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

**NBD-PWD-2016/M0519**

**Source: NBD-PWG**

**Status: Draft**

**Title: Web chat from Meeting of April 12, 2016**

**Chat Log D:\\_wo\1DMG\2015\\_BigDataWG\Docs\ChatLog 2016\_04\_12 14\_44.rtf**

**Geoffrey Fox (to Everyone)**: 1:02 PM: yes

**Russell Reinsch (to Organizer(s) Only)**: 1:06 PM: Using Organizer chat channel for meeting notes: Agenda: Co chair version 2 updates. Geoffrey: someone in Piyush organization is checking on use case [wiki?].

**Russell Reinsch (to Organizer(s) Only)**: 1:09 PM: Mark: taking version 1 IRBs, assigning a couple of tasks.

**Russell Reinsch (to Organizer(s) Only)**: 1:13 PM: Let me try a different machine

**Mark Underwood (to Everyone)**: 1:13 PM: Lee Anne (at all) I skimmed your doc in light of this briefing, and there imay be some ways here to narrow the focus a bit: http://www.slideshare.net/drmaltman/altman-simons-privacy

**Mark Underwood (to Everyone)**: 1:13 PM: \*and all

**Mark Underwood (to Everyone)**: 1:17 PM: Frank: Slide 3 on my deck incorporates some work items from our talk on conformance. Not necessarily assignable to you, but a goal for V2 drafts

**Lee Anne Davies (to Everyone)**: 1:19 PM: Thanks Mark - My quick scan didn't pick up any 'consent' but more 'privacy/anonymization'. I will see what else this person has written and also look at the references in the slides. I appreciate help narrowing scope! I need to tease apart 'consent lifecyle' (I think).

**Bob Marcus (to Everyone)**: 1:22 PM: Lambda Architecture at http://lambda-architecture.net/

**Russell 2 (to Organizer(s) Only)**: 1:22 PM: Manning: nathan martz.

**Russell 2 (to Organizer(s) Only)**: 1:23 PM: Martin Ordurtski. created SCALA. Spark almost always means lambda.

**Mark Underwood (to Everyone)**: 1:24 PM: Tim - I assume you mean healthcare.gov ~ lambda

**Russell 2 (to Organizer(s) Only)**: 1:25 PM: Focus on getting ingest data into \_ fast. Pointers operate at times on the log.

**Mark Underwood (to Everyone)**: 1:25 PM: "Healthcare.gov uses a combination of Windows, Red Hat Enterprise Linux, Node.js, MySQL, Go and Postgres workloads on Amazon's Elastic Compute Cloud (EC2) and Relational Database Service, as well as Direct Connect to link the CMS network to the cloud."

**Bob Marcus (to Everyone)**: 1:26 PM: Lambda Architecture and Spark http://blog.cloudera.com/blog/2014/08/building-lambda-architecture-with-spark-streaming/

**Tim Zimmerlin (to Everyone)**: 1:26 PM: Mark, yes.

**Tim Zimmerlin (to Everyone)**: 1:27 PM: Both Amazon AWS and Microsoft Azure provide their native Lambda "services" (sans reserving specific servers).

**Tim Zimmerlin (to Everyone)**: 1:28 PM: Mark, downstream the data could reside in any number of data ponds in support of specific workloads. Timing and prioritization and protecting new data are key concerns.

**Russell 2 (to Everyone)**: 1:29 PM: Tim: what do you think about IoT devices presenting windows in an interface that makes the data look like tables

**Russell 2 (to Everyone)**: 1:29 PM: Have you seen that anywhere

**Tim Zimmerlin (to Everyone)**: 1:30 PM: Russell, I typically use XML for messages: formatting, parsing, error detection, version control, parsing, meta data, etc.

**Russell 2 (to Everyone)**: 1:31 PM: How are people cross referencing streaming data types with static / resting data types

**Tim Zimmerlin (to Everyone)**: 1:32 PM: Russell, event logs.

**Tim Zimmerlin (to Everyone)**: 1:33 PM: Study Michael Stonebraker's papers.

**Mark Underwood (to Everyone)**: 1:33 PM: Concept for Sarafudeen (sp) - design pattern for moving from traditional data warehouse (Teradata, SQL Server, Oracle) to big data, since that is his sweet spot AFAIK

**Mark Underwood (to Everyone)**: 1:35 PM: http://bigdatastandards.com/research-projects/big-data-projects-at-indiana-university/

**Lee Anne Davies (to Everyone)**: 1:35 PM: Where does Geoffrey teach?

**Mark Underwood (to Everyone)**: 1:35 PM: Indiana U

**Tim Zimmerlin (to Everyone)**: 1:36 PM: Russell, research Splunk. They traverse and parse msg logs. Their language is educational.

