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

**NBD-PWD-2016/M0530**

**Source:    NBD-PWG**

**Status:     Draft**

**Title:       NBD-PWG Meeting Agenda for May 31, 2016**

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

Meeting logistics

Date/time: May 31, 1:00PM – 3:00PM EDT

Web conferencing tool: <https://global.gotomeeting.com/join/790820565>

Audio: Use your microphone and speakers (VoIP) - a headset is recommended, or, call in using your telephone (US, long distance): +1 (646) 749-3122, access code/meeting ID: 790-820-565

## **Agenda**

1. Planning face-to-face NBD-PWG workshop at NIST, Dec. 14 – 16, 2016 (tentative)
2. Tentative timelines for our V2 are (A BIG assumption that we have sufficient V2 content):

June 15: public announcement for the Dec. 14-16 workshop on Big Data Interop. Framework

Nov. 14: make available rough DRAFT-1 for Dec. workshop discussion

Dec. 14 – 16: NBDIF workshop (could be two days or two and half days)

Dec. 19: incorporate all input and produce DRAFT-2 for public comments

Feb. 28, 2017: begin DRAFT-2 for public comment for 45 days; comment close by ~mid-April

Mid-April: incorporate all comments and produce DRAFT-3 for ITL review

Mid-May: incorporate ITL comments and produce DRAFT-4 for WERB (NIST editor board) review

Mid-June: begin WERM review

Mid-Aug: incorporate WERB comments and produce V2 release

1. Continue Subgroup Co-Chairs report – Draft V2 outline/content:

Action Items - Continue to identify and review V2 contributions:

(Contributors: please work with the subgroup co-chairs for the items highlighted in light blue; looking forward to discuss your contributions at next telecon)

1. From Vol. 1 Definitions, Section 1.5 Future Work (with updates) – Nancy
2. Enhance existing V1 definitions

* Data ownership (sharing??) and privacy
* What does data ownership mean?
* Users can contribute content but company (e.g. Facebook will own them)
* Data ownership involves with intellectual property (licensing, etc.)
* Cavan: Data ownership: what and who allows access…
* Lee Anne – Informed consent (moved to SnP)

Thomas – share thoughts/outline specific contribution

Dan Gillman – share thoughts/outline specific contribution

1. Revisit Big Data and associated definitions
2. Redo the Data Science section (Nancy feels this has been too diluted in the current version)

Ann – will look into the data science stack

Tim – will look into data discovery and pattern analytics (from the perspective of Andrew Ng)

1. Redo data lifecycle to analytics lifecycle or something else

Nancy – updating to CRISP-DM (list of what lifecycle not)

Frank – Datasets lifecycle

Cavan – provenance on datasets

1. Decide how to address the data-information-knowledge distinctions
2. Defining the different patterns of communications between Big Data resources to better clarify the different approaches being taken;
3. Bob Marcus Internet of Things (IoT) (e.g., streaming analytics, other new scenarios)

Bob – will provide draft text on IoT

1. Bob’s earlier use cases
2. Application provider with 3V’s (TC69 relations)
3. Improving the discussions of governance, value, and data ownership;

Tim - shared his outline doc submitted to reflector(?) 2-16: Governance Provenance Curation. Sourced Morningstar. Mark suggests [pillar] section #6 is probably the most important. Next / action item for group - Take contributions on relationship to previous work by NARA, maybe something from Quyen

Quyen: work with Russell’s feedback for Vol. 6 Section 5 on Big Data Lifecycle

Rose/MITRE team – share thoughts/outline specific contribution

1. Governance – architecture view (from velocity, volume, etc.); big data inside/outside ??, Rose
2. Developing the Management section; (Scope? Obviously systems, etc.)

Russell – share thoughts/outline specific contribution by March 01.

