Title of Capstone Here

Author's Full Name Here, Including Middle Initial

Western Governors University

Table of Contents

[A. Proposal Overview 4](#_Toc2137083979)

[A.1 Research Question or Organizational Need 4](#_Toc1815560889)

[A.2 Context and Background 4](#_Toc1384268585)

[A.3 and A3A Summary of Published Works and Their Relation to the Project 4](#_Toc2079458267)

[Review of Work 1 5](#_Toc694593955)

[Review of Work 2 5](#_Toc475195288)

[Review of Work 3 5](#_Toc1156669356)

[A.4 Summary of Data Analytics Solution 5](#_Toc374341712)

[A.5 Benefits and Support of Decision-Making Process 6](#_Toc1116633891)

[B. Data Analytics Project Plan 6](#_Toc1561262715)

[B.1 Goals, Objectives, and Deliverables 6](#_Toc1217623361)

[B.2 Scope of Project 7](#_Toc427690838)

[B.2.A Included in Project Scope 7](#_Toc1243851001)

[B.2.B Not included in Project Scope 7](#_Toc2036717847)

[B.3 Standard Methodology 7](#_Toc204762054)

[B.4 Timeline and Milestones 8](#_Toc1049840195)

[B.5 Resources and Costs 8](#_Toc313275103)

[B.6 Criteria for Success 9](#_Toc1337049307)

[C. Design of Data Analytics Solution 9](#_Toc764513894)

[C.1 Hypothesis 9](#_Toc1169566000)

[C.2 and C.2.A Analytical Method 9](#_Toc1622461191)

[C.3 Tools and Environments 10](#_Toc53928856)

[C.4 and C.4.A Methods and Metrics to Evaluate Statistical Significance 10](#_Toc1386867513)

[C.5 Practical Significance 11](#_Toc157862960)

[C.6 Visual Communication 11](#_Toc1263386613)

[D. Description of Dataset 12](#_Toc34323025)

[D.1 Source of Data 12](#_Toc1867631323)

[D.2 Appropriateness of Dataset 12](#_Toc418252330)

[D.3 Data Collection Methods 12](#_Toc1665164903)

[D.4 Observations on Quality and Completeness of Data 12](#_Toc57902398)

[D.5 and D.5.A Data Governance, Privacy, Security, Ethical, Legal, and Regulatory Compliances 13](#_Toc1649406101)

[References 14](#_Toc1322159537)

[Appendix A 14](#_Toc1046046101)

[Title of Appendix 15](#_Toc804972029)

[Appendix B 15](#_Toc128761729)

[Title of Appendix 16](#_Toc1200922991)

[Appendix C 16](#_Toc1897962072)

[Title of Appendix 17](#_Toc1862588384)

[Appendix D 17](#_Toc1292381938)

[Title of Appendix 18](#_Toc390418831)

# A. Proposal Overview

## A.1 Research Question or Organizational Need

My 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. In the event that 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

**Rubric A3:** An accurate summary of 3 different published works that relate to the research question or organizational need in part A1 is provided, and the summary of *each* work includes essential details.

Any citable work produced by an industry or academic professional is acceptable, e.g., whitepaper, online articles, manuals, videos, etc.

**Rubric A3A:** The description logically addresses how *each* published work relates to the research question or organizational need from part A1.

The work only needs to relate to an aspect of your project, e.g., methodology, tools, statistical methods, models, etc.

For each work, start with an overview of the article, summarize the major points discussed, and include an APA style in-text citation, e.g., (Author, year). Next, describe how the work relates to the research question or organizational need provided in A.1.

### Review of Work 1

Include a summary, APA style in-text citation, e.g., (Smith, 2023), and relation to the research question or organizational need here.

### Review of Work 2

Include a summary, APA style in-text citation, e.g., (Purdue University, n.d.), and relation to the research question or organizational need here.

### Review of Work 3

Include a summary, APA style in-text citation, e.g., (Scribbr, 2022), and relation to the research question or organizational need here.

