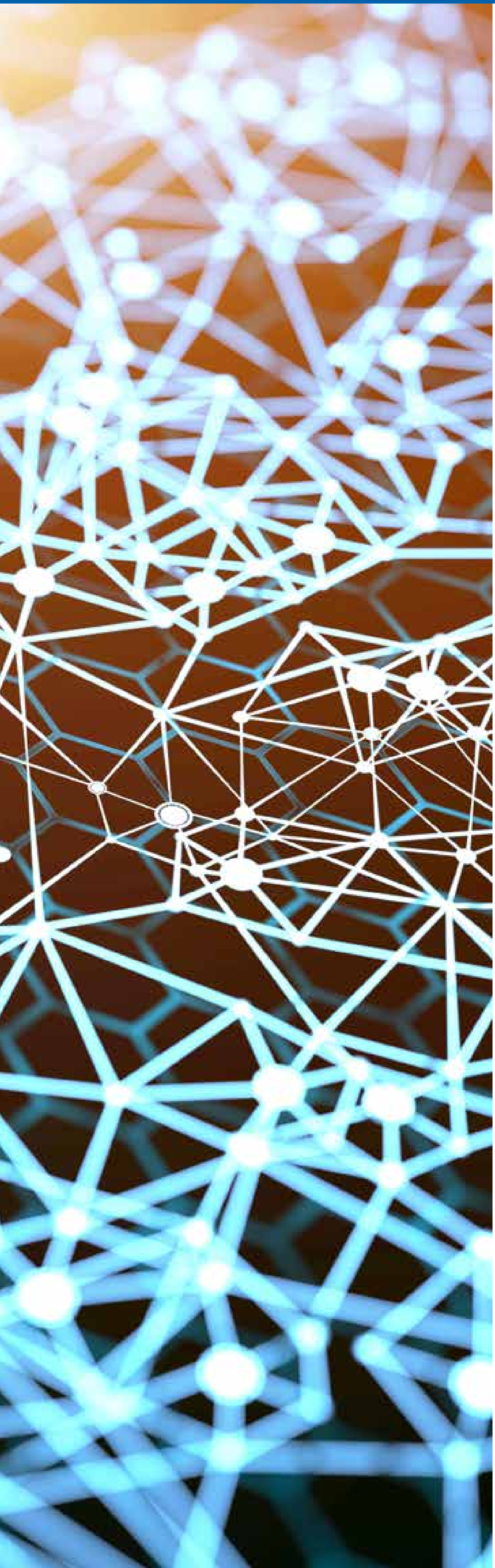


ISO/IEC JTC 1

Information technology

Big data

Preliminary Report 2014

Our vision

To be the world's leading provider of high quality, globally relevant International Standards through its members and stakeholders.

Our mission

ISO develops high quality voluntary International Standards that facilitate international exchange of goods and services, support sustainable and equitable economic growth, promote innovation and protect health, safety and the environment.

Our process

Our standards are developed by experts all over the world who work on a volunteer or part-time basis. We sell International Standards to recover the costs of organizing this process and making standards widely available.

Please respect our licensing terms and copyright to ensure this system remains independent.

If you would like to contribute to the development of ISO standards, please contact the ISO Member Body in your country:

www.iso.org/iso/home/about/iso_members.htm

This document has been prepared by:

ISO/IEC JTC 1, *Information technology*

Cover photo credit: ISO/CS, 2015

Copyright protected document

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. +41 22 749 01 11
Fax. +41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

1 Scope

This document does the following:

— Survey the existing ICT landscape for key technologies and relevant standards/models/ studies/ use cases and scenarios for Big Data from JTC 1, ISO, IEC and other standard setting organizations;

— Identify key terms and definitions commonly used in the area of Big Data; and

— Assess the current status of Big Data standardization market requirements, identify standards gaps, and propose standardization priorities to serve as a basis for future JTC 1 work.

2 Terms and definitions

2.1 Terms defined elsewhere

This document uses the following terms defined elsewhere.

2.1.1
capability
quality of being able to perform a given activity

[SOURCE: ISO 15531-1:2004]

2.1.2
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

[SOURCE: Recommendation ITU-T Y.3500 | ISO/IEC 17788:2014]

2.1.3
framework
structure expressed in diagrams, text, and formal rules which relates the components of a conceptual entity to each other

[SOURCE: ISO 17185-1:2014]

2.1.4
Internet of Things
integrated environment, inter-connecting anything, anywhere at anytime

[SOURCE: ISO/IEC JTC 1 SWG 5 Report:2013]

2.1.5
lifecycle
evolution of a system, product, service, project or other human-made entity from conception through retirement

[SOURCE: ISO/IEC/TR 29110-1:2011]

2.1.6
ownership

legal right of possession, including the right of disposition, and sharing in all the risks and profits commensurate with the degree of ownership interest or shareholding, as demonstrated by an examination of the substance, rather than the form, of ownership arrangements

[SOURCE: ISO 10845-5:2011]

2.1.7
privacy

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

[SOURCE: ISO/TS 17574:2009]

2.1.8
provenance

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

[SOURCE: ISO 19153:2014]

2.1.9
relational model

data model whose structure is based on a set of relations

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

2.1.10
repository

collection of all software-related artefacts belonging to a system or the location/format in which such a collection is stored

[SOURCE: ISO/IEC IEEE 24765:2010]

2.1.11
role

set of activities that serves a common purpose

[SOURCE: Recommendation ITU-T Y.3502 | ISO/IEC 17789:2014]

2.1.12
security

all aspects related to defining, achieving, and maintaining confidentiality, integrity, availability, non-repudiation, accountability, authenticity, and reliability of a system

[SOURCE: ISO/IEC 15288:2008]

2.1.13
sensor
device that observes and measures a physical property of a natural phenomenon or man-made process and converts that measurement into a signal

[SOURCE: ISO/IEC 29182-2:2013]

2.1.14
smart grid
electric grid system, which is characterized by the use of communication networks and the control of grid components and loads

[SOURCE: ISO/IEC/TR 27019:2013]

2.1.15
streaming data
data passing across an interface from a source that is operating continuously

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

3.1.16
traceability
property that allows the tracking of the activity of an identity, process, or an element throughout the supply chain

[SOURCE: ISO/IEC 27036-3:2013]

2.2 Terms defined in this report

This document defines the following terms:

2.2.1
Big Data Analytics
analytical functions to support the integration of results derived in parallel across distributed pieces of one or more data sources. This is a rapidly evolving field both in terms of functionality and the underlying programming model

2.2.2
Big Data Engineering
storage and data manipulation technologies that leverage a collection of horizontally coupled resources to achieve a nearly linear scalability in performance

2.2.3
Big Data Models
logical data models (relational and non-relational) and processing/computation models (batch, streaming, and transactional) for the storage and manipulation of data across horizontally scaled resources

2.2.4
Big Data Paradigm
distribution of data systems across horizontally-coupled independent resources to achieve the scalability needed for the efficient processing of extensive datasets

2.2.5
NoSQL
datastores and interfaces that are not tied to strict relational approaches

Note 1 to entry: Alternately called “no SQL” or “not only SQL”

2.2.6
Non-Relational Models
logical data models such as document, graph, key value and others that are used to provide more efficient storage and access to non-tabular datasets

2.2.7
schema-on-read
big data is often stored in a raw form based on its production, with the schema, needed for organizing (and often cleansing) the data, is discovered and transformed as the data is queried

2.3 Abbreviated terms and conventions

This document uses the following abbreviated terms and acronyms.

