Department of the Air Force

Integrity - Service - Excellence

Data Analytics at SAF/FMC

Sarah Green
Sept 2024

Outline

- Why VAULT?
- VAULT Architecture/ Tech stack
- Traditional cost processes vs. future vision
- Status and Process for Data Pipelines
- Metadata
- Use cases/ Dataiku

SAF/FMC’s Platform Comparison

	VAULT	Advana	Envision
Platform highlights	- Specializes in analytical tools & custom application sharing - Flexibility on toolsets & “Bring your own Tool” process - End-to-end analysis no coding required but capability for data scientist to do advanced analysis	- Specializes in authoritative data ingestion & visualization -Toolsets fixed by Advana (some flexibility for Enclaves)	- Specializes in “quick start” access to integrated tool suite & established data connections - Palantir tools only (Vendor lock-in)
Ownership	SAF	OSD	SAF
Paradigm	Platform as a Service	Platform as a Service	Software as a Service
Org Workspaces	Tenants	Enclaves	Project Folders
Data ingestion/ PROPIN data handling	Primarily self-service: essential w/ cost community’s PROPIN data	Contracted services	Primarily contracted services
Ease of analytics	-Ability for users to easily pull raw data into end-to-end analytical toolsets to perform data transformations, analytics & custom modeling without coding required -Deployment of custom apps -Easy data extraction	-Requires advanced data science skillsets/programming to perform end-to-end analytics and custom modeling -No custom plug-ins/ custom apps -User feedback indicates difficult data extraction	-Requires advanced data science skillsets/programming to perform end-to-end analytics and custom modeling -No custom plug-ins/ custom apps

Legend	
🟩 X	No code required
🟨 X	Low code required
🟦 X	High Code required
X	Capability- coding n/a
	No capability

Platform	Tool	Data Loading/ Connection*	Data Transformations	ML/AI modeling and prediction	Dashboard Building	Custom Analysis/ Applications	Customizations- Deployed	Project Management Tools
VAULT	Dataiku	🟩 X	🟩 X	🟩 X	🟩 X	🟦 X	🟩 X	X
	Plotly	🟦 X	🟦 X	🟦 X	🟦 X	🟦 X	🟩 X	
VAULT & ADVANA	Tableau	🟩 X	🟩 X	🟦 X	🟩 X			
	Databricks	🟦 X	🟦 X	🟦 X	🟦 X	🟦 X	🟦 X	
ADVANA	Qlik	🟩 X	🟨 X	🟨 X	🟩 X			
	DataRobot	🟦 X	🟦 X	🟩 X				
Envision	Palantir Foundry	🟩 X	🟩 X	🟦 X	🟩 X	🟦 X	🟨 X	X

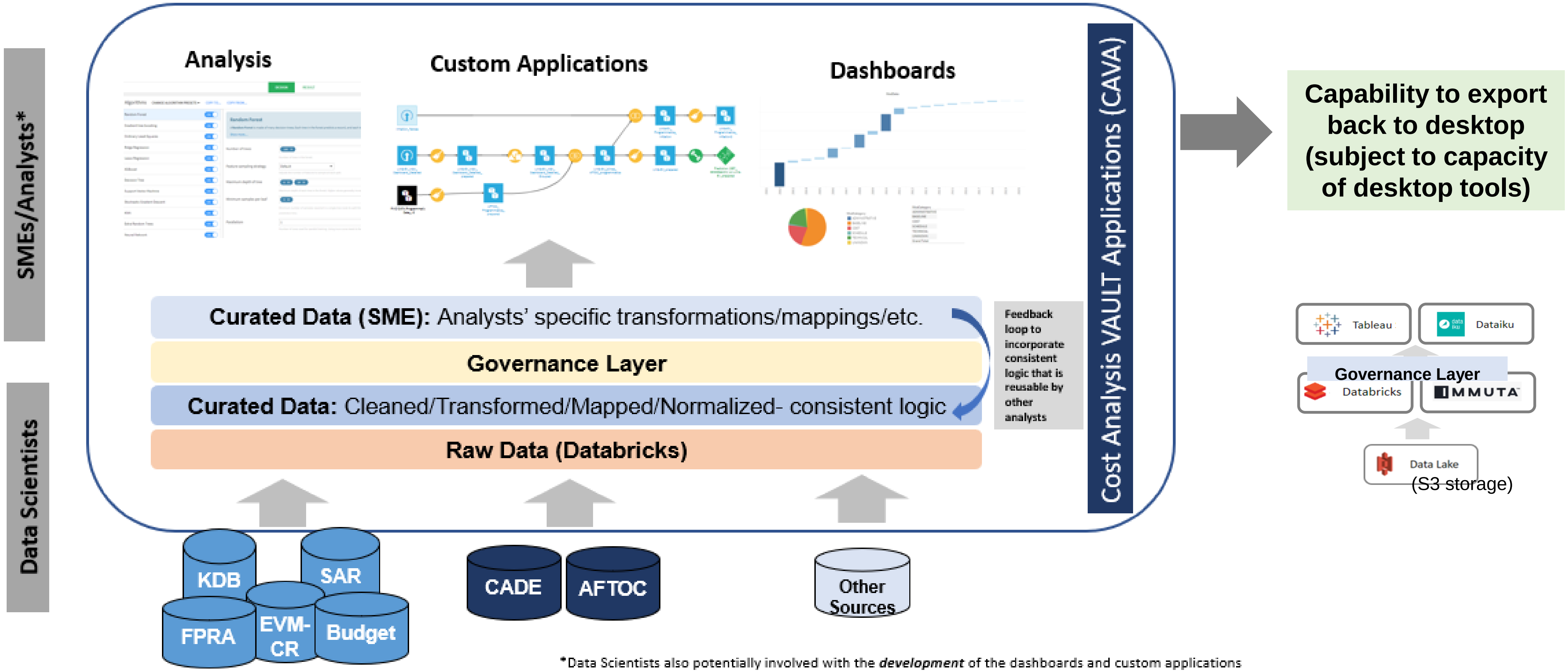
As of Jan 2023

SAF/FMC/AFCAA
Platform Requirements Analysis

Conclusion: Vault is Preferred DAF FM Cost Estimating and Analysis Community Platform

- **Pros of Vault over Advana for AFCAA**
  - Proprietary data protection – government analysts can perform application/dashboard development and data governance - a MUST for industry data
  - Emphasis on analytic tool accessibility and allowance for “Bring Your Own Tools” – a significant requirement/benefit to cost community
  - AFCAA solutions developed in Vault - cannot be replicated in Advana without capability degradation
  - AF Operations Research community platform of choice- Enhancing AF cost community synergistic Analytics Sharing
- **Findings**
  - As-is Advana does not meet cost community data analytic needs
  - DAF FM Budget/Financial community has different platform needs
  - Current Advana development priorities not focused on our cost requirements- waiting for those would significantly derail current cost tool development path
- **AFCAA Vault use is funded; funding for use by AF/SF and DoD CADE cost community will be worked**
- **NIPR/SIPR VAULT are currently operational, JWICS on roadmap for consistency with toolsets across classification environments**

Cost Analysis Data Operating Model (CA DOM) Architecture Overview



SMEs/Analysts*

Data Scientists

Analysis

Custom Applications

Dashboards

Curated Data (SME): Analysts' specific transformations/mappings/etc.

Governance Layer

Curated Data: Cleaned/Transformed/Mapped/Normalized- consistent logic

Raw Data (Databricks)

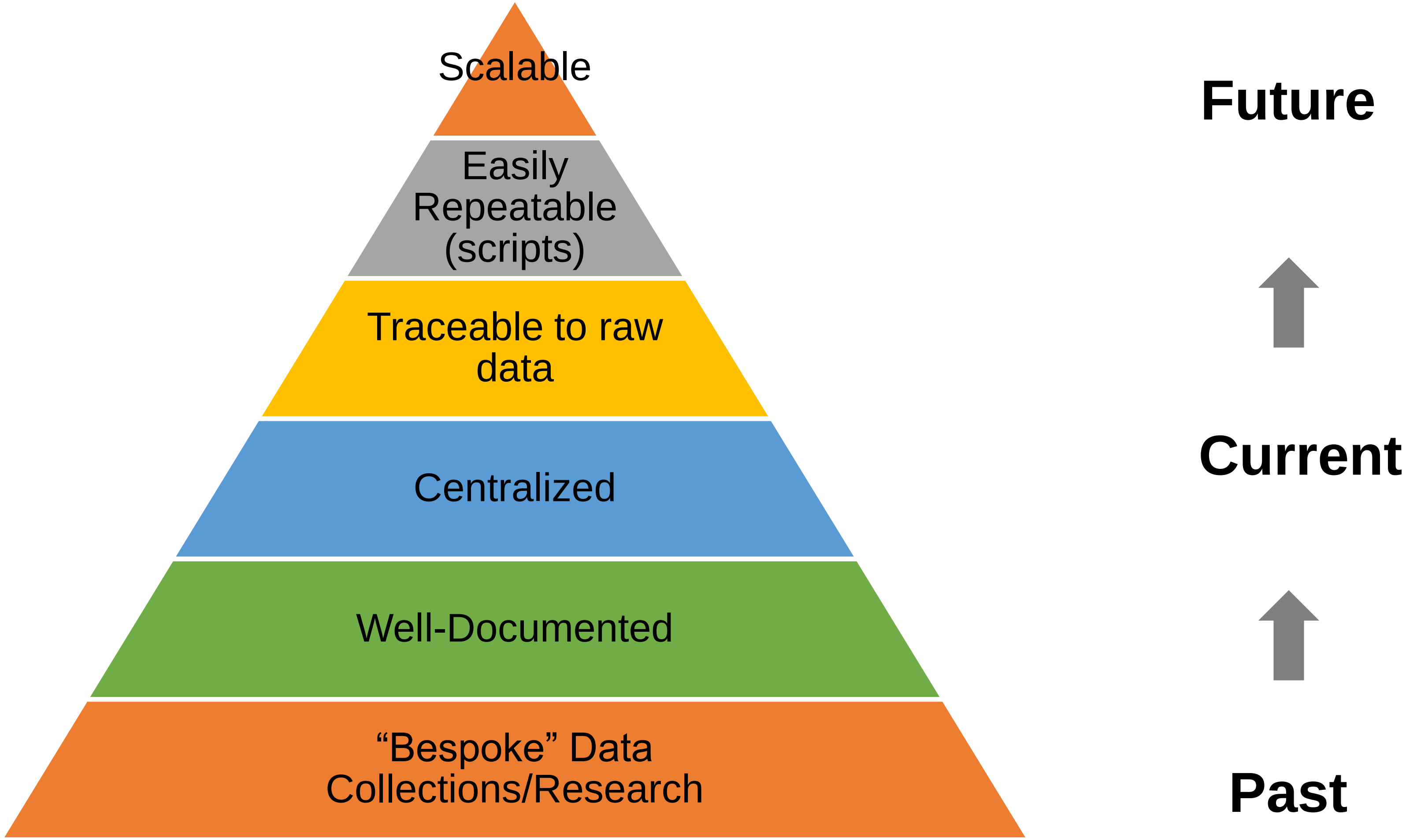
Feedback loop to incorporate consistent logic that is reusable by other analysts

Cost Analysis VAULT Applications (CAVA)

Capability to export back to desktop (subject to capacity of desktop tools)

*Data Scientists also potentially involved with the *development* of the dashboards and custom applications

Why VAULT?



Scalable

Easily Repeatable (scripts)

Traceable to raw data

Centralized

Well-Documented

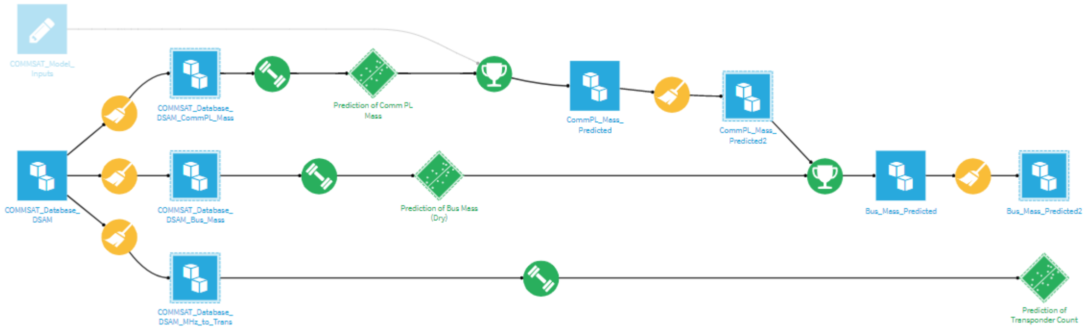
“Bespoke” Data Collections/Research

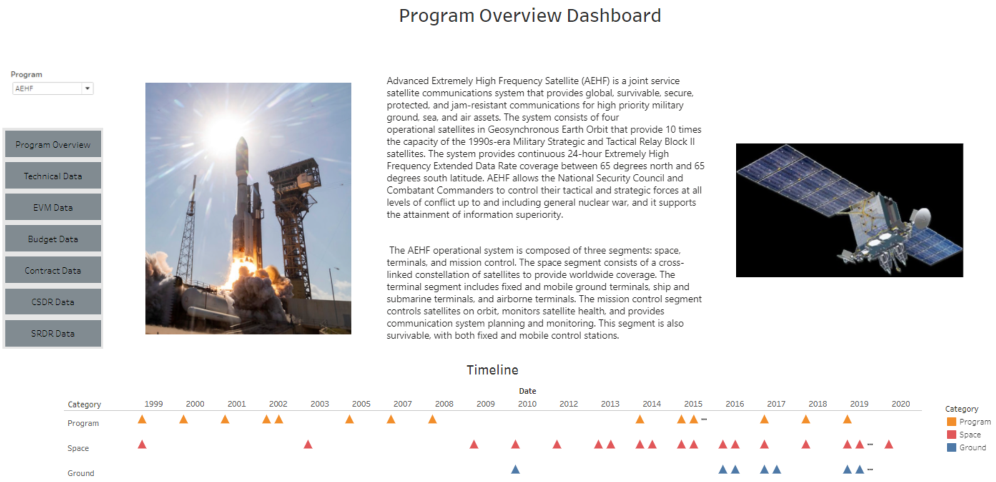
Future

Current

Past

Example VAULT Products


COMMSAT_Model_Inputs
COMMSAT_Database_DSAM_CommPL_Mass
Prediction of Comm PL Mass
CommPL_Mass_Predicted
CommPL_Mass_Predicted2
COMMSAT_Database_DSAM
COMMSAT_Database_DSAM_Bus_Mass
Prediction of Bus Mass (Dry)
Bus_Mass_Predicted
Bus_Mass_Predicted2
COMMSAT_Database_DSAM_MHz_to_Trans
Prediction of Transponder Count

