Ab initio GDE

Lab Book

1. Filter by expression & Reformat

**Reformat:**

1. Create a graph to drop a column from an input file.
2. Create a graph to concatenate values from 2 field,

- first 3 characters from empid field

- first 3 characters from name field

Concatenated result must be stored in newly created column called as userid.

1. Create a graph which will read input from below given input file and will convert name of employee in upper case in the output file.
2. Create a graph which will create a new column known as slno, which creates sequence number. (Hint: use next\_in\_seq)

**Filter By Expression:**

1. Create a graph to filter all records who belongs to Pune city.
2. Create a graph to filter records who year of joining is greater than 2016 and belonging to HR dept. Create below file and populate some sample data.

Emp.dat

empid,empname,yearofjoin,department

**Sort**

1. Create a graph to sort employee details based on Salary.
2. Create a graph to sort employee details based on Salary, If salaries are same, next it should sort data based on designation.
3. Create a graph to find employee details who is paid with highest salary and employee paid with lowest salary.

Joins:

**Problem Statement:** From given input file, find the sim id’s whose status is in Active mode and is not cancelled.

|  |  |
| --- | --- |
| Input Files | Simstatus.dat |
| Output file | Activestatus.dat |
| dmlfile | Simstatus.dml |
| Components | input file, output file, filter by expression,joins. |

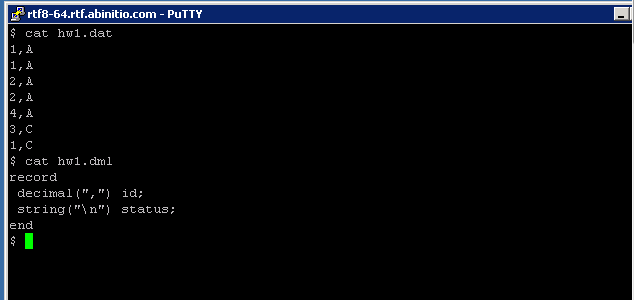
**INPUT FILE**

|  |  |
| --- | --- |
| ID | Status |
| 1 | A |
| 1 | A |
| 2 | A |
| 2 | A |
| 4 | A |
| 3 | C |
| 1 | C |

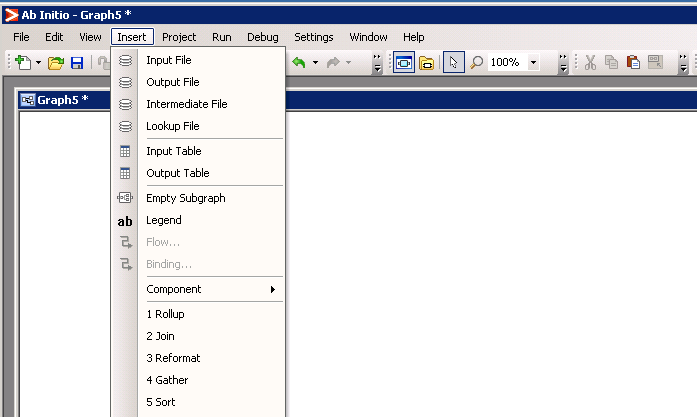
**OUTPUT FILE**

|  |  |
| --- | --- |
| **ID** | **Status** |
| 2 | A |
| 2 | A |
| 4 | A |

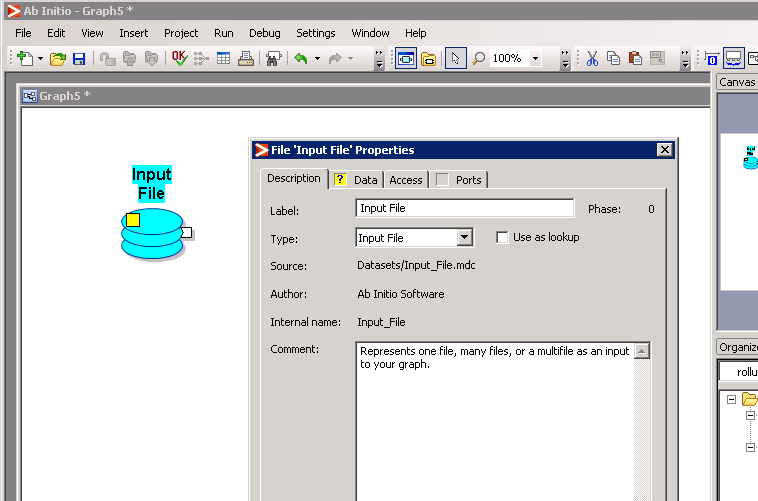
**STEP 1 :** Create an input file and .dml file as below.



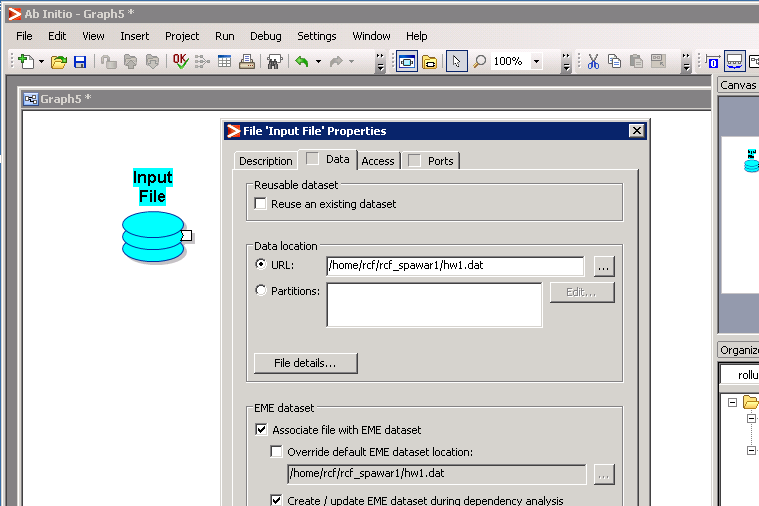
**STEP 2**: Open the GDE Environment and click on FILE tab and click on new to create a new graph and then click on INSERT and double click on Input file component to add it to the graph.

