

## Digital Exhaust: An Insight on Big Data Projects

Dr. Prajakta Dhote, Dr. Pradeep Butey, Sandhya Dahake

<sup>1</sup>(prajaktadhote@gmail.com, Computer Department, City Premier College, Nagpur, India)

<sup>2</sup>(buteypradeep@gmail.com, Computer Science Department, Kamla Nehru Mahavidyalaya, Nagpur, India)

<sup>3</sup>(sandhya.dahake@raisoni.net, GHRIIT, Nagpur, India)

---

**Abstract:** At present the maximum industries are concentrating on managing their organization's big data efficiently. It can be a big competitive differentiator. For many companies today it is considered as mission-critical entity. Recent survey shows that there are a significant number of Big Data Projects underway in a wide variety of industries. While carrying out such Big Data Projects we need to take into account various architectural, organizational, and security considerations before deciding on it. This research paper throws an insight on various critical factors coming in the way of designing Big Data projects. Few recommendations and suggestions are given. We have tried to address both the perspective i.e. organizational as well as technical.

**Keywords** – Big Data, Big Data Governance, Big Data Projects, Big Data Technology, Business Intelligence,

---

### I. INTRODUCTION

In the present world of big data and internet economy, the way organizations are managing its data is an important competitive differentiator. For many companies today enterprise- wide data management is mission critical. As per survey nearly 85% of companies are engaged in some type of Big Data Projects [1] . There is a need of powerful, new and latest technologies to support the innovation and change. Technology which supports multiplatform architectures and storage capabilities is demanded. Such technologies and software are difficult and at the same time expensive too. It is very difficult to build a cost- effective approach to deal with big data. The process requires different approach for data sourcing and data preparation. It is also requires different technologies, additional skills and change to processes and organization culture.

It is world known that Big Data can be needed in various speciality areas. With the advent of the industrial internet many millions of sensors and self reporting components are now available throughout a make-and-move companies supply chain. Nearly every orientation from retails to charities is trying to leverage customer data for marketing and sales purpose. Prominent use of mobile phones is the new source to reach to the perspective customer.

Concept of Big Data is varying from person to person and from organization to organization. Some people treat it just a bigger Information Technology and to others it is a biggest asset of organization. Most of the companies today found that their current access to the relevant data is not adequate. That means company managers think that there is still a lot of untapped values in the data and there is a need for better data. On the other hand, most CIO's are more worried about the collection of this ever-growing data. They treat it as a "digital tsunami". If the companies have problems in data integration today, Big Data Projects may only exacerbate those fundamental data management problems. In this paper in this paper we will give you a close look on what is mean by Big Data Projects, why it is needed, various strategy considerations while developing it, a little discussion on required skill sets, the organizational considerations and data governance and security issues.

### II. BIG DATA PROJECTS

Larger organizations are the ones that have moved rapidly unto the Big Data arena. According to the Newvantage Partners survey 2016 the percentage of firms that now view Big Data as very important or critical to success has grown sharply from a bare majority of 54.5% in 2014 to a critical mass of 69.9% in 2015. The percentage of firms that see Big Data initiatives as being mission critical has risen from 23.2% in 2014 to 32.1% [7].

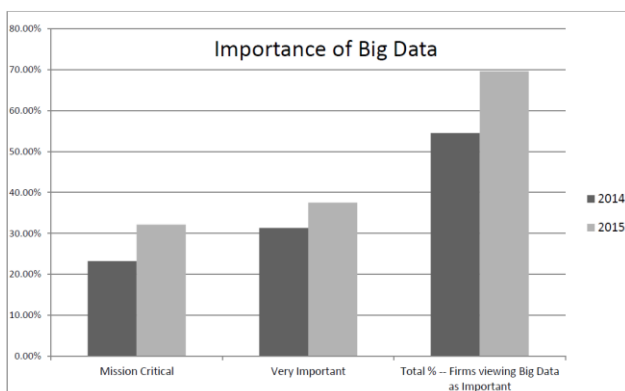


Fig. 1 Importance of Big Data

Importance of Big Data	2014	2015
Mission Critical	23.2%	32.1%
Very Important	31.3%	37.5%
Total % - Firms viewing Big Data as Important	54.5%	69.6%

Investment in Big Data initiatives is projected to grow sharply from 2014 through 2017. The percentage of firms reporting an expected investment in Big Data of greater than \$50MM is expected to grow nearly 5x from 5.4% to 26.8%. Firms investing greater than \$100MM is projected to increase from a small 1.8% in 2014 to a significant 8.9% by 2017. Firms investing \$50MM-\$100MM is projected to increase from 3.6% in 2014 to a very significant 17.9% by 2017.

