Air Force Chief Data Office (SAF/CO)

Data Services Reference Architecture

March 2019

U.S. AIR FORCE

EXECUTIVE SUMMARY

The Air Force faces a rapidly evolving technological landscape, marked by the growing power of interconnected systems. Big data analytics are transforming societies and economies, and expanding the power of information and knowledge. Future warfare will respond to these advances, and provide unparalleled advantages to militaries that can gather, share, and exploit vast streams of rich data. To compete against great power adversaries, the Air Force must harness the power of this data revolution and develop the underlying structures to dominate a new and emerging character of war. Our “Digital Air Force” initiative drives purposefully toward this future and, with this Data Services Reference Architecture, outlines the path to a more lethal, innovative, and connected force.

Today, the Air Force’s data rests in legacy systems connected by an antiquated and brittle architecture. Rather than freely push information to the users who need it, these systems employ “data jails” to stovepipe and restrict access. Not only are these systems costly to maintain, they actively inhibit timely sharing of data and frustrate the analytic needs of our warfighters. Instead, we must become a data-driven organization that prioritizes our information ecosystem. To do so, we must institute policies and procedures that make our data Secure, Visible, Accessible, Understandable, Linked, and Trusted (SVAULT). This document outlines our approach to implement these SVAULT principles and prepare our force for a new model of business operations and high-velocity, networked warfighting.

I expect commanders and materiel leaders at all levels to implement this guidance and comply with these principles. It is only by working together, throughout the force, that we will ensure all Airmen have uninterrupted access to the data they need, where and when they need it. This connectivity is not a luxury or nicety, but a fundamental necessity for warfighting dominance. This Data Services Reference Architecture forms the essential foundation for this dominance, and initiates the critical reforms we need to compete, deter, and win over any future adversary.

Matthew P. Donovan
Under Secretarty of the Air Force

TABLE OF CONTENTS

FIGURES

TABLES

1 INTRODUCTION

1.1 Methodology

The objective of this Reference Architecture document is to provide clear guidance for the design, development, implementation, and use of Air Force Major Commands (MAJCOM)/ Functional Data Platforms. A Reference Architecture is an organizational asset that provides common language for the various stakeholders, provides consistency of implementation of technology to solve problems, supports the validation of proposed solutions, and encourages adherence to common standards, specifications, and patterns. It also serves as an authoritative about a specific subject area that guides and inform the instantiations of multiple architectures and solutions.

This Air Force Data Services Reference Architecture is below the Enterprise Reference Architecture level and crosses mission areas and portfolios. It is intended to demonstrate a capability-oriented architecture and support the implementation of diverse Solution Architectures for scalable data management and for data and analytics as service capabilities.

The document is organized into the following major sections:

- **Mission Need:** describes the Air Force business need(s) that the Air Force Data Services Reference Architecture is intended to address.
- **Strategic Purpose:** outlines the strategic purpose of the Air Force Data Services Reference Architecture described in this document.
- **Technical Positions and Principles:** describes the technical principles that are recommended for specific designs and implementations of this Reference Architecture for a business need.
- **Technical Patterns and Templates:** contains Functional architectural patterns to guide design, implementation, validation, and verification of solution implementations.
- **Use of the Reference Architecture Solution:** describes implementation and workflow patterns for the application of the general Reference Architecture to the business needs.
- **Documentation Patterns:** describes minimum effective documentation within specific solution implementations of the architecture.
- **Workflow Implementation Template:** contains a notional process for design, description, and specification of configured workflows within a solution implementation of the Reference Architecture.
- **References and Supporting Information:** contains standards used throughout the document, standard language and reference terminology for use in understanding the Reference Architecture.

2 MISSION NEED

The United States Air Force has delivered unmatched capabilities in its core missions. We have evolved from an Air Force of bombers, fighters, and airlifts to a juggernaut that operates below the surface of the earth to the highest orbits of space. As we have evolved, so have our

adversaries as well as threats to our environment. To prevail in this new competitive era, we must instead recognize that success requires the Air Force to maximize its ability to sense, share, synthesize and act upon data. We must leverage information across all domains, at all levels of operations, and synchronize with our Joint Forces, strategic allies, and coalition partners to deliver military advantage at a rate and scale that outpaces our adversaries. In short, we must value data as a strategic asset and view it as a force multiplier. We must make Air Force data visible, accessible, understandable, linked, and trusted which permits it to be used to enable mission execution in this quickly changing, demanding environment.

3 STRATEGIC PURPOSE

The purpose of this document is to identify the common practices, principles, and technical standards that support the implementation of the Air Force Chief Data Office Principles and Objectives. This Reference Architecture serves as a guide for Air Force MAJCOM and Functional communities as they develop data services platforms. It also enables the Air Force's transformation into a data-driven organization. Data services platforms are specific solutions that deliver data management and analytic capabilities. By adhering to these standards, both metadata and data will be made more visible, accessible, understandable, linked and trusted. Data jails will be avoided and instead, data will be shared across platforms. The data's trustworthiness on different platforms will be exposed and the ability to link data together to support Use Case implementations will be increased.

4 TECHNICAL POSITIONS AND PRINCIPLES

4.1 Air Force Chief Data Office Principles and Objectives

The Air Force Data Services Reference Architecture is intended to reflect the Air Force Chief Data Office's (SAF/CO) key guiding principles. This Reference Architecture, including design and development principles and technical templates and patterns, is intended to reflect these core values:

- **Collective Data Ownership:** Data is a strategic enterprise asset collectively owned by the Air Force and shared with other Services, Agencies, Allies and strategic partners. Air Force data must be managed and leveraged in combination with other non-Air Force data (including publicly available information) wherever necessary to enable well-informed decisions to ensure mission success. This must be done in a manner that assures appropriate compliance safeguards (such as privacy and security) are maintained throughout all phases of the data lifecycle.
- **Data Collection:** As data is collected, it must be transformed, harmonized, and provisioned in a way that enables maximum value to be derived from it. This includes the application of advanced analytic techniques including machine learning and other forms of artificial intelligence that enable predictive or prescriptive analyses or perform routine tasks that free up humans to make informed judgements on the analyses presented. As such, the Air Force will increase the collection, tagging and preservation of data (and metadata) from a variety of formats and ensure it is available and useable for sharing and analytics.
- **Enterprise Wide Data Access and Availability:** Data, information, and knowledge are

currently managed in stovepipes across the Air Force. This precludes an enterprise view. Enterprise-wide data access and availability will be considered throughout the data and systems lifecycle. While data architectures may be adjusted within specific functional communities or Air Force components to meet specific needs, architectures will support and enable enterprise-wide data availability by ensuring discoverability, accessibility and usability of data.

- **Data Sharing:** Users must be able to leverage data across multiple systems and domains to derive insights and identify options for optimizing performance in both warfighting and business operations. In compliance with the Air Force Chief Data Office Principles and Objectives, this Reference Architecture follows an "open first" design approach. This approach establishes free and open source (FOSS) tools (defined as public access to source code and unrestrictive licenses) tools as the preferred option, to provide a vendor-agnostic, cost effective, scalable suite of capabilities.
- **Data Fit for Purpose:** To be "fit for purpose," data must be reliable, consistent, and timely. Exposing data directly from its authoritative source with clear and complete metadata included with it, speeds up the ability to understand and leverage that data for a specific purpose. Ensuring that data is fit for purpose reinforces all other guiding principles.

The Air Force Chief Data Office Principles and Objectives identify key functional capabilities fit for any solution and intend to help enable the goals listed above. The key enabling capabilities are listed below:

- **Data Visibility:** Data cannot be leveraged if no one knows it exists. Authoritative data is particularly important because it is, by definition, the most valid, trusted source data that exists for any element of interest. The first goal of the Air Force Chief Data Office is to identify, register and expose all authoritative data in a way that makes it easily discoverable by Airmen and/or machines across the enterprise.
- **Data Accessibility:** Data access is often limited to a small number of credentialed users responsible for the data on a day to day basis. Access limitations are typically a result of concerns about protecting data from compromise, mis-use or mis-interpretation. While these concerns are valid, the Air Force will establish protected means and mechanisms for all credentialed users to have timely authorized access to the right data when and where it is needed. This includes access to meaningful metadata that clarifies the historical contextual meaning of the data.
- **Data Understandability:** Data must be put into context to establish its meaning. Air Force data will be described with clearly defined dictionaries, glossaries, and ontologies to provide a more holistic understanding of the lineage and meaning of each data element, and the relationships to other data elements. This includes establishing a governed set of standards for tagging and querying data, and ensuring that upon creation, metadata contains mission essential security markings, access control claims, and associated record management disposition.
- **Data Linkages and Explicit Relationships:** Data will have increased operational relevance and value when it can be linked to other data to derive deeper insights and understanding. Data relationships and dependencies of data will be established, and data standards will be codified allowing data to be correlated, shared, and used across functional communities and operational domains, when permissible by law.

- **Data Trustworthiness:** Data must be trusted to deliver value in a high-speed, multi-domain operating environment. The Air Force will establish ways to measure, record and protect the veracity and reliability of data at the source. Data will be validated and de-conflicted to establish the level of confidence that can be placed on it. Data timelines will be measured and captured in metadata to establish context, validity, and maturation over time.

4.2 Principles of Solution Design and Development

To achieve the Air Force Chief Data Office's principles and objectives, this Reference Architecture contains principles related to platform implementation and use. These principles can be used to ensure that specific solution implementations of this Reference Architecture achieve the goals and principles of the Chief Data Office.

4.2.1 SOLUTION DESIGN PRINCIPLES

These design principles are applied throughout the Reference Architecture to promote design patterns that support consistent, flexible, useful, and affordable implemented solutions. Applying these principles reduces cost and risk and increases utility and reusability.

- **Make Discoverable** – implemented solutions enable the discovery of data, data products, and analytics across multiple implemented solution instances regardless of where they are hosted.
- **Use Common Language** – consistent use of syntax and terminology (if not necessarily naming promotes discoverability and interoperability across disparate implemented instances.
- **Use Distributed Services and Microservices** – applying a service architecture approach to implementation that enables the ability to share data and analytics across multiple platforms and implementation instances of the Reference Architecture and supports horizontal scaling models.
- **Standardize Interface Configurations** – provide patterns and lexicons for service publication and discovery to ensure consistency. Provide tools and assistance to aid adherence.
- **Standardize Workflow Pattern Configurations** – provide validated patterns, templates, and metadata syntax for analytics and data product design and use to aid in user understanding across implementations of this Reference Architecture.
- **Decentralized, Organic Architecture** – minimize centralized approvals and roadblocks for solution scaling, deployment, and use. Use central registries for service and product publication and discovery while reducing dependence on administrative processes.
- **Design for Evolution, Extensibility and Scalability** – design and implement service architectures that scale up and down quickly, are intended to be reconfigurable, and separate service interfaces form service execution. Enables components to be replaced and scaled quickly and independently: This supports multiple data products, analytics, and users. The original configuration is designed to support future enhancements without significant refactoring.

- **Technology independence** – focus on Air Force needs and capability rather than tools. Demonstrate multiple solution implementations using multiple platforms and tools, which reduces dependence on vendors and specific technologies.
- **Use Governance to Support Capability** – establish communities of Data Stewards and Users to collaborate on curating data, administering metadata and access control policies, managing user credentials and access requests, evaluating data products for fitness and authoritativeness; and driving the capability roadmap. Governance focuses on the quality of products in the solution implementation and prioritizing capability rather than acting as an approval authority.

4.2.2 TECHNOLOGY SELECTION PRINCIPLES

In evaluating and selecting technologies for use in a solution implementation of the Air Force Data Services Reference Architecture, it is critical to identify components that maximize flexibility, scalability, and availability while also maintaining extensibility and minimizing the maintenance and support burden. As such, this Reference Architecture recommends that specific solution implementations strive to adhere to key technology selection principles:

- **Agility** – Identify and select technologies that are open, extensible and lend themselves to implementation and configuration using an Agile development approach.
- **Prioritize Open, Modern Solutions** – Prioritize services that are cloud native to maximize scalability and minimize support burden. Where these are not available, prioritize the use of FOSS tools wherever possible, defined as public access to source code and unrestrictive licenses. If no other compliant option is available, use proprietary software.
- **Minimize Dependencies on Proprietary and Vendor Technologies** – Design an architecture that supports the rapid evolution and replacement of specific services to minimize dependencies on any one product or company.
- **Minimal Impact to Mission** – Architect a solution implementation for easy adoption of new tools and services by minimizing their impact on the underlying data and existing mission systems.
- **Portability** – Design a solution implementation that can easily use both on-premise and offpremise services and be quickly and cleanly ported between them.

4.2.3 DATA MANAGEMENT PRINCIPLES

Foundational to the Air Force Data Services Reference Architecture is the application of Data Management principles through the Reference Architecture – such that Data Management is implicit in the functionality of the system and within individual platform implementations rather than conducted as an external activity as an afterthought. Scalable data management in an Air Force Data Services Reference Architecture solution implementation seeks to embody the following guiding principles:

- **Source Independence:** Critical business processes can be executed across functional domains and systems with secured, timely, accurate and relevant data regardless of the source or location of the data.
- **Common Lexicons and Metadata:** Enterprise data assets, data products, and analytics are published, exposed, and consumed based on common or semantically compatible

vocabularies to the greatest degree possible using a common descriptive model.

- **Reduce Replication:** Single purpose copies of datasets are reduced except in the context of their authoritative source and for documented business purposes. Reduce the proliferation of point to-point interfaces. When replicated data does exist, the business purpose, pedigree and lineage, and modifications are fully documented.
- **Central Metadata Registry:** Enterprise metadata is centralized, governed, and integrated at all levels to enable consumers to discover authoritative data and fully understand the business, operational and technical dimensions of enterprise data assets while also being able to preserve their own domain specific language.
- **Leverage Reference Architectures:** A coherent, fully integrated suite of technology enablers and design patterns are employed in a consistent manner to construct all aspects Enterprise Data Management in a cost-effective manner.
- **Leverage Shared Services:** Enterprise data assets are exposed in a centralized self-service, multitenant analytics and decision support platform in which descriptions of the available data is exposed through the enterprise metadata management repository.

4.2.4 SOLUTION DEVELOPMENT PRINCIPLES

Solution development principles are intended to be used during the platform implementation of the Reference Architecture to ensure the overall quality, usability, and maintainability of a given solution implementation of the Reference Architecture. When implemented, consider the following principles:

- **Provide User Configurable Workflows:** Different missions, organizations and users work differently and systems that try to impose uniformity are unpopular and often unused. Consistency is critical for data integrity and data reuse but does not need to drive how users interact with the system or create workflows. Allow for user configurable interfaces and workflows.
- **Interoperable Metadata Model and Interface Architecture:** Implement the common metadata model and interface architecture to support downstream interoperability by allowing data to be mapped to a fully normalized concept model for discovery and operations. This model does not rely on developing new standards and lexicons, but rather a more thoughtful implementation of existing standards and lexicons.
- **Microservices-based Evolutionary Architecture:** Air Force Data Services Reference Architecture breaks away from classical monolithic systems architecture models. This Reference Architecture outlines system organization that is hierarchically and functionally decomposed to minimum viable units that can be deployed as microservices using open interfaces with standard interface and metadata configurations. Logically separate the service interfaces and the execution of the services - design services and interfaces with change in mind. Best practices for unit testing, regression testing, and scaling profiles account for Air Force Data Services Reference Architecture system services being changed, replaced, modernized and upgraded – and that these activities not affect system performance.
- **Test-Driven Development and Operations:** Platform implementations of this Reference Architecture incorporate Test-Driven Design and Development and Operations (DevOps)

principles. A guiding principle is to first create a test module/procedure for all functional components, and this be used to continuously validate modifications and changes. Source code control and configuration management are integrated with continuous integration, deployment/container orchestration and deployment automation.

- **Interface with Existing Legacy Systems:** The Air Force Data Services Reference Architecture allows for integration and interoperability with existing Air Force legacy systems through interface extension and enablement. Create a data and analytics workflow system that implements data management and does not necessitate replacing or substantially modifying existing systems. Solution implementations support and enable the retrofitting of legacy systems with compatible, standardized interfaces for integration into the Air Force Data Services Reference Architecture system.
- **Automate Documentation:** Integrate DevOps and configuration management with systems metadata to create a system that is largely self-documenting, ensuring the appropriate documentation of services, data assets, and data products using the Metadata model: identifying stewards, relevant metadata, and supporting the stand up of governance and management of access control while also minimizing the burden of this process by leveraging Air Force Data Services Reference Architecture automation capabilities.

4.2.5 METADATA PRINCIPLES

Core to the Air Force Data Services Reference Architecture is the implementation of an Air Force Enterprise Information Model (EIM) which acts as the core registry for the data and analytics assets managed in the Air Force Data Services Reference Architecture. The EIM is critical to the Air Force Data Services Reference Architecture, as it allows users and applications to publish and discover data and analytics, provide technical information necessary for access and consumption, and provide contextual metadata necessary to ascertain fitness for purpose. The EIM consist of two core components, each of which is enabled through functional services:

- **Metadata Model:** The core metadata model used to store and extend the resource (data or analytical product) metadata
- **Metadata Registry:** The physical storage of the metadata that complies with the metadata model

Together, these components create the Metadata Catalog. To be effective, each asset record (data or analytics service) contains the following types of metadata, to the greatest level of detail as predetermined by the CDO for each asset/asset class.

