Indonesian Journal of Electrical Engineering and Computer Science
Vol. 1, No. 1, January 2016, pp. 191 ~ 200
DOI: 10.11591/ijeecs.v1.i1.pp191-200

Big Data Platforms and Techniques

Salisu Musa Borodo*[1], Siti Mariyam Shamsuddin[2], Shafaatunnur Hasan[3]
[1,2,3]Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Malaysia
[2,3]UTM Big Data Centre, Ibnu Sina Institute for Scientific and Industrial Research Universiti Teknologi Malaysia, 81310 Johor Malaysia
*Corresponding author, email: salisuborodo@gmail.com, mariyam@utm.my, shafaatunnur@gmail.com

Abstract

Data is growing at unprecedented rateand has led to huge volume generated; the data sources include mobile, internet and sensors.This voluminous data is generated and updated at high velocity by batch and streaming platforms. This data is also varied along structured and unstructured types. This volume, velocity and variety of data led to the term big data. Big data has been premised to contain untapped knowledge, its exploration and exploitation is termed big data analytics. This literature reviewed platforms such as batch processing, real time processing and interactive analytics used in big data environments. Techniques used for big data are machine learning, Data Mining, Neural Network and Deep Learning. Some big data architectures are offered from Microsoft, IBM and National Institute of Standards and Technology. Big data potentials can transform economies and reducerunning cost of institutions. Big data has challenges such as storage, computation, security and privacy.

Keywords*: Big data, big data architecture, big data techniques, big data platform*

1. Introduction

The concept big data came into being through explosion of data from the Internet, cloud, data center, mobile, Internet of things, sensors and domains that possess and process huge datasets. Volume, velocity and variety are the main features of big data (1). These features make traditional computing models ineffective. A premise of tremendous value in the huge datasets is the motive for big data exploration and exploitation. Big data has been identified with potential to revolutionize many aspects of life (2); applications of big data in some domains have practically changed their practices (3). Data has penetrated each industry and all business functions; it is now considered a major factor in production (4).

Big data would inspire a lot of innovative models, companies, products and services. This is due to the strategic insight it behold for the IT industry and businesses. The IT industry would create new products and target untapped markets. Businesses would optimize existing businesses and have new business models. A lot of academic research is being conducted in the field of big data; ranging from applications, tools, techniques and architecture. The research is interdisciplinary and generally called data science. In view of the divergent interest in big data from distinct domains, a clear and innate appreciation of its definition, advancement, constituent technologies and challenges becomes paramount (5). In this regard, this paper would give a literature review on some aspectsof big data; there are definitions, features, opportunities, challenges, platforms, techniques and architectures. The review is unique from existing ones due to its wide coverage of big data architecture offerings from companies and the unified reference architecture. These architectures are the most important blueprint to help navigate the big data terrain. The rest of this paper is outlined as follows; section two discuss the big data context; section three examines big data platforms while section four analyze big data techniques; section fiveelucidate big data architectures. Section six is theconclusion of the literature review.

2. The Big Data Context

This section tries to put big data in its proper context. Numerous definitions of big data as provided by three perspectives are provided. Five features that distinguish big data from traditional computing are also explained. These are volume, velocity, variety, veracity and value.

Received August 21, 2015; Revised November 24, 2015; Accepted Decmmber 17, 2015

The opportunities brought by big data such as transformation of economies and cost saving for government institutions are provided equally. Challenges confronting big data such as storage, privacy, security and computation are covered as well.

2.1. Big Data Definition

There are a lot of definitions of big data due to its wide usage in numerous fields. There are the product oriented, process oriented, cognition oriented and social movement perspective (6). The product oriented perspective looks at features of the data; especially its volume, velocity and variety. Big data imply contemporary technologies and architectures that have been designed to efficiently deduce benefit from huge and variety of datasets (7). Another definition states big data involves huge volume, heterogeneous, localized control and finds intricate and dynamic correlations between data (8). A third definition states as a result of exponential increase in global data, big data signifies huge datasets. Relative to conventional data, it includes huge unstructured data that need real time analysis. It also comes with new opportunities for dissecting value and deeper understanding (9). The second perspective of big data leverages the processes involved in its operation.