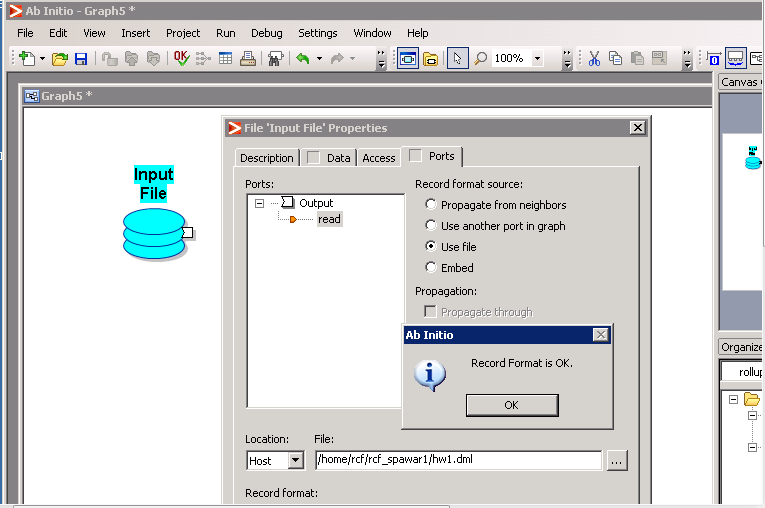
****

**STEP 3:** Double click on the input file and the following window opens.

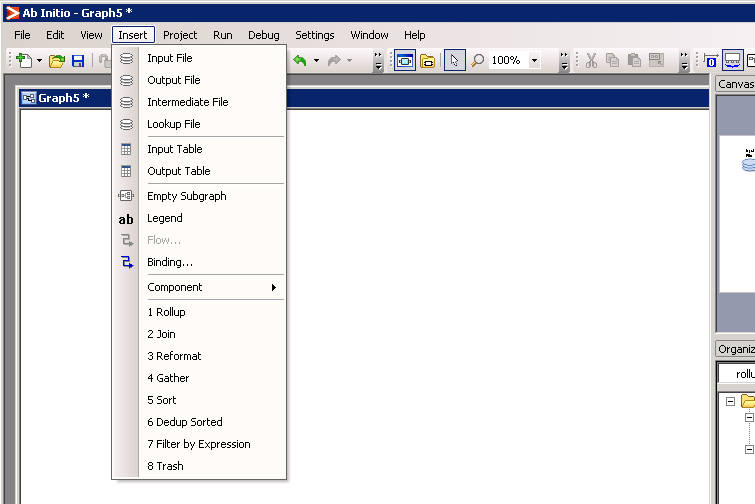


**STEP 4:** Click on Data and in the URL field give the path of your input file. Similarly in Ports select USE FILE and mention the path of your dml file and click on validate.

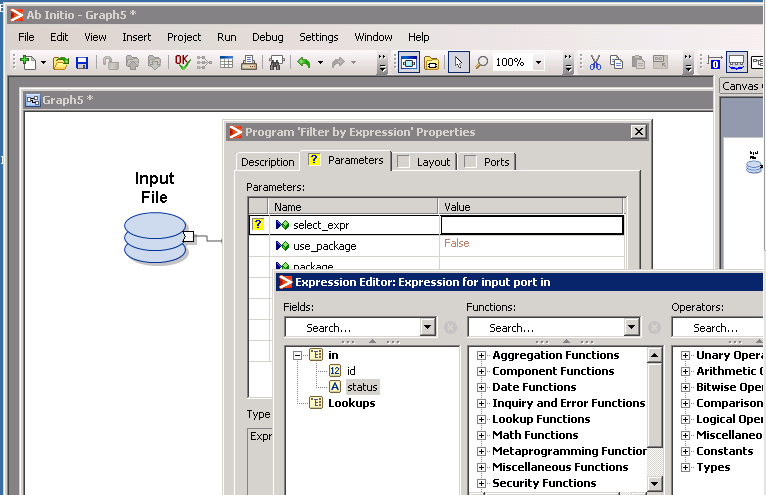




STEP 5: Navigate to the Insert tab and double click on the Filter by expression component to add it to the graph.



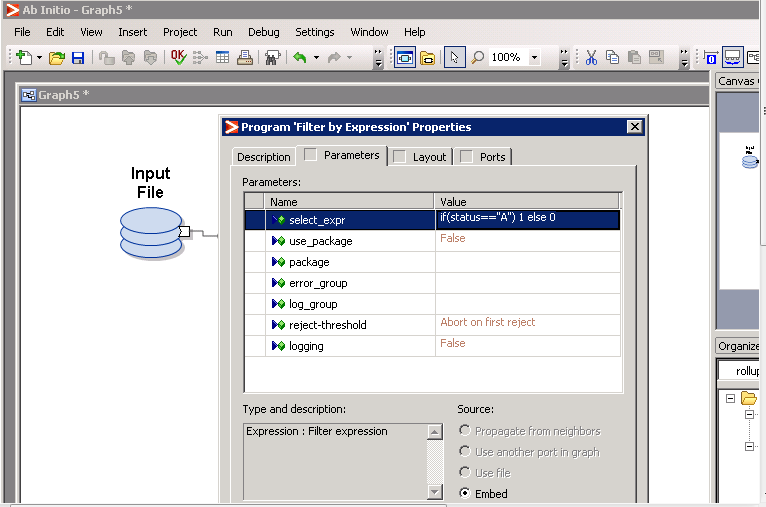
STEP 6: Double click on the filter by expression component and navigate to Parameters and double click on select expression to mention the condition which filters the data.