1. Developing the Security and Privacy section; (Arnab, Mark)

Rose/MITRE team – share thoughts/outline specific contribution

Tim: willing to be “interpreter” between NBD and MITRE

1. Improve discussion of new behavior (new design patterns?) in big data
2. Concurrency
3. Emergent behavior (inference) “Matrix Effect” (Mosaic effect) where for example now have PII concerns

Cavan – share thoughts/outline specific contribution (Mosaic effect)

Ann – write up something with the “inference is emerging”

1. Memory innovations (walls: memory, network, power)

Tim – share thoughts/outline specific contribution

1. Whitepaper: Implication of Big Data in social, business, public, knowledge extraction, and lifecycle, etc.

Ann – Share outline/document

1. Discuss the topic of value
2. Architectures (briefly introduce, use GFox descriptions, etc. Brief here, fleshed out elsewhere in Taxonomy. Tim recommends.)
3. Orchestrator / Orchestration: Refer to additional detail in the MindMap, (a discussion ensued re: Orchestrator as role, but also as conceptual underpinning for enterprise (“business”) process in the RA. (Remarks from Ann, Russell, Bob, Dave, Nancy).
4. Cross references to other volumes…
5. INCOSE bib. site
6. From Vol. 2 Taxonomies, Section 1.5 Future Work (with updates) – Nancy
7. Align with the other v1 documents (Nancy: “not mature” at this stage). See also the Mindmap, a resource for this work. [Russell]

Ann – share thoughts/outline specific contribution

Dan Gillman – share thoughts/outline specific contribution

Thomas – share thoughts/outline specific contribution

1. Mindmap – <https://www.mindmeister.com/322462463>
2. Translate mindmap into text

Russell – share thoughts/outline specific contribution

1. The Subgroup is continuing to explore the changes in both Management and in Security and Privacy. As changes in the activities within these roles are clarified, the taxonomy will be developed further.
2. The Privacy taxonomy draft is a small fork/task (SnP subgroup)
3. In addition, a fuller understanding of Big Data and its technologies should consider the interactions between the characteristics of the data and the desired methods in both technique and time window for performance. These characteristics drive the application and the choice of tools to meet system requirements. Investigation of the interfaces between data characteristics and technologies is a continuing task for the NBD-PWG Definitions and Taxonomy Subgroup and the NBD-PWG Reference Architecture Subgroup.
4. Finally, societal impact issues have not yet been fully explored. There are a number of overarching issues in the implications of Big Data, such as data ownership and data governance, which need more examination. Big Data is a rapidly evolving field, and the initial discussion presented in this volume must be considered a work in progress.
5. Explore the taxonomy’s ability to work with Geoffrey’s blend of HPC and Big Data
6. Explore formal methods to define taxonomy (Dave suggested) [Russell] DODAF like?
7. Efforts should leverage definitions produced elsewhere to avoid ocean-boiling or unresolvable concerns that are less salient for big data. E.g., taxonomy of application patterns (Dave). Emphasize building blocks, roles (Wo).
8. From Vol. 3 Use Cases & Requirements, Section 1.5, Future Work (with updates) – Geoffrey and Piyush

Lee Anne – share thoughts/outline specific contribution

William – share thoughts/outline specific contribution

[Need to come up with a strong motivation reason why people would submit new use cases. It is also hard to get security and privacy use cases]

1. Review, finalize, and begin collecting new use cases based from UC Template V2 (Challenge is to foster motivation)

(In parallel: review and identify security and privacy related issues/requirements from Bob’s 10 scenarios and/or the existing 51 use cases)

(Mark is willing to take time to interview people who have BD applications; Mark could Piyush after the call for NASA’s applications)

1. How to analyze new use cases with SnP info? How to coordinate vol. 4?
2. Draw on the use case classification to suggest classes of software models and system architectures [1][2][3][4][5]
3. A more detailed analysis of reference architecture based on sample codes that are being implemented in a university class. [6]
4. Collect benchmarks that capture the “essence” of individual use cases.
5. Additional work may arise from these or other NBD-PWG activities. Other future work may include collection and classification of additional use cases in areas that would benefit from additional entries, such as Government Operations, Commercial, Internet of Things, and Energy. Additional information on current or new use cases may become available, including associated figures. In future use cases, more quantitative specifications could be made, including more precise and uniform recording of data volume. In addition, further requirements analysis can be performed now that the reference architecture is more mature.
6. From Vol. 4 Security & Privacy, Section 1.5, Future Work (with updates) – Arnab and Mark
7. How to coordinate with vol.3 new use cases on SnP?
8. Developing the unified security and privacy taxonomy:

William – share thoughts/outline specific contribution

1. Developing the connection between the security and privacy fabric and the NBDRA (should we rename Vol. 4 to “Security and Privacy Fabric”?);
2. Exploring governance, risk management, data ownership, and valuation with respect to Big Data ecosystem, with a focus on security and privacy;

Rose/MITRE team – share thoughts/outline specific contribution

1. Contextualizing the content of Appendix B in the NBDRA; and
2. Expanding the privacy discussion within the scope of this volume;
3. Exploring privacy in actionable terms based on frameworks such as those described in NISTIR 8062 [7] with respect to the NBDRA.
4. Whitepaper:

Cavan and team – Continue to review Privacy in Big Data System whitepaper outline

1. Sarafudheen: Data Warehouse Big Data Migration
2. Ajit Gaddam: Crosswalk to SABSA/Open Group Security Services Catalog

Other complement topics (Mark):

1. **V2 “Synthesis” Document**. The scope as proposed is likely beyond the reach of the WG resources. Proposing a first step in lieu of a full V2 as an overview which: (1) summarizes V1 in a more digestible format; (2) addresses conflicts in terminology across the documents; (3) better distinguishes where we to contribute vs. other standards groups; (4) focus on fewer but more detailed use cases (below)
2. **Two primary use cases.** Selecting a big science and a PII-rich use case which can be implemented in the NIST hybrid cloud and highlight Big Data Variety and Velocity. Probably a streaming use case that simulates IoT end points and uses analytics aggressively.
3. **Implementer Guide** that could be adapted by agencies wanting to get their feet wet in Big Data; identify where additional support might be needed (e.g., what a supporting SOW might need to have in it);
4. **Flesh out Orchestration** (6) walkthrough and/or actual example(s) of orchestration across logical boundaries in the RA, probably superimposed on Docker (see also Amazon SWF, Google Kubernetes, Apache Falcon, Spring Cloud Dataflow, Bluemix Websphere Cast Iron, Apcera, Mesos, etc.). Focus on data movem40ent across RA and organizational boundaries. Our examples need to incorporate risk, ownership, provenance, etc. and go beyond cloud orchestration alone by addressing, e.g., sensor data and on-premises applications. Rose/MITRE team – share thoughts/outline specific contribution
5. **SnP Version 2** Two tracks: one Big Data SnP Fabric: Technical, one a narrative: Big Data SnP Fabric: Emerging Processes. The narrative track adapts takes the crosswalk standards document (the long Word doc with a list of standards or recommended practices) and offers a non-proscriptive discussion which references the RA whenever possible. (It can morph to something that supports conformance documents, but that might not happen in the proposed timeframe). The narrative incorporates improvements to taxonomy and makes clearer where policy and technology solutions for SnP can be separated (this is a recurring source of confusion).
6. **Value, Governance, Ownership** There are overlaps with systems management (SysMan), but key aspects are related to Big Data SnP. Create a SysMan section for each major component in the RA, then address value, governance, ownership, privacy etc. in that context. The fabric concept tends to emerge from this level of discussion rather than speaking of privacy or ownership in the abstract.

