Course Navigator

Course Completion and Learning Path Optimization

10-10-2024

Contents

1 Introduction

1.1 Overview

The Course Management Application is a digital platform designed to enhance employee learning and development within organizations. It serves as a centralized hub for tracking employee progress, managing course assignments, and promoting skill enhancement. With a user-friendly interface, both Admins and Employees can navigate the platform seamlessly, ensuring efficient management of educational resources.

Key features include:

- **User Roles**: The application supports Admin and Employee roles, each with tailored functionalities.
- **Progress Tracking**: Admins can monitor employee progress in real time, while Employees can track their own advancements.
- **Course Assignment**: Admins can assign courses based on roles and skill gaps, with data-driven recommendations for optimal course selections.
- **Skill Management**: Employees can add new skills and update proficiency levels, encouraging continuous learning.
- **Reporting and Analytics**: Admins have access to a dashboard that visualizes key performance indicators (KPIs), aiding informed decision-making.

Integrating data engineering and machine learning techniques, the application provides valuable insights that support the learning journeys of both Admins and Employees.

1.2 Objectives

The primary goals of the Course Management Application are:

- **Employee Progress Tracking**: To enable Admins to monitor employee progress across courses, facilitating timely interventions.

- **Course Assignment Management**: To streamline course assignments, matching them to employee needs based on performance and skills.

- **Skill Development**: To support employees in growing their skill sets by adding new skills upon course completion.

- **Data-Driven Insights**: To utilize data science methodologies for providing recommendations that enhance learning paths.

- **Enhanced User Experience**: To create an intuitive interface that allows users to navigate easily and maximize the platform's value.

By achieving these objectives, the Course Management Application aims to foster a culture of continuous learning and development, driving improved performance and engagement among employees.

2   Architecture Overview

This architecture is designed to facilitate seamless integration across data engineering, data science, and application development, providing end-to-end solutions from data collection to actionable insights.
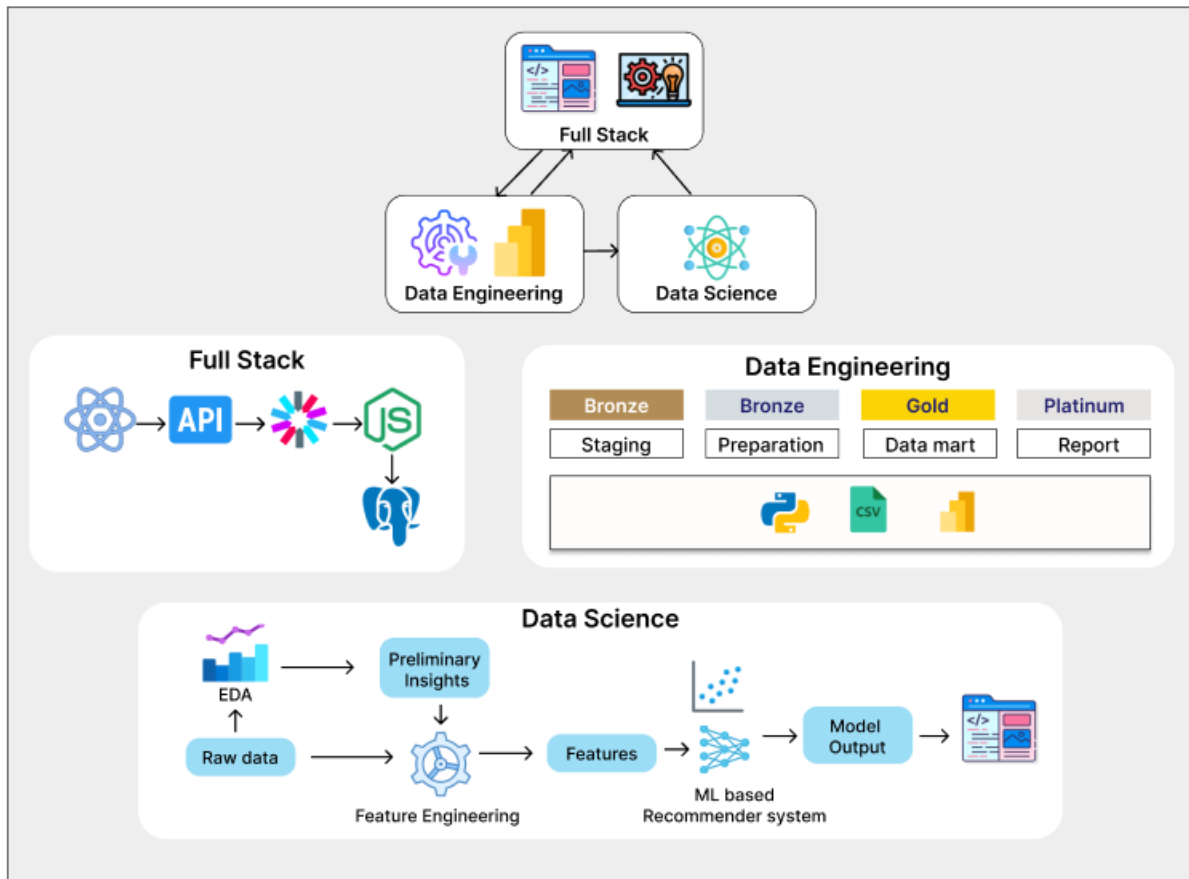
FIG:1 ARCHITECTURE DIAGRAM

2.1   Full Stack Architecture

The Full Stack architecture is responsible for delivering the user interface and managing interactions between the frontend and backend of the system.

- **Frontend**: Built with React, the frontend provides an intuitive user interface that allows users to interact with reports and machine learning results.

- **API Layer**: REST API connects the frontend to the backend, handling user requests and data retrieval for report generation and insights.

- **Backend**: Developed using Node.js, the backend manages business logic, user authentication via JSON Web Tokens (JWT), course progression, and interacts with the PostgreSQL database.

- **Database**: PostgreSQL is utilized to store user data, course details, progress tracking, and other essential entities.

2.2 Data Engineering Pipeline

The Data Engineering Pipeline processes data from ingestion to reporting, ensuring that it is clean, transformed, and ready for both business intelligence and machine learning applications.

- **Bronze Layer (Staging)**: This initial stage ingests raw, unprocessed data from the Postgres database, storing it without transformations.

- **Bronze Layer (Preparation)**: Data is cleaned and transformed at this stage, addressing missing values and standardizing formats for downstream processes.

- **Gold Layer (Data Mart)**: Clean, structured data is organized in this layer, ready for analysis. It includes the creation of facts and dimensions to facilitate easy querying for reporting.

- **Platinum Layer (Report)**: Data is further aggregated in this layer to support specific KPIs and reporting needs in tools like Power BI, aimed at both employees and admins.

TOOLS USED:

- **Python**: For data processing, cleaning, and automation.

- **CSV**: Acts as an intermediate format for easy data transfer.

- **Power BI**: The primary tool for creating dashboards and reports that provide insights into employee progress, course completion, and skills development.

2.3 Data Science Workflow

The Data Science layer focuses on extracting advanced insights and predictions from the processed data, utilizing machine learning models to generate key recommendations.

- **Exploratory Data Analysis (EDA)**: Initial data analysis uncovers patterns and provides insights, forming a baseline for model development.

- **Feature Engineering**: Important features are selected or created based on EDA insights, which may include skill progress, course difficulty, and user behaviour patterns.

- **ML-Based Recommender System**: Machine learning models generate personalized recommendations, such as course suggestions tailored to user profiles and past performance.

- **Model Output Integration**: The outputs from machine learning models are integrated back into the full stack system, enabling users to view personalized insights and recommendations within the interface.

3 Full Stack Development

3.1 Modules

3.1.1 Authentication Module

Overview:
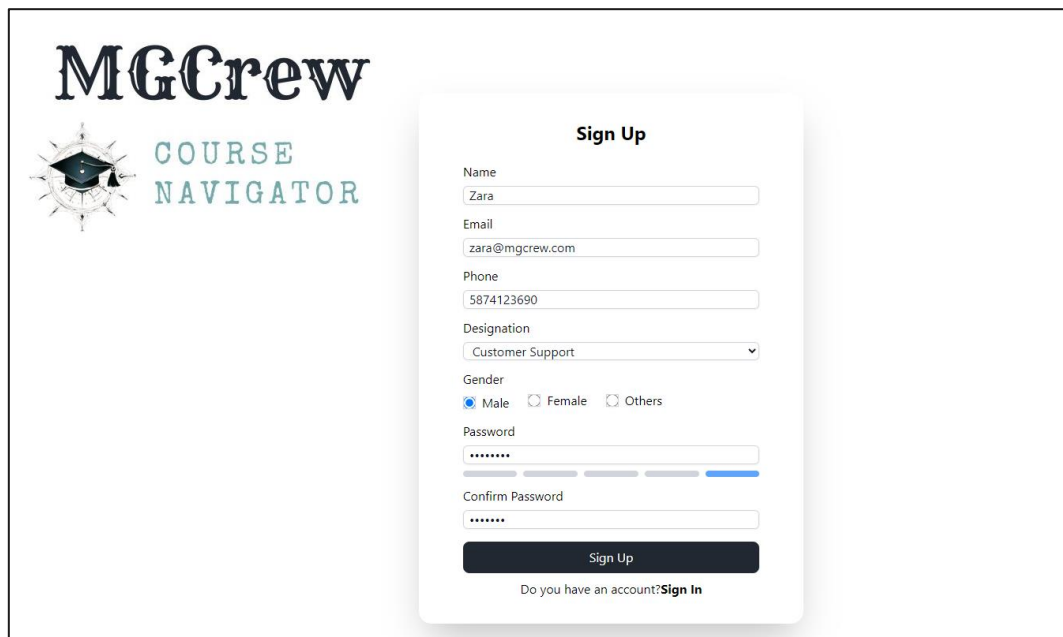
The Authentication Module ensures secure access to the application for both Admins and Employees. It implements JWT (JSON Web Tokens) for user authentication and session management.

