2020

Big Data Reference Architectures, a systematic literature review

Pouya Ataei
Auckland University of Technology, pouya.ataei@aut.ac.nz

Alan T. Litchfield
Auckland University of Technology, Alan.litchfield@aut.ac.nz

Big Data Reference Architectures, a systematic literature review

Completed research paper

Pouya Ataei
School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
Email: pouya.ataei@aut.ac.nz

Alan Litchfield
Service and Cloud Computing Research Lab
Auckland University of Technology
Auckland, New Zealand
Email: alan.litchfield@aut.ac.nz

Abstract

Today, we live in a world that produces data at an unprecedented rate. The significant amount of data has raised lots of attention and many strive to harness the power of this new material. In the same direction, academics and practitioners have considered means through which they can incorporate data-driven functions and explore patterns that were otherwise unknown. This has led to a concept called Big Data. Big Data is a field that deals with data sets that are too large and complex for traditional approaches to handle. Technical matters are fundamentally critical, but what is even more necessary, is an architecture that supports the orchestration of Big Data systems; an image of the system providing with clear understanding of different elements and their interdependencies. Reference architectures aid in defining the body of system and its key components, relationships, behaviors, patterns and limitations. This study provides an in-depth review of Big Data Reference Architectures by applying a systematic literature review. The study demonstrates a synthesis of high-quality research to offer indications of new trends. The study contributes to the body of knowledge on the principles of Reference Architectures, the current state of Big Data Reference Architectures, and their limitations.

Keywords Big Data, Data analytics, Big Data for business, Data-intensive applications, Decision making

1 Introduction

This paper presents the findings from a Systematic Literature Review (SLR) of the use of Reference Architecture (RA) to address issues in Big Data (BD) solution design and development. The study has a particular focus on the use of BD in digital commercial activities.

Driven by the need to harness the potential hidden within the patterns buried deep in the enormous volumes of data produced daily, the term BD emerged as a concept that provides the means for realizing those opportunities (Rada et al. 2017). In modern world, users are the ceaseless generators of structured, semi-structured, and unstructured data that if analyzed, may reveal game-changing patterns (Erevelles et al. 2016). The growth in the application of analytics is witnessed by an extensive amount of attention toward data and predictive analytics from scholars and practitioners (Mikalef et al. 2017). The adoption of the concept of BD has precipitated a fundamental transformation within the industry and triggered technological innovation, business shifts, and other possibilities but also entails challenges. The application of BD in extracting insights and driving organizational advantages has introduced a new era that has yet to mature. We find that traditional approaches to data analytics fail to address the characteristics of BD and consequently, industry is in a state of change (Sivarajah et al. 2017).

As BD system development is a practice of technology orchestration, RAs can address some of the problems that emerge (Cloutier et al. 2010). RAs are abstract artefacts that allow for high-level contextualization of the system and its comprising components. Practitioners that build complex systems, software engineers, and system designers use RAs so that a collective understanding of system components, functionalities, data-flows and patterns that shape the overall qualities of system are made clear. Furthermore, the RA provides a foundation to adjust a system design to better meet business objectives (Cloutier et al. 2010; Kohler and Specht 2019). An RA provides predefined architectural patterns that address classes of problem or issue which enables an overall context of the system to be defined, its major component is identified, and its quality attributes assessed.

On that account, there is a need for more research in the area of BD RAs, and SLR can be a good academic effort to provide means by which current best evidence from literature can be combined, interpreted and explained. SLRs can play a momentous role in disseminating knowledge, developing new theories, supporting evidence-based practice and overall future research studied in the field (Paré et al. 2016).

2 Review methodology

The SLR method applied in this study follows the guidelines presented in Kitchenham et al. (2004) and Shamseer et al. (2015). We've utilized Kitchenham et al. (2004) framework because of its clear instructions on critically appraising evidence for validity, impact and applicability. Complementary to that, we've used guidelines provided by Shamseer et al. (2015) on Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). PRISMA provided means for increasing systematicity, transparency, and prevention of bias.

Of particular importance is the quality of the evidence collected as part of the data gathering process. Evidence is defined as the synthesis of high-quality researches that has been collected. SLR has been applied to the literature review because it directs the study toward an understanding of current state of global awareness of the subject. SLR provides a multi-faceted perspective toward the subject by reviewing gaps, highlighting critical areas, and identifying emerging concepts. Furthermore, SLR provides a transparent and reproducible procedure that addresses the research question and elicits patterns, trends, relationships, and portrays the overall picture of the subject (Borrego et al. 2014).

