Progress Report Draft

1. SDLM: Plan Stage

In the Plan stage of the USGS Science Data Lifecycle Model (SDLM), I am defining the datasets and also the methods for organizing them. I am selecting both an organizational framework (SDLM) and a technical structure (Star Schema Data Model) to guide how raw data will be curated, integrated, and prepared for analysis in this project.[1] [2]

More specifically, the SDLM provides an organizational framework for end-to-end data stewardship, ensuring that activities such as Plan, Acquire, Process, Analyze, Preserve, and Publish are performed in a transparent and reproducible way.[1] In contrast, the Star Schema Model is a technical design that defines how data is structured to support efficient storage and usage.[2]

More specifically, in the Process and Analyze phases I plan to transform raw data into the schema below.

1. **Fact Table:** The integrated C-MAPSS time-series data will form the central Fact Table. This contains the core, continuous operational cycles and sensor readings required for Remaining Useful Life (RUL) calculation.[3]
2. **Dimension Tables:** The FAA and ASRS data will be processed to create Dimension Tables. I hope to adjust these data sources to mimic data that provide descriptive attributes (metadata) and maintenance context that enrich the Fact Table. For example, FAA Aircraft Registration data will provide contextual attributes (e.g., engine type, thrust) to conceptually link simulated C-MAPSS unit numbers with real-world aircraft metadata. NASA ASRS reports will undergo NLP to extract structured features (e.g., maintenance functions, failure modes), offering qualitative insights that enrich the Fact Table.

For stages Process and Analyze, I plan to use a python dataframe to analyze the data and for stages Preserve and Publish, I plan to export the data to a persistent table. By planning this SDLM + Star Schema design in advance, I can ensure that datasets can be transformed into a curated, multi-modal structure that is transparent and reproducible.

2. SDLM: Acquire Stage

I have completed this stage by downloading the source data from the relevant websites.

Dataset	Acquisition Activity	Data Format Notes

C-MAPSS	Downloaded the four degradation sub-datasets (FD001-FD004) from the NASA data portal.	The data are provided as a text file with 26 columns of numbers, separated by spaces.[3]
FAA Registration	Downloaded the main Aircraft Registration Master File and associated reference files (e.g., Engine Reference File) from the FAA website.	Comma-delimited text files where data elements are defined by fixed position locations (e.g., Engine Type code is at position 251-252 in the Master File).[4]
NASA ASRS	Downloaded a sample of confidential aviation incident reports from the ASRS database. Need to include the use case example.	Reports are primarily narrative text but are indexed by structured codes based on the ASRS Taxonomy (e.g., Component, Problem). Reports are voluntary and de-identified to protect privacy.[5]

3. SDLM: Process Stage (Cleaning, Transformation, and Integration)

The **Process** stage prepares the collected data for analysis, which includes the design of a database, integration of disparate datasets, and necessary transformation.[1] This stage transforms the raw files into the integrated Star Schema.

3.1. C-MAPSS Fact Table Creation (Time Series Processing)

The C-MAPSS data (columns 1-26) will undergo several transformations to prepare it as the core Fact Table:

- **Initial Cleaning and Loading:** The raw, space-separated C-MAPSS text files will be parsed and loaded into a structured DataFrame. This process will include basic schema validation to ensure the expected 26 columns are present and that data types are numerical.
- **Target Variable Transformation:** The core RUL target variable will be calculated for the training set by determining the number of remaining cycles before failure. This calculated value, rather than a raw sensor reading, becomes the predictive target

variable necessary for the subsequent **Analyze** stage.[2]

- **Feature Engineering:** Sensor measurements will be normalized based on the three operational settings provided in the data (columns 3-5). This transformation step is a critical component of preparing time-series data for predictive modeling.

3.2. FAA Dimension Table Creation (Tabular Processing)

The FAA data requires specialized parsing due to its fixed-position, delimited format [1]:

- **Parsing and Extraction:** The Master File will be parsed to extract contextual metadata elements relevant to a turbofan engine (Type Engine code **5** for Turbo-fan, Engine Mfr Mode Code, Year Mfr).[1]
- **Normalization:** The FAA Engine Reference File will be used as a lookup table to translate raw codes (e.g., Engine Manufacturer Code) into standardized, human-readable names for the Dimension Table.[1]
- **Conceptual Integration:** A synthetic identifier will be generated for the simulated C-MAPSS units. This conceptual link will allow the FAA Dimension Table (e.g., containing the engine manufacturer name and thrust pounds) to be joined to the C-MAPSS Fact Table, demonstrating a robust data integration capability.

3.3. ASRS Feature Table Creation (Qualitative Text Processing)

The narrative data from ASRS reports requires an NLP sub-workflow to extract structured features, creating a Feature Dimension Table:

- **Entity Extraction:** The reports will be processed using text mining techniques to extract structured entities based on the ASRS Coding Form, focusing on fields like **Component** (specifically 'Aircraft Component'), **Problem** (e.g., 'Failed', 'Malfunctioning', 'Improperly Operated'), and **Function.Maintenance**.[1]
- **Feature Quantization:** The extracted qualitative entities (e.g., a report of 'Improperly Operated' by a maintenance technician) will be quantified into new numerical or categorical features (e.g., Boolean flags, frequency counts) that can be integrated as features into the RUL predictive model. This step is a direct method of fusing quantitative sensor data with qualitative human and maintenance context.

4. Cross-Cutting Activities (Describing, Managing Quality, Backup & Secure)

The USGS SDLM requires continuous performance of cross-cutting activities throughout the linear stages [1]:

- **Managing Quality:** Quality assurance measures will be implemented using a data validation framework, such as **Great Expectations (GX)**. This tool will perform automated checks during the Process stage to ensure consistency, accuracy, and completeness. Checks will include:
    - **Schema Validation:** Ensuring the integrated Star Schema tables conform to expected column types and names.
    - **Numerical Range Checks:** Validating that sensor measurements within the C-MAPSS Fact Table fall within realistic or expected operational bounds.
- **Describing:** Detailed documentation (metadata) will be created for the resulting multi-modal dataset. This includes drafting the data dictionary to define variables, units, and the methodology used to derive features (e.g., RUL calculation, ASRS feature quantization).[1]
- **Backup & Secure:** To manage the risk of data loss and ensure reproducibility, all Python scripts used for the Acquire and Process stages, along with the final curated Star Schema tables, will be managed and version-controlled in a Git repository.

Works cited

1. Plale, B., & Kouper, I. (2017). The Centrality of Data: Data Lifecycle and Data Pipelines. In C. R. Sugimoto, H. Ekbia, & M. Mattioli (Eds.), *Big Data is Not a Monolith: Policies, Practices and Problems*. MIT Press.
2. Star Schema, accessed September 29, 2025, https://en.wikipedia.org/wiki/Star_schema
3. CMAPSS Jet Engine Simulated Data - Dataset - Catalog, accessed August 29, 2025, https://catalog.data.gov/dataset/cmapss-jet-engine-simulated-data
4. Download the aircraft registration database - Federal Aviation Administration, accessed August 29, 2025, https://www.faa.gov/licenses_certificates/aircraft_certification/aircraft_registry/releasable_aircraft_download
5. ASRS Database Online - Aviation Safety Reporting System, accessed August 29, 2025, https://asrs.arc.nasa.gov/search/database.html

https://asrs.arc.nasa.gov/search/dbol/samples.html#Example1