ISO/IEC JTC 1/SC 32 N 2388b

Date: 2013-06-05

REPLACES: –

ISO/IEC JTC 1/SC 32

Data Management and Interchange

Secretariat: United States of America (ANSI)
Administered by Farance Inc. on behalf of ANSI

DOCUMENT TYPE	Officer's Contribution (Contribution from Chairman, Convener, Rapporteur, etc.)
TITLE	Report of study group on next generation analytics and big data
SOURCE	Keith Hare - study group rapporteur
PROJECT NUMBER	1.32.
STATUS	preliminary report of study group on Next Generation Analytics and Big Data to SC32 and to JTC1 Special Working Group on Planning
REFERENCES	
ACTION ID.	FYI
REQUESTED ACTION	
DUE DATE	--
Number of Pages	20
LANGUAGE USED	English
DISTRIBUTION	P & L Members SC Chair WG Conveners and Secretaries

Dr. Timothy Schoechle, Secretary, ISO/IEC JTC 1/SC 32
Farance Inc *, 3066 Sixth Street, Boulder, CO, United States of America
Telephone: +1 303-443-5490; E-mail: Timothy@Schoechle.org
available from the JTC 1/SC 32 WebSite http://www.jtc1sc32.org/
*Farance Inc. administers the ISO/IEC JTC 1/SC 32 Secretariat on behalf of ANSI

1 Executive Summary

This is a preliminary report of on-going discussions being conducted in the SC32 (Data Management and Interchange) *ad hoc* Study Group on *Next Generation Analytics and Big Data*. It is directed to the JTC1 Special Working Group on Planning, and is primarily intended to provide an indication of the range of discussion in SC32 and to provide background and context. This study was initiated by JTC1 Resolution 33 from its San Diego plenary meeting (2011-11-12) which asked SC32 to consider opportunities in the area of "next generation analytics" and "social analytics." The study group was initiated at the SC32 plenary meeting in Berlin (2012-04-08) and continued at the recent SC32 plenary meeting in Gyeongju (2013-05-31).

1.1 Purpose of this study of Data Analytics and Big Data

The initial interest at JTC1 resulted from a Gartner report on the *Top 10 Strategic Technologies for 2012* (2011-10-18). Since that time, the topic has greatly expanded in the IT industry and the public media under the general theme of "Big Data". The emergence of this theme may be attributed to several causes, among which are the dramatic drop in data storage costs, the popularity of mobile Internet devices, the rise of social networking applications, and the availability of a proliferation of consumer data.

A result has been the rapid emergence of a plethora of application developers seeking to monetize the data by a variety of techniques including novel database and analytical approaches. It is not clear where all this is going or how it relates to established database and metadata technologies, but SC32 is actively engaged in studying the topic.

1.2 Summary of this report

The report begins with an *Introduction* to the immediate perspective of the SC32 experts. A *Background* is provided describing the history of the report and of the Analytics and Big Data topics in general. A third section describes the basic *Concepts* involved, then sections on *Existing Standards*, and on *Potential Standards* are provided. Finally some preliminary *Recommendations* are offered, both to SC32 and to JTC1. Additionally, sections are appended for *References* and *Definitions*, as well as *Big Data Life Cycle* proposal, and list of active *Participants* in the SC32 *Ad Hoc.*

1.3 Conclusion

The reader needs to recognize that this report is a very preliminary snapshot and that the topic is a moving target. There is understandably a lot of hype in the industry at this early stage and much posturing and promotion by those seeking to open new markets for their products. We can assume that more data is coming soon and that new applications for Big Data and Analytics will emerge over the coming year. We can also expect that a number of problems will emerge, including those in areas of consumer privacy and security. Also we can expect challenges in competing or conflicting formal and non-formal standards, and in potential overlap among various standards activities. Some of these may include cloud computing, sensor networks, smart grid, "Internet of things", and other emerging activities that relate to the creation and analysis of Big Data.

SC32 has renewed the Next Generation Analytics Study Group for another year and expects the study group to continue to track these on-going changes and enhance this report.

2 Introduction

Next generation analytics and social analytics cover a very wide range of requirements in terms of both the kinds of analysis and the kinds of data to which it is applied. It is appropriate (for SC32) to consider standardizing any of these range of approaches to the analytics problem. However, SC32 has extensive experience in the standardization of database languages, utilities and metadata models that support the people who are responsible for the development of systems for analytics.

The standardization of interfaces to support such tools promote the interoperability of components and thereby aim to make the development of analytic solutions easier/simpler/quicker for the developer and to make their operation more efficient and reliable.

As Gartner [Gartner] points out, analytics is developing to include on-line, real-time, analysis, the integration of historical data with current data to provide predictions and will evolve to include more complex data types into the analysis as well as the involvement of multiple actors both human and non-human to allow collaborative analysis and brainstorming.

SC32 has traditionally supported analysis efforts and has over time developed the following mechanisms:

- the OLAP functionality introduced and steadily enhanced in the SQL language,
- the recent introduction of support for bi-temporal data into the SQL language which is attracting considerable interest within the data warehousing community,
- the definition of complex data types and their associated operations,
- the metadata registries and data representation work of WG2,
- the ability to integrate real-time data streams with other data as defined in the Management of External Data (ISO/IEC 9075-9).

