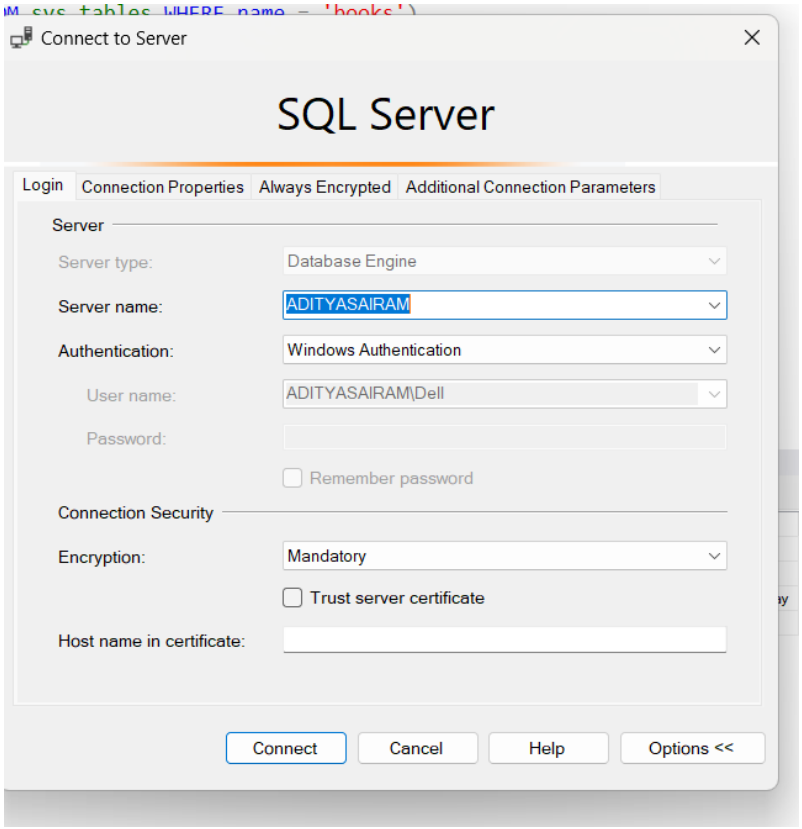
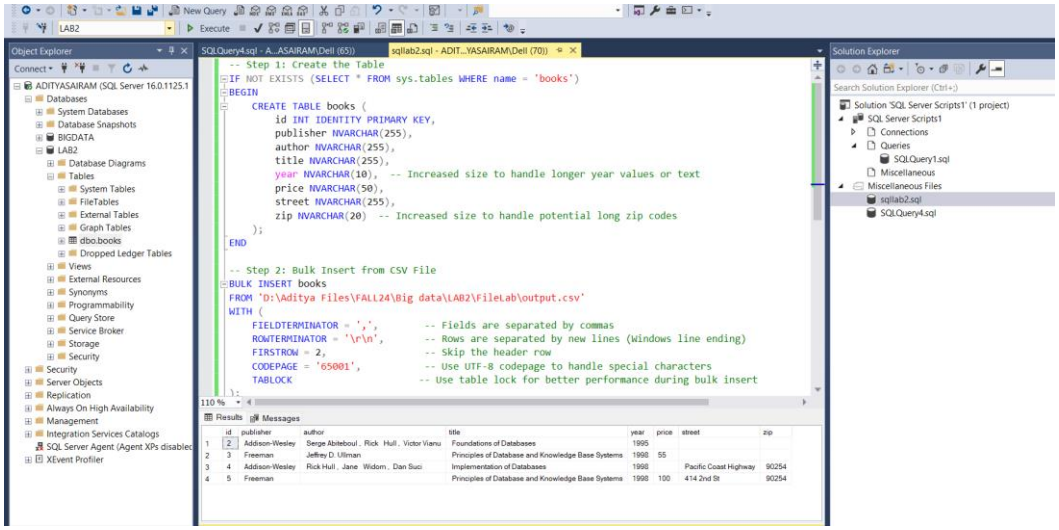
Lab 3 Assignment on XML

- **Platform setup(SQL server for SQL code)**



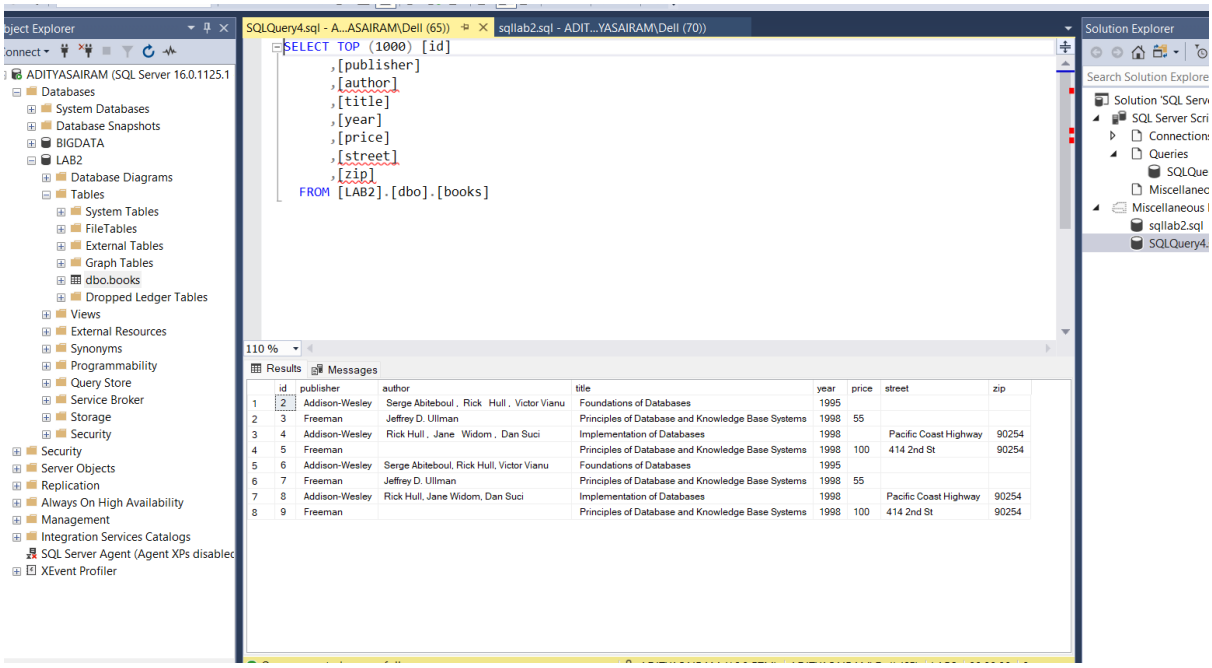


Sql code:

```
-- Step 1: Create the Table
IF NOT EXISTS (SELECT * FROM sys.tables WHERE name = 'books')
BEGIN
    CREATE TABLE books (
        id INT IDENTITY PRIMARY KEY,
        publisher NVARCHAR(255),
        author NVARCHAR(255),
```

```
        title NVARCHAR(255),
        year NVARCHAR(10),  -- Increased size to handle longer year values or
text
        price NVARCHAR(50),
        street NVARCHAR(255),
        zip NVARCHAR(20)  -- Increased size to handle potential long zip codes
    );
END

-- Step 2: Bulk Insert from CSV File
BULK INSERT books
FROM 'D:\Aditya Files\FALL24\Big data\LAB2\FileLab\output.csv'
WITH (
    FIELDTERMINATOR = ',',        -- Fields are separated by commas
    ROWTERMINATOR = '\r\n',       -- Rows are separated by new lines (Windows
line ending)
    FIRSTROW = 2,                 -- Skip the header row
    CODEPAGE = '65001',           -- Use UTF-8 codepage to handle special
characters
    TABLOCK                      -- Use table lock for better performance
during bulk insert
);

-- Step 3: Verify Inserted Data (Optional)
SELECT * FROM books;
```

Output:



	id	publisher	author	title	year	price	street	zip
1	2	Addison-Wesley	Serge Abiteboul , Rick Hull , Victor Vianu	Foundations of Databases	1995			
2	3	Freeman	Jeffrey D. Ullman	Principles of Database and Knowledge Base Systems	1998	55		
3	4	Addison-Wesley	Rick Hull , Jane Widom , Dan Suci	Implementation of Databases	1998		Pacific Coast Highway	90254
4	5	Freeman		Principles of Database and Knowledge Base Systems	1998	100	414 2nd St	90254
5	6	Addison-Wesley	Serge Abiteboul, Rick Hull, Victor Vianu	Foundations of Databases	1995			
6	7	Freeman	Jeffrey D. Ullman	Principles of Database and Knowledge Base Systems	1998	55		
7	8	Addison-Wesley	Rick Hull, Jane Widom, Dan Suci	Implementation of Databases	1998		Pacific Coast Highway	90254
8	9	Freeman		Principles of Database and Knowledge Base Systems	1998	100	414 2nd St	90254

Parse.py

```
import xml.etree.ElementTree as ET
import csv

def parse_xml_to_csv(xml_file, csv_file):
    tree = ET.parse(xml_file)
    root = tree.getroot()

    with open(csv_file, 'w', newline='', encoding='utf-8') as file:
        writer = csv.writer(file)
        writer.writerow(["Publisher", "Author", "Title", "Year", "Price",
"Street", "Zip"])  # Header

        for bib in root.findall('bib'):
            for book in bib.findall('book'):
                publisher = book.find('publisher').text if
book.find('publisher') is not None else ''
                year = book.find('year').text.strip() if book.find('year') is
not None else ''
                title = book.find('title').text if book.find('title') is not
None else ''
                price = book.get('price', '').strip()  # Attribute price

                authors = []
                for author in book.findall('author'):
                    if author.find('first-name') is not None and
author.find('last-name') is not None:
                        authors.append(f"{author.find('first-
name').text.strip()} {author.find('last-name').text.strip()}")
                    else:
                        authors.append(author.text.strip())
                author = ", ".join(authors)

                address = book.find('author/address')
                street = address.find('street').text.strip() if address is not
None and address.find('street') is not None else ''
                zip_code = address.find('zip').text.strip() if address is not
None and address.find('zip') is not None else ''

                writer.writerow([publisher, author, title, year, price,
street, zip_code])

if __name__ == "__main__":
    parse_xml_to_csv("bibXMLInputNoDup.xml", "output.csv")
```

sql.py

```
import pyodbc
import csv

def create_table(cursor):
    cursor.execute('''
        IF NOT EXISTS (SELECT * FROM sys.tables WHERE name = 'books')
        CREATE TABLE books (
            id INT IDENTITY PRIMARY KEY,
            publisher NVARCHAR(255),
            author NVARCHAR(255),
            title NVARCHAR(255),
            year NVARCHAR(10),  -- Increased the size to handle longer year
values or text
            price NVARCHAR(50),
            street NVARCHAR(255),
            zip NVARCHAR(20)  -- Increased size to handle potential long zip
codes
        )
    ''')
    cursor.commit()

def insert_data_from_csv(cursor, csv_file):
    with open(csv_file, 'r', encoding='utf-8') as file:  # Added encoding to
handle special characters
        reader = csv.DictReader(file)

        # Prepare the bulk insert data without the 'id' column
        data = [(row['Publisher'], row['Author'], row['Title'],
row['Year'].strip(), row['Price'], row['Street'], row['Zip'])
                for row in reader if row['Publisher'] and row['Title'] and
row['Year']]  # Ensure mandatory fields are present

    # Bulk insert into SQL Server
    insert_query = '''
        INSERT INTO books (publisher, author, title, year, price, street, zip)
        VALUES (?, ?, ?, ?, ?, ?, ?)
    '''
    cursor.executemany(insert_query, data)
    cursor.commit()

if __name__ == "__main__":
    # Replace these parameters with your actual SQL Server configuration
    server = 'ADITYASAIRAM'
    database = 'LAB2'

    connection_string = f'DRIVER={{SQL
Server}};SERVER={server};DATABASE={database};Trusted_Connection=yes;'
```

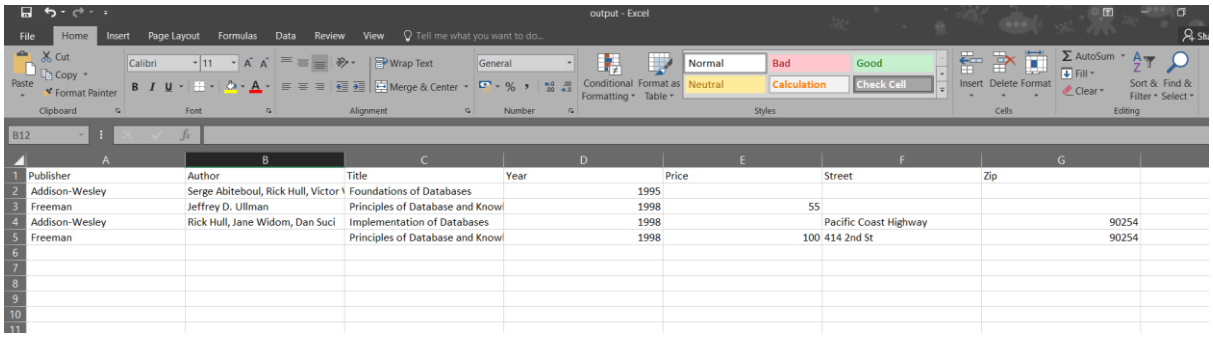
```
    try:
        conn = pyodbc.connect(connection_string)
        cursor = conn.cursor()

        # Create the table and insert data
        create_table(cursor)
        insert_data_from_csv(cursor, "output.csv")

    except pyodbc.Error as ex:
        print("Error:", ex)
    finally:
        cursor.close()
        conn.close()
```

xml file:

```
<bibs>
    <bib>
        <book>
            <publisher> Addison-Wesley </publisher>
            <author> Serge Abiteboul </author>
            <author>
                <first-name> Rick </first-name>
                <last-name> Hull </last-name>
            </author>
            <author> Victor Vianu </author>
            <title> Foundations of Databases </title>
            <year> 1995 </year>
        </book>
        <book price="55">
            <publisher> Freeman </publisher>
            <author> Jeffrey D. Ullman </author>
            <title> Principles of Database and Knowledge Base Systems </title>
            <year> 1998 </year>
        </book>
    </bib>
    <bib>
        <book>
            <publisher> Addison-Wesley </publisher>
            <author> Rick Hull </author>
            <author>
                <first-name> Jane </first-name>
                <last-name> Widom </last-name>
                <address>
                    <street> Pacific Coast Highway </street>
                    <zip> 90254 </zip>
```

```
                </address>
            </author>
            <author> Dan Suci </author>
            <title> Implementation of Databases </title>
            <year> 1998 </year>
        </book>
        <book price="100">
            <publisher> Freeman </publisher>
            <author>
                <name> Jeffrey D. Ullman </name>
                <address>
                    <street> 414 2nd St </street>
                    <zip> 90254 </zip>
                </address>
            </author>
            <title> Principles of Database and Knowledge Base Systems </title>
            <year> 1998 </year>
        </book>
        <paper price="15">
            <publisher> ACM Press </publisher>
            <author>
                <name> Jeffrey Ullman </name>
                <address>
                    <street> 200 Sepuveda </street>
                    <zip> 90245 </zip>
                </address>
            </author>
            <title> Principles of Database and Knowledge Base Systems </title>
            <year> 1998 </year>
        </paper>
        <paper price="10">
            <publisher> IEEE Press </publisher>
            <author> Jeffrey D. Ullman </author>
            <title> Cloud Azure </title>
            <year> 2010 </year>
        </paper>
    </bib>
</bibs>
```

CSV file output:



	A	B	C	D	E	F	G
1	Publisher	Author	Title	Year	Price	Street	Zip
2	Addison-Wesley	Serge Abiteboul, Rick Hull, Victor V	Foundations of Databases	1995			
3	Freeman	Jeffrey D. Ullman	Principles of Database and Knowl	1998	55		
4	Addison-Wesley	Rick Hull, Jane Widom, Dan Suci	Implementation of Databases	1998		Pacific Coast Highway	90254
5	Freeman		Principles of Database and Knowl	1998	100	414 2nd St	90254
6							
7							
8							
9							
10							
11							