Draft chapter

Literature Review

Fred Spoons

9 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.R. 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 (H. 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 (R. 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, (M. 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.

Table 1.1: my caption

Contibution	Multi-column	
(Luo, Wu, Gopukumar & Zhao, 2016)	Application of big data in health care	*Healthcare*
(Murdoch & Detsky, 2013)	Adoption of big data in health care	
(Y. Zhang, Qiu, Tsai, Hassan & Alamri, 2015)	Application of big data in cloud in healthcare cyber-physical system	
(Y. Zhang et al., 2015)	Application of big data in cloud in healthcare cyber-physical system	
(Bates, Saria, Ohno-Machado, Shah & Escobar, 2014)	Exerting big data analytics to identify and manage high-risk and high-cost patients	
(K. Lin, Xia, Wang, Tian & Song, 2016)	Designing systems for emotion-aware healthcare using big data	
(K. Lin et al., 2016)	Designing systems for emotion-aware healthcare using big data	

(Srinivasan & Arunasalam, 2013)	Analyzing health insurance claims to detect frauds, errors, waste, abuse	
(Mehta & Pandit, 2018)	Analysis of application of big data in healthcare	
(Firouzi et al., 2018)	Amalgamation of Internet of Things (IoT) and Big Data for a smarter healthcare	
(Firouzi et al., 2018)	Analysis of application of big data in healthcare	
(M. Chen et al., 2018)	Analysis of application of big data in healthcare	
(Asur & Huberman, 2010)	Predictive analytics using social networks	*Social*
(Dodds et al., 2011)	Revealing temporal patterns of happiness	
(Guo & Vargo, 2015)	Utilizing big data to examine message networks such as Twitter and traditional news media	
(Jin et al., 2015)	Developing a novel distributed community structure mining framework	
(Durahim & Coşkun, 2015)	Analysis of overall happiness in Turkey through twitter analysis and big data	
(Bohlouli et al., 2015)	Knowledge discovery from big data	
(Chainey et al., 2008)	Predicting of spatial patterns of crime using big data	*Crime and Fraud*
(Li et al., 2012)	Using bank data to identify patterns of fraud	

(Liao et al., 2015)	Predicting abusiveness in online commentaries and preventing them	
(Tran et al., 2018)	Data driven approaches for credit card fraud detection	
(Sigala, 2019)	A book on big data and how it can bring innovation to tourism business	*Tourism*
(M. Zhang et al., 2019)	Analyzing the application of big data in tourism	
(Dezfouli, Shahraki & Zamani, 2018)	Developing a tour model using big data	
(Qin et al., 2019)	Utilization of big data with Call Detail Record (CDR) data and mobile real-time location data to monitor the tourist flow and travel behavior	

1.8 Business Benefits and Challenges

Big data and the benefits it brought along has resulted in new approaches of decision-making, which can be called data grounded decision-making (Comuzzi & Patel, 2016) or a datocratice decision-making.These new organizational adaptations target compelling interventions in both internal and external aspects, especially in areas that has formerly been dominated by gut and intuition rather than by data and exactitude (Wamba et al., 2017).

As the data-oriented strategies, philosophies, artifacts, technologies, tools, and

methodologies emerge, they will revolutionize long-established ideologies toward the nature of decision-making, business process management, and predictive models (Popovič, Hackney, Tassabehji & Castelli, 2018).

There is and will be a colossal attention toward the undeniable benefits of big data to lay out customer's behavioral patterns, experimental orientation, organizational functions, business acumen, predictive decision-making, and a far more informative business and marketing plans comparing to traditional resolutions (van den Driest, Sthanunathan & Weed, 2016).

But as with any other great change, there comes the hurdles. Aside from the technical difficulties and talent shortage, turning into a data-based management, or data grounded decision-making system, would require a clear vision to align business objectives with insights, and there arise a big possibility for error (Ranjan, 2019).

To invest in such hype as big data and to extract insights sounds fundamentally important to any competitive business. Nevertheless, what is even more important is to ascertain that extracted insights are in parallel with business objectives. From planning to brainstorming, mentoring, strategizing, networking, training, and even communicating, data can be utilized to create a coherent view that integrates and synthesizes the main aspects of the field, so executives can put into perspective the new directions that are more likely to result in success.

