Journal of Computer Science 10 (12): 2658-2665, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.2658.2665 Published Online 10 (12) 2014 (http://www.thescipub.com/jcs.toc)

FROM DATA MINING AND KNOWLEDGE DISCOVERY TO BIG DATA ANALYTICS AND KNOWLEDGE EXTRACTION FOR APPLICATIONS IN SCIENCE

Subana Shanmuganathan

Auckland University of Technology (AUT), New Zealand

Received 2014-10-13; Revised 2014-10-15; Accepted 2014-11-28

ABSTRACT

“Data mining” for “knowledge discovery in databases” and associated computational operations first introduced in the mid-1990 s can no longer cope with the analytical issues relating to the so-called “big data”. The recent *buzzword* big data refers to large volumes of diverse, dynamic, complex, longitudinal and/or distributed data generated from instruments, sensors, Internet transactions, email, video, click streams, noisy, structured/unstructured and/or all other digital sources available today and in the future at speeds and on scales never seen before in human history. The big data also being described using 3 Vs, volume, variety and velocity (with an additional 4th V for “veracity” and more recently with a 5th V for “value”), requires a set of new technologies, such as high performance computing i.e., exascale, architectures (distributed or grid), algorithms (for data clustering and generating association rules), programming languages, automated and scalable software tools, to uncover hidden patterns, unknown correlations and other useful information lately referred to as “actionable knowledge” or “data products” from the massive volumes of complex raw data. In view of the above facts, the paper gives an introduction to the synergistic challenges in “data-intensive” science and “exascale” computing for resolving “big data analytics” and “data science” issues in four main disciplines namely, computer science, computational science, statistics and mathematics. For the realisation of vital identified foundational aspects of an effective cyber infrastructure, basic problems need to be addressed adequately in the respective disciplines and are outlined. Finally, the paper looks at five scientific research projects that are urgently in need of high performance computing; this is in contrast to the earlier situations where private business enterprises were the drivers of better modern and faster technologies.

Keywords: Unstructured Data, High Performance Computing, Data Science

1. INTRODUCTION

“Data mining” for “knowledge discovery in databases” and associated computational operations first introduced in the mid-1990 s can no longer be used to analyse the so-called “big data”. This more recent *buzzword* refers to large volumes of diverse, dynamic, complex, longitudinal and/or distributed data generated from instruments, sensors, Internet transactions, email, video, click streams, noisy, structured/unstructured and/or all other digital sources available today and in the future (Fan and Bifet, 2012). The big data also described using 3 Vs for volume, variety and velocity (with an additional 4th V for “veracity” and lately with a 5th V for “value”) requires a new set of technologies, such as high performance computing i.e., exascale, architectures (distributed or grid), algorithms (for data clustering and generating association rules describing the relationships between different variables), programming languages, automated and scalable software tools, to uncover hidden patterns, unknown correlations and other useful information or lately referred to as “actionable knowledge” or “data products” embedded in massive volumes of complex raw data (O'Leary, 2013; Cuzzocrea and Gaber, 2013; Chen *et al*., 2013; Demchenko *et al*., 2013).

Up until very recently, business enterprises have been in the forefront in demanding faster and intelligent data processing methodologies for producing up-to-the-second information on almost every aspect of their business for critical decision making. In contrast, today high performance computing is required for scientific research for understanding vital complex phenomena relating to national security i.e., nuclear stockpile, climate change, neutron fission, medical device innovations and more than anything else to understand the brain anatomic and physiological functioning (Arocena *et al*., 2013; Stevens, 2013; ASCRCM, 2013). These projects are urgently in need of the next stage of development in high performance computing known as Exascale (i.e. 1018 operations per second or 1018 bytes of storage, 100 times faster than the currently available high performance computing).

In view of the above facts, the paper looks at some major synergistic challenges we are facing today in “data-intensive science” and “exascale” computing spheres for resolving present/ future global critical issues relating to new domains called “big data analytics” and “data science”. The issues are described to be associated with four main disciplines namely, computer science, computational science, statistics and mathematics. For the realisation of some vital identified foundational aspects of an effective cyber infrastructure, basic problems need to be addressed adequately in the respective disciplines and are detailed in the paper. Finally, the paper looks at five scientific research projects that are urgently in need of high performance computing in contrast to earlier situations where private business enterprises were the drivers of better, modern and faster technologies.

