Draft chapter

Literature Review

Fred Spoons

6 July 2020

Chapter 1

Literature Review

1.1 Introduction

The literature review design of this PHD research constitutes of 3 parts, 2 systematic literature reviews (SLR) on the topics ‘Big data reference architectures’ and ‘E-commerce systems” and one generic literature review on the topic ‘Big data’.

Systematic literature reviews take shape by embarking on an extensive search for topic-related articles within the years 2010-2020. Most literature chosen for the purposes of this research are within the years 2016-2020 as they provided with recent, and more relevant information. Albeit, some old studies dating back to 2010, helped clarifying some basic matters that existed and how they correlated to big data. world most renowned online libraries for quality research have been selected such as IEEE, MIS Quarterly, Science Direct, Elsevier, Springer, ACM, AISeL and Emerald insight.

Every library provided with a vast sea of research and inordinate amount of information to absorb. Arguably, different publications provided with different sort of mental framework and so did the authors. For instance, it’s been found that many high-quality information system researches are published in MIS quarterly, whereas Elsevir and SpringerLink provided with quality big-data literature.

A combination of long-tail and short-tail keywords are chosen to target literature that are related to the current state of art. Keywords chosen for ‘Big data reference architectures’ SLR are ‘big data reference architectures’, and ‘reference architectures. Keywords chosen for ‘E-commerce systems’ are ‘e-commerce system architectures’, ‘smart e-commerce systems’, and ‘e-commerce and big data’. Each systematic literature review is conducted in a span of three weeks.

In what follows, first the generic big data literature review will be conducted, second ‘big data reference architecture’ SLR and finally ‘E-commerce systems’ SLR will take place.

1.2 State of the art

We’ve come a long way with technology, and specifically software development. In fact, the rapid advancements left many spaced out. From the emergence of the first computer Eniac in 1946 to 8-core 5.0GHz processing core speed in 2019; From document-oriented waterfalls to agile two-weeks sprints; from punch cards to fancy transpilers and dynamic programming languages. Computers were first perceived as calculation engines and has been used to focus entirely on algorithms and mathematics. It was during the mid-1950, that it became commercially available and businessmen start to pick it up to produce value for business. Along the lines, once people started using computers for real-life purposes, many leftover data has been produced, as these data increased, people started realizing the value of it and began to store it (Grad & Bergin, 2009).

That’s where the industry came up with a concept of a Database Management System (DBMS), and humanity began to store data for various purposes. In 1968, as a result of a NATO-sponsored conference, the term software engineering emerged, referring to a highly systematic approach to software development and maintenance (Wirth, 2008).

Since the beginning of 1968, the advancement began on the areas of tools generation, testing, automation and systematizing. During the same years, in 1960s the history of computer hardware started by conversion of vacuum tubes to solid-states. Todays, the word 'bug' is quite a common phrase among engineers and programmers to refer to a fault, failure or a flaw. We ow this word to a literal moth that were caught inside a tube before the transition to solid states. It is hardly conceivable that we've progressed from absolutely no understanding for data to devices that can produce zettabytes of them in a span of 60 years. Along this track, software engineering has passed several major phases. Recent polyglot approaches with nascent lambda functions, functional paradigms and micro-services have come to take the industry by storm. This is the only time in human history, where the computing resources and the necessary data is available to harness the hidden patterns behind every momentum or dynamism. Being so focused on development of more maintainable and scalable software, and microchips and hardware's and devices that can perform faster and last longer, we have lost the track of the output of all these entities and peripherals, and that's the void that current industry is facing. Abundance of computer power, the emergence of open source community, and the ubiquity of internet has brought us with a new material to harness. A material, that is complex and random in nature.

It was not until 2005 that the term big data has been coined (Long, 2015), and Web 2.0 emerged which referred to a large set of data that is impossible to process with the traditional data management systems. Within the same year, Yahoo created Hadoop, Google came up with MapReduce. In 2009, the Indian government took a revolutionary step and decided to take an iris scan of its 1.2 billion inhabitants. In 2011 McKinesy published the title "Big Data: the next frontier of innovation" and startups and companies started investing heavily in this field. The big data revolution is ahead of us, and yet there is a big chasm both in practice and academia (McKinsey et al., 2011).

