VA Enterprise Design Patterns:

3.0 Interoperability and Data Sharing Design Pattern

3.2 Hybrid Data Access Increment 2

Office of Technology Strategies (TS)
Architecture, Strategy, and Design (ASD)
Office of Information and Technology (OI&T)

Version 1.0

Date Issued: December 2015

THIS PAGE INTENTIONALLY LEFT BLANK FOR PRINTING PURPOSES

APPROVAL COORDINATION

TIMOTHY L MCGRAIL 111224

Digitally signed by TIMOTHY L MCGRAIL 111224
DN: dc=gov, dc=va, o=internal, ou=people, 0.9.2342.19200300.100.1.1=tim.mcgrail@va.gov, cn=TIMOTHY L MCGRAIL 111224
Date: 2015.12.24 10:06:07 -05'00'

Date:

Tim McGrail
Senior Program Analyst
ASD Technology Strategies

PAUL A. TIBBITS 116858

Digitally signed by PAUL A. TIBBITS 116858
DN: dc=gov, dc=va, o=internal, ou=people, 0.9.2342.19200300.100.1.1=paul.tibbits@va.gov, cn=PAUL A. TIBBITS 116858
Date: 2015.12.28 17:04:22 -05'00'

Date:

Paul A. Tibbits, M.D.
DCIO Architecture, Strategy, and Design

REVISION HISTORY

Version	Date	Organization	Notes
0.1	1/21/15	ASD TS	Initial Draft of Increment 2
0.2	2/10/15	ASD TS	Added information on the “as-is” state of VA enterprise data, based on stakeholder input and research
0.3	2/27/15	ASD TS	Added outlines for use cases and additional details in the main body of the paper, based on conversations with vendors and internal stakeholders
0.4	3/10/15	ASD TS	Developed first complete draft of use cases and fleshed out “Design Pattern Key Concepts” section
0.5	3/12/15	ASD TS	Incorporated feedback from government lead into “Design Pattern Key Concepts” section, fleshed out “As-Is State” section
0.6	3/19/15	ASD TS	Changed some terminology (including “DaaS” to “HDA”) and added new principles/constraints.
0.7	3/24/15	ASD TS	Updated Section 1 to provide explanation of hybrid data access and reference data lakes: changed content related to MVI in the “As-Is” section.
0.8	3/26/15	ASD TS	Added more terms and definitions to Appendix B, referenced authoritative sources in the Business Case, changed “dynamic schema” to “schema on read,” added “schema on read” description to Key Concepts, and began fleshing out the To-Be state section.
0.9	3/31/15	ASD TS	Made edits based on colleague feedback, added more detailed language on how HDA will facilitate sharing of ADS records, added architecture charts/graphics, and revised the “As-Is” section
0.95	5/1/15	ASD TS	Integrated feedback from stakeholders (submitted during and after the public forum) into the design pattern document.
0.96	6/16/15	ASD TS	Re-integrated use cases back into main document as requested by client.
0.97	6/22/15	ASD TS	Made revisions to address feedback by client.
0.98	7/14/15	ASD TS	Addressed additional feedback volunteered by a new stakeholder.

REVISION HISTORY APPROVALS

Version	Date	Approver	Role
0.4	3/11/15	Nicholas Bogden	Hybrid Data Access Design Pattern Lead
0.5	3/19/15	Nicholas Bogden	Hybrid Data Access Design Pattern Lead
0.6	3/23/15	Nicholas Bogden	Hybrid Data Access Design Pattern Lead
0.7	3/25/15	Nicholas Bogden	Hybrid Data Access Design Pattern Lead

Version	Date	Approver	Role
0.8	3/31/15	Nicholas Bogden	Hybrid Data Access Design Pattern Lead
0.9	4/23/15	Nicholas Bogden	Hybrid Data Access Design Pattern Lead
0.95	5/20/15	Nicholas Bogden	Hybrid Data Access Design Pattern Lead
0.96	6/16/15	Nicholas Bogden	Hybrid Data Access Design Pattern Lead
0.97	6/23/15	Nicholas Bogden	Hybrid Data Access Design Pattern Lead
0.98	7/14/15	Nicholas Bogden	Hybrid Data Access Design Pattern Lead
1.0	12/15	Tim McGrail	TS Design Pattern Final Reviewer

TABLE OF CONTENTS

FIGURES

TABLES

1 INTRODUCTION

1.1 Problem Statement

Secretary Robert A. MacDonald observed in his 2014 MyVA Presentation that "Assessments informing the [2014-2020] strategic plan told us the VA often provides a fragmented, disjointed experience that results in poor customer service and frustrated Veterans and beneficiaries." The Secretary described how Service members, Veterans, and beneficiaries take on the burden of serving as their own integration point for "multiple VAs." Veterans and beneficiaries contend with the following issues in their interactions with the Department:

- Must initiate contact with different parts of VA (through multiple call centers, hotlines, and websites) in order for VA to "know" them and see discrepancies in how different parts of VA understand them
- Fill out multiple forms to provide the same data multiple times to different parts of VA without knowing where particular forms should be sent
- Lack of consistent visibility into their own data at VA (including benefits/services history), and no single way to "write back" updates, corrections, or feedback to VA data sources
- Receive uncoordinated, duplicative, and sometimes inappropriate outreach from different parts of VA

These issues stem from fundamental, systemic problems around data management and sharing at VA (and many other large organizations), such as:

1. VA does not know how many databases are in the enterprise, where they are, or what information they contain
2. Lack of central policies and standards around data and databases has resulted in barriers to data sharing – siloed data stores cannot "talk" to each other
3. Data silos containing the same information (i.e., addresses) must be updated manually on an individual basis, which is time-consuming and creates opportunities for error
4. No enterprise-wide data quality requirements or "official" definitive data records – inconsistencies are hard to catch and almost impossible to correct
5. Scattered records about individual Veterans have not been tied to a single VA-wide identity (until recently), making it difficult to obtain a unified view of Veterans across VA

A business and technical solution that integrates and provides standardized, transparent access to VA's data stores can help address the data integration issues experienced by Veterans and VA staff who serve them.

The future capabilities described in section 3 will specifically note where the design pattern contributes to addressing these data problems.

1.2 Business Case

Hybrid Data Access (HDA) will reduce the burden of data integration and manual data management on Service members, Veterans, and beneficiaries and VA.

HDA benefits to Veterans and beneficiaries:

- Supports access to health and benefits information through a variety of Web interfaces, applications, and mobile devices using a single portal.
- Allows Veterans to update their information in a single place at a single point in time, rather than requiring them to fill out and submit the same information to multiple VA offices.
- Ensures that Veterans and beneficiaries receive coordinated services and timely, appropriate outreach from VA.

HDA Benefits to VA Employees:

- Reduced redundancy, time, and effort to update Veteran records in individual data stores.
- Facilitates enterprise-wide use of designated Authoritative Data Sources (ADSs) by serving as a standardized, scalable, and centrally managed integration point.
- Ensures enterprise-wide availability of accurate-up-to-date records, allowing VA to streamline back-end operations and provide VA employees with additional time and effort to devote to other tasks.

The VA EA data layer with HDA capability will meet VA's current data integration needs and will serve as a foundation for advanced capabilities to include processing streaming data from wearable devices, trend analysis, and automatic enrollment of separating Service members into VA benefit programs.

2 CURRENT CAPABILITIES AND LIMITATIONS

VA is a large, diverse, geographically distributed organization in which individual lines of business, regional offices, and operating units have traditionally developed and maintained their own IT systems.[1] This means a large number of systems in VA's inventory consist of databases tightly coupled with the business applications. These systems rely on customized and/or proprietary software that did not anticipate service-oriented computing or shared services. Often, they are not interconnected and operate in isolation from each other.

VA has recently launched an enterprise-wide effort to transform its IT architecture from a set of stove-piped systems to an integrated, modern, service oriented architecture (SOA) environment. This effort includes modernizing existing applications and creating a framework for standardized implementations of future applications using shared enterprise services and data using the VA EA SOA infrastructure. HDA will form the data layer of this infrastructure. The simplified VA SOA architecture diagram below focuses on the data layer and highlights some of its internal components.

[1] For purposes of this document, the word "system" means a FISMA-reportable system, unless otherwise noted.

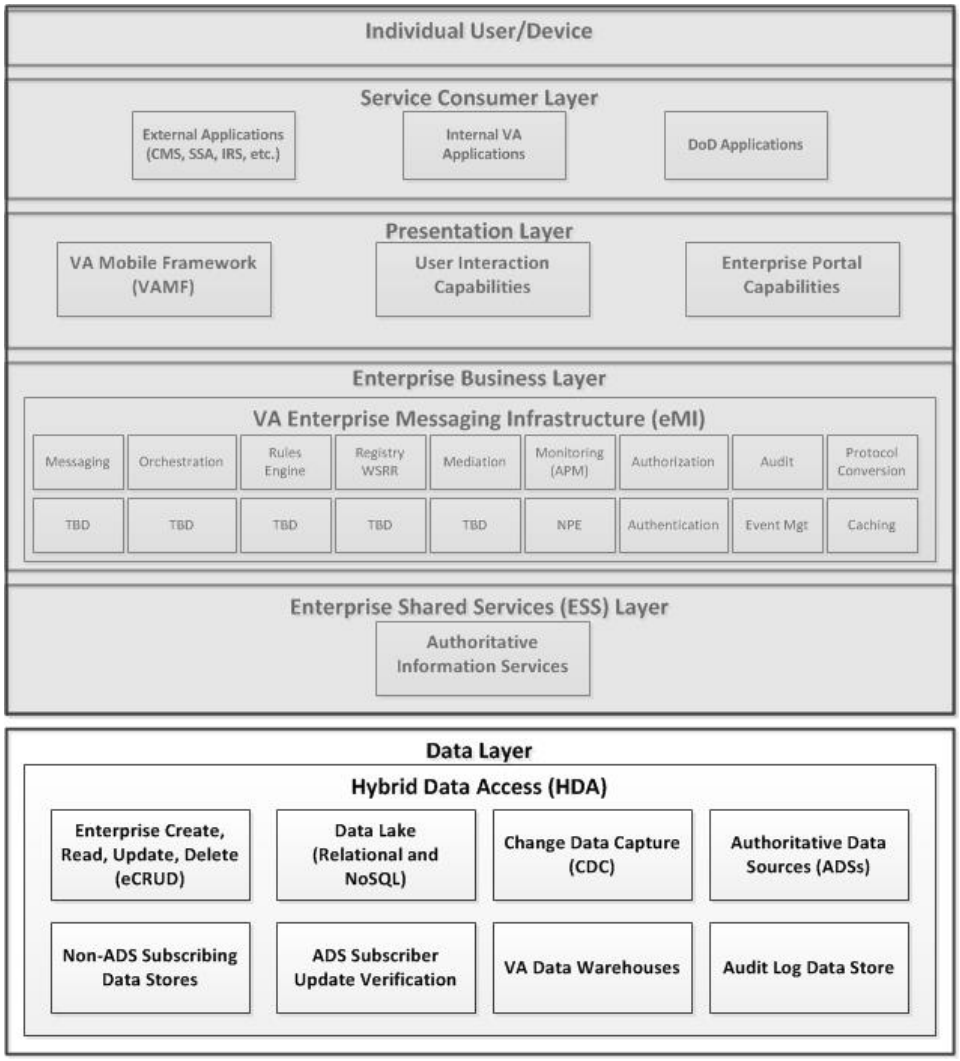

Figure 1: Illustration of HDA's Position in the VA EA SOA Infrastructure

Within the data layer, HDA will aggregate and standardize data from across VA, leveraging existing integration and standardization solutions where available. HDA will support a transparent connection point providing systems, applications, and users with on-demand access to VA's enterprise data. HDA will reduce data silos by enabling enterprise-wide data sharing and simplifying and accelerating application development by providing a standardized portal for data access.

