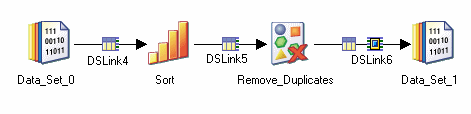
The Remove Duplicates stage is a processing stage. It can have a single input link and a single output link.

The Remove Duplicates stage takes a single sorted data set as input, removes all duplicate rows, and writes the results to an output data set.



Removing duplicate records is a common way of cleansing a data set before you perform further processing. Two rows are considered duplicates if they are adjacent in the input data set and have identical values for the key column(s). A key column is any column you designate to be used in determining whether two rows are identical.

The data set input to the Remove Duplicates stage must be sorted so that all records with identical key values are adjacent. You can either achieve this using the in-stage sort facilities available on the Input page Partitioning tab, or have an explicit Sort stage feeding the Remove Duplicates stage.

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_remove_duplicate_stages.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_Inputs_Page_remove_duplicate_stage.html?view=kc). This is where you specify details about the data set that is having its duplicates removed.
* [**Output Page**](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Output_Page_remove_duplicate_stages.html?view=kc). This is where you specify details about the processed data that is being output from the stage.

In the example the data is a list of GlobalCo's customers. The data contains some duplicate entries, and you want to remove these.

The first step is to sort the data so that the duplicates are actually next to each other. As with all sorting operations, there are implications around data partitions if you run the job in parallel (see ["Copy Stage,"](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Copy_Stage.html?view=kc) for a discussion of these). You should hash partition the data using the sort keys as hash keys in order to guarantee that duplicate rows are in the same partition. In the example you sort on the CUSTOMER\_NUMBER columns and the sample of the sorted data shows up some duplicates:

"GC13849","JON SMITH","789 LEDBURY ROAD","2/17/2007"

"GC13933","MARY GARDENER","127 BORDER ST","8/28/2009"

"GC13933","MARY GARDENER","127 BORDER ST","8/28/2009"

"GC14036","CHRIS TRAIN","1400 NEW ST","9/7/1998"

"GC14127","HUW WILLIAMS","579 DIGBETH AVENUE","6/29/2011"

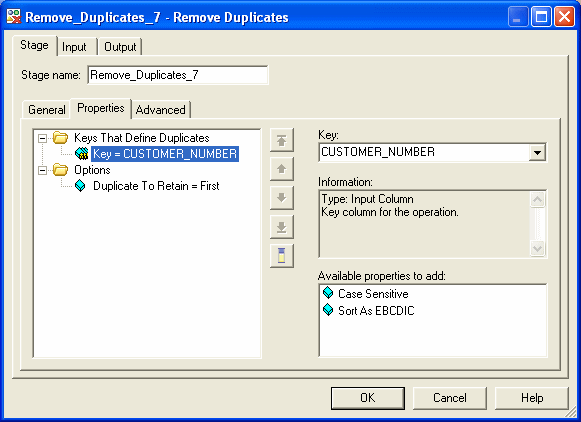
"GC14263","SARA PEARS","45 ALCESTER WAY","4/12/2008"

"GC14263","SARA PEARS","45 ALCESTER WAY","4/12/2008"

"GC14346","LUC TEACHER","3 BIRMINGHAM ROAD","11/7/2010"[Copy](javascript:void(0);)

Next, you set up the Remove Duplicates stage to remove rows that share the same values in the CUSTOMER\_NUMBER column. The stage will retain the first of the duplicate records:

*Figure 1. Property settings*



Here is a sample of the data after the job has been run and the duplicates removed:

"GC13849","JON SMITH","789 LEDBURY ROAD","2/17/2007"

"GC13933","MARY GARDENER","127 BORDER ST","8/28/2009"

"GC14036","CHRIS TRAIN","1400 NEW ST","9/7/1998"

"GC14127","HUW WILLIAMS","579 DIGBETH AVENUE","6/29/2011"

"GC14263","SARA PEARS","45 ALCESTER WAY","4/12/2008"

"GC14346","LUC TEACHER","3 BIRMINGHAM ROAD","11/7/2010"

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

**About this task**

InfoSphere® DataStage® has many defaults which means that it can be very easy to include Remove Duplicates 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.

