RESEARCH ARTICLE OPEN ACCESS

A Report on Big data Analysis in Development Sector

V.Satish Kumar [1], D.Wasiha Tasneem [2], Shaik Neeha [3]
Department of Computer Science and Engineering
Dr.K.V.Subba Reddy College of Engineering for Women
Kurnool, Andhra Pradesh - INDIA

ABSTRACT
Big data means really a big data; it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it. Systems, sensors and mobile devices transmit it. To extract meaningful value from big data, you need optimal processing power, analytics capabilities and skills. Big data refers to the dynamic, large and disparate volumes of data being created by people, tools and machines; it requires new, innovative and scalable technology to collect, host and analytically process the vast amount of data gathered in order to derive real-time business insights that relate to consumers, risk, profit, performance, productivity management and enhanced shareholder value. This paper presents the analysis of big data in development sector and also discussed about the issues that capture the potential of big data.

Keywords :— Big data, data analysis, potential, data sets, security.

I. INTRODUCTION

Organizations are increasingly generating large volumes of data as result of instrumented business processes, monitoring of user activity, web site tracking, sensors, finance, accounting, among other reasons. With the advent of social network Web sites, users create records of their lives by daily posting details of activities they perform, events they attend, places they visit, pictures they take, and things they enjoy and want. This data deluge is often referred to as Big Data a term that conveys the challenges it poses on existing infrastructure with respect to storage, management, interoperability, governance, and analysis of the data.

In today's competitive market, being able to explore data to understand customer behaviour, segment customer base, offer customized services, and gain insights from data provided by multiple sources is key to competitive advantage [1]. Although decision makers would like to base their decisions and actions on insights gained from this data, making sense of data, extracting non obvious patterns, and using these patterns to predict future behaviour are not new topics. Knowledge Discovery in Data (KDD) aims to extract non obvious information using careful and detailed analysis and interpretation. Data mining, more specifically, aims to discover previously unknown interrelations among apparently unrelated attributes of data sets by applying methods from several areas including machine learning, database systems, and statistics. Analytics comprises techniques of KDD, data mining, text mining, statistical and quantitative analysis, explanatory and predictive models, and advanced and interactive visualization to drive decisions and actions.

Big data analysis is most important because the data is continuously changing based on interval of time to store big data most of companies are using cloud setup [2].

Characteristics of Big data:

Big data includes information garnered from social media, data from internet-enabled devices (including smartphones and tablets), machine data, video and voice recordings, and the continued preservation and logging of structured and unstructured data [3]. It is typically characterized by the four "V's":

A. ***Volume:*** Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would've been a problem – but new technologies (such as Hadoop) have eased the burden.

B. ***Velocity:*** Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.

C. ***Variety:*** Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions.

D. ***Veracity:*** Data sources (even in same domain) are of different qualities with differences accuracy, coverage and timelineness.

