The International Technology Management Review, Vol. 5 (2015), No. 4, 202-210

Emerging Trends in Technologies for Big Data

Nitin Singh

KPMG India
E-mail: nitinsingh@kpmgindia.com

Abstract

There is a need for more understanding of the technology in Big Data and researching key challenges from the information technology perspective. In this paper, we study the current status, challenges/issues and emerging trends in Big Data. We present key findings and unfolding landscape of technology. Over the next decade, we foresee interesting developments like Analytics-as-a-Service emerging as sustainable business models.

Keywords: Enterprise Data, Business Analytics, Information Management, Big Data, Hadoop.

1. Introduction

Digital data is increasing exponentially. It is estimated that world's data will amount to 44 zettabytes by 2020 as predicted by a study done by International Data Corporation (1 zettabyte = 1 trillion gigabyte).[1] This has come about due to vast deployment of automation and sensory devices which capture variety of data. Obviously enterprises would like to 'lever' the data and be benefitted by improving their key performance indicators like sales and profit.[2] To 'lever' Big Data enterprises require technology to record, store and analyze data. Evidently, Big Data seems to be big business. Challenges are there but they present opportunities too. Apart from the technology to 'lever' the data, talent is needed to use this technology effectively. It is estimated that there will be a 'talent gap' of around 140,000 – 190,000 in USA itself in Big Data.[3] A recent Gartner study suggests that only 13% respondents stated that they have Big Data deployments. In contrast, 75% respondents stated that they are investing in some way in Big Data. It points us towards the fact that a large majority are still exploring, and only a few are in implementation stage of Big Data projects. In fact, few years back, firms worked in a world where enterprise software was quite different from what researchers used. Now that that has changed and we have a completely different way of software development and style of building systems. Earlier, enterprises would run on large mainframes and software developers would work on individual workstations. It has changed now and we find enterprises are really adopting open source approach to development. Big Data technology has been benefitted utmost by this approach to development and it has opened new vistas for enterprises and software experts to create a collaborative ecosystem. Data is enabling enterprises to better understand themselves and customers. However, technology adoption cannot grow any faster than talents/skill which is trained to use it. Data is indeed driving growth in Big Data technology but that also depends on availability of talent.

International Data Corporation also suggests that by 2020, business transactions on Internet will increase to 450 billion transactions per day.[4] It is large volume of activity and it will, in turn, generate higher volumes of data. Obviously, such data can be used but we would more powerful ways of storing, extracting and analyzing this data in a faster and 'user-friendly' way.

One of surprising outcomes of Big Data paradigm is the shift of where the value can be found in data. Not far (till about 2005), inherent hypothesis was that bulk

Published by Atlantis Press

of value could be in found on structured data, which usually constitute about 20% of total data. Other 80% is unstructured in nature and was viewed as having limited value. This perception has changed. It was the analyses of that unstructured data that has led to click-stream analytics and predictions that search engines to now and launched much of Big Data movement.

In this paper, we outline emerging developments in Big Data which are having remarkable imprint on businesses. Our observations are indicative in nature and shed light on the trajectory these developments are taking. We also investigate how Big Data is helping enterprises in different domain, challenges, opportunities and the future in this field.

2. Literature Review

A review of literature suggests that heterogeneity of data sources that can potentially be exploited for analytics is perhaps one of the most daunting challenges facing researchers. The approach described by Patching et al involves integrating data from media reports, expert curated sources and official alerts with geospatial data for disease surveillance.[5] Most widely cited challenges of Big Data are pervasive in almost all industries and domain. The paper by Razavian et al develops a model that works on electronic health records. In their paper, they integrate Big Data using datasets sourced from electronic health records of 4 million patients over 4 years, represented by 42,000 variables. They develop a learning model which analyses Big Data to predict the onset of type 2 diabetes.[6]