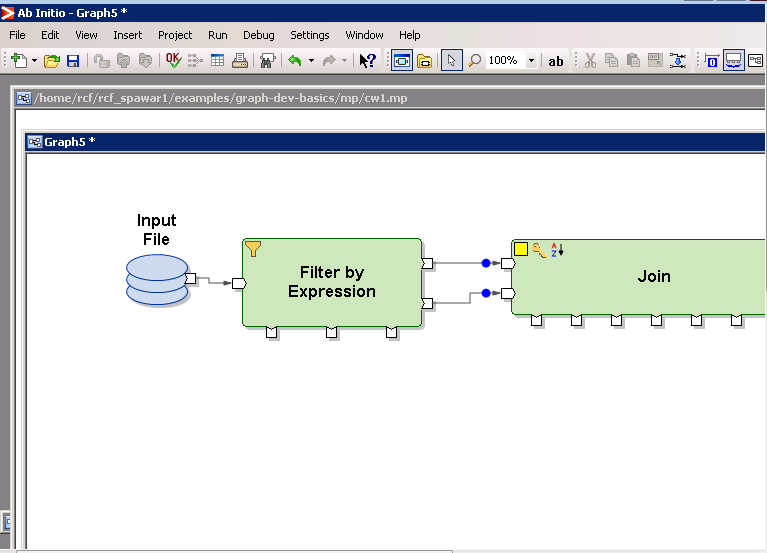
STEP 7 : The condition to filter the data is as follows :

if(status==”A”) 1 else 0.

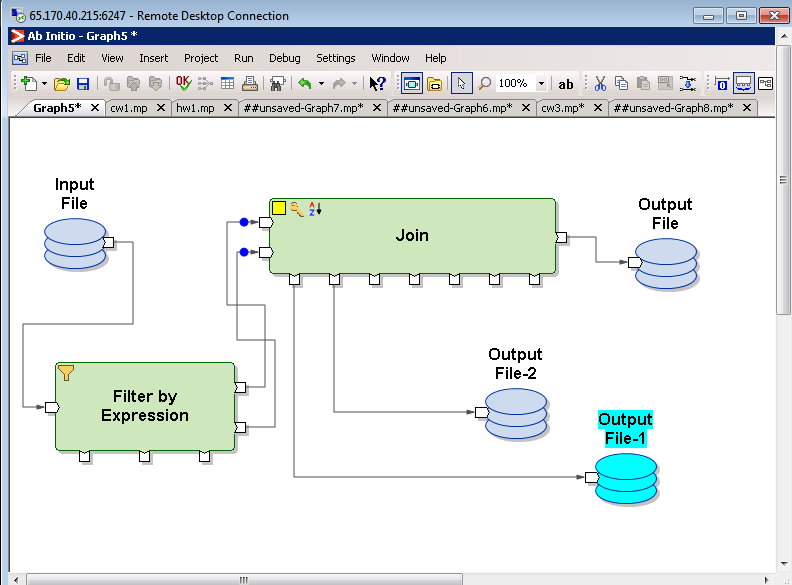
This condition separates the records with status A into port 1 and status C to the second port. Click on validate.



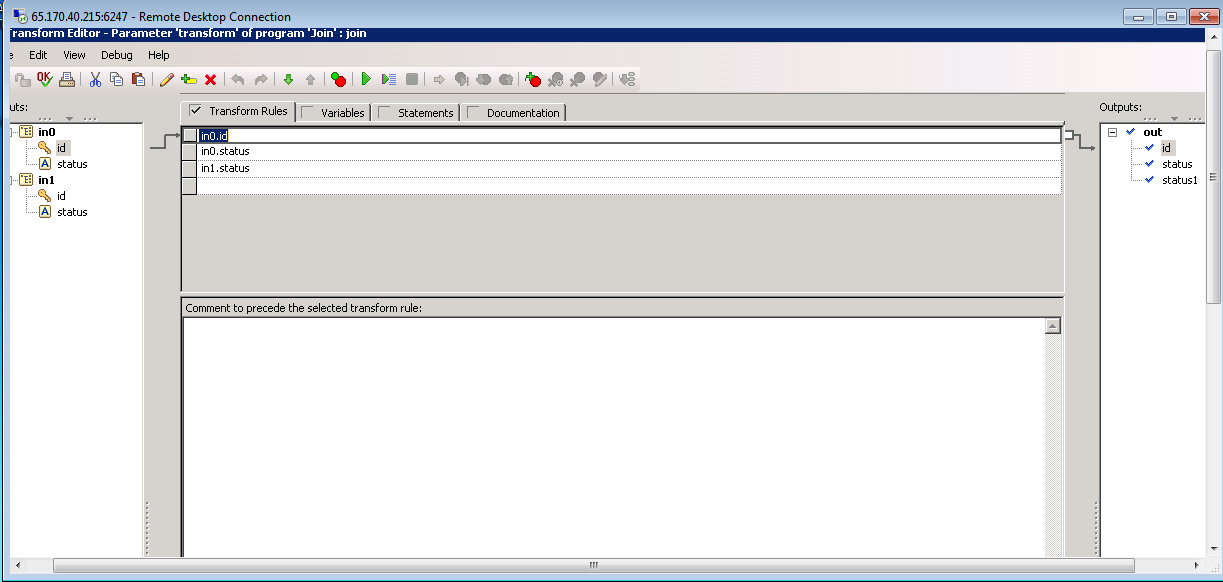
STEP 8: Navigate to Insert and double click on JOIN component to add it and join the output ports of filter by expression as input to the join component.

