International Journal of Advanced Computer Research, Vol 14(69)
ISSN (Print): 2249-7277   ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2023.1362047

Research Article

A business-driven reference architecture for big data analytics implementation by public sector organizations: a case study of Uganda

Matendo Didas*, Frederick Henri Chali and Noe Elisa
College of Information and Virtual Education (CIVE), The University of Dodoma, Dodoma, Tanzania

Received: 22-March-2024; Revised: 01-November-2024; Accepted: 09-November-2024

Abstract
Advanced data analytics especially big data analytics (BDA) has proved to have a great potential to improve public service delivery both in urban and rural or remote areas. However, many of the BDA implementation projects in public sector organizations are failing. Early efforts to the successful implementation of BDA, suggest that business-driven reference architecture (RA) for these projects is among the viable solutions to have BDA project failure rate addressed in public sector organizations. The current study aimed at designing business-driven RA for big data analytics implementation (BRABDAI) to act as a frame of reference (blueprint) for delivering big data analytics implementation projects in public sector organizations. Uganda as a case study, three public organizations were considered (Ministry of Health, Ministry of Education and Sports, and Uganda Bureau of Statics). To accomplish the study's goals, the design science research methodology (DSRM) was used supported by a questionnaire survey and mini literature review for data collection. The developed RA offers a mechanism to systematically consider business use cases and processes in the first place before assembling BDA technical aspects and procedures. Hence, the main result of this paper is BRABDAI for public sector organizations. The practical implication of the paper is that feasibility study, requirement engineering, design, implementation, deployments, and maintenance of big data analytics projects in public sector organizations can now be driven by the need of the organization in question, other than assembling technology first which can result into detrimental misalignment between organizational vision and technology.

Keywords
Big data analytics, Reference architecture, Design science research methodology , Business-driven reference architecture.

1.Introduction

Regardless of big data analytics (BDA) technologies being widely spread, there is a high failure rate associated with BDA implementation projects in public sector organizations. In addition to the huge capital invested in these projects, they fail to provide the anticipated outcomes in public sector organizations [1, 2]. Even if BDA can significantly improve the operations of public sector organizations, BDA implementation projects have a record of failing [3–5]. According to several studies [6–8], many public sector organizations have not even started working on BDA programs that get past the proof of concept (PoC) stage. Organizations in the public sector are aware of the limitless prospects offered by BDA, but they don't seem to know how to go about putting these cutting-edge data technologies into practice [9].

One of the prominent reasons for the poor performance of these projects is the implementation methodology adopted [10]. The reference architectures (RAs) in existence such as Azure, Google Cloud reference architecture (RA), and Nippon Telegraph and Telephone (NTT)Data among others do not suit the mandates, business use cases, missions, and values concerning a particular BDA tool as well as business limitations associated with the selected BDA technology in public sector organizations. They tend to pivot toward their services; their metadata component is ill-discussed among other weak links associated with them. This study selected a different approach where business processes need to be identified and analyzed first and thereafter, the big data (BD) assets (BD sources) and technologies are defined to serve the given business process and purpose.

*Author for correspondence

Furthermore, these RAs are broad and generalized, and public sector organizations are unique in their own way. For example, the majority of use cases/business requirements used to inspire and create the existing BDA RA are found in web-based business models or at the very least in the business domain. They don't take into account use cases from the scientific, academic, or medical fields. This issue may be resolved by incorporating elements of BDA RA into domain-specific RAs or by designing BDA RAs for these domains that are not linear [11]. Generic designs for BDA implementation also only aid in identifying essential technological elements, and hence, in making technology decisions. Domain-specific specifications and issues are less reusable, have a lower level of abstraction, and require less work to instantiate since they are more anchored in reality. When comparing the approaches of Klein [12] and the national institute of science and technology (NIST) [13], the distinction between technical-oriented RAs and those that are more domain-oriented becomes clear. While the former is domain-specific for the security domain, the latter is domain-agnostic and focuses on technological components [14], among others. The benefit of the current study's product-independent and business-driven RA would be that public sector organizations could choose the parts first, business processes, and then the corresponding products in a flexible way without risking costly vendor lock-in.

The study was guided by the following Research Question (RQ) and Research Objective (RO).

RQ-What suitable business-driven RA is needed for BDA project implementation in public sector organizations?

RO-To develop business-driven RA for BDA implementation to utilize BD public sector organizations, taking Uganda as a Case study.

Geographically, Uganda was chosen as a case study and the RA incorporates aspects mainly from 3 public sector organizations: the ministry of health (MOH), the Uganda bureau of statistics (UBOS), and the ministry of education and sports (MoES). These organizations are known to be the most data producers in Uganda, which require an innovative approach to deliver advanced data analytics [15].

The rest of the paper is arranged as follows: section 2 provides brief relevant literature on the subject. Section 3 describes the methodology of the study, followed by the results and discussion in section 4. Conclusions, limitations, and future work suggestions are provided in section 5.

2.Literature review

The concept of BD and BDA

BD and associated analytics technologies take the first position as chief causes of the whole continuum of technology societies ranging from data society to information society to knowledge society [16]. BDA technologies and their potential advantages have exhibited a large current consideration. It is also acknowledged that BDA is the chief factor for the technological movement of enormous modern technologies for instance Fourth Industrial Revolution Technologies (4thIRTs) which include artificial intelligence (AI), internet of things (IoTs), block chain technology (BCT), assistance driving systems (ADSs), and additive manufacturing (AM) among others. The nexus of 4thIRTs and technology societies with BDA is that both trends need large-scale data for their accurate results and models to aid the decision-making process and intended purpose.

A glance at what BD means is given by the Gartner research firm; the firm defines BD as a dataset with high Volume, high velocity, and high variety [17]. The firm further adds that the capability of managing BD assets demands less cost-advanced technologies for processing and analyzing these assets to properly justify critical aspects of these technology projects. These aspects include but are not limited to associated huge capital investment, insight derivation sufficiency, informed decision-making process with sufficient accuracy, and process automation requirements among others. It can be deduced from this glimpse that BD is nothing other than data assets produced in massive capacities with several presentations from multiple sources in a fast fast-paced time landscape. Processing and examining these massive capacities of data is what is termed BDA. The potential prospects and benefits of BDA range from traditional (conventional) to new market trends in terms of improved public service delivery through the generation of accurate and proper indicators in various domains such as education (enabling customized learning), food safety (disaster prediction and management), and health (providing personalized medicine) among others [18–20]. Given the business-driven approach to the utilization of BA and its analytics, the public sector is bound to advance on huge prospects and benefits to serve society better [21]. With the above preamble, there is still a question regarding which approach public sector organizations can use to implement BDA to utilize ever-increasing data to better serve the citizenry. This question is what the research intends to answer by developing a business-oriented RA for the said purpose.

Business-driven implementation of BDA implementation projects and Trends

As stated earlier, business-driven implementation of BDA is an enabler towards achieving successful implementation of BDA projects for better public service delivery because it makes use of ever-increasing data for the right purpose/business use case [22]. Apart from its importance, the high failure rate of BDA implementation projects are still among the challenges facing public sector organizations for service delivery in the world today. But how does this challenge emerge? Some of the causes of this failure identified by different researchers in the public sector include the following: Mismatch among business processes & BDA, Inadequate practices of data governance, quality and value control, integration and transformation complexities, security issues security, lack of innovative and business-driven approach for BD, poor information communication and technology (ICT) Infrastructure among others [23–26].

By 2018, the rise of visual discovery tools had halted, and product commoditization had set in. By 2020, new cloud pricing models catered to specialized analytics workloads, which contributed to a five-fold rise in cloud versus on-premises analytics spending growth. By 2026, G2000 businesses will have policies in place to prevent unintended consequences of cognitive systems, such as noncompliance and ethical quandaries [27]. The volume of data in the world is growing at a 40% yearly rate. With the advent of the IoT, AI, frameworks, and other 4.0 industrial innovations, one of the main BDA Trends of 2018 and beyond is the integration of multiple new sources of BD and its analytical technologies into the Data Management environment with the hope to solve BDA implementation challenges.

Benefits of BD analytics in public sector organizations

So many advantages of BDA in public sector organizations have been identified by several researchers, some of which include reducing the for-cost public service delivery, reducing the incidence of errors, contributing to more effective and efficient citizenry care, providing timely access to information and enables informed decision and personalized service delivery [28–31]. Furthermore, BDA provides improved communication, enables public institutions to make informed decisions based on social media network interactions by the citizens, a basis for managing knowledge in terms of knowledge preservation and dissemination of any form, enables the enacting of informed policies and frameworks that trigger the creation of knowledge, fostering the citizenry's participation in government planning and policy-making activities [32, 33]. It should not be overlooked that citizens are the focus to serve for the existence of public institutions, BDA is known for enhancing the healthcare domain to another advanced level in terms of both the quality and throughput of the services offered by the sector, BDA can help to make the urban planning and development process better and smart [34]. BDA can improve the performance of public sector organizations such as improving the performance of the healthcare domain, supporting disease surveillance and real-time analytics and results among others [35, 36].

Existing relevant RAs

It is worth defining the term architecture before describing what RA means, architecture covers the organization of a given system contained in its components, associations among its components as well as surroundings and tenets that guide its design development, and deployment according to [37, 38], from institute of electrical and electronics engineers (IEEE). However, [39], from the open group architecture framework (TOGAF) summarizes the architecture as a detailed sketch of a particular system but at the level of component for the reason of guiding the implementation of the respective system. Combining the above two definitions of architecture from both IEEE and TOGAF, we concluded that architecture is just a structure full of components, their associations and tenets, and guidance to manage the design, development, and deployment of the respective system over some time. This discussion is also complemented by [39], who define architecture as a group of artifact objects capable of explaining and describing the organization of the particular system by specifying the connections between the artifact objects. The pertinent RAs that were taken into consideration for assessment in this study are described below;

- **Microsoft azure RA (MARA):** The architecture is designed by Microsoft and presents only BDA technologies for BD processing in a cloud setting [40]. It is also called Windows Azure for processing BD in a cloud environment to offer services for Microsoft industrial-oriented operations [41, 42]. It is product-dependent based on the use cases of Microsoft situated in both cloud and on-premise. The RA serves well for cloud BD processes.

- **NTT Data reference architecture (NDRA):** This is the work of Nippon telegraph and telephone (NTT) Corporation [43] and Everis in Spain [44]. The architecture aimed to show the various mechanisms of data utilization in the world by presenting individual technical components such as hadoop distributed file system (HDFS), and complex event processing (CEP). Though the architecture covers all the phases of the complex data pipeline, it uses a rigid approach to data management similar to traditional data management tools. Furthermore, the architecture is only good at integrating and utilizing numerous conventional data and software tools for instance relational databases.
- **Google reference architecture (GRA):** This can be sometimes referred to as Google Cloud Platform (GCP) architecture authored by [45, 46]. The focus of the architecture is on Google Cloud service which operates on the infrastructure that Google applies for the end user's products. These services can include Google search engines as well as tube services [46, 47]. This implies that the architecture is meant to guide the implementation and deployment Google oriented BD services and products. Though all the stages of the data pipeline are present in the architecture, the presentation of data is based on a web setting hence it is only for organizations whose business processes can be served in the web environment**.** Furthermore, no conventional data management software that supports relational databases and rules of engine processing methods. Generally, the architecture supports both the product and industrial environment in contrary to the public sector operating environment.
- **Reference architecture by Nirmala & Raj (NRRA):** This was designed by [48] with a focus on knowing the technical components and elements of BDA usually applied in numerous organizations. Though it covers stream as well as batch functionalities with machine learning (ML) layers along with all phases of the data pipeline, the key components to utilize these industrially commonly BDA tools in the organization's vision. Missions and objectives are missing. Also, the architecture is fit for technological, online, and industrial organizations since it only gathers technical components to guide their assembling process for executing BDA processes hence a technology-driven approach. Web visualization of analytics results is also ill-defined in the architecture layout.
- **NIST RA:** This architecture was designed by the NIST with the focus of making an extensive comparison of available technical RAs related to BD to discover and define components that do exist mostly in these architectures. The architectures covered were nine (9); Lexis Nexl, 9sight Consulting, systems applications and products (SAP), electromagnetic compatibility (EMC), Oracle, International Business Machines (IBM), University of Amsterdam, Microsoft, and strategic architecture.

