TECHNICAL REPORT

ISO/IEC TR 20547-2

First edition
2018-01

Information technology — Big data reference architecture —

Part 2: Use cases and derived requirements

Technologies de l'information — Architecture de référence des big data —

Partie 2: Cas pratiques et exigences dérivées

Reference number
ISO/IEC TR 20547-2:2018(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

Published in Switzerland

Contents

Page

Foreword ..... **v**

Introduction ..... **vi**

1 Scope **1**

2 Normative references **1**

3 Terms and definitions **1**
3.1 Terms defined elsewhere 1
3.2 Terms defined in this document 1
3.3 Abbreviated terms 1

4 Use case properties for survey **6**
4.1 Overall description 6
4.2 Current solution 7
4.3 Big data characteristics 7
4.4 Big data science 7
4.5 Overall big data issues 8
4.6 Big data use case Template 8

5 Use cases summaries **9**
5.1 Use case development process 9
5.2 Government operation 10
5.2.1 Use case 1: Census 2010 and 2000 — Title 13 big data 10
5.2.2 Use case 2: NARA Accession, Search, Retrieve, Preservation 10
5.2.3 Use case 3: Statistical survey response improvement 11
5.2.4 Use case 4: Non-Traditional Data in Statistical Survey Response Improvement (Adaptive Design) 11
5.3 Commercial 12
5.3.1 Use case 5: Cloud Eco-System for Financial Industries 12
5.3.2 Use case 6: Mendeley — An International Network of Research 12
5.3.3 Use case 7: Multi-media streaming service 13
5.3.4 Use case 8: Web Search 13
5.3.5 Use case 9: Big data Business Continuity and Disaster Recovery Within a Cloud Eco-System 14
5.3.6 Use case 10: Cargo Shipping 14
5.3.7 Use case 11: Materials Data for Manufacturing 14
5.3.8 Use case 12: Simulation-Driven Materials Genomics 15
5.4 Defense 16
5.4.1 Use case 13: Cloud Large-Scale Geospatial Analysis and Visualization 16
5.4.2 Use case 14: Object Identification and Tracking from Wide-Area Large Format Imagery or Full Motion Video—Persistent Surveillance 16
5.4.3 Use case 15: Intelligence Data Processing and Analysis 17
5.5 Health care and life sciences 17
5.5.1 Use case 16: Electronic Medical Record Data 17
5.5.2 Use case 17: Pathology Imaging/Digital Pathology 18
5.5.3 Use case 18: Computational Bioimaging 18
5.5.4 Use case 19: Genomic Measurements 19
5.5.5 Use case 20: Comparative Analysis for Metagenomes and Genomes 19
5.5.6 Use case 21: Individualized Diabetes Management 19
5.5.7 Use case 22: Statistical Relational Artificial Intelligence for Health Care 20
5.5.8 Use case 23: World Population-Scale Epidemiological Study 20
5.5.9 Use case 24: Social Contagion Modeling for Planning, Public Health, and Disaster Management 21
5.5.10 Use case 25: Biodiversity and LifeWatch 21
5.6 Deep Learning and Social Media 22
5.6.1 Use case 26: Large-Scale Deep Learning 22