API	Applications Programming Interface
BSON	Binary JSON, representing simple data structures and associative arrays
CAGR	Compound Annual Growth Rate
DAPS	Distributed Application Platforms and Services
GPS	Global Positioning System
HTML	Hypertext Markup Language
ICT	Information and Communication Technology
ITU-T	International Telecommunications Union-Telecommunications Standardization Sector
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
ISO/IEC JTC 1/SG 1	Smart Cities
ISO/IEC JTC 1/SG 2	Big Data
ISO/IEC JTC 1/SWG 5	Internet of Things, (IoT)
ISO/IEC JTC 1/SC 6	Telecommunications and information exchange between systems
JTC 1	Joint Technical Committee 1
JSON	Java Script Object Notation
MPP	Massively Parallel Processing
NB	National Body

NIST	National Institute of Standards and Technology
NoSQL	Not Only Structured Query Language
OGC	Open Geospatial Consortium
OASIS	Organization for the Advancement of Structured Information Standards
POSIX	Portable Operating System Interface
RA	Reference Architecture
RFID	Radio-Frequency Identification
SC	Standards Committee
SDO	Standards Development Organization
SG	Study Group
SOA	Service-Oriented Architecture
SQL	SQL Query Language
SQL/MM	SQL Multimedia
SWG	Special Working Group
TPC	Transaction Processing Performance Council
W3C	World Wide Web Consortium
XML	Extensible Markup Language

3 Introduction to Big Data

In recent years, the term Big Data has emerged to describe a new paradigm for data applications. New technologies tend to emerge with a lot of hype, but it can take some time to tell what is new and different. While Big Data has been defined in a myriad of ways, the heart of the Big Data paradigm is that is too big (volume), arrives too fast (velocity), changes too fast (variability), contains too much noise (veracity), or is too diverse (variety) to be processed within a local computing structure using traditional approaches and techniques. The technologies being introduced to support this paradigm have a wide variety of interfaces making it difficult to construct tools and applications that integrate data from multiple Big Data sources. This report identifies potential areas for standardization within the Big Data technology space.

3.1 General concept of Big Data

Big Data are used as a concept that refers to the inability of traditional data architectures to efficiently handle the new data sets. Characteristics that force a new architecture to achieve efficiencies are the data set-at-rest characteristics ***volume***, and ***variety*** of data from multiple domains or types; and from the data-in-motion characteristics of ***velocity***, or rate of flow, and ***variability*** (principally referring to a change in velocity). Each of these characteristics results in different architectures or different data lifecycle process orderings to achieve needed efficiencies. A number of other terms (often starting with the letter 'V') are also used, but a number of these refer to the analytics and not big data architectures.

The new big data paradigm occurs when the scale of the data at rest or in motion forces the management of the data to be a significant driver in the design of the system architecture. Fundamentally the big data paradigm represents a shift in data system architectures from monolithic systems with vertical scaling (faster processors or disks) into a horizontally scaled system that integrates a loosely coupled set of resources. This shift occurred 20-some years ago in the simulation community when the scientific simulations began using massively parallel processing (MPP) systems. In different combinations of splitting the code and data across independent processors, computational scientists were able to greatly extend their simulation capabilities. This of course introduced a number of complications in such areas as message passing, data movement, and latency in the consistency across resources, load balancing, and system inefficiencies while waiting on other resources to complete their tasks. In the same way, the big data paradigm represents this same shift, again using different mechanisms to distribute code and data across loosely-coupled resources in order to provide the scaling in data handling that is needed to match the scaling in the data.

The purpose of storing and retrieving large amounts of data is to perform analysis that produces additional knowledge about the data. In the past, the analysis was generally accomplished on a random sample of the data.

> **Big Data Paradigm** consists of the distribution of data systems across horizontally-coupled independent resources to achieve the scalability needed for the efficient processing of extensive data sets.

With the new Big Data Paradigm, analytical functions can be executed against the entire data set or even in real-time on a continuous stream of data. Analysis may even integrate multiple data sources from different organizations. For example, consider the question "What is the correlation between insect borne diseases, temperature, precipitation, and changes in foliage". To answer this question an analysis would need to integrate data about incidence and location of diseases, weather data, and aerial photography.

While we certainly expect a continued evolution in the methods to achieve efficient scalability across resources, this paradigm shift (in analogy to the prior shift in the simulation community) is a one-time occurrence; at least until a new paradigm shift occurs beyond this "crowdsourcing" of processing or data system across multiple horizontally-coupled resources.

> **Big Data Engineering** is the storage and data manipulation technologies that leverage a collection of horizontally coupled resources to achieve a nearly linear scalability in performance.

New engineering techniques in the data layer have been driven by the growing prominence of data types that cannot be handled efficiently in a traditional relational model. The need for scalable access in structured and unstructured data has led to software built on name-value/key-value pairs or columnar (big table), document-oriented, and graph (including triple-store) paradigms.

Non-Relational Models refers to logical data models such as document, graph, key value and others that are used to provide more efficient storage and access to non-tabular data sets.

NoSQL (alternately called "no SQL" or "not only SQL") refers to datastores and interfaces that are not tied to strict relational approaches.

Big Data Models refers to logical data models (relational and non-relational) and processing/computation models (batch, streaming, and transactional) for the storage and manipulation of data across horizontally scaled resources.

Schema-on-read big data are often stored in a raw form based on its production, with the schema, needed for organizing (and often cleansing) the data, is discovered and transformed as the data are queried. This is critical since in order for many analytics to run efficiently the data must be structured to support the specific algorithms or processing frameworks involved.

Big Data Analytics is rapidly evolving both in terms of functionality and the underlying programming model. Such analytical functions support the integration of results derived in parallel across distributed pieces of one or more data sources.

The Big Data paradigm has other implications from these technical innovations. The changes are not only in the logical data storage, but in the parallel distribution of data and code in the physical file system and direct queries against this storage.

The shift in thinking causes changes in the traditional data lifecycle. One description of the end-to-end data lifecycle categorizes the steps as collection, preparation, analysis and action. Different big data use cases can be characterized in terms of the data set characteristics at-rest or in-motion, and in terms of the time window for the end-to-end data lifecycle. Data set characteristics change the data lifecycle processes in different ways, for example in the point of a lifecycle at which the data are placed in persistent storage. In a traditional relational model, the data are stored after preparation (for example after the extract-transform-load and cleansing processes). In a high velocity use case, the data are prepared and analysed for alerting, and only then is the data (or aggregates of the data) given a persistent storage. In a volume use case the data are often stored in the raw state in which it was produced, prior to the application of the preparation processes to cleanse and organize the data. The consequence of persistence of data in its raw state is that a schema or model for the data are only applied when the data are retrieved, known as schema on read.

A third consequence of big data engineering is often referred to as "*moving the processing to the data, not the data to the processing*". The implication is that the data are too extensive to be queried and transmitted into another resource for analysis, so the analysis program is instead distributed to the data-holding resources; with only the results being aggregated on a different resource. Since I/O bandwidth is frequently the limited resource in moving data, another approach would be to embed query/filter programs within the physical storage medium.

At its heart, Big Data refers to the extension of data repositories and processing across horizontally-scaled resources, much in the same way the compute-intensive simulation community embraced massively parallel processing two decades ago. In the past, classic parallel computing applications utilized a rich set of communications and synchronization constructs and created diverse communications topologies. In contrast, today, with data sets growing into the Petabyte and Exabyte scales, distributed processing frameworks offering patterns such as map-reduce, offer a reliable high-level, commercially viable compute model based on commodity computing resources, dynamic resource scheduling, and synchronization techniques.

