



# Data Quality Assessment Made Easy with Teradata Profiler

**Frank Capobianco**

**Senior Data Mining Consultant  
Teradata Data Mining Lab**





**PARTNERS**



---

*The Teradata User Group*

Teradata  
a division of  NCR



**YOUR  
GROWTH  
CONNECTION** **2004**



*\$600 Billion*

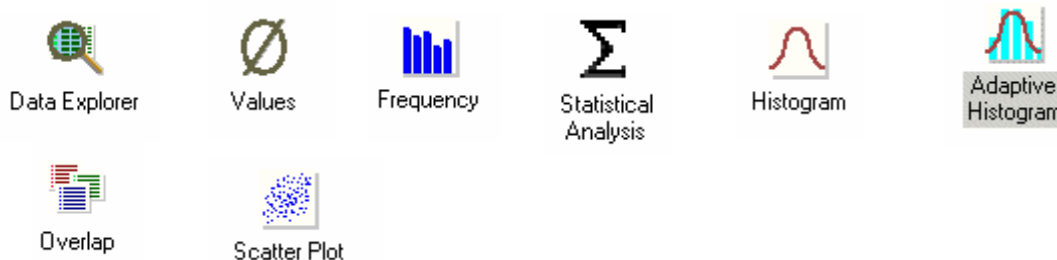
## ***Background***

- **Data Quality: *the suitability of data for its intended purpose***
- **The Data Warehousing Institute estimates that data quality problems cost US businesses more than \$600 B per year<sup>1</sup>**
- **A Data Warehousing Institute survey also indicates SQL queries are the most common methodology for data quality assurance<sup>1</sup>**
- **Teradata Profiler can make data quality assessment easy by eliminating such query writing**

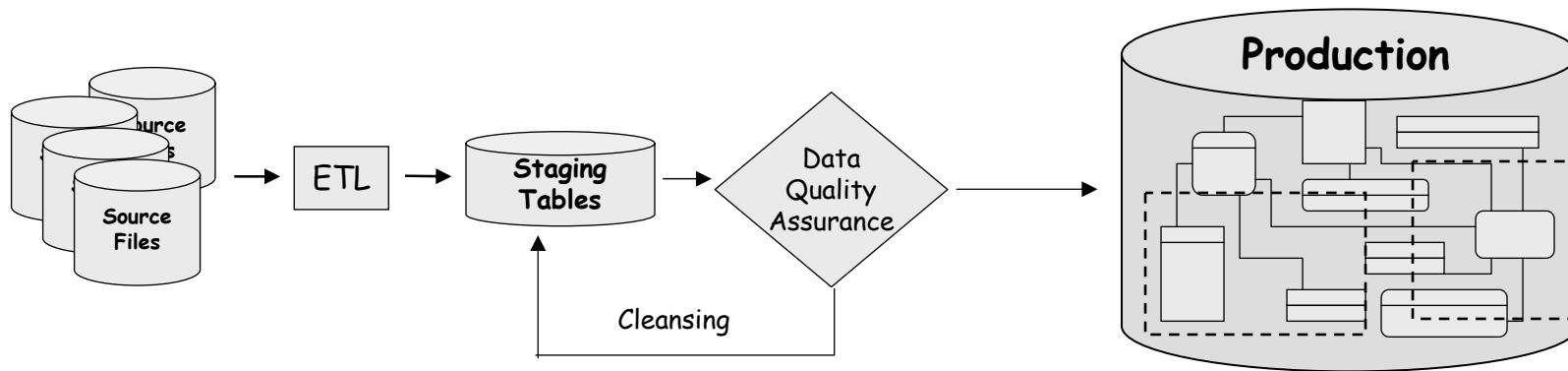
<sup>1</sup> “Data Quality and the Bottom Line”, Wayne Eckerson, The Data Warehousing Institute, 2002

# What is Teradata Profiler?

- **In-database data profiling software**
  - ODBC-based client tool requiring no additional hardware, no data movement
- **Enabled by the power of Teradata SQL**
  - All Warehouse Miner functions result in the generation and execution of Teradata SQL
- **Offers several data quality assessment functions**



