Volume 3, Issue 9, September 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Real Time Approach with BIG Data – A review

Gaurav Vaswani
Student, Second Year Computer Technology,
VESIT, Mumbai, India

Anuradha Bhatia ,
Faculty, Computer Technology Department,
Mumbai, India

Abstract:** The term "big data" is pervasive, and yet still the notion engenders confusion. Big data has been used to convey all sorts of concepts, including: huge quantities of data, social media analytics, next generation data management capabilities, real-time data, and much more. Whatever the label, organizations are starting to understand and explore how to process and analyze a vast array of information in new ways. In doing so, a small, but growing group of pioneers is achieving breakthrough business outcomes. Big Data provides opportunities for business users to ask questions they never were able to ask before. How can a financial organization find better ways to detect fraud? How can an insurance company gain a deeper insight into its customers to see who may be the least economical to insure? How does a software company find its most at-risk customers those who are about to deploy a competitive product? They need to integrate Big Data techniques with their current enterprise data to gain that competitive advantage. Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link previously created data. Data analysis, organization, retrieval, and modelling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge.

***Keywords:** Big data, Knowledgebase, Data environment, obstacles, volume, variety, velocity*

I. INTRODUCTION

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guesswork, or on pain staking constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, Retail, manufacturing, financial services, life sciences, and physical sciences. Scientific research has been revolutionized by Big Data.

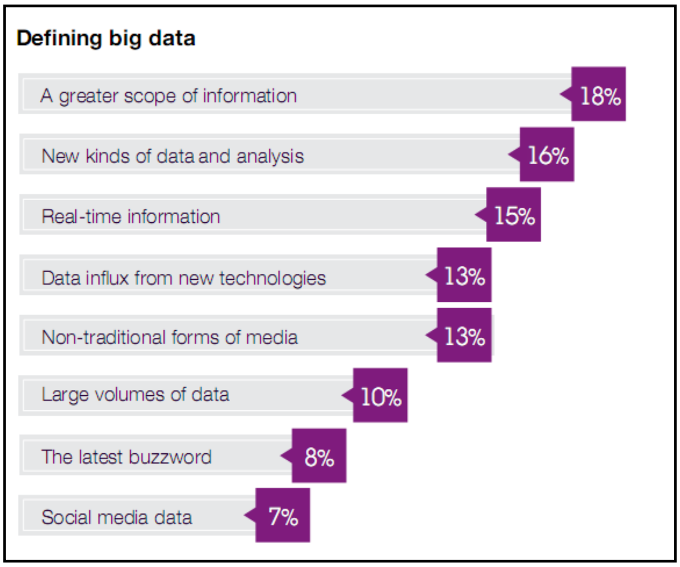

Fig 1: Defining Big Data

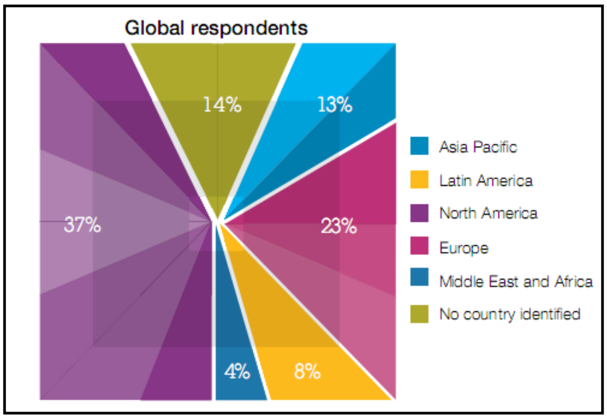

Fig 2: Global Respondents of Big Data

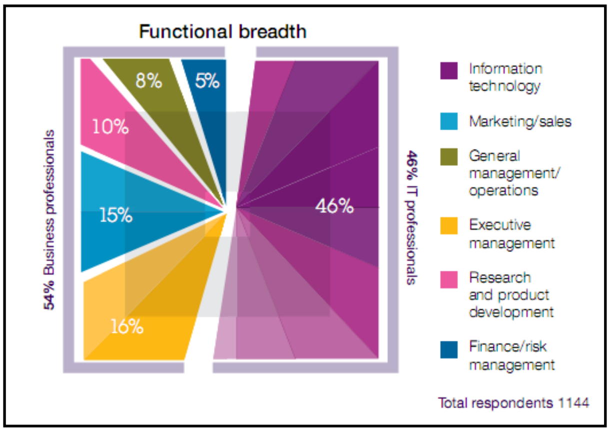

Fig 3: Functional Breadth of Big Data

II. THE BIG DATA WORKFLOW

The notion of exploring Wikipedia's view of history is a classic Big Data application: an open-ended exploration of "what's interesting" in a large data collection leveraging massive computing resources. While quite small in comparison to the hundreds-of-terabytes datasets that are becoming increasingly common in the Big Data realm of corporations and governments, the underlying question explored in this Wikipedia study is quite similar: finding overarching patterns in a large collection of unstructured text, to learn new things about the world from those patterns, and to do all of this rapidly, interactively, and with minimal human investment.

