

|  |  |  |
| --- | --- | --- |
| Research Paper | doi: 10.17705/1pais.xxxxx | Volume xx, Issue x (Year) |

**PAJAIS Submission Style Guide**

**Author Author1, Author Author2,\***

1Affiliation, Country, Email Address

2Affiliation, Country, Email Address

***Abstract***

|  |
| --- |
| *All PAJAIS manuscripts should be submitted in the style of a structured abstract with no more than 300 words. Write an informative abstract that addresses key points. The abstract should not include references. The structured abstract has four labeled sections3.*  **Background:**  **Method:**  **Results:**  **Conclusions:**  **Keywords:** Keyword1, Keyword2, Keyword3, Keyword4, Keyword5. |

Citation: Author, A., & Author, A. (Year). PAJAIS submission style guide. *Pacific Asia Journal of the Association for Information Systems, xx*(x), xx-xx. https://doi.org/10.17705/1pais.xxxxx

Copyright © Association for Information Systems.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3 Jiang, J., & Tsai, J. C. A. (2019). Constructing an Effective Abstract: Guidelines and New Standards in PAJAIS. *Pacific Asia Journal of the Association for Information Systems*, *11*(3), 1-4.

# 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 (Rada, Ataeib, Khakbizc, & Akbarzadehd, 2017). 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 (Ataei & Litchfield, 2020).

The unprecedented proliferation of data have emerged a new ecosystem of technologies; one of these technologies is big data (Mannering, Bhat, Shankar, & Abdel-Aty, 2020; Rad & Ataei, 2017). Big data is a term emerged to describe large amount of data that comes in various forms from different channels. Within the years, big data has attained a lot of attention from academia and industry, and many strive to benefit from this new material (Erevelles, Fukawa, & Swayne, 2016). Howbeit, adopting big data requires the absorption of great deal of complexity and many traditional systems cannot cope with characteristics of this domain.

Based on various reports and surveys published within the last decade, approximately 75% of big data projects have failed (AI, 2019; Gartner, 2014; Manyika et al., 2011; Nash, 2015; Partners, 2019; White, 2019). Among the challenges of adopting big data, the most frequently mentioned are 1) Architectural and system development challenges 2) Rapid technology change challenges and 3) Organizational challenges (Bashari Rad, Akbarzadeh, Ataei, & Khakbiz, 2016; Chen, Kazman, Garbajosa, & Gonzalez, 2017; Singh, Lai, Vejvar, & Cheng, 2019).

Today, majority of big data systems are designed underlying ad-hoc and complicated architectural solutions that do not favour many principles of software engineering (Gorton & Klein, 2015; Hummel, Eichelberger, Giloj, Werle, & Schmid, 2018; Nadal et al., 2017). As the systems grow bigger and new technologies are introduced, software architectures will have harder time to design the suitable solution for any given context. This is a foundation for an immature architecture that is hard to scale and maintain.

Since the approach of ad-hoc design to big data systems is undesirable and leaves many engineers in the dark, there is a need for more software engineering research for big data systems. To this end, this study presents a systematic literature review (SLR) on big data (BD) reference architectures (RAs). Conceptualisation of the system as a reference architecture, helps with understanding of the system’s key components, behaviour, composition and evolution, which in turn affect quality attributes such as maintainability, scalability and performance (Hilliard). Therefore RAs can be a good standardisation artefact and a communication medium that not only results in a concrete architecture for big data systems, but also provide stakeholders with unified elements and symbols to discuss and progress big data projects (Galster & Avgeriou, 2011) (Angelov, Grefen, & Greefhorst, 2009).

# Review Methodology

This research has been designed following the guidelines demonstrated by B. Kitchenham et al. (2009) and Shamseer et al. (2015). B. A. Kitchenham, Dyba, and Jorgensen (2004) framework is used because of its clear instructions on critically appraising evidence for impact, validity and applicability. In addition, to further increase systemacity , transparency and to prevent bias, we used the guidelines provided by Shamseer et al. (2015) on Preferred Reporting Items for Systematic Reviews and Meta-Analysis ( PRISMA ).

Of high importance is the quality of evidence collected in data gathering phase. Here, evidence is defined as the composition of quality literature. SLR has been chosen because it is qualitative research methodology that is aimed at driving knowledge and understanding about the subject matter and the elements around it. Besides, SLR provides with a transparent and reproducible procedure that is in-ling with research question and elicits patterns, relationships, trends, and delineates the overall picture of the subject (Borrego, Foster, & Froyd, 2014).   
  
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 (figure 1). In first phase, research questions are stated, literature are identified and pooled, and exclusion and inclusion criteria are defined. In second phase, literatures are assessed for their quality based on inclusion/exclusion criteria and relevance to research questions. Thirdly, selected pool of literature is coded based on research questions. Lastly, findings are synthesized, trends and patterns understood and delineated.

This SLR is based on the following research questions:

1. What are the fundamental concepts of RAs?
2. How can RAs help BD system development?
3. What are current BD RAs?
4. What are some common approaches to creating BD RAs?
5. What are the challenges of creating BD RAs?
6. What are the common architectural components of these BD RAs?
7. What are the limitations of current BD systems?

|  |
| --- |
|  |
| **Figure 1 – PRISMA flowchart of the SLR (Moher, Liberati, Tetzlaff, Altman, & The, 2009)** |

## **Identification**

In identification phase, we selected literature from the years 2010-2021. Most literature selected were within the years 2016-2021 as they provided with most recent and relevant information. Howbeit, some studies dating back to 2010 have been included to provide fundamental knowledge regarding big data systems.

