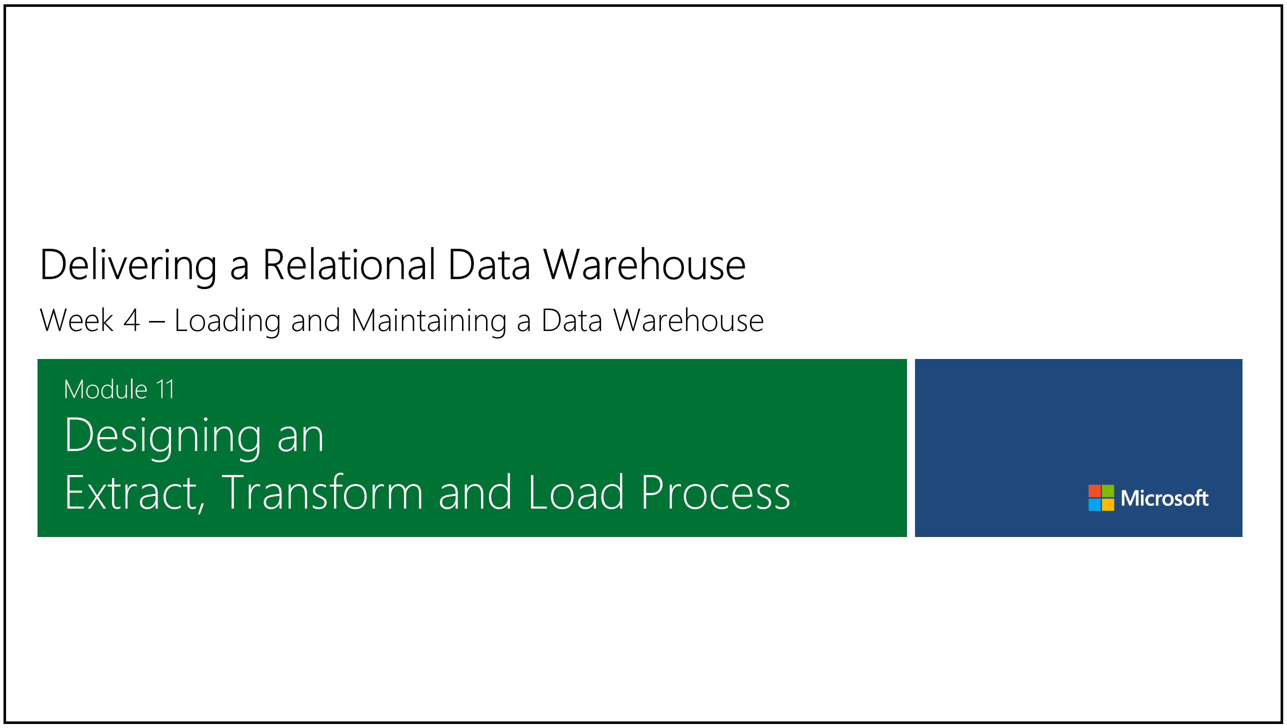


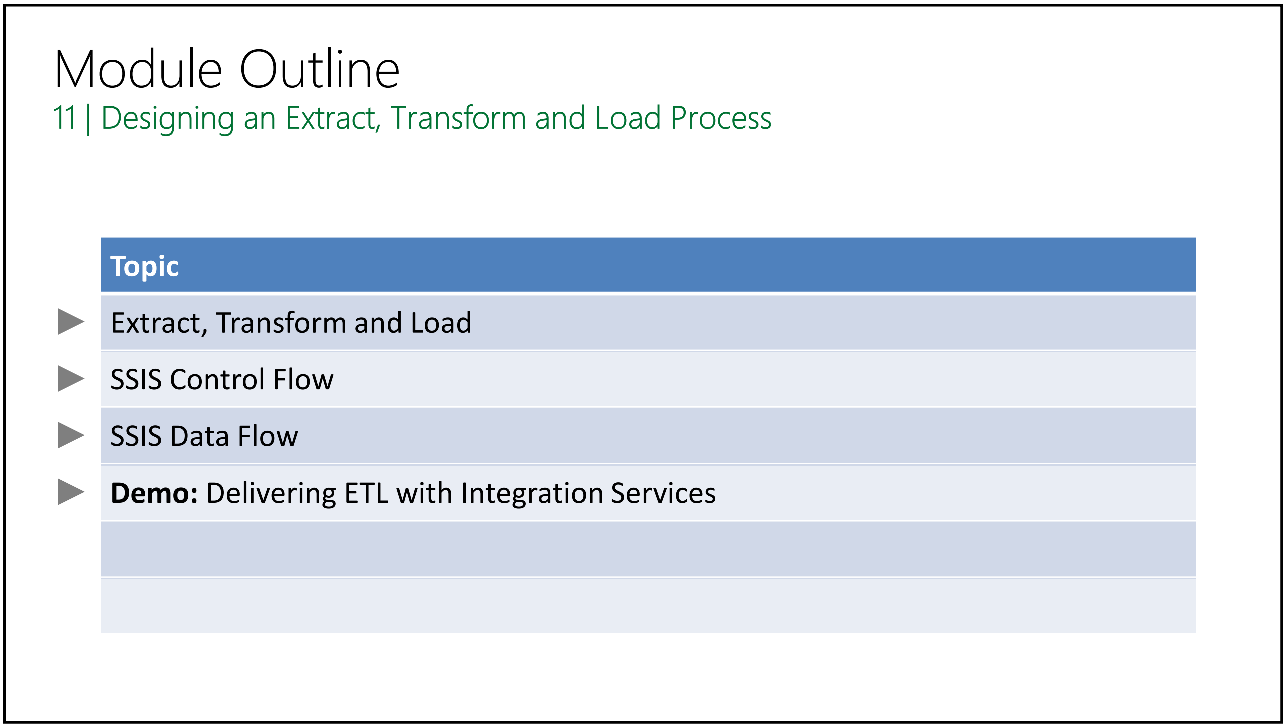
Delivering a Relational Data Warehouse

Week 4 – Loading and Maintaining a Data Warehouse

Module 11

Designing an Extract, Transform and Load Process

Microsoft



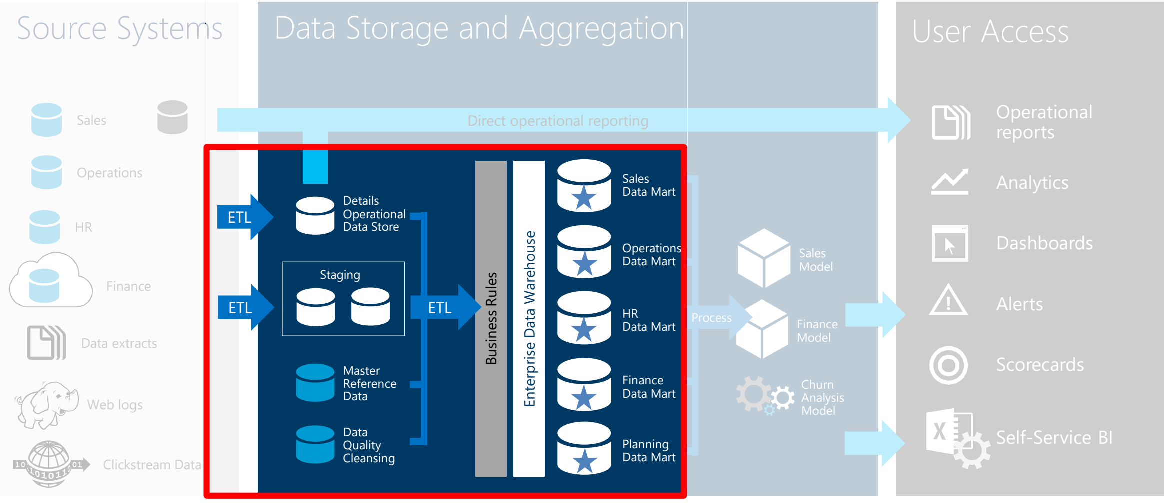
Module Outline

11 | Designing an Extract, Transform and Load Process

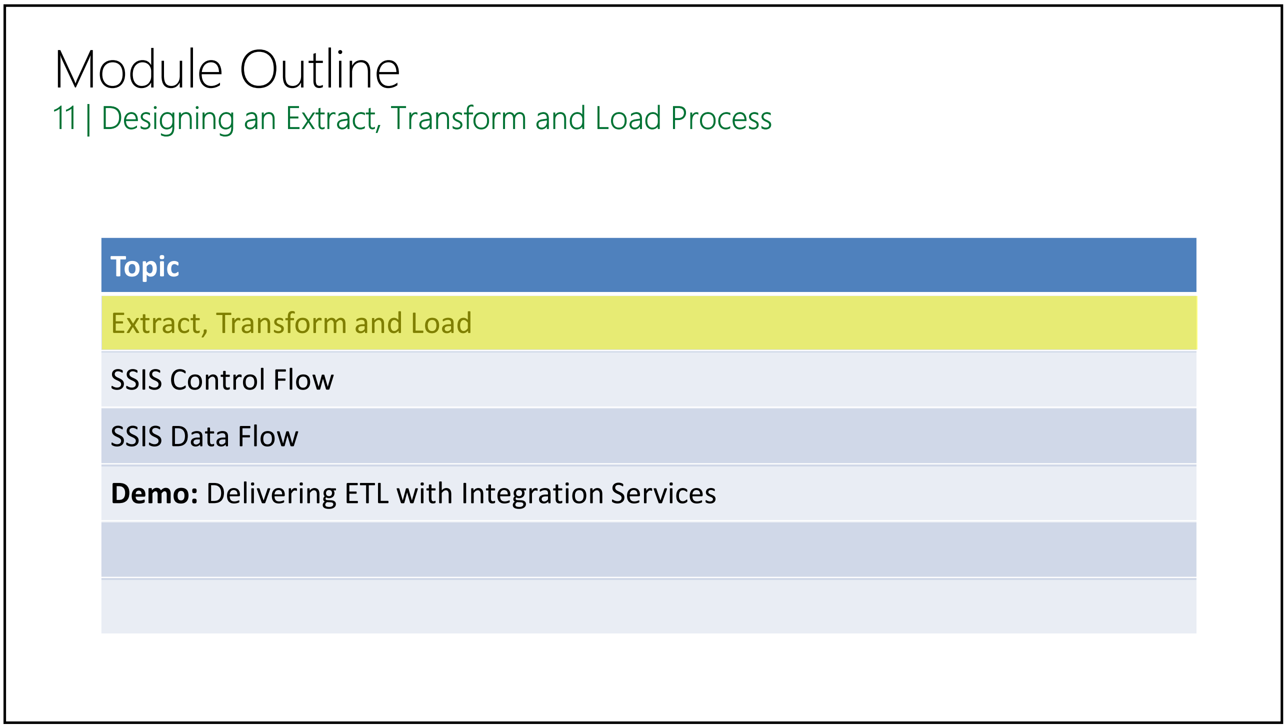
	Topic
▶	Extract, Transform and Load
▶	SSIS Control Flow
▶	SSIS Data Flow
▶	**Demo:** Delivering ETL with Integration Services

Module Outline

11 | Designing an Extract, Transform and Load Process



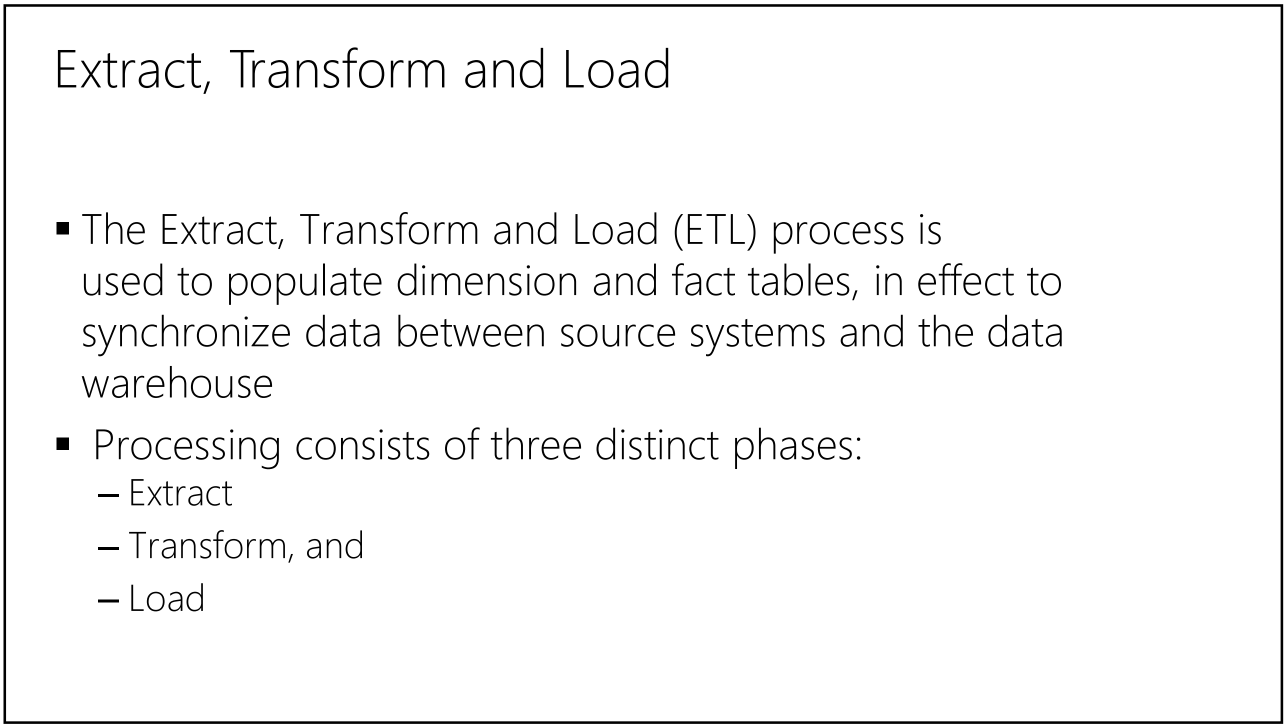
Microsoft



Module Outline

11 | Designing an Extract, Transform and Load Process

Topic
Extract, Transform and Load
SSIS Control Flow
SSIS Data Flow
Demo: Delivering ETL with Integration Services



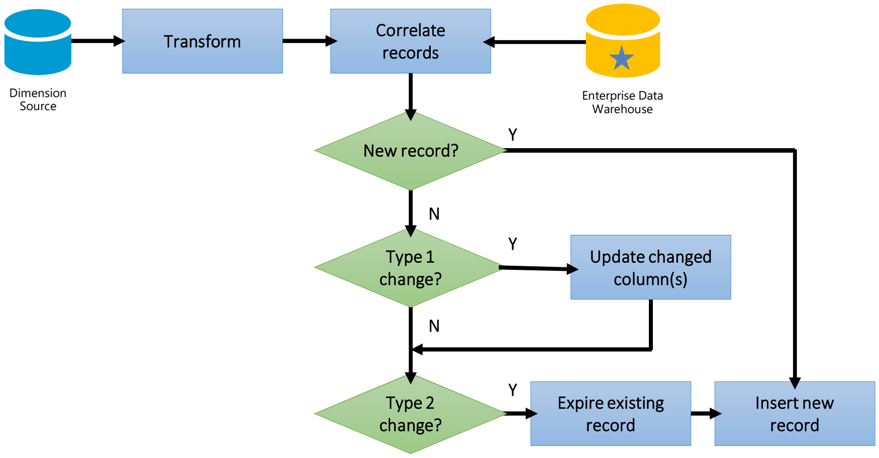
Extract, Transform and Load

- The Extract, Transform and Load (ETL) process is used to populate dimension and fact tables, in effect to synchronize data between source systems and the data warehouse
- Processing consists of three distinct phases:
  - Extract
  - Transform, and
  - Load

Extract, Transform and Load
(Continued)

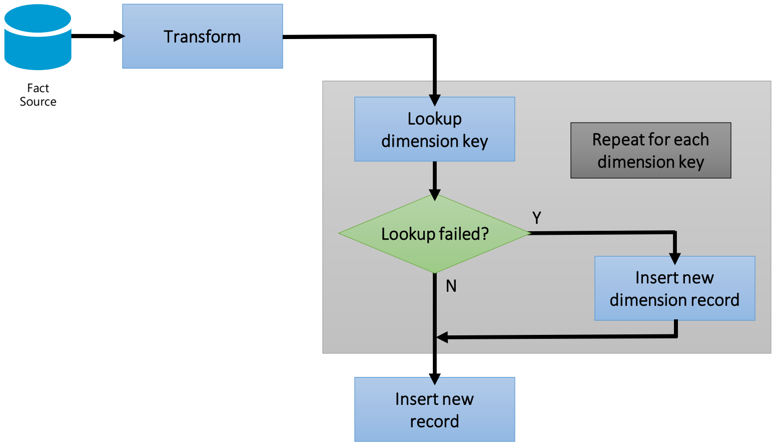
- Common challenges:
  - Retrieving and integrating data from multiple sources
  - Cleansing and transforming the data
  - Loading the data into appropriate data stores for analysis and reporting
- Enterprises spend 60%–80% of their resources developing, testing and maintaining their ETL processes
- ETL processes usually require dedicated monitoring and maintenance

Extract, Transform and Load
Populating Dimension Tables



Extract, Transform and Load

Populating Fact Tables



Fact Source

Transform

Lookup dimension key

Repeat for each dimension key

Lookup failed?

Y

N

Insert new dimension record

Insert new record

Extract, Transform and Load

SQL Server Integration Services

- Integration Services (SSIS) is a SQL Server service primary designed to implement ETL processes
- Provides a robust, flexible, fast, scalable and extensible architecture
- Its capabilities are useful in many other scenarios:
  - Assessing data quality
  - Cleansing and standardizing data
  - Merging data from heterogeneous data stores
  - Implementing ad hoc data transfers
  - Automating administrative tasks

Extract, Transform and Load

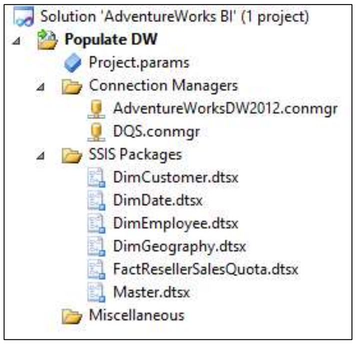
SQL Server Integration Services ► Architecture

- The services has two distinct runtime engines:
  - Control flow
  - Data flow

Extract, Transform and Load

SQL Server Integration Services ► Development

- Solutions are developed by using the Integration Services Project template
- A project can consist of:
  - Parameters
  - Connection Managers
  - Packages

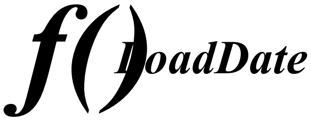


- Solution 'AdventureWorks BI' (1 project)
  - **Populate DW**
    - Project.params
    - Connection Managers
      - AdventureWorksDW2012.conmgr
      - DQS.conmgr
    - SSIS Packages
      - DimCustomer.dtsx
      - DimDate.dtsx
      - DimEmployee.dtsx
      - DimGeography.dtsx
      - FactResellerSalesQuota.dtsx
      - Master.dtsx
    - Miscellaneous

Extract, Transform and Load

SQL Server Integration Services ► Parameters

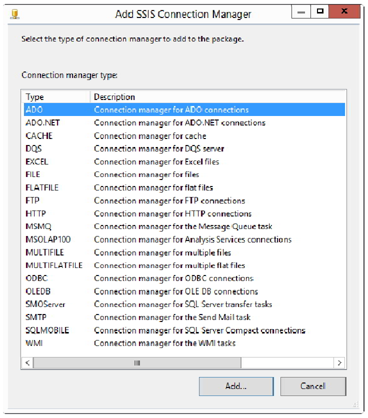
- Allow the assignment of values to properties within packages at package execution time
  - Parameter values, once passed, cannot be modified
- Properties:
  - Name
  - Data type
  - Value (default)
  - Sensitive
  - Required



Extract, Transform and Load

SQL Server Integration Services ► Connection Managers

- Logical representation of a connection
- Created at project or package level
  - Project connection managers are available to all project packages
- Used by package components



Select the type of connection manager to add to the package.

Connection manager type:

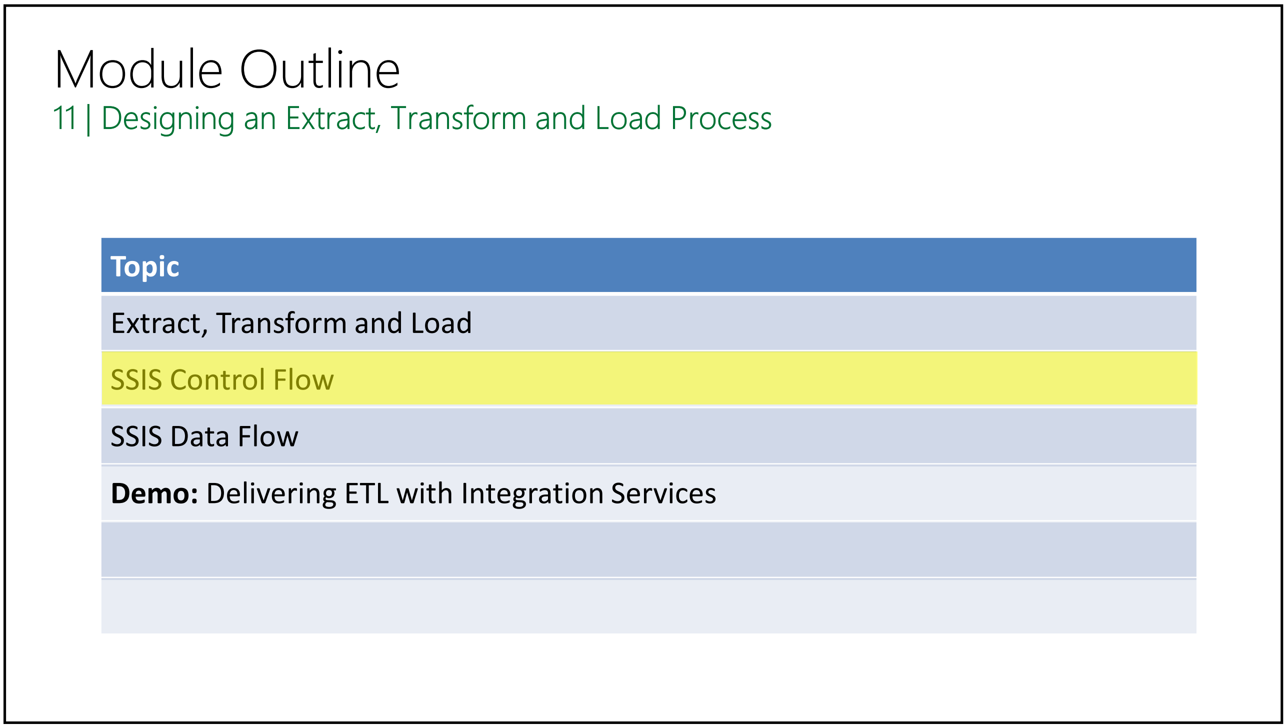
Type	Description
ADO	Connection manager for ADO connections
ADO.NET	Connection manager for ADO.NET connections
CACHE	Connection manager for cache
DQS	Connection manager for DQS server
EXCEL	Connection manager for Excel files
FILE	Connection manager for files
FLATFILE	Connection manager for flat files
FTP	Connection manager for FTP connections
HTTP	Connection manager for HTTP connections
MSMQ	Connection manager for the Message Queue task
MSOLAP100	Connection manager for Analysis Services connections
MULTIFILE	Connection manager for multiple files
MULTIFLATFILE	Connection manager for multiple flat files
ODBC	Connection manager for ODBC connections
OLEDB	Connection manager for OLE DB connections
SMOServer	Connection manager for SQL Server transfer tasks
SMTP	Connection manager for the Send Mail task
SQLMOBILE	Connection manager for SQL Server Compact connections
WMI	Connection manager for the WMI tasks

Extract, Transform and Load

SQL Server Integration Services ► Package

- The basic unit of design, deployment and execution
- An organized collection of:
  - Connection managers
  - Parameters
  - Variables
  - Control flow components, linked by precedence constraints
  - Data flow components, linked by data paths to form a pipeline
- Designed graphically by using the package designer

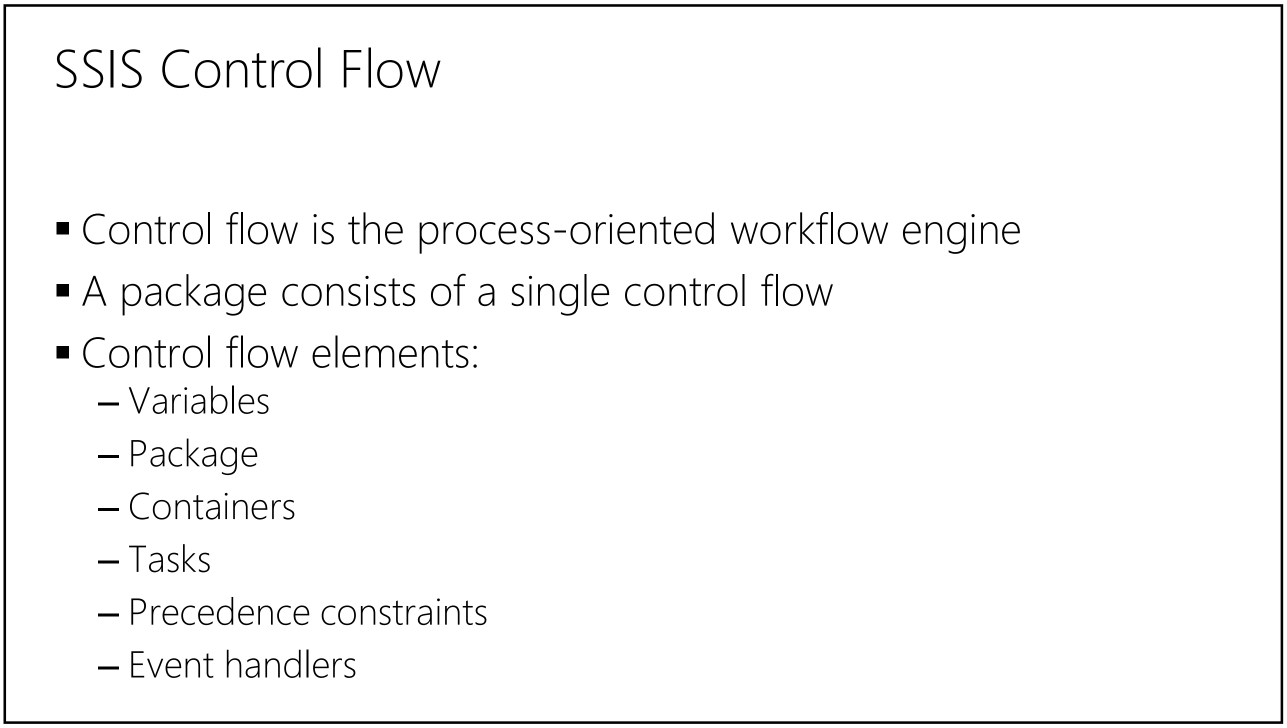
Microsoft



Module Outline

11 | Designing an Extract, Transform and Load Process

Topic
Extract, Transform and Load
SSIS Control Flow
SSIS Data Flow
Demo: Delivering ETL with Integration Services



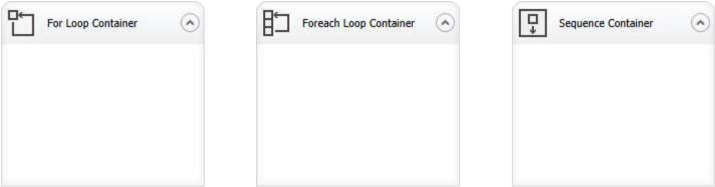
SSIS Control Flow

- Control flow is the process-oriented workflow engine
- A package consists of a single control flow
- Control flow elements:
  - Variables
  - Package
  - Containers
  - Tasks
  - Precedence constraints
  - Event handlers

SSIS Control Flow

Containers

- Provide structure and services for:
  - Grouping tasks
  - Implementing repeating flows
- Can also manage variable and transactional boundaries



SSIS Control Flow

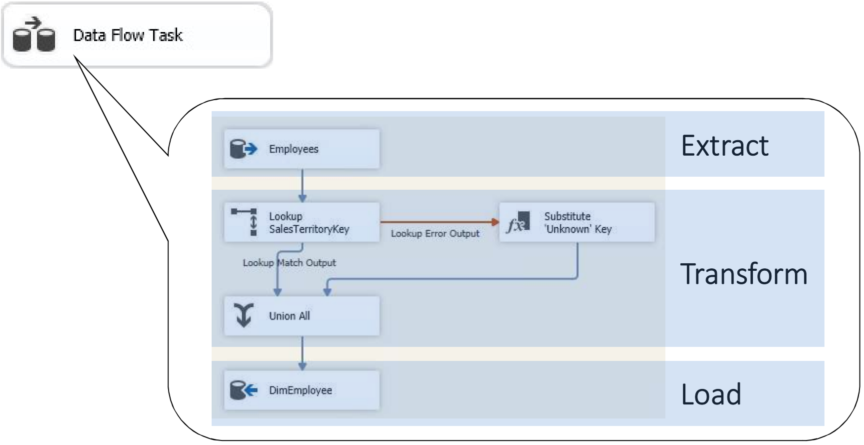
Tasks

- Perform discrete operations
- Categories:
  - Data Flow
  - Data Preparation
  - Process Communication
  - Execute SQL
  - Analysis Services
  - Scripting
  - Miscellaneous

SSIS Control Flow

Tasks ► Data Flow

- Encapsulates the data flow engine



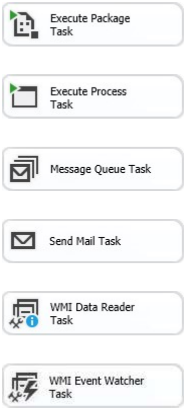
SSIS Control Flow

Tasks ► Data Preparation

Data Profiling Task	▪ Assess data characteristics and quality
File System Task	▪ Copy files and directories
FTP Task	▪ Download or upload files and data
XML Task	▪ Apply operations to XML documents
Web Service Task	▪ Execute Web Service methods

SSIS Control Flow

Tasks ► Process Communication



- Run packages
- Run programs or batch files
- Send and receive asynchronous messages
- Send email messages
- Read Windows Management Instrumentation (WMI) data
- Watch for WMI events

SSIS Control Flow

Tasks ► Execute SQL

- Run SQL statements or stored procedures
  - Can use Excel, OLE DB, ODBC, ADO, ADO.NET, or SQLMOBILE connection managers
  - Parameters can be passed in and out
  - Variable values can be initialized
  - Entire result sets can be stored to a variable
- Examples:
  - Truncate a table in preparation for inserting data
  - Create, alter, or drop database objects like tables and indexes

SSIS Control Flow

Tasks ► Analysis Services

Analysis Services Processing Task

- Process dimensions, cubes and mining models

Analysis Services Execute DDL Task

- Execute Data Definition Language (DDL) statements, typically to create cube partitions

Data Mining Query Task

- Execute a data mining query

SSIS Control Flow

Tasks ► Scripting

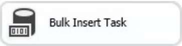
Script Task

- Implement custom logic by using either VB.NET or C#
- Developed by using Visual Studio Tools for Applications (VSTA)
- Features:
  - IntelliSense
  - Color coding
  - Integrated help
  - References to .NET assemblies
  - Debugging

SSIS Control Flow

Tasks ► Miscellaneous

- Set the value of a variable



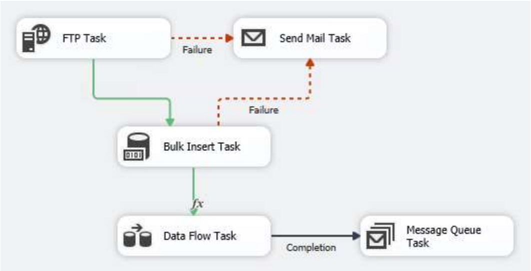
- Efficiently load data into SQL Server from a text file

- Control the extraction of data from Change Data Capture (CDC) tables
  - Works in conjunction with the CDC data flow components

SSIS Control Flow

Precedence Constraints

- Link containers and tasks to control the order of execution
- Configure conditions that determine whether the constrained executable runs:
  - Success, Failure, or Completion constraints
  - Expressions
  - Logical AND/OR for multiple constraints

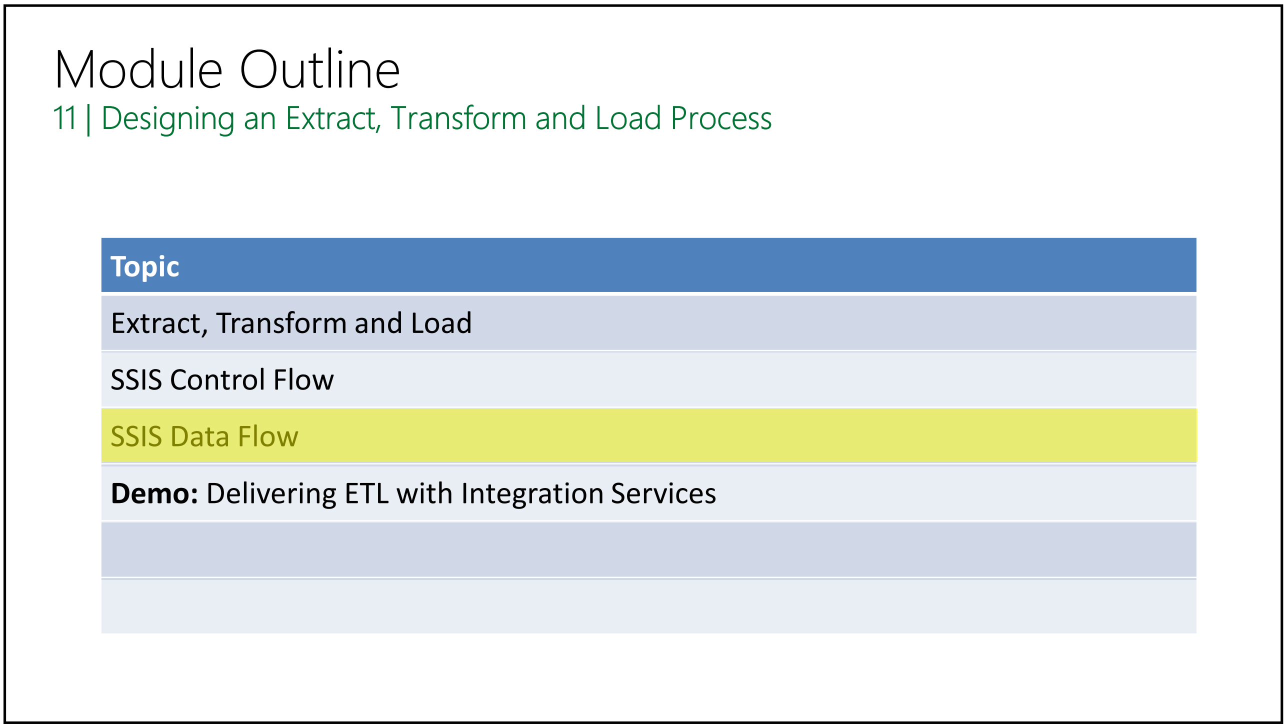


SSIS Control Flow

Event Handlers

- Executables raise events at run time
- Event handlers can be created to respond to these events
- Creating an event handler is based on control flow
- Common events used to trigger event handlers:
  - OnPreExecute, OnPostExecute, and OnError
- Examples:
  - Retrieve system information to assess resource availability before the package runs
  - Send an email message when an error occurs

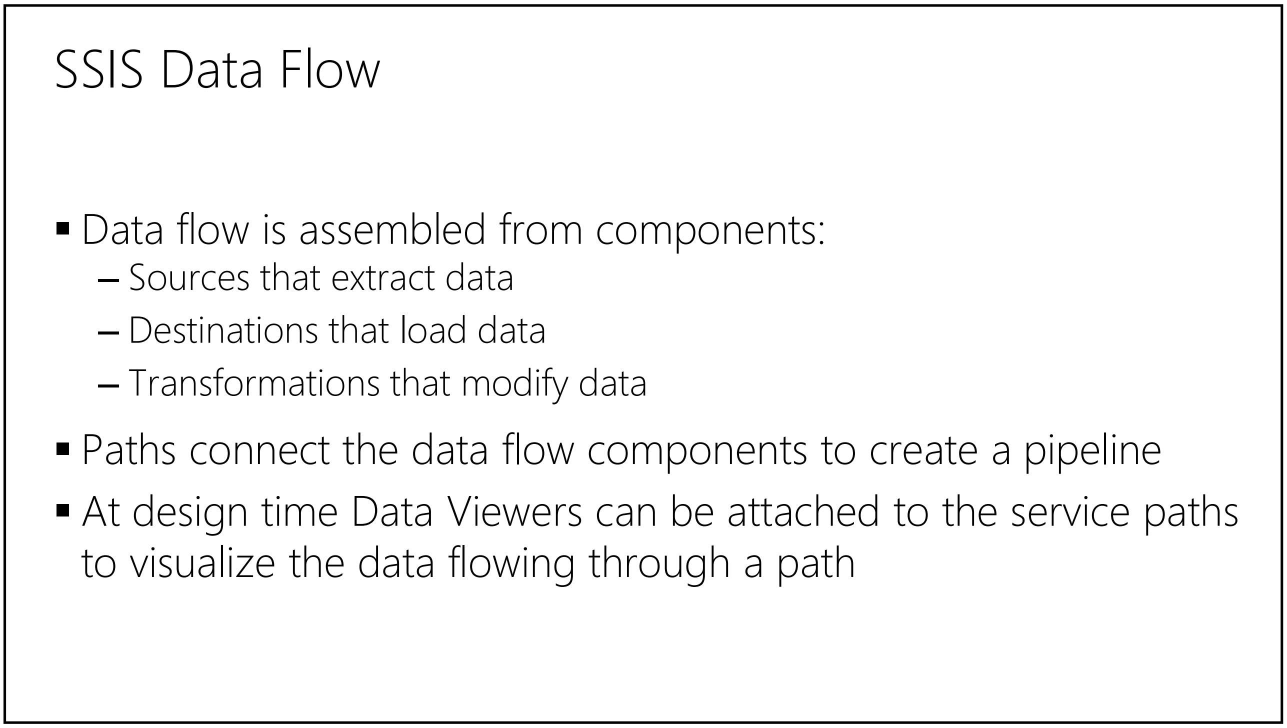
Microsoft



Module Outline

11 | Designing an Extract, Transform and Load Process

Topic
Extract, Transform and Load
SSIS Control Flow
SSIS Data Flow
Demo: Delivering ETL with Integration Services



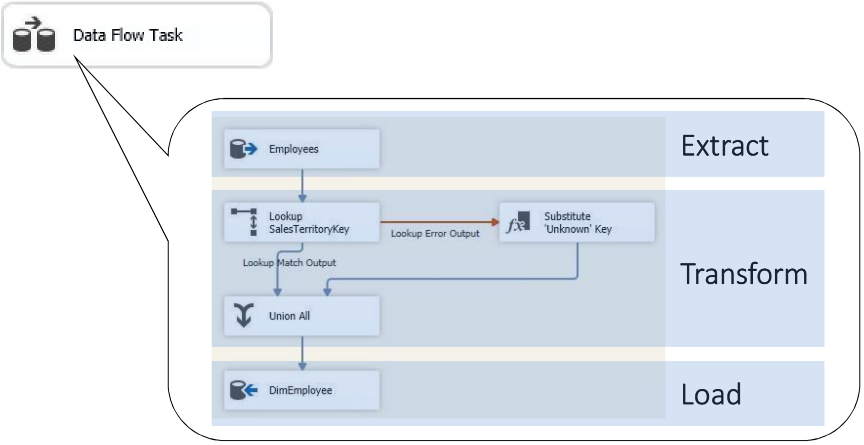
SSIS Data Flow

- Data flow is assembled from components:
  - Sources that extract data
  - Destinations that load data
  - Transformations that modify data
- Paths connect the data flow components to create a pipeline
- At design time Data Viewers can be attached to the service paths to visualize the data flowing through a path

SSIS Data Flow

Data Flow Task

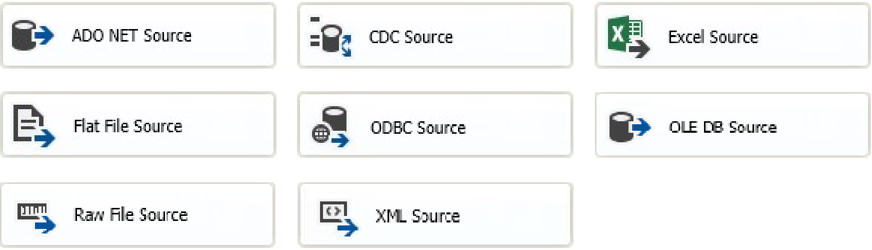
- Encapsulates the data flow engine



SSIS Data Flow

Sources

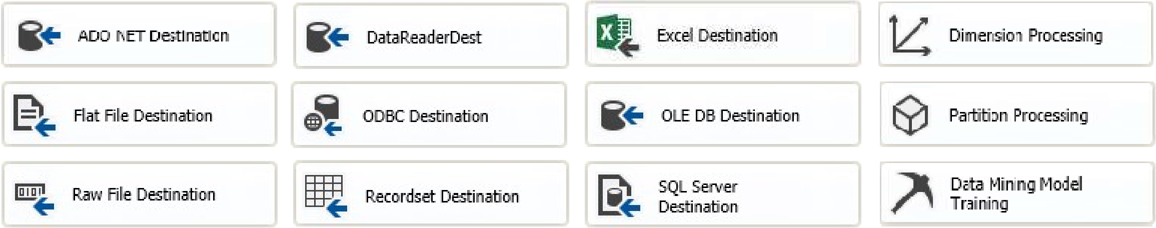
- Sources extract data from:
  - Relational tables and views
  - CDC tables
  - Files



SSIS Data Flow

Destinations

- Destinations load data to:
  - Relational tables and views
  - Data readers and recordsets
  - Analysis Services database objects



ADO NET Destination
DataReaderDest
Excel Destination
Dimension Processing
Flat File Destination
ODBC Destination
OLE DB Destination
Partition Processing
Raw File Destination
Recordset Destination
SQL Server Destination
Data Mining Model Training

SSIS Data Flow

Transformations

- Perform discrete operations
- Categories:
  - Row
  - Rowset
  - Routing and Lookup
  - Business Intelligence
  - Slowly Changing Dimension
  - Scripting
  - Miscellaneous

SSIS Data Flow

Transformations ► Row

- Update column values or create new columns
- Transform each row in the pipeline input

Character Map

- Modify strings, typically for code page changes

Copy Column

- Copy columns to new output columns

Data Conversion

- Data casting

Derived Column

- Define new columns, or override values in an existing column

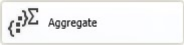
OLE DB Command

- Execute a command against an OLE DB connection manager

SSIS Data Flow

Transformations ► Rowset

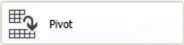
- Create new outputs that can aggregate, sort, sample, pivot or unpivot input data


Aggregate

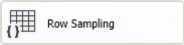
- Summarizes numeric columns

Percentage Sampling

- Samples a random proportion of the data


Pivot

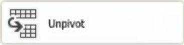
- Pivots (rows to columns)


Row Sampling

- Samples a fixed number of rows

Sort

- Sorts and de-duplicates data


Unpivot

- Unpivots (columns to rows)

SSIS Data Flow

Transformations ► Routing or Lookup

- Split, merge, and join rows, make copies and perform lookup operations

Conditional Split	▪ Uses conditions to route rows to different outputs
Lookup	▪ Lookups a value against a reference set
Merge	▪ Unions two sorted input to one sorted output
Merge Join	▪ Joins two sorted inputs (inner, left or full outer)
Multicast	▪ Duplicates all rows to multiple outputs
Union All	▪ Unions two or more inputs to one non-sorted output

SSIS Data Flow

Transformations ► Business Intelligence

- Execute predictions, cleanse data, and text mining preparation

Data Mining Query	▪ Retrieves a data mining prediction for each row
Fuzzy Grouping	▪ Identifies duplicate rows based on fuzzy matching on external data
Fuzzy Lookup	▪ Identifies duplicate rows based on fuzzy matching on the same set of data
Term Extraction	▪ Extracts English terms for text mining
Term Lookup	▪ Creates custom word lists and statistics for text mining

SSIS Data Flow

Transformations ► Slowly Changing Dimension

- Wizard-based configuration promotes rapid ETL development for dimension packages
- Supports:
  - Type 0 (Fixed Attribute)
  - Type 1 (Changing Attribute)
  - Type 2 (Historical Attribute)
  - Inferred member management
- Automatically constructs the downstream data flow

SSIS Data Flow

Transformations ► Slowly Changing Dimension (Continued)

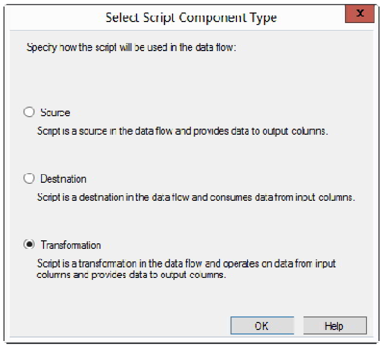
- Note: While the SCD Transformation provides simple and rapid configuration of Type 1 and Type 2 changes, it does not perform well for large volumes of data correlation (> 10,000 rows)

SSIS Data Flow

Transformations ► Scripting

Script Component

- Implements custom transformation logic by using either VB.NET or C#
- Can be configured as a:
  - Source
  - Destination, or
  - Transformation
- Developed by using VSTA
- Can be debugged



Specify how the script will be used in the data flow:

○ Source
Script is a source in the data flow and provides data to output columns.

○ Destination
Script is a destination in the data flow and consumes data from input columns.

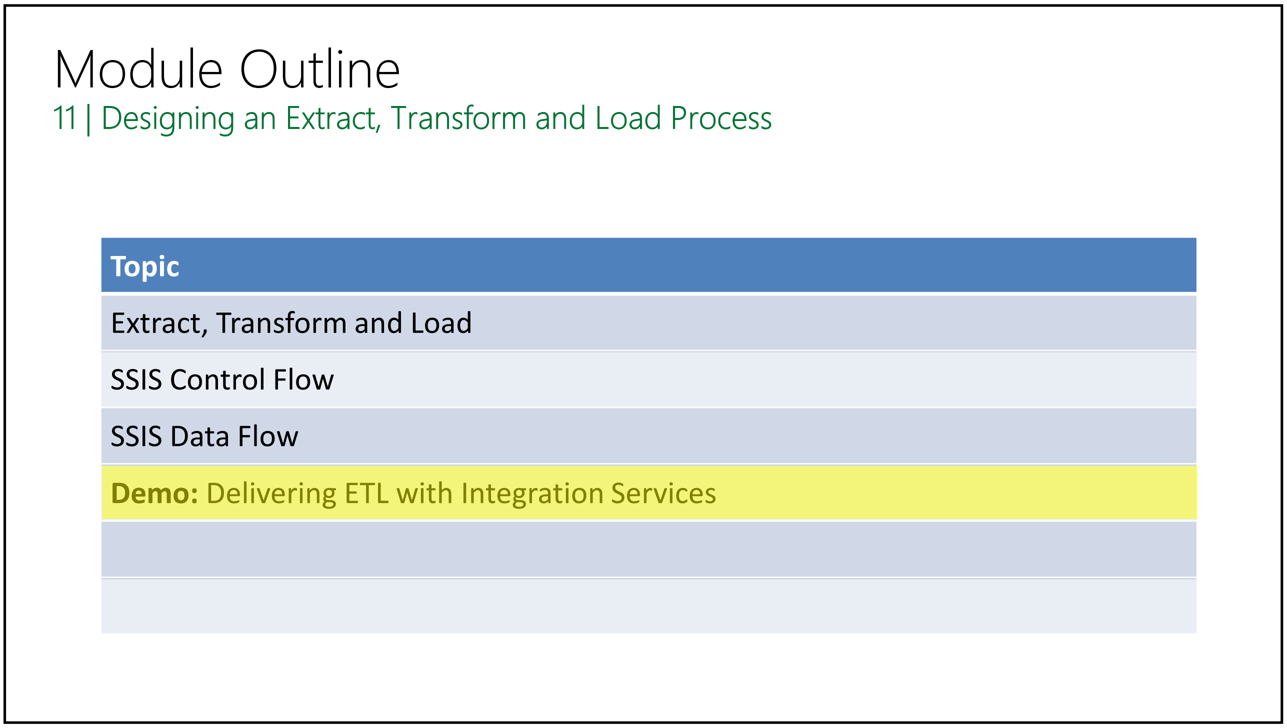
● Transformation
Script is a transformation in the data flow and operates on data from input columns and provides data to output columns.

SSIS Data Flow

Transformations ► Miscellaneous

Audit	Adds audit columns to the output
Cache Transform	Prepares a cache for the Lookup transformation
CDC Splitter	Splits a single input of change rows from a CDC Source into different outputs for insert, update and delete operations
DQS Cleansing	Implements data cleansing by using a DQS knowledge base
Export Column	Creates a binary file for each input row
Import Column	Retrieves binary data into the data flow from a file for each input row
Row Count	Stores the number of rows in a variable

Microsoft



Module Outline

11 | Designing an Extract, Transform and Load Process

Topic
Extract, Transform and Load
SSIS Control Flow
SSIS Data Flow
Demo: Delivering ETL with Integration Services

Demo

Delivering ETL with Integration Services

Demo objectives:

1. Load a dimension table
2. Load a fact table

Microsoft