**NIST BD Ref. Arch. + Tech. Roadmap meeting minutes for Jan 23, 2014**

Prepared by Nancy Grady, SAIC

**Agenda**

A. Def&Tax: identify areas that need further development.

Examples include:

1. Framework Provider: drill down into each of the sub-blocks: Processing, Platforms, and Infrastructures. (The discussion on the list is a great starting point)

2. Privacy: identify the terminology used to describe: data types (user content, user account info, etc.); security measures/techniques (encryption, de-identification, etc.) data usage (internal, 3rd party, etc.)

3. Map Apache+HPC specific technologies onto our mind map as examples how they fit into our taxonomy

B. RefArch:

1. Identify what “pattern” do we want to include or work on

2. How do we get the “pattern” definitions and when?

3. Need to think hard “how” to communicate/interface between RA key components

C. TechRoadmap: any issues to be discussed?

**Action Items**

1. Need to define what we mean by the term pattern vs use cases
2. Need to extract patterns from use cases and create candidate patterns from solution attributes (simple as possible)
3. Need to begin to map these patterns to the reference architecture components
4. Nancy post mindmap to document repository
5. Next meeting in two weeks
6. Use the email list to continue discussion

**Notes**

Quick overview of what has been done so far, to see where we are as we start this phase 2.

Review the latest mindmap at

<https://www.mindmeister.com/322462463>

Our next focus will be on the Framework Provider. We want to evaluate what needs to be added and why, with text descriptions. This will allow us to ensure that we’re all clear on the nature of the new elements.

We need to consider additional elements describing data processes such as encryption inline where these processes relate.

We want to consider adding in specific tools/technologies to the taxonomy mind map to help for clarification and to let folks analyze the entries for completeness. We will continually be evaluating whether some concepts are needed for the discussion of big data. Given that we don’t want to duplicate work done elsewhere, we don’t want to describe the world, but only enough of the existing concepts that are needed for describing the landscape of the new big data approaches.

Discussion of the location of different concepts, Governance is contained in the Orchestrator; Policy is between Orchestrator and Data Provider, etc

Want to look at patterns as a mechanism to categorize use cases.

Want to consider a sandbox to clarify the interactions between tools. While a full sandbox is an ambitious agenda, it would be useful for clarifying system component interactions.

Q: How should we begin to collect the patterns we can extract from the use cases. One suggestion is that these should reside within the system architecture discussion. Another is that the use case subgroup should take a first cut to distill the use cases into a much smaller number of functional patterns. Then these distilled patterns could be placed as an overlay on the reference architecture to see how it fits.

Use cases represent a particular user problem, patterns are intended to represent the solutions.

One important process is quality.

One possible pattern relates to the high availability attribute of big data. To ensure availability, data is replicated across nodes, with some form of fail-over approach.

Discussion of how we can create the sandbox, try-and-learn setup.

Steps in the sandbox effort with RDA:

1. Description of the scenario
2. How do the steps map into the architecture components
3. Implement in the sandbox in the future group

Takes 30 days to set up, so these meetings will begin in the future

**Online Attendee List**

1. Wo Chang
2. Orit Levin
3. Nancy Grady
4. Any Hu
5. David Boyd
6. Chetan
7. John Rogers
8. Sanjay Mishra
9. Venkat Reddy
10. Geoffrey Fox
11. Nenad Ivezic
12. Felix Njeh
13. Ian Gorton
14. Manoj Srivastava
15. Tim Zimmerlin
16. Andrey Shevel
17. Eugene Luster
18. Ron Batdorf
19. William Reith
20. Bill Mandrick
21. K. Eric Harper
22. Venkat Reddy

**Online Chat log**

(1:15 PM) chetan: hi Wo Is this meeting getting recorded then we can take a look at it later point of time

(1:16 PM) Wo (guest): Chetan: no, session is not recorded due to privacy issues.

(1:22 PM) John Rogers (HP): Nancy, can you provide us with the link to the mind map so that we read the layers?

(1:24 PM) Andrey Shevel: Hi Nancy, it seems to me there is another topic you did not mention. It is BigData moving from one site to another one.

(1:25 PM) Wo (guest): For those would like to review and help edit the mind map, please send Wo an email so that I could give you permission.

(1:31 PM) Manoj Srivastava (CyberIQ): I agree that using technologies/products as example is a great idea. Using examples does not violate principle of staying tech agnostics

(1:33 PM) Ron Batdorf: What EA are you using to show how big data even comes into play? Tools are interfaced by the language used such as xml etc. Are you looking for what the tools have or are you thinking about what plug ins are we able to use.

