Big Data: An inventory and the constant search for a definition

Thomas Meier, M.A. and Doc. Ing. Helena Makyšová, PhD.

Abstract: Big data has long been an established vocabulary in industry and science. Although this buzzword is on everyone's lips, the exact meaning is still obscured by many conceptual ambiguities. The term is used synonymously to describe a variety of technological concepts, such as storing data, aggregating and processing it. The cultural, social and economic change goes hand in hand with regular floods of information. Inconsistent, formal definitions of the term led to a discrepancy between researchers and practitioners and thus to an efficient further development of the topic. In this article, the existing literature is checked and existing definitions are analyzed. The hermeneutic approach is intended to show the status quo of current research and thus represent the starting point for showing future developments. Furthermore, a generally valid description of the term is to be provided, which emerges from the essence of existing literature.

Big Data – a term that has been enjoying increasing popularity since 2011/2012, is spreading exponentially and is on everyone's lips, unlike in the pure, defined realm of computer science, even in the media and even in public. There is talk of active media work with data, of large amounts of data – Big Data.[1]Socially, the discrepancy between the social and societal well-being and the misuse of Big Bata repeatedly makes headlines. “Big data: the greater good or invasion of privacy?” [2]Big Data has long since grown beyond the boundaries of the pure IT industry and covers almost all areas of economic and social life. The education sector, medicine, administration and, of course, industry and business face enormous challenges that require a clear definition of the term.

But what exactly is meant by this term? “To make money, you’ve got to predict twothings—what’s going to happen and what people think is going to happen.” (Hal Varian).[3]When the chief economist at Google and professor at the University of California at Berkeley use such words, the topic of Big Bata is not far away. In 2019 alone, the results list included around 73,000 hits of the term Big Data when the term was entered in Google Scholar. The topic has long ceased to be a fad, but has been increasingly the subject of scientific discussion since 2011. At this time, around 530 academic articles were published on the main term Big Bata. This year, 2020, therearealreadyover 4,000.[4]Algorithms hold not only move in the scientific area, they penetrate for a long time till the everyday life. The following probably very often published history shows impressively that the application of Big Data is omnipresent in the everyday life:

The father of a young girl angrily entered a branch of a supermarket chain near Minneapolis (USA). His daughter, who was still in high school, had received coupons for baby clothes and cribs in the mail. The father scolded, 'Do you want to encourage her to get pregnant?' The branch manager was surprised when he saw the girl on the mailing list, apologized and finally removed her from the list. A few weeks later, the father contacted the branch again, informing them that the daughter had actually been pregnant. Apparently the supermarket chain knew about the upcoming events in front of the father. [5]

The concept Big Data is as such, defines as already mentioned, not clearly and just as little the authorship. [6]Basically Big Bata amounts are connected with two ideas: The data storage and the next data analysis. The definition that it would be with Big Data only about the memories of Big Data amounts is enough only beginning-wise for the description and exhausts the potential of Big Data only beginning-wise. This signifies, stored data without their analysis are useless. Consequently this at first rudimentary beginning of a definition is insufficient.

The attempts at definition by some companies

An extended specification for the definition of terms can be defined with the frequently used “three V's” (Volume, Velocity and Variety. This definition can be traced back to the consulting firm Gartner, a provider of market research results and analyzes of developments in the IT industry. The following figure visualizes the concept of the 3 V's and shows their dimensions.

