CONF-IRM 2022 Proceedings

International Conference on Information Resources Management (CONF-IRM)

10-2022

An Exploratory-Descriptive Review of Main Big Data Analytics Reference Architectures – an IT Service Management Approach

Manuel Mora

Jorge Marx Gomez

Paola Reyes-Delgado

Oswaldo Diaz

6. An Exploratory-Descriptive Review of Main Big Data Analytics Reference Architectures – an IT Service Management Approach

Manuel Mora
Autonomous University of Aguascalientes
jose.mora@edu.uaa.mx

Jorge Marx Gomez
University of Oldenburg
jorge.marx.gomez@uni-oldenburg.de

Paola Reyes-Delgado
Autonomous University of Aguascalientes
pyrd25@hotmail.com

Oswaldo Diaz
INEGI
oswaldo.diaz@inegi.org.mx

Abstract

Big Data Analytics (BDA) aims to create decision-making business value by applying multiple analytical procedures from the Statistics, Operations Research and Artificial Intelligence disciplines to huge internal and external business datasets. However, BDA requires high investments in IT resources – computing, storage, network, software, data, and environment -, and consequently the selection of the right-sized implementation is a hard business managerial decision. Parallelly, IT Service Management (ITSM) frameworks have provided best processes-practices to deliver value to end-users through the concept of IT services, and the provision of BDA as Service (BDAaaS) has now emerged. Consequently, from a dual BDA-ITSM perspective, delivering BDAaaS demands the design and implementation of a concrete BDAaaS architecture. Practitioner and academic literature on BDAaaS architectures is abundant but fragmented, disperse and uses a non-standard terminology. ITSM managers and academics involved on the problematic to deliver BDAaaS, thus, face the lack of mature practical guidelines and theoretical frameworks on BDAaaS architectures. In this research, consequently, with an exploratory-descriptive purpose, we contributed with an updated review of three main non-proprietary BDAaaS reference architectures to ITSM managers, and with a hybrid functional-deployment architectural view to the BDAaaS literature. However, given its exploratory status, further conceptual and empirical research is encouraged.

Keywords: Big Data Analytics as a Service (BDAaaS), IT Service Management (ITSM), BDAaaS architectures, Reference Architecture, NIST Big Data Reference Architecture V3.0.

1. Introduction

In the modern business environment, thousands of Big Data Analytics (BDA) projects are pursued by multiple business organizations given the expected organizational value to be generated (Oesterreich et al., 2022; Fortune, 2022). Big Data business relevance has been recognized from about one decade (Davenport et al., 2012; McAfee & Brynjolfsson, 2012). For instance, Davenport et al. (2012) realized the potential value of the 3V - volume, velocity and variety - Big Data attributes, and McAfee and Brynjolfsson (2012) qualified Big Data as a critical input to improving the modern business decision-making process due to the data richness provided by Big Data – i.e. high data variety, faster generation of data, and huge data volume –. At present days, 5V Big Data model has also included veracity attribute – i.e. quality and trust on data and data sources - and has done explicit the value attribute (Wamba et al., 2015).

Business value of Big Data only can be generated when Big Data is analyzed by human and/or machine decision makers (Klee et al., 2021). For this aim, multiple Artificial Intelligence – including Data Mining and Machine Learning approaches -, Statistics, and Database Management techniques (Phillips-Wren et al., 2015) are used through the umbrella of Analytics. Analytics was defined as the organizational ability to "collect, analyze, and act on data" (Davenport, 2006; p. 1) before its convergence with Big Data, but currently the joint Big Data Analytics approach is fundamental for modern business organization to support data-driven decision-making and creating business value (Klee et al., 2021). Big Data Analytics is the joint approach to creating data-based business value by applying analytics techniques to complex high-volume, high-velocity and high-variety data sets that require advanced technologies for their gathering and transmission, pre-processing and storage, veracity management, processing, analysis, and visualization. However, despite the highest potential business value of Big Data Analytics, its realization counter demands high investments in IT resources – computing, storage, network, software, applications, data, and IT environment – (Rao et al., 2019).

Parallelly, IT Service Management (ITSM) frameworks and standards have provided to business organizations with the best processes-practices to deliver value to end-users through the concept of IT services (Hunnebeck, 2011; TSO, 2018; ISO/IEC, 2019), and Big Data Analytics as Service (BDAaaS) (Delen & Demirkan, 2013; Wang et al., 2017) has emerged from the convergence of three components – Big Data Analytics, Big Data Analytics IT resources, and ITSM frameworks -. From the perspective of ITSM managers and academics, delivering BDAaaS implies a hard design effort given that despite the abundant literature on BDAaaS architectures (Sena et al., 2017; Ataei & Litchfield, 2020), it is highly fragmented, disperse and uses a non-standard and formal terminology (ISO/IEC/IEEE, 2011).

Consequently, to design-select the right-sized BDAaaS architecture implementation for a business organization is a hard business managerial-technical decision for ITSM managers. Similarly, the diversity, fragmentation, and lack of compliance to the standard and formal terminology of the system architecture ISO/IEC/IEEE 42010 standard (ISO/IEC/IEEE, 2011) delay the scholastic maturation of this research stream. In this research, thus, with an exploratory-descriptive purpose, we report an updated review of three main non-proprietary BDAaaS reference architectures to ITSM managers and contribute to the literature with an integrative hybrid functional-deployment architectural view. This paper continues as follows. In Section 2, we describe the research approach. In Section 3, the theoretical foundations of Big Data Analytics capabilities, and IT service architecture models are reported. In Section 4, the selected three main non-proprietary BDAaaS reference architectures are exploratory reviewed. In Section 5, a discussion of contributions is presented. Finally, in Section 6 the conclusions of this research are reported.