3.2 Definition of Big Data

The term "Big Data" is used in a variety of contexts with a variety of characteristics. To understand where standards will help support the big data paradigm, we have to reach some level of consensus on what the term really means. This report uses the following working definition of "Big Data":

> **Big Data** is a data set(s) with characteristics (e.g. volume, velocity, variety, variability, veracity, etc.) that for a particular problem domain at a given point in time cannot be efficiently processed using current/existing/established/traditional technologies and techniques in order to extract value.

The above definition distinguishes Big Data from business intelligence and traditional transactional processing while alluding to a broad spectrum of applications that includes them. The ultimate goal of processing Big Data is to derive differentiated value that can be trusted (because the underlying data can be trusted). This is done through the application of advanced analytics against the complete corpus of data regardless of scale. Parsing this goal helps frame the value discussion for Big-Data use cases.

— **Any scale of operations and data:** Utilizing the entire corpus of relevant information, rather than just samples or subsets. It's also about unifying all decision-support time-horizons (past, present, and future) through statistically derived insights into deep data sets in all those dimensions.

— **Trustworthy data:** Deriving valid insights either from a single-version-of-truth consolidation and cleansing of deep data, or from statistical models that sift haystacks of "dirty" data to find the needles of valid insight.

— **Advanced analytics:** Faster insights through a variety of analytic and mining techniques from data patterns, such as "long tail" analyses, micro-segmentations, and others, that are not feasible if you're constrained to smaller volumes, slower velocities, narrower varieties, and undetermined veracities.

A difficult question is what makes "Big Data" big, or how large does a data set have to be in order to be called big data? The answer is an unsatisfying "it depends". Part of this issue is that "Big" is a relative term and with the growing density of computational and storage capabilities (e.g. more power in smaller more efficient form factors) what is considered big today will likely not be considered big tomorrow. Data are considered "big" if the use of the new scalable architectures provides improved business efficiency over other traditional architectures. In other words the functionality cannot be achieved in something like a traditional relational database platform.

Big data essentially focuses on the self-referencing viewpoint that data are big because it requires scalable systems to handle it, and architectures with better scaling have come about because of the need to handle big data.

3.3 Organizational drivers of Big Data

The key drivers for Big Data in organizations are about realizing value in any of several ways:

— Insight: enable discovery of deeper, fresher insights from all enterprise data resources

— Productivity: improve efficiency, effectiveness, and decision-making

— Speed: facilitate more timely, agile response to business opportunities, threats, and challenges

— Breadth: provide a single view of diverse data resources throughout the business chain

— Control: support tighter security, protection, and governance of data throughout its lifecycle

— Scalability: improve the scale, efficiency, performance, and cost-effectiveness of data/analytics platforms

3.4 Key characteristics of Big Data

The key characteristics of Big Data focus on volume, velocity, variety, veracity, and variability. The following subclauses go into further depth on these characteristics.

3.4.1 Volume

Traditionally, the data volume requirements for analytic and transactional applications were in sub-terabyte territory. However, over the past decade, more organizations in diverse industries have identified requirements for analytic data volumes in the terabytes, petabytes, and beyond.

Estimates produced by longitudinal studies started in 2005[8] show that the amount of data in the world is doubling every two years. Should this trend continue, by 2020, there will be 50 times the amount of data as there had been in 2011. Other estimates indicate that 90 % of all data ever created, was created in the past 2 years [7]. The sheer volume of the data are colossal - the era of a trillion sensors is upon us. This volume presents the most immediate challenge to conventional information technology structures. It has stimulated new ways for scalable storage across a collection of horizontally coupled resources, and a distributed approach to querying.

Briefly, the traditional relational model has been relaxed for the persistence of newly prominent data types. These logical non-relational data models, typically lumped together as NoSQL, can currently be classified as Big Table, Name-Value, Document and Graphical models. A discussion of these logical models was not part of the phase one activities that led to this document.

3.4.2 Variety

Traditionally, enterprise data implementations for analytics and transactions operated on a single structured, row-based, relational domain of data. However, increasingly, data applications are creating, consuming, processing, and analysing data in a wide range of relational and non-relational formats including structured, unstructured, semi-structured, documents and so forth from diverse application domains.

Traditionally, a variety of data was handled through transforms or pre-analytics to extract features that would allow integration with other data through a relational model. Given the wider range of data formats, structures, timescales and semantics that are desirous to use in analytics, the integration of this data becomes more complex. This challenge arises as data to be integrated could be text from social networks, image data, or a raw feed directly from a sensor source. The “Internet of Things” is the term used to describe the ubiquity of connected sensors, from RFID tags for location, to smart phones, to home utility meters. The fusion of all of this streaming data will be a challenge for developing a total situational awareness. Big Data Engineering has spawned data storage models that are more efficient for unstructured data types than a relational model, causing a derivative issue for the mechanisms to integrate this data. It is possible that the data to be integrated for analytics may be of such volume that it cannot be moved in order to integrate, or it may be that some of the data are not under control of the organization creating the data system. In either case, the variety of big data forces a range of new big data engineering in order to efficiently and automatically integrate data that is stored across multiple repositories and in multiple formats.

3.4.3 Velocity

The Velocity is the speed/rate at which the data are created, stored, analysed and visualized. Traditionally, most enterprises separated their transaction processing and analytics. Enterprise data analytics were concerned with batch data extraction, processing, replication, delivery, and other applications. But increasingly, organizations everywhere have begun to emphasize the need for real-time, streaming, continuous data discovery, extraction, processing, analysis, and access.

In the big data era, data are created in real-time or near real-time. With the availability of Internet connected devices, wireless or wired, machines and devices can pass-on their data the moment it is created. Data Flow rates are increasing with enormous speeds and variability, creating new challenges to enable real or near real-time data usage. Traditionally this concept has been described as streaming data. As such there are aspects of this that are not new, as companies such as those in telecommunication have been sifting through high volume and velocity data for years. The new horizontal scaling approaches do however add new big data engineering options for efficiently handling this data.

3.4.4 Variability

Variability refers to changes in data rate, format/structure, semantics, and/or quality that impact the supported application, analytic, or problem. Specifically, variability is a change in one or more of the other Big Data characteristics. Impacts can include the need to refactor architectures, interfaces, processing/algorithms, integration/fusion, storage, applicability, or use of the data.

The other characteristics directly affect the scope of the impact for a change in one dimension. For, example in a system that deals with petabytes or exabytes of data refactoring the data architecture and performing the necessary transformation to accommodate a change in structure from the source data may not even be feasible even with the horizontal scaling typically associated with big data architectures. In addition, the trend to integrate data from outside the organization to obtain more refined analytic results combined with the rapid evolution in technology means that enterprises must be able to adapt rapidly to data variations.

3.4.5 Veracity

Veracity refers to the trustworthiness, applicability, noise, bias, abnormality and other quality properties in the data. Veracity is a challenge in combination with other Big Data characteristics, but is essential to the value associated with or developed from the data for a specific problem/application. Assessment, understanding, exploiting, and controlling Veracity in Big Data cannot be addressed efficiently and sufficiently throughout the data lifecycle using current technologies and techniques.

3.5 Roles in Big Data ecosystem

The different functional roles within a typical Big Data ecosystem are as follows:

— **Data Provider:** introduces new data or information feeds into the ecosystem

— **Big Data Application Provider:** executes a life cycle (collection, processing, dissemination) controlled by the system orchestrator to implement specific vertical applications requirements and meet security and privacy requirements

— **Big Data Framework Provider:** establishes a computing fabric (computation and storage resources, platforms, and processing frameworks) in which to execute certain transformation applications while protecting the privacy and integrity of data

— **Data Consumer:** includes end users or other systems who utilize the results of the Big Data Application Provider

— **System Orchestrator:** defines and integrates the required data application activities into an operational vertical system

— **Security and Privacy:** the role of managing and auditing access to and control of the system and the underlying data including management and tracking of data provenance

— **Management:** the overarching control of the execution of a system, the deployment of the system, and its operational maintenance

3.6 Security and privacy for Big Data

Security and Privacy issues arise in any distributed computing environment. These issues are exacerbated by Big Data for a number of reasons.

3.6.1 Issues

Much of the value of Big Data comes from combining data from different sources. Combining data in this manner can provide context. Thus, data that may not have been intelligible on its own can be mined for private information given enough context.

Some of the Big Data comes from social media and medical records and inherently contains private information. While social media sites may not do much to protect their users, analysis of such data, particularly in the presence of context, must protect privacy.

Big Data may be gathered from diverse end points and brought together for a variety of applications. There may be more types of actors than just providers and consumers—primarily, data owners, such as mobile users and social network users. Some "actors" may be devices that ingest data streams for still different data consumers. Moreover, the volume of Big Data necessitates storage in multi-tiered storage media some of which may store aggregated data. Aggregation and the movement of data between applications and tiers can lose provenance and metadata information and open the door to privacy violations.

Security and Privacy are important for both data quality and for protection. Big Data frequently moves across individual boundaries to group, community of interest, state, national, and international boundaries. Provenance, a component of veracity, addresses the problem of understanding the data's original source and what has been done with the data. One approach is through the use of metadata, though the problem extends beyond metadata maintenance. Provenance also encompasses information assurance for the methods through which information was collected. For example, when sensors are used, traceability to calibration, version, sampling and device configuration are needed.

The universal attribute of data ownership must be addressed in the context of the security and privacy of Big Data. Ownership is an attribute (which may or may not be visible to users) that associates data with one or more entities who own or can influence what can be done with the data (For example, you influence but cannot change your credit record). In databases, ownership confers the privileges to create, read, update, and delete data. Transparency of ownership enables trust and control for data owners, as well as openness and utility for enterprises and society. Maintaining data provenance allows traceability through the data lifecycle and tracks data ownership and change.

Distributed programming frameworks developed to support volume and velocity were not necessarily designed with security in mind. Malfunctioning computing nodes might leak confidential data. Partial infrastructure attacks could compromise a significantly large fraction of the system due to high levels of connectivity and dependency. If the system does not enforce strong authentication among geographically distributed nodes, rogue nodes can be added that can eavesdrop on confidential data.

Data search and selection can lead to privacy or security policy concerns because results can be provided without provenance and access control policies may be lost in the search and selection process. It is often unclear what capabilities are provided by a provider in this respect. A combination of user competency and system protections is likely needed, including the exclusion of databases that enable re-identification. Because there may be disparate processing steps between the data owner, provider, and data consumer, the integrity of data coming from end points must be ensured. End-to-end information assurance practices for Big Data—for example, for verifiability—are not dissimilar from other systems, but must be designed on a larger scale.

Retargeting traditional relational database security to non-relational databases has been a challenge. These systems were not designed with security in mind, and security is usually relegated to middleware.

The movement and aggregation of data between applications has led to a requirement of systematically analysing the threat models and research and development of novel techniques. The threat model for network-based, distributed, auto-tier systems includes the following major scenarios: confidentiality and integrity, provenance, availability, consistency, collusion attacks, roll-back attacks and recordkeeping disputes.

The flip side of having volumes of data are that analytics can be performed to detect security breach events. This is an instance where Big Data technologies can fortify security.

Big Data systems will exert stresses upon security and privacy aspects of “conventional” applications and data produced by those applications. The potential of Big Data analytics, whether a current option or merely a future possibility, creates a natural bias against discarding data. Inconvenient archives could be relegated to specialized uses, rather than a recognized design pattern which relegates data to an intentionally degraded access modality. Security and privacy frameworks will evolve as Big Data systems are deployed more widely, but much Big Data may be collected through legacy applications that did not benefit from those frameworks and did not anticipate Big Data uses.

Requirements development for Big Data systems will emphasize extensibility and scalability in ways that set the stage for greater threats to security and privacy. For example, systems will be architected to one day incorporate real time feeds from devices that are part of the Internet of Things – even if those feeds are not yet available.

While US. Safe Harbor privacy principles have been criticized as inadequate, a few of its principles serve to highlight areas in which Big Data systems are likely to be tested: notice, choice, onward data transfer, security, data integrity, the access of an individual to correct or delete data, and effective enforcement of these guidelines. Each of these areas is challenging enough in traditional IT settings. The productive use of derived, indirect and correlated data in Big Data will further amplify the need for increased control; however, current trust exchange technologies do not address many Safe Harbor needs.

The human element in privacy and security for Big Data will also be transformed in ways not easily anticipated. As more data becomes available through Big Data analytics engines, there will be more “analysts,” some of them less well versed in best practices for preserving security and privacy. Similarly, analysts will likely gain access to data whose provenance and usage they are comparatively unfamiliar with. The security and privacy problems created through human agents will vary from benign to accidental to malicious.

Security and privacy measures in Big Data must scale nonlinearly. Consider first the scope of existing regulations, such as the EU General Data Protection Regulation, APEC Cross Border Privacy Rules, the Privacy Act of 1974 and the California Right to Privacy. Then consider what new regulations are likely to emerge to address perceived and real risks as the public and regulators become aware of Big Data capabilities. For instance, the HIPAA guidance "minimum necessary use and disclosure" could not have anticipated the many possible uses – salutary or otherwise – for personal health records.

Architects who believed they understood the full scope of audit, forensic, compliance, civil rights and risk elements of security and privacy may come to feel otherwise. The relative comfort of many isolated information systems will, for many practitioners, become a relic of a less ubiquitously connected past. Information assurance – including responsibility for resilience and reliability – summons different specializations within computing, but Big Data are likely to give them a security and privacy face to the public. Technical solutions must take this into account.

3.6.2 Recommendations

It is of paramount importance that Big Data systems be designed with security in mind from the ground up rather than have it emerge as an afterthought - which often leads to adoption of ad hoc solutions with unsystematic and vague threat models in mind. If there is no global perspective on security then fragmented solutions to address security may not offer any security at all and often impart a false sense of safety.

Data aggregation and dissemination should be secured inside the context of a formal, understandable framework. This process should be an explicit part of a data consumer's contract with the data owner. Privacy-preserving mechanisms are needed for Big Data, such as for Personally Identifiable Information so that provenance information and data access policies are not lost. Anonymization and obfuscation of some data values can be used to protect sensitive information. For example, geographic location may be generalized to a village or a town rather than the exact coordinates.

The availability of data and its current status to data consumers is often an important aspect of Big Data. In some settings, this may dictate a need for public or closed-garden portals and ombudsman-like roles for data at rest.

While the context of the data, in terms of its structure, might be a standard Big Data problem, the payload might be encrypted to enforce confidentiality. However, traditional encryption technology hinders organization of data based on semantics. The aim of encryption is to provide semantic security, which means that the encryption of any value is indistinguishable from the encryption of any other value.

Data encrypted using known standard and/or commercial algorithms cannot be searched, ordered, added, etc. While some basic processing operations can be performed on the data encrypted using the emerging homomorphic algorithms, it will take time until this approach matures and becomes applicable to real-life scenarios.

4 Relevant standardization activities

This clause describes related standardization activities with Big Data including ISO/IEC JTC 1 in order to identify standards gaps. The current content is based on an informal survey by this Study Group and contributions from other SDOs.

Specific Big Data standards are being developed by a variety of well-established SDOs and industry consortia as outlined in Table 1. The following sublcauses provide additional details on activities by those organizations that relate to Big Data.

