# Secure big data ecosystem architecture: challenges and solutions

Memoona J. Anwar[1*], Asif Q. Gill[1], Farookh K. Hussain[1] and Muhammad Imran[2]

*Correspondence:
memoona.j.anwar@uts.
edu.au
[1] School of Software,
Faculty of Engineering,
and Information Technology,
The University of Technology
Sydney, Ultimo, NSW,
Australia
Full list of author information
is available at the end of the
article

## Abstract

Big data ecosystems are complex data-intensive, digital–physical systems. Data-intensive ecosystems offer a number of benefits; however, they present challenges as well. One major challenge is related to the privacy and security. A number of privacy and security models, techniques and algorithms have been proposed over a period of time. The limitation is that these solutions are primarily focused on an individual or on an isolated organizational context. There is a need to study and provide complete end-to-end solutions that ensure security and privacy throughout the data lifecycle across the ecosystem beyond the boundary of an individual system or organizational context. The results of current study provide a review of the existing privacy and security challenges and solutions using the systematic literature review (SLR) approach. Based on the SLR approach, 79 applicable articles were selected and analyzed. The information from these articles was extracted to compile a catalogue of security and privacy challenges in big data ecosystems and to highlight their interdependencies. The results were categorized from theoretical viewpoint using adaptive enterprise architecture and practical viewpoint using DAMA framework as guiding lens. The findings of this research will help to identify the research gaps and draw novel research directions in the context of privacy and security in big data-intensive ecosystems.

**Keywords:** Big data, IoT, Big data ecosystem, Privacy and security, Big data ecosystem challenges, Big data ecosystem privacy and security solutions

## 1 Introduction

The history and relevance of big data can be backtracked to the origin of the Internet. The Internet can be considered as a global network of machines comprising data and applications. However, the term "big data" was first used in 1999 in an academic paper, which led to further detailed characterization of big data in 2003 [1]. As the volume of data was gradually increasing, this resulted in the emergence of open-source big data technologies and applications, such as Apache Hadoop in 2005. According to a report by the next frontier for innovation, competition and productivity by the McKinsey Global Institute, a typical US company with 1,000 employees could store 200 terabytes of data by 2009 [2]. In 2011, RFID came into existence and the global market doubled in 2013. The frequent use of mobile devices by businesses and consumers resulted in the explosion of the volume of daily data [3, 4]. According to McKinsey Global Institute

Anwar *et al. J Wireless Com Network*     (2021) 2021:130

Page 2 of 30

(2011) [2], the data volume rises by 40% per year, and it is expected to increase 44 times at this rate between 2009 and 2020 [5]. Data is produced from devices such as mobile phones, satellites and other sensors used in different industries such as healthcare [6–8]. The interconnection between large number of diverse devices generates massive data [9, 10]. More data was generated in the last two years than in the entire human history before that [11]. Obtaining, incorporating, handling and utilizing IoT for these data sets are becoming urgent and vital research problems for organizations to meet their aims [12–14].

With increase in demand for digital transformation, many businesses seek to benefit from the technological developments by employing smart devices. The amount of data produced by these devices is massive and requires real-time processing. It is referred to as big data (BD), a terminology used for huge data sets which have big and diverse formats [15]. This inherent nature of big data makes it challenging for traditional methods of storage, analysis and visualization of data for advance processing [16]. There are many definitions of big data. Madden (2015) defines BD as "data that's too big, too fast, or too hard for existing tools to process" [17]. In spite of being complex and hard to handle, big data is of utmost importance for most businesses in terms of making effective decisions and accomplishing their goals [18]. Beer (2016) stated that "big data is a concept that has achieved a profile and vitality that very few concepts attain" [19].

In today's smart world, data is continuously generated due to increase in number of connected devices. IoT and advancement in technology have made people more connected than they were in the past. Digital footprints may be considered as personally identifiable information (PII) as it has the potential to identify the individuals. This points to the most pressing challenge, the need to maintain individual confidentiality in the BD environment [20]. Privacy and security, among many challenges associated with BD, directly affect the persons involved. Some "technologists argue that privacy is dead" [21]. Due to the volume and sensitivity of the stored data, breach incidents with big data can result in extremely upsetting consequences. This is due to larger number of people getting affected by it.

In BD, "data are rather a "fuel" that "powers" the whole complex of technical facilities and infrastructure components built around a specific data origin and their target use" [19]. This complex is called big data ecosystem (BDE). Ecosystems are complex systems of networked organizations dealing with different sizes, speeds and types of big data. A BDE is a collection of interrelated components that work together toward the evolution and improvement of data, models and associated infrastructure, not just at one stage but throughout the entire life cycle of data [19]. The objective of any BD project is to positively impact the business. Though BD seems to offer several benefits to organizations, it also presents challenges due to the implicit nature of the procedures done at the time of collection, storage and use. Research challenges in the BDE continue throughout the data life cycle, starting right at data creation and leading to problems with storage and transport, conversion and processing, and lastly usage and destruction [22]. It has been observed that serious privacy and security conditions were encountered when dealing with collection, storage and usage of huge volumes of data [17, 23, 24]. Understanding the architecture of BDE and how interactions take place within it, is critical to boosting BD investment.

Anwar *et al. J Wireless Com Network*    (2021) 2021:130

Page 3 of 30

There are not many academic papers on BD; in most instances, they are concentrating on a specific technology (such as data analytics, artificial intelligence and machine learning) or solution that mirrors only a fraction of the entire problem. To the best of our knowledge, no journal article systematically reviews the literature from the perspective of identifying big data challenges, solutions and their interdependency. Hence, we address this critical issue by conducting a systematic literature review (SLR) to synthesize and draw attention toward noteworthy challenges of big data. This paper uncovers the important concerns of privacy and security. The results of this research aim to present a unified view of the BD related challenges, solutions and their interdependencies to initiate a broad discussion rather than specific to a smaller context or technology. This will help industry practitioners to design secure BDE architecture keeping in mind the challenges and leveraging existing solutions. It will also help researchers to further explore the challenge areas where there is no solution. Therefore, the scope of this research is limited to the two questions in Table 1.

This paper is organized as follows. Firstly, it presents the research method used in carrying out the SLR. Secondly, it presents the research results. In the end, it discusses the research results and concludes with opportunities for further research.

## 2  Research method

Based on the guidelines suggested by Kitchenham and Charters (2007) [25], an SLR approach was used to address the research questions in hand (see Table 1). The SLR approach is helpful to ensure that the related articles are not overlooked. In this paper, we organize our SLR into the following steps:

*Step 1: Development of Review Protocol*

In this step, the research questions are formulated, data sources are identified, and search terms are defined.

*Step 2: Inclusion and Exclusion criteria*

The most relevant studies are extracted based on certain criteria.

*Step 3: Study Selection Process*

Studies for inclusion in the SLR are selected at this step.

*Step 4: Data Extraction*

After reviewing the selected studies, data is extracted and recorded.

**Table 1** Research questions addressed in this manuscript and their motivations

| Research question | Motivation |
| --- | --- |
| *RQ1*: What are the key privacy and security challenges in big data ecosystems? | To discover the privacy and security related challenges, which are affecting big data ecosystems the most |
| *RQ2*: What are the future research directions in relation to the privacy and security of big data ecosystems? | To explore existing models, techniques, frameworks and methodologies to make big data more secure. To highlight the measures that can be taken in the future to make big data ecosystems more secure and trustworthy |

## 3  Step 5: data synthesis

Review results are presented.

### 3.1  Step 1: development of review protocol

The following electronic scientific databases were selected and used to collect literature for this review.

1. IEEE Xplore (www.ieexplore.ieee.org/Xplore/)
2. Elsevier ScienceDirect (www.sciencedirect.com/)
3. Google Scholar (www.scholar.google.com.au/)
4. ProQuest Science and Technology (www.proquest.com/)
5. Scopus

These well-known databases were selected because they provide sufficient relevant literature coverage for the review. After the first search round, the identified literature (175 studies) was filtered by a thorough review process to identify the relevance of each article to the research questions. Table 2 summarizes the search categories and keywords used to search for relevant literature. We used the search strings "big data ecosystem", "big data challenges", "privacy of big data ecosystems" and "security of big data ecosystems" based on the research questions in hand. We searched different combinations of items from all the search categories using the Boolean operator "AND" to retrieve the relevant studies. Similar terms are linked using the "OR" operator to obtain maximum coverage. The search statement is divided into two key parts. The union of big data-related terms makes up the first sub-part. The second sub-part is the union of big data ecosystem challenges and big data tools. The second sub-part includes interdisciplinary terms that are primarily related to the privacy and security of big data in an ecosystem. As a result, to obtain the union of results, the combination of these two sections is applied an "OR" Boolean operator. Consequently, the following search string is generated:

[("Iot" **OR** "Smart buildings" **OR** "Massive Data" **OR** "Big Data" **OR** "Massive Data Storage" **OR** "Data Formats" **OR** "Data Analytics" **OR** "Predictive Analytics" **OR** "Decision Making" **OR** "Actionable Knowledge" **OR** "Open Data" **OR** "Big Data Architecture Framework" **OR** "Big Data Lifecycle") **AND** ("Digital Physical Ecosystem" **OR** "Cloudera" **OR** "Hadoop" **OR** "Real-time Data Analytics" **OR** "Privacy Ecosystem" **OR** "Big Four" **OR** "e-Governance" **OR** "Digital Governance" **OR** "Open Governance" **OR** "Openness")]**OR**("V's Big Data Characteristics" **OR** "Big Data Structure" **OR** "Big Data

**Table 2** Search categories and keywords used in SLR

| Search category | Keyword |
| --- | --- |
| Big data sources | IoT smart buildings, massive data, predictive analytics, data processing, decision making, actionable knowledge, data formats, relative data |
| Big data ecosystem | Digital physical ecosystem, Cloudera, Hadoop, real-time data analytics, eco-analytics, privacy ecosystems, big four |
| Big data ecosystem challenges | V's big data characteristics, storage, structures, speed, value |
| Security and privacy | e-governance, e-government, digital governance, open data, openness, transparency, privacy, security, trust, data anonymization, differential privacy, notice and consent, information security ecosystem, third part intermediary, privacy-security trade off, authorization, access control, information security |

Speed" **OR** "Value" **OR** "Transparency" **OR** "Privacy" **OR** "Security" **OR** "Trust" **OR** "Data Privacy" **OR** "Anonymization" **OR** "Differential Privacy" **OR** "Notice and Consent" **OR** "Information Security Ecosystem" **OR** "Privacy Security Trade-off").

### 3.2 Step 2: inclusion and exclusion criteria

To review the most relevant studies, a certain criterion is used to decide if a certain study will be added to the review process or not. The inclusion or exclusion of a study was based on the factors listed below.