2. Research Method

This research applies a Conceptual Review and Analysis (CRA) research methodology adapted from Glass et al. (2004) and Mora et al. (2008), where an exploratory-descriptive was pursued. This CRA was performed through four general activities: CRA.1 Research Definition; CRA.2 Research Purpose and Method; CRA.3 Conceptual Data Collection; and CRA.4 Conceptual Analysis and Synthesis.

CRA.1 activity corresponds to Sections 1 and 3. Section 1 describes the context, knowledge gap, motivation, and methodological justification for conducting this research. Section 3 presents the technical theoretical foundations to support this research. CRA.2 and CRA.3 activities correspond to Section 2. In CRA.2, the exploratory-descriptive purpose was stated as "**to provide an updated review of top-three non-proprietary BDAaaS Reference Architectures useful to ITSM Managers relying in an integrative BDAaaS hybrid functional-deployment architectural view**". A Selective Literature Review (SeLR) method was used instead of a Systematic Literature Review (SLR) (Pare et al., 2015). SLR is usually conducted for mature domains to generate quantitative-based summaries of attributes-topics from the vast generated knowledge rather than to provide deep conceptual reviews on a small but representative group of studies (Boell & Cecez-Kecmanovic, 2015). Because BDAaaS Reference Architectures research stream is still under developing, we consider worthy a SeLR method. We applied three steps: SeLR.1 Definition of Sources and Search Statements; SeLR.2 Definition of Study Selection Criteria; and SeLR.3 Search Execution and Study Selection.

In SeLR.1 step, we defined GoogleScholar and ACM Digital Library as the two sources for searching studies. The generic search statement was defined as "TitleIncludes("big data" AND "reference architecture") AND Period(2010-2022)". In SeLR.2 step, we defined the study selection criteria as "C.1 OR C.2". C.1 was defined as ("study is published in a journal JCR or Scopus indexed journal" AND "study has been highly cited (at least 100 citations) AND "study does not address a specific domain"). C.2 was defined as ("study is reported by a trustable international association") AND "reference architecture is non-proprietary"). In SeLR.3 step GoogleScholar and ACM Digital Library located 45 and 2 studies respectively, and we applied the selection criteria (C.1 OR C.2), and two studies satisfied these conditions (Pääkkönen & Pakkala, 2015; NIST, 2019). Research team added manually a third study that was considered highly relevant given that the publisher of the manuscript is an international association that groups the main international providers of BDAaaS (Cloud Standards Consumer Council, 2017). Despite this SeLR collected only three documents, they provided a representative sample of high-quality and mature studies. First manuscript was reported in a premier journal, and it is the highest cited study on this topic (300+ times). Second manuscript is the unique and most referenced formal standard of the practice issued by the National Institute of Standards and Technology at the USA, and this study is the most extensive detailed study (65 pages in the volume six; the full standard includes nine volumes). Third study is endorsed by an international association from Cloud and Big Data Analytics professional enterprises. SeLR.2 and SeLR.3 steps, thus, implicitly used a non-random judgment (purposive) sampling approach to select units of study (Zikmund et al., 2012). Finally, CRA.4 activity correspond to sections 4 and 5.

3. Theoretical Basis

3.1 On Big Data Analytics Capabilities

In "Big Data Analytics" concur two data-based computational approaches (Phillips-Wren et al., 2015). The "Big Data" side refers to the stages of 1) Raw Data Sources Identification and Acquisition, 2) Raw Data Pre-Processing, and 3) Data Storing and Processing, and the "Analytics" side to the stages of 4) Data Modeling and Analysis, and 5) Data Access and Usage. This flow of stages is known as the "Big Data Analytics Pipeline".

To summarize, the "Big Data" side is responsible for making available processed Big Data sets with the potential of creating business value, and the "Analytics" side for providing business value through the application of analytics procedures to the processed Big Data sets. Regarding the Data Modeling and Analysis stage, there are three types of analytics procedures. Exploratory and Descriptive Analytics refers to procedures to report summary metrics and graphs of the Big Data sets that represent historical and current status of the business processes and systems related to the big data sets. Predictive Analytics refers to procedures to create data-driven models that permit estimating future status of the business processes and systems related to the Big Data sets. Prescriptive Analytics refers to procedures to create data-driven models that determine the optimal solutions or best viable alternative solutions. Table 1 reports the stages, purpose, main activities, key issues and main involved information and communication technologies (ICT) for a generic "Big Data Analytics Pipeline", adapted from the main literature (Jagadish et al., 2014; Phillips-Wren et al., 2015).

