International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-3, Issue-7)

Research Article | **July 2014**

Big Data Process Analytics: A Survey

Sameera Siddiqui
Cse, Rkgit,
Ghaziabad, India

Deepa Gupta
Amity Institute of Information Technology,
Noida, India

ABSTRACT-

This paper is the review paper which gives the summary of various surveys done by companies like TCS and IDC Enterprise on Big Data. It has become big news overnight and there are no signs that interest is decreasing. Over the last four years, companies around the world have awakened to a new direction: Big Data brings with it Big Responsibility. Big Data is thus very important to increase productivity growth in the entire world since it is affecting not only software-intensive industries but also public domains like education,health field ,education and administrative sectors. Big data refers to voluminous data which ranges in Exabyte's (10^{18}) and beyond. It is defined as the amount of data just beyond technology's capability to store, manage and process efficiently. The paradigm of processing huge datasets has been shifted from centralized architecture to distributed architecture. In this paper, we provide an extensive survey of Big data analytics research, while highlighting the specific concerns in Big data world. We present a taxonomy based on the key issues in this area, and discuss the different methods to tackle these issues. Based on this survey study many midmarket organizations report a need for tools ranging from real-time processing to predictive analytics, data cleansing, and data visualization .

Keywords— Big data, data visualization , data analytics, hadoop.

I. INTRODUCTION

As the current technology enables us to efficiently store and query large datasets, the focus is now on techniques that make use of the complete data set, instead of sampling. This has tremendous implications in areas like pattern recognition, machine learning and classification, to name a few. Therefore, there are a number of requirements for moving beyond standard data mining techniques:

- a a robust exploratory establishment to have the capacity to select an adequate method or design ;
- a new algorithm;
- a technology platform and adequate development skills to be able to implement it;
- a genuine ability to understand not only the data structure (and the usability for a given processing method), but also the business value.

As a result, building multi-disciplinary teams of "Data scientists" is often an essential means of gaining a competitive edge.
More than ever, intellectual property and patent portfolios are becoming essential assets.
One of the obstacles to widespread analytics adoption is a lack of understanding on how to use analytics to improve the business [1]

"Big Data" is a term encompassing the use of techniques to capture, process, analyse and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. By extension, the platform, tools and software used for this purpose are collectively called "Big Data technologies".[12]

In recent years, Big Data has become a major topic in the field of ICT. It is evident that Big Data means business opportunities, but also major research challenges. According to McKinsey & Co[2] Big Data is "the next frontier for advancement, competition and productivity"[2]. The effect of Big Data gives not only a huge potential for competition and growth for individual companies, but the right use of Big Data also can increase productivity, advancement, and competitiveness for entire sectors and economies.

Big Data has the potential to revolutionize not just research, but also education [3]. A recent detailed quantitative comparison of different approaches taken by 35 charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide instruction [4]. There is a strong trend for massive Web deployment of educational activities, and this will create an increasingly huge amount of detailed data about students' performance. [10]

It is widely believed that the use of information technology can reduce the cost of healthcare while improving its quality[5], by making care more preventive and personalized and basing it on more extensive (home-based) continuous monitoring. McKinsey estimates[6] a savings of 300 billion dollars every year in the US alone.

Similarly there have been persuasive cases made for the value of Big Data for urban planning (through fusion of high-fidelity geographical data), intelligent transportation (through analysis and visualization of live and detailed road network data), environmental modeling (through sensor networks ubiquitously collecting data) [7],financial systemic risk analysis (through integrated analysis of a web of contracts to find dependencies between financial entities) [8], homeland security (through analysis of social networks and financial transactions of possible terrorists), computer security (through analysis of logged information and other events, known as Security Information and Event Management (SIEM)), and so on.