1. Does the article have a segment that argues about four, defined search categories (Table 2): Big data sources, big data ecosystem, big data ecosystem challenges, and security and privacy?
2. Is it an academic, experimental or commercial project?
3. Does the article date between 2014 and 2019?
4. Is the full text available and written in English?
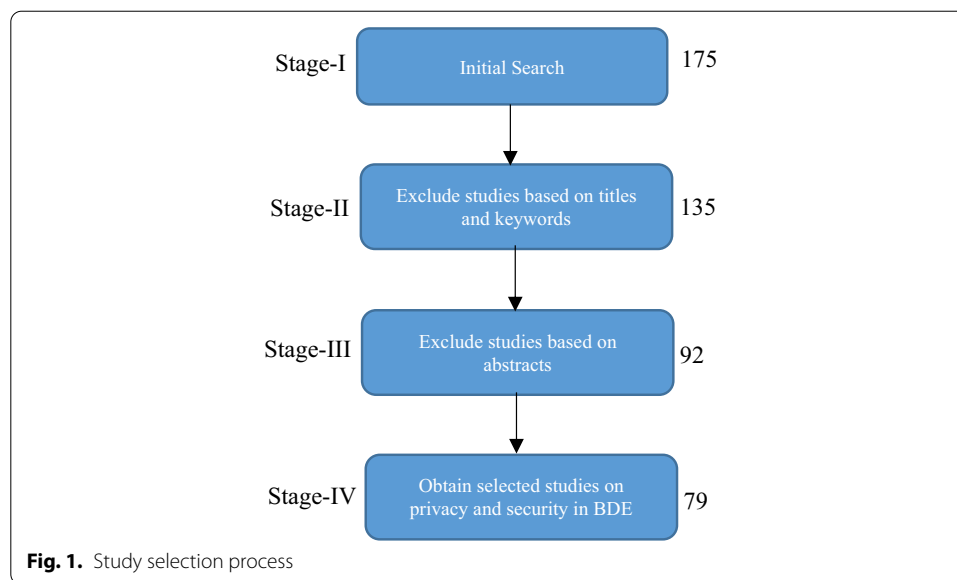5. Is the article peer reviewed?

Studies that either do not answer the research question or meet the following exclusion criteria are excluded from this study.

1. Papers not in English
2. Magazines, newspapers, websites, podcasts, blogs and wire feeds
3. Duplicated studies

### 3.3 Step 3: study selection process

Endnotes were used to store the related references from step1.These references were then exported to Microsoft Word in tabular form, containing detail of each reference and relevant inclusion/exclusion decision was noted. Endnote databases and word tables were individually created for every step of the selection process. To start with, all studies are classified from the investigated databases based on all keywords. During the second stage of the selection process, the titles of all 175 papers were reviewed and studies that were clearly not relevant were excluded. Some titles were not very clear in terms of whether the article should be included or excluded at this stage; therefore, they were pushed to the next stage for assessment. At stage 3, the studies were filtered based on a review of the abstract. Finally, only 79 articles were deemed to be aligned to our research topic and suitable for detailed review, as shown in Fig. 1. Table 3 summarizes each filtration stage.

This study targeted academic publications from 2014 to 2019. The reason for including studies between 2014 and 2019 is due to the rise in mobile machine usage during this period which caused a huge surge in volume of data [6]. The search excluded articles that included non-academic surveys, news, article summaries and discussions. This was done to ensure the relevance and academic quality of this research. An initial search resulted in 235 hits across all the selected databases. After eliminating the irrelevant and identical studies, the number dropped to 175. All the studies found during the first stage of the

Anwar *et al. J Wireless Com Network*     (2021) 2021:130

Page 6 of 30



**Fig. 1.** Study selection process

**Table 3** stages of filtration for selecting relevant papers for SLR

| Filtration stage | Method | Assessment criteria |
|---|---|---|
| Stage-I | Classify related studies from investigated databases | All keywords inclusive (N = 175) |
| Stage-II | Eliminate studies based on titles | Title = search term<br>Include else exclude (N = 135) |
| Stage-III | Eliminate studies based on abstracts | Abstract = big data ecosystem challenges<br>Include else exclude (N = 92) |
| Stage-IV | Retrieve and critically evaluate selected studies | Address security and privacy in big data ecosystem (N = 79)<br>Yes = accepted<br>No = rejected |

search are evaluated to determine whether they should be part of the literature review or not. The reasons for inclusion and exclusion are documented, and a list of inclusion and exclusion criteria is created and discussed (see Step 2).

Sources such as journals, conference proceedings and the Internet are also used because lone digital libraries may not provide basis for a complete SLR. Therefore, in this research, five electronic resources were explored. Only those resources which are easily accessible and widely available were included in this research. Studies were included if their motivation was about BD challenges, mainly security and privacy challenges, and if they presented empirical data. This study includes papers not older than 2014, industry, government and academic experimental software projects and all sized business ecosystems. The included papers are those which were published in English only. Table 4 lists the articles that were included from different databases after each filtration.

### 3.4  Step 4: data extraction
We analyzed the final set of 79 papers and extracted the big data challenges and solutions. To accurately extract the most relevant information from each of the selected papers, the papers were evaluated using several quality assessment criteria to ensure that

**Table 4** Count of search results for the SLR from scientific database

| Database | 1st filtration | 2nd filtration | 3rd filtration | Studies selected | Percent selected |
|---|---|---|---|---|---|
| IEEE Xplore | 51 | 31 | 26 | 25 | 31.6% |
| ProQuest | 45 | 39 | 31 | 27 | 34% |
| Google Scholar | 39 | 35 | 14 | 11 | 14% |
| Science Direct | 25 | 17 | 15 | 13 | 16% |
| Scopus | 15 | 13 | 6 | 3 | 3.7% |
| Total | 175 | 135 | 92 | 79 | 100 |

**Table 5** Quality assessment criteria to evaluate the final 79 papers

| Quality assessment criteria |
|---|
| 1. Does the study set up an appropriate context for the related research? |
| 2. Did the study details suitable research method to achieve its aims? |
| 3. How suitably it answers the research question in hand? |
| 4. Are the results and applications described and discussed thoroughly? |
| 5. Does the study provide clear findings with justifiable results and conclusions? |
| 6. Does the study provide future directions? |

the selection of the final publications was not biased. Table 5 lists the quality assessment criteria that were used in this study.

Each selected paper was analyzed and a score of 1–6 was assigned across each of the seven criteria. A score of six (6) denotes high applicability to the research question and one (1) signifies low applicability. These criteria were used to appraise the quality of studies included. If a study meets all six criteria, it is suitable for our literature review. Each of the six criteria was given a score of "1" or "0" where 1 means yes and 0 means no. The results are shown in Table 6.

We only included research papers in our study; hence, each paper is assigned a score of "1" in the research column. Twenty-three studies did not mention any clear research aim. Most of the studies included a brief description of the context in which they were conducted. Twenty-three studies failed to meet the relevance criteria. Eleven studies did not clearly describe what they discovered from the research, and twelve of the reviewed studies did not provide future research directions. Fourteen studies received a score of six as they met all six criteria. Overall, each study obtained a score greater than or equal to three which was deemed to be an acceptable level of quality for this research.

### 3.5 Step 5: data extraction

All the data collected from the chosen articles was synthesized in tabular form against each of the two research questions. An extensive investigation of data extracted from selected literature led to the classification of different categories related to challenges or solution concept.

The results of this SLR were categorized from industry as well as theoretical viewpoint. We used the adaptive enterprise architecture (EA) framework for digital ecosystems [26] (Fig. 2) as a theoretical framework and DAMA [27] (Table 7) as a practical industry framework to derive and code the categories used for structuring, analyzing and synthesizing the results of this study. Adaptive EA covers the entire

Anwar *et al. J Wireless Com Network*    (2021) 2021:130

Page 8 of 30

**Table 6** Assessment of selected 79 articles against the quality assessment criteria

| Study | Research | Aim | Context | Relevance | Findings | Future Direction | Total |
|-------|----------|-----|---------|-----------|----------|------------------|-------|
| s1 | 1 | 1 | 1 | 0 | 1 | 1 | 5 |
| s2 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s3 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s4 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s5 | 1 | 1 | 0 | 1 | 1 | 1 | 5 |
| s6 | 1 | 1 | 0 | 1 | 1 | 1 | 5 |
| s7 | 1 | 0 | 0 | 1 | 1 | 0 | 3 |
| s8 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s9 | 1 | 1 | 0 | 1 | 1 | 1 | 5 |
| s10 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s11 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| s12 | 1 | 1 | 0 | 1 | 1 | 1 | 5 |
| s13 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s14 | 1 | 1 | 1 | 1 | 0 | 1 | 5 |
| s15 | 1 | 0 | 1 | 1 | 1 | 1 | 5 |
| s16 | 1 | 1 | 0 | 0 | 0 | 1 | 3 |
| s17 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| s18 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s19 | 1 | 1 | 0 | 1 | 0 | 1 | 3 |
| s20 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s21 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s22 | 1 | 1 | 0 | 1 | 1 | 0 | 4 |
| s23 | 1 | 1 | 0 | 1 | 1 | 0 | 4 |
| s24 | 1 | 1 | 0 | 0 | 1 | 1 | 4 |
| s25 | 1 | 1 | 1 | 1 | 1 | 0 | 4 |
| s26 | 1 | 1 | 1 | 1 | 0 | 1 | 5 |
| s27 | 1 | 1 | 0 | 1 | 1 | 1 | 6 |
| s28 | 1 | 0 | 1 | 0 | 1 | 1 | 4 |
| s29 | 1 | 1 | 1 | 0 | 1 | 1 | 5 |
| s30 | 1 | 0 | 1 | 0 | 1 | 1 | 4 |
| s31 | 1 | 0 | 1 | 1 | 0 | 1 | 4 |
| s32 | 1 | 1 | 0 | 1 | 1 | 1 | 5 |
| s33 | 1 | 1 | 0 | 1 | 1 | 1 | 5 |
| s34 | 1 | 0 | 0 | 1 | 1 | 0 | 3 |
| s35 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s36 | 1 | 1 | 0 | 1 | 0 | 1 | 4 |
| s37 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s38 | 1 | 1 | 1 | 0 | 1 | 0 | 4 |
| s39 | 1 | 0 | 1 | 0 | 1 | 1 | 4 |
| s40 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s41 | 1 | 1 | 0 | 0 | 0 | 1 | 3 |
| s42 | 1 | 0 | 1 | 1 | 1 | 1 | 5 |
| s43 | 1 | 1 | 0 | 0 | 0 | 1 | 3 |
| s44 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| s45 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s46 | 1 | 0 | 1 | 0 | 0 | 1 | 3 |
| s47 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| s48 | 1 | 1 | 0 | 1 | 1 | 1 | 5 |
| s49 | 1 | 0 | 1 | 1 | 1 | 0 | 4 |

