

REVIEW

Open Access

Software architectures for big data: a systematic literature review



Cigdem Avci^{1*} , Bedir Tekinerdogan¹ and Ioannis N. Athanasiadis^{1,2}

* Correspondence: cigdem.avci@wur.nl

¹Information Technology Group, Wageningen University and Research, Hollandseweg 1, 6706 KN Wageningen, The Netherlands
Full list of author information is available at the end of the article

Abstract

Big Data systems are often composed of information extraction, preprocessing, processing, ingestion and integration, data analysis, interface and visualization components. Different big data systems will have different requirements and as such apply different architecture design configurations. Hence a proper architecture for the big data system is important to achieve the provided requirements. Yet, although many different concerns in big data systems are addressed the notion of architecture seems to be more implicit. In this paper we aim to discuss the software architectures for big data systems considering architectural concerns of the stakeholders aligned with the quality attributes. A systematic literature review method is followed implementing a multiple-phased study selection process screening the literature in significant journals and conference proceedings.

Keywords: Big data, Software architecture, Systematic literature review

Background

Various industries are facing challenges related to storing and analyzing large amounts of data. Big Data Systems become nowadays a very important driver for innovation and growth, by means of the insights and information that is obtained via the excessive processing of data. The business and application requirements vary depending on the application domain. Software architectures of big data systems have been previously studied sporadically/extensively. However, it is not easy to suggest a suitable software architecture for big data systems, when considering also both the application requirements and the stakeholder concerns [1].

The interactions and relations among the elements and all the elements as a whole that are necessary to reason about the system define the architecture of that system [2]. The architecture is constructed considering the driving quality attributes therefore it is important to capture those and analyze how these are satisfied by an architecture [3]. The requirements that are satisfied with the given architecture shall also match with the quality attributes.

In this study, we provide a systematic literature review (SLR) focused on the Software Architectures of the Big Data Systems in terms of the application domain, architectural



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

viewpoints, architectural patterns, architectural concerns, quality attributes, design methods, technologies and stakeholders. The challenging part of the study was screening the publications from various domains. The variety of the application areas of big data systems brings along the dissimilar representations of the system architectures with flexible terminologies. In order to achieve the requirements provided by different stakeholders which derive different architectural configurations, a proper architectural design with consistent terminology is essential. We aim to focus on the software architectures for big data systems considering architecture design configurations derived by architectural concerns of the stakeholders aligned with the quality attributes which are implicit in design of various systems.

The application areas of the big data systems vary from aerospace to healthcare [4, 5], and depending on the application domain, the functional and non-functional concerns vary accordingly, influencing both the architectural choices and the implementation of big data systems. To shed light on the experiences reported in the recent literature with deploying big data systems in various domain applications, we conducted a systematic literature review. Our aim was to consolidate reported experience by documenting architectural choices and concerns, summarizing the lessons learned and provide insights to stakeholders and practitioners with respect to architectural choices for future deployment of big data systems.

The study aims to investigate the big data software architectures based on application domains assessing the evidence considering the interrelation among the data extraction area and the quality attributes with the systematic literature review methodology which is the suitable research method. Our research questions are derived to find out in which domains big data is applied, the motivation for adopting big data architectures and to identify the existing software architectures for big data systems. We identified 622 papers with our search strategy. Forty-three of them are identified as relevant primary papers for our research. In order to identify various aspects related to the application domains, we extracted data for selected key dimensions of Big Data Software Architectures, such as current architectural methods to deal with the identified architectural constraints and quality attributes. We presented the findings of our systematic literature review to help researchers and practitioners aiming to understand the application domains involved in designing big data system software architectures and the patterns and tactics available to design and classify them.

Big data

The term “Big Data” usually refers to data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. In general, Big Data can be explained according to three V's: Volume (amount of data), Velocity (speed of data), and Variety (range of data types and sources). The realization of Big Data systems relies on disruptive technologies such as Cloud Computing, Internet of Things and Data Analytics. With more and more systems utilizing Big Data for various industries such as health, administration, agriculture, defense, and education, advances by means of innovation and growth have been made in the application areas. These systems represent major, long-term investments

requiring considerable financial commitments and massive scale software and system deployments.

The big data systems are applicable to the data sets that are not tolerable by the ability of the generic software tools and systems [6]. The contemporary technologies within the area of cloud computing, internet of things and data analytics are required for the implementation of the big data systems. Such massive scale systems are implemented using long term investments within the industries such as health, administration, agriculture, defense and education [7].

Big data systems analytic capability strongly depends on the extreme coupling of the architecture of the distributed software, the data management and the deployment. Scaling requirements are the main drivers to select the right distributed software, data management and deployment architecture of the big data systems [8]. Big data solutions led to a complete revolution in terms of the used architecture, such as scale-out and shared-nothing solutions that use non-normalized databases and redundant storage [9].

As a sample domain, space business already benefits from the big data technology and can continue improving in terms of, for instance horizontal scalability (increasing the capacity by integrating software/hardware) to meet the mission needs instead of procuring high end storage server in advance. Besides multi-mission data storage services can be enabled instead of isolated mission-dedicated warehouse silos. Improved performance on data processing and analytics jobs can support activities such as early anomaly detection, anomaly investigation and parameter focusing. As a result, big data technology is transforming data-driven science and innovation with platforms enabling real time access to the data for integrated value.

The trend is to increase the role of information and value extracted from the data by means of improving the technologies for automatic data analysis, visualization and use facilitating machine learning and deep learning or utilizing the spatio-temporal analytics through novel paradigms such as datacubes.

Systematic reviews

The systematic literature review is a rigorous activity that is applied screening the identified studies and evaluating such studies based on the defined research questions, topic areas or phenomenon of interest. As a result of the evidence gathered for a particular topic, the gaps can be investigated further with supporting studies.

Evidence-based research is successfully conducted initially in the field of medicine and similar approaches are adopted in many other disciplines. Among the goals of the evidence-based software engineering, the quality improvement, assessing the application extent of the best practices for the software-intensive systems can be listed. Besides the evidence based guidelines can be provided to the practitioners as a result of such studies. Considering the benefits of the evidence based research, its application is valuable also in the software engineering field.

The systematic literature review shall be transparent and objective. Defining clear inclusion/exclusion criteria for the selected primary studies is critical for the accuracy and consistency of the output of the review. Well defined inclusion/exclusion criteria minimizes the bias and simplifies the integration of the new findings.

Software architectures

The software architecture is the high-level representation and definition of a software system providing the relationships between architectural elements and sub-elements with a required level of granularity [3, 10]. Views and beyond is one of the approaches to define and document the software architectures [11]. Viewpoints are generated to focus on relevant quality attributes based in the area of use for the stakeholder and more than one viewpoint can be adopted depending on the complexity of the defined system. In order to solve common problems within the architecture, architectural patterns are designed within the relevant context. Architectural patterns, templates and constraints are consolidated and described in viewpoints.

Research method

In this study, the SLR is applied for the software architectures of big data systems following the guidelines proposed in [12, 13] by Kitchenham and Charters. The review protocol that is followed is defined in the following sections.

Review Protokol

In order to apply the systematic literature review, a review protocol shall be defined with the methods to be used for reducing the overall bias. Figure 1 below shows the review protocol that is followed throughout this study:

The research questions are defined using the objectives of the systematic review as discussed in section 3.2 which is followed by drawing the scope (time range and publication resources) and the strategy (section 3.3). The search strategy is shaped by conducting pilot searches to form the actual search strings.

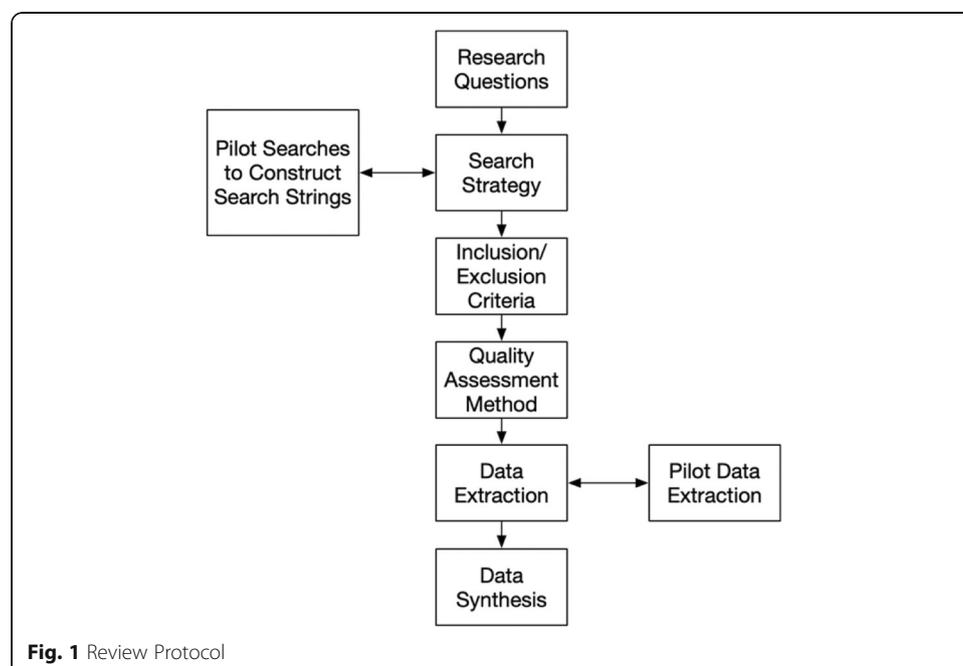
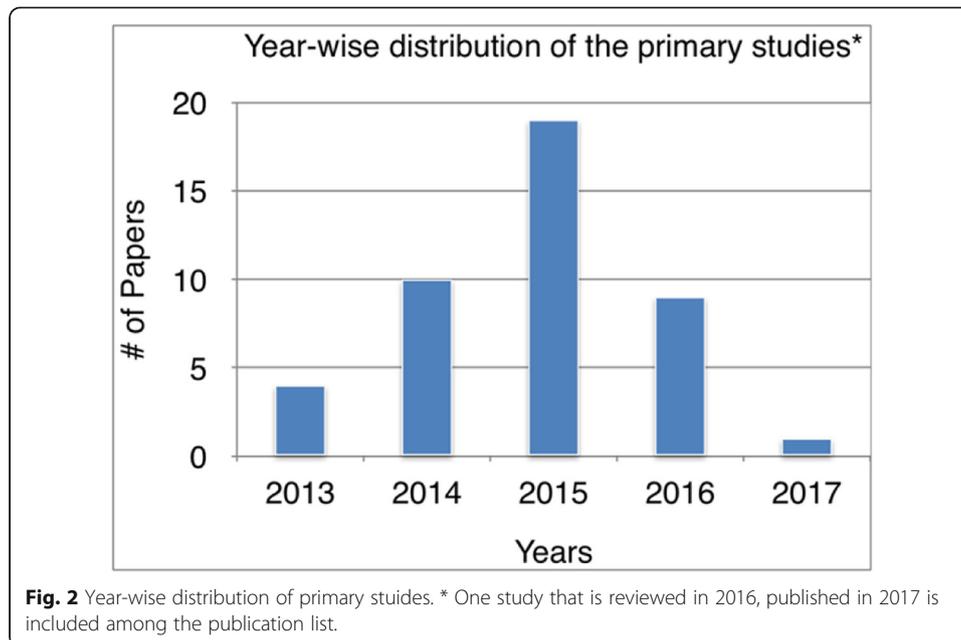


Fig. 1 Review Protokol



The appropriate definition of the search string reduces the bias and helps to achieve the target precision. The inclusion/exclusion criteria (section 3.4) is defined as the next step. The primary studies are filtered applying the inclusion/exclusion criteria. The success of the study selection process is assessed via the peer reviews of the authors.

The selected primary studies are passed through a quality assessment (section 3.5). Afterwards the data extraction strategy is built to gather the relevant information from the selected set of studies (section 3.6). The data extraction form is constructed and filled with the corresponding output to present the results of the data synthesis.

Research questions

Constructing the research questions in the right way increases the relevancy of the findings and the accuracy of the SLR output. Validity and significance of the research questions is critical for the target audience of the SLR. Considering the fact that we are investigating the software architectures of the big data systems, the following research questions are defined to examine the evidence:

RQ.1: In which domains is big data applied?

RQ.2: Why are the big data architectures applied?

RQ.3: What are the existing software architecture approaches for big data systems?

RQ.4: What is the strength of evidence of the study?

Search strategy

In this section, our search strategy is defined to find as many primary studies as possible regarding the research questions listed in section 3.2.

Scope The search scope of our study consists of the publication period as January 2010 and December 2017 and search databases such as: IEEE Xplore, ACM Digital Library, Wiley Inter Science Journal Finder, ScienceDirect, Springer Link. Our targeted search items were both journal and conference papers.

Method Automatic and manual search are applied to search the databases.

In order to gather the right amount of relevant studies out of a high number of search process outputs, the selection criteria shall be aligned with the objectives of the SLR. A search strategy with high recall causes false positives and a precise search strategy will narrow down the outcome.

Initially a manual survey is conducted to analyze and bring out the search strings. Using this outcome, search queries are formed and run to obtain the right set of studies with optimum precision and recall rates.

The right method shall be applied to design the search strings with the relevant set of keywords which is critical for optimum retrieval of the studies. The keywords within the references section shall be eliminated and the keywords of the authors shall have higher weight. By means of the concrete set of keywords, the final search string is formed.

After the construction of the search strings, they are semantically adapted to the electronic data sources and extended via OR and AND operators. A sample search string is presented below:

Query 1:

```
((("Abstract": "Big Data" OR "Publication Title": "Big Data") AND (p_Abstract: "Software Architecture" OR "Abstract": "System Architecture" OR "Abstract": "Cloud Architecture" OR "Publication Title": "Architecture")))
```

Other search strings can be found in [Appendix 1](#). Eliminating the duplicate publications, 662 papers are detected.

Study selection criteria

In order to omit the studies that are irrelevant, out of scope or false positive, aligned with the SLR guidelines, we apply the following exclusion criteria:

- EC 1: Papers that does not state a big data architecture description, or a big data application that applies an architecture.
- EC 2: Papers that are not related to a field of computer science.
- EC 3: Papers are written in different language than English
- EC 4: Workshop papers
- EC5: Papers that does not discuss (or discuss partially) the big data architecture
- EC6: Papers don't explicitly present the architectural representation/view/model

After the exclusion criteria is applied, the reduced amount of studies are presented in [Table 1](#) where after applying EC1-EC5, which narrowed down our corpus to 341 papers. After applying criterion EC6, we concluded with 43 papers.

Table 1 Search results after the application of the elimination criteria

Source	After Applying Search Query	After EC1-EC5 Applied	After EC6 Applied
IEEE Xplore	113	34	20
ACM Digital Library	111	17	9
Wiley Interscience	100	2	2
Science Direct	48	5	9
Springer	250	20	3
Total	622	341	43

Study quality assessment

The quality of the selected studies shall also be assessed based on a checklist. The aim of the quality assessment is the improvement of the relevance and importance of the results via fine tuning the inclusion/exclusion criteria, driving the interpretation of the results (data extraction and synthesis) and recommendations. The checklist should be constructed considering the factors for each study. The factors that have a biasing effect on the outcome are used to form the quality checklist presented in Table 2. The studies are ranked according to the three point scale with the corresponding assigned scores (yes = 1, somewhat = 0.5, no = 0). The assessment results can be found in [Appendix 2](#) (List of Primary Studies).

Data extraction

The data is extracted from the 43 studies selected targeting the review questions and study quality criteria. The standard columns such as title, date, author are included in the data extraction form, in addition to the data extraction columns aligned with the research questions which are application domain, architectural viewpoints, patterns, concerns, quality attributes, etc. The field categories and fields are listed in Table 3.

Data synthesis

After gathering the data aligned with the data extraction form, the data is synthesized to obtain the answers for the predefined research questions. The qualitative

Table 2 Quality Checklist

No	Question
Q1	Are the aims of the study is clearly stated?
Q2	Are the scope and context of the study clearly defined?
Q3	Is the proposed solution clearly explained and validated by an empirical study?
Q4	Are the variables used in the study likely to be valid and reliable?
Q5	Is the research process documented adequately?
Q6	Are the all study questions answered?
Q7	Are the negative findings presented?
Q8	Are the main findings stated clearly in terms of creditability, validity and reliability?
Q9	Do the conclusions relate to the aim of the purpose of study?
Q10	Does the report have implications in practice and results in research area for big data software architecture?

Table 3 Data Extraction Columns

Category	Data Extraction Columns
Application Domain Categories	Cyber security IOT/Smart cities Social big data Incident/Anomaly detection Healthcare Aerospace
Stakeholders	Enterprise managers Strategic suppliers Customers Manufacturers Operators Managers Technical staff Business analysts Data scientists Operation managers Data platform designers
Key Concerns	Integration concerns Functional concerns Non-functional concerns
Motivation for adopting a big data architecture	Description/Motivation
Architectural Approaches	Hybrid and others
Architectural Models/Viewpoints	Decomposition, deployment and others
Architectural Tactics/Patterns	Flow chart, layered, cloud based and others

assessment is covered interpreting the content of the data and assessing its relevance and relation with other categories/columns while the quantitative assessment is accomplished calculating the quality score for reporting, relevance, rigor and credibility.

