The Change Capture stage is a processing stage that compares two data sets and makes a record of the differences.

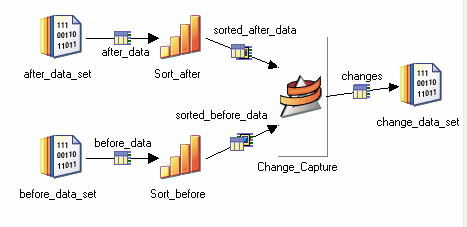
The Change Capture Stage is a processing stage. The stage compares two data sets and makes a record of the differences.

The Change Capture stage takes two input data sets, denoted before and after, and outputs a single data set whose records represent the changes made to the before data set to obtain the after data set. The stage produces a change data set, whose table definition is transferred from the after data set's table definition with the addition of one column: a change code with values encoding the four actions: insert, delete, copy, and edit. The preserve-partitioning flag is set on the change data set.

The compare is based on a set a set of key columns, rows from the two data sets are assumed to be copies of one another if they have the same values in these key columns. You can also optionally specify change values. If two rows have identical key columns, you can compare the value columns in the rows to see if one is an edited copy of the other.

The stage assumes that the incoming data is key-partitioned and sorted in ascending order. The columns the data is hashed on should be the key columns used for the data compare. You can achieve the sorting and partitioning using the Sort stage or by using the built-in sorting and partitioning abilities of the Change Capture stage.

You can use the companion Change Apply stage to combine the changes from the Change Capture stage with the original before data set to reproduce the after data set (see [Switch stage](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Switch_Stage.html?view=kc)).



The Change Capture stage is very similar to the Difference stage described in [Difference stage](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Difference_Stage.html?view=kc).

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_change_capture_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_change_capture_stage.html?view=kc). This is where you specify details about the data set 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_Outputs_Page_change_capture_stage.html?view=kc). This is where you specify details about the processed data being output from the stage.

This example shows a *before* and *after* data set, and the data set that is output by the Change Capture stage.

This example shows a *before* and *after* data set, and the data set that is output by the Change Capture stage when it has compared them.

This is the *before* data set:

| **bcol0** | **bcol1** | **bcol2** | **bcol3** | **bcol4** |
| --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 0 | a |
| 1 | 7 | 1 | 1 | b |
| 2 | 2 | 2 | 2 | c |
| 3 | 3 | 3 | 3 | d |
| 4 | 5 | 4 | 4 | e |
| 5 | 2 | 5 | 5 | f |
| 6 | 6 | 6 | 6 | g |
| 7 | 7 | 7 | 7 | h |
| 8 | 8 | 8 | 8 | i |
| 9 | 9 | 9 | 9 | j |
| *Table 1. Before data set* | | | | |

This is the *after* data set:

| **bcol0** | **bcol1** | **bcol2** | **bcol3** | **bcol4** |
| --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 0 | a |
| 1 | 1 | 1 | 1 | b |
| 2 | 2 | 2 | 2 | c |
| 3 | 3 | 3 | 3 | d |
| 4 | 4 | 4 | 4 | e |
| 5 | 5 | 5 | 5 | f |
| 6 | 6 | 6 | 6 | g |
| 7 | 7 | 7 | 7 | h |
| 8 | 8 | 8 | 8 | i |
| 9 | 9 | 9 | 9 | j |
| *Table 2. After data set* | | | | |

This is the data set output by the Change Capture stage (bcol4 is the key column, bcol1 the value column):

| **bcol0** | **bcol1** | **bcol2** | **bcol3** | **bcol4** | **change\_code** |
| --- | --- | --- | --- | --- | --- |
| 1 | 1 | 1 | 1 | b | 3 |
| 4 | 4 | 4 | 4 | e | 3 |
| 5 | 5 | 5 | 5 | f | 3 |
| *Table 3. Change data set* | | | | | |

The change\_code indicates that, in these three rows, the bcol1 column in the after data set has been edited. The bcol1 column carries the edited value.