The process oriented perspective looks at processing activities that are involved or essential in dealing with big data. These processes are searching, aggregating, storage and analysis of data. These processes in the context of big data are novel in terms of architecture, tools and techniques. This perspective to big data further highlights the unique technological infrastructure; most importantly tools and programming methodologies required in creating big data. A definition developed at University of California Berkeley resonates within the process oriented perspective. It states big data is when the traditional computing technology cannot meet timely, cost effective and quality answers that are demanded by data driven questions requested by users (10). The next perspective to big data definition involves the cognition dimension.

The third school of thought is the cognition oriented perspective. Big data is defined in relation to the limited ability of the human mind to understand, hence a need for technological infrastructure, numerous techniques from diverse disciplines and visualization techniques in conveying meaning of data.

2.2. Big Data Features

Volume, variety and velocity are the main features that distinguish big data. Additional features also exist such as veracity and value. These five V's make big data outstanding from traditional computing data. These features reflect huge volume, high velocity, high variety, low veracity and high value(11).

The huge volume of data being generated is astonishing due to the ease at which data is created, shared and stored.Some of these huge data generation schemes are the LSS Telescope which generates 150 terabytes weekly (12). Simulations in data-intensive science require and generate huge data (13). E-Commerce allows ten times more data about a transaction to be collected (1). Internet companies have lot of customer data. Facebook generates 3 billion contents daily; Walmart generates 267 million transactions from its 6000 global outlets (2). An urban Smart project in China generated 200 PB of data. There would be 209 billion RFID tags collecting data by 2021 (9). Data volume would continue to rise in the years ahead. The second distinguished feature of big data is its velocity.

Big data velocity involves the timing required to collect, store, preprocess, process and present data in a reasonable amount of time. The velocity of the data leads to different kinds of processing; such as batch, neartime, real time and stream. Batch processing systems have low time requirement and are the most deployed big data analytics. Applications such as stock prices and weather forecasting have strict time requirements; hence they need real time or stream processing (14). Social network applications also need real time processing for its publishers, stakeholders and observers (15). Due to the much demand for real time big data processing; solution providers have been providing stream and complex event processing systems. Due to the increase in the rate of unstructured data, the variety aspect of big data is also paramount.

The variety of big data sources is another strategic feature. Conventional databases deal with structured data, where by the schema is predefined to reflect the data model; this makes processing, indexing and querying easy due to the standardized nature of the data. The

recent big data types are unstructured and semi structured. These data types lack data schema entirely or the schemais not enforced strictly. Examples of varied data sources are social media logs, email and internet logs. Not all data sources have acceptable quality criteria in them; this makes the veracity feature of big data important.

Veracity involves the reliability of the data; this is challenging due to the numerous data sourcesof big data. This affects the quality and predictability of the data due to inconsistency, incompleteness, ambiguities, latency and model approximations. Criteria would need to be in place to gauge the quality of the data, when the acceptability criteria are met; the data isthen used for preprocessing and analytics. The last outstanding feature of big data is value.

The value in big data comes from use cases in companies, governments and research. Companies such as Facebook and Amazon have taken the big data leap. An in-depth study by Mckingsey (4) showed big data can boost productivity, increase competitive advantage and increase economic surplus of consumers. When big data is deployed creatively and effectively, the study project a value of $300 billion for USA healthcare, 60% increase in profit for USA retails and savings of $149 billion for governments in Europe. Big data would herald new innovation, competition, productivity, growth and value capture; while the demand side is from consumers, companies and different economic sectors (4). Epidemic outbreak forecast and discounted ticket sales are other big data values targeted by Google (16) and Microsoft (9) respectively.

2.3. Big Data Opportunities and Challenges

Big Data opportunities have been well articulated in the McKinsey study (4). There are increased productivity and cost savings for numerous organizations. Innovative business models such as the booming sharing economy and on-demand services would rely on big data. New companies would keep on springing up and provide cutting edge services for consumers. Big data has also been premised to refine the scientific method (13). Big data has also opened the door for data scientists; these are skilled people that effectively and creatively deduce big data value for their organizations and clients.There are however a number of challenges to be surmounted to harness big data value.