The convergence of these four dimensions helps both to define and distinguish big data:

Volume: The amount of data. Perhaps the characteristic most associated with big data, volume refers to the mass quantities of data that organizations are trying to harness to improve decision-making across the enterprise. Data volumes continue to increase at an unprecedented rate. However, what constitutes truly "high" volume varies by industry and even geography, and is smaller than the peta bytes and zeta bytes often referenced. Just over half of respondents consider datasets between one terabyte and one peta byte to be big data, while another 30 percent simply didn't know how big "big" is for their organization. Still, all can agree that whatever is considered "high volume" today will be even higher tomorrow.

Variety: Different types of data and data sources. Variety is about managing the complexity of multiple data types, including structured, semi-structured and unstructured data. Organizations need to integrate and analyze data from a complex array of both traditional and non-traditional information sources, from within and outside the enterprise. With the explosion of sensors, smart devices and social collaboration technologies, data is being generated in countless forms, including: text, web data, tweets, sensor data, audio, video, click streams, log files and more.

Velocity: Data in motion. The speed at which data is created, processed and analyzed continues to accelerate. Contributing to higher velocity is the real-time nature of data creation, as well as the need to incorporate streaming data into business processes and decision making. Velocity impacts latency – the lag time between when data is created or captured, and when it is accessible. Today, data is continually being generated at a pace data in dimensions that is impossible for traditional systems to capture, store and analyze. For time-sensitive processes such as real-time fraud detection or multi-channel "instant" marketing, certain types of data must be analyzed in real time to be of value to the business

Veracity: Data uncertainty. Veracity refers to the level of reliability associated with certain types of data. Striving for high data quality is an important big data requirement and challenge, but even the best data cleansing methods cannot remove the inherent unpredictability of some data, like the weather, the economy, or a customer's actual future buying decisions. The need to acknowledge and plan for uncertainty is a dimension of big data that has been introduced as executives seek to better understand the uncertain world around them (see sidebar, "Veracity, the fourth 'V.'").

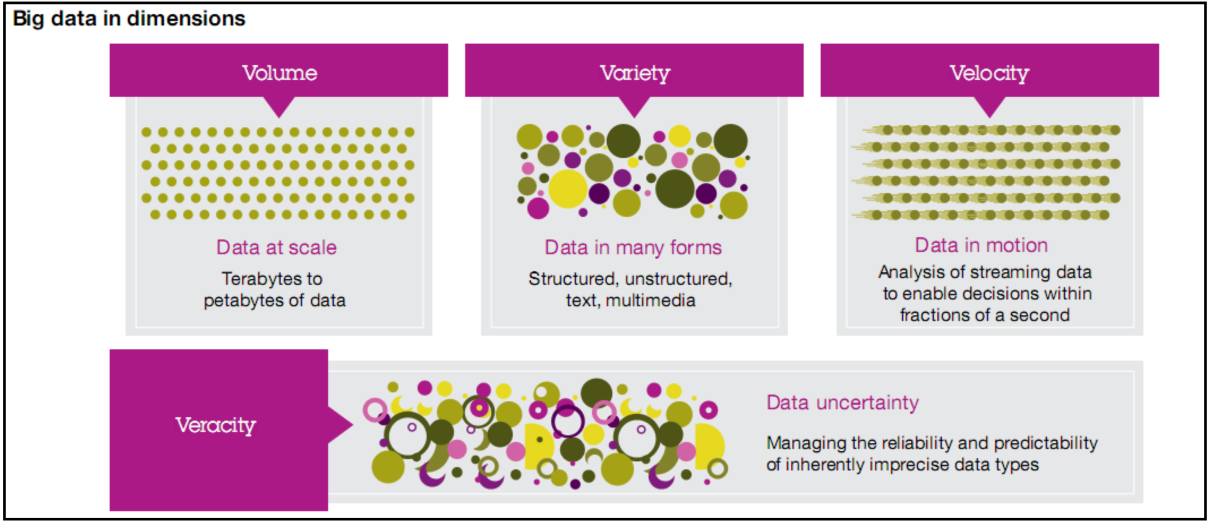
Big data in dimensions
Volume
Data at scale
Terabytes to petabytes of data
Variety
Data in many forms
Structured, unstructured, text, multimedia
Velocity
Data in motion
Analysis of streaming data to enable decisions within fractions of a second
Veracity
Data uncertainty
Managing the reliability and predictability of inherently imprecise data types

Fig 4: Big Data Dimensions

The emerging pattern of big data adoption is focused upon delivering measurable business value .To better understand the big data landscape, we asked respondents to describe the level of big data activities in their organizations today. The results suggest four main stages of big data adoption and progression along a continuum that we have labelled Educate, Explore, Engage and Execute.