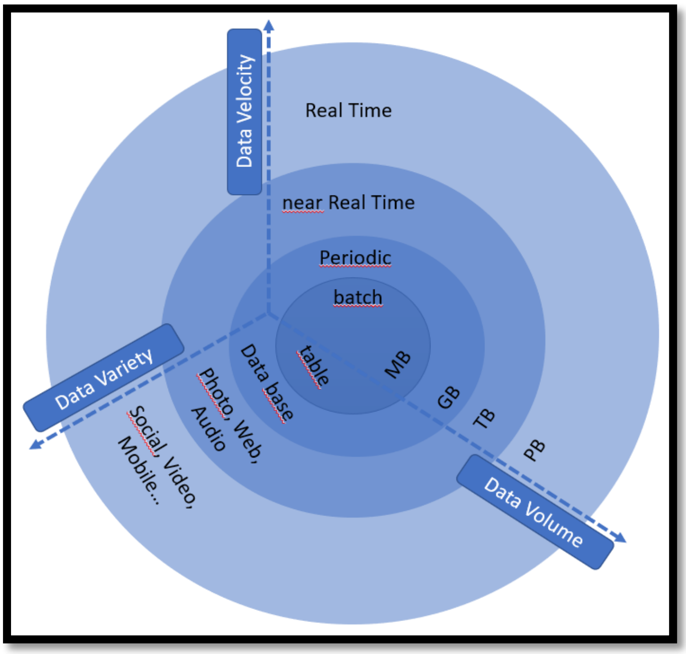

Figure 1: Big Data 3V model - The conventional V‘s of Big Data(own representation)

The characteristic of the 'first' V’s 'volume is already hidden in the name of Big Data and defines a large volume of data. For example, a text file consists of a few kilobytes, an audio file can easily reach a few megabytes in size, and a full-length movie usually consists of a few gigabytes. Data Velocity describes the enormous speed of data processing and Data Variety the data diversity. In the early days of data processing, companies analyzed their data using the so-called batch process. Here, a data block was sent to a server as part of an order and then waited for the result to be transmitted. This procedure works if the incoming data rate is slower than the batch processing rate and the result can be used meaningfully despite the delay. With the increase in data sources, such as social and, above all, mobile applications, the batch process has reached its technical limits and is no longer valid. The data is now streamed to the server in real time and the delay in transmitting the results is negligible. This requirement is essential for the benefit of the results. High-frequency trading on the stock exchange provides examples of this. The variety of data has changed from simple databases or Excel tables to dozens of formats. The structure of the input format could be imposed in the past to keep control over the analysis. This doctrine is history.

Gartner postulates further: „Technology alone is insufficient to solve the bigger problem of data and analytics governance“. [7]According to Gartner analyst Doug Laney, Big Data is simply a definition of the amount of data that appears larger than you are used to. [8]Another Gartner survey found that companies believe poor data quality cost them an average of $ 11.8 million in 2018.[9]

The three characteristics were later supplemented by the perspectives of value and freedom from contradictions. [10]Other authors added attributes, how the truthfulness or credibility (Veracity). [11]

Current state of research

Now that the Gartner 3 V model has been expanded to 5 V’s and later to 7 V’s, we are now talking about 10 properties of Big Data that are intended to define the term in more detail:[12] Volume, Velocity, Variety, Variability, Veracity, Validity, Vulnerability, Volatility, Visualization and Value. Some of the “V’s” have already been explained, in summary follows a descriptive overview:

- Volume: 90% of the today's data were provided during the last years. For instance, the total number of the stored photos amounted in 2016 to 3.9 trillions from what only in 2016 1.1 trillions photos were taken up.
- Velocity: Google processes about 40000 searching inquiries per second.
- Variability: Describes the diversity or the different origins of the data.
- Veracity: In a broader sense, refers to the validity or trust in the data. This usually decreases when one or all of the properties mentioned increase.
- Validity: Is the measure of how accurate the data is for the intended purpose. According to Forbes, data cleansing causes a lot of time for data scientists. [13]
- Vulnerability: Big Data causes worldwide security doubt. Offense in connection with Big Data is therefore already in agenda. For hacker's attacks a whole list only of the last three years can be provided.
- Volatility: Besides, the question after the age of the data plays a determining role, when they can be classified no more as useful, have become outdated thus. However, regulations count to many areas of the economy and the right how long data must be kept. On account of the high volumes and processing speed, must be weighed out carefully which data must be available at which time – an aspect which pulls a cost question after itself in the consequence.

- Visualization: Represents the challenge, e.g. Visualize billions of data points. Conventional diagrams fail here. Data clusters or the use of tree maps, sunbursts or cone trees help here.
- Value: Describes the most important property for the business world: the economic benefit. This can be to better understand customers, to optimize and improve processes and machines.

Oracle: A different view of things

Again, the US software and hardware manufacturer Oracle completely avoids the use of V's in definition approaches. [14]The group describes Big Data as a derivation of the values from traditional data bank decisions, supplemental by structured data from new springs like Blogs, images and other forms of data which vary in size, structure or if necessary other factors. Oracles definition of Big Data is therefore the inclusion of additional data springs in existing operations and focuses therefore the infrastructure of the data. Besides, Oracle places on technologies like NoSQL, Hadoop, HDFS, R and relational data banks. Basically the definition of Oracle is easier to apply than, for instance, those of the house Gardner, lacks, however, it her also of the quantification. The question after „What is big? “, also remains with Oracle unanswered.

