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

**NBD-PWD-2014/M0286**

**UC & Requirements + Security & Privacy Meeting Minutes for Jan. 21, 2014**

Prepared by: Arnab Roy

# Agenda included:

1. Review 10 possible unique Big Data characteristic applications proposed by Bob Marcus
2. Batch process BD analytic system
3. BD system requires move data from external
4. Move data from BD framework to traditional enterprise data warehouse
5. Real time BD analytic system
6. Visualize data extracted from BD analytic processed system
7. Extract, process, and move data from BD data stores to archives
8. Interactive Analytic system for optimized database
9. Multi-users BD interactive queries system
10. Combine data from cloud with BD data stores for analytics, data mining, etc.
11. Orchestrate multiple sequential and parallel data transformation and/or analytic processing using a workflow manager
12. Compare and select unique Big Data scenarios from our 60 (51 general + 9 SnP) submitted use cases for actual implementations
13. Seek closer collaboration with the identified use case submitters

# Action Items from the meeting:

1. Pick a use case from the 60 (51 general + 9 SnP) submitted use cases.
2. Work with domain expert to specify interaction from the System Orchestrator to rest of the Reference Architecture components.
3. Construct something small, manageable, implementable in a small confined environment and model interaction between key components.

# Fragments of conversation on the bridge:

Geoffrey: What is meant by implementation?

Wo: <points to website on summary and use cases.> Pick real scenario from 51 use cases

Geoffrey: Large Hadron Collider. 300,000 cores. Assume we choose this.

Use 39: Particle Physics

Wo:

Work with the domain expert from the perspective of a domain expert. How can we build a Big Data system?

Geoffrey:

Implementation would be extremely resource consuming.

Wo:

We won’t do real implementation, but learn characteristics. Construct something small, manageable, implementable in a small confined environment and model interaction between key components. Scale back. Fingerprint searching example. Actual size in TBs. Use few MBs. System Orchestrator perspective using the RA.

David Boyd:

Maybe we are getting into too much detail. We can use existing Big Data implementations to attempt mapping to the RA.

Wo:

Future group projects. Workflow – exact interaction between components – take traditional DB – convert to Hive etc

Orit:

Suggestion to go forward - Identifying pattern – simple generic use cases – implementing some of them – useful for the use cases group to map current practices, challenges to RA.

Wo:

<points to use case 39>. Can we come up with a template for mapping?

Wo:

Capture scenario – warehouse scenario – map a scenario into RA

Geo:

Bob suggested Patterns – map use cases to Bob’s patterns and also add patterns

Orit:

Identify and document a list of patterns.

David:

How to port traditional DB to be Big Data distributed DB.

Wo:

Identify goal – data source, nosql db

Actual query

* + - 1. Detailed Scenario : MySQL -> Hive
      2. I need computing from framework provider. Need to specify interaction from SO to rest of the RA components.
      3. Actual implementation

Geoffrey: Ought to have a use case.

Wo: We can use Bob’s ten patterns.

William: The Cargo Shipping Use Case can be used.

David: Use Transaction Processing Benchmark.

Eugene: data.gov

Arnab: We should have use cases where SnP is non-trivial.

William: Intelligent Transportation system.

Wo: Use case -> patterns.

Mark: InBloom.org – K-12 data

# Bob Marcus’ list of patterns:

**1. Multiple users performing interactive queries and updates on a database with basic availability and eventual consistency (BASE)**

   Big Data File Systems as a data resource for batch and interactive queries

   NoSQL (and NewSQL) DBs as operational databases for large-scale updates and queries

   NoSQL DBs for storing diverse data types NoSQL DBs for storing diverse data types

   Databases optimized for rapid updates and retrieval (e.g. in memory or SSD)

**2. Perform real time analytics on data source streams and notify users when specified events occur**

[Operations Analysis](http://www-01.ibm.com/software/data/bigdata/use-cases/operations-analysis.html)

      Stream Processing and ETL

      Real Time Analytics (e.g. Complex Event Processing)

**3. Move data from external data sources into a highly horizontally scalable data store, transform it using highly horizontally scalable processing (e.g. Map-Reduce), and return it to the horizontally scalable data store (ELT)**