Educate: Building a base of knowledge (24 percent of respondents). In the Educate

Stage, the primary focus is on awareness and knowledge development. Almost 25 percent of respondents indicated they are not yet using big data within their organizations. While some remain relatively unaware of the topic of big data, our interviews suggest that most organizations in this stage are studying the potential benefits of big data technologies and analytics, and trying to better understand how big data can help address important business opportunities in their own industries or markets. Within these organizations, it is mainly individuals doing the knowledge gathering as opposed to formal work groups, and their learnings are not yet being used by the organization.

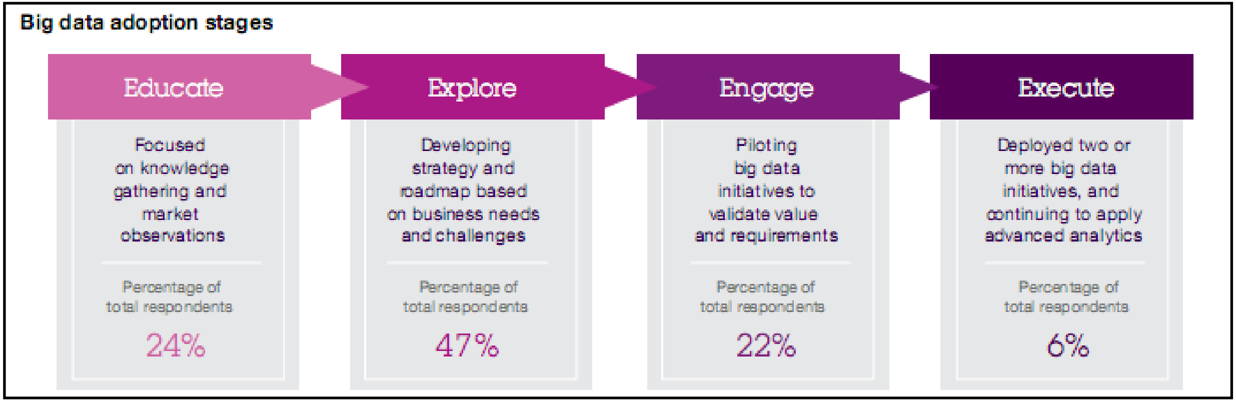
Big data adoption stages
Educate
Focused on knowledge gathering and market observations
Percentage of total respondents
24%
Explore
Developing strategy and roadmap based on business needs and challenges
Percentage of total respondents
47%
Engage
Piloting big data initiatives to validate value and requirements
Percentage of total respondents
22%
Execute
Deployed two or more big data initiatives, and continuing to apply advanced analytics
Percentage of total respondents
6%

Fig 5: Big Data Adoption Stages

As a result, the potential for big data has not yet been fully understood and embraced by the business executives.

Explore: Defining the business case and roadmap (47 percent)

The focus of the Explore stage is to develop an organization’s roadmap for big data development. Almost half of respondents reported formal, ongoing discussions within their organizations about how to use big data to solve important business challenges. Key objectives of these organizations include developing a quantifiable business case and creating a big data blueprint. This strategy and roadmap takes into consideration existing data, technology and skills, and then outlines where to start and how to develop a plan aligned with the organization’s business strategy.

Engage: Embracing big data (22 percent)

In the Engage stage, organizations begin to prove the business value of big data, as well as perform an assessment of their technologies and skills. More than one in five respondent organizations is currently developing proofs-of-concept (POCs) to validate the requirements associated with implementing big data initiatives, as well as to articulate the expected returns. Organizations in this group are working – With in a defined, limited scope – to understand and test the technologies and skills required to capitalize on new sources of data.

Execute: Implementing big data at scale (6 percent)

In the Execute stage, big data and analytics capabilities are more widely operational zed and implemented within the organization. However, only 6 percent of respondents reported that their organizations have implemented two or more big data solutions at scale – the threshold for advancing to this stage. The small number of organizations in the Execute stage is consistent with the implementations we see in the marketplace. Importantly, these leading organizations are leveraging big data to transform their businesses and thus are deriving the greatest value from their information assets. With the rate of enterprise big data adoption accelerating rapidly – as evidenced by 22 percent of respondents in the Engage stage, with either POCs or active pilots underway – we expect the percentage of organizations at this stage to more than double over the next year.

III. DATA AVAILABILITY

As depicted in Figure 10, we see how data availability requirements change dramatically as companies mature their big data efforts. Analysis of responses revealed that no matter the stage of big data adoption, organizations face increasing demands to reduce the latency from data capture to action. Executives, it seems, are increasingly considering the value of timely data in making strategic and day-to-day business decisions. Data is no longer just something that supports a decision; it is a mission-critical component in making that decision.

