(An appropriate title of the report /5)

*Note: Sub-titles are not captured in Xplore and should not be used

Stone Fang (Student ID: 19049045)
Computers and Information Sciences
Auckland University of Technology
Auckland, New Zealand
fnk7060@autuni.ac.nz

***Abstract*—xxxxxx xxxxxxxxxxxxxxxxx xxxxxx > Abstract /10 > The abstract should be one paragraph and it should cover the whole theme of the report.**

This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.

***Index Terms*—Big Data, scalability**

I. Introduction

/10 The introduction part should introduce the topic and its importance. It should also include at least one diagram related to the topic. One page is sufficient for this section.

A. ~~what is Big Data.~~

In recent years, exponential increasing volumes of data have been generated by a variety of sources such as e-businesses, communications, mobile phones, social media sites, web servers, sensor networks, cameras, banks, stock markets, and so on [1], [2]. Nowadays, data are being generated at an unprecedented rate, bringing us into a era of Big Data or "data deluge" [2], [3]. The types of data vary from text to multimedia including image, audio, and video. The format can be structured, semi-structured and unstructured. In the real world, more than 90% of the overall data generated are unstructured [2].

B. ~~value, importance~~

Big Data has enormous potential values with the expectation of transforming the humans society, and is consequently regarded as the "new oil" by some researchers [3]. It not only can brings large amount of revenues to businesses and values to consumers, but also has great potential applications in a range of industries. For instance, health or medical data analysis has many benefits including personalised health service, disease evolution monitoring, and adaptive public health plans [1]. Another example is the real-time analysis of data generated by smart meters, sensors and control devices on smart grid, which can help in incident detection, risk identification, and energy consumption forecast [1].

> "Google, Amazon, Facebook and Twitter gained enormous advantages from big data methodologies and techniques." [4]
>
> "For instance, the opportunities include value creation (Brown, Chui, & Manyika, 2011), rich business intelligence for better- informed business decisions (Chen & Zhang, 2014), and support in enhancing the visibility and flexibility of supply chain and resource allocation (Kumar, Niu, & Re, 2013)." [2]

C. ~~challenges, especially on scalability~~

Before Big Data evolution, it is difficult to store, manage or analyse data sets in large volumes because of the limited store capacity and lack of scalability, flexibility and performance in traditional technologies [1]. Relational database management system (RDBMS), the main technology in traditional data management, does not fit the requirements in Big Data analysis. The reasons of this mismatch are twofold, including:

- Data structure: RDBMS only supports structured data, while has little capability in storage and analysis of semi-structured and unstructured data [3].
- Scalability: RDBMS only scales up at high costs of hardware, and is very difficult to scale out, which makes it incapable in continuously growing data scenarios [2], [3].

D. ~~popularity~~

As a result of the big values and big challenges inside Big Data, interests from both academia and industry are dramatically increasing in recent years. A wide range of issues have been studied at different levels including data storage, cleaning, analysis, visualisation, and so on, some of them still open to research [1]. In the industry, many companies have their own Big Data platforms, for example, Google's large data storage Google File System(GFS) and cloud based data management system Fusion Table [4]. Many Big Data systems and platforms including open-source ones have been being developed, for instance, NoSQL Databases, BigQuery, MapReduce, Hadoop, HiveQL, Spark, to mention but a few [2]–[4]. Some projects have also been launched by governments of countries such as USA and Japan to catch Big Data opportunities [1].

II. Background/Motivation

/10 Background is important to understand the topic in depth while motivation presents the importance

of the topic statement of objectives; two themes identified for the report should be clearly stated. One page is sufficient for this section.

A. ~~*in depth understanding of Big Data*~~

The main characteristics of Big Data are described as three Vs, namely Volume, Velocity and Variety [1], [3]. First of all, the large volume of data is an essential difference between Big Data and traditional data [3]. Second, the velocity at which the data are being generated implies that the processing and analysis of datasets should be carried out at a comparable rate to the data production [3]. Third, Big Data are produced in various format including both text and multimedia from various data sources, resulting in high heterogeneity and diversity [1], [5].

