DOUGLAS COLLEGE

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

CSIS 4495-050: APPLIED RESEARCH PROJECT

Progress Report 2:

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

1.0 Work Hours

Student Name: Bruno do Nascimento Beserra

Date	Number of Hours	Description of Work Done
09/26/2025	4	Study General Databricks Concepts: ● Platform ● Workspaces ● Notebook ● Delta Lake
09/27/2025	2.5	Study General Databricks Concepts: ● Clusters ● Job ● ABFSS Paths
09/28/2025	3.5	Create Folders, and Delta Lakes in Databricks Workspace
09/29/2025	1.5	Review ETL and ELT Concepts for Project's Data Flow
10/01/2025	1.5	Review Medallion Architecture concepts
10/01/2025	1	Set Databricks workspace to comply with Medallion Architecture
10/03/2025	2.5	Start working with Notebook for project sample dataset in Databricks
10/04/2025	1.5	Finish creation of first dataset and decide parameters for historical analysis
10/05/2025	0.5	Export Dataset from Databricks to Github
10/09/2025	0.5	Export Python Notebook of dataset creation from Databricks to Github
10/09/2025	2.5	Create Progress Report 2 Document

2.0 Description of Work Done

Between September 26th and October 1st, I focused on consolidating my understanding of Databricks and applying it directly to our project implementation. I started studying general concepts of Databricks, like the platform fundamentals, workspaces, notebooks, Delta Lake, clusters, jobs, and ABFSS paths that gave me knowledge to manage the environment created before. I then reviewed ETL and ELT principles to align our project's data flow with modern data engineering practices and studied the Medallion Architecture to ensure our pipeline design followed a structured framework.

From October 1st to 4th, I created folders and Delta tables inside the workspace to organize data effectively and began working with notebooks to process a sample dataset, we decided to create a sub dataset from our source with 5000 employees as that's the average for a middle size company and would serve its purpose, I also defined the key parameters we're going to keep track during the period of our analysis like promotions, resigns, department changes, and new hires.

Finally, I exported both the dataset and the corresponding Python notebook to GitHub for version control and documentation purposes. These files can be found under the folders data/landing and databricks/notebooks, both under Implementation folder.