**NIST Big Data Working Group (NBD-WD)**

**NBD-WD-2013/M0095**

|  |  |
| --- | --- |
| **Source:** | Requirements Subgroup |
| **Title:** | Requirements Subgroup Meeting Minutes |
| **Date:** | August 6, 2013 |
| **Author(s):** | Geoffrey Fox (U. of Indiana) |

**Agenda**

1. Network Implications of BD
2. Need good vertical representative Use Cases
3. Discuss process for interlocking with Cloud Computing WG
4. Discuss additional actions to be performed after use cases collected or while they are being collected
5. Consider individuals and organizations who can contribute. Solicit use cases from them
6. Discuss interlock with other SG, particularly Reference Architecture and Roadmap

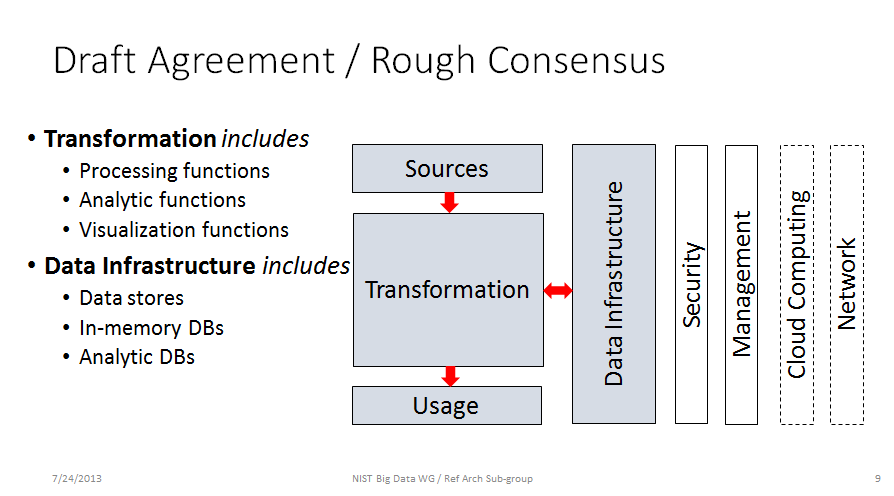


Figure-1: High-level Ref. Architecture (draft)

The meeting was dogged by poor audio but had better attendance than last week. Co-chair Tsegereda Beyene was unable to make meeting and so we postponed consideration of agenda topic “Network Implications of BD” where she is expert. We discussed two new submitted use-cases “cargo shipping” and “pathology” in detail. Cargo shipping is an “Internet of Things” application with analogies to Smart Grid” and the distributed sensors used to monitor machines such as modern aircraft engines. There was some discussion if it fitted “big data” definition and it was noted that it fell into “real time analytics” / “Complex Event Processing” categories.

We discussed at length taking use cases and relating them to features in architecture diagram (see above Figure-1, M0059) -- in particular “source” “transformations” “usage” and “data infrastructure” areas. Fox did this for CReSIS Radar processing use case showing it worked well if we broke use case into 4 stages -- each stage then had clean entries in reference architecture. Fox will do this for the 4 use cases he contributed; get comments from Orit; iterate and then send out to existing use cases submitters for them to add architecture discussion.

Bob Marcus requested group identify top 10 requirements from current use cases. This was end of session. Fox will discuss this with Chang Thursday.

Chat log:

(10:57 AM) Bob Marcus(ET-Strategies) joined.

(11:01 AM) Scott Brim (Internet2) joined.

(11:05 AM) Eugene Luster (DISA CTO/R2AD) joined.

(11:06 AM) Orit Levin (Microsoft) joined.

(11:06 AM) \_USER\_NAME\_(\_COMPANY\_) joined.

(11:07 AM) \_Cherry Tom\_(\_IEEE-SA\_) joined.

(11:09 AM) Scott Brim (Internet2): I will type here instead of speaking

(11:10 AM) \_Cherry Tom\_(\_IEEE-SA\_) disconnected.

(11:10 AM) \_Cherry Tom\_(\_IEEE-SA\_) joined.

(11:12 AM) Alicia Zuniga-Alvarado joined.

(11:14 AM) Orit Levin (Microsoft): Yes, I am on the call.

(11:18 AM) Scott Brim (Internet2): Orit: (1) Hi. (2) it seems to me that infrastructure should be in the center, with sources/transformation/usage all circled around it.

(11:18 AM) Pw Carey(Compliance Partners, LLC) joined.

(11:19 AM) William Miller (MaCT USA) joined.

(11:19 AM) Bob Marcus(ET-Strategies) disconnected.

(11:19 AM) Bob Marcus(ET-Strategies) joined.

(11:19 AM) PavithraKenjige joined.

(11:21 AM) Pw Carey(Compliance Partners, LLC): The 'transformation' takes place via MapReduce across the entire enterprise....no?

(11:22 AM) Pw Carey(Compliance Partners, LLC): Rather than within on individual 'data warehouse'...no?

(11:24 AM) Scott Brim (Internet2): OK, I can accept this if "transformation" is a very general term, and moving data in and out of the infrastructure is itself a transformation.

