Big Data Infrastructure, Data Visualisation and Challenges

Ramanathan Venkatraman
Institute of Systems Science
Singapore
rvenkat.iss@gmail.com

Sitalakshmi Venkatraman
Melbourne Polytechnic
Melbourne
sitavenkat@mp.edu.au

ABSTRACT
The importance of Big Data is being realised worldwide with the advancement of information technologies, leveraging the capabilities of virtualization and cloud computing. Big Data infrastructure and the use of its tools and applications will significantly transform the data centers of businesses in the next decade. Data analytics is evolving with the new real-time capability of Big Data solutions to provide business intelligence for timely and effective decision making. However, Big Data poses various challenges related to the infrastructure and resource constraints, and other issues including security and privacy. This paper takes an initial step in recognizing the value of creating Big Data infrastructure for delivering high performance and scalable business intelligence in an organization. It presents the state-of-the-art tools and technologies for Big Data infrastructure and NIST framework. The advantages of data visualisation are illustrated thorough industry case scenarios. The Big Data trends and challenges are also discussed. Overall, this paper contributes to providing valuable insights unto the Big Data journey of an organization to enable a scalable infrastructure for achieving mission critical decision-making through data visualisation.

CCS Concepts
• Information systems→Data management systems→Database management system engines

Keywords
Big Data; Big Data infrastructure; cloud; NoSQL database; Hadoop; data visualisation.

1. INTRODUCTION
The world with various advancements in technology is changing dramatically as businesses experience an unprecedented data explosion [1][2]. With the rise of Internet of Things (IoT), an increased enterprise aggregation of huge data and user statistics collected from diverse geographic locations, sensors and other sources of Big Data could be used effectively for real-time data analytics and visualisation [3]. From the increase in data growth to the way it is structured and used, businesses look towards leveraging Big Data to make accurate business decisions using data visualisation. Hence, Big Data has become a recent area of strategic investment for businesses by providing extremely powerful business intelligence when data is properly synthesized, analyzed and visualised. Though the increasing global Internet population to use new technologies for personal communication is driving the Big Data trend, there are limitations for businesses to make full utilization of the benefits of Big Data due to their resource constraints, cost and flexibility [4]. Big data presents a lot of challenges in terms of infrastructure and security. However, due to its capability of providing various business opportunities, industries do not consider all the challenges and risks before venturing into Big Data applications [5][6]. This forms the key motivation in this research paper to provide various insights in adopting the right Big Data infrastructure for business intelligence for organisations.

Big Data technology and services involve a variety of hardware/software resources, tools and techniques such as NoSQL databases, Hadoop or MapReduce file systems, virtualization, cloud platforms and related software ad well as analytics solutions [7][8]. These will impact an organisation's information technology (IT) related to server, storage, and networking infrastructure that are specifically designed to leverage and optimize the business services in different industry applications [9][10]. Since, the infrastructure forms the cornerstone of Big Data architecture for the successful use of data analytics in businesses, in this paper, we closely examine the infrastructural platforms, the analytical tools for data visualisation and the challenges such as security to better understand the Big Data landscape. The innovation of this research is that it is unique in exploring the big data infrastructures with the objective of achieving successful data visualization.

The rest of the paper is organized as follows. Section 2 provides the background information on the Big Data concepts, trends and challenges. In Section 3, we describe the prominent Big Data infrastructure technologies. Section 4 presents data visualisation with illustrations from industry case scenarios. Finally, Section 5 gives the conclusion and future work.

2. TRENDS AND CHALLENGES
Big data is a term that refers to gigantic datasets that are huge (volume), having more wide-ranging compound structure (variety) and constantly produced and dynamically changing (velocity) [11]. Occasionally, the data also perishes at the equivalent high speed as it is produced. Infrastructural technology is considered as the basis of the Big Data ecosystem for the storage, analytics and visualisation of data [12]. Big data has become a hot topic for businesses in this digital world as they face growing challenges that deal with large volumes of structured and unstructured data that are complex to process using traditional database and software methods. The growth of big data exceeds the capabilities of traditional IT infrastructures and represents a largely undeveloped area of computing and data management problems in different industry applications [9][10][13]. Big data has the potential to help companies grow, improve business operations, making faster, and more intelligent decisions. Hence, many Big Data technology

BDIOT 2019, August 22–24, 2019, Melbourne, VIC, Australia

DOI: https://doi.org/10.1145/3361758.3361768

providers offer resources in the cloud environment for organisations to reap the benefits of their business data.

