ICEB 2015 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-6-2015

From Big Data To Knowledge – Good Practices From Industry

Amit T. Chowdhury

Guan Chong

Sean Z. X. Lee

Ravi S. Sharma

FROM BIG DATA TO KNOWLEDGE – GOOD PRACTICES FROM INDUSTRY

Amit T. Chowdhury, Nanyang Technological University, Singapore, tridib001@e.ntu.edu.sg
Chong Guan, SIM University, Singapore, guanchong@unisim.edu.sg
Sean Z. X. Lee, University at Buffalo, U.S.A., zhanxion@buffalo.edu
Ravi S. Sharma, Nanyang Technological University, Singapore, asrsharma@ntu.edu.sg

ABSTRACT

Recent advancements in data gathering technologies have led to the rise of a large amount of data through which useful insights and ideas can be derived. These data sets are typically too large to process using traditional data processing tools and applications and thus known in the popular press as 'big data'. It is essential to extract the hidden meanings in the available data sets by aggregating big data into knowledge, which may then positively contribute to decision making. One way to engage in data-driven strategy is to gather contextual relevant data on specific customers, products, and situations, and determine optimised offerings that are most appealing to the target customers based on sound analytics. Corporations around the world have been increasingly applying analytics, tools and technologies to capture, manage and process such data, and derive value out of the huge volumes of data generated by individuals. The detailed intelligence on consumer behaviour, user patterns and other hidden knowledge that was not possible to derive via traditional means could now be used to facilitate important business processes such as real-time control, and demand forecasting. The aim of our research is to understand and analyse the significance and impact of big data in today's industrial environment and identify the good practices that can help us derive useful knowledge out of this wealth of information based on content analysis of 34 firms that have initiated big data analytical projects. Our descriptive and network analysis shows that the goals of a big data initiative are extensible and highlighted the importance of data representation. We also find the data analytical techniques adopted are heavily dependent on the project goals.

Keywords: Big data, analytical techniques, data-type, data-spectrum, case-study.

INTRODUCTION

Development in technologies and increase in interactions among business, consumers and suppliers have resulted in generation of vast amounts of information from which useful insights can be derived almost in real time. This occurrence is popularly known as the big data revolution [4]. Big data is not only large quantities of data sets, but information that possess characteristics like variety, velocity and value. It can be static in the form of historical information or may even evolve in real-time. Big data analytics have become increasingly important in both the academic and the business communities over the past two decades. Industry reports have highlighted this significant development. Based on a survey of over 4,000 information technology (IT) professionals from 93 countries and 25 industries, the IBM Tech Trends Report [17] identified business analytics as one of the four major technology trends in the 2010s. In another survey with 930 respondents from across the globe on the current state of business analytics within organizations by Bloomberg Businessweek Research Services [3], 97% of companies with revenues exceeding $100 million were found to use some form of business analytics, up from 90% in 2009. Big data analytics has been a fast growing market. According to the International Data Corporation (IDC), the global revenue of major players involved in big data grew by 35% to €6.1 billion in 2012 and is expected to continue rising at a similar rate until 2016 [18]. The Economist maintains that, "In recent years, Oracle, IBM, Microsoft and SAP have spent more than €11.3 billion on buying software firms specializing in data management and analytics. This industry is estimated to be worth more than €65 billion and growing at almost 10% a year, roughly twice as fast as the software business as a whole" [11]. Researchers have undertaken extensive analysis on the role of data in promoting innovation, digital literacy, economic growth and well-being [28].

The opportunities associated with data analytics in organizations have generated significant interest, which is often in the domains of techniques, technologies, systems, practices, methodologies and applications that analyse various data to help an enterprise better understand its business and market and make critical decisions [10]. Application of big data analytical tools can be found across a wide range of industries including healthcare, manufacturing, retail, market intelligence, e-government and security.

Not all organizations were embracing data-driven decision making, but when adopted, big data analytics can improve firm performance dramatically. Across the analyses conducted by McAfee and Brynjolfsson [23], they found that the more companies characterized themselves as data-driven, the better they performed on objective measures of financial and operational results. In particular, companies in the top third of their industry with data-driven decision making were 5% more productive and 6% more profitable than their competitors on average, after accounting for the variance contributed by labour, capital, outsourced services, and traditional ICT investment. The impact was statistically significant and economically important.

While there is little doubt on the impact that deriving knowledge out of big data can have on industry innovation and decision making, there are also many challenges that have been identified. In spite of the well documented rise of the data revolution, many businesses are yet to realign their processes to make the best of this data driven environment. For example, Isik, Jones and Sidorova [19] suggest that technological capabilities such as data quality, user access and the integration of business intelligence with other systems are necessary for its success. Ross [26] and Sharma & Bhattacharya [27]have pointed out that most companies not only fail at doing a good job with the existing information, but also do not know how to manage, analyse and translate it into their own knowledge. In order to develop companies' competencies, they first need to learn how to use the data already embedded in their core operating systems. Until a company learns how to use data and analysis to support its operation decisions, it will not be in a position to benefit from big data [26]. Another survey of more than 3000 executives suggested that data and the relating techniques of storage and analysis were not the biggest issue. Instead, what exactly is needed from the data -- the management goal -- is the biggest challenges to set [21]. To help the industries and the public sectors translate data into knowledge, there is substantial work needed to be done in the academic domain. In an effort to help organisations understand how to perform big data analytics, transform data into actionable knowledge, and develop the ability to benchmark effectively, we aggregate the analysis of 34 cases of big data solutions in various industries in order to distil "common good practices" on its effective application using content analysis.

The remainder of the paper is organized as follows. The next section provides a review of big data analytics. Particularly, we discuss the definition, the various techniques and the applications in various industries, highlighting its many challenges and opportunities. Subsequently, we proposes an empirical study that aggregate the content analysis 34 cases of big data solutions in various industries. We then describe the data coding, data analysis and discuss the results. Conclusions and future research are given in the last section.