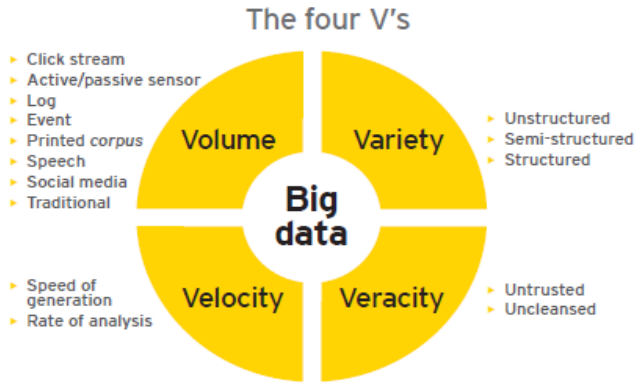

Fig.1 Properties of Big data

II. BIG DATA FOCUS FOR THE DEVELOPMENT SECTOR

Big data shows potential to advance development work in a variety of ways. In the first section above, several examples were provided which highlighted the ways in which big data could be used as a proxy for conventional official statistics, thereby enhancing institutional awareness of the conditions of a population; to better organize governmental processes thereby delivering more effective services; or to enhance understanding of the drivers of health epidemics, thereby guiding policy decisions [4]. Any point in the framework can be used to stimulate the imagination on the horizon of possibilities for big data in development. Moreover, this section describes several examples of data sets utilized by medium as well as by purpose and report information on the first World Bank attempts to leverage big data to address development challenges in Central American countries. By cross referencing primary media with the primary purpose of the use of big data--awareness, understanding, or forecasting--one can easily see [5] how big data projects can take a variety of configurations depending on the context. Then a summary is presented detailing what institutions and individuals are saying about where big data shows promise for development. Finally, recommendations for next steps in advancing the application of big data for development are provided.

Examples by medium and relevant data set:

A. Mobile | Call Detail Records. Although usage of call detail record (CDR) data for development is still in early phases, applications such as using Digicel's data to track population displacement after the Haiti earthquake and modeling of infectious disease spread show great promise. One study in Afghanistan showed that CDR data could be used to detect impacts from small-scale violence, such as skirmishes and improvised explosive devices, in terms of their impacts on communication methods and patterns of mobility. Another project done by the lead researcher in the Afghanistan study was to capture seasonal and temporary migration, usually overlooked by traditional survey models, permitting a more precise quantification of its prevalence. An ongoing project which builds upon these results aims to measure precisely the extent to which wage disparities in Rwanda, Afghanistan, and Pakistan are arbitrated by migration [5].

B. Satellite | remote sensing images. Usage of satellite data abounds. For example, the United Nations University engaged in a project using satellite rainfall data combined with qualitative data sources and agent-based modeling to understand how rainfall variability affects migration as well as food and livelihood security in South and Southeast Asia, Sub- Saharan Africa and Latin America [6]. In Stockholm, GPS-equipped vehicles provided real-time traffic assessments and, when combined with other data sets such as those pertaining to weather, made traffic predictions. Such analyses inform urban planning and also can increase time and cost savings for drivers [7].

C. Internet | Search Queries. The internet stores a vast amount of information, much of which is unstructured. Search queries present one source of data on the internet. In this vein, Google searches for "unemployment" were, found, for example, to correlate with actual unemployment data. Similar data was used to notice changes in the Swine Flu epidemic before official US Centers for Disease Control and Prevention data sources reflected it [8]. The Bank of England uses search queries related to property, for example, to infer housing price changes [9]. Colombia's Ministry of Finance uses the information generated by Google searches to assess short-term GDP trends in Colombia and publish monthly macroeconomic reports which discuss the results of the model developed [10].

D. Internet | Text. Text analysis is critical for data generated via the internet not only for sentiment analysis (e.g. favorable/unfavorable views on a policy) but also for lexical analysis to understand elements of culture. One group analyzed the concept of honor in the Middle East, for example, and found how it differed by region and changed over time in response to the events of September 11th. Such analysis could inform the appropriate selection of language in, say, diplomacy or educational materials. Further applications in this regard could include, for example, developing a contextual lexicon on financial literacy in order to tailor microlending by region [10]. By combining topic modeling methods--whereby one explores and understands concepts and topics from text--with sentiment analysis, one can gain a richer understanding of unstructured text data [11].

E. Social Media | Tweets. Similar to the example of analyzing search queries above, social media data such as Twitter tweets can be used as an early indicator of an unemployment hike or to evaluate crisis-related stress [12]. Another case utilized tweets to know about a cholera outbreak in Haiti up to two weeks prior to official statistics. Both of these cases demonstrate the ability to reduce reaction time and improve process with which to deal with various crises. Tweets have been used by hedge fund managers to make

better stock exchange decisions and by researchers to predict a film's success at the box office or a person's likelihood to get flu shots.

F. Financial | Credit Card Transactions. Credit card companies have increasingly been using their massive stores of data to enhance their services. In several cases, companies use purchase data to identify unusual behavior in real time and quickly address potential credit card fraud [13]. In other cases, financial institutions have been cited as being able to predict whether someone is dating [14] or even infer the strength of a marriage.

