

COMMERCE & BUSINESS ADMINISTRATION

COMPUTING STUDIES & INFORMATION SYSTEMS

COMPUTER AND INFORMATION SYSTEMS (PBD)

CSIS 4495-050: APPLIED RESEARCH PROJECT

Mid Term Report:

**End-to-End Data Engineering Solution for HR Analytics**

**Team:** Bruno do Nascimento Beserra | 300392300

Jay Clark Bermudez | 300380540 (Team Leader)

Matheus Filipe Figueiredo | 300389657

**Instructor:** Dr. Bambang Sarif

NEW WESTMINSTER/BC

FALL/2025

## Contents

[**Contents**](#_hx4avysy15vb) **1**

[**1.0 Introduction**](#_gcc65htj0hmh) **2**

[**2.0 Proposed Research Project**](#_wayzqmoelasw) **2**

[**3.0 Project Planning and Timeline**](#_yd7cla9f9dtf) **3**

[3.1 Phase Overview and Milestones](#_b6isksa1uqa7) 3

[3.1.1 Requirement Analysis and Planning](#_bhv1ysyd5z7f) 3

[3.1.2 System Design and Modelling](#_l3ieerxa6pha) 4

[3.1.3 Implementation and Development](#_vkewd63dnea0) 4

[3.1.4 Testing and Validation](#_3rnxgk2cjfpr) 4

[3.1.5 Deployment, Visualization, and Documentation](#_33678yxzq7vh) 4

[3.2 Weekly Work Plan and Schedule](#_ws0ildfe6byc) 4

[**4.0 Implemented Feature Power BI**](#_t552cn49n207) **5**

[4.1 Introduction/Overview of the feature Power BI](#_9rx80ndvqb5f) 5

[4.2 Details of Implementation](#_jzjoyxlc8ccj) 5

[**5.0 Implemented Feature Y**](#_on3ywuaaskir) **5**

[5.1 Introduction/Overview of the feature Y](#_bthr8s7v8t4t) 6

[5.2 Details of Implementation](#_qpxw9za8qmy9) 6

[**6.0 Work Hours**](#_vx12xxlaqczx) **6**

[**7.0 Work Hours**](#_sfj8ovakvsbn) **6**

[**8.0 Acknowledgement**](#_aepipu9sq4ug) **7**

[**9.0 References**](#_d1om02rxg0mi) **7**

## 

## 1.0 Introduction

## 

This project looks at a challenge with Dayforce, a SaaS platform used to manage HR data like employee information and payroll. The problem is that Dayforce does not keep historical records. When an employee leaves, their data is deleted, and when updates are made, older records are replaced. This makes it hard to do historical analysis, track workforce trends, or study issues like employee turnover (Dayforce, 2024). This is not just a Dayforce issue, but a common limitation with HR SaaS platforms (Solutions, 2025).

Platform3 Solutions (Solutions, 2025) notes that not keeping payroll and HR records can lead to compliance issues, problems during audits, and even legal trouble. They stress that companies need a clear plan for keeping and archiving data so it stays available when needed and costs stay under control.

Research shows that without strong historical archives, companies struggle with workforce planning and decision-making (Madden, 2025). Using data engineering techniques like Slowly Changing Dimensions Type 2 and platforms like Databricks and Delta Lake can fix this problem by allowing data to be captured, stored, and analyzed over time (WJARR, 2025).

To solve this problem, the project will build a data pipeline that automatically collects, processes, and saves historical HR data. The pipeline will run in Databricks, using Python and PySpark for transformations and Delta Lake for reliable storage. On top of this, a web app will be built with Django and React to show the results of the analysis.

The finished system will help organizations keep and study HR history in a more efficient way. It will make storage use better, allow faster queries, support long-term workforce analysis, and improve decision-making by giving insights that are not possible with the current setup.

## 2.0 Proposed Research Project

The goal of this project is to design and implement an end-to-end data engineering solution that preserves and enables analysis of historical HR data. The project addresses a key limitation in current HR data management, where employee records are deleted after termination or overwritten when changes occur, making historical analysis impossible.