SC32 is already progressing proposals which are also relevant to analytics such as:

- row pattern recognition in the SQL language
- array data

Areas of where further opportunities may arise include, but are not confined to, the use of ontologies to assist in the integration of data from disparate sources that span multiple related domains into a single analysis and to enable the user to discover and correlate the data, the standardization of additional vocabularies and/or data representations, the standardization of more complex data types to assist in the integration of these into the analysis, the extension of the statistical facilities of the SQL language.

SC32, which is clearly involved in this area, has taken the opportunity implied in Resolution 33 to study the opportunities for standardization in the areas of Next Generation Analytics, Social Analytics, and Big Data in a more holistic manner than has been the situation up to now.

This report is the current state of the SC32 study of *Next Generation Analytics and Big Data*. It uses several conceptual models to identify the interfaces between pieces of the problem, documents the existing SC32 standards efforts related to those interfaces, and outlines several potential areas for additional work both within SC32 and in cooperation with other JTC1 groups.

3 Background

3.1 History resulting in this report

[SC32N2181], Resolution 33 – Potential Technology Areas for JTC 1 Subcommittees, says:

> JTC 1 thanks those SCs that have responded to Resolution 29 of the Belfast Plenary and provided input on potential new technology areas. Referring to JTC 1 N10513 and JTC 1 N10722, JTC 1 requests that:
>
> 1) JTC 1/SC 32 consider whether to formally study opportunities for JTC 1 in the area of 'Next Generation Analytics'
> 2) JTC 1/SC 32 consider whether to formally study opportunities for JTC 1 in the area of 'Social Analytics', especially metadata aspects
> 3) …
> 4) …
> 5) …
>
> JTC 1 requests that each SC and JTC 1 WG listed above submit reports on the above to SWG on Planning by 1 June 2012.

At the June, 2012 SC32 Plenary in Berlin, the SC32 Chair, Jim Melton, appointed an Ad Hoc to respond to the JTC1 request. The Ad Hoc was made up of participants from all four SC32 working groups:

- WG1 – E-business
- WG2 – Metadata
- WG3 – Database Languages
- WG4 – Multimedia

At the 2012 closing plenary, a Study Group was assigned to progress the discussions, but made little progress.

At the May, 2013 Plenary in Gyeongju, Korea, the Study Group/Ad Hoc was urged to make further progress.

3.2 Next Generation Analytics and Big Data

“Next Generation Analytics” is a frequently used term with a fluid definition. [SC32N2198] links next generation analytics to “Big Data”:

> (6) Next-Generation Analytics → link to „Big Data (7), …“
> (1) From structured data and traditional analytics to analysis of complex information of many types
> (2) Shift to cloud and exploit cloud resources
> (7) Big Data → link to JTC1/SC 06 Telecommunications and information exchange between systems, JTC1/SC 32 Data management and interchange, JTC1/SC 39 Sustainability for and by Information Technology, …
> (1) Manage of large data volume, DBMS,
> (2) Management of multiple sources

Social Analytics basically pertains to applying "analytic" approaches and tools to "social media" information (e.g., that found on Facebook, You Tube, Twitter, etc.). Note that where the "big data" and/or "analytics" involve recorded information on or about identifiable individual external constraints of a public policy nature apply (e.g., consumer protection, privacy protection, etc.) Here in a data management and interchange context and from a JTC1/SC32 perspective, [PrivacyRequirements], "..privacy protection..", identifies requirements of this nature and are applicable, especially where any activity of a big data or analytic nature relates to business transactions.

3.3 Gartner

Since the original request from JTC1 referenced a report by the US industry analyst Gartner Group, it is useful to review what Gartner has to say. [Gartner] calls out both "Next-Generation Analytics" and "Big Data" as Strategic Technologies for 2012:

> **Next-Generation Analytics.** Analytics is growing along three key dimensions:
>
> > From traditional offline analytics to in-line embedded analytics. This has been the focus for many efforts in the past and will continue to be an important focus for analytics.
> >
> > From analyzing historical data to explain what happened to analyzing historical and real-time data from multiple systems to simulate and predict the future.
> >
> > Over the next three years, analytics will mature along a third dimension, from structured and simple data analyzed by individuals to analysis of complex information of many types (text, video, etc…) from many systems supporting a collaborative decision process that brings multiple people together to analyze, brainstorm and make decisions.
>
> Analytics is also beginning to shift to the cloud and exploit cloud resources for high performance and grid computing.
>
> In 2011 and 2012, analytics will increasingly focus on decisions and collaboration. The new step is to provide simulation, prediction, optimization and other analytics, not simply information, to empower even more decision flexibility at the time and place of every business process action.
>
> **Big Data**. The size, complexity of formats and speed of delivery exceeds the capabilities of traditional data management technologies; it requires the use of new or exotic technologies simply to manage the volume alone. Many new technologies are emerging, with the potential to be disruptive (e.g., in-memory DBMS). Analytics has become a major driving application for data warehousing, with the use of MapReduce outside and inside the DBMS, and the use of self-service data marts. One major implication of big data is that in the future users will not be able to put all useful information into a single data warehouse. Logical data warehouses bringing together information from multiple sources as needed will replace the single data warehouse model.