While, 41% of firms are still investing less than \$10MM in Big Data in 2014, this number is projected to drop by more than half to 19.6% by 2017. Big Data spending is on the rise [1].

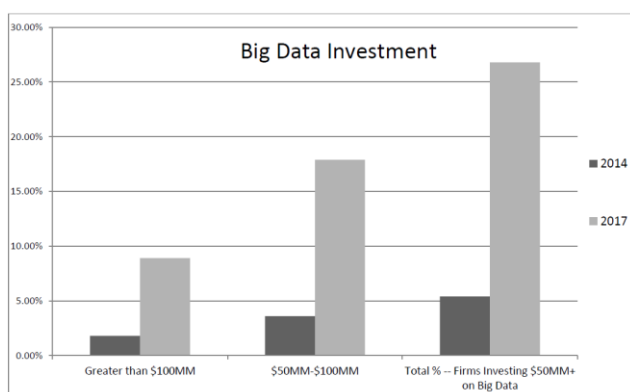


Fig. 2 Investment in Big Data Projects

Big Data Investment	2014	2017
Greater than \$100MM	1.8%	8.9%
\$50MM-\$100MM	3.6%	17.9%
Total % - Firms Investing \$50MM+ on Big Data	5.4%	26.8%

That's a Big Data for big companies at big cost and reflects the integral nature of data collection. That is not to say that small and medium-sized enterprises (SMEs) are not interested in Big Data projects, but most surveys indicate that small companies are more likely to opt for the prebuilt analytics tools available to them, both off-the-shelf and through the cloud, that are specifically designed for their particular vertical industry – health care, small manufacturing, transportation, and so on – and provide turn-key solutions that don't require huge investments in staff training, project management, or technologies.

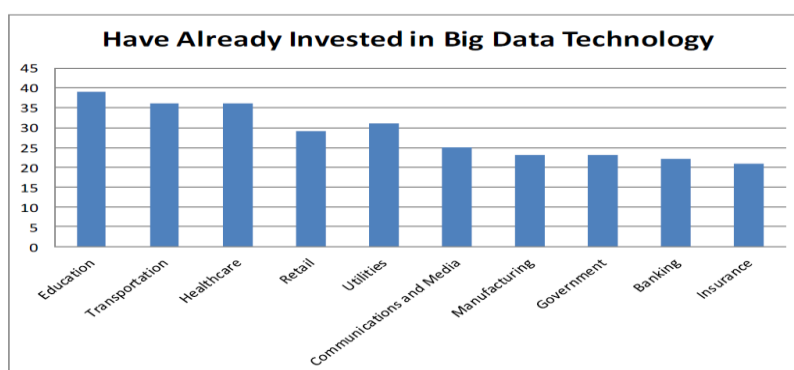


Fig. 3 Big Data Industry Focus

Data Source: Gartner

It is hard to say that this approach provides companies near to the potential of Big Data or not. To understand what Big Data means to various companies, it helps to look at the types of benefits that companies are looking for in a Big Data project. In the New Vantage survey, companies described a variety of benefits they hope to achieve, from operational efficiency to new product innovations. Interestingly, and probably a reflection of the number of financial services and big pharmaceuticals taking part, 64% of companies responding said they use Big Data for new product development and innovation. Companies also cited reduced risk and higher quality products or as goals. Many of the benefits they sought reflected better decision making capabilities. But when asked to identify the biggest opportunity, more than half returned to the theme of better customer insight, increased sales, and customer loyalty. Many companies hope to use Big Data projects to overcome data integration and data access issues. One of the most revealing statistics from the New Vantage survey is that more than half the companies that responded said that their main goal with a Big Data project is to be able to access “relevant, accurate, and timely” data. Sixty percent of respondents said that they want to apply Big Data technologies to data that they are already collecting. In short, a Big Data project, they hoped, would help them to overcome the data silos and lack of access to relevant data that still plague so many companies today.

Most of the companies in the survey accepted that their current access to so called relevant, accurate and timely data is not adequate [3]. A similar survey taken by Avanade found that 43% of all companies said they are dissatisfied with their current tools to filter out irrelevant data, and 46% said they had made poor business decisions because of bad or outdated data [4].

### **III. BIG DATA PROJECT USERS**

Most of the Big Data Projects are driven by marketing and sales than by the tech shop. That is why there is a tremendous growth in digital marketing and the fact that everything from mobile consumer apps to customer analytics (this is a key application of Big Data) is driven and used by the marketing department [5]. It also highlights the growing importance of digital competence within the marketing function. More than 80% of companies with annual revenue of \$500 million now have chief marketing technology officers [6]. Forrester recently estimated that IT spending by the Marketing department within companies is set to rise two to three times faster than IT spending by the organization as a whole.