Stage	Purpose	Main Activities	Main Issues	Main Involved ICT
1. Raw Data Sources Identification and Acquisition	To identify the set of raw data sources for the big data analytics pipeline, agree legally on its accessibility, collect the agreed raw data, transmit them, and register them.	1.1 Identification of the available raw data sources. 1.2 Analysis of the available raw data sources. 1.3 Selection and legal agreement of raw data sources. 1.4 Raw data collection and transmission. 1.5 Raw data registering.	Variety of raw data formats (structured, text, image, audio, video, device signal). Velocity (generation rates of raw data). Volume (raw data size). Veracity (trust level of raw data). Value (business need for raw data). QoS metrics for LAN/WAN/Internet data transmission systems. Variety, velocity and volume of raw data.	Business ERP systems. Business devices. External IoT. Social networks. External open data repositories. External commercial data repositories. LAN/WAN/Internet data transmission systems. Cloud Platforms (OpenStack, Apache CloudStack, OpenNebula). Streaming/CEP engines (Kafka, Flink, Storm). IoT sensors (IoTDB). Data Lakes platforms (Hudi, Delta).
2. Raw Data Pre-Processing	To apply pre-processing procedures to raw data.	2.1 Raw data pre-processing (compression / decompression, cleaning, redundancy elimination, transformation).	Performance metrics for pre-processing platforms. Data security issues.	IT cluster management (Mesos, YARN). Big Data pre-processing tools (CKAN, Apache Griffin, Open Refine, DataCleaner).
3. Data Storage and Processing	To pull data of interest, apply them processing procedures, and load them in the persistent storage platforms.	3.1 Data Integration, aggregation, and representation. 3.2 Data replication. 3.3 Processed data ingestion/ETL.	Performance metrics for storage server cluster, cloud storage services, and processing server clusters. Performance metrics for processing platforms. Data security and privacy issues.	Storage servers clusters (Hadoop/HDFS). Storage processing engines (Apache Pig). Big Data warehouses (Hive, Impala, BigQuery, Presto). Big Data non-SQL databases (MongoDB, Cassandra, HBase).
4. Data Modeling and Analysis	To elaborate data-driven models and apply them analytics procedures for specific business goals.	4.1 Exploratory and descriptive analytics (OLAP, descriptive statistics, descriptive charts/graphs). 4.2 Predictive analytics (classification, regression, clustering, association). 4.3 Prescriptive analytics (optimization, simulation, heuristic methods, expert systems).	Performance metrics for processing servers clusters, and analytics platforms. Taxonomy of exploratory-descriptive, predictive and/or prescriptive analytics procedures.	Big Data Analytics servers clusters (Mahout, Apache Drill, Spark, MLlib, RHadoop, RHive, TensorFlow, Pytorch, Keras). Big Graphs engines (GraphX, GraphLab, neo4j, Giraph, ArangoDB).
5. Data Access and Usage	To use data-driven models in stand-alone and/or embedded into end-user or automatic control systems for specific business goals.	5.1 Visual interactive analytics. 5.2 Development of end-user big data analytics systems. 5.3 Development of automatic control big data analytics systems.	QoS metrics for LAN/WAN/Internet data transmission systems. Usability metrics. Performance metrics. Business goals metrics.	LAN/WAN/Internet data transmission systems. Laptops, desktops, mobile devices, IoT devices, workstations. Web programming languages. Visual interactive analytics packages (Kibana, Google Data Studio, MS Power BI, RStudio).

Table 1: A Generic Big Data Analytics Pipeline

3.2 On Systems Architectures and IT Service Management Frameworks

The concept of system architecture is fundamental for achieving a high-quality and cost-efficient IT service (ISO/IEC/IEEE, 2011). According to the ISO/IEC/IEEE 42010 standard (ISO/IEC/IEEE, 2011), the architecture of a system conveys the “fundamental concepts or properties of a system in its environment embodied in its elements, relationships, and in the principles of its design and evolution” (ISO/IEC/IEEE, 2011; p. 2). The architecture of a system, which is abstract, is manifested through the functional and non-functional properties of the system. Systems Engineering discipline (ISO/IEC/IEEE, 2011; ISO/IEC/IEEE, 2015) defines a system as a set of interacting elements integrated for achieving a purpose. Systems Engineering addresses systems that are “man-made and may be configured with one or more of the following: hardware, software, data, humans, processes (e.g., processes for providing service to users), procedures (e.g. operator instructions), facilities, materials and naturally occurring entities” (ISO/IEC/IEEE, 2015; p.1).

To guide systems architects in the design of a system architecture, have been proposed Architecture Frameworks (AF), Reference Architectures (RA), and Architecture Design Processes and Practices (ADPP) (ISO/IEC/IEEE, 2011; Angelov et al., 2012). An AF “establishes a common practice for creating, interpreting, analyzing and using architecture descriptions within a particular domain of application or stakeholder community” (ISO/IEC/IEEE, 2011; p. 9). A RA refers to “a generic architecture for a class of systems that is used as a foundation for the design of concrete architectures from this class” (Angelov et al., 2012; p. 417). ADPP define the activities and practices for analyzing the functional and non-functional architectural requirements, designing candidate architectures, and selecting the solution architecture. According to (ISO/IEC/IEEE, 2011), the architecture of a system can be designed and represented through an architecture description (AD) document. An AD document reports stakeholders and their concerns, architecture decisions and rationale, and architecture views and viewpoints. Stakeholders are any entity that will be affected by the system of interest. Concerns are the expected system properties of interest for the stakeholders. Architecture decisions and rationale are the architectural design selections done and their justifications. Architecture views are diagrams – called architecture models – governed by architecture viewpoints that depict a set of specific concerns.

The main ITSM frameworks and standards – ITIL v2011, ITIL v4, and ISO/IEC 20000:2019– do not provide Architecture Frameworks nor Reference Architectures for IT services (Hunnebeck, 2011; TSO, 2018; ISO/IEC, 2019). However, the main ITSM frameworks and standards have provided the best processes-practices to deliver business value to IT users through the concept of IT services. An IT service can be defined as a functionality enabled to IT users that delivers business value and that is provided by an IT service system composed of IT resources, IT processes-practices, and IT people (Hunnebeck, 2011; ISO/IEC, 2018; TSO, 2018). Value is realized when the expected IT service utility (fit for purpose) and IT service warranty (fit for use) are achieved. The utility of an IT service refers to what the service does that is valued by the customer. The warranty for an IT service refers to how well it is delivered – i.e. how well are reached the levels of availability, capacity, continuity, and security agreed -. Figure 1 - adapted from (Hunnebeck, 2011) – illustrates the concept of IT service and IT service system. The specific elements of the IT service system are: IT resources (APP: end-user applications; SW: software base; HW: hardware equipment; NW: network devices; DATA: datasets; and ENV: physical environment); IT processes-practices (applied by IT Teams and IT Suppliers to manage the IT resources to provide the IT services), and IT people (IT Teams; IT Suppliers).