To solve this, the team will develop a data pipeline that ingests daily data from Dayforce and applies data engineering techniques such as Slowly Changing Dimensions (Type 2) to track changes over time, the medallion architecture to structure data into quality layers, and Kimball data modelling to reduce redundancy and simplify queries.

The solution will be built in Databricks using Python, PySpark, and Delta Lake, and will be paired with a Power BI dashboard that highlights key workforce metrics. Together, these components will demonstrate how historical HR records can be effectively preserved, organized, and analyzed to support long-term workforce insights.

## 3.0 Project Planning and Timeline

### 3.1 Phase Overview and Milestones

Figure 3.1 Gantt Chart

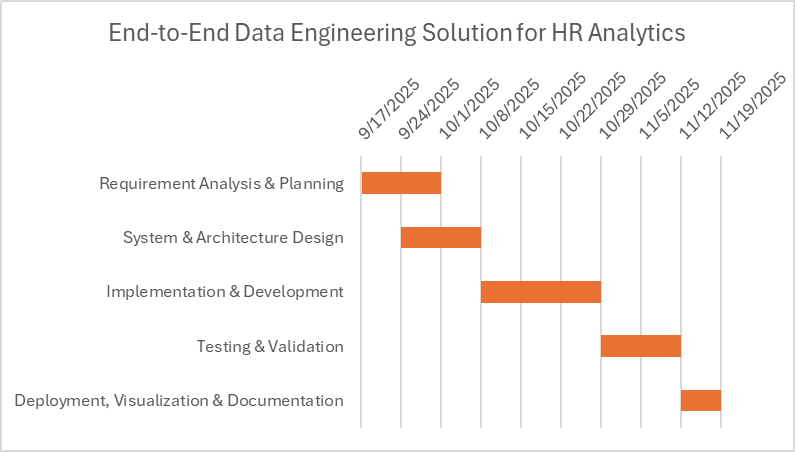


Figure 3.1 shows the Gantt chart, which outlines the five phases of the project: Requirement Analysis, System Design, Implementation, Testing, and Deployment, along with their timelines and dependencies. It shows how the earlier phases provide the foundation for development and testing, which ensures readiness for final deployment.

#### 3.1.1 Requirement Analysis and Planning

The team will assess Dayforce’s data limitations, focusing on record loss after updates or employee exits. From these observations, requirements such as daily ingestion, historical record preservation, and trend analysis will be defined. The scope will outline Databricks, PySpark, Delta Lake, and Django as core tools, with acceptance criteria based on accurate ingestion, retention, and longitudinal querfying.

#### 3.1.2 System Design and Modelling

The system will be designed using the medallion architecture with Bronze, Silver, and Gold layers. SCD Type 2 logic will be specified to capture employee history, and Kimball-style modelling will define fact and dimension tables for analysis. Deliverables include schema definitions and a system design document. Additionally, the web application design will be planned, with Django handling backend processes and React/D3.js enabling interactive dashboards for visualization.

#### 3.1.3 Implementation and Development

A Databricks pipeline will be built to ingest simulated Dayforce data into the Bronze layer, transform and apply SCD Type 2 in the Silver layer, and organize fact/dimension tables in the Gold layer. Delta Lake features such as incremental ingestion, schema enforcement, and merge operations will ensure reliability. The web application will be developed with the backend processing data and the frontend displaying HR metrics such as headcount trends and payroll history.

#### 3.1.4 Testing and Validation

Testing will cover both the pipeline and the web application. Unit tests will confirm that ingestion and transformation processes work correctly, while integration tests will ensure data flows smoothly across all layers. User acceptance testing (UAT) will check that historical records are preserved accurately and that analysis tasks return correct results. The web application will also be tested for responsiveness and accuracy of visualizations.

#### 3.1.5 Deployment, Visualization, and Documentation

The pipeline will be deployed in Databricks and connected to a custom web application for dashboards displaying metrics such as workforce trends, employee retention and turnover, and other key metrics. Final deliverables include technical documentation, versioned code in Git, a project report, and a presentation of the solution.