III. TECHNOLOGY MEGATRENDS

Big data is generating an intense amount of attention among businesses, media and even consumers, along with analytics, cloud-based technologies, digital channels and data visualization. These are all part of the current diverse ecosystem created by the technology megatrends. Some even herald the potential transformative power of the current trends as rivaling that of the internet. Yet, as in the early days of the internet, there is uncertainty about just what big data is, its potential benefits and the associated risks.

EY's 2013 Global Information Security Survey results indicate that while adoption and use of big data is not yet widespread, there is growing confidence and familiarity with the technology. Respondents ranked big data technologies as being "around the corner" (i.e., those that have been on organizations' radar for a period of time but may not yet be implemented or widely adopted) as average in terms of level of importance, familiarity and confidence in their capabilities to address related cyber risks [15][16]. Organizations typically view these technologies as offering opportunities to improve their performance and create competitive advantage. This is where familiarity and confidence in capabilities needs to increase today, as the importance of these technologies is likely to grow significantly in the near future.

The term Big Data has become a major theme of the technology media, but it has also increasingly made its way into many compliance, internal audit and fraud risk management-related discussions. In EY's Global Forensic Data Analytics Survey 2014, 72% of respondents believe that emerging big data technologies can play a key role in fraud prevention and detection [17]. Yet only 7% of respondents were aware of any specific big data technologies, and only 2% were actually using them.

Forensic data analytics (FDA) technologies are available to help companies keep pace with increasing data volumes, as well as business and regulatory complexities; examples can include real-time analytical processing engines that make rapid business decisions, such as stopping a potentially improper payment or business transaction, or leveraging anti-fraud/anti-corruption monitoring controls that integrate data visualization, statistical analysis and text mining [18]. Yet despite their availability, many companies have not scaled up their data usage to take advantage of these effective tools, and may be missing important fraud prevention and detection opportunities by not mining larger data sets to more robustly monitor business activities [19].

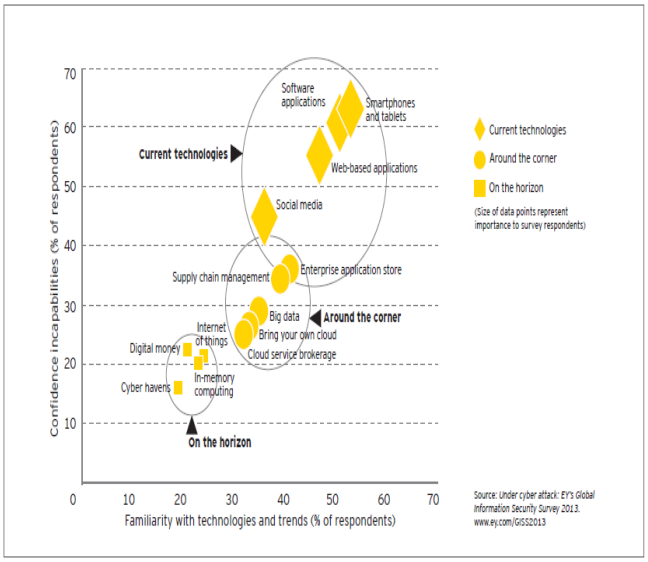

Fig. 2 Emerging technologies and trends

IV. ISSUES THAT CAPTURE THE POTENTIAL OF BIG DATA

A. Data policies. As an ever larger amount of data is digitized and travels across organizational boundaries, there is a set of policy issues that will become increasingly important, including, but not limited to, privacy, security, intellectual property, and liability [20]. Clearly, privacy is an issue whose importance, particularly to consumers, is growing as the value of big data becomes more apparent. Personal data such as health and financial records are often those that can offer the most significant human benefits, such as helping to pinpoint the right medical treatment or the most appropriate financial product. However, consumers also view these categories of data as being the most sensitive. It is clear that individuals and the societies in which they live will have to grapple with trade-offs between privacy and utility.

Another closely related concern is data security, e.g., how to protect competitively sensitive data or other data that should be kept private [21]. Recent examples have demonstrated that data breaches can expose not only personal consumer information and confidential corporate information but even national security secrets.

