Review: Big Data Techniques of Google, Amazon, Facebook and Twitter

Thulara N. Hewage, Malka N. Halgamuge, Ali Syed, and Gullu Ekici
School of Computing and Mathematics, Charles Sturt University, Melbourne, Victoria 3000, Australia
Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, VIC 3010
Email: malka.nisha@unimelb.edu.au, {ASyed, gekici}@studygroup.com

Abstract—Google, Amazon, Facebook and Twitter gained enormous advantages from big data methodologies and techniques. There are certain unanswered questions regarding the process of big data, however, not much research has been undertaken in this area yet. This review will perform a comparative analysis based on big data techniques obtained from sixteen peer-reviewed scientific publications (2007-2015) about social media companies such as Google, Amazon, Facebook and Twitter to undertake a comparative analysis. Google has invented many techniques by using big data methods to strategize against competitors. Google, Facebook, Amazon and Twitter are partially similar companies that use big data despite their own business model requirements. As an illustration, Google required the data "ware housing" approach to store trillion of data related to Facebook, since Facebook owns more than one billion users and Twitter owns 300 million active users correspondingly equally to Amazon. Since all these organization required data ware house approach, Google has preferred the variation of data ware house storages (Spanner, Photon, Fusion table) variation of data transaction methods. By using these data ware house storage approaches (F1 for execute queries via SQL) and communication of different approached such as, Yedalog. Facebook and Twitter are both the only social media companies that have different requirements. The requirement of big data is high and these entire requirements partially depend on each another as it is completely isolated. This study is a useful reference for many researchers to identify the differences of big data approaches and technological analysis in comparison to Google, Facebook, Twitter and Amazon big data techniques and outline their, variations and similarities analysis.

Index Terms—Big data, big data techniques, amazon, social networks

I. Introduction

The accustomed computers need manual setups to retrieve information, they are also designed to learn through human interaction and, through continuous feedback, and essentially, they do not reprogram themselves. However, the revolutionary big data provides comprehensive predictive software that illuminates the effort requiring feedback and planning that eventually becomes unfeasible with demand. Further, big data strategize futuristic business moves with user performance analytics. In the early 2000s the concept of big data aroused, especially in the earlier periods, as storage of data was a problem. However, the newly invented technologies such as Hadoop, MapReduce provided some solution for particular issues. Big Data provides solution as it reduces cost, time, and provides prospect to opens ways for new products. For accurate decision making, companies have become reliant on big data. Big data provides skilled analytics and instantly that deciphers failures and problems that may occur in near future or in current situations. Simply, big data detects risks early allowing time to take preventative measures.

Companies like Google, Facebook, Amazon and Twitter, have their entire core functions related to big data. This paper discusses whether and how Google, Facebook, Amazon and Twitter apply big data techniques to their business models. These companies have different business models, as Google is a search engine based business model. However, organizations like Google are prone to use big data techniques in with regard to improve their storage ability with accurate output that captures client queries, and maintains query logs and etc. Facebook organization (social network) business model is completely different when compared to search engine based organization (Google) yet these companies rely on big data since the main necessities are storage, analysis and accuracy.

Based on scientific results presented in the literature; Google File System has successfully overcome the problem of Google data storage. GFS is the largest cluster that provides hundreds of terabytes of storage across thousands of disks over thousand machines and continuously requests [1] client requirements. Spanner is a Google database that globally distributes the Database. This product can share data within machines across the world within data centers and provides intelligence ability to respond to failures and balance the requested load. This development expands into millions of machines across hundreds of datacenters and trillion [2] of database rows. Another development of Google is, "Google's F1 database management system" provides consistency and high availability allowing the user to execute queries via SQL. Google advertisers can bid, budgets, get involved in campaigns that change and provide immediate feedback

Manuscript received May 19, 2017; revised January 15, 2018.
Corresponding author email: malka.nisha@unimelb.edu.au.
doi:10.12720/jcm.13.2.94-100

perform primary user [4] events (search queries). Colum store database concept internally started by Google and became very famous in the last decade. The concept is beneficial as it allows investigation of lager sets in datasets (billions of rows, log records) within few seconds. The columns store concepts that could work on thousands of machines known as [5] Dremel. However, Yedalog allows programmers to code, process and to write data on the same pipeline, and the same formation could run in different platforms [6].