(11:25 AM) Scott Brim (Internet2): and Carey, I guess this is abstract, and infrastructure can be multiple instances

(11:25 AM) Pw Carey(Compliance Partners, LLC): The 'scattered raw data' is transformed into a useable 'format'....that's the benefit MapReduce brings to this equation....

(11:26 AM) Pw Carey(Compliance Partners, LLC): Yes, correct....multiple instances...multiple buckets of data w/o a good structure...prior to the MapReduce transformation.....of this same data....

(11:27 AM) Alicia Zuniga-Alvarado: excellent Orit

(11:27 AM) Scott Brim (Internet2): yes

(11:29 AM) PavithraKenjige disconnected.

(11:29 AM) Scott Brim (Internet2): Always better to dial into these meetings instead of using inline audio

(11:29 AM) PavithraKenjige joined.

(11:33 AM) Pw Carey(Compliance Partners, LLC): We just hung up our dial in and are now listening in via the Web tool.....also the following citation will bring everyone up to a baseline regarding Big Data & Hadoop:Hadoop and the Data Warehouse:When to Use WhichDr. Amr Awadallah, Founder and CTO, Cloudera, Inc.Dan Graham, General Manager, Enterprise Systems, Teradata Corporation...hope this helps....

(11:36 AM) Pw Carey(Compliance Partners, LLC): Big Data does not re-place RDMS or DBMS but is an adjunct to these based upon the volume of data being extracted, analyzed, and consumed...including data preservation, archiving and data destruction....one aspect of this difference is the lack of 'chain of custody' requirements within Hadoop/Big Data....

(11:42 AM) Bob Marcus(ET-Strategies): The 4 component architecture i very high level. It will probably be necessary to drill down to a more detailed architecture to capture the specific requirements.

(11:42 AM) Pw Carey(Compliance Partners, LLC): You talking is fine.....Respectfully yours, Pw

(11:46 AM) William Miller (MaCT USA): the reference architecture is a first shot on a top down basis we should now review it from a bottoms up basis

(11:47 AM) Pw Carey(Compliance Partners, LLC): Rather, your talking is fine....Respectfully yours, Pw

(11:47 AM) William Miller (MaCT USA): there are things that are in the details that relate to architecture for big data

(11:47 AM) William Miller (MaCT USA): i sent a use case to Wo on cargo shipping industry which is goo example

(11:48 AM) William Miller (MaCT USA): i stillneed to supply more inforamtion on the use case that i've submittted

(11:49 AM) Pw Carey(Compliance Partners, LLC): A good reference regarding big data architecture is: \* “MapReduce and the Data Scientist”, Colin White, 2012

(11:58 AM) PavithraKenjige disconnected.

(11:58 AM) PavithraKenjige joined.

(12:11 PM) Pw Carey(Compliance Partners, LLC): Your voices are breaking up...however, feel free to assign us a task and we'll follow through, and please include the deadlines....Pw

(12:12 PM) PavithraKenjige disconnected.

(12:13 PM) Pw Carey(Compliance Partners, LLC): We did discuss the shipping diagram yesterday....

(12:14 PM) PavithraKenjige joined.

(12:15 PM) Geoffrey Fox: I don't think we discussed in requirements group before

(12:15 PM) PavithraKenjige disconnected.

(12:17 PM) PavithraKenjige joined.

(12:17 PM) PavithraKenjige disconnected.

(12:18 PM) PavithraKenjige joined.

(12:21 PM) Alicia Zuniga-Alvarado disconnected.

(12:23 PM) Alicia Zuniga-Alvarado joined.

(12:28 PM) Scott Brim (Internet2): OK I get these requirements: Data collection, data management, timeliness of processing ...

(12:28 PM) Scott Brim (Internet2): interworking of data (RDA) ...

(12:32 PM) Orit Levin (Microsoft): I agree with Scott, it is data collection, data intergration, data matching...

(12:33 PM) Pw Carey(Compliance Partners, LLC): and data persistence....so, where does data leakage enter into this scenario....?

(12:34 PM) PavithraKenjige disconnected.

(12:34 PM) PavithraKenjige joined.

(12:49 PM) Orit Levin (Microsoft): :feedback" not approval ;-)

(12:53 PM) Pw Carey(Compliance Partners, LLC): We'll have our DR Disaster Recovery Use Case ready by this Friday, Aug. 9th

(12:55 PM) Eugene Luster (DISA CTO/R2AD) disconnected.

(12:56 PM) PavithraKenjige disconnected.

(12:56 PM) \_USER\_NAME\_(\_COMPANY\_): agreed with the caller

(12:57 PM) PavithraKenjige joined.

(12:59 PM) \_USER\_NAME\_(\_COMPANY\_) disconnected.

(12:59 PM) Pw Carey(Compliance Partners, LLC): Thank you....

(12:59 PM) \_Cherry Tom\_(\_IEEE-SA\_) disconnected.

(12:59 PM) Alicia Zuniga-Alvarado: thanks,and bye!

(12:59 PM) Alicia Zuniga-Alvarado disconnected.