**September 30, 2013 NIST Big Data Workshop Summary**

**Oct. 29, 2013**

NIST hosted a Big Data Workshop for the NIST Big Data Public Working Group (NBD-PWG) on September 30, 2013. Over 70 industry, academia, and government representatives attended the workshop. The program schedule and presentations are available at: <http://bigdatawg.nist.gov/workshop.php>.

The key objectives of this workshop were to present the NBD-PWG working draft documents and to plan future activities for the next three to six months. The report below is combined with the subgroup co-chairs meeting held in Oct. 21 and 23.

1. **Working Drafts V1.0 Publication**

The working draft documents were the results of over 140 hours of intensive debates over 3 ½ months of Web teleconferencing with over 240 submitted documents; 51general use cases (in addition, there are 9 security and privacy use cases) which are detailed use cases ranging from government operation to healthcare and life sciences; from deep learning and social media to security and privacy; from the research ecosystem to astronomy and physics and many more. From these use cases, the NBD-PWG has extracted 437 specific requirements and 35 general requirements. The working draft documents included:

1. Big Data Definitions
2. Big Data Taxonomies
3. Big Data Use Cases and Requirements
4. Big Data Security and Privacy Requirements
5. Big Data Reference Architectures Survey – White Paper
6. Big Data Reference Architectures
7. Big Data Security and Privacy Reference Architectures
8. Big Data Technology Roadmap

The goal is to publish the above documents as V1.0 under the NIST Special Publication within the next six months, including technical editing, external comments, and internal review. Due to the government shutdown for over two weeks, we have now shifted the working drafts publication process for a month.  New schedule is (some documents may complete earlier than the others):

* Oct. 31: drafts completion
* Nov. : technical writer editing
* Dec. – Jan.: RFI
* Feb. – Mar.: adjudicate comments
* Apr. – NIST: internal review
* May: publication

1. **Next Steps for NBD-PWG**

At the September 30 workshop (and earlier), there was an overwhelming interest to map use cases (or patterns of unique scenarios) to the NBD Reference Architecture (NBD RA), see Section 3 for Big Data Workshop Brainstorming on Mapping Use Cases to NBD RA.  However, details of the approach and how deep the mapping would be were not fully discussed which is why the subgroup co-chairs meeting was needed.

**Use Cases/Patterns Mapping**

Thanks to Chaitan for willing to contribute three patterns from BDBC for us to use; two of them are readily available.  In addition, Geoffrey will identify few other simple unique scenarios (real time streaming, batch processing, etc.) from the 51 use cases received.  The idea is to get the use case submitters as collaborators in the mapping and reference implementation (see below) of the NBD RA with any necessary tools and technologies (under other potential collaboration organization, see below).

Action Item:

1. Chaitan will send two readily available use cases to [bigdatasgc@nist.gov](mailto:bigdatasgc@nist.gov) by Oct. 31.
2. Geoffrey will identify few other unique big data scenarios out from the 51 submitted use cases by Oct. 31.

**Use Cases/Patterns Implementation (duration: six months)**

An approach to validate the NBD RA is to collaborate with other consortium or alliance such as the Research Data Alliance to implement selected use cases using the NBD RA. The specific responsibility between NBD-PWG and the potential collaboration organization are:

NBD-PWG

1. Identify appropriate unique use case scenarios to map use cases onto our NBD RA
2. Invite NBD-PWG use case submitters to work closely with us to specify workflow between use case and NBD RA key components
3. Work with the potential collaboration organization to provide guidance on how to map use case scenarios onto the NBD RA
4. Identify high-level interface between NBD RA key components

Potential Collaboration Organization

1. Work with NBD-PWG to implement the identified unique use case scenarios using NBD RA with any desired necessary vendor solutions or technologies
2. Create a best practices implementation guide based on the NBD RA

Action Item:

1. Wo will continue to seek potential collaboration organization such as RDA by mid Nov.

**Refresh NBD subgroups with new tasks – new call for volunteers (sometime in early Nov.)**

Since the commitment for last-call volunteers for subgroup co-chairs was from July to end of September and new tasks require a new duration of commitment (six months), the NBD-PWG will try to issue another round of call for volunteers as subgroup co-chairs once the new charters are defined (early Nov.). I strongly hope and encourage all existing subgroup co-chairs to re-apply. Most likely thee will be bi-weekly meetings for the new tasks. It can be adjusted more/less frequent depending on the need.

Action Item:

1. NBD-PWG Co-Chairs will issue a new call for volunteers in early Nov.
2. **Big Data Workshop Brainstorming on Mapping Use Cases to NBD RA**

* Abstract use cases and then develop concept. Develop specific pattern
* It is hard to move various use cases into a few well-defined (but monolithic use cases)
* Conceptually, how you distribute across clusters, but you do need to communicate between each cluster. Can we identify how that inter-process communication occurs?
* There is quite a bit of complexity in the reference architecture. Can we identify underlying patterns as they will not change?
* Come up with patterns and principles that you can bubble up…but may require to drill down
* Because BD is so diverse, applications are so different and fast moving (dynamic), going forward and in order to get performance, the applications are going to use different interfaces (based on their needs). Standardizing single interfaces may not be practical.
* A lot of the use cases may be hybrid (with mapReduce, with column restore)…different patterns….If we can boil down specific patterns, (synchronous vs asynchronous) based on a specific problem being solved.
* So, do we stay at conceptual level and define patterns? For e.g., real-time processing, batch-processing so we know what patterns are?
* Create patterns and specific data source in terms of 3V’s (or 5V’s). What exactly these means for a pattern? For this patterns, for these kinds of source, we can specify recommendations (relatively high-level)…Definition of pattern is specify the data provider and life cycle (what commands are coming from orchestrator) and then map those to Road-map.
* Create conceptual representation of the use cases and then distribute to a wider audience for input if the conceptual representation of a use case matches or fit the polled audience’s scenarios?
* Develop patterns based on the requirements that generalize requirements into specific implementable patterns. You can have patterns addressing, real-time, batch processing and/or hybrid….
* Two patterns have emerged, data warehouse (traditional relational database) and deep analytics pipeline (ingestion of real-time stream, or batch processing or machine leaning). There is a 3rd proposal from Intel on real-time analytics.
  + Transaction Processing Council, for e.g., there is one dealing with machine learning, TPC Machine Learning