# Big Data Redux: New Issues and Challenges Moving Forward

J. Alberto Espinosa
American University
alberto@american.edu

Stephen Kaisler
SHK & Associates
skaisler1@comcast.net

Frank Armour
American University
fjarmour@gmail.com

William H. Money
The Citadel
wmoney@citadel.edu

## Abstract

*As of the time of this writing, our HICSS-46 proceedings article has enjoyed over 520 Google Scholar citations. We have published several HICSS proceedings, articles and a book on this subject, but none of them have generated this level of interest. In an effort to update our findings six years later, and to understand what is driving this interest, we have downloaded the first 500 citations to our article and the corresponding citing article, when available. We conducted an in-depth literature review of the articles published in top journals and leading conference proceedings, along with articles with a high volume of citations. This paper provides a brief summary of the key concepts in our original paper and reports on the key aspects of interest we found in our review, and also updates our original paper with new directions for future practice and research in big data and analytics.*

## 1. Introduction

As of the time of this writing, our HICSS article titled *"Big Data: New Issues and Challenges Moving Forward"* [1] has enjoyed over 560 Google Scholar citations (see: https://tinyurl.com/yb9qd8bl). We argued that *"big data is the new business and social science frontier"* and that the amount of information and knowledge that can be extracted from the digital universe continued to expand rapidly. We also concluded that *"big data is just the beginning of the problem. Technology evolution and placement guarantee that in a few years more data will be available in a year than has been collected since the dawn of man."* This is proven to be true, and the new challenge we face is that big data is becoming ***"really big data"*** as larger volumes of more complex and dynamic keep accumulating in gigantic databases.

Our original paper discussed important issues and challenges that would require concerted research efforts. We concluded these major challenges had to be addressed within the next decade. Six years later, we concurred that it was important to evaluate where the big data field has evolved and where it is going. The high citation level of our original HICSS-46 article offers us the opportunity to do that. In order to understand the specific issues that have generated the most interest in our original article, we conducted a targeted literature review on its first 500 citations. We skimmed all articles available.Our goals for the present study are to: (1) better understand the big data issues that matter the most to practitioners and scholars; (2) identify new issues that authors have uncovered; and (3) update our findings to provide new insights into the future direction of research in big data and analytics. In this paper we provide a summary of the key concepts and issues reported in our original paper, followed by a summary of our review findings for each. We then discuss the new issues reported in the citing articles, and present our conclusions.

## 2. Method

For this study, we selected: all articles published in premier journals and leading conference proceedings; all articles with 50 or more Google Scholar citations; and all HICSS articles. This selection process yielded 107 articles. We read and analyzed these paper, but noted that 53 of them did not contribute anything new, but just repeated content from other articles. Thus, we used all 107 articles for topic counts, but only the 54 remaining articles for in depth analysis.

## 3. Issues and Challenges Revised

### 3.1 Definitions of Big Data.

Our prior HICSS article reported that the early conceptualization of big data was large volumes of data just beyond technology's capability to store, manage and process efficiently by traditional database methods and tools. We concluded that this definition was not useful, making the definition of big data a *moving target* because data technologies improve dramatically every year. We favored a definition based on the 3V's below.

*Review:* Many authors have provided big data definitions, but the majority agree with our perspective that a definition based on the 3V's provide a better grasp of what big data is about.

### 2.2 The 3V's of Big Data

We reported that most big data researchers had adopted Doug Laney's classic perspective from 2001 [2], that the main big data challenges resided along key dimensions, commonly referred as the ***3V's of big data***

URI: https://hdl.handle.net/10125/59546
ISBN: 978-0-9981331-2-6

– ***Volume, Velocity*** and ***Variety***. The intuitive appeal of these 3V's led to some consensus for a definition of big data as *"high volume, high velocity, and/or high variety information assets"* [3].

*Review:* The 3 V's continue to be dominant. The simplicity and appealing concept of the original 3 V's has generated a gold rush (by us and others) to discover and propose new V's (e.g., value, variability, veracity, validity, etc.). However, the original 3V's seem to be the most stable and fundamental dimensions of big data that really matter to most. Of the new V's proposed, *"veracity"* is gaining more acceptance because of the big data cleansing challenges. We argue that "*validity*" is captured in veracity, *"variability"* is already captured in velocity and variety, and "*value*" is dependent on the user, not on the informational content. One author proposed that it is better to think of big data along a continuum, from small data, to bigger data, to big data depending on the magnitude of each of the 3 V's [4].

## 2.3 Main Challenges of Big Data

Table 1 and Figure 1 show, the issues and challenges that have received most attention include: growth rate; tools and talent; management and analytic challenges. The issues and challenges that received very little or no attention include: time to information; data ownership; needle in the haystack; turning straw into gold and predicting the world. Also, a large proportion of the articles raised new issues and challenges discussed below. Each challenge includes the number of papers that discussed (in parenthesis) the issue in depth, followed by a summary from our HICSS 2013 paper and a review of the new literature on each issue.

