SIFMA Society: IAS Seminar

Analyzing Data to Increase Audit Efficiency

A Look at Data Mining from a Business Perspective

October 25, 2011

PricewaterhouseCoopers

Table of Contents

Section

Introduction

Regulatory and Accounting Analytics Group

We are a PricewaterhouseCoopers specialist team comprised of 150+ technology specialists globally. Our team are cross discipline and cross industry, and mainly provide services to assurance, regulatory, and forensic engagements.

Capabilities	Credentials	Industries
1. Data Mining	1. CPA , CISA, CIA	Banking & Insurance
2. C.A.A.T's	2. MCSD, MCDBA, Oracle DBA	Healthcare and Life Science
3. Continuous Monitoring	3. CFA	Law and Litigation
4. Forensic Investigation	4. CFE	Retail and Consumer
5. Regulatory Analytics	5. MBA & PHD	Technology

Trend is a Continued Focus on Data Mining by Internal Audit

For the last three years (2007-2009), the top three areas of considered by Internal Audit departments to need improvement, from a skillset perspective were*:

1. Computer Assisted Audit Techniques
2. Data Analysis Tools – Statistical and Data Manipulation
3. Continuous Auditing

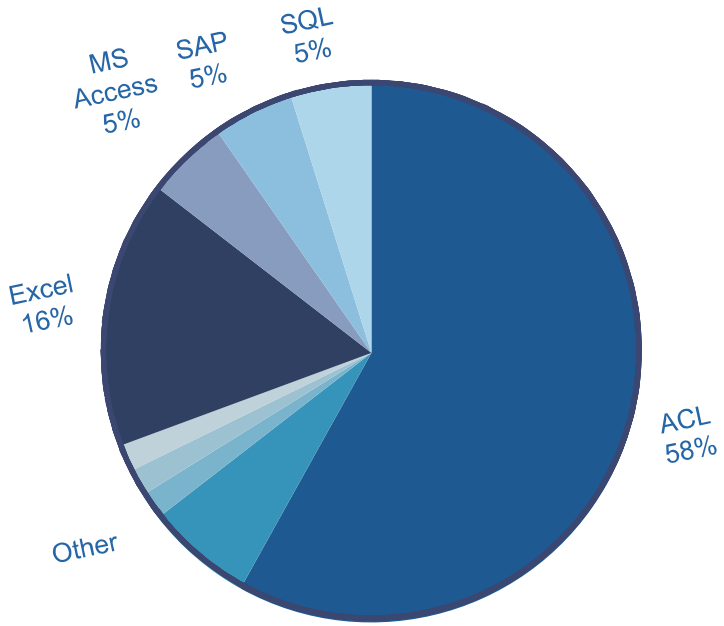
Results were consistent across size of respondent and industry, with the above occupying values being in the top 5 in all sizes\industries.

* Based on Protiviti 2010 Survey

What Software Tools are being used?

Although there is a focus on DM skill sets, the type of tools being used vary. One third of respondents are not using Data Mining in any form.

Do you use software for extraction and analysis?	
Yes	63.0 %
No	32.6 %
Na	4.3 %



Source: Based on GAIN 2009 IT Benchmarking Survey

Data Mining Basics

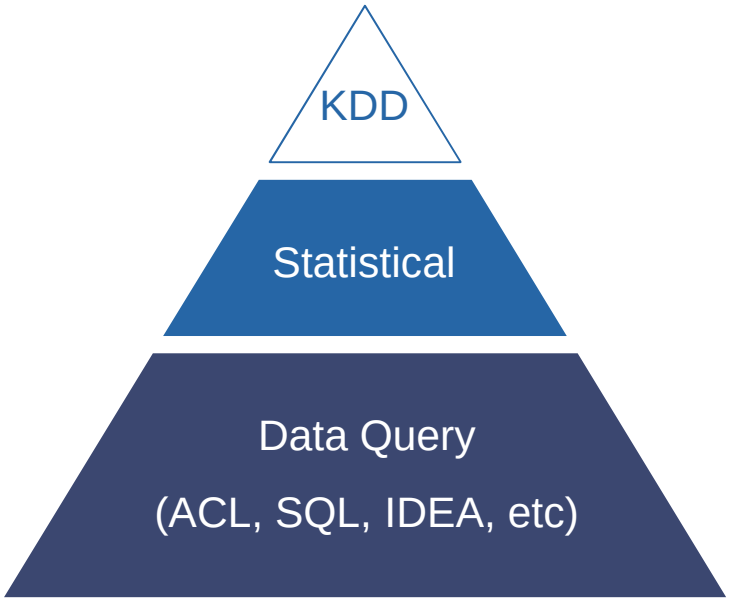
What is Data Mining?

- Computer Assisted Audit Techniques (CAATS) are ways in which the auditor may use a computerized information system to gather or assist in gathering audit evidence
  - Integrated Test Facility (ITF)
  - System Control Audit Review File (SCARF)
  - Snapshot
  - Parallel Simulation
- Data Mining is a subset of CAATS, and is focused on testing and monitoring of risk through the use of data analysis techniques.