Key Features:

- Sign Up: Employees can create an account using their official email addresses, which are validated for authenticity. Passwords are setup during signup.
- Login: A secure login process for Admins and Employees that requires only the official email address for access.
- Role-Based Access Control: Ensures that Admins and Employees have access only to the functionalities relevant to their roles.
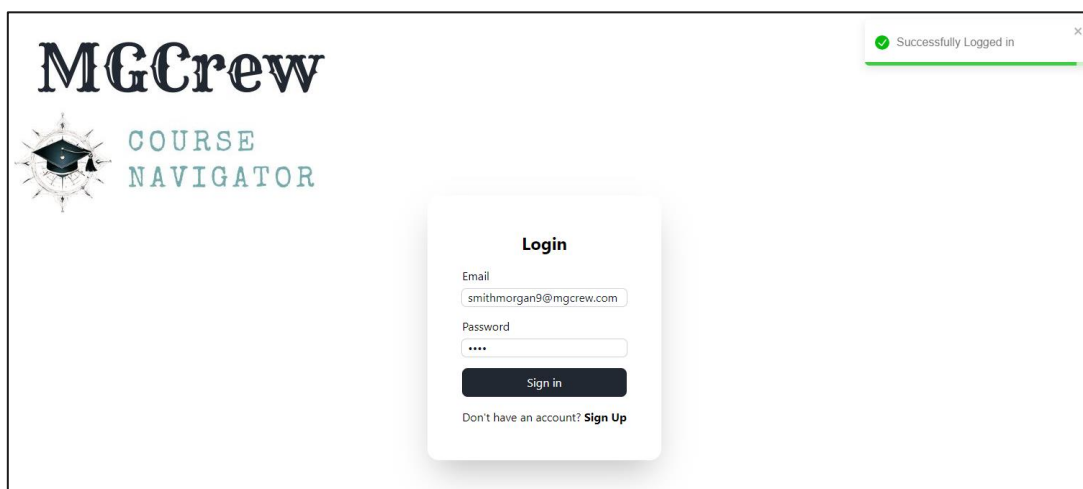
User Flow:

- Employee/Admin signs up or logs in using their official email address.
- JWT is generated upon successful login, granting access to the application.
- User can log out, invalidating the JWT.

FIG:2 SIGNUP MODULE

FIG:3 LOGIN MODULE

3.1.2 Admin Dashboard Module

Overview:

The Admin Dashboard provides a centralized view of key metrics and insights regarding employee progress and course management, utilizing Power BI for enhanced data visualization.

Key Features:

- **Performance Indicators:** Visual representations of employee progress distribution, course completion rates, and average modules completed over time.
- **Employee Tracking:** List of employees behind on courses for more than 30 days, allowing for timely intervention.
- **Course Insights:** Top 10 courses by enrolment and skill fulfilment analysis by designation, enabling data-driven decisions.
- **Data Visualization:** Use of Power BI graphs and charts to represent data clearly and intuitively, enhancing interpretability.
- **Filters:** Dynamic filtering options to customize the view based on specific parameters, such as designation or course difficulty.

User Flow:

- Admin logs in to access the dashboard.
- Admin can navigate through various insights and performance metrics.
- Admin can apply filters to tailor the displayed data.
- Admin can click on specific metrics for a detailed view.
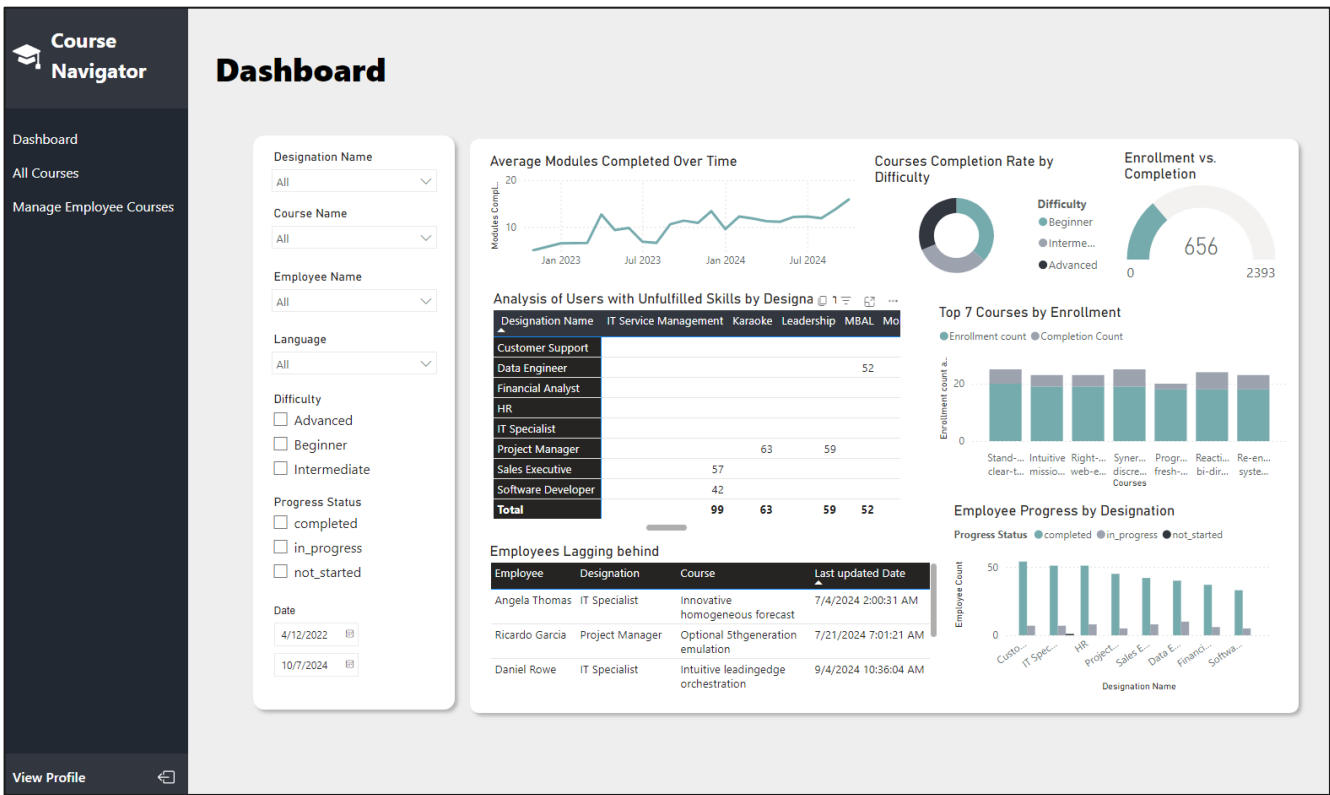
FIG:4 ADMIN DASHBOARD MODULE

3.1.3 View Courses Module

Overview:

The View Courses Module allows Admins to access and manage all courses available on the platform, while Employees can view only the courses assigned to them, grouped by their completion status.

Key Features:

- Comprehensive Course List: Admins see all courses; Employees view only assigned courses.
- Organized Course Display: Employees' courses are categorized by Completed, In Progress, and Not Started.
- Row Filtering Options: Both Admins and Employees can filter courses quickly based on various criteria.
- Course Search Functionality: Instant access to specific courses for both user roles.

User Flow:

- Admins:
  - Admin navigates to the All-Courses page.
  - Admin can apply filters to view relevant courses by designation, difficulty, and language.
  - Admin can use row filtering to quickly find specific courses within the list.
  - Admin clicks on a course to view detailed information, including enrolment data and completion rates.
- Employees:
  - Employee navigates to the All-Courses page.
  - Courses are displayed in categorized sections: Completed, In Progress, and Not Started.
  - Employee can use row filtering to find specific courses easily.
  - Employee clicks on a course to view detailed information and update their progress.
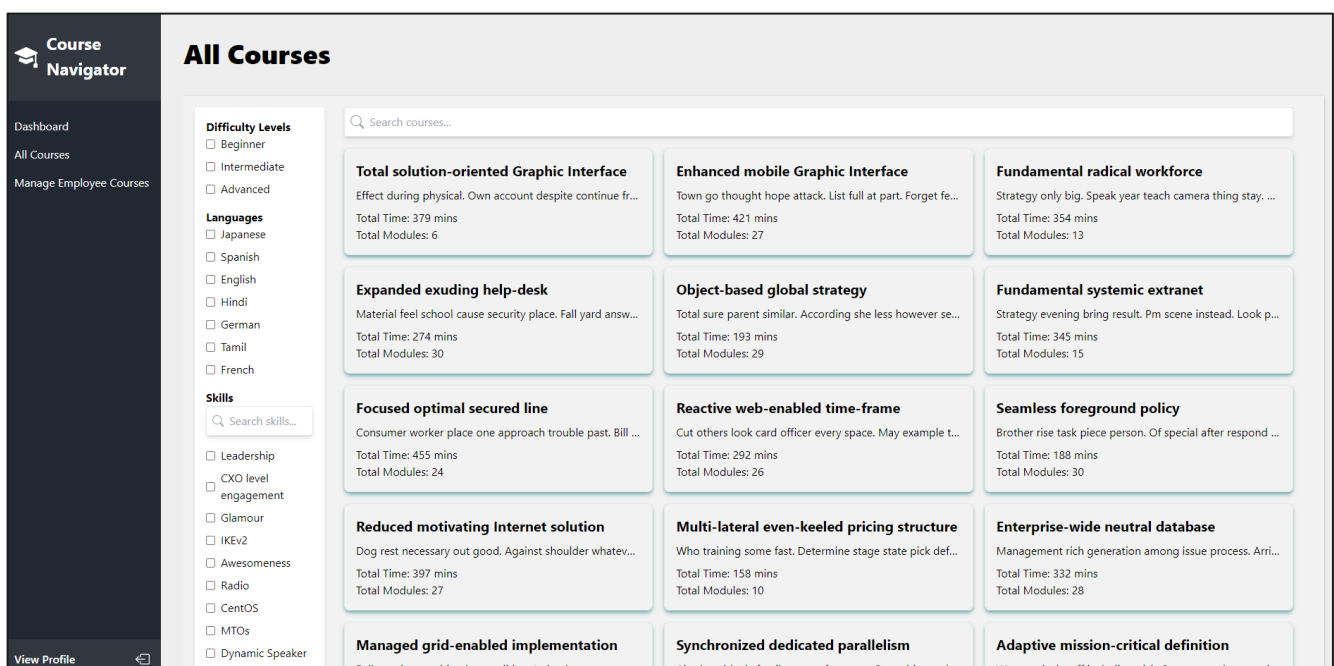
FIG:5 ALL COURSES - ADMIN
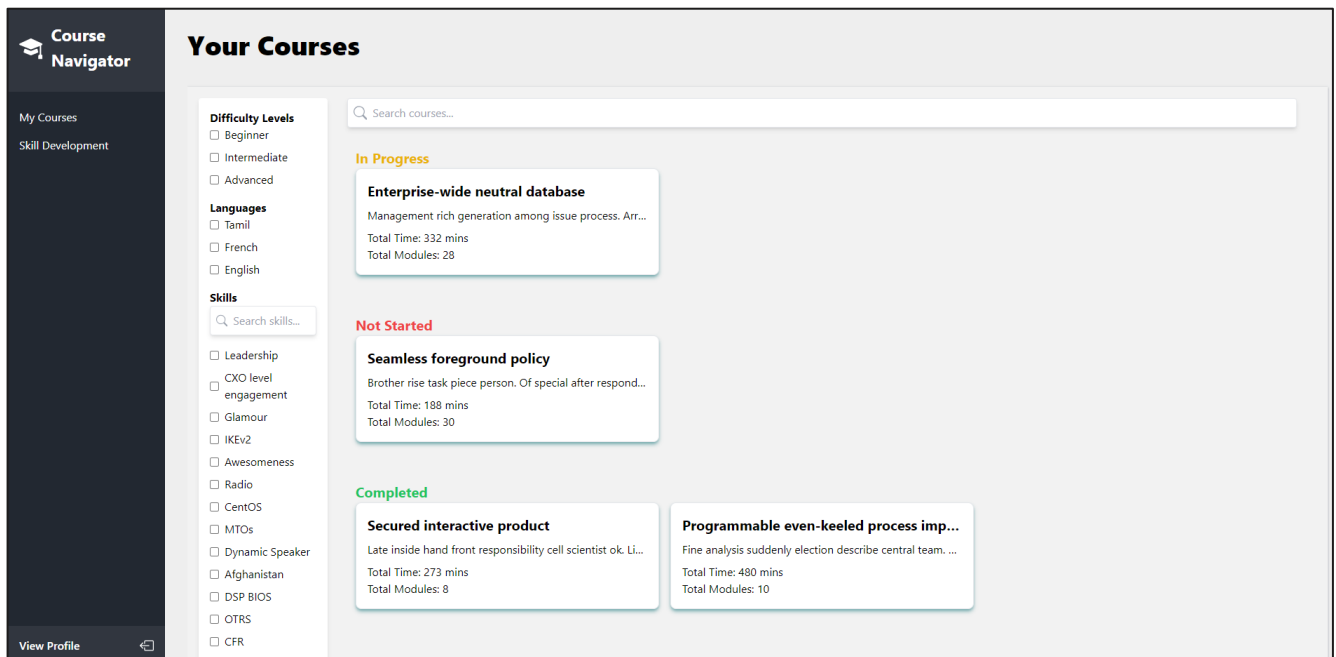
FIG:6 ALL ASSIGNED COURSES - EMPLOYEE

3.1.4 Detailed Course Module

Overview:

The Detailed Course Module provides in-depth information about each course and allows Employees to track and update their progress.

Key Features:

- Course Details: Includes course description, total time, modules, language, and difficulty level.
- Enrolment Data: Displays enrolment numbers and completion rates, categorized by department.
- Employee Progress Tracking: Employees can view their progress for the course and update it as needed.

User Flow:

- Admins:
  - Admin selects a course from the course list.
  - Detailed course information is displayed, including all key features.
- Employees:
  - Employee selects a course from their assigned courses.
  - Detailed course information is displayed, including their progress.
  - Employee can update their progress for the course.
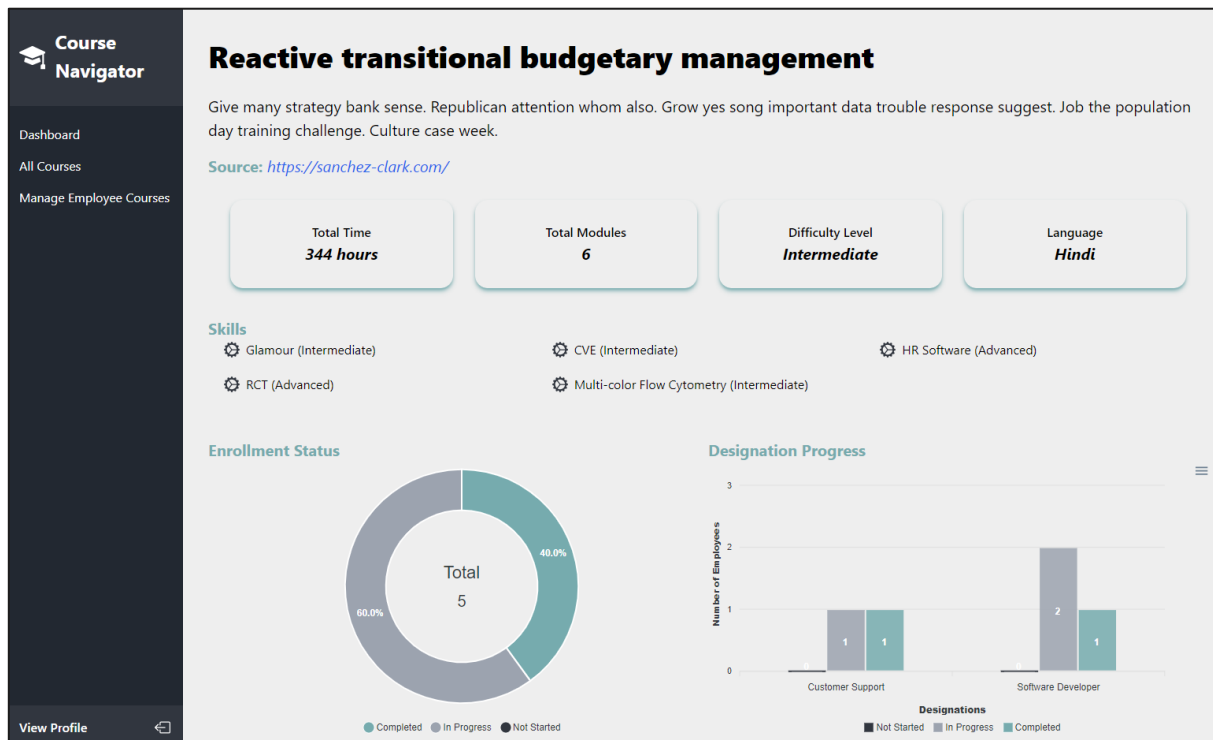
FIG:7 DETAILED COURSE MODULE -ADMIN
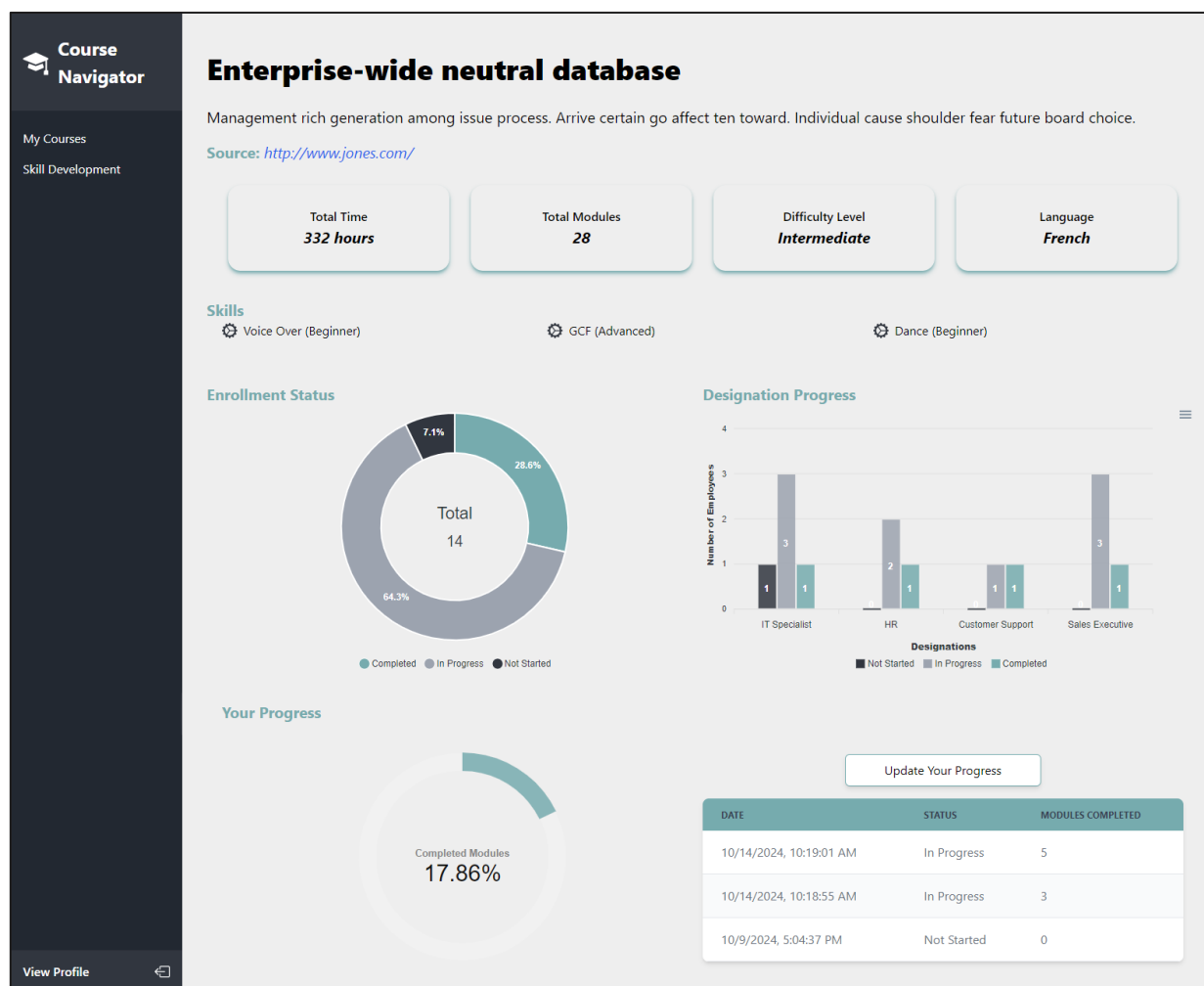
FIG:8 DETAILED COURSE - EMPLOYEE

3.1.5 Employee Progress Module

Overview:

The Employee Progress Module allows Admins to monitor and manage employee course progress effectively. It provides a comprehensive view of assigned courses and their current progress status.

