VA Enterprise Design Patterns
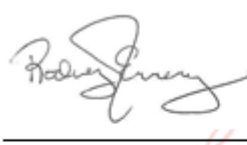**Interoperability and Data Sharing**

Enterprise Data Analytics

OFFICE OF TECHNOLOGY STRATEGIES (TS)

OFFICE OF INFORMATION AND TECHNOLOGY (OI&T)

VERSION 1.0

DATE ISSUED: FEBRUARY 2016

APPROVAL COORDINATION

Rodney Emery
Director, Technology Strategies and GEAC, ASD

PAUL A.
TIBBITS
116858

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

REVISION HISTORY

Version	Date	Approver	Notes
1.0	February 16, 2016	Nicholas Bogden	Incorporated stakeholder feedback from during and after public forum and added diagrams to use cases

CONTENTS

1 INTRODUCTION

The Department of Veterans Affairs' (VA) patchwork of analytics capabilities does not adequately meet the data needs of its changing operational environment or its burgeoning enterprise-level programs and services. VA's emerging data needs are driven by two trends:

- Implementation of new technologies that generate novel types of data (e.g., NoSQL data) which are not compatible with traditional SQL-based analytics solutions.
- MyVA and VA Enterprise Architecture (EA) initiatives are driving a sea change in favor of shared resources, enterprise-level programs, and more centralized management.

1.1 Business Need

VA's traditional approach to analytics is based on the data needs of individual Administrations, offices, projects, and lines of business (LOB) using structured query language (SQL)-based database management systems (DBMS). Most of VA's analytics capabilities and warehouses are artifacts of that approach. Both the approach and environment are at odds with MyVA's emphasis on shared services, standardization, and holistic enterprise management. As a result, VA contends with three interrelated critical gaps in the area of analytics:

1. VA cannot meet its emerging data needs without radically changing its approach to analytics.
2. Absence of enterprise analytics governance creates organizational barriers to data collection and sharing between Administrations, LOBs, and programs.
3. VA's collection of "little data" capabilities are not plugged into the existing enterprise-level "big data" capability (the VA Information and Analytic Ecosystem).

1.2 Approach

The goal of this Enterprise Design Pattern (EDP) is to establish the architectural principles, guidelines, and constraints regarding a VA enterprise "big data" capability that will meet this demand and maximize the value of VA's business intelligence. The enhanced analytics capabilities proposed in this EDP support more efficient and effective service delivery to Veterans and beneficiaries. Specifically, this capability will provide:

- Continued support for existing analytics activities
- Consolidation of analytics data into logically separate Public, VA Internal, and Private warehouses within the VA Information and Analytic Ecosystem
- Balance between standardization and flexibility
    - Standardized, repeatable processes for data collection and transfer
    - Clearly defined enterprise rules for data ownership/responsibility

o   Configurable, automated sharing with different audiences
- A shared service/resource for advanced analytics capabilities (e.g., predictive analytics, context sensing, machine learning)
- Integration with the VA Enterprise Architecture (EA) data layer including Enterprise Create, Read, Update, Delete (eCRUD) service and the operational data lake[1]

This capability will also support VA organizational goals and missions that rely on cross-cutting data collection and sharing, to include:

- **Customer Data Integration (CDI) Initiative:** Harmonize and sequence the exchange and use of digital information within the department computing environment, and between the VA and its mission partners in the delivery of benefits to Veterans.
- **Veteran Relationship Management (VRM) Program:** Build and publish complete Veteran interaction histories.
- **National Center for Veteran Analysis and Statistics (NCVAS):** Provide Department-wide statistics and reports to Congress, other Federal agencies, and the public.
- **Identity and Access Management (IAM) Program:** Correlate and harmonize Veteran and beneficiary data to create consistent authoritative identity records.
- **Veterans Benefits Administration (VBA)**: Collect data from VHA to support cost analysis and service delivery improvement.
- **Veterans Health Administration (VHA):** Share clinical data and conduct research with partner organizations.

2   CURRENT CAPABILITIES AND LIMITATIONS

2.1 Capabilities

Many of VA's analytics shops (including warehouses, tools, and expertise) reside in silos owned by individual Administrations, LOBs, offices, projects, and programs. In 2012, the VA Office of Information and Technology (OI&T) Business Intelligence Service Line (BISL) deployed an analytics infrastructure called the VA Information and Analytic Ecosystem. BISL operates and manages the Ecosystem in partnership with its largest customer, VHA.[2] Figure 1 below illustrates the high-level operational concept of the Analytic Ecosystem, including the Corporate Data Warehouse (CDW) and Regional Data Warehouses (RDW).

---

[1] As described in the *Interoperability and Data Sharing 3.3: Hybrid Data Access (HDA) Enterprise Design Pattern*.
[2] VHA's data needs were in fact the primary driver for development of the Analytic Ecosystem, which is why they use it intensively and are deeply involved in its governance and management.

Each warehouse has an attached "report farm" where analytics reports are published. The CDW, which contains a copy of all the data in the RDWs, provides access to BISL's catalog of analytics applications through an "Analytics App Store."

A variety of master enclaves are coupled to the Ecosystem. Each master enclave contains a copy of the CDW. BISL has a repeatable, customizable approach to integrating existing systems as enclaves or deploying entirely new ones. The master CDW analytic enclaves are:

- General Purpose (GP)
- Business Intelligence (BI)
- Analytics and Informatics (AN)
- Health Services R&D (RD)/ Veterans Informatics and Computing Infrastructure (VINCI)
- Field Reporting (FR)
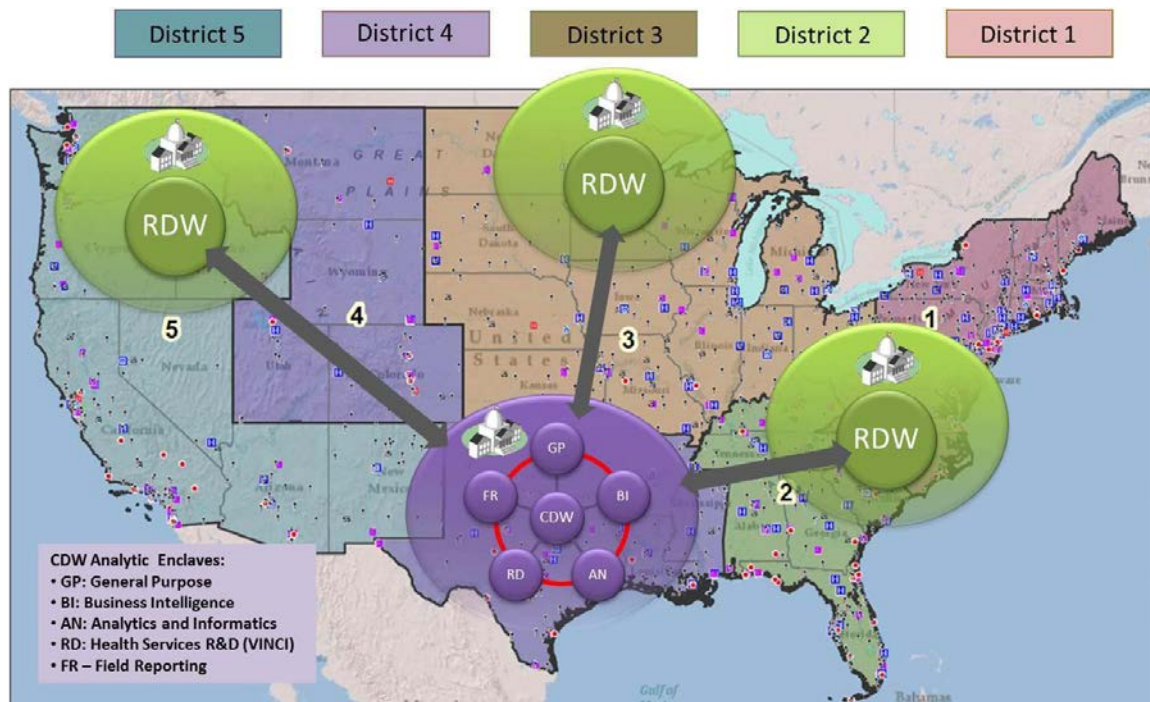
FIGURE 1: ARCHITECTURE OF THE VA INFORMATION AND ANALYTIC ECOSYSTEM

BISL also maintains dedicated master enclaves for specialized, repeated, and/or ongoing analytics efforts. An example of this type of enclave is VINCI, which VHA uses for clinical research and development projects. VINCI contains not only data from the CDW, but also datasets from other federal healthcare organizations such as the Department of Health and Human Services (HHS). It supports initiatives pursuing state-of-the-art analytic capabilities in Natural Language Processing (NLP), Machine Learning, Predictive Analytics, and Clinical Decision Support. VINCI findings and analysis are directly translatable to the VA Analytic Ecosystem for larger VA analysis and implementation.

Consumers in the VA Analytic Ecosystem can combine Tier 1 (raw) data from the CDW with their own data to create data marts of Tier 2 (refined and usable) data. The Analytic Ecosystem provides some rudimentary capabilities for publishing data from these data marts. One of BISL's strategic goals (independent of but complementary to the proposed capabilities in this EDP) is to develop and implement a seamless data-mart-based sharing capability.

2.2 Limitations

2.2.1 VA's Emerging Data Needs Require a New Approach to Analytics

The technology behind newer applications such as the Burn Pit Registry and the Electronic Health Management Platform (eHMP) is driving a rapid growth in the volume, variety, and complexity of data that the enterprise generates and uses. The largely SQL-based analytics technologies that have served VA in the past are not able to keep pace with new services and mission-critical applications. VA's analytics technology needs include:

- Specialized analytics tools for enterprise resource planning (ERP) and customer relationship management (CRM)
- Support for Not Only SQL (NoSQL) and semi-structured/unstructured data
- Streaming and real-time analytics
- Clinical decision support and medical case management analytics
- Streaming analytics to support growing number of Internet of Things (IoT) devices
- Predictive analytics
- Machine learning capabilities

These technology needs cannot be adequately addressed in the fragmented landscape of VA analytics capabilities. More urgently, that fragmented landscape itself is increasingly disconnected from and unsuited for the needs and activities of VA analysts.

Analysts within LOBs and enterprise programs must resort to ad-hoc, labor-intensive processes to obtain the data they need, because there is no alternative available. Even VA enterprise offices and programs with regular collection schedules (like NCVAS) must rely on informal agreements and haphazard data collection practices. If an analyst in one Administration wants to collect data from another Administration, she has no established channels or processes for doing so. Instead she must find a point of contact (POC) in the target organization that is able and willing to help her find and obtain the data she needs.