To use a Remove Duplicates stage:

* In the Stage page [**Properties Tab**](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Properties_Tab_remove_duplicate_stages.html?view=kc) select the key column. Identical values in this column will be taken to denote duplicate rows, which the stage will remove. Repeat the property to specify a composite key.
* 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_remove_duplicate_stages.html?view=kc), specify how output columns are derived.

You can specify aspects of the Remove Duplicates stage from the Remove Duplicates stage: Stage page.

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. The NLS Locale tab appears if your have NLS enabled on your system. It allows you to select a locale other than the project default to determine collating rules.

Use the Properties tab to specify how the Remove Duplicates stage operates.

The Properties tab allows you to specify properties which determine what the stage actually does. Some of the properties are mandatory, although many have default settings. Properties without default settings appear in the warning color (red by default) and turn black when you supply a value for them.

The following table gives a quick reference list of the properties and their attributes. A more detailed description of each property follows.

| **Category/Property** | **Values** | **Default** | **Mandatory?** | **Repeats?** | **Dependent of** |
| --- | --- | --- | --- | --- | --- |
| Keys that Define Duplicates/[Key](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Keys_that_Define_Duplicates_Category.html?view=kc) | Input Column | N/A | Y | Y | N/A |
| Keys that Define Duplicates/[Sort as EBCDIC](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Keys_that_Define_Duplicates_Category.html?view=kc) | True/False | False | N | N | Key |
| Keys that Define Duplicates/[Case Sensitive](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Keys_that_Define_Duplicates_Category.html?view=kc) | True/False | True | N | N | Key |
| Options/[Duplicate to retain](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_remove_duplicate_stages.html?view=kc) | First/Last | First | Y | N | N/A |
| *Table 1. Properties* | | | | | |

Use the keys that define duplicate category to specify how the Remove Duplicates stage operates.

**Key**

Specifies the key column for the operation. This property can be repeated to specify multiple key columns. You can use the Column Selection dialog box to select several keys at once if required. Key has dependent properties as follows:

* **Sort as EBCDIC**

To sort as in the EBCDIC character set, choose True.

* **Case Sensitive**

Use this to specify whether each key is case sensitive or not, this is set to True by default, that is, the values "CASE" and "case" would not be judged equivalent.

Use the Options category to specify how the Remove Duplicate stage operates.

## Duplicate to retain

Specifies which of the duplicate columns encountered to retain. Choose between First and Last. It is set to First 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 **Set** or **Clear** from the previous stage. You can explicitly select **Set** or **Clear**. Select **Set** to request that next stage in the job 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.

For the Remove Duplicates stage, the NLS Locale tab appears if you have NLS enabled on your system. It lets you view the current default collate convention, and select a different one for this stage if required.

You can also use a job parameter to specify the locale, or browse for a file that defines custom collate rules. The collate convention defines the order in which characters are collated. The Remove Duplicates stage uses this when it is determining the sort order for the key column(s). Select a locale from the list, or click the arrow button next to the list to use a job parameter or browse for a collate file.

The Input page allows you to specify details about the data coming in to be sorted. Choose an input link from the **Input name**drop down list to specify which link you want to work on.

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 Remove Duplicates 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 operation is performed.

By default the stage uses the auto partitioning method.

If the Remove Duplicates 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 Remove Duplicates 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 Remove Duplicates 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 Remove Duplicates 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 Remove Duplicates 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 Remove Duplicates 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 with 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 Remove stage.

The Output page allows you to specify details about data output from the Remove Duplicates stage. The stage only has one output link.

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 Remove Duplicates stage and the output columns. The Advanced tab allows you to change the default buffering settings for the output link.

Details about Remove Duplicates 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 Remove Duplicates stage, the Mapping tab allows you to specify how the output columns are derived, that is, what input columns map onto them.

The left pane shows the columns of the input data. These are read only and cannot be modified on this tab. This shows the meta data from the incoming link.

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