RESEARCH ARTICLE OPEN ACCESS

Theoretical Understanding of Big Data Analytics Challenges and Future Scope

Raghu Ram Chowdary Velevela

Assistant Professor, Department of Information Technology,
Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru

ABSTRACT

Big Data Analytics (BDA) usage in the industry has been increased markedly in recent years. A huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. Big Data containing unique features that cannot be handled and processed using the conventional methods. This has presented a significant challenge to the industry. The basic objective of this paper is to explore the big data challenges. It further discusses the future scope in particular the future direction for Big Data research.

Keywords **-** Big Data, Big Data Characteristics, Big Data Issues, Big Data Challenges, Massive data, Structured Data, Unstructured Data.

I. INTRODUCTION

In 2010, enterprises and users stored more than 13 exabytes of new data; this is over 50,000 times the data in the Library of Congress. The potential value of global personal location data is estimated to be $700 billion to end users, and it can result in an up to 50% decrease in product development and assembly costs, according to a recent McKinsey report [McK2011]. McKinsey predicts an equally great effect of Big Data in employment, where 140,000-190,000 workers with "deep analytical" experience will be needed in the US; furthermore, 1.5 million managers will need to become data-literate. Not surprisingly, the recent PCAST report on Networking and IT R&D [PCAST2010] identified Big Data as a "research frontier" that can "accelerate progress across a broad range of priorities." Even popular news media now appreciates the value of Big Data as evidenced by coverage in the Economist [Eco2011], the New York Times [NYT2012], and National Public Radio [NPR2011a, NPR2011b].

While the potential benefits of Big Data are real and significant, and some initial successes have already been achieved (such as the Sloan Digital Sky Survey), there remain many technical challenges that must be addressed to fully realize this potential. The sheer size of the data, of course, is a major challenge, and is the one that is most easily recognized. However, there are others. Industry analysis companies like to point out that there are challenges not just in Volume, but also in Variety and Velocity [Gar2011], and that companies should not focus on just the first of these. By Variety, they usually mean heterogeneity of data types, representation, and semantic interpretation. By Velocity, they mean both the rate at which data arrive and the time in which it must be acted upon. While these three are important, this short list fails to include additional important requirements such as privacy and usability.

A study on the bibliometrics of Big Data reveals that there has been a phenomenal growth in the number of researches on Big Data. Based on an analysis on the trends of publication, up to 2011, less than 38 publications were conducted on 'Big Data'; however, by the year 2017, the number had grown rapidly to 3890 publications. Big Data's time trend is denoted in the fig. 1.

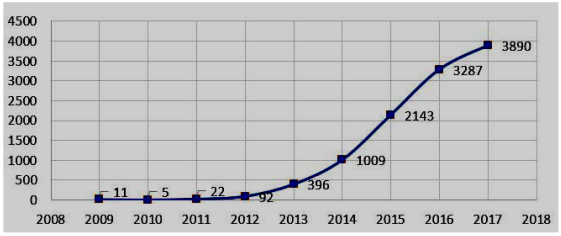


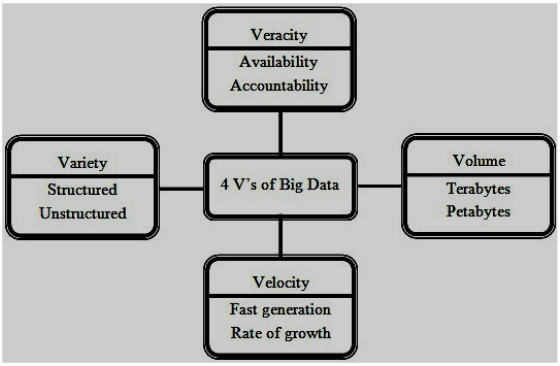
Fig. 1: Time Trend Of Big Data Research

II. CHARACTERISTICS OF BIG DATA

Recent years big data has been accumulated in several domains like health care, public administration, retail, biochemistry, and other interdisciplinary scientific researches. Web-based applications encounter big data frequently, such as social computing, internet text and documents, and internet search indexing. Social computing includes social network analysis, online communities, recommender systems, reputation systems, and prediction markets where as internet

search indexing includes ISI, IEEE Xplorer, Scopus. Thomson Reuters etc.