(1:34 PM) K. Eric Harper (ABB): Would it help to classify the types of analytics? E.g. batch, streaming, etc.

(1:35 PM) David Boyd (Data Tactics): What Nancy is saying was where my thought processes were going in laying out the types.

(1:36 PM) David Boyd (Data Tactics): Some tools will likely cross types and involve multiple types. Columnar/BigTable implementations are really enhanced implementations of KeyValue stores.

(1:37 PM) Ron Batdorf: queing is a function therefore more centered on applications not the data; so I would keep it separately and independent of the data, IMO

(1:37 PM) Nancy Grady (SAIC): The mindmap is at https://www.mindmeister.com/322462463. if you want edit access you need to contact Wo

(1:38 PM) David Boyd (Data Tactics): Reference to queues while they have been around for a long time, recent advances have looked to provide horizontal scaling attributes.

(1:39 PM) John Rogers (HP): Perhaps rather than concentrating on Data Consumer and Data Providers as actors, we might regroup these into Consumption models: Search, Transaction, Transformation, and Storage/Persistance.

(1:40 PM) David Boyd (Data Tactics): For some of the existing leaves on the mind map I think some detailed definitions would be helpful to better deliniate the differences.

(1:40 PM) K. Eric Harper (ABB): I see know that you have the topic Processing Frameworks

(1:41 PM) K. Eric Harper (ABB): How about the capabilities for workflow / algorithm design and specification?

(1:42 PM) K. Eric Harper (ABB): Might you expand the node for Data Scrubbing?

(1:42 PM) David Boyd (Data Tactics): @Eric - I belive those tools would sit in the APplication provider framework.

(1:42 PM) Nancy Grady (SAIC): @Ron Batdorf: sorry I'm behind the chat. I think we first need to identify the components, then we want to identify how they interoperate

(1:44 PM) John Rogers (HP): I agree that we should address the issues that are bundled under System Orchestrator. Perhaps we should bundle the Data Provider issues under System Orchestration as well. That way we put all of the policy interactions in one place.

(1:44 PM) Nancy Grady (SAIC): @K. Eric Harper: You raise a good point, we didn't want to create an analytics taxonomy since most all techniques pre-date big data, but we do need to break the functionality as you say batch, real-time, interactive because those are functional differences among solutions

(1:44 PM) Ron Batdorf: @Nancy Grady you are somewhat on track but the problem we have is that the business models gives us the components and when we only look at the components without the business models we miss some of the important attributes of data.

(1:46 PM) Nancy Grady (SAIC): @ Eric: we would welcome you having a look at scrubbing for its completeness

(1:46 PM) Nancy Grady (SAIC): @Eric: I think workflow etc should be expended in the architecture design in the system orchestrator

(1:48 PM) K. Eric Harper (ABB): I was thinking that Data Scrubbing would also include data quality expectations.

(1:48 PM) Nancy Grady (SAIC): @John: you are correct that policy is a negotiation between the data provider (may be in another organization) and the system orchestrator (operating on behalf of the owner of the system we're describing)

(1:48 PM) K. Eric Harper (ABB): Maybe that is an expectation of the Data Provider?

(1:50 PM) K. Eric Harper (ABB): Where does the non-functional requirements / quality attributes come in?

(1:50 PM) Nancy Grady (SAIC): @Ron Batdorf: we have had prior discussion about whether we're able to represent different business models using our architecture. I think just as the use cases will help clarify tools used and how they inter-operate, we could consider some business models, say for example when you have a system built using the resources from multiple actors (sits on cloud, some part uses a hosted application, etc)

(1:51 PM) K. Eric Harper (ABB): That might help filter the architecturally significant use cases, what do you thiink?

(1:53 PM) K. Eric Harper (ABB): Nancy, you have some good 3Vs patterns in your presentation at the Government Big Data Symposium.

(1:54 PM) Ron Batdorf: Use cases have to be based on users not systems. This opens up the flow of data that is not hard wired. A user centric view imo should be a starting position.

(1:55 PM) Nancy Grady (SAIC): @ Eric Harper: yes it looks like data scrubbing is sitting out on its own. It needs to be placed in security (for the ones listed), and there should be different categories more related to quality that should be inside the transformation, where the application provider has to reach into their toolbox of resources to address application-specific issues

(1:55 PM) David Boyd (Data Tactics): Wo is on track. THe idea was to create generalized versions of the use cases.

