

Ascential DataStage™

Merge Stage Guide

Version 1.2



This document, and the software described or referenced in it, are confidential and proprietary to Ascential Software Corporation ("Ascential"). They are provided under, and are subject to, the terms and conditions of a license agreement between Ascential and the licensee, and may not be transferred, disclosed, or otherwise provided to third parties, unless otherwise permitted by that agreement. No portion of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of Ascential. The specifications and other information contained in this document for some purposes may not be complete, current, or correct, and are subject to change without notice. NO REPRESENTATION OR OTHER AFFIRMATION OF FACT CONTAINED IN THIS DOCUMENT, INCLUDING WITHOUT LIMITATION STATEMENTS REGARDING CAPACITY, PERFORMANCE, OR SUITABILITY FOR USE OF PRODUCTS OR SOFTWARE DESCRIBED HEREIN, SHALL BE DEEMED TO BE A WARRANTY BY ASCENTIAL FOR ANY PURPOSE OR GIVE RISE TO ANY LIABILITY OF ASCENTIAL WHATSOEVER. THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT OF THIRD PARTY RIGHTS. IN NO EVENT SHALL ASCENTIAL BE LIABLE FOR ANY CLAIM, OR ANY SPECIAL INDIRECT OR CONSEQUENTIAL DAMAGES, OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE. If you are acquiring this software on behalf of the U.S. government, the Government shall have only "Restricted Rights" in the software and related documentation as defined in the Federal Acquisition Regulations (FARs) in Clause 52.227.19 (c) (2). If you are acquiring the software on behalf of the Department of Defense, the software shall be classified as "Commercial Computer Software" and the Government shall have only "Restricted Rights" as defined in Clause 252.227-7013 (c) (1) of DFARs.

© 2004 Ascential Software Corporation. All rights reserved. DataStage®, EasyLogic®, EasyPath®, Enterprise Data Quality Management®, Iterations®, Matchware®, Mercator®, MetaBroker®, Application Integration, Simplified®, Ascential™, Ascential AuditStage™, Ascential DataStage™, Ascential ProfileStage™, Ascential QualityStage™, Ascential Enterprise Integration Suite™, Ascential Real-time Integration Services™, Ascential MetaStage™, and Ascential RTI™ are trademarks of Ascential Software Corporation or its affiliates and may be registered in the United States or other jurisdictions.

Adobe Acrobat is a trademark of Adobe Systems, Inc. UniData and UniVerse are registered trademarks of IBM Corporation. Microsoft, Windows, Windows NT, and Windows Server are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. UNIX is a registered trademark in the United States and other countries, licensed exclusively through X/Open Company, Ltd. Other marks mentioned are the property of the owners of those marks. The software delivered to Licensee may contain third-party software code. See *Legal Notices* ([LegalNotices.pdf](#)) for more information.

How to Use This Guide

DataStage Merge is a passive stage that joins data from two sequential files. A custom GUI is included in this stage to provide a graphical, easy-to-use interface for the join operation. Version 1.2 of the Merge stage is compatible with Ascential DataStage Release 7.5.1.

Audience

This guide is intended for DataStage designers who create or modify jobs that use Merge.

How This Book is Organized

The following table lists topics that may be of interest to you and it provides links to these topics.

To learn about	Read...
Functionality	"Functionality" on page 1
Installation	"Installing the Plug-In" on page 1
Adding Merge to a DataStage job	"Adding Merge to a DataStage Job" on page 1
Invoking the GUI	"Invoking the GUI" on page 2
Defining stage properties	"Defining Stage Properties" on page 3
Adjusting for input file size	"Adjusting for Input File Size" on page 5
Defining output properties	"Defining Output Properties" on page 6

Related Documentation

To learn more about documentation from other Ascential products as they relate to the Merge stage, refer to the following section/table.

Ascential Software Documentation

Guide	Description
<i>Ascential DataStage Designer Guide</i>	General principles for designing jobs
<i>Ascential DataStage Server Job Developer's Guide</i>	Techniques for designing server jobs
<i>Ascential MetaStage User's Guide</i>	Information about Ascential MetaStage™
<i>Ascential DataStage NLS Guide</i>	Information about NLS and techniques for character-set mapping
<i>Ascential DataStage Plug-In Installation and Configuration Guide</i>	Information required to configure your system and install this stage

Conventions

Convention	Used for...
bold	Field names, button names, menu items, and keystrokes. Also used to indicate filenames, and window and dialog box names.
<code>user input</code>	Information that you need to enter as is.
<code>code</code>	Code examples
<code>variable</code> or <code><variable></code>	Placeholders for information that you need to enter. Do not type the greater-/less-than brackets as part of the variable.
<code>></code>	Indicators used to separate menu options, such as: Start >Programs >Ascential DataStage
[A]	Options in command syntax. Do not type the brackets as part of the option.
B...	Elements that can repeat.

Convention	Used for...
A B	Indicator used to separate mutually-exclusive elements.
{ }	Indicator used to identify sets of choices.

Contacting Support

To reach Customer Care, please refer to the information below:

Call toll-free: 1-866-INFONOW (1-866-463-6669)

Email: support@ascentialsoftware.com

Ascential Developer Net: <http://developernet.ascential.com>

Please consult your support agreement for the location and availability of customer support personnel.

To find the location and telephone number of the nearest Ascential Software office outside of North America, please visit the Ascential Software Corporation website at <http://www.ascentialsoftware.com>.

Contents

Audience	iii
How This Book is Organized	iii
Related Documentation	iv
Ascential Software Documentation	iv
Conventions	iv
Contacting Support	v
Introduction	1
Functionality	1
Installing the Plug-In	1
Adding Merge to a DataStage Job	1
Invoking the GUI	2
Defining Stage Properties	3
Defining Directories for Input and Working Files	3
Defining Character Set Mapping	5
Adjusting for Input File Size	5
Defining Output Properties	6
Specifying the File Names	7
Specifying the Type of Join	7
Specifying the Tracing Level	7
Specifying Input File Format	8
Specifying Input File Columns	10
Saving Column Information in a Table	12
Specifying Keys for the Join	12
Specifying Output Columns	13
Defining the Name and Format of the Output Columns	14

Introduction

DataStage Merge allows you to merge two sequential files into one or more output links. A Merge stage is a passive stage that can have no input links and one or more output links.

The custom GUI allows you to use point-and-click techniques to represent the merge operation. Merge is part of the Ascential DataStage Designer, which lets you select stage icons, drop them onto the Designer work area, and add links.

This technical bulletin describes how to use the Merge GUI to define the input files and the output link.

For more information on using a stage in a DataStage job, see Ascential DataStage documentation.

Functionality

DataStage Merge allows you to define the join operation that is used to merge two sequential files. The two input files to be merged must be sequential text files.

The DataStage Merge stage has the following functionality:

- Support for NLS (National Language Support) in automatic mode only. For more information, see *Ascential DataStage NLS Guide*.
- Support for Ascential MetaStage™. For more information, see *Ascential MetaStage User's Guide*.

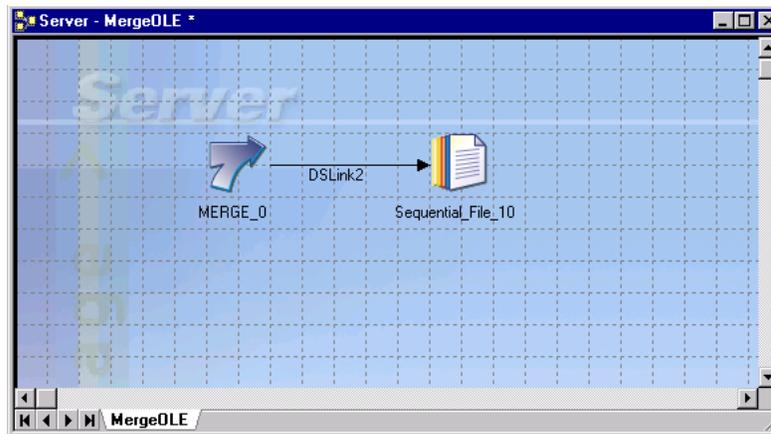
Installing the Plug-In

For instructions and information supporting the installation, see *Ascential DataStage Plug-In Installation and Configuration Guide*.

Adding Merge to a DataStage Job

You must add Merge to a DataStage job before you can use the Merge custom GUI. To add Merge to a DataStage job:

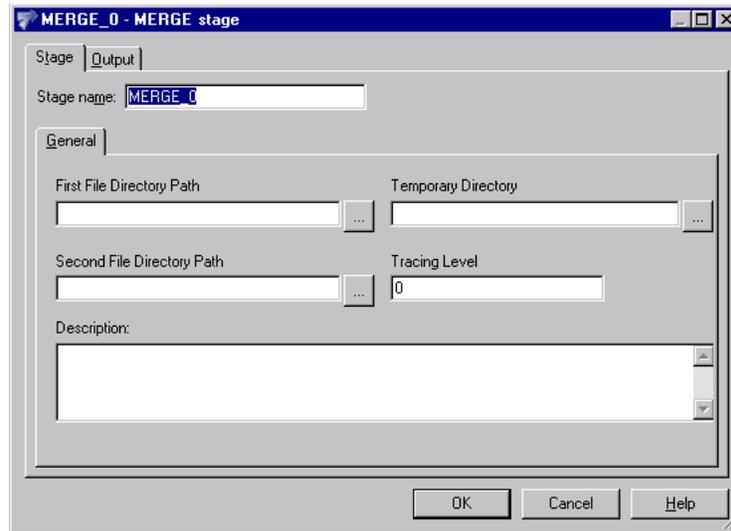
- 1 Start the DataStage Designer by doing one of the following:
 - Choose **Start > Programs > Ascential DataStage > DataStage Designer**.
 - Double-click the **Ascential DataStage** icon in the Ascential DataStage program folder.
- 2 Choose **Processing** from the **Palette**. A drop-down list of available stages is displayed.
- 3 From the list, select **MERGE Plug-in** and click **OK** to create a Merge stage. The Merge stage is now added to the DataStage job.
- 4 You also need to create a link and an output file. Once you have created them, your DataStage Designer window should resemble the following:



Invoking the GUI

Invoke the GUI from the DataStage Designer by double-clicking the Merge stage, in this case **MERGE_0**, on the Diagram window. Alternatively, you can choose **Job Properties** from the Edit menu.

When you run the GUI editor, the MERGE_0 - MERGE Stage dialog box appears with the **General** page at the front of the **Stage** page:



Defining Stage Properties

Use the **Stage** page to define the Stage properties. The **Stage** page contains the following tabs:

Tab	Function
General	Defines input file and working file directories and log file information.
NLS	Specifies a character set map for the stage.

The following sections describe how to perform the functions of each tab.

Defining Directories for Input and Working Files

When you use Merge, you must define the directory for both input files. You also must specify a directory that is used for working files that are deleted when the job is successfully completed.

Use the **General** tab on the **Stage** page to specify the path and directory for the first and second input files. You can specify a directory path in the following ways:

- Click the ... button to browse directories.
- Enter the directory path in the respective text entry box.

Using Browse

If you use Browse (... button), the Browse directories dialog box appears. The following table describes the elements in the dialog box:

Element	Description
Directory on field	Displays the directory path. This field is automatically updated with the drive and directory you choose. You can also enter a directory path directly in this field.
Directory list	Displays the directories on the chosen drive. Double-click the directory you want.
Drive on list, <i>displayed only when connected to a Windows server</i>	Displays the mounted drives on the DataStage server. Choose the drive you want from the drop-down list. The Directory list box is automatically updated when you choose a drive.
OK button	Accepts the text in the Directory on field and closes the Browse directories dialog box.
Cancel button	Closes the dialog box without specifying the directory path.
Help button	Starts the Help system.

Entering the Directory Path

Fill in the fields on the **General** page as described in the following table:

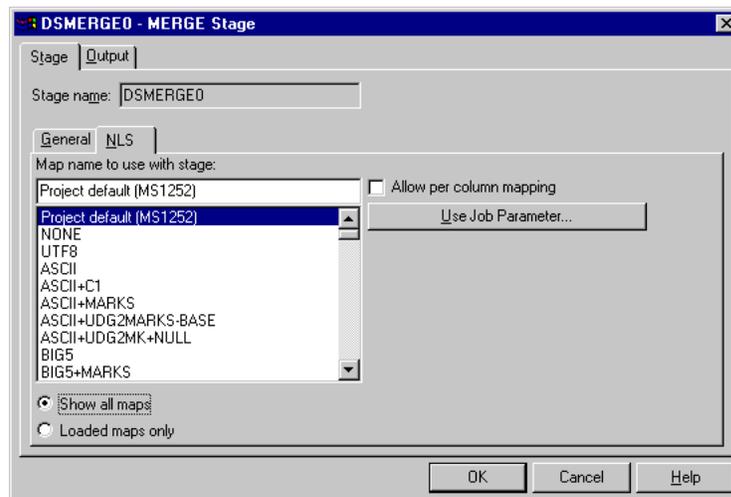
Field	Default	Description
First File Directory Path	None	The path and directory of the first sequential file.
Second File Directory Path	None	The path and directory of the second sequential file.
Temporary Directory	Current working directory	The complete path and directory in which temporary files are stored. These temporary files are created while a job is running and deleted when the job is complete.
Tracing Level	0	Specifies the type of information to be included in the job log file. You can specify the following tracing levels: 0 No information is written to the log file. 1 Stage properties are written to the log file.
Description		An optional description of the stage properties.

You can also include a job parameter in the directory path. For information on how to define and use job parameters, see DataStage documentation.

Defining Character Set Mapping

Using the **NLS** tab, you can define a character set map for a stage and specify whether the mapping is to be done by column. The **NLS** tab appears only if NLS is installed.

The **NLS** tab appears on the Stage page as shown:



The default character set map is defined for the project or the job. For more information, see DataStage documentation. You can select another character set map from the **Map name to use with stage** drop-down list.

Columns within a record can use different maps within the metadata. To enable character set mapping on a column basis, check the **Allow per-column mapping** box.

Adjusting for Input File Size

The Merge stage supports 64-bit files. But you must change the value of the property **Max Space in VM for Hash Table** to accommodate extremely large input files. Failure to do so results in abnormal termination of jobs. The default value of **Max Space in VM for Hash Table** is 12. This value is appropriate for many file sizes. As the size of the larger of the two input files grows, you must increase the value of **Max Space in VM for Hash Table**. For files of 2 GB or larger, you must set the value of **Max Space in VM for Hash Table** to its maximum value of 512.

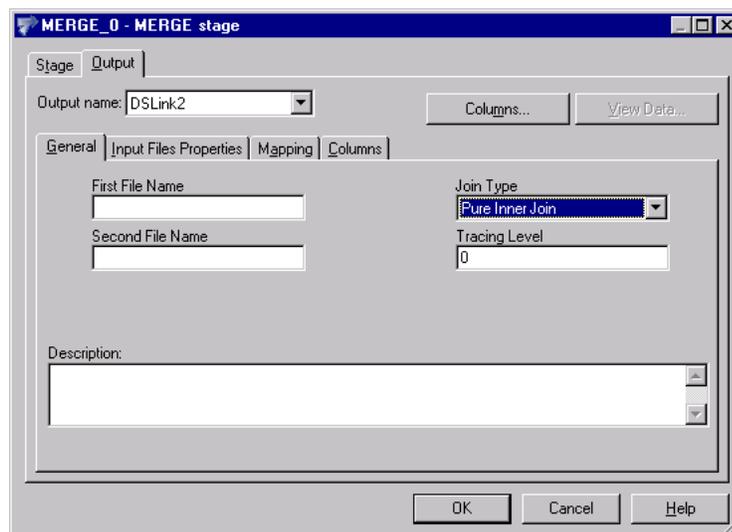
To access **Max Space in VM for Hash Table**, right click the **Merge** icon on the canvas, and select **Grid Style**. The grid-style editor appears. Go to the **Properties** tab of the **Output** page. Scroll the list of properties until you come to **Max Space in VM for Hash Table**.

Defining Output Properties

The **Outputs** page in the MERGE_0 - MERGE Stage dialog box lets you specify properties for the output link. Output properties describe different characteristics of your input files and the output link, such as the following:

- Names of the first and second input files
- Output link tracing level
- Format of the first and second input files
- Column names and characteristics of the first and second input files, including character set mapping
- Column information to be saved to a table
- Type of join operation to be performed
- Keys used in the join operation
- Content of columns in the output link
- Column names and formats in the output link, including character set mapping

Each task is described in the order in which you might perform it. To perform these tasks, click the **Outputs** tab. The **General** page appears at the front of the **Outputs** page as shown:



Note The **Columns...** button lists the columns in the output link and is included only for compatibility with other stages.

Specifying the File Names

You must enter file names for the first and second files.

To specify the file names, enter the directory path and file names in the **First File Name** and **Second File Name** text boxes on the **General** page. Both files must be sequential text files. You can also include a job parameter in the directory path. For information on how to define and use job parameters, see DataStage documentation.

Specifying the Type of Join

Use the **General** page on the **Outputs** page to specify the type of join you want to perform on the two input files. You can choose one of the following types of join operations:

Type of Join	Operation	Description
Pure Inner Join	A AND B	Merges only those rows with the same key values in both input files.
Complete Set	A OR B	Merges all rows from both files.
Right and Left Only	A NOR B	Merges all rows from both files except those rows with the same key values.
Left Outer Join	A	Merges all rows from the first file (A) with rows from the second file (B) with the same key value.
Right Outer Join	B	Merges all rows from the second file (B) with rows from the first file (A) with the same key value.
Left Only	A NOT B	Merges all rows from the first file except rows with the same key value in the second file (B).
Right Only	B NOT A	Merges all rows from the second file except rows with the same key value in the first file (A).

Specifying the Tracing Level

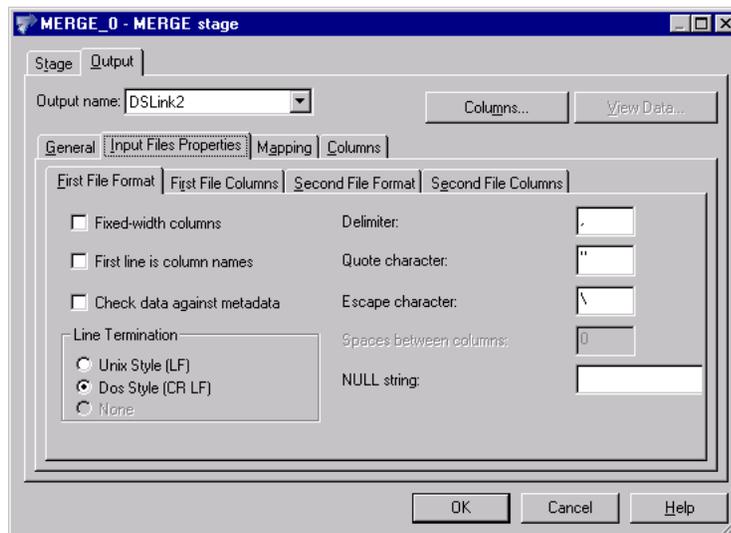
Similar to the stage properties, you can specify a tracing level for the output link. The tracing level specifies the type of information to be

included in the job log file. You can specify the following tracing levels:

Tracing Level	Description
0	No information is written to the log file.
1	Output link properties are written to the log file.

Specifying Input File Format

You must specify the file format of the first and second input files. To specify the file format, click the **Input Files Properties** tab on the **Outputs** page. The **First File Format** page appears at the front of the **Input Files Properties** page, as shown:



To specify the format of the second input file, click the **Second File Format** tab. The fields and check boxes are identical for the second file. The following table describes each field and check box on the **First File Format** or **Second File Format** pages:

Field	Default	Description
Fixed-width columns	Cleared	Indicates whether the file has fixed-width columns.
First line is column names	Cleared	Indicates whether the first line of the first sequential file is column names.

Field (Continued)	Default	Description (Continued)
Check data against metadata	Cleared	Indicates whether to use metadata definitions to read data from the file instead of using a line terminator for the end of a row. Data is read until the metadata is exhausted. For fixed-width data, this means the total of the column lengths plus spaces. For delimited data, this means the number of columns. (Output link only.) If cleared, the end of row is determined by the end-of-line sequence.
Delimiter	, (comma)	Specifies the delimiter that separates the data fields in the file. This option is enabled if Fixed-width columns is cleared. You can enter an unquoted single character or the ASCII value of the character you want to use.
Quote character	" (double quotation marks)	Specifies the single character used to enclose a data value that contains the delimiter character as data. This option is enabled if Fixed-width columns is cleared. You can also enter the ASCII value for the character you want to use. You can suppress Quote character by not entering a value.
Escape character	\ (backslash)	Specifies a single character to be interpreted as an escape character. This option is enabled if Fixed-width columns is cleared.
Spaces between columns	0	Specifies the number of spaces between columns in a sequential file with fixedwidth columns.
NULL string	None	Specifies the string used for the SQL null value.
Unix Style (LF)	Cleared	Specifies whether a line-feed character is used to indicate the end-of-line sequence in the input file.
Dos Style (CR LF)	Cleared	Specifies whether a combination of carriage-return and line-feed characters is used to indicate the end-of-line sequence in the input file.

Field (Continued)	Default	Description (Continued)
None	Cleared	Specifies whether to use an end-of-line terminator. None is enabled if Fixedwidth columns and Check data against metadata are selected.

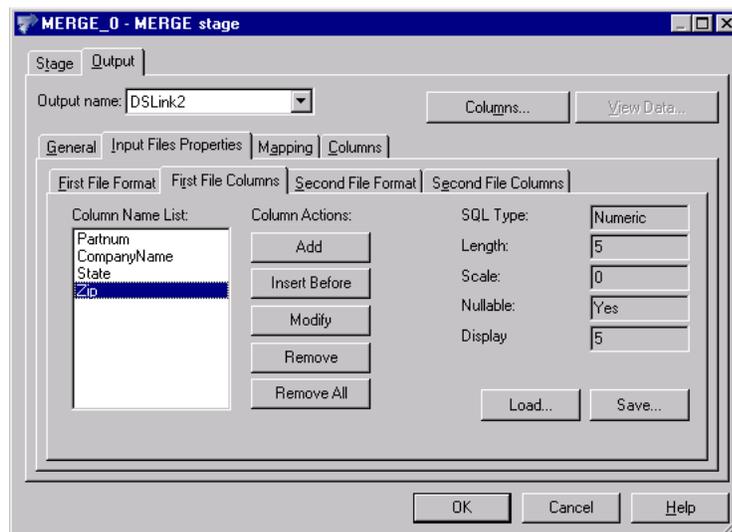
Specifying Input File Columns

Using the **First File Columns** and **Second File Columns** pages, you can specify the following:

- Column names of the first and second sequential input files
- Sequential file characteristics, including SQL type, length, scale, nullable, and display of the column
- Character set map used for the column

Click the **First** (or **Second**) **File Columns** tab on the **Input Files Properties** page.

The **First** (or **Second**) **File Columns** page appears, as shown:



You have two options when entering information about the columns:

- You can use information from an existing table to specify the input file columns.
- You can enter the column information manually.

Using Column Information from an Existing Table

You can use information from an existing table to define the columns in the first and second input files. Table definitions specify the data

used at each stage of a DataStage job and are stored in the Repository. For more information on tables, see DataStage documentation.

To transfer information about columns from an existing table:

- 1 Click **Load...** . The Table Definition dialog box appears.
- 2 Use the mouse to select the table definition in the left pane and click **OK**. The listed tables are already defined in the Repository.
 - a If you don't know the table definition, click **Find...** . The Find dialog box appears.
 - b In the **Find what** field, enter a text string. The first table definition that contains the text string you specify is highlighted in the left pane.
- 3 Once you select the file name, click **OK**.

Entering Column Information Manually

You can enter information about columns manually, by entering the information on the **First File Columns** page.

Enter a column name in the **Column Name List** and use the **Column Actions** buttons (**Add**, **Insert Before**, **Modify**, **Remove**, or **Remove All**) to specify where to put the names in the **Column Name List**.

You are then prompted to enter the information described in the following table:

Field	Default	Description
Column Name List	None	The names of each column in either the first or second files. These names are used in the Mapping page that defines the output link.
SQL Type	None	The SQL data type.
Length	0	Defines the data precision. It is the length for CHAR data or the maximum length for VARCHAR data. For numeric data, it is the number of digits of precision.
Scale	0	The data scale factor. For numeric data, it is the number of digits to the right of the decimal point.
Nullable	Yes	Specifies whether the column can contain null values.
Display	0	The maximum number of characters required to display the column data.

Field	Default	Description
NLS map		Specifies a different mapping for the column if per-column mapping is enabled (see "Defining Character Set Mapping" on page 5). Select a map from the drop-down list.

Saving Column Information in a Table

Once you have specified the column names and corresponding information the way you want, you can write that information to a new table. To save column information in a table:

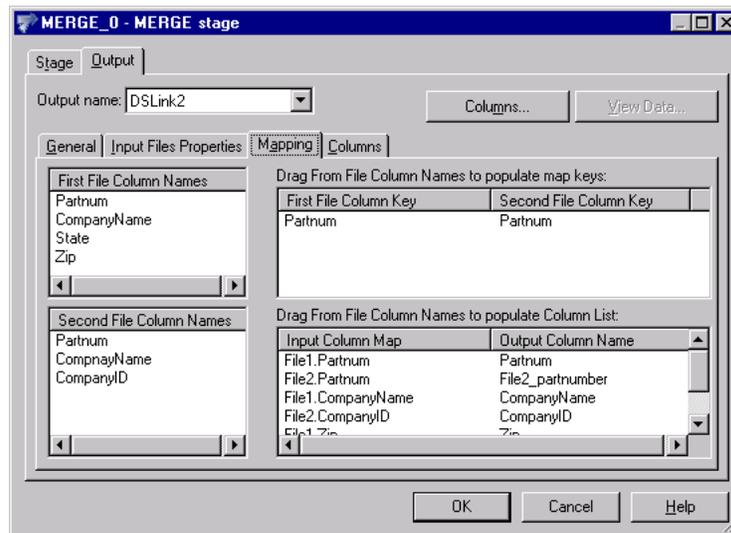
- 1 Click **Save...** . The Save Table Definition dialog box appears.
- 2 Complete the dialog box as follows:

Field	Default	Description
Data source type	Saved	The type of data written to the table. The data source type can be an ODBC data source, a UniVerse table, a hashed (UniVerse file), a UniData file, a sequential file, or a stage. The table definition is stored according to the data source in the Table Definitions branch.
Data source name	Link name	Forms the second part of the table definition identifier and provides the name of the branch created under the data source type. It provides a means to track where the data definition originated.
Table/file name	Link name	The table or file name containing the data.
Short description	Time and date saved	An optional brief description of the data.
Long description	None	An optional long description of the data.

Specifying Keys for the Join

You must specify the keys in the first and second sequential input files to be used in the join operation. To specify the keys, click the **Mapping** tab on the **Outputs** page.

The **Mapping** page appears as shown:



Select the keys from **First** (and **Second**) **File Column Names** on the left side of the page and drag them over to **First** (and **Second**) **File Column Key** on the right. These keys are used in the join operation to compare the two files.

You can specify multiple keys for the join operation. If you use multiple keys, you must have the same number of keys in the **First File Column Key** and **Second File Column Key** lists.

To delete an entry you made, select it and then right-click and choose **Clear Entry** from the pop-up menu.

Specifying Output Columns

You must specify the contents of the columns to be included in the output link. Use the **Mapping** page to specify the contents of these columns.

In the **Mapping** page, the **First File Column Names** and **Second File Column Names** are already defined. You defined these in the **Input Files Properties** page.

In the **Mapping** page, you must specify which columns from the input files you want included in the output link. To specify the contents of a column in the output link, select a column from the **First File Column Names** or **Second File Column Names** list box and drag the column to the **Input Column Map** list. The **Output Column Name** is automatically generated. The properties of the columns in the output link are derived from those in the input file. You must include a **First File Column Key** and **Second File Column Key** in the **Column List**.

If you want to explicitly specify the names and properties of the columns in the output link, go to the **Columns** page as described in "Defining the Name and Format of the Output Columns" on page 14.

You can select multiple columns at once to be dragged from the **First** (or **Second**) **File Column Names** list to the **Input Column Map** list. To select multiple columns, select the first column you want and hold down the **Ctrl** key until all the columns you want are highlighted. Or you can hold down the **Shift** key and click to select multiple columns.

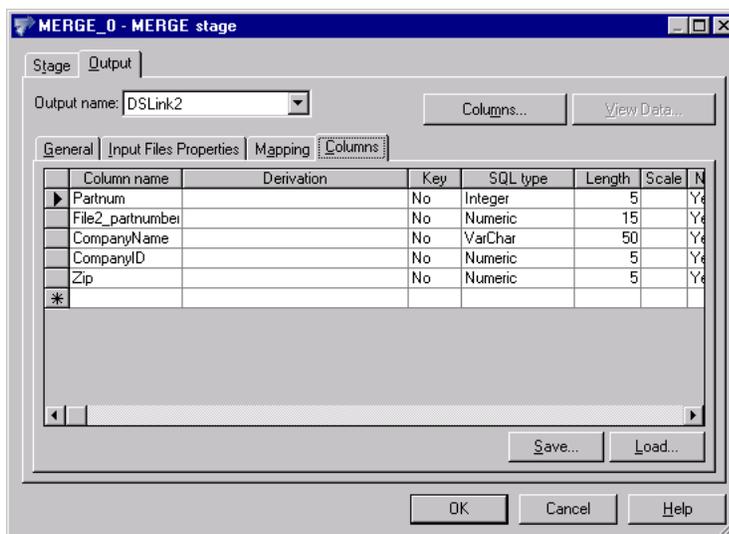
You can right-click to delete any item from the **Input Column Map** list. To delete columns from the output link, click the **Columns** tab, and delete the columns as described in "Defining the Name and Format of the Output Columns" on page 14.

Note If you change the **First File** (or **Second File**) **Column Names** on the left side of the page, you may need to verify the mapping information (that is, the map keys and column list) on the right side of the page. If the column names on the right side of the page do not match those on the left, drag the correct column names from the left side to the right side.

Defining the Name and Format of the Output Columns

You can use the **Columns** page to specify the name and the format of the columns in the output link. You can also use the **Columns** page to specify a different character set map for the column so that columns within a record can use different maps.

The following screen shows the **Columns** page:



The grid shown displays the standard fields used for all output links regardless of whether they are created in other stages.

As described in ["Using Column Information from an Existing Table" on page 10](#), you can use information from an existing table to specify the columns. Refer to that section for an explanation of how to use the **Load...** button to transfer information from a table.

Note You must set all columns to “Nullable,” except when the merge is set to **Pure Inner Join** as described in ["Specifying the Type of Join" on page 7](#).

Complete the cells in the grid as follows:

Field	Default	Description
Column name	None	The name of the column whose format you are defining.
Group	No	Specifies whether you want to group by this column.
Derivation	None	Allows you to specify that you want to summarize using this column.
Key	No	Defines whether the column is a key.
SQL type	(Unknown)	The SQL data type.
Length	None	Defines the data precision. It is the length for CHAR data or the maximum length for VARCHAR data. For numeric data, it is the number of digits of precision.
Scale	None	The data scale factor. For numeric data, it is the number of digits to the right of the decimal point.
Nullable	No	Specifies whether the column can contain null values. Must be Yes unless you are performing a Pure Inner Join.
Display	None	The maximum number of characters required to display the column data.
Data element	None	The type of data in the column.
Description	None	An optional text description of the column.
NLS map	Project default (MS1252)	If per-column mapping is enabled, the mapping performed for the column. Select one of the map names from the drop-down list.

Deleting Columns in the Output Link

Using the **Columns** page, you can delete columns you defined in the output link. To delete columns in the output link:

- 1 Select the row you want to delete.
- 2 Press the **Delete** key.