2.2.2 No Enterprise Organization or Governance for Analytics

The disorganized and ad-hoc nature of VA's enterprise data collection efforts is due to an absence of enterprise governance for and ownership of analytics. BISL and the governance boards that manage the CDW and RDWs Analytic Ecosystem have no general, department-wide authority over analytics at VA. There is no formal enterprise working group or collaboration channel to support cooperation between analytics shops or help analysts navigate the CDW to find the data they need.[3]

The challenges of enterprise analytics data collection are compounded by the difficulty of cleansing raw data to make it usable for analytics. Data gathered from certain VA sources contains pervasive anomalies such as missing birthdates and gender indicators, misspelled names, and incorrect Social Security Numbers. These anomalies – introduced at the point of data entry – are created by an absence of enterprise data quality standards. Correcting these anomalies at the point of analytics data collection entails a significant workload that diverts time and resources from other, potentially more productive activities.[4]

2.2.3 VA's Existing Enterprise Capability is Under-utilized

The VA Information and Analytic Ecosystem were intended to serve as VA's enterprise analytics infrastructure. In principle, it is the ideal platform to support:

- Enterprise analytics collection and data sharing
- Cross-organizational collaboration on analytics projects
- A holistic view of the Veteran population across all Administrations and LOBs
- Centrally managed, shareable analytics tools and advanced capabilities

Many VA organizations do not use the Analytic Ecosystem, because there is no policy requirement to do so. VBA depends on ad-hoc agreements and willing partners to facilitate its (increasingly frequent) efforts to gather and use analytics data from VHA.

Certain VA organizations that operate at an enterprise level understand the value of the Analytic Ecosystem and are working to integrate with it. VRM is increasingly using the

---

[3] The Health Service Research and Development (HSRD) listserv is the unofficial collaboration mechanism for the Analytic Ecosystem/CDW. Despite what the name implies, it is open to all VA users regardless of organization.

[4] Data quality deficiencies are a critical problem in an operational context as well as an analytics context. For example, quality anomalies interfere with the correlation of identity records in the Master Veteran Index (MVI) and its consuming applications (see Interoperability and Data Sharing 3.3: Utilizing Enterprise Identities Design Pattern).

Ecosystem to support its enterprise-wide data collection activities. It is also building a capability to publish Veteran interaction histories to the CDW for consumption by other VA organizations. Analytics shops within the Office of Policy and Planning (OPP) – including NCVAS – are working with BISL to leverage the Analytic Ecosystem for streamlined Department-wide data collection. They are also moving their datasets into enclaves within the Ecosystem.

These organizations (and VA as a whole) cannot realize the full potential benefit of the Analytic Ecosystem unless all the Department's analytics shops integrate with it. Until that happens, VA will continue losing time, money, and potentially valuable insights on haphazard, labor-intensive data collection processes in a siloed environment.

3 FUTURE CAPABILITIES

Many organizations within VA recognize that meeting their emerging data needs requires not only new technologies, but an entirely different approach to managing analytics. They also recognize that they cannot make the necessary changes without leadership and direction from the VA enterprise level.

This Enterprise Design Pattern addresses VA's current analytics gaps and anticipated needs by addressing the following considerations:

- VA enterprise roles and responsibilities for analytics governance, policy, standards, training, provisioning, and management
- Policy and governance framework for analytics, particularly data sharing in a "circle of trust" model
- Guidelines for integrating existing analytics silos and operational systems with the VA Analytic Ecosystem
- Evaluating, selecting, and using analytics technologies, to include advanced analytics capabilities (e.g., predictive analytics, machine learning)

These capabilities will align VA's analytic environment with its technological, strategic, and operational landscape, partly by serving as the analytics portion of a future "data backbone." VA will also be better able to respond to changing technological needs and opportunities, and allow everyone in the Department to fully leverage newly deployed analytics capabilities.

3.1 Organizational and Personnel Support for VA Analytics

Problem(s) Addressed: No Enterprise Organization or Governance for Analytics

Integrating VA analytics silos into an enterprise-wide infrastructure, and sustaining that infrastructure, requires coordinated organizational and personnel support in the form of:

- Designated governance roles and responsibilities for analytics
- Dedicated subject matter experts (SME) and collaboration mechanisms to help users locate and leverage resources
- A coordinated analytics training program that allows analysts and data scientists to keep up with technology trends

3.1.1 Governance Roles and Responsibilities

Many large private-sector organizations create executive ownership positions responsible for governing and managing analytics: a Chief Data Officer and a Chief Analytics Officer. The two roles are complementary and even dependent on each other. The Analytics Officer requires input (data) furnished by the Data Officer, while the Data Officer requires feedback from the Analytics Officer to effectively manage data collection, storage, and operational use. Figure 2 below illustrates their responsibilities with regard to analytics.

Chief Data Officer	Chief Analytics Officer
• Maintain global awareness of organizational data sources • Acquire and ingest data from organizational data sources • Establish requirements and duties for data owners/stewards • Set and enforce standards and processes for data (e.g., modeling, architecture, quality)	• Acquire, develop, and maintain assets (hardware/software) used for analytics • Manage, coordinate, direct, and train analysts and data scientists • Select and apply appropriate tools/methods for processing and analyzing data • Deliver information derived from analytics using tools and techniques appropriate to the content and audience

FIGURE 2: DUTIES OF A CHIEF DATA OFFICER AND CHIEF ANALYTICS OFFICER

VA's Chief Data Officer[5] will:

- Help BISL refine and implement its repeatable process for integrating analytics silos with the VA Analytic Ecosystem, to be executed by analytics integrated project teams (IPT)

---

[5] As of this writing (January 2016), VA has recently created a Chief Data Officer position, but it has not been filled.

- Maintain global awareness of VA assets (i.e., data sources, programs, and projects) that may align with or impact BISL's functions
- Establish requirements and duties for data owners/stewards
- Develop and enforce data standards (including quality, modeling, and architecture standards) for VA's operational data stores
- Implement policies and standards created by the VA Data Governance Council (DGC)
- Ensure that current and future VA data assets integrate with and provide sufficient, usable data to the VA Analytic Ecosystem

Until (or unless) VA creates a Chief Analytics Officer position, BISL and the CDW/RDW governance boards will perform the equivalent functions.

BISL and the CDW/RDW governance boards will continue to perform their current functions, to include developing, managing, and improving the VA Analytic Ecosystem and its capabilities. The governance boards will provide high-level planning, policy, and standards for VA analytics. They will also select and prioritize new analytics technologies for the Technical Reference Model (TRM) evaluation for implementation in the VA Analytic Ecosystem. BISL will implement or enforce their directives as appropriate.

To effectively manage analytics for the entire Department and integrate all VA analytics capabilities and functions into the Ecosystem, they will require:

- Policy and resource support from VA leadership, including the ability to enforce established policies and standards
- Mandate for BISL to take ownership of all present and future enterprise licenses for VA analytics tools and technologies approved by the TRM[6]
- Governance board representation/participation from all VA Pillars (including the EPMO), enterprise programs, and Administrations[7]

They will also implement or help to implement the capabilities and solutions described in this Enterprise Design Pattern (particularly Section 3.2) which are not otherwise assigned to another role.

---

[6] See Section 3.3.1 for additional details on tool consolidation and license management.
[7] At this time (December 2015), NCA and VBA have little or no involvement in the governance boards, since they do not use the VA Analytic Ecosystem as intensively as VHA.

3.1.2   Dedicated Subject Matter Experts and Collaboration Mechanisms

The CDW and RDWs already contain vast amounts of data, and it is difficult for analysts who are unfamiliar with the warehouses to find what they need for a particular project. There are some tools available for searching the contents of the warehouses (such as VHA's Reports and Measures Portal), but the contents are still difficult to sift through.

While the HSRD listserv is a helpful resource for anyone who needs help locating a particular dataset, it is not a formal CDW resource and many analysts outside VHA are unaware of it. As more VA analytics shops transition into the VA Analytic Ecosystem, they will need dedicated, easily reachable data librarians to help them navigate the CDW and locate useful data. The Chief Data Officer will work with BISL, the warehouse governance boards, and HSRD to establish one or more formal data librarians and/or a data helpdesk for the CDW.

BISL and HSRD will turn the HSRD listserv into an enterprise collaboration resource for VA analytics, both in general and for specific tools and capabilities. They will also create, or allow the HSRD listserv community to create, other collaboration sites and mechanisms. These collaboration mechanisms will allow analysts to exchange knowledge about, and consequently leverage, techniques and capabilities that were previously unknown to them.

3.1.3   Analytics Training Program

BISL provides training materials, classroom courses, and other resources to train VA users and analysts in how to use the VA Analytic Ecosystem and the tools available in it. As one of the principal governance bodies for analytics in the VA enterprise, they will be responsible for developing (or acquiring) and providing all analytics training in the Department. They will provide training on any new tools or capabilities acquired through consolidation of analytics silos. In particular, they will train VA analysts to use advanced and/or newly acquired technologies and capabilities that are unfamiliar to them.

3.2 Policy and Governance Framework for Analytics

Problem(s) Addressed: No Enterprise Organization or Governance for Analytics

VA's analytics governance bodies will build on existing VA Analytic Ecosystem standards and policies to create an enterprise-wide analytics policy and governance framework. This framework will enable:

- Standardized long-term agreements for analytics data sharing and disclosure between VA organizations, particularly BISL and its consumers[8]
- Consistent organization-wide adherence to processes, business rules, and protocols for analytics data collection and sharing within the VA Analytic Ecosystem
- Resolving disputes and other issues

This section provides essential elements, principles, and concerns to use as a starting point for building an enterprise analytics policy and governance framework, specifically:

- A conceptual "circles of trust" model for data classification and sharing
- Provider and consumer responsibilities in the circles of trust model
- Minimal data quality requirements

3.2.1 The Circles of Trust Model

The VA Analytic Ecosystem will allow consumers to impose broad access controls on analytics data and data products[9] they own by classifying them into one of three "circles of trust":

- **Private:** The narrowest and most restrictive circle of trust. Data and data products in this circle are accessible only to internal users of the consumer that owns them.
- **VA Internal:** The middle circle of trust. Data and data products in this circle are accessible to all VA organizations.
- **Public:** The widest circle of trust. Data and data products in this circle are accessible to third parties outside of VA.

In this model, the default classification for consumer data is Private, and consumers have control over what data they will share (or not) with which circles of trust. Consumers may make their sharing decisions based on considerations such as:[10]

- Organizational mission or business needs
- Policy, regulatory, or statutory requirements to publicize or protect certain data
- Contextual concerns, for example completeness (or lack thereof) of the dataset

---

[8] In the context of this Enterprise Design Pattern, "consumers" are internal VA organizations (e.g., Administrations, LOBs, projects) that use or will use the VA Analytic Ecosystem for their analytics activities.
[9] Data products are insights, information, and business intelligence (BI) derived from analytics.
[10] The "Sharing Analytics Data With Different Circles of Trust" use case (Section 4.1) provides an example of an LOB classifying its analytic data assets based on these considerations.

Any data that a consumer shares with a broader circle of trust will also be accessible to the narrower circle(s) of trust. That means anything shared with the Public circle of trust is also shared within the VA Internal circle of trust.

