COMPETENCIES

**4123.1.1** : **Capstone**

The learner integrates and synthesizes competencies from across the degree program, thereby demonstrating the ability to participate in and contribute value to the chosen professional field.

INTRODUCTION

This task will consist of a proposal for a data analytics project that was approved by an instructor. The proposal will identify a problem and propose a data solution to the problem. Your proposal will also explain how you plan to implement your project to successfully resolve the problem, including a methodology for project development.  
  
Your work for Task 2 will not be evaluated until the appropriate forms in Task 1 have been submitted and evaluated.

REQUIREMENTS

Your submission must represent your original work and understanding of the course material. Most performance assessment submissions are automatically scanned through the WGU similarity checker. Students are strongly encouraged to wait for the similarity report to generate after uploading their work and then review it to ensure Academic Authenticity guidelines are met before submitting the file for evaluation. See [Understanding Similarity Reports](https://cm.wgu.edu/t5/Frequently-Asked-Questions/Understanding-Similarity-Reports/ta-p/252) for more information.    
  
**Grammarly Note:**   
Professional Communication will be automatically assessed through Grammarly for Education in most performance assessments before a student submits work for evaluation. Students are strongly encouraged to review the Grammarly for Education feedback prior to submitting work for evaluation, as the overall submission will not pass without this aspect passing. See [Use Grammarly for Education Effectively](https://cm.wgu.edu/t5/Academic-Coaching-Center/Use-Grammarly-for-Education-Effectively/ta-p/52276) for more information.    
  
**Microsoft Files Note:**  
Write your paper in Microsoft Word (.doc or .docx) unless another Microsoft product, or pdf, is specified in the task directions. Tasks may not be submitted as cloud links, such as links to Google Docs, Google Slides, OneDrive, etc.  All supporting documentation, such as screenshots and proof of experience, should be collected in a pdf file and submitted separately from the main file. For more information, please see [Computer System and Technology Requirements.](https://cm.wgu.edu/t5/WGU-Student-Policy-Handbook/Computer-System-and-Technology-Requirements/ta-p/78)    
 *You must use the rubric to direct the creation of your submission because it provides detailed criteria that will be used to evaluate your work. Each requirement below may be evaluated by more than one rubric aspect. The rubric aspect titles may contain hyperlinks to relevant portions of the course.*

**Project Overview**

**A.  Create a proposal for a data analytics project by doing the following:**

**1.  Describe a research question or organizational need that your project will solve.**

Can we 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?

**2.  Describe the context and background for your project.**

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. Ensuring completeness and consistency of asset records is essential for accurate vulnerability assessment and regulatory compliance.

**3.  Summarize three published works that relate to the research question or organizational need in part A1.**

1. Robotron: Top-down Network Management at Facebook Scale
   1. FULL REFERENCE: 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/>
   2. SUMMARY: This paper describes Facebook’s approach to large-scale network asset management using a unified data model and automated reconciliation techniques. It highlights the organizational and technical challenges of keeping asset records accurate and consistent across complex, distributed systems—mirroring the core challenge addressed in this project.
2. CRISP-DM
   1. FULL REFERENCE: MethodologyChapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. <https://www.scirp.org/reference/referencespapers?referenceid=1592779>
   2. SUMMARY: This guide describes the Cross-Industry Standard Process for Data Mining (CRISP-DM), a widely adopted framework for organizing data analytics and machine learning projects from problem definition through deployment.
3. Udacity ML DevOps Nanodegree (Course)
   1. FULL REFERENCE: Udacity. (2024). *Machine Learning DevOps Engineer Nanodegree Program* [Online course]. Udacity, Inc.
   2. SUMMARY: This professional program covers end-to-end machine learning workflows, including data validation, pipeline automation, version control, model training, and evaluation—all of which informed the technical implementation and best practices used in this project.
4. Describe how *each* published work informs the development of the project.
   1. Robotron - It demonstrates the necessity and value of systematic data validation, reconciliation, and automation for asset management in large organizations, directly supporting the rationale for the analytical approach used in this capstone.
   2. CRISP-DM - The project follows CRISP-DM’s structured phases to ensure reproducibility and clarity: defining objectives, understanding and preparing data, modeling, evaluating, and documenting results.
   3. Udacity ML DevOps Nanodegree - The concepts and skills learned in this course directly influenced the project’s approach to building modular, automated, and reproducible ML pipelines.