The sections below highlight efforts to address data silos and/or management issues at VA. Each section describes specific issues and highlights where the solution effort contributes to this design pattern.

It is important to note that some of the problems described are historical cases, and some of the efforts to address them are notional or in development. Also, this does not represent a comprehensive list of all VA data integration and/or management solutions.

2.1 VistA Service Assembler (VSA) Platform

The Veterans Information Systems and Technology Architecture (VistA) is VA's largest IT system, and supports a variety of core VA functions related to healthcare, financial transactions, and benefits. Among other things, VistA is the primary data store for VHA patient records and the back-end for the Computerized Patient Record System (CPRS).

Instances of VistA are siloed based on the VA Medical Center in which they are physically hosted. The CPRS allows clinicians to retrieve patients' medical records from a different VistA instance, but those records are read-only. If clinicians want read/write access to patient records from a different VistA, they have to use a different set of applications than the ones they use on a daily basis. According to stakeholders who have received care in the VA medical system and changed VAMCs, many clinicians do not take the training necessary to use these applications.

This means that while the capability exists to move a Veteran's medical records from one VistA to another, that capability is rarely if ever used. Veterans are responsible for transferring their own records between VistAs, either as physical copies or as PDFs sent using My Health*e*Vet's "Blue Button" function. Like records from other VistAs retrieved using CPRS, the Veteran's health records must be manually entered into the target VistA to allow read/write access, which impedes clinicians' access to important health information and creates significant potential for error and information loss.

To solve this problem, VHA is transitioning its VistA instances to a SOA architecture.[2] The VSA platform, currently in development, will encapsulate VistA instances so they use a single unified interface. VA clinicians will have seamless and transparent access to data in any VistA instance regardless of physical location, through the same set of applications they use on a daily basis.

[2] For additional details on this effort, refer to the VistA Evolution SOA Design Pattern.

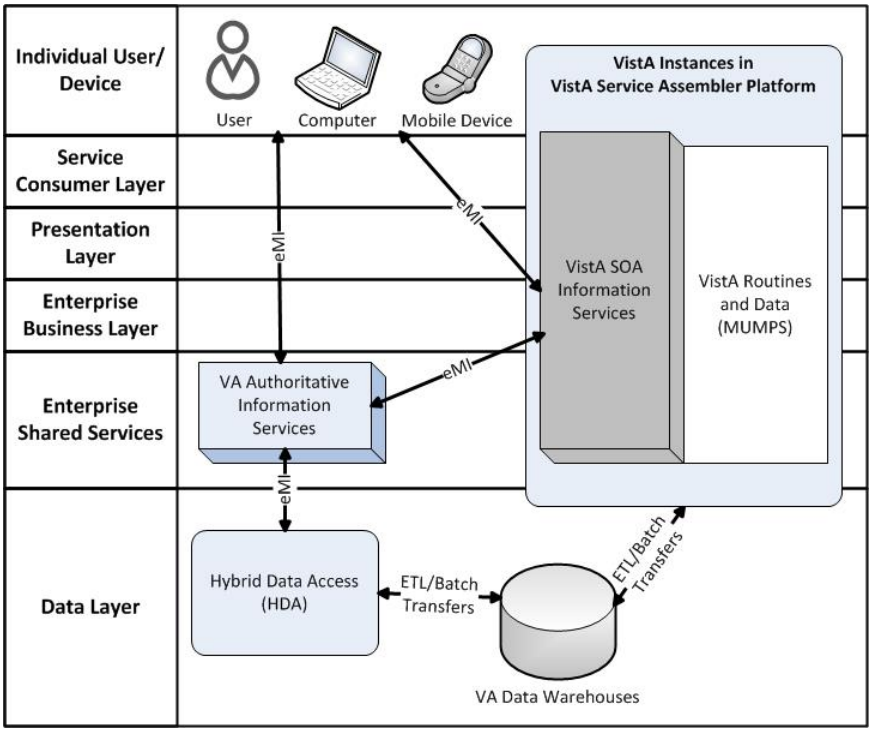

Figure 2: HDA and VistA in the VA SOA Architecture

The encapsulation of VistA instances through VSA means that VistA data stores will not be integrated into the VA EA data layer, but will be accessible through shared authoritative information services. Figure 2 illustrates the relationship between the data layer, the encapsulated VistA instances, and shared authoritative information services in the VA EA framework. Both the data layer and VistA feed directly into VA's data warehouses.

2.2 VA Data Access Service (DAS)

VistA alone cannot fulfill all of VHA's patient-related data needs. It is not flexible enough to store all the varieties of data VHA needs to support its operations, such as:

- Electronic case files from DoD
- Unstructured document data (e.g., questionnaires, scanned treatment records)
- Patient-generated data (PGD) from Veteran-facing applications

Similarly, VistA is not scalable or flexible enough to serve as the foundation of a SOA application stack and support the functionality that Veterans and beneficiaries have come to expect from Web and mobile applications.

To meet its needs, VHA implemented the Data Access Service (DAS) to serve as the back-end to its own SOA architecture. The core of DAS is a group of virtualized, scalable NoSQL datastores containing data in a variety of structured and unstructured formats.

The data stores are encapsulated into a single logical object by an Enterprise Create, Read, Update, and Delete (eCRUD) logical wrapper. eCRUD provides and mediates access through standardized, secure Representational State Transfer (REST)-based Application Programming Interfaces (APIs). It is responsible for data ingest, access control, data persistence, and logging and auditing functionality for the DAS data stores. Other VA programs are developing or have already implemented their own versions of DAS and its related components (including eCRUD).

The proposed VA EA data layer with the HDA capability draws many elements from DAS, particularly eCRUD and the MongoDB data stores (the latter can be directly incorporated into the data lake). The development and operation of DAS also provide best practices for using and accessing a NoSQL data store.

2.3 Customer Relationship Management (CRM) Unified Desktop

As a service-oriented organization, VA has a large number of call centers belonging to various operating units within different lines of business. These call centers share common business processes and resource requirements. Also, there are many scenarios in which a Veteran or Beneficiary may need to interact with multiple call centers in different lines of business to address a complex problem, such as applying for disability benefits.

Historically, VA call centers have operated in isolation from each other. This led to a variety of issues that negatively impacted internal VA operations and service delivery to Veterans:

- Redundant expenditures of cost and effort on similar data stores, IT infrastructure, and applications for individual call centers.
- No visibility into Veterans' complete customer service history, even between call centers in the same line of business – creating delays and difficulties in resolving a Veteran's issues.
- Lack of communication and information sharing between call centers, i.e., a Veteran who was routed between multiple VA customer service representatives would have to provide the same information multiple times.

To address these issues, the Veteran Relationship Management (VRM) program, developed the Customer Relationship Management (CRM) Unified Desktop, based on the Microsoft Dynamics CRM application. The CRM Unified Desktop integrates desktop applications to support capabilities including:

- A single consolidated view of the Veteran/Beneficiary for VA call center staff, ensuring all staff have accessible up-to-date information on callers seeking assistance.
- Ability to collect, preserve, and track the context of all interactions with Veterans and beneficiaries, giving VA customer service representatives immediate access to details on callers' issues and efforts made to address them.
- Access to a set of shared resources such as data and authoritative business systems, significantly reducing duplicative effort and cost.

While the Unified Desktop was developed and implemented by VBA, the solution is used by VA call centers across all Administrations and lines of business.

CRM's information flow models and business processes can be a source of best practices and lessons learned for the proposed data layer. The CRM Reference Architecture designates ADSs (beyond those already named by VA) for particular elements of Veteran and Beneficiary data, providing criteria for selecting ADSs to be used in an enterprise-wide solution.[3] Finally, the forthcoming CRM Enterprise (CRMe) release includes a data integration layer that will facilitate access to and sharing of Veteran data for customer service purposes. The HDA capability can serve in that role.

2.4 Burial Operations Support System (BOSS) Enterprise Database

NCA maintained exclusively paper-based records of burial locations and memorial inscriptions up until the early 2000s. These records were scattered across more than a hundred national cemeteries, monument sites, and regional offices. Similarly, NCA's business processes – in particular, handling memorial and burial applications and providing information on where particular individuals were buried – were entirely paper-based and manual. NCA needed a database solution to consolidate, organize, and provide easy access to all of its records for both internal and external users.

NCA's solution is the Burial Operations Support System (BOSS) Enterprise Database, a virtualized Oracle data store hosted in the Quantico Information Technology Center. Due to resource constraints, the staff responsible for developing and maintaining NCA's systems found it necessary to modify and expand NCA's single data store in order to support new applications and capabilities, rather than setting up additional databases. As a result, NCA effectively uses its BOSS Enterprise Database as a shared data service and does not suffer from the data siloing issues found in other lines of business in VA. In addition, since all of NCA's processes and applications use the same set of data, NCA has been able to implement many automatic or "pass-through" processes, in which data output from one workflow triggers and provides input to another workflow.

The BOSS Enterprise Database serves as a model of practices that can be implemented in the data layer, such as:

- Ensuring the availability and consistency of data for a geographically distributed population of internal and external users.
- Leveraging database virtualization for scalability, flexibility, performance improvements, and continuity of operations planning (COOP).
- Using a shared data store to support pass-through processing and automated functions.

[3] Refer to the Microsoft Dynamics CRM Platform Reference Architecture v5.1 for further details.

2.5 Master Veterans Index (MVI)

Historically, VA did not have standardized enterprise-wide unique identifiers for individual Veterans and beneficiaries. As a result there was no single, consistent VA enterprise identity to serve as an organizing principle for individuals' records across multiple siloed VA data stores. Under those circumstances, identity confusion and subsequent recordkeeping errors were common. It was also impossible for VA to support a secure single sign-on (SSO) capability for external users that would give them access to all VA applications.

OI&T addressed this problem by adapting VHA's Master Patient Index (once part of VistA) into an overarching Master Veterans Index (MVI). MVI acts as the mandated ADS for identity information related to over 22 million Veterans, beneficiaries, and other persons of interest to VA. Each person is assigned a globally unique MVI Integrated Control Number (ICN) that corresponds to the "Primary View" or definitive "gold copy" of their identity data. Identities in MVI form the basis of the user accounts that VA external users employ for SSO to all VA applications.

The MVI Identity Service (Ids) broadcasts identity trait updates to systems of interest to which the person identity record is correlated. VA applications are mandated to integrate with MVI and use it to locate and retrieve records associated with a Veteran or Beneficiary. VHA Data Quality Healthcare Identity Management (HC IdM) Program, MVI's data stewards, ensure the quality and correctness of MVI's data and tie legacy records to individual identities.

However, MVI's underlying system was not designed to support enterprise shared services. This means that the Identity Services Integrated Technical Team (Ids ITT) has had to build new interfaces to MVI that will accommodate a large variety of different platforms and technologies, using a set of adaptable integration design patterns. They also have to coordinate with the owners of other VA systems and applications, which can be challenging when both parties have conflicting technical, business, and security requirements.