**Table 6** (continued)

| Study | Research | Aim | Context | Relevance | Findings | Future Direction | Total |
|-------|----------|-----|---------|-----------|----------|------------------|-------|
| s50 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| s51 | 1 | 0 | 1 | 0 | 1 | 1 | 4 |
| s52 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| s53 | 1 | 1 | 1 | 1 | 0 | 1 | 5 |
| s54 | 1 | 0 | 0 | 1 | 1 | 1 | 4 |
| s55 | 1 | 0 | 1 | 1 | 0 | 1 | 4 |
| s56 | 1 | 1 | 0 | 1 | 1 | 1 | 5 |
| s57 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |
| s58 | 1 | 0 | 0 | 0 | 1 | 0 | 2 |
| s59 | 1 | 0 | 0 | 0 | 1 | 1 | 3 |
| s60 | 1 | 1 | 0 | 0 | 1 | 1 | 4 |
| s61 | 1 | 1 | 0 | 1 | 0 | 1 | 4 |
| s62 | 1 | 1 | 1 | 0 | 1 | 0 | 4 |
| s63 | 1 | 1 | 0 | 0 | 1 | 0 | 3 |
| s64 | 1 | 0 | 0 | 1 | 0 | 0 | 2 |
| s65 | 1 | 1 | 1 | 1 | 0 | 1 | 5 |
| s66 | 1 | 0 | 1 | 1 | 0 | 1 | 4 |
| s67 | 1 | 0 | 0 | 1 | 1 | 1 | 4 |
| s68 | 1 | 0 | 1 | 0 | 1 | 1 | 4 |
| s69 | 1 | 1 | 0 | 1 | 0 | 0 | 3 |
| s70 | 1 | 1 | 0 | 0 | 0 | 0 | 2 |
| s71 | 1 | 0 | 1 | 1 | 0 | 1 | 4 |
| s72 | 1 | 1 | 1 | 0 | 1 | 0 | 4 |
| s73 | 1 | 1 | 1 | 0 | 1 | 1 | 5 |
| s74 | 1 | 0 | 0 | 1 | 1 | 0 | 3 |
| s75 | 1 | 0 | 1 | 1 | 0 | 1 | 4 |
| s76 | 1 | 1 | 0 | 1 | 1 | 1 | 5 |
| s77 | 1 | 0 | 0 | 1 | 0 | 1 | 3 |
| s78 | 1 | 1 | 1 | 0 | 1 | 0 | 4 |
| s79 | 1 | 1 | 1 | 1 | 1 | 1 | 6 |
| Total | 79 | 56 | 48 | 56 | 59 | 57 | |

ecosystem layers, whereas DAMA is related to data/information only. The focus of this SLR is on big data ecosystem which is a combination of data and ecosystem layers.

Based on a comprehensive literature review, this study synthesizes past research efforts in the privacy and security of BD and identifies important factors that influence the development of a secure BDE (e.g., challenges, solutions and their interdependencies). This approach is suitable for categorizing our results as it provides adequate coverage of the privacy and security challenges and available solutions to those challenges as well as their interdependencies. Furthermore, a frequency analysis of each category was conducted to identify the strength and trend of research interest in that area. For example, the most important challenge of privacy and security was discussed by 78% of the total number of studies reviewed in this paper.

**Fig. 2.** Adaptive EA theoretical framework—selected categories
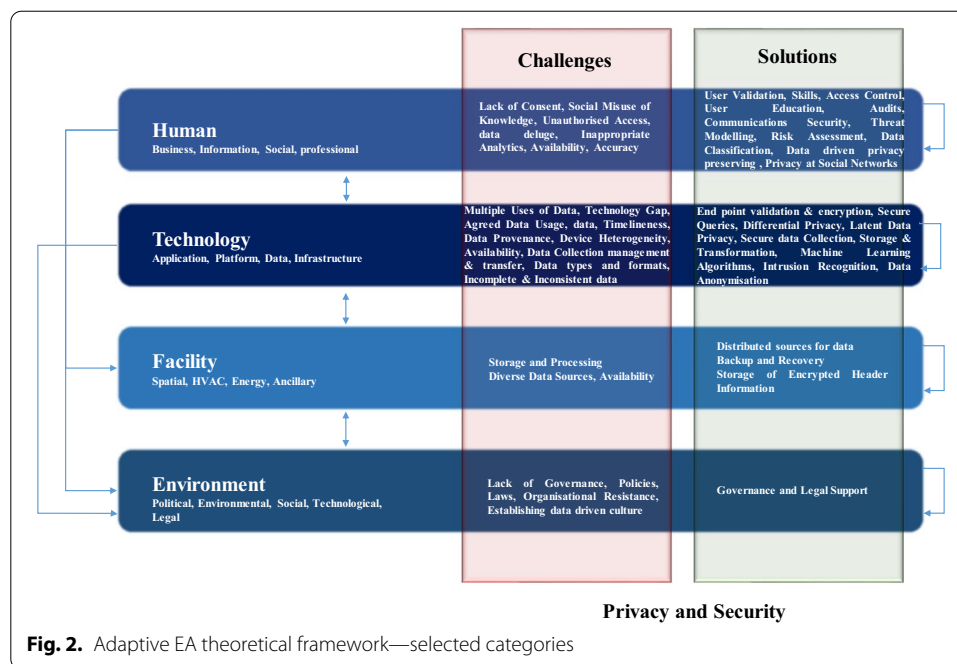
**Table 7** Assessment of selected 79 articles against the quality assessment criteria

| DAMA knowledge area | Challenges | Solutions |
|---|---|---|
| Data governance | Privacy at social networks | Policies, laws and governance |
| Data architecture | Timeliness, device heterogeneity | Architecture security |
| Data modeling and design | Data deluge, inappropriate analytics, technology gap | Communication security, threat modeling/risk assessment, intrusion recognition |
| Data storage and operations | Data storage and sharing, data collection, availability | Endpoint validation and response capabilities, secure queries, secure data collection, storage and transaction logs, distributed sources for data, data transformation, machine learning algorithms, data classification |
| Data security | Privacy and security, confidentiality, lack of skills | Encryption, user validation, confidentiality, skills, differential privacy, latent data privacy, data-driven privacy preserving |
| Data integration and interoperability | Data capture, data transfer | |
| Documents and contents | Complexity, different data types and formats | Granular access control, data anonymization, granular audit, user education |
| Reference and master data | | Recovery |
| Data warehousing and business intelligence | Establishing a data-driven culture, cost/budget, organizational resistance | |
| Metadata | | Storage of encryption header information, provenance metadata |
| Data quality | Accuracy, incompleteness and inconsistency, vagueness | |

Anwar *et al. J Wireless Com Network*     (2021) 2021:130

Page 11 of 30

## 4 Results

This section presents the SLR results. We carefully selected and reviewed 79 studies (s1–s79) and identified 21 key privacy and security challenges relevant to big data ecosystems using a SLR approach. Most of the articles provide coverage for all facets of the research questions (challenges, privacy and security, solutions). The SLR carried out for this research covers all the key research outlets. A summary of the chosen publications as per the publication channel shows that most of the papers (47) were taken from journals and a rest were from conference proceedings (32). We used the well-known DAMA framework [27] as an industry guiding lens to analyze, identify and group the challenges from the selected studies (see Table 7). The final version of DAMA DMBOK was presented in 2014. This is one of the reasons to select studies not older than 2014. Similar to other fields such as Information Technology (ITIL, COBIT, CMMi, and ISO17799), the DAMA framework comprised of the 11 knowledge areas as discussed in the paper. The reason for using DAMA framework is that it offers practical guidelines for implementing data management. DAMA provides coverage around "WHAT, WHO and WHY of data management and its various knowledge areas" [27]. In addition, it describes how data management knowledge areas are described in industry, what terminologies are used and what are tried and tested industry best practices. The focus of this study is BDE; therefore, the results are further discussed as per categories derived from DAMA. We have also chosen adaptive EA [26] as a theoretical lens, which discusses the end-to-end ecosystem architecture layers (human, technology, facility and environment). Adaptive EA is about vital "elements (concepts or properties) of integrated adaptive human (BIPS: business, information, professional, social), technology (ADPI: application, data, platform, infrastructure) and facility (SEHA: spatial, energy, HVAC, ancillary) system or ecosystem (value network of systems) in its secure environment (PESTLE: political, economic, sociological, technological, legal and environment), relationships (type, strength), and the principles (adaptive design) and evolution" [28]. The rationale for choosing adaptive EA framework is that it is an overarching framework consisting of important layers of BDE (human, technology, facility, environment, interaction and privacy). These layers are connected to each other via interaction layer. These all layers have an overarching security layer. The results of this study are categorized according to adaptive architecture in Fig. 2.

All the challenges identified in Table 8 are important; however, it is clear from the analysis of the selected studies (see Table 6) that the most significant challenges facing BDE are related to privacy and security. We further analyzed the privacy and security challenges within BDE and their relevant solutions in detail, as this is our area of study and scope of this paper.

Table 8 summarizes the challenges of BDE, their frequency and relevant data sources. One of the challenges highlighted is accuracy. BD analysis does not always produce accurate results; hence, the challenge of accuracy arises in BDE. A fact learned as a result of this SLR is that BD is more around complexity rather than size. The complexity of data must be taken into account, particularly when there are diverse data sources [29]. In addition to this, a major challenge in BDE is dealing with large data sets. Every two years, the size of data in storage systems across the world is getting doubled [30]. Not only is data growth rapid, but its arrival speed is also fast too. Higher speed of incoming data

