**NYC Payroll Data Integration**

**Introduction**

The City of New York embarked on a project to integrate payroll data across all its agencies. The objective was to develop a Data Analytics platform to achieve two primary goals which include Financial Resource Allocation Analysis and Transparency/Public Accessibility.

**Project Scope**

The scope of the project encompassed the following

1. Designing a Data Warehouse for NYC
2. Developing a scalable and automated ETL Pipeline
3. Ensuring data quality and consistency
4. Creating a public user with limited privileges
5. Documenting processes for reproducibility
6. Maintaining a cloud-hosted repository of the codebase

**Data Warehouse Design**

The Schema design include creation of tables which entails : Employee, Agency, Title, DateDimension, PayrollTransactions

CREATE TABLE Employee (

EmployeeID INT PRIMARY KEY,

LastName VARCHAR(255),

FirstName VARCHAR(255),

);

CREATE TABLE Agency (

AgencyID INT PRIMARY KEY,

AgencyName VARCHAR(255),

);

CREATE TABLE Title (

TitleCode VARCHAR(255) PRIMARY KEY,

TitleDescription VARCHAR(255),

);

CREATE TABLE DateDimension (

DateID INT PRIMARY KEY,

Date DATE,

Month INT,

Year INT,

Quarter INT

);

CREATE TABLE PayrollTransactions (

TransactionID INT PRIMARY KEY,

EmployeeID INT,

AgencyID INT,

TitleCode VARCHAR(255),

DateID INT,

FiscalYear INT,

PayrollNumber INT,

BaseSalary DECIMAL(10, 2),

PayBasis VARCHAR(255),

RegularHours DECIMAL(10, 2),

RegularGrossPaid DECIMAL(10, 2),

OTHours DECIMAL(10, 2),

TotalOTPaid DECIMAL(10, 2),

TotalOtherPay DECIMAL(10, 2),

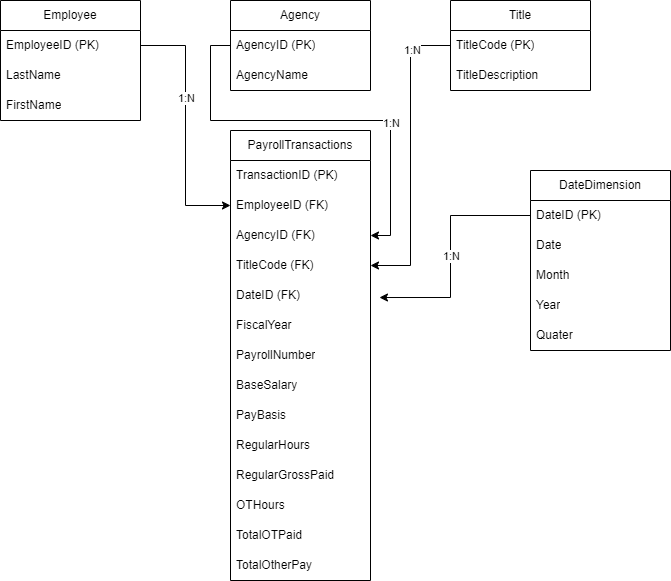
FOREIGN KEY (EmployeeID) REFERENCES Employee(EmployeeID),

FOREIGN KEY (AgencyID) REFERENCES Agency(AgencyID),

FOREIGN KEY (TitleCode) REFERENCES Title(TitleCode),

FOREIGN KEY (DateID) REFERENCES DateDimension(DateID)

**);**



CREATE TABLE Employee (

EmployeeID INT PRIMARY KEY,

LastName VARCHAR2(255),

FirstName VARCHAR2(255)

);

CREATE TABLE Agency (

AgencyID INT PRIMARY KEY,

AgencyName VARCHAR2(255)

);

CREATE TABLE Title (

TitleCode VARCHAR2(255) PRIMARY KEY,

TitleDescription VARCHAR2(255)

);

CREATE TABLE DateDimension (

DateID INT PRIMARY KEY,

TransactionDate DATE,

Month INT,

Year INT,

Quarter INT

);

CREATE TABLE PayrollTransactions (

TransactionID INT PRIMARY KEY,

EmployeeID INT,

AgencyID INT,

TitleCode VARCHAR2(255),

DateID INT,

FiscalYear INT,

PayrollNumber INT,

BaseSalary NUMBER(10, 2),

PayBasis VARCHAR2(255),

RegularHours NUMBER(10, 2),

RegularGrossPaid NUMBER(10, 2),

OTHours NUMBER(10, 2),

TotalOTPaid NUMBER(10, 2),

TotalOtherPay NUMBER(10, 2),

CONSTRAINT fk\_employee FOREIGN KEY (EmployeeID) REFERENCES Employee(EmployeeID),

CONSTRAINT fk\_agency FOREIGN KEY (AgencyID) REFERENCES Agency(AgencyID),

CONSTRAINT fk\_title FOREIGN KEY (TitleCode) REFERENCES Title(TitleCode),

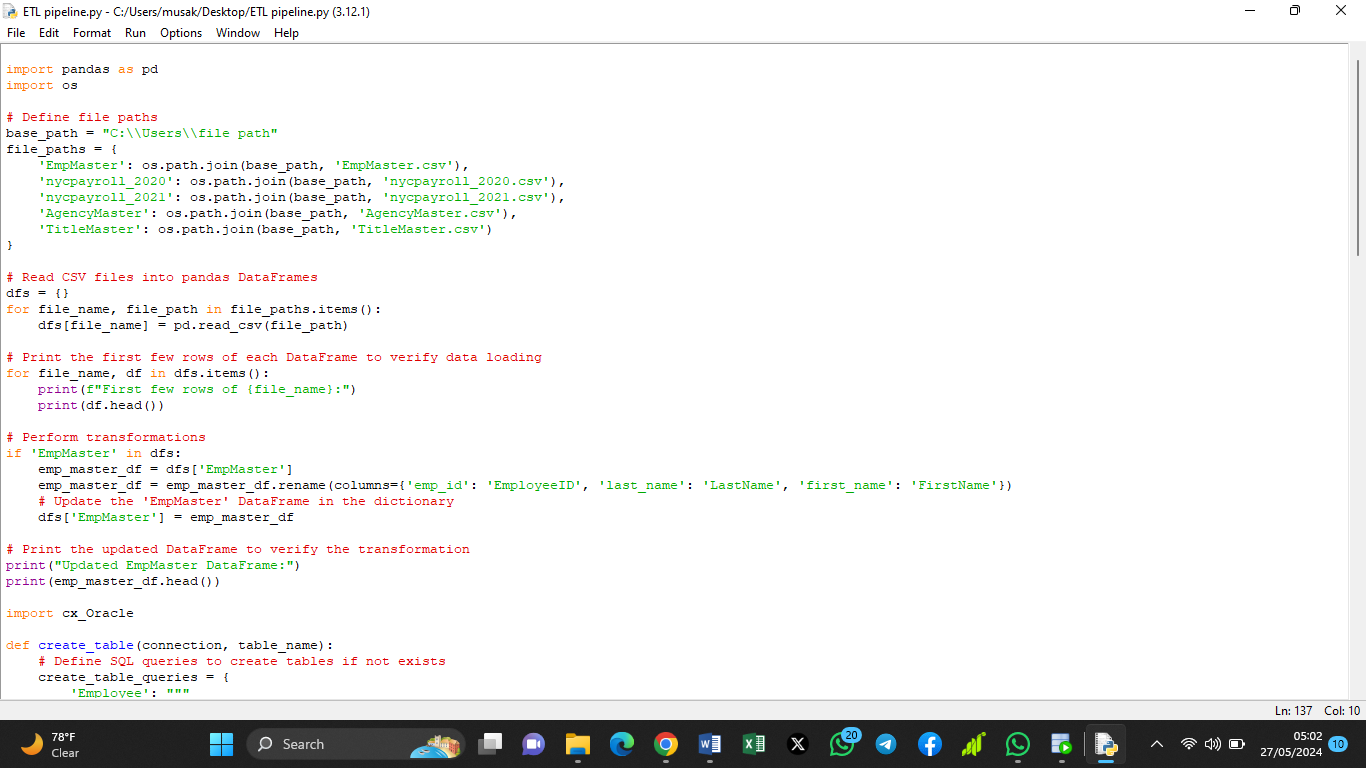
CONSTRAINT fk\_date FOREIGN KEY (DateID) REFERENCES DateDimension(DateID)

);

**Utilization of Oracle Database for storage**

*ETL Pipeline Development*

Python-based ETL pipeline pandas was used for data manipulation and cx\_Oracle for database interaction. It was also used for automated extraction, transformation, and loading of CSV data into Oracle database tables, error handling and logging for data quality assurance.



*Data Quality Assurance*

To ensure the integrity and quality of our data, we've implemented robust data validation checks throughout the ETL (Extract, Transform, Load) process. These checks meticulously validate the data against predefined rules, including data types, constraints, and business logic.

During the execution of the ETL pipeline, we meticulously handle errors and exceptions that may arise, ensuring smooth operation even in the face of unexpected issues. Our error handling mechanisms are designed to gracefully manage any errors encountered, allowing the pipeline to continue processing data without disruption.

In addition, we've implemented logging mechanisms to facilitate troubleshooting and error analysis. These logs provide detailed information about the execution of the ETL process, including any errors encountered, data transformations applied, and successful data loads. By maintaining comprehensive logs, we ensure transparency and facilitate efficient debugging when necessary.

*User Access Control*

To ensure secure and controlled access to our data, we've implemented a comprehensive system of roles and privileges within the Oracle Database. This system allows us to manage user permissions effectively, granting appropriate access levels to different types of users. As part of this strategy, we established a public user role with restricted access to sensitive data. This role is carefully configured to provide the necessary access to non-sensitive information while safeguarding confidential data from unauthorized access. Furthermore, we have enabled public access to specific datasets, ensuring that essential information is available to the public. This approach balances transparency with security, allowing us to share valuable insights with the community while maintaining stringent data protection standards.

**Conclusion**

The successful implementation of the NYC payroll data integration project has provided show the financial resource allocation and enhanced transparency for the public. The developed ETL pipeline, with robust data quality assurance measures and user access controls, ensures the reliability and integrity of the data. By maintaining thorough documentation and leveraging version control, the project remains scalable and reproducible for future enhancements and collaborations.