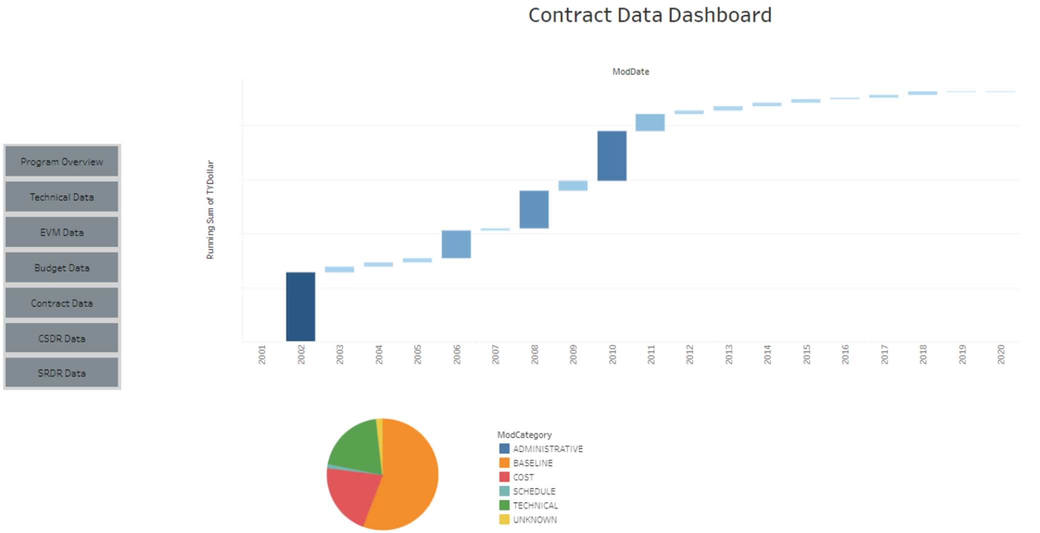

Program Overview Dashboard

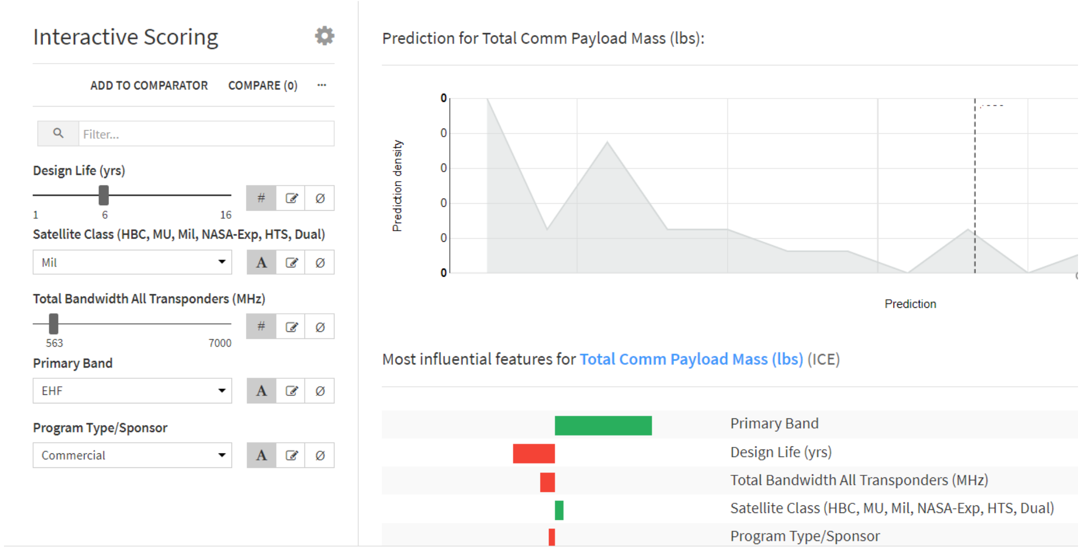
Program: AEHF

Advanced Extremely High Frequency Satellite (AEHF) is a joint service satellite communications system that provides global, survivable, secure, protected, and jam-resistant communications for high priority military ground, sea, and air assets. The system consists of four operational satellites in Geosynchronous Earth Orbit that provide 10 times the capacity of the 1990s-era Military Strategic and Tactical Relay Block II satellites. The system provides continuous 24-hour Extremely High Frequency Extended Data Rate coverage between 65 degrees north and 65 degrees south latitude. AEHF allows the National Security Council and Combatant Commanders to control their tactical and strategic forces at all levels of conflict up to and including general nuclear war, and it supports the attainment of information superiority.

The AEHF operational system is composed of three segments: space, terminals, and mission control. The space segment consists of a cross-linked constellation of satellites to provide worldwide coverage. The terminal segment includes fixed and mobile ground terminals, ship and submarine terminals, and airborne terminals. The mission control segment controls satellites on orbit, monitors satellite health, and provides communication system planning and monitoring. This segment is also survivable, with both fixed and mobile control stations.

Timeline


Contract Data Dashboard


Interactive Scoring

ADD TO COMPARATOR   COMPARE (0)   ···

Filter...

Design Life (yrs): 6 (range 1–16)

Satellite Class (HBC, MU, Mil, NASA-Exp, HTS, Dual): Mil

Total Bandwidth All Transponders (MHz): 563 (range 563–7000)

Primary Band: EHF

Program Type/Sponsor: Commercial

Prediction for Total Comm Payload Mass (lbs):

0
0
0
0
0
0
Prediction density
Prediction

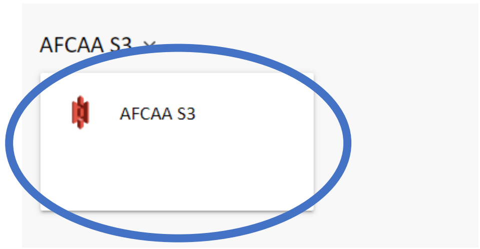
Most influential features for Total Comm Payload Mass (lbs) (ICE)

Primary Band
Design Life (yrs)
Total Bandwidth All Transponders (MHz)
Satellite Class (HBC, MU, Mil, NASA-Exp, HTS, Dual)
Program Type/Sponsor

Integrity - Service - Excellence

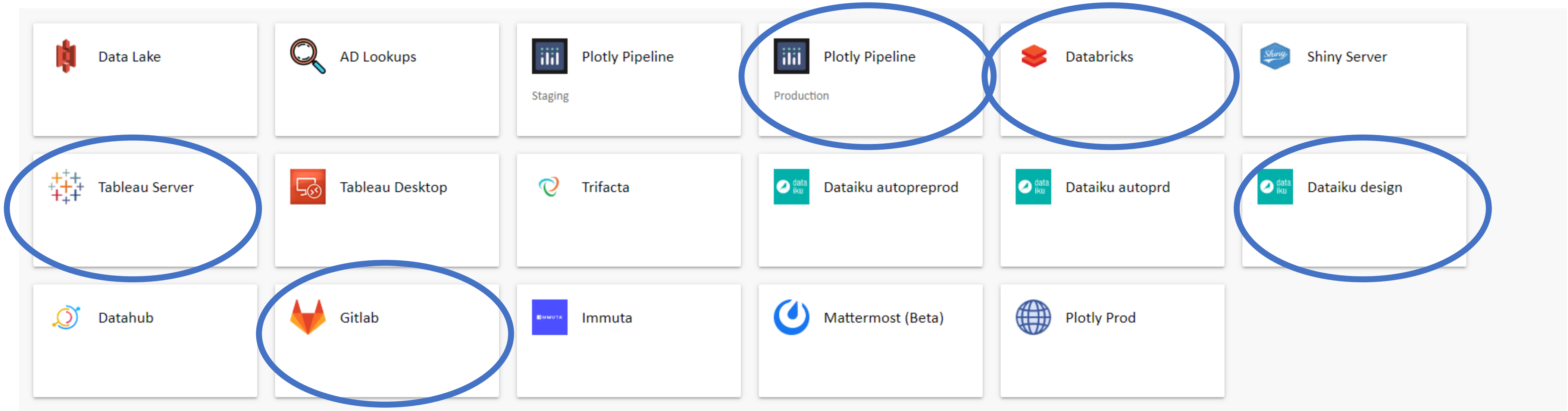
UNCLASSIFIED

VAULT Homepage



AFCAA S3

AFCAA S3



Data Lake

AD Lookups

Plotly Pipeline
Staging

Plotly Pipeline
Production

Databricks

Shiny Server

Tableau Server

Tableau Desktop

Trifacta

Dataiku autopreprod

Dataiku autoprd

Dataiku design

Datahub

Gitlab

Immuta

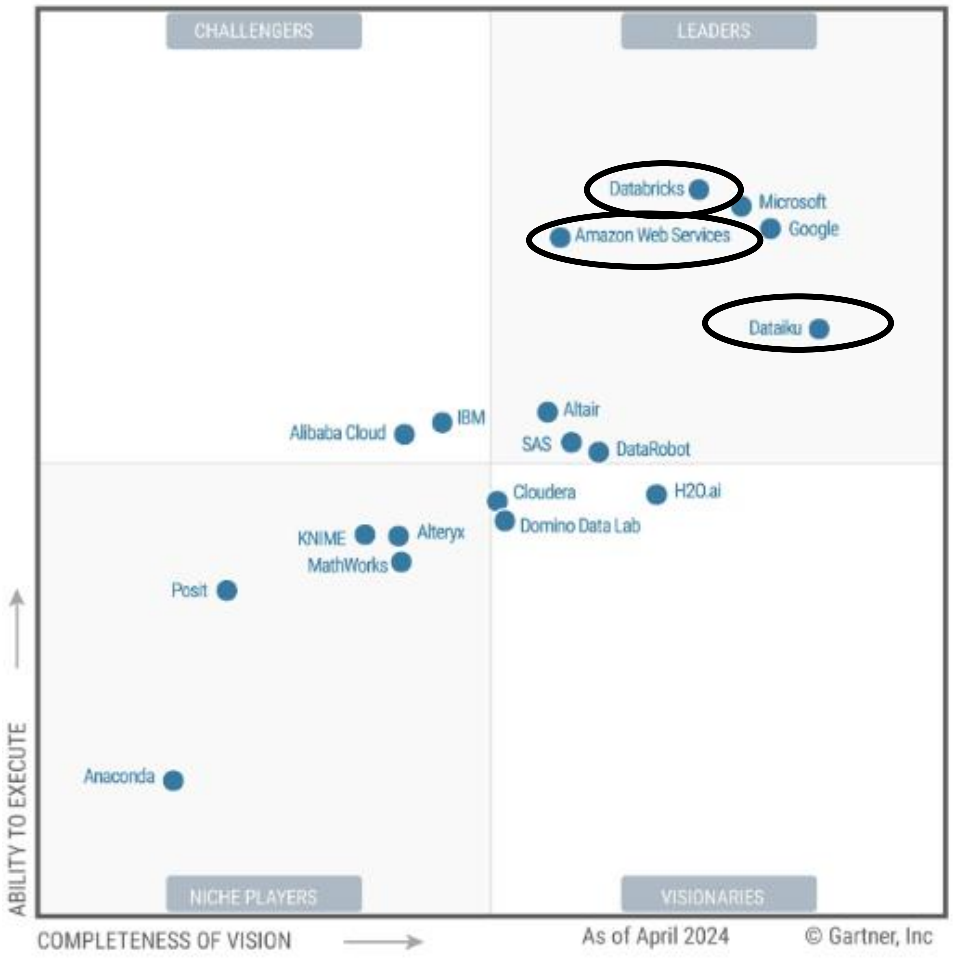
Mattermost (Beta)

Plotly Prod

Industry Assessment of Data Science Platforms

- **Gartner sets the industry standard for assessment of various tools**
- **Our entire tech stack is in the upper quadrant for their 2024 assessments**
  - AWS
  - Databricks
  - Dataiku
  - Tableau (on the Analytics and Business intelligence quadrant)
- **Annually reassessing**
- **Data Pipelines not tool dependent- built with SQL, Python, R which are supported in all DoD cloud based platforms**