**Table 8** Detailed analysis of the big data ecosystem challenges and their frequency and distribution

| DAMA knowledge area | Challenge | Study | Frequency | Percentage |
|---|---|---|---|---|
| Data architecture | Timeliness | s4, s5, s9, s12, s13, s15, s19, s24, s28, s37, s38, s40, s44, s47 | 14 | 18 |
| | Device heterogeneity | s14, s22, s30, s41 | 4 | 5 |
| Data modeling and design | Data deluge | s11, s12, s13, s17, S23, s24, s27, s28S37, s38, s44, s46, s54 | 13 | 16 |
| | Inappropriate analytics | s6, s8, s13, s15, s18, s21, s23, s24, s44 | 9 | 11 |
| | Technology gap | s6, s13, s21, s23, s24, s36, s37, s38, s47 | 9 | 11 |
| Data storage and opera-tions | Data storage and sharing | s4, s12, s13, s17, s21, s23, s24, s26, s28, s36, s38 | 11 | 13 |
| | Data collection | s2, s4, s8, s12, s13, s18, s23, s24, s28, s68 | 9 | 11 |
| | Availability | s3, s7, s8, s9, s11, s14, s19, s21, s22, s26, s28, s38, s39, s48, s54, s57 | 16 | 20 |
| Data security | **Privacy and security** | **s2, s4, s5, s7, s9, s11, s13, s14, s16, s17, s18, s20, s21, s22, s24, s26, s28, s29, s30, s31, s37, s38, s39, s42, s44, s48, s49, s52, s53, s55, s56, s57, s59, s61, s62, s64, s66s67, 69, s70, s71, s73, s74, s78, s79** | **45** | **53** |
| | **Confidentiality** | **s9, s16, s21, s30, s41, s74** | **6** | **8** |
| | **Lack of skills** | **s5, s9, s13, s15, s18, s22, s23, s27, s34, s37** | **10** | **17** |
| Data integration and interoperability | Data management | s23, s28, s29, s30, s36, s37 | 6 | 8 |
| | Data transfer | s4, s19, s23, s28, s37, s41, s44 | 7 | 9 |
| Documents and contents | Complexity | s9, s13, s15, s22, s24, s35, s40, s42, s54 | 9 | 11 |
| | Different data types and formats | S38, s37, s36, s30, s24, s14, S13, s9, s13, s14, s24, s30, s36, s37, s38, s55 | 16 | 20 |
| Reference and master data | | | | 0 |
| Data warehousing and business intelligence | Establishing a data-driven culture | S2, s15, s20, s21, s24, s31, s35, s39 | 8 | 10 |
| | Organizational resistance | s15, s24, s31 | 3 | 4 |
| | Cost/budget | s6, s9, s13, s15, s22, s24, s26, s35, s37, s38, s40, s45, s46, s53, s54 | 15 | 19 |
| Meta-data | | | | |
| Data quality | Accuracy | S9, s13, s16, s23, s24, s35, s38, s67, s71 | 7 | 9 |
| | Incompleteness and inconsistency | s8, s13, s16, s17, s23, s24, s38, s42, s44 | 9 | 11 |
| | Vagueness | s2, s6, s10, s13, s29, s42, s44 | 7 | 9 |
| Data Governance | Governance bodies | s1, s9, s13, s19, s21, s24, s25, s28, s29, s31, s35, s38, s39, s44, s45, s47, s49 | 17 | 22 |

underscores faster processing needs. Much of the data is unstructured or semi-structured. Therefore, different data types and data formats give rise to issues of data collection, data storage, data management, data transfer and vagueness. Since data is stored in different formats (photos, videos etc.) and comes from various devices, such as sensors etc., sometimes it is incomplete and inconsistent and processing such data for reliable or accurate decision making becomes a challenge. Huge costs are incurred during collecting, storing and processing heterogeneous data. Several studies show that dealing with challenging BD is not the same as dealing with traditional data (s11, s12, s27, s38). There is often a dearth of workers who are skilled in handling big data applications (s5, s9, s34, s35). In addition, there is a large technology gap regarding BDE. The security techniques used for traditional data cannot be considered appropriate to cater with the volume, velocity, veracity and complexity of BD  [31]. Technological aspects of BD are not the only challenge; people can be an issue as well (s15, s24, s31). These challenges may give rise to resistance in an organization to the adoption of BD and related approaches. It is challenging to establish a BD-driven culture in an organization due to these challenges.

It is clear from the results in Tables 6 and 8 that the most mentioned challenge in the selected studies is privacy and security. Data confidentiality also comes under the umbrella of privacy and security. It is an arduous task to be sure about the privacy and security of stored BD. It requires appropriate government rules, laws and policies to deal with security breaches and privacy violations. The challenges of big data are classified according to the DAMA knowledge areas shown in Fig. 2 and listed in Table 9.

### 4.1 Privacy and security challenges of big data

The detailed analysis of the studies highlights the following privacy- and security-related challenges. We identified 10 challenges as presented in Table 9:

One of the major challenges encountered by BDE is the validity of the inferences drawn from BD. The mishmash of PII with non-PII data can lead to new results interpreted in such a way that can lead to disclosure of identity of a person without his knowledge and consent. It is possible to use standard data, which does not contain

**Table 9** Identification of big data privacy and security challenges and their distribution in the literature

| Challenge | Study | Frequency | Percentage |
|---|---|---|---|
| Inference | s9, s12, s16, s19, s21, s24, s28, s35, s40, s51, s52, s57, s58, s59, s64, s73, s79 | 17 | 22 |
| Lack of consent | s7, s8, s9, s19, s24, s24, s29, s30, s50, s52, s56, s57, s59, s69 | 14 | 18 |
| Social misuse of knowledge | s9, s18, s19, s24, s28, s44, s52, s64, s69 | 9 | 11 |
| Diverse data sources | s2, s5, s8, s12, s28, s35, s36, s37, s38, s51, s55, s76 | 13 | 17 |
| Multiple uses of data | s9, s13, s19, s24, s29, s30, s52, s56 | 8 | 10 |
| Storage and processing | s13, s28, s63 | 2 | 3 |
| Technology gap | s13, s24, s36, s38 | 4 | 5 |
| Agreed data usage | s19, s29, s30 | 3 | 4 |
| Unauthorized Access | s15, s21, s24, s26, s30, s33 | 6 | 8 |
| Lack of governance | s1, s19, s21, s24, s25, s28, s29, s31, s38, s44, s61 | 11 | 14 |
| Data provenance | s24, s26, s28, s33, s36, s40, s48, s54, s55, s57 | 9 | 12 |

any PII, to predict sensitive and personal facts about individuals such as bank details and sexual interests [32]. In the mid-90s, Latanya Sweeney, a PhD student at MIT, correctly identified patients by comparing and correlating anonymous health data with a voter database [33]. Lack of consent is a big problem in the way of privacy. Sensitive personal information is often taken without data owner's knowledge and is used to gain commercial benefits. It has been reported that iPhones and Android phones are sending individual's location information to other vendors (Apple and Google) without the consent of the user [32]. In the context of big data, "the concept of notice and consent underlying the data protection laws around the world is no longer suitable as it is often either too restrictive to unearth data's latent value or too empty to protect individuals' privacy" [21]. Unwanted consequences are especially high for consumers who are not much familiar with technology or are poor and naive [32]. This leads to a social misuse of data by more informed users. Multiple sources of information pose challenges and may affect people's privacy in BDE. PII such a health records and financial statements can be accessed via diverse sources that may result in trust alarms. Multiple users can misuse data collected about an individual. The extent to which the data will be used cannot be controlled at the time of collection especially when no consent is taken [2]. The data is not necessarily used for beneficial purposes always. Data owners do not have control over data about themselves and hence are uncertain about the possible use of this data [34]. Finally, the infrastructure of BDE is expected to uphold end-to-end security. This will warrant that no data loss can happen throughout the BD lifecycle. In addition to the need for appropriate technology, there is a need for suitable governance framework as well. The absence of suitable governance framework in BDE can result in deceptive analysis of data and hence can cause high costs [35]. The absence of legal support for application of data policies [36] in relation to BD needs immediate research attention. This section has systematically and methodologically retorted the first research question related to the documentation of prominent privacy and security challenges. Uncertain provenance is a bottleneck to privacy and security. Data provenance must be available and certified [37]. All the challenges related to privacy and security fall under the knowledge area of data security in the DAMA framework.

The most cited challenges reported in the existing literature are inference (22%), diverse data sources (17%) and lack of consent (18%). There are no boundaries on data usage at the time of collection. In some cases, it is used for purposes other than the one for which it was originally collected [2]. The least mentioned challenge in the context of privacy and security is storage and processing (3%); however, although this may not represent the importance of this challenge, it indicates its importance in the selected studies. Further, this study identifies the proposed privacy and security solutions from the selected studies. This is done to identify the gaps and areas for further research and development in privacy and security. In the next section, we present solutions for the identified challenges.

### 4.2 Solutions

Despite the technological and procedural advancements, privacy and security are still the main concerns of many organizations [38]. There should be appropriate procedures and guidelines for BD adoption. BD poses many problems, such as complexity, high cost

and data transfer issues. Thus, many organizations find it challenging to manage and use BD to its full potential.

The privacy and security problems in BDE require immediate attention and a strong roadmap. A successful roadmap starts from decision makers and their concerns, useful data sources and technologies. This means privacy and security problems span the lifecycle of data in ecosystems. It is likely that if privacy and security challenges are not well addressed, the concept of BD cannot be widely accepted [39]. There is no magic to cure privacy and security challenges. As Moura and Serrao [40] state, there is a need to understand the ways through which huge sets of complex heterogeneous data can be protected. In our work, the proposed solutions for privacy and security challenges in BDE are identified and synthesized using the DAMA framework [27] as shown in Table 7 which has eleven different knowledge areas of data management. The solutions found in the selected studies are modeled according to the DAMA framework. For the purpose of simplification, the results are summarized in Table 10.

The general organization of data and relevant resources together make enterprise architecture [41]. Hence, in order to secure data, architecture security should be in place. The analysis, architecture and design, implementation and ongoing maintenance of data are part of data modeling and the design of DAMA knowledge areas [27]. The challenges encountered in the data modeling and design area can be solved via communication security, threat modeling and illegal intrusion detection before a security threat occurs. Most security breaches occur at the time of data collection and when it is stored for later use [22]. A number of studies (s28, s33, s36, s54) suggest incorporating end-point validation and response capabilities in ecosystems to warrant the security and privacy of data. In addition, strategies like encrypted secure queries and making data collection secure by logging the transaction and distributing the data over multiple storage resources are reported to be helpful in achieving security in this area. To conduct efficient and secure operations on BD, data transformation helps to make data more meaningful for user representation and system analysis [42]. In addition to this, there is need for more innovative machine learning algorithms to process huge volumes of data proficiently. In order to ensure security in BDE, encrypted queries to interact with data stores might help. CryptDB allows developers to write queries for encrypted data. In 2014, Google implemented the "Encrypted Big Query Client" on the basis of CryptDB. It enables encrypted big queries against Google's BigQuery service, making use of existing processing power of Google's infrastructure [22]. The reviewed studies suggest close attention should be paid to the issue of user authentication by creating user profiles through an identity-based encryption scheme. The ordinary person is most concerned about data privacy [43], and the most frequently employed solution for securing data privacy in BDE is cryptography [22]. Some authors focus on encryption schemes that ensure privacy while others focus on processing data that is already encrypted [22]. Restricting non-desirable access to the system is another important technique to secure a system. However, even better approach is to safeguard the information using suitable access control policy and by allowing encryption/decryption only when the user is authorized to do so [44]. Protecting data using cryptography and granular access control is the minimum security requirement [45]. Anonymizing data can ensure that privacy

