Name: Girum Obse
Date: 02/18/2025
Course: BIDD 320 A /Data Migration Techniques (ETL Processing)
Assignment: A05_G.Obse

ETL Workflow for Phone Number Validation with No-SQL Scripts and SSIS

1. Introduction

2. Extracting Source Data

2.1 Source Data Location

- File path and structure of TestData.csv.

2.2 Explanation of Source Data

- Description of fields in the CSV file.

3. Validating Phone Numbers

- Required phone number format (111-222-3333).
- Separation of valid and invalid records.

3.1 Screenshot of ValidData.csv

3.2 Screenshot of BadData.csv

4. Loading Data into SQL Server

4.1 Create Staging Table

- SQL script to create the CustomersStaging table.

4.2 Explanation of Staging Table

- Description of table structure and columns.

4.3 Screenshot of SQL Table Creation

5. SSIS Package Setup

6. Summary

1. Introduction

This document provides a step-by-step guide to an ETL (Extract, Transform, Load) process designed to validate phone numbers from a CSV file. The process ensures that phone numbers adhere to a standardized format before being stored in a structured database. By implementing this validation, organizations can maintain data integrity, improve operational efficiency, and ensure consistency in their customer records.

The ETL process involves three main phases:

1. Extract: The data is retrieved from a source file located at C:\Data\TestData.csv.

2. Transform: The phone numbers are validated against a defined pattern, separating valid and invalid records.

3. Load: The valid records are stored in a SQL Server staging table for further processing, while invalid records are logged separately for review and correction.

This document is intended for data engineers, ETL developers, and support teams responsible for maintaining data quality and troubleshooting the process. Detailed explanations of each step, along with relevant screenshots, are provided to facilitate understanding and implementation.

2. Extracting Source Data

2.1 Source Data Location

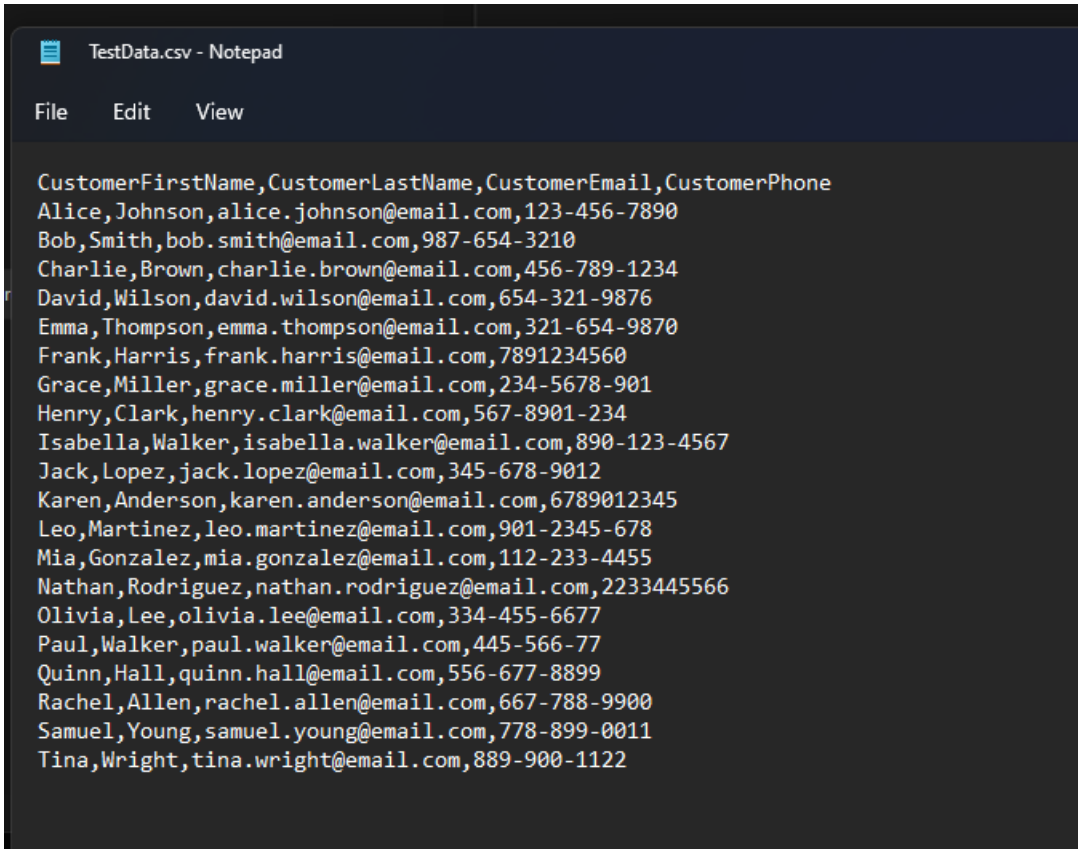
The input data file is stored at: C:\Data\TestData.csv

2.2 Explanation of Source Data

The source data file is a CSV (Comma-Separated Values) file that consists of multiple customer records. Each row represents a customer and includes the following fields: Customer First Name, Customer Last Name, Customer Email, and Customer Phone Number.

2.3 Screenshot of TestData.csv

Picture 1,Screenshot of TestData.csv



3. Validating Phone Numbers

Phone numbers are expected to follow the format: 111-222-3333.

Any entry that does not match this pattern is considered invalid.

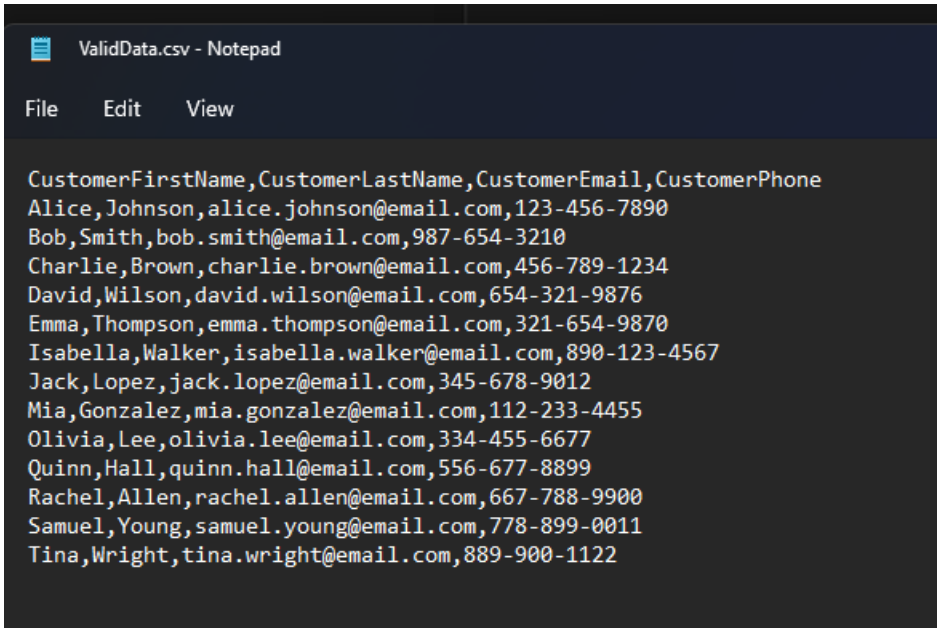
The transformation phase separates valid and invalid phone numbers, storing them in two different CSV files:

- Valid Data: Stored in C:\Data\ValidData.csv

- Invalid Data: Stored in C:\Data\BadData.csv

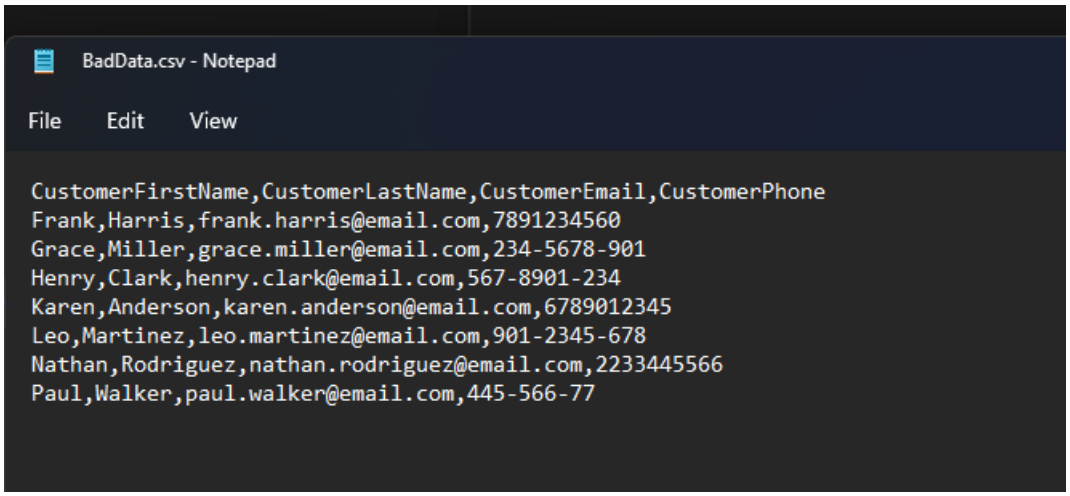
3.1 Screenshot of ValidData.csv

Picture 2, Screenshot of valid data



3.2 Screenshot of BadData.csv

Picture3, Screenshot of BadData.



4. Loading Data into SQL Server

4.1 Create Staging Table

To store the valid customer records, a staging table is created in SQL Server. The table schema ensures that customer data is structured correctly and can be efficiently queried for further processing.

```
Use TempDB;
Go

Create Table CustomersStaging
```

```
(
   CustomersStaging int identity Primary Key,
   CustomerFirstName nvarchar(100),
   CustomerLastName nvarchar(100),
   CustomerEmail nvarchar(100),
   CustomerPhone nvarchar(100)
);
```

4.2 Explanation of Staging Table

The table consists of the following columns:

- CustomersStaging: Auto-incremented primary key.

- CustomerFirstName: Stores the first name of the customer.

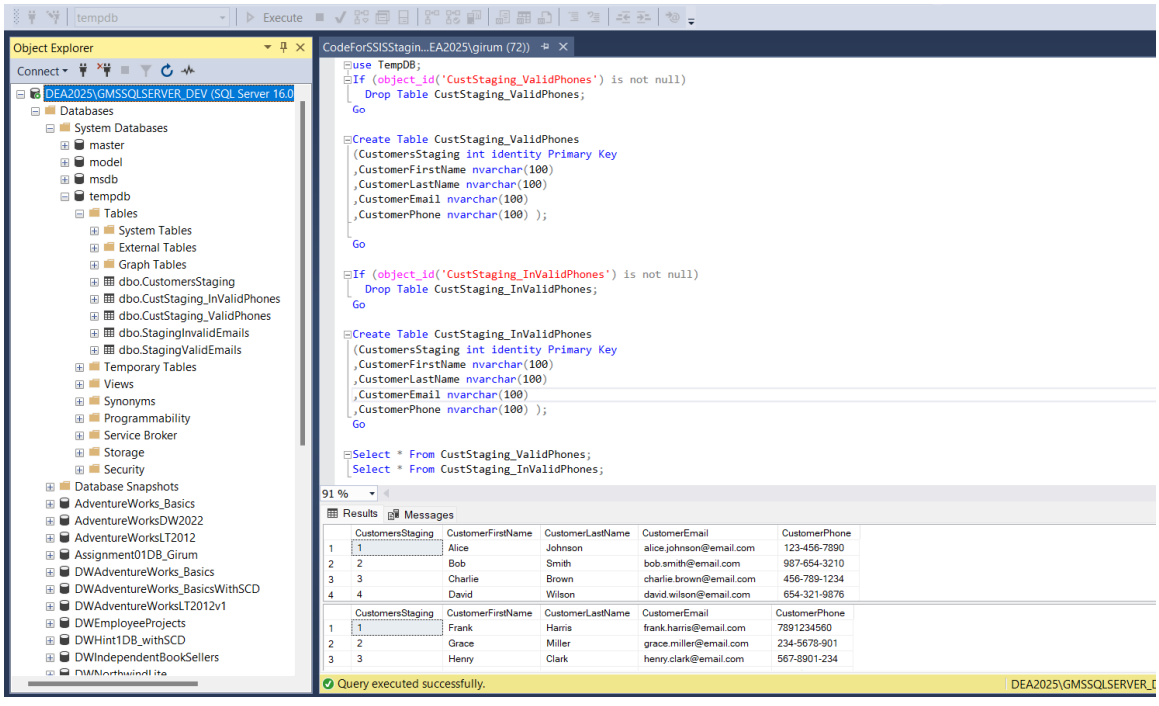
- CustomerLastName: Stores the last name of the customer.

- CustomerEmail: Stores the email address.

- CustomerPhone: Stores the validated phone number.

4.3 Screenshot of SQL Table Creation

Picture 4, screenshot of the tables creation in tempdb



5. SSIS Package Setup

5.1 Overview of SSIS Package

An SSIS (SQL Server Integration Services) package is used to automate the ETL process. The package consists of multiple tasks that execute sequentially to extract, validate, and load the data.

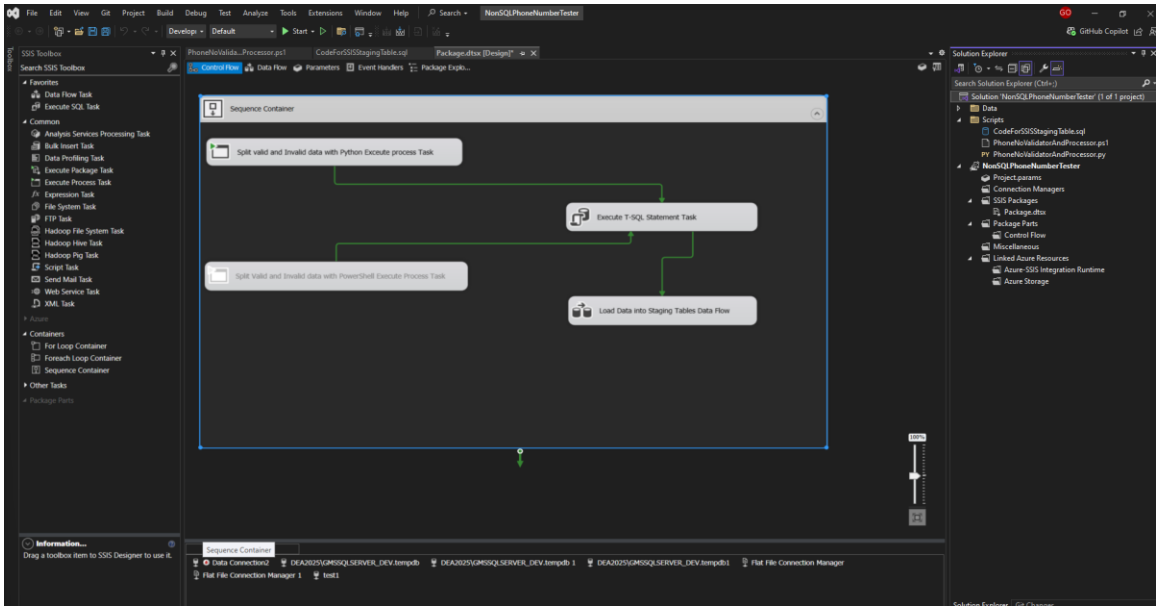
5.2 Execute Process Task

- Runs the validation process to filter valid and invalid phone numbers.

- Ensures only correctly formatted phone numbers proceed to the database.

5.3 Screenshot of SSIS Package Control Flow

Picture 5,screenshot of the SSIS Package Control



5.4 Data Flow Task

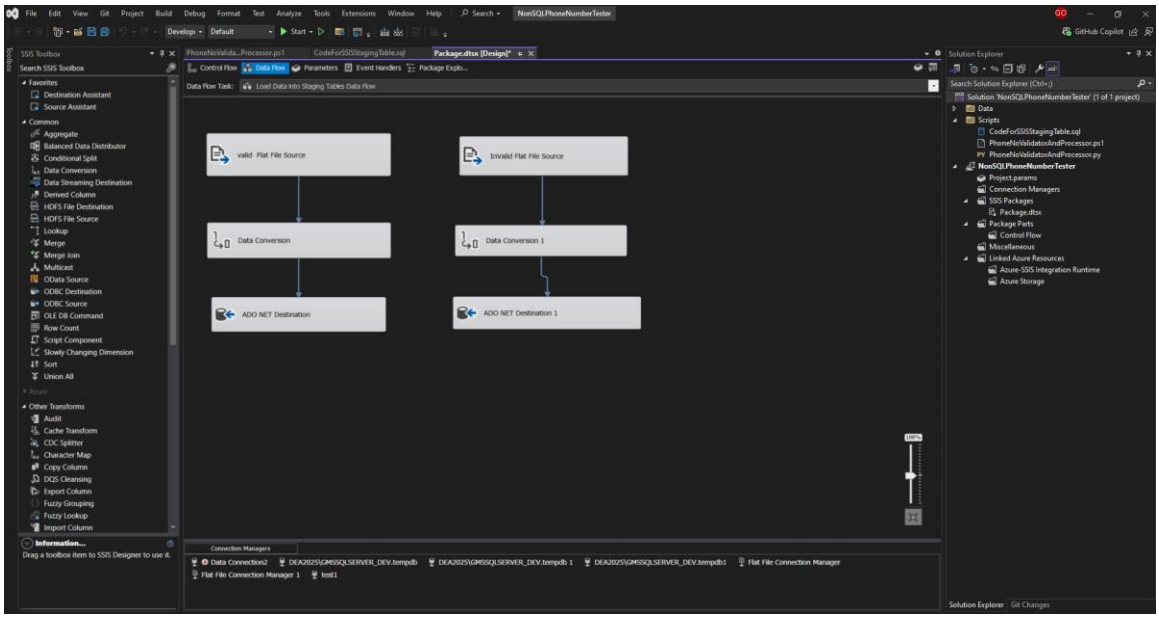
- Extracts data from ValidData.csv.

- Transforms and maps the data to match the staging table schema.

- Loads the processed data into SQL Server.

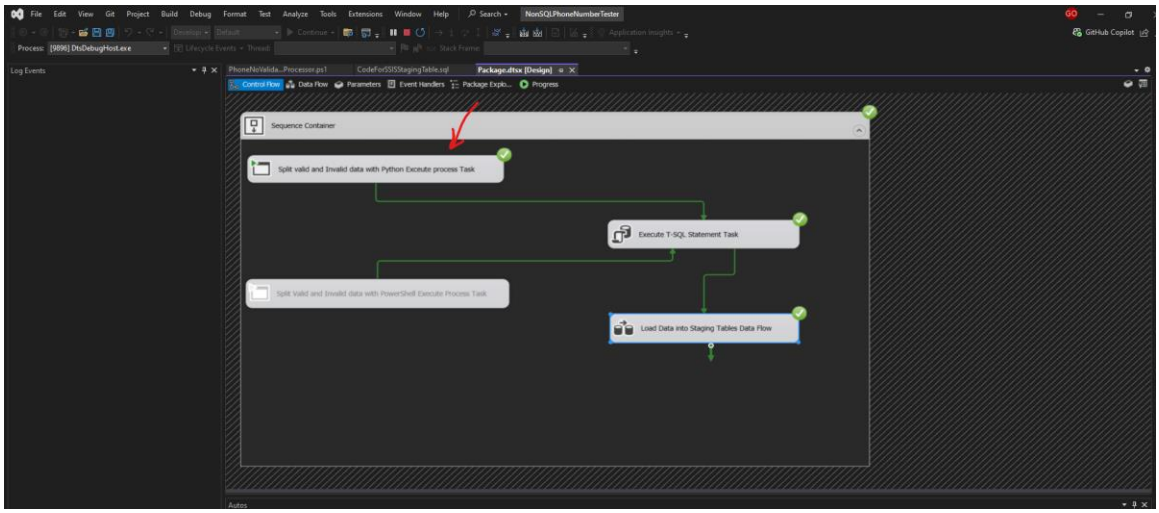
5.5 Screenshot of SSIS Package Data Flow

Picture 6, screenshot of the SSIS Package Data flow

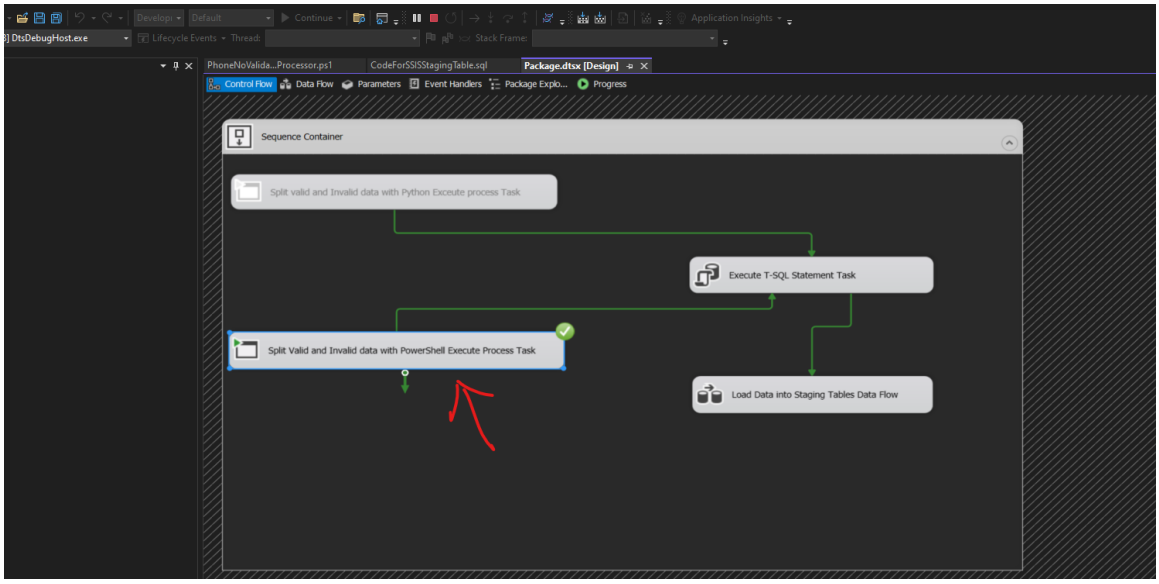


5.6 Entire SSIS Package Overview using Python and Power shell Excute process

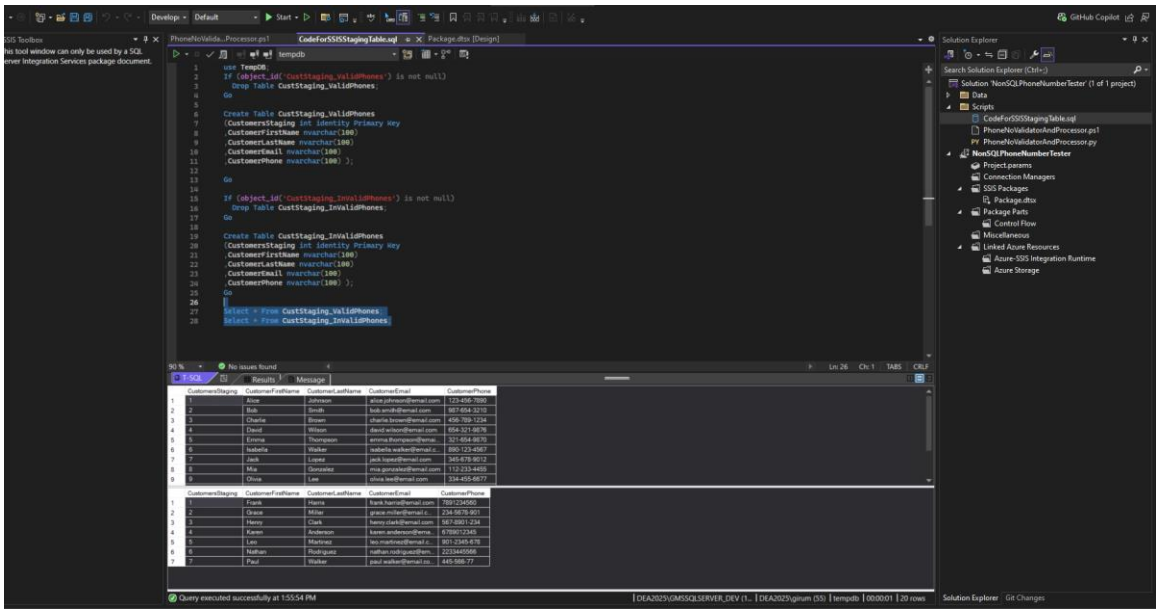
Picture7, screenshot of the SSIS using Python Execute process



Picture 8, screenshot of the ssis using Powershell execute process.



Picture 9, screenshot of the tables showing valid and bad date



7. Summary

This document serves as a comprehensive guide to the ETL process for validating and loading phone number data. By implementing this ETL workflow, organizations can maintain high data integrity, streamline customer data management, and reduce errors in communication and reporting.

By following the structured steps outlined in this document, users can effectively manage and troubleshoot the ETL process, ensuring seamless execution and data accuracy. The use of SSIS enhances automation, minimizing manual intervention and improving efficiency.

This document is designed to be a reference guide for new hires, ETL developers, and support technicians who work with data validation and integration tasks. Detailed explanations, along with screenshots, ensure clarity and ease of use, making troubleshooting and implementation straightforward.