Databases searched were ScienceDirect, IEEE Explore, SpringerLink, AISeL, Elsevier, MIS Quarterly 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 Scopus, Google Scholar, Web of Science, and Research Gate have been utilized.

In this phase, it becomes apparent that AISeL and Elsevier are good sources for good quality big data literature, whereas MIS Quarterly provided with the highest quality of Information Systems (IS) research.   
  
A combination of short-tail keywords and long-tail keywords based on research question has been used. These keywords are as followings:

* *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*
* *Reference Architectures in the domain of Big Data*
* *Big Data specific Reference architectures*

## Screening and Eligibility

In this phase, first we removed duplicates and then screened records for relevancy to our research questions. As a result, 15 studies excluded.

From there on, a full-text assessment took place based on our inclusion and exclusion criteria. These criteria are as following:

**Inclusion Criteria:**

* The study Includes detailed analysis, preferably in practice
* The study showed considerable case studies that aimed to explore data-business context
* Quantitative or qualitative research that illuminates on industry gaps
* Illuminates on RA concepts
* Indicates the current state of RAs in the field of BD and demonstrates possible outcomes
* Explore BD RAs extensively, and discusses the ecosystem, drivers and challenges
* Studies that are within the specified time range
* Studies that are scholarly publications, book, book chapter, thesis, white paper, dissertation, or conference proceedings

**Exclusion Criteria:**

* Studies that are not English
* Very short studies (less than 8 pages)
* Studies that do not aim to explore practice, or discuss practice related concepts
* Studies that provided with low quality information
* Studies that do not directly address the research questions
* Duplicated studies

After excluding papers based on inclusion and exclusion criteria, we assessed studies based on their quality. The quality assessment took place based on the following three factors.

* Is the study rich in terms of its relevant to practice and case studies?
* Does the study provide with ample information/data?
* Is the study discussing contemporary trends in BD RAs?

Regarding the first quality factor, richness is defined as quality and volume of information provided. Primary studies that have an international focused was one of the important facets of the quality assessment. For example, studies that followed research methodologies with good pedigrees and aimed to solve a complex problem in an actual industrial setup with a prototype that is extensively evaluated has been considered rich in terms of relevance to practice and cast studies.   
  
In regard to the second quality factor, studies that revolved around either creation of novel or exploration and examination of current RAs have been perceive as quality researched and have been included in the pool.

Lastly, any study that discussed the recent trends in BD RAs have been added to the pool. A lot of attention has been paid to research methodology and evaluation.

## Data Collection and Synthesis

In the first phase (identification) of this SLR, a total of 84 literature has been pooled. 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 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 did not either discuss a BD RA, or its ecosystem or limitations, it was excluded. As a result of this phase, 15 papers excluded.

In the next phase, we’ve applied the quality framework against the remaining literature. The clear criteria set in the framework has helped eliminating bias. In this phase, 2 further study excluded with reason.

By this stage, research questions have been set, inclusion and exclusion criteria demarcated, the quality assessment framework is developed and applied to the pool of studies, and the research embarked on actual synthesis of data. For this purpose, the software Nvivo has been utilized. Nvivo, being primarily developed for qualitative research, was used to label, code, and classify studies.   
  
In Nvivo, we defined 6 nodes for this SLR. These nodes are as followings;

1. big data reference architecture
2. big data reference architecture limitations
3. reference architecture concepts
4. big data challenges
5. big data reference architecture gaps
6. big data RA development

Consequently, all the studies were coded and classified based on the defined nodes. After coding all studies, we synthesized this coded information to induce findings and point out patterns. The synthesis took place in-line with the research questions and objectives.

## SLR Statistics

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 2013-20, and the rest to years 2010-2013. Google scholar entails all other academic databases that we found relevant literature and the white papers.

These stats are portrayed inf Figure 2.

|  |
| --- |
|  |
| **Figure 2 – SLR Statistics** |

# Findings

In this section, we map our findings against the research questions in a series of sub-sections.

## **What are the fundamental concepts of RAs?**

The proliferation of software systems and the increasing demand for solving problems with technology, have increased the complexity of man-made systems to an unprecedented level. Procedures, principles, and concepts of software architecture are increasingly applied to address the complexity faced by practitioners (Hilliard). A system abstracted and expressed in terms of architectural concepts, facilitates the understanding of system’s essence, properties revolving around it, evolution, and composition, which in turn affects quality attributes such as performance, maintainability, and scalability.

In recent years, IT architectures played a pivotal role in the progress and evolution of system development and gained acceptance in maintenance, planning, development, and cost reduction of complex systems (Martínez-Fernández et al. 2014; van Engelenburg et al. 2019).

To address ambiguity about what should be developed to address what needs, an architecture can play an overarching role by portraying the fundamental components of the system and the means and ways in which these components communicate to achieve the overall goal of the system (Sievi-Korte et al. 2019). This in turn creates manageable components that can be used to address different aspect of the problem and provides stakeholders with an abstract artefact to observe, reflect upon, contribute, and communicate with (Kohler and Specht 2019).

