Lesson Summary

■■ The Star schema is the most common design for a DW.

■■ The Snowflake schema is more appropriate for POC projects.

■■ You should also determine the granularity of fact tables, as well as auditing and lineage

needs.

Lesson Summary

■■ In a dimension, you have the following column types: keys, names, attributes, member

properties, translations, and lineage.

■■ Some attributes form natural hierarchies.

■■ There are standard solutions for the Slowly Changing Dimensions (SCD) problem.

Lesson Summary

■■ Fact tables include measures, foreign keys, and possibly an additional primary key and

lineage columns.

■■ Measures can be additive, non-additive, or semi-additive.

■■ For many-to-many relationships, you can introduce an additional intermediate

dimension.

**Quick Check**

■■ How can SQL Server help you with values for your surrogate keys?

**Quick Check Answer**

■■ SQL Server can autonumber your surrogate keys. You can use the IDENTITY property

or sequence objects.

Lesson Summary

■■ In this lesson, you learned about implementing a data warehouse.

■■ For a data warehouse database, you should use the Simple recovery model.

■■ When creating a database, allocate enough space for data files and log files to prevent

autogrowth of the files.

■■ Use surrogate keys in dimensions in which you expect SCD Type 2 changes.

■■ Use computed columns.

*Exam Tip*

Opt for an integer autonumbering surrogate key as the clustered primary key for all DW

tables, unless there is a really strong reason to decide otherwise.

*IMPORTANT* **Minimize Usage of Nonclustered Indexes in a DW**

Analyze the need for every single nonclustered index in a DW thoroughly. Never create a

nonclustered index in a DW without a good reason.

*Exam Tip*

Unicode compression is applied automatically when you apply either row or page compression.

**Quick Check**

1. How many columnstore indexes can you have per table?

2. Should you use page compression for OLTP environments?

**Quick Check Answers**

1. You can have one columnstore index per table.

2. No, you should use age compression only for data warehousing environments.

Lesson Summary

■■ In this lesson, you learned how to optimize data warehouse query performance.

■■ In a DW, you should not use many nonclustered indexes.

■■ Use small, integer surrogate columns for clustered primary keys.

■■ Use indexed views.

■■ Use columnstore indexes and exploit batch processing.

■■ **Partition function** This is an object that maps rows to partitions by using values

from specific columns. The columns used for the function are called *partitioning columns*.

A partition function performs logical mapping.

■■ **Partition scheme** A partition scheme maps partitions to filegroups. A partition

scheme performs physical mapping.

■■ **Aligned index** This is an index built on the same partition scheme as its base table.

If all indexes are aligned with their base table, switching a partition is a metadata operation

only, so it is very fast. Columnstore indexes have to be aligned with their base

tables. *Nonaligned indexes* are, of course, indexes that are partitioned differently than

their base tables.

■■ **Partition elimination** This is a Query Optimizer process in which SQL Server accesses

only those partitions needed to satisfy query filters.

■■ **Partition switching** This is a process that switches a block of data from one table or

partition to another table or partition. You switch the data by using the ALTER TABLE

T-SQL command. You can perform the following types of switches:

■■ Reassign all data from a nonpartitioned table to an empty existing partition of a

partitioned table.

■■ Switch a partition of one partitioned table to a partition of another partitioned

table.

■■ Reassign all data from a partition of a partitioned table to an existing empty nonpartitioned

table.

*Exam Tip*

Make sure you understand the relationship between columnstore indexes and table partitioning

thoroughly.

**Quick Check**

■■ How many partitions can you have per table?

**Quick Check Answer**

■■ In SQL Server 2012, you can have up to 15,000 partitions per table.

Lesson Summary

■■ Table partitioning is extremely useful for large fact tables with columnstore indexes.

■■ Partition switch is a metadata operation only if an index is aligned with its base table.

■■ You can add lineage information to your dimensions and fact tables to audit changes

to your DW on a row level.

*Exam tip*

Plan data movements carefully, and consider the benefits as well as the shortcomings of

the Import and Export Wizard.

**Quick Check**

1. What is the SQL Server Import and Export Wizard?

2. What is the principal difference between simple and complex data movements?

**Quick Check Answers**

1. The Import and Export Wizard is a utility that provides a simplified interface

for developing data movement operations where data is extracted from a

source and loaded into a destination, without the need for any transformations.

2. In simple data movements, data is copied from one data store into another one

unmodified, whereas in complex data movements, data is modified (transformed)

before being loaded into the destination data store.

*Exam tip*

You should understand the difference between the option to copy data from one or more

tables and the option to use a query to specify the data to transfer, so that you can select the

best option in a particular situation.

For instance, copying from tables and views allows multiple data flows but no additional

restrictions, whereas copying data by using a query allows restrictions to be specified, but

only allows a single data flow.

*Exam Tip*

Even though this may not be apparent from the file name, SSIS package definitions are

stored in XML format and can be reviewed and edited with most text editing tools.

Lesson Summary

■■ The SQL Server Import and Export Wizard can be used for simple data movement

operations.

■■ The wizard allows you to create the destination database.

■■ Multiple objects can be transferred in the same operation.

■■ If the destination objects do not already exist, they can be created by the process.

■■ The SSIS package created by the wizard can be saved and reused.

Lesson Summary

■■ SSIS projects are developed by using SSDT, a specialized version of Visual Studio.

■■ SSDT provides the complete integrated development environment (IDE) required for

efficient development of SSIS packages.

■■ The SSIS toolbox is context-aware and will either allow you access to control flow tasks

or data flow components, depending on whether you are designing the control flow or

the data flow.

*Exam Tip*

Never confuse the control flow with the data flow. Control flow determines the operations

and the order of their execution. Data flow is a task in the control flow that determines the

ETL operation.

*Exam Tip*

SSIS parameterization represents a vital element of SSIS development; its principal role

might not be apparent until the SSIS solution is deployed, but it must be considered from

the very start of development.

**Quick Check**

1. What is a control flow?

2. What is a data flow?

**Quick Check Answers**

1. In SSIS packages, the control flow defines the tasks used in performing data

management operations; it determines the order in which these tasks are executed

and the conditions of their execution.

2. In SSIS packages, the data flow is a special control flow task used specifically in

data movement operations and data transformations.

Lesson Summary

■■ Existing SSIS packages can be added to SSIS projects in SQL Server Data Tools (SSDT).

■■ Control flows contain the definitions of data management operations.

■■ Control flows determine the order and the conditions of execution.

■■ SSIS package settings can be parameterized, which allows them to be changed without

direct access to SSIS package definitions.

*Exam Tip*

Become familiar with all standard connection managers; learn about their purpose, usability,

and the benefits and possible drawbacks of their use. Using inappropriate connection

managers might prevent you from completing your work or might cause you to run out

of time.

*Note* **ADO.NET Connect ion Mana ger**

When using stored procedures or parameterized queries against a SQL Server database

in an Execute SQL Task, consider using the ADO.NET data provider, because it provides

a much better usability and manageability experience compared to the OLE DB data

provider:

■■ With ADO.NET, you can use parameter names in queries, instead of question

marks as parameter placeholders.

■■ When you are using stored procedures, ADO.NET allows you to set the query

type appropriately (for example, by setting the IsQueryStoredProcedure property

to True) —you provide the name of the procedure and define the parameters

in the Task Editor (in any order, with or without the optional parameters),

and the query statement is assembled automatically.

■■ ADO.NET has better support for data types compared to OLE DB (for example,

Xml, Binary, and Decimal data types are not available in OLE DB, and there are

problems with the SQL Server large object data types VARCHAR(MAX) and

VARBINARY(MAX)).

*Important* **CONNECTION MANAGER NAMES**

If a package connection manager and a project connection manager use the same name,

the package connection manager overrides the project connection manager.

**Quick Check**

1. What is the purpose of connection managers in SSIS at design time?

2. What is the purpose of connection managers in SSIS at run time?

3. How does connection manager scope affect their use?

**Quick Check Answers**

1. At design time, connection managers are used by the SSIS developer to configure

a connection to a data source.

2. At run time, connection managers are used by the SSIS engine to establish connections

to data sources.

3. A project-scoped connection manager is available to all packages of a particular

SSIS project, whereas a package-scoped connection manager is only available

to the package in which it was created.

*Exam Tip*

Learn about SQL Server security best practices and recommendations, and think about how

to implement parameterization of sensitive settings such as connection strings to keep

your environment secure, while at the same time utilizing the benefits of parameterization.

*EXAM TIP*

Make sure you understand the difference between package and project connection managers

and how naming them affects their usability.

Lesson Summary

■■ Connection managers are used to establish connections to data sources.

■■ Different data sources require different types of connection managers.

■■ The usability of a connection manager within an SSIS project or an SSIS package is

determined by its scope.

***Data cleansing*** Unwanted or invalid pieces of data are discarded or replaced with

valid ones. Many diverse operations fit this description—anything from basic cleanup

(such as string trimming or replacing decimal commas with decimal points) to quite

elaborate parsing (such as extracting meaningful pieces of data by using Regular

Expressions).

■■ ***Data normalization*** In this chapter, we would like to avoid what could grow into a

lengthy debate about what exactly constitutes a scalar value, so the simplest definition

of normalization would be the conversion of complex data types into primitive

data types (for example, extracting individual atomic values from an XML document or

atomic items from a delimited string).

■■ ***Data type conversion*** The source might use a different type system than the

destination. Data type conversion provides type-level translation of individual values

from the source data type to the destination data type (for example, translating a .NET

Byte[] array into a SQL Server VARBINARY(MAX) value).

■■ ***Data translation*** The source might use different domains than the destination.

Translation provides a domain-level replacement of individual values of the source domain

with an equivalent value from the destination domain (for example, the character

“F” designating a person’s gender at the source is replaced with the string “female”

representing the same at the destination).

■ ***Data validation*** This is the verification and/or application of business rules against

individual values (for example, “a person cannot weigh more than a ton”), tuples (for

example, “exactly two different persons constitute a married couple”), and/or sets (for

example, “exactly one person can be President of the United States at any given time”).

■■ ***Data calculation* and *data aggregation*** In data warehousing, specifically, a common

requirement is to not only load individual values representing different facts or

measures, but also to load values that have been calculated (or pre-aggregated) from

the original values (for example, “net price” and “tax” exist at the source, but “price

including tax” is expected at the destination).

■■ ***Data pivoting* and *data unpivoting*** Source data might need to be restructured or

reorganized in order to comply with the destination data model (for example, data in

the entry-attribute-value (EAV) might need to be restructured into columns or viceversa).

*Exam Tip*

You should have a very good understanding of what constitutes data transformations.

Knowing whether the data needs to be transformed or not will help you determine not

only which tasks are appropriate in your work, but also how to define the order and the

conditions of their execution.

*Exam Tip*

Become familiar with all standard SSIS tasks: understand their intended purpose and their

capabilities as well as their limitations, and always think hard about whether the business

problem can be solved with just a single task or with a system of two or more tasks.

File System task This task provides operations on file system objects (files and folders), such as

copying, moving, renaming, deleting objects, creating folders, and setting object

attributes.

FTP task This task provides operations on file system objects on a remote file store via the

File Transfer Protocol (FTP), such as receiving, sending, and deleting files, as well

as creating and removing directories.

Typically, the FTP task is used to download files from the remote file store to be

processed locally, or to upload files to the remote store after they have been processed

(or created) in the SSIS solution.

Web Service task This task provides access to web services; it invokes web service methods, receives

the results, and stores them in an SSIS variable or writes them to a file

connection.

XML task This task provides XML manipulation against XML files and XML data, such as

validation (against a Document Type Definition or an XML Schema), transformations

(using XSLT), and data retrieval (using XPath expressions). It also supports

more advanced methods, such as merging two XML documents and comparing

two XML documents, the output of which can consequently be used to create a

new XML document (known as a DiffGram).

Data Profiling task This task can be used in determining data quality and in data cleansing. It can be

useful in the discovery of properties of an unfamiliar data set.

You will learn more about the Data Profiling Task in Chapter 17.

*Note* **THE FILE SYSTEM TASK**

The operations provided by the File System task target individual file system objects. To

use it against multiple objects, you should use the Foreach Loop Container (discussed later

in this chapter).

Execute Package task This task executes other SSIS packages, thus allowing the distribution of programmatic

logic across multiple SSIS packages, which in turn increases the reusability

of individual SSIS packages and enables a more efficient division of labor

within the SSIS development team.

You will learn more about the Execute Package task in Chapter 6, ”Enhancing

Control Flow.”

Execute Process task This task executes external processes (that is, processes external to SQL Server).

The Execute Process task can be used to start any kind of Windows application;

however, typically it is used to execute processes against data or data stores

that cannot or do not need to be more closely integrated with the SSIS process

but still need to be performed as part of it.

Message Queue task This task is used to send and receive messages to and from Microsoft Message

Queuing (MSMQ) queues on the local server.

Typically, the Message Queue task would be used to facilitate communication

with other related processes that also utilize MSMQ, such as other SSIS processes

or external processes.

With MSMQ queues, you can distribute your automated data management

processes

across the entire enterprise.

Send Mail task The task allows the sending of email messages from SSIS packages by using the

Simple Mail Transfer Protocol (SMTP).

Typically, the Send Mail task would be used to send information or files,

although it could also be used to send messages regarding its execution.

You will learn more about notifications related to SSIS solution deployment in

Chapter 10, “Auditing and Logging” and Chapters 11 and 12.

WMI Data Reader task This task provides access to Windows Management Instrumentation (WMI)

data, allowing access to information about the environment (such as server

properties, resource properties, and performance counters).

Typically, the WMI Data Reader task would be used to gather WMI data for

further use (to be processed and loaded into a database, for example), or to

monitor the state of the environment in order to determine the behavior of SSIS

processes or SSIS tasks (whether to run them at all or to configure them dynamically

in line with the current state of the environment, for example).

WMI Event Watcher task This task provides access to WMI events.

Typically, the WMI Event Watcher task would be used to trace events in the

environment, and based on them to control the execution of SSIS processes or

SSIS tasks (for example, to detect the addition of files to a specific folder in order

to initiate the SSIS process that relies on these files).

Expression task This task is used in the workflow to process variables and/or parameters and to

assign the results to other variables used by the SSIS process.

Typically, the Expression task is used to assign values to variables without the

overhead of using the Script task for the same purpose.

CDC Control task This task controls the life cycle of SSIS packages that rely on the SQL Server

2012 Change Data Capture (CDC) functionality. It provides CDC information

from the data source to be used in CDC-dependent data flows.

Bulk Insert task This task allows the loading of data from formatted text files into a SQL Server database

table (or view); the data is loaded unmodified (because transformations are not

supported), which means that the loading process is fast and efficient. Additional settings

(such as using table lock, disabling triggers, and disabling check constraints) are

provided to help reduce contention even further.

Execute SQL task This task executes SQL statements or stored procedures against a supported data store.

The task supports the following data providers: EXCEL, OLE DB, ODBC, ADO, ADO.NET,

and SQLMOBILE, so keep this in mind when planning connection managers.

The Execute SQL task supports parameters, allowing you to pass values to the SQL

command dynamically.

Also see the note about ADO.NET connection managers in Lesson 1, earlier in this

chapter.

Data flow task This task is essential to data movements, especially complex data movements, because

it provides all the elements of ETL (extract-transform-load); the architecture of the data

flow task allows all of the transformations to be performed in flight and in memory,

without the need for temporary storage.

Chapter 5, “Designing and Implementing Data Flow,” is dedicated to this most vital

control flow task.

*Important* **THE BULK INSERT TASK AND PERMISSIONS**

The Bulk Insert task requires the user who is executing the SSIS package that contains this

task to be a member of the sysadmin fixed server role. If your security policy does not

allow the SSIS service account to have elevated permissions, consider using a different account

when connecting to the destination server.

*Note* **THE DATA FLOW TASK**

In contrast to pretty much any other task, the operational characteristics of the data flow

task are not defined solely through properties and parameters; the actual operation of a

data flow task is defined by a system of data flow components that it encapsulates.

Transfer Database task Use this task to copy or move a database from one SQL Server instance to

another or create a copy of it on the same server. It supports two modes of

operation:

■■ In online mode, the database is transferred by using SQL Server

Management Objects (SMO), allowing it to remain online for the

duration of the transfer.

■■ In offline mode, the database is detached from the source instance,

copied to the destination file store, and attached at the destination

instance, which takes less time compared to the online mode, but

for the entire duration the database is inaccessible.

Transfer Error Messages task Use this task to transfer user-defined error messages from one SQL Server

instance to another; you can transfer all user-defined messages or specify

individual ones.

Transfer Jobs task Use this task to transfer SQL Server Agent Jobs from one SQL Server instance

to another; you can transfer all jobs or specify individual ones.

Transfer Logins task Use this task to transfer SQL Server logins from one SQL Server instance to

another; you can transfer all logins, logins mapped to users of one or more

specified databases, or individual users.

You can even copy security identifiers (SIDs) associated with the logins. The

built-in sa login cannot be transferred.

Transfer Master Stored

Procedures task

Use this task to transfer user-defined stored procedures (owned by dbo)

from the master database of one SQL Server instance to the master database

on another SQL Server instance; you can transfer all user-defined

stored procedures or specify individual ones.

Transfer SQL Server Objects

task

Use this task to transfer objects from one SQL Server instance to another;

you can transfer all objects, all objects

Back Up Database task Use this task in your maintenance plan to automate full, differential,

or transaction log backups of one or more system and/or user databases.

Filegroup and file level backups are also supported.

Check Database Integrity task Use this task in your maintenance plan to automate data and index

page integrity checks in one or more system and/or user databases.

Execute SQL Server Agent Job task Use this task in your maintenance plan to automate the invocation of

SQL Server Agent Jobs to be executed as part of the maintenance plan.

Execute T-SQL Statement task Use this task in your maintenance plan to execute Transact-SQL scripts

as part of the maintenance plan.

You should not confuse the very basic Execute T-SQL Statement Task

with the more advanced Execute SQL Task described earlier in this

lesson. The Execute T-SQL Statement Task only provides a very basic

interface, which will allow you to select the connection manager and

specify the statement to execute; parameters, for instance, are not

supported in this task.

History Cleanup task Use this task in your maintenance plan to automate the purging of

historical data about backups and restore operations, as well as SQL

Server Agent and maintenance plan operations on your SQL Server

instance.

Maintenance Cleanup task Use this task in your maintenance plan to automate the removal of

files left over by maintenance plan executions; you can configure the

task to remove old backup files or maintenance plan text reports.

Notify Operator task Use this task in your maintenance plan to send email messages to SQL

Server Agent operators.

Rebuild Index task Use this task in your maintenance plan to automate index rebuilds for

one or more databases and one or more objects (tables or indexed

views).

Reorganize Index task Use this task in your maintenance plan to automate index reorganizations

for one or more databases and one or more objects (tables or

indexed views).

Shrink Database task Use this task in your maintenance plan to automate database shrink

operations.

Update Statistics task Use this task in your maintenance plan to automate updates of statistics

for one or more databases and one or more objects (tables or

indexed views).

*Important* **THE SHRINK DATABASE TASK**

Shrinking the database will release unused space from the database files back to the operating

system. To achieve this, SQL Server will probably need to rearrange the contents of

the file in order to place unused portions at the end of the file. This might cause fragmentation,

which in turn may have a negative impact on query performance. In addition, a

large modification operation against the database (such as a large insert) performed after

the shrinking of the database might require more space than is available, which will require

the database to grow automatically. Depending on the space requirements, an auto-grow

operation could take a long time, which in turn could cause the modification operation to

reach the timeout and roll back, which could effectively render the server unresponsive.

Therefore, you should avoid shrinking databases, and—most importantly—never automate

the shrinking process unless absolutely necessary, and even then you should make sure to

reserve enough free space to avoid auto-grows!

*Note* **WHEN TO USE THE SCRIPT TASK**

Avoid resorting to the Script task until you have eliminated all possibilities of solving the

business problem by using one or more standard tasks.

Compared to the Script task, standard tasks provide a much better deployment and maintenance

experience. The developers following in your footsteps and taking your work over

from you might find it much easier to understand a process that uses standard tasks, however

complex, than to ”decode” a lengthy Script task.

*EXAM TIP*

Although the typical “procedural” approach to programming, in which a single item is processed

at a time, should generally be avoided in favor of ”set-oriented” programming, in which

an entire set of items is processed as a single unit of work, some operations still require the

procedural approach— to be executed in a loop.

Study all three containers in SSIS well to understand the differences between them, so that

you can use looping appropriately in your SSIS solutions.

For Loop container This container executes the encapsulated tasks repeatedly, based on an

expression—the looping continues while the result of the expression is

true; it is based on the same concept as the For loop in most programming

languages.

Foreach Loop container This container executes the encapsulated tasks repeatedly, per each item

of the selected enumerator; it is based on the same iterative concept as the

For-Each loop in most contemporary programming languages.

The Foreach Loop container supports the following enumerators: the

ADO enumerator, the ADO.NET Schema Rowset enumerator, the File

enumerator, the Item enumerator, the Nodelist enumerator, and the

SMO enumerator.

Sequence container This container has no programmatic logic other than providing structure

to encapsulate tasks that form a logical unit, to provide a scope for SSIS

variables to be accessible exclusively to a specific set of tasks or to provide

a transaction scope to a set of tasks.

**Quick Check**

1. What tasks is the Foreach Loop container suited for?

2. How can the current item or its properties be made available to the tasks inside

a Foreach Loop container?

3. Is it possible to change the settings of an SSIS object at run time?

**Quick Check Answers**

1. It is suited for executing a set of operations repeatedly based on an enumerable

collection of items (such as files in a folder, a set of rows in a table, or an

array of items).

2. You can assign the values returned by the Foreach Loop container to a variable.

3. Yes, it is. Every setting that supports expressions can be modified at run time.

Lesson Summary

■■ A rich collection of tasks supporting the most common data management operations

is provided by the SSIS design model.

■■ Control flow is defined by precedence constraints that determine the order and conditions

of execution.

■■ Tasks representing logical units of work can be grouped in containers.

■■ Loop containers allow a unit of work to be executed repeatedly.

**Quick Check**

1. Can SSIS execution be redirected from one task to another?

2. Can multiple precedence constraints lead from the same preceding task?

3. What is the principal difference between a success constraint and a completion

constraint?

**Quick Check Answers**

1. Yes, by using different conditions in precedence constraints, the order of

execution can be directed to the following tasks in one branch or to another

branch.

2. Yes, multiple precedence constraints can lead from a single task to the following

tasks, but only one precedence constraint can exist between two distinct

tasks.

3. A success constraint will only allow the process to continue to the following

task if the preceding task completed successfully, whereas a completion

constraint will allow the process to continue as soon as the preceding task has

completed, regardless of the outcome.

Lesson Summary

■■ Precedence constraints determine the order of execution and the conditions that must

be met for the process to either continue or stop.

■■ Precedence constraints can even be used to allow the process to recover from failures.

*EXAM TIP*

At run time, the data flow task builds an execution plan from the data flow, and the data

flow engine executes the plan.

*IMPORTANT* **USING THE NEW ODBC SOURCE OVER OLE DB SOURCE FOR SQL SERVER**

In previous versions of SQL Server, I always used the OLE DB provider when working with

SQL Server as a data source or data destination. In 2011, Microsoft released a statement

announcing that it is aligning with ODBC for native relational data access and that SQL

Server 2012 will be the last release to support OLE DB. Based on this, I suggest that you use

the new ODBC source adapter when creating your SSIS packages.

*EXAM TIP*

Many SSIS objects have a ValidateExternalMetadata property that you can set to False if

the object being referenced (such as a table) does not exist when the package is being

designed. This property is most commonly used for source or destination adapters when,

for example, a destination table is created during package execution.

*EXAM TIP*

You can configure the OLE DB destination adapter (and now, with SQL Server 2012, also the

ODBC destination adapter) to insert data from the data flow through bulk batches of data,

instead of one row at a time. To use this destination-optimization technique for OLE DB,

edit the OLE DB destination and set the data access mode to Table Or View—Fast Load. For

ODBC, edit the ODBC destination and set the data access mode to Table Name—Batch. If

the destinations are not configured with fast load or batch load, only one row at a time will

be inserted into the destination table.

**Quick Check**

■■ You need to migrate a user-created Microsoft Access database to SQL Server, but

the data flow SSIS Toolbox does not contain an Access source adapter. How do you

import this data into SQL Server?

**Quick Check Answer**

■■ Although it is not listed in the toolbox, Access is one of the many database sources

and destinations that SSIS works with. To extract data from Access, you have three

choices: You can use the new ODBC source adapter, create a package connection

to the Microsoft Jet OLE DB Provider, or use the OLE DB source adapter.

*Important* **Optimizing Packages Using Fast Parse**

Fast Parse has limited functionality, because it works only for specific data types. But if

you have a lot of columns of date, time, or integer data types or very large files, Fast Parse

is the fastest method for importing data files. On the projects we have worked on, we got

more than 20 percent faster loading times, so use it when you have to optimize your packages.

Note that Fast Parse is also available in the Data Conversion transformation.

*Note* **Undo and Redo Operations Now Su pported in SQL Server 2012**

Versions of SQL Server before SQL Server 2012 did not support Undo and Redo operations.

After you performed an operation, you could not undo it. Now, in SSIS 2012, you can use

the Undo and Redo functionality when creating a data flow.

Lesson Summary

■■ Use appropriate data source or data destination adapters.

■■ Always extract only the columns you need.

■■ Use Fast Load or Batch mode when inserting data by using an ODBC or OLE DB destination

adapter.

■■ Use a Raw File destination if you have to temporarily store data to be used by SSIS

later.

There are three types of blocking:

■■ In *non-blocking transformations*, each row is passed through the transformation without

any waits.

■■ A *partial-blocking transformation* waits until a sufficient number of rows is stored and

then it passes them through.

■■ In a *blocking transformation*, all rows must be read into the transformation before the

transformation can pass them through.

Derived Column Creates or replaces a column for each row based on a

specified SSIS expression. This is the most often used

logical row-level transformation because it enables

the replacement of column values or the creation of

new columns based on existing columns, variables, and

parameters.

N

Export Column Exports binary large objects (BLOB) columns, one row at a

time, to a file.

N

Import Column Loads binary files such as images into the pipeline; intended

for a BLOB data type destination.

N

Row Count Tracks the number of rows that flow through the transformation

and stores the number in a package variable after

the final row.

N

*EXAM TIP*

When inserting data into data warehouse tables, you should check for NULL inside each

column and replace it with a value that represents an “unknown” or default value. With the

Derived Column transformation, you can check for NULL by using an SSIS conditional expression.

You replace the existing column col1 value with the expression ISNULL(col1) ? 0 : col1.

The expression will check first to see whether col1 is NULL and if it is, the expression will put

the value 0 for the column col1; if it is not NULL, the col1 column value will stay the same. In

SQL Server 2012 SSIS, you can do this more elegantly by using the new REPLACENULL (col1, 0)

function.

On the other hand, if you would like to put NULL inside the column, you must use the appropriate

SSIS NULL function for the specific data type. For example, if you would like to

store NULL inside a 4-byte signed integer column, you must use the NULL(DT\_I4) function.

CDC Splitter Splits a single flow of changed rows from the CDC source

component into multiple data flows based on the type of the

source data change (that is, whether it is an insert, update,

or delete operation). CDC Splitter routes the data based on

the *\_\_$operation* column into three possible outputs. this

transformation is like a specific version of the Conditional

Split transformation that automatically handles the standard

values of the *\_\_$operation* column.

N

Conditional Split Routes or filters data based on a Boolean expression to one

or more outputs, from which each row can be sent out only

one output path.

N

Lookup Performs a lookup operation between a current row and an

external dataset on one or more columns. Additional columns

can be added to the data flow from the external dataset.

N

Merge Combines the rows of two similar sorted inputs, one on top of

the other, based on a defined sort key.

P

Merge Join Joins the rows of two sorted inputs based on a defined join

column or columns, adding columns from each source.

P

Multicast Generates one or mode identical outputs, from which every

row is sent out every output. This transformation creates a

logical copy of the data.

N

Union All Combines one or more similar inputs, stacking rows one on

top of another, based on matching columns. The number of

rows in the output of Union All is the combined row counts of

all the inputs.

P

*EXAM TIP*

The Merge Join transformation can match more than one row across the join columns. It

behaves the same as the T-SQL Join clause; you can specify an inner join, a full outer join, or

a left outer join. Remember that this transformation can be used only if both source inputs

are sorted on the same column or columns.

**Quick Check**

■■ What is the difference between the Union All and the Merge transformation?

**Quick Check Answer**

■■ The Merge transformation is similar to Union All, but with Merge, the sources have

to be sorted and the sort position is preserved.

Aggregate Associates rows based on defined grouping and generates

aggregations such as SUM, MAX, MIN, and COUNT.

B

Percent Sampling Filters the input rows by allowing only a defined percent to

be passed to the output path.

N

Pivot Takes multiple input rows and pivots the rows to generate an

output with more columns based on the original row values.

P

Row Sampling Generates a fixed number of rows, sampling the data from

the entire input, no matter how much larger than the defined

output the input is.

B

Sort Orders the input based on defined sort columns and sort

directions. The Sort transformation also allows the removal

of duplicates across the sort columns.

B

Unpivot Takes a single row and generates multiple rows, moving

column values to the new row based on defined columns.

P

*Note* **Remember Which Transformations Ar e Fully Blocked**

Remember which transformations are fully blocked and try to use them only when absolutely

necessary, because they often require more memory and processor capacity. If you

are aggregating or sorting a data source input that will not fit into the server memory, the

performance will degrade by a factor of 100, because swapping to disk will occur.

*IMPORTANT* **Using the New Pivot Component**

When using the Pivot component, you must explicitly name all new columns based on

the distinct values of the column that is specified as the Pivot Key. If you select the Ignore

Un-Matched Pivot Key Values And Report Them After DataFlow Execution check box, the

component will not produce an error if some of the possible values were not specified as

new columns.

You cannot create “dynamic” pivoting (automatically adding new columns) by using this

component, because the data flow engine must have exact information about each column

that will be present in the data flow. The only workaround is to store all possible values in a

single column as XML or as a delimited set of values. In this way, you can create “dynamic”

pivoting by using the Script component.

DQS Cleansing Validates rows by automatically performing

data cleansing using an existing

knowledge base in Data Quality

Services (DQS).

P

OLE DB Command Performs database operations such as

updates or deletions, one row at a time,

based on mapped parameters from

input rows.

N

Slowly Changing Dimension Generates transformations necessary

to support loading dimension tables

in data warehouse scenarios. This

transformation handles SCD (Slowly

Changing Dimension) Type 1 and Type

2 and also has support for inferred

members. Chapter 7 focuses on this

transformation.

N

Data Mining Query Applies input rows against a data mining

model for prediction.

P

Fuzzy Grouping Performs de-duplication based on

similarity of string values in selected

columns.

B

Fuzzy Lookup Joins a data flow input to a reference

table based on column similarity. The

Similarity Threshold setting specifies

the closeness of allowed matches—a

high setting means that matching values

are close in similarity.

B

Script Component Applies custom .NET scripting

capabilities against rows, columns,

inputs, and outputs in the data

flow pipeline. This is the most

powerful component. Chapter 19,

“Implementing

Custom Code in

SSIS Packages” looks at some of its

possibilities.

N

Term Extraction Analyzes text input columns for

English-language nouns and noun

phrases.

P

Term Lookup Analyzes text input columns against a

user-defined set of words for association.

P

*EXAM TIP*

With the Script Component, you can apply almost any kind of transformation to the data

flow pipeline. Its application can range from replacing multiple expressions by using multiple

Derived Column components to complex transformations using .NET code. You can use

Microsoft Visual Basic .NET (VB.NET) and Microsoft Visual C# as programming languages

with this component. SQL Server 2012 uses Microsoft Visual Studio Tools for Applications

(VSTA ) 3.0 as the environment in which to write the scripts, and the component supports

Microsoft .NET 4.

Lesson Summary

■■ Remember which transformations are non-blocking, partly-blocking, and blocking.

■■ Use the Resolve References dialog box to solve mapping errors.

■■ Use Derived Column transformation to add new columns or replace the value in

existing ones.

*Full cache* is the default mode. In this mode, the database is queried once during the

pre-execute phase of the data flow, and the entire reference set is stored into memory.

Lookup operations will be very fast during their execution, but you need to have

enough memory to fit the needed dataset.

■■ *Partial cache* means that the lookup cache is empty at the beginning of the data

flow. When a new row comes in, the Lookup transformation checks its cache for the

matching values. If no match is found, it queries the database. If the match is found

at the database, the values are cached so they can be used the next time a matching

row comes in. In SQL Server 2008 and SQL Server 2012, it is also possible to set up the

Miss Cache feature, which allows you to allocate a certain percentage of your cache to

remembering rows that had no match in the database.

■■ The *no cache* mode will store only the last matched row, which means that the Lookup

transformation will query the database for each row. This mode should be avoided in

data warehouse scenarios when you are loading a fact table, because it will work too

slowly.

*EXAM TIP*

The most important setting of the Lookup transformation is the cache mode. In data warehouse

environments, try to have enough memory to use the Full Cache option, because it

will perform very fast. Remember also that the Lookup transformation is a non-blocking

transformation; this is also important when you design your ETL strategy. The only drawback

is that the whole dataset has to be written into memory first, in the pre-execute

phase of the data flow. One thing to note is that the lookup will not swap memory out to

disk, so your data flow will fail if you run out of memory.

*IMPORTANT* **Lookup Transformation Case Sensitivity**

The lookups performed by the Lookup transformation are case sensitive. Use either the

Character Map transformation to convert the data to uppercase or lowercase and use appropriate

SQL functions such as UPPER or LOWER for the referenced dataset.

**Quick Check**

■■ What is the difference between the Lookup and Merge Join transformations?

**Quick Check Answer**

■■ The Lookup transformation does not need a sorted input; it is a non-blocking

transformation, and in cases when more matching rows exist from the referenced

dataset, only the first one will be retrieved. This means that the transformation will

never retrieve more rows than exist in the input rows. With the Merge Join transformation

more rows can be retrieved, because all matching data is retrieved.

*EXAM TIP*

A cache is created in a standard data flow, which means that you can use any data source

that SSIS can connect to as a source for the Lookup transformation. With the Cache connection

manager, you are no longer bound to an OLE DB connection to create a lookup

dataset.

Lesson Summary

■■ Use sorting on the database layer as much as possible.

■■ When joining large tables, consider doing so on the database layer.

■■ Insert the data that needs to be updated into a temporary table and then perform a

set-based update using SQL.

*Important* **PARAMETERS**

Always parameterize properties that cannot be determined any other way, except for being

set by the administrator (such as file paths, server and database names, or entire connection

strings).

Avoid parameterizing properties that can also be determined automatically (such as the

number of CPUs); furthermore, do not parameterize automatically determinable properties

that, if set by the administrator, could cause the solution to fail (such as data source

queries) or cause it to underperform (such as batch sizes).

*Important* **REUSABILITY**

Maximize data reusability: avoid retrieving configuration data from an external store (such

as a database or a file) more than once, unless any changes to this data that could occur

during execution must be reflected in the SSIS package (such as the amount of available

system memory).

*NOTE* **SSIS PROPERTY PATHS**

As you will learn in Chapter 12, “Executing and Securing Packages,” the values of SSIS

object properties can be set explicitly upon SSIS package execution; however, one needs

to be intimately familiar with the internal structure of the package in order to do so.

Compared to property paths, SSIS parameters provide a much more transparent alternative:

they allow the developer to decide which settings will actually be configurable,

and they allow the administrator to configure the package without the need for any

detailed knowledge of its internal organization.

*EXAM TIP*

The intended purpose of SSIS parameters is to supply values that either should or must be

determined outside the SSIS process—mostly because they cannot be determined automatically.

The pivotal purpose of SSIS variables, on the other hand, is to improve reusability

and optimize data retrieval. SSIS variables are therefore not subject to the same sort of

restrictions as SSIS parameters.

*Exam Tip*

You should be familiar with the available SSIS variable data types, understand how and

when to use them in your variables, and also understand their limitations very well (especially

with respect to numeric data types and their decimal precision).

*Important* **MULTIPLE RESULT SETS**

If the query (such as one used by an Execute SQL Task) returns more than one row set, only

the first row set is returned and the rest are discarded. If multiple results are required, use

the FOR XML directive in your query to return the result as an XML document and assign it

to an SSIS variable of the Object or String data type.

*Note* **REUSING ROW SETS IN SSIS**

A lot of additional programmatic logic might be required in your SSIS package to consume

row sets stored in SSIS variables, especially when they are multiple row sets stored

in a single variable. A best practice is to use standard, built-in tasks and components (for

example, by splitting up the processing of multiple data sets into as many individual data

flows or by using nested Foreach Loop containers). Using standard tasks and components

simplifies the development (including debugging) and deployment as well as the maintenance

of your SSIS solutions.

If, on the other hand, none of the standard methods seem to be appropriate for the requirements

of a particular case, consider using the Script Task or the Script Component to

extend the SSIS functionalities, or even go as far as to develop your own custom control

tasks or custom data flow components.

*Exam Tip*

To understand variable behavior, you really need to understand variable scope—variable

accessibility inside the SSIS object hierarchy.

Scope allows you to restrict variable accessibility to a specific branch of the package hierarchy,

and it also allows you to effectively “override” variables when necessary.

**Quick Check**

1. Which SSIS objects can access container-scoped variables?

2. How many namespaces are available for SSIS variables?

3. How can you modify the value of a system variable?

**Quick Check Answers**

1. Container-scoped variables are only accessible to the container and the SSIS

objects it contains.

2. SSIS variables can exist in two namespaces: user variables in the User

namespace, and system variables in the System namespace.

3. System variables are read-only; their values are determined by SSIS.

Lesson Summary

■■ You can use variables in SSIS packages to determine certain values once, and then reuse

them multiple times.

■■ Variable values can be assigned literally or by using expressions. SSIS implements a

variety of data types to be used in SSIS variables.

■■ SSIS variables can be assigned dynamically as the package is executed—either once

per execution or iteratively.

■■ Variable accessibility is determined by scope.

■■ SSIS variables can be used to parameterize SSIS object properties.

*Note* **DELAYED VALIDATION**

Design time validation can be disabled for every individual SSIS object by setting the

DelayValidation property to True.

**Quick Check**

1. Which .NET Framework programming language is used for SSIS expressions?

2. What are SSIS expressions typically used for?

**Quick Check Answers**

1. SSIS expressions use a special, proprietary expression language that is only

available for SSIS development and is therefore not part of the .NET Framework.

2. SSIS expressions allow you to determine values needed in SSIS execution dynamically

at run time, rather than having to assign constants to them at design time.

Lesson Summary

■■ The SSIS runtime provides information about the execution environment and other

system information via SSIS system variables.

■■ Expressions can be used to compute the values of SSIS object properties at run time.

■■ Variables and expressions can also be used to extend the elementary functionality of

precedence constraints based on information available at run time that usually is not

available at design time.

*Exam Tip*

Study all workflow tasks carefully in order to understand which of them can be used in a master

package, which situations they are most suited for, and how they should be implemented

in various business cases.

Individual workflow tasks are discussed in more detail in Chapter 4.

**Quick Check**

1. What is the principal purpose of a master package?

2. Can child package properties be set from the master package?

**Quick Check Answers**

1. A master package provides centralized control flow and configuration to SSIS

solutions using multiple SSIS packages.

2. Yes, master package parameters and variables can be used to set the child

package parameters.

Lesson Summary

■■ SSIS operations can be distributed across multiple SSIS packages by using the master

package concept.

■■ The master package (also referred to as the parent package) can execute dependent

packages (or child packages) by using the Execute Package task; this task can also be

used to set child package parameters from the parent package at run time.

*Note* **Using Su rrogate Keys**

In cases when you have to implement Type 2 SCD, you will need to have a new data warehouse

key in the dimension. This key is called a surrogate key.

*EXAM TIP*

Use the Slowly Changing Dimension Wizard to quickly build the ETL process for updating

dimension tables with a small number of rows.

*IMPORTANT* **The Slowly Changing Dimension Wizard LIMI TATION**

The Slowly Changing Dimension Wizard only supports connections to SQL Server.

**Quick Check**

■■ Which SCD types does the Slowly Changing Dimension Wizard support?

**Quick Check Answer**

■■ Types 1 and 2

Lesson Summary

■■ Define each attribute’s SCD type in the data modeling phase of the data warehouse

project.

■■ Use the Slowly Changing Dimension Wizard for dimensions with a small number

of rows.

■■ Use an alternative solution to the Slowly Changing Dimension Wizard for larger

dimensions to solve the problem of updating the Type 1 and Type 2 SCD attributes

of a dimension.

*eXAM TIP*

In SQL Server 2012, you can also use parameters in SSIS to store information that you

would like to change at run time.

*EXAM TIP*

The value of a parameter cannot change within the execution instance of a package. This

means that you cannot change its value inside the package, because its value must remain

the same for the entire execution of the package.

*Note* **Initial and Incremental Loads**

When using CDC functionality in SSIS, you usually have separate packages for the initial and

incremental loads. Remember also that it is possible to set the value of the TaskOperation

property of the CDC Control task dynamically by using expressions, which means that with

some additional logic you can have a single package for both initial and incremental load

when you are using CDC functionality.

**Quick Check**

■■ Which are the new data flow components in SQL Server 2012 for implementing

CDC functionality?

**Quick Check Answer**

■■ The CDC source, for reading the changed rows, and the CDC splitter, for splitting

the input rows into different outputs.

Lesson Summary

■■ You can use expressions to dynamically set properties for tasks and transformations

inside the data flow task.

■■ Use CDC when you cannot define an exact rule for changed data and when your

source systems support this functionality.

■■ Use partitions for your fact tables and apply partition switching instead of deleting data.

**Quick Check**

■■ Can you add an error flow path to the OLE DB destination adapter to store all rows

that could not be loaded into a destination table?

**Quick Check Answer**

■■ Yes, this is a very useful feature when you are designing a robust ETL process.

Lesson Summary

■■ Error flow is a very powerful method for making your data flow task robust when you

are loading data into a data warehouse.

■■ Use error flows to capture problematic rows and store them in a table or file for business

users to inspect.

■■ **Required** If a transaction already exists, join it; if not, start a new transaction.

■■ **Supported** If a transaction exists, join it (this is the default setting).

■■ **NotSupported** The package, container, or task should not join an existing

transaction.

*Exam Tip*

At times, you might want to enable a transaction for a container but exclude some of the

tasks within the container. For example, if you have a couple of Execute SQL tasks that

are used for auditing purposes in a container and the TransactionOption property for the

container is set to Required, the logging tasks will also be rolled back if an error occurs. To

prevent the auditing tasks from rolling back, set the TransactionOption property for those

tasks to Not Supported. This will still let the other tasks in the container be in the transaction,

but it will not include the auditing tasks as part of the transaction.

■■ **Unspecified** A different isolation level than the one specified is being used, but the

level cannot be determined. Set this value to an entire package to override the isolation

levels inside the package and then use specific levels for containers or tasks inside

the package.

■■ **ReadUncommitted** Does not lock the records being read. This means that an uncommitted

change can be read and then rolled back by another client, resulting in a

local copy of a record that is not consistent with what is stored in the database. This is

called a *dirty read* because the data is inconsistent.

■■ **Chaos** Behaves the same way as ReadUncommitted, but checks the isolation level

of other pending transactions during a write operation so that transactions with more

restrictive isolation levels are not overwritten.

■■ **ReadCommitted** Locks the records being read and immediately frees the lock as

soon as the records have been read. This prevents any changes from being read before

they are committed, but it does not prevent records from being added, deleted, or

changed by other clients during the transaction.

■■ **RepeatableRead** Locks the records being read and keeps the lock until the transaction

completes. This ensures that the data being read does not change during the

transaction.

■■ **Serializable** Locks the entire data set being read and keeps the lock until the transaction

completes. This ensures that the data and its order within the database do not

change during the transaction. This is the default value in SSIS.

■■ **Snapshot** The data read within a transaction will not reflect changes made by other

simultaneous transactions. The transaction uses the data row versions that exist when

the transaction begins. No locks are placed on the data when it is read.

*Note* **Setting Is olation Levels of a Transaction in SSIS**

The IsolationLevel property in SSIS objects applies only when you explicitly open a transaction

inside SSIS by setting the TransactionOption property to Required.

**Quick Check**

■■ Suppose you have a package to which you add a sequence container that contains

several tasks, one that calls a command on a legacy system and another that is a

data flow task that imports data into SQL Server. Both tasks have the TransactionOption

property set to Required. Even with the MSDTC service started and transactions

turned on, your sequence container fails before tasks even run. What is the

problem?

**Quick Check Answer**

■■ The transactions featured in SSIS use the MSDTC service. However, not all systems

support the MSDTC, and a transaction cannot be forced on a noncompliant

system, so the container will fail. You should remove the legacy task from the

sequence container or set the TransactionOption property to Not Supported.

Lesson Summary

■■ You can enlist a container or a task to be part of a transaction.

■■ It is important to understand different transaction isolation levels.

**Quick Check**

■■ Can you have multiple checkpoint files for one package?

**Quick Check Answer**

■■ No, you can only have one checkpoint file per package.

*Exam Tip*

If you set the CheckpointUsage property to Always, the checkpoint file must be present or

the package will not start. In addition, using checkpoints is not allowed if you have set the

TransactionOption of the package to Required.

Lesson Summary

■■ Enable checkpoints to restart your package from the last successful step.

■■ Enable checkpoints on the package level and then for each task inside the control flow.

OnError Runs when an executable component reports an error

OnExecStatusChanged Runs for all tasks and containers when the execution status changes

to In Process, Success, or Failed

OnInformation Runs when SSIS displays information messages during the validation

and execution of a task or container

OnPostExecute Runs after a container or task successfully completes

OnPostValidate Executes after the task or container has been successfully validated

OnPreExecute Runs before an executable component is executed

OnPreValidate Runs before a component is validated by the engine

OnProgress Executed when a progress message is sent by the SSIS engine, indicating

tangible advancement of the task or container

OnQueryCancel Invoked when an Execute SQL task is canceled through manual intervention,

such as stopping the package

OnTaskFailed Similar to OnError, but runs when a task fails rather than each time

an error occurs

OnVariableValueChanged Runs when the value changes in a variable for which the

RaiseChangeEvent property is set to True

OnWarning Runs when a task returns a warning event such as a column not being

used in a data flow

**Quick Check**

■■ Can you use a data flow task for a specific event handler?

**Quick Check Answer**

■■ Yes. You can use all the tasks from the SSIS Toolbox.

*Note* **Developing a Custom Auditing and Logging Process**

You can use event handlers to develop a custom auditing and logging process. Each package

contains a set of system variables that are updated for the various levels in the package

during package execution. With event handlers, you can capture these variables and

values, which provide contextual information such as the ErrorCode, ErrorDescription, and

SourceName (the task), when the event fires.

*Exam Tip*

You can turn off event handlers for any task or container by setting the DisableEventHandlers

property of the task or container to True. So if you have an event handler

defined

but you do not want it to be invoked for a specific task, you can turn off event

handlers for that task only.

Lesson Summary

■■ Event handlers enable you to have more control over the execution of a package.

■■ You can use all the control flow tasks and containers when creating event handlers.

■■ Use event handlers if you need to integrate the execution information of SSIS packages

into a central logging database.

*EXAM TIP*

Parameter design values are stored in the project file.

*Exam Tip*

You cannot change the value of a parameter while a package is running.

**Quick Check**

■■ When are property expressions evaluated as a package is running?

**Quick Check Answer**

■■ Unlike parameters that are read at the start of package execution, property

expressions are updated when the property is accessed by the package during

package execution. A property expression can change the value of a property in

the middle of package execution, so that the new value is read when the property

is needed by the package.

*Exam Tip*

You can also update the value of connection properties while a package is running. This

capability is especially useful when you have a Foreach Loop that is iterating over files. You

can use a variable to indicate that the full path is being captured and update the connection

for the file with a property expression. Remember that you cannot use a parameter in

this case to store the full file path, because parameter values cannot change while a package

is running.

Lesson Summary

■■ Use parameters to set up connection properties at run time.

■■ Parameters and project-level connection mangers can only be used with the new project

deployment model introduced with SSIS in SQL Server 2012.

■■ Use property expressions to change the control flow properties at run time.

Package Configuration Types

**Type Description**

XML Configuration File Stores configuration settings in an XML file in the file system. Use this option

if you are comfortable working with configuration files and your project requirements

let you store configuration information in a file system file. Note

that you can store multiple configurations in a single XML file.

Environment Variable Saves the configuration information inside the system’s global variables collection,

which is called an *environment variable*. Only one property can be

stored in each Environment Variable configuration.

Registry Entry Lets you save package properties and settings in your computer’s registry.

Parent Package Variable Provides a way for your package to inherit the value of a variable from a

parent package. When a package is executed from another SSIS package by

using the Execute Package task, the values of its variables are available to

the child package through the Parent Package Variable configuration. With

this configuration type, you can choose only one package property setting

at a time.

SQL Server Stores configuration settings in a SQL Server table. You can store multiple

configurations in a single table.

*Exam Tip*

In a package deployment model, you can pass a variable value from one package to

another only by using package configurations and selecting the Parent Package Variable

package configuration type.

*IMPORTANT* **PARAMETERS OR PACKAGE CONFIGURATIONS?**

If you are using SQL Server 2012, use the new project deployment model with parameters

and project-level connection managers to support moving your solution from one environment

to another. These new features provide better package management and flexibility

in package development compared to package configurations. These new features are

positioned as an evolution of Integration Services deployment and administration in SQL

Server.

*Exam Tip*

Indirect configurations are useful when the location of the file changes between the development

and the production server. To use the indirect configuration, you first need to

create the file by using the wizard and then go back and edit the configuration to assign

the environment variable.

Lesson Summary

■■ Package configurations are available in the package deployment model.

■■ Use package configurations if you are using previous versions of SSIS to set connection

properties at run time.

■■ Use a combination of XML and SQL Server configurations to provide additional portability

for your packages.

■■ ***Integration Services Logging*** This method allows you to configure an SSIS package

to automatically log information about each execution. With this method, you can

identify a variety of *events* that take place during package execution, including exceptions

(such as validation and execution errors and failures).

The amount of available information as well as the structure of each log entry depend

on the actual event—some events are quite verbose, while others are less so. Also note

that some events that can be traced don’t originate in SSIS; they can originate from

any one of the components participating in an SSIS process (such as data providers,

database management systems, and operating systems).

■■ ***Integration Services Auditing*** Provided by a dedicated data flow transformation

component, auditing provides additional capabilities that can be used to extend the

collection of information accompanying the captured events with additional information

about the state of the SSIS process, among other things.

The Audit component can be used to add *system variables* to the data flow, providing

contextual information about the SSIS process in which a particular row of data

was prepared. Auditing information can be stored with the data being processed or

separately.

*Exam Tip*

Study the differences in the operation of the available log providers and how the information

in each corresponding log type can be used later. Some log types provide better

integration with the operating system, whereas others provide better integration with the

data in a database. Not every log type is best suited for every intended use.

**Quick Check**

1. How are SSIS events propagated to the environment?

2. Which SSIS object can be configured for logging?

3. Can SSIS variables create log entries?

**Quick Check Answers**

1. SSIS events can be consumed by log providers, which send them to destinations

in the environment (such as tables in a database, files in the file system, or

operating system event handlers).

2. Any SSIS package executable (control flow container or task) can be configured

for logging, including the package itself.

3. Variables can be configured to trigger an event when their values change, and

this event can then be propagated to a log provider.

Lesson Summary

■■ Before enabling SSIS package logging, you should determine how much information

you actually need to log and choose the appropriate log providers.

■■ Depending on the selected log provider, log information is written to the Windows

logs, to a SQL Server Profiler trace, to a database, or to files in the file system.

*Exam Tip*

In order to use auditing and logging appropriately, you should understand the differences

between them very well.

Rather than trying to gather as much information as possible using just one or the other of

the functionalities, you should focus on using each for the specific features it provides, and

balance their capabilities to achieve a common objective.

**Quick Check**

1. Does the OnError event generated by a data flow task contain enough information

to identify the erroneous rows?

2. Is it possible to correlate log entries generated by SSIS events with data generated

by SSIS auditing?

**Quick Check Answers**

1. The quantity and quality of information available in an event depend on its

origin (for example, SSIS run time, data providers, or database management

systems).

2. Yes, both logging as well as auditing can be configured to generate information

that can be used to correlate log entries generated by an SSIS executable

with the corresponding audit entries.

Lesson Summary

■■ SSIS auditing complements SSIS logging by providing additional information about

how the data in the database was affected by the SSIS process.

■■ The SSIS run time provides information about the current execution instance, which

you can include in your audits.

*Note* **SSIS PACKAGE TEMPLATES IN PRACTICE**

If you are part of an SSIS development team, make sure your manager is aware of the benefits

of using SSIS package templates; also ask him or her to periodically dedicate some time

for training to allow the team members to share their experience in developing SSIS solutions.

The combined knowledge of all team members is one of the principal assets of any

team. Allow your teammates to learn from you, and learn as much as you can from them in

return.

**Quick Check**

1. What is an SSIS package template?

2. Can log configurations be copied from one SSIS object to another?

**Quick Check Answers**

1. An SSIS package template is a regular SSIS package that has been stored in a

specific location and can be used by SSDT when a developer adds new items to

an SSIS project.

2. Yes. Log configurations can be exported to files and reapplied.

Lesson Summary

■■ You can prepare SSIS package templates to simplify SSIS package development.

■■ When you share SSIS package templates with your colleagues, your team can implement

common SSIS development practices.

*Exam Tip*

You should understand SSIS installation requirements very well, especially from the perspective

of the target environments. Typically, a production environment is governed

by a set of very strict rules and principles determining the security and the collection of

features to be deployed. On the other hand, in a development environment, some features

would not be required at all.

*Exam Tip*

Remember that SSDT is a 32-bit-only application; although it can be used in a 64-bit environment,

all the data providers it uses are 32 bit. To enable the SSIS packages to be used in

64-bit environments, you must have the appropriate editions of the data providers; if you

don’t, they will have to be executed using the 32-bit run time.

*Exam Tip*

Before attempting to upgrade an existing installation of SQL Server, you should be familiar

with the limitations regarding versions and editions of SQL Server, as well as whether the

target environment is a 32-bit or 64-bit environment.

*Exam Tip*

Become familiar with relevant SSIS tools, and understand what they are used for. If you

have worked with SSIS in previous versions of SQL Server, remember that some tools are

only provided for legacy reasons.

**Quick Check**

1. What are SQL Server Integration Services (SSIS)?

2. How can SSIS be installed?

**Quick Check Answers**

1. SSIS is a feature of SQL Server that hosts SSIS deployment, maintenance, execution,

and monitoring.

2. SSIS can be installed together with other SQL Server features, added to an existing

SQL Server installation, or used to upgrade an earlier version of SSIS.

Lesson Summary

■■ Before attempting the installation of SSIS, you need to know whether the target environment

is going to be used in development and testing or in production, whether

the environment is a 32-bit or 64-bit environment, and whether SQL Server has already

been installed on the target server.

■■ Service accounts demand special consideration for security and manageability reasons.

*Exam Tip*

Become familiar with the internal layout and organization of the SSISDB database. Understand

the purposes of folders, projects, packages, parameters, environments, and

operations, as well as how they interact when SSIS projects are deployed, maintained, and

executed.

*Exam Tip*

If you have worked with SSIS in a previous version of SQL Server, you will have to “unlearn”

some of the characteristics of the “old” package deployment model when moving on to the

new project deployment model in SQL Server 2012.

You should think of the SSIS package as a set of operations that need to be performed

together, and think of the SSIS project as a collection of packages that are developed,

deployed, and maintained as multiple parts of the same solution.

**Quick Check**

1. What is SSISDB?

2. What is SSIS server?

3. How can SSIS projects be deployed?

**Quick Check Answers**

1. SSISDB is a special database provided by SQL Server 2012 to be used as the

principal SSIS solution repository.

2. SSIS server is a name used to refer to an instance of SQL Server hosting the

SSISDB catalog. Any instance of SQL Server 2012 can be used as the SSIS server,

except SQL Server Express.

3. Under the project deployment model, SSIS project deployment is integrated into

SQL Server Data Tools (SSDT), as well as SQL Server Management Studio (SSMS).

*IMPORTANT* **CLR INTEGRATION**

CLR integration allows a SQL Server instance to use user assemblies, such as userdefined

SQL CLR types, user-defined SQL CLR functions, SQL CLR procedures, or

triggers. CLR integration is disabled by default.

Once enabled, the feature is available throughout the SQL Server instance and can be

used in any database.

Lesson Summary

■■ In SQL Server 2012, deployed SSIS projects are stored in the SSISDB catalog.

■■ The SSISDB catalog is used to store SSIS project and package definitions, as well as

other SSIS metadata.

**FULLY READ AND REVIEW CHAPTER 12**

*Exam Tip*

You should understand the differences between individual SSISDB logging levels in order

to be able to choose the most appropriate one for a particular purpose. For instance, by

turning logging off, you could slightly improve execution performance at the cost of losing

all ability to detect execution errors. On the other hand, using verbose logging will provide

a lot of information about the execution but will lead to less-than-optimal performance.

**Quick Check**

1. Is it possible to execute a deployed SSIS package manually?

2. How can SSIS packages be executed programmatically?

3. Is it possible to monitor SSIS executions?

**Quick Check Answers**

1. Yes. After deployment, SSIS packages can be executed on demand by using

SSMS or by using the DTExec command-line utility.

2. Programmatic access to deployed SSIS packages is possible through Transact-SQL,

Windows PowerShell, and the SSIS managed API.

3. Yes. The execution of SSIS processes can be monitored by using the Active Operations

viewer in SSMS, as well as by using a variety of standard built-in reports.

Additionally, custom reports can be developed, and integrated with SSMS.

*Important* **VALIDATION LIMITATIONS**

SSISDB validation only checks project or package configurations. Do not rely on SSISDB

validation alone to predict successful completion of SSIS processes.

*Note* SQL SERVER AGENT JOBS AND REMOTE SSIS SERVER

SQL Server Agent job steps can reference SSIS packages that are stored on remote servers.

However, SSIS packages stored in an SSISDB catalog are always executed on the SQL

Server instance where the SSISDB catalog is located. To scale SSIS execution to additional

servers, an SSISDB catalog containing the appropriate SSIS projects and/or packages

must be available at each one of them.

Lesson Summary

■■ SSIS packages can be executed manually (for instance, by using SSMS or the DTExec

utility) or programmatically (for instance, by using Windows PowerShell or proprietary

applications). SSIS execution can even be automated—for instance, by using SQL Server

Agent jobs.

■■ You are free to determine the most appropriate methods to execute SSIS processes in

your environment.

■■ SSIS execution monitoring is available in SSMS through standard built-in reports, but

you can also implement your own custom reports if needed.

*Exam Tip*

You should understand the principles of the general SQL Server security model before

learning about the specifics of the SSIS security model. Although both of them have a lot

in common, there are significant differences between them that you should understand

well—particularly with respect to securables and permission granularity.

*Important* **SSISDB FOLDERS**

Only members of the ssis\_admin role are permitted to create folders in the SSISDB catalog.

SSISDB users must be granted the Read and Create Objects permissions on at least one

existing SSISDB folder in order to be able to deploy projects to the SSISDB catalog.

**Quick Check**

1. Who can access SSISDB objects after they have been deployed?

2. How can permissions on various SSISDB objects be controlled?

**Quick Check Answers**

1. By default, access to SSISDB objects is limited to the users who created them

and the members of the ssis\_admin database role. Of course, for any user other

than the SSISDB database owner to be able to create SSISDB objects, the administrator

must first allow that user access to the SSISDB catalog.

2. Permissions can be controlled explicitly on folders, projects, and environments,

but not on packages, environment references, or variables. Permissions on the

latter are inherited from the containing object (the project or the environment

to which the object belongs).

*Important* **EXECUTE AS**

The EXECUTE AS Transact-SQL statement allows you to test the permission assignments

for a particular user, in order to ensure that the SSISDB objects have been secured appropriately.

Of course, to be able to use the statement accordingly, you yourself need

permission to impersonate the user in question.

Lesson Summary

■■ SSISDB catalog security is determined based on information about which SSISDB principal

(SSISDB user) is allowed access to which SSISDB securables (folders, projects, and

environments) and what the nature of this access is (to view, to execute, to modify, or

to remove them).

■■ Only members of the ssis\_admin database role have full access to any of the objects

stored in the SSISDB catalog.

■■ Permissions on SSIS folders, projects, and environments are managed explicitly; permissions

on packages, environment references, and variables, are inherited from the

objects they belong to.

*IMPORTANT* **ALLOWING MULTIPLE ERRORS IN PACKAGE EXECUTION**

The MaximumErrorCount control flow property allows a package to continue to run to

completion even after errors have occurred. This property specifies the number of errors

that can occur before the package will stop executing and report failure.

**Quick Check**

■■ On which levels in the SSISDB catalog are the main predefined reports available?

**Quick Check Answer**

■■ On the SSISDB, project, and package levels

*eXAM TIP*

Data taps cannot be defined at design time.

*Note* **Reading err or messages**

Sometimes you cannot read the full error message on the Progress tab. One way to

address this is to right-click the message, select Copy Message Text, and paste the information

to some other program, such as Notepad, so that you can read the full error

message.

Lesson Summary

■■ Use breakpoints when debugging packages at design time.

■■ SQL Server 2012 provides new ways to efficiently monitor and troubleshoot packages

in a production environment.

■■ **Estimated Row Size** This is not a specific data flow property but is something that

SSIS calculates based on the metadata that it collects about your source data at design

time. You can shrink the row size by identifying the smallest possible data types for

all of your columns as early in the data flow as possible and selecting only needed

columns. This is especially important for flat file data sources, because each column is

automatically read into SSIS as a string data type unless you configure the column’s

data type explicitly.

■■ **DefaultMaxBufferRows** This setting of the data flow task is automatically set at

10,000 records. SSIS multiplies Estimated Row Size by DefaultMaxBufferRows to get a

rough sense of your dataset size per 10,000 records.

■■ **DefaultMaxBufferSize** This is also a setting of the data flow task. It is set to 10 MB

by default. As you configure this setting, keep in mind that its upper bound is constrained

by an internal SSIS parameter called MaxBufferSize, which is set to 100 MB

and cannot be changed.

■■ **MinBufferSize** This is not configurable, but it is still important to know about

because SSIS uses this internal parameter to gauge whether you have set the Default-

MaxBufferSize too low. MinBufferSize is defined by the granularity of your operating

system’s virtual memory allocation functionality. Typically, this is set to 65,536 bytes,

but it differs from machine to machine.

You can configure the DefaultMaxBufferRows

*Note* **Changing Buffer Settings**

When you have the proper data flow design in place, you can start changing the buffer

settings. Tweak the values for DefaultMaxBufferRows and DefaultMaxBufferSize to get as

many records into a buffer as possible. Setting these values too low causes SSIS to create

many small buffers instead of fewer but larger buffers. Fewer but larger buffers is a great

scenario if you have enough memory. After you apply the changes, always test your design

and measure the results on the production server to see the impact.

*eXAM TIP*

You can change the DefaultMaxBufferRows and DefaultMaxBufferSize properties to influence

how SSIS manages buffers.

**Quick Check**

■■ Which performance counters display information that the data flow has started

swapping to disk storage?

**Quick Check Answer**

■■ The Buffers Spooled counter

Lesson Summary

■■ Package design is crucial for good performance.

■■ Monitor package execution with the new functionality of SSISDB catalog views.

**Quick Check**

■■ Is trust a hard data quality dimension?

**Quick Check Answer**

■■ No, trust is a typical soft data quality dimension.

Lesson Summary

■■ Data quality issues can be categorized into data quality dimensions.

■■ Data governance is the activity of taking care of data quality, and data stewards are

people responsible for the quality of particular data.

The Data Quality Server component includes three databases:

■■ DQS\_MAIN, which includes DQS stored procedures. The DQS stored procedures make

up the actual DQS engine. In addition, DQS\_MAIN includes published knowledge

bases. A published KB is a KB that has been prepared to be used in cleansing projects.

■■ DQS\_PROJECTS, which includes data for knowledge base management and data

needed during cleansing and matching projects.

■■ DQS\_STAGING\_DATA, which provides an intermediate storage area where you can

copy source data for cleansing and where you can export cleansing results.

*Exam Tip*

Remember what is stored in each of the three Data Quality Server databases.

**Quick Check**

1. What are the prerequisites for installing Data Quality Server?

2. What are the prerequisites for installing Data Quality Client?

**Quick Check Answers**

1. SQL Server 2012 Database Engine Services are needed for Data Quality Server. In

addition, it is highly recommended that you install Management Tools as well.

2. For Data Quality Client, you need to have the .NET Framework 4 and at least

Internet Explorer 6 SP1 installed on your computer. The .NET Framework 4 is

installed automatically during Data Quality Client setup if it was not already

installed on the machine.

Lesson Summary

■■ Data Quality Services consists of Data Quality Server and Data Quality Client.

■■ There are some prerequisites for installing both components.

■■ You start installation with SQL Server Setup.

■■ You finish the installation with the DQSInstaller.exe application.

With Data Quality Client, you can:

■■ Monitor DQS activities.

■■ Configure reference data service settings.

■■ Configure threshold values for the cleansing and matching.

■■ Enable or disable notifications.

■■ Configure logging.

■ **DQS Administrator (dqs\_admininstrator)** Members of this role can do everything

in the scope of Data Quality Services. They can create or edit a knowledge base, create

and execute a DQS project, terminate a running project or other activity, stop a process,

and change the Reference Data Services and other settings.

■■ **DQS KB Editor (dqs\_kb\_editor)** Members of this role can edit and execute a project

and create and edit a knowledge base. They can monitor all DQS activity; however,

members of this role cannot stop an activity or perform other administrative tasks.

■■ **DQS KB Operator (dqs\_kb\_operator)** Members of this role can edit and execute

a project. However, they cannot create or change a knowledge base. They can monitor

all DQS activity; however, members of this role cannot stop an activity or perform

other administrative tasks.

**Quick Check**

■■ How can you delete old DQS log files?

**Quick Check Answer**

■■ You have to use Windows Explorer to delete old log files. There is no user interface

for this task in Data Quality Client or SQL Server Management Studio.

Lesson Summary

■■ DQS administration involves setting up reference data sources, configuring general

and log settings, monitoring activity, managing security, and backing up and restoring

DQS databases.

■■ You perform all of these tasks except for security management and backup and restoration

with Data Quality Client.

■■ You manage security and backup and restoration of DQS databases with SQL Server

Management Studio or with Transact-SQL commands.

*Exam Tip*

Cardinality, complexity, volatility, auditing, versioning, and reusage determine the definition

of the master data for a company.

**Quick Check**

1. Are constraints in a relational database enough to enforce data accuracy?

2. Do fact tables in a data warehouse contain master data?

**Quick Check Answers**

1. No, constraints enforce data integrity; they cannot enforce data accuracy as well.

2. No, fact tables in a data warehouse typically contain transactional data, which

is usually not considered to be master data.

Lesson Summary

■■ Master data management is a set of coordinated processes, policies, tools, and technologies

used to create and maintain accurate master data.

■■ Master data management has to deal with many challenges.

■■ Master data typically needs auditing, versioning, or some other kind of maintenance of

history.

■■ Reusage increases the value of master data.

■■ If you use a data warehouse, you should map your master data to dimensions in the

data warehouse.

■■ SQL Server 2012 Master Data Services (MDS) is a specialized master data management

application.

*Exam Tip*

Remember that the Master Data Services web application requires specific Windows Web

Server roles and role services, and you must have specific Windows features enabled. It is

not enough to use only SQL Server Setup to install the MDS web application.

**Quick Check**

■■ How can you export data from your MDS database?

**Quick Check Answer**

■■ You can query the subscription views you create, or you can integrate your application

directly to Master Data Services through the Master Data Manager web

service.

Lesson Summary

■■ During the pre-installation tasks, you have to prepare your system for the MDS database

and the web application.

■■ During the installation operations, you use SQL Server Setup to install MDS.

■■ During the post-installation tasks, you configure the MDS database, the Master Data

Manager web application, and the MDS add-in for Excel.

Domain-based attributes form natural hierarchies, which in MDS terminology are known as

*derived hierarchies*. Therefore, for a derived hierarchy, relationships between entities must already

exist in a model. You can think of a hierarchy as an inverted tree. At the lowest level are

the leaf-level members. Members from a lookup entity define the grouping for members of a

base entity; members of the base entity are thus leaf-level members of this derived hierarchy.

Derived hierarchies can be recursive. A hierarchy is *recursive* when a recursive relationship

exists; this happens when an entity has a domain-based attribute based on the entity

itself. Consider, for example, an Employee entity that has a Manager attribute. The Manager

attribute is domain based, and domain values come from the Employee entity itself. If theManager attribute has no value—that is, if it is NULL—then this member is on the top of the

hierarchy. Recursive hierarchies have the following limitations:

■■ You can have only one recursive relationship in a derived hierarchy.

■■ In a recursive hierarchy, you cannot assign member permissions. You can think of

member permissions as row-level security in a table.

■■ You cannot have circular relationships.

In an *explicit hierarchy*, entity members can be organized in any way. The hierarchy structure

can be ragged, which means that the hierarchy can end at different levels. You create

*consolidated members* for the purpose of grouping other members. A consolidated member

belongs to a single explicit hierarchy. The leaf-level members can be under a single level or

under multiple levels of consolidated members; however, you can include each leaf member

in an explicit hierarchy only once. A consolidated member might not even have any leaf

members. In order to use explicit hierarchies, you must enable them on an entity. An explicit

hierarchy can be mandatory, in which all leaf-level members must be part of the hierarchy, or

non-mandatory, where some leaf-level members are not part of the hierarchy. The leaf-level

members that do not belong to a consolidated member remain in a special system-created

node (consolidated member) called Unused.

*Exam Tip*

Make sure you understand the difference between derived hierarchies, explicit hierarchies,

and collections.

**Quick Check**

■■ You want to improve the quality of your master data. You want to prevent the insertion

of inaccurate values into an attribute, and furthermore, you want to notify

a data steward responsible for this attribute when an inaccurate value is inserted.

How can you achieve these two tasks?

**Quick Check Answer**

■■ You can create one or more change value and/or validation business rules to prevent

the insertion of inaccurate values or to correct inaccurate values automatically.

You can create a business rule for an external action such as sending an email

message in order to notify the data steward responsible for the attribute.

Lesson Summary

■■ In MDS, models are containers for other objects.

■■ Entities have attributes, derived and explicit hierarchies, collections, and attribute

groups.

■■ Attributes can be free-form, file, or domain based.

■■ You can define several business rules for each attribute.

*Exam Tip*

The vast majority of companies already have line of business applications in production in

which the master data is maintained. Therefore, importing batches of data is a very important

part of an MDS deployment and maintenance.

**Quick Check**

■■ Can you deploy an existing MDS model to another MDS instance with data

included?

**Quick Check Answer**

■■ Yes. You have to extract the package with the MDSModelDeploy command.

Lesson Summary

■■ You can use model deployment packages to export and import metadata, and in some

cases data as well.

■■ You can use the staging process for importing batches of data.

■■ Applications can use the MDS web service and subscription views to get master data

from an MDS instance.

There are two levels of administrators in MDS: the *MDS System Administrator* and *Model*

*Administrators*.

In order

to change the System Administrator, you need to:

1. Use SSMS to query the *mdm.tblUser* table. You need to find the user who will be the

new administrator and copy the value in the *SID* column for this user.

2. Use the following piece of code (replace *DOMAIN\user\_name* with the new administrator's

user name, and replace *SID* with the value you got in the previous step).

EXEC mdm.udpSecuritySetAdministrator @UserName='DOMAIN\user\_name', @SID = 'SID', @

PromoteNonAdmin = 1

Model Administrators are the users who have Update permissions on the complete model—

that is, on the top-level object—and no other assigned permissions on lower levels, such as

the entity level.

*Exam Tip*

Be sure you have a thorough understanding of how MDS determines effective permissions

for a user.

**Quick Check**

■■ A user is a member of two groups. You give Read-Only permission for an object to

the first group, Update permission for the same object to the second group, and

Read-Only permission for the same object to the user. What effective permission

does the user have for that object?

**Quick Check Answer**

■■ The user has the Update effective permission for that object.

Lesson Summary

■■ Keep security settings simple.

■■ An end user must have at least Explorer functional area access and permissions on at

least one object from a model in order to get access to the model.

■■ Overlapping permissions make security more complicated.

**Quick Check**

■■ How can you update batches of MDS data without using the staging process?

**Quick Check Answer**

■■ Use Excel 2012 with MDS Add-in for Excel to update batches of data in an MDS

model.

Lesson Summary

■■ With MDS Add-in for Excel, you can edit MDS data in batches by using all Excel

capabilities.

■■ Advanced users can even create new entities and manage attributes from Excel.

■■ MDS maintains security and validation for Excel data the same way it does for data

updated with Master Data Manager.

Building a DQS KB involves the following processes:

■■ ***Knowledge discovery*** A computer-guided process that uses a data sample

■■ ***Domain management*** An interactive process in which you manually verify and

extend domains in a KB

■■ ***Reference data services*** A process in which you validate domain data against

external data maintained and guaranteed by an external provider

■■ ***Matching policy*** A process in which you define rules to identify duplicates

*Exam Tip*

Understand the difference between synonyms and term-based relations.

Besides defining correct, invalid, erroneous, and corrected values for a domain, you can

also define domain rules. A *domain rule* is a condition that DQS uses to validate, standardize,

and/or correct domain values. You can define rules such that the value must be greater than

a selected value, must begin with a value, must comply with a pattern or a regular expression,

must contain a value, must be in a list of values you specify, and more. A single rule can have

multiple conditions connected with logical AND or logical OR operators.

In addition to synonyms, you can create term-based relations. You use a *term-based*

*relation* to correct a term that is part of a domain value and not the complete domain value.

You define a term-based relation once, and DQS uses it for all values in a domain. For example,

you could define that the string Inc. should always be expanded to Incorporated.

**Quick Check**

■■ You want to use a knowledge base that exists in one DQS instance in another DQS

instance. Should you re-create the knowledge base on the second DQS instance

manually?

**Quick Check Answer**

■■ No, you should export the knowledge base from the first DQS instance and import

it into the other.

Lesson Summary

■■ You can start building a knowledge base with knowledge discovery.

■■ You can manually edit domains after the discovery process is finished.

■■ You can import and export knowledge bases and domains.

*Exam Tip*

Use the DQS Cleansing transformation for batch cleaning during your extract-transformload

(ETL) process.

A cleansing project has the following stages:

1. **Mapping** In this stage, you map source columns to KB domains.

2. **Computer-assisted cleansing** In this stage, DQS uses the KB with built-in algorithms

and heuristics to find the best match of an instance of data you are cleansing to known

data domain values.

3. **Interactive cleansing** In this stage, you review the results of the computer-assisted

cleansing and additionally correct data. You see the changes that DQS proposes and

decide whether to approve or reject the changes.

4. **Export** In this stage, you export the cleansed data. You can export the data to SQL

Server tables or Excel files (.xlsx, .xls, and .csv). You can also standardize output if

you defined standardized output in the appropriate domain of the knowledge base

used. You can decide to export data only, or data and cleansing information. The data

cleansing information includes source value, output value, reason for correction, confidence

for correction, and the status of the operation performed on the data.

**Quick Check**

■■ Which are the two cleansing phases of a DQS project?

**Quick Check Answer**

■■ The two cleansing phases of a DQS project are computer-assisted cleansing and

interactive cleansing.

Lesson Summary

■■ A DQS project has four stages: mapping, computer-assisted cleansing, interactive

cleansing, and export.

■■ You can cleanse data that comes from SQL Server or from Excel.

■■ You never modify the source data directly; you can export the data with the cleansing

info as the result of your cleansing project.

SSIS Data Profiling Task

With the SSIS Data Profiling task, data profiling is very simple but also limited. The Data

Profiling task saves the result in XML form, which cannot be used directly in an SSIS package.

You can write custom code in the SSIS Script task in order to consume the results in the same

package. For an overview of the result, an application called Data Profile Viewer is included

within the SQL Server suite. You will learn how to read the results of the Data Profiling task

inside your SSIS package in Chapter 19, ”Implementing Custom Code in SSIS Packages.”

*Exam Tip*

You need to use custom code inside an SSIS package to use the results of the Data Profiling

task in the same package.

You can use the SSIS Data Profiling task for profiling the following:

■■ **Column Length Distribution** This helps you find strings of unexpected length.

■■ **Column Null Ratio** Use this to find the percentage of NULLs in a column.

■■ **Column Pattern** This is a very powerful profile that expresses patterns in strings as

regular expressions and then calculates the distribution of these regular expressions.

■■ **Column Statistics** This gives you the minimum, maximum, average, and standard

deviation for numeric columns, and the minimum and maximum for datetime columns.

■■ **Column Value Distribution** This gives you the distribution of values for discrete

columns.

■■ **Candidate Key** This profile gives you the percentage of unique values in columns,

thus helping you identify columns that are candidates for keys.

■■ **Functional Dependency** This profile reports how much the values in a dependent

column depend on the values in a set of determinant columns.

■■ **Value Inclusion** This profile finds the extent to which column values from one table

have corresponding values in a set of column values of another table, thus helping you

find potential foreign keys.

**Quick Check**

■■ Can you use the SSIS Data Profiling task to cleanse your data?

**Quick Check Answer**

■■ No, with the SSIS Data Profiling task you can only profile your data.

Lesson Summary

■■ In addition to DQS, you can use many other tools from the SQL Server suite to improve

the quality of your data.

■■ The SSIS Data Profiling task is easy to use for quick profiling of your data.

What Is Data Mining?

What does the term *data mining* mean? In short, by using data mining (DM) you can deduce

hidden knowledge by examining, or *training*, the data with data mining algorithms.

**Quick Check**

■■ How do you perform a prediction by using a data mining model?

**Quick Check Answer**

■■ You need to create a DMX prediction query that joins patterns stored in an SSAS

mining model with your data.

Lesson Summary

■■ You create data mining models with SSDT by using the Analysis Services Multidimensional

and Data Mining Project template.

■■ You can use mining model predictions in SSIS packages.

With the Term Extraction transformation, you can retrieve the key terms from a Unicode

string or text column. The transformation has built-in knowledge about the grammar and

syntax of the English language. Therefore, you should use this transformation for English text

only. If you use it for other languages, the results might be worse than you expect.

The Term Lookup transformation uses a dictionary of terms that is stored in a column in a SQL

Server table and applies this dictionary to an input data set that contains documents stored as

a Unicode string or text column. The transformation counts how many times a term appears

in a document. It extracts words from a document in an input column by using the same

knowledge about the English language as the Term Extraction transformation

*Exam Tip*

Any table with a column with key terms can be used as the reference dictionary for the

Term Lookup transformation. Although it is quite convenient to create the dictionary with

the Term Extraction transformation, these two transformations do not need to be connected

in any way. You can also edit the dictionary extracted with the Term Extraction

transformation manually.

**Quick Check**

■■ Which languages are supported by the two text mining transformations?

**Quick Check Answer**

■■ Both the Term Extraction and Term Lookup transformations are limited to the

English language only.

Lesson Summary

■■ You can extract terms with the Term Extraction transformation.

■■ You can count how many times a term appears in a document with the Term Lookup

transformation.

*Exam Tip*

You typically train data mining on a sample from a large data set.

**Quick Check**

■■ How can you select a predefined number of rows randomly?

**Quick Check Answer**

■■ You can use the SSIS Row Sampling transformation to select a predefined number

of rows randomly.

Lesson Summary

■■ You have to have a lot of knowledge to prepare data for data mining.

■■ The SSIS Percentage Sampling and Row Sampling transformations help you select

rows randomly.

Configuring the Script Task

There are many uses of the script task. For example, you might need to browse Active

Directory Domain Services, and there is no such task out of the box. You might need to

generate a custom list of objects for the Foreach Loop. You might want to check whether a

file you need to import is empty before starting the import. Another good example is using

the results of the Data Profiling task further in your package. As you already know, the Data

Profiling task writes results to an XML file. You can use the Data Profile Viewer application to

view the results; however, you cannot incorporate the results in your package directly.

*Exam Tip*

Before coding the script, you need to configure the script task.

**Quick Check**

■■ Can you log the execution of your script in the script task?

**Quick Check Answer**

■■ Yes, you can use the Log property of the Dts object to log the execution of your

script.

Lesson Summary

■■ You can extend the control flow of your packages with the script task.

■■ Before writing the code, you configure the script task.

*Exam Tip*

Note that the script component can return synchronous or asynchronous outputs.

**Quick Check**

■■ What does it mean if the value of the SynchronousInputID property of a script

component is the component’s input ID?

**Quick Check Answer**

■■ This means that the component returns synchronous outputs.

Lesson Summary

■■ For the script component, you need to configure inputs and outputs.

■■ An output of an SSIS component can be synchronous or asynchronous.

■■ The run-time methods of the script component are the same as the run-time methods

of a custom component.

**Quick Check**

1. When should custom components be used instead of script components?

2. When is it necessary to use asynchronous outputs?

3. Why do custom components need design-time programmatic logic?

**Quick Check Answers**

1. Generally, the number-one reason for using custom components is reusability.

If the same custom operation needs to be implemented in multiple SSIS packages,

custom components with package-independent development, deployment,

and maintenance capabilities provide a more appropriate alternative.

2. Whenever a data transformation operation produces zero, one, or more rows

per each input row, asynchronous outputs need to be used, because additional

rows cannot be added to a synchronous output, nor can they be removed from

them.

3. Before being used in an executing SSIS package to perform the principal data

management operation, each custom component is used by the SSIS package

developer—it needs to be placed in the data flow, configured, and validated.

Lesson Summary

■■ If a particular problem cannot be solved by using any of the standard built-in SSIS data

flow components, you can design your own custom component: a custom source, a

custom destination, or a custom transformation component.

■■ Before custom components can be used in SSIS development, they need to be deployed;

this means that they must be copied to the workstation and registered in the

workstation’s global assembly cache. Of course, to be used in deployed SSIS processes,

the components must also be deployed on the target server.

**Quick Check**

■■ Which two SSIS transformations are useful for identity mapping and de-duplication?

**Quick Check Answer**

■■ The SSIS Fuzzy Lookup transformation is useful for identity mapping, and the SSIS

Fuzzy Grouping transformation is useful for de-duplication.

Lesson Summary

■■ You can use DQS and SSIS from the SQL Server suite for identity mapping and deduplicating

tasks.

■■ Data merging has to be done based on similarities of strings.

■■ When performing the matching, you have to avoid cross joins of huge amounts of

data.

■■ You can even use MDS functions for your own mapping procedures.

*Exam Tip*

Remember that you should cleanse your data before matching.

**Quick Check**

■■ How can you influence which column in a table should be more important for

matching than other columns?

**Quick Check Answer**

■■ When creating a matching KB, you can define weight for each domain. Give higher

weight to the domain mapped to the column you want to be more important for

matching.

Lesson Summary

■■ You can use the DQS Cleansing transformation to take advantage of DQS knowledge

bases to correct data in the SSIS data flow.

■■ For DQS matching, use a matching policy DQS knowledge base.

■■ You can use sample data to speed up the learning process when creating a matching

policy DQS knowledge base.

**Quick Check**

■■ Which SSIS transformation is useful for de-duplication?

**Quick Check Answer**

■■ You can use the SSIS Fuzzy Grouping transformation for de-duplication.

Lesson Summary

■■ SSIS Fuzzy Lookup and Fuzzy Grouping can handle bigger amounts of data than

DQS can.

■■ DQS matching is very useful for master data matching when you are using MDS Add-in

for Excel to edit your master data.