LITERATURE REVIEW

Big Data Analytics

In the current research, we adopt the definition of big data as "data sets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse" [22]. The transformative process from big data to knowledge varies by sector and even practitioner, according to the kind of tools available and the sizes of dataset achieved in the particular industry [13]. Overall, big data in different applications range from a few terabytes to thousands of petabytes or even zetabytes. The amount of data has been exponentially growing in the recent years. During 2012, around 2.5 exabytes of data were generated daily, and the number is doubling each 40 months [23]. Data is available and created everywhere, in each department, each company, each industry, and each economy. However, big data is not only about the sheer volume. Two other significant features have been highlighted including velocity and variety [19]. The speed of data creation is a key factor in data generation, especially in channels such as online social network platforms, global position system (GPS), and real-time transaction tracking. With the real-time information, it is possible for an organization to make more accurate decisions. Another key factor in big data is the variety of the data generated. Compared with traditional relational dataset, these new data usually comes in as various forms and are unstructured, which makes them not possible to be organized in traditional relational database management system (RDBMS) databases. This new trend in decision support is evocative of what we saw in 1990s with the emergence of data mining; with the new emphasis being on data with a large number of dimensions and much higher complexity [4]. As big data is very different from traditional data, new technologies and techniques are required to serve the purpose of storage, indexing and analysis. Big Data analytics refers to technologies that carry out data mining and statistical analysis. These techniques rely on relational Database Mangement Systems (DBMS), data warehousing, extracting transforming and loading, Online Analytical Processing and Business Process Management [5].

From Big Data to Knowledge

Application of big data analytics tools can be found across a wide range of industries including healthcare, manufacturing, retail, utilities and public sector. On the other hand, financial giants like the Citigroup and GE Ventures have invested in big data startups like Ayasdi to explore the possibilities arising out of processing big data [20]. Big data has found application in defence research and the power sector too [16, 30]. In the following sections, we review the impact of Big Data analytics on organisations in the different industries.

E-Commerce

Arguably, e-commerce communities have been the pioneers in generating Big Data. E-commerce giants such as Amazon and eBay have lead recent surge of web analytics use. Social media platforms such as Facebook and Twitter have also proved to be a gold mine of Big Data on the internet. Web 2.0 content that is user-generated on online platforms such as forums, social media, and crowd-sourcing frameworks allows practitioners to access the views of customers, employees, investors, and media [12, 25]. In today's digital world, e-commerce systems collect data that are less structured than that of transaction records, and often includes data of consumer behavioural information. With such data, there comes a need for new analytical techniques, and some of such techniques include association rule mining, database segmentation and clustering, anomaly detection and graph mining [1].

Government

Government and political processes are becoming more transparent in today's world, and the participation of the people online has allowed great opportunity for Governmental organisations to tap into the growing big data online. Social networking sites have proven to be a great ground for opinion mining and analysis, which can then support online political participation, forums analysis, e-government service delivery and process accountability [7, 9].

Healthcare

Patient points of contact and web-based communities have driven the amount of Big Data available on the internet currently. However, Gelfand [14] identified problems posed with extracting data from such databases such as regulations from the Institutional Review Board (IRB) due to privacy concerns. Additionally, its potential has not been fully taken advantage of also due to the lack of scalable analytical methods or computational platforms [24]. Two main sources of such data are genomics-driven data such as genotyping and sequencing, and payer-provider data which include electronic health and insurance records, prescriptions and customer feedback [24]. The potential of Big Data in the healthcare sector can help clinical decision making as well as building up vast knowledge bases for the several areas of healthcare [8, 29].

Security and Public Safety

Since the tragic events of September 11, 2001, security research has gained much attention, especially given the increasing dependency of business and our global society on digital enablement. Facing the critical missions of international security and various data and technical challenges, the need to develop the science of "security informatics" was recognized, with its main objective being the development of advanced information technologies, systems, algorithms, and databases for security related applications, through an integrated technological, organizational, and policy-based approach [6]. Table 1 summarizes these promising impacts and applications of Big Data analytics in the various industries.

Table 1. Use-Cases of Big Data Analytics in the Various Industries

E-Commerce	**E-Government**	**Healthcare**	**Security**
Recommender systems Social media monitoring and analysis Crowd-sourcing systems Social and virtual games	Ubiquitous government services Equal access and public services Citizen engagement and participation Political campaign and e-polling	Human and plant genomics Healthcare decision support Patient community analysis	Crime analysis Computational criminology Terrorism informatics Open-source intelligence Cyber security

METHOD

Data Collection

Based on case-studies discussed in academic sources, trade journals and industry white papers, we selected a total of 34 featured use cases where big data has been used to achieve various business goals. The selection of the cases is based on reported usage of big data analytics and the availability of the data. These use cases and their information have been taken from credible reliable third party sources, such as the online big data knowledge platforms like the Big Data Start-ups [2]. In addition to these single point sources of case studies, we referred to the multiple back links that were indicated in these articles and found credible sources to gather the data from. This ensures an unbiased account of the cases. For each case to be included in the sample, we collected data on the availability of company background, such as the industry, size and annual revenue etc. and big data activity was being undertaken, such as the characteristics of data and techniques and processes used based on secondary sources.

Data Coding and Analysis

Based on the unstructured nature of the data collected, we performed a content analysis on these cases to arrive on insights that help us obtain a deep understanding. To perform the content analysis, we first defined the content types that we will be focussing on. Then we apply content segregation methods to bring more structure to the data in order to identify patterns and insights. To start, we identified the background of every single use case. This involved the company name, the industry and the need for which the big data activity was being undertaken. We also researched and gathered the basic company information like annual revenue and the number of employees to understand the company background. We then looked at the characteristics of the data which was being collected, treated and analysed. Chang et al [4] in their study of paradigm shift of research methods to computational social science in the presence of big data discussed the ways in which data sets can be segregated during collection of data for analysis in a high volume, high velocity and high variety environment. They introduced the concepts of data type and data spectrum in order to better understand the environment in which we are analysing the data and the characteristics of the data. We then analysed the techniques and the tools used for each big data analysis exercise. Finally, we studied the barriers and tried to identify the insights that we could gather from each use case. The basic information can be found in Appendix A.

