TECHNICAL REPORT

ISO/IEC TR 20547-1

First edition
2020-08

Information technology — Big data reference architecture —

Part 1: Framework and application process

Technologies de l'information — Architecture de référence des mégadonnées —

Partie 1: Cadre méthodologique et processus d'application

Reference number
ISO/IEC TR 20547-1:2020(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword ..... **iv**
Introduction ..... **v**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Abbreviated terms **2**
5 Document overview **3**
6 Big data standardization: motivation and objectives **3**
7 Conceptual foundations **5**
7.1 General 5
7.2 Reference architecture concepts 5
7.3 Reference architecture structure 6
8 Big data reference architecture elements **7**
8.1 Overview 7
8.2 Stakeholders 8
8.3 Concerns 9
8.4 Views 9
8.4.1 User view 10
8.4.2 Functional view 10
9 Big data reference architecture application process **10**
9.1 Overview 10
9.2 Identify stakeholders and concerns 11
9.3 Map stakeholders and concerns to roles and subroles 11
9.4 Develop detailed activity descriptions and map to concerns 12
9.5 Define functional components to implement activities 13
9.6 Cross walk activities/functional components back to concerns 13
Bibliography **14**

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see http://patents.iec.ch).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 42, *Artificial intelligence*.

A list of all parts in the ISO/IEC 20547 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

The big data paradigm is a rapidly changing field with rapidly changing technologies. This dynamic situation creates two significant issues for potential implementers of the technology. First, there is a lack of standard definitions for terms including the core concept of big data. The second issue is that there is no consistent approach to describe a big data architecture and implementation. The first issue is addressed by ISO/IEC 20546. The ISO/IEC 20547 series is targeted to the second issue and provides a framework and reference architecture which organizations can apply to their problem domain to effectively and consistently describe their architecture and its implementations with respect to the roles/actors and their concerns as well as the underlying technology. This document describes the reference architecture framework and provides a process for mapping a specific problem set/use case to the architecture and evaluating that mapping.

Information technology — Big data reference architecture —

Part 1: Framework and application process

1 Scope

This document describes the framework of the big data reference architecture and the process for how a user of the document can apply it to their particular problem domain.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC/IEEE 42010, *Systems and software engineering — Architecture description*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC/IEEE 42010 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1
big data
extensive datasets — primarily in the characteristics of volume, variety, velocity, and/or variability — that require a scalable technology for efficient storage, manipulation, and analysis

Note 1 to entry: Big data is commonly used in many different ways, for example as the name of the scalable technology used to handle big data extensive datasets.

[SOURCE: ISO/IEC 20546:2019, 3.1.2]

3.2
reference architecture
in the field of software architecture or enterprise architecture, provides a proven template solution for an architecture for a particular domain, as well as a common vocabulary with which to discuss implementations, often with the aim of stressing commonality

[SOURCE: ISO/TR 14639-2:2014, 2.65]

3.3
framework
particular set of beliefs, or ideas referred to in order to describe a scenario or solve a problem

[SOURCE: ISO 15638-6:2014, 4.30]

3.4
security
protection against intentional subversion or forced failure. A composite of four attributes — confidentiality, integrity, availability, and accountability — plus aspects of a fifth, usability, all of which have the related issue of their assurance

[SOURCE: ISO/IEC/IEEE 15288:2015, 4.1, 31]

3.5
privacy
right of individuals to control or influence what information related to them may be collected and stored and by whom that information may be disclosed

[SOURCE: ISO/IEC TR 26927:2011, 3.34]

3.6
provenance
information on the place and time of origin, derivation or generation of a resource or a record or proof of authenticity or of past ownership

[SOURCE: ISO/IEC 11179-7:2019, 3.1.10]

3.7
SQL
database language specified by ISO/IEC 9075

Note 1 to entry: SQL is sometimes interpreted to stand for Structured Query Language but that name is not used in the ISO/IEC 9075 series.