[1]

1.3 Big data

1.3.1 What is big data?

To define big data for the course of this PHD thesis, we will first look at available definitions in academia.

Kaisler, Armour, Espinosa and Money define big data as “the amount of data which is beyond technology’s capability to store, manage and process efficiently.Srivastava referred to big data as “the use of large data sets to handle the collection or reporting of data that serves business or other recipients in decision making”.

Sagiroglu and Sinanc define big data as “a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results”. Inspired by these definitions, we define big data as “an endeavor to harness the patterns behind vast amount of data for the purposes of improvement, control, and prediction of business matters”.

1.4 The Hype of Emerging Technologies

The term big data, was initially coined to refer to the gradual growth and availability of data (Lycett, 2013).

The ubiquity of digital devices and capability of users to produce different forms of data, have consolidated the interconnected links among suppliers, customers, affiliates, partners, and stakeholders (Bughin, 2016). With recent emergence of 5G technology

[1] A transpiler is a sort of a compiler that translates source codes from one language to another, or another version of the same language. For example Babel (a Javascript Library) transpiles the latest syntax of Javascript (ES6) into older version of it (ES5), thus all the browsers can support the system.

and its launch in the UK, we are experiencing a fundamental network shift that is unprecedented in human history (Ahmad et al., 2020)

Opposed to general belief of 5G being only faster than its elder brother 4G, 5G has come to offer bi-directional large bandwidth shaping, large broadcasting of data in gigabits which supports wearable devices with AI capabilities, pervasive networks providing ubiquitous computing (the user can be seamlessly connected to several wireless access technologies), traffic statistics, IPV6 utilization and finally 25Mbps of connectivity speed (Gohil, Modi & Patel, 2013).

In a world where we have the average processing power of 1.5 GHz on smart phones and up to 8 GHz on desktops running on network infrastructures that will support up to 25Mbps of transmission per second, data becomes the new oil, the atom, the dot that lays the foundation of the nexus (Rad & Ataei, 2017). It is astonishing to witness data being produced by netizens in every second .According to live internet statistics website , there are 4 billion internet users currently active, that produce 8,522 Tweets, 920 Instagram photos, 1,540 Tumbler posts, 3,868 Skype calls, 74,993 Google searches, 79,099 Youtube videos, 2,806,143 emails, and 73,693 GB of internet traffic per second (Stats, 2017). That implies, if it has taken 3 seconds to read the preceding paragraph, in the interim, 221,79 GB of traffic has been produced. Howbeit, how useful are these data? And how far have we gone with harnessing its power?

1.5 The Value of Big Data

The value of big data is no longer under the hood. In fact, the concept has been repeatedly discussed in various reports, statistics, researches and conferences (Chen, Chiang & Storey, 2012). The outburst is driven by the colossal investment of companies such as Google, Facebook, Netflix and Amazon (Rada, Ataeib, Khakbizc & Akbarzadehd, 2017).

A study of Netflix Prize recommender system provided details on employment of big data in order to induce better, more accurate results (Amatriain, 2013). The research has explicitly stated the notion of using various pools of data to further optimize recommendations. Data produced by queries, ratings, queues, search terms, and metadata alongside impression, social, external, demographic, location, language, and finally temporal data has been taken in use for predictive models (Amatriain, 2013). Using big data enforced recommendation systems, the company has managed to increase TV series consumption by the factor of four (Amatriain, 2013).

The Taiwanese government leveraged its national health insurance database and merged it with custom and immigration datasets to forge a big data initiative (C. J. Wang, Ng & Brook, 2020). This initiative resulted in improved case identification by generating real-time alerts during clinical visits. These alerts have been created by the analysis of clinical symptoms, travel history, and other data that could be found. Proactively seeking out patients that may be infected by COVID-19 was one of the reasons that Taiwanese government managed to handle the epidemic effectively.

Shell uses big data to reduce costs energy resources exploration (Marr, 2016). The company uploads data to analytics system and compare it with data from drilling sites around the world. The closer the results match where abundant resources have been found, the better decision will be made. Before big data, company had huge problems to identify energy resources. Waves of energies traveled through the earth's crust registered differently on sensors, depending on whether they are travelling through gaseous material, liquids, or solid rocks. Formerly, company employed the traditional hit and miss approach to confirm the findings of the initial survey which was expensive and time-consuming.