2. DATA MINING AND KNOWLEDGE DISCOVERY OF THE 1990 s

Data mining, first introduced in the mid-1990 s, was defined as ”…a step in the KDD process that consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data…”. Meanwhile, Knowledge Discovery in Databases (KDD) was defined as the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in the raw data (Fayyad *et al*., 1996). The following are the major approaches still in use for DM and KDD.

Classification trees and rules: This is considered as a popular technique in data mining. It is used to classify a dependent categorical variable based on measurements of one or more predictor variables. The results produced in a tree form with nodes and links between the nodes, can be also converted into *if* and *then* rules.

Logistic regression: This is a statistical technique that is little different to the standard regression however, it extends the concept to deal with classification. The logistic regression produces a formula that predicts the probability of the occurrence as a function of the independent variables.

Artificial neural networks (ANNs): ANNs are biologically inspired networks of basic units called neurons that mimic animal and human brain structure and functioning to process information. ANNs provide a means to incorporate heuristics into conventional computational modelling. A neural network consists of 3 basic layers (an input, one/many hidden and an output), of nodes that are connected to each other based on the chosen network architecture. By a trial and error method, the network learns the correlations between the input and output pairs and then applies them to predict unknown output for new sets of input data. There are different learning and recall algorithms to train and use the knowledge learned by the ANN. The ANN’s black box approach to resolving problems came under criticisms and this major drawback was eventually overcome by methods introduced to extract knowledge from the trained networks in the form of rules.

Clustering techniques: They are useful in identifying groups of similar records within the raw data. For example, the K-nearest neighbour technique calculates the distances between the record and points in the historical (training) data and it then assigns this record to the class of its nearest neighbour in the data set.

3. BIG DATA ANALYTICS AND DATA SCIENCE

Big data, one of today’s buzzwords and it is being used to describe massive volumes of both structured and unstructured data. The volume of big data is exceptionally large and so it is difficult to process this data using traditional database and software techniques (Zikopoulos *et al*., 2012). Big data can be defined as large, diverse, complex, longitudinal and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams and/or all other digital sources available today and in the future (NSF, 2014).

Meanwhile, the area of study called "data science" relates to the generalised extraction of knowledge from data though it incorporates various elements and builds on techniques and theories from many fields, not just statistics. The fields currently used include; signal processing, mathematics, probability models, machine learning, computer programming, data engineering, pattern recognition and learning, visualization, uncertainty modelling, data warehousing and high performance computing. The main goal is to extract meaningful information and hence called as the creation of data products. Data Science need not be always for big data, however, the fact that data is scaling up makes big data an important aspect of data science (Dhar, 2012). In this context, recent big data initiatives and discipline specific needs and issues are discussed here onwards.

Directorate for Computer and Information Science and Engineering (CISE), National Science Foundation (NSF), USA has recently (2012-2013) initiated funding to boost research in Core Techniques and Technologies for Advancing Big Data Science and Engineering (BIGDATA) (DCISE, 2014). The BIGDATA solicitation program facilitated by CISE described the program aim as "to advance the core scientific and technological means of managing, analysing, visualising and extracting useful information from large, diverse, distributed and heterogeneous data sets so as to: Accelerate the progress of scientific discovery and innovation; lead to new fields of inquiry that would not otherwise be possible; encourage the development of new data analytic tools and algorithms; facilitate scalable, accessible and sustainable data infrastructure; increase understanding of human and social processes and interactions; and promote economic growth and improved health and quality of life".

Collaborative environments have been suggested to direct the big data research for dealing with recent scientific research specific needs. To achieve success through this proposition the following have been envisaged; data analytics and interpretation to be highly interdisciplinary requiring collaborations that will need techniques for representations, new modelling techniques and tools that allow for collaborations across individuals looking at complex data sets or across disciplines using multiple representations that make sense within the respective disciplines. These were described as foundational aspects of an effective cyber infrastructure and these basic problems have to be addressed in the respective disciplines urgently to overcome some major barriers encountered in contemporary scientific research efforts.