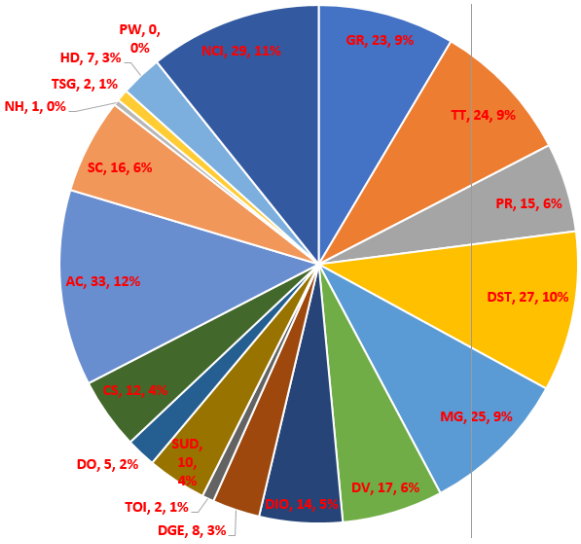


**Figure 1: Issue Popularity**

| **Topic** (rank ordered) | **Count** |
|---|---|
| Analytics Challenges | 33 (35.5%) |
| Data Storage and Transport | 27 (29.0%) |
| Management | 25 (26.9%) |
| Tools and Talent | 24 (25.8%) |
| Growth Rate | 23 (24.7%) |
| Data Validation | 17 (18.3%) |
| Privacy | 15 (16.1%) |
| Data Input and Output Process | 14 (15.1%) |
| Compliance and Security | 12 (12.9%) |
| Structured and Unstructured Data | 10 (10.8%) |
| Data Growth and Expansion | 8 (8.6%) |
| Hybrid Data | 7 (7.5%) |
| Data Ownership | 5 (5.4%) |
| Time to Information | 2 (2.2%) |
| Turning Straw into Gold | 2 (2.2%) |
| Needle in the Haystack | 1 (1.1%) |
| Predicting the World | 0 (0%) |

**Table 1: Number of Papers Discussing Old Topics**

***Growth Rate (GR, 23).*** The staggering rates of increase in the amount and complexity of data collected up to 2013 pale in comparison to today's growth rate. We reported that this overwhelming growth rate was fast exceeding our ability to both design appropriate systems to handle the data effectively and analyze it to extract relevant meaning for decision-making.

*Review:* Most papers we reviewed discussed the staggering growth of data. There has been much improvement in big data and analytical tools, technologies and methods, and this growth continues today, with many challenges, such as scarcity of big data talent , analytics talent and slow growth in IT budgets limiting available resources and computational power. This growth hinders the ability to design suitable cloud computing platforms [5] and restricts our capability to store, manage, transport and analyze data. This growth comes from things like sensor data, financial transactions, social media, digital content and GPS data, among other things [6]. An author characterized this growth, not as a *"tidal wave"*, but as an overwhelming ***"tsunami"*** [7]. For example, online social networks contain massive amounts of rapidly expanding data [8].

The greatest challenges are not just the staggering amounts of data generated, but also about the opacity, noise, and the relational nature of the data, making it very difficult to move it, analyze it and protect it from corruption or loss [9]. Volume is about the amount of data *"available"* to an organization, not just what it owns . On the technical side, the challenges lie in the features making big data difficult to handle and to extract from different systems and platforms [10]. Many authors concurred that big data grows at exceptional rates. Some authors cited the zettabyte level as the next threshold to be achieved by the early 2020s. However, this number is misleading since it does not

account for duplications such as those generated by blockchain, massive retweets and the innumerable copies of images and videos. We propose differentiating big data volume into original data and duplicated data to grapple with possible discrepancies in data volume that may lead to different results during processing.

***Tools and Talent (TT, 24).*** We noted in our article that the U.S. Government's big data initiatives assumed that users would be more successful, but had concerns about lack of tools and trained personnel to properly work with big data. Despite the proliferation of big data and analytics educational programs, this continues to be a concern.

*Review:* The shortages predicted in the popular McKinsey Global Institute report of 190K data scientists and 1.5M analytical managers by 2018 [11] proved to be true or even larger [12]. Several authors reiterated this challenge [13-15]. The skill set of the labor force today is improving, but there continues to be skilled labor shortages [14, 16]. Tools and talent that require changes in training and improvements in capabilities were recognized as important by over half of the papers, leading in some cases to reluctance to adopt a big data solution [17]. Another study [18] recognized the large physical challenges for a *"highly secured critical infrastructure systems, a scalable, reliable and robust threat monitoring and detection system."* They noted that *"big data from threat monitoring systems pose serious challenges for cyber operations because an ever growing number of devices in the system and the amount of complex monitoring data collected from critical infrastructure systems require scalable methods to capture, store, manage, and process the big data."* They proposed a *"cloud computing based network monitoring and threat detection system to make critical infrastructure systems secure."*

***Privacy (PR, 15).*** Another concern [19] was that personal data from social media, health and patient records, phone logs, and government records would create new invasive privacy incursions. These concerns are on target, given the litigations, sanctions and new regulations spreading over Europe and the U.S.

*Review:* Privacy continues to be one of the biggest challenges [5]. One study points to a social stratification problem, where savvy users with big data skills, including those in government and law enforcement, can take advantage of unskilled individuals [14]. Furthermore, new privacy issues are not being addressed due to the scarcity of educational opportunities and curricula on privacy [6]. Privacy issues associated with storing and processing big data and using the results continues to be an increasing problem as shown by recent storage breachess. Some [5] suggested encryption as a means of ensuring privacy, but this does not address privacy given the willingness of many users to share personal data online.

***Data Storage and Transport (DST, 27).*** This challenge poses serious limitations each time we see a burst in data growth. Technologies that generate massive amounts of data include, the Web, social media, Internet of Things, digital content streaming, Blockchain, etc. Data volumes of multiple terabytes can quickly overwhelm current data storage capacities, choke bandwidths, and impair data transfer rates of most networks. As data volumes surge to exabytes, current and anticipated data storage technology will not be able to handle these huge volumes. We noted two solutions attempting to address these issues, including processing the data *"in place"* and transmitting just the portions of the data needed for downstream analysis.