Some big data challenges are privacy, security, storage and processing. Privacy is the most challenging aspect of big data; this is due to the age old sensitive nature of peoples' data. When publicly available data of an individual is subjected to inference techniques, insightful informationof an individual can be deduced. Legal, technical and policy strategies can be put in place to safeguard the privacy of individuals' data. Laws could guide data collection, processing, usage and transfer to third party. Technical mechanisms include encrypting data at rest and on transmission. Organization policy should ensure workers clearly know their responsibilities, limitations and applicable sanctions (17) (18) (19). Security is challenging to big data due to its pervasive nature. Secure computation need to be incorporated for distributed programming frameworks such as Map Reduce (19). Storage and processing areother challenges of big data; the amount of effort required on networks, servers to store and process data is enormous.The processing of datawould warranteffort to distribute the stored data for processing. Another processing challenge is the size of data for machine learning, this requires several scan of the data which is costly because learning objectives and assessment measures are non-linear, non-smooth, non-convex and non-decomposable over samples (18). The viable solution is to bring processing to the storage point or taking only the relevant data for processing through reduction techniques. Cloud computing is anoverall viable option to the storage and processing dilemma, this is due to its prospect of scalable storage and processing capacity (17).

3. Big Data Platforms

Big data consist of three types of platforms. There are batch processing, real time processing and interactive analytics. Batch processing platforms take time to process data, there are the most deployed big data platforms, they carry out extensive computations; real time processing require faster processing of data, streaming is needed in applications that require minimal latency; interactive analytics allow users to access dataset remotely and carry out numerous operations as needed. The batch processing platform Apache Hadoop is the most deployed.

Apache Hadoop is an open source implementation of MapReduce. MapReduce was designed at Google and used on applications likelarge scale machine learning and clustering problems; it was inspired by the map and reduce primitives of the functional language Lisp (20). A problem is broken down by Hadoop recursively into the smallest unit to be solved. The small chunks are then distributed on system nodes for execution through the coordination of a master node and working nodes (15). Hadoop is a software library with modules such as MapReduce, HDFS and YARN. Hadoop also has abstractions that operate in its ecosystem; they are Hive, Hbase, Flume, Sqoop, Chukwa and Zookeeper. These Hadoop modules and abstractions operate across the big data value chain; from acquisition, storage, computation, analysis and management.Hadoop is deployed on a huge scale and scope in companies due to scalability, cost effectiveness, flexibility and fault tolerance.

The second type of big data platforms are stream processing systems, there are also called real time or near real time systems. These systems are needed in environments that have continuous data streams and require immediate processing of data. Real time systems require fast query and analysis of data in warehouses for users. Stream systems require processing of continuous data without necessarily need for its storage. Such applications are whether and transport systems. The number of networked objects with sensors is projected to reach 25 to 50 billion by 2020, majority of these objects would be smart devices, smartphones and be interconnected (21). The devices data would have known and novel uses to optimize wellbeing of people, processes, environment, systems and organizations. Stream processing would be required for these devices. Examples of stream processing platforms are Storm and SAP HANA. The last big data platform available is interactive analytics.

Interactive analytics allow users to issue direct queries to data, data scientists can formulate, assess hypothesis and see instant results. Users get connected to a system directly and interact with data, users with minimal skills become more productive because the tools are very easy. The data is explored through user defined, default filtering, query and numerous transformations (22). Interactive analytics hasten scientific work validation and improvement. Some interactive analytics platforms are Network Repository (22) and Apache Drill (23).

4. Big Data Techniques

There are many techniques for analyzing big data, the techniques cut across disciplines and overlap in some cases. The disciplines range from data mining, machine learning, neural networks and pattern recognition (2). A technique from a discipline or combination of them is used to gain valuable insight from big data. These techniques in the context of big data are called analytics. Machine learning techniques are known for their predictive capability, given a previous instance of data, the technique is able to infer appropriate response to subsequent occurrence. With this ability at hand, machine learning techniques have solved traditional computing tasks. There are numerous learning techniques for diverse problem types. There is no “one size fits all” technique; this is due to unique ability, limitations of each technique and nature of problems (24). Machine learning techniques are primarily hindered by scale of data and computation in big data. We would look at some improved learning techniques used in big data context. A MapReduce framework for real time traffic processingwas designed in a study (25); the challenges were distributed processing, distributed local learning and model fusion. The framework used Expectation Maximization algorithm for local learning process, the global model is compared to the distributed models in order to merge into a final model for traffic forecasting (25). Support Vector Machine was also improved with a probabilistic distribution for a large scale data application; the problem required an active learning mechanism due to sample complexity (26). Cross domain learning incorporating Support Vector Machine was used to develop a new and efficient method for learning classification of images across domains, the method works despite a small training data and different distributions. The method was evaluated against other cross domain ones on large video datasets (27). Other scaled machine learning algorithms are fast learning for ranking (28) and sparse approximations through boosting (29). Machine learning are used by another set of techniques called Data Mining in big data problems. Data mining techniques are used for deducing information from data. There are used for clustering, regression and association rule mapping. Data mining techniques equally face computational and data complexity inherent in big data. Sampling techniques are used to

augment their weaknesses to meet big data challenges (2). Neural networks are one of the most used machine learning techniques.