In the above initiatives, the steering committee also identified the following as urgently needed for the effective use of big data in biomedical applications:

- Various techniques for analysing structural and functional correlatives, interactions and networks, various protein interaction networks, network of neurons
- Example analysis on social media for understanding local, original and national health
- New techniques for mining literature and other types of data to get an understanding of the biomedical research landscape, techniques for analysing multiple clinical research data sets
- Predictive modelling in biology that are related to human health and treating disease
- Clinical science to generate hypotheses using already available background knowledge
- New techniques for disseminating scientific knowledge beyond traditional publications. Methods to link the publications to data sets, simulations one can actually replicate the study reported in the publication

An effective use of big data within neonatal intensive care units has been presented in (McGregor, 2013). The big data has been described as having great potential to support a new wave of clinical discovery leading to earlier detection and prevention of a wide range of deadly medical conditions. The lack of medical informatics research has hampered progress required in real-time analysis of high-frequency physiological data to further improve healthcare as genomics research. The following are the areas identified where more research is urgently required:

- Online extraction, secure transmission, processing and storage of massive streaming data along with additional derived data/information and abstracts
- Easy access to retrospective and prospective clinical studies using systemic and standardised approaches
- Better visualizations for the various critical care roles
- Store and provide as and when needed new discoveries through clinical trials

- Additional methods to translate extracted knowledge to the patient bedside without much more complications
- Additional quality-improvement initiatives to clinical treatment and training guidelines
- Additional tools to monitor healthcare service through the adoption of new clinical guidelines

In view of resolving the online streaming big data analysis issues, a conceptual framework for knowledge-discovery life-cycle for Big Data (**Fig. 1**) has been put forward in (Chen *et al*., 2013). The major stages in this conceptual framework are:

- Data generation
- Data processing and organisation
- Mining, discovery and predictive modelling
- Derive insights, refine, plan and execute

The design of these stages will have to adequately support operations/functions to incorporate three other online processes in parallel, the other processes being (1) big data feeder, (2) data management and (3) historic data analysis, modelling and prediction. The new knowledge-discovery life-cycle for big data requires the compute-intensive architecture to be data-intensive computer (**Fig. 2**) as suggested in (Chen *et al*., 2013) in the 2017 timeframe. Accordingly, it requires shifts in the performance parameters (**Fig. 3**), such as data references, bus accesses, instruction decodes, resources related stalls, ALU instructions, L1 and L2 misses, branches and branch mis-predictions.

Finally, the framework as well have to meet the different data-generation requirements for different domains. For example, smaller distributed instruments, such as field work sensors, i.e., atmospheric, environmental monitoring, plant response sensing, will have to deal with the processing of a massive number of distributed devices and sensors. They will require local processing/derivations/analytics and integration of massive data (possibly at an exascale system/data centres. Storage and post processing of such instruments will have to deal with raw data, derived data subsets and distributed copies. Sharing and distribution will involve a large number of geographically distributed scientists. The visualisation aspects could be massive, with high dimensional large scale graphs, patterns, clustering and scalability. Meanwhile, Discovery Oriented Exascale Simulations, such as climate, cosmology, requires Integration of data generated from simulation and observations. Data reduction for post processing, Time series, statistics and advanced analytics will be the processing issues. Storage and post processing will have to deal with raw data and well organized DBs both enabled for queries. Here again, the data storage and sharing requirements will be high involving a large number of geographically distributed scientists. Visualisation will require pattern detection, correlation, clustering, ensemble visualisation and uncertainty.

All the above requirements point to the computer science side of issues, such as constraints relating to hardware which is reaching a critical phase as we prepare for post Moore Era, in a race for the next level performance/Exascale systems (Snir *et al*., 2011).

In view of the above facts, Defence Advanced Research Projects Agency (DRAPA) and Ubiquitous High Performance Computing (UHPC) launched research aimed at achieving peta-scale performance in a single rack system consuming only 57 KW. Meanwhile, in 2008 report presented in (Kogge, 2008) looked at the major constraints with regard to achieving Exascale X1000 in 2015; not just high end, floating point intensive and supercomputers ("exoflops" machines) but across the board spectrum of three classes (1) data center-sized systems (2) departmental-sized systems and 3) embedded systems and identified the following challenges.

- Energy and Power Challenge
- Memory and Storage Challenge
- Concurrency and Locality Challenge
- Resiliency Challenge

The areas suggested for more research in the co-development and optimization of Exascale were:

- Hardware Technologies and Architectures
- Architectures and Programming Models
- Algorithms, Applications, Tools and Run-times
- Development of a deep understanding of how to architect Resilient Exascale Systems.