*Review:* This continues to be a recognized problem, even in the cloud environment [5]. It is primarily an issue of capacity and performance [14]. The size of big data ranging from terabytes to exabytes, exceeds the capacity of most processing systems. It takes longer to transmit big data over current networks from a collection or storage point to a processing point, than it takes to process it. For example, transferring an exabyte of data can take a few thousand hours over a fast network with a sustained data transfer rate [21]. Several authors support, *"bringing the code to the data"*, unlike the traditional method of *"bringing the data to the code"* [22].

One study [23] noted that the large physical challenges for a *"highly secured critical infrastructure systems, a scalable, reliable and robust threat monitoring and detection system."* The authors also noted that *"big data from threat monitoring systems pose serious challenges."* One paper addressed these challenges and discussed issues such as scalability, availability, data integrity, and governance as key issues [5]. Another study considered the problem in the context of electoral data from contested races in India [13]. Many authors realized that limitations in hardware and networks affect the ability of organizations to embrace big data.

One study [24] noted that big data vendors are not optimistic that there is a single solution for big data storage. It suggests that cloud storage can increase the fault tolerance of a system because it provides system administrators the ability to manage resources from a single point. Another study [25] proposed a dynamic architecture – the Storage Resource Broker - for big data storage and management overlaid on top of a "grid computing" solution. They assert that data storage has two aspects – spatial and temporal locality. Temporal location occurs when data is moved from several locations to a specific processing point. They suggest

that reshaping access patterns to data and subsequent movement may improve performance. Another study [26] noted that storing data in the cloud makes it subject to different unadjudicated legal constraints, protections and rules, depending on where the data resides.

***Management (MG, 25).*** Management challenges include data distribution, access schemes and governance. The key issues we raised included: How is the data distributed geographically? Who owns which part of it? Who has access to what? What are the security, access and legal issues? We noted that the amount of data collected, stored and analyzed far exceeded the capacity of humans to apply effective governance protocols to ensure the accuracy, validity, and ethical utilization of the data.

*Review:* One study [27] questioned if leverage from big data can actually create value for the organizations implementing it. The paper notes that no scientifically grounded and unbiased method exists to affect such an assessment and to guide implementation. It described a methodology based on IT value theory and workgroup ideation, guiding big data idea generation, idea assessment and implementation management. Another study [28] discussed how to improve the techniques available to analyze data and reports on the results of an experiment comparing four different methodologies showing significant differences based on the methodology used.

Other management challenges discussed [29] are that managers who track business intelligence and analytics costs believe they have already established a higher degree of big data implementations. Another paper [30] discussed big data-driven innovations as key in improving healthcare system sustainability through public-private partnerships (PPPs).

Some authors discussed other management and governance issues. One study [31] focused on the management of data collection by urban sensors. There were significant challenges in using sensors including fusion of data, formats, and ranges with respect to values. Another study [32] noted that there *"are no universal methods, which would guide any organization or individual on how to manage and interpret the meaning of information available on Twitter."* We believe this to be true of most other social media as well because of the diversity of presentation of ideas in different forms. One study [33] is noteworthy for its tabular presentations of the issues and factors affecting data governance.

***Data Validation (DV, 17).*** We noted that data validation often focused on missing data and outliers. But, we also commented that data is often very fine-grained and, given the large volumes involved, it would be impossible or impractical to validate every datum. More systematic approaches to data cleansing and validation are necessary.

*Review:* Data quality from data arriving fast from a variety of sources and in a variety of formats continues to be a great challenge for validation [5]. Several studies [5, 18] discussed the big data validation challenges. One main challenge is determining if the data to be analyzed have no issues, such as mismatching or missing values, needing cleansing, or determining if the data meets quality needs [28]. We noted that the acceptable level of data quality depends on the purpose of that analysis. For example, trend analysis may not require the precision of other predictions [34]. One study [35] discussed big data sharing and Internet of Things (IoT) caching data throughout the network and argued that when distributed this poses a challenge on flexible authorization and identity verification. It proposed exploiting a process for authorization and "identity-based signature" (IBS) for the distributed verification of the identities.

***Data Input and Output Processes (DIO, 14).*** Big challenges reside in updating, searching and retrieving relevant data efficiently for analysis in the Petabyte and Exabyte range, in a variety of formats [36].

*Review:* With our increased reliance on big data, the need for fault tolerant systems increases beyond what has been necessary for traditional systems [14]. A study [5] discussed cloud computing advantages as promising for data input and output processes. Another study [37] also noted that big data presents challenges for digital earth to store, transport, process, mine and serve the data; and that cloud computing provided fundamental support to address the challenges with required shared computing resources.

Some authors identified streaming as an emerging problem in big data, with input streaming being the focus of most discussions. One study [24] reiterated that organizations need to develop criteria to assess the relevance and criticality of streaming data when they cannot afford to store it all. Another paper [38] proposed a model for designing and developing energy-efficient networks that may alleviate some of the bottlenecks in organizations that receive big data in near real-time. One paper [39] noted that *"rapid increases in the volume and velocity of data streaming pose technical challenges for capturing, storing, and analyzing massive amount of data"*.

***Data Growth vs. Expansion (DGE, 8).*** Many organizations do not consider data expansion, which occurs when the data records grow in richness and variety, as they evolve over time with additional new types of information, processes and demands.

*Review:* One study [40] addressed two challenges in integrating heterogeneous data: (1) horizontal evolution that concerned the number of data sources

which are always increasing exponentially, and (2) vertical evolution that concerned the quantity of data in each data source, which was also growing rapidly. However, another paper [41] argued that "the concept of big is problematic to pinpoint, not least because a dataset that appears to be massive today will almost surely appear small in the near future".

