Lightspeed Asset Inventory Analysis

Jacob L. Porter

Western Governors University

Table of Contents

A. Proposal Overview

A.1 Research Question or Organizational Need

Lightspeeds organizational need is to see if we can validate and predict which network assets are missing from one or more critical asset management systems, so that data quality issues can be identified and corrected prior to a major security audit.

A.2 Context and Background

Lightspeed operates three core asset management systems—Observability, Inventory, and IPAM—to track all network devices. Discrepancies between these systems can result in incomplete asset visibility, leading to security gaps and audit failures. If records are missing, engineers responsible for remediation actions are unable to proceed with the remediation due to a lack of complete documentation. As such, ensuring completeness and consistency of asset records is essential for accurate vulnerability assessment and regulatory compliance.

A.3 and A3A Summary of Published Works and Their Relation to the Project

The literature reviewed for this project provides a foundation for both methodology and technical implementation. Kim et al. (2020) describe Facebook's approach to network asset management at scale, emphasizing the importance of automated data validation and reconciliation, which directly supports this project's focus on improving asset completeness through analytics. Chapman et al. (2000) present the CRISP-DM methodology, a structured framework that this project follows to ensure a clear, reproducible analytics workflow from data generation through evaluation. Finally, the Udacity Machine Learning DevOps Engineer Nanodegree (Udacity, 2024) provides practical guidance on building automated machine

learning pipelines, version control, and reproducibility, which informed the pipeline design and technical best practices used in this project.

Review of Work 1: Robotron: Top-down Network Management at Facebook Scale

Kim et al. (2020) describes Facebook's approach to managing network assets at scale using a unified data model and automated reconciliation techniques. The paper highlights the technical and organizational challenges of maintaining accurate asset records across distributed systems, emphasizing the importance of automation and structured validation for operational efficiency.

Relation to Project:

This work directly supports the rationale for this project by demonstrating the value of systematic data validation and reconciliation for asset management in large organizations. It reinforces the need for a reproducible analytics pipeline to improve data quality and ensure readiness for audits.

Review of Work 2: CRISP-DM 1.0: Step-by-step Data Mining Guide

Chapman et al. (2000) present the Cross-Industry Standard Process for Data Mining (CRISP-DM), a widely adopted framework for organizing data analytics projects into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The methodology provides a structured approach that improves project reproducibility and clarity.

Relation to Project:

This project applies CRISP-DM to ensure a clear, systematic workflow from problem definition through evaluation. By following this methodology, the project aligns with

industry best practices for analytics and machine learning, ensuring repeatability and traceability.

Review of Work 3: Udacity Machine Learning DevOps Engineer Nanodegree Program

The Udacity Machine Learning DevOps Engineer Nanodegree provides practical training in building modular, automated, and reproducible machine learning pipelines. It covers critical topics such as data validation, pipeline automation, model versioning, and performance evaluation (Udacity, 2024).

Relation to Project:

The skills and concepts from this program directly informed this project's technical implementation. It influenced the design of the MLflow pipeline and reinforced the use of best practices for automation and reproducibility in machine learning operations.

A.4 Summary of Data Analytics Solution

The deliverables that will be included in this solution will incorporate several different aspects. First, there will be a python repository that will contain all the code necessary to calculate the presence of assets across all three systems. Part of the code will be synthetic data generation that emulates the datapoints from the source systems identified, which ensures that no proprietary data or CPNI is at risk. In addition, the code will also provide for a trained supervised model (Random Forest classifier) that will predict asset missingness root cause based on other data features in each dataset. We will incorporate visualizations that break down the feature labels to highlight where the failures most likely exist. We will also deliver two discrete analyses based on different data generation scenarios and weight adjustments to the synthetic data

generator. Lastly, we will have a summary report with key data quality metrics in a python notebook for ease of review.

A.5 Benefits and Support of Decision-Making Process

The most immediate benefit will be a data driven benchmark informing us if we can proceed with our larger project, our security audit. In addition, we will also be able to identify and prioritize areas with high risk of missing asset data. This should help us drive remediations of data quality issues PRIOR to our security audits, which will in turn ensure that we reduce the risk of audit failure. This should also support the organization by evidencing targeted process improvement through pinpointing the main drivers of missing records. Lastly, this will provide an ongoing framework for continuous data validation and improvement in asset management.