Neural Network (NN) consists of mature techniques that mimic human brain function. The techniques have successful applications used for clustering and classification. NN consists of nodes and layers for its operation, with more layers and nodes, better accuracies are obtained but at a higher computational cost. Big data with its complexity causes high computational cost for neural networks (30). The second challenge faced by NN in big data is the algorithms perform poorly. Two approaches are used to overcome these challenges. First is to reduce data size by sampling techniques; the second is to scale Neural Network for distributed computing. A research trained neural networks on large scale data, incorporating a hash based implementation of maximum entropy model. The method achieved lower computational cost and 10% reduction in word error rate (31). Another research implemented a parallel and pipelined neural network on FPGA hardware; the resulting prototype had an advantage of no limitation to the network size and topology of the network (32). Deep learning is a new technique that leverage neural networks and are able to produce compelling results based on multiple layer feature extraction. Deep convolution neural networks were used to classify images of MNIST dataset that yielded near human recognition by reducing error rate up to 30% (33). Neural deep learning was able to achieve 15.8% accuracy on a 10 million image recognition problem (34).

5. Big Data Architecture

In view of the enormous potential and challenges of big data; there is a need to decompose big data into an understandable architecture that conveys key components inherent in it. These stem from the following; a lack of consensus on what attributes constitute a big data solution; how does big data differ from traditional computational environments; what are the essential characteristics of big data environment; how does big data integrate with existing architectures; what scientific, technological and standardization challenges need to be addressed to hasten deployment of big data solutions (35). There is also the need to better understand data life cycle for users, industry and policy makers; another need for big data architecture is to identify important components and functions so as to outline relevant boundaries, interactivity of components and security concerns (36). Based on these premise, an initiative was put in place to come up with a reference architecture of big data (35).This section looks at specific big data architecture offerings andthe vendor neutral big data reference architecture.

5.1. Booz Allen Hamilton Cloud Analytics Reference Architecture

A reference architecture offering is from Booz Allen (37). The architecture is a four layered horizontal architecture. The top layer is the human insights and actions layer, this layer consist of custom interfaces and visualization of data. The second layer of the Booz Allen offering is the Analytics and Services layer, this consist of tools for analysis, modeling, testing and simulations of data. The third layer is the data management; this is the secure data repository that handles all data sources and metadata. The bottom layer is the infrastructure layer; this involves the technology platform that stores and manages the data.

5.2. IBM Big Data & Analytics Reference Architecture

The IBM Big data reference architecture shown in Figure 1 is a vertically layered architecture (38). The left most layer is the data sources; this layer accepts all structured and unstructured data types. The second left most layer is called big data platform capabilities; this is where all data integration and real time analytic capabilities take place. Information governance, security and business continuity also operate at this layer. The third layer is the advanced analytics and insights layer; this is where cognitive learning, prescriptive analytics, descriptive analytics and predictive analytics take place. The last layer is called New and enhanced applications. This layer hosts all types of applications that make use of analytics result for their operations.

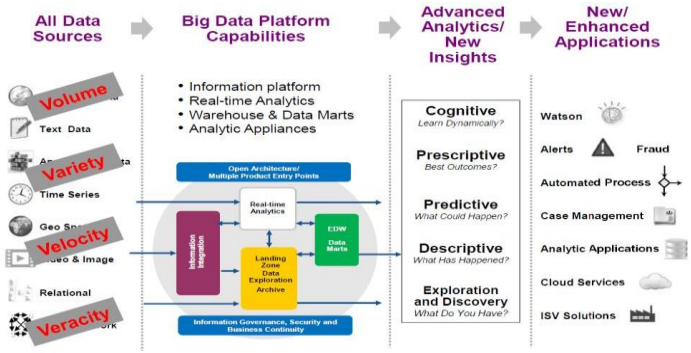

Figure 1. IBM Big Data & Analytics Reference Architecture – source: (38)

5.3. Big Data Ecosystem Reference Architecture

The next big data reference architecture is from Microsoft (36); the architecture consist of horizontal and vertical layers. The top horizontal layer is the data sources, this consist of all structured and unstructured data. The next layer is a vertical component called data transformation, this consist of techniques that collect, aggregate, match and mine the data. The other vertical component is the data infrastructure; this consists of database software, networking, servers and storage. The bottom layer is the data usage layer; this layer provides the analyzed data in various formats to diverse users.