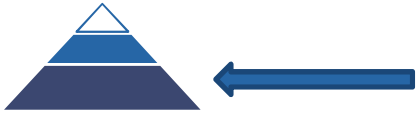
Data Analysis Techniques Pyramid

Data mining can be thought of in three distinct categories, based on the type of question\risk you are addressing. In general terms:

When risk can be defined by an attribute and pattern, which is the majority of testing, data querying is sufficient. Statistical techniques take precendence , when attributes are known, but patterns are not. Knowledge Discovery and Data Mining are used when both are not known.



Data Querying



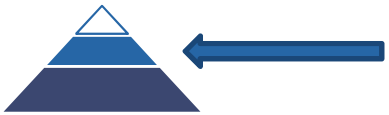
Data Querying is used when the auditor is seeking an answer to a specific question or risk. This is used when both the attribute (a trait or value) and pattern (relationship) is known. Questions such as:

- Do front office personnel have access to my GL?
- What percentage of my level 1 equity have zero\stale price?
- Does my position and balance system reconcile to my custodian(s)?

Example Tools

- ACL, IDEA, MS Excel, Access, RDBMS

Statistical Analysis



Statistical analysis overlaps heavily with data querying, and is best used to obtain information about the population. Using our previous metaphor, this is used when the attribute (a trait or value) is known but when the pattern (relationship) is not known or proven.

Questions such as:

- Is there a correlation of better price or better execution between trades executed through X entity versus Y entity?
- Based on black scholes inputs, what positions are outliers based on attributes those?
- Predicting and establishing thresholds and limits for capacity or trading volume based on ANOVA (analysis of variance)

Example Tools

- SAS, Matlab, RDBMS

Knowledge Discovery and Data Mining

KDD is used when neither the risk is known, but neither the attribute nor pattern is known. Technically, KDD techniques use some form of machine or automation to discover patterns within the data. It is the least mentioned in IA literature, and techniques designed are proprietary to the IA departments using them.

Questions\Objectives such as:

- What are potentially fraudulent entries in my general ledger?
- Identify duplicate payments to vendors with similar name and addresses
- Identify credit card transactions that pattern wise, have a high fraud rate

Example Tools

- IBM Data Miner, Clementine, TBD

Key Benefits and Challenges for using Data Mining

Benefits

- Remove sampling risk by gaining audit coverage over100% of population
- Increase independence from Information system functions (developers, technology administration)
- Decreased cost over time due to reusability
- Provide real-time audit opinions
- Efficiently test certain assertions such as completeness and accuracy

Challenges

- First year costs can be higher than a manual test
- Auditor needs a strong understanding of the operation, financial or technology supporting the data
- Perception by information providers that data cannot be extracted or used.

