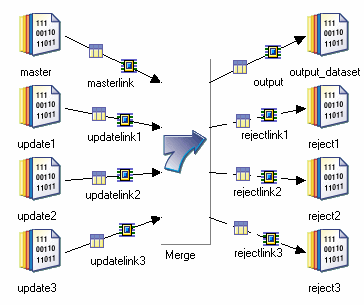
The Merge stage is a processing stage. It can have any number of input links, a single output link, and the same number of reject links as there are update input links. The Merge stage combines a master data set with one or more update data sets.

The Merge stage is one of three stages that join tables based on the values of key columns. The other two are:

* [Join stage](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Join_Stage.html?view=kc)
* [Lookup Stage](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Lookup_Stage.html?view=kc)

The three stages differ mainly in the memory they use, the treatment of rows with unmatched keys, and their requirements for data being input (for example, whether it is sorted).

The Merge stage combines a master data set with one or more update data sets. The columns from the records in the master and update data sets are merged so that the output record contains all the columns from the master record plus any additional columns from each update record that are required. A master record and an update record are merged only if both of them have the same values for the merge key column(s) that you specify. Merge key columns are one or more columns that exist in both the master and update records.



The data sets input to the Merge stage must be key partitioned and sorted. This ensures that rows with the same key column values are located in the same partition and will be processed by the same node. It also minimizes memory requirements because fewer rows need to be in memory at any one time. Choosing the auto partitioning method will ensure that partitioning and sorting is done. If sorting and partitioning are carried out on separate stages before the Merge stage, InfoSphere® DataStage® in auto partition mode will detect this and not repartition (alternatively you could explicitly specify the Same partitioning method).

As part of preprocessing your data for the Merge stage, you should also remove duplicate records from the master data set. If you have more than one update data set, you must remove duplicate records from the update data sets as well. See [Remove Duplicates Stage](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Remove_Duplicates_Stage.html?view=kc) for information about the Remove Duplicates stage.

Unlike Join stages and Lookup stages, the Merge stage allows you to specify several reject links. You can route update link rows that fail to match a master row down a reject link that is specific for that link. You must have the same number of reject links as you have update links. The **Link Ordering** tab on the Stage page lets you specify which update links send rejected rows to which reject links. You can also specify whether to drop unmatched master rows, or output them on the output data link.

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_merge_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_Inputs_Page_merge_stage.html?view=kc). This is where you specify details about the data sets being merged.
* [**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_merge_stage.html?view=kc). This is where you specify details about the merged data being output from the stage and about the reject links.

This example shows what happens to a master data set and two update data sets when they are merged.

The key field is Horse, and all the data sets are sorted in descending order. Here is the master data set:

| **Horse** | **Freezemark** | **Mchip** | **Reg\_Soc** | **Level** |
| --- | --- | --- | --- | --- |
| William | DAM7 | N/A | FPS | Adv |
| Robin | DG36 | N/A | FPS | Nov |
| Kayser | N/A | N/A | AHS | N/A |
| Heathcliff | A1B1 | N/A | N/A | Adv |
| Fairfax | N/A | N/A | FPS | N/A |
| Chaz | N/A | a296100da | AHS | Inter |
| *Table 1. Master data set* | | | | |

Here is the Update 1 data set:

| **Horse** | **vacc.** | **last\_worm** |
| --- | --- | --- |
| William | 07.07.02 | 12.10.02 |
| Robin | 07.07.02 | 12.10.02 |
| Kayser | 11.12.02 | 12.10.02 |
| Heathcliff | 07.07.02 | 12.10.02 |
| Fairfax | 11.12.02 | 12.10.02 |
| Chaz | 10.02.02 | 12.10.02 |
| *Table 2. Update 1 data set* | | |

Here is the Update 2 data set:

| **Horse** | **last\_trim** | **shoes** |
| --- | --- | --- |
| William | 11.05.02 | N/A |
| Robin | 12.03.02 | refit |
| Kayser | 11.05.02 | N/A |
| Heathcliff | 12.03.02 | new |
| Fairfax | 12.03.02 | N/A |
| Chaz | 12.03.02 | new |
| *Table 3. Update 2 data set* | | |

Here is the merged data set output by the stage:

| **Horse** | **Freeze mark** | **Mchip** | **Reg. Soc** | **Level** | **vacc.** | **last worm** | **last trim** | **shoes** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| William | DAM7 | N/A | FPS | Adv | 07.07.02 | 12.10.02 | 11.05.02 | N/A |
| Robin | DG36 | N?A | FPS | Nov | 07.07.02 | 12.10.02 | 12.03.02 | refit |
| Kayser | N/A | N/A | AHS | Nov | 11.12.02 | 12.10.02 | 11.05.02 | N/A |
| Heathcliff | A1B1 | N/A | N/A | Adv | 07.07.02 | 12.10.02 | 12.03.02 | new |
| Fairfax | N/A | N/A | FPS | N/A | 11.12.02 | 12.10.02 | 12.03.02 | N/A |
| Chaz | N/A | a2961da | AHS | Inter | 10.02.02 | 12.10.02 | 12.03.02 | new |
| *Table 4. Merged data set* | | | | | | | | |

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

**About this task**

InfoSphere® DataStage® has many defaults which means that Merges can be simple to set up. 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.