Many successful IT artefacts exist today that stemmed from an effective RA. A few good examples are 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). In fact, every system goes with an architecture, either known or unknown, and it is in the architecture that the overall qualities of the system are defined (Angelov et al. 2013).

Whereas there are various definitions to what constitutes an RA, they all share the same principle that the concept of patterns plays a significant role (Cloutier et al. 2010). 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 defined as generic schema that can be instantiated and configured for a particular class of systems (Derras et al. 2018).

In software engineering, RAs can be defined as an artefact that transfers software engineering knowledge as a family of solutions to a problem domain (Klein et al. 2016). In another terms, RAs are artefacts that embody domain relevant concepts and qualities, break down solutions and a create a ubiquitous language to facilitate effective communication, and illuminate various stakeholders.

Taking all into consideration, and to answer RQ1, five concepts of RA is identified; these concepts are as the following;

1. **RAs are at the highest level of abstraction:** In comparison to concrete architectures, RAs aim to capture the essence of the practice as an abstraction that portrays elements necessary for communication, standardization, implementation and maintenance (Cloutier et al. 2010). Hence, RAs aim to provide software engineering knowledge as a set of high level of architectural patterns and do not provide micro implementation details such as specific frameworks, vendors or environments
2. **RAs emphasize heavily on architectural qualities:** RAs, sitting at a higher level of abstractions are artifacts created for a wider audience and a bigger context, and are usually used by solution architects to deduce a concrete architecture in a specific environment (Angelov et al. 2008; Stricker et al. 2010). As a result, RAs pay more attention to architectural qualities and in specific quality attributes.
3. **In reference architectures, stakeholders are not clearly defined:** Stakeholders are usually people of the same company involved in the actual design and implementation of the system and do get involved in the product creation in various phases. Different stakeholders have different concerns and are crucial to the creation of the overall product (Geerdink 2013). A stakeholder c an be a developer, a designer, a product owner, a data scientist or a business analyst. Notwithstanding, due to the generic nature of the RAs, it is not feasible to indicate all stakeholders a priori. In addition, RAs are at a higher level of abstraction and tend to provide solutions for a class of problems and not a specific context. Therefore, defining and introducing stakeholders into RAs can potentially decrease their effectiveness (Chang and Boyd 2018)(Ataei & Litchfield, 2020) .
4. **RAs promote adherence to common standards:** The design of an RA is usually guided by existing architectural patterns, common pitfalls in practice, system development and literature. For this reason, RAs convey standard approaches and patterns that avoid known pitfall, facilitate reuse, and decreased complexity (Ataei & Litchfield, 2020).
5. **RAs are effective artefacts for system development and communication:** RAs are powerful artefacts that can are used by architects that design, manage, and utilize complex system to improve communication, enabling them to work in an coherent, integrated fashion (Hilliard). RAs are created as assets that codify the best practice and conventions of the industry and often include architectural descriptions and standards.

## How can RAs help BD system development?

Despite the high failure rate of BD projects, IT giants such as Google or Amazon have developed exclusive BD systems with complicated data pipelines, data management, procurement and batch and real-time analysis capabilities (Kohler and Specht 2019). Having the resources required, these companies attract the best of talent from around the globe to manage the complexity involved in development of big data systems. Notwithstanding, that’s not the reality of majority of organizations that are trying to benefit from big data analytics.

Big data systems sail away from the traditional small data analytics paradigms and bring various challenges including rapid technology change challenges (Chen et al. 2016), system development and architecture challenges (Jagadish et al. 2014) and organizational challenges (Rada et al. 2017; Vassakis et al. 2018). BD does not only mean ‘big’ amount of data, or just volume; other characteristics of BD such as velocity, variety, veracity and variability bring significant challenges on the table. Although these challenges do not only belong to domain of BD systems, BD exacerbates these challenges because of the following reasons;

1. Distributed scaling is required to address batch and stream processing demands
2. There is a need for real near-time performance (stream processing)
3. Complex technology orchestration is required to create effective communication channels between components and data flow
4. Continuous delivery is required to continually disseminate patterns and insights into various business domains
5. Two different approaches are required for data processing, stream and batch processing; or fast and delayed processing (Jagadish et al. 2014).

To provide a solution to these challenges, one has to realize the core fundamentals of BD systems. Academic and practitioners of BD, describe BD as an interplay of methodology (workflow, organization), software engineering (data engineering, storage, etc.), and analysis (math, statistics) (Akhtar et al. 2019; Rad and Ataei 2017). Therefore, one can deduce that technology orchestration is a focal in BD system development and maintenance.

Positioned on top of this rationale, and to address RQ2, RAs can be considered an effective artefact that help with component delineation, interface definition, technology orchestration, variability management, scalability, and maintenance of BD systems (Chang and Boyd 2018; Nadal et al. 2017). The purpose of RAs is to create an integrated environment in which fragmented processes around the system are optimized, responsiveness to change is assured, and delivery of architectural strategies is supported.

Most authors and practitioners agree that issues around BD software engineering and system development are severe and that this justifies the use of RAs for BD systems (Chang and Boyd 2018). Starting with a grounded RA means that the software architect can refer to an already designed orchestration of components, interfaces, inter-communications, and variability points and map them against the organization’s capability framework, desired quality attributes, and business drivers and vision. This also means that the software architecture or the software architecture group is no longer challenged to model a new architecture from an array of independent components that needs to be assembled through effective interface, cache mechanisms, storage, etc.

