TATA CONSULTANCY SERVICES
Experience certainty.

Training Material –Datastage 8.1

November 28, 2012

Datastage

Domain/ Service/ Technology /Geo:Fn:Datastage 8.1

Name of the Project/Support Fn:Datastage 8.1

Name of the Author: Minakshee Patil, Praveen Nair

Date Created:07-07-2012

Agenda

- Overview of Datastage stages 8.1

Datastage Introduction

- IBM's ETL tool
- Extracts, Transforms and Loads data
- Used in Data migration and Data warehousing projects

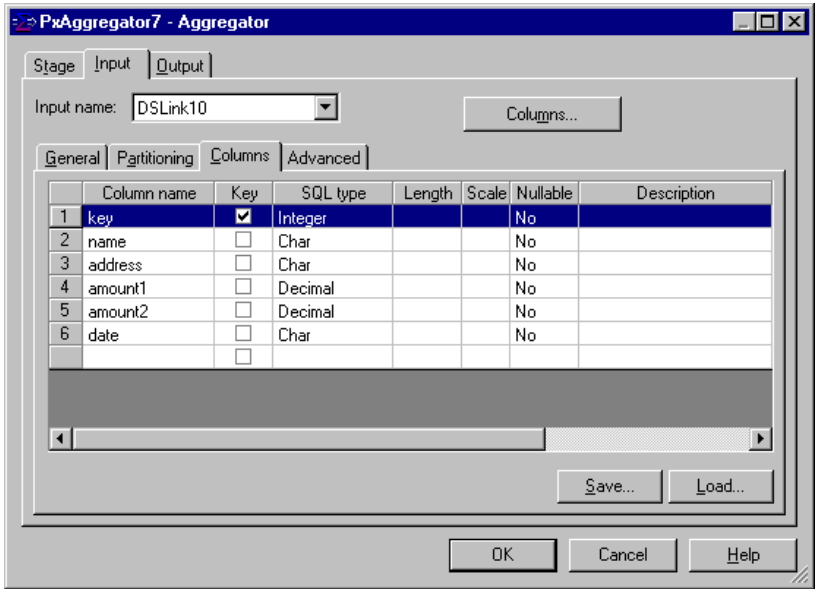
Aggregator Stage :

Definition : Aggregator classifies data rows from a single input link into groups and calculates totals or other aggregate functions for each group. The summed totals for each group are output from the stage thro' output link.

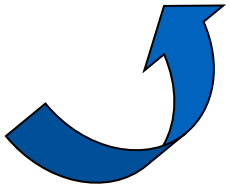
Group is a set of record with the same value for one or more columns.

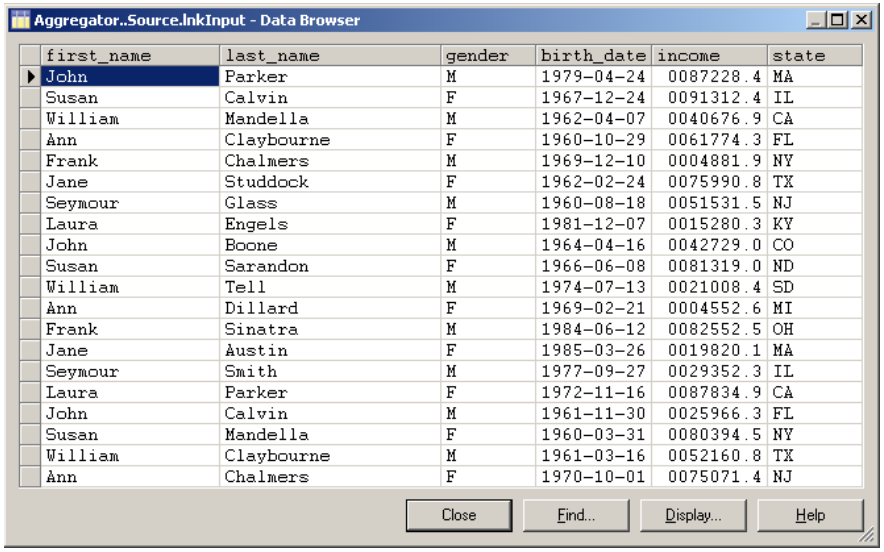
Example : Transaction records might be grouped by both day of the week and by month. These groupings might show the busiest day of the week varies by season.

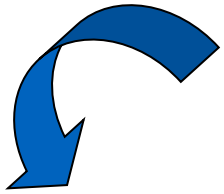
Input & View data :



The **INPUT** page shows you the metadata of the incoming data.



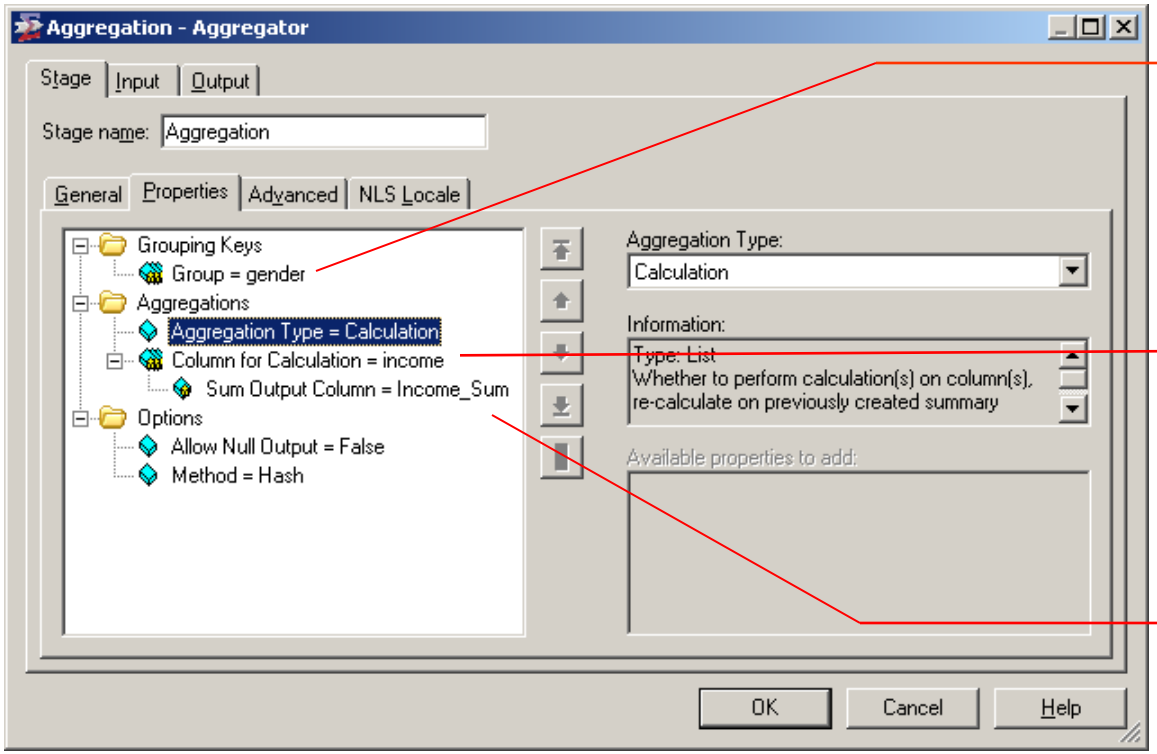




The input data look like this...

Properties :

When "Aggregation Type = Calculation" ...

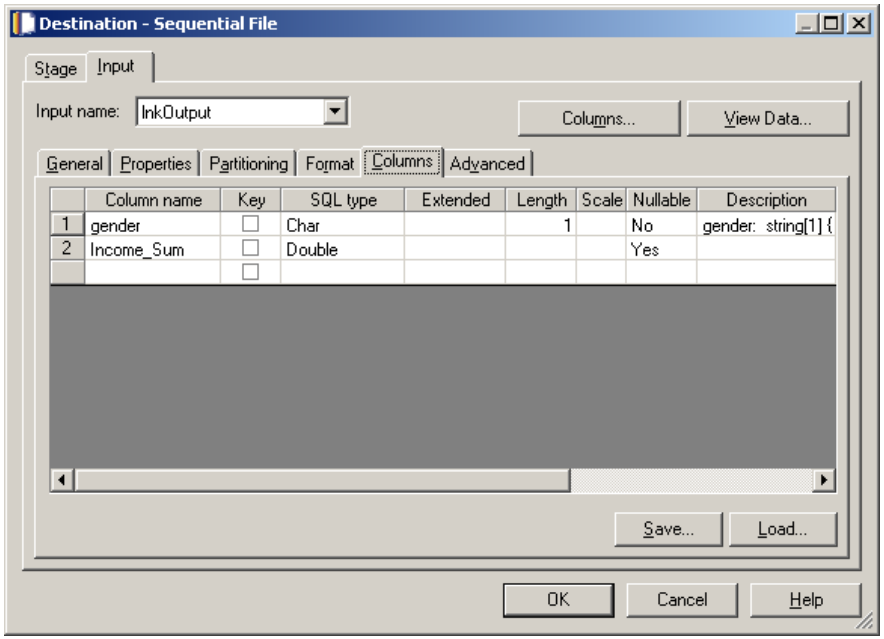


Here, we group by **"Gender"**.

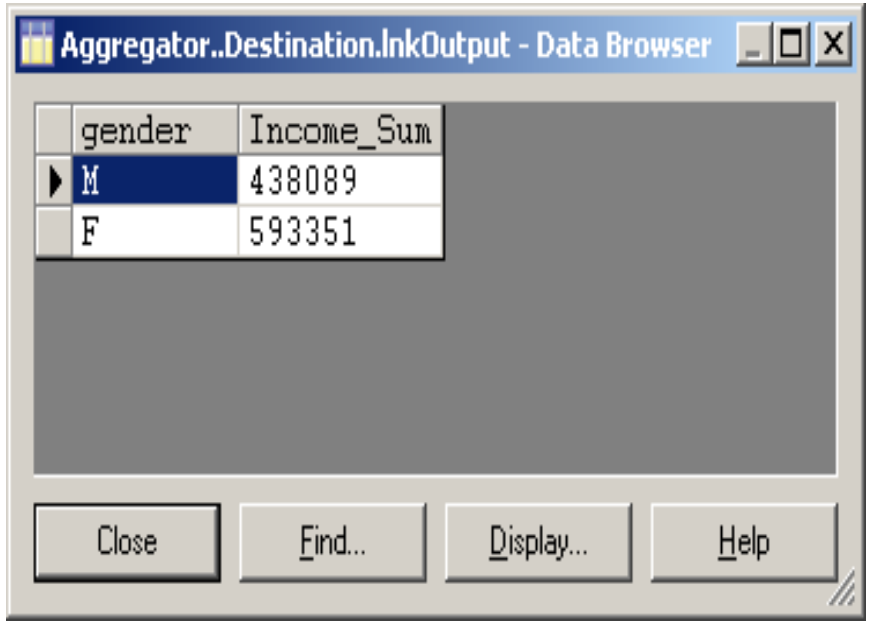
The column to aggregate.

User defined column to collect the aggregated values.

Output & View data :

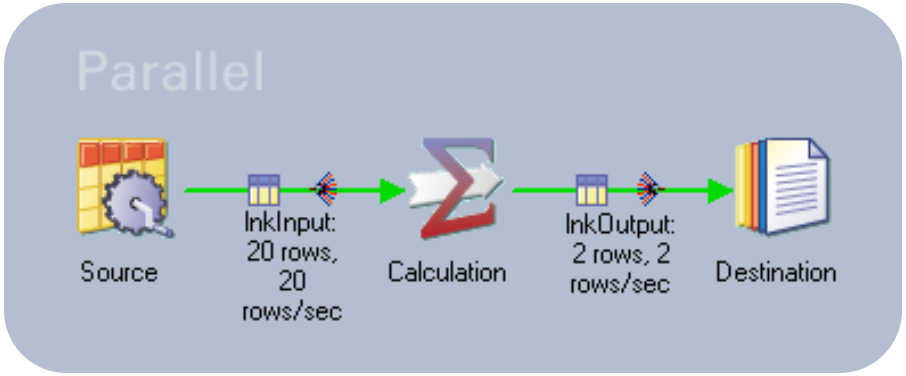


The **OUTPUT** page shows only those columns used to group and aggregate.



As we have grouped by "**Gender",** the incomes of Males and Females are summed up and shown here.

Execution Mode :

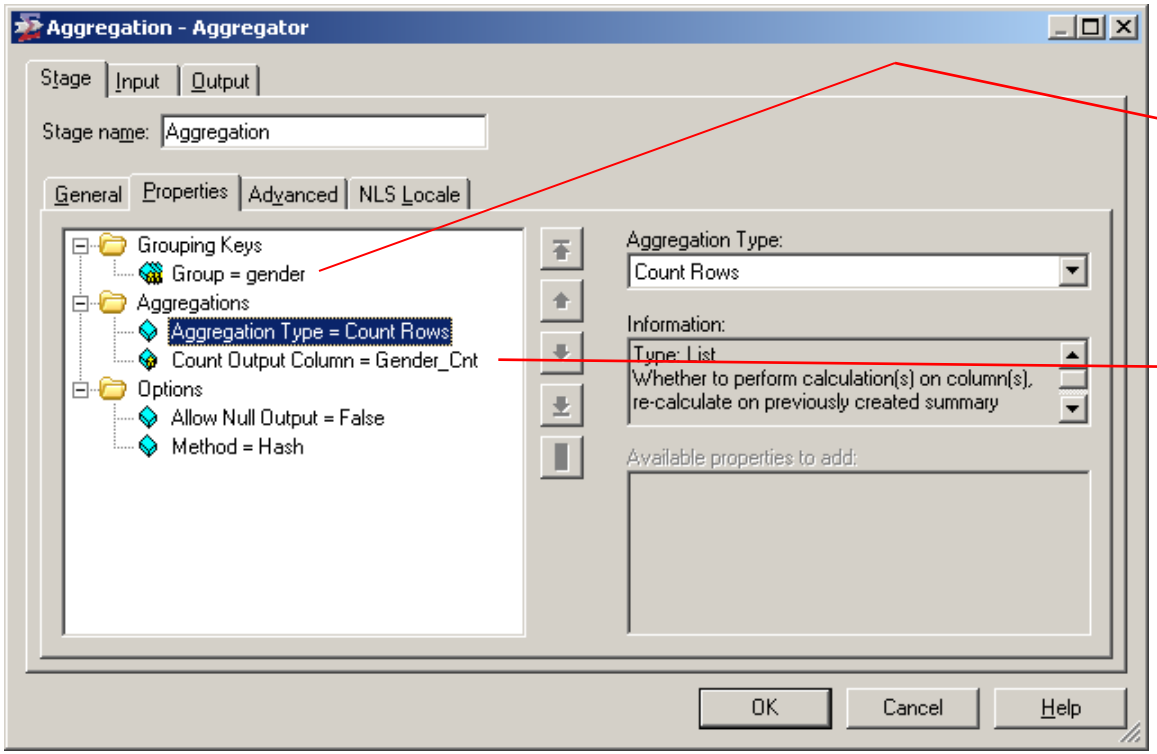


Note :

The Aggregator stage can have only one output link.

Properties :

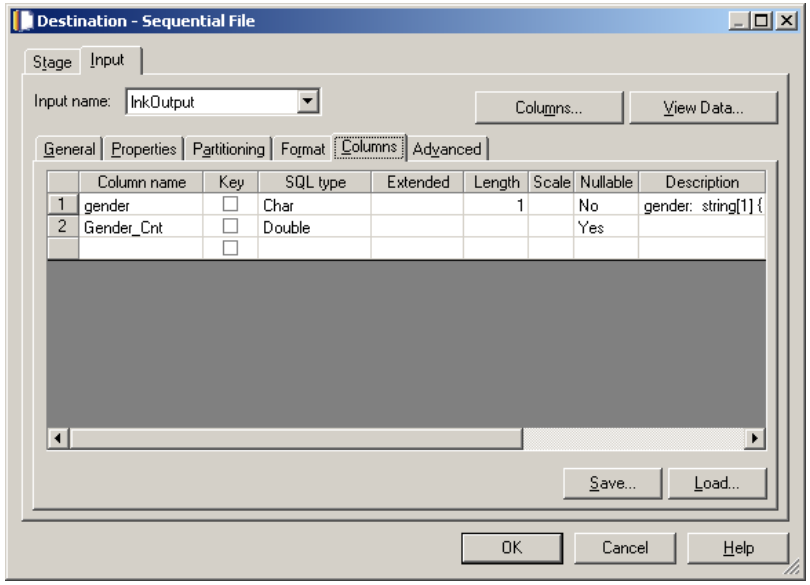
When "Aggregation Type = Count Rows" ...



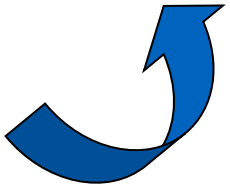
Here, we group by **"Gender"**.

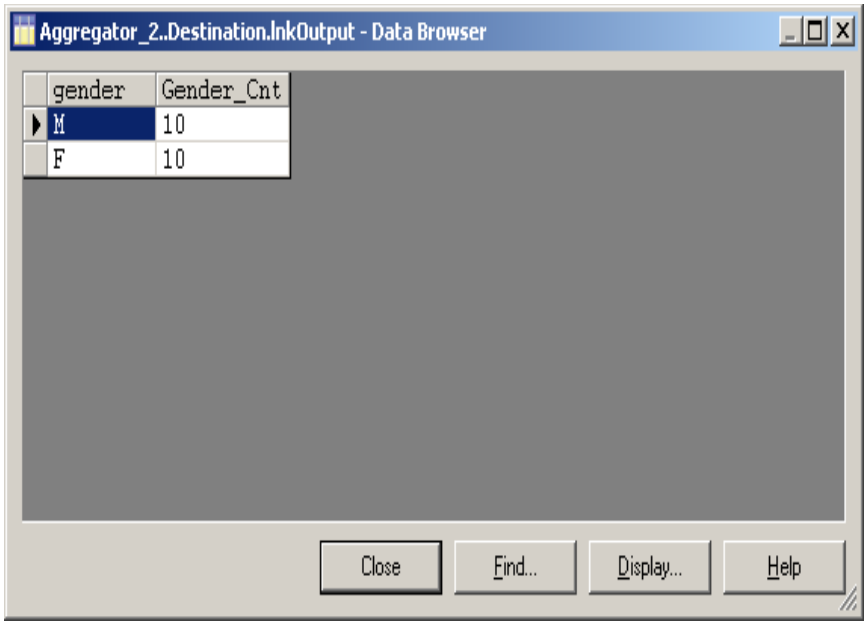
The column to be counted.

