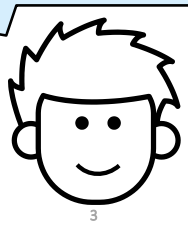
Using Match Code Fields to Join Data

Data Joining with Match Codes

Data Joining

Data jobs have available the Data Joining node (in the Data Integration grouping of nodes) that can be used to join two inputs.

Records from two inputs can be paired if a common primary key value exists for both inputs. If the inputs do not have a common key, two inputs can be joined if one or more field values are the same.



3

sas

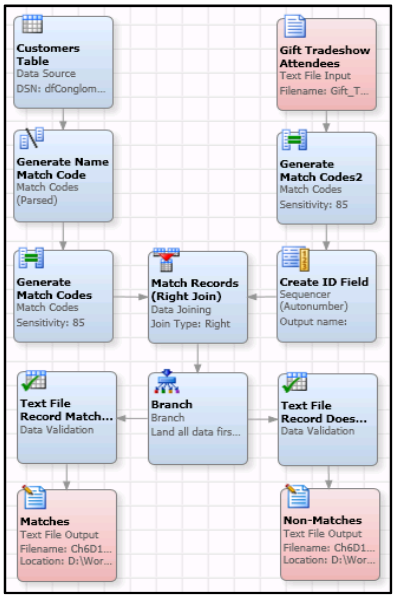
Data needed for a report or a needed analysis are often not all found in a single file. It is common to need to combine information from two or more inputs.

If the inputs have a common key field (or a set of fields that make up a composite key) then the join of the records can be simple.

If the inputs do not have a common key field or fields then a join could occur if one or more field values are the “same”. This “sameness” would have to accommodate for minor differences in the field values. For example, one input may have a company field value of DataFlux, while a second input may have a company field value of DataFlux Inc. Thus, join criteria for two inputs with no common key could use match code fields.

Customer Matches Example

- This data job reads from two inputs (Customers table and Gift_Tradeshow_List).
- The two inputs do not have a common primary key.
- Match codes are created on several fields in the two input tables.
- The generated match codes are used as surrogate keys to join the two inputs.



4

Assume that you have data that was collected at a recent trade show that contains information about potential new customers. Naturally, this data has no way to connect to your existing customer database. You want to identify which attendees visited your tradeshow booth and are existing customers. The problem is that you have only their names and some limited contact information. There is no way to link them to your existing customer database.

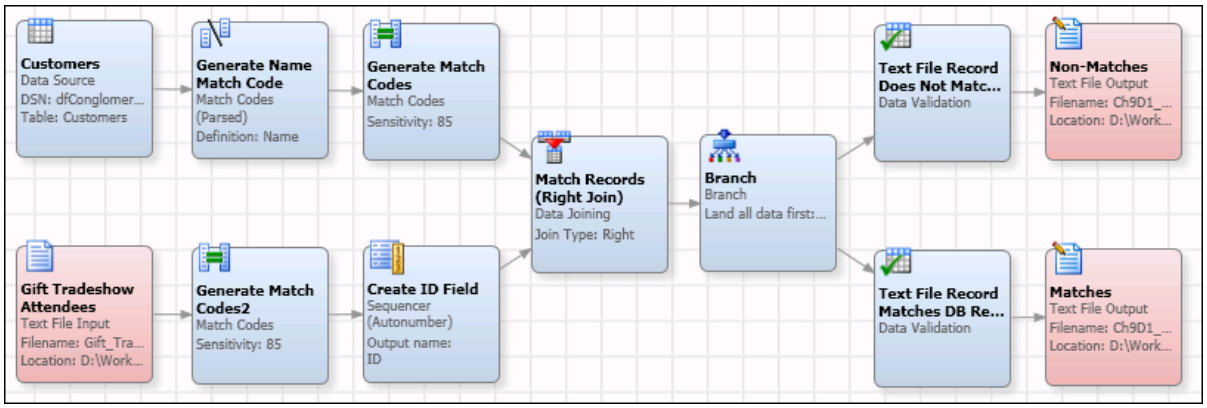
The data job shown demonstrates how to take two inputs without a common key field and use match code fields to join the two inputs together.

Investigating Data Joining: Customer Matches

This demonstration illustrates the steps that are necessary to examine a data job with two inputs and two outputs. The data job generates match codes for a variety of fields from each of the sources. Then the data job performs a data join using conditions that involve the generated match code fields. Matches and non-matches are written to separate text files.

1. If necessary, access Data Management Studio.
   a. Select **Start** ⇨ **All Programs** ⇨ **DataFlux** ⇨ **Data Management Studio 2.7**.
   b. Click **Cancel** to close the Log On window.
2. Open an existing job.
   a. If necessary, click the **Home** tab.
   b. If necessary, click the **Folders** riser bar.
   c. Expand **Basics Solutions**.
   d. Click **batch_jobs**.
   e. Double-click the data job named **Ch9D1_CustomerMatches**.

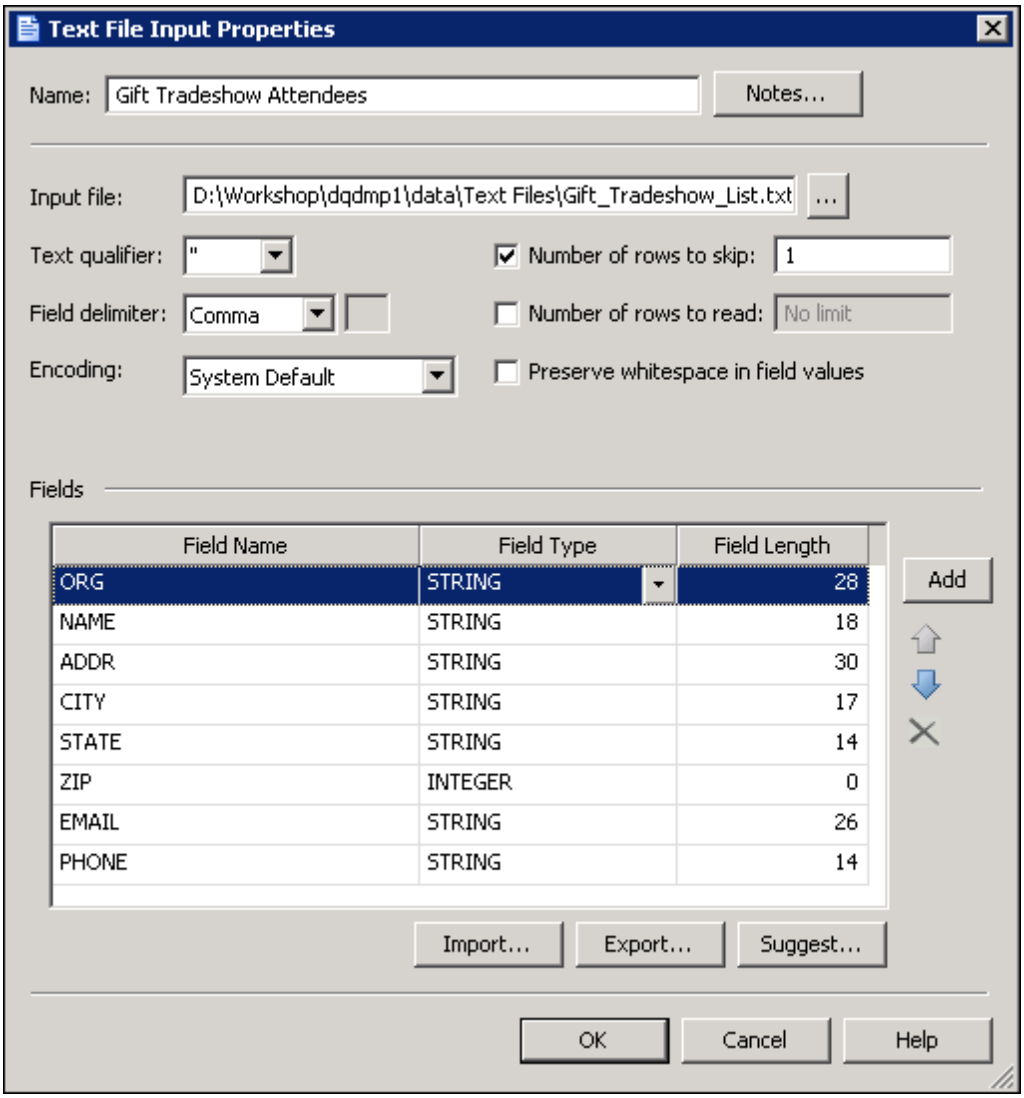
   The data job appears on a new tab. The data job diagram should resemble the following:



Note: The data flow diagram could be vertical. In the above picture, the job diagram is horizontal for display purposes. In addition, the data flow diagram can have *sticky-note* objects that are not displayed in this picture.

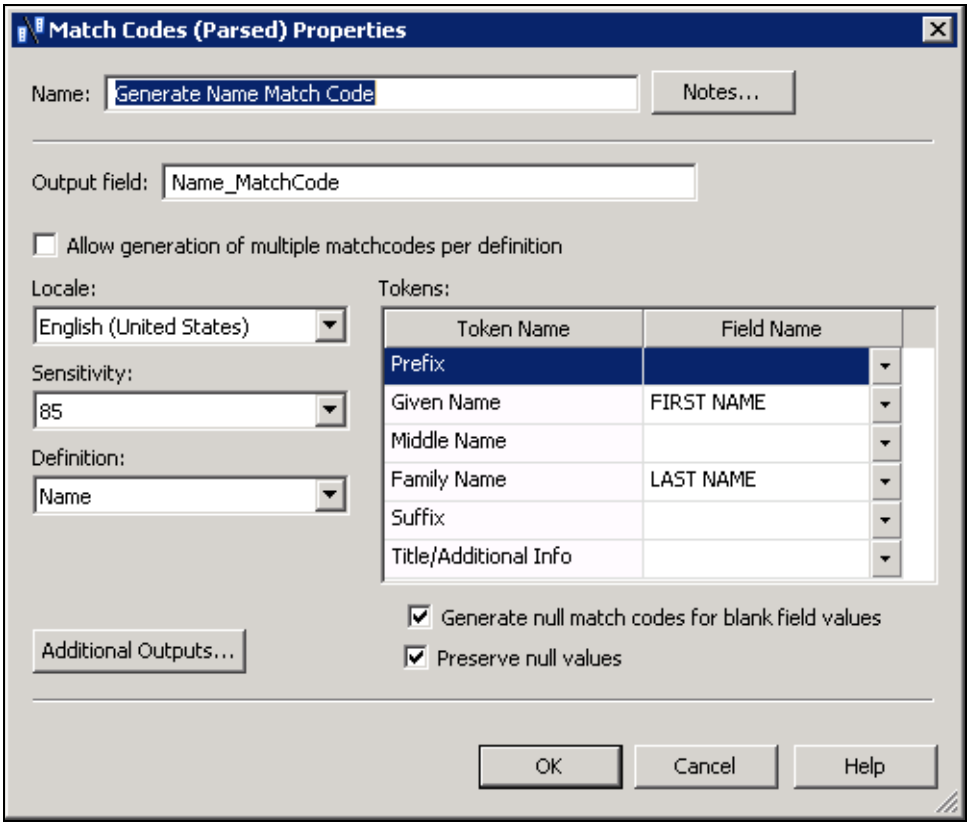
3. Review the properties for the Data Source node.
   a. Right-click the **Data Source** node and select **Properties**.
   b. Verify that the **Input table** field displays the **Customers** table from **dfConglomerate Gifts**.
   c. Verify that all fields from the **Customers** table are selected.
   d. Click **Cancel** to close the Data Source Properties window.

4. Review the properties for the Text File Input node.
   a. Right-click the **Text File Input** node and select **Properties**.
   b. Review the input file specifications.
   c. Verify that eight fields are specified in the Fields area.



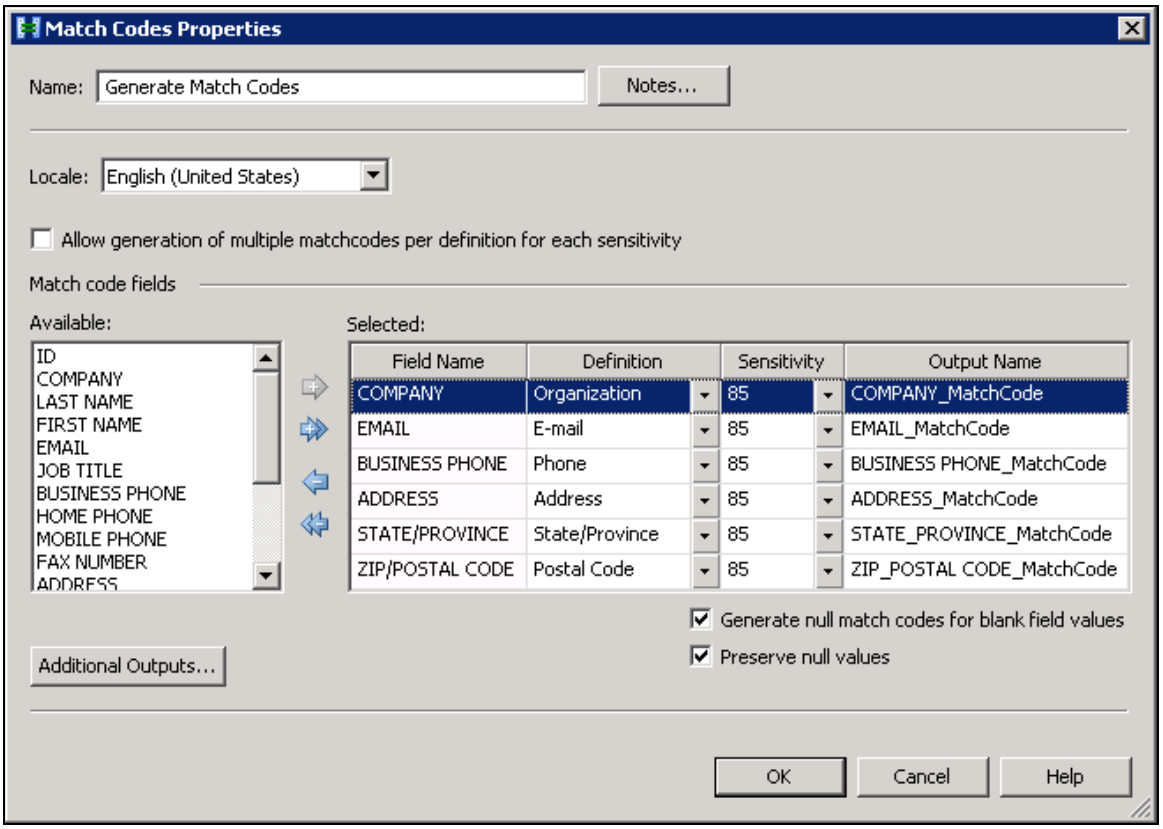
   d. Click **Cancel** to close the Text File Input Properties window.