- **Technical Metadata:** The metadata pertaining to technical access to a resource, including format, structure, type, location, configuration, volume, primary keys, necessary libraries for processing, and necessary system variables.
- **Administrative Metadata:** The metadata pertaining to stewardship, administrative access to a resource, including dissemination/access control, data lifecycle stage, and pedigree and lineage.
- **Operational/Contextual (Business) Metadata:** The metadata pertaining to the application context of the data in the resource, including standardized descriptive language using a business lexicon, information about unique identifiers or attributes, linkages to documentation, and the ability for users to “meta tag” resources.

To judge the overall efficacy and maturity of a Metadata Catalog in an Air Force Data Services Reference Architecture implementation, it is evaluated for enabling the following activities:

- **Documentation of Pedigree & Lineage and Discovery of Authoritative Source:** Allow a consumer to understand the data context and make an informed judgment about quality, timeliness, and fitness for purpose.
- **Contextual Understanding of the Asset:** Support the application of quality, risk, and trust constructs to the asset and enable appropriate use by the end consumer, including minimizing the need to discover additional metadata.
- **Collect a Minimum Set:** Balance the burden on implementers while still providing potential consumers with all necessary descriptive information.
- **Demonstrate a Flexible and Easily Extendible Model:** Support adding additional elements as needed without breaking the core structure.
- **Support Conformance to Standards and Enterprise Vocabulary:** Utilize, wherever possible, existing reference models and demonstrate the ability to extend through the addition of and mapping to additional models and dictionaries.
- **Demonstrate Openness:** Support transparency of assets including both machine and human readable formats.
- **Support Necessary Access Restrictions:** If needed, control the release of information to credentialed, appropriate consumers.

4.3 Principles of Architectural Organization

This Reference Architecture uses a set of organizational approaches intended to manage the balance between ensuring completeness and quality of user capability and experience while also minimizing the total number and complexity of base components, which improves the reliability, maintainability, and cost profile of any solution implementation.

- **Hierarchy** - The Air Force Data Services Reference Architecture is deliberately a modular one; designed to focus on functional principles describing the use of existing tools wherever possible that can be deployed when ready and re-used in extensible workflows to support multiple current and future programmatic needs. The Air Force Data Services Reference Architecture is therefore hierarchical: it is grouped into Capability Layers that are further broken out into Functional groups and segmented by a system boundary.
- **Capability Layers and Functional groups** - Capability Layers represent groups of logically similar functions that serve a common purpose and are managed as a group – these also equate to a similar set of user activities and skills for staffing and development. Within a Capability Layer, there are multiple Functional groups. A Functional group represents a logical group of functions or operations within the Reference Architecture that corresponds to a set of discrete modules, applications, or software products requiring specific configuration, management and close interoperation to deliver the overall functionality of the Functional group.
- **Value-Added Services and Microservices Views –** In order to effectively describe and translate across uses of the system and enabling components, this Reference Architecture consists of two views: A value-added Services view describing base reference

configurations of system workflows that track to specific, necessary user activities, and a Microservices view that describes the functional components necessary to enact the Value-Added Services configuration.

4.4 Implementation Templates and Design Patterns

When creating a Platform Implementation of the Air Force Data Services Reference Architecture, to minimize technical and performance risk, maximize user satisfaction, and ensure consistent development, deployment, and configuration of a platform, the Functionals and MAJCOMS that implement this document will develop standard templates and patterns to support development, deployment, and configuration of data assets, analytic assets, data products, and Use Cases.

This Reference Architecture includes Use Case templates intended to provide a standard operating procedure for Data Scientists and Business Analysts. The intent is to provide clear instructions and configuration guidance to allow Data Scientists and Business Analysts to use the modular data, analytics, and Business Intelligence Functions of a platform implementation to create multiple Use Cases using the same reusable components.

This Reference Architecture also includes: design patterns for metadata model, catalog, and data architectures within Data Operations Analytics Framework and Data Lake Storage Capability Functional groups and for the configuration and deployment of data resources and analytics in the solution in a manner consistent with Data Management and Metadata principles and the Data Lifecycle. These design patterns allow Data Scientists and Developers to rapidly reuse existing objects in the system in a manner consistent with overall design.

4.5 Constraining Criteria

One of the objectives of Air Force Data Services Reference Architecture is to ensure that the system meets Air Force programmatic and business needs with the minimum amount of disruption to existing operations, while keeping system extensions to a minimum and achieving a system that meets the objectives and goals outlined above. As a result, when engaging in a Platform Implementation of this Reference Architecture, it is a recommended practice for each Functional group in this architecture to reference the relevant specific constraining criteria to that group in the following categories:

- **Compliance Constraints:** Identify the specific controls and reporting required to ensure compliance, and the production of necessary compliance artifacts.
- **Architectural Constraints:** Identify architectural constraints, such as scaling profile, operating system, or configuration requirements that may conflict with current supported scaling and deployment models, including understanding approved / non-approved software products.
- **Support Constraints:** Identify areas where additional skills or support may be necessary to support an extended Functional group.
- **Network Constraints:** Identify the networking requirements of the Functional group, and areas where this may require traversing network boundaries, which may require specific approvals and configurations.
- **Security Constraints:** Reference the specific Air Force security policies applicable to the Functional group and identify configurations and tests required for compliance.

4.5.1 COMPLIANCE CONSTRAINTS

The Air Force has specific policies pertaining to Risk Management, Information Integrity, and System Security. Prior to generating specifications for a Platform Implementation, the implementer reviews these policies, identifies applicable controls, and then maps controls to the proposed solution – identifying gaps and remediation. This analysis package will then become the basis of the submission for a formal accreditation package.

4.5.2 ARCHITECTURAL CONSTRAINTS

The primary Air Force Data Services Reference Architecture constraints are to 1) ensure consistency with current Air Force hosting and deployment preferences, 2) understand the approved (and preferred) models for application and code deployment, 3) understand the availability of approved tools, and to develop rationales for additional tools if deemed necessary, and 4) to ensure consistency with the current capacity management model for the Air Force shared services environment. These architectural constraints are accounted for and documented in the proposed Platform Implementation.

4.5.3 SUPPORT CONSTRAINTS

Platform Implementations of Air Force Data Services Reference Architecture will be developed, hosted, and managed in customer designated locations. Therefore, prior to developing and launching a Platform Implementation Functional group, a support plan is in place that maps the tools in the Functional group to existing skills in these respective locations or provides a plan for availability of support.

4.5.4 NETWORK CONSTRAINTS

To comply with Air Force security and governance, and to ensure user access to data assets and the ability to publish analytics and data products, any Platform Implementation of Air Force Data Services Reference Architecture describes and maintains network boundaries around data release. The Air Force Data Services Reference Architecture structure, design principles, and specifications describe the network boundaries and the locations of the Functional groups within these boundaries.

5 TECHNICAL PATTERNS AND TEMPLATES

5.1 Capability Layer Definitions

The technical patterns and templates are provided as a best practice for this Reference Architecture and are organized around four capability layers, all of which consist of both Value-Added and Micro services.

- **Data Product Consumer Services (DPCS)** interact directly with the end consumers (users) of an Air Force Data Services Reference Architecture platform implementation. DPCS consume services from EDAS, EMS, and DPFS.
- **Enterprise Data and Analytics Services (EDAS)** provide data storage and operations, provide services to the DPCS, provide services to and consume them from the Enterprise Metadata Services, and consume services from the DPFS. The abstraction suggested here is sometimes referred as the Analytical Processing Engine throughout this document.

- **Enterprise Metadata Services (EMS)** support publication and discovery of data assets, data products, analytics, and services across a solution implementation. They consume DPFS and EDAS and provide services to EDAS and DPCS.
- **Data Platform Foundation Services (DPFS)** are general purpose services consumed by the other three layers.

5.1.1 SECURITY CONSTRAINTS

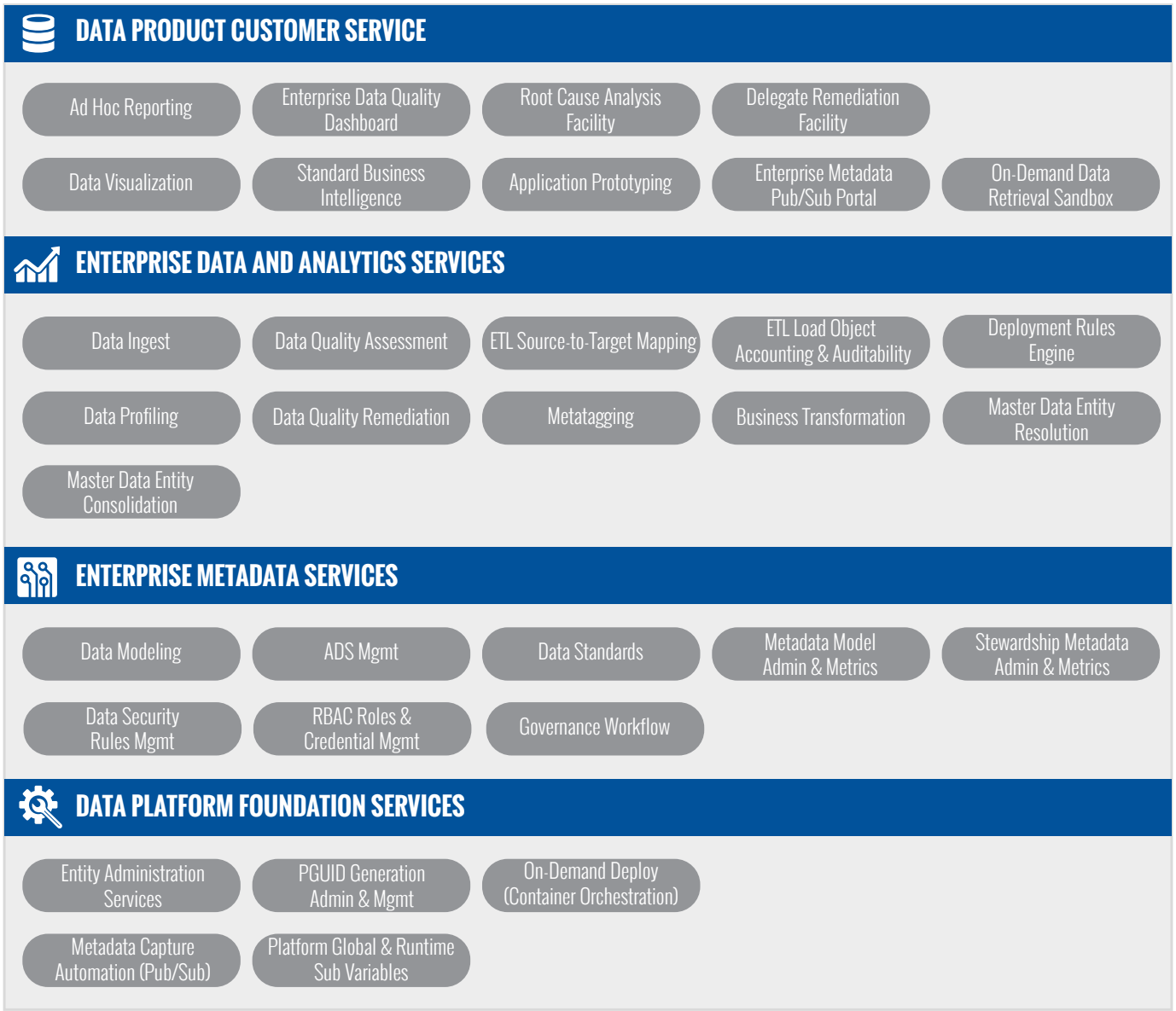
An implementer of an Air Force Data Services Reference Architecture Platform Implementation understands applicable security controls, including networking controls, application controls, authentication and audit standards, and applicable principles and regulations pertaining to data dissemination, retention, and release.

5.2 Value-Added Services View

Value-Added Services (VAS) represent value-added user/consumer activities that Air Force Data Services Reference Architecture is intended to support. Each VAS contains a description and functional standards and can be rendered by combining microservices into standard configurations and workflows as reusable, compound services. For configuration and implementation, each VAS has a defined workflow and a bill of materials consisting of the component microservices, data assets, and analytic assets needed for the activity to fulfil the business need.

Figure 1 below details the VAS for each Capability Layer, and the descriptions and standards for each are detailed within this section.

FIGURE 1: AIR FORCE DATA SERVICES REFERENCE ARCHITECTURE VALUE-ADDED SERVICE (VAS) VIEW



5.2.1 DATA PRODUCT CUSTOMER SERVICES

5.2.1.1 Ad-Hoc Reporting

Ad-Hoc Reporting Services provide consumers with an interactive facility to query data lake accessible datasets. Data Engineers, Developers and Data Scientists have access to command line facilities supporting interactive queries using lower level query languages. Business Users and tech savvy Business Analysts are provided facilities to support visual query tools based on predefined business-oriented views of the data with options to also use lower level query languages as necessary. The Ad-Hoc reporting services provide a means to retain report definition metadata for future use. In addition, users can publish and/or share generated datasets (report data) for use or review by other users.

5.2.1.2 Standard Business Intelligence

Business Intelligence (BI) Services constitute a class of analysis and reporting services which facilitate and streamline the construction, development and rendering of business analytics to support business decision making. The Standard Business Intelligence Services defined in this Reference Architecture is narrowed to reporting and visualization services supporting descriptive, diagnostic and (in some cases) predictive analytics. The platform provides these services understanding the variety of BI software suites available and their loosely coupled relationship to platform data. As a result, this standard can be met by a variety of FOSS or commercial off the shelf (COTS) software packages wherein the specific selection is based on localized options. It is a standard practice, however, that the implementation of any such tools never creates data asset level dependencies which would exclusively rely on use of that tool over time wherein using other tools or capabilities would necessitate re-formatting, translation, migration or transformation of platform data in any manner.

5.2.1.3 Data Visualization

Data Visualization Services will provide users with the ability to support meaningful, relevant and understandable analysis and story-telling with graphical depictions of analysis results. As with the BI tools, it is a best practice that the implementation of a visualization tool never creates data asset level dependencies which would only allow exclusive use of that tool over time wherein using other tools or capabilities would necessitate re-formatting, translation, migration or transformation of the platform data in any manner.

5.2.1.4 Enterprise Data Quality Dashboard

Effective management of Data Quality at an Enterprise level is essential to achieve the stated operational and strategic mission outcomes enabled by this architecture. In an Air Force context, Data Quality issues are well documented as factors impacting many Air Force initiatives including Enterprise Resource Planning (ERP) modernization efforts, Financial Audit Readiness Compliance and the reliability of predictive analytics models. The ability to measure, monitor, prioritize and remediate Data Quality issues with an enterprise perspective represents an end-state envisioned by this Reference Architecture. The Enterprise Data Quality Dashboard is an important construct of this envisioned end-state in which Data Quality Business Rules, evaluated datasets and the resulting metrics are reported and published along multiple dimensions aligned to the business metadata within the EIM: Domain, Organization, System, Business Process, Information Asset, Data Service, etc. The reporting structure also includes categorization of business rules and identified anomalies into the industry defined dimension of Data Quality including Completeness, Accuracy, Consistency, Currency, Precision, Integrity and Conformity. The Enterprise Data Quality Dashboard provides summary level reporting of the overall health of enterprise data assets complete with drill through capabilities eventually exposing record level images of the data violating one or more business rules. Business Rule metadata is sourced from the EIM and reflects the current state of EIM metadata. The Dashboard implementation is a reusable construct with the ability to be configured and used for organizational and project level data quality initiatives that could be used for a single system or single data migration effort. The re-usable construct will create a layer of common and shared metadata and reporting structure such that metrics gathered at a lower project or system level can be natively reported up to the enterprise dashboard. The enterprise dashboard is linked functionally to the Root Cause Analysis Portal and Delegate Remediation Portal described within this Reference Architecture.

5.2.1.5 Root Cause Analysis Facility

The Root Cause Analysis facility defined in this Reference Architecture supports collaboration among parties responsible for the identification, measurement, remediation and Root Cause Analysis of specific anomalies. The facility is organized around specific sets of anomalous data violating a specific business rule. The ultimate root cause is documented within this facility along with the steps and timeline needed for remediation. This facility is functionally linked to the Enterprise Data Quality Dashboard and Delegate Remediation Portal services also described in this Reference Architecture.

5.2.1.6 Delegated Remediation Facility

Delegated Remediation facilitates the remediation of data quality issues in federated fashion without violating the overall chain-of-custody implicit in the defined lines of data ownership, stewardship and governance. Delegated Remediation is an activity designed to minimize the churn between an authoritative source of data (usually an Air Force IT system) and the consumer of that data which has identified anomalies within the data sent. For example, in the context of an interface the consumer will identify a data quality issue with a known resolution. If the consumer remediates the data, then the overall chain-of-custody and auditability of that data is invalidated. Platform capabilities facilitate delegated remediation activities avoiding the need to modify and re-transmit data from the source while still retaining auditability and chain-of-custody. In one application, delegation empowers the consumer to remediate the data with the authority of the sending system. A second application is the ability to expose received anomaly data directly to its data owners (sending system) through a facility which would enable the data owners to authenticate and change the data directly.