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.

B. Technology and techniques. To capture value from big data, organizations will have to deploy new technologies (e.g., storage, computing, and analytical software) and techniques (i.e., new types of analyses). The range of technology challenges and the priorities set for tackling them will differ depending on the data maturity of the institution. Legacy systems and incompatible standards and formats too often prevent the integration of data and the more sophisticated analytics that create value from big data. New problems and growing computing power will spur the development of new analytical techniques [21]. There is also a need for ongoing innovation in technologies and techniques that will help individuals and organizations to integrate, analyze, visualize, and consume the growing torrent of big data.

C. Organizational change and talent. Organizational leaders often lack the understanding of the value in big data as well as how to unlock this value. In competitive sectors this may prove to be an Achilles heel for some companies since their established competitors as well as new entrants are likely to leverage big data to compete against them [22]. And, as we have discussed, many organizations do not have the talent in place to derive insights from big data. In addition, many organizations today do not structure workflows and incentives in ways that optimize the use of big data to make better decisions and take more informed action.

D. Access to data. To enable transformative opportunities, companies will increasingly need to integrate information from multiple data sources. In some cases, organizations will be able to purchase access to the data [23]. In other cases, however, gaining access to third-party data is often not straightforward. The sources of third-party data might not have considered sharing it. Sometimes, economic incentives are not aligned to encourage stakeholders to share data [24]. A stakeholder that holds a certain dataset might consider it to be the source of a key competitive advantage and thus would be reluctant to share it with other stakeholders. Other stakeholders must find ways to offer compelling value propositions to holders of valuable data.

E. Industry structure. Sectors with a relative lack of competitive intensity and performance transparency, along with industries where profit pools are highly concentrated, are likely to be slow to fully leverage the benefits of big data. For example, in the public sector, there tends to be a lack of competitive pressure that limits efficiency and productivity; as a result, the sector faces more difficult barriers than other sectors in the way of capturing the potential value from using big data [25]. US health care is another example of how the structure of an industry impacts on how easy it will be to extract value from big data. This is a sector that not only has a lack of performance transparency into cost and quality but also an industry structure in which payors will gain (from fewer payouts for unnecessary treatment) from the use of clinical data [26][27]. However, the gains accruing to payors will be at the expense of the providers (fewer medical activities to charge for) from whom the payors would have to obtain the clinical data. These examples suggest, organization leaders and policy makers will have to consider how industry structures could evolve in a big data world if they are to determine how to optimize value creation at the level of individual firms, sectors, and economies as a whole.

V. CONCLUSION

The main objective of this paper is to describe the analysis of big data in development sector. The paper first defined what is meant by big data to combine the data sets. We also extended with different definitions of big data, highlighting the fact that size is only one dimension of big data. We have highlighted the issues that are raised potential of big data. Big data for development comprehends a whole range of applications and services that are changing the way decisions are made to improve global development. The intersection of all this data with something more digestible that can guide smart decision making is probably the biggest advance for governance in centuries. Major innovations in analytical techniques for big data have not yet taken place; one anticipates the emergence of such novel analytics in the near future. For instance, real-time analytics will likely become a prolific field of research because of the growth in location-aware social media and mobile apps. Since big data are noisy, highly interrelated, and unreliable, it will likely lead to the development of statistical techniques more readily apt for mining big data while remaining sensitive to the unique characteristics.

REFERENCES

[1] D. Fisher, R. DeLine, M. Czerwinski and S. Drucker, "Interactions with big data analytics," Interactions, vol. 19, no. 3, pp. 50-59, 2012.

[2] A. McAfee and E. Brynjolfsson, "Big data: the management revolution," Harvard business review, vol. 90, no. 10, pp. 60-68, 2012.

[3] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins and N. Kruschwitz, "Big data, analytics and the path from insights to value," MIT Sloan Management Review, vol. 21, 2014.

[4] A. R. Syed, K. Gillela and C. Venugopal, "The Future Revolution on Big Data," Future, vol. 2, no. 6, 2013.