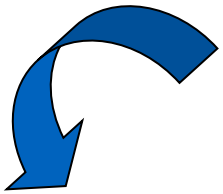
Output & View data :



The **OUTPUT** page shows only the grouping column and the column to be counted.

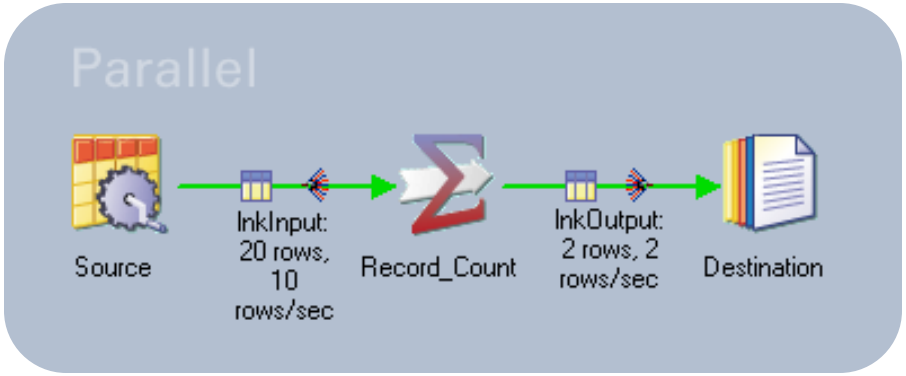






As we have grouped by **"Gender",** the number of records in Males and Females are totaled and shown here.

Execution Mode :

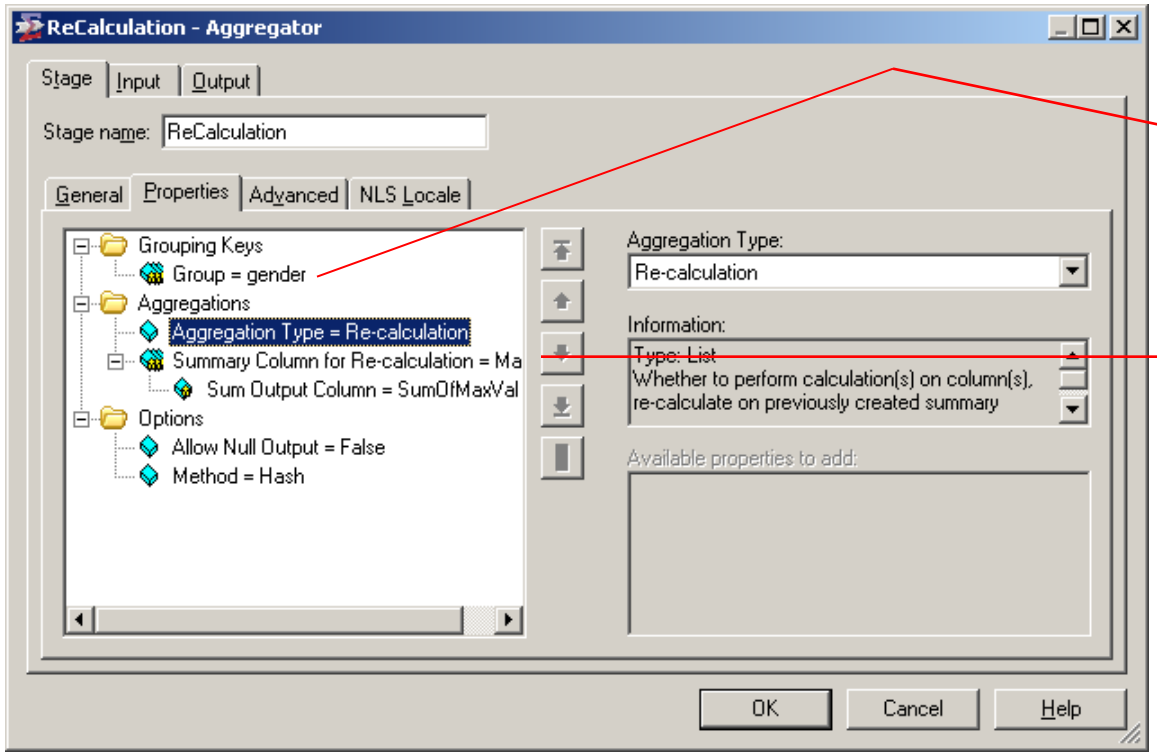


Note :

The Aggregator stage can have only one output link.

Properties :

When "Aggregation Type = Re-calculation" ...



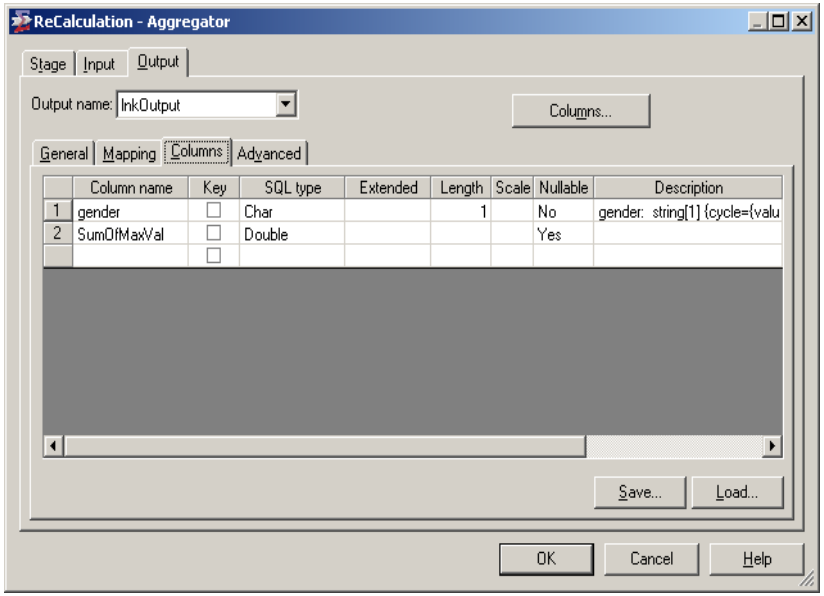
Here, we group by **"Gender"**.

The column to preserver the summary of Recalculation.

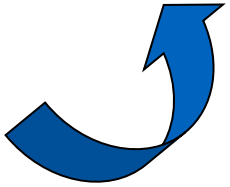
Note :

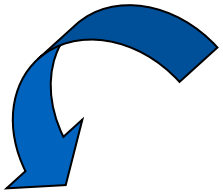
When the "Aggregation Type = Re-calculation" then place an extra aggregator to aggregate a column, first. The second aggregator will re-calculate the previously calculated column.

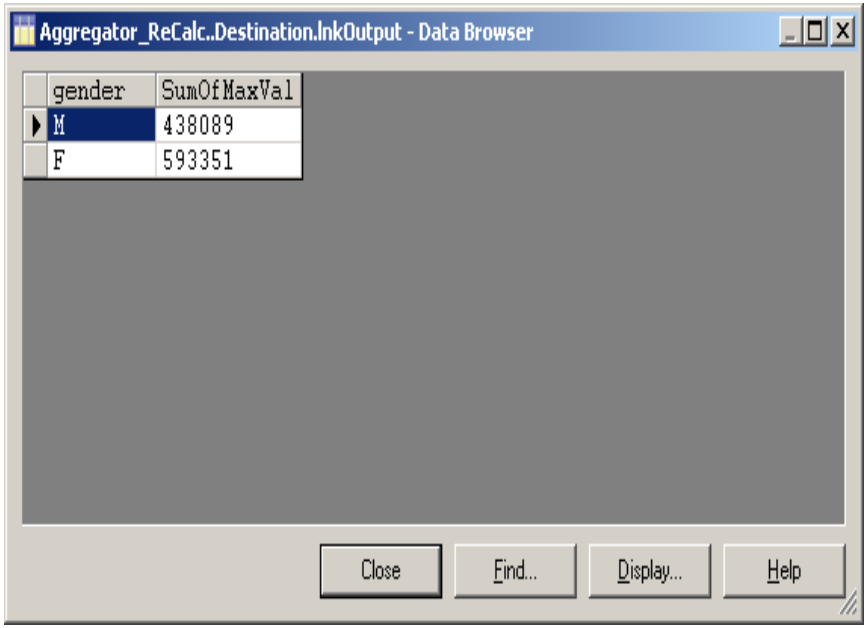
Output & View data :



The **OUTPUT** page shows only the grouping column and the column for recalculation.

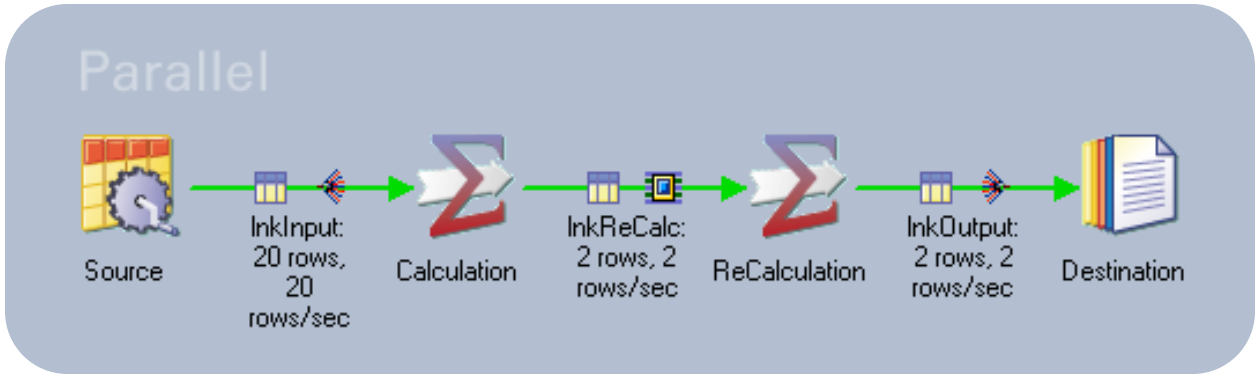






The column "MaxVal" is recalculated as "SumOfMaxVal"

Execution Mode :

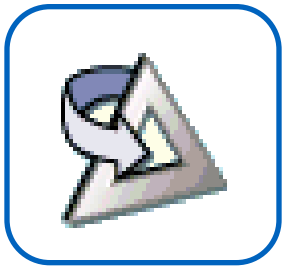


Note :

The Aggregator stage can have only one output link.

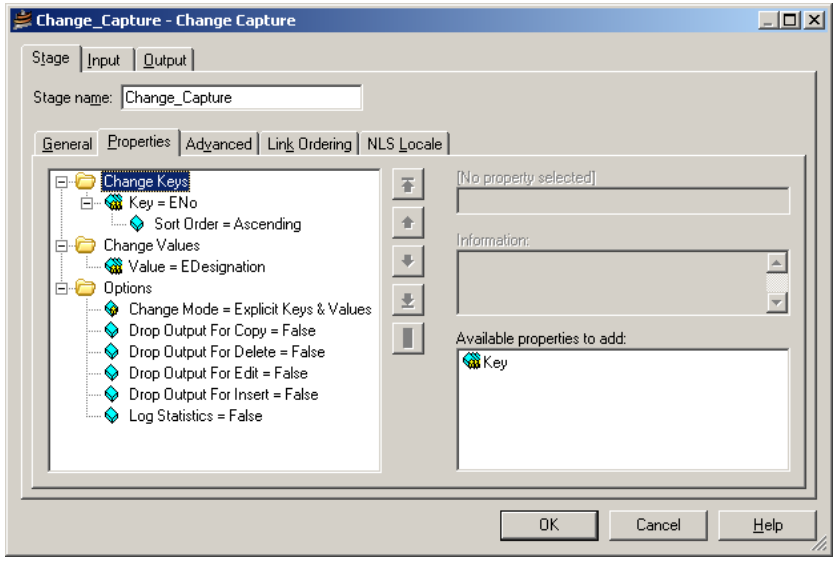
Change Apply Stage :

Definition : Takes the change data set, that contains the changes in the before and after data sets, from the Change Capture stage and applies the encoded change operations to a before data set to compute an after data set.

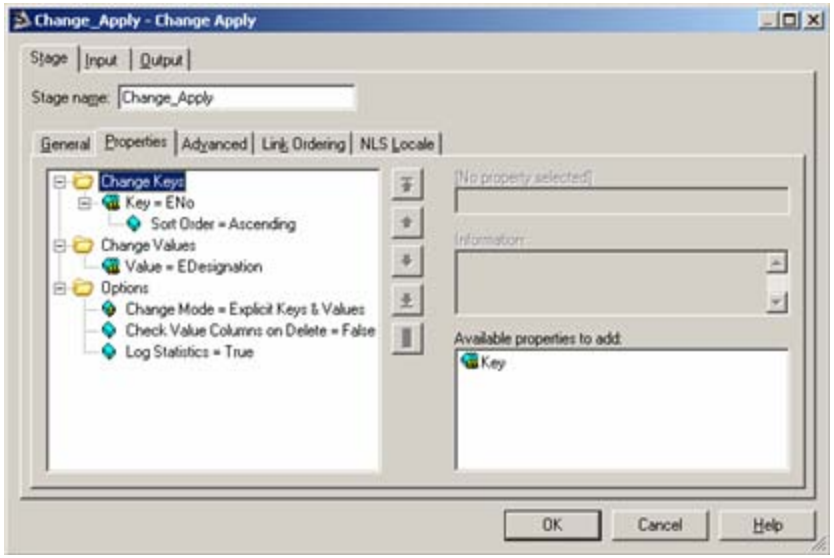


The Change Apply stage read a record from the change data set and from the before data set, compares their key column values, and acts accordingly.

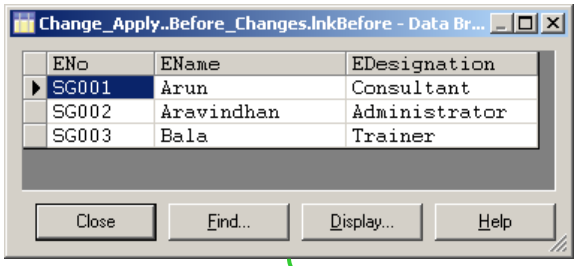
Change Capture Property :



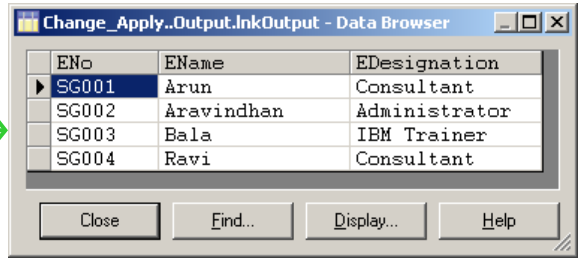
Change Apply Property :

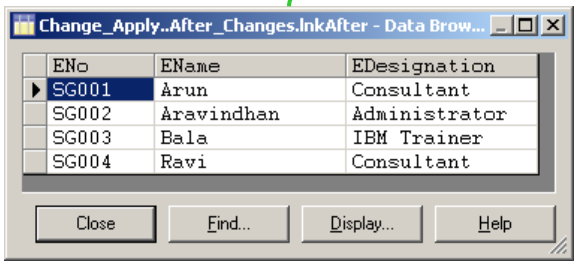


Input (Before Changes) :



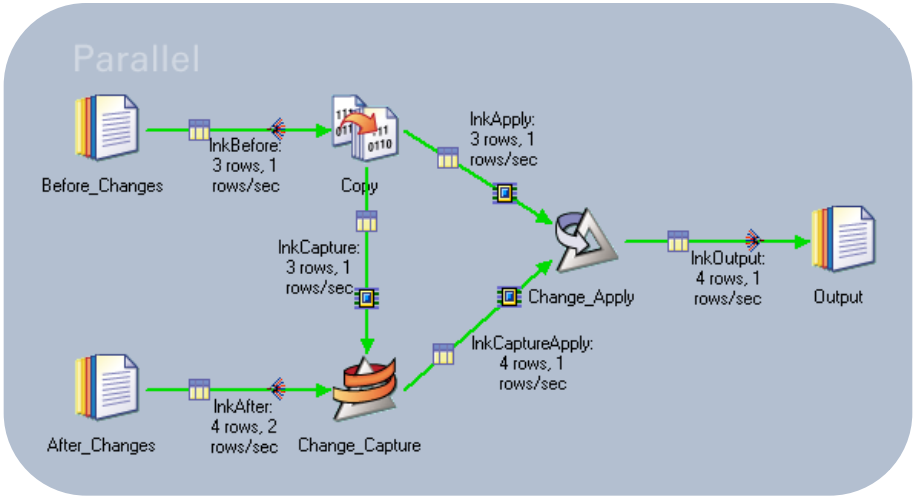
Output :





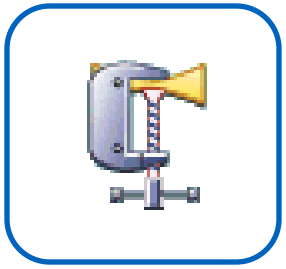
Input (After Changes) :

Job :

