**NBD-Requirements Meeting Minutes July 30 2013**

The meeting was dogged by poor audio and modest attendance but progress was made. The first half of meeting was devoted to discussion of current use template and some issues were discussed.

a) Ambiguity between Data Velocity and Data Variability fields -- we decided to leave as is

b) Ambiguity between Data Quality and Data Veracity fields. It was suggested that Quality refer to semantics (meaning) of data and Veracity to technical (syntax) correctness.

c) Significance of workflow in Data Variety -- no changes were made.

We agreed to consider use case template agreed and it has already been sent to full working group mailing list.

The Security subgroup has much more details on issues in “Security & Privacy Requirements” field in 3 verticals: Health, Retail and Social. We did not identify other sources of use cases. We discussed an emergency response use case (DHS) that would be good to have.

We spent the rest of session discussing Remote Sensing (mainly) and Web Search examples to identify general requirements for other subgroups. This was an active discussion. The constraint that radar signal processing dominantly Matlab produced a lively discussion. Is this a “legacy software” anomaly or a profound constraint. We also realized that both use cases really had several distinct activities that perhaps should be entered separately.

Remote sensing has field expedition processing and batch processing of full data sample.

Web Search has Crawling, document clean-up, Production of inverted index, calculation of relevance (as in Pagerank) and search.

**Log of Chat Session**

(7:55 AM) Bob Marcus(ET-Strategies) joined.

(8:01 AM) Wo Chang joined.

(8:03 AM) Geoffrey Fox joined.

(8:05 AM) Geoffrey Fox disconnected.

(8:06 AM) Alicia Zuniga-Alvarado/AZA joined.

(8:07 AM) Geoffrey Fox joined.

(8:07 AM) Geoffrey Fox: I was disconnected!

(8:14 AM) Tony joined.

(8:14 AM) Tony disconnected.

(8:15 AM) Geoffrey Fox: Any comments on velocity v variability

(8:17 AM) Wo Chang: I am here but with no audio mike

(8:19 AM) William Miller (MaCT USA) joined.

(8:19 AM) William Miller (MaCT USA): yes

(8:20 AM) William Miller (MaCT USA): did not finish it yet

(8:20 AM) Wo Chang: I need to send the Use Case template to the Big Data WG.

(8:20 AM) William Miller (MaCT USA): template is ok

(8:20 AM) William Miller (MaCT USA): should contain diagrams if avaiable

(8:21 AM) William Miller (MaCT USA): attach if avaialbe

(8:22 AM) Wo Chang: I will do it soon.

(8:23 AM) PavithraKenjige joined.

(8:23 AM) Wo Chang: I agree with Geoffrey, I will send the Use Case Template with those 5 samples.

(8:23 AM) Wo Chang: since we have 2 extra use cases this time, can we go over the exercise to extract the requirements?

(8:26 AM) Wo Chang: I should be done with my current meeting in ~20 minutes then I can actively engage with discussion for this telecon.

(8:29 AM) William Miller (MaCT USA): alot of manual processes

(8:29 AM) William Miller (MaCT USA): as far as interpretating of the information

(8:31 AM) William Miller (MaCT USA): note NIST will have a meeting of the NIST Data Science and Anlytics on Nov. 18-29 - They want to define metrics and standards for Brig Data for example

(8:33 AM) William Miller (MaCT USA): a couple of questions might be are they using metadata and what type

(8:33 AM) William Miller (MaCT USA): may be hard to determine tho

(8:33 AM) Wo Chang disconnected.

(8:35 AM) William Miller (MaCT USA): <http://www.uicds.us/>

(8:35 AM) William Miller (MaCT USA): this is an application area

(8:35 AM) William Miller (MaCT USA): the data is held by first responder organizations

(8:35 AM) William Miller (MaCT USA): check out the link

(8:35 AM) Wo Chang joined.