As a general rule analytic products will go into the same circle of trust as the most sensitive pieces of data from which they were derived (the "highest classification" approach). Table 1 below shows default circle of trust assignments for analytic products, based on the circle of trust assignments for the data from which they were derived.

TABLE 1: DEFAULT CIRCLE OF TRUST ASSIGNMENTS FOR ANALYTICS PRODUCTS

	Private	VA Internal	Public
Private	Private	Private	Private
VA Internal	Private	VA Internal	VA Internal
Public	Private	VA Internal	Public

The highest classification approach is intended as a guideline rather than a rule: there are valid business cases for making exceptions. Owners of data and analytics processes will decide when it is appropriate to make such exceptions, provided that they are consistent with applicable laws, regulations, and policies.

3.2.2 Provider and Consumer Responsibilities within the Circles of Trust Model

- Within this framework, BISL (i.e., the provider) will:
- Ensure that any otherwise "unclassified" data ingested into the Ecosystem is automatically classified as Private to the owner of the data source.
- Enable data owners and stewards to classify specific datasets and/or data elements in the VA Analytic Ecosystem as Private, VA Internal, or Public (for data they own).
- Enforce classifications by data owners in the VA Analytic Ecosystem, consistent with the model presented in this Enterprise Design Pattern.
  - o Encrypt and control access to Private data sets or elements in transit or at rest within the VA Analytic Ecosystem.[11]
  - o Share Public data through enclaves or platforms that are logically separate from the CDW/RDWs and Private enclaves.

---

[11] By necessity, all data ingested into the VA Analytic Ecosystem must pass through the RDWs and/or CDW, even if it is Private. Since many organizational big data environments operate in a similar way, there are a variety of tools available to mask or redact selected data sets/elements in such environments.

- o Adhere to a blanket non-disclosure agreement for Private data accessed as part of analytics management, configuration, and troubleshooting activities.
- Provide data owners with the means to further restrict access to their own Private data within their organization.

VA consumers will:

- Classify their datasets and/or data elements (as Private, VA Internal, or Public).
- Further restrict access to their own Private data to certain individuals, groups, or roles within their organization (if desired) using tools provided or approved by BISL.
- Ensure that their classifications and access control decisions comply with applicable statutes, regulations, policies, security requirements, and business processes.

The VA analytics governance entities will work with the Office of General Counsel (OGC) to develop a policy framework for data sharing, ownership, and responsibility within and across the circles of trust. They will also develop standardized, long-term data sharing and collaboration agreements based on this framework.

In addition, OGC and the analytics governance entities will answer the following questions:

- What information about data provenance and lineage needs to be tracked for all data within the VA Analytic Ecosystem?
- Who is responsible for maintaining data integrity in the Private, VA Internal, and Public circles of trust? Is it BISL, the consumer/data owner, or some combination of both?
- Can data owners ever impose additional duties/obligations on BISL for their data in the VA Internal or Public circles of trust? If so, what additional obligations can they impose?
- If a data owner releases certain data into the VA Internal and/or Public circles of trust and they subsequently want to change it or remove it:
    - o Are there any circumstances under which they should be able to do so?
    - o If they can, what process must they go through to affect the change/removal, either before or after the fact?
- Can VA organizations share their Private data (and related analytics products) solely with each other, rather than the VA Internal circle of trust?
- Where do the Department of Defense (DoD) and/or Federal agency partners fit into this model? They could be treated as VA consumers with their own Private circles of trust, part of the VA Internal circle of trust, or placed in some special category.
- Can BISL (or any VA organization) ever contest overrule another organization's classification of data? For example, can BISL compel a VA organization to make some of

its Private data VA internal, or refuse to release certain data into the Public circle of trust?

3.2.3   Minimal Data Quality Requirements

Most of the analytic data in the VA Analytic Ecosystem belongs to VHA and conforms to VHA's stringent data quality standards. Other VA organizations may have few if any data quality standards, and as a consequence their operational and analytic data are affected by significant quality issues. The most critical issues tend to involve missing, inconsistent, or incorrect identity and demographic data.[12]

Fully addressing data quality assurance will require implementation of data quality standards in the operational environment (which is beyond the scope of this Enterprise Design Pattern). Until then, VA analytics governance bodies will develop minimal quality assurance standards for structured data in the VA Analytic Ecosystem. BISL will enforce these standards to the greatest extent practicable. Primarily, enforcement will involve refining data in staging areas or "data pools" before releasing it into the CDW/RDWs and analytic enclaves (i.e., making it available for analysis).[13]

There are two exceptions to this quality assurance requirement. The first is analytic processes that are designed to use unstructured or raw data, such as streaming analytics, sentiment analysis, or data exploration. The input for these processes does not have to be accurate or consistent (in fact it is assumed not to be), so adherence to quality standards is not a concern.

The second exception is an emergency situation in which it is necessary to pull raw structured data from operational systems without refining (e.g., cleansing, normalizing, de-duplicating) it first. The VA Analytic Ecosystem supports this capability for the Veterans Health Information Systems and Technology Architecture (VistA). BISL will develop a universal protocol for emergency raw data pulls from other operational systems, based on its procedures for VistA data pulls. The protocol will provide consumers with needed flexibility while preventing overuse or abuse of the raw data pull capability.

---

[12] These quality issues are described in Section 2.2.2 of this document. A more detailed description can be found in Interoperability and Data Sharing 3.3: Utilizing Enterprise Identities Design Pattern.
[13] The use of staging areas and refining tools is described in greater depth in Section 3.3.2 of this document.

3.3 Integrating VA Analytics Silos and Data Sources with the VA Analytic Ecosystem

Problem(s) Addressed:

- Emerging Data Needs Require a New Approach to Analytics
- Existing Enterprise Capability is Under-utilized

The extensible, loosely coupled architecture of the VA Analytic Ecosystem makes it an ideal foundation for a VA enterprise analytics capability. Consolidating VA analytics operations into (or at least close to) the Analytic Ecosystem necessitates attention to the following issues:

- Consolidating VA analytics tools and capabilities
- Connecting warehouses to the Analytic Ecosystem as enclaves
- Using staging areas for decoupling and refining
- Linking operational data stores with the Analytic Ecosystem
    - Prioritizing data stores for connection to the Analytic Ecosystem
    - Leveraging the future operational data lake

BISL will refer to the VA Data Inventory (VADI) and Data Architecture Repository (DAR) for documentation regarding consumer data, metadata, and related business processes. Data owners are responsible for cataloging and documenting their data and metadata in VADI and DAR. They will also provide BISL with any additional information necessary to facilitate integrating their warehouses with the VA Analytic Ecosystem.

When the planned eCRUD service is implemented as part of the VA EA data layer (described in the HDA Enterprise Design Pattern), the VA Analytic Ecosystem will use it to support:

- Access control, to include mediating access to analytics data and data products through/by applications and production systems
- Logging and auditing
- Data ingest from applications, sensors, social media feeds, and other sources external to the VA EA data layer[14]

This section refers to concepts described in Appendix B: Concepts for Service-Oriented Analytics.

---

[14] eCRUD supports the capability to incorporate or integrate with data ingestion technologies, so any ingestion capabilities in use prior to the deployment of eCRUD can be added to it in the future.

3.3.1 Consolidating VA Analytics Tools and Capabilities

Consumers access BISL's catalog of analytics software tools through the Analytics App Store. While it is extensive, this catalog may not include all the TRM-approved analytics tools in use at VA. BISL will add these tools to its catalog and make them available to consumers as it takes ownership of the enterprise licenses for them (as described in Section 3.1.1). If there is a sufficient level of demand and/or business case for implementing the requested tool or capability, BISL will acquire it and make it available to consumers.

BISL will assume all future licensing costs for analytic technology as part of its enterprise approach to license, support, and training management. It will use blanket enterprise licensing agreements (consistent with existing VA Analytic Ecosystem practices) to:

- Reduce unnecessary and duplicative costs associated with technology acquisition and support
- Ensure that analytics tools are readily available to all consumers through the VA Analytic Ecosystem App Store
- Acquire and provide high-quality training for analytics tools and technologies
- Foster cooperation and knowledge transfer in the VA user community
- Complement efforts to use cloud- and Web-based analytics tools (rather than client-based tools)

In a similar vein, BISL will incorporate siloed analytics processing frameworks into the VA Analytic Ecosystem as enclaves.

3.3.2 Using Staging Areas for Decoupling and Refining

BISL refers to raw ingested data as "Tier 1" data and refined, stored data as "Tier 2" data. All Tier 1 data in the VA Analytic Ecosystem will be routed through one or more intermediate staging areas[15] before being stored in the CDW, RDWs, and/or enclaves as Tier 2 data. This requirement applies to data input from all sources, to include:

- Operational VA data stores that support one or more VA applications
- Legacy (formerly siloed) warehouses integrated with the VA Analytic Ecosystem as Private enclaves
- Sensors, IoT devices, monitors, and other streaming data sources
- The future operational data lake described in the HDA Enterprise Design Pattern

---

[15] BISL uses the term "staging areas": these may also be called "data pools" or "landing zones."

Staging areas will be used to:

- Serve as dedicated platforms for refining of ingested data prior to storage
- Decouple data ingest, refining, and storage functions in the analytic environment, facilitating easier management and troubleshooting
- Apply tags/metadata to ingested (Tier 1) or cleansed (Tier 2) data
- Correlate or link one dataset with other datasets
- Facilitate replicating the same data to different routes/stores for different analytics processes and use cases
- Support maintenance and improvement of different refining operations, including resequencing of operations for individual data flows

In addition to staging areas for automated "pass-through" functions, the Ecosystem will include holding areas for data input that requires human/manual attention. If BISL has reason to suspect that data from a particular source has issues that make it unsuitable for use in analytics, they will divert input from that source to a holding area. Reasons for diverting data to a holding area include:

- The data needs to be reviewed/evaluated as part of due diligence because:
  - It is the first input[16] from a newly connected or integrated source
  - The mechanisms or sequence used for ingesting and/or refining the data have changed
- Lack of compliance with applicable quality or integrity standards (from BISL, the governance boards, the data owner/consumer, DGC, VA)
- Possible leakage of sensitive data
- Unanticipated and undesirable changes to the data (e.g., corruption) caused by some unknown issue

The diverted data will remain in the holding area until:

- Issues with the data, data source, ingestion and/or refining mechanisms are resolved, and the data is released into the CDW/RDWs for use in analytic operations
- The data is deemed irrecoverable and deleted

---

[16] Depending on the type of data ingest used, this may be the first batch transfer or a sample (e.g., the first 10 gigabytes or 24 hours) of input.

Most refining operations – and the technologies that execute them – will likely be used in multiple data flows and analytics use cases. BISL will select, implement, and maintain technologies (to include cloud services, where security requirements permit) for refining in staging areas on an operation-by-operation basis.[17] BISL will determine and configure the appropriate routing of data inputs and data flows through staging areas, based on:

- Compliance with established minimal data quality requirements
- Quality, consistency, and formatting requirements specified by the consumers of analytic processes that use a particular data source/flow
- Functional, performance, and business needs of the VA Analytic Ecosystem

Consumers may have read access to the staging areas, including any contents they own and/or have permission to view. BISL may grant consumer's temporary write access to specific staging areas for a particular purpose, e.g., to resolve anomalies in data retained in a holding area.

