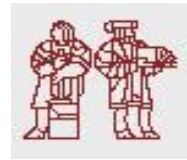




The MIT Information Quality Industry Symposium, 2007



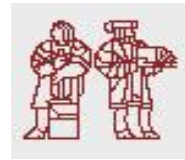
Solutions... from the Data Up

Presented by
Chuck Backus
CTO, Qbase Inc.

Date 06/04/2007



The MIT Information Quality Industry Symposium, 2007



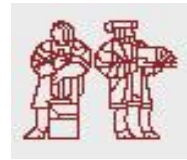
Agenda

- About Qbase
- Solutions... from the Data Up
- Data Strategy
- Data, Information and BI
- Data Challenges
- Profiling Data
- Rapid Data Analysis
- Summary

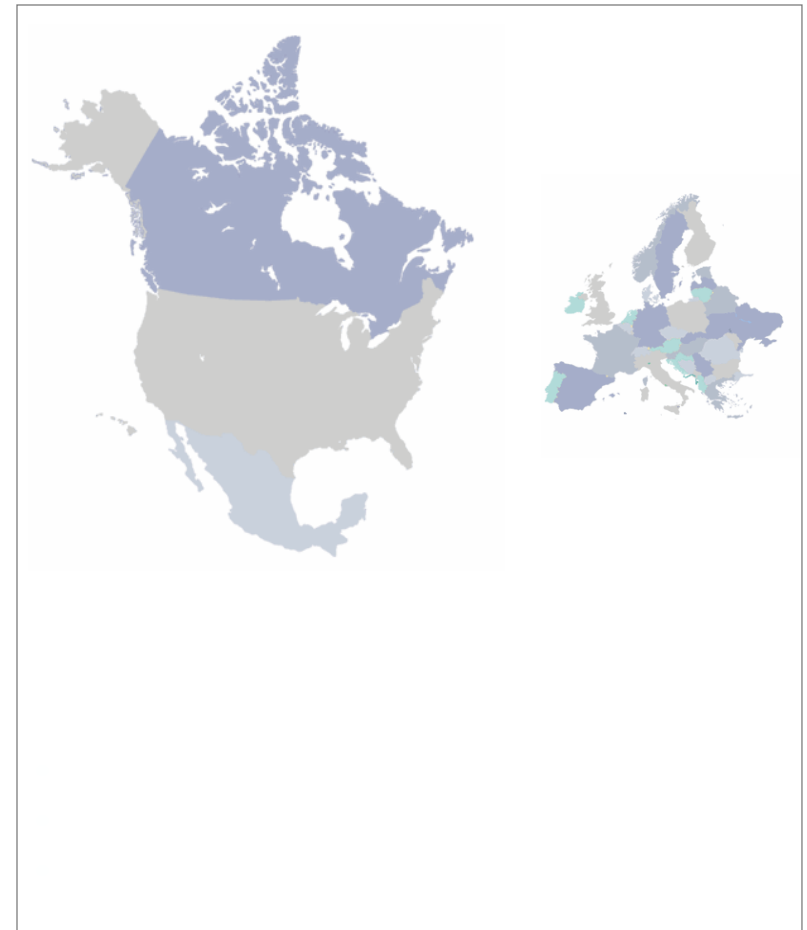


The MIT Information Quality Industry Symposium, 2007

About Qbase



- Technology and leadership team from LexisNexis, Lockheed Martin, Cox Publishing, and national premier nonprofits
- We serve nonprofit organizations, state and federal government agencies, US military, higher education institutions, healthcare facilities and provide direct marketing solutions.
- Markets built around industry expertise





The MIT Information Quality Industry Symposium, 2007
Solutions... from the Data Up



“...it was estimated that poor quality customer data costs U.S. businesses a staggering \$611 billion a year in postage, printing, and staff overhead.”

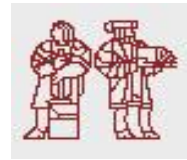
The Data Warehousing Institute's (TDWI) 2002 Data Quality and The Bottom Line Report

“Clean data is the key to focused campaigns and will prevent you from spending money on dead-end leads. Unfortunately, only 61% of companies believe their data is accurate enough for decision-making, and 27% agree that the information they need isn't there.”

The Direct Marketing Association's 2005 Annual Report



The MIT Information Quality Industry Symposium, 2007



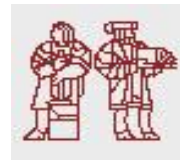
Solutions... from the Data Up

- Data/information have a critical role in business
- Data usually gets the least focus in an enterprise
- Data challenges can make it very difficult to leverage significant investments in infrastructure and operations
- Planning for data quality and data's role in operations can help avoid pitfalls
- Building solutions “from the data up” ensures appropriate focus on data's role



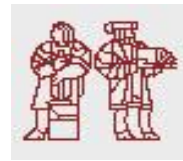
Enterprise data collections are numerous and diverse

- Customer database
- Transactions (e.g., sales)
- Accounting systems
- Personnel
- Regulatory (e.g., audit trails)
- And many more...
- Data is often in “stovepiped” systems
- Data integration amplifies data value



Data Strategy

- Data collections in enterprises are built over time, and rarely are they organized holistically
- It makes sense to approach enterprise data *strategically*:
 - Consider future information needs
 - Engineer data solutions to solve specific needs
 - Keep an eye on extensibility
- Develop a data governance strategy
 - Determine how and when data interacts
 - Ensure data sources can be integrated
- Data governance is a must for Business Intelligence