The report suggested the following phased research agenda to accelerate the research:

- A System Architecture Exploration Phase
- A Technology Demonstration Phase
- A Scalability Slice Prototyping Phase

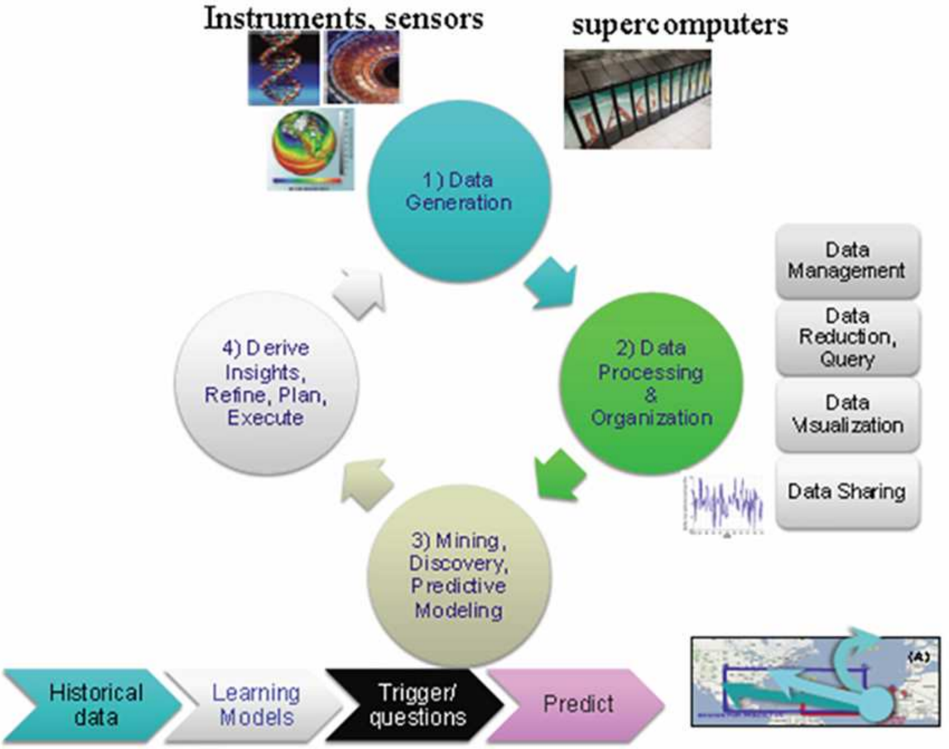
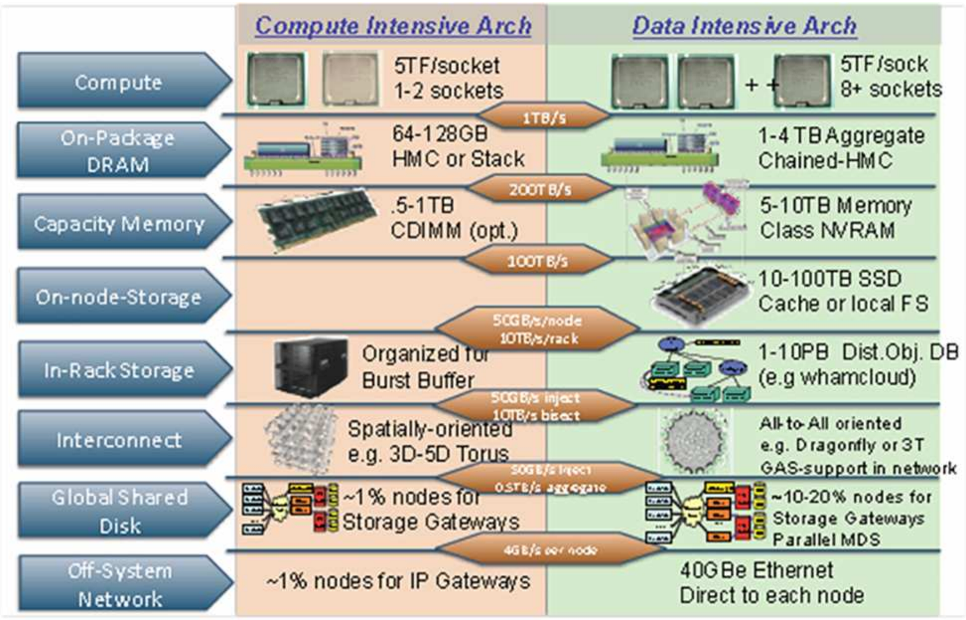
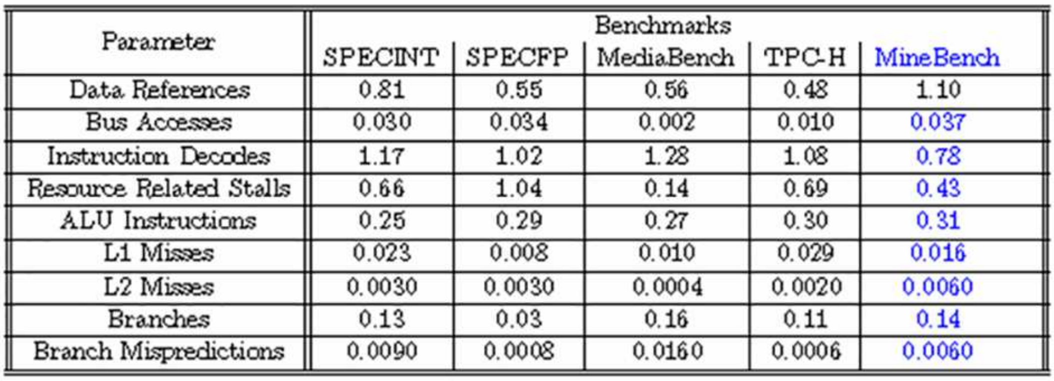