For majority of businesses, the business plan is the key ingredient of a good strategy and execution. The business plan serves as a guiding light to realize 'where to play' and 'how to win' (van den Driest et al., 2016).

Within the business plan, resources are allocated, financials are calculated, and the roadmap is determined against goals. This is where, insights can contribute a lot to drive strategy, to layout activities, and to schedule business phases (Y.-S. Chen, 2018).

Despite the hype of big data, in majority of cases, these planning are based on executive's mature judgment that is inevitably affected by the selective past experiences,

mental patterns, emotions, motives and ideologies. Whereas, if there are insights injected into this game-changing phase, the activities and business functions can be precisely aligned with goals and objectives. That is why, overperforming companies inject insights to decision-making in all the major phases, and cultivate a data-driven culture (van den Driest et al., 2016).

As well as that, there has been considerably fewer researches on the areas of big data architecture, reference architectures for data-driven systems. Majority of studies to date, have concentrated on big data analytics capability, pattern recognition, and big data challenges, whereas other complementary areas such as insight orchestration, required socio-technological developments, big data reference architectures and suitable data-oriented artifacts are neglected (Mikalef, Pappas, Krogstie & Giannakos, 2018).

What is repeatedly observed within the researches, is that the notion of 'insight' and the ways in which it provides a means to desired ends, is missing (van den Driest et al., 2016; H.-M. Chen, Kazman & Matthes, 2016; H.-M. Chen, Kazman, Haziyev & Hrytsay, 2015). Manifold aspects within the particularities of data, and its interdependencies to induce value systematically, requires an established architecture used in the IT-business value domain.

The lack of sufficient research on the areas of big data system architectures and system development, impedes the growth of the conceivable potential, and leaves practitioners in uncharted territories (Mikalef et al., 2018). Unquestionably, with any great stream of change, comes the ambiguity, and today's challenges of becoming a data driven organization can be immense and requires a transition that needs to engage with today's market (McAfee & Brynjolfsson, 2012).

1.9 Is Big Data for Everyone?

The concept of big data has not yet matured enough to have its grounded unanimous understanding, and inevitably, ambiguity and challenges are the constants (Akhtar, Frynas, Mellahi & Ullah, 2019).

On one hand big data represents an opportunity for an organization to grow and expand rapidly by taking data into the decision making process (H.-M. Chen, Kazman & Matthes, 2016). On the other hand, big data, if not fully understood, can be considered "a hammer looking for nail". Under the conditions that organization is not capable of innovation, is conservative, or simply lacks the desire for innovation, complexity tolerance will be decreased, success rate becomes negligible, and much of a resource will be wasted. Big data can be a perceived as an educated and adventurous approach towards decision making, and such an approach isn't suitable for every organization.

An informed and proactive decision-making process for the future is plausible and desirable in the first sight, but as the projects progress, hidden problems emerge and potentially increase complexity. In such states, most traditional approaches of technology orchestration and most common architectures fail. In spite of big data potential, majority of big data's internal mechanisms are not clear from the start, and there is a void of knowledge among both practitioners and researchers. Big data can create domains of knowledge that is expandable to new areas of understanding. To embark on this process and to benefit from it, there is a high demand of dynamism and plasticity.

1.10 An Architecture-Centric Approach

As data flows grow and emerge, our traditional system development become less and less effective. Traditional system development is mainly focused on relational models

that revolve around ANSI standard 3 tier DBMS architecture, which explicitly states data/program independence for development of data systems (Elmasri, 2017).

The process of data system development consists of 7 major phases, two first phases are about requirement analysis and conceptual design. Third phase is the selection of the DBMS. Fourth and fifth phases are logical design, physical design. And three last phases are prototyping/testing, implementation and performance evaluation (Elmasri, 2017).

In a default setup, an agile methodology is applied where progress occurs through Kanban and usually two weeks sprints. The industry has a clear idea of the architectures, references and patterns to refer to when needed. Majority of software's today run on a multitier or multilayered architecture, where presentation, application, and data management functions are separated (H.-M. Chen, Kazman & Haziyev, 2016).

