International Journal of Computer Systems (ISSN: 2394-1065), Volume 01– Issue 02, November, 2014
Available at http://www.ijcsonline.com/

Big Data Challenges: A Survey

Kajal Garg, Sonal Somani

Department of Computer Science,
Rajasthan Technical University,
5-Jha-22, Jawahar Nagar, Jaipur, Rajasthan, India - 302004

Abstract

The era of Big Data has begun. Big data is often compared to a tsunami. Today's five billion cell phone users and nearly one billion Facebook and Skype users generate unprecedented volumes of data, and they represent only part of the global online population. We are moving quickly toward the "Internet of things," in which vast numbers of networked sensors in businesses, homes, cars, and public places drive data generation to almost unimaginable levels. The advent of the age of big data poses opportunities and challenges for businesses and previously unavailable forms of data can now be saved, retrieved, and processed. However, this emergent field is challenged by inadequate attention to methodological and conceptual issues. A review of key methodological and conceptual challenges including lack of clarity with regard to sampling, universe and representativeness, the missing ecology of information flows etc is done. But will Big Data cause more problems than it solves? Or will it improve social, political, and economic intelligence? Will its rise facilitate things like "now-casting" (real-time "forecasting" of events)? Or will it engender false confidence in our predictive powers? Can we say that analysis of Big Data will be misused by powerful people and institutions? Or will it be used to improve social protection ? Such analysis is covered to some extent in this paper with the focus upon the major concerns and challenges rose by "Big Data for Development", and some ways are also suggested in which few of these challenges can be addressed.

I. INTRODUCTION

Even twenty or thirty years ago, data on economic activity was relatively scarce. In just a short period of time, this has changed dramatically. One reason is the growth of the internet. Practically everything on the internet is recorded. When you search on Google or Bing, your queries and subsequent clicks are recorded. When you shop on Amazon or eBay, not only every purchase, but every click is captured and logged. When you read a newspaper online, watch videos, or track your personal finances, your behavior is recorded. The recording of individual behavior does not stop with the internet: text messaging, cell phones and geo-locations, scanner data, employment records, and electronic health records are all part of the data footprint that we now leave behind us.

The term Big Data is used almost anywhere these days; from news articles to professional magazines, from tweets to YouTube videos and blog discussions. The term coined by Roger Magoulas from O'Reilly media in 2005, refers to a wide range of large data sets almost impossible to manage and process using traditional data management tools – due to their size, but also their complexity.

The amount of data generated is expected to double every two years, from 2500 exabytes in 2012 to 40,000 exabytes in 2020. Security and privacy issues are magnified by the volume, variety, and velocity of Big Data. Large-scale cloud infrastructures, diversity of data sources and formats, the streaming nature of data acquisition and high volume inter-cloud migration all create unique security vulnerabilities. Many significant challenges extend beyond these, for example, Big Data has to be managed in context, which may be noisy, heterogeneous and not include an upfront model.

The purpose of this paper is to highlight the challenges of Big Data. To do so, the working group utilized a process to arrive at the top challenges in Big Data:

1) Published solutions were studied.

2) Discussion of both, what has already been done and what challenges remain as we seek to exploit Big Data, was done.

In this paper, a brief description of each challenge and possible solution of few of them is provided.

II. SOME CONSIDERATIONS

A. The Missing Denominator: We know who clicked but we don't know who saw or who could

One of the biggest methodological dangers of big data analyses is insufficient understanding of the underlying sample and denominators. It's often not enough to understand how many people have ―liked" a Facebook status update, clicked on a link, or ―re-tweeted" a message without having a sense of how many people saw and chose to –or not to– take that option. That kind of normalization is rarely done, or may even be actively decided against because the results start appearing more complex or more trivial [1].

While the precise denominator is often not calculable, in many cases, it may be possible to calculate preliminary estimates. For example, Facebook has allowed its researchers to disclose information about potential audiences for status updates –for example, the mean and

median fraction of a user's friends that see the post is about 34-35% of the universe of friends, though the distribution of the variable seems to have a large spread [2].