The goal of this study is to analyze the current state of BD RAs, point out major concepts, and discuss limitations. SLR has been found suitable because of the volume of information it is able to handle, findings that emerges, and the perspectives it draws upon. The process of SLR comprises four phases (Figure 1): First phase is identification. In this phase, research questions are stated, publications are selected, and pooled and inclusion and exclusion criteria are set, then mapped against the pooled literature; Second, the quality of the literature is assessed by considering relevance, selection/rejection criteria, and coverage; Third, with coding of literature, reviewing and relevant information captured; and Fourth, synthesis take place, current trends and patterns are understood and delineated.

The SLR data search is guided by a series of questions:

RQ1. What are the fundamental concepts of RA?

RQ2. How can RA help BD system development?

RQ3. What are current BD RAs? What are their main components?

RQ4. How BD RAs are developed? What are the challenges?

RQ5. What are the limitations of current BD RAs?

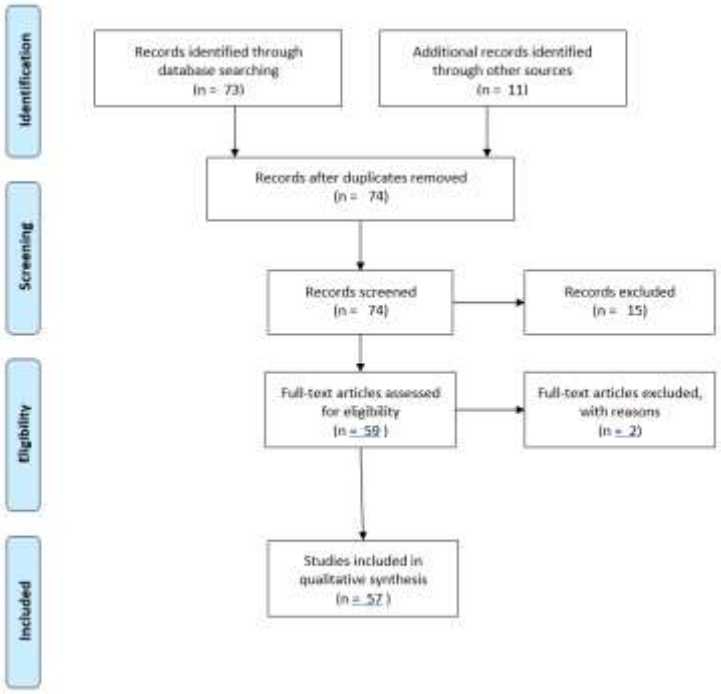

Figure 1: The SLR's PRISMA flowchart (Moher et al. 2009)

2.1 Identification

The Identification phase included a search for articles within the years 2010-2020. Most literature chosen for the purposes of this research are within the years 2016-2020 as they provided recent and more relevant information. Some studies dating back to 2010 helped to clarify fundamental concerns related to BD.

Databases searched were IEEE Explore, ScienceDirect , SpringerLink, and ACM library. To avoid overlooking researches and to make sure that all the studies are captured, abstract and citation databases such as Scopus, Web of Science, Google Scholar, and Research Gate have been utilized.

The highest quality Information Systems (IS) research were found in in MIS Quarterly, whereas Elsevier and AISeL provided good quality BD related sources. A combination of long-tail and short-tail keywords to target literature were used.

The keyword search strings relate the questions above: `Big Data Reference Architectures; Reference Architectures in the domain of Big Data; Reference Architectures and Big Data; Reference Architectures concepts; the concept of Reference Architectures.`

The following inclusion and exclusion criteria were applied.

Inclusion criteria:

- Includes detailed analysis and practice collaborative research
- Includes substantial case studies that dig deep into the data-business context
- Qualitative or quantitative research that identifies industry gaps
- Describe RA concepts
- Demonstrates the current state of BD RAs and discusses possible outcomes
- Provides extensive discussion of BD RA, it's ecosystem and drivers

- Is recent, within the publication range specified
- Is a scholarly publication, conference proceedings, book, book chapter, white paper, thesis or dissertation.

Exclusion criteria:

- Provide poor quality information, possibly identified through short length of article
- Not concerned with practice
- Has been duplicated elsewhere in the data asset
- Does not directly address the research questions
- Not written in English

2.2 Quality Assessment

Four factors to assess quality have been applied as follows:

- Is the article rich in terms of its case studies and relevance to practice?
- Does the article provide sufficient data/information?
- Does the article discuss recent trends in BD RA?

Regarding the first quality factor, richness is defined as the volume and quality of the information. It was important that selected studies should be based on primary data, be internationally focused. For instance, researches that revealed complexities and have been evaluated by prototyping in actual business, has been considered rich in terms of case studies and relevance to practice. In regard to second factor, studies that involved in either creation of a new RA or examination of current RAs have been selected as quality researches and added into the pool. Lastly, articles that aimed at state-of-the-art BD RAs have been added to the pool.