Google invented Hive to write codes on map reduce program as an open source data warehouse solution. Allows SQL based queries – HiveQL [7] allow customaries to map reduce script plug into quires. Facebook has more than 1.59 billion active users at the moment, and Facebook has also invented Scuba as a data management tool. This product at the moment consumes millions of rows per second and expires millions of data per second. Facebook owns more than hundred servers with the capacity of 144 GB RAM [8] therefore Scuba store data in the memory of these servers. As a solution for high growth of data Amazon built Dynamo product, which is the highest available key value storage system. The Amazon web site functions needs, a primary key to access the best seller list, shopping cart, customer preference, session management, sales rank and product catalog to accomplish the requirements of Dynamo [9] in pursuance to provide an interface with a simple primary key. Twitter is a social network system with 1.3 billion active users, and in the first stages, Twitter used application specific logging system nevertheless they have introduced unified log format. When the analytics task considers the client session as basis of analysis Twitter comes up with session sequences. This method summarizes answers for large classes of common queries [10] as much as possible. The results from these studies have not been without controversy. As discussed above the mentioned techniques selected to draw data from research papers (2007-2015) in regard to eliminate these controversies.

II. Material and Methods

In our analysis, we include research experiments results from big data techniques used by Google, Facebook, Twitter and Amazon. We further investigated big data methods and techniques such as MapReduce, Paxos, Flattening technique, SQL, Tree Structured data model, Spark, Hadoop, Classical Reed Solomon codes, etc.

The raw data presented in Tables I-III specifies the variables that were in the Google, Facebook, Twitter and Amazon. Identification of big data techniques and how Google, Facebook, Twitter and Amazon uses these techniques coherently as a business models that has clearly been described in Table I. The demonstration of the analysis of big data techniques used in various companies has also been outlined. The categorization of data concerns and characteristics of techniques demonstrated in Table II. Furthermore, clarification according to the characteristics and techniques of each technology is based on data ware house properties that are described in Table III.

A. Collection of Raw Data

This analysis was made to pool data of scientific research of whether and how big data techniques are used by Google, Facebook, Twitter and Amazon. The analysis excluded the values of scientific research experiments. The technological aspect of big data and their variations, differentiation and similarities of technologies used by Google, Facebook, Twitter and Amazon were included.

B. Analysis and Comparison of Raw Data

For the analysis, we used comparison method to pool big data techniques used by Google, Facebook, Twitter and Amazon. According to the requirements of business and the use of techniques, the technological products are invented by Google, Facebook, Twitter and Amazon. Category of the technique, subsidiaries are the architecture of the technique, data model, the API of the technique, security of the invented product based of the big data technique and portioning and replication of the technology.

C. Descriptive Analysis and Comparison

Big data techniques vary from one another; however there are similarities in many aspects. The technologies we consider in this research analysis and comparison are according to the information technology categorizations. All the other categories are excluded in this research paper nevertheless all the required and provided aspects are included a clear analysis and comparisons are further included. In terms of categorization and comparison of data, the ware house (models of databases) and data communication (query methods), API of each technique, Replication, Architecture is included.

III. Results

Table I illustrates the overview of published articles by year and number of their publications dates. The analyzed content of the study is collected by Authors and published year, Big Data technique name, Big Data techniques used by the company and Description about the technique are also outlined. Table II describes the Author, published year, Big Data technique, Big Data technique used in the company, Base Technology, Categories of the technology and supporting areas. Overview of Author and publication year, Big Data technique name, Big Data technique used company. The conclusion is drawn, followed by recommendations that are provided in Table IV. In our final analysis, we describe our findings and supported areas categorized by the Big Data technologies, performances and techniques that are used by each company displayed in Table III.

TABLE I: THE ANALYSIS OF BIG DATA TECHNIQUES USED BY VARIOUS COMPANIES, DRAWN FROM SIXTEEN PEER REVIEWED SCIENTIFIC ARTICLES PUBLISHED IN 2007-2015.