The major characteristics of Big Data are clearly described by Gartner, according to their definition, Big Data is a high-volume, high-velocity as well as high-variety data resource that requires new types of handling to empower improved decision makings, knowledge disclosure, and process maximization. The three Vs (volume, velocity, and variety) are the fundamental characteristics of Big Data. However, various organizations and institutes have come out with different definitions.



III. CHALLENGES IN BIG DATA

A. Data Storage and Analysis

In recent years the size of data has grown exponentially by various means such as mobile devices, aerial sensory technologies, remote sensing, radio frequency identification readers etc. These data are stored on spending much cost whereas they ignored or deleted finally because there is no enough space to store them. Therefore, the first challenge for big data analysis is storage mediums and higher input/output speed. In such cases, the data accessibility must be on the top priority for the knowledge discovery and representation. The prime reason is being that, it must be accessed easily and promptly for further analysis. In past decades, analyst use hard disk drives to store data but, it slower random input/output performance than sequential input/output. To overcome this limitation, the concept of solid state drive (SSD) and phrase change memory (PCM) was introduced. However the available storage technologies cannot possess the required performance for processing big data.

B. Data Heterogeneity:

Another significant challenge that researches are facing is to integrate data from various sources to optimize their value. A large amount of data that is generated by social media, blogospheres, and websites, each source is diverse in terms of format, semantics, and source of data. Data structure from these sources varies from very organized data (databases) to unstructured data (heterogeneous reports).

C. Scalability:

Scalability is another challenge in Big Data especially in the analysis stage. Incremental strategies have great scalability features in terms of analysis of big Data. Since the size of the data scales faster than the CPU speed. This challenge prompts the advancement of parallel computing. Real time applications such as social networks, navigations, finance, timeliness, web search, etc. necessitated computing requires parallel processing.

D. Preprocessing:

Preprocessing is an important part of Data Mining. This is mainly due to the fact that real-world databases are greatly affected by the existence of noise, missing values, conflicting and unnecessary data. Preprocessing is the set of strategies, selection of feature, defective data, imbalanced learning, etc. which is utilized before the utilization of the data mining strategy and it is one of the significant concerns within the infamous process of knowledge discovery from data.

E. Analytics:

Big Data carries with it large analytical challenges. Big Data analytics is the way toward inspecting Big Data. This will help with revealing hidden trends, obscure relationships and other valuable information that can be utilized for better decision making. A high level of technical skills is needed to carry out these types of analysis on vast amounts of data that are unstructured, semi-organized and organized. Moreover, it is yet unclear how an ideal architecture of analytics techniques ought to be able to manage historic information with real-time data simultaneously.

F. Information Security:

In 2015, a comprehensive review concluded that it is extremely difficult to store and examine Big Data with conventional applications. Additionally, there are problems with privacy and security issues. Encryption plans, firewalls, access permissions, transport layer security can be insecure; The provenance of data can be obscure,

even anonymous data can be re-distinguished. The research came to conclusion that privacy and security of Big Data are the issues that should be researched further. Despite the increasing number of R&D, recent analysis in 2018 demonstrated that privacy and security are still huge problems for Big Data and its execution.

IV. FUTURE SCOPE IN BIG DATA ANALYTICS

Big data analytics and data science are becoming the research focal point in industries and academia. Data science aims at researching big data and knowledge extraction from data. Applications of big data and data science include information science, uncertainty modelling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing and signal processing. Effective integration of technologies and analysis will result in predicting the future
drift of events. Main focus of this section is to discuss open research issues in big data analytics.

A. Iot For Big Data Analytics

Internet has restructured global interrelations, the art of businesses, cultural revolutions and an unbelievable number of personal characteristics. Currently, machines are getting in on the act to control innumerable autonomous gadgets via internet and create Internet of Things (IoT). Thus, appliances are becoming the user of the internet, just like humans with the web browsers. Internet of Things is attracting the attention of recent researchers for its most promising opportunities and challenges. It has an imperative economic and societal impact for the future construction of information, network and communication technology. The new regulation of future will be eventually, everything will be connected and intelligently controlled. The concept of IoT is becoming more pertinent to the realistic world due to the development of mobile devices, embedded and ubiquitous communication technologies, cloud computing, and data analytics. Moreover, IoT presents challenges in combinations of volume, velocity and variety. In a broader sense, just like the internet, Internet of Things enables the devices to exist in a myriad of places and facilitates applications ranging from trivial to the crucial. Conversely, it is still mystifying to understand IoT well, including definitions,
content and differences from other similar concepts. Several diversified technologies such as computational intelligence and big-data can be incorporated together to improve the data management and knowledge discovery of large scale automation applications. Much research in this direction has been carried out by Mishra, Lin and Chang. Knowledge acquisition from IoT data is the biggest challenge that big data professional are facing.