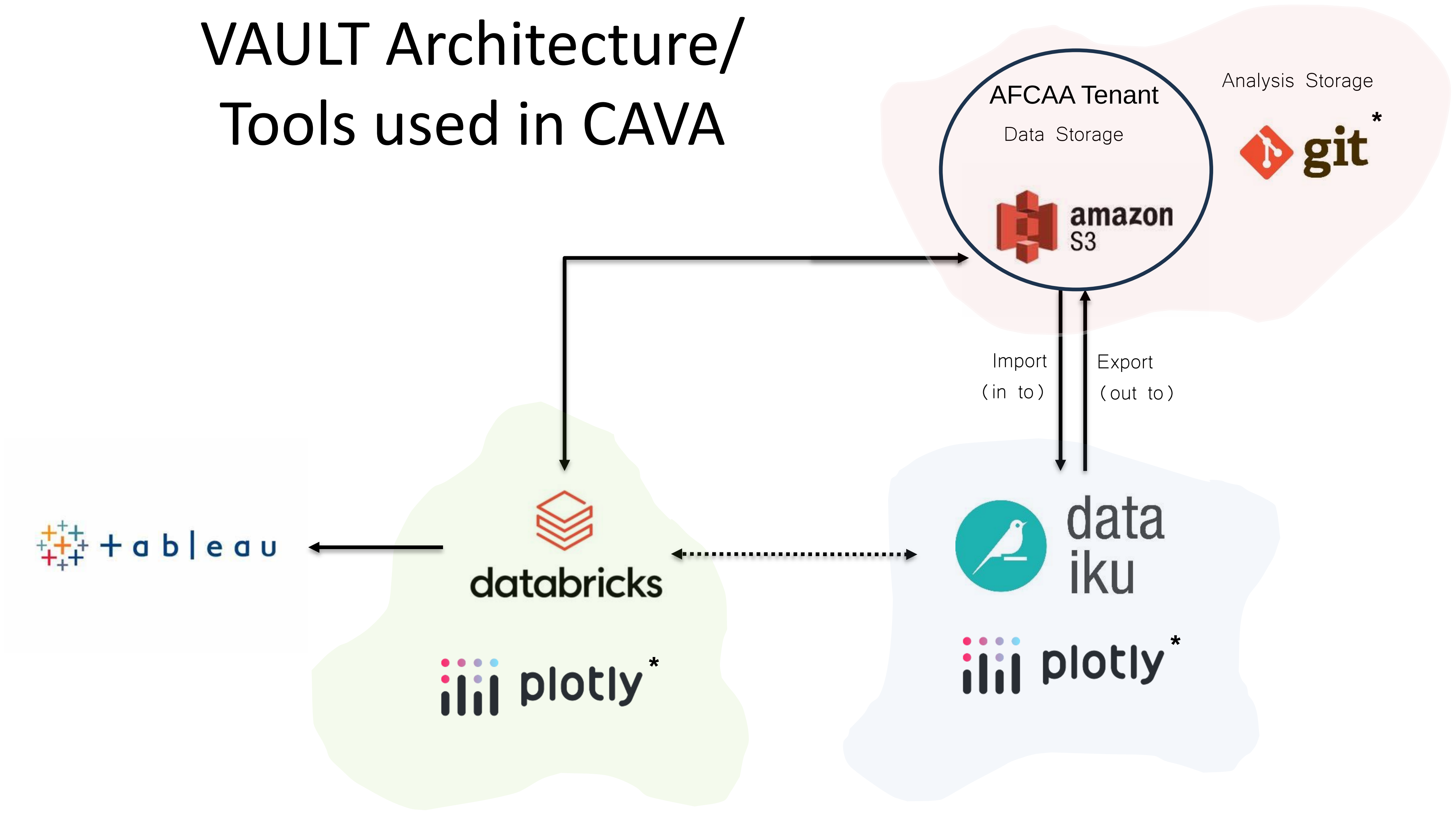
Figure 1: Magic Quadrant for Data Science and Machine Learning Platforms



Source: Gartner (June 2024)

Gartner.

VAULT Architecture/ Tools used in CAVA



Data Location

Provide me with…

- …all Air to Ground (AGM) missile programs
- …all F-35 production reports
- …all Sidewinder production reports, and subset the data to the Propulsion WBS element
- … all Space System ground components by a certain contractor

Data Modeling

Plot for me…

- **…total cost for all F-35 production reports, in Lot order, normalized for work scope by reporting entity and adjusted for escalation**

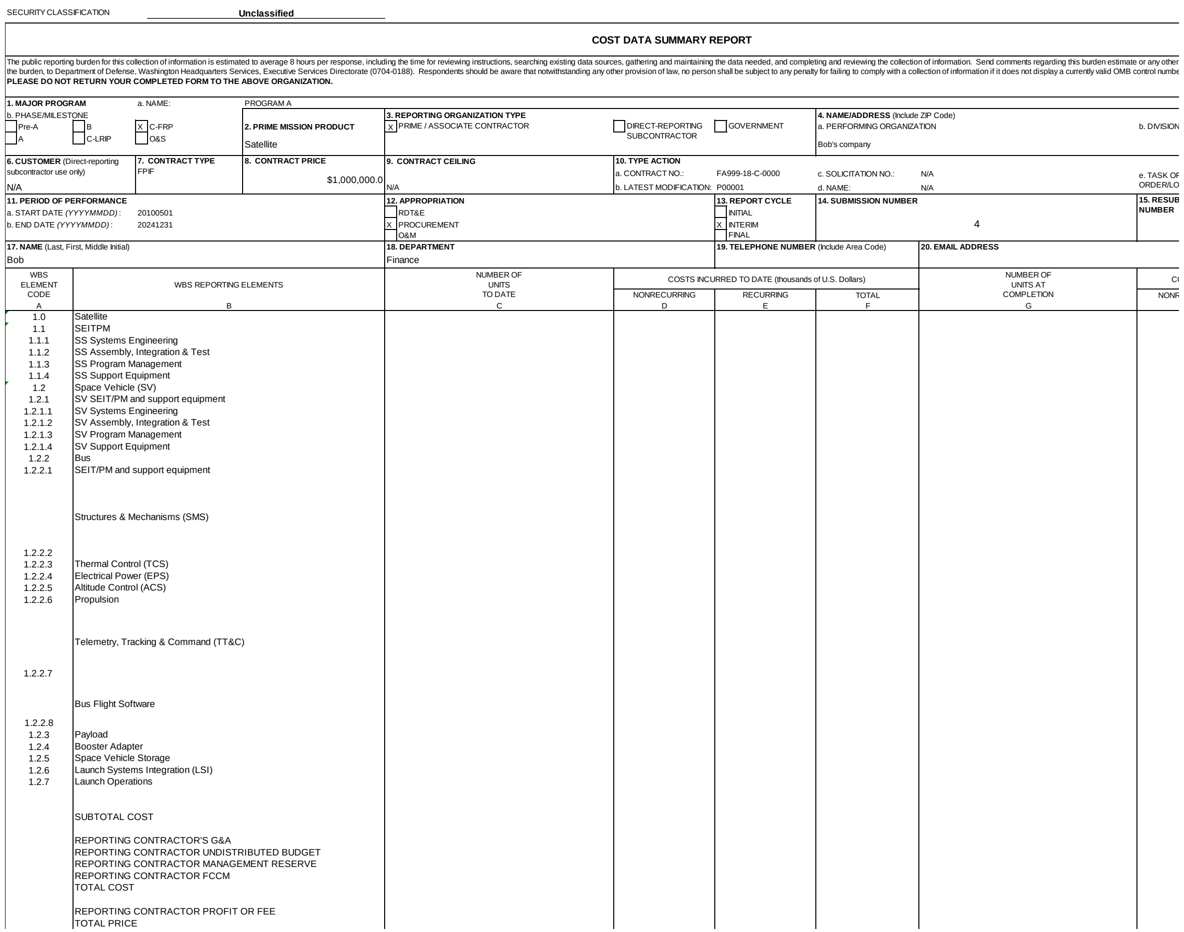
Provide me data that I can use to…

- **…estimates material cost for all missiles programs, stratified by missile type, across time**

Data Modeling labeling provides a much deeper set of tags that would conducive modeling with minimal further manual data preparation or transformation work

Historical Process

1921= REPORT



SECURITY CLASSIFICATION **Unclassified**

COST DATA SUMMARY REPORT

The public reporting burden for this collection of information is estimated to average 8 hours per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other the burden, to Department of Defense, Washington Headquarters Services, Executive Services Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control numbe
PLEASE DO NOT RETURN YOUR COMPLETED FORM TO THE ABOVE ORGANIZATION.

1. MAJOR PROGRAM a. NAME: PROGRAM A
b. PHASE/MILESTONE
□ Pre-A □ B ☒ C-FRP
□ A □ C-LRIP □ O&S

2. PRIME MISSION PRODUCT
Satellite

3. REPORTING ORGANIZATION TYPE
☒ PRIME / ASSOCIATE CONTRACTOR □ DIRECT-REPORTING SUBCONTRACTOR □ GOVERNMENT

4. NAME/ADDRESS (Include ZIP Code)
a. PERFORMING ORGANIZATION
Bob's company
b. DIVISION

6. CUSTOMER (Direct-reporting subcontractor use only)
N/A

7. CONTRACT TYPE
FPIF

8. CONTRACT PRICE
$1,000,000.0

9. CONTRACT CEILING
N/A

10. TYPE ACTION
a. CONTRACT NO.: FA999-18-C-0000
b. LATEST MODIFICATION: P00001
c. SOLICITATION NO.: N/A
d. NAME: N/A
e. TASK OF ORDER/LO

11. PERIOD OF PERFORMANCE
a. START DATE *(YYYYMMDD)* : 20100501
b. END DATE *(YYYYMMDD)* : 20241231

12. APPROPRIATION
□ RDT&E
☒ PROCUREMENT
□ O&M

13. REPORT CYCLE
□ INITIAL
☒ INTERIM
□ FINAL

14. SUBMISSION NUMBER
4

15. RESUB NUMBER

17. NAME (Last, First, Middle Initial)
Bob

18. DEPARTMENT
Finance

19. TELEPHONE NUMBER (Include Area Code)

20. EMAIL ADDRESS

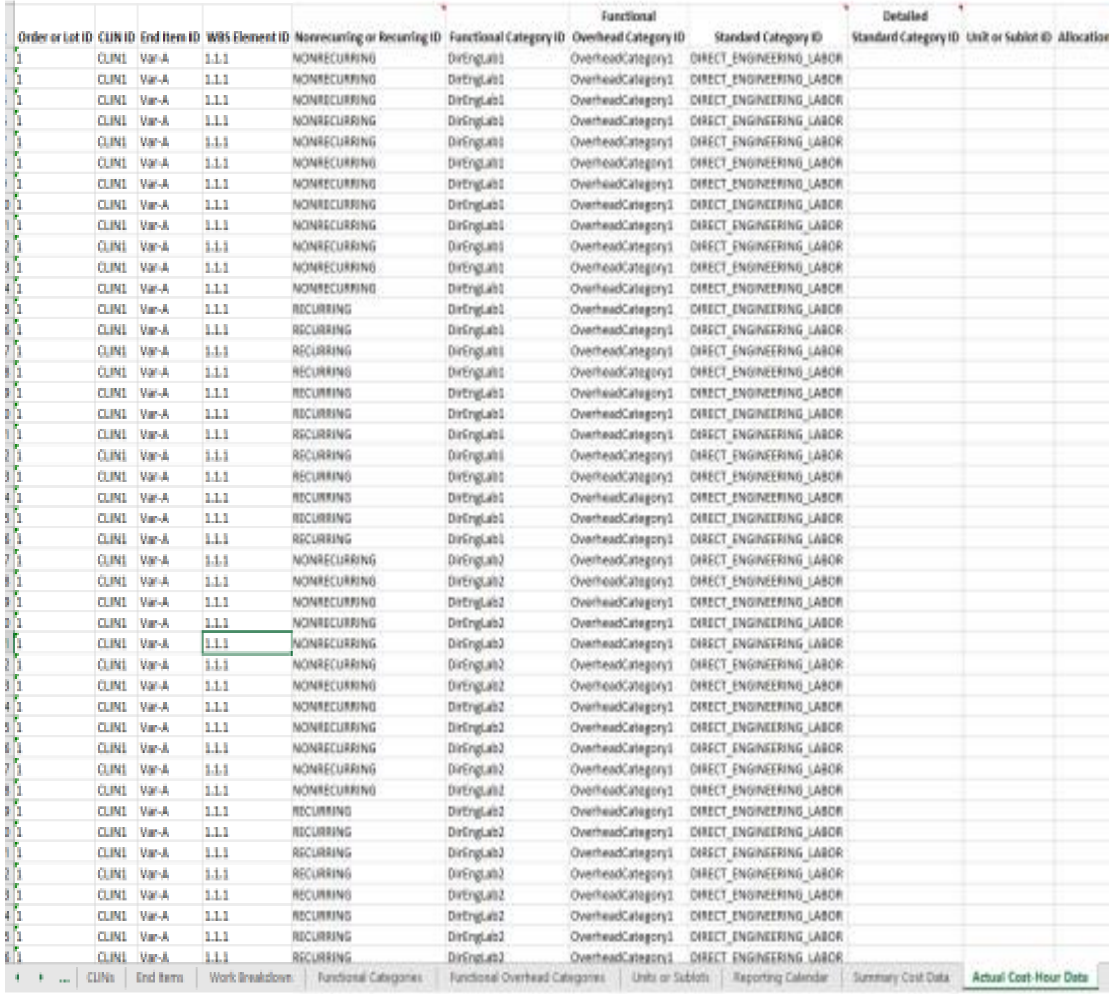
WBS ELEMENT CODE	WBS REPORTING ELEMENTS	NUMBER OF UNITS TO DATE	COSTS INCURRED TO DATE (thousands of U.S. Dollars)			NUMBER OF UNITS AT COMPLETION	CO
			NONRECURRING	RECURRING	TOTAL		NONR
A	B	C	D	E	F	G	
1.0	Satellite						
1.1	SEITPM						
1.1.1	SS Systems Engineering						
1.1.2	SS Assembly, Integration & Test						
1.1.3	SS Program Management						
1.1.4	SS Support Equipment						
1.2	Space Vehicle (SV)						
1.2.1	SV SEIT/PM and support equipment						
1.2.1.1	SV Systems Engineering						
1.2.1.2	SV Assembly, Integration & Test						
1.2.1.3	SV Program Management						
1.2.1.4	SV Support Equipment						
1.2.2	Bus						
1.2.2.1	SEIT/PM and support equipment						
	Structures & Mechanisms (SMS)						
1.2.2.2							
1.2.2.3	Thermal Control (TCS)						
1.2.2.4	Electrical Power (EPS)						
1.2.2.5	Altitude Control (ACS)						
1.2.2.6	Propulsion						
	Telemetry, Tracking & Command (TT&C)						
1.2.2.7							
	Bus Flight Software						
1.2.2.8							
1.2.3	Payload						
1.2.4	Booster Adapter						
1.2.5	Space Vehicle Storage						
1.2.6	Launch Systems Integration (LSI)						
1.2.7	Launch Operations						
	SUBTOTAL COST						
	REPORTING CONTRACTOR'S G&A						
	REPORTING CONTRACTOR UNDISTRIBUTED BUDGET						
	REPORTING CONTRACTOR MANAGEMENT RESERVE						
	REPORTING CONTRACTOR FCCM						
	TOTAL COST						
	REPORTING CONTRACTOR PROFIT OR FEE						
	TOTAL PRICE						

