|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Item #** | **Type** | **Page #** | **Line #** | **Section** | **Comment (with rationale)** | **Suggested Change** |  |
| 1 | E | 10 | 394 | 3.3 | Suggest we include content from Section 2.1 of Volume 7 describing a definition of big data, in Table 1 of Volume 1. | Add Frank Farance as Author and  Add: “[A] definition of Big Data can be constructed by considering the essential technical characteristics in the field of study. These characteristics tend to cluster into the following four distinct segments:   1. Irregular or heterogeneous data structures, their navigation, query, and data-typing (i.e., variety); 2. The need for computation and storage parallelism and its management during processing of large data sets (i.e., volume); 3. Descriptive data and self-inquiry about objects for real-time decision-making (i.e., validity/veracity); 4. The rate of arrival of the data (i.e., velocity); and 5. Aggregation and presentation of such data sets (i.e., visualization)”   To the Definition column of Sampling of Definitions Attributed to Big Data list.  Not sure what Concept term, if any, would apply. | Modify  Suggest we insert this above 377 with a new Section  3.3 Big Data Field  Big Data as a technical field of science and engineering involves aspects of the analytical lifecycle in the context of Big Data, as well as the new techniques and technologies.  Big Data can be constructed by considering the essential technical characteristics in the field of study. These characteristics tend to cluster into the following five distinct segments:  (then items 1-5) |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

**Nancy Grady proposed the following text to expand the field of study:**

## 3.3 Big Data Science and Engineering – a Field of Study

Big Data has been described conceptually using the “v” terms in common usage. These common terms are conceptual, referring to the different qualitative characteristics of datasets. They help to understand in general when the handling of the data benefits from the use of the new scalable technologies.

Big Data Science and Engineering concerns a new *field of study*, just as chemistry, biology, statistics, and terminology are fields of study.Thisfield of study deals with the convergence of problems. The “v” word for each sub-field is provided to help align the field of study with the concepts discussed previously in Section 3.2.

### 3.3.1 Data Structures

The field of Big Data is concerned with irregular or heterogeneous data structures, their navigation, query, and data typing – relating to the variety concept. Note that Big Data is not necessarily about a large amount of data because many of the above concerns can be demonstrated with small (less than gigabyte) data sets. Big Data concerns typically arise in processing large amounts of data because some/all of the four main characteristics (irregularity, parallelism, real-time metadata, presentation/ visualization) are unavoidable in such large data sets. *Irregularity* means that the structure and shape are not known/agreed-upon in advance. The structure can be described as *structured* (both datatyping and navigation are known in advance), *semi*-*structured* (either datatyping and navigation are known, but not both), *unstructured* (neither datatyping nor navigation are known). To know the datatyping means to know the datatype as defined in ISO/IEC 11404 General Purpose Datatypes. To know the *navigation* means to know the path to complete the reference to the data, e.g., a URI is a kind of navigation, and so is  
  
 **employee[17].home\_address.address\_line[2]**  
  
The *shape* is the dimensions on a multi-dimensional array, e.g., as 4-by-5-by-6 array is a three-dimensional array (i.e., its rank is 3), a scalar is 0-dimensional array (i.e., its rank is 0, its dimensions are the empty vector {}). NoSQL techniques are used for data storage of such irregular data. JSON and XML are examples of serializing such irregular data.

### 3.3.2 Parallelism

A Big Data subfield is new engineering for computational and storage parallelism and its management during processing of large data sets – volume. Computational 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. *Parallelism* concerns the distribution of a task into multiple subtasks and a partition of resources, the coordination and scheduling of the tasks and access/contention of those resources, and the consolidation of those results. A *task* is conceived in its broadest sense: a computer, a process, a subroutine, a block of code, and a single programming language statement are all examples of a task in this sense. The Map-Reduce technique is an example of parallelism, but certainly not the only method in common use.

### 3.3.3 Metadata

Big Data Science requires descriptive data and self-inquiry about objects for real-time decision-making – validity and veracity. Descriptive data is also known as metadata. Self-inquiry is known as reflection or introspection in some programming paradigms. *Metadata* is descriptive data, e.g., describing an interface in JSON, 11179 metadata describing data structures, Dublin Core describing resources, and so on. *Real-time decision-making* is necessary because of the *irregularity* of data, e.g., rather than processing each element of data in the same way, the irregularity means that some real-time decisions are made concerning how the data is processed. An example might by: as one is walking a medical record looking for particular X-ray images in a series, one must walk into subfolders and make sense of which ones are relevant to the query/question/study in progress. As another example, the processing might change, based upon the datatype (numeric vs. string processing), the way in which the data element was observed, or the quality of data (e.g., accumulating quantities of differing precision/accuracy), and so on.

### 3.3.4 Flow rate

The Big Data field is also concerned with the rate of arrival of the data – velocity. Streaming data is the traditional description for high velocity data. This has long been a specialization in particular in such domains as telecommunications or in the credit industry. In both domains incidents of fraud needed to be detected in near real-time in order to take a mitigating action as quickly as possible.

### 3.3.5 Visual Communication

Additional challenges in Big Data projects concerns the presentation and aggregation of such data sets – visualization. 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.