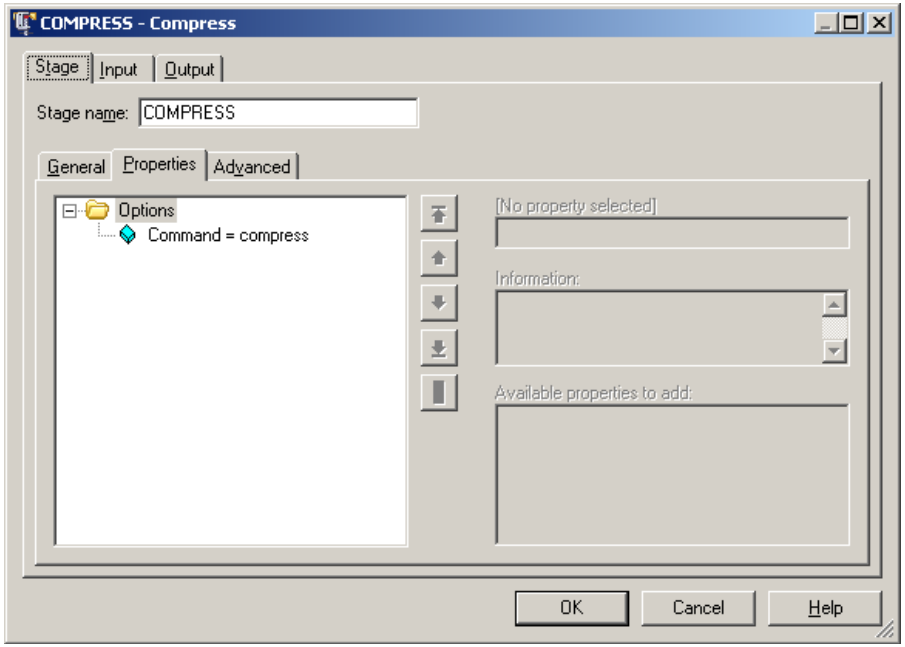

Parallel
Before_Changes
lnkBefore:
3 rows, 1
rows/sec
Copy
lnkApply:
3 rows, 1
rows/sec
lnkCapture:
3 rows, 1
rows/sec
Change_Apply
lnkOutput:
4 rows, 1
rows/sec
Output
lnkCaptureApply:
4 rows, 1
rows/sec
After_Changes
lnkAfter:
4 rows, 2
rows/sec
Change_Capture

Compress Stage :

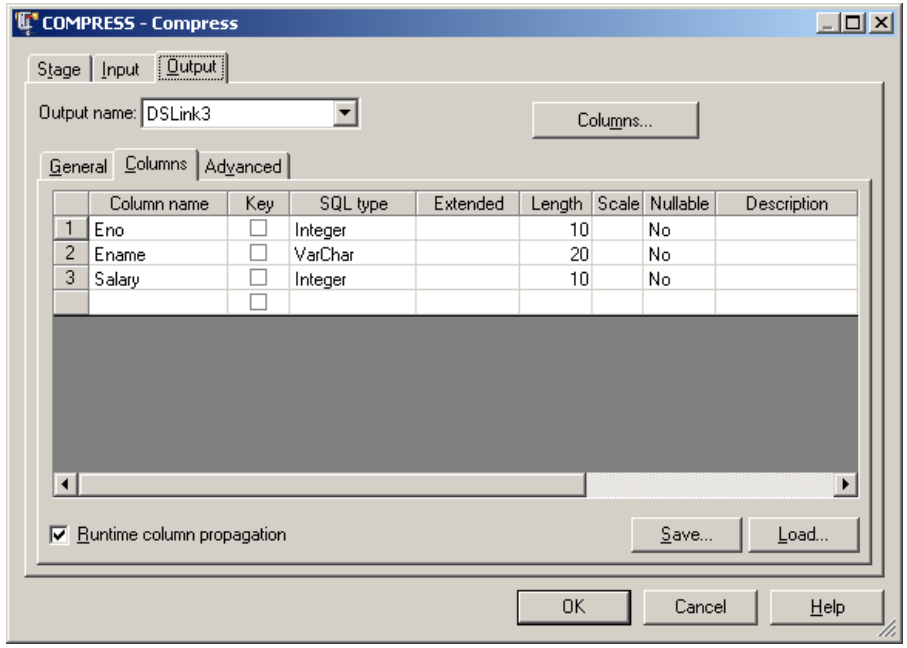
The Compress stage uses the UNIX compress or GZIP utility to compress a data set. It converts a data set from a sequence of records into a stream of raw binary data.



Steps :

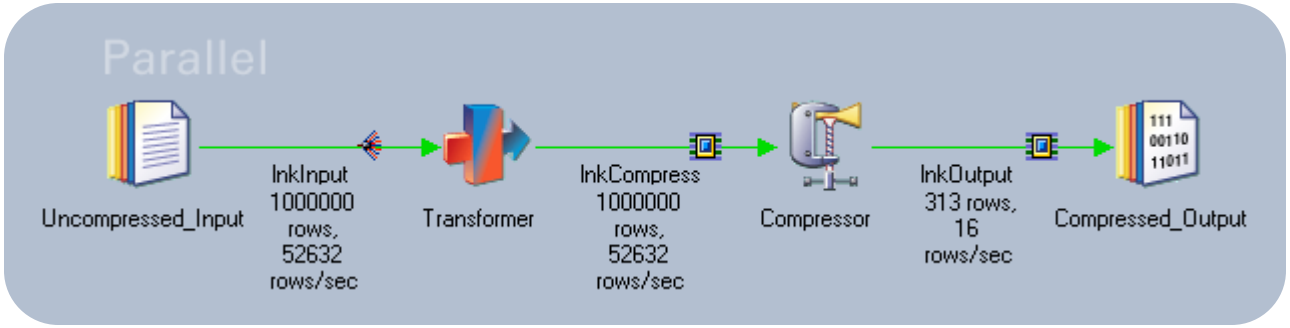


- Set the Stage Properties : "Command = Uncompress"



Load the Metadata in the Output tab...

Example :



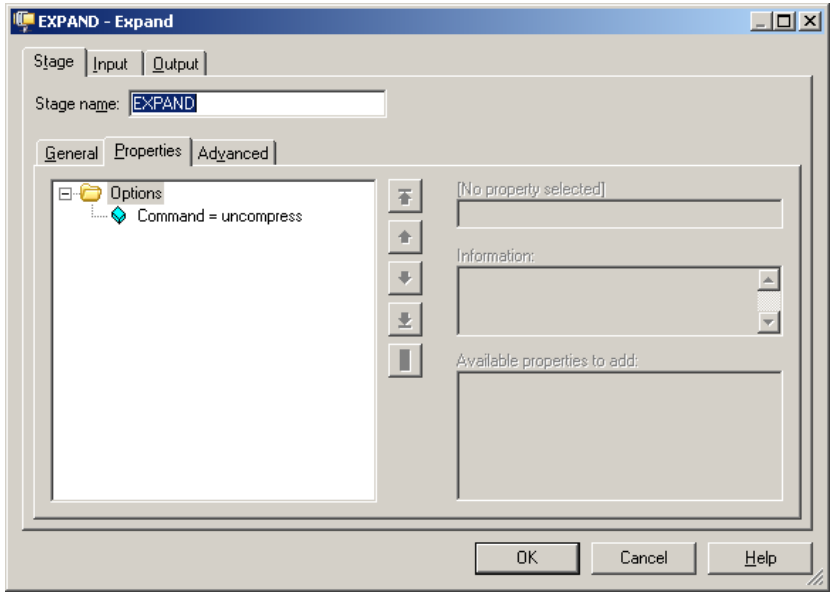
Limitations :

A compressed data set cannot be processed by many stages until it is expanded, i.e., 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.

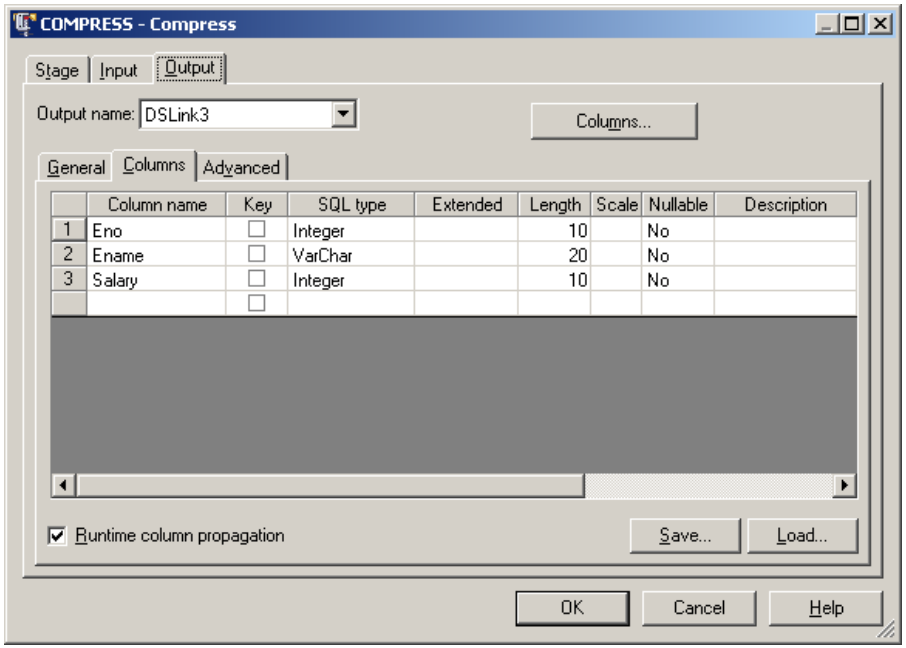
Expand Stage :

The Expand stage uses the UNIX compress or GZIP utility to expand the data set. It converts a data set from a stream of raw binary data into sequence of records.

Steps :

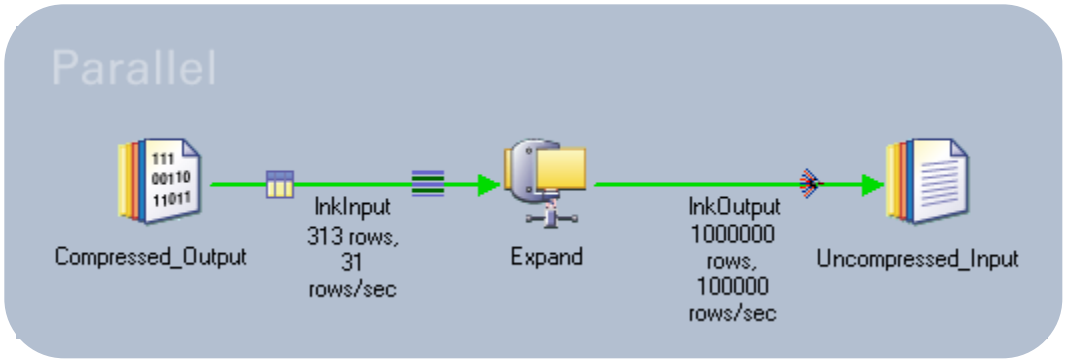


- Set the Stage Properties : "Command = Uncompress"



Load the Metadata in the Output tab...

Example Job :

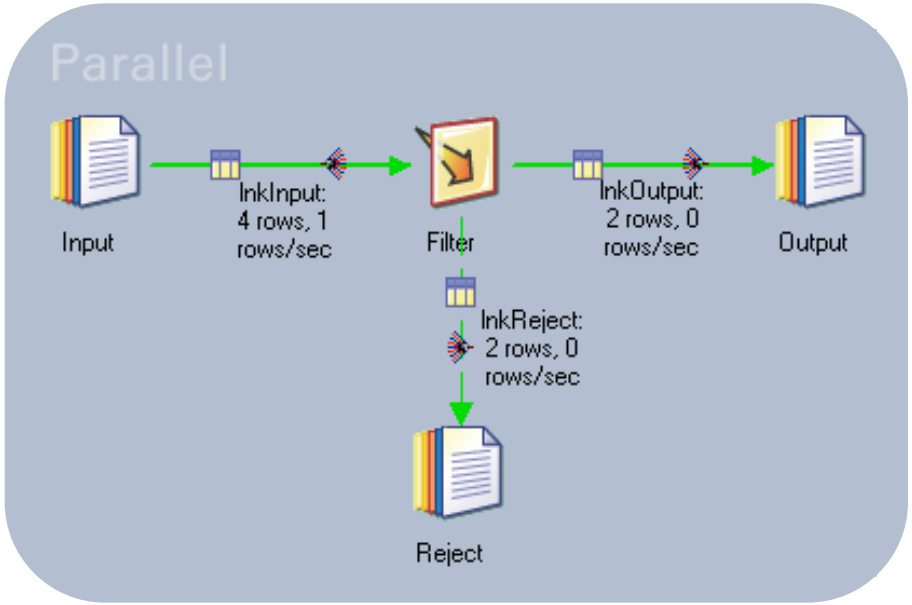

Parallel
111
00110
11011
Compressed_Output
lnkInput
313 rows,
31
rows/sec
Expand
lnkOutput
1000000
rows,
100000
rows/sec
Uncompressed_Input

Filter Stage :

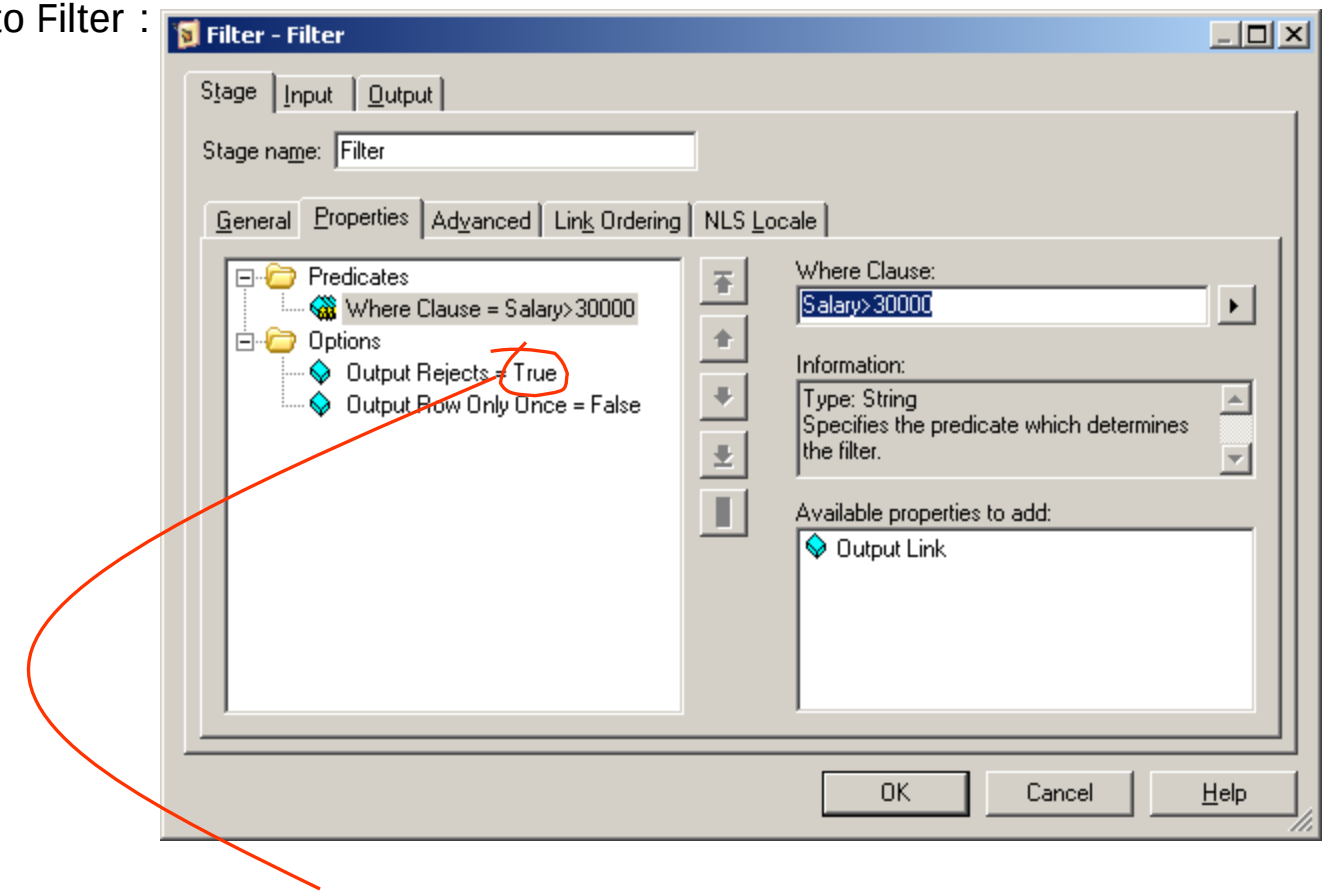
Definition : The Filter stage transfers, unmodified, the records of the input data set which satisfy the specified requirements and filters out all other records.

Filter stage can have a single input link and a any number of output links and, optionally, a single reject link. You can specify different requirements to route rows down different output links. The filtered out records can be routed to a reject link, if required.

Simple Job :


Parallel
Input
lnkInput:
4 rows, 1
rows/sec
Filter
lnkOutput:
2 rows, 0
rows/sec
Output
lnkReject:
2 rows, 0
rows/sec
Reject

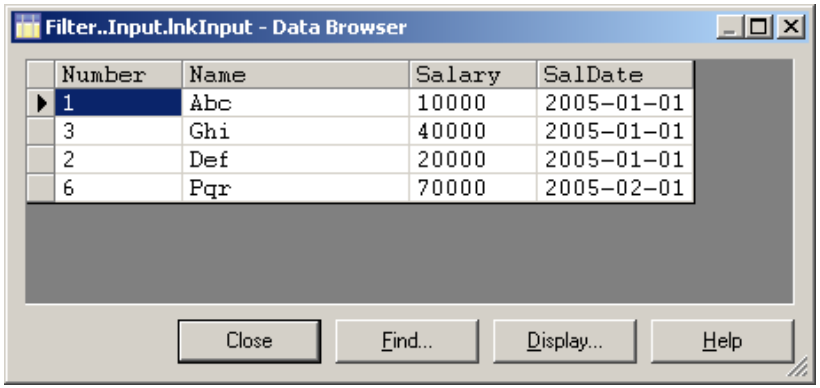
Criteria to Filter :



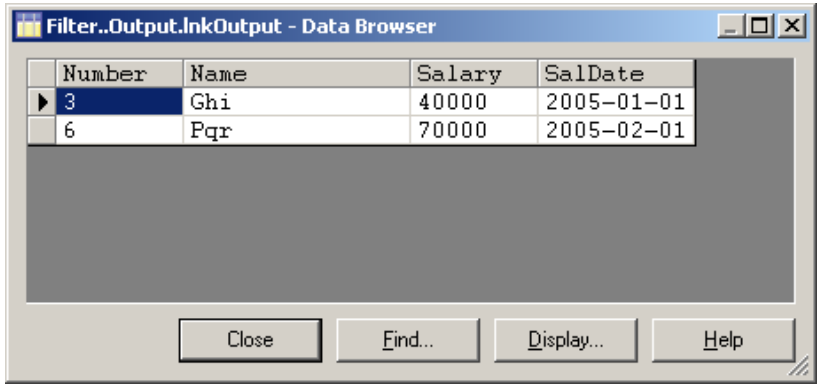
Note : Only if the "Output Rejects=True" the rejected rows are collected separately, otherwise those rows that fails to satisfy the criteria will be ignored.