5.2.1.7 Enterprise Pub/Sub Metadata Portal

Within the defined Reference Architecture, accommodations are made to expose the metadata associated with all providers and consumers of all data assets involved in Publish/Subscribe integrations. (Point-to-Point interface metadata is published in the same manner wherein the understood distinction is a single provider and a single consumer are documented) Metadata includes Publisher Details, Subscriber Details, Topic Name, Volume statistics, etc. In the context of the Reference Architecture, the goal is to promote data asset visibility and overall demand of specific data assets.

5.2.1.8 On-Demand Data Retrieval Sandbox

To promote data asset accessibility this Reference Architecture promotes Data Storefront capabilities which allow authenticated users to browse the Data Catalog (EIM) and download specific, discrete physical data assets available from the platform data lake context. Where applicable, the service provides access to specific versions or installments of the data.

5.2.1.9 Analytical Workflow and Data Services Prototyping

Platform capabilities provide the ability to construct data asset and analytical workflow exposure services within a services testbed. This would enable downstream consumers and trading partners to test service interactions in a non-production capacity typically using sample datasets. The Reference Architecture acknowledges the need to promote rapid prototyping and development activities by providing network accessible test services wherein the only change for a production implementation is a change to the service endpoint.

5.2.2 ENTERPRISE DATA AND ANALYTICS SERVICES

5.2.2.1 Data Ingest and Onboarding

Data ingestion services described in this Reference Architecture enable the import of data for immediate use via storage in a designated central repository, such as a data lake; via real-time streams or batches for ad hoc consumption; or through persistent services for on-demand access. Real-time streaming enables the consumption of data as it is produced by the source. Batches allow for data to be accessed in discrete chunks at periodic intervals of time. Ingest services provide the ability to prioritize data sources, validate files, and route data sources to the correct destination within the data lake repository. Persistent services may “ingest” data by onboarding a service co-located with the source data that enable analytic to be pushed to co-located compute and results returned. As a best practice, the platform can rapidly on-board a variety of files and formats including structured and unstructured data for immediate analysis and opportunities. Ingest services provide a declarative, configurable ingestion workflow facility which streamlines and consolidates the common tasks associated with constructing, maintaining and monitoring ingest pipelines. Ingest services also provide schema inference capabilities, perform tagging based on pattern recognition and apply validation and standardization rules to schema elements.

5.2.2.2 Data Profiling

Data profiling services compile statistical metadata about a data set into a technical profile so that one can understand the contents of the data. The platform enables the creation of technical profiles that calculate record counts, identify distinct values stored in each data element, frequency of occurrence and percentage occurrence of each distinct value, and identify data type/format information. Generated statistics also include min, max, median and density measure. These statistical characteristics provide analysts with important information about, and insight into, the essential constitution and complexion of the physical data assets under evaluation. Profile results promote an exploratory and often iterative process in which analysts observe and explore patterns often leading to questions about fitness for purpose, quality, integrity and consistency. By default, the profiling service captures profile results for each physical data asset at the time of ingestion or whenever the data asset’s schema is known, understood and applied. Versioned Profile Results can then be compared and analyzed for changes over time. This is especially important for monitoring remediation and application refactoring initiatives.

5.2.2.3 Data Quality Assessment

Data quality assessment services provide data practitioners the capabilities to evaluate the degree to which data within and transiting through systems is accurate, complete, timely, relevant and consistent with all standards and business rules. The platform supports data quality processes that evaluate data to determine its “fitness for use” for specific business purposes based on published, authoritative rules published in the EIM. The platform supports the design of business rules, code development that implements data quality business rules, ability to execute the data quality business rules, and report generation of anomalies. The Data Quality Assessment service operates on a common metadata model to ensure the consistency and fidelity of data quality metrics across the enterprise. Data Quality metrics and anomaly data will be exposed through the Enterprise Data Quality Dashboard also described within this Reference Architecture.

5.2.2.4 Data Quality Remediation

Data quality remediation services support processes to resolve anomalies identified during data quality assessment. The remediation service determines how, by whom, and when anomalies are resolved. The platform includes ability to track the status of anomalies involved in the remediation process. Remediation Services will operate on the same set of microservices which enable the Business Transformation service described within this Reference Architecture.

5.2.2.5 Source-to-Target Mapping

The Source-to-Target mappings are the blueprint of data curation (data collection, transformation, structuring, and service publication) activities. Source-to-Target mapping facility should be flexible tosupport mappings to instantiate specific overload schemas. The service provides ability to capture load frequency, join instructions, data types, conversion logic, and applicable business rules for anomaly remediation. Source and Target entities and attributes are obtained directly from published EIM metadata to ensure enterprise consistency across multiple transformation programs. Moreover, the Source-to-Target Mapping service metadata is stored in a centralized, queryable repository accessible to business and system analysts.

5.2.2.6 Metatagging

The metatagging service supports both manual and automated tagging methods for all physical data assets inducted into the platform data lake as well as tagging of assets exposed to downstream consumers. Automated tagging techniques are based on pattern recognition of metadata of both data element names or data values using reasoning. Pattern recognition metadata is associated to EIM recognized domain types and classifiers. Metatags are searchable to aid analysts and engineers in identifying and locating data assets.

5.2.2.7 Business-Driven Data Transformation

The Business Transformation service is a value-added service which specializes in fit-for-purposes transformation of data assets based on business defined needs or data quality remediation rules. A broad spectrum of specific transformation techniques and capabilities are enabled by micro-services enablers defined within this Reference Architecture. The Business Transformation service will provide metadata linkage to specific Source-to-Target mappings with upstream linearity to documented project specifications or other functional design and testing artifacts. This metadata can then be used to support standards and principles for traceability matrices and other traceability reporting. The metadata associated with the transformation can be used to auto-generate templates associated with preferred implementation methodologies that can be published as on-demand services for use within the platform or as a web service by external consumers. Transformation metadata is also captured within the platform’s data lineage facility in the EIM.

5.2.2.8 Load Object Accounting & Auditability

The Load Object Accounting services provides facilities to gather various counts and compile metrics associated with data acquisition, data validation, data cleansing, data transformation and general data curation and publish/consume processes using simple accounting jobs available on the platform. The resulting metrics ultimately provide managers and decision makers insights into the expected number entities or records converted and loaded into the target data environment.

5.2.2.9 Deployment Rules Engine

The deployment rules engine service provides the unique capability to support site-by-site or organization-by-organization data migration activities without the need to modify, re-compile or redeliver conversion code. Leveraging specific platform micro-services and metadata, the Rules Engine combines its Organization/Location Tagging capabilities (OrgLoc) with an exposed layer of event specific metadata to enable managers to define business rules which dictate how migrating datasets will be automatically filtered for a specific deployment event. Collectively, this capability is referred to as “Transparent Parameterization”. Beyond Data Migration activities, Transparent Parameterization can be used for a variety of scenarios in which specific domain specific data values leveraged to seamlessly publish domain specific datasets whose values can be configured and/or filtered appropriately for a given event.

5.2.2.10 Master Data Entity Resolution

The Master Data Entity service provides the ability to define, sequence and relate rules in a manner which provides determinations of identity wherein the possible outcome supports a match to an existing business entity or, a no-match condition in which a new entity is created. Enhancements to this service include facilities to support probabilistic match algorithms in which the algorithm produces a percentage-based confidence level that a given entity is a match to an existing entity. Confidence levels above a certain percentage can be configured to execute the match while other lower thresholds may trigger a data steward review and acceptance of the match. Combined with the Master Data Entity Consolidation Service, this Reference Architecture acknowledges the importance and impact of Master Data Management (MDM) challenges across the Air Force enterprise. In general terms, MDM Solutions consist of the people, processes, and technology that ensure business-critical master data is available to enterprise consumers in a single, consolidated and authoritative view. While many COTS applications exist within the MDM space, the goal of this Reference Architecture is to encapsulate essential MDM best practices and design patterns into reusable components based on FOSS and cloud-native technology enablers.

5.2.2.11 Master Data Entity Consolidation

Master Data Entity Consolidation serves as a component service designed to aid in the construction of MDM solutions requiring entity consolidation based on cell-level survivorship enabled by source specific, column-level trust value assignments. Enhancements to the service include facilities to support gold record creation based on Best Value of the Truth calculations accounting for trust value assignments, trust validation rules, precedence and trust decay over time. For distributed systems and platform consumers, these capabilities can be published as an on-demand service at run-time for resolution.

5.2.3 ENTERPRISE METADATA SERVICES

5.2.3.1 Data Modeling

The EIM provides Data Modeling facilities based on new or existing entities and attributes registered within the EIM. The modeling services are structured to support the modeling of both data-at-rest and data-in-flight wherein the baseline model in both cases is the Conceptual Model covering the major business entities within the subject’s functional domain. For data-at-rest,

the Conceptual Model may then be decomposed into an Enterprise, Logical Model converting the core entities and attributes supporting domain specific information standards. This Logical Model would then be translated to a physical model for application specific implementation in a database repository. For data-in-flight (or messaging/interface implementations) the Conceptual Model is decomposed into a Canonical Model typically consisting of a denormalized, object-oriented model expressed in an eXtensible Markup Language (XML) Schema Definition. Messaging specific sub-schemas are ultimately derived from canonicals as needed and used operationally in web service and Application Programming Interface (API) exposures of the subject datasets.

5.2.3.2 ADS Management

The Authoritative Data Sources (ADS) management service supports the capture, maintenance, exposure, and governance of ADS within the EIM context. ADS designations provide data consumers and practitioners with the most trusted and reliable source for information, as determined by an appropriate governing body. The platform supports the ability to denote the different ADS designation types: source ADS, aggregation ADS, and suitable ADS. The enabling platform provides mechanisms and utilities to identify different levels of ADS, to include element and information asset designations. The capability enables linking of ADS to data dictionary data elements and Information Asset Catalog assets within the EIM.

5.2.3.3 Data Standards

In the Reference Architecture, the EIM will serve as the authoritative enterprise source for how specific data standards apply to specific information assets and their constituent data dictionary terms, as well as the authoritative source for enterprise metadata. In this role, the EIM should be able to link to and bridge across various functional and contextual data standards and entity models, leveraging and integrating existing models to the greatest extent possible. Examples of data standards and entity models in the Air Force context include standards such as Real Property Information Model (RPIM), Enterprise Energy Information Management (EEIM), Standard Financial Information Structure (SFIS), Defense Logistics Management Standards (DLMS) and Financial Improvement and Audit Remediation (FIAR). The EIM, in these cases, would link these standards to the specific attribute and information assets governed by a specific standard. In general terms, the following data-standard areas are governed as part of the EIM:

- Data Exchange and Interoperability Standards to ensure semantic interoperability
- Data Retention and Archiving Standards
- Data Access and Security Standards
- Naming convention for XML schemas, message formats, entities, and attributes
- Service-oriented architecture (SOA)-specific standards and messaging protocols to standardize service definitions
- Canonical XML schemas to help data exchange across the organization
- Reference compliance standards (such as privacy and security standards)

5.2.3.4 Business Rule Management

Effective management of business rules is essential to establish a measurement of data on

various dimensions of quality from functional and technical perspectives. Business rules describe the operations, definitions and constraints applying to an organization that are in place to help the organization achieve its goals. The platform supports the capture, management, execution, maintenance, and monitoring of Logical Business Rules (LBR) and Physical Business Rules (PBR). The service enables both business experts and technical developers to maintain a repository of business rules for reuse, create candidate business rules, link business rules to data dictionary elements, execute business rules, and expose business rule results. The service centrally organizes, maintains, and exposes the business rules within the EIM construct.

5.2.3.5 Metadata Model Administration and Metrics

The EIM provides an extensible metadata model to support modifications to the defined metadata standards over time. As data assets, data products, and users are onboarded onto the Air Force Data Services Reference Architecture, the metadata model will need to be extended and new lexicon will need to be added and extend the Air Force Data Value lexicons or be mapped to existing synonyms. This is a combined process of automated discovery, deliberate maintenance, and metadata governance. This service will enable this governance and administration function by allowing Air Force Data Services Reference Architecture user to visualize the metadata model in a graphical context, and to query related elements and lexicon terms as well as key metrics pertaining to data in the solution implementation that will assist in governance decisions pertaining to model and lexicon maintenance, extension, and deprecation.

5.2.3.6 Stewardship Metadata Administration and Metrics

The ability to capture, track and maintain stewardship metadata attributes about data stewards, stewardship processes, and responsibility assignments within the EIM repository. Data stewards ensure metadata is accurate and is of high quality across the enterprise as well as establish and monitor sharing of data. Data governance practitioners can centrally manage data stewardship metadata including: data steward roles and responsibilities, Functional and technical Subject Matter Experts (SME's), data owners, data users, contact information on data stewards and SME's, regulatory bodies; governance organization structure and responsibilities, and contextual metadata pertaining to operational / business context of the data asset, data product, or analytic asset.

5.2.3.7 Governance Workflow

The EIM provides configurable workflow facilities to enable the business processes associated with the full lifecycle management of the data assets, data products, and analytics assets in the context of the business, technical and operational metadata contained within EIM as well as the business rules. This governance workflow will support ADS Management, Data Modeling, and configuration and risk management of items with a solution implementation of the Air Force Data Services Reference Architecture.

5.2.3.8 Data Security Rules Management

Data Security Rule Management is a capability that manages rules for the dissemination and restriction of data assets, data products, and analytical assets. Protection and dissemination control of Air Force data is critical to the Air Force mission and to protection mission and data assets. Since the platform may involve the creation of new data assets, this protection and control may need to be included in the design and development phase of new data products

with greater sensitivity than the inputs. Decisions regarding the sensitivity (classification), organizational (compartment), and role (functional) restrictions and credentials that will be enforced through the Role Based Access Control microservice will be determined through the Air Force Data Service Reference Architecture governance function.

5.2.3.9 Role Based Access Control Roles and Credential Management

The Role Based Access Control (RBAC) Roles and Credential Management serves as a set of tools to administer the management of assigning credentials to users to manage data dissemination controls. The assigning of these credentials will be determined through the Air Force Data Service Reference Architecture governance function.

5.2.4 DATA PLATFORM FOUNDATION SERVICES

5.2.4.1 Entity Administration Service

The Entity Administration Service allows for the ingestion and maintenance of entity reference sets, defined as formal sets of codes or terms maintained by functional users and consumers of platform services. Examples could include financial or maintenance codes, geographic place names, or equipment names or codes (such as Federal Supply Class (or National Stock Number codes)). This service supports both ingestion of entity data as well as the ability to call external services that provide entity data.

5.2.4.2 Metadata Capture Automation (Pub/Sub)

This service facilitates the automated capture of metadata (technical, administrative, operational/ contextual) about data and analytic assets and data products through automated interaction with other services within the Air Force Data Services Reference Architecture and through defined interactions with external publishers/consumers.

5.2.4.3 PGUID Generation, Administration & Management

A system of generating, managing and monitoring Globally Unique IDs are implemented as a shared service in the platform context. PGUIDs are globally unique across a defined enterprise or organizational namespace. PGUIDs are retrievable via a service call and accessible in all contexts. PGUID length and formatting are documented.

5.2.4.4 On-Demand Deployment (Service Orchestration)

This capability is developed and deployed using on-demand deployment technologies such as continuous integration and service/container orchestration. This ensures flexibility, quality control, speed and most importantly, that all service modifications and configurations are under source control from day one.