(8:35 AM) William Miller (MaCT USA): Unified Incident Command and Decision Support

(8:35 AM) William Miller (MaCT USA): DHS

(8:36 AM) William Miller (MaCT USA): They are still working on putting up the architecture

(8:36 AM) William Miller (MaCT USA): I am a technology provider to UICDS

(8:36 AM) William Miller (MaCT USA): and DHS

(8:37 AM) William Miller (MaCT USA): i can try i have alot of information but they may not conisder themselve big data at the moment

(8:37 AM) William Miller (MaCT USA): they share emergency repsonse information when needed

(8:37 AM) William Miller (MaCT USA): by mutual agreement

(8:38 AM) William Miller (MaCT USA): it is growing many applications

(8:38 AM) William Miller (MaCT USA): we are still trying to tie-in sensors directly

(8:38 AM) William Miller (MaCT USA): web cams are already used

(8:38 AM) William Miller (MaCT USA): confident to those who have a mutual agreement

(8:40 AM) William Miller (MaCT USA): <http://uicds.kzoplatform.com/swf/player/1219&c=webcast>

(8:40 AM) William Miller (MaCT USA): this is a link to the introduction to UICDS

(8:41 AM) Geoffrey Fox: wo -- I can hear you

(8:41 AM) William Miller (MaCT USA): shows they types of information that may be considered big data

(8:41 AM) William Miller (MaCT USA): not normally avaialbe to the public

(8:41 AM) Wo Chang: Arnab: why don't you finish what you tried to say.

(8:41 AM) William Miller (MaCT USA): they due use social medai and smart devices

(8:50 AM) William Miller (MaCT USA): use raw data

(8:51 AM) William Miller (MaCT USA) disconnected.

(8:51 AM) William Miller (MaCT USA) joined.

(8:51 AM) William Miller (MaCT USA): <http://www.mathworks.com/help/matlab/import_export/supported-file-formats.html>

(8:52 AM) William Miller (MaCT USA): these are the file type supported by Matlab

(8:52 AM) William Miller (MaCT USA): they also support XML

(8:52 AM) William Miller (MaCT USA): check out the web site

(8:52 AM) William Miller (MaCT USA): should not be a problem

(8:53 AM) William Miller (MaCT USA): They need to have metadata to map to the files so the correct information can be located

(8:53 AM) William Miller (MaCT USA): still requires alot of manual intervention

(8:55 AM) Wo Chang: sorry my Internet is very poor

(8:56 AM) Geoffrey Fox: CReSIS is a big project producing comparatively homogenous dat. POlar Science has lots of "long tail" science typically stored in Excel spreadsheets. NASA (NSIDC) and NSF have prjects improving access/metaadata

(8:57 AM) Wo Chang: I will try to call in via audio bridge

(8:57 AM) William Miller (MaCT USA) disconnected.

(8:57 AM) Bob Marcus(ET-Strategies) disconnected.

(9:00 AM) PavithraKenjige: i can not hear anything

(9:01 AM) William Miller (MaCT uSA) joined.

(9:01 AM) William Miller (MaCT uSA): keep getting kicked off sorry

(9:01 AM) Arnab Roy (Fujitsu, Guest): audio bridge is breaking up

(9:01 AM) William Miller (MaCT uSA): Too much congestion

(9:02 AM) PavithraKenjige: Is Use Case Description same as Scenerio

(9:02 AM) William Miller (MaCT uSA): cannot hear on the web via VoIP

(9:02 AM) Geoffrey Fox: I think it includes scenario as "use case description" etc.?

(9:03 AM) Wo Chang: Arnab: can you hangup and redial in using your Moderator code?

(9:03 AM) Arnab Roy (Fujitsu, Guest): sure

(9:03 AM) Geoffrey Fox: William can you hear me or should I switch to phone?

(9:04 AM) Wo Chang: I think using the audio brdige would be better

(9:04 AM) Wo Chang: i could hear Arnab very clear

(9:04 AM) Arnab Roy (Fujitsu, Guest): I see ... but your voice is breaking up Wo

(9:04 AM) Geoffrey Fox: So could I.

(9:05 AM) Geoffrey Fox: Should I call in at 206-402-0823, Participant code: 272-30-504

(9:05 AM) Wo Chang: Geoffrey: could you try to call in using audio bridge?

(9:05 AM) Wo Chang: yes

(9:06 AM) Alicia Zuniga-Alvarado/AZA: now we can hear you

(9:06 AM) Wo Chang: I could hear you Geoffrey.

(9:06 AM) Alicia Zuniga-Alvarado/AZA: on the phone line

(9:07 AM) Wo Chang: yes the meeting is still on

(9:07 AM) PavithraKenjige: I looked at the use case template that is displayed

(9:07 AM) PavithraKenjige: and it looks fine

(9:07 AM) PavithraKenjige: We can hear now.. audio seems to be working

(9:07 AM) Wo Chang: looks like my audio coming is pretty poor so I will type instead.