Criteria : Salary>30000

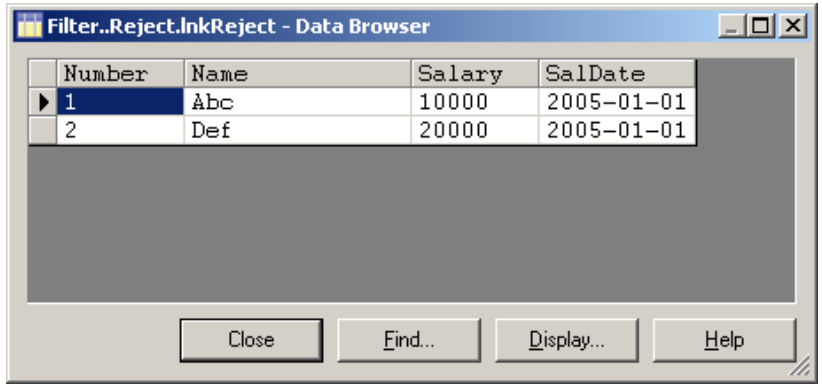
Input :



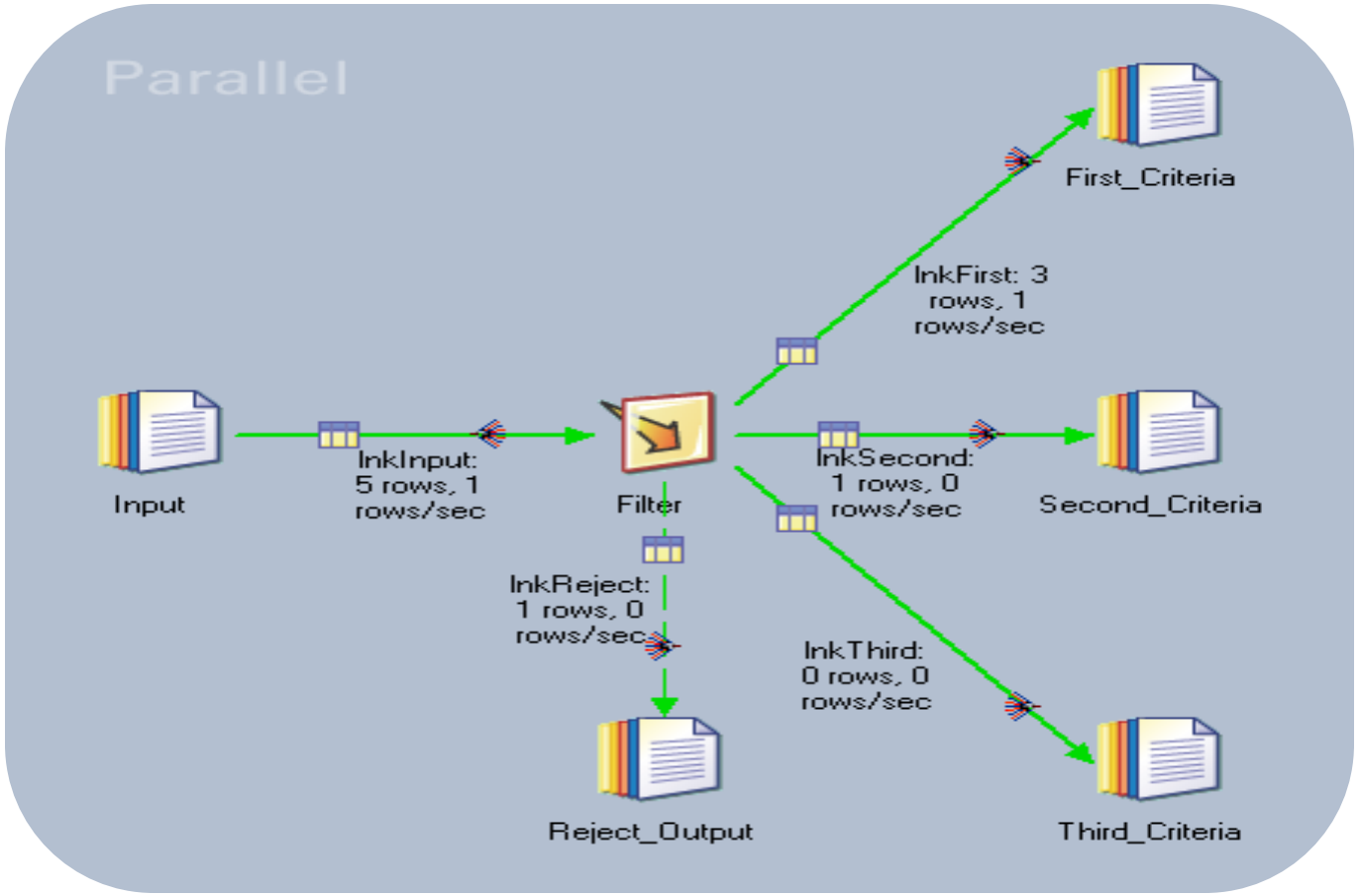
Output :



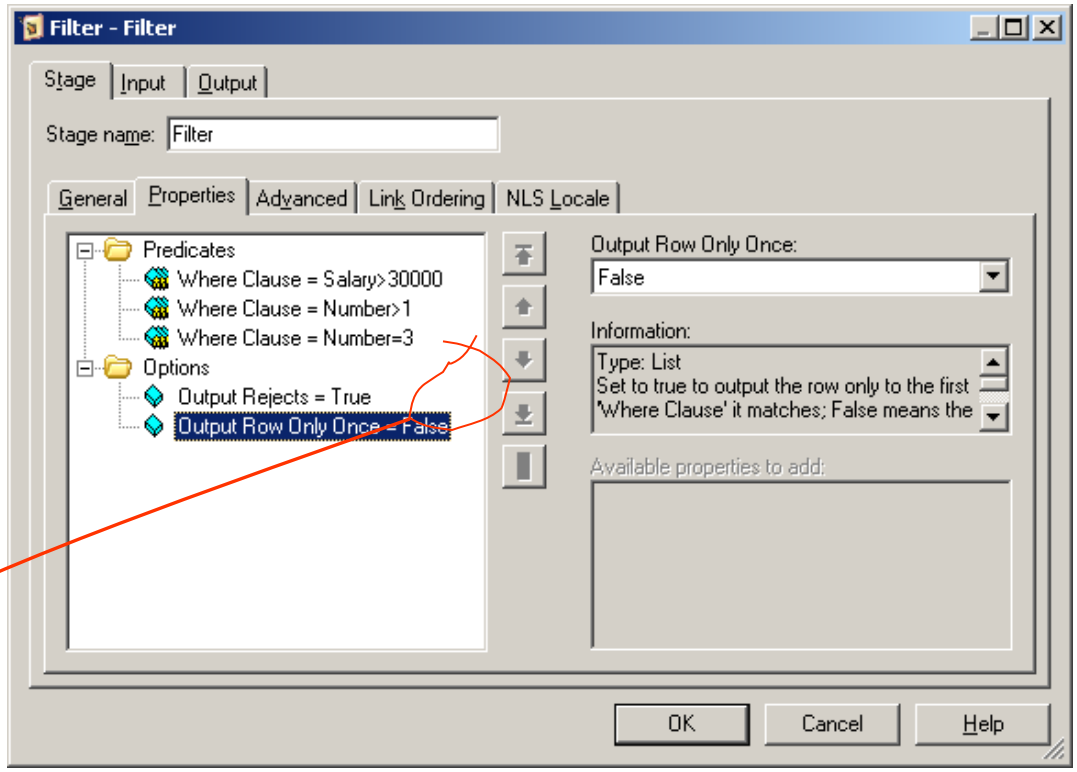
Reject Rows :



Complex Job :

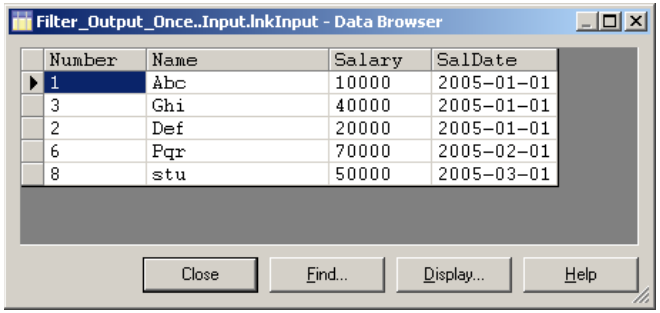

Parallel
First_Criteria
InkFirst: 3
rows, 1
rows/sec
Input
InkInput:
5 rows, 1
rows/sec
Filter
InkSecond:
1 rows, 0
rows/sec
Second_Criteria
InkReject:
1 rows, 0
rows/sec
InkThird:
0 rows, 0
rows/sec
Reject_Output
Third_Criteria

Criteria to Filter :



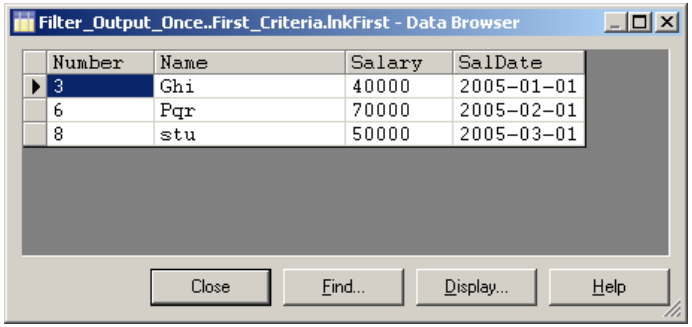
Note : "Output Row Only Once=False" means, every single input row is forced to satisfy each and every criteria given. In other words, Row that satisfies a criteria is forced to undergo another criteria. In such case, every row gets more than a single chance to output.

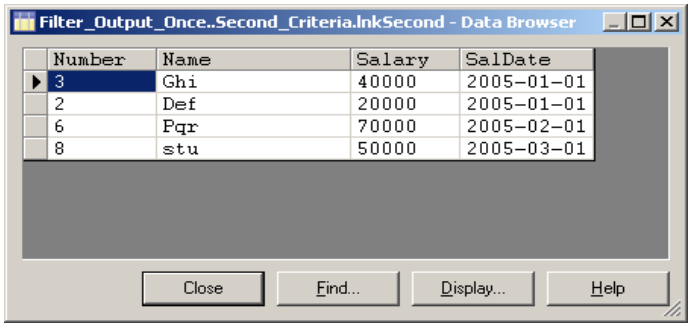
Input Data



Criteria 1 : Salary>30000

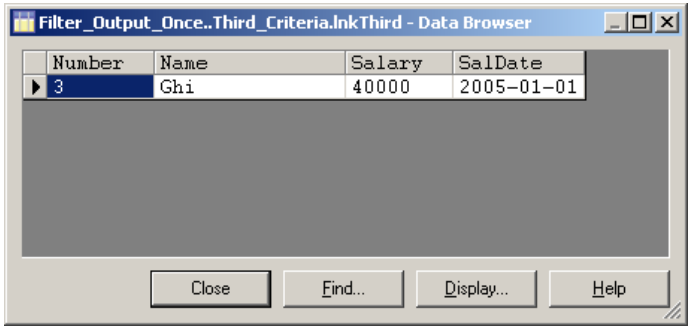
Criteria 2 : Number>1

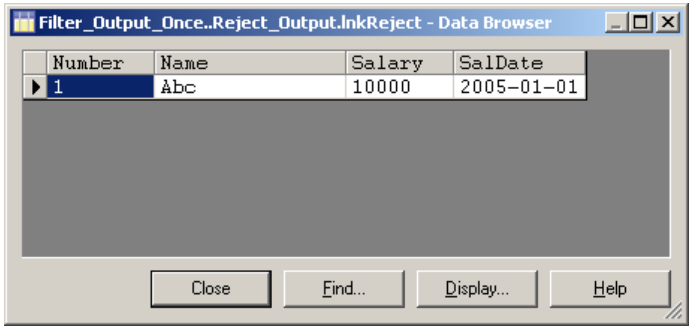




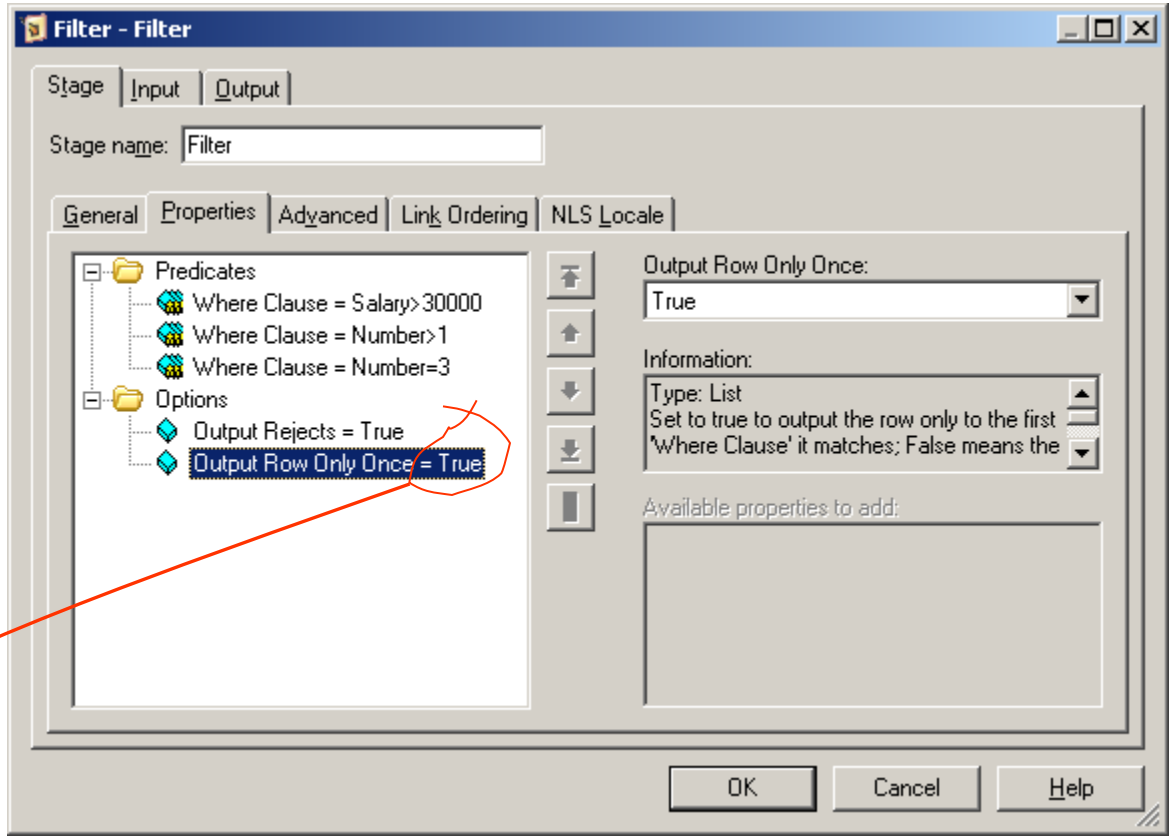
Criteria 3 : Number=3

Reject Rows :



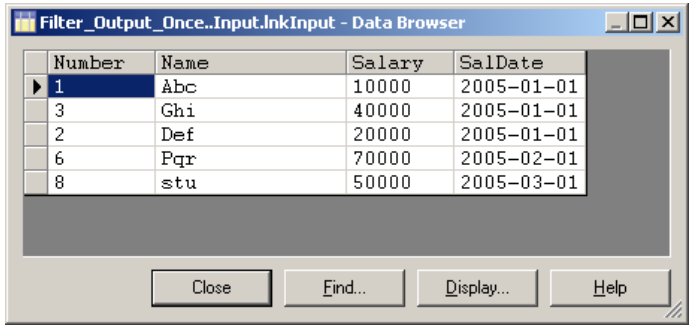


Criteria to Filter :

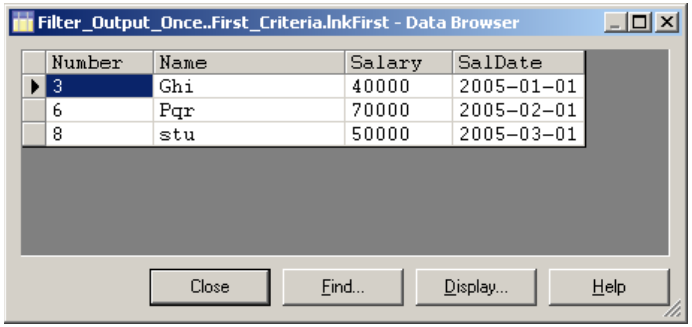


Note : "Output Row Only Once=True" means, every single input row is not forced to undergo each and every criteria given. In other words, Rows that satisfy at least one criteria is not forced to satisfy another criteria. In such case, every row gets a single chance to output.