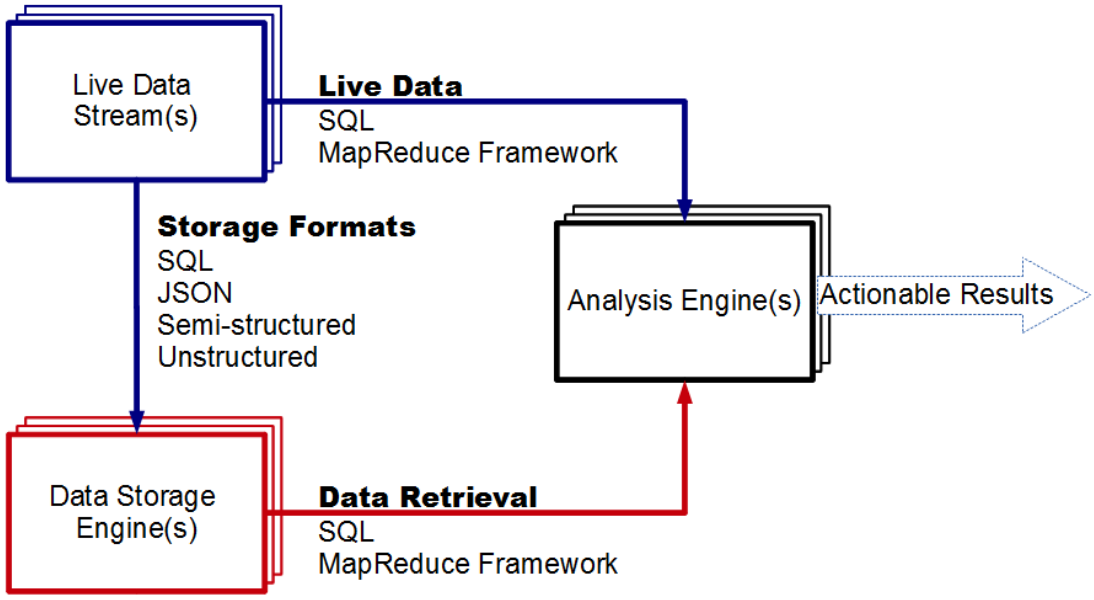
These descriptions are not particularly helpful in limiting the scope so the next section presents conceptual models at several levels of detail.

4 Conceptual Models

Conceptual models provide a framework for discussing “Next Generation Analytics”, its component pieces, the interfaces between those pieces, where applicable standards currently exist, and where there is room for additional standardization.

4.1 High Level Conceptual Model

The following high-level conceptual model provides a



To isolate the data storage engines from the data source (producer) and data retrieval (consumer) a separate dispatch layer can be introduced. This layer stores information about the location of the data.

4.1.1 Data Streams

Data about something comes from somewhere and is passed directly to the Analysis engine(s) and/or to the data storage engine(s). This description is intentionally vague to emphasize the concept that the data can be any variety of data.

4.1.2 Data Storage Engine

The data storage engine(s) need accept data in a variety of formats, structured, semi-structured, or unstructured, preserve that data as long as needed, retrieve the data, and present it to the analysis engine(s).

The data storage engine/layer needs to have the following characteristics:

- Scalable
- Distributed
- Dynamic distribution/partitioning

The storage engines can consist of one of more of the following types:

- SQL Database(s)
- NoSQL Database(s)
- Graph Database(s)

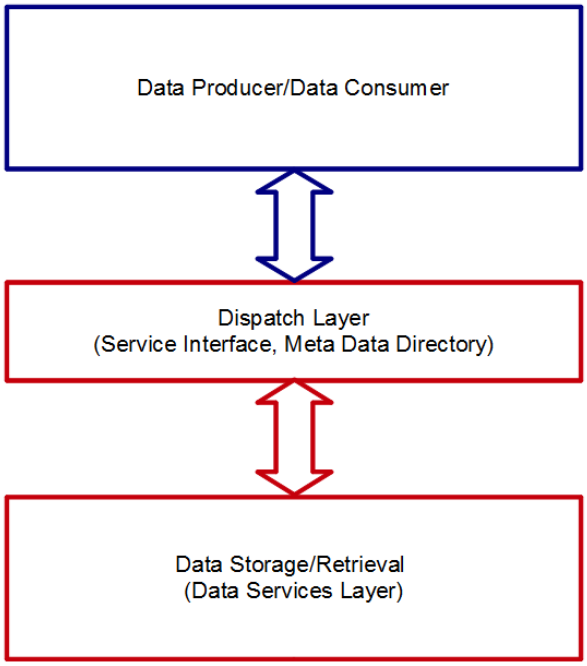
- Something else

Much of the current discussion of Next Generation Analytics and Big Data focuses on details of the data storage engine because the applications that store and retrieve the data are intertwined with the details of the physical storage.