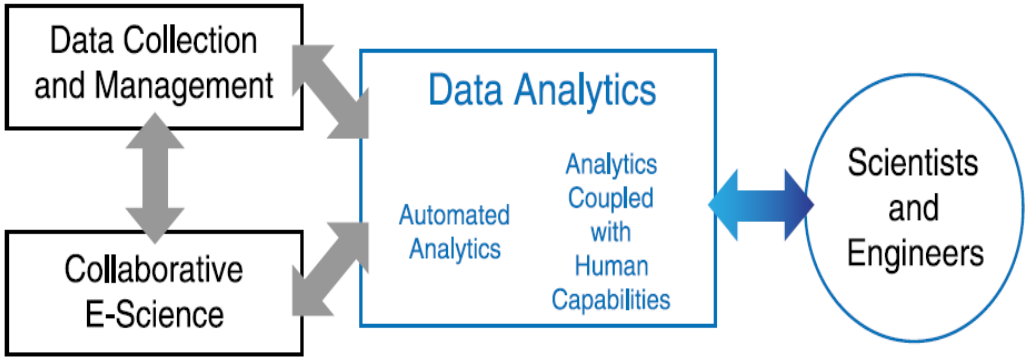

Fig 1. For regulatory science, one of the most important aspects of big data research is Data Analytics, which is a key for moving from data to human insights.[12]

The goal of this paper is to discuss in detail the current research that addresses these issues. We review the proposed solutions, and study the upcoming research challenges in Big data Analytics.

II. BIG DATA CHALLANGES

The biggest challenges of big data is facing the four V's which is [13]

(i) Volume which is the most visible aspect of big data referring to the fact that the amount of generated data has increased tremendously the past years. The natural expansion of internet has created an increase in the global data production. A response to this situation has been the virtualization of storage in data centres, amplified by a significant decrease of the cost of ownership through the generalization of the cloud based solutions. The noSQL database approach is a response to store and query huge volumes of data heavily distributed.

(ii) Velocity which captures the growing data production rates. More and more data are produced and must be collected in shorter time frames. The daily addition of millions of connected devices (smart phones) will increase not only volume but also velocity. Real-time data processing platforms are now considered by global companies as a requirement to get a competitive edge.

(iii) Variety is explained with the multiplication of data sources where comes the explosion of data formats, ranging from structured text to free text. The necessity to collect and analyse non-structured or semi-structured data goes against the traditional relational data model and query languages. This reality has been a strong motivation to create new kinds of data stores able to support flexible data models.

(iv) Value is highly subjective aspect refers to the fact that until recently, large volumes of data where recorded or regulatory but not exploited. Big Data technologies are now seen as enablers to create or capture value from otherwise not fully exploited data. In essence, the challenge is to find a way to transform raw data into information that has value, either internally, or for making a business out of it.

III. RESEARCH CHALLENGES

Everyone is talking about Big Data .The world data is doubling every 1.2 years .There are 7 billion people in the world,5.1 Billion of them own a cell phone. Each day we send over 11 billion texts, watch over 2.8 billion You tube videos and perform almost 5 billion Google searches and we are not just consuming it we're creating it. The data agents – generate over 2.5 Quintillion bytes everyday from communication devices, consumer transactions, online behaviour and streaming services.

In 2012 the world's information totalled over two zeta bytes that's 2 trillion gigabytes .By 2020 we will need 10 times more servers, 50 times more data management, 75 times more files to handle it all .If we see most companies they aren't ready. 80% of this new data is unstructured , it is too complex and too diagnosed to be analyzed by traditional tools

.There are 500k computer scientists yet only 3k mathematicians .We will fall short of the talent needed to understand big data by at least 100k .To find opportunities in big data we need new tools and new talent to mine this information and find value .We need big data analytics which is more than just technology .It's a new way of thinking which will help companies better understand customers , find hidden opportunities even help our government better serve citizens and mitigate fraud . It will inspire hundred , thousand and even million of new start-ups .We are at the beginning of the big data revolution .

IV. THE PAPER SURVEYS

Over the last three years,many research work has undergone in Big Data. Hundreds of articles have appeared in the general business press (for example, Forbes, Fortune, Bloomberg Business Week, The Wall Street Journal, The Economist)[9]. A March 2013 search on Amazon.com surfaces more than 250 books, articles and e-books on the topic. The technology research community: Gartner, Forrester, IDC are all involved into Big Data Study.