Author and Published Year	Big Data Technique Name	Big Data Technique used Company	Description about Technique
Corbett *et al*. (2012) [2]	Spanner	Google	- Google Globally Distributed Database. - Shared data set into Paxos[1] state and expands to millions of machines across hundreds of datacenters and trillion of database rows. - Data centers balance the load and respond to failures.
DeCandia *et al*. (2007) [9]	Dynamo	Amazon	- Key value storage system - Provides an interface with simple primary key. - Run continually underneath the failure situations.
Abraham *et al*. (2013) [8]	Scuba	Facebook	- Data management tool - Consume millions of rows of data per second and expires millions of data per second. - Analysis live data
Ghemawat *et al*. (2003) [1]	Google File System	Google	- Storage management. - The largest cluster up to the data provides hundreds of terabytes of storage across thousands of disks over thousand machines and continuously request clients' requirements. - Run under the circumstance of fault tolerance on inexpensive commodity hardware
Lee *et al*. (2012) [10]	Unified Logging Infrastructure for Data Analytics	Twitter	- Introduced unified log format. - Capture messages in common and well-structured format. - Based on session sequences and summarizer large class of common.
Rae *et al*. (2013) [3]	F1	Google	- Relational database management system - Allow user to execute queries via SQL.
Ananthanara yanan *et al* (2013) [4]	Photon	Google	- Make the availability of real time data - Perform primary user's events such as search query with following event.
Melnik *et al*. (2010) [11]	Dremel	Google	- Query engine - Dominance relation and semi flattening - Flattening technique is proposed method for maps.
Madhavan *et al*. (2012) [12]	Fusion Table	Google	- Cloud based data management system. - Sharing, collaboration, exploration, visualization, web publishing and provision visualizations, such as maps, timelines, and network graphs which can be implanted on any web belongings.
Gupta *et al*. (2015) [13]	F1, Mesa and Photon	Google	- Processing and maintain advertisement related facts and send critical report to Google's Ad user and clients. - Including performance of their Ad campaigns and budgeting of the live serving system
Hall *et al*. (2012) [5]	Processing a Trillion cells per mouse click	Google	- Colum store database technology. - Used OLAP or OLTP like SQL interfaces - Additional approach for establish products like MonetDB [14], Netezza [15] and QlikTech [16]
Afrati *et al*. (2014) [17]	Dremel	Google	- TreeStructured data model. - One or more than one relations. - Example JSON data format, Google's protocol buffers[17], Nested relations recent developments (combination of relational and TreeStructured) Dremel, F1
Chin *et al*. (2012) [6]	Yedalog	Google	- Google finding solution to assemble digital knowledge and search engine query logs. -MapReduce and Spark main drawback is not automated. - Search input parsed using dependency
Cheng *et al*. (2014) [18]	Cascades	Facebook	- Prediction of re-sharing pattern - Cascades of re-share content focused on analyzing and characterizing.
Thusoo *et al*. (2009) [7]	Hive	Facebook	- Open source warehouse solution - SQL based queries – HiveQL - Customaries map reduce script plug into quires. - Hive base on Hadoop system.
Maheswaran *et al* (2013) [19]	XORing Elephants	Facebook	- Coding technique use as saving storage with redundancy. - Classical Reed Solomon codes.

TABLE II: THE ANALYSIS OF BIG DATA TECHNIQUES USED BY VARIOUS SOCIAL MEDIA COMPANIES. THE CONCERNING DATA CHARACTERISTICS OF TECHNIQUES FROM SIXTEEN PEER REVIEWED SCIENTIFIC ARTICLES PUBLISHED IN 2007-2015.