Along the lines, Rolls Royce harness the power of big data by capturing internal data from sensors fitter on the company's aircraft products. The data is received through a wireless transmission medium and contains multitudes of performance reports. These reports shed lights on various key phases such as take-off, engine power climax, steady state (climb and cruise), dynamisms, and maintenance (Marr, 2016). The company uses the data to detect degradation, to induce diagnosis and prognosis, and to minimize the false-positive as well.

1.6 Datocracy

The availability of data at an unprecedented frequency and the hidden patterns behind this nexus of interconnections has resulted in a new world, a datocratic world. Before getting further, is it essential to grasp the meaning of the new term ‘Datocracy’ proposed to correctly address the lingual needs for this research. To clarify the meaning of the word, it is helpful to understand the etymology behind the common term “Democracy”. Deomcracy comes from the combination of two ancient Greek words namely “demos” meaning ‘people’, and the post fix “-Kratia” meaning ‘to rule’. By the same line the combination of the Greek word “Datum” and the post fix “-Kratia” generates the word Datocracy, meaning “data to rule”.

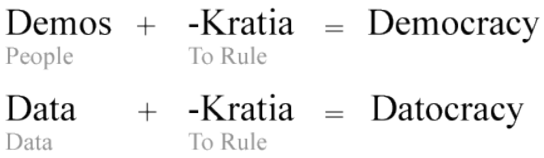

Figure 1.1: Datocracy

1.7 Ubiquity

Recent technology shifts and the computing power that each person carries along, has brought along a new business material, a datocracy. In a conference held in Abu Dhabi in 2013, Joseph S. Nye, a former US assistance secretary of defense and a university professor at Harvard, proposed the idea of future governance in the age of information (Nye, 2013).

He proposed the scenario in which the central government will use big data to fortify control. On the other hand, there is an estimation of 7121 publications on the fields big data regarding different dimensions, such as mathematical techniques, decision-making techniques, data characteristics, technical challenges and adoption failures (H. Wang, Xu, Fujita & Liu, 2016).

Paying clear attention to recent social, commercial and industrial trends will yield the evidence of big data ubiquity. In the domain of social network, there has been study for understanding temporal patterns of happiness by using a data set of 46 billion words contained in nearly 4.6 billion expression by 63 million unique users posted over a 33 months span (Dodds, Harris, Kloumann, Bliss & Danforth, 2011).

Furthermore, in another research, big data analytics and semantic network analysis were utilized to examine the largest data set collected on Twitter during 2012 U.S presidential election (Guo & Vargo, 2015). The study concluded that the news media could determine the public's identification of a certain candidate.

Other academicians have used big data to develop a novel distributed community structure mining framework. The framework makes use of local information data alongside MapReduce, and well-known algorithms such as FastGN, and Radetal to address scalability, velocity, and accuracy (Jin et al., 2015). On a bigger, more social-oriented studies, there has been researches regarding the overall well-being Turkish citizens by adopting a sentiment analysis model (Durahim & Coşkun, 2015).

Along the lines, the very sentiment analysis model has been taken by other researchers to discover general knowledge from social media (Bohlouli, Dalter, Dornhöfer, Zenkert & Fathi, 2015), and to evaluate and infer enhanced marketing advantage and to shed lights on areas in which the business is leading and lagging to further improve customer-business relation (He, Wu, Yan, Akula & Shen, 2015).

Similar researches have been conducted by analyzing suspended spam accounts on Twitter in terms of the profile's properties and interactions. These researchers were aimed to point out spammers and malicious users by using big data (Almaatouq et al., 2016). Chainey, Tompson and Uhlig have conducted a research on hotspot mapping and its usage to identify spatial patterns of crime. The study concluded that by utilizing a data from the past, hotspot mappings can identify where crimes most densely occur. From there on, there has been the proposition of target enforcement and prevention resources in the crime areas for mitigating crimes.

By the same token, (Li, Yen, Lu & Wang, 2012) used a large dataset from the bank of Taiwan and developed a big data system to identify signs and patterns of fraudulent accounts. They've developed a detection system by applying the Bayesian Classification and Association Rule. Along the lines, there has been other researches to predict negative behaviors spreading dynamics (Liao, Squicciarini & Griffin, 2015), emotional response detection by browsing Facebook (Lin & Utz, 2015), as well as identifying the impacts of national security by using the US intelligent community datasets (Crampton, 2015).