5.2.4.5 Platform Global & Runtime Substitution Variables

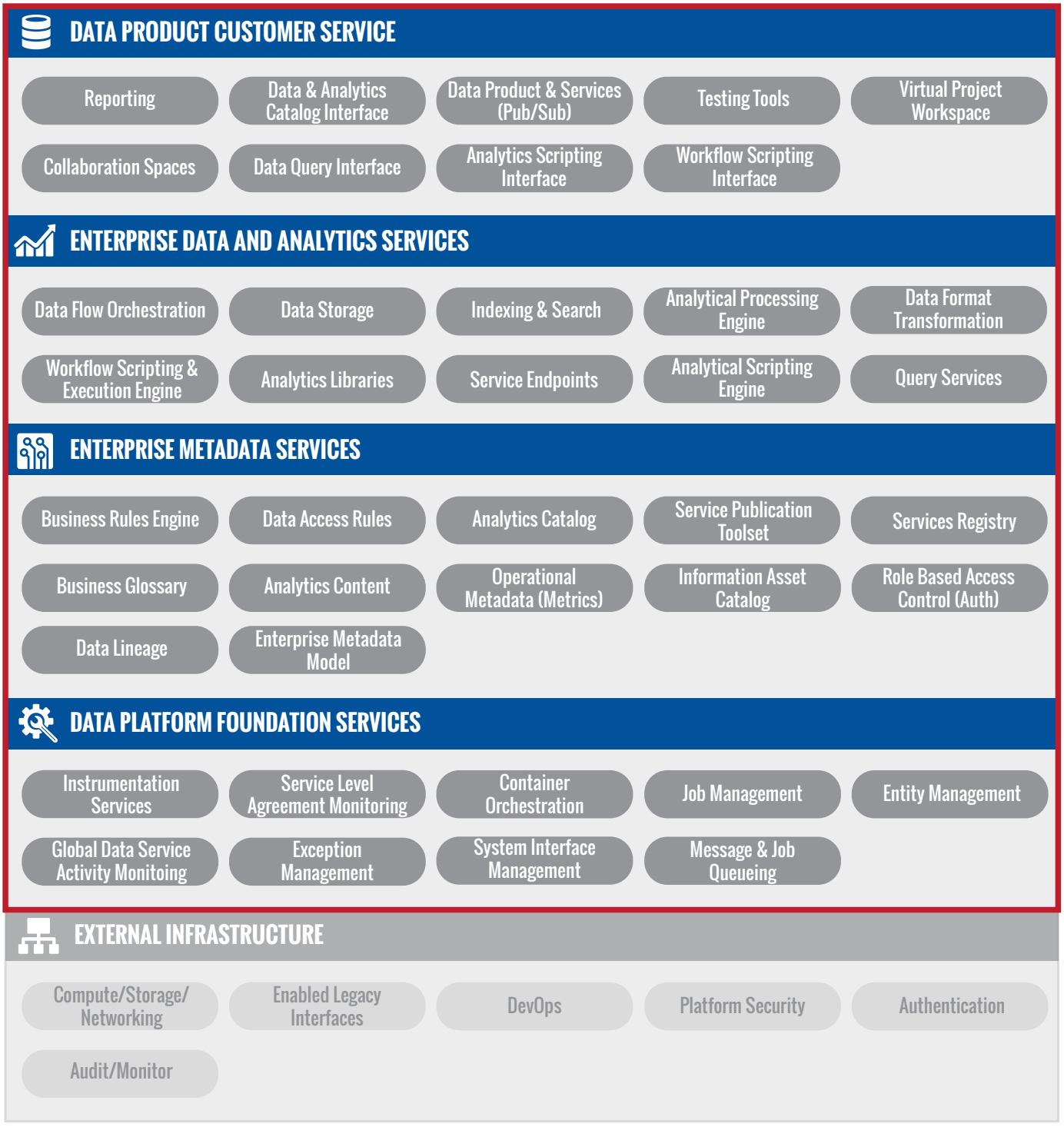
The Platform Global and Runtime Substitution Variables provide a means for defining, initializing, setting and interrogating global and runtime substitution parameters. Variables are defined within specific, user defined namespaces and provide a programmatic means of variable interrogation as well as indirect references via substitution markup.

5.3 Microservices View

Microservices (Micro) detail the fundamental core functions of a solution implementation of Air Force Data Services Reference Architecture and represent the software components of the solutions that are installed/developed, tested, and exposed as services within the Air Force Data Services Reference Architecture to support the VAS. Each Micro has clear standards, constraining criteria, and benchmark performance targets that can be used for test-driven development and continuous integration. Combining these microservices into standard configurations and workflows as reusable, compound services creates VASs.

Figure 2 below details the Micros for each Capability Layer, and the descriptions and standards for each are detailed within this section.

FIGURE 2: AIR FORCE DATA SERVICES REFERENCE ARCHITECTURE MICROSERVICES (MICRO) VIEW



The Microservices View consists of the functions and services necessary to present and display information and data to Data Scientists, Business Analysts, and Consumer actors within Air Force Data Services Reference Architecture platform implementation. It consists of both the scripting and querying user interfaces, as well as interfaces to configure and publish dashboards. This group is supported using COTS or FOSS rather than custom development. All Framework Code in the Microservices View is managed under Change Control Board (CCB) control.

5.3 DATA PRODUCT CUSTOMER SERVICES

5.3.1.1 Reporting

A report generation service based upon a COTS/FOSS tool for presenting reports as interactive dashboards. The service allows Business Analysts to self-service create and publish report Data Products for Consumers. These products are saved and added to the Metadata Catalog so that they can be shared with colleagues, discovered, and reused. Reporting includes the following Systematic Functional Standards:

- Supports multiple users
- Allows users to share reports with each other
- Ability to support front-end user authentication and role-based access
- Ability to configure both standard and ad hoc report objects
- Ability to save products as objects for reuse
- Ability to associate create and edit privileges for a product with a named user

The Data Visualization and Dashboarding Functional group incorporate the following design criteria:

- Use API Services, Queries, and Data Products from the platform Service Management and Querying Functional groups
- Store products as objects and add the Uniform Resource Identifier (URI) to the Metadata Catalog for discovery and re-use
- Store product objects in Data Storage

5.3.1.2 Data and Analytics Catalog Interface

A catalog interface that includes both a graphical user interface as well as a web service interface for discovery of data assets, data products, and analytics assets. The catalog interface supports search and query across technical, administrative, and contextual metadata. It supports domain specific language queries and the ability to retrieve metadata and pass that metadata to other services within the Air Force Data Services Reference Architecture for execution.

5.3.1.3 Data Product and Service Pub/Sub Interface

A service interface (both graphical and web service) for subscribing and publishing data and analytic assets as well as data products allows the publisher to specify dissemination control criteria and a service level for Air Force Data Services Reference Architecture enforcement. The service interface provides the product to the consumer and enforces dissemination controls and service level. This interface should support internal service publication and consumption within

the platform, from the platform to external consumers, and publishing by external consumers to the platform, such that the services can support multiple service delivery models. Examples of multiple service delivery models include data ingested into a data lake for analysis, a subset of data relayed to a remote service, or the remote service instantiation of a job on co-located compute for local execution with a returned result.

5.3.1.4 Data Query Interface

A user friendly graphical User Interface (UI) for interacting with the data platform Query functions. Provides users with a simple user interface to create high availability, high scale queries and save these queries as objects. The Query User Interface Functional group contains the following Systematic Functional Standards:

- Ability to create new indices and document types to support queries
- Ability to use query to generate visualizations, dashboards, and reports
- Ability to save and reuse pre-defined queries through the user interface
- In-depth view of data on command
- Supports new query creation through code completion
- Full feature learning of UI is easy and obtainable within a day or two of training
- Admins and developers can dual as a diagnostics tool

The Query UI is a user friendly graphical interface and natively interfaces with the Querying Function Group.

The Query UI incorporates the following *constraining criteria*:

- Interface with Air Force Single-Sign-On authentication
- Comply with Air Force user interface support guidelines

5.3.1.5 Analytics Scripting Interface

Analytics Scripting graphical interface that can provide Job Scripting for analytics. Ability to identify and retrieve APIs, Data Products, and Analytics from the Metadata Catalog as well as analytics from the Analytics Libraries. The Analytics Scripting UI is intended to provide Data Scientists and Business Analysts with a user-friendly interface to take advantage of the Job Scripting Functional group. The Analytics Scripting UI Functional group includes the following Systematic Functional Standards:

- Graphical user interface with a guided workflow
- Allows users to share workflows and analytical workbooks with each other
- Ability to retrieve Analytics Libraries as objects for use in workflows
- Ability to retrieve API Services as objects for use in workflows, including:
    - Queries
    - Data Products
    - Analytical Jobs

- Natively interface with Workflow Scripting service
- Supports multiple users
- Allows users to share workflows and analytical workbooks with each other
- Ability to support front-end user authentication and role-based access

The Analytics Scripting UI is a user friendly graphical interface and can natively interface with the Job Scripting Functional group to retrieve services and read/write from/to the Metadata Catalog.

The Analytics Scripting UI incorporates the following *constraining criteria*:

- Interface with Air Force authentication
- Comply with Air Force user interface support guidelines
- Support a minimum set of analytical modes

5.3.1.6 Workflow Scripting Interface

Workflow Scripting graphical interface that allows scripting of data asset retrieval, analysis, and publication of data product workflows. Provides the ability to identify and retrieve APIs, Data Assets, Data Products, and Analytic Assets from the Metadata Catalog as well as analytics from the Analytics Libraries. The Workflow Scripting UI is intended to provide Data Scientists and Business Analysts with a user-friendly interface to take advantage of the Job Scripting Functional group. The Workflow Scripting UI includes the following Systematic Functional Standards:

- Graphical user interface to create a guided workflow
- Tools to enforce reference and logical integrity of workflows
- Allows users to share workflows with each other, as well as publish
- Ability to retrieve Data Assets for use in workflows
- Ability to retrieve Data Products for use in workflows
- Ability to retrieve Analytic Assets for use in workflows
- Ability to retrieve Analytics Libraries as objects for use in workflows
- Ability to retrieve API Services as objects for use in workflows, including:
    - Queries
    - Data Products
    - Analytical Jobs
- Natively interfaces with Analytics Scripting service
- Supports multiple users
- Allows users to share workflows and analytical workbooks with each other
- Ability to support front-end user authentication and role-based access

The Workflow Scripting UI is a user friendly graphical interface and can natively interface with the

Analytics Scripting service to retrieve services and read/write from/to the Metadata Catalog.

The Workflow Scripting UI incorporates the following constraining criteria:

- Interface with Air Force authentication
- Comply with Air Force user interface support guidelines
- Support for a minimum set of workflow modes

5.3.1.7 Testing Tools

A suite of testing tools to validate that Data Products, Analytic Assets, and Workflows comply are technically/logically compatible with Air Force Data Services Reference Architecture services and that they comply with Air Force data quality standards.

5.3.1.8 Virtual Project Workspace

User-specific virtual workspaces in which to create and manage projects that use the user interface tools within the suite.

5.3.1.9 Collaboration Spaces

Collaboration tools that allow users of the system to communicate and share documentation in a manner that is integrated with object metadata. Collaboration includes chat functions, self-service documentation publishing functions, and comment functions that are integrated with the data asset, data product, or analytic asset being referenced.

5.3.2 ENTERPRISE DATA AND ANALYTICS SERVICES

5.3.2.1 Data Flow Orchestration

A service for orchestrating data flows throughout the system. This orchestration service is aware of services and APIs within Air Force Data Services Reference Architecture and can manage data flows that support functions and workflows in the most efficient manner possible. The orchestration service interfaces with the RBAC to enforce dissemination controls.

5.3.2.2 Data Storage

A storage service for storing all data at rest. Storage supports both discrete Data Products and provides storage capability for other data platform functions and all users/consumers of platform services. Storage solutions are intended to act as a commodity shared service behind registered web services and Value Added Services. The design objective of the Storage service is to provide a single storage environment with a single set of service endpoints, so that this can be scaled and managed effectively. Storage includes the following Systematic Functional Standards:

- Atomicity, Consistency, Isolation, Durability (ACID) Compliance
- Ability to scale without downtime
- Ability to interact directly with the Data Operations Framework and Query engine to support read/write in different formats
- Support Restful Services (REST), and Open Database Connectivity (ODBC)/Java Database Connectivity (JDBC) interfaces

- Integrate with Service Management
- Storage components have well developed documentation and preferably, a large user base
- Joins, selects, and views capability
- Natively integrate with cloud-based storage options

The data platform Storage service reflects specific design principles intended to make it easier to maintain, support, configure, and update:

- Data Products exposed as web services through the API Service Management and Metadata Catalog Functional groups – database functions are not exposed directly to non-Infrastructure or non-Developer System Actors
- Store all data platform data for as long as retention policy demands
- Support a shorter-term (180 day) high performance query store that supports rapid, user friendly queries and a graphical query interface
- The Data Storage Functional group takes advantage of a direct push to the Querying Functional group to ensure data consistency
- Storage can be a dedicated environment, such as a data lake, or a virtualized environment, such as a networked, service-enabled collection of web services attached to legacy systems

The Storage Functional group recognizes and accounts for the following *constraining criteria*:

- Support security protocols such as Secure Sockets Layer (SSL), Transport Layer Security (TLS), and encryption at rest and in motion
- Support for virtualized deployment methodologies such as a virtual machine, machine image, or container based deployment and scaling models
- Evaluation considers existing deployed/supported data stores in the Air Force
- A serialization capability to store multiple data structures
- Support unlimited data types

5.3.2.3 Workflow Scripting and Execution Engine

An engine for scripting workflows that include data assets, analytical assets, data products, and visualizations. The Workflow Scripting and Execution Engine service interfaces with the Workflow Scripting User Interface and contain the logic for validation of workflows for logical and referential integrity. The Workflow Scripting and Execution Engine is aware of the Metadata pertaining to data assets, analytical assets, data products, and visualizations through Metadata in the Information Asset Catalog and assists in maintaining timeliness of product and fitness for use. The Workflow Scripting and Execution Engine interfaces with the RBAC and is aware of, and enforces, dissemination controls. The Workflow Scripting and Execution Engine can schedule and execute jobs in the Analytical Processing Engine.

5.3.2.4 Indexing and Search

An indexing and search service that supports discovery and retrieval through the Metadata

Catalog and Querying functions by managing and exposing primary and secondary keys and indexed elements that can be used to select and modify requests. As a service, Indexing extends across both the Querying and Data Storage services and provides the ability to configure queries and search for individual records. Therefore, Indexing has both a Storage and Query configuration. Indexing includes the following Systematic Functional Standards:

- Ability to add new records to Indices in both Storage and Query post ingestion
- Ability to take advantage of Entity Management indices
- Indices are persisted in all logical partitions (Raw, Quality Assurance (QA)/ Quality Control (QC), Enriched), Query stores, and Metadata, as appropriate
- Indices are easily created and/or extended
- Supports full text analysers, as appropriate

The Indexing Functional group accounts for the following constraining criteria:

- Need to integrate services with Data Storage and Querying in a manner the supports expected query response times

5.3.2.5 Analytics Libraries

Analytics libraries represent reference libraries of code objects that support generalized analytical functions that can be executed in the Analytical Processing framework and scripted in the Analytics Scripting environment. The objective of the analytics libraries is to populate the Analytics Scripting environment with established libraries with external support that allow us to maintain, patch, and upgrade the libraries with the minimum amount of internal development. This also enforces use of common analytical objects, ensuring consistent data operations that minimize discrepancies between Data Products. The Analytics Libraries service includes the following Systematic Functional Standards:

- Rely on established open-source analytics libraries with sufficient committers and support
- Libraries are compiled languages to support rapid integration into the Job Scripting functional
- Libraries support the Job Scripting Functional group

The primary design criteria for the Analytics Libraries are that they be natively compatible with the Analytics Scripting Functional group and maintenance of these libraries is conducted using Continuous Integration.

The Analytics Libraries Functional group reflects the following constraining criteria:

- Compliance with approved libraries

5.3.2.6 Query Services

A query service that can create and retrieve collections, data sets, and views from Storage. The design objectives of the Querying service are to identify the set of functions and utilities necessary to expose data querying to Data Scientists and Business Analysts in a way that is highly responsive and scalable yet reduces performance and integrity risk to Data Storage.

Querying includes the following Systematic Functional Standards:

- Querying tool includes a domain specific language (DSL):
  - The DSL is already well known or extremely easy to learn
  - DSL has strong documentation
  - Preferably a strong user base
- Ability to save and reuse queries
  - Easily extensible
- Ability to combine queries to produce reports easily
  - Query results are exportable
- Ability to add saves queries to the Metadata Catalog for discovery
  - Able to externally link to defined/saved queries as a URI
- Graphical User Interface to create queries
  - Preferably Code completion
- When queries fail, error messages are explicit and easy to debug
- Easily support queries necessary for creation of metrics
- Highly scalable (horizontally) and fault tolerant
- Queries support multiple languages with minimum dependencies

The querying Functional group includes specific designs, such as:

- Ability to add defined/saved queries as a URI in an external registry, including relevant metadata
- Fast interface for query retrieval
- As closely integrated with Data Storage as possible
- Support a shorter-term (180 day) high performance query store that supports rapid, user friendly queries and a graphical query interface
- Index and document design patterns that a priori constrain the ability of users to create malformed queries
- Reflect the correlation and normalization schemes in the index and document designs
- Point and click functionality for query construction
- Ability to publish queries as reports for consumption by the Data Reporting and Visualization Functional group

The Querying Functional group accounts for the following constraining criteria:

- Network connectivity could reduce query speed, so design patterns are used that reduce network load

- Support security protocols (e.g., SSL, TLS)
- Strong documentation and user base with high end-user supportability
- Support a shorter-term (180 day) high performance query store that supports rapid, user friendly queries and a graphical query interface
- Fault tolerant architecture
- Support for joins and nested logic
- Can retrieve and export multiple file formats

5.3.2.7 Service Endpoints

Service Endpoints ensure that all microservices, data assets, data products, and analytical assets are callable as web services. Therefore, the solution implementation provides for an initial set of base service endpoints upon initial deployment.

5.3.2.8 Data Format Transformation

A service for translating data assets across format, structure, and type that maintains logical and referential integrity across services. The input/output configurations of the service are controlled by the Enterprise Metadata Model and comply with metadata in the Information Asset Catalog.