On that account, RA is an artefact that facilitates development and homogenization of BD systems. Using RA to address complex problems have been successfully applied for Database Management Systems (DBMS) (Piñeiro et al. 2019) and Distributed Database Management Systems (DDBMS) (Rahimi and Haug 2010).

## **What are current BD RAs?**

As a result of this SLR and to answer RQ3, 23 BD RA has been found, among which 23 RAS are from academia, 4 from practice, and one through the collaboration of academia and practice. These are described further in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1 – BD RAs** | | | |
| **No.** | **Reference Architecture** | **Domain** | **Year** |
| 1 | Towards a big Data reference architecture (Maier et al. 2013) | Academia | 2013 |
| 2 | A reference architecture for Big Data solutions introducing a model to perform predictive analytics using Big Data technology (Geerdink 2013) | Academia | 2013 |
| 3 | a proposal for a reference architecture for long-term archiving, preservation, and retrieval of Big Data (Viana and Sato 2014) | Academia | 2014 |
| 4 | Questioning the Lambda architecture; Kappa Architecture (Kreps 2014) | Academia | 2014 |
| 5 | Big data architecture framework (Demchenko et al. 2014) | Academia | 2014 |
| 6 | Big Data driven e-commerce architecture (Ghandour 2015) | Academia | 2015 |
| 7 | The solid architecture for real-time management of big semantic data; Solid architecture (Martínez-Prieto et al. 2015) | Academia | 2015 |
| 8 | The Big Data research reference architecture (Pääkkönen and Pakkala 2015) | Academia | 2015 |
| 9 | A Reference Architecture for Big Data Systems (Sang et al. 2016) | Academia | 2016 |
| 10 | A reference architecture for Big Data systems in the national security domain (Klein et al. 2016) | Academia | 2016 |
| 11 | A Reference Architecture for Supporting Secure Big Data Analytics over Cloud-Enabled Relational Databases (Cuzzocrea 2016) | Academia | 2016 |
| 12 | Managing Cloud-Based Big Data Platforms: A Reference Architecture and Cost Perspective (Heilig and Voß 2017) | Academia | 2017 |
| 13 | Scalable data store and analytic platform for real-time monitoring of data-intensive scientific infrastructure (Suthakar 2017) | Academia | 2017 |
| 14 | A software reference architecture for semantic-aware Big Data systems; Bolster Architecture (Nadal et al. 2017) | Academia | 2017 |
| 15 | Towards a secure, distributed, and reliable cloud-based reference architecture for Big Data in smart cities (Kohler and Specht 2019) | Academia | 2019 |
| 16 | Reference Architectures and Standards for the Internet of Things and Big Data in Smart Manufacturing (Ünal 2019) | Academia | 2019 |
| 17 | Lambda architecture (Marz 2011) | Practice | 2011 |
| 18 | IBM - Reference architecture for high performance analytics in healthcare and life science (Mysore et al. 2013) | Practice | 2013 |
| 20 | Microsoft - Big Data ecosystem reference architecture (Levin, 2013) | Practice | 2013 |
| 21 | Oracle - Information Management and Big Data: A Reference Architecture (Oracle 2014) | Practice | 2014 |
| 22 | SAP - NEC Reference Architecture for SAP HANA & Hadoop (SAP 2016) | Practice | 2016 |
| 23 | NIST Big Data interoperability framework (Chang and Boyd 2018) | Hybrid | 2018 |

Within the past 10 years, there has been a considerable attention to the BD domain, and in specific BD system development (Li et al. 2019). For instance, in March 2012, White House announced an initiative for BD research and development (House, 2012). The goal of this initiative was to accelerate the speed of science and engineering discovery, to improve national security, and to improve the knowledge extraction from large and complicated sets of data (Chang & Grady, 2015). This project has been supported by six federal departments and has been given more than $200 million USD with the goal of substantial progress in the tools and techniques to handle big data (Chang & Grady, 2015).

A year later, in June 2013, National Institute of Standards and Technology (NIST) Big Data Public Working Group (NBD-PWG) was launched with considerable participation from across the nation. Practitioners, researchers, agents, government representatives, and none-profit organizations joined in this momentum (Chang & Grady, 2015). One of the results of this project was NIST Big Data Reference Architecture (NBDRA). According to US Department of Defense, one of the main objectives of NBDRA was to provide with an authoritative source of information on big data that restraint and guides the overall practice (Chang & Boyd, 2018). This is arguably one of the most comprehensive and recent RAs available on the fields of big data. NBDRA is made up of two fabrics encompassing five functional logical components connected by various interfaces, representing intertwined nature of security and privacy and management (Bashari Rad et al., 2016).

Along the lines, other giant IT vendors published their own RAs for big data. In this SLR, 5 BD RA has been collected from the practice, and mostly through white papers. Among these RAs, arguably Lambda architecture is the most prominent. It is also worth mentioning that there has been other BD RAs found in practice, but they were rather too short or did not reflect the contemporary state of BD analytics and has been eliminated as described in the research methodology section.