5.6.2 Use case 27: Organizing Large-Scale, Unstructured Collections of Consumer Photos 22
5.6.3 Use case 28: Truthy—Information Diffusion Research from Twitter Data 23
5.6.4 Use case 29: Crowd Sourcing in the Humanities as Source for Big and Dynamic Data 23
5.6.5 Use case 30: CINET—Cyberinfrastructure for Network (Graph) Science and Analytics 23
5.6.6 Use case 31: NIST Information Access Division — Analytic Technology Performance Measurements, Evaluations, and Standards 24
5.7 The Ecosystem for research 24
5.7.1 Use case 32: DataNet Federation Consortium 24
5.7.2 Use case 33: The Discinnet Process 25
5.7.3 Use case 34: Semantic Graph Search on Scientific Chemical and Text-Based Data 25
5.7.4 Use case 35: Light Source Beamlines 26
5.8 Astronomy and physics 26
5.8.1 Use case 36: Catalina Real-Time Transient Survey: A Digital, Panoramic, Synoptic Sky Survey 26
5.8.2 Use case 37: DOE Extreme Data from Cosmological Sky Survey and Simulations 27
5.8.3 Use case 38: Large Survey Data for Cosmology 27
5.8.4 Use case 39: Particle Physics—Analysis of Large Hadron Collider Data: Discovery of Higgs Particle 28
5.8.5 Use case 40: Belle II High Energy Physics Experiment 29
5.9 Earth, environmental and polar science 29
5.9.1 Use case 41: European Incoherent Scatter Scientific Association 3D Incoherent Scatter Radar System 29
5.9.2 Use case 42: Common Operations of Environmental Research Infrastructure 30
5.9.3 Use case 43: Radar Data Analysis for the Center for Remote Sensing of Ice Sheets 31
5.9.4 Use case 44: Unmanned Air Vehicle Synthetic Aperture Radar (UAVSAR) Data Processing, Data Product Delivery, and Data Services 31
5.9.5 Use case 45: NASA Langley Research Center/ Goddard Space Flight Center iRODS Federation Test Bed 32
5.9.6 Use case 46: MERRA Analytic Services (MERRA/AS) 32
5.9.7 Use case 47: Atmospheric Turbulence – Event Discovery and Predictive Analytics 32
5.9.8 Use case 48: Climate Studies Using the Community Earth System Model at the U.S. Department of Energy (DOE) NERSC Center 33
5.9.9 Use case 49: DOE Biological and Environmental Research (BER) Subsurface Biogeochemistry Scientific Focus Area 33
5.9.10 Use case 50: DOE BER AmeriFlux and FLUXNET Networks 34
5.10 Energy 34
5.10.1 Use case 51: Consumption Forecasting in Smart Grids 34
5.10.2 Use case 52: Home Energy Management System 34

6 Use cases derived technical considerations 35
6.1 Use case specific technical considerations 35
6.2 Summary of requirements analysis 35
6.3 Features of use cases 37

Annex A Submitted use case studies 40

Annex B Summary of Key Properties 197

Annex C Use case technical considerations summary 207

Annex D Use case detail technical considerations 225

Bibliography 252

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. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

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).

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

For an explanation on 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 the following URL: www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/IEC JTC 1, *Information Technology*.

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

Introduction

This document is focuses on forming a community of interest from industry, academia, and government, with the goal of developing a consensus list of big data technical considerations across all stakeholders. This included gathering and understanding various examples of use cases from diversified areas (i.e., application domains). To achieve this goal, the following tasks were done:

— gathered input from all stakeholders regarding big data technical considerations;

— analyzed and prioritized a list of challenging use case specific technical considerations that may delay or prevent adoption of big data deployment;

— developed a comprehensive list of generalized big data technical considerations for ISO/IEC 20547-3, *Information technology – Big data reference architecture - Part 3: Reference architecture*; and

— documented the findings in this document.

Information technology — Big data reference architecture —

Part 2: Use cases and derived requirements

1 Scope

This document provides examples of big data use cases with application domains and technical considerations derived from the contributed use cases.

2 Normative references

The following documents, in whole or in part, are normatively referenced in this document and are indispensable for its application. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 20546*Information technology — Big data — Definition and vocabulary*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 20546 and the following apply.

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

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

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

3.1 Terms defined elsewhere

None.