Big Data Analytics as Service (BDAaaS) can be delivered through an on-Premise or a Cloud-based deployment model (Rao et al., 2019). Independently of the type of BDAaaS deployment, BDAaaS can be delivered in three different service models (Mell & Grance, 2011): BDASaaS (BDA software as a service), BDAPaaS (BDA platform as a service), or BDAIaaS (BDA infrastructure as a service).

BDAIaaS refers to the customer agreement for paying the utilization of physical and virtual IT resources. The cloud provider owns and hosts the physical IT resource layer, but the BDAIaaS customer remotely manages them. In this BDAIaaS provision model, the customer is free and responsible to install and manage the upper cloud layers. BDAPaaS provision model refers to the customer agreement for paying the utilization of the required cloud layers for developing BDA systems. These cloud layers are Big Data Cluster Management, Big Data Analytics Cluster Management, and Big Data Analytics Development Tools. The two lower cloud layers are considered black boxes, and the next upper layer is responsibility of the customer. Finally, BDASaaS refers to the customer agreement for paying the utilization of an end-user Big Data Analytics system. All lower cloud layers are black boxes for the customer. Figure 2 illustrates the three IT service models for BDAaaS using a hybrid functional-deployment architectural view from a cloud-based IT service provider viewpoint. Figure 2 maps also the generic Big Data Analytics pipeline reported in Table 1.

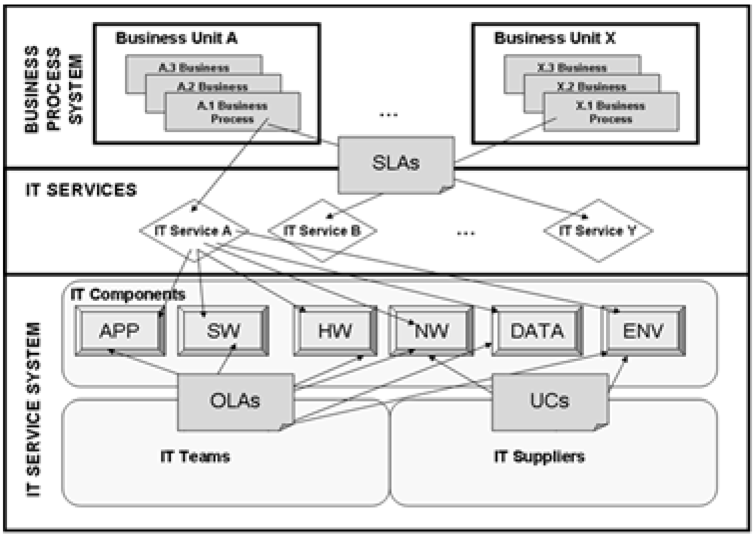

Figure 1: IT Service and IT Service System Concepts

(Source: adapted from Hunnebeck, 2011)

4. Exploratory-Descriptive Review of main BDAaaS Reference Architectures

3.1 BDAaaS Reference Architecture Conceptual Lenses

To conduct this exploratory-descriptive review, we have derived from the main literature a BDAaaS hybrid functional-deployment architectural view from a cloud-based IT service provider viewpoint – Fig. 2 – with six functional layers (Physical IT Resources, Virtual IT Resources, Big Data Storage Cluster Management, Big Data Analytics Cluster Management, Big Data Analytics Development Methods, and End-User Big Data Analytics Systems). The two bottom layers correspond to the BDAIaaS. The next three layers correspond to the BDAPaaS, and the last top layer corresponds to the BDASaaS.

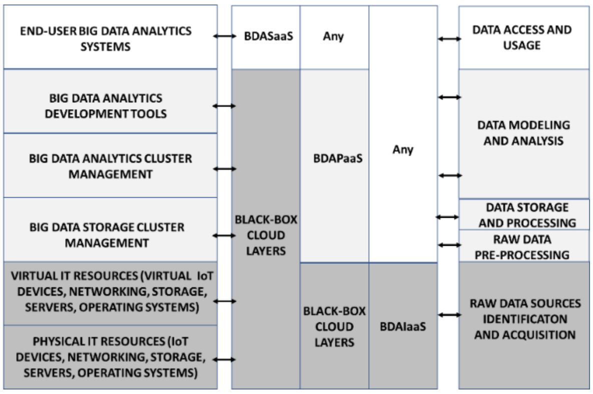
END-USER BIG DATA ANALYTICS SYSTEMS
BIG DATA ANALYTICS DEVELOPMENT TOOLS
BIG DATA ANALYTICS CLUSTER MANAGEMENT
BIG DATA STORAGE CLUSTER MANAGEMENT
VIRTUAL IT RESOURCES (VIRTUAL IoT DEVICES, NETWORKING, STORAGE, SERVERS, OPERATING SYSTEMS)
PHYSICAL IT RESOURCES (IoT DEVICES, NETWORKING, STORAGE, SERVERS, OPERATING SYSTEMS)
BDASaaS
Any
BLACK-BOX CLOUD LAYERS
BDAPaaS
Any
BLACK-BOX CLOUD LAYERS
BDAIaaS
DATA ACCESS AND USAGE
DATA MODELING AND ANALYSIS
DATA STORAGE AND PROCESSING
RAW DATA PRE-PROCESSING
RAW DATA SOURCES IDENTIFICATON AND ACQUISITION

Figure 2: A BDAaaS Hybrid Functional-Deployment Architectural View from a Cloud-based IT Service Provider Viewpoint (Source: authors)

In this hybrid functional-deployment architectural view, we have included a generic 5-stage Big Data Analytics pipeline – Table 1-. The first stage of Raw Data Sources Identification and Acquisition is mapped to the two bottom layers of virtual and physical IT resources. Internal and external, structured and non-structured, and batch, interactive or stream data sources from business enterprise systems, business devices, external IoT networks, social networks, external open data repositories, and external commercial data repositories, need to be identified and acquired. LAN/WAN/Internet data transmission systems, cloud platforms, streaming/CEP engines (such as Kafka, Flink, or Storm), IoT sensors databases (such as IoTDB), and data lakes platforms (such as Hudi, Delta) are design components that must be also considered for the first stage. For this aim in the first stage, the two mapped bottom cloud layers refer to the virtual and physical IT resources that enable access to these data sources. These two cloud layers correspond to the BDAIaaS delivering model.