A wander into different areas provides with interesting ideas about how far the progress has been with the adoption of big data and proves a truly datocratic world. One good example is a comparative study conducted to document how big data can help with multifaceted aspects of international accreditations for two universities, namely Plekhanov Russian University of Economic and HAN University of Applied Science (Arnhem Business School) (Popescu, Iskandaryan & Weber, 2019).

In addition, (Zhang, Liu & Feng, 2019) conducted a research on the application of big data for tours and creative agencies. The objective of the study was to extract behavioral data and to form strategic objects that can be later applied for business benefit.

As witnessed hereinabove, there are abundant number of researches on the application of big data in various industries. Table 1 portrays an overview of the aforementioned studies and even further.

Contibution	Research Focus	Test 90

References

Ahmad, W. S. H. M. W., Radzi, N. A. M., Samidi, F., Ismail, A., Abdullah, F., Jamaludin, M. Z. & Zakaria, M. (2020). 5g technology: Towards dynamic spectrum sharing using cognitive radio networks. *IEEE Access*, *8*, 14460–14488. doi: 10.1049/iet-com.2018.6129

Almaatouq, A., Shmueli, E., Nouh, M., Alabdulkareem, A., Singh, V. K., Alsaleh, M., ... Alfaris, A. (2016). If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts [Journal Article]. *International Journal of Information Security*, *15*(5), 475-491. doi: 10.1007/s10207-016-0321-5

Amatriain, X. (2013). Beyond data: from user information to business value through personalized recommendations and consumer science [Conference Proceedings]. In (p. 2201-2208). ACM. doi: 10.1145/2505515.2514691

Bohlouli, M., Dalter, J., Dornhöfer, M., Zenkert, J. & Fathi, M. (2015). Knowledge discovery from social media using big data-provided sentiment analysis (somabit) [Journal Article]. *Journal of Information Science*, *41*(6), 779-798. doi: 10.1177/0165551515602846

Bughin, J. (2016). Big data, big bang? [Journal Article]. *Journal of Big Data*, *3*(1), 2. doi: 10.1186/s40537-015-0014-3

Chainey, S., Tompson, L. & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime [Journal Article]. *Security journal*, *21*(1-2), 4-28. doi: 10.1057/palgrave.sj.8350066

Chen, H., Chiang, R. H. & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact [Journal Article]. *MIS quarterly*, *36*(4), 1165. doi: 10.2307/41703503

Crampton, J. W. (2015). Collect it all: National security, big data and governance [Journal Article]. *GeoJournal*, *80*(4), 519-531. doi: 10.1177/2053951716661366

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: hedonometrics and twitter [Journal Article]. *PLoS One*, *6*(12), e26752. Retrieved from `https://www.ncbi.nlm.nih.gov/pubmed/22163266` doi: 10.1371/journal.pone.0026752

Durahim, A. O. & Coşkun, M. (2015). iamhappybecause: Gross national happiness through twitter analysis and big data [Journal Article]. *Technological Forecasting and Social Change*, *99*, 92-105. doi: 10.1016/j.techfore.2015.06.035

Gohil, A., Modi, H. & Patel, S. K. (2013). 5g technology of mobile communication: A survey [Conference Proceedings]. In *2013 international conference on intelligent systems and signal processing (issp)* (p. 288-292). IEEE. doi: 10.1109/issp.2013.6526920

Grad, B. & Bergin, T. J. (2009). Guest editors' introduction: History of database management systems [Journal Article]. *IEEE Annals of the History of Computing*, *31*(4), 3-5. doi: 10.1109/mahc.2009.99

Guo, L. & Vargo, C. (2015). The power of message networks: A big-data analysis of the network agenda setting model and issue ownership [Journal Article]. *Mass Communication and Society*, *18*(5), 557-576. doi: 10.1080/15205436.2015.1045300

He, W., Wu, H., Yan, G., Akula, V. & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks [Journal Article]. *Information and Management*, *52*(7), 801-812. doi: 10.1016/j.im.2015.04.006