In the realm of academia, there has been numerous efforts including a postgraduate master’s dissertation for creating big data RAs. In addition, few universities have published their own RA. For instance, university of Amsterdam published the BD architecture framework (Chang & Mishra, 2015), and a postgraduate student published a BD RA at the university of Eindhoven (Maier, Serebrenik, & Vanderfeesten, 2013).

Last but not least, there has been numerous reference architectures developed recently for specific domains. These studies have been usually published as short journal papers, and many have promised future publication of the full reference architecture as a book. For instance, Klein, Buglak, Blockow, Wuttke, and Cooper (2016) developed a BD Ra in the national security domain, and Weyrich and Ebert (2015) worked on a BD RA in the domain of internet of things (IOT).

Through the process of literature review for this SLR, scarcity of big data reference architectures has been witnessed. The studies listed above are prominent research, with great potential to induce concrete architectures. But with all, they are mostly published as short journals and provide with little information about architectural qualities, metadata management, and security, privacy concerns. In another terms, they are notion or brief discussions on reference architectures in very particular domains.

## **What are some common approaches to creating BD RAs?**

The findings gained from this study led to the understanding that there are not many frameworks available for design and development of RAs. Nevertheless, to address RQ4, we sought to find the research methodology and approaches chosen to develop RAs. One of the most commonly used approaches to developing RAs is ‘Empirically grounded Reference Architectures’ by Galster and Avgeriou (2011). The research methodology is well-received because of its emphasis on empirical validity and empirical foundation. This methodology is comprising of 6 step process which are respectively 1) Selecting the type of the RA, 2) Selection of the design strategy, 3) Empirical acquisition of data, 4) Construction of the RA, 5) Enabling RA with variability, 6) Evaluation of the RA.

Another seminal work in this area is a framework for analysis and design of software RAs created by (Angelov, Grefen, & Greefhorst, 2012). The framework utilizes a e multi-dimensional space for the classification of RAs and as a result presents 5 major types. It is developed with the objective of supporting analysis of RAs with regards to their architectural specification/design, goal, and context. This is achieved through three major dimensions, each having their own corresponding subdimensions of design, goal, and context. These dimensions and sub-dimensions are derived by interrogatives of ‘why’, ‘where’, ‘who’, ‘when’, ‘what’, and ‘how’, which is a well-established practice for problem analysis. The interrogative why addresses the goal of the RA, who, when, where address the context, and how and what address the design dimensions (Galster and Avgeriou 2011). This framework categorizes RAs in two major groups: facilitation RAs and standardization RAs.

Volk et al. (2019) utilized Software Architecture Comparison Analysis Method (SCAM) to compare and examine RAs based on their applicability. This result of this work was a decision-support process for selection of BD RAs.

Two standards that have been observed the most were ISO/IEC 25010 for choosing quality software products for RAs (Iso 2011), and ISO/IEC 42010 for architecture description.

Surprisingly, based on the evidence gained from this SLR, most researchers and practitioners use informal architectural description methods like boxes and lines, except for the works of Geerdink (2013). In this study, the author used Archimate (Josey, Lankhorst, Band, Jonkers, & Quartel, 2016) as the modeling language which is a formal and standard modeling language that is accepted and recommended in ISO/IEC 42010 as well. Informal methods of modeling can introduce inconsistency issues between system design and implementation of the system (Zhu, 2005), do not adhere to a well-established standard and do not promote the development of modeling approaches. Therefore, on can argue that there is a need for more emphasis on the modeling language with which different researchers and practitioners describe ontologies.

Lastly, Hevner's information systems research framework (Hevner, March, Park, & Ram, 2004) has been used for the development of RA presented by Geerdink (2013), which is a suitable research design, since a BD RA is an information system artefact based on existing literature and business needs.

## **What are the challenges of creating BD RAs?**

Among the challenges of developing RAs, perhaps evaluation is the most significant (Maier et al. 2013) (Cioroaica et al. 2019). Two fundamental pillars of the evaluation is the correctness and the utility of the RA and how efficient it can be adapted and instantiated (Galster & Avgeriou, 2011). RAs and concrete architectures come with a different level of abstraction and have divergent qualities.

Whereas there are many well-established evaluation methods for concrete architectures such as Architecture Level Modifiability Analysis (ALMA), Scenario-based Architecture Analysis Method (SAAM) (Kazman et al. 1996), (Bengtsson et al. 2004), Architecture Trade-off Analysis Method (ATAM) (Kazman et al. 1998) and Performance Assessment of Software Architecture (PASA) (Williams and Smith 2002), none of the these can really be directly applied to RAs.

For instance, ATAM is reliant on participation of stakeholders in early stages for creation of utility tree, and RAs, being highly abstract, do not have a clear group of stakeholders at that stage. In addition, many of evaluation methodologies listed make use of scenarios, whereas RAs are highly abstract and are potentially adopted for various contexts, therefore making scenario creation a difficult and sometimes invalid. Either a few general scenarios is developed to cover all aspects, or a large number of specific scenarios are developed to cover various aspects of the RA.   
  
Based on three problems discussed above, available methods of architecture analysis are not sufficient in evaluating the RA. This has been addressed by various researchers in the industry. In one study Angelov, Trienekens, and Grefen (2008), modified ATAM and extended it to resonate well with RAs. This process took place by invitation of representatives from leading industries for the evaluation process, and the selection of various contexts and defined scenarios for these contexts.