The second and third stages of Raw Data Pre-Processing, and Data Storage and Processing, were mapped to the third cloud layer of Big Data Cluster Management. This cloud layer refers to the IT Big Data tools for managing the SQL- and non-SQL based storage through a cluster of storage nodes, as well as for pre-processing (compression / decompression, cleaning, redundancy elimination, transformation) and processing (integration, aggregation, representation, replication, and processed data ingestion/ETL) tasks. In this third cloud layer, the design components are IT cluster management systems (such as Mesos, YARN), Big Data pre-processing tools (such as CKAN, Apache Griffin, Open Refine, DataCleaner), Storage Servers clusters (such as Hadoop/HDFS), Storage Processing engines (such as Apache Pig), Big Data warehouses (such as Hive, Impala, BigQuery, Presto), and Big Data non-SQL databases (such as MongoDB, Cassandra, HBase).

The fourth stage of Data Modeling and Analysis was mapped to the fourth and fifth cloud layers of Big Data Analytics Cluster Management, and Big Data Analytics Development Tools. These cloud layers enable the design and building of data-driven models and the application of analytics procedures for specific business goals. Analytics procedures can be Exploratory and Descriptive (e.g. OLAP, descriptive statistics, and descriptive charts/graphs), Predictive (e.g. classification, regression, clustering, and association), and Prescriptive (e.g. optimization, simulation, heuristic methods, and expert systems). In these fourth and fifth cloud layers, the design components are Analytics Servers clusters, Big Analytics engines (such as Mahout, Apache Drill, Spark, MLlib, RHadoop, RHive, TensorFlow, Pytorch, Keras), and Big Graphs engines (such as GraphX, GraphLab, neo4j, Giraph, ArangoDB). These third, fourth and fifth cloud layers correspond to the BDAPaaS delivering model.

The fifth stage of Data Access and Usage was mapped to the sixth cloud layer of End-User Big Data Analytics Systems. This top cloud layer enacts the remote access and utilization of the data-driven models in stand-alone applications and/or embedded into end-user or automatic control systems for specific business goals. This sixth cloud layer corresponds to the BDASaaS delivering model.

4.2 Review of BDAaaS Reference Architectures

For BDAaaS, several proprietary Reference Architectures from IT business consulting companies have been proposed. From the non-proprietary side, three main BDAaaS Reference Architectures are available. These are: Reference Architecture for Big Data Systems (RABDS) (Pääkkönen & Pakkala, 2015), Cloud Customer Architecture for Big Data and Analytics V2.0 (CCABDA) (Cloud Standards Consumer Council, 2017), and NIST Big Data Reference Architecture (NBDRA) V3.0 (NIST, 2019).

RABDS (Pääkkönen & Pakkala, 2015) was proposed from an inductive design from seven real cases. RABDS includes seven primary layers (Data Sources, Data Extraction, Data Loading and Pre-processing, Data Processing, Data Analysis, Data Loading and Transformation, and Interfacing and Visualization) and two support layers (Data Storage, Job Model and Specification). RABDS architectural views are reported as a Big Data Pipeline. From a BDAaaS perspective, no information is provided. Figure 3 – derived from (Pääkkönen & Pakkala, 2015) – illustrates a functional architectural view of RABDS.

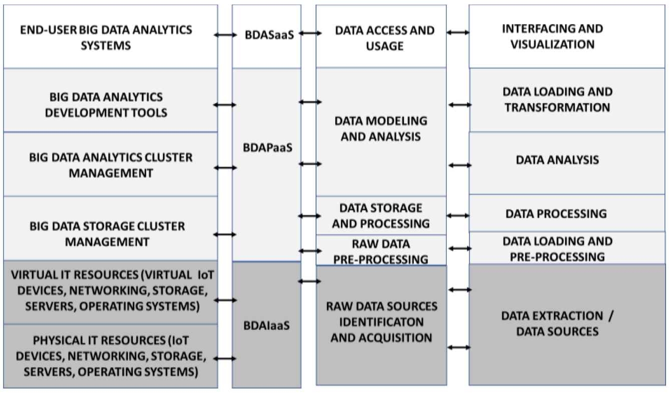
END-USER BIG DATA ANALYTICS SYSTEMS
BIG DATA ANALYTICS DEVELOPMENT TOOLS
BIG DATA ANALYTICS CLUSTER MANAGEMENT
BIG DATA STORAGE CLUSTER MANAGEMENT
VIRTUAL IT RESOURCES (VIRTUAL IoT DEVICES, NETWORKING, STORAGE, SERVERS, OPERATING SYSTEMS)
PHYSICAL IT RESOURCES (IoT DEVICES, NETWORKING, STORAGE, SERVERS, OPERATING SYSTEMS)
BDASaaS
BDAPaaS
BDAIaaS
DATA ACCESS AND USAGE
DATA MODELING AND ANALYSIS
DATA STORAGE AND PROCESSING
RAW DATA PRE-PROCESSING
RAW DATA SOURCES IDENTIFICATON AND ACQUISITION
INTERFACING AND VISUALIZATION
DATA LOADING AND TRANSFORMATION
DATA ANALYSIS
DATA PROCESSING
DATA LOADING AND PRE-PROCESSING
DATA EXTRACTION / DATA SOURCES