Fig. 1. A knowledge-discovery life-cycle for big data (Chen *et al*., 2013)



	Compute Intensive Arch	Data Intensive Arch
Compute	5TF/socket 1-2 sockets	5TF/sock 8+ sockets
On-Package DRAM	64-128GB HMC or Stack	1-4 TB Aggregate Chained-HMC
Capacity Memory	.5-1TB CDIMM (opt.)	5-10TB Memory Class NVRAM
On-node-Storage		10-100TB SSD Cache or local FS
In-Rack Storage	Organized for Burst Buffer	1-10PB Dist.Obj. DB (e.g whamcloud)
Interconnect	Spatially-oriented e.g. 3D-5D Torus	All-to-All oriented e.g. Dragonfly or 3T GAS-support in network
Global Shared Disk	~1% nodes for Storage Gateways	~10-20% nodes for Storage Gateways Parallel MDS
Off-System Network	~1% nodes for IP Gateways	40GBe Ethernet Direct to each node

Fig. 2. Strawman compute-intensive vs. data-intensive computer architectures in the 2017 timeframe (Chen *et al*., 2013)



Parameter	Benchmarks				
	SPECINT	SPECFP	MediaBench	TPC-H	MineBench
Data References	0.81	0.55	0.56	0.48	1.10
Bus Accesses	0.030	0.034	0.002	0.010	0.037
Instruction Decodes	1.17	1.02	1.28	1.08	0.78
Resource Related Stalls	0.66	1.04	0.14	0.69	0.43
ALU Instructions	0.25	0.29	0.27	0.30	0.31
L1 Misses	0.023	0.008	0.010	0.029	0.016
L2 Misses	0.0030	0.0030	0.0004	0.0020	0.0060
Branches	0.13	0.03	0.16	0.11	0.14
Branch Mispredictions	0.0090	0.0008	0.0160	0.0006	0.0060

Fig. 3. A comparison of selected performance parameters for deferent benchmarks with data analytics and mining workloads (Ozisikyilmaz *et al*., 2006)

The present day Complementary Metal-Oxide-Semiconductor (CMOS) technology, widely utilized in manufacturing digital integrated circuits or IC's in microprocessors, memories and digital Application-Specific Integrated Circuits (ASIC) is slowing down without a viable alternative technology to continue (Snir, 2011). Based on Stein's Law, something cannot go on forever, 7.5 nm feature size for CMOS is considered as the limit and getting there by 2024 is described as "needs several small miracles" (Snir, 2011). With CMOS Technology, power consumption is the big challenge, its leakage or static component will become a major industry crisis between now and 2024, threatening the survival of CMOS Technology itself, similar to the bipolar technology that was in the end discarded some decades ago. Even if 7.5 nm feature size for CMOS is achieved, the technology will plateau in the next decade or so without no immediate replacement. The other technologies being investigated, such asspintronics, Rapid Single Flux Quantum (RSFQ) Logic (requires cryogenic cooling) have their own issues.

In this context, one may wonder why we need high performance computing. The answer is, the world needs high level performance systems by the end of this decade or early next the latest to achieve progress in the simulation of the following as suggested in (Snir *et al*., 2011).

