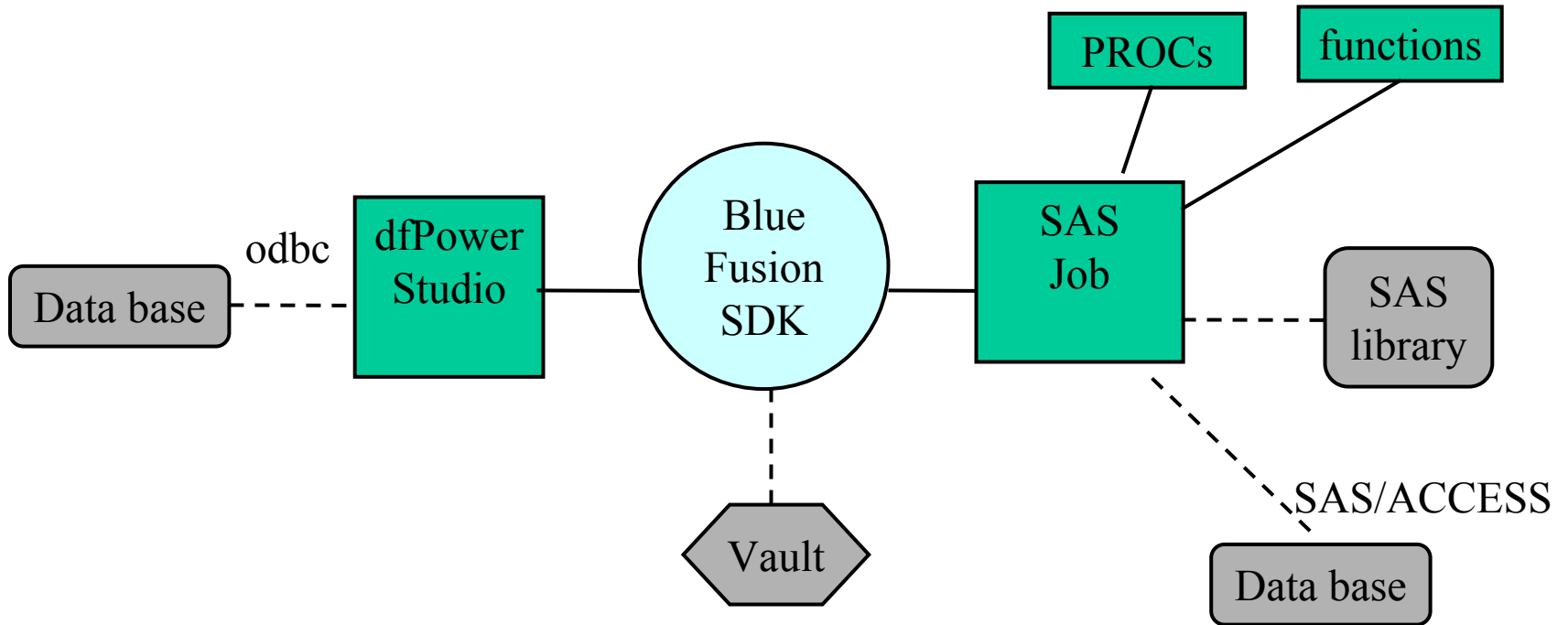


# Do you have valid, consistent and accurate data? Consider a data quality solution.



Bill Fehlner, Education, SAS  
416 307-4513 bill.fehlner@sas.com

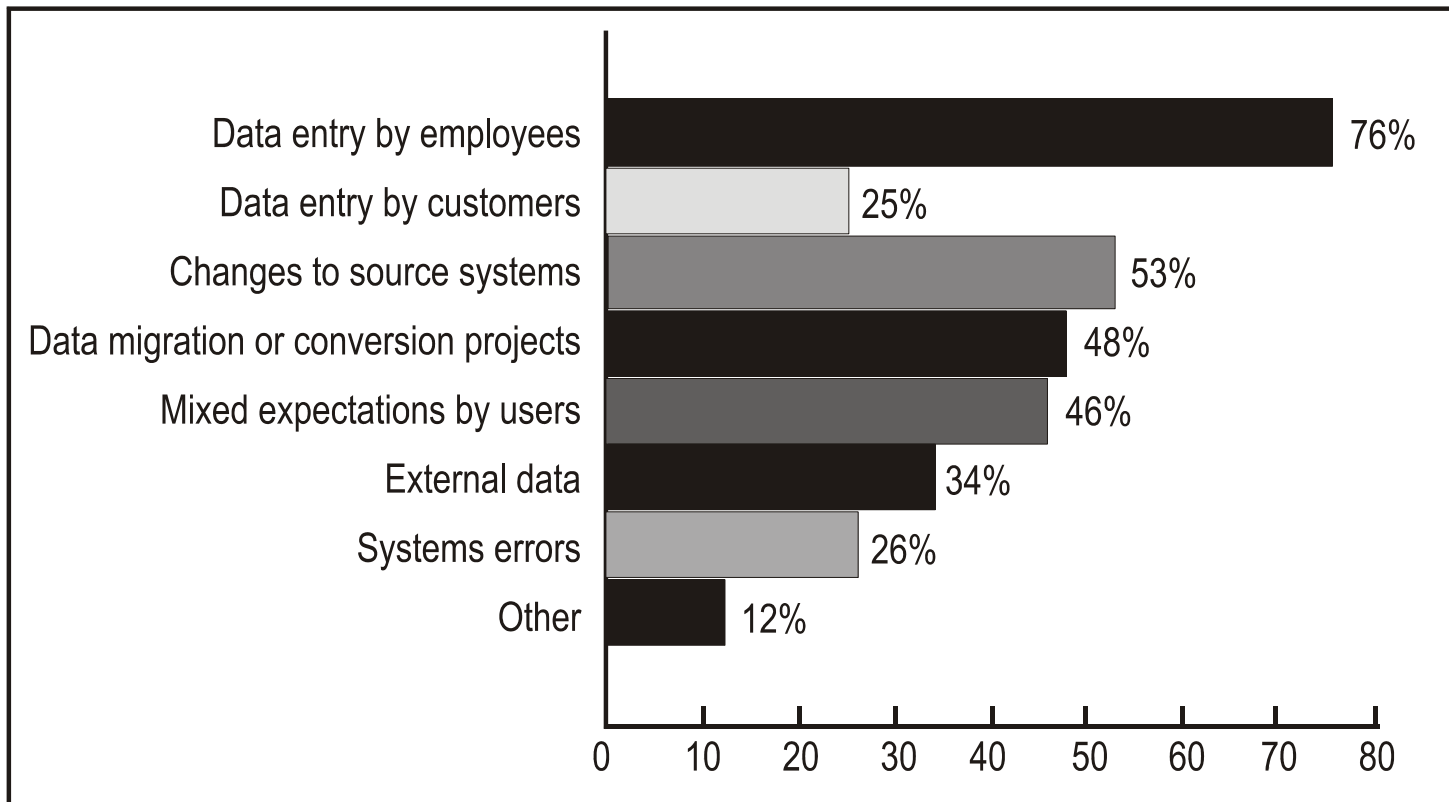
# Agenda

- Why worry about data quality?
- Data samples before and after cleaning.
- Interactive Data Cleaning
- Data Cleaning with programs
- Where to learn more

# Why Worry about Data Quality?

- META Group – Ten to twenty percent of the data used to build data warehouses is corrupt or incomplete.
- Data Warehousing Institute – estimates that data quality problems cost US corporations more than \$600 billion per year (2002).

# Primary Sources of Data Quality Problems



- Source: *The Data Warehousing Institute, Data Quality and the Bottom Line, 2002*

# Data Quality Characteristics

- Data quality affects several attributes associated with data:
  - **Accuracy** – Is it realistic or believable?
  - **Consistency** – Is it consistently defined and maintained?
  - **Validity** – Is the data valid, based on business or industry rules and standards?

# Agenda

- Why worry about data quality?
- **Data samples before and after cleaning.**
- Interactive Data Cleaning
- Data Cleaning with programs
- Where to learn more

# Definitely Inconsistent, and not that Accurate

	Customer Name	Address
	William Jaackson	515 E Broad St., Suite 12
	Bill Jackson	515 E. Broad St., Suite 12
	Bill F. Jackson	515 E. Broad St., Suite 12
	Billy Jackson	515 E. Broad St., Ste. 12
	William P Jackson	515 E. Broad St., Ste 12
	Mr. William Jackson	515 E. Broad St., Suite 12
	William Jackson	515 E. Broad St., Suite 12
	Mr Bill Preston Jackson	515 E Broad St., Suite 12
	William P. Jackson	515 East Broad St., Suite 12
	Mr. William Jacksonn	515 E. Broad St., Suite 12
	Will Jackson	515 E. Broad St., Suite 12
	Mr. William P Jackson	515 E. Broad St., Suite 12
	BILLY JACKSON	515 East Broad St., Suite 12
	Mr. Bill Jackson	515 East Broad Street Suite 12
	Mr. William Jackson	515 E. Broad St., Suite 12
	bill jackson	515 e broad st., suite 12
	Bill Jackson	515 E. Broad St., Suite 12
	Bill Jackson	515 E. Broad St., Suite 12

# The issues here

- Inconsistent use of name prefixes
- Inconsistent capitalization
- Use of nicknames for given names
- Misspelling of last names