Studies done by research organizations have reported that technologies on Big Data are continually changing. One such study conducted by Accenture reports that 92% of executives are satisfied with results by applying big data to their businesses.[7] It also reports that 89% executives feel big data as 'very important' to their businesses. More importantly, the survey also reports that executives feel that domain understanding is critical to the success in leveraging Big Data. Hence the executives must know what data matters and what doesn't. Knowing the domain, helps us create the right hypotheses which can be tested. Analytics and domain understanding complement each other so that we know where we are going. Internet and ERP technologies have created large data repositories and streams of data flows. It has spawned yet another field which has come to be known as Big Data. Unstructured data which consists of images, video, clickstream data captured through social media sites brings additional layer of complexity to analytical process. Image processing technologies for face recognition and data integration during geospatial processing are becoming the norm in processing unstructured data. If this is complemented with traditional analyses on structured data like financial and accounting, it becomes easy to see that deductions would be of great significance for enterprises and Governance.

Proliferation of large-scale data sets is beginning to change businesses. Let us interweave certain examples of real-life applications in few domains. Traditionally agriculture is practiced by performing a particular task, such as planting or harvesting based on certain schedule of seasons.[8] But, if real-time is collected and analyzed on weather, soil and air quality, crop maturity, equipment and labor costs and availability, smarter decisions can be made. Basically, the opportunity to identify patterns reaches far beyond single view of data-sets. Now, tools are available to slice and dice massive data-sets that will help in better agriculture management. Urban planning is also increasingly adopting Big Data in these terms. Consider the bus network system where buses need to carry passengers who wait a few minutes to be transported over a few kilometers. Collecting the data related to the time in between buses at each stop, possibly together with the number of passengers waiting, gives the planner the basis for a feedback control solution. Such data is analyzed to enforce desired standards of service, quickly place more or fewer units in service where these parameters start to deviate from the ideal metrics, and the quality of service as measured by per person waiting times will improve.

It is increasingly being found that data analytics could be used effectively in audit engagements thus leading to lower losses. It so happens that lax audit can actually trigger fraud incidents and so higher rigor in audit is sought by regulatory bodies.[9] In this context, Data analytics could make it possible for external auditors to improve audit by testing he whole of accounting data rather than just samples. Better risk control is also achieved through data analytics through identification of anomalies and trends pointing auditor toward items they need to deep-dive. For example,

annual accounting data in many firms for a year like General Ledger (GL) and Invoices, Goods Receipt Note (GRN) etc. is huge and falls under the category of Big Data with line-items (records) going upwards of hundreds of millions. In this case, Big Data technology can be used to gather better audit evidence through comprehensive analyses of accounting data. Public Company Accounting Oversight Board (PCAOB), a premier regulatory body for audit has also cited the importance of audit and has urged audit companies to leverage data analytics in audit process. Audit teams usually face challenges in obtaining timely, complete and accurate data. Another challenge is to obtain relevant data. Consistent issues normally encountered related to multiple legacy systems, irrelevant data, data quality, multiple format and limited technologies and even skills to download required data. More importantly, audit teams also need to overcome the hurdles of skill sets and embedding data throughout the audit cycle.

Another area of interest that has emerged through leveraging Big Data is that of Genome analysis. Genome data is huge and computational analysis of this data used to take years without the help of Big data technology.[10] individual DNA analysis of genomic data is now available at $1000. Bioinformatics companies are now able to sequence an individual's genome in a matter of few days. This is indeed an emphatic development given that it used to take years before the emergence of Big data technology. Such a development has pave the path for improved health diagnosis and personalized health care.

Online stores like Ebay heavily depend on user profiles for personalized product recommendations to make a targeted sales pitch to customers. Better user profiles can turn into higher Click Thru Rate (CTR) for the advertisements.[11] A user profile is a measure to classify a given user into a user segment to capture his/her online behavior. To create user profiles, large data sets are analyzed which include user's registration preferences online, his/her interactions on social networking sites, log files from search engines and user's clickstream data during browsing sessions.[12] However, the task of consolidating data from various sources and creating coherent datasets suitable for meaningful analyses is a challenge. The next challenge is to apply right analyses that will get correct insights to be used for business.