5. Review the properties for the Match Codes (Parsed) node.
   a. Right-click the **Match Codes (Parsed)** node and select **Properties**.
   b. Verify that **Output field** has a value of **Name_MatchCode**.
   c. Verify that **English (United States)** is selected as the locale.
   d. Verify that **85** is selected as the sensitivity value.
   e. Verify that **Name** is selected as the definition.
   f. Verify that two tokens are used to form the output match code field.
      1) **FIRST NAME** field is paired with the **Given Name** token.
      2) **LAST NAME** field is paired with the **Family Name** token.
   g. Verify that **Generate null match codes for blank field values** is selected.
   h. Verify that **Preserve null values** is selected.



   i. Click **Cancel** to close the Match Codes (Parsed) Properties window.

6. Review the properties for the Match Codes node that follows the Match Codes (Parsed) node.
   a. Right-click the **Match Codes** node (labeled **Generate Match Codes**) and select **Properties**.
   b. Verify that **English (United States)** is selected as the locale.
   c. Verify that match codes are generated for six fields and that appropriate definitions are applied in the Match code fields area.
   d. Verify that **85** is selected as the sensitivity value for all six fields.
   e. Verify that **Generate null match codes for blank field values** is selected.
   f. Verify that **Preserve null values** is selected.



   g. Click **Cancel** to close the Match Codes Properties window.

7. Review the properties for the Match Codes node that follows the Text File Input node.

   a. Right-click the **Match Codes** node (labeled **Generate Match Codes2**) and select **Properties**.

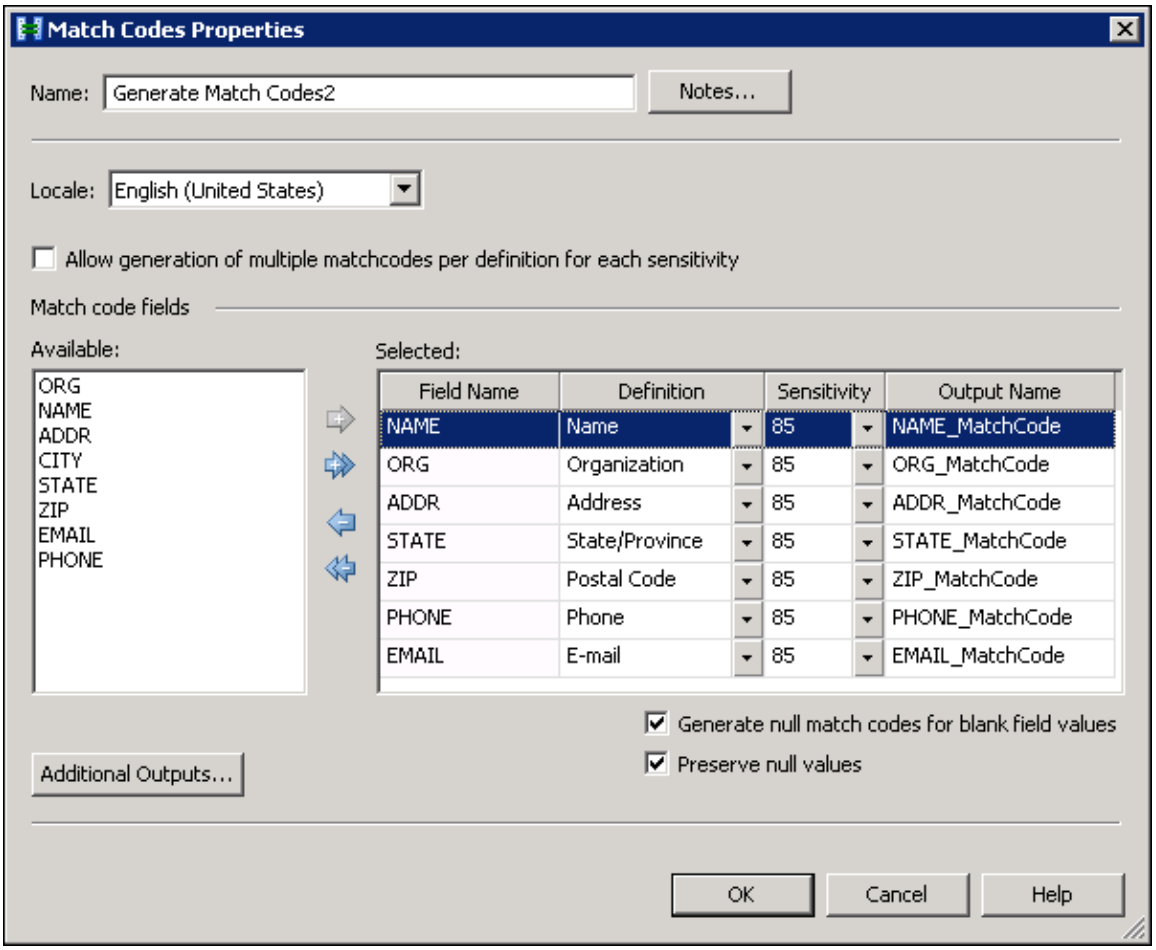
   b. Verify that **English (United States)** is selected as the locale.

   c. Verify that match codes are generated for seven fields and that appropriate definitions are applied in the Match code fields area.

   d. Verify that **85** is selected as the sensitivity value for all seven fields.

   e. Verify that **Generate null match codes for blank field values** is selected.

   f. Verify that **Preserve null values** is selected.