We investigated whether the qualitative results can lead us to explain quantitative results. We realized that application architectures are seldom based on a reference architecture in the papers that we reviewed. The analyzed papers mainly elaborate and evaluate the target area using test cases, experiments and other methods that are quantitative in nature however the data used is not available for each case. They also target the research audiences beyond computer science, and the reported data from the computer science point of view is rather limited. The coverage of a possible statistical analysis is not sufficient, however qualitative analysis is applicable for our case.

The data synthesized is transferred to tabular and graphical representations to present the reader an enriched and meaningful translation of the findings which enables and simplifies the process of comparisons across categories, application areas and studies. Both qualitative and quantitative findings are valid inputs for future application areas within big data software architectures.

Grading of recommendation assessment, development and evaluation (GRADE)

GRADE (Grading of Recommendations, Assessment, Development and Evaluations) [2] framework is a systematic approach with a transparent framework to gather and present evidence and measure the quality of it. The method assesses the likelihood of bias at the outcome and is widely adopted by ensuring a transparent link between the evidence and recommendations. The results of application of the GRADE methodology is presented in Section 4.4 for research question RQ4.

Results

Overview of the reviewed studies

The selected 43 primary studies are briefly summarized below:

- **Study 1:** This article proposes a five-level of fusion model, in order to process the big datasets with complex magnitudes. Hadoop Processing Server is used. A four-layered network architecture is presented.
- **Study 2:** presents AsterixDB, a Big Data Management System. Its target application areas can be listed as web data warehousing, social data storage and analysis, etc. It implements a flexible NoSQL style data model and transaction support similar to that of a NoSQL store.
- **Study 3:** A Big Data architecture for construction waste analytics is proposed. A graph database (Neo4J) and Spark is employed. Building Information Modelling (BIM) is investigated for possible extensions.
- **Study 4:** In order to design and deploy the scientific applications into the cloud in an agile way, the Scientific Platform for the Cloud (SPC) is developed. The platform embodies a web interface, job scheduling, real-time monitoring etc. Population Genetics, Geophysics, Turbulence Physics, DNA analysis, and Big Data can be listed among the application domains.
- **Study 5:** The software architecture presented in this paper is developed to support gathering of IoT sensor-based data in a cloud-based system. The use case is the SMARTCAMPUS project.
- **Study 6:** A scalable workflow-based cloud platform is implemented based on Hadoop, Spark, Cassandra, Docker, and R. High performance and productivity is aimed. Data storage and management, data mining and machine learning capabilities are involved.
- **Study 7:** WaaS is a standard and service platform architecture for big data. Four service layers implements four components accordingly.
- **Study 8:** The study describes the architecture-centric agile big data analytics which is a methodology that combines big data software architecture analysis and design together with agile practices.
- **Study 9:** The system architecture of the City Data and Analytics Platform is introduced in this paper. A smart city testbed, SmartSantander, is implemented based on this architecture.
- **Study 10:** This paper discusses how to design big data system architectures using architectural tactics considering the design tradeoffs. A healthcare informatics use case is illustrated.

- **Study 11:** Private cloud computing platform which is developed for the China Centre for Resources Satellite Data and Application (CCRSDA) and its architectural design is discussed in this paper.
- **Study 12:** Semantic-based heterogeneous multimedia retrieval architecture is described in this paper. A NoSQL-based approach to process multimedia data in distributed in parallel and a map-reduce based retrieval algorithm are employed.
- **Study 13:** A cloud computing-based system architecture is presented for implementation of a production tracking and scheduling system. A prototype system is implemented and validated in terms of its efficiency.
- **Study 14:** A distributed system architecture for text-based social data (Twitter, YouTube, The New York Times etc) is introduced in this paper. HDFS, Map-reduce, and message service analysis are utilized to analyze reputation, social trends, and customer reactions.
- **Study 15:** The Alexandria provides a framework and platform for big-data analytics and visualisations mainly for text-based social media data. REST-based service APIs are heavily used within the system architecture.
- **Study 16:** Software architectures for Web Observatories are discussed in this paper, for processing real time web streams.
- **Study 17:** A generic system architecture is proposed in this paper, which focuses on running big data workflows in the cloud. Big data workflows are investigated in Amazon EC2, FutureGrid Eucalyptus and OpenStack clouds.
- **Study 18:** Big Data and data warehousing architectures and design are discussed in this book for the next-generation data warehouse.
- **Study 19:** A general system architecture for big data analytics is proposed in this paper, focusing on manufacturing industries.
- **Study 20:** This paper discusses the big data with the concept of e-learning and academic environment. A three-step system architecture presented based on a Cloud environment. Graphical Gephi tool is used for analyzing unstructured data.
- **Study 21:** An agent oriented architecture is presented in this paper and the proposed for the IoT domain.
- **Study 22:** CityWatch framework, which is designed for data sensing and dissemination by using the data collected from Dublin. Two prototype applications are implemented.
- **Study 23:** This paper discusses real time big data application architecture challenges. Initial implementation is Hadoop-based which is later replaced with a custom in-memory processing engine.
- **Study 24:** This paper focuses on the analysis of the data produced by camera sensors for intruder detection and construction of barrier. A three-layered big data analytics architecture is designed for the study.
- **Study 25:** A Big Data architecture system design is introduced in this paper for global financial institutions. Hadoop and no-SQL are applied within the architecture, besides the architecture complies with the data integration, transmission and process orchestration requirements of the application domain.
- **Study 26:** A cloud architecture for healthcare is proposed in this paper. In order to use heterogenous devices as data sources, cloud middleware is utilized. Besides different healthcare platforms are integrated via the cloud middleware.

The paper also mentions the security and management concerns and emphasizes the importance of the standardized interfaces for the integration with medical devices.

- **Study 27:** The paper discusses the social CRM by means of architectural perspectives using five layers.
- **Study 28:** In this paper, a technology independent reference architecture is proposed for big data systems. Real use cases are investigated and implementation technologies and products are classified.
- **Study 29:** Within the domain of educational technology, based on the Experience API specification, a big data software architecture is introduced in this paper. The data generated as a result of the learning activities of a course is used for the data analytics.
- **Study 30:** A big data analytical architecture for a remote sensing satellite application is described in this paper. The data gathered from an earth observatory system is analysed in real time and stored using Hadoop.
- **Study 31:** A novel mobile-based end-to-end architecture is described in this study, for the healthcare domain. The architecture is specialized for live monitoring and visualization of life-long diseases. The architecture is based on web services and SOA and a supporting Cloud infrastructure.
- **Study 32:** This paper proposes a two-layered cloud architecture for real-time public opinion monitoring model.
- **Study 33:** Based on a search cluster for data indexing and query, a cloud service architecture is introduced in this paper. The architecture has the capability to integrate with Hadoop and Spark. REST APIs are employed for access.
- **Study 34:** An analytical big data framework is presented in this paper for the smart grid domain. EU funded project BIG and the German funded project PEC are the case studies.
- **Study 35:** A big data application architecture for smart cities is implemented within this study. Identify and responding to anomalous and hazardous events in real time is the main goal of the designed architecture. Sensor data is used and sequential learning algorithms are adopted.
- **Study 36:** The architecture presented within this paper is for both offline and real time processing and applied for the recommender systems.
- **Study 37:** A cloud based big data software application architecture is presented in this paper. The target application domain is research/science. Open source software paradigm is emphasised.
- **Study 38:** A real time data-analytics-as-service architecture with RESTful web services is presented in this paper. The architectural challenges are discussed by means of big data processing frameworks, reliability, real time performance and accuracy.
- **Study 39:** In this paper, Banian system's 3-layer system architecture is discussed. The layers are listed as follows: storage, scheduling, application. The results are compared with Hive.
- **Study 40:** A novel architecture of big data for assessing the city traffic state is proposed in this paper. A real time, highly scalable system is among the architectural goals. The implementation is based on Hadoop and Spark. Various clustering methodologies like DBSCAN, K-Means, and Fuzzy C-Means are implemented.

- **Study 41:** Embodying the big data analytics and service oriented patterns, a big data based analytics system architecture is presented in this paper. The availability and accessibility are the main architectural goals.
- **Study 42:** The Cloud Grid (CG) is discussed in in this study for the cloud-based power system operations within the smart grid domain. CG covers the concepts of internet of things (IoTs) together with service-oriented cloud computing and big data analytics. Besides, the architectural constraints related to high performance computing and smart grid are covered within the capabilities of the CG.
- **Study 43:** The complex event processing framework H2O is presented in this study. The framework has the capability of supporting the queries over realtime data which are hybrid online and on-demand.

Figure 2 presents the number of selected published 43 papers per year.

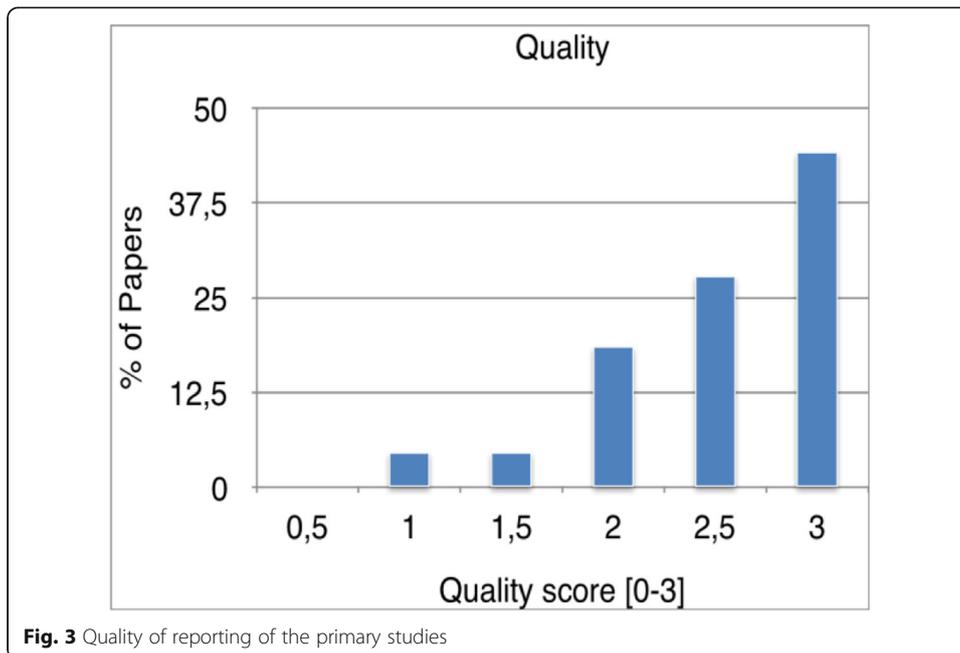
Table 4 presents the publication channel, publisher and the type of the selected primary studies as an overview. It can be derived from Table 4 that the selected primary studies mostly published by IEEE, Elsevier and ACM that are accepted as highly ranked publication sources. While “Conference on Quality of Software Architectures.” and “SIGMOD International Conference on Management of Data” are significant conferences, “Network and Computer Applications” and “VLDB Endowment” are remarkable journals for the software engineering domain. Besides, it can be indicated that the publication channels that have high impact in the domains other than software engineering are raising the number of papers with emphasis on big data system architectures within their publications. “Renewable and Sustainable Energy”, “Journal of Cleaner Production” and “Journal of selected topics in applied earth observations and remote sensing” can be listed among the remarkable publication changes from other domains.

Research methods

Research method has a critical role within the empirical studies. In order to converge valid and reliable outcomes, clear cut research methodologies should be applied and reported in the selected primary studies. The types of the research methods can be listed as “Case Study” (in depth investigation with a real life context), “Experiment” (scientific procedure to test a hypothesis) and “Short Example”. It can be derived from Table 5 that there is not a tendency towards a research method, considering the fact that the gap between the percentages of the methodologies is not wide. Nevertheless, experimentation is used more often comparing to case studies and short examples to evaluate the system architectures.

Methodological quality

We evaluated the selected primary studies quality using 4 dimensions of quality which are the quality of reporting (Fig. 3), rigor (Fig. 4), relevance (Fig. 5) and credibility (Fig. 6). The questions are grouped as follows: Q1, Q2 and Q3 assess the quality of reporting, while Q4, Q5 and Q6 focus on the rigor. Q7 and Q8 are for assessing credibility, and finally Q9 and Q10 search for relevance. The overall quality checklist results can be found in [Appendix 3](#).



The trustworthiness of the primary studies were assessed in the context of rigor. The distribution of the quality scores of the primary studies from the dimension of rigor is presented in Fig. 4. We observe that the quality of rigor of the primary studies scored around the average values. While none of the papers scored below 0.5, the top scored papers are less than 10%. 30% of the studies scored 1 and similarly, the primary studies scored 1.5 are marginally above then 30%. The overall rigor quality appears as average.

The third quality dimension to report is relevance, which is illustrated in Fig. 5. As it can be inferred from Fig. 5, the primary studies are quite relevant to their

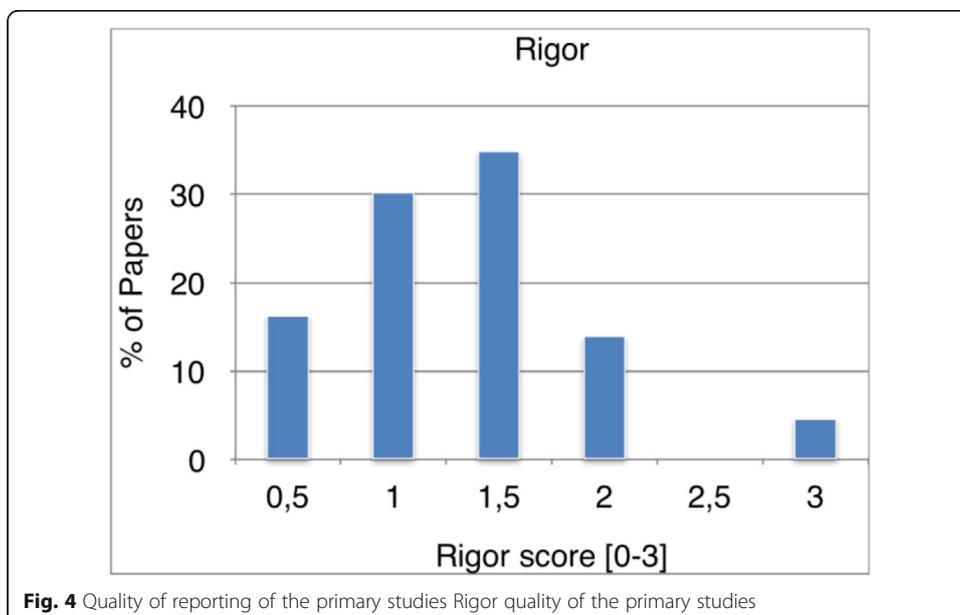


Table 4 Distribution of the studies over publication channel

Publication Channel	Publisher	Type	#
Computers in biology and medicine	Elsevier	Article	1
Journal of Network and Computer Applications	Elsevier	Article	1
IEEE Software	IEEE	Article	1
Journal of Building Engineering	Elsevier	Article	1
Proceedings of the VLDB Endowment	ACM	Article	1
Transactions on Embedded Computing Systems (TECS)	ACM	Article	1
Journal of Systems and Software	Elsevier	Article	1
Journal of Cleaner Production	Elsevier	Article	1
Wisdom Web of Things	Springer	Chapter	1
Modeling and processing for next-generation big-data technologies	Springer	Chapter	1
Data Warehousing in the Age of Big Data	Elsevier	Chapter	1
Digital Libraries	ACM	Conference	1
Procedia Economics and Finance	Elsevier	Conference	1
Multimedia Big Data (BigMM)	IEEE	Conference	1
Cloud Computing and Big Data	IEEE	Conference	2
Big Data (BigData Congress)	IEEE	Conference	3
International Conference on Utility and Cloud Computing	ACM	Conference	1
Communications (COMM)	IEEE	Conference	1
Cloud Computing Research and Innovation	IEEE	Conference	1
IEEE World Congress on Services (SERVICES)	IEEE	Conference	1
Big Data Analysis (ICBDA)	IEEE	Conference	1
Services Computing (SCC)	IEEE	Conference	1
ASE Big Data & Social Informatics	ACM	Conference	1
International Conference on Big Data	IEEE	Conference	1
Conference on e-Business, e-Services and e-Society.	Springer	Conference	1
Frontiers in Education Conference	IEEE	Conference	1
Web Intelligence (WI) and Intelligent Agent Technologies (IAT)	IEEE	Conference	1
Conference on Quality of Software Architectures	ACM	Conference	1
SIGMOD International Conference on Management of Data	ACM	Conference	1
Proceedings of the 24th International Conference on World Wide Web	ACM	Conference	1
Proceedings of the 2015 Conference on research in adaptive and convergent systems.	ACM	Conference	1
Future internet of things and cloud (FiCloud), 2014 international conference on.	IEEE	Conference	1
Transactions in GIS	Wiley	Journal	1
Renewable and Sustainable Energy	Elsevier	Journal	1
Tsinghua Science and Technology	IEEE	Journal	2
Transactions on Emerging Telecommunications Technologies	Wiley	Journal	1
Big Data Research	Elsevier	Journal	1
Journal of selected topics in applied earth observations and remote sensing	IEEE	Journal	1
IEEE Transactions on Big Data	IEEE	Journal	1

research questions. About 50% of the studies scored the highest relevance score (i.e. 2), whereas the remaining studies mostly scored around 1–1.5 and only a few studies had a very low score. Therefore, we conclude that the selected primary studies are of high quality relevance.

