I n t e g r i t y - S e r v i c e - E x c e l l e n c e

Air Force Data Reference Architecture and Platform

Ms. Jackie Murray
11 Oct 2018

U.S. AIR FORCE

AF Data Challenges

- **Large number of legacy systems with numerous point-to-point interfaces that are expensive to implement and maintain**

- **Antiquated, brittle architectures inhibit legacy system enhancements required to make data accessible**

- **Many legacy systems employ "data jails" that prevent timely data access that supports the analytics needs of Functional communities**

- **The time is now to employ new, commercially available data services that will change the landscape of how data enables mission effectiveness**
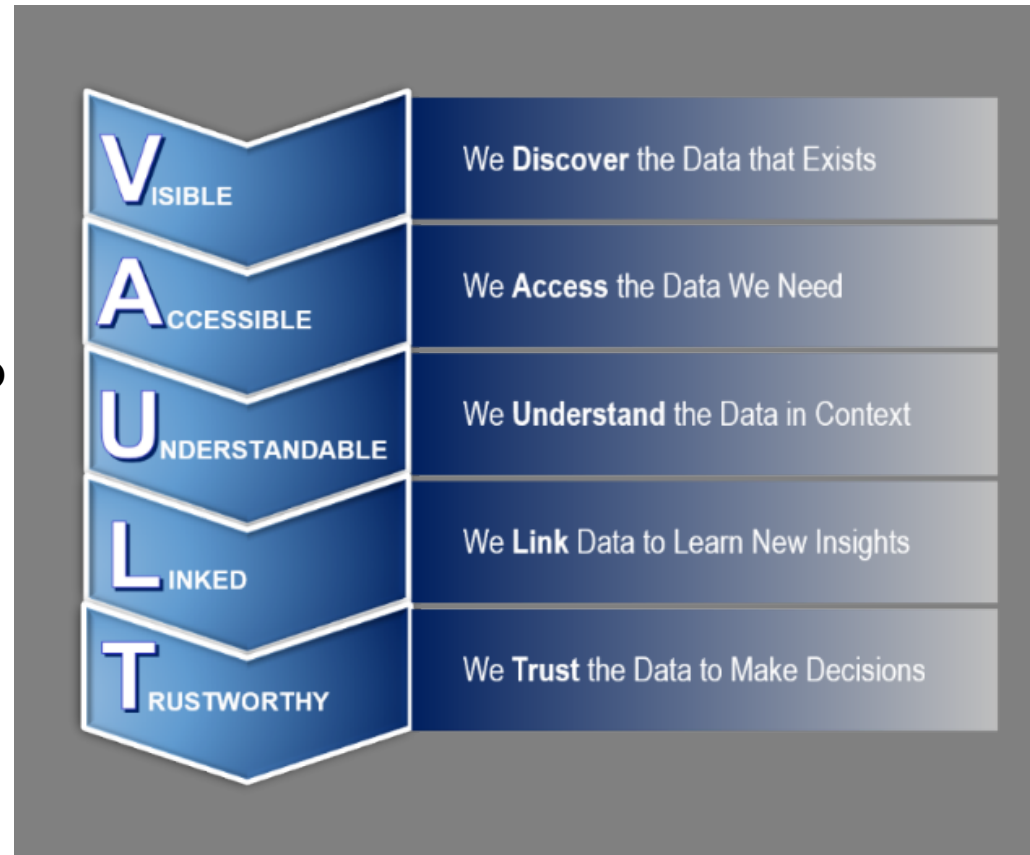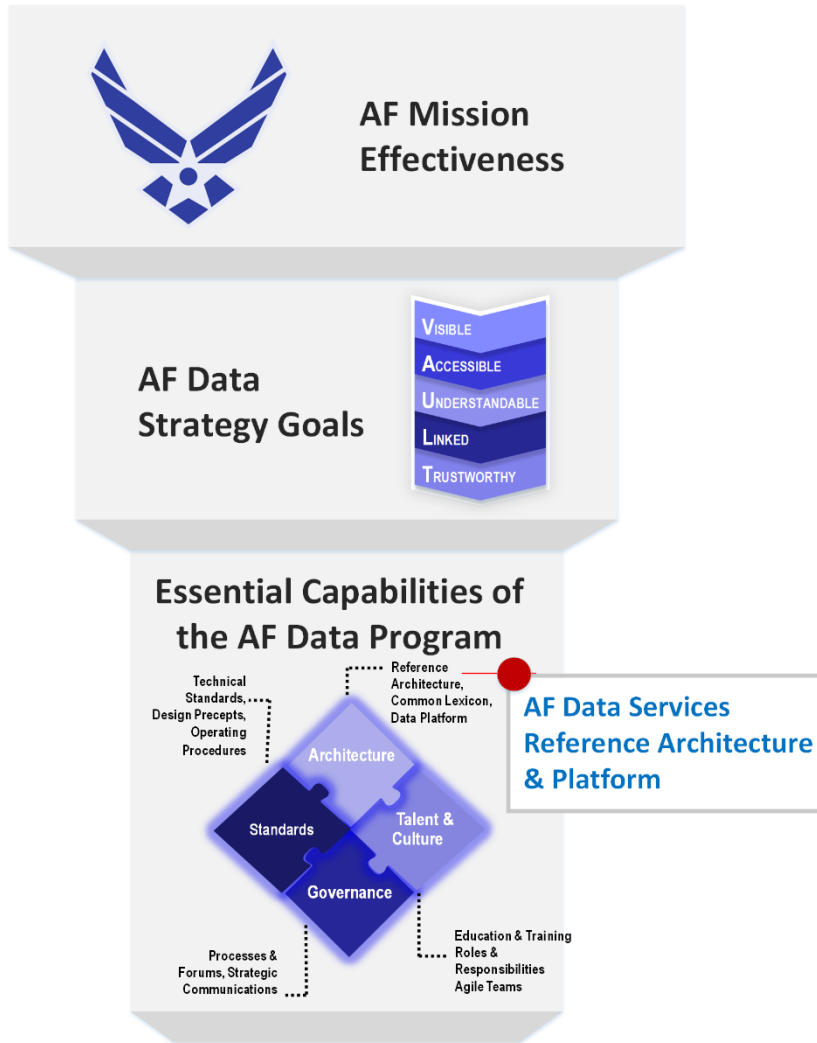
U.S. AIR FORCE

