Big data: The next frontier for Innovation, Competition and Productivity

Pankaj Sareen[#1], **Parveen Kumar**[#2]
#1 Assistant Professor, SPN College Mukerian
#2 Assistant Professor, SPN College Mukerian
[1]pankaj.sareen.mca@gmail.com
[2]parveenspn20@gmail.com

Abstract: **Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on hand database management tools or traditional data processing applications. Big data is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity. Despite these problems, big data has the potential to help companies improve operations and make faster, more intelligent decisions.**

Keywords-: **Big Data, Hadoop, RFID, Data Fusion, HBase, Cassandra.**

I. INTRODUCTION

Big Data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying and inf ormation privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making. And better decisions can result in greater operational efficiency, cost reduction and reduced risk [1].

Analysis of data sets can find new correlations to "spot business trends, prevent diseases, and combat crime and so on." Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, connectomics, complex physics simulations, biology and environmental research. Data sets are growing rapidly in part because they are increasingly gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exa bytes ($2.5×10^{18}$) of data are created. One question for large enterprises is determining who should own big data initiatives that affect the entire organization. Relational database management systems and desktop statistics and visualization packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information.

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale.in a 2001 research report and related lectures, META Group (now Gartner) analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Gartner, and now much of the industry, continue to use this "3Vs" model for describing big data. In 2012, Gartner updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." Additionally, a new V "Veracity" is added by some organizations to describe it.

“Big data” refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data—i.e., we don’t define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes). We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes).[2] Big data can be characterized by 3Vs: the extreme volume of data, the wide variety of types of data and the velocity at which the data must be must processed. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data, much of which cannot be integrated easily.

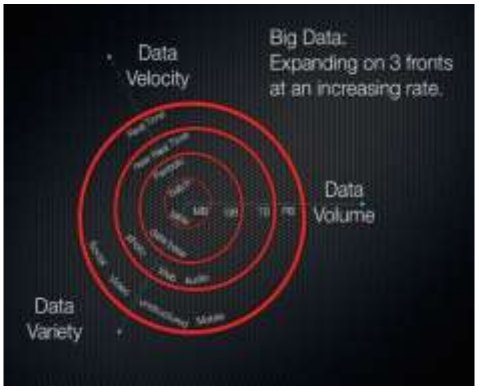

Fig.1 Data Is Growing at an Exponential Rate

Because big data takes too much time and costs too much money to load into a traditional relational database for analysis, new approaches to storing and analyzing data have emerged that rely less on data schema and data quality. Instead, raw data with extended metadata is aggregated in a data lake and machine learning and artificial intelligence (AI) programs use complex algorithms to look for repeatable patterns. Big data analytics is often associated with cloud computing because the analysis of large data sets in real-time requires a platform like Hadoop to store large data sets across a distributed cluster and MapReduce to coordinate, combine and process data from multiple sources. Although the demand for big data analytics is high, there is currently a shortage of data scientists and other analysts who have experience working with big data in a distributed, open source environment.

In the enterprise, vendors have responded to this shortage by creating Hadoop appliances to help companies take advantage of the semi-structured and unstructured data they own. Big data can be contrasted with small data, another evolving term that's often used to describe data whose volume and format can be easily used for self-service analytics. A commonly quoted axiom is that "big data is for machines; small data is for people. Gartner's definition of the 3Vs is still widely used, and in agreement with a consensual definition that states that "Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value".

II. CHARACTERISTICS [3]

Big data can be described by the following characteristics:

A. Volume

The quantity of generated data is important in this context. The size of the data determines the value and potential of the data under consideration, and whether it can actually be considered big data or not. The name 'big data' itself contains a term related to size, and hence the characteristic.

B. Variety

The type of content, and an essential fact that data analysts must know. This helps people who are associated with and analyze the data to effectively use the data to their advantage and thus uphold its importance.

C. Velocity

In this context, the speed at which the data is generated and processed to meet the demands and the challenges that lie in the path of growth and development.

D. Variability

The inconsistency the data can show at times—-which can hamper the process of handling and managing the data effectively.

E. Veracity

The quality of captured data, which can vary greatly. Accurate analysis depends on the veracity of source data.

F. Complexity

Data management can be very complex, especially when large volumes of data come from multiple sources. Data must be linked, connected, and correlated so users can grasp the information the data is supposed to convey.