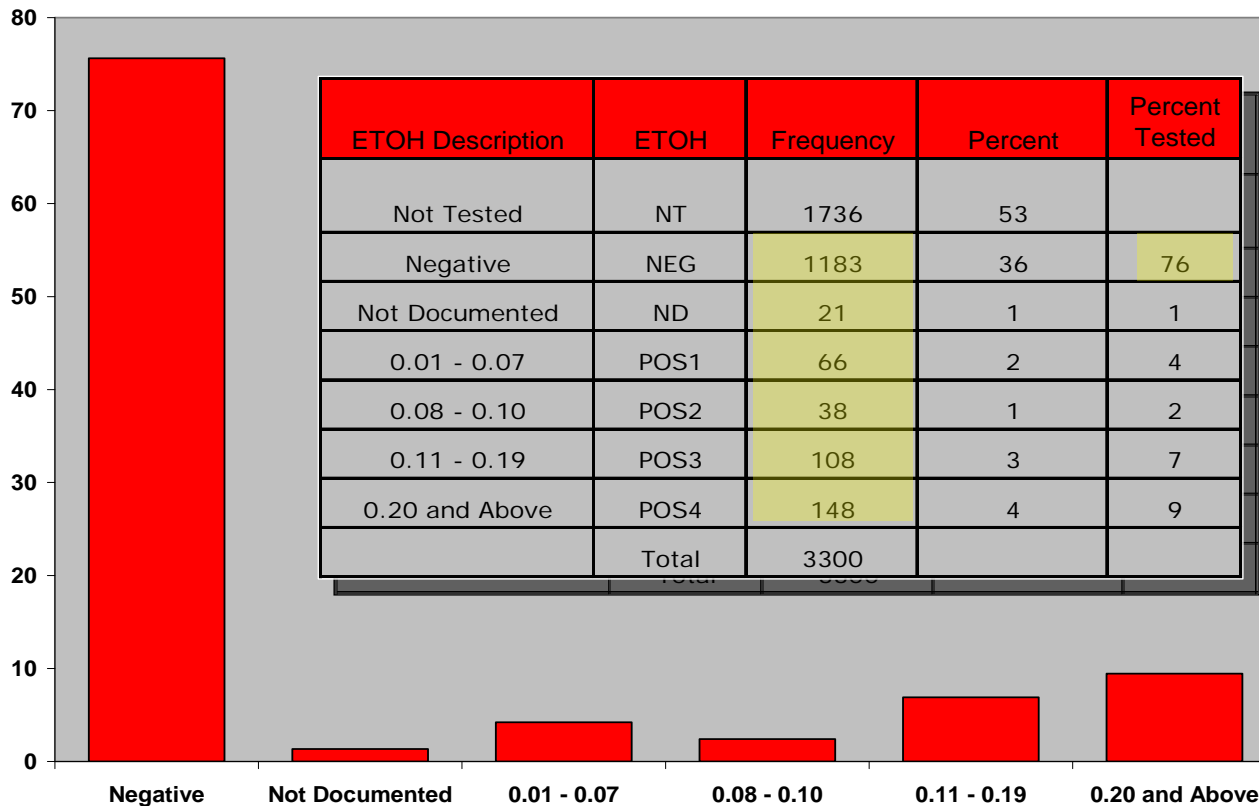


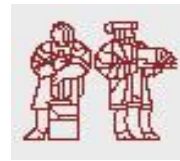
- High level data mining process:
 - ✓ Define what is to be mined... the goal.
 - ✓ Decide on appropriate modeling type, if necessary
 - ✓ **Analyze and prepare data sources**
 - ✓ Conduct data mining
 - ✓ Interpret results
 - ✓ Take action
- Data quality is critical!
- **Enterprises that deploy data mining without first understanding their data run the risk of being seriously misguided**



Data Challenges

- **Impact of poorly captured data**
- From a study of events captured in a trauma registry
Conclusion: **76% tested negative for ETOH**





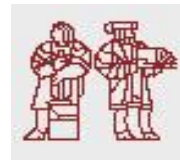
Data was an issue

- There were *actually* 3,818 cases (not 3,300)
 - 518 cases had incorrectly recorded ETOH value
 - ETOH should be 1 of 7 values, instead found 135 values

20 MOST FREQUENT VALUES (ALL VALUES)				
135 UNIQUE VALUES				
NUMBER	VALUE	COUNT	% COUNT	CUMULATIVE % COUNT
1	NT	1,736	45.47%	45.47%
2	NEG	1,183	30.98%	76.45%
3	[empty]	348	9.11%	85.57%
4	POS4	148	3.88%	89.44%
5	POS3	108	2.83%	92.27%
6	POS1	66	1.73%	94.00%
7	POS2	38	1.00%	95.00%
8	ND	21	0.55%	95.55%
9	NT#POS1#NEG#RNA#ND#NT#POS1#POS2#POS3#POS4#RNA	7	0.18%	95.73%
10	24	4	0.10%	95.84%
11	RNA	4	0.10%	95.94%
12	224	3	0.08%	96.02%
13	175	3	0.08%	96.10%
14	67	3	0.08%	96.18%
15	Y#NEG#RNA###NT##ND#NT#POS1#POS2#POS3#POS4	3	0.08%	96.25%
16	204	3	0.08%	96.33%
17	130	2	0.05%	96.39%
18	397	2	0.05%	96.44%
19	215	2	0.05%	96.49%
20	167	2	0.05%	96.54%