## A.4 Summary of Data Analytics Solution

## The deliverabls that will be included in this solution will incorporate several different aspects. First, there will be a python repository that will contain all of 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 proprieraty 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 a given dataset. We will incorporate visualizations that breakdown the feature lables to highlight where the failures most likely exist. We will also deliver two discrete analysis 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 Genaeration & Preperation | 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 applies 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 and C.4.A 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 (H₀):** 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:** α = 0.05. If p ≤ 0.05, the null hypothesis will be rejected, indicating that completeness meets the organizational requirement.

For *each* model, provide the following information:

* The type of model, e.g., supervised regression, supervised classification, etc.
* The algorithm(s) to be used to develop the model.
* The metric(s) to be used to assess performance.
* The benchmark to which the above metric(s) will be compared to determine success of the model(s), e.g., “If the correlation coefficient is ≥ .6, the model will be considered successful…”

**Rubric C4A:** The submission justifies the chosen methods and metrics, including specific, logical, and well-supported reasons for why the chosen methods and metrics are appropriate for the data analytics solution.

For *each* statistical test or model, describe why it is an appropriate choice. This may repeat parts of section C.2.A.

## C.5 Practical Significance

**Rubric C5:** The submission describes how the practical significance of the data analytics solution will be assessed, including specific criteria regarding whether the solution has provided the expected benefits and supported a decision-making process in the context of the chosen research question or organizational need.

Practical significance refers to how meaningful your findings are in practical application. Results are practically significant when the difference is large enough to be meaningful in real life. This is subjective. But at minimum discuss some criteria to judge the practical significance and how this will be used to support the research question or organizational need from A1. Consider including an example of how the client might apply your work discussed in sections C1 through C.4.A.

## C.6 Visual Communication

**Rubric C6:** The submission describes key details about *each* tool and graphical representation that will visually communicate the findings of the data analytics solution, and the described tools and graphical representations will effectively communicate the expected findings.

Task 3, the Project Report, must include graphic visualizations (at least two) for visually communicating elements of your project (see Task 3: G2). Describe a plan to include at least two visualizations of the data, statistical test(s), or model(s). Specifically, name the types of graphs, what they will visualize, and the tools you’ll use to generate the images.

# D. Description of Dataset

## D.1 Source of Data

**Rubric D1:** *Each* source of the data is correctly identified.

Identify each data source. The minimum number of data sources is one.

## D.2 Appropriateness of Dataset

**Rubric D2:** The discussion provides reasons why the dataset is appropriate for the stated goals of the project.

Describe why each data source provided in section D.1 is appropriate for supporting the research question or organizational need from section A.1.

## D.3 Data Collection Methods

**Rubric D3:** The described data collection methods are thorough.

Describe how each data source listed in section D.1 was collected, e.g., “the data was collected by downloading the .csv file from www.kaggle.com/data\_source\_link.html.”

## D.4 Observations on Quality and Completeness of Data

**Rubric D4:** The summary includes logical and accurate observations on *both* the quality and completeness of the data.

Describe both the quality and completeness of the data and any accommodation needed. Often, data is already clean and complete, but it is still necessary to comment on both.

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

**Rubric D5:** The discussion accurately addresses the data governance, privacy and security, ethical, legal, and regulatory compliance considerations, and *all* of these relate to the dataset and the proposed project.

Specifically, address how *each* of the following relates to your data and project:

* Data governance
* Privacy
* Security
* Ethical, legal, and regulatory compliance considerations

**Rubric D5A:** *Each* described precaution includes specific details about working with and communicating about the data, and there is a precaution described for *each* of the considerations discussed in part D5. *Each* precaution reasonably manages the risk associated with the considerations discussed in part D5.

Describe any necessary precautions. In cases where an item is not relevant, you must explain why. You only need to discuss measures for handling human data if you collected that data.

# References

Purdue University. (n.d.). Retrieved from APA Formatting and Style Guide (7th Edition) - Purdue OWL: https://owl.purdue.edu/owl/research\_and\_citation/apa\_style/apa\_formatting\_and\_style\_guide/index.html

Scribbr. (2022, December 21). *Free Citation Generator*. Retrieved from Scribbr: https://www.scribbr.com/citation/generator/

Smith, J. (2023). A Generic Journal Article Example. *Generic Journal*, 50-62.

# Appendix A

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix B

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix C

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.

# Appendix D

# Title of Appendix

Put any supporting material in these appendices. Add additional or delete superfluous appendices as needed.