INTERNATIONAL STANDARD

ISO/IEC 20546

First edition
2019-02

Information technology — Big data — Overview and vocabulary

Technologies de l'information — Mégadonnées — Vue d'ensemble et vocabulaire

Reference number
ISO/IEC 20546:2019(E)

COPYRIGHT PROTECTED DOCUMENT

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

Published in Switzerland

Contents

Page

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 Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 42, *Artificial intelligence*.

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.htm

Introduction

The big data paradigm is a rapidly changing field with rapidly changing technologies.

The term big data implies datasets that are extensive in volume, velocity, variety and/or variability. The term does not, however, represent data that is simply larger than before, since this has happened on a regular basis for decades. The specific occurrence that has led to the widespread usage of the term big data is that in the mid-2000s, extensive datasets could no longer be handled using extant data systems. The big data techniques represented a shift at that time to use distributed data management and processing through horizontal scaling to achieve the needed performance efficiency at an affordable cost.

In the evolution of data processing systems, there have been a number of times when the need for efficient, cost-effective data analysis has forced a change in existing technologies. For example, the move to a relational model occurred when methods to reliably handle changes to structured data led in the 1980s to the shift to relational databases that modelled relational algebra. That was a fundamental shift in data handling. The revolution in technologies referred to as big data has arisen because the relational model could no longer efficiently handle all the needs for analysis of large and often unstructured datasets. It is not just that data is larger than before, as data has been steadily getting larger for decades. The big data revolution is instead a one-time fundamental shift in architecture towards parallelization, just as the shift to the relational model was a one-time shift. As relational databases evolved to greater efficiencies over decades, so too will big data technologies continue to evolve. Many of the conceptual underpinnings of big data have been around for years, but the years since the mid-2000s have seen an explosion in scaling technologies and their maturation and application to scaled data systems.

The term big data is overloaded in common usage and is used to represent a number of related concepts, in part because several distinct system dimensions are consistently interacting with each other. To understand this revolution, the interplay of the following aspects needs to be considered: the data and processing characteristics of the datasets, the analysis of the datasets, the performance of the systems that handle the data, the business considerations of cost effectiveness, and the new engineering and analysis techniques for distributed data processing using horizontal scaling.

Annex A provides an overview of several concepts from the broader computing domain which are cross-cutting with respect to big data.

Information technology — Big data — Overview and vocabulary

1 Scope

This document provides a set of terms and definitions needed to promote improved communication and understanding of this area. It provides a terminological foundation for big data-related standards.

This document provides a conceptual overview of the field of big data, its relationship to other technical areas and standards efforts, and the concepts ascribed to big data that are not new to big data.

2 Normative references

There are no normative references in this document.

3 Terms, definitions and abbreviated terms

For the purposes of this document, the following terms and definitions 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 Terms and definitions

3.1.1
benefit
advantage to the organization of the actionable knowledge derived from an analytic system

Note 1 to entry: Benefit is often ascribed to big data due to the understanding that data has potential value that was typically not considered previously.

3.1.2
big data
extensive *datasets* (3.1.11) — primarily in the *data* (3.1.5) characteristics of volume, variety, velocity, and/or variability — that require a scalable technology for efficient storage, manipulation, management, 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.

3.1.3
cloud computing
paradigm for enabling network access to a scalable and elastic pool of shareable physical or virtual resources with self-service provisioning and administration on-demand

Note 1 to entry: Examples of resources include servers, operating systems, networks, software, applications, and storage equipment.

[SOURCE: ISO/IEC 17788:2014, 3.2.5]

3.1.4
cluster
<distributed data processing> set of functional units under common control

[SOURCE: ISO/IEC 2382:2015, 2120586]

3.1.5
data
reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing

Note 1 to entry: Data can be processed by humans or by automatic means.

[SOURCE: ISO/IEC 2382:2015, 2121272]

3.1.6
data analytics
composite concept consisting of data acquisition, data collection, data validation, *data processing* (3.1.9), including data quantification, data visualization, and data interpretation

Note 1 to entry: Data analytics is used to understand objects represented by *data* (3.1.5), to make predictions for a given situation, and to recommend on steps to achieve objectives. The insights obtained from analytics are used for various purposes such as decision-making, research, sustainable development, design, planning, etc.