Intel:A little more concrete statement with regard to the data size

One of few enterprises which published concrete figures concerning the subject is the solid-state manufacturer Intel. The enterprise connects Big Data with “generating a median of 300 terabytes (TB) of data weekly”.[15]Intel quantifies, differently than Oracle, with the help of the experiences and measurements of his business partners the data amounts. The enterprise refers to the fact that it concerns at the questioned enterprises primarily unstructured data and the main focus lays therefore on the analysis of the data. Thus the analysis of the data moves at the enterprises cooperating with Intel in the focus which are generated with a rate by 500 TB per week. To the comparison: The big optical measurement telescope (LSST) in Chile generates every week approx. 210 TB data during ten years of durable sky investigation. [16]

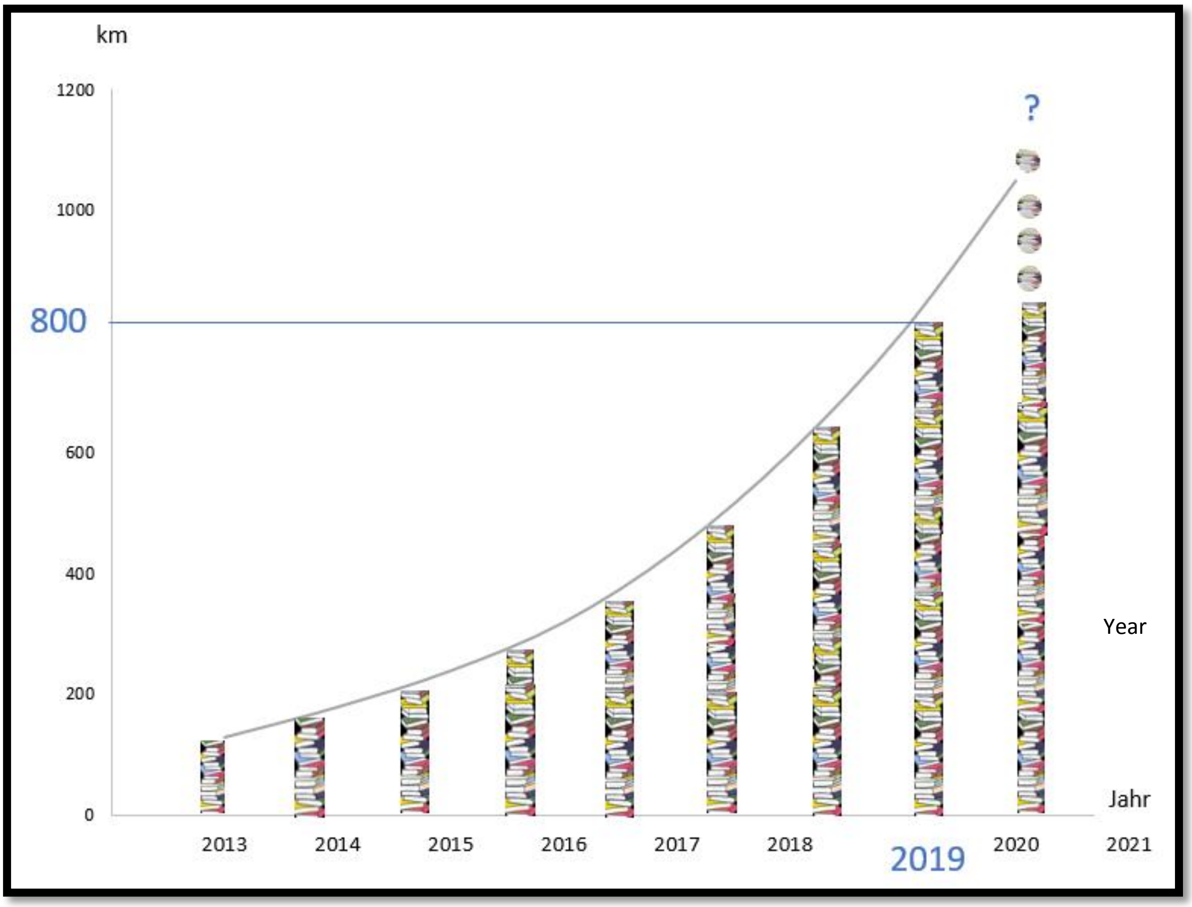

Figure 2: The data amounts develop exponentially (own representation)

