The state of big data reference architectures: a systematic literature review

Abstract

This template helps you to create a properly formatted LaTeX manuscript.

Keywords: `elsarticle.cls`, LaTeX, Elsevier, template

2010 MSC: 00-01, 99-00 s

1. Introduction

The rapid development of software technologies, the proliferation of digital devices and networking infrastructure of today, have by and large, augmented user's capability to generate data [1]. In the age of information, users are unceasing generators of structured, semi-structured, and unstructured data that if collected and crunched correctly, may reveal game-changing patterns [2].

The unprecedented proliferation of data have emerged a new ecosystem of technologies; one of these ecosystems is big data (BD)[3]. BD is a term emerged to describe large amount of data that comes in various forms from different channels. Within the years, BD has attained a lot of attention from academia and industry, and many strive to benefit from this new material. Howbeit, adopting BD requires the absorption of great deal of complexity and many traditional systems cannot cope with characteristics of this domain.

A recent survey published by Databricks in partnership with MIT Technology Review Insights, stated that only 13% of companies excel at delivering on their data strategy [4]. In the same vein, Vintage Partners highlighted that only 24% of companies have successfully adopted BD [5]. Sigma computing report presented that 1 in 4 business experts have given up on getting insights they needed because the data processing took too long [6]. Moreover, Gartner

approximated that only 20% of companies have successfully adopted BD.

Some of the most highlighted challenges of BD is 'lack of business context', 'organizational challenges', 'BD architecture', 'data engineering', 'rapid technology change', and 'lack of talent' [7]. Whereas similar issues may exist in other domains, it is exacerbated when it comes to BD systems. This is due the inherent complexity of BD engineering, the need for real-time processing, the scalability requirement of these systems, and the sensitivities around data.

Today, majority of BD systems are designed underlying ad-hoc and complicated architectural solutions [8], that do not seem to adhere to similar patterns. This will challenge software architects to design a suitable solution for any given context, creates a foundation for an immature architectural decision, and does not promote the growth and development of BD systems as a whole.

Therefore, since the approach of ad-hoc design to BD systems is undesirable and leaves many engineers in the dark, there is a need for more software engineering research for BD systems. To this end, this study presents a systematic literature review (SLR) on BD (BD) reference architectures (RAs).

2. Why reference architectures?

Conceptualization of the system as an RA, helps with understanding of the system's key components, behavior, composition and evolution of it, which in turn affect quality attributes such as maintainability, scalability and performance [9]. Therefore RAs can be a good standardization artefact and a communication medium that not only results in concrete architectures for BD systems, but also provide stakeholders with unified elements and symbols to discuss and progress BD projects.

This approach to system development is not new to practitioners of complex system. In software product line (SPL) development, RAs are utilized as generic artifacts that are instantiated and configured for a particular domain of systems [10]. In software engineering, IT giants like IBM have referred to RAs as the 'best of best practices' to address complex and unique system design challenges

2

[9]. In other international standardization, RAs have been repeatedly used to standardize an emerging domain, a good example of this is BS ISO/IEC 18384-1 RA for service oriented architectures [11].

3. State of the art

Despite the undeniable benefits of RAs, and their potential to solve some of the complex issues of BD systems, we think that this area is underdeveloped and needs more attention from both academia and practice. This insight is derived from our preliminary systematic review in academia, and a search for available big data RAs ([2]).

To the best of our knowledge, one of the most comprehensive BD RA published, is the National Institute of Standards and Technology (NIST) BD RA. This RA is published by Big Data Public Working Group (NBD-PWG) with large set of contributors from academia, industry, non-profit organizations, agents, and government representatives. This was announced as an initiative from White house in March 2012, and the the RA was published under the title 'NIST Big Data Interoperability Framework: Volume 6, Reference Architecture' in October 2019.