Until the proposed eCRUD service is deployed, BISL will use special staging "airlocks" to facilitate data sharing with external organizations and systems (see Section 3.5.5).

3.3.3   Integrating Existing Warehouses with the VA Analytic Ecosystem

The VA Analytic Ecosystem will leverage and support existing data collection activities in the short-to-medium term by integrating siloed VA warehouses into the Ecosystem as Private enclaves. BISL (with the support of the EPMO) will use their repeatable, customizable integration processes to integrate all siloed VA warehouses in this way. It may not be technically feasible to integrate some warehouses into the Ecosystem without seriously compromising functionality or performance. In such cases, the owners of the warehouses will work with BISL to transition their analytics data and activities to one or more Private enclaves.

Consumers who want to share some of their data with the VA Internal and/or Public circles of trust will sign an agreement with BISL that includes:

- Description and specification of the data to be shared, and with which circle(s) of trust
- Specification of how data will be transferred from the warehouse to the CDW/RDWs (e.g., periodic batch transfers, transactional updates)
- Permission for BISL to access the data for management, configuration, and troubleshooting purposes

---

[17] A sample (not comprehensive) list of refining operations can be found in Section B.2.2.

- Statements or assertions that the data owner will accept responsibility for:
  - Appropriateness of access control decisions (including circle of trust classifications)
  - Helping BISL address any currently known or to-be-discovered anomalies or quality issues in the data sets to be shared

Initially, the warehouse will be treated as a new source of data, meaning that the first input provided by the data source must be diverted to a holding area for evaluation.

3.3.4   Linking Operational Data Stores with the Analytic Ecosystem

In the long term, all operational VA data collected for analytics purposes will be ingested directly into the VA Analytic Ecosystem. Data sources feeding into siloed data warehouses will be connected directly to the Ecosystem over time. Any future applications, tools, sensors, or other data sources will rely exclusively on the VA Analytic Ecosystem for analytics data collection and support. All data collected from operational data stores of any kind will be "Private" to the data owner unless otherwise specified. The data owner is responsible for deciding which data sets and/or data elements are appropriate for sharing within the VA Internal and Public circles of trust.

Since there are many existing data sources to connect with the Ecosystem and many new ones to come, BISL will prioritize data sources for connection in the following order:

- Designated authoritative data sources (ADS). These sources will also be used for master data management, correlation, and de-duplication within the VA Analytic Ecosystem.
- Sources from which enterprise programs and Department-level offices (e.g., OPP, NCVAS, VRM) collect analytics data.[18]
- Newly implemented data sources, to include sensors/wearables, application databases, social media feeds, etc.
- The operational data lake described in the HDA Enterprise Design Pattern (when implemented). While this data lake will perform some of the same functions as staging areas, it is not exempt from the staging area requirement applied to other data sources.
- Data stores from which analytics data is already being collected in non-BISL warehouses.

---

[18] As of December 2015, BISL is already working to provide these programs and offices with easier access to analytics data through the VA Analytic Ecosystem.

3.4 Evaluating, Selecting, and Using Analytics Technologies

Problem(s) Addressed:

- Emerging Data Needs Require a New Approach to Analytics
- Existing Enterprise Capability is Under-utilized

VA will use multiple interoperable technology components through which data flows can be routed as appropriate to their respective use cases. All analytics technologies used by VA will support:

- Deployment, management, and provisioning within or from the VA Analytic Ecosystem as described in Sections 3.2 and 3.3
- Interoperability with common/open standards and BI tools, particularly those already deployed or in common use within VA

The following subsections address selecting or designing context-appropriate technologies and approaches for data storage, analytics processing, and information delivery/visualization. Possible options for technologies at any stage include cloud services vetted and used by BISL, where security requirements permit.

This section uses terms and concepts described in Appendix B: Concepts for Service-Oriented Analytics.

3.4.1   General Principles for Technology Selection

Analysts, data scientists, data owners, and other decision makers will architect analytic processes according to functional and business requirements, rather than building these processes around specific technologies. They will determine requirements, develop approaches, and select technologies for:

- Data storage (Section 3.4.2)
- Processing/analysis (Section 3.4.3)
- Visualization and delivery (Section 3.4.4)

Technologies or cloud services for these functions may be selected in any order or independently of each other, so long as the selections can interoperate.

Decision makers will, whenever practicable, choose from among technologies already deployed in the VA Analytic Ecosystem and/or cloud services used by BISL. If they have a use case that

cannot be supported with currently deployed technologies, they will work with the governance boards to obtain the technologies they need.

3.4.2   Data Storage

Decision makers will select a type of data storage that best meets the functional and business requirements of their analytic use case. The response surface in Figure 3 below plots possible storage types in terms of data temperature (X-axis) and structural complexity (Y-axis).
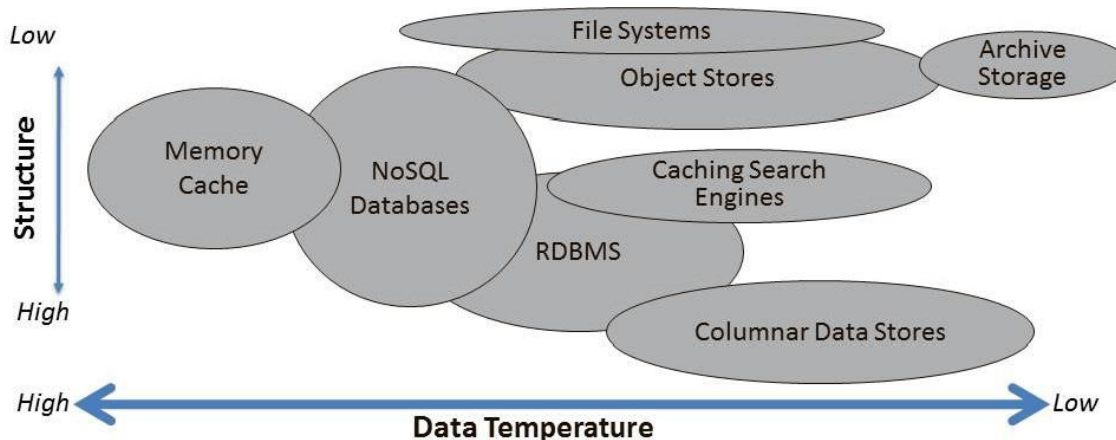
FIGURE 3: RESPONSE SURFACE FOR SELECTING DATA STORAGE TYPE

To determine the temperature of the data under consideration in the use case, refer to Table 6 in Section B.1. The table provides guidance for determining data temperature based on characteristics such as volume, latency, and request rate. Plot the options for the relevant data temperature range on a decision grid that compares data structure complexity and query structure complexity, as shown in Figure 4.[19]

Choose the decision grid quadrant that best corresponds to the data structure and query requirements for the analytic process. If there is more than one possible option in the selected quadrant, evaluate the low-level characteristics of each option against the functional requirements of the analytic process to find the best fit.[20]

---

[19] The example decision grid in this figure contains storage types for multiple data temperatures, but the decision grid used in an actual selection process will only include options from a single temperature range.
[20] In-depth guidance on selecting data storage types (for analytics and other purposes) are a possible topic for future EDP increments.
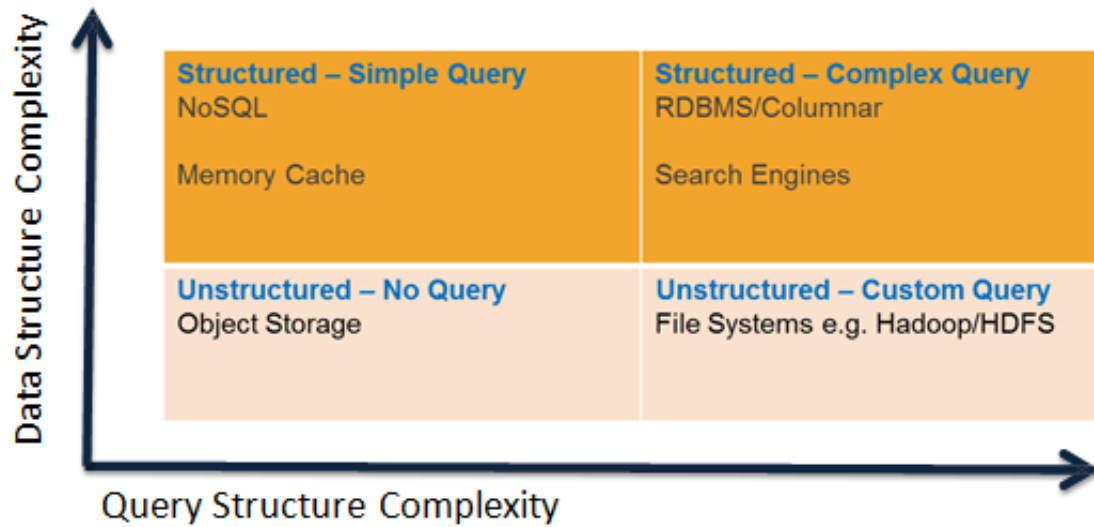
FIGURE 4: DECISION GRID FOR EVALUATING STORAGE OPTIONS

3.4.3 Analytic Processing Frameworks

Decision makers will evaluate and select technologies that best meet their needs in terms of processing type, attributes, and supporting storage technology.

The two major types of processing frameworks are:

- **Batch:** Queries a large amount of cool or cold data and asks complex questions, taking minutes or hours to generate answers. Used for operations in which accuracy is more important than speed of response.
- **Stream:** Queries a small amount of hot data with simple, highly specific questions, taking milliseconds, seconds, or minutes to generate answers. Used for real-time (or near-real-time) analytics where speed of response is more important than accuracy.

The attributes to consider for analytic processing frameworks differ depending upon whether the processing framework under consideration is batch or stream. Figure 5 below shows relevant attributes for batch and stream processing.

Batch

- **Query latency:** Amount of time required to execute a query. The larger the volume of data being processed, the higher the latency will be. In use cases with cool or cold data, high query latency is not a critical issue, but it is a concern in cases with hot data.
- **Durability:** Degree to which committed data transactions will survive permanently.
- **Storage:** The type of data storage used by the processing framework. Corresponds to the type of storage technology chosen in the data storage types selection approach.

Stream

- **Scale/throughput:** The ingestion bandwidth of the processing framework. Typically depends on the number of nodes used.
- **Fault tolerance:** Capacity to recover from node crashes or other errors and continue operating normally.
- **Programming languages:** Which programming languages the framework supports.

Both Batch and Stream

- **Data volume:** Maximum volume of data that can be processed by the framework at any one time. In some batch frameworks and most stream frameworks, the maximum volume depends on the number of nodes used.