Picture 2: The data amounts develop exponentially (own representation)

Worldwide, the volume of data is growing exponentially, which is shown in Figure 2. For example, over 2 trillion searches are made online every year.[17]Assuming that a search query contained an average of four words and an approximately 4 cm thick book had approximately 400,000 words, a tower of 800 km would be created if all search queries were stacked in this way.[18]
The most familiar data type are no pictures like with the telescope, but commercial transactions which are stored in relational data banks. They are complemented by e-mails, sensor data, Blogs and social media.

Microsoft – Computing power as a decisive factor

The software giant Microsoft also defines the term Big Data quite succinctly. “Big data is the term increasingly used to describe the process of applying serious computing power - the latest in machine learning and artificial intelligence - to seriously massive and often highly complex sets of information“. [19]Microsoft has clearly focused on computing power since this time of rewriting Big Bata. This will be necessary to structure and analyze large amounts of data. In this context, two other terms are listed: machine learning and artificial intelligence. So far, these two aspects have not yet been associated with Big Data, but should in future be crucial components of a definition.

The conceptual puzzle is further completed

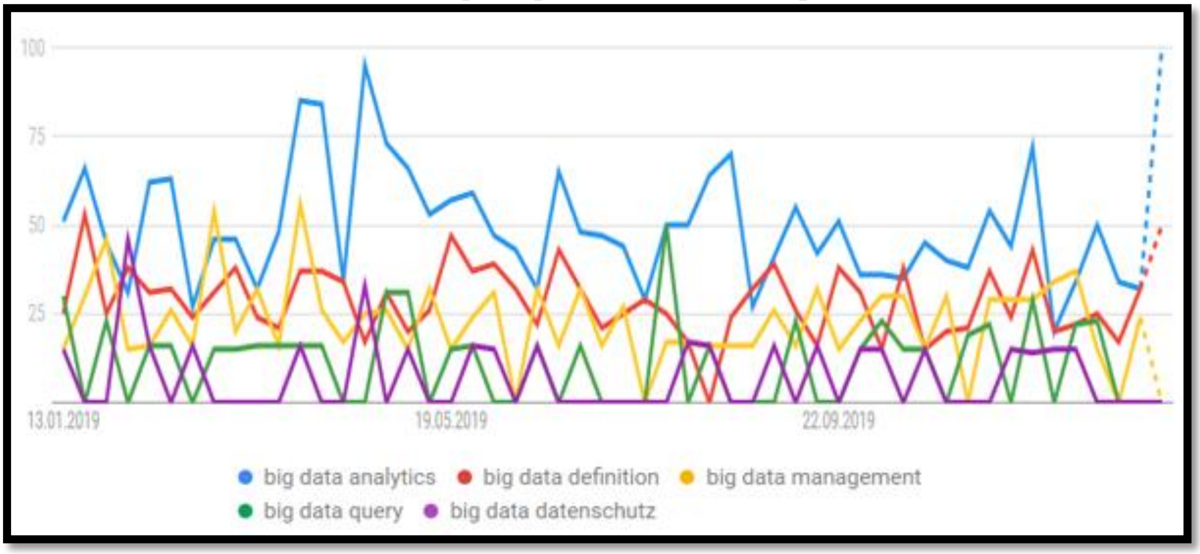

Figure 3: Business interest over time (Google Trends 2019)

The search terms that were requested in connection with Big Data as related topics at Google Trends in the period from January to Dec 2019 in Germany provide information. [20] These were in descending order Big Data Analytics, Big Data Definition, Big Data Management, Big Data Query and Big Data Data Protection.

The diagram shows the relative query intensity for the main topic Big Data. The topics asked mainly provide information about what a definition of Big Data should contain in the core. They also underline Microsoft's definition approaches.