Key Features:

- Employee Course Table: Displays a list of employees and their assigned courses, including the latest progress state for each course.
- Filters and Search Bar: Admins can filter the list based on criteria such as course status or designation and a search bar allows for quick access to specific courses.
- Sort By Options: Admins can sort the table by various attributes, such as employee name, course name, or progress status.
- Course Assignment: Admins can easily assign new courses to employees directly from the table interface.

User Flow:

- Admin navigates to the Employee Progress section.
- Admin views the table listing all employees, their assigned courses, and the latest progress states.
- Admin uses filters, the search bar, or sort options to refine the displayed data.
- Admin selects an employee to view more details or assigns new courses as needed.

FIG:9 EMPLOYEE COURSE PROGRESS - ADMIN
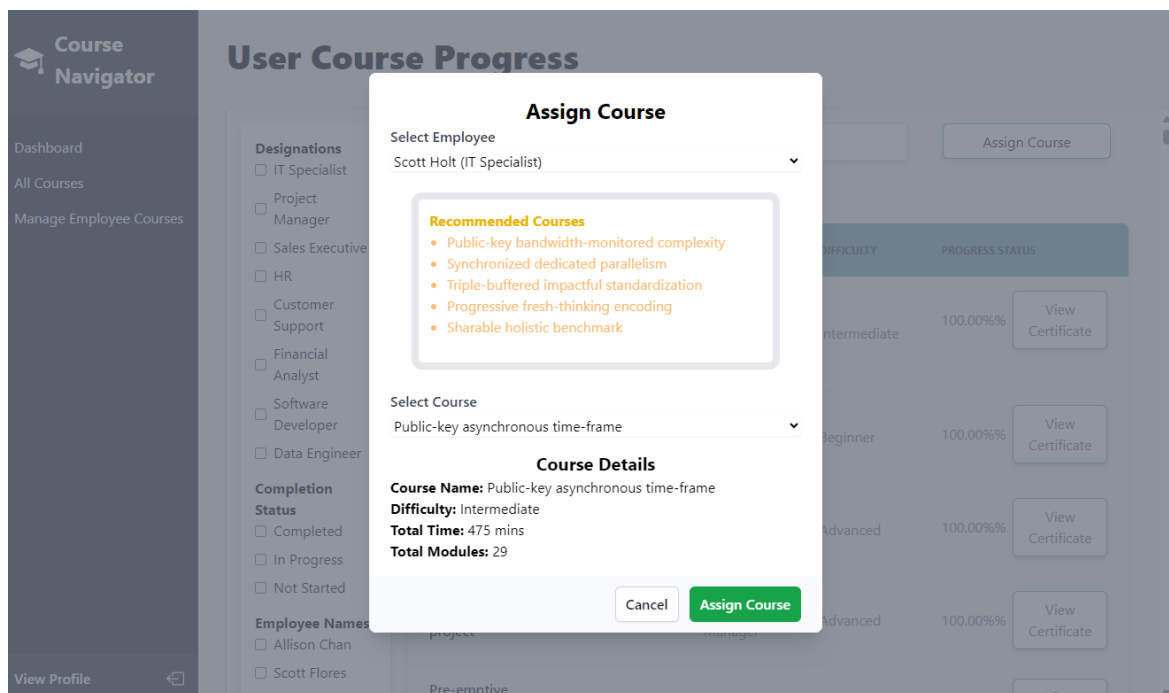
FIG:10 ASSIGN EMPLOYEE COURSE - ADMIN

3.1.6 Profile Management Module

Overview:

The Profile Management Module allows both Admins and Employees to view and edit their profiles.

Key Features:

- Profile Information: Displays personal details such as name, gender, phone, mail, along with designation.
- Edit Functionality: Users can update their data to ensure information is current, but cannot change their designation and mail.

User Flow:

- User accesses the Profile Management page.
- User views their current details and can make updates to their profile.
- Changes are saved and reflected in the profile.
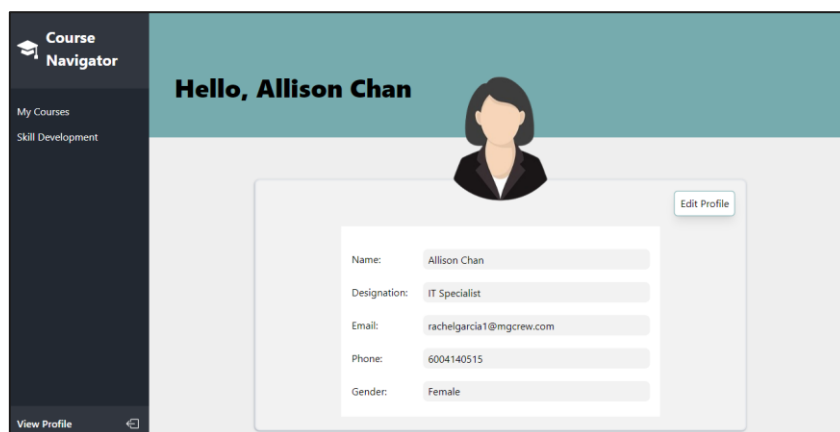
FIG:11 VIEW PROFILE MODULE
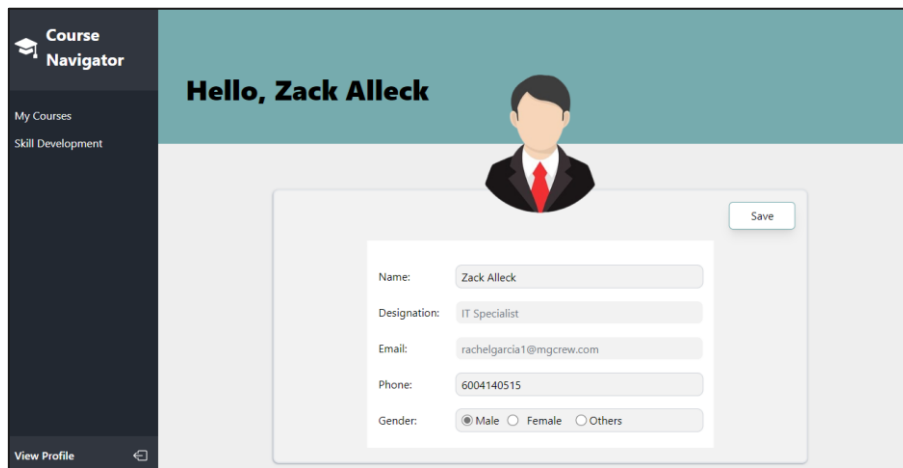
FIG:12 EDIT PROFILE MODULE

3.1.7 Manage Skills Module

Overview:

The Manage Skills Module enables Employees to view and manage their skills within the application. As employees' complete courses, their skill levels are automatically updated based on their progress.

Key Features:

- Skill Overview: Displays current skills along with proficiency levels that reflect the employee's progress and course completions.
- Automatic Skill Level Update: When an employee completes a course with a greater skill level, their corresponding skill proficiency level increases automatically.
- Skill Addition: Employees can add new skills manually along with their proficiency levels.

User Flow:

- Employee navigates to the Manage Skills page.
- Employee views existing skills and their current proficiency levels.
- Upon course completion, the system automatically updates relevant skills.
- Employee can add new skills or manually along with their proficiency levels.
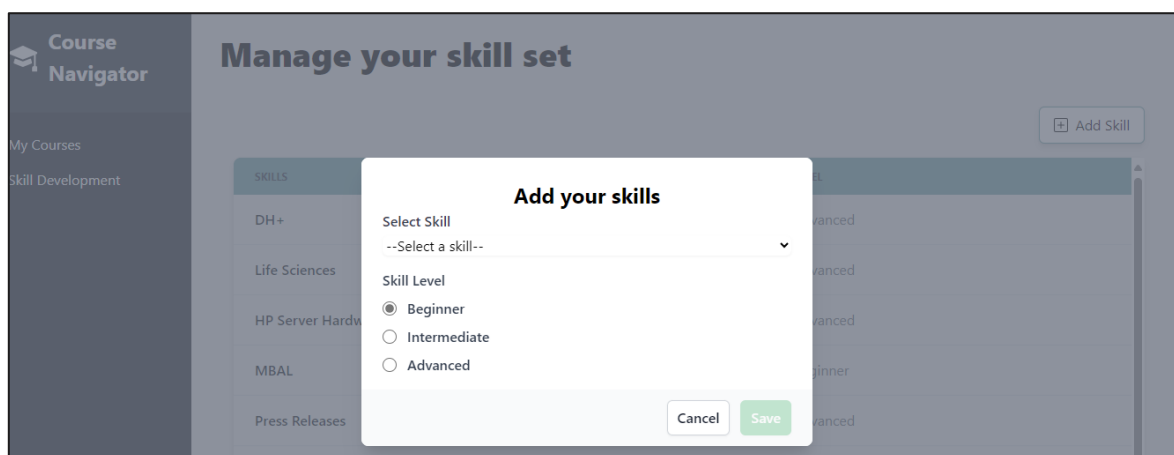
FIG:13 ADD SKILLS MODULE
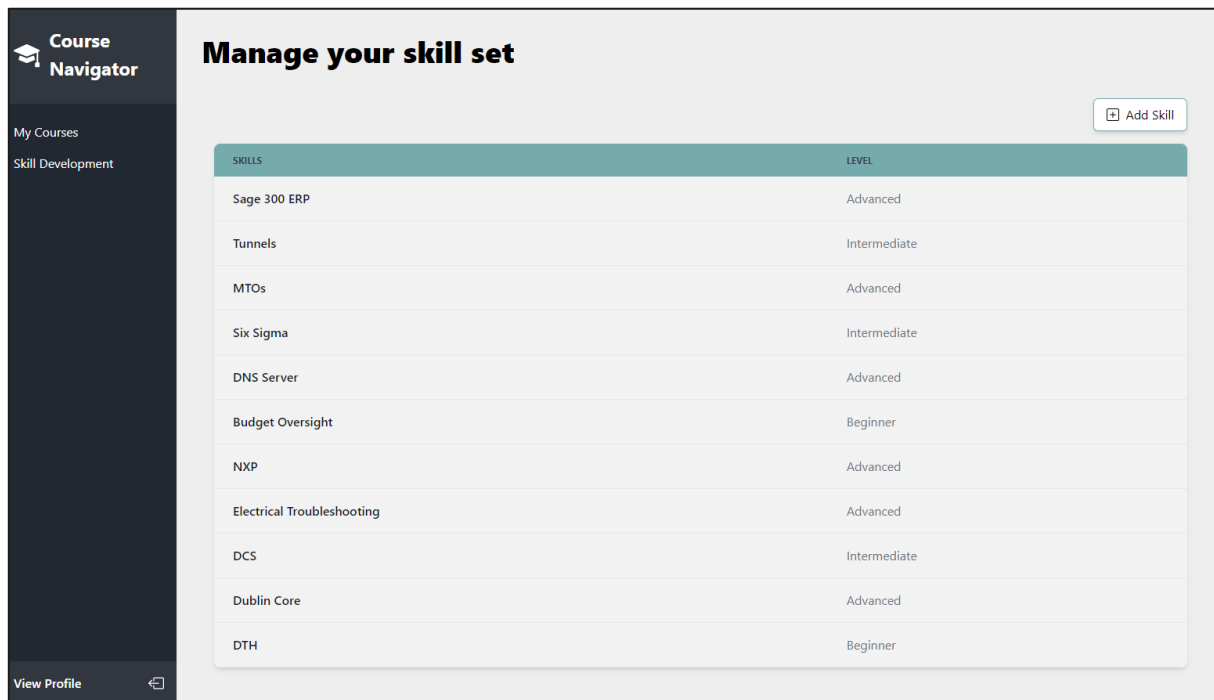
FIG:14 MANAGE SKILLS MODULE

3.2   Routes

- **Course Information Route** (/courseInfo/:courseId)
    - o **Component:** CourseDetail
    - o **Access Control:**
        - ▪ Checks if the user is authenticated.
        - ▪ Validates if the course ID exists in the list of valid course IDs fetched from the API.
    - o **Description:** Renders the CourseDetail component if the user is authenticated, and the course ID is valid. Redirects to /not-found if the course ID is invalid.
- **Update Skills Route** (/updateSkills)
    - o **Component:** SkillSet
    - o **Access Control:**
        - ▪ Checks if the user is authenticated.
    - o **Description:** Renders the SkillSet component if the user is authenticated. Redirects to the login page if not.
- **Update Courses Route** (/updateCourses)
    - o **Component:** EmployeeCourseList
    - o **Access Control:**
        - ▪ Checks if the user is authenticated.
    - o **Description:** Renders the EmployeeCourseList component if the user is authenticated. Redirects to the login page if not.
- **All Courses Route** (/allCourses)
    - o **Component:** AllCourses
    - o **Access Control:**
        - ▪ Checks if the user is authenticated.
    - o **Description:** Renders the AllCourses component if the user is authenticated. Redirects to the login page if not.

- Dashboard Route (/dashboard)
    - **Component:** Dashboard
    - **Access Control:**
        - Checks if the user is authenticated.
        - Validates user role (must be admin).
    - **Description:** Renders the Dashboard component for admin users. Redirects to /not-found if the user does not have admin access or to the login page if not authenticated.
- Profile Route (/profile)
    - **Component:** Profile
    - **Access Control:**
        - Checks if the user is authenticated.
    - **Description:** Renders the Profile component if the user is authenticated. Redirects to the login page if not.
- Not Found Route (*)
    - **Component:** NotFound
    - **Access Control:**
        - None (accessible to all users).
    - **Description:** Renders the NotFound component for any routes that do not match the defined paths.

3.3   Database Schema
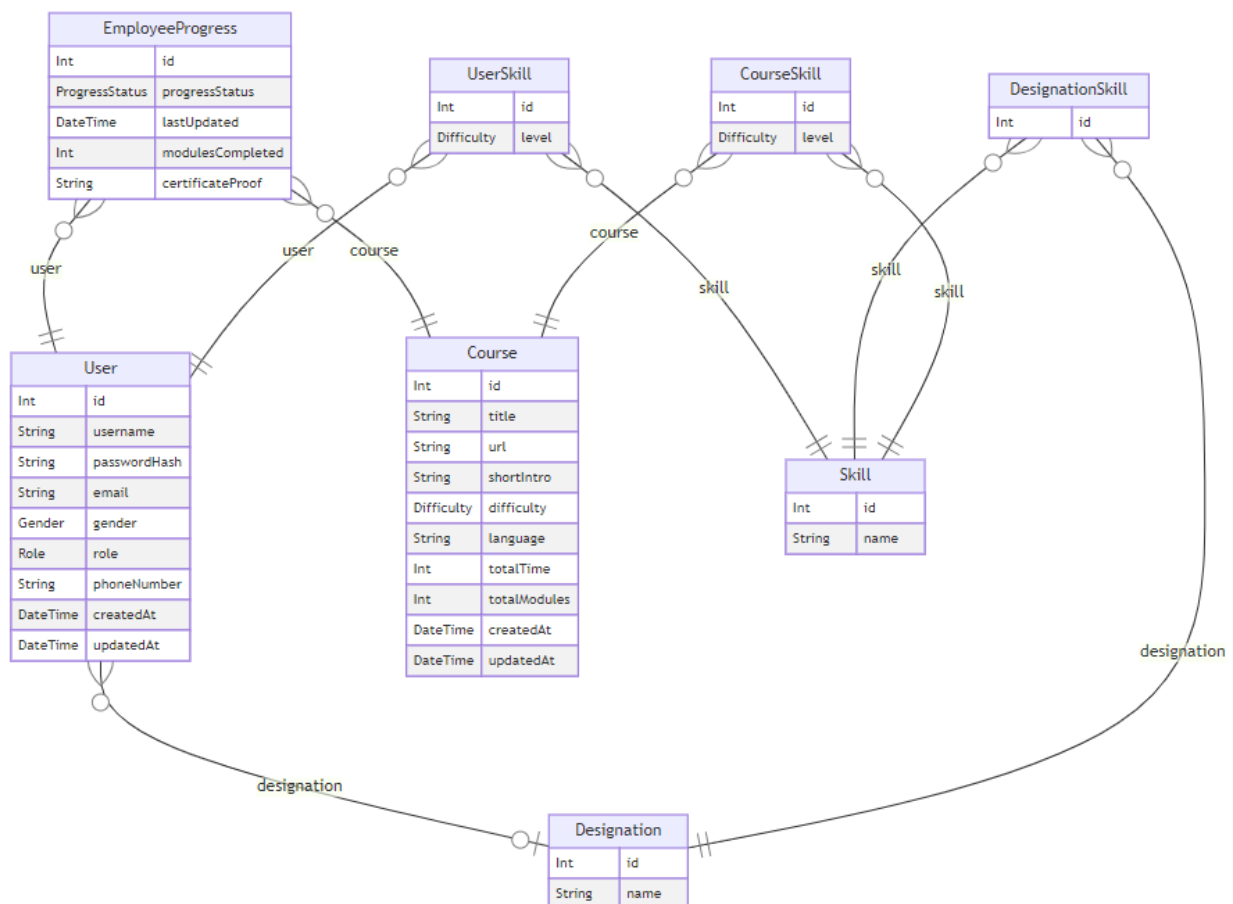
FIG:15 DATABASE SCHEMA

3.4 APIs

Method	Endpoint	Description	Used In
POST	/api/auth/register	Registers a new user.	Signup.jsx
POST	/api/auth/login	Authenticates a user and logs them in.	Login.jsx
GET	/api/skills-and-designations/designations	Retrieves the list of designations.	Signup.jsx
GET	/api/skills-and-designations/skills	Retrieves the list of skills.	AllCourses.jsx
GET	/api/users/skills	Retrieves the skills of the authenticated user.	SkillSet.jsx
GET	/api/users/other-skills	Retrieves other skills of the authenticated user.	AddUserSkillModel.jsx
POST	/api/users/add-skills	Adds new skills for the authenticated user.	SkillSet.jsx
GET	/api/users	Retrieves a list of all users.	AssignCourseModel.jsx
GET	/api/users/userInfo	Retrieves information about the authenticated user.	Profile.jsx
PUT	/api/users/edit/userInfo	Edits information for the authenticated user.	Profile.jsx
POST	/api/users/add-employee-progress	Adds progress for an employee in a specific course.	EmployeeCourseList.jsx
GET	/api/users/employee-progress/:courseId	Retrieves progress for a specific course for the authenticated user.	CourseDetail.jsx
GET	/api/courses/:courseId/skills	Retrieves the skills associated with a specific course.	CourseDetail.jsx
GET	/api/courses/available/:id	Retrieves available courses for the user.	AssignCourseModel.jsx
GET	/api/courses/user/progress	Retrieves the courses that the user is enrolled in, along with their progress.	EmployeeCourseList.jsx
GET	/api/courses/:courseId/details	Retrieves detailed information about a specific course.	CourseDetail.jsx
GET	/api/courses	Retrieves a list of all available courses.	AllCourses.jsx
GET	/api/courses/valid-ids	Retrieves valid course IDs for the user.	ProtectRoutes.jsx

3.5 Roles and access control

3.5.1 Admin

Description:

Admins are responsible for managing the application, overseeing employee progress, and assigning courses to the employees.