**Table 10** Overview of solutions to key big data security and privacy issues

| DAMA knowledge area | Solution | Study | Frequency | Percentage |
|---|---|---|---|---|
| Data architecture | Architecture security | s11, s13, s18, s22, s28, s30, s35, s40, s54 | 9 | 12 |
| Data modeling and design | Communication security | s11, s21, s22, s23, s30, s35, s40, s41 | 8 | 10 |
| | Threat modeling/risk assessment | s10, s18, s19, s21, s24, s29, s30, s36, s39, s48, s54, s70, s78 | 13 | 17 |
| | Intrusion recognition | s13, s18, s39 | 3 | 4 |
| Data storage and operations | Incorporate end point validation and response capabilities | s28, s33, s36, s54 | 4 | 5 |
| | Secure queries | s12, s16, s28 | 3 | 4 |
| | Secure data collection, storage, usage, transaction logs and destruction | s8, s12, s18, s23, s57s28, s33, s35, s36, s37, s39, s54,  s57, s69, s71, s72, s73 | 16 | 20 |
| | Distributed sources for data | s10, s23, s28, s36, s38 | 5 | 7 |
| | Data transformation | s13, s23, s28, s36 | 5 | 7% |
| | Machine learning algorithms | s39, s46, s55 | 3 | 4 |
| | Data classification | s69, s72, s73 | 3 | 4 |
| Data security | Encryption | s4, s10, s11, s12, s16, s21, s22, s26, s28, s30, s31, s35, s36, s38, s40, s41, s42, s45, s47, s53, s54, s57, s59, s65, s66, s77, s78 | 27 | 34 |
| | User validation | s11, S19, s26, s28, s30, s37, s54, s77 | 8 | 10 |
| | Confidentiality | s16, s26, s38, s39, s48, s57, s66, s69 | 8 | 10 |
| | Skills | s13, s15, s18, s27, s37 | 5 | 7 |
| | Differential privacy | s7, s10, s11, s40, s57, s69, s71, s73, s75 | 9 | 11 |
| | Latent data privacy | s64 | 1 | 2 |
| | Data-driven privacy preserving | s67, s68, s71, s76 | 4 | 5 |
| Data integration and interoperability | | | 0 | 0 |
| Documents and contents | Granular access control | s10, s11, s15, s16, s21, s24, s26, s28, s29, s30, s31, s33, s36, s45, s48, s54, s55, s66, s69, s76 | 20 | 25 |
| | Data anonymization | s7, s10, s12, s14, s16, s19, s21, s29, s31, s35, s42, s44, s45, s47, s48, s51, s57, s58, s59, s68, s69, s71, s73, s77 | 24 | 31 |
| | Granular audits | s21, s28, s33, s54, s72, s76, s78 | 7 | 9 |
| | User education | s28, s44 | 2 | 3 |
| Reference and master data | Recovery | s15, s21 | 2 | 3 |
| Data warehousing and business intelligence | | | 0 | 0 |
| Meta-data | Storage of encrypted header information | S35 | 1 | 2 |

**Table 10** (continued)

| DAMA knowledge area | Solution | Study | Frequency | Percentage |
|---|---|---|---|---|
| | Provenance metadata | s4, s23, s42, s67, s72 | 5 | 6 |
| Data quality | Integrity | s4, s10, s14, s16, s22, s23, s26, s28, s30, s35, s38, s39, s42, s44, s48, s51, s53, s55, s57, s69 | 20 | 25 |
| Data governance | Privacy at social network | s13, s19, s23, s28, s38, s49, s52, s53, s78 | 9 | 11 |
| | Policies, laws or government | s1, s8, s10, s13, s15, s19, s21, s59, s62s22, s23, s25, s28, s29, s31, s35, s37, s38, s40, s41, s44, s49, s51, s56, s57, s69, s72, s73, s76 | 27 | 34 |

threats are addressed as all sensitive information is removed from the data set and is not in identifiable form. Existing privacy preserving methods, such as differential privacy, *k*-anonymity, *l*-diversity, are considered suitable for attaining inherent-data privacy, but they are not capable for achieving latent-data privacy which is vulnerable to inference attacks [46]. Data anonymization, access control, audits and educating users all contribute to data storage, protection, indexing and facilitating usage of unstructured data and offering it in integrated and interoperable form with standard structured data [47]. Hence, these fall under the category of document and content in the DAMA knowledge areas. Storing encrypted header information and metadata provenance makes securing big data easier. At the time of data collection, if information related to metadata is not recorded, the data stored has no traceability and becomes useless for decision makers [48]. Most privacy breaches occur from public data which is exposed via social networks. If an additional layer of privacy is employed on social networks, a lot of personal information, which can expose a person's identity, can be made private or de-identified. This can be achieved if there is a well-defined strategy including all or elements of planning, monitoring and hold over the data management and usage [27]. Data governance covers all these steps toward a secure and privacy-aware BDE.

The most cited knowledge area with respect to solutions is data security for BDE, and the least cited is data architecture. This indicates the need for more work in the area of data security architecture. Finding a means to guard data at rest is not the only important thing, it is also important to find out how safely the data is collected in the first place [22]. Moreno, Serrano and Fernandez-Medina [22] suggest that the storage of data in BDE can be made secure by dividing the data into different chunks and then storing them in separate cloud storage service providers. Appropriate policies and governance can make data protection a more organized process. Preparing beforehand for expected threats might be helpful to contend with potential data loss and security breaches. A number of studies (s9, s12, s32) suggest that modeling the expected threats and devising a mitigation strategy can reduce security and privacy breaches. Some of the selected papers (s8, s13, s19, s22, s32) describe early attack detection, maintaining data integrity and keeping track of data origins for recovery purposes as ways to attain privacy and security.

## 5 Interdependence between challenges and solutions for privacy and security in BDE

This study reports on a number of challenges and solutions, which were organized and analyzed using the DAMA framework. We further analyzed and identified the interdependency between challenges (see Table 11), the interdependency between solutions (see Table 12) and the interdependency between challenges and solutions (see Table 13).

### 5.1 Interdependency between challenges

The interdependence indicates that existence of one challenge might be an indicator that other related and interdependent challenges also exist. Table 11 shows the interdependence between challenges. As an example, the inference from non-PII can lead to discovery of PII which is facilitated by diverse sources of data. Similarly, due to social misuse of knowledge, consent can be given on behalf of original data owner without their knowledge. Lack of consent gives rise to other challenges such as the social misuse of data, multiple sources of data and agreed use of data. Technology gap may result in storage and processing challenges and unauthorized data access. To model this interdependency, if the existence of one challenge gives rise to other challenges, a cross is placed in the corresponding column, as shown in Table 11. Hence, to address one challenge, all other interdependent and related challenges need to be taken care of.

### 5.2 Interdependency between solutions

Further, solutions and their interdependencies are carefully analyzed and mapped in Table 12. For instance, communication security, architecture security, user validation, confidentiality, privacy at social networks and recovery are attained if proper access control is in place. Similarly, authentication, encryption, confidentiality, access control and anonymization are all inter-related and are used to gain common advantages. Furthermore, the privacy at social networks can be ensured if a proper access control mechanism is in place and the user is well educated about what content to make public and what not to make public.

### 5.3 Interdependency between challenges and solutions

Finally, we analyzed and mapped the challenges to the proposed solutions. This will provide more detailed information on how to address the challenges using the proposed solutions. Table 13 shows the relationship between solutions and challenges and which security measures would help to ensure security against a particular challenge.

## 6 Discussion and implications

This SLR study reviewed several papers using the well-known DAMA [27] framework as a guiding industry lens and adaptive EA [26] as theoretical lens. This study provides valuable insights into the recent research on BDE, in particular security and privacy challenges and solutions. The results of this SLR show that there is growing interest

Anwar *et al. J Wireless Com Network*     *(2021) 2021:130*

Page 19 of 30

**Table 11** Modeling the dependency between the challenges to privacy and security in big data ecosystem

| Challenges | Challenge mapping | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Inference | Lack of consent | Social misuse of knowledge | Diverse sources of data | Multiple use of data | Storage and processing | Technology gap | Agreed data usage | Unauthorized access | Lack of governance |
| Inference | | | | × | | | | | | |
| Lack of consent | | | × | | × | | | × | × | × |
| Social misuse of knowledge | | | | | | | | | | |
| Diverse sources of data | | | | | | × | × | | | |
| Multiple use of data | | | × | | | × | × | | | |
| Storage and processing | | | | | | | | | | |
| Technology gap | | | | | | × | | | × | |
| Agreed data usage | | | | | | | | | × | × |
| Unauthorized access | | | | | | × | | × | | × |
| Lack of governance | | | | | | | | × | | |
| Data provenance | × | | × | × | | × | | | | |

Anwar *et al. J Wireless Com Network*     (2021) 2021:130

Page 20 of 30

**Table 12** Mapping of solution to the challenges for privacy and security in the big data ecosystem

|  | Solution mapping |
| --- | --- |
| *Solutions* |  |
| 1.Communication security | 6 |
| 2. Architecture security | 1, 3, 4, 5, 6, 7, 10, 12, 15 |
| 3. User validation | 6 |
| 4. Encryption | 7, 9 |
| 5.Confidentiality | 6, 22 |
| 6.Access control | 3, 4, 21 |
| 7. Data anonymization | 4, 20 |
| 8.Privacy at social networks | 6, 17 |
| 9.Secure queries | 4 |
| 10. Secure data collection, storage and transaction logs | 3, 24 |
| 11.Policies, laws or governance | 13 |
| 12.Distributed sources of data | 11 |
| 13.Threat modeling/risk assessment | 1, 5, 14 |
| 14.Intrusion recognition | 13 |
| 15.Integrity | 10,  12 |
| 16.Recovery | 3, 6 |
| 17.User education | 5 |
| 18. Incorporate end point validation and response capabilities | 3, 4, 10 |
| 19. Granular audits | 10 |
| 20. Data transformation | 2, 7 |
| 21. Machine learning algorithms | 18 |
| 22. Differential privacy | 5, 7, 26 |
| 23.Storage of encrypted header information | 4 |
| 24.Provenance metadata | 7, 10 |
| 25. Latent data privacy | 7 |
| 26. Data-driven privacy preserving | 22,24 |
| 27. Data classification | 7,10,24 |

among the research community in this topic. For instance, our initial research on this topic returned 175 articles across five well-known databases. Based on the paper selection criteria, we selected and reviewed 79 papers for this study. The review of the selected studies indicates a total set of 22 challenges, such as data access and sharing, data capture, governance, heterogeneity, scale, skills and relevant solutions. Our findings reveal that privacy and security is the most reported (53%) challenge in BDE. Other challenges such as heterogeneity, scale, cost, data access, skill gaps and timeliness are also mentioned but have a lesser impact compared to privacy and security concerns.