FIGURE 5: RELEVANT ATTRIBUTES FOR BATCH VS. STREAM PROCESSING FRAMEWORKS

Plot the candidate batch or streaming technologies on a decision grid. The example decision grids in this Enterprise Design Pattern use dummy products (e.g., Product 1) as placeholders for TRM-approved technologies.

Table 2 is an example of a decision grid for batch processing technologies, in which the options are ordered in terms of query latency (lowest to the left, highest to the right).

TABLE 2: EXAMPLE DECISION GRID SELECTING BATCH PROCESSING FRAMEWORKS

	Product 1	Product 2	Product 3	Product 4	Product 5
Query Latency	Low	Low	Low	Low-Medium	Medium-High
Durability	High	High	High	High	High
Data Volume	1PB Max	# Nodes	500 TB Max	# Nodes	# Nodes
Storage	Native	HDFS	NoSQL	HDFS/Object	HDFS

Table 3 is a decision grid for stream processing technologies. Options in this table may be ordered depending upon the most critical attribute for the use case (e.g., fault tolerance, programming languages supported).

	Product 1	Product 2	Product 3
Scale/Throughput	# Nodes	# Nodes	# Nodes
Data Volume	# Nodes	# Nodes	# Nodes
Fault Tolerance	Built-in	Built-in	Optional
Programming languages	Java, Python, Scala	Java, Scala, Clojure	Python, Java

Identify the technologies that most cost-effectively support the required attributes for the use case. If there is more than one possible option, evaluate the low-level characteristics of each technology against the functional requirements of the use case to find the best fit.

3.4.4 Delivery and Visualization

The goal of delivery is to integrate analytics into operations in such a way that findings from analytics drive concrete actions and improvements. Decision makers will use a business requirements definition approach to selecting or designing tools and approaches for delivering analytic products to end users. "End users" in this case are not analysts or data scientists, but other organizational roles who will act on information derived from analytic activities.

To define the business requirements for delivery and visualization, begin by identifying the user population(s) to be served by the analytic process. With respect to each user group identified, investigate and answer the following questions:

- What information do they want from the analytic process?
- When do they want it?
  - On a schedule (e.g., weekly, monthly, quarterly)
  - In real time or near-real time
  - When they perform a certain task or operation
- Where do they want it?
  - In a report or dashboard
  - As an alert through email, text, or another messaging system
  - In an application interface to facilitate decision support[21]
- Do they want (and can we provide) some level of interactivity for the delivery mechanism, such as:
  - The ability to build/customize a dashboard or report using predefined analytic logic elements and/or snippets of code

---

[21] Delivering analytics for real-time decision support entails collaborating with application developers and/or providing them with mechanisms (e.g., message channels) to integrate analytics data into their applications.

o Summary and detail views of information
o Visibility into the analytic process/logic that produced the information

Whenever possible and appropriate, leverage "watering holes" that target user population(s) already use or visit on a frequent basis, including preexisting reports and analytics dashboards. Other possibilities include newsletters, e-mails (for alerts that do not require an immediate response), periodic status meetings, and widgets on internal websites.

In some cases the end users for the dataflow will be analysts or third parties who want "direct data delivery." This means that they want the data itself (rather than a data product) so that they can run their own analysis. This requires a different kind of channel or mechanism than conventional delivery, which may involve one or more of the following:

- Publishing items to a library, catalog, or inventory
- Tagging and/or indexing items so they can be located with a search function
- Pushing out alerts to subscribers/interested parties when new items become available

Consider using cloud platforms (consistent with VA's cloud strategy) for direct data delivery, particularly when delivering Public data.

3.5 Analytic Capabilities in High Demand

Problem(s) Addressed: _Emerging Data Needs Require a New Approach to Analytics_

This section of the document lists and summarizes analytics capabilities that stakeholders have requested to meet emerging data needs.

These capabilities will be deployed in the VA Analytic Ecosystem as enclaves, App Store offerings, or both.[22] Some of the listed capabilities are already present in VA's technology environment, but only on a limited or partial basis. Others are not yet deployed in VA at all and may take some years to implement, especially those which are dependent on changes to operational systems. BISL and the governance boards will determine the timeline and order of prioritization for implementing these capabilities.

3.5.1 Next-Generation Analytics Technologies

Many of the capabilities described in the following subsections will depend upon or greatly benefit from some combination of next-generation analytics technologies.

---

[22] TS (in collaboration with BISL) may explore these capabilities in future EDP increments

- **Streaming analytics** ingest and process continuous streams of data from various sources, particularly sources that have little or no data storage capacity of their own. By nature, streaming analytics are critical to ingest and make sense of data from IoT devices. They are also a prerequisite for real-time analytics.
- **Real-time and near-real-time analytics** are valuable in situations where the window of opportunity to act on information is small. One example is providing emergency alerts based on sensor (e.g., medical monitoring device) data. Real-time analytics also enable streamlined real-time management reporting and point-of-service decision support.
- **Predictive analytics** make projections based on historical and current data, making them extremely valuable for financial planning and enterprise resource planning (ERP). Combined with real-time analytics, predictive analytics enhance alerting and point-of-service decision support capabilities.
- **Context sensing** uses cues such as time, location, and user roles to enable, disable, or modify certain functions so that they behave in a situationally appropriate manner. Geofencing is a context sensing capability: so is automatically providing silent (rather than audible) alerts when a user's Outlook calendar indicates that they are in a meeting.
- **Machine learning**[23] is the capability to discover and recognize patterns in data input, with or without human supervision. An example of supervised machine learning is analyzing user input in the performance of a manual task to build algorithms and business rules that will automate that task. Unsupervised learning is using a machine learning capability to explore a body of data in search of patterns and insights without human instruction.

3.5.2   User Experience Analytics

The VA Analytic Ecosystem will facilitate collection and analysis of user experience data for all platforms, including mobile platforms. Many mobile application management/development platforms have the capacity to collect such data: VA will acquire and deploy a mobile application stack that supports this capability. BISL will collaborate with human-computer interaction specialists and interface designers in VA to develop strategies and approaches to collecting and analyzing user experience data.

3.5.3   Data Laboratories

The VA Analytic Ecosystem will support provisioning of analytic environments/enclaves specifically for data exploration and research. By definition, these data laboratories will not

---

[23] Sometimes used interchangeably with the term "pattern cognition"

deliver output to operational systems and functions (e.g., decision support, alerting). Unsupervised machine learning will only be conducted in data laboratories.

3.5.4   Improved Integration of Operational and Analytics Architectures

BISL will assist data and system owners in integrating their transactional systems with the VA Analytic Ecosystem to facilitate using analytics in decision support within application interfaces. If possible (given the variety of operational systems and data stores), they will develop a repeatable process for doing so. BISL and system owners will focus particularly on ways to integrate operational and analytics systems for real-time decision support.

The planned eCRUD service and authoritative information services will serve as common portals for everything in the VA EA data layer.[24] When implemented, they will help to facilitate this type of integration.

3.5.5   Exchanging Data with Trusted Partners and Services

- The VA Analytic Ecosystem will support secure, streamlined data sharing with designated organizations, systems, and services, to include:
- Subscribing to analytic data sets, knowledge bases, and periodic transfers/updates from external sources
- Publishing data to external systems and services for public consumption

To support this capability, the governance boards will designate:

- Market-compatible standards that the VA Analytic Ecosystem will employ for external data collection and data sharing
- Trusted "publisher" organizations from whom VA will accept analytic data transfers, data sets, knowledge bases, etc., to include:
    - o   DoD
    - o   Federal departments and agencies
    - o   Medical research institutions
- "Subscriber" organizations and services to whom VA will publish data placed in the Public circle of trust, to include:
    - o   Trusted publisher organizations, at their request
    - o   Partner organizations, to include DoD, Federal agencies, and VSOs

---

[24] Refer to the HDA Enterprise Design Pattern for a description. Note that the Data Layer will not include VistA systems, which are being encapsulated in the VistA Service Assembler (VSA) Platform.

      o   Cloud services and applications, in compliance with VA's cloud strategy and applicable Enterprise Design Patterns

The VA Analytic Ecosystem will use eCRUD to mediate data transfers between it and external systems. Until eCRUD is deployed, the VA Analytic Ecosystem will use staging airlocks to facilitate secure data sharing. Each airlock will consist of a pair of staging areas: one will reside within the VA Analytic Ecosystem and its partner will reside in the VA network demilitarized zone (DMZ). Incoming data transfers will flow through the DMZ staging area into the Ecosystem staging area: outgoing data transfers will flow in the opposite direction.

BISL and the governance boards will evaluate cloud services as potential sharing or delivery mechanisms for exchanges within VA, with partner organizations, and with the public. They will develop repeatable approaches – consistent with VA's cloud strategy and security requirements – to using private, community, public, and hybrid clouds for data exchanges.

3.5.6   Combined Clinical/Operational Data Analytics

While VHA has invested heavily in developing its clinical analytics capabilities, its operational analytics capabilities (particularly with regards to financial analysis) are not nearly as mature. VHA has also recognized that decoupling these two spheres from each other makes it difficult to assess the cost-effectiveness and long-term outcomes of medical treatments.

To facilitate combining operational and clinical analytics, the VA Analytic Ecosystem will support the following capabilities:

- Tools for health care system (patient, provider, and payer) health care analytics
- Exploration and testing of different data models for bodies of unstructured data
- Real-time application of clinical models

3.5.7   Integrated Financial and Management Analytics

BISL and the governance boards will acquire, develop, and/or improve analytics capabilities for financial analysis and management, to include enterprise resource planning (ERP). These capabilities will depend on collecting and integrating data from various financial and accounting systems,[25] streaming real-time analytics, and predictive analytics.

One critical need in this area is providing effective long-term cost management recommendations as part of point-of-care decision support. These recommendations will be

---

[25] Unless and until the operational systems are integrated and/or overhauled.

based on the total lifetime costs of certain treatment options (e.g., medications) rather than point of purchase costs.

3.6 Alignment to the One-VA Technical Reference Model (TRM)

This section provides examples of TRM-approved tools that are currently used, or may be used, in the VA Information and Analytic Ecosystem. Future analytics capabilities are bound by approved technologies and standards cataloged in the TRM.