Thus the **content types** that we focussed on for data collection were as follows:

- Company name
- Annual revenue of the company
- Total number of employees
- Industry
- Project need
- Project start date
- Solution
- Benefit
- Data entity
- Data type
- Data spectrum
- Dominating data characteristic – Volume / Velocity / Variety
- Techniques used
- Tools used
- Barriers
- Insights

The coding of each content type was executed as follows. Annual revenue, employee count and project start date are measured by ratio scale which did not require any further segregation. Due to the incremental nature of big data projects where the projects often start with small pilot activations and then transform into major projects, the exact date of start of a project is often not specified. In these cases, we either looked at major events that signalled the start of project (eg: migration of all company data into a Hadoop cluster) or dates for which the earliest records of the project is shown.

We do note that the concept of data-type and data-spectrum has not been universally identified to segregate data sources, often these data characteristics are not mentioned in the case-studies. In these events we made our assumptions from the material available and our initial understanding. Thus the identification of data-type, data-spectrum and dominating characteristics of the data is based on our knowledge and understanding of the respective cases and data. In terms of project needs, we classified them into three broad categories: customer service, improve process and derive insight. These categories are developed based on the cases collected. The solutions and benefits have been combined into Project Outcomes, which are further segregated into four broad categories. The outcome of a typical big data project is either a product/application, reporting or visualization of data points, or an informed strategic decision. These are the three categories in which the outcomes are segregated.

To satisfactorily understand and classify data, we have followed the classification suggested by Chang et al [4] in their paper which describes data through two property types – data-type and data-spectrum. Data type is the attribute assigned to the data entity to describe it under varies environments. These types can be contextual, spatial or temporal. Contextual data is the data that is generated in context with an environment. Its data has valuable meaning only in the presence of the environment it was generated in. Spatial data is the data in which the geographical location of its activity is a crucial property. Temporal data involves time as the assisting attribute. Data spectrum is a property indicates the span of the data. They can either be micro, meso or macro-data. Micro-data has the highest level of granularity which result from technology mediated human action in social or machine settings. For instance, individual tweets from a person can be called micro-data. Meso-data have a higher level of aggregation than micro-data. Meso-data may indicate the statistics related to the tweets of a user, or an overview of all the tweets concerning a particular subject or a hash-tag. Macro-data has the highest level of aggregation. It may indicate properties of a geographical area, industry or even an economic sector. Macro data on the forest area in a country will give us an account of the ecosystem balance in the region.

For the purpose of data coding on Big data analytical techniques in the current research, we needed to use the most appropriate analysis technique categories which were broad enough to include most of the techniques in the sample and focussed enough to give us a clear direction on which techniques are to be listed under which category. For this purpose, we adapted from AT Kearney's broad categorization of big data in their white paper big data models as summarised in Table 2 [15].

Table 2. Analytical Capabilities to Realize Big Data's Potential

Analytical Process Maturity		
Anticipatory	Segmentation analysis	What are the unique drivers?
	Statistical analysis	What is happening?
	Sentiment analysis	How do they feel?
Predictive	Simulation	What would happen if…?
	Predictive modelling	What could happen next?

The coded data and data evidences are provided in the Appendix B.

RESULTS AND DISCUSSION

Our sample consists of 34 companies from a range of diverse industries and sub-sectors, including manufacturing, telecommunications, retail, healthcare, hospitality, public services, entertainment, finance, food & beverage (F&B), and so on.

The annual revenues range from US$ 224.5 million to 465 billion, with two of the companies' annual revenues exceeding US$ 450 billion (Shell and Walmart). 50% of companies in the sample have an annual revenue that is lower than US$10 billion. The numbers of employees range from 122 to 2.2 million. About 50% of the companies have more than 45000 employees. The earliest start date of the big data analytic projects is 2001 and the latest one started in 2012. Particularly, nearly one quarter (23.5%) of the projects from the sample started in 2012, indicating that big data analytics is starting to pick up its momentum in recent years. The insights that we have derived from the data are based on the patterns identified through descriptive statistics and visualizations of the coded data. The pioneers that initiated data analytical projects prior to 2006 include Catalyst IT, Amazon, Mastercard and Time Warner Cable, all of which are the major players of their own industry. Three out of the four pioneering firms are in the information and communications technology (ICT) related industry, which might have provided them with the necessary technological infrastructure and human resources for data analytical projects.