Since the NIST workgroup already offers a thorough overview of common architectural components, the same technique is also applied in this dissertation. NIST is rich in BDA implementation components acquired from numerous technical RAs. *Tables 1* and *2* show the aspects (variables) taken into consideration while evaluating the existing RAs with proposed architecture as well as architectural components derived from the assessed architectures, especially NIST architecture.

Table 1 Comparison of assessed relevant existing architectures with proposed architecture

RA variable	**NRRA**	**GRA**	**NDRA**	**MARA**	**NIST**	**Proposed architecture**
Technology-Oriented	√	√	√	√	√	×
Metadata Management Layer	×	√	×	×	×	√
Business Process-Driven	×	×	×	×	×	√
Technical, Online, and Industrial	√	√	√	√	√	×
Scalability	×	×	×	×	√	√
Product Independence /Own Services	×	×	×	×	×	√
Whole Data Structure Continuum	×	×	×	√	√	√
Complex Data Pipeline	√	√	√	√	√	×
Performance	×	×	×	×	√	√
Web-Visualization	×	√	×	√	×	√
User Satisfaction	×	√	×	√	×	√

Table 2 Summary of derived architectural components

Layer	Component element
Infrastructure	Infrastructure Services components which include computer Hardware, network hardware among others Infrastructure related to management as well as monitoring Infrastructure related to security services Infrastructure related to systems management Infrastructure related to privacy management Business Process identification management
Interfaces (Analytics and Applications)	extract, transformation, and loading (ETL) as well as mining data Various BDA such as descriptive, predictive, prescriptive, diagnostic, sentimental analytics, analyzing texts, batch analytics, near or real-time analysis, and ML, among others Presentation & visualization such as key performance Indicators (KPI), scorecards, business intelligence (BI) presentations, actions, and signals among others Control which includes quality of data, summarizing and profiling of data among others Hybrid of scorecards and metrics, and BI reporting and events involved
Storage and Management	Sources of data such as Legacy/Entity relationship planning systems/Customer relation management systems, sensors, weblogs, satellites, and machine-generated data among others BD repositories/Storage technologies such as not only structured query language (NoSQL), HDFS relational-oriented databases, in-memory databases and data warehouses, and data lakes, among others Integration process which involves data cleansing, data conversions, and transformations, extractions among others Processing & discovery of data that covers aspects like Stream processing and information discovery Management related to Metadata Auditing, checking, cataloging, and logging

Table 3 gives a summary of the success and failure scenarios of BDA projects in various sectors. Since BDA implementation projects demand a huge amount of investment and resources in general, public-sector organizations should be able to display the showcases for outcomes of such projects through economic and operational feasibilities (Tabesh et al., 2019). On the other hand, the return on investment (ROI) for such projects should be reasonable enough to justify and sustain such projects with their benefits.

Given such an understanding of BDA implementations in public sector organizations, it should be a custom for public sector organizations to assess first infrastructures, the problem at hand, and use cases for which BD and its analytics are needed [49]. Furthermore, it is recommended that any public sector organization intending to implement a BDA project should have well-defined business problems as well as technology architecture and infrastructure for which BDA must address in line with the corresponding organizations, strategy [50, 51].

Table 3 Summary of success and failure scenarios of BDA implementation projects in various sectors across the globe

Sector	Project	Problem	Description Summary	F/S
Public	COMPSTAT Project (USA)	Lack of real reporting of crime cases	In New York, lives were being lost. Then Compstat was introduced. Early on, maps played a critical role by helping commanders identify where crimes were occurring and, more importantly, determine whether arrest patterns aligned with crime patterns. In the interests of civilians, the ultimate goal was to reduce crimes committed and achieve other security objectives. BDA implementation for Accountability was an important aspect of the initiative.	S
Public	Jobs versus Jobs (USA)	How to predict the USA unemployment rate	The researchers devised a category of many words that pertained to unemployment, including jobs. They culled tweets that contained these words and then looked for correlations between the total number of words per month in this category and the monthly unemployment rate.	F
Public	Satellite-Based Yield Measurement (Uganda)	How to predict future yields for farmers	A novel technique that links satellite-based data to plot-level ground yield measurements in Uganda. This allows for future yield projections, which can assist farmers in enhancing productivity	S
Public	NHS (United Kingdom (UK)	Integrate all patient medical records into a central	A project was started by the UK National Health Service to incorporate all patient medical records into a single, distributed healthcare database. This analytical platform's development was hampered by the absence of specific business objectives at the outset. Healthcare providers continued to operate	F

Sector	Project	Problem	Description Summary	F/S
		distributed healthcare database.	in their respective silos since the leadership team did not fully grasp, explain, or communicate the advantage of this initiative to those affected. Technical issues resulting from a lack of analytics team knowledge posed another hurdle for the initiative. The hints that were ignored in this situation were Problem Identification, Specific Skills, and No Data Silos	
Public	Golden Tax Project (China)	How to cover the national taxation authorities, to closely monitor the value added tax (VAT) special invoices and the VAT payment status of enterprises.	China's Golden tax system (GTS) is a BDA project that uses the online network of tax authorities to closely control VAT special invoices, monitor the corporate VAT tax status, and ensure taxpayers' compliance. The key idea was to enhance the effectiveness and efficiency of administering tax through BDA technologies [52]	S
Public	Scottish Independence Vote (Scotland)	The fastest way of presenting the voting results	Cable news network (CNN) aired a graphic showing that 110% of Scotland's population had been polled during a referendum report. The graph indicates that 58% of Scots voted yes and 52% voted no. This little failure might have been avoided with the advice of rational results.	F
public	Predictive Policing (India)	Identify targets for the prevention of crimes through police intervention or the use of statistical predictions to solve crimes.	Predictive policing BDA project usually works in the following four ways; a) predicting places and times with an increased risk of crime, b) predicting potential future offenders; c) creating profiles for past crimes, and d) predicting groups of individuals likely to be victims of crimes. Data is collected from different sources that is to say crime data from police station databases, environmental data including crime seasonal patterns, neighborhood composition, call data records, and other mobile phone data. The next step is the Analysis, where the data collected is analyzed based on a predictive method.	S
Private	Suicide-Prevention App (Australia)	Avoiding potential suicidal cases	To inform users when someone they follow on Twitter says something possibly suicidal, such as "hate me" or "tired of being alone this BDA project goal." After receiving complaints that the software may be used to harass users during their most vulnerable times, the group promptly uninstalled it. Problem Identification was ignored in this example; similarly, to the Target case, they should have carefully estimated the ramifications of using the response. Also, Domain Knowledge may have aided in raising awareness of the possibility of stress in such scenarios.	F
Public	Project Shakti (India)	How to link party supporters to their mobile numbers, through a simple message system (SMS)	The plan was to find out who worked for the party and connect with them. The main issue was data quality, with 50%-70% of the data being fraudulent depending on the location (party volunteers were pushed to insert as many persons as possible into the system, resulting in random names being inserted). The data was then managed in silos, blocking access to those who may benefit from it. Finally, when the system was utilized to survey the linked people, the questions were poorly designed, causing respondents to repeat the views of the party's leaders.	F
Private and public	Google Flu (USA)	How to provide real-time monitoring of flu cases	As mentioned earlier, the number of people infected with the flu was closely tied to the number of people who sought flu-related material on the internet. While the notion appeared to be a fantastic approach to demonstrate the potential of BD, Google got it wrong. Google built its BDA on the assumption that there was a strong association between the number of individuals who searched for flu-related information on the web and the number of people who got the flu. Google Flu, on the other hand, exaggerated the prevalence of flu by more than 50%. The issue was caused by the data used for analytics.	F
Private	Spotify (Sweden)	How to use BDA to improvise music	Spotify uses predictive analysis through BDA to add zeal for enjoyment and fun to the user experience. It also uses a huge stream of BD to predict the winners at Grammy Awards with a prediction accuracy of 67%. Spotify has	S

Sector	Project	Problem	Description Summary	F/S
		engagement and render a personalized experience	one of the biggest Hadoop clusters with 694 heterogeneous nodes running close to 7000 jobs in a day.	
Public	Citizen's engagement and analysis system (CEAS) (Africa)	Awareness of pandemic outbreaks	This project was looking at the pandemic outbreaks such as Ebola, and corona virus disease nineteen-19 (COVID-19) among others in numerous African regions like Sierra Leon, and Nigeria among others (Palfreyman, 2014). The project was to allow the public to lay out information to the government directly regarding any outbreak for analytics purposes and decision-making. The media through which communities were expected to use to communicate with the government such as radio, mobile voice, and others were limited in scope to be accessible to the open government for the project. The radios, mobile voices of toll-free digits, as well as simple messaging services of toll-free digits through telecom companies, were limited by the narrow scope of ML and top categorization to find the class of issues as well as maps for the spatial-temporal content.	F

The most frequently overlooked signals in genuine failures are business and information technology (IT) working together, implementation approaches as a whole, domain knowledge, data quality, and the absence of data silos, all of which demonstrate their importance for the success of BDA implementation initiatives [53]. It may be concluded that, despite public sector organizations' initiatives and efforts, there are various problems in implementing BDA projects in these organizations [54]. Furthermore, [54] states that the characteristics of 3 vs are introduced by Doug Laney from Meta Group, who stated that new techniques, analytics, and new architectures support BDA implementation projects for decision-making in public sector organizations.

3.Materials and methods

As highlighted in the introduction, the study was guided by design science research methodology (DSRM) as it is recommended for studies that aim to provide knowledge and solutions through creating and evaluating innovative artifacts [55]. Data collection and analysis: Mini-literature review (LR) analysis and questionnaire survey were used to identify requirements needed for designing the research architecture. The study adopted a questionnaire as a way to gather data from target participants because it assists in the collection of data on preferences, intentions, attitudes, and opinions from a large number of individuals in a short period [56]. They involve a "series of questions designed to elicit specific information" [57, 58]. Furthermore, this technique was chosen because it can be administered to large groups of persons at a distance and is limited to specific questions [59]. It is worth noting that questionnaires can generate both numerical and qualitative data, thereby supporting the chosen study methodology. The questionnaire included subsections designed to better understand and consider previous efforts by other scholars regarding RA for implementing BDA projects in public sector organizations. These subsections covered demographic features, an assessment of knowledge about current data practices and data ranges, the extent of traditional data management usage, awareness of BDA technologies, types of data analytics, and considerations related to data security and privacy. To develop an appropriate diagnosis of an organization's data management issues, a comprehensive grasp of the present data ranges and processing techniques, the extent of utilization of traditional data analysis tools, and awareness of BDA technologies were required. A variety of techniques were utilized to create the statements that were included in the questionnaire, including brainstorming and thinking, LR, content analysis, and interviews with subject matter experts. For instance, the literature shows a variety of conventional data analysis tools used to varied degrees [60]. Also, a variety of technologies and solutions for enabling BDA-oriented business operations that people may or may not be aware of were evaluated [61, 62].

On the other hand, a LR involves the analysis and assessment of existing documentation about a given subject or topic [63, 64]. It is suggested that the use of the LR at various stages in the design science process [65]. Although the LR was used in finding the problem and understanding the state of the art concerning the implementation of BDA, it was also used as a qualitative research method to establish both requirements and architecture components for designing of RA of the study. The mini-LR method was applied with the aid of the preferred reporting items for systematic reviews and meta-analyses (PRISMA) technique [66]. The mini-review follows a similar fundamental structure to a comprehensive systematic review [67]. Mini-systematic reviews are one approach of condensing current relevant research findings focused on one particular issue, utilizing a constrained number of databases, and a focus on a constrained time frame, in contrast to systematic reviews, which frequently cover numerous questions on a topic [68]. PRISMA is a minimal set of criteria for reporting in systematic reviews that are based on evidence. The mini-review format was chosen because it uses a small number of databases, focuses on a short period, and is one way to summarize current relevant research articles that address a single subject [68]. This mini-review's major goal was to examine the present state of BDA RAs, identify their common architectural components, and highlight their shortcomings. The following is the question that guided mini-LR:

What are the common limitations of current BDA RAs?