- **Ensure the Air Force can leverage data as a strategic asset in every facet of military operations and support functions**

- **Enable and empower Airmen to use data as a critical strategic asset when and where it is needed to accelerate informed decisions at speed and scale**

- **Transform the Air Force to a data-driven organization**

VISIBLE	We **Discover** the Data that Exists
ACCESSIBLE	We **Access** the Data We Need
UNDERSTANDABLE	We **Understand** the Data in Context
LINKED	We **Link** Data to Learn New Insights
TRUSTWORTHY	We **Trust** the Data to Make Decisions

U.S. AIR FORCE

AF Mission Effectiveness

AF Data Strategy Goals

Visible
Accessible
Understandable
Linked
Trustworthy

Essential Capabilities of the AF Data Program

Technical Standards, Design Precepts, Operating Procedures

Reference Architecture, Common Lexicon, Data Platform

AF Data Services Reference Architecture & Platform

Architecture

Standards

Talent & Culture

Governance

Processes & Forums, Strategic Communications

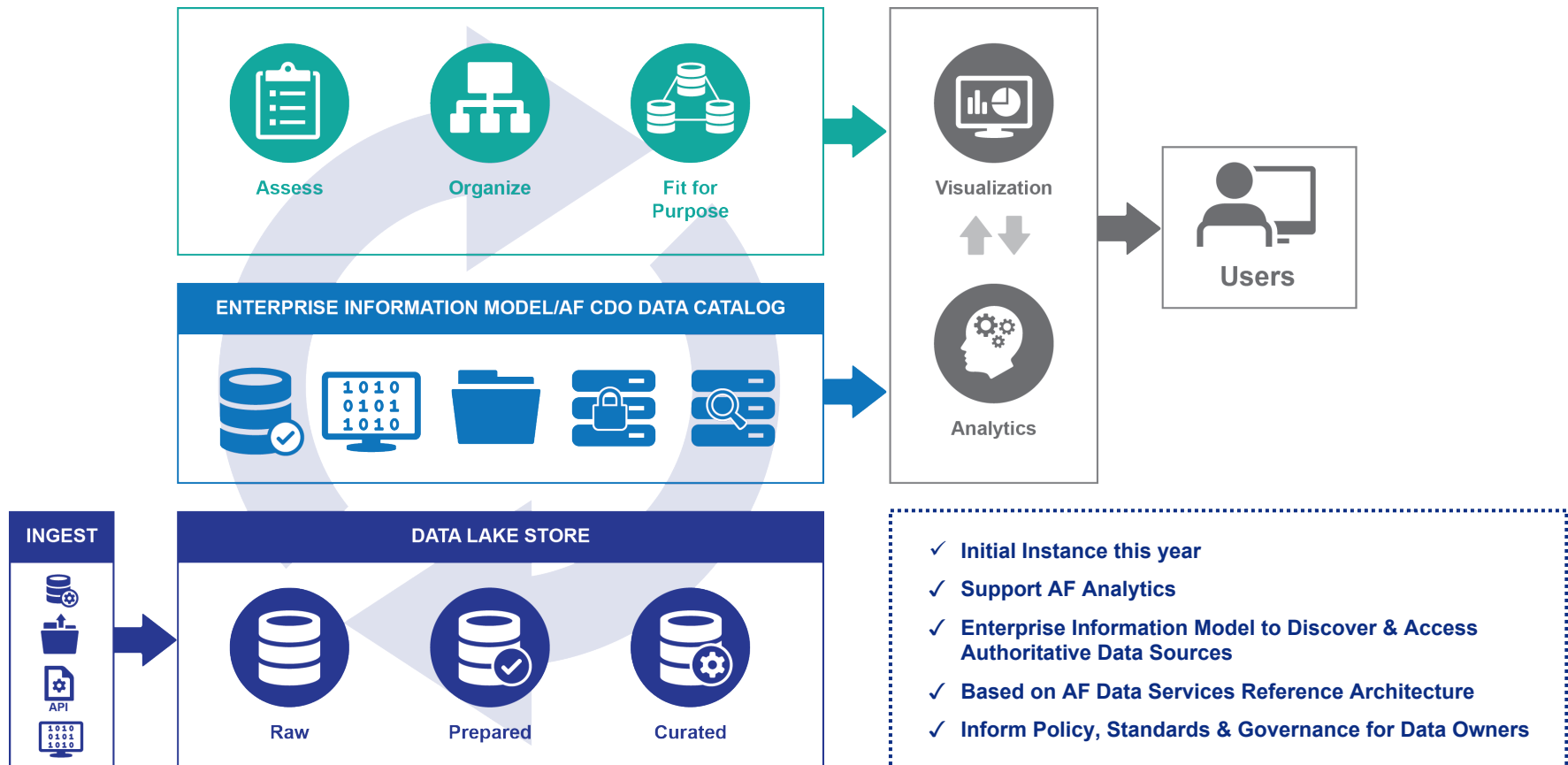Education & Training Roles & Responsibilities Agile Teams

Architecture & CDO Platform Key Tenants

- **Agile Approach**

- **Solution Priorities**
  - Cloud-Native Services First
  - Open Source Software Second
  - COTS Products Third -- Minimize Licensed Software
  - No Proprietary Solutions

- **Vendor Agnostic –** Must not be reliant on one company

- **Product Agnostic –** Must allow for access to a variety of tools best suited to meet individual mission demands

- **Architect for easy adoption of new tools and services while minimizing impact to the underlying data**

- **Utilize with on-prem or off-prem services**

- **Decoupling of compute and storage**

- **Minimizes RMF impact**

Delivering Data Through the AF Data Platform

U.S. AIR FORCE

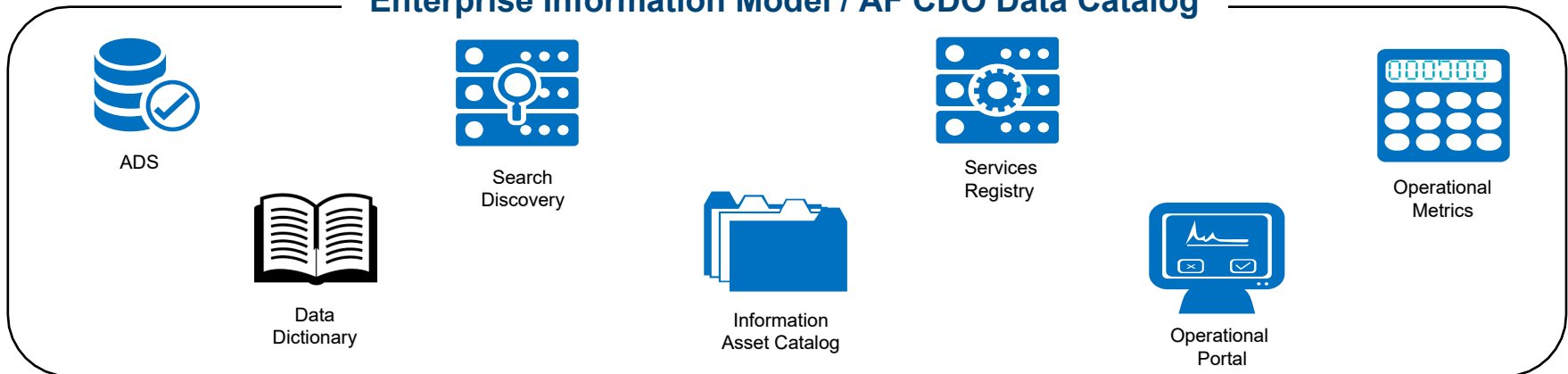**Assess** · **Organize** · **Fit for Purpose**