Project Needs and Outcomes

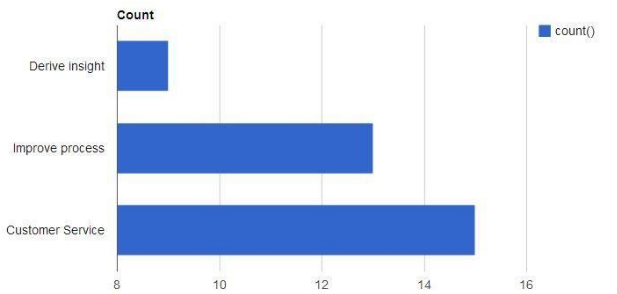

Fig 1. Number of Companies with Different Project Need

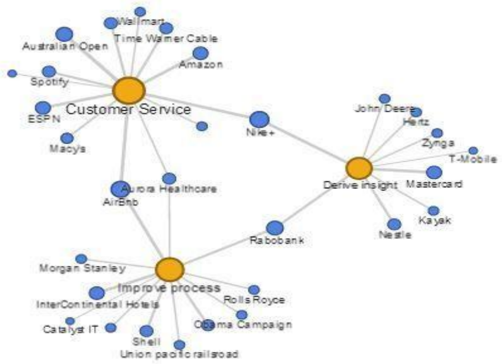

Fig 2. Company – Project Need Relational Graph

Except for four cases (AirBnb, Aurora Healthcare, Nike+ and Rabobank), all of the other cases focus on one single project need. Out of the use-cases studied, close to half of the projects (15 cases or 44.1%) are geared towards improving the customer services (see Figure 1). Through analysis of micro data gather at every customer touch-point, the companies can equip the customer service executives with relevant information about the customers and/or provide personalized responses to the consumers. The second most popular project needs is process improvement, with 13 cases or 38.2% of the companies adopting such needs. When the data describing a complex organizational, service-delivery, manufacturing, or continuous or batch manufacturing environment are made available, modeling and simulation can empower companies in conducting "what-if" analyses under variable resource configurations, and provide optimised performance solutions to operational or other business decisions. The remaining cases use data analytics to generate general business insights. The project needs of AirBnb and Aurora Healthcare consist of both customer service and process improvements. Nike+ used the date for both customer service and deriving insights, while Rabobank used its data for both process improvement and deriving insights (see Figure 2).

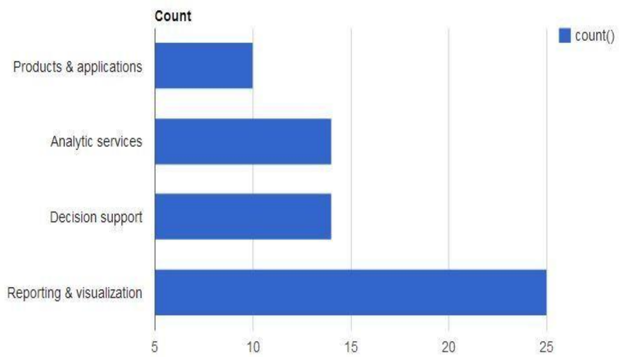

Fig 3. Number of Companies with Each Project Outcome

Fig 4. Project Need–Project Outcome Network Graph

Although most of the projects started with a single need in mind, majority of them (70.6%) finished with multiple project outcomes. As shown in Figures 3, reporting and visualization is a dominating outcome of a big data project. Though this may seem like a direct outcome to the need for providing customer support, reporting and visualization are also used by internal teams to make sense of data that the company is harvesting. Beside reporting and visualization, decision support and providing analytical services are two other outcomes mentioned by 14 cases. Lastly, out of the 34 cases in the sample, only ten analytical projects directly benefited their products. As reflected in Figure 4, each need of a big data project is connected with multiple outcomes and vice-versa. This indicated there is no fixed outcome for a particular project need. For example, the most popular project need, customer service, is related with all four categories of project outcomes, with the strongest linkage with data reporting/ visualization and products.

To summarise, while only four organisations started off with multiple needs, a majority of organisations end up with multiple outcomes. This indicates that though organisations may initiate a big data project with a limited set of goals, the projects often result in seemingly unrelated outcomes. This may stem from our previously made observation that the needs and the outcomes are cross-linked.

Analytical Techniques, Data-Types and Data Spectrum

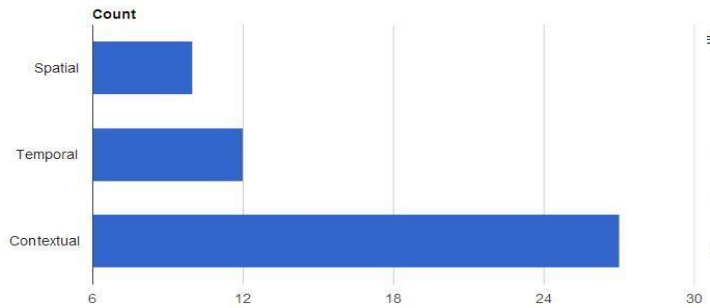

Fig 5. Data-types

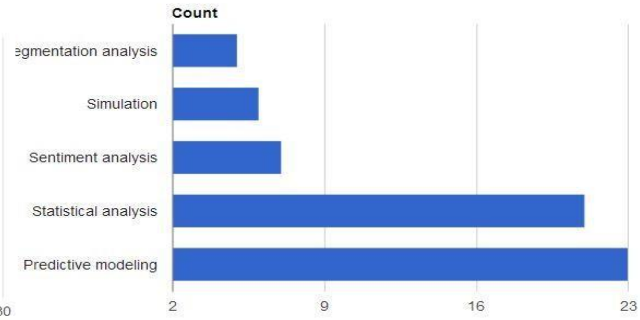

Fig 6. Big data analytical techniques

Predictive modelling and statistical analysis are the most prominent analytical techniques because of their wide range of applications. Contextual data is the most commonly harvested Data-Type. This may be due to the fact that contextual data are more granular in nature and thus they may create a higher number of record sets. This increases the likelihood of big data sets being formed out of contextual data.