In 2010, VA instituted a requirement for MVI integration to be addressed early in the design phase of new IT projects to address this issue in future applications and systems. The adaptable integration patterns for MVI can be leveraged for the proposed data layer, and MVI itself will continue to be a key ADS for the VA enterprise. The standards and processes used to ensure data quality for MVI can be applied to the HDA data lake. However, the technical, cultural, and business process challenges involved in getting VA system owners to integrate with MVI are also a valuable resource: these challenges highlight the potential pitfalls of implementing a shared service without adequate consideration of how to integrate it with the existing technology and business environment.

2.6 VA Data Inventory (VADI) and Data Architecture Repository (DAR)

VA does not currently have a single comprehensive inventory of its data stores, their contents, and their physical schemas. Without complete knowledge of its data assets, VA is unable to develop or implement feasible, effective mechanisms for central data management or enterprise data integration. To facilitate its enterprise data and SOA development efforts, VA OI&T's Enterprise Architecture (EA) group is

launching an effort to develop an enterprise-wide data inventory and create the first complete map of VA's data landscape.

The VA EA group is expanding the VA Enterprise Architecture Repository (VEAR) to include an authoritative VA Data Inventory (VADI). The VADI builds upon the existing VA Systems Inventory (VASI) and the VA Data Architecture Repository (DAR), the latter of which will be absorbed into the VEAR. VASI is a searchable enterprise-wide authoritative systems inventory that includes information on the databases associated with each system. Current DAR capabilities include a VA-wide metadata catalog that allows users to search, report, and manage characteristics of data and data stewardship across platforms. DAR content includes database designs for a subset of existing VA databases, such as VistA, the Administrative Data Repository (ADR), Beneficiary Identification Record Locator System (BIRLS), VBA Corporate Database (CORPDB), and Burial Operations Support System (BOSS).

The VADI will contain a complete list of VA data stores, their physical schemas, and relationships to data model concepts and data stewardship metadata. By serving as a detailed "map" of the VA data landscape, the VADI will enable understanding of the persisted data environment; facilitate the identification of authoritative data stores and reduction of unnecessary redundant data storage; promote data standardization; and facilitate enterprise data integration and centralized data management.

2.7 Customer Data Integration (CDI) and the VA Data Architecture

Certain pieces of customer information about Veterans and beneficiaries are used throughout VA, across administrations and lines of business. Some examples of common customer information include name, date of birth, SSN, address, military service history. However, while these pieces of customer information may be used across the enterprise in similar ways, they are not formatted in a uniform or consistent way – meaning that, in the current environment, commonly used customer data cannot be shared.

OI&T is taking steps to make sharing of this data possible, starting with the Customer Data Integration (CDI) initiative. CDI initiative was created to address the lack of consistent, reliable, authoritative data about Veterans and their beneficiaries. The CDI will leverage existing efforts and strategies (such as information in DAR) to provide trusted and authoritative Veteran data in cases where information is captured once and used many times. CDI will enable VA to maintain authoritative common data on each customer that is securely shared across the enterprise for proactive, seamless delivery of benefits and services with reduced burden to the Veteran, their family, and VA staff. This will allow VA to dramatically improve its delivery of benefits and services. CDI has developed a formal inventory of common customer data elements, which they sort into four categories: Identity, Military Service History, Contact Information, and Demographic and Socioeconomic Data.

☐ The VA Data Architecture Framework provides an integrated approach to describe, understand, address, and align VA data. The Framework is comprised of Conceptual, Logical and Physical Data Models. The VA Conceptual Data Model (CDM) categorizes and defines VA-wide common data concepts

and the data concepts unique to each Administration. VA's Enterprise Logical Data Model (ELDM), which documents data entities, their relationships, and attributes, extends the description of the information identified in the VA CDM v1.0. The ELDM, along with the CDM, informs and promotes VA data requirements and consistency with each Administration's data concepts. The EA Office will work with the CDI team to support the analysis of enterprise-wide standards around common VA customer data. The ELDM team is working with VA stakeholders to review ELDM content. The goal of this effort is to use the ELDM as the basis for enterprise-wide standards around common VA customer data, which will inform the design of a technical customer data sharing solution.

One of the purposes of the HDA design pattern is to describe how the ELDM will be put into operation and used as the common VA model for formatting and structuring the common high-value customer data identified by CDI. The use of schema-on-read functionality in the proposed solution will make it possible to use defined and approved ELDM entities and attributes in a production environment (with their associated business rules) as they become available. At this time, considerable work will be required before the ELDM is mature and sufficiently articulated to be used as the basis of operations in the data layer.

2.8 Enterprise Information Management (EIM) Policy

Besides MVI (the authoritative source for customer identity data), there is no official data source for Veteran and Beneficiary data or enterprise-wide quality standards applied to that data. VA needs to select and mandate the use of master records for customer data to enable data sharing and establish policies for maintaining master record quality.

To accomplish this, the VA Assistant Secretary for Information and Technology signed the Enterprise Information Management (EIM) policy in early 2015. EIM will drive efforts to establish ADSs for the entities defined in the ELDM. The policy requires VA Administrations and Offices (in coordination with OI&T) to designate additional ADSs and data stewards to maintain them. It specifies the duties and responsibilities of ADS data stewards with regards to making their data available and maintaining its quality. The policy has not yet been implemented, and a timeline for implementation has not been established.

The proposed HDA capability will work with the EIM policy to support effective enterprise data sharing within VA. The EIM policy creates the necessary preconditions for an enterprise-wide data solution, while HDA will be a mechanism for enabling compliance with EIM within the data layer. The mechanism by which HDA will publish ADS records to subscribing VA data sources is described in Section 3.2.3.

3 FUTURE CAPABILITIES THAT ADDRESS CURRENT LIMITATIONS

The HDA design pattern "to-be" state will be implemented in two phases:

- Phase One: A data integration strategy that enables sharing common, high-value customer data.

- Phase Two: A "big data" strategy that supports sophisticated service delivery and analytic capabilities, including automatic enrollment of separating Service members in VA benefit programs and near-real-time processing of streaming data from wearables and medical devices.

The capacity to reach Phase One and subsequently evolve to Phase Two will depend on the data lake's scalability and capability of storing unstructured data and supporting "schema-on-read" functionality. A full description of the required capabilities for the data lake is provided in Section 3.2.2.

Aspects of the data layer and HDA, described in Sections 3.1 and 3.2, are each tied to the relevant data problem(s) enumerated on the list in Section 1.1.

3.1 Pre-Implementation

In order to ensure that the proposed HDA capability and the data layer provide sufficient return on investment, the following issues must be addressed prior to deploying the solution into production:

- Develop a comprehensive inventory of VA data stores, their contents, and their schema
- Implement the EIM policy across the enterprise
- Develop enterprise-wide records management strategies for:
    - Enterprise data retention/persistence
    - Creation, management, and retention of audit logs
- Establish common enterprise metadata, schema, and related tools to support data access/consumption by EA enterprise services

3.1.1 Map the VA Data Landscape

A full map of the data landscape must be built to support the enterprise data solution. VA requires an inventory of data stores, who owns them, where they are located, what database management system (DBMS) they are running, and what kinds of data are stored on them. VASI contains a partial list but a complete record of VA data stores is needed.[4]

HDA and the data layer will leverage the VADI, which will initially focus on data stores containing customer data so that these stores can be subscribed to ADSs. This does not mean that every data store in the inventory will need to be integrated. If a system is slated for transition/decommissioning or uses a DBMS not approved by the TRM, it may be practical to decommission that system.

The VADI will identify data stores, who owns them, where they are located, what database management system (DBMS) they are running, and what kinds of data are stored on them. The VADI will also clearly indicate encapsulated (i.e., federated or virtualized) groups of databases.

VA Data Problems Addressed: #1

[4] A list of DBMSes currently in use at VA, originally compiled by the Technical Reference Model (TRM) team, can be found in Appendix F.

3.1.2 Enterprise Data Management and Governance

VA needs formalized governance processes and modeling concepts for enterprise shared data to shape the operations of the data layer. Otherwise, integration in the data layer would accomplish nothing more than automating chaos. Fortunately, some of the governance and modeling elements necessary to make the data layer function are already in development (if not yet implemented) at VA.

One such element is the EIM policy signed in 2015, which mandates that VA formally designate ADSs that will host the master records for common customer data elements (as mapped and defined by the ELDM). Having a full inventory of data elements will guide the Data Governance Council (DGC) in identifying, evaluating, and selecting potential ADS candidates. Once ADSs have been designated, VA must coordinate with data stewards to establish enterprise business requirements and rules for common customer data.

The Enterprise Architecture Council (EAC), in partnership with DGC, must further develop the ELDM before it can be used as the basis for operations in the data layer. Entity attributes, business rules, and associations must be clearly defined. In addition, data obligations – i.e., rules around CRUD for particular types of data elements – must be articulated so they can be operationalized and enforced in the data layer. Data obligations are particularly important for data that is provided by and/or shared with partner organizations. VHA's Business Information Architecture (BIA) program has a large collection of tools, repeatable processes, and best practices that can potentially be leveraged to perform the work of further developing the ELDM and its related business rules and obligations. The DGC should establish the priorities for the evolution of the ELDM to meet operational requirements.

VA Data Problems Addressed: #2, #4

3.1.3 Records Management

A records management strategy is essential for governing and managing an enterprise data solution. Common enterprise customer data (defined by the ELDM and published by ADSs to subscribers) will persist in the data lake. Other types of data can be retained provided there is a valid business case. Any type of data that is not explicitly targeted for persistence in the data lake will be overwritten.

A records management strategy for the creation, use, and retention of logs is required. Lessons learned from DAS demonstrate that without such a strategy, the rapidly growing volume of log files will become unmanageable and prohibitively expensive to store. A practical and effective strategy answers the following questions:

- Why is the organization conducting auditing (e.g., performance monitoring, security)?
- What are the requirements for audit information?
- What utility will the organization realize from its logs, i.e., what will be done with them and what benefit will it provide?
- When does a log qualify as an official record or personally identifiable information (PII) or protected health information (PHI) in and of itself? In such cases, what are the legal requirements for retaining/protecting the log?

- What are the forensic requirements for security logs? What are the requirements for logs to serve as evidence in a court of law?
- If a security breach occurs, how far back will the security logs need to go to allow organizational security personnel to discover the root cause of the breach?

The records management strategy will address archiving for both operational data and log data. The trigger for moving data from "hot" on-demand storage to archival storage (e.g., time-based vs. event-based) will differ depending upon factors including the particular type of data and how it is used.

VA Data Problems Addressed: #2

3.1.4 Develop Schema, Metadata, and Transformation Tools

When ADSs, enterprise business rules, and auditing policies have been established, VA can start developing common enterprise schema for customer data based on approved entities and definitions from the ELDM. Because the data lake will use a schema-on-read approach, it will be feasible to use one or more "starter schema" and make incremental modifications as needed with little or no disruption.

VA needs to develop metadata consistent with its business rules for enterprise data. The metadata will provide the capability to attach provenance and lineage information to each data record and/or message, and additional information needed to support logging, auditing, and performance monitoring.

VA will need to configure the transformation capabilities of the data layer. HDA will require the capacity to transform data from any format used in the VA enterprise to any other, including a common enterprise data format consumable by all VA SOA applications.

All of the above elements – the schema, metadata, data lake, and transformation capabilities – can be prototyped and refined before production in a secure, easily configurable cloud environment using historical and anonymized warehouse data.