3. Research Method

The objective of this study is to explore emerging trends in trend process including tracking emerging trends in Big Data. Because, enterprises and open-source development have different resources, constraints and ways to gather and apply trend information in order to survive in competitive market, another approach is to fund and compare suitable options for Big Data technology development in open source and technology adoption in enterprises. Cases were collected from 30 large enterprises in Fortune 100 list and trend research organizations like Gartner and International Data Corporation, academic research in literature and technology development enterprises. Data and Information was collected and analyzed directly from the Internet and overall content analyses of information available in public domain through above mentioned enterprises and research organizations. In specific cases, databases and reports from research organizations like Gartner and International Data Corporation was analyzed to draw conclusions about current state of practice, issues and emerging trends in Big Data. We made a conscious endeavor to cite selective and relevant enterprises and literature that is core to the area we are studying. We could interweave challenges and opportunities in Big Data technology, current state of practice and emerging trends therein through this method and conducted specific content analyses of information,

4. Overview of Emerging Technology

Many analytic applications depend on the user interaction clickstream data. Such data can be aggregated to build customer profiles which, if grouped provide valuable insights into user behavior. However, there are large number of users and high rate of interactions and so maintaining profile clusters has high processing and memory resource requirements. Mehmet et al apply distributed stream processing on telecommunications calling data.[13] In the presence of distributed state, they also report major challenge to partition the profiles over nodes such that memory and computation balance is maintained, while keeping the clustering accuracy high. Distributed processing while maintain memory and computation balance is a major challenging space where technology development

centered. We have seen adoption of Big Data technology in complex and large scale projects like security and National Intelligence usually funded by Governments. Commercial deployments have also begun to happen in social media, ad delivery, accounting audit, supply chain management and others. A key development has been in the area of NoSQL Database Management Systems (DBMS).[14] The capacity of DBMS to provide more storage and access has improved over the year and, in turn, has given a fillip to Big Data analytics software Hadoop. In this endeavor we are discussing below key technologies and what use they are being put to. It is not exhaustive list but the explanation of how different technologies in Big Data work in tandem to deliver Analytics. Hadoop was developed as an open source and was first used for large scale data analyses at Yahoo and Facebook.[15] Faster data analysis is done by breaking a set of data into smaller sets of data where individual elements are broken down into pairs.[16] Advantage is that processing is happening in parallel (during the sub-setting stage) in the same machine or could be in different machines. Evidently, it saves lot of time and could be done on shared servers or simple computers. Interesting developments have also taken in mining and analyzing Big Data in social networks. To study social networks which is an important source of big data, HYDRA is developed which captures real data from different social networks platforms and links user's different accounts from different social networks platforms.[17] This is very beneficial for firms in online business since it can work with different information sources (profile and social structure) and different time spans.

4.1. Seamless data connectivity

Since data resides across platforms, it is a hugely daunting task to pull it from diverse platforms and bring into one place before any meaningful analyses is undertaken. Many companies have grown over the years through mergers, acquisitions, partial buy-outs and various other ways and so they carry different kind of legacy IT platforms. Companies require that data discovery, assembly and transformation of data are possible across platforms before running analytical processes. Such a possibility increases the efficiency of data scientists. They do not need to spend efforts in moving data between platforms (importing, appending and consolidating). Instead they can visualize it through various charts, histograms etc. without data movement. New technologies are allowing this to happen which is a key competitive niche.[18] Spatial and temporal data has grown by leaps and bounds and companies endeavor to mine it and get insights on their business. To address such issues, frameworks like TODMIS have been developed which combines additional information with raw trajectory data and creates multiple similarity metrics. Data analyses through such framework, for example, can help to monitor customer trajectories in a shopping mall and city-scale taxi movement data. Pattern matching on Big Data is also being researched and newer protocols are being developed. It is a well-known problem and extensive research has been conducted for performing effective and efficient search. In a mobile environment (e.g., mobile phone networks), one person's pattern could be separately stored in a number of different stations, and such a local pattern is incomplete compared with the global pattern. Communication efficient and search effective solution are being developed to address this issue.[19] A key beneficial attribute is that there are easier ways for data scientists and IT specialists to collaborate on analytics problems. It reduces time and brings more meaningful insights and business value to customers requiring analytical solutions.