2.3 Data Collection & Synthesis

In the first phase of the SLR (identification), a pool of 84 literature has been selected. Some of which has been added to the pool by the result of backward searching. For instance, we've found Oracle, Facebook, an Amazon RA from NIST RA's references. In phase 2 (screening), literatures have been analyzed for duplicates. We've looked for similar RAs, or similar concepts that has been presented as the findings. We've excluded records that were not in line with our inclusion and exclusion criteria. For instance, if the literature did not discuss main RA concepts, or did not introduce or analyze any BD RA, or its limitations, the paper was excluded. In this study, 15 papers were excluded in phase 2. In the next phase, the quality framework has been applied against the studies. This framework helped eliminating bias by setting clear criteria. In this phase, 2 studies excluded with reason. After setting research questions, demarcating the inclusion criteria, and developing a quality assessment framework, study embarked on a data collection & synthesis process. Nvivo was used to code, label and classify articles. We defined 6 nodes for the purpose of this SLR, namely 'big data reference architecture', 'big data reference architecture limitations', 'reference architecture concepts', 'big data challenges', 'big data reference architecture gaps', 'big data RA development'. These nodes are mapped against the research questions. Studies are coded and classified based on these 5 nodes. Once the articles are all coded, then they are synthesized to induce findings. This process took place with clear consideration on what data is required and how it should be synthesized. The data elicited were; the source, main topic, summary, and research questions. Figure 1 provides a PRISMA flowchart of the study selection process:

3 Findings

In this section, the initials findings are presented. By the result of this work, 57 articles have been selected comprising of proceedings, journal articles, book chapters, and white papers. Out of the pool of articles, 33.3% are from IEEE Explore, 5.2% from ScienceDirect, 24.5% from SpringerLink, 15.7% from ACM, and 21% from Google Scholar. 30 journal articles, 13 conference proceedings, 12 book chapters, 1 white paper, and 1 Master's Thesis were selected. 51% of the articles were selected from the years 2016-2020, 33% belonged to years 2013-2015, and the rest to years 2010-2013.

3.1 The concept of Reference Architecture

In recent years, IT architectures that play a pivotal role in the growth and progress of system development gained acceptance in planning, development, and maintenance of complex systems (Martínez-Fernández et al. 2014; van Engelenburg et al. 2019). When there exists ambiguity about what should be developed to address needs, then an architecture can play an overarching role by describing the building blocks of the system and the ways in which these blocks communicate to achieve the overall goal of the system (Sievi-Korte et al. 2019). This in turn produces manageable modules that address aspects of the problem and provides stakeholders with a high-level medium to observe, reflect upon, communicate with, and contribute to (Kohler and Specht 2019).

Many of the prevalent technologies in IT exist because of an effective RA, for instance the Open Systems Interconnection model or OSI (Zimmermann 1980), Open Authentication or OATH (OATH 2007), Common Object Request Broker Architecture or CORBA (OMG 2014), and WMS or workflow management systems (Grefen and de Vries 1998). This presents the argument then that every system has an architecture and it is in the architecture that the overall qualities of the system are defined and other aspects emerge (Angelov et al. 2013).

The synthesis of findings from this SLR handed over various definitions of what constitute an RA. Howbeit, they all share the same principle that the concept of patterns plays a significant role (Cloutier et al. 2010). For instance, Reed (2002) defines RA as "a predefined architectural pattern, or set of patterns, possible, partially or completely instantiated, designed, and proven for use in particular business and technical contexts, together with supporting artifacts to enable their use".

In Software Product Line (SPL) development, RAs are generic schemas that can be configured and instantiated for a particular domain of systems (Derras et al. 2018). In software engineering, an RA can be defined as a means for representing and transferring knowledge that bridges from the problem domain to a family of solutions (Klein et al. 2016). RAs serve as a mechanism that embodies domain relevant qualities and concepts, breaks down the solutions and generates a terminology to facilitate effective communication, and illuminates various stakeholders and system designers (Klein et al. 2016).

Thus, to answer RQ1, four concepts of RA is identified; these concepts are as the following;

a. **Reference architectures are a high level of abstraction:** Compared to fixed design solutions, RAs tend to capture the essence of the practice and to provide abstractions about elements critical for standardization, communication, implementation and maintenance (Cloutier et al. 2010). Thus, they aim to provide software engineering knowledge as high level architectural patterns and do not provide details regarding specific vendors, platforms or environments.

b. **Reference architectures describe qualities:** RAs are designed for a wider audience and a larger context where design solutions are usually applied to a particular context (Angelov et al. 2008; Stricker et al. 2010). Thus, architectural qualities are bolder in the case of RAs.

c. **Stakeholders are not clearly grouped:** Stakeholders that take part in a solution design process are usually people in the same company that is involved in the project, such as the developer, designer, marketer, data scientist, etc (Geerdink 2013). However, because of the generic nature of an RA, it is not possible to indicate all stakeholders initially. Aside from that, RAs are of a higher level of abstraction and try to provide solution to a class of problems, thus, defining stakeholders can decrease their effectiveness (Chang and Boyd 2018).

d. **Reference architectures encourage adherence to common standards:** The design of an RA is often guided by previous practice, architectures, system developments, and patterns. Therefore, the RA promotes standard approaches that avoid known pitfalls, reduces complexity, and improve reuse (Chang and Boyd 2018).