Figure 3: RABD mapped to the BDAaaS Hybrid Functional-Deployment Architectural View from a Cloud-based IT Service Provider Viewpoint (Source: authors)

CCABDA (Cloud Standards Consumer Council, 2017), provides a reference for deploying BDAaaS using three network zones: public network, provider cloud, and enterprise network. The core components of the public network are Public Data Sources and SaaS Applications. The core components of the provider cloud are Streaming Computing, Data Repositories, Cognitive Assisted Data Integration, Cognitive Analytics Discovery and Exploration, Cognitive Actionable Insights, API Management, Transformation and Connectivity, and Security. The core components of the private enterprise network are Enterprise Data, and Enterprise Applications. CCABDA (Cloud Standards Consumer Council, 2017), promotes explicitly BDASaaS and implicitly BDAPaaS. BDAIaaS is not promoted explicitly but it is referred as a capability infrastructure functionality required for BDAaaS. Capability infrastructure refers to “platform tools that enable connectivity, load balancing, routing, and the like, or hardware resources such as suitable storage, compute, and networking.” (Cloud Standards Consumer Council, 2017; p. 20). Figure 4 – derived from (Cloud Standards Consumer Council, 2017) – illustrates a functional architectural view of CCABDA.

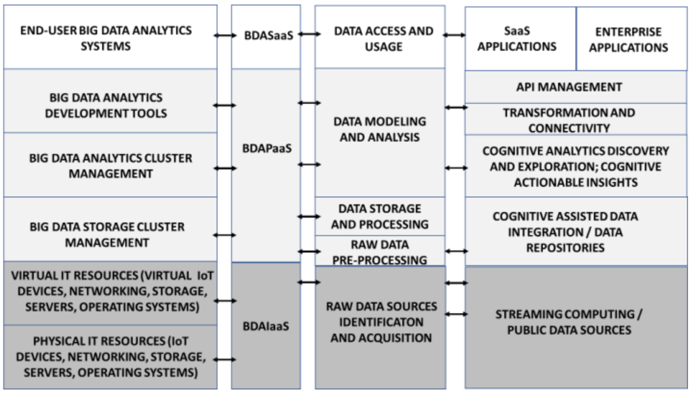
END-USER BIG DATA ANALYTICS SYSTEMS
BIG DATA ANALYTICS DEVELOPMENT TOOLS
BIG DATA ANALYTICS CLUSTER MANAGEMENT
BIG DATA STORAGE CLUSTER MANAGEMENT
VIRTUAL IT RESOURCES (VIRTUAL IoT DEVICES, NETWORKING, STORAGE, SERVERS, OPERATING SYSTEMS)
PHYSICAL IT RESOURCES (IoT DEVICES, NETWORKING, STORAGE, SERVERS, OPERATING SYSTEMS)
BDASaaS
BDAPaaS
BDAIaaS
DATA ACCESS AND USAGE
DATA MODELING AND ANALYSIS
DATA STORAGE AND PROCESSING
RAW DATA PRE-PROCESSING
RAW DATA SOURCES IDENTIFICATON AND ACQUISITION
SaaS APPLICATIONS
ENTERPRISE APPLICATIONS
API MANAGEMENT
TRANSFORMATION AND CONNECTIVITY
COGNITIVE ANALYTICS DISCOVERY AND EXPLORATION; COGNITIVE ACTIONABLE INSIGHTS
COGNITIVE ASSISTED DATA INTEGRATION / DATA REPOSITORIES
STREAMING COMPUTING / PUBLIC DATA SOURCES

Figure 4: CCABDA mapped to the BDAaaS Hybrid Functional-Deployment Architectural View from a Cloud-based IT Service Provider Viewpoint (Source: authors)

NBDRA V3.0 (NIST, 2019) consists of a vendor-neutral, technology- and infrastructure-agnostic conceptual model and two architectural views (activity view and functional view). It was designed by NITS (National Institute of Standards and Technology, USA) after several rounds of sessions in the NIST Big Data Public Working Group (NBD-PWG) with participants from industry, academia, and government agencies. According to NIST (2019; p. 3) a reference architecture provides "an authoritative source of information about a specific subject area that guides and constrains the instantiations of multiple architectures and solutions.". NBDRA supports the requirements of interoperability, portability, reusability, extensibility, data usage, analytics, and technology infrastructure. NBDRA is structured with five main functional components (System Orchestrator, Data Provider, Big Data Application Provider, Big Data Framework Provider (Infrastructures Frameworks, Processing Frameworks, Data Platforms Frameworks), and Data Consumer) and two fabrics (Management Fabric, and Security and Privacy Fabric) that that provide critical internal support services for the five functional components. Figure 5 – derived from (NIST, 2019) – illustrates a functional architectural view of NBDRA.