Manually updated Datasets

Compartmentalized Analysis

Charts/ Graphs (normally static/ PowerPoint)

Optimal Process



						Functional		Detailed		
Order or Lot ID	CLIN ID	End Item ID	WBS Element ID	Nonrecurring or Recurring ID	Functional Category ID	Overhead Category ID	Standard Category ID	Standard Category ID	Unit or Sublot ID	Allocation
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab1	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	NONRECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			
1	CLIN1	Var-A	1.1.1	RECURRING	DirEngLab2	OverheadCategory1	DIRECT_ENGINEERING_LABOR			

OUTPUTS

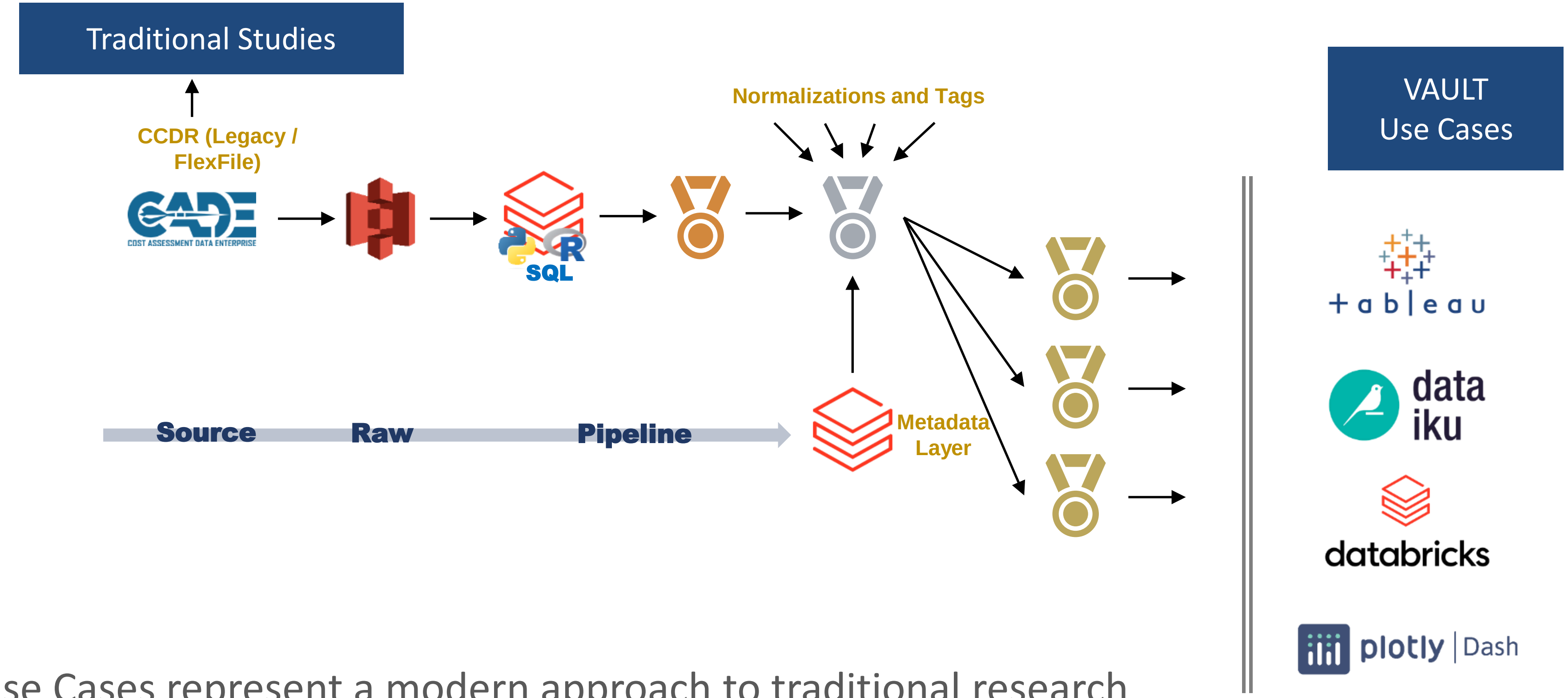
1921 – like reports

Dynamically updated Dashboards

Models connected to data

Applications

Analysis



- VAULT Use Cases represent a modern approach to traditional research
- Dynamically linked to the data pipeline to enable rapid and consistent analyses
  - Built upon a singular source of cleaned and normalized data (i.e., Silver/Gold layers)
  - Enables replicability across programs, commodities, or other data subsets
  - Curates data in a way that can be leveraged across multiple data products
- Tools can be anything (e.g., Tableau, Dataiku, or even back to Excel based on an export)
- Does not fully automate an analysis
  - Provides some efficiency, be it small or large depending on the use case
  - Still includes heavy analyst in the loop steps to ensure the correct data is being used

Selecting the Optimal Solution

If some of the following are true..

- Original, raw data source size is large
- Data is relational in nature (tables related to other tables by common fields)
- Data requires a set of steps from raw source to final form for analysis (transformation/mapping/cleaning etc)
- Original raw data set is updated frequently (more than ~3xs a year)
- Data is from a original source that has consistent fields structured in flat tables
- Resulting model tends to be based on parametrics
- Models tend to have consistent logic for elements as a starting point
- Models that over time require many iterations/alternatives or require frequent updates
- Steps of the analysis could be easily reused by another program/ commodity/ etc. (i.e. common factors)
- Model is generally based on objective data

…Then **VAULT/data science** might be the best solution/approach

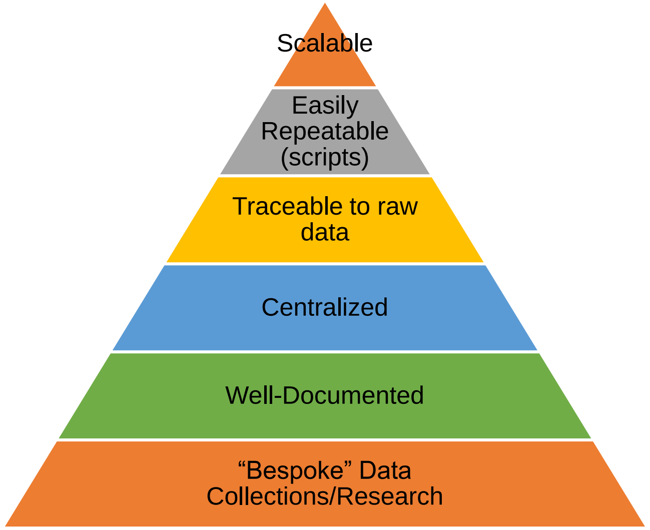
If most of the following are true..

- Original, raw data source size is small
- Dataset is stand-alone and not relational in nature
- Data is used as-is from the original source
- Original raw data set is updated infrequently (less than ~3xs a year)
- Data is from an original source that does not have consistent fields structured in flat tables
- Resulting model tends to be based on a very small dataset or analogy
- Models that do not have any consistent logic for elements even as a starting point for analysis
- Models that do not require many iterations and are stable without a need to update with more recent data
- Steps of the analysis are not reusable by another program/ commodity/ etc (i.e. very niche/specific)
- Model is based on highly subjective data (i.e., each point subject to interpretation by SME)

…Then **current desktop tools** might be the best solution/approach

Why Dataiku?

- Power of Data Science in a no-code/ low code environment
- Scalable (can run any number of excursions simultaneously)
- Easily repeatable
- Centralized/ Highly collaborative
- Dynamic documentation
- Customizable – can deploy custom “plug-ins” and applications for cost estimation


Scalable
Easily Repeatable (scripts)
Traceable to raw data
Centralized
Well-Documented
“Bespoke” Data Collections/Research

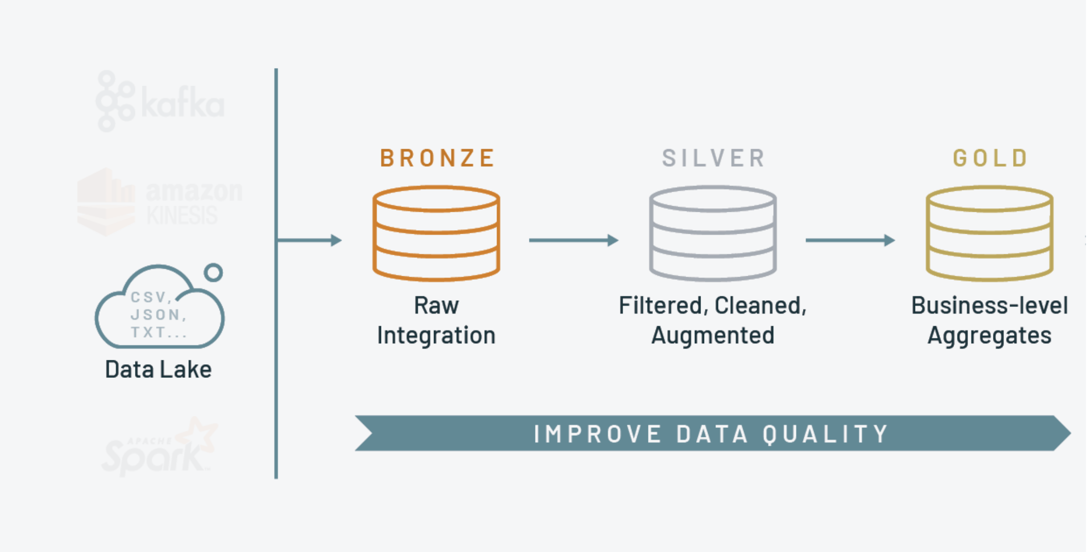
Excel	Dataiku
VLOOKUP	Join
SUMIF	Group By
Cut and Paste	Stack
Sort	Sort
VBA	Repeatable Scripts
Complex formulas	Window recipe
Complex formulas/ Manual / VBA	Prepare Recipe (100 processors)

Status of Data Pipelines

Data Layer	Data Included	Current Status	Notes
Metadata	Metadata	🟪	Metadata architecture 95% - Implementation needed for each layer at silver level
CSDR	FlexFiles	🟪	Missles 90% / Space 70%/ Aircraft 15%
	1921	🟪	
	1921-1	🟪	
	1921-2	🟪	
	1921-3	🟪	
	SRDR	🟧	Planned to start 2024 with PCIP/PAQ
	TDR	🟧	
	Maintenance & Repair	🟧	
SAR	SAR	🟪	
EVM	IPMDAR	🟨	currently only fields limited to export provided
	EVM Format 1	🟨	Lower level detail for AF only; working with EVM on other services
	EVM Format 2	🟧	
	EVM Format 3	🟧	
	EVM Format 4	🟧	
	EVM Format 5	🟧	
Contract	Contracts	🟦	Most mature data layer due to work on CADE contract- dashboards in production
Budget	Budget	🟧	
AFTOC	Flying Hours	🟪	as of FY24 Q1
	PAA	🟪	as of FY24 Q1
	TAI	🟪	as of FY24 Q1
	End Strength	🟪	as of FY23 Q4
	Total Cost Combined	🟪	as of FY23 Q4
	CAIG/CAPE	🟪	as of FY23 Q4
	Maintenance	🟧	not yet provided by AFTOC
	Supply	🟧	
	Personnel	🟧	
	Indirect	🟧	
FPRA	FPRA	🟧	
Gov Test	Government Test	🟧	

🟦	bronze/silver done - some gold
🟪	bronze/ significant progress on silver
🟨	bronze/ some progress on silver
🟧	haven't started yet