Impact: Instead of 76% being negative, it is actually **57%**

(Excludes not-tested, includes incorrect cases)

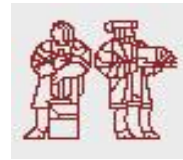


Data Challenges

- Impact of disconnected systems/stores
- Frequency of occurrence of patient safety adverse events

NUMBER	VALUE	COUNT	% COUNT	CUMULATIVE%
1	Emergency	461,009	57.74%	57.74%
2	Service Delays	136,325	17.08%	74.82%
3	Medical	34,909	4.37%	79.19%
4	Surgical	27,823	3.48%	82.68%
5	Maternal	27,723	3.47%	86.15%
6	Medication Errors	22,962	2.88%	89.02%
7	Laboratory	17,347	2.17%	91.20%
8	Service Feedback	15,466	1.94%	93.13%
9	Patient Falls	12,875	1.61%	94.75%
10	Device Complications	7,805	0.98%	95.72%
11	Nosocomial Infections	7,571	0.95%	96.67%
12	Env Safety / Security	6,586	0.82%	97.50%
13				
14				
15				
16				
17				
18				
19				
20				

- Problem: Cost data not captured or connected to adverse events in information system
- Result: Unable to prioritize actions to achieve best cost/benefit



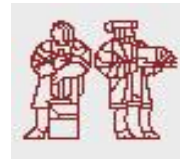
Profiling Data Sources

- Data profiles - establish a *baseline* for data sources
 - Completeness
 - Missing records?
 - Missing fields?
 - Timeliness
 - Is the data current?
 - What is the update nature of the data?
 - Pedigree
 - Is this data *the* master source?
 - What data sources contribute to this?



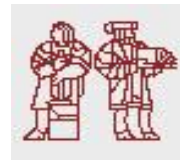
Profiling Data Sources

- Data profiles - establish a *baseline* for data sources
 - Field characteristics
 - Type (string, integer, etc.)
 - Semantic type (date, dollar amount, etc.)
 - Population
 - Shape/distribution
 - High & low values
 - Minimum and maximum length
 - Conformity (normalized, standardized)
 - Composite or *atomic* field?



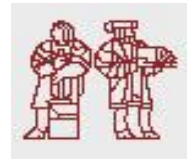
Profiling Data Sources

- Data profiles - establish a *baseline* for data sources
 - Integrity
 - Are there duplicate records?
 - Is this a redundant store?
 - Modifications/Permissions
 - Who can change the data?
 - Are there access restrictions?
 - Storage
 - Where is the data kept?
 - What sort of file structure?



Profiling Data Sources

- Data profiles - establish baseline for data sources
 - Bonus analysis:
 - From an enterprise perspective, document how each data source is linkable to others
 - Determine which fields can serve as foreign keys and ensure their integrity
 - Force linkability among sources, or recognize that isolated sources exist
- *Data baselines are necessary for successful ETL and support effective BI*



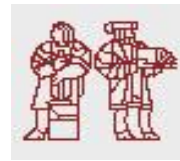
Summary

- Data quality issues are costly, prevalent and becoming more intense
- Establish a data governance policy – enterprise-wide if possible
- Plan ahead to avoid discovering data issues after significant investment has already been made
- Baseline data sources and keep baselines current; know your data
- Building business solutions “from the data up” ensures appropriate focus on data quality