In principle, scientific definitions have only appeared to a small extent.[21]The term polystructured data known in German-speaking countries makes it clear that not only classic structured data from ERP systems and other sources are used, but also videos, images, texts or other publications. [22]The partially or unstructured data experience their actual value through convergence. In principle, it is no longer a matter of generating aggregated total tables, but of including the individual process except for one document. [23]In this way, patterns can be recognized and correlations can be drawn from them, from which conclusions can be drawn about buying behavior, or better future buying behavior, price developments, logistical processes or even political developments. As a result of this development, our vocabulary is adapting more and more. Today we use terms that didn't exist 10 years ago. An aspect that is of great importance in the definition of terms.The technical periodical „ComputerBild“ published a Big Data glossary, with those the most common words: [23]

- Ad targeting - the effort to attract potential customers' attention through targeted advertising.
- Algorithm and Analytics - the mathematical and technical prerequisite for analyzing data with software.

- Automatic Identification and Capture (AIDC) - any method of automatically identifying and collecting data about a given situation.
- Behavioral Analytics - gathering information about human behavior.
- Business Intelligence (BI) - the definition for the identification, origin and analysis of the data.
- Cassandra - this distributed database management system is responsible for very large structured databases ('NoSQL' database system) on an open source basis (Apache).
- Clickstream Analytics - describes the evaluation of a user's web activity.
- Data aggregation - collecting data from different sources.
- Distributed Object - means software that allows another computer to work with distributed objects.
- Predictive Analytics - describes a form of analysis that uses statistical functions to identify trends and predict events.
- Recommendation engine - algorithms evaluate customer orders in order to be able to offer additional products.
- Sentiment Analysis - Entries or posts from users of social networks about a product or a company are statistically evaluated here.
- Variable pricing - this requires real-time monitoring of consumption and inventory. The purchase price of a product follows supply and demand.

This extract from the concept world of Big Data makes clear the complexity of this subject. Countless areas of application can be derived from it and a clear containment develops to date as inadequate.

Resume

The previously mentioned attempts to adequately define Big Data aimed essentially at the amount of data and its analysis. A definition that focuses on the complexity of the data is less common. The method for an integrated knowledge environment (MIKE 2.0), an open source delivery method, could lead to conflicting ideas. “Big data can be very small and not all large datasets are big”.[24]This approach advocates the complexity of data. The MIKE project also argues that a high level of permutation and interaction within a data set is critical when defining Big Data. It follows that handling with conservative tools is difficult or that these conventional tools cannot be used sensibly at all.

This definition is supported by NIST (National Institute of Standards and Technology), who postulate that this is to be understood as Big Data data that: “… exceed (s) the capacity or capability of current or conventional methods and systems.” [25]However, the second time one looks this definition is not comprehensive enough. She subordinates as it were the attempt to reduce the amount of the attacking data and to compensate for this with on and on improved calculation standards.

Despite their scope and differences, the definitions shown so far show clearly identifiable similarities. The size of the data plays the most dominant role when it comes to conceptual limitation. Although the literature also uses a different weighting scale.

Furthermore, the structure, or rather the complexity, is another important argument for a definition. In particular, the permutation of a data set is to be considered as a critical factor since this is crucial for the usability of the data.

As a third point, according to the companies, the technology that enables the processing of a data set is a critical aspect. This third point opens the door to further topics that will move around the topic of big data like satellites.

AI (artificial intelligence) or BI (Business Intelligence) have long been separate areas, which are in direct connection with Big Data and its implementation will be Substantially dependent on technology.

In summary, it can be said that Big Data includes the storage and analysis of large amounts of data, it is structured with a variety of techniques and mathematical-statistical foundations and is integrated into social and economic life. Big data is no longer a stand-alone buzzword, but rather a framework of our society. A billion dollar industry has developed around this term. For example, 'Data Scientist' is one of the most desirable jobs of the past decade. [26]From an analytical point of view, it can be observed and also expected that Big Data, with all its facets, tools, technologies and know-how, will establish itself in all of our lives. So far, however, numerous companies have failed due to the challenges that the broad field - Big Data - brings with it. It will be increasingly necessary to analyzeBigData in the context of different industries and to break down the resulting knowledge into individual processes.

The article is a part of the research project VEGA Nb.1/0373/18 "Big data analytics as a tool for increasing the competitiveness of enterprises and supporting informed decisions"

Literatur

[1]. Dander, V.: [KRB] 2014c.«Die Kunst des Reg (istr) ierens mit Big Data. Ein Versuch über Digitale Selbstverteidigung und Aktive Medienarbeit mit Daten». medienimpulse. Beiträge zur Medienpädagogik 4 (Steuerung, Kontrolle, Disziplin/Medienpädagogische Perspektiven auf Medien und/der Überwachung): 1–13. **MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung** (2018).