**Bob Marcus (to Everyone)**: 1:36 PM: Suggestion: Use Geoffrey's ABDS-HPC stack as an intermediary between 37 Use Cases and NIST Big Data Reference Architecture

**Mark Underwood (to Everyone)**: 1:36 PM: BTW Sarafudeen has Splunk

**Tim Zimmerlin (to Everyone)**: 1:39 PM: Russell, are you doing "data discovery" or "pattern discovery"?

**Tim Zimmerlin (to Everyone)**: 1:39 PM: "Data discovery" is most efficiently supported by search engines in big data ecosystems.

**Russell 2 (to Everyone)**: 1:40 PM: Fascinating split on the term [ discovery]

**Tim Zimmerlin (to Everyone)**: 1:41 PM: "Pattern discovery" generates 3 datasets: training, cross vadilation, and test before using "all the data".

**Tim Zimmerlin (to Everyone)**: 1:42 PM: …cross validation...

**Lee Anne Davies (to Everyone)**: 1:44 PM: #35 is interesting - but no technologies identified

**Lee Anne Davies (to Everyone)**: 1:45 PM: Sorry about my font size-I'm really not yelling! Don't know why.

**Mark Underwood (to Everyone)**: 1:47 PM: Bob - Can you push a link to the patterns slide that Geoffrey alluded to?

**Mark Underwood (to Everyone)**: 1:49 PM: Quite a few of the student projects have PII or PII-like data

**Russell 2 (to Organizer(s) Only)**: 1:53 PM: NIST has a smart city \_. No agreement on how iot connects to the cloud.

**Russell 2 (to Organizer(s) Only)**: 1:54 PM: Ck CPS SoS

**Russell 2 (to Organizer(s) Only)**: 1:56 PM: RA assumes cloud implementation, but its all sort of in one system [vs. system of systems].

**Bob Marcus (to Everyone)**: 2:01 PM: Fox ABDS-HPC Framework http://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/fox.pdf

**Mark Underwood (to Everyone)**: 2:03 PM: There can be a meta-orchestrtor tho

**Mark Underwood (to Everyone)**: 2:03 PM: If not, there is no S&P fabric

**Russell 2 (to Everyone)**: 2:06 PM: Frank: Bob is talking about what Frank describes as the parallel element of BD definition.

**Mark Underwood (to Everyone)**: 2:06 PM: Yes, cloud is a catch-all

**Tim Zimmerlin (to Everyone)**: 2:07 PM: Internet of Smart X's…overlays as far as the eye can see.

**Mark Underwood (to Everyone)**: 2:09 PM: Agree of the value Bob suggests; especially useful for cross-organizational big data; e.g., consumer device (frig?) to utility to analytics resource

**Tim Zimmerlin (to Everyone)**: 2:09 PM: Real world clouds have terabits per second Internet access links. Each AWS data center has 25 to 125 Tbps extenal links.

**Geoffrey Fox (to Everyone)**: 2:09 PM: My impression is that Apache Stack does not directly address fog (which I agree is necessary). Kinesis (Amazon) Heron (Twitter) and Cloud Data Flow (Apache Beam, Google) do not seem to have fog directly

**Russell 2 (to Everyone)**: 2:11 PM: SOS acronym for system of systems. IaaS = delivery model; hybrid = deployment model.

