Office of Technology Strategies
DaaS – Hybrid Data Access Design Pattern
Public Forum
April 23, 2015

Executive Summary

Purpose: The purpose of this Public Forum was to provide an opportunity for VA and external experts and stakeholders to learn about the DaaS – Hybrid Data Access Design Pattern, and provide feedback prior to the Design Pattern's finalization, publication and implementation.

Background: The Technology Strategies Design Patterns Team has been soliciting input for the development of the DaaS – Hybrid Data Access Design Pattern. This involved collaboration efforts with a variety of stakeholders, including internal VA subject matter experts (SMEs), external government SMEs, industry vendors, and members of academia. This Public Forum represented the final stage of stakeholder engagement in the Design Pattern development process.

Overview: The TS Design Pattern Team presented the final draft of their updated Design Pattern. An introduction of the department was provided by Stephany Perez-Cohen followed by an overview presented by Nicholas Bogden of the Office of Technology Strategies. Hannah Wald of the Design Patterns Team then presented the draft increment of the Design Pattern and covered the details of its content and development. Stakeholders from various OI&T offices and organizations, both internal and external to VA, shared their questions, comments, and insights related to the content covered in the presentation. The presentation and Q&A session lasted approximately 2.25 hours.

Key Discussion Points:

- Enterprise data from the VA EA data layer/HDA and VistA will be accessible through the authoritative information services illustrated in the design pattern and the architecture diagram.
- Walt Grant asked: Is the Enterprise Messaging Infrastructure (eMI) going to carry the information coming from VistA directly to the user?
    - Hannah Wald responded: eMI carries everything within VA Enterprise Architecture (VA EA) stack. If it is a VA EA-based application, then yes. If it is a VistA application that clinicians use now then it may not be because those are not SOA-based. For anything SOA-based it's eMI.
- Nick: There is a SOA design pattern that goes into more detail on the Vista Service Assembler (VSA) platform and how VistA and the VA EA authoritative information services talk to each other.
- Damon Feldman: I see eMI to authoritative services and another eMI to VistA; what is the distinction between the two? Why not use the eMI call directly to the HDA?
    - Hannah: eMI is the general carrier for communications between components in the VA EA framework. The authoritative information services are more specific to data access (through eCRUD) and provide extra safeguards against applications and people doing things that they shouldn't do. Also, the authoritative information services handle exchanges between HDA and VSA, and provide seamless access to both for applications that pull data from both.

DRAFT

  - Hannah: However, the specifics of how authoritative data services work are outside the scope of our design pattern. We do reference them because they provide access to HDA, but that's all.
- Damon Feldman: Does the data layer consist of business entities with raw data?
  - Hannah: Yes. Having all these entities and data encapsulated into the data layer makes it more modular and reduces the chance of unintended breakage due to a change somewhere else. It also internally handles the sharing of data from authoritative data sources (ADSs) without having an impact on other things in the VA EA stack.
- Nick emphasized that when he talks about the VA's data problems in the context of the HDA design pattern, he is not laying blame; just trying to identify the problems so that they can be fixed.
  - These problems are not unique to VA. Many large government and commercial organizations have these problems too.
- The starting point of the design pattern was the MyVA presentation (from December 2014) delivered by Secretary MacDonald, in which he described the results of a survey asking Veterans about issues they have when interacting with VA.
- The root causes of the issues described in the MyVA presentation are deficiencies in data sharing and management, such as:
  - No complete inventory of VA databases: the VA Systems Inventory (VASI) and Technical Reference Model (TRM) have partial lists.
  - No central management of data.
  - No enterprise-wide quality standards or designated "official" definitive data records.
  - No standards around technology or interconnection, leading to data silos.
  - Until recently, no single VA-wide identity for Veterans and Beneficiaries that could tie individuals' records together.
- Part of the purpose of HDA is to standardize policies and processes around data management/data sharing and provide the underlying technology for making that work.
- Identity and Access Management (IAM) is the data steward for the Master Veteran Index (MVI), the ASD for identity traits associated with Veterans, Beneficiaries, and other persons of interest to VA.
  - MVI's stewards have data quality standards and procedures around the creation and maintenance of identity records.
  - MVI has helped address some VA data management problems by providing enterprise-wide Veteran identities, but it has some problems and does not easily support use as an enterprise shared service.
  - An unidentified participant noted that making MVI a true shared service would be helpful since that would allow other people to write to it (currently only the MVI data stewards can write to it).
- Bob Bishop: The problems you are describing are business problems, not really technical problems. This pattern can be the impetus for solving the business problem but the idea of a holistic view of your customer is/has been a business problem for a long time.
  - Hannah: HDA actually implements the VA Enterprise Logical Data Model (ELDM), which is designed to serve as an enterprise-wide common customer model of Veterans, Beneficiaries, etc. It takes into account how common customer information is used in VA business processes, and also includes rules about how data should be used and handled. ELDM will guide the business processes that HDA supports.