[2]. Chatterjee, P.: Big data: the greater good or invasion of privacy. **The Guardian** 12 (2013).

[3]. Buhl, H; Röglinger, M; Moser, F; Heidemann, J.: Big Data—Ein (ir-) relevanter Modebegriff für Wissenschaft und Praxis? **Wirtschaftsinformatik & Management** 5 (2013) 2, S. 24–31.

[4]. Kolany-Raiser, B; Heil, R; Orwat, C; Hoeren, T.: Big Data und Gesellschaft: Eine multidisziplinäre Annäherung. Springer Fachmedien Wiesbaden 2018.

[5]. Kashmir Hill: How target figured out a teen girl was pregnant before her father did 2012.

[6]. Klein, D; Tran-Gia, P; Hartmann, M.: Big Data. **Informatik Spektrum** 36 (2013) 3, S. 319–323.

[7]. Susan Moore: MDM is a complex undertaking. 2018.

[8]. Beyer, M; Laney, D.: The importance of'big data': A definition. Stamford, CT: Gartner Inc. Report ID number: G00235055 2012.

[9]. Gartner Special Reports. https://www.gartner.com/en/products/special-reports. Zugriff am 21.01.2020.

[10]. Bachmann, R; Kemper, G; Gerzer, T.: Big Data - Fluch oder Segen?: mitp Professional. Verlagsgruppe Hüthig Jehle Rehm, s.l. 2014.

[11]. Beyer, M; Laney, D.: The importance of'big data': A definition. Stamford, CT: Gartner. Retrieved June 22, 2014 2012.

[12]. The 10 Vs, Issues and Challenges of Big Data. ACM 2018.

[13]. Wilbik, A.: FSS++ Workshop Report: Handling Uncertainty for Data Quality Management. **arXiv preprint arXiv:1810.02091** (2018).

[14]. Dijcks, J.-P.: Oracle: Big data for the enterprise. **Oracle white paper** (2012), S. 16.

[15]. Ward, J; Barker, A.: Undefined By Data: A Survey of Big Data Definitions 2013.

[16]. Telescope, L.: Press Releases | Legacy Survey of Space and Time. https://www.lsst.org/news/press_releases. Zugriff am 17.01.2020.

[17]. Sullivan, D.: Google now handles at least 2 trillion searches per year. **Search Engine Land** 24 (2016).

[18]. Kersting, K; Lampert, C; Rothkopf, C. (Hrsg.): Wie Maschinen lernen. Künstliche Intelligenz verständlich erklärt. Springer Fachmedien Wiesbaden GmbH; Springer, Wiesbaden 2020.

[19]. Redmond, W.: The Big Bang: How the Big Data Explosion Is Changing the World 2012.

[20]. Google Trends 2019.

[21]. Loos, P; Lechtenbörger, J; Vossen, G; Zeier, A; Krüger, J; Müller, J; Lehner, W; Kossmann, D; Fabian, B; Günther, O.: In-Memory-Datenmanagement in betrieblichen Anwendungssystemen. **Wirtschaftsinformatik** 53 (2011) 6, S. 383–390.

[22]. Bange, C; Grosser, T; Janoschek, N.: Big Data Survey Europe-Nutzung, Technologie Und Budgets Europäischer Best Practice Unternehmen. **'^'eds.'): Book Big Data Survey Europe-Nutzung, Technologie Und Budgets Europäischer Best Practice Unternehmen, Business Application Research Center (BARC), Würzburg** (2013).

[23]. Kriemhilde Klippstätter: Big Data FAQs: Was ist was bei Big Data? **Computerwoche** 2016 (2016).

[24]. Rindler, A; McLowry, S; Hillard, R.: Big Data Definition. MIKE2. 0, the open source methodology for Information Development 2013.

[25]. NIST: "NIST Big Data Working Group (NBD-WG). [online] http://bigdatawg.nist.gov/home.php" 2019.

[26]. Ranjan, J.: The 10 Vs of Big Data framework in the Context of 5 Industry Verticals. **Productivity** 59 (2019) 4.