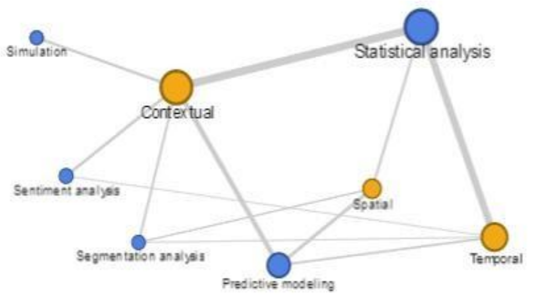

Fig 7. Analysis Technique – Data Type Network Graph

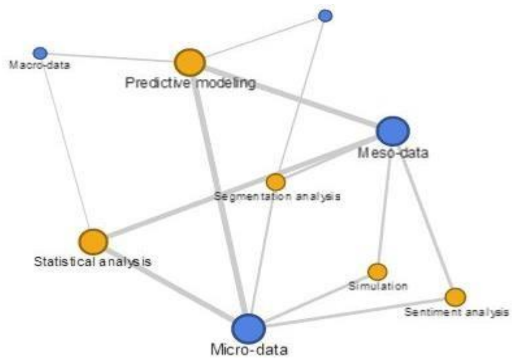

Fig 8. Data analytical Technique – Data Spectrum Network Graph

As shown in Figure 7, simulation is the only technique that was applied to a single data type – contextual data. All the other techniques were applied to multiple data types. For example, statistical analysis, segmentation analysis and predictive modelling techniques have been adopted for all three data types. Sentiment analysis was not applied on the spatial data type. In general, we may gather that big data analytical techniques are not data-type diagnostic. A similar argument may also hold when we consider data-spectrum with respect to data analytical techniques. It can be seen from Figure 8 that all analytical techniques have been applied to multiple data spectrums. In fact, techniques like predictive modelling and statistical modelling were applied to all three data spectrums: macro-, meso-, and micro-data. Segmentation analysis and sentiment analysis may not process macro-data as shown in Figure 8. Yet, the results show that they are applicable to the other two data spectrums, namely, meso- and micro-data.

CONCLUSIONS

The content analysis of the various cases has brought to our attention some network characteristics of various attributes of the big data ecosystem. To summarise, our findings have highlighted four important insights. Firstly, goals of a Big Data Initiative are *extensible*. Efforts to achieve a single goal in a big data initiative may lead to multiple benefits. A particular set of data can deliver many insights and this can be attributed to the fact that data is application diagnostic. For example, while Disney is using spatial data harvested from the wristband to deliver customised service to the visitors, the same data can be used by Disney to detect traffic bottlenecks. Second, *Representing Processed Data* is as important as Effective Analysis. One of the most common outcomes of a big data initiative is enhanced reporting and visualisation of the information that the processed data delivers. This goes to point out that no matter how sophisticated the analysis of a given data set is, it will be of little use unless the results of the analysis is appropriately represented. This puts considerable emphasis on data visualisation technologies and techniques. Third, the Big Data analytical techniques to be used are heavily *dependent on project goals*. Data analytical techniques to be used may also depend on the kind of data that is being treated to a certain extend. Lastly, Big Data analytics are suited to data of higher granularity and lower levels of aggregation. A majority of the big data initiatives that we have studied have made use of contextual data-type and micro-meso-data spectrum. These kinds of data have lower levels of aggregation and are suited to big data processes since the number of data records tend to be higher for these data types.

As data analytical techniques are intrinsically linked with the project goals and data that is being harvested, it is essential for companies to have a thorough understanding of the decision making scenarios and the data before proceeding to select the data analytical techniques. To realise the full potential of big data analytics, managers should harvest granular data-types that are more prominent in various business processes and apply the findings in an age of customisation and advanced decision making. It should also be highlighted to the higher management in the organisations that big data can be used to not only solve problems, but also identify issues and derive insights to help make strategic decisions for long-term business growth.

REFERENCES

[1] Adomavicius, G. & Tuzhilin, A. (2005) 'Toward the next generation of recommender systems: A survey of the state-of the-art and possible extensions', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 6, pp. 734-749.

[2] Big data best practices, [Online], available at http://www.bigdata-startups.com/best-practices/.

[3] Bloomberg Businessweek. (2011) *The current state of business analytics: Where do we go from here*, Bloomberg Businessweek, available at http://www.sas.com/resources/asset/busanalyticsstudy_wp_08232011.pdf (accessed September 7 2015).

[4] Chang, R. M., Kauffman, R. J., & Kwon, Y. (2013) 'Understanding the paradigm shift to computational social science in the presence of big data', *Decision Support Systems*, Vol. 63, No. 0, pp. 67-80.

[5] Chaudhuri, S., Dayal, U., & Narasayya, V. (2011) 'An overview of business intelligence technology', *Communications of the ACM*, Vol. 54, No. 8, pp. 88-98.

[6] Chen, H. (2006) *Intelligence and security informatics for international security: Information sharing and data mining*, Springer, New York, NY.

[7] Chen, H. (2009) 'Ai, e-government, and politics 2.0', *IEEE Intelligent Systems*, Vol. 24, No. 5, pp. 64-67.

[8] Chen, H. (2011) 'Smart health and wellbeing', *IEEE Intelligent Systems*, Vol. 26, No. 5, pp. 78-79.

[9] Chen, H., Brandt, L., Gregg, V., Traunmuller, R., McIntosh, A., Dawes, S., Hovy, E., & Larson, C. A., Eds. (2007) *Digital government: E-government research, case studies, and implementation*. Springer, New York, NY.

[10] Chen, H., Chiang, R. H. L., & Storey, V. C. (2012) 'Business intelligence and analytics: From big data to big impact', *MIS Quarterly*, Vol. 36, No. 4, pp. 1165-1188.

[11] Cukier, K. (2010) *Data, data everywhere: A special report on managing information*, Economist Newspaper.

[12] Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011) 'Crowdsourcing systems on the world-wide web', *Communications of the ACM*, Vol. 54, No. 4, pp. 86-96.

[13] Foo, S., Sharma, R., & Chua, A. (2015) *Knowledge management tools and techniques*, 3rd ed. Prentice Hall, Singapore.

[14] Gelfand, A. (2012) 'Privacy and biomedical research: Building a trust infrastructure—an exploration of data-driven and process-driven approaches to data privacy', *Biomedical Computation Review*, Vol. ,Winter, pp. 23-28.

[15] Hagen, C., Ciobo, M., Wall, D., Yadav, A., Khan, K., Miller, J., & Evans, H. (2013) *Big data and the creative destruction of today's business models*, AT Kearney Publishing, Chicago.

[16] Howard, C. E. (2012) 'Dealing with big data', *Military & Aerospace Electronics*, Vol. 23, No. 12, pp. 18-23.

[17] IBM. (2011) *The 2011 ibm tech trends report: The clouds are rolling in...Is your business ready?*, IBM, available at http://www.ibm.com/developerworks/techntrendsreport (accessed September 7 2015).