Visualization

Users

ENTERPRISE INFORMATION MODEL/AF CDO DATA CATALOG

Analytics

INGEST

API

DATA LAKE STORE

Raw · Prepared · Curated

- ✓ **Initial Instance this year**
- ✓ **Support AF Analytics**
- ✓ **Enterprise Information Model to Discover & Access Authoritative Data Sources**
- ✓ **Based on AF Data Services Reference Architecture**
- ✓ **Inform Policy, Standards & Governance for Data Owners**

Policy & Standards
Data quality, Data tagging, Metadata management, Data loss prevention, Ingest methods, ADS registration, Information asset definition, Data dictionary management, Change data capture
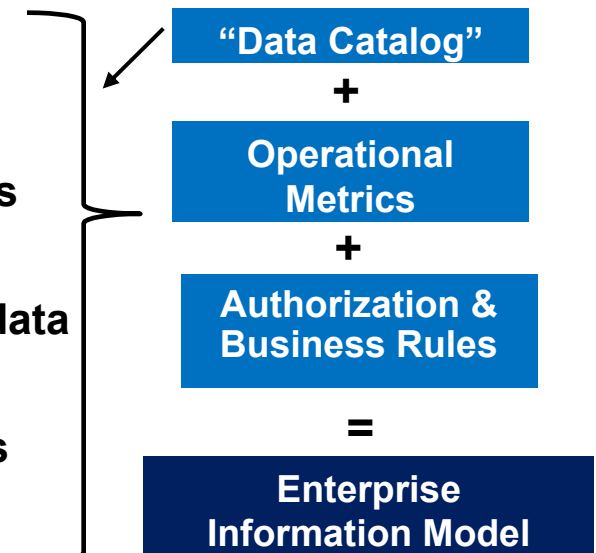
U.S. AIR FORCE

Enterprise Information Model / AF CDO Data Catalog

- ADS
- Search Discovery
- Services Registry
- Operational Metrics
- Data Dictionary
- Information Asset Catalog
- Operational Portal

- *Data Dictionary* – **Data element level glossary (rationalization of Enterprise vocabulary)**

- *Information Asset Catalog* – **Aggregated data elements designated as AF data assets**

- *ADS Registry* – **Listing of originating sources for the data elements and information assets**

- *Services Registry* – **Description of Enterprise Services utilizing cataloged information assets**

"Data Catalog"
+
Operational Metrics
+
Authorization & Business Rules
=
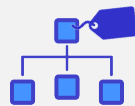**Enterprise Information Model**

U.S. AIR FORCE

1 — INGEST

a) Data is inducted into the lake and ingest pipelines are created using Apache Kylo and Apache NiFi
b) Data is profiled, indexed and loaded into Apache Hive table
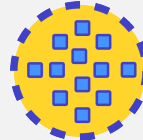c) Data is added to the Catalog

2 — CURATION

a) Data cleansing and validation policies are set – an optimized target format is selected
b) Create additional business metadata and tags for searching
c) A feed schedule is finally established to check for new files
d) Profile details

3 — WRANGLING

a) Using R in a Zeppelin notebook key columns are imported
b) A Spark data frame is created
c) Libraries are imported, column names and data types are properly assigned
d) Additional wrangling removes invalid sequences and manipulates the data into a tidy format

4 — ANALYTICS

a) The R script continues by calculating the mean and standard deviation of each NIIN by FY QTR
b) A final table of statistics is aggregated, collected, and verified in preparation for plotting

5 — VISUALIZATION

a) Additional wrangling shapes the stats table into proper format for plotting
b) Coding calls libraries to display high end publishable graphics from the statistical analysis

QUESTIONS