Defining Big Data

Isitor Emmanuel
Staffordshire University
Stoke on Trent
Staffordshire
+44(0)1782 353463
Emmanuelemeke.isitor@research.staffs.ac.uk

Dr Clare Stanier
Staffordshire University
Stoke on Trent
Staffordshire
+44(01782) 353463
c.stanier@staffs.ac.uk

ABSTRACT
As Big Data becomes better understood, there is a need for a comprehensive definition of Big Data to support work in fields such as data quality for Big Data. Existing definitions of Big Data define Big Data by comparison with existing, usually relational, definitions, or define Big Data in terms of data characteristics or use an approach which combines data characteristics with the Big Data environment. In this paper we examine existing definitions of Big Data and discuss the strengths and limitations of the different approaches, with particular reference to issues related to data quality in Big Data. We identify the issues presented by incomplete or inconsistent definitions. We propose an alternative definition and relate this definition to our work on quality in Big Data.

CCS Concepts
•**Information systems→Database design and models**

Keywords
Big Data; Data Quality; Data Quality Dimensions; Big Data characteristics

1. INTRODUCTION
This paper discusses the difficulty of understanding what is meant by Big Data, reviews existing definitions of Big Data and proposes an alternative definition of Big Data. The motivation for this work on the definition of Big Data comes from our work on data quality in Big Data. Data quality dimensions (DQDs), which describe the characteristics data should possess to be regarded as of good quality, are an accepted tool in the literature on data quality [1]. Developing data quality dimensions for Big Data is challenging because of the lack of agreement as to what constitutes Big Data and how to recognise Big Data and also because some well known Big Data definitions include elements which we argue are data quality characteristics rather than Big Data characteristics. This introduces redundancy into Big Data DQDs and makes the process of identifying Big Data quality concepts and tools more difficult. In this paper we develop a definition of Big Data which recognises the complexity of Big Data and distinguishes between the characteristics of Big Data and data quality characteristics.

The paper is structured as follows: Section 2 discusses the issues involved in developing a definition for Big Data, Section 3 discusses current definitions of Big Data; Section 4 develops a definition for Big Data based on the literature; Section 5 gives conclusions and proposals for future work.

DOI: http://dx.doi.org/10.1145/12345.67890

2. Defining Big Data
The term 'Big Data' is in general use to describe the collection, processing, analysis and visualisation associated with very large data sets but the concept of Big Data has proved difficult to define [2,3,4]. There are a range of formal [5,6,7] and informal [8,9] definitions, which typically have some elements in common and some areas of difference. It has been suggested that all the alternative definitions of Big Data should be embraced [10] but it is argued here that the number and variety of definitions means that this would introduce inconsistency and duplication and that a shared understanding of Big Data is needed to support work in fields such as Big Data data quality and Big Data analytics. The development of DQDs for Big Data, for example, requires an understanding of the elements which are regarded as specific to Big Data and the elements which are regarded as data quality markers for Big Data. The approach known as the 3Vs (volume, velocity, variety) is widely used, particularly in the practitioner and technical literature. Volume, Velocity and Variety are not by themselves regarded as sufficient to define Big Data [7] and the terms also require definition. 'Volume' for example is understood differently in different contexts. The 3 Vs approach focuses on the characteristics of data and does not consider the wider Big Data environment.

'Big Data' is used as an umbrella term to cover a range of data, technologies and applications. This contrasts with previous data management approaches which are typically based around data models that define the structure of and operations on a database [12] and specify elements such as data structures and data operators. The best known example of a data model in this sense is the relational data model [13] while other approaches include development based on an underpinning theory such as graph theory [14] or on a programming paradigm such as object oriented. Object oriented database (OODB) development illustrates the role of the Big Data environment since OODB assume an understanding not only of object databases but of the object oriented environment and the concepts and characteristics that support OO development. OODB cannot be understood separately from the OO paradigm and we argue that, similarly, Big Data cannot be understood only in terms of data and must be seen in the context of the environment of Big Data. Big Data is implementation driven and continually evolving and is not based around a single unifying theory or paradigm. This means that defining Big Data presents a number of challenges in that the process is bottom up and iterative and is potentially open ended.