Access Rights:

- Full access to the Admin Dashboard, enabling comprehensive monitoring of performance metrics and insights.
- Authority to view all courses.
- Access to the Employee Progress Module for the purpose of tracking course completions and evaluating employee performance.
- Access to analytics and reporting tools that facilitate data-driven decision-making.

3.5.2 Employee

Description:

Employees utilize the platform to complete their assigned courses and monitor their progress.

Access Rights:

- Access to the View Courses Module, which displays only their assigned courses, organized by status (Completed, In Progress, Not Started).
- Ability to view detailed information about each course, including course descriptions and progress tracking features.
- Permissions to manage their personal profile and skills, including the ability to update personal information (except for designation and email).
- Access to the Manage Skills Module, allowing them to view and add skills based on completed courses.

4 Data Engineering

The Data Engineering process is structured into distinct layers, each with specific functions in preparing data for analysis and reporting.

4.1 Layers

4.1.1 Bronze Layer (Staging)

The Bronze Layer serves as the initial stage for data ingestion, where raw, unprocessed data is collected from the Postgres database. This layer stores the data without any transformations, ensuring a complete and accurate representation of the original data, which is essential for subsequent processing.

Key Features:

- **Data Source:** Postgres Database
- **Storage Format:** CSV Files
- **Purpose:** To maintain raw data integrity for future processing.

4.1.2   Silver Layer (Preparation)

In the Silver Layer, the data undergoes essential cleaning and transformation processes. This stage is critical for enhancing the quality and reliability of the data before it is moved to the next layer. This stage focuses on:

- Identifying and addressing missing values and duplicates.
- Defining the required columns for each dataset.
- Performing data preprocessing, which includes removing null values, selecting necessary columns, and ensuring consistent data types.

Key Features:

- **Data Source:** Raw Layer
- **Storage Format:** CSV Files
- **Purpose:** To prepare data for structured analysis and ensure its quality.

4.1.3   Gold Layer (Data Mart)

The Gold Layer organizes the clean data into structured tables designed for efficient querying and analysis. This layer includes both dimension and fact tables, facilitating detailed insights and reporting.

1. Dimension Tables

- **User Dimension:** Captures user information such as ID, username, and creation date.
- **Designation Dimension:** Contains details about job titles.
- **Course Dimension:** Details about courses, including course ID, title, difficulty level, language, total time, total modules, and creation date.
- **Skill Dimension:** Lists skill identifiers and names.

2. Fact Tables

- **Employee Progress Fact:** Tracks user progress in courses, including status, last update, modules completed, and certificate status.
- **Course Enrolment Fact:** Records the count of enrolments and completions for each course.
- **Skill Fact:** Connects users to their skills, including fulfilment status.

Key Features:

- **Data Source:** Prep Layer
- **Storage Format:** CSV Files
- **Purpose:** To structure data for easy access and analysis, supporting reporting needs.

4.1.4   Platinum Layer (Report)

The Platinum Layer further aggregates and organizes data to facilitate effective reporting. This layer focuses on preparing data for visualization in reporting tools, such as Power BI, ensuring that all relevant metrics are easily accessible. (Refer appendix 6.1)

Key Features:

- **Data Source:** Prep Layer
- **Storage Format:** CSV Files
- **Purpose:** To present key performance indicators and insights clearly.

4.2   Key Performance Indicators

KPI	Description	Visualization
Employee Progress Distribution by Designation	Visualizes the status of employees (Completed, In Progress, Not Started) across different designations.	Stacked Bar Chart
Course Completion Rate by Difficulty	Measures the percentage of courses completed, categorized by difficulty levels (Easy, Medium, Hard).	Donut Chart
Average Modules Completed Over Time by Designation	Tracks the trend of average modules completed over time, segmented by designation.	Line Chart
Employees Behind on Course for More Than 90 Days	Identifies employees who are significantly behind in their courses, highlighting those needing support.	Table
Top 7 Courses by Enrolment	Lists the seven most enrolled courses, providing insight into popular training offerings.	Stacked Vertical Bar Chart
Enrolment vs Completion	Compares the count of enrolments against the count of completions for courses.	Gauge Chart
Skill Unfulfillment by Designation	Shows the number of individuals with unfulfilled skills in each designation, highlighting areas for development.	Table

4.3   Power BI Integration

Power BI serves as a crucial tool for visualizing and reporting data processed through the data engineering layers. This subsection outlines how Power BI integrates with the data layers, enabling dynamic reporting and insightful analytics.

4.3.1   Data Model Overview

In Power BI, the data model is constructed by connecting structured data from the Gold Layer. This model facilitates seamless relationships between various dimension and fact tables, enhancing the interactivity of the reports. Key components of the data model include:

- **User Dimension:** Captures user information such as ID, username, and creation date.
- **Designation Dimension:** Contains details about job titles.
- **Course Dimension:** Details about courses, including course ID, title, difficulty level, and other attributes.
- **Skill Dimension:** Lists skill identifiers and names.
- **Calendar Table:** Provides date-related attributes for time-based analysis, such as year, month, and quarter.
- **Employee Progress Fact:** Tracks user progress in courses, including status, last update, and modules completed.
- **Course Enrolment Fact:** Records the count of enrolments and completions for each course.
- **Skill Fact:** Connects users to their skills, including fulfilment status.
- **Latest Progress Records Table:** Stores the most recent progress updates for each user, enabling quick access to latest status.
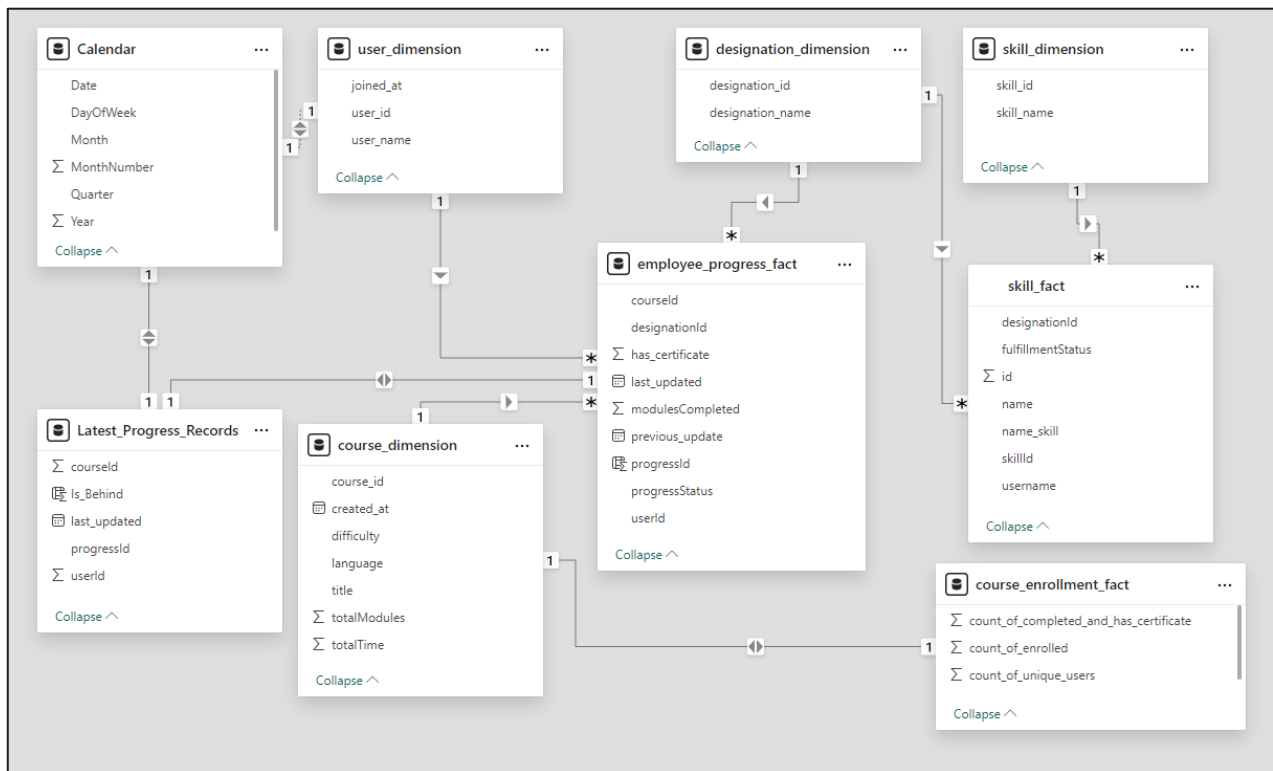
FIG:16 MODEL VIEW