Author and published year	Big Data technique used company	Big Data technique	Base Technology	Categories of the technology	Subsidiary area
Rae *et al*. (2013) [3]	Google	F1	Hybrid database technologies	Database	Google Ad Work

Ananthanarayanan *et al*. (2014) [4]	Google	Photon	Query events	Database	Google joint data steams
Hall *et al*. (2012) [5]	Google	Processing a Trillion cells per mouse click	Composite range partition, Column oriented database system, ad hoc queries	System environment	Google single mouse click into trillion datasets producing process.
Afrati *et al*. (2014) [17]	Google	Dremel	Tree Structured, Schemas	Programming model	Google query language features.
Cheng *et al*. (2014) [18]	Facebook	Cascades prediction	Data mining	Programming model	Facebook cascade predicted framework
Sathiamoorthy *et al*. (2013) [19]	Facebook	XORing Elephants	Hadoop HDFS	Programming model	Facebook overcomes Reed-Solomon codes limitations
Thusoo *et al*. (2009) [7]	Facebook	Hive	Hadoop	System environment	Facebook in warehousing solution
Chin *et al*. (2015) [6]	Google	Yedalog	Logic programming, Data structured and nested records	Programming model	Google overcome MapReduce and Spark technology had limitation as man powered coding.
Corbett *et al*. (2012) [2]	Google	Spanner	Versioned key-value store into a temporal multi-version database	Database (data ware house environment)	Data ware house and transaction of data
Ghemawat *et al*. (2003) [1]	Google	Google File System	Traditional file system	System environment	Data ware house and transaction of data
Gupta *et al*. (2015) [13]	Google	F1, Mesa and Photon	Bigtable	Database	Multi-homing
Madhavan *et al*. (2010) [12]	Google	Google Fusion Tables	Google maps	Database	Google map visualizations, interactive maps
DeCandia *et al*. (2007) [9]	Amazon	Dynamo	Primary key, decentralized techniques	Database	Amazon data warehousing
Abraham *et al*. (2013) [8]	Facebook	Scuba	Hadoop	Database	Facebook data warehousing
Lee *et al*. (2012) [10]	Twitter	Unified Logging Infrastructure for Data Analytics	Hadoop-based, System running on a cluster of several	System environment	Twitter system environment solutions

Despite the number of considerable publications, demonstrating the use of big data techniques used by well-known companies, there are still some authors that cancel out the growth and performance of these techniques. Our aim is to provide advance knowledge of big data techniques used by well-known companies. Consequently, we focus on conferred growth of data and performances of techniques since some of the reported positive findings are flawed by data represent limitation and shortcomings.

TABLE III: THE ANALYZE OF BIG DATA TECHNIQUES USED IN VARIOUS COMPANIES. DATA CONCERNING THE DATABASE RELATED CHARACTERISTICS OF TECHNIQUES TAKEN FROM SIXTEEN PEER REVIEWED SCIENTIFIC ARTICLES PUBLISHED IN 2007-2015.

Technique	Architecture	Data model	API (Function)	Security	Partitioning	Replication
Dynamo	Decentralized	Key-value	Get, put	No security	Consistent hashing	Successor nodes in ring
BigTable	Centralized	Multidimensional sorted map	Get, scan, put, delete	Access control	tablet server	Chunk server in GF
Megastore	Semi relational	Access control	Create, update, delete	Access control	Hashing	Synchronous
Spanner	Semi relational, True Time	Schematized	Paxos algorithm	Access control	Hierarchies of tables	Synchronous
F1	Decentralized	Hierarchical schema	Create, update, delete	Access control	Relational	Not applied
Dynamo	Map reduce	Key value	Create, update, delete, etc.	Access control	Multiple (sort key, partition key, etc.)	Synchronous, cross region
Scuba	Hadoop base	Semi-structured and sparse	Ad hoc queries	Access control	Not applied	Not applied
Mesa	Decentralized	Novel batch-update	Paxos algorithm	Access control	Not applied	Not applied

Photon	Decentralized	Near-exact semantics	Paxos algorithm	Access control	Not applied	Synchronous
Google Fusion Tables	Decentralized	Map-reduce	Selection, projection, grouping, aggregation, equijoin	Access control	Not applied	Map data servers

All these techniques are related to the databases. Some of the techniques are deliberated by early coding languages [20] that cascade prediction [18] etc. In above comparison, we discussed database related techniques. According to the company requirements, all the techniques or the inventions have to be adopted into the organization. At the same time within same organization some techniques differ from each other due to the complexity of the organization. As an example, Google has a search engine based company therefore database is a massive requirement [4] on the other hand, the database categorizes into several types of data warehouses [7].