The subject was crucial to our study because it throws insight into the limits of existing BDA reference designs as well as prevalent components. The following electronic databases were searched for this study: sciences direct, IEEE explore springer link, journal storage (JSTOR), and association for computing machinery (ACM) Library. Abstract and citation databases and search engines such as Google Scholar and Research Gate (RG) were used to pursue the objective of discovering relevant literature on the issue and to prevent neglecting worthwhile research. The BDA AND RA OR BD, architecture were the keywords for the databases chosen. We added architecture to keywords since the words RA and architecture may have been used interchangeably, and architecture at the abstraction level of the RA may have been referred to simply as architecture. As a result, it was vital that we define these criteria precisely and then categorize research based on these definitions.

Analysis and interpretation of quantitative data was carried out using Microsoft Excel and online forms software. Microsoft Excel was used in the study because of its adaptability, simplicity, versatility, and widespread usage in both the business and academic worlds [69]. The foundation of descriptive statistics was the comparison and investigation of correlations between contested variables. The process of how data obtained from the survey was analyzed is given in *Figure 1*.

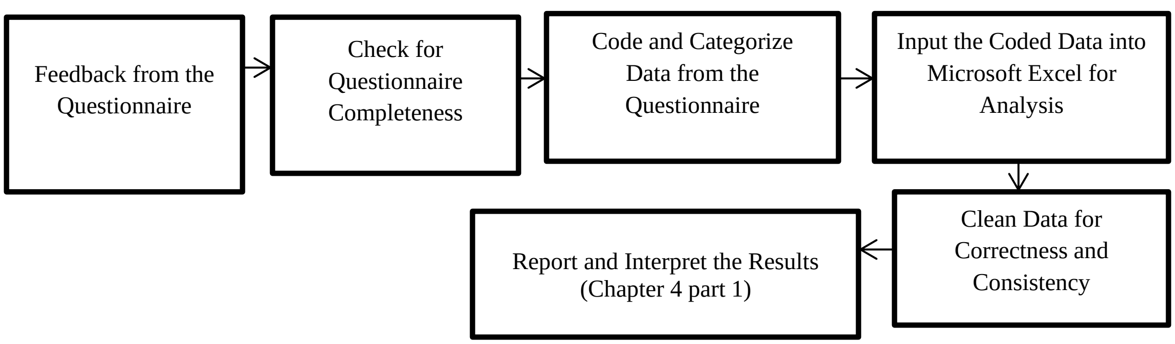

Figure 1 Data analysis process

a) Feedback from the Questionnaire: After the administration of questionnaires to IT and Managerial (M) staff (Target sample), the respondents filled out the questionnaires and returned them. We, later on, collected the questionnaires from the respondents for further analysis of the feedback as per the individual responses. **b) Check for questionnaire completeness:** After collecting the questionnaires from the respondents, we embarked on checking the questionnaires to find out whether all the questions were fully answered and completed. All respondents were able to complete all parts of the questionnaire's pages. Therefore, checking for completeness was aimed at eliminating all those questionnaires that were not fully answered or completed by the respondents. **c) Code and categorize data from the questionnaire:** Coding is the process that involves compiling; organizing, sorting data, and assigning a character or numerical code to all the responses for every question in the questionnaire, to enable the data analyst to enter data into the analysis software. Therefore, data coding was done by the researcher by reading through the responses from the questionnaires and grouping them into categories of related responses. Then numeric codes were assigned to each category of data including respondent's bio-data (like age, gender, education, and organization among others), knowledge of data practices and features, data life cycle tools, data analysis types, and security and privacy. **d) Input the coded data into Ms. Excel for analysis:** After assigning each element of the questionnaire a given numeric code, hard copies of the questionnaires were organized, and all the feedback or data from each questionnaire was entered into the Ms. Excel software using the same codes/numeric figure as it was answered by the respondents. **e) Clean data for correctness and consistency:** Data cleaning is the process that involves finding missing data and eliminating data entry errors, to make sure that entries are correct. Cleaning data further involves consistency checks which serve to identify data that is out of range and has extreme values and fixing of the missing responses. Therefore, data cleaning was done to eliminate some of the entry errors of faulty data entries. While entering a given range of numbers like one (1) or two (2) to represent male or female in the gender field, any other figure other than the defined ones could have been mistakenly typed in the field of gender. So, cleaning data was mainly done to eliminate such errors and finally, **f) Report and interpret the results:** During this phase, the researcher performed various analysis tests on the input data to derive the relevant information through several reports. Some of the analysis tests included computing frequencies, percentages, and means (averages) tables for the various demographic groups and other major study variables of the study as seen in the following sections.

On the side of qualitative data analysis, the goal was to find broad generalizations about the correlations between various categories of data [70]. Content analysis was used to examine qualitative data gathered from the LR. The interpretation of findings formed a basis for some requirements and components which eventually set the ground for design decisions needed for the development of the RA. When a LR is used as a research method, it is a qualitative research method. *Table 4* Show the inclusion/exclusion criteria adopted in study. The *Figure 2* depicts how the mini-LR was carried out using the PRISMA flowchart as guidance.

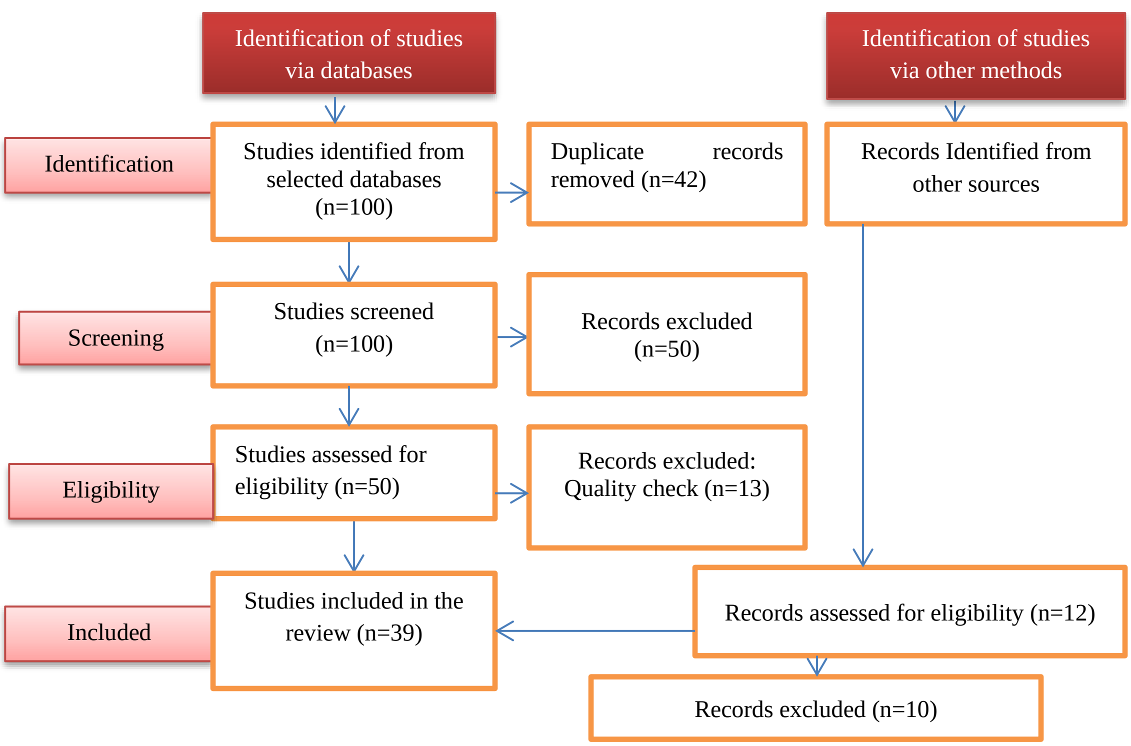

Figure 2 PRISMA flowchart for mini-LR

Table 4 Mini-LR exclusion and inclusion criteria

Inclusion	Exclusion
Papers between 2015 and 2023. Scholarly articles and publications that show the current state of RAs in the realm of BDA and possible results. Books, theses, and conference proceedings, white papers with knowledge of BD reference designs.	Informal literature surveys with no clearly defined research questions or research procedure. Duplicate research reports (a conference and journal version of the same paper) Short articles (less than 5 pages) Studies produced in a language other than English and Grey literature

Target population and sampling procedures: The study target population was the IT and M staff from three (3) Uganda's public sector organizations; Uganda Bureau of Statics (UBOS), the Ministry of Health (MoH), and Ministry of Education and Sports (MoES) (*Figure 3*). The literature rationale for the selection of these three (3) organizations as the top producers of data in Uganda with exceptional interest in utilizing BD for their daily public business according UBOS report 2016. Moreover, M staff members were selected because they are core stakeholders in the decision-making process, which is the center for BD utilization. IT staff members were chosen because the research belongs to the IT domain which requires participants who are aware of the technology under investigation. *Table 5* shows sample size from each target population unit.

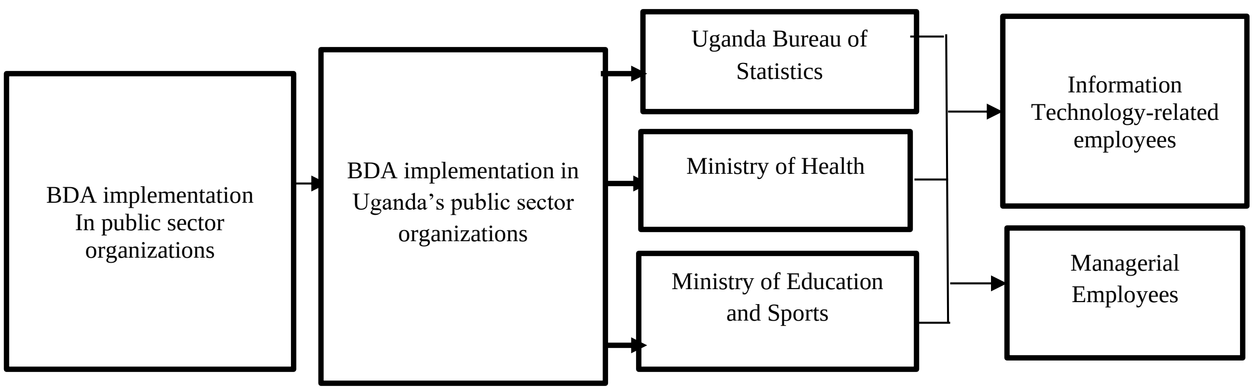

Figure 3 Sample selection procedure

Sampling refers to the practice of using a subset of the population to represent the entire population [71]. There are about 29 public sector organizations in Uganda [72]. However, only the three public sector entities were purposively taken into consideration for this study. The study applied a purposive sampling technique from a non-probability sampling group of techniques. The selection of the technique was made based on the fact that the technique allows the capturing of only participants with vivid exposure and knowledge of a particular topic under investigation [73, 74].

With a rigorous review of existing methods of determining the sample sizes; the census method applicable for the small population, published tables method, benchmarking the size of samples for related studies, and formulas [75, 76]. This study employed the method of published tables for estimating the size of the sample for the study. This is because these tables have already established numerous numbers of populations with their respective sample sizes from where the researcher can easily pick a sample size that matches the study population under investigation. The method is simple and less time-consuming [77]. The table published by Krejcie and Morgan was considered for determining the sample size [78,79]. According to the UBOS report of employees in public sector organizations (2020), the number of IT and Management (M) staff totaled in three (3) selected organizations is approximately forty (40). According to [79], for any population of forty (40) participants, the sample size of thirty-six (36) is sufficient for the study as shown in *Table 5.*

After analysis of data obtained from questionnaire survey and mini-LR the following requirements were derived , 1) Need to first identify the business process/Use case for which data is needed so that these business processes generate added business value, 2) Need to support low, medium, and high volumes of data, 3) Need to support collecting data from multiple sources such as centralized or distributed data sources, social media, surveys, and secondary data from ministries among others, 4) Need to make sure that the data sources for a given decision can be extended easily and quickly by adding new data sources, 5) Need to support slow, medium, and high-throughput data transmissions between data sources and the BD platform like transactional systems, 6) Need to incorporate BDA tools throughout the post data collection management, 7) Need to support diversified processing and analysis (like predictive, batch-oriented, real-time, and stream, among others), 8) Need to ensure compliance with central data privacy laws, 9) Need to protect and preserve security and privacy on highly sensitive data for mitigating data and IT-Security risks to a great extent, 10) Need to support high-level/advanced data security measures such as time multiplexing and Key management are some of the requirements that need to be addressed for successful implementation of BDA in public sector organization for utilization of BD. The results of this study were mapped to carefully selected architecture components which in turn were considered as a basis for the design and development of the business-driven RA for BDA implementation (BRABDAI) in public sector organizations. Details of the developed BRABDAI are presented in the following result section.