***Time-to-Information (TOI, 2).*** This challenge is about decision making with real-time devices and processes that generate near-continuous data, such as radio frequency identifiers (RFIDs), GPS data for location based services, and Internet-of-Things (IoT) devices. Rather than analyzing all available data, which will drive the processing system architecture and the characteristics of computational engines and algorithms, we must consider the value contributed by relevant data and the time needed for effective dissemination.

*Review:* Surprisingly, none of the articles we reviewed directly discussed this issue.

***Structured vs. Unstructured Data (SUD, 10).*** As we reported, translation and interpretation of structured data is relatively straightforward through data dictionaries, schemas and relational database technologies. Also, analytics tools and open source computational software like R and Python are readily available to analyze structured data. The main challenges were with unstructured data (e.g., free text, graphics, multi-media, etc.), which add complexity and demand for end-to-end processing performance. The emergence of non-relational, distributed, data management platforms like NoSQL database systems, such as MongoDB and Cassandra, are showing some promise to store, manage and analyze unstructured data up to the Petabyte range.

*Review:* One study [42] examines the basic model for the extraction of structured data from unstructured information sources, the basic steps involved, and analyzed the methods used in the extraction process.

***Data Ownership (DO, 5).*** We identified this is as a critical and ongoing challenge, particularly in the social media and similar public arenas. Individuals contribute data continually and freely, while the data is stored by entities that profit from mining it. While petabytes of social media data reside on the servers of providers like Facebook and Twitter, it is not so clear who owns which data when it's aggregated and used for public reporting and marketing. One study [43] addressed this issue with respect to cloud computing. Furthermore, unverified data can also lead to erroneous linkages among entities and unethical use of publicly available data.

*Review:* Few papers discussed data ownership. One article [30] discussed interesting challenges and best practices in managing big data-driven healthcare innovations in the Netherlands (see MG above). Data ownership continued to be a recognized challenge because it brings about policy, security, privacy and technical issues, particularly when disparate databases are used [41]. Data sensitive to one organization may not be sensitive to another, which can lead to conflict in decision-making. Data ownership also has cost and operational expenditure issues when the data is stored and processed in different locations. One study [26] addressed some of the legal issues related to storing data in the cloud and concluded that the legal concept of property and data ownership is problematic. This is particularly true if the service provider can move data from one repository to another without the user's knowledge or consent as long as the service provider continues to give access to the data. The legal community has not given sufficient attention to cloud computing and its impact on property ownership. But the authors believe this will result in significant litigation and difficulties with groups such as the European Union.

***Compliance and Security (CS, 12).*** We reported that, as more data is collected, there was the fear that certain organizations would know too much about individuals. One prominent example is with the storage of individual health records. A major threat to personal security is the unregulated accumulation of data by numerous social media companies. International Data Corporation (IDC) coined the term "digital shadow" to reflect the amount of data concerning an individual that is collected, organized, and perhaps analyzed to form an aggregate "picture" of the individual.

*Review:* There are still many unresolved CS issues related to big data, which are magnified in cloud environments [5]. One study [35] discussed big data sharing and "Internet-of-Things" (IoT) data caching throughout the network and argue that this poses a challenge on flexible authorization and identity verification and noted that *"Ciphertext-Policy Attribute-Based Encryption (CP-ABE) has been identified as a promising approach."* They propose *"Verifiable and Flexible Data Sharing"* (VFDS) authorization and *"Identity Based Signature"* (IBS) for distributed verification of identities.

***Analytics Challenges (AC, 33).*** In our original paper we discussed how Big Data Analytics (BDA) is a major challenge. Analytics comes in different flavors, including descriptive, predictive, prescriptive, decision modeling and optimization. It also comes in various types and forms, including quantitative, classification, visual, etc. There are also specialized domains of analytics with their own idiosyncrasies, including: graph theory, social network analysis, behavioral analytics, econometric modeling, game theory, control theory, evolutionary computation and simulation models, among many others [44].

*Review:* Analytics is by far the aspect of big data with the most challenges reported in the literature. Existing analytics algorithms are very limited when it comes to big data analytics. There is a clear need to develop more scalable algorithms with increasing computational resources [5]. Take for example "social big data" where a combinatorial explosion occurs once we incorporate the large number of possible links and relationships among actors, requiring high computational complexity and muscle for analysis. Another example is text mining where text data tends to be sparse, fuzzy and multi-dimensional [6]. One study raised the issue of outliers with BDA and proposed an outlier-robust method for extreme machine learning [45]. Another author argued that spurious correlations in BDA can lead to very wrong conclusions [4].

Both, the academic and industrial communities are generating and collecting data at unprecedented rates, resulting in a great opportunity to advance domain knowledge decision quality [46]. But in order to tap into this potential we need more secure critical infrastructures, more scalable, reliable and robust systems [23] and more effective diagnostics for error prevention [47]. A few studies [5, 14, 42, 48] have examined the basic models for the extraction of structured data from unstructured information sources and articulated the related challenges with technologies like Hadoop and MapReduce. Another study [49] has also discussed efficient platforms for BDA and identified related performance issues. One more study [50] also proposed a taxonomy and analysis of indexing techniques to organize access to big data under varying conditions. Finally, in a follow up paper [51] the authors provided a taxonomy of analytic classes and described issues and challenges associated with BDA.

***Scalability (SC, 16).*** In our original study we reported that a critical issue is whether or not an analytic process scales seamlessly as a dataset increases by orders of magnitude. An open research question we asked was whether there is a fundamental limit to the scalability of algorithms.