# Data Quality Assessment Settings



- **Some users use Teradata Profiler to assess data in existing databases**
  - to establish quality baselines
  - to ensure quality standards are maintained over time
- **Some users may find Teradata Profiler particularly effective to apply to data once it has been loaded into *staging* tables and as a condition for loading into *production* tables**

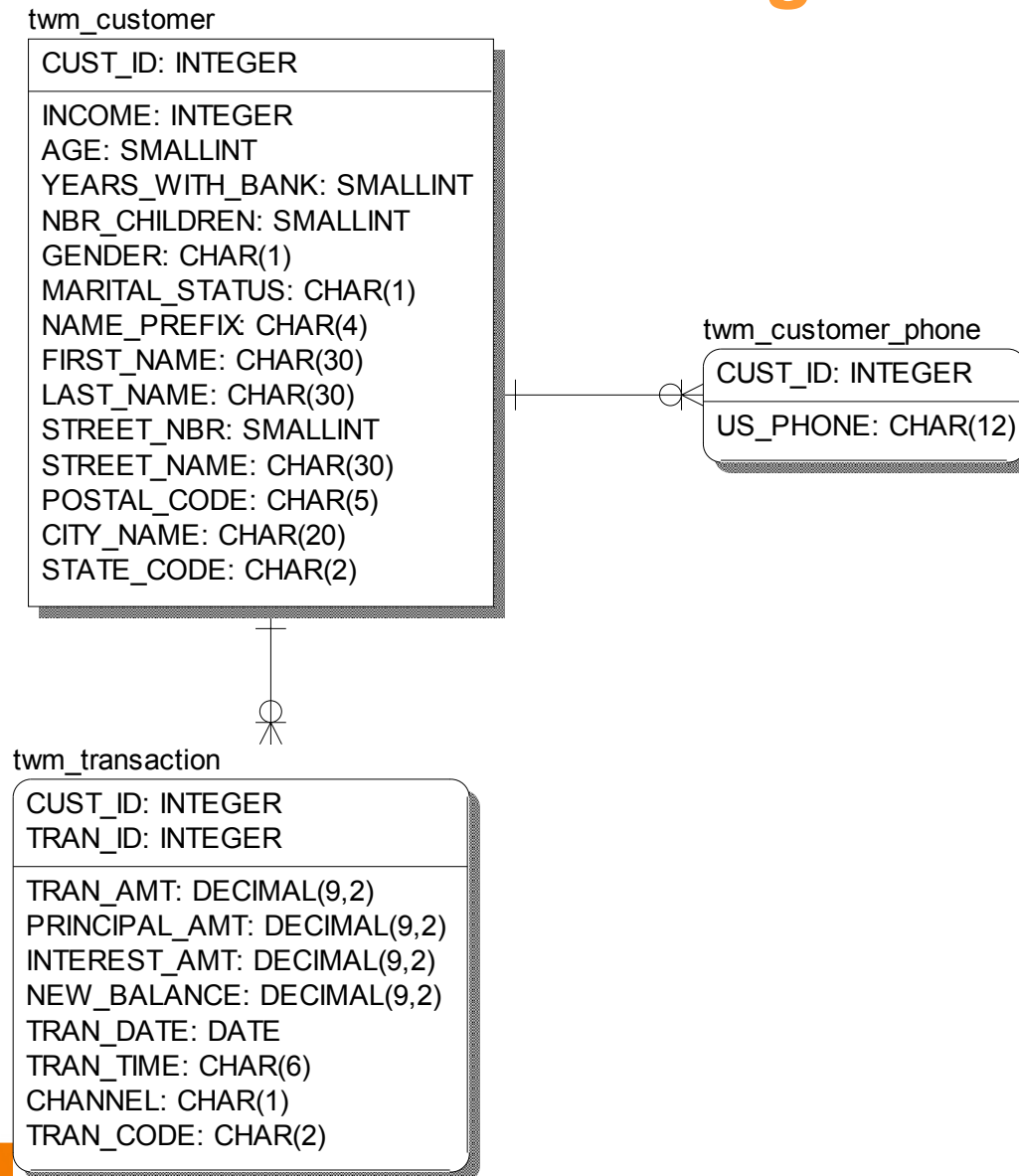
## *Data Quality Assessment Planning*

- **A data quality assessment plan is a project plan**
  - defining data quality in specific, measurable terms, and
  - describing who, how, where, and when the assessment of actual data against this definition is to be carried out
- **Sample Plan Excerpts:**
  - Conduct broad-based profiling of the following tables, comparing results against the documentation represented by the data model and data dictionaries . . .
  - Populate the Customer Data Quality Metadata Table, which includes date-of-run, table profiled, column profiled, % missing values, % unique values, . . . .
  - Develop a monthly data load filter to flag any variances from the profile for the previous five month's data