3.2 Terms defined in this document

3.2.1
use case
typical application stated at a high level for the purposes of extracting technical considerations or comparing usages across fields

3.3 Abbreviated terms

2D	two-Dimensional
3D	three-Dimensional
6D	six-Dimensional
AOD	Analysis Object Data

API	Application Programming Interface
ASDC	Atmospheric Science Data Center
ASTM	American Society for Testing and Materials
AWS	Amazon Web Services
BC/DR	Business Continuity and Disaster Recovery
BD	Big data
BER	Biological and Environmental Research
BNL	Brookhaven National Laboratory
CAaaS	Climate Analytics as a Service
CADRG	Compressed ARC Digitized Raster Graphics
CBSP	Cloud Brokerage Service Provider
CERES	Clouds and Earth's Radiant Energy System
CERN	European Organization for Nuclear Research
CESM	Community Earth System Model
CFTC	U.S. Commodity Futures Trading Commission
CIA	Confidentiality, Integrity, and Availability
CINET	Cyberinfrastructure for Network (Graph) Science and Analytics
CMIP	Coupled Model Intercomparison Project
CMIP5	Climate Model Intercomparison Project
CMS	Compact Muon Solenoid
COSO	Committee of Sponsoring Organizations
CPU	Central Processing Unit
CReSIS	Center for Remote Sensing of Ice Sheets
CRTS	Catalina Real-Time Transient Survey
CSP	Cloud Service Provider
CSS	Catalina Sky Survey proper
CV	Controlled Vocabulary
DFC	DataNet Federation Consortium
DHTC	Distributed High Throughput Computing
DNA	DeoxyriboNucleic Acid
DOE	U.S. Department of Energy

DOJ	U.S. Department of Justice
DPO	Data Products Online
EBAF–TOA	Energy Balanced and Filled–Top of Atmosphere
EC2	Elastic Compute Cloud
EDT	Enterprise Data Trust
EHR	Electronic Health Record
EMR	Electronic Medical Record
EMSO	European Multidisciplinary Seafloor and Water Column Observatory
ENVRI	Common Operations of Environmental Research Infrastructures
ENVRI RM	ENVRI Reference Model
EPOS	European Plate Observing System
ESFRI	European Strategy Forum on Research Infrastructures
ESG	Earth System Grid
ESGF	Earth System Grid Federation
FDIC	U.S. Federal Deposit Insurance Corporation
FI	Financial Industries
FLUXNET	Flux Tower Network
FMV	Full Motion Video
FNAL	Fermi National Accelerator Laboratory
GAAP	U.S. Generally Accepted Accounting Principles
GB	Giga Byte
GCM	General Circulation Model
GEOS-5	Goddard Earth Observing System version 5
GeoTiff	Geo Tagged Image File Format
GEWaSC	Genome-Enabled Watershed Simulation Capability
GHG	Green House Gas
GMAO	Global Modeling and Assimilation Office
GPFS	General Parallel File System
GPS	Global Positioning System
GPU	Graphics Processing Unit
GRC	Governance, Risk management, and Compliance

GSFC	Goddard Space Flight Center
HDF5	Hierarchical Data Format
HDFS	Hadoop Distributed File System
HPC	High-Performance Computing
HTC	High-Throughput Computing
HVS	Hosted Virtual Server
I/O	Input Output
IaaS	Infrastructure as a Service
IAGOS	In-service Aircraft for a Global Observing System
ICD	International Classification of Diseases
ICOS	Integrated Carbon Observation System
IMG	Integrated Microbial Genomes
INPC	Indiana Network for Patient Care
IPCC	Intergovernmental Panel on Climate Change
iRODS	Integrated Rule-Oriented Data System
ISACA	International Society of Auditors and Computer Analysts
isc2	International Security Computer and Systems Auditors
ISO	International Organization for Standardization
ITIL	Information Technology Infrastructure Library
JGI	Joint Genome Institute
KML	Keyhole Markup Language
kWh	kilowatt-hour
LaRC	Langley Research Center
LBNL	Lawrence Berkeley National Laboratory
LDA	latent Dirichlet allocation
LHC	Large Hadron Collider
LPL	Lunar and Planetary Laboratory
LSST	Large Synoptic Survey Telescope
MERRA	Modern Era Retrospective Analysis for Research and Applications
MERRA/AS	MERRA Analytic Services
MPI	Message Passing Interface