Given the substantial investment on BD RAs, one might infer the value of these artifacts, and this can in turn highlights the necessity for more research in this domain. Another factor that worths mentioning is how vaguely the phrase 'reference architecture' is defined and institutionalized. For instance, the difference between a 'concrete architecture' and an RA is hardly discussed, and different domains seem to have defined the artifact slightly differently. For instance, Cloutier et al ([9]) defined RAs as 'Reference Architectures capture the essence of existing architectures, and the vision of future needs and evolution to provide guidance to assist in developing new system architectures'. This definition is derived from the system engineering domain and by the means of collaborative forum from Steven's institute of technology.

In another effort, Muller et al ([12]) defines RA as 'artifacts that captures

3

the essence of architecture of a collection of systems. This definition is driven from the product line engineering domain'. Moreover, the difference between RAs and concrete architectures is rarely discussed. Another definition by Bass et al ([13]) stated that 'A reference architecture is a reference model mapped onto software elements (that cooperatively implement the functionality defined in the reference model) and the data flows between them'.

Angelov et al ([14]) defined RAs proposed that 'A reference architecture is a generic architecture for a class of information systems that is used as a foundation for the design of concrete architectures from this class'. Although different authors may have defined RAs with different syntax, the essence remains the same: to reuse the software engineering knowledge for a class of systems, particularly in relation to architecture.

Given the failure rate of BD projects, we posit RAs as potential solution to facilitate system development and BD architecture, and aim to explore this area through a systematic literature review. Up to date, there's only one SLR that explored this area ([2]), which is outdated, suffers from methodological clarity, and is published as a conference paper, which implies lack of detail.

Based on this, the objective of this review is to find and collate the BD RAs available from the body of evidence, highlight their architectural commonality and point out the limitations. This study can be considered a useful primer for practitioners or academics who are interested in partaking in a BD project.

The research questions are formulated as the following;

1. What are current BD RAs available in academia and industry?

2. What are major architectural components of these BD RAs?

3. What are the limitations of current BD RAs?

4. Review Methodology:

This research follows the guidelines of PRISMA ([15]). In addition, we adopted PRISMA-S ([16]) to improve our search strategy and lastly we have used Barbara et al's guidelines for evidence based software engineering and

4

systematic reviews [17]. Although PRISMA is a comprehensive guidelines on conducting a systematic literature review, it is derived from the healthcare community and sometimes makes assumptions that may not be relevant to software engineering and information system researchers. Barbara et al [17] has translated many of these assumptions to the domain of software engineering and included many guidelines for lone researchers and projects with small number of researchers.

We have therefore utilized PRISMA as the underpinning of our research design, with complementary studies to reduce bias, improve transparency and systematiticity. SLR has been chosen because it is a qualitative research methodology that is aimed at driving knowledge and understanding about the subject matter and the elements surrounding it. Besides, SLR provides a transparent and reproducible procedure that elicits patterns, relationships, trends, and delineates the overall picture of the subject [18].

The main objective of this study is to assess the current state of BD RAs, identify their major architectural components, point out fundamental concepts and discuss their limitations. This objective is achieved in four phases. In first phase, research questions are stated, literature are identified and pooled, exclusion and inclusion criteria are defined, and the quality framework is developed. In second phase, the title of the studies are assessed based on the inclusion and exclusion criteria. After that, the filtered studies are once more assessed based on their title, abstract, introduction and conclusion. After this, full analysis of the studies took place by running each study against the criteria defined in the quality framework. Thirdly, selected pool of literature is coded based on research questions. Lastly, findings are synthesized by the means of thematic synthesis, and themes realized are depicted.

This study builds on the SLR conducted by Ataei et al [2] and aims to improve it by covering the years 2020 to 2022. Unlike Ataei's work, this paper aims to employ thematic synthesis, and provide a more detailed view of BD RAs and their properties.

4.1. Identification

The first phase of the SLR began, by adoption of PRISMA-S ([16]) to develop a robust multi-database search strategy. This extension of PRISMA provided us with a framework of 12 items to increase transparency, systematiticity, and reduce bias. For the purposes of this study, following electronic databases were search: ScienceDirect, IEEE Explore, SpringerLink, AISeL, JSTOR and ACM library. To pursue to goal of finding all literature available on the topic, and to avoid overlooking valuable research, abstract and citation databases and search engines such as Google Scholar, and Research Gate was used.