IV. Discussion

The conclusion drawn from this analysis shows that big data techniques are used by various companies. This study has collected data concerning the characteristics and techniques, of particular data methods, after a thorough analysis drawn from sixteen peer reviewed scientific articles published in 2007-2015. Nevertheless, as discussed, the analysis and comparison of Google, Amazon, Facebook and Twitter big data techniques are outlined and contrasted with many other similar research papers. However Advanced Light Source (ALS) is used by MongoDM (a document oriented data store) to understand the performance, scalability and fault tolerance as a comparison of MongoDB and Hadoop, and this is a partially completed understanding of scientific analysis of two partially similar big data techniques [21]. MonogoDB stores Meta data and also provides query language. MapReduce allow users to write map and reduce functions.

This paper has discussed Google File system based on MapReduce technology. Big data can be used for its many technological aspects. Google uses big data for Google maps [22]. Geometry of road map in Tunisha is used for categories and traces of the roads and the use of big GPS data which makes the divider of the data to handle unstructured data [23]. MapReduce has a function called parallel mode and sequential mode. Parallel processes used for the implementation to find roads and trace for vehicles. [14] conducted a research experiment used 10GB raw data for process, using big data technique MapReduce (for continuously growing data). Google and Face, etc process trillion of data in a second or a minute nevertheless some countries gain advantage of big data by implementing popular techniques of big data (MapReduce, Hadoop) Google invented F1, Mesa and Photon used the BigTable technique in big data to provide multi homing facility for Google search engine process. Big Table implemented on top of MapReduce technique. Furthermore, big data techniques connect with each other, as F1, Mesa and Photon are three different products implemented by Google for different functions nevertheless they have been used for the same big data implementation technique. Big data proves advantages that are ample, however there are minor drawbacks.

One drawback of big data classification of modern big data technologies is that it is questionable and challengeable. Big data classification techniques are representation of learning, supervised learning and machine lifelong learning. Big Data technologies are Hadoop, Hive etc. Suggested solution for this challenge is integrated with Hadoop distributed file system with representation learning techniques. Furthermore, this integration solves the prediction network [24] of big data classification strategy and solves continuity parameters.

Twitter for the front-end processes JSON logs can be applicable as a fast solution. Data category, integration hooks, robust data dependency and work flow scheduling schemas are all beneficial nonetheless applying schema techniques and implementing a framework is required. Together JSON logs and schema provide a solution to stand and overcome Twitter data mining [10]. Requirements of big data vary from one organization to another depending on the requirement of the company and techniques in use. We clearly considered this research experiment in the result section as an analysis and as a comparison of views. Furthermore, as an instance Google built Dynamo and BigTable characteristic partially similar in API nonetheless key aspects such as Architected and data model are completely different. Moreover, investigating various techniques for Big Data Databases [20], [22], [25], security, [26]-[28] prediction and pattern analysis [14] could be an interesting path to explore in future.

V. Conclusion

Google, Amazon, Facebook and Twitter have gained enormous returns since big data methodologies and techniques, lack research in this field. In this review, we have completed a comparison-based analysis of big data techniques used by various companies. The data has been collected from sixteen peer reviewed scientific publications from 2007 to 2015. The collected data is in relation to the characteristics and techniques of data storage. The analysis and comparison carried out in the literature discuses different experimental analysis by using the (category of technology, data model, etc.). In this review, we clearly differentiated all big data techniques according to their mode of techniques that differ from each other. Both Facebook and Twitter are social media based companies that have different

requirements. Facebook requires a cascade prediction system and coding language for data transition, besides Twitter requires a system infrastructure for handle millions of Tweets (per minute data transaction). Nevertheless, Facebook and Twitter both require data from a ware house as a solution. When considered, the requirement of Google is similar to Facebook and Twitter, the data ware house solution of Google implemented Google File system and Spanner, etc. Besides that, Google owns YouTube for advertising and marketing, F1 make all the process for budgeting, ad clicking and for customer and user log handling. All these functions connect to big data for connectivity to be used for big data technologies as MapReduce, Hadoop, etc. For instance, XORing Elephants is Facebook programming model Yedalog is Google programming model. Facebook built XORing Elephants for overcome Reed-Solomon codes limitations beside Google built Yedalog for overcome man powered coding limitation had in MapReduce and Spark technologies. This research study is based on comparison and analysis of big data techniques. This study should be useful as a reference for many researchers as this study provides technological analysis and comparison of Google, Facebook, Twitter and Amazon big data techniques.

