Module 6

Creating an ETL Solution

Zyead Ahmed.
Aspring To Learning Data Engineer.

Module Overview

- Introduction to ETL with SSIS
- Exploring Source Data
- Implementing Data Flow

Lesson 1: Introduction to ETL with SSIS

- Options for ETL
- What Is SSIS?
- SSIS Projects and Packages
- The SSIS Design Environment
- Upgrading from Previous Versions

Options for ETL

- SQL Server Integration Services
- The Import and Export Data Wizard
- Transact-SQL
- The bcp utility
- Replication

What Is SSIS?

- A platform for ETL operations
- Installed as a feature of SQL Server
- Control flow engine:
  - Runtime resources and operational support for data flow
- Data flow engine:
  - Pipeline architecture for buffer-oriented rowset processing

SSIS Projects and Packages

- Project Deployment Model:
  - Multiple packages are deployed in a single project
- Package Deployment Model:
  - SSIS packages are deployed and managed individually

The SSIS Design Environment

- Solution Explorer
- Properties pane
- Control Flow design surface
- Data Flow design surface
- Parameters tab
- Event Handlers design surface
- Package Explorer
- Connection Managers pane
- Variables pane
- SSIS Toolbox

Upgrading from Previous Versions

- Upgrading Packages in a Project:
  - Open project file in SQL Server Data Tools
  - SSIS Package Upgrade Wizard will automatically be activated
- Upgrading a Single Package:
  - Open package file in SQL Server Data Tools
  - Package will automatically be upgraded
- Scripts:
  - Migrated VSA scripts are automatically updated to VSTA
  - Microsoft ActiveX scripts are no longer supported and must be replaced

Lesson 2: Exploring Source Data

- Why Explore Source Data?
- Examining Source Data
- Demonstration: Exploring Source Data
- Profiling Source Data
- Demonstration: Using the Data Profiling Task

Why Explore Source Data?

- Understand business data:
  - What business entities are represented
  - How to interpret values and codes
  - Relationships between business entities
- Examine data for:
  - Column data types and lengths
  - Data volume and sparseness
  - Data quality issues

Examining Source Data

- Extract a sample of data:
  - Run queries in SSMS
  - Create an SSIS package that extracts a sample of data
  - Use the Import and Export Data Wizard
- Examine the data using an appropriate application such as Excel

Demonstration: Exploring Source Data

In this demonstration, you will see how to:

- Extract data by using the Import and Export Data Wizard
- Explore the data

Profiling Source Data

- Use the Data Profiling task in SSIS to report data statistics:
  - Candidate key
  - Column length distribution
  - Column null ratio
  - Column pattern
  - Column statistics
  - Column value distribution
  - Functional dependency
  - Value inclusion
- View the profile in the Data Profile Viewer

Demonstration: Using the Data Profiling Task

In this demonstration, you will see how to:

- Use the Data Profiling Task
- View a Data Profiling Report

Lesson 3: Implementing Data Flow

- Connection Managers
- The Data Flow Task
- Data Source Components
- Data Destination Components
- Data Transformation Components
- Optimizing Data Flow Performance
- Demonstration: Implementing a Data Flow

Connection Managers

- A connection to a data source or destination:
  - Provider (for example, ADO.NET, OLE DB, or flat file)
  - Connection string
  - Credentials
- Project or package level:
  - Project-level connection managers can be shared across multiple packages
  - Package-level connection managers exist only in that package

The Data Flow Task

- The core control flow task in most SSIS packages
- Encapsulates a data flow pipeline
- Define the pipeline for the task on the Data Flow tab

Data Source Components

- The source of data for a data flow:
  - Connection manager
  - Table, view, or query
  - Columns that are included
- Many sources supported:
  - Database
  - File
  - Custom

Data Destination Components

- Endpoint for a data flow:
  - Connection manager
  - Table or view
  - Column mapping
- Multiple destination types:
  - Database
  - File
  - SQL Server Analysis Services
  - Rowset
  - Other

Data Transformation Components

- Perform operations on rows of data
- Use inputs and outputs
- Multiple transformation types:
  - Row Transformations
  - Rowset Transformations
  - Split and Join Transformations
  - Auditing Transformations
  - BI Transformations
  - Custom Transformations
- Blocking types:
  - Non-blocking
  - Partial-blocking
  - Blocking

Optimizing Data Flow Performance

- Optimize queries:
  - Select only the rows and columns that you need
- Avoid unnecessary sorting:
  - Use pre-sorted data where possible
  - Set the **IsSorted** property where applicable
- Configure Data Flow task properties:
  - Buffer size
  - Temporary storage location
  - Parallelism
  - Optimized mode

Demonstration: Implementing a Data Flow

In this demonstration, you will see how to:

- Configure a data source
- Use a derived column transformation
- Use a lookup transformation
- Configure a destination