We also searched the grey literature on the topic, using the search string "big data" AND "reference architecture*" on Google (in June 2022). The first 40 results were selected for screening. This was done in 'incognito mode' to avoid any personal customization of the google search pages. Reference lists of included studies were manually screened to identify additional studies. This is to achieve the critical component of 'completeness' as suggested by Kitchenham et al [17].

The platform search capabilities varied, but our search strategy remained uniform for most parts. For instance, if a platform did not support wildcards (like asterisk), we just searched twice for the singular and plural version of the word. The only exception that made the selection process longer was Springer-Link, because it did not support bulk download of references in BibTex format. The reproducible search for the chosen databases is as follows:

- ("Document Title":big data) AND ("Document Title":reference architecture) OR ("Document Title":big data architecture)

The reason we included architecture is due to the fact that terms *reference architecture* and *architecture* may have been used interchangeably, and an architecture that is at the abstraction level of an RA, might have been called just an architecture. Therefore it was critical for us to firmly define these terms and then categorize studies based on these definitions. These definitions and our findings are depicted in the findings section.

Our initial search was set to year 2020 to year 2022, as the work of Ataei et al [2] covered the years 2010-2020. Nevertheless, we still included the years 2010 to 2020 to make sure no research is left out or overlooked. These years are chosen firstly because more contemporary researches are focused on the facilitation of big data system development, and secondly there's no SLR that has covered those.

It is worth mentioning that what we refer to by *limit* here should, not be confused with *filters* or inclusion criteria. To achieve these limits, we have utilized databases features. All databases supported the selection of year range, and the language limit was automatically applied by doing an advanced search with the aforementioned keywords.

Our approach to systematic collection of evidence was to search databases using the keywords aforementioned and then bulk download the BibTex files. Majority of the databases supported bulk downloading of BibTex files except for SpringerLink, Google Scholar, and Research Gate. For SpringerLink we downloaded the studies in CSV format and then converted them to a BibTex using a custom script. For Google Scholar and ResearchGate, unfortunately, we had to take the manual path of creating a bib file for the studies.

Once all the bib files have been created, we merged them into one large bib file and imported it to a software called JabRef ([19]) for deduplication. 172 studies are pooled initially, out of which 6 duplicates have been identified. We removed the SLR that this study is based on, and also another paper that we could not find the citation for. In the other hand, we found 5 white papers and 4 website blogs and added them to the selection pool. At the end of this phase, 173 studies have been pooled.

4.2. Screening and Eligibility

Stage 1 of screening started with assessing the title, abstract, and keywords of the pooled studies. For grey literatures simply the title. This was achieved based on our inclusion and exclusion criteria;

- Primary and secondary studies (including grey literature) between Jan 1st 2010 and June 1st 2022 on the topics of BD RAs, BD models, and BD architectural components were included.

- Research that Indicates the current state of RAs in the field of BD and demonstrates possible outcomes

- Studies that are scholarly publications, book, book chapter, thesis, dissertation, or conference proceedings

- Grey literature such as white paper that includes extensive information on BD RAs

And the studies with the following topics were excluded:

- Informal literature surveys without any clearly defined research questions or research process

- Duplicate reports of the same study (a conference and journal version of the same paper)

- Short papers (less than 5 pages)

- Studies that are not written in English

Disagreement among researchers were resolved using Krippendorff's alpha ([20]). Our aim was not to get involved in a very complicated statistics model, so we've done most of the computations using SPSS, specifically with Hayes' Macro. We made sure that a separate file is created for each variable, and inserted coders as variables and not a constant value. Our

$$\alpha \tag{1}$$

value was within the acceptable range (above 80), and any disagreement was solved by inviting a third person or a moderator. When

$$\alpha \tag{2}$$

value was very low (indicating a low reliability), we stopped the process, and
tried to clarify fundamental concepts and categories. The final computed

$$\alpha \qquad\qquad (3)$$

value was 89.9%.