AUTHOR CONTRIBUTION

T.N.H. and M.N.H. conceived the study idea and developed the analysis plan. T.N.H. analyzed the data and wrote the initial paper. M.N.H. helped to prepare the tables and finalizing the manuscript. All authors read the manuscript.

REFERENCES

[1] S. Ghemawat, H. Gobioff, and S. Leung, "The Google file system," *ACM SIGOPS Operating Systems Review*, vol. 37, no. 5, p. 29, 2003.

[2] J. Corbett, P. Hochschild, W. Hsieh, S. Kanthak, E. Kogan, H. Li, A. Lloyd, *et al.*, "Spanner: Google's globally distributed database," *ACM Transactions on Computer Systems*, vol. 31, no. 3, pp. 1-22, 2013.

[3] I. Rae, E. Rollins, J. Shute, S. Sodhi, and R. Vingralek, "Online, asynchronous schema change in F1," *Proceedings of the VLDB Endowment*, vol. 6, no. 11, pp. 1045-1056, 2013.

[4] R. Ananthanarayanan, S. Venkataraman, V. Basker, S. Das, A. Gupta, *et al*., "Photon: Fault-tolerant and scalable joining of continuous data streams," in *Proc. International Conference on Management of Data - SIGMOD '13*, 2013.

[5] A. Hall, O. Bachmann, R. Büssow, S. Gănceanu, and M. Nunkesser, "Processing a trillion cells per mouse click," *Proceedings of the VLDB Endowment*, vol. 5, no. 11, pp. 1436-1446, 2012.

[6] B. Chin, D. Dincklage, V. Ercegovac, P. Hawkins, *et al.*, "Yedalog: Exploring Knowledge at Scale," in *Proc. 1st Summit on Advances in Programming Languages*, 2015, pp. 63-78.

[7] A. Thusoo, J. Sarma, N. Jain, Z. Shao, P. Chakka, *et al*., "Hive: A warehousing solution over a map-reduce framework," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1626-1629, 2009.

[8] L. Abraham, S. Subramanian, J. Wiener, O. Zed, J. Allen, O. Barykin, *et al*., "Scuba: Diving into data at FaceBook," *Proceedings of the VLDB Endowment*, vol. 6, no. 11, pp. 1057-1067, 2013.

[9] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, *et al.*, "Dynamo: Amazon's highly available key-value store," *ACM SIGOPS Operating Systems Review*, vol. 41, no. 6, p. 205, 2007.

[10] G. Lee, J. Lin, C. Liu, A. Lorek, and D. Ryaboy, "The unified logging infrastructure for data analytics at Twitter," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 1771-1780, 2012.

[11] S. Melnik, A. Gubarev, J. Long, G. Romer, S. Shivakumar, M. Tolton, and T. Vassilakis, "Dremel: Interactive analysis of web-scale datasets," *Communications of the ACM*, vol. 54, no. 6, p. 114, 2011.

[12] J. Madhavan, S. Balakrishnan, K. Brisbin, *et al*., "Big data storytelling through Interactive maps," *IEEE Data Eng. Bull.*, 2012.

[13] A. Gupta and J. Shute, "High-Availability at massive scale: Building Google's data infrastructure for ads," in *Proc. Workshop on Business Intelligence for the Real Time Enterprise*, 2015.

[14] A. Gupta, A. Mohammad, A. Syed, and M. Halgamuge, "A comparative study of classification algorithms using data mining: Crime and accidents in Denver city the USA," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 7, 2016.

[15] IBM Netezza Data Warehouse Appliances – The Simple Data Warehouse Appliance for Serious Analytics. (2017). [Online]. Available: https://www-01.ibm.com/software/data/netezza/.