But that's definitely not the case of big data. The relational data models despite being dominant since the 80's with prominent implementations such as Microsoft SQL servers, MySQL, and Oracle databases, fail to effectively address big data characteristics (Moniruzzaman & Hossain, 2013).

These short comings occur in different domains. The gradual growth of data would definitely require a horizontal scaling and a distributed architecture. Here we clearly distinguish the horizontal scaling by delegating tasks to other nodes of the network, that will exert the necessary computation and return the results once done. and vertical scaling, by simply increasing the computer power of a single node or workstation to handle the necessary computation.

. In the world of big data, a massive-parallel data processing is the focal point of scaling (Moniruzzaman & Hossain, 2013). Furthermore, the variety of unstructured, semi-structured and pseudo-structured data, makes it impossible to consider any RDBMS-centered approach for a successful big data project. IT leaders such as Facebook, Amazon, and Google had reached the tipping point where the conventional

RDBMS solutions could not cope and had to go through storms to innovate new technologies. The rain after the storm were the great technologies that are becoming more and more common among practitioners these days.

Some of these technologies are;

- **Presto**: a distributed SQL Query Engine for Big Data
- **Relay**: a JavaScript framework for building data-driven applications
- **GraphQL**: a data query language and runtime
- **Airflow**: a platform to programmatically author, schedule, and monitor data pipelines.
- **AresDB**: a GPU-powered real-time analytics storage and query engine
- **Hadoop**: open-source software for reliable, scalable, distributed computing

Thus, the finalization could be that the traditional RDBMS can not address big data characteristics for three major reasons;

1. **Data size**: Traditional RDBMS solutions, being inherently centralized and vertically scaled, would fail to handle petabytes of data
2. **Data variety**: Traditional RDBMS solutions have been developed and implemented to support only and solely the structured data. Thus, they fail to successfully support semi-structured and un-structed data.
3. **Data velocity**: Tradition RDMBS solutions are not suitable for high velocity data. Even if they're implemented to handle high velocity data, they would not be efficient.

There has been mentions of different characteristics of big data, especially with regards to volume, velocity, and variety. In next section, these characteristics will be analyzed and elaborated.

1.10.1 Volume

Simply, the volume and multitude of data can bring about major technical challenges. An architecture needs to be elastic enough to address various volumes at different rates. Storing and computing large volume of data with attention to efficiency is a complex process that has not been fully addressed. In this domain, there is a need for a scalable configurable architecture for a massive distributed and parallel processing (H.-M. Chen, Kazman & Haziyev, 2016).

1.10.2 Variety

Up until the year 2000, there were lack of support for unstructured data. In 2000, technologies such as BLOB (binary large object) has been developed to address unstructured data (Bahrami & Singhal, 2015). BLOB solved a lot of problems of engineers and developers as it allowed the storage of multimedia data (video, image, sound). They were also used to store programs and chunks of codes.

1.10.3 NoSQL

With rise of newer and newer types of data, traditional SQL databases became less effective in storing and computing. Big data precipitated the emergence of NoSQL databases. NoSQL emerged to handle large volumes of unstructured data. Whereas SQL databases run on rational models, NoSQL technologies are often schema less or non-tabular. This allows for an increased flexibility. In SQL world a unit of storage is a table that is constituent of columns and rows, whereas in NoSQL world, a unit of

storage can be a document, a key-value pair, a wide column or a graph. NoSQL database types are classified under 4 major categories; document-oriented databases (such as MongoDB), key-value store databases (such as Redis), column-oriented databases (such as MariaDB), and lastly graph databases (such as Neo4J) (Han, Haihong, Le & Du, 2011).

These technologies mainly differ on the domains of availability, consistency and partition tolerance. This is mainly based on the CAP theorem proposed by professor Eric Brewer in 2005, who believed that in a distributed setup system cannot meet the three district needs concurrently, but can at max meet two (Brewer, 2012).

1.10.4 Polyglot Persistence

Aside from NoSQL databases, Big data has also given rise to a polyglot persistence (K. Srivastava & Shekokar, 2016). The term polyglot is coined behind the idea that different datasets require different databases as to solve different problems (Sadalage & Fowler, 2013).