Table 5 Target population, population size, sample size, and method for sampling

S/N	Target population	Population size	Sample size	Sampling method
1	UBOS	15	19	Purposive
2	MoH	15	11	Purposive
3	MoES	10	6	Purposive
	3	**40**	**36**	

4. Results and discussion

The results of the study primarily focus on the designing of BRABDAI in Public Sector Organizations, taking Uganda as a case site. The process of designing BRABDAI started with the establishment of design decisions from the derived requirements (*Table 6*). Design decisions D6 and D10 were drawn from requirements obtained from the mini-LR of the existing relevant and current publications on RAs. The mini-LR also further helped in identifying architecture components particularly those discovered by the NIST workgroup. Through rigorous literature assessment, architectural components (*Table 2*) were derived on which design decisions were mapped. The second phase of designing BRABDAI was mapping the design decisions to the identified architecture components (*Table 7*).

Table 6 A Summary of derived design decisions based on derived requirements

S/N	Requirements	S/N	Design decisions
R1	Need to ensure that these business processes produce additional business value, it is necessary to first define the business need in terms of business process/operation.	D1	Ensure that the delivery of BDA is driven by business, not technology.
R2	It is necessary to support low, medium, and high data amounts	D2	Put in place the technologies required to handle the data volume fluctuations
R3	Need to ensure support for data collection from various sources is required, including social media, surveys, dispersed and centralized data sources, and secondary data from government agencies, among others	D3	Implement the essential tools that can handle the volume, variety, and velocity of BD.
R4	Need to ensure that the data sources for a particular choice may be quickly and readily expanded by including emerging sources of data assets	D4	Make sure that flexible data management tools are being used.
R5	The BDA approach must handle transactional systems, such as slow, medium, and rapid data exchanges.	D5	Deploy the technologies that are required and capable of addressing the huge data velocity (v) at any rate.
R6	Tools for BDA must be integrated throughout the data pipeline.	D6	Providing suitable advanced analytics tools across the data management lifecycle
R7	Support for a variety of processing and analysis is required (including stream, batch-oriented, real-time,	D7	Support for any conceivable processing and analytical insights

S/N	Requirements	S/N	Design decisions
	and predictive, among others).		
R8	Support for advanced data security mechanisms like temporal multiplexing and Key management is required.	D8	Offer the necessary security and privacy infrastructures for data quality, recovery, and loss prevention in addition to a workflow with resource management.
R9	Highly sensitive data must be protected and kept private to significantly reduce data and IT security threats.	D9	Provide the necessary security and privacy infrastructures for data quality, recovery, and loss prevention, in addition to a workflow with resource management.
R10	It is necessary to verify adherence to national data privacy and security regulation	D10	In addition to a workflow with resource management, offers a security and privacy architecture for data quality, recovery, and loss prevention.

Table 7 Mapping design decisions to architectural components

S/N	Architectural Component(s)
D1	Business Process identification management
D2	Engines for processing, technologies for loading and storing data, and result visualizations
D3	Engines for processing, technologies for loading and storing data, and result visualizations
D4	Data Transformations, Data loading, Event Triggering, Integration
D5	Processing Engines, Data loading Storage Technologies, Result Visualizations
D6	Data integrity, loss prevention, event triggering, recovery Workflow and resource management, storage technologies, and data lifecycle (DLc) management
D7	Processing engines, ML, Application Programming Interface
D8	Data Security and Privacy
D9	Data Security and Privacy
D10	Data Security and Privacy

The third phase of BRABDAI designing was the merging of design decisions according to architectural components. *Table 7* shows that specific design choices are related to specific architectural elements (one-to-one mapping relationship) while others share the same group of architectural components like D2, D3 & D5, as well as D8, D9 & and D10. Design decisions that share the same group of architectural components were merged to form the final design decisions on which chosen architectural components are mapped. The architectural components are the basis of RA development. Design decisions code numbers were adjusted for that reason as shown in *Figure 4*.

The last phase was to determine how BRABDAI can be visualized. It is well known that when it comes to designing an artifact, there is always the matter of what notations to use when portraying architecture in a diagram. There are several different notations, such as unified modeling language (UML) and business process modeling notation (BPMN), as well as tools, such as Sparx Enterprise Architecture, and each has benefits and drawbacks. It was necessary to be able to display a data flow and workflow steps where data is processed using a combination of UML sequence, component, and activity diagrams to reflect the new RA. Given the lack of standardized notations for this use, generic notations (boxes and lines) were chosen, as seen in *Figure 5*.

The new RA can be seen as a top-level modular system from where any components can be picked for BDA solution implementation for a particular BDA business process in public sector organizations. Since the RA is based on several requirements for operationalizing business processes in public sector organizations through the implementation of BDA, it is made sure that the developed RA easily provides BDA architectural components needed for implementing each business process. The highest level one (1) of the BRABDAI in public sector organizations shows a top-level view of all the architectural layers in the RA. It covers business processes, data sources, core components, and cross-functional components. As it has been pointed out in the comparison of existing BDA RAs, a data pipeline should always be built following the same structure with four main phases. Therefore, the core components consist of ingestion, storage, analytics & processing, and presentation & operationalization. Ingestion and storage deal with loading, transforming, and storing data as it comes from various data sources. In analytics & processing the data that is now available after it has passed the two previous layers is analyzed by executing analytical or predictive algorithms on processing engines. The goal of this layer is to derive insights from the available data. These insights can then be visualized or sent for post-processing to the presentation & operationalization layer where a data consumer (applications) uses them. To group the architectural components and improve the readability of the architectural diagram, several categories of architectural components are introduced. The four phases within the core components and the cross-functional components can be regarded as functional BDA methods. These processes, particularly within the core components are implemented using BDA applications, systems, and components. Data arrives in a BDA application that is built using the business-driven RA from several sources with some examples visualized in *Figure 6* as well. Both the core components and the cross-functional processes are executed on infrastructure components, which form the last category.