*Review:* Cloud computing is gaining substantial strength coping with scalability issues [14]. One study [52] discussed the challenge of data arriving at a much faster rate than computational speeds can handle. This article noted that scalable machine learning algorithms already exist. But challenges remain, since data accuracy tends to degrade with large-scale longitudinal data sets [53]. Scalability has different connotations in a big data environment. One aspect relates to the processing of larger data, but another one relates to improving computational performance as the complexity of the data increases.

***Needle in the Haystack (NH, 1).*** We discussed that this is a problem in which the right answer is very difficult to determine in advance, but very easy to verify once you know where the needle is [54]. The challenge with this type of problem was that the likelihood of being wrong is quite high.

*Review:* Interestingly, none of the articles we reviewed discussed this issue.

***Turning Straw into Gold (TSG, 2).*** We described the processing of a large set of discrete data points (e.g., a Facebook network) into a more valuable but smaller subset [55]. As the number of network ties emanating from "central" sub-group members increases, the overall complexity of the sub-group network increases exponentially. A related challenge we noted was the time-varying nature of networks, which requires more intensive computations and different algorithms for near-real-time analytics.

*Review:* None of the articles we reviewed discussed this issue.

***Hybrid Data (HD, 7).*** We discussed that with large heterogeneous datasets, a major challenge was to figure out what parts of the data to analyze and how to analyze it. Some [56] argued, it was not always clear that a particular group of analytic methods could provide appropriate answers. Others [57] noted a related problem with unpredicted effects arising from the presences of outliers in the data, because the past can no longer predict the future accurately.

*Review:* Please refer to our review in 2.2 above.

***Predicting the World (PW, 0).*** The challenge here was whether we could predict complex, macro-level events (e.g., global weather; political or economic world events; epidemics). Underlying this challenge were questions of modeling complex systems present in various fields, like natural sciences, large-scale social and cultural interactions, global trade, natural disasters, and world poverty and hunger, among other things. These are all "wicked problems" [58], which are those with incomplete, contradictory and changing requirements [59]. Because of the complex interdependencies of their elements, it is often difficult to recognize that one has achieved even a partial solution.

*Review:* None of the articles we reviewed discussed this issue.

## 4. New Challenges and Issues

***General Issues (NCI).*** Several papers have articulated new BDA challenges and issues. One study [60] discussed big data's evolution from numbers and characters created and collected by individuals, to unstructured data types collected by devices and proposed a new concept – *"big data as objects"* with some degree of intelligence or "smartness." Other studies also discussed: big data sharing and flexible

authorization and identity verification [35]; the strategic impact factors of big data [61]; and degradation of data accuracy in large-scale longitudinal data sets [53]. We now discuss more specific new issues and challenges receiving attention. The list of new challenges and issues discussed in the recent research literature are shown in table 2.

| **Topic** (rank ordered) | **Count** |
|---|---|
| New Challenges and Issues | 29 (31.2%) |
| Functional Domains | 12 (12.9%) |
| Business Value | 5 (5.4%) |
| Data Staging | 3 (3.2%) |
| Cloud Computing | 2 (2.2%) |
| Collaborations | 1 (1.1%) |
| Digital Divide | 1 (1.1%) |

**Table 2: Number of Papers Discussing New Topics**

***Cloud Computing (CC).*** One study [5] concluded from five industry and five academic case studies that cloud computing presented substantial advantages for big data operations due to benefits like scalability, availability, data integrity, transformation (into suitable forms for analysis), data quality, heterogeneity, privacy, legal/regulatory issues and governance. Some cloud popularity has been driven by mobile devices, providing remote access to large data storage, processing and computational capabilities to otherwise limited devices.

***Data Staging (DS).*** This continues to be a big challenge stemming from the heterogeneous nature of data these days [5]. A growing problem with big data getting bigger is data cleansing [62].

***Business Value (BV).*** While several authors have proposed "value" as one of the big data V's, there is no agreed upon conceptualization of what this means nor how to study it empirically. One suggested approach is to use dynamic capabilities theories to study the impact of BDA from a business value chain perspective, rather than a from data or system perspective [63]. Their empirical study finds evidence that BDA applications enable internal and external knowledge management to help create organizational agility. A larger question about business value is how to evaluate BDA systems [62]. Another study of seven cases concluded that many companies are still not extracting sufficient value from BDA [64] and another empirical study noted that it takes a very long time to derive business value from BDA [16]. This is partly due to the fact that most people still don't understand BDA or how it can yield value to business. One study proposed the concept of the "big data value chain (BDVC)" as a way to understand the business value of big data, with sequential BDVC activities – data generation, collection, transmission, pre-processing, storage and analysis, and their respective sub-activities [65].

***Functional Domains (FD).*** BDA is permeating just about every discipline and functional domain out there and becoming quite ubiquitous. Some of the functional domains applying BDA reported in the studies we reviewed included: social media [6]; marketing [66]; value chain [63]; finance and materials management [45]; healthcare, telecommunications, retail, law enforcement, insurance and education [65]; energy smart data grids [7]; and the public sector, where there is an abundance of data and great needs for BDA skills and not enough talent supply [4]. The authors reported various areas of public sector domains needing BDA, such as energy, decision making, sentiment analysis, etc.

Specific areas of specialization discussed, which can benefit from BDA included: knowledge management and social commerce [67]; decision making [68]; big social data (BSD) as a combination of social computing, computational social science, big data science and data analytics [69]; online social networks, which pose unique data and analytical challenges associated with things like graph analysis, social network analysis, collaborative filtering, trend detection, and sentiment analysis [8]. These authors also pointed to several limitations of machine learning and other analytical methods with the complexities of online social network data.

***Collaboration in BDA (COL).*** One study raised the importance of understanding how to collaborate across data within organizations, particularly with public organizations, which tend to operate in silos [4].