[16] QlikTech. *MongoDB.* (2017). [Online]. Available: https://www.mongodb.com/partners/software/qliktech

[17] F. Afrati, D. Delorey, M. Pasumansky, and J. Ullman, "Storing and querying tree-structured records in Dremel," *Proceedings of the VLDB Endowment*, vol. 7, no. 12, pp. 1131-1142, 2014.

[18] J. Cheng, L. Adamic, P. Dow, J. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proc. 23rd International Conference on World Wide Web*, 2014.

[19] M. Sathiamoorthy, M. Asteris, D. Papailiopoulos, A. Dimakis, R. Vadali, S. Chen and D. Borthakur, "XORing elephants," *Proceedings of the VLDB Endowment*, vol. 6, no. 5, pp. 325-336, 2013.

[20] Vargas, A. Syed, A. Mohammad, and M. Halgamuge, "Pentaho and jaspersoft: A comparative study of business intelligence open source tools processing big data to evaluate performances," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 10, 2016.

[21] S. Khalid, A. Syed, A. Mohammad, and M. Halgamuge, "Big-Data NoSQL databases: Comparison and analysis of 'Big-Table', 'DynamoDB', and 'Cassandra'," in *Proc.*

IEEE 2nd International Conference on Big Data Analysis, 2017.

[22] K. Kaur, A. Syed, A. Mohammad, and M. Halgamuge, "Review: An evaluation of major threats in cloud computing associated with big data," in *Proc. IEEE 2nd International Conference on Big Data Analysis*, 2017.

[23] S. Suthaharan, "Big data classification," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 4, pp. 70-73, 2014.

[24] J. Lin and D. Ryaboy, "Scaling big data mining infrastructure," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 2, p. 6, 2013

[25] S. Munugala, G. K. Brar, A. Syed, A. Mohammad, and M. N. Halgamuge, "The much needed security and data reforms of cloud computing in medical data storage," *Applying Big Data Analytics in Bioinformatics and Medicine*, IGI Global, Chapter 5, pp. 99-113, February 2017.

[26] D. V. Pham, A. Syed, and M. N. Halgamuge, "Universal Serial Bus Based Software Attacks and Protection Solutions," *Digital Investigation*, vol. 7, no. 3, pp. 172-184, Feb. 2011.

[27] D. V. Pham, A. Syed, A. Mohammad, and M. N. Halgamuge, "Threat analysis of portable hack tools from USB storage devices and protection solutions," in *Proc. International Conference on Information and Emerging Technologies*, pp. 1-5, Karachi, Pakistan, June 14-16, 2010.

[28] D. V. Pham, A. Syed, and M. N. Halgamuge, "Universal serial bus based software attacks and protection solutions," *Digital Investigation*, vol. 7, no. 3, pp. 172-184, Feb. 2011.

Thulara N. Hewage was born in Western Province, Sri Lanka in 1990. She received the B.Sc degree from the University of Middlesex, London, in 2012 and the M.I.T degree from the University of Charles Sturt University (CSU), Melbourne, in 2016, both in Information Technology. Her research interest includes Information Technology enhanced theories, online communication and social media, processes of data analytics and related computer visions.

Malka N. Halgamuge is a Research Fellow with the department of Electrical and Electronic engineering, University of Melbourne. She received the PhD degree from the same department in 2007. Since then she has published in areas including wireless communication, life sciences and data science/big data. She is an experienced researcher and educator with a demonstrated history of working with highly reputed research institutes all over the world on life sciences.

Ali Syed has wide experience as a lecturer and examiner of undergraduate and postgraduate courses in the field of Information Systems Management, Business Studies and has been actively involved with course development and delivery. He is a member of several academic and practitioner communities. His experience and involvement with the Industry contributes a strong flavor to Academia. His research interests are in the areas of Systems development, Systems Security, Knowledge Management, Ethical Issues in Information Systems and the implications of information systems for people and their work environments.

Gullu Ekici is a PhD. Monash University (in progress - due to complete in 2018). Masters of Education, (TESOL), Monash University. Bachelor of Arts (Linguistics), Monash University. Monash University I have more than 15 years' experience in teaching/researching and academic skills support experience at tertiary level in a range of settings.