    Data input and output to Big Data File System (ETL, ELT)

    Stream Processing and ETL

**4. Perform batch analytics on the data in a highly horizontally scalable data store using highly horizontally scalable processing(e.g Map-Reduce) with a user-friendly interface (e.g. SQL like)**

[Big Data Exploration](http://www-01.ibm.com/software/data/bigdata/use-cases/exploration.html)

[Data Warehouse Augmentation](http://www-01.ibm.com/software/data/bigdata/use-cases/data-warehouse.html)

   Big Data File Systems as a data resource for batch and interactive queries

**5. Perform interactive analytics on data in analytics-optimized database**

[Big Data Exploration](http://www-01.ibm.com/software/data/bigdata/use-cases/exploration.html)

[Data Warehouse Augmentation](http://www-01.ibm.com/software/data/bigdata/use-cases/data-warehouse.html)

   Databases optimized for complex ad hoc queries

   Databases optimized for rapid updates and retrieval (e.g. in memory or SSD)

   Big Data File Systems used as a data resource for interactive queries

**6. Visualize data extracted from horizontally scalable Big Data score**

[Big Data Exploration](http://www-01.ibm.com/software/data/bigdata/use-cases/exploration.html)

   Visualization Tools for End-Users

**7. Move data from a highly horizontally scalable data store into a traditional Enterprise Data Warehouse**

    Data input and output to Big Data File System (ETL, ELT)

    Data exported to Databases from Big Data File System

[Data Warehouse Augmentation](http://www-01.ibm.com/software/data/bigdata/use-cases/data-warehouse.html)

**8. Extract, process, and move data from data stores to archives**

**9. Combine data from Cloud databases and on premise data stores for analytics, data mining, and/or machine learning**

**10. Orchestrate multiple sequential and parallel data transformations and/or analytic processing using a workflow manager**

[Big Data Exploration](http://www-01.ibm.com/software/data/bigdata/use-cases/exploration.html)

[Enhanced 360º View of the Customer](http://www-01.ibm.com/software/data/bigdata/use-cases/enhanced360.html)

[Security/Intelligence Extension](http://www-01.ibm.com/software/data/bigdata/use-cases/security-intelligence.html)

# Attendees:

1. Wo Chang
2. Gerard Fernando
3. Geoffrey Fox
4. Arnab Roy
5. Manoj Srivastava
6. Nancy Grady
7. Jon Rogers
8. Sanjay Mishra
9. Orit Levin
10. Robin (Deloitte Consulting)
11. Phil
12. David Boyd
13. Marck Underwood (Krypton Bros)
14. Rabi
15. Phil
16. Andrey Shevel
17. PhilMM
18. Tim Zimmerlin
19. Terie MM
20. Greg Wieting
21. Orest Swystun (HP)
22. William Miller
23. Peter Bajcsy
24. Eugene Luster (R2AD/DISA CTO)
25. Bill Mandrick (Data Tatics)

# Web Chat Log:

(10:01 AM) Gerard Fernando joined.

(10:01 AM) Geoffrey Fox joined.

(10:01 AM) Manoj Srivastava joined.

(10:02 AM) Nancy Grady (SAIC) joined.

(10:03 AM) John Rogers (HP) joined.

(10:03 AM) Sanjay Mishra (Verizon) joined.

(10:04 AM) Orit Levin (Microsoft) joined.

(10:04 AM) Robin joined.

(10:05 AM) Robin disconnected.

(10:05 AM) Robin ( Deloitte Consulting) joined.

(10:06 AM) Orit Levin (Microsoft): Good Morning / Good Day! No audio on web (yet)

(10:06 AM) Phil joined.

(10:07 AM) David Boyd: Ok, I was going to ask if there was web audio.

(10:07 AM) Mark Underwood (Krypton Bros) joined.

(10:14 AM) Rabi joined.

(10:17 AM) Phil disconnected.

(10:17 AM) Andrey Shevel joined.

(10:18 AM) PhilMM joined.

(10:18 AM) tim zimmerlin joined.

(10:20 AM) Andrey Shevel disconnected.

