My Datastage Notes

Pages

- Home
- Different Versions
- All Datastage Stages
- Configuration File
- Sequentional_Stage
- Dataset
- Transformer Stage
- Sort Stage
- Aggregator_Stage
- Join Stage
- Lookup_Stage
- Merge_Stage
- Filter_Stage
- Copy Stage
- Funnel_Stage
- Column Generator
- Surrogate_Key_Stage
- SCD
- Pivot_Enterprise_Stage
- Sequence_Activities
- Datastage Study Material/Interview Questions
- Datastage Errors and Resolution
- Datastage Scenarios and solutions
- Unix Shell Scripting
- SQL/Database
- Datawarehousing Concepts

Blog Archive

▼ 2014 (46)

▼ September (40)

Monday, September 8, 2014

Scenario: To get the Unique and Duplicates values from Input Data

Input :

There is a input file which contains duplicates data, Suppose :

13
22
95
37
78
87
29
33
33
13
12
87
21
32
13

In this file :
Unique values are : 22 95 37 78 29 12 21 32
Duplicate values are : 13 33 87

Now, we need 3 kind of outputs:

Job1:

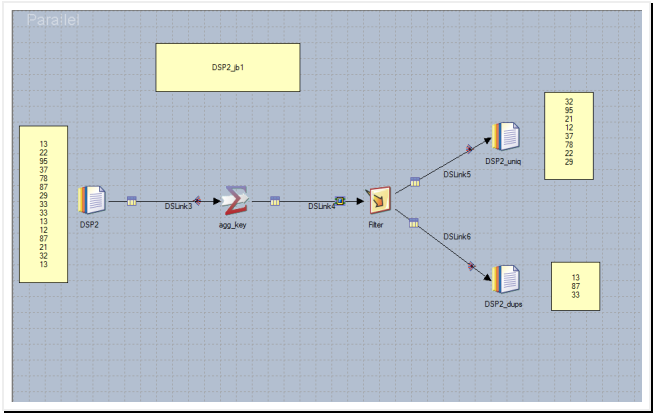
We need 2 o/p file
o/p1 --> Contains Uniq values
o/p2 --> Contains Duplicate Values (each once) i.e - 13 33 87

DataStage Scenario - Design 2 - job1

Solution Design :

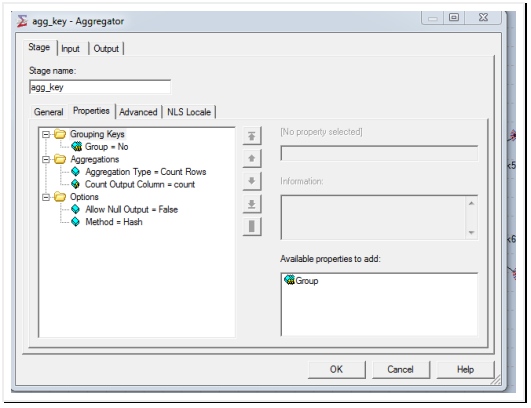
a) Job Design :

Below is the design which can achieve the output as we needed. Here, we are reading seq file as a input, then data is passing through Aggregator and Filter stage to achieve the output**.**

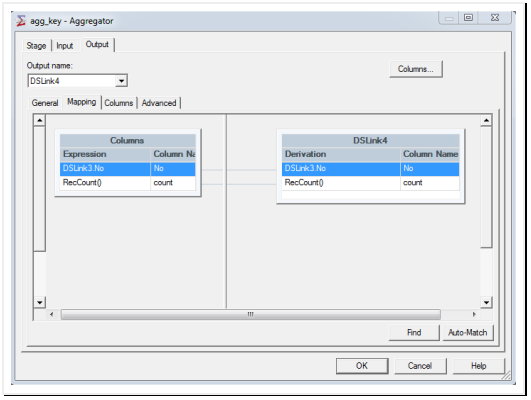


b) Aggregator Stage Properties

Input data contains only one column "No" **,** In Aggregator stage, we have group the data on the "No" column and calculate the rows for each Key (No).