The conceptual model in [SQL/FOUNDATION] isolates the storage and retrieval language from the physical storage of the data. This allows the underlying storage and access data structures to be tuned (manually or automatically) without disrupting the code that stores and retrieves the data.

The data storage engines need to accept data in a variety of structured, semi-structured and unstructured formats, including (but not limited to) SQL, XML, and Java Script Object Notation (JSON). JSON is specifically mentioned here because it is referenced in a number of instances as if it were faster/quicker/better than other data representations. From the point of view of this conceptual model, JSON data is just another representation that can be accommodated in a standard way with a small amount of effort.

If the storage and retrieval model is abstracted a bit more, a dispatch layer can be introduced between the data producer/consumer and the actual storage:



This dispatch layer needs to provide a metadata directory that has knowledge of what data is where, and stores and retrieves the data as requested by the consumer. Conceptually, this sort of layer exists in any application. In some applications, it exists in the same code module as the application logic. In other applications, greater separation exists.

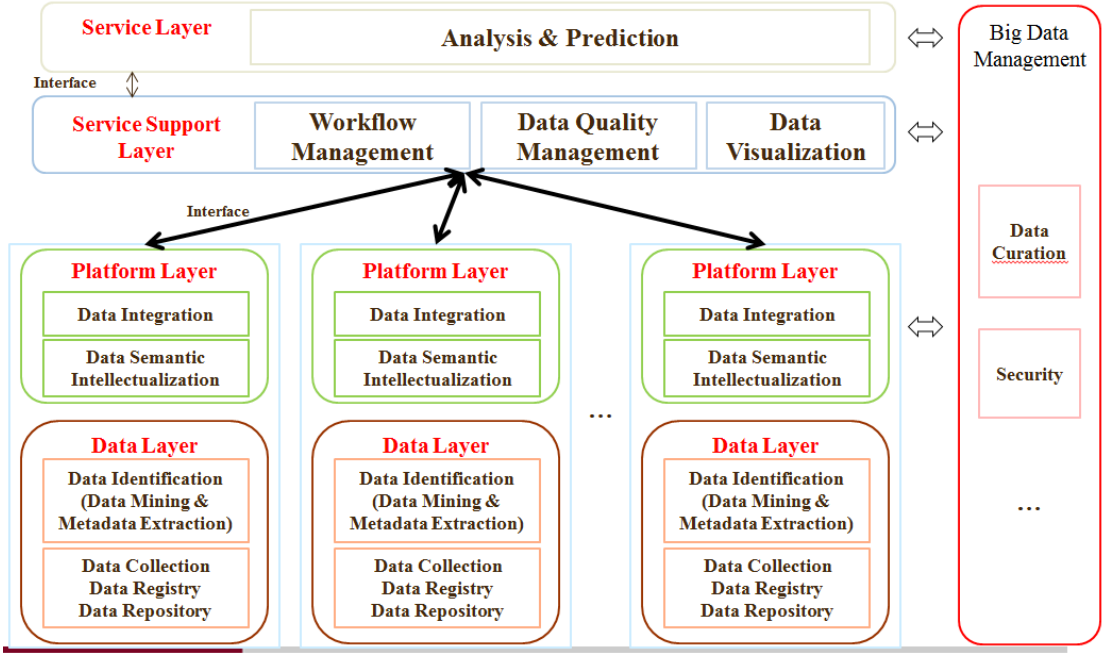
The dispatch layer/service interface model could coexist with and benefit from the Service Oriented Architecture work being done in SC38.

4.1.3 Analysis Engine(s)

In this model, the Analysis engine is depicted as a black box that accepts a wide variety of data and produces some sort of actionable results. The analysis engine could be a machine learning engine, a statistics package (Project R or a proprietary package such as SPSS or SAS), perhaps coupled with data visualization tools. These tools will benefit from a standard interface that allows them to sufficiently understand the data and metadata being accessed.

4.2 A More Detailed Conceptual Model

[NGA-REFMOD] proposes a more detailed reference model for Big Data that adds additional important aspects and identifies standards originating in SC32 that address components of the model.



A number of existing SC32 standards address pieces of this model:

- Data Interface: 9075 – Database Language SQL
- Data Registry: 11179 – Metadata Registry
- Data Mining: 13249 part 6 – SQL/MM Data Mining
- Service Layer: 19763 – Metamodel framework for interoperability (MFI)

4.3 Data Visualization

Presenting complex data in such a way that a human user can understand and act on that data is a complex specialized problem. This is outside the scope of working currently being done in SC32. SC32 needs to identify groups within JTC1 that are working in the area of data visualization and coordinate so that the necessary interfaces are identified and standardized.

4.4 Metadata

Many of the tools currently associated with Big Data projects are touted as “schema-less.” The claimed benefit of schema-less is that application efforts do not have to take the time up front to articulate a schema design. In practice, this usually means that the schema or metadata is implemented in the application or call interface built specifically for the project or even as the data is stored. The challenge comes when data from multiple data stores must be integrated.

The SC32 WG2 Metadata Registry work (ISO/IEC 11179) provides mechanisms for articulating and registering Metadata that can be queried and then incorporated into other applications.

