Article

Big Data: A Study of Its Issues and Challenges

Mudasir Manzoor Kirman[1], Syed Mohsin Saif[2], Ahmad Talha Siddiqui[3]

[1] *Assistant professor SKUAST-K, Srinagar, JK India*

[2] *Research Scholar, Department of CS & IT, School of Computer Science and Information Technology, MANUU, Hyderabad, India*

[2] *Assistant professor, Department of CS & IT, School of Computer Science and Information Technology, MANUU, Hyderabad, India*

*Author to whom correspondence should be addressed; E-Mail: sms.manuu@gmail.com

Abstract: Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. Big data is not merely data, rather it has become a complete subject, which involves various tools, techniques and frameworks. Big Data refers to the data volumes in the range of 10-10,000 tera bytes and beyond. Such volumes exceed the capacity of current storage systems and processing system. One cannot claim that technology is 100 percent good, beneficial and effective to support the mammoth size of data. Every emerging technology has both advantages and disadvantages associated with it and as time passes it gets reframed to get better out of best. It was observed that all the developmental prospects have environmental, social, and human consequences that go far beyond the immediate purposes of the technical devices and practices themselves. The era of Big Data has begun. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and many others are demanding for access to the massive quantities of information produced by "internet of things". In this paper a review study is made to analyze the issues and challenges are inline with adopting this technology and benefits that are end product of using big data in special reference to Social media.

Keywords: bigdata, social media, big data security.

1. Introduction

Big Data introduces two new popular types of social networks derived from data traces: 'articulated networks' and 'behavioral networks.' Articulated networks are those that result from people specifying their contacts through a mediating technology. There are three common reasons in which people articulate their connections: to have a list of contacts for personal use; to publicly display their connections to others; and to filter content on social media. These articulated networks take the form of email or cell phone address books, instant messaging buddy lists, 'Friends' lists on social network sites, and 'Follower' lists on other social media genres. The motivations that people have for adding someone to each of these lists varies widely, but the result is that these lists can include friends, colleagues, acquaintances, celebrities, friends-of-friends, public figures, and interesting strangers.

2. Big Data in Search Engines & Social Media

We looked at how search engines process a tremendous amount of data including search queries, location-specific searches and scans for new content. Google's search algorithm is a Big Data company that processes over 20 petabytes of data each year. Beyond the search engines, social networks also process a massive dataset, which is an important consideration due to the positive correlation of social signals on search rankings[20].

There are currently over 1 billion Face book users, nearly 300 million Twitter users and 156 million public blogs, 350 Google+. Face book processes 2.5 billion pieces of content, 300 million photos, 2.7 billion like actions a day and over 500 terabytes of data each day. Face book in average after every 20 minutes share 1 million Links, 2 million Friend requests and 3 million Message disposals. Twitter, our client and partner, processes over 400 million tweets daily[21].

When we look at video, YouTube users upload 48 hours of new video every minute. What's more, with over 5 billion mobile phone devices in use, the convergence of social, mobile and local data creates massive amounts of data to process and make sense of.

3. Data Integrity, Security, Reliability & Scalability in the Age of Big Data

Specifically, there are four key factors that determine the return from Big Data:

i.   Is the Big Data accurate?

ii.   Is our Big Data secure?

iii. Is our Big Data available at all times?

iv.   Does our Big Data scale?

Collating and creating Big, *Valuable* Data is a very expensive process and requires lots of investment and massive engineering resources to create a rigorous and high-quality set of data streams. Currently, 75% of Fortune 500 companies use cloud-based solutions, and the IDC predicts that 80% of new commercial enterprise apps will be deployed on cloud platforms [22].

4. Ensure Your Data Is As Accurate As Possible

As a search marketer, you are among the most data-driven people on this planet. You make important decisions around keywords, pages, content, link building and social media activity based on the data you have on hand.

Before gaining insight and building a plan of action based on Big Data, it's important to know that you can trust this data to make the right decisions. While this might seem like a daunting exercise, there are a few fairly achievable steps you can take.

5. Ensure Your Data Is Secure

You have, on your hands, unprecedented amounts of data on users and their behavior. You also have precious marketing data that has a direct impact on your business results.

With great amounts of knowledge comes even greater responsibility to guarantee the security of this data. Remember, you and your technology provider together are expected to be the trusted guardians of this data. In this extended geography, you have a legal obligation to safeguard this data.

6. Ensure Your Data Scales with User Growth

This is the part that deals with the *Big* in Big Data and how zeta bytes of data already exist and more data is being generated at an astounding pace by billions of Internet users and transactions every day. For you to understand these users and transactions, your technology should have the ability to process such huge volumes of data across channels and keep up with the growth of the Internet.

Scale should matter even if the organizations are not big then. Think about this analogy – even if you are searching for a simple recipe on Google, Google has to parse through huge volumes of data to serve the right results.

Similarly, your technology should be able to track billions of keywords and pages, large volumes of location-specific data and social signals to give you the right analytics. Be sure the technology you rely on is made for scale.

7. Data Ownership

As big data has now emerged as biggest driver and production factor in this corporate world wherein countless data bytes are flushed across the domains. If one can have look towards the data production and dissemination in social media houses and the extent of the data being surfaced across and stored. The data being generated across these platforms might have varying degrees of security levels. Who cares the data security and who claims the ownership are the challenging questions to be addressed. The most popular rather inevitable parameters like data ownership, right to generate, use to monetise big data are yet be established. Other legal issues like data privacy, database rights, IP rights antitrust laws as well as the basic civil rights.