[SOURCE: ISO/IEC 20546:2019, 3.1.36]

3.8
lifecycle
evolution of a system, product, service, project or other human-made entity from conception through retirement

[SOURCE: ISO/IEC/IEEE 15288:2015, 4.1.23]

4 Abbreviated terms

BDA	big data auditor
BDAcP	big data access provider
BDAnP	big data analytics provider
BDAP	big data application provider
BDCP	big data collection provider
BDFP	big data framework provider
BDIP	big data infrastructure provider
BDPlaP	big data platform provider
BDPreP	big data preparation provider
BDProP	big data processing provider

BDSD	big data service developer
BDSO	big data system orchestrator
BDSP	big data service partner
BDRA	big data reference architecture
BDVP	big data visualization provider
GDPR	general data protection regulation
JSON	Javascript object notation
RDF	resource description framework
SQuaRE	systems and software quality requirements and evaluation
XML	extensible markup language

5 Document overview

This document is designed to introduce the reader to certain big data reference architecture concepts so that they can apply the other documents in the ISO/IEC 20547 series to their specific system and problem set.

Clauses 6 to 9:

— give the motivation and objectives behind big data standards;

— provide an introduction to reference architectures and their purpose;

— provide an overview of the BDRA and an explanation of its key concepts;

— provide a process on application of the BDRA to a problem domain.

This document can be leveraged in various ways when reading and applying the ISO/IEC 20547 series:

a) if the user intends to read only this document to gain a general understanding of the BDRA and its applicability to his/her problem space, he/she can concentrate on Clauses 5, 6, and 7;

b) if the user is developing a big data architecture and wishes to align it to the BDRA, then he/she can follow the process in Clause 8.

6 Big data standardization: motivation and objectives

In a 2019 report, IDC forecast worldwide revenues for big data and data analytics of 189,1 billion USD, a 12 % increase over 2018 and predicts a five-year compound annual growth rate of 13,2 % with revenues in 2022 exceeding 274,3 billion USD[15].

In addition, buyers and implementers of big data systems deal with an exploding number of technologies and options — many of which get wrapped by the vendors in the buzz words including the undefined term big data. In order for the stakeholders of big data systems to understand what they are buying and implementing, a clear framework for communications with potential technology and service vendors is needed to support robust and accurate communication.

NOTE 1 "Big data system" means a system that leverages big data engineering and employs a big data paradigm to process big data.

NOTE 2 "Big data engineering" means advanced techniques that harness independent resources for building scalable data systems when the characteristics of the datasets require new architectures for efficient storage, manipulation, and analysis.

NOTE 3 "Big data paradigm" means distribution of data systems across horizontally coupled, independent resources to achieve the scalability needed for the efficient processing of extensive datasets.

While the potential value for analyzing big data is what attracts organizations to implementation of big data systems, these organizations need to understand the potential issues and liabilities associated with managing and controlling this data. IDC estimates that enterprises have liability or responsibility for nearly 80 % of the information in the digital universe and should be prepared to deal with issues of compliance, copyright and privacy. IDC further predicts that, by 2020, over 40 % of the information in the digital universe will require explicit protection and the amount of this data is growing faster than the total digital universe[15]. These risks mean that organizations should both be able to identify, define and articulate the policies for data security, provenance, and governance as well as implementing and documenting the technical controls to enforce those policies in order to protect the organization as a whole from liability for compromise or misuse of the data they control.

Finally, very few organizations dealing with big data operate solely on data organic to that organization. This means that systems that collect and analyze big data need to be able to securely and reliably interoperate and share data. In fact, the sheer volume associated with big data frequently makes it impractical to transfer between systems necessitating that, in many cases, the analytics need to be moved to the data requiring not just interoperability at the data level but at the software and application level between systems.

The existing big data landscape, market requirements for big data standardization were examined and the standardization priorities below were identified:

a) big data use cases, definitions, vocabulary and reference architectures (e.g. system, data, platforms, online/offline, etc.);

b) specifications and standardization of metadata including data provenance;

c) application models (e.g. batch, streaming, etc.);

d) query languages including non-relational queries to support diverse data types (XML, RDF, JSON, multimedia, etc.) and big data operations (e.g. matrix operations);

e) domain-specific languages;

f) semantics of eventual consistency;

g) advanced network protocols for efficient data transfer;

h) general and domain specific ontologies and taxonomies for describing data semantics including interoperation between ontologies;

i) big data security and privacy access controls;

j) remote, distributed, and federated analytics (taking the analytics to the data) including data and processing resource discovery and data mining;

k) data sharing and exchange;

l) data storage, e.g. memory storage system, distributed file system, data warehouse, etc.;

m) human consumption of the results of big data analysis (e.g. visualization);

n) energy measurement for big data;

o) interface between relational (SQL) and non-relational (NoSQL) data stores;

p) big data quality and veracity description and management[13].