VA Data Problems Addressed: #2, #3

3.2 Phase One

The results of the pre-work described in Section 3.1 will inform the configuration and implementation of the data lake and its associated components. Phase One will accomplish the following goals:

- Implement the EIM policy by publishing customer records from designated ADSs to designated subscribers in the VA enterprise, thus providing all VA users (internal and external) with access to current, correct records about Veterans and beneficiaries.
- Provide, through eCRUD, a single common access point for enterprise data to be used by VA enterprise shared services, SOA applications, partner organizations, and authorized third-party applications.
- Implement the necessary capabilities for authorized internal and external users to easily view, update, and provide feedback on information relating to a particular Veteran or Beneficiary.

- Support the continued operation of existing VA data stores (including data warehouses) and their associated applications and systems.

The three key components of the data layer will provide these capabilities:

- **eCRUD**
    - Ingest for the data lake
    - Provision of unified, standardized, secure read and write-back access to the VA data layer (specifically the data lake and ADSs) through RESTful APIs
    - Auditing and logging for operations in the data layer
- **The Data Lake**
    - Aggregation and persistence of enterprise data in multiple formats
    - Publication of master records from ADSs to subscribing data stores
    - Harmonization of data with established common enterprise schema based on ELDM
    - Provision of data to consuming VA enterprise services
- **Change Data Capture**
    - Detection and propagation of updates to master records in ADSs
    - Triggering of automated workflows based on creation of or updates to records in VA data stores

Figure 3 below illustrates the components of the data layer and the flow of information between them, including ADSs, non-ADS data stores, VA data warehouses, and archival data stores.

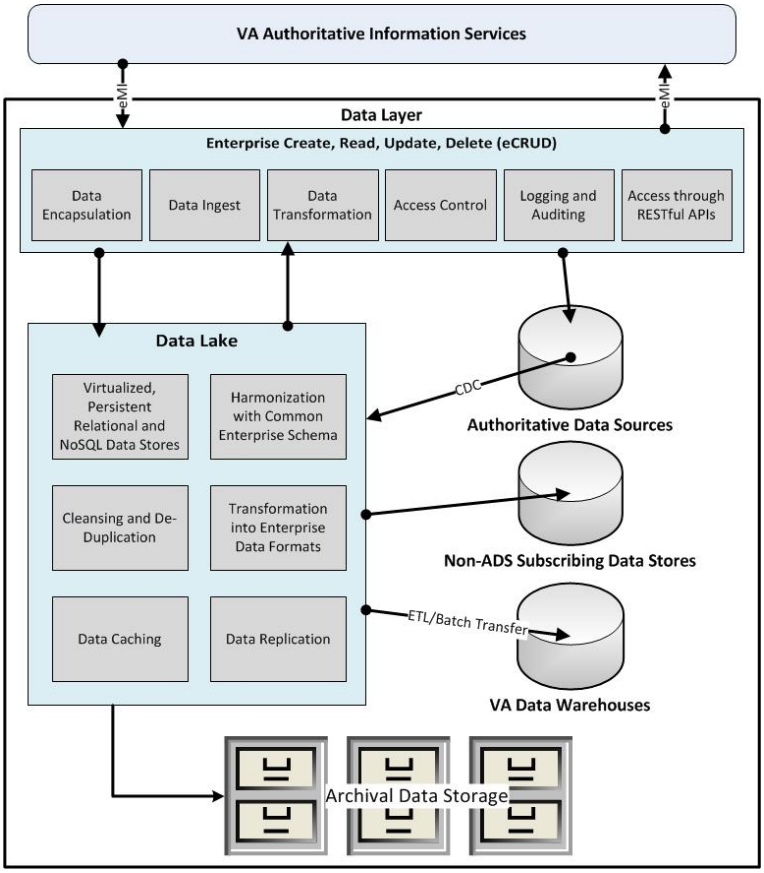

Figure 3: Components of the VA EA Data Layer

3.2.1 eCRUD

Although there are multiple versions of eCRUD used in various SOA application stacks within VA, the canonical, definitive version is associated with the DAS data store owned by VHA. Currently, eCRUD is used on a relatively small scale and is specifically intended for use with a MongoDB data store, but it can potentially be scaled up, extended, and otherwise modified to serve as the provider and mediator of access to an enterprise data lake.

eCRUD will perform the same key functions as it does for DAS, such as data ingest, logging/auditing, enforcing access control policies, and queuing data messages. By exposing the data lake and ADS data stores as standardized, RESTful APIs, eCRUD will facilitate read and write-back access to the data layer for VA information services, VA partner organizations, and Veteran-owned third-party applications. It will perform data cleansing and de-duplication on write operations from these sources in order to maintain data integrity for the data lake and ADSs. eCRUD will also handle writes that are “internal” to the data layer, i.e., from ADSs to the data lake. The collection of adaptable interface patterns created by Ids ITT for MVI will be used as a model for interfaces between eCRUD and ADSs.

VA Data Problems Addressed: #3, #5

3.2.2 The Data Lake

To provide sufficient scalability, maximize availability, and ensure continuity of operations, each instance of the data lake will be virtualized on clusters of co-located machines. There will be at least two instances, each in different geographic locations. The data lake will have scalable and extensible storage capacity to accommodate any structured or unstructured data from authorized sources. It will contain local caches of frequently used data from preexisting VA data stores, particularly ADSs. The data lake will support persistent virtualization of additional databases, particularly virtualized instances of databases from systems that have been retired/decommissioned. It will have sufficient capacity to process high volumes of incoming and outgoing data messages at high speed.

The data lake will serve as the (logical) halfway point between data stores in the enterprise. As with eCRUD, the collection of adaptable interface patterns created by Ids ITT for MVI will be used as a model for data lake interfaces to other VA data stores. Where there are existing database integration solutions (such as virtualized or federated databases), the data lake will by default be connected to their designated interfaces, rather than to the individual data stores that compose them. Through these connections, the data lake will replicate published ADS updates to subscribing data stores and will be evaluated for possible batch transfers to VA data warehouses.[5]

In addition, the data lake will read newly ingested and loaded "raw" input using transformation tools to harmonize it, if possible, with the common enterprise schema (singular or plural). The harmonized data will be made available for consumption by enterprise SOA services.[6] This data can also be transformed and replicated to subscribing data stores. The data lake will include cleansing and de-duplication capabilities (or at a minimum, the capacity to add those capabilities) to support large-scale processing, merging of data sets, and quality assurance for any data stores that are hosted exclusively in the data lake.

While the data lake will perform many functions in the VA enterprise data solution, it will not be used as a data warehouse (i.e., to support analytics). It will instead replicate data (raw, transformed, or both, as desired) to existing data warehouses for analysis. Likewise, audit logs for the data layer as a whole will be stored separately from the data lake.

VA Data Problems Addressed: #2, #3

3.2.3 CDC for ADS

Changes to definitive ADS records by SOA applications will be routed through eCRUD. Legacy applications that are directly tied to existing data stores will not use eCRUD (they cannot be modified to use an enterprise SOA infrastructure), but will still be supported by the data layer. If a legacy application (as opposed to a SOA application) writes to an ADS, the resulting change will be detected by a CDC tool,

[5] This will require coordination with the owners of those warehouses.

[6] Any data cleansing and de-duplication conducted within the VA's enterprise data solution must be consistent with applicable data policies (e.g., those related to the legal Electronic Health Record).

which will trigger replication of the update to a virtualized "mirror" of the ADS in the data lake.[7] Changes in the ADS mirror will trigger additional replication to subscribing systems, including transformation of the updated data into the format used by those systems.

Subscribers to ADS updates will have the capability to verify and validate they received ADS updates sent to them from the data lake, and also to indicate whether or not those updates were successful. If the data lake pushes an update to a subscriber and does not receive an acknowledgement of the update and/or the acknowledgement indicates the update was not written correctly, the data lake will attempt to push the update again. Repeated failures will trigger an alert to the owners of both the data lake and the subscribing system at issue.

Using CDC tools to route updates from ADSs to the data lake will introduce some delay in publishing those updates, but will make it possible for ADS owners to offer their data as a shared enterprise resource without taking on the burden of turning their own systems into shared services, as IdS ITT had to do with MVI. ADS owners can simply ensure they have a working interface to one data store – the data lake – instead of establishing and maintaining interfaces to dozens of subscribers across VA. Similarly, the owners of subscribing systems can connect to the data lake and rely on it as their source for ADS updates, rather than having to integrate directly with the ADSs themselves.

VA Data Problems Addressed: #3, #4, #5

3.3 Phase 2

3.3.1 Expand Schema to Support New Applications

The use of schema-on-read will allow adding new data elements to the common enterprise schema as needed with minimal time, effort, or disruption. However, adequate change management processes must still be applied to schema updates. Testing of proposed updates prior to their implementation in production systems is strongly recommended.

The stewards for the data lake will coordinate with the ELDM team to ensure the data layer's common enterprise schema are consistent with the most recent release of the ELDM. When a new version of the ELDM is released, corresponding updates to the common enterprise schema will be made within a reasonable amount of time. If the data lake stewards perceive demand for changes to the common enterprise schema, they will work with the ELDM team to incorporate those changes. The common enterprise schema in the data lake will always be "synchronized" with the ELDM.

3.3.2 Support Additional Types of Data

While the initial purpose of the data layer will be to facilitate sharing of and access to common customer data, it can (and should) eventually support other types of data as well. This will require expanded data

[7] The same security policies enforced in an ADS or other data store (particularly with regards to access control) must be enforced for its mirror and/or caches in the data lake.

transformation capabilities and new capabilities within the data lake itself (e.g., a terminology service for clinical content mapping and harmonization within the data lake).

3.3.3 Develop Sophisticated Analytics Capabilities

As the number of VA data producers and consumers connected to the data layer increases, eCRUD will be scaled up to meet demand. It will need to be modified and extended to support new, advanced capabilities, such as processing streaming data from medical devices and performing near-real-time analytics.

3.3.4 Consolidate Data into the Lake

In the very long term, VA will use the data lake and/or the underlying virtual platform to consolidate and virtualize its (non-VistA) operational databases. In the nearer term, the data lake can serve as a long-term archive for static data from legacy databases still in operation so that they can continue to support the non-SOA applications that are connected to them, with a minimal amount of dedicated infrastructure.

Ultimately, legacy data stores (or at least the datasets they contain) will be transitioned entirely into the data lake itself, and any new operational VA data stores will be created there (unless they have technical, security, or other requirements that cannot be met by the data lake). This means that the data lake can potentially became an ADS by hosting a set of persistent "master record" data.

Stewards of data stores that have been transitioned to or created in the data lake will have the access rights and capabilities necessary to perform their duties, particularly maintaining the quality of their data and supporting their business processes. They will be able to modify their individual data store as required, provided the effects of their modifications do not impact the functioning of the data lake and/or data layer as a whole. Access control and other security policies will continue to be enforced, just as they were in the original database. In that respect, the data lake will be similar to NCA's BOSS enterprise database: the underlying DBMS, platform, and infrastructure are managed by OI&T, but users in NCA offices and cemeteries have full, transparent access to the data itself.

Consolidating VA's data stores into a virtual environment will facilitate central management and governance while giving system and application owners the flexibility needed to generate and use data for business processes. It is also compatible with VA's long-term strategic goal of transitioning to a cloud-based technology environment.

3.4 Enterprise HDA Constraining Principles and Strategic Guidance

The following table summarizes the constraining principles associated with HDA/data layer concepts and architectural attributes. These principles will be used to drive implementation guidance for the HDA capability, the data layer, and any components (services, applications, and data stores) that will integrate with it.