5.4. SAP Big Data Reference Architecture

The software company SAP also has a Big Data Reference Architecture (39). The architecture shown in Figure 2 consists of vertical and horizontal layers.The first vertical layer consists of data sources, the second vertical layer is the data ingestion and provisioning; this ensures diverse data sources are aggregated. The top horizontal layer consists of various applications such as analytic, predictive, web and mobile. The second horizontal layer is SAP HANA; this involves stream processing and accelerated analytics. The bottom horizontal layer is the Hadoop Data Lake. It involves Storage, data processing and Deep analytics.

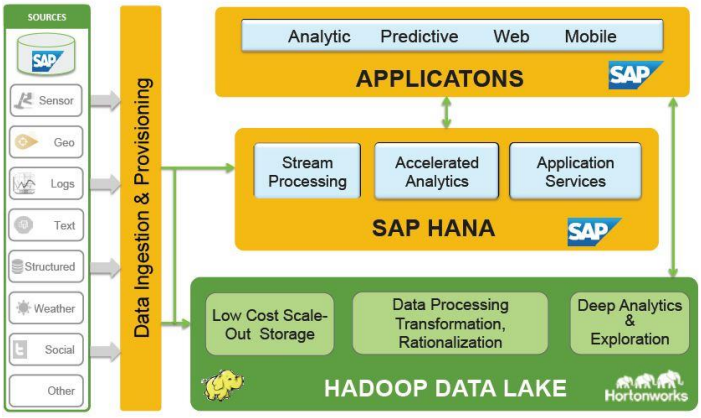

Figure 2. SAP Big Data Reference Architecture – source: (39)

5.5. Oracle Architecture Framework for Big Data

Oracle also has a reference architecture offering. The architecture consists of two bottom horizontal layers and six vertical layers. The first vertical layer from left to right is data, this consist of structured, semi structured and unstructured data. The second vertical layer is the acquire layer, data is acquired here in different formats such asfiles, HDFS, NoSQL and

Streaming Events. The third vertical layeris called Organize; this layer prepares data (MapReduce) for the analytical operations and ensures data quality assurance. The fourth vertical layer is the Analyze; this layer carries out analytical operations such as In-place analytics, Faceted Analytics and SQL Analytics. The fifth layer is the decide layer, this involves the processed data presented in formats such as visualization, recommendations, alerts and dashboards. The last vertical layer is the Management, Security and Governance. Next are the horizontal layers; the top layer here is the specialized hardware which involves technology platforms. The bottom horizontal layer provides an Industry Integration layers that integrates different technology offerings, this makes the operation of diverse technologies seamless(40).

5.6. GPUMLIB Big Data Architecture Framework

GPUMLIB is an acronym for Graphic Processing Unit Machine Learning Library (41). GPUMLIB providesan architecture framework to handle big data problems using machine learning algorithms. GPUMLIB consist of four vertical layers. The first layer is the multi criteria data sources; covering both structured and unstructured data. The second layer provides a preprocessing pipeline for data. This is where cleaning of data takes place prior to analytical operations. The third layer provides predictive knowledge discovery engine that host numerous machine learning algorithms. This layer runs on Graphic Processing Units, CUDA architecture and C++programs to execute theclassifiers. The last layer provides post predictive knowledge discovery multi criteria decision output. The GPUMLIB architecture is shown in Figure 3

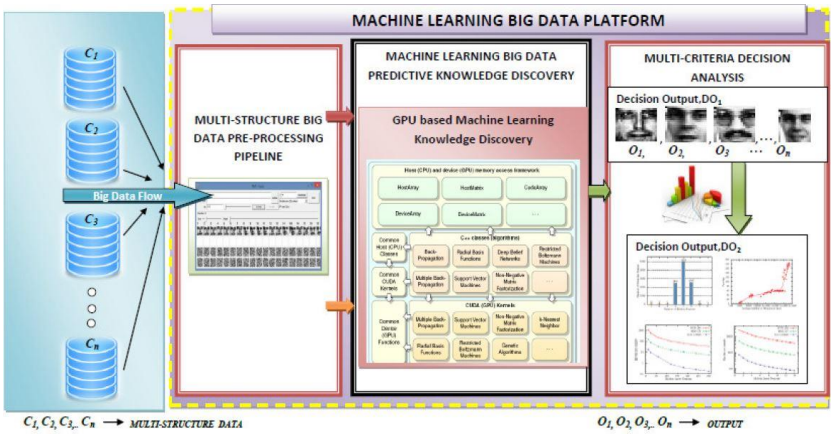

Figure 3. GPUMLIB Big Data Architecture Framework – source: (41)