3.1.7
database
collection of *data* (3.1.5) organized according to a conceptual structure describing the characteristics of these *data* and the relationships among their corresponding entities, supporting one or more application areas

[SOURCE: ISO/IEC 2382:2015, 2121413]

3.1.8
data model
pattern of structuring *data* (3.1.5) in a *database* (3.1.7) according to the formal descriptions in its information system and according to the requirements of the database management system to be applied

[SOURCE: ISO/IEC 2382:2015, 2125519]

3.1.9
data processing
systematic performance of operations upon *data* (3.1.5)

Note 1 to entry: Example: Arithmetic or logic operations upon data, merging or sorting of data, or operations on text, such as editing, sorting, merging, storing, retrieving, displaying, or printing.

Note 2 to entry: The term data processing should not be used as a synonym for information processing.

[SOURCE: ISO/IEC 2382:2015, 01.01.06]

3.1.10
data science
extraction of actionable knowledge from *data* (3.1.5) through a process of discovery, or hypothesis and hypothesis testing

3.1.11
data set
dataset
identifiable collection of *data* (3.1.5) available for access or download in one or more formats

[SOURCE: Adapted from ISO 19115-2:2009, 4.7]

3.1.12
data type
datatype
defined set of *data* (3.1.5) objects of a specified data structure and a set of permissible operations, such that these *data* objects act as operands in the execution of any one of these operations

Note 1 to entry: Example: An integer type has a very simple structure, each occurrence of which, usually called value, is a representation of a member of a specified range of whole numbers and the permissible operations include the usual arithmetic operations on these integers.

Note 2 to entry: The term "type" may be used instead of "data type" when there is no ambiguity.

Note 3 to entry: Data type; datatype: terms and definition standardized by ISO/IEC [ISO/IEC 2382-15:1999].

Note 4 to entry: 15.04.01 (17.05.08) (2382).

[SOURCE: ISO/IEC 2382:2015, 2122374]

3.1.13
data variability
changes in transmission rate, format or structure, semantics, or quality of *datasets* (3.1.11)

3.1.14
data variety
range of formats, logical models, timescales, and semantics of a *dataset* (3.1.11)

Note 1 to entry: Data variety refers to irregular or heterogeneous data structures, their navigation, query, and data typing.

3.1.15
data velocity
rate of flow at which *data* (3.1.5) is created, transmitted, stored, analysed or visualised

3.1.16
data veracity
completeness and/or accuracy of *data* (3.1.5)

Note 1 to entry: Data veracity refers to descriptive data and self-inquiry about objects to support real-time decision-making.

3.1.17
data volatility
characteristic of *data* (3.1.5) pertaining to the rate of change of these data over time

[SOURCE: ISO/IEC 2382:2015, 17.06.06]

3.1.18
data volume
extent of the amount of *data* (3.1.5) relevant to impacting computation and storage resources and their management during data processing

Note 1 to entry: Data volume becomes important in dealing with large *datasets* (3.1.11), including their.

3.1.19
distributed data processing
data processing (3.1.9) in which the performance of operations is dispersed among the nodes in a computer network

[SOURCE: ISO/IEC 2382:2015, 18.01.08]

3.1.20
distributed file system
system which manages files and folders across multiple networked systems

3.1.21
file
named set of records treated as a unit

[SOURCE: ISO/IEC 2382:2015, 04.07.10]

3.1.22
gather
consolidation of results from multiple nodes in a cluster

Note 1 to entry: See *scatter-gather* (3.2.33).

3.1.23
horizontal scaling
providing a single logical unit through the connection of multiple hardware and software

Note 1 to entry: The example of horizontal scaling is increasing the performance of distributed data processing through the addition of nodes in the cluster for additional resources.

Note 2 to entry: Horizontal scaling for increasing performance is also referred to as scale-out.

3.1.24
metadata
data (3.1.5) about data or data elements, possibly including their data descriptions, and data about data ownership, access paths, access rights and *data volatility* (3.1.17)

[SOURCE: ISO/IEC 2382:2015, 17.06.05]

3.1.25
non-relational database
database (3.1.7) that does not follow a *relational model* (3.1.31)

Note 1 to entry: NoSQL, which is typically translated as "non SQL" or "not only SQL", is the term in common usage to refer to databases that do not conform to a relational model.