3.2 Reference architectures and BD

Major technology companies such as Amazon or Google have developed exclusive BD systems with sophisticated data management, procurement and analysis capabilities (Kohler and Specht 2019). Being resourceful, these companies have attracted talent from around the globe to manage the complexity of BD systems. However, that's not the reality of many other organizations. BD analytics sails away from traditional small data analytics and brings various challenges, including rapid technology changes (Chen et al. 2016), organizational limitations (Rada et al. 2017; Vassakis et al. 2018), and tech nical challenges (Jagadish et al. 2014). This is due to the inherent characteristics of BD, namely, velocity, variety, value, veracity and volume.Whereas these challenges do not only belong to the BD domain, BD exacerbates these issues because (1) it requires near real-time performance, (2) scalablity (has to scale

when needed), (3) technology orchestration (the communication between components and data flow), and (4) continuous delivery (insight dissemination to components, business units, etc) (Jagadish et al. 2014). Based on the findings gained from this SLR, it is estimated that approximately 75% of the BD projects have failed within the last decade (AI 2019; Gartner 2014; McKinsey 2016; Nash 2015; Partners 2019; White 2019). BD practitioners and academics, report that BD is an interplay of analysis (statistics, math), technology (software, tools) and methodology (organization, workflow) (Akhtar et al. 2019; Rad and Ataei 2017). Thus, the deciding factor and the key element of picking up this new magical wand lies in its technology orchestration. Positioned on top of the rationale discussed and to answer RQ2, RAs can be perceived as effective artefacts that help with technology orchestration, variability management, interface definition, component delineation, communication and lastly maintenance of BD systems (Chang and Boyd 2018; Nadal et al. 2017).. RAs open the door to efficient and effective deployment by providing a systematic way of deriving and synthesizing BD systems that meet the requirements of the context (Nadal et al. 2017). Most authors agree that issues around BD system development and software engineering processes are severe and that this justified the use of RAs (Chang and Boyd 2018).

Using an RA would mean that the software architect is no longer challenged to form a new architecture from a set of independent components that needs to be assembled, and instead they can refer to an already created orchestration of components, their relations, variability points and map them against organizational resources, quality attributes and the domain. Therefore, the RA is a step towards elucidation and homogenization of BD systems. This approach has been successfully applied for Distributed Database Management Systems (DDBMS) (Rahimi and Haug 2010), and Database Management Systems (DBMS) (Piñeiro et al. 2019).

3.3 Current BD Reference Architectures?

At the present time, many organizations including governmental agencies have taken steps to standardize BD systems development through RAs(Li et al. 2019). The main product of this SLR is the collection of 23 RAs from the extant literature; 18 RAs from academia, 4 from practice, and 1 through the collaboration of both domains has been found. Table 1 provides an overview of these RAs.

Whereas most RAs were usually in the form of short papers and did not provide with much detail, some are very detailed, such as NIST. Many of the RAs analyzed have been inspired by other RAs, and this provides the notion that "RAs can be proven more effective when they are created out of a studied domain and available knowledge rather than from scratch". To address RQ3, RAs listed in Table 1 are compared and reviewed to highlight commonalities. Three key components have been identified as follows;

a. **BD Management and Storage** - SQL and NoSQL, Distributed file system, NewSQL, polyglot persistence, data lake, data finery, data swamp
b. **BD Analytics and Application Interfaces** - Real-time, Batch analytics, Reporting, Descriptive, predictive and spatial analysis
c. **BD Infrastructure** - In memory data grids, latency, optimal data transformation

RAs from Practice	**Domain**
SAP - NEC Reference Architecture for SAP HANA & Hadoop (SAP 2016)	Practice
IBM - Reference architecture for high performance analytics in healthcare and life science (Mysore et al. 2013)	Practice
Oracle - Information Management and Big Data: A Reference Architecture (Oracle 2014)	Practice
Microsoft - Big Data ecosystem reference architecture (Microsoft)	Practice
Lambda architecture (Marz 2011)	Practice
The solid architecture for real-time management of big semantic data; Solid architecture (Martínez-Prieto et al. 2015)	Academia
Questioning the Lambda architecture; Kappa Architecture (Kreps 2014)	Academia

