Big Data Issues and Challenges

Data analysis, storing, processing, issues, challenges and future scope

Mrs. Sneha Gupta
Department of Computer Science & Engineering
Priyadarshini Bhagwati College of Engineering, PBCOE
Nagpur, India
e-mail: snehathebest90916@gmail.com

Mr. Manoj S. Chaudhari
Department of Computer Science & Engineering
Priyadarshini Bhagwati College of Engineering, PBCOE
Nagpur, India
e-mail: manojchaudhary2@gmail.com

Abstract— Big data refers to large volume of data which requires new technologies and architectures to make possible to extract value from it by capturing and analysis process. Big Data has emerged because we are living in a society which makes increasing use of data intensive technologies. Due to such large size of data, it becomes very difficult to perform effective analysis using the existing traditional techniques. Since Big data is a recent upcoming technology in the market which can bring huge benefits to the business organizations, it becomes necessary that various challenges and issues associated in bringing and adapting to this technology are need to be understood.

***Keywords** - Big data, Hadoop, MapReduce, HDFS, NoSQL, Data Security, Data Analytics, Data Analysis, Big data research*

*****

I. Introduction

Progress in digital data acquisition and storage technology has resulted in the growth of huge databases. This has occurred in all the areas of human endeavor, from the mundane (such as supermarket transaction data, credit card usage records, telephone call details and government statistics) to the more exotic (such as images of astronomical bodies, molecular databases, and medical records). Developed economies make increasing use of data-intensive technologies. There are 4.6 billion mobile-phone subscriptions worldwide and between 1 billion and 2 billion people accessing the internet. The interest has grown in the possibility of tapping these data and extracting from them the information that bring value and profit to organization [1].

The today world is alive with the electronic information. Every second of the day, computers and other electronic system are creating, processing, transmitting and receiving huge volumes of information. We create around 2 exabytes (10^{18} bytes) of data every two days. This huge volume includes 2 million searches processed by Google each minute, 4,000 hours of video uploaded into youtube every hour and 144 billion emails sent around the world every day. This equate to the entire contents of the US Library of Congress passing across the internet every 10 seconds [2].

The concept of big data has been endemic within computer science since the earliest days of computing. "Big Data" originally meant the volume of data that could not be processed efficiently by traditional database methods and tools. Each time a new storage medium was invented, the amount of data accessible exploded because it could be easily accessed. The original definition focused on structured data, but most researchers and practitioners have come to realize that most of the world's information resides in massive, unstructured information, largely in the form of text and imagery. The explosion of data has not been accompanied by a corresponding new storage medium.

We define "Big Data" as the amount of data just beyond technology's capability to store, manage and process efficiently. These imitations are only discovered by a robust analysis of the data itself, explicit processing needs, and the capabilities of the tools (hardware, software, and methods) used to analyze it. As with any new problem, the conclusion of how to proceed may lead to a recommendation that new tools need to be forged to perform the new tasks. As little as 10 years ago, we were only thinking of tens to hundreds of gigabytes of storage for our personal computers. Today, we are thinking in tens to hundreds of terabytes. Thus, big data is a moving target. Put another way, it is that amount of data that is just beyond our immediate grasp, we have to work hard to store it, access it, manage it, and process it. The current growth rate in the amount of data collected is staggering. A major challenge for IT researchers and practitioners is that this growth rate is fast exceeding our ability to both:

- Design appropriate systems to handle the data effectively.
- Analyze it to extract relevant meaning for decision making.

In this paper we identify critical issues associated with data storage, management, and processing. To the best of our knowledge, the research literature has not effectively addressed these issues [3].

A. Defination of Big Data

"Big data" is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

The definition consists of 23 words, 181 characters with quotation marks. The latter is a hint that Gartner believes "big data" will be the new normal in the very foreseeable future. We also like that this definition reflects relativity of big data[4].

B. Importance of Big Data

In August 2010, the White House, OMB, and OSTP proclaimed that big data is a national challenge and priority along with healthcare and national security [8]. The National Science Foundation, the National Institutes of Health, the U.S. Geological Survey, the Departments of Defense and Energy, and the Defense Advanced Research Projects Agency announced a joint R&D initiative in March 2012 that will invest more than $200 million to develop new big data tools and techniques. Its goal is to advance our "…understanding of the technologies needed to manipulate and mine massive amounts of information; apply that knowledge to other scientific fields "as well as address the national goals in the areas of health energy defense, education and researcher" [9].