DRAFT

  - Hannah: The Enterprise Information Management (EIM) policy that was signed in February 15 is a policy driver for HDA. It requires that the VA designate new ADSs for "master records" of Veteran and Beneficiary data (i.e., which database contains the "official" record for data elements that everyone must use). HDA will effectively enable compliance with this policy by facilitating sharing of data between ADSs and their subscribers.
  - Hannah: We see this design pattern and the public forum as a way to raise awareness of this policy. It will touch everyone in VA because a system you own could be designated as an ASD for something, or alternatively you may be told "you have to integrate with system X to get these data elements."
- Bob Bishop: What you refer to as an HDA problem is a business problem and HDA is a technical solution.
  - Hannah: It's actually both a business and technical problem, and the solution has both business and technical elements.
- Ken Laskey: For the nine problems you identified that HDA is supposed to address, can you change the bullets to numbers and then in the document point back to specific problems that are solved with different aspects of the technical solution?
  - Nick: Great idea.
  - Ken: You may be able to harvest that; do a summary "these are the solutions that address these problems."
  - Nick: Fair comment, OK.
- Hannah: In the long term, HDA can be used to encapsulate the VA's data stores and make them hardware-independent and transparently accessible, which will help the VA achieve its long-term goal of transitioning to a cloud-based environment (and this can mean a private VA-owned cloud).
- Hannah: Right now the only ADS in the VA is MVI, which is the ADS for Veteran/Beneficiary identity traits.
- Bob Bishop: VHA has designated data stewards with extensively defined responsibilities.
  - Hannah: I was using MVI/Identity Services Integrated Technical Team (Ids ITT) as the ADS/data steward example here, but I know there are other data stewards in other parts of the VA (including VHA). They just aren't officially designated as such.
  - Hannah: The EIM policy signed in February mandates that the VA designate more ADSs and data stewards for them. One of the things we want to do with HDA is make it easier for ADSs to share records and for non-ADSs to use those records.
- Hannah: Having ADSs, data stewards, and other data management and governance elements in place is critical, because otherwise HDA would just be automating chaos.
- Hannah: We want to emphasize that HDA is going to use schema-on-read, which provides fault tolerance and makes it easier to adjust the system.
- Bob Bishop: It seems like there could be a lot of integrity issues with schema-on-read.
  - Hannah: Yes, there could be. This will have to be implemented carefully to preserve data integrity. But it's really the only way to go. There are just so many data sources and moving parts that HDA will need schema-on-read to handle all the things that could potentially change. If we use traditional schema-on-write in this situation things will break.
- Bob Bishop: Regarding provenance and lineage, at VHA we have the concept of data obligations, which is similar. Might want to make a note of it if you weren't previously aware.
  - Hannah: I didn't know about that. Please email us documentation.