* 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_merge_stage.html?view=kc) specify the key column or columns that the Merge will be performed on.
* 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_merge_stage.html?view=kc) set the Unmatched Masters Mode, Warn on Reject Updates, and Warn on Unmatched Masters options or accept the defaults.
* In the Stage page [**Link Ordering Tab**](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Link_Ordering_Tab_merge_stage.html?view=kc), check that your input links are correctly identified as "master" and "update(s)", and your output links are correctly identified as "master" and "update reject". Reorder if required.
* Ensure required column meta data has been specified (this might be done in another stage, or might be omitted altogether if you are relying on Runtime Column Propagation).
* 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_merge_stage.html?view=kc), specify how the columns from the input links map onto output columns.

You can specify aspects of the Merge stage from the Merge 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 Merge stage operates.

The Properties tab allows you to specify properties that 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** |
| --- | --- | --- | --- | --- | --- |
| Merge Keys/[Key](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Merge_Keys_Category.html?view=kc) | Input Column | N/A | Y | Y | N/A |
| Merge Keys/[Sort Order](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Merge_Keys_Category.html?view=kc) | Ascending/ Descending | Ascending | Y | N | Key |
| Merge Keys/[Nulls position](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Merge_Keys_Category.html?view=kc) | First/Last | First | N | N | Key |
| Merge Keys/[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_Merge_Keys_Category.html?view=kc) | True/False | False | N | N | Key |
| Merge Keys/[Case Sensitive](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Merge_Keys_Category.html?view=kc) | True/False | True | N | N | Key |
| Options/[Unmatched Masters Mode](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_merge_stage.html?view=kc) | Keep/Drop | Keep | Y | N | N/A |
| Options/[Warn On Reject Masters](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_merge_stage.html?view=kc) | True/False | True | Y | N | N/A |
| Options/[Warn On Reject Updates](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_merge_stage.html?view=kc) | True/False | True | Y | N | N/A |
| *Table 1. Properties* | | | | | |

Use the Merge keys category to specify how the Merge stage operates.

## Key

This specifies the key column you are merging on. Repeat the property to specify multiple keys. You can use the Column Selection dialog box to select several keys at once if required. Key has the following dependent properties:

* **Sort Order**

Choose Ascending or Descending. The default is Ascending.

* **Nulls position**

By default columns containing null values appear first in the merged data set. To override this default so that columns containing null values appear last in the merged data set, select Last.

* **Sort as EBCDIC**

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

* **Case Sensitive**

Use this to specify whether each merge 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 Merge stage operates.

## Unmatched Masters Mode

Set to Keep by default. It specifies that unmatched rows from the master link are output to the merged data set. Set to Drop to specify that rejected records are dropped instead.

## Warn On Reject Masters

Set to True by default. This will warn you when bad records from the master link are rejected. Set it to False to receive no warnings.

## Warn On Reject Updates

Set to True by default. This will warn you when bad records from any update links are rejected. Set it to False to receive no warnings.

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 which results from ORing the settings of the input stages, that is, if any of the input stages uses **Set** then this stage will use **Set**. You can explicitly select **Set** or **Clear**. Select **Set** to request the next stage in the job attempts 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.

This tab allows you to specify which of the input links is the master link and the order in which links input to the Merge stage are processed.

You can also specify which of the output links is the master link, and which of the reject links corresponds to which of the incoming update links.

By default the links will be processed in the order they were added. To rearrange them, choose an input link and click the up arrow button or the down arrow button.

For the Merge 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 Merge stage uses this when it is determining the order of the key fields. 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 merged. Choose an input link from the **Input name**drop down list to specify which link you want to work on.

The Input page allows you to specify details about the data coming in to be merged. 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 Merge 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 merge is performed.

By default the stage uses the auto partitioning method. If the Preserve Partitioning option has been set on the previous stage in the job, this stage will warn you if it cannot preserve the partitioning of the incoming data. Auto mode ensures that data being input to the Merge stage is key partitioned and sorted.

If the Merge 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 Merge 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 Merge 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 Merge 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 collection method for the Merge 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 Merge stage. Normally, when you are using Auto mode, InfoSphere DataStage will eagerly read any row from any input partition as it becomes available. In the case of a Merge stage, Auto will also ensure that the collected data is sorted.
* **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 merge is performed (you might use this if you have selected a partitioning method other than auto or same). 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.

The Output page allows you to specify details about data output from the Merge stage.

The Merge stage can have only one master output link carrying the merged data and a number of reject links, each carrying rejected records from one of the update links. Choose an output link from the **Output 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 output link. The Columns tab specifies the column definitions of incoming data. The Mapping tab allows you to specify the relationship between the columns being input to the Merge stage and the Output columns. The Advanced tab allows you to change the default buffering settings for the output links.

Details about Merge 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.

Use the Reject links to specify how the Merge stage operates.

You cannot change the properties of a Reject link. They have the meta data of the corresponding incoming update link and this cannot be altered.

For the Merge 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 merged data. These are read only and cannot be modified on this tab. This shows the meta data from the master input link and any additional columns carried on the update links.

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.

In the above example, the left pane represents the incoming data after the merge has been performed. The right pane represents the data being output by the stage after the merge operation. In this example the data has been mapped straight across.