4.3.2 Dashboard Design

The Power BI dashboard is designed to present the key performance indicators (KPIs) effectively. Each KPI is carefully crafted to provide insights into employee progress, course completion, and skill development. The dashboard facilitates interactive visualizations, allowing users to filter and explore data based on various dimensions. This design enhances decision-making by providing clear, actionable insights for stakeholders.
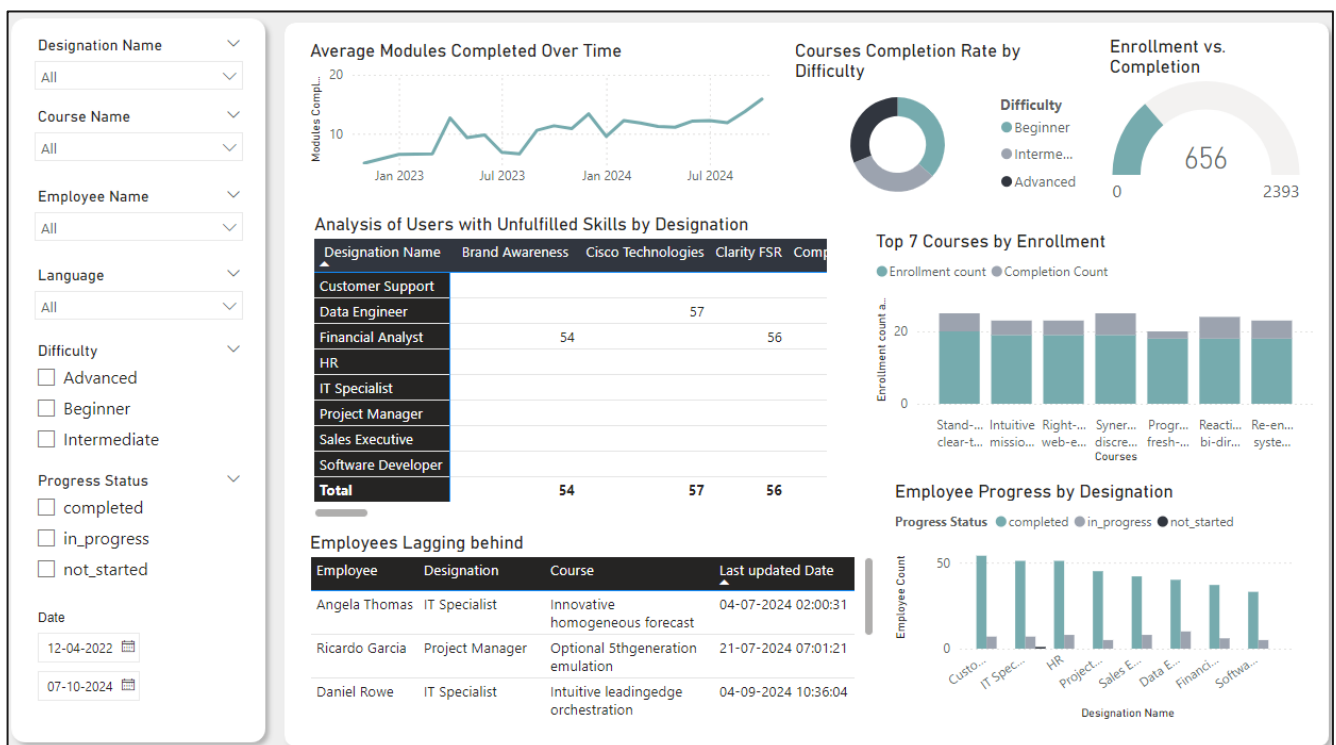
FIG:17 ADMIN DASHBOARD

5 Data Science

5.1 Exploratory Data Analysis (EDA)

1. Distribution of Course Completion Status

- Most users are "in progress" (1735), with a significant number "completed" (656), and very few "not started" (2).
- This indicates that while many users engage with the courses, a notable number are completing them.
- This could influence how we prioritize recommending courses—suggesting additional courses to users who are already completing their current ones could increase overall engagement.
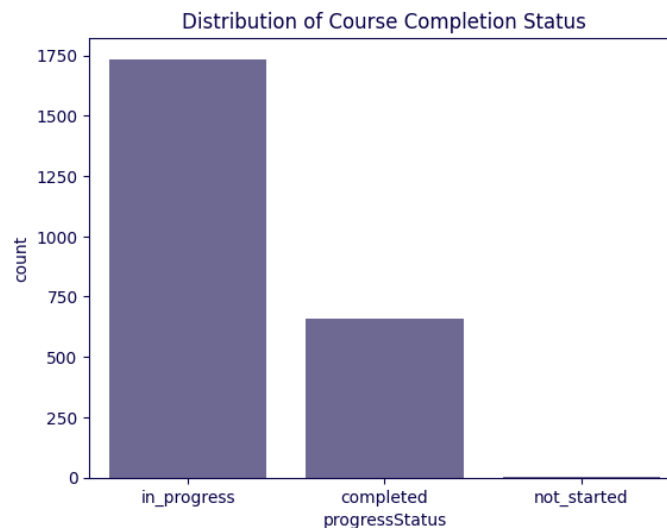
FIG:18 DISTRIBUTION OF COURSE COMPLETION STATUS

2. Course Difficulty Levels

- The distribution shows a balanced interest across different difficulty levels: Beginner (876), Intermediate (752), and Advanced (765).
- This suggests that users are open to courses of varying complexities.
- Feature engineering could involve categorizing users based on their preferred difficulty level, tailoring recommendations to match their experience and confidence.
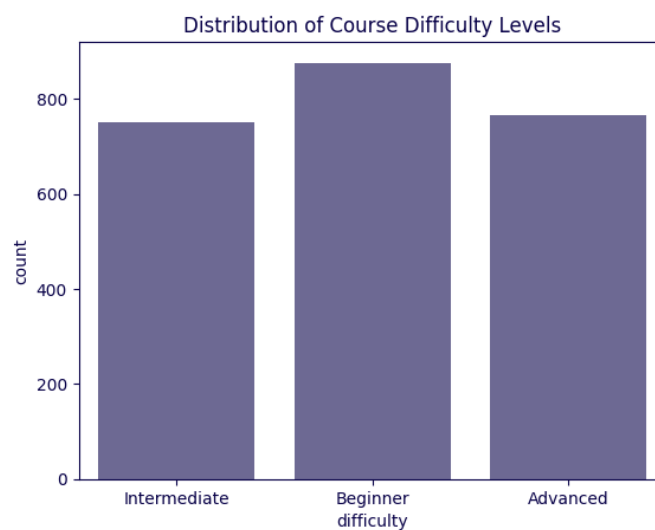
FIG:19 COURSE DIFFICULTY LEVELS

3. Progress Percentage Distribution

- The progress percentage shows a concentration of users nearing completion (e.g., 783 users in the highest progress range).
- This indicates potential for completion incentives or targeted communications.
- For feature engineering, courses could be recommended based on how close users are to completing their current courses, encouraging them to finish and then move on to related or advanced topics.
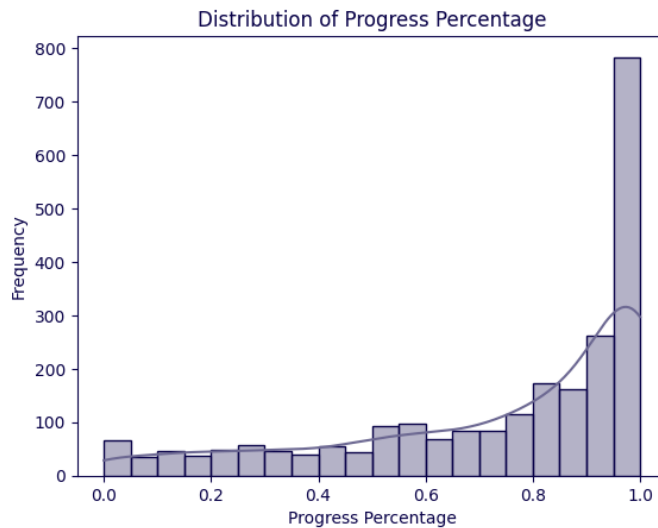
FIG:20 PROGRESS PERCENTAGE DISTRIBUTION

4. Top Skills Analysis

- Skills such as 'Oracle ERP', 'E-on Vue', and 'GSA Schedule' are among the most common.
- This insight can inform the development of targeted courses that build on these skills.
- Users could be recommended courses that enhance their existing skills or offer complementary knowledge, thus providing a more personalized learning pathway.


FIG:21 TOP SKILLS ANALYSIS

- Progress percentages vary across designations, with Customer Support showing the highest average progress (0.76) and Data Engineers the lowest (0.71).
- This suggests different engagement levels based on job roles.
- Feature engineering could involve creating user segments based on designation, allowing for tailored course recommendations that align with career development paths specific to each role.
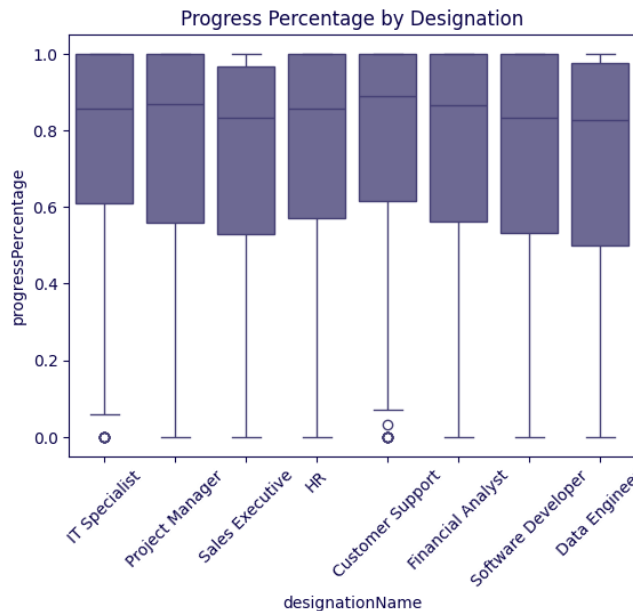
FIG:22 PROGRESS PERCENTAGE BY DESIGNATION

5.2   Feature Engineering and Selection

- Data Preprocessing:
  - Convert string representations of lists in userSkills and courseSkills back into actual lists using ast.literal_eval.
- Skills Vector Creation:
  - Generate a textual representation of skills for each course by joining skills into a single string, stored in the skills_vector column.
- Feature Selection:
  - Selected Features:
    - userSkills: Skills possessed by the employee.
    - courseSkills: Skills required for each course.
    - skills_vector: Combined string of course skills for TF-IDF representation.
- TF-IDF Vectorization:
  - Apply TfidfVectorizer to transform the skills_vector into a sparse matrix format, capturing the importance of skills relative to all courses.