ISO/IEC 20546 and the ISO/IEC 20547 series were developed with the intention to address those gaps.

This document specifically addresses framework and application process, big data use cases and requirements [gap a) above], reference architectures [gap a) above], and security and privacy [gap i) above], and standards roadmap. In addition, organizations with big data analytic requirements cannot wait for big data specific standards before they can implement their systems. Because big data is essentially a subset of all data, and almost every information technology standard deals with data in some respect, there are a large number of standards in place or underdevelopment today that address a number of big data issues. To address this need, ISO/IEC 20547-5 is a standards roadmap that aligns existing standards to the roles within the reference architecture to provide big data system stakeholders some guidance on how they can apply those standards to their problems today. Clause 7 describes each of the other parts in this series.

7 Conceptual foundations

7.1 General

The ISO/IEC 20547 series is designed to provide a foundation to a range of stakeholders in a given system to effectively and unambiguously describe and communicate about the characteristics and attributes of a given big data system. Based on the definitions provided in ISO/IEC 20546 for big data, a big data system is a system that:

— processes extensive data sets — primarily in the characteristics of volume, variety, velocity, and/or variability — that require a scalable architecture for efficient storage, manipulation, and analysis;

— leverages advanced techniques that harness independent resources for building scalable data systems when the characteristics of the datasets require new architectures for efficient storage, manipulation, and analysis;

— employs a paradigm where distribution of data systems across horizontally coupled and independent resources to achieve the scalability needed for the efficient processing of extensive datasets.

The broad and unconstrained nature of big data systems necessitates that the reference architecture provided in the ISO/IEC 20547 series be sufficient to represent the wide range of potential use cases implemented by big data systems.

7.2 Reference architecture concepts

In order to understand what a reference architecture covers, it is necessary to first define what a reference architecture means. Since it is an architecture, the reference architecture necessarily possesses all the characteristics of an architecture as defined by ISO/IEC/IEEE 42010 (see 3.2). The big data reference architecture should also be generalized enough to cover the variety of potential big data systems architectures.

Examined from an object-oriented view point, the reference architecture would be considered the abstract class from which specific instances of architectures derive their structure and attributes.

ISO/TR 14639-2 defines a reference architecture as in the field of software architecture or enterprise architecture, provides a proven template solution for an architecture for a particular domain, as well as a common vocabulary with which to discuss implementations, often with the aim of stressing commonality.

Based on this reasoning, a reference architecture is an architectural framework as defined by ISO/IEC/IEEE 42010, including the structure, rules and constraints common to all big data systems. Thus, a big data reference architecture provides a series of conventions, principles and practices for describing big data system architectures.

Reference architectures are developed to meet a wide variety of objectives as shown in Figure 1 taken from Reference [14], which goes on to describes that the core purpose of a reference architecture is to be forward looking and should be used (referenced) as the basis for future implementations.