A software reference architecture for semantic-aware Big Data systems; Bolster Architecture (Nadal et al. 2017)	Academia
Towards a big Data reference architecture (Maier et al. 2013)	Academia
Scalable data store and analytic platform for real-time monitoring of data-intensive scientific infrastructure (Suthakar 2017)	Academia
Big data architecture framework (Demchenko et al. 2014)	Academia
Towards a secure, distributed, and reliable cloud-based reference architecture for big data in smart cities (Kohler and Specht 2019)	Academia
A reference architecture for big data systems in the national security domain (Klein et al. 2016)	Academia
The big data research reference architecture (Pääkkönen and Pakkala 2015)	Academia
NIST big data interoperability framework (Chang and Boyd 2018)	Practice-Academia
Managing Cloud-Based Big Data Platforms: A Reference Architecture and Cost Perspective (Heilig and Voß 2017)	Academia
a proposal for a reference architecture for long-term archiving, preservation, and retrieval of big data (Viana and Sato 2014)	Academia
A Reference Architecture for Big Data Systems (Sang et al. 2016)	Academia
A reference architecture for big data solutions introducing a model to perform predictive analytics using big data technology (Geerdink 2013)	Academia
Big Data driven e-commerce architecture (Ghandour 2015)	Academia
Reference Architectures and Standards for the Internet of Things and Big Data in Smart Manufacturing (Ünal 2019)	Academia
A Reference Architecture for Supporting Secure Big Data Analytics over Cloud-Enabled Relational Databases (Cuzzocrea 2016)	Academia

Table 1: Articles that provide Ras and their alignment to practice and academic authorship

3.4 How are BD RAs developed?

Our findings bring forth the fact that there are not many frameworks available for developing RAs. Nonetheless, to answer RQ4, we aimed to search for most utilized approaches for developing BD RAs. One of the accepted approaches for developing RAs that has been repeatedly noted is 'Empirically-grounded Reference Architectures' by Galster and Avgeriou (2011). This approach is well received because it emphasizes empirical foundations and empirical validity. It is a 6-step process that starts with a decision on the type of RA, selection of design strategy, empirical acquisition of data, construction of the RA, enabling RA with variability and ends with evaluation of the RA. Along these lines and to aid in deciding on what type the RA should be, the RA classification framework created by Angelov et al. (2012) has been used commonly.

Based on the classification framework by Angelov et al. (2012), five types of RAs are defined. This framework has been developed with the goal of supporting analysis of RAs with regards to context, goal, and the architecture specification/design. As a result, the framework is based on three dimensions, each having their own corresponding sub-dimensions: context, goals, and design. These dimensions and sub-dimensions are derived by interrogatives where the use of interrogatives is a well-established practice for problem analysis. The interrogatives When, Where, and Who address the context, Why addresses the goal, and How and What address the design dimension (Galster and Avgeriou 2011).

This study categorizes RAs as two types; Standardization RAs and Facilitation RAs. Moreover, Volk et al. (2019) describes a decision-support process for the selection of BD RAs, which investigated 6 BD RAs and used the Software Architecture Comparison Analysis Method (SCAM) to examine and compare RAs based on their applicability. Additionally, ISO/IEC 25010 is the most commonly used standard among researchers for choosing quality software products for their RAs (Iso 2011), and Archimate is the most common tool for designing RAs and creating ontologies (Syynimaa 2018).

3.4.1 What are the challenges?

One of the main challenges of RA development is evaluation (Maier et al. 2013) (Cioroaica et al. 2019). Two pillars of evaluation are the correctness and the utility of the RA, and how efficient it can be adapted and instantiated (Galster and Avgeriou 2011). The quality of an RA can be assessed by how it can be transformed into an effective organization-specific concrete architecture. RA and concrete architecture have distinct qualities and while there are many well-established methods for assessing concrete architectures such as Scenario-based Architecture Analysis Method (SAAM) (Kazman et al. 1996), Architecture Level Modifiability Analysis (ALMA) (Bengtsson et al. 2004), Performance Assessment of Software Architecture (PASA) (Williams and Smith 2002) and Architecture Trade-off Analysis Method (ATAM) (Kazman et al. 1998), none of these methods can be directly applied to the evaluation of RAs. One of the main problems for applying existing evaluation methods to an RA is the lack of a clearly defined group of stakeholders (Angelov et al. 2008), but ATAM and other methods are highly dependent on the participation of stakeholders for evaluation (Kazman et al. 1998). Secondly, evaluation frameworks and methods for concrete architectures make use of scenarios. However, due to the RA level of abstraction, the creation of a usable scenario is difficult (Angelov et al. 2008). Either a large set of scenarios should be developed, covering all the aspects of the RA with regards to a specific domain, or a more general scenario should be developed to cover all the aspects. Lastly, RAs are about architectural qualities, rather than comparing with concrete architectures and existing evaluation methods fall short in when considering high-level architectural qualities (Angelov et al. 2008).