(9:08 AM) Wo Chang: Looks like from the NASA's case, there may be a requirement to support legacy software such as C/C++. Matlab could be tough to support for the MR environemtn

(9:09 AM) PavithraKenjige: Use case templates needs to support multiple scenerios

(9:09 AM) PavithraKenjige: including alternate and exceptions

(9:09 AM) PavithraKenjige: If they are detail level

(9:09 AM) PavithraKenjige: I did not see options for that

(9:10 AM) Wo Chang: I don't hear any audio: is this mean nobody is speaking?

(9:10 AM) Wo Chang: yes

(9:10 AM) Geoffrey Fox: why aren't multiple scenarios multiple use cases?

(9:10 AM) PavithraKenjige: BUt such things seems to be covered in descriptions

(9:11 AM) Wo Chang: I could hear you clear Geoffrey

(9:11 AM) PavithraKenjige: I am not suggesting changes

(9:11 AM) PavithraKenjige: I am trying to understand

(9:11 AM) PavithraKenjige: the way this is formatted

(9:12 AM) PavithraKenjige: You have use Case Desciptions

(9:13 AM) Alicia Zuniga-Alvarado/AZA: lots of eko

(9:14 AM) Wo Chang: You may want to disable your computer mic

(9:15 AM) Geoffrey Fox: I think you are sdaying similar thing to Wo; we need to draw high level generic requirements

(9:16 AM) William Miller (MaCT uSA) disconnected.

(9:16 AM) Alicia Zuniga-Alvarado/AZA: cant hear nothing on the bridge or via web

(9:17 AM) Wo Chang: I can hear you

(9:17 AM) Geoffrey Fox: i'm muted -- I can hear you!

(9:18 AM) Alicia Zuniga-Alvarado/AZA: yes I can hear you her too

(9:18 AM) Geoffrey Fox: yes continue

(9:19 AM) William Miller (MaCT USA) joined.

(9:22 AM) Wo Chang: Let's look at the use case template another way: let's assume we are the big data computing provider and we have customer submitting their needs and how we identify issues that we can help customer's problem.

(9:22 AM) William Miller (MaCT USA): Communications constraints

(9:23 AM) Geoffrey Fox: Yes this system is fully operational

(9:24 AM) William Miller (MaCT USA): mchaines are obsolete will have trouble getting parts and may be incompatiable with latest software

(9:26 AM) William Miller (MaCT USA): what is the OS?

(9:27 AM) Bob Marcus joined.

(9:27 AM) William Miller (MaCT USA): good

(9:28 AM) William Miller (MaCT USA): how about SATCOM?

(9:30 AM) Wo Chang: I think Matlab can produce c/c++ library at least compile to executable

(9:31 AM) Geoffrey Fox: Yes but it's not so smooth

(9:31 AM) Wo Chang: supporting Matlab or not is an issue but overall the big data ecosystem may need to support common legacy software/libs

(9:31 AM) William Miller (MaCT USA): r we going to propose some solutions

(9:31 AM) William Miller (MaCT USA): or just leave it open

(9:32 AM) Geoffrey Fox: I think this example has some quite useful specific requirements? We "just" need to get more like it?

(9:33 AM) Wo Chang: Yes, our goal is to get the use cases' requirements

(9:34 AM) Wo Chang: Let's look at the "Visualization" requiremetn: is there anything we can draw from this use case?

(9:34 AM) Geoffrey Fox: This is low power field cluster, Matlab, disk shipping, GIS, lots of jobs

(9:35 AM) Wo Chang: do they have any spefici output format or pre-rendering needs?