[18] International Data Corporation (IDC). (2012) *Worldwide big data technology and services 2012–2016 forecast*, International Data Corporation (IDC), available at http://www.idc.com/getdoc.jsp?containerId=238746 (accessed September 9 2015).

[19] Isik, O., Jones, M. C., & Sidorova, A. (2013) 'Business intelligence success: The roles of bi capabilities and decision environments', *Information & Management*, Vol. 50, pp. 13-23.

[20] Konrad, A. (2013) *Ayasdi scores $30m series b round as citigroup and ge look for answers in big data mapping*, Konrad, A., available at http://www.forbes.com/sites/alexkonrad/2013/07/16/ayasdi-series-b-big-data-mapping/ (accessed September 5 2015).

[21] Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011) 'Big data, analytics and the path from insights to value', *MIT Sloan Management Review*, Vol. 52, No. 2, pp. 21-32.

[22] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011) *Big data: The next frontier for innovation, competition and productivity.*, Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H., available at http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (accessed September 6 2015).

[23] McAfee, A. & Brynjolfsson, E. (2012) 'Big data', *Harvard Business Review*, Vol. 90, No. 10, pp. 61-67.

[24] Miller, K. (2012) 'Big data analytics in biomedical research', *Biomedical Computation Review*, Vol. ,Winter, pp. 15-21.

[25] O'Reilly, T. (2005) *What is web 2.0? Design patterns and business models for the next generation of software*, O'Reilly, T., available at http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (accessed September 6 2015).

[26] Ross, J. W., Beath, C. M., & Quaadgras, A. (2013) 'You may not need big data after all', *Harvard Business Review*, Vol. 91, No. 12, pp. 90-98.

[27] Sharma, R. & Bhattacharya, S. (2013) 'Knowledge dilemmas within organizations: Resolutions from game theory', *Knowledge-Based Systems*, Vol. 45, pp. 100-113.

[28] Sharma, R. & Mokhtar, I. A. (2005/2006) 'Bridging the digital divide in asia - challenges and solutions', *International Journal of Technology, Knowledge & Society*, Vol. 1, No. 3, pp. 15-30.

[29] Wactlar, H., Pavel, M., & Barkis, W. (2011) 'Can computer science save healthcare?', *IEEE Intelligent Systems*, Vol. 26, No. 5, pp. 79-83.

[30] Wolf, G. (2013) 'Big data unleashes more capacity', *Transmission & Distribution World*, Vol. 65, pp. 8-11.

APPENDIX A COMPANY PROFILES

Company Name	**Industry**	**Data entity**	**Annual Revenue (in Millions)**	**Number of employees**
AirBnb	Online / Travel	Multiple	1,000	600
Amazon	Electronic commerce	User behavior	74,450	117,300
Aurora Healthcare	Healthcare	Financial data, lab data, pharmacy data, procedure data	4,344.5	25,087
Australian Open	Sports event	Plays' historical performance and social networking websites	N/A	N/A
Catalyst IT	Recruiting/ Technology outsourcing	Job applicant data, real-time process data, domain public data, interaction data	N/A	N/A
Ceasars Entertainment	Hospitality	Transaction data, customer preference data, surveillance data, insurance data	8,830	70,000
Crowd control management	Public services	local tweets; Geographical Information; smartphone broadcasts	N/A	N/A
Disney	Entertainment	Real-time location data, transaction data, customer preferences	45,041	175,000
ESPN	Media / Entertainment	web traffic, fantasy league, game statistics,	10,975	7,000
Hertz	car rental	customer survey (website, email and other messages)	2,436	29,350
InterContinental Hotels	Hospitality	Transaction data, guest data, external partner data	1,903	120,000
John Deere	Manufacturers of agricultural machinery	Historical and real-time data regarding weather prediction, soil conditions, crop features and many other data sets.	26,005	55,700
Kayak	Online / Travel / Meta	Search query data, airline scheduling, availability	224.5	133
Macy's	Retail	Customer data	9,370	175,000
Mastercard	Retail / Finance	Transaction data	7,391	8,200
Morgan Stanley	Finance	Market movements and statistics. Public & social data, wire data	36,840	55,794
Morgan Stanley Smith Barney (MSSB)	Finance	Market movements, social media data, industry data	13,423	17,649
Nestle	Food & Beverage	Social media and online interactions regarding the brands	92,000	333,000
Nike+	Retail/Sports	Consumer input, vendor data, internal process data	25,300	44,000
Nordstrom	Luxury retail	Web traffic, point-of-sale data, user data (fb fans), spending behavior	8,770	52,000
Obama Campaign	Election	everything	N/A	N/A
Purdue University	Education	Student academic preparation, session data, past performance records	1,263.4	18,872
Rabobank	Finance	Internal data, social data, Internet data, public data	2,779.8	56,870
Rolls Royce	engine and power systems	components, systems or sub-systems (pressure, temperature, speed)	15,505	55,000
Shell	Energy	sensor data, technical data	451,235	87,000
Southwest Airlines	Airline	Multiple	16,790	45,000
Spotify	Music	Music listening behavior, email response tracking	576	1,200
Time Warner Cable	Telecommunications	public data sets and local viewing	22,100	51,600

T-Mobile	wireless network operator	customer data, product and service data, customer experience data, business operations data, supply chain, and network data	24,420	38,000
TomTom	manufacture of automotive navigation systems	Speed limits, new street names, blocked roads, new traffic directions and altered traffic directions. Trip information	1,460.4	3,500
Union Pacific Railroad	Rail transportation	Temperature, acoustic and visual data	20,926	45,400
US Xpress	Time-definite truckload service	petrol usages, tyres, brakes, engine operations, geo-spatial data and driver comments	1,540	10,885
Walmart	Retail	Public data, social data, transaction data	465,294	2,200,000
Zynga	Entertainment/online game	Game data, server data	873.3	2,034