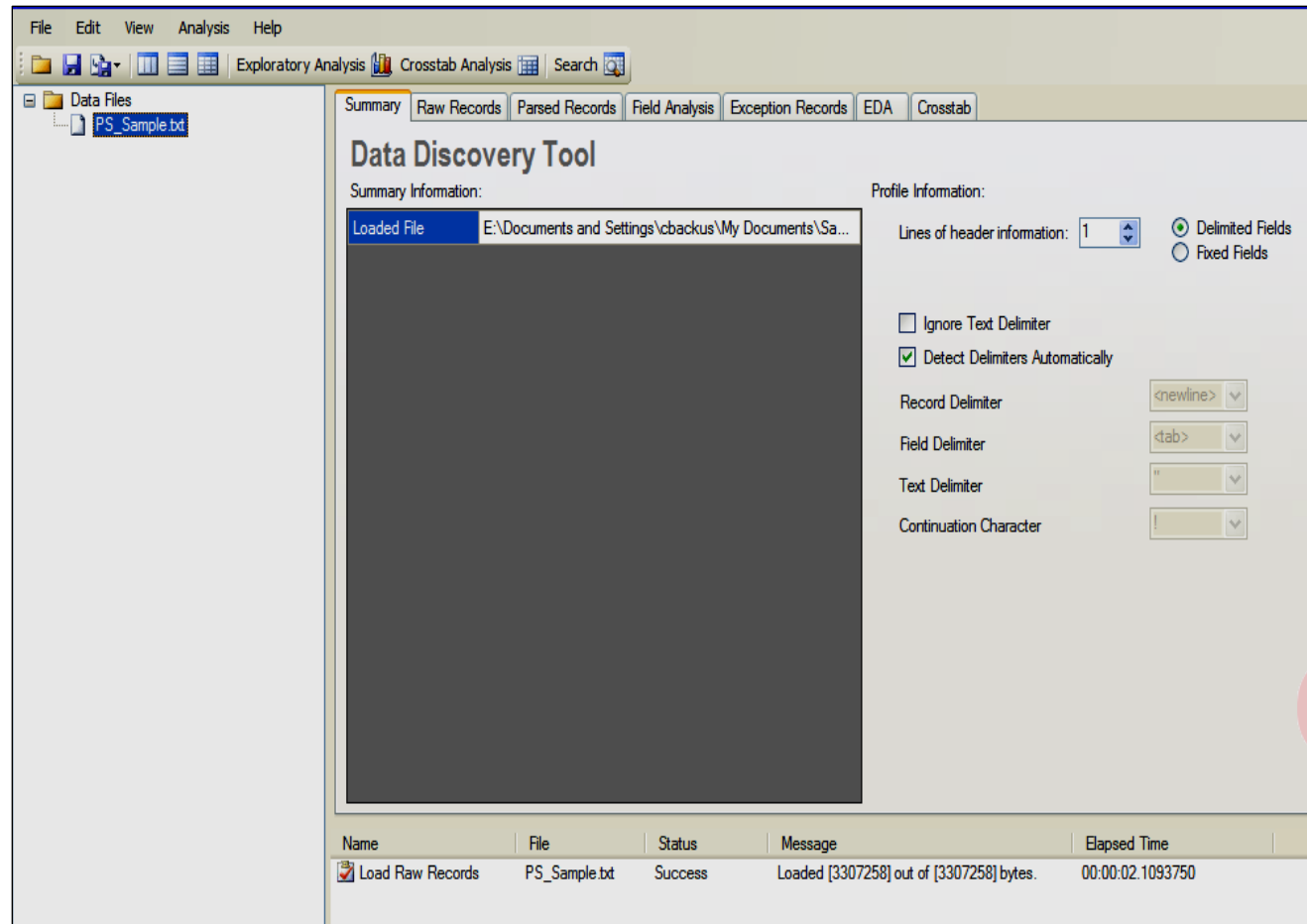


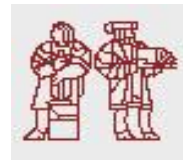
The MIT Information Quality Industry Symposium, 2007

Rapid Data Analysis



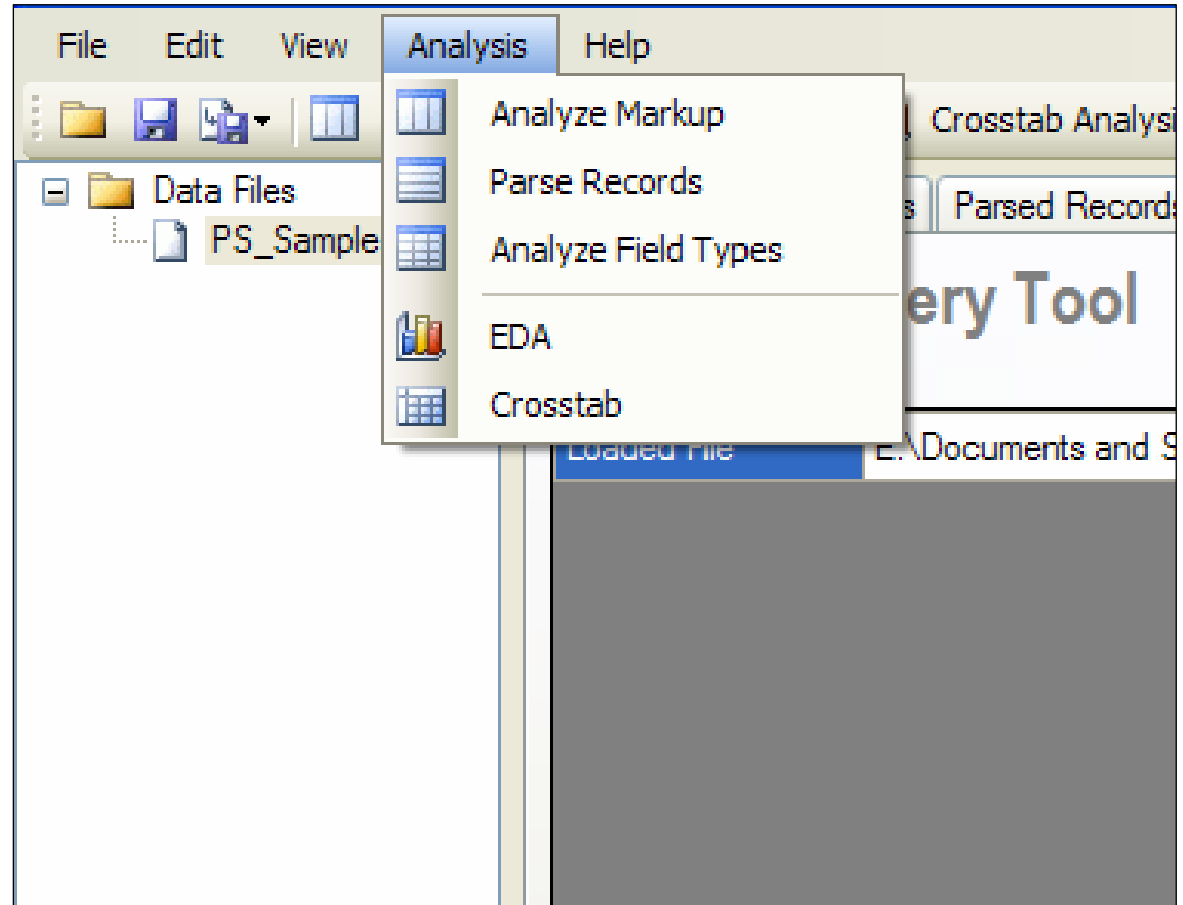
- Data analysis can be achieved quickly and inexpensively
- Qbase uses proprietary Data Discovery Tool (DDT)
- A quick tour...

