ISO/IEC JTC 1
Information technology
Secretariat: ANSI (United States)

Document type: Text for CD ballot or comment

Title: Text of CD 20546, Information Technology— Big Data— Overview and Vocabulary

Status: Please submit your vote via the online balloting system.

Date of document: 2016-04-12

Source: Project Editor

Expected action: VOTE

Action due date: 2016-07-13

Email of secretary: lrajchel@ansi.org

Committee URL: http://isotc.iso.org/livelink/livelink/open/jtc1

ISO/IEC JTC 1 N12986

2016-04-12

Replaces:

ISO/IEC JTC 1
Information Technology

Document Type:	CD for ballot
Document Title:	Text of CD 20546, Information Technology— Big Data— Overview and Vocabulary
Document Source:	Project Editor
Project Number:	
Document Status:	Please submit your vote via the online balloting system.
Action ID:	Vote
Due Date:	**2016-07-13**
Pages:	

Secretariat, ISO/IEC JTC 1, American National Standards Institute, 25 West 43rd Street, New York, NY 10036; Telephone: 1 212 642 4932; Facsimile: 1 212 840 2298; Email: lrajchel@ansi.org

ISO/IEC CD 20546	
Date: **2016-04-12**	Reference number: ISO/IEC JTC 1 N12986
Supersedes document	

THIS DOCUMENT IS STILL UNDER STUDY AND SUBJECT TO CHANGE. IT SHOULD NOT BE USED FOR REFERENCE PURPOSES.

ISO/IEC JTC 1 INFORMATION TECHNOLOGY Secretariat: USA (ANSI)	Circulated to P- and O-members, and to technical committees and organizations in liaison for: - discussion at - comment by - **voting by (P-members only)** **2016-07-13** Please return all votes and comments via the online balloting system.

ISO/IEC JTC 1

Title: Text of CD 20546, Information Technology— Big Data— Overview and Vocabulary

Project: 1.20546

Introductory note:

Recipients of this document are invited to submit notification of any relevant patent rights of which they are aware and to provide supporting documentation.

Reference number of working document: **ISO/JTC 1/ N XXXXX**

Date: 2016-03-xx

Reference number of document: **ISO/IEC CD 20546**

Committee identification: ISO/JTC 1/WG 9

Secretariat: ANSI

Information Technology— Big Data— Overview and Vocabulary

Warning

This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

Document type: International standard
Document subtype:
Document stage: (30) Committee Draft
Document language: E

Copyright notice

Contents

Page

Foreword ..... **iv**
Introduction ..... **v**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
3.1 Terms Defined Elsewhere **1**
3.2 Terms Defined in this International Standard **3**
3.3 Abbreviations **3**
4 Overview of Big Data **3**
4.1 Volume **4**
4.2 Velocity **4**
4.3 Variety **4**
4.4 Variability **4**
Annex A (informative) **Additional Related Concepts** **5**
A.1 Volatility **5**
A.2 Distributed File System **5**
A.3 Non-relational Models **5**
A.4 Non-relational Database **5**
A.5 Scatter-gather **5**
A.6 Computational Portability **5**
A.7 Data Science **5**
A.8 Value **5**
A.9 Veracity **6**
A.10 Unstructured Data **6**
Annex B (informative) **Cross-Cutting Concepts of Big Data** **7**
B.1 Metadata **7**
B.2 Analytics/Statistics **7**
B.3 Cluster Computing **7**
B.4 Cloud Computing **7**
B.5 Security **7**
B. 6 Privacy **7**
B.7 SQL **8**
B.8 Parallel Computing **8**
B.9 Internet of Things **8**
B.10 Programming Languages **8**
Bibliography **9**

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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

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.

ISO/IEC 20546 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information Technology*.

Introduction

The Big Data paradigm is a rapidly changing field with rapidly changing technologies. This standard will provide the normative definitions and the vocabulary needed to promote improved communication and understanding of this emerging area.

This international standard will provide the conceptual overview of the emerging 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. This international standard is expected to be time sensitive and will need revision over time.

Information Technology — Big Data — Overview and Vocabulary

1 Scope

This International Standard provides an overview of Big Data along with a set of terms and definitions. It provides a terminological foundation for Big Data-related standards.

2 Normative references

The following International Standards contain provisions which, through reference in this text, constitute provisions of this International Standard. At the time of publication, the editions indicated were valid. All Standards are subject to revision, and parties to agreements based on this International Standard are encouraged to investigate the possibility of applying the most recent edition of the Standards listed below. Members of IEC and ISO maintain registers of currently valid International Standards.

2.1 Identical International Standards

None.

2.2 Paired International Standards

None.

2.3 Additional references

None.

3 Terms and definitions

3.1 Terms Defined Elsewhere

For the purposes of this document, the following terms and definitions apply.