Furthermore, ATAM has been extended to evaluate completeness, buildability and applicability. Howbeit the selection of the right candidate and involving them in the process is a time-consuming and daunting task (Angelov et al., 2008). In Another study by Maier et al. (2013) as a postgraduate thesis in Eindhoven University of Technology, the evaluation of the RA has been conducted by mapping it against existing concrete architectures described in industrial whitepapers and reports. Along the lines, Galster and Avgeriou (2011) suggested reference implementations, prototyping and incremental approach for the validation of the RA.

By the virtue of the findings from this SLR, and by studying the approaches from Bosch (2000), Avgeriou (2003), and Derras et al. (2018), an evaluation framework for a RA can be done through architectural prototype evaluation, which means a concrete architecture of the RA is generated and then evaluated through a well-grounded method such as ATAM.

## **What are the common architectural components of these BD RAs?**

Some of the RAs collected were in in the form of a short paper and provided with not much detail, whereas some of the other such as NIST were quite comprehensive. Majority of RAs have been inspired or based on other RAs, and this signified the notion that “RAs can be perceived more effective when they are created out of available knowledge, studied domain, and existing RAs rather than from scratch”. To address RQ5, RAs listed in table 1 was reviewed and compared to deduce common architectural components of BD RAs.

We describe these architectural components as three major categories, namely ‘BD management and storage’, ‘BD analytics and application interfaces’, and ‘

### BD Management and Storage

Storage involves variety of different database, or sometimes the practice of polyglot persistence (Khine & Wang, 2019), as data come in various formats in BD systems. For instance, when it comes to dynamic data, NoSQL databases such as MongoDB is a suitable choice because of their non-tabular nature (Banker, Garrett, Bakkum, & Verch, 2016), and when there is a need for complex relationship between entities, graph databases such as Neo4J are more suitable because of their tree traversal performance (Van Bruggen, 2014).

In the same vein, when there is a requirement for a large dataset that can be distributed across multiple database nodes, column wide databases such as Cassandra can be a good architectural component (Carpenter & Hewitt, 2020), and when there is a need for text analysis, one can choose ElasticSearch (Tsaousi, 2021). Choosing the right database or databases, is an important architectural decision that the architect makes. This can also include patterns for data access, storage and caching, and highlights the importance of RAs even more. For example, the practitioners of distributed system that are specialized in micro-services architecture do make user of Command Query Responsibility Segregation (CQRS) pattern for high performance applications (Márquez & Astudillo, 2018).

Another architectural component that is popular in BD systems is data lake. Data lake can be perceived as an ingestion framework that can be given various types of data including internal and external data. The data stored in the data lake can then be stored for cleansing, preparation and modeling through data pipelines (Sawadogo & Darmont, 2021).

Similar to the way that Business Intelligence (BI) and BD differ in their source data types both in terms of granularity and data structure of it, a data lake and data warehouse share same of those same differences. In the case of a data warehouse, a relational database is utilized which decreases flexibility when it comes to analysis. In the case of data lake, data of different kind can be stored without the engineer needing to define the schema in advance. This increases the flexibility.  
  
Nevertheless, this has its own downside and can be abused by data engineers. One can just throws different data sets without much regard at all for how they’re structured, which leads to what people refer to as data swamp (Khine & Wang, 2018). Data governance can alleviate some of these issues which is a kind of an architectural work (Josey, 2016).

Every BD RA analyzed provided the notion of data pipelines, even when the exact phrase wasn’t used. This is due to the fact that in BD systems, data usually goes through several phases before it’s ready for statistical analysis (Geerdink, 2013).

### BD Analytics and Application Interfaces

The result of this SLR shed lights on two major data processing approaches that a big data architecture encompasses. These processes are firstly stream processing (fast processing) and batch processing.

**Document Layout**

***Page Layout***

* A4 size
* 1” (2.54cm) from top and bottom margins
* 1” (2.54cm) from left and right margins

\* All items such as paragraph, table, and figure should fit in this width. Otherwise, use horizontal orientation

***Page Alignment***

* Title and authorship: Centered
* Main text: justified

***Type Font, Spacing, and Indent***

* **Title:** Arial 16, Bold
* **Abstract:** heading: Arial 14, Bold, Italic; text: Arial 11, Italic
* **Keywords:** text: Arial 11, three to five keywords
* **Heading 1** (e.g., introduction): Arial 14, Bold (\* Headings do not have roman numbers)
* **Heading 2:** Arial 12, Bold, Italic (\* Headings do not have roman numbers)
* **Heading 3:** Arial 11, Bold (\* Headings do not have roman numbers)
* Line spacing: single
* All other text: Arial 11
* No indent: additional one-line spacing separates paragraphs

|  |
| --- |
| **Heading 1**  ***Heading 2***  **Heading 3** |

***Page Number***

* Do not insert the page number.

***Document Format***

Submitted file should be in Microsoft Word format.

**Manuscript Material Ordering**

1. Title
2. Abstract
3. Keywords
4. Main Body

Typically beginning with introduction and ending with conclusions; Section numbering not allowed. Suggested reading: Getting Published in PAJAIS: A Practical Guide from the Perspectives of Editors (Jiang & Tsai, 2019).