Popular Big Data technology providers offer great processing potential and management of large quantities of data through platforms such as Apache Hadoop and Google MapReduce. The Big data technology is growing, and the vision in the use of big data is the way relevant, large and fast-growing data can be captured from any source and analysed to assist the organisation to gain useful insight towards helping their overall goal. Businesses look towards using big data to gain competitive advantages and to help them achieve their business goal such as increase revenues, improve customer satisfaction, and enhance their productivity [14]. As businesses continue to store large volumes of data, they look towards more sophisticated tools to mine and analyse data into meaningful way. Organisations are starting to realise that big data is more about business transformation and making the change to exploit data. Big data allows businesses to gain a deeper understanding of the dynamics of their business by analysing and visualising big data and integrating the results with traditional information so as to get a new perspective on their day-to-day operations. However, Big Data pose various challenges [15][16].

The first and foremost challenge is the lack of scalability due to poor infrastructure management, whether it is on-premise infrastructure or on the cloud [4]. Organisations do not want to maintain and pay for substantially more Big Data infrastructure when it is underutilized currently. However, if the infrastructure does not increase in size as the business data grows, they will not be able to gain any value from Big Data. Another issue is that many applications and data analytics software tools do not make us of optimal data transformation, efficient analysis and appropriate data visualisation [3][6]. While performing data transformation, if the quality of data is lost, whatever be the infrastructure used, it will not meet the organisation's Big Data needs. Above all these challenges, the security and privacy issue is the most compelling one since any data hosted by third party can raise questions about the security and privacy of an organisation's confidential information [17][18].

With new technology emerging for Big Data, organisations must be prepared to face challenges in supporting their dependence on Big Data due to the high costs involved, technological complexities, data availability, privacy and integrity concerns. Using Big Data infrastructure without understanding these issues may not necessarily be the right way for any organisation as Big Data forms an essential component of management decision making that requires new capabilities, as well as organisational and culture change [19]. The next section describes the industry standards and tools in Big Data infrastructure that can benefit organisations to create an implementation plan.

3. BIG DATA INFRASTRUCTURE

The first and foremost requirement of an orgnaisation before plunging into the Big Data landscape is to understand the infrastructural tools and technologies: what are they, how they operate and what is best used for. Some of the popular technologies for Big Data architecture are described below:

3.1 Hadoop

Hadoop is a readily available open source framework that uses a coste-effective programming model to allow distributed processing of big datasets by efficiently breaking it and distributing smaller parts for parallel or concurrent processing and analysing of them. Hadoop permits distributed parallel processing of gigantic amounts of data through economical, industry-standard servers that stores and processes the data. A HDFS storage layer is used and MapReduce component executes a variety of analytic functions to analyse thedata efficiently. Hadoop uses YARN for cluster management and scheduling applications for the user. In-depth analysis of data using machine learning algorithms could be done using SPARK on top of HDFS. However, with such opensource technologies there are additional risks of ongoing maintenance and support.

3.2 NoSQL

NoSQL stands for 'Not Only SQL' and refers to non-relational database technologies such as Cassandra, Neo4j, Redis, and MongoDB, which are also effective and economic choices for Big Data infrastructure. NoSQL databases are better tailored to handle dynamic and semi-structured data with low latency. While NoSQL is better suited for operational and analytical tasks to process selective criteria-based data in real-time, Hadoop is more employed for harnessing all data and in-depth analysis with high-throughput. Since both Hadoop and NoSQL have different advantages and purposes, both can be used simultaneously as in the case of HBase. However, security is one of the major concerns of NoSQL.

3.3 In-Memory Database (IMDB)

An IMDB is also known as a main memory database system (MMDB) that is popularly used in high-volume environments where response time is very critical. Since data resides in the memory of the system rather than in the disk storage, data access time and processing is very fast. Hence, IMDBs have become popular in recent years for handling High-Performance Computing (HPC) and Big Data applications. Related to this is also in-memory data grid (IMDG), which is the real-time analytics engine that produces real-time changes to data providing smart grid features. Using such technologies, In Big Data applications, huge quantities of data for processing can be stored in-memory, while the original and persistent data could be residing on an external disk.

3.4 Massively Parallel Processing (MPP)

MPP technology is a form of collective processing of massive amounts of data using several processors working on different parts of the same program. Each processor takes up different threads of the program to execute its own operating system and memory. A messaging interface in necessary to organize and manage the thread handling of the different processes involved in the MPP architecture. Many MPP technologies have partnerships with other major players among the Big Data technology providers. Hence, MPP technologies also have a crossover with other Big Data technologies.