- Occasional use of middle name

# Accurate and Consistent

	standardname	Customer Name	Address
123	Mr. William Jackson	William Jaackson	515 E Broad St., Suite 12
124	Mr. William Jackson	Bill Jackson	515 E. Broad St., Suite 12
125	Mr. William Jackson	Bill F. Jackson	515 E. Broad St., Suite 12
126	Mr. William Jackson	Billy Jackson	515 E. Broad St., Ste. 12
127	Mr. William Jackson	William P Jackson	515 E. Broad St., Ste 12
128	Mr. William Jackson	Mr. William Jackson	515 E. Broad St., Suite 12
129	Mr. William Jackson	William Jackson	515 E. Broad St., Suite 12
130	Mr. William Jackson	Mr Bill Preston Jackson	515 E Broad St., Suite 12
131	Mr. William Jackson	William P. Jackson	515 East Broad St., Suite 12
132	Mr. William Jackson	Mr. William Jacksonn	515 E. Broad St., Suite 12
133	Mr. William Jackson	Will Jackson	515 E. Broad St., Suite 12
134	Mr. William Jackson	Mr. William P Jackson	515 E. Broad St., Suite 12
135	Mr. William Jackson	BILLY JACKSON	515 East Broad St., Suite 12
136	Mr. William Jackson	Mr. Bill Jackson	515 East Broad Street Suite 12
137	Mr. William Jackson	Mr. William Jackson	515 E. Broad St., Suite 12
138	Mr. William Jackson	bill jackson	515 e broad st., suite 12
139	Mr. William Jackson	Bill Jackson	515 E. Broad St., Suite 12

# When name processing is not enough

	standardname	Customer Name	Address
165	Mr. Donnie Williams	Mr. Donnie Williams	307 Boltstone Ct
166	Mr. Donnie Williams	MR DON F WILLIAMS	25670 W HEDGEWOOD DRIVE
167	Mr. Donnie Williams	Don Williams	6512 Six Forks Rd Suite 404B
168	Mr. Donnie Williams	DON WILLIAMS	25670 W HEDGEWOOD DR
169	Mr. Donnie Williams	DONALD F. WILLIAMS	25670 WEST HEDGEWOOD DR
170	Mr. Donnie Williams	DON WILIAMS	25670 W HEDGE WOOD
171	Mr. Donnie Williams	DONALD WILLIAMS	25670 W HEDGE WOOD DR.
172	Mr. Donnie Williams	DONNY WILLIAMS	25670 W. HEDGEWOOD DR.
173	Sheryl Wellman	Sheryl Wellman	PO Box 3887

# The issues here

- Inconsistent use of name prefixes
- Inconsistent capitalization
- Use of nicknames for given names
- Misspelling of last names
- Occasional use of middle name
- Some names apparently the same have different addresses

# More issues with addresses

- Inconsistent use of punctuation
- Inconsistent reference to direction
- Inconsistent use of street extensions
- Misspelling of street names