Acquiring Data

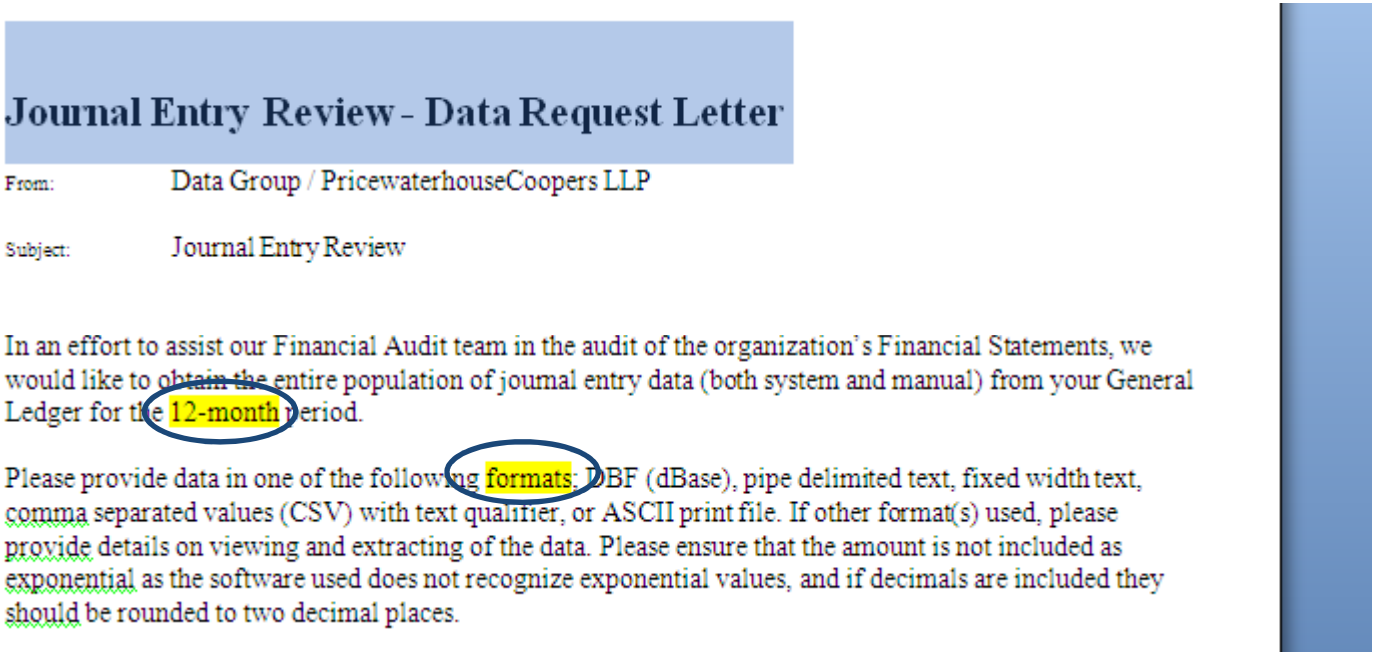
Data Acquisition – Acquiring Data

The data request, more than any other single factor, determine the success or failure of a data mining test.

Before making the request, consider the following:

- Source – Report vs Database vs Extract
- Time period
- Format
- Size of Request
- Target Platform
- Confidentiality
- Control Total\Validation Reports
- Method of Transfer

Data Acquisition – Sample Request



Journal Entry Review - Data Request Letter

From: Data Group / PricewaterhouseCoopers LLP

Subject: Journal Entry Review

In an effort to assist our Financial Audit team in the audit of the organization's Financial Statements, we would like to obtain the entire population of journal entry data (both system and manual) from your General Ledger for the 12-month period.

Please provide data in one of the following formats: DBF (dBase), pipe delimited text, fixed width text, comma separated values (CSV) with text qualifier, or ASCII print file. If other format(s) used, please provide details on viewing and extracting of the data. Please ensure that the amount is not included as exponential as the software used does not recognize exponential values, and if decimals are included they should be rounded to two decimal places.

Data Acquisition – Sample Request

In addition, please provide a response for the following regarding the organization's General Ledger system:

General Ledger System (i.e., Lawson, PeopleSoft, etc.):	
Can a one sided journal entry be posted within the General Ledger:	
Are there statistical accounts included in the General Ledger and how do we identify them:	
What is the numerical account range for the Assets, Liabilities, Equity, Revenue, and Expense account numbers within the General Ledger:	
Can a Journal Entry Number be reused within the General Ledger? If so, what makes a Journal Entry Number unique (i.e. journal number + location code + date):	
How can one identify a journal entry that is automated vs. manual:	
Is there any additional information we should know regarding the General Ledger and our ability to perform analysis on the data requested:	
Contact information of the IT person, if any clarification is needed:	Name: Phone: Email:

Data Acquisition – Sample Request

In summary, we should receive a total of five files:

1) Journal entries (line item level) containing all fields listed above (01/01/07 through 12/31/07)
2) Trial balance @ 01/01/07 containing all fields listed above.
3) Trial balance @ 12/31/07 containing all fields listed above.
4) Record count/Control total, if available.
5) General ledger roll up (mapping of accounts to the financial statement)

Data Acquisition – Plain Text - Common Formats

Generally, ASCII text is a universal common denominator for most systems*. Basic options include:

- Fixed Width vs Delimited
- Text Qualifier

```
Delimited by pipe (|)

"Cheng"|"Glenn"|"300 Madison"|$39.99
```

However, if the source system can target your analysis platform, then that can save you time and effort

* Mainframe can present unique obstacles due to packed fields, and data conversion issues.

3 – Acquisition and Analysis Techniques

Data File Formats - Delimited

Special characters called delimiters are used to separate each field. Common characters are: comma, tab, semicolon, carat (^), tilde (~), and pipe (|). The character used as the delimiter cannot be in the data itself. Fields that were data type character in the source system are often in quotes while numeric fields are not.

```
1222991,"LANC","$HL 0426","265047",20040623,47102,"PRIMUS AUTOMOTIVE/LANDROVER",31.83,35.01,"RADIATOR HOSE",1,"L",350,"M"
1222991,"LANC","$HL 0426","265047",20040623,47102,"PRIMUS AUTOMOTIVE/LANDROVER",3.55,3.91,"FREIGHT CHARGE",1,"L",350,"M"
1936824,"MBC","$MU 0527","313956",20040526,15895,"MERCEDES-BENZ CREDIT",5.00,5.00,"SENSOR",1,"L",350,"C"
1936824,"MBC","$MU 0527","313956",20040526,15895,"MERCEDES-BENZ CREDIT",59.00,59.00,"SHOE",1,"L",350,"C"
1936824,"MBC","$MU 0527","313956",20040526,15895,"MERCEDES-BENZ CREDIT",85.00,85.00,"DISC",1,"L",350,"C"
2427163,"HANN","%MC 6/29","560186",20040622,31635,"HANN FINANCIAL SERVICE CORP",20.00,24.00,"BATTERY",1,"L",350,""
3731612,"BMW","%KP 0316","F05324",20040312,39983,"BMW PICK UP",6.09,7.00,"VALVES",1,"L",350,"P"
3731612,"BMW","%KP 0316","F05324",20040312,39983,"BMW PICK UP",9.00,10.35,"HUBS",1,"L",350,"P"
```