B. Data Analytics Project Plan

B.1 Goals, Objectives, and Deliverables

- **Goal 1:**

  Ensure that Lightspeed's network asset data is complete and consistent across all critical management systems in order to support audit readiness and informed business decisions.

  - **Objective 1.1:**

    Quantify the percentage of network assets that are present in each required system (Observability, Inventory, and IPAM).

    - **Deliverable 1.1.1:**

      Data quality summary report showing presence rates for each system and across all systems.

    - **Deliverable 1.1.2:**

      Visualizations and tables that clearly communicate these presence rates.

- **Objective 1.2:**

  Identify and explain the key drivers of missing asset records using machine learning techniques.

  - **Deliverable 1.2.1:**

    A trained predictive model (Random Forest classifier) that highlights which asset characteristics most contribute to missingness.

  - **Deliverable 1.2.2:**

    Feature importance visualizations and explanatory narrative linking model results to business risk.

- **Objective 1.3:**

  Test and demonstrate how changes in system or process risks affect asset data completeness.

  - **Deliverable 1.3.1:**

    Scenario-based analysis using configurable parameters to simulate different risk conditions.

  - **Deliverable 1.3.2:**

    Documentation and supporting code that allow reproducible scenario testing and results comparison.

B.2 Scope of Project

This project is limited to synthetic data simulating network asset records from three core systems (Observability, Inventory, IPAM). The scope includes data generation, cleaning, labeling, descriptive analysis, model training and evaluation, scenario testing, and reporting. It does not include integration with production systems or deployment of live dashboards.

B.3 Standard Methodology

This project uses the **CRISP-DM (Cross-Industry Standard Process for Data Mining)** methodology to organize and guide all phases of implementation. Each step of the methodology aligns directly with the project workflow:

Business Understanding: Clearly define project goals, audit readiness criteria, and the business need for complete asset records across all systems.

Data Understanding: Generate synthetic asset datasets and perform exploratory analysis to verify structure, field content, and initial data quality.

Data Preparation: Clean and standardize the data, label missingness using scenario-driven parameters, and engineer relevant features for modeling (e.g., encoding device type, region).

Modeling: Develop and train a supervised classification model (Random Forest) to predict asset missingness based on the prepared data.

Evaluation: Evaluate model performance using metrics such as accuracy and feature importance and run scenario tests to assess model robustness under different risk conditions.

Deployment: Package all results, code, and supporting documentation, ensuring the solution is reproducible and ready for use in ongoing audit preparation and decision-making.

By following CRISP-DM, the project remains structured, transparent, and adaptable to new scenarios or organizational needs.

B.4 Timeline and Milestones

Milestone or deliverable	Duration	Status	Completion Deadline
Project Planning & Design	2 Hours	Complete	*7/27/2025*
Data Generation & Preparation	4 Hours	Complete	*7/28/2025*
Modeling & Scenario Testing	6 Hours	Complete	*7/29/2025*
Results & Reporting	5 Hours	Complete	*7/30/2025*
Review & Final Submission	3 Hours	Complete	*7/31/2025*

B.5 Resources and Costs

Personel, Technology or Infrastructure	Cost
Personal Laptop	n/a
Python	n/a
MLflow, Pandas, Scikit-Learn, Faker, Jupyter	n/a
VSCode or JupyterLab	n/a
Estimated work hours: 20-25	n/a

The resources necessary for this project do not need to be purchased. The code, libraries and tools are all opensource. No outside staff or labor is necessary for this project.

B.6 Criteria for Success

Project execution will be evaluated using several specific, objective criteria. Success will be measured by the ability to quantify the completeness of asset data—specifically, by calculating the percentage of assets present in all required systems, with a target threshold of at least 75%. The performance of the predictive model will be assessed using standard classification metrics such as accuracy, precision, recall, and F1-score, ensuring the model reliably predicts missingness. Reproducibility will be demonstrated by running all code end-to-end with the provided configuration files and data, confirming that results can be independently replicated. The clarity and completeness of deliverables will be assessed by ensuring that all reports, visualizations, and documentation are understandable by non-technical stakeholders. Finally, the project's success will include the ability to perform scenario analysis, demonstrating how changes in risk parameters impact outcomes and providing actionable insights for decision-makers.