Five ways to leverage big data [5]

- Big Data can unlock significant value by making information transparent. There is still a significant amount of information that is not yet captured in digital form, e.g., data that are on paper, or not made easily accessible and searchable through networks. We found that up to 25 percent of the effort in some knowledge worker workgroups consists of searching for data and then transferring them to another (sometimes virtual) location. This effort represents a significant source of inefficiency.
- As organisations create and store more transactional data in digital form, they can collect more accurate and detailed performance information on everything from product inventories to sick days and therefore expose variability and boost performance. In fact, some leading companies are using their ability to collect and analyse big data to conduct controlled experiments to make better management decisions.
- Big Data allows ever-narrower segmentation of customers and therefore much more precisely tailored products or services.
- Sophisticated analytics can substantially improve decision-making, minimise risks, and unearth valuable insights that would otherwise remain hidden.
- Big Data can be used to develop the next generation of products and services. For instance, manufacturers are using data obtained from sensors embedded in products to create innovative after-sales service offerings such as proactive maintenance to avoid failures in new products

C. *Characteristics of Big Data*

Big data can be described by the following characteristics [6]:

- **Volume -** The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.
- **Variety -** The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.
- **Velocity -** The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.
- **Variability -** This is a factor which can be a problem for those who analyze the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.
- **Veracity -** The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.
- **Complexity -** Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.

D. *Why Big Data Matters's?*

The real issue is not that you are acquiring large amounts of data. It's what you do with the data that counts. The hopeful vision is that organizations will be able to take data from any source, harness relevant data and analyze it to find answers that enable 1) cost reductions, 2) time reductions, 3) new product development and optimized offerings, and 4) smarter business decision making. For instance, by combining big data and high-powered analytics, it is possible to: [6]

- Determine root causes of failures, issues and defects in near-real time, potentially saving billions of dollars annually.
- Optimize routes for many thousands of package delivery vehicles while they are on the road.
- Analyze millions of SKUs to determine prices that maximize profit and clear inventory.
- Generate retail coupons at the point of sale based on the customer's current and past purchases.
- Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.
- Recalculate entire risk portfolios in minutes.
- Quickly identify customers who matter the most.
- Use clickstream analysis and data mining to detect fraudulent behavior.

E. *Big Data Technologies and Tools*

Intimidating or not, Big Data is a natural result of our collective obsession with technology. Put simply, it is a reference to mountainous volumes of data that can be amassed by companies or even individuals for that matter. Around a decade ago, it was only the scientific community that actually had what took to wade through reams of data. Today, the need for this kind of diligence is something that is faced by almost every stream [7].

Though Big Data may sound futuristic, it does need certain exceptional technologies to efficiently process huge volumes of data in a good span of time. Here are some of the technologies that can be applied for handling big data.

- **Massively Parallel Processing (MPP)** – This involves a coordinated processing of a program by multiple processors (200 or more in number). Each of the processors makes use of its own operating system and memory and works on different parts of the program. Each part communicates via messaging interface. An MPP system is also known as "loosely coupled" or "shared nothing" system.
- **Distributed file system or network file system** allows client nodes to access files through a computer network. This way a number of users working on multiple machines will be able to share files and storage resources. The client nodes will not be able to access the block storage but can interact through a network protocol. This enables a restricted access to the file system depending on the access lists or capabilities on both servers and clients which is again dependent on the protocol.
- **Apache Hadoop** is key technology used to handle big data, its analytics and stream computing. Apache Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It can be scaled up from a single server to thousands of machines and with a very high degree of fault tolerance. Instead of relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer.
- **Data Intensive Computing** is a class of parallel computing application which uses a data parallel approach to process big data. This works based on the principle of collocation of data and programs or algorithms used to perform computation. Parallel and distributed system of inter-connected stand alone computers that work together as a single integrated computing resource is used to process / analyze big data.

According to IBM, 80 per cent of world's data is unstructured and most businesses don't even attempt to use this data to their advantage. Once the technologies to analyze big data reach their peak, it will become easier for companies to analyze massive datasets, identify patterns and then strategically plan their moves based on consumer requirements that identified through historic data.