5.7. National Big Data Reference Architecture

The National Institute of Standard and Technology wanted avendor neutral, infrastructure and technology independent framework; which optimizes the identification and use of best analytics tool by big data stakeholders on a platform of their choice. This necessity prompted the setup of the Big Data Public Working Group to achieve the stated goals.This group has sub groups working under it to achieve the stated big data goals. The goals were broken intodefinitions, taxonomies, use cases and general requirements, security and privacy requirements, architectures white paper survey, reference architecture and technology roadmap. Work on each sub task resulted into a volume of work. The published Volume on Reference Architecture (35) is a working document as it evolves based on new developments in the big data ecosystem. It is called the National Big Data Reference Architecture (NBDRA) and shown in Figure 4. The reference architecture got inputs from big data use case analysis (42), survey of existing reference architectures (43) and thebig data taxonomy (44). The architecture consists of five logical components connected by interoperable services and two overall management fabrics. These components reflect technical roles in any big data ecosystem.

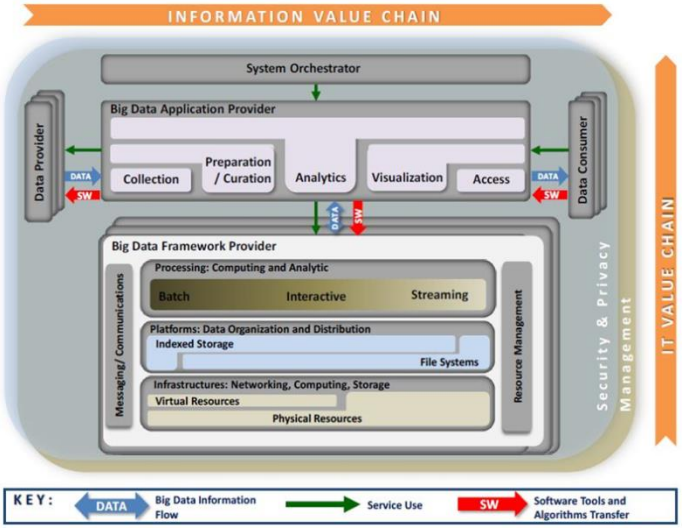

Figure 4. National Big Data Reference Architecture – source: (35)

The first component is data provider; this is an entity or organization that provides information for collection and processing by the big data ecosystem. The second component is the big data application provider, this is where the core functionality of big data system is implemented, and this involves data collection, multiple transformations and diverse usage of data. The specific functions are collection, preparation, analytics, visualization and access. The application provider ensures compliance with all management, security and privacy requirements. The big data framework provider is the third component; this layer provides the IT value chain components that include diverse technologies in a certain hierarchy. These components are used by the application provider to achieve the big data goal. The components in this layer are infrastructure, platforms and processing frameworks.The fourth component is the data consumers. Data consumer can be a user or another system. Data consumer accesses services provided bythe application provider component through interfaces. The services involved are initiation, data transfer and termination. The last functional component is system orchestrator. This is the role of defining and administering the data application activities into an operational system. The system orchestrator is either a human or a software component; it can also involve a combination of both. The system orchestratorin an enterprise system can be mapped to the role of a system governor; this role stipulates requirements and constraints within which the system must operate. The last two components are management layers. There are system management and data life cycle management. System management involves the processes of setting up, oversight and upkeep of big data infrastructure. This management component is challenging due to the inherent complexity of the big data ecosystem. The management of numerous processing frameworks is the most challenging because they are expected to scale, be safe and resilient in their operations. The last component of the NBRDA is Data Life Cycle Management. Due to the diverse data requirements of big data systems, there is a need for data life cycle management. This component is similar to the data provider in functionality; but has a broader scope across all the other four architecture components.