Steps in this direction are likely to be complex and difficult, but without such efforts, our ability to interpret raw numbers will remain limited as we won't even have an estimate of the denominators. Coupled with the lack of representativeness and (often) lack of random sampling in the sources of big data –as discussed above—the denominator and sampling issues raise many troubling questions about both the representativeness and fairness of generalizing from many available big data sets.

B. Missing the Ecology for the Platform

Most existing big data analyses of social media are confined to a single platform, often Twitter, as discussed above. However, most of the topics of interest to such studies, such as influence or information flow can rarely be confined to the Internet, let alone to a single platform. Understandable difficulty in obtaining high-quality multi-platform data does not mean that we can treat a single platform as a closed and insular system, as if human information flows were all gases in a chamber. They are not.

The emergent media ecology is an integrated mix of old and new media rather than strictly segregated by platform or even device. Many —viral" videos take off on social media only after being featured on broadcast media and this step itself is often preceded by being highlighted on intermediary sites such as Reddit or Buzzfeed. This example shows the object of analysis should be this integrated ecology, that rather than stay at —which site" or —which link" [8].

These challenges do not mean that nothing valuable can be used from single-platform analyses. However, all such analyses must take into account that they are not examining a closed system. More research is needed to understand the actual multi-platform connectivity patterns. It's possible that the solution to this —big data" limitation may not be solvable by —big data" alone. Sometimes, the way to study people is… to study people.

C. Data dredging

As suggested by Nathan eagle, an adjunct assistant professor at HSPH, one of the most prominent is data dredging, which involves searching for patterns in huge datasets. A traditional social-science study might assert that the results are significant with 95 percent confidence. That means in one out of 20 instances, when dredging for results, we might get results that are statistically significant purely by chance. So, although this is true for any statistical finding, the enormous number of potential correlations in very large datasets substantially magnifies the risk of finding spurious correlations. Eagle agrees that —you don't get good scientific output from throwing everything against the wall and seeing what sticks" [10]. No matter how much data exists, researchers still need to ask the right questions to create a hypothesis, design a test, and use the data to determine whether that hypothesis is true. Also, there aren't enough people comfortable dealing with petabytes of data. The move is from a big data problem to a really big data problem.

III. CHALLENGES OF BIG DATA

While big data can yield extremely useful information, it also presents new challenges with respect to how much data to store, how much this will cost, whether the data will be secure, and how long it must be maintained and so on.

So, although —Big data" has arrived, but big insights have not. The challenge now is to solve new problems and gain new answers.

Some of the major challenges are discussed below:

A. Data storage: Is big file data really a big deal?

When a storage system hits a point where it simply won't scale any larger, or it becomes so bottlenecked that access times and throughput are unacceptable, it's usually time to migrate to a new system. But with a big file data application, a migration may be nearly impossible. Few businesses or organizations have enough downtime to move petabytes of data, especially when new data is still flowing into the system. Like the proverbial "immovable object," large file archives can get so big that they become unmanageable within a traditional infrastructure.

Similar to a foundation of a building, once these infrastructures are set up and put into use, it's often too late to change. For this reason, big file data storage infrastructures must be designed for maximum flexibility with the ability to be upgraded as non-disruptively as possible with their data in place.

B. Maintenance issues: What happens over the long haul?

Keeping data for a long time isn't unique to big file data applications, but when each file adds another 10 GB (or 100 GB or more), retention quickly becomes an issue. Data retention isn't related to file size per se, but many of the files that people and companies want to keep are image based. Digital content such as video and audio are good examples, as are digital snapshots (Shutterfly maintains nearly 100 PB of photo data) and video surveillance files.

Long-term retention has historically been driven by regulatory compliance, but now data is just as likely to be kept for its possible reuse or for security. A good example is surveillance videos. Historically archived for legal reasons, these files are now being used to help analyze customers' shopping behavior. Storing this type of data for extended, often open-ended, timeframes creates operational cost issues. Maintaining the disk space to keep tens or hundreds of terabytes for years isn't trivial, but it's nothing compared to the power and floor space required to support petabytes of data on even the lowest cost disk available.