II. BIG DATA ISSUES

There are three fundamental issue areas that need to be addressed while dealing with big data. Each of these represents a large set of technical research problems in its own right [3]

A. *Storage and Transport Issues*

Due to the introduction of the various social media websites which allows users to upload the data in any format (such as text, images, video, audio, files). The volume of the user generated data has increased to the level which are hard to manage. Even with the introduction of the new disk technologies, the limitation to store data in the single disk can be maximum to some terabytes; so to process the petabytes or exabytes of the data, large amount of the disks and processor are required to work in parallel. Also if you want to transfer the petabytes of the data over the network towards the processing system, its needs a lot of time to transfer the data and processed it. As a result of this now a days the data is processed in the storage location itself so that the network cost can be minimized and data can be process faster. In other words, "bring the code to the data", vs. the traditional method of "bring the data to the code."

B. *Management Issues*

Management of the big data is one the most difficult problem to address. All the business are already using the traditional ways of the data management with the drastic addition of the new data, it's become hard to manage data. The problem may go worst if the data was distributed geographically and "owned" and "managed" by multiple entities. As of now there is no perfect big management solution yet.

C. *Processing Issues*

The amount of the data that can be processed by the single processor is limited even with the use of the recent technologies. In order to process the exabytes of data lot of parallel processing is needed. For simplicity, assume the data is chunked into blocks of 8 words, so 1 exabyte = 1K petabytes. Assuming a processor expends 100 instructions on one block at 5 gigahertz, the time required for end-to-end processing would be 20 nanoseconds. To process 1K petabytes would require a total end-to-end processing time of roughly 635 years. Thus, effective processing of exabytes of data will require extensive parallel processing and new analytics algorithms in order to provide timely and actionable information.

III. BIG DATA CHALLENGES

There are numerous challenges and long term research requires to working with big data. The system that need to be designed to handle the big need to have the clear understanding of the users, data, sources, and value that need to be drawn from it. There are unknown challenges that will arise with each increase in scale and development of new analytics [3].

A. *Data Input and Output Processes*

It was found that getting the data is easier than processing. The data analysis needs to understand the complete the behavior of the raw data, its structure & the value that need to be drawn from it. Processing data stored in the relational database is quite easy and various tools are readily available for it to process. But as the volume increases & data behavior changes; the relational database technologies cannot be used. So it very challenging for the data scientist to devise the algorithm that can fully understand the raw data and process it.

B. *Quality versus Quantity*

An emerging challenge for big data users is "quantity vs. quality". As users acquire and have access to more data (quantity), they often want even more. For some users, the acquisition of data has become an addiction. Perhaps, because they believe that with enough data, they will be able to perfectly explain whatever phenomenon they are interested in.

Conversely, a big data user may focus on quality which means not having all the data available, but having a (very) large quantity of high quality data that can be used to draw precise and high-valued conclusions.

C. Data Growth versus Data Expansion

Most organizations expect their data to grow over their lifetime as the organization increases its services, its business and business partners and clients, its projects and facilities, and its employees. Few businesses adequately consider data expansion, which occurs when the data records grow in richness, when they evolve over time with additional information as new techniques, processes and information demands evolve. Most data is time-varying – the same data items can be collected over and over with different values based on a timestamp. Much of this data is required for retrospective analysis – particularly that which is used in estimative and predictive analytics.

D. Speed versus Scale

As the volume of data grows, the "big" may morph from the scale of the data warehouse to the amount of data that can be processed in a given interval, say 24 hours. Gaining insight into the problem being analyzed is often more important than processing all of the data. Time-to-information is critical when one considers (near) real-time processes that generate near-continuous data, such as radio frequency identifiers (RFIDs – used to read electronic data wirelessly, such as with EZPass tags) and other types of sensors. An organization must determine how much data is enough in setting its processing interval because this will drive the processing system architecture, the characteristics of the computational engines, and the algorithm structure and implementation. That said, another major challenge is data dissemination. The bottleneck is the communications middleware. While communication hardware speeds are increasing with new technologies, message handling speeds are decreasing only slowly. The computation versus communication dichotomy has not been fully resolved by large data store systems such as HDFS or Accumulo for exabyte-sized data sets.