Rose/MITRE team – share thoughts/outline specific contribution

1. **Big Data Analytics / Data Science** Not mentioned in Wo’s list directly, but the use of “synthetic categories” and classifications through deep learning or other means is a Big Data concern. See this lay discussion on the subject of Big Analytics Failure. It’s related to provenance, but that’s not the whole story.
2. From Vol. 5 White Paper Survey, No Future Work for now
3. From Vol. 6 Reference Architecture, Section 1.5, Future Work (with updates) – David

Thomas – share thoughts/outline specific contribution

1. Reference Architecture Refinement
2. Establish activity and functional component views beyond the current conceptual view

All – seek feedback on activity and functional diagrams, David

1. Define high level and general activities and functional components within each view
2. Identify high level stakeholder and map their concerns to activities and functional components
3. Reference Architecture application
4. Establish white paper template

All – seek feedback on preliminary white paper template, Wo

1. Implement the NIST identified six use cases and/or other use cases from the 62 (51 generals and 11 security and privacy) collected use cases or others
2. Identify development environment (NIST hybrid cloud) for hosting the use case implementations
3. Create white papers by working with domain experts to identify workflow and interactions among the NBDRA components and fabrics
4. Review, analyze white papers high-level interactions and workflows and aggregate them into preliminary general interfaces
5. Conformance approach (emphasized by Frank & Mark)
6. From Vol. 7 Standards Roadmap, Section 1.5, Future Work (with updates) – Russell

Thomas – share thoughts/outline specific contribution

1. Examine all version 1 volumes and:
2. Identify available standards, and those under development. Status: work is in progress.
3. Gap the differences between version 1 volumes and list of standards, and continue to build and refine the gap analysis and document the findings;
4. Extend related SDO listing and establish criteria on how to select relevant SDOs
5. Support harmonizing of terminology between volumes. Status: work has started.
6. Identify early adopters from academia, government and industry. Status: work has started.
7. Identify barriers to big data adoption. Doc M0499 uploaded to reflector on Feb 16.
8. Identify where standards may accelerate the adoption and interoperability of Big Data technologies;
9. Further map standards to NBDRA components and the interfaces between them.
10. Enhance gap analysis on how to enable the RA
11. Document vision and recommendations for future standards activities
12. Engage communities to attract additional BDWG participation and drive consensus on how big data should move forward. List of whom to engage and email template are 90% complete and first handful of outreach communications went out prior to Feb 23.

Other contributions: M0499v4, M0501, M0502 – Russell

[1] Shantenu Jha, Judy Qiu, Andre Luckow, Pradeep Mantha, and Geoffrey C. Fox, “A Tale of Two Data-Intensive Approaches: Applications, Architectures and Infrastructure, in 3rd International IEEE Congress on Big Data Application and Experience Track,” Cornell University Library, June 27- July 2, 2014, <http://arxiv.org/abs/1403.1528>.

[2] Judy Qiu, Shantenu Jha, Andre Luckow, and Geoffrey C. Fox, “Towards HPC-ABDS: An Initial High-Performance Big Data Stack,” Indiana University, August 8, 2014. <http://grids.ucs.indiana.edu/ptliupages/publications/nist-hpc-abds.pdf>.

[3] Geoffrey Fox, Judy Qiu, and Shantenu Jha, “High Performance High Functionality Big Data Software Stack, in Big Data and Extreme-scale Computing (BDEC),” Indiana and Rutgers Universities, 2014. <http://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/fox.pdf>.

[4] Geoffrey C. Fox, Shantenu Jha, Judy Qiu, and Andre Luckow, “Towards an Understanding of Facets and Exemplars of Big Data Applications,” Indiana University, July 20, 2014. <http://grids.ucs.indiana.edu/ptliupages/publications/OgrePaperv9.pdf>.

[5] Geoffrey Fox and Wo Chang, “Big Data Use Cases and Requirements,” Indiana University, August 10, 2014. <http://grids.ucs.indiana.edu/ptliupages/publications/NISTUseCase.pdf>.

[6] Geoffrey Fox. “INFO 590 Indiana University Online Class: Big Data Open Source Software and Projects,” Indiana University, 2014 [accessed December 11, 2014], <http://bigdataopensourceprojects.soic.indiana.edu/>.

[7] DRAFT Privacy Risk Management for Federal Information Systems

<http://csrc.nist.gov/publications/drafts/nistir-8062/nistir_8062_draft.pdf>