According to a well-accepted system engineering methodology in industry, the Big Data value chain is decomposed into four consecutive stages [3]:

- **Data generation** refers to the processes that data are generated from various sources.
- **Data acquisition** focuses on the obtaining and collection of data.
- **Data storage** concerns the persistent data storage and effective data management.
- **Data analytics** is the stage concerning the extraction of value from data by exploring, transforming, modelling and visualising data with analytical tools.

B. ~~*motivation & importance of Big Data*~~

C. ~~*challenges of opportunities of Big Data*~~

The mismatch between the requirements of Big Data and existing data management hardware and software platforms raises many challenges to both industry and research community. Many researches and practices, especially relating to the scalability, have been conducted among the four phases, including:

- Network architectures and protocols with high throughput, low latency and optimal energy consumption for large-scale data transmission [3]
- Scalable data cleaning, aggregation and duplication removal method for huge dataset at reasonable speed but still with acceptable accuracy. It is essential for big data quality and reliability [1], [3]
- Infrastructures, file systems and database technologies for distributed and scalable data storage. More specific issues include data partitioning and replication scheme, scalable data indexing and query, CAP option (consistency, availability and partition tolerance), concurrency control mechanism, parallel and distributed programming model [3], [6].
- Scalable machine learning on large dataset, including deep learning on large dataset, online (or stream) learning, parallel reinforcement learning, computation framework for machine learning, and so on [1], [6]
- Real-time or near real-time analysis of large increasing volume of data [1]
- Imbalanced Big Data analysis [1]

D. ~~*themes to dive in*~~

III. Related work/literature review

/15 Related work should comprise the review of current state of knowledge relevant to the topic. Comparison and contrasting between different authors/approaches should be a clear. Page length is 1 to 2.

reviews of surveys on Big Data
reviews of publications on the themes

IV. Discussion/Opinion

/20 This is an important section in which the student will criticise the existing work and will present his/her own opinion about how to improve it further? This section should reflect some research insight developed by the student.

opinion on theme 1
opinion on theme 2
...

V. Conclusion

/10 In this section student will draw conclusions on the given topic. In other words, it is a brief summary about work presented in the research report.

VI. Future issues

/10 This section should discuss at least two future issues on the topic.

VII. References

/10 All references should be of peer-reviewed journal and conferences. They must be clickable in the document. Each report should include at least 6 peer-reviewed references.

References

[1] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "Big data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4, pp. 431 –448, 2018, ISSN: 1319-1578. DOI: https://doi.org/10.1016/j.jksuci.2017.06.001.

[2] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods," *Journal of Business Research*, vol. 70, pp. 263 –286, 2017, ISSN: 0148-2963. DOI: https://doi.org/10.1016/j.jbusres.2016.08.001.

[3] H. Hu, Y. Wen, T. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2014.2332453.

[4] T. Hewage, M. Halgamuge, A. Syed, and G. Ekici, "Review: Big data techniques of google, amazon, facebook and twitter," English, *Journal of Communications*, vol. 13, no. 2, pp. 94–100, Feb. 2018, ISSN: 1796-2021. DOI: 10.12720/jcm.13.2.94-100.

[5] S. Pouyanfar, Y. Yang, S.-C. Chen, M.-L. Shyu, and S. S. Iyengar, "Multimedia big data analytics: A survey," *ACM Comput. Surv.*, vol. 51, no. 1, Jan. 2018, ISSN: 0360-0300. DOI: 10.1145/3150226.

[6] P. Gupta, A. Sharma, and R. Jindal, "Scalable machine-learning algorithms for big data analytics: A comprehensive review," *WIREs Data Mining and Knowledge Discovery*, vol. 6, no. 6, pp. 194–214, 2016. DOI: 10.1002/widm.1194. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1194.