**4.  Describe the deliverables (e.g., reports, visuals, apps, or models) for the data analytics solution you will implement to address the research question or organizational need described in part A1.**

* Summary report with key data quality metrics (asset presence rates across all systems)
* Visuals (tables and plots) highlighting missingness and feature importance
* A trained supervised model (Random Forest classifier) to predict asset missingness
* Scenario-based results showing how risk assumptions affect data quality outcomes
* All code and configuration files for reproducibility

**5.  Explain how the data analytics solution will benefit the organization and support a decision-making process.**

This solution will allow Lightspeed to:

* Identify and prioritize areas with high risk of missing asset data
* Drive remediations of data quality issues before security audits, reducing risk of audit failure
* Support targeted process improvement by pinpointing the main drivers of missing records
* Provide an ongoing framework for data validation and continuous improvement in asset management

**Data Analytics Project Plan**

**B.  Describe your data analytics project plan by doing the following:**

**1.  Describe the goals, objectives, and deliverables for the project.**

**Goal:**  
Ensure data quality and readiness for security audit by validating the completeness and consistency of network asset records across all critical systems.

**Objectives:**

* Quantify the percentage of assets present in all required systems.
* Identify key drivers of missingness using machine learning.
* Test how changes in system/process risk affect data completeness.

**Deliverables:**

* Data quality summary report
* Visuals/tables showing presence rates and feature importance
* Trained predictive model for asset missingness
* Scenario-based analysis using configurable parameters
* All supporting code and documentation for reproducibility

**2.  Describe the scope of the 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.

**3.  Explain how you will use a project planning methodology (e.g., ADDIE, SDLC, Agile, CRISP-DM, or SEMMA) to organize and implement your project.**

The project will follow the **CRISP-DM** (Cross-Industry Standard Process for Data Mining) methodology:

* **Business Understanding:** Define project goals and success criteria.
* **Data Understanding:** Generate and explore the synthetic asset data.
* **Data Preparation:** Clean, label, and engineer features as needed.
* **Modeling:** Train and evaluate a supervised classification model.
* **Evaluation:** Assess model results and scenario outcomes.
* **Deployment:** Package results, code, and reports for organizational use.

4.  Provide a timeline with milestones for your project, including the duration and start and end dates for *each* milestone.

|  |  |  |  |
| --- | --- | --- | --- |
| Milestone | Estimated Hours | Status | Completion Deadline |
| Project Planning & Design | 2 | Completed | 07/27/2025 |
| Data Generation & Preparation | 4 | Completed | 07/28/2025 |
| Modeling & Scenario Testing | 6 | Completed | 07/29/2025 |
| Results & Reporting | 5 | Completed | 07/30/2025 |
| Review & Final Submission | 3 | Completed | 07/31/2025 |

5.  Provide a list of resources and *any* associated costs needed to implement the project (e.g., hardware, software, work hours, third-party services).

**Resources:**

* Personal laptop or desktop computer (already available)
* Python (free, open source)
* MLflow, Pandas, Scikit-learn, Faker, Jupyter (free, open source)
* VSCode or JupyterLab (free)
* Estimated work hours: ~20-25 (self-directed)
* No paid third-party services or additional hardware required

**Total cost:** $0 (no additional costs anticipated)

6.  Describe the measurable criteria you will use to evaluate the success of project execution.

* **Data completeness:** % of assets present in all required systems (target: ≥75%)
* **Model performance:** Metrics such as accuracy, precision, recall, F1-score on missingness prediction
* **Reproducibility:** All code runs end-to-end with provided configs and data
* **Clarity of deliverables:** Reports, visuals, and documentation are complete and understandable by non-technical stakeholders
* **Scenario analysis:** Ability to demonstrate how changing risk parameters impacts outcomes

**Design of Data Analytics Solution**

C.  Describe the data analytics solution you will use to address the research question or organizational need identified in part A by doing the following:

1.  Identify the hypothesis of the project.

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.

