**INCITS Big Data Ad Hoc BD-00057r3**

**ISO/IEC JTC1/SGBD N<tba>**

Date: August 8, 2014

**INCITS Big Data Ad Hoc**

**ISO/IEC JTC 1/Study Group on Big Data**

|  |  |
| --- | --- |
| **DOCUMENT TYPE** | Change Proposal |
| **TITLE** | Additions and Revisions to Section 4 |
| **SOURCE** | USNB |
| **PROJECT NUMBER** |  |
| **STATUS** | Proposed |
| **REFERENCES** |  |
| **ACTION ID.** |  |
| **REQUESTED ACTION** |  |
| **DUE DATE** |  |
| **Number of Pages** |  |
| **LANGUAGE USED** | English |
| **DISTRIBUTION** | P & L Members  SC Chair & Secretariat  WG Conveners and Secretaries |

**REFERENCES**

[JTC1SGBD N0079] “Draft SGBD Report to JTC 1 v4.0”, July 23, 2014

Note to reader – The following conventions have been used in the “proposed change” column:

* Unchanged existing text is in black
* Existing text to be deleted is in ~~blue with strike-through~~
* Inserted text is in red

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| US-009 |  | 4.1 General Concept of Big Data |  | te | Section 4.1 “General concept of Big Data” describes a variety of Big Data concepts but does not describe why one would use Big Data. The initial paragraphs need some additional text indicating that Big Data is being used as input for analytical functions to learn something about the data. Analytics are also mentioned in standardization gap 10 “Remote, distributed, and federated analytics (taking the analytics to the data) including data and processing resource discovery and data mining” so it would be useful to include analytics with other important concepts. | In Section 4.1 “General Concepts of Big Data” add the following text after the third paragraph:  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.  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. |  |
| US-010 |  | 4.1 General Concept of Big Data |  | ed | In Section 4.1 “General concept of Big Data”, the paragraphs describing the general concepts should be indented, or otherwise delineated to make it clear that the paragraph starting with “The Big Data paradigm has other implications…” is not part of “Schema-on-read” | Section 4.1 “General concept of Big Data”, indent the paragraphs associated with:   * The Big Data Paradigm * Big Data Engineering * Non-Relational Models * Big Data Models * Schema-on-read |  |
| US-011 |  | 4.1 General Concept of Big Data |  | te | Section 4.1 “General concept of Big Data” needs another paragraph describing Big Data Analytics to provide context for Standardization Gap 10 | In Section 4.1 “General Concept of Big Data”, add the following text following the paragraph that starts with “**Schema-on-read** is the recognition…”  **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. |  |
| US-012 |  | 4.2 Definition of Big Data |  | te | Section 4.2 “Definition of Big Data” would benefit from a value discussion for Big Data use cases. | Insert the following text at the end of section 4.2 “Definition of Big Data”:  The above definition distinguishes Big Data from business intelligence and traditional transaction~~al~~ 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**: Big data is all about utilizing the entire corpus of relevant information, rather than justsamples 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**: Big data is all about 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**: Big data is all about 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. |  |
| US-013 |  | Section 4 |  | te | The drivers for organizations (businesses, government, etc.) to process big data should be elaborated in the document. | Add the following new section as prior to the current section 4.3 “Key Characteristics of Big Data”: 4.3 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 |  |
| US-014 |  | 4.3 Key characteristics of Big Data |  | te | The discussion of the Big Data characteristics would benefit from a more thorough initial introductory definition. | Revise the text of the current Section 4.3 “Key characteristics of Big Data” to read as follows. 4.3 Keycharacteristics of Big Data ~~The Big Data paradigm is often associated with a set of characteristics known collectively as the three, four, or even five V’s. While not every Big Data application encompasses all of these characteristics, understanding these characteristics can provide a base knowledge useful for understanding where Big Data will benefit from standards.~~  ~~The followings sections are the key characteristics of Big Data.~~  The key characteristics of Big Data focus on volume, velocity, variety, veracity, and variability. The following sections go into further depth on these characteristics. 4.3.1 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.  ~~Volume is the characteristic of data at rest that is most associated with big data.~~ Estimates produced by longitudinal studies started in 2005 [8] show that the amounts 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 is 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 at 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. 4.3.2 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.  ~~Variety means that the data represents a number of data domains and a number of data types.~~ 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 smartphones, 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~~ This variety has spawned data storage models that are more efficient for the range of 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 is 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. 4.3.3 Velocity The Velocity is the speed/rate at which the data is created, stored, analyzed 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 is 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. 4.3.4 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. ~~4~~.3.5 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. |  |