There are multiple definitions of Big Data but it is possible to identify a number of themes in the Big Data Literature. Hu [10] arranged definitions of Big Data into three categories, comparative, based around the comparison of relational and Big Data

characteristics, attribute based developed from the characteristics of Big Data and architectural which emphasises computer architecture and technical elements. In the following section we examine existing definitions of Big Data, using the Comparative and Attribute based headings suggested by Hu but for reasons discussed in section 3.3, using an Environmental rather than an Architectural category.

3. Existing Definitions of Big Data

3.1 Comparative Approach to Definition

The term Big Data, in its present sense, is said to have been introduced for the first time in 2008 by the Gartner Group [15]. At the start of the Big Data era, the most commonly used data management systems were based on the relational data model and one approach to understanding the new paradigm was to discuss Big Data in terms of differences with relational systems and the opportunities offered by the new technology. An influential report by the McKinsey Institute described Big Data as datasets that "*cannot be processed stored and analysed by traditional data management technologies"* [16, p.2]. Traditional, in this context, implies relational. Jacobs [17] suggests a rolling definition in which Big Data is data whose "size forces us to look beyond the tried-and-true methods that are prevalent at that time" [17, p.44]. The comparative approach to defining Big Data is designed to explore the possibilities of Big Data rather than providing a formal definition and the focus is primarily upon relative processing capacities. Developments in technologies mean that enhanced relational systems can handle much larger data volumes than previous relational systems and Big Data Business Intelligence capabilities (BI) are increasingly being incorporated into traditional, relational BI systems meaning that the boundaries between processing capabilities in Big Data and traditional systems is less clear. The chief limitations of a comparative approach to the definition of Big Data are that it relies on a consensus as to what constitutes traditional or state of the art without defining what is meant by these terms and that it defines Big Data in the context of (primarily) relational functionality and capabilities. This limits the scope and potentially restricts the understanding of Big Data. With reference to measuring data quality in Big Data, for example, discussing Big Data in terms of relational capabilities makes it more difficult to identify data issues specific to Big Data.

3.2 Attribute Based Definitions

The three Vs, Volume, Velocity and Variety, are widely accepted as the basis for the definition of Big Data [4,5,6,7,18, 19,20] usually in the Gartner glossary sense of high volume, high velocity and high variety [15]. The initial 3Vs definition, developed to explain the technical and business implications of newer data management strategies, has been extended over time to reflect different approaches to Big Data. IBM, in a vendor context proposed Veracity as a fourth V [21]. Saha, working in a data quality context, also proposed veracity as the fourth V [22]. A 5Vs approach, which added value and veracity was proposed in the context of scientific data [6]. From a commercial perspective, the 3 Vs were extended with Veracity, Variability and Visualization [23]. A 7Vs approach, which uses Health Care as an example case study, extends the 3Vs with Veracity, Validity, Volatility and Value [4]. An alternative 7Vs definition proposes Volume, Variety, Velocity (the 3 Vs) and Value, Veracity, Variability and Complexity [5].

There is a large degree of consensus around the definition of the original 3Vs. Volume has been described as the most visible big data characteristic [2] and is usually understood as data generated from different sources [5] or as referring to size and scale of data [6], summarised as the magnitude of data [20] although the same source noted that definitions of volume are relative and that it is not possible to define a specific threshold for data volumes in Big Data. The difficulty of defining volume in a Big Data context means that although recognised as a characteristic or attribute of Big Data, it is sometimes defined in terms of traditional (relational) systems [24], reintroducing a comparative element. Velocity is understood in relation to the speed at which data is received, stored, processed and analysed [5,20] sometimes with specific reference to real time or near real time [6] and to the streaming of data [25]. Chen [19] suggests a different but related definition of velocity, arguing that data collection and analysis must be conducted rapidly and in a timely manner to allow for the commercial use of Big Data. Definitions of variety usually emphasise the role of unstructured and semi-structured data [6,20,24] sometimes in the sense that variety is understood in opposition to relational data [4]