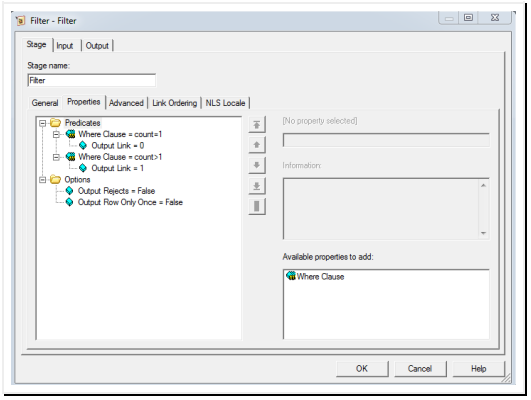


When we have used the "Count Rows" aggregation type, it will generate a new column which contain the count for each Key (No). Here we have given the column name - "count" and assigned to output as below.

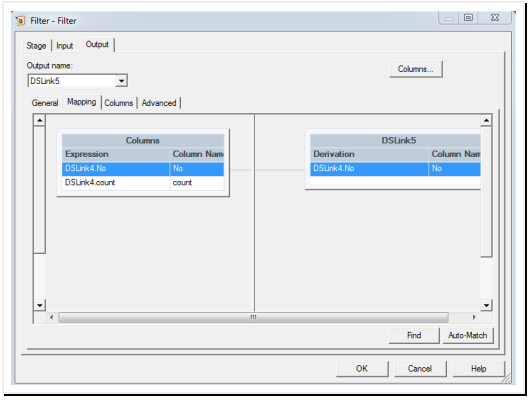


c) Filter Stage Properties

In Filter stage, we put 2 where condition **count=1** and **count>1**. and assigned different output files to both conditions.



Assigned the data (column No) to output tab.



d) Output File

We got two output from the jobs

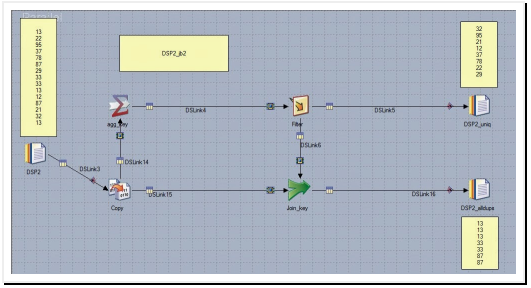
i) Contains where count=1 (unique values in input)
ii) Contains where count>1 (dups values in input)

Job2

We need 2 o/p file
o/p1 --> Contains Uniq values
o/p2 --> Contains Duplicate Values (no of times they appear) i.e - 13 13 13 33 33 87 87
DataStage Scenario - Design2 - job2

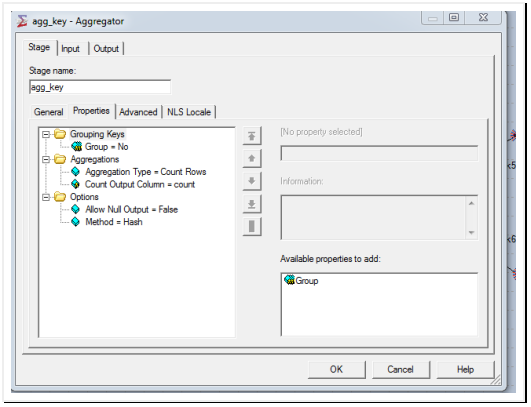
Solution Design :

a) Job Design :
In job design, we are using Copy, Aggregator, Filter and Join stage to get the output.

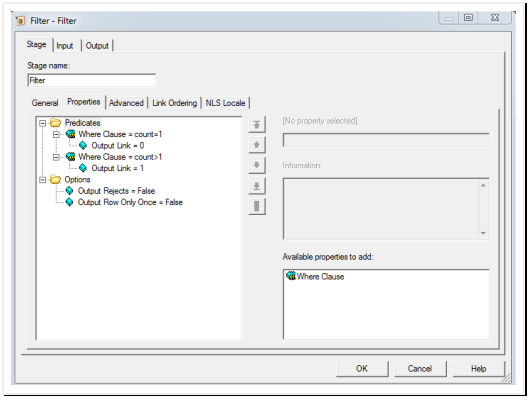


b) Copy Stage Properties :
Simply map the input to both link output. first link goes to Aggregator and second link goes to Join stage.

c) Aggregator Stage Properties :
Input data contains only one column "No" **,** In Aggregator stage, we have group the data on the "No" column and calculate the rows for each Key (No).