**Tim Zimmerlin (to Everyone)**: 2:11 PM: Amazon AWS supports systems of systems with AWS services.

**Russell 2 (to Everyone)**: 2:12 PM: Can we zoom in on slide 16

**Russell 2 (to Organizer(s) Only)**: 2:12 PM: Can we zoom in on slide 16

**Bob Marcus (to Everyone)**: 2:15 PM: Note: DoD has been CPS SoS for many years. I am going to see how to leverage some of this work. Google C4ISR to see some examples

**Mark Underwood (to Everyone)**: 2:15 PM: Bob - did you distribute this deck already?

**Bob Marcus (to Everyone)**: 2:15 PM: Slides are at http://www.slideshare.net/bobmarcus/reference-architectures-for-layered-cps-system-of-systems

**Bob Marcus (to Everyone)**: 2:16 PM: Everyone is welcome to steal anything

**Mark Underwood (to Everyone)**: 2:16 PM: Bob - ack'd

**Russell 2 (to Organizer(s) Only)**: 2:16 PM: Tim: smart cities under Sec of Commerce.

**Bob Marcus (to Everyone)**: 2:17 PM: Steal any of the slides from the slide deck accessible from http://www.slideshare.net/bobmarcus/inventory-of-my-cps-slide-sets

**Russell 2 (to Organizer(s) Only)**: 2:20 PM: Wo: ICD has three archs. Q: how to utilize them.

**Russell 2 (to Organizer(s) Only)**: 2:21 PM: Working w people at AWS, for Hadoop project, maybe Arango, [hybrid] intent to make it useful for the use cases.

**Mark Underwood (to Everyone)**: 2:24 PM: Tim: 5G standard as future use cases for BD

**Russell 2 (to Organizer(s) Only)**: 2:24 PM: Tim: 5G standards, community butting heads w legacy wifi people. 5G will play a role in Bob's network low duty cycle orchestrated edge devices. shared boxes.

**Mark Underwood (to Everyone)**: 2:25 PM: Bob - What attracted vendor interest? Best way to leverage that?

**Tim Zimmerlin (to Everyone)**: 2:25 PM: Mark, profits@

**Russell 2 (to Everyone)**: 2:25 PM: Frank: I sent the textalyzer email because of the mention about Apple / iPhone hack

**Mark Underwood (to Everyone)**: 2:27 PM: Yuri - mentions upcoming EU documents forthcoming in APR this yr

**Bob Marcus (to Everyone)**: 2:27 PM: Mark: Vendors thought that NIST was going to take a leading role in the IoT to Cloud area

**Mark Underwood (to Everyone)**: 2:27 PM: Bob ' ack'd

**Russell 2 (to Everyone)**: 2:29 PM: Dave: not sure if you remember there are a bunch of captures of missing connections in RA i M0311

**REYLING, ROBERT A CTR USAF AFMC AFLCMC/HN (to Everyone)**: 2:29 PM: Yes, we would like Uri's unput

**Russell 2 (to Everyone)**: 2:29 PM: \*in

**Russell 2 (to Everyone)**: 2:34 PM: not real hybrid. Simply meaning access to private cloud by going thru Wo. Mark: can we clone the design resource into our own account.

**Tim Zimmerlin (to Everyone)**: 2:34 PM: Github and Yum are key shared repositories for big data. NIST, NSF, NIH, and other orgs will find both necessary.

**Russell 2 (to Everyone)**: 2:35 PM: Wo: replicate the dataset and the container. Mark: only want the design pattern [perhaps]. Keep the cost down.

**Russell 2 (to Everyone)**: 2:37 PM: Mark: V2 not really thinking literally about the example spaces.

**Mark Underwood (to Everyone)**: 2:38 PM: We discussed how best to use the NIST example instances to make our work more tangilble -- or perhaps a subset of thta work

**Mark Underwood (to Everyone)**: 2:38 PM: i.e., code sharin

**Mark Underwood (to Everyone)**: 2:39 PM: \*sharing

**Mark Underwood (to Everyone)**: 2:40 PM: I asked about CRISP-DM https://en.wikipedia.org/wiki/Cross\_Industry\_Standard\_Process\_for\_Data\_Mining and SEMMA https://en.wikipedia.org/wiki/SEMMA

**Russell 2 (to Organizer(s) Only)**: 2:41 PM: Wo: commitments SEMMA and crisp - DM people use these methodologies to try models, TC69 uses crisp for wide. SEMMA more narrow.

**Mark Underwood (to Everyone)**: 2:42 PM: Per Wo, Nancy's goal is to map CRISP activities into the app layer of the RA

**Russell 2 (to Organizer(s) Only)**: 2:42 PM: For App provider layer, crisp ~ activities.

**Mark Underwood (to Everyone)**: 2:43 PM: "activity" <=> orchestration (?)