Human consumption: There are also institutional and technical challenges when data is stored in places and ways that make it difficult to be accessed, transferred, etc. Like in a lot of big data analytics apps, computers perform the analysis, so data is often stored in the same data center that houses the database servers or in the same servers themselves, as with Hadoop clusters. But in big file data use cases, the data is often analyzed by people -- and people don't live in data centers. When the processing engine wants to consume data on a tablet from home or a smartphone on the road, the storage infrastructure must deliver that data appropriately.

Most of the files are consumed in order, so they need to be streamed (often through a low-bandwidth connection), and can't be chopped up and reassembled upon delivery. To support that kind of consumption pattern, many big file data repositories need a random-access storage tier that can quickly send enough content to get the streaming process started, and then buffer the rest of the file. But that disk storage tier must be very large and very scalable, since it has to contain the first portions of the files in a very large archive and keep up when that archive grows. Maintenance of such large datasets becomes a challenge in the long haul.

C. Big cost of big data: Is it?

As traditional database approaches don't scale or write data fast enough to keep up with the speed of creation. Additionally, purpose-designed data warehouses are great at handling structured data, but there's a high cost for the hardware to scale out as volumes grow.

A key enabler for Big Data is the low-cost scalability of Hadoop. For example, a petabyte Hadoop cluster will require between 125 and 250 nodes which costs ~$1 million. The cost of a supported Hadoop distribution will have similar annual costs (~$4,000 per node), which is a small fraction of an enterprise data warehouse ($10-$100s of millions). On initial evaluation, Big Data on Hadoop appears to be a great deal. Innovative enterprises have Hadoop today – the question is how will they leverage it and at what pace will it become mission-critical?

However, the real cost lies in the operation and the overall management or integration of Big Data within the existing ecosystem, and as Big Data environments scale, such as at Yahoo, managing 200 petabytes across 50,000 nodes require that more be added to deliver additional storage capacity [9]. To be able to determine whether the cost generated from value of Big Data outweighs the operational cost factors in the long run remains a challenge.

D. Data privacy: Ethical challenges to be dealt

Big data also presents new ethical challenges. Corporations are using big data to learn more about their workforce, increase productivity, and introduce revolutionary business processes. However, these improvements come at a cost: tracking employees' every move and continuously measuring their performance against industry benchmarks introduces a level of oversight that can quash the human spirit. Such monitoring might be in the best interest of a corporation but is not always in the best interest of the people who make up that corporation.

In addition, as big multimedia datasets become commonplace, the boundaries between public and private space will blur. However, unlike surveillance cameras, smartphones and wearable devices afford no privacy protection to innocent bystanders who are captured in a video at the right place at the wrong time. For example, in the wake of the recent Boston bombings, images of several people photographed at the scene were mistakenly identified as suspects on social media sites.

In fact, one of the major challenges of big data is preserving individual privacy. As we go about our everyday lives, we leave behind digital footprints that, when combined, could denote unique aspects about ourselves that would otherwise go unnoticed, akin to digital DNA. Examples include our use of language and punctuation in blog and forum posts, the places we frequent—do we spend our Sunday morning's outdoors, indoors online, etc.? Something as innocuous as when and how we use energy in our homes reveals many details about us.

An important direction is to think security for information sharing in Big Data use cases. Many online services today require us to share private information (think of Facebook applications), but beyond record-level access control we do not understand what it means to share data, how the shared data can be linked, and how to give users fine-grained control over this sharing [3].

Big data analytics will draw on aspects of our home, work, and social lives to make assumptions beyond typical "market segmentations" and delve deep into ontological questions such as, "Who are you?" This has metaphysical implications. For example, people will consciously alter their online activity, and will modify their behavior in surveilled spaces, to protect their privacy [4]. Big data will change how we live in both small and large ways. Are we on a trajectory toward an uberveillance society?

So, managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of big data.

E. Data Security: Safeguarding data

Today, Big Data is cheaply and easily accessible to organizations large and small through public cloud infrastructure. Software infrastructures such as Hadoop enable developers to easily leverage thousands of computing nodes to perform data-parallel computing. Combined with the ability to buy computing power on-demand from public cloud providers, such developments greatly accelerate the adoption of Big Data mining methodologies. As a result, new security challenges have arisen from the coupling of Big Data with public cloud environments characterized by heterogeneous compositions of commodity hardware with commodity operating systems, and commodity software infrastructures for storing and computing on data.

As Big Data expands through streaming cloud technology, traditional security mechanisms tailored to securing small-scale, static data on firewalled and semi-isolated networks are inadequate. For example, analytics for anomaly detection would generate too many outliers. To be able to secure the Big Data is a big challenge.

In order to secure the infrastructure of Big Data systems, the distributed computations and data stores must be secured. To secure the data itself, information dissemination must be privacy-preserving, and sensitive data must be protected through the use of cryptography and granular access control.

F. Timelineness: Makes Big deal out of big data?

The flip side of size is speed. The larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually

meant when one speaks of Velocity in the context of Big Data. Rather, there is an acquisition rate challenge and a timeliness challenge.

There are many situations in which the result of the analysis is required immediately. For example, if a fraudulent credit card transaction is suspected, it should ideally be flagged before the transaction is completed – potentially preventing the transaction from taking place at all. Obviously, a full analysis of a user's purchase history is not likely to be feasible in real- time. Rather, we need to develop partial results in advance so that a small amount of incremental computation with new data can be used to arrive at a quick determination. Given a large data set, it is often necessary to find elements in it that meet a specified criterion. In the course of data analysis, this sort of search is likely to occur repeatedly. Index structures are created in advance to permit finding qualifying elements quickly [11]. With new analyses desired using Big Data, there are new types of criteria specified, and a need to devise new index structures to support such criteria. For example, consider a traffic management system with information regarding thousands of vehicles and local hot spots on roadways. The system may need to predict potential congestion points along a route chosen by a user, and suggest alternatives. Doing so requires evaluating multiple spatial proximity queries working with the trajectories of moving objects. New index structures are required to support such queries.

Designing such structures becomes particularly challenging when the data volume is growing rapidly and the queries have tight response time limits.

G. Heterogeneity and incompleteness: Is structuring of data becoming a necessity?

When humans consume information, a great deal of heterogeneity is comfortably tolerated. In fact, the nuance and richness of natural language can provide valuable depth. However, machine analysis algorithms expect homogeneous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step in (or prior to) data analysis. Consider, for example, a patient who has multiple medical procedures at a hospital. We could create one record per medical procedure or laboratory test, one record for the entire hospital stay, or one record for all lifetime hospital interactions of this patient. With anything other than the first design, the number of medical procedures and lab tests per record would be different for each patient. The three design choices listed have successively less structure and, conversely, successively greater variety. Greater structure is likely to be required by many data analysis systems. However, the less structured design is likely to be more effective for many purposes – for example questions relating to disease progression over time will require an expensive join operation with the first two designs, but can be avoided with the latter. However, computer systems work most efficiently if they can store multiple items that are all identical in size and structure. Efficient representation, access, and analysis of semi-structured data require further work.

Consider an electronic health record database design that has fields for birth date, occupation, and blood type for each patient. What do we do if one or more of these pieces of information is not provided by a patient? Obviously, the health record is still placed in the database, but with the corresponding attribute values being set to NULL. A data analysis that looks to classify patients by, say, occupation, must take into account patients for which this information is not known. Worse, these patients with unknown occupations can be ignored in the analysis only if we have reason to believe that they are otherwise statistically similar to the patients with known occupation for the analysis performed. For example, if unemployed patients are more likely to hide their employment status, analysis results may be skewed in that it considers a more employed population mix than exists, and hence potentially one that has differences in occupation-related health-profiles [5].