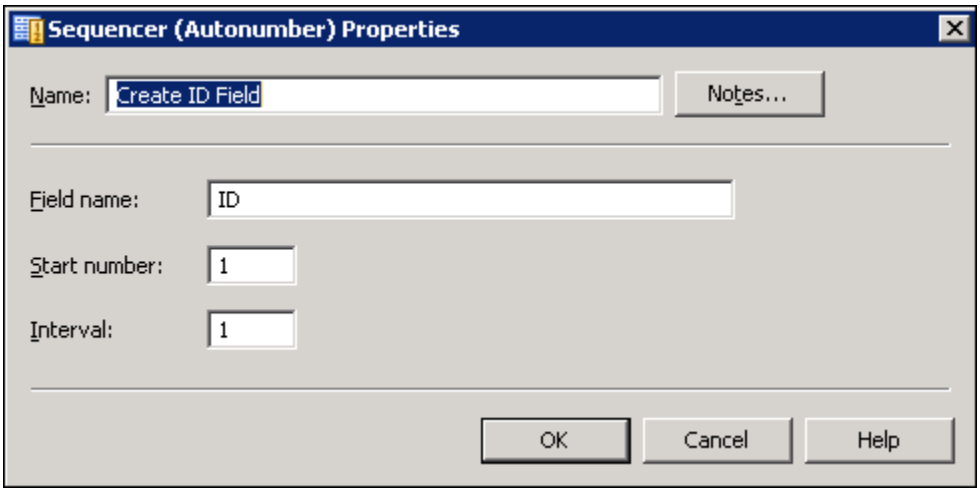
   g. Click **Cancel** to close the Match Codes Properties window.

8. Review the properties for the Sequencer (Autonumber) node.

   a. Right-click the **Sequencer (Autonumber)** node and select **Properties**.

   b. Verify that **ID** is in **Field name** field.

   c. Verify that **1** is in the **Start number** field and that **1** is in the **Interval** field.



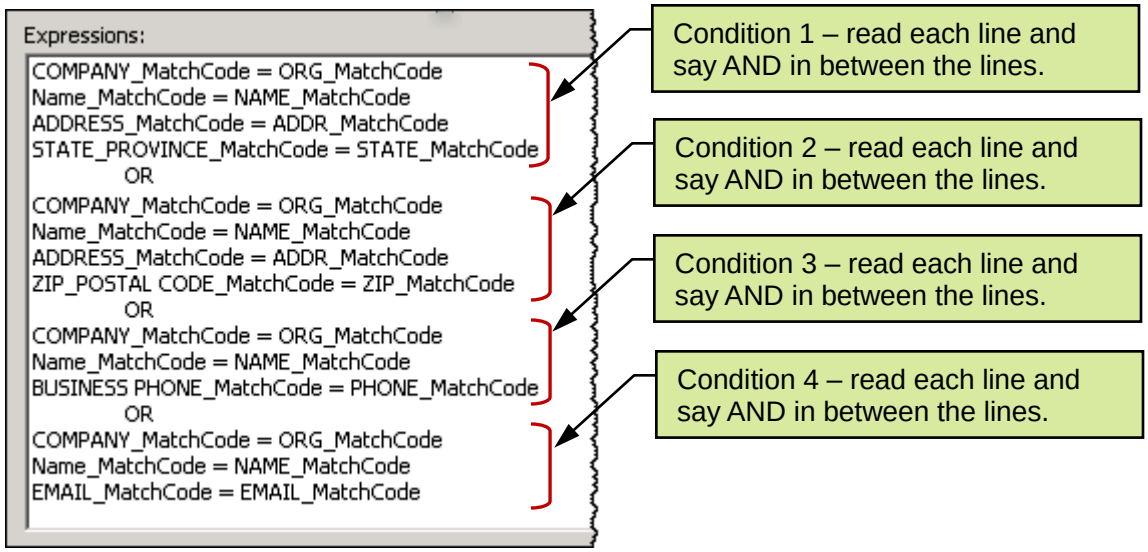
   d. Click **Cancel** to close the Sequencer (Autonumber) Properties window.

9. Review the properties for the Data Joining node.

   a. Right-click the **Data Joining** node and select **Properties**.

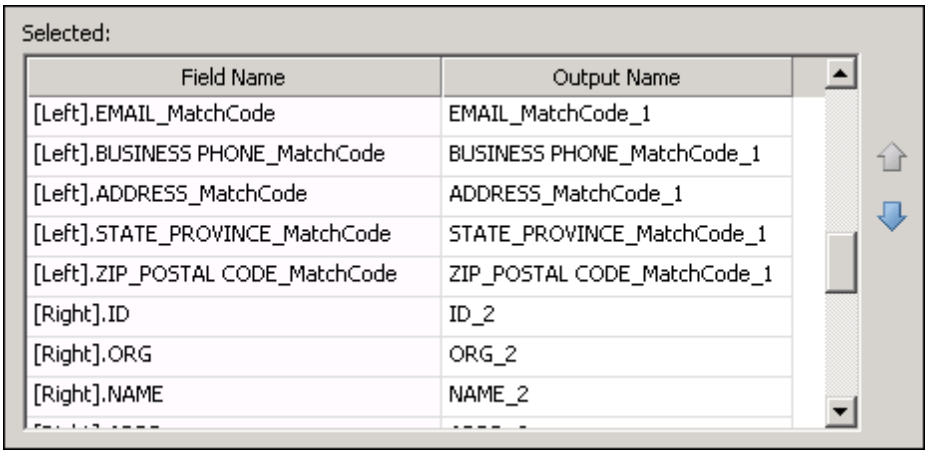
   b. Verify that **Right** is selected for the **Join type** and **Right Table** is selected for the **Memory load** option.

   c. Verify that four join conditions are specified in the Expressions area.



d. In the Output fields area (in the Selected list), verify that the fields from the left table (**Customers**) have output names that end with **_1**.

e. In the Output fields area (in the Selected list), verify that the fields from the right table (the text file) have output names that end with **_2**.



f. Click **Cancel** to close the Data Joining Properties window.

Typically, the output from the Data Joining node would be limited to far fewer fields, and fields selected would be determined based on intended use of the output or result set from this node. In this example we have output all fields from both inputs for illustration purposes.

Recall that the input Customers table has an ID field. Assume that the ID field is a valid primary key field (that is, the ID field has unique and non-null values). The above steps specify that the ID field from the Customers table will be output from the Data Joining node as the field named ID_1.