DRAFT

- Damon Feldman: About Change Data Capture (CDC), you described it as both a design pattern and a tool.
    - Hannah: It's a design pattern and also a function that follows the design pattern. There are tools and software that do it.
- Damon Feldman: For CDC in this pattern, is there a database trigger or a CDC monitor that has tendrils in all the databases?
    - Hannah: Actually, it could be either or a combination of both. NCA's Burial Operations Support System (BOSS) has database trigger CDC. All their applications run off one database, so a change in a record that is the output of one workflow can be a trigger and input for another. This is what we want to do with CDC in HDA. We could automate things like, say, when a Veteran updates his address, we can find his dependents who live with him and update their addresses too.
- Michael Shinozaki: You mentioned processing of streaming data. I think of this like extracting information from a continuous flow of data and acting on the whole stream (think stock market or telemetry data you're monitoring for patterns). This is different from having lots of triggered actions. They're similar but the actions are different between stream and transactional.
    - Bob Bishop: The streaming data doesn't persist.
    - Michael: There's a difference between flowing data and discrete transactions.
    - Hannah: Processing streaming data would be done by Enterprise Create Read Update Delete (eCRUD), and that's kind of a long-term objective anyway. CDC would be used to trigger actions based on transactional data. Right now the transactional data is what VA needs to share and it's what we're focused on.
- Hannah: There are certain pieces of transactional data that are used throughout VA. They are relatively static and don't change frequently. They're very important because they drive all kinds of functions like outreach and mail-order medications and such. We call this high-value customer data, and the first thing we want HDA to do is share it and make sure it's consistent everywhere.
- Hannah: eCRUD will facilitate direct write-back from VA EA applications to ADSs and the data lake. CDC makes sure that changes to ADSs from legacy applications are reflected in the data lake.
- Hannah: HDA will ensure that non-ADS subscribing data stores reflect what's in the ADSs. No need to set up ADSs as shared services: just send ADS updates to the data lake and the data lake will make sure that information is available to enterprise applications and subscribing data stores to support legacy applications.
- Hannah: The HDA solution will support continued feeds to the VA Corporate Data Warehouse. We want to point out that the data stores for warehouses and security logs will be separate from the data lake. It is a best practice to keep those workloads in separate places.
- Damon Feldman: eCRUD is a design pattern, correct? And it's implemented to some extent?
    - Hannah: Yes it is, for the VHA Data Access Service (DAS). But it's not provisioned as an ESS. It needs adequate provisioning and sustainment so it can be scaled up and modified to work that way.
    - Hannah: There are actually a couple of different versions of eCRUD in operation for different SOA application stacks, but the canonical definitive one is in DAS.
- Damon Feldman: As the data lake grows, will eCRUD process data in the lake?
    - Hannah: eCRUD acts as the ingester for the data lake. We have it broken out as a separate component from the rest of the lake because that makes both eCRUD and the lake more scalable and configurable.

DRAFT

- Damon Feldman: If some transaction doesn't go through eCRUD it might not get logged?
    - Hannah: One of the principles of the design pattern is that all of the access from higher up in the VA EA stack must go through eCRUD. We can't really control what happens with legacy applications and their data stores: that has to be handled by existing solutions.
    - Hannah: Also, this illustration doesn't reflect it this way, but eCRUD is actually an all-around wrapper for the data lake, so anything that goes in or out can be logged. It's just that we couldn't manage a diagram that showed it as a wrapper and as an access gateway at the same time.
- Ken Laskey: Can a data lake be an ADS?
    - Hannah: We don't see it as an ADS. It just contains local mirrors of ADSs, which hare pre-existing data stores. But it might someday contain an ADS. You could virtualize a data source and put it there.
- Ken Laskey: So the lake is more for caching. But it does cleansing/de duplication. Wouldn't that be done before you write back to the authoritative source?
    - Hannah: This is a standard capability in any data lake. You could use that as the basis for detecting when you have an ADS conflict, as in "whoops, there are two different data store publishing this information and only one of these can be the ADS. We need to alert somebody." Also, in the future you could have some kind of ADS in there.
- Ken Laskey: You need the results of cleansing/de-duplication to write data back to non-persistent sources. The data lake will have to do the processing of incoming write before ADSs get them.
    - If we use MVI to see where a Veteran has records and create a temporary collection for a Veteran's visit to a VA facility and then we do some cleansing/de-duplication for a correct record where do we write this back some place. Want to be sure to capture that.
    - Do these conclusions need to be captured somewhere? An audit trail isn't enough.
- Ken Laskey: Regarding data persistence policies, you can't store everything forever? To what extent is it a cache and to what extent is it a data store? As a cache the data lake will be dropping data all the time when it may need to be available.'
    - Hannah: The data lake will provide some persistent data storage; isn't shown here. Cleansing and de-duplication would be used to analyze and trigger other processes. This function can be used to resolve issues of people trying to write to the same record at the same time.
- Damon Feldman: The lake could hold provisional data and determine which record should be added to ADS. Not every record is of sufficient quality.
- Hannah: This diagram focuses on using the data lake to facilitate enterprise data sharing from authoritative data sources. But the data lake could also be used to contain data stores for new SOA applications, and that data would be available for other applications. The cleansing and de-duplication facility is there to support that kind of thing.
- Bob Bishop: Didn't hear you mention archiving.
    - Hannah: I'd put in a reference but I believe the best place for that is the data warehouse.
- Bob Bishop: We archive records for 75 years in VA and need to be able to pull data out of those records. I'm not sure the data warehouse will support it. Maybe that is the responsibility of the data owners.
    - Hannah: We could add archival data stores in the data layer. We will have to look into that.