## *Common Data Quality Assessment Tasks*

- **Broad-based assessment using Data Explorer**
  - Four functions packaged as one
  - Uses Teradata Session Pooling for concurrent table analysis
- **Targeted assessment using individual functions**
  - Enabled by SQL Clause options on individual functions
  - Frequency (Listing Specific Values, Specific Integrity Violations)
- **Constraint Checking**
  - Overlap (Referential Integrity)
- **Interactive explorations, using graphic features**
  - Most Profiler functions offer graphic output
- **As a tool development vehicle**
  - SQL can be extracted and run outside of Profiler
  - Profiler can be run in batch mode
  - Tables created by Profiler can be accessed DYNAMICALLY using Excel, for instance

# Data to be Used in Performing these Tasks



# Teradata Profiler

## - 7 Basic Steps to Results

- **there are 7 basic steps in the use of Teradata Profiler**
  - connect to an ODBC data source with appropriate permissions
    - set the database connection properties
  - create a new, (or open an existing) *Project*
  - add at least one *Analysis* to the *Project*
  - set input options
    - select table(s) and columns to be analyzed
    - set analysis parameters
    - set expert options ▶
  - set output options
  - execute the *Analysis* (using the run icon )
  - examine, interpret, and use results of interest
- **that's it!**

# Assessing all tables with DATA EXPLORER



Teradata Profiler - [twm\_tables - Data Explorer]

File View Project Tools Window Help

twm\_tables date created: 6/30/2004 last time run: 6/30/2004  
 date modified: 6/30/2004 last complete run: 6/30/2004

INPUT OUTPUT RESULTS

data graphs

demo1\_values (27 rows) Sort Format Select All Copy

tbl	col	xtype	xcnt	xnull	xunique	xblank	xzero	xpos
twm_customer	age	SMALLINT	749.00	0.00	79.00		0.00	748.00
twm_customer	city_name	CHAR(20) C	749.00	0.00	69.00	0.00		
twm_customer	cust_id	INTEGER	749.00	0.00	749.00		0.00	749.00
twm_customer	first_name	CHAR(30) C	749.00	0.00	406.00	0.00		
twm_customer	gender	CHAR(1) CH	749.00	0.00	3.00	0.00		
twm_customer	income	INTEGER	749.00	0.00	641.00		102.00	647.00
twm_customer	last_name	CHAR(30) C	749.00	0.00	345.00	0.00		
twm_customer	marital_statu	CHAR(1) CH	749.00	0.00	4.00	0.00		
twm_customer	name_prefix	CHAR(4) CH	749.00	0.00	6.00	506.00		
twm_customer	nbr_children	SMALLINT	749.00	0.00	7.00		468.00	281.00
twm_customer	postal_code	CHAR(5) CH	749.00	0.00	442.00	0.00		
twm_customer	state_code	CHAR(2) CH	749.00	0.00	33.00	0.00		
twm_customer	street_name	CHAR(30) C	749.00	0.00	122.00	0.00		
twm_customer	street_nbr	SMALLINT	749.00	0.00	466.00		0.00	749.00
twm_customer	years_with_b	SMALLINT	749.00	0.00	11.00		88.00	661.00
twm_customer_phone	cust_id	INTEGER	0.00		0.00			
twm_customer_phone	us_phone	CHAR(12) C	0.00		0.00			
twm_transaction	channel	CHAR(1) CH	46206.00	0.00	10.00	10767.00		
twm_transaction	cust_id	INTEGER	46206.00	0.00	521.00		0.00	4620.00
twm_transaction	interest_amt	DECIMAL(9,2)	46206.00	0.00	654.00		44319.00	1887.00
twm_transaction	new_balance	DECIMAL(9,2)	46206.00	0.00	32081.00		272.00	4593.00
twm_transaction	principal_amt	DECIMAL(9,2)	46206.00	0.00	20154.00		10043.00	6815.00
twm_transaction	tran_amt	DECIMAL(9,2)	46206.00	0.00	20798.00		8156.00	8702.00

Project Explorer

- Data Quality Demonstration
  - twm\_tables

Execution Status

Analysis	Status	Message
twm_tables	Complete	Execution complete

# Reviewing Explorer Results

- **Things to look for in Explorer results in tabular form:**

- Values:

- Number of unique values

Does this agree with the available documentation, such as from a data dictionary or Erwin model?