TABLE 4: ALIGNMENT TO THE TECHNICAL REFERENCE MODEL

Tool Category	Example Approved Technology
Business Intelligence Platforms	• BI Office • Business Objects • Cognos Enterprise – v10.2 • Oracle Business Intelligence Publisher • SAS Online Analytical Processing Server • Tableau Desktop – v9.0
Dashboard/Scorecard Tools	• Cognos Enterprise – v10.2 • iDashboards Enterprise Suite – v8.x • Rational Insight – v1.1
Data Analytics (Statistical Analysis, Prediction, and Modeling)	• BI Office, Cognos Enterprise – v10.2 • Comprehensive Meta-Analysis (CMA) • Datawatch Desktop • EpiData – Analysis 2.2, Entry 3.1, EpiC 3.1 • iDataFax • JMP – v10.0.x, 11.0.x • MAXQDA • MedCalc – v13.0, 13.2 • R for Statistical Computing – v3.x
Data Mining Tools	• BI Office • Factor • Linguistic Inquiry and Word Count (LWC) – v1.17 Mac, v1.14 Windows • PolyAnalyst – v6.0 • PowerGREP – v4.6x • Qbase Data Transformer (QDT) • Qualitative Data Analysis (QDA) Miner

32

Data Warehouses	• Enterprise Elements (EE) • SnapWeb
Geospatial Tools	InVision Site
Point of Care Analytical	RAPID for RIICAM Software – v8.0
Unstructured Data/Natural Language Processing	• ATLAS.ti • Linguistic Inquiry and Word Count (LWC) – v1.17 Mac, v1.14 Windows • MAXQDA • Nuance Recognizer
Web Reporting Tools	• Business Objects • Crystal Reports – 2013 v14.1 • Datawatch Desktop • Google Analytics • IBM Business Monitor
Extract, Transform, Load (ETL)	• Exiftool • Infosphere DataStage – v11.3.x • Pentaho Data Integration – v5.0 • Qbase Data Transformer (QDT) • Scribe Insight
Columnar DBMS	N/A[26]
Data Quality Management	• AutoDelivery • DataFax • Exiftool • Freely Extensible Biomedical Record Linkage (FEBRL) • Informatica Data Quality – v9.6
Master Data Management	• Occupational Access System (OASYS) • Protege – v4.3 • Repliweb for Enterprise File Replication (EFR)
Non-Relational Database	FIS-GTM

---

[26] Although the TRM does not currently include any approved columnar DBMS tools, vendors of widely used (and TRM-approved) relational databases are now developing and releasing columnar DBMS products.

Object-Oriented DBMS	• Cache – 2012.1, 2012.2, 2014.1.3
	• Cache Management Portal SQL Interface - 2012.2, 2013.1, 2014.1
	• Cache Objects – 2012.2
Relational DBMS	• Microsoft SQL Server – 2012
	• Oracle Database – v12.1.x

4 USE CASES

The following use cases demonstrate the application of capabilities/recommendations described in this document.

4.1 Sharing Analytics Data with Different Circles of Trust

4.1.1 Purpose

This use case illustrates how a VA analytics services consumer (in this case, an LOB) applies the circles of trust concept in sharing analytics data. Unlike most Enterprise Design Pattern document use cases, it is a narrative rather than a step-by-step process.

4.1.2 Assumptions

- At the outset, all of the LOB's analytic data assets are uncategorized, so they are automatically placed into the Private circle of trust.
- The LOB's decisions regarding sharing of analytic data are correct and align with statutory, regulatory, and Department requirements.
- The LOB's analytic data assets may be data sets, individual data elements, or a combination of both.
- The VA Analytic Ecosystem:
  - Supports configurable sharing in a circles of trust model
  - Implements/enforces circles of trust designated by the LOB

4.1.3 Use Case Description

A VA LOB has recently moved all of its analytics data, tools, and operations into the VA Analytic Ecosystem. After ensuring that its operations in the Ecosystem are working as intended, it investigates the possibility of sharing some of its data with the rest of VA and even the public at large.

The LOB's data steward compiles a summary of the LOB's analytic data assets (A, B, C, D, E, and

F) and their suitability for sharing with different audiences:

Data Assets	Sharing Determination	Justification
A	Prohibited from sharing this.	Contains a lot of detailed information on individual Veterans and Beneficiaries, and qualifies as PII. Privacy laws prevent us from sharing it outside our own organization.
B	Don't want to share at this point in time.	Not subject to any legal or regulatory restrictions. However, it is associated with a research project that's still in the early stages, so it's not complete and probably not fully trustworthy.
C	Share this with the rest of VA, but not outside VA.	Another VA organization has requested access to this data for a project they're doing, and it could be useful to other parts of VA as well. We've already determined that it's safe to share with other parts of VA – just not with anyone else.
D	Required to share this with other organizations in VA, but not outside VA.	Operational data that we are required to submit to certain VA enterprise programs.
E	Want to share this with the Federal government and the public.	Research data that we have shared with Federal partners and VSOs in the past. The project leads have always wanted to share it with the general public as well. There are no security concerns preventing that.
F	Required to share this with the public.	Performance data that we are collecting to comply with a statute. That same statute will require us to share the performance data with the public starting next year.

Having made these determinations, the LOB places each data asset in the appropriate circle of trust, as shown in Figure 6. Every data asset that is shared with a less restrictive circle of trust (i.e., VA Internal and Public) is also shared with the more restrictive circle(s) of trust.

Private (LOB)
A, B, C, D, E, F

VA Internal
C, D, E, F

Public
E, F
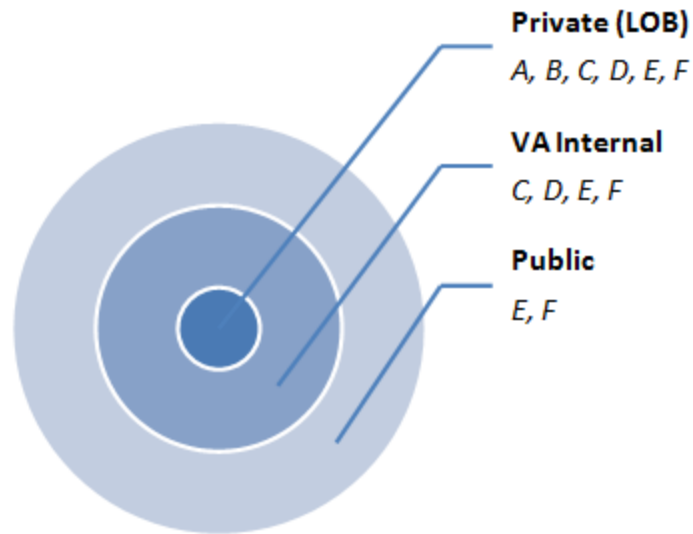
FIGURE 6: AN ILLUSTRATION OF DATA SHARING IN A CIRCLES OF TRUST MODEL

4.2 Event-Driven Special Clinical Process

4.2.1 Purpose

This use case elaborates on a scenario described by VHA analytics specialists in which advanced analytic capabilities are employed for clinical decision support and case management. The use case is triggered when a VA Clinician enters a diagnosis of hypertension in a patient record through eHMP. The system then generates a special clinical process with action items for the VA Clinician and the Patient. If the action items are not completed within a prescribed deadline, the system automatically generates alerts or follow-up workflows.

The capability for the system to initiate follow-up workflows for missed actions is critical for effective case management. Case managers are medical support staff who plan and coordinate healthcare services for patients with long-term or chronic conditions (including disabilities).[27] They help patients monitor their condition, make and keep doctors' appointments; obtain and take prescribed medications; and comply with doctors' recommendations and treatment plans.

Essentially, case managers help patients take the preventative measures necessary to prevent their chronic conditions from causing series health crises.

---

[27] Case managers are typically trained medical staff such as nurse practitioners, licensed practical nurses, or physician assistants/extenders.

Note that the use case describes a simplified version of a special clinical process for demonstration purposes only.

4.2.2   Assumptions

- VHA has formally designated case managers or some equivalent role
- The Patient:
    - Has received all of his medical services from VHA over the past 20 years
    - Uses VA mobile applications such as MyHealth*e*Vet
- The Clinical Decision Support System (CDSS) used by the Clinician and care team:
    - Is integrated with point-of-care applications that they use on a day-to-day basis
    - Supports near-real-time analytics that process a patient's newly entered clinical information along with his/her past history and demographic information
    - Provides explicit, transparent reasoning for recommended courses of action, either by surfacing them or displaying them with a click/mouse over
    - Employs a knowledge base and rules engine that are frequently updated/tuned with new information and best practices
- Clinicians and some members of the Patient's care team can modify action items on the clinical process or mark them as complete, but the Patient cannot

4.2.3   Use Case Description

1. During a wellness exam, a VA clinician discovers that her Patient has hypertension (chronic high blood pressure), which if left untreated can lead to cardiovascular disease. She enters the patient's diagnosis into eHMP.
2. The CDSS processes the new input in the context of:
    a. CDSS medical knowledge base and business rules engine
    b. Patient's medical history (e.g., weight, other conditions, other medications)
    c. Patient's demographic data (e.g., age, sex, ethnicity)
3. The CDSS returns a list of action items, which are immediately visible to the Clinician. A "patient view" of the action items is made available to the Patient through Web services and mobile applications. The action items are as follows:
    a. For the Clinician:
        i. Prescribe Patient a once-daily dose of a specific blood pressure medication appropriate to the Patient's age, weight, and ethnicity[28]

---

[28] Ethnicity is a factor in determining the appropriate blood pressure medication for a particular patient. People of African or Asian descent respond differently to some medications than people of Caucasian descent. The VA Clinician can view the reasoning behind this recommendation and even explain it to the patient.

ii. Refer Patient to a dietician who will help them develop a lifestyle/weight management plan to better manage their hypertension

iii. Review all action items with Patient, ensure that he understands what he has to do, and answer any questions the Patient has

b. For the Patient

i. Fill blood pressure medication prescription and take it once daily

ii. Schedule a follow-up appointment with Clinician within the next ten days for an evaluation and tests

iii. Attend scheduled follow-up appointment with the Clinician and provide blood and urine samples for tests

iv. Make an appointment with dietician referred by the Clinician within the next three weeks

4. The CDSS also assigns the Patient a Case Manager and forwards her the Patient's relevant information (including action items). The CDSS also gives the Case Manager a list of action items:

a. Call Patient in the next three days to see how he is doing and whether he is taking his medication as prescribed

b. Ensure that Patient schedules recommended follow-up appointments with Clinician and dietician

c. Call Patient before each scheduled appointment to remind him and confirm that he will attend

5. Patient, with reminders from his Case Manager, takes medication as prescribed and visits the Clinician for follow-up and tests. Clinician marks the related clinical process action items as complete.

6. Patient fails to make an appointment with a dietician within three weeks of the initial diagnosis. Missing this milestone triggers an alert to the designated Case Manager.

7. Prompted by the alert, the Case Manager calls the Patient to check in and remind them that they still have to make an appointment with the dietician. The Case Manager works with the Patient to schedule an appointment over the phone.

8. Within the next ten days, the Patient consults with a dietician to develop a diet and exercise plan that will help control the Patient's hypertension. The dietician accesses the patient's list of action items in the CDSS and marks the appropriate item complete.

4.2.4   Use Case Context Diagram

The cycle diagram in Figure 7 below illustrates how the CDSS facilitates case management for a patient with a chronic condition (such as hypertension, as described in the case study above). The Clinician enters patient data in the CDSS, which generates a list of health action items for all parties involved (i.e., Clinician, Patient, and the Patient's Case Manager). The health action

items inform action alerts, which are sent to the Case Manager. The Case Manager acts on these alerts by reaching out to the Patient, thus ensuring that the Patient attends follow-up appointments with one or more Clinicians. The Clinician then uses the CDSS to generate an updated list of action items, starting the cycle again.