In stage 2, After excluding papers based on inclusion and exclusion criteria, and as suggested by Kitchenham et al [17], we assessed studies based on their quality. Quality of the evidence collected as a result of this SLR has direct impact on the quality of the findings, making quality assessment an important undertaking. Therefore it is imperative for us to realize how much confidence we can place in the conclusions and findings arising from the evidence collected to form a great whole, that is themes and models in this case.

However, this process comes with some well-known complexities. The most fundamental ones are perhaps firstly defining the term 'quality', and secondly trying to appraise the quality of conference papers that rarely provide enough detail on research methodology and evaluation. Generally, a quality of a study is tightly associated to its research method and the validity of its findings. From this perspective, and inspired by the works of Noblit and Hare on meta-ethnography ([21]), and Dyba et al ([22]), quality of studies is assessed by the extent to which the conduct, design and analysis of a research is susceptible to systematic errors or bias ([23]). That is, the more bias in the selected literature, the more chance to create miss-leading conclusions.

Considering the rather heterogeneous nature of software engineering and information systems (IS) papers, and difficulty of defining quality in studies with varying nature, we first analyzed a few well-established checklists such as Critical Appraisal Skills Programme (CASP [24]), and JBI's critical appraisal tool ([25]). Whereas these checklists could potentially account for the requirements of this study, we opted for something that is more specific to software engineering and IS. We realized for example that, Runeson et al ([26]) provided a checklist designated to help researchers reading and undertaking software engineering case studies. In the same vein, Dyba et al ([22]) proposed a quality criteria based

9

on CASP checklist for qualitative studies in software engineering systematic reviews.

Nevertheless, the challenge is that our study includes a large number of different study types that needs to go through a single checklist. To address this, we developed a criteria made up of 11 elements. These criteria are informed by those proposed by CASP for assessing the quality of qualitative research ([24]) and by guidelines provided by Kitchenham ([27]) on empirical research in software engineering. The 7 criteria tested literature on 4 major areas that can critically affect the quality of the studies. These categories and the corresponding criteria are as following;

1. *Minimum quality threshold:*

    (a) Does the study report empirical research or is it merely a 'lesson learnt' report based on expert opinion ?

    (b) The objectives and aims of the study is clearly communicated, including the reasoning for why the study was undertaken ?

    (c) Does the study provide with adequate information regarding the context in which the research was carried out ?

2. *Rigour:*

    (a) Is the research design appropriate to address the objectives of the research ?

    (b) Is there any data collection method used and is it appropriate ?

3. *Credibility:*

    (a) Does the study report findings in a clear and unbiased manner ?

4. *Relevance:*

    (a) Does the study provides value for practice or research

Taken all together, these 7 criteria gave us a measure of the extent to which a particular study's findings could make a valuable contribution to the review. These criteria was disseminated as a checklist among researchers with value for each property being dichotomous, that is 'yes' or 'no' in two phases. In the first phase, researchers only assess the quality based on the first major area

10

(minimum quality threshold). If the study passed the first phase, it would then go into the second phase, where it was assessed for credibility, rigour and relevance. The quality is agreed if 75% of the responses are positive for any given study with at least 75% inter-rater reliability.

Disagreements regarding the quality was usually resolved through a meeting. While, the meeting could not address the disagreements, a moderator has been invited to the process. Lastly, it is worth mentioning that this quality framework was not used for grey literature. Grey literature were only assessed through inclusion and exclusion criteria.