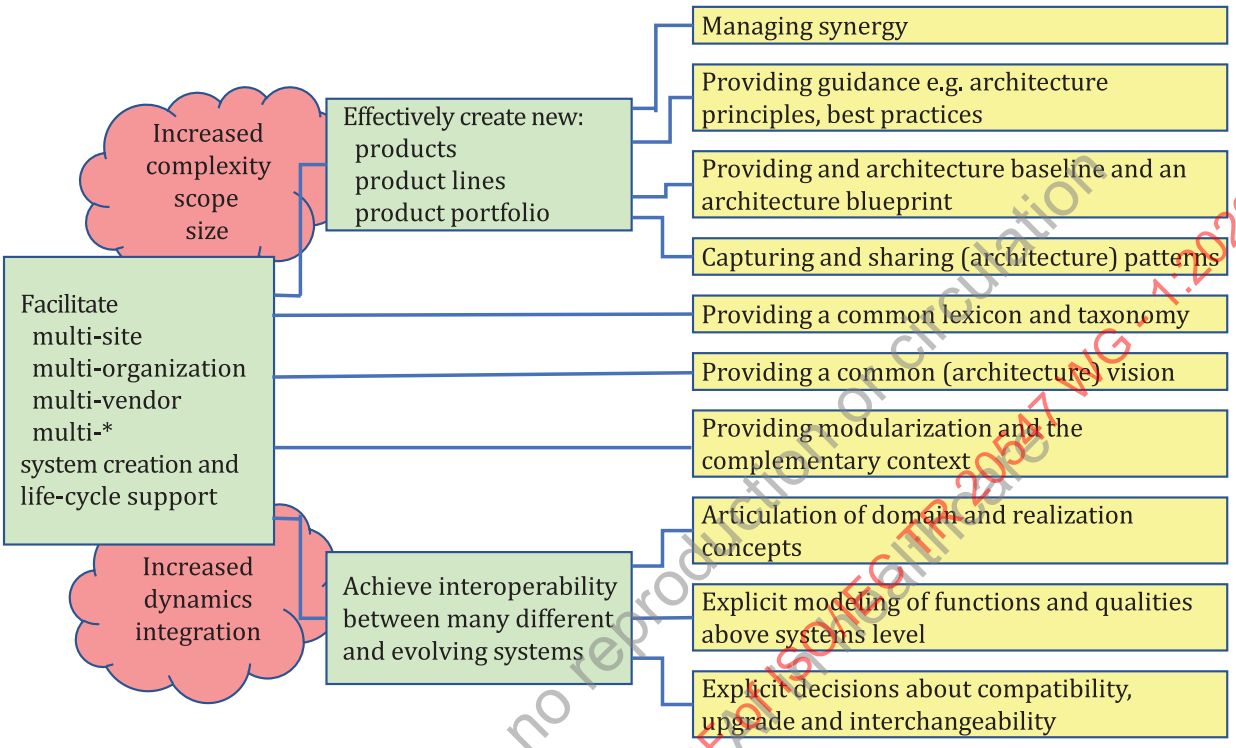

Figure 1 — The concept of reference architectures

7.3 Reference architecture structure

Figure 2 combines concepts and structures from ISO/IEC/IEEE 42010 to depict the outline for a reference architecture structure.

A reference architecture is defined for a domain. for this reference architecture, the domain is big data.

The domain in turn defines the environment, in the case of big data the environment is primarily defined by the core characteristics of big data — volume, velocity, variety, variability (see ISO/IEC 20546).

The stakeholders in this environment includes all the common stakeholders (users, owners, architects, etc.) for any system along with anyone having a concern related to the data and its characteristics.

The environment bounds the concerns. Since the environment is defined by the big data characteristics the concerns are bound by those characteristics and each concern should relate to one or more of those characteristics along with the stakeholder(s) which have that concern.

The reference architecture is described using an architecture framework. This framework is described in ISO/IEC 20547-3 and is presented in terms of two view points:

— roles and activities — user view;

— functional components — functional view.

Each of these viewpoints in turn addresses one or more concerns.

Those concerns can be addressed by one or more roles, activities and functional components within those architecture views.

EXAMPLE In a credit monitoring system, everyone with a record in that system is a stakeholder. Most have a concern that their privacy is maintained, and the security of their personal information is protected.

Within both the user and functional views of the big data reference architecture, there is a security and privacy cross-cutting aspect. This cross-cutting aspect has a relation with the activity "perform audit" and functional component "audit framework" which addresses that concern.