***Digital Divide (DD).*** One study also pointed out this challenge relating to the unequal access to data resources and uneven analytical skills in the market, which provide advantages to those with BDA skill and access to data [4].

## 5. Conclusions – What's the Same and What has Changed?

In sum, it was true in 2013 that there was no perfect big data management solution at that time. This remains true today. Many advances have been possible in the last 6 years in terms of big data storage and retrieval, as well as with the corresponding conventional analytics methods and tools. But once we move beyond the conventional into very large scale, complex, dynamic data, the present technologies, methods and tools fall short. Our review indicates that the major areas of research interest fall into several categories, leading us to the following concluding remarks.

***Old vs. New:*** By far, the old topics we raised in 2013 continue to dominate in the research literature, with the most attention given to analytics challenges, data storage and transport, management issues, tools

and talent, data validation and privacy. Several papers discussed new challenges, but they were very general for the most part. The most discussed new challenge was functional domains, highlighting the pervasiveness nature of BDA and the importance to understand and use BDA in the context of specific functional domains.

***Really Big Data:*** As big data keeps growing into ***"really big data"*** the issues and challenges will continue to grow exponentially. The big promise of benefits from BDA have generated a rush to generate and store *"data gold"* to be mined for prosperity. But yesterday's solutions are always tomorrow's problems. This growth continues to create substantial challenges, particularly in terms of: data storage and transport, analytic complexity, privacy, security, governance, compliance, and data ownership.

***System and Software Architecture:*** Single computer systems will not be able to store, manage, and analyze data in the Exabyte range in the near future. Cloud computing is constrained by a few parallelism models emphasizing homogeneous computational structures. System architectures able to handle hundreds of petabytes and beyond will require enhanced or new network topologies connecting multiple computing systems, perhaps with heterogeneous processors, to achieve the computational power necessary to process and analyze these volumes of data. New software architectures and programming languages driven by domain knowledge will be needed to construct high performance systems that can yield effective time-to-information results to support near-real time decision-making. These architectures will need to focus on end-to-end global performance rather than individual system performance.

***Analytic Tools and Methods****:* The plethora of analytic tools that emerged over the past decade has consumed the low-hanging fruit. Open source technologies and software languages like Hadoop, Python and R, along with the thousands of analytical packages for predictive modeling and machine learning, are bringing high analytical power to just about everyone's desktop. However, these sophisticated tools and technologies don't resolve the analytical challenges we have noted in growing BDA environments. More attention is needed to address the analytical challenges we have identified and find better ways to resolve complex problems that require new approaches to process larger and more complex data.

***Predictive Analytics Evolves****:* For predictive analytics to be useful, analytic tools must evolve beyond statistical and probabilistic models using numerical data to encompass unstructured non-numerical data.

***Prescriptive Analytics Emerge****:* Prescriptive analytics recommend courses of action that must be derived from the outcomes of predictive analytics, which must be carefully weighed against behavioral changes and possible disruptions to normal business operations. Little has been written about the analytical challenges of developing sound prescriptive models that can leverage the potential of BDA.

## 5. References


[1] S. H. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big Data: Issues and Challenges Moving Forward," in *46th. Hawaii International Conference on System Sciences*, Maui, Hawaii, 2013.

[2] D. Laney, "3D Data Management: Controlling Data Volume, Velocity, and Variety," Stamford, CTFebruary 6 2001.

[3] M. A. Beyer and D. Laney, "The importance of 'big data': a definition," *Stamford, CT: Gartner*, pp. 2014-2018, 2012.

[4] K. C. Desouza and B. Jacob, "Big data in the public sector: Lessons for practitioners and scholars," *Administration & Society*, vol. 49, pp. 1043-1064, 2017.

[5] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, "The rise of "big data" on cloud computing: Review and open research issues," *Information Systems*, vol. 47, pp. 98-115, 2015.

[6] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Information Fusion*, vol. 28, pp. 45-59, 2016.

[7] J. Corbett and J. Webster, "Organizational Sensemaking and Big Data Frames: Opportunity, Control, and Data Limitation," in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, 2015, pp. 4772-4781.

[8] A. Sapountzi and K. E. Psannis, "Social networking data analysis tools & challenges," *Future Generation Computer Systems*, 2016.

[9] H. Ekbia, M. Mattioli, I. Kouper, G. Arave, A. Ghazinejad, T. Bowman, *et al.*, "Big data, bigger dilemmas: A critical review," *Journal of the Association for Information Science and Technology*, vol. 66, pp. 1523-1545, 2015.

[10] H.-G. Fill and F. Johannsen, "A knowledge perspective on big data by joining enterprise modeling and data analyses," in *System Sciences (HICSS), 2016 49th Hawaii International Conference on*, 2016, pp. 4052-4061.

[11] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, *et al.*, "Big data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global InstituteMay 2011.

[12] N. Henke, J. Bughin, M. Chui, J. Manyika, T. Saleh, B. Wiseman, *et al.*, "The age of analytics: Competing in a data-driven world," *McKinsey Global Institute report*, 2016.

[13] G. Jagdev and A. Kaur, "Excavating Big Data associated to Indian Elections Scenario via Apache Hadoop," *International Journal of Advanced Research in Computer Science*, vol. 7, 2016.

[14] A. Katal, M. Wazid, and R. Goudar, "Big data: issues, challenges, tools and good practices," in *Contemporary Computing (IC3), 2013 Sixth International Conference on*, 2013, pp. 404-409.

[15] N. Khan, I. Yaqoob, I. A. T. Hashem, Z. Inayat, M. Ali, W. Kamaleldin, *et al.*, "Big data: survey, technologies, opportunities, and challenges," *The Scientific World Journal*, vol. 2014, 2014.