6. Conclusion

Big data platforms and tools are the cornerstone of the data revolution taking place. Volume, velocity and variety are the main features thatmake big data outstanding from traditional computing. Big data has been premised to make profound insights in all domains. Innovative models, products, services and huge cost savings are some of the opportunities big data present. Big data due to its huge potential and user base has multiple definitions reflecting wide perspective of its stakeholders. Platforms for big data have diverse purposes with regard to speed at which data needs to be produced and processed. The platforms are batch processing, stream processing and interactive analytics. The techniques involvedin big data are many and

from different domains; but eachtechnique exhibit characteristics that can be mapped to a problem pattern. The multidisciplinary nature of big data warrants collaboration of diverse fields; this can lead tohybrid techniques for solving a particular problem.

The need for a standardized reference architecture due to diverse offerings from numerous Industry players led to the National Big Data Reference Architecture. The reference architecture is a working document as it continues to evolve; reflecting the dynamic nature of the big data ecosystem. Big data has many challenges confronting it at this infancy stage of its growth. Some of them are privacy, security, storage and processing.

Privacy gaps can be resolved by extending the capacity of encryption schemes. Novel encryption techniques that are industry ready can equally help, incorporating and operationalizing institutional governance can help as well. Security can be improved by improving current security architectures reliability, this would be an on going effort due to the dynamic nature of big data ecosystem. Storage of data is a challenge due to the huge volume generated. Hadoop can equally fit the storage gap due to its scalability. Cloud computing can provide a limitless storage capacity for data as well. Processing involves the techniques used for analyzing data, challenges faced by machine learning techniques are high sample dimension, huge data size and problem complexity. Possible solution strategies are dimensionality reduction, instance selection, incremental learning and adapting techniques for distributed and parallel computing.

Acknowledgement

This work was partially supported by The Ministry of Higher Education (MOHE), Universiti Teknologi Malaysia (UTM) and UTM Big Data Centre (UTMBDC) with grant numbers FRGS 4F802, FRGS 4F786, FLAGSHIP 02G38, FLAGSHIP 02G50 and FLAGSHIP 03G17.

References

[1] Laney D. 3D Data Management: Controlling Data Volume, Velocity and Variety. *Application Delivery Strategies*. 2001.
[2] Philip Chen CL, Zhang CY. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Inf Sci*. 2014; 275: 314–347.
[3] Huang T, Lan L, Fang X, An P, Min J, Wang F. Promises and Challenges of Big Data Computing in Health Sciences. *Big Data Res*. 2015; 2(1): 2–11.
[4] James M, Michael C, Brad B, Jacques B, Richard D, Charles R. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Glob Inst. 2011.
[5] Hu H, Wen Y, Chua T-S, Li X. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. *IEEE Access*. 2014; 2: 652–687.
[6] McKerlich R, Ives C, McGreal R. Measuring use and creation of open educational resources in higher education. *Int Rev Res Open Distance Learn*. 2013; 14(4): 90–103.
[7] Gantz BJ, Reinsel D. Extracting Value from Chaos State of the Universe : An Executive Summary. *IDC iView*. 2011: 1–12.
[8] Wu X, Zhu X, Wu G-Q, Ding W. Data mining with big data. *IEEE Trans Knowl Data Eng*. 2014; 26(1): 97–107.
[9] Chen M, Mao S, Liu Y. Big data: A survey. *Mob Networks Appl*. 2014; 19(2): 171–209.
[10] Kraska T. Finding the needle in the big data systems haystack. *IEEE Internet Comput.* 2013; 17(1): 84–6.
[11] Jin X, Wah BW, Cheng X, Wang Y. Significance and Challenges of Big Data Research. *Big Data Res*. 2015; 1: 3–8.
[12] Steering the future of computing. *Nature*. 2006; 440(7083): 383.
[13] Collins J. The fourth paradigm. *The fourth paradigm*. XVIII ISA World Congress of Sociology. Yokohama. 2014.
[14] Assunção MD, Calheiros RN, Bianchi S, Netto M a. S, Buyya R. Big Data computing and clouds: Trends and future directions. *J Parallel Distrib Comput*. 2014; 3–15.
[15] Chardonnens T, Cudre-Mauroux P, Grund M, Perroud B. *Big data analytics on high Velocity streams: A case study*. 2013 IEEE Int Conf Big Data. 2013: 784–787.
[16] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009; 457(7232): 1012–1024.
[17] Katal A, Wazid M, Goudar RH. *Big data: Issues, challenges, tools and Good practices*. 6th Int Conf Contemp Comput IC3 2013. Noida. 2013; 404–409.
[18] Michael K, Miller K. *Big data: New opportunities and new challenges*. IEEE Computer Society Long