Data File Formats – Fixed

Every field has a predetermined length. Fixed length files must be accompanied by a record layout indicating the start position and length for each field. Headers and other information can be included in the data file as long as the detail lines are clearly recognizable.

```
C100100000001 2004SA    11/17/200411/17/2004    11EBARIUS   FB50                USD 001    S              82.25 USD
C100100000001 2004SA    11/17/200411/17/2004    11EBARIUS   FB50                USD 002    H              82.25 USD
C100100000176 2004ZR    12/15/200412/15/2004    12RLANCAST  FBD5                USD 002    H         334,416.66 USD
C100100000177 2004ZR    12/15/200412/15/2004    12RLANCAST  FBD5                USD 001    S          63,775.00 USD
C100100000177 2004ZR    12/15/200412/15/2004    12RLANCAST  FBD5                USD 002    H          63,775.00 USD
C100100000178 2004ZR    12/15/200412/15/2004    12RLANCAST  FBD5                USD 001    S          47,233.34 USD
C100100000178 2004ZR    12/15/200412/15/2004    12RLANCAST  FBD5                USD 002    H          47,233.34 USD
C100100000179 2004ZR    12/15/200412/15/2004    12RLANCAST  FBD5                USD 001    S          43,716.66 USD
C100100000179 2004ZR    12/15/200412/15/2004    12RLANCAST  FBD5                USD 002    H          43,716.66 USD
C100100000180 2004ZR    12/15/200412/15/2004    12RLANCAST  FBD5                USD 001    S          16,308.35 USD
C100100000334 2004AB    12/31/200412/31/2004    12RLANCAST  FB08                USD 024    S           3,397.00 USD
C100100000334 2004AB    12/31/200412/31/2004    12RLANCAST  FB08                USD 025    H           9,040.00 USD
```

Data File Formats – Flat/Report/Print

Also called a “print to unconverted text” or “report spooled to file ” or “an ascii report”. A flat file usually includes headers, footers, page numbers, etc. The data in these files is often overlapping to the point that a special utility such as Monarch must be used to get the data out.

```
DAILY04                                                MANHEIM AUTO AUCTION
                                                  PRINT GENERAL LEDGER TRIAL BALANCE
                                                     POSTING YEAR: 2004   PERIOD#:  06
                                                              7/12/2004
ACCT     ACCT DESCRIPTION                    Y-T-D THIS YEAR                Y-T-D LAST YEAR
----     ------------------------------      ---------------                ---------------
1004     CASH IN BANK - GE PAYROLL                    740.30                      18,921.19
1005     CASH IN BANK - DEPOSITORY             18,129,669.01                  51,874,694.84
1006     CASH IN BANK - PYMT TO CONSIGNORS     12,480,351.02-                 30,696,881.97-
1007     CASH IN BANK - GEN'L DISBURSEMENT      3,831,589.58-                  4,615,673.40-
1010     CASH ON HAND                              31,984.00                      26,826.00
                                               ---------------                ---------------
                                                1,850,452.71                  16,607,886.66
```