III. ARCHITECTURE

In 2000, Seisint Inc. developed a C++-based distributed file-sharing framework for data storage and query. The system stores and distributes structured, semi-structured, and unstructured data across multiple servers. Users can build queries in a modified C++ called ECL. ECL uses an "apply schema on read" method to infer the structure of stored data at the time of the query. In 2004, LexisNexis acquired Seisint Inc. and in 2008 acquired Choice Point, Inc. and their high-speed parallel processing platform. The two platforms were merged into HPCC Systems and in 2011, HPCC was open-sourced under the Apache v2.0 License. Currently, HPCC and Quantcast File System are the only publicly available platforms capable of analyzing multiple exabytes of data [4].

In 2004, Google published a paper on a process called MapReduce that used such architecture. The MapReduce framework provides a parallel processing model and associated implementation to process huge amounts of data. With MapReduce, queries are split and distributed across parallel nodes and processed in parallel (the Map step). The results are then gathered and delivered (the Reduce step). The framework was very successful, so others wanted to replicate the algorithm. Therefore, an implementation of the MapReduce framework was adopted by an Apache open-source project named Hadoop.

MIKE2.0 is an open approach to information management that acknowledges the need for revisions due to big data implications identified in an article titled "Big Data Solution Offering". The methodology addresses handling big data in terms of useful permutations of data sources, complexity in interrelationships, and difficulty in deleting (or modifying) individual records [4].

Recent studies show that the use of multiple-layer architecture is an option for dealing with big data. The Distributed Parallel architecture distributes data across multiple processing units, and parallel processing units provide data much faster, by improving processing speeds. This type of architecture inserts data into a parallel DBMS, which implements the use of MapReduce and Hadoop frameworks. This type of framework looks to make the processing power transparent to the end user by using a front-end application server.

Big Data Analytics for Manufacturing Applications can be based on 5C architecture (connection, conversion, cyber, cognition, and configuration). The data lake allows an organization to shift its focus from centralized control to a shared model to respond to the changing dynamics of information management. This enables quick segregation of data into the data lake, thereby reducing the overhead time.

IV. TECHNIQUES THAT REQUIRE THE USE OF BIG DATA

There are many techniques that draw on disciplines such as statistics and computer science (particularly machine learning) that can be used to analyze datasets. In this section, we provide a list of some categories of techniques applicable across a range of industries. This list is by no means exhaustive. Indeed, researchers continue to develop new techniques and improve on existing ones, particularly in response to the need to analyze new combinations of data.

We note that not all of these techniques strictly require the use of big data-some of them can be applied effectively to smaller datasets (e.g., A/B testing, regression analysis). However, all of the techniques we list here can be applied to big data and, in general, larger and more diverse datasets can be used to generate more numerous and insightful results than smaller, less diverse ones[5].

A. A/B testing

A technique in which a control group is compared with a variety of test groups in order to determine what treatments (i.e., changes) will improve a given objective variable, e.g., marketing response rate. This technique is also known as split testing or bucket testing. An example application is determining what copy text, layouts, images, or colors will improve conversion rates on an e-commerce Web site. Big data enables huge numbers of tests to be executed and analyzed, ensuring that groups are of sufficient size to detect meaningful (i.e., statistically significant) differences between the control and treatment groups. When more than one variable is simultaneously manipulated in the treatment, the multivariate generalization of this technique, which applies statistical modeling, is often called “A/B/N” testing.

B. Association rule learning

A set of techniques for discovering interesting relationships, i.e., “association rules,” among variables in large databases. These techniques consist of a variety of algorithms to generate and test possible rules. One application is market basket analysis, in which a retailer can determine which products are frequently bought together and use this information for marketing, Used for data mining.

C. Cluster analysis

A statistical method for classifying objects that splits a diverse group into smaller groups of similar objects, whose characteristics of similarity are not known in advance. An example of cluster analysis is segmenting consumers into self-similar groups for targeted marketing. This is a type of unsupervised learning because training data are not used. This technique is in contrast to classification, a type of supervised learning.