5.3.2.9 Analytical Processing Engine

An Analytical Processing Engine using a parallel processing environment that can retrieve data from storage, interface with Air Force MAJCOM/Functional Data Platform services, receive jobs from Workflow Scripting and Execution, and perform manipulations on the data to render a new Data Product. The Analytical Processing Engine acts as a management framework and engine for scheduling, executing, controlling, and distributing analytical processing jobs, either in a data lake or in a co-located computing environment. The Analytical Processing Engine includes the following Systematic Functional Standards:

- Supports streaming and batch processing capabilities
- Supports job scheduling
- Supports multiple languages for analytics
- Supports exactly once delivery semantics
- Functionality and operation is independent of any analytics or processing it performs
    - Any functionality added can be done without modifying the source code
    - Functionality is modular and implemented in a microservices fashion
- Supports multiple data functions, in concert with the Analytics Libraries
- Ability to configure to ingest and process multiple file formats
- Ability to configure to ingest and process multiple data structures
- Ability to configure and process multiple data types

- Ability to integrate with the Job Management Functional group for initialization and management of jobs
- Ability to expose data operations as Web Services
- Ability to consume data from Message Queuing and Batch Upload Functional groups
- Ability to write to Data Storage Functional group
- Ability to parallelize data operations

Analytical Processing reflected critical design criteria intended to ensure flexibility, scalability, reuse and documentation:

- Analytical Processing jobs can be created and loaded through the Workflow Scripting service
- Analytical Processing jobs can be controlled, scheduled and managed through the Workflow Scripting and Execution Engine service, which is ideally closely coupled to Analytical Processing
- Analytical Processing jobs can be written to the Data Storage service
- Analytical Processing jobs can be distributed to registered co-located compute locations
- All Data Products can be added to the Metadata Catalog concurrent with being written to Data Storage; using standard typing
- Standard enrichments and metrics are calculated and emitted upon ingestion
    - Enrichments and Metrics are added to the Catalog
- Analytical Processing functionality and operation is independent of any analytics or processing it performs
    - Any functionality added can be done without modifying the source code
    - Functionality is modular and implemented in a microservices fashion
- The critical nature of the Analytical Processing Engine means it must be deployed as a horizontal peer-based service that holistically recovers from failures in any given node or region

Analytical Processing scales horizontally to process the necessary workload (both streaming and batch).Therefore, the Data Operations Framework utilizes a multi-region, multi-node deployment approach, with workload estimation and proactive deployment of additional nodes to ensure adequate parallel processing capacity as well as peering to ensure job continuity if a node or region fails. Further, streaming and batch nodes can be managed separately, as the streaming load is more predictable and therefore can be insulated from the more variable batch Analytical Processing workload.

A deployed, configured Analytical Processing framework can account for the following *constraining criteria*:

- Current data quality, completeness, and accuracy are measured and surpassed (The exact qualifications for this are yet to be determined)

- Workflow Scripting and Execution Engine service is closely coupled with Analytical Processing
- Analytics and data sets can be registered quickly
  - Analytics can be used within hours
  - Data sets can be used within days
- The Data Operations Framework is a central component for Air Force Data Services Reference Architecture and its failure will negatively impact the entire architecture
  - It can be considered a single point of failure

5.3.2.10 Analytical Scripting Engine

Analytical Scripting Engine defines, writes, scripts, and tests jobs before being executed in the Analytical Processing framework and added to Job Management and the Metadata Catalog. Analytical Scripting ensures consistent use of Analytics Libraries, consistent deployment to the Analytical Processing framework, standardized consumption of web services, automated documentation, and automated addition to the Metadata Catalog. To maximize these benefits, the Analytical Scripting service includes the following Systematic Functional Standards:

- Ability to operate as a server supporting multiple scripting clients
- Ability to store scripts in a structured manner in Data Storage
- Ability to publish scripts to the Metadata Catalog
- Ability to consume Web Services and Data Products from within Air Force MAJCOM/ Functional Data Platforms, through the Querying and API Service Management Functional groups
- Ability to retrieve API Services as objects for use in workflows, including:
  - Queries
  - Data Products
  - Analytical Jobs
- Natively contains multiple standard analytics libraries
- Ability to consume Metadata to ensure submitted analytics and data used are of matching types
- Ability to call the Job Management function to initiate jobs in the Analytical Processing Framework

The Job Scripting server reflects specific design criteria, including the need to interact with multiple system components:

- Job Scripting acts as the broker between the Data Scientist and / or Business Analyst and the Analytical Processing Framework, Data Products, and API Services Management
- Job Scripting is integrated with the Metadata Catalog to discover Data Products, Queries, and Analytical Jobs

- Job Scripting is configured to allow access to services, and to Data Products, through the Job Scripting Functional group
- Job Scripting is configured to present configuration managed Analytics Libraries to users
- Job Scripting automatically publish all saved jobs to the Metadata Catalog with appropriate technical and application metadata

The Job Scripting Functional group needs to account for the following *constraining criteria*:

- Support approved analytical libraries

5.3.3 ENTERPRISE METADATA SERVICES

5.3.3.1 Business Rules Engine

Business Rules Engine service supports the specific execution of business rules, either as services or as scheduled jobs that are executed either by logical or temporal triggers within the system. The Business Rules Engine interfaces with the Analytical Processing Engine and the Workflow Scripting and Execution Engine.

5.3.3.2 Business Glossary

Capability to import, publish, and apply one or more business glossaries (lexicons) through Analytical Workflows and with integration into the Information Asset Catalog. Business glossaries are stored in the Storage service as Data Assets in their own right.

5.3.3.3 Data Access Rules

Documented data access rules for all Data and Analytical Assets and Data Products. These rules specify dissemination controls by sensitivity, organization, role/use, and are describable to the field level as needed. The rules comply with the Enterprise Metadata Model design and are capable of being implemented by the RBAC and documentation in the Enterprise Information Asset Catalog service.

5.3.3.4 Analytics Content

Analytics Content refers to the essential metadata required to fully describe completed analytics

projects such that the project is fully documented, auditable, consumable and reusable. The intent is to capture analytics project content based on a standard metamodel and expose that metadata for search and consumption from the Analytics Catalog services described in section 5.3.3.5. Project content included in the metamodel includes but is not limited to the following:

- Detailed Project Charter describing the Business Need and Scope
- Data Source Description and Details
- Description of Physical Data Structures and essential Business Context
- New or Existing Data Ingest Pipelines used
- Target Data Model against which Analytics were Performed
- Technical Validations Applied to data

- Business Rule Validations Applied to data
- Data Remediations Performed
- Data Transformation Rules Applied
- Algorithms and Analytics Techniques Applied
- Sequencing of Applied Techniques including Conditional Branching and Iterations

5.3.3.5 Analytics Catalog

The Analytics Catalog is an enterprise scoped, searchable repository of the Analytics Content described in section 5.3.3.4. The catalog is based on the Analytics Content metamodel and will classify Analytics Content based on consumer needs and apply a flexible system of tagging to ensure analytics resources can be located and reused as required. Where applicability scopes beyond a single project, the catalog will include data ingest pipelines, cleansing routines and other transformations.

5.3.3.6 Role Based Access Control (Authorization)

Role Based Access Control service uses the Information Asset Catalog service, the Data Access Rules service and supports the Query services, Service Publication Toolset services, Workflow Scripting, Execution Engine, and Service Endpoints to automatically capture and enforce data dissemination controls in compliance with policy and Steward determination of access restrictions. The RBAC uses a generalizable dissemination control model that minimizes the need for custom rules and exceptions.

5.3.3.7 Operational Metadata (Metrics)

Operational Metadata Metrics service uses the Analytical Processing Engine and the Reporting UI to maintain a real-time report on the status, usage, and performance of Data Assets, Analytical Assets, Workflows, and Data Products within the Air Force Data Services Reference Architecture.

5.3.3.8 Service Publication Toolset

Service Publication and Discovery service manages the maintenance, documentation, access rights, and Service Level Agreement (SLA) enforcement for data platform web services, both within the data platform and to External System Consumers. Traditionally, workflows between system components and services have been developed and maintained through “glue code” – often highly custom middleware. With the advent of open-source tools for CI (Continuous Integration) and for publication, subscription,and management of APIs glue code can be replaced with standardized, supported frameworks. The Service Publication and Discovery service needs to include the following specific Systematic Functional Standards:

- Ability to manage publication of APIs/Web Services
- Ability to manage consumption of APIs/Web Services
- Ability to enforce SLAs
- Ability to integrate with Role/User Management
- Ability to restrict service consumption to designated counterparts

- A set of tools to assist Developers with API publication and consumption that result in common and discoverable documentation and common interface configuration parameters
- A set of tools to facilitate consumer self-service discovery and provisioning of access to services, within the role-based access control framework
- Ability to integrate with Identity and Access Management for user/actor authentication

The Service Publication and Discovery service is implemented to reflect critical design principles:

- A common service for data platform web services, both internal to the data platform and external
- Standardized interface syntax and interface conventions, reflected in a standardized Interface
- Control Document (ICD) that is managed by the API Service Management tools
- Registration of all services in the Metadata Catalog
- Support exposure of service interfaces by all data platform service groups
- Reflect service interface design by restricting service interfaces where appropriate

The Service Publication and Discovery service scaling profile is fundamentally different from that of Storage and the Data Operations Framework in that it does not store or process data per se. The Service Publication and Discovery Function scaling profile is similar to the Job Management scaling: it supports discovery, authentication, and access to web services within and without data platform. The services could act as a proxy for web service access, however this has the potential to turn it into a bottleneck. Therefore, the Service Publication and Discovery Function interfaces with the Metadata Catalog to facilitate discovery, then authenticates service requests before handing requests off to the web service, and monitors service performance to identify SLA violations. The Service Publication and Discovery Function essentially functions in the same way as a domain name system server. Service Publication is implemented with acknowledgement of the following *constraining criteria*:

- Need to interface with existing Air Force enterprise authentication services
- Ability to interface with existing network monitoring tools
- Services are engineered to minimize transit across network boundaries
- Rely on existing FOSS tools wherever possible; defined as public access to source code and unrestrictive licenses
- Service is engineered to have multiple redundant Service Publication and Discovery nodes that support a directed, validated peer-to-peer service access profile – essentially acting as a *directory rather* than a *clearing house*

5.3.3.9 Information Asset Catalog

Information Asset Catalog provides a central documentation repository with pointers to Analytics, Queries, Reports, Indices, and Dashboards. The Catalog is intended to provide a user friendly faceted search capability to publish and discover Data Operations, Data Products, and other assets based on application metadata (what it is used for) and technical metadata (what it is).

The Information Asset Catalog is intended to provide this discovery function for Developers, Data Scientists, Business Analysts Consumers, and External System Call users. The Catalog includes the following Systematic Functional Standards:

- Ability to receive Application and Technical metadata from any data platform service and add this data to the Metadata Catalog
- A basic standardized metadata schema that conforms all metadata to a common data structure with standardized fields
- Ability to expose all metadata to users through the Querying service and Query UI service
- Ability to hand off service requests from the Metadata Catalog using technical metadata to the Job Management, Querying, API Service Management services for execution

The data platform Information Asset Catalog reflects from basic design principles that ensure minimum development effort and maximum flexibility and usability:

- The standardized metadata schema has a common data structure with standardized fields that can support extensible arrays for storing multiple activities
- The standardized metadata schema combines enumerated values for technical metadata, and hierarchical extensible values for application metadata
- Metadata collection and addition to the catalog are automated through a metadata collection API
- Metadata includes provenance history of the object
- Metadata supports all Data Products, Data Operations, and web services within the data platform
- Metadata is exposed both as a web service and through a graphical user interface
- Service and service interface design should allow an enterprise pub/sub metadata syndication model on the minimum set of enterprise metadata

The Metadata Catalog is implemented with acknowledgement of the following *constraining criteria*:

- The Metadata Catalog supports the CCB, Documentation and Review (D&R), and Documentation Only (DO) processes and therefore contains a complete transactional history of the registered objects
- The Metadata Catalog does not rely on any additional software components beyond the basic data platform build
- The Metadata Catalog is highly automated
- The Metadata Catalog leverages a standard, controlled metadata model that reflects the metadata principles of the Air Force Data Services Reference Architecture

5.3.3.10 Services Registry

The Services Registry maintains and manages access to services published through the publication service. The Registry service interfaces with the Information Asset Catalog to ensure that service metadata is current and used to manage service provisioning. For the EIM to serve as the SAF/CO defined “Data Storefront”, a Service Registry component is added to store information about services residing en-platform or elsewhere in the Air Force enterprise. Service Registry metadata is used for the selection, invocation, management, governance and reuse of web services and API exposing specific physical data assets for enterprise consumption. Services Registry entries are linked to the specific EIM registered information assets and data elements.

5.3.3.11 Data Lineage

Data Lineage provides visibility to details regarding the origins, movement, transformation and usage of a data asset from relative points of observation. Applicable details at a point of observation serve to establish visibility, understanding and trust to potential consumers of the data asset.

5.3.3.12 Enterprise Metadata Model

An instantiation of the Enterprise Metadata Model captures Administrative, Technical, and Operational/Contextual metadata for all Data Assets, Analytics Assets, Workflows, and Data Products in the Air Force Data Services Reference Architecture. The Enterprise Metadata Model is designed to be a “minimum necessary” extensible core schema that can support extension through the mapping of Business Glossaries to core logical entities within the Enterprise Metadata Model. The Enterprise Metadata Model will provide the schema for the Information Asset Catalog, the schema for service description in the Enterprise Metadata Catalog and capture of that data in the Service Publication Toolset service and supports the capture of Data Access Rules for use in the Role Based Access Control service. The model will be designed to support distributed metadata services to support specific mission needs where a central catalog is technically impractical, or a given functional group requires catalog extensions that are not shareable either due to mission or security requirements.

5.3.4 DATA PLATFORM FOUNDATION SERVICES

5.3.4.1 Instrumentation Services

Basic instrumentation services monitor system performance, provide real time visibility (dashboards) into system performance, and can provision additional resources for auto-scaling as needed to maintain SLA performance.

5.3.4.2 Global Data Service Activity Monitoring

Basic monitoring services monitor use of Data Asset, Data Product, Analytical Asset, and Workflow services within the system. This service provides real time visibility (dashboards) into product usage and consumption and supports the system governance and planning functions.

5.3.4.3 Service Level Agreement Monitoring

Basic monitoring services monitor SLA performance of services within the system. This service provides real time visibility (dashboards) into product usage and consumption and supports the system governance and planning functions.

5.3.4.4 Exception Management

Exception Management service maintains performance of services within the system and make the system fault tolerant.

5.3.4.5 Deployment Orchestration

Local Deployment Orchestration service, allowing for maintenance of services from within the platform using continuous integration or container technology. This supports the rapid importation and implementation of new services that comply with the interface specification, allows for the modernization and replacement of services with little to no downtime while maintaining quality, and enforces the system-wide use of version control and configuration management.

5.3.4.6 System Interface Management

Interface Management provisions access to data prior to ingestion through the messaging group. It includes access to both streaming and batch data, and the necessary skills include: transiting network boundaries, managing message queues, creating redundancy and failover, managing message topics, ensuring data integrity, and managing endpoint security. Because of the risk of negative systemic outcomes when system interfaces are modified, all Hosting and Environment Administration Functional groups are under full CCB control.

Interface Management in the data platform leverages external Legacy Interface Enablement, Services Publication and Discovery, Job Management, Messaging and Authentication services.

5.3.4.7 Job Management

Job Management service supports the queuing, distribution, and management of Data Operations; as scheduled, triggered, and ad hoc jobs. The objective of Job Management is to ensure that Data Operations jobs are managed in a way that maximizes performance, minimizes risk of malfunction, and provides a mechanism to manage equities in compute resources (priority, significance, complexity, programmatic need) relative to available compute resources. The Job Management service includes the following Systematic Functional Standards:

- Ability to define jobs using standard components, data objects, and workflows
- Support a graphical user interface for job definition, as well as job scripting
- Ability to manage jobs as objects
- Ability to assign metadata to jobs and add to catalog
- Ability to prioritize jobs based upon described equities
- Ability to launch jobs as objects from an object repository

Job Management, like any other component, has a set of specific design criteria and a design approach. In the case of Air Force MAJCOM/Functional Data Platforms, Job Management will execute jobs that are scripted using the Job Scripting Functional group (including the Job Scripting UI) in the Data Operations Functional group, will store the Job Scripts in the Storage Functional group, and will execute in the Data Operations Framework, using data platform services.

Jobs consist of executing data operations for streaming ingestion (including those based on event triggers and timers), batch ingestion, calculation and application of analytics as batch jobs, and publication of reports (including reports based on timers).