Bronze-Silver-Gold



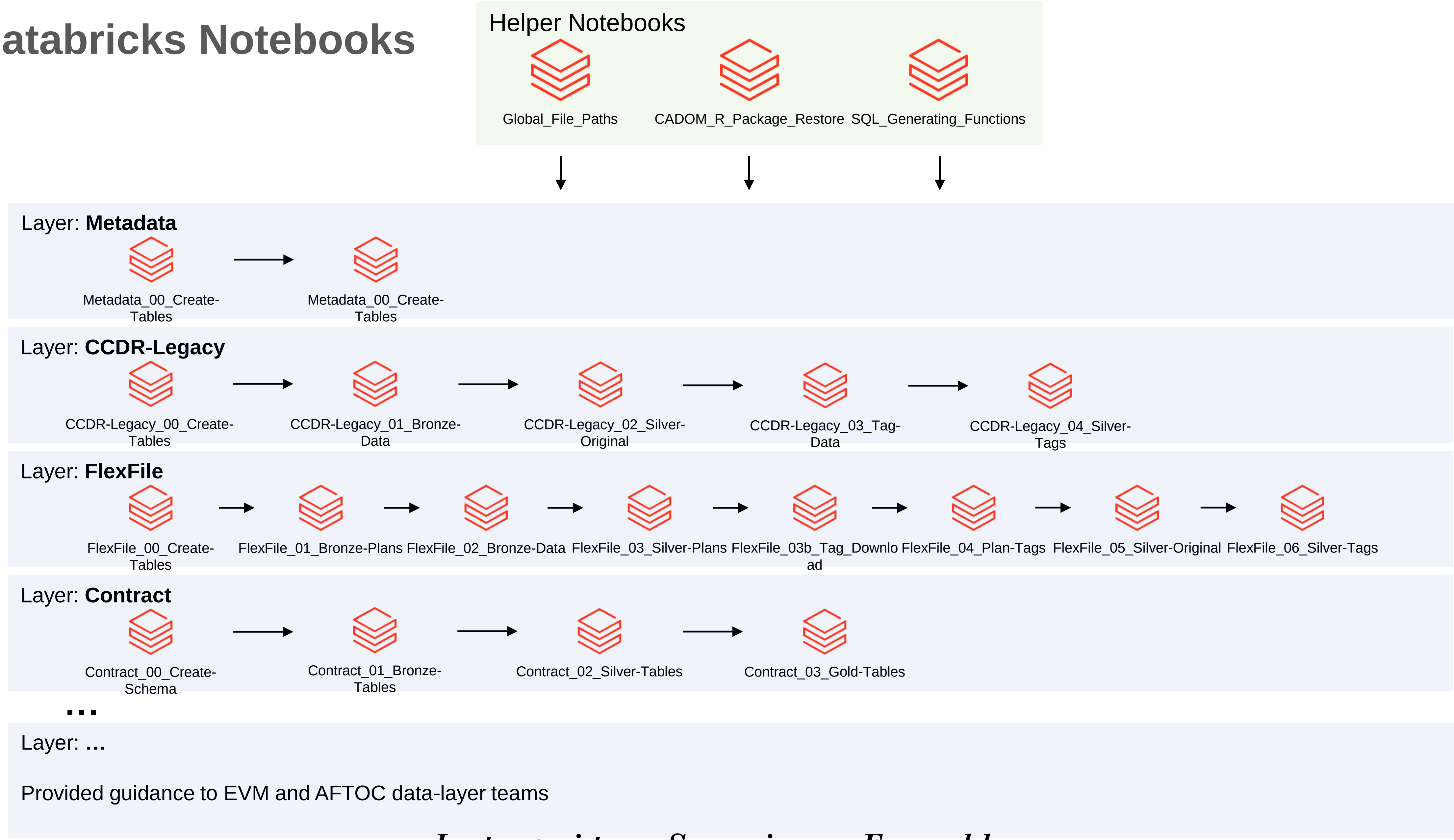
Gold: production-grade data that your entire company can rely on for business intelligence, descriptive statistics, and data science / machine learning

Silver: the raw data get cleansed (think data quality checks), transformed and potentially enriched with external data sets

Bronze: the initial landing zone for the pipeline. We recommend copying data that's as close to its raw form as possible to easily replay the whole pipeline from the beginning, if needed

https://www.databricks.com/blog/2020/06/02/monitor-your-databricks-workspace-with-audit-logs.html

Databricks Notebooks



Helper Notebooks

Global_File_Paths

CADOM_R_Package_Restore

SQL_Generating_Functions

Layer: **Metadata**

Metadata_00_Create-Tables → Metadata_00_Create-Tables

Layer: **CCDR-Legacy**

CCDR-Legacy_00_Create-Tables → CCDR-Legacy_01_Bronze-Data → CCDR-Legacy_02_Silver-Original → CCDR-Legacy_03_Tag-Data → CCDR-Legacy_04_Silver-Tags

Layer: **FlexFile**

FlexFile_00_Create-Tables → FlexFile_01_Bronze-Plans → FlexFile_02_Bronze-Data → FlexFile_03_Silver-Plans → FlexFile_03b_Tag_Download → FlexFile_04_Plan-Tags → FlexFile_05_Silver-Original → FlexFile_06_Silver-Tags

Layer: **Contract**

Contract_00_Create-Schema → Contract_01_Bronze-Tables → Contract_02_Silver-Tables → Contract_03_Gold-Tables

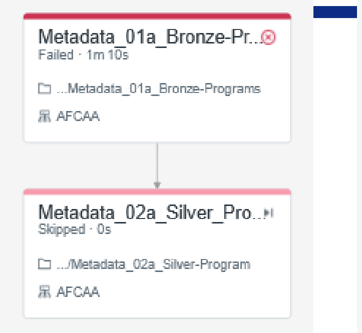
...

Layer: **...**

Provided guidance to EVM and AFTOC data-layer teams

VAULT Infrastructure

- **A collection of steps that trigger each other based on successful run**
- **Can be scheduled (i.e., run nightly)**
- **Set up for core CCDR & Metadata activities**
- **Provide an <u>auditable way to track progress, trap errors, and restart flows for data and infrastructure updates</u>**

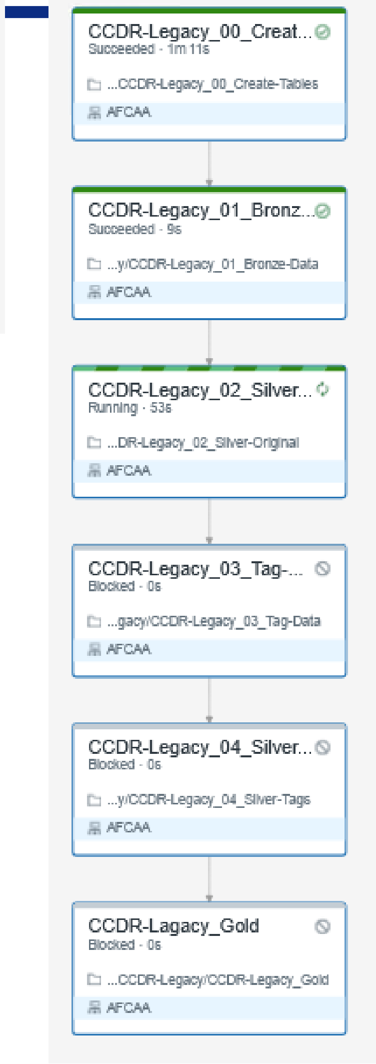


Failed

Metadata_01a_Bronze-Pr...
Failed · 1m 10s
...Metadata_01a_Bronze-Programs
AFCAA

Skipped

Metadata_02a_Silver_Pro...
Skipped · 0s
.../Metadata_02a_Silver-Program
AFCAA



Complete

CCDR-Legacy_00_Creat...
Succeeded · 1m 11s
...CCDR-Legacy_00_Create-Tables
AFCAA

Complete

CCDR-Legacy_01_Bronz...
Succeeded · 9s
...y/CCDR-Legacy_01_Bronze-Data
AFCAA

In progress

CCDR-Legacy_02_Silver...
Running · 53s
...DR-Legacy_02_Silver-Original
AFCAA

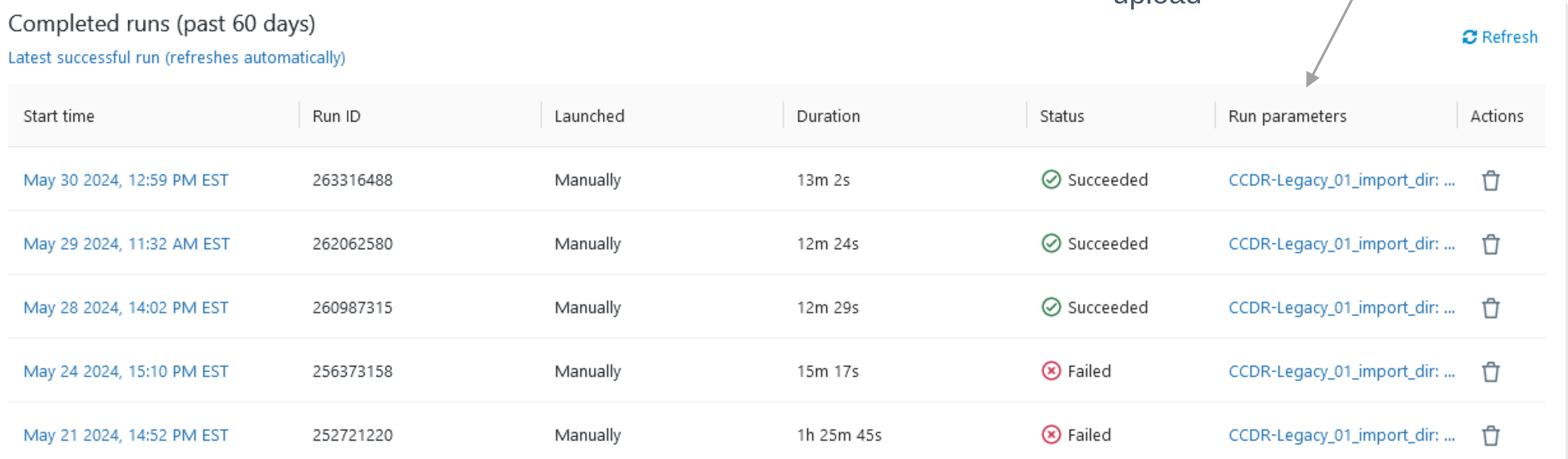
Pending

CCDR-Legacy_03_Tag-...
Blocked · 0s
...gacy/CCDR-Legacy_03_Tag-Data
AFCAA

CCDR-Legacy_04_Silver...
Blocked · 0s
...y/CCDR-Legacy_04_Silver-Tags
AFCAA

CCDR-Lagacy_Gold
Blocked · 0s
...CCDR-Legacy/CCDR-Legacy_Gold
AFCAA

Traceable to S3 location for data upload



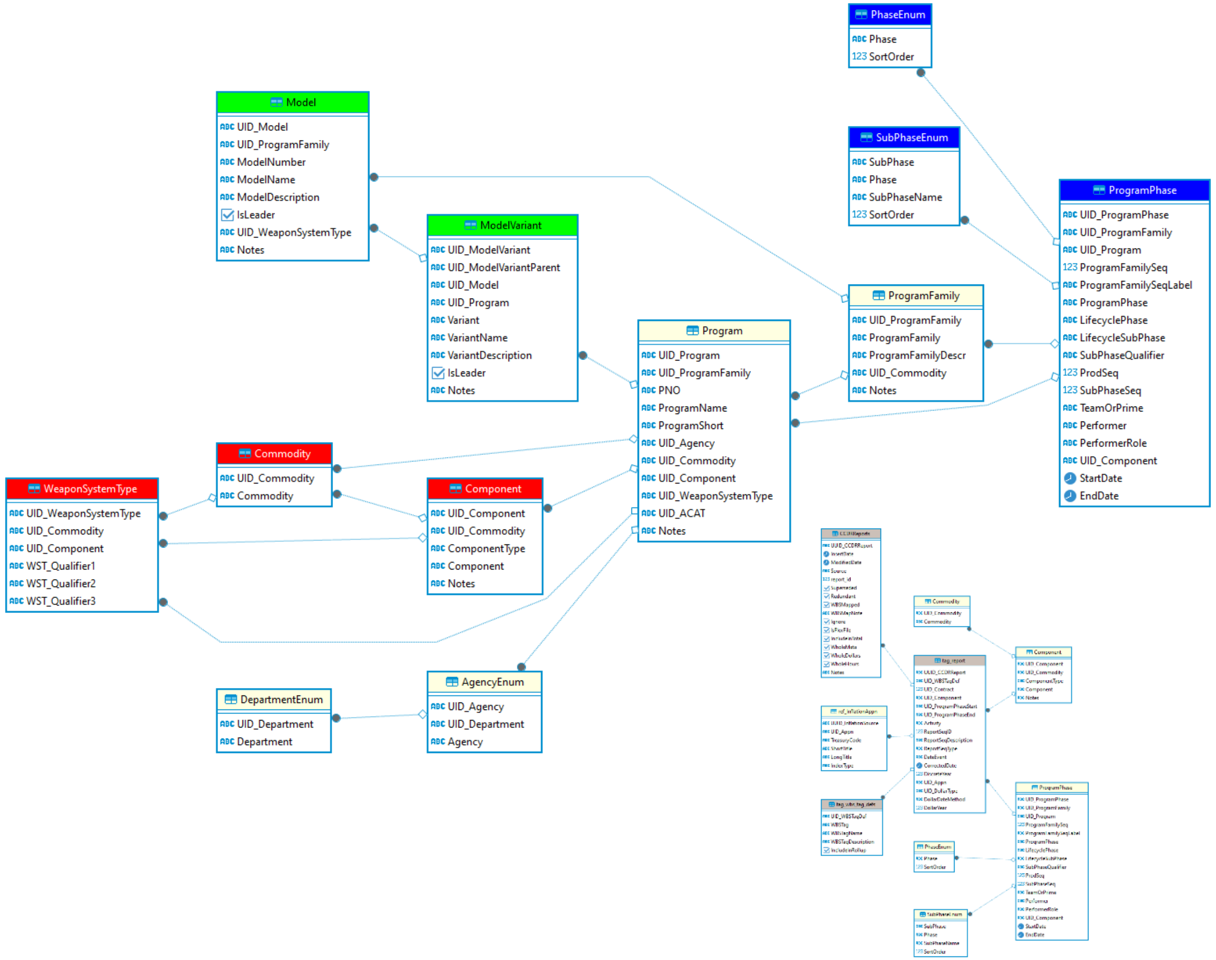
Completed runs (past 60 days)

Latest successful run (refreshes automatically)

