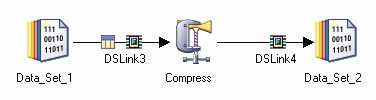
The Compress stage is a processing stage. This stage uses the UNIX compress or GZIP utility to compress a data set.

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

The Compress stage uses the UNIX compress or GZIP utility to compress a data set. It converts a data set from a sequence of records into a stream of raw binary data. The complement to the Compress stage is the Expand stage, which is described in [Expand Stage](https://www.ibm.com/support/knowledgecenter/SSZJPZ_11.7.0/com.ibm.swg.im.iis.ds.parjob.dev.doc/topics/c_deeref_Expand_Stage_expand_stage.html?view=kc).

A compressed data set is similar to an ordinary data set and can be stored in a persistent form by a Data Set stage. However, a compressed data set cannot be processed by many stages until it is expanded, that is, until its rows are returned to their normal format. Stages that do not perform column-based processing or reorder the rows can operate on compressed data sets. For example, you can use the Copy stage to create a copy of the compressed data set.

Because compressing a data set removes its normal record boundaries, the compressed data set must not be repartitioned before it is expanded.



DataStage® puts the existing data set schema as a subrecord to a generic compressed schema. For example, given a data set with a schema of:

a:int32;

b:string[50];[Copy](javascript:void(0);)

The schema for the compressed data set would be:

record

( t: tagged {preservePartitioning=no}

( encoded: subrec

( bufferNumber: dfloat;

bufferLength: int32;

bufferData: raw[32000];

);

schema: subrec

( a: int32;

b: string[50];

);[Copy](javascript:void(0);)

Therefore, when you are looking to reuse a file that has been compressed, ensure that you use the 'compressed schema' to read the file rather than the schema that had gone into the compression.

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_compress_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_compress_stage.html?view=kc). This is where you specify details about the data set being compressed.
* [**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_compress_stage.html?view=kc). This is where you specify details about the compressed data being output from the stage.

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

**About this task**

InfoSphere® DataStage® has many defaults which means that it can be very easy to include Compress 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. 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_compress_stage.html?view=kc) choose the compress command to use. Compress is the default but you can also choose gzip.
2. Ensure column meta data is defined for both the input and output link.

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 Compress stage operates.

The Properties tab allows you to specify properties which determine what the stage actually does. The stage only has a single property which determines whether the stage uses compress or GZIP.

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

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

## Command

Specifies whether the stage will use compress (the default) or GZIP.

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 **Set** by default. You can explicitly select **Set** or **Clear**. Select **Set** to request the next 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 compressed. There is only one input link.

The Input page allows you to specify details about the data set being compressed. 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 Compress 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 compress is performed.

By default the stage uses the auto partitioning method.

If the Compress 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 Compress 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 Compress 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 Compress 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 Compress 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 Compress 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 compression 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 Outpage page, you can specify details about data output from the Compress stage.

The Output page allows you to specify details about data output from the Compress 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 Advanced tab allows you to change the default buffering settings for the output link.

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 tabs.