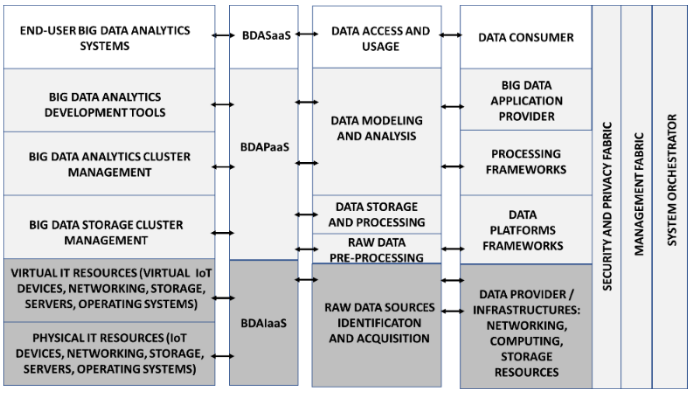
END-USER BIG DATA ANALYTICS SYSTEMS
BIG DATA ANALYTICS DEVELOPMENT TOOLS
BIG DATA ANALYTICS CLUSTER MANAGEMENT
BIG DATA STORAGE CLUSTER MANAGEMENT
VIRTUAL IT RESOURCES (VIRTUAL IoT DEVICES, NETWORKING, STORAGE, SERVERS, OPERATING SYSTEMS)
PHYSICAL IT RESOURCES (IoT DEVICES, NETWORKING, STORAGE, SERVERS, OPERATING SYSTEMS)
BDASaaS
BDAPaaS
BDAIaaS
DATA ACCESS AND USAGE
DATA MODELING AND ANALYSIS
DATA STORAGE AND PROCESSING
RAW DATA PRE-PROCESSING
RAW DATA SOURCES IDENTIFICATON AND ACQUISITION
DATA CONSUMER
BIG DATA APPLICATION PROVIDER
PROCESSING FRAMEWORKS
DATA PLATFORMS FRAMEWORKS
DATA PROVIDER / INFRASTRUCTURES: NETWORKING, COMPUTING, STORAGE RESOURCES
SECURITY AND PRIVACY FABRIC
MANAGEMENT FABRIC
SYSTEM ORCHESTRATOR

Figure 5: NBDRA mapped to the BDAaaS Hybrid Functional-Deployment Architectural View from a Cloud-based IT Service Provider Viewpoint (Source: authors)

Table 2 reports a summary of the main findings found in this research for the three Reference Architectures analyzed.

	BDAaaS Reference Architectures		
Issue	**Reference Architecture for Big Data Systems**	**Cloud Customer Architecture for Big Data and Analytics V2.0**	**NIST Big Data Reference Architecture V3.0**
Formal ISO/IEC/IEEE 42010 terminology?	**No.** A high-level implementation view is considered. Stakeholders' concerns, architecture decisions and rationale, and additional views and viewpoints are not reported.	**Partial.** A high-level functional view is considered. Stakeholders' concerns are reported. Architecture decisions and rationale, and additional views and viewpoints are not reported.	**Yes.** Stakeholders' concerns, architecture decisions and rationale, and diverse views and viewpoints are reported (high-level conceptualization view, activities view, and functional components view).
Big Data Analytics pipeline stages?	**Yes.** Seven main stages and two support stages.	**No.** No explicit Big Data Analytics pipeline is reported. An implicit one can be derived from the high-level functional view.	**Yes.** Five main stages in the Big Data Application Provider component is reported.
BDAaaS delivering models?	**No.** BDAIaaS, BDAPaaS, and BDASaaS are not reported.	**Partial.** Only the BDASaaS is considered.	**Yes.** Cloud deployment issues are reported.

IT Service Management terminology?	**Yes**. Essential issues are considered.	**Yes.** Essential issues are considered.	**Yes**. Essential issues are considered.
BDA technologies for design components?	**Yes.** BDA technologies are considered for the stages.	**No.** BDA technologies are not considered for the stages.	**No.** BDA technologies are not considered for the stages.
Contribution to ITSM managers?	**Yes.** It provides a RA for BDA systems and analysis of seven real BDA platforms.	**Partial.** It provides a RA for BDA systems but a BDA pipeline is not reported.	**Yes.** It provides a full comprehensive RA for BDA systems.
Contribution to BDAaaS literature?	**Yes.** It provides a RA for BDA systems, and a classification of BDA technologies.	**Partial.** It provides a RA for BDA systems but limited to BDASaaS type.	**Yes.** It provides a RA for BDA systems well-documented using the ISO/IEC/IEEE 42010 standard.

Table 2: Summary of Findings

4.3 Contributions

We consider this exploratory-descriptive review provides contributions to ITSM practitioners and literature focused on designing BDAaaS architectures. Previously, two Systematic Literature Review (SLR) studies (Sena et al., 2017; Ataei & Licthfield, 2020), have provided important contributions to this research stream. These SLR studies located 19 and 23 final studies respectively – after several filters-. Both SLR studies identified an accounting of several expected architectural quality requirements – Consistency, Scalability, Real-Time Operation, High Performance, Security, Availability, Modularity, and Interoperability, all of them mapped to the ISO/IEC 25010 standard (ISO/IEC, 2011) -, and five common expected architectural layers (L1 Data Sources, L2, Data Integration, L3 Data Storage, L4 Data Analytics, and L5 Data Visualization). However, SLR uses a shallow quantitative-oriented analysis with summarization purpose (Boell & Cecez-Kecmanovic, 2015), and thus their insights are limited. In these two SLR studies, most of the reported studies were short papers, did not provide sufficient technical design details, did not use the terminology and concepts from the system architecture ISO/IEC/IEEE 42010 standard (ISO/IEC/IEEE, 2011), did not consider the ITSM approach neither the relevant BDAaaS concept, and some of them are proprietary models requiring additional high consulting costs to accessing them, with some particular exceptions (NIST, 2011; Pääkkönen & Pakkala, 2015).

This research provides an updated descriptive review of the three main BDAaaS Reference Architectures reported at present, which helps ITSM managers to acquire a better understanding on the architectural design implications for delivering BDAaaS. ITSM managers, thus, can use this review for elaborating a high-level design for a required BDAaaS, avoiding to adding extra unnecessary architectural layers or omitting required layers. ITSM managers have also a brief but informative list of the main IT technologies possible to deliver BDAaaS. This research also contributes to the BDAaaS literature providing a hybrid functional-deployment architectural view that includes an updated integrative generic Big Data Analytics Pipeline. This research makes sense also that ITSM core literature on IT service architecture design required maturation toward the utilization of formal systems architectures standards. Finally, we consider this research contributes scholastically providing implicitly a didactical resource that organizes the vast but fragmented, disperse and using informal terminology literature on BDAaaS.

