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

**NBD-PWD-2015/M0489**

**Source: NBD-PWG**

**Status: Draft**

**Title: Web chat from Meeting of Jan. 12, 2016**

**Chat Log D:\\_wo\1DMG\2015\\_BigDataWG\Docs\ChatLog 2016\_01\_12 15\_00.rtf**

**David W. Boyd (to Everyone)**: 1:18 PM: dboyd@incadencecorp.com

**Tim Zimmerlin (to Everyone)**: 1:21 PM: Frank, what about "newSQL" in INCITS RefArch?

**Tim Zimmerlin (to Everyone)**: 1:24 PM: Frank, what of provenance metadata?

**Tim Zimmerlin (to Everyone)**: 1:32 PM: Wo, is OpenGov explicitly using EarthCube?

**David W. Boyd (to Everyone)**: 1:36 PM: The whole Multipart thing is up in the air. Several objections and changes to the split in the JTC1 voting. Will email you some information. Big thing is that there is disagreement on where the S&P part should go (WG9 or SC27).

**David W. Boyd (to Everyone)**: 1:37 PM: All: last message was to be private to Mark. Sorry to distract folks.

**Russell Reinsch (to Everyone)**: 1:39 PM: I have to handle someting here for a minute; I will be offline for a couple.

**Tim Zimmerlin (to Everyone)**: 1:42 PM: Wo, Geoffrey Fox has your version 2 SW APIs in his Apache BD stack. This is the best organized and most comprehensive AFAIK.

**Cavan Capps (CENSUS) (to Everyone)**: 1:43 PM: sorry late - phone problems

**Russell Reinsch (to Everyone)**: 1:44 PM: Have returned

**Tim Zimmerlin (to Everyone)**: 1:44 PM: Note this Apache BD stack relies on the Java VM, just like Hadoop, Solr/Lucene, Hutch.

**Cavan Capps (CENSUS) (to Everyone)**: 1:49 PM: Noaa always has large datasets that are public

**Cavan Capps (CENSUS) (to Everyone)**: 1:52 PM: I would also be interested in testing extreme multi-user environments (like to explore monitoring it and extending graceful management)

**Tim Zimmerlin (to Everyone)**: 1:52 PM: Note the JVM faces many core chips (e.g., >50), and there is no universal solution. Stroustrup, Odersky, Stepanov, and Fox are working on post Von Neumann general machines.

**Geoffrey Fox (to Everyone)**: 1:53 PM: We have Hadoop runnin well on 100 nodes http://dsc.soic.indiana.edu/publications/LDA\_optimization\_paper.pdf with plug in Hadoop faster than Spark

**Mark Underwood (to Everyone)**: 1:53 PM: Having a Spark DevOps environment seems critical for getting industry interest

**Cavan Capps (CENSUS) (to Everyone)**: 1:54 PM: agree with Mark

**Russell Reinsch (to Everyone)**: 1:55 PM: 1+ the Spark

**Tim Zimmerlin (to Everyone)**: 1:56 PM: Geoffrey, these experts are using out of the box Apache and Open Stack and other open source. I understand your papers clearly point to future options.

**Mark Underwood (to Everyone)**: 1:56 PM: Geoffrey - thx for the link

**Tim Zimmerlin (to Everyone)**: 1:57 PM: Geoffrey, real world BD users are naive and prone to failure, so experts provide them safe SW tools, with training wheels.

**Mark Underwood (to Everyone)**: 1:58 PM: Cavan - does your remark about multi-user fit into the systems management space or is it straightforward scalability?

**Mark Underwood (to Everyone)**: 2:04 PM: We have some work to do to link up our V1 fabric analysis with the issues Cavan has brought to us

**Cavan Capps (CENSUS) (to Everyone)**: 2:04 PM: I think both in reality

**Mark Underwood (to Everyone)**: 2:04 PM: E.g., not speaking of applications as sec / priv components omits a core element of how we integrate the RA with "fabric"

**Mark Underwood (to Everyone)**: 2:05 PM: It's a nontrivial todo!

**Tim Zimmerlin (to Everyone)**: 2:06 PM: The EU Commision is focused exclusively on complete control of information. They intend to control the Internet, public media, public access to data. When pushed hard, their concrete proposals are scary echoes from the past.