Table 5 The studies corresponding research methods

Research Method	Studies	Number	Percent
Case Study	4, 6, 7, 8, 13, 17, 18, 19, 23, 25, 34, 38	12	28%
Experiment	2, 5, 11, 12, 14, 15, 21, 22, 24, 31, 32, 33, 35, 37, 39, 40	16	37%
Short Example	1, 3, 9, 10, 16, 20, 26, 27, 29, 28, 30, 36, 41, 42, 43	15	35%

The credibility quality dimension is summarized in Fig. 6. The studies mostly have slightly below average credibility of evidence. Around 50% of the studies achieved score 1, which we considered fair, and around 30% scored 0.5 indicating a poor quality of credibility. We therefore conclude that the studies barely discuss major conclusions, and poorly list positive and negative findings.

The overall quality scores are shown in Fig. 7, incorporating the quality scores for quality of reporting, relevance, rigour and credibility of evidence. Around 70% of the studies are above average quality (i.e. with a score greater than 4.5). 11% of the papers is in the category of poor quality (< 5) and 29% of the papers have high quality scores (> 7).

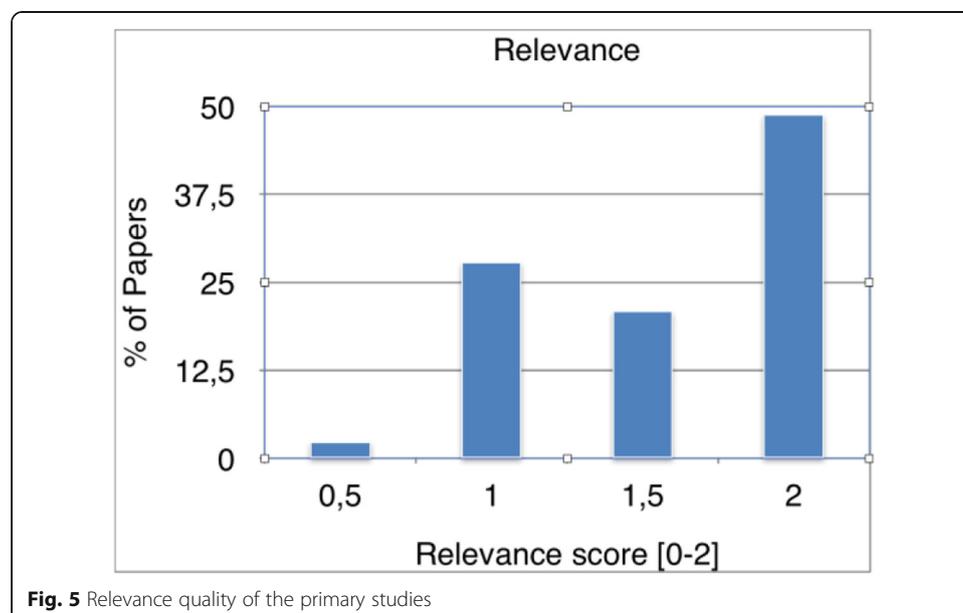
The distribution of the quality attributes per domain is presented in Fig. 8:

Systems investigated

In this section, we present the results which are extracted from 43 selected primary studies in order to answer the research questions.

RQ.1: in which domains is big data software architectures applied?

After screening the selected 43 primary studies, we extracted seven target domains and other domains that have less number of occurrence within the primary studies. The main domains can be listed as follows: Social Media, Smart Cities, Healthcare, Industrial Automation, Scientific Platforms, Aerospace and Aviation, and Financial Services (See Fig. 9).

**Fig. 5** Relevance quality of the primary studies

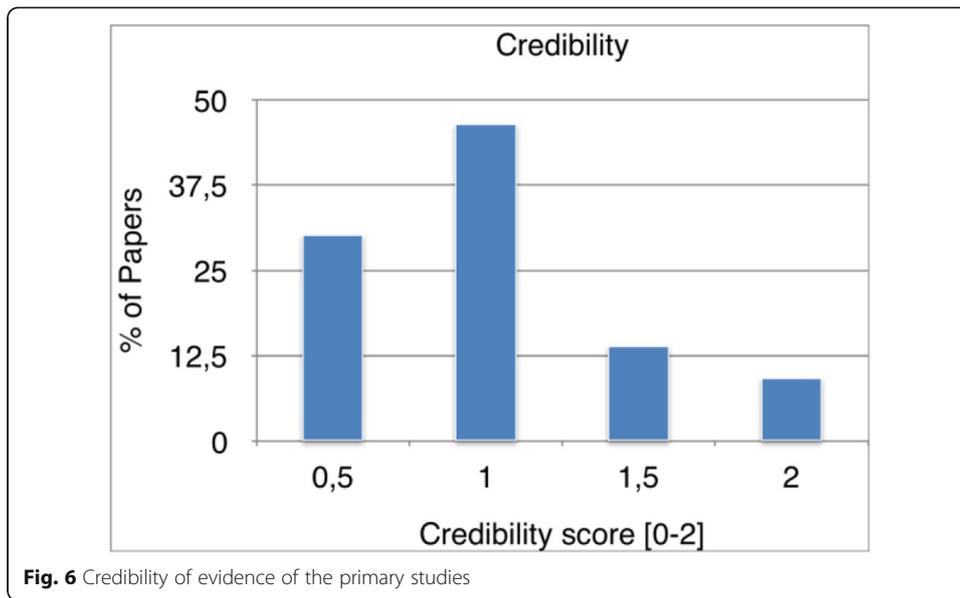
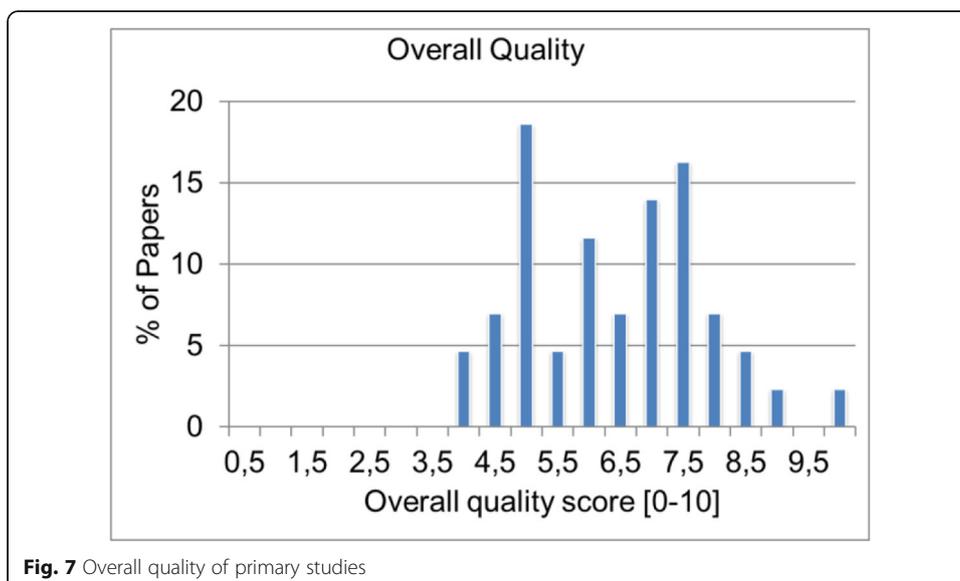


Table 6 shows the domain categories and their subcategories. For the smart cities domain, the subcategories are smart grid, surveillance system, traffic state assessment, smart city experiment testbed, network security and wind energy. Under the smart grid category, study 34 discusses a smart home cloud-based system for analyzing energy consumption and power quality, while study 42 describes a power system with a cloud-based infrastructure. Within the surveillance systems subcategory, study 24 presents a barrier coverage and intruder detection system, and study 18 introduces a system to track potential threats in the perimeters and border areas. Study 40 presents a cloud-based real-time traffic state assessment system. For the smart city experiment testbed, studies 5, 9, 22 discuss system infrastructures that analyze real-time and historical data from the perspectives of parking occupation, heating and traffic regulation. Study 35 is listed under the network security subcategory for smart pipeline monitoring system.



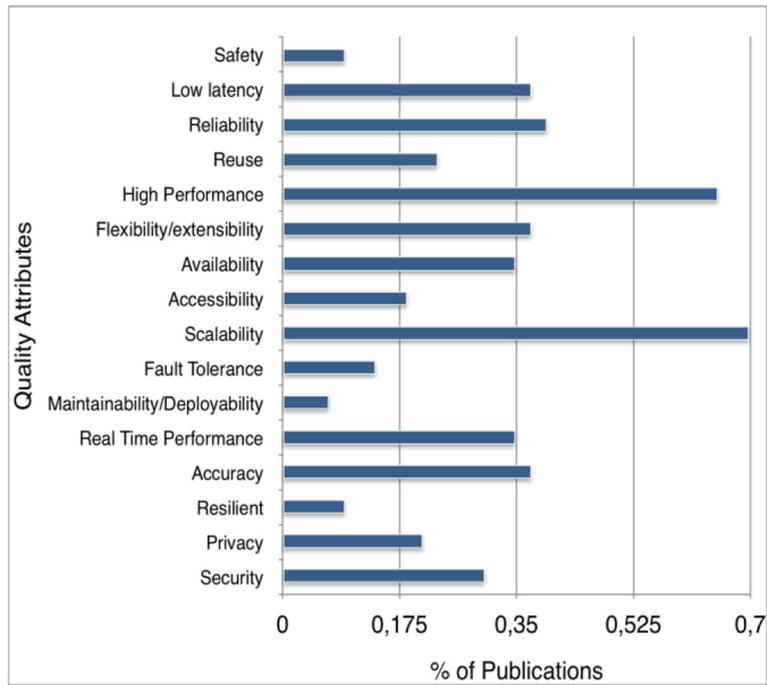


Fig. 8 Quality attribute distribution for all domains

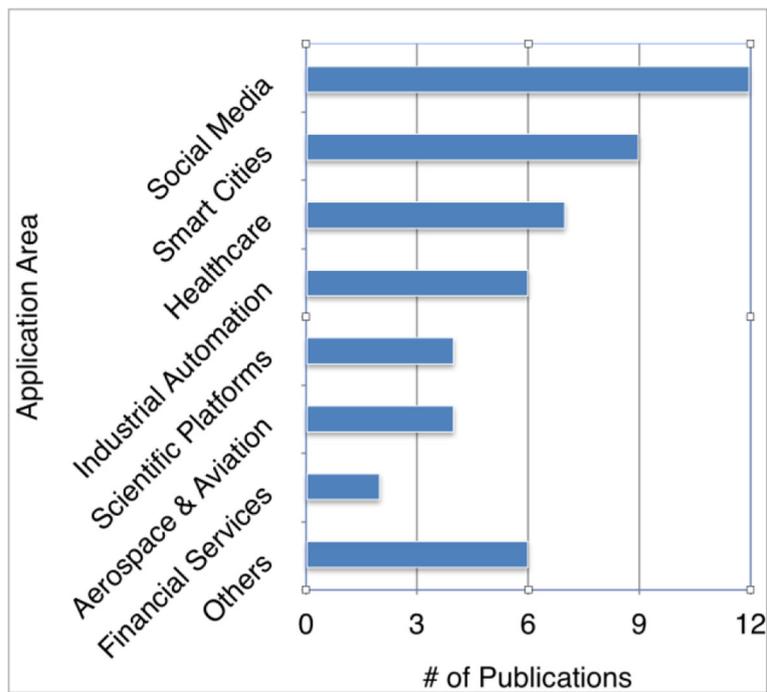
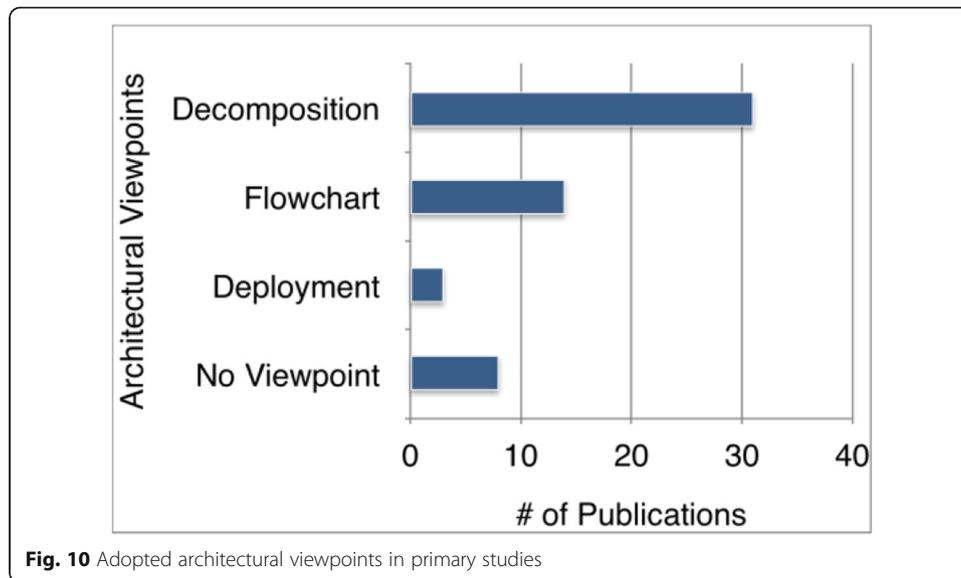


Fig. 9 Domain distribution of primary studies



Under the sub-category of wind energy, study 18 discusses a system which uses climate data to predict the most optimal usage of wind energy.

The category of social media consists of the following subcategories: public opinion monitoring, query suggestion and spelling correction, reference architectures of social media systems, web observatories, travel advising, semantic-based heterogeneous multimedia retrieval, web data warehousing, social data storage and analytics, social network analysis. Studies 15 and 32 are listed under the public opinion monitoring subcategory which covers exploration and visualization of social media data in connection with a given domain. Study 23 falls under the sub-category of query suggestion and spelling correction and describes the architecture behind Twitter's real-time related querying service. In study 28, a technology independent big data system reference architecture is presented within the social media domain. Web observatories are introduced in study

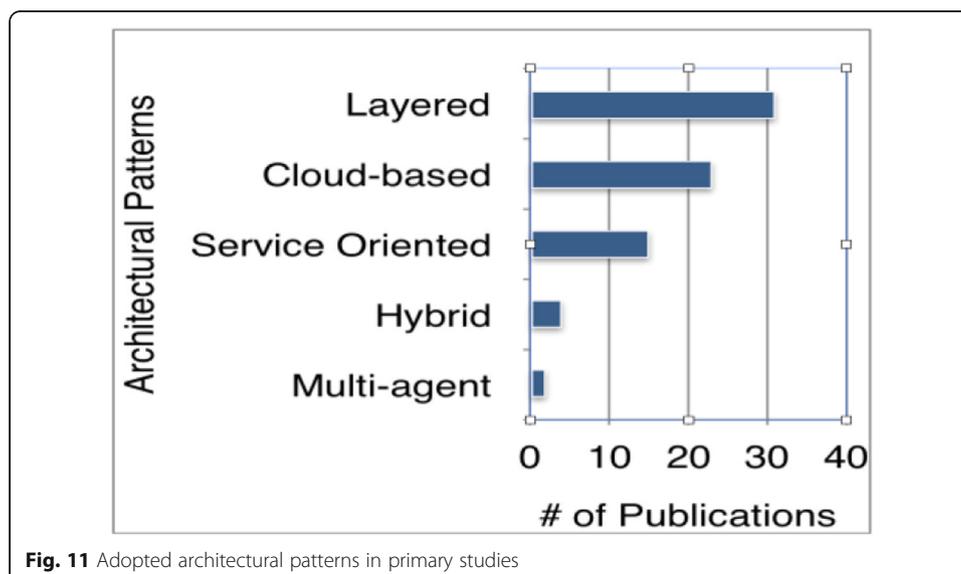


Table 6 Identified domains of big data software systems

Domain	Identified Subcategory	Details	Study Citation #
	Smart Grid	Smart home energy consumption and power quality into a cloud-based infrastructure	34
		Power system with a cloud-based infrastructure	42
Smart Cities	Surveillance System	Barrier coverage, Intruder detection	24
		The potential threats are to be detected/classified/located to secure the perimeters and border areas	18
	Traffic State Assessment	Cloud based, real time	40
	Smart City Experiment Testbed	Real-time and historical data, parking occupation, heating regulation, traffic etc.	5, 9, 22
	Network Security	Smart Pipeline Monitoring, Fiber optic sensors, detect events threatening pipeline safety	35
	Wind Energy	Prediction of the optimum wind energy use (Climate data)	18
Social Media	Public Opinion Monitoring	Social media data is explored and visualized (in a given domain)	15, 32
		Query Suggestion and Spelling correction	Twitter's architecture supporting the real-time query suggestion and spelling correction service
	Reference Architecture for big data systems in social media domain	Technology independent	28
	Web Observatories	Web scale data gathering, storing and analysis	16
	Travel Advising	Monitor, Troubleshoot, Management Reports, Anomalies, etc.	8
	Semantic-based Heterogeneous Multimedia Retrieval	Flickr, Wikipedia and Youtube are the sites for data acquisition	12
	Web Data Warehousing, Social Data Storage and Analytics	Cell phone event analytics, tweet analytics, analyzing the information streams (behavioral, events)	2
	Social Network Analysis	Process the social data in real time (terms and sentiment analysis)	14
Industrial Applications	Environmental Sustainability	Big data based analytics for product lifecycle. Cleaner sustainable production.	41
		Construction waste, Mobile app	3
	Production Tracking and Scheduling	Remote production data capture and tracking, and intelligent optimization	13
	Manufacturing	Event-based predictions of their manufacturing processes	19
	Automotive Industry	Analyzing Driving Competency from the Vehicle Data	17
	Electric power industry	Short term electricity load is predicted with the historical data, big data workflows	6
Healthcare	Brain and health monitoring system	Brain health and mental disorders	7
		Heart rate, ECG, and body temperature are the datasets considered for analysis	1
		Monitoring and visualization states of epileptic diseases	31
	Improving healthcare quality and costs	Complex data processing, clinical quality measure analytics, proactive care management analytics.	18, 10
	Patient Monitoring	Emergency situations and medical records	21
	Interconnection of Healthcare Platforms	Integration of the medical devices with cloud computing middleware services via standardized	26