5. Conclusions

This research reviewed three of the main Reference Architectures for BDAaaS and illustrated their correspondence with a hybrid functional-deployment architectural view from a cloud-based IT service provider viewpoint derived from the core literature. This correspondence also included an updated generic Big Data Analytics Pipeline, and a brief but succinct exemplification of BDA technologies that can be used as design components for the BDAaaS architecture. From a practitioner perspective, the three architecture descriptions provided useful practical insights (i.e. a high-level conceptualization, main functional components, BDA pipeline stages, and BDA technologies). From a theoretical perspective (i.e. architecture of systems), only the NIST Big Data Reference Architecture V3.0 (NIST, 2019) description is reported formally (i.e. it uses the expected terminology and conceptual structures from the systems architecture literature).

Hence, this research contributes with an updated review of three main non-proprietary BDAaaS Reference Architectures to ITSM managers, and adds to the BDAaaS literature, a hybrid functional-deployment architectural view that includes an updated integrative generic Big Data Analytics Pipeline. However, further conceptual and empirical research to reach a mature theoretical BDAaaS Reference Architecture, and their associated application guidelines for ITSM managers is required.

References

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

Ataei, P., and Litchfield, A. (2020). Big data reference architectures - a systematic literature review. ACIS 2020 Proceedings, 30. Https://aisel.aisnet.org/acis2020/30

Boell, S. K., and Cecez-Kecmanovic, D. (2015). On being ‘systematic’ in literature reviews in IS. *Journal of Information Technology*, 30, 161-173.

Cloud Standards Consumer Council (2017). *Cloud Customer Architecture for Big Data and Analytics V2.0.* Cloud Standards Consumer Council.

Davenport, T. (2006). Competing on analytics. *Harvard Business Review, 84*(1), 98-107.

Davenport, T., Barth, P., and Bean, R. (2012). How Big Data Is Different. *MIT Sloan Management Review, 54*(1), 22-24.

Delen, D. and Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems, 55,* 359-363.

Fortune (2022). *Infographics on Big Data Analytics Market*. Retrieved 15 August 2022, from https://www.fortunebusinessinsights.com/infographics/big-data-analytics-market-106179

Glass, R., Ramesh, V. and Vessey, I. (2004). An analysis of research in computing disciplines. *Communications of the ACM, 47*(6), 89-94.

Hunnebeck, L. (2011). *Service Design*, London: The Stationary Office (TSO).

ISO/IEC/IEEE (2011). *ISO/IEC/IEEE: 42010: 2011 systems and software engineering, architecture description*. Geneva, Switzerland: International Organization for Standardization.

ISO/IEC/IEEE (2015). *ISO/IEC/IEEE 15288: 2015 – Systems and software engineering–System life cycle processes*. Geneva, Switzerland: International Organization for Standardization.

ISO/IEC (2019). *ISO/IEC 20000-2:2019, Information technology — Service management — Part 2: Guidance on the application of service management systems*. Geneva, Switzerland: International Organization for Standardization.

Klee, S., Janson, A. and Leimeister, M. (2021). How data analytics competencies can foster business value–A systematic review and way forward. *Information Systems Management, 38*(3), 200-217.

Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J., Ramakrishnan, R. and Shahabi, C. (2014). Big Data and its Technical Challenges. *Communications of the ACM, 57*(7), 86-94.

McAfee, A. and Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review, 90*(10), 1-9.

Mell, P. and Timothy Grance. (2011). *The NIST definition of cloud computing. Special Publication 800-145*. Gaithersburg, MD: National Institute of Standards and Technology.

Mora, M., Gelman, O., Paradice, D. and Cervantes, F. (2008). *The Case for Conceptual Research in Information Systems*. In CONF-IRM 2008 Proceedings, 52. Http://aisel.aisnet.org/confirm2008/52

NIST (2019). *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture Version 3. NIST Special Publication 1500-6r2*. Gaithersburg, MD: National Institute of Standards and Technology.

Oesterreich, T. D., Anton, E., and Teuteberg, F. (2022). What translates big data into business value? A meta-analysis of the impacts of business analytics on firm performance. *Information & Management*, 103685.

Pääkkönen, P. and Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems. *Big Data Research 2*(4), 166-186.

Pare, G., Trudel, M., Jaana, M. and Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, 52, 183-199.

Phillips-Wren, G., Iyer, L., Kulkarni, U. and Ariyachandra, T. (2015). Business analytics in the context of big data: A roadmap for research. *Communications of the Association for Information Systems, 37*(1), 448-472.

Rao, T., Mitra, P., Bhatt, R. and Goswami, A. (2019). The big data system, components, tools, and technologies: a survey. *Knowledge and Information Systems, 60*(3), 1165-1245.

Sena, B., Allian, A. P., and Nakagawa, E. Y. (2017). Characterizing big data software architectures: a systematic mapping study. In *Proceedings of the 11th Brazilian Symposium on Software Components, Architectures, and Reuse* (pp. 1-10).

TSO (2018). *ITIL 4, Create, Deliver, Support*. London: The Stationary Office (TSO).

Wamba, S., Akter, S., Edwards, A., Chopin, G. and Gnanzou, D. (2015). How 'big data'can make big impact: findings from a systematic review and a longitudinal case study. *International Journal of Production Economics, 165*, 234-246.

Wang, X., Yang, L., Liu, H. and Deen, M. (2017). A big data-as-a-service framework: State-of-the-art and perspectives. *IEEE Transactions on Big Data, 4*(3), 325-340.