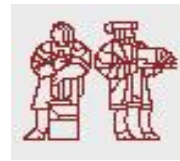




Typical steps include:

- ✓ Analyzing markup
- ✓ Parsing records
- ✓ Analyzing field types
- ✓ Exploratory analysis





Open data file

- ✓ point to file
- ✓ DDT shows *raw data*
- ✓ see every row and column
- ✓ not much fun to look at raw data

The screenshot shows the Qbase software interface. The 'Data Files' pane on the left shows 'PS_Sample.txt' selected. The main window displays the 'Raw Records' view, showing a list of 19 records. Each record consists of a unique identifier, a date, and a description of contributing factors. A status bar at the bottom indicates that the raw records were successfully loaded from the file.

Record ID	Date	Description	Category
1	011/7/2001	RMEDADMDISPNICUL011/7/200163183CONTRIBUTING FACTORS:	DEL
2	011/9/2001	RMEDDISPERRORPICUL011/9/200164049CONTRIBUTING FACTORS:	D
3	011/10/2001	RMFALLRM1023WL011/10/200164529CONTRIBUTING FACTORS:	DELIN
4	011/20/2001	RMEDDISPERROR3WL011/20/200164064CONTRIBUTING FACTORS:	DE
5	012/5/2001	RMLABLAB104EDL012/5/200164790CONTRIBUTING FACTORS:	DELIN
6	01/3/2002	RMEDDISPERRORIMCUL01/3/200264864CONTRIBUTING FACTORS:	DE
7	01/8/2002	RMPROCDELAY3WL01/8/200264768CONTRIBUTING FACTORS:	DELIN
8	03/15/2002	RMPROCDELAYORL03/15/200268756CONTRIBUTING FACTORS:	DELIN
9	03/19/2002	RMPROCORDERTESTHOL03/19/200269594CONTRIBUTING FACTORS:	DE
10	03/23/2002	RMEDADMERRORSURGERYL03/23/200271769CONTRIBUTING FACTORS:	
11	03/25/2002	RMOTHERRM199SURGERYL03/25/200268811CONTRIBUTING FACTORS:	
12	04/24/2002	RMPROCPROCOTHERLABL04/24/200271318CONTRIBUTING FACTORS:	D
13	05/9/2002	RMPROCAMALABL05/9/200271306CONTRIBUTING FACTORS:	DELIN
14	06/29/2002	RMLABLAB102EDL06/29/200273106CONTRIBUTING FACTORS:	DELIN
15	07/8/2002	RMCONFIDRMCONFIDQRML07/8/200274103CONTRIBUTING FACTORS:	D
16	07/24/2002	RMPROCCOUNTNEEDORL07/24/200274014CONTRIBUTING FACTORS:	DE
17	08/6/2002	RMEQUIPEQUIPERRORAHUL08/6/200274929CONTRIBUTING FACTORS:	
18	08/13/2002	RMEDADMDISPDIETL08/13/200274870CONTRIBUTING FACTORS:	DEL
19	08/15/2002	RMEDADMMONEDL08/15/200274877CONTRIBUTING FACTORS:	DELIN

Status Bar: Load Raw Records | PS_Sample.txt | Success | Loaded [3307258] out of [3307258] bytes. | 00:00:02.10s



Analyzing markup

- ✓ detect file structure
- ✓ use layout for fixed-field
- ✓ list number of fields per record

The screenshot shows the Data Discovery Tool interface. The menu bar includes File, Edit, View, Analysis, and Help. The toolbar contains icons for file operations and analysis functions like Exploratory Analysis, Crosstab Analysis, and Search. Below the toolbar are tabs for Summary, Raw Records, Parsed Records, Field Analysis, Exception Records, EDA, and Crosstab. The main window displays the title "Data Discovery Tool" and "Summary Information:" followed by a table of file details.

Property	Value
Loaded File	E:\Documents and Settings\cbackus\My Documents\Samples\PS_Sample.txt
Header Lines	1
Field Delimiter	<tab>
Text Delimiter	<none>
Fields per Record	84
Loaded Schema	E:\Program Files\Qbase\DataDiscoveryTool\Schema\ECreditSchema.xml



Parse Records

- ✓ use header if provided
- ✓ organize data for easy browse
- ✓ all columns and rows
- ✓ sortable columns