Table 6 Identified domains of big data software systems (*Continued*)

Domain	Identified Subcategory	Details	Study Citation #
		interfaces	
Scientific Platforms	Digital libraries for scientific data management	Big data management infrastructure, reuse, public cloud OASIS	37
	Scientific Platform for the Cloud	Rapid design and deployment of scientific apps to cloud (DNA Analyser, population genetics, dynamics of planetary mantles)	4
	E-learning	Applying Big Data to e-Learning and analyze influence on the academic environment	20
	Learning Analytics	Learner's performance prediction, exploring the patterns and learning paths and learner's behavior paths	29
Earth Observation & Aviation	Earth Observation	Advanced Synthetic Apertures Radar- earth data (ASAR), product classifications (i.e., Land and Sea areas), Satellite Images, download, process, wies	1, 11, 30
	Aviation Maintenance and Optimisation	Real time fault diagnosis, fuel consumption optimisation, and prediction of the maintenance need	10
Financial Services	Banking	Cost reduction preserving the system scalability and flexibility, more regulatory requirements and various data sources	25
	Social Customer relationship management	Social CRM, Enterprise 2.0 ^o , CRM 2.0, social commerce	27
Other	Ambient Intelligence	Smart House	21
	Recommendation System	Near real time optimum online recommendation	36
	Anomaly detection system	Logs and monitoring metrics real time analysis	38
	Hive alternative	interactive cross-platform query	39
	Query Engine	Hybrid online on-demand query engine	43
	Trace Analyzer	Trace data is analyzed to predict the workload and levels, trends and seasonality factors are interpreted	33

The table is presented respecting the content of the primary studies

16 where gathering, storing and analyzing the data at web scale is the main focus. To monitor and troubleshoot a travel advising system, a big data architecture is defined in study 8. Semantic-based heterogeneous multimedia retrieval domain subcategory appears in study 12, in which a big data system is utilized for acquisition and analysis of data from specific websites such as Flickr, Youtube and Wikipedia. Study 2 includes the web data warehousing, social data storage and analytics subdomain. It covers cell phone event analytics, tweet analytics, behavioral data analysis of information streams about events. Study 14 is applied on social network analysis sub-domain, presenting a system that processes the social data in real time.

The industrial applications domain includes 5 subcategories which are environmental sustainability, production tracking and scheduling, manufacturing, automotive and electric power. Study 41 discusses big data analytics for product lifecycle and cleaner manufacturing. Study 3 targets construction waste analytics. Both are listed in the subdomain environmental sustainability. Production tracking and scheduling subdomain appears in the study 13, discussing a system for capturing and analysing the remote production data in terms of tracking and intelligent optimisation. Within the subdomain manufacturing, study 19 covers a system that makes event-based predictions of

manufacturing process. Study 17 is presented under the automotive subdomain, introducing a system for analyzing driving competency from the vehicle data. Electric power subdomain includes study 6, discussing a big data system that uses historical data to predict short term electricity load in a certain area.

Four subdomain categories appear in healthcare domain, listed as follows: brain and health monitoring, improving healthcare quality and costs, patient monitoring and interconnection of healthcare platforms. Studies 1, 7, 31 are included within the brain and health monitoring subcategory. Study 1 analyses heart rate, ECG and body temperature. Study 7 analyses brain health and mental disorders. Likewise, study 31 monitors and visualizes epileptic disease-related data. Improvement of healthcare quality and costs subdomain category appears in the studies 18 and 10, involving complex data processing, clinical quality measure analytics and proactive care management analysis. Study 21 is applied on patient monitoring subcategory which covers a system for the analysis of the emergency situations and medical records. In study 26, interconnection of the healthcare platforms is discussed and an overview of the required cloud computing middleware services and standardized interfaces for the integration with medical devices is presented.

The domain scientific platforms involves four subdomain categories as follows: digital libraries for scientific data management, scientific platforms for the cloud, e-learning and learning analytics. Study 37 is listed under digital libraries for scientific data management subdomain as it reports a use-and-reuse-driven big data management infrastructure. Within the scientific platform for the cloud subdomain, study 4 introduces a framework to support rapid design and deployment of scientific applications to cloud. The learning analytics domain appears in study 29, presenting a system to predict the learner's performance, discovering the real learning paths and extracting the learner's behavior patterns. Study 20 is included in the e-learning domain and analyses the influence of big data technologies on the academic platforms.

The sixth domain category is earth observation and aviation, which has two subdomains: earth observation, aviation maintenance and optimization. Studies 1, 11, 30 are within the earth observation domain subcategory analyzing earth data, downloading, processing and viewing satellite images. In [12] a cloud platform is presented with a processing chain model for satellite images with the focus of providing interactive real time services. A real time big data analytical architecture is proposed in [14] for remote sensing satellite application. Besides in [14], a multidimensional big data fusion approach is implemented with a big data architecture and tested with satellite data. Aviation maintenance and optimization domain appears in study 10 and focuses on diagnosing faults in real time, optimizing fuel consumption and predicting maintenance need.

The last target domain category is Financial Services and it is applied into subcategories that are banking with study 25 focusing on cost reduction, scalability and availability of the infrastructures and social customer relationship management with study 27 which presents an architecture consisting of five layers aiming the understanding and implementation of social CRM aspects and dependencies.

Other subdomains which are not listed under any target domain are ambient intelligence (21), recommendation systems (36), anomaly detection (38), trace analyzer (33) and query engine (43). An online and on demand query engine implementing complex event processing to cover a variety of data for querying in real time is discussed in 43

with the target domains e-commerce and energy. Insights about how the backend systems work or for the cloud monitoring systems, traces are analyzed in 33 which can be applicable to any domain. Study 38 targets creating common and reusable services in order to make real time analytics as a service for an anomaly detection system. Modelling lambda architecture as a multi agent heterogenous system, a recommendation system is discussed in 36. Another multi agent architecture is proposed within the direction of internet of things and a case study on ambient intelligence is applied for a smart house in 21.

RQ.1.1: who are the stakeholders?

Answering this research question, we aim to identify the stakeholders that are targeted in different application domains. Various stakeholders are mentioned within in the papers from the following application domains: Industrial Applications, Smart Cities, Social Media, Scientific Platforms. Managers appear frequently as a stakeholder in the studies from the industrial application domain. Whereas for the smart cities domain, depending on the subdomain, the stakeholders significantly differ. In Table 4, a subset of the application domains and the corresponding stakeholders are listed: Table 7

RQ.2: what is the motivation for adopting big data architectures?

Here, we aimed to identify the motivation for adopting big data architectures within the papers examined:

Supporting analytics processes Effective processing and management of massive volumes of data to support data analytics processes is one of the main motivations behind adopting a big data architecture. The input for the big data analytics processes often involves multimedia data, including text, sensor-born data, or music/video streams in order to carry out comparative analysis and identify the emerging patterns and associated relationships in the various domains of application. Big data architectures, infrastructures and tools enable the systems to provide with better decision support.

Improving efficiency Another main motivation for adopting big data architectures is efficiently processing massive volume of heterogenous data with flexible, semi-structured data models and wide range of query sizes while ensuring the fault tolerance of the deployed solution. Monitoring massive information efficiently is also emphasized in the selected primary studies. Execution of join queries on different big data platforms and different big datasets efficiently and interactive querying in timely fashion are also among the goals for adopting big data architectures.

Improving real-time data processing capability The third main reason behind applying big data systems is to gain the ability to deal with the unprecedented speed of real time data generation and the associated needs of processing it. The Internet of Things is a driver for the intensive deployment of sensors, which subsequently generate data streams that are gathered, monitored and processed via big data tools for making event based predictions, querying (complex and ad-hoc) and complex event processing. The big data architecture shall be effectively meeting the latency requirements in such cases.

Table 7 Subset of the application domains and stakeholders for big data systems

Domain	Subdomain	Citation #	Stakeholders	Concerns
Industrial applications	Environmental sustainability	3	Designers, Managers	High performance computation, large scale data storage, complex, voluminous, heterogenous, incomplete dataset
	Production tracking	13	Production Management Shop Managers Users Suppliers Manufacturers	Accurately tracking and determining the order of the production
	Manufacturing	19	Manufacturers	Response time, near real time analysis, accurate forecasts
	Environmental sustainability	41	Enterprise Manager Strategic Support Customer Manufacturer	Accurate and complete data acquisition, data availability, monitor and track in real time, multi source heterogenous data
Smart cities	Smart city experiment testbed	5	Customers Suppliers Manufacturers	Heterogenous data formats, sensors and protocols. High throughput
		9	Platform Designers	Real-time data, diversity of different data sources.
	Surveillance systems	18	Business IT Executives	
	Smart grid	42	Operators, Managers Technical Staff	Scalable user friendly, multiple programming environments
Social media	Travel advising (More than 30 stakeholders...)	8	Customers Reviewers Business Analysts Data Scientists Operation Managers Collaborators Engineering Teams	Value/Cost
Scientific platforms	Digital libraries for scientific data management	37	System Admin Sys. Eng.	Horizontal and vertical scalability, physical bandwidth, no limitation on data, no prescription (multidisciplinary data use and reuse), immediate data access and public access, scaling and flexible schemas

The table is presented respecting the content of the primary studies

Reduce development costs Another main reason is to reduce the costs for system deployment or operation. For example, in the financial sector, market conditions change abruptly, which triggers the urge of processing high volumes of data in short time. Similarly in [13], to improve the user experience, an effective and economical architecture is designed considering time and storage costs. Minimizing costs of both sensors and data storage are at the main focus in [15]. Reducing the development cost of analytical services for citizens and decision makers, efficient use of natural and manmade resources is targeted in [16] and mining big data is used as a valuable source to achieve these targets.

Enable new kind of services Providing new services to support the rapid design and deployment of scientific applications is the primary goal of the scientific platform described in [17]. Service oriented architecture and the semantic web are in the light of this study. The platform adopts software-as-a-service approach and enables the execution, packaging, uploading and configuring of the scientific software applications. In order to support the collection of the data from sensors, in [11] a new kind of big data architecture is defined. This architecture resolves the problems related to data storage, data processing, sensor heterogeneity and high throughput and addresses the data-as-a-service requirement of the system with the support of a reception middleware. As another approach, workflow web services with special analysis processes (speech tagging, named entity recognition etc.) are implemented in [10] to support data scientists to rapidly implement data mining applications.

Data management and system orchestration The last main reason is enabling the system to manage and orchestrate big data sets. In [5], an architecture centric approach is presented to control continuous big data delivery, discussing big data system design and agile analytics development. It focuses on the orchestration of the technologies, prototypes and benchmarks each technology and uses conceptual data modelling method to extend the architecture [4]. presents a system architecture which fosters the system orchestration utilizing REST-based services. The system not only supports data collection, processing and analytics but also enables integration to the other social media analytics systems. The details of the data management concerns for the other studies are listed in Table 8.

RQ.3: what are the existing approaches for software architecture for big data?

Three main approaches are observed for designing the software architectures for big data systems throughout the screening process of the selected primary studies: Adopt a reference architecture, follow an architectural design methodology and use a reference model. The first approach is adopting a reference architecture. In studies 8, 34 and 36, lambda architecture appears as the reference architecture which enables efficient real-time and historical analytics via a robust framework. As another approach, Prometheus methodology which supports the design of multi-agent systems based on plans, goals, behaviours and other aspects, is used in study 25. In study 38, the OAIS reference model is followed to design the software architecture. The OAIS Reference Model provides a conceptual framework for service oriented architectures. Finally, study 28, differentiated replication research methodology is applied to design the reference architecture. Most papers did not

Table 8 Data management concerns for big data software systems

Motivation	Domain, Subdomain	Citation #	Details (explanation, data size, sensor type)
Supporting Analytics Process	Industrial applications, Environmental Sustainability	41	Sensor types: temperature, pressure, velocity I. Decision making for the coordination and optimization (lifecycle management)
	Social media, Web Data Warehousing, Social Data Storage and Analytics	39	A query language at the level of SQL, parallel runtime querying, various query sizes, continuous data ingestion... 100 nodes 5 gb -> 500 gb 100 nodes 12 tb -> 1.2 pb
Improving Efficiency	Health, Brain and Health Monitoring System	1	Improvement of the fusion and the analysis of the big data (>2gb in ~ 70 s)
	Other, Trace Analyzer	33	To optimize the workload and investigate the usage pattern via the analysis of the monitoring data
	Industrial Applications, Electric Power Industry	42	To improve the efficiency and safety of the power systems while keeping the flexibility and availability on demand Sensors: occupancy sensor
	Other, Hive Alternative	39	Performance improvement up to 30 times with higher scalability and availability
	Healthcare, Interconnection of Healthcare Platforms	26	Improve the processing performance, having a resilient cloud storage and indexes for unstructured data and metadata for efficient search
	Industrial Applications, Automotive Industry	17	Facilitated cloud services, cost and performance improvement, scalability on demand ● Data size may exceed 17 Eb per year
	Industrial Applications, Environmental Sustainability	3	Focused on integration considering design optimization and exploration. Models in size of 50GB in size (3D, encoded, compressed, in diverse formats)
Improving Real-time Processing	Industrial Applications, Environmental Sustainability	41	Improvement of the decision-making procedures for coordination and optimization II. Sensors: temperature, pressure, velocity
	Smart Cities, Network Security	35	Optimal responses in real time Latency sensitive applications with fog computing
	Social Media, Social Network Analysis	14	Social data processing in real time.
Reduce development costs	Smart Cities, Traffic State Assessment	40	Real-time traffic situation prediction III. 5gb per day
	Financial Services, Banking	25	Cost reduction keeping the required level of flexibility and scalability
	Social media, Semantic-based Heterogeneous Multimedia Retrieval	12	Heterogenous multimedia data, low cost store and retrieval IV. 10 machines, 10 processors, 20 GB memory, 10 disks and 10 slave data nodes
	Smart cities, surveillance systems	24	Optimizing the number of camera sensors ● Up to 100 gb ● Microwave sensor, boundary/non-boundary camera sensor
	Smart Cities, Smart Grid	34	Efficiency of energy usage and minimization of pollution via managing traffic and the city
	Smart Cities, Smart City Experiment Testbed	9	Testbed with a wireless network topology, reliable data transmission and battery lifetime. Processing both real time and historical data: 50 GB data, 112 sensor nodes, 9 different sensor types.

Table 8 Data management concerns for big data software systems (Continued)

Motivation	Domain, Subdomain	Citation #	Details (explanation, data size, sensor type)
New Kinds of Services	Scientific Platforms, Scientific Platform for the Cloud	4	Scientific applications are designed and deployed to the cloud. Generic infrastructure, web interface, post processing and plotting, monitoring real time V. Use of big machines (An r3.8, 32 vCPUs and 244GB of RAM)
	Smart cities, Smart City Experiment Testbed	5	Architecture for collection of sensor-based data in the context of the IoT. VI. 2gb per year for one sensor (Scenarios with 150–200 GB) VII. 1 measurement per second VIII. Sonar/temperature sensors
	Smart Cities, Smart Grid	34	Enable new kind of services, data accuracy data to assist decision making <ul style="list-style-type: none"> • Social sensors (twitter, blogs etc.) • Smart home sensors (18 sensor measurement per second, citywide 360,000 measurements per second)
	Industrial Applications, Electric Power Industry	6	Data analysis and statistics functions for specific tasks and services such as time series. Performance evaluation with certain evaluation metrics.
	Healthcare, Brain and Health Monitoring System	7	Smart monitoring services. Brain monitoring and models for accurate diagnosis, personalized service modules.
Data Management and Orchestration	Social Media, Public Opinion Monitoring	15	Flexible, knowledge-worker driven iterative exploration, rapid integration
	Social media, Travel Advising	8	Achieving, strategic control, continuous big data value delivery for WBS. <ul style="list-style-type: none"> • Various systems from 90 TB to 1 PB

The table is presented respecting the content of the primary studies

explicitly report on the software architecture approach they adopted. This does not imply that such an approach was not used. It was not reported, as many of these studies were not addressing the software architecture community.

RQ.3.1: What are the adopted architectural models/viewpoints?

The adopted architectural viewpoints (Fig. 10) are the decomposition (presents elements, relations and topology assigning responsibilities to modules), flowchart (displays tasks in a network diagram style) and the deployment (aspects of the system ready to go in live) viewpoints. Eight studies do not include a viewpoint. The decomposition viewpoint is the most applied among the appeared viewpoints. Note that 90% of the architectures that adopt the decomposition viewpoint are layered architectures.

RQ.3.2: What are the adopted architectural tactics/patterns?