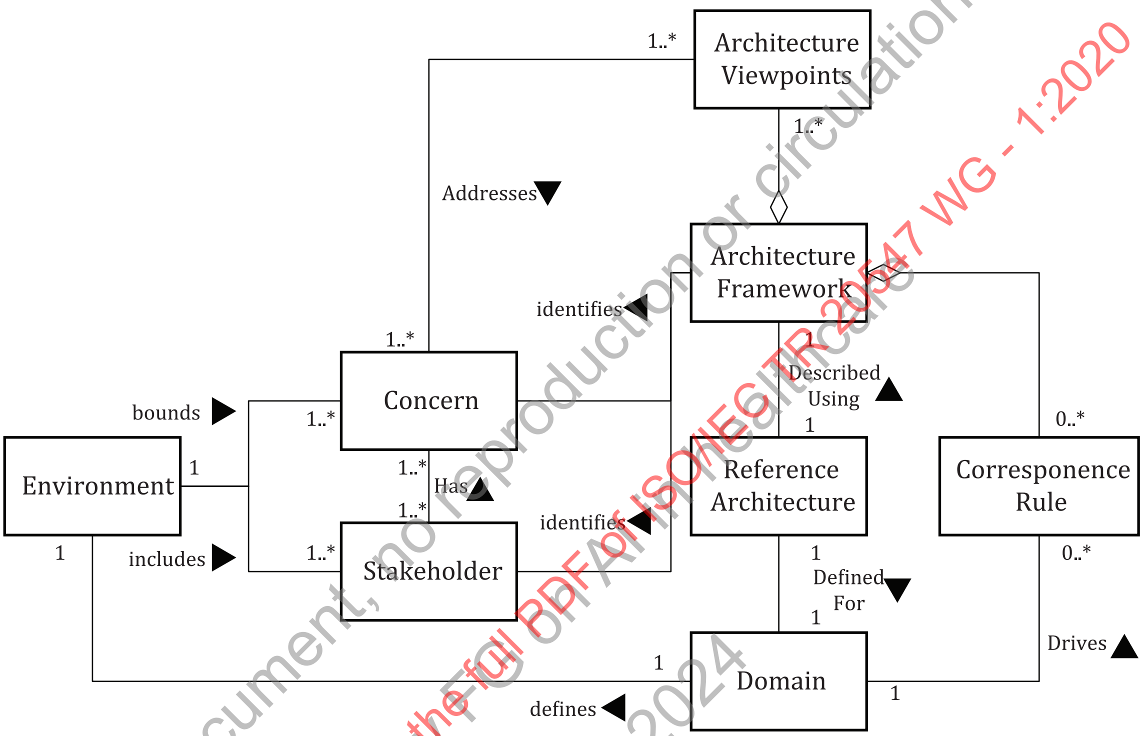

Figure 2 — Basic outline of a reference architecture structure based on ISO/IEC/IEEE 42010

8 Big data reference architecture elements

8.1 Overview

For the big data system environment, this document provides BDRA structure by providing the scope of each part of the ISO/IEC 20547 series, logical relationships of each document, and application process of BDRA. The scope of each document in the ISO/IEC 20547 series is as follows:

— ISO/IEC 20547-1: Framework and application process describes the framework of the big data reference architecture and the process for how a user of the standard can apply it to their particular problem domain;

— ISO/IEC 20547-2: Use cases and derived requirements provides examples of big data use cases with application domains and technical considerations derived from the contributed use cases;

— ISO/IEC 20547-3: Reference architecture specifies the big data reference architecture (BDRA). The reference architecture includes concepts and architectural views (user view and functional view);

— ISO/IEC 20547-4: Security and privacy specifies the security and privacy aspects applicable to the Big Data Reference Architecture (BDRA) including the big data roles, activities, and functional components, and also provides guidance on security and privacy operations for big data;

— ISO/IEC 20547-5: Standards roadmap describes big data relevant standards, both in existence and under development, along with priorities for future big data standards development based on gap analysis.