3.1.1
data processing
the systematic performance of operations upon data

NOTE 1 - Example: Arithmetic or logic operations upon data, merging or sorting of data, assembling or compiling of programs, or operations on text, such as editing, sorting, merging, storing, retrieving, displaying, or printing.

NOTE 2 - The term data processing should not be used as a synonym for information processing.

[SOURCE: ISO/IEC 2382-1;1999EF, 01.01.06]

3.1.2
distributed data processing
data processing in which the performance of operations is dispersed among the nodes in a computer network

[SOURCE: ISO/IEC 2382-18;1999EF, 18.01.08]

3.1.3
data volatility
characteristic of data pertaining to the rate of change of these data over time

[SOURCE: ISO/IEC 2382-17:1999, 17.06.06]

3.1.4
cluster (in distributed data processing)
A set of functional units under common control

[SOURCE: ISO/IEC 2382-18;1999EF 18.05.06

3.1.5
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 - Example: The parallel transmission of the bits of a computer word along the lines of an internal bus.

[SOURCE: ISO/IEC 2382-3: 1987EF 03.02.01]

3.1.6
relational algebra
algebra for expressing and manipulating relations

[SOURCE: ISO/IEC 2382-17:1999, 17.04.08]

3.1.7
relational model
data model whose structure is based on a set of relations

[SOURCE: ISO/IEC 2382-17:1999]

3.1.8
relational database
A database in which the data are organized according to a relational model

[SOURCE: ISO/IEC 2382-17 17.04.05]

3.1.9
data type
defined set of data 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 - 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. The permissible operations include the usual arithmetic operations on these integers.

[SOURCE: ISO/IEC 2382-17:1999, 17.05.08]

3.1.10
streaming data
data passing across an interface from a source that is operating continuously

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

3.1.11
metadata
data about data or data elements, possibly including their data descriptions, and data about data ownership, access paths, access rights and data volatility

[SOURCE: ISO/IEC 2382-17:1999, 17.06.05]

3.2 Terms Defined in this International Standard

3.2.1
big data
extensive datasets − primarily in the characteristics of volume, variety, velocity, and/or variability − that require a scalable architecture for efficient storage, manipulation, and analysis

NOTE - 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.

3.2.2
horizontal scaling
increasing the performance of distributed data processing through the addition of nodes in the cluster for additional resources

NOTE - horizontal scaling is also referred to as scale-out

3.2.3
vertical scaling
increasing the performance of data processing through improvements to processors, memory, storage, or connectivity

NOTE - vertical scaling is also referred to as scale-up

3.3 Abbreviations

For the purposes of this International Standard, the following abbreviations apply:

None.

4 Overview of Big Data

The term[1)] *big data* implies datasets that are extensive in volume, velocity, or variety. The term does not, however, represent data that is simply bigger 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 2000's, extensive datasets could no longer be handled using extant data system architectures. The new big data techniques represented a shift at that time to use *distributed data processing* through *horizontal scaling* to achieve the needed performance efficiency at an affordable cost.

In the evolution of data 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 1980's 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 bigger than before, as data has been steadily getting larger for decades. The big data revolution is instead a one-time

1) Note that all terms referenced in Clause 3 are italicized.

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 years since the mid 2000's 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 commonly 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 must be considered: the 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*.

The guidance for the choice of *big data* architectures is driven by four *big data* characteristics.

4.1 Volume

Volume is one of the characteristics 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.

4.2 Velocity

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. Dealing with high velocity data is commonly referred to as techniques for *streaming data*.

4.3 Variety

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 extract features that would allow integration with other data. The wider range of data formats, logical models, timescales, and semantics, which is desirous to use in analytics, complicates the integration of the variety of data. *Metadata* is increasingly used to aid in the integration

4.4 Variability

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.

Annex A
(informative)

Additional Related Concepts

The big data vocabulary consists of a number of additional terms and concepts that are in common usage.

A.1 Volatility

Volatility in Big Data usage refers to *data volatility*, or the rate of change of data over time

A.2 Distributed File System

Given the use of *horizontal scaling*, one new concept is a named set of records distributed across nodes of a *cluster* treated as a unit.

A.3 Non-relational Models

Logical data models that do not follow *relational algebra* for the storage and manipulation of data

A.4 Non-relational Database

A non-relational database is a database that does not follow a *relational model*. NoSQL, which is typically translated as “no-SQL” or “not only SQL”, is the term in common usage to refer to databases that do not conform to a *relational model*.

A.5 Scatter-gather

Processing large data sets that are distributed across nodes of a *cluster* requires an algorithmic change to do alter the algorithm to process the data on each node, and then return the results to process across the entire dataset. For example, MapReduce is an implementation of a scatter-gather process for data processing

A.6 Computational Portability

Computational portability is the vocabulary for the movement of the computation to the location of large volume datasets, rather than moving a large dataset to a different processing location.