E. Structured versus Unstructured Data

Translation between structured data with well defined data definitions (often in tables) as stored in relational databases, and unstructured data (e.g., free text, graphics, multi-media, etc.) suitable for analytics can impede end-to-end processing performance. The emergence of non-relational, distributed, analytics relational databases, analytics oriented databases such as NoSQL, MongoDB, SciDB and linked data DBs provides dynamic flexibility in representing and organizing information.

F. Data Ownership

Data ownership presents a critical and ongoing challenge, particularly in the social media arena. While petabytes of social media data reside on the servers of Facebook and Twitter, it is not really owned by them (although they may contend so because of residency). Certainly, the "owners" of the pages or accounts believe they own the data. This dichotomy will have to be resolved in court. Kaisler, Money and Cohen addressed this issue with respect to cloud computing as well as other legal aspects that we will not delve into here [11].

G. Compliance and Security

In certain domains, such as social media and health information, as more data is accumulated about individuals, there is a fear that certain organizations will know too much about individuals. For example, data collected in electronic health record systems in accordance with HIPAA/HITECH provisions is already raising concerns about violations of one's privacy. Developing algorithms that randomize personal data among a large data set enough to ensure privacy is a key research problem. Perhaps the biggest threat to personal security is the unregulated accumulation of data by numerous social media companies. This data represents a severe security concern, especially when many individuals so willingly surrender such information. Questions of accuracy, dissemination, expiration, and access abound. For example, the State of Maryland became the first state to prohibit by law employers asking for Facebook and other social media passwords during employment interviews and afterwards.

H. The Value of "Some Data" versus "All Data"

Not all data is created equal; some data is more valuable than other data – temporally, spatially, contextually, etc. Previously, storage limitations required data filtering and deciding what data to keep. Historically, we converted what we could and threw the rest away (figuratively, and often, literally).

I. Distributed Data and Distributed Processing

The allure of hardware replication and system expandability as represented by cloud computing along with the MapReduce and Message Passing Interface (MPI) parallel programming systems offers one solution to these challenges by utilizing a distributed approach. Even with this approach, significant performance degradation can still occur because of the need for communication between the nodes.

J. Data Visualization

To fully take advantage of visual analysis and analytics, organizations will need to address several challenges related to visualization of big data. Here we have outlined some of those key challenges.

- Meeting the need for speed
- Understanding the data
- Addressing data quality
- Displaying meaningful results
- Dealing with outliers

IV. BIG DATA ANALYTICS CHALLENGES

Processing big data is a major challenge, perhaps more so than the storage or management problem. There are many types of analytics: descriptive, estimative, predictive, and prescriptive, leading to various types of decision and optimization models. Kaisler [10] presents another decomposition of analytics into 16 categories based on the types of problems to be addressed, including econometric models, game theory, control theory, evolutionary computation, and simulation models. The new normal is agile, advanced, predictive analytics that adapt readily to changing data sets and streams and yield information and knowledge to improve services and operations across academia, industry, and government.

A. Scaling

Processing big data is a major challenge, perhaps more so than the storage or management problem. There are many types of analytics: descriptive, estimative, predictive, and prescriptive, leading to various types of decision and optimization models [3]. A critical issue is whether or not an analytic process scales as the data set increases by orders of magnitude. Every algorithm has a "knee" – the point at which the algorithm's performance ceases to increase linearly with increasing computational resources and starts to plateau or, worse yet, peak, turn over, and start decreasing. Solving this problem requires a new algorithm for the problem, or rewriting the current algorithm to "translate" the knee farther up the scale. An open research question is whether for any given algorithm, there is a fundamental limit to its scalability. These limits are known for specific algorithms with specific implementations on specific machines at specific scales. General computational solutions, particularly using unstructured data, are not yet known.

B. Finding the Needle in the Haystack

This challenge focuses on finding the key data that provides leverage for decision-making within a problem space. A needle-in-a-haystack problem is one in which the right answer is very difficult to determine in advance, but very easy to verify once you know where the needle is [9]. Suppose we characterize it as finding the one right answer within a pool of 1,000,000 wrong answers. If the decision process is wrong 0.1% of the time, then of the 100 answers proposed, there are 101 answers, only one of which is 'correct". As Felten notes, any research area depending on this approach will suffer this problem, especially if it relies on statistical analysis. There are only two ways out of this problem: reduce the size of the haystack, or improve our search, analysis and decision-making procedures.