Data ownership presents a critical and ongoing challenge, particularly in the social media arena. While petabytes of social media data reside on the because of residency. Certainly, the "owners" of the pages or accounts believe they own the data. Kaisler, Money and Cohen [12] addressed this issue with respect to cloud computing as well as other legal aspects. There is still no conclusive outcome of the legal discussion and legislation around allocation of ownership of anonymous data or at least of the right to use and to exclude others from using such data. Arguments indicate, though, that the entity or individual controlling the production of anonymous data should be entitled to use them.The addition of unverified data: compromises the fidelity of the dataset; may introduce non-relevant entities; and may lead to erroneous linkages among entities. As a result, the accuracy of conclusions drawn from processing this mixed data varies widely.

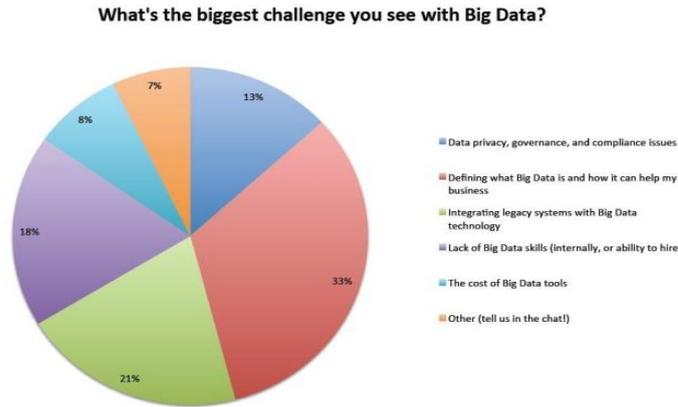**Fig. 1:** Big data challenges (Source: TDWI Predictive analytics study, 2015)

What's the biggest challenge you see with Big Data?

Fig. 2: What is the biggest challenge (Source: TDWI Predictive analytics study, 2015)

Table 1: Challenges that are to addressed while holding big data and like application processing

	Search Engine	**Social Media (i.e Twitter, Face Book)**	**Big Data Analytics Professionals**
Challenge No-1	Being able to handle the large volume, and variety of big data	Being able to handle the large volume, velocity and variety of big data	Being able to handle the large volume, velocity and variety of big data
Challenge No-2	Finding the optimal way to organize big data activities	Determining what data both structured and unstructured to use for different business decisions	Finding and hiring data scientists who can manage large amounts of structured and unstructured data and create insights.
Challenge No-3	Getting business units to share information across organizational silos.	Getting business units to share information across organizational silos.	Getting business units to share information across organizational silos.
Challenge No-4	Determining what to do with the insights that are created from big data	Getting top management in the company to approve investments in big data and its related investment	Determining what to do with the insights that are created from big data
Challenge No-5	Building high levels of trust between the data scientists who present insights on big data and the functional managers	Building high levels of trust between the data scientists who present insights on big data and the functional managers	Understanding where big data investments in the company should be focused.

6. Conclusion

There is no doubt that Big data has revolutionised both industrial and corporate data warehouse management systems. It has provided more challenging and more interesting results which has made other domains to think about the adaption of Big data has now become like lingua franca in the premises of information and communication technology. Whether being online on "internet of things" or offline. Big Data is the data or the information owned by corporate organisations, public sectors organisations, obtained and processed through new techniques to produce value in the best way possible. Apache Hadoop is one such technology, and it is generally the software most commonly associated with Big Data. Apache calls it "a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. 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. There might emerge different challenges to manage, secure, and ensure persistence in data integrity, accessibility and other challenges. The need of the hour is to spend much more research in categorising the levels of security and data integrity.

References

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

[2] Ayres, I. 2007. *Supercrunchers*, Bantam Books, New York, NY.

[3] Boyd, D. and K. Craford. 2011. Six Provocations for Big Data, Oxford Internet Institute's A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society.

[4] The Economist. 2010. *Data, Data Everywhere*, (online edition, February 28) http://www.economist.com/node/15557443.

[5] Felten, E. 2010. *Needle in a Haystack Problems*, https://freedom-to-tinker.com/blog/felten/needle-haystackproblems/.

[6] Fox, B. 2011. *Leveraging Big Data for Big Impact, Health Management Technology*, http://www.healthmgttech.com/

[7] Freeman, K. 2011. http://en.wikipedia.org/wiki/File:Kencf0618Facebook Network.jpg.

[8] Gantz, J. and E. Reinsel. 2011. *Extracting Value from Chaos*, IDC's Digital Universe Study, sponsored by EMC.

[9] Jacobs, A. 2009. *Pathologies of Big Data*, Communications of the ACM, 52(8):36-44.

[10] JASON. 2008. *Data Analysis Challenges*, The Mitre Corporation, McLean, VA, JSR-08-142.

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

[12] Kaisler, S., Money, W. 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.

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

[14] Mervis, J. 2012. Agencies Rally to Tackle Big Data, *Science*, 336(4): 22.

[15] Popp, R., S. Kaisler, et al. 2006. *Assessing Nation-State Fragility and Instability*, IEEE Aerospace Conference, 2006, Big Sky, MT.

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

[17] Rittel, H. and M. Webber. 1973. Dilemmas in a General theory of Planning, *Policy Sciences*, 4: 155-169.

[18] Stonebraker, M. and J. Hong. 2012. Researchers' Big Data Crisis; Understanding Design and Functionality, *Communications of the ACM*, 55(2):10-11.

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

[20] Jim Yu, 2013, *The Importance of Big Data, Integrity & Security In Enterprise SEO*, http://searchengineland.com/the-importance-of-big-data-integrity-and-security-enterprise-seo-143066.

[21] http://www.statisticbrain.com/facebook-statistics