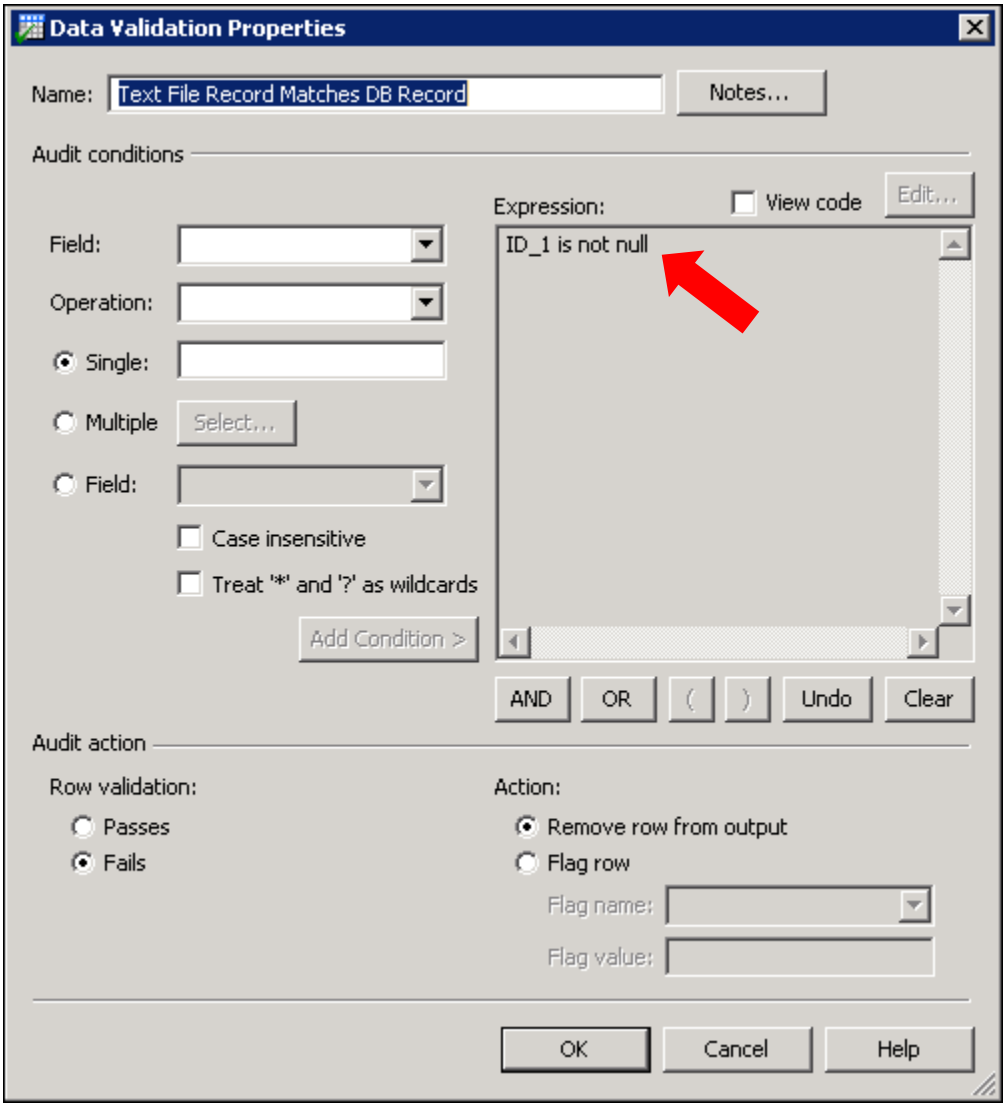
Consider a record from the input text file matching a record from the input Customers table. For all the fields in the Customers table, we will be guaranteed a value will exist for the ID field (since it is a primary key field). Thus, if a match occurs, then the ID_1 field output from the Data Joining node will be non-null.

Consider a record from the input text file ***not*** matching a record from the input Customers table. Since no match occurred then all the output fields with names ending in _1 will be null. And, in particular, ID_1 will be null.

10. Review the properties for the Data Validation node labeled **Text File Record Matches DB Record**.

    a. Right-click the **Data Validation** node and select **Properties**.

    b. Verify that **ID_1 is not null** is in the Expression pane.



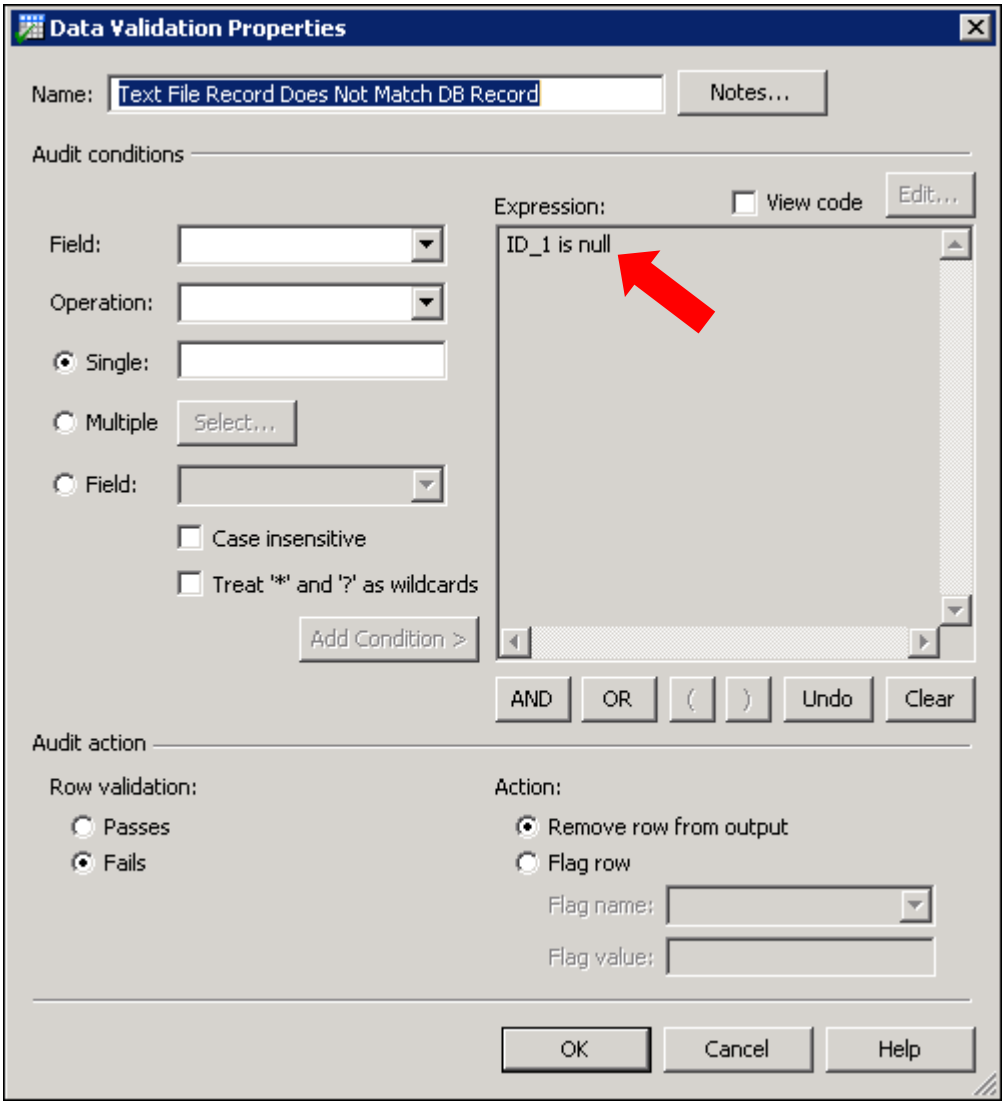
Note: Because this is a right join, if a match occurs (that is, if at least one of the four conditions specified in the Data Joining node is met), then the **ID_1** field has a value.

    c. Click **Cancel** to close the Data Validation Properties window.