3.1.26
non-relational model
logical *data model* (3.1.10) that does not follow a *relational model* (3.1.31) for the storage and manipulation of *data* (3.1.5)

3.1.27
parallel
pertaining to a process in which all events occur within the same interval of time, each one handled by a separate but similar functional unit

Note 1 to entry: Example: The parallel transmission of the bits of a computer word along the lines of an internal bus.

[SOURCE: ISO/IEC 2382:2015, 03.02.01]

3.1.28
partially structured data
data (3.1.5) that has some organization

Note 1 to entry: Partially structured data is often referred to as semi-structured data by industry.

Note 2 to entry: examples of partially structured data are records with free text fields in addition to more structured fields. Such data is frequently represented in computer interpretable/parsible formats such as XML or JSON.

3.1.29
relational algebra
algebra for expressing and manipulating relations

[SOURCE: ISO/IEC 2382:2015, 17.04.08]

3.1.30
relational database
database (3.1.7) in which the data are organized according to a *relational model* (3.1.31)

[SOURCE: ISO/IEC 2382:2015, 17.04.05]

3.1.31
relational model
data model (3.1.10) whose structure is based on a set of relations

[SOURCE: ISO/IEC 2382:2015, 17.04.04]

3.1.32
scatter
distribution of processing across multiple nodes in a *cluster* (3.1.4)

Note 1 to entry: See *scatter-gather* (3.2.33).

3.2.33
scatter-gather
form of processing of large *datasets* (3.1.11) where the computation required is divided up and distributed across multiple nodes in a cluster and the overall result is combined from the results from each node

Note 1 to entry: Scatter-gather processing typically requires an algorithmic change to the processing software. An example of scatter-gather data processing is MapReduce.

3.1.34
streaming data
data (3.1.5) passing across an interface from a source that is operating continuously

[SOURCE: ISO/IEC 19784-4:2011, 4.4]

3.1.35
structured data
data (3.1.5) which are organized based on a pre-defined (applicable) set of rules

Note 1 to entry: The predefined set of rules governing the basis on which the data is structured needs to be clearly stated and made known.

Note 2 to entry: A pre-defined data model is often used to govern the structuring of data.

3.1.36
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

3.1.37
unstructured data
data (3.1.5) which are characterized by not having any structure apart from that record or file level

Note 1 to entry: On the whole unstructured data is not composed of data elements.

EXAMPLE An example of unstructured data is free text.

3.1.38
vertical scaling
act of increasing the performance of data processing through improvements to processors, memory, storage, or connectivity

Note 1 to entry: Vertical scaling for increasing performance is also referred to as scale-up.

3.2 Abbreviated terms

JSON	Javascript Object Notation
PII	Personally Identifiable Information
XML	Extensible Markup Language

4 Key characteristics of big data

4.1 General

The guidance for the choice of big data system is driven by four data characteristics, volume, velocity, variety and variability (see 4.2). The handling of these data characteristics is governed by processing characteristics as described in 4.3.

4.2 Key data characteristics

4.2.1 Data volume

Data volume represents the extensive amount of data available for analysis to extract valuable information. The massive amounts of data generated through internet activity was one of the primary drivers for the development of the big data processing techniques.

4.2.2 Data velocity

Data velocity is 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. An example of dealing with high velocity data is commonly referred to as techniques for streaming data.

4.2.3 Data variety

Data variety represents 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 identify 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 data analytics, complicates the integration of the variety of data. Metadata is increasingly used to aid in the integration. Impacts of variety on big data include that it requires the semantics of the data to be machine readable.

4.2.4 Data variability

Data variability refers to 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. In addition, a variability in data volumes implies the need to scale-up or scale-down virtualized resources to efficiently handle the additional processing load.

4.3 Key data processing characteristics

4.3.1 Data science

Data science refers to the process for extracting knowledge from data – the approach can be either through exploration or by hypothesis testing. Data science refers to the complete data analytics lifecycle where data analytics is understood as defined in 3.1.5.

4.3.2 Data volatility

Data volatility refers to a limited time span in which data values remain relevant for a particular analysis, expressed as a rate of change over time. In real-time data analytics situations, it is critical to operate on data immediately for decision-making. This is most evident in high velocity situations such as stock markets or telecommunications. However, data that is no longer valid for a particular time sensitive analytic due to time decay may still be valid for other non-time sensitive analytics.