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

**About this task**

InfoSphere® DataStage® has many defaults which means that it can be very easy to include Change Capture 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. Go to the [**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_change_capture_stage.html) on the Stage page.
2. Specify the key column. You can repeat this property to specify a composite key. *Before* and *after* rows are considered to be the same if they have the same value in the key column or columns.
3. Optionally specify one or more Value columns. This enables you to determine if an *after* row is an edited version of a *before*row.

**Note**You can also set the Change Mode property to have InfoSphere DataStage treat all columns not defined as keys treated as values, or all columns not defined as values treated as keys.

1. Specify whether the stage will output the changed row or drop it. You can specify this individually for each type of change (copy, delete, edit, or insert).
2. 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_change_capture_stage.html), specify which of the two links carries the *before* data set and which carries the *after*data set.
3. If the two incoming data sets aren't already key partitioned on the key columns and sorted, set InfoSphere DataStage to do this on the **Input Page** [**Partitioning Tab**](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Partitioning_Tab_change_capture_stage.html).
4. 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_change_capture_stage.html), specify how the change data columns are mapped onto the output link columns.

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 Link Ordering tab allows you to specify which input link carries the before data set and which the after data set. 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 Change Capture 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** |
| --- | --- | --- | --- | --- | --- |
| Change Keys/[Key](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Change_Keys_Category_change_capture_stage.html?view=kc) | Input Column | N/A | Y | Y | N/A |
| Change 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_Change_Keys_Category_change_capture_stage.html?view=kc) | True/False | True | N | N | Key |
| Change 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_Change_Keys_Category_change_capture_stage.html?view=kc) | Ascending/ Descending | Ascending | N | N | Key |
| Change 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_Change_Keys_Category_change_capture_stage.html?view=kc) | First/Last | First | N | N | Key |
| Change Values/[Value](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Change_Value_category_change_capture_stage.html?view=kc) | Input Column | N/A | N | Y | N/A |
| Change Values/[Case Sensitive](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Change_Value_category_change_capture_stage.html?view=kc) | True/False | True | N | N | Value |
| Options/[Change 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_change_capture_stage.html?view=kc) | Explicit Keys & Values/All keys, Explicit values/Explicit Keys, All Values | Explicit Keys & Values | Y | N | N/A |
| Options/[Log Statistics](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_change_capture_stage.html?view=kc) | True/False | False | N | N | N/A |
| Options/[Drop Output for Insert](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_change_capture_stage.html?view=kc) | True/False | False | N | N | N/A |
| Options/[Drop Output for Delete](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_change_capture_stage.html?view=kc) | True/False | False | N | N | N/A |
| Options/[Drop Output for Edit](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_change_capture_stage.html?view=kc) | True/False | False | N | N | N/A |
| Options/[Drop Output for Copy](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_change_capture_stage.html?view=kc) | True/False | True | N | N | N/A |
| Options/[Code Column Name](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_change_capture_stage.html?view=kc) | string | N/A | N | N | N/A |
| Options/[Copy Code](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_change_capture_stage.html?view=kc) | number | 0 | N | N | N/A |
| Options/[Deleted Code](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_change_capture_stage.html?view=kc) | number | 2 | N | N | N/A |
| Options/[Edit Code](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_change_capture_stage.html?view=kc) | number | 3 | N | N | N/A |
| Options/[Insert Code](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Options_Category_change_capture_stage.html?view=kc) | number | 1 | N | N | N/A |
| *Table 1. Properties* | | | | | |

Specifies the name of a difference key input column.

**Key**

This property can be repeated to specify multiple difference key input columns. You can use the Column Selection dialog box to select several keys at once if required. Key has the following dependent properties:

* **Case** **Sensitive**

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

* **Sort Order**

Specify ascending or descending sort order.

* **Nulls Position**

Specify whether null values should be placed first or last.

**Value**

Specifies the name of a value input column. You can use the Column Selection dialog box to select values at once if required . Value has the following dependent properties:

* **Case** **Sensitive**

Use this to property to specify whether each value is case sensitive or not. It is set to True by default; for example, the values "CASE" and "case" would not be judged equivalent.