There are five architectural patterns reported within the selected primary studies which are listed as follows: Layered (data is forwarded from one level of processing to another in a defined order), cloud based (architectural elements are in cloud), hybrid (combination of different architectural patterns) and multi-agent (a container/component architecture, containers are the environment and components are the agents) (Fig. 11).

In [18], the system architecture proposed for cleaner manufacturing and maintenance is composed of 4 layers that are data layer (storing big data), method layer (data mining and other methods), result layer (results and knowledge sets) and application layer (uses the results from result layer to achieve the business requirements). In [19], the traditional 3 layered architecture of the financial systems was adopted: front office (interaction with external entities, data acquisition and data integrity check), middle office (data processing), back office (aggregation and validation). While at least a 3-layered approach is applied for most of the application domains and two layers with processing and application layer driving the results via aggregation and validation is consistent for all domains, the layers on top are adopted depending on the application domain.

For web-based systems, lambda architecture is implemented in [4] with the batch (non-relational) and streaming layer (real-time data) completely isolated, scalable and fault tolerant. For machine to machine communication, a 4 layer architecture is presented with the service (business rules, tasks, reports), processing (Hadoop, HDFS, MapR), communication (m2m, binary/text I/O) and data collection layers. The layered architecture of the AsterixDB (an open source big data management system) is shown in [2]. Hyracs layer and Algebrics Algebra layer are layers that are represented within the software stack. Hyracs layer accepts and manages the parallel data computations (processing jobs and output partitions). Algebrics layer which is data model neutral and supports high level data languages, aims to process queries. Banian system architecture which is described in [20] consists of 3 layers which are storage, scheduling and execution and application layer and the system provides better scalability and concurrency. The architecture proposed in [15] is for intruder detection in wireless sensor networks. Three layered big data analytics architecture is designed: wireless sensor layer (wireless sensors are deployed), big data layer (responsible for streaming, data processing, analysis and identifying the intruders) and cloud layer (storing and visualizing the analyzed data).

Cloud based architectures are also frequently observed among the selected primary studies. In [2], a scalable and productive workflow based cloud platform for big data analytics is discussed. The architecture is based on the open source cloud platform CloudFlows. Model view controller (MVC) architectural pattern is applied. The main components are data storage, data analytics and prediction and data visualization which are accessible via a web interface. The architecture of [17] uses the cloud environment (Amazon EC2 cloud service) to store the data collected from the sensors and host the middleware. Overall system is composed of sensors, sensor boards, bridge and middleware. Another cloud architecture is used to construct a cloud city traffic state assessment system in [21] with cloud technologies, Hadoop and Spark. Clustering methods such as K-Means, Fuzzy C-Means and DBSCAN are applied to detect the traffic jam. The architecture has 2 high level components which are data storage and data analysis and computation. While data storage is based on Hadoop HDFS and NoSQL, data analysis and computation part utilizes Spark for high speed real time computation. For all of the big data systems applying cloud based architectures, the cloud is used to resolve the scalability problem of the data collection.

In order to provide the users interactive real time processing of the satellite images, a cloud computing platform is introduced for the China Centre for Resources Satellite Data and Application (CCRSDA) in [12]. The platform aims low latency, disk-space

customization and remote sensing image processing native support. The architecture consists of application software including image search, image browsing, fusion and filter, web portal containing private file center, data center, app center, route service and work service, virtual user space management, Moosefs, ICE and Zookeeper and virtual machine management (3 service levels, SaaS, PaaS and IaaS respectively).

One of the primary studies [22] discusses a multi-agent architecture for real time processing. The lambda architecture is modelled as a heterogeneous multi-agent system in this study as 3 layers (batch, serving and speed layer). The communication among the components within the layers is achieved via agents with message passing. The multi agent approach simplifies the integration.

Service oriented architectures are frequently applied for big data systems. In [23] a cloud service architecture is presented. It has three major layers which are an extension of semantic-wiki, rest api and SolrCloud cluster. The architecture explores a search cluster for indexing and querying. Another system architecture described in [5] is based on a variety of rest-based services for the flexible orchestration of the system capabilities. The architecture includes domain modelling and visualization, orchestration and administration services, indexing and data storage.

Other state of the art approaches

Cybersecurity

Software systems are developed and integrated aligning with a software application architecture and deployed when the system is mature enough satisfying the acceptance criteria for the system release and deployment. If the maturity of the system is measurable, the quality metrics are utilized to assess the performance of the system. While a system is performing, the vulnerabilities rooted in the system architecture, deployment configuration or the network architecture enables an external or internal entity to perform malicious activities. Tracing or pre-detecting the vulnerabilities residing within the system could support the decision process for maintenance, risk analysis, implementation or system extension processes. Not only for the system performance but also for the vulnerability analysis which could directly have an impact on the performance itself, system specific metrics could be selected and defined. However, due to the rapid technological developments, system specific and implementation specific codes, artefacts and configurations and maintenance activities, resulting with the right set of metrics is a challenge.

According to [24] “Resilience – i.e., the ability of a system to withstand adverse events while maintaining an acceptable functionality – is therefore a key property for cyberphysical systems”. Primary approaches to measure the resilience could be model based or metric based. As a metric based approach, resilience indexes are defined to be extracted from system data such as logs and process data as a quantitative general-purpose methodology [24].

Resilience readiness level metrics are proposed in [25], as shown in Table 9 and as a matter of fact, the aspects that the big data systems are related to the readiness levels from the cybersecurity point of view are outlined and discussed.

Another study in the survey format is composed in 2018 which is called “Big Data Meets Cyberphysical Systems” [28], that summarizes the impact of the increasing variety of the cyberphysical systems and the amount of sensor data produced. The study also discusses the cyber attacks targeting such systems. Centroid based clustering and

Table 9 Resilience readiness and big data cyber security aspects

Proposal of Resilience Readiness Level Metrics [25]	Big Data and Cybersecurity Aspects
Responsibility	HMI configuration to meet the big data cybersecurity needs. Tailoring of the big data analytics results considering the cyber security concerns. F.e. malicious attacker manipulating the HMI to cause an incorrect action [26], injecting false data or invalid commands. The actors shall have sufficient level of data analytics knowhow to distinguish the false data on UI. The possible commands, command names, action names, updates and event names could be derived via data analytics from the system data and the UI/HMI could be having validity checks. The timing/duration of the action/update in HMI/UI could be compared with the update timing/duration/effort statistics/historical traces/logs of the system that have been performing with real data (duration of querying, UI update etc.).
Mutual Impacts	Threat model [27], common system view, standardization? Security view as an architectural view? Design patterns defined for security? Security as a service?
Situational Intelligence	Security as a lifecycle issue. How to coordinate security practices in requirements, design, code analysis and test stages? Are there common high level security requirement sets for big data systems? Common code fallacies causing vulnerabilities or common test approaches to detect them?
Operation Resources	Attacks targeting the recovery and replication management, that are specific for big data systems. The secure strategy for data replication and its effects on performance. Attacks that cause data replication or unexpected recovery. Attacks against system configuration that injects error to the configuration managed by the operator, validation of the configuration (shall be automated or manual?). The detections of noise injection at data fusion, what are the data analytics methodologies for this? Are there any software libraries that verifies and validates the data analytics process against attacks, f.e. at the time of fusion?
Mutability	The encryption management strategy would be application architecture specific. What are the encryption management strategies applied for the security of the big data system? Are there any communication route or structure adaptation of the system to meet the cybersecurity requirements? What are the attack detection strategies, i.e. checking the mean time between failures? Is the system compatible with the contemporary cybersecurity tools, i.e. ease of modification, integration or monitoring?
Modularity	Is the system compatible with the contemporary cybersecurity tools, i.e. ease of modification, integration or monitoring? Could the component availability be measured as an attack parameter? What are the security criteria considered while applying the data and information refinement?
Event Mechanisms	What are the cybersecurity qualifications considered while adopting a driver driver, kernel function? Which are the secure functions and how to assess the maturity of the function from the cybersecurity aspect? How is the security ensured for the manual modes, i.e. training of the individual or system adaptations such as command verification or peer review?

hierarchical clustering are listed as two groups of clustering methods. K-means is an example of the clustering methods and it has the empty clustering problem. For the hierarchical clustering, the clusters are defined based on similarity measures such as distance matrices and the clustering speed and accuracy is higher comparing to the other algorithms like k-means.

Integration is a concern in cyberphysical systems in critical infrastructures due to the computational challenges observed while applying techniques for data confidentiality and privacy protection [34]. Semi or fully autonomous security management could be adapted according to the needs of the application to be implemented. The solutions could have high cost by means of latency, power consumption or management complexity.

Deep learning

Application of deep learning in big data is discussed in [15], stating the challenges as:

- Estimating the portion/amount of the big data to be used for the deep learning approach
- Overcoming the gap between test and the training data via having generalized learnt patterns.

- Determining the criteria that is representative for the data.
- The interpretation of the complex result.
- Labeled data is required for good performance.
- Open questions are:
 - The way to fuse the conflicting data.
 - The effect of enlarged modalities on system performance.
 - The architectural approaches for feature fusion and heterogenous data.
- Data with high velocity, how to approach the variety of the data distribution with respect to time.

In [19] emotion recognition is achieved via the fusion of the outputs of convolutional neural networks (CNN) and extreme learning machines (ELM) and for final classification SVM is used. The architectural approach could be characterized as hybrid application architecture having CNN and SVM with ELM fusion. Achieving high accuracy with this approach, it is observed that data augmentation improved the accuracy further.

Sentiment analysis

In [29] is analysed based on topics of sensitive information. In order to accomplish the analysis, bidirectional recurrent neural network (BiRNN) and LSTM are combined to form BiLSTM to ensure having context information continuously. The architecture has a layered structure.

The brand authenticity analysis is carried out in [30]. The quality commitments for the tweets are instantly sharing sentiments, sharing complaints, processing complaints and the quality of ingredients. Statistica 13 software is used for SVM analysis.

RQ.4: What is the strength of evidence of the study?

In order to state the plausibility of the results, within this research question we will be discussing to which extent the audience of this study can rely upon the outcomes. Among the various definitions to address the strength of evidence, for this SLR we selected the Grading of Recommendation Assessment, Development and Evaluation (GRADE). As it can be observed from the Table 10 (adopted from [17]), there four grades which are high, moderate, low and very low to assess the strength of evidence which takes into consideration the quality, consistency, design and directness of the study. Comparing to the observational studies, experimental studies are graded higher by the GRADE system. Among the primary studies in this review, 16 (37%) are experimental. The average quality score of these studies is 6,4 which means that our studies have a moderate strength of evidence based on the design (Table 11).

Table 10 Definitions used for grading the strength of evidence

Grade	Definition
High	Further research is very unlikely to change our confidence in the estimate of effect
Moderate	Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate
Low	Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate
Very low	Any estimate of effect is very uncertain.

Table 11 Average quality scores of experimental studies

Experimental studies	2, 5, 11, 12, 14, 15, 21, 22, 24, 31, 32, 33, 35, 37, 39, 40
Number of studies	16
Mean quality score	6406
Standard deviation of quality score	1.09

Most of the primary studies we analyzed do not include explicitly a quality assessment by means of our quality criteria which are rigor, credibility and relevance. Therefore there is a risk of bias implied by the low quality scores.

In terms of quality, from the rigor perspective, we can observe a variety of presentation structure and reporting methods which complicates the comparison of the content. For most of them the aim, scope and context are clearly defined, however for some of them the results are not clearly validated by an empirical study or the outcomes are not quantitatively presented. Besides, throughout many studies, research process is documented explicitly but some of the research questions remained unanswered. Considering credibility, while the studies tend to discuss credibility, validity and reliability, they generally avoid discussing negative findings. The conclusions are quite relating to the purpose of the study, and the results are relevant while not always practical.

Considering the fact that the presentation of the research questions and the results extremely varying from study to study, it is very complicated to analyze the consistency of the outputs of the primary studies. As a result, sensitivity analysis and synthesis of the quantitative results were not feasible.

With respect to directness, the total evidence is moderate. According to Atkins et al., (2004) a directness is the extent to which the people, interventions, and outcome measures are similar to those of interests. The people were experts from academy or industry which are within the area of interest. The outcomes are not restricted for this literature survey. A considerable amount of primary studies answers the research questions and validates the outcome quantitatively.

Assessing all elements of the strength of evidence, the overall grade of the impact of the big data system architectures presented throughout the primary studies is moderate.

Discussion

Various dimensions of improvements are analyzed, implemented and experimented throughout the studies listed in previous chapters. In AsterixDB, the performance of the functions within dataflows are planned to be improved via introducing further parallelism for function evaluation in pipeline.

The performance is measured in the presented big data architectures for construction waste analytics by means of accuracy. Application areas within the internet of things field the challenge is scaling the platform, privacy and security both for the data and the system. Assessing the maturity of a big data system and which metrics to measure the maturity is another area to be explored.

Hybrid architectures are not often observed among the primary studies, however could be benefited more as in the application architecture having CNN and SVN for the sentiment analysis. Integration concerns for big data systems within the cyberphysical domain is to be further investigated (Tables 12, 13, 14 and 15).

Threats to validity

In order to have a valid systematic literature survey, we should make sure that the research protocol and the constructed research questions ensure elimination of the publication bias, in which the positive results tend to be presented by the researchers. In our study the search is conducted automatically, therefore the search string is also aligned with the target. The keyword list within the search string can result with incompleteness, which can be resolved using an iterative approach to construct the keyword list. Combining the search results in digital libraries with the results from the search engines we achieve a better coverage of the search results. Keyword list is incrementally expanded or shrunk to cover target studies. Considering the irrelevant searches introduced by the inefficient search algorithms of the digital libraries, manual selection criteria are defined.

Modelling the data extraction method is critical to derive the correct results from the selected primary studies to overcome the data extraction bias. The selected papers are screened considering the previously defined research questions to form the initial data extraction model. We iterate through the selected papers by adding and removing the fields to the data extraction model until we eliminate irrelevant results and have enough coverage within the final data extraction model [32].

Conclusion validity

The systematic literature review methodology has a significant control on the primary studies that are derived as an output of the screening process. The data extraction columns are peer reviewed in order to have a common and generic objective for the study. The results are based on a solid mapping to the selected primary studies which is traceable via the data extraction table.

Internal validity

Exclusion criteria that is applied for the selection of the studies has the highest effect on the result. Exclusion criteria is also peer reviewed to ensure the precision and recall.

Construct validity

The goal of the study is analyzing different aspects of the primary studies such as architectural methods/viewpoints, stakeholders, key concerns etc. focusing on the application domain. The outcomes of the analyzed aspects are presented in Section 4 Results.

External validity

The results are applicable for the application domains listed in Section 4. In case of having overlapping key concerns or quality goals, the results can be implemented in other application domains.

Related work

Hu, Han, et al. (2014) in "Toward scalable systems for big data analytics: A technology tutorial." conducted a literature survey and provided a tutorial on big data analytics platforms to introduce the overall picture about the big data solutions [10]. They introduced a big data technology map which visualizes the exemplary technologies over the past 10

Table 12 Benefits and limitations of the primary studies^a

Study Citation #	Benefit	Limitation
1	Complex magnitudes can be altered into smaller data subsets using 5 level fusion model	Not compatible with social media applications
2	Efficiency for smaller as well as large queries. Scalable new runtime engine, all-LSM-based data storage, with B+ tree, R tree, and keyword and n-gram indexes; rich set of primitive types, including spatial, temporal, textual types to handle Web and social media data; support for fuzzy selections and joins; a built in notion of data feeds for continuous ingestion.	Continuous queries are not supported.
3	Minimization of the construction waste. The intended tool will equip designers with well-informed and data driven insights to optimize design for designing out waste	This paper limits discussions to horizontal scaling Big Data platforms, particularly, Apache TM Hadoop and Berkeley Big Data Analytics Stack (BDAS). This selection is mainly influenced by the data and computational requirements of construction waste analytics, which include iterative algorithms, compute-intensive tasks, and near real-time visualisation.
4	Manage the cloud infrastructure, including an interface to create modified input descriptions, job scheduling, plotting of output data, and file management	No support for more complex plotting capabilities, such as contour plots, no workflow management system, no command-line installer
5	Sensors to data management, and supports a user who wants to set up a research or production infrastructure to collect very large datasets in the context of the IoT	Project is still at its beginning. As a consequence, the work done in this architecture focused on data collection and storage
6	A scalable and productive platform to facilitate data scientists' work. The Docker light-weighted visualization container, to support multiple programming environments embedded within the workflow interface. Spark infrastructure, scalable to big data set, which is transparent to the end user. Web interface provides a user- friendly data analytics environment with access- anywhere, – anytime, and any devices feature. Cloud platform is also be able to scale up and down based on the requirements of user's data analytics work.	The workflow interface is to be enhanced to make it more open to data scientists, who will be able to revise and add widgets more conveniently
7	Web intelligence (WI) may be viewed as an enhancement or extension of artificial intelligence (AI) and IT on the Web. A prototype of a portable brain and mental health-monitoring system (brain-monitoring system, for short) to support the monitoring of brain and mental disorders.	The technological architecture of security and privacy protection should be fit for different application environments, including the Internet, IoT, and MI.
8	Wisdom Web of Things (W2T), where the "wisdom" means that each of the "things" in IoT and WoT is aware of both itself and others to provide the right service for the right object at the right time and context.	A design method or development methodology, no matter how thorough, can never guarantee success. The application of an architecture-centric methodology like AABA requires discipline and creativity, which may be a tall order for organizations that do not have the required discipline and innovation mindset.
9	A valuable example to future Smart City platform designers so that they can foresee some practice issues and refer to this solution when building their own smart city data platforms.	It is not possible for us to identify the concrete reasons why those sensor become abnormal. But this observation indicates for a smart city with a large number of deployed sensors, detecting anomalies in collected data must be seriously considered. That is why we implemented some anomaly detection algorithms as external processing tasks
10	Systematic design using tactics	The tactics could be characterized. F.e. tactics that have dependency on each other or complex