4.2. Quantum computing

Research on quantum computing has implications for Big Data. There has been developments happening in the area of quantum computing but its commercialization would take time. In fact, its impact would be a game changer in all parts of life and not just on Big Data. Be it instrumentation, automobiles, mobile devices, control systems, hand-held devices of all kind. We have heard quite a bit of it and people in business conclaves and conferences touch upon it but this will take time. Quantum computing uses a different concept altogether which is faster and error-free if it becomes feasible and practical.[20]

4.3. In memory processing

There is also a movement towards making the analytic engine more user friendly so that data scientists can slice and dice data in a variety of ways and produce meaningful analyses from terabytes (and sometime petabytes) of data easily. Another development that has taken place which has really pushed Big Data analytics

in commercial use is that of 'in-memory' processing. We now have more powerful processors which are affordable and also carry more storage capability.[21] Technology is now allowing them to put specific data into memory for faster processing. For example, SAP's HANA, (High-Performance Analytic Appliance) is an in-memory, column-oriented, relational database management system.[22]

5. Highlights of Key Findings

It is clear that big data is a robust technology that is being used by enterprises and which avails itself to various analytical tools to generate insights for big data. Tis section examines the research findings from the previous sections to illustrate what can be gleaned from big data. Three major highlights includes issues of big data quality, data storage and analysis, and data privacy. In the next subsections, we discuss these findings with a reference to opportunities and challenges that they post.

Big Data is receiving wild popularity that is going exponential with the advent of newer tools to capture both structured and unstructured data. However, in the middle of all this fanfare, certain important challenges get masked – technical, operational and financial. As we look ahead to the next decade, we attempt to make a trajectory of technology challenges and relevant advances that are going to take place to handle those challenges.

Often Data management systems don't talk to one another in many cases. A case most cited is that of online behavior coupled by offline buying process. A prospective customer may start the shopping process online by registering, browsing, searching etc. but calls the toll-free number for assistance in purchasing process. Here, the link with online analytics goes away. Analytics done on the online data will have to report real-time so that offline behavior can be influenced too. Basically, data points from offline interactions have to be appended to online data so that companies have appropriate customer IDs to approach the customer on any given channel. To handle these challenges technologies are being developed as well. Data quality, data storage & access and data privacy present challenges as well as opportunities. There are several vendors who are addressing these challenges and developing solutions. We are putting below the relevant issues within each of these below and the solutions therein.

5.1. Data quality

The amount of data that we generate and work with daily keeps increasing in volume. Companies intend to keep this data clean so they can get the most value out of its analysis.[23] For example, we can imagine how hard it would be to perform aggregations on sales per vendor if the vendor is identified by name, and most of the vendors are entered with several different spellings in their name, depending on the data source. Here is where Data Quality Services come in, as a necessary part of Extraction, Transformation and Loading processes working with Big Data.

Data Quality Services applications like DQS of SQL Server program allows rapid data cleansing and improving data quality.[24] For example, individual with incorrectly spelled names in different sources of data can be identified thereby appropriate demographic profiling. If we wish to analyze large data sets, it could be possible that data is extracted from different sources – ERP, Point of Sales, Radio Frequency Identification, Biometric data capture systems, telecommunication data systems etc. In such scenario, data is in variety of format and complexity. Before any meaningful data analyses can happen, the challenge is bring it all together in a uniform format. For example, data collected through application devices such as mobile phone are huge source of outlier or burst detection. Such anomalies in data are handled through data mining and knowledge discovery.[24] Then we have the challenge of handling missing values and data import which has also been cited as one of the top five big data research issues.[25] In various cases, data is not fully captured when it is being entered manually.[26] For example, customer IDs or product IDs, if manually entered, can be incorrect sometimes. These have to be corrected before analyses. Data also needs to be consolidated since it may come in pieces. To cite another case accounting data on invoices may be coming in parts i.e. for each week, month or quarter. It will require appending of multiple line items into a single consolidated database. Software vendors have identified this as a mine of business opportunity and created solutions around it. This process of data cleansing takes huge effort and sometimes costs 50% - 60% of time in a Big Data Analytics assignment.