[16] P. Mikalef, V. Framnes, F. Danielsen, J. Krogstie, and D. H. Olsen, "Big data analytics capability: antecedents and business value," in *Proceedings of the 21st Pacific Asia conference on information systems (PACIS)*, 2017.

[17] S. N. Brohi, M. A. Bamiah, and M. N. Brohi, "IDENTIFYING AND ANALYZING THE TRANSIENT AND PERMANENT BARRIERS FOR BIG DATA," *Journal of Engineering Science and Technology*, vol. 11, pp. 1793-1807, 2016.

[18] G. Chen, Q. Cai, and Y. Zhan, "Approaches on personal data privacy preserving in cloud: a survey," in *Proceedings of The Third International Conference on Data Mining, Internet Computing, and Big Data, Konya*, Konya, Turkey, 2016, pp. 36-43.

[19] D. Boyd and K. Crawford, "Six Provocations for Big Data," Massachusetts2011.

[20] U. Marjit, K. Sharma, and P. Mandal, "Data Transfers in Hadoop: A Comparative Study," *Open Journal of Big Data (OJBD)*, vol. 1, pp. 34-46, 2015.

[21] N. Mathur and R. Purohit, "Issues and challenges in convergence of big data, cloud and data science," *International Journal of Computer Applications*, vol. 160, 2017.

[22] R. H. Bajaj and P. Ramteke, "Big data–the new era of data," *International Journal of Computer Science and Information Technologies*, vol. 5, pp. 1875-1885, 2014.

[23] Z. Chen, G. Xu, V. Mahalingam, L. Ge, J. Nguyen, W. Yu, *et al.*, "A cloud computing based network monitoring and threat detection system for critical infrastructures," *Big Data Research*, vol. 3, pp. 10-23, 2016.

[24] H. Nawaz and T. R. Soomro, "Private Cloud Storage in Big Data," *International Journal of Computer Applications*, vol. 142, 2016.

[25] K. Kumar, S. Kumar, and M. Shrivastava, "Novel dynamic and scalable storage management architecture," *International Journal of Computer Applications*, vol. 125, 2015.

[26] C. Bartolini, C. Santos, and C. Ullrich, "Property and the cloud," *Computer Law & Security Review*, 2017.

[27] M. Vanauer, C. Böhle, and B. Hellingrath, "Guiding the introduction of big data in organizations: A methodology with business-and data-driven ideation and enterprise architecture management-based implementation," in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, 2015, pp. 908-917.

[28] J. Saltz, I. Shamshurin, and C. Connors, "Predicting data science sociotechnical execution challenges by categorizing data science projects," *Journal of the Association for Information Science and Technology*, 2017.

[29] R. Grytz and A. Krohn-Grimberghe, "Service-Oriented Cost Allocation for Business Intelligence and Analytics: Helping Service Consumers to Increase Business Value," *International Journal of Systems and Service-Oriented Engineering (IJSSOE)*, vol. 7, pp. 40-57, 2017.

[30] E. Witjas-Paalberends, L. van Laarhoven, L. van de Burgwal, J. Feilzer, J. de Swart, E. Claassen, *et al.*, "Challenges and best practices for big data-driven healthcare innovations conducted by profit–non-profit partnerships–a quantitative prioritization," *International Journal of Healthcare Management*, pp. 1-11, 2017.

[31] S. E. Bibri, "The IoT for Smart Sustainable Cities of the Future: An Analytical Framework for Sensor-Based Big Data Applications for Environmental Sustainability," *Sustainable Cities and Society*, 2017.

[32] R. Juric, I. Kim, H. Panneerselvam, and I. Tesanovic, "Analysis of ZIKA Virus Tweets: Could Hadoop Platform Help in Global Health Management?," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[33] S. U. Lee, L. Zhu, and R. Jeffery, "Designing Data Governance in Platform Ecosystems," in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[34] J. S. Saltz, "The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness," in *Big Data (Big Data), 2015 IEEE International Conference on*, 2015, pp. 2066-2071.

[35] R. Li, H. Asaeda, J. Li, and X. Fu, "A Verifiable and Flexible Data Sharing mechanism for Information-Centric IoT," in *Communications (ICC), 2017 IEEE International Conference on*, 2017, pp. 1-7.

[36] A. Jacobs, "The pathologies of big data," *Communications of the ACM*, vol. 52, pp. 36-44, 2009.

[37] C. Yang, Q. Huang, Z. Li, K. Liu, and F. Hu, "Big Data and cloud computing: innovation opportunities and challenges," *International Journal of Digital Earth*, vol. 10, pp. 13-53, 2017.

[38] R. Li, H. Harai, and H. Asaeda, "An aggregatable name-based routing for energy-efficient data sharing in big data era," *IEEE Access*, vol. 3, pp. 955-966, 2015.

[39] N. Cho, M. Ku, X. Rui, and D. Min, "An Analysis of Deployment Models of HBase-based Hadoop Platform in Virtualized Computing Environment," in *Proceedings of the 2015 International Conference on Big Data Applications and Services*, 2015, pp. 156-160.

[40] H. Abbes and F. Gargouri, "MongoDB-Based Modular Ontology Building for Big Data Integration," *Journal on Data Semantics*, pp. 1-27, 2017.

[41] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," *Journal of Business Research*, vol. 70, pp. 263-286, 2017.

[42] A. Gupta, S. Anand Shankar, and C. Manjunath, "A Comparative Study on Data Extraction and Its Processes," *International Journal of Applied Engineering Research*, vol. 12, pp. 7194-7201, 2017.