(9:40 AM) Wo Chang: So, the whole scenario is: lots of collected radar data on removable disk and they have matlab software for analysis but they want to run their data and analysis in a big data environment and expect the output in either matlib or jpeg images

(9:42 AM) Geoffrey Fox: Then jpegf/ml format gets run thru image processing which is mix of "undergrads" and image understanding algs

(9:44 AM) Geoffrey Fox: <http://grids.ucs.indiana.edu/ptliupages/presentations/mitchell_RESworkshop_05061.pdf> is typical talk

(9:44 AM) Wo Chang: so the requirements are: a) need to be able to transfer data from legacy data source either through tape or removalbe dsik, b) be able to run legacy softeware ml in a parallel fashion, c) be able to output the reuslt in either legacy code (ml) or jpeg image

(9:45 AM) William Miller (MaCT USA): this is a very difficult use case

(9:45 AM) Geoffrey Fox: BTW I wouldn't say legacy data source. This is highly active expanding area using best practice for "radart informatics"

(9:47 AM) Wo Chang: sorry I don't mean legacy data source; I mean legacy data media (tape/removalbe disk)

(9:47 AM) Geoffrey Fox: The layer data goes into boundary conditions of PDE solvers that predict melting of ice caps and rise in sealevel

(9:48 AM) Wo Chang: Now we are talking the workflow how data been processed from one tool to antoher, etc. This is pretty typical

(9:48 AM) Geoffrey Fox: It might be interesting to know when we will have pervasive high bandwidth links so disk transfer can be removed

(9:49 AM) Geoffrey Fox: I agree this is a nice example of workflow as seen in many cases

(9:49 AM) Geoffrey Fox: This discussion suggests we should aim at more detail?

(9:50 AM) Wo Chang: So another requirement is: the RA needs to support data processing workflow for various analysis needs and the order of the workflow can be important

(9:50 AM) Bob Marcus disconnected.

(9:51 AM) Wo Chang: since we only have about 10 minutes left: can we switcfh to the web search use case?

(9:52 AM) Geoffrey Fox: Note data volumes change. There is raw radar data (large), processed "information) as large image files and "tiny amount of knowledge" (depths)

(9:52 AM) Geoffrey Fox: OK to search

(9:53 AM) Geoffrey Fox: Here I am just a reader of tutorials on web

(9:53 AM) Geoffrey Fox: I listed URL's for these

(9:56 AM) Wo Chang: Are those Pre-process data distributed at diff. sites?

(9:56 AM) Geoffrey Fox: I think Google etc. bring back crawled data to "central cloud"

(9:57 AM) Wo Chang: If yes, then one possible requirement is: be abel to connect multiple data sources

(9:57 AM) Geoffrey Fox: A lot of work could be distributed but inverted index needs to bring together (reduce phase in MapReduce)

(9:58 AM) Geoffrey Fox: As volume says there are an estimate of 45 billion (10 billion for Bing) external web pages

(9:59 AM) Geoffrey Fox: The 500 million photos each day are dominantly facebook

(9:59 AM) Geoffrey Fox: There source are distributed users going to FB cloud?

(10:00 AM) Wo Chang: I think at this point: I would prefer separate the actual analytic processing from the interaction between data source and the transformer (analytic engine)

(10:01 AM) Wo Chang: Just try to see what kind of requirements that we can leaarn/draw from this use case.

(10:01 AM) Wo Chang: It seems those pages can coming in as a streaming data

(10:01 AM) PavithraKenjige disconnected.

(10:02 AM) Geoffrey Fox: This casse like CReSIS has multiple phases which can be considered separately

(10:02 AM) Wo Chang: agree.

(10:02 AM) William Miller (MaCT USA) disconnected.

(10:03 AM) Geoffrey Fox: There is discussion of handling of streaming data and only build full inverted index ebery now and then

(10:03 AM) Wo Chang: OK, looks like our time is up but I will try to send out the use case template to the NBD-Wg for more input.

(10:03 AM) Geoffrey Fox: Thank you

(10:03 AM) Wo Chang: thanks ARnab for your help!!!

(10:04 AM) Alicia Zuniga-Alvarado/AZA: were you discussion the CReSIS doc

(10:05 AM) Wo Chang: use case itself

(10:05 AM) Alicia Zuniga-Alvarado/AZA: Thanks

(10:05 AM) Wo Chang: arnab: can you capture the chat log and send it to Geoffrey ?

(10:05 AM) Wo Chang: Thanks much Arnab!

(10:05 AM) Arnab Roy (Fujitsu, Guest): Sure Wo

(10:05 AM) Arnab Roy (Fujitsu, Guest): Glad to help!

(10:06 AM) Wo Chang: sorry about the audio part; talk to you later.

(10:06 AM) Wo Chang: bye

(10:06 AM) Alicia Zuniga-Alvarado/AZA: bye

(10:06 AM) Alicia Zuniga-Alvarado/AZA disconnected.