Table 1 — The mission and key members of major Consortia for Big Data standardization

SDO/Consortium	Interests area on standardization	Main deliverables
ISO/IEC JTC 1/SC 32	Data management and interchange, including database languages, multimedia object management, metadata management, and e-Business.	e-Business standards, including role negotiation; metadata repositories, model specification, metamodel definitions; SQL; and object libraries and application packages built on (using) SQL.
ISO/IEC JTC 1/SC 38	Standardization for interoperable Distributed Application Platform and Services including Web Services, Service Oriented Architecture (SOA), and Cloud Computing	Cloud Data Management Interfaces, Open Virtualization Format, Web Services Interoperability
ITU-T SG13	Cloud computing for Big Data	Cloud computing based big data requirements, capabilities, and use cases.
W3C	Web and Semantic related standards for markup, structure, query, semantics, and interchange.	Multiple standards including ontology specification standards, data markup, query, access control, and interchange.
Open Geospatial Consortium	Geospatial related standards for the specification, structure, query, and processing of location related data.	Multiple standards related to the encoding, processing, query, and access control of geospatial data.
Organization for the Advancement of Structured Information Standards	Information access and exchange.	A set of protocols for interacting with structured data content such as OData (https://www.oasis-open.org/standards#odatav4.0), standards for security, Cloud computing, SOA, Web services, the Smart Grid, electronic publishing, emergency management, and other areas
Transaction Processing Performance Council	Benchmarks for Big Data Systems	Specification of TPC Express, Benchmark^{TM} for Hadoop system and the related kit
TM Forum	Enable enterprises, service providers and suppliers to continuously transform in order to succeed in the digital economy	Share experiences to solve critical business challenges including IT transformation, business process optimization, big data analytics, cloud management, and cyber security.

4.1 ISO/IEC JTC 1/SC 32

ISO/IEC JTC 1/SC 32, titled “Data management and interchange”, currently works in several distinct, but related, areas of Big Data technology.

- SQL is already adding new features to support Big Data. In addition, SQL has been supporting bi-temporal data, two forms of semi-structured data (XML and JSON), and multidimensional arrays. SQL implementations are known to exist, which utilize storage engines that are built using several of the NoSQL technologies, including name-value pairs, big table, and document.
- Metadata efforts have focused on two major areas: (1) the specification and standardization of data elements, including the registration of those data elements (essentially, a repository for data element definitions); and (2) the definition of metamodels (to describe data and application models) and definitions of those models themselves.

4.2 ISO/IEC JTC 1/SC 38

ISO/IEC JTC 1/SC 38, titled “Distributed application platforms and services (DAPS)”, currently works in several areas related to areas of the Big Data Paradigm:

- Cloud Data Management Interfaces;
- Open Virtualization Format;
- Web Services Interoperability.

4.3 ITU-T SG13

ITU-T SG13 Question17 has initiated a new draft Recommendation on Big Data (Y.Bigdata-reqts)[6] with the title of “Requirements and capabilities for cloud computing based big data” in July 2013. The scope of Y.BigData-reqts is:

- Overview of cloud computing based big data;
- Cloud computing based big data requirements;
- Cloud computing based big data capabilities;
- Cloud computing based big data use cases and scenarios.

4.4 W3C

Most W3C work revolves around the standardization of Web technologies. Given that one of the primary contributors to the growth of Big Data has been the growth of the Internet and World Wide Web (WWW) many of the developing standards around web technologies must deal with the challenges inherent in Big Data.

Currently the following examples of W3C standard efforts relate to big data technologies and interests:

- Model for Tabular Data and Metadata on the Web
- Delivery Context Ontology
- Efficient XML Interchange
- Linked Data
- Mathematical Markup Language
- OWL Web Ontology Language
- Platform for Privacy Preferences
- Protocol for Web Description Resources (POWDER)
- Provenance
- Relational Database to Resource Description Framework (RDB2RDF)
- Resource Description Framework (RDF)
- Rule Interchange Format (RIF)
- Service Modelling Language
- Sparse Query Language (SPARQL)
- Extensible Markup Language (XML) and associated technologies (XQuery, XPath, etc.)

Furthermore, W3C has the following data activities which may relate to Big Data:

- Big Data CG
- ETL Markup Language CG
- Resource Description Framework (RDF) WG
- Linked Data Platform (LDP) WG
- Government Linked Data (GLD) WG
- CSV (comma-separated values) on the Web WG
- Provenance WG

Additional information on these standards efforts can be found at: http://www.w3.org

4.5 Open Geospatial Consortium (OGC)

The Open Geospatial Consortium is an international industry consortium of companies, government agencies, research institutes and universities participating to develop standards for interoperable "geo-enable" solutions on the Web, wireless and location-based services. Geospatial data (particularly sensor imagery, simulation output, and statistics data) represents a common big data problem due to the sheer size (volume and number of records) of the data involved. With the extensive growth of location data collection from mobile devices and geo-sensors, OGC was faced with key big data problem that must be address. For these reasons, OGC established Big Data Domain Working Group (BigData DWG) in 2014 [14].

OGC BigData DWG aims to clarify:

- Foundational terminologies in the context of data analytics understanding differences/overlaps with terms like data analysis, data mining, etc.
- A systematic classification of analysis algorithms, analytics tools, data and resource characteristics, and scientific queries.

Currently the following examples of OGC standard efforts relate to big data technologies and interfaces:

- Data Model Extension standards (e.g. netCDF, HDF, GeoTIFF, geoISON)
- Registry Services
- Metadata Profiles
- GeoAPI Implementation Standard (joint effort with ISO Technical Committee 211)
- Geospatial eXtensible Access Control Markup Language Encoding Standard (GeoXACML)
- GeoSPARQL – A geographic query language for RDF
- Web Coverage Processing Service (WCPS) Interface Standard, OGC's spatio-temporal raster query language

Additional information on these and other potentially applicable OGC standards can be found at http://www.opengeospatial.org/standards/is. Note: there are other SDOs such as ISO TC 204 (Intelligent Transport Systems) and TC 211 (Geographic information and Geomatics) working on the related geospatial projects.

4.6 Organization for the Advancement of Structured Information Standards (OASIS)

Organization for the Advancement of Structured Information Standards is a non-profit consortium that drives the development, convergence and adoption of open standards for the global information society.

OASIS promotes industry consensus and produces worldwide standards for security, Cloud computing, SOA, Web services, the smart grid, electronic publishing, emergency management, and other areas. OASIS open standards offer the potential to lower cost, stimulate innovation, grow global markets, and protect the right of free choice of technology.

The following OASIS technical committees and activities are relevant to Big Data:

- OASIS Advanced Message Queuing Protocol (AMQP) TC: Defining a ubiquitous, secure, reliable and open internet protocol for handling business messaging.
- OASIS Key-Value Database Application Interface (KVDB) TC: Defining an open application programming interface for managing and accessing data from database systems based on a key-value model
- OASIS Message Queuing Telemetry Transport (MQTT) TC: Providing a lightweight publish/subscribe reliable messaging transport protocol suitable for communication in M2M/IoT contexts where a small code footprint is required and/or network bandwidth is at a premium.
- OASIS XML Interchange Language (XMILE) for System Dynamics TC: Defining an open XML protocol for sharing interoperable system dynamics models and simulations.
- Cross-Enterprise Security and Privacy Authorization (XSPA)
- Darwin Information Typing Architecture (DITA)
- Directory Services Markup Language (DSML)
- eXtensible Access Control Markup Language (XACML)
- XRD (Extensible Resource Descriptor)
- Open Data Protocol (OData)
- Search Web Services (SWS)
- Service Provisioning Markup Language (SPML)
- Topology and Orchestration Specification for Cloud Applications (TOSCA)
- Universal Description, Discovery and Integration (UDDI)
- Unstructured Information Management Architecture (UIMA)

Additional information on these activities and standards can be found at: https://www.oasis-open.org/standards

4.7 Transaction Processing Performance Council (TPC)

The TPC defines transaction processing and database benchmarks to provide objective, verifiable TPC performance data to industry.