4.3. Data Collection and Synthesis

In the first phase (identification) of this SLR, a total of 138 literature has been pooled from academia, and 24 from grey literature. Some of this literature has been added to the pool by the process of forward and backward searching. For instance, by reading NIST RA, we found out about Oracle, Facebook, and Amazon RAs and included those in the pool of the literature as well.
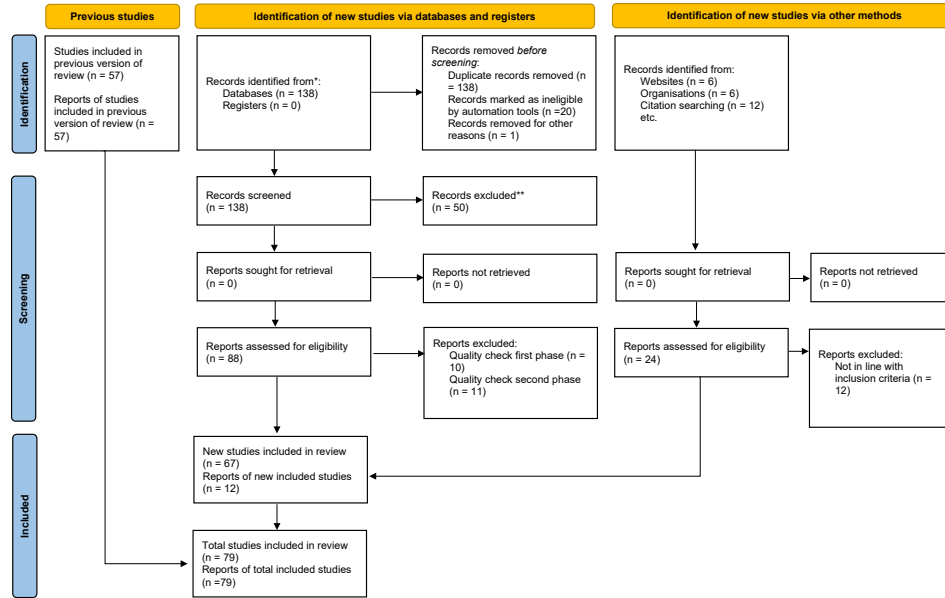
In the screening phase, the literature that were not in-line with our inclusion and exclusion criteria have been eliminated. For example, if the paper was very short and was not on the topic of BD RA, or its ecosystem or limitations, it was excluded. As a result of this phase, 50 papers excluded. In the next phase, by assessing studies against the quality framework, 21 studies from academia, and 12 studies from grey literature pool has been eliminated.

By the end of the selection and screening process, 79 papers have been pooled. The detail of this process is depicted in

5. Improvements

1. The current writing style looks like a summary description, lacks new insight on the topic. The overall contribution needs to be enhanced.

2. In general, each larger-scale system requires a more understanding of architectural components, owing largely to the complex nature of system

11

Previous studies	Identification of new studies via databases and registers		Identification of new studies via other methods

Identification

Studies included in previous version of review (n = 57)

Reports of studies included in previous version of review (n = 57)

Records identified from*:
Databases (n = 138)
Registers (n = 0)

Records removed *before screening*:
Duplicate records removed (n = 138)
Records marked as ineligible by automation tools (n =20)
Records removed for other reasons (n = 1)

Records identified from:
Websites (n = 6)
Organisations (n = 6)
Citation searching (n = 12)
etc.

Screening

Records screened (n = 138)

Records excluded** (n = 50)

Reports sought for retrieval (n = 0)

Reports not retrieved (n = 0)

Reports sought for retrieval (n = 0)

Reports not retrieved (n = 0)

Reports assessed for eligibility (n = 88)

Reports excluded:
Quality check first phase (n = 10)
Quality check second phase (n = 11)

Reports assessed for eligibility (n = 24)

Reports excluded:
Not in line with inclusion criteria (n = 12)

Included

New studies included in review (n = 67)
Reports of new included studies (n = 12)

Total studies included in review (n = 79)
Reports of total included studies (n =79)

architects. However, I cannot find a case that the authors demonstrate the uniqueness of BD systems, and the actual development challenges in BD systems.

3. The findings yielded by investigating the research questions of this SLR should constitute many discussion points around the research and practice of BD systems. However, the manuscript is completely missing a discussion section. One should expect that the results of SLR can inform the current knowledge and provide several research directions for future research.

4. Last, one of the core challenges with the paper is to situate it within an ongoing scholarly conversation. The authors currently reference a fairly diverse set of papers, but remain at a fairly abstract level when it comes to elaborating how your work builds upon and expands existing work. In turn, this makes it difficult to appreciate theoretical implications of your work.

References

[1] B. Bashari Rad, N. Akbarzadeh, P. Ataei, Y. Khakbiz, Security and privacy challenges in big data era, International Journal of Control Theory and Applications 9 (43) (2016) 437–448.