Societal impacts of weather, environmental change: Using large scale simulations, it is possible to significantly increase our ability to develop alternative energy sources. The high performance computing could aid simulations of novel concepts and designs to increase the efficiency of current energy consumption in turn reducing global warming. Contemporary assessments on the damage caused by global warming has significant error margins and it has made the implementation of any single mitigation measure impossible. In the meantime, both inaction and inadequate action for mitigating global warming or adapting to it is estimated costs millions of lives and trillions of dollars. With Exascale climate simulations the error margins can be reduced significantly provided that simulation applications are made available at affordable costs and timeframe (running time).

Continued certification of nuclear stockpile: The need to certify the nuclear stockpile based on large scale simulations alone has been a main motivator for DOE (Department of Energy). Such large scale simulations are found to be useful when approving the current nuclear weapons that are divergent from already tested designs. They provide a means to overcome the knowledge gap though more accurate simulations with a broader range of scenarios and models for better uncertainty quantification however, invariably require fast growth in the performance of supercomputers. The other benefit is that reduction in size of the nuclear stockpile can be achieved through large scale simulations.

Combustion simulation: A major portion (70%) of the crude oil consumed each day is used in internal combustion engines that also produce undesirable emissions, e.g., nitric oxides, particulates and CO2 production. In an atmosphere of worldwide pressure to mitigate the negative environmental and health implications, drastic changes in the fuel constituents and operational characteristics of automobiles and trucks are expected to take place over the next few decades. The world is required to transition away from petroleum-derived transportation fuels. The urgent for a concerted effort to develop non-petroleum-based fuels and their efficient, clean utilization in transportation is warranted by concerns over energy sustainability, energy security, competitiveness and global warming. There is legislation that mandates reductions in fuel usage per kilo meter by 50% in new vehicles by 2030 and greenhouse gases by 80% by 2050.

National security: Continuing the progress towards achieving higher performance is essential for national security. In this era, information is considered as the main weapon of war, hence the US and other super powers view that they cannot afford to be out computed any more than they can afford to be outgunned. The use of supercomputing in national security is, for obvious reasons, not well documented in the open literature when compared with its use in science and engineering. However, national security agencies are the major customers of leading supercomputing systems similar to any other technology and this will continue in the foreseeable future.

To understand brain functioning to learn about the brain disorders: The general view of scientists involved in this significantly important venture called the Human Brain Project (HBP) is that "… current computer technology is insufficient to simulate complex brain function. But within a decade, supercomputers should be sufficiently powerful to begin the first draft simulation of the human brain" (Markram, 2014a). The HBP is "… an attempt to build completely new computer science technology that will enable us to collect all the information we have built up about the brain over the years," (Markram, 2014b). This is anticipated to "…open a whole new array of possibilities in the diagnosis and treatment of brain disorders and diseases, such as Alzheimer´s, Parkinson`s and Schizophrenia" (Markram, 2014a).

Meanwhile, US President Barack Obama has launched a $100 m project to map the "enormous mystery" of the human brain. Brain Research through Advancing Innovative Neurotechnologies (BRAIN) project is hoped to aid us to understanding how the brain works and to learn more about diseases such as Alzheimer's (Mooney, 2013).

As it appears, as far as brain research is concerned despite the extensive scientific understanding of how the human body works gained over the recent decades, the brain functioning still remains elusive. However, more recent attempts through new technologies have enabled scientists to peer deep inside a living brain that are leading to discoveries about our most complex organ. The efforts being described as among ground breaking have led to explore and learn more on how the brain's functions are connected to behaviour and disease and many more, such as how memories are encoded and how to enhance our cognitive abilities by playing video games, building robotic technology controlled by one's thoughts.

4. CONCLUSION

The paper looked at the major issues relating to the use of Data Mining (DM) and Knowledge Discovery in Databases (KDD) methodologies for big data analysis. Big data requires a set of new technologies, such as high performance computing i.e., exascale, architectures, algorithms, programming languages, automated and scalable software tools, to uncover hidden patterns, unknown correlations and other useful information or "actionable knowledge" or "data products" from the massive volumes of complex raw data. The new big data analytics technologies are needed for five important scientific research projects in contrast to the earlier situations where private business enterprises were the drivers of better modern and faster technologies.

5. ACKNOWLEDGEMENT

The author wishes to acknowledge AUT's support extended to attend the conference.