### **IV. STRATEGY CONSIDERATIONS**

There are many routes to strategy development. But before anything one should have a clear thought on why your company is considering a move toward Big Data. Many companies want to be able to gain insights by capturing more and varied data. Others (nearly 60%) want the ability to do better analysis with the data they already have available. Still other organizations see a move toward Big Data as necessary to preemptively manage issues that may confront the organization in the future, such as building in scalability or better security. The most important first steps to any Big Data strategy project are to understand what data are needed and what data are available. It can be helpful to start the process by reviewing the report requests recently received by IT from the various business units.

Next point of thought is origins of data your company will need. Most data sets taken from customer-related web activities are subject to legal controls of some type (data privacy or regulatory compliance, usage rights, and so on) and may require customer consent through a legally vetted Terms of Agreement contract.

In addition to above considerations think about issues of accuracy and completeness (data hygiene) or data latency. Even if the company is hoping to take advantage of large data sets, the more consistent and clean that data is, the more beneficial it will be. An organization needs a policy not just for how to keep data but also for how to get rid of it.

Once it is clear about what data your organization needs, it is important to attempt a first cut at capacity planning and to calculate as accurately as possible how much data, and of what type, will be gathered over the next three to five years. From that, you can begin to consider the implications in terms of cost, storage capacity, and architectural needs. Finally, consider who in the company needs to see that data and in what format. It is important to identify the users of the data and think about the best way of making it available to them. Most advanced analytics will require not only data scientists and data analysts to configure the report generators but to interpret and decipher the reports themselves.

### **V. ARCHITECTURAL STRATEGY CONSIDERATIONS**

As per Big Data Executive survey 2016 firms cite the need to integrate more data – from new sources as well as legacy source. This focus on integrating greater varieties of data is often referred to as the “long tail” of Big Data. While 40% of firms cite “variety” (more sources) as the primary driver of Big Data investment, only 14.5% name “volume” (more data) or “velocity” (faster data).

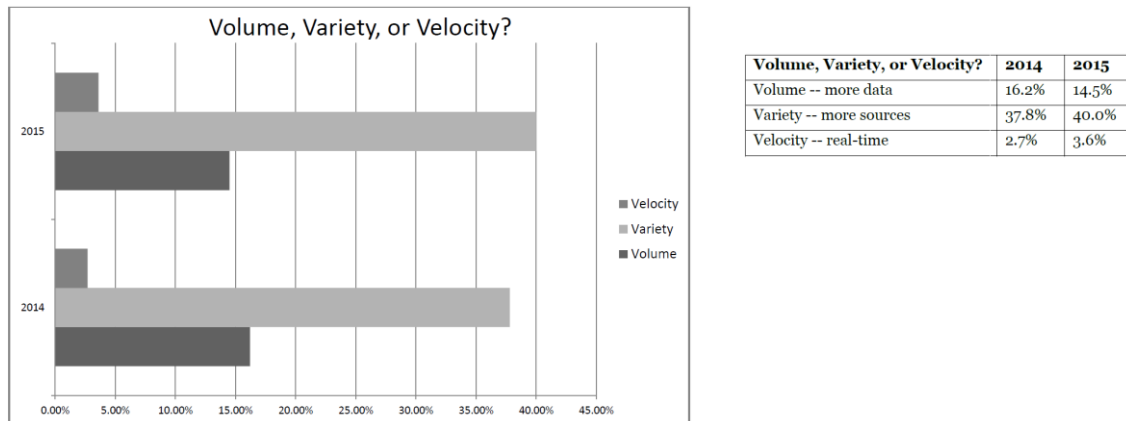


Fig. 4 Comparison of Volume, Variety and Velocity aspects of data

One of the most important considerations facing an organization contemplating adapting Big Data technologies is how to acquire new Big Data capabilities at a reasonable cost and with as little disruption to the company as possible. The biggest question before organizations is how to reconcile these two regimes: the clean and the chaotic, the contained and the distributed, the discrete and the enormous, the old and the new. The biggest concern of most organizations is that their current systems are already equipped with data warehouses and analytics but are based on traditional in-house storage and retrieval technologies – relational databases, SQL, and so on. These are the backbone or nervous system of the organization.