Indeed, business process and application depend on clear, consistent data. However, there could be several ways the data may not be suited to business needs. For example sales service representatives might misplace values among various fields. Consumers can also submit incomplete records via web services of E-Commerce platforms. Software vendors have come out with solutions for data cleaning and standardization. For example, customer's data is changed and standardized for names and address. Likewise, product data is changed & standardized for product codes, brands, model number, catalog numbers etc. Few names in this include the likes of SQL DQS, Trillium Systems and others are entering in this space and few more already operating.[27, 28, 29, 30] We are not mentioning all of them here since the objective is not to provide an exhaustive review of software but to highlight emerging trends in Big Data. Indeed, the importance of Big Data initiatives in terms of analytics and the business value is a given. However, Big Data projects have to effectively interweave both Data Governance and Data Quality in the implantation phase. Data Quality Services (DQS) is a fairly new part of SQL Server (available in Enterprise, Business Intelligence and Developer editions since SQL Server 2012),[31] which performs the tasks of monitoring and maintaining the new coming data in good condition.

5.2. Data storage

Another challenge that organizations usually face is: how to shorten the time from data collection and business decision? Machine learning-based systems working on Big Data are helping here in big way. For example, Big Data analyses software like Yarn and SQL Server Information Systems can plug into Big Data and run customized queries to build reports quickly.[32] Companies are expecting that reports on Big Data to be embedded in cloud with a deployment period running into few weeks only with the onset of digital revolution,. Big data implementation are becoming foster with the help of various platform and not alone like itself recognized as an Ecosystem: consisting of web console, data mining, job workflow and scheduling, Analytical language, column NoSQL Database& HDFS, commonly known as a Zookeeper,[33] which is a centralized service for maintaining and Configuration information providing distributed synchronization. A set of tools to build distributed applications that can safely handle partial failures. ZooKeeper ensures coordination between various applications within Hadoop ecosystem.[34]

A solution for large scale data processing, storage and distribution is that of cloud computing. Cloud computing provides on-demand, pay-as-per-use computer and storage facility. Traditional WAN based transport method could not move strategies of data at the speed dedicated by business. Companies can now leverage On-Demand technologies to meet their Big Data infrastructure needs. One of the most interesting trends in the computer world during the past few years has been the rapid growth of NoSQL databases. NoSQL databases don't use SQL in order to store and retrieve data, but that's about where the commonalities end.[35] Companies are using NoSQL database management systems extensively to store their data, in particular unstructured data (audio, video data). For example, Cisco UCS and Oracle NoSQL database solution can integrate traditional oracle database on Cisco UCS with infrastructure for deploying NoSQL database. The computer nodes, cluster inter connects and storage access all work together and are managed from a single management domain. User companies can partner with instantly storage vendors for their storage needs.

5.3. Data privacy

We come across several occasions where we are bombarded by product promotions and advertisements. Companies call it enhanced user experience which is an offshoot of predictive analytics that levers customer data that is shared, tracked and monitored. This data is captured by plethora of information vendors in exchange for free opinions, applications, devices and free use of Internet. In a way, predictive analytics is being used to know exactly where customer is going to be, what he will say or do next. There are issues of data privacy and governance.[36] Customers wish to have enhanced user experience (more targeted sales, better discounts, promotions, offers, better deals etc.) but they also need to reconcile it with privacy that they desire.[37] More efforts are also put towards monetizing 'incidental' data sets. Data like Point of Sales data which is collected in retail stores, restaurant and variety of scanning and biometric devices is also very valuable to companies. Such data is being collected in the routine course by sensors (scanning devices at airports, railway stations, retail outlets etc.) and holds immense

value for companies. Herein data privacy permissions are to be worked out too since the datasets hold private information about users (customers). This data is collected, stored and sold by third parties.[38]