### 3.2 Weekly Work Plan and Schedule

| Phase | Duration | Due Date | Key Deliverables |
| --- | --- | --- | --- |
| Requirement Analysis & Planning | 2 weeks | Oct 1 | * Approved project proposal * Draft use case diagram * Initial dataset collection and exploration |
| System Design & Modelling | 2 weeks | Oct 8 | * System architecture design * Data engineering pipeline design * Web application architecture design * UI/UX design * Draft test cases * Setup of GitHub repository * Setup Databricks workspace |
| Implementation & Development | 3 weeks | Oct 29 | * Databricks ingestion and transformation pipeline * Lakehouse configuration for Bronze, Silver, Gold layers * Backend development * Frontend development * Unit testing of pipeline and components |
| Testing & Validation | 2 weeks | Nov 12 | * Test execution * Regression testing for pipeline and web app * Bug tracking and resolution * Test summary report |
| Deployment, Visualization & Documentation | 1 week | Nov 19 | * Final pipeline deployment * Web application deployment * Project documentation * Final project presentation and demo |

### 

## 4.0 Implemented Feature Power BI

### 4.1 Introduction/Overview of the feature Power BI

### 4.2 Details of Implementation

## 5.0 Implemented Feature Y

### 5.1 Introduction/Overview of the feature Y

### 5.2 Details of Implementation

## 6.0 Work Hours

### 

## 7.0 Work Hours

**Student Name: Bruno do Nascimento Beserra**

| **Date** | **Number of Hours** | **Description of Work Done** |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Student Name: Jay Clark Bermudez**

| **Date** | **Number of Hours** | **Description of Work Done** |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Student Name: Matheus Filipe Figueiredo**

| **Date** | **Number of Hours** | **Description of Work Done** |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 8.0 Acknowledgement

We would like to acknowledge our instructor, Dr. Bambang Sarif, for guidance and support throughout the project proposal.

## 9.0 References

Asanka, D. (2021). Implementing slowly changing dimensions (scds) in data ware-

houses. Retrieved from https://www.sqlshack.com/implementing-slowly-changing-dimensions-scds-in-data-warehouses/

Databricks, T. (2020). Medallion architecture. Retrieved from https://www.databricks.com/glossary/medallion-architecture

Dayforce. (2020). Dayforce main website. Retrieved from https://www.dayforce.com/

Microsoft. (2022). What is git? Retrieved from https://learn.microsoft.com/en-us/devops/develop/git/what-is-git

Nguyen, H., Pham, H., & Chin, C. (2020). The analytics setup guidebook. Holistics.

Retrieved from https://www.holistics.io/books/setup-analytics/

D3 by Observable. (2020). What is d3? Retrieved from https://d3js.org/what-is-d3

Django Software Foundation. (2018). Django: The web framework for perfectionists

with deadlines. Retrieved from https://www.djangoproject.com/

W3Schools. (2017). React introduction. Retrieved from https://www.w3schools.com/react/react\_intro.asp

Kaggle, T. (2025). Hr dataset (multinational company). Retrieved from https://www.kaggle.com/datasets/rohitgrewal/hr-data-mnc

Dayforce. (2024). Personal data retention policies. Retrieved from https://help.dayforce.com/r/ImplementationGuide/Dayforce-Implementation-Guide/Personal-Data-Retention-Policies

HRBrain. (2024). Hr saas platform analytics and reporting. Retrieved from https://hrbrain.ai/blog/hr-saas-platform-analytics-and-reporting/

Madden, S. (2025). Combating turnover using people analytics. Retrieved from

https://www.scottmadden.com/insight/combating-turnover-using-people-analytics/

Solutions, P. (2025). Payroll data retention management: How to do it right. Retrieved from https://platform3solutions.com/blog/solution-for-keeping-payroll-records-data-retention-compliance/

WJARR. (2025). Developing scalable hr analytics platforms for smes with scd2.

Retrieved from https://journalwjarr.com/sites/default/files/fulltext\_pdf/WJARR-2025-2920.pdf