Adaptive EA [26] describes an end-to-end ecosystem as four main layers. Privacy considerations should be taken individually at each layer; hence, this study describes challenges of privacy and security specific to four layers of digital ecosystem (Fig. 2). The existing solutions for the mentioned challenges are also categorized based upon ecosystem layers in Fig. 2. Privacy and security challenges and relevant solutions are categorized and mapped to provide deep insights into this critical concern. This research study provides a catalogue of challenges that are specific to privacy and security. With respect to privacy and security threats mentioned in the literature, BD challenges are

Anwar *et al. J Wireless Com Network*     (2021) 2021:130

Page 21 of 30

**Table 13** Mapping of challenges to the proposed solutions of privacy and security in big data ecosystems

| Challenges | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Solutions | Inference | Lack of Consent | Social misuse of knowledge | Diverse sources of data | Multiple use of data | Storage and processing | Technology gap | Agreed data usage | Unauthorized Access | Governance | Data Provenance |
| Communication security | | | | × | | × | × | | | | |
| Architecture security | | | | × | | | | | | | |
| User validation | | | | × | × | | | × | | | |
| Incorporate end point validation and response capabilities | | | | × | × | | | | | | |
| Encryption | | | | | | × | | | × | | |
| Confidentiality | | | | | | | | × | | | |
| Access control | | | | | | | | × | | | |
| Data anonymization | | | | | | | | | | | |
| Privacy at social networks | × | | × | | | | | | | | |
| Secure queries | | | | | | × | | | × | | |
| Secure data collection, storage, and transaction logs | | | | | | × | | | | | |
| Policies, laws or governance | × | | × | | × | | | | | × | |
| Distributed sources of data | | | × | | | × | | | | | |
| Threat modeling | | | | | | | × | | | | |
| Granular audits | | | | | | | | | × | | |
| Attack detection | | | | | | | | | | | |
| Integrity | | | | | × | × | | | | | |
| Recovery | | | | × | | | | | | | |
| User education | | × | × | | | | | | | × | |
| Data transformation | × | | | | | × | | | | | |
| Machine learning algorithms | | | | | | | | | × | | |

Anwar *et al. J Wireless Com Network*   (2021) 2021:130

Page 22 of 30

**Table 13** (continued)

| Challenges | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Solutions | Inference | Lack of Consent | Social misuse of knowledge | Diverse sources of data | Multiple use of data | Storage and processing | Technology gap | Agreed data usage | Unauthorized Access | Governance | Data Provenance |
| Differential privacy | × | | | | | | | | × | | |
| Storage of encrypted header Information | | × | | | | × | | | | | × |
| | | | | | | × | | | | | |
| Provenance metadata | | | × | | | × | | | | | × |
| Latent data privacy | | | × | | × | × | | | | | |
| Data-driven privacy preserving | | | × | | | | | × | × | | |
| Data classification | | | | | | | | × | × | × | × |

categorized into eleven knowledge areas based on the DAMA framework. Some studies report on challenges in one or two knowledge areas of DAMA. Data architecture (timeliness and device heterogeneity) and data governance (governance bodies) are the knowledge areas which comprise 22–23% of the total reported challenges. The most covered area in the articles selected in this review is data security (privacy and security, confidentiality, and technology gap) contributing 78% of the challenges. Data modeling and design (data deluge, inappropriate analytics and lack of skills) and data storage and operations (data storage and sharing, data collection and availability) are the second most reported knowledge areas after privacy and security, contributing 38% and 43% of the challenges. Data integration and interoperability (data management) is discussed in 17% of the articles. No study appears to address security issues centered on master data and metadata; thus, this identifies the need for further research in this area. Furthermore, some papers (s4, s12, s15, s28, s34, s40, s54) discuss issues in more than one category and therefore belong to more than one category. Documents and contents (complexity and different data types and formats) are discussed in 31% of the literature, whereas data warehousing and business intelligence (establishing a data-driven culture, cost/budget and organizational resistance) is addressed in 33% of the literature.

The results of this research indicate that certain challenges are repeatedly mentioned in the literature. These are related to timeliness, data deluge, availability, privacy and security, cost/budget and governance bodies. Other challenges such as device heterogeneity, organizational resistance, accuracy and vagueness of data were rarely reported; however, they have an effect on the overall working of BDE. A number of studies have reported one major concern in BDE, i.e., privacy and the security of data. It has been reported that there are certain stages and areas during a data life cycle that give rise to problems regarding privacy and security. Lack of consent from the user (18%), inference (22%), social misuse of knowledge (11%), diverse sources of data (17%), multiple uses of data (10%), ineffective storage and processing (3%), technology gap (5%), utility of data (4%), unauthorized access (8%), lack of governance (14%) and data provenance (12%) make up the landscape of all the security and privacy challenges of big data ecosystems.

Privacy and security of BD remains an area of interest through the entire lifecycle of BD in an ecosystem. Therefore, the challenges and their solutions have a strong interdependence. There are certain issues which, if not solved, can give rise to other issues, for example, a user's lack of consent can cause multiple uses of data without the user's knowledge. This, in turn, can give rise to the social misuse of knowledge. Similarly, if technological issues are not solved, then the secure storage and processing of data is not assured. Moreover, it might not be possible to avoid issues regarding unauthorized access to data with weak security at the storage and processing level. Furthermore, a lack of governance may result in inferences being made.

In the existing literature, a range of solutions has been proposed; however, these solutions lack concrete implementation guidelines and do not discuss application case studies and their impact on privacy and security improvement. This warrants further research, development and the evaluation of solutions in different organizational and user contexts and solutions are needed that can be developed and deployed easily in BDE and are reusable. According to the results presented in this paper, data passes through four stages. These are data source, the data itself, data storage, results and

output devices. The solutions are divided into eleven knowledge areas [27]. The first knowledge area is data architecture which includes securing the overall architecture (12%). The second knowledge area is data modeling and design, which requires secure communication (10%), threat modeling/ risk assessment (17%) in advance to have some preparedness and intrusion recognition (4%). The next knowledge area is data storage and operations, which includes incorporating end-point validation and response capabilities (5%), writing encryption and secure queries (4%), securing data collection, storage and transaction logs (20%), make data sources distributed (7%), transforming data before use (7%), incorporating machine learning algorithms (4%) and conducting data classification (4%). Data security is the next knowledge area in the DAMA framework and includes encryption (34%), user validation (10%), confidentiality (10%), appropriate skills (7%), differential privacy (11%), latent data privacy (2%) and data-driven privacy preserving schemes (5%). The knowledge area, documents and contents cover granular access control (25%), data anonymization (31%), granular audits (9%) and user education (3%). The knowledge area of data quality only includes integrity (25%). Reference and master data includes recovery (3%) to secure BD. The area of metadata includes storage of encrypted header information (2%) and provenance metadata (6%). No solution found in this SLR falls in the category of data integration and interoperability and data warehousing and business intelligence. Lastly, data governance which is central point of the DAMA framework includes privacy at social networks (11%) and incorporating policies, laws and government (34%) rules for big data.

According to the quality settings formulated, few selected studies do not include the details regarding implementation methodology used for preserving privacy and security of BDE. Thus, more quality empirical studies are needed which include high-grade implementation methodologies in the context of the privacy and security of BD. As a result, such studies may augment the practical approaches for a secure and privacy-aware BDE.

This SLR study has suggestions for both research and practice. For practitioners, the findings feature imperative research voids that need more scrutiny, particularly the development and implementation of methods to achieve a secure and privacy-aware big data ecosystem. As shown in the results in Table 9, there is a need for more practically tested solutions for privacy and security of BD in an ecosystem. As a result of this research study, it is expected that practical, efficient, and effective security and privacy solutions to address the identified gaps will be developed and deployed.

Moreover, it has been noted that the reviewed studies broadly emphasize on the challenges of privacy and security in BDE and methods to handle these challenges. The results of this review paper indicate that there is a pattern of dependency among different solutions. Some solutions are interlinked which means the presence of one ensures the other and vice versa. For instance, if we can achieve architecture security, this will ensure communication security, user validation, encryption, access control, data anonymization, secure data collection, storage and transaction logs and integrity. Access control can be employed if authentication and encryption are already in place. Another consideration in relation to privacy and security is organizational compliance with information security laws. Legal considerations for handling data must not be overlooked [49]. Policies, laws and governance are related to threat modeling. Threat

Anwar *et al. J Wireless Com Network*    (2021) 2021:130

Page 25 of 30

modeling is a technique for enhancing security by identifying threats and vulnerabilities, and then stating mitigation approaches to stop or lessen the adverse impact to the system [26]. Future research may also investigate whether there is one complete solution that addresses the majority of privacy and security challenges in BDE. In BDE, research challenges are not limited to any one stage; rather, they range from data creation to data usage and all the stages in between (data storage and transportation, data transformation and processing). In order to support this lifecycle, a distributed and highly capable framework is needed [22]. Solutions should be considered in the context of the whole rather than individual solutions; thus, integrated solutions which support the different lifecycle stages of data ecosystems are required. This is the focus of our further research.

Further, we realize that there are a plenty of subjective and exploratory studies but not confirmatory and descriptive studies. This research study highlights the need for more confirmatory and action-oriented research because this could help a security measure that is applicable through the entire life cycle of data in an ecosystem to be devised. Furthermore, to obtain a trustworthy big data ecosystem, previous authors have emphasized end-to-end data privacy and security. There is a need for a complete security framework that covers the entire lifecycle of big data. Designing such a framework without negotiating the data's information state along with retaining its volume, variety and velocity is a challenge in itself.

Finally, it has been noted that although there have been a large number of theoretical descriptions on managing the privacy and security of big data ecosystems over the past few years, concrete implementation strategies and their impact on privacy and security improvement have not received much consideration in the existing literature. This is a critical gap that needs urgent research attention and effort to address this gap. Indeed, the fact that security issues have not been investigated at the origin of big data [22] highlights the need for a more in-depth understanding of the challenges and the provision of an end-to-end solution [43].

Our findings also indicate that more real-time techniques are required in the future for synching with the ever-increasing volume and heterogeneity of BD. The procedures should be capable of dealing with big data sets in a reasonably acceptable amount of time. Most of the technologies that have been developed for big data processing, including Hadoop, Cassandra, Hive, PigLatin, MapReduce, Mahout and Storm, do not have satisfactory security safeguards [46]. Ensuring security and high performance simultaneously is questionable. Due to the characteristics of big data, it is necessary to pay immediate attention to all security aspects [22].

## 7 Limitations

Like any study, this research has few limitations which are discussed in this section. The results from this study should be viewed in light of these limitations. Firstly, the sources used in this study are limited to only five databases. However, these are very well-known scientific databases that provide excellent coverage of the existing literature. Secondly, it is vital to report that because of the scope of this research project, a limited number of search strings were used. We may have missed certain findings on some of the challenges and possible solutions to security and privacy issues because they were not included in the search strings.

To ensure a correct interpretation and syntheses of the results from theoretical as well as practical viewpoint, we applied adaptive EA and the DAMA framework. However, we focused on analysis of result from industry viewpoint (i.e., DAMA framework) only. An extension to this research can be analysis of research from a theoretical viewpoint (i.e., adaptive EA) as well. Further, the results and categorization of concepts (challenges/ solutions) were reviewed and checked by the second author to ensure the quality of results. This was also done to avoid any possible omission or researcher bias. To ensure that correct search studies and databases were selected, we pilot tested of the search strings across different databases. This was important to identify and select the relevant data sources and studies for this research.