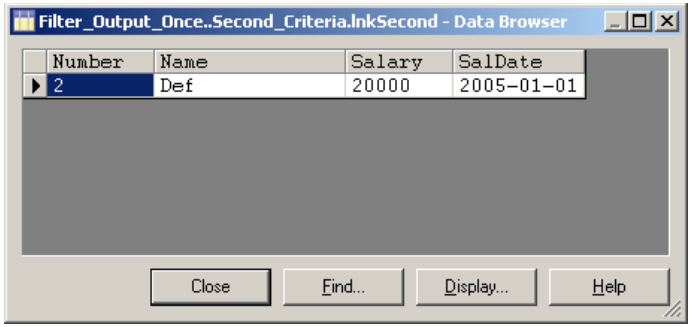
Input Data



Criteria 1 : Salary>30000

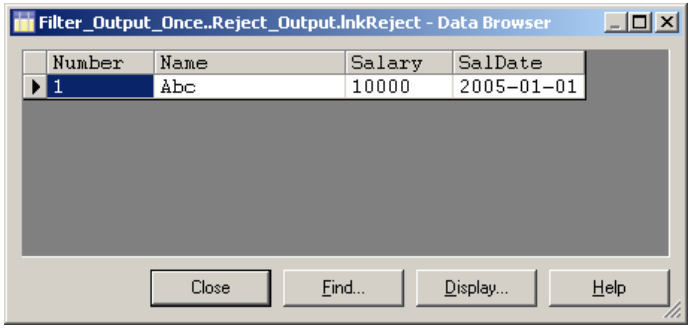


Criteria 2 : Number>1



Criteria 3 : Number=3

Reject Rows :

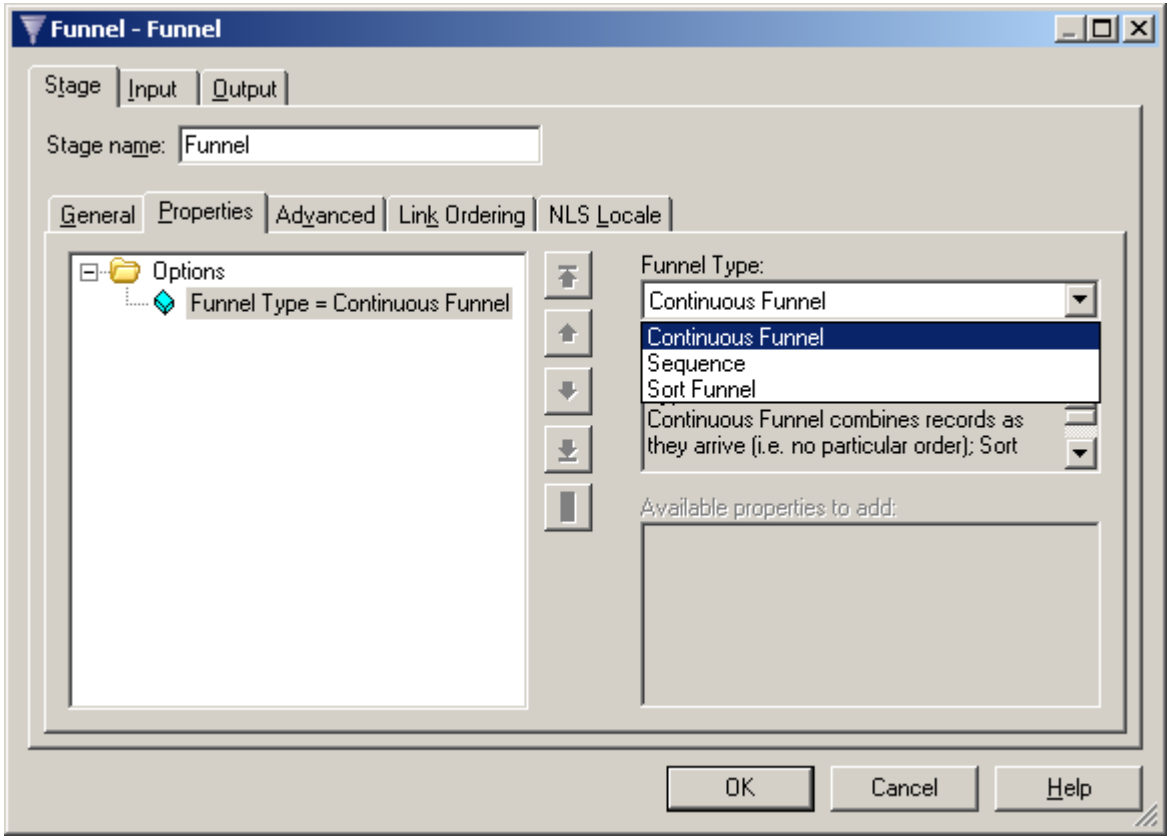


Note : Though there is a row that satisfies this criteria, it is not outputted as it is already been outputted for satisfying the Criteria – 1.

Funnel Stage :

Definition : Funnel Stage copies multiple input data sets to a single output data set. This operation is useful for combining separate data sets into a single large data set. The stage can have any number of input links and a single output link.

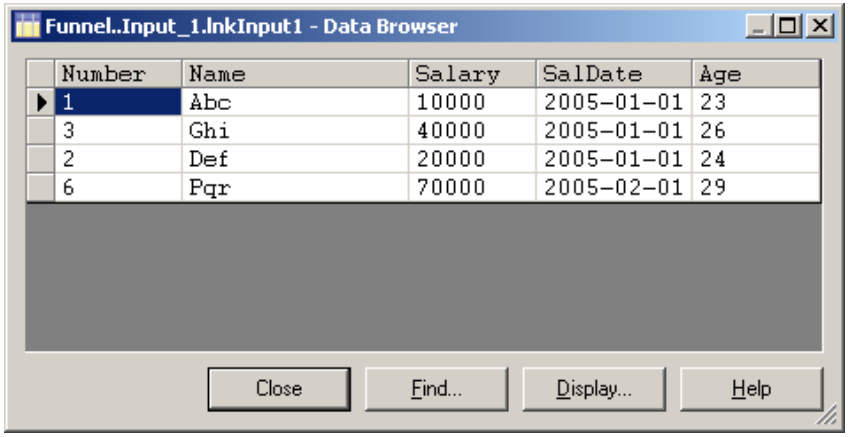
Type – 1 : Continuous Funnel



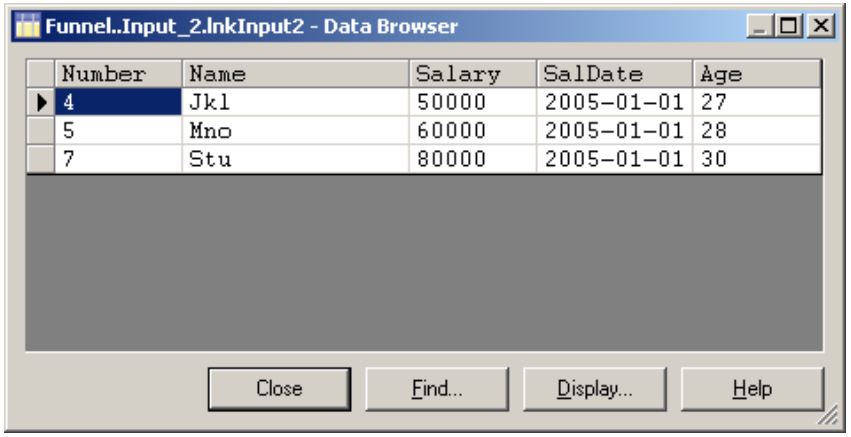
Note : This type of Funnel combines the records of the input data in no guaranteed order. It takes one record from each input link in turn. If data is not available on an input link, the stage skips to the next link rather than waiting.

Continuous Funnel type ...

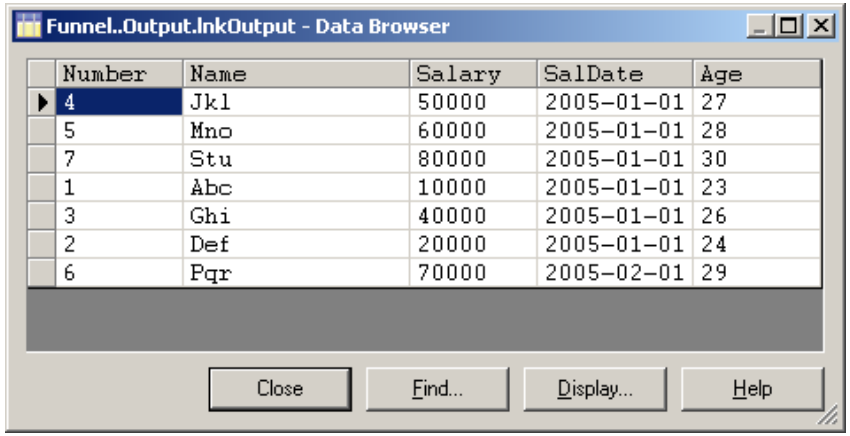
Input - 1 view data :



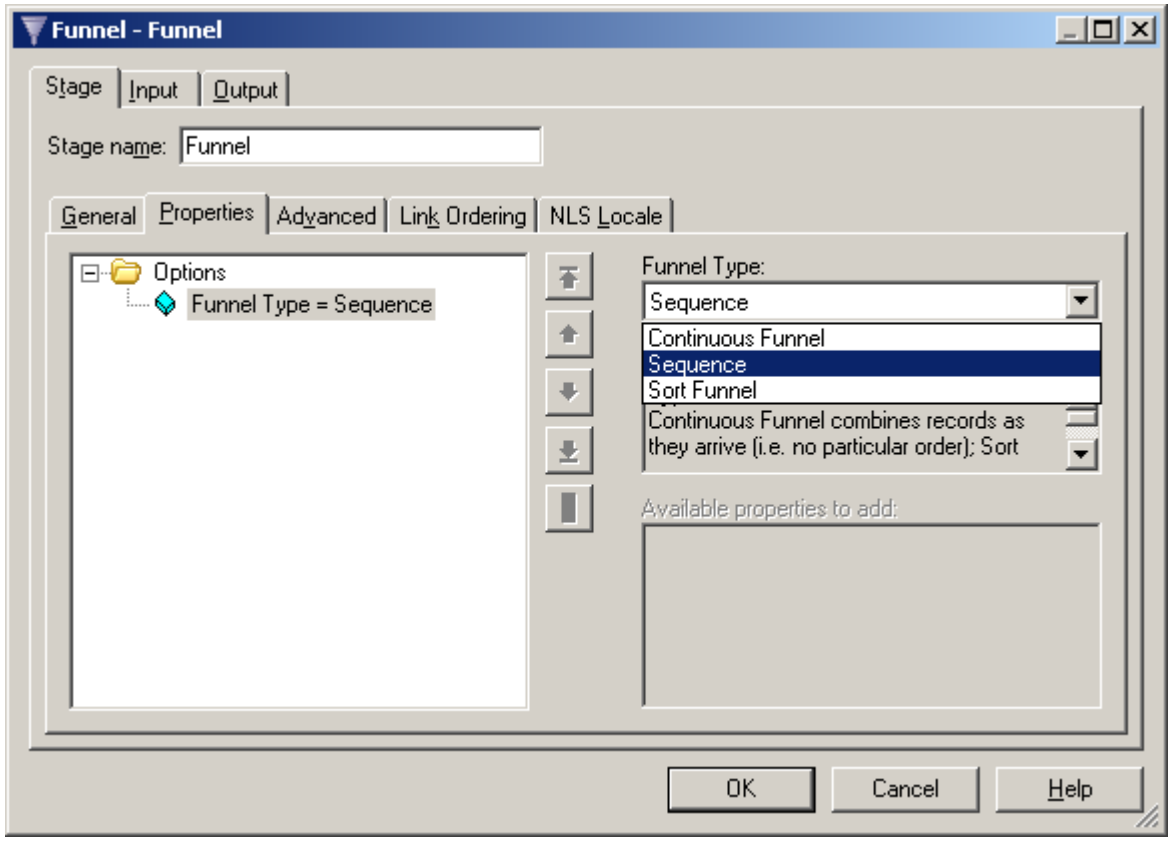
Input - 2 view data :



Output view data :



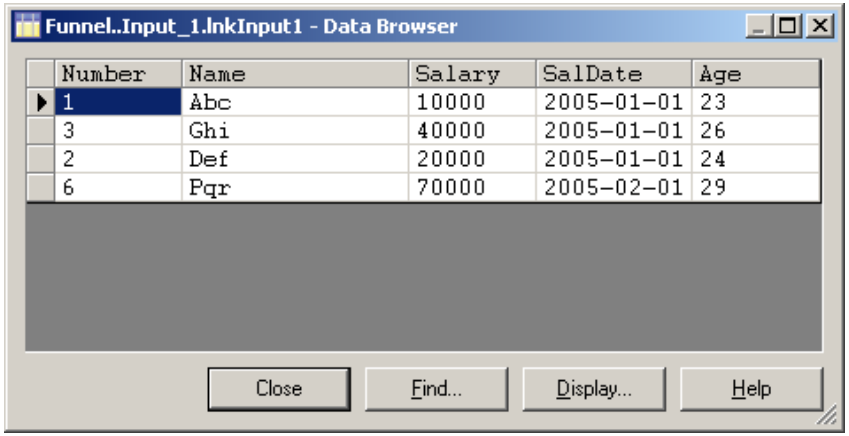
Type – 2 : Sequence



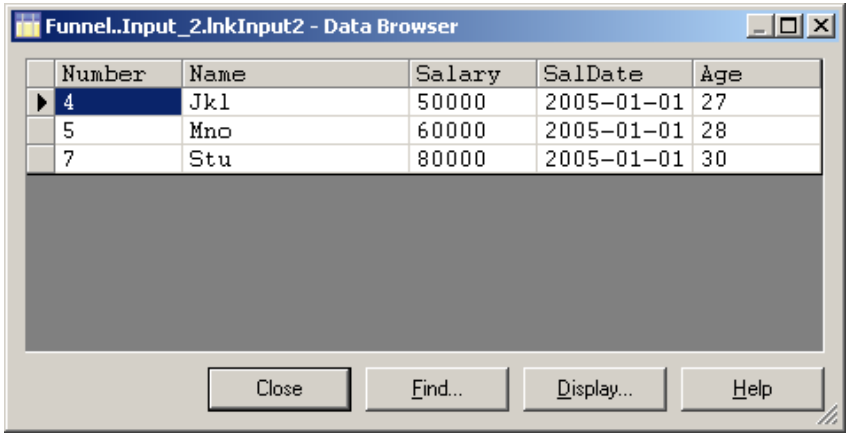
Note : This type of Funnel copies all records from the first input data set to the output data set, then all the records from the second input data set, and so on.

Sequence type ...

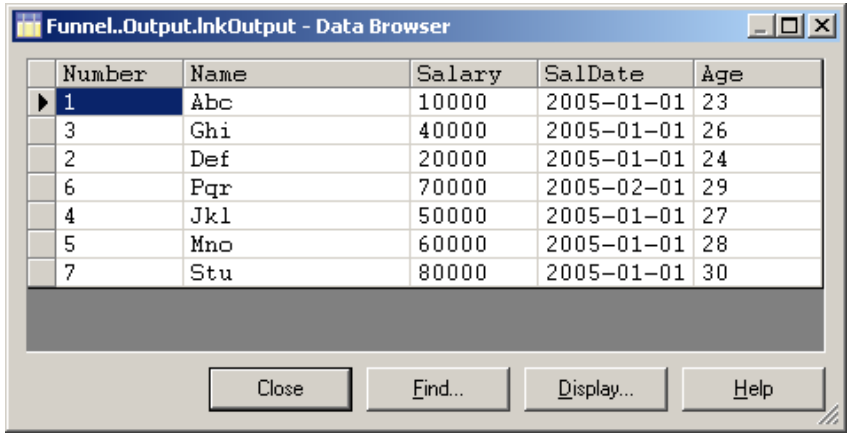
Input - 1 view data :



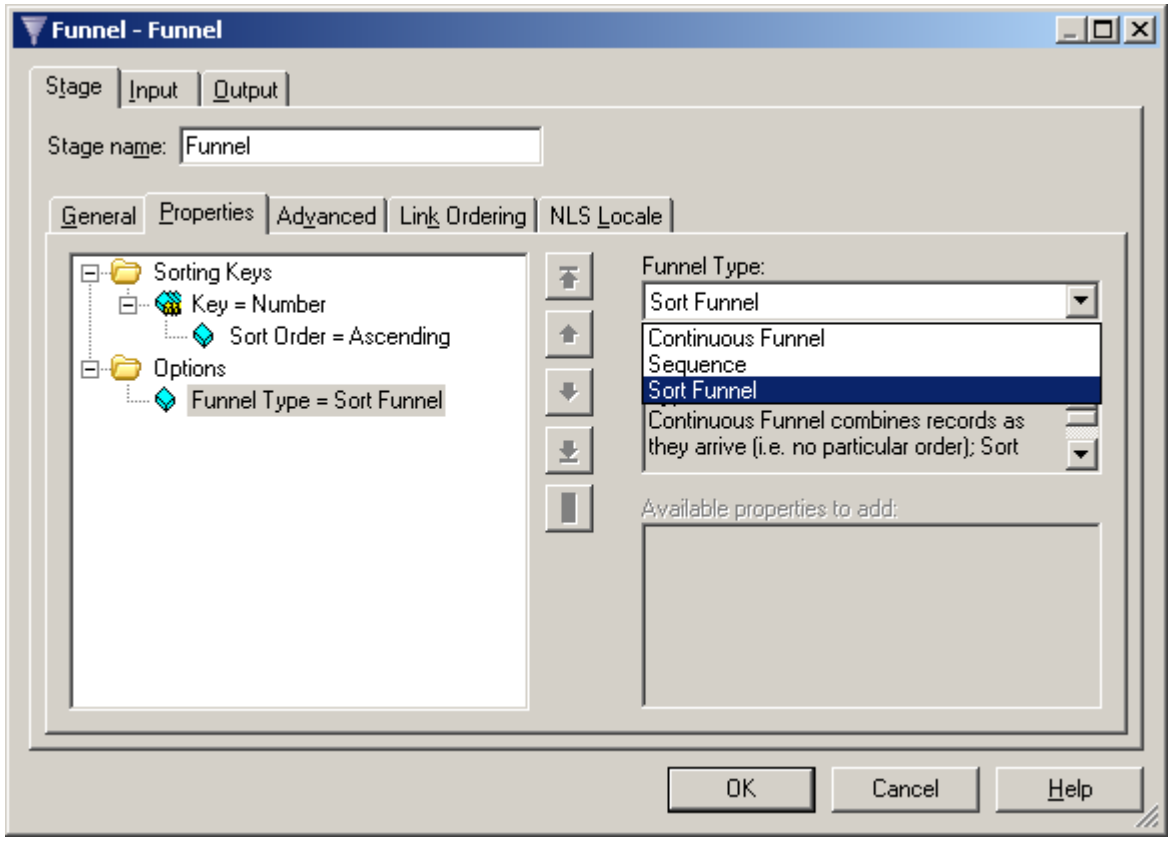
Input - 2 view data :