The 2014 IDG Enterprise Big Data research was completed with the goal of gaining a better understanding of organizations' big data initiatives, investments and strategies[11].
Key Findings Include[11]:

- Organizations are seeing exponential development in the amount of data managed with an expected increase of 76% within the next 12-18 months.
- Companies are escalating their efforts to derive value through big data initiatives with nearly half (49%) of respondents already implementing big data projects or in the process of doing so in the future; however, enterprise organizations are ahead of the curve in implementation plans compared to SMB organizations.
- CEOs are centered around on the value of big data and are partnering with IT executives who will purchase/manage/execute on the strategies.
- Organizations are investing in developing or buying software applications, additional sever hardware, and hiring staff with analytics skills in preparation for big data initiatives.
- Organizations are facing numerous challenges with big data initiatives and limited availability of skilled employees to analyze and manage data tops the list.
- In the next 12-18 months, organizations plan to invest in skill sets necessary for big data deployments, including data scientists (27%), data architects (24%), data analysts (24%), data visualizers (23%), research analysts (21%), and business analysts (21%).
- Half of respondents indicated there is no clear thought leader in the big data solution space.

IDG Enterprise's 2014 Big Data research was conducted online among the audience of six IDG Enterprise brands – CIO, Computerworld, CSO, InfoWorld, IT world and Network World – via web pop-up, forum posts, and email invitations. Results are based on 751 respondents.

Late in 2012, TCS launched its own study on Big Data. It focussed on six core issues, which needed attention[9].

- How much are companies investing in Big Data, and what kinds of returns are they achieving on their spending?
- What are companies in 12 industries doing with Big Data? That is, in which business functions and specific activities are they focusing their investments?
- What kind of digitized data are they finding to be most important ?
- How are they organizing the professionals who process and analyze Big Data (e.g., embedded in business functions, in a central analytics group, etc.), and what are the upsides and downsides of those reporting relationships?
- What are the biggest challenges of turning Big Data into insights that enable the company to make far better and faster decisions[9]?
- What is the current state of the technology, and where is it going?

A. They're Spending a Lot on Big Data[9]

The investments these companies made in Big Data were sizable. We measure those investments in two ways: by the median and the average survey respondent:

- Median spending on Big Data was $10 million, which was 0.14% of revenue (based on median revenue of survey respondents: $6.9 billion). We believe the median spending numbers provide a more accurate picture of spending on Big Data than the mean (or average) numbers here since the mean was skewed because of a number of respondents (7% of the ones we asked for spending data) who spent more than $500 million on Big Data in 2012.
- The average survey respondent spending on Big Data was $88 million in 2012 , which was 0.5% of average revenue (of $19 billion). Again, we believe this is a less reliable indicator of what companies are spending on Big Data.

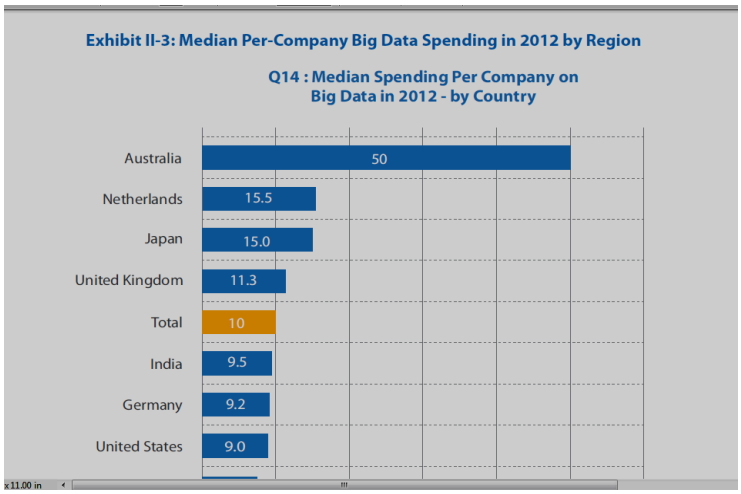

Fig 2. Median Per Company Big data spending in 2012 by region [9]

By the year 2015, companies across the surveyed regions expect to spend 75% more on Big Data, with Australia and U.K. companies projecting the highest spending per company. Median spending across all countries is projected to increase by 75% to $17.5 million. (Fig 3.) [9]