Is the number of unique foreign key values in a given table at least as great as the number of unique values it assumes in its primary table?



Values

- Constant columns

Generally such have no application value

- Missing values

Are there potentially multiple conventions for representing missing values, such as nulls and blanks in character fields?

- (Obvious) domain violations

Negative values in count or amount columns

- Frequency:

- Follow-up, drill-down answers to questions raised by Values results

- Disproportionately high or low counts for a given column

Very low counts could indicate a data entry or transformation error

- Differences in coding conventions



Frequency

# Reviewing Explorer Results

- **Things to look for in Explorer results in tabular form:**

- **Statistics:**

- Outliers, although the associated box-and-whiskers plots may provide a quicker means of identification
- If the assessment plan calls for an understanding of the distribution of the data, and the appropriate advanced options – such as kurtosis and skewness were selected – look for the shape of the distribution



Statistical  
Analysis

- **Histograms:**

- If bins were chosen as the analysis option, the results will divide the range of values the column may assume into equal size intervals. Look for intervals with disproportionately high or low frequencies.
- If quantiles were chosen as the analysis option, the results will divide the range of values the column may assume into intervals with approximately the same frequencies in each. Look for intervals that are disproportionately wide or narrow.



Histogram

# What do Explorer's VALUES Results tell you in this case?

Teradata Profiler - [twm\_tables - Data Explorer]

twm\_tables date created: 6/30/2004 last time run: 6/30/2004  
date modified: 6/30/2004 last complete run: 6/30/2004

demo1\_values (27 rows)

xtbl	xcol	xtype	xcnt	xnull	xunique	xblank	xzero	xpos	xneg
twm_customer	age	SMALLINT	749.00	0.00	79.00		0.00	748.00	1.00
twm_customer	city_name	CHAR(20) C	749.00	0.00	69.00	0.00			
twm_customer	cust_id	INTEGER	749.00	0.00	749.00		0.00	749.00	0.00
twm_customer	first_name	CHAR(30) C	749.00	0.00	406.00	0.00			
twm_customer	gender	CHAR(1) CH	749.00	0.00	3.00	0.00			
twm_customer	income	INTEGER	749.00	0.00	641.00		102.00	647.00	0.00
twm_customer	last_name	CHAR(30) C	749.00	0.00	345.00	0.00			
twm_customer	marital_status	CHAR(1) CH	749.00	0.00	4.00	0.00			
twm_customer	name_prefix	CHAR(4) CH	749.00	0.00	6.00	506.00			
twm_customer	nbr_children	SMALLINT	749.00	0.00	7.00		468.00	281.00	0.00
twm_customer	postal_code	CHAR(5) CH	749.00	0.00	442.00	0.00			
twm_customer	state_code	CHAR(2) CH	749.00	0.00	33.00	0.00			
twm_customer	street_name	CHAR(30) C	749.00	0.00	122.00	0.00			
twm_customer	street_nbr	SMALLINT	749.00	0.00	466.00		0.00	749.00	0.00
twm_customer	years_with_bank	SMALLINT	749.00	0.00	11.00		88.00	661.00	0.00
twm_customer_phone	cust_id	INTEGER	0.00		0.00				
twm_customer_phone	us_phone	CHAR(12) C	0.00		0.00				
twm_transaction	channel	CHAR(1) CH	46206.00	0.00	10.00	10767.00			
twm_transaction	cust_id	INTEGER	46206.00	0.00	521.00		0.00	46206.00	0.00
twm_transaction	interest_amt	DECIMAL(9,2)	46206.00	0.00	654.00		44319.00	1887.00	0.00
twm_transaction	new_balance	DECIMAL(9,2)	46206.00	0.00	32081.00		272.00	45934.00	0.00
twm_transaction	principal_amt	DECIMAL(9,2)	46206.00	0.00	20154.00		10043.00	6815.00	29348.00
twm_transaction	tran_amt	DECIMAL(9,2)	46206.00	0.00	20798.00		8156.00	8702.00	29348.00

Execution Status

Analysis	Status	Message
twm_tables	Complete	Execution complete

Note negative value for age, the presence of 3 unique values for gender, and 506 blanks for name\_prefix in twm\_customer. Note twm\_customer\_phone is an empty table.