Typically, the TPC produces benchmarks that measure transaction processing and database performance in terms of how many transactions a given system and database can perform per unit of time, e.g. transactions per second or transactions per minute.

The lack of easily verifiable performance claims and the absence of a neutral industry-wide benchmark for Big Data has led the TPC to create a Big Data Working Group (TPC-BDWG) tasked with developing industry standards for benchmarking Big Data systems.

Additional information on these and other potentially applicable TPC standards can be found at http://www.tpc.org/default.asp

4.8 TM Forum

TM Forum is a global trade association to help enterprises, service providers and suppliers continuously transform to succeed in the digital economy. They are creating the tools, best practice guidance for success in Big Data Analytics.

The following TM Forum activities are relevant to Big Data:

- The Framework provides operational standards, best practices and tools for leveraging Big Data and Analytics.
- TM Forum's IPDR provides the protocol for collecting and managing large volumes of usage data operating across any digital service infrastructure.

Additional information on these and other TM Forum standards can be found at http://www.tmforum.org

5 Market estimation

Many research institutions have chosen Big Data as a core technology of ICT and published the market forecast. Because Big Data business is in its embryonic stage now, there is a deviation in scale as Figure 1.

- Big data are a top business priority and drives enormous opportunity for business improvement. Wikibon's own study projects that big data will be a $50 billion business by 2017[10].
- Market research firm IDC has released a forecast that shows the big data market is expected to grow from $3.2 billion in 2010 to $16.9 billion in 2015[11].

IDC a leading market research firm estimates the big data market to exceed 16.1 billion in 2014. Included in this figure are infrastructures (servers, storage, etc., the largest and fastest growing segment at 45 % of the market), services (29 %) and software (24 %) [8][9]. Another analyst with the Wikibon project estimates the market to grow to over $50.1 billion by 2017 and breaks out that growth as shown below[10].

A McKinsey&Company 2011 report estimated that in the developed economies of Europe, government administrators could save more than $149 billion in operational efficiency improvements alone by using big data, not including using big data to reduce fraud and errors and boost the collection of tax revenues[12].

As stated in a 2013 TechAmerica Foundation report US Federal IT officials say real-time analytics of Big Data can help the government cut at least 10 percent ($380 billion) annually from the federal budget or more[12].

Standards for big data while critical to the success of each of these growth areas are even more essential to the interoperability between those areas. Just as the SQL standard enables interoperability between relational databases similar standards are necessary to support interoperability between Infrastructure Software, Applications,

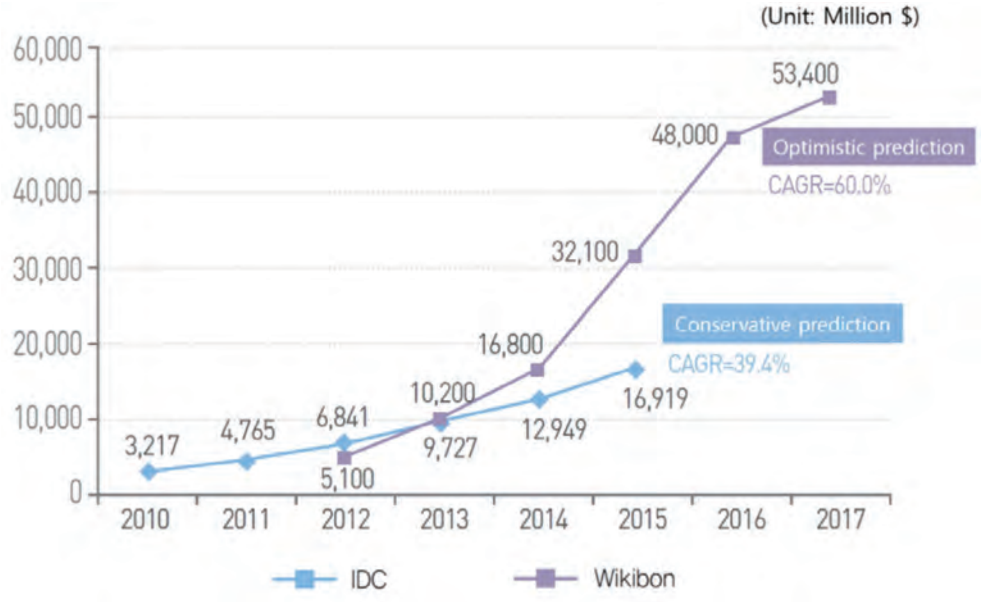


Figure 1 — Big Data market forecast

and Analytics and the underlying Compute, Storage, and Networking resources. For example, Infrastructure Software used to implement distributed file systems is able to function across a wide range of compute and storage resources due to the existing POSIX standards. Similarly, query languages and standard APIs will allow applications and analytics to function across a range of infrastructure services. In addition, extensions to existing standards will enable better interoperability. For example, APIs to query and control data replication and distribution on distributed file systems may be appropriate additions to the POSIX standards.

6 Potential gaps in standardization

This clause discusses potential gaps in Big Data standardization. Its goal is to describe broad areas that may be of interest to JTC 1 versus specific areas and issues described in Clause 7 below.

The identified gaps in standardization activities related to big data are in the following areas:

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

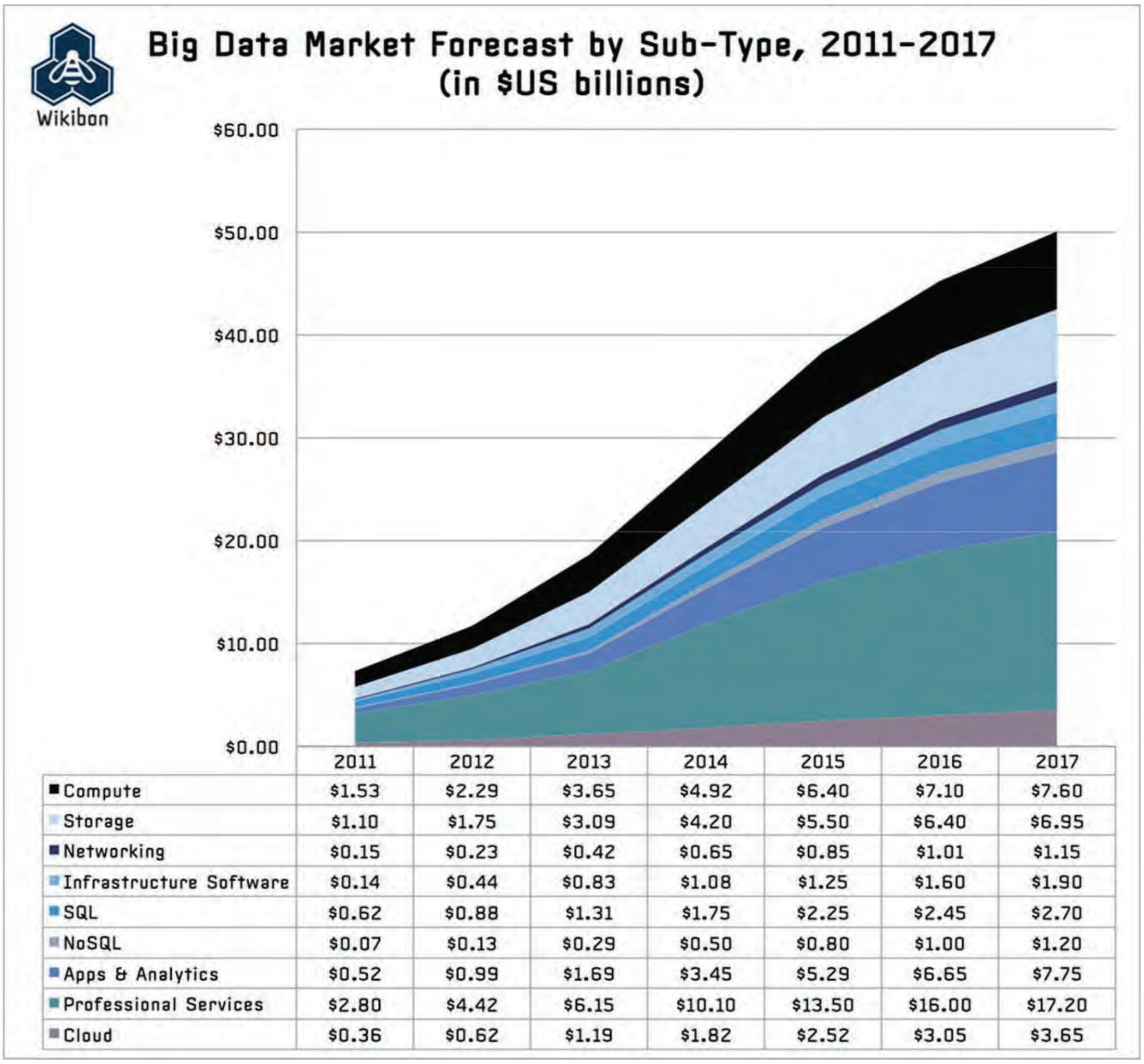
b) Specifications and standardization of metadata including data provenance;