Table 1: HDA Constraining Principles and Strategic Guidance

#	Data Layer Concept/Attribute	Principles & Strategic Guidance
1	Data Access	Enterprise data will be accessible through non-proprietary protocols and communication methods.
2	Data Access	The data layer will support the use of dynamic schema (also called "schema-on-read") to support extensibility and fault tolerance.
3	Data Access	Proposed modifications to common enterprise schema will be tested before being implemented in a production environment.
4	Data Aggregation	The data layer will provide functions such as data aggregation, data de-duplication/rationalization, and data synchronization through the use of the data lake, eCRUD, and CDC tools.
5	Data Aggregation	All common enterprise schema and business processes used within the data layer will correspond with OI&T's Enterprise Logical Data Model (ELDM).
6	Data Aggregation	The underlying platform of the data lake within the data layer will support virtualization and continued use of databases from systems being transitioned/divested.
7	Data Aggregation	Development and sustainment efforts associated with the data layer and its components will encourage and facilitate the consolidation and standardization of both new and existing data stores.
8	Data Aggregation	The data lake will not be used as a data warehouse for historical analytics purposes, nor will it contain security/audit logs or archives: it is solely intended for use as an operational/transactional data store.
9	Data Security	Upon the creation of records or collection of records from joins, federation, or aggregation, the security and access levels will raise to the level of the data element with the highest security classification.
10	Data Security	Implemented data services and enterprise data stores will provide full, end-to-end FIPS 140-2-compliant encryption modules to ensure that data is encrypted both in transit and at rest.
11	Data Security	Data layer and data lake security, including database security and exception handling, will comply with VA 6500 Handbook, NIST and DISA STIG security policies.
12	Data Security	At the level of patient generated data (PGD) encryption (i.e., PII/PHI), data at rest and in transit will comply with HIPAA controls in addition to VA 6500 Handbook, NIST, and DISA STIG security policies.
13	Data Security	Within the data layer, the data lake will only accept data messages from, or replicate to, predetermined and explicitly authorized (i.e., whitelisted) systems and software.
14	Scalability	The data layer and data lake will provide a high level of availability for service consumers and facilitate outward scalability.

4 USE CASES

The three (3) use case scenarios in this section involve user access to VA enterprise data in the proposed VA EA data layer. These use cases all take place in the context of a larger scenario in which a Veteran moves from Washington, DC to San Francisco, CA, which leads to not only a change of address but a change from one VA Veterans Integrated Service Network (VISN) to another. The first use case scenario describes the initial change of address for the Veteran, which creates the conditions for the two subsequent use cases.

4.1 Use Case 1: Veteran Initiates a Change of Address

4.1.1 Purpose

In the "as-is" environment of VA, a Veteran cannot change his or her address in one interaction and have the change propagate across all the databases belonging to all the VA organizational units that provide him or her with services and benefits. Instead, the Veteran must notify multiple VA organizational units of his or her change of address. This is burdensome for the Veteran and creates significant opportunity for errors (e.g., misheard street names, mistyped characters during data entry).

This use case provides an example of how the proposed capabilities of the VA EA data layer will allow a Veteran (or a VA staffer acting on the Veteran's behalf) to update his or her address once, and have the update automatically propagated across VA without any further action on the Veteran's part.

4.1.2 Assumptions

Assumptions for this use case include:

- Regardless of the Web service or application employed by the user, the same core set of services are used for the address change.
- User Authentication and Authorization processes have already been completed successfully, allowing the Veteran to update his/her contact information or a VA staffer to update the information on his/her behalf.
- BIRLS is the ADS for locating beneficiaries of a particular Veteran.

4.1.3 Use Case Description

Basic Flow of Events

1. The Veteran, or a VA Staffer acting on the Veteran's behalf, initiates a change of address through a Web application to reflect the Veteran's move from the Washington, DC area to the San Francisco, CA area.
2. A data message is sent from the application through authoritative information services down to the data layer. This data message contains the following elements:
    a. The Identity and Access Management (IAM) security ID of the user responsible for the address change (Veteran or VA Staffer)
    b. Veteran's MVI ICN
    c. Veteran's previous address

    d. When the new address record was created, and whether it was through an internal or external user application
3. The message is processed by the eCRUD service, which writes the address change to the data lake and also to the ADS for Veteran addresses.
    a. Because the address update was performed by a SOA application and is already in a format consistent with a common enterprise schema, the data lake does not have to harmonize the data.
    b. The records in both data stores will now reflect the Veteran's new address.
4. The change of address in the data lake is detected by CDC capabilities, triggering multiple preprogrammed actions (described below).
5. The change of address is replicated to target data stores that subscribe to updates of master records in the ADS for Veteran addresses.
    a. The data lake waits for a specified amount of time to receive acknowledgement of the update from the subscribing data stores.
    b. All but one of the subscribing data stores sends an acknowledgement to the data lake indicating that the address update was received and written successfully.
    c. The data lake sends a second update message to the target data store that failed to acknowledge the first message, and waits for a reply.
    d. The data store at issue does not acknowledge the second update message, so the data lake sends an alert to the owners of the subscribing data store and the data lake stewards informing them of the issue.
6. The data lake sends instructions to automated outreach systems to send follow-up physical and/or electronic email to the Veteran that:
    a. Gives the Veteran information on the VA Health and Benefits centers closest to his/her new address (along with links to their websites).
    b. Informs the Veteran that his/her dependents (listed individually by name) will have their contact information changed to match the Veteran's.
    c. Supplies contact information for the appropriate VA offices the Veteran can reach out to if he/she has any issues associated with his/her move to a new address.
7. The data lake ensures that the Veteran's change of address is extended to the Veteran's dependents by taking the following actions:
    a. Queries BIRLS for the MVI ICNs of dependents associated with the Veteran (e.g., spouses, children).
    b. Upon receiving requested ICNs from BIRLS, queries its own database for records that:
        i. Have the ICNs returned by BIRLS.
        ii. Have a physical address and home phone number matching the Veteran's previous address and phone number.
    c. Updates the contact information of the Veteran's dependents in its own data stores and sends the update to the appropriate ADS and ADS subscribers. Again, provenance and lineage metadata is attached to indicate the nature, time, and cause of the change(s).

i. As in Step 5, the data lake waits for acknowledgement of the update from the subscribers. This time it also expects an acknowledgement from the ADS. It receives acknowledgements from all subscribers except for the one that was experiencing problems before.

d. Uses authoritative information services to notify VistA[8] that the Veteran's address has changed and he/she and his/her dependents should be aligned to a new VAMC on the basis of his/her new ZIP code.

Results

The Veteran's VA benefits and records travel seamlessly with him/her and his/her dependents. The Veteran receives mail and phone calls from VA at his/her new address and phone number in San Francisco, and no communications are mistakenly sent to his/her previous address. The Veteran's complete records are easily accessible to VA clinicians and benefits counselors in the VISN where the Veteran now resides.

4.1.4 Use Case Context Diagram

Figure 4 illustrates the flow of information through the components of the VA EA architecture and the data layer, as described above.

[8] Refer to the *Enterprise Messaging Infrastructure (eMI) Design Pattern* for details on the interaction between services in the service layer. See the *VistA Evolution SOA Design Pattern* for specifics on data access within VistA systems.

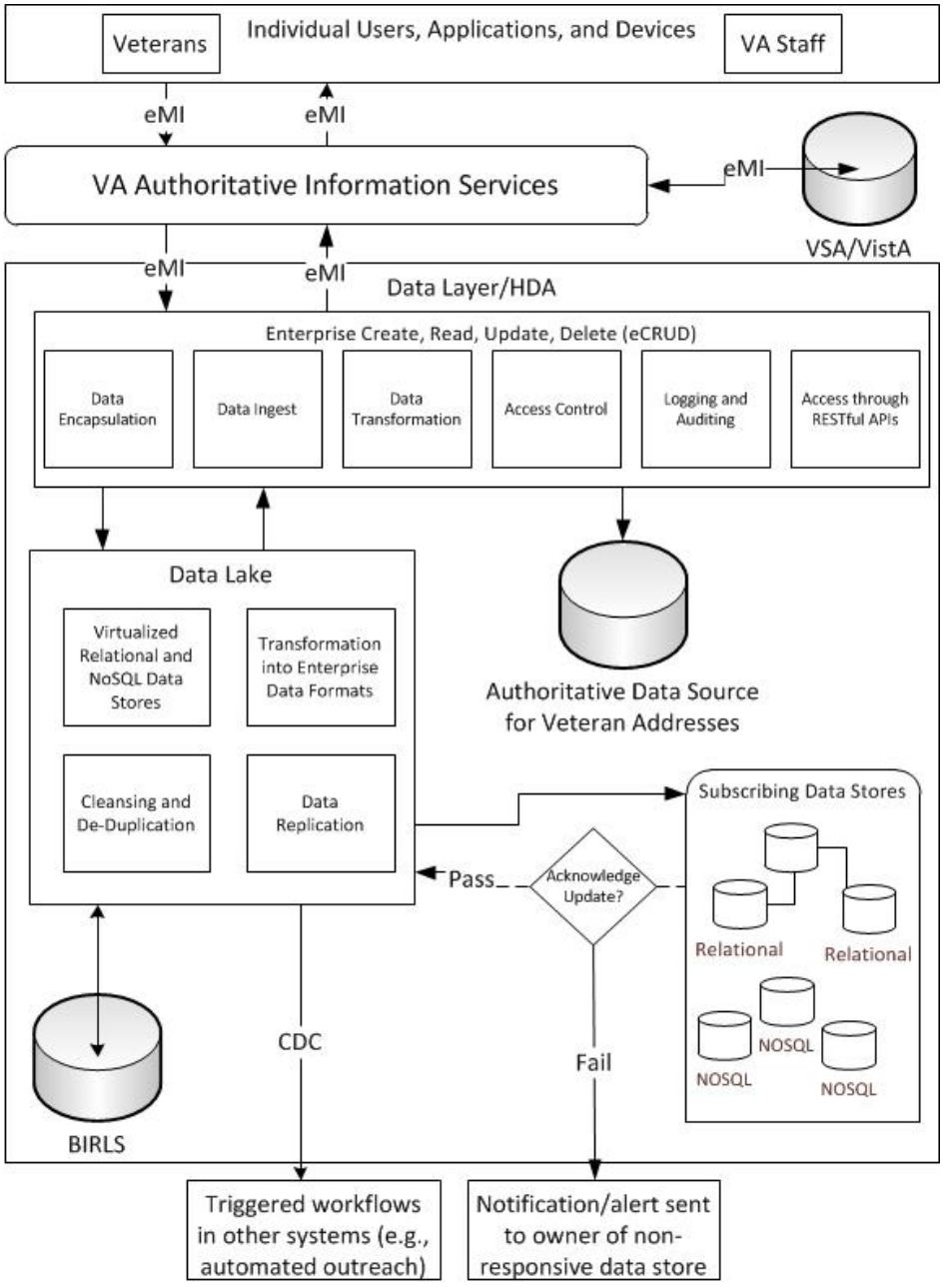

Figure 4: HDA Use Case #1 - Veteran Initiates a Change of Address

4.2 Use Case 2: Clinician Accesses Veteran's Health Records from Another Region

4.2.1 Purpose

In the "to-be" state of the VA IT environment, instances of VistA and their associated data will be encapsulated using VSA, as described in Section 2.1.[9] VSA and the VA EA data layer will be accessible (to

[9] For additional details, refer to the *VistA Evolution SOA Design Pattern*.

applications, services, and each other) through shared authoritative information services. This use case describes a clinician using an application that, through those services, accesses data from both VistA and the data layer, providing contextual information that helps ensure continuity of care when a Veteran moves between different VA Medical Centers.