C. Turning Straw into Gold

This challenge focuses on processing a large set of discrete data points into high-valued data. Consider a data visualization of anyone Facebook Friends which represents a small subset of the hundreds of millions or so people using Facebook. As the number of edges emanating from "central" nodes increases, the overall mesh complexity increases nonlinearly. Finding sub graphs within graphs with particular sets of features may not be a linearly computationally tractable problem using standard graph-traversal and analysis algorithms. One approach to solving the representation problem is to parse semistructured text and convert it to linked data using the Resource Description Framework (RDF) triple format. The explosion of resulting text is often on the order of 10:1 due to the use of RDF tags to identify components of an RDF structure. This translates the problem from processing semistructured text to finding relationships over a very large, real-world, partially connected mesh. Extracting mesh structural features is critical to identifying patterns and anomalies. Inference across mesh substructures is akin to "guilt by association", e.g., if a person is a drug abuser, it is likely his friends are as well. Beliefs are propagated across the mesh resulting in a further explosion of data [12].

Another challenge is the time-varying nature of very large graphs. Determining the change between two snapshots – either statically or continuously for a given interval – is a computationally explosive problem. This type of problem occurs frequently, but requires more intensive computation and new algorithms when applied in near-real-time analytics such as network attack monitoring. To us, it is clear: **Gold Mining is not equal to Data Mining!** Different algorithms with greater reliance on reasoning (machine learning – symbolic, not statistical), computational social science, and domain-based analysis are essential to seeing the "big picture" in order to interpret and extract actionable patterns of behavior, meaning and nuggets of intelligence for informed decision-making.

D. A Hybrid of Techniques

Given a very large heterogeneous data set, a major challenge is to figure out what data one has and how to analyze it. Unlike the previous two sections, hybrid data sets combined from many other data sets hold more surprises than immediate answers. To analyze these data will require adapting and integrating multiple analytic techniques to "see around corners", e.g., to realize that new knowledge is likely to emerge in a non-linear way. It is not clear that statistical analysis methods, as Ayres argues, are or can be the whole answer. Nassim Taleb [16] addressed the potential for undirected and unpredicted effects arising from events that are outliers that lie outside the realm of regular expectations, because nothing in the past can convincingly point to their possibility. Such events often have an extreme impact – a "shock" to the system that can force new behaviors. Because we do not expect it, we cannot predict it. Thus, we can only try to explain what happened retrospectively. With more data, the likelihood of identifying such events rises and will force us to re-evaluate our estimative and predictive analytical tools.

E. Know the World

With the (over)abundance of data available to us, a major research question is: can we model world systems, say, on the order of our desire to model and predict the weather? For example, can we forecast global political and/or economic stability at a given temporal interval? Underlying this challenge are questions of modeling natural science, social and cultural interactions at different scales; understanding how societies function and the causes of global unrest; and understanding how human societies produce and consume resources and the resource flows around the world. Such questions are important to transnational and global businesses determining where to allocate resources and invest in infrastructure. Creating models in a computer is standard science. But, creating world-encompassing models (or, even, domain-encompassing) models is not yet feasible. Nevertheless, the challenge is to begin building such models that will allow us to comprehend systems at both the scope and granularity necessary to answer fundamental questions of cause and effect. Consider the effects of natural disasters (such as the tsunami affecting Japan's decisions regarding nuclear power), economic system failure (the US housing and banking crisis), the recession/depression and the Arab Spring causing major evolution in governments, or technological innovation (such as the rise of social media). The key question for many decision makers – business, academia, government - is, "what does it all mean?" followed by "what is likely to happen next?" These are all "wicked problems" [14] as defined by Ritchey. A wicked problem is one which has incomplete, contradictory and often changing requirements [15]. Because of the complex interdependencies of their elements, it is often difficult to recognize that one has achieved even a partial solution. Moreover, while attempting to solve a wicked problem, the partial solution often reveals or creates even more complex problems. The underlying systems are emergent, adaptive systems meaning that the system dynamically changes its behavior and its ability to adapt to new situations. Modeling these types of systems must continually evolve in order to support the decision-maker's wide area situation awareness.