[2] P. Ataei, A. T. Litchfield, Big data reference architectures, a systematic literature review (2020).

[3] B. B. Rada, P. Ataeib, Y. Khakbizc, N. Akbarzadehd, The hype of emerging technologies: Big data as a service (2017).

[4] Databricks.
URL https://databricks.com/

[5] N. Partners, Big data and ai executive survey 2021 (2021).
URL https://www.supplychain247.com/paper/bi_data_and_ai_executive_survey_2021/pragmadik

[6] S. Computing, Bridging the gap between data and business teams (2020).
URL https://www.sigmacomputing.com/resources/data-language-barrier/

[7] B. B. Rad, P. Ataei, The big data ecosystem and its environs, International Journal of Computer Science and Network Security (IJCSNS) 17 (3) (2017) 38.

[8] I. Gorton, J. Klein, Distribution, data, deployment, STC 2015 (2015) 78.

[9] R. Cloutier, G. Muller, D. Verma, R. Nilchiani, E. Hole, M. Bone, The concept of reference architectures, Systems Engineering 13 (1) (2010) 14–27.

[10] M. Derras, L. Deruelle, J.-M. Douin, N. Levy, F. Losavio, Y. Pollet, V. Reiner, Reference architecture design: A practical approach, in: IC-SOFT, pp. 633–640.

13

[11] I. Iso, Information technology — reference architecture for service oriented architecture (soa ra) — part 1: Terminology and concepts for soa, International Organization for Standardization (2016) 51.

URL https://www.iso.org/standard/63104.html

[12] G. Muller, A reference architecture primer, Eindhoven Univ. of Techn., Eindhoven, White paper (2008).

[13] L. Bass, I. Weber, L. Zhu, DevOps: A software architect's perspective, Addison-Wesley Professional, 2015.

[14] S. Angelov, P. Grefen, D. Greefhorst, A classification of software reference architectures: Analyzing their success and effectiveness, in: 2009 Joint Working IEEE/IFIP Conference on Software Architecture & European Conference on Software Architecture, IEEE, 2009, pp. 141–150.

[15] M. J. Page, D. Moher, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al., Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews, Bmj 372 (2021).

[16] M. L. Rethlefsen, S. Kirtley, S. Waffenschmidt, A. P. Ayala, D. Moher, M. J. Page, J. B. Koffel, Prisma-s: an extension to the prisma statement for reporting literature searches in systematic reviews, Systematic reviews 10 (1) (2021) 1–19.

[17] B. A. Kitchenham, D. Budgen, P. Brereton, Evidence-based software engineering and systematic reviews, Vol. 4, CRC press, 2015.

[18] M. Borrego, M. J. Foster, J. E. Froyd, Systematic literature reviews in engineering education and other developing interdisciplinary fields, Journal of Engineering Education 103 (1) (2014) 45–76.

[19] [link].

URL https://www.jabref.org/

14

[20] K. Krippendorff, Computing krippendorff's alpha-reliability (2011).

[21] G. W. Noblit, R. D. Hare, R. D. Hare, Meta-ethnography: Synthesizing qualitative studies, Vol. 11, sage, 1988.

[22] T. Dybå, T. Dingsøyr, Empirical studies of agile software development: A systematic review, Information and software technology 50 (9-10) (2008) 833–859.

[23] M. Cumpston, T. Li, M. J. Page, J. Chandler, V. A. Welch, J. P. Higgins, J. Thomas, Updated guidance for trusted systematic reviews: a new edition of the cochrane handbook for systematic reviews of interventions, Cochrane Database Syst Rev 10 (10.1002) (2019) 14651858.

[24] [link].
URL https://casp-uk.net/casp-tools-checklists/

[25] [link].
URL https://jbi.global/critical-appraisal-tools

[26] P. Runeson, C. Andersson, T. Thelin, A. Andrews, T. Berling, What do we know about defect detection methods?[software testing], IEEE software 23 (3) (2006) 82–90.

[27] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam, J. Rosenberg, Preliminary guidelines for empirical research in software engineering, IEEE Transactions on software engineering 28 (8) (2002) 721–734.