1. Acknowledgments\*
2. References
3. Appendices
4. About the Authors\*

\*The author information, acknowledgment, and other information about manuscript should be filled out in the title page of online manuscript systems during review process. PAJAIS logo, header, footer is not allowed to use when the manuscript is under review.

***Elements Style***

1. Footnotes

Footnotes should be avoided if possible. If they are absolutely necessary, footnotes are referred to by superscript number, and displayed on the same page as referred in text; font is Arial 10, single-spaced.

1. In-Text Reference

In-text reference should be referred to in text within parentheses, as follows.

* + - Single author - Fichman (2004)… ; …(Fichman, 2004).
    - Two authors - Lyytinen and King (2004)… ; …(Lyytinen & King, 2004).
    - More than two authors, including first citation - Rossi et al. (2004)… ; (Rossi et al., 2004).

1. Figures

Figures should be inserted at the end of the paragraph in which they are first referred. If there is not sufficient space for full display, it can appear on the next page.

* Figure should be boxed.
* Figure should be aligned center (both vertically and horizontally).
* The font of entries is Arial 10 point
* Figure titles should be at the bottom of the figure box.
* Figure titles should be horizontally left aligned, vertically center aligned; white font, Arial 11 point, bold on a black box; the height of the black box for figure title is 0.7cm” for single line title, and 1.2cm” for double line title.
* Notes should be added at bottom of figure box in Arial 9 point.

|  |
| --- |
|  |
| **Figure 1 - Example of Figure Appearance** |

1. Tables

Tables should be inserted at the end of the paragraph in which they are first referred. If there is not sufficient space for full display, it can appear on the next page.

* Table should be boxed.
* Table should be aligned center (both vertically and horizontally), but the entry alignments are at authors’ disposal.
* The font of entries is Arial 10 point
* Table titles should be at the head of tables.
* Table titles should be horizontally left aligned, vertically center aligned; white font, Arial 11 point, bold on a black box. The height of the black box for table title is 0.7cm for single line title, and 1.2cm for double line title.
* Notes should be added (1) at bottom of table in Arial 9 point or (2) in the last raw of table without indent.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1 - Example of Table Appearance** | | | |
|  |  | Mean | St.dev |
| TEL | 1985 | -- | -- |
| 1991 | 15.9 | 14.0 |
| 2001 | 17.7 | 11.5 |
| CALL | 1985 | -- | -- |
| 1991 | 0.1 | 0.06 |
| 2001 | 0.1 | 0.09 |
| URBAN | 1985 | 64.8 | 20.3 |
| 1991 | 67.1 | 19.6 |
| 2001 | 69.9 | 18.7 |
| TEL = Average monthly telephone subscription cost;  CALL = Average cost of local call;  URBAN = Size of urban population, as a percentage of total population; | | | |

Notes: Year dummies are included in all regressions. Standard errors are in parentheses. \*p<0.1. \*\*p<0.05.\*\*\*p<0.01

**References**

All references should follow the American Psychological Association (2010) guide. The only major difference between APA and PAJAIS style does not use Digital Object Identifiers (DOIs).

* **Journal Papers:**

Hevner, A., March, S., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75-105.

* **Magazine Articles:**

Ackerman (2007). The perfect blend. *Consumer Goods Technology Magazine*. Retrieved from http://consumergoods.edgl.com/media/publicationsarticle/art-mar07-6.pdf

* **Books:**

Alvesson, M., & Deetz, S. (2000). *Doing critical management research*. Sage Publications: Thousand Oaks, California.

* **Book Chapters:**

Keren, G. (1990). Cognitive aids and debiasing methods: Can cognitive pills cure cognitive ills? In J. Caverni, J. Fabre, & M. Gonzalez (Eds.), *Cognitive biases* (pp. 523-555). Amsterdam, North Holland: Elsevier.

* **Conference:**

Forget, A., Chiasson, S., & Biddle, R. (2007). Persuasian as education for computer security. In T. Bastiaens & S. Carliner (Eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007* (pp. 882-82). Chesapeake, VA: AACE.

* **Working Papers:**

Acha, V. (2004). *Technology frames: The art of perspective and interpretation in strategy* (SPRU Electronic Working Paper No. 109). Brighton: University of Sussex.

AI, V. (2019). Why do 87% of data science projects never make it into production? Retrieved from <https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/>

Angelov, S., Grefen, P., & Greefhorst, D. (2009). *A classification of software reference architectures: Analyzing their success and effectiveness.* Paper presented at the 2009 Joint Working IEEE/IFIP Conference on Software Architecture & European Conference on Software Architecture.

Angelov, S., Grefen, P., & Greefhorst, D. (2012). A framework for analysis and design of software reference architectures. *Information and software technology, 54*(4), 417-431.

Angelov, S., Trienekens, J. J., & Grefen, P. (2008). *Towards a method for the evaluation of reference architectures: Experiences from a case.* Paper presented at the European Conference on Software Architecture.

Ataei, P., & Litchfield, A. T. (2020). Big Data Reference Architectures, a systematic literature review.

Avgeriou, P. (2003). Describing, instantiating and evaluating a reference architecture: A case study. *Enterprise Architecture Journal, 342*, 1-24.

Banker, K., Garrett, D., Bakkum, P., & Verch, S. (2016). *MongoDB in Action: Covers MongoDB version 3.0*: Simon and Schuster.