The SC32 WG3 Database Language SQL work (ISO/IEC 9075) incorporates structural metadata for database tables and objects. This 9075 metadata can queried to support dynamic generation of code to store, modify, and retrieve.

There are opportunities in these areas to better integrate metadata work as well as challenges in articulating the long term benefits and cost savings of utilizing the capabilities included in the SC32 standards. In particular, efforts are needed to:

- Integrate the production of metadata that assists with discovery, access, integration and interpretation of the data
- Ensure that metadata is readily accessible
- Support the ability of tools to be able to read data by reading its metadata
- Support the inclusion of provenance in metadata to enable users of the data to evaluate 'trust' or veracity
- Support standardization of core information models that are integrated at a high level, and can be easily extended to enable data to be aggregated and interpreted across disparate domains.

4.5 Storage and Retrieval of Complex, Semi-structured, and Unstructured Data

The SC32 WG3 work in the ISO/IEC 9075 family of standards provides extensive capabilities for the storage and retrieval of structured data supporting relational tables, user defined objects, and XML data. The SC32 WG4 work in the ISO/IEC 13249 family of standards adds support for more complex datatypes such as spatial, full text, and images.

4.6 Data Privacy

Privacy protection pertains to a set of external constraints and requirements of laws and regulation of jurisdictional domains which apply whenever the “data” pertains to an identifiable individual, i.e. “personal information” on data (of any kind) legal and regulatory. Privacy protection requirements apply to personal information of any kind and from any source irrespective of the technology used for its creation, processing, communications, or medium on which it is recorded, etc.

JTC1/SC32/WG1 work incorporates privacy protection requirements in use of EDI, i.e., where personal information is recorded and interchanged as electronic data and thus involves EDI. Where any “big data” and “analytics” implementation involve either EDI and/or personal information, the existing JTC1/SC32/WG1-eBusiness standards will be of use, and especially the ISO/IEC 15944-8 “…privacy protection” standard should be referenced and used.

Implementation of “Big Data and Analytics” which involve the creation or “capture” of personal information will be well served to use the ISO/IEC 15844-8 standard for a number of reasons including avoidance of public push-back of a “Big Brother” fear nature, sanctions by Privacy or Data Commissioners in jurisdictional domains, possible lawsuits Including those of a class action nature, etc.

4.7 Data Security

Standardization pertaining to the provisioning of security services and techniques is outside the domain of JTC1/SC32 standards development. SC32 work includes support for security restrictions although in practice these restrictions are often implemented at application or operating system level. Whenever possible, support for Analytical and Big Data applications needs to incorporate the work on security techniques from JTC1 SC27.

5 Existing Standards

In general, standards will have the most benefit in the interfaces between other pieces. The Next Generation Analytics and Big Data models discussed in this document contain a large number of interfaces, leaving room for a variety of types of standards. The ISO/IEC JTC 1/SC 32 Standards in these areas are:

5.1 Metadata

The aim of the metadata standards developed within Working Group 2 of SC32 is to provide an information and software services infrastructure that supports communities that wish to interoperate.

The ISO/IEC 11179 series of standards provides specifications for the structure of a metadata registry and the procedures for the operation of such a registry. These standards address the semantics of data (both terminological and computational), the representation of data, and the registration of the descriptions of that data. It is through these descriptions that an accurate understanding of the semantics and a useful depiction of the data are found. These standards promote:

- Standard description of data
- Common understanding of data across organizational elements and between organizations
- Re-use and standardization of data over time, space, and applications
- Harmonization and standardization of data within an organization and across organizations
- Management of the components of data
- Re-use of the components of data

The ISO/IEC 20943 series of technical reports provide a set of procedures for achieving consistency of content within a metadata registry.

The ISO/IEC 19763 series of standards provides specifications for a metamodel framework for interoperability. In this context interoperability should be interpreted in its broadest sense: the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units (ISO/IEC 2382-1:1993). ISO/IEC 19763 will eventually cover:

- A core model to provide common facilities
- A basic mapping model to allow for the common semantics of two models to be registered
- A metamodel for the registration of ontologies
- A metamodel for the registration of information models
- A metamodel for the registration of process models
- A metamodel for the registration of models of services, principally web services
- A metamodel for the registration of roles and goals associated with processes and services
- A metamodel for the registration of form designs

The ISO/IEC 19763 series of standards will also include a technical report describing on-demand model selection based on roles, goals, processes and services and a standard for a registry of registries.