There is less consensus around the extensions to the 3Vs , but the most widely used extensions include veracity and value. Veracity has been defined as relating to data certainty and trustworthiness in terms of collection, processing methods, trusted infrastructure and data origin [5]. It has been suggested that veracity in the sense originally defined by IBM relates to the fact that Big Data is required to deal with imprecise and uncertain data [20]. In the context of scientific data, Veracity was partly defined as 'data consistency .. what can be defined by their statistical reliability' [6, p. 50] which might not be a valid definition in the context of user generated data such as twitter feeds. The understanding of veracity appears to be context dependent. Validity was proposed as an addition to veracity, where veracity means the truthfulness of data and validity means the correctness and accuracy of the data with regard to the intended usage [4]. Value has been defined in terms of 'low value density', the concept that the value of Big Data is low in relation to its volume but that high value can be achieved by processing large volumes of data [20] or as the desired outcome of data processing [4] or as the added value that the data can contribute [6]. The concepts are similar but subtly different in that low value density is a distinguishing characteristic of the type of analysis operations carried out while added value is an end product. Volatility was proposed as a 6th 'V' in relation to data retention policies [4] but in this context it is a data management issue not a characteristic of Big Data. Variability was proposed to reflect the fact that there may be peaks in data load [5] but variability in this sense is not specific to Big Data.

Attribute based definitions have three important limitations. Firstly, attributes can be added to without restriction meaning that there is no stopping point for extensions to the 3 Vs. As discussed in the previous section, multiple definitions may be in use. Secondly, as there is no single agreed definition of attributes, particularly for the extensions to the 3 Vs, the same attribute may be understood differently, as with value and veracity. This means that multiple definitions may be in use, limiting shared understanding. Thirdly, the extensions to the original 3 Vs suggest a confusion between the attributes of Big Data and data quality characteristics. This makes it difficult to distinguish between the identifying characteristics of Big Data and quality characteristics which are desirable in all data.

Data quality attributes are defined through Data Quality Dimensions, a widely used tool for data quality [26,27]. The naming convention for Big Data, which requires Big Data characteristics to begin with the letter V, gave rise to the terms Veracity [5,6,20] and validity [6], proposed as extensions to the 3 Vs. The nomenclature in DQDs is different meaning that the terms validity and veracity are not in use but the concepts represented by these terms are present in DQDs through a range of characteristics

such as accuracy, reliability, completeness, fitness and usability [1, 26,27,28,29]. Value [20,4] or Value-added [6], has been added as a fourth or fifth V to Big Data definitions but value-added exists as a contextual DQD in the earliest [1] and more recent work on DQDs [19]. Volatility, also proposed as an addition to the 3 Vs, is present in DQDs as an element of the DQD of timeliness [1]. Including data quality elements in the definition of Big Data reduces the distinctiveness of Big Data and makes the identification of data quality issues in Big Data more difficult. It raises the question of whether Veracity, for example, should be regarded as an intrinsic element of Big Data, without which Big Data is not Big Data, or whether it is a quality goal which can be achieved provided certain criteria are met.

3.3 Environmental Definitions

Hu [10] listed a definition category described as an architectural approach to Big Data definition and illustrated this with reference to a NIST (National Institute of Standards and Technology) presentation that linked the 3 Vs to the requirement to use horizontal scaling for efficient processing of data [30]. A more formal definition was developed as part of the NIST Big Data Interoperability Framework; this definition split the Big Data concept into two key elements, the characteristics of Big Data and the Big Data Paradigm [31]. A related definition, described by the authors as structural [7], gave a more comprehensive definition of Big Data which covered a wider range of elements. We also include in the environmental category an approach which linked the 3 Vs to the requirements and processes of Big Data analytics [20]. We use the term environmental to reflect the fact that the understanding of Big Data is increasingly moving beyond identifying the attributes that define Big Data to include recognition of the architectures, processing and applications of Big Data.

The NIST Taxonomies subgroup definition of the characteristics of Big Data, emphasised the processing and architectural consequences of working with datasets built on the 3Vs and that extensive datasets "require a scalable architecture for efficient storage, manipulation and analysis" [31, p. 5]. The NIST approach sees the Big Data paradigm in terms of a shift away from monolithic vertically scaled systems to distributing data across horizontally scaled independent resources to achieve scalability [31]. The NIST approach makes it clear that the Big Data Paradigm is seen as explicitly non relational [31] but horizontal scaling and parallel processing are also used in distributed relational systems. There is clear recognition that Big Data will have implications for data analytics but Big Data analytics does not form part of the NIST definition [31].