MRI	Magnetic Resonance Imaging
NARA	National Archives and Records Administration
NARR	North American Regional Reanalysis
NaaS	Network as a Service
NASA	National Aeronautics and Space Administration
NCAR	National Center for Atmospheric Research
NCBI	National Center for Biotechnology Information
NCCS	NASA Center for Climate Simulation
NERSC	National Energy Research Scientific Computing Center
NetCDF	Network Common Data Form
NEX	NASA Earth Exchange
NFS	Network File System
NIKE	NIST Integrated Knowledge Editorial Net
NIST	National Institute of Standards and Technology
NITF	National Imagery Transmission Format
NLP	Natural Language Processing
NRT	Near Real Time
NSF	National Science Foundation
ODP	Open Distributed Processing
OGC	Open Geospatial Consortium
PB	PetaByte
PCA	Principal Component Analysis
PCAOB	Public Company Accounting and Oversight Board
PID	persistent identification
PII	Personally Identifiable Information
PNNL	Pacific Northwest National Laboratory
RDBMS	relational database management system
RDF	Resource Description Framework
RECOVER	Rehabilitation Capability Convergence for Ecosystem Recovery
ROI	return on investment
RPI	Repeat Pass Interferometry

RPO	Recovery Point Objective
RTO	Response Time Objective
SAN	storage area network
SAR	Synthetic Aperture Radar
SDN	software-defined networking
SIOS	Svalbard Integrated Arctic Earth Observing System
SPADE	Support for Provenance Auditing in Distributed Environments
SSH	Secure Shell
SSO	Single Sign-On
TB	TeraByte
tf-idf	term frequency–inverse document frequency
UA	University of Arizona
UAVSAR	Unmanned Air Vehicle Synthetic Aperture Radar
UC	Use Case
UI	User Interface
UPS	United Parcel Service
UQ	Uncertainty Quantification
VASP	Vienna Ab initio Simulation Package
vCDS	virtual Climate Data Server
VO	Virtual Observatory
VOIP	Voice over IP
WALF	Wide Area Large Format Imagery
WLCG	Worldwide LHC Computing Grid
XBRL	eXtensible Business Related Markup Language
XML	Extensible Markup Language
ZTF	Zwicky Transient Factory

4 Use case properties for survey

4.1 Overall description

— **Use case title**: Title provided by the use case author

— **Vertical (area)**: Intended to categorize the use cases. However, an ontology was not created prior to the use case submissions so this field was not used in the use case compilation.

- **Author/company/email**: Name, company, and email (if provided) of the person(s) submitting the use case
- **Actors/ stakeholders and their roles and responsibilities**: Description of the players and their roles in the use case
- **Goals**: Objectives of the use case
- **Use case description**: Brief description of the use case

4.2 Current solution

Current solutions describe current approach to processing big data at the hardware and software infrastructure and analytics level.

- **Compute (System)**: Computing component of the data analysis system
- **Storage**: Storage component of the data analysis system
- **Networking**: Networking component of the data analysis system
- **Software**: Software component of the data analysis system

4.3 Big data characteristics

Big data Characteristics describe the properties of the (raw) data including the four major 'V's' of big data.

- **Data source**: The origin of data, which could be from instruments, Internet of Things, Web, Surveys, Commercial activity, or from simulations. The source(s) can be distributed, centralized, local, or remote.
- **Data destination**: If data transformed in use case, where the final results end up.
- **Volume**: The characteristic of datasets that is most associated with big data. Volume represents the extensive amount of data available for analysis to extract valuable information. The assumption that you can extract the most value by analysing as much of the volume of data as possible was one of the primary drivers for the creation of the new scaling technologies.
- **Velocity**: The rate of flow at which the data is created, stored, analysed, or visualized. Big data velocity means a large quantity of data needs to be processed in a short amount of time. Dealing with high velocity data is commonly referred to as techniques for streaming data.
- **Variety:** The need to analyse data from a number of domains and a number of data types. The variety of data was handled through transformations or pre-analytics to extract features that would allow integration with other data. The wider range of data formats, logical models, timescales, and semantics, which is desirous to be used in analytics, complicates the integration of the variety of data. Metadata is increasingly used to aid in the integration.
- **Variability**: Changes in data rate, format/structure, semantics, and/or quality that impact the supported application, analytic, or problem. Impacts can include the need to refactor architectures, interfaces, processing/algorithms, integration/fusion, storage, applicability, or use of the data.