## 8  Conclusion and future directions

In this study, we investigated the current state of research in BDE using the well-known SLR approach. From the primary studies, we extracted 175 relevant studies published in the last five years (2014–2019). Using inclusion and exclusion criterion, 79 primary studies with the highest assessment scores were selected.

This study identified a number of challenges from the standpoint of privacy and security in BDE. The outcomes of this work have been described in two steps. Firstly, it discusses the current research emphasis and directions as documented in the chosen papers. Secondly, in order to answer the research question, it debates about the data that was analyzed and inferred from the selected literature. The study enabled us to explore the existing research in BDE using the well-known SLR approach. In particular, this study highlighted the most pressing areas of focus both in challenges (e.g., confidentiality, privacy and security, governance) and solutions (e.g., encryption, access control and anonymization). Due primarily to the newness of BD, no end-to-end empirical solution for the most mentioned challenges of privacy and security is reported. Existing research mostly advocates facts that are conceptual in nature and lack an empirical methodology. This study also highlights the lack of practical security and privacy implementation in BD environments. This indicates more work is required in both directions. Issues concerning availability, data formats and governance remain challenging areas of research in BD. Thus, research on data governance, heterogeneous data formats and availability in the domain of BD management would be highly insightful. There is also great potential in relation to data integrity and policy-making which may greatly augment the research on big data privacy and security.

A unique and critical research finding from this study is that it also identifies and analyses a pattern of interdependency within challenges and solutions and among them. These findings of this SLR will help in developing a comprehensive and end-to-end security framework to handle the volume and diversity of BDE. Further, the challenges and solutions are categorized according to the DAMA knowledge areas in order to obtain a standard industry view of data management. The results are also listed according to the adaptive EA layers to get a theoretical view of a secure BDE. The exposition in this study will raise awareness of the key challenges and solutions to achieve privacy and security in BDE. It highlights the many gaps in existing research and develops a roadmap for further research. Based on these conclusions, there is a need for a complete framework that strives to improve the issues of privacy and security. These findings will be very valuable

Anwar *et al. J Wireless Com Network*   (2021) 2021:130

Page 27 of 30

for future researchers in the area of BDE in focusing their research efforts on topics where this is needed the most and could have maximum impact.

## 9 Appendix 1: Studies included in SLR

| Study number | Reference |
| --- | --- |
| s38 | I. A. T. Hashem et al., Inf. Syst. **47**, 98 (2015) |
| s15 | A. Chopra JIMS, A. Chopra, and S. Madan, *Big Data: A Trouble or A Real Solution? "Big Data: A Trouble or a Real Solution?" View Project Data Encryption and Decryption View Project Big Data: A Trouble or A Real Solution?* (n.d.) |
| s7 | A. Gosain and N. Chugh, Int. J. Comput. Appl. **100**, (2014) |
| s54 | O. Awodele et al., Int. J. Comput. Appl. **133**, (2016) |
| s56 | A. B. Munir et al., Int. Sch. Sci. Res. Innov. 9 (2015) |
| s16 | B. Nelson and T. Olovsson, in *Big Data (Big Data), 2016 IEEE Int. Conf.* (IEEE, 2016), pp. 3693–3702 |
| s27 | C. Constantine, Netw. Secur. **2014**, 18 (2014) |
| s53 | C. Stergiou and K. E. Psannis, Multimed. Tools Appl. **76**, 22,803 (2017) |
| s8 | D. Broeders et al., Comput. Law Secur. Rev. **33**, 309 (2017) |
| s5 | D. Bumblauskas et al., Bus. Process Manag. J. **23**, 703 (2017) |
| s6 | D. Oprea, Rev. Cercet. si Interv. Soc. **55**, 112 (2016) |
| s8 | D.-J. Cho, Int. Inf. Inst. (Tokyo). Inf. **19**, 605 (2016) |
| s48 | E. Bertino, in *Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* (Springer Verlag, 2014), pp. 9–13 |
| s52 | G. Hull, Ethics Inf. Technol. **17**, 89 (2015) |
| s31 | G. Lafuente, Netw. Secur. **2015**, 12 (2015) |
| s26 | H. Es-Samaali, A. Outchakoucht, and J. P. Leroy, Int. J. Comput. Networks Commun. Secur. **5**, 137 (2017) |
| 36 | H. Singh and G. Singh, i-Manager's J. Inf. Technol. **6**, 25 (2016) |
| s19 | I. S. Rubinstein, Int. Data Priv. Law **3**, 74 (2013) |
| s1 | Holt and Malčić, J. Inf. Policy **5**, 155 (2015) |
| s11 | J. Moreno, M. A. Serrano, and E. Fernández-Medina, Futur. Internet **8**, 44 (2016) |
| s28 | J. Moura and C. Serrão, arXiv Prepr. arXiv1601.06206 (2016) |
| s43 | J. W. Crampton, GeoJournal **80**, 519 (2015) |
| s55 | J. Sänger et al., in *Database Expert Syst. Appl. (DEXA), 2014 25th Int. Work.* (IEEE, 2014), pp. 278–282 |
| s42 | K. V. S. N. R. Rao, M. Pranava, and A. Mounika, **10**, (2015) |
| s39 | M. S. Al-Kahtani, *Security and Privacy in Big Data* (2017) |
| s25 | M. Awais and A. Gill, Int. Conf. Inf. Syst. Dev. (2016) |
| s9 | N. A. Shozi and J. Mtsweni, in *11th Int. Conf. Cyber Warf. Secur. ICCWS2016* (Academic Conferences and publishing limited, 2016), p. 296 |
| s24 | N. Kshetri, Telecomm. Policy **38**, 1134 (2014) |
| s21 | O. Hamami, Bus. Intell. J. **19**, 20 (2014) |
| s35 | O. Tene and J. Polonetsky, Stan. L. Rev. Online **64**, 63 (2011) |
| s10 | P. Jain, M. Gyanchandani, and N. Khare, J. Big Data **5**, (2018) |
| s33 | P. Johri et al., in *Comput. Commun. Autom. (ICCCA), 2017 Int. Conf.* (IEEE, 2017), pp. 268–272 |
| s51 | P. Leonard, Int. Data Priv. Law **4**, 53 (2013) |
| S40 | P. Pathak, N. Vyas, and I. Joshi, Int. J. Adv. Res. Comput. Sci. **8**, (2017) |
| s41 | P. Porambage et al., IEEE Cloud Comput. **3**, 36 (2016) |
| s22 | A. Q. Gill et al., in *ACM Int. Conf. Proceeding Ser.* (Association for Computing Machinery, 2017) |
| s47 | R. Lu et al., IEEE Netw. **28**, 46 (2014) |
| s4 | R. M. Alguliyev, R. T. Gasimova, and R. N. Abbasli, Int. J. Mod. Educ. Comput. Sci. **9**, 28 (2017) |
| s29 | R. Meijer, P. Conradie, and S. Choenni, J. Theor. Appl. Electron. Commer. Res. **9**, 32 (2014) |
| s34 | S. Bagriyanik and A. Karahoca, Glob. J. Inf. Technol. **6**, 107 (2016) |

Anwar *et al. J Wireless Com Network*    (2021) 2021:130

Page 28 of 30

| Study number | Reference |
| --- | --- |
| s44 | S. Chauhan, N. Agarwal, and A. K. Kar, Info **18**, 73 (2016) |
| s20 | S. Fosso Wamba and D. Mishra, Business Process Management Journal **23**, 477 (2017) |
| s23 | S. Kaisler et al., in *Syst. Sci. (HICSS), 2013 46th Hawaii Int. Conf.* (IEEE, 2013), pp. 995–1004 |
| s32 | S. Madden, IEEE Internet Comput. **16**, 4 (2012) |
| s17 | S. P. Menon and N. P. Hegde, in *Intell. Syst. Control (ISCO), 2015 IEEE 9th Int. Conf.* (IEEE, 2015), pp. 1–7 |
| s37 | S. Sagiroglu and D. Sinanc, in *Collab. Technol. Syst. (CTS), 2013 Int. Conf.* (IEEE, 2013), pp. 42–47 |
| s45 | S. Spiekermann et al., Electron. Mark. **25**, 161 (2015) |
| s30 | S. U. Rehman et al., Int. J. Commun. networks Inf. Secur. **8**, 147 (2016) |
| s13 | U. Sivarajah et al., J. Bus. Res. **70**, 263 (2017) |
| s2 | V. Parmar and J. Yadav, Int. J. Adv. Res. Comput. Sci. **8**, (2017) |
| s50 | B. W. Schermer, B. Custers, and S. van der Hof, Ethics Inf. Technol. **16**, 171 (2014) |
| s46 | X.-W. Chen and X. Lin, IEEE access **2**, 514 (2014) |
| s14 | Y. Perwej, Sci. Educ. **4**, 14 (2017) |
| s49 | Y. Wang, TechTrends **60**, 381 (2016) |
| s57 | S. H. Begum and F. Nausheen, in *Proc. 2nd Int. Conf. Inven. Syst. Control. ICISC 2018* (Institute of Electrical and Electronics Engineers Inc., 2018), pp. 512–516 |
| s58 | M. Mito et al., in *2018 9th Int. Conf. Aware. Sci. Technol. ICAST 2018* (Institute of Electrical and Electronics Engineers Inc., 2018), pp. 319–323 |
| s59 | T. Kittmann, J. Lambrecht, and C. Horn, in *IEEE Int. Conf. Emerg. Technol. Fact. Autom. ETFA* (2018), pp. 1067–1070 |
| S60 | J. Walker, L. Rock, and L. Vieira, in *SoutheastCon 2018* (Institute of Electrical and Electronics Engineers (IEEE), 2018), pp. 1–1 |
| s61 | P. Dunphy, L. Garratt, and F. Petitcolas, in *Proc.—3rd IEEE Eur. Symp. Secur. Priv. Work. EURO S PW 2018* (Institute of Electrical and Electronics Engineers Inc., 2018), pp. 75–78 |
| s62 | H. Moon et al., in *Proc.—2017 IEEE 6th Int. Congr. Big Data, BigData Congr. 2017* (Institute of Electrical and Electronics Engineers Inc., 2017), pp. 525–528 |
| s63 | L. M. Tanczer et al., in *IET Conf. Publ.* (Institution of Engineering and Technology, 2018), pp. 33 (9 pp.)-33 (9 pp.) |
| s64 | Z. He, Z. Cai, and J. Yu, IEEE Trans. Veh. Technol. **67**, (2018) |
| s65 | K. David Strang and Z. Sun, in *Proc.—2016 IEEE Int. Conf. Big Data, Big Data 2016* (Institute of Electrical and Electronics Engineers Inc., 2016), pp. 4035–4037 |
| s66 | S. Khuntia and P. S. Kumar, in *2018 9th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2018* (2018) |
| s67 | A. Cuzzocrea and E. Damiani, in *Proc.—18th IEEE/ACM Int. Symp. Clust. Cloud Grid Comput. CCGRID 2018* (Institute of Electrical and Electronics Engineers Inc., 2018), pp. 675–681 |
| s68 | Y. Canbay, Y. Vural, and S. Sagiroglu, in *Int. Congr. Big Data, Deep Learn. Fight. Cyber Terror. IBIGDELFT 2018—Proc.* (Institute of Electrical and Electronics Engineers Inc., 2019), pp. 24–29 |
| s69 | K. Patel and G. B. Jethava, in *Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2018* (Institute of Electrical and Electronics Engineers Inc., 2018), pp. 194–199 |
| s70 | J. Hernandez-Serrano et al., in *2018 Glob. Internet Things Summit, GIoTS 2018* (2018) |
| s71 | A. Cuzzocrea, in *IEEE Int. Conf. Data Min. Work. ICDMW* (IEEE Computer Society, 2017), pp. 992–994 |
| s72 | J. L. C. Sanchez, J. B. Bernabe, and A. F. Skarmeta, in *IEEE World Forum Internet Things, WF-IoT 2018—Proc.* (Institute of Electrical and Electronics Engineers Inc., 2018), pp. 41–46 |
| s73 | J. A. Shamsi and M. A. Khojaye, IT Prof. **20**, 73 (2018) |
| s74 | S. Saxena, J. Information, Commun. Ethics Soc. **15**, 385 (2017) |
| s75 | P. Jain, M. Gyanchandani, and N. Khare, J. Big Data **5**, (2018) |
| s76 | T. Rantala, K. Palomäki, and K. Valkokari, ISPIM Conference Proceedings 1 (2018) |
| s77 | K. Abouelmehdi, A. Beni-Hessane, and H. Khaloufi, J. Big Data **5**, 1 (2018) |
| s78 | B. Duncan, M. Whittington, and V. Chang, in *Proc. 2017 Int. Conf. Eng. Technol. ICET 2017* (2017), pp. 1–7 |