# Cross-Reference with Explorer's FREQUENCY Results

Teradata Profiler - [twm\_tables - Data Explorer]

twm\_tables      date created: 6/30/2004      last time run: 6/30/2004  
date modified: 6/30/2004      last complete run: 6/30/200

demo1\_frequency (196 rows)      Sort      Format      Select All      Copy

xtbl	xcol	xval	xcnt	xpct
twm_customer	city_name	Tampa	2.00	0.27
twm_customer	city_name	Arlington	1.00	0.13
twm_customer	city_name	Buffalo	1.00	0.13
twm_customer	city_name	Virginia Beach	1.00	0.13
twm_customer	gender	F	419.00	55.94
twm_customer	gender	M	328.00	43.79
twm_customer	gender	1	2.00	0.27
twm_customer	last_name	Pieper	8.00	1.07
twm_customer	marital_status	2	354.00	47.26
twm_customer	marital_status	1	277.00	36.98
twm_customer	marital_status	4	70.00	9.35
twm_customer	marital_status	3	48.00	6.41
twm_customer	name_prefix		506.00	67.56
twm_customer	name_prefix	Ms.	92.00	12.28
twm_customer	name_prefix	Mr.	82.00	10.95
twm_customer	name_prefix	Mrs.	31.00	4.14
twm_customer	name_prefix	Dr.	23.00	3.07
twm_customer	name_prefix	Miss	15.00	2.00
twm_customer	nbr_children	0	468.00	62.48
twm_customer	nbr_children	1	114.00	15.22
twm_customer	nbr_children	2	110.00	14.69
twm_customer	nbr_children	3	38.00	5.07
twm_customer	nbr_children	4	9.00	1.20

Execution Status

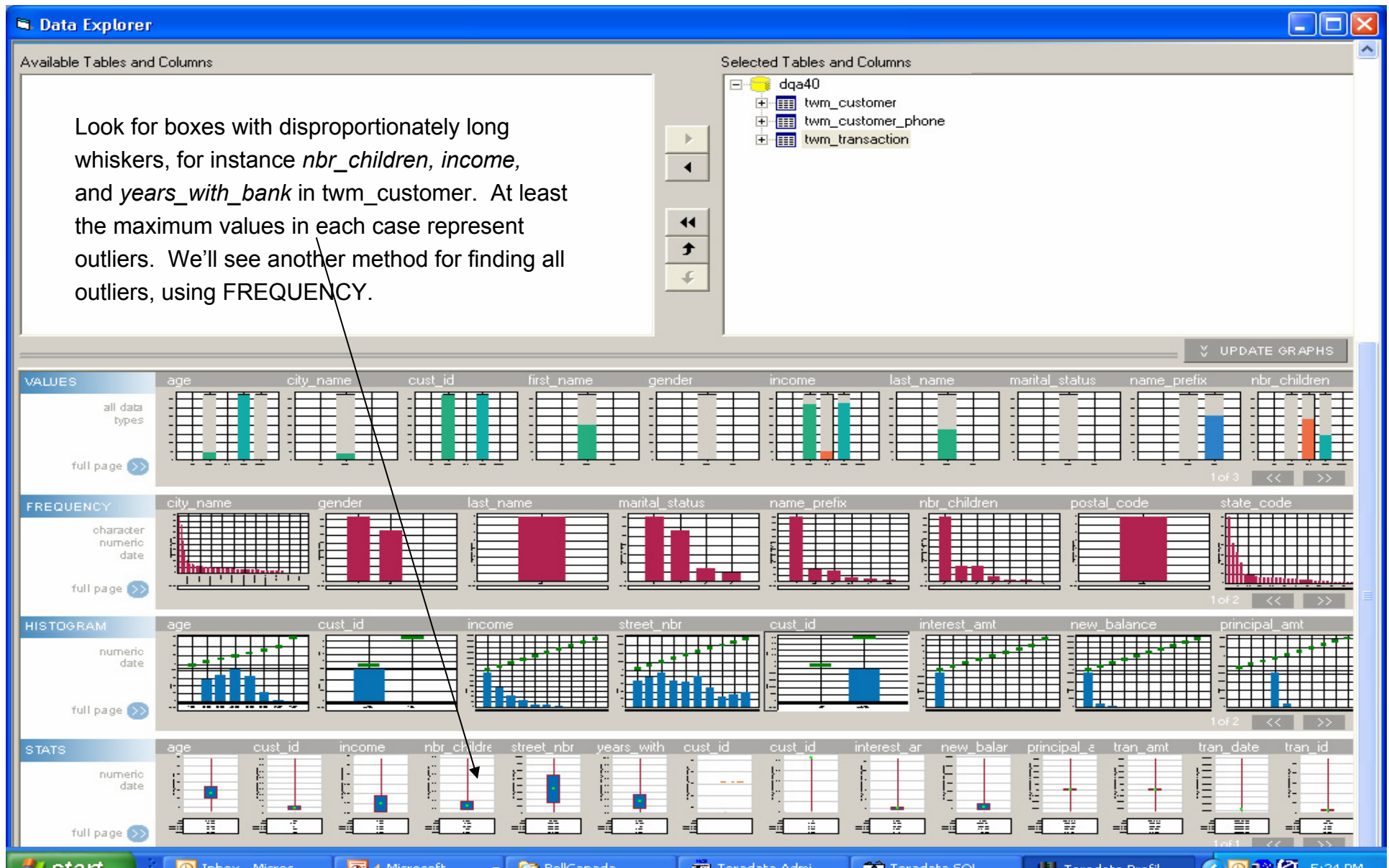
Analysis	Status	Message
twm_tables	Complete	Execution complete