Exclude	EVENT_GRP	EVENT_CD	LOCATION	SEVERITY	DATE	REVIEW_ID
<input checked="" type="checkbox"/>	RMMED	ADMDISP	NICU	L0	11/7/2001	63183
<input type="checkbox"/>	RMMED	DISPERROR	PICU	L0	11/9/2001	64049
<input type="checkbox"/>	RMFALL	RM102	3W	L0	11/10/2001	64529
<input type="checkbox"/>	RMMED	DISPERROR	3W	L0	11/20/2001	64064
<input type="checkbox"/>	RMLAB	LAB104	ED	L0	12/5/2001	64790
<input type="checkbox"/>	RMMED	DISPERROR	IMCU	L0	1/3/2002	64864
<input type="checkbox"/>	RMPROC	DELAY	3W	L0	1/8/2002	64768
<input type="checkbox"/>	RMPROC	DELAY	OR	L0	3/15/2002	68756
<input type="checkbox"/>	RMPROC	ORDERTEST	HO	L0	3/19/2002	69594
<input type="checkbox"/>	RMMED	ADMERROR	SURGERY	L0	3/23/2002	71769
<input type="checkbox"/>	RMOTHER	RM199	SURGERY	L0	3/25/2002	68811
<input type="checkbox"/>	RMPROC	PROCOTHER	LAB	L0	4/24/2002	71318
<input type="checkbox"/>	RMPROC	AMA	LAB	L0	5/9/2002	71306



The MIT Information Quality Industry Symposium, 2007

Rapid Data Analysis



Analyze Fields

- ✓ detect types
- ✓ count nulls
- ✓ count unique values
- ✓ count exceptions
- ✓ min, max field length
- ✓ sortable columns

The screenshot shows the Qbase software interface with a 'Field Analysis' table. The table has columns for Exclude, Field Number, Original Name, Data Type, Semantic Type, Qbase Field Name, Null Values, Unique Values, Unique Exceptions, Shortest Field, and Longest Field. The 'SEVERITY' field (row 3) is highlighted, and a pop-up window titled 'Unique values for SEVERITY' is open, showing a list of values and their frequencies.

Exclude	Field Number	Original Name	Data Type	Semantic Type	Qbase Field Name	Null Values	Unique Values	Unique Exceptions	Shortest Field	Longest Field
<input type="checkbox"/>	0	EVENT_GRP	string	String		22	22	0	5	9
<input type="checkbox"/>	1	EVENT_CD	string	String		22	121	0	2	10
<input type="checkbox"/>	2	LOCATION	string	String		42	68	0	2	10
<input type="checkbox"/>	3	SEVERITY	string	String		181	23	0	2	6
<input type="checkbox"/>	4	DATE	date	Date		22	2307	0	8	10
<input type="checkbox"/>	5	REVIEW_ID	int64	Integer		22	8	0		
<input type="checkbox"/>	6	F7	string	String		22	2	0		
<input type="checkbox"/>	7	CF1	string	String		7163	1	0		
<input type="checkbox"/>	8	CF1A	string	String		7163	1	0		
<input type="checkbox"/>	9	CF2	string	String		8484	5	0		
<input type="checkbox"/>	10	CF2A	string	String		8484	5	0		
<input type="checkbox"/>	11	CF3	string	String		8755	3	0		
<input type="checkbox"/>	12	CF3A	string	String		8755	3	0		
<input type="checkbox"/>	13	CF4	string	String		8833	2	0		
<input type="checkbox"/>	14	CF4A	string	String		8833	2	0		
<input type="checkbox"/>	15	CF5	string	String		8858	7	0		
<input type="checkbox"/>	16	CF5A	string	String		8858	7	0		
<input type="checkbox"/>	17	CF6	string	String		8866	4	0		
<input type="checkbox"/>	18	CF6A	string	String		8866	4	0	29	43
<input type="checkbox"/>	19	CF7	string	String		8867	3	0	6	8
<input type="checkbox"/>	20	CF7A	string	String		8867	3	0	29	47
<input type="checkbox"/>	21	CF8	string	String		8868	2	0	8	10
<input type="checkbox"/>	22	CF8A	string	String		8868	2	0	8	10

Field Value	Frequency
L3	1974
L1	1570
L4	1009
LEVEL2	733
PRLOW	623
LEVEL3	520
L5	472
L2	275
LEVEL1	272
L0	225



The MIT Information Quality Industry Symposium, 2007

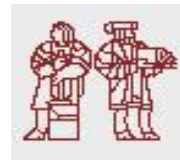


Thank You



The MIT Information Quality Industry Symposium, 2007

Exploratory Data Analysis



File Edit View Analysis Help

Exploratory Analysis Crosstab Analysis Search

Summary Raw Records Parsed Records Field Analysis Exception Records EDA Crosstab

EDA Field Analysis

ADVANCED DATA MANAGEMENT SOLUTIONS

FIELD NAME: [EVENT_GRP] DATA TYPE: [String]

POPULATED	MISSING VALUES	WHITESPACE ONLY	INVALID FORMAT	INVALID VALUES	MINIMUM LENGTH	MAXIMUM LENGTH
8,847	22	0	0	0	5	9
99.75%	0.25%	0.00%	0.00%	0.00%	N/A	N/A