3.5 Cloud Computing

Big players of Big Data infrastructure providers offer cloud computing that cover a range of products, technologies and services to various organisations in order to jump start with their Big Data ventures. All the resources and applications are hosted in cloud, and is considered to have minimal cost implications as organisations can pay based on the infrastructure, platform or software services used. Amazon, Microsoft, Oracle, IBM are some of the big players offering cost-effective Big Data architectures in the cloud. While cloud computing can deliver data insights seamlessly for organisations to benefit from, security and privacy issues of confidential and sensitive data are of great concern.

The rapid technological developments in Big Data could overwhelm traditional computing frameworks in businesses. Hence, the National Institute of Standards and Technology (NIST) has

provided a high-level conceptual framework as shown in Figure 1 [20]. The purpose of this framework is to serve as a reference model to facilitate understanding of the operational intricacies, design structures and requirements in Big Data. The advantage of the model is that it can be adopted by any organization as it is not tied to any specific vendor products, services, or reference implementation.

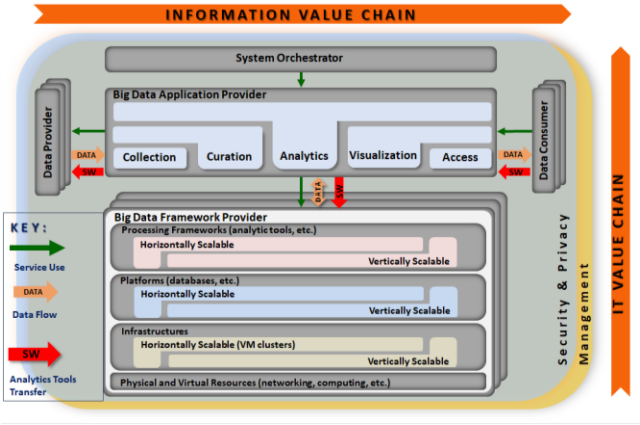

Figure 1. NIST Big Data Reference Architecture

Use Case Title		
Vertical (area)		
Author/Company/Email		
Actors/Stakeholders and their roles and responsibilities		
Goals		
Use Case Description		
Current Solutions	Compute(System)	
	Storage	
	Networking	
	Software	
Big Data Characteristics	Data Source (distributed/centralized)	
	Volume (size)	
	Velocity (e.g. real time)	
	Variety (multiple datasets, mashup)	
	Variability (rate of change)	
Big Data Science (collection, curation, analysis, action)	Veracity (Robustness Issues, semantics)	
	Visualization	
	Data Quality (syntax)	
	Data Types	
	Data Analytics	
Big Data Specific Challenges (Gaps)		
Big Data Specific Challenges in Mobility		
Security & Privacy Requirements		
Highlight issues for generalizing this use case (e.g. for ref. architecture)		
More Information (URLs)		
Note: <additional comments>		

Figure 2. Big Data requirement use case template (NIST).

The NIST framework is useful for organisations to plan for Big Data and the scalability of their resources as they would want to make the most of what they have and not to hastily commit significantly on investments of new technology that can lead to risks. They could also make use of the Big Data Use case template provided by NIST as shown in Figure 2 in order to understand their requirements.

4. DATA VISUALISATION

With the fast developments in Big Data technologies and application solutions, organisations are gaining meaningful data insights that can transform their businesses by utilising the large volumes of data for efficient decision-making and management. Organisations can use different analytical strategies such as predictive analytics to reveal patterns and provide decision making effectively. However, Big Data can add value only if the following key elements are planned well:

- data collection,
- data storage,
- data analysis, and
- data visualisation/output.

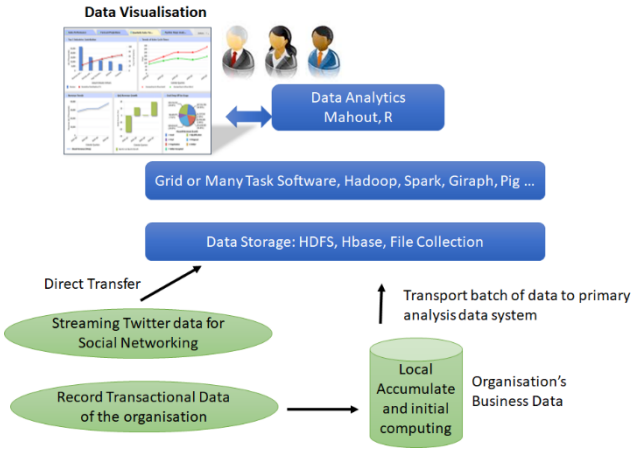

Figure 3. Typical Model for Big Data Visualisation.