Use FREQUENCY results to drill-down on the three values of *gender*.

Look also for high/low frequencies of a given column value.

# Examine Explorer's STATISTICS results, looking for potential outliers

Look for boxes with disproportionately long whiskers, for instance *nbr\_children*, *income*, and *years\_with\_bank* in *twm\_customer*. At least the maximum values in each case represent outliers. We'll see another method for finding all outliers, using FREQUENCY.



# Why is age “missing”, as it occurs in the FREQUENCY results for twm\_customer?

**Data Explorer Expert Options instruct the software**

- what to analyze
- what to skip
- how to analyze it

**Note the default of 20, as the maximum unique numeric/date values for frequency analysis.**

Execution Status

Analysis	Status	Message
Data Explorer1	Complete	Execution complete

## *Using Stand-Alone Functions for Drill-Down*

- **Teradata Profiler also provides Values, Frequency, Statistics, and Histogram as standalone analyses**

### **Why?**

- **Each of these analysis offers the data profiler SQL options not present as part of the Data Explorer function**
- **These SQL options greatly expand the capabilities of these individual analyses**
  - to restrict the scope of an analysis, such as via a WHERE clause
  - to return results, based on categorization using a GROUP BY clause
  - to filter results – such as in showing duplicates only – using a HAVING clause
- **As standalone analyses, the generated SQL is exposed for further leverage as well**

# Data Quality Assessment using FREQUENCY

- **Warehouse Miner FREQUENCY function**
  - Arguably the Warehouse Miner function most useful for data quality assessment, next to Data Explorer
  - What does it do? LIST & COUNT !
    - Provide simple counts stand-alone or as part of Data Explorer
    - Provide counts of duplicate or near duplicate columns, alone or in combination
    - Provide counts of referential integrity violations
    - Provide counts of outliers - values which may be incorrect



# FREQUENCY for combined column counts

The screenshot shows the Teradata Profiler interface for a 'Frequency1' analysis. The window title is 'Frequency1 - Frequency'. The 'INPUT' tab is active, showing configuration options. Under '1) Select Frequency Style', the 'Crosstab' style is selected. Under '2) Select Columns From One Table', the 'Available Tables' dropdown is set to 'twm\_customer'. The 'Available Columns' list includes 'city\_name', 'cust\_id', 'first\_name', 'gender', 'income', 'last\_name', 'marital\_status', 'name\_prefix', 'nbr\_children', 'postal\_code', 'state\_code', 'street\_name', 'street\_nbr', and 'years\_with\_bank'. The 'Selected Columns' list shows 'street\_nbr', 'street\_name', and 'city\_name' selected from the 'twm\_customer' table. The 'Project Explorer' on the right shows the project structure: 'Data Quality Demonstration' > 'twm\_tables' > 'Frequency1'. The 'Execution Status' window at the bottom is empty.