6. ADDITIONAL INFORMATION

6.1. Funding Information

The 2nd International Conference on Computational and Network Technologies 2013 (ICCNT 2013) Organisers are acknowledged for funding the author's travel to present the paper at the conference in Perth, Australia.

6.2. Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

7. REFERENCES

Arocena, P.C., R.J. Miller and J. Mylopoulos, 2013. The Vivification Problem in Real-Time Business Intelligence: A Vision. In: Enabling Real-Time Business Intelligence, Castellanos, M., U. Dayal and E.A. Rundensteiner (Eds.)., Springer Berlin Heidelberg, ISBN-10: 978-3-642-39871-1, pp: 37-49.

ASCRCM, 2013. Accelerating Scientific Knowledge Discovery Working Group Report. US Department of Energy (DOE) Office of Advanced Scientific Computing.

Chen, J., A. Choudhary, S. Feldman, B. Hendrickson and D. Williams *et al*., 2013. Synergistic challenges in data-intensive science and exascale computing. DOE ASCAC data subcommittee report. Office of Science, U.S. Department of Energy.

Cuzzocrea, A. and M.M. Gaber, 2013. Data Science and Distributed Intelligence: Recent Developments and Future Insights. In: Intelligent Distributed Computing VI, Fortino, G., C. Badica, M. Malgeri and R. Unland (Eds.)., Springer Berlin Heidelberg, ISBN-10: 978-3-642-32523-6, pp: 139-147.

DCISE, 2014. The National Science Foundation. BIGDATA Webinar. Directorate for Computer and Information Science and Engineering.

Demchenko, Y., P. Grosso, C. de Laat and P. Membrey, 2013. Addressing big data issues in scientific data infrastructure. Proceedings of the International Conference on Collaboration Technologies and Systems, May 20-24, IEEE Xplore Press, San Diego, CA, pp: 48-55. DOI: 10.1109/CTS.2013.6567203

Dhar, V., 2012. Data Science and Prediction. KD Nuggets.

Fan, W. and A. Bifet, 2012. Mining big data: Current status and forecast to the future. ACM SIGKDD Exp. Newslett., 14: 1-5. DOI: 10.1145/2481244.2481246

Fayyad, U.M., G. Piatetsky-Shapiro and P. Smyth, 1996. From Data Mining to Knowledge Discovery: An Overview. In: Advances in Knowledge Discovery and Data Mining, Fayyad, U.M. (Eds.)., AAAI Press, Menlo Park, ISBN-10: 0262560976, pp: 611.

Kogge, P., 2008. ExaScale computing study: Technology challenges in achieving exascale systems. DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod. AFRL contract number FA8650-07-C-7724. This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

Markram, H., 2014a. Billion pound brain project under way.

Markram, H., 2014b. Unraveling the human brain. European Media Center.

McGregor, C., 2013. Big data in neonatal intensive care. Computer, 46: 54-59.

Mooney, A., 2013. Obama seeks $100M to unlock mysteries of the brain, CNN

NSF, 2014. Directorate of Computer and Information Science and Engineering (CISE). Event BIGDATA webinar/Core Techniques and Technologies for Advancing Big Data Science and Engineering (BIGDATA). National Science Foundation.

O'Leary, D.E., 2013. Artificial intelligence and big data. Intelli. Syst., 28: 96-99. DOI: 10.1109/MIS.2013.39

Ozisikyilmaz, B., R. Narayanan, J. Zambreno, G. Memik and A.N. Choudhary *et al*., 2006. An architectural characterization study of data mining and bioinformatics workloads. Proceedings of the IEEE International Symposium on Workload Characterization, Oct. 25-27, IEEE Xplore Press, San Jose, CA, pp: 61-70. DOI: 10.1109/IISWC.2006.302730

Snir, M., 2011. Exascale and climate simulation; The evolution of HPC architectures and the implications for climate simulation.

Snir, M., W. Gropp and P. Kogge, 2011. Exascale research: Preparing for the post-moor era. University of Illinois.

Stevens, R., 2013. Statement of rick Stevens. Speech made before the subcommittee on energy of the committee on science, space and technology. U.S. House of Representative.

Zikopoulos, P.C., C. Eaton, D. deRoos, T. Deutsch and G. Lapis, 2012. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. 1st Edn., McGraw Hill Professional, New York, ISBN-10: 0071790535, pp: 176.