Regulators, legislatures, interest group and citizens have begun to voice concern about the impact of big data on privacy- from the misuse or unauthorized use of personnel information and surveillance. Many firms use privacy techniques and principles, like 'Privacy by Design' to protect personal information while driving innovation at the same time 'Privacy by Design' is an approach to protecting privacy by embedding into the design specifications of technologies, business protection and infrastructure.[39] Data privacy is complicated by cross-border transfer issues and the difference in privacy laws around the world. These laws are complex and pose myriad obligation to multinational enterprise. International efforts are being put into place. For example, International Compendium of Data Privacy Laws is a reference guides that outlines the case requirements in place when dealing with international data reach so that client and customers would know what steps to take to minimize their companies' exposure.[40]

6. Contribution to Literature

This paper has given an overview of big data technology and illustrated emerging trends in this area. It has been found and also cited in the academic circles that big data has a complex ecosystem of various applications for data extraction, storage analysis, visualization etc. We have reviewed various relevant literatures and brought to bear significant objectives that they address. We have also analytically investigated key issues in big data by following the research method as cited in Section 3. Since it is a very practical oriented field, we have specifically reviewed key reports and white papers of permanent research agencies while also exhaustively scanning information in public domain about big data technologies. Discussion of findings is presented in previous sections wherein we have highlighted key challenges and opportunities. Data quality and data piracy are evidently most critical bridge to cross. There is focus attention given to this area by enterprises and specific researches have also been conducted. There will be more frame work to build better data quality model that come out though directed research. Likewise, there is a wide scope of cross-functional research in data privacy that would have teams from legal, technology and management. This paper provided a framework and can form a conceptual basis to conduct further research and publish in aforesaid areas.

7. Managerial Implications

When enterprises work with large data sets, frequently, getting the data to the processors itself becomes a bottleneck. If we consider a quick back-of-the-envelope calculation, a typical disk data transfer rate is approximately 80-85 MB/second for a standard microprocessor. With this kind of rate, it would take almost 20 minutes to transfer 100 GB of data to the processor. Apparently, it is by far inordinately long time to transfer the data itself in a world where enterprises and consulting companies are struggling to work with cutting costs by optimizing the utilization of their employees' time and resources. In today's world, it is usual to get 100 GB of data and enterprises expect that such data is analyzed to the core so as to get the insights that can help them improve their business. Big Data technologies are indeed a big facilitator to help them this objective. An understanding of issues in this technology can help them leverage it to the best possible extent. We have reported the current state of practice and emerging issues in Big Data. The issues of data quality, storage and analytics challenges and data privacy are critical pain points for Big Data practitioners. However, it is imperative for practitioners to use the opportunity for better customer/user experience by harvesting Big Data. Usually, the Business Analytics practice in corporate world uses their expertise in econometrics and technology to respond to business issues faced by their clients or leaders. Understandably, they have to draw on draw on data from corporate sources (Point-of-Sales, market research, click-stream data, log files, accounting data etc.), which is sometimes complemented by information in public domain. The practitioners are looking towards improving the experience of doing analyses with Big Data. There are challenging issues of managing data quality and efficient analytics. For example, direct manipulating data is far easier than scripting. Analysts are accustomed to working with spreadsheet based tools that allows them to work with small data jobs. Gone are

those days now since data itself has grown so big that it cannot even be in stored and analyzed in spreadsheets. It is a call to action in that we highlight important challenges in Big Data. There is a great opportunity to make the analysis easier to do and faster.