APPENDIX B CODED DATA AND DATA EVIDENCES

Company Name	Project Start date	Coded project need(s)/ goal(s)	Data evidence on project needs	Data-type	Data-spectrum	Volume/ Velocity/ Variety	Techniques used	Coded project outcome(s)	Data evidence on project solutions	Data evidence on project benefits
AirBnb	2009	Improve process Customer Service	Optimize resource allocation & sales	Contextual, Temporal (website optimisation)	Micro-data, Meso-data	Variety, Velocity	Regression analysis, predictive modeling, segmentation analysis	Analytic services Reporting & visualization	Trend analysis of content and location based data	Optimize market rates, improve sales
Amazon	2003	Customer Service	Customize the webpages for customers	Contextual	Micro-data, Meso-data	Volume, Velocity	Predictive analysis, segmentation analysis	Reporting & visualization Decision support Products & applications	Collect all data from customers, use recommender system	Offer superior service, improve the relationship with customers; monitor, track and secure items.
Aurora Healthcare	2009	Improve process Customer Service	Improve decision-making and service standard, get nationally recognized measure of the clinical performance	Temporal, contextual	Micro-data, Meso-data	Volume, variety	Pattern recognition; Predictive analysis	Reporting & visualization	Created a hybrid business intelligence system combining 18 different streams of data for storage and analysis	Predict and improve patient outcomes and treatments; cost saving of $6mil; helping doctors to analyse the outcome and recommend different procedures
Australian Open	2012	Customer Service	Fans want to know all about their favourite tennis player	Contextual, Temporal	Micro-data	Variety, Velocity	Data visualisation, statistical analysis, sentiment analysis preventive scaling	Reporting & visualization Products & applications	Data is captured, analysed and shared in real-time on multiple platforms and multiple devices.	Full-interaction during the events; improved data reliability and server robustness
Catalyst IT	2001	Improve process	Speedup candidates screening process and maintain employee	Spatial, contextual	Micro-data, Meso-data	Volume, Variety	Data mining, pattern recognition	Decision support	The company came up with an algorithms for analysing applicant based on how an obligatory survey is completed. The analysis is based on data, not perceptions.	Employee turnover is brought down to only half of the market average.
Ceasars Entertainment	2009	Improve process	Improve service standard an bring personalized guests experience; improving employee performance and satisfaction	Spatial, temporal	Micro-data	Volume, variety	Data mining, pattern recognition	Reporting & visualization Decision support	Based on the Total Rewards program, guests are tracked through the journey and data were used to analyse and provide insights on their behaviour.	Better customer satisfaction; more efficient manpower deployment and better employee satisfaction
Crowd control management	2012	Customer Service	Ethe safety among all the visitors	Spatial, Temporal	Micro-data, Meso-data	volume, Variety, Velocity	Data mining, sensitivity analysis, information visualisation, statistical analysis	Reporting & visualization Analytic services	Using three tools to monitor the real-time pictures of the situation.	Find emergencies happening; a real-time situational awareness overview of a complete area; control the amount of people
Disney	2010	Customer Service	Understand visitors to the theme park, provide personalized	Spatial, Temporal	Micro-data	Volume, variety	Data mining, predictive analysis	Analytic services	Use wireless-tracking wristband to collect real-time data on every visitor; tied to	Better visitor analysis, traffic flow, personalized service

			services and tailor its marketing.						credit card and other service points	
ESPN	2012	Customer Service	Provide most relevant telecasts and news to viewers	Temporal	Meso-data	Velocity	Information visualisation, statistical analysis, sentiment analysis	Reporting & visualization Decision support	Over 50 data scientists analysing real-time data to provide relevance. ESPN API	Relevant stories and telecasts to the user
Hertz	2011	Derive insight	Keep track of customer touch points a day	Contextual, Temporal	Micro-data	Volume, Velocity	Sentiment analysis	Analytic services Reporting & visualization	Real-time analyses on the captured unstructured data	Understand real-time customer's opinion, improve service and retain customers
InterContinental Hotels	2010	Improve process	Better understanding of overal l organization's performance and	Spatial, Temporal	Micro-data, Meso-data	Variety, Velocity	Operational analytics, regression and correlation alaysis, predictive analytics	Analytic services Reporting & visualization Decision support	Big data analysis system with more than 650 variables on unstructured and structured data.	Real-time analysis on its hotels, guests and other internal/ external data
John Deere	2012	Derive insight	Help farmers increase productivity	Contextual, Spatial	Micro-data	Volume, Variety, Velocity	Predictive analysis	Analytic services Reporting & visualization	add sensors to equipment	Increase the productivity and efficiency of the crops that will in the end lead to higher production and revenue.
Kayak	2006	Derive insight	Derive insights from aggregated data	Temporal	Meso-data	Volume	Predictive analytics, A/B testing (to improve website and user experience)	Reporting & visualization	Predictive analysis and feedback measurement to provide travellers with insights	Offer rich user experience through data aggregation
Macy's	2007	Customer Service	personalized shopping experience	Contextual	Micro-data	Variety	Predictive analytics	Products & applications Decision support Analytic services	dynamic, data-driven and integrated website	customized content granular pricing market strategy
Mastercard	2005	Derive insight	Find consumer spending patterns	Spatial, Temporal	Micro-data	Volume, Variety	Statistical analysis, Predictive analysis, Segmentation analysis	Reporting & visualization Analytic services	Detailed customer segments from transaction data	Offer insights to merchants
Morgan Stanley	2011	Improve process	Traditional databases and grid computing was no longer supporting the vast data	Temporal, Contextual	Macro-data	Volume, Velocity	Statistical analysis	Analytic services Reporting & visualization	use Hadoop to store mass logs to discover problems and predict impact of events on the system	Hadoop helps with mission critical investment projects
Morgan Stanley Smith Barney (MSSB)	2010	Customer Service	Improve their recommendations for their investments in stocks, municipal bonds and fixed income	Temporal	Micro-data	Velocity, Variety	Predictive analysis	Decision support	Use of predictive analysis of big data to improve recommendations	All the information is used to recommendations regarding whether to buy or sell stock based on real-time positions and market conditions.
Nestle	2006	Derive insight	Understand the brand sentiment of multiple	Contextual	Micro-data	Volume, Variety	Data visualisation, sentiment analysis	Reporting & visualization	Digital acceleration team - a 24/7 monitoring center that	Understand the sentiment of consumers and take actions

