



VAULT Platform Presentation

DAFITC 2023

Clare Bolin, Mark Bryant, Christopher Gillie

SAF/CND, Air Force Chief Data and Artificial Intelligence Office

0900-0950, Montgomery Room 7



Disclaimer



The discussion of non-federal entities, methods, products, or services does not imply any endorsement by the United States Government, the Department of Defense, or the Department of the Air Force.



VAULT Data Platform



- Agenda
 - Foundation & Overview (Chris)
 - Tenant Space Offerings (Clare)
 - BYOT Overview & Process (Clare)
 - Data Acquisition (Clare)
 - Costs (Clare)
 - Demo Use Case (Mark)
 - Summary (Chris)
 - How To Access VAULT (Chris)



VAULT Data Platform (VDP)

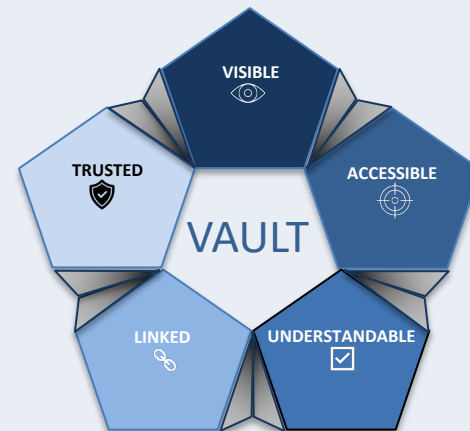
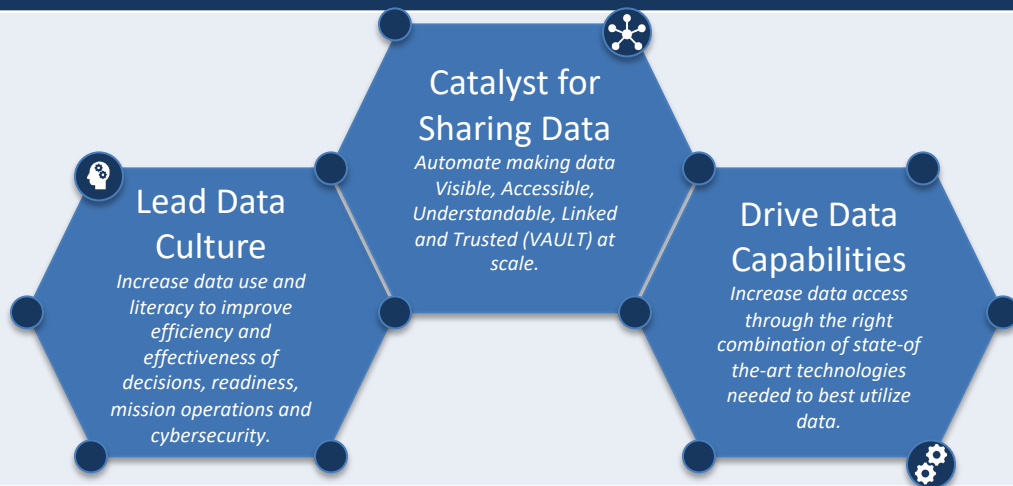


Vision & Mission

VDP Vision – Provide users the platform to work on current-state and emerging technology tools with real data in a safe and secure environment – matching need with innovation.

VDP Mission – Empower and posture the Department of the Air Force to harness data for competitive military advantage.

Strategic Goals



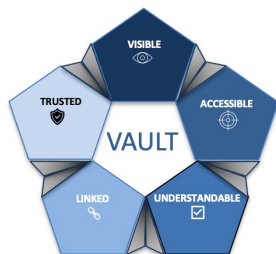


VDP Impact – Customer Focus

Enhance Data Sharing: Support an evolving DAF Data Ecosystem. Sharing data between the Big 6 platforms thru automations. Support advancement toward a virtualized data fabric while supporting current day data sharing needs thru innovative technologies.

Design for Evolution, Extensibility & Scalability: Drive DevSecOps methodologies and P1 alignment (BigBang and deployment methodologies). CI/CD build/deployment pipelines, GitOps for applications, and focus on Kubernetes. Deliver tools to the platform but without enabling vendor lock within the platform.

Data as a Strategic Asset: Ensure data is visible and accessible thru automated data pipelines to acquire, scan, enhance, promote & egress data. Provision data to users within their tenant and the platform.



Defense
Department
Data Strategy

CDO Data
Service
Reference
Architecture

Air Force Data
Management
Policy

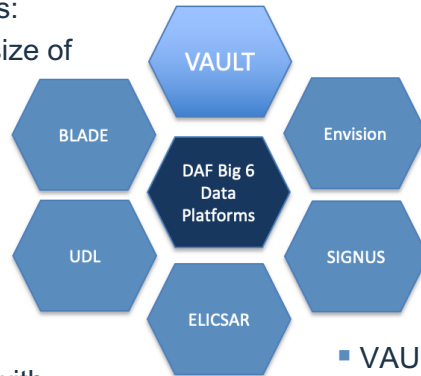
VAULT foundation is built upon 3 core policies/directives.



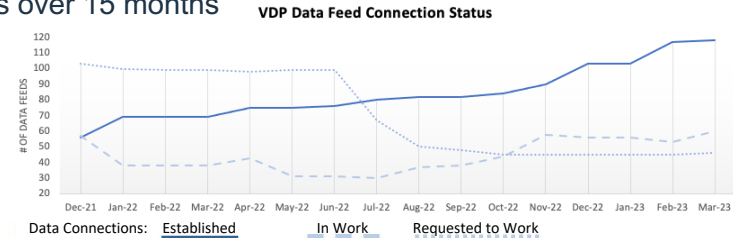
Vault Data Platform – Data & the Big 6



- VAULT Data Platform is one of the **Department of the Air Force Big 6 Data Platforms**
- VAULT is a constantly expanding ecosystem and a key pillar to the DAF Data Fabric
- VAULT Supports Data Connections to many different platforms
- Data can be shared via several methods:
 - Each connection type varies in type, size of data, level of effort and speed
 - DISA SFG
 - SFTP
 - S3 to S3
 - AWS Snowball
 - API <In Work>
 - End game is **data federation**
- VAULT has an established connection with Envision
- VAULT is working a BLADE connection



- VAULT has more than **doubled** the amount of available Data Feeds over 15 months



VAULT Data Platform Data Statistics

- 120 Total Data Feeds
 - 59 Actively in Progress
 - 46 Requested to Work
 - 86 Separate Data Sources
 - ~473 Million Files / 1,570 TB of Data within VAULT
-
- Legend: ■ Fully Automated ■ Manual
- VAULT is automating the data sharing capability with Envision – allowing the user to initiate data sharing between the platforms
 - VAULT is expanding to support API connections – enabling the user to request more connections in a self service manner

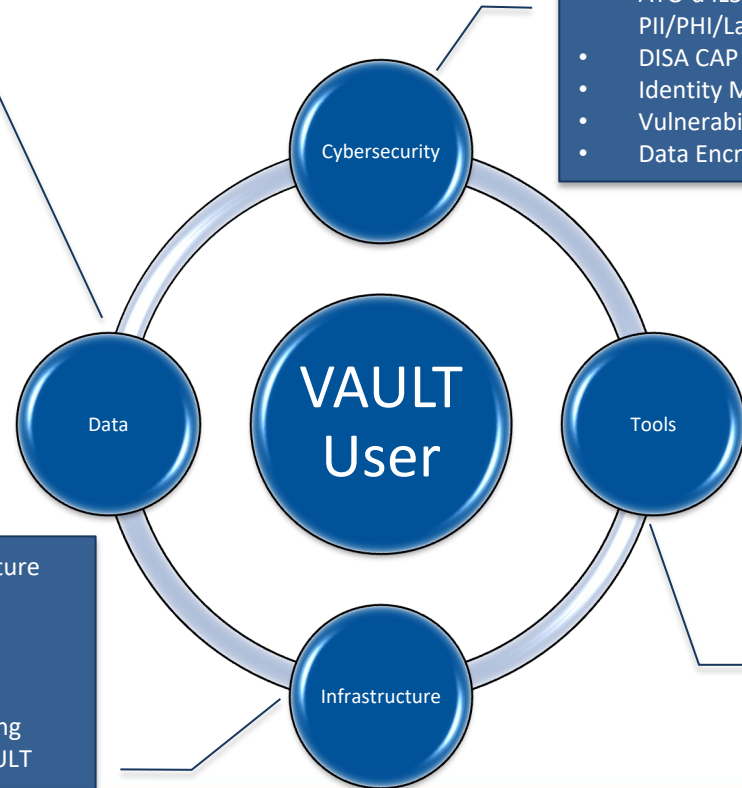


VDP Impact – Core Capabilities



- VAULT enables:
 - Data Access and Visibility thru implementation of Data Pipelines
 - Data Access thru automated Data Acquisition
 - Data Curation, Analysis, Discovery and Wrangling thru provisioning of the Enterprise and BYOT Tool Suites
- VAULT implements Data Access Controls to properly manage access to Limited Visibility/Limited Distribution Data thru RBAC
 - ABAC Access Control Implementation <in work>

- Cloud based Scalable Architecture
- Automated CI/CD deployment pipelines
- Platform 1 Alignment and Kubernetes Deployments
- Tenant Space Offerings enabling private workspaces within VAULT



- ATO'd IL5 and IL6 boundary with PII/PHI/Law Enforcement Overlay
- DISA CAP Connection
- Identity Management
- Vulnerability Scanning
- Data Encryption

- VAULT enables microservices to support plug and play open source and COTS tools – preventing vendor lock on the program
- VAULT provides three major areas of tools:
 - Infrastructure
 - Enterprise
 - Bring Your Own Tool
- Infrastructure tools support the key infrastructure capabilities (I.e. Keycloak)
- Enterprise tools are provided to the platform users as an enterprise service
- Bring Your Own Tool (BYOT) enables users to bring a tool of their choice to the platform in their tenant space.



VAULT Data Platform Evolution

- NIPR VAULT Contract Award – July 2018
- IL4 ATO

2018

- IL6 ATO
- NIPR – 25 Tenants / 300+ Accounts
- SIPR – 3 Tenants / 200+ Accounts

2020

- NIPR – 100 Tenants / 13000+ Accounts
- SIPR – 13 Tenants / 900+ Accounts

2022

- SIPR VAULT Contract Award
- IL5 ATO
- NIPR – 10 Tenants / 200 Accounts

2019

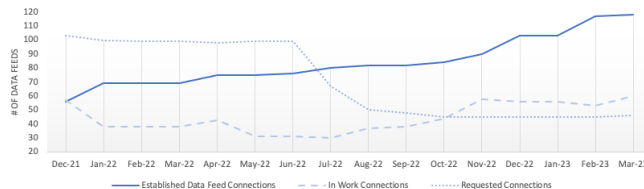
- NIPR – 50 Tenants / 6000+ Accounts
- SIPR – 13 Tenants / 600 Accounts

2021

- NIPR – 111 Tenants / 15000+ Accounts
- SIPR – 13 Tenants / 1000+ Accounts

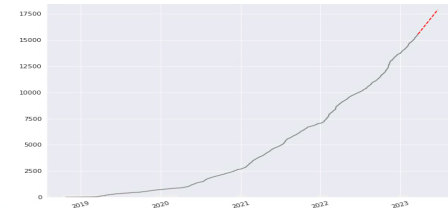
2023

VDP Data Feed Connection Status



120 Data Feeds
59 In Process
46 In Backlog

Vault Users





VAULT Data Platform - Metadata



VAULT Minimum Metadata – Captured in Data Pipelines
100% Aligned to Department of Defense Metadata Guidance Memo (March 2023)

Data Acquisition Pipeline

Data Promotion Pipeline

Min³ Metadata

- Data Sponsor Contact & Org Info
- Data Source Steward Contact & Org Info*
- VAULT Tenant Name
- Enterprise or Tenant Specific Data
- Justification for non-Enterprise Data
- Strategic/Operational Impact of Data
- Limited Distribution Requirements*
- Limited Distribution Justification
- Data Location
- Data Source Description / Information*
- Data Set Description / Information*
- Security Classification*
- System Information
- DITPR ID

Min2 Metadata

- Origin Info (Chief, PM and Org Info)*
- Data Set Steward Contact & Org Info*
- Incident Response Contact & Org Info
- Disclosure & Releasability*
- Data Source Information*
- Data Set Information*
- Data Set Program Info
- Frequency
- Bureau Info