3.5 Limitations of Current BD RAs

Next, to address RQ5, RAs collected by the result of this SLR have been appraised, and a few limitations have been identified. First, the notion of metadata management has been scarcely discussed in the literature. Except for one case, metadata management has not been accentuated and metadata layers are not defined. Metadata plays a significant role in BD system development as it addresses a wide range of issues such as security, privacy, linear analysis, and data provenance (Eichler 2019). Any BD system development can benefit from a clear architectural view of metadata as a means of bridging data stored in different platforms such as on Cloud or on premise (Chang and Boyd 2018). This can in turn reduce complexity, facilitate data governance, facilitate access management, facilitate change management, and help with defining the general scope of the system. Secondly, white papers published from big IT companies tend to pivot their RA around their services, which hinders its openness and reduces its applicability. Alternative technologies or approaches have typically not been discussed and practitioners are left with a small pool of choices.

Academic studies have motivated aspects of this research, but they suffer from the same limitations stated above. Metadata management is poorly discussed, and some have not taken recent emerging architectures into account, but this may be a factor of age in the publication. For instance, the RA described by Maier et al. (2013) has got limitations in the area of technology description. The concept of technology description is coarse-grained and emphasizes mostly the class of products, which can be troublesome and potentially confusing to system designers and engineers. Furthermore, the study revolves around infrastructure software, leaving crucial areas such as analytical methods, and architectural qualities almost untouched.

Moreover, two major factors that are crucial to any BD system development, privacy and security, are scarcely touched. With regards to recent movements of data privacy, BD systems should now be designed underlying the shadow of regional data privacy policies and aim to achieve good security standards (Bashari Rad et al. 2016). Many security and privacy concerns may arise that can bring BD system development into bottleneck, there is no clear guidelines found in the RAs studies and the notion of security and privacy is scarcely discussed.

4 Conclusion

The SLR has sought to find all available BD RAs in practice and academia. The findings bring the understanding that RAs can be one good solution to tackle complex BD system developments and serve as a guide. RAs are conceptual artefacts that promote the application of software engineering knowledge

and patterns into development. As BD system development is inherently a technology orchestration, RAs can be an effective start to design and implementation. RAs help by defining functional and non-functional requirements and delineate variability points. This helps the organization to tackle BD projects successfully and avoid common pitfalls. The limitations in existing studies are around security, privacy, and metadata management. These areas require more attention and future research.

5 References

AI, V. 2019. "Why Do 87% of Data Science Projects Never Make It into Production?", 2019, from https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/

Akhtar, P., Frynas, J. G., Mellahi, K., and Ullah, S. 2019. "Big Data-Savvy Teams' Skills, Big Data-Driven Actions and Business Performance," *British Journal of Management* (30:2), pp. 252-271.

Angelov, S., Grefen, P., and Greefhorst, D. 2012. "A Framework for Analysis and Design of Software Reference Architectures," *Information and Software Technology* (54:4), pp. 417-431.

Angelov, S., Trienekens, J., and Kusters, R. 2013. "Software Reference Architectures-Exploring Their Usage and Design in Practice," *European Conference on Software Architecture*: Springer, pp. 17-24.

Angelov, S., Trienekens, J. J., and Grefen, P. 2008. "Towards a Method for the Evaluation of Reference Architectures: Experiences from a Case," *European Conference on Software Architecture*: Springer, pp. 225-240.

Bashari Rad, B., Akbarzadeh, N., Ataei, P., and Khakbiz, Y. 2016. "Security and Privacy Challenges in Big Data Era," *International Journal of Control Theory and Applications* (9:43), pp. 437-448.

Borrego, M., Foster, M. J., and Froyd, J. E. 2014. "Systematic Literature Reviews in Engineering Education and Other Developing Interdisciplinary Fields," *Journal of Engineering Education* (103:1), pp. 45-76.

Chang, W. L., and Boyd, D. 2018. "Nist Big Data Interoperability Framework: Volume 6, Big Data Reference Architecture."

Chen, H.-M., Kazman, R., and Haziyev, S. 2016. "Agile Big Data Analytics for Web-Based Systems: An Architecture-Centric Approach," *IEEE Transactions on Big Data* (2:3), pp. 234-248.

Cloutier, R., Muller, G., Verma, D., Nilchiani, R., Hole, E., and Bone, M. 2010. "The Concept of Reference Architectures," *Systems Engineering* (13:1), pp. 14-27.

Cuzzocrea, A. 2016. "A Reference Architecture for Supporting Secure Big Data Analytics over Cloud-Enabled Relational Databases," *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*: IEEE, pp. 356-358.

Demchenko, Y., De Laat, C., and Membrey, P. 2014. "Defining Architecture Components of the Big Data Ecosystem," *2014 International Conference on Collaboration Technologies and Systems (CTS)*: IEEE, pp. 104-112.