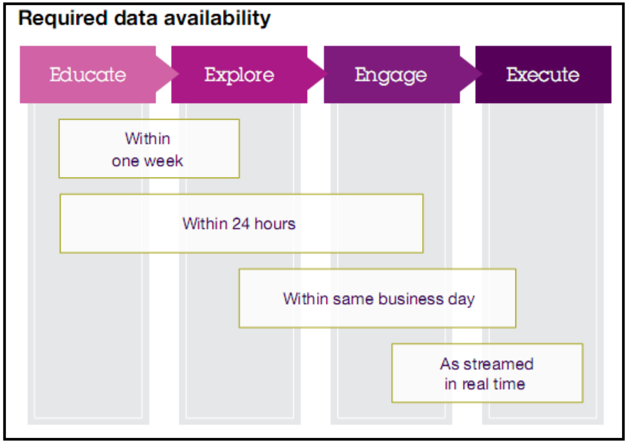

Fig 6: Big Data required Data Availability

IV. BIG DATA OBSTACLES

Challenges that inhibit big data adoption differ as organizations move through each of the big data adoption stages. But our findings show one consistent challenge – regardless of stage – and that is the ability to articulate a compelling business case (see Figure 11). At every stage, big data efforts come under fiscal scrutiny. The current global economic landscape has left businesses with little appetite for new technology investments without measurable benefits – a requirement that, of course, is not exclusive to big data initiatives. After organizations successfully implement POCs, the biggest challenge becomes finding the skills to operationalize big data, including: technical, analytical and governance skills.

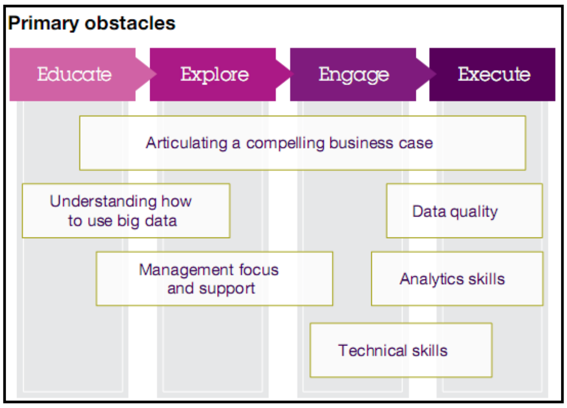
Primary obstacles
Educate
Explore
Engage
Execute
Articulating a compelling business case
Understanding how to use big data
Data quality
Management focus and support
Analytics skills
Technical skills

Fig 7: Big Data Obstacles

V. RECOMMENDATIONS: CULTIVATING BIG DATA ADOPTION

Work Study findings provided new insights into how organizations at each stage are advancing their big data efforts. Driven by the need to solve business challenges, in light of both advancing technologies and the changing nature of data, organizations are starting to look closer at big data's potential benefts. To extract more value from big data, we offer a broad set of recommendations to organizations as they proceed down the path of big data. Mass digitization, one of the forces that helped to create the surge in big data, has also changed the balance of power between the individual and the institution. If organizations are to understand and provide value to empowered customers and citizens, they have to concentrate on getting to know their customers as individuals. They will also need to invest in new technologies and advanced analytics to gain better insights into individual customer interactions and preferences.

VI. CONCLUSION

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

REFERENCES

1. Anuradha Bhatia and Gaurav Vaswani, “Big Data– An Overview”, International Journal of Engineering Sciences & Research Technology, Vol 2, No8,August (2013). (www.ijesrt.org).
2. Hey, T., Tansley, S. & Tolle, K. (2009) The Fourth Paradigm. “Data-intensive scientifc discovery”, Microsoft.
3. Hilbert, M. & Lopez, P. (2011) “The world’s technological capacity to store, communicate and compute information”, Science 332, 1 April 2011, 60-65.
4. IDC (2010) “IDC Digital Universe Study, sponsored by EMC”, May 2010, available at http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm
5. ICSU Strategic Plan 2006-2011, International Council for Science, Paris, 64pp
6. All the reports are available at the ICSU website, www.icsu.org
7. http://www.mysql.com/
8. Leetaru, K. (2011) “Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space”, First Monday. 16(9).
9. http://frstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040
10. Bellomi, F. & Bonato, R. (2005) “Network Analysis for Wikipedia”, Proceedings of Wikimania.
11. Leetaru, K. (forthcoming). “Fulltext Geocoding Versus Spatial Metadata For Large Text Archives: Towards a Geographically Enriched Wikipedia”,
12. http://www.tei-c.org/index.xml
13. http://history.state.gov/historicaldocuments
14. Leetaru, K. (forthcoming). “Fulltext Geocoding Versus Spatial Metadata For Large Text Archives: Towards a Geographically Enriched Wikipedia”, D-Lib Magazine.