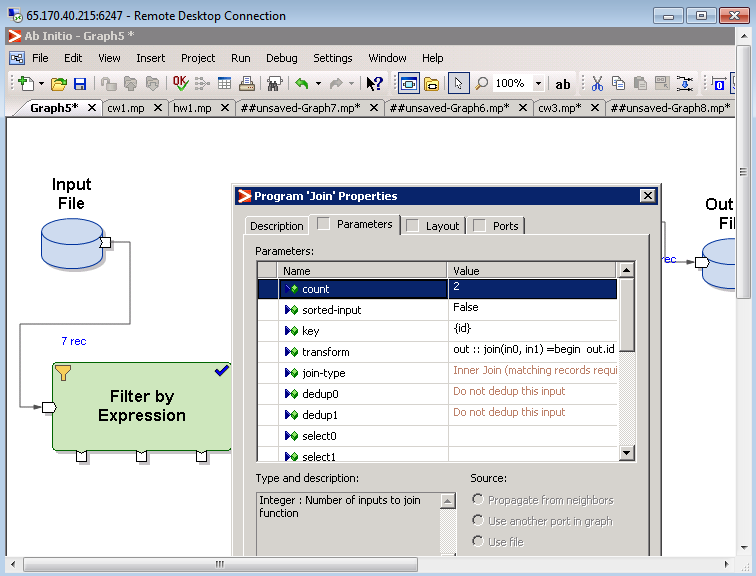


STEP 9 : Navigate to the Insert tab and insert three output file components and connect one of the output file components to the out port and another output file component to the unused 0 port.

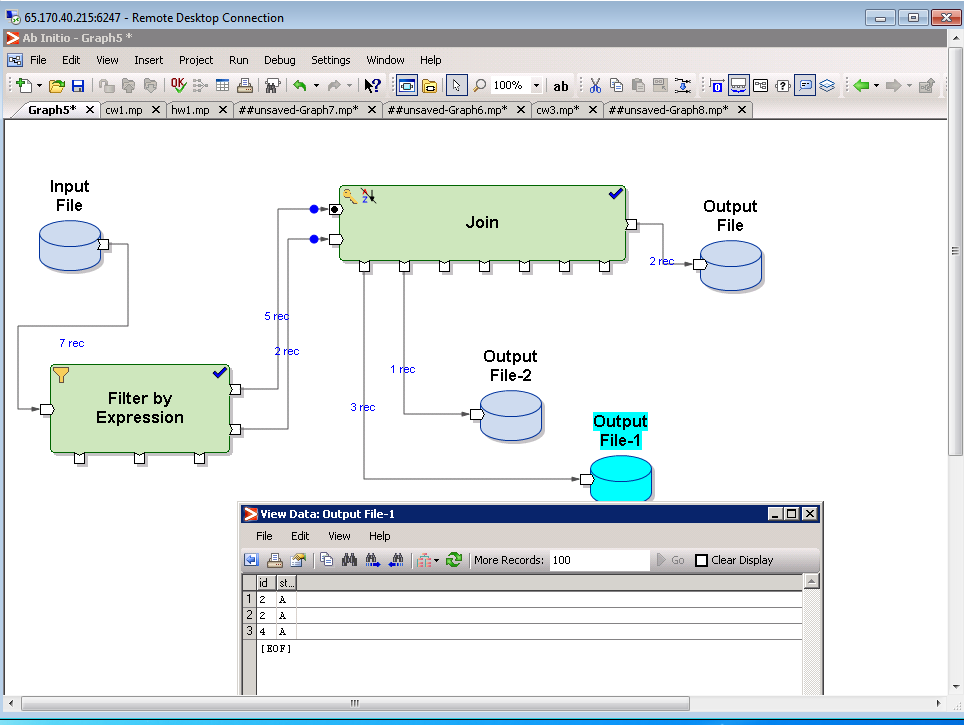


STEP 10: Double click on the JOIN component ,go to the Parameters tab and assign the key value to the ID, sorted-input to False and then double click on transform so that the grid view opens and you can make the joins. Click on validate and Ok.





STEP 11: The Final step is to RUN the graph and verify the output.



Working with Multiple files

Problem Statement: Create a grap to read data from multiple files and store it in the output file.

|  |  |
| --- | --- |
| Input Files | Pune.dat, bglr.dat, Mumbai.dat |
| Output file | All\_dept.dat |
| dmlfile | Filename.dml, file.dml |
| Components | input file, output file, read multiple files. |

Step 1. Create four files, Pune.dat,Bglr.dat, Mumbai.dat and filenames.dat with following data

Pune.dat

1,101,Hinjewadi

2,102,Kalyaninagar

3,101,Talewadi

4,103,Bhosari

Bglr.dat

8,106,whitefield

9,107,electroniccity

Mum.dat

5,104,CGKP

6,105,Vikroli

Filenames.dat

Pune.dat

Bglr.dat

Mum.dat

Step 2. Create below dml files with following definition on Unix server using Putty,