D. Data fusion and data integration

A set of techniques that integrate and analyze data from multiple sources in order to develop insights in ways that are more efficient and potentially more accurate than if they were developed by analyzing a single source of data. Signal processing techniques can be used to implement some types of data fusion. One example of an application is sensor data from the Internet of Things being combined to develop an integrated perspective on the performance of a complex distributed system such as an oil refinery. Data from social media, analyzed by natural language processing, can be combined with real-time sales data, in order to determine what effect a marketing campaign is having on customer sentiment and purchasing behavior[14].

E. Data mining

A set of techniques to extract patterns from large datasets by combining methods from statistics and machine learning with database management. These techniques include association rule learning, cluster analysis, classification, and regression. Applications include mining customer data to determine segments most likely to respond to an offer, mining human resources data to identify characteristics of most successful employees, or market basket analysis to model the purchase behavior of customers.

F. Natural language processing (NLP)

A set of techniques from a subspecialty of computer science (within a field historically called "artificial intelligence") and linguistics that uses computer algorithms to analyze human (natural) language. Many NLP techniques are types of machine learning. One application of NLP is using sentiment analysis on social media to determine how prospective customers are reacting to a branding campaign.

V.BIG DATA TECHNOLOGIES [6]

There are a growing number of technologies used to aggregate, manipulate, manage, and analyze big data. We have detailed some of the more prominent technologies but this list is not exhaustive, especially as more technologies continue to be developed to support big data techniques, some of which we have listed:

A. Big Table

Proprietary distributed database system built on the Google File System. Inspiration for HBase.

B. Business intelligence (BI)

A type of application software designed to report, analyze, and present data. BI tools are often used to read data that have been previously stored in a data warehouse or data mart. BI tools can also be used to create standard reports that are generated on a periodic basis, or to display information on real-time management dashboards, i.e., integrated displays of metrics that measure the performance of a system.

C. Cassandra

An open source (free) database management system designed to handle huge amounts of data on a distributed system. This system was originally developed at Facebook and is now managed as a project of the Apache Software foundation.

D. Cloud Computing

A computing paradigm in which highly scalable computing resources, often configured as a distributed system, are provided as a service through a network.

E. Distributed System

Multiple computers, communicating through a network, used to solve a common computational problem. The problem is divided into multiple tasks, each of which is solved by one or more computers working in parallel. Benefits of distributed systems include higher performance at a lower cost (i.e., because a cluster of lower-end computers can be less expensive than a single higher-end computer), higher reliability (i.e., because of a lack of a single point of failure), and more scalability (i.e., because increasing the power of a distributed system can be accomplished by simply adding more nodes rather than completely replacing a central computer).

F. Hadoop

An open source (free) software framework for processing huge datasets on certain kinds of problems on a distributed system. Its development was inspired by Google's MapReduce and Google File System. It was originally developed at Yahoo! and is now managed as a project of the Apache Software Foundation.

G. Mashup

An application that uses and combines data presentation or functionality from two or more sources to create new services. These applications are often made available on the Web, and frequently use data accessed through open application programming interfaces or from open data sources.

VI. ADVANTAGES OF BIG DATA [7]:

Big data identified five broadly applicable ways to leverage big data that offer transformational potential to create value and have implications for how organizations will have to be designed, organized, and managed:

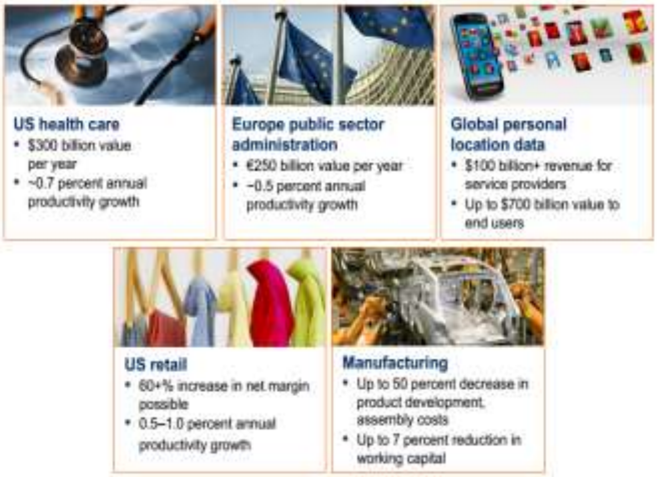
US health care
• $300 billion value per year
• ~0.7 percent annual productivity growth
Europe public sector administration
• €250 billion value per year
• ~0.5 percent annual productivity growth
Global personal location data
• $100 billion+ revenue for service providers
• Up to $700 billion value to end users
US retail
• 60+% increase in net margin possible
• 0.5–1.0 percent annual productivity growth
Manufacturing
• Up to 50 percent decrease in product development, assembly costs
• Up to 7 percent reduction in working capital