Even after data cleaning and error correction, some incompleteness and some errors in data are likely to remain. This incompleteness and these errors must be managed during data analysis [6]. Doing this correctly is a challenge.

H. Data access and sharing

Although much of the publicly available online data (data from the open web) has potential value for development, there is a great deal more valuable data that is closely held by corporations and is not accessible. One challenge is the reluctance of private companies and other institutions to share data about their clients and users, as well as about their own operations.

Many of the novel data belongs to private companies. Accessing private company data creates several issues for researchers. In most enterprises, the data generated by a functional area ends up being the property of that group. This leads to two problems. First, it's difficult to get a ―complete" view of the data. Consider all the silos and systems that hold data: CRM, ticketing, bug tracking, fulfillment, etc. Getting all the relevant systems to even talk to each other is a huge challenge. Second, there's significant cultural dissonance within organizations [7]. Typically, each group controlling a data silo ends up caring more about their power and place in a department rather than the success of the organization as a whole. The owners of huge datasets are very nervous about sharing even anonymized, population-level information like the call records a company uses. For the companies that hold it, there is a lot of downside to making this data open to researchers. But will sharing the data not pose security and privacy challenges? There is possible scope of research in this area to undertake the development versus challenges issue.

Ways need to be figured out to mitigate that concern and craft data-usage policies in ways that make these large organizations more comfortable with sharing these data, which ultimately could improve the lives of the millions of people who are generating it—and the societies in which they are living." Organizations need to pool their data to find the answers to and get a complete view of their data.

IV. CONCLUSION

The true value of data comes from being able to contextualize and understand it, in order to deliver insights. Big file data creates several challenges. Maintaining a storage system that can hold a mountain of data and still

provide decent throughput and access performance is all but impossible using traditional infrastructures. These systems must be extremely flexible, able to support multiple types of random-access storage and tape, and allow users to upgrade and modify them while leaving the data in place. They also need a highly scalable, secure, manageable and semi-structured format.

We can live with many of these uncertainties for now with the hope that the benefits of big data will outweigh the harms, but we shouldn't blind ourselves to the possible irreversibility of changes—whether good or bad—to society.

Further research can be carried out in order to overcome these challenges.

REFERENCES

[1] Cha, Meeyoung, H. Haddadi, F. Benevenuto, and P. K. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy", ICWSM 10, 2010, pp. 10-17.

[2] Bernstein, Michael, E. Bakshy, M. Burke, and B. Karrer, "Quantifying the invisible audience in social networks.", In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp. 21-30 , 2013.

[3] Pulse, UN Global. "Big Data for Development: Opportunities & Challenges." White paper, Global pulse, New York, USA, May 2012.

[4] Aday, Sean, H. Farrell, M. Lynch, J. Sides, and D. Freelon, "New Media and Conflict after the Arab Spring", Washington DC: United States Institute of Peace, 2012.

[5] G. Halevi, and H. Moed, "The evolution of big data as a research and scientific topic: overview of the literature.", Research Trends, Special Issue on Big Data, 30, pp. 3-6, 2012.

[6] P. D. Allison, ed. Missing Data. No. 136. SAGE, 2002.

[7] Boyd, Danah, and Kate Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." Information, Communication & Society 15, no. 5, pp.662-679, 2012.

[8] Mitchell, Amy, and P. Hitlin. "Twitter reaction to events often at odds with overall public opinion." Pew Research Center 4 , 2013.

[9] Zikopoulos, Paul, and Chris Eaton, ―Understanding big data: Analytics for enterprise class hadoop and streaming data", McGraw-Hill Osborne Media, 2011.

[10] Lazer, David, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis et al, "Life in the network: the coming age of computational social science." Science (New York, NY) 323, no. 5915 (2009): 721.

[11] Ferguson, Mike, "Architecting A Big Data Platform for Analytics." White paper, IBM , 2012.