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

**NBD-PWD-2014/M0298**

**Source: NBD-PWG**

**Status: Draft**

**Title: Ref. Architecture + Tech. Roadmap Meeting Minutes for March 4, 2014**

**Author: NBD-PWG Co-Chairs**

Meeting logistics**:**

Dial-in Phone: 206-402-0823, Participant code: 272-30-504

Web conferencing tool: <http://nistdd.megameeting.com/guest/#&id=115333>

# Agenda

1. Action Items
2. Review SnP use cases contributed by Mark
3. Review fraud detection use case by Atul
4. Discussion mapping open source solutions to Reference Architecture
5. Use Case Implementation Platform and Approach
   1. Preparing VirtualBox (or others??) VM for Puppet Master
   2. Using Github for revision control
   3. Trying to get 5 computing nodes (1 Puppet Master, 4 Puppet Agents)
   4. Still pursuing with CS graduate students for coding
   5. Another approach is to formulate coding team within NBD
6. Selection Use Case for Implementation: let’s start with one use case first
   1. Out from RDA/BDA-IG (NASA case with datasets and domain experts)
   2. Out from 51 use cases (need to have datasets)
   3. Out from 9 Sn& use cases (need to have datasets)
7. V1.0 Technical Editing status
   1. Four documents (Definitions, Use Cases & Requirements, SnP Requirements, and Ref. Architecture) are under editing. Next will be Taxonomies and White Paper, then the Technology Roadmap.
   2. Need to identify attributions for those seven documents
8. ISO/IEC JTC 1 Study Group on Big Data
   1. Abstract (~300 words) due date for San Diego Meeting: March 11

# Notes:

Discussion by Mark Underwood on Use Cases for small projects that would be feasible to address.

The biggest question is how to get access to server log data

## Use Case mapping to Reference Architecture

Discuss the sample Use Case document provided by Bob.

The value is in parsing the use cases to determine if they represent real-time analytics or batch, ACID-compliant or eventually consistent.

One possible ISO standard discussion would be a definition of eventual consistency.

Good Article in Communications of the ACM

<http://cacm.acm.org/magazines/2013/12/169945-replicated-data-consistency-explained-through-baseball/fulltext>

and

<http://cacm.acm.org/magazines/2014/3/172505-eventually-consistent/fulltext>

As an example for the orchestrator, Puppet can be used to describe what is needed.

We wanted to see about having grad students set up tools in the environment to analyze, but some funding is needed because most students are already funded.

The LDA has core data sets and experts.

Open Science Data Cloud is another place to look for public datasets

US Census is 1.8TB

How could we model the problem?

Describe data in transmission, in processing or at rest?

Analytics life cycle is it real-time, interactive, or batch?

…

Discussion, what is our end goal, how do we move towards understandings

Need something that looks like 800-53 for different characteristics, and then evaluate the systems

We could extend the architecture to describe the interaction processes between the application provider and framework provider, as well as communication (instructions/queries and data) processes among the horizontally scaled resources

ISO could have a technical report to describe the different characteristics.

For security and privacy, the concern is that the Use Cases are too vanilla, so there is insufficient information to clarify the security fabric.

Nancy, John, PW work on validity of a process taxonomy and suggestions for metrics

## Phase 1 Documents

We need lists of contributors,

Authors

Reviewers

…

Send Wo an email listing which documents you contributed to.

Subgroup Chairs also need to list contributors

## ISO meeting

Need to submit 200-300 word abstracts for ideas for discussion

## Actions:

Mark – web logs

Robert – go to his organization for linux or windows logs

Mark – think about analytics

Nancy and Robert and John – create characteristics abstract for JTC1

Will

## Attendees:

Alicia M. Zuniga-Alvarado

Amy

Andrew Watson

Arnab Roy

Bob Marcus

Dan Samarov

David Boyd

Felix Njeh

Geoffrey Fox

John Rogers

Kristen Summers

Manoj

Mark Underwood

Nancy Grady

Ovace Mamnoon

Quyen Nguyen

Robert Whetsel

Wo Chang

# Cleansed Chat Log

(1:15 PM) Robert Whetsel: I am working a problem at work dealing with log files in an enterprise system

(1:18 PM) David Boyd (Data Tactics): Wo I would like to sit in as well.

(1:20 PM) Wo (guest): Thanks Robert and David.

(1:46 PM) David Boyd (Data Tactics): A good example of a potential standard is a more precise definition of eventual consistency.

(1:47 PM) David Boyd (Data Tactics): There was a good article in Communications of the ACM about that this month.

(1:47 PM) Robert Whetsel: hadoop is not the only file system that can be used

(1:47 PM) Robert Whetsel: for example, hammer file system

(1:48 PM) David Boyd (Data Tactics): Another example, from what Robert just said is there a need for additional Filesystem standards beyond the POSIX requirements.

