**NIST Big Data Public Working Group (NBD-PWG)**

**NBD-PWD-2016/M0521**

**Source: NBD-PWG**

**Status: Draft**

**Title: Web chat from Meeting of April 19, 2016**

**Chat Log D:\\_wo\1DMG\2015\\_BigDataWG\Docs\ChatLog 2016\_04\_19 15\_03.rtf**

**Russell 2 (to Everyone)**: 1:00 PM: I am speaking

**Russell 2 (to Everyone)**: 1:00 PM: So now I have 2 machines with o microphone. Yes I can hear you

**Russell 2 (to Everyone)**: 1:01 PM: I am going to get a headset

**Russell 2 (to Everyone)**: 1:03 PM: Looking for it

**Yuri Demchenko (to Everyone)**: 1:13 PM: I need to leave in 10 min.

**Mark Underwood (to Everyone)**: 1:15 PM: We can always use bigdatastandards.com in the meantime

**Yuri Demchenko (to Everyone)**: 1:16 PM: European Commission published a set of documents today, one of them on standardisation. I will read, we discuss locally and I can provide short information on one of the next calls. The document may facilitate european invlvement including in this group.

**Mark Underwood (to Everyone)**: 1:16 PM: Yuri - great that you're keeping abreast of that work

**Ann Racuya-Robbins (to Everyone)**: 1:17 PM: who is administrating the bigdatastandards.com?

**Mark Underwood (to Everyone)**: 1:18 PM: me

**Mark Underwood (to Everyone)**: 1:19 PM: Ann - the cochairs curate content for their areas

**Mark Underwood (to Everyone)**: 1:19 PM: It's pretty thin at this point, but growing

**Ann Racuya-Robbins (to Everyone)**: 1:21 PM: Bob Marcus or anyone—I have been trying to find meeting times and notes for NIST cyberphysical and cloud groups without success. Are these groups still meeting?

**Ann Racuya-Robbins (to Everyone)**: 1:23 PM: mark could we have a implications section or subsection?

**Mark Underwood (to Everyone)**: 1:24 PM: @Ann R U thinking S+P? There are many technical implications as well, of course

**Bob Marcus (to Everyone)**: 1:24 PM: Ann: Cloud Working Groups at http://collaborate.nist.gov/twiki-cloud-computing/bin/view/CloudComputing/WebHome Check out "New" groups to see active meetings.

**Ann Racuya-Robbins (to Everyone)**: 1:26 PM: @mark, both

**Cavan (to Everyone)**: 1:29 PM: Mark: I am sitting in a De-Identification meeting in Cambridge from 11 - 12:30 on Friday. I wouldn't like to miss your subcomittee meeting

**Cavan (to Everyone)**: 1:29 PM: @Yuri: Thanks for keeping up with that

**Frank Farance (to Everyone)**: 1:30 PM: Big Data: field of study based on convergence of problems in: (1) irregular or heterogeneous data structures, their navigation, query, and datatyping; (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

**Frank Farance (to Everyone)**: 1:31 PM: Note 1: Big Data is not necessarily about a large amount of data because many of these 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.

**Frank Farance (to Everyone)**: 1:31 PM: 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.

**Frank Farance (to Everyone)**: 1:31 PM: Note 3: Descriptive data is also known as metadata. Self-inquiry is known as reflection or introspection in some programming paradigms.

**Frank Farance (to Everyone)**: 1:31 PM: 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.

**Frank Farance (to Everyone)**: 1:32 PM: As computer scientists, the four main characteristics of this definition of Big Data are: irregularity, parallelism, real-time metadata, and/or presentation /visualization. These notions are further described:

**Tim Zimmerlin (to Everyone)**: 1:32 PM: Frank, you have a "good enough" definition. However, the term "big data" is a marketing buzzword…not an engineering or math concept per se.