# Joint processing of name and address

	standardname	Customer Name	Address
46	Mr. William Jackson	William Jaackson	515 E Broad St., Suite 12
47	MR DON F WILLIAMS	MR DON F WILLIAMS	25670 W HEDGEWOOD DRIVE
48	MR DON F WILLIAMS	DON WILLIAMS	25670 W HEDGEWOOD DR
49	MR DON F WILLIAMS	DONALD F. WILLIAMS	25670 WEST HEDGEWOOD DR
50	MR DON F WILLIAMS	DON WILIAMS	25670 W HEDGE WOOD
51	MR DON F WILLIAMS	DONALD WILLIAMS	25670 W HEDGE WOOD DR.
52	MR DON F WILLIAMS	DONNY WILLIAMS	25670 W. HEDGEWOOD DR.
53	Rob Beckett	Rob Beckett	392 S. Main St. PO Box 2270
54	Rob Beckett	Rob Beckett	392 S. Main St. PO Box 2270

# Common Processes in a Data Quality Initiative

- Consistency Analysis
- Standardization Schemes
- Gender Analysis
- Entity Analysis
- Data Parsing
- Data Casing

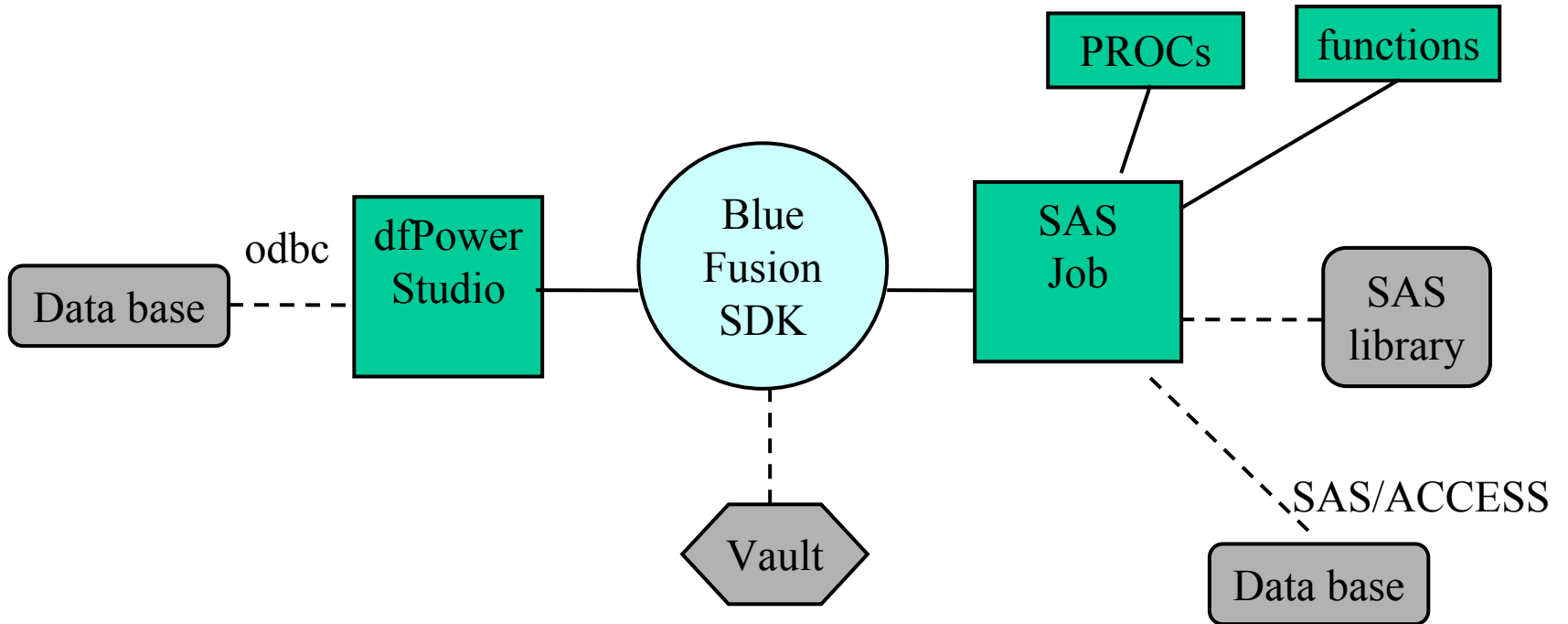
# Tasks simplified by a Data Quality Initiative

- Matching rows in multiple tables.
- De-duplication of rows in one table.
- Address Verification.

# Agenda

- Why worry about data quality?
- Data samples before and after cleaning.
- **Interactive Data Cleaning**
- Data Cleaning with programs
- Where to learn more