5.2 Data Storage and Retrieval

SC32 WG3 has defined the SQL database language in the 9075 family of standards. Additional work is needed to needed to accommodate new types of objects such as JSON. Additional work may be needed in call level interfaces (CLI) to better support distributed access

- ISO/IEC 9075-1:2011 Information technology -- Database languages -- SQL -- Part 1: Framework (SQL/Framework)
- ISO/IEC 9075-2:2011 Information technology -- Database languages -- SQL -- Part 2: Foundation (SQL/Foundation)
- ISO/IEC 9075-3:2008 Information technology -- Database languages -- SQL -- Part 3: Call-Level Interface (SQL/CLI)
- ISO/IEC 9075-4:2011 Information technology -- Database languages -- SQL -- Part 4: Persistent Stored Modules (SQL/PSM)
- ISO/IEC 9075-9:2008 Information technology -- Database languages -- SQL -- Part 9: Management of External Data (SQL/MED)
- ISO/IEC 9075-10:2008 Information technology -- Database languages -- SQL -- Part 10: Object Language Bindings (SQL/OLB)
- ISO/IEC 9075-11:2011 Information technology -- Database languages -- SQL -- Part 11: Information and Definition Schemas (SQL/Schemata)
- ISO/IEC 9075-13:2008 Information technology -- Database languages -- SQL -- Part 13: SQL Routines and Types Using the Java TM Programming Language (SQL/JRT)
- ISO/IEC 9075-14:2011 Information technology -- Database languages -- SQL -- Part 14: XML-Related Specifications (SQL/XML)

5.3 Support for Complex Data Types

SC32 WG4 has defined standards for complex data storage and retrieval:

- ISO/IEC 13249-2 SQL/MM Part 2: Full Text provides full information retrieval capabilities and complement SQL and SQL/XML. SQL/XML provides facilities to manage XML structured data while MM Part 2 provides contents based retrieval.
- ISO/IEC 13249-3 Part 3: Spatial provides all the functionalities required to support geo applications. Most big data application now includes processing of GPS data together with geographic information. Thus Part 3: Spatial is also one of the key components of big data applications.
- ISO/IEC 13249-5 Part 5: Still Image provides basic functionalities for Image data management.
- ISO/IEC 13249-6 Part 6: Data Mining provides all the functionalities required to support statistical data mining applications. SQL/OLAP functionality provide simple online analytic processing while MM Part 6: provides sophisticated statistical data mining functionalities.

6 Potential Standards Efforts

The SC32 Study Group on Next Generation Analytics have identified the following opportunities for Standards enhancements:

- Review the metadata standards to ensure the required support exists for Analytical and Big Data projects and tools.
- Review the data storage standards to ensure the required support exists for storing and retrieving the volume and diversity of data required by Analytical and Big Data projects and tools.
- Review the support for complex, semi-structured, and unstructured datatypes to ensure the required support exists for storing and retrieving the volume and diversity of data required by Analytical and Big Data projects and tools.
- Review the integration of all SC32 standards to ensure the standards work well together to provide the required support for Analytical and Big Data projects and tools

7 Recommendations to SC32

This work is preliminary and incomplete.

Recommendation 1: SC32 forward this document, SC32N2388 to JTC1 and SWG-P in response to their instruction to SC32. SC32 request its secretary to forward this document immediately to ensure timely input to the June 19-20, 2013 SWG-P meeting.

Recommendation 2: SC32 continue a study group to further develop the models described in this document and the areas of potential standardization relating to Next Generation Analytics, Social Analytics, and the underlying technologies for their support. As a part of their efforts, the study group should establish a document with a variety of Next Generation Analytics/Big Data Use Cases.

8 References

8.1 ISO/IEC JTC 1/SC 32 References

[SC32N2181] ISO/IEC JTC 1/SC 32 N2181, “Resolutions and topics from the recent JTC 1 meeting of particular interest to SC 32 participants”, SC32 Chair – Jim Melton, 16 November 2011.

[SC32N2198] ISO/IEC JTC 1/SC 32 N 2198, “Analysis of 2012 Gartner Technology Trends”, JTC1 SWG-P - Mario Wendt – Convenor, 12 January 2012.

[SC32N2199]ISO/IEC JTC 1/SC 32 N 2199, “Discussion: SC 32 Response to 2011 JTC 1 Resolution 33”, SC32 Chair – Jim Melton, 19 March 2012.

[PrivacyRequirements] ISO/IEC 15944-8:2012 “Information technology- Business Operational View – Part 8: Identification of privacy protection requirements as external constraints on business transaction”, 29 March 2012

[SQL/Foundation] ISO/IEC 9075-2:2011, “Information technology -- Database languages -- SQL -- Part 2: Foundation (SQL/Foundation)”, 15 December 2011.

[SQLMM/DataMining] ISO/IEC 13249-6:2006, “Information technology -- Database languages -- SQL multimedia and application packages -- Part 6: Data mining”, 2006.

[DTR-20943-5] ISO/IEC 20943, Information Technology -- Procedures for Achieving Metadata Registry Content Consistency -- Part 5: Metadata Mapping Procedure (Under development)

[MDRModules] ISO/IEC 19773:2011, “Information technology -- Metadata Registries (MDR) modules”, 1 September, 2011

[NGA-Report-Berlin] ISO/IEC JTC 1/SC32 N2241”SC32 Ad Hoc on Next Generation Analytics”, June 2013

[NGA-Tutorial] ISO/IEC JTC 1/SC32 N2383 “Big Data & Next Generation Analytics – Tutorial”, May 2013

[NGA-REFMOD] ISO/IEC JTC 1/SC32 N2386 “Next Generation Analytics & Big Data (A Reference Model for Big Data)”

8.2 External References

[Gartner] “Gartner Identifies the Top 10 Strategic Technologies for 2012”, http://www.gartner.com/it/page.jsp?id=1826214, 18 October 2011, accessed 4 June 2012.