**Ann Racuya-Robbins (to Everyone)**: 1:32 PM: @bob, thank you. is the cyberphysical considered a new group?

**Mark Underwood (to Everyone)**: 1:33 PM: Ann - The wiki Bob lists is also at http://bigdatastandards.com/associations-and-consortia/

**Ann Racuya-Robbins (to Everyone)**: 1:33 PM: bueno

**Frank Farance (to Everyone)**: 1:35 PM: · "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.

**Frank Farance (to Everyone)**: 1:36 PM: · "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.

**Frank Farance (to Everyone)**: 1:36 PM: · "metadata" is descriptive data, e.g., JSON's describing an interface, 11179 metadata describing data structures, Dublin Core describing resources, and so on.

**Frank Farance (to Everyone)**: 1:36 PM: · "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.

**Frank Farance (to Everyone)**: 1:39 PM: Popular notions of Big Data include ill-specified ideas, such as Volume, Veracity, Velocity, and other V-words. While these words might sound catchy (possibly only to English-speakers because of the alliteration of words beginning with the letter V than do not translate similarly to non-English languages), they fail two important tests of terminological methodology:

**Ann Racuya-Robbins (to Everyone)**: 1:39 PM: @mark, might we try such a section(s)

**Frank Farance (to Everyone)**: 1:39 PM: (1) these V-words are not essential characteristics of Big Data, e.g., for the purported notion of "Velocity", telephone carriers such as Verizon, AT&T, and T-Mobile easily process a petabyte of information per second, but are not characterized as Big Data scenarios; Big Data scenarios can be demonstrated on much smaller scales less than a gigabyte

**Frank Farance (to Everyone)**: 1:40 PM: (1) these V-words are not distinguishing characteristics of Big Data, e.g., many of the notions also apply to non-Big Data scenarios, i.e., these characteristic do not distinguish Big Data from other non-Big Data concepts

**Frank Farance (to Everyone)**: 1:40 PM: (1) these V-words are not distinguishing characteristics of Big Data, e.g., many of the notions also apply to non-Big Data scenarios, i.e., these characteristic do not distinguish Big Data from other non-Big Data concepts

**Tim Zimmerlin (to Everyone)**: 1:40 PM: Frank, please send your work in an email to bigdataarch reflector for the global audience.

**Ann Racuya-Robbins (to Everyone)**: 1:41 PM: Cavan I have a few questions about your Statistical privacy work if we have time.

**Tim Zimmerlin (to Everyone)**: 1:41 PM: Russell, MapReduce implements the embarrassingly parallel algorithm. Read Geoffrey Fox's works.

**Tim Zimmerlin (to Everyone)**: 1:42 PM: Spark implements the bulk synchronous parallel (BSP) algorithm by Leslie Valiant.

**Ann Racuya-Robbins (to Everyone)**: 1:42 PM: Is it possible that the parallel approach may largely replace structured data approach?

**Mark Underwood (to Everyone)**: 1:42 PM: @Ann - Sorry, did you resp to the Q about whether this is S+P or general?

**Cavan (to Everyone)**: 1:43 PM: @Ann, sure - when?

**Ann Racuya-Robbins (to Everyone)**: 1:43 PM: @mark, I did, no problem I said both

**Cavan (to Everyone)**: 1:44 PM: Russell, thanks for update on lambda

**Ann Racuya-Robbins (to Everyone)**: 1:44 PM: @Cavan, perhaps today or next week.

**Tim Zimmerlin (to Everyone)**: 1:45 PM: Russell, Lambda architecture focuses on the event log as the core data structure; physical or virtual servers are secondary; data pools like SQL views are also secondary.

**Cavan (to Everyone)**: 1:46 PM: @Ann after the call?

**Ann Racuya-Robbins (to Everyone)**: 1:47 PM: @Cavan, during the meeting

**Bob Marcus (to Everyone)**: 1:49 PM: How is batch processing differentiated from stream processing in the Reference Architecture?

**K. Eric Harper (to Everyone)**: 1:58 PM: I have to drop out for another meeting. Thanks for the update!

**Bob Marcus (to Everyone)**: 2:00 PM: How are stream, interactive, and batch processing Use Cases mapped to the RA? Specifically how does Lambda Architecture map to RA? See http://lambda-architecture.net/img/la-overview\_small.png

**Ann Racuya-Robbins (Private)**: 2:01 PM: I think Cavan could discuss the Statistical Privacy paper a bit today unless you want to wait?

