DOUGLASCOLLEGE

COMMERCE & BUSINESS ADMINISTRATION
COMPUTING STUDIES & INFORMATION SYSTEMS
COMPUTER AND INFORMATION SYSTEMS (PBD)

CSIS 4495-050: APPLIED RESEARCH PROJECT

Progress Report 1:

End-to-End Data Engineering Solution for HR Analytics

Student Name: Bruno do Nascimento Beserra | 300392300

Instructor: Dr. Bambang Sarif

NEW WESTMINSTER/BC
FALL/2025

Work Hours

Student Name: Bruno do Nascimento Beserra

Date	Number of Hours	Description of Work Done
09/09/2025	2	Initial Research on the topic (ResourceMatch).
09/09/2025	0.5	Write about the Project Goal (ResourceMatch).
09/10/2025	1.5	Write part of the Methodology (ResourceMatch).
09/11/2025	1	Create the project git Repository, Readme and folder Structure (ResourceMatch).
09/12/2025	1.5	Initial Research on the new project.
09/13/2025	2.5	Update the Project Goal,Tech Requirements and Methodology for the new project.
09/20/2025	1	Search for dataset to use in the project
09/21/2025	2	Update the Methodology for the new project
09/21/2025	1	Create databricks workspace
09/22/2025	1	Create new project git Repository, Readme and folder Structure

Description of Work Done

During the first week, I focused on the ResourceMatch project, which was our initial proposal. I began with initial research about the topic, checking the documents from the Riipen project and reviewing key concepts about matching algorithms. Based on this research, I wrote the project goal, and started to write the methodology, where I chose and explained about the classifiers we could test, like logistic regression and XGBoost, and the evaluation metrics we would use for our model comparison. After that first week, we had concerns about the partner company and how the project would proceed in the future, so together with the professor we decided to give up on the project and start another idea more aligned with the data stream of our program.

Starting on September 12th, we began working on a new project related to data engineering. This project describes the whole process of creating a project from scratch and applying key data engineering features to the design. I started with a initial research on state-of-the-art principles and, with previous knowledge acquired from my internship, we decide to implement features like Slowly Changing Dimension type 2, to aggregate historical data without redundancy, Medallion Architecture to improve data quality and governance by separating raw, cleaned, and curated layers, and kimball modelling to optimize querying and reporting. To demonstrate these concepts, we decided to focus on a solution for HR pipelines since this is some of the most valuable data inside a company. Based on these metrics, I updated the project goal, technical requirements for the data part, and methodology with the new guidelines. One challenge that arose was that we did not have access to a real company employee dataset, which made us search for alternative sources. We agreed on using a Kaggle HR dataset as our foundation. Additionally, I created a Databricks workspace to support experimentation and testing, and then built a new GitHub repository and folder structure for the new project. By the beginning of this week, the methodology was revised and finalized.

Repo Check-in of Implementation Completed

This week, I re-created the Github repository with the information of the new project, so the first steps involved setting up a clear and organized repository structure and adding README files to explain the purpose of each folder. I also added sample texts to indicate the types of files expected in the data and databricks folders, as well as created the main README to the root of the repository that describes the overall repository.

Additionally, this progress report has been checked into the GitHub repository under Documents folder