(1:56 PM) Nancy Grady (SAIC): @Eric Harper: those types of 3V's patterns to me are helpful to catagorize use cases

(1:58 PM) David Boyd (Data Tactics): The general cases can then be expanded. For example in the general scenerio I posted to the reflector I did not address privacy but that scenerio can easily be expanded to add privacy requirements and elements.

(1:59 PM) David Boyd (Data Tactics): The other day in discussing the implementation one idea that was floated was to leverage the TPC data sets and generators as a base data set.

(2:02 PM) Geoffrey Fox: I think use case group could collect and record mapping of patterns to use cases but all choices resonable

(2:03 PM) Geoffrey Fox: Ref arch as suggested is fine! I hope patterns are related to use cases

(2:04 PM) Geoffrey Fox: There is enough work that we should all groups that want to should contribute

(2:05 PM) Manoj Srivastava (CyberIQ): @nancy. I think that's a good point on how to identify these patterns. once patterns are identified they can be enumerated in Ref Arch.

(2:07 PM) K. Eric Harper (ABB): Would not a focus on the quality attributes bring focus to what is needed in the reference architecture?

(2:12 PM) Geoffrey Fox: There is a concept of "software defined systems" generalizing SDN . There are many tools and languages such as opentosca

(2:14 PM) Ron Batdorf: You might want to look at LEADing Practice which has done much of the work of combining all architectures with business model in general and they have already associated the patterns within an Enterprise approach which is more networked instead of linear. Just a suggestion.

(2:15 PM) Wo (guest): Ron: if you can provide a link to it, that will be great!

(2:16 PM) Ron Batdorf: http://www.leadingpractice.com/wp-content/uploads/2013/05/LEADing-Practice-FAQ.pdf> This gets you started with the FAQ's of it.

(2:17 PM) K. Eric Harper (ABB): There is a way to trace non-functional requirements back to QAs.

(2:17 PM) Wo (guest): Tim: can you post the title of the book?

(2:17 PM) K. Eric Harper (ABB): Thank you for the kind words.

(2:18 PM) K. Eric Harper (ABB): How will you know when you are done if you do not have some expectations?

(2:19 PM) Ron Batdorf: Pushing bits is a problem until you knows what is the business models first. We keep building infrastructure without integration of business models. We have this same problem within DoD we are building JIE now without the details.

(2:21 PM) K. Eric Harper (ABB): If the focus is only on functionality, then IMHO the current approach is OK. If you are interested in suitability of the architecture, an extension of the existing approach is needed. What do you think?

(2:22 PM) Andrey Shevel: Hi Nancy, I am still not sure that your map is available to be read.

(2:22 PM) K. Eric Harper (ABB): Looks like you need to buy a license.

(2:22 PM) Andrey Shevel: I do have personal license.

(2:23 PM) Nancy Grady (SAIC): Mindmeister can be read if Wo gives your email (username) permission. A personal license is only required to extract the data. I purchased one so I can pull down the data and save copies of the mindmap file.

(2:24 PM) Geoffrey Fox: There is a famous parallel computing concept of dwarves described by berkeley and helpful in that field

(2:26 PM) Manoj Srivastava (CyberIQ): @nancy is it possible to make the .mind file available at the NBD website?

(2:26 PM) Andrey Shevel: I se. I sent request to Wo.

(2:27 PM) Geoffrey Fox: Quality was in use case template

(2:27 PM) David Boyd (Data Tactics): So I am looking at this a bit like a Mr. Potato head. THere are nearly infinate ways many of the sub-patterns can be put together. What is important is what combination meets the functional and business patterns of the specific use case.

(2:28 PM) Nancy Grady (SAIC): yes, I can post a copy of the .xmind file (or some other like .mm if folks can't translate)

(2:29 PM) Ron Batdorf: Mr. Potato head might not be a bad reference. Because agility comes from when humans make or change decisions and if the data can be extracted for this change in sequencing would be what we need for the Warfighter.

(2:30 PM) K. Eric Harper (ABB): .mm is good for me.

(2:30 PM) Geoffrey Fox: We can add new use cases if important "pattern" not present today

(2:31 PM) Ron Batdorf: Patterns need to be understood from both the sequencing of the functions and data besides the functions and data.

(2:32 PM) David Boyd (Data Tactics): @Ron - good point. There are multiple ways these things can be put together.