**Wo Chang (to Ann Racuya-Robbins)**: 2:02 PM: Yes, I talked to Cavan and he agrees to discuss on that as well as take away from his conference.

**Mark Underwood (to Everyone)**: 2:07 PM: The ontology folks have entire upper ontologies on the time and event representations

**Tim Zimmerlin (to Everyone)**: 2:09 PM: Mark, ha ha ha hah!

**Russell 2 (to Everyone)**: 2:10 PM: Bob: I wonder if it would be helpful to say Lambda is NRT, vs r/t

**Tim Zimmerlin (to Everyone)**: 2:11 PM: Ann, the stakeholders are responsible for defining time and all its proxies. For most vertical application domains, everyone has already agreed on time's definition (e.g., a worker's timecard is stamped).

**Mark Underwood (to Everyone)**: 2:13 PM: Tim :)

**Mark Underwood (to Everyone)**: 2:14 PM: @Tim the contempt is fully bidirectional between the SQL folks & the ontologists

**Tim Zimmerlin (to Everyone)**: 2:14 PM: Ann, the stakeholders must agree. Fairness is part of governance, provenance, and curation.

**Wo Chang (to Cavan)**: 2:15 PM: After this discussion with David, I would like to turn the time to you on your stat. privacy and lessons learned from your conf. that you attened. Is this ok?

**Tim Zimmerlin (to Everyone)**: 2:16 PM: Mark, there is no contempt. There are fundamental inconsistencies. These differences are ironed out as new systems arrive in the marketplaces. Evolve or die.

**Cavan (Private)**: 2:16 PM: Sure, it was a DARPA PI meeting on Security and Privacy

**Wo Chang (to Cavan)**: 2:17 PM: great and perfect.

**Russell 2 (to Everyone)**: 2:18 PM: I would like to get the newer multiple views doc please

**Russell 2 (to Everyone)**: 2:19 PM: Dave: what was the email date you sent the multiple views doc?

**David Boyd (to Everyone)**: 2:19 PM: Russel 5 April I sent it to the BigDataArch reflector

**Russell 2 (to Everyone)**: 2:19 PM: Thanks

**Mark Underwood (to Everyone)**: 2:29 PM: Ann, not contradicting Cavan, and Arnab is the expert here, but this may be of interest:

**Mark Underwood (to Everyone)**: 2:29 PM: https://www.coursera.org/learn/data-results/lecture/50Y9k/weaknesses-of-differential-privacy

**Ann Racuya-Robbins (to Everyone)**: 2:33 PM: great thanks mark

**Mark Underwood (to Everyone)**: 2:40 PM: Cavan: discussing provenance issues w/ commercial datasets, e.g., housing stock

**Tim Zimmerlin (to Everyone)**: 2:43 PM: Cavan, US government needs service level agreements (SLAs) for big data aggregators.

**Mark Underwood (to Everyone)**: 2:44 PM: There are SO many of me out there that I had to concede SEO failure :(

**Tim Zimmerlin (to Everyone)**: 2:45 PM: Mark, how many of those are actually real people versus avatars?

**Mark Underwood (to Everyone)**: 2:47 PM: Tim -heh, more provenance issues!

**Cavan (to Everyone)**: 2:57 PM: Mark, would love to hear about your provenance work with housing data sets. We are working with the data as well. Curious about what you are finding

**REYLING, ROBERT A CTR USAF AFMC AFLCMC/HN (to Everyone)**: 2:58 PM: Security and Privacy policy is definitely in the legal realm

**Mark Underwood (to Everyone)**: 3:00 PM: Cavan - What I run into is smallish data, but provenance is hard even here: http://www.millersamuel.com/

**Cavan (to Everyone)**: 3:02 PM: There is no incentive to provide origin provenance

**REYLING, ROBERT A CTR USAF AFMC AFLCMC/HN (to Everyone)**: 3:02 PM: End of Meeting at 3:03 PM