File.dml

record

decimal(“,”) deptid;

decimal(“,”) mgrid;

string(“\n”) location;

end

filename.dml

record

string(“\n”) path

end

Step 3. To verify if dml is correctly created as per file, execute following command on Unix,

$ m\_dump file.dml Mum.dat

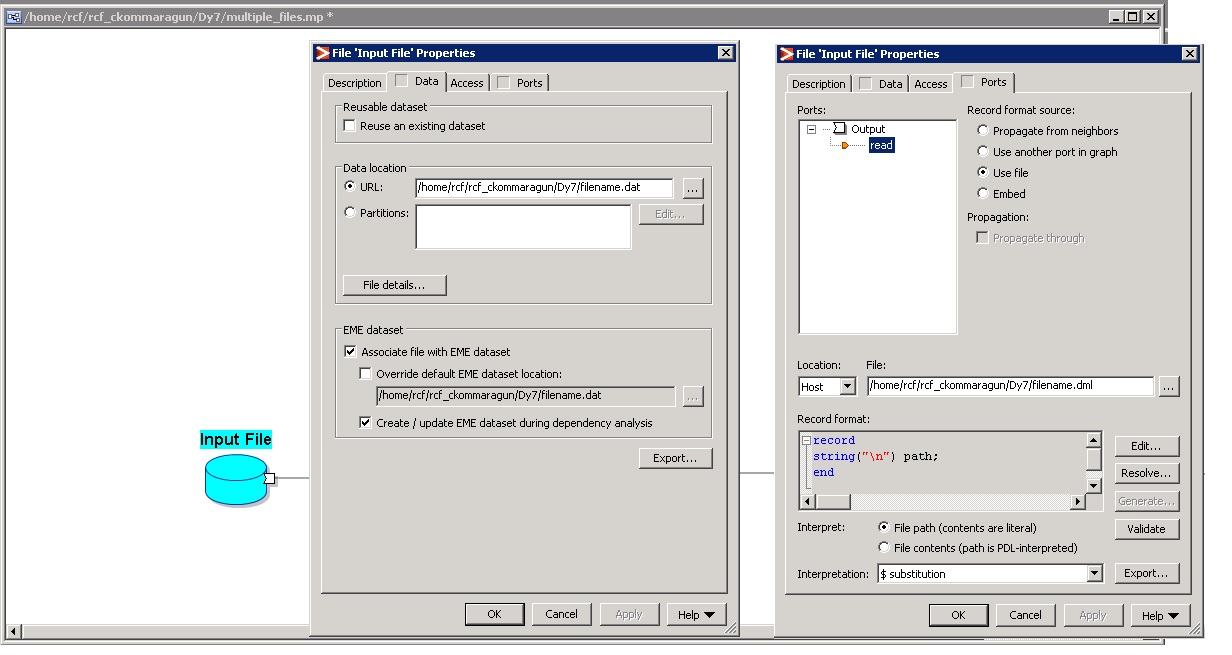
(This would throw error if your definition and data file is not matching)

Step 3. Open Ab Initio GDE. Create a new graph. Save the graph with .mp externsion.

step 4. Drag input file from Datasets components. Perform following settings on it,

* Double click on input file and perform following,
  + Select data tab and provide url of input file which was created on Unix server (input file here would be file with list of file names, I.e filenames.dat).
  + Next, select ports tab and select use file, and provide url of dml file (In this case, filename.dml).

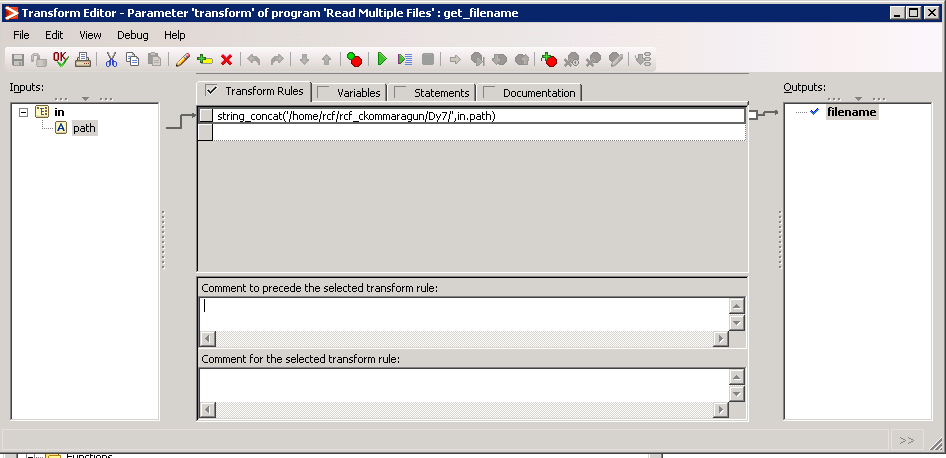
This is shown below.



Step 4: Right click on input file and select, view data->formatted data. Verify if data is propely extracted into input file.

Step 5:Drag “read multiple files” from components and place on graph. Perform the following ,

* Double click on component, Go to parameter tabs and select transform.
* Double click on transform value, you will be asked to select pakage. Select “get\_filename”.
* Drag path from in port to file name, out port. Use string concate with entire path and file name as shown below.



* Select port tab and provide path of file.dml.

step 6: Drag output file from Datasets components. Perform following settings on it,

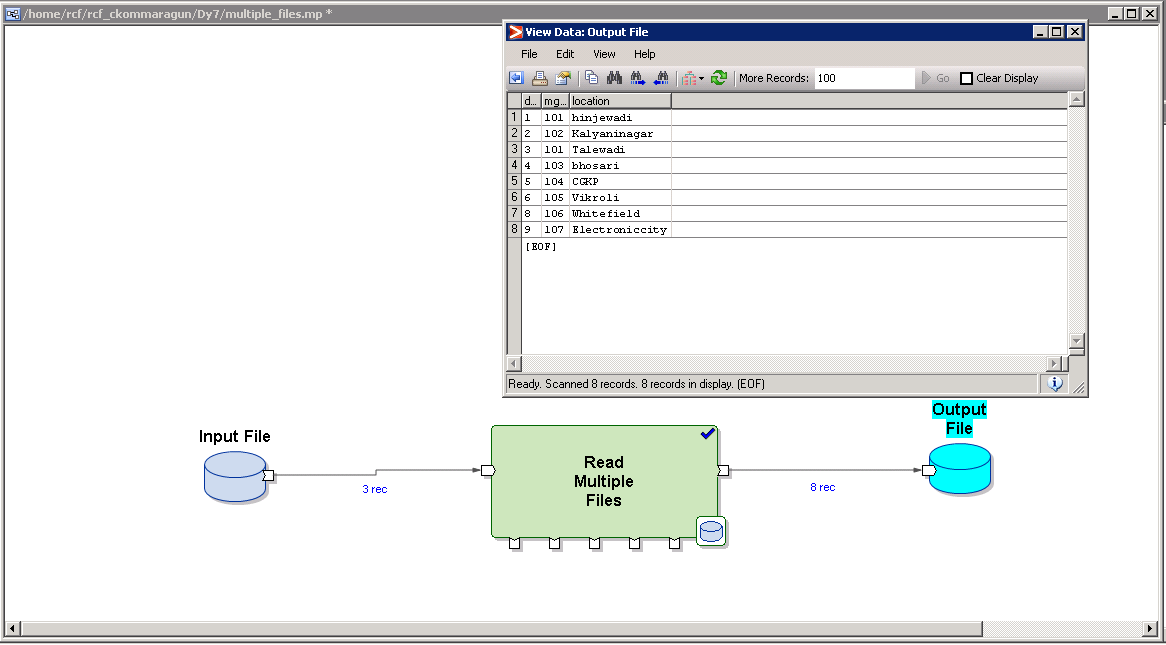
* Double click on output file and perform following,
  + Select data tab and provide url of output file which will be created on Unix server.
* Next, select ports tab and select use file, and provide url of dml file.

(Also , propogate from neighbors can be used or embed can be used.)

Step 6. Connect all the components with flow. And run the Project.

Step 7. Verify the output in output file, by right click and view data->formatted view.

Below is final mapping:



1. **Sort & Dedup**

A feed file from a banking system consists of the various transactions taking place over a period of time.

Process this file as below.

1) Obtain the transactions with the highest transaction amount for each account.

2) Obtain the transactions with the lowest transaction amount for each account.

3) Obtain the accounts which have undergone only a single transaction over this period.

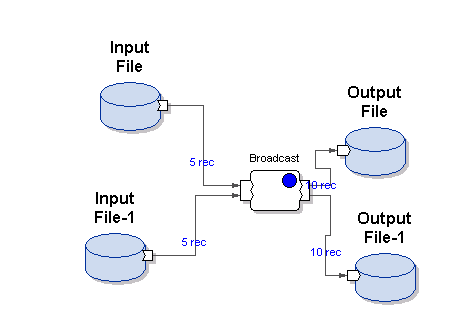
References

1) Input file dml

****

**BROADCAST:**

It combines the record from all input files into a single flow and copies it in each output file .



BROADCAST GRAPH

Steps of implementation -

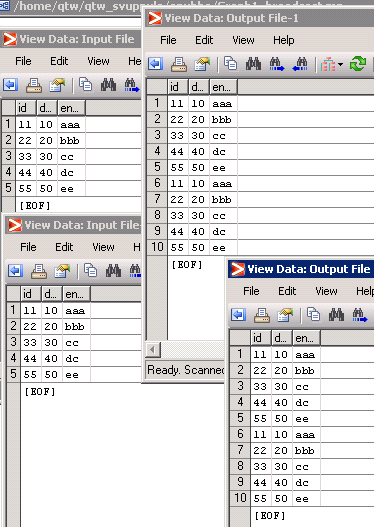
1. Select input file from Datasets Component and set its properties i.e. **Data and Ports.**

2. Select output file from Datasets Component and set its properties i.e. **Data and Ports.**Record format source is Propagate from neighbours.

3. Select Broadcast component from Partitioning.No need to set any property.

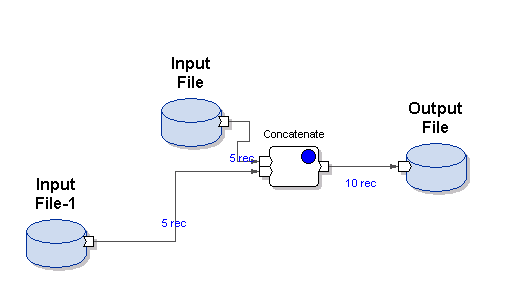
4. Press F5 to execute the graph.

5. View the output through ViewData by using right click on output file.



**CONCATENATE:**

It is used to append the input records one after another.



CONCATENATE GRAPH

Steps of implementation -

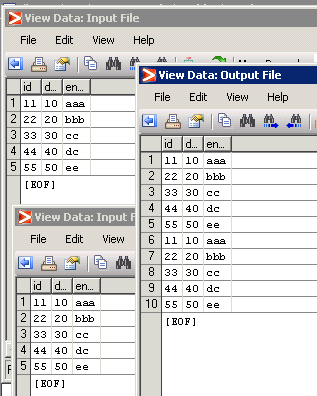
1. Select input file from Datasets Component and set its properties i.e. **Data and Ports.**

2. Select output file from Datasets Component and set its properties i.e. **Data and Ports.**Record format source is Propagate from neighbours.

3. Select Concatenate component from Partitioning.No need to set any property.

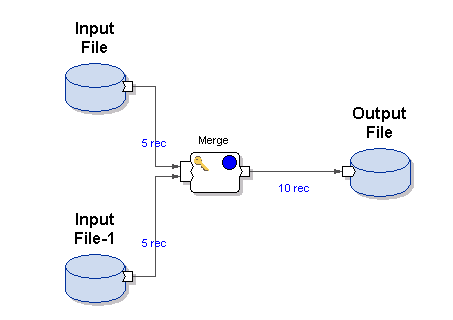
4. Press F5 to execute the graph.

5. View the output through ViewData by using right click on output file.



**MERGE:**

It combines the input record from multiple files and maintains the sort order on the basis of the selected key column.



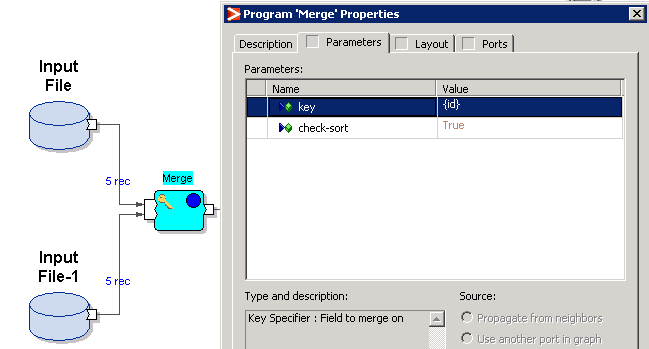
MERGE GRAPH

Steps of implementation -

1. Select 2 input files from Datasets Component and set its properties i.e. **Data and Ports.**

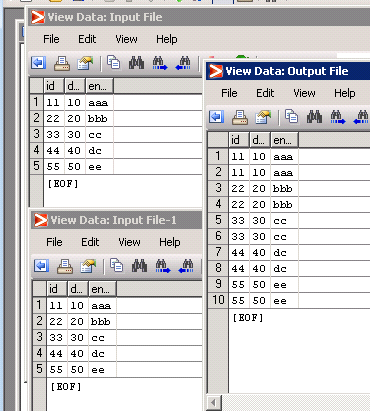
2. Select output file from Datasets Component and set its properties i.e. **Data and Ports.**Record format source is Propagate from neighbours.

3. Select Merge component from Partitioning.Set the key property according to which we want to sort the merged data.



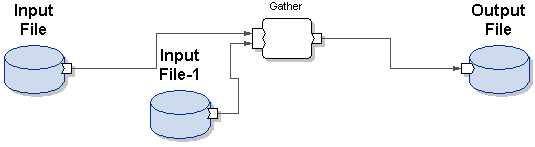
4. Press F5 to execute the graph.

5. View the output through ViewData by using right click on output file.



**GATHER:**

It randomly combines data from multiple input files and send it as a single stream of data to output file.



GATHER GRAPH

Steps of implementation -

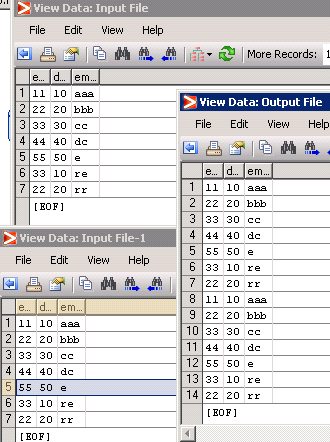
1. Select 2 input files from Datasets Component and set its properties i.e. **Data and Ports** and files should be with same structure and data type.

2. Select output file from Datasets Component and set its properties i.e. **Data and Ports.**Record format source is Propagate from neighbours.

3. Select Gather component from Partitioning.No need to set any property for it.

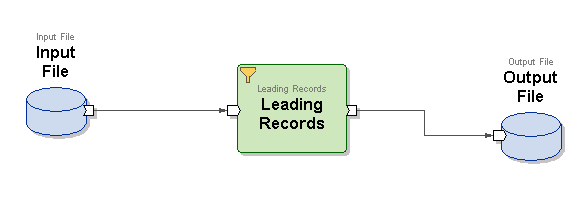
4. Press F5 to execute the graph.

5. View the output through ViewData by using right click on output file.



**LEADING RECORDS:**

It is used to select the required number of records from the input file from the top by specifying the number\_records.



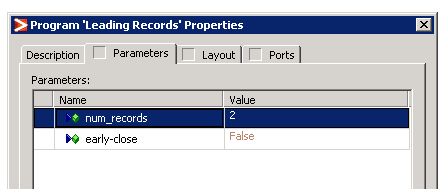
LEADING RECORDS GRAPH

Steps of implementation -

1. Select input file from Datasets Component and set its properties i.e. **Data and Ports.**

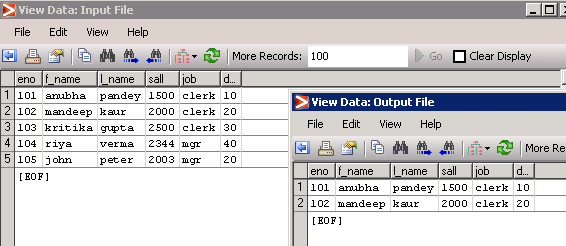
2. Select output file from Datasets Component and set its properties i.e. **Data and Ports.**Record format source is Propagate from neighbours.

3. Select Leading Records component from Miscellaneous folder.We specify the number of records required from top.



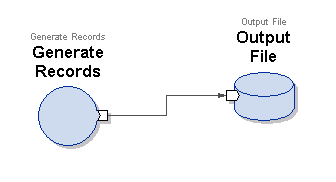
4. Press F5 to execute the graph.

5. View the output through ViewData by using right click on output file.



**GENERATE RECORDS:**

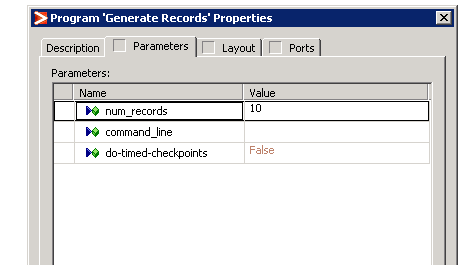
It is used to generate the records for file whose structure is same as that specified in generate component but may contain any random values.



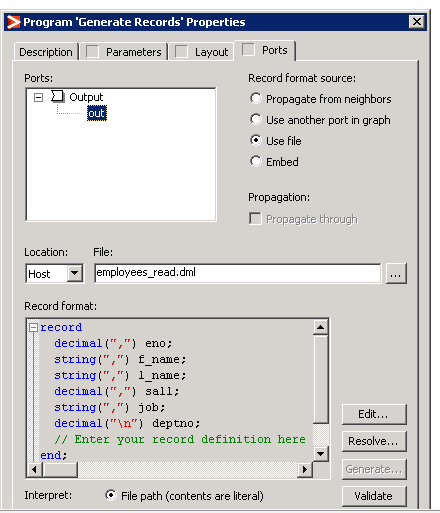
GENERATE RECORDS GRAPH

Steps of implementation -

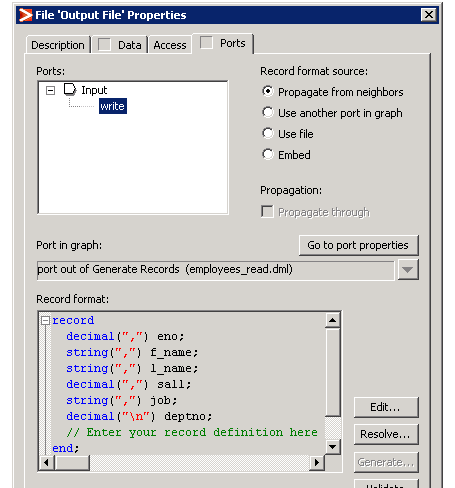
1. Set generate records component from Validate folder and set the number\_records.



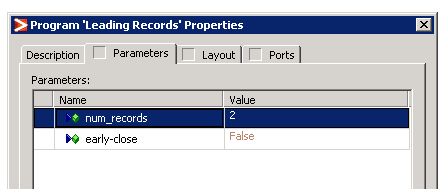
Set the Use file from Ports option to get the structure of the file desired at output file.



2. Select output file from Datasets Component and set its properties i.e. **Data and Ports.**Record format source is Propagate from neighbours.

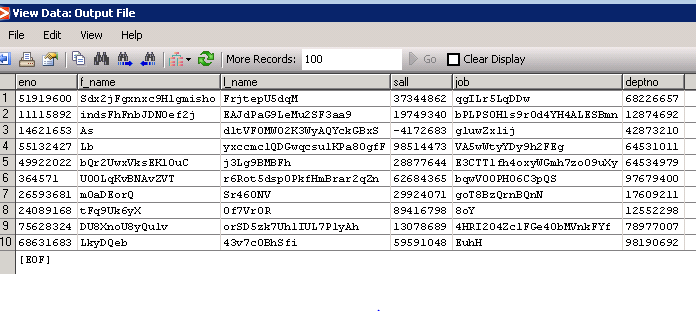


3. Select Leading Records component from Miscellaneous folder.We specify the number of records required from top.



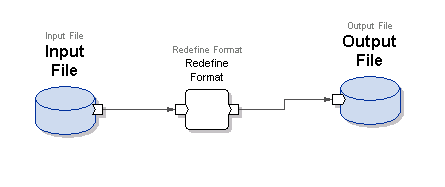
4. Press F5 to execute the graph.

5. View the output through ViewData by using right click on output file.



**REDEFINE FORMAT:**

It is used to copy the fields from in port to out port without changing the values of records but we can rename column names as per the requirement.

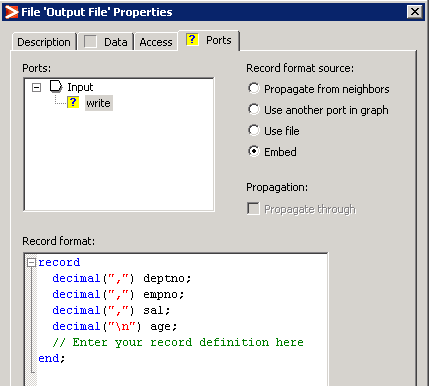


REDEFINE FORMAT GRAPH

Steps of implementation -

1. Select input files from Datasets Component and set its properties i.e. **Data and Ports.**

2. Select output file from Datasets Component and set its properties i.e. **Data and Ports.**Record format source is Embed.



3. Select Redefine component from internal folder of miscelleaneous.

4. Press F5 to execute the graph.

5. View the output through ViewData by using right click on output file.