# DataFlux Functionality



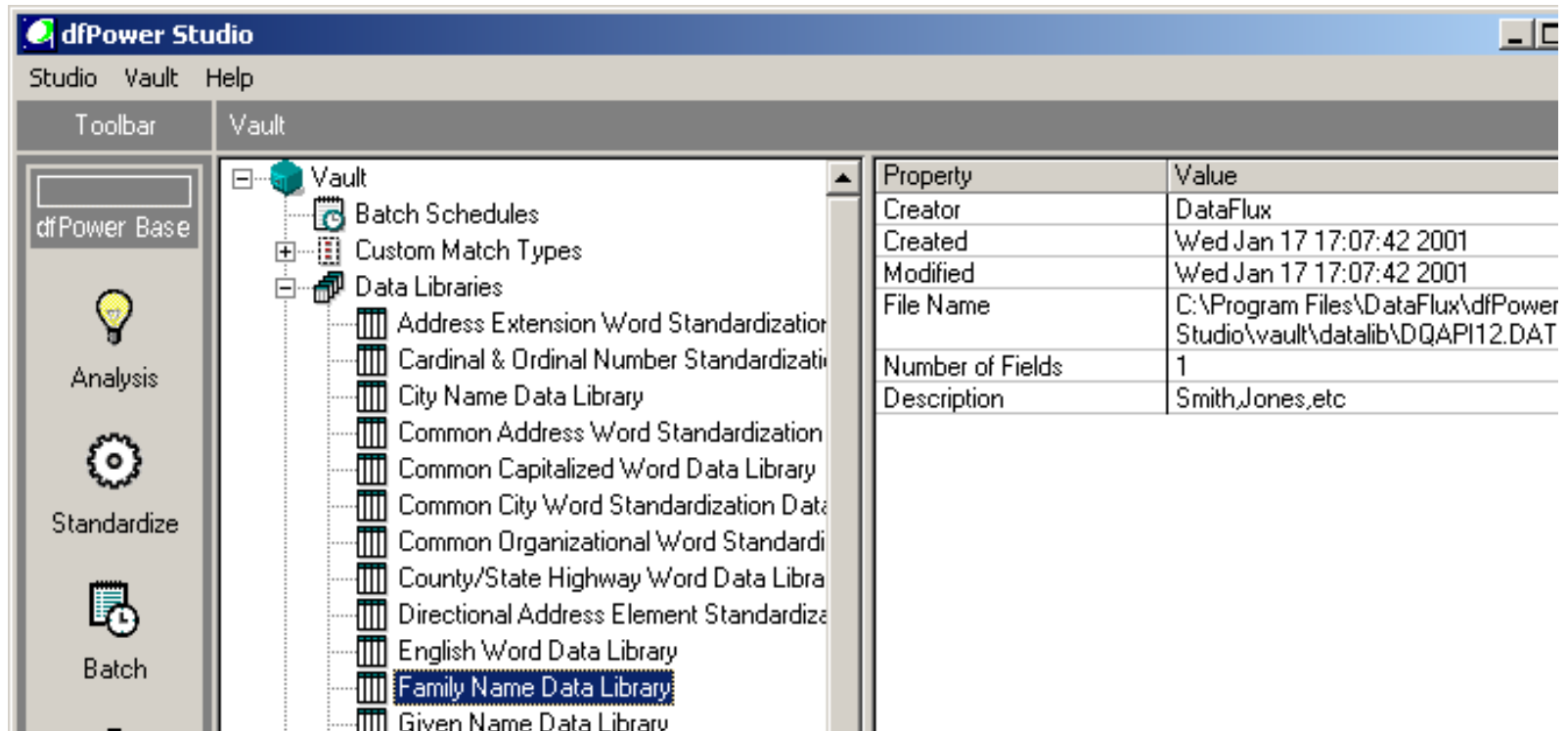
# dfPower Studio's Vault

The screenshot displays the dfPower Studio application window. The title bar reads "dfPower Studio" and the menu bar includes "Studio", "Vault", and "Help". The interface is divided into three main sections:

- Toolbar:** Located on the left, it contains a "dfPower Base" button, an "Analysis" button with a lightbulb icon, and a "Standardize" button with a gear icon.
- Vault Tree:** A hierarchical tree view on the right side of the toolbar. The root node is "Vault", which is expanded to show several sub-nodes: "Batch Schedules", "Custom Match Types", "Data Libraries" (highlighted with a blue selection box), "Jobs", "References", "Reports", and "Standardization Schemes".
- Properties Table:** A table on the far right, titled "Property" and "Value", showing details for the selected "Data Libraries" node.

Property	Value
Creator	DataFlux
Children Type	Data Library
Number of Children	22

# dfPower Studio's Vault



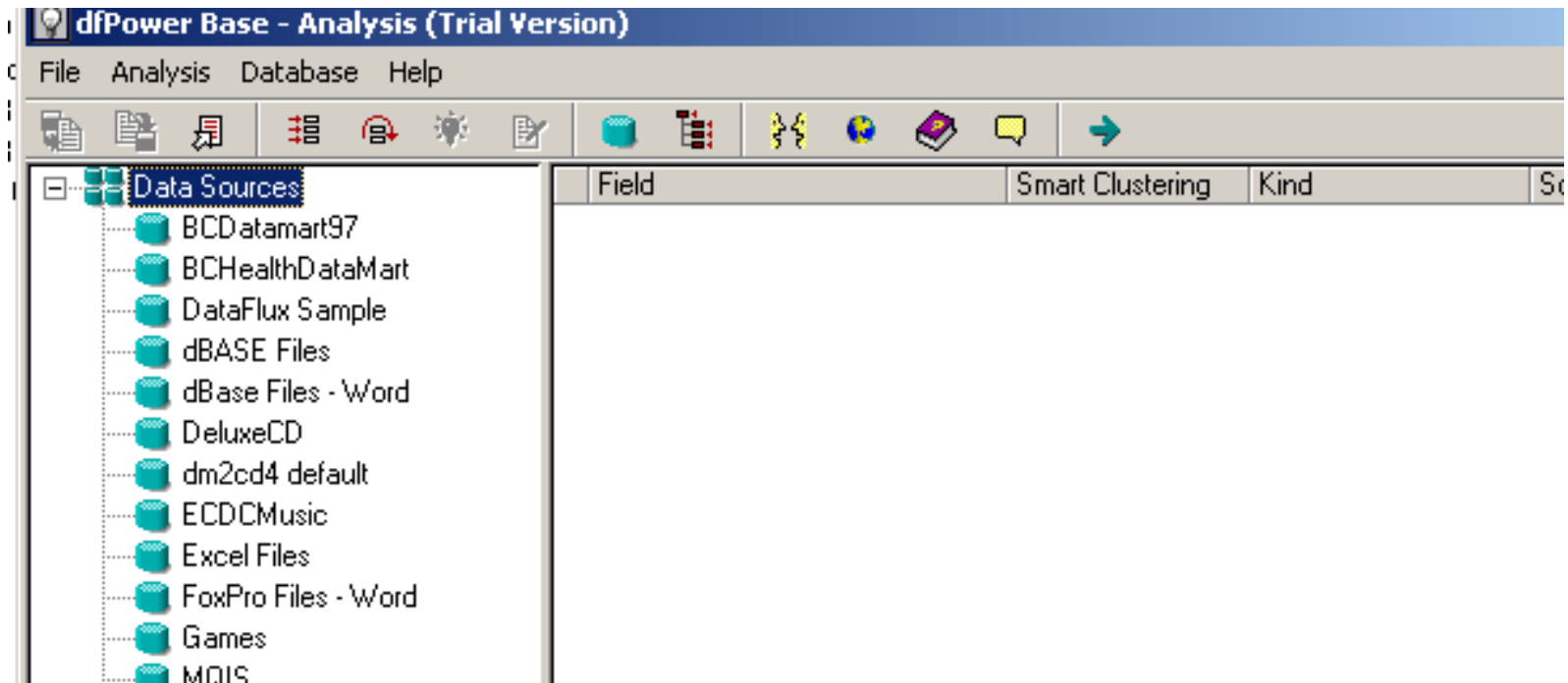
The screenshot displays the dfPower Studio interface. The main window is titled "dfPower Studio" and has a menu bar with "Studio", "Vault", and "Help". Below the menu bar is a toolbar with three icons: a lightbulb for "Analysis", a gear for "Standardize", and a document with a clock for "Batch". The "Vault" pane shows a tree view of the data library structure:

- Vault
  - Batch Schedules
  - Custom Match Types
  - Data Libraries
    - Address Extension Word Standardization
    - Cardinal & Ordinal Number Standardization
    - City Name Data Library
    - Common Address Word Standardization
    - Common Capitalized Word Data Library
    - Common City Word Standardization Data
    - Common Organizational Word Standardization
    - County/State Highway Word Data Library
    - Directional Address Element Standardization
    - English Word Data Library
    - Family Name Data Library**
    - Given Name Data Library

The "Family Name Data Library" is selected, and its properties are displayed in the table on the right:

Property	Value
Creator	DataFlux
Created	Wed Jan 17 17:07:42 2001
Modified	Wed Jan 17 17:07:42 2001
File Name	C:\Program Files\DataFlux\dfPower Studio\vault\datilib\DQAPI2.DAT
Number of Fields	1
Description	Smith,Jones,etc

# dfPower Base - Analysis



# dfPower Base - Analysis

The screenshot shows the 'dfPower Base - Analysis (Trial Version)' application window. The interface includes a menu bar (File, Analysis, Database, Help) and a toolbar with icons for file operations, analysis, and navigation. The left pane displays a tree view of 'Data Sources' with 'DataFlux Sample' expanded to show 'Contacts' and 'NC\_Customer'. The right pane displays a table of fields with columns for 'Field', 'Smart Clustering', 'Kind', and 'Sort Mode'.

Field	Smart Clustering	Kind	Sort Mode
ID			
COMPANY			
CONTACT			
ADDRESS			
CITY			
STATE			
PHONE			
OS			
DATABASE			
MATCH_CD			
DELETE_FLG			

# dfPower Base - Analysis

Field	Smart Clustering	Kind	Sort Mode
ID			
COMPANY	Organization	Phrase Analysis	Alphabetical
CONTACT			
ADDRESS			
CITY			
STATE			
PHONE			
OS			
DATABASE			
MATCH_CD			
DELETE_FLG			

# dfPower Base – Analysis Report

dfPower Base - Analysis Editor

File Analysis Schemes Help

Report: None

Type: Phras

Permutation	Occurrences
First Interstate Bank-Project Management	1
First Marret Bank	1
First Merit Corp	10
>First Merit Corp	2
First Merit Corp.	2
First Merit Bank	13
1st Merit Bank Corp	1
1st Merit Bank	1
First Merit	11
first merit	2
1st Merit	10
First Total Systems Inc	1
First Trust Corporation	1

Data

# dfPower Base – Analysis Report

		Type: <input type="text" value="Phrase"/>	Scheme: Company Standardiz
	Occurrences	Data	Standard
	1	First Interstate Bank	First Interstate Bank
	1	First Interstate Services Company	First Interstate Bank
	<b>10</b>	first merit	First Merit Bank
	2	>First Merit Corp	First Merit Bank
	2	1st Merit Bank	First Merit Bank
	13	1st Merit	First Merit Bank
	1	First Merit Corp	First Merit Bank
	1	First Merit Bank	First Merit Bank
	11	First Merit	First Merit Bank
	2	1st Merit Bank Corp	First Merit Bank
	10	First Merit Corp.	First Merit Bank
	1	First Merit Medical Co.	First Merit Medical Co.