FIGURE 7: HYPERTENSION CASE MANAGEMENT CYCLE FACILITATED BY CDSS

4.3 Streamlined Disability Benefits Application Process

4.3.1   Purpose

Veterans who have filed a claim for disability benefits must undergo an examination by a VHA clinician to determine the type, extent, and severity of his/her disability or disabilities. VBA claims processors evaluate the clinician's exam report against the schedule for rating disabilities[29] to determine the Veteran's disability rating (expressed as a percentage). The disability rating, along with other factors (i.e., the Veteran's age), are used to determine the amount of the Veteran's disability pension benefits.

A Veteran cannot schedule a disability examination until VBA claims processors have established that his/her disability is (or is likely to be) linked to a service-related injury or illness. VBA claims processors typically do not receive a clinician's disability exam report until 30-60

---

[29] 38 CFR Book C, Schedule for Rating Disabilities, available at http://www.benefits.va.gov/warms/bookc.asp.

days after the fact. The claims processors must review dozens of pages of medical records to determine what disabilities the Veteran has and establish the Veteran's disability rating.

A shared VA enterprise analytics infrastructure with text analysis capabilities may be used to streamline this process. In particular, these capabilities would:

- Allow the Veteran to initiate his/her disability claims process through VHA (with the assistance of a clinician), rather than "bouncing" between VBA and VHA
- Reduce the turnaround time for receipt of disability assessment reports once they are submitted by a VA Clinician
- Automatically calculate the Veteran's disability rating using text-based analytics supported by a knowledge base and business rule set consistent with 38 CFR Book C

4.3.2 Assumptions

- VHA and VBA are capable of seamlessly exchanging analytics data through the VA Analytic Ecosystem/CDW
    - o Both have an ongoing data use agreement in place
    - o VBA has a Disability Benefits enclave within the VA Analytic Ecosystem which is used for this process as well as other analytic operations
- VHA is equipped to help Veterans initiate the disability benefits claims process
    - o Clinicians can provide Veterans with a referral/recommendation for a disability assessment, without any involvement from VBA
    - o Clinicians can flag reports/results associated with disability assessments in a point-of-care or other clinical application
    - o VHA office staff and/or case managers can help Veterans submit an electronic VA Form 21-526 to a VBA Disability Claims Processing Application (DCPA)
- IAM services support the ability to associate a Veteran's Person Identifier (PID) with his existing record in the Master Veteran Index (MVI)[30]

4.3.3 Use Case Description

1. A Specialist performs an interview and examination with a Veteran who was referred to her by the Veteran's primary care doctor.

---

[30] The PID is one of the corresponding identifiers for a person's identity record in MVI. Specifically, it corresponds to their records in the CDW. The PID supports linking a record to a real person without retraining any of their personally identifiable information (PII) in the CDW itself.

a. Specialist finds that the Veteran suffers from limited mobility, chronic pain, and moderate peripheral neuropathy due to injuries below the knee.

b. Specialist recommends that the Veteran speak with a case manager about submitting an electronic VA Form 21-526 for disability compensation.

2. Following the appointment, the Veteran and Specialist make their respective submissions to VBA Claims Processing.

a. Specialist compiles and files an exam report on her examination of the Veteran, which includes:

   i. Description of and/or diagnostic codes for the Veteran's conditions/disabilities

   ii. A flag indicating that the exam report is a disability assessment and should be entered into the VBA disability claims process[31]

b. Veteran, with assistance from a case manager, submits an electronic disability claim to the DCPA. VBA claims processors can begin investigating the Veteran's service history upon receipt of the claim, while waiting for the exam report to be processed.

3. The exam record is transferred to the CDW along with other updates/changes to VHA's patient data from the same day.

4. A copy of the CDW is propagated to the VBA Disability Benefits enclave, where a scan of new/updated data picks up the "disability assessment" flag in the exam record.[32]

5. The flag triggers a special disability claims process (which may involve copying the data to a different enclave or sub-enclave) that is executed on the record. In this process:

a. Using text analytics, the record is evaluated against a knowledge base of terms/diagnostic codes corresponding to conditions/disabilities in 38 CFR Book C. Any terms or diagnostic codes that match something in the knowledge base are extracted from the record.

b. The extracted terms/codes are rendered down (i.e., normalized and de-duplicated) to produce a summary list of disabilities. For example, the textual reference to an injury and the diagnosis code(s) for the same injury are reduced to a single reference/item. The final list may have only one item (for a single disability) or multiple items (for multiple disabilities).[33]

---

[31] Flagging the record may be as simple as clicking a checkbox/radio button or selecting an item from a drop-down menu. Alternatively, flagging may require additional confirmation.

[32] Assuming that the flag is time stamped, it can be used to track metrics such as how many disability assessment reports are received in a particular period, the lag between entry into VHA systems and arrival in the VBA analytic enclave, the arrival gap between the assessment report and the corresponding VA Form 21-526, etc.

[33] This process can preserve modifiers such as "mild," "moderate," or "severe" that may affect the ratings for

c. The Veteran's list of disabilities is evaluated against a knowledge base of disabilities and their associated percentage ratings.

d. If there is only one disability on the final list, the rating for that disability is the Veteran's overall disability rating. Otherwise, the following steps are used to calculate a combined disability rating:

   i. Individual disabilities are ordered from most severe (highest rating) to least severe (lowest rating). The most severe disability will be #1 on the list, the next most severe will be #2, and so on.

   ii. Disabilities are evaluated against a combined disability rating table. Disability #1 is evaluated with Disability #2; their combined rating is evaluated with Disability #3, etc.

e. The original disability assessment report is bundled into a single data object together with:

   i. The extracted terms/codes

   ii. The de-duplicated summary list of disabilities with their individual percentage ratings

   iii. The Veteran's combined disability rating (if applicable), with a brief explanation of how the rating was calculated

6. The bundled report and extracts are placed into a long-term repository accessed by the DCPA (to which the Veteran submitted his disability claim). The arrival of the bundle prompts the DCPA to attempt to "match" it to the corresponding VA Form 21-526.

   a. If the Veteran's disability benefits claim has already been submitted to the application, it will be paired with the bundled report and extracts.

   b. If the Veteran's claim has not been submitted yet, the bundle will be flagged to indicate that it is awaiting a match. The next time a VA Form 21-526 is submitted to the DCPA, it will search the flagged bundles for the appropriate match.[34]

VBA Claims Processor handling the Veteran's case receives the disability rating. The Claims Processor can "drill down" to see how the rating was calculated and has the original disability assessment report available for verification and reference.

---

certain conditions. For example, "mild mobility impairment" may have a lower rating than "severe mobility impairment," so in that case the analytic process should distinguish between degrees of disability.

[34] This is to ensure that a given Veteran's claim and exam/disability rating are promptly bundled together regardless of which one reaches the DCPA first.

APPENDIX A. SCOPE

This Enterprise Design Pattern describes an enterprise "big data" and analytics capability to streamline data collection, support configurable sharing, and maximize the value of VA's business intelligence. Topics include:

- Governance bodies for VA Enterprise analytics
- Analytics governance fundamentals
    - The circles of trust model
    - Provider and consumer responsibilities
    - Minimal data quality standard
- Integrating VA data sources (including warehouses) with the VA Analytic Ecosystem
    - Consolidating analytics capabilities
    - Using staging areas for refining ingested data
- Evaluating, selecting, and using analytics technologies
- Analytics capabilities in high demand

The following concepts are outside the scope of this design document:

- Requirements and specifications for conceptual, logical, and physical data models used in the analytic environment
- Specific mechanisms for securing the analytic environment[35]
    - Logging and auditing
    - Data messaging security and authenticity
    - Access control and authorization decisions
- Detailed guidance and/or requirements for:
    - Architecting and applying next-generation analytics technologies (e.g., streaming analytics, machine learning)
    - Selecting appropriate data storage types for a particular use case
- Specifics of applications and services that will support analytics
- Infrastructure and hardware design specifications
- Vendor-specific technologies

---

[35] These mechanisms (e.g., access control, authorization, and encryption) are addressed in other Enterprise Design Patterns that apply to the VA EA data layer and/or the VA EA at large.

Intended Audience

The primary audience for this document consists of VA stakeholders who manage and/or conduct analytics activities on behalf of their organization (e.g., office, program, LOB). Specifically, these stakeholders are:

- System and application owners/stewards
- Data analysts, data scientists, and statisticians

This document is also intended for those in leadership roles who can establish governance mechanisms and policies related to analytics.

Document Development and Maintenance

This document was developed collaboratively with internal stakeholders from across the Department and included participation from VA OI&T, Product Development (PD), Office of Information Security (OIS), Architecture, Strategy and Design (ASD), and Service Delivery and Engineering (SDE). Extensive input and participation was also received from VHA, VBA and the National Cemetery Administration (NCA). 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 are coordinated with the Government lead for this document, which will also facilitate stakeholder coordination and subsequent re-approval depending on the significance of the change.

APPENDIX B. CONCEPTS FOR SERVICE-ORIENTED ANALYTICS

This appendix describes ways of thinking about analytics that facilitate building a multi-stage, modular architecture for data analytics in which:

- Data flows can be re-routed and re-sequenced through different modules/components as needed
- Interoperable components can be modified and/or replaced with little or no downstream impact
- Individual components can be easily repurposed for different analytic processes

The concepts described below are referenced in Sections 3.3 and 3.4 of the Enterprise Design Pattern.

Data Temperature

Data temperature is a critical consideration in designing analytic processes and selecting appropriate technologies, particularly storage technologies. The temperature of data – hot, warm, or cold – is the combination of the following data characteristics:

- **Latency:** Amount of time for a query to be executed (milliseconds, seconds, minutes, hours)
- **Volume:** Total quantity of stored data used in each analytic processing operation (MB, GB, TB, PB)
- **Item Size:** Size of individual data elements or data sets ingested into the analytic system in a given instance (B, KB, MB, TB)
- **Request Rate:** Frequency at which analysis of data is performed (Low, Medium, High, Very High)
- **Durability:** Degree to which committed data transactions will survive permanently (Low, Medium, High, Very High)

Data structure does not factor into data temperature. Table 6 illustrates different data temperatures based on combinations of relevant data characteristics.

TABLE 6: DATA TEMPERATURES

	Hot	Warm	Cold
Average Latency	milliseconds	milliseconds, seconds, minutes	hours, days

Data Volume	GB-TB	GB-PB	GB-PB
Item Size	B-MB	KB-TB	GB-TB
Request Rate	Very High	Low-Very High	Very Low
Durability	Low-Very High	High-Very High	Very High

"Hot" data is typically associated with high-speed, low-volume stream analytics where the window of opportunity to take action on information gleaned from the data is very small. "Cold" data is typically associated with analytics involving historical or archival data, where the accuracy of information derived from analysis is important, but the speed of analysis is not. The cost per unit for hot data storage tends to be high, while the cost per unit for cold data tends to be low.