MINIMUM VALUE	MAXIMUM VALUE
RADINCPHY	RMSAFETY

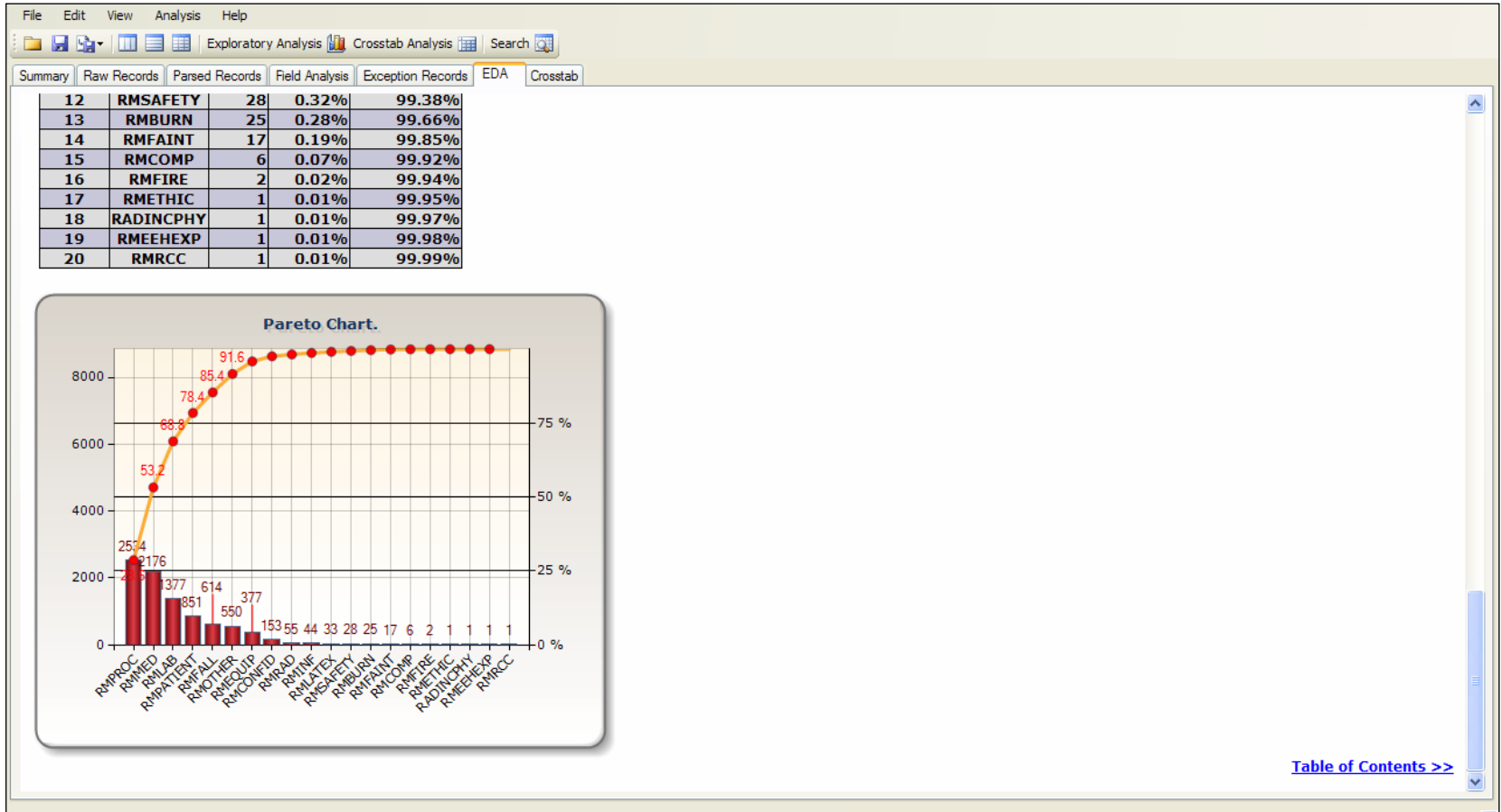
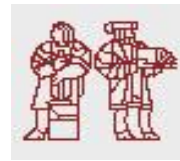
10 SHORTEST VALUES				
NUMBER	VALUE	COUNT	% COUNT	CUMULATIVE % COUNT
1	RMRAD	55	0.62%	0.62%
2	RMLAB	1,377	15.53%	16.15%
3	RMINF	44	0.50%	16.64%
4	RMRC	1	0.01%	16.65%
5	RMED	2,176	24.53%	41.19%
6	RMPROC	2,534	28.57%	69.76%
7	RMBURN	25	0.28%	70.04%
8	RMFIRE	2	0.02%	70.06%
9	RMCOMP	6	0.07%	70.13%
10	RMFALL	614	6.92%	77.05%

10 LONGEST VALUES				
NUMBER	VALUE	COUNT	% COUNT	CUMULATIVE % COUNT
1	RADINCPHY	1	0.01%	0.01%



The MIT Information Quality Industry Symposium, 2007

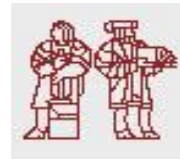
Exploratory Data Analysis





The MIT Information Quality Industry Symposium, 2007

Exploratory Data Analysis



File Edit View Analysis Help

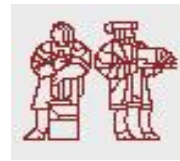
Exploratory Analysis Crosstab Analysis Search