V. Conclusion and Future Work

Big data is the "new" business and social science frontier. The amount of information and knowledge that can be extracted from the digital universe is continuing to expand as users come up with new ways to manage and process data. Moreover, it has become clear that "more data is not just more data", but that "more data is different".

"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. If Facebook and Twitter are producing, collectively, around 50 gigabytes of data per day, and tripling every year, within a few years (perhaps 3-5) we are indeed facing the challenge of "big data becoming really big data".

We as a global society are evolving from a data-centric to a knowledge-centric community. Our knowledge is widely distributed and equally widely accessible. One program that is addressing this problem is The Federal Semantic Interoperability Community of Practice (SICoP) which supports an evolving model: Citizen Centric Government Systems That Know; Advanced Analytics Systems That Learn; and Smart Operations Systems That Reason. These systems will require big data. The data will not be stored in one or even a few locations; it will not be just one or even a few types and formats; it will not be amenable to analysis by just one or a few analytics; and there will not be thus, it is an exemplar of some of the issues we have addressed in this paper. Solving the issues and challenges addressed in this paper will require a concerted research effort - one which we expect to evolve over the next several years.

This paper initiates a collaborative research effort to begin examining big data issues and challenges. We identified some of the major issues in big data storage, management, and processing. We also identified some of the major challenges – going forward – that we believe must be addressed within the next decade and which will establish a framework for our Big Data minitrack in future HICSS sessions.

Our future research will concentrate on developing a more complete understanding of the issues associated with big data, and those factors that may contribute to a need for a big data analysis and design methodology. We will begin to explore solutions to some of the issues that we have raised in this paper through our collaborative research effort.

References

[1] David Hand, Heikki Mannila & Padhraic Smyth – "Principles of Data Mining" Eastern Economy Edition., pp. 1.

[2] Will Garside and Brain Cox – " Big Data storage for Dummies "- EMC^{2,} pp -4

[3] Stephen Kaisler, Frank Armour, J. Alberto Espinosa , William Money - "Big Data: Issues and Challenges Moving Forward" paper published - 2013 46th Hawaii International Conference on System Sciences (http://www.computer.org/csdl/proceedings/hicss/2013/4892/00/4892a995.pdf) [published]

[4] Gartner – "Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s" Forbes.com , By Svetlana Sicular

Gartner , Inc. March 2013

[5] Tim McGuire, James Manyika, and Michael Chui - "Why Big Data is the new competitive advantage " Ivey Business Journal – July-August 2012

[6] META Group. "3D Data Management: Controlling Data Volume, Velocity, and Variety." February 2001.

[7] ERP Know More - "Big Data technologies – a quick way to get through more " - Ramco.com Blog , 2011

[8] American Institute of Physics (AIP). 2010. College Park, MD, (http://www.aip.org/fyi/2010/)

[9] Felten, E. 2010. "Needle in a Haystack Problems", https://freedom-to-tinker.com/blog/felten/needle-haystack-problems/

[10] Kaisler, S. 2012. "Advanced Analytics", CATALYST Technical Report, i_SW Corporation, Arlington, VA

[11] Kaisler, S., W. Money, and S. J. Cohen. 2012. "A Decision Framework for Cloud Computing", 45th Hawaii International Conference on System Sciences,Grand Wailea, Maui, HI, Jan 4-7, 2012

[12] Kang, U. 2012. "Mining Tera-scale Graphs with MapReduce: Theory, Engineering, and Discoveries", PhD. Thesis, Computer Science, Carnegie-Mellon University, Pittsburgh, PA

[13] Mervis, J. 2012. "Agencies Rally to Tackle Big Data", Science, 336(4):22, June 6, 2012

[14] Ritchey, T. 2005. "Wicked Problems: Structuring Social Messes with Morphological Analysis", Swedish Morphological Society, http://www.swemorph.com/wp.html

[15] Rittel, H. and M. Webber. 1973. "Dilemmas in a General theory of Planning", in Policy Sciences, Vol. 4, Elsevier Scientific, Amsterdam, the Netherlands, pp. 155-169

[16] Taleb, N. 2010. The Black Swan: The Impact of the Highly Improbable, Random House, New York, NY