Data Acquisition – Sample Data Extract

Fixed Width

Intelligent Field

```
          0        10        20        30        40        50        60        70        80        90        100
 1 3400775  0      USD       NONE      0         H         N         Tek  [DTS.ANY.LIUSD2MD.TEL], MDSAPI warning [JPMTIB_
 2 4601783  0      USD       NONE      0         H         N         Tek  [DTS.ANY.LIUSD5MD.TEL], MDSAPI warning [JPMTIB_
 3 1846106  1      EUR       NONE      0         H         N         Tek  [PB.ANY.IDAY_RFJPEUEE.contrib], MDSAPI warning
 4 1848512  1      EUR       NONE      0         H         N         Tek  [PB.ANY.IDAY_RFJPEUBE.contrib], MDSAPI warning
 5 1848514  1      GBP       NONE      0         H         N         Tek  [PB.ANY.IDAY_RFJPUKEE.contrib], MDSAPI warning
 6 1848515  1      GBP       NONE      0         H         N         Tek  [PB.ANY.IDAY_RFJPUKBE.contrib], MDSAPI warning
 7 1848526  1      JPY       NONE      0         H         N         Tek  [PB.ANY.IDAY_RFJPJPBE.contrib], MDSAPI warning
 8 1848527  1      USD       NONE      0         H         N         Tek  [PB.ANY.IDAY_RFJPUSEE.contrib], MDSAPI warning
 9 1848528  1      USD       NONE      0         H         N         Tek  [PB.ANY.IDAY_RFJPUSBE.contrib], MDSAPI warning
10 1848533  1      USD       NONE      0         H         H         Mark Price Is Zero      0         0         0         0
11 3419935  1      USD       NONE      0         H         N         Tek  [IDN_RDF.ANY.DDNF.N], MDSAPI warning [JPMTIB_ST
12 3601214  1      USD       NONE      0         H         N         Tek  [IDN_RDF.ANY.CN.N], MDSAPI warning [JPMTIB_STAT
13 3624642  1      USD       NONE      0         H         N         Tek  [IDN_RDF.ANY.AFRE.N], MDSAPI warning [JPMTIB_ST
14 35085780         1         USD       NONE      0         S         S       OK        1.050000        0         0         1.0C
15 13058    2      USD       NONE      0         H         N         Tek  [IDN_RDF.ANY.TDS.A], MDSAPI warning [JPMTIB_STA
16 18474    2      USD       NONE      0         S         S         OK       38.660000        38.580000        38.680000
17 3413803  2      USD       NONE      0         S         S         OK        1.050000         1.000000         1.050000
18 3414050  2      USD       NONE      0         S         S         OK        1.300000         1.280000         1.310000
19 3414286  2      USD       NONE      0         S         S         OK       36.110000        36.100000        36.170000
20 3414306  2      USD       NONE      0         H         N         Tek  [IDN_RDF.ANY.BL.A], MDSAPI warning [JPMTIB_STAT
21 3433238  2      USD       NONE      0         S         S         OK       11.120000        11.020000        11.150000
22 3433616  2      USD       NONE      0         S         S         OK        4.520000         4.420000         4.550000
23 3441943  2      USD       NONE      0         S         S         OK        0.460000         0.450000         0.460000
24 3451328  2      USD       NONE      0         H         N         Tek  [IDN_RDF.ANY.JPI_pa.A], MDSAPI warning [JPMTIB_
25 3451606  2      USD       NONE      0         S         S         OK        0.510000         0.510000         0.540000
26 3455051  2      USD       NONE      0         S         S         OK        6.850000         6.850000         6.890000
27 3460499  2      USD       NONE      0         S         S         OK        1.850000         1.830000         1.870000
28 3463611  2      USD       NONE      0         S         S         OK        0.900000         0.880000         0.940000
29 3463919  2      USD       NONE      0         S         S         OK        3.500000         3.480000         3.510000
30 3480344  2      USD       NONE      0         S         S         OK      118.990000       118.950000       119.060000
31 3505176  2      USD       NONE      0         S         H         Out Of Tolerance (0.1) [ 2.142 < 3.75 < 2.618 ] 3.75
32 3505182  2      USD       NONE      0         S         S         OK        4.250000         4.150000         4.300000
33 3518981  2      USD       NONE      0         H         N         Tek  [IDN_RDF.ANY.ERSO.A], MDSAPI warning [JPMTIB_ST
34 3667993  2      USD       NONE      0         S         S         OK        0.140000         0.130000         0.140000
35 3672504  2      USD       NONE      0         S         H         Out Of Tolerance (0.1) [ 0.261 < 0.23 < 0.319 ] 0.23
36 3673261  2      USD       NONE      0         S         S         OK       20.020000        20.010000        20.070000
37 3675835  2      USD       NONE      0         S         S         OK        7.810000         7.760000         7.820000
38 3676350  2      USD       NONE      0         S         S         Stale Price IDN_RDF.ANY.GVP.A   5.480000         5.45
39 3678422  2      USD       NONE      0         S         S         OK        3.930000         3.920000         3.940000
40 3678901  2      USD       NONE      0         S         S         OK        9.960000         9.870000         9.980000
41 3678991  2      USD       NONE      0         S         S         OK        6.290000         6.290000         6.310000
```

Data Acquisition – Sample Report – Structured Data

PARTICIPANT: XXX THE DEPOSITORY TRUST COMPANY PAGE 1
XXX PARTICIPANT POSITION STATEMENT -- 12/31/09 XXX

G8259U 10 7 GPFDREGSSOLSTIC BE-
CHILLS - COD W/T DEPOSIT
LAST ACTIVE 10/15/08

ACCOUNT SUMMARY - SDFS EQUITY - CLOSING BALANCES

(010) **TOTAL** 7,000

16117M 13 1 WTSCHTER113014 BE+
CHILLS - COD W/T DEPOSIT
LAST ACTIVE 12/09/09

ACCOUNT SUMMARY - SDFS EQUITY - CLOSING BALANCES

(010) **TOTAL** 46,736

05873K 11 6 WTSBALLY083114 BE+
CHILLS - COD W/T DEPOSIT
LAST ACTIVE 12/18/09

ACCOUNT SUMMARY - SDFS EQUITY - CLOSING BALANCES

(010) **TOTAL** 4,667

16117M 30 5 CHARTERCOMMUNIC BE+
CHILLS - COD W/T DEPOSIT
LAST ACTIVE 12/18/09

ACCOUNT SUMMARY - SDFS EQUITY - CLOSING BALANCES

(010) **TOTAL** 166,447

05873K 20 7 BALLY TOTAL FIT BE+
CHILLS - COD W/T DEPOSIT
LAST ACTIVE 12/18/09

ACCOUNT SUMMARY - SDFS EQUITY - CLOSING BALANCES

218681 10 4 CORE-MARK HLDG CO.
LAST ACTIVE 12/31/09

ACCOUNT SUMMARY - SDFS EQUITY - CLOSING BALANCES

(010) **TOTAL** 723

Data Acquisition - Monarch

Monarch is Windows-based ‘Report Mining’ software that easily extracts data from existing reports produced by any information system, along with easy data analysis, graphing, and exporting of data to other applications such as Excel and Access.” (monarch.datawatch.com)

Monarch template: used to extract fields from a report. Monarch provides four template types:

1 - Detail

2 - Append (similar to Header in ACL)

3 - Footer

4 - Page Header

Monarch Trap: Combination of the templates below can be saved as a single Trap that can be reused on a similar file.