Table 12 Benefits and limitations of the primary studies^a (Continued)

Study Citation #	Benefit	Limitation
		tactics etc.
11	processing chain model is proposed for satellite images on a private cloud computing platform	Currently our fault tolerance mechanism depends fully on the structure of ZooKeeper; all nodes are identical and there is no centralized control. When the route service fails, the work services in ZooKeeper will automatically recommend an alternative as the route service.
12	Architecture named SHMR (Semantic-based Heterogeneous Multimedia Retrieval) to support heterogeneous multimedia big data retrieval. Solves type heterogeneity.	Experimental dataset acquisition is from some specific websites such as Flickr, Wikipedia and Youtube, the semantic provision by social users is still a simulation. The experiments will be in real Internet environment and the retrieval speed will be increased.
13	Cloud and RFID technologies are integrated for remote and real-time production data capture and tracking while intelligent techniques are used to generate effective production scheduling solutions	Better supply chain coordination and better production scheduling decisions can be achieved
14	Generating meaningful information from text-based social data	Improve the efficiency of multi-processing. The dynamic process controller will work as a load balancer in our system to mitigate the gaps depending on the system resources such as usages of memory and CPU. We expect that the controller dynamically will control the number of processes according to their hardware resources
15	The system provides tools to help with constructing “domain models” (i.e., families of keywords and extractors to enable focus on tweets and other social media documents relevant to a project), to rapidly extract and segment the relevant social media and its authors, to apply further analytics (such as finding trends and anomalous terms), and visualizing the results	Optimizations are underway, including a shift to SPARK for management and pre-processing of the background corpora that support the rapid domain scoping. Tools to enable comparisons between term generation strategies and other scoping tools are under development. A framework to enable “crowd-sourced” evaluation and feedback about the accuracy of extractors is planned. The team is working to support multiple kinds of documents (e.g., forums, customer reviews, and marketing content), for both background and foreground analytics. The team is also developing a persistent catalog for managing sets of topics and extractors; this will be structured using a family of industry-specific ontologies.
16	Web Observatories with rich, timely resources for observation and analysis. Individually, these feeds provide a resource to measure the current state - or health - of a social machines, and combined, they have the potential to provide a collective pulse of the Web	A wider analysis of the current and proposed metrics for measuring social machine activity, and how they contribute to understanding different classes of social machines is required.
17	Provide scientific workflows to help remove technical burdens from researchers, allowing them to focus on solving their domain-specific problems.	workflow scheduling techniques Are to be explored
18	N/a	N/a
19	Proposes and examines the concept of event-based process predictions and outlines its potentials for planning, forecasting and eventually controlling business processes	Current techniques and systems available data cannot be analyzed in a reasonable time frame to make sufficient business value out of it
20	Help education in the near future, by changing the way we approach the e-learning process, by encouraging the interaction between students and teachers, by allowing the fulfilment of the individual requirements and goals of learners.	Further performance analysis to be done against high workload

Table 12 Benefits and limitations of the primary studies^a (Continued)

Study Citation #	Benefit	Limitation
21	The use of multi-agent systems in software development has two major benefits given by the re-usability and composability of the agents and by the higher level of abstraction introduced by the agent oriented programming paradigm	Backed by the Prometheus Design Tool (PDT) ⁴ an Eclipse plug-in,
22	Enable the integration of disparate urban sensing systems, including individually owned data through participatory sensing.	Framework to include more features, such as opportunistic task assignment by dynamically finding out the most suitable group of sensing participants to gather information about a specific issue, sensor stream quality validation and improved privacy and security.
23	Provides a better understanding of how fundamental assumptions in Hadoop's design make it a poor fit for real-time applications	–
24	To analyze the data generated during the construction of the barrier and detection of the intruder using camera sensors	An intruder is detected, if it intersects with the sensor's path along with the sensing range of the sensor. However, it could be possible that an intruder is undetected, if it is not within the sensing range of a sensor and also even if the intruder is detected, the sensor cannot communicate instantly with other sensors to pass the information
25	Architecture system design, based in open distributed computing paradigms like Hadoop map-reduce, offering horizontal scalability and no-SQL flexibility while at the same time meeting the stringent quality and resilience requirements of the banking software standards. Benefits: 1) the orchestration double layer architecture allows for an effective decoupling of the external from the internal processing workflows. Changes in the workflows due to external business requirements could be easily implemented without affecting the data processing structure. 2) The segmentation of map-reduce jobs in the triad barrier/map-reduce job/barrier together with the orchestration database provided an effective mechanism to orchestrate non-trivial data processing logic. 3) The orchestration database containing data processing status at configurable granularity level (both on data entities or processing steps) provides a reliable tool for the implementation of error monitoring, backup and disaster recovery procedures. 4) Following this pattern, the introduction of new processing steps like new XSLT transformations on already defined data requires minimal implementation effort, obtaining an already parallelized process.	Among the observed pitfalls the introduction of an external orchestration engine with advanced capabilities and a reliable database increases the cost, both in terms of platform infrastructure and development. At the present time, Oozie workflows are represented as simple directed acyclical graphs, which impose its limitations on the workflow data processing complexity that can be implemented.
26	An overview of cloud middleware services for interconnection of healthcare platforms	–
27	Separates the concerns of social CRM using architectural perspectives and aims at building a better understanding. The research method is a literature review in which artefacts are gathered and assigned to five layers, which are business, process, integration, software, and technology. The conclusion states that social CRM is an emergent research field and comprises a call for more artefacts that concretise abstracted components of the business-layer.	–
28	Technology independent reference architecture for big data systems, which is based on analysis	A limitation of the proposed classification is concentration on selected technologies in the

Table 12 Benefits and limitations of the primary studies^a (Continued)

Study Citation #	Benefit	Limitation
	of published implementation architectures of big data use cases. An additional contribution is classification of related implementation technologies and products/services, which is based on analysis of the published use cases and survey of related work.	survey. However, other authors have covered other technological topics in earlier surveys: batch processing, machine learning, data mining, storage systems, statistical tools, and document-oriented databases. Another limitation of this work is that the reference architecture should be evaluated with a real big data use case, which would complete step 6 of the research method.
29	A big data software architecture that uses an ontology, based on the Experience API specification, to semantically represent the data streams generated by the learners when they undertake the learning activities of a course, e.g., in a course. Th	To be improved with an Enterprise Service Bus able to integrate different data stream sources and a big data-oriented message queue to increase the activity stream performance.
30	Real-time Big Data analytical architecture for remote sensing satellite application, capability of dividing, load balancing, and parallel processing of only useful data	Not compatible for Big Data analysis for all applications, e.g., sensors and social networking.
31	Mobile-based monitoring and visualization architecture for life-long diseases	To be expanded to detect and analyze other life-long disorders, such as Alzheimer and Parkinson's disease
32	A mode of using double cloud architecture and optimizes clustering algorithm to monitor the massive network information in real time.	–
33	Cloud service architecture that explores a search cluster for data indexing and query	More analysis methods are required for the architecture extension to make the architecture generic. The architecture is to be expanded with support of running MapReduce-based analysis methods.
34	Contributions of this chapter are threefold: (1) we provide an overview of Big Data and Internet of Things technologies including a summary of their relationships, (2) we present a case study in the smart grid domain that illustrates the high level requirements towards such an analytical Big Data framework, and (3) we present an initial version of such a framework mainly addressing the volume and velocity challenge.	Extend the analytical framework with the necessary mechanisms to achieve such uniform processing
35	To support the integration of massive number of infrastructure components and services in future smart cities.	-To secure future communities, it is necessary to build large-scale, geospatial sensing networks, perform big data analysis, identify anomalous and hazardous events, and offer optimal responses in real-time.
36	Processing of Big Data in real-time based on multi-agent system paradigms.	in the presented approach it is strongly recommended to use the same event representation in a both processing: batch and online. it is argued the approach is general purpose.
37	Use and reuse driven big data management approach that fuses the data repository and data processing capabilities in a co-located, public cloud.	Although much still needs to be done to fully realize the vision of use and reuse driven data management, the evaluations presented in section 7 have clearly demonstrated the technical feasibility to manage big data in the cloud.
38	A real time data-analytics-as-service architecture that uses RESTful web services to wrap and integrate data services, dynamic model training services (supported by big data processing framework), prediction services and the product that uses the models.	The machine learning algorithms supported are limited to the applied machine learning library and big data frameworks.

Table 12 Benefits and limitations of the primary studies^a (Continued)

Study Citation #	Benefit	Limitation
39	An efficient system for managing and analyzing PB level structured data called Banian	To achieve higher processing performance and scalability, Banian does not support the partial update and deletion of table data, and its support for transaction consistency is not very strong. Therefore, it is not yet a full-fledged replacement for parallel database. In the future, the above-mentioned weaknesses of Banian will be addressed with further research efforts.
40	Real-time bus location and real-time traffic situation, especially the real-time traffic situation nearby, through open data, GPS, GPRS and cloud technologies. With the high-scalability cloud technologies, Hadoop and Spark, the proposed system architecture is first implemented successfully and efficiently.	In the future, we expect to apply this system to all roads in Taichung and to improve the accuracy of estimates.
41	To make better Product Lifecycle Management and Cleaner Production decisions based on these data collected from smart sensing devices	Without proper data preparation and accurate model, data mining is apt to generate useless information.
42	Service-oriented operation model of China's power system, which integrates the concepts and techniques of cloud computing, big data analytics, internet of things (IoT), high performance computing, smart grid and other advanced information and communication technologies (ICTs).	There are huge spaces for the future development of CG in China. We should also note that CG is a complex system engineering. The implementation of CG need more mature techniques, as well as the collaborative supports of government, industry and academia. Smart grid is in its infancy, and the implementation of CG is also a gradual process. The application of CG needs more comprehensive theoretical support and more widely experimental demonstrations.
43	A stateful complex event processing framework, to support hybrid online and on-demand queries over realtime data.	The scalability of the in-memory state persistence model. When the stateful CEP system processes a large number of long running online queries, the system memory may be drained out.

^aThe table is presented respecting the content of the primary studies

years, relating it with the data value chain that consists of the data generation, data acquisition, data storage and analytics. Alternatively, they present the layered architecture for the big data systems which is decomposed into 3 layers which are infrastructure layer, computing layer and application layer. The study discusses the big data system challenges, data sources, frameworks and their applications. Compared to our study, Hu, Han, et al. (2014) reviewed the technologies mapping to the big data system architectures, while we focused instead on the architectural perspectives and big data application domains.

In the literature review of Tan, et al. (2015), the focus is on the big data architectures for pervasive healthcare systems that aims to deliver healthcare services to patients anywhere and anytime, including data collection via mobile devices and sensor network [1]. Besides it discusses the relationship between the research directions and the compiled big data architecture. While our study screens various application domains, the literature review of Tan, et al. (2015) discusses the big data architectures based on a single application domain. Aligned with our study, the data interoperability, security and privacy are among the key concerns for the big data system architectures in healthcare domain.

In “How ‘big data’ can make big impact: Findings from a systematic review and a longitudinal case study”, Jamba et al. present a framework which analyses the big data

Table 13 Difference and importance of the primary studies^a

Study Citation #	Difference	Importance
1	Data fusion model with 5 processing levels	Based on partitioning and aggregation technique for big data. Focuses on improving the computational efficiency.
2	Native data storage and indexing as well as querying of datasets in HDFS or local files. Open data model that handles complex nested data as well as flat data and use cases ranging from “schema first” to “schema never”	Open source. Full function BDMS that is best characterized as a cross between a big data analytics platform, a parallel RDBMS, and a NoSQL store, yet different from each.
3	Not applicable	The first big data based architecture for construction waste analytics.
4	–	There are still no available frameworks or middleware solutions that are dedicated to supporting scientific applications in such a way that (1) users can easily upload their program to the cloud, (2) have a user-friendly interface automatically generated for them to run
5	Aims to equip an academic campus with sensors and supports the definition of innovating application exploiting these data	It triggers interesting challenges about scalability of a community-driven usage of such an open data platform, the evolution capabilities of the <i>Data as a Service</i> API, as well as privacy and security issues.
6	Integrate a web-based workflow interface with Spark to support big data analytics; 2) utilize Docker to create a light-weight virtualization environment to support a variety of program development environments, and facilitate user program/widget management; 3) demonstrate the workflow-based data analytics platform capabilities with real-world electric power industry data	Having a productive data analytics cloud platform by integrating a variety of data analytics tools and packages with a high-level workflow interface
7	Cloud computing primarily focuses on the system resource architecture of IT applications—that is, infrastructures, platforms, and software (developing and scheduling abilities). For the large scale converging of intelligent IT applications, it’s necessary to develop an open and interoperable intelligence service architecture for the contents of IT applications—the data, information, knowledge, and wisdom (DIKW).	Supports the challenges of: Investigation of human BI via research on holistic intelligence? Collecting, managing, and mining BI Big Data to gain a systematic investigation and understanding of human intelligence?
8	Agile big data analytics for web-based systems: An architecture-centric approach. Agile big data analytics for web-based systems: An architecture-centric approach.	The first of its kind, AABA fills a methodological void by adopting an architecture-centric approach, advancing and integrating software architecture analysis and design, big data modeling and agile practices.
9	A citywide testbed, with regard to wireless network topology, reliable data transmission, battery lifetime and programmability of deployed sensor nodes	There is still a gap between what a big data platform for smart cities looks like at the high level and how it should be properly realized. To fill this gap, this paper presents a concrete and valuable example by introducing our city data and analytics platform named <i>CIDAP</i> .
10	Existing catalogs do not contain tactics specific to big data systems	Expanding the collection of architecture tactics presented in this paper and encoding these in an environment that supports navigation between quality attributes and tactics, making crosscutting concerns for design choices explicit
11	There are three major differences. First, the types of images are much more diverse in our environment, including classify-image, pan image, DEM, etc. Second, the number of bands is possibly more than three. For instance, the	A new computing model, the Remote Sensing On-Demand Computing (RS-Demand) model that overcomes these challenges. The key idea behind RS-Demand is to treat remote sensing image processing as chain computing procedures in

Table 13 Difference and importance of the primary studies^a (Continued)

Study Citation #	Difference	Importance
	multispectral data in ZY-3 satellite has four bands. Third, many coordinate systems coexist in our system. For example, the ZY-3 satellite has its own RPC parameters.	memory. Image tiles go through the algorithm node and reach the end-user screen on-the-fly. no software installation, transparent processing and storage, and low bandwidth requirements – are critical in emergency applications.
12	Excellent economic efficiency.	Excellent economic efficiency.
13	No research has examined the production scheduling problems in a distributed manufacturing company from a holistic perspective	Make-to-order labor-intensive manufacturing to improve information visibility and transparency
14	Firstly, the data storages of the previous approaches are based on a relational database that may cause performance issues when a huge amount of datasets ranging from a few terabytes to multiple petabytes needs to be handled. Secondly, they do not support distributed processing, which may slow down processing time. Lastly, they collect data from only a single source channel, such as Twitter.	Own sentiment analysis model in previous research, which guarantees higher accuracy. Previous approaches mainly used a relational database as a main data storage.
15	The key novelties in the system are: (a) enabling iterative rapid domain scoping that takes advantage of several advanced text analytics tools, and (b) the development of a data-centric approach to support the overall lifecycle of flexible, iterative analytics exploration in the social media domain.	Alexandria advances the state of the art of social media analytics in two fundamental ways (see also Section VIII). First, the system brings together several text analytics tools to provide a broad-based environment to rapidly create domain models. This contrasts with research that has focused on perfecting such tools in isolation. Second, Alexandria applies data-centric and other design principles to provide a working platform that supports ad hoc, iterative, and collaborative exploration of social media data
16	Working on dynamic methods of integrating streams. One approach involves monitoring the overall message rate of a given set of streams (i.e. posts per minute), and using fluctuations in stream volumes as an early indicator for combining streams undergoing similar changes	From a social machines researchers perspective, the ability to access unified, and in certain circumstances, integrated real-time streams of activity is an essential resource to understand, analyse, and possibly make predictions about the current state of a social machines health
17	For Big Data Workflows in the Cloud more generic, implementation-independent solution	For Big Data Workflows in the Cloud more generic, implementation-independent solution
18	–	–
19	Data base is potentially available for event-based predictions of its manufacturing processes	Outlined the potentials of event-based predictions in order to plan and eventually control business processes. Besides outlining these potentials, a general concept for event-based predictions has been conceived and the current state of the art was discussed
20	A three-step system architecture for a consortium of universities	Efficient - The entire solution described is efficient, because activities are separated on levels and resources, the traffic is managed by Hadoop in Clouds, and the analysis is able to add graphic representations to other types of results.
21	The Prometheus has proven to be a practical agent oriented methodology	The Prometheus has proven to be a practical agent oriented methodology
22	Existing research literature lacks a generic data collection and dissemination system. Existing approaches are application-specific, hindering their scalability and reuse.	Gathering real-time information produced by such disparate existing systems can improve the management of city resources
23	–	Illustrating the challenges of real-time data processing
24	Inadequate research is not only on the quality	No barrier construction protocol proposed so far

Table 13 Difference and importance of the primary studies^a (Continued)

Study Citation #	Difference	Importance
	measurement of the image in terms of width and resolution but also on the limited mobility on camera sensors. To get rid of this problem, we propose an energy efficient barrier construction algorithm where all camera sensors are having limited mobility. Also we provide a better solution for intruder detection with the help of this barrier line.	considers all the three functionalities such as node mobility, rotation of the camera sensors and Quality of Measurement of WSN to detect the intruder efficiently. Moreover, camera sensors are normally expensive and efficient detection of an intruder with a minimum number of camera sensors is a challenging research issue.
25	A new distributed computing paradigm based in highly scalable and fault tolerant map-reduce model, running on commodity class servers, a new opportunity h	An architectural and design pattern for the adoption of these new technologies in the solution of massive data processing and analytics tasks of investment and financial institutions, adapted to the strict requirements imposed by the banking technological model: rich and complex workflows, massive volumes, enormous variety of data structures that must be combined together and stringent requirements of reliability, consistency (every single record counts), data back-up and persistency
26	–	Fills a gap in the electronic healthcare register literature by providing an overview of cloud computing middleware services and standardized interfaces for the integration with medical devices.
27	–	–
28	Conceptual work integrating the approaches into one coherent reference architecture has been limited, others but they did not focus specifically on architectural issues or explicit classification of technologies and commercial products/services.	concentrated on reference architecture
29	No proposal that (i) collects in real time the large volume of information generated during a course; (ii) represents and stores this information following a standard specification to facilitate its interoperability with learning analytics services; (iii) enables these services to effectively access to the information generated in the learner's activities; and (iv) offers a set of intelligent learning analytics services that provide new and valuable information to teachers in order to take better decisions for improving the quality of the learning and teaching processes.	Teachers could be aware of what learners are doing, making difficult to improve or correct any deficiencies
30	Majority of work have been done in the various fields of remote sensory satellite image data, such as change detection [6], gradient-based edge detection [7], region similarity- based edge detection [5], and intensity gradient technique for efficient intraprediction [31]	Efficiently analyzing real-time remote sensing Big Data using earth observatory system
31	Data intensive, availability	Introduce a mobile system to monitor Code patients, while receiving professional Design healthcare
32	A faster operating speed, strong reliability and faster convergence rate, especially with the increase of the amount of data, the advantages of the speed of convergence is more obvious.	Uses double cloud architecture to make full use of cloud resources and network bandwidth.
33	Existing tools are mostly provided as part of IaaS or PaaS cloud services. These monitoring systems are provisioned by the cloud service providers.	