4.3.3 Data veracity

Data veracity refers to the completeness and accuracy of the data and relates to the vernacular "garbage-in, garbage-out" description for data quality issues in existence for a long time. If the data analytics are causal, then the quality of every data element is extremely important. If the data analytics are correlations or trending over massive volume datasets, then individual bad elements could be lost in the overall counts and the trend could still be accurate.

4.3.4 Benefit

Benefit refers to the extent that the application outcome meets the goals for those building the big data processing system.

4.3.5 Data visualization

Data visualization refers to the presentation of data to allow the viewer to discern information about the data being displayed. Big data has required new techniques for large volume datasets, including aggregation and summarization to make the data more accessible visually. Big data also requires more attention to the visual presentation to decision-makers – to convey easily understood results while also communicating the complexity, accuracy, and probabilistic error range on the results.

4.3.6 Structured and unstructured data

Unstructured data has been increasing in both volume and prominence. While relational databases tend to have support for these types of data elements, their ability to directly analyse, index, and process them has tended to be both limited and accessed via non-standard SQL extensions. The need to analyse unstructured data has been present for many years. However, the big data paradigm shift has increased the emphasis on the value of unstructured data. Unstructured data also places an emphasis on new and different engineering methods that can analyse such data more efficiently.

4.3.7 Scaling

Big data refers to the extensibility of data repositories and data processing across resources working in parallel, in the same way that the compute-intensive simulation community embraced massively parallel processing. By working out methods for communication among resources, the same scaling is now available to data-intensive applications. Vertical scaling implies increasing the system parameters of processing speed, storage, and memory for greater performance. This approach is limited by physical capabilities whose improvements have been described by Moore's Law, requiring ever more sophisticated elements (e.g., hardware, software) that add time and expense to the implementation. The alternative method is to use horizontal scaling, to make use of distributed individual resources integrated to act as a single system. It is this horizontal scaling that is at the heart of the big data revolution. While the methods to achieve efficient scalability across resources will continually evolve,

this paradigm shift (in analogy to the prior shift in the simulation community to parallel processing) is a one-time occurrence.

4.3.8 Distributed file system

In distributed file systems, multi-structured (object) datasets are distributed across the computing nodes of the server cluster(s). The data may be distributed at the file/dataset level, or more commonly, at the block level, allowing multiple nodes in the cluster to interact with different parts of a large file/dataset simultaneously. Big data systems are frequently designed to take advantage of data locality to each node when distributing the processing, which avoids any need to move the data between nodes. In addition, many distributed file systems also implement file/block level replication where each file/block is stored multiple times on different machines nodes for both reliability/recovery (data is not lost if a node in the cluster fails), as well as enhanced data locality. Any type of data and many sizes of files can be handled without formal extract, transformation, and load conversions, with some technologies performing markedly better for large file sizes.

4.3.9 Distributed data processing

The popular framework for distributed computing consists of a storage layer and processing layer combination that implements a multiple-class, algorithm-programming model. Low-cost commodity servers supporting the distributed file system that stores the data can dramatically lower the storage costs of computing on a large scale of data (e.g., web indexing). In distributed data processing a query is scattered across the processors and the results are gathered into a central processor. Processing results are typically then loaded into an analysis environment. Multiple nodes (e.g. client nodes, data nodes, replica nodes) are arranged in a master-slave architecture to achieve efficiency, reliability, high availability, and fault-tolerance of the system.

4.3.10 Non-relational databases

In horizontally scaled systems, the data is distributed across the nodes of a cluster, while having a single logical structure. The new non-relational model database paradigms are typically referred to as NoSQL (Not Only SQL or No SQL) systems. The problem with identifying big data storage paradigms as NoSQL is, first, it describes the storage of data with respect to a set theory-based language for query and retrieval of data, and second, there is a growing capability in the application of the query languages similar to SQL against the new non-relational data repositories. While NoSQL is in such common usage that it will continue to refer to the new data models beyond the relational model, the term refers to databases that do not follow a relational model. Examples of non-relational database models include the column, sparse table, key-value, key-document, and graphical models.

Annex A
(informative)

Cross-cutting concepts of big data

A.1 General

The development of big data systems has implications for a number of technological areas of discussion and standardization. This annex discusses the relationships of big data to other standards development areas.