4.4 Big data science

Big data science describes the high level aspects of the data analysis process.

- **Veracity and data quality**: This covers the completeness and accuracy of the data with respect to semantic content as well as syntactical quality of data (such as presence of missing fields or incorrect values).

— **Visualization**: Refers to the way data is viewed by an analyst making decisions based on the data. Typically, visualization is the final stage of a technical data analysis pipeline and follows the data analytics stage.

— **Data types**: Refers to the style of data such as structured, unstructured, images (e.g., pixels), text (e.g., characters), gene sequences, and numerical.

— **Metadata**: Comments on quality and richness of metadata.

— **Curation and governance**: Comment on process to ensure good data quality and who is responsible.

NOTE The use case template has a separate item to describe security and privacy issues.

— **Data analytics**: Refers broadly to tools and algorithms used in processing the data at any stage including the data to information or knowledge to wisdom stages, as well as the information to knowledge stage.

4.5 Overall big data issues

— **Other big data issues**: Did we miss something important that your use case highlights? Your chance to address questions which we should have asked.

— **User Interface and mobile access issues**: Refers to issues in accessing or generating big data from clients including smart phones and tablets.

— **List key features and related use cases**: Put use case in context of related use cases. What features generalize and what are idiosyncratic to this use case.

— **Project future**: How do you expect application, and approach (hardware, software, analytics) to change in future?

— **More project information (URLs)**: Put a collection of useful links.

4.6 Big data use case Template

This clause provides one blank use case template. The below blank use case template was used for the purpose of capturing use cases to derived technical consideration.

NOTE The terms used in this template may or may not match with ISO/IEC 20546 and other parts of the ISO/IEC 20547-series.

Use case title		
Vertical (area)		
Author/company/email		
Actors/stakeholders and their roles and responsibilities		
Goals		
Use case description		
Current solutions	**Compute(System)**	
	Storage	
	Networking	
	Software	

Big data characteristics	**Data source (distributed/ centralized)**	
	Volume (size)	
	Velocity (e.g. real time)	
	Variety (multiple datasets, mashup)	
	Variability (rate of change)	
Big data science (collection, curation, analysis, action)	**Veracity (Robustness Issues, semantics)**	
	Visualization	
	Data quality (syntax)	
	Data types	
	Data analytics	
Big data specific challenges (Gaps)		
Big data specific challenges in mobility		
Security and privacy technical considerations		
Highlight issues for generalizing this Use case (e.g. for ref. architecture)		
More information (URLs)		
NOTE <additional comments>		

5 Use cases summaries

5.1 Use case development process

A use case is a typical application stated at a high level for the purposes of extracting technical considerations or comparing usages across fields. In order to develop a consensus list of big data technical considerations across all stakeholders, publicly available information was collected for various big data architectures. After collection of use cases, application domains were identified to better organize the collection of use cases.

NOTE 1 The list of application domains reflects the use cases submitted and is not intended to be exhaustive.

The nine application domains were as follows:

— **Government operation** (4): National Archives and Records Administration, Census Bureau;

— **Commercial** (8): Finance in Cloud, Cloud Backup, Citations, Multi-media streaming, Web Search, Digital Materials, Cargo Shipping;

— **Defense** (3): Sensors, Image Surveillance, Situation Assessment ;

— **Healthcare and life sciences** (10): Medical Records, Graph and Probabilistic Analysis, Pathology, Bioimaging, Genomics, Epidemiology, People Activity Models, Biodiversity;

— **Deep learning and social media** (6) Self-driving cars, Geolocate Images, SNS, Crowd Sourcing, Network Science, Benchmark Datasets;