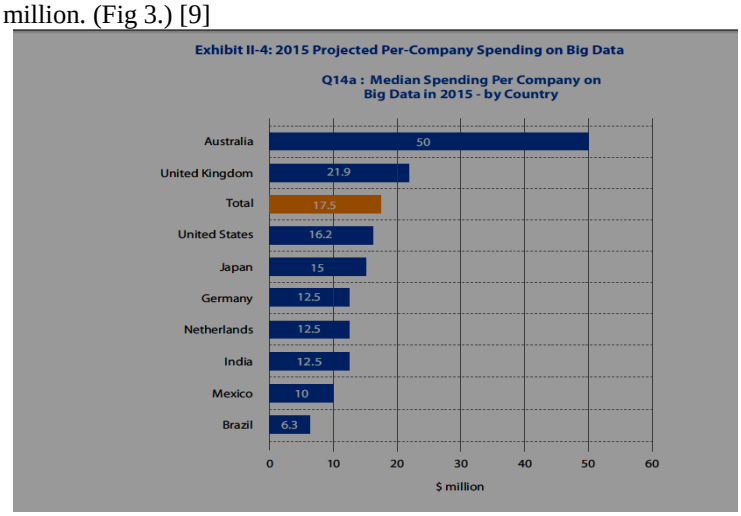

Fig 3: 2015 Projected Per- company Spending on Big Data[9]

B. Big Data is Improving Decisions in Most Companies[9]

Slicing and dicing huge volumes and varieties of digital data can keep data scientists busy for days or even weeks. But if the insights they derive do not provide useful guidance – or if business managers don't utilize that guidance – all that spending is futile. To find out, the survey first asked participants whether their initiatives had improved decision making in the business. The answer for the clear majority – 80% -- was indeed yes. The lowest percentage of improvement in decision making was in the U.S. (77%) while the highest was in Latin America (86%). (See Fig 4.)[9]

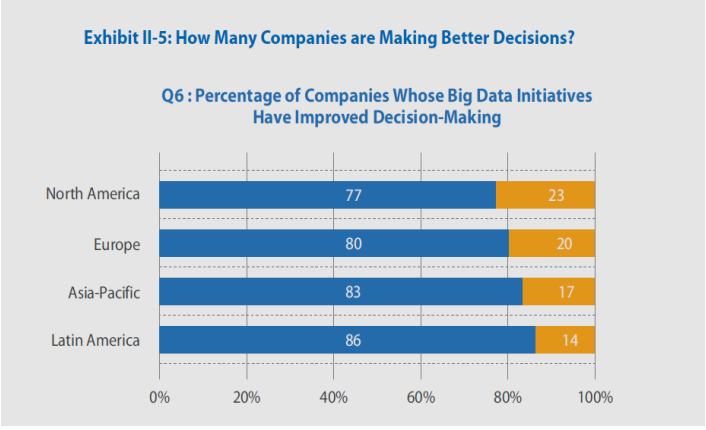

Fig 4. Percentage of companies whose big data initiatives have improved decision making[9]

- The company has a Big Data initiative(s) in place, and it has improved decision-making in the business.
- The company has a Big Data initiative(s) in place, and it hasn't yet improved decision-making in the business.

C. What Kinds of Digital Data are Companies Using[9]?

One way that Big Data experts such as Tom Davenport distinguish between the eras of 'big' and 'little' data is on the type of data companies are using. Big Data is more associated with unstructured and external data

D. Defining Types and Sources of Digital Data[9]

In our research, we defined data along two dimensions: structured versus unstructured and internal versus external. Given below are the definitions we used.
On the dimension of data structure:

- Structured – Data that resides in fixed fields (for example, data in relational databases or in spreadsheets)
- Unstructured – Data that does not reside in fixed fields (for example, free-form text from articles, email messages, untagged audio and video data, etc.)
- Semi-structured – Data that does not reside in fixed fields but uses tags or other markers to capture elements of the data (for example, XML, HTML-tagged text)

On the dimension of data source:

- Internal - from a company's sales, customer service, manufacturing, and employee records; from visits to the company's website, etc.
- External - from sources outside a company such as third-party data providers, public social media sites such as Facebook, Twitter and Google+, etc.

A much higher than anticipated percentage of data was not structured – either unstructured or 'semi-structured' (when combined, about half) [9]. (See Fig 5.)