Output view data :



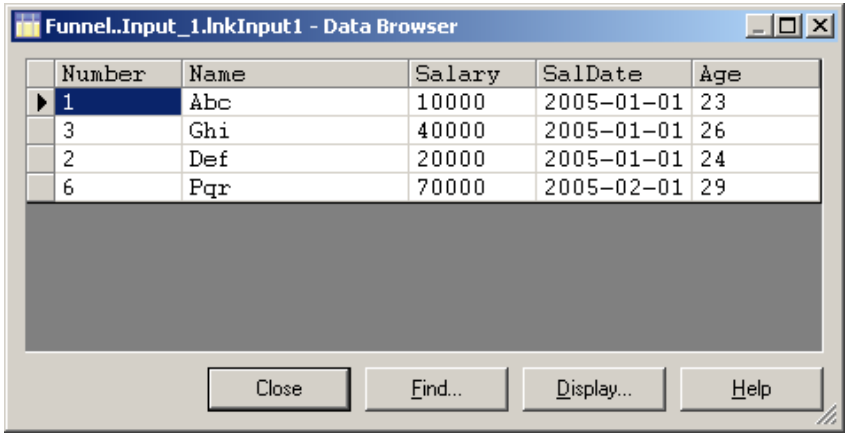
Type – 3 : Sort Funnel



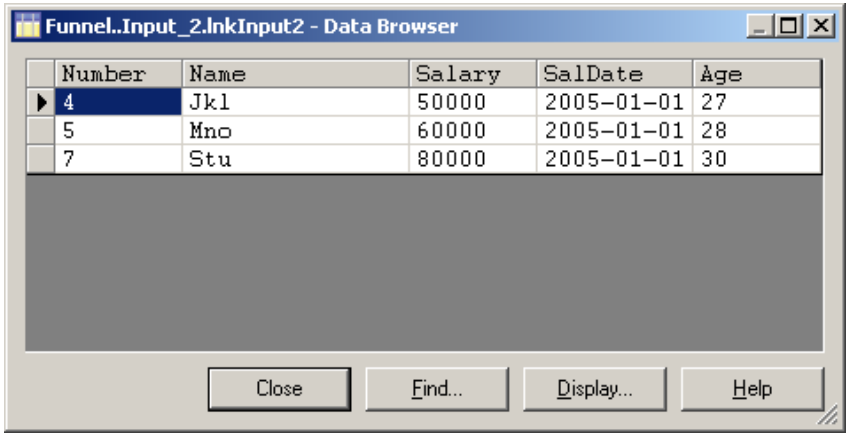
Note : This type of Funnel combines the input records in the order defined by the value(s) of one or more key columns and the order of the output records is determined by these sorting keys.

Sort Funnel type ...

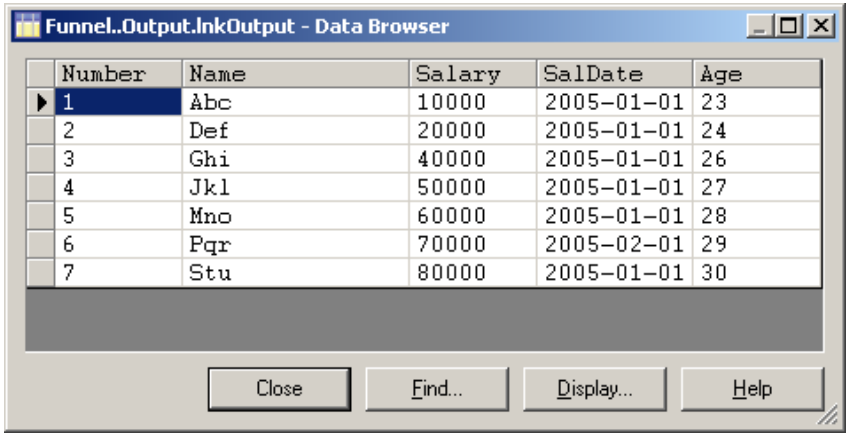
Input - 1 view data :



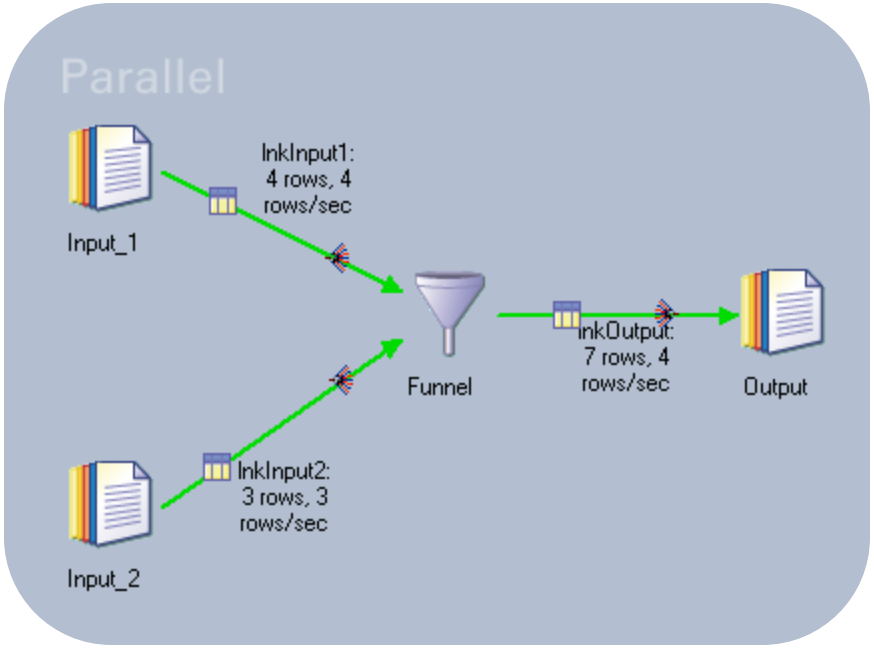
Input - 2 view data :



Output view data :

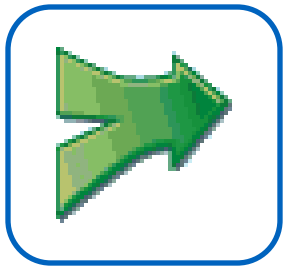


Job :


Parallel
Input_1
InkInput1:
4 rows, 4
rows/sec
Funnel
InkOutput:
7 rows, 4
rows/sec
Output
Input_2
InkInput2:
3 rows, 3
rows/sec

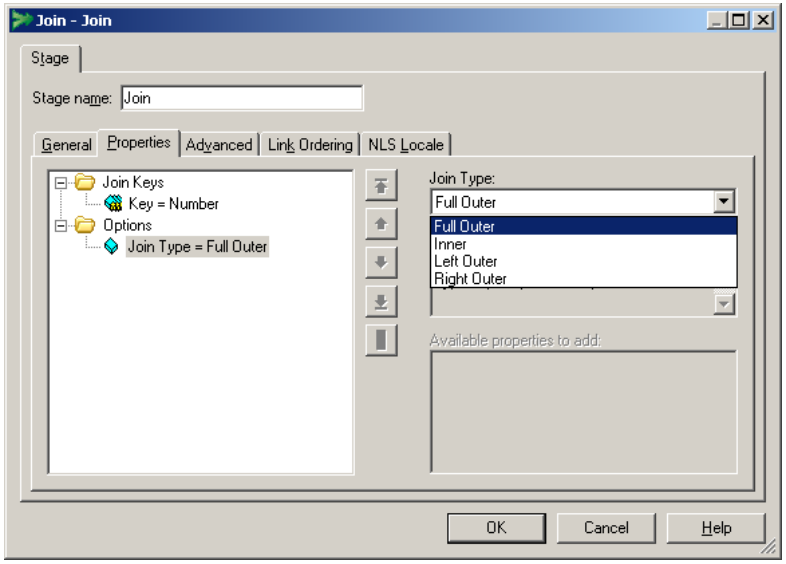
Join Stage :

Definition : Join Stage performs join operations on two or more data sets input to the stage and then outputs the resulting data set.

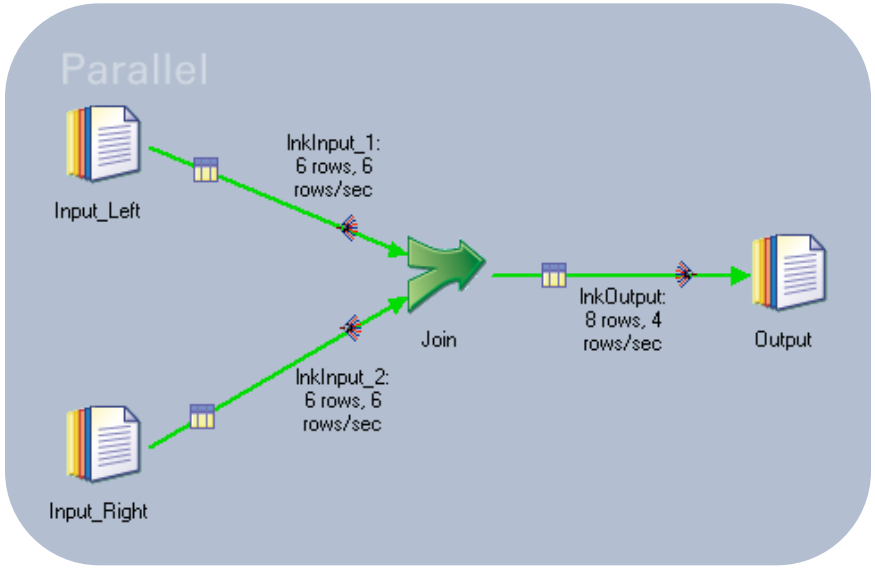


The input data sets are notionally identified as the "right" set and the "left" set, and "intermediate" sets. It has any number of input links and a single output link.

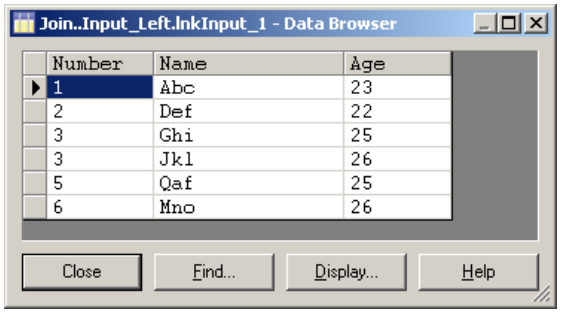
Join Type = Full Outer



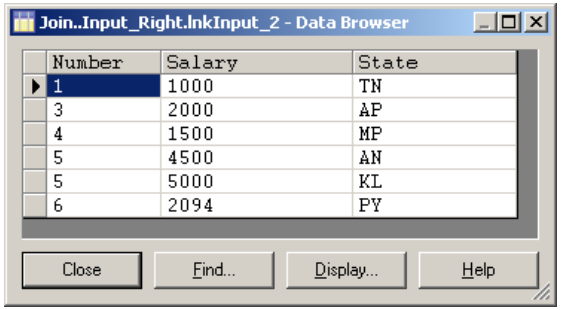
Job :



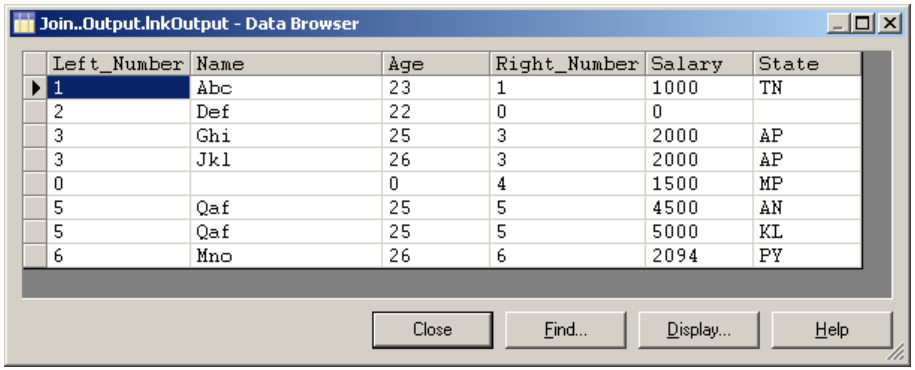
Left Input :



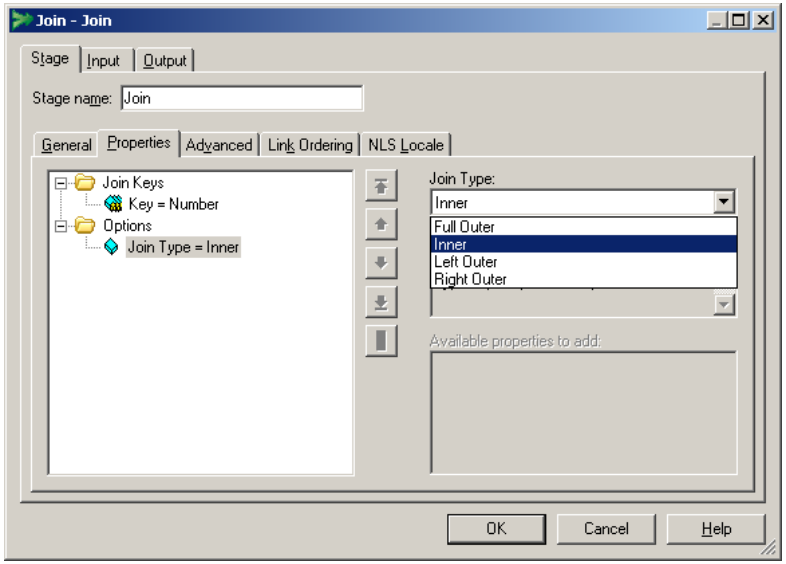
Right Input :



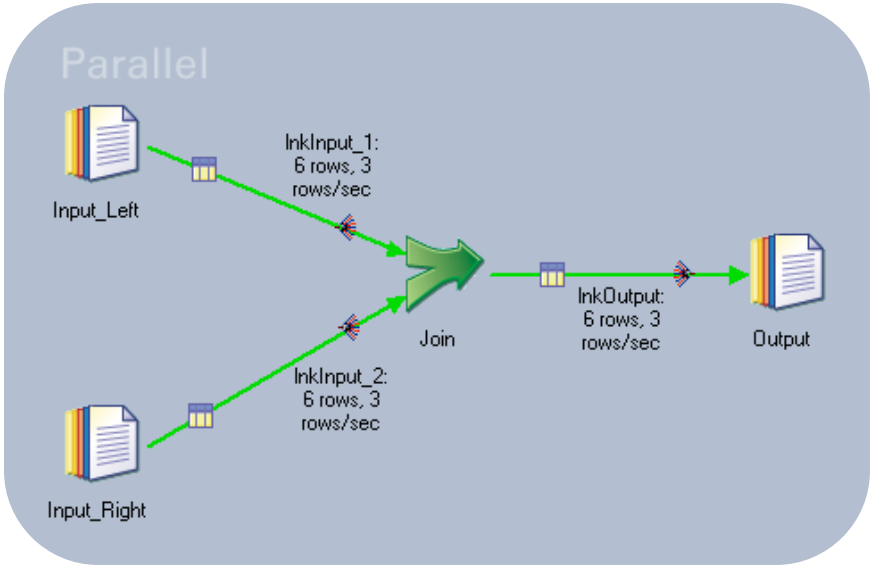
Output :



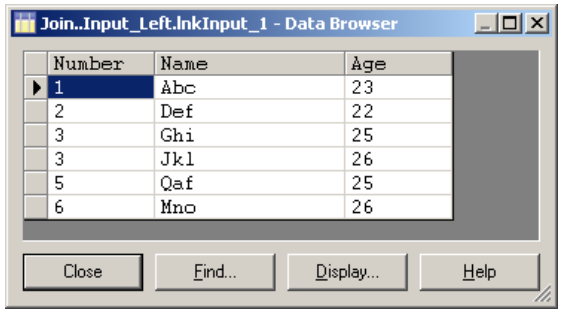
Join Type = Inner



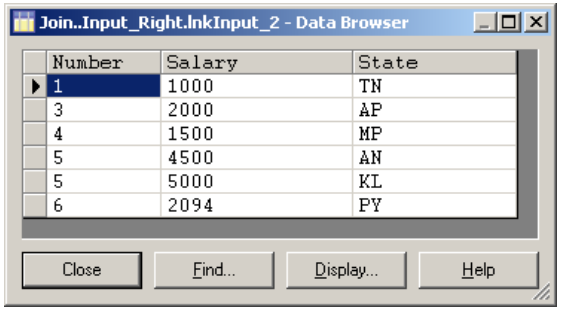
Job :



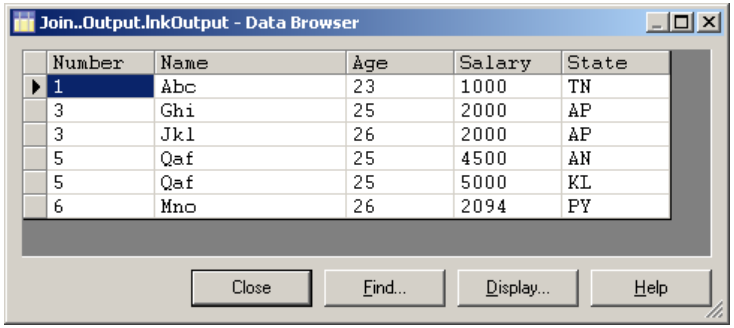
Left Input :