The temperature of data is determined with respect to a particular use case and not the source(s) of the data itself. Data from one source may be employed in multiple use cases, each with a different temperature. For example, data ingested from a medical monitoring device has a different temperature with respect to each of the following use cases:

- **Hot:** Alerting clinicians to potentially dangerous changes in a patient's vital signs, so they can respond quickly.
- **Warm:** Tracking changes in a patient's condition over the course of days of days or a week to evaluate how they are responding to medication.
- **Cold:** Tracking long-term outcomes for similar populations with similar conditions.

To support these use cases, the data from the medical monitoring device will be replicated to three distinct routes for storage, analysis, and delivery.

Stages in an Analytic Data Flow

While analytic data flows differ depending on the characteristics of the data type and use case, they consist of five basic stages, as illustrated in Figure 8.

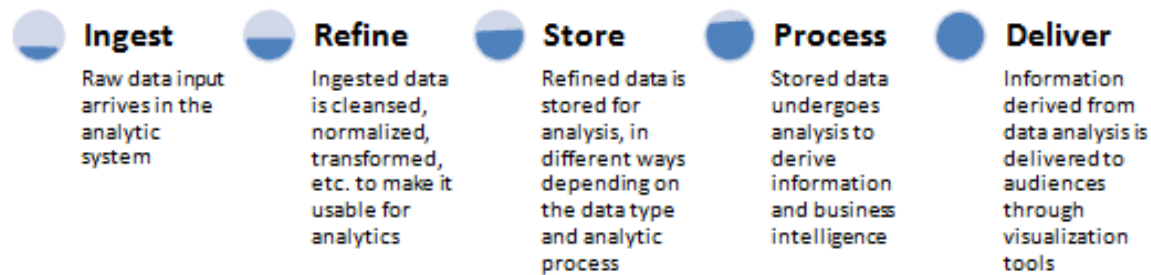| Ingest	Refine	Store	Process	Deliver
Raw data input arrives in the analytic system	Ingested data is cleansed, normalized, transformed, etc. to make it usable for analytics	Refined data is stored for analysis, in different ways depending on the data type and analytic process	Stored data undergoes analysis to derive information and business intelligence	Information derived from data analysis is delivered to audiences through visualization tools

FIGURE 8: THE FIVE STAGES OF AN ANALYTIC DATA FLOW IN SEQUENCE

Data temperature is a critical consideration at each stage of the process, and is the primary factor in determining the amount of time between ingest and delivery. Some data flows do not always move sequentially through the five stages. As shown in Figure 9, it is possible to cycle through certain stages more than once instead of proceeding through them in a linear fashion.
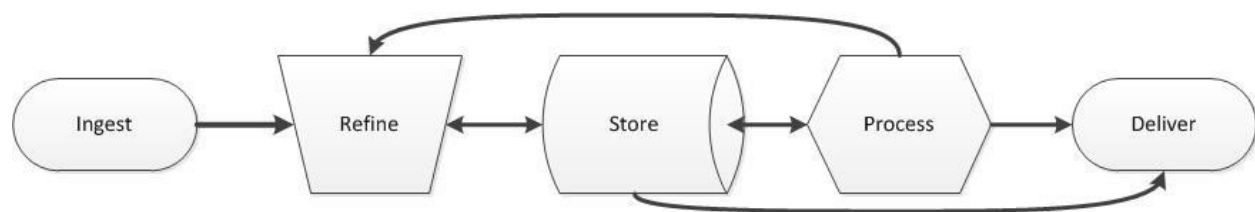

FIGURE 9: POTENTIAL PATHS FOR NON-SEQUENTIAL DATA FLOWS

Ingest

Although data comes in many different formats, there are essentially three different types of data ingest:

- Transactional (individual database reads/writes)
- Files (batch transfers, logs, objects)
- Stream (sensor data, mobile)

The most appropriate type of ingest for each use case depends on the source or producer of the data, and has a direct bearing on later stages of analytics processing, particularly storage.

BISL refers to raw ingested data as "Tier 1" data.

Refine

While some analytic processes use raw ingested data, most do not. Some type and degree of work is usually required to close the gap between data in its raw "as-is" state upon ingestion to

the desired "to-be" state for storage and analysis. Depending on the type of ingests, data temperature, quality requirements, and use case characteristics, refining may include one or more of the following operations:

- Stream sequencing
- Tagging with metadata
- Transformation
- Master data management
- Cleansing
- Normalization/de-duplication
- Integration, correlation, and/or linking with other data sets
- Replication to multiple data stores

These capabilities may also be used to change, enhance, tag, or link data that has already been stored for use by analytic processes. Some possible operations of this type are:

- Linking elements of one dataset with elements of another dataset
- Applying metadata tags to and/or indexing stored data
- Replicating and transforming data stored for one analytic process to make it usable for a different analytic process
- Bringing stored data into compliance with new requirements regarding format, field values, etc.
- Removing certain elements from a dataset in order to reduce its sensitivity level and make it suitable for access by a wider audience
- Propagating new datasets produced by analysis into a variety of different data stores for use by future analytic operations

BISL refers to refined data as "Tier 2" data.

Store

Once ingested data has been refined, it can be stored for later analysis (how much later depends on the data temperature). Data storage for a particular use case depends upon data temperature, structure, and query types.

New datasets produced by analysis may also be stored for use by future analytic operations and/or delivered to certain parties for their own analysis

Process

At this point, data undergoes processing to derive useful insights and business intelligence. The results of analysis may be fed back into the system as ingested data for further analysis or correlation with other data.

The results of analysis may be analytic products delivered to end users, or new datasets that are themselves stored for use in other analytic processes.

Deliver

Delivery typically means delivering analytic products to the intended audience – not analysts and data scientists, but people who act on information derived through the work of analysts and data scientists. Information may be delivered through reports (as in traditional analytics), dynamic real-time displays, alerts, or applications. Stored data itself may also be delivered to third parties for use in their own analytic operations.

Whether the object of delivery is a report, visualization, or stored data, part of this stage may involve advertising that the object is available. Delivering and advertising items (i.e., data or data products) may involve one or more of the following:

- Publishing items to a library, catalog, or inventory
- Tagging and/or indexing items so they can be located with a search function
- Pushing out alerts to subscribers/interested parties when new items become available

APPENDIX C. DEFINITIONS

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

Key Term	Definition
Data Flow	Describes the lifecycle and movement of data in an analytic system with respect to a particular process or use case. A data flow begins with collection/ingestion from data sources and ends with the presentation of information extracted from the data using reports, visualization tools, applications, etc.
Enterprise Create, Read, Update, Delete (eCRUD)	The eCRUD service was initially created as part of the Veteran Lifetime Electronic Record (VLER) Data Access Service (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.
Ingest, Ingesting, or Ingestion	The entry of raw data input into the analytic system from data sources, to include applications, operational data stores, feeds, sensors, etc.
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.
Master Veterans Index (MVI)	VA ADS for the identity information of all VA persons of interest, including Veterans, beneficiaries, employees, and contractors. Each identity is cross-referenced with records from the DoD Defense Enrollment Eligibility Reporting System (DEERS) database.

Key Term	Definition
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.
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.
Refine or Refining	Essentially similar to refining raw materials for use in a physical manufacturing process, refining refers to converting raw data into a form suitable for storage and subsequent use by analytic processes. Depending on the characteristics of the data and the use case, refining may involve some combination of transformation, sequencing, tagging/annotation (with metadata), normalization, cleansing, de-duplication, and correlation with other data.

APPENDIX D. ACRONYMS

The following table provides a list of acronyms that are applicable to and used within this document.

Acronym	Description
AA&A	Authentication, Authorization, and Access
ADS	Authoritative Data Source
AN	Analytics and Informatics
ASD	Architecture, Strategy and Design
BI	Business Intelligence
BIRLS	Beneficiary Identification Records Locator System
CDI	Customer Data Integration
CDSS	Clinical Decision Support System
CDW	Corporate Data Warehouse
DAR	Data Architecture Repository
DAS	Data Access Service
DCPA	Disability Claims Processing Application
DEERS	Defense Enrollment Eligibility Reporting System
DGC	Data Governance Council
DoD	Department of Defense
EA	Enterprise Architecture
eHMP	Electronic Health Management Platform
EHR	Electronic Health Record
eMI	Enterprise Messaging Infrastructure
EPMO	Enterprise Program Management Office
ERP	Enterprise Resource Planning
ESS	Enterprise Shared Services
ETSP	Enterprise Technology Strategic Plan
FR	Field Reporting
GP	General Purpose
HHS	Department of Health and Human Services
HSRD	Health Services Research and Development
IAM	Identity and Access Management
IoT	Internet of Things
IPT	Integrated Project Team
IT	Information Technology
LOB	Line of Business
MVI	Master Veteran Index
NCA	National Cemetery Administration

Acronym	Description
NLP	Natural Language Processing
NoSQL	Not Only SQL
OGC	Office of General Counsel
OI&T	Office of Information and Technology
OIS	Office of Information Security
PHI	Protected Health Information
PID	Person Identifier
PII	Personally Identifiable Information
POC	Point of Contact
RD	Health Services R&D
RDW	Regional Data Warehouse
SOA	Service-Oriented Architecture
SQL	Structured Query Language
SSA	Social Security Administration
TRM	Technical Reference Model
VADI	VA Data Inventory
VBA	Veteran Benefits Administration
VHA	Veteran Health Administration
VINCI	Veterans Informatics and Computing Infrastructure
VistA	Veterans Health Information Systems and Technology Architecture
VLER	Veteran Lifetime Electronic Record
VSA	VistA Service Assembler
VSO	Veteran Service Organization

APPENDIX E. REFERENCES, STANDARDS, AND POLICIES

This EDP 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 Enterprise Technical Architecture (ETA):

#	Issuing Agency	Applicable Reference/Standard	Purpose
1	VA OIS	VA 6500 Handbook	Directive from the OI&T OIS for establishment of an information security program in VA, which applies to all applications that leverage ESS.
2	OI&T BISL	Corporate Data Warehouse (CDW) Update	Describes the VA Analytic Ecosystem, VINCI, and their positive impact on VHA's healthcare and administrative operations.
3	VBA	38 CFR Book C, Schedule for Rating Disabilities	Policy/procedures for assigning disability ratings to Veterans for their service-connected injuries and/or illnesses. These disability ratings, in combination with other factors, are used to determine an appropriate disability pension for the affected Veteran.
4		VA Data Inventory (VADI)	VADI is the authoritative source for VA Data Store metadata. VADI allows users to search for and navigate through VA Data Store metadata.
5		Data Architecture Repository (DAR)	DAR provides a means to catalog, search, report, and manage VA metadata via a web-accessible portal.