Therefore, it is essential to develop infrastructure to analyse the IoT data. An
IoT device generates continuous streams of data and the researchers can develop tools to extract meaningful information from these data using machine learning techniques. Understanding these streams of data generated from IoT devices and analysing them to get meaningful information is a challenging issue and it leads to big data analytics. Machine learning algorithms and computational intelligence techniques is the only solution to handle big data from IoT prospective. Key technologies that are associated with IoT are also discussed in many research papers. Fig. 2 depicts an overview of
IoT big data and knowledge discovery process.

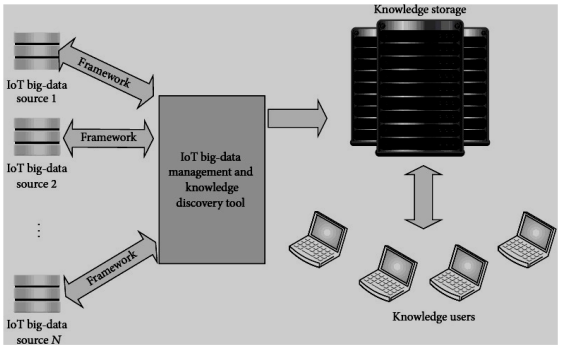


Fig. 2: IoT Big Data Knowledge Discovery

B. Cloud Computing For Big Data Analytics

The development of virtualization technologies have made supercomputing more accessible and affordable. Computing infrastructures that are hidden in virtualization software make systems to behave like a true computer, but with the flexibility of specification details such as number of processors, disk space, memory, and operating system. The use of these virtual computers is known as cloud computing which has been one of the most robust big data technique. Big Data and cloud computing technologies are developed with the importance of developing a scalable and on demand availability of resources and data. Cloud computing harmonize massive data by on demand access to configurable computing resources through virtualization techniques. The benefits of utilizing the Cloud computing include offering resources when there is a demand and pay only for the resources which is needed to develop the product. Simultaneously, it improves availability and cost reduction. Open challenges and research issues of big data and cloud computing are discussed in detail by many researchers which highlights the challenges in data

management, data variety and velocity, data storage, data processing, and resource management. So Cloud computing helps in developing a business model for all varieties of applications with infrastructure and tools.

Big data application using cloud computing should support data analytic and development. The cloud environment should provide tools that allow data scientists and business analysts to interactively and collaboratively explore knowledge acquisition data for further processing and extracting fruitful results.

This can help to solve large applications that may arise in various domains. In addition to this, cloud computing should also enable scaling of tools from virtual technologies into new technologies like spark, R, and other types of big data processing techniques.

C. Bio-Inspired Computing For Big Data Analytics

Bio-inspired computing is a technique inspired by nature to address complex real world problems. Biological systems are self organized without a central control. A bio-inspired cost minimization mechanism search and find the optimal data service solution on considering cost of data management and service maintenance. These techniques are developed by biological molecules such as DNA and proteins to conduct computational calculations involving storing, retrieving, and processing of data. A significant feature of such computing is that it integrates biologically derived materials to perform computational functions and receive intelligent performance. These systems are more suitable for big data applications. Huge amount of data are generated from variety of resources across the web since the digitization. Analysing these data and categorizing into text, image and video etc will require lot of intelligent analytics from data scientists and big data professionals. Proliferations of technologies are emerging like big data, IoT, cloud computing, bio inspired computing etc; whereas equilibrium of data can be done only by selecting right platform to analyse large and furnish cost effective results.