(10:20 AM) Andrey Shevel joined.

(10:21 AM) Gerard Fernando disconnected.

(10:23 AM) Mark Underwood (Krypton Bros): Something like a design walkthrough, but acknowledges that CERN's real use case is too big itself?

(10:24 AM) tim zimmerlin disconnected.

(10:24 AM) Andrey Shevel disconnected.

(10:24 AM) Andrey Shevel joined.

(10:25 AM) tim zimmerlin joined.

(10:25 AM) Andrey Shevel disconnected.

(10:25 AM) Terie MM joined.

(10:26 AM) Terie MM disconnected.

(10:26 AM) Greg Wieting joined.

(10:26 AM) Andrey Shevel joined.

(10:27 AM) Andrey Shevel disconnected.

(10:28 AM) Andrey Shevel joined.

(10:28 AM) tim zimmerlin: I cannot hear audio...

(10:29 AM) Andrey Shevel: can not here anything too

(10:30 AM) Andrey Shevel disconnected.

(10:30 AM) Mark Underwood (Krypton Bros): Audrey: Likewise, call into Phone: 206-402-0823, Participant code: 272-30-504

(10:31 AM) Orest Swystun of HP joined.

(10:34 AM) tim zimmerlin disconnected.

(10:34 AM) Terie MM joined.

(10:35 AM) Tim Zimmerlin joined.

(10:35 AM) Terie MM disconnected.

(10:40 AM) Tim Zimmerlin disconnected.

(10:41 AM) Tim Zimmerlin joined.

(10:41 AM) Mark Underwood (Krypton Bros): From the security & privacy perspective, some of the use cases should be mature enough to have representative privacy and security problems that a domain expert in that space would recognize as legitimate

(10:42 AM) Tim Zimmerlin disconnected.

(10:42 AM) Tim Zimmerlin joined.

(10:44 AM) Tim Zimmerlin disconnected.

(10:50 AM) William Miller joined.

(10:56 AM) Peter Bajcsy joined.

(10:56 AM) Manoj Srivastava disconnected.

(10:56 AM) PhilMM disconnected.

(10:57 AM) Nancy Grady (SAIC) disconnected.

(10:59 AM) Mark Underwood (Krypton Bros): In document M0281, it seems that patterns and use cases are conflated a bit. Maybe there is some sorting-out to be done there

(11:00 AM) John Rogers (HP): Geoffrey: Here's how I organize the patterns. Think of a rubics cube. The y-axis is the consumption model with three rows for Sarch, Transaction type, and Transformation type. The X-Axis shows three compute and storage models: Static EDW, Streaming, and Batch processing. The Z-Axis indicates the avaialle Data Treatment Types: Key-Value, Columnar, Document oriented, Graph, and perhaps Images. WE should be able to map all use cases to the cubes in the larger Rubic's Cube.

(11:03 AM) Eugene Luster (R2AD/DISA CTO) joined.

(11:04 AM) Robin ( Deloitte Consulting) disconnected.

(11:14 AM) Bill Mandrick (Data Tactics) joined.

(11:17 AM) Geoffrey Fox: To John Rogers: I would like to map use cases into patterns.

(11:19 AM) Bill Mandrick (Data Tactics) disconnected.

(11:20 AM) Bill Mandrick (Data Tactics) joined.

(11:21 AM) Bill Mandrick (Data Tactics) disconnected.

(11:25 AM) Peter Bajcsy disconnected.

(11:29 AM) John Rogers (HP): To Geoffrey: Which patterns. It seems that the group is having difficulty clarifying the patterns. I'll send you my grapic separately.

(11:32 AM) Geoffrey Fox: Patterns are clearly not well defined or perhaps not agreed. I like use cases as they are by definition "real"

(11:38 AM) Mark Underwood (Krypton Bros): inbloom.org

(11:40 AM) Mark Underwood (Krypton Bros): https://inbloom.org/developer-documentation

(11:42 AM) Arnab Roy (Fujitsu): M229

(11:50 AM) Geoffrey Fox: M167 page 2 has a breakup of use case 43 into stages that have a simple pattern