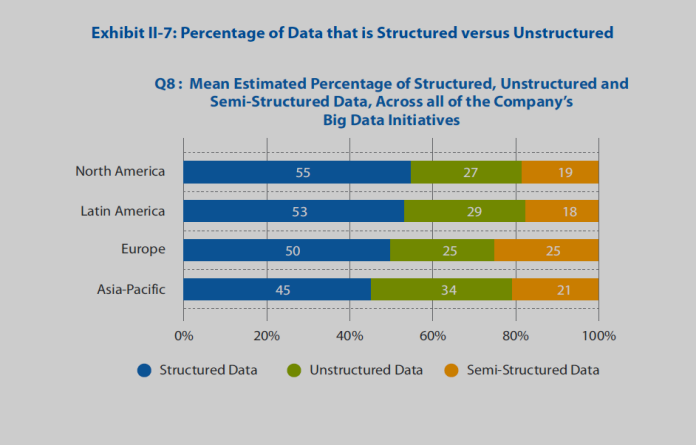

Fig. 5 Percentage of Data that is structured versus Unstructured[9]

"Studies have been done on electronic records that show, on average, 80%-90% or more of data in records is unstructured data," one health care executive said. "

E. Who is Selling Their Big (Digitized) Data[9]?

In 2012, about one-quarter of the companies we surveyed (27%) were capitalizing on this opportunity: selling their digital data. U.S. companies profited least from such data, with only 22% doing so. In contrast, half the Asia-Pacific companies we polled said they sell their digital data. About one-quarter of European and Latin American companies sold their digital data in 2012[9]. (See Fig 6.)

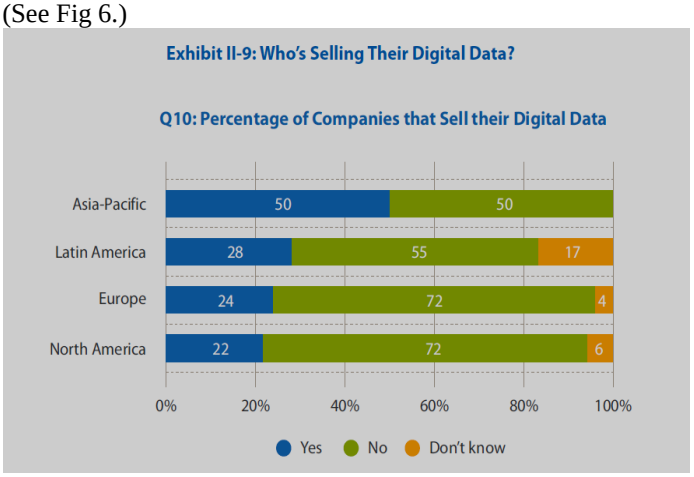

Fig 6. Percentage of companies that sell their digital data[9]

F. Views of the visionaries[9]:

To get some insights into what the technology makes possible today and what it may make possible in the near future, TCS interviewed their leading pioneers of Big Data technologies: Joseph Heller stein of the University of California at Berkeley.

Here are the highlights of those discussions.

"We're in the Early Days of Big Data – Like the Early 1900s' Era Before Washing Machines"

Joseph Heller stein, Chancellor's Professor of Computer Science, UC Berkeley, EECS Computer Science Division.

Joseph Heller stein likens today's times for Big Data to the early 1900s before the advent of the washing machine. (The first electric washing machines began appearing in the first decade of that century.) Back then, women spent an average 60 hours a week manually washing clothes[9]. Cleansing Big Data is in a similar state, Heller stein believes. He and several colleagues interviewed 35 analysts in companies across industries. They told them they spent 60% to 80% of their time on data preparation. "We're getting data from all over the place and it's not prepared for analysis or to be integrated with other data and analysis tools[9]," he says. "The tools available are not designed for analysts." Heller stein sees a big opportunity in bringing data cleansing into the modern-day equivalent of the electric washing machine. He is founder and CEO of a data analysis tools start-up called Trifacta[9].

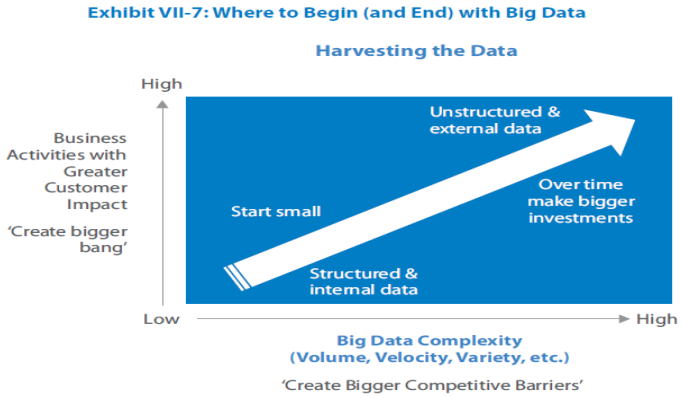

Fig 7. Big data Complexity[9]

G. Survey Demographics: Getting a 360-Degree View on Big Data[9]

To get a better picture of how companies are using Big Data, we designed the study to collect data from IT, business functions, and analytics managers. Nearly one third were IT managers; 62% were from eight business functions(marketing,sales,service,production/manufacturing, logistics, research & development, finance, and human resources). And the remainder (7%) operated in analytics groups[9]. (See Fig 8.) In all, 88% either headed one of those functions or reported to the head of it.