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;



	2011	2012	2013	2014	2015	2016	2017
■ Compute	$1.53	$2.29	$3.65	$4.92	$6.40	$7.10	$7.60
■ Storage	$1.10	$1.75	$3.09	$4.20	$5.50	$6.40	$6.95
■ Networking	$0.15	$0.23	$0.42	$0.65	$0.85	$1.01	$1.15
■ Infrastructure Software	$0.14	$0.44	$0.83	$1.08	$1.25	$1.60	$1.90
■ SQL	$0.62	$0.88	$1.31	$1.75	$2.25	$2.45	$2.70
■ NoSQL	$0.07	$0.13	$0.29	$0.50	$0.80	$1.00	$1.20
■ Apps & Analytics	$0.52	$0.99	$1.69	$3.45	$5.29	$6.65	$7.75
■ Professional Services	$2.80	$4.42	$6.15	$10.10	$13.50	$16.00	$17.20
■ Cloud	$0.36	$0.62	$1.19	$1.82	$2.52	$3.05	$3.65

Figure 2

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;

Refer to Annex A for some specific suggestions relevant to JTC 1 SCs.

7 JTC 1 prospective standardization areas and issues

Clause 4 discusses current initiatives to create Big Data related standards. However, these activities are at an early stage, are narrowly focused and do not have a global in perspective. Many of these activities are being carried out by special interest groups and industry consortia and are not at a national body level so far.

Many of the Big Data issues and gaps listed in Clause 6 are within the scope of existing JTC1 SCs. Annex A documents a mapping of these gaps with the scope and current activities of existing SCs (the numbers in brackets in the last column correspond to the bullet numbers listed in Clause 6). Annex B documents several topics that do not fall within the scope of any existing SC but are nevertheless relevant to Big Data standardization.

From a JTC 1 perspective, it is appropriate to identify new work items and to develop the international standards for Big Data with a more global point of view under the NB-level. In particular, it is desirable to consider the standards for adoption of Big Data in various public sectors such as e-Government. It is also appropriate to consider liaisons with other relevant SDOs.

Annex A

Mapping table between SCs and Big Data issues

This Annex provides a mapping between standard gaps identified in this report and existing JTC 1 SCs.

SC No.	The title of SCs	Current scope and activities relevant to big data	Suggestions for future lines of investigation
SC 6	Telecommunication and information exchange between systems	— Networking technologies	— Standards and protocols for efficient transfer of Big Data (7)
SC 22	Programming Languages	— Potential new language for Big Data applications	— Domain-specific languages (5)
SC 24	Computer graphics, imaging processing, and environmental data representation	— Potential new methods of presenting data	— Visualization in Big Data analytics (13)
SC 27	IT Security techniques	— Big Data creates a large number of Security and privacy issues	— Metadata and provenance standards (2,9)

SC No.	The title of SCs	Current scope and activities relevant to big data	Suggestions for future lines of investigation
SC 32	Data management and interchange	— Database languages and systems related to Big Data	— Definition of standard interfaces (e.g. language, API) to support non-relational datastores (4) — Definition of SQL extension to support exchange and integration between SQL and non-SQL datastores (11, 15) — Metadata and provenance standards (2,9) — SQL and NoSQL standards for data mining (10) — Support for large complex data structures in SQL and/or SQL/MM (4,11) — Support for operations on complex data structures and defined operations on such structures (e.g. add, multiply union) (4,5) — Standards for eventual consistency and acceptable consistency (6) — Support for massive parallelism (10) — Definition and registration of application and processing models (3) — Representation of Big Data Veracity and Quality description and management attributes (16)
SC 34	Document description and processing languages	— A large number of the descriptions and processing languages supported by SC 34 are leveraged in Big Data systems and architectures	— Scalability of these languages and implementations (5) — General and domain specific ontologies. Taxonomies for describing data semantics including ontology interoperation (8)
SC 38	Distributed application platforms and services (DAPS)	— SOA, Web Services, Cloud Computing	— Standards for horizontal scalability (10) — Security, privacy and access controls for distributed file systems (9) — Standards for data replication and distribution (6) — Standards for identification and access to distributed object stores; APIs to access data and attributes (2,12)

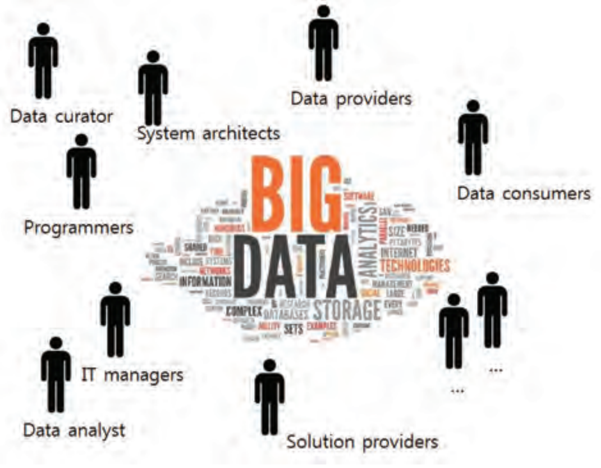
SC No.	The title of SCs	Current scope and activities relevant to big data	Suggestions for future lines of investigation
SC 39	Sustainability for and by Information Technology	— Resource efficient data centre and green ICT	— Measurement for energy cost of Big Data (14)

Annex B

Topics and technologies relevant to Big Data standardization

This Annex identifies activities that are not in the scope of existing JTC 1 SCs. The following tables identify two recommended activities to be undertaken by the new JTC 1 Working Group recommended in this report.