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

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

Bosch, J. (2000). *Design and use of software architectures: adopting and evolving a product-line approach*: Pearson Education.

Carpenter, J., & Hewitt, E. (2020). *Cassandra: the definitive guide: distributed data at web scale*: O'Reilly Media.

Chang, W. L., & Boyd, D. (2018). *NIST Big Data Interoperability Framework: Volume 6, Big Data Reference Architecture*. Retrieved from

Chang, W. L., & Grady, N. (2015). *NIST Big Data Interoperability Framework: Volume 1, Big Data Definitions*. Retrieved from

Chang, W. L., & Mishra, S. (2015). *NIST Big Data Interoperability Framework: Volume 5, Architectures White Paper Survey*. Retrieved from

Chen, H.-M., Kazman, R., Garbajosa, J., & Gonzalez, E. (2017). *Big Data Value Engineering for Business Model Innovation.* Paper presented at the Proceedings of the 50th Hawaii International Conference on System Sciences.

Derras, M., Deruelle, L., Douin, J.-M., Levy, N., Losavio, F., Pollet, Y., & Reiner, V. (2018). *Reference Architecture Design: A Practical Approach.* Paper presented at the ICSOFT.

Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. *Journal of Business Research, 69*(2), 897-904.

Galster, M., & Avgeriou, P. (2011). *Empirically-grounded reference architectures: a proposal.* Paper presented at the Proceedings of the joint ACM SIGSOFT conference--QoSA and ACM SIGSOFT symposium--ISARCS on Quality of software architectures--QoSA and architecting critical systems--ISARCS.

Gartner. (2014). Survey Analysis: Big Data Investment Grows but Deployments Remain Scarce in 2014. Retrieved 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.* Paper presented at the 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013).

Gorton, I., & Klein, J. (2015). Distribution, Data, Deployment. *STC 2015*, 78.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75-105.

Hilliard, R. ISO/IEC/IEEE 42010. In.

House, W. (2012). Big Data is a Big Deal. Retrieved from <https://obamawhitehouse.archives.gov/blog/2012/03/29/big-data-big-deal>

Hummel, O., Eichelberger, H., Giloj, A., Werle, D., & Schmid, K. (2018). *A collection of software engineering challenges for big data system development.* Paper presented at the 2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA).

Josey, A. (2016). *TOGAF® Version 9.1-A Pocket Guide*: Van Haren.

Josey, A., Lankhorst, M., Band, I., Jonkers, H., & Quartel, D. (2016). An introduction to the ArchiMate® 3.0 specification. *White Paper from The Open Group*.

Khine, P. P., & Wang, Z. (2019). A review of polyglot persistence in the big data world. *Information, 10*(4), 141.

Khine, P. P., & Wang, Z. S. (2018). *Data lake: a new ideology in big data era.* Paper presented at the ITM web of conferences.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology, 51*(1), 7-15.

Kitchenham, B. A., Dyba, T., & Jorgensen, M. (2004). *Evidence-based software engineering.* Paper presented at the Proceedings of the 26th international conference on software engineering.

Klein, J., Buglak, R., Blockow, D., Wuttke, T., & Cooper, B. (2016). *A reference architecture for big data systems in the national security domain.* Paper presented at the 2016 IEEE/ACM 2nd International Workshop on Big Data Software Engineering (BIGDSE).

Levin, B. (2013). Big data ecosystem reference architecture. *Microsoft Corporation*.

Maier, M., Serebrenik, A., & Vanderfeesten, I. (2013). Towards a big data reference architecture. *University of Eindhoven*.

Mannering, F., Bhat, C. R., Shankar, V., & Abdel-Aty, M. (2020). Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic methods in accident research, 25*, 100113.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

Márquez, G., & Astudillo, H. (2018). *Actual use of architectural patterns in microservices-based open source projects.* Paper presented at the 2018 25th Asia-Pacific Software Engineering Conference (APSEC).

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The, P. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine, 6*(7), e1000097. doi:10.1371/journal.pmed.1000097

Nadal, S., Herrero, V., Romero, O., Abelló, A., Franch, X., Vansummeren, S., & Valerio, D. (2017). A software reference architecture for semantic-aware Big Data systems. *Information and software technology, 90*, 75-92.

Nash, H. (2015). CIO SURVEY 2015. *Association with KPMG*.

Partners, N. (2019). Big Data and AI Executive Survey 2019. *Data and Innovation*.

Rad, B. B., & Ataei, P. (2017). The big data Ecosystem and its Environs. *International Journal of Computer Science and Network Security (IJCSNS), 17*(3), 38.

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

Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems, 56*(1), 97-120.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., . . . Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *Bmj, 349*.

Singh, N., Lai, K.-H., Vejvar, M., & Cheng, T. (2019). Big data technology: Challenges, prospects and realities. *IEEE Engineering Management Review*.

Tsaousi, K. D. (2021). Elasticity of Elasticsearch. In.

Van Bruggen, R. (2014). *Learning Neo4j*: Packt Publishing Ltd.

Weyrich, M., & Ebert, C. (2015). Reference architectures for the internet of things. *IEEE software, 33*(1), 112-116.

White, A. (2019). Top Data and Analytics Predicts for 2019.