Jin, S., Lin, W., Yin, H., Yang, S., Li, A. & Deng, B. (2015). Community structure mining in big data social media networks with mapreduce [Journal Article]. *Cluster computing*, *18*(3), 999-1010. doi: 10.1007/s10586-015-0452-x

Kaisler, S., Armour, F., Espinosa, J. A. & Money, W. (2013). Big data: Issues and challenges moving forward [Conference Proceedings]. In *2013 46th hawaii international conference on system sciences* (p. 995-1004). IEEE. doi: 10.1109/hicss.2013.645

Li, S.-H., Yen, D. C., Lu, W.-H. & Wang, C. (2012). Identifying the signs of fraudulent accounts using data mining techniques [Journal Article]. *Computers in Human Behavior*, *28*(3), 1002-1013. doi: 10.1016/j.chb.2012.01.002

Liao, C., Squicciarini, A. & Griffin, C. (2015). Epidemic behavior of negative users in online social sites [Conference Proceedings]. In *Proceedings of the 5th acm conference on data and application security and privacy* (p. 143-145). ACM. doi: 10.1145/2699026.2699129

Lin, R. & Utz, S. (2015). The emotional responses of browsing facebook: Happiness, envy, and the role of tie strength [Journal Article]. *Comput Human Behav*, *52*, 29-38. Retrieved from `https://www.ncbi.nlm.nih.gov/pubmed/26877584` doi: 10.1016/j.chb.2015.04.064

Long, C. (2015). Data science and big data analytics: Discovering, analyzing, visualizing and presenting data [Journal Article]. *Indianapolis, Indiana*. doi: 10.1109/bgdds.2018.8626811

Lycett, M. (2013). *'datafication': making sense of (big) data in a complex world* (Vol. 22). Taylor and Francis. doi: 10.1057/ejis.2013.10

Marr, B. (2016). *Big data in practice: how 45 successful companies used big data analytics to deliver extraordinary results* [Book]. John Wiley and Sons. doi: 10.1109/bigdata.2018.8622333

McKinsey, G. et al. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, 158-184. doi: 10.7591/9781501734328-007

Nye, J. S. (2013). Governance in the information age [Journal Article]. *Governance*, *113*, 19-22. doi: 10.1525/curh.2014.113.759.19

Popescu, F., Iskandaryan, R. & Weber, T. (2019). Big data and international accreditations in higher education: A dutch–russian case study [Journal Article]. doi: 10.5220/0007572102070214

Rad, B. B. & Ataei, P. (2017). The big data ecosystem and its environs [Journal Article]. *International Journal of Computer Science and Network Security (IJCSNS)*, *17*(3), 38. doi: 10.1007/978-3-319-94301-5_16

Rada, B. B., Ataeib, P., Khakbizc, Y. & Akbarzadehd, N. (2017). The hype of emerging technologies: Big data as a service [Journal Article]. doi: 10.1109/bigdata.2017.8258302

Sagiroglu, S. & Sinanc, D. (2013). Big data: A review [Conference Proceedings]. In *2013 international conference on collaboration technologies and systems (cts)* (p. 42-47). IEEE. doi: 10.1109/cts.2013.6567202

Srivastava, R. (2018). Big data: Issues and challenges [Journal Article]. *International Journal Of Scientific And Innovative Research*(6), 1. doi: 10.1007/978-981-13-8759-3_10

Stats, I. L. (2017). Internet users [Journal Article]. , *10*, 7. doi: 10.3390/fi10010007

Wang, C. J., Ng, C. Y. & Brook, R. H. (2020). Response to covid-19 in taiwan: big data analytics, new technology, and proactive testing. *Jama*, *323*(14), 1341–1342.

Wang, H., Xu, Z., Fujita, H. & Liu, S. (2016). Towards felicitous decision making: An overview on challenges and trends of big data [Journal Article]. *Information Sciences*, *367*, 747-765. doi: 10.1016/j.ins.2016.07.007

Wirth, N. (2008). A brief history of software engineering [Journal Article]. *IEEE Annals of the History of Computing*, *30*(3), 32-39. doi: 10.1109/mahc.2008.33

Zhang, M., Liu, J. & Feng, L. (2019). The application of big data technology in creative travel [Conference Proceedings]. In *2019 international conference on intelligent transportation, big data and smart city (icitbs)* (p. 317-319). IEEE. doi: 10.1109/icitbs.2019.00083