Fig.2 Big data can generate significant financial value across sectors

A. Creating transparency

Simply making big data more easily accessible to relevant stakeholders in a timely manner can create tremendous value. In the public sector, for example, making relevant data more readily accessible across otherwise separated departments can sharply reduce search and processing time. In manufacturing, integrating data from R&D, engineering, and manufacturing units to enable concurrent engineering can significantly cut time to market and improve quality.

B. Enabling experimentation to discover needs, exposes variability, and improves performance

As they create and store more transactional data in digital form, organizations can collect more accurate and detailed performance data (in real or near real time) on everything from product inventories to personnel sick days. IT enables organizations to instrument processes and then set up controlled experiments. Using data to analyze variability in performance-that which either occurs naturally or is generated by controlled experiments-and to understand its root causes can enable leaders to manage performance to higher levels.

C. Segmenting populations to customize actions

Big data allows organizations to create highly specific segmentations and to tailor products and services precisely to meet those needs. This approach is well known in marketing and risk management but can be revolutionary elsewhere—for example, in the public sector where an ethos of treating all citizens in the same way is commonplace. Even consumer goods and service companies that have used segmentation for many years are beginning to deploy ever more sophisticated big data techniques such as the real-time micro segmentation of customers to target promotions and advertising.

D. Replacing/supporting human decision making with automated algorithms

Sophisticated analytics can substantially improve decision making, minimize risks, and unearth valuable insights that would otherwise remain hidden. Such analytics have applications for organizations from tax agencies that can use automated risk engines to flag candidates for further examination to retailers that can use algorithms to optimize decision processes such as the automatic fine-tuning of inventories and pricing in response to real-time in-store and online sales. In some cases, decisions will not necessarily be automated but augmented by analyzing huge, entire datasets using big data techniques and technologies rather than just smaller samples that individuals with spreadsheets can handle and understand. Decision making may never be the same; some organizations are already making better decisions by analyzing entire datasets from customers, employees, or even sensors embedded in products.[12]

E. Innovating new business models, products, and services

Big data enables companies to create new products and services, enhance existing ones, and invent entirely new business models. Manufacturers are using data obtained from the use of actual products to improve the development of the next generation of products and to create innovative after-sales service offerings. The emergence of real-time location data has created an entirely new set of location- 6 based services from navigation to pricing property and casualty insurance based on where, and how, people drive their cars.

F. Use of big data will matter across sectors:

Computer and electronic products and information sectors (Cluster A), traded globally, stand out as sectors that have already been experiencing very strong productivity growth and that are poised to gain substantially from the use of big data. Two services sectors (Cluster B)-finance and insurance and government-are positioned to benefit very strongly from big data as long as barriers to its use can be overcome. Several sectors (Cluster C) have experienced negative productivity growth, probably indicating that these sectors face strong systemic barriers to increasing productivity. Among the remaining sectors, we see that globally traded sectors (mostly Cluster D) tend to have experienced higher historical productivity growth, while local services (mainly Cluster E) have experienced lower growth. While all sectors will have to overcome barriers to capture value from the use of big data, barriers are structurally higher for some than for others (Exhibit 3). For example, the public sector, including education, faces higher hurdles because of a lack of data-driven mind-set and available data. Capturing value in health care faces challenges given the relatively low IT investment performed so far. Sectors such as retail, manufacturing, and professional services may have relatively lower degrees of barriers to overcome for precisely the opposite reasons.

A significant constraint on realizing value from big data will be a shortage of talent, particularly of people with deep expertise in statistics and machine learning, and the managers and analysts who know how to operate companies by using insights from big data [11] In the United States, we expect big data to rapidly become a key determinant of competition across sectors. But we project that demand for deep analytical positions in a big data world could exceed the supply being produced on current trends by 140,000 to 190,000 positions (Exhibit 4). Furthermore, this type of talent is difficult to produce, taking years of training in the case of someone with intrinsic mathematical abilities. Although our quantitative analysis uses the United States as illustration, we believe that the constraint on this type of talent will be global, with the caveat that some regions may be able to produce the supply that can fill talent gaps in other regions [8].