2.  Identify the analytical methods (i.e., descriptive, diagnostic, predictive, or prescriptive) you will implement in your data analytics solution.

This project uses **descriptive** and **predictive** analytics:

* **Descriptive:** To quantify asset presence and identify completeness rates across systems.
* **Predictive:** To build a supervised machine learning model that predicts which assets are most likely to be missing from Inventory or IPAM, based on asset features.

1. Justify why the analytical methods identified in part C2 are appropriate for your project.

Descriptive analytics is essential for directly answering the completeness research question and establishing current state. Predictive analytics enables proactive identification of risk factors for missing records, allowing the organization to target data quality improvements before the audit.

3.  Describe the tools and environments that you will use to produce the data analytics solution, including any applicable third-party code.

* **Programming Language:** Python
* **Libraries:** Pandas, Scikit-learn, Faker, MLflow
* **Environment:** Jupyter Notebook, VSCode
* **Artifacts:** CSV files (data), JSON files (config), PNG images (reports)
* **All third-party code is open-source and appropriately cited in the project files.**

4.  Describe the methods and metrics you will use to evaluate the output of your data analytics solution or model.

* **Completeness Rate:** Percentage of observability assets also found in Inventory and IPAM.
* **Model Metrics:** Accuracy, precision, recall, F1-score, and feature importance rankings for the supervised model predicting missingness.

1. Justify why the chosen methods and metrics are appropriate for evaluating the output of your data analytics solution or model.

Completeness rate directly measures whether the organizational threshold (75%) is met. Classification metrics are standard for evaluating supervised models, and feature importance reveals actionable drivers for missing records.

5.  Describe how you will assess the practical significance of the data analytics solution, including specific criteria to determine whether it has provided the expected benefits and supported a decision-making process.

Practical significance will be assessed by:

* Determining if the >75% presence threshold is met (supports audit readiness).
* Using model insights to inform targeted remediation and process improvement (e.g., focusing on high-risk device types or regions).
* Enabling leadership to make informed go/no-go decisions about audit timing and resource allocation.

6.  Describe the tools and graphical representations you will use to visually communicate the findings of your data analytics solution.

* **Bar charts and summary tables** showing asset presence rates.
* **Feature importance plots** from the machine learning model.
* **Markdown narratives and annotated visuals** in Jupyter Notebook reports.
* All visuals produced using Python’s plotting libraries (e.g., matplotlib or pandas built-ins).

**Description of Dataset(s)**

D.  Write a description of the data by doing the following:

1.  Identify the source(s) of the data.

The dataset was fully generated using Python scripts and the Faker library to simulate realistic network asset records across three systems (Observability, Inventory, and IPAM). No real or proprietary data was used.

2.  Discuss why this dataset is appropriate for the stated goals of your project.

This synthetic dataset is designed to closely mimic the structure, features, and operational realities of actual asset management systems in large organizations. It includes typical device fields (IP address, hostname, region, vendor, model) and realistic patterns of missingness, which supports both descriptive and predictive data quality analysis.

3.  Describe the data collection methods you used.

All records were generated programmatically using repeatable scripts. Controlled probabilities for missingness were set via configuration files to simulate data quality issues. The entire data generation and labeling process is automated, ensuring consistency and reproducibility.

4.  Summarize your observations on the quality and completeness of the data.

The data is 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.

5.  Discuss data governance; data privacy and security; and ethical, legal, and regulatory compliance considerations that relate to the dataset and the proposed project.

* **Data Governance:** All data is synthetic, and the codebase documents all transformation steps for transparency.
* **Data Privacy and Security:** No sensitive, personal, or proprietary information is present. No risk of privacy breach exists.
* **Ethical, Legal, and Regulatory Compliance:** No real customer, employee, or company data is included. The use of synthetic data eliminates regulatory and ethical risks.

1. Describe the precautions you will need to take when working with and communicating about the data for *each* of the considerations in part D5.

 Clearly label all data and reports as synthetic in any communication or documentation.

 Ensure that scripts and reports are free of any references to actual production or proprietary data sources.

 Maintain transparency in data generation and analysis to uphold ethical standards.

E.  Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

F.  Demonstrate professional communication in the content and presentation of your submission.