A.2 Metadata

Metadata is descriptive data, including for example the description of the processing history of the data. As big data systems are architected to perform distributed data processing including data that is external and not under the control of the big data system, the use of metadata becomes an increasingly important concept. As big data is reused for purposes far removed from its collection, it is important that metadata be associated with any data that is made available to others. Metadata also includes the source of the data and its usage. It can be categorized into business and technical metadata.

A.3 Algorithms

The development of algorithms for the analysis of big data need to consider the requirements of distributed data processing, the data was typically held locally. For big data, algorithms to process data across nodes need to be adapted to horizontal scaling to explicitly accommodate the particular distribution of data across nodes.

A.4 Cluster computing

Cluster computing refers to the distribution of processes across the network of machines. The machines leverage a software to use the physical system as a unit. If a service layer is inserted on top of the physical system, then the advantages of cloud computing are achieved."

NOTE In this rephrased definition of cluster computing the cluster is understood as "a combination of a set of interconnected machines/servers.

A.5 Cloud computing

Cloud computing is one of the paradigms for the availability and management of resources for big data systems. There are several key characteristics often present in cloud computing deployments including: broad network access, measured service, multi-tenancy, on-demand self-service, rapid elasticity and scalability, and resource pooling. Big data systems can leverage cloud computing deployments for infrastructure, platforms, or applications.

A.6 Data security

Big data systems have additional security concerns due to the distributed nature of the data processing. Additional vulnerabilities arise, for example, with the distributed ownership and control of the physical computer and network infrastructure, as well as the control across each level of the software and storage frameworks. Usually encryption, masking, and role-based access are implemented in a distributed data processing environment to ensure comprehensive data security at all levels including

transmission over the network. Some examples of datasets where high security is mandated include: sensitive customer information, product information, account details, trade data of the business, financial transactions, patient medical records, and defence records.

A.7 Privacy Protection Requirements

There are legal and regulatory requirements which impact and govern the use of personal information. An increasing amount of information on or about personal identifiable information is available from the web, social media, sensors, and so forth. Privacy protection in a wide sense is a set of legal and regulatory requirements which provide individuals with the right to control not only the use of their personal information but also its veracity, life cycle aspects, (including mandatory expungement), etc. In addition, a key privacy protection right is that of "informed consent" of the individual with respect to the use of their personal information. The integration of datasets from heterogeneous sources may well create sets of personal identifiable information or introduce a new use of personal information other than that of the goal of which the use of such personal information was given informed consent by that individual. It is therefore a legal and fiduciary requirement of any organization developing and using big data systems to ensure that where these involve data processing of personal information that applicable privacy protection requirements are fully supported and implemented.

A.8 SQL

SQL is a standard (see the ISO/IEC 9075 series), interactive programming language designed for querying, updating, and managing data and datasets in the database. SQL is designed for manipulating structured data, and it provides a mature and comprehensive framework for data access and supports a broad range of advanced analytical features. The extensions of SQL databases support the discovery of columns across a wide range of datasets: not only relational table/views, but also XML, JSON, spatial objects, image-style objects (Binary Large Objects and Character Large Objects), and semantic objects. NoSQL data management systems, which are intended to provide support for non-tabular structured data as well as unstructured and semi-structured data, have not yet settled on a common access language. Many NoSQL implementations have adopted SQL-like languages involving some subset of standard SQL with extensions that support specific features of NoSQL implementations.

A.9 Parallel computing

Big data typically refers to distributed data-intensive processing across the nodes of a cluster. The simulation community has been developing methods for compute-intensive processing across large clusters of nodes for many years. Given that both approaches represent extreme cases for large scale computation and large scale data analysis, techniques from both will be leveraged for spectrum of capabilities needing both compute-intensive and data-intensive computation.

A.10 Internet of things

More and more data is being created, along with computing systems capable of analysing data. Users want to leverage the amount of data available from a variety of sensors and other data generators. This provides efficient predictive data analytics to manage and control networked solutions. Typical technological advances in sensors, and the deployment of IPV6 to provide Internet connectivity to sensors creates the need for a big data system that can handle high velocity streaming data from a number of sources. This is in contrast to high volume big data systems that typically run batch jobs over a relatively small number of large datasets. This difference in the characteristics of the datasets has direct implications on the architecture and methods used for data analysis.

A.11 Programming languages

An analysis of extended data by using statistical computing is the fundamental approach for big data. Users can develop big data analytics systems by using general-purpose programming languages. The