Pick the Crosstab style, and then the columns shown. This will list and count all such address combinations.

# FREQUENCY for Duplicate Detection

The screenshot displays the Teradata Profiler interface. The main window is titled "Frequency1 - Frequency" and shows the configuration for a data quality analysis. The "Optional HAVING clause text" field contains the SQL clause "xcnt > 1". The "Optional WHERE clause text" and "Optional QUALIFY clause text" fields are empty. The "Execution Status" window at the bottom shows the analysis is complete.

Teradata Profiler

File View Project Tools Window Help

Frequency1 - Frequency

date created: 7/1/2004 last time run: 7/1/2004  
date modified: 7/1/2004 last complete run: 7/1/2004

INPUT OUTPUT RESULTS

data selection analysis parameters expert options

Optional WHERE clause text:

Optional HAVING clause text:  
xcnt > 1

Optional QUALIFY clause text:

Project Explorer

Data Quality Demonstration

twm\_tables

Frequency1

Execution Status

Analysis	Status	Message
Frequency1	Complete	Execution complete

This one small SQL clause will cause duplicates only to be listed

# Results of Duplicate Detection

Teradata Profiler

File View Project Tools Window Help

Frequency1 - Frequency

date created: 7/1/2004 date modified: 7/1/2004 last time run: 7/1/2004 last complete run: 7/1/2004

Table (196 rows)

street_nbr	street_name	city_name	xcnt	xpct
13154	Sunset	Los Angeles	6.00	0.80
16437	37th	Long Beach	5.00	0.67
3204	Kingsbury	Portland	4.00	0.53
3830	Tenth	Honolulu	4.00	0.53
4490	E	Portland	4.00	0.53
5364	Capital	Philadelphia	4.00	0.53
5845	Eighth	Los Angeles	4.00	0.53
5874	Old Cliffs	Los Angeles	4.00	0.53
5975	Lilac	San Francisco	4.00	0.53
6082	23rd	Albuquerque	4.00	0.53
6099	Lima	St Paul	4.00	0.53
6797	Alamo	Tulsa	4.00	0.53
7625	35th	Austin	4.00	0.53
9335	Division	New York City	4.00	0.53
11307	C	Houston	4.00	0.53
11583	Rose	Chicago	4.00	0.53
11763	106th	Long Beach	4.00	0.53
18391	J	Sacramento	4.00	0.53
731	Marguerite	Toledo	3.00	0.40
1223	Vine	New York City	3.00	0.40
1944	Hoover	Tucson	3.00	0.40
1980	106th	Chicago	3.00	0.40

Project Explorer

Data Quality Demonstration

twm\_tables

Frequency1

Execution Status

Analysis	Status	Message
Frequency1	Complete	Execution complete

For this data set, here are the duplicate cities, streets, and street numbers – along with the duplicate count for each triple.

# FREQUENCY for referential integrity assessment

- checking whether there are any orphan *cust\_id* values

Frequency1 - Frequency

Frequency1

date created: 7/1/2004 last time run:  
date modified: 7/1/2004 last complete run:

INPUT OUTPUT RESULTS

data selection analysis parameters expert options

1) Select Frequency Style

Crosstab (compute the frequencies of every permutation of v.

2) Select Columns From One Table

Available Databases: dqa40

Available Tables: twm\_transaction

Available Columns:

- channel
- cust\_id
- interest\_amt
- new\_balance
- principal\_amt
- tran\_amt
- tran\_code
- tran\_date
- tran\_id
- tran\_time

Selected Columns:

Frequency Columns

- dqa40
  - twm\_transaction
    - cust\_id
    - tran\_id

For *twm\_transaction*, set up a Crosstab Frequency on *cust\_id* and *tran\_id*.