Job Management is implemented with acknowledgement of the following *constraining criteria*:

- Job Management interfaces with other MAJCOM/Functional Data Platform system components
- Job Management leverages a FOSS or COTS component and be readily supported
- Job Management is either a configured component within one of the Functional groups or a tool that can operate as a discrete service

5.3.4.8 Message and Job Queuing

Message Queuing service consumes streaming and batch data ingestion and place it onto a queue for consumption by Air Force MAJCOM/Functional Data Platform Data Operations Framework and application of Data Operations jobs. The objective of this Functional group is to support the message queuing, management, and onramp to the Data Operations Framework for Ingestion Processing. Message queuing/streaming services include the following Systematic Functional Standards:

- Asynchronous queue management
- Decoupled architecture (so Air Force MAJCOM/Functional Data Platforms do not rely on explicit modifications if the data platform has specific queue management standards or practices)
- Move correlation and ingestion processing farther back into Air Force MAJCOM/Functional Data Platforms and decouple data platform specific needs
- Logical separation of queue management and ingestion processing
- Delivery and ordering guarantees
- Supportable instance
- Supports multiple named message formats: XML, JSON, TXT, tuples, etc.
- Supports multiple physical consumption methods, including web services (restful services, SOAP), transfers (file transfer protocol, email), and manual placing of a file into a monitored directory

Current processing/correlation functions would be instantiated and managed within the Data Operations Framework, taking advantage of the redundancy, resiliency, and scaling features of the Data Operations Framework. Upon ingestion, the Data Operations Framework would simultaneously write a copy of the data to the Data Lake Storage Capability, pass the data to triggered QA/QC and Enrichment jobs, and add the Data Products created to the Metadata Catalog.

Message queuing is implemented with acknowledgement of the following *constraining criteria*:

- Design transits the network boundary and comply with security criteria

- Design supports non-auto-scaling
- Design supports multiple message formats

Batch upload onto the messaging bus consists of the services and APIs necessary to perform ad hoc Ingest, QA/QC, Indexing, and data storage of new, non-streaming data by Developers, Data Scientists, and Business Analysts. Batch upload is the ingestion interface for static files that are uploaded based on an event trigger (such as an analyst calling the upload service and adding data). Batch upload includes the following Systematic Functional Standards:

- Support for ad hoc data mark-up
- Data will become available for query immediately
- Does not disturb (pause, stop, etc.) streaming ingestion
- Batches may be sub-batched for optimal ingestion
- Batch ingest will be simple to perform and will provide feedback
- Batch upload will be an explicit service
- Batch upload exposes a service that can be called by a Business Analyst through a Graphical User Interface
- Batch upload support the following file formats: XML, JSON, TXT Flat Files in CSV format, XLS Flat Files, Raster, etc.

Batch upload is implemented with acknowledgement of the following *constraining criteria*:

- Design integrates with Identity and Access Management and Role / User Management at the User Interface/Web Service level
- Design transits the network boundary and comply with security criteria
- Design supports multiple scaling models

5.3.4.9 Entity Management

Entity Management service provides a collection of scripted jobs applied to Data Resources on Ingest in the Data Operations Framework to ensure that each record is resolved to the correct identity, and that identity is indexed to multiple appropriate identifiers based on perspective. The objective of entity management is to create a central “correlation” service that allows System Actors to query records from a single authoritative source and use pre-calculated normalization schemes. So, while all Data Products do not have to use the same normalization scheme, Data Products are created using a common set of defined correlation and normalization schemes, reducing discrepancies and saving Business Analyst time. Entity Management includes the following Systematic Functional Standards:

- Ability to correlate to multiple keys simultaneously
- Ability to cross-walk seamlessly between different key/code systems
- All entities indexed at time of resolution and added to Query Functional group services
- Multi-key query capability (ability to query for a given entity using multiple key systems)

- Ability to manage reference code sets / master keys
- Ability to automatically resolve at least a basic portion of the metadata standards of the EIM
- Entities inherit contextual metadata from the Metadata Catalog

Entity Management incorporates the following design principles into the final design:

- All correlation and entity resolution are processed at ingest and included in data written to a Data Product in the Data Storage Functional group
- Entity resolution jobs are created using Job Scripting and controlled through Job Management using event triggers when new records are ingested into Air Force MAJCOM/ Functional data platforms
- Users have the ability to query data using a specific, documented correlation and normalization schema
- Air Force MAJCOM/Functional data platform Data Scientists and Business Analysts are able to perform on-the-fly normalization of records through query specification of correlation keys instead of data transformation and using contextual metadata
- Air Force MAJCOM/Functional data platform Entity Resolution includes QA/QC rules that perform completeness checks to ensure that all records needed to normalize an entity are present

Entity Management is implemented with acknowledgement of the following constraining criteria:

- Entity resolution most occur at ingest and demonstrate improved completeness, accuracy, and reduced Business Analyst time compared to the current system (Air Force MAJCOM/ Functional data platform reports have validated accuracy to existing reports that is equivalent or better, as measured by a concordance analysis with existing report objects)
- Multiple primary and secondary keys are supported
- Entity resolution is consistent, and provide alerts when incomplete record sets are received

5.3.5 EXTERNAL INFRASTRUCTURE SERVICES

External Services are those services that Air Force MAJCOM/Functional Data Platforms can access and / or consume to maintain availability to all Air Force users, meet uptime and availability standards, ensure compliance with hosting and networking standards, and ensure compatibility with legacy systems and remote hosted interfaces.

5.3.5.1 Enabled Legacy Interfaces

To achieve the full potential of Air Force Data Services Reference Architecture, the system can consume data as a service from existing legacy systems and expose services that allow it to return data products to those systems. It serves as a methodology and toolkit to support distal service retrofitting of legacy systems. This ensures that these systems can be published to the service registry and have current and accurate metadata. By enabling automatically access to the Air Force Data Services Reference Architecture and its users, will also enforce Role Based Access Control.

5.3.5.2 Authentication

Identity and access management (IDAM) consists of the suite of services necessary to support authentication of system actors; and management of authorization to use Air Force MAJCOM/ Functional Data Platforms system components as externally callable services. IDAM supports system actor authentication to each of the Capability Groups. IDAM provides:

- Ability to identify a named user or external system actor and authenticate identity through a single service
- Focus on the authentication side of access management, while the SAF/CO EIM focuses on the authorization side

Role / User Management (RUM) provides a service to manage the Roles and Identities of system actors. RUM provides:

- Interface with IDAM to associate a named user or external system actor with one or more defined Air Force MAJCOM/Functional Data Platform actors/roles
- Inherit and apply role properties to read/write permissions on various Air Force MAJCOM/ Functional Data Platform services

5.3.5.3 Compute/Storage/Networking

Hosting and Environment Administration represents those skills necessary to maintain the basic hosting environments for Air Force MAJCOM/Functional Data Platforms that ensure the specified performance for Air Force MAJCOM/Functional Data Platforms. This includes system monitoring, managing environment scaling (in terms of compute and storage), deployment of virtual machines, managing user authentication, roles, and authorizations, maintenance of Framework Code, and providing the test, control, code and configuration management procedures necessary to ensure consistent Air Force MAJCOM/Functional Data Platform performance and the artifacts necessary to support root cause analysis in the event of system issues. Because of the risk of negative systemic outcomes when changes are made to the hosting environment, all Hosting and Environment Administration Functional groups are under full CCB control.

System Monitoring consists of the components and applications necessary to monitor system performance, support dynamic scaling of Air Force MAJCOM/Functional Data Platform Functional components, and provide metrics of system performance to ensure compliance with Air Force MAJCOM/Functional Data Platform objectives and success criteria. Strength Maintenance Management System (SMMS) provides:

- Logged software exceptions for Air Force MAJCOM/Functional Data Platform components
- Logged compute and storage load, capable of supporting load forecasting
- Scheduling to deploy, initialize, and shut down component instances

Finally, Air Force MAJCOM/Functional Data Platform hosting includes the organizational activities associated with maintaining accreditation for a Platform Implementation instance.

The network Functional group represents those skills necessary to maintain Transmission ControlProtocol/Internet Protocol connectivity between Air Force MAJCOM/Functional Data Platform services, ensure connectivity to external services and consumers, ensure compliance with network boundary standards, and ensure necessary data transport between network

enclaves. Because of the risk of negative outcomes when changes are made to the networking environment, all Network Functional groups are under full CCB control.

Network Access and connectivity encompasses both connectivity between servers/services within Air Force MAJCOM/Functional Data Platforms, connection to external Data Resources, exposing web services for connection by external systems, and providing web-based user interfaces.

5.3.5.4 DevOps/Code Management

The DevOps Functional group is focused on the management and deployment of code to Air Force MAJCOM/Functional Data Platform core functions for deployment, maintenance, patching, and upgrades. DevOps is focused on maintenance of the codebase for Air Force MAJCOM/Functional Data Platform frameworks, rather than configurations loaded within them. DevOps support the CCB governed activities within Air Force MAJCOM/Functional Data Platform, while configurations are supported within the Metadata Catalog and the D&R and DO processes. DevOps provides:

- Code repositories, including collaborative code management and integration
- Tools for builds and dependency/ variable management
- Automated deployment tools
- Configuration management

5.3.5.5 Platform Security

The platform undergoes review and complies with applicable security controls per Air Force policy.

5.3.5.6 Audit/Monitor

The platform will be reviewed, audited, and monitored regularly to ensure compliance with security controls and Air Force policy.

5.4 Logical Interface Patterns

The Logical Interface Patterns are intended to describe the recommended standards and practices for Interfaces and Data Flows between Functional groups to support documentation and specification of these interfaces. All interfaces will be exposed as general interfaces for consumption by other system components.

The objective is to break Air Force MAJCOM/Functional Data Platform services and development down into general functional interfaces that can service multiple service-to-service interactions; interfaces are unit tested, and then individual workflows are tested for compliance.

Table 1 below describes an example interface standards table, including proposed connections (within a loose coupling model) between Air Force MAJCOM/Functional Data Platform Functional groups. These interface standards were developed by extracting the design patterns and interface diagrams from the Logical Architecture Description and represents an initial listing of published Air Force MAJCOM/Functional Data Platform interfaces. A solution implementation of Air Force Data Services Reference Architecture will include a documented interface standards table.

TABLE 1: AIR FORCE MAJCOM/FUNCTIONAL DATA PLATFORM EXAMPLE INTERFACE TABLE FOR PRECONFIGURED WORKFLOWS

FUNCTIONAL GROUP	SERVICE INTERFACE STANDARDS	CONSUMES FROM	PUBLISHES TO
Message Queuing	1) Mirror Active MQ topics	External Sources	Analytical Processing Engine through Active MQ
Batch Upload	2) Job Management Call 3) Push to Analytical Processing Engine	Analytics Scripting UI File Source	Job Management Analytical Processing Engine
Job Management	4) API Call Function 5) General Record Write Function 6) External Control Interface	Job Scripting Querying API Service Management	Storage Metadata Catalog Analytical Processing Engine
API Service Management	7) Consume IDAM Credentials 8) API Management Publication Interface 9) API Management Subscription Interface 10) General Record Write Function	IDAM All Publishing Services	Metadata Catalog All Consuming Services
Entity Management	11) Entity management queries 12) Entity management job scripts 13) General Record Write Function	Storage	Job Management Metadata Catalog Analytical Processing Engine
Metadata Catalog	14) Record write function 15) Record read function	Indexing Querying Job Scripting Analytical Processing Engine Data Visualization and Dashboarding Tool	Storage Analytics Scripting UI Data Visualization and Dashboarding Tool Querying/Query UI Indexing Job Scripting
Storage	16) Record write function 17) Record read function 18) Indexing function	Analytical Processing Engine Metadata Catalog	Querying Indexing

FUNCTIONAL GROUP	SERVICE INTERFACE STANDARDS	CONSUMES FROM	PUBLISHES TO
Querying	19) Remote query function 20) Store query as service 21) Record write function 22) Data synchronization with storage	Storage	Query UI Analytics Scripting UI Data Visualization and Dashboarding Tool Metadata Catalog
Indexing	23) Create index from storage 24) Expose index as function	Storage	Analytics Scripting UI Querying UI Metadata Catalog
Data Operations Framework	25) Accept jobs 26) Consume topics 27) Retrieve records from Querying 28) Retrieve records from Storage 29) Record write function	Job Management Storage Querying Message Queuing Batch Upload	Storage Metadata Catalog
Job Scripting	30) Execute Analytics Libraries 31) API Call Function 32) Record write function 33) User Interface	Analytics Libraries Metadata Catalog	Job Management Storage Metadata Catalog
Analytics Librarie	34) Expose to Job Scripting	Storage	Job Scripting
Analytics Scripting UI	35) Expose Job Scripting	Job Scripting	Job Scripting
Data Visualization and Dashboarding Tool	36) Query retrieval 37) Record write function	Querying	Storage Metadata Catalog

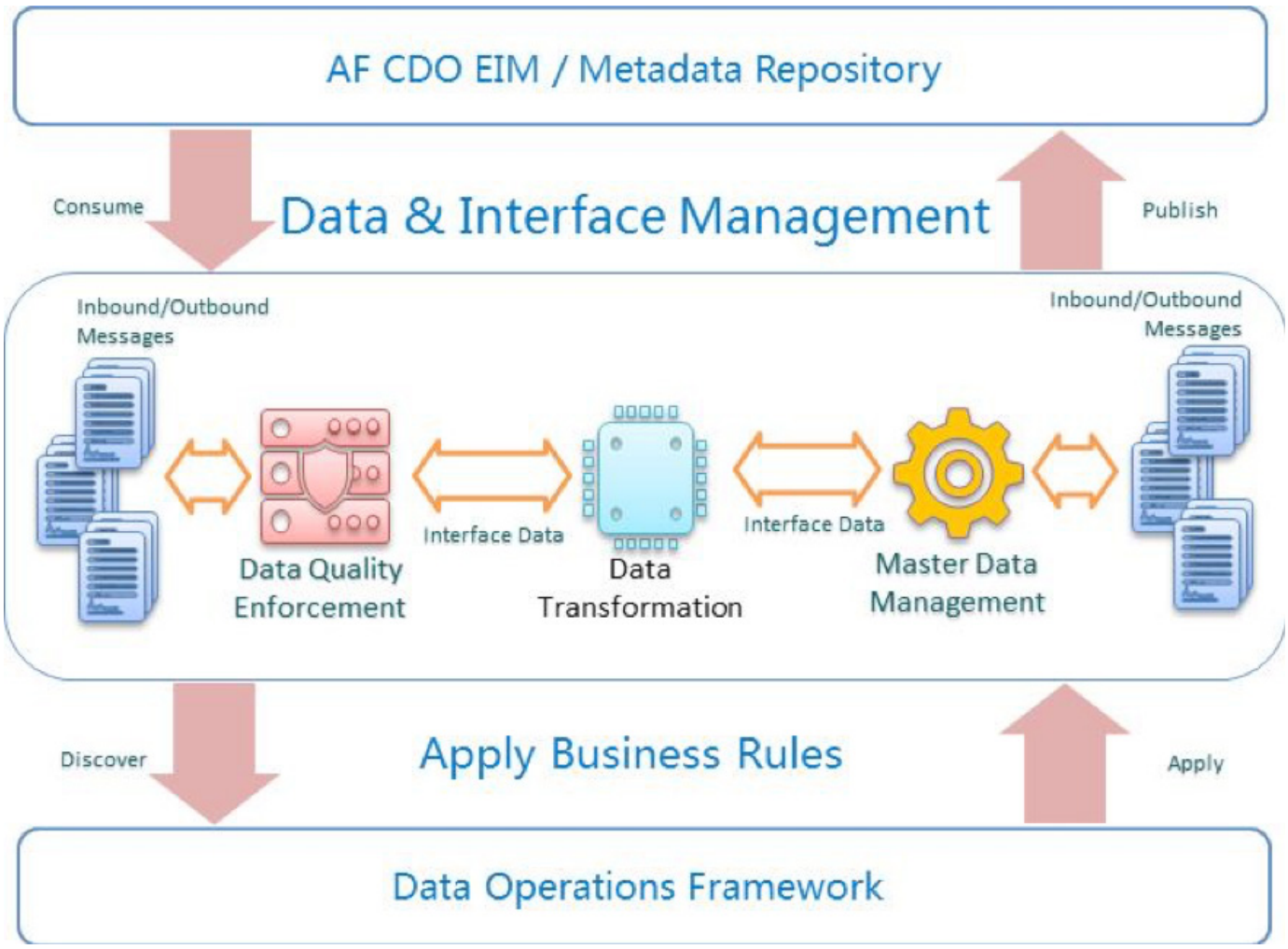
6 EXAMPLE USE OF THE REFERENCE ARCHITECTURE SOLUTION

The sections below describe appropriate use models of the Air Force Data Services Reference Architecture Reference Architectures as templates for a given Platform Implementation. These discussions focus on the relationship between interface and data management, the data lifecycle, and the development and publication of data products.

6.1 Data Management and Interface Management

Data Management and Interface Management are critical components of the Air Force Data Services Reference Architecture, and support and interact with each other. Interfaces provide data, from legacy systems and from the data lake, to which business rules are applied that perform quality assurance and curation on data to create curated, authoritative data products for use in analysis and the ultimate creation of other data products. As such, interface management, business rules management, and metadata management are critically interrelated. Figure 1 below describes how the functions of interface management and data management operate together, leveraging the Metadata Catalog and Data Operations Framework.