Note: related to the prior concept of agents

A.7 Data Science

Data Science is the extraction of actionable knowledge from data through a process of discovery, or hypothesis and hypothesis testing.

A.8 Value

Value represents the benefit to the organization of the actionable knowledge derived from an analytic system. This term is not new to *big data*, but is often ascribed to *big data* due to the understanding that data has potential value that was typically not considered previously. Value arises as an output of the implementation of *big data* systems.

A.9 Veracity

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

A.10 Unstructured Data

Unstructured data, such as text, image, video, and relationship data, have been increasing in both volume and prominence. While modern *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 or semi-structured data has been present for many years. However, the *big data* paradigm shift has increased the emphasis on the value of unstructured or relationship data, and also on different engineering methods that can handle data more efficiently.

Annex B
(informative)

Cross-Cutting Concepts of Big Data

The development of Big Data systems has implications for a number of technological areas of discussion and standardization. In this section we discussion the relationships to other areas of investigation.

B.1 Metadata

Metadata is data about data or data elements, including 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 open 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.

B.2 Analytics/Statistics

The development of algorithms for the analysis of data previously did not consider the requirements of *distributed data processing*, the data was typically held locally. For *big data*, algorithms must be adapted to explicitly accommodate the particular distribution of data across nodes.

B.3 Cluster Computing

Big data systems can use the coordination of nodes in a cluster to achieve scalability in the *distributed data processing.* The nodes can be a physical compute system leveraging software to cause the nodes to function as a unit, or it can insert a virtualized interface on top of the physical systems to realize the advantages of cloud computing.

B.4 Cloud Computing

Cloud computing is a 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. There are several key characteristics often present in for cloud computing deployments including: broad network access, measured service, multi-tenancy, on-demand self-service, rapid elasticity and scalability, and resource pooling. Cloud computing is an infrastructure model for the development of a Big Data system. To achieve the needed scalability, systems can leverage infrastructure as a service (IaaS), data platform software can to provide a platform as a service (PaaS), or an application can be provided as software as a service (SaaS)

B.5 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 compute and network infrastructure, as well as the control across each level of the software and storage frameworks.

B. 6 Privacy

Privacy concerns are heightened now that so much data is available about individuals from the web, social media, sensors, and so forth. Privacy in a wide sense of the term is information control by an individual, which is not only to prevent data usage disadvantageous to the individual data subjects but also to facilitate data usage beneficial to them. A typical issue in the former respect is that integration of datasets may create personally identifiable information (PII), even if none of these datasets contains PII independently. On the other hand, the major problem in the latter respect is that personal data is mostly owned by organizations, and typically not by the data subjects themselves. This makes it hard for individuals to use their data for their own

sake, and for organizations to use personal data owned by other organizations, so that noone can utilize big and deep personal data.

B.7 SQL

SQL (Structured Query Language) is a standard interactive and programming language designed for querying, updating, and managing data and data set in the database management system. SQL was first published as ISO International Standard (ISO/IEC 9075) in 1987, and it has been revised to include a larger set of features as the query language for Information technology. SQL is designed for manipulating structured data, and it is also fast becoming the default language for big data analytics. SQL provides a mature and comprehensive framework for data access supporting a broad range of advanced analytical features. Analytics is a must-have functional component of data warehouse and big data. Modern SQL databases support the discovery of columns across a wide range of data set: not only relational table/views, but also XML, JSON, spatial objects, image-style objects (Binary Large Objects and Character Large Objects), and semantic objects.

B.8 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 extrema 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.

B.9 Internet of Things

More and more data are 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 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.

B.10 Programming Languages

An analysis of extended data by using statistical computing is the fundamental approach for *big data*. Customers can develop big data analytics system by using general-purpose programming languages. On the other hand, R is a programming language and software environment for statistical computing and graphics for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. R is freely available under the GNU General Public License since 1995, and is adopted onto various operating systems. R is widely used for developing statistical computing applications that analyse big data.

The needs for *distributed data processing* have led to a number of new programming and query languages suited to the development of *big data* systems, as well as new processes. An example of new programming languages and frameworks is the Spark framework for analytics. New data processing languages such as Pig have also been developed. New query languages address data in NoSQL systems. New processes include scatter-gather for *distributed data processing*.

Bibliography

[1] ISO 2382-1:1993, *Information technology – Vocabulary — Part 1: Fundamental terms*

[2] ISO 2382-3:1987, *Information processing systems - Vocabulary — Part 3: Equipment technology*

[3] ISO 2382-17:1999, *Information technology - Vocabulary — Part 17: Databases*

[4] ISO 2382-18:1999, *Information technology - Vocabulary — Part 18: Distributed data processing*

[5] ISO 19784-4:2011, *Information technology – Biometric application programming interface — Part 4: Biometric sensor function provider interface*