Figure 3 shows the relationships and iteration cycle of each part of the ISO/IEC 20547 series. Based on the contributions from enterprises, organizations and experts of the related research and academia related, ISO/IEC TR 20547-2 collects use cases and derives technical considerations. ISO/IEC 20547-3 defines reference architecture for big data by reflecting these technical considerations. ISO/IEC 20547-4, in particular, specifies the security and privacy aspect to support big data. ISO/IEC 20547-5 provides the applicable list of standards at the BDRA perspective. The ISO/IEC 20547 series represents a point-in-time view of big data systems and architectures. As big data implementations are created and evolve based on ISO/IEC 20547-1, ISO/IEC TR 20547-2, ISO/IEC 20547-3 and ISO/IEC 20547-4, they will reference and make use of the standards documented in ISO/IEC 20547-5. Those new systems can be documented in ISO/IEC TR 20547-2 as new use cases leading to new technical considerations. The technical considerations introduced by those use cases can lead to new standardization activities resulting in new standards to be documented in ISO/IEC 20547-5.

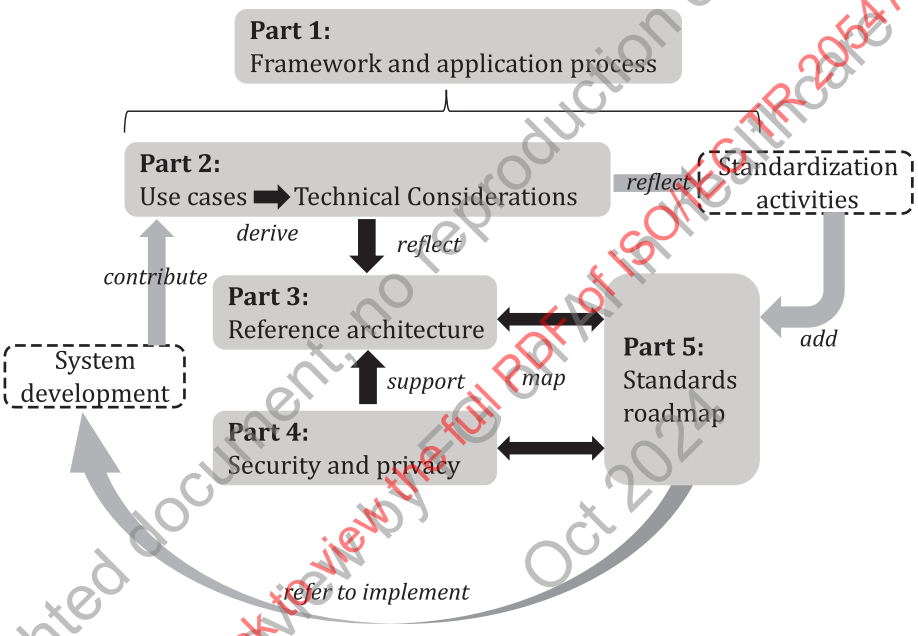

Figure 3 — Relationships between the parts of the ISO/IEC 20547 series

In order to apply this framework to a specific use case, it is necessary to understand the overall environment in which the big data system will be implemented, who the stakeholders are for that system, and what are the concerns of those stakeholders. Subclauses 8.2 to 8.4 describe each of these key components.

8.2 Stakeholders

ISO/IEC/IEEE 42010 defines a stakeholder as any individual, team, organization or classes thereof, having an interest in a system. Common stakeholders include the system owners, customers, system implementors and others. In the case of big data systems, the stakeholders also include anyone with an interest in the data being processed by the system. This includes the data owners who can be providing data to the big data system, the data consumers who are making decisions based on the data coming from the big data system, and also those people or organizations who can be described by the data. Identification of the stakeholders and their concerns is the first step in developing a big data architecture. ISO/IEC 20547-3 refers to the stakeholders of a big data system as parties within the user view.

8.3 Concerns

Any interest in a system relevant to one or more of its stakeholders is a concern. A concern pertains to any aspect of the big data system to include technical, business, operational, legal and even social influences on a system in its environment as described in ISO/IEC/IEEE 42010. The environment of a system is defined and bounded by the stakeholders and their concerns with that system. Some concerns can be codified in terms of other international standards.

NOTE 1 For example, the distribution transparencies described in the Reference Model of Open Distributed Processing (ISO/IEC 10746-1) are concerns that would be relevant to the operation of a big data system since horizontal scaling and distributed processing are core aspects of big data systems.

NOTE 2 Software properties as described in SQuaRE (ISO/IEC 25010:2011, 4.2) name concerns about the quality of the software in a big data system to include effectiveness, efficiency, trust, risk and risk mitigation, flexibility.