4.2.2 Assumptions

Assumptions for this use case include:

- The Veteran has moved his or her residence from one state to another state – this can be a permanent move or a seasonal "snowbird" move.
- The clinician accessing the Veteran's medical records is doing so because he/she will be meeting the Veteran for the first time during a routine checkup or consultation appointment.
- The clinician accessing the Veteran's medical records has successfully completed the user Authentication and Authorization process for the Web interface or application that provides access Veteran's records.

4.2.3 Use Case Description

Basic Flow of Events

1. Clinician opens an application that provides access to his/her appointment schedule. One of the listed appointments is flagged with an indicator showing that the patient has not been to Clinician's VAMC before, based on the Veteran's change of address (in the first use case) and resulting alignment from a VA Medical Center in the Washington, DC area to the current VA Medical Center in the San Francisco, CA area.
2. Clinician selects the appointment in order to access a "Person View" of information pertaining to the new patient, such as name, age, branch and dates of service, etc. The Person View contains links or buttons that Clinician can use to access medical information about the Veteran.
3. Clinician selects option to view prescription information on the patient. Application calls on authoritative information services to access VistA for prescription records associated with the Veteran's VHA PatientID. Clinician reviews Veteran's current prescriptions.
4. Clinician returns to the Person View and selects an option to view Veteran's notes entered into MyHealtheVet. Application calls on services with access to HDA, which retrieves Veteran's PGD stored in the best available instance of the data lake for Clinician to review.
5. By accessing both VistA and non-VistA data through the application, Clinician is able to get an overall picture of Veteran's current condition, medical history, and health-related concerns.

Results

By accessing both VistA and non-VistA data through the application, Clinician is able to get an overall picture of Veteran's current condition, medical history, and health-related concerns, ensuring that the Clinician can provide continuity of care.

4.2.4 Use Case Context Diagram

Figure 5 illustrates how the events of Use Case #1 affect what the Clinician sees in his/her application, and how that application accesses data from both the VA EA data layer and VistA in the VSA platform.

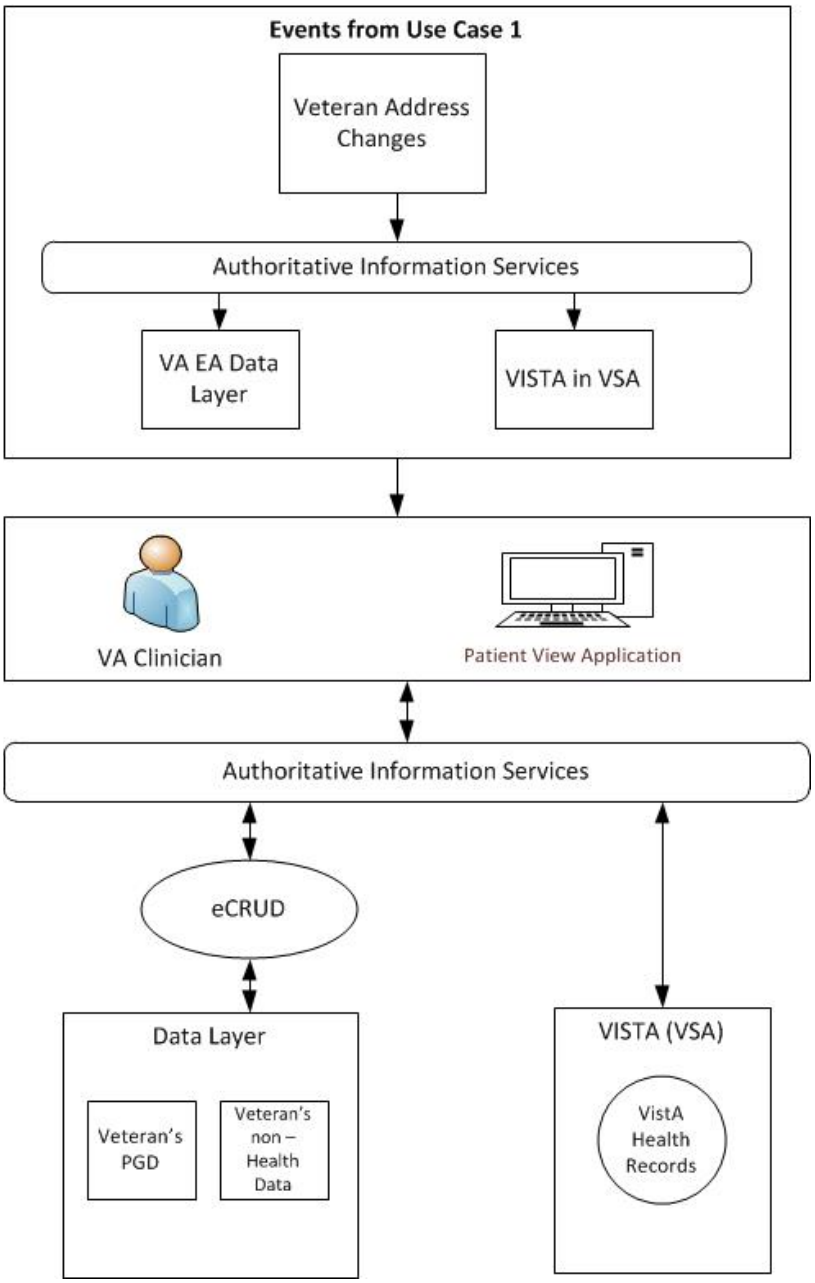

Figure 5: HDA Use Case #2 - Clinician Accesses Veteran's Health Records from Another Region

4.3 Use Case 3: VA Staffer Checks Veteran's Status Prior to Outreach

4.3.1 Purpose

Uncoordinated or inappropriate outreach is a persistent problem created by VA data silos. Some examples of outreach issues caused by incorrect or outdated information regarding a particular Veteran include:

- Inability to relay critical information in a timely fashion.
- Sending sensitive documents to the wrong address.

- Attempting to make contact with a Veteran after he or she has died.

The proposed capabilities for the VA EA data layer include a mechanism to propagate changes to the master records in ADSs across all other connected VA databases, ensuring all instances of the record within the data stores are consistent and correct. This use case describes how automatic change propagation and related triggers will allow VA staff to provide better coordinated outreach and services to Veterans.

4.3.2 Assumptions

Assumptions for this use case include:

- The Veteran has confirmed a change of address (as described in the first use case), updating his/her record across VA and setting off pre-programmed triggers and alerts.
- The VA staffer reviewing the Veteran's contact information has successfully completed the user Authentication and Authorization process for the VRM application he/she is using.

4.3.3 Use Case Description

Basic Flow of Events

1. VA staffer selects a mailing list associated with mass e-mails for a VA program in the Washington, DC metropolitan area.
2. VA staffer initiates a "POC Check" function that will query various tables in the data lake (including those with the official records for Veteran/Beneficiary contact information) to find entries on the mailing list that meet certain preprogrammed conditions.
3. Based on the results of the query, the VRM application displays a subset of individuals on the mailing list who have been "flagged" because they meet the preprogrammed conditions specified for the POC Check function.
4. VA staffer reviews details on why each individual is flagged. He/she sees that some individuals on the mailing list are flagged because their physical addresses on file are not in the Washington, DC metropolitan area.
5. VA staffer unsubscribes "out of range" individuals from the e-mail list.

Results

Without having to unsubscribe from, or ask to be removed from, a mailing list for Veterans in the Washington, DC metropolitan area, the Veteran who changed his/her address in the first use case will no longer receive mail from this list, which is no longer relevant to him/her.

4.3.4 Use Case Context Diagram

Figure 6 illustrates the sequence of events as the VA staffer checks and acts on the status of "flagged" mailing list recipients in an outreach application.

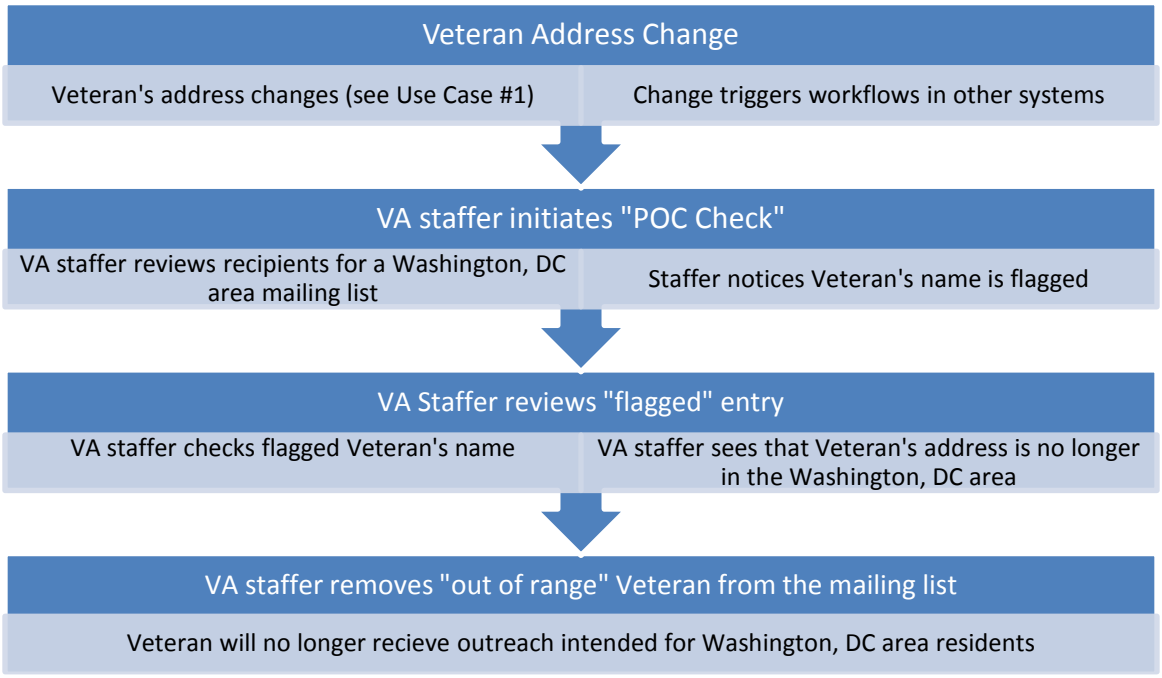

Figure 6: HDA Use Case #3 - VA Staffer Checks Veteran's Status Prior to Outreach

Appendix A. BACKGROUND INFORMATION

This appendix contains the following high-level information for the Hybrid Data Access design pattern:

- Document scope
- Intended audience
- How to use this document
- Document development and Maintenance

A.1 Scope

This increment of the HDA Design Pattern document will expand on the concepts and constraints described in *VA Enterprise Design Patterns: Data-as-a-Service (DaaS)* by describing how they will be applied in a VA enterprise data solution. Specifically, this document addresses:

- Interactions between middleware and different types of databases (e.g., Relational, NoSQL).
- Aggregation and processing of information retrieved from multiple data sources of different types into a single message consumable by services/applications connected to HDA.
- Ensuring data can be accessed and/or edited quickly (availability) while maintaining its correctness and integrity (consistency) across data sources.
- Providing access to VistA records through the VistA Service Assembler (VSA), the future replacement for the current access service (Medical Domain Web Services or MDWS).[10]

The following concepts are outside the scope of this document:

- Ensuring data messaging security and authenticity.
- Access control and authorization decisions.
- Specifics of applications/services used to access HDA.
- Infrastructure and hardware design specifications.
- Specifics of services and applications that will support HDA.
- Vendor-specific technologies.