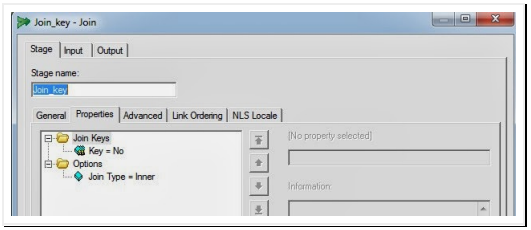
d) Filter Stage Properties :
In Filter stage, we put 2 where condition **count=1** and **count>1**. and assigned different links to both conditions.



From filter Stage, first link (count=1) map to output file (which contains the unique records)
and second link we map with Join stage.

e) Join Stage Properties :

In join stage, we join the both input on key column (No).



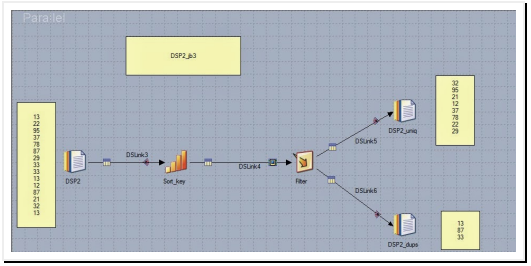
Output from Join map with second output files which contains all the dups as occur in input.

Job3

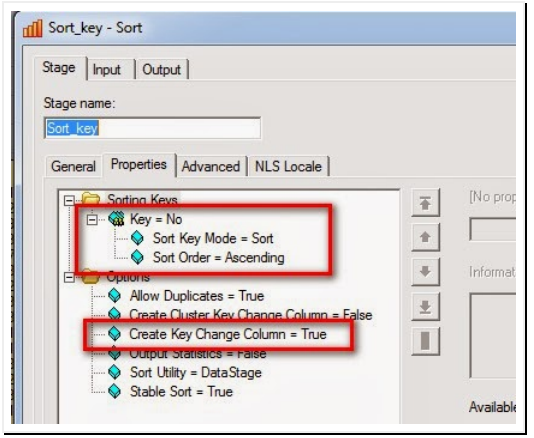
We need 2 o/p file
o/p1 --> Contains all values once each i.e - 22 95 37 78 29 12 21 32 13 33 87
o/p2 --> Contains remaining values - 13 13 33 87
DataStage Scenario - Design2 - job3

Solution Design :

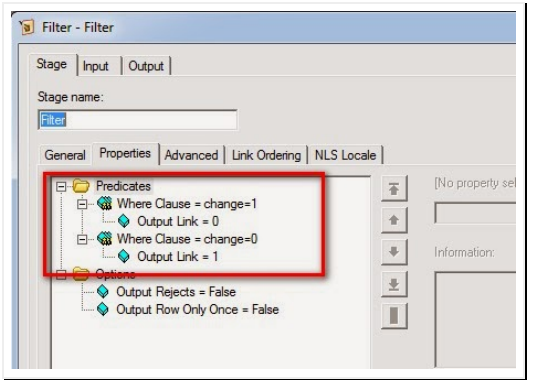
a) Job Design :



b) Sort Stage Properties :
In sort stage, we sort the data on Key column (no) and generate the change key column.



c) Filter Stage Properties :
filer the data on "Change" column generated in sort stage.



In filter stage, condition (change =1) gives you all values (each once) from input and condition (change=0) gives the all duplicate occurrence from input.

Posted by manohar at 1:47 AM

No comments:

Post a Comment

Newer Post Home Older Post

Subscribe to: Post Comments (Atom)