D. Quantum Computing For Big Data Analysis

A quantum computer has memory that is exponentially larger than its physical size and can manipulate an exponential set of inputs simultaneously. This exponential improvement in computer systems might be possible. If a real quantum computer is available now, it could have solved problems that are exceptionally difficult on recent computers, of course today's big data problems. The main technical difficulty in building quantum computer could soon be possible. Quantum computing provides a way to merge the quantum mechanics to process the information. In traditional computer, information is presented by long strings of bits which encode either a zero or a one. On the other hand, a quantum computer uses quantum bits or qubits. The difference between qubit and bit is that, a qubit is a quantum system that encodes the zero and the one into two distinguishable quantum states. Therefore, it can be capitalized on the phenomena of superposition and entanglement. It is because qubits behave quantumly. For example, 100 qubits in quantum systems require 2100 complex values to be stored in a classic computer system. It means that many big data problems can be solved much faster by larger scale quantum computers compared with classical computers.

Hence it is a challenge for this generation to built a quantum computer and facilitate quantum computing to solve big data problems.

V. TOOLS FOR BIG DATA PROCESSING

Large numbers of tools are available to process big data. In this section, we discuss some current techniques for analysing big data with emphasis on three important emerging tools namely MapReduce, Apache Spark, and Storm. Most of the available tools concentrate on batch processing, stream processing, and interactive analysis. Most batch processing tools are based on the Apache Hadoop infrastructure such as Mahout and Dryad. Stream data applications are mostly used for real time analytic. Some examples of largescale streaming platform are Strom and Splunk. The interactive analysis process allow users to directly interact in real time for their own analysis. For example Dremel and Apache Drill are the big data platforms that support interactive analysis. These tools help us in developing the big data projects. A fabulous list of big data tools and techniques is also discussed by much researchers. The typical work flow of big data project discussed by Huang et al is highlighted in this section and is depicted in Fig 4.

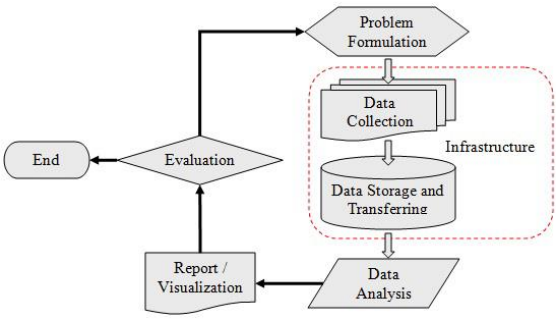


Fig. 4: Workflow of Big Data Project

A. Apache Hadoop and MapReduce

The most established software platform for big data analysis is Apache Hadoop and Mapreduce. It consists of Hadoop kernel, mapreduce, hadoop distributed file system (HDFS) and apache hive etc. Map reduce is a programming model for processing large datasets is based on divide and conquer method. The divide and conquer method is implemented in two steps such as Map step and Reduce Step. Hadoop works on two kinds of nodes such as master node and worker node. The master node divides the input into smaller sub problems and then distributes them to worker nodes in map step. Thereafter the master node combines the outputs for all the subproblems in reduce step. Moreover, Hadoop and MapReduce works as a powerful software framework for solving big data problems. It is also helpful in fault-tolerant storage and high throughput data processing.

B. Apache Mahout

Apache mahout aims to provide scalable and commercial machine learning techniques for large scale and intelligent data analysis applications. Core algorithms of mahout including clustering, classification, pattern mining, regression, dimensionality reduction, evolutionary algorithms, and batch based collaborative filtering run on top of Hadoop platform through map reduce framework. The goal of mahout is to build a vibrant, responsive, diverse community to facilitate discussions on the project and potential use cases. The basic objective of Apache mahout is to provide a tool for elevating big challenges. The different companies those who have implemented scalable machine learning algorithms are Google, IBM, Amazon, Yahoo, Twitter, and Facebook.

C. Apache Spark

Apache spark is an open source big data processing framework built for speed processing, and sophisticated analytics. It is easy to use and was originally developed in 2009 in UC Berkeleys AMPLab. It was open sourced in 2010 as an Apache project. Spark lets you quickly write applications in java, scala or python. In addition to map reduce operations, it supports SQL queries, streaming data, machine learning, and graph data processing. Spark runs on top of existing hadoop distributed file system (HDFS) infrastructure to provide enhanced and additional functionality. Spark consists of components namely driver program, cluster manager and worker nodes. The driver program serves as the starting point of execution of an application on the spark cluster. The cluster manager allocates the resources and the worker nodes to do the data processing in the form of tasks. Each application will have a set of processes called executors that are responsible for executing the tasks. The major advantage is that it provides support for deploying spark applications in an existing hadoop clusters. Fig 5 depicts the architecture diagram of Apache Spark.