C. Design of Data Analytics Solution

In this part, you will discuss the design details of your Capstone data analytics solution.

C.1 Hypothesis

If an asset is present in the Observability platform, it will also be found in both the Inventory and IPAM systems at least 75% of the time.

C.2 and C.2.A Analytical Method

This project should apply both descriptive and predictive analytical methods to address the research question. Descriptive analytics will be used to quantify asset presence and calculate completeness rates across Inventory and IPAM systems. This method is suitable for establishing a baseline view of current data accuracy and identifying gaps between systems.

Predictive analytics will be implemented using a supervised machine learning model to estimate the likelihood of an asset being missing from either Inventory or IPAM based on attributes such as vendor, region, and role. This approach enables proactive identification of at-risk assets and provides actionable insights to reduce future discrepancies.

C2A: Justification of Analytical Method

Descriptive analytics is justified because it provides a clear, measurable understanding of the existing state of asset data, which is necessary for validating the scope of the problem and informing subsequent predictive modeling. Without this foundational assessment, the organization would lack the context to interpret the results of more advanced methods.

Predictive analytics is appropriate because it aligns directly with the project's goal of improving data completeness by enabling the organization to forecast which assets are most likely to be missing. Supervised machine learning is a suitable technique because the project uses historical labeled data, allowing the model to learn patterns and make accurate predictions that support decision-making. Together, these methods provide a balanced approach that first quantifies the current problem and then enables proactive resolution.

C.3 Tools and Environments

This project will be implemented using Python as the primary programming language due to its extensive support for data analytics and machine learning. Key libraries include Pandas for data manipulation and preparation, Scikit-learn for developing and evaluating machine learning models, Faker for generating synthetic test data, and MLflow for experiment tracking and reproducibility. The development environment consists of Jupyter Notebook for exploratory

analysis and validation of intermediate results, along with Visual Studio Code (VSCode) for structured development of the end-to-end pipeline. Project artifacts will include CSV files for dataset storage, JSON files for configuration management, and PNG images for visual reports, such as model performance charts and feature importance plots. All third-party libraries and tools used in this project are open-source and have been properly cited within the project files. This toolset ensures a controlled, reproducible, and transparent environment for building, testing, and delivering the data analytics solution, aligning directly with the project's goal of improving data completeness and supporting informed decision-making.

C.4 Methods and Metrics to Evaluate Statistical Significance

This project will evaluate its results using both statistical tests and model performance metrics.

Completeness Rate:

- **Null Hypothesis (H_0):** The percentage of observability assets also found in Inventory and IPAM is less than 75%.
- **Statistical Test:** A one-sample proportion z-test will be used to determine whether the observed completeness rate meets or exceeds the 75% organizational threshold.
- **Metrics:** The z-statistic and the corresponding p-value will be calculated.
- **Alpha Value:** $\alpha = 0.05$. If $p \leq 0.05$, the null hypothesis will be rejected, indicating that completeness meets the organizational requirement.

Predictive Model:

- **Model Type:** Supervised classification model.
- **Algorithm:** Random Forest Classifier.

- **Metrics:** Accuracy, precision, recall, F1-score, and feature importance rankings will be used to assess model performance.
- **Benchmark:** The model will be considered successful if the F1-score is ≥ 0.80, indicating strong predictive capability. Feature importance will be reviewed to ensure that the model produces interpretable results aligned with known business drivers for missing records.

C4A. Justification of Analytical Method.

The one-sample proportion z-test is appropriate because it provides a statistically valid method for determining whether the observed completeness rate is significantly greater than or equal to the organizational benchmark of 75%. By applying this test with an alpha level of 0.05, the project can objectively evaluate whether improvements in data completeness are meaningful rather than the result of random variation.

The Random Forest Classifier is an appropriate predictive modeling technique because it can handle categorical and numerical features without extensive preprocessing and provides robust feature importance scores, which are critical for actionable insights. Accuracy, precision, recall, and F1-score are standard metrics for supervised classification and ensure balanced evaluation of the model's performance across both majority and minority classes. The F1-score threshold of 0.80 provides a clear, measurable benchmark for success, while feature importance rankings will allow stakeholders to understand and act on the factors that drive missingness in Inventory and IPAM.