Thus, using a single database engine for all the requirements leads to inefficient implementation and technical bottlenecks. Polyglot persistence was coined in 2006 by Neal Ford, to express the ideology that persistence should be attained by utilizing different databases systems, programming languages and tools (Sadalage & Fowler, 2013). These nascent technologies have brought much engineering discipline that challenges traditional DB development.

1.10.5 Velocity

Data come at different rate and with different volume, this is what refers to velocity. Challenge in this area arises from the choices of real-time process and batch-processing. Processing of data to expedite the decision-making process quickly on one hand and

handling the variety of data and storing them for batch-processing on other hand brings along a great technical challenge. One of the recent endeavors to addressee such challenges is the famous Lambda Architecture (Marz & Warren, 2015). This proves that there is a need for an architectural design to address the big data application accordingly.

1.10.6 Veracity

Data is not always appreciated. In fact, a bad, dirty data is just a hunch on top of the server. This brings along challenges in the areas of cleansing, modeling, and governance (H.-M. Chen, Kazman & Matthes, 2016).

Moreover, many of the data fetched may be illegal which causes privacy and security issues. Veracity includes two main dimensions; namely, data trustworthiness that is defined by a number of variables such as the collection method, processing, medium, protocol, data origin and platform; and data consistency which can be asserted by statistical reliability (Demchenko, Grosso, De Laat & Membrey, 2013). Big data veracity is concerned with authenticity and usability of data.

This is where huge concepts like data cleaning, and data detection and metadata come into play. Different big data mechanisms are exerted in the data lifecycle, from the collection of data from trusted sources to, crunching, cleansing, protecting and storing it. There have been researchers who have devoted their career only to a single aspect of this lifecycle. Nevertheless, based on the studies conducted by Kaisler et al.; Demchenko et al.; Snijders, Matzat and Reips; Bughin; H.-M. Chen, Schütz, Kazman and Matthes following elements are defined and are necessary to ensure data veracity:

- Accountability of data sources
- Proven trustworthiness of the platform
- Trusted origin

- A clear identification of data lifecycle
- Data integration
- Timeliness and availablity

Traditional DB system developments were mostly used in a well-known context; thus, the validation rules could be easily made and exerted. However, big data veracity as elaborated hereinabove comes with a hard-to-define context for interpretation. This adds architectural complexity and highlights the importance of a clear predefined architecture.

1.10.7 Value

Lastly and most importantly, value encompasses all its four brothers as gleaning, crunching and extracting value from data, requires an integrated approach of storage and computing. Value is extraction of knowledge that depend on events or processes and interdependencies. These events and process can come in different forms such as stochastic, probabilistic, regular or random (Demchenko et al., 2013).

In this domain, there is always the challenge of bridging old architecture with the new ones. Architecture-centric approach can provide with sufficient abstraction to address such problems. Notwithstanding, architecture can be perceived as a fulcrum of technicalities and business needs (H.-M. Chen, Kazman & Haziyev, 2016).

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

Akhtar, P., Frynas, J. G., Mellahi, K. & Ullah, S. (2019). Big data-savvy teams' skills, big data-driven actions and business performance [Journal Article]. *British Journal of Management*, *30*(2), 252-271. doi: 10.1111/1467-8551.12333

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

Asur, S. & Huberman, B. A. (2010). Predicting the future with social media. In *2010 ieee/wic/acm international conference on web intelligence and intelligent agent technology* (Vol. 1, pp. 492–499). doi: 10.1109/wi-iat.2010.63

Bahrami, M. & Singhal, M. (2015). The role of cloud computing architecture in big data [Book Section]. In *Information granularity, big data, and computational intelligence* (p. 275-295). Springer. doi: 10.1201/9781315155678-8

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A. & Escobar, G. (2014). Big data in health care: using analytics to identify and manage high-risk and high-cost patients [Journal Article]. *Health Affairs*, *33*(7), 1123-1131. doi: 10.1377/hlthaff.2014.0041

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

Brewer, E. (2012). Cap twelve years later: how the [Journal Article]. *Computer*, *45*(2), 23-29. doi: 10.1109/mc.2012.37