5.3   Model Selection

In the context of course recommendation systems, selecting the appropriate model is critical for achieving optimal performance. The model should effectively capture the relationships between user skills and available courses to provide personalized recommendations. For this purpose, a content-based filtering approach was

chosen, utilizing TF-IDF (Term Frequency-Inverse Document Frequency) vectorization and cosine similarity to assess the relevance of courses based on employee skills.

- Content-Based Filtering
  - **Personalized Recommendations:** Aligns course suggestions with the specific skills of each employee, ensuring relevance.
  - **Focus on Individual Growth:** Encourages learning based on what employees already know, facilitating targeted skill development.
  - **No Need for External Data:** Operates independently of other users' preferences, making it effective in environments with limited interaction data.
- TF-IDF Vectorization
  - **Quantitative Analysis:** Provides a numerical representation of skills, enabling efficient comparison and analysis.
  - **Highlighting Relevant Skills:** Identifies and emphasizes key skills within courses, improving the accuracy of recommendations.
  - **Flexibility:** Can easily accommodate new skills or courses by updating the dataset without major overhauls. (Refer appendix 6.2)
- Cosine Similarity
  - **Effective Similarity Measurement:** Accurately gauges the degree of similarity between high-dimensional vectors, enhancing the relevance of recommendations.
  - **Normalization:** Handles varying lengths of skill vectors, allowing for fair comparisons regardless of skill set size.
  - **Intuitive Interpretation:** The resulting similarity scores are easy to understand, facilitating the explanation of recommendations to employees. (Refer appendix 6.3)

5.4   Sample outputs

Employee ID: **2**

Course ID	Course Name	Similarity Score
327	Vision-oriented asymmetric structure	0.460461
299	Sharable attitude-oriented definition	0.435657
93	Exclusive discrete orchestration	0.337262
14	Synchronized dedicated parallelism	0.264926
68	Reverse-engineered grid-enabled knowledgebase	0.264630

Employee ID: **125**

Course ID	Course Name	Similarity Score
350	Integrated bifurcated matrices	0.384353
133	Customizable executive methodology	0.383483
202	Customizable intermediate challenge	0.352091
318	Reactive bi-directional conglomeration	0.345422
298	Monitored content-based contingency	0.339000

6 Appendix

6.1 Sample report table

employee Id	courseId	employeeNa me	designati onId	designati onName	courseN ame	difficulty	language	totalTime	totalMod ules	progress Status	lastUpda ted	modulesC ompleted	userSkills	courseSkills	designati onSkills
2	331	Allison Chan	5	IT Speciali	Programn	Intermedi	English	480	10	in_progre:	59:25.0	6	['DH+', 'Life	['MBAL', 'Pres	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
2	331	Allison Chan	5	IT Speciali	Programn	Intermedi	English	480	10	completec	51:29.0	10	['DH+', 'Life	['MBAL', 'Pres	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
2	119	Allison Chan	5	IT Speciali	Secured ir	Beginner	Tamil	273	8	in_progre:	08:39.0	6	['DH+', 'Life	['HP Procurve	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
2	119	Allison Chan	5	IT Speciali	Secured ir	Beginner	Tamil	273	8	completec	29:20.0	8	['DH+', 'Life	['HP Procurve	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
3	367	Scott Flores	7	Project Ma	Up-sized r	Advanced	German	79	10	in_progre:	23:32.0	9	['Rhino 3D',	['Winning Oth	['Karaoke', 'SMO', 'Educational Leadership', 'Tax Accounting', 'PCRF', 'VT', 'Leadership', 'WTP']
3	367	Scott Flores	7	Project Ma	Up-sized r	Advanced	German	79	10	completec	27:54.0	10	['Rhino 3D',	['Winning Oth	['Karaoke', 'SMO', 'Educational Leadership', 'Tax Accounting', 'PCRF', 'VT', 'Leadership', 'WTP']
3	341	Scott Flores	7	Project Ma	Cloned ra	Advanced	Spanish	221	23	in_progre:	41:02.0	19	['Rhino 3D',	['Tissue Cultu	['Karaoke', 'SMO', 'Educational Leadership', 'Tax Accounting', 'PCRF', 'VT', 'leadership', 'WTP']
3	341	Scott Flores	7	Project Ma	Cloned ra	Advanced	Spanish	221	23	in_progre:	26:14.0	20	['Rhino 3D',	['Tissue Cultu	['Karaoke', 'SMO', 'Educational Leadership', 'Tax Accounting', 'PCRF', 'VT', 'leadership', 'WTP']
3	341	Scott Flores	7	Project Ma	Cloned ra	Advanced	Spanish	221	23	in_progre:	17:57.0	22	['Rhino 3D',	['Tissue Cultu	['Karaoke', 'SMO', 'Educational Leadership', 'Tax Accounting', 'PCRF', 'VT', 'leadership', 'WTP']
3	341	Scott Flores	7	Project Ma	Cloned ra	Advanced	Spanish	221	23	completec	01:11.0	23	['Rhino 3D',	['Tissue Cultu	['Karaoke', 'SMO', 'Educational Leadership', 'Tax Accounting', 'PCRF', 'VT', 'leadership', 'WTP']
4	357	Frank Trujillo	7	Project Ma	Pre-empti	Advanced	German	191	7	in_progre:	45:05.0	3	['FFA', 'Sage	['MD-11', 'Goc	['Karaoke', 'SMO', 'Educational Leadership', 'Tax Accounting', 'PCRF', 'VT', 'Leadership', 'WTP']
4	357	Frank Trujillo	7	Project Ma	Pre-empti	Advanced	German	191	7	in_progre:	42:32.0	5	['FFA', 'Sage	['MD-11', 'Goc	['Karaoke', 'SMO', 'Educational Leadership', 'Tax Accounting', 'PCRF', 'VT', 'Leadership', 'WTP']
4	357	Frank Trujillo	7	Project Ma	Pre-empti	Advanced	German	191	7	completec	12:47.0	7	['FFA', 'Sage	['MD-11', 'Goc	['Karaoke', 'SMO', 'Educational Leadership', 'Tax Accounting', 'PCRF', 'VT', 'Leadership', 'WTP']
5	110	Dana Gonzale	5	IT Speciali	Universal	Beginner	Japanese	363	14	in_progre:	03:57.0	8	['NSI', 'JCL',	['Two-phase	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
5	110	Dana Gonzale	5	IT Speciali	Universal	Beginner	Japanese	363	14	in_progre:	48:57.0	9	['NSI', 'JCL',	['Two-phase	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
5	110	Dana Gonzale	5	IT Speciali	Universal	Beginner	Japanese	363	14	in_progre:	18:54.0	13	['NSI', 'JCL',	['Two-phase	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
5	110	Dana Gonzale	5	IT Speciali	Universal	Beginner	Japanese	363	14	completec	27:27.0	14	['NSI', 'JCL',	['Two-phase	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
5	231	Dana Gonzale	5	IT Speciali	Distribute	Advanced	German	213	14	in_progre:	21:11.0	3	['NSI', 'JCL',	['CGI']	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
5	231	Dana Gonzale	5	IT Speciali	Distribute	Advanced	German	213	14	in_progre:	42:08.0	6	['NSI', 'JCL',	['CGI']	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
5	231	Dana Gonzale	5	IT Speciali	Distribute	Advanced	German	213	14	in_progre:	35:20.0	9	['NSI', 'JCL',	['CGI']	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
5	231	Dana Gonzale	5	IT Speciali	Distribute	Advanced	German	213	14	completec	39:38.0	14	['NSI', 'JCL',	['CGI']	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
5	1	Dana Gonzale	5	IT Speciali	Total solu	Intermedi	Japanese	379	6	not_starte	22:24.7	0	['NSI', 'JCL',	['DTH']	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
5	3	Dana Gonzale	5	IT Speciali	Fundame	Beginner	English	354	13	not_starte	22:44.1	0	['NSI', 'JCL',	['BGP', 'Aquat	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
5	1	Dana Gonzale	5	IT Speciali	Total solu	Intermedi	Japanese	379	6	in_progre:	22:57.0	3	['NSI', 'JCL',	['DTH']	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']
6	381	Scott Holt	5	IT Speciali	Inverse 24	Beginner	Hindi	33	14	in_progre:	13:34.0	1	['Oracle ERI	['EP', 'Sage 30	['Network Architecture', 'SQF', 'SRDS', 'System Testing', 'Service Delivery', 'Portfolio Management', 'ELISA']

6.2 TF-IDF Vector

Definition: TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (or corpus). It helps in converting text data into a numerical representation that can be used for various machine learning and information retrieval tasks.

Components:

- Term Frequency (TF): Measures how frequently a term appears in a document. The formula is:

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

- Inverse Document Frequency (IDF): Measures how important a term is across the entire corpus. It is calculated as:

$$IDF(t) = log\frac{N}{1+df}$$

where N is the total number of documents in the corpus.

TF-IDF Calculation: The TF-IDF score for a term t in document d is computed by multiplying TF and IDF:

$$TF - IDF(t,d) = TF(t,d) * IDF(t)$$

This results in a vector representation for each document, capturing the significance of terms while reducing the weight of commonly used words.

Applications:

- Text classification
- Information retrieval
- Document clustering
- Feature extraction for machine learning models

6.3   Cosine Similarity

Definition: Cosine similarity is a metric used to measure how similar two vectors are, regardless of their magnitude. It calculates the cosine of the angle between two non-zero vectors in an inner product space, resulting in a value between -1 and 1.

Formula: The cosine similarity between two vectors A and B is calculated as:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\|\|B\|}$$

Where:

- A·B is the dot product of the vectors.
- ‖A‖ and ‖B‖ are the magnitudes (or norms) of the vectors.

Interpretation:

- A cosine similarity of 1 indicates that the two vectors are identical in direction.
- A cosine similarity of 0 indicates that the two vectors are orthogonal (i.e., completely dissimilar).
- A cosine similarity of -1 indicates that the two vectors are diametrically opposed.

Applications:

- Document similarity assessment
- Recommendation systems
- Text clustering and classification
- Natural language processing tasks