Summary Raw Records Parsed Records Field Analysis Exception Records EDA Crosstab

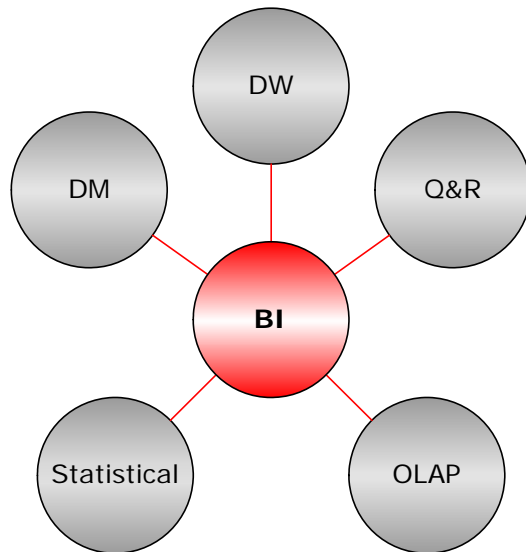
6	PRHIGH	37	0.42%	20.67%
7	PRLOW	623	7.02%	27.69%
8	PRMOD	167	1.88%	29.57%
9	MOD	1	0.01%	29.59%
10	IV1	158	1.78%	31.37%

20 MOST FREQUENT VALUES (ALL VALUES)				
23 UNIQUE VALUES				
NUMBER	VALUE	COUNT	% COUNT	CUMULATIVE % COUNT
1	L3	1,974	22.26%	22.26%
2	L1	1,570	17.70%	39.96%
3	L4	1,009	11.38%	51.34%
4	LEVEL2	733	8.26%	59.60%
5	PRLOW	623	7.02%	66.63%
6	LEVEL3	520	5.86%	72.49%
7	L5	472	5.32%	77.81%
8	L2	275	3.10%	80.91%
9	LEVEL1	272	3.07%	83.98%
10	L0	225	2.54%	86.51%
11	[empty]	181	2.04%	88.56%
12	LEVEL4	179	2.02%	90.57%
13	PRMOD	167	1.88%	92.46%
14	IV2	166	1.87%	94.33%
15	IV1	158	1.78%	96.11%
16	IV3	106	1.20%	97.31%
17	LEVEL5	92	1.04%	98.34%
18	IV4	68	0.77%	99.11%
19	L6	37	0.42%	99.53%
20	PRHIGH	37	0.42%	99.94%

1 MOST FREQUENT VALUES (INVALID VALUES)				
1 UNIQUE VALUES				
NUMBER	VALUE	COUNT	% COUNT	CUMULATIVE % COUNT



- Data is a key, critical asset of an enterprise
- Careful planning drives creation of *strategic information assets...* (think this way!)
- Business Intelligence (BI) - drawing full value from strategic information assets



The BI Umbrella

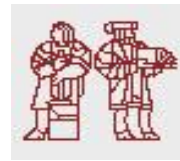
(DW) Data Warehousing

(DM) Data Mining (DM)

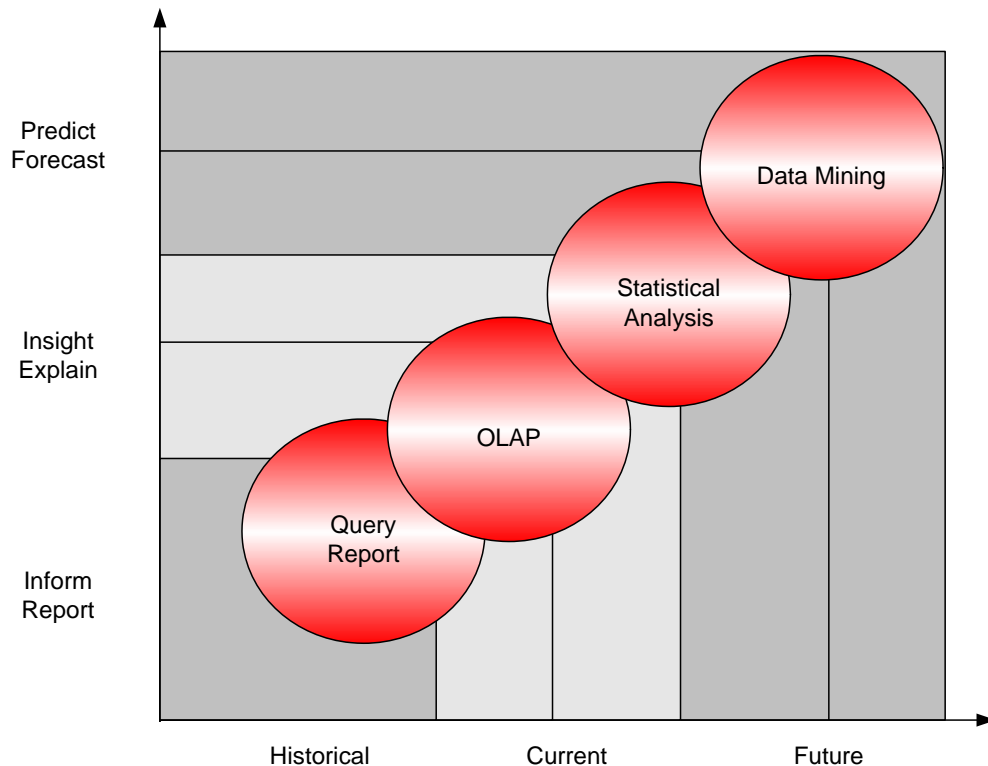
(Q&R) Query and Reporting

(OLAP) On-Line Analytics Processing

Statistical Analysis



- BI provides the complete temporal spectrum: from historical to future perspective



BI useful for:

- Informing and reporting
- Explanation and insight
- Forecasting and predicting