Figure 3 provides a typical model for developing data visualisation with Big Data.

An organisation's data collection consists of each and every transactional record of business operation that would include history of sales, marketing promotions, emails, customer database, feedback, social media, and any data required for monitoring, and measuring in facilitating decision-making. Some data are already available in the local storage, some could be from the cloud and some would be accumulated or processed data. Data could be captured using Internet of Things (IoT) devices and sensors, customer apps, websites and social media profiles. However, collection of certain type and data formats require special data storage.

The data storage of an organization houses the data gathered from various sources. It includes traditional servers, data warehouses, data lakes, and other distributed/cloud-based storage systems. With cloud-based storage, organisations do not require physical systems on-site and it is flexible and cost-effective saving from maintenance and information security costs [21]. It is also considerably cheaper than investing in expensive dedicated systems and data warehouses. Organisations could choose in-house data storage for confidential data and cloud storage for data that requires low privacy in order to minimize the impact of possible security breaches and privacy risks in the cloud [22].

The Data analysis element consists of three main steps: 1. data preparation (identifying, cleaning and transforming the data into the required format for analysis); 2. analytical model development (employing the relevant model classified under predictive, descriptive and prescriptive analytical models) and 3. insights for

decision-making (based on varying paraments and different criteria for analysis from the chosen models). The output of the analysis must be visually comprehensible.

Data visualisation is the key to the success of Big Data. Unless the final output is in the form acceptable by people who need the data to be analysed, the whole Big Data venture is of no value. Huge reports or complicated graphics that seldom people understand will result in no meaningful decision-making or actions. There are various visualisation tools that include management dashboards and commercial data visualisation platforms that output attractive charts and graphs for clear and concise communication in order to gain data insights.

Figure 4 gives a simple data visualisation chart showing that the peak flu season in Australia did not occur until August from data collected for 6 years. However, the chart does show that there is a dramatic increase in number of patients affected by flu in 2017 as compared to previous years. Hence, hospitals across Australia can plan increasing staff in hospital workforce accordingly.

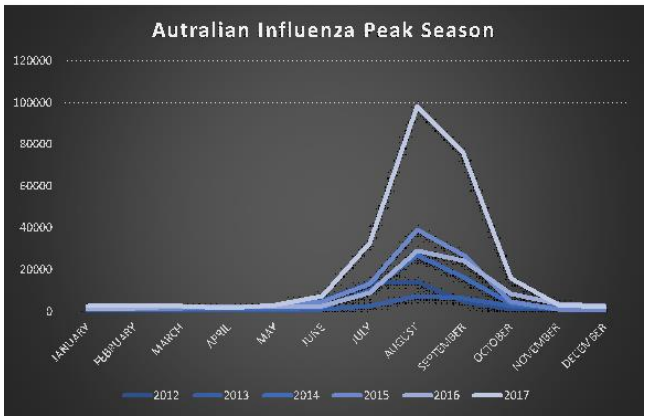

Figure 4. Data visualisation of time series data of flu patients.

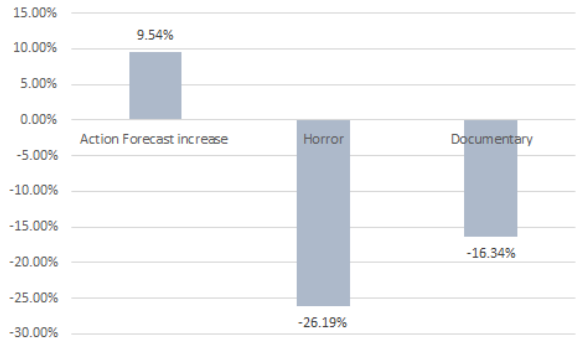

Figure 5. Data visualisation of movie genres prediction.

Figure 5 shows a visualisation of the prediction of movie genres for box office hit in a particular year using a forecasting model. It shows that while action movies have an increasing trend, horror movie genre is the worst, showing highest decreasing trend. Another example in Figure 6 provides rich data about influenza flu demographics in a country. Various sensitive information and decision parameters behind the data visualisation are used by the data analytics model to provide big data insights to aid in various drill-down analysis and decision-making. However, they are just a click away to anyone who has access to the visual tools and artefacts. Hence, even visual security is an important concern in Big Data.

Figure 6. Data insights - multiple views/decision parameters.