Start time	Run ID	Launched	Duration	Status	Run parameters	Actions
May 30 2024, 12:59 PM EST	263316488	Manually	13m 2s	✓ Succeeded	CCDR-Legacy_01_import_dir: ...	🗑
May 29 2024, 11:32 AM EST	262062580	Manually	12m 24s	✓ Succeeded	CCDR-Legacy_01_import_dir: ...	🗑
May 28 2024, 14:02 PM EST	260987315	Manually	12m 29s	✓ Succeeded	CCDR-Legacy_01_import_dir: ...	🗑
May 24 2024, 15:10 PM EST	256373158	Manually	15m 17s	☒ Failed	CCDR-Legacy_01_import_dir: ...	🗑
May 21 2024, 14:52 PM EST	252721220	Manually	1h 25m 45s	☒ Failed	CCDR-Legacy_01_import_dir: ...	🗑

Documentation

- Further expanded documentation and definitions
  - CCDR Metadata field definitions and examples
  - Metadata-layer definitions
  - Commodity-specific guides
- Developed relationship diagrams
  - Layer-specific (i.e., Metadata)
  - Cross-layer (i.e., CCDR interaction with Metadata)



PhaseEnum
Phase
SortOrder

SubPhaseEnum
SubPhase
Phase
SubPhaseName
SortOrder

ProgramPhase
UID_ProgramPhase
UID_ProgramFamily
UID_Program
ProgramFamilySeq
ProgramFamilySeqLabel
ProgramPhase
LifecyclePhase
LifecycleSubPhase
SubPhaseQualifier
ProdSeq
SubPhaseSeq
TeamOrPrime
Performer
PerformerRole
UID_Component
StartDate
EndDate

Model
UID_Model
UID_ProgramFamily
ModelNumber
ModelName
ModelDescription
☑ IsLeader
UID_WeaponSystemType
Notes

ModelVariant
UID_ModelVariant
UID_ModelVariantParent
UID_Model
UID_Program
Variant
VariantName
VariantDescription
☑ IsLeader
Notes

ProgramFamily
UID_ProgramFamily
ProgramFamily
ProgramFamilyDescr
UID_Commodity
Notes

Program
UID_Program
UID_ProgramFamily
PNO
ProgramName
ProgramShort
UID_Agency
UID_Commodity
UID_Component
UID_WeaponSystemType
UID_ACAT
Notes

WeaponSystemType
UID_WeaponSystemType
UID_Commodity
UID_Component
WST_Qualifier1
WST_Qualifier2
WST_Qualifier3

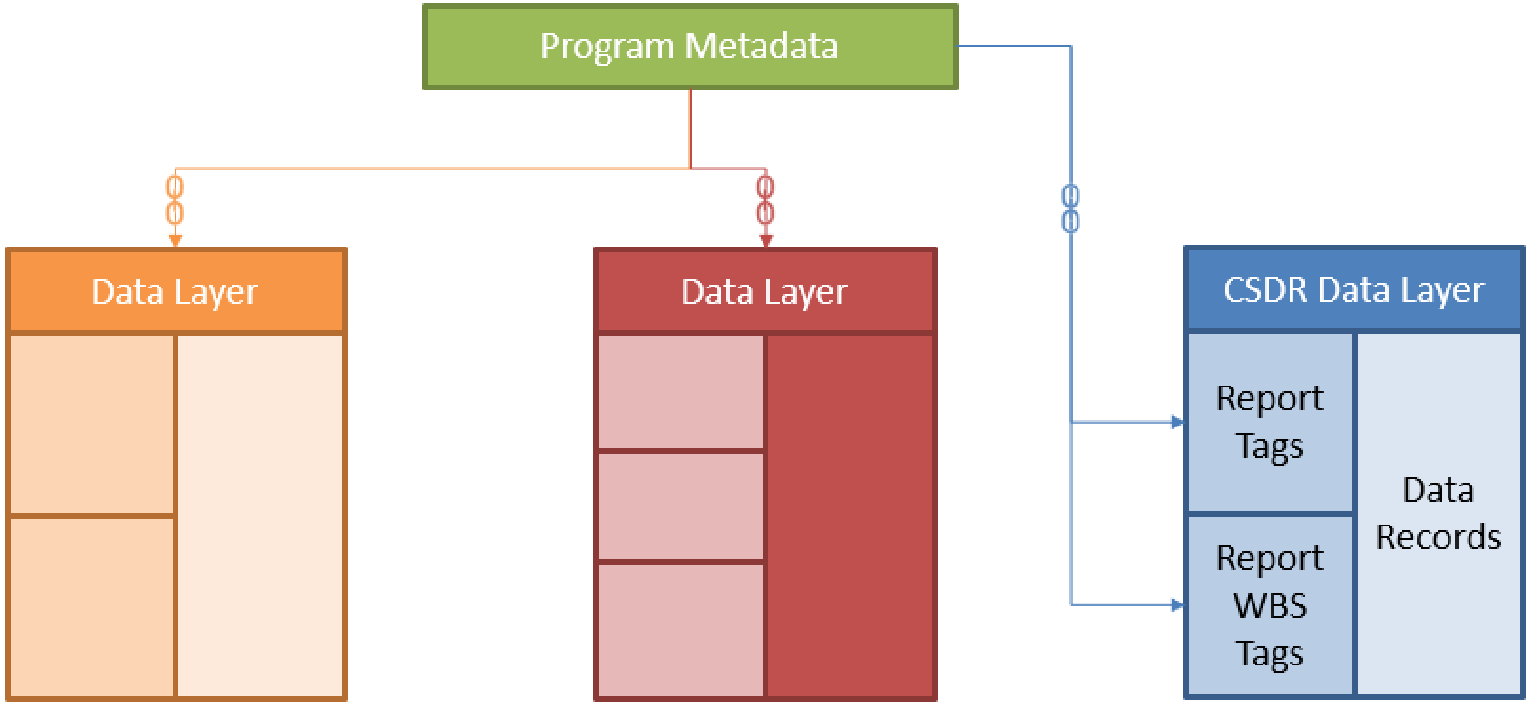
Commodity
UID_Commodity
Commodity

Component
UID_Component
UID_Commodity
ComponentType
Component
Notes

DepartmentEnum
UID_Department
Department

AgencyEnum
UID_Agency
UID_Department
Agency

Metadata



- Program-level metadata layer provides linkages between each data layer
- Each data layer has a consistent set of metadata that allows them to talk to each other as well as it’s unique normalizations that allow it to be used to it’s full potential

Use Cases Highlight

- **NLP Applications**
  - WBS Mapping Tool- utilizing ML with analyst in the loop (ML has 85-90% accuracy)
- **Other Applications/Use Cases/ Plug-ins: Dataiku**
  - Escalation/ Inflation Plug-in – ability to centrally load inflation/escalation and apply to individual datasets – MVP
  - Learning Curve Plug in – ability to perform Cumulative Avg or Unit LC/Rate analysis – MVP currently and iterating to improve
  - Monte Carlo Plug in – mature capability to run monte carlo simulations within Dataiku
  - Conversion of O&S LACE model in VAULT
    - Lots of lessons learned- and this is currently informing our silver/gold table build
  - Aircraft SEPM Study (in development utilizing CSDR data layer)

Dataiku Overview

Dataset Types in Dataiku

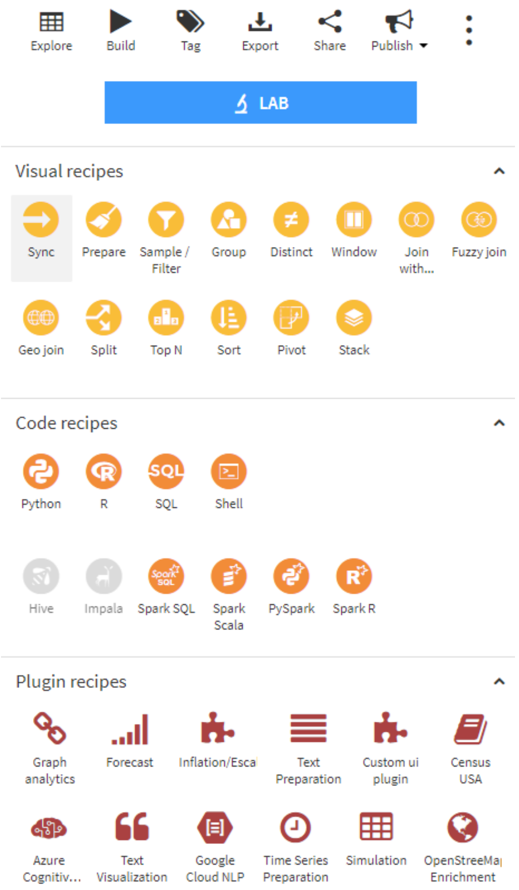
Uploaded Directly in Dataiku project

JDBC connection (for us-Databricks “live” tables)

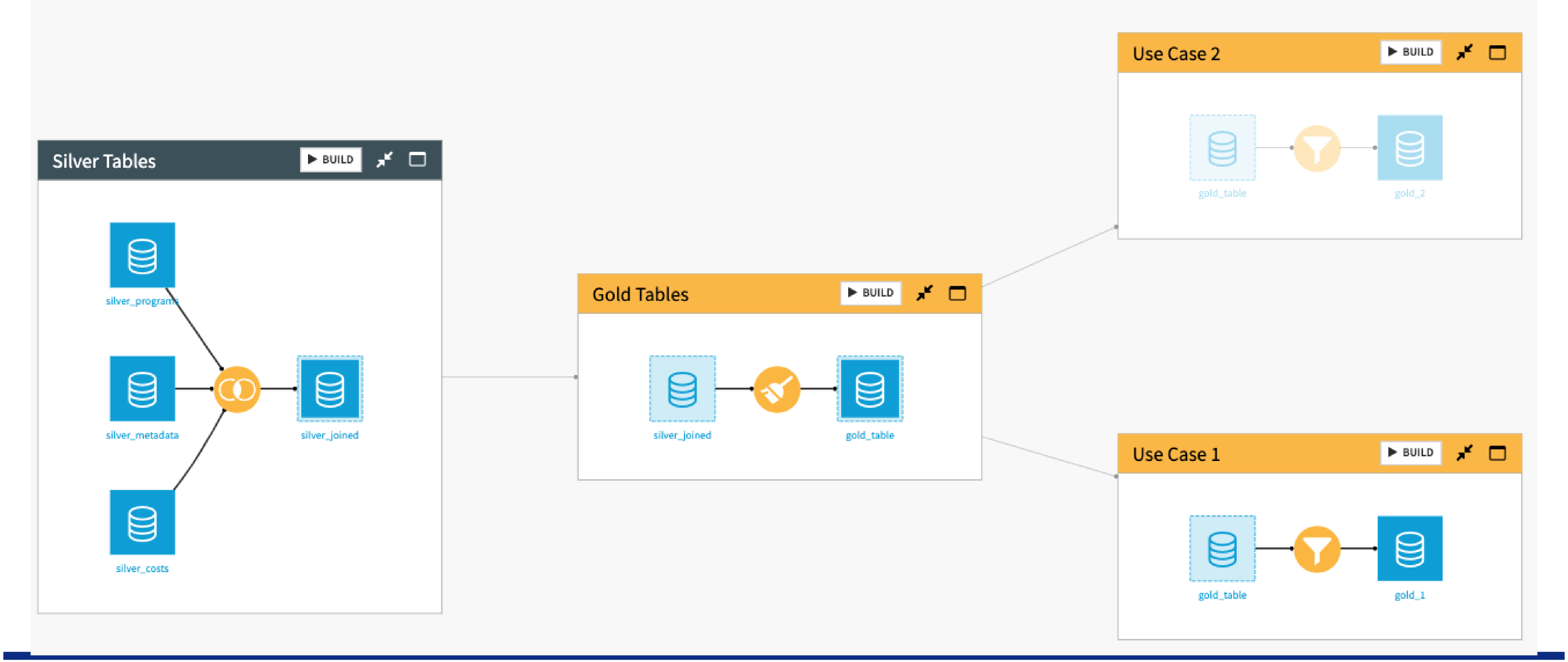
File coming from S3 bucket (these are static files)

Editable dataset- can interact with this similar to Excel

Dataiku Navigation Pane



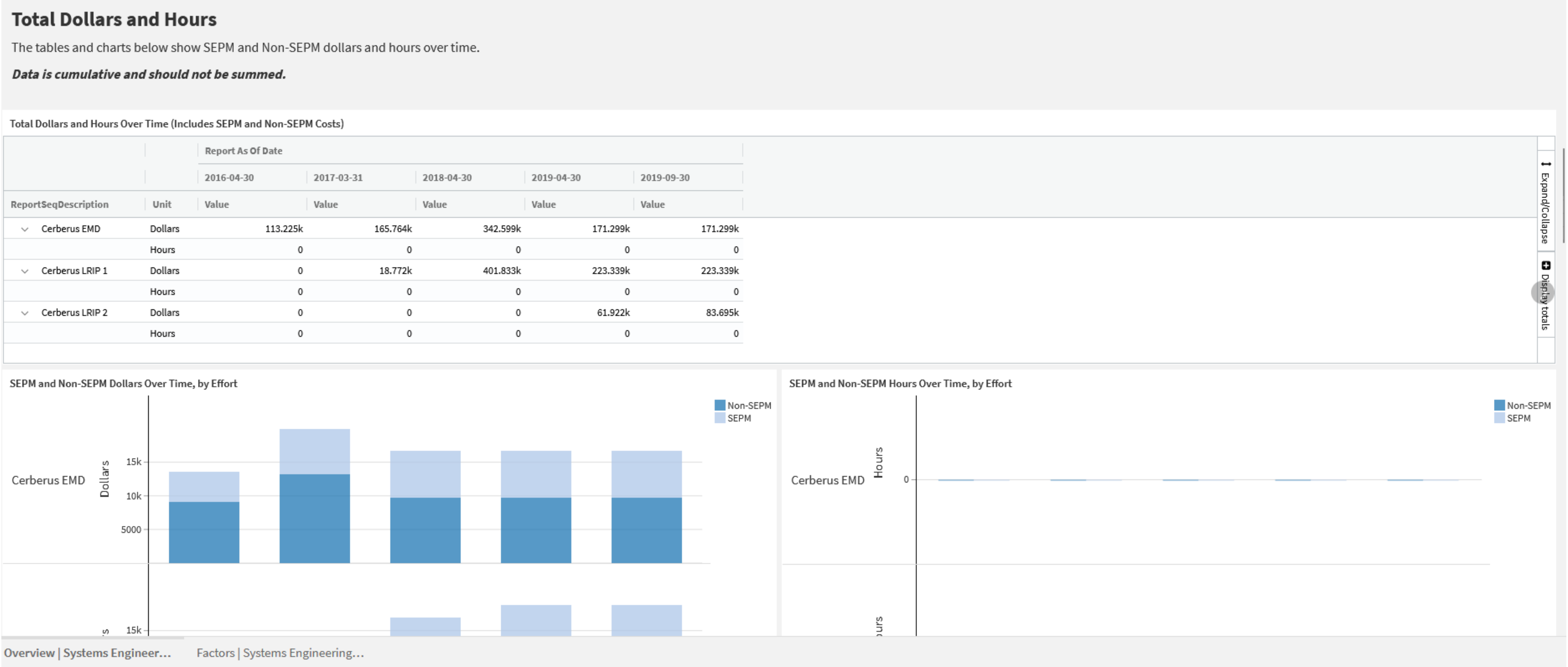
Data Pipeline to Use Cases



Integrity - Service - Excellence

UNCLASSIFIED

Aircraft SEPM Factors



Total Dollars and Hours

The tables and charts below show SEPM and Non-SEPM dollars and hours over time.