Data Acquisition – Acquiring Data

Other considerations:

- Repeatability of tests – structure extract to be repeatable
- Impact on production environment, especially transaction systems
- Effort required to obtain data
- Accessing read-only version of production

Analysis Techniques

Data Analysis– Analysis Techniques

Data Mining provides a means of adjusting the auditors initial approach and react to findings real time.

Basic Steps

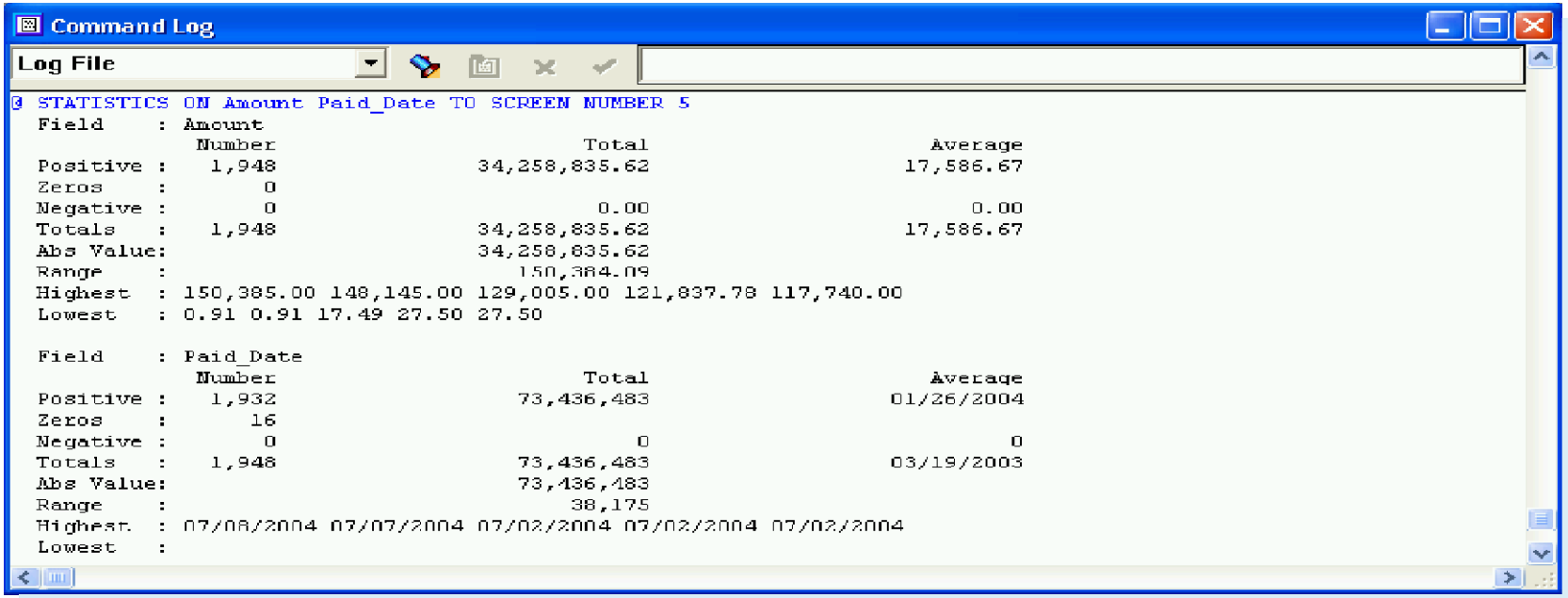
We will use ACL as an example, but can be replicated with almost all analysis and Database tools.

- Statistics \ Profiling
- Stratification
- Summarization
- Pivot Tables & OLAP
- Sampling

Data Analysis– Analysis Techniques

The STATISTICS command generates simple descriptive statistics for numeric fields and date fields.