C.5 Practical Significance

Practical significance will be assessed by evaluating how well the analytics solution supports audit readiness and drives actionable business decisions. Specifically:

- **Threshold Validation:** Confirming whether the asset completeness rate exceeds the 75% benchmark, demonstrating sufficient data quality to proceed confidently with audit activities.
- **Model-Driven Insights:** Using predictive model results to identify high-risk device types, regions, or vendors, enabling leadership to prioritize remediation efforts where they will have the most impact.
- **Decision Support:** Providing clear, data-backed criteria that allow leadership to make informed go/no-go decisions regarding audit timing, resource allocation, and process improvements based on current completeness levels.

For example, if the analysis shows that completeness falls below the threshold in specific regions, leadership could delay the audit for those areas and reassign resources to address the identified gaps, reducing audit risk and ensuring readiness.

C.6 Visual Communication

The project will include multiple graphical representations to effectively communicate the findings. Bar charts will visualize asset presence rates across Inventory, IPAM, and both systems combined, enabling stakeholders to quickly assess completeness relative to the 75% audit-readiness threshold. Feature importance plots generated from the machine learning models will highlight the top predictors of missing assets, such as region or vendor, to support targeted remediation planning. Summary tables will accompany these visuals, presenting exact counts, percentages, and statistical test results, such as z-test outcomes, for reference. All visualizations will be produced programmatically in Jupyter Notebook using Python libraries such as matplotlib and pandas, with markdown annotations to explain the results and their significance, ensuring both high-level overviews for decision-makers and detailed evidence for data analysts.

D. Description of Dataset

D.1 Source of Data

The dataset will be fully generated using Python and the Faker library to simulate realistic asset records across Observability, Inventory, and IPAM, with no real or proprietary data used.

D.2 Appropriateness of Dataset

This synthetic dataset is appropriate for the stated goals of the project because it closely mirrors the structure, features, and operational realities of actual asset management systems in large organizations. It contains typical device attributes such as IP address, hostname, region, vendor, and model, along with realistic patterns of missingness, making it well-suited to support both the descriptive completeness analysis and the predictive modeling required to address the research question and organizational need.

D.3 Data Collection Methods

All records will be programmatically generated using Python scripts and the Faker library. Controlled probabilities for missingness shall be defined through configuration files to simulate data quality issues realistically. The entire data generation and labeling process will be automated, ensuring consistency and full reproducibility.

D.4 Observations on Quality and Completeness of Data

The data will be of high quality and completeness by design, except for intentionally introduced missing records. The proportion of missingness is set to test specific audit scenarios, and the dataset supports accurate modeling, analysis, and reproducibility.

D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances

All relevant data governance, privacy, security, and compliance considerations should be addressed. The dataset should be entirely synthetic, and every transformation step documented within the codebase to ensure transparency and reproducibility. No sensitive, personal, or proprietary information will be included, eliminating the risk of privacy breaches. Because no real customer, employee, or company data is used, there are no applicable ethical, legal, or regulatory risks, and the use of synthetic data ensures full compliance.

D5A: Precautions

Precautions will be implemented to ensure proper handling of the dataset:

- **Data Governance:** All processing steps shall be tracked in the source code and version-controlled for full transparency and reproducibility.
- **Privacy and Security:** The project will use only synthetic data, and access to files is restricted to the development environment to prevent unauthorized use.
- **Ethical, Legal, and Regulatory Compliance:** No real-world data will be used, removing the need for additional regulatory safeguards; however, the project is designed to comply with standard software development and data management best practices.

References

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide.* Retrieved from https://www.scirp.org/reference/referencespapers?referenceid=1592779

Kim, J., Naous, J., Lim, H., Yoo, T., Radhakrishnan, S., Wu, J., ... & Mahajan, R. (2020). Robotron: Top-down network management at Facebook scale. *ACM SIGCOMM Computer Communication Review, 50*(4), 426–439. https://research.facebook.com/publications/robotron-top-down-network-management-at-facebook-scale/

Udacity. (2024). *Machine Learning DevOps Engineer Nanodegree Program* [Online course]. Udacity, Inc. https://www.udacity.com/course/machine-learning-dev-ops-engineer-nanodegree--nd0821