Beach. 2013; XX(X): 1–20.
[19] Carter P. *Top Ten Big Data Security and Privacy Challenges*. Int Data Corp. 2011.
[20] Dean J, Ghemawat S. MapReduce. *Commun ACM*. 2008; 51(1): 107–113.
[21] Kamburugamuve S, Fox G, Leake D, Qiu J. Survey of Distributed Stream Processing for Large Stream Sources. *Grids Ucs Indiana Edu*. 2013.
[22] Rossi RA, Ahmed NK. Interactive Data Repositories : From Data Sharing to Interactive Data Exploration & Visualization. *Purdue University*. 2015; 78–82.
[23] Hausenblas M, Nadeau J. Apache Drill: Interactive Ad-Hoc Analysis at Scale. *Big Data*. 2013; 1(2): 100–104.
[24] Steve O. *Machine Learning, Cognition, and Big Data*. CA Technologies. 2012
[25] Chen C, Liu Z, Lin WH, Li S, Wang K. Distributed modeling in a mapreduce framework for data-driven traffic flow forecasting. *IEEE Trans Intell Transp Syst*. 2013; 14(1): 22–33.
[26] Mitra P, Murthy CA, Pal SK. A Probabilistic Active Support Vector Learning Algorithm. *IEEE Trans Pattern Anal Mach Intell*. 2004; 26(3): 413–418.
[27] Jiang W, Zavesky E, Chang SF, Loui A. *Cross-domain learning methods for high-level visual concept classification*. Proc - Int Conf Image Process ICIP. San Diego. 2008; 161–164.
[28] Raykar VC, Duraiswami R, Krishnapuram B. A fast algorithm for learning a ranking function from large-scale data sets. *IEEE Trans Pattern Anal Mach Intell*. 2008; 30(7): 1158–1170.
[29] Sun P, Yao X. Sparse approximation through boosting for learning large scale kernel machines. *IEEE Trans Neural Netw*. 2010; 21(6): 883–894.
[30] Ranger C, Raghuraman R, Penmetsa A, Bradski G, Kozyrakis C. *Evaluating MapReduce for multi-core and multiprocessor systems*. Proc - Int Symp High-Performance Comput Archit. 2007; 13–24.
[31] Deoras A, Povey D. *Strategies for Training Large Scale Neural Network Language Models*. Asru 2011.
[32] Ahn JB. Neuron Machine : Parallel and Pipelined Digital Neurocomputing Architecture. *Cybernetics Com.* Bali. 2012; 143–147.
[33] Ciresan D, Meier U, Schmidhuber J. Multi-column Deep Neural Networks for Image Classification. *Comput Vis Pattern Recognit (CVPR)*, 2012 IEEE Conf. 2012; 3642–3649.
[34] Le QV, Ranzato M, Monga R, Devin M, Chen K, Corrado GS, et al. *Building high-level features using large scale unsupervised learning*. 29th International Conference on Machine Learning. Edinburgh. 2011
[35] NIST Special Publication. XXX-XXX.*DRAFT NIST Big Data Interoperability Framework: Reference Architecture*. National Institute of Standards and Technology. Maryland. 2015.
[36] Levin BO, *Big Data Ecosystem Reference Architecture.* Microsoft Corporation.2013;
[37] *Cloud Analytics Playbook*. Booz Allen Hamillton. 2012.
[38] Ralph B, *IBM Big Data Platform*. IBM Corporation. 2013.
[39] We H, *Hadoop D. SAP and Hortonworks Reference Architecture*. SAP AG. 2014.
[40] *An Enterprise Architect's Guide to Big Data*. Oracle Corporation. 2015.
[41] Lopes N, Ribeiro B. GPUMLib: An Efficient Open-source GPU Machine Learning Library.*J Comput Inf Sys*. 2011;3:355–362.
[42] NIST Special Publication. 1500-3. *DRAFT NIST Big Data Interoperability Framework : Use Cases and General Requirements*. National Institute of Standards and Technology. Maryland. 2015
[43] NIST Special Publication. XXX-XXX. *DRAFT NIST Big Data Interoperability Framework:Architectures White Paper Survey*. National Institute of Standards and Technology. Maryland. 2015.
[44] NIST Special Publication. 1500-2.*DRAFT NIST Big Data Interoperability Framework: Big Data Taxonomies DRAFT NIST Big Data Interoperability Framework*. National Institute of Standards and Technology. Maryland. 2015.