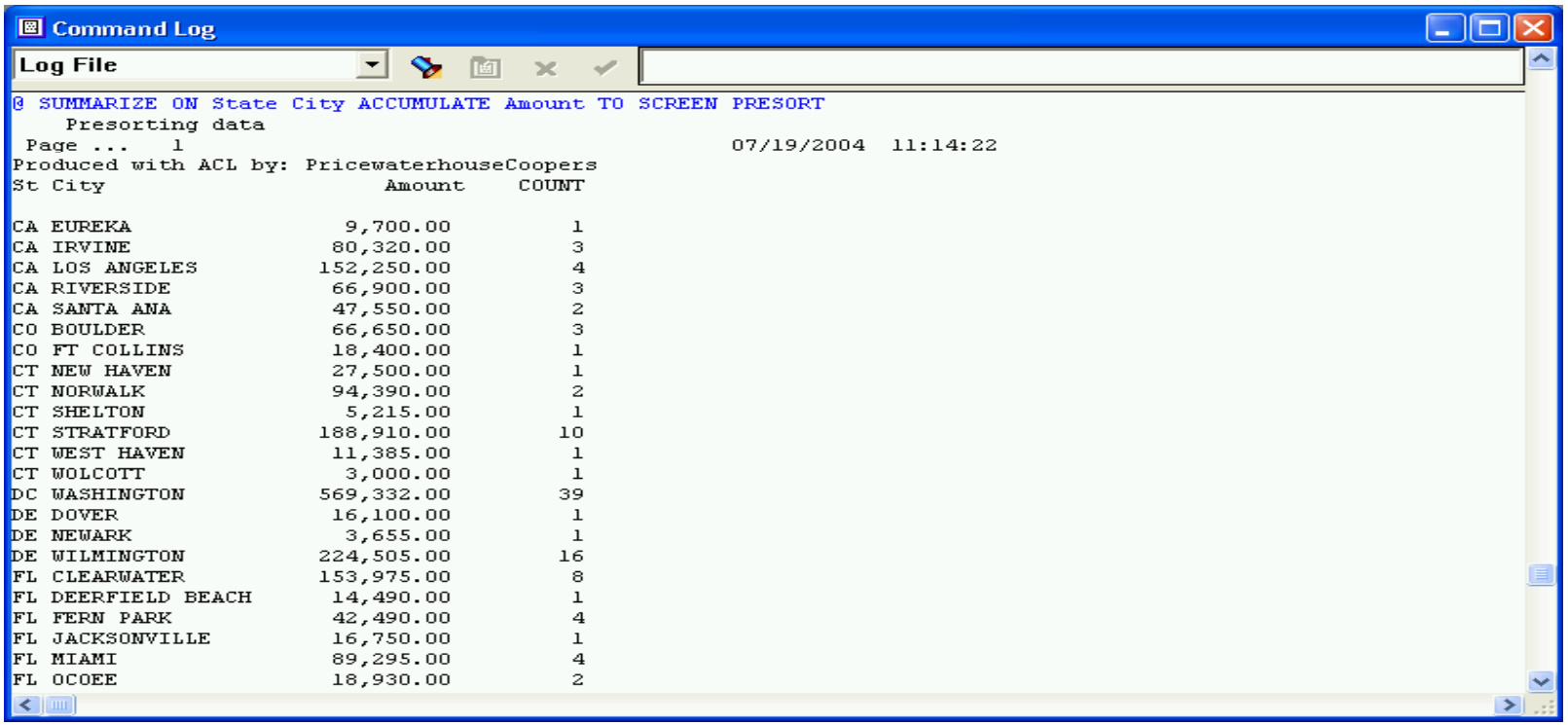
Tools → Analyze → Statistics



Data Analysis– Analysis Techniques

The SUMMARIZE function is very similar to the GROUP BY clause in a SQL statement. It aggregates a value, by a designated field.

