**Minutes of NBD-PWG 2/25/2014 – UC+SnP Meeting**

*Prepared by Arnab Roy, Fujitsu*

**Agenda**

1. Action Items
2. Place all the use cases in one place on the use cases web page to avoid confusion. Geoffrey
3. Make SnP use cases available from the SnP Requirements as part of the item #1.  Arnab
4. Make fraud detection and prevention use case available.  Atul
5. Use Case selection (5 – 6 unique Big Data scenarios with datasets and domain experts available from the submitters) and discussion
   1. 51 general summary use cases (<http://bigdatawg.nist.gov/_uploadfiles/M0291_v1_7418211592.pdf>)

[May defer database ones to Chaitan and Bob Marcus ]

1. 7, 8 can be done as open source data but better with industry support
2. 13-15, 44 can be done as earthquake disaster problem. I did in past
3. 17 can do. I know Fusheng Wang well
4. 18 can do. Lots of images these days but 17 also images
5. 19,20 genomics can do
6. 23, 24 Epidemiology can do -- I know Virginia Tech team well
7. 26 certainly, I know Stanford group well
8. 27, 28 We have projects here already
9. 30 graph processing -- I know Virginia Tech team well
10. 31 is source of data for 7, 8, 26
11. 36-38 Astronomy. Can do
12. 39-40 Particle Physics. Can do
13. 43 CReSIS can do
14. 47 -- maybe do with Reidel
15. 50 51 Internet of Things. Can do with different data (SDSC, IU)
    1. Nine SnP use cases (<http://bigdatawg.nist.gov/_uploadfiles/M0292_v1_2455397040.docx>)
16. Consumer Digital Media Usage
17. Nielsen Homescan
18. Web Traffic Analytics
19. Health Information Exchange
20. Genetic Privacy
21. Pharma Clinic Trial Data Sharing [3]
22. Cyber-security
23. Military - Unmanned Vehicle sensor data
24. Education - “Common Core” Student Performance Reporting
25. Aviation Industry
26. Cargo Shipping

**Summary of Meeting**

* Arnab went through the 11 SnP Use Cases – Document 292
* Mark Underwood talked about cyber-security use case and potential dataset availability
* Dan Samarov referred to the Yelp Dataset Challenge - <http://yelp.com/dataset_challenge>
  + The objective is to learn useful metrics about businesses from consumer contributed text
  + However, the dataset is only 135MB
* Tim Zimmerlin suggested looking at data.gov
  + One concern is that the datasets are small.
* Ilkay talked about general use cases – Document 291
* Bob Marcus suggested that the best way to validate the Reference Architecture and explore possible standardizations is to implement simple generic use cases. Diving deeply into specific applications will be very resource consuming and not be generalizable or useful to Reference Architecture enhancement and standards development. The right abstraction level will allow us to validate the Reference Architecture and possible standardizations based on best practices for Big Data solutions. One of the key questions is how do open source solutions map to the Reference Architecture and can they be a starting point for future standardizations.

**Attendees**

* + - 1. Steven McGee
      2. Arnab Roy
      3. Bob Marcus
      4. Mark Underwood
      5. Nancy Grady
      6. Tim Zimmerlin
      7. Manoj
      8. Jon Hudson
      9. Ilkay Altintas
      10. PavithraKenjige
      11. Andrey Shevel
      12. Prachi Sukhatankar
      13. Ovace A. Mamnoon
      14. Atual Mathur
      15. Pw
      16. Alicia M. Zuniga-Alvarado

**Web Chat Log**

(10:05 AM) Steven McGee SAW Concepts LLC joined.

(10:07 AM) Arnab Roy (Fujitsu, Host) joined.

(10:10 AM) Bob Marcus disconnected.

(10:10 AM) Bob Marcus joined.

(10:10 AM) Bob Marcus disconnected.

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

(10:10 AM) Bob Marcus joined.

(10:11 AM) Nancy Grady(SAIC): is the VOIP connected?

(10:11 AM) Bob Marcus: I can't hear anything

(10:12 AM) Tim Zimmerlin joined.

(10:12 AM) Nancy Grady(SAIC): phone it is i guess

(10:12 AM) Mark Underwood (Krypton): Dial-in Phone: 206-402-0823, Participant code: 272-30-504

(10:13 AM) Manoj (CyberIQ): Where is this usecase doc posted?

(10:13 AM) Nancy Grady(SAIC): The bridge to VOIP now seems to be connected

(10:13 AM) \_Jon\_Hudson\_(\_Brocade\_) joined.

(10:19 AM) Ilkay Altintas joined.

(10:19 AM) Mark Underwood (Krypton): Yelp data challenge? News to me

(10:21 AM) Mark Underwood (Krypton): http://yelp.com/dataset\_challenge

(10:27 AM) PavithraKenjige joined.

(10:36 AM) Andrey Shevel joined.

(10:37 AM) Prachi Sukhatankar (SAIC) disconnected.

(10:38 AM) Ovace A. Mamnoon (HP) joined.

(10:41 AM) Bob Marcus disconnected.

(10:42 AM) Bob Marcus joined.

(10:47 AM) Mark Underwood (Krypton): I got the Yelp DB - It's ony 135MB (tar)

(10:47 AM) Mark Underwood (Krypton): \*only

(10:48 AM) Bob Marcus: I think that the best way to validate the Reference Architecture and explore possible standardizations is to implement simple generic use cases.

(10:50 AM) Bob Marcus: I am concerned that diving deeply into specific applications will be very resource consuming and not be generalizable or useful to Reference Architecture enhancement and standards development.

(10:54 AM) Atul Mathur disconnected.

(10:58 AM) 115333 joined.

(11:00 AM) Ilkay Altintas disconnected.

(11:00 AM) 115333: We're interested in working on the Fraud Detection Use Cases area....as we're currently designing a tool for fraud detection, analysis, forecasting and prosecution.....Pw

(11:01 AM) 115333: Via a Cloud/Big Data Eco-System of course....

(11:01 AM) Mark Underwood (Krypton): Maybe someone who's clear on Wo's objective with "live" data use could put the task of threading this needle into text. It seems to go in and out of focus for me.

(11:03 AM) 115333: We believe Mr. Chang is refering to 'real world' data that is publically available.....our thoughts.....

(11:05 AM) Manoj (CyberIQ): @11533 I sure hope that is the case i.e. publically available data. getting hands on big data would be very difficult IMO. Any suggestions/examples of these publically available data sources?

(11:05 AM) 115333: Yes....government agencies has lots of publically available data.....

(11:06 AM) Manoj (CyberIQ): any specific URLs/Source you can point us to?

(11:08 AM) 115333: NIH, FDA, NASA, DoD, Government Archives.....State and Federal Agencies......et al.....Pw Carey, Compliance Parters, LLC (we forgot to sign-in at the beginning...)

(11:08 AM) Andrey Shevel: many sci and gov institutions have a lot of big data however to get practical access to you need to be in contact with appropriate persons.

(11:10 AM) Bob Marcus: The right abstraction level will allow us to validate the Reference Architecture and possible standardizations based on best practices for Big Data solutions

(11:12 AM) Bob Marcus: One of the key questions is how do open source solutions map to the Reference Architecture and can they be a starting point for future standardizations

(11:16 AM) Alicia M. Zuniga-Alvarado disconnected.

(11:16 AM) 115333: Bob we'll send you some open source info.....Pw