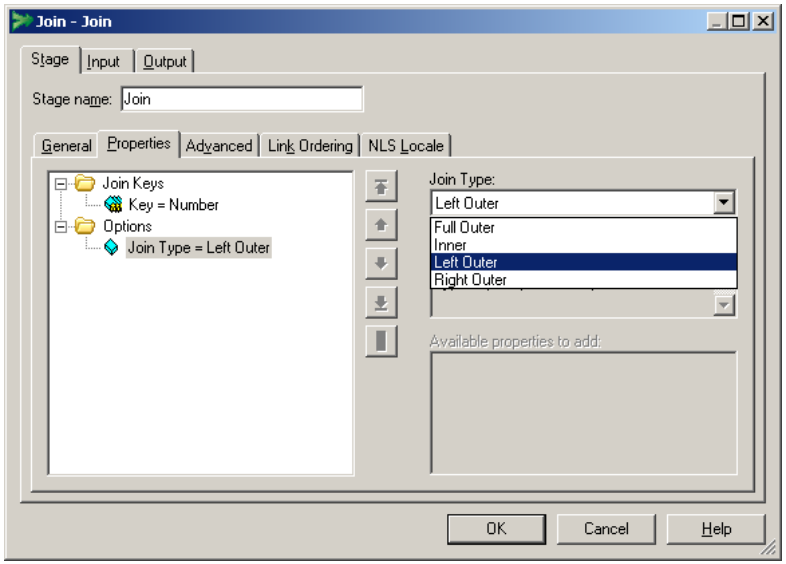
Right Input :



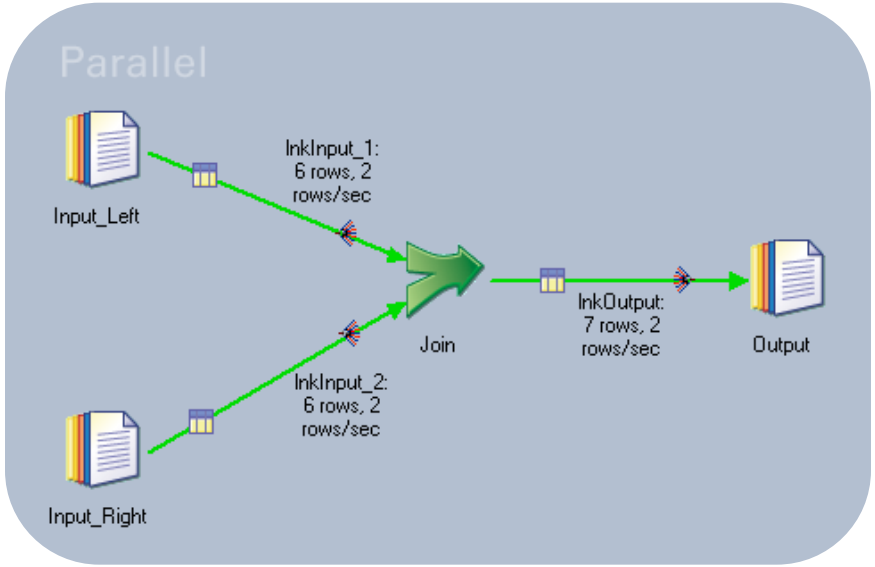
Output :



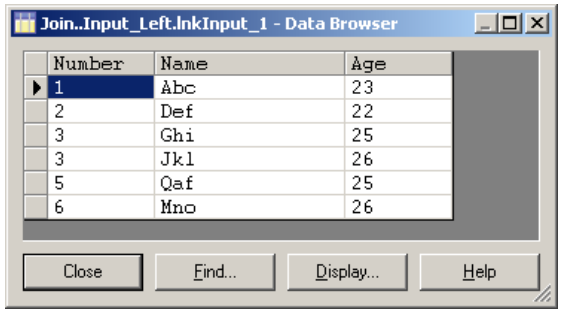
Join Type = Left Outer



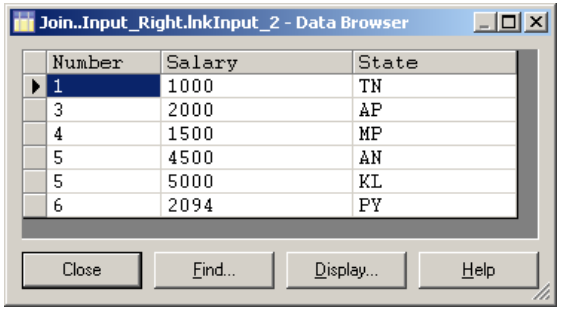
Job :



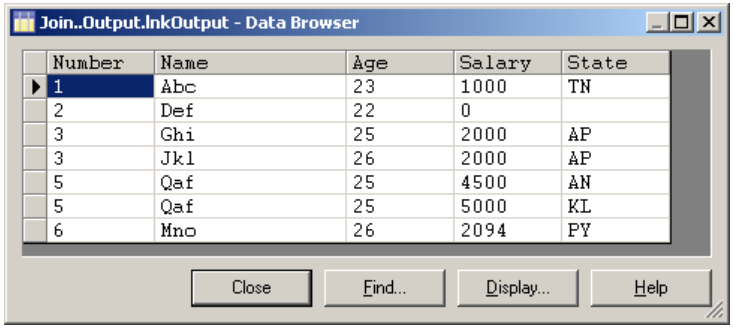
Left Input :



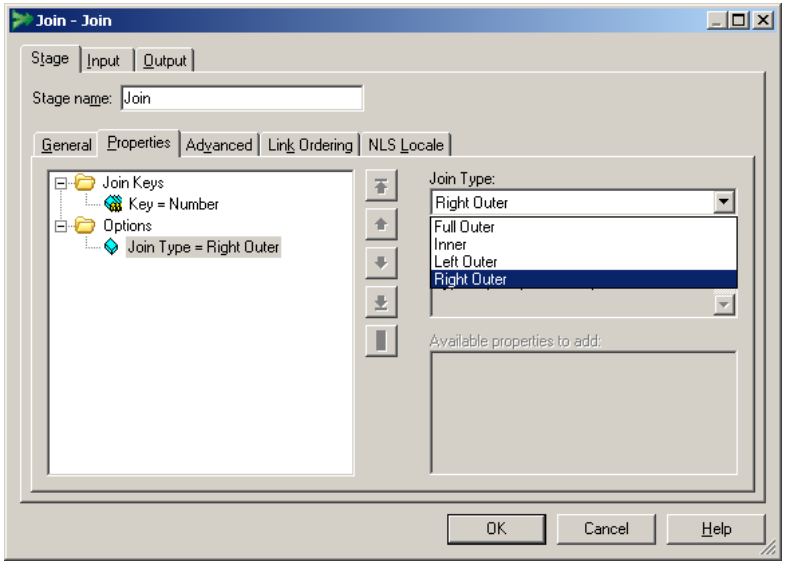
Right Input :



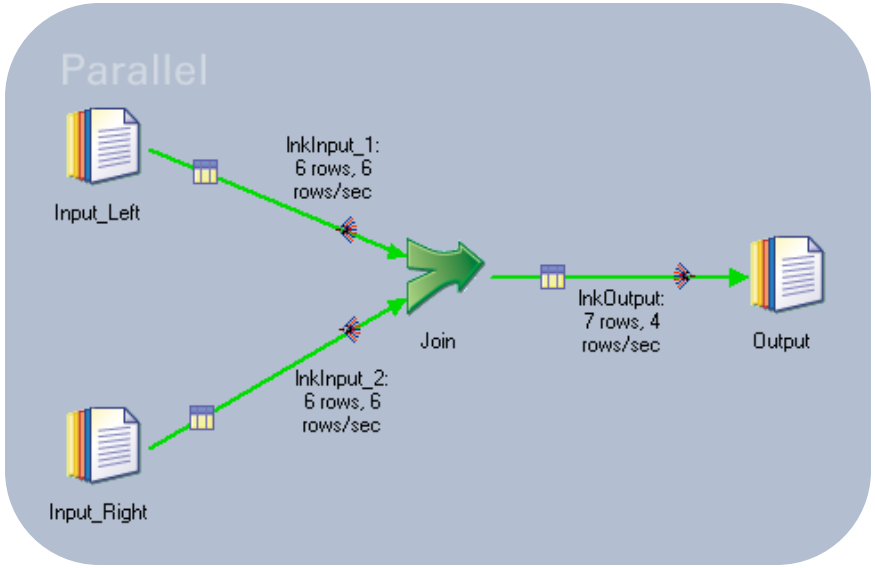
Output :



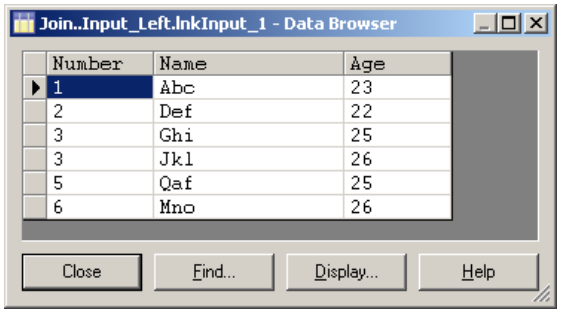
Join Type = Right Outer



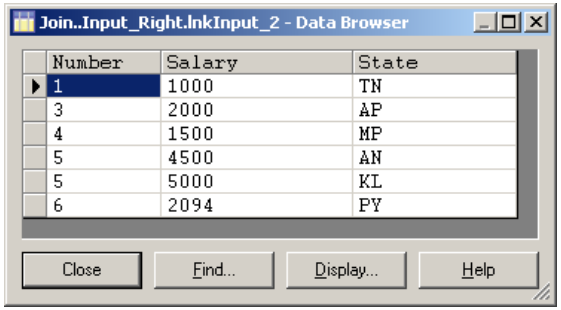
Job :



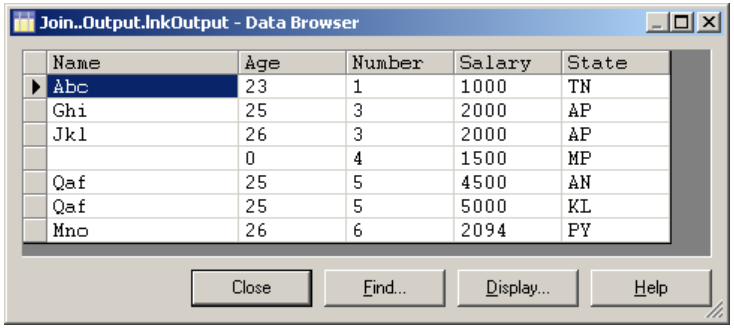
Left Input :



Right Input :



Output :

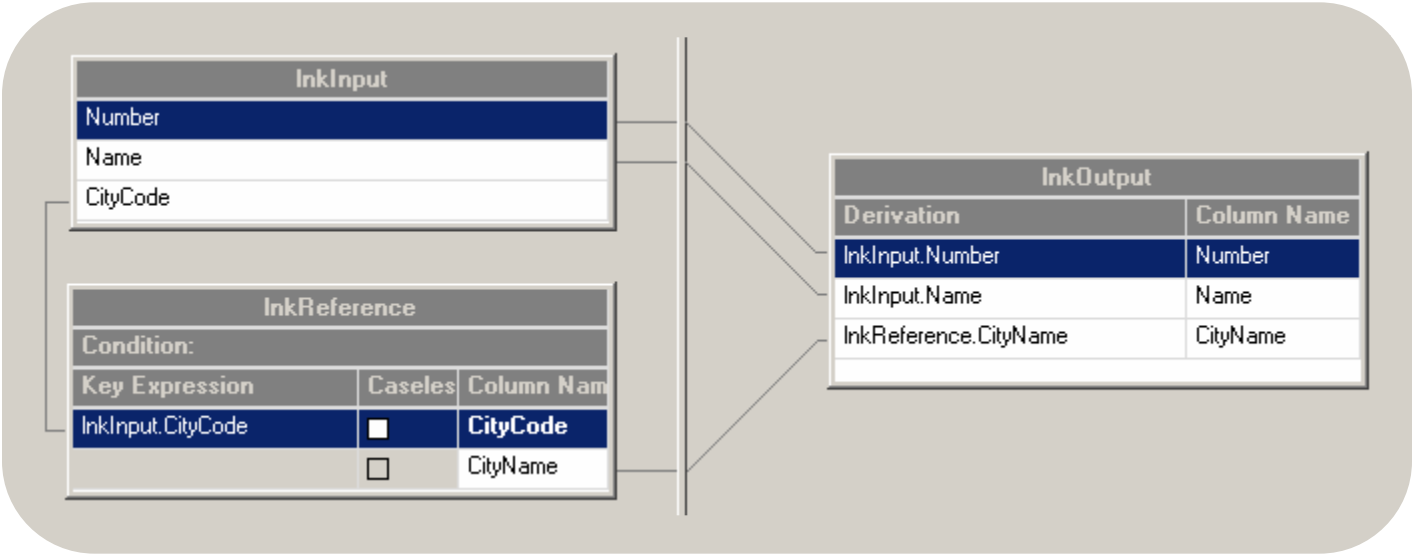


Lookup Stage :

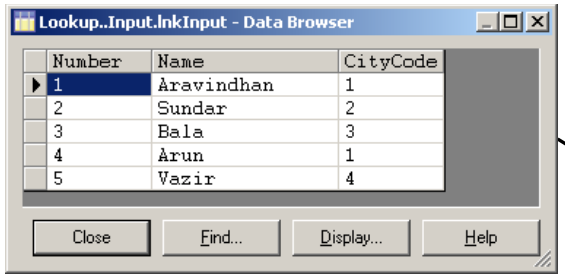
Definition : Lookup Stage used to perform lookup operations on a data set read into memory from any other Parallel job stage that can output data.

It can also perform lookups directly in a DB2 or Oracle database or in a lookup table contained in a Lookup File Set stage.

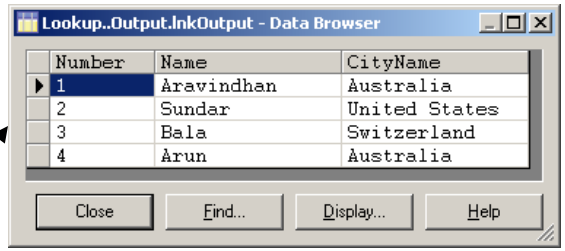
Mappings :


lnkInput
Number
Name
CityCode
lnkReference
Condition:
Key Expression
Caseles
Column Nam
lnkInput.CityCode
CityCode
CityName
lnkOutput
Derivation
Column Name
lnkInput.Number
Number
lnkInput.Name
Name
lnkReference.CityName
CityName

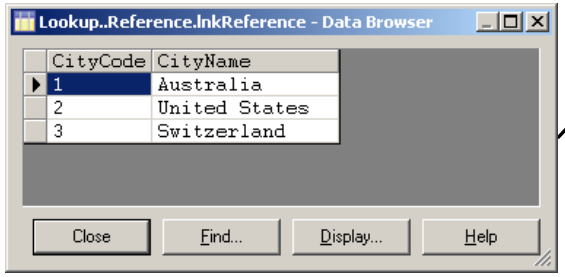
Input :



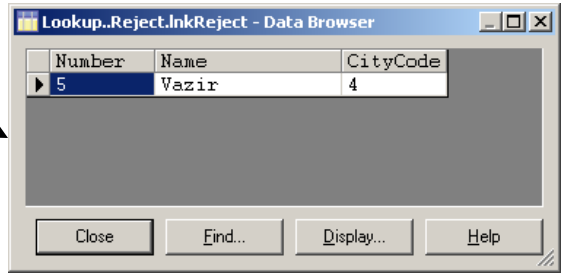
Output :



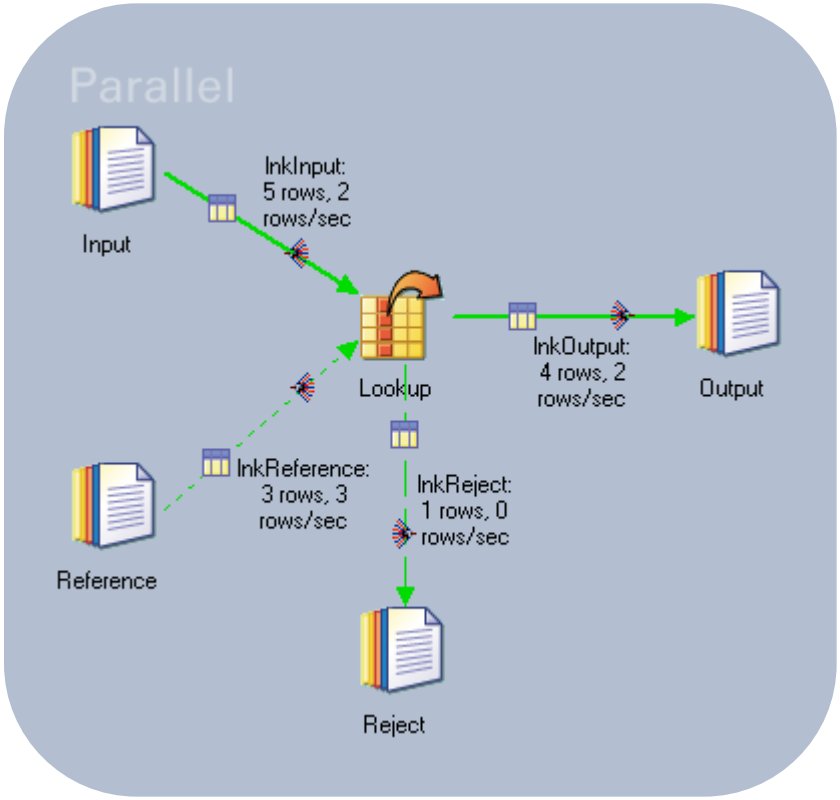
References :



Reject :



Job :


Parallel
Input
lnkInput:
5 rows, 2
rows/sec
Lookup
lnkOutput:
4 rows, 2
rows/sec
Output
Reference
lnkReference:
3 rows, 3
rows/sec
lnkReject:
1 rows, 0
rows/sec
Reject

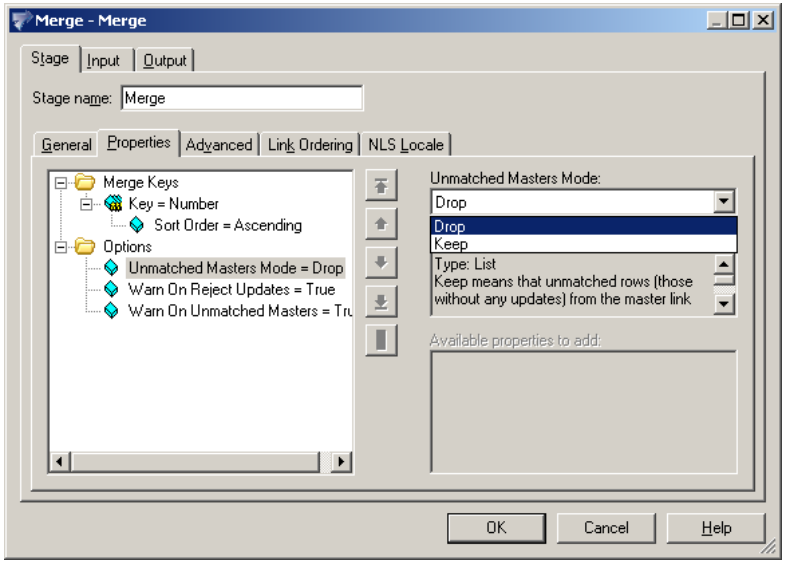
Merge Stage :

Definition : Join Stage combines a sorted 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.