Min Metadata

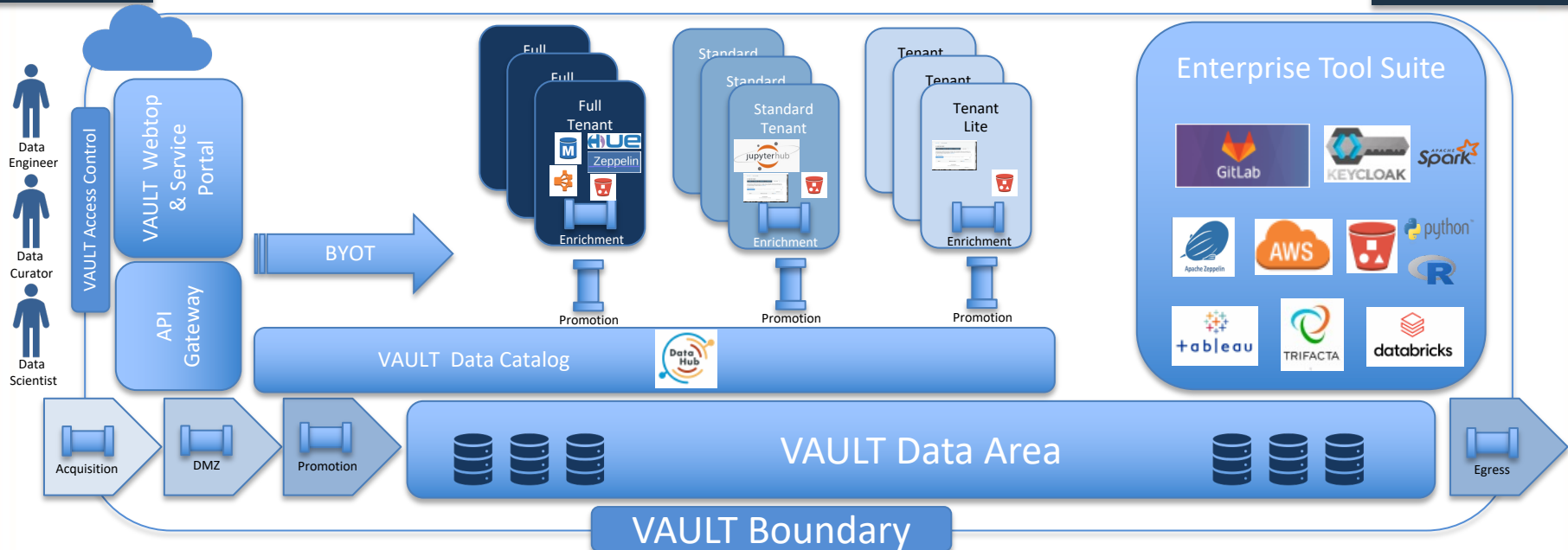
- Data Dictionary Information
- Authorization Reference / Legal Authority*
- Office of Record / Custodian*
- Date/Time Cataloged*
- Update Time
- Lifecycle Management Info
- Frequency
- Refactoring Rules
- Tags
- Data Schema/Type/Format*
- Unique ID*

Metadata Management is the foundational to a data-centric, zero trust environment.
Empowering data visibility, data accessibility and data linking.

* Indicates alignment to DoD Metadata Guidance Memo from March 2023



VAULT Data Platform



The VAULT Data Platform provides cyber-secure, cloud-based tools to connect, find, share, and learn from AF data – Enables Executive Order on Maintaining American Leadership in Artificial Intelligence



VAULT Data Platform – Tenant Spaces



- A tenant space is a private workspace for storing, building, and maintaining organization-specific data and workflows
- As the VAULT Data Platform continues to grow and expand number of tenants and BYOT, a charge back model is needed to offset SAF/CND's AWS cloud and compute costs
- VAULT Data Platform offers the following options when it comes to a tenant space:
 - Full Tenant
 - Standard Tenant (coming soon)
 - Tenant Lite
- Each tenant option supports different use cases and comes with a different cost model



"Full" Tenant

- Hue and Zeppelin enable data analytics, experimentation, and cleansing
- Elastic MapReduce (EMR) and Relational Database Service (RDS) are used by Hue and Zeppelin for processing
- Databricks
- Typical compute stack in NIPR and SIPR consists of:
 - Two (2) r5a.2xlarge EC2 servers (8 vCPUs and 64GB RAM per server)
 - One (1) db.t2.xlarge RDS server (4 vCPUs and 16 GB RAM)
- S3 buckets for data storage



"Standard" Tenant

- **Coming Soon!**
- Jupyterhub
- S3 console access still allows tenants users to upload and download data
- S3 buckets for data storage



Tenant "Lite"

- S3 console access still allows tenants users to upload and download data
- S3 buckets for data storage



Vault Data Platform Impact - BYOT



VDP offers a unique and highly requested Bring Your Own Tool (BYOT) process to enable tenants to bring custom applications to their tenant spaces for utilization in data manipulation and visualization.

- VDP has over 20 unique BYOT applications currently deployed
- VDP currently has a backlog of ~20 unique BYOT applications requesting deployment thru various phases of the BYOT process

Sample BYOTs on VAULT

DAF Data Fabric	VESPR	Swimlane
ITIPS	Immuta	Galvanize
Appian	Dataiku	Enveil
COGYNT	Plotly Enterprise	DAFBOP / PERforM
Rhombus Guardian: SLRD & Force Structure	A1 myData – Erwin & Informatica Tools (PowerCenter, Data Quality & Address Doctor, Axon, EDC, MDM)	

Mission Critical BYOT Examples:

- A1 myData – directly supporting the transition of A1 TMDE to the VDP
- SAF/CDM – mission impacts supporting insider threat initiatives
- SAF/MG – supporting direct provisioning of the PERforM capabilities
- DAF Data Fabric – deployment of the DAF Data Fabric to enable virtualization and cataloging across the DAF
- ITIPS – enabling ITIPS access thru VAULT

EN|VEIL
ENCRYPTED VEIL

COGILITY



data
iku

Informatica

erwin
by Quest



plotly

VESPR

appian

IMMUTA

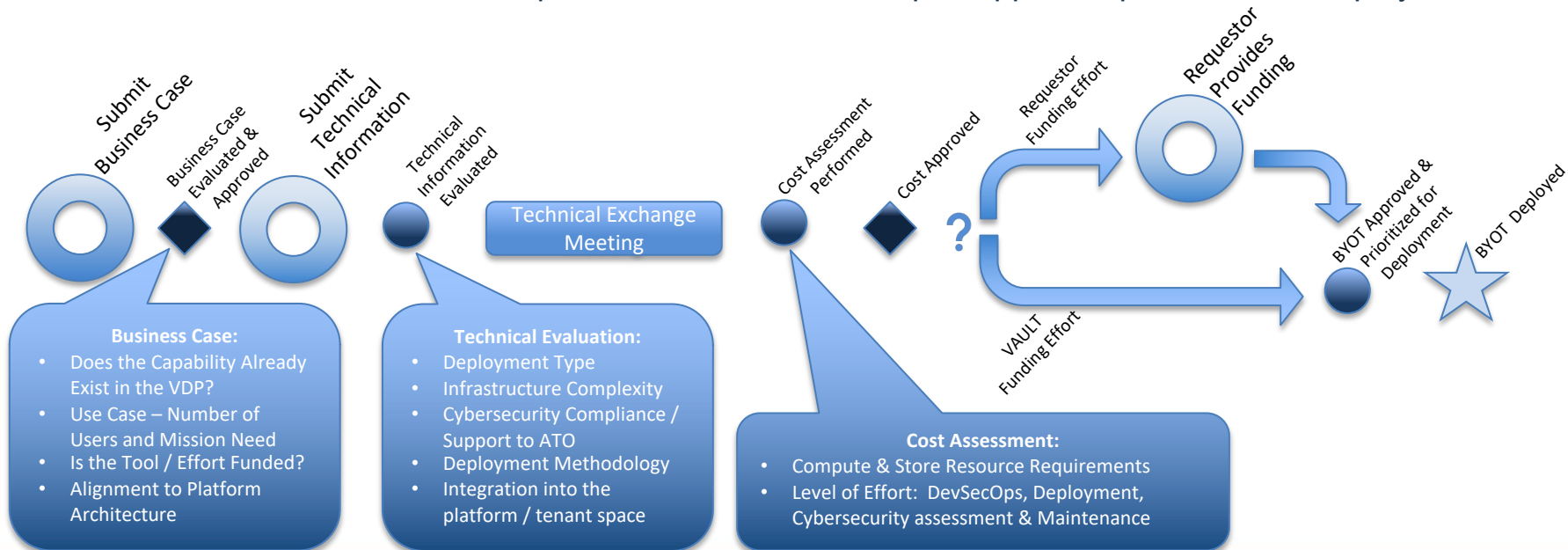




Vault Data Platform - BYOT



- VAULT BYOT process is designed to capture and evaluate the business need and technical requirements – enabling a thorough understanding and plan to deployment
 - BYOTs are routed thru the process, evaluated and, upon approval, prioritized for deployment

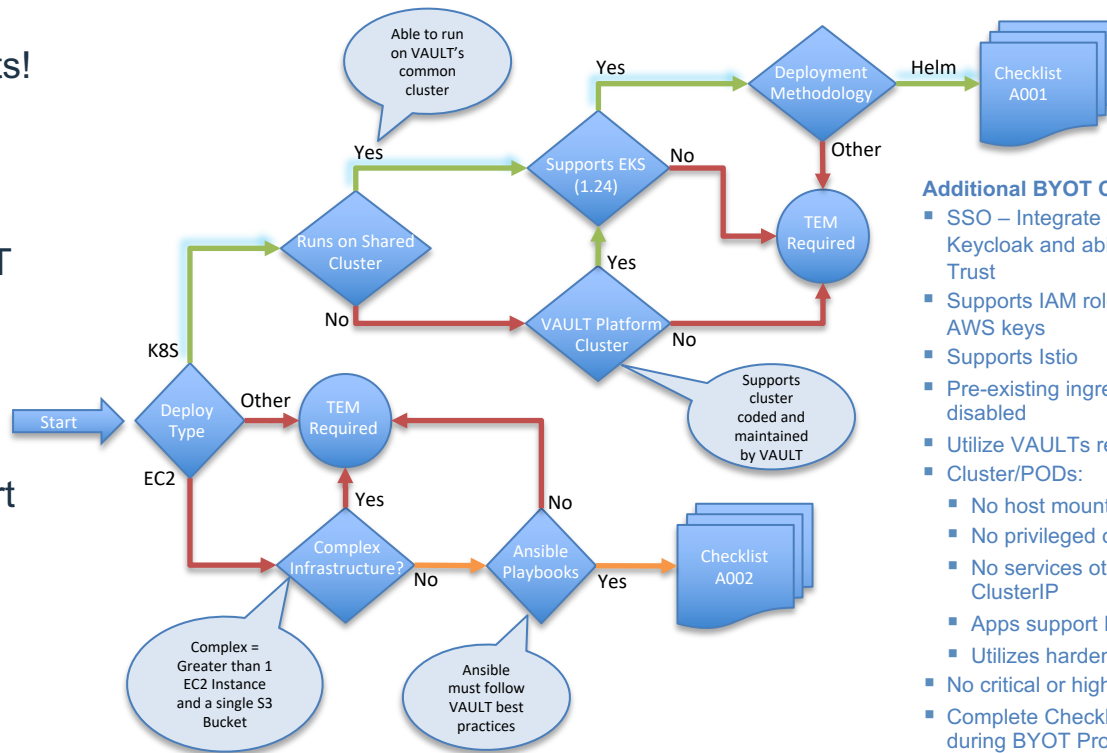




VAULT Data Platform - BYOT



- BYOT Fast Lane Deployments!
 - Optimized Deployment in alignment with VAULT Platform
- BYOT Deployments to VAULT can vary greatly based upon technical complexity.
 - BYOTs can take the fast track or may require additional technical support to deploy



Additional BYOT Considerations:

- SSO – Integrate with VAULT's Keycloak and ability to pass CA Trust
- Supports IAM roles vs requiring AWS keys
- Supports Istio
- Pre-existing ingress can be disabled
- Utilize VAULT's registry
- Cluster/PODs:
 - No host mounts
 - No privileged containers
 - No services other than ClusterIP
 - Apps support Istio injection
 - Utilizes hardened containers
- No critical or high vulnerabilities
- Complete Checklist Provided during BYOT Process



VAULT Data Platform – Data Acquisition (1 of 2)



- Data can be acquired and brought into the VAULT Data Platform via several methods
 - Varies in terms of connection, speed and size of data to be transferred, level of effort required on data provider and the VAULT team, and the speed of delivering/receiving data

Method	Description	Supported File Size	Level of Effort	Typical Turnaround Time*
Direct upload via Apache Hue, S3 Console, Databricks	Users can directly upload data via the Hue application	Up to 500MB	Low (requires user to have a VAULT account)	Data is available once upload is complete
Secure File Gateway (SFG)	DISA Secure File Gateway (SFG) ensures secure data transfer between DoD system and those interfacing with the DoD system on the NIPR SIPR networks. Relays and folders are created for each customer. Data provides are granted access and can drop files to the relay into the Vault Platform. DISA Secure File Gateway (SFG) ensures secure data transfer between DoD system and those interfacing with the DoD system on the NIPR SIPR networks. Relays and folders are created for each customer. Data provides are granted access and can drop files to the relay into the Vault Platform	<= 100GB	Medium (data provider requires DISA account be created + obtaining access to DISA's SFG relay)	2-4 weeks (dependent on DISA)
Secure File Transfer Protocol (SFTP)	Secure File Transfer Protocol (SFTP) is a network protocol for accessing, transferring, and managing files on remote systems. SFTP allows securely transfer of data files.	<= 100GB	Medium (Requires creation of RSA Public / Private Keys and possible Firewall changes)	5+ weeks (VAULT team creates public keys and, update Firewall rules, establish connection and test the SFTP connection)



VAULT Data Platform – Data Acquisition (2 of 2)



- Data can be acquired and brought into the VAULT Data Platform via several methods
 - Varies in terms of connection, speed and size of data to be transferred, level of effort required on data provider and the VAULT team, and the speed of delivering/receiving data

Method	Description	Supported File Size	Level of Effort	Typical Turnaround Time*
AWS Cloud S3-to-S3	Amazon Simple Storage Service (Amazon S3) is an object storage solution that provides scalability, data availability, security, and performance. Data provider can upload their data to specific S3 buckets and make them available for the Vault Platform to pick up. This storage space supports all data formats.	<5TB	Medium (Requires access to be granted to the specific bucket for both Data provider, and Vault Platform.)	<Up to 2 weeks (VAULT team creates required bucket, grants appropriate permissions and makes bucket available)
AWS Snowball	Amazon Web Services will ship a Snowball to the provided location. You will be required to load your data to the Snowball and mail it to the location specified in the instructions provided with the Snowball	Up to 80 TB per snowball	High (Requires approval by data provider and data provider.)	5+ weeks (Order of snowball by Vault takes about 2-3 weeks, Loading of data, shipping back to AWS, and uploading to Vault takes 1.5 – 2 weeks. Approval by provider hosting provider can vary in duration.)
Removeable Hard Drive	Data downloaded to an external hard drive, then handed over to proper personnel.	<= 500 GB	Medium (Requires getting an approved hard drive, having someone with USB exception on data provider side.)	4+ weeks (Acquire approved hard drive, downloading of data, doing the hand-off, and uploading of data.)
Application Programming Interface (API)	Application Programming Interface (API) to federate Metadata and data between system (Under development)			



VAULT Data Platform – Licensing



- AWS GovCloud compute / store
 - For a full-service tenant on NIPR, this cost is \$2,400/month (\$28.8K/FY)
 - For a tenant-lite configuration on NIPR, this cost is \$500/mo (\$6K/FY)
 - SIPR Compute/Store costs are currently the same as NIPR
- MySQL RDS Instance
 - A single MySQL RDS Instance using a db.t2.xlarge server (default for a tenant) is \$482/month.
- Databricks
 - Utilization is a *shared* resource across both NIPR and SIPR
 - Cost is \$1.00/unit for utilization < 100,000 DBUs
 - Cost is \$0.69/unit for utilization > 100,000 DBUs
 - Utilization of Databricks will increase monthly compute / store costs
- Tableau
 - Tenants are required to purchase their own license utilizing the Blanket Purchase Agreement (BPA) already in place
 - Requirement is for license end dates to be aligned to end of FY (Sept) or end of CY (Dec)
 - One license is required for every license of Viewer, Explorer, and Creator
 - Data Management add-on license required for each license-type
 - ~\$133/month for Tableau Creator, Explorer, Viewer & Data Manager license



VAULT Use Case Demo – Pilot and Aircraft Readiness



VAULT Webtop



Welcome to VAULT

[Logout](#)

15:54:47

HelpDesk ▾



Help Desk & Requests

VAULT-DEFAULT ▾



VAULT Default Hue



VAULT Default
Zeppelin

Other Resources ▾



Plotly Dash Enterprise



Data Lake



Axon



Plotly Dash



AD Lookups



Databricks



Shiny Server



VAULT Help Desk



Welcome to VAULT

 Logout

15:56:23

HelpDesk ▾



Help Desk & Requests



VAULT Help Desk: Tenant, Data, Apps



GET STARTED WITH VAULT

ALL REQUEST ITEMS

Tenant Management Requests

Applications

PII Access

Bring your own Tool

Tenant Spaces

Data Requests

Quick Links

DAF Data Fabric

Vault Community Forum

Information Hub

Submit Help Desk Ticket

Browse All Request Items

How can we help?

SECRETARY OF THE AIR FORCE
CHIEF DATA OFFICE

QUICK LINKS

Vault Community Forum
Community Forum

Webtop
VAULT Webtop

DAF Data Fabric
External Site

Information Hub
HelpDesk Articles

Submit Help Desk Ticket
Submit a trouble ticket for issues in VAULT

MY ACTIVE ITEMS

0 Tasks

RECOMMENDED FOR YOU

Request Access to App

POPULAR TOPICS

MY REQUESTS

View Open requests Search open requests

Request	State	Updated
Databricks REQ0027918 SERIDEH MAHARATI	Open	12d ago
PII Certification Update REQ0024108 SERIDEH MAHARATI	Open	5mo ago
Databricks REQ0024039 SERIDEH MAHARATI	Open	5mo ago



VAULT Help Desk: Knowledge Base

[MY TASKS](#)[MY REQUESTS](#)[SM](#)[TOURS](#)[GET STARTED WITH VAULT](#)[ALL REQUEST ITEMS](#)[Home](#) > [Search](#)

Sources

[All](#)[Questions and Answers](#)[Knowledge Bases](#)[Catalogs](#)

FILTERS

Knowledge Bases

[Clear All](#)

KNOWLEDGE BASES RESULTS FOR "TICKET"

Create an Incident/Request

a Help Desk & Request **ticket** Log into VAULT at <https://afdatalab.af.mil> (recommend utilizing Chrome and then select 'Submit Help Desk **Ticket**' OR Select the 'Submit Help Desk **Ticket**' tile from the QUIC

Article: KB0010014 · Published: 9mo ago

Steps to Access Data in the VAULT

it is refreshed. Submit a trouble **ticket** with the name of the dataset and how often it should refreshed including to see if the view/report is located there. If in Envision, submit a trouble **ticket** wi

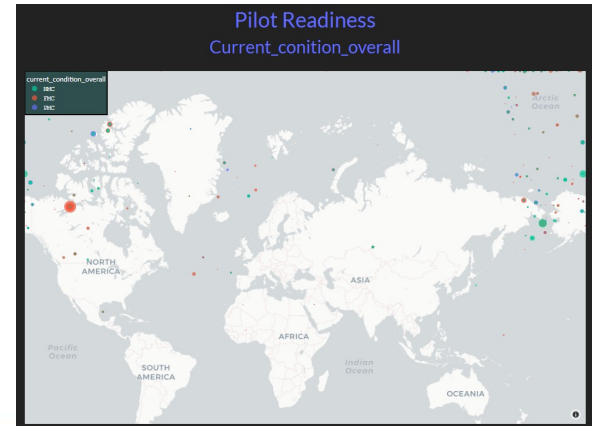
Article: KB0010102 · Published: about a year ago



Sample Use Case



- Goals:
 - Determine the current readiness levels of Pilots and Aircraft across the department
 - Use a classic ML model to predict readiness levels
- Data Sources:
 - G081: Provides unclassified maintenance data for Mobility airlift aircraft
 - ARMS: Provides aircrew training records used to determine readiness
- NOTE: Synthetic data for demonstration purposes





Sample Tooling Pipelines





Sample Tooling Pipeline



What is the Delta Lake?

- An open-source data management layer which provides reliability and performance to our data lakehouse

Why Delta?

- ACID transactions and logs
- Metadata handling
- Time travel – rollbacks and audit trails
- Schema enforcement
- Supports both structured and unstructured data



leau

+ a b l e a u

tabricks

plotly | Dash

data
iku

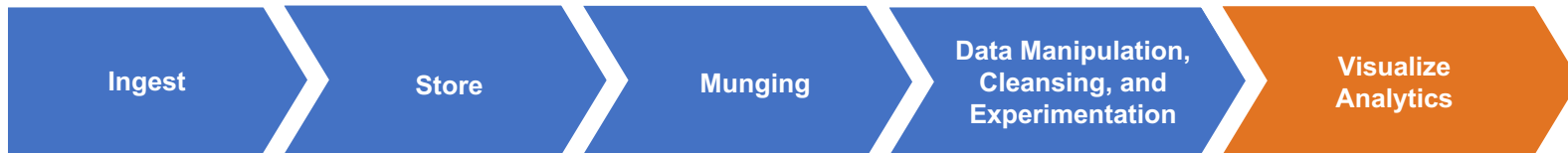
data
iku

Shiny





Tenant Lite Sample Use-Case Pipeline



Utilizing curated, or better yet, trusted data in Delta Lake can save data analysis tremendous time



Time to complete data cleaning work in Databricks? For basic use-case, as little as a couple hours.



- Time to write code and configure/customize Plotly Dash Application? Half a day or less for basic use-case
- Time to deploy? Plotly Enterprise is self-service so app can be deployed same day. Plotly Dash will require a help desk ticket.

Time to Completion? $\approx$ 1 day



Databricks: Delta Tables



Data Science & E...

+

Create

Workspace

Recents

Search

Data

Compute

Jobs

Help

Settings

1595009789.GILLI...

Menu options

Database TablesDBFS

Create Table

Databases

Filter Databases

1229700703weaverda_spar...

1229700703weaverda_spar...

1260155470_harrell_r_7qf4...

1501141557larkinsh_spark...

1606301584bolincla_spark...

1606301584bolincla_spark...

4tes_training_db

a9_covid_tracker

a9a_stage

a9test

a9test2022

aaax_db

abacdemo

active_directory

afa4lm_db

afac_cda_db

afac_cda_db

Tables

Filter Tables

g081_aircraft

g081_aircraft_schedule

g081_aircraft_status_history

g081_bases_with_sran

g081_component

g081_debrief

g081_deviation

g081_discrepancy

g081_end_item

g081_engine

g081_engines_join_aircraft

g081_flight

g081_geoloc_icao

g081_mdc

g081_mds

g081_mission

g081_mission_design

Refresh

https://databricks.afdatalab.af.mil/#table/vault_g081_curated/g081_aircraft

UNCLASSIFIED: Distribution A

Integrity | Service | Excellence

NOTE: Synthetic data for demonstration purposes

27



vault_g081_curated.g081_aircraft

Public

Details History

Description:

Created at: 2023-01-05 17:29:26

Last modified: 2023-05-26 04:44:52

Partition columns:

Number of files: 1

Size: 162 kB

Schema:

Showing all 47 rows.

Sample Data:

aircraft_serial_number ▲	possessing_command ▲	mission_design_series ▲	mission_design ▲	aircraft_tail_number ▲	call_sign ▲	model_code ▲	current_aircraft_status ▲	nm
--------------------------	----------------------	-------------------------	------------------	------------------------	-------------	--------------	---------------------------	----



Databricks: Read in Delta Table

Final_G081_Geomap & ARMS Project Python

SAFCO

File Edit View: Standard Permissions Run All Clear

Cmd 6

Python

1 df22 = spark.read.table("vault_g081_curated.g081_aircraft")
2 display(df22)

(2) Spark Jobs

df22: pyspark.sql.dataframe.DataFrame = [aircraft_serial_number: string, possessing_command: string ... 42 more fields]

	4	possessing_command_abbreviation_code	command_code_nomenclature	possessing_organization_no	possessing_organization_name	possessing_organization_type
1		ANG	AIR NATIONAL GU	0176	WGH	WG
2		MTC	AIR FORCE MATER	661	AES	S
3		AFR	AF RESERVE	0089	ALF	SQ
4		AET	EDUCATION TRAIN	0097	MBY	WG
5		RAF	ROYAL AIR FORCE	0002	OFG	SQ
6		RAF	ROYAL AIR FORCE	0002	OFG	SQ
7		AFE	USAF IN EUROPE	0086	ALF	WG

Truncated results, showing first 1000 rows.

Command took 1.60 seconds -- by 1617327232.ESC0BAR.G at 5/22/2023, 2:54:09 PM on SAFCO

Cmd 7

1 row =df22.select('aircraft_tail_number').distinct().count()
2 #row
3 df22.count()

(6) Spark Jobs

UNCLASSIFIED: Distribution A

Integrity | Service | Excellence

NOTE: Synthetic data for demonstration purposes

29



Databricks: Data Prep



G081_Geomap copy (Python)

SAFCO

File Edit View: Standard Permissions Run All Clear

Cmd 13

```
1 #####to get count of last landing, grouped by geohash and command code
2
3 #g081_aircraft_count = ((spark.read.table("vault_g081_curated.g081_aircraft")
4 #                               .selectExpr("aircraft_serial_number", "last_landing_geohash", 'command_code_nomenclature')
5 #                               .groupBy("last_landing_geohash", 'command_code_nomenclature')
6 #                               .count()
7 #                               ))
8 #g081_aircraft_count.display()
```

Command took 0.02 seconds -- by 1617327232.ESCOBAR.G at 5/1/2023, 3:41:53 PM on SAFCO

Cmd 14

```
1 #####to drop null geohash values
2 g081_aircraft_count1 = g081_aircraft_count1.dropna(subset=['last_landing_geohash'])
3 ##### convert pyspark sql dataframe to pandas
4 g081_aircraft_count_pandas = g081_aircraft_count1.toPandas()
5 #####to separate geohash into latitude and longitude
6 lat = []
7 lon = []
8
9 # For each row in a variable,
10 for row in g081_aircraft_count_pandas['last_landing_geohash']:
11     # Try to,
12     try:
13         # Split the row by comma and append
14         # everything before the comma to lat
15         lat.append(row.split(',')[0])
16         # Split the row by comma and append
17         # everything after the comma to lon
18         lon.append(row.split(',')[1])
```



Databricks: Machine Learning



Final_G081_Geomap & ARMS Project Python

SAFCO

File Edit View: Standard Permissions Run All Clear

ML Implementation

Cmd 31

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import sklearn
5 from sklearn.tree import DecisionTreeClassifier
6 from sklearn.model_selection import cross_val_score, train_test_split
7 from sklearn.preprocessing import OneHotEncoder
8
9 X = morsDfnumeric
10 y = morsDf['BMC']
11 y1 = morsDf["CMR"]
12 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)
13 print("Input features: ", list(X_train.columns))
14 print("\nTarget features: ", y_train.name)
15
16 #Tree Classifier
17 tree_clf = DecisionTreeClassifier(max_depth=8)
18 #Fit Classifier
19 tree_clf.fit(X_train, y_train)
20 #Create Predictions
21 preds = tree_clf.predict(X_test)
22
23 #Weighted F1 Score for Predictions, Cross Validated
24 tree_scores = cross_val_score(tree_clf, X_train, y_train, scoring='f1_weighted', cv=10)
25 tree_scores
```



Databricks: Machine Learning



Final_G081_Geomap & ARMS Project Python

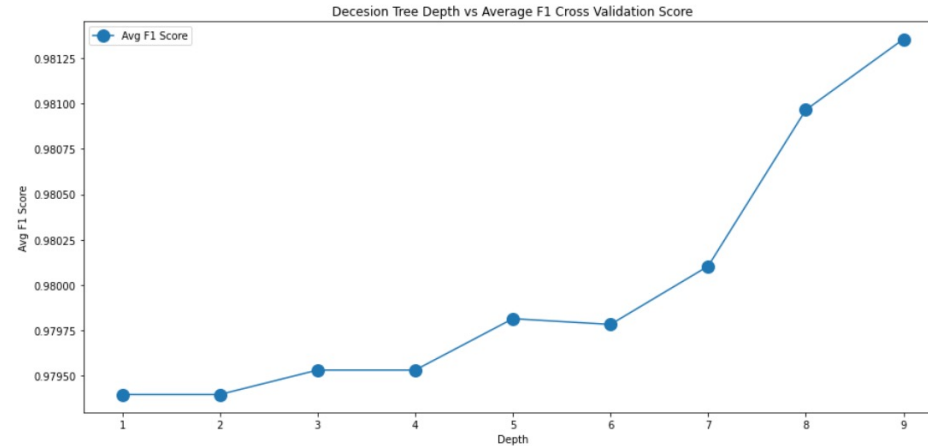
SAFCO File Edit View: Standard Permissions Run All Clear

Schedule

Comments Experiment Revision history

```
1 plt.figure(figsize=(15,7))
2 plt.plot(depth_df.depth,depth_df.avg_f1, linestyle='-', marker="o", mew=4, markersize=9, label="Avg F1 Score")
3 plt.legend()
4 plt.title("Decesion Tree Depth vs Average F1 Cross Validation Score")
5 plt.xlabel('Depth')
6 plt.ylabel('Avg F1 Score')
```

Out[370]:



Text(0, 0.5, 'Avg F1 Score')



Databricks: Plotly Dash Graph Configuration

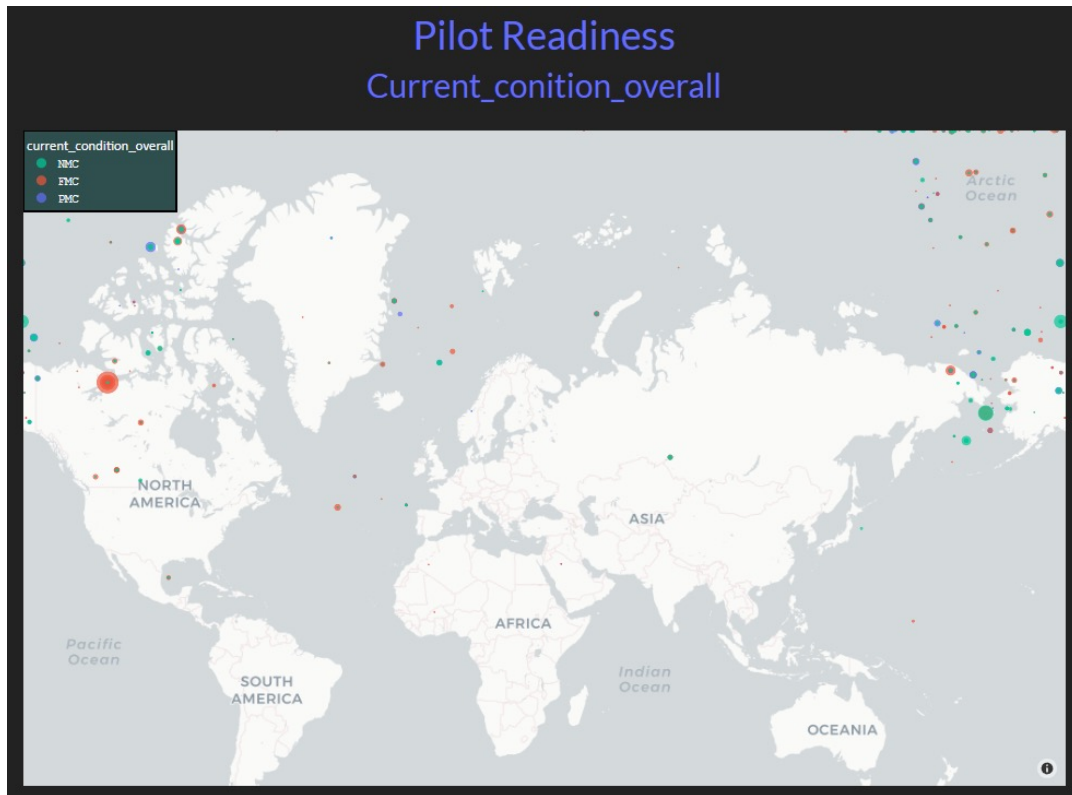


```
Final_G081_Geomap & ARMS Project Python
SAFCO
Command took 0.02 seconds -- by a user at 5/10/2023, 1:46:16 PM on unknown cluster
Cmd 103
Python

1  #-----## New layout ##-----
2  ###First Page
3
4  #g081_aircraft_count_pandas
5  this_condition = g081_aircraft_count_pandas["current_condition_overall"]
6  Default = g081_aircraft_count_pandas
7  PMC = g081_aircraft_count_pandas[g081_aircraft_count_pandas["current_condition_overall"]=="PMC"]
8  FMC = g081_aircraft_count_pandas[g081_aircraft_count_pandas["current_condition_overall"]=="FMC"]
9  NMC = g081_aircraft_count_pandas[g081_aircraft_count_pandas["current_condition_overall"]=="NMC"]
10
11 #condition = g081_aircraft_count_pandas["current_condition_overall"]
12 ### Home Page
13 #Title
14 start_layout = dbc.Container([
15     html.H1("Pilot Readiness", style = {'textAlign': 'center', 'color': '#636EFA'}),
16     html.H2("Current_conitio_n_overall", style = {'textAlign': 'center', 'color': '#636EFA'}),
17
18     html.Br(),
19
20     #Graph
21     dbc.Row([
22         dbc.Col([
23             dcc.Graph(id = 'map', figure = px.scatter_mapbox(g081_aircraft_count_pandas, lat="latitude",lon="longitude",color = 'current_condition_overall',size = 'count',hover_name =
24             ['latitude','longitude','count','command_code_nomenclature']).update_layout(legend=dict(x=0,y=1,traceorder="reversed",title_font_family="Calibri",font=dict(family="Courier",size=14,c
25             olor="ghostwhite"),bgcolor="darkslategray",bordercolor="Black",borderwidth=2,height=800,mapbox={"style": "carto-positron", "zoom":2}, margin={"l": 0, "r": 0, "t": 0, "b": 0}))
26         ]),
27     html.Br(),
28     dbc.Row([
```




Plotly Dash

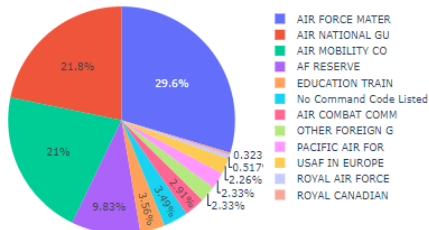




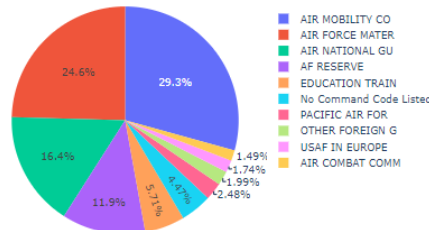
Plotly Dash



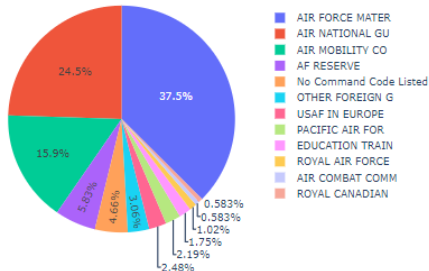
Percentage of mission design series per location



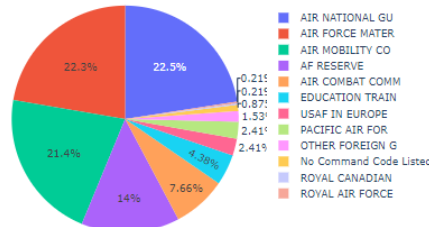
Percentage of PMC per location



Percentage of FMC per location



Percentage of NMC per location

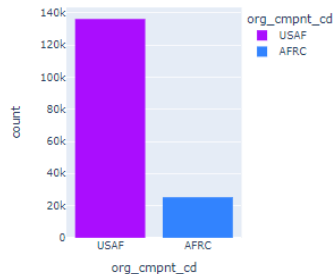




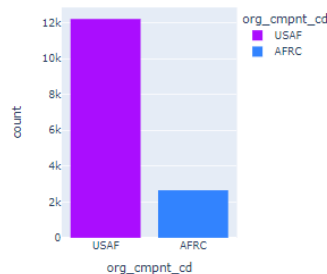
Plotly Dash



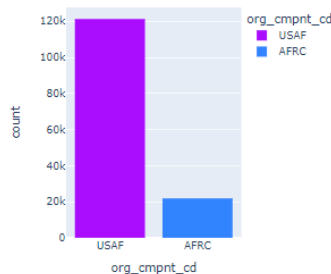
Count of F035A Pilots per Org



Count of Simulators per Org



Count of Sorties per Org



BMC Pilot per Year



CMR Pilot per Year





Other Use-Case Pipelines: Tableau and Dataiku



Time to Completion? With just a few clicks we will have a geographical scatter map



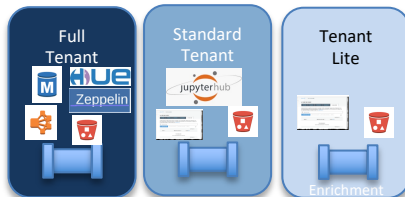
VAULT Data Platform - Summary



VAULT enables users data access and utilization thru multiple offerings: dedicated tenant spaces, data acquisition, enterprise data, an enterprise tool suite and Bring Your Own Tools and a suite of tools that enable data exploitation activities.

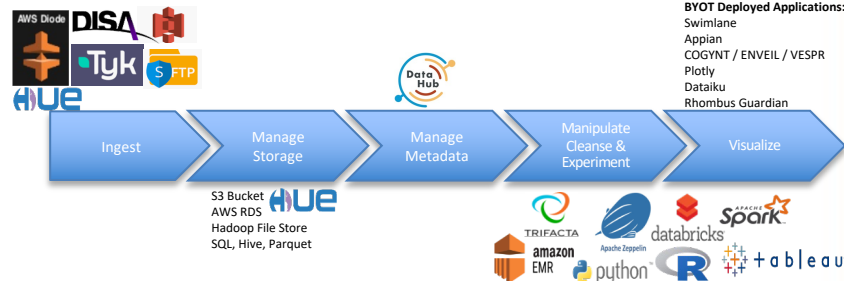


- Users can **self provision** to VAULT
 - NIPR: <https://afdatalab.af.mil/>
 - SIPR: <https://afdatalab.af.smil.mil>
- Users can work in the following areas:
 - VAULT Default Tenant – access to Enterprise Data Area & Enterprise Tools (Cannot access limited/controlled data)
 - New Tenant Space or Existing Tenant Space
 - Access to Enterprise Data, Enterprise Tools, Limited/Controlled Data <as applicable>
 - Supports BYOT Deployment



- A VAULT Tenant space is a private workspace for storing, building and maintaining organization-specific data and workflows
- VAULT Tenant Space Offerings
 - Full** <VAULT Default> = S3, Hue, Zeppelin, RDS & EMR
 - Standard** = S3, JupyterHub <in work>
 - Lite** = S3 and Console
- Each tenant option supports different use cases and comes with different cost models

- VAULT offers a unique capability to **Bring Your Own Tool (BYOT)** to the platform and the User Tenant Space.
 - These tools are GOTS, COTS or custom code tools that a tenant space may need to support data exploration and visualization



- VAULT offers a constantly expanding **Enterprise Tool Suite** to support a full lifecycle of data exploitation activities customizable to the users.
 - Note: Some tools require licensing or participation in cost sharing

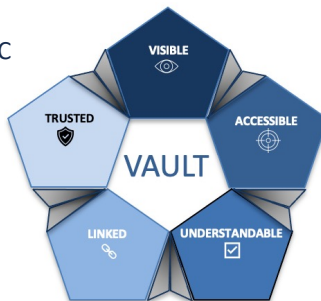


VAULT Data Platform – Mission Support



VAULT Supports Critical Missions & Delivers Value to the Department of the Air Force
VAULT Data Platform Supports over 110 Dedicated Workspaces “Tenants” across NIPR & SIPR

38th Engineering Squadron	A1ZR	AFGSC	AFRL-RYWA	Analytics & Assessment Directorate	Cyberworx	FLTS	MROI	SAF/CND
40 FLTS	ACC	AFIMSC	AFSC	AWPPAO	DAFCINT	HAF/DSI	myFDdata	SAF/CO
412 TW-PA	ACC-A3	AFIT-USAFA	AFSOC	BESPIN	DSAF	HBAT	NGBFM	SAF/SA
417 FLTS	ACTF	AFMAA	AFSVC		ElTaaS	HPSRF	OMBPC	SAFMG
46TS-TGBB	AETC A1	AFMC-FM				HQ AETC CSS	OSD-CAPE	SMARTFM
76sweg	AF/HO	AFNWC-NC3				Kessel Run	OSI	TruDat
96 CTG Data WG	AFA4LM	AFOSI-SFMIS				KRADOS	PMTEC-Tableau	UDTAI
96 RNCS	AFAC	AFOTEC	AFTC	C-19	F-22	LogUot III	Readiness Pathfinder	USAF-TestPilotSchool
A1 myData	AFBIT	AFRC	AFWIC	CON-IT	F15EXPO	MDACOST	Rhombus Guardian	USSOCOM
A1XD	AFCAA	AFRL-CDAO	AMCA2Y	CTIO/S9	F35	Minute Man III	SAF-IE	Vulcan

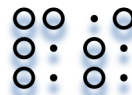




VAULT Data Platform: Fact Sheet



Number of Accounts:
16,000+



Data Feeds: ~120



Tenants:
~124



Data Files: ~473 Million



VDP Instances:
NIPR, SIPR, JWICS <In Work>



Data Storage/Size: ~1,570 TB



Bring Your Own Tool (BYOT):
~20 Deployed



VAULT Enterprise Tools: ~20

**Data represents NIPR & SIPR data



VAULT Data Platform - Access Info



- To access VAULT Data Platform:
 - On NIPR, go to ---> <https://afdatalab.af.mil/>
 - Note: make sure you're connected via .mil domain or VPN
 - On SIPR, go to --> <https://afdatalab.af.smil.mil/>
- To request a VAULT user account
 - Click the "Request an Account" tile
- Need assistance?
 - Email afcdoprod@servicenowservices.com -- this will generate a ticket and the appropriate person will contact you
 - Email AF.Vaultcustomer.successteam@us.af.mil -- this will put you in touch with our customer success team to help with tool usage or other user-centric support



VAULT Data Platform - Contact Info



Government POCs	Contractor POCs
Mr. Matt Cocuzzi – Chief, Data Ops & Capabilities Development Division – SAF/CND • matthew.cocuzzi@us.af.mil	Ms. Marlu Oswald – Seed Innovations CEO & President • marluoswald@seed-innovations.com
Lt Col William Parker – VAULT Director – SAF/CND • william.parker.3@us.af.mil	Ms. Krista Wright – Seed Innovations COO & VAULT Program Manager • kristawright@seed-innovations.com
Lt Col David Tungesvik – VAULT Liaison – SAF/CND • david.tungesvik.1@spaceforce.mil	Mr. Toby Oswald – Seed Innovations CIO & VAULT Chief Architect • tobyoswald@seed-innovations.com
Mr. Jay Roberts – VAULT Technical Lead – SAF/CND • isaac.roberts.2@us.af.mil Ms. Clare Bolin, Mr. Mark Bryant, Mr. Gerson Escobar, Mr. Christopher Gillie, Ms. Sepideh Maharati – VAULT SMEs – SAF/CND • clare.bolin@us.af.mil, mark.bryant.16@us.af.mil, gerson.escobar@us.af.mil, christopher.gillie.2@us.af.mil, sepideh.maharati@us.af.mil	



QUESTIONS?



BACKUP SLIDES

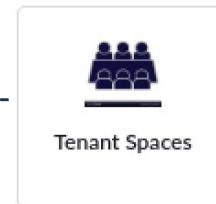
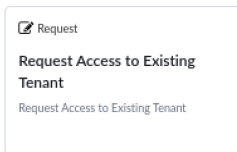


VAULT Data Platform - Access Info



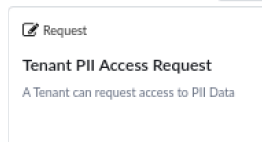
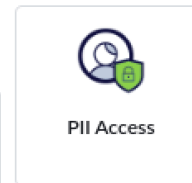
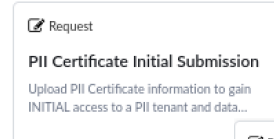
- Request to be added to an existing VAULT Tenant:

- Select "Tenant Spaces" from the Popular Links section on the main page -
- Select "Request Access to Existing Tenant" -



- Submit a New/Updated PII Certificate and Request Tenant PII Access

- Select "PII Access" from the Popular Links section on the main page -
 - Select "Initial Submission" or "Certification Update" -
- Select "Tenant PII Access Request" -



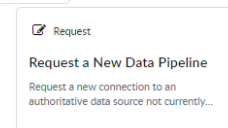
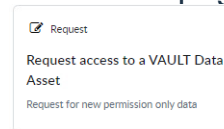
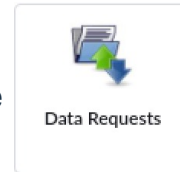


VAULT Data Platform - Access Info



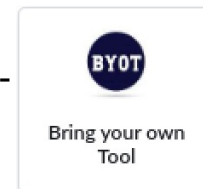
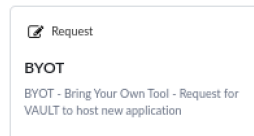
■ To Request Data in VAULT

- Select "Data Requests" from the Popular Topics section on the main page
 - Select "Request Access to a VAULT Data Asset"
 - For data already available within the VAULT
 - Select "Request a New Data Pipeline"
 - To request a new data asset be provisioned into VAULT
- Complete the forms
 - Upon submittal - the data request will be queued by priority to be worked



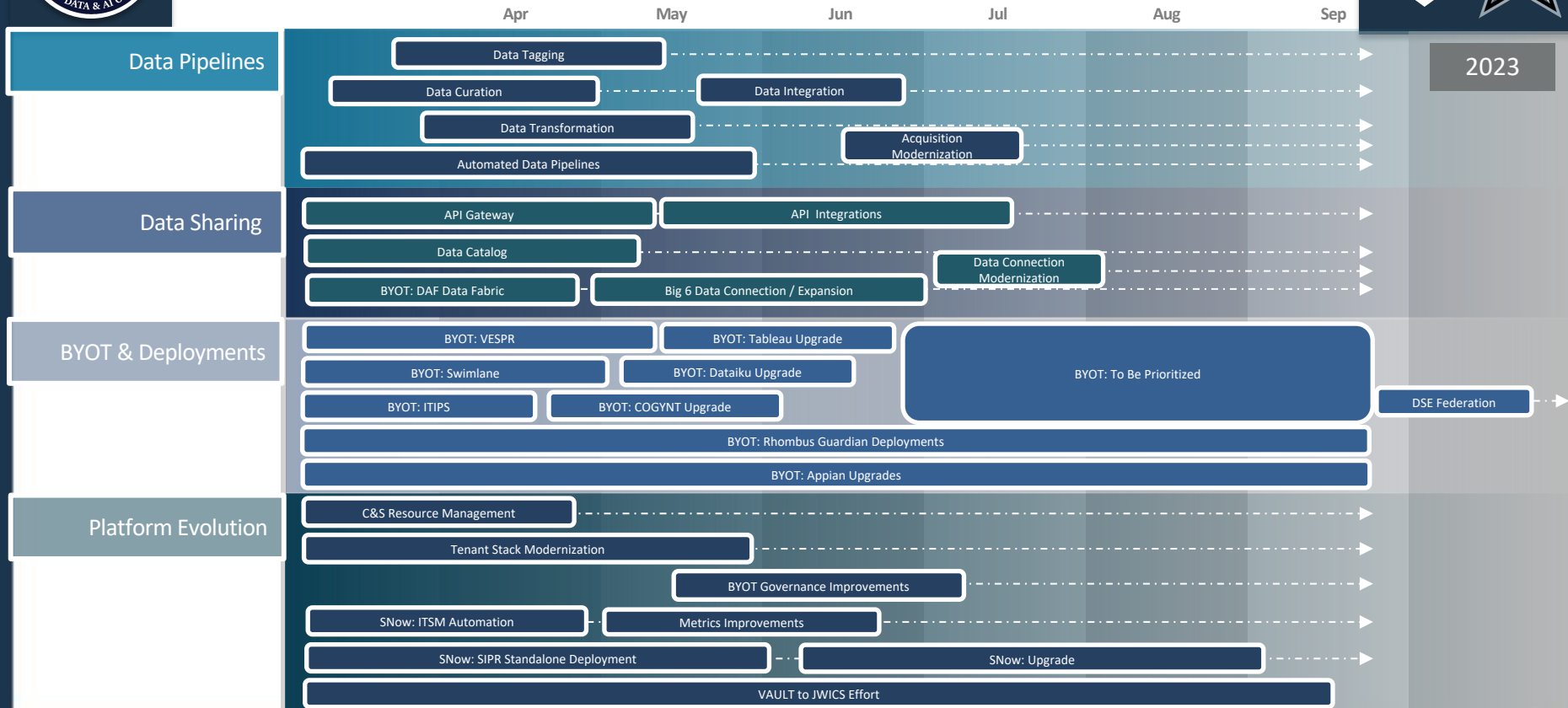
■ To submit a BYOT request

- Select "Bring Your Own Tool" from the Popular Topics section on the main page -
- Select "BYOT" -
- Fill out the form to begin the process





VAULT Data Platform - Roadmap

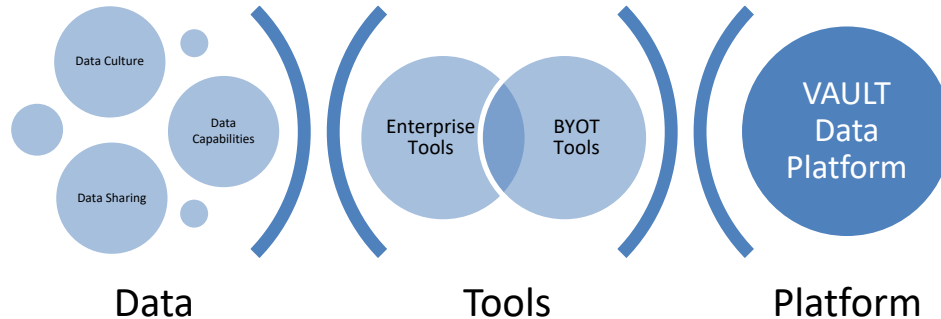




VAULT Data Platform - Summary



- VAULT Focus = Users Supporting the Mission!
 - Visible – Accessible – Understandable – Linked – Trusted
 - VAULT enables users by provisioning key enterprise tools to support data exploration, curation and analysis on the platform
 - VAULT supports more than **16000** accounts and continues to grow!
 - VAULT enables data sharing and has provisioned **120** data feeds with 59 more in progress and ~20 more ready to be delivered when available
 - VAULT has established an automated data sharing tool between Envision and VAULT
 - VAULT currently implementing an API Gateway solution
 - VAULT provides unique and mission focused applications for the users thru the **BYOT** offering!





VDP Impact – Data Provisioning



VAULT Data Execution

VAULT enables data sharing thru automated acquisition and data discovery, to user enabled enrichment and visualization using enterprise & BYOT tools and finally promoting and sharing the data to the platform and the enterprise.

Acquisition

- Enables self service acquisition of data feeds
- Captures Business and Data Source & Set Metadata
- Automated Pipeline established for Envision Data Sharing

DMZ

- VDP Protection Pipeline
- Protects VDP from security threats hidden within the data
- Performs Scanning of Data
- Quarantine & Archiving of Data

Promotion

- Promotes Data to the Enterprise Data Area or directly to the tenant space
- Verify/Validate and tag data based upon information collected during the acquisition phase
- Captures File Based Metadata
- Performs Data Transformation <if required>



Enrichment

- Data Enrichment Pipeline
- Performed by the VAULT users
- Performed using Enterprise and BYOT tool suites
- Data Maturation, Curation, Linking, Wrangling
- Enrichment Pipeline is within a VAULT user space

Egress

- Pipelines to enable Data Egress from VAULT
- Data Egress from VAULT or Cross Domain
- Enables Auditing, Tagging and Cataloging

VAULT Data Catalog

VAULT Data Area



VAULT Data Platform – Foundational Capability



VAULT drives key behaviors to demonstrate “Visible, Accessible, Understandable, Linked & Trusted”

- Enabling **INTEROPERABILITY** through a collaborative model to support Data Sharing across data platforms and the enterprise!
- Enable **SELF-SERVICE** through the deployment of tools to support user needs without creating vendor lock in the platform
- Automated and rapid deployments to ensure platform stability and relevancy following the **DoD Enterprise DevSecOps Reference Design**
 - CI/CD Pipelines
 - Kubernetes Deployments enabling an automatically scaling platform

VDP Foundational Capability

Self Provisioning to VAULT Data Platform	Identity Control Access Management (ICAM)	API Provisioning* API Gateway & API/Tableau Integration
Constantly Expanding Enterprise Tool Offering	Attribute Based Access Control*	Data Access Management
Bring Your Own Tool Enabling users to customize their data exploration experience	Data Provisioning & Data Sharing Automated / Self-Service data sharing pipeline with Envision Data Feed pipelines to tenant space and/or Enterprise Area	Automated Data Pipelines Data Acquisition Pipeline Data Scanning and Quarantine Transform Pipeline* Promotion Pipelines*
Jupyter Hub Capabilities* Provides robust capability to VAULT users for complex and large data computations.	Data Catalog * Data Catalog containing all platform data information Capable of integrating with Enterprise data catalog	Automated Metadata Definition*
Automated Plotly CI/CD Pipelines	Data Tagging*	* Capability available soon



PlotlyDash Application and the data source in Envision



CUJ/NF

1390346286 > Score > Data > App: SUS

Try out the new and improved search experience. [Turn on](#)

Search: g081

Clear X

Files 50+ Objects 7.3k People 2 Apps 0

Only show

- ☐ Catalog items
- ☐ In Trash

Tags

Filter by tag

Projects

Filter by project

In path

Created by

Search for people...

Types Sorted descending

Add additional filters

G081	/Envision_NIPR/Envision (Air Force) PM Team/Exports for VAULT/...
g081_mission	/Envision_NIPR/G081_Datasource/exports/vault-exports-parque...
g081_mds	/Envision_NIPR/G081_Datasource/exports/vault-exports-parque...
g081_mission	/Envision_NIPR/G081_Datasource/exports/parquet-datasets/g08...
g081_mds	/Envision_NIPR/G081_Datasource/exports/parquet-datasets/g08...
g081_flight	/Envision_NIPR/G081_Datasource/exports/parquet-datasets/g08...
g081_discrepancy	/Envision_NIPR/G081_Datasource/exports/parquet-datasets/g08...
g081_aircraft	/Envision_NIPR/G081_Datasource/exports/parquet-datasets/g08...
g081_debrief	/Envision_NIPR/G081_Datasource/data/api/g081_debrief
g081_aircraft	/Envision_NIPR/G081_Datasource/data/api/g081_aircraft

Report issue

No description

Created Dec 8, 2021, 9:22 AM by [Clare Bolin](#)

Location /Envision_NIPR/Envision (Air Fo...

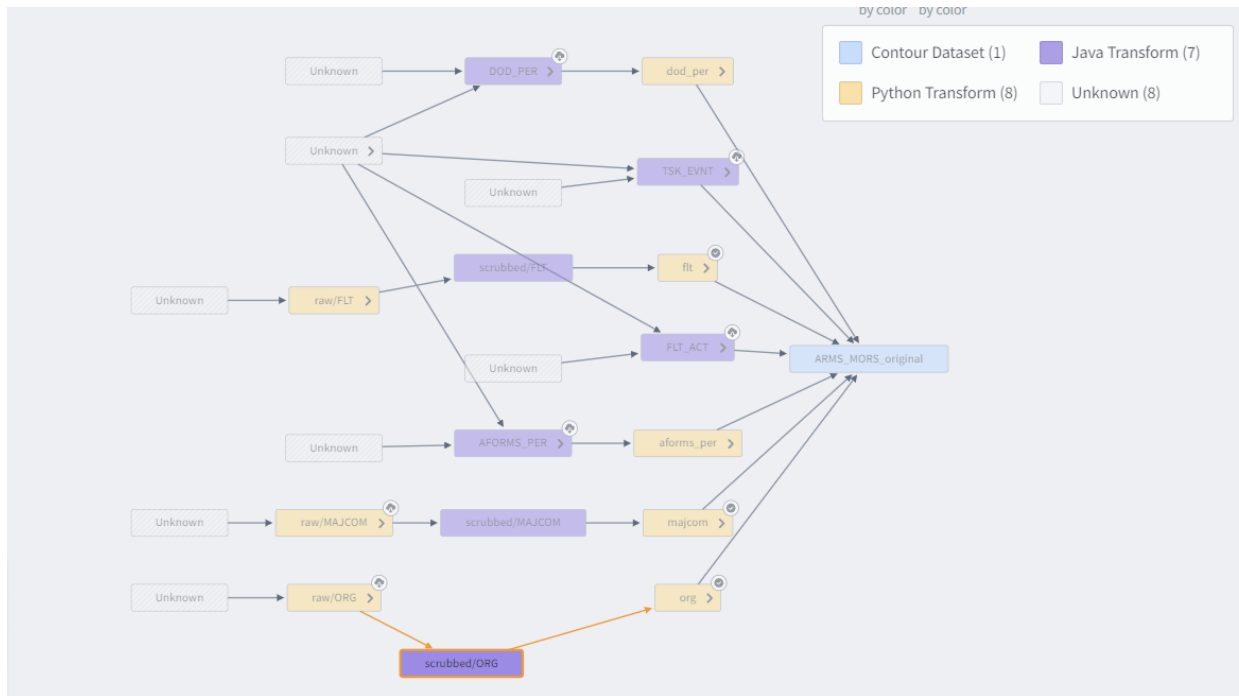
Type Folder

[Show more](#)

Explore data lineage [Open](#)



PlotlyDash Application and the data source in Envision





Databricks: Pip Installs/Imports



Final_G081_Geomap & ARMS Project Python

SAFCO

File Edit View: Standard Permissions Run All Clear Comments Experiment Revision history

```
1 pip install dash-bootstrap-components
```

Show result

Cmd 5

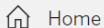
```
1 # imports and variables
2 import dash_bootstrap_components as dbc
3 from pyspark.sql.functions import *
4 import pandas as pd
5 import numpy as np
6 from pyspark.sql.window import Window
7 from pyspark.sql.functions import col
8 import pandas as pd
9 from datetime import datetime
10 import calendar
11 from dash import Dash, html, dcc
12 import plotly.express as px
13 import pandas as pd
14 from pyspark.sql.types import *
15 import dash
16 from dash import dcc
17 import dash_html_components as html
18 import dash_table as dt
19 from dash.dependencies import Input, Output
20 import plotly.graph_objects as go
21 import plotly.express as px
22 import random
23 import string
24 import flask
```



Tableau Server: Home



Default



Home



Favorites



Recents



Shared with Me



Recommendations



Personal Space



Collections



Explore



External Assets

Search for views, metrics, workbooks, and more



M

Home

Your Tableau site—where analytics and your organization meet

Harness your data. Discover opportunities. Elevate your insights.

New

Project

Workbook

Flow

Published Data Source

Virtual Connection

Upload Workbook

Tableau Desktop | Download Tableau Prep Builder

User can refresh their Desktop and Prep keys on their own. Expect new keys available by 7 Jan 2021.

/pro/desktop/en-us/license_subscriptions_enterprise.htm

Favorites

See All



Tableau Server: Connect to Databricks



FileDataWorksheetDashboardAnalysisMapFormatHelp

New Workbook

Publish As...

MARK.BRYAN...

×

←

Connect to Data

×

Connect to the data you need to visualize. [Learn more](#)

On This SiteFilesConnectors

Alibaba AnalyticDB for MySQL	Amazon EMR Hadoop Hive	Azure Synapse Analytics	Denodo
Alibaba Data Lake Analytics	Amazon Redshift	Box	Dremio
Alibaba MaxCompute	Apache Drill	Cloudera Hadoop	Dropbox
Amazon Athena	Azure Data Lake Storage Gen2	Databricks	Exasol
Amazon Aurora for MySQL	Azure SQL Database	Datorama	Firebird 3

Data Source

Sheet 1





vault_g081_curated.g081_aircraft

File Data Help Publish MARK.BRYAN... Show Me

Connections
databricks.afdatalab.af.mil
Databricks

Catalog
SPARK

Database
vault_g081_curated

Table
Enter table name
☐ Exact ☒ Contains ☐ Starts with
☐ Check permissions

New Custom SQL

g081_aircraft (vault_g081_curated.g081_aircraft) (vault_g081_curated)

Calculation1
FLOAT(SPLIT((REPLACE([Last Landing Geohash], ",", "")), ", ", " "))

The calculation is valid.

Need more data?
Add tables here to relate them. [Learn more](#)

>	c g081_aircraft Signed Organization T...	Abc g081_aircraft Current Condition Statu...	Abc g081_aircraft Current Condition Overall	Abc g081_aircraft Current Condition Detail	Abc g081_aircraft Current Condition Sche...	Abc g081_aircraft Current Condition Use R...	Abc g081_aircraft Current Condition Sum...
G	A	null	NMC	NMCB	Unscheduled	Restricted	NMC-BU
G	B	FMC	FMC	Scheduled	Restricted	FMC	
G	G	NMC	NMCB			NMC-BU	
G	G	PMC	PMCM			PMC-M	
G		FMC	FMC			FMC	

Data Source G081_Geomap G081_Geomap_Dashboard



Tableau Server: Create Scatter Map

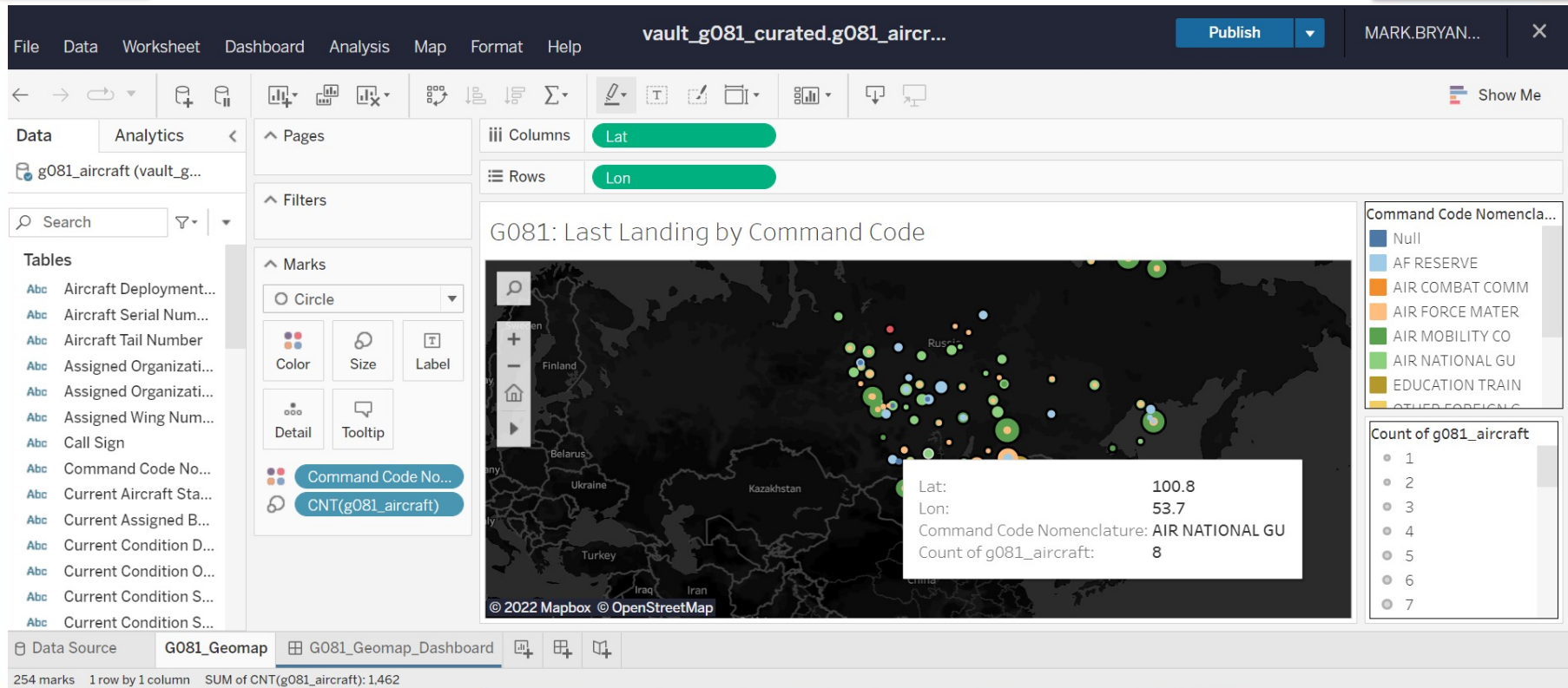
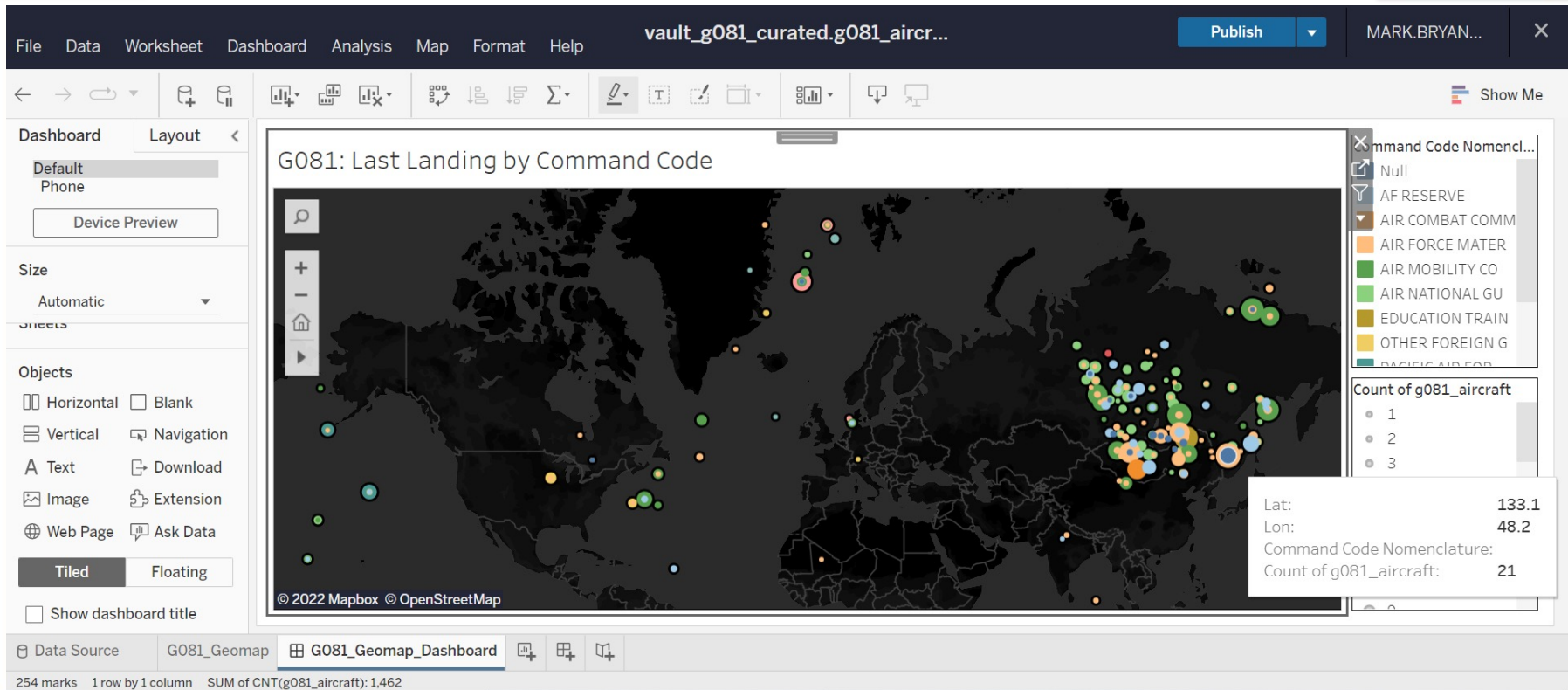




Tableau Server: Dashboard





Dataiku



Dataiku DSS

Filter items...

AFAC

Search DSS...

1

MY ITEMS

See all > ☆

g081 aircraft count ☆
G081_Geomap_Demo

g081 aircraft count prepared ☆
G081_Geomap_Demo

compute g081 aircraft count prepared ☆
G081_Geomap_Demo

New chart on g081 aircraft count prepared ☆
G081_Geomap_Demo

PROJECTS

See all > ☆

G081_Geomap_Demo ☆
The project G081_Geomap was created by MARK.BRYANT.1596494559 on Apr 1
Sandbox

G081 ☆
The project G081 was created by MARK.BRYANT.1596494559 on Apr 08th 2022
Sandbox

Credit Card Fraud Training ☆
The project credit-card-fraud-bryan was created by BRYAN.FAGAN on Nov 29th 2021
+3

Blank project
Create a new blank project

DSS tutorials
Create a tutorial project

Sample projects
Create a project from a sample

Import project
Import a project into the system

APPLICATIONS

See all > ☆

No applications have been shared with you.

GETTING STARTED WITH DSS ^



Dataiku: Project Summary



 G081



AFAC

 Search DSS...

Summary

Activity

Metrics

 ACTIONS



G081

The project *G081* was created by MARK.BRYANT.1596494559 on Apr 15th 2022

1

master

+ Add Projects Markings + Add tags

Sandbox

WATCH 1

STAR 0

Flow

Lab

Dashboards

Wiki

Tasks

 0 DATASETS

 0 RECIPES

 0 NOTEBOOKS

 0 ANALYSES

 1 DASHBOARD

 0 ARTICLES

 3 TASKS

 0 MODELS

+ IMPORT YOUR FIRST DATASET

<https://dataikudesign.afdatalab.af.mil/projects/G081/datasets/new/>

TIMELINE

TODAY

1

You created dashboard

G081's default dashboard

11:40

1

You created project

G081

11:40



Dataiku: Import Data Set



G081

Datasets

AFAC

Search DSS...

1

New Uploaded Files dataset

New dataset name: CREATE

Files

Format / Preview

Schema

Partitioning

Advanced

/g081_aircraft_count.csv

12.04 KB

Drag and drop your files here or ADD A FILE

Creates a single dataset: multiple files must have the same schema.

Store on

Default (in DSS data dir.)

[Show Advanced options](#)

TEST

✓ Used `/g081_aircraft_count.csv` (12.04 KB) to parse data

✓ Used format csv and found 3 columns

PREVIEW >



Dataiku: New Data Preparation Recipe



The screenshot displays the Dataiku interface with a 'New data preparation recipe' dialog box open. The dialog box is divided into two main sections: 'Input dataset' and 'Output dataset'. In the 'Input dataset' section, a dropdown menu shows 'g081_aircraft_count' as the selected dataset. The 'Output dataset' section contains three fields: 'Name' with the value 'g081_aircraft_count_prepared', 'Store into' with the value 'filesystem_managed', and 'Format' with the value 'CSV'. Below these fields, there is a toggle switch labeled 'NEW DATASET | USE EXISTING DATASET'. At the bottom right of the dialog box, there are two buttons: 'CANCEL' and 'CREATE RECIPE'. The background of the screenshot shows the Dataiku interface with a search bar at the top, a list of datasets on the left, and a sidebar on the right containing various recipe types such as Sync, Prepare, Sample/Filter, Group, Distinct, Window, Join with..., Fuzzy join, Split, Top N, Sort, Pivot, Stack, and Code recipes (Python, R, SQL, Shell).



Dataiku: Begin Visual Analysis



G081 **compute** **Visual Analyses** **copy_prepared** **AFAC** **Search DSS...**

Script **Design Sample** **Script output** **Summary** **Script** **Input / Output** **Advanced** **History** **SAVE** **ACTIONS**

Script 5 steps **Design Sample** 261 rows 9 cols **Script output** 254 rows, 6 cols (254 → 7)

Remove rows where last_landing_geohash is null -7

Remove columns Lon, Lat, lon, lat 254

Create GeoPoint from Latitude_scrubbed & Longitude_scrubbed 254

Move Latitude_scrubbed after Longitude_scrubbed

Rename 2 columns 254

Renamings

- Longitude_scr → Lat_scrubbed
- Latitude_scr → Lon_scrubbed

+ ADD RENAMING

Raw text edit

RUN

command_code_nomenclature	count	Lat_scrubbed	Lon_scrubbed	Geopoint_scrubbed
string Text	bigint Integer	double Decimal	double Decimal	string GeoPoint
AIR NATIONAL GU	5	-107.316803	48.06110001	POINT(-107.316803 48.06110001)
AF RESERVE	7	-58.56300354	50.86389923	POINT(-58.56300354 50.86389923)
AIR FORCE MATER	1	-68.02200317	49.21160126	POINT(-68.02200317 49.21160126)
AF RESERVE	6	-101.0538025	46.8927002	POINT(-101.0538025 46.8927002)
AIR FORCE MATER	1	40.6519928	54.25149918	POINT(40.6519928 54.25149918)
AIR NATIONAL GU	3	-94.264	51.8256	POINT(-94.264 51.8256)
AIR FORCE MATER	1	-90.30670166	49.33580017	POINT(-90.30670166 49.33580017)
AIR FORCE MATER	3	-98.79789734	57.87239838	POINT(-98.79789734 57.87239838)
AIR NATIONAL GU	9	-109.1766968	46.92210007	POINT(-109.1766968 46.92210007)
AIR FORCE MATER	1	-109.3115997	-56.4382019	POINT(-109.3115997 -56.4382019)
null	1	-96.15979767	50.15969849	POINT(-96.15979767 50.15969849)
AIR NATIONAL GU	13	-87.85030365	55.08309937	POINT(-87.85030365 55.08309937)
AF RESERVE	4	136.8404999	78.45269966	POINT(136.8404999 78.45269966)
AIR FORCE MATER	1	-93.84790039	49.35189819	POINT(-93.84790039 49.35189819)
AF RESERVE	1	-58.07299805	51.73730087	POINT(-58.07299805 51.73730087)
AIR FORCE MATER	2	-80.145401	57.57920074	POINT(-80.145401 57.57920074)

254 matching rows



Dataiku: Data Prep



G081

Recipes

AFAC

Search DSS...

1

compute_g081_aircraft_count_prepared

Summary

Script

Input / Output

Advanced

History

SAVE

ACTIONS

Script

Design Sample

Viewing design sample

no step

261 rows 3 cols

261 rows, 3 cols

DISPLAY

Table View

Grid View

Chart View

No step yet in your script.

You can add steps either by:

Clicking on column headers.

John Doe

Highlighting some text.

Clicking on cells.

John Doe

Paste if you've copied steps.

OR

+ ADD A NEW STEP

RUN

Local stream

Split column on null

Replace null by ...

Extract text like null...

Remove rows containing text like null...

Keep rows containing text like null...

Flag rows containing text like null...

Remove rows containing null

Keep only rows containing null

Remove rows equal to null

Keep only rows equal to null

Clear cells equal to null

Filter on null

Filter on null

Show complete value

Toggle row highlight

Copy row as JSON

	clature	count
		bigint
		Integer
61.2		2
61.2		2
61.2		6
61.2		1
64.6		8
null		51
null		6
null		2
null	AIR COMBAT COMM	57
null	AIR MOBILITY CO	6
null	AIR FORCE MATER	15
null	AIR NATIONAL GU	1



Dataiku: Convert Column Types



Summary Script Input / Output Advanced History SAVE ACTIONS

DISPLAY Table Grid Chart

254 matching rows

command_code_nomenclature	count	Latitude_scrubbed	Longitude_scrubbed
string Text	bigint Integer	string Decimal	double Decimal
AIR NATIONAL GU	1	Auto-detect	-82.73179626
AIR MOBILITY CO	2	Text	-83.24079895
null	2	Decimal	-30.19299316
AIR FORCE MATER	14	Boolean	Latitude
null	5	Date	Longitude
AIR MOBILITY CO	12	Object	GeoPoint
AIR NATIONAL GU	9	Array	Geometry
AIR MOBILITY CO	2	Natural lang.	Country
AIR NATIONAL GU	5	Geospatial	US State
AF RESERVE	7	Web	-64.96600342
AIR FORCE MATER	1	Other	-107.316803
AF RESERVE	6	User-defined...	-58.56300354
AIR FORCE MATER	1	49.21160126	-68.02200317
AF RESERVE	6	46.8927002	-101.0538025
AIR FORCE MATER	1	54.25149918	40.6519928



Dataiku: Split Geohash into Lat/Lon



compute_g081_aircraft_count_prepared

Script | Design Sample | Script output

1 step | 261 rows 3 cols | 254 rows, 3 cols (-7)

Remove rows where last_landing_geohash is null (-7)

+ ADD A NEW STEP

+ ADD A GROUP

RUN Local stream

Script output: 254 matching rows

last_landing_geohash	command_code_nomenclature	count
-27.6406,1		7
-33.393001		1
-33.561798		1
-43.489398		2
11.547300		4
13.4815,2		1
13.584,144		5
13.584,144		3
13.584,144		1
13.584,144		1
17.701900		1
18.439399		2
21.32062,-157.924228	PACIFIC AIR FOR	3



Dataiku: Create Column with type Geopoint from Lat/Lon



Summary Script Input / Output Advanced History SAVE ACTIONS

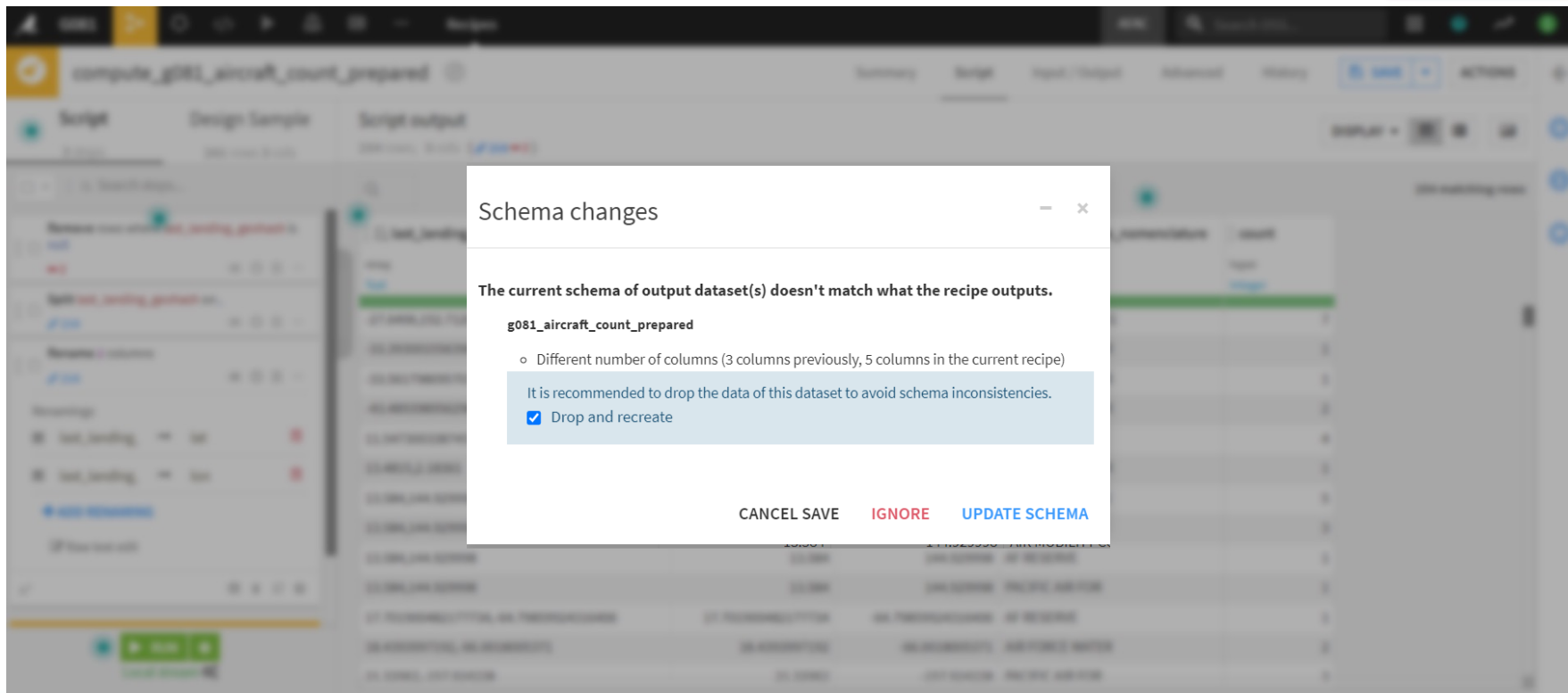
DISPLAY Table Grid Chart

254 matching rows

command_code_nomenclature	count	Lat_scrubbed	Lon_scrubbed	Geopoint_scrubbed
string	bigint	double	double	string
Text	Integer	Decimal	Decimal	GeoPoint
AIR FORCE MATER	1	-68.02200317	49.21160126	POINT(-68.02200317 49.21160126)
AIR NATIONAL GU	5	-107.316803	48.06110001	POINT(-107.316803 48.06110001)
AF RESERVE	7	-58.56300354	50.86389923	POINT(-58.56300354 50.86389923)
AIR FORCE MATER	1	-68.02200317	49.21160126	POINT(-68.02200317 49.21160126)
AF RESERVE	6	-101.0538025	46.8927002	POINT(-101.0538025 46.8927002)
AIR FORCE MATER	1	40.6519928	54.25149918	POINT(40.6519928 54.25149918)
AIR NATIONAL GU	3	-94.264	51.8256	POINT(-94.264 51.8256)
AIR FORCE MATER	1	-90.30670166	49.33580017	POINT(-90.30670166 49.33580017)
AIR FORCE MATER	3	-98.79789734	57.87239838	POINT(-98.79789734 57.87239838)
AIR NATIONAL GU	9	-109.1766968	46.92210007	POINT(-109.1766968 46.92210007)
AIR FORCE MATER	1	-109.3115997	-56.4382019	POINT(-109.3115997 -56.4382019)
null	1	-96.15979767	50.15969849	POINT(-96.15979767 50.15969849)
AIR NATIONAL GU	13	-87.85030365	55.08309937	POINT(-87.85030365 55.08309937)
AF RESERVE	4	136.8404999	78.45269966	POINT(136.8404999 78.45269966)
AIR FORCE MATER	1	-93.84790039	49.35189819	POINT(-93.84790039 49.35189819)
AF RESERVE	1	-58.07299805	51.73730087	POINT(-58.07299805 51.73730087)



Dataiku: Save as Prepared Data Set





Dataiku: Create Scatter Map



G081 **AFAC** Search DSS... **1**

Analyze g081_final **Summary** **Script** **Charts** **Models** **DEPLOY SCRIPT** **ACTIONS**

Columns **Sampling & Engine**

Search...

- # Count of records
- A last_landing_geohash
- A command_code_nomenclature
- # count
- # Latitude_scrubbed
- # Longitude_scrubbed
- Geopoint_scrubbed

Maps

Scatter

- Scatter Map
- Geometry Map

Administrative Map

- Bubbles
- Filled

Binned

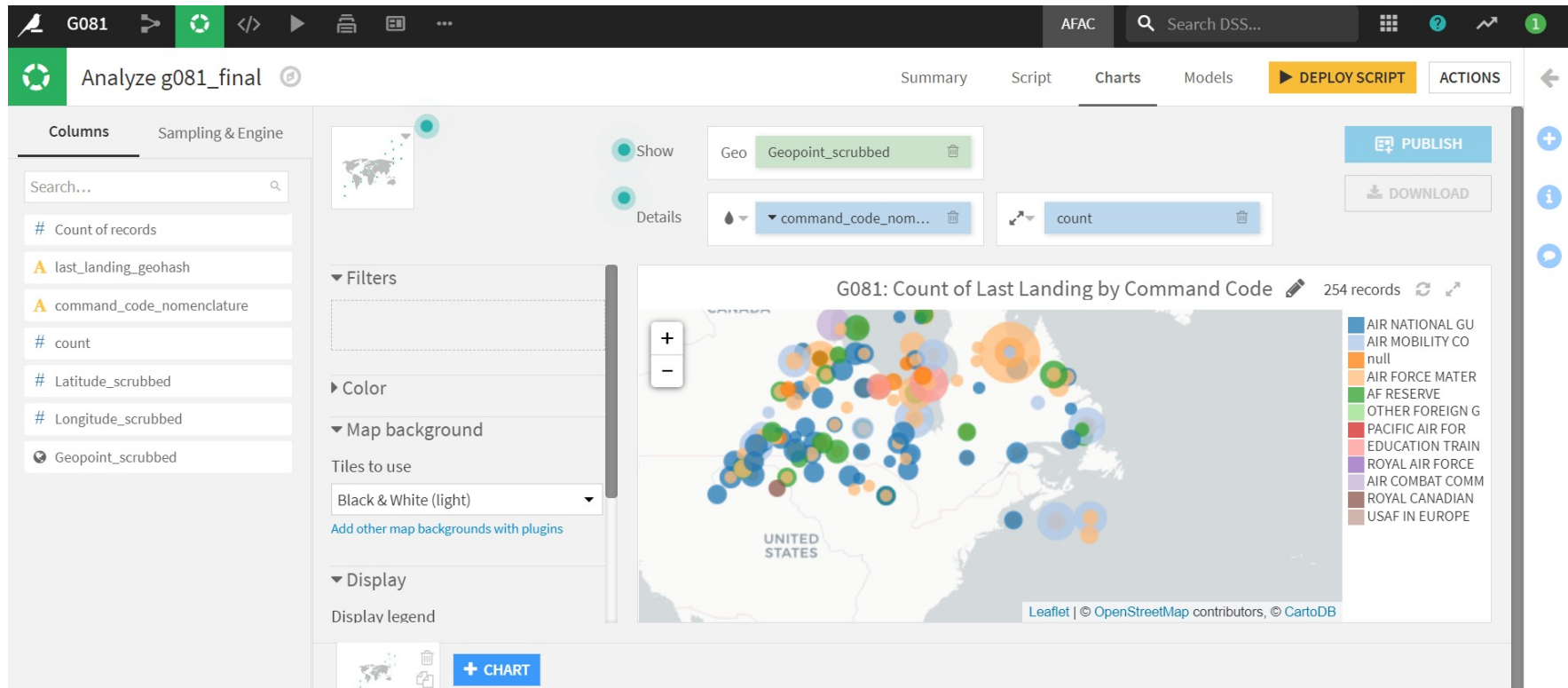
- Grid Map

OTHER

+ CHART



Dataiku: Configure Scatter Map





Dataiku: Publish to Dashboard



g081

Datasets

AFAC

Search DSS...

Summary Explore **Charts** Statistics Status History Settings

PARENT RECIPE ACTIONS

Columns Sampling &...

Search...

Count of records

A last_landing_geohash

A command_code_nomenclature

count

Latitude_scrubbed

Longitude_scrubbed

Geopoint_scrubbed

Create insight and add to dashboard

New chart

☒

This source is not yet shared with dashboard-only users.

☒ Add this source to authorized objects

Dashboards

Slides

1. g081_scrubbed_map - Slide 1

+ ADD TO ANOTHER DASHBOARD

CANCEL CREATE

4 records

- AIR NATIONAL GU
- AIR MOBILITY CO
- null
- AIR FORCE MATER
- AF RESERVE
- OTHER FOREIGN G
- PACIFIC AIR FOR
- EDUCATION TRAIN
- ROYAL AIR FORCE
- AIR COMBAT COMM
- ROYAL CANADIAN
- USAF IN EUROPE

Placement

Disk*

+ CHART

Leaflet | © OpenStreetMap contributors, © CartoDB



Dataiku: Publish to Dashboard