[BigDataDef] Big Data Definition, Wikipedia, http://en.wikipedia.org/wiki/Big_data

[TechCrunch] Big Data Right Now: Five Trendy Open Source Technologies
http://techcrunch.com/2012/10/27/big-data-right-now-five-trendy-open-source-technologies/

[BigDataNow] “Big Data Now: 2012 Edition”
http://oreilly.com/data/radarreports/big-data-now-2012.csp
The electronic version of this book is free, but requires registration.

[EntBigData] Enterprise Big-data
http://wikibon.org/wiki/v/Enterprise_Big-data

[BigDataDef-4] What is Big Data? 4 Definitions
http://siliconangle.com/blog/2011/09/08/what-is-big-data-4-definitions/

[BigDataApp] Characteristics of Big Data Application Platform
http://nosql.mypopescu.com/post/9849679819/characteristics-of-big-data-application-platform

[Chaudhuri] Chaudhuri, S., "What next?: A half-dozen data management research goals for big data and the cloud", In Proceedings of the 31st Symposium on Principles of Database Systems, ACM, 2012.

[USA-BigData] “Fact Sheet: Big Data Across the Federal Government” March 29, 2012
http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final.pdf

9 Definitions

The following definitions are useful in understanding the arena of Next Generation Analytics and Big Data. Note that these definitions are working definitions have not yet reached sufficient consensus to be considered normative.

9.1 Actionable Results

In general, the results of an analysis needs to be something that the organization or enterprise can use to make decisions. That is, the results are used to direct and drive action.

In many cases, it is critical to narrow time gap between data acquisition and acting on a business decision based on the data.

9.2 Analytics / Prediction

9.3 Big Data

Definition 1

Big Data can be defined by some combination of the following five characteristics:

- Volume – The amount of data is sufficiently large so as to require special considerations.
- Variety – The data consists of multiple types of data potentially from multiple sources. This variety can be a combinations of:
    - Structured data – tables, objects, etc. for which the metadata is well defined
    - Semi-structured data – documents, etc. for which the metadata is contained internally, for example Java Script Object Notation (JSON)
    - Unstructured data – Photographs, video, binary data
- Velocity – the data is produced at high rates, operating on stale data is not valuable
- Value – the data has perceived or quantifiable benefit to the enterprise or organization using it
- Veracity – the correctness of the data can be assessed

Note that this set of characteristics allows the term “Big Data” to be used in a wide variety of ever-changing ways. In fact, this set of characteristics can be used to describe existing efforts to collect, manage, and manipulate data.

Definition 2

Big Data: field of study based on convergence of problems in: (1) irregular or heterogeneous data structures, their navigation, query, and data typing; (2) computation parallelism and its management during deployment or execution; (3) descriptive data and self-inquiry about objects for real-time decision-making; and/or (4) presentation and aggregation of data that exceed visual limitations of a single page

> **Note 1:** Big Data is not necessarily about a large amount of data because many of the concerns can be demonstrated with small (less than gigabyte) data sets. Big Data concerns typically arise in processing large amounts of data because the four main characteristics (irregularity, parallelism, real-time metadata, and/or presentation/visualization) are unavoidable in such large data sets.
> **Note 2:** Computation parallelism issues concern the unit of processing (thread, statement, block, process, node, etc.), contention methods for shared access, and begin-suspend-resume-completion-termination processing.

Note 3: Descriptive data is also known as metadata. Self-inquiry is known as reflection is some programming paradigms.
Note 4: The visual limitations concern how much information a human can usefully process on a single display screen or sheet of paper. For example, the presentation of a connection graph of 500 nodes might require more than 20 rows and columns, along with the connections (relationships) among each of the pairs. Typically, this is too much for a human to comprehend in a useful way. Big Data presentation/visualization issues concern reformulating the information in a way that can be presented for convenient human consumption.

9.4 Collection/Identification

9.5 Data Curation

9.6 Data Quality

Degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions, ISO/IEC 25012:2008(en)

9.7 Data Scientist/Data Engineer

9.8 Integration

Process of physical and functionally combining lower-level data products to produce a new data product.

Process of physically and functionally combining lower-level products (hardware or software) to obtain a particular functional configuration, ISO 10795:2011(en), 1.117

9.9 MapReduce

MapReduce is a programming paradigm that allows for parallel query and retrieval of data that from distributed data storage. Data is retrieved in parallel (Map), processed to integrated the returned data (Reduce) then and presented to the to the data requestor.

Both the Map operation and the Reduce operation are custom programming created using whatever language is required by the underlying data store.

9.10 Registry

Facility for recording and storing object models, interfaces and data concepts within an organised and managed environment, ISO 14813-5:2010(en), B.1.127

9.11 Repository

Physical facility for storing object models, interfaces, and implementations
ISO 14813-5:2010(en), B.1.130

9.12 Semantic Intellectualization

9.13 Schema-less databases, or NoSQL databases

There are several database types that fit into this category, such as key-value stores and document stores, which focus on the storage and retrieval of large volumes of unstructured, semi-structured, or even structured data. They achieve performance gains by doing away with some (or all) of the restrictions traditionally associated with conventional databases, such as read-write consistency, in exchange for scalability and distributed processing.