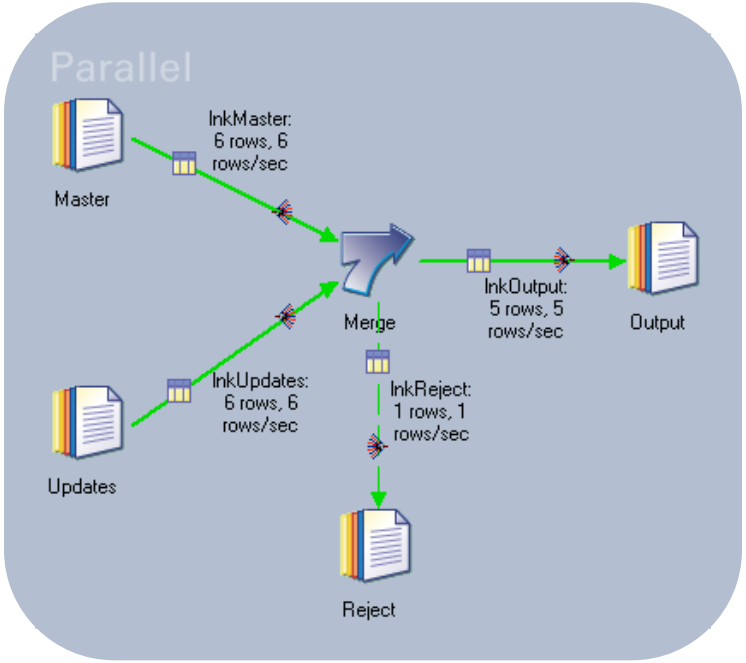
A master record and an update record are merged only if both of them have the same values for the merge key column(s) that you specify. Merge key columns are one or more columns that exist in both the master and update records.

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.

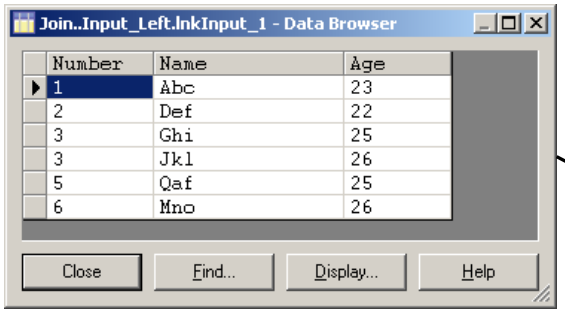
Unmatched Masters Mode = Drop



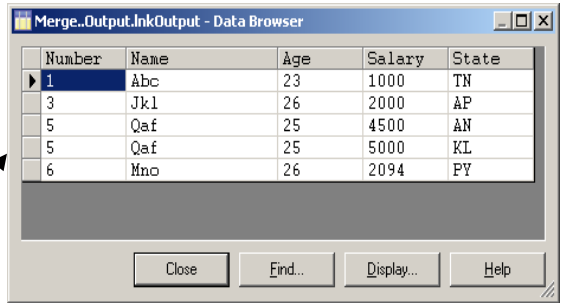
Job :



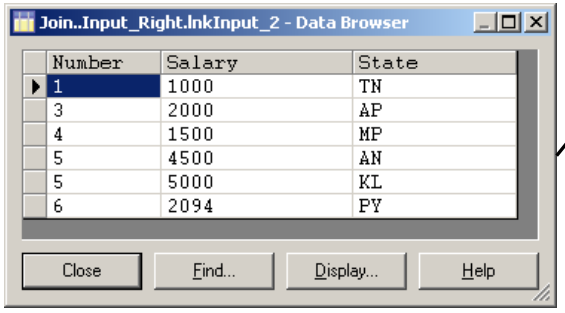
Master :



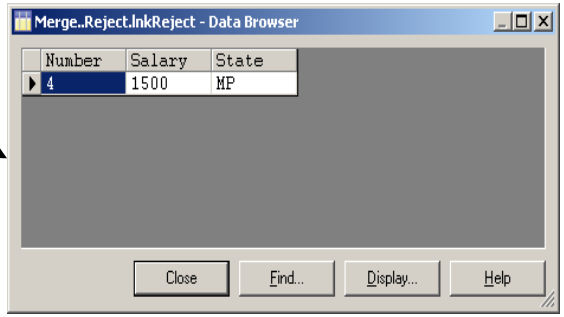
Output :



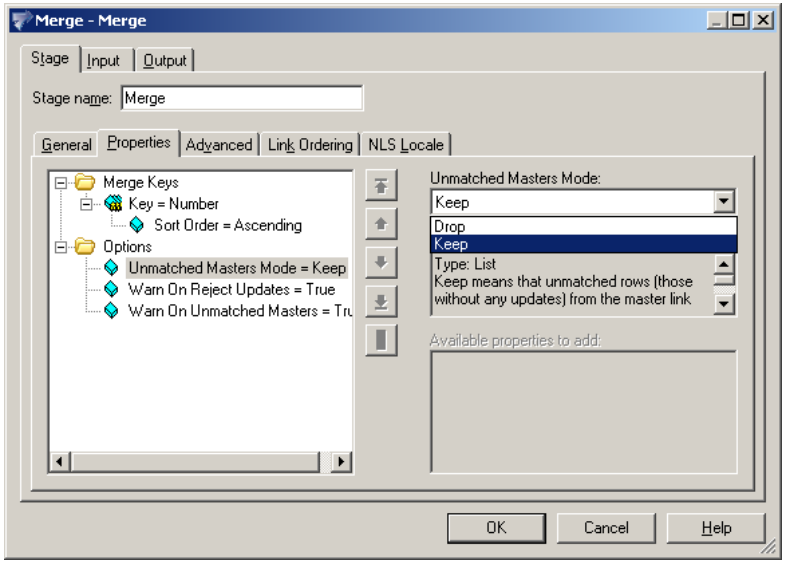
Updates :



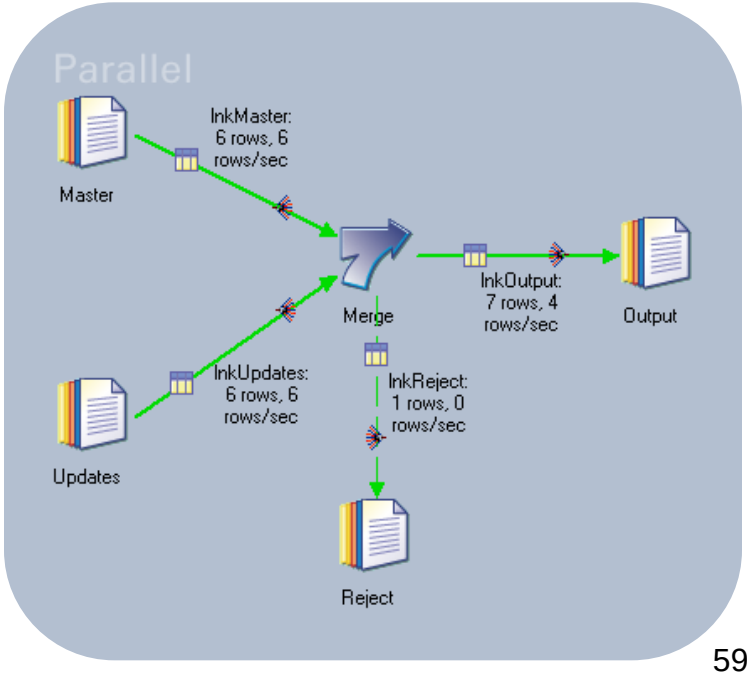
Reject :



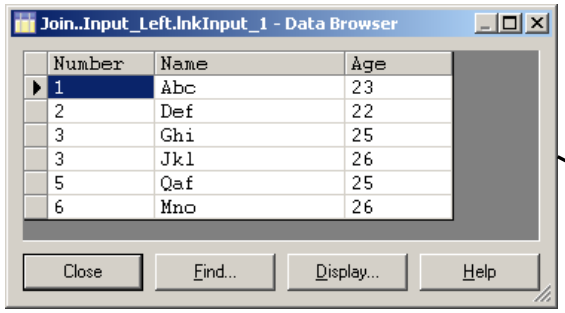
Unmatched Masters Mode = Keep



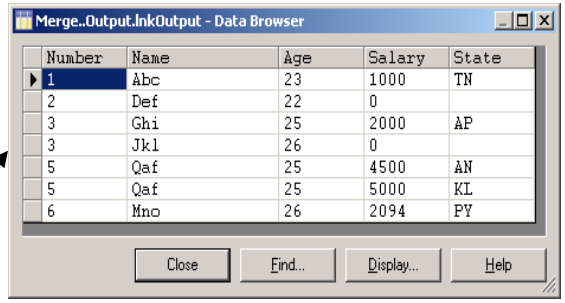
Job :



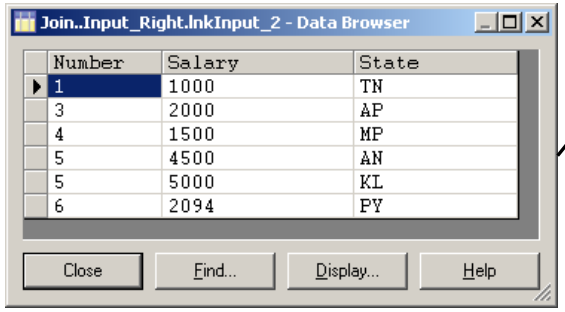
Master :



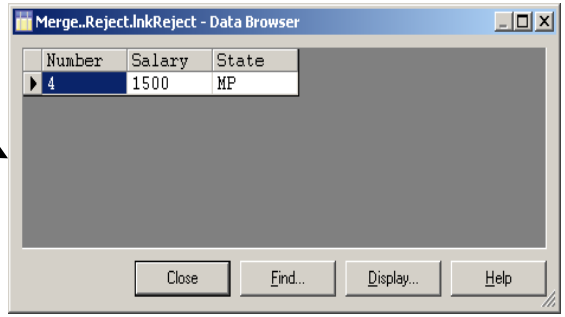
Output :



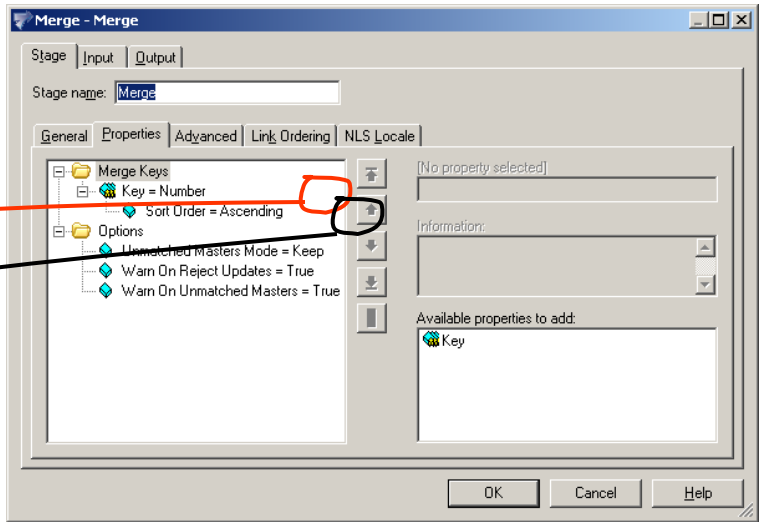
Updates :



Reject :

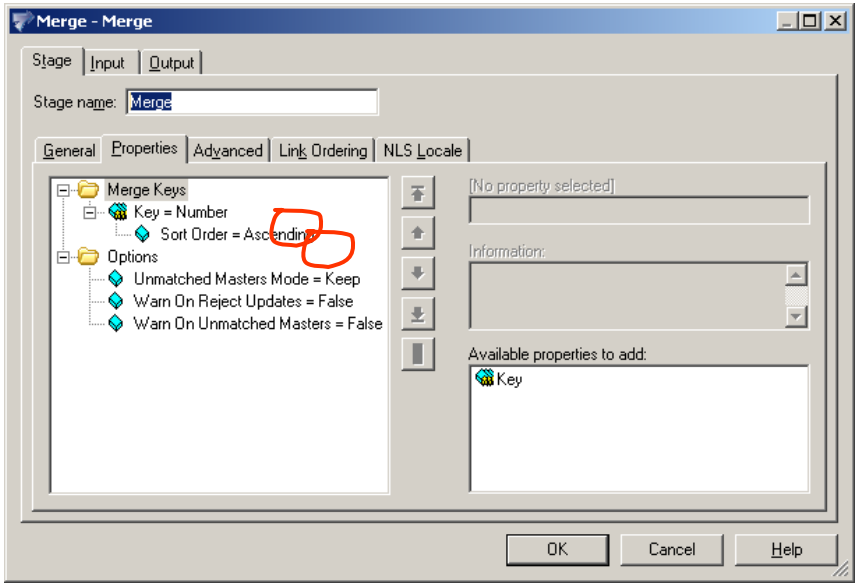


Note : If the options "Warn on Reject Updates = True" and "Warn on Unmatched Masters = True" then the log file shows the warnings on Reject Updates and Unmatched Data from Masters.



>Occurred	>On date	Type	Event
2:08:25 PM	1/6/2004	Control	Starting Job Merge. (...)
2:08:25 PM	1/6/2004	Info	Environment variable settings: (...)
2:08:25 PM	1/6/2004	Info	Parallel job initiated (...)
2:08:25 PM	1/6/2004	Info	Parallel job default NLS map ISO-8859-1, default locale OFF
2:08:26 PM	1/6/2004	Info	main_program: Ascential DataStage(tm) Enterprise Edition 7.5 (...)
2:08:26 PM	1/6/2004	Info	main_program: orchgeneral: loaded (...)
2:08:28 PM	1/6/2004	Info	main_program: APT configuration file: C:/Ascential/DataStage/Configurations/default.apt (...)
2:08:28 PM	1/6/2004	Info	Updates,0: Import complete. 6 records imported successfully, 0 rejected.
2:08:29 PM	1/6/2004	Info	Master,0: Import complete. 6 records imported successfully, 0 rejected.
2:08:29 PM	1/6/2004	Warning	Merge,0: Master record (1) has no updates
2:08:29 PM	1/6/2004	Warning	Merge,0: Master record (3) is a duplicate
2:08:29 PM	1/6/2004	Warning	Merge,0: Update record (2) of data set 1 rejected (no master record)
2:08:29 PM	1/6/2004	Info	Output,0: Export complete. 7 records exported successfully, 0 rejected.
2:08:29 PM	1/6/2004	Info	Reject,0: Export complete. 1 records exported successfully, 0 rejected.
2:08:29 PM	1/6/2004	Info	main_program: Step execution finished with status = OK.
2:08:29 PM	1/6/2004	Info	main_program: Startup Time 0:02 Production Run Time 0:00
2:08:29 PM	1/6/2004	Info	Parallel job reports successful completion
2:08:29 PM	1/6/2004	Control	Finished Job Merge.

Note : If the options "Warn on Reject Updates = False" and "Warn on Unmatched Masters = False" then the log file do not shows the warnings on Reject Updates and Unmatched Data from Masters.



>Occurred	>On date	Type	Event
2:34:09 PM	1/6/2004	Control	Starting Job Merge. (...)
2:34:09 PM	1/6/2004	Info	Environment variable settings: (...)
2:34:09 PM	1/6/2004	Info	Parallel job initiated (...)
2:34:09 PM	1/6/2004	Info	Parallel job default NLS map ISO-8859-1, default locale OFF
2:34:10 PM	1/6/2004	Info	main_program: Ascential DataStage(tm) Enterprise Edition 7.5 (...)
2:34:10 PM	1/6/2004	Info	main_program: orchgeneral: loaded (...)
2:34:12 PM	1/6/2004	Info	main_program: APT configuration file: C:/Ascential/DataStage/Configurations/default.apt (...)
2:34:12 PM	1/6/2004	Info	Master,0: Import complete. 6 records imported successfully, 0 rejected.
2:34:13 PM	1/6/2004	Info	Updates,0: Import complete. 6 records imported successfully, 0 rejected.
2:34:13 PM	1/6/2004	Info	Output,0: Export complete. 7 records exported successfully, 0 rejected.
2:34:13 PM	1/6/2004	Info	Reject,0: Export complete. 1 records exported successfully, 0 rejected.
2:34:13 PM	1/6/2004	Info	main_program: Step execution finished with status = OK.
2:34:13 PM	1/6/2004	Info	main_program: Startup Time 0:03 Production Run Time 0:00
2:34:13 PM	1/6/2004	Info	Parallel job reports successful completion
2:34:13 PM	1/6/2004	Control	Finished Job Merge.

TATA CONSULTANCY SERVICES
Experience certainty.

Thank You

IT Services
Business Solutions
Outsourcing