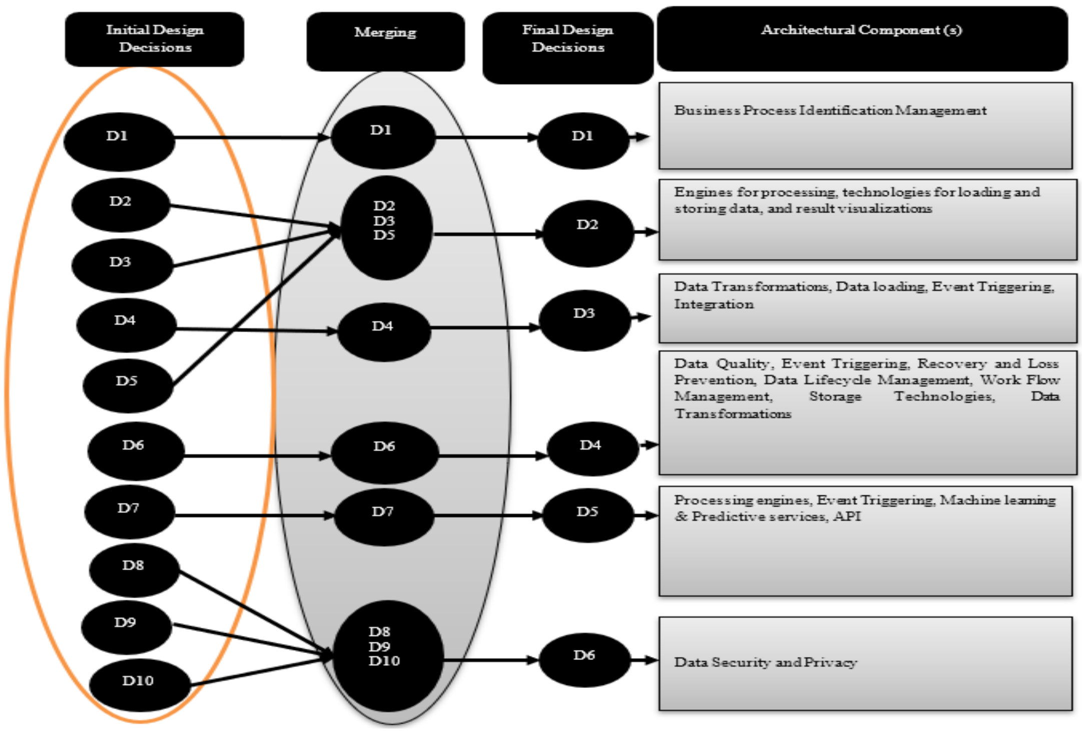

Figure 4 Merging design decisions

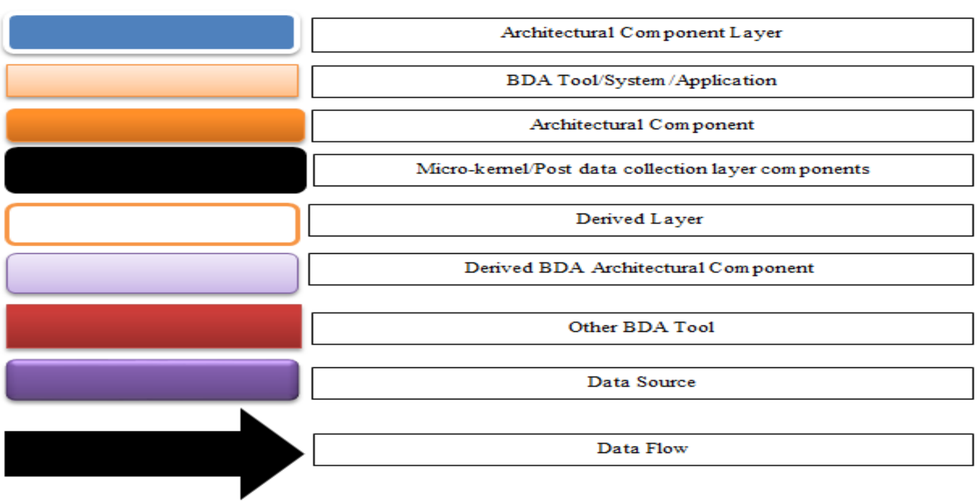

Figure 5 RA generic notations for architecture visualization

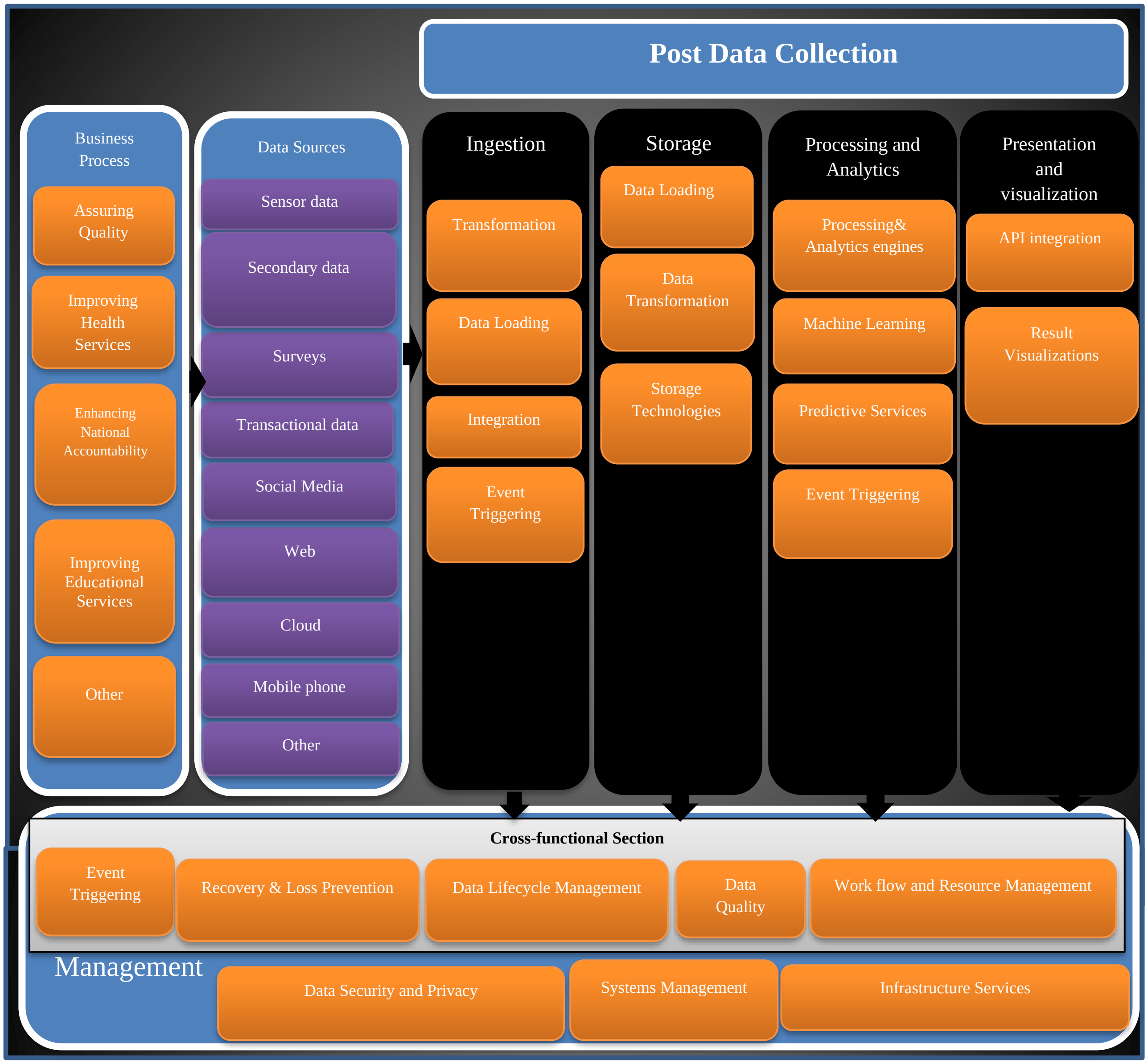

Figure 6 The initial BRABDAI in public sector organization -the highest level

Description of the BRABDAI-Level Two (2): The developed RA layers of level two (2) are described in detail below:

a) Component of business processes layer: What do public sector organizations hope to achieve with BDA? Does it want to improve decision-making processes, better understand its customers/citizenry, or find new revenue opportunities? Once it knows what it wants to accomplish, it can start the BDA implementation project. These are the questions that this layer of the designed architecture answers. After merging, this component answered design decision D1, which in turn satisfied the requirement (R1). As a result, the business process component is focused on determining the business use cases/processes that justify the requirement for which a specific dataset is required that later should decide on which BDA technology to implement. Therefore, the approach is to assess what information is required for what objectives for those purposes to provide business value in the public sector organizations. Data sources, from which information about a specific business process can be accessed, are an architectural component required for meeting this requirement. b) Component of data sources layer: What data does the public sector organization have, and where does it come from? Organizations will need to think about structured, semi-structured, and unstructured data and internal and external data sources. After merging, this component addressed D3 which satisfies R4. The sources of datasets offer the necessary quantity of information for gaining insights, enabling BDA to contribute value and support the targeted business process for the specified public organizations. The generated RA only provides a few examples; the sources vary depending on the offered public sector enterprises' business processes. The RA includes images of social media, surveys, the web, the cloud, mobile devices, secondary data, sensor data, and transactional data (information from current operational systems). Public sector organizations mostly employ three (3) sources: secondary data, public transaction activities, and surveys. Only using these three (3) sources provides little insight into evidence-based decision-making. Consequently, there is a requirement for using different sources of data to make evidence-based judgments.

c) Post-data collection layer component: Ingestion Component: Design decision D2 was addressed by the ingestion component, which is translated to requirements R2, R3, and R5. The part also addressed design choice D4, which satisfied requirement R6. The extension was created because data must first be fed into the BDA tool from the data sources. Data can be added to the BDA tool in several methods, including batch loading and stream loading. Batch loading is the process of loading data from a queue into a BDA tool at a predetermined, programmable time. This method can be favored when importing data into the HDFS or Data Lake storage technologies from current transactional systems. The apache sqoop tool, which moves data from relational databases into Hadoop, may be used for this. Data is consumed into the BDA tool by stream-loading as soon as it is produced. As a result, it is utilized in the majority of internet of things (IoT)-focused business processes as well as other speed-based technologies that rely on quickly deriving insights from data. Although the term "streaming" has a broad definition, in this context, it solely refers to the act of absorbing data without processing it. Although there are many alternative apps available for stream-loading data, Apache Kafka is the most well-known. Data from data sources referred to as producers, is written to a distributed log using this streaming platform, which functions as a publish/subscribe queue. The log is divided into subjects, and those topics are further divided into partitions. Applications that use the data (referred to as consumers) can subscribe to various subjects on the streaming platform to acquire the data they require. Finally, traditional ETL technologies are currently in place in the majority of public sector enterprises for moving data into well-known enterprise data storages like warehouses may also be used to load data into the BDA tool. Data can change after entering the BDA tools. The positioning of the data transformation components, however, is highly influenced by the business process, goal, and governance standards. For instance, data may be stored in the HDFS in its raw form, and modifications can only be made to it when it is extracted for analysis. Here, any data transformations that are unique to business processes may be handled. To avoid the storage technology, like Hadoop/Data Lake, from becoming a data swamp, it would make sense to apply cross-business process purification processes and validations to all data before it is placed in the storage zone's BD tool. Data transformations can range from altering data format to cleaning and verifying data to augmenting and combining data. In summary, this component is responsible for collecting and storing data from various sources. In BDA, the data ingestion component processes and extracts data from various sources and loads it into a data repository. Data ingestion is the essential component of the developed business-driven RA because it determines how data will be ingested, transformed, and stored. Ingestion BDA Tools include Apache Storm and Trident, Kafka, and Sqoop among others, Storage Component: With the ingestion component, the component shared design choices (D2) and (D4) as well as requirements (R2, R3, and R5) and R6. Hadoop for BD storage serves as the primary illustration of this component in the storage layer. Its essential component is the actual HDFS, which houses all the data needed for analysis in the public sector organization. Depending on the data that needs to be stored and the associated business procedure, different storage technologies, including column databases (CDB), key-value databases (KVDB), graph databases (GDB), document databases (DDB), and analytical databases (ADB), are used to implement the physical Hadoop. The component indicates the locations where all the data needed for analysis is kept. Neo4j, MongoDB, and Not Only Structured Query Language Database (NoSQLDB) are a few examples of different storage BDA solutions. To prevent everyone inside the public sector organization from having access to every piece of data, the storage technology might be divided into several zones. This also ensures adherence to data privacy rules, such as the prohibition on the storage, combination, or public disclosure of information from personal health records. The data science data lake, which is used to store data needed for sandboxing, is one zone that is specifically designated in this business-driven RA model. This area was created because it is not always permitted for data scientists to directly access the production data that operational systems and other sources provide to the data lake when they are investigating and building new models. To comply with compliance data regulations, this zone contains data that has been partially anonymized, but it also enables data scientists to create accurate models. An access control component is in place and coupled to the Identity & Access Management process from the cross-functional layer to guarantee that no one can access the data inside the relevant storage zone without authorization. Additionally, the cross-functional layer's procedures for data quality, data privacy & security, and data lifecycle management are tightly connected with storage technologies. However, as it must be available for all storage zones of the storage system used, metadata can be kept in the database. It seems reasonable to present it as a separate component and isolate it from the physical data storage. Through querying connections, data may be accessible from any business process' analytics & processing layer. After extracting data from the HDFS, transformations can be done depending on the state of the data. It is also possible to execute further data aggregations, validations, and enrichments. Consequently, the HDFS also has a connection to transformations. The sample storage platforms within the storage layer are shown in *Figure 7*. The data storage part is in charge of keeping the data in a way that makes it simple to retrieve, process, and analyze it. The other forward components must have access to the data, and that is made possible by this layer.

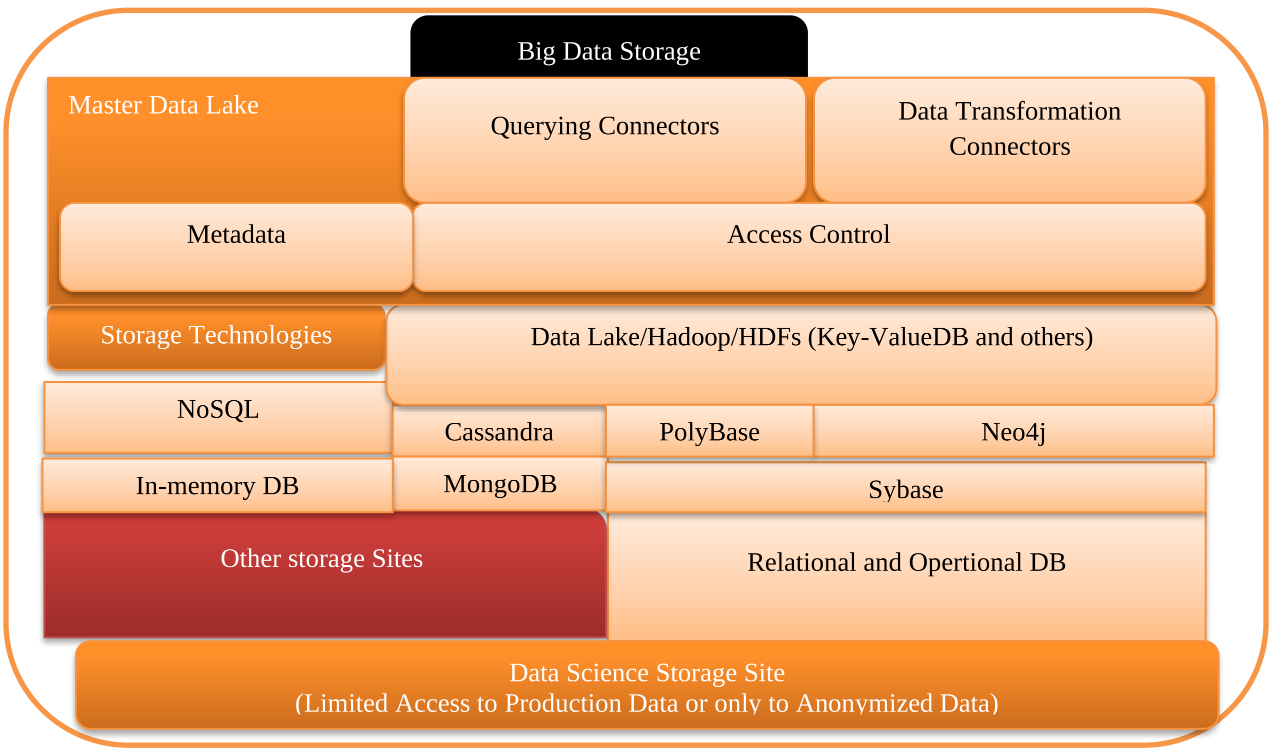

Figure 7 The BRABDAI – level two (2): Storage layer

d)Processing & analytics component: This part responded to design choice D5, which in turn satisfied requirement R7. Data is processed so that insights can be drawn from it using analytical and predictive algorithms and then used by data consumers (applications that require data for additional processing). Utilizing already available ML principles, new ML models can be built, trained utilizing BD sets, and then implemented as a prediction service. Common ML languages like Python, R, or Scala are supported when building models for processing and analytics. Transfer learning can be used on existing ML application programming interfaces like Google Cloud Vision or natural language application programming interfaces for common ML tasks like text analytics or picture recognition. The predictive service, that holds the trained models, has a component for tracking and evaluating the model's precision and accuracy throughout production performance. Model parameters can be adjusted as needed for maximum outcomes, or the model can be retrained. Methods for unsupervised ML can be used to cluster data and discover fresh, undiscovered patterns. All analytical and predictive algorithms, whether in the training or production phases, require processing engines to function. Depending on the processing speed and underlying business process needs, many processing engines or frameworks exist. They can be categorized into several different groups, including batch processing, stream processing, and others. In summary, the processing part of this component is responsible for collecting, cleaning, and preparing the data for analysis. This processing part is critical for ensuring that the data is of high quality and ready to be used in the future. *Figure 8* represents the processing & analytics component.