(2:32 PM) Ian Gorton, CMU SEI: many many ways - David is right

(2:32 PM) David Boyd (Data Tactics): I would really like to see an example of what everyone here thinks is a pattern.

(2:32 PM) Ian Gorton, CMU SEI: and a particular composition of patterns can only make sense in teh context of a given problem

(2:33 PM) Ian Gorton, CMU SEI: David - ne too ...

(2:34 PM) David Boyd (Data Tactics): I see patterns in individual elements of the Ref Arch. Eg. I have a data warehouse provider or I have a real time streaming provider or I have a transaction oriented provider.

(2:34 PM) David Boyd (Data Tactics): On the Analytic provider side I may need Complex Event processing or regression, or pattern analysis.

(2:35 PM) David Boyd (Data Tactics): @Ian - Right On.

(2:36 PM) Ron Batdorf: You basically have four use cases; human to human; human to machine; machine to human; and machine to machine. If you pass data and use functions (either machine embedded or human) you can come up with all the requirements for these four views, IMO

(2:37 PM) David Boyd (Data Tactics): To whomever is talking I agree - you need to pick the solutions/patterns that meet your specific requirements as close as possible - then customize. The requirements include business limitations (funds, bandwidth, equipment, etc.)

(2:38 PM) Geoffrey Fox: Sorry need to leave

(2:41 PM) K. Eric Harper (ABB): One approach we have considered is to look at not just one architecture, but one that has small, medium, and large sizes. What do you think?

(2:43 PM) David Boyd (Data Tactics): Have we done any mapping of the decomposed requirements from the use cases to the ref arch components?

(2:43 PM) Ian Gorton, CMU SEI: David - not that I know of - that would be interesting ...

(2:44 PM) David Boyd (Data Tactics): That might help us define patterns within the ref arch components

(2:44 PM) Tim Zimmerlin: Wo, I found the text on Amazon as "Release It!" by Nygard.

(2:44 PM) Nancy Grady (SAIC): @David and @Ian I think that would be a good first exercise for this group

(2:45 PM) Wo (guest): Thanks Tim.

(2:45 PM) Ian Gorton, CMU SEI: @David, Nancy - yep - it might give us a framework for doing this exercise. It all seems rather unconstrained to me right now, and its just going to cause mor e'busy owrk'

(2:45 PM) David Boyd (Data Tactics): We could define the saliant characteristics of each arch component.

(2:46 PM) David Boyd (Data Tactics): based on the mapped requirements.

(2:54 PM) David Boyd (Data Tactics): Wo can you please post information on participating in the RDA meetings.

(2:56 PM) K. Eric Harper (ABB): IHMO this effort will lack focus until we choose an application domain and specific architecturally significant scenarios. What do you think?

(2:56 PM) K. Eric Harper (ABB): How will be data warehouse be used?

(2:57 PM) K. Eric Harper (ABB): be -> the

(2:58 PM) Ian Gorton, CMU SEI: Eric - i think if we picked say 3-ish use cases and focused on their quality attribut erequirements (availabilityy, performance, consistency, scalability for example), we could make some good progress

(2:58 PM) K. Eric Harper (ABB): I agree completely.

(2:58 PM) David Boyd (Data Tactics): @Eric - Correct - that is why I offered up the simple data warehouse case. That case can then be evolved and expanded to include additional requirements.

(2:59 PM) Ron Batdorf: Business model; application reference model; performance reference model;etc. There are six layers to the EA and these layers need to be integrated for the data is also a reference model within the EA. IMO

(2:59 PM) David Boyd (Data Tactics): E.g. - Add privacy issues in the data.

(2:59 PM) Ron Batdorf: privacy and/or security..

(3:01 PM) David Boyd (Data Tactics): I think use the reflector for discussions and then meet every two weeks.

(3:02 PM) David Boyd (Data Tactics): Orit's suggestion is good also.

(3:04 PM) David Boyd (Data Tactics): Thurs work best for me.

(3:05 PM) Manoj Srivastava (CyberIQ): Thursday works

3:08 PM) Wo Chang (Host, NIST): As the scheudle goes: the group agrees to have the subgroups meeting every week but alternating between Reqs+SnP and RA+TR.

(3:09 PM) Wo Chang (Host, NIST): We also agreed these weekly meetings will take place on Thursdays starting with Jan. 30 with Requirements + SnP then follow RA+TR the next Thursday.