Big data’s increasing economic importance also raises a number of legal issues, especially when coupled with the fact that data are fundamentally different from many other assets. Data can be copied perfectly and easily combined with other data. The same piece of data can be used simultaneously by more than one person. All of these are unique characteristics of data compared with physical assets.

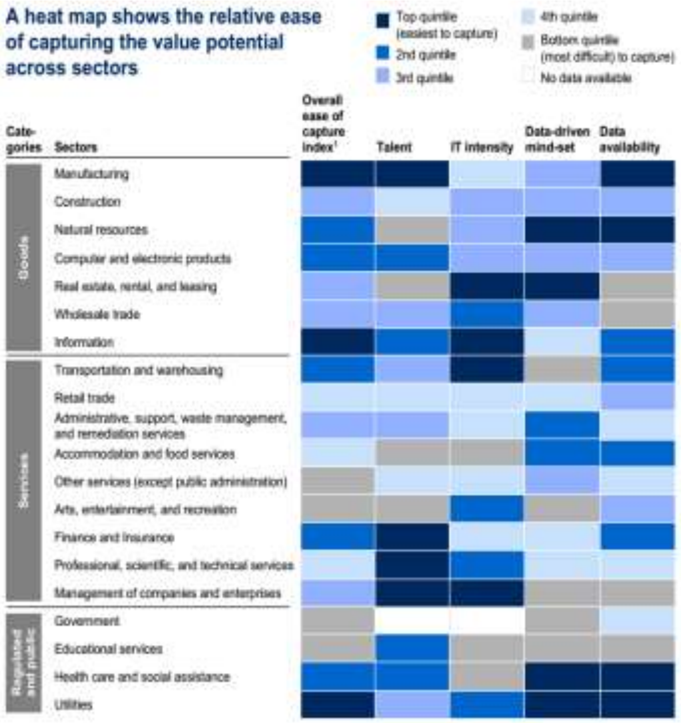

Fig.3 detailed definitions and metrics used for each of the criteria with big data

VII. CRITIQUES OF THE BIG DATA PARADIGM

"A crucial problem is that we do not know much about the underlying empirical micro-processes that lead to the emergence of the typical network characteristics of Big Data". In their critique, Snijders, Matzat, and Reips point out that often very strong assumptions are made about mathematical properties that may not at all reflect what is really going on at the level of micro-processes. Mark Graham has leveled broad critiques at Chris Anderson's assertion that big data will spell the end of theory: focusing in particular on the notion that big data must always be contextualized in their social, economic, and political contexts. Even as companies invest eight- and nine-figure sums to derive insight from information streaming in from suppliers and customers, less than 40% of employees have sufficiently mature processes and skills to do so. To overcome this insight deficit, "big data", no matter how comprehensive or well analyzed, must be complemented by "big judgment," according to an article in the Harvard Business Review [9].

Much in the same line, it has been pointed out that the decisions based on the analysis of big data are inevitably "informed by the world as it was in the past, or, at best, as it currently is". Fed by a large number of data on past experiences, algorithms can predict future development if the future is similar to the past. If the systems dynamics of the future change, the past can say little about the future. For this, it would be necessary to have a thorough understanding of the systems dynamic, which implies theory. As a response to this critique it has been suggested to combine big data approaches with computer simulations, such as agent-based models and Complex Systems. Agent-based models are increasingly getting better in predicting the outcome of social complexities of even unknown future scenarios through computer simulations that are based on a collection of mutually interdependent algorithms. In addition, use of multivariate methods that probe for the latent structure of the data, such as factor analysis and cluster analysis, have proven useful as analytic approaches that go well beyond the bi-variate approaches (cross-tabs) typically employed with smaller data sets[15].

In health and biology, conventional scientific approaches are based on experimentation. For these approaches, the limiting factor is the relevant data that can confirm or refute the initial hypothesis. A new postulate is accepted now in biosciences: the information provided by the data in huge volumes (omics) without prior hypothesis is complementary and sometimes necessary to conventional approaches based on experimentation. In the massive approaches it is the formulation of a relevant hypothesis to explain the data that is the limiting factor. The search logic is reversed and the limits of induction ("Glory of Science and Philosophy scandal", C. D. Broad, 1926) are to be considered.