11. Review the properties for the Data Validation node labeled **Text File Record Does Not Match DB Record**.

    a. Right-click this **Data Validation** node and select **Properties**.

    b. Verify that **ID_1 is null** is in the Expression pane.



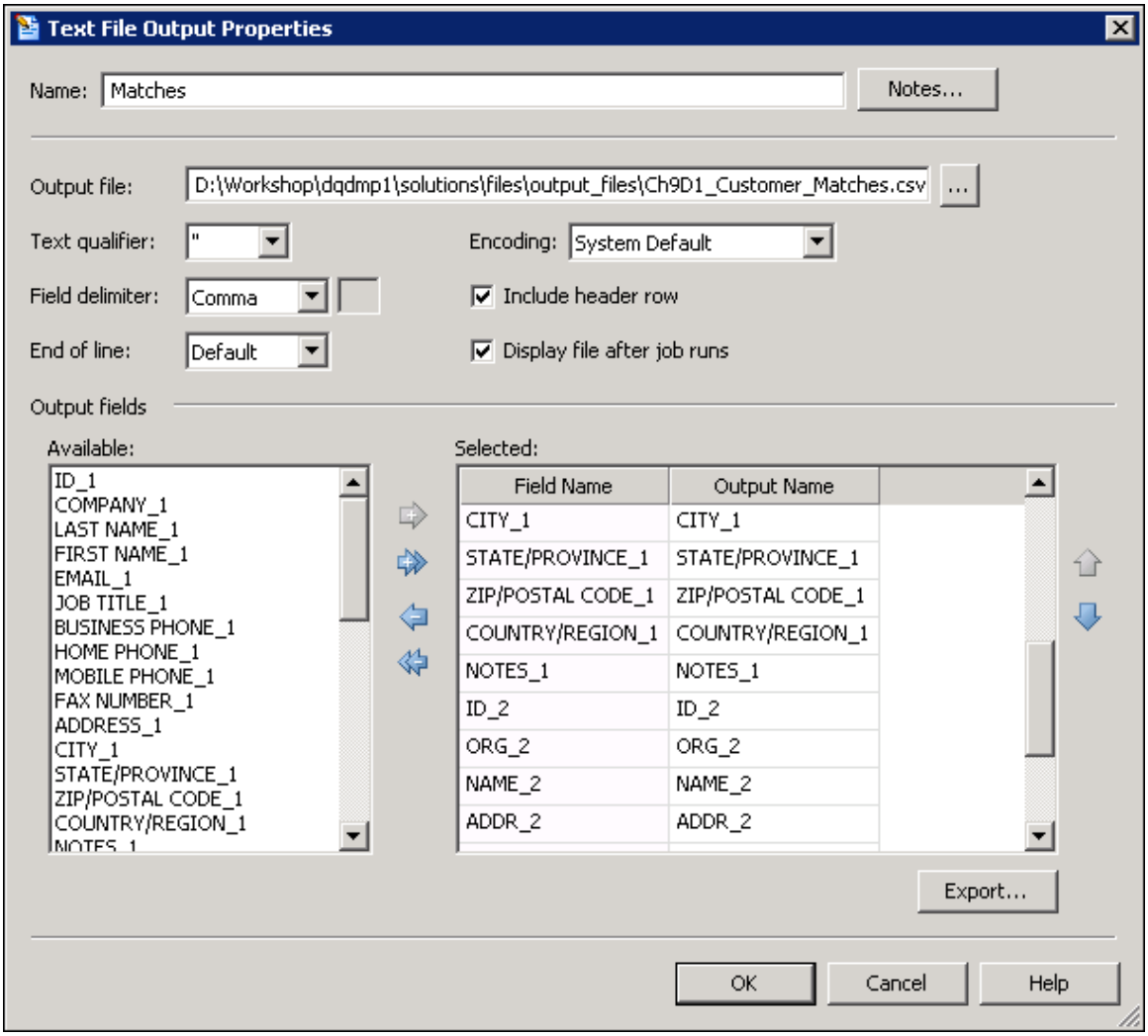
Note: Because this is a right join, if a match does not occur (that is, if at least one of the four conditions specified in the Data Joining node is not met), then the **ID_1** field does not have a value.

    c. Click **Cancel** to close the Data Validation Properties window.

12. Review the properties for the Text File Output node labeled **Matches**.

    a. Right-click the **Text File Output** node and select **Properties**.

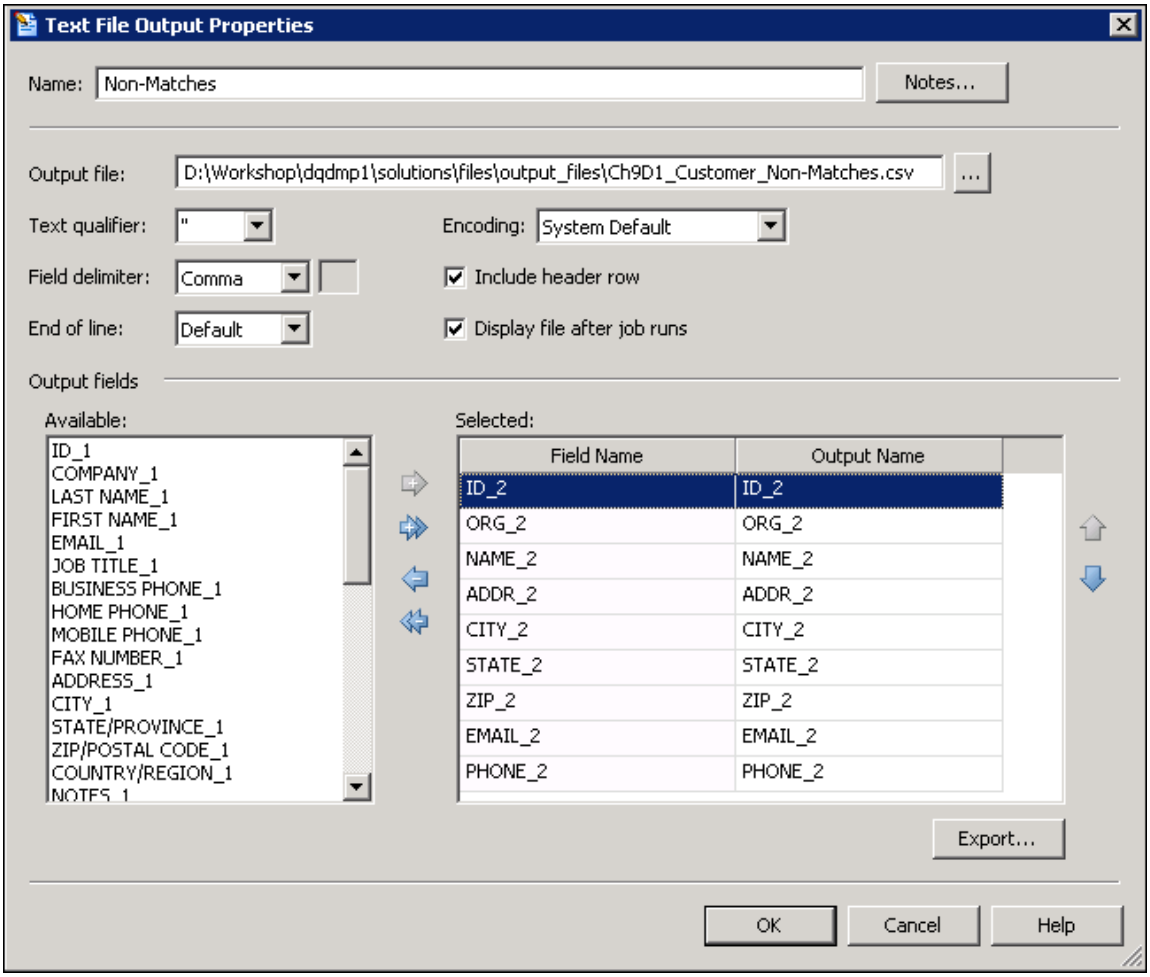
    b. Review the output file specifications.