**Cavan Capps (CENSUS) (to Everyone)**: 2:09 PM: The folks I talked to who helped developed Hadoop (at Strata 4 years ago) said not to use SAN storage, however cloud providers have pushed to relax that

**Cavan Capps (CENSUS) (to Everyone)**: 2:10 PM: And most of them (back then) recommended to run on bare metal (i.e. not elastic)

**Tim Zimmerlin (to Everyone)**: 2:11 PM: Cavan, HDFS expands on RAID 0+1 as often implemented in a SAN. HDFS and SAN are redundant and extra overhead. For Hadoop/Spark use HDFS for share nothing and minimized LAN traffic.

**Tim Zimmerlin (to Everyone)**: 2:12 PM: Cavan, Amazon AWS solved all these bare vs VM, local storage vs SAN years ago. Just use AWS services, after test driving them.

**Tim Zimmerlin (to Everyone)**: 2:14 PM: Chinese chair ITU today. There is a lot of dust kicked up.

**Tim Zimmerlin (to Everyone)**: 2:16 PM: Previously, Putin tried to take over the Internet with his puppet chair of ITU. Putin failed because he did not understand how the Internet is governed. Today, the attack continues on a new initiative.

**Cavan Capps (CENSUS) (to Everyone)**: 2:17 PM: I don't truly understand how AWS impliments hadoop at a physical level - yet. I know it is a standard use now.

**Mark Underwood (to Everyone)**: 2:18 PM: So China will handle security and privacy? That would be so Manchurian Candidate

**Tim Zimmerlin (to Everyone)**: 2:20 PM: Cavan, Amazon will not tell you. Policy.

**Cavan Capps (CENSUS) (to Everyone)**: 2:20 PM: figures

**Tim Zimmerlin (to Everyone)**: 2:21 PM: Cavan, business survival strategy.

**Tim Zimmerlin (to Everyone)**: 2:25 PM: Mark, China is understands the strategic value of Great Walls.

**Tim Zimmerlin (to Everyone)**: 2:26 PM: Cavan, based on what I know, AWS uses the Lambda architecture for most services. Lambda is documented in a few textbooks and articles.

**Mark Underwood (to Everyone)**: 2:27 PM: Tim, and not walled gardens

**Frank Farance (to Everyone)**: 2:28 PM: Article 10: Processing not allowing identification

If the data processed by a controller do not permit the controller to identify a natural person, the controller shall not be obliged to acquire additional information in order to identify the data subject for the sole purpose of complying with any provision of this Regulation.

**Mark Underwood (to Everyone)**: 2:30 PM: Frank is describing a use case from our v1 doc

**Arnab Roy (to Everyone)**: 2:32 PM: Agree that policy requirements need to be delineated from technologies enforcing them

**Arnab Roy (to Everyone)**: 2:32 PM: What we need is policy as examples of requirements

**Arnab Roy (to Everyone)**: 2:32 PM: Then describe which technologies address the specific requirements

**Tim Zimmerlin (to Everyone)**: 2:33 PM: Arnab, policy can be distilled into cross cutting concerns and requirement elemental specifications.

**Tim Zimmerlin (to Everyone)**: 2:34 PM: Arnab, the NSF's four BD hubs have healthy SnP communities, including ethics.

**Arnab Roy (to Everyone)**: 2:35 PM: Tim, if you have pointers pls share

**Tim Zimmerlin (to Everyone)**: 2:36 PM: Arnab, I have listed the four links to the hub sites. You can google North East, South, Midwest, West big data regional hubs with NSF.

**Arnab Roy (to Everyone)**: 2:37 PM: OK thanks Tim

**Mark Underwood (to Everyone)**: 2:45 PM: I realize this doesn't deliver immediately useful results for Census. There are interim helpful results that might depart from our standards-guided instincts but still be helpful

**Mark Underwood (to Everyone)**: 2:49 PM: http://iot.ieee.org/conferences-events/ieee-end-to-end-trust-and-security-for-the-internet-of-things.html

**Arnab Roy (to Everyone)**: 2:57 PM: http://www.ieee-security.org/TC/SP2016/

**Russell Reinsch (to Everyone)**: 2:57 PM: Tim do you have a pointer to something on the concept of public records on Dr and lawyers history of...

**Arnab Roy (to Everyone)**: 2:58 PM: http://csis.pace.edu/BigDataSecurity/BigDataSecurity2016/

**Cavan Capps (CENSUS) (to Everyone)**: 2:58 PM: Thanks for the links