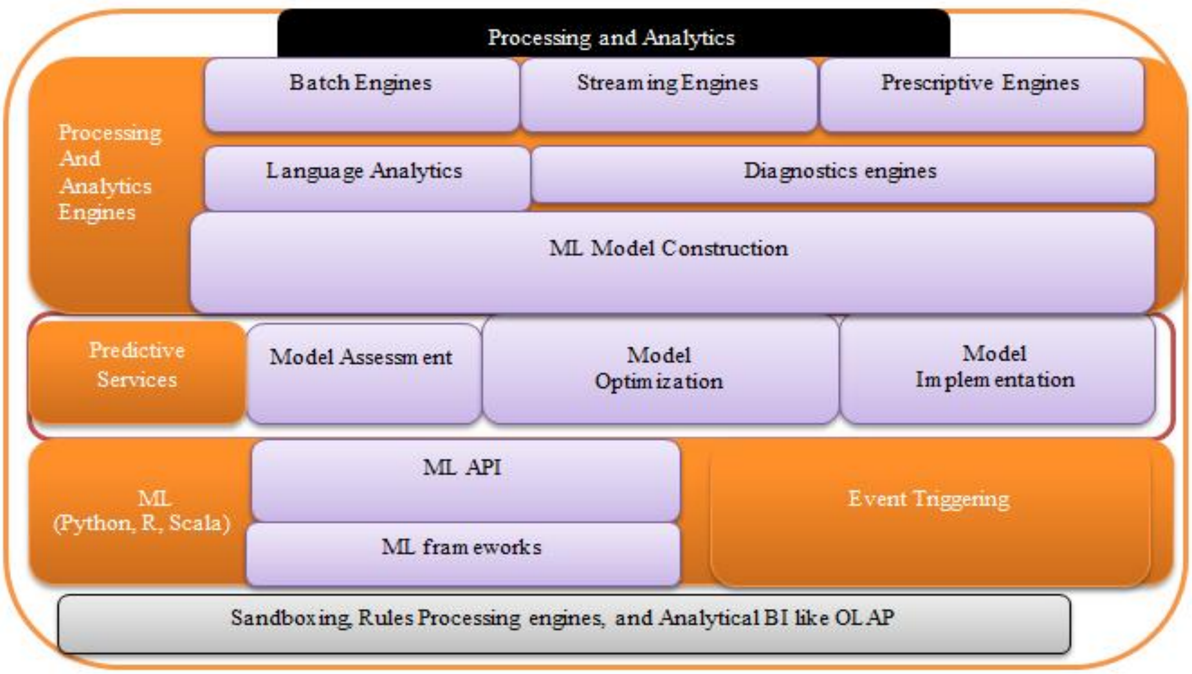

Figure 8 The BRABDAI–level two (2): Processing and analytics layer

e) Component of presentation & operationalization: The foundation of the component is the requirement for post-processing applications to view and use the results of calculations made in the analytics and processing phase to derive insights from data (*Figure 9*). Dashboards, diagrams, and report creation are examples of typical visualization tools that are used in data visualization. Using the information gathered during the analytics and processing phase, reports are automatically created in natural language. A BDA tool can initiate operations in other systems or transfer its findings there via application programming interfaces, enabling this further processing, through application integration. Additionally, tools for visual data exploration like sandboxing should be accessible; two examples are the cloud data lab on the GCP and the Azure ML Studio. Data analysis outputs must be processed further to bring value to the majority of business operations in public sector enterprises, in addition to being displayed. Through application integration, a BDA solution may start processes in other systems and send its findings there via application programming interfaces, enabling this post-processing. Making the findings available for manual checks and validations before handing them off to the subsequent system is the responsibility of another crucial component. This is crucial for managing complicated situations and assuring the quality of results, such as fraud detection or claims settlement in the legal and security sectors. It has to be tightly connected with visualization tools so that the interested parties may examine the outcomes of the data analysis and learn more about the relevant occurrences. In summary, the data presentation and operationalization components are responsible for creating visualizations of the data that humans can easily understand. This component is important for making the data accessible for real-world decisions.

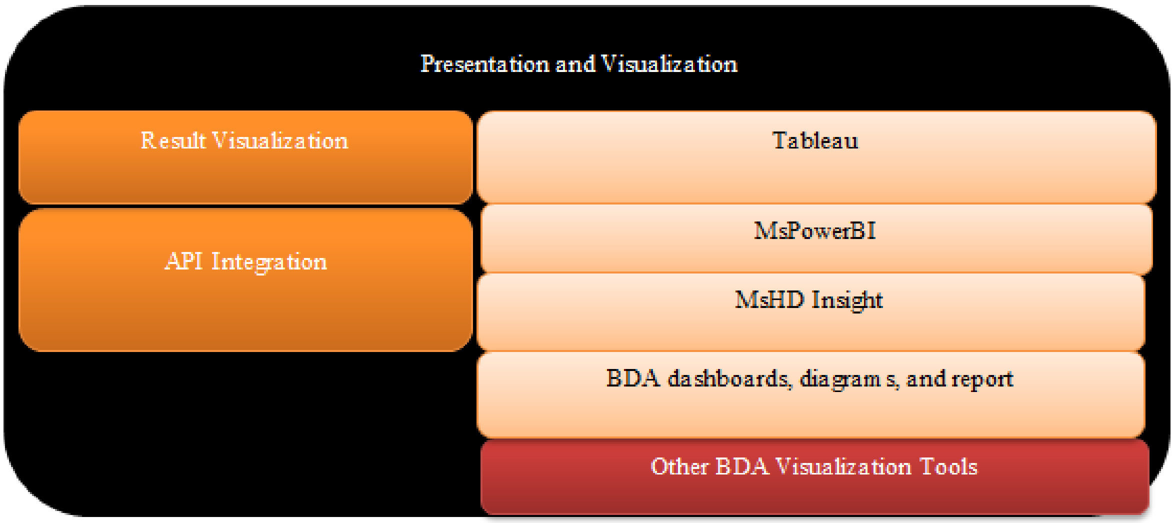

Figure 9 The BRABDAI-level two (2): Presentation and operationalization layer

f) Infrastructure/Management Layer: In the new RA, the infrastructure layer is kept, along with components for data privacy and security. It was expanded, nevertheless, by the addition of a cross-functional part that covers all functional activities that must support the essential elements and span the full data pipeline. Infrastructure/Management Layer includes cross-functional section component: The design decision D6 was handled by this cross-functional section component, which also satisfies the requirements R8, R9, and R10. This component or layer includes infrastructure as well as all operational procedures that support the fundamental elements of the BRABDAI. It spans the full post-data collection components. Here, data privacy and security concerns are addressed, assuring system security and compliance with data protection legislation (Metadata concerns). Identity and access management, a component of this, defines roles and privileges for data access. To ensure that the BDA platform complies with the Regulation for General Data Protection (RGDP) by following the rules of purpose for data consumption, data stewardship explains why a role has access to specific data sets and how it uses this data.

Anonymization provides abilities for anonymizing data sets, which is mainly required for business processes to ensure compliance with the RGDP and other data protection rules when personal identification information (PII) is involved. Data validations, verifications, audits, and quality checks are carried out, especially on data saved in the storage tool, to assure data quality throughout the BDA platform and across the post-data collection layer. Furthermore, quality assurance checks must be made on the outcomes generated by the analytical and predictive models. By including replication techniques in the BDA platform, further steps like recovery and loss prevention ensure that data is not lost in the event of outages and also provide assistance for restoring crucial components after an outage to keep operations up and running. According to factors like, for example, access frequency, data type, or data structure, the management of DLc addresses issues like how long data may be maintained on a platform and where to retain specific data sets. The management of available resources as well as the coordination of processes workflow and resource management are topics covered inside the BDA platform. The latter is especially crucial for a BDA platform because resource consumption is frequently high and it must be ensured that there are sufficient amounts of them accessible to prevent post-data collection components from breaking down if one task begins consuming all resources. The BDA platform must have the capabilities to operate in a dispersed environment, which infrastructure delivers. This comprises the physical components or the server nodes on which BDA applications can be installed, as well as the networks that link them. The infrastructure layer with the cross-functional section component is depicted in *Figure 10*.

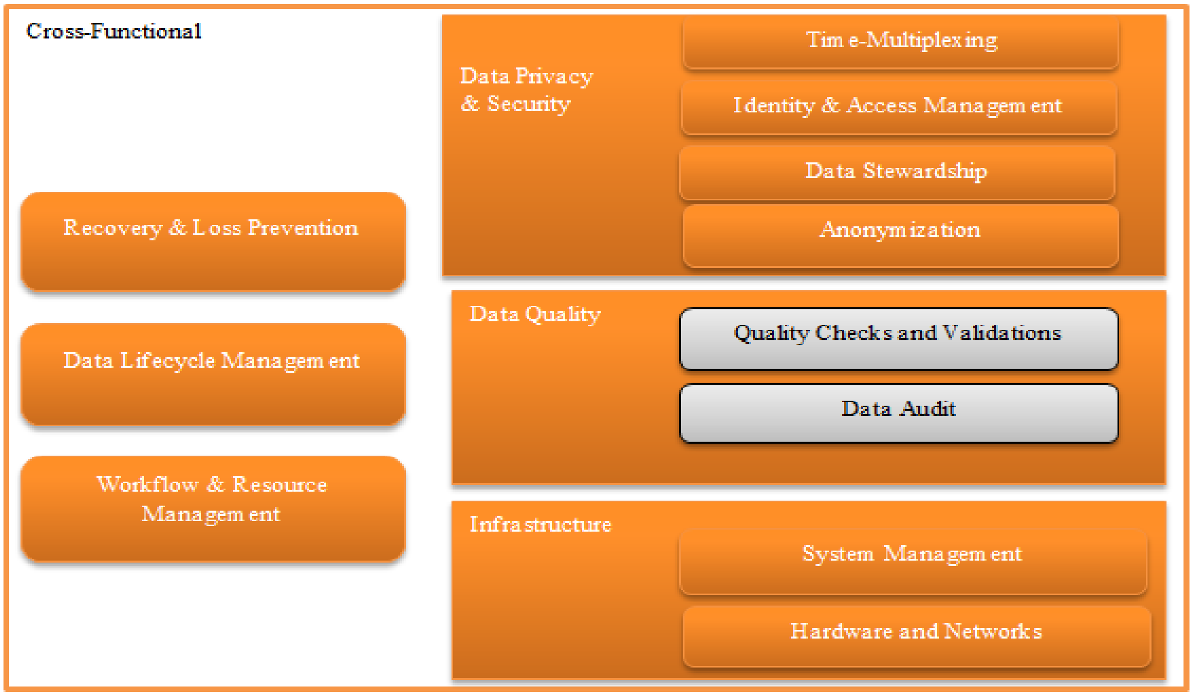

Figure 10 The BRABDAI–level two (2): Cross-function section layer

From the LR, BDA in Public Sector Organizations can play a vital role in improving decision-making processes, service quality, or cost reductions, which should be included to substantiate the claims. BDA is the emerging set of tools and methods to manage and analyze the explosive growth of digital information and includes visualization, ML techniques, and algorithms [79, 80]. BDA is further termed as a process of harnessing, finding, and harvesting information and knowledge from a large number of datasets for informed decision-making [81]. We can say that BDA is about two (2) things; BD and analytics as well as how the two (2) band up to create one of the most profound trends in BI today.

Public sector organizations are known with solid history to have potential numerous avenues from which big volumes of data can be obtained such as surveys, web, mobile applications, and records among others [81, 82]. However, most of these avenues for data are not utilized by the public sector to address societal citizenry problems. Public sector organizations are uncertain about how to harness the advanced analytics tools to utilize ever-increasing data volumes and sources for the benefit of the public of improved public service delivery, justice and transparency, and saving money among others.