Table 13 Difference and importance of the primary studies^a (Continued)

Study Citation #	Difference	Importance
	They often have limitations in adding analysis methods beyond simple aggregations and threshold-based settings. In this paper, we present a cloud architecture leveraging SolrCloud, the open source search-based cluster that supports large monitoring data storage, query, and processing. This architecture is integrated with Semantic MediaWiki that allows documenting, structuring, and sharing the source of cloud monitoring data as well as any analysis results.	
34	The terms are increasingly used interchangeably and the corresponding solutions follow similar principles.	Addressing the lack of an analytical framework that pulls all these components together such that services for urban decision makers can easily be developed.
35	In contrast to the Cloud, the Fog not only performs latency-sensitive applications at the edge of network, but also performs latency-tolerant tasks efficiently at powerful computing nodes at the intermediate of network.	To secure future communities, it is necessary to build large-scale, geospatial sensing networks, perform big data analysis, identify anomalous and hazardous events, and offer optimal responses in real-time.
36	–	Shown how autonomous agents can enhance the architecture and provide capabilities for robust processing of data in real-time.
37	Although seismology data repositories exist, they usually mandate data access methods, processing tools, and have very limited search options that mostly gear towards seismology researches. In contrast, we purposefully avoid prescribing and limiting what the data shall be used for and how they are used.	It answers to the urgent data management needs from the growing number of researchers who don't fit in the big science/small science dichotomy.
38	Other big data batch processing frameworks and their machine-learning layer are usually not designed to be easily invocable as well-designed model training services by different users. It is also out of the scope of these frameworks to enable the resulting learned models to be used by different external products, not to mention dealing with real time requirements of these products.	Main contribution made in this paper is a real time data analytics service architecture design where it allows a machine learning model to be continually updated by real time data and it wraps big data processing framework as reusable services.
39	Banian overcomes the storage structure limitation of relational database and effectively integrates interactive query with large-scale storage management.	By combining HDFS with the splitting and scheduling model, Banian effectively integrates large-scale storage management with interactive query and analysis.
40	Our proposed architecture will combine Spark with the distribution computation of Hadoop YARN to enhance performance.	Because the real-time data collected is huge and from different attributes, this thesis utilizes a novel cloud architecture of big data to store, process, and analyze a huge amount of real-time data and thus provides useful information.
41	The researches above mainly focus on how to apply the IoT related techniques on one stage of PLM (such as manufacturing process of BOL), and the overall solution for the whole lifecycle is seldom investigated. There is lack of systematic solution of automatic identification and capturing for lifecycle data	A novel concept of integrating big data analytics with product service
42	Other research works have paid less attention to the overall penetration of cloud computing service model to the whole process of power system operation.	A new operational model of power system. It can support the cost-efficient and environmentally friendly operation and development of China's power industry
43	Traditional CEP systems do not consider data <i>variety</i> and only support online queries.	A semantically enriched event and query model to address data <i>variety</i> .

^aThe table is presented respecting the content of the primary studies

Table 14 Benefits and limitations of the Other State of Art Studies^a focusing on the application domain

Citation #	Benefit	Limitation
19	A <i>model-free, quantitative, and general-purpose</i> evaluation methodology to extract <i>resilience indexes</i> from, e.g., system logs and process data	Furthering the investigation into the combination of several FOM functions or resilience indexes in systems with several observed variables and more complex hierarchical structures
20	Review the main enabling technologies included under the concept of Industry 4.0, identifying the local security threats against those areas and their most representative attack vectors	–
23	An overview of the different security solutions proposed for Cyber Physical Systems (CPS) big data storage, access, and analytics. We also discuss big data meeting green challenges in the contexts of CPS.	–
24	A comprehensive survey on what is Big Data, comparing methods, its research problems, and trends. Then a survey of Deep Learning, its methods, comparison of frameworks, and algorithms is presented. And at last, application of Deep Learning in Big Data, its challenges, open research problems and future trends are presented	–
25	An audio-visual emotion recognition system using a deep network to extract features and another deep network to fuse the features.	Using other deep architectures to improve the performance of the system
26	The sensitive information topics-based sentiment analysis method for big data is proposed. This method integrates topic semantic information into text representation through a neural network model.	The extension of the sentiment dictionary and the emoticons will be considered to improve accuracy of the sentiment dictionary tagging method, thereby losing less texts.
27	examine the sentiments toward a brand, via brand authenticity, to identify the reasons for positive or negative sentiments on social media	could use retrospective data to access sort of data (number of likes, retweets) to see if there are sentiment differences between popular tweets and others.

^aThe table is presented respecting the content of the primary studies

from the definition perspective [4]. Besides, presenting a general taxonomy, the paper also enables the reader to understand the big data systems and how business value is derived out of it. The study is a comprehensive literature review which does not discuss the big data system architectures in depth, rather focuses on the business and practical aspects of the big data systems.

“A general perspective of Big Data: applications, tools, challenges and trends” is another study presenting the main trends, technical domains and tools for big data systems and summarizes the state of art in big data [12]. The study screened 457 papers and classifies them into 6 categories which are capture, store, search, share, analysis and visualization and mentions analysis as the most important category. Besides the widely applied frameworks are listed as Hadoop, Mahout, Storm, Spark, S4, Drill, MapReduce, Dryad, MOA, SpagoBI and D3.js.

The study “Big Data and virtualization for manufacturing cyber-physical systems: A survey of current status and future outlook” was authored by Babiceanu et al. Manufacturing cyber physical systems are monitored by means of simulation and data processing simultaneously with the actual physical world operations. The study reviews the

Table 15 Difference and Importance of the Other State of Art Studies^a focusing on the application domain

Citation #	Difference	Importance
19	Does not require a mathematical model of system dynamics, but only knowledge of (un) desired values for process variables.	Improved on the current state of the art in resilience evaluation by providing experimental data showing that it is possible to summarize the resilience of a system through numerical indexes that ensure model freedom and generality.
20	–	–
23	–	–
24	–	–
25	The existing systems were not evaluated in Big Data	The proposed system outperformed other similar systems
26	In traditional machine learning approaches, the features and goals are independent. The traditional RNN model has defects such as vanishing gradient problem (gradient approximates zero) and exploding gradient problem (gradient is very high). The vanishing gradient will make the learning process difficult to converge, and exploding gradient will lead to the instability of the learning process. Due to the shortcomings of the traditional RNN model, Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) were proposed.	It is important for public opinion supervision to identify sensitive information topics through topic models and conduct sentiment analysis based on sensitive information topics.
27	Quality commitments are used as the first dimension of brand authenticity. In addition to the four dimensions of brand authenticity –quality commitment, heritage, uniqueness, and symbolism—we added an alternative option for our coders to choose from. When they found that a tweet is related to brand authenticity but the four dimensions are not able to capture it, they chose the none of them category.	Fill the current gaps in the literature, and contribute both in terms of better precision, and of providing firms with valuable insights about the way people interact with their brands.

^aThe table is presented respecting the content of the primary studies

application of the big data analytics, virtualization and cloud based services for planning and control of the manufacturing operations [33].

Previous work did not review big data software architectures with a methodological approach such as systematic literature review as we provided within this study. The data extraction methodology, elimination criteria and the research questions on the domain analysis and architectural aspects such as patterns, viewpoints, quality attribute are uniquely represented for the big data software architecture field within our study.

Conclusions

In this paper, we have presented the results of a systematic literature survey on software architectures for big data systems. We screened the application domains that the big data software architectures were applied, and as a target, identified the current domains, architectural concerns, aspects and future research and application areas. There has been no previous systematic literature survey study performed yet on big data system architectures for this purpose in the known literature. The systematic literature survey is carried out covering the published literature since 2013. Starting with a corpus of 622 papers from the searching literature, we narrowed them down to 43 primary studies to address our research questions.

We have analyzed the current big data software architectures for various domains and presented the results to support the researchers and readers in terms of having a consolidated information and identifying the future research areas. We can conclude that big data software architectures are applied in various application domains. We identified recurring common motivations for adopting big data software architectures, such as supporting analytics process, improving efficiency, improving real-time data processing, reducing development costs and enabling new kind of services, including collaborative work.

As a result of the final set of primary studies quality attribute have a clear impact on the big data software architectures. The business constraints vary for each application domain, therefore targeting a big data software application for a specific application domain requires tailoring of the generic reference architectures to a domain specific reference architecture to better support derivation of the application architectures. Among the primary studies, none of the reference architectures is indicated or suggested for a specific application domain. Considering the fact that a detailed application architecture is often missing in the primary studies, an overall evaluation of the architecture is not feasible for them.

Having a uniform platform, flexibility, sparsity of the data, hiding the details of the sensor nodes and sensor heterogeneity, dynamic decision making, common service interfaces and being accessible to non-technical users are listed among the concerns that are associated with the architectures for the application domains of our selected primary studies. The architectural concerns derive the quality attributes such as safety, low latency, reliability, reuse, high performance, availability, resilience and scalability. Gathering the stakeholders' needs and applying the appropriate architectural design methodology which can be based on a reference architecture, architectural styles or patterns such as layered, cloud based, multi-agent or service oriented and architectural viewpoints like deployment, flowchart or decomposition can be utilized. The technologies to be adopted and the integration concerns at the system and system of systems level are also to be defined while deriving the application architecture from the described big data system software architecture based on an application domain.

In this study, we focused on the most relevant primary studies in the known literature of the big data software architecture domain. We analyzed the studies in terms of key concerns, application domains, stakeholders, motivations, architectural approaches, models viewpoints and discussed the strength of evidence of the and threads to validity of the results. The reliability and credibility of data and having low latency and providing real time information is the key for smart cities, while in social media efficiency, load balancing and user friendliness is critical. Similarly, scientific applications are expected to be scalable, flexible and user friendly. Within the aerospace and aviation domain, real time and offline decision making is expected. Depending on the application domain, the stakeholders vary, such as for industrial applications, designers, managers, suppliers, manufacturers and customers are among the stakeholders while for scientific platforms the stakeholders are the system administrators and engineers. The main motivation is scalability and high performance where maintainability and deployment are given less emphasis. From the architectural approach point of view, layered, cloud based, service oriented and multi-agent architectures are applied to the big data systems.

As a future work, we will analyze big data software architectures of different use cases from various application domains against our results and discuss identified challenges and possible enhancements.

Appendix 1

Table 16 A Search Strings

Electronic Database	Search String
IEEE Xplore	((("Abstract": "Big Data" OR "Publication Title": "Big Data") AND (p_Abstract: "Software Architecture" OR "Abstract": "System Architecture" OR "Abstract": "Cloud Architecture" OR "Publication Title": "Architecture")))
ACM Digital Library	((acmdlTitle:(+"Big Data") OR recordAbstract:(+"Big Data")) AND (acmdlTitle:(+Architecture) OR recordAbstract:(+"Software Architecture") OR recordAbstract:(+"System Architecture") OR recordAbstract:(+"Cloud Architecture")))
Wiley Interscience	((("Big Data" in Publication Titles OR "Big Data" in Abstract) AND (Architecture in Publication Titles OR "Software Architecture" in Abstract OR "System Architecture" in Abstract OR "Cloud Architecture" in Abstract) AND (Not poster in Article Titles))
Science Direct	(TITLE ("Big Data") or ABSTRACT ("Big Data")) and (TITLE (Architecture) or ABSTRACT ("Software Architecture") or ABSTRACT ("System Architecture") or ABSTRACT ("Cloud Architecture"))
Springer	(ti: ("Big Data") or abs: ("Big Data")) and (ti: ("Architecture") or abs: ("Software Architecture") or abs: ("System Architecture") or abs: ("Cloud Architecture"))

Appendix 2

List of primary studies

1. Ahmad, A., Paul, A., Rathore, M., Chang, H, "An efficient multidimensional big data fusion approach in machine-to-machine communication", *ACM Transactions on Embedded Computing Systems*, 2016, (TECS), 15(2), 39.
2. Alsubaiee, S., Altowim, Y., Altwaijry, H., Behm, A., Borkar, V., Bu, Y., Gabrielova, E, "AsterixDB: A scalable, open source BDMS" *Proceedings of the VLDB Endowment*, 2014, 7(14), 1905–1916.
3. Bilal, M, Oyedele, L, O, Akinade, O, O, Ajayi, S, O, Alaka, H, A, Owolabi, H, A., Bello, S, "Big data architecture for construction waste analytics (CWA): A conceptual framework", *Journal of Building Engineering*, 6, 2016, 144–156
4. Brewer, W., Scott, W., Sanford, J, "An Integrated Cloud Platform for Rapid Interface Generation, Job Scheduling, Monitoring, Plotting, and Case Management of Scientific Applications", 2015, *International Conference on Cloud Computing Research and Innovation (ICCCRI)*, 2015, pp. 156–165, IEEE.
5. Cecchinell, C., Jimenez, M., Mosser, S., Riveill, M, "An architecture to support the collection of big data in the internet of things", *IEEE World Congress on Services*, 2014, pp. 442–449.
6. Chen, C., Yan, Y., Huang, L., Dong, X, "A scalable and productive workflow-based cloud platform for big data analytics", *IEEE International Conference on Big Data Analysis (ICBDA)*, 2016, pp. 1–5.
7. Chen, J., Ma, J., Zhong, N., Yao, Y., Liu, J., Huang, R., Gao, Y, "WaaS—Wisdom as a service", *Wisdom Web of Things*, 2016, pp. 27–46, Springer, Cham.
8. Chen, H, M, Kazman, R, Haziyevev, S, "Agile big data analytics for web-based systems: An architecture-centric approach", *IEEE Transactions on Big Data*, 2(3), 2016, 234–248.