Use the Options category to specify how the Change Capture stage operates.

## Change mode

This mode determines how keys and values are specified. Choose Explicit Keys & Values to specify the keys and values yourself. Choose All keys, Explicit values to specify that value columns must be defined, but all other columns are key columns unless excluded. Choose Explicit Keys, All Values to specify that key columns must be defined but all other columns are value columns unless they are excluded.

## Log statistics

This property configures the stage to display result information containing the number of input rows and the number of copy, delete, edit, and insert rows.

## Drop output for insert

Specifies to drop (not generate) an output row for an insert result. By default, an output row is always created by the stage.

## Drop output for delete

Specifies to drop (not generate) the output row for a delete result. By default, an output row is always created by the stage.

## Drop output for edit

Specifies to drop (not generate) the output row for an edit result. By default, an output row is always created by the stage.

## Drop output for copy

Specifies to drop (not generate) the output row for a copy result. By default, an output row is not created by the stage.

## Code column name

Allows you to specify a different name for the output column carrying the change code generated for each record by the stage. By default the column is called change\_code.

## Copy code

Allows you to specify an alternative value for the code that indicates the after record is a copy of the before record. By default this code is 0.

## Deleted code

Allows you to specify an alternative value for the code that indicates that a record in the before set has been deleted from the after set. By default this code is 2.

## Edit code

Allows you to specify an alternative value for the code that indicates the after record is an edited version of the before record. By default this code is 3.

## Insert Code

Allows you to specify an alternative value for the code that indicates a new record has been inserted in the after set that did not exist in the before set. By default this code is 1.

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 pools 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 input link carries the *before* data set and which carries the *after* data set.

By default the first link added will represent the *before* set. To rearrange the links, choose an input link and click the up arrow button or the down arrow button.

For the Change Capture 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 Change Capture stage uses this when it is determining the sort order for key columns. 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 incoming data sets. The Change Capture expects two incoming data sets: a before data set and an after data set.

The Input page allows you to specify details about the incoming data sets. The Change Capture expects two incoming data sets: a before data set and an after data set.

The General tab allows you to specify an optional description of the input link. The Partitioning tab allows you to specify how incoming data is partitioned before being compared. 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 Change Capture 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.

* [**Change Capture stage: Partitioning tab**](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/r_deeref_Partitioning_Tab_change_capture_stage.html?view=kc)

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

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

It also allows you to specify that the data should be sorted before being operated on.

By default the stage partitions in Auto mode. This 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. In the case of the Change Capture stage, InfoSphere® DataStage® will determine if the incoming data is key partitioned. If it is, the Same method is used, if not, InfoSphere DataStage will hash partition the data and sort it. You could also explicitly choose hash and take advantage of the on-stage sorting.

If the Change Capture stage is operating in sequential mode, it will first collect the data 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 Change Capture 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 Change Capture 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 Change Capture 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 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 partitioning method for the Change Capture 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 Change Capture stages. For the Change Capture stage, InfoSphere DataStage will ensure that the data is sorted as it is collected.
* **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 being compared. 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 Change Capture stage. The tabs in this stage allow you to specify an optional description, the relationship between the columns being input and the Output columns, and change the default buffer settings for the output link. You can also view the column definitions.

The Output page allows you to specify details about data output from the Change Capture stage. The Change Capture stage can have only 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 Change Capture stage and the Output columns. The Advanced tab allows you to change the default buffering settings for the output link.

Details about Change Capture 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.

* [**Change Capture stage: 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_change_capture_stage.html?view=kc)

For the Change Capture stage, the Mapping tab allows you to specify how the output columns are derived, that is, what input columns map onto them and which column carries the change code data.

For the Change Capture stage, the Mapping tab allows you to specify how the output columns are derived, that is, what input columns map onto them and which column carries the change code data.

The left pane shows the columns from the before/after data sets plus the change code column. These are read only and cannot be modified on this tab.

The right pane shows the output columns for each 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. By default the data set columns are mapped automatically. You need to ensure that there is an output column to carry the change code and that this is mapped to the Change\_code column.