On the other hand the Big Data environment consist of cloud computing, NoSQL, Hadoop-like technologies, and advanced analytics. It is simply a different world. It can use the structured data that your company already collects, but to get real value from the search, storage, and analytics, you need to augment that structured data with various other types of much less stable, unstructured data. Apart from these two extremes there is middle ground: Small data or “smart data” applications that combine some of the search and data management flexibility of Hadoop-like technologies with cloud-based applications and storage are emerging every day. They have their own value, but even with this approach there are migration costs and subscription fees, and it still often involves extracting and storing data separately from your main databases and enterprise systems, and “dumping” that data into separate, (often) cloud-based storage facilities.

Due to these chaoes many organizations are in a quandary about whether the advantages of Big Data search and analytics justify such an infrastructure upheaval and are wondering about the best approach to combining these two different frameworks for their particular organization. There are three broad configuration choices available to them.

1. Built on your own solution in addition to present IT structure. ( more than a half of large organization will have begun this type of do-it-yourself project [8].
2. Outsource it using cloud (cloud base companies need to make their offerings as easy, flexible, and powerful as possible) [9].
3. Run parallel Database Framework.

## VI. SKILLS CONSIDERATIONS

Organizations which are moving to Big Data Projects will need experts understanding how to value and collect the source of data available to the organization. He should be specialized in using Hadoop- like technologies, should know how to mine manipulate data and equally aware of analyzing and interpreting the output from Big Data analytics tools. Experts having these skill sets are not readily available in the organization today. It means a company will probably need to hire expensive employees and devise new, dedicated roles. Few of them are described in short as follows.

1. **Data Engineers** – Data engineers are the technology experts who work with analyst to source and mine the data, create and manage the NoSQL/SQL or Hadoop infrastructure, and coax useful information out of the process.
2. **Data Analysts/ Data Scientist** - This specialized people determine what data goes into that Hadoop framework and interpret the data when it comes out. According to Dice’s 2014-2015 salary survey highest paying IT salaries were for these Big-related skills [10], These data analysts are the ones who decide what data can be used from which sources, and apply the statistical structure that allows the Hadoop engineer to create the logical schema for that type of data. They’re called data scientists because they need to have the methodical, hypothesis-based reasoning that allows them to design the statistical models and interpret the algorithms that reveal correlation, patterns, and insight. They also play a key role in soliciting the data

needs from various departments around the organization and providing those departments with a way to visualize that information that is most helpful to them. That means these analysts are needed not just centrally within the IT function but within key use areas like marketing, sales, procurement, logistics, inventory, or production, throughout the organization.

Only one-third of companies, according to Gartner, are confident that they can move ahead with a Big Data initiative using their current staff [11]. Even the largest companies may not be able to cover all the skills needed in a Big Data initiative and so will need to turn to outside consultants or particularly to the services of the technology/tool providers themselves.

## **VII. ORGANIZATIONAL CONSIDERATIONS**

Some 80% of companies responding to the NewVantage survey said that Big Data initiatives involved reaching across multiple business lines or functions [8]. Much of the new data sought by companies tends to come from unstructured sources – the Web, chat lines, social media – it is now possible for business lines – especially marketing and sales functions – to simply turn to cloud applications independent of the supervision of the central IT department. A recent poll by IDC found that 32% of departments use cloud-based services “in part” and 12% “very comprehensively” without the support of the company’s central IT function. That may indicate a new level of departmental independence, but it also reflects their frustrations. Far from IT complaining that the departments had moved unilaterally, in fact, 40% of those department-led initiatives complained that they had received no support from central IT [12]. According to Gartner, by 2020 nearly 90% of technology spending will originate outside the central IT department [13].

This type of decentralized model, where each department or function goes its own way, is welcome freedom and simply reflects the shifting authority from the CIO to the Chief Marketing Technology Officer. Certainly cloud-based applications do promote a more agile use of IT resources, but departmental independence has obvious drawbacks. The lack of central IT control can mean duplication of resources, and a lack of data sharing among silo-focused departments, each with its own set of specialist analysts and its own data collection procedures, can be inefficient and lack strategic continuity. It can also create significant risks to good governance and data security.

## **VIII. GOVERNANCE AND SECURITY ISSUES**

Collecting and selling customer data, is the focus of a Big Data initiative and it is a serious thing in terms of governance and privacy. Venture capitalist Ted Schlein made an interesting observation recently, noting that there are two types of companies: those that have had a data breach and know it, and those that have been breached and don’t know it [14]. Now company databases are filled with all manner of customer data, data security has become integral to customer relations and the company’s core business. Hacking and data breaches harmful to the customers whose data is stolen, but these types of data thefts have a significant effect on the company being breached. The servers of smaller vendors and suppliers are usually less well protected in the first place and have increasingly become the target of e-mail phishing fraud and data theft. The more customer data a company collects, and the more it attempts to make that data available to all its employees, anytime they need it, the more exposed they are to financial and reputational damage when their systems are breached. Following are the aspects of Big Data which is also responsible for maximum data breach –

1. Collecting as much as data possible and making it available to all departments.
2. Concentration on giving data through Single Company View.
3. Making data access available to noncompany employee.
4. Allowing departments to collect and store customer data.
5. Allowing that data to be input and access through mobile devices.