The structural definition referred to above was produced by a member of the NIST working group separately [7]. This was a five part definition which linked the 6V approach (Volume, Variety, Velocity, Value, Variety, Veracity) to a requirement for cost-effective innovative analytics to provide enhanced insight, supported by new data models and new infrastructure tools to support data acquisition from a range of sources and data delivery in different formats to different data consumers [7]. Value, Veracity and Variability were seen as acquired features which depend on a specific process or model and are dependent on the data context [7]. This approach combines the goals of Big Data operations with a description of Big Data but provides an overview of Big Data and Big Data targets rather than a definition. Elements such as cost-effective and enhanced insight are arguably the intended outcomes/goals of all data processing and are not exclusive to Big Data. The reference to new data models and new infrastructure introduces a comparative element as the definition of new will change over time. The inclusion of the role of analytics recognises the need to include operations upon Big Data as part of the discussion of Big Data.

3.4 Discussion

Existing definitions of Big Data have been discussed under the headings of comparative, attribute based and environmental definitions. The comparative approach relies on a consensus as to what constitutes traditional approaches and does not provide clear boundaries for distinguishing between Big Data and enhanced relational processing; the focus is on identifying the potential of Big Data rather than defining the elements that constitute Big Data. The attribute based approach identifies some of the key characteristics of Big Data but terms may have multiple definitions and there is a tendency for the number of attributes in the definition to grow. Some of the attributes used are better viewed as data quality dimensions rather than as properties specific to Big Data. Attribute based definitions do not take account of the wider context of Big Data. The environmental approaches discussed in 3.3 illustrate a developing consensus that attribute based definitions alone are not sufficient and that Big Data cannot be fully understood in isolation from the technical environment and the uses of Big Data. However, the environmental definitions discussed have a number of limitations in that they include attributes which we regard as DQD elements rather than Big Data characteristics. Environmental definitions may be linked to implementation elements which are not Big Data specific or to specific Big Data implementations which may change over time. Discussion of the uses of Big Data is heavily focused on analytics and there is limited recognition of other types of applications for Big Data. One limitation common to all Big Data definitions is that definitions are empirical, and extrapolated from existing systems; this reflects the implementation driven evolution of Big Data.

4. Definition of Big Data

4.1 Definition Approach

We describe Big Data in terms of the *data characteristics* of Big Data, the 'what' of Big Data; the *processing and supporting architectures* used with the data, the 'how' of Big Data; the *applications of Big Data*, the 'why' of Big Data. Fig. 1 illustrates our approach

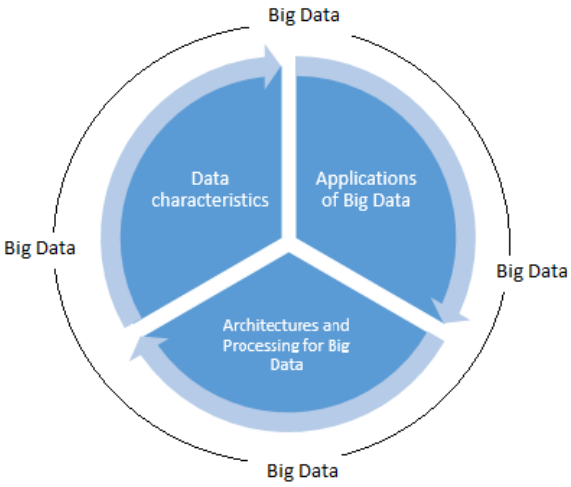

Figure 1: The components of a Big Data Definition

As discussed in section 2, Big Data cannot be defined like relational data with reference to a single unifying underpinning theory. The *data characteristics* element of the definition describes the data component of Big Data. *Data Characteristics,* as discussed in sections 3.2 and 3.4 do not by themselves provide a sufficient definition of Big Data since data attributes may not be unique to Big Data and understanding of data attributes may evolve over time. Technical advances mean that 'Volume', for example, is a constantly evolving concept. We argue that the data characteristics of Big Data cannot be understood separately from the Big Data environment which includes the application context and for this reason include *Architectures and processing* and *Applications of Big Data* as part of our definition. This reflects the fact that the elements of Big Data are interdependent and need to be considered together rather than in isolation. The volume of Big Data, for example, determines the processing requirements but the processing determines how data is supplied to applications which in turn drive the demand for data volume.