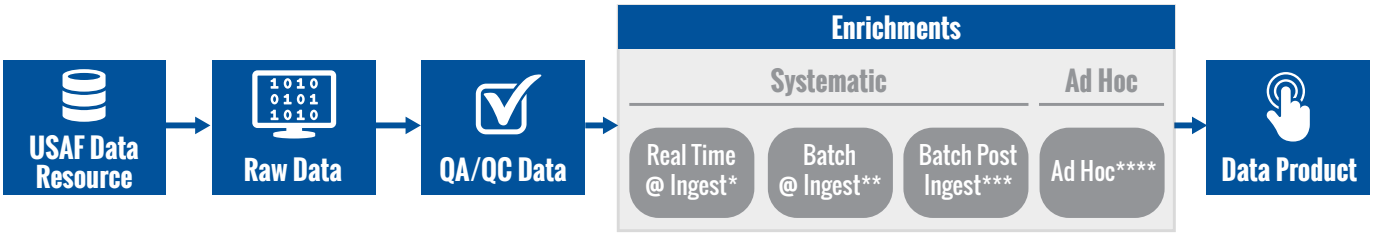
FIGURE 3: AIR FORCE DATA QUALITY AND MANAGEMENT PROCESS ENABLED THROUGH INTERFACE MANAGEMENT, DATA OPERATIONS, AND METADATA



6.2 Data Product Lifecycle and Operations

Data and analytic assets will collocate within an Air Force MAJCOM/Functional Data Platform at various stages of curation and publication. Understanding the concepts of a Data Product and a Data Product Lifecycle will ensure that quality outputs are provided to an end user.

FIGURE 4: AIR FORCE MAJCOM/FUNCTIONAL DATA PLATFORM - DATA PRODUCT LIFECYCLE AND LOGICAL FLOWS



Existing, external data sources that are ingested into AF Data Vault.

Data assets that have crossed the AF Data Vault system boundary, without modifcation.

Raw Data checked and transformed to ensure that it meets USAF Data Quality Standards, published as a data product in USAF Data Vault.

Enrichments are additional data products created from QA/QC data, such as metafields added to records, enhances data products that provide additional information, or calculated outputs of business and analytical routines.

Systematic enrichments are conducted automatically as de facto standard system processes and support authorization source data, data quality management, and authorization data products.

Data Products are any documented product container raw or enriched data that is provided to an end user: reports, dashboards, visualizations, or simply enriched data itself.

* Added in-line as QA/QCed data is ingested into USAF Data Vault
** Added as batch processes concurrent to ingest
*** Added as batch processes anytime subsequent to ingest
**** Batch enrichments added in a non-systematic manner

TABLE 2: AIR FORCE MAJCOM/FUNCTIONAL DATA PLATFORM DATA PRODUCT LIFECYCLE MAPPED TO SYSTEM ACTORS, LOGICAL PARTITIONS, AND FUNCTIONAL GROUPS

DATA STAGE	LOGICAL LOCATION	ACTORS	FUNCTIONAL GROUPS RECOMMENDED TO EXECUTE
Data Resource	External to Air Force MAJCOM/Functional Data Platform	Infrastructure Developer	Data Access: Message queuing and Batch Upload Metadata Catalog (Application Information / Admin) Network Access
Raw Data	Raw Data Product in the Metadata Catalog	Infrastructure Developer Data Scientist External Data Stream External Data Push Timed External Data Pull	Message queuing Batch Upload Data Operations Framework Metadata Catalog Job Management Storage Indexing

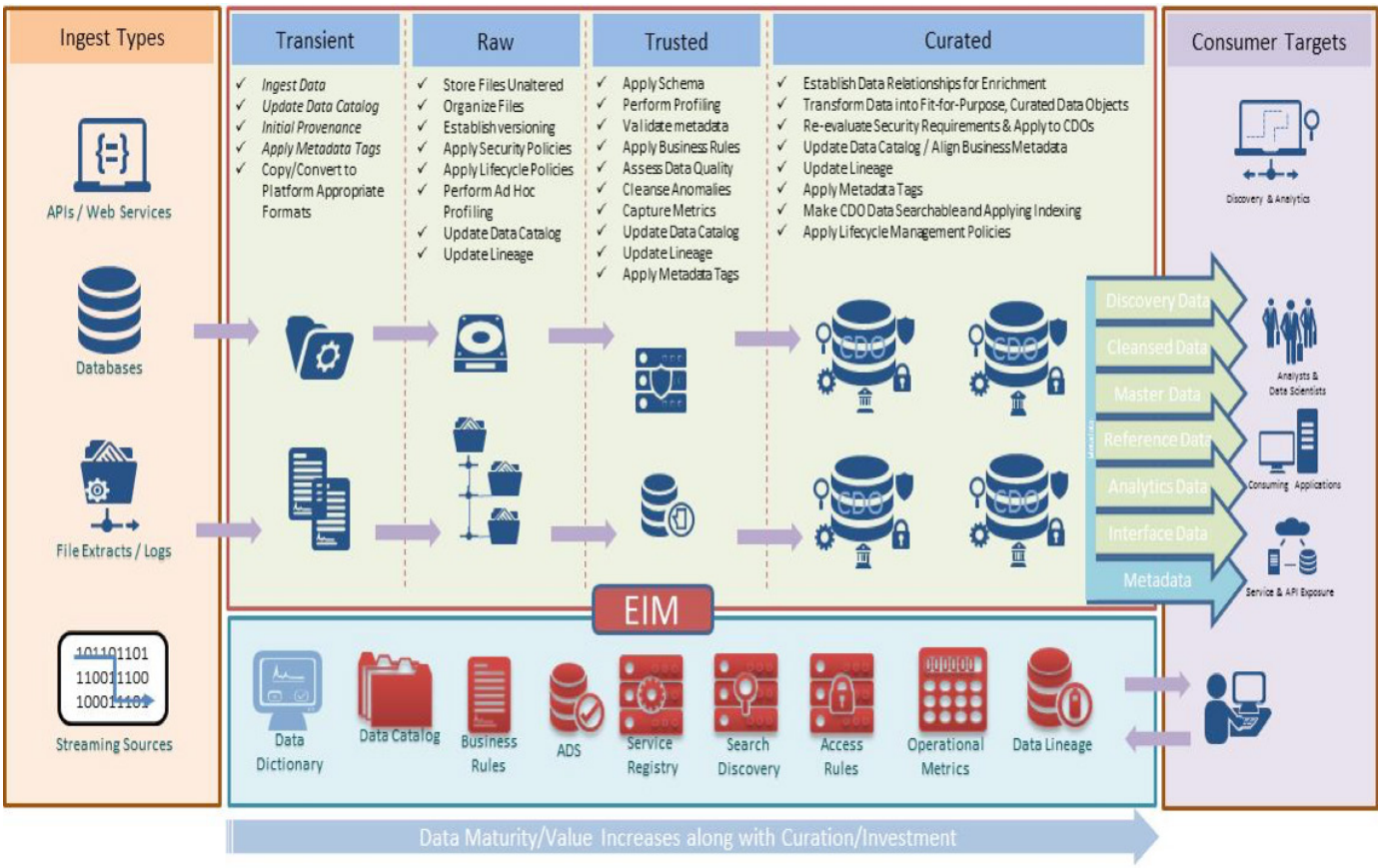
DATA STAGE	LOGICAL LOCATION	ACTORS	FUNCTIONAL GROUPS RECOMMENDED TO EXECUTE
QA/QCed Data	QA/QC Data Product in the Metadata Catalog	Developers Data Scientist Business Analysts Internal AF MAJCOM/ Functional Data Platform Data Operations Timer Internal AF MAJCOM/ Functional Data Platform Data Operations Event Trigger	Message queuing Batch Upload Data Operations Framework Metadata Catalog Job Management Storage Indexing Query Job Scripting Entity Management API Service Management Analytics Libraries
Enriched Data	Enriched Data Product in the Metadata Catalog	Data Scientist Business Analyst Internal Air Force MAJCOM/Functional Data Platform Data Operations Timer Internal Air Force MAJCOM/Functional Data Platform Data Operations Event Trigger External System Call	Data Operations Framework Metadata Catalog Job Management Storage Indexing Query Query UI Job Scripting Analytics Scripting UI Entity Management API Service Management
Metrics	Stored Metadata and Indices	Business Analyst External System Call	Metadata Catalog Job Management Storage Indexing Query Query UI Entity Management Data Visualization and Dashboarding Tool

DATA STAGE	LOGICAL LOCATION	ACTORS	FUNCTIONAL GROUPS RECOMMENDED TO EXECUTE
Reports	Published as a Data Product Service	Business Analyst External System Call	Metadata Catalog Job Management Storage Indexing Query Query UI Entity Management Data Visualization and Dashboarding Tool
Visualizations	Stored Queries and Visualization Configurations	Business Analyst External System Call	Metadata Catalog Job Management Storage Indexing Query Query UI Entity Management Data Visualization and Dashboarding Tool

6.3 Development and Publication of Data Products

The Capability and Functional groups within the Air Force Data Services Reference Architecture allow for users to leverage these microservices, in the context of the Data Product lifecycle, to create value-added data products.

FIGURE 5: AIR FORCE MAJCOM/FUNCTIONAL DATA PLATFORM - DATA PRODUCT CREATION PROCESS



7 DOCUMENTATION PATTERNS

Air Force MAJCOM/Functional Data Platform Design patterns are intended to provide standard guides for service description and implementation. They are intended to provide a Minimum Necessary standard for interface, component, and workflow implementations to ensure consistency and interoperability.

7.1 Interface Metadata Standards

All interfaces in Air Force MAJCOM/Functional Data Platforms comply with a minimum standard for design, description, and publication. It is the intent for Air Force MAJCOM/Functional Data Platforms to be able to support multiple modes of interface discovery, including discovery of interfaces through the Metadata Catalog, through API Service Management, or review of code repositories. To ensure that APIs are more easily discoverable, reusable, and extensible, all APIs/Web Services developed / published by Air Force MAJCOM/Functional Data Platforms are documented using a World Wide Web Application Description Language (WADL) document to

provide the consumer with clear expectations of server side behavior. Table 3 below details the recommended classes and sub-classes for Air Force MAJCOM/Functional Data Platforms web services.

TABLE 3: AIR FORCE MAJCOM/FUNCTIONAL DATA PLATFORMS DRAFT ICD DESCRIPTORS

CLASS SUB CLASS	STRUCTURE	DESCRIPTION
Application Description Schema	String	Reference WADL standard
Resource URI	String	Describes URI of the web service
Technical	Superclass	Supports multiple technical objects
Authentication	String	Authentication protocol
Service Level	Array	SLA restrictions on the server side
Request/Command	Array	Request methods supported
Parameters	String	Individual method parameters
Parameter Format	Array	Describes format, structure, and type of data inputs recommended for the parameter
Responses	Array	Response methods supported
Response	String	Individual response parameters and definitions
Response Format	Array	Describes format, structure, and type of data response
Application	Superclass	Describes contextual application metadata per the Metadata Catalog schema
Programmatic Area	Array	Describes all programmatic application areas (e.g., Metrics) supported
Entities Supported	Array	Describes all standard entity indexing schemes supported by this interface

7.2 Product Metadata Model Patterns

To support discovery of all objects in the Air Force MAJCOM/Functional Data Platforms system (Web services, Data Products, Scripted Analytics, Queries, Indices, Dashboards), Air Force MAJCOM/Functional Data Platforms will support a Metadata Catalog integrated with the Query Functional group. The Metadata Catalog is a specific Metadata Index Pattern intended to facilitate discovery. Table 4 describes the classes of metadata that are captured on each object in the system.

TABLE 4: AIR FORCE MAJCOM/FUNCTIONAL DATA PLATFORMS GENERAL METADATA CLASSES

CLASS SUB CLASS	STRUCTURE	DESCRIPTION
Application Description Schema	String	Reference AF MAJCOM/Functional Data Platforms metadata schema
Technical	Superclass	Supports multiple technical objects
Description	String	Description of resource
Resource URI	String	Describes URI of the object
Access Restrictions	Array	Lists Air Force MAJCOM/Functional Data Platforms system actors not permitted to access the resource
Resource Type	String	Lists the object type from an enumerated list (Data Operation, Data Product, Query, Dashboard)
Application Metadata	Superclass	Superclass for metadata describing the programmatic application of the object
Metrics	Array	Lists all applicable metrics supported by the object, as enumerated
Report	Array	Lists all applicable report types supported by the object, as enumerated
Analytic Type	Array	Lists all applicable analytic types supported by the object, as enumerated
Responses	Array	Response methods supported
Provenance	Superclass	Describes the pedigree of the object
Stewardship	Array	Names the steward of the object, along with contact data
Provenance Activities	Array	Describes the history of actions on the object
Data Lifecycle	Superclass	Describes the relationship of the object to the AF MAJCOM/Functional Data Platforms data lifecycle
State	Array	Describes the state of the object upon completion, from an enumerated list
Data Format, Structure, Type	Array	Describes the data formats, structures, and types of inputs and outputs
Entity Metadata	Entity Metadata	Describes enumerated entity code schemes supported by the object

Enumerated lists will be developed during Air Force MAJCOM/Functional Data Platforms development based upon the Use Case and Data Discovery activities.

7.3 Reference Database Architectures

In current database architectures, there are six principal categories: relational, document, graph, keyvalue, hybrid and geospatial:

- Relational stores use a traditional table/key storage model with a schema used to describe relationships between primary and foreign keys across tables. Relational stores use Structured Query Language (SQL) for programming and configuration and are usually optimized for highly structured data and non-hierarchical data that does not rely on having large collections of binary objects. Per NIST Special Publication 1500-1 Guidelines for unstructured data, relational designs are often ill-suited for rapidly changing data sources and unstructured data because of the effort necessary to revise schema.
- Document stores are the most commonly used of the “noSQL” data store options and are designed to store records as discrete documents rather than as rows in a table; schema in a document store relates to document identifiers (such as Global Unique Identifiers, or GUIDS), field typing and standards, and field indexing. Document stores tend to be more flexible than relational stores, as they are optimized to store document collections, however they do not easily support many traditional relational data operations and do not support transactional systems well. While document stores are reasonable well suited to storing and retrieving large binary objects, they are not the most scalable solution for large collections.
- Graph stores are a highly specialized type of store for reading and writing graph structures (vertices and edges) at high speed. Because graph structures are uniquely different from other data structures and often have high computational complexity, graph stores are optimized for this type of data structure. A subset of graph stores is Resource Description Framework (RDF) stores, with specific functions designed to store data in RDF format. Graph stores are not designed to store large binary objects, and do not readily support easy storage, indexing, and retrieval of documents or relational (tabular) data.
- Key-value stores are probably the most scalable and flexible of all the data store types, supporting the native storage of documents, records, and large binary objects. However, keyvalue stores are designed and optimized to handle extremely large data sets, and as a result, lag relational stores in transactional speed, document stores in ease of use, and lack many of the basic functions and ACID compliance that are provided in relational and document stores. Keyvalue stores necessitate a high level of configuration, often have a specialized hosting model, and therefore have a relatively high support burden.
- Hybrid stores are designed to combine the features of one or more of the other data store types, such as key-value and relational, or document and relational. Hybrid stores tend to scale well and provide a readily supportable platform. However, because they are not optimized to a specific data structure, hybrids will not maximize performance in any given category. Further, while they enjoy more ready features than key-value stores, a relatively focused Use Case has resulted in a smaller support community.

- Geospatial storage and operations are designed to combine features to replace master content stores. The consolidation of this data eliminates duplication of effort and reduces major costs of deployment and data maintenance. The data is stored once and maintained in a central repository and can be used by several applications. These operations are necessary, so Air Force MAJCOM/Functional Data Platforms can utilize geospatial functions in relational or document stores.

Air Force MAJCOM/Functional Data Platforms architecture would indicate that a combination of relational and document stores is most appropriate for Air Force MAJCOM/Functional Data Platforms, given data inventory and constraining criteria. Given a lack of many large binary objects, nor the programmatic need to represent graph structures, key-value, graph, and hybrid stores are not appropriate data stores for Air Force MAJCOM/Functional Data Platforms. To the extent that geospatial storage and operations are necessary, Air Force MAJCOM/Functional Data Platforms can utilize geospatial functions in relational or document stores.