DRAFT

- Michael Shinozaki: You have to think about the origin of the record. You get the original point of entry and then stuff being pulled into data lakes. It might still be live data (e.g., patient being treated, Veteran still alive). How often does the data lake transfer to the corporate data warehouse? It can be multiple times per day.
- Michael Shinozaki: When are you done with an element of data for all purposes, who makes that call, and how do you handle that?
- Bob Bishop: For FOIA purposes, you may have to know the state of the data (active data or not, what was the state 24 months ago).
- Michael Doane: The solution can include operational capabilities to move data to lower-cost storage when it ages out. Another option is to use HGFS as one of your storage options; it's a semi-archived state. Tiered storage approach gives best of both worlds.
- Bob Bishop: Archiving could be future sub-pattern. Talk to archiving software people; it's very complex.
- Michael Shinozaki: Tools are different than policies. Does the VA information officer/data steward establish a data life cycle for a given piece of information? When can data be destroyed? Who makes life cycle decisions?
    - Hannah: This is something that data stewards would work out with VA OIS.
- The HDA Design Pattern is located on the OMBMax site. Please provide comments over the next week using the tracking matrix.

Next Steps:

The TS Design Pattern Team will finalize the Design Pattern document to address feedback from this Public Forum, and begin formal approval staffing. Updates will include additional inputs regarding architectural guidance and references to implementation guidance. The document will be edited to ensure Section 508 compliance, and the final approved version will be posted onto the ASD TS public-facing website: http://www.techstrategies.oit.va.gov/docs_design_patterns.asp

Appendices:

A. Attendee List
B. Presentation Slide Deck

Appendix A: Attendee List

Attendee	Organization
Marcus Baggs	VA – OI&T
Bill Barnes	NCA
Charles Bates	
John Blankenship	OI&T PD
Robert Bishop	VHA Office of Informatics & Analytics
Nicholas Bogden	VA ASD
John Brekke	VHA
Joe Brooks	VA TS
Lynne Case	Systems Made Simple (VHA)
Walter Coryell	OI&T TRM

DRAFT

Robert Crawford	VHA Office of Informatics & Analytics
Michael Dance	BAH
Michael Doane	MarkLogic
Arlisha Dukes	VA ICAM
Theodore Eccleston	Deloitte
Damon Feldman	MarkLogic
Dennis Fowler	Systems Made Simple (VHA)
Walter Grant	OI&T PD
Adrian Gordon	BAH
Nidhi Gupta	Deloitte
Thomas Henry	VHA Office of Informatics & Analytics
Russell Holt	OI&T Product Development
Michael Huddleston	BAH
Gordon Iglehart	VA – VBA
Sonya T. Johnson	DRC/Engility (OI&T)
Kenneth Laskey	MITRE
Liz Lake	Deloitte
Steve Mahalec	PwC
John Mandala	BAH
Jackie Matassa	BAH
Afzal Memon	Oracle
Jason Nixof	AWS
Steven Oster	OI&T-PD Health Delivery Program
John Payne	VA OSP
Stephany Perez Cohen	BAH
Dennis Peterson	VA
William Proffer	VA – SAIC
Darin S Satterthwaite	MITRE
Nimish Shah	BAH
Mary Solak	Leidos
Michael Shinozaki	Microsoft
Alan Tolk	Leidos
Curtis Trimble	VA ASD
Hannah Wald	BAH
Kevin Weinstein	BAH

Appendix B: Presentation Slide Deck

HDA_Public_Forum_
Presentation_2015_(

DRAFT