Studies of BD applications in the public sector indicate that, the worldwide regretful problem of "food scarcity" can be eliminated by harnessing and leveraging BDA implementation. They referred to this solution as "BDA to feed the earth". Moreover, BDA Innovative strategies can be used to predict and prevent disease outbreaks such COVID-19 pandemic from spreading [83]. Moreover, for public sector organizations not to be able to place BD and associated analytics into work, the consequences are bound to be pretty severe and negative. The Gartner, BD research firm also adds that by 2023, sixty percent (60%) of business integrity abuses will occur through unsuitable implementation approaches of BDA tools in the public sector.

Therefore, from the above discussions, it can be observed that there is a need for implementing BDA in public sector organizations in a way that is in line with public sector organizations' mandates, interests, and business processes. It is asserted that eighty-five percent (85%) of the BDA implementation projects fail to bring the positive revolutions anticipated [84]. Other organizations are even avoiding projects related to BDA altogether in fear of the registered high failure rate [85–88]. BDA projects such as the Brexit project that was developed for predicting elections in the United Kingdom (UK) and, the flue trend (FT) project by Google is some of the known scenarios of BDA implementation failures in the global sphere. It can be deduced that designed BRABDAI can guide the successful BDA implementation projects in public sector organizations.

The study had some limitations like any other investigation. Despite the findings' answers to the research questions, the researcher encountered limitations, which are detailed here. The majority of the respondents had never heard of the topic under study; thus, the researcher was obliged to rewrite the majority of the questions that called for the use of sophisticated data tools to draw in and keep participants throughout the testing of the questionnaire. In addition, due to time and resource limitations, the study's conclusions and implications were based on data gathered from purposefully chosen respondents in Uganda's public sector organizations. However, given that BDA is a widely used technology, even in the commercial sector, for the sophisticated use of data for decision-making, this may have overlooked some information. Future studies can be expanded to include more respondents from private-sector organizations to remove any bias that may have developed from solely including respondents from public-sector organizations. BDA is a dynamic phenomenon that has a wealth of literature or documentation. By consulting with and receiving validation from BDA practitioners and experts, the researcher overcame this. In addition to reviewing pertinent research articles on BDA, the researcher searched only for pertinent supporting documentation.

A complete list of abbreviations is listed in *Appendix I.*

5. Conclusion

The development of the BRABDAI was done by critically selecting architectural components from NIST work group architecture to address the derived design decisions which in turn fulfilled established requirements. The outcome of applying the BRABDAI in practice is expected to include but is not limited to an improved decision-making process, quality of public service, and reduced cost in delivering government services to the citizens along with citizenry sentimental analytics among others. Public sector organizations' key stakeholders and executives should decide immediately which business processes require which BD to harness and how to align BDA technologies with those business processes/Use cases. By deciding on the business process first, the purpose for which BD is collected, stored, processed analyzed presented visualized and disseminated using BDA can be realized seamlessly. This means that public sector organizations should start both exploratory and large-scale projects and programs for operationalizing BDA business processes and solutions. With the methodical path being set, it is up to public sector stakeholders and executives to decide which business processes and their corresponding BDA tools their organizations should implement.

The study also suggests some avenues for future investigations like; it considering business processes in the public sector, which would be addressed by another investigation, considering business processes from the private sector. The proposed architecture would benefit much if the sample of the considered organizations is increased. Exploring more new data sources or testing the architecture in different public sector contexts, would provide clearer guidance for future research efforts. The researcher suggests furthering the investigation by carefully examining and assessing each BDA technology, such as HDFS, Cassandra, and Red Minder, to make trade-offs. Perceptions of BDA by statisticians and management need to be investigated to validate the use and acceptance of BDA technologies in public sector organizations. This research can also tap into concerns regarding the application of BDA tools due to their non-traditional methods of use and analysis. Moreover, all components explored in this study along with all use cases/business processes are related to public sector organizations, hence further research on other industries with their use cases will provide further input and validation. Even though this will require extensive work and research, it enables us to discover key elements, challenges, techniques, and methods, hence would allow us to make more references. In addition to exploring further cases, the future can work on data analytics related to other industry 4.0 technologies and services including open sources and commercial offers, quantum computing, and cloud computing. Future research should also address the role of public sector key stakeholders and executives in harnessing BDA and how to align BDA technologies with those business processes. This has a nexus with research into infrastructure issues especially in developing countries should be thought about and decided upon in a short time to avoid unjustifiable capital investments associated with BDA projects whose RoI is hardly quantifiable and justifiable. Since BDA and AI will continue to play a significant role in public sector organizations, the nexus between the two should be investigated. Furthermore, it is essential to look into new options since well-known digital companies like Google and Amazon will seek out new business prospects in the public sector and launch new products with cutting-edge, attractive citizen experiences. Through the practical deployment of one of the BDA tools, such as HDFS or Cassandra, the industrial application and assessment of the BRABDAI if investigated will contribute to knowledge and validity while exploring the scenarios of one or more business processes identified in this study. The study finally recommends further studies focused on developing a framework for enhancing the use and acceptance of BDA implementation in public sector organizations. The future research areas can also include examining and generating a distinct set of data for validity and reliability, hence other research might use a different approach, such as participative, human centered design (HCD), behavioral approaches.

Acknowledgment
This work was carried out with the aid of a grant from the Artificial Intelligence for Development in Africa Program, a program funded by Canada's International Development Research Center, Ottawa, Canada, and the Swedish International Development Cooperation Agency.

Conflicts of interest
The authors have no conflicts of interest to declare.

References
[1] Hidalgo-sanchis P. UN global pulse: a UN innovation initiative with a multiplier effect. Data Science for Social Good: Philanthropy and Social Impact in a Complex World. 2021:29-40.
[2] Klievink B, Romijn BJ, Cunningham S, De BH. Big data in the public sector: uncertainties and readiness. Information Systems Frontiers. 2017; 19(2):267-83.
[3] Henrion M. Why most big data analytics projects fail: how to succeed by engaging with your clients. ORMS Today, Informs. 2019.
[4] Rahmani AM, Azhir E, Ali S, Mohammadi M, Ahmed OH, Ghafour MY, et al. Artificial intelligence approaches and mechanisms for big data analytics: a systematic study. Peer J Computer Science. 2021; 7:e488.
[5] Chen B, Tu Y, Song Y, Theobald DM, Zhang T, Ren Z, et al. Mapping essential urban land use categories with open big data: results for five metropolitan areas in the United States of America. ISPRS Journal of Photogrammetry and Remote Sensing. 2021; 178:203-18.
[6] Guenduez AA, Mettler T, Schedler K. Technological frames in public administration: what do public managers think of big data. Government Information Quarterly. 2020; 37(1):101406.
[7] Desouza KC, Jacob B. Big data in the public sector: lessons for practitioners and scholars. Administration & Society. 2017; 49(7):1043-64.
[8] Munné R. Big data in the public sector. New Horizons for a Data-driven Economy. 2016:195-208. Springer Open.
[9] Wang Y, Kung L, Byrd TA. Big data analytics: understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change. 2018; 126:3-13.
[10] Aggarwal S, Manuel N. Big data analytics should be driven by business needs, not technology. McKinsey & Co; 2016.
[11] Ataei P, Litchfield AT. Big data reference architectures, a systematic literature review. Australasian conference on information systems 2020 (pp. 1-11).
[12] Gorton I, Klein J. Distribution, data, deployment: software architecture convergence in big data systems. IEEE Software. 2014; 32(3):78-85.
[13] Interoperability NB. NIST big data public working group definitions and taxonomies subgroup in: framework: definitions. NIST Special Publication. 2015.
[14] Muchenje G, Seppänen M. Unpacking task-technology fit to explore the business value of big data analytics. International Journal of Information Management. 2023; 69:102619.
[15] Ntawiha W, Anderson B. Uganda's data ecosystem. 2016.
[16] Grover V, Chiang RH, Liang TP, Zhang D. Creating strategic business value from big data analytics: a research framework. Journal of Management Information Systems. 2018; 35(2):388-423.
[17] Gärtner B, Hiebl MR. Issues with big data. In the Routledge companion to accounting information systems 2017 (pp. 161-72). Routledge.
[18] Van DVHG, Klievink AJ, Arnaboldi M, Meijer AJ. Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making. Government Information Quarterly. 2019; 36(1):27-38.
[19] Marr B. Data strategy: how to profit from a world of big data, analytics and the internet of things. Kogan Page Publishers; 2017.
[20] Davenport TH, Dyché J. Big data in big companies. International Institute for Analytics. 2013.

[21] Mishra D, Gunasekaran A, Papadopoulos T, Childe SJ. Big data and supply chain management: a review and bibliometric analysis. Annals of Operations Research. 2018; 270:313-36.

[22] Barham H, Daim T. The use of readiness assessment for big data projects. Sustainable Cities and Society. 2020; 60:102233.

[23] Kaufmann M. Big data management canvas: a reference model for value creation from data. Big Data and Cognitive Computing. 2019; 3(1):19.

[24] Weibl J, Hess T. Success or failure of big data: insights of managerial challenges from a technology assimilation perspective. In MKWI 2018 (pp. 47-58).

[25] Alharthi A, Krotov V, Bowman M. Addressing barriers to big data. Business Horizons. 2017; 60(3):285-92.

[26] Spangenberg N, Wilke M, Franczyk B. A big data architecture for intra-surgical remaining time predictions. Procedia Computer Science. 2017; 113:310-7.

[27] Naganathan V. Comparative analysis of big data, big data analytics: challenges and trends. International Research Journal of Engineering and Technology. 2018; 5(05):1948-64.

[28] Li L, Ma S, Wang R, Wang Y, Zheng Y. Citizen participation in the co-production of urban natural resource assets: analysis based on social media big data. Journal of Global Information Management. 2021; 30(6):1-21.

[29] Zhuang YT, Wu F, Chen C, Pan YH. Challenges and opportunities: from big data to knowledge in AI 2.0. Frontiers of Information Technology & Electronic Engineering. 2017; 18:3-14.

[30] Jordan L. The problem with big data in translational medicine. a review of where we've been and the possibilities ahead. Applied & Translational Genomics. 2015; 6:3-6.

[31] Henninger M. The value and challenges of public sector information. Cosmopolitan Civil Societies: An Interdisciplinary Journal. 2013; 5(3):75-95.

[32] Lande D, Subach I, Puchkov A. A system for analysis of big data from social media. Information & Security. 2020; 47(1):44-61.

[33] Kim GH, Trimi S, Chung JH. Big-data applications in the government sector. Communications of the ACM. 2014; 57(3):78-85.

[34] Bassoo V, Ramnarain-seetohul V, Hurbungs V, Fowdur TP, Beeharry Y. Big data analytics for smart cities. Internet of Things and Big Data Analytics Toward Next-Generation Intelligence. 2018:359-79.

[35] Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. Journal of Big Data. 2019; 6(1):1-25.

[36] Mcdermott S, Turk MA. What are the implications of the big data paradigm shift for disability and health. Disability and Health Journal. 2015; 8(3):303-4.

[37] Elsheikh AS. A brief outlook of enterprise architecture role in the digital age. In proceedings of the federated Africa and middle east conference on software engineering 2022 (pp. 51-58). ACM.

[38] Sofyana L, Putera AR. Business architecture planning with TOGAF framework. In journal of physics: conference series 2019 (pp. 1-10). IOP Publishing.

[39] Saputra H, Gustiana I, Soeryanto SE. Design of enterprise information system using TOGAF framework for public service agency. Journal of Engineering Science and Technology. 2022; 17(4).

[40] KR RB. Design and development of QoS guaranteed cloud scheduling. Doctoral Dissertation, Cochin University of Science and Technology. 2019.

[41] Tamke M, Nicholas P, Zwierzycki M. Machine learning for architectural design: practices and infrastructure. International Journal of Architectural Computing. 2018; 16(2):123-43.

[42] Breivold HP. Internet-of-things and cloud computing for smart industry: a systematic mapping study. In 5th international conference on enterprise systems (ES) 2017 (pp. 299-304). IEEE.

[43] Peterson L, Al-shabibi A, Anshutz T, Baker S, Bavier A, Das S, et al. Central office re-architected as a data center. IEEE Communications Magazine. 2016; 54(10):96-101.

[44] https://www.nttdata.com/global/ja/-/media/nttdataglobal-ja/files/investors/library/ar/ar12_j_all.pdf. Accessed 28 June 2024.