We also wanted people in these functions who had intimate knowledge of their company's Big Data activities. The majority (58%) said they played supporting roles in this endeavour, and 23% played leading roles. The rest (19%) said they had no role but substantial knowledge about what their company was doing with Big Data.

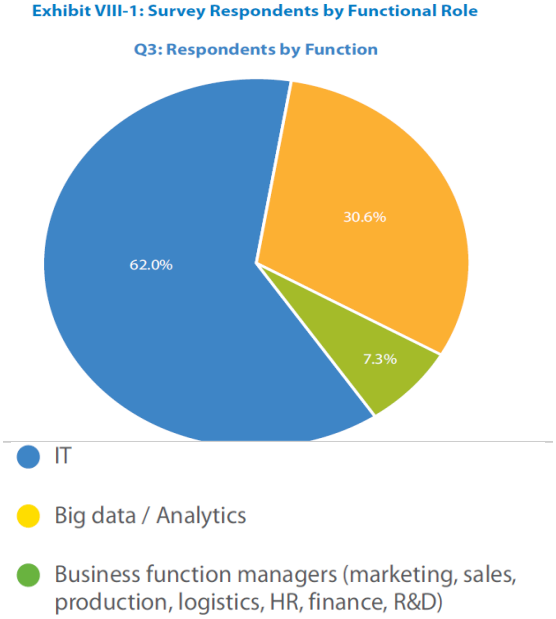

Fig 8. Survey Respondents by functional Role[9]

V. CONCLUSION AND FUTURE WORK

As we have entered in an era of Big Data, there is the potential for making profits in many scientific disciplines and enterprises through better analysis of the large volumes of data that is becoming available. However, many technical challenges like data Visualization and implementation is to be taken into consideration in future. This is just the survey paper which shows the demand of big data and how big companies are taking interest in it. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data. As Hadoop extends into new markets and sees new use cases with security and compliance challenges, the benefits of processing sensitive and legally protected data with all Hadoop projects and HBase must be coupled with protection for private information that limits performance impact.

REFERENCES

[1] LaValle et al: *Big Data, Analytics and the Path From Insights to Value*, (Dec 2010)
[2] McKinsey Global Institute, *Big Data: The next frontier for innovation, competition and productivity* (June 2011)
[3] Advancing Personalized Education. Computing Community Consortium. Spring 2011,jan1.
[4] Getting Beneath the Veil of Effective Schools: Evidence from New York City. Will Dobbie, Roland G. Fryer, Jr. NBER Working Paper No. 17632. Issued Dec. 2011.
[5] Smart Health Wellbeing. Computing Community Consortium. Spring 2011.
[6] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. 15 May 2011.
[7] A Sustainable Future. Computing Community Consortium. Summer 2011.
[8] Using Data for Systemic Financial Risk Management. Mark Flood, H V Jagadish, Albert Kyle, Frank Olken, and Louiqa Raschid. Proc. Fifth Biennial Conf. Innovative Data Systems Research, Jan. 2011.
[9] The Emerging Big returns on big data,TCS-Big-Data-Global-Trend-Study-2013,mar 21,2013
[10] Challenges and Opportunities with Big Data, Divyakant Agrawal, Philip Bernstein, Elisa Bertino ,Susan Davidson, Umeshwas Dayal, 1-1-2011
[11] IDG ENTERPRISE RESEARCH REPORTS,jan 6,2014 ,http://www.idgenterprise.com/report/big-data-2
[12] Grand Challenge: Applying Regulatory Science and Big Data to Improve Medical Device Innovation, Arthur G. Erdman∗, Daniel F. Keefe, Senior Member, IEEE, and Randall Schiestl, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 60, NO. 3, MARCH 2013
[13] Big Data A new World of Opportunities ,Nessi White Paper , December 2012.