Anwar *et al. J Wireless Com Network*    (2021) 2021:130

Page 29 of 30

| Study number | Reference |
|---|---|
| s79 | N. Fabiano, in *Proc.—2017 IEEE Int. Conf. Internet Things, IEEE Green Comput. Commun. IEEE Cyber, Phys. Soc. Comput. IEEE Smart Data, IThings-GreenCom-CPSCom-SmartData 2017* (Institute of Electrical and Electronics Engineers Inc., 2018), pp. 727–734 |

### Abbreviations
RFID: Radio frequency identification; BD: Big data; PII: Personally identifiable information; BDE: Big data ecosystem; SLR: Systematic literature review; IoT: Internet of things; DAMA: Data management association; ICO: Initial coin offering.

### Author details
[1]School of Software, Faculty of Engineering, and Information Technology, The University of Technology Sydney, Ultimo, NSW, Australia. [2]College of Applied Computer Science, King Saud University, Riyadh, Saudi Arabia.

### References
1. S. Bryson et al., Commun. ACM **42**, 82 (1999)
2. J. Manyika et al., Big data: the next frontier for innovation, competition, and productivity (2011)
3. T.H. Jetzek, The sustainable value of open government data: uncovering the generative mechanisms of open data through a mixed methods approach (2015)
4. S. Tanwar, S. Tyagi, N. Kumar (ed.), Multimedia Big Data Computing for IoT Applications (Springer, Singapore, 2020). https://doi.org/10.1007/978-981-13-8759-3
5. A.B. Munir, S.H. Mohd Yasin, F. Muhammad-Sukki, WASET Int. J. Soc. Educ. Econ. Manag. Eng. **9**, 355 (2015)
6. U. Selvi, D.S. Pushpa, A Review of Big Data and Anonymization Algorithms (2015).
7. Y. Perwej, Sci. Educ. **4**, 14 (2017)
8. M.I. Razzak, M. Imran, G. Xu, Neural Comput. Appl. **32**, 4417 (2020)
9. B.B.P. Rao et al., in *Proc. Int. Conf. Sens. Technol. ICST* (2012), pp. 374–380
10. F.S. Gürses (2010)
11. N. Galov, Hosting Tribunal (2020)
12. H. Cai et al., IEEE Internet Things J. **4**, 75 (2017)
13. A. Kumari et al., IET Netw. **8**, 155 (2019)
14. M.A. Amanullah et al., Comput. Commun. **151**, 495 (2020)
15. P. Jain, M. Gyanchandani, N. Khare, J. Big Data **3** (2016)
16. S.P. Menon, N.P. Hegde, in *Intell. Syst. Control (ISCO), 2015 IEEE 9th Int. Conf.* (IEEE, 2015), pp. 1–7
17. S. Madden, IEEE Internet Comput. **16**, 4 (2012)
18. W.A. Günther et al., J. Strateg. Inf. Syst. (2017)
19. Y. Demchenko, C. De Laat, P. Membrey, in *2014 Int. Conf. Collab. Technol. Syst. CTS 2014* (IEEE Computer Society, 2014), pp. 104–112
20. K. Michael, K.W. Miller, Computer **46**, 22 (2013)
21. A.B. Munir et al., Int. Sch. Sci. Res. Innov. 9 (2015)
22. J. Moreno, M.A. Serrano, E. Fernández-Medina, Futur. Internet **8**, 44 (2016)
23. A. Karim et al., Futur. Gener. Comput. Syst. **107**, 942 (2020)
24. F. Amalina et al., IEEE Access **8**, 3629 (2020)
25. B. Kitchenham, B. Kitchenham, S. Charters (2007)
26. A.Q. Gill, in *The Gill Framework: Adaptive Enterprise Architecture Toolkit* (CreateSpace Independent Publishing Platform, 2012)
27. P. Cupoli, S. Earley, D. Henderson, *DAMA-DMBOK2 Framework Production Editor* (2014)
28. A.Q. Gill, *Adaptive Cloud Enterprise Architecture* (WORLD SCIENTIFIC, 2015).
29. A. Izang, S.O. Kuyoro, F.Y. Osisanwo, Int. J. Comput. Appl. **133**, 975 (2016)
30. J. Gantz, D. Reinsel, THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadow s, and Biggest Grow Th in the Far East (2012).
31. U. Sivarajah et al., J. Bus. Res. **70**, 263 (2017)
32. N. Kshetri, Telecomm. Policy **38**, 1134 (2014)
33. PETER BUTTLER, Dataconomy (2017)
34. G. Hull, Ethics Inf. Technol. **17**, 89 (2015)
35. G. Lafuente, Netw. Secur. **2015**, 12 (2015)
36. S. Chauhan, N. Agarwal, A.K. Kar, Info **18**, 73 (2016)
37. E. Bertino, in *Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* (Springer Verlag, 2014), pp. 9–13
38. N. A. Shozi, J. Mtsweni, in *11th Int. Conf. Cyber Warf. Secur. ICCWS2016* (Academic Conferences and publishing limited, 2016), p. 296
39. R. Lu et al., IEEE Netw. **28**, 46 (2014)

Anwar *et al. J Wireless Com Network*    (2021) 2021:130

Page 30 of 30

40.  J. Moura, C. Serrão, arXiv Prepr. arXiv1601.06206 (2016)
41.  G.J. Selig, *Implementing Effective IT Governance and IT Management* (Van Haren Publishing, 2015).
42.  H. Singh, G. Singh, i-Manager's. J. Inf. Technol. **6**, 25 (2016)
43.  I.S. Rubinstein, Int. Data Priv. Law **3**, 74 (2013)
44.  N. F.-2017 I. I. C. on I. of and undefined 2017, ieeexplore.ieee.org (n.d.).
45.  D. Beer, Big Data Soc. **3**, 205395171664613 (2016)
46.  Z. He, Z. Cai, and J. Yu, IEEE Trans. Veh. Technol. **67**, (2018).
47.  M. Fleckenstein *et al.*, in *Mod. Data Strateg.* (Springer International Publishing, 2018), pp. 55–59.
48.  R.M. Alguliyev, R.T. Gasimova, R.N. Abbasli, Int. J. Mod. Educ. Comput. Sci. **9**, 28 (2017)
49.  S. Sagiroglu and D. Sinanc, in *Collab. Technol. Syst. (CTS), 2013 Int. Conf.* (IEEE, 2013), pp. 42–47

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Memoona J. Anwar**    is currently a PhD student in the faculty of engineering and information technology at the University of Technology, Sydney. She is also working as Head of Compliance and Digital Strategy in a Australia based company providing electronic identity verification services in APAC region. Her research focuses on privacy and security of PII, Digital Ecosystems, Blockchain, big data security, compliance and information security regulations and standards.

**Asif Q. Gill**    is a result-oriented adviser, consultant, inventor, and trainer with extensive experience in successfully delivering multi-million-dollar projects in various sectors including banking, consulting, education, finance, government, nonprofit, software and telco. He is also the Director of the SoS DigiSAS Lab (P.K.A. COTAR).

**Farookh K. Hussain**    is an Associate Professor at the School of Software and the Centre for Artificial Intelligence (CAI) of the University of Technology Sydney (UTS). He is also the Head of Discipline (Software Engineering) in the School of Software. Within CAI, he formed and provides leadership to the Cloud Computing group. His key research interests are in Cloud-of-Things, Cloud Computing, Internet-of-Things, Cloud-driven Data Analytics, Fog Computing and Block chain.

**Muhammad Imran**    is an Associate Professor in the College of Applied Computer Science at King Saud University, Saudi Arabia. He received a Ph. D in Information Technology from the University Teknologi PETRONAS, Malaysia in 2011. His research interest includes Internet of Things, Mobile and Wireless Networks, Big Data Analytics, Cloud computing, and Information Security. His research is financially supported by several grants. He has completed a number of international collaborative research projects with reputable universities. He has published more than 250 research articles in peer-reviewed, well-recognized international conferences and journals. Many of his research articles are among the highly cited and most downloaded. He served as an Editor in Chief for European Alliance for Innovation (EAI) Transactions on Pervasive Health and Technology. He is serving as an associate editor for top ranked international journals such as IEEE Communications Magazine, IEEE Network, Future Generation Computer Systems, and IEEE Access. He served/serving as a guest editor for about two dozen special issues in journals such as IEEE Communications Magazine, IEEE Wireless Communications Magazine, Future Generation Computer Systems, IEEE Access, and Computer Networks. He has been involved in about one hundred peer-reviewed international conferences and workshops in various capacities such as a chair, co-chair and technical program committee member. He has been consecutively awarded with Outstanding Associate Editor of IEEE Access in 2018 and 2019 besides many others.