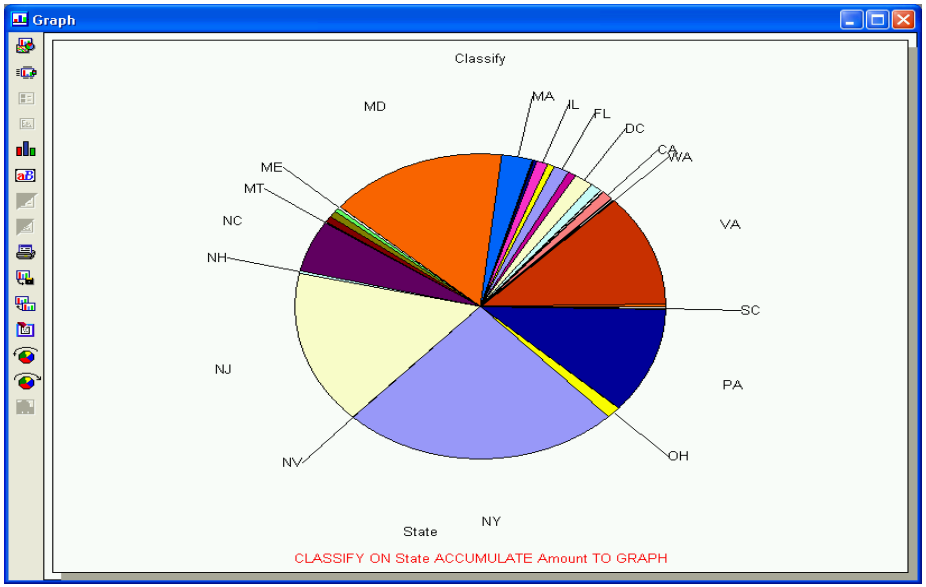
Tools → Data → Summarize

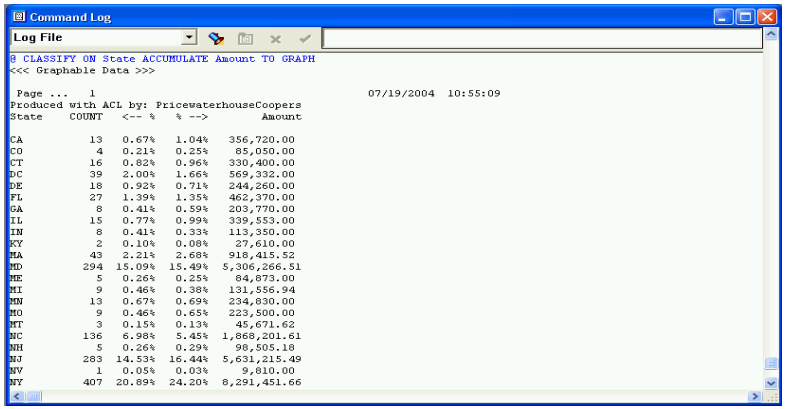


Data Analysis– Analysis Techniques

The CLASSIFY command is very similar to the SUMMARIZE command, but can be used to summarize data on one key field only. It can, however accumulate many numeric fields, just like SUMMARIZE.

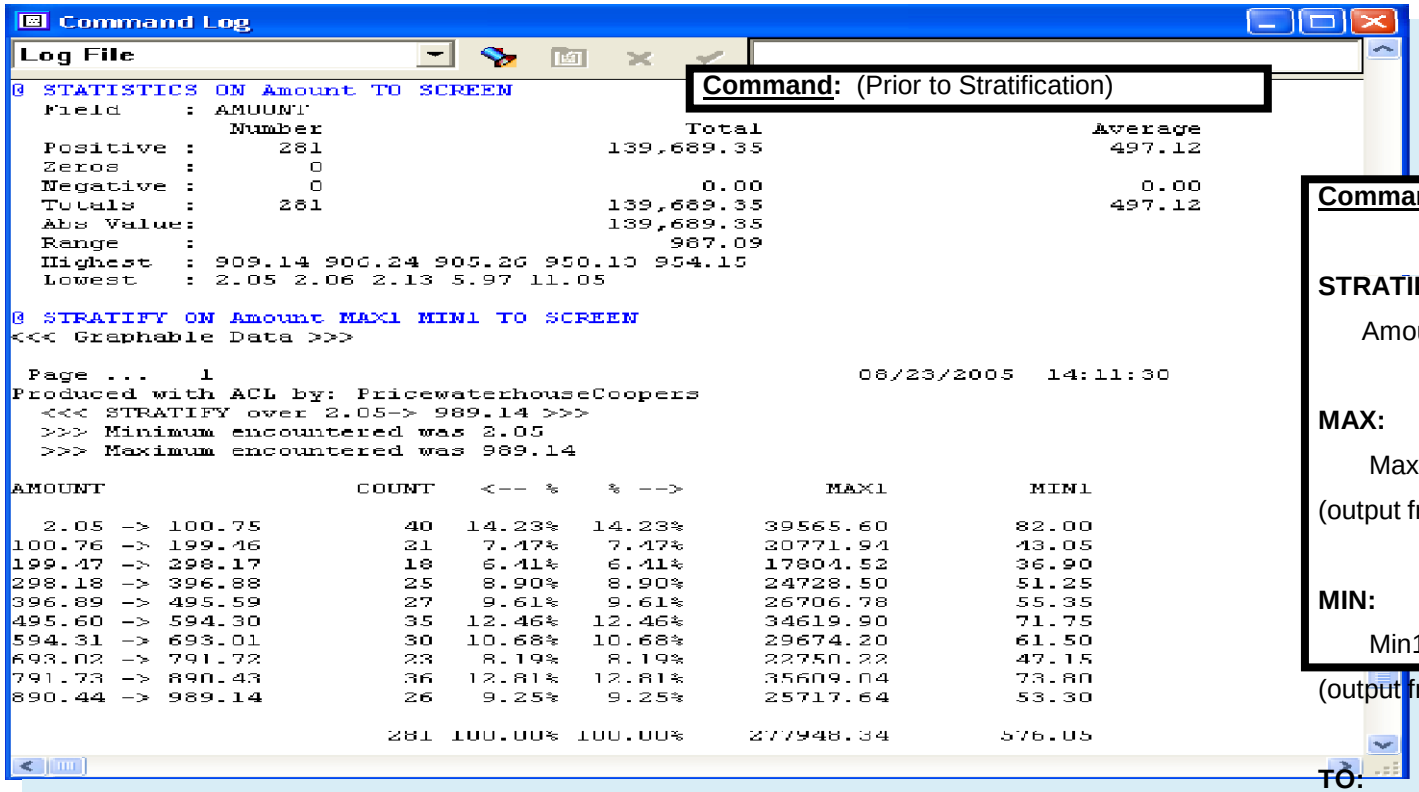
Tools → Data → Classify





Data Analysis– Analysis Techniques

The STRATIFY command summarizes numeric fields into specified intervals (or buckets), and accumulates one or more numeric fields for each interval.