8. Conclusion

We have attempted to draw out a trajectory that illustrates as to where we see technology going in the years to come. Big Data analytics space will see more of consolidation. As of now, space is crowded with many small players, who are supporting similar services and trying to increase market share. Big vendors like IBM, SAP, Microsoft, Oracle etc. will go in for more acquisitions thus allowing a place for Big Data service for their customers. Big Data analytics will be benefited more by increased convergence of In-memory processing, distributed architecture and parallel computing technologies. The applications will have a wide range of capabilities including dash boarding and reports that can provide a clear and easy to understand visual representation of quantitative data. They will also provide newer capabilities to process different data that may come from social media, mobile applications and sensors like Point-of-Sales devices. Applications will also move towards easier-to-use functionality. Traditionally, the use of such applications needed some programming knowledge but now they provide Graphical User Interface that have drag-and-drop facility to import data and create models. Analytics as a Service (AaaS) is also attracting more research interests and companies are putting the best of their IT budget by using AaaS which could be supported by internal private cloud or a public cloud. Intel, for instance, is providing implementation support to companies who are using AaaS as an option. As a cost effective option of running analytics, more and more companies will go this route. We foresee application developments such that data analysts will be able to leverage easy-to-use tools and provide insights and recommendations in shorter span of time. Determining the right mix would indeed depend on the IT expense appetite of company weighed against the risk mitigation and control. If the data is off-premise (on a public cloud), company has lesser control over it and thus higher the risk. However, the advantage would be that cost of storage, processing and analyses is lower if it is off-premise.

9. Future Research Directions

If we consider current, unbounded and high speed streaming data which is recorded real time from sensory devices that posits the highest opportunity to mine so as to provide most revealing insights about our world. To mine this data, a traditional approach to data analyses will fail since that does not assume an avalanche of streaming data. Research in this area would help in developing methods to build data analytics applications which provide predictive power to managerial decisions on a teal-time basis. Enterprises' ability to build Big Data capabilities is a function of its information governance processes. Increasingly, enterprises will spend more effort towards this direction. Most enterprises do not have a program to coordinate and architect the processes that make up information management strategy. For example deduplication efforts that occur before data enters Database. To bridge the gap, we foresee increasing attention to forming cross-functional teams. Such teams will comprise of people from functions like Finance, HR, operations etc. and would also have Big Data experts. We see this convergence increasingly and, in fact, such teams will form a task force to put precise and de-duplicated data through Big Data ecosystems. It will emphatically facilitate data management and consistency issues and that is vital to reap the benefits from Big Data. The combination of commodity hardware and software will allow more effective processing of Big Data. We expect that Big Data technology will bring information symmetry in enterprises. Information silos will be removed since fewer clusters will be created for more economical analyses. Evidently, such a system will allow executives to have more visibility into information and not be limited by department or functional silos. We also foresee opportunities for enterprises to have the opportunity to find insights for themselves or clients by utilizing advanced research, knowledge transfer and creating training program for talent and engage in more partnerships with open-source foundations.