Derras, M., Deruelle, L., Douin, J.-M., Levy, N., Losavio, F., Pollet, Y., and Reiner, V. 2018. "Reference Architecture Design: A Practical Approach," *ICSOFT*, pp. 633-640.

Eichler, R. K. 2019. "Metadata Management in the Data Lake Architecture."

Erevelles, S., Fukawa, N., and Swayne, L. 2016. "Big Data Consumer Analytics and the Transformation of Marketing," *Journal of Business Research* (69:2), pp. 897-904.

Galster, M., and Avgeriou, P. 2011. "Empirically-Grounded Reference Architectures: A Proposal," *Proceedings of the joint ACM SIGSOFT conference--QoSA and ACM SIGSOFT symposium--ISARCS on Quality of software architectures--QoSA and architecting critical systems--ISARCS*: ACM, pp. 153-158.

Gartner. 2014. "Survey Analysis: Big Data Investment Grows but Deployments Remain Scarce in 2014.", from http://www.gartner.com/document/2841519

Geerdink, B. 2013. "A Reference Architecture for Big Data Solutions Introducing a Model to Perform Predictive Analytics Using Big Data Technology," *8th International Conference for Internet Technology and Secured Transactions (ICITST-2013)*: IEEE, pp. 71-76.

Ghandour, A. 2015. "Big Data Driven E-Commerce Architecture," *International Journal of Economics, Commerce and Management* (3:5), pp. 940-947.

Grefen, P., and de Vries, R. R. 1998. "A Reference Architecture for Workflow Management Systems," *Data & Knowledge Engineering* (27:1), pp. 31-57.

Heilig, L., and Voß, S. 2017. "Managing Cloud-Based Big Data Platforms: A Reference Architecture and Cost Perspective," in *Big Data Management*. Springer, pp. 29-45.

Iso, I. 2011. "Iec25010: 2011 Systems and Software Engineering–Systems and Software Quality Requirements and Evaluation (Square)–System and Software Quality Models," *International Organization for Standardization* (34), p. 2910.

Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., and Shahabi, C. 2014. "Big Data and Its Technical Challenges," *Communications of the ACM* (57:7), pp. 86-94.

Kitchenham, B. A., Dyba, T., and Jorgensen, M. 2004. "Evidence-Based Software Engineering," *Proceedings of the 26th international conference on software engineering*: IEEE Computer Society, pp. 273-281.

Klein, J., Buglak, R., Blockow, D., Wuttke, T., and Cooper, B. 2016. "A Reference Architecture for Big Data Systems in the National Security Domain," *2016 IEEE/ACM 2nd International Workshop on Big Data Software Engineering (BIGDSE)*: IEEE, pp. 51-57.

Kohler, J., and Specht, T. 2019. "Towards a Secure, Distributed, and Reliable Cloud-Based Reference Architecture for Big Data in Smart Cities," in *Big Data Analytics for Smart and Connected Cities*. IGI Global, pp. 38-70.

Kreps, J. 2014. "Questioning the Lambda Architecture. The Lambda Architecture Has Its Merits, but Alternatives Are Worth Exploring." O'Reilly Media on line, July.

Li, Q., Xu, Z., Chan, I., Yang, S., Pu, Y., Wei, H., and Yu, C. 2019. "Big Data Architecture and Reference Models," Cham: Springer International Publishing, pp. 15-24.

Maier, M., Serebrenik, A., and Vanderfeesten, I. 2013. "Towards a Big Data Reference Architecture," *University of Eindhoven*).

Martínez-Fernández, S., Ayala, C., Franch, X., and Marques, H. M. 2014. "Artifacts of Software Reference Architectures: A Case Study," *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, pp. 1-10.

Martínez-Prieto, M. A., Cuesta, C. E., Arias, M., and Fernández, J. D. 2015. "The Solid Architecture for Real-Time Management of Big Semantic Data," *Future Generation Computer Systems* (47), pp. 62-79.

Marz, N. 2011. "How to Beat the Cap Theorem," *Thoughts from the Red Planet*).

McKinsey. 2016. "The Age of Analytics: Competing in a Data-Driven World."

Microsoft. "Microsoft Big Data Solution Brief." from http://download.microsoft.com/download/F/A/1/FA126D6D841B-4565-BB26-D2ADD4A28F24/

Microsoft_Big_Data_Solution_Brief.pdf

Mikalef, P., Pappas, I. O., Krogstie, J., and Giannakos, M. 2017. "Big Data Analytics Capabilities: A Systematic Literature Review and Research Agenda," *Information Systems and e-Business Management*), pp. 1-32.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and The, P. G. 2009. "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The Prisma Statement," *PLOS Medicine* (6:7), p. e1000097.

Mysore, D., Khupat, S., and Jain, S. 2013. "Big Data Architecture and Patterns," *IBM, White Paper*).