# dfPower Studio's Vault

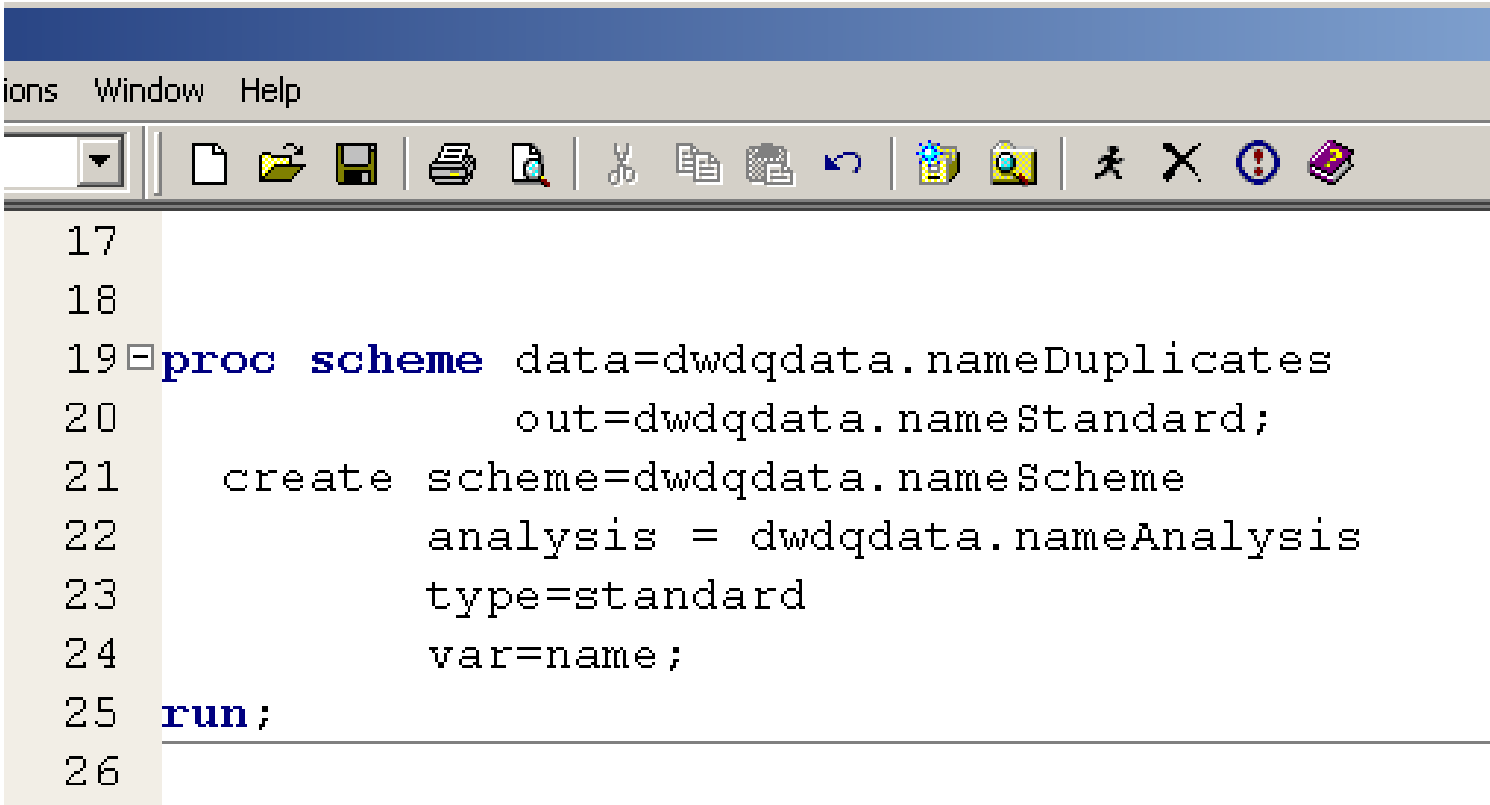
The screenshot displays the dfPower Studio application window. The title bar reads "dfPower Studio" and the menu bar includes "Studio", "Vault", and "Help". Below the menu bar is a "Toolbar" area with three icons: a lightbulb labeled "Analysis" and a gear labeled "Standardize". The main area is titled "Vault" and shows a tree view of the vault structure. The "Vault" folder is expanded, showing sub-items: "Batch Schedules", "Custom Match Types", "Data Libraries" (highlighted), "Jobs", "References", "Reports", and "Standardization Schemes". To the right of the tree view is a table with two columns: "Property" and "Value".

Property	Value
Creator	DataFlux
Children Type	Data Library
Number of Children	22

# Agenda

- Why worry about data quality?
- Data samples before and after cleaning.
- Interactive Data Cleaning
- **Data Cleaning with programs**
- Where to learn more