9. Cheng, B, Longo, S, Cirillo, F, Bauer, M, Kovacs, E. "Building a big data platform for smart cities: Experience and lessons from Santander", *IEEE International Congress on Big Data*, 2015, pp. 592–599.
10. Gorton, I, Klein, J, "Distribution, data, deployment: Software architecture convergence in big data systems", *IEEE Software*, 2014, 32(3), 78–85.
11. Guo, W, She, B, Zhu, X, "Remote Sensing Image On-Demand Computing Schema for the China ZY-3 Satellite Private Cloud-Computing Platform", *Transactions in GIS*, 18, 2014, 53–75.
12. Guo, K, Pan, W, Lu, M, Zhou, X, Ma, J, "An effective and economical architecture for semantic-based heterogeneous multimedia big data retrieval", *Journal of Systems and Software*, 2015, 102, 207–216.
13. Guo, Z, X, Yang, C, "Development of production tracking and scheduling system: A cloud-based architecture", *International Conference on Cloud Computing and Big Data*, 2013, pp. 420–425.
14. Han, Y, Lee, H, Kim, Y, "A real-time knowledge extracting system from social big data using distributed architecture", *Conference on research in adaptive and convergent systems*, 2015, pp. 74–79, ACM.
15. Heath, F, F, Hull, R., Khabiri, E., Riemer, M., Sukaviriya, N, Vaculín, R, "Alexandria: extensible framework for rapid exploration of social media", *IEEE International Congress on Big Data*, 2015, pp. 483–490, IEEE.
16. Tinati, R, Wang, X, Brown, I, Tiropanis, T, Hall, W. "A streaming real-time web observatory architecture for monitoring the health of social machines", *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 1149–1154, ACM.
17. Kashlev, A, Lu, S, "A system architecture for running big data workflows in the cloud", *IEEE International Conference on Services Computing*, 2014, pp. 51–58.
18. Krishnan, K, "Data warehousing in the age of big data", Newnes, 2013.
19. Krumeich, J, Jacobi, S, Werth, D, Loos, P, "Big data analytics for predictive manufacturing control-A case study from process industry", *IEEE International Congress on Big Data*, 2014, pp. 530–537.
20. Logica, B, Magdalena, R, "Using big data in the academic environment", *Procedia Economics and Finance*, 2015, 33, 277–286.
21. Manate, B, Fortis, F, Moore, P, "Applying the Prometheus methodology for an Internet of Things architecture" *IEEE/ACM 7th International Conference on Utility and Cloud Computing*, 2014, pp. 435–442.
22. Manzoor, A, Patsakis, C, Morris, A, McCarthy, J, Mullarkey, G, Pham, H, Bouroche, M, "CityWatch: exploiting sensor data to manage cities better", *Transactions on Emerging Telecommunications Technologies*, 2014, 25(1), 64–80.
23. Mishne, G, Dalton, J, Li, Z, Sharma, A, Lin, J, "Fast data in the era of big data: Twitter's real-time related query suggestion architecture", *ACM SIGMOD International Conference on Management of Data*, 2013, pp. 1147–1158, ACM.
24. Mohapatra, S, K, Prasan K, S, Shih-Lin, W, "Big data analytic architecture for intruder detection in heterogeneous wireless sensor networks." *Journal of Network and Computer Applications* 66, 2016, 236–249.
25. Munar, A, Chiner, E, Sales, I, "A big data financial information management architecture for global banking" *International Conference on Future Internet of Things and Cloud*, 2014, pp. 385–388, IEEE.

26. Ochian, A, Suciu, G, Fratu, O, Voicu, C, Suciu, V, "An overview of cloud middleware services for interconnection of healthcare platforms", *10th International Conference on Communications (COMM), 2014*, pp. 1–4, IEEE.
27. Rosenberger, M. (2015, October). "Social customer relationship management: an architectural exploration of the components", *e-Business, e-Services and e-Society, 2014*, pp. 372–385. Springer, Cham.
28. Pääkkönen, P, Pakkala, "Reference architecture and classification of technologies, products and services for big data systems" *Big data research*, 2(4), 2015, 166–186.
29. Rabelo, T, Lama, M, Amorim, R, R, Vidal, J, C, "SmartLAK: A big data architecture for supporting learning analytics services", *IEEE Frontiers in Education Conference (FIE), 2015*, pp. 1–5, IEEE.
30. Serhani, M, A, El Menshawy, M, Benharref, A, "SME2EM: Smart mobile end-to-end monitoring architecture for life-long diseases", *Computers in biology and medicine*, 68, 2016, 137–154.
31. Shi, G, Wang, H, "Research on big data real-time public opinion monitoring under the double cloud architecture", *IEEE Second International Conference on Multimedia Big Data (BigMM), 2016*, pp. 416–419, IEEE.
32. Tang, B, Chen, Z, Hefferman, G, Wei, T, He, H, Yang, Q, "A hierarchical distributed fog computing architecture for big data analysis in smart cities", *ASE BigData & SocialInformatics, 2015*, p. 28, ACM.
33. Xie, Zhiwu, et al. "Towards use and reuse driven big data management." Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries. 2015.
34. Xu, D, Wu, D, Xu, X, Zhu, L, Bass, L, "Making real time data analytics available as a service", *11th International ACM SIGSOFT Conference on Quality of Software Architectures (QoSA), 2015*, pp. 73–82, IEEE.
35. Zhou, K, Yang, S "A framework of service-oriented operation model of China's power system", *Renewable and Sustainable Energy Reviews*, 50, 2015, 719–725.
36. Zhou, Q, Simmhan, Y, & Prasanna, V, "Towards hybrid online on-demand querying of realtime data with stateful complex event processing" *IEEE International Conference on Big Data*, 2013, pp. 199–205

Appendix 3**Table 17** Study quality assessment

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
1	0,5	0,5	1	0,5	1	0	1	1	1	0,5
2	1	1	1	0,5	1	0	0,5	0,5	1	1
3	1	1	0,5	0,5	0,5	0,5	0	1	1	1
4	1	1	1	1	0,5	0	0,5	0,5	1	1
5	0,5	0,5	0	0,5	0,5	0	0	1	1	1
6	1	1	0,5	0,5	0,5	0	0	1	1	0,5
7	0,5	0,5	0	0,5	0	0	0,5	0	1	1
8	1	1	0,5	0,5	0,5	1	0,5	0,5	1	1
9	1	1	0,5	0,5	0,5	1	1	0,5	1	1
10	1	1	0	0,5	0,5	0	0	1	0,5	0,5
11	1	1	1	0,5	0	0	0	1	1	1
12	1	1	1	0,5	1	0	1	0,5	1	1
13	1	1	0,5	0,5	0	0	0	1	0,5	0,5
14	1	1	1	0,5	0,5	0	0	0,5	1	0,5
15	1	1	0,5	0,5	1	0	0,5	0,5	1	0,5
16	1	1	0,5	0,5	0,5	0	0,5	0,5	0,5	0,5
17	1	1	1	0,5	0,5	0,5	1	0,5	0,5	1
18	1	1	1	1	1	1	1	1	1	1
19	1	1	0,5	0,5	0,5	1	0	0,5	0,5	0,5
20	1	0,5	0,5	0,5	0	1	0	0,5	1	1
21	1	1	1	0,5	0,5	0	0	0,5	1	1
22	0,5	0,5	0,5	0,5	0,5	0	0	0,5	0,5	0,5
23	1	1	1	0,5	1	0	0,5	0,5	1	1
24	1	1	1	1	0,5	0	1	1	1	1
25	1	1	1	0,5	0,5	0	0	0,5	0,5	0,5
26	1	1	0,5	0	0,5	0	0	0,5	1	0,5
27	1	1	0	0,5	0,5	1	0,5	1	1	0,5
28	1	1	0	0,5	0,5	1	0,5	0,5	1	1
29	1	1	0	0,5	0,5	0	0	0,5	0,5	0,5
30	1	1	1	0,5	0,5	0,5	1	0,5	0,5	0,5
31	1	1	1	1	0,5	0	0,5	0,5	1	1
32	1	0,5	0,5	0,5	0	0	0	0,5	1	0,5
33	1	1	1	1	1	0	1	0,5	1	1
34	1	1	0,5	0,5	0	0	1	0	0,5	0,5
35	0,5	0,5	0,5	0,5	1	0	0	0,5	1	0,5
36	1	1	0	0,5	0,5	0	0,5	0,5	0,5	0,5
37	1	1	1	0,5	0,5	0,5	0,5	0,5	1	1
38	1	1	1	0,5	1	0	0,5	0,5	0,5	0,5
39	1	1	1	0,5	0	1	0,5	0,5	1	1
40	1	1	0,5	0,5	0	0	0,5	0,5	0,5	0,5
41	1	1	1	1	1	1	0,5	0,5	1	1
42	1	1	1	0,5	0,5	0	1	1	1	1
43	1	1	0,5	0,5	0,5	0	0	0,5	0,5	0,5

Table 17 Study quality assessment (*Continued*)

	Quality of Reporting	Rigor	Credibility	Relevance	Total
1	2	1,5	2	1,5	7
2	3	1,5	1	2	7,5
3	2,5	1,5	1	2	7
4	3	1,5	1	2	7,5
5	1	1	1	2	5
6	2,5	1	1	1,5	6
7	1	0,5	0,5	2	4
8	2,5	2	1	2	7,5
9	2,5	2	1,5	2	8
10	2	1	1	1	5
11	3	0,5	1	2	6,5
12	3	1,5	1,5	2	8
13	2,5	0,5	1	1	5
14	3	1	0,5	1,5	6
15	2,5	1,5	0,5	1,5	6
16	2,5	1	1	1	5,5
17	3	1,5	1,5	1,5	7,5
18	3	3	2	2	10
19	2,5	2	0,5	1	6
20	2	1,5	0,5	2	6
21	3	1	0,5	2	6,5
22	1,5	1	0,5	1	4
23	3	1,5	1	2	7,5
24	3	1,5	2	2	8,5
25	3	1	0,5	1	5,5
26	2,5	0,5	0,5	1,5	5
27	2	2	1,5	1,5	7
28	2	2	1	2	7
29	2	1	0,5	1	4,5
30	3	1,5	1,5	1	7
31	3	1,5	1	2	7,5
32	2	0,5	0,5	1,5	4,5
33	3	2	1,5	2	8,5
34	2,5	0,5	1	1	5
35	1,5	1,5	0,5	1,5	5
36	2	1	1	1	5
37	3	1	1	2	7
38	3	1,5	1	1	6,5
39	3	1,5	1	2	7,5
40	2,5	0,5	1	0,5	4,5
41	3	3	1	2	9
42	3	1	2	2	8
43	2,5	1	0,5	1	5

Appendix 4

Table 18 Data extraction

Field Area	Name
General Information	Title
	Source
	Type
	Year
	Method
	Description
Application Area	Cyber Security
	IOT/Smart Cities
	Social Big Data
	Incident/ Anomaly Detection
	Health- care
	Predictive Manufac-turing
	Banking
	Transportation
	Education
	Aerospace
	Machine to machine communication
	Recommendation System
	Environmental Sensing
	CRM
Architectural Viewpoints	Decomposition
	Deployment
	Flowchart
	No Viewpoint
Architectural Patterns	Layered
	Cloud-based
	Multi-agent
	Decoupled frontend-backend
	Service Oriented
Architectural Concerns	Hybrid
	Functional Concerns
Requirements	Non Functional Concerns
	High-level Functional Requirements
Quality Attributes	Safety
	Low latency
	Reliability
	Reuse
	High Performance
	Flexibility/extensibility
	Availability
	Accessibility
	Scalability
	Fault Tolerance

Table 18 Data extraction (*Continued*)

Field Area	Name
	Maintainability/Deployability
	Real Time Performance
	Accuracy
	Resilient
	Privacy
	Security
Methodology	Architectural Design Method
	Architectural Evaluation
Integration	Integration
	Integration Concerns
Technologies	Technologies
Stakeholders	Stakeholders
Analysis	Data analysis & Analytics (Descriptive/Prescrip)/Algorithms;
Reference Architecture	Adopted Reference Architecture

Acknowledgements

All authors would like to thank to the reviewers for the valuable comments.

Authors' contributions

All authors have contributed equally to finish this paper. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

All the material is accessible (please see references and the primary studies sections).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Information Technology Group, Wageningen University and Research, Hollandseweg 1, 6706 KN Wageningen, The Netherlands. ²Geo-Information Science and Remote Sensing Lab Wageningen University and Research, Droevendaalsesteeg 3, 6708 PB Wageningen, The Netherlands.

Received: 20 November 2019 Accepted: 30 June 2020

Published online: 14 August 2020

References

1. Gorton I, Klein J. Distribution, data, deployment: software architecture convergence in big data systems. *IEEE Softw.* 2014;32(3):78–85.
2. Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. *BMJ.* 2004;328(7454):1490. <https://doi.org/10.1136/bmj.328.7454.1490>.
3. Angelow S, Grefen P, Greefhorst D. A classification of software reference architectures: analyzing their success and effectiveness. In: Joint working IEEE/IFIP conference on Software Architecture & European Conference on software architecture; 2009. p. 141–50.
4. Gölzer P, Cato P, Amberg M. Data Processing Requirements of Industry 4.0 - Use Cases for Big Data Applications. In: Proceedings of the 23th European Conference on Information Systems (ECIS), paper 61; 2015.
5. Tan et al., 2015 C. Tan, L. Sun, K. Liu Big data architecture for pervasive healthcare: a literature review Proceedings of the Twenty-Third European Conference on Information Systems, Münster, Germany, 2015:26–29.
6. Perry DE, Wolf AL. Foundations for the study of software architecture. *ACM SIGSOFT Software Eng Notes.* 1992;17(4):40–52.
7. Rodríguez-Mazahua L, Rodríguez-Enríquez CA, Sánchez-Cervantes JL, Cervantes J, García-Alcaraz JL, Alor-Hernández G. A general perspective of big data: applications, tools, challenges and trends. *J Supercomput.* 2016;72(8):3073–113.
8. Hu H, et al. Toward scalable systems for big data analytics: a technology tutorial. *IEEE Access.* 2014;2:652–87.
9. Jin X, Wah BW, Cheng X, Wang Y. Significance and challenges of big data research. *Big Data Res.* 2015;2(2):59–64.
10. Garlan D, Shaw M. An introduction to software architecture. In: *Advances in software engineering and knowledge engineering*, 1.3.4; 1993.
11. Bachmann F, Bass L, Klein M. "Architectural tactics: a step toward methodical architectural design", technical report CMU/SEI-2003-TR-004, Pittsburgh, PA; 2003.
12. Kitchenham B, Charters S. Guidelines for performing systematic literature reviews in software engineering Technical Report, EBSE; 2007.
13. Kitchenham B, Budgen D, Brereton OP, Turner M, Bailey J, Linkman S. Systematic literature reviews in software engineering - a systematic literature review. *Inf Softw Technol.* 2009;51(1):7–15. <https://doi.org/10.1016/j.infsof.2008.09.009>.
14. Rathore M, M U, Paul A, Ahmad A, Chen BW, Huang B, et al. Real-time big data analytical architecture for remote sensing application. *IEEE J Selected Topics Appl Earth Observ Remote Sensing.* 2015;8(10):4610–21.
15. Gheisari M, Wang G, Bhuiyan MZA. A survey on deep learning in big data. In: 2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC), vol. 2: IEEE; 2017.
16. Strohbach M, Ziekow H, Gazis V, Akiva N. Towards a big data analytics framework for IoT and smart city applications. In: *Modeling and processing for next-generation big-data technologies*. Cham: Springer; 2015. p. 257–82.
17. Clements P, Garlan D, Bass L, Stafford J, Nord R, Ivers J, et al. *Documenting software architectures: views and beyond*: Pearson Education; 2002.
18. Zhang Y, Ren S, Liu Y, Si S. A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. *J Clean Prod.* 2017;142:626–41.
19. Hossain MS, Muhammad G. Emotion recognition using deep learning approach from audio-visual emotional big data. *Inf Fusion.* 2019;49:69–78.
20. Xu T, Wang D, Liu G. Banián: a cross-platform interactive query system for structured big data. *Tsinghua Sci Technol.* 2015;20(1):62–71.

21. Yan YZ, Liu RH, Yang CT, Chen ST. Cloud city traffic state assessment system using a novel architecture of big data. In: *International conference on cloud computing and big data (CCBD)*; 2015. p. 252–9. IEEE.
22. Twardowski B, Ryzko D. Multi-agent architecture for real-time big data processing. In: *2014 IEEE/WIC/ACM international joint conferences on web intelligence (WI) and intelligent agent technologies (IAT)*, vol. 3: IEEE; 2014, August. p. 333–7.
23. Singh S, Liu Y. A cloud service architecture for analyzing big monitoring data. *Tsinghua Sci Technol.* 2016;21(1):55–70.
24. Murino G, Armando A, Tacchella A. Resilience of cyber-physical systems: an experimental appraisal of quantitative measures. In: *2019 11th international conference on cyber conflict (CyCon)*, vol. 900: IEEE; 2019.
25. Rubio JE, Roman R, Lopez J. Analysis of cybersecurity threats in industry 4.0: the case of intrusion detection. In: *International conference on critical information infrastructures security*. Cham: Springer; 2017. p. 119–30.
26. Grubel BC, Reid DSD. U.S. patent no. 9,712,551. Washington, DC: U.S. Patent and Trademark Office; 2017.
27. Li Y, et al. Intelligent cryptography approach for secure distributed big data storage in cloud computing. *Inf Sci.* 2017; 387:103–15.
28. Atat R, et al. Big data meet cyber-physical systems: a panoramic survey. *IEEE Access.* 2018;6:73603–36.
29. Xu G, et al. Sensitive information topics-based sentiment analysis method for big data. *IEEE Access.* 2019;7:96177–90.
30. Shirdastian H, Laroche M, Richard M-O. Using big data analytics to study brand authenticity sentiments: the case of Starbucks on twitter. *Int J Inf Manag.* 2019;48:291–307.
31. Wamba SF, Akter S, Edwards A, Chopin G, Gnanzou D. How 'big data' can make big impact: findings from a systematic review and a longitudinal case study. *Int J Prod Econ.* 2015;165:234–46.
32. Zhang H, Babar MA, Tell P. Identifying relevant studies in software engineering. *Inf Softw Technol.* 2011;53(6):625–37. <https://doi.org/10.1016/j.infsof.2010.12.010>.
33. Babiceanu RF, Seker R. Big data and virtualization for manufacturing cyber-physical systems: a survey of the current status and future outlook. *Comput Ind.* 2016;81:128–37.
34. Lee J, Bagheri B, Kao H-A. A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manuf Lett.* 2015;3:18–23.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