D. Dryad

It is another popular programming model for implementing parallel and distributed programs for handling large context bases on dataflow graph. It consists of a cluster of computing nodes, and an user use the resources of a computer cluster to run their program in a distributed way. Indeed, a dryad user use thousands of machines, each of them with multiple processors or cores. The major advantage is that users do not need to know anything about concurrent programming. A dryad application runs a computational directed graph that is composed of computational vertices and communication channels. Therefore, dryad provides a large number of functionality including generating of job graph, scheduling of the machines for the available processes, transition failure handling in the cluster, collection of performance metrics, visualizing the job, invoking user defined policies and dynamically updating the job graph in response to these policy decisions without knowing the semantics of the vertices.

E. Storm

Storm is a distributed and fault tolerant real time computation system for processing large streaming data. It is specially designed for real time processing in contrasts with hadoop which is for batch processing. Additionally, it is also easy to set up and operate, scalable, fault-tolerant to provide competitive performances. The storm cluster is apparently similar to hadoop cluster. On storm cluster users run different topologies for different storm tasks whereas hadoop platform implements map reduce jobs for corresponding applications. There are number of differences between map reduce jobs and topologies. The basic difference is that map reduce job eventually finishes whereas a topology processes messages all the time, or until user terminate it. A storm cluster consists of two kinds of nodes such as master node and worker node. The master node and worker node implement two kinds of roles such as nimbus and supervisor respectively. The two roles have similar functions in accordance with jobtracker and tasktracker of

map reduce framework. Nimbus is in charge of distributing code across the storm cluster, scheduling and assigning tasks to worker nodes, and monitoring the whole system. The supervisor complies tasks as assigned to them by nimbus. In addition, it start and terminate the process as necessary based on the instructions of nimbus. The whole computational technology is partitioned and distributed to a number of worker processes and each worker process implements a part of the topology.

F. Apache Drill

Apache drill is another distributed system for interactive analysis of big data. It has more flexibility to support many types of query languages, data formats, and data sources. It is also specially designed to exploit nested data. Also it has an
objective to scale up on 10,000 servers or more and reaches the capability to process petabytes of data and trillions of records in seconds. Drill use HDFS for storage and map reduce to perform batch analysis.

G. Jaspersoft

The Jaspersoft package is an open source software that produce reports from database columns. It is a scalable big data analytical platform and has a capability of fast data visualization on popular storage platforms, including MangoDB, Cassandra, Redis etc. One important property of Jaspersoft is that it can quickly explore big data without extraction, transformation, and loading (ETL). In addition to this, it also have an ability to build powerful hypertext markup language (HTML) reports and dashboards interactively and directly from
big data store without ETL requirement. These generated reports can be shared with anyone inside or outside user's organization.

H. Splunk

In recent years a lot of data are generated through machine from business industries. Splunk is a real-time and intelligent platform developed for exploiting machine generated big data. It combines the up-to-the-moment cloud technologies and big data. In turn it helps user to search, monitor, and analyze their
machine generated data through web interface. The results are exhibited in an intuitive way such as graphs, reports, and alerts. Splunk is different from other stream processing tools. Its peculiarities include indexing structured, unstructured machine generated data, real-time searching, reporting analytical results and dashboards. The most important objective of Splunk is
to provide metrices for many application, diagnose problems for system and information technology infrastructures, and
intelligent support for business operations.

VI. CONCLUSION

A steadily expanding number of organizations has been endeavored to utilize Big Data and organizational analytics to analyze available data and assist with decision-making. For these organizations, influence the full potential that Big Data and organizational analytics can present to acquire competitive advantage. In any case, since Big Data and
organizational analytics are generally considered as new innovative in business worldview, there is a little exploration on how to handle them and leverage them adequately. While past literature has shown the advantages of utilizing Big Data in various settings, there is an absence of theoretically determined research on the most proficient method to use these solutions to acquire competitive advantage. This research recognizes the need to explore BDA through a comprehensive approach. Therefore, we focus on summarizing with the proposed development related to BDA themes on which we still have a restricted observational arrangement.