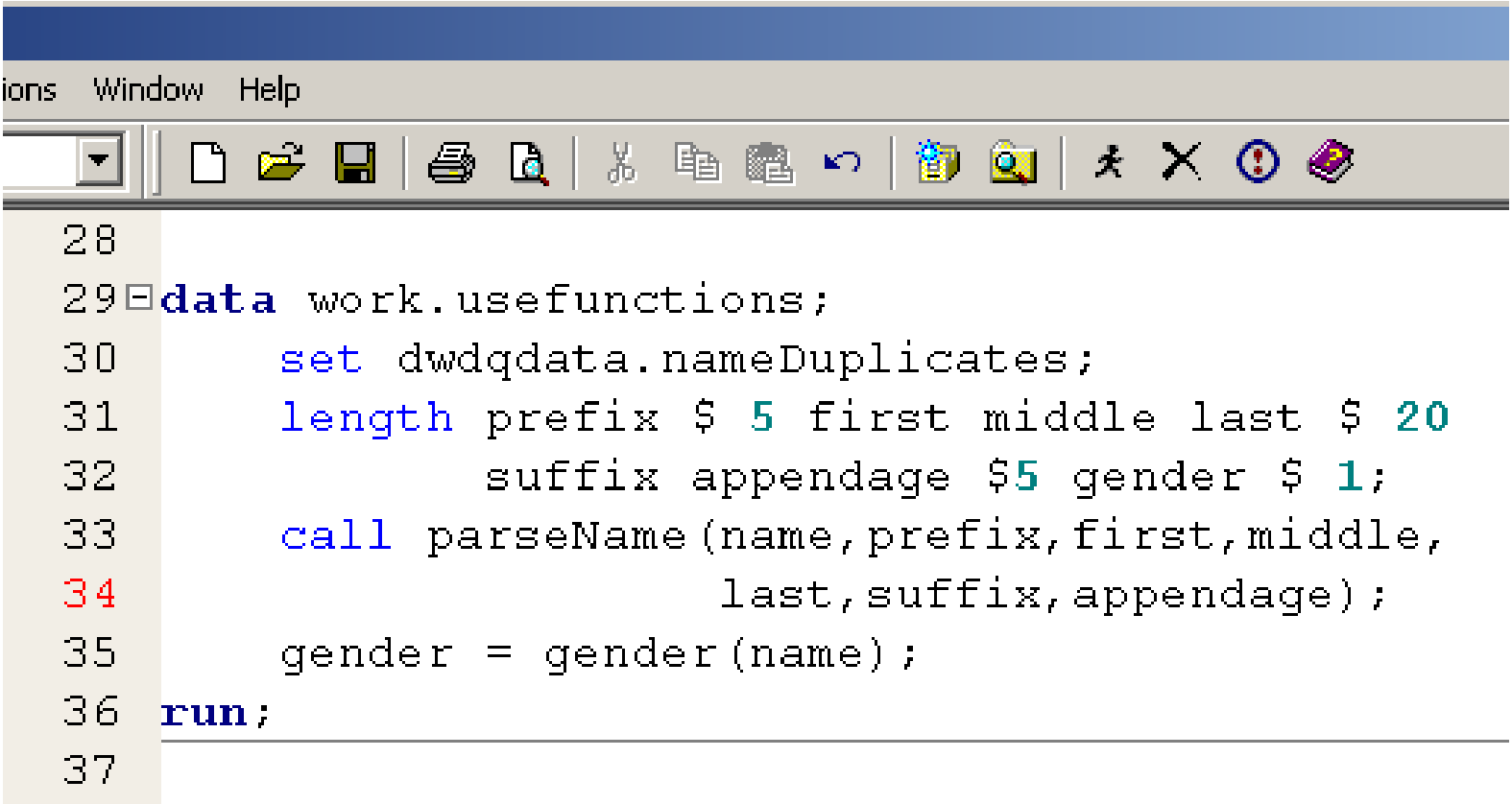
# SAS – Proc Scheme



The image shows a screenshot of the SAS software interface. At the top, there is a menu bar with 'ions', 'Window', and 'Help'. Below the menu bar is a toolbar with various icons for file operations (new, open, save, print, search), editing (cut, copy, paste), and navigation (undo, redo, home, end). The main area of the window displays a SAS code block with line numbers 17 through 26. The code is as follows:

```
17  
18  
19 proc scheme data=dwdqdata.nameDuplicates  
20           out=dwdqdata.nameStandard;  
21   create scheme=dwdqdata.nameScheme  
22     analysis = dwdqdata.nameAnalysis  
23     type=standard  
24     var=name;  
25 run;  
26
```

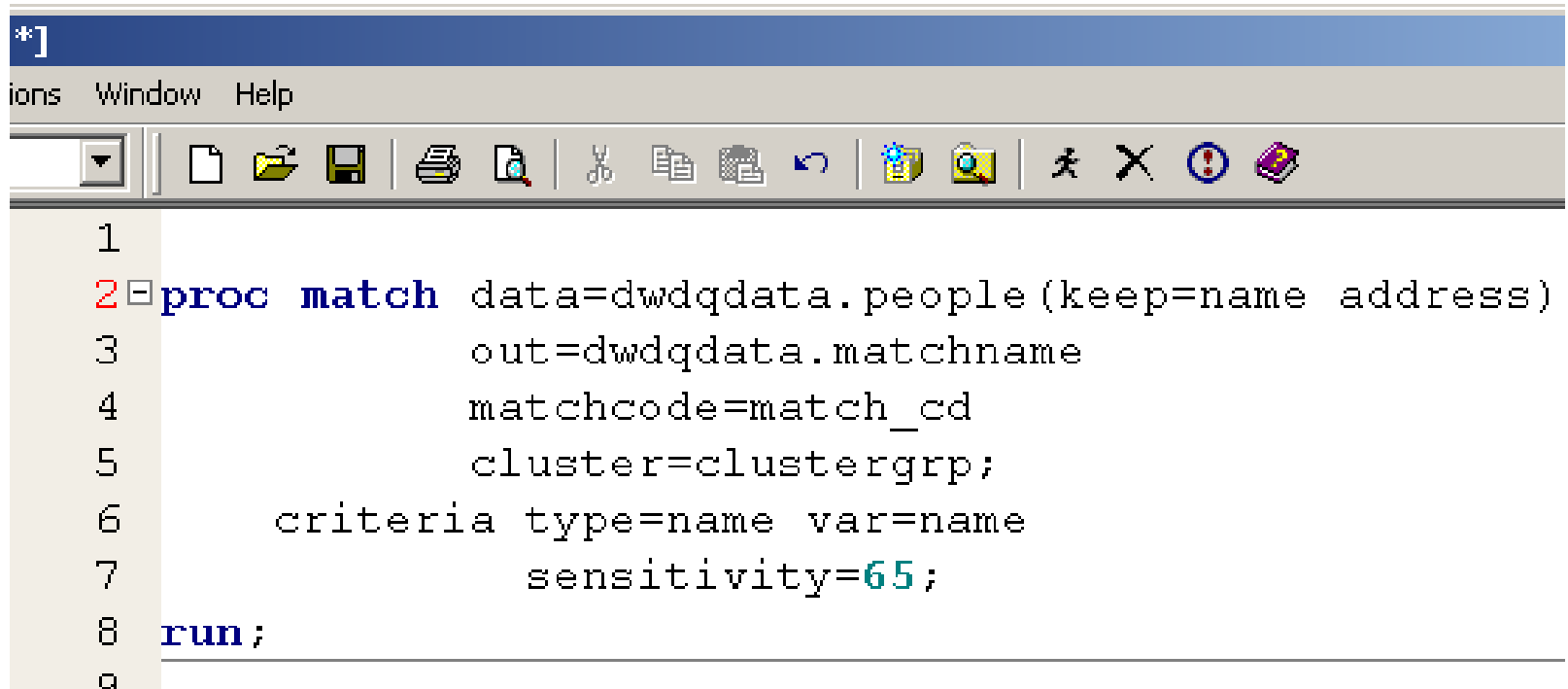
# SAS – DQ functions



The screenshot shows the SAS software interface. At the top, there is a menu bar with 'ions', 'Window', and 'Help'. Below the menu bar is a toolbar with various icons for file operations (new, open, save, print, find), editing (cut, copy, paste, undo, redo), and navigation (home, search, stop, refresh). The main area is a code editor with a light gray background. The code is as follows:

```
28
29 data work.usefunctions;
30     set dwdqdata.nameDuplicates;
31     length prefix $ 5 first middle last $ 20
32           suffix appendage $5 gender $ 1;
33     call parseName(name, prefix, first, middle,
34                   last, suffix, appendage);
35     gender = gender(name);
36 run;
37
```

# SAS – Proc Match process name only



The image shows a screenshot of the SAS software interface. At the top, there is a menu bar with 'ions', 'Window', and 'Help'. Below the menu bar is a toolbar with various icons for file operations (new, open, save, print, find) and editing (cut, copy, paste, undo, redo). The main area displays a SAS code block with the following text:

```
1  
2 proc match data=dwdqdata.people (keep=name address)  
3     out=dwdqdata.matchname  
4     matchcode=match_cd  
5     cluster=clustergrp;  
6     criteria type=name var=name  
7     sensitivity=65;  
8 run;  
9
```