When dealing with big data, additional concerns evolve from the big data characteristics of volume, velocity and variety.

NOTE 3 For example, a concern can relate to potential data loss due to the velocity of the data.

In addition, there are a number of concerns related to the data itself to include provenance, pedigree and protection. Concerns related to security and privacy with big data are significant enough that ISO/IEC 20547-4 focuses directly on those aspects. For example, the ability to fuse multiple sources of data using big data technology to deanonymize data is a specific privacy concern.

The concerns identified for a big data system in turn drive the activities of the system and the functional components that implement those activities.

8.4 Views

As described above a big data system architecture can be defined in terms of views. The BDRA defines two primary views:

— user view, which describes the roles, sub-roles, activities, and cross-cutting aspects necessary to meet the concerns of the stakeholders;

— functional view, which describes the functional layers, functional components, and multi-layer functions necessary to implement the activities and cross-cutting aspects defined in the user view.

Figure 4 illustrates the relationship of stakeholders, and concerns to these views and their elements.

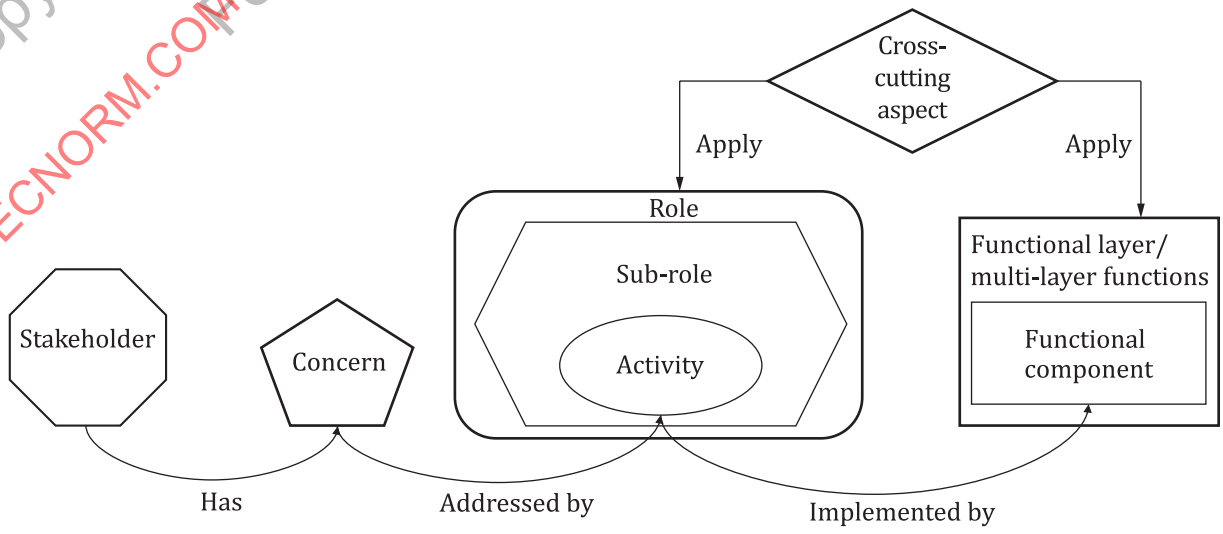

Figure 4 — Relationships between the elements of the BDRA views

8.4.1 User view

The user view addresses the ecosystem of big data with following concepts:

— parties: a party is a natural person or legal person, whether or not incorporated, or a group of either. Parties in a big data ecosystem are its stakeholders;

— roles and sub-roles: a role is a set of big data activities that serves a common purpose. A sub-role is a subset of the big data activities for a given role, and different sub-roles can share the big data activities associated with a given role;

— activities: an activity is defined as a specified pursuit or set of tasks. Big data activities need to have a purpose and deliver one or more outcomes and these are conducted using functional components;

— cross-cutting aspects: cross-cutting aspects are behaviours or capabilities which need to be coordinated across roles and implemented consistently in a big data ecosystem. Cross-cutting aspects can be shared and can impact multiple roles, big data activities and functional components. Cross-cutting aspects apply to multiple individual roles or functional components.

