

---

---

**Information technology — Big data  
reference architecture —**

**Part 1:  
Framework and application process**

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

**iTeh STANDARD PREVIEW**  
*Partie 1: Cadre méthodologique et processus d'application*  
**(standards.iteh.ai)**

[ISO/IEC TR 20547-1:2020](https://standards.iteh.ai/catalog/standards/sist/65b6ad7a-d32a-4400-84b5-0fd4b4c0c1b5/iso-iec-tr-20547-1-2020)

<https://standards.iteh.ai/catalog/standards/sist/65b6ad7a-d32a-4400-84b5-0fd4b4c0c1b5/iso-iec-tr-20547-1-2020>



**iTeh STANDARD PREVIEW**  
**(standards.iteh.ai)**

ISO/IEC TR 20547-1:2020  
<https://standards.iteh.ai/catalog/standards/sist/65b6ad7a-d32a-4400-84b5-0fddb4c0c1b5/iso-iec-tr-20547-1-2020>



**COPYRIGHT PROTECTED DOCUMENT**

© ISO/IEC 2020

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office  
CP 401 • Ch. de Blandonnet 8  
CH-1214 Vernier; Geneva  
Phone: +41 22 749 01 11  
Email: [copyright@iso.org](mailto:copyright@iso.org)  
Website: [www.iso.org](http://www.iso.org)

Published in Switzerland

# Contents

	Page
Foreword.....	iv
Introduction.....	v
<b>1 Scope.....</b>	<b>1</b>
<b>2 Normative references.....</b>	<b>1</b>
<b>3 Terms and definitions.....</b>	<b>1</b>
<b>4 Abbreviated terms.....</b>	<b>2</b>
<b>5 Document overview.....</b>	<b>3</b>
<b>6 Big data standardization: motivation and objectives.....</b>	<b>3</b>
<b>7 Conceptual foundations.....</b>	<b>5</b>
7.1 General.....	5
7.2 Reference architecture concepts.....	5
7.3 Reference architecture structure.....	6
<b>8 Big data reference architecture elements.....</b>	<b>7</b>
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
<b>9 Big data reference architecture application process.....</b>	<b>10</b>
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
<b>Bibliography.....</b>	<b>14</b>

## 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](http://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](http://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](http://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](http://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.

## iTeh STANDARD PREVIEW (standards.iteh.ai)

[ISO/IEC TR 20547-1:2020](https://standards.iteh.ai/catalog/standards/sist/65b6ad7a-d32a-4400-84b5-0fd4b4c0c1b5/iso-iec-tr-20547-1-2020)

<https://standards.iteh.ai/catalog/standards/sist/65b6ad7a-d32a-4400-84b5-0fd4b4c0c1b5/iso-iec-tr-20547-1-2020>

**iTeh STANDARD PREVIEW**  
**(standards.iteh.ai)**

ISO/IEC TR 20547-1:2020

<https://standards.iteh.ai/catalog/standards/sist/65b6ad7a-d32a-4400-84b5-0fd4b4c0c1b5/iso-iec-tr-20547-1-2020>

# 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 (standards.iteh.ai)

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.

STANDARD PREVIEW  
(standards.iteh.ai)

[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

ISO/IEC TR 20547-1:2020  
<https://standards.iteh.ai/catalog/standards/sist/65b6ad7a-d32a-4400-84b5-0f1db4c0c1b5/iso-iec-tr-20547-1-2020>

[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



BDSB	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.

**iTeh STANDARD PREVIEW**

Clauses 6 to 9:

**(standards.iteh.ai)**

- 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<sup>[15]</sup>.

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<sup>[15]</sup>. 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<sup>[13]</sup>.

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.