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

Chen, H.-M., Kazman, R. & Haziyev, S. (2016). Agile big data analytics development: An architecture-centric approach [Conference Proceedings]. In *2016 49th hawaii international conference on system sciences (hicss)* (p. 5378-5387). IEEE. doi: 10.1109/hicss.2016.665

Chen, H.-M., Kazman, R., Haziyev, S. & Hrytsay, O. (2015). Big data system development: An embedded case study with a global outsourcing firm [Conference Proceedings]. In *Proceedings of the first international workshop on big data software engineering* (p. 44-50). IEEE Press. doi: 10.1109/bigdse.2015.15

Chen, H.-M., Kazman, R. & Matthes, F. (2016). Demystifying big data adoption: Beyond it fashion and relative advantage [Conference Proceedings]. In *Proceedings of pre-icis (international conference on information system) digit workshop.* doi: 10.1109/hicss.2016.631

Chen, H.-M., Schütz, R., Kazman, R. & Matthes, F. (2017). How lufthansa capitalized on big data for business model renovation [Journal Article]. *MIS Quarterly Executive*, *16*(1). doi: 10.24251/hicss.2017.713

Chen, M., Yang, J., Zhou, J., Hao, Y., Zhang, J. & Youn, C.-H. (2018). 5g-smart diabetes: Toward personalized diabetes diagnosis with healthcare big data clouds [Journal Article]. *IEEE Communications Magazine*, *56*(4), 16-23. doi: 10.1109/mcom.2018.1700788

Chen, Y.-S. (2018). E-business and big data strategy in franchising [Book Section]. In *Encyclopedia of information science and technology, fourth edition* (p. 2686-2696). IGI Global. doi: 10.4018/978-1-5225-2255-3.ch234

Comuzzi, M. & Patel, A. (2016). How organisations leverage big data: A maturity model [Journal Article]. *Industrial management and Data systems*, *116*(8), 1468-1492. doi: 10.1108/imds-12-2015-0495

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

Demchenko, Y., Grosso, P., De Laat, C. & Membrey, P. (2013). Addressing big data issues in scientific data infrastructure [Conference Proceedings]. In *2013 international conference on collaboration technologies and systems (cts)* (p. 48-55). IEEE. doi: 10.1109/cts.2013.6567203

Dezfouli, M. B., Shahraki, M. H. N. & Zamani, H. (2018). A novel tour planning model using big data [Conference Proceedings]. In *2018 international conference on artificial intelligence and data processing (idap)* (p. 1-6). IEEE. doi: 10.1109/idap.2018.8620933

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

Elmasri, R. (2017). *Fundamentals of database systems* [Book]. doi: 10.1007/978-1-4614-8265-9_80674

Firouzi, F., Rahmani, A. M., Mankodiya, K., Badaroglu, M., Merrett, G. V., Wong, P. & Farahani, B. (2018). *Internet-of-things and big data for smarter healthcare: From device to architecture, applications and analytics* (Vol. 78). Elsevier. doi: 10.1016/j.future.2017.09.016

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

Han, J., Haihong, E., Le, G. & Du, J. (2011). Survey on nosql database [Conference Proceedings]. In *2011 6th international conference on pervasive computing and applications* (p. 363-366). IEEE. doi: 10.1109/icpca.2011.6106531

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, K., Xia, F., Wang, W., Tian, D. & Song, J. (2016). System design for big data application in emotion-aware healthcare [Journal Article]. *IEEE Access*, *4*, 6901-6909. doi: 10.1109/access.2016.2616643

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

Luo, J., Wu, M., Gopukumar, D. & Zhao, Y. (2016). Big data application in biomedical research and health care: A literature review [Journal Article]. *Biomed Inform Insights*, *8*, 1-10. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/26843812 doi: 10.4137/BII.S31559

Lycett, M. (2013). *'datafication': making sense of (big) data in a complex world* (Vol. 22). Taylor 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

Marz, N. & Warren, J. (2015). *Big data: Principles and best practices of scalable real-time data systems* [Book]. New York; Manning Publications Co. doi: 10.1109/tcss.2020.2995497