9.14 Visualization

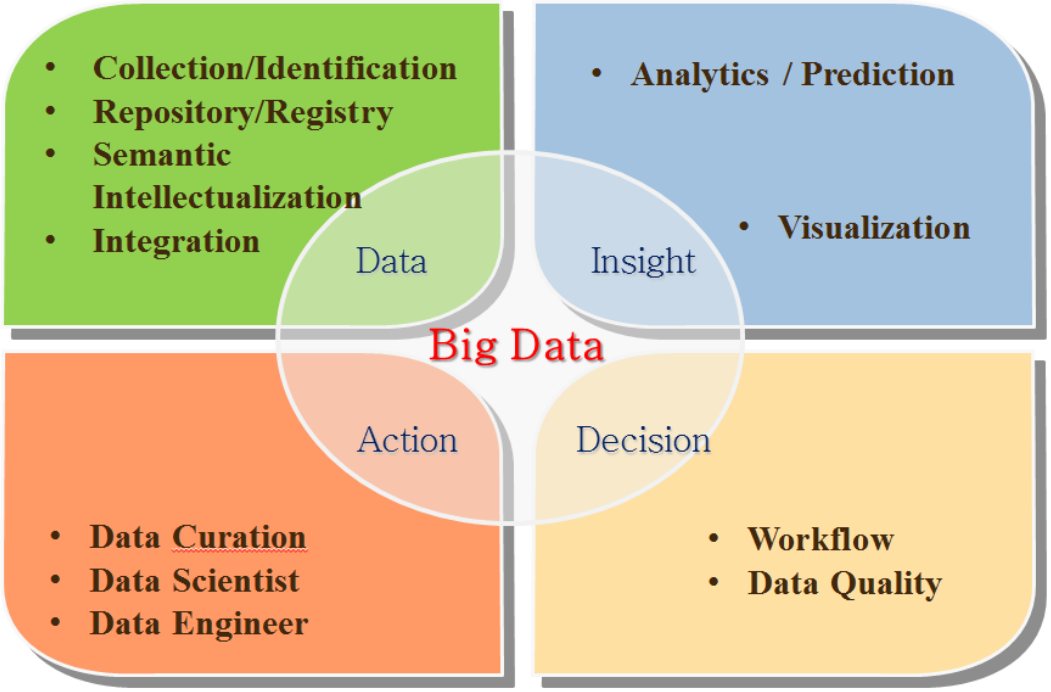
The use of computer graphics and image processing to present models or characteristics of processes or objects for supporting human understanding.
ISO/IEC 2382-13:1996, 13.01.07

9.15 Workflow

Depiction of the actual sequence of the operations or actions taken in a process
ISO 18308:2011(en)

10 Big Data Life Cycle

[NGA-REFMOD] proposes the following model to assist in understanding the life cycle of Big Data:



Note that this life cycle is multi-directional and continuous.

Also note that this diagram is included for discussion purposes and is not universally accepted as a valid description of a Big Data life cycle. However it identifies the various steps in the Life cycle of Big Data, and will help identify any special additional steps we need to take that are not found in the traditional life cycle.

11 Participants

The following people participated in this *Ad Hoc* during the 2013 SC32 plenary in Gyeongji, Korea.

Name	Email
Keith Hare – Convenor	keith@jcc.com
Baba Pripani	babap@attglobal.net
Doo-Kwon Baik	baikdookwon@gmail.com
Chong Wang	cwang@whu.edu.cn
Dongwon Jeong	djeong@kunsan.ac.kr
Frank Farance	frank@farance.com
Fei He	hefei.kh@whu.edu.cn
Hajime Horiuchi	hori@tiu.ac.jp
Jan-Eike Michels	janeike@us.ibm.com
Jangwon Gim	jangwon@kisti.re.kr
Jian Wang	jianwang@whu.edu.cn
Jörn Bartels	joern.bartels@oracle.com
Keith Gordon	kfgordon@bcs.org.uk
Eui Jong Lee	kongjjagae@gmail.com
Krishna Kulkarni	krishnak@us.ibm.com
Sukhoon Lee	Leha82@korea.ac.kr
Ying Li	liying@cesi.cn
Masashi Tsuchida	masashi.tsuchida.ax@hitachi.com
Jake Knoppers	mpereira@istar.ca
Masao Okabe	okabe.masao@gmail.com
Phil Brown	Pr_brown@btinternet.com
Jim Melton	SheltieJim@xmission.com
Kohi Shibano	shibano@aa.tufs.ac.jp
Shi Rui	shirui@cesi.cn
Tatsumi Adachi	t-adachi@gv.jp.nec.com
Takashi Kotera	takashi.kotera.xa@hitachi.com
Tim Schoechle	timothy@schoechle.org
Fenglin Wei	weifl@cesi.cn
Yang Ying	yangying@cesi.cn
Zaiwen Feng	zwfeng@whu.edu.cn