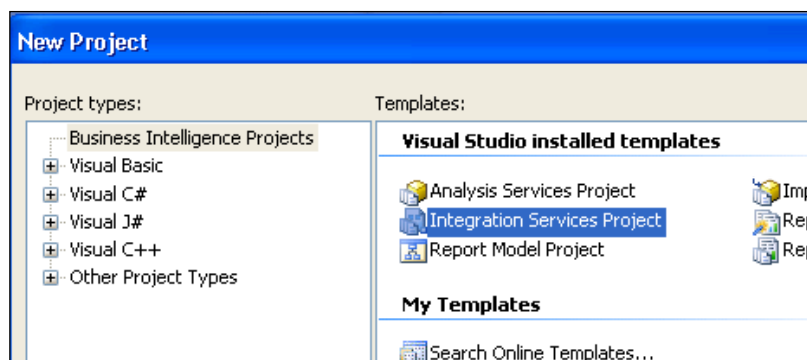
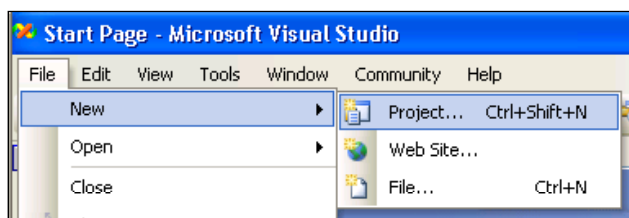


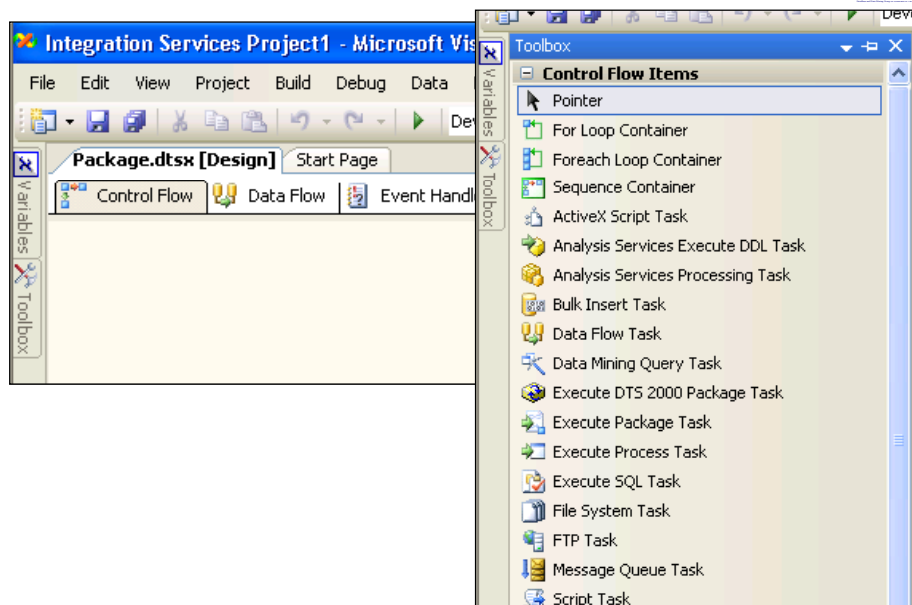
Integration Services project

- An Integration Services project allows managing all ETL processes
- It is based on Business Intelligence projects of type “Integration Services”
- Open Visual Studio and create a new Business Intelligence projects of type “Integration Services Project”



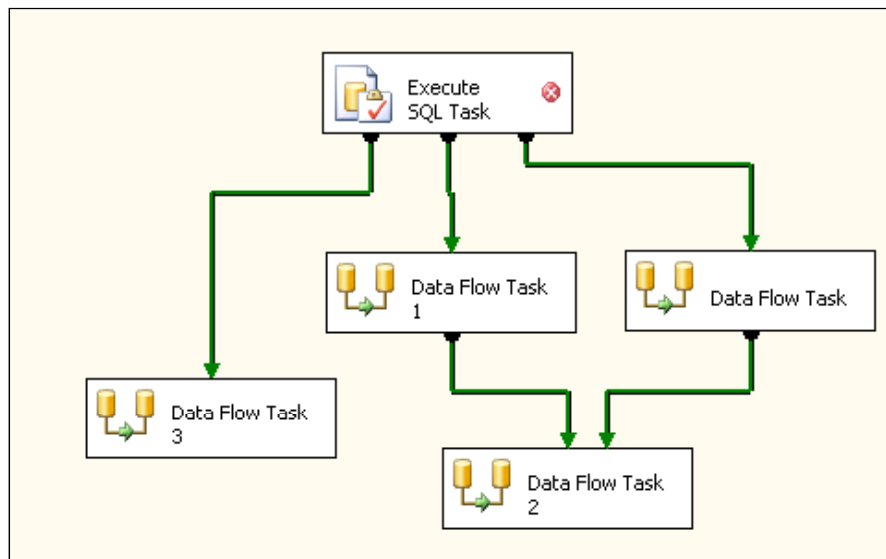
Control Flow

- A Control Flow defines the flow of tasks by means of the following tools
 - Data Flow Task
 - Basic block to execute data transfer from a source to a destination using transformations
 - Execute SQL Task
 - SQL statements execution
 - Execute Package
 - External package execution



Creating a Control Flow

- To create a new Control Flow
 - drag and drop the objects in the middle of the main window
 - connect the objects using the connection arrows
- Connection arrows define the order of precedence among tasks
- Usually there is no data exchange among Control Flow tasks
 - You can use variables



Database and data mining group, Politecnico di Torino

DBG

Control flow - Execute SQL Task

- Execute SQL Task allows the execution of SQL statements/scripts
- Usually used to
 - Create tables
 - Delete data
 - Insert information into Auditing tables

SQL Server 2005 Integration Services - 8

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Database and data mining group, Politecnico di Torino

DBG

Execute SQL Task

Execute SQL Task Editor

Configure the properties required to run SQL statements and stored procedures using the selected connection.

General

Parameter Mapping

Result Set

Expressions

General

Options

Result Set

SQL Statement

Name

Description

TimeOut

CodePage

ResultSet

ConnectionType

Connection

SQLSourceType

SQLStatement

IsQuery/StoredProcedure

BypassPrepare

Execute SQL Task

Execute SQL Task

0

1252

None

OLE DB

Direct input

False

False

SQLStatement

Specifies the query to be run by the task.

Connection to DB

SQL statement to execute

SQL Server 2005 Integration Services - 9

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Control flow - Execute SQL Task

- Before executing a SQL statement, a connection to a DB must be defined
- Creating a new connection
 - Name of the server
 - Name of the data base
 - Authentication method

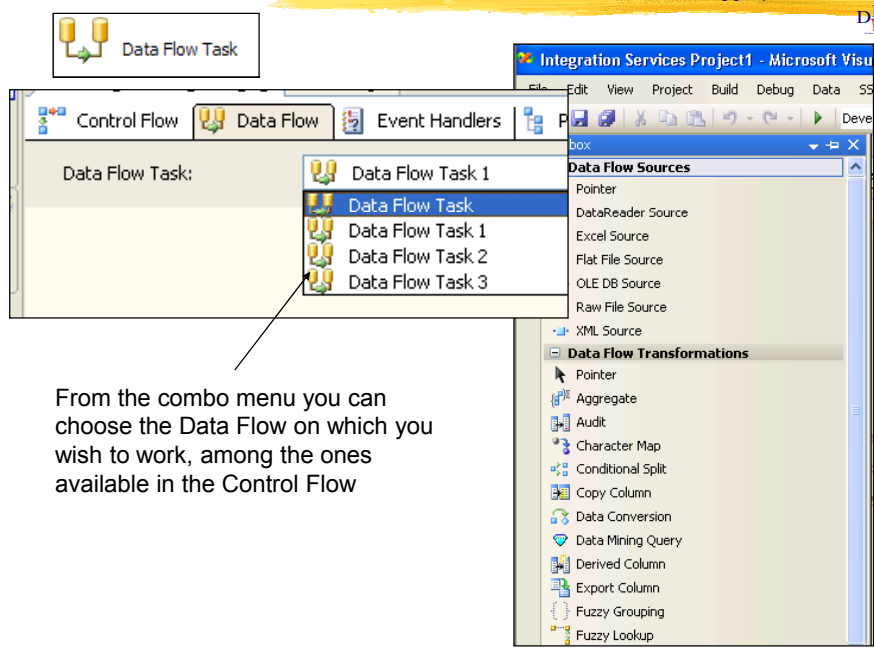
The screenshot shows the 'SQL Statement' task configuration window. The 'Connection' tab is active, displaying a list of connections on the left. The 'ConnectionType' is set to 'OLE DB'. The 'SQLSourceType' is set to '<New connection...>'. The 'SQLStatement' field is empty. The 'Server name' is set to 'localhost'. The 'Log on to the server' section has 'Use Windows Authentication' selected. The 'Connect to a database' section has 'Select or enter a database name' selected, and 'mio_database' is chosen from the dropdown. The 'Test Connection' button is visible at the bottom left.

Control Flow – Data Flow

- Data Flow blocks handle the data processing by means of objects that
 - Define sources and destinations
 - Transform and add attributes
 - Merge the content of different sources
 - ...

Database and data mining group, Politecnico di Torino

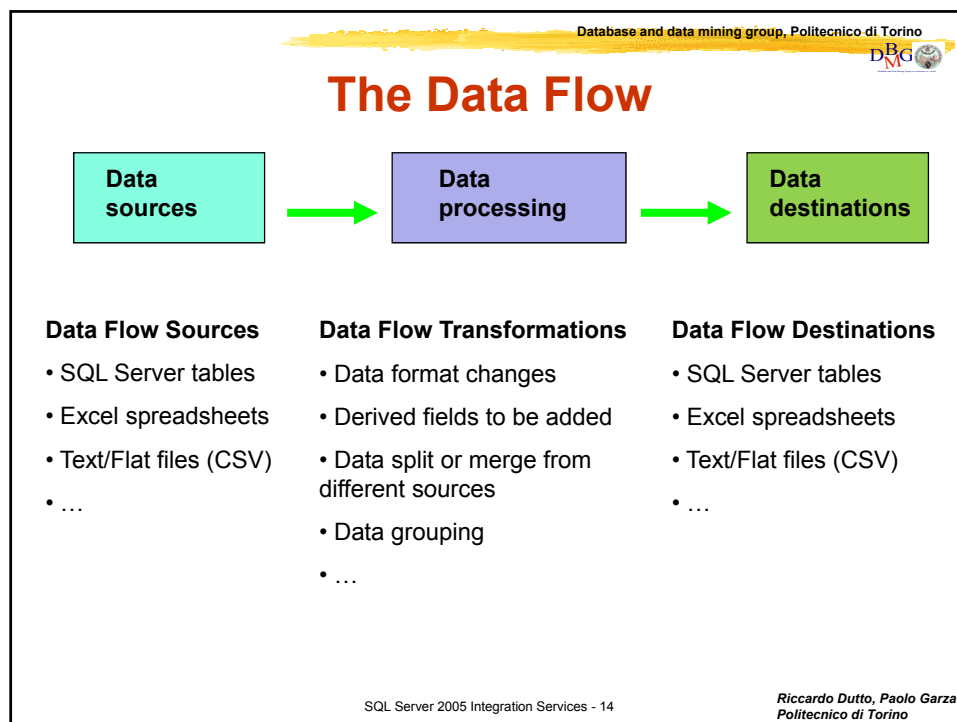
DBG



From the combo menu you can choose the Data Flow on which you wish to work, among the ones available in the Control Flow

SQL Server 2005 Integration Services - 13

Riccardo Dutto, Paolo Garza
Politecnico di Torino



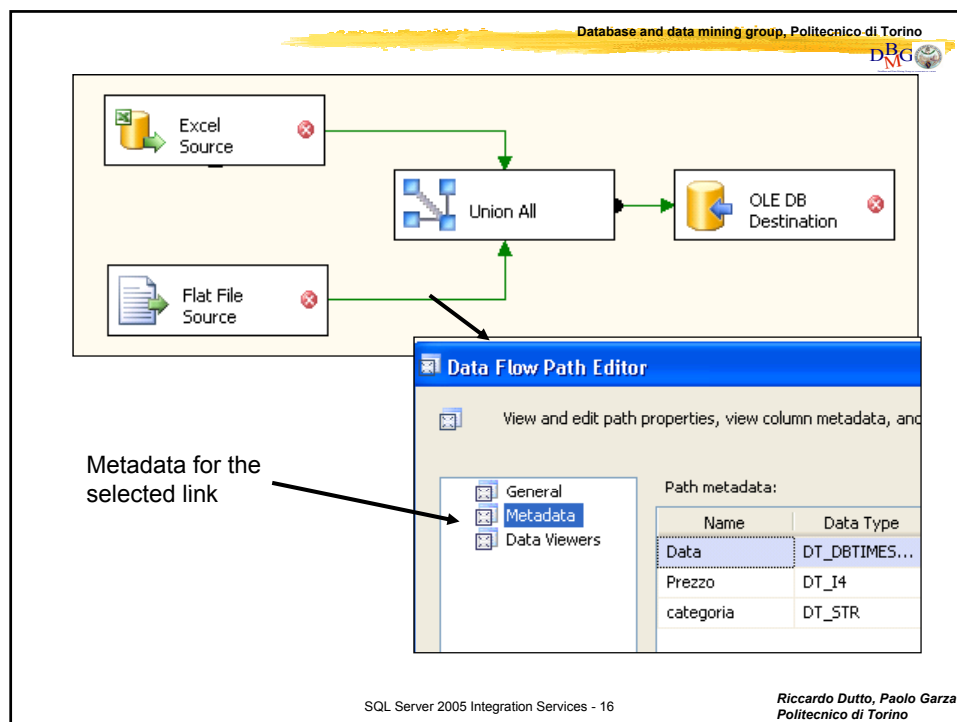
Database and data mining group, Politecnico di Torino
DBG

Control Flow – Data Flow

- In the Data Flow, connection arrows indicate physical data exchange
- Double clicking on a link (Data Flow Path) you can
 - Read the metadata of transferred data
 - Define data viewers to visualize data being transferred on the link itself at runtime
 - Useful for debugging

SQL Server 2005 Integration Services - 15

*Riccardo Dutto, Paolo Garza
Politecnico di Torino*



Database and data mining group, Politecnico di Torino

Data Flow Source - OLE DB source

- OLE DB source
 - Defines a connection to a SQL Server source
 - Specifies which data are used

SQL Server 2005 Integration Services - 17

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Data Flow Source - OLE DB source

- Choose the connection to the desired DB
 - If the connection does not exist, create a new connection
- Choose the table to access to
 - Allows accessing to the whole table content
- Alternatively, specify a SQL query
 - Allows accessing to selected data from one or more tables

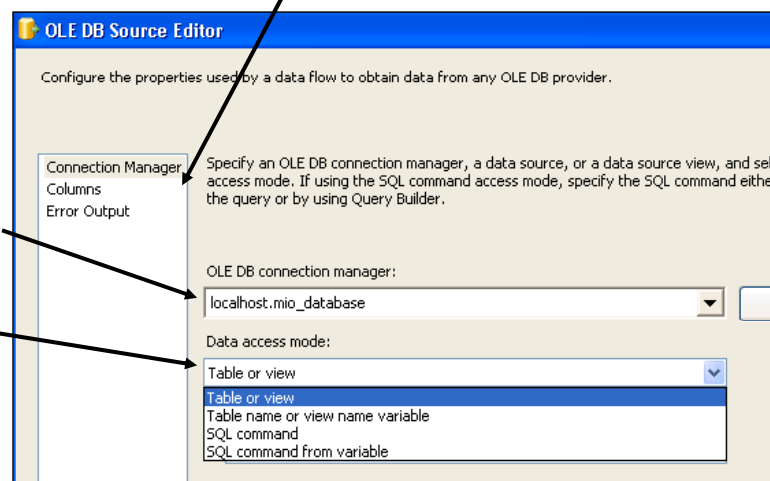


Selected columns

Connection to DB in use

Table access method

- whole table
- SQL query



Database and data mining group, Politecnico di Torino

DBG

Data Flow Transformations

Data Conversion


- Used to change the data format
 - Change the string format
 - E.g., Unicode to non-Unicode
 - Numerical precision conversion
 - ...
- Creates a new column for each transformation

SQL Server 2005 Integration Services - 20

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Database and data mining group, Politecnico di Torino

DBG

 Data Conversion

Available Input Colu...

Name
<input checked="" type="checkbox"/> DATA
<input checked="" type="checkbox"/> CATEGORIA
<input checked="" type="checkbox"/> PREZZO

Input Column	Output Alias	Data Type
DATA	DATA_NEW	database timestamp [DT_DBTIMESTAMP]
CATEGORIA	CATEGORIA_NEW	string [DT_STR]
PREZZO	PREZZO_NEW	four-byte signed integer [DT_I4]

Column with input data

New columns generated by the conversion block

Output column format

SQL Server 2005 Integration Services - 21

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Database and data mining group, Politecnico di Torino

DBG

Data Flow Transformations

Union all

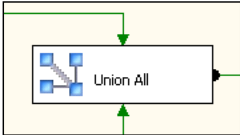
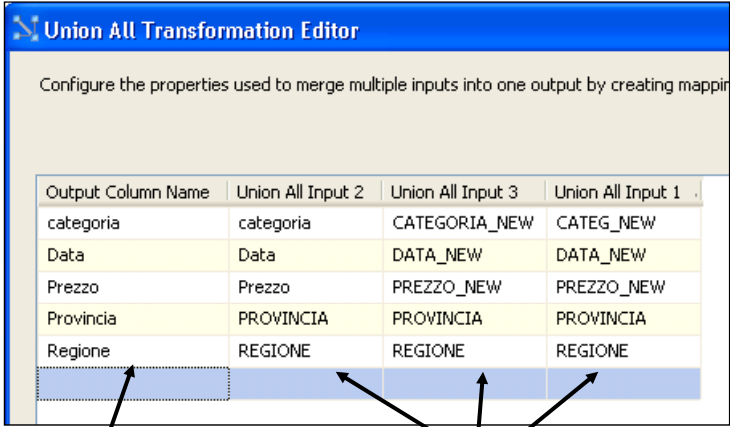
- Two or more sources are merged to create a common output
 - Input data must have the same schema and format to allow union

SQL Server 2005 Integration Services - 22

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Database and data mining group, Politecnico di Torino

DBG

Output Column Name	Union All Input 2	Union All Input 3	Union All Input 1
categoria	categoria	CATEGORIA_NEW	CATEG_NEW
Data	Data	DATA_NEW	DATA_NEW
Prezzo	Prezzo	PREZZO_NEW	PREZZO_NEW
Provincia	PROVINCIA	PROVINCIA	PROVINCIA
Regione	REGIONE	REGIONE	REGIONE

Output attribute name Input attribute names

SQL Server 2005 Integration Services - 23

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Database and data mining group, Politecnico di Torino

DBG

Data Flow Transformations

Derived columns


- Creates new columns
 - For each row the content of new columns may be
 - a fixed value
 - the result of a function applied to other columns
- Replaces the content of a column with a new value
 - The new value may be
 - a fixed value
 - the result of a function

SQL Server 2005 Integration Services - 24

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Database and data mining group, Politecnico di Torino

DBG

 Derived Column

Derived Column Transformation Editor

Specify the expressions used to create new column values, and indicate whether the values are new columns.

Variables

Columns

- Data
- Prezzo
- categoria
- Regione
- Provincia

Mathematical Functions

String Functions

Date/Time Functions

NULL Functions

Type Casts


Operators

Derived Column Name	Derived Column	Expression	Data Type
ANNO	<add as new column>	YEAR(Data)	four-byte s
MESE	<add as new column>	MONTH(Data)	four-byte s
AUTORE	<add as new column>	"PIPPO"	Unicode st
Provincia	Replace 'Provincia'	UPPER([Provincia])	string [DT...

SQL Server 2005 Integration Services - 25

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Database and data mining group, Politecnico di Torino



Data Flow Transformations


Aggregate


- Aggregates the rows according to the aggregating attributes
- Returns the value of aggregated functions for each group
 - sum, average, count, ...
- When there is only one data source, the same result can be obtained using the group by clause

SQL Server 2005 Integration Services - 26

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Database and data mining group, Politecnico di Torino



 Aggregate

Aggregating attributes

Results of aggregating functions

Input Column	Output Alias	Operation	Com
Regione	Regione	Group by	
categoria	categoria	Group by	
ANNO	ANNO	Group by	
MESE	MESE	Group by	
Prezzo	Incasso	Sum	
(*)	Quantita	Count all	

SQL Server 2005 Integration Services - 27

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Database and data mining group, Politecnico di Torino

DBG

Data Flow Transformations

Lookup


- Returns a given field of another “table”
 - it behaves as a join, but
 - can be applied to data flows which come from different sources

SQL Server 2005 Integration Services - 28

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Database and data mining group, Politecnico di Torino

DBG



Lookup

Choose a connection and a table on which to lookup

Choose the join condition and the field to return as result

Lookup Transformation Editor

This transform enables the performance of simple e

Reference Table Columns Advanced

Specify a data source to use. You can select a t
connection, or the results of an SQL query.

OLE DB connection manager:
localhost.DW_STABILIMENTI_BALNEARI

Use a table or a view:
[dbo].[LOCALITA]

Reference Table Columns Advanced

Specify join columns and use of reference columns here.

Available Input Col...

Name
Quantità
categoria
regione
provincia
Incasso
anno
me

Available Lookup Col...

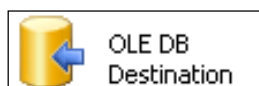
Name
<input checked="" type="checkbox"/> COD_L
<input type="checkbox"/> REGIONE
<input type="checkbox"/> PROVINCIA

SQL Server 2005 Integration Services - 29

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Data Flow Destination OLE DB

- Defines a connection to a SQL Server destination
- Specifies which data are used
- Last block of all Data Flows



Specify the connection to the output DB

Define the column mapping

Connection Manager
Mappings
Error Output

Specify an OLE DB connection manager, a data access mode. If using the SQL command access the query or by using Query Builder. For fast-load

OLE DB connection manager:
localhost.DW_STABILIMENTI_BALNEARI

Data access mode:
Table or view - fast load

Name of the table or the view:
[dbo].[FATTOAFFITTI]

Available Input

Name
Quantità
categoria
regione
provincia
Incasso
anno
mese
COD_L
COD_C

Available Destination

Name
COD_T
COD_C
COD_L
INCASSO
QUANTITA

Database and data mining group, Politecnico di Torino

DBG

Data Flow - Execute Package Task

- Executes a child package inside a given package
- Variable values can be passed from parent to child packages

SQL Server 2005 Integration Services - 32

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Database and data mining group, Politecnico di Torino

DBG

Execute Package ...

Location of the saved package

Information on the package to be executed

Execute Package Task Editor

Configure the properties used to execute an SSIS package created in SQL Server 2005.

General

Package

Execution

Location	SQL Server
Connection	
PackageName	
Password	*****
ExecuteOutOfProcess	False

Location
Specifies the storage location type of the package to be run.

OK

Cancel

Help

SQL Server 2005 Integration Services - 33

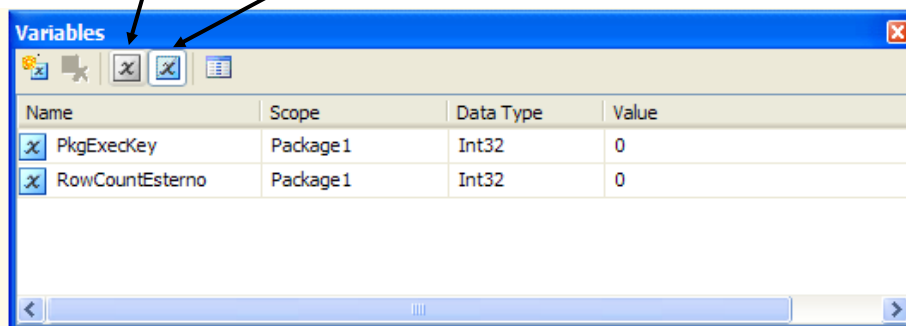
Riccardo Dutto, Paolo Garza
Politecnico di Torino

Variables

- System variables
 - Predefined variables with system information
 - PackageName
 - ExecStartDT
 - ...
- User variables
 - User defined
 - Used to manage
 - Local/relative paths
 - Data counters, ...

System variables

User variables

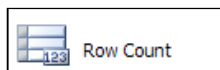


Name	Scope	Data Type	Value
PkgExecKey	Package1	Int32	0
RowCountEsterno	Package1	Int32	0

Data Flow Transformations

Row count

- Counts the number of rows flowing through a link
 - the block must be put on the link to analyze
- The result is stored in a user variable
 - Useful to debug and log operations



User variable name
where the result of the
count is stored

Advanced Editor for Row Count

The advanced editor provides access to the low-level properties of data flow components. Additionally, the advanced editor can be used to configure components that do not have a custom user interface.

Component Properties | Input Columns | Input and Output Properties

Specify advanced properties for the data flow component.

Properties:

Common Properties	
ComponentClassID	{DE50D3C7-41AF-4804-9247-CF1DEB147971}
ContactInfo	Row Count;Microsoft Corporation;Microsoft SQL Server v9; (0
Description	Counts the rows in a dataset.
ID	SSS
IdentificationString	component "Row Count" (SSS)
IsDefaultLocale	True
LocaleID	English (United States)
Name	Row Count
PipelineVersion	0
UsesDispositions	False
ValidateExternalMetadata	True
Version	0
Custom Properties	
VariableName	User::RowCountEsterno

Name
Specifies the name of the component.

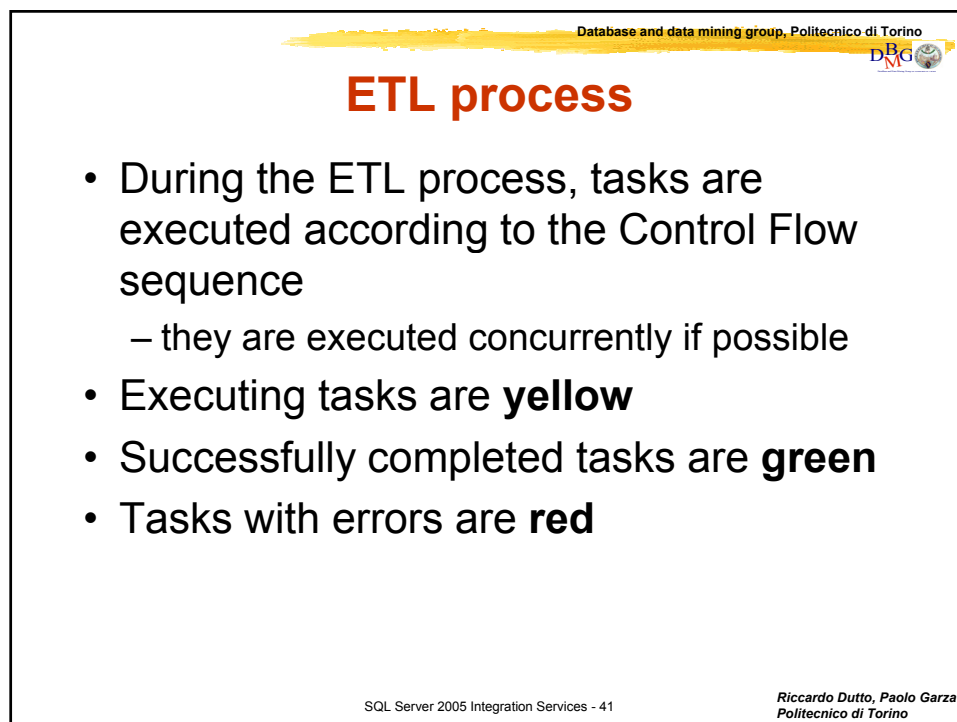
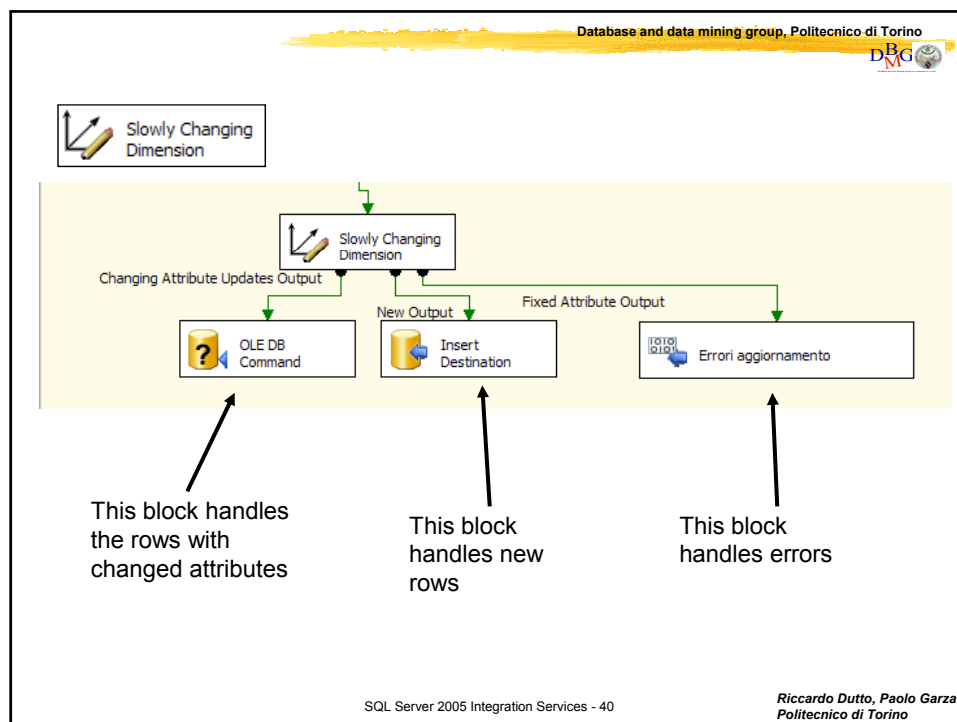
Refresh OK Cancel Help

Incremental data loading

- Dimensions
 - identify additions and changes
 - SQL Server 2005 has a component dedicated to changing dimensions
- Fact table
 - identify the new rows
 - usually user variables are used to identify the temporal period of interest
 - rows in source tables must have the insert/update timestamp

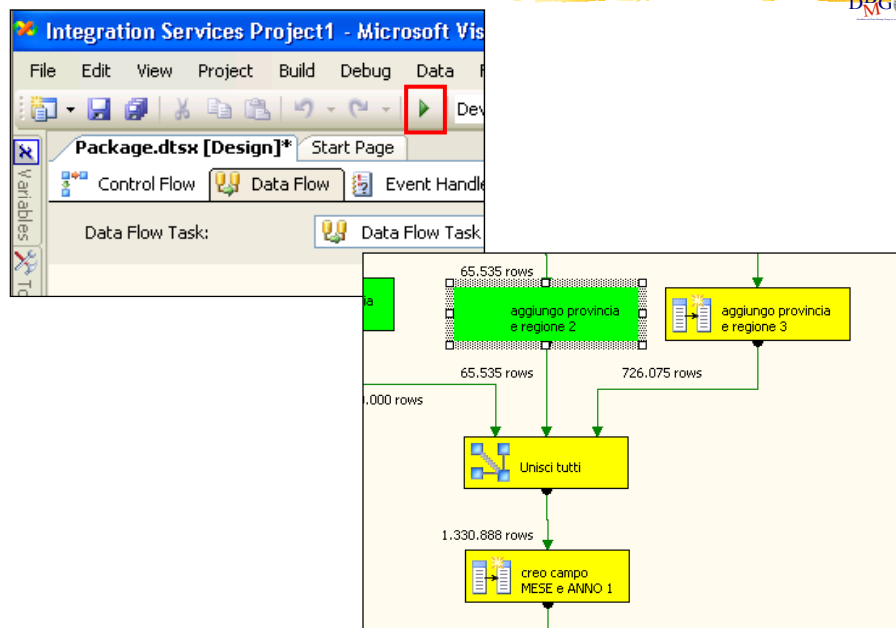
Data Flow Transformations Slowly Changing Dimension

- Automatically generates the blocks needed to handle dimension changes
- The user chooses how to handle dimension changes for each attribute of the changing dimension
 - The choice, among the available types, must be done during the conceptual design of the data warehouse



ETL process

- On the links among the Data Flow blocks, the number of processed rows is shown
- Data Viewers allow in depth analysis of data flowing on links
 - useful to debug



Automatic ETL processing

- The ETL process (as a package) can be automatically executed using
 - dtexec
 - SQL Agent

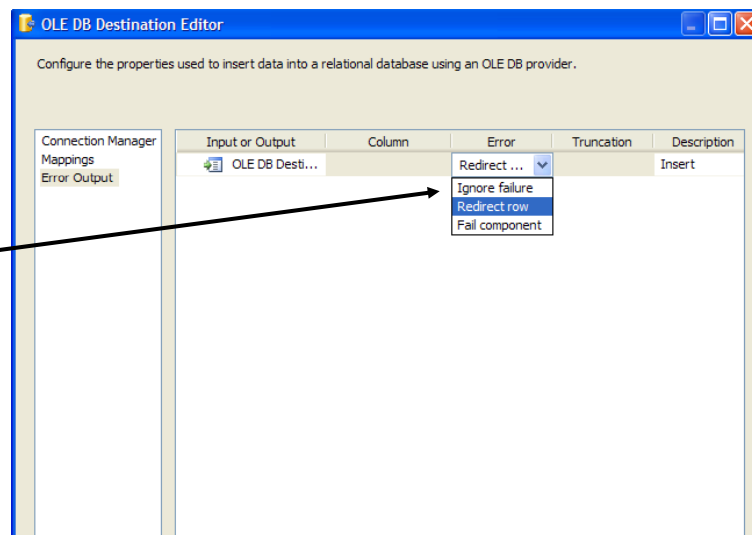
Error handling

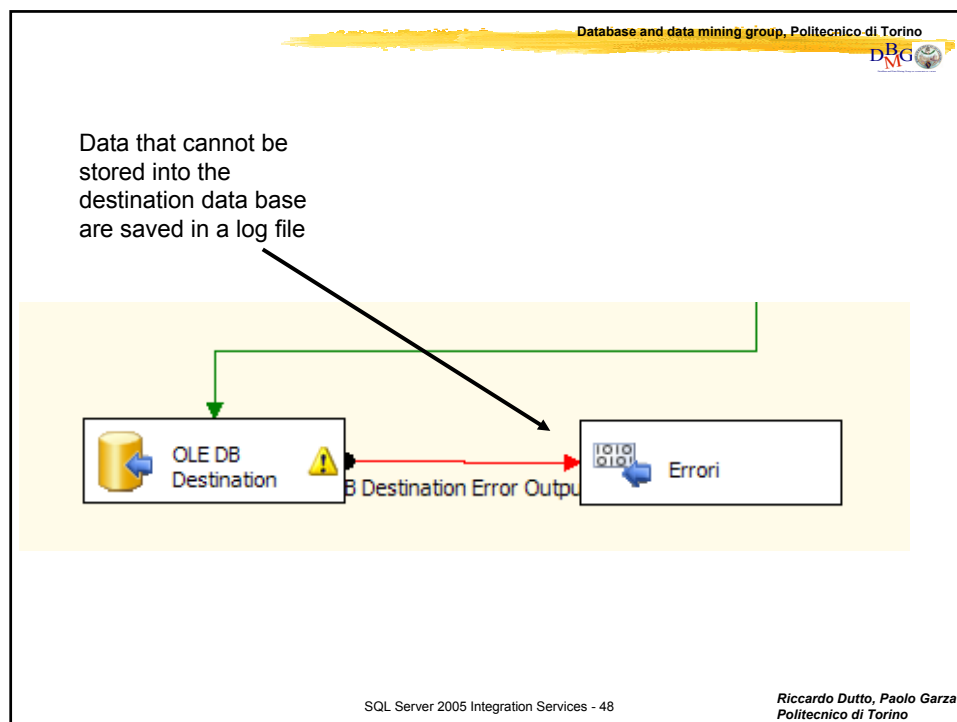
- Errors must be handled, otherwise
 - the whole package fails
 - you do not know where the error lies
- Handling errors allows to
 - successfully load the data that do not generate errors
 - save the data that do generate errors (e.g., in a log file)

Error handling

- The “Error” preference of blocks can be set to the following values
 - Redirect row
 - The rows that generate errors are redirected to the error link (i.e., the red output line of the block)
 - Ignore failure
 - Rows that generate errors are ignored, but the process goes on with the following rows
 - Fail component (default value)
 - The component fails
 - No data are loaded

Type of
error
handling





Database and data mining group, Politecnico di Torino

DBG

Metadata - Auditing

- During the ETL, executed processes (packages) must be tracked
 - State
 - Execution time
 - Number of rows analyzed by each process
 - Number of errors
 - ...

SQL Server 2005 Integration Services - 49

Riccardo Dutto, Paolo Garza
Politecnico di Torino

Metadata - Auditing

- Auditing information / Metadata are usually stored inside a relational table
 - Easy to update during the ETL process execution
 - Easy to analyze by means of SQL queries

Metadata - Auditing

- In the Control Flow, insert a block at the beginning of each process saving information before execution
- Insert a block at the end of each process saving information of execution and result
 - Number of processed rows
 - Number of errors

Metadata - Auditing

- User variables
 - Information of the process execution and completion, at the end of the block
 - Number of processed rows
 - Number of errors
- Text files or data base
 - Rows that generated errors

SSIS Log

- Automatic process log system provided by SQL Server 2005
- Allows to access and save the basic process/package information
- Information can be saved in files, a SQL Server data base, ...
- Information that cannot be saved
 - Number of processed rows
 - Rows that generated errors

Database and data mining group, Politecnico di Torino
DBG

Configure SSIS Logs: DimPromotion

Create and configure a new log to capture log-enabled events that occur at run time.

Containers:

☒ DimPromotion

☒ Caricamenti dati

☒ Delete data from

☒ Inizializzazione

☒ Metadata- Fine

Providers and Logs

Details

Select the events to be logged for the container:

Events	Description
<input type="checkbox"/> OnError	Handles error events. Use to define actions to perfo...
<input type="checkbox"/> OnExecStatusChanged	Handles changes of execution status. Use to define ...
<input type="checkbox"/> OnInformation	Handles information events. The meanings of inform...
<input type="checkbox"/> OnPipelinePostEndOfRowset	A component will be given the end of rowset signal.
<input type="checkbox"/> OnPipelinePostPrimeOutput	A component has returned from its PrimeOutput call.
<input type="checkbox"/> OnPipelinePreEndOfRowset	A component will be given the end of rowset signal.
<input type="checkbox"/> OnPipelinePrePrimeOutput	PrimeOutput will be called on a component.
<input type="checkbox"/> OnPipelineRowsSent	Rows were provided to a data flow component as in...
<input type="checkbox"/> OnPostExecute	Handles post-execution events. Use to define post-...
<input type="checkbox"/> OnPostValidate	Handles post-validation events. Use to define post-...
<input type="checkbox"/> OnPreExecute	Handles pre-execution events. Use to define pre-pr...
<input type="checkbox"/> OnPreValidate	Handles pre-validation events. Use to define pre-pr...
<input type="checkbox"/> OnProgress	Handles progress notifications. Use to define action...
<input type="checkbox"/> OnQueryCancel	Handles cancel events. Called periodically to determi...
<input type="checkbox"/> OnTaskFailed	Handles task failures. Use to define actions to perfo...

Advanced >>

Load...

Save...

OK

Cancel

Help

Task to log

Type of events to log

SQL Server 2005 Integration Services - 54

Riccardo Dutto, Paolo Garza
Politecnico di Torino