Nadal, S., Herrero, V., Romero, O., Abelló, A., Franch, X., Vansummeren, S., and Valerio, D. 2017. "A Software Reference Architecture for Semantic-Aware Big Data Systems," *Information and software technology* (90), pp. 75-92.

Nash, H. 2015. "Cio Survey 2015," *Association with KPMG*).

OATH. 2007. "Oath Reference Architecture, Release 2.0 Initiative for Open Authentication " *OATH*).

OMG. 2014. ""Common Object Request Broker Architecture:

Core Specification," *OMG, Inc*).

Oracle. 2014. "Information Management and Big Data." from http://www.oracle.com/technetwork/database/bigdataappliance/overview/bigdatarefarchitecture-2297765.pdf

Pääkkönen, P., and Pakkala, D. 2015. "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems," *Big data research* (2:4), pp. 166-186.

Paré, G., Tate, M., Johnstone, D., and Kitsiou, S. 2016. "Contextualizing the Twin Concepts of Systematicity and Transparency in Information Systems Literature Reviews," *European Journal of Information Systems* (25:6), pp. 493-508.

Partners, N. 2019. "Big Data and Ai Executive Survey 2019," *Data and Innovation*).

Piñeiro, C., Morales, J., Rodríguez, M., Aparicio, M., Manzanilla, E. G., and Koketsu, Y. 2019. "Big (Pig) Data and the Internet of the Swine Things: A New Paradigm in the Industry," *Animal Frontiers* (9:2), pp. 6-15.

Rad, B. B., and Ataei, P. 2017. "The Big Data Ecosystem and Its Environs," *International Journal of Computer Science and Network Security (IJCSNS)* (17:3), p. 38.

Rada, B. B., Ataeib, P., Khakbizc, Y., and Akbarzadehd, N. 2017. "The Hype of Emerging Technologies: Big Data as a Service,").

Rahimi, S. K., and Haug, F. S. 2010. *Distributed Database Management Systems: A Practical Approach*. John Wiley & Sons.

Reed, P. 2002. "Reference Architecture: The Best of Best Practices." from https://www.ibm.com/developerworks/rational/library/2774.html

Sang, G. M., Xu, L., and De Vrieze, P. 2016. "A Reference Architecture for Big Data Systems," *2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)*: IEEE, pp. 370-375.

SAP. 2016. "Nec Reference Architecture for Sap Hana & Hadoop,").

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., and Stewart, L. A. 2015. "Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (Prisma-P) 2015: Elaboration and Explanation," *Bmj* (349).

Sievi-Korte, O., Richardson, I., and Beecham, S. 2019. "Software Architecture Design in Global Software Development: An Empirical Study," *Journal of Systems and Software* (158), p. 110400.

Sivarajah, U., Kamal, M. M., Irani, Z., and Weerakkody, V. 2017. "Critical Analysis of Big Data Challenges and Analytical Methods," *Journal of Business Research* (70), pp. 263-286.

Stricker, V., Lauenroth, K., Corte, P., Gittler, F., De Panfilis, S., and Pohl, K. 2010. "Creating a Reference Architecture for Service-Based Systems-a Pattern-Based Approach," *Future Internet Assembly*, pp. 149-160.

Suthakar, U. 2017. "A Scalable Data Store and Analytic Platform for Real-Time Monitoring of Data-Intensive Scientific Infrastructure." Brunel University London.

Syynimaa, N. 2018. "Essence: Reference Architecture for Software Engineering-Representing Essence in Archimate Notation," *International Conference on Enterprise Information Systems*: SCITEPRESS Science And Technology Publications.

Ünal, P. 2019. "Reference Architectures and Standards for the Internet of Things and Big Data in Smart Manufacturing," *2019 7th International Conference on Future Internet of Things and Cloud (FiCloud)*: IEEE, pp. 243-250.

van Engelenburg, S., Janssen, M., and Klievink, B. 2019. "Design of a Software Architecture Supporting Business-to-Government Information Sharing to Improve Public Safety and Security," *Journal of Intelligent information systems*), pp. 1-24.

Vassakis, K., Petrakis, E., and Kopanakis, I. 2018. "Big Data Analytics: Applications, Prospects and Challenges," in *Mobile Big Data*. Springer, pp. 3-20.

Viana, P., and Sato, L. 2014. "A Proposal for a Reference Architecture for Long-Term Archiving, Preservation, and Retrieval of Big Data," *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*: IEEE, pp. 622-629.

Volk, M., Bosse, S., Bischoff, D., and Turowski, K. 2019. "Decision-Support for Selecting Big Data Reference Architectures," *International Conference on Business Information Systems*: Springer, pp. 3-17.

White, A. 2019. "Top Data and Analytics Predicts for 2019."

Zimmermann, H. 1980. "Osi Reference Model-the Iso Model of Architecture for Open Systems Interconnection," *IEEE Transactions on communications* (28:4), pp. 425-432.