit questions that you do not wish to answer.

By consenting to this online survey, you do not give up your legal rights and do not release the researchers from their professional responsibilities. Please retain a copy of this consent information for your records.

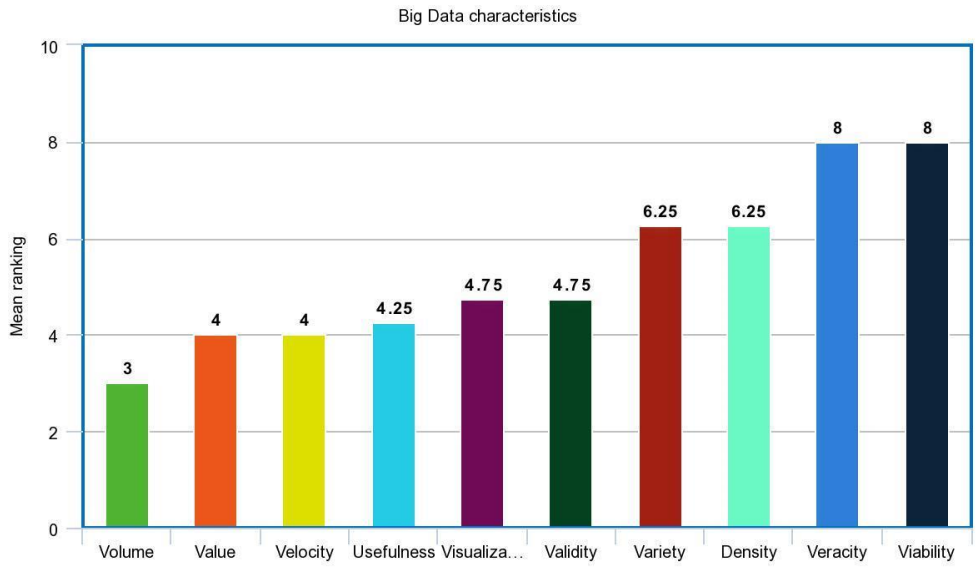
I am aware of the research procedure and I agree to participate in the survey

- o I agree
- o I do not agree

How would you categorize yourself?

- Big Data practitioner - uses Big Data tools and technologies in an organization
- Big Data researcher - does academic research related to Big Data

The following figure represents the result of survey response (mean ranking of individual characteristics of Big Data) received from my study using "Delphi" method.



On a scale of 1-7, how would you agree with the result of the figure above?

1(Extremely disagree) 2(Mostly disagree) 3 (Somewhat disagree) 4(Neither agree nor disagree) 5(Somewhat agree) 6(Mostly agree) 7(Extremely agree)

1 level of agreement 7

Can you please provide your justification for the above rating?

Participants' response

<table>
<tr><th>Round 1</th><th>Round 2</th><th>Round 3</th></tr>
<tr>
<td><u>Participant 1</u>

Big in Volume, Velocity, Variety</td>
<td>Validity
Value
Variety
Velocity
Veracity
Visualization
Volume</td>
<td>
<table>
<tr><td>Visualization</td><td>8</td></tr>
<tr><td>Volume</td><td>1</td></tr>
<tr><td>Validity</td><td>5</td></tr>
<tr><td>Value</td><td>4</td></tr>
<tr><td>Variety</td><td>3</td></tr>
<tr><td>Velocity</td><td>2</td></tr>
<tr><td>Veracity</td><td>6</td></tr>
<tr><td>Viability</td><td>7</td></tr>
<tr><td>Density</td><td>10</td></tr>
<tr><td>Usefulness</td><td>9</td></tr>
</table>
</td>
</tr>
<tr>
<td><u>Participant 2</u>

Large data sets, importance of predicting trends and revealing importance information, statistics</td>
<td>Validity
Variety
Visualization
Volume</td>
<td>
<table>
<tr><td>Visualization</td><td>1</td></tr>
<tr><td>Volume</td><td>2</td></tr>
<tr><td>Validity</td><td>4</td></tr>
<tr><td>Value</td><td>5</td></tr>
<tr><td>Variety</td><td>6</td></tr>
<tr><td>Velocity</td><td>7</td></tr>
<tr><td>Veracity</td><td>9</td></tr>
<tr><td>Viability</td><td>10</td></tr>
<tr><td>Density</td><td>8</td></tr>
<tr><td>Usefulness</td><td>3</td></tr>
</table>
</td>
</tr>
</table>

<table>
<tr><th>Round 1</th><th>Round 2</th><th>Round 3</th></tr>
<tr>
<td><u>Participant 3</u>

Collected originally for one purpose but (re)used/analyzed for different purposes; low insight-to-data ratio, that is, individual rows mean little, but meaning is derived from aggregating and analyzing many data points together; requires batch processes (Hadoop) for typical analysis; typically written once, read many times (such as log files, social media posts, sensor data); vast size makes retrieval of individual records challenging, hence the move toward NoSQL key-value databases.</td>
<td>Value

Velocity

Viability

Visualization

Volume

I would call it "density" or "sparsity" = the ratio of meaningful information to data points. In "small data", like a customer database, each record has meaning of its own. "Big data" is characterized by combining lots of data points together to find a small number of insights. For example, millions of transactions might be analyzed to generate a much smaller number of product recommendations</td>
<td>
<table>
<tr><td>Visualization</td><td>6</td></tr>
<tr><td>Volume</td><td>2</td></tr>
<tr><td>Validity</td><td>7</td></tr>
<tr><td>Value</td><td>5</td></tr>
<tr><td>Variety</td><td>8</td></tr>
<tr><td>Velocity</td><td>3</td></tr>
<tr><td>Veracity</td><td>9</td></tr>
<tr><td>Viability</td><td>10</td></tr>
<tr><td>Density</td><td>1</td></tr>
<tr><td>Usefulness</td><td>4</td></tr>
</table>
</td>
</tr>
<tr>
<td><u>Participant 4</u>

Privacy issues, visualization techniques, sense making</td>
<td>Validity

Value

Veracity

Viability

Visualization

Volume

Usefulness, is it worth analyzing it?</td>
<td>
<table>
<tr><td>Visualization</td><td>4</td></tr>
<tr><td>Volume</td><td>7</td></tr>
<tr><td>Validity</td><td>3</td></tr>
<tr><td>Value</td><td>2</td></tr>
<tr><td>Variety</td><td>8</td></tr>
<tr><td>Velocity</td><td></td></tr>
<tr><td>Veracity</td><td></td></tr>
<tr><td>Viability</td><td>5</td></tr>
<tr><td>Density</td><td>6</td></tr>
<tr><td>Usefulness</td><td>1</td></tr>
</table>
</td>
</tr>
</table>