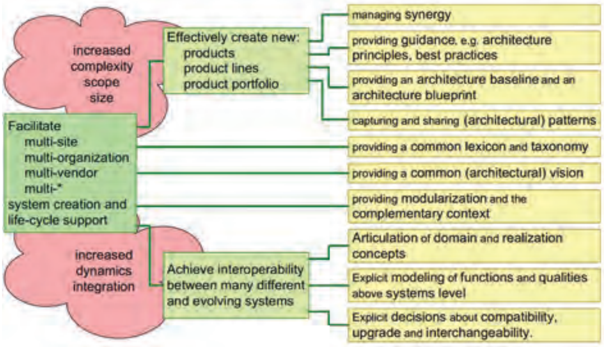
B.1 Definition and vocabulary of Big Data

Title	Description
Goals and objectives	Definition and vocabulary of Big data
Topic description	"Big data" is being defined and classified in a variety of different ways by industry, academia, and research institutes. This is a field in flux, and different people may have different conceptions of the meaning of terms. Therefore, the use of standardized terminologies is essential to clearly and accurately communicates with each other. These definitions and controlled vocabulary should be broad enough to support the full range of stakeholders in a Big Data implementation.
High-level figure describing the use case(with actors)	 Data curator System architects Data providers Data consumers Programmers BIG DATA IT managers Data analyst Solution providers ... ...
Big Data characteristics	— This effort should result in clear definitions of all the Big Data characteristics.

Title	Description
Related standards and standardization activities	— Relevant International Standards: — ISO 704:2009, Terminology work — Principles and methods — ISO 860:2007 Terminology work — Harmonization of concepts and terms — ISO 10241-1:2011, Terminology entries in standards — Part 1: General requirements and examples of presentation — ISO 10241-2:2011, Terminology entries in standards — Part 2: Adoption of standardised terminological entries — ISO/IEC 17788:2014 Information tTechnology — Cloud computing — Overview and vocabulary — Standardization activities: — None
Standardization issues and priority	Candidate/Preliminary new work items with a priority(high, normal, low) — N0025, KNB Proposal for NWIP on Big Data Definition and Vocabulary
Related "N" documents of **SGBD**	— N0028: NBD-PWG_Volume 1 - NIST Big Data Definition V1.0 Draft Pre-release — N0029: NBD-PWG_Volume 2 - NIST Big Data Taxonomies V1.0 Draft Pre-release

B.2 Big Data Reference Architecture

Title	Description
Goals and objectives	Providing a baseline architecture and blueprint for Big Data
Topic description	There are many architectures, frameworks or ecosystems associated with Big Data already. It is difficult to compare them to each other or for them interoperate, due to the different points of view and level of details. For this reason, a standardized reference architecture is needed that will facilitate a shared understanding across multiple products, organizations, and disciplines about current architectures and future direction. The high level view of reference architecture will also provide a framework for understanding how Big Data complements and differs from existing analytics, Business Intelligence, databases and systems. A reference architecture is needed with the aim to achieve the following: A. Illustrate the various Big Data components, processes, and systems to establish a common language for the various stakeholders B. Provide a technical reference for the industry to understand, discuss, categorize, and compare Big Data solutions C. Encourage adherence to common standards, specifications, and patterns by facilitating the analysis of candidate standards for interoperability, portability, reusability, and extendibility The resultant reference architecture needs to be applicable to a variety of business environments including tightly-integrated enterprise systems, as well as loosely-coupled vertical industries that rely on the cooperation of independent stakeholders.

Title	Description
High-level figure describing the use case(with actors)	**Reference Architectures; Why, What and How**  increased complexity scope size Effectively create new: products product lines product portfolio Facilitate multi-site multi-organization multi-vendor multi-* system creation and life-cycle support increased dynamics integration Achieve interoperability between many different and evolving systems managing synergy providing guidance, e.g. architecture principles, best practices providing an architecture baseline and an architecture blueprint capturing and sharing (architectural) patterns providing a common lexicon and taxonomy providing a common (architectural) vision providing modularization and the complementary context Articulation of domain and realization concepts Explicit modeling of functions and qualities above systems level Explicit decisions about compatibility, upgrade and interchangeability. White Paper, Architecture Forum, 2007.
Big Data characteristics	The reference architecture should align with the full range of Big Data characteristics.
Related standards and standardization activities	— Relevant Standards: — ISO/IEC 17789 Information Technology - Cloud Computing - Reference Architecture — Related Standardization activities: — ITU-T SG 13, Y.BigData-reqts “Requirements and capabilities for cloud computing based big data”
Standardization issues and priority	Candidate/Preliminary new work items with a priority(high, normal, low) — N0026: KNB Proposal for NWIP on Big Data Reference Architecture (high)
Related “N” documents of SGBD	— N0032: NBD-PWG_Volume 5 - NIST Big Data Architecture White Paper Survey V1.0 Draft Pre-release — N0033: NBD-PWG_Volume 6 - NIST Big Data Reference Architecture V1.0 Draft Pre-release

Bibliography

[1] DRAFT NIST Big Data Interoperability: Volume 1: NIST Big Data Definitions, Version 1.0 April 23, 2014

[2] DRAFT NIST Big Data Interoperability: Volume 2: NIST Big Data Taxonomies, Version 1.0 April 23, 2014

[3] DRAFT NIST Big Data Interoperability: Volume 3: NIST Big Data Use Case & Requirements, Version 1.0 April 23, 2014

[4] DRAFT NIST Big Data Interoperability: Volume 4: NIST Big Data Security and Privacy Requirements, Version 1.0 April 23, 2014

[5] DRAFT NIST Big Data Interoperability. Volume 5: NIST Big Data Architectures White Paper Survey, Version 1.0 April 23, 2014

[6] DRAFT NIST Big Data Interoperability: Volume 6: NIST Big Data Reference Architecture, Version 1.0 April 23, 2014

[7] DRAFT NIST Big Data Interoperability: Volume 7: NIST Big Data Technology Roadmap, Version 1.0 April 23, 2014

[8] White paper, "Big Data Meets Big Data Analytics", by SAS, June 2012

[9] Wikipedia. http://en.wikipedia.org/wiki/Big_Data

[10] Free resources for technology professionals, http://www.techtarget.com/html/it_pro_list.html

[11] "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data", June 27, 2011

[12] Draft Recommendation I.T.U.-T.Y. Bigdata-reqts, *Requirements and capabilities for cloud computing* based *big data.* Available at http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=9853

[13] SINTEF. 2013, May 22). Big Data, for better or worse: 90% of world's data generated over last two years. ScienceDaily. Retrieved April 8, 2014 from http://www.sciencedaily.com/releases/2013/05/130522085217.htm

[14] IDC IVIEW. 2012, Dec). THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East Retrieved April 8, 2014 from http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

[15] IDC, Big Data Predictions 2014: Beyond Irrational Exuberance - Opportunities in the Big Data and Analytics Markets Web Conference Dec 11, 2013.Framingham, MA United States

[16] Wikibon Project, Big Data Vendor Revenue and Market Forecast 2013-2017, Feb 2014, Retrieved April 8, 2014 from http://wikibon.org/wiki/v/Big_Data_Vendor_Revenue_and_Market_Forecast_2013-2017

[17] IDC. Worldwide BigData Technology and Services 2012. Forecast. 2015, **2012** p. 3

[18] McKinsey&Company, Big Data: The next frontier for innovation, competition, and productivity, 2011

[19] TechAmerica Foundation. Big Data. Can Save Money and Lives Say Government IT Officials, 2013, pp. 2.

[20] OGC BigData DWG. http://external.opengeospatial.org/twiki_public/BigDataDwg/, 2014

International Organization for Standardization

ISO Central Secretariat
Ch. de Blandonnet 8
CP 401
CH – 1214 Vernier, Genève
Switzerland

iso.org

International Electrotechnical Commission

IEC Central Office
3, rue de Varembé
P.O. Box 131
CH – 1211 Genève 20
Switzerland

iec.ch