7.4 Data Operations / Analytics Design Patterns

Data Operations created in the Data Operations Framework will perform operations on data moving between all Air Force MAJCOM/Functional Data Platforms data lifecycle states. Within Air Force MAJCOM/Functional Data Platforms, there are five basic analytics patterns, which can be applied to create data products at various stages in the Air Force MAJCOM/Functional Data Platforms data lifecycle:

- Basic Data Transformations: Provides simple math operands, joins, and selects based upon the data sample
- Descriptive Analysis: Provides a statistical summarization to describe what is happening or has happened based upon the data sample
- Diagnostic Analysis: Provides defined analytical rules to describe what is happening or happened, and suggestions as to why, given system context
- Predictive Analysis: Prioritizes statistical analysis and modeling to estimate an outcome or future value, includes machine learning
- Prescriptive Analysis: Prioritizes or recommends actions based on a combination of statistical modeling and contextual business rules

TABLE 5: AIR FORCE MAJCOM/FUNCTIONAL DATA PLATFORM ANALYTICS PATTERNS

DATA LIFECYCLE STATE	ANALYTIC APPLIED TO THIS STATE	TRANSITIONS TO
Data Resource		
Streaming/Queuing	Basic	QA/QC Intermediate Results
Batch	Basic Descriptive	QA/QC Intermediate Results
Raw Data	Basic	QA/QC
QA/QC Data Product	Basic Descriptive Diagnostic Predictive	Intermediate Results Basic and Complex Enrichment Analytic Output Metrics
Enriched Data Product: Intermediate Results	Basic Descriptive Diagnostic Predictive	Basic and Complex Enrichment Analytic Output Metrics Reports Visualizations
Enriched Data Product: Basic and Complex Enrichment	Basic Descriptive Diagnostic Predictive Prescriptive	Metrics Reports Visualizations
Enriched Data Product: Analytic Output	Basic Descriptive Diagnostic	Metrics Reports Visualizations
Metrics; as tagged Data Products	Basic	Metrics are an end-state and are not transformed
Reports; as Queries using Metrics and other Data Products	Basic	Reports are an end-state andare not transformed
Visualization; as a graphical representation of a report	Basic	Visualizations are an end-state and are not transformed

8 USE CASE WORKFLOW IMPLEMENTATION TEMPLATE

In implementing Air Force MAJCOM/Functional Data Platforms Programmatic Use Cases, Data Scientists and Business Analysts can leverage an implementation template – initially as a checklist but eventually as a workflow. The initial checklist is described in this section.

1) What programmatic need does this use case satisfy?

- a) What are the programmatic standards and best practices?
- b) What are the programmatic constraints?
- c) What are the programmatic success criteria?
- d) What information is needed to create a clearly defined problem statement?

2) Data onboarding: Identifying and adding a resource to Air Force MAJCOM/Functional Data Platform

- a) Which resource am I onboarding?
- b) Does it support the Use Case?
- c) Does it already exist in the Metadata Catalog?
- d) Is it streaming or batch?
  - i) If streaming, it necessitates a developer for onboarding
  - ii) If batch, it necessitates a Data Scientist
- e) Which Entity Management code sets are already represented in the data?

3) Data/Metadata Management: Ensuring the Data Products of the Use Case are in proper context

- a) What applications does the workflow support?
- b) What entities are identified as vital?
- c) What Data States are involved in the Use Case?
- d) What level of configuration management is necessary?

4) Analytics Configuration: What Data Operations are necessary for the Use Case?

- a) Type of Analytics needed at each Data State?
- b) Data States input / output standards?
- c) Which Data Products do these Analytics consume?
- d) Do these analytics already exist within Air Force MAJCOM/Functional Data Platform?
- e) Which stakeholders will be consuming the resulting data product?

f) What analytics libraries are needed? Are they already supported in The Air Force MAJCOM/ Functional Data Platform?

g) What level of configuration management is necessary?

5) Report design, generation, and tagging: creating queries for use by visualizations and other Air Force MAJCOM/Functional Data Platform System Actors

a) What types of reports / queries are needed in order to enable the Use Case?

b) Do these queries already exist within the Air Force MAJCOM/Functional Data Platform?

c) Which Data Products do these queries/reports consumer?

d) What applications do these queries support?

e) Which stakeholders will be consuming the reports and queries?

f) What level of configuration management is necessary?

6) Business Intelligence: creating final visualizations for presentation to users

a) What types of dashboards are needed?

b) Do these dashboards already exist within the Air Force MAJCOM/Functional Data Platform?

c) Which reports or queries do the dashboards need to consume?

d) What applications do these dashboards support?

e) Which stakeholders / users will be consuming the dashboards?

f) What level of configuration management is necessary?

9 APPENDIX 1 - GLOSSARY OF REFERENCES AND SUPPORTING INFORMATION

9.1 References

AFI 17-140, Architecting, dated 29 June 2018

AFI 33-322_AFGM2018-01, Records Management Program, dated 26 April 2018.

DoD Direction 5000.01 The Defense Acquisition System, dated November 2007. http://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/500001p.pdf

DoD Direction 5000.02: Operation of the Defense Acquisition System, August 2017 http://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500002_dodi_2015.pdf

DoD Instruction 5200.01: DoD Information Security Program and Protection of SCI, dated June 2011. http://www.dtic.mil/whs/directives/corres/pdf/520001p.pdf 24

DoD Manual 5200.01 Vol 1: DoD Information Security Program: Overview, Classification and Declassification, dated February 2012. http://www.dtic.mil/whs/directives/corres/pdf/520001_vol1.pdf

DoD Manual 5200.01 Vol 2: DoD Information Security Program: Marking of Classified Information, dated March 2013. http://www.dtic.mil/whs/directives/corres/pdf/520001_vol2.pdf 26

DoD Manual 5200.01 Vol 3: DoD Information Security Program: Protection of Classified Information, dated March 2013. http://www.dtic.mil/whs/directives/corres/pdf/520001_vol3.pdf

DoD Manual 5220.22 Manual: National Industrial Security Program: Operating Manual (NISPOM), dated March 2013. http://www.dtic.mil/whs/directives/corres/pdf/522022m.pdf 23

DoD Manual 8310.01 Manual: Information Technology Standards in the DoD, 2 February 2015

DoD Directive 8320.02, “Data Sharing in a Net-Centric Department of Defense,” 5 August 2013

DoD Directive 8320.03, “Unique Identification (UID) Standards for a Net-Centric Department of Defense,” 4 November 2015

DoD Directive 8330.01 Interoperability of Information Technology (IT), Including National Security Systems (NSS), 21 May 2014

DoD Instruction 8500.01: Cybersecurity, dated 14 March 2014. http://dtic.mil/whs/directives/corres/pdf/850001_2014.pdf 17

DoD Instruction 8510.01: Risk Management Framework (RMF) For DoD Information Technology (IT), dated 12 March 2014. http://dtic.mil/whs/directives/corres/pdf/851001_2014.pdf

DoD Cloud Computing Security Requirements Guide (SRG), 6 March 2017

DoD Chief Information Officer, Updated Guidance on the Acquisition and Use of Commercial Cloud Computing Services, 15 December 2014. http://iase.disa.mil/Documents/commercial_cloud_computing_services.pdf 16

NIST SP 500-292: NIST Cloud Computing Reference Architecture, dated September 2011.http://www.nist.gov/customcf/get_pdf.cfm?pub_id=909505

NIST SP 800-53: Recommended Security Controls for Federal Information Systems and Organizations, Revision 4, dated April 2013.
http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf
Note: http://csrc.nist.gov/publications/PubsSPs.html contains additional documents relating to SP 800-53.

NIST SP 800-59: Guideline for Identifying an Information System as a National Security System, dated August 2003. http://csrc.nist.gov/publications/nistpubs/800-59/SP800-59.pdf

NIST SP 800-66, Revision 1: An Introductory Resource Guide for Implementing the Health Insurance Portability and Accountability Act (HIPAA) Security Rule, dated October 2008.
http://csrc.nist.gov/publications/nistpubs/800-66-Rev1/SP-800-66-Revision1.pdf

NIST SP 800-88, Revision 1: Draft: Guidelines for Media Sanitization, dated September 2012.
http://csrc.nist.gov/publications/drafts/800-88-rev1/sp800_88_r1_draft.pdf

NIST SP 800-122: Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), dated April 2010. http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf

NIST SP 800-144: Guidelines on Security and Privacy in Public Cloud Computing, dated December 2011. http://csrc.nist.gov/publications/nistpubs/800-144/SP800-144.pdf

NIST SP 800-145: The NIST Definition of Cloud Computing, dated September 2011.
http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

NIST SP 800-37, Revision 1: Guide for Applying the Risk Management Framework to Federal Information Systems, dated February 2010.
http://csrc.nist.gov/publications/nistpubs/800-37-rev1/sp800-37-rev1- final.pdf

NIST SP 1500-1, Big Data Interoperability Framework: Volume 1, Definitions, dated September 2015. https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.1500-1.pdf

International Organization for Standardization (ISO) 15704-2000: Industrial Automation Systems— Requirement for Enterprise-Reference Architectures and Methodologies, Published 2000-06

2009. The DAMA Guide to the Data Management Body of Knowledge - DAMA-DMBOK. Technics Publications, LLC, USA.

2014. The Data Maturity Model (DMM). Capability Maturity Model Integration (CMMI). USA

Federal Information Security Modernization Act (FISMA), dated 18 December 2014
https://www.whitehouse.gov/wp-content/uploads/2017/11/FY2017FISMAReportCongress.pdf

9.2 User Classes and Characteristics

These defined terms will be used throughout the Air Force MAJCOM/Functional Data Platform Reference Architecture.

TABLE 6: AIR FORCE MAJCOM/FUNCTIONAL DATA PLATFORM 2.0 SYSTEM ACTORS

ACTOR	DEFINITION
Infrastructure	Those technical personnel who maintain the core technology infrastructure, including hosting, base software component maintenance, implementation of coded data operations and installation of packages / capabilities – comfortable with operating at the command line.
Developer	Leverage the infrastructure to add capabilities; those who extend the basic infrastructure capabilities and add new base operations including support for moving data enrichment into the data ingestion topologies – comfortable operating at the command line and in an Integrated Development Unit: may also provide Data Scientist capabilities but is intended to focus on extension of the infrastructure.
Data Scientist (extending development)	Use the basic infrastructure, along with analytical and data subject matter expertise, to extend analytical capabilities and pull Data Enrichment into more automated capabilities; acting as a narrow extension of the developer – familiar with the Data Operations, Storage, and querying tools.
Data Scientist (power business analyst)	Provide analytical support to the Business Analyst / SME by supporting development of queries, indices, analytical applications and approaches, new visualizations for Business Intelligence and automation of data enrichment; acting as a more technical extension of the Business Analyst / SME.
Business Analyst / SME	Consume QA/QCed and enriched (metrics, enrichment, analytic outputs) data products to create additional enriched data products, reports, and/ or Business Intelligence dashboards (visualization/ presentation); will combine and select data products (pre-defined); may also create or flag simple metrics; not responsible for data quality or implementing analytical functions.
Data Engineer	Responsible for defining, building and managing the essential services which ingest, validate, remediate, transform and store physical data assets required for analytics or other data management functions.
Consumer	Consume / view Business Intelligence dashboards (visualization/ presentation), including reports and metrics, interact through simple select/filter and visualization tool.
External Data Stream	An external streaming data system creates a Data Ingestion action when a message is published to queue.
External Data Push	An external data resource that pushes data to the AF MAJCOM/ Functional Data Platform on a scheduled or ad hoc basis, creates a Data Ingestion action when data is pushed to the AF MAJCOM/Functional Data Platform.
Timed External Data Pull	A timed ingestion action from an external Data Resource.
Internal AF MAJCOM/Functional Data Platform Data Operations Timer	A timed Data Operation within the AF MAJCOM/Functional Data Platform.

ACTOR	DEFINITION
Internal AF MAJCOM/Functional Data Platform Data Operations Event Trigger	A Data Operation within The AF MAJCOM/Functional Data Platform that is triggered by an event, such as the creation of an OR record.
External System Call	An action triggered by a call from an external system to a AF MAJCOM/ Functional Data Platform API.

TABLE 7: AF MAJCOM/FUNCTIONAL DATA PLATFORM LOGICAL BUSINESS ARCHITECTURE DEFINED TERMS

ACTOR	DEFINITION
Risk Management	Ensuring that management of AF MAJCOM/Functional Data Platform development, deployment, and configuration activities are appropriate to the level of system and user risk.
Configuration Management	An activity that introduces a formal external review, control, and documentation process for changing AF MAJCOM/Functional Data Platform system components, configurations, or versions prior to making modifications.
Documentation	The formal documentation of a configuration, model, or component in a manner that is electronically searchable and retrievable by all AF MAJCOM/Functional Data Platform users.
Documentation and Review	The formal documentation of a configuration, model, or component in a manner that is electronically searchable and retrievable by all AF MAJCOM/Functional Data Platform users, after review and feedback by an external process.
Capability Group	A logical group of system components and functions that groups AF MAJCOM/Functional Data Platform Functional groups into a portfolio of services that have similar characteristics in terms of what they are used for and the skill sets recommended to maintain them.
Functional Group	A logical group of system components and functions that groups AF MAJCOM/Functional Data Platform functions into a discrete AF MAJCOM/Functional Data Platform system component that can be maintained as a specific module with enumerated functions and services.
Functions	A callable AF MAJCOM/Functional Data Platform service or operation that ingests, stores, retrieves, manipulates, or presents data. A Function is a subset of a Functional group.
Conceptual Architecture	The overall organization of Capability and Functional groups.
Logical Architecture	Describes the interface points between Functional groups.
AF MAJCOM/Functional Data Platform System Boundary	Describes the logical data and networking boundary between AF MAJCOM/Functional Data Platform and other systems in the hosting environment.
Use Case Implementation Template	A standard operating procedure that can be followed to implement one or more Air Force MAJCOM/Functional Data Platform Use Cases using modular, reusable Air Force MAJCOM/Functional Data Platform functions.
Design Pattern	A suggested implementation model for Data Storage or a Data Operation.

ACTOR	DEFINITION
Framework Code	This represents installation, modification, code base/library upgrades or patches, and base installation and environment configurations specific to one of the tools supporting each of the Functional groups that impact how it manages scaling, job management, or base functions and classes as a general environment to fulfill Systematic Standards.
Configurations	Represent specific jobs, workflows, configuration files, or scripts that utilize the Functional group tools' scaling and management procedures, base functions, and classes to provide a specific analytic, data operation, data product, or visualization to enact a Programmatic Standard.

9.3 Acronym Glossary

TABLE 8: KEY ACRONYMS

ACRONYM	DEFINITION
ACID	Automaticity, Consistency, Isolation, Durability
ADS	Authoritative Data Sources
AF	Air Force
AF MAJCOM	Air Force Major Commands
AMI	Ambient Intelligence
API	Application Programming Interface
BI	Business Intelligence
CCB	Change Control Board
CDO	Air Force Chief Data Officer
CI	Continuous Integration
CM	Configuration Management
COTS	Commercial Off the Shelf
D&R	Documentation and Review

ACRONYM	DEFINITION
DevOps	Development and Operations
DLMS	Defense Logistics Management Standards
DO	Documentation Only
DOF	Data Operations Framework
DPCS	Data Product Consumer Services
DPFS	Data Platform Foundation Services
DSL	Domain Specific Language
EDAS	Enterprise Data and Analytics Services
EEIM	Enterprise Energy Information Manage
EIM	Enterprise Information Model
EMS	Enterprise Metadata Services
ERP	Enterprise Resource Planning
ETL	Extract, Transform, and Load
FIAR	Financial Improvement and Audit Remediation
FISMA	Federal Information Security Management Act
FOSS	Free and Open-Source Software
GUIDS	Global Unique Identifiers
IA	Information Asset – a contextual data product
ICD	Interface Control Document
IDAM	Identity and Access Management
JDBC	Java Database Connectivity

ACRONYM	DEFINITION
LBR	Logical Business Rules
MDM	Master Data Management
ODBC	Open Database Connectivity
PBR	Physical Business Rules
QA	Quality Assurance
QC	Quality Control
RBAC	Role Based Access Control
RBCM	Risk Based Configuration Management
RCA	Root Cause Analysis
RDF	Resource Description Framework
REST	Representational state transfer
RMF	Risk Management Framework
RPIM	Real Property Information Model
RUM	Role / User Management
SAF/CO	Air Force Chief Data Office
SFIS	Standard Financial Information Structure
SLA	Service Level Agreement
SME	Subject Matter Expert
SMMS	Strength Maintenance Management System
SOA	Service-Oriented Architecture
SQL	Structured Query Language

ACRONYM	DEFINITION
SSL	Secure Sockets Layer
SVAULT	Secure, Visible, Accessible, Understood, Linked and Trusted
TLS	Transport Layer Security
UI	User Interface
URI	Uniform Resource Identifier
VAS	Value-Added Services
VM	Virtual Machine
WADL	Web Application Description Language

9.4 Interoperability Key Guidelines

The key to enabling the data-driven organization is the ability for systems to communicate and share data. By following the guidelines set forth in this Reference Architecture platform data interoperability can be achieved. Table 9 provides a summary of key concepts with which platform owners must comply to provide platform and data interoperability:

TABLE 9: PLATFORM AND DATA INTEROPERABILITY CONCEPTS

PLATFORM INTEROPERABILITY
• Driven by the exposure of services – Web API or Web Services • All functions, microservices and value-added services exposed through Web API or Web Service • Document and publish web services to the Enterprise Information Model • Web services parameterized and published to the services registry using a Standardized Interface Control Document (a reference example of which is in Section 7.1) for describing methods and parameters.
DATA INTEROPERABILITY

• Employ a loosely coupled architecture that separates compute, storage and Enterprise Information Model layers
• Use open file formats vs. proprietary formats
  ○ Self-describing metadata infused formats such as CSV, XML, etc.
  ○ Big Data Ecosystem formats such as Parquet, Apache ORC, Apache AVRO
  ○ Flat, fixed-width, variable-width are admissible if supplemented with schema metadata
• Employ a service referenceable object store that can be consumed by Web, API or Web Service
• Network resolvable
• In addition to service referenceable object store, data can be consumed through traditional web service Pub/Sub topic or message queue

U.S. AIR FORCE