(2:01 PM) Robert Whetsel: big data has three states: in transmission, in processing or at rest

(2:20 PM) Bob Marcus: Robert Whetset: Please send me some information about your research. Thanks.

(2:22 PM) Bob Marcus: I think that we should avoid implementing applications using large domain-specific data sets and analytic tools. This will require a lot of resources and may not generalize to other application domains.

(2:24 PM) Bob Marcus: In general, real world Use Cases can be decomposed and abstracted to provide generic building blocks that can be easilly implemented and provide generic value

(2:27 PM) Bob Marcus: Simole Use Cases can be used to validate the Reference Architecture and suggest possible standardizations

(2:28 PM) David Boyd (Data Tactics): So in the Tech roadmap I had some fairly deep discussions on the common attributes of the various layers. For example on the framework provider I discussed Relational, Key Value, Graph, Columnar, etc.

(2:29 PM) David Boyd (Data Tactics): @Bob - I agree.

(2:30 PM) Bob Marcus: To guide users planning Big Data projects, it will be necessary in the future to map existing tools and packages (e..g. from Apache) to the NISt Reference Architecture

(2:31 PM) Nancy Grady (SAIC): One good outcome would be to follow the attributes in the roadmaps to another level to define metrics. Where would we measure latency? How would we measure the performance difference in the way information flows from one resource to another

(2:33 PM) Robert Whetsel: Those things are in the poerpoint I sent you

(2:33 PM) Robert Whetsel: powerpoint

(2:33 PM) David Boyd (Data Tactics): Nancy - correct. For example latency is critical - what is the impact of latency on an application? What does eventual consistency really mean to this app?

(2:43 PM) Nancy Grady (SAIC): @Robert, could you post the powerpoint to the NIST site?

(2:45 PM) Robert Whetsel: sure

(2:48 PM) John Rogers (HP): Nancy: I can be reached at john.w.rogers@hp.com

(2:48 PM) Nancy Grady (SAIC): @john, email me at gradyn@saic.com to talk about a process taxonomy

(2:50 PM) Nancy Grady (SAIC): @PW, can you send me your email also?

(2:55 PM) Robert Whetsel: robert.c.whetsel.civ@mail.mil

(2:59 PM) Robert Whetsel: sorry got to go good meeting

# Extracted from Chat Log

(1:01 PM) Arnab Roy (Fujitsu) joined.

(1:02 PM) Wo Chang (Host, NIST) joined.

(1:03 PM) Quyen Nguyen joined.

(1:04 PM) Geoffrey Fox joined.

(1:04 PM) Geoffrey Fox disconnected.

(1:04 PM) Geoffrey Fox joined.

(1:07 PM) Mark Underwood (Krypton Bros): i will phone in - I was trying to use the bridge

(1:08 PM) Geoffrey Fox: Wo -- I need to leave at 1.55

(1:10 PM) Mark Underwood (Krypton Bros) disconnected.

(1:12 PM) Dan Samarov (NIST) joined.

(1:14 PM) Ovace Mamnoon (HP) joined.

(1:18 PM) David joined.

(1:21 PM) Alicia M. Zuniga-Alvarado joined.

(1:24 PM) Geoffrey Fox disconnected.

(1:27 PM) Geoffrey Fox joined.

(1:35 PM) Wo (guest) disconnected.

(1:35 PM) John Rogers (HP) joined.

(1:36 PM) Felix Njeh (COMINT) joined.

(1:54 PM) Geoffrey Fox disconnected.

(1:57 PM) John Rogers (HP) disconnected.

(1:57 PM) Mark Underwood joined.

(1:57 PM) Mark Underwood disconnected.

(1:58 PM) John Rogers (HP) joined.

(2:00 PM) Geoffrey Fox joined.

(2:01 PM) Mark Underwood joined.

(2:09 PM) Dan Samarov (NIST) disconnected.

(2:14 PM) Mark Underwood disconnected.

(2:15 PM) John Rogers (HP) disconnected.

(2:15 PM) John Rogers (HP) joined.

(2:18 PM) David disconnected.

(2:20 PM) Geoffrey Fox disconnected.

(2:35 PM) amy disconnected.

(2:52 PM) Bob Marcus disconnected.

(2:57 PM) John Rogers (HP) disconnected.

(2:59 PM) Ovace Mamnoon (HP) disconnected.

(3:02 PM) Robert Whetsel disconnected.

(3:03 PM) Quyen Nguyen disconnected.

(3:03 PM) Alicia M. Zuniga-Alvarado disconnected.

(3:03 PM) Manoj (CyberIQ) disconnected.

(3:03 PM) Felix Njeh (COMINT) disconnected.

(3:03 PM) Kristen Summers disconnected.

(3:04 PM) Andrew Watson  disconnected.

(3:07 PM) Arnab Roy (Fujitsu) disconnected.

(3:16 PM) David Boyd (Data Tactics) disconnected.