Privacy advocates are concerned about the threat to privacy represented by increasing storage and integration of personally identifiable information; expert panels have released various policy recommendations to conform practice to expectations of privacy.

VIII. CRITIQUES OF BIG DATA EXECUTION

Big data has been called a "fad" in scientific research and its use was even made fun of as an absurd practice in a satirical example on "pig data". Researcher danahboydhas raised concerns about the use of big data in science neglecting principles such as choosing a representative sample by being too concerned about actually handling the huge amounts of data. This approach may lead to results bias in one way or another. Integration across heterogeneous data resources—some that might be considered "big data" and others not—presents formidable logistical as well as analytical challenges, but many researchers argue that such integrations are likely to represent the most promising new frontiers in science. In the provocative article "Critical Questions for Big Data", the authors title big data a part of mythology: "large data sets offer a higher form of intelligence and knowledge with the aura of truth, objectivity, and accuracy". Users of big data are often "lost in the sheer volume of numbers", and "working with Big Data is still subjective, and what it quantifies does not necessarily have a closer claim on objective truth". Recent developments in BI domain, such as pro-active reporting especially target improvements in usability of Big Data, through automated filtering of non-useful data and correlations [10].

Big data analysis is often shallow compared to analysis of smaller data sets. In many big data projects, there is no large data analysis happening, but the challenge is the extract, transform, load part of data preprocessing.

Big data is a vague term but at the same time an "obsession" with entrepreneurs, consultants, scientists and the media. Big data showcases such as Google Flu Trends failed to deliver good predictions in recent years, overstating the flu outbreaks by a factor of two. Similarly, Academy awards and election predictions solely based on Twitter were more often off than on target. Big data often poses the same challenges as small data; and adding more data does not solve problems of bias, but may emphasize other problems. In particular data sources such as Twitter are not representative of the overall population, and results drawn from such sources may then lead to wrong conclusions. Google Translate—which is based on big data statistical analysis of text—does a good job at translating web pages. However, results from specialized domains may be dramatically skewed.

IX. CONCLUSION

The availability of Big Data, low-cost commodity hardware, and new information management and analytic software have produced a unique moment in the history of data analysis. The convergence of these trends means that we have the capabilities required to analyze astonishing data sets quickly and cost-effectively for the first time in history. These capabilities are neither theoretical nor trivial. They represent a genuine leap forward and a clear opportunity to realize enormous gains in terms of efficiency, productivity, revenue, and profitability. The Age of Big Data is here, and these are truly revolutionary times if both business and technology professionals continue to work together and deliver on the promise.

References

[1]. https://en.wikipedia.org/wiki/Big_data

[2]. http://www.datasciencecentral.com/profiles/blogs/top-10-big-data-and-analytics-references

[3]. https://resources.sei.cmu.edu/asset_files/.../2014_017_101_89659.pdf

[4]. http://thinkbig.teradata.com/leading_big_data_technologies/big-data-reference-architecture

[5]. http://www.sciencedirect.com/science/article/pii/S2214579615000027

[6]. http://resources.sei.cmu.edu/library/asset-view.cfm?assetid=89485

[7]. A special report on managing information: Data, data everywhere," *The Economist*, February 25, 2010; and special issue on "Dealing with data," *Science*, February 11, 2011.

[8]. "Internet of Things" refers to sensors and actuators embedded in physical objects, connected by networks to computers. See Michael Chui, Markus Löffler, and Roger Roberts, "The Internet of Things," *McKinsey Quarterly*, March 2010.

[9]. https://resources.sei.cmu.edu/asset_files/.../2014_017_101_89659.pdf

[10]. www.acadgild.com/Big-Data-&-Hadoop

[11]. bigdatawg.nist.gov/_uploadfiles/M0047_v2_3966324695.pdf

[12]. www.ibmbigdatahub.com/ibm-solutions/customer-references

[13]. http://h20195.www2.hp.com/V2/getpdf.aspx/4AA5-6141ENW.pdf

[14]. https://www.ibm.com/developerworks/community/groups/service/html/communityview?communityUuid=c0a9d736-a0f4-42f4-8664-d3168bbff284

[15]. http://www.slideshare.net/robdthomas/ibm-big-data-references