REFERENCES

[1] Adams, M.N.: Perspectives on Data Mining. International Journal of Market Research 52(1), 11–19 (2010)

[2] Asur, S., Huberman, B.A.: Predicting the Future with Social Media. In: ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 492–499 (2010)

[3] Bakshi, K.: Considerations for Big Data: Architecture and Approaches. In: Proceedings of the IEEE Aerospace Conference, pp. 1–7 (2012)

[4] Cebr: Data equity, Unlocking the value of big data. in: SAS Reports, pp. 1–44 (2012)

[5] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., Welton, C.: MAD Skills: New Analysis Practices for Big Data. Proceedings of the ACM VLDB Endowment 2(2), 1481–1492 (2009)

[6] Cuzzocrea, A., Song, I., Davis, K.C.: Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! In: Proceedings of the ACM

International Workshop on Data Warehousing and OLAP, pp. 101–104 (2011)

[7] Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In: Capgemini Reports, pp. 1–24 (2012)

[8] Elgendy, N.: Big Data Analytics in Support of the Decision Making Process. MSc Thesis, German University in Cairo, p. 164 (2013)

[9] EMC: Data Science and Big Data Analytics. In: EMC Education Services, pp. 1–508 (2012)

[10] He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z.: RCFile: A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems. In: IEEE International Conference on Data Engineering (ICDE), pp. 1199–1208 (2011)

[11] Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S.: Starfish: A Self-tuning System for Big Data Analytics. In: Proceedings of the Conference on Innovative Data Systems Research, pp. 261–272 (2011)

[12] Kubick, W.R.: Big Data, Information and Meaning. In: Clinical Trial Insights, pp. 26–28 (2012)

[13] Lee, R., Luo, T., Huai, Y., Wang, F., He, Y., Zhang, X.: Ysmart: Yet Another SQL-to-MapReduce Translator. In: IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 25–36 (2011)

[14] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big Data: The Next Frontier for Innovation, Competition, and Productivity. In: McKinsey Global Institute Reports, pp. 1–156 (2011)

[15] Mouthami, K., Devi, K.N., Bhaskaran, V.M.: Sentiment Analysis and Classification Based on Textual Reviews. In: International Conference on Information Communication and Embedded Systems (ICICES), pp. 271–276 (2013)

[16] Plattner, H., Zeier, A.: In-Memory Data Management: An Inflection Point for Enterprise Applications. Springer, Heidelberg (2011)

[17] Russom, P.: Big Data Analytics. In: TDWI Best Practices Report, pp. 1–40 (2011)

[18] Sanchez, D., Martin-Bautista, M.J., Blanco, I., Torre, C.: Text Knowledge Mining: An Alternative to Text Data Mining. In: IEEE International Conference on Data Mining Workshops, pp. 664–672 (2008)

[19] Serrat, O.: Social Network Analysis. Knowledge Network Solutions 28, 1–4 (2009)

[20] Shen, Z., Wei, J., Sundaresan, N., Ma, K.L.: Visual Analysis of Massive Web Session Data. In: Large Data Analysis and Visualization (LDAV), pp. 65–72 (2012)

[21] Song, Z., Kusiak, A.: Optimizing Product Configurations with a Data Mining Approach. International Journal of Production Research 47(7), 1733–1751 (2009)

[22] TechAmerica: Demystifying Big Data: A Practical Guide to Transforming the Business of Government. In: TechAmerica Reports, pp. 1–40 (2012)

[23] Van der Valk, T., Gijsbers, G.: The Use of Social Network Analysis in Innovation Studies: Mapping Actors and Technologies. Innovation: Management, Policy & Practice 12(1), 5–17 (2010)

[24] Zeng, D., Hsinchun, C., Lusch, R., Li, S.H.: Social Media Analytics and Intelligence. IEEE Intelligent Systems 25(6), 13–16 (2010)

[25] Zhang, L., Stoffel, A., Behrisch, M., Mittelstadt, S., Schreck, T., Pompl, R., Weber, S., Last, H., Keim, D.: Visual Analytics for the Big Data Era—A Comparative Review of State-of-the-Art Commercial Systems. In: IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 173–182 (2012)