The use cases associated with this design pattern can be found in the companion document *Hybrid Data Access Increment 2 Use Cases.*

A.2 Purpose

The purpose of this document is to provide strategic direction and enterprise-level capability guidance that identifies best practices for developing an enterprise hybrid data access (HDA) capability that provides seamless, unified read and write-back access to VA's data stores within the VA Enterprise Architecture (EA) data layer.

[10] For a detailed description of VSA, refer to the VistA Evolution SOA Design Pattern.

HDA focuses on sharing and providing access to high-value customer information about individual Veterans and beneficiaries. This type of information is used across VA Administrations and lines of business, but not in a consistent, enterprise-wide format. Data elements of high-value customer information include, but are not limited to:

- First name
- Last name
- Date of birth
- Military service history
- Address[11]
- Contact information

Most of the data at issue is relatively static in nature – meaning that once a record of it is created, that record is likely to remain unchanged for weeks, months, or years at a time. What is critical and currently problematic for VA is making sure that when these changes *do* occur, the changes are reflected in all the records used by all the Administrations and lines of business in the enterprise.

Integrating and standardizing this high-value customer information is the key to addressing the systemic data access and management problems identified in the previous section. Maintaining a single access point for data across VA enables the Department to provide greater accessibility and transparency for Service members, Veterans, beneficiaries, and partner organizations. Integrating VA's data stores will also reduce redundancy in both organizational data and processes.

A.3 Intended Audience

All programs will use this document to guide efforts toward using Enterprise Shared Services (ESS) for standardized, enterprise-wide access to enterprise data. This will help programs meet data sharing requirements utilizing enterprise data stores while:

- Developing new VA applications
- Modifying existing production systems
- Acquiring and integrating Commercial Off the Shelf (COTS) (including open-source) applications

A.4 How to Use This Document

This document can be used to guide and inform the following related activities:

- Development, acquisition, and/or implementation of products and services with the capabilities required for VA enterprise data integration and analytics.
- Creation and use of data management and governance policies, standards, and mechanisms.

[11] A single individual may have multiple types of addresses, for example, a permanent home address, a temporary address, and/or a prescription address. This document uses the term "address" in the sense of one's permanent residence/primary address.

- Defining, articulating, and enacting requirements and business processes around enterprise data use and management.
- Long-term development of sophisticated enterprise analytics capabilities.

A.5 Document Development and Maintenance

This document was developed collaboratively with internal stakeholders from across VA with participation from the Office of Information and Technology (OI&T), Product Development (PD), Office of Information Security (OIS), Architecture, Strategy and Design (ASD), and Service Delivery and Engineering (SDE). In addition, the development effort included engagements with industry experts to review, provide input, and comment on the proposed pattern. This document contains a revision history and revision approval logs to track all changes. Updates will be coordinated with the Government lead for this document, who will facilitate stakeholder coordination and subsequent re-approval depending on the significance of the change.

Appendix B. Design Pattern Key Concepts

This appendix provides high-level overviews of key concepts associated with HDA (and the data layer) that apply to solving recurring problems within the current state of the VA IT environment, such as:

- Lack of a unified, consistent view into records for a Veteran or Beneficiary
- Necessity for Veterans and beneficiaries to provide the same information (e.g., address) multiple times to different parts of VA
- Manual, piecemeal write/update processes for siloed data stores, resulting in long delays and significant potential for error
- Duplicative or inappropriate outreach based on incorrect or outdated information

The concepts outlined below provide the context for the data layer "to-be" architectural attributes described in Section 3. These attributes will establish the design constraints to be applied to solution architectures developed in VA.

Related database concepts (with definitions) that are not included in this section can be found in Appendix C: Definitions.

B.1 Availability and Consistency

Availability is concerned with the accessibility of data to authorized consumers (e.g., applications, services). This includes performance considerations including the amount of time required for a query/response to complete successfully.

Consistency is concerned with the quality, accuracy, and "up-to-date-ness" of data. Achieving consistency throughout an enterprise with numerous distributed data stores is particularly challenging. This requires synchronizing records in data stores across the enterprise with authoritative sources as frequently and rapidly as possible.

Data consistency and data availability can be mutually exclusive under some circumstances, because ensuring consistency reduces the speed at which data can be delivered. An effective enterprise data solution is balanced between these two properties consistent with the enterprise's mission.

B.2 Data Management and Governance

B.2.1 Data Stewards

ADSs are sources of data or information designated and officially recognized as trusted, timely, secure and used within VA's information environment in support of VA business processes. The DGC designates these sources and the Office of Information and Technology (OI&T) develops and maintains technology solutions (e.g. services) that use these sources.

A data steward is a designated role responsible for maintaining the quality of data in an ADS, and is typically an organizational unit that is synonymous with or designated by the business owner of the ADS.

Data stewards develop and implement rules and processes around how data in ADSs should be entered, changed, audited, or otherwise handled in order to ensure its quality.

B.2.2 Data Models and Schema

The enterprise data architecture (which also applies to metadata) documents VA data requirements and establishes a source for their data entity definitions, attribute definitions, and characteristics. A high-level conceptual data model describes general groupings of data in terms of entities[12] and the relationships between those entities. It informs a logical data model which describes data elements, their attributes, and their associated rules. The logical data model forms the basis for a common schema or blueprint for the structure of data elements used by enterprise services.

Traditional relational databases use a "schema-on-write" approach to input. Any input to the database must be in a predetermined format recognized by the database before it can actually be written. If the structure of the input is not what the database "expects" – i.e., organized using the database schema – the write operation will fail and errors will result. NoSQL databases can employ a "schema-on-read" approach, accepting any input regardless of its structure (or lack thereof), transform any elements of the input they recognize, and ignore everything else.

B.2.3 Provenance and Lineage

An important aspect of data management and governance is using metadata to track the origin and history – or provenance and lineage – of records and changes to those records. All records should be "bundled" with metadata indicating their provenance and lineage.

Tracing the provenance of records and updates is essential to preserving data quality in the context of an enterprise data service. Provenance supports data integrity by ensuring only changes made by authorized users, processes, applications, and/or systems are propagated through the enterprise data service and committed to ADSs and local data stores. It should be detailed enough to trace the complete "chain of custody" of a record back to its point of origin.

Data lineage is the history of a data record. It tracks deltas (i.e., changes) between the current state of the record and the previous version or versions, with timestamps to indicate when changes were made. Lineage is essential for rollback functionality and for queuing changes to data elements in a record to ensure the data are committed in the correct order.

B.3 Technical Components of an Enterprise Data Solution

In addition to the data stores themselves, the proposed HDA capability will include three key components: a change data capture (CDC), a data lake, and a data ingester.

[12] The Unified Modeling Language (UML) is the prevailing language for developing health information models.

B.3.1 Change Data Capture (CDC) Tool

Tools or applications with CDC functionality detect changes to a data store or record and initiate preprogrammed actions as a result of that change. In the proposed HDA capability, CDC is the preferred mechanism for maintaining consistency between publishers (in this case, ADSs) and subscribers throughout VA. A CDC-capable tool will "listen" for ADS updates (i.e., new records or changes to existing records) and push those updates to one or more selected targets. The CDC tool can be used to initiate other actions based on changes to data, including sending alerts or automatically triggering workflows.

B.3.2 Data Lakes

Data lakes[13] are storage repositories that hold vast amounts of data in native formats until needed. A data lake typically contains multiple structured and unstructured data stores. However, these databases are encapsulated so as to present a transparent logical interface to the services and applications that interact with them. Instances of data lakes are hosted in clusters of co-located servers. Duplicate instances can be distributed geographically to facilitate local data access and ensure continuity of operations.

Data in the data lake can be cleansed (i.e., standardized, de-duplicated, and corrected) and indexed to support an enterprise-wide search capability. Some or all of the data can be harmonized by transforming it from its original format into a format consistent with one or more common enterprise schema. Harmonized data can be served to consuming enterprise services and applications, replicated to a data warehouse for analytics, or both.

Data lakes can serve as the target store for virtualized databases from decommissioned systems, allowing those systems to be taken offline without disrupting the applications and services that depend on them. This capability facilitates long-term transitions from a traditional IT environment to a virtualized and/or cloud-based environment.

The data lake should *not* contain a data warehouse or its own security audit logs. It is an industry best practice for enterprise data solutions to use separate, dedicated data stores for workloads associated with operations, analytics, and logging, since they have very different (and sometimes conflicting) security, performance, and resource requirements.

B.3.3 Data Ingestion and Data Ingesters

Data ingestion is a function (or series of related functions) that serves as the intake mechanism for a data lake. The data ingestion function handles large volumes of data messages from various sources across an enterprise, queues them to ensure they are committed in the correct order, and performs preprocessing before writing the data into the data lake's storage.

In some enterprise data solutions, data ingestion is a capability of the data lake itself, while others decouple data ingestion from the data lake into a "data ingester" component with its own dedicated

[13] Equivalent terms include data hubs, data reservoirs, landing zones, or operational data warehouses.

cluster of hardware. HDA employs a data ingester, partly because it can support write-back from SOA applications to ADSs (not just the data lake). A data ingester can also be scaled to accommodate growing data message volume without impacting the performance of the data lake. Data ingesters can be extended to support advanced capabilities including processing streaming data from wearables and medical devices.

A data ingester can also perform data transformation to some degree: it can transform some data messages en route to the data lake while letting others pass through to be transformed after they are loaded into storage. The proportion of total transformation work performed by the ingester versus the data lake itself depends upon the characteristics of the IT environment and the requirements or business processes of the enterprise.

Appendix C. DEFINITIONS

This appendix provides definitions for terms used in this document, particularly those related to databases, database management, and data integration.

Atomicity, Consistency, Isolation, Durability (ACID): Set of properties that guarantee that database transactions are processed reliably. Some DBMSes provide better guarantees of ACID properties than others.

Authoritative Data Source (ADS): A source of data or information designated and recognized as official that is trusted, timely, secure and used within VA's information environment in support of VA business processes. Administrations and staff offices designate these sources within domains for which they are the stewards (i.e., owners/responsible parties). OI&T develops and maintains technology solutions (e.g. services) that use these sources.

Change Data Capture (CDC): The capability to initiate certain preprogrammed actions based on changes to data stores and/or data records, e.g., "When a new record is created in Database A, create an associated record in Database B."

Conceptual data model: A data model that is presented at a high level of abstraction, hiding the underlying details, and making it easier for people to comprehend.

Consistency: In the context of data and data management, consistency is concerned with the quality, accuracy, and "up-to-date-ness" of data. Achieving consistency throughout an enterprise with many distributed data stores can be particularly challenging. It requires synchronizing records in data stores across the enterprise with authoritative sources as frequently and rapidly as possible.

Data ingester: A component in an enterprise data solution that serves as the intake pipeline for a data lake. It takes in large volumes of data messages from various sources across an enterprise, queues them to ensure that they are committed in the correct order, and performs some level of preprocessing before writing them into the data lake's storage. The data ingester may perform data transformation on some incoming data messages, while others are transformed in the data lake itself. The proportion of total transformation work performed by the ingester versus the data lake itself depends upon the characteristics of the IT environment and the requirements or business processes of the enterprise.

