**Database Systems Project Part I – Modelling**

**Course Title:** Database Systems **Course Number:** CSCI-GA.2433-001

**Instructor:** Jean-Claude Franchitti

**Team Members:**

|  |  |  |
| --- | --- | --- |
| **Name:** Dedeepya Guntaka  **Net ID:** dg4489  **NYU ID:** N18304471 | **Name:** Shivam Dheer  **Net ID:** sd6080 **NYU ID:** N18103074 | **Name:** Varun Putta  **Net ID:** vsp7221  **NYU ID:** N10489122 |

**Database Management Systems Project Specification - Part I - Data Modeling**

***Project Objective:*** *Designing an Enterprise Data Architecture for Integrated, Scalable, and Governed Data Management to Predict and Reduce Employee Attrition.*

**Problem Statement**

Employee attrition is a significant challenge, affecting productivity, morale, and costs. This project aims to design a comprehensive EDA that enables efficient data management across traditional storage systems and supports predictive analytics to identify high-risk employees likely to leave. The company can tailor interventions that improve employee retention, optimise workplace conditions, and mitigate attrition-related costs by leveraging a structured data environment.

**Objectives**

**1. Develop a Centralised Data Model for Employee Insights and Attrition Prediction**

**Goal:** Establish a cohesive enterprise data model that captures and structures employee-related data across various dimensions, including job satisfaction, compensation, performance, and work patterns, enabling predictive analytics for attrition.

* **Approach:** Create an Entity-Relationship Diagram (ERD) capturing critical entities such as Employee, Compensation, Job Satisfaction, Work Patterns, and Performance, providing a single source of truth for employee data.

**2. Implement Data Governance Standards for Attrition Prediction and Insights**

**Goal:** Define data retention and quality standards specific to employee data, ensuring accurate, bias-free, and secure data for predictive modelling.

* **Approach:** Incorporate policies for data ageing, metadata management, and data quality assurance to support predictive analytics and model reliability while safeguarding sensitive employee data.

**3. Enable Efficient Data Integration and Analysis through an Operational Data Store (ODS)**

**Goal:** Develop an ODS to manage real-time employee data, allowing for daily snapshots of changes in critical factors affecting job satisfaction and attrition risk.

* **Approach:** Integrate the ODS with existing relational databases, providing a central resource for analysing employee metrics in real-time enabling dynamic updates for predictive modelling.

**4. Optimize Reporting and Analytics with Data Warehousing and Data Marts for Employee Insights**

**Goal:** Design a Data Warehouse (DW) and Data Marts (DM) to store historical employee data and support in-depth analytics, tracking changes in job satisfaction, performance, and retention trends over time.

* **Approach:** Implement ETL processes to cleanse and transform data, making it consistent and reliable for attrition analysis and reporting, with separate marts for Compensation, Performance, and Work Patterns.

**5. Integrate Predictive Analytics for Identifying High Attrition Risk and Personalized Interventions**

**Goal:** Develop predictive models within the EDA framework that identify employees at high risk of attrition based on historical patterns and provide actionable insights for managers.

* **Approach:** Apply machine learning algorithms to analyse critical factors (e.g., job satisfaction, travel demands, compensation) and predict attrition risk, allowing for interventions such as role reassignments, remote work options, or customised career growth plans.

**System Requirements and Dependencies**

To build a scalable, secure, and real-time Enterprise Data Architecture (EDA) for attrition prediction, we would implement the following system requirements and dependencies:

1. **Database Management Systems (DBMS)**

* Relational Database: SQL Server or MySQL for transactional data, managing employee information and supporting high data volumes with quick retrieval.
* Data Warehouse: Amazon Redshift, Google BigQuery, or Snowflake to store historical data, enabling trend analysis and scalability.
* Operational Data Store (ODS): Azure Synapse or PostgreSQL for real-time updates, ensuring a central source of dynamic employee metrics.

1. **ETL Tools and Data Integration**

* ETL Tools: Apache Nifi, Talend, or Azure Data Factory for data extraction, transformation, and load processes from diverse sources, maintaining data consistency.
* Real-Time Integration: Apache Kafka or Amazon Kinesis to support continuous data streaming for timely data synchronization.

1. **Machine Learning and Analytics**

* Programming Language: Python, chosen for its libraries and machine learning capabilities, ensuring seamless model building and integration.
* Machine Learning Libraries: Scikit-learn, TensorFlow, or PyTorch for predictive modeling, providing flexibility in algorithm selection.
* Visualization Tools: Tableau or Power BI for user-friendly dashboards that monitor attrition trends and provide insights.

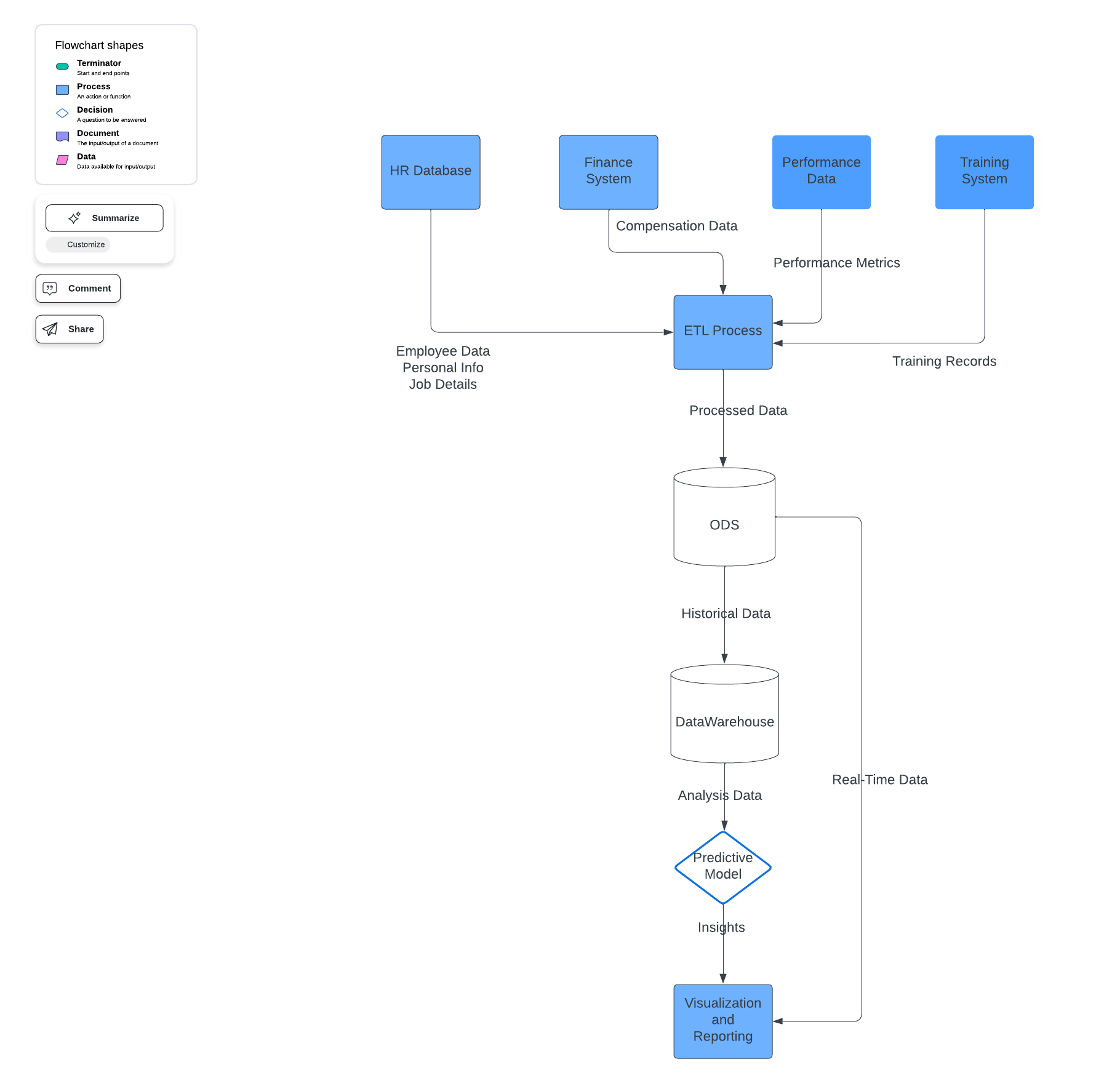
1. **Data Governance and Security**

* Metadata Management: Apache Atlas or Informatica for consistent data tracking and lineage, ensuring structured documentation.
* Data Quality and Compliance: Talend Data Quality or Microsoft Purview is used for quality checks, privacy, and GDPR compliance.
* Access Control and Security: Securing sensitive data using role-based access control (RBAC) and encryption via Azure Active Directory or AWS IAM.

**Dependencies**

* HR Data Access: Integration with HR systems for core employee metrics.
* API Integrations: Real-time data access from operational databases (e.g., attendance, payroll).
* IT and Infrastructure Support: Support for setup, maintenance, and cloud or on-premises configurations based on privacy needs.

**Data Flow Diagram**



**Conceptual Model**

The model will inform job satisfaction, work-life balance, department, travel demands, and compensation factors. Entities will include:

1. **Employee:** Captures demographic and employment details (e.g., Age, Gender, Department).
2. **Job Satisfaction:** Stores satisfaction metrics (e.g., Job Satisfaction, Work-Life Balance).
3. **Compensation:** Contains income details (e.g., Monthly Income, Salary Hike).
4. **Work Patterns:** Tracks travel and location data (e.g., Business Travel, Distance from Home).
5. **Performance:** Holds performance metrics (e.g., Performance Rating, Total Working Years).

**ER Diagram:**

### 

### **Entities and Attributes:**

* **PERSONAL INFO**:

Attributes: EmployeeNumber (PK, FK), Age, EducationField, Education, RelationshipSatisfaction.

* + This entity holds personal information about an employee, including details like age, education, and relationship satisfaction.
* **MANAGER**:

Attributes: ManagerID (PK), EmployeeNumber (FK), ManagementLevel, TeamSize, ResponsibilityArea.

* + This entity represents managers within the company, associating them with an employee (likely to identify which employee is a manager) and detailing their managerial attributes.
* **COMPENSATION**:

Attributes: EmployeeNumber (PK, FK), DailyRate, HourlyRate, MonthlyRate, MonthlyIncome, StockOptionLevel, Benefits, Bonus.

* + Stores financial and benefit details about employees, such as pay rates, income, benefits, and bonuses.
* **EMPLOYMENT HISTORY**:

Attributes: HistoryID (PK), EmployeeNumber (FK), NumCompaniesWorked, TotalWorkingYears, PreviousCompany, StartDate, EndDate.

* + Tracks each employee's employment history, including the number of companies they have worked for, total working years, previous companies, and dates of employment.
* **JOB DETAILS**:

Attributes: EmployeeNumber (PK), JobRole, JobLevel, JobInvolvement, JobSatisfaction, EnvironmentSatisfaction, BusinessTravel, OverTime, YearsInCurrentRole, YearsSinceLastPromotion, YearsAtCompany, YearsWithCurrManager.

* + Contains detailed information about the job, such as role, level, involvement, satisfaction, business travel requirements, and tenure-related attributes.
* **PERFORMANCE**:

Attributes: PerformanceID (PK), EmployeeNumber (FK), PerformanceRating, PercentSalaryHike, ReviewDate, Feedback, Goals.

* + Tracks the performance metrics of employees, including ratings, salary hikes, and performance feedback.
* **EMPLOYEE**:

Attributes: EmployeeNumber (PK), Over18, Gender, MaritalStatus, DistanceFromHome, StandardHours, WorkLifeBalance.

* + This central entity represents the employee's core details, including personal information like age, gender, marital status, and work-life balance.
* **DEPARTMENT**:

Attributes: DepartmentID (PK), Department, EmployeeCount, Location, Budget.

* + Represents various departments within the company, detailing the department name, employee count, location, and budget.
* **PROJECT**:

Attributes: ProjectID (PK), DepartmentID (FK), ProjectName, StartDate, EndDate, Status, Priority.

* + Tracks information about departments' projects, including project timelines, status, and priority levels.
* **TRAINING**:

Attributes: TrainingID (PK), EmployeeNumber (FK), TrainingName, TrainingTimesLastYear, CompletionDate, Status, Certification.

* + It contains training-related data for employees, detailing the types of training they attend, frequency, completion status, and certification.

### 

### **Relationships**

* **PERSONAL INFO - EMPLOYEE**:

One-to-One (EmployeeNumber as FK in both entities).

* + Each employee has one personal information record associated with them.
* **MANAGER - EMPLOYEE**:

One-to-Many (one manager supervises multiple employees).

* + Managers supervise multiple employees, establishing a hierarchical structure within the organisation.
* **EMPLOYEE - COMPENSATION**:

One-to-One (EmployeeNumber as FK).

* + Each employee receives specific compensation details.
* **EMPLOYEE - EMPLOYMENT HISTORY**:

One-to-Many.

* + An employee can have multiple employment history records, tracking changes over time.
* **EMPLOYEE - JOB DETAILS**:

One-to-One (EmployeeNumber as FK).

* + Each employee has specific job details related to their role, level, and job satisfaction.
* **EMPLOYEE - PERFORMANCE**:

One-to-Many (EmployeeNumber as FK).

* + An employee can have multiple performance records, reflecting evaluations over time.
* **EMPLOYEE - DEPARTMENT**:

Many-to-One (DepartmentID as FK in EMPLOYEE).

* + Each employee belongs to one department, while a department may have multiple employees.
* **DEPARTMENT - PROJECT**:

One-to-Many (DepartmentID as FK in PROJECT).

* + Each department manages multiple projects.
* **DEPARTMENT - EMPLOYEE** (manages):

One-to-Many.

* + The department manages employees, likely indicating that a department is responsible for managing a subset of employees.
* **EMPLOYEE - TRAINING**:

One-to-Many (EmployeeNumber as FK in TRAINING).

* + An employee can attend multiple training sessions.

### 

### **Attributes of Note**

* **Primary Keys (PK):**

Each entity has a unique primary key to identify individual records (e.g., EmployeeNumber for EMPLOYEE, ManagerID for MANAGER).

* **Foreign Keys (FK):**

Many entities reference the EmployeeNumber as a foreign key to link data back to the EMPLOYEE entity.

**Key Deliverables**

1. Entity-Relationship Diagram (ERD): A detailed ERD is created for the EDA, capturing employee-related entities for attrition prediction and structured data management.
2. Architecture Documentation covers data model design, governance policies, and predictive model specifications.
3. Predictive Model and Report: A prototype model identifying high-risk employees with documentation on critical business insights and recommended interventions.

**Business Logic for Actionable Insights**

1. Identifying High Attrition Risk: Analyse patterns among past attrition cases (Attrition = Yes) to identify trends in job dissatisfaction, travel demands, or compensation issues.
2. Role Reassignment Recommendations: Recommend job changes based on role satisfaction data and retention trends.
3. Flexible Work Options: Identify high-risk employees with demanding travel or commute patterns and recommend remote or hybrid work options.
4. Personalised Growth Plans: Tailor career progression or training plans for employees with low satisfaction in their current roles or high commute demands.

**Impact and Insights**

* **Insights for Managers**

Managers can use this EDA's insights to proactively adjust high-risk employees' conditions, such as offering flexible work arrangements, optimising role assignments, or increasing benefits for high-turnover roles.

* **Long-term Impact**

By implementing this predictive EDA, the company can proactively reduce turnover, save on recruitment costs, retain experienced talent, and foster a healthier, more sustainable workplace culture.