Data is cumulative and should not be summed.

Total Dollars and Hours Over Time (Includes SEPM and Non-SEPM Costs)

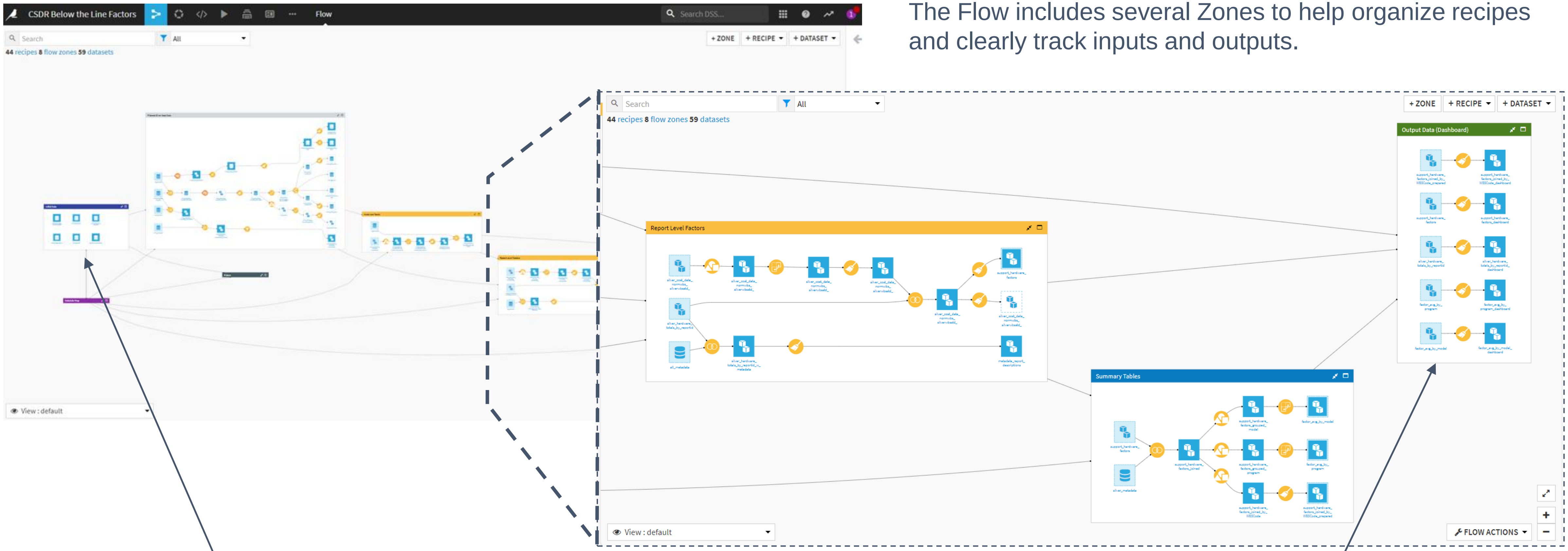
		Report As Of Date				
		2016-04-30	2017-03-31	2018-04-30	2019-04-30	2019-09-30
ReportSeqDescription	Unit	Value	Value	Value	Value	Value
Cerberus EMD	Dollars	113.225k	165.764k	342.599k	171.299k	171.299k
	Hours	0	0	0	0	0
Cerberus LRIP 1	Dollars	0	18.772k	401.833k	223.339k	223.339k
	Hours	0	0	0	0	0
Cerberus LRIP 2	Dollars	0	0	0	61.922k	83.695k
	Hours	0	0	0	0	0

SEPM and Non-SEPM Dollars Over Time, by Effort

SEPM and Non-SEPM Hours Over Time, by Effort

Below the Line Factors

The Flow includes several Zones to help organize recipes and clearly track inputs and outputs.



Inputs: Silver Tables from CCDR layer

Outputs: Custom Factor Tables for Visualization Dashboard

Backup

Database Structures

- It all starts with database structure
- Data should be structured and stored so that easy for the computer to use and manipulate
- Data are reported to be easy to consume by human
- There are certain best practices when working with databases and tables in the VAULT (or any modern toolset or programming language)
- Material borrowed extensively from "Don't just use your data… Exploit it" – Technomics ICEAA 2019

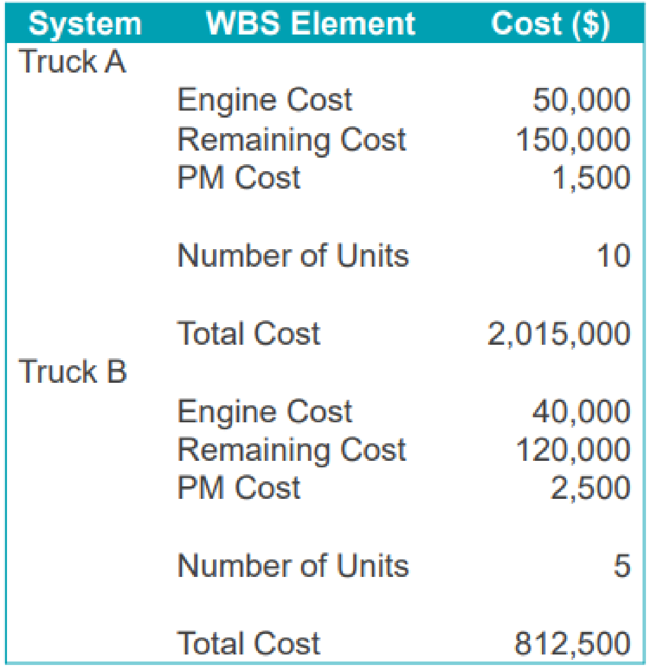
Structuring Tables in Data Collection

- When collecting data (even in Excel) – be intentional about building tables so that it facilitates data analysis with modern toolsets
- General rule of thumb- choose to organize so that you have 1 or more tables with a set of mostly unique fields
- Ensure each table has at 1 common field (unique ID)
- DO's and DON'Ts
  - **DON'T** separate out various subsets of the data into different tabs
  - **DO** consider adding separate tables such as "metadata" if the fields are getting too repetitive

Flat File Concept

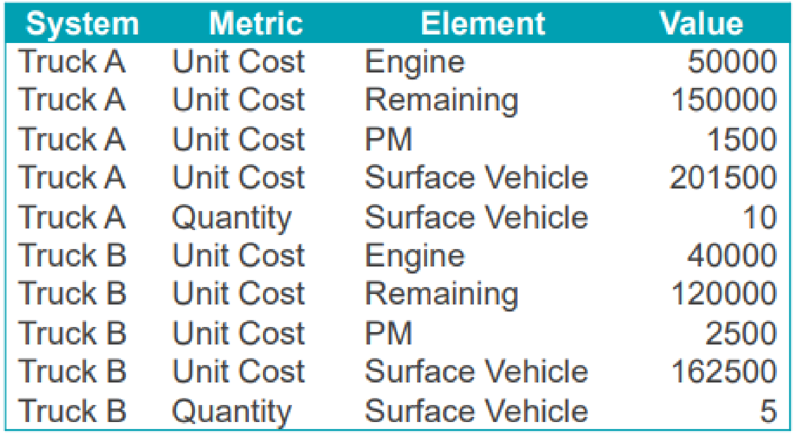
Rows represent unique data observations, **columns** represent variables

Bad



System	WBS Element	Cost ($)
Truck A		
	Engine Cost	50,000
	Remaining Cost	150,000
	PM Cost	1,500
	Number of Units	10
	Total Cost	2,015,000
Truck B		
	Engine Cost	40,000
	Remaining Cost	120,000
	PM Cost	2,500
	Number of Units	5
	Total Cost	812,500

Better

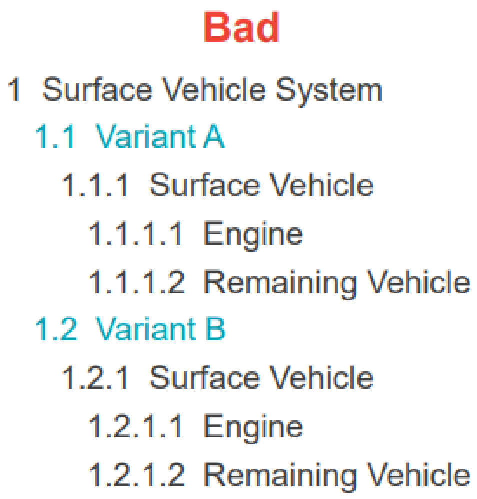


System	Metric	Element	Value
Truck A	Unit Cost	Engine	50000
Truck A	Unit Cost	Remaining	150000
Truck A	Unit Cost	PM	1500
Truck A	Unit Cost	Surface Vehicle	201500
Truck A	Quantity	Surface Vehicle	10
Truck B	Unit Cost	Engine	40000
Truck B	Unit Cost	Remaining	120000
Truck B	Unit Cost	PM	2500
Truck B	Unit Cost	Surface Vehicle	162500
Truck B	Quantity	Surface Vehicle	5

- **Also avoid merged columns as header labels in raw dataset**
- **Similar concept to the format needed to create pivot table in Excel**

Single Purpose Variables

- **A column (or variable) should only contain ONE piece of information**
- **Single purpose variable**



Bad

1 Surface Vehicle System

1.1 Variant A

1.1.1 Surface Vehicle

1.1.1.1 Engine

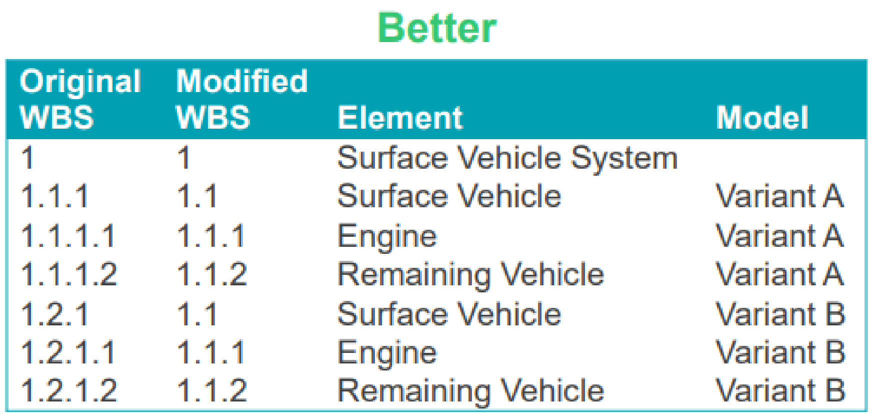
1.1.1.2 Remaining Vehicle

1.2 Variant B

1.2.1 Surface Vehicle

1.2.1.1 Engine

1.2.1.2 Remaining Vehicle

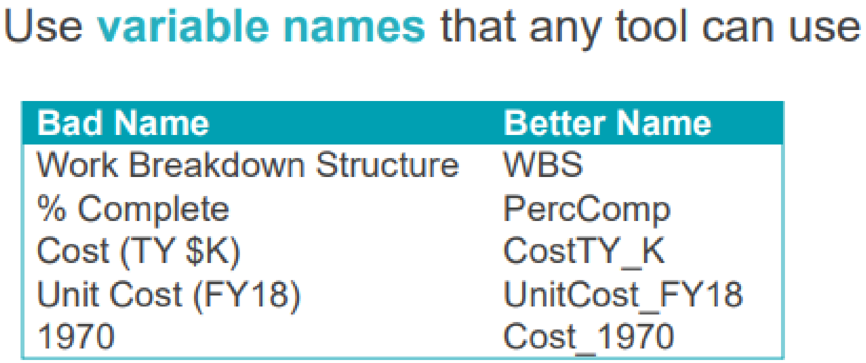


Better

Original WBS	Modified WBS	Element	Model
1	1	Surface Vehicle System	
1.1.1	1.1	Surface Vehicle	Variant A
1.1.1.1	1.1.1	Engine	Variant A
1.1.1.2	1.1.2	Remaining Vehicle	Variant A
1.2.1	1.1	Surface Vehicle	Variant B
1.2.1.1	1.1.1	Engine	Variant B
1.2.1.2	1.1.2	Remaining Vehicle	Variant B

Naming Conventions

- **Best practice for Variable names (i.e. column or field names)**
  - No special characters
  - Unique (don’t name 2 fields the same thing)
  - No spaces (common to replace with underscore)
  - Start with letter