[5] M. Hilbert, "Big Data for Development: From Information-to Knowledge Societies," Working Paper, 2013.

[6] Bharti Thakur, Manish Mann ―Data Mining for Big Data: A Review‖ ,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014 ISSN: 2277 128X .

[7] H. Herodotou, H. Lim, G. Luo, N. Borisov, L. Dong, F.B. Cetin, S. Babu, Starfish: A Self-tuning System for Big Data Analytics, in: Proceedings of the 5th Biennial Conference on Innovative Data Systems Research (CIDR 2011), 2011, pp. 261–272.

[8] M. Rahman, X. Li, H. Palit, Hybrid Heuristic for Scheduling Data Analytics Workflow Applications in Hybrid Cloud Environment, in: Proceedings of the IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011, pp. 966–974.

[9] Bernice Purcell "The emergence of "big data" technology and analytics" Journal of Technology Research 2013.
[10] Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta, Kumar N " Analysis of Bidgata using Apache Hadoop and Map Reduce" Volume 4, Issue 5, May 2014" 27.
[11] O'Driscoll, A., Daugelaite, J. and Sleator, R.D. (2013) 'Big Data', Hadoop and Cloud Computing in Genomics. Journal of Biomedical Informatics, 46, 774-781.
[12] Zhang, Jinso, Huang, Mao Lin (3-5 Dec. 2013), "5Ws Model for Big Data Analysis and Visualization", Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference, Pages 1021 – 1028.
[13] B. Brown, M. Chui, and J. Manyika, "Are you ready for the era of 'big data'?" McKinsey Q., vol. 4, pp. 24–35, 2011.
[14] P. Russom, "Big data analytics," TDWI Best Pract. Rep. Fourth Quart., 2011.
[15] M. K.Kakhani, S. Kakhani and S. R.Biradar, Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2(8) (2015), pp.228-232.
[16] A. Gandomi and M. Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management, 35(2) (2015), pp.137-144.
[17] K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big data analytics, Journal of Parallel and Distributed Computing, 74(7) (2014), pp.2561-2573.
[18] T. K. Das and P. M. Kumar, Big data analytics: A framework for unstructured data analysis, International Journal of Engineering and Technology, 5(1) (2013), pp.153-156.
[19] M. D. Assuno, R. N. Calheiros, S. Bianchi, M. a. S. Netto and R. Buyya, Big data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing, 79 (2015), pp.3-15.
[20] D. P. Acharjya, S. Dehuri and S. Sanyal Computational Intelligence for Big Data Analysis, Springer International Publishing AG, Switzerland, USA, ISBN 978-3-319-16597-4, 2015.
[21] W.-H. Weng and W.-T. Lin, "A Scenario Analysis of Big Data Technology Portfolio Planning," in International Journal of Engineering Research and Technology, 2013.
[22] F. Zulkernine, P. Martin, Y. Zou, M. Bauer, F. Gwadry-Sridhar, and A. Aboulnaga, "Towards Cloud- Based Analytics-as-a Service (CLAaaS) for Big Data Analytics in the Cloud," in Big Data (BigData Congress), 2013 IEEE International Congress on, 2013, pp. 62-69.
[23] F. Chang, J. Dean, S. mawat, et al., "Bigtable: A distributed storage system for structured data," ACM Trans. on Computer Systems, vol. 26, no. 2, pp. 1-39, June 2008.
[24] Bashed Kapil, Considerations for Big Data –Architecture and Approach, IEEE, 2012
[25] V. Shukla, P. K. Dubey," Big Data: Moving Forward with Emerging Technology and Challenges ", International Journal of Advance Research in Computer Science and Management Studies, Vol.2 (9), Sept 2014, pp. 187-193
[26] F. Chang, "Bigtable: A distributed storage system for structured data", ACM Transactions on Computer Systems (TOCS), vol. 26, no. 2, (2008), pp. 4.
[27] Puneet Singh Duggal, Sanchita Paul, "Big Data Analysis: Challenges and Solutions", International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV.