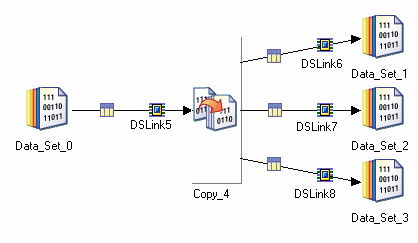
The Copy stage is a processing stage. The Copy stage copies a single input data set to a number of output data sets. Some SMP systems allow scalability of disk I/O.

Each record of the input data set is copied to every output data set. Records can be copied without modification or you can drop or change the order of columns (to copy with more modification - for example changing column data types - use the Modify stage as described in [Modify stage](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Modify_Stage.html?view=kc)). Copy lets you make a backup copy of a data set on disk while performing an operation on another copy, for example.



Where you are using a Copy stage with a single input and a single output, you should ensure that you set the Force property in the stage editor TRUE. This prevents InfoSphere® DataStage® from deciding that the Copy operation is superfluous and optimizing it out of the job.

The stage editor has three pages:

* [**Stage Page**](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Stage_Page_copy_stage.html?view=kc). This is always present and is used to specify general information about the stage.
* [**Input Page**](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Input_Page_copy_stage.html?view=kc). This is where you specify details about the input link carrying the data to be copied.
* [**Output Page**](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Outputs_Page_copy_stage.html?view=kc). This is where you specify details about the copied data being output from the stage.

In this example you are going to copy data from a table containing billing information for GlobalCo's customers.

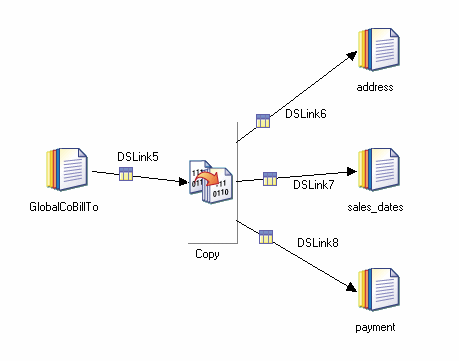
You are going to copy it to three separate data sets, and in each case you are only copying a subset of the columns. The Copy stage will drop the unwanted columns as it copies the data set.

The column names for the input data set are as follows:

* BILL\_TO\_NUM
* CUST\_NAME
* ADDR\_1
* ADDR\_2
* CITY
* REGION\_CODE
* ZIP
* ATTENT
* COUNTRY\_CODE
* TEL\_NUM
* FIRST\_SALES\_DATE
* LAST\_SALES\_DATE
* REVIEW\_MONTH
* SETUP\_DATE
* STATUS\_CODE
* REMIT\_TO\_CODE
* CUST\_TYPE\_CODE
* CUST\_VEND
* MOD\_DATE
* MOD\_USRNM
* CURRENCY\_CODE
* CURRENCY\_MOD\_DATE
* MAIL\_INVC\_FLAG
* PYMNT\_CODE
* YTD\_SALES\_AMT
* CNTRY\_NAME
* CAR\_RTE
* TPF\_INVC\_FLAG,
* INVC\_CPY\_CNT
* INVC\_PRT\_FLAG
* FAX\_PHONE,
* FAX\_FLAG
* ANALYST\_CODE
* ERS\_FLAG

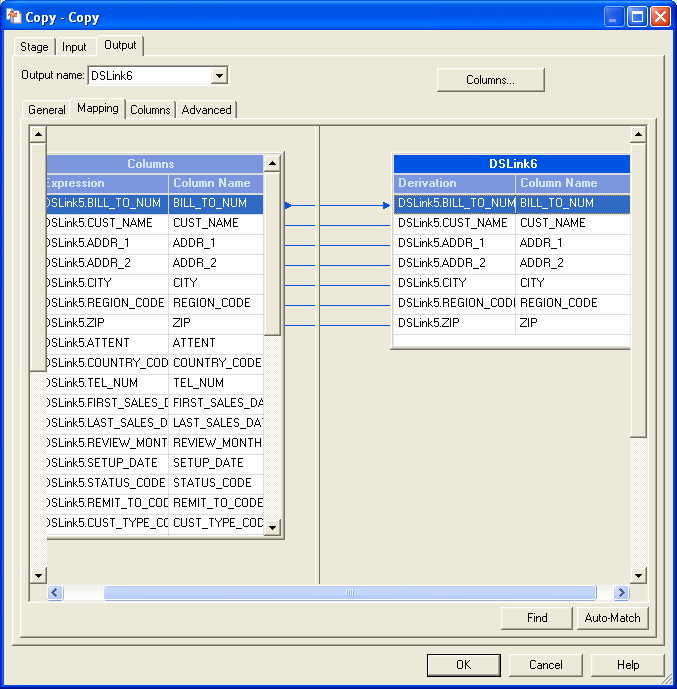
Here is the job that will perform the copying:

*Figure 1. Example job*

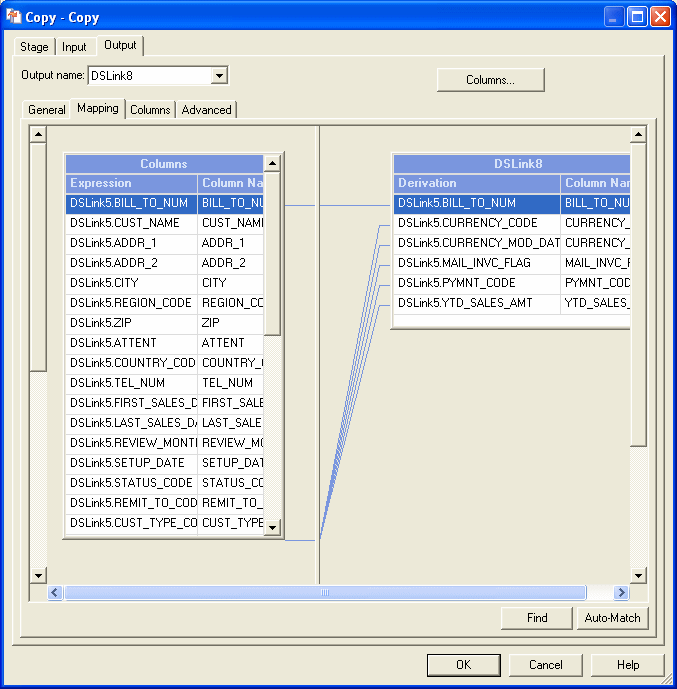


The Copy stage properties are fairly simple. The only property is Force, and you do not need to set it in this instance as you are copying to multiple data sets (and InfoSphere® DataStage® will not attempt to optimize it out of the job). You need to concentrate on telling InfoSphere DataStage which columns to drop on each output link. The easiest way to do this is using the Output page Mapping tab. When you open this for a link the left pane shows the input columns, simply drag the columns you want to preserve across to the right pane. You repeat this for each link as follows:

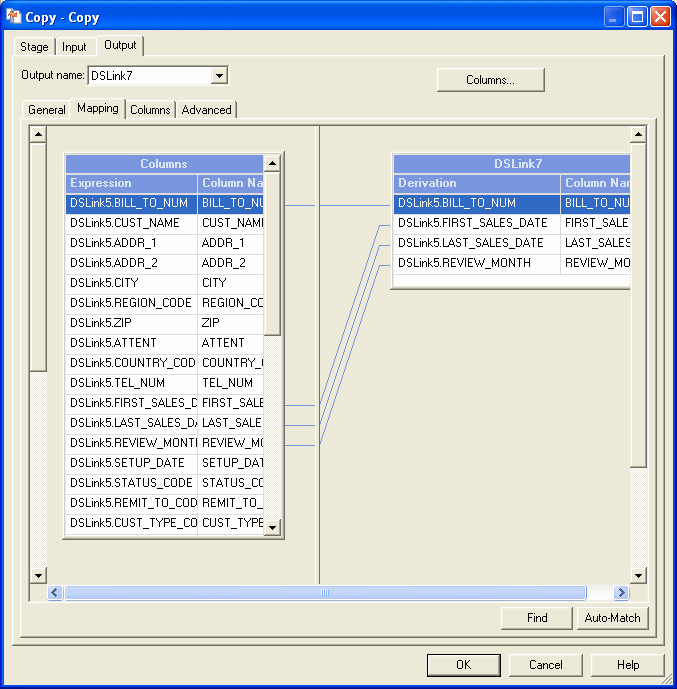
*Figure 2. Mapping tab: first output link*



*Figure 3. Mapping tab: second output link*



*Figure 4. Mapping tab: third output link*



When the job is run, three copies of the original data set are produced, each containing a subset of the original columns, but all of the rows. Here is some sample data from each of the data set on DSLink6, which gives name and address information:

"GC13849","JON SMITH","789 LEDBURY ROAD"," ","TAMPA","FL","12345"

"GC13933","MARY GARDENER","127 BORDER ST"," ","NORTHPORT","AL","23456"

"GC14036","CHRIS TRAIN","1400 NEW ST"," ","BRENHAM","TX","34567"

"GC14127","HUW WILLIAMS","579 DIGBETH AVENUE"," ","AURORA","CO","45678"

"GC14263","SARA PEARS","45 ALCESTER WAY"," ","SHERWOOD","AR","56789"

"GC14346","LUC TEACHER","3 BIRMINGHAM ROAD"," ","CHICAGO","IL","67890"[Copy](javascript:void(0);)

This section specifies the minimum steps to take to get a Copy stage functioning.

**About this task**

InfoSphere® DataStage® has many defaults which means that it can be very easy to include Copy stages in a job. InfoSphere DataStage provides a versatile user interface, and there are many shortcuts to achieving a particular end, this section describes the basic method, you will learn where the shortcuts are when you get familiar with the product.

**Procedure**

1. Ensure that meta data has been defined for input link and output links.
2. In the **Output Page** [**Mapping Tab**](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Mapping_Tab_copy_stage.html?view=kc), specify how the input columns of the data set being copied map onto the columns of the various output links.

The General tab allows you to specify an optional description of the stage. The Properties tab lets you specify what the stage does. The Advanced tab allows you to specify how the stage executes.

Use the Properties tab to specify how the Copy stage operates.

The Properties tab allows you to specify properties which determine what the stage actually does. The Copy stage only has one property.

| **Category/Property** | **Values** | **Default** | **Mandatory?** | **Repeats?** | **Dependent of** |
| --- | --- | --- | --- | --- | --- |
| Options/[Force](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_copy_stage.html?view=kc) | True/False | False | N | N | N/A |
| *Table 1. Properties* | | | | | |

Use the Options category to specify how the Copy stage operates.

## Force

Set True to specify that InfoSphere® DataStage® should not try to optimize the job by removing a Copy operation where there is one input and one output. Set False by default.

This tab allows you to specify options.

This tab allows you to specify the following:

* **Execution Mode**. The stage can execute in parallel mode or sequential mode. In parallel mode the input data is processed by the available nodes as specified in the Configuration file, and by any node constraints specified on the Advanced tab. In Sequential mode the entire data set is processed by the conductor node.
* **Combinability mode**. This is Auto by default, which allows InfoSphere® DataStage® to combine the operators that underlie parallel stages so that they run in the same process if it is sensible for this type of stage.
* **Preserve partitioning**. This is **Propagate** by default. It adopts the setting of the previous stage.You can explicitly select **Set**or **Clear**. Select **Set** to request the stage should attempt to maintain the partitioning.
* **Node pool and resource constraints**. Select this option to constrain parallel execution to the node pool or pools or resource pool or pools specified in the grid. The grid allows you to make choices from drop down lists populated from the Configuration file.
* **Node map constraint**. Select this option to constrain parallel execution to the nodes in a defined node map. You can define a node map by typing node numbers into the text box or by clicking the browse button to open the **Available Nodes** dialog box and selecting nodes from there. You are effectively defining a new node pool for this stage (in addition to any node pools defined in the Configuration file).

**Note**In the **Node map constraint** text box, you can enter jobs parameters as well as numbers. You can enter a single parameter, for example #testnode#, or you can enter a comma separated lists of parameters, for example #testnode#, #testnode2#. The browse button next to the text box will display a list of the node names from the last configuration file that was referenced by the job, but the browse button will not display the node names that were specified by the job parameters.

The Input page allows you to specify details about the data set being copied. There is only one input link.

The General tab allows you to specify an optional description of the link. The Partitioning tab allows you to specify how incoming data on the source data set link is partitioned. The Columns tab specifies the column definitions of incoming data. The Advanced tab allows you to change the default buffering settings for the input link.

Details about Copy stage partitioning are given in the following section. See ["Stage Editors,"](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Stage_Editors.html?view=kc) for a general description of the other tabs.

The Partitioning tab allows you to specify details about how the incoming data is partitioned or collected before the copy is performed.

By default the stage uses the auto partitioning method.

If the Copy stage is operating in sequential mode, it will first collect the data before writing it to the file using the default auto collection method.

The Partitioning tab allows you to override this default behavior. The exact operation of this tab depends on:

* Whether the Copy stage is set to execute in parallel or sequential mode.
* Whether the preceding stage in the job is set to execute in parallel or sequential mode.

If the Copy stage is set to execute in parallel, then you can set a partitioning method by selecting from the **Partition type** drop-down list. This will override any current partitioning.

If the Copy stage is set to execute in sequential mode, but the preceding stage is executing in parallel, then you can set a collection method from the **Collector type** drop-down list. This will override the default auto collection method.

The following partitioning methods are available:

* **(Auto)**. InfoSphere® DataStage® attempts to work out the best partitioning method depending on execution modes of current and preceding stages and how many nodes are specified in the Configuration file. This is the default method for the Copy stage.
* **Entire**. Each file written to receives the entire data set.
* **Hash**. The records are hashed into partitions based on the value of a key column or columns selected from the **Available**list.
* **Modulus**. The records are partitioned using a modulus function on the key column selected from the **Available** list. This is commonly used to partition on tag fields.
* **Random**. The records are partitioned randomly, based on the output of a random number generator.
* **Round Robin**. The records are partitioned on a round robin basis as they enter the stage.
* **Same**. Preserves the partitioning already in place.
* **Db2®**. Replicates the Db2 partitioning method of a specific Db2 table. Requires extra properties to be set. Access these properties by clicking the properties button.
* **Range**. Divides a data set into approximately equal size partitions based on one or more partitioning keys. Range partitioning is often a preprocessing step to performing a total sort on a data set. Requires extra properties to be set. Access these properties by clicking the properties button.

The following Collection methods are available:

* **(Auto)**. This is the default collection method for the Copy stage. Normally, when you are using Auto mode, InfoSphere DataStage will eagerly read any row from any input partition as it becomes available.
* **Ordered**. Reads all records from the first partition, then all records from the second partition, and so on.
* **Round Robin**. Reads a record from the first input partition, then from the second partition, and so on. After reaching the last partition, the operator starts over.
* **Sort Merge**. Reads records in an order based on one or more columns of the record. This requires you to select a collecting key column from the **Available** list.

The Partitioning tab also allows you to specify that data arriving on the input link should be sorted before the remove duplicates operation is performed. The sort is always carried out within data partitions. If the stage is partitioning incoming data the sort occurs after the partitioning. If the stage is collecting data, the sort occurs before the collection. The availability of sorting depends on the partitioning or collecting method chosen (it is not available for the default auto methods).

Select the check boxes as follows:

* **Perform Sort**. Select this to specify that data coming in on the link should be sorted. Select the column or columns to sort on from the **Available** list.
* **Stable**. Select this if you want to preserve previously sorted data sets. This is the default.
* **Unique**. Select this to specify that, if multiple records have identical sorting key values, only one record is retained. If stable sort is also set, the first record is retained.

If NLS is enabled an additional button opens a dialog box allowing you to select a locale specifying the collate convention for the sort.

You can also specify sort direction, case sensitivity, whether sorted as ASCII or EBCDIC, and whether null columns will appear first or last for each column. Where you are using a keyed partitioning method, you can also specify whether the column is used as a key for sorting, for partitioning, or for both. Select the column in the **Selected** list and right-click to invoke the shortcut menu.

In the Output page, you can specify details about data output from the Copy stage. The tabs in this stage allow you to specify an optional description and the relationship between the columns being input and the Output columns. You can change the default buffer settings for the output link and view the column definitions.

The Output page allows you to specify details about data output from the Copy stage. The stage can have any number of output links, choose the one you want to work on from the **Output name** drop down list.

The General tab allows you to specify an optional description of the output link. The Columns tab specifies the column definitions of the data. The Mapping tab allows you to specify the relationship between the columns being input to the Copy stage and the output columns. The Advanced tab allows you to change the default buffering settings for the output links.

Details about Copy stage mapping is given in the following section. See ["Stage Editors,"](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Stage_Editors.html?view=kc) for a general description of the other tabs.

For the Copy stage, the Mapping tab allows you to specify how the output columns are derived, that is, what copied columns map onto them.

The left pane shows the copied columns. These are read only and cannot be modified on this tab.

The right pane shows the output columns for the output link. This has a **Derivations** field where you can specify how the column is derived. You can fill it in by dragging copied columns over, or by using the Auto-match facility.