Data lake: A type of data store that is a critical component of enterprise data integration and/or "big data" solutions. It serves as the aggregation, integration, and distribution center for data from sources across the enterprise and makes it available to consuming services and applications. Also called a data sink, data reservoir, or landing zone.

Data model: See *Conceptual data model* and *Logical data model*.

Data store: Often used interchangeably with “database,” but also encompasses types of data repositories that are not strictly databases, such as data warehouses and data lakes.

Data warehouse: An archival data store where enterprise data is replicated and processed for analytics and the creation of test data. Unlike a database, a data warehouse does not hold “live” production data.

Database federation: Federation is an established method of encapsulating independent databases into a single logical entity with a common interface. Queries coming through the interface are “broken” into subqueries, which are sent to the appropriate databases in the federation. Responses to the queries are reassembled into a single message for whatever application/service initiated the original query.

Database virtualization: The practice of hosting databases in a virtual environment. Like database federation, this can be used to encapsulate multiple databases into a single logical entity. Virtualization provides other advantages, such as faster query/response performance, platform independence, centralized management/governance, and analytics capabilities.

Enterprise Create, Read, Update, Delete (eCRUD): The eCRUD service was initially created as part of the DAS project. eCRUD provides an interface that allows enterprise services to perform create, read, update or delete (CRUD) operations on data in the VA SOA data access layer/HDA solution. It also supports numerous adapters for data transformation, notification of data changes, and custom event handlers.

Enterprise Messaging Infrastructure (eMI): Communications infrastructure that supports transactions in the VA SOA environment.

Integrated Control Number (ICN): The primary key and unique identifier associated with individuals in MVI.

Lineage: Information (usually metadata) associated with a data element that provides a record of changes made to that element, both in terms of what the changes were and when they were committed. Lineage information supports rollback functionality, queuing, and correct ordering of data updates in an enterprise data solution.

Logical data model: An entity-relationship data model including data attributes that represents the inherent properties of the data, including names, definitions, structure, and integrity rules, independent of software, hardware, volumetrics, frequency of use, or performance considerations.

Master Veterans Index (MVI): VA ADS for the identity information of approximately 17 million Veterans and beneficiaries. Identity traits are cross-referenced with DoD records.

Not Only SQL (NoSQL): Type of DBMS that structures data in a non-tabular/non-relational format. NoSQL database types include key-value, column-family, document, and network. Some NoSQL databases can also store unstructured data.

Object-oriented database: Type of DBMS that organizes data into objects (in the sense of object-oriented programming). These databases are designed to integrate with object-oriented programming languages. For example, the InterSystems Caché DBMS (used for VHA's VistA systems) integrates with the MUMPS programming language, storing objects written in MUMPS so that they can be reused or reassembled. Caché structures data in hierarchical multidimensional arrays, and does not use schema.

Provenance: Information (usually metadata) indicating the origin of a data element/record or changes to that element/record. Supports the capability to establish, record, and trace a clear "chain of custody" for a piece of data.

Relational database: Type of DBMS in which data is organized in a tabular fashion – i.e., in tables, one for each type of entity described in the database. Each row in a table is an individual instance of a type of entity, and each column is an attribute of the entity. For example, an Oracle database at VA may contain a table in which each row represents an individual Veteran, and the columns contain attributes about the Veteran such as date of birth, gender, branch of service, etc.

Schema: The logical "blueprint" for how data is structured in a relational database. Schema determines the format of data in the data store and of data messages passed between systems, services, and applications within an enterprise. Most contemporary DBMSes support the use of multiple schema, making it possible to organize and view a single dataset in different ways for different purposes.

Structured Query Language (SQL): Common language for accessing and manipulating data stored in relational databases.

Appendix D. ACRONYMS

Table 2: Acronyms

Acronym	Definition
ACID	Atomicity, Consistency, Isolation, Durability
ADR	Administrative Data Repository
ADS	Authoritative Data Source
API	Application Programming Interface
BIRLS	Beneficiary Identification Record Locator System
BOSS	Burial Operations Support System
CDC	Change Data Capture
CDM	Conceptual Data Model
CPRS	VHA Computerized Patient Record System
COOP	Continuity of Operations Plan
CorpDB	VBA Corporate Database
CRUD	Create, Read, Update, Delete
DAR	VA Data Architecture Repository
DAS	VA Data Access Service
DBMS	Database Management System
eCRUD	Enterprise Create, Read, Update, Delete
EIM	Enterprise Information Management
ELDM	Enterprise Logical Data Model
eMI	Enterprise Messaging Infrastructure
ESB	Enterprise Service Bus
ESS	Enterprise Shared Services
HC IdM	VHA Data Quality Healthcare Identity Management
HDA	Hybrid Data Access
IAM	Identity and Access Management
ICN	Integration Control Number
Ids ITT	VA Identity Services Integrated Technical Team
MVI	Master Veterans Index
NoSQL	Not Only SQL (database)
PGD	Patient-Generated Data
PHI	Protected Health Information
PII	Personally Identifiable Information
REST	Representational State Transfer
SDD	System Design Document
SQL	Structured Query Language
UML	Unified Modeling Language
VEAR	VA Enterprise Architecture Repository
VA EA	VA Enterprise Architecture
VADI	VA Data Inventory
VISN	Veterans Integrated Service Network

Acronym	Definition
VistA	Veterans Information Systems and Technology Architecture

Appendix E. References, Standards, and Policies

This Enterprise Design Pattern is aligned to the following VA OI&T references and standards applicable to all new applications being developed in the VA, and are aligned to the VA ETA:

#	Issuing Agency	Applicable Reference/Standard	Purpose
1	VA OIS	VA 6500 Handbook http://www1.va.gov/vapubs/	• Defines the overall security framework for VA including data storage, retrieval, and exchange.
2	VA ASD	VistA Evolution SOA Design Pattern – COTS Application and Non-COTS Applications http://www.techstrategies.oit.va.gov/docs/designpatterns/OnePager_SOA%20VISTA%20inc%201_v4_09122014.pdf	• Provides references to the use of enterprise CRUD services and persistent data stores as part of the integration with SOA support infrastructure services. These documents are intended to standardize and constrain the solution architecture of all healthcare applications in the VA. • HDA represents the "data layer" of the VA SOA design pattern
3	VA	SOA Layer Implementation Guide Section 4.1	• Provides implementation guidance for application developers with regard to a virtual data access layer, focusing on the Java EE platform but conceptually is applicable to all platforms using data services.
4	VA ASD	Enterprise Application Architecture (EAA) Sections 2 and 5 http://vaww.ea.oit.va.gov/enterprise-architecture/enterpr	• Section 5 provides technical underpinnings for cross-cutting development concerns for VA applications including data architecture. • Section 2 prescribes the architecture principles for all applications that form the basis of enterprise design patterns. Specifically, Principle 2 regarding the separation of presentation, business, and data

#	Issuing Agency	Applicable Reference/Standard	Purpose
		ise-technical-architecture/	logic applies to the to-be vision of virtual data access and aggregation via Data-as-a-Service
5	VA ASD	SOA Technical Framework (SOA-TF) Section 7 http://vaww.ea.oit.va.gov/enterprise-architecture/enterprise-technical-architecture/	• Provides technical underpinnings for data logic concerns for VA applications. • Includes the concepts of data access for direct access to a data store owned by an application, and data interchange for performing CRUD operations on a data store owned by an application. Data interchange requires the use of the ESB to route messages and perform data validation.
6	VA ASD	ESS Strategy Document and Directive http://vaww.ea.oit.va.gov/enterprise-shared-services-service-oriented-architecture/	• Provides the overarching strategy for developing, deploying, and managing ESS throughout the VA
7	VA ASD	OneVA Enterprise Technology Strategic Plan (ETSP) http://www.techstrategies.oit.va.gov/docs/itroadmap/OneVA_ETSP_508C.PDF	• The HDA design pattern will help programs develop applications in alignment with IT Vision Attributes 6, 8, 9, 10, 12 (scale-out of enterprise data stores)
8	VA	Data Access Service (DAS) System Design Document (SDD)	• Describes how the VA DAS will be constructed and translates identified business needs, business requirements, functional requirements, and non-functional requirements into a document from which developers can create the system
9	VA	VA Directive 6518: Enterprise	• Establishes official policy for the implementation of Authoritative Data Sources (ADSs) in VA

#	Issuing Agency	Applicable Reference/Standard	Purpose
		Information Management http://www.va.gov/vapubs/viewPublication.asp?Pub_ID=791&FType=2	• HDA provides access to and synchronization with enterprise ADSs in compliance with this directive
10	VA	VA Memorandum: VA Identity Management Policy (VAIQ 7011145) https://www.voa.va.gov/DocumentView.aspx?DocumentID=3073	• Establishes the Master Veterans Index (MVI) as the authoritative source for identity traits of Veterans and all other persons of interest to the VA • Mandates a unique identifier for all Veterans and requires that all VA applications integrate with MVI
11	VA	Data Access Service (DAS) eCRUD Service Interface Control Document	• Describes the functions, capabilities, and requirements of the eCRUD service
12	VA	Microsoft Dynamics Customer Relationship Management (CRM) Platform Reference Architecture v5.1	• Describes principles, requirements, processes, and best practices used by the VA CRM Platform, an enterprise shared service for VA call centers.
13	VA	VA Directive 6051: Department of Veterans Affairs (VA) Enterprise Architecture (EA) – http://www.va.gov/	• Prescribes Department-wide policy for the establishment and implementation of an integrated One-VA EA to be used for the development and management of all information assets. • Establishes operating principles for the VA EA and assigns specific responsibilities for its development

#	Issuing Agency	Applicable Reference/Standard	Purpose
		vapubs/viewPublication.asp?Pub_ID=3&FType=2	and implementation.

Appendix F. Database Management Systems Used at VA

The list of DBMSes in the table below was derived from assessments conducted by the ASD team responsible for development and maintenance of the VA TRM. The table also indicates in the "TRM Status" column whether a DBMS is currently approved, unapproved, or slated for divestiture/retirement under the TRM. Unapproved DBMSes are those that cannot be reliably secured, and DBMSes slated for divestiture are either legacy technologies or "desktop databases" (e.g., Microsoft Access) that cannot be adequately managed for security, quality, or business continuity.

To maintain consistency with the TRM and VA policy, the proposed HDA solution will only connect to databases with a DBMS that is listed as "Approved" or "Approved with constraints."

#	DBMS	DBMS Type	TRM Status
1	Microsoft Access	Relational desktop	Divest
2	Redis	Non-relational in-memory	Unapproved
3	FIS-GTM	Non-relational	Approved with Constraints
4	Integrated Data Store II	Non-relational	Unapproved
5	MarkLogic Server	Non-relational	Unapproved
6	MongoDB	Non-relational	Approved with Constraints
7	Cache	Non-relational object-oriented	Approved with Constraints
8	PostgreSQL	Object-oriented	Divest
9	DB2	Relational	Approved with Constraints
10	Filemaker Pro	Relational desktop	Divest
11	InterBase	Relational	Approved with Constraints
12	Java DB	Relational embedded	Approved with Constraints
13	Model 204 (M204)	Non-relational	Approved with Constraints
14	Microsoft SQL Server	Relational	Approved with Constraints
15	MySQL	Relational	Approved with Constraints
16	Oracle DB	Relational	Approved with Constraints
17	SAP Sybase SQL Anywhere	Relational	Approved with Constraints
18	CA IDMS	Non-relational	Approved with Constraints
19	SQLite	Relational	Approved with Constraints
20	Apache Cassandra	Non-relational	Unapproved