[45] Moreno J, Serrano MA, Fernandez-medina E, Fernandez EB. Towards a security reference architecture for big data. In DOLAP 2018 (pp. 1-9).

[46] Velásquez N, Estevez EC, Pesado PM. Cloud computing, big data and the industry 4.0 reference architectures. Journal of Computer Science & Technology. 2018.

[47] Varghese B, Buyya R. Next generation cloud computing: new trends and research directions. Future Generation Computer Systems. 2018; 79:849-61.

[48] Nirmala MB, Raj P. Big data computing and the reference architecture. In big data: concepts, methodologies, tools, and applications 2016 (pp. 56-72). IGI Global.

[49] Singh N, Lai KH, Vejvar M, Cheng TC. Big data technology: challenges, prospects, and realities. IEEE Engineering Management Review. 2019; 47(1):58-66.

[50] García S, Ramírez-gallego S, Luengo J, Benítez JM, Herrera F. Big data preprocessing: methods and prospects. Big Data Analytics. 2016; 1:1-22.

[51] Becker DK. Predicting outcomes for big data projects: big data project dynamics (BDPD): research in progress. In international conference on big data (big data) 2017 (pp. 2320-30). IEEE.

[52] Baran ML. Mixed methods research design. In research anthology on innovative research methodologies and utilization across multiple disciplines 2022 (pp. 312-33). IGI Global.

[53] Xing W, Whalley J. The golden tax project, value-added tax statistics, and the analysis of internal trade in China. China Economic Review. 2014; 30:448-58.

[54] Saltz J, Shamshurin I, Connors C. A framework for describing big data projects. In business information systems workshops: BIS 2016 international

workshops, Leipzig, Germany 2017 (pp. 183-95). Springer International Publishing.
[55] Ibrahim SA, Hamzah N, Wahab AR, Abdullah JM, Malim NH, Sumari P, et al. Big brain data initiative in universiti sains Malaysia: challenges in brain mapping for Malaysia. The Malaysian Journal of Medical Sciences: MJMS. 2020; 27(4):1-8.
[56] Wieringa RJ. Design science methodology for information systems and software engineering. Springer; 2014.
[57] Bartram B. Using questionnaires. In practical research methods in education 2019 (pp. 1-11). Routledge.
[58] Ikart EM. Survey questionnaire survey pretesting method: an evaluation of survey questionnaire via expert reviews technique. Asian Journal of Social Science Studies. 2019; 4(2).
[59] Krosnick JA. Questionnaire design. The Palgrave handbook of survey research. 2018:439-55.
[60] Lund B. The questionnaire method in systems research: an overview of sample sizes, response rates and statistical approaches utilized in studies. VINE Journal of Information and Knowledge Management Systems. 2023; 53(1):1-10.
[61] Ong MH, Puteh F. Quantitative data analysis: choosing between SPSS, PLS, and AMOS in social science research. International Interdisciplinary Journal of Scientific Research. 2017; 3(1):14-25.
[62] Yang M. Big data: issues, challenges, tools. Thesis, Centria University of Applied Sciences. 2020.
[63] Alzahrani SM. Big data analytics tools: twitter API and spark. In international conference of women in data science at Taif university 2021 (pp. 1-6). IEEE.
[64] Hiebl MR. Sample selection in systematic literature reviews of management research. Organizational Research Methods. 2023; 26(2):229-61.
[65] Booth A, James MS, Clowes M, Sutton A. Systematic approaches to a successful literature review. Sage Publications; 2021.
[66] Lame G. Systematic literature reviews: an introduction. In proceedings of the design society: international conference on engineering design 2019 (pp. 1633-42). Cambridge University Press.
[67] Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Systematic Reviews. 2021; 10:1-9.
[68] Nti IK, Quarcoo JA, Aning J, Fosu GK. A mini-review of machine learning in big data analytics: applications, challenges, and prospects. Big Data Mining and Analytics. 2022; 5(2):81-97.
[69] Mohd SMH, Fitri PA, Ngalimat MS, Sim KJ. A mini literature review on current advancements in protein purification techniques. In biology and life sciences forum 2022 (pp. 1-10). MDPI.
[70] Bree RT, Gallagher G. Using microsoft excel to code and thematically analyse qualitative data: a simple, cost-effective approach. All Ireland Journal of Higher Education. 2016; 8(2).
[71] Gibbs GR. Analyzing qualitative data. Sage Publications; 2018.
[72] Prayag G. Post-positivism and mixed methods research in rural tourism. In handbook on tourism and rural community development 2023 (pp. 62-75). Edward Elgar Publishing.
[73] Didas M, Chali F, Elisa N, Matogoro J. Requirements for building a business-driven reference architecture for implementation of big data analytics by public sector organizations: a case study of Uganda. In first international conference on the advancements of artificial intelligence in African context 2023 (pp. 1-12). IEEE.
[74] Andrade C. The inconvenient truth about convenience and purposive samples. Indian Journal of Psychological Medicine. 2021; 43(1):86-8.
[75] Baltes S, Ralph P. Sampling in software engineering research: a critical review and guidelines. Empirical Software Engineering. 2022; 27(4):94.
[76] Walters WH. Survey design, sampling, and significance testing: key issues. The Journal of Academic Librarianship. 202; 47(3):102344.
[77] Sharma G. Pros and cons of different sampling techniques. International Journal of Applied Research. 2017; 3(7):749-52.
[78] Nanjundeswaraswamy TS, Divakar S. Determination of sample size and sampling methods in applied research. Proceedings on Engineering Sciences. 2021; 3(1):25-32.
[79] Chaokromthong K, Sintao N. Sample size estimation using Yamane and Cochran and Krejcie and morgan and green formulas and Cohen statistical power analysis by G* power and comparisions. Apheit International Journal. 2021; 10(2):76-86.
[80] Tamym L, Benyoucef L, Moh AN, El OMD. Big data analytics-based approach for robust, flexible and sustainable collaborative networked enterprises. Advanced Engineering Informatics. 2023; 55:101873.
[81] Shi Y. Advances in big data analytics. Adv Big Data Anal; 2022. Springer.
[82] Watson HJ. Tutorial: big data analytics: concepts, technologies, and applications. Communications of the Association for Information Systems. 2014; 34(1):65.
[83] Fredriksson C, Mubarak F, Tuohimaa M, Zhan M. Big data in the public sector: a systematic literature review. Scandinavian Journal of Public Administration. 2017; 21(3):39-61.
[84] Kiganda C, Akcayol MA. Forecasting the spread of COVID-19 using deep learning and big data analytics methods. SN Computer Science. 2023; 4(4):374.
[85] Coyne EM, Coyne JG, Walker KB. Big Data information governance by accountants. International Journal of Accounting & Information Management. 2018; 26(1):153-70.
[86] Joubert A, Murawski M, Bick M. Measuring the big data readiness of developing countries–index development and its application to Africa. Information Systems Frontiers. 2023; 25(1):327-50.
[87] Arora B. Big data analytics: the underlying technologies used by organizations for value

generation. Understanding the Role of Business Analytics: Some Applications. 2019:9-30.
[88] Sang GM, Xu L, De VP. Simplifying big data analytics systems with a reference architecture. In collaboration in a data-rich world: 18th IFIP WG 5.5 working conference on virtual enterprises, Vicenza, Italy, 2017 (pp. 242-9). Springer International Publishing.

Matendo Didas received his Certificate and Diploma in Computer Science from Makerere University Business School and Nakawa Institute of Commercial Studies, Uganda, respectively. He holds Cisco Certified Networking Associate (CCNA) and Cisco Certified Networking Professional (CCNP) certifications from Makerere University, Uganda. He earned his BSc in Software Engineering from Makerere University, Uganda, and holds an MSc in Information Technology from the University of Dodoma, Tanzania, as well as an MSc in Information Systems from Makerere University, Uganda. He specializes in Artificial Intelligence, Blockchain Technology, Big Data Analytics, and Metaverse Technology Management. Currently, he is an Assistant Lecturer at St. John's University of Tanzania and a remote AI researcher at AI4D, Tanzania.
Email: matendodidas@gmail.com

Frederick Henri Chali received his BSc in Information and Communication Technology Management from Mzumbe University, Tanzania, in 2007 and his MSc in Computer Science from the University of Dodoma, Tanzania, in 2011. He holds a Ph.D. in Information Communication Systems and Engineering (ICSE), specializing in Information Technology Development Systems and Management (ITDSM), from the School of Computational and Communication Systems and Engineering (CoCSE) at The Nelson Mandela African Institution for Science and Technology (NM-AIST). He is currently a Lecturer in the Department of Information Systems at the School of Informatics, part of the College of Informatics and Virtual Education (CIVE), at the University of Dodoma. His research interests include Mobile Computing, Web Technologies, and Data Mining.
Email: chali.frederick@gmail.com

Noe Elisa is a former Ph.D. researcher at Northumbria University, Newcastle upon Tyne, UK. He holds a Master of Technology in Computer Networks and Information Security, with a major in Distributed Systems Security, from Jawaharlal Nehru Technological University, India, and a BSc in Telecommunication Engineering from the University of Dar es Salaam (UDSM), Tanzania. He is currently a Lecturer in the Department of Information Systems at the School of Informatics, part of the College of Informatics and Virtual Education (CIVE), at the University of Dodoma. Noe is an experienced researcher and engineer in cybersecurity, artificial intelligence, and data science, with over six years of expertise in network and web security. His specialization includes applying AI and data science to cybersecurity challenges, particularly in the design and implementation of network intrusion detection systems. He also has experience in leveraging blockchain technology to design and develop secure and privacy-preserving distributed systems.
Email: noelisa1983@gmail.com

Appendix I

S.No.	Abbreviation	Description
1	ACM	Association For Computing Machinery
2	ADB	Analytical Databases
3	ADSs	Assistance Driving Systems
4	AI	Artificial Intelligence
5	AM	Additive Manufacturing
6	BD	Big Data
7	BDA	Big Data Analytics
8	BI	Business Intelligence
9	BPMN	Business process Modeling Notation
10	BRABDAI	Business-driven Reference Architecture for Big Data Analytics Implementation
11	CDB	Column Databases
12	CEAS	Citizen's Engagement and Analysis System
13	CEP	Complex Event Processing
14	CNN	Cable News Network
15	COVID-19	Corona Virus Disease Nineteen
16	D	Design
17	DB	Database
18	DD	Design Decisions
19	DDB	Document Databases
20	DLc	Data Lifecycle
21	DSRM	Design Science Research Methodology
22	EMC	Electromagnetic Compatibility
23	ETL	Extract, Transformation, and Loading
24	FT	Flue Trend
25	GCP	Google Cloud Platform
26	GDB	Graph Databases
27	GRA	Google Reference Architecture
28	GTS	China's Golden Tax System
29	HCD	Human Centered Design
30	HDFS	Hadoop Distributed File System
31	IEEE	Institute of Electrical and Electronics Engineers
32	IT	Information Technology
33	ICT	Information and Communication Technology
34	JSTOR	Journal Storage
35	KPI	key performance Indicators
36	KVDB	Key-Value Databases
37	LR	Literature Review
38	M	Managerial
39	MARA	Microsoft Azure Reference Architecture

40	ML	Machine Learning
41	MoES	Ministry of Education and Sports
42	MoH	Ministry of Health
43	NDRA	NTTData Reference Architecture
44	NHS	National Health Service
45	NIST	National Institute of Science and Technology
46	NIST	NIST Reference Architecture
47	NoSQL	Not Only Structured Query Language
48	NoSQLDB	Not Only Structured Query Language Database
49	NRRA	Reference Architecture by Nirmala & Raj
50	NTT	Nippon Telegraph and Telephone
51	PII	Personal Identification Information
52	PoC	Proof of Concept
53	PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
54	R	Requirements
55	RA	Reference Architecture
56	RAs	Reference Architectures
57	RGDP	Regulation for General Data Protection
58	ROI	Return on Investment
59	RO	Research Objective
60	RQ	Research Question
61	SAP	Systems Applications and Products
62	SMS	Simple Message System
63	TOGAF	The Open Group Architecture Framework
64	UBOS	Uganda Bureau of Statistics
65	UK	United Kingdom
66	UML	Unified Modeling Language
67	USA	United States of America
68	VAT	Value Added Tax