NOTE 1 An example of a cross-cutting aspect is security.

NOTE 2 A party can assume more than one role at any given point in time and can engage in a specific subset of activities of that role. Examples of parties include, but are not limited to, large corporations, small and medium sized enterprises, government departments, academic institutions and private citizens.

8.4.2 Functional view

The functional view is a technology-neutral view of the functions necessary to form a big data system. It describes the distribution of functions necessary for the support of big data activities.

The functional architecture also defines the dependencies between functions.

The functional view addresses the following big data concepts:

— functional components: a functional component is a functional building block needed to engage in an activity, backed by an implementation;

— functional layers: a functional layer is a set of functional components that provide similar capabilities or serve a common purpose. The functional architecture is partially layered (i.e. has layers and a set of multi-layer functions);

— multi-layer functions: the multi-layer functions include functional components that provide capabilities that are used across multiple functional layers, and they are grouped into subsets.

NOTE Not all layers or functional components are necessarily instantiated in a specific big data system.

9 Big data reference architecture application process

9.1 Overview

This clause provides the reader with a stepwise process for applying the reference architecture to develop an architecture description for a given big data system implementation. While the BDRA is extremely general and designed to apply to a wide range of systems, due to the broad variety of potential big data systems and components that can comprise them, the process is designed to support extension of the BDRA to meet unique requirements of the given system. The primary nature of these extensions is the identification of additional activities related to roles and/or the assignment of activities to different roles/sub-roles. Throughout this process, the reader is advised to leverage other relevant ISO standards including ISO/IEC/IEEE 15288 for systems engineering, ISO/IEC/IEEE 12207 for software lifecycle engineering, the quality measurement processes defined with in the ISO 9000

family of standards, and associated standards in order to verify that the resulting architecture does in fact cover and address the full range of concerns.

Before undertaking this process, the architect is urged to define what tools would be used to capture and manage the data generated.

9.2 Identify stakeholders and concerns

The first step in this process is to identify the stakeholders and concerns associated with the big data system. This step is the first part of the stakeholder requirements analysis process described in ISO/IEC/IEEE 15288. The outcomes of this step include:

a) the required characteristics and context of use of system services are specified;

b) the constraints on a conformant system are defined;

c) traceability of stakeholder requirements to stakeholder needs is achieved;

d) the basis for establishing the system requirements is described;

e) the basis for validating the conformance of the system services is defined;

f) a basis for negotiating and agreeing to supply a system service or product is provided.

As described in ISO/IEC/IEEE 15288 stakeholder concerns (requirements) are expressed in terms of the needs, wants, desires, expectations and perceived constraints of identified stakeholders. They are expressed in terms of a model that can be textual or formal, that concentrates on system purpose and behaviour, and that is described in the context of the operational environment and conditions.

Stakeholder concerns should include the needs and requirements imposed by society (e.g. privacy expectation) and take into consideration government regulations (e.g. GDPR in the European Union).

The stakeholders and concerns need to be captured in a model that can later be used to provide traceability from the system activities and components in order to support verification of the process.

Specifically, stakeholders and concerns each need to be uniquely identified to support the traceability in future steps. Where feasible, concerns shared by stakeholders should be collapsed into a single concern and mapped to each stakeholder with that concern.

A review of the use cases and derived requirements from ISO/IEC TR 20547-2 in this step can aid the architect in identifying those stakeholders and concerns from other big data use cases that can be applicable to this use case.

As with any good requirements and architectural process, the outputs of this process should be reviewed with stakeholders to confirm the accuracy of what was captured.

9.3 Map stakeholders and concerns to roles and subroles

The goal of this step is to organize the stakeholders and their concerns into a common framework of reference for the big data system. This step is especially critical for big data systems in that, in many cases, the concerns should be addressed through a system-of-systems solution (e.g. multiple systems coordinating together to meet a requirement).

ISO/IEC 20547-3 defines the following roles and sub-roles:

— big data application provider (BDAP)

  — big data collection provider (BDCP)

  — big data preparation provider (BDPreP)

  — big data analytics provider (BDAnP)