			brands in real-time					Decision support	listens to all conversations about the products on SM	to avert reputation crisis
Nike+	2012	Derive insight Customer Service	Engage, understand and encourage customers to share and commit into running	Spatial, contextual	Micro-data, Meso-data	Volume, variety	Data mining, predictive analysis	Products & applications Decision support Reporting & visualization	Gamification platform Nike+ Accelerator to encourage users to share; centralized material-for-production database for designer; open data sharing to build "vendor index"	Coherent user community engaged to sharing and exchanging experiences; Nike could change behaviour of many consumers; better design ideas rooted from users.
Nordstrom	2012	Customer Service	Improve integrated customer experience through innovation via big data. Identify marketing strategies	Contextual, Spatial	Micro-data	Variety	Predictive analysis, data visualisation	Products & applications Analytic services	Nordstrom innovation lab Cross channel inventory	Increase in same store sales Identify products to market
Obama Campaign	2007	Improve process	Build an analyst driven organisation and an environment for smart people to freely pursue their ideas.	Spatial, Temporal, contextual	Micro-data, Meso-data, Macro-data	Volume, Variety, Velocity		Decision support	Divide campaign team into different channels; use MPP database	find the influences of this campaign, find the influences of this campaign
Purdue University	2011	Improve process	Predict academic and behavioural issues and notifies teachers/students when action is required	Temporal, contextual	Micro-data	Volume, Velocity	Predictive modelling, data mining	Reporting & visualization	Couse Signal - a platform that analyses the student's behavior and academic performance based on their activities and past records.	Risk profile of student's academic performance could be generated as early as 2nd week of the semester, with suggested materials/solutions to help students improve. Better academic performance was observed since its implementation
Rabobank	2006	Improve process Derive insight	Improve business process, understand customer, create new opportunities	Spatial, contextual	Micro-data, Meso-data	Volume, variety	data mining, pattern recognition, content analysis	Products & applications Decision support	Based on the business ICT forecast, the bank implemented big data storage and analysis in stages with precise goal in mind	Fast consumer analysis, real-time autocompletion of forms and feedback gathering
Rolls Royce	2010	Improve process	Whenever a small error is noticed, it can be corrected.	Contextual	Micro-data	Velocity, Variety		Reporting & visualization Decision support	All engines are equipped with sensors; All minor details are sent via satellite to a computer that analyses the data.	Early warning; minimizes disruption and delays for customers; make engines safer and more reliable
Shell	2008	Improve process	Improve operations and increase oil and gas output of the wells	Spatial, Temporal	Micro-data	Volume, velocity	Regression analysis, predictive modeling, information visualisation	Analytic services Reporting & visualization	With sensors laid into every well and a group of analyst working at the backend, the company could know the progress and output projection of the well.	3D or 4D map of the oil reservoirs; immediate visualization for analysis and reporting
Southwest Airlines	2008	Customer Service	improve customer service, air safety, optimise product offers	Contextual	Micro-data	Variety	psychographic analysis, predictive analysis	Analytic services	psychographics using big data analytics	Multiple business output optimisations
Spotify	2012	Customer Service	User of data like log messages to provide	Contextual	Meso-data	Volume, velocity	Predictive analysis, A/B	Reporting & visualization	An eco-system of multiple big data technologies to leverage	Creation of a billion dollar industry built on data

			music recommendations				testing - email communication effectiveness	Products & applications	the Spotify data	analysis and storage
Time Warner Cable	2005	Customer Service	To optimize viewer's experience	Audience metrics	Micro-data, Meso-data	Variety	Cross-platform analysis	Analytic services Products & applications	Combine public data sets and local viewing habits to measure user preference	Personalised advertising
T-Mobile	2012	Derive insight	store, analyse search and visualize data. Determine most valuable customer.	Contextual	Micro-data, Meso-data	Volume, Variety, Velocity	Sentiment analysis	Reporting & visualization	Using different data zones that are connected to business objectives	improve customer satisfaction and revenue
TomTom	2009	Customer Service	providing up-to-date mapping	Spatial, Temporal	Micro-data	Volume, Variety, Velocity		Reporting & visualization Products & applications	When users docked their SatNav their anonymized information was sent to TomTom.	improve the driving experience of customers
Union Pacific Railroad	2011	Improve process	Optimize operations	Temporal	Macro-data	Volume	Predictive analysis	Analytic services Reporting & visualization	Network of sensors connected by optical cables	Predict train derailment weeks before it occurs
US Xpress	2008	Improve process	Monitor in real-time; better control	Contextual, Spatial	Micro-data, meso-data	Volume, Variety, Velocity	Geospatial analysis, statistical analysis	Analytic services Reporting & visualization	On-stop: combine all different data streams into one interface	Improve efficiency and save money
Walmart	2011	Customer Service	Understand and engage customers, recommend suitable products via social media platform.	contextual, temporal	Micro-data, meso-data	Volume, variety, velocity	content analysis, data mining	Products & applications Reporting & visualization	In house products Social Genome and ShoppyCat could help better understand the customers' feedback and favored products and provide consumer intelligence for marketing	Walmart could optimize the local assortment of stores based on neighbourhoods' comments online; in-store mobile navigation could steer customers through aisles of products of their interest
Zynga	2008	Derive insight	Analyze the large amount of player data and provide insights for consumer intelligence	Spatial, contextual	Micro-data	Volume, velocity	Pattern recognition; Predictive analysis	Decision support	Matrix-driven data access and analysis on player's behavior, providing feedback for future development.	