We next describe each of the components of the definition of Big Data in terms of data characteristics, architectures and processing and the applications of Big Data and then provide a definition of Big Data

4.2 Data Characteristics of Big Data

For the *data characteristics* element of the definition, the *'what'* of Big Data, we adopt the traditional 3 'V's approach (Volume, Variety, Velocity) and give a definition of these elements below. We do not include any of the proposed extensions to the 3 'V's in the data characteristic element as we regard characteristics such as value, for example, as data quality characteristics rather than elements uniquely descriptive of Big Data.

4.2.1 Volume

Volume is widely accepted as a key characteristic of Big Data. However, it is also accepted that volume cannot be defined in terms of a fixed quantity of data [20] and for this reason, volume is usually defined in the literature in comparative terms, as volumes too large to be handled by traditional processing capabilities [16] or in terms of the architecture or processing required [31]. As discussed in 4.1, the data characteristics of Big Data cannot be understood separately from the Big Data environment and application context. We understand 'Volume' as the requirement for high volume storage or processing space to support high volumes of data.

4.2.2 Velocity

Velocity is classically seen as a Big Data characteristic since systems dealing with high volumes of data are associated with high velocity of data processing [5, 20]. Velocity is also often understood to refer to high velocity of data acquisition but not all data has velocity in this sense. We understand velocity as a Big Data characteristic in the accepted sense of data that is generated and/or processed at high velocity and we understand the data characteristic of velocity in Big Data as support for data which is generated and processed at high velocity.

4.2.3 Variety

Variety is sometimes used as a shorthand to indicate that Big Data supports unstructured and semi-structured data as against (relational) structured data [4]. As discussed in 3.2, we argue that Big Data may include structured and relational data and that different data formats may exist within the same Big Data set. We understand the data characteristic of variety in Big Data as support for variety in data formats.

4.3 Architectures and Processing

Definitions of Big Data which include architectural elements are typically at a lower level of abstraction than attribute based definitions and may specify physical elements such as horizontal scaling and parallel processing [31] and even specific algorithms such as Hadoop MapReduce. Specifying a particular implementation strategy becomes a snapshot of the state of the art in Big Data rather than a definition. The MapReduce algorithm, for example, is strongly associated with Big Data but has acknowledged limitations [34] and is likely to be amended or replaced over time; variants of Hadoop MapReduce such as Spark have been discussed for a number of years [35]. As architectures and processing evolve to meet new challenges, implementation based definitions become out of date. A high level of abstraction is needed in this element.

We adopt a data driven approach to specifying the *'how'* of the Big Data environment. This recognises that Big Data cannot be understood in isolation from the Big Data implementation environment and views the architectural and processing element as driven by the data requirements rather than as characteristics which define Big Data. In this approach to definition, Hadoop, for example, is a technology required to support the data characteristics of Big Data rather than a technology which defines Big Data. This approach means that the understanding of Big Data is not tied to a particular implementation technology, recognising that technologies tend to converge.

We understand Big Data architectures and Big Data processing as scalable architectures which support the processing requirements of data which has high volume and which may have a variety of data formats and may include high velocity data acquisition and processing.

4.4 Applications of Big Data

The applications of Big Data, form the *'why'* of Big Data, the third element in our definition of Big Data. Most definitions of Big Data focus on data characteristics only or on data characteristics and architectures and processing although a discussion of Big Data analytics forms part of a number of definitions of Big Data [7, 20]. We include applications of Big Data in our definition because the data characteristics, architectures, processes and uses of Big Data interact. Architecture and processing is driven by the way in which Big Data is used as well as by the characteristics of Big Data. Big Data applications are not limited to analytics and include a range of use cases such as Big Data OLTP [32] and applications which are better described as reporting than as an analysis applications. The use cases of Big Data are best seen as a spectrum ranging from, at the reporting end, Big Data OLTP [32] to analytics which are closer to artifical intelligence at the other end of the spectrum [33]. For this reason we describe the use cases of Big Data as the Applications of Big Data rather than as Big Data Analytics.