Major threat to data is when it is in motion. Permission and authorization rules must be implemented well. Data privacy laws altered every other day in nations all over the world. These data governance and security policies must be followed carefully and strictly.

## **IX. CONCLUSION**

Big Data Projects are the major concern for the Fortune 1000 companies. These projects give rise to new technological developments and innovations. Marketing and sales department need Big Data Projects. They want each and every bit of data to be collected and used to reach their customers and know their latest and changing demands. This results in “data tsunami”. Organization should find their way to throw out outdated and unwanted data. While developing and implementing Big Data Projects organization must understand their need and find answer that whether implementing this solution will answer their query as this is very expensive matter in terms of money as well as time. Major concern of these projects is security and governance because data is in

motion, moving freely using mobile technologies. These projects are very expensive and time consuming. To operate and decipher the result special experts are required. Organization should double assure before implementing this change. Once it is decided to go for it then study each and every consideration carefully.

#### REFERENCES

- [1] *NewVantage Partners Big Data Executive Survey: Themes & Trends*, 2016, <http://newvantage.com/wp-content/uploads/2016/12/NVP-Big-Data-Survey-Themes-Trends.pdf>.
- [2] *Ibid.*, NewVantage Partners
- [3] Avanade, *Global Survey: The Business Impact of Big Data*, November 2010, [http://www.avanade.com/Documents/Research%20and%20Insights/Big%20Data%20Executive%20Summary%20FINAL%20SEO v.pdf](http://www.avanade.com/Documents/Research%20and%20Insights/Big%20Data%20Executive%20Summary%20FINAL%20SEO%20v.pdf).
- [4] Taylor, Paul, *Financial Times Corporate 'Digital IQ' Linked to Performance*, March 26, 2014, <http://www.ft.com/intl/cms/s/0/8ad17f4c-b52e-11e3-af92-00144feabdc0.html?siteedition=intl#axzz2xAvlKJlp>.
- [5] Brinker, Scott, Chief Digital Officer, *81% of Big Firms Now Have a Chief Marketing Technologist*, January 31, 2014, <http://chiefdigitalofficer.net/81-of-big-firms-now-have-a-chief-marketing-technologist/>.
- [6] NewVantage Partners, *Big Data Executive Survey 2016*.
- [7] *Ibid.*, *The Economist*
- [8] Olavsrud, CIO *Thor, Red Hat and Horton works Expand Strategic Big Data Alliance*, February 10, 2014, [http://www.cio.com/article/748045/Red\\_Hat\\_and\\_Hortonworks\\_Expand\\_Strategic\\_Big\\_Data\\_Alliance?page=2&taxonomyId=600010](http://www.cio.com/article/748045/Red_Hat_and_Hortonworks_Expand_Strategic_Big_Data_Alliance?page=2&taxonomyId=600010).
- [9] Bednarz, Ann, CIO *Big Data Skills Pay Top Dollar*, February 7, 2015, [http://www.cio.com/article/747927/Big\\_Data\\_Skills\\_Pay\\_Top\\_Dollar?taxonomyId=600010](http://www.cio.com/article/747927/Big_Data_Skills_Pay_Top_Dollar?taxonomyId=600010).
- [10] Riley, James, *ComputerWeekly.com Cut Big Data Down to Size*, July 2013, <http://www.computerweekly.com/feature/Cut-big-data-down-to-size>.
- [11] CIO *Strategic Cloud Computing Demands Cooperation*, October 28, 2013, <http://zerodistance.cio.com/2013/10/vcloud/#sthash.x5ijM4Ib.dpuf>.
- [12] Flinders, Karl, *ComputerWeekly.com, Gartner Outsourcing Summit 2013: How Digitisation Is Shaking Up IT Outsourcing*, September 10, 2013, <http://www.computerweekly.com/news/2240205134/Gartner-Outsourcing-Summit-2013-How-digitisation-is-shaking-up-IT-outsourcing>.
- [13] Rosenbush, Steve, *Wall Street Journal Kleiner's Ted Schlein on Cyber Risk: 'It Only Gets Worse*, February 4, 2014, <http://blogs.wsj.com/cio/2014/02/04/kleiners-ted-schlein-on-cyber-risk-it-only-gets-worse/>.