    c. Verify that all fields are selected for output.

    d. Click **Cancel** to close the Text File Output Properties window.

13. Review the properties for the Text File Output node labeled **Non-Matches**.
    a. Right-click the **Text File Output** node and select **Properties**.
    b. Review the output file specifications.

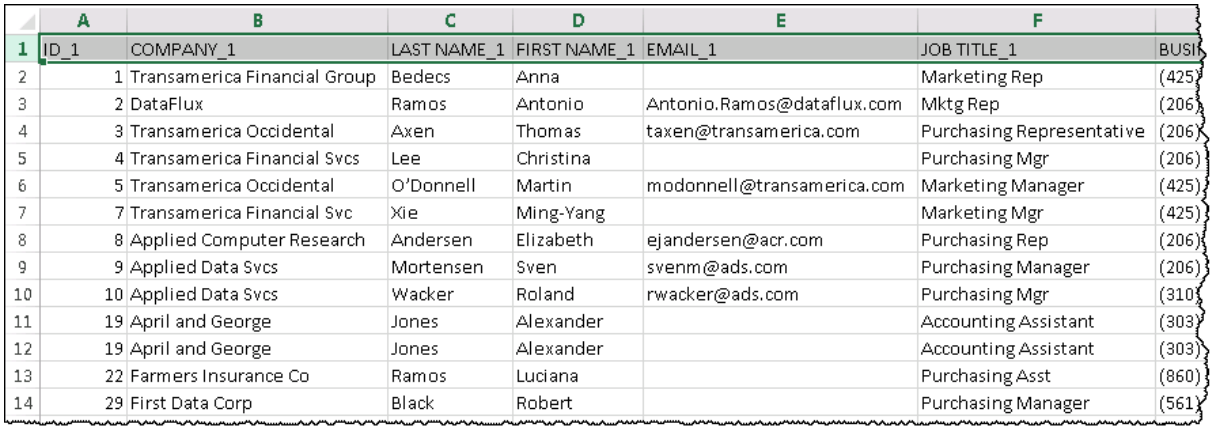


    c. Verify that only the fields with names ending in **_2** are selected for output.
    d. Click **Cancel** to close the Text File Output Properties window.

14. If necessary, select **File** ⇨ **Save** to save the data job.

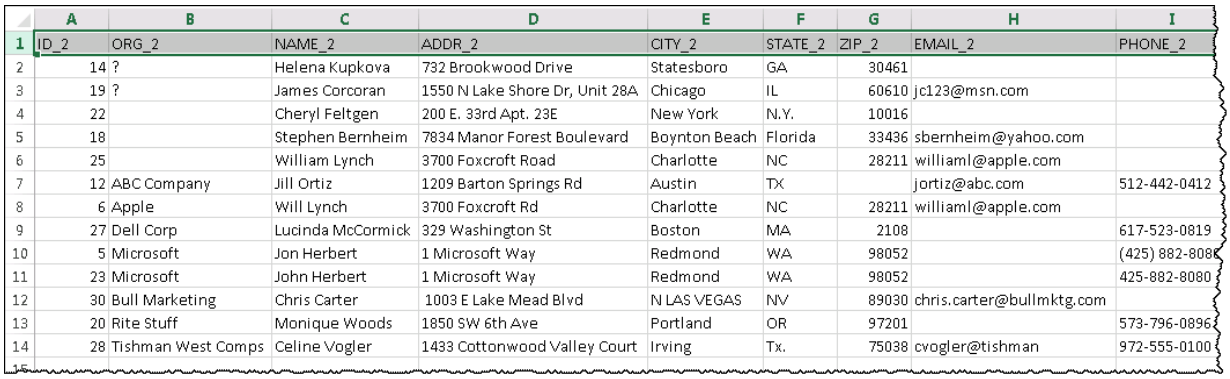
15. Run the job.
    a. Verify that the **Data Flow** tab is selected.
    b. Select **Actions** ⇨ **Run Data Job**.
    c. Verify that the two text files open in Microsoft Excel.

MATCHES:



	A	B	C	D	E	F	
1	ID_1	COMPANY_1	LAST NAME_1	FIRST NAME_1	EMAIL_1	JOB TITLE_1	BUSI
2	1	Transamerica Financial Group	Bedecs	Anna		Marketing Rep	(425)
3	2	DataFlux	Ramos	Antonio	Antonio.Ramos@dataflux.com	Mktg Rep	(206)
4	3	Transamerica Occidental	Axen	Thomas	taxen@transamerica.com	Purchasing Representative	(206)
5	4	Transamerica Financial Svcs	Lee	Christina		Purchasing Mgr	(206)
6	5	Transamerica Occidental	O'Donnell	Martin	modonnell@transamerica.com	Marketing Manager	(425)
7	7	Transamerica Financial Svc	Xie	Ming-Yang		Marketing Mgr	(425)
8	8	Applied Computer Research	Andersen	Elizabeth	ejandersen@acr.com	Purchasing Rep	(206)
9	9	Applied Data Svcs	Mortensen	Sven	svenm@ads.com	Purchasing Manager	(206)
10	10	Applied Data Svcs	Wacker	Roland	rwacker@ads.com	Purchasing Mgr	(310)
11	19	April and George	Jones	Alexander		Accounting Assistant	(303)
12	19	April and George	Jones	Alexander		Accounting Assistant	(303)
13	22	Farmers Insurance Co	Ramos	Luciana		Purchasing Asst	(860)
14	29	First Data Corp	Black	Robert		Purchasing Manager	(561)

NON-MATCHES:

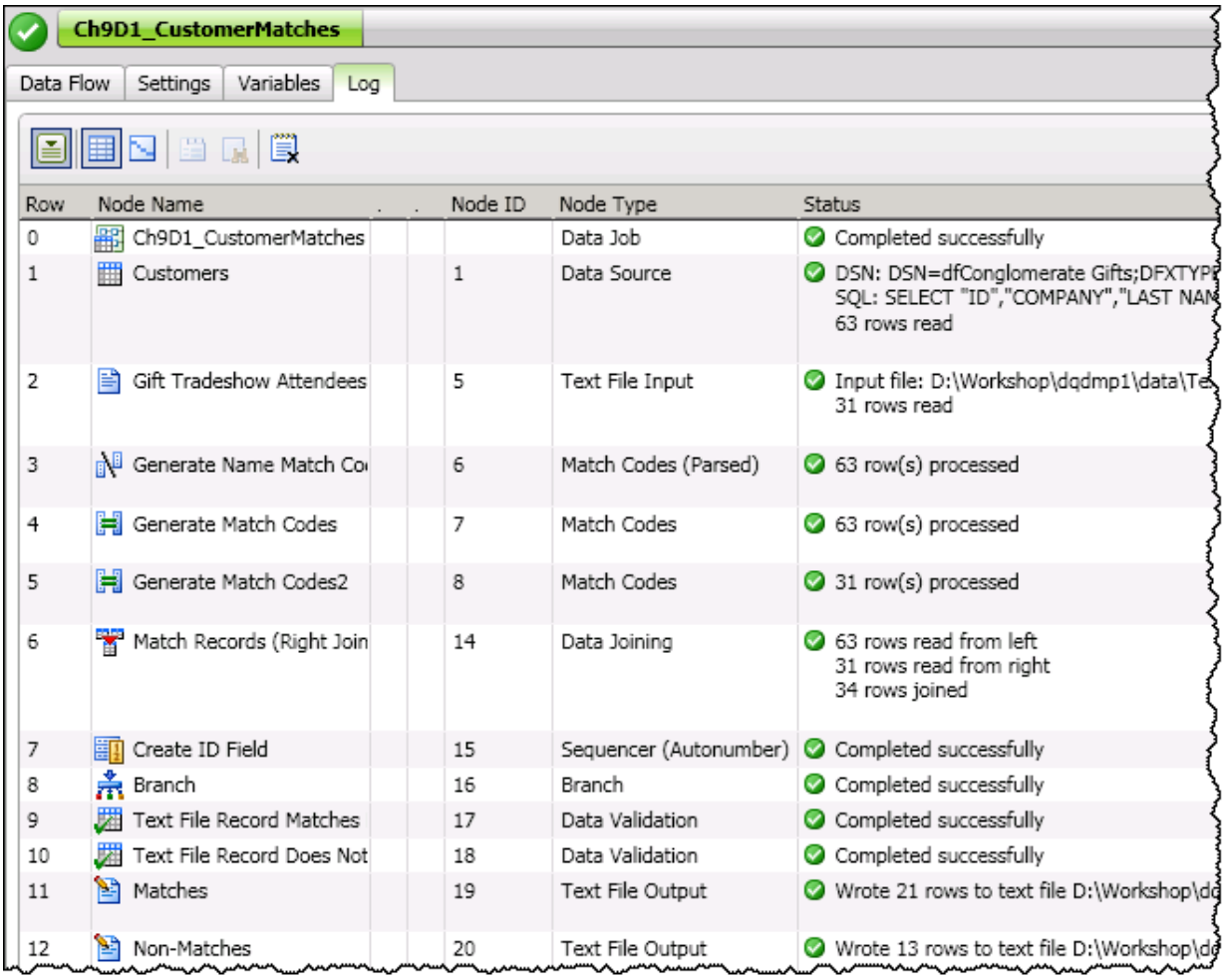


	A	B	C	D	E	F	G	H	I
1	ID_2	ORG_2	NAME_2	ADDR_2	CITY_2	STATE_2	ZIP_2	EMAIL_2	PHONE_2
2	14	?	Helena Kupkova	732 Brookwood Drive	Statesboro	GA	30461		
3	19	?	James Corcoran	1550 N Lake Shore Dr, Unit 28A	Chicago	IL	60610	jc123@msn.com	
4	22		Cheryl Feltgen	200 E. 33rd Apt. 23E	New York	N.Y.	10016		
5	18		Stephen Bernheim	7834 Manor Forest Boulevard	Boynton Beach	Florida	33436	sbernheim@yahoo.com	
6	25		William Lynch	3700 Foxcroft Road	Charlotte	NC	28211	williaml@apple.com	
7	12	ABC Company	Jill Ortiz	1209 Barton Springs Rd	Austin	TX		jortiz@abc.com	512-442-0412
8	6	Apple	Will Lynch	3700 Foxcroft Rd	Charlotte	NC	28211	williaml@apple.com	
9	27	Dell Corp	Lucinda McCormick	329 Washington St	Boston	MA	2108		617-523-0819
10	5	Microsoft	Jon Herbert	1 Microsoft Way	Redmond	WA	98052		(425) 882-8080
11	23	Microsoft	John Herbert	1 Microsoft Way	Redmond	WA	98052		425-882-8080
12	30	Bull Marketing	Chris Carter	1003 E Lake Mead Blvd	N LAS VEGAS	NV	89030	chris.carter@bullmktg.com	
13	20	Rite Stuff	Monique Woods	1850 SW 6th Ave	Portland	OR	97201		573-796-0896
14	28	Tishman West Comps	Celine Vogler	1433 Cottonwood Valley Court	Irving	Tx.	75038	cvogler@tishman	972-555-0100

d. Select **File** ⇨ **Close** to close both text files. Do not save any changes.

16. View the detailed log.

    a. Click the **Log** tab.



Row	Node Name	Node ID	Node Type	Status
0	Ch9D1_CustomerMatches		Data Job	Completed successfully
1	Customers	1	Data Source	DSN: DSN=dfConglomerate Gifts;DFXTYP SQL: SELECT "ID","COMPANY","LAST NAM 63 rows read
2	Gift Tradeshow Attendees	5	Text File Input	Input file: D:\Workshop\dqdmp1\data\Te 31 rows read
3	Generate Name Match Co	6	Match Codes (Parsed)	63 row(s) processed
4	Generate Match Codes	7	Match Codes	63 row(s) processed
5	Generate Match Codes2	8	Match Codes	31 row(s) processed
6	Match Records (Right Join	14	Data Joining	63 rows read from left 31 rows read from right 34 rows joined
7	Create ID Field	15	Sequencer (Autonumber)	Completed successfully
8	Branch	16	Branch	Completed successfully
9	Text File Record Matches	17	Data Validation	Completed successfully
10	Text File Record Does Not	18	Data Validation	Completed successfully
11	Matches	19	Text File Output	Wrote 21 rows to text file D:\Workshop\dq
12	Non-Matches	20	Text File Output	Wrote 13 rows to text file D:\Workshop\dq

17. Close the data job.

    a. Click the **Data Flow** tab.

    b. Select **File** ⇨ **Close**.

End of Demonstration