5. CONCLUSIONS AND FUTURE WORK

Big Data plays a critical role in the industry and continues to grow exponentially. The data-driven business revolution adds new levels of complexity for analysing data to match with the velocity at which data is generated from diverse sources. Big data technology developments have driven the transformation of organisations that look to leveraging Big Data for competitive advantage and to facilitate in achieving business goals. However, understanding Big Data technology, and modelling data visualisation with the data captured and analysed in a meaningful and intelligent way are important for the planning and management of Big Data in an organization. This paper provided guidelines for Big Data infrastructure using NIST framework and the importance of data visualisation for effective decision-making using illustrations from industry scenarios.

This paper has made a modest initial step to bring out the opportunities and challenges of Big Data. Future work would have a focus on the security and privacy concerns, in particular, with reference to the proliferation of IoT and blockchain technologies.

6. REFERENCES

[1] Frizzo-Barker J, Chow-White PA, Mozafari M, Ha D *An empirical study of the rise of big data in business scholarship.* International Journal of Information Management 36(3), (2016), 403–413.

[2] Chen M. et al., Big Data: A Survey, *Mobile Networks and Applications*, 19(2), (2014), 171-209.

[3] Gorodov E. Y. and Gubarev V. V. *Analytical review of data visualization methods in application to big data*. Journal of Electrical and Computer Engineering (4), (2013), 1-7

[4] Tian W. and Zhao Y., *Big data technologies and cloud computing*, Optimized Cloud Resource Management and Scheduling Theory and Practice, (2015), 17–49.

[5] McNeely CL, Hahm, J. *The big (data) bang: policy, prospects, and challenges*. Review of Policy Research 31(4), (2014), 304–310.

[6] Gandomi A, Haider M. *Beyond the hype: Big data concepts, methods, and analytics.* International Journal of Information Management 35(2), (2015),137–144.

[7] Xindong W., Xingquan Z., Gong-Qing W., Wei, D. *Data Mining with Big Data*, IEEE Transactions on Knowledge and data Engineering, 26(1), (2014), 97-107.

[8] Chang, V. A. *A model to compare cloud and non-cloud storage of Big Data*, Future Generation Computer Systems, 57, (2016), 56–76.

[9] Goli-Malekabadi, Z. Sargolzaei-Javan, M. Akbari, M. K. *An effective model for store and retrieve big health data in cloud computing*, Computer Methods and Programs in Biomedicine, 132, (2016), 75–82.

[10] Kumar, N. Vasilakos, A. V. and Rodrigues, J. J. *A multi-tenant cloud-based DC nano grid for self-sustained smart buildings in smart cities*, IEEE Communications Magazine, 55(3), (2017), 14–21.

[11] Laney, D. *3D Data Management: Controlling Data Volume Velocity and Variety*, Tech. rep. META Group, (2001).

[12] Gronwald, K.-D. *Big Data Analytics*, In: Integrated Business Information Systems A Holistic View of the Linked Business Process Chain ERP-SCM-CRM-BI-Big Data, (2017), 127-157.

[13] Huang T., Lan L., Fang X., An P., Min J., and Wang F., *Promises and challenges of big data computing in health sciences*, Big Data Research, 2(1), (2015), 2–11.

[14] Kshetri N. *The emerging role of Big Data in key development issues*: Opportunities, challenges, and concerns. Big Data & Society 1(2), (2014), 1-20.

[15] Jing, P. *A new model of data protection on cloud storage*, Journal of Networks, 9(3), (2014), 666–671.

[16] Nelson B, Olovsson T *Security and privacy for big data: A systematic literature review*. In: Big Data (Big Data), 2016 IEEE International Conference on, IEEE, (2016), 3693–3702

[17] Li-chuan M., Qing-qi P., Hao L., Hong-ning L.. *Survey of Security Issues in Big Data*, Radio Communications Technology, 41(1), (2015), 1-7.

[18] Deng-Guo F., Min Z., Hao L. *Big Data Security and Privacy Protection*, Chinese Journal of Computers, 37(1), (2014), 246-258.

[19] Jina, X.. Waha B., Chenga X., and Wanga Y., *Significance and challenges of big data research*, Big Data Research, 2, (2015), 59–64.

[20] NIST, Big Data Interoperability Framework: Volume 6, Reference Architecture, NIST, USA (2018).

[21] Subashini S. and Kavitha V., *A survey on security issues in service delivery models of cloud computing*, Journal of Network and Computer Applications, 34(1), (2011), 1–11.

[22] Cheng H., Wang W., and Rong C., *Privacy protection beyond encryption for cloud big data*, in Proceedings of the 2nd International Conference on Information Technology and Electronic Commerce, (ICITEC ’14), (2014), 188–191, IEEE, Dalian, China.