References

1. IDC, *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things* (International Data Corporation, 2014).

2. Gartner Report, Survey Analysis: *Big Data Adoption in 2013 Shows Substance Behind the Hype* (2013).
3. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. Hung Byers, *Big Data: The Next Frontier for Innovation, Competition and Productivity* (Mckinsey Global Institute. May 2011).
4. IDC, *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things* (International Data Corporation, 2014).
5. H. M. M. Patching, L. M. Hudson, W. Cooke, A. J. Garcia, S. I. Hay, M. Roberts, C. L. Moyes, A supervised learning process to validate online disease reports for use in predictive models, *Big Data* **3**(4) (Jan 2016) 230–237.
6. R. Razavian et al., Population-level prediction of Type 2 diabetes from claims data and analysis of risk factors, *Big Data* **3**(4) (2016) 277–287.
7. Accenture, *Companies are Satisfied with Business Outcomes from Big Data and Recognize Big Data as Very Important to Their Digital Transformation* (Accenture White Paper, Sept. 2014).
8. H. Baldwin, *Big Data Hits the Dirt* (Forbes, Dec 2014).
9. M. L. Murphy and K. Tysiac, Data analytics helps auditors gain deep insight, *J. Accountancy* (2015).
10. Q. Yiming et al., The current status and challenges in computational analysis of genomic big data, *Big Data Research* **2**(1) (2015) 12–18.
11. T. Tullis and W. Albert, *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics* (Morgan Kaufmann, 2008).
12. M. Scheuer, H. Roitman and Y. Mass, *Konopnicki D, Extracting User Profiles from Large Scale Data, Proceedings of Massive Data Analytics on the Cloud* (ACM, 2010).
13. A. Mehmet, G. Buğra and F. Hakan, Aggregate profile clustering for telco analytics, *The Computer J.* (2015).
14. R. K. Lomotey and R. Deters, Terms analytics service for CouchDB: a document-based NoSQL, *International Journal of Big Data Intelligence* **2**(1) (2015) 23–36.
15. B. Lublinsky, K. T. Smith and A. Yakubovich, *Professional Hadoop Solutions* (John Wiley & Sons, Indianapolis, Ind., 2013).
16. K. Schmidt and C. Phillips, *Programming Elastic MapReduce Using AWS Services to Build an End-to-End Application* (O'Reilly Media, 2013).
17. S. Y. Liu et al., HYDRA: Large-scale social identity linkage via heterogeneous behavior modeling, In *41 ACM SIGMOD Int. Conf. Management of Data* (ACM SIGMOD, Snowbird, 2014).
18. B. Baesens and Hoboken, *Analytics in a Big Data World the Essential Guide to Data Science and its Applications* (John Wiley & Sons, 2014).
19. S. Y. Liu et al, How to conduct distributed incomplete pattern matching, *IEEE T Parall. Distr*. **1** (2013) 1.
20. Y. Nakamura et al., O'Brien, Quantum computers, *Nature* **464** (2014) 45–53.
21. H. Plattner and A. Zeier, *In-Memory Data Management Technology and Applications* (Springer, Berlin Heidelberg, 2012).
22. G. Vey, T. Krojzl and I. Krutov, *In-Memory Computing with SAP HANA on IBM eX5 Systems United States: IBM* (International Technical Support Organization, 2013).
23. A. Bucella, A. Cedchich and D. Domingo, Analyzing and improving data quality, *J. Comput. Sci. Tech.* **8**(2) (2008) 57–83.
24. L. Siyuan, L. Chen and N. Lionel, Anomaly detection from incomplete data, *ACM Transactions on Knowledge Discovery from Data* **9**(2) (2014).
25. X. Jin et al., Significance and challenges of big data research, *Big Data Research* **2**(2) (2015) 59–64.
26. V. Ganti and A. D. Sarma, *Data Cleaning: a Practical Perspective* (Morgan & Claypool, San Rafael, 2013).
27. DQS Group, Data quality services, https://msdn.microsoft.com/en-us/library/ff877925.aspx (accessed, Sept. 13, 2015).
28. J. Harper, *Big Data Governance: Big Data Quality* (Dataversity. April 2015).
29. Trillum Big Data Quality Services, Big data, http://www.trilliumsoftware.com/products/big-data/ (accessed, Sept. 2015).
30. WinPure, WinPure Clean and match, http://www.winpure.com/cleanmatch.html (accessed, Sept. 13, 2015).
31. R. K. Lomotey and R. Deters, Terms analytics service for CouchDB: a document-based NoSQL, *International Journal of Big Data Intelligence* **2**(1) (2015) 23–36.
32. Cloudera and Hadoop, Hadoop ecosystem, http://www.cloudera.com/content/cloudera/en/about/hadoop-and-big-data.html (accessed, Sept. 13, 2015).
33. A. D. Barrachina1 and A. O'Driscoll, A big data methodology for categorising technical support requests using Hadoop and Mahout, *J. Big Data* **1**(1) (2014).
34. J. R. Lourenço et al., Choosing the right NoSQL database for the job: a quality attribute evaluation, *J. Big Data* **2**(18) (2015).
35. R. Lerner, PostgreSQL, the NoSQL database, *Linux J.* (2015).
36. G. Chen, S. Wu and Y. Wang, The evolvement of big data systems: from the perspective of an information security application, *Big Data Research* **2**(2) (2015) 65–73.
37. C. N. Davis and D. Cuillier, *Transparency 2.0: Digital Data and Privacy in a Wired World* (Peter Lang, 2014).
38. J. Lane et al., *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (Cambridge: University Press, 2014).
39. H. Kenyon, Privacy by design: protect uses data from 'Get-Go', *Information Week* (2015).
40. B. Hostetler, *International Compendium of Data Privacy Laws* (2015).