We understand Big Data applications as a family of applications which operates upon Big Data and which includes, but is not limited to, Big Data analytics.

4.5 Defining Big Data

Based on the existing definitions of Big Data reviewed in this paper, we understand Big Data as:

The term Big Data describes a data environment in which scalable architectures support the requirements of analytical and other applications which process, with high velocity, high volume data which may have a variety of data formats and which may include high velocity data acquisition.

5. Conclusions and Future Work

We identified as one of the limitations of some existing Big Data definitions that the focus on data characteristics meant that the importance of the Big Data environment was not recognised. We propose an understanding of Big Data which recognises the role of the Big Data environment but provides a high level view which is not dependent upon current implementation technologies. We include in our discussion the applications of Big Data to emphasise the importance of the use cases of Big Data but do not link this to any specific implementation technology or use case, recognising that the uses of Big Data may change over time. We base our definition on the traditional 3 Vs approach but extend this approach to recognise that while Big Data has the potential to support velocity and variety, not all Big Data applications will require the characteristics of velocity and variety. This is particularly significant given that structured data is increasingly being used in Big Data applications. We exclude from our definition data characteristics which would form part of a Data Quality Dimension and this recognises the distinction between the essential characteristics of Big Data and characteristics which are targets or goals or markers for data quality. Future work, building on the understanding of Big Data discussed in this paper, is to develop Data Quality Dimensions for Big Data to support our work on Big Data quality.

6. REFERENCES

[1] Wand, Y & Wang R.Y. (1996) Anchoring Data Quality Dimensions in Ontological Foundations *Communications of the ACM* 39, 86-95 http://doi.org/10.1145/240455.240479

[2] Gupta, P., Tyagi, N., 2015. An approach towards big data; A review, *2015 International Conference on Computing, Communication Automation (ICCCA)*. Presented at the 2015 International Conference on Computing, Communication Automation (ICCCA), 118–123. http://doi.org/10.1109/CCAA.2015.7148356

[3] Suresh, J. (2014) Bird's Eye View on Big Data Management *2014 Conference on IT in Business, Industry and Government (CSIBIG)* 1-5 http://doi.org/ 10.1109/CSIBIG.2014.7056930

[4] Khan,M.,Uddin, M. & Gupta N. (2014) Seven V's of Big Data; Understanding Big Data to extract value. *Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education "Engineering Education: Industry Involvement and Interdisciplinary Trends" ASEE Zone 1 2014* http://doi.org/10.1109/ASEEZone1.2014.6820689

[5] Bedi, P., Jindal, V., & Gautam, A. (2014) Beginning with big data simplified. *2014 International Conference on Data Mining and Intelligence Computing (ICDMIC) 1-7*, http://doi.org/10.1109/ICDMIC.2014.6954229

[6] Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). Addressing big data issues in Scientific Data Infrastructure. In *2013 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 48–55). http://doi.org/10.1109/CTS.2013.6567203

[7] Demchenko, Y., Gruengard, E. & Klous, S., 2014. Instructional Model for Building Effective Big Data Curricula for Online and Campus Education. *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, pp.935–941. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7037787

[8] Marr, B., 2014. Big Data: The 5 Vs Everyone Must Know. *LinkedIn Pulse*. https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know

[9] Press, G., 2014. 12 Big Data Definitions: What's Yours? *Forbes*. http://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#611ecbcc21a9

[10] Hu, H., Wen, Y., Chua, T.-S., & Li, X. (2014). Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access*, *2*, 652–687. http://doi.org/10.1109/ACCESS.2014.2332453

[11] Durham,E.E., Rosen, A. & Harrison R.W. (2014) . A model architecture for Big Data applications using relational databases *2014 International Conference on Big Data* 9-16. http://doi.org/10.1109/BigData.2014.700

[12] Navathe, S.B., (1992). Evolution of Data Modeling for Databases. *Communications ACM 35, 112–123.* doi:10.1145/130994.131001