# Match Codes – name only

	standardname	Customer Name	Address	MATCH_CD	CL
164	Stacy Wyndam	Stacey Windom	5100 Ming Ave	LP84~	
165	Mr. Donnie Williams	Mr. Donnie Williams	307 Boltstone Ct	LWB8@	
166	Mr. Donnie Williams	MR DON F WILLIAMS	25670 W HEDGEWOOD DRIVE	LWB8@	
167	Mr. Donnie Williams	Don Williams	6512 Six Forks Rd Suite 404B	LWB8@	
168	Mr. Donnie Williams	DON WILLIAMS	25670 W HEDGEWOOD DR	LWB8@	
169	Mr. Donnie Williams	DONALD F. WILLIAMS	25670 WEST HEDGEWOOD DR	LWB8@	
170	Mr. Donnie Williams	DON WILIAMS	25670 W HEDGE WOOD	LWB8@	
171	Mr. Donnie Williams	DONALD WILLIAMS	25670 W HEDGE WOOD DR.	LWB8@	
172	Mr. Donnie Williams	DONNY WILLIAMS	25670 W. HEDGEWOOD DR.	LWB8@	
173	Sheryl Wellman	Sheryl Wellman	PO Box 3887	LWBJ2	

# SAS – Proc Match

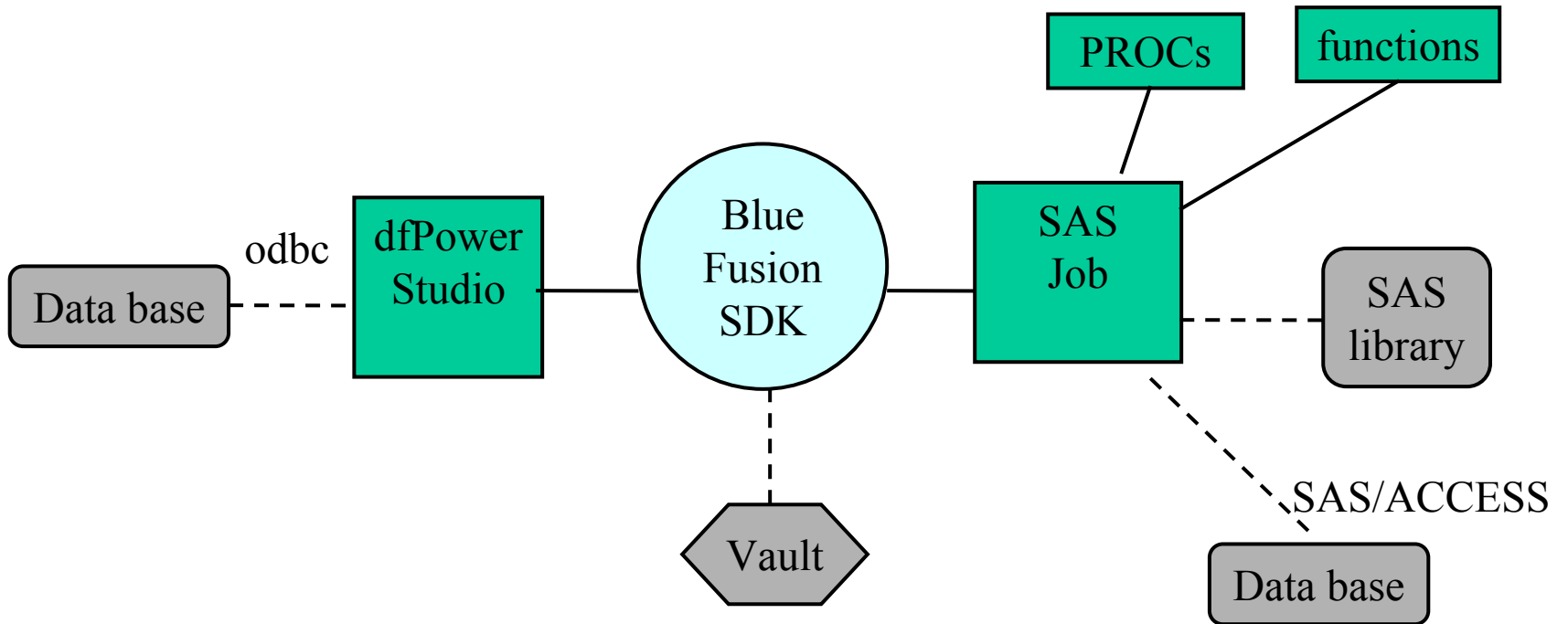
## process name and address

```
*]
ions Window Help
[Icons] [File] [Open] [Save] [Print] [Find] [Cut] [Copy] [Paste] [Undo] [Redo] [Run] [Stop] [Help] [Error]
36
37 proc match data=dwdqdata.people (keep=name address)
38     out=dwdqdata.matchnameaddress
39     matchcode=match_cd
40     cluster=clustergrp;
41     criteria type=name var=name
42         sensitivity=65;
43     criteria type=address var=address
44         sensitivity=65;
45 run;
46
```

# Match Codes – name and address

	standardname	Customer Name	Address	MATCH_CD
46	Mr. William Jackson	William Jaackson	515 E Broad St., Suite 12	CX4M7!5Z5\$MY8ZH
47	MR DON F WILLIAMS	MR DON F WILLIAMS	25670 W HEDGEWOOD DRIVE	LWB8@!H56I2CL\$\$
48	MR DON F WILLIAMS	DON WILLIAMS	25670 W HEDGEWOOD DR	LWB8@!H56I2CL\$\$
49	MR DON F WILLIAMS	DONALD F. WILLIAMS	25670 WEST HEDGEWOOD DR	LWB8@!H56I2CL\$\$
50	MR DON F WILLIAMS	DON WILIAMS	25670 W HEDGE WOOD	LWB8@!H56I2CL\$\$
51	MR DON F WILLIAMS	DONALD WILLIAMS	25670 W HEDGE WOOD DR.	LWB8@!H56I2CL\$\$
52	MR DON F WILLIAMS	DONNY WILLIAMS	25670 W. HEDGEWOOD DR.	LWB8@!H56I2CL\$\$
53	Rob Beckett	Rob Beckett	392 S. Main St. PO Box 2270	M3~M@!K-H\$BP\$HH
54	Rob Beckett	Rob Beckett	392 S. Main St. PO Box 2270	M3~M@!K-H\$RP\$HH

# DataFlux Functionality



## How to Learn More

- Instructor based training:
  - <http://support.sas.com/training/Canada>
  - “SAS Data Quality-Cleanse”, a two-day course, starting on April 22 in Toronto

# How to Learn More

- Technical Support

- <http://support.sas.com/rnd/warehousing/quality>