McAfee, A. & Brynjolfsson, E. (2012). Big data: the management revolution [Journal Article]. *Harv Bus Rev*, *90*(10), 60-6, 68, 128. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23074865

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

Mehta, N. & Pandit, A. (2018). Concurrence of big data analytics and healthcare: A systematic review [Journal Article]. *Int J Med Inform*, *114*, 57-65. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/29673604 doi: 10.1016/j.ijmedinf.2018.03.013

Mikalef, P., Pappas, I. O., Krogstie, J. & Giannakos, M. (2018). Big data analytics capabilities: a systematic literature review and research agenda [Journal Article]. *Information Systems and e-Business Management*, *16*(3), 547-578. doi: 10.1109/educon.2018.8363273

Moniruzzaman, A. & Hossain, S. A. (2013). Nosql database: New era of databases for big data analytics-classification, characteristics and comparison [Journal Article]. *arXiv preprint arXiv:1307.0191*. doi: 10.1109/icrito.2015.7359207

Murdoch, T. B. & Detsky, A. S. (2013). The inevitable application of big data to health care [Journal Article]. *JAMA*, *309*(13), 1351-2. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23549579 doi: 10.1001/jama.2013.393

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

Popovič, A., Hackney, R., Tassabehji, R. & Castelli, M. (2018). The impact of big data analytics on firms' high value business performance [Journal Article]. *Information Systems Frontiers*, *20*(2), 209-222. doi: 10.1016/j.jbusres.2018.12.072

Qin, S., Man, J., Wang, X., Li, C., Dong, H. & Ge, X. (2019). Applying big data analytics to monitor tourist flow for the scenic area operation management [Journal Article]. *Discrete Dynamics in Nature and Society*, *2019*, 1-11. doi: 10.1155/2019/8239047

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

Ranjan, J. (2019). The 10 vs of big data framework in the context of 5 industry verticals [Journal Article]. *Productivity*, *59*(4), 324-342. doi: 10.32381/prod.2019.59.04.2

Sadalage, P. J. & Fowler, M. (2013). *Nosql distilled: a brief guide to the emerging world of polyglot persistence* [Book]. Pearson Education. doi: 10.1109/aiccsa .2015.7507130

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

Sigala, M. (2019). *Big data and innovation in tourism, travel, and hospitality: Managerial approaches, techniques, and applications* [Book]. ieeexplore. doi: 10.1007/978-981-13-6339-9_4

Snijders, C., Matzat, U. & Reips, U.-D. (2012). " big data": big gaps of knowledge in the field of internet science [Journal Article]. *International Journal of Internet Science*, *7*(1), 1-5. doi: 10.1037/13620-017

Srinivasan, U. & Arunasalam, B. (2013). Leveraging big data analytics to reduce healthcare costs [Journal Article]. *IT professional*, *15*(6), 21-28. doi: 10.1109/mitp.2013.55

Srivastava, K. & Shekokar, N. (2016). A polyglot persistence approach for e-commerce business model [Conference Proceedings]. In *2016 international conference on information science (icis)* (p. 7-11). IEEE. doi: 10.1109/infosci.2016.7845291

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

Tran, P. H., Tran, K. P., Huong, T. T., Heuchenne, C., HienTran, P. & Le, T. M. H. (2018). Real time data-driven approaches for credit card fraud detection [Conference Proceedings]. In *Proceedings of the 2018 international conference on e-business and applications* (p. 6-9). ACM. doi: 10.1145/3194188.3194196

van den Driest, F., Sthanunathan, S. & Weed, K. (2016). Building an insights engine [Journal Article]. *Harvard business review*, *94*(9), 15. doi: 10.2501/jar-2016-029

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J.-f., Dubey, R. & Childe, S. J.

(2017). Big data analytics and firm performance: Effects of dynamic capabilities [Journal Article]. *Journal of Business Research*, *70*, 356-365. doi: 10.1016/j.jbusres.2016.08.009

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

Zhang, Y., Qiu, M., Tsai, C.-W., Hassan, M. M. & Alamri, A. (2015). Health-cps: Healthcare cyber-physical system assisted by cloud and big data [Journal Article]. *IEEE Systems Journal*, *11*(1), 88-95. doi: 10.1109/jsyst.2015.2460747