[43] S. Kaisler, W. H. Money, and S. J. Cohen, "A decision framework for cloud computing," in *System Science (HICSS), 2012 45th Hawaii International Conference on*, 2012, pp. 1553-1562.

[44] S. Kaisler, "Advanced analytics," *CATALYST Technical Report, i_SW Corporation, Arlington, VA*, 2012.

[45] K. Zhang and M. Luo, "Outlier-robust extreme learning machine for regression problems," *Neurocomputing*, vol. 151, pp. 1519-1527, 2015.

[46] S. Bonner, A. S. McGough, I. Kureshi, J. Brennan, G. Theodoropoulos, L. Moss, *et al.*, "Data quality assessment and anomaly detection via map/reduce and linked data: a case study in the medical domain," in *Big Data (Big Data), 2015 IEEE International Conference on*, 2015, pp. 737-746.

[47] K. Gai, M. Qiu, L.-C. Chen, and M. Liu, "Electronic health record error prevention approach using ontology in big data," in *High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), 2015 IEEE 12th International Conferen on Embedded Software and Systems (ICESS), 2015 IEEE 17th International Conference on*, 2015, pp. 752-757.

[48] J. A. Morente-Molinera, R. Wikström, E. Herrera-Viedma, and C. Carlsson, "A linguistic mobile decision support system based on fuzzy ontology to facilitate knowledge mobilization," *Decision Support Systems*, vol. 81, pp. 66-75, 2016.

[49] S. Prabhu, A. P. Rodrigues, G. Prasad, and H. Nagesh, "Performance enhancement of Hadoop MapReduce framework for analyzing BigData," in *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2015.

[50] A. Gani, A. Siddiqa, S. Shamshirband, and F. Hanum, "A survey on indexing techniques for big data: taxonomy and performance evaluation," *Knowledge and information systems*, vol. 46, pp. 241-284, 2016.

[51] S. H. Kaisler, J. A. Espinosa, F. Armour, and W. H. Money, "Advanced Analytics--Issues and Challenges in a Global Environment," in *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, 2014, pp. 729-738.

[52] M. A. Wani and S. Jabin, "Big Data: Issues, Challenges, and Techniques in Business Intelligence," in *Big Data Analytics*, ed: Springer, 2018, pp. 613-628.

[53] M. S. Weber and H. Nguyen, "Big Data?: Big Issues Degradation in Longitudinal Data and Implications for Social Sciences," in *Proceedings of the ACM Web Science Conference*, 2015, p. 6.

[54] E. Felten, "Needle in a Haystack Problems," *Retrieved on November*, vol. 1, p. 2013, 2010.

[55] K. Freeman. (2011). *Data Visualization of Facebook Friendship Network*. Available: http://en.wikipedia.org/wiki/File:Kencf0618FacebookNetwork.jpg

[56] I. Ayres, *Super crunchers: Why thinking-by-numbers is the new way to be smart*: Bantam Books, 2007.

[57] N. N. Taleb, *The black swan: The impact of the highly improbable* vol. 2: Random house, 2007.

[58] T. Ritchey, "Wicked Problems Structuring Social Messes with Morphological Analysis." swemorph. com," ed: Online, 2005.

[59] H. W. Rittel and M. M. Webber, "Dilemmas in a general theory of planning," *Policy sciences*, vol. 4, pp. 155-169, 1973.

[60] S. Kaisler, W. Money, and S. Cohen, "Smart Objects: An Active Big Data Approach," in *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[61] R. Schmidt, M. Möhring, S. Maier, J. Pietsch, and R.-C. Härting, "Big data as strategic enabler-insights from central european enterprises," in *International Conference on Business Information Systems*, 2014, pp. 50-60.

[62] L. Rodríguez-Mazahua, C.-A. Rodríguez-Enríquez, J. L. Sánchez-Cervantes, J. Cervantes, J. L. García-Alcaraz, and G. Alor-Hernández, "A general perspective of Big Data: applications, tools, challenges and trends," *The Journal of Supercomputing*, vol. 72, pp. 3073-3113, 2016.

[63] N. Côrte-Real, T. Oliveira, and P. Ruivo, "Assessing business value of Big Data Analytics in European firms," *Journal of Business Research*, vol. 70, pp. 379-390, 2017.

[64] O. Al-Hujran, R. Wadi, R. Dahbour, M. Al-Doughmi, and M. M. Al-Debei, "Big Data: Opportunities and Challenges," in *The Fifth International Conference on Business Intelligence and Technology, ISBN*, 2015, pp. 978-1.

[65] A. K. Bhadani and D. Jothimani, "Effective Big Data Management and Opportunities for Implementation, IGI Global, Pennsylvania, USA," in *Big Data: Challenges, Opportunities, and Realities*, M. K. Singh and D. G. Kumar, Eds., ed: IGI Global, 2016, pp. 1-24.

[66] S. Fan, R. Y. Lau, and J. L. Zhao, "Demystifying big data analytics for business intelligence through the lens of marketing mix," *Big Data Research*, vol. 2, pp. 28-32, 2015.

[67] W. He, F.-K. Wang, and V. Akula, "Managing extracted knowledge from big social media data for business decision making," *Journal of Knowledge Management*, vol. 21, pp. 275-294, 2017.

[68] A. Intezari and S. Gressel, "Information and reformation in KM systems: big data and strategic decision-making," *Journal of Knowledge Management*, vol. 21, pp. 71-91, 2017.

[69] E. Olshannikova, T. Olsson, J. Huhtamäki, and H. Kärkkäinen, "Conceptualizing big social data," *Journal of Big Data*, vol. 4, p. 3, 2017.