[13] Codd, E.F., 1970. A Relational Model of Data for Large Shared Data Banks. *Communications. ACM*, 13(6), 377–387

[14] Angles, R., & Gutierrez, C. (2008). Survey of Graph Database Models. *ACM Comput. Surv.*, *40*(1), 1:1–1:39. http://doi.org/10.1145/1322432.1322433

[15] Gartner Research http://www.gartner.com/it-glossary/big-data/

[16] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., (2011) Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute* http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

[17] Jacobs, A. (2009). The Pathologies of Big Data. *Queue*, *7*(6), 10. http://doi.org/10.1145/1563821.1563874Jacobs (2009)

[18] Gantz, B. J., & Reinsel, D. (2011). Extracting Value from Chaos State of the Universe : An Executive Summary. *IDC iView*, (June), 1–12. Retrieved from http://idcdocserv.com/1142Gantz & Reinsel, 2011

[19] Chen,M., Mao,S. & Liu, Y.(2014) Big Data: A survey *Mobile Networks and Applications*,19,171-209. http://doi.org/10.1007/s11036-013-0489-0

[20] Gandomi, A. & Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), pp.137–144. Available at: http://www.sciencedirect.com/science/article/pii/S0268401214001066.

[21] IBM Big Data & Analytics Hub http://www.ibmbigdatahub

[22] Saha, B., & Srivastava, D. (2014). Data quality: The other face of Big Data. *Proceedings - International Conference on Data Engineering*, 1294–1297. http://doi.org/10.1109/ICDE.2014.6816764

[23] Datafloq https://datafloq.com/read/3vs-sufficient-describe-big-data/166

[24] Sagiroglu, S. & Sinanc, D., 2013. Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*. pp. 42–47

[25] Xhafa, F., Naranjo, V., Barolli, L., & Takizawa, M. (2015). On Streaming Consistency of Big Data Stream Processing in Heterogenous Clutsers. *2015 18th International Conference on Network-Based Information Systems*, 476–482. http://doi.org/10.1109/NBiS.2015.122

[26] Batini, C., & Scannapieco, M. (2006). Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications). Secaucus, NJ, USA: Springer-Verlag New York, Inc.

[27] Juddoo, S. (2015). Overview of data quality challenges in the context of Big Data. In *2015 International Conference on Computing, Communication and Security (ICCCS)* (pp. 1–9). http://doi.org/10.1109/CCCS.2015.7374131

[28] Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*(0), 2. http://doi.org/10.5334/dsj-2015-002

[29] Strong, D.M., Lee, Y.W., Wang, R.Y., (1997). Data Quality in Context. *Communications ACM* 40, 103–110. doi:10.1145/253769.253804

[30] Cooper,M., Mell, P. (2012) Tackling Big Data *NIST Computer Security Resource Centre* (fcsm_june2012_cooper_mell).

[31] NIST Big Data Public Working Group, & Subgroup, T. (2015). NIST Special Publication XXX-XXX DRAFT NIST Big Data Interoperability Framework : Volume 1 , Definitions DRAFT NIST Big Data Interoperability Framework : Volume 1 , Definitions, *1*.

[32] Cetintemel, U. et al., 2014. S-Store: a streaming NewSQL system for big velocity applications. *Proceedings of the VLDB Endowment*, 7(13), pp.1633–1636. Available at: http://dl.acm.org/citation.cfm?doid=2733004.2733048.

[33] Moreno, A., & Redondo, T. (2016). Text Analytics: the convergence of Big Data and Artificial Intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence*, *3*(6), 57. http://doi.org/10.9781/ijimai.2016.369

[34] Weets, J., Kakhani, M. K., & Kumar, A. (2015). Limitations and Challenges of HDFS and MapReduce. In *2015 International Conference on Green Computing and Internet of Things (ICGCIoT), 2015*. http://doi.org/10.1109/ICGCIoT.2015.7380524

[35] Sakr, S., Liu, A., & Fayoumi, A. G. (2013). The Family of MapReduce and Large-Scale Data Processing Systems. *ACM Computing Surveys*, *46*(1), 1–44. http://doi.org/10.1145/2522968.2522979