Use **variable names** that any tool can use

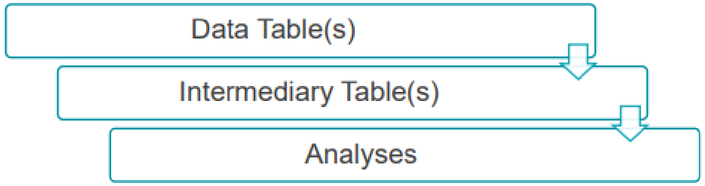
Bad Name	Better Name
Work Breakdown Structure	WBS
% Complete	PercComp
Cost (TY $K)	CostTY_K
Unit Cost (FY18)	UnitCost_FY18
1970	Cost_1970

Other Basic Rules 1/2

Avoid storing **redundant information**

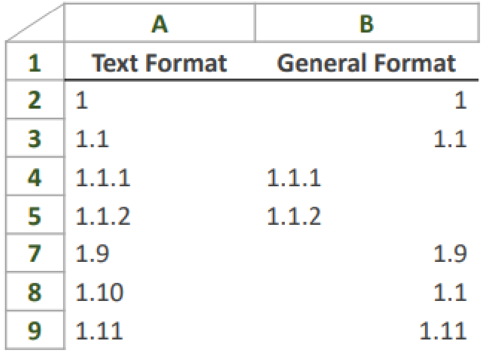
- ✓ Only store child elements
- ✗ Do not store subtotals / totals
- ✗ Do not store calculated variables

Use **intermediary** tables for calculations



Be mindful of **data types**

- Numeric, date, and text
- Excel will make (sometimes wrong) assumptions



	A	B
1	Text Format	General Format
2	1	1
3	1.1	1.1
4	1.1.1	1.1.1
5	1.1.2	1.1.2
7	1.9	1.9
8	1.10	1.1
9	1.11	1.11

Other Basic Rules 2/2

- **Should not have to rely on the physical sequencing – data in table should be unordered-meaning they can be shuffled randomly without any loss of information**
  - Might need to define a variable to define the ordering

More on Data Structures

- **Begin with the end in mind before even starting**
  - Similar to building a cost model – you need to have a purpose/ end goal: what are you trying to accomplish?
- **Obey the data structure rules**
- **The best structure for your data DEPENDS what you're doing**
- **General Rule of Thumb**
  - LONG (more rows) → visuals
  - WIDE (more columns) → analysis
  - Easy to either fold or pivot the data (i.e. turn rows into columns or columns into rows which will be covered in the Dataiku hands on portion)

Data Analytics User Group Overview

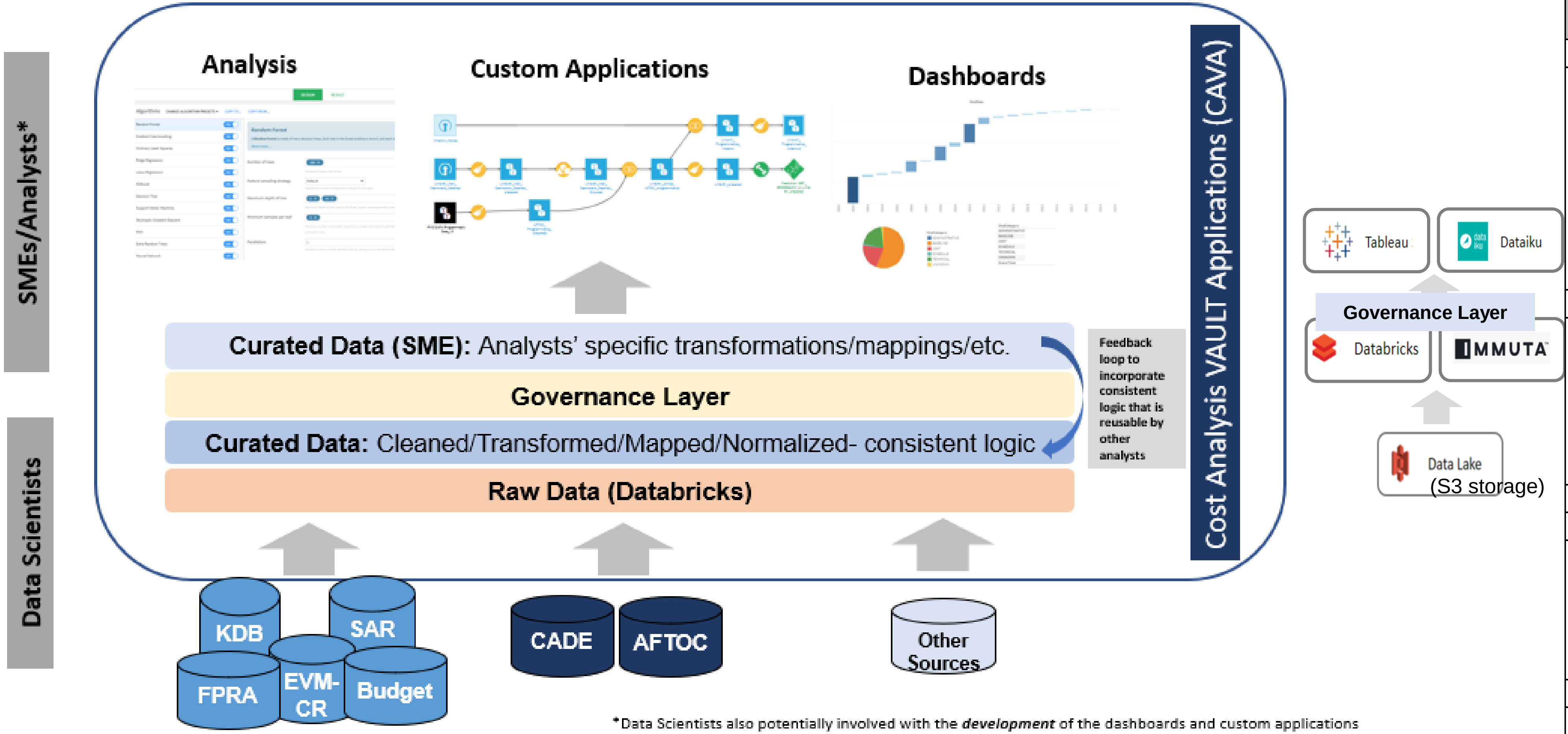
- **What this User Group IS**
  - Analysts from a wide spectrum of different government organizations that are **CURRENT users of advanced data analytic tools** and can represent their organization by talking specifically about them and ideally be able to demonstrate exactly how they are using those tools
  - Liaising with Data Tools Tiger Team (whose mission is to better inform leadership on what is needed to identify, procure and adopt the right tools for the cost community)
  - Government civilians
  - Contractors directly supporting a government organization
  - Meets regularly every 3 weeks
- **What this User Group IS NOT**
  - Providing training for novice users
  - Making authoritative decisions about which tools should be used
  - Industry contractors

Data Analytics User Group Goals

- **Short term Goals :**
  - Create a community of analysts using data analytics tools to collaborate
  - Discuss each tool in detail to include the different ways that each organization is using the tools to their advantage
  - Demonstrate and share results with the group so we can consolidate best practices and lessons learned
  - Collaborate in order to avoid duplicative data analytic efforts and leverage work that has already been done to the greatest extent possible
  - More widespread outreach to the cost community to help with adoption of tools
- **Potential longer-term goals:**
  - Collaboration on products to be eventually hosted in the DTM Hub

Email me for more information or to request to join: sarah.green.10@us.af.mil

Cost Analysis Data Operating Model (CA DOM) Architecture Overview


SMEs/Analysts*
Data Scientists
Analysis
Custom Applications
Dashboards
Curated Data (SME): Analysts' specific transformations/mappings/etc.
Governance Layer
Curated Data: Cleaned/Transformed/Mapped/Normalized- consistent logic
Raw Data (Databricks)
Feedback loop to incorporate consistent logic that is reusable by other analysts
Cost Analysis VAULT Applications (CAVA)
KDB
SAR
FPRA
EVM-CR
Budget
CADE
AFTOC
Other Sources
Tableau
Dataiku
Governance Layer
Databricks
IMMUTA
Data Lake
(S3 storage)

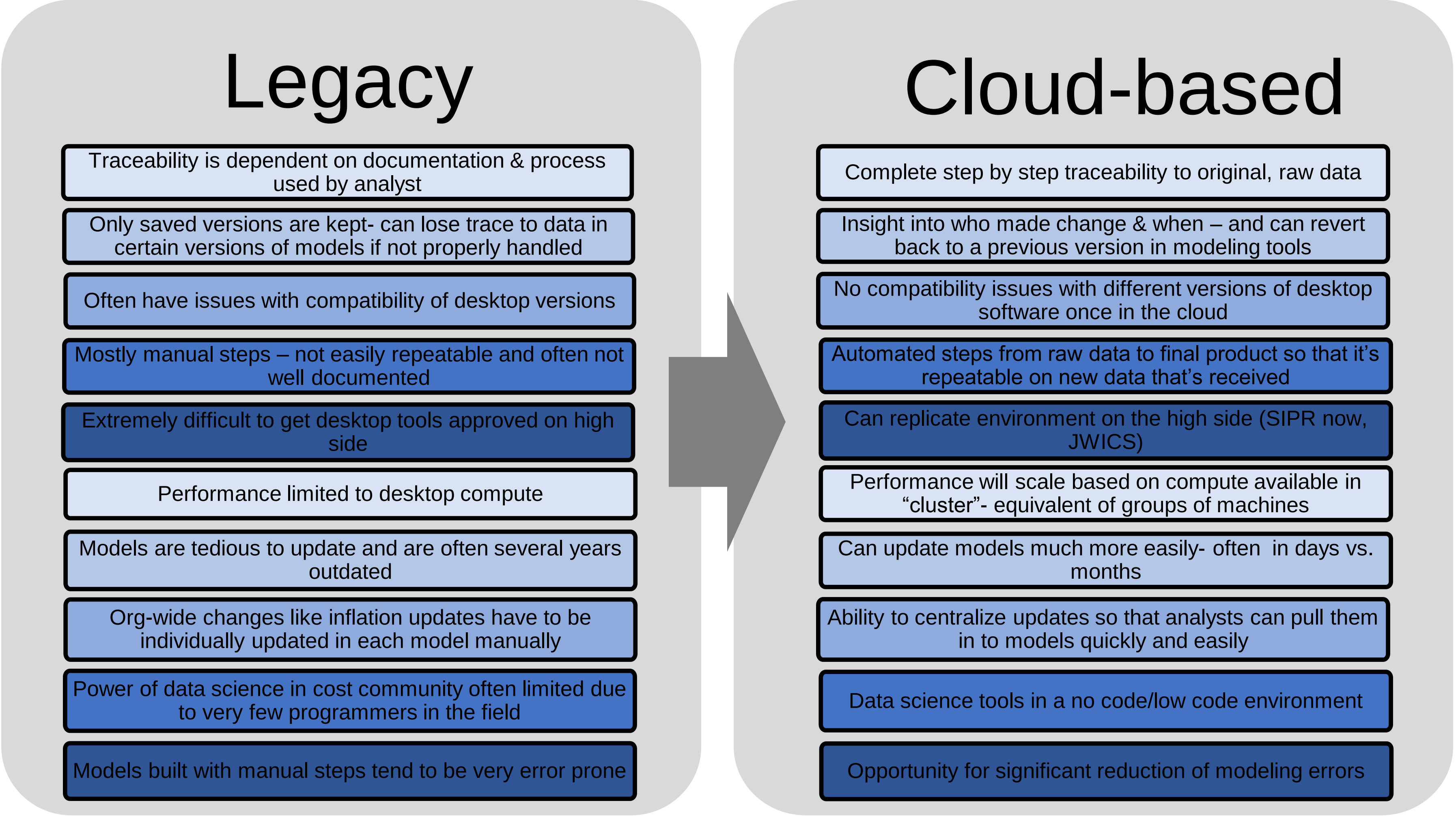
*Data Scientists also potentially involved with the *development* of the dashboards and custom applications

Data Layer	Data Included
Metadata	Metadata
CSDR	FlexFile
	1921s
	1921-1
	1921-2
	1921-3
	SRDR
	TDR
	Maintenance & Repair
SAR	SAR
EVM	IPMDAR
	EVM Format 1
	EVM Format 2
	EVM Format 3
	EVM Format 4
	EVM Format 5
Contracts	Contracts
Budget	Budget
AFTOC	Appropriations Universe
	CAPE Universe
	SDT/MDT
	Indirects/MI6 universe
	Other AFTOC
FPRA	FPRA
Gov Test	Government Test

Integrity - Service - Excellence

UNCLASSIFIED

Realized Benefits



Legacy	Cloud-based
Traceability is dependent on documentation & process used by analyst	Complete step by step traceability to original, raw data
Only saved versions are kept- can lose trace to data in certain versions of models if not properly handled	Insight into who made change & when – and can revert back to a previous version in modeling tools
Often have issues with compatibility of desktop versions	No compatibility issues with different versions of desktop software once in the cloud
Mostly manual steps – not easily repeatable and often not well documented	Automated steps from raw data to final product so that it's repeatable on new data that's received
Extremely difficult to get desktop tools approved on high side	Can replicate environment on the high side (SIPR now, JWICS)
Performance limited to desktop compute	Performance will scale based on compute available in "cluster"- equivalent of groups of machines
Models are tedious to update and are often several years outdated	Can update models much more easily- often in days vs. months
Org-wide changes like inflation updates have to be individually updated in each model manually	Ability to centralize updates so that analysts can pull them in to models quickly and easily
Power of data science in cost community often limited due to very few programmers in the field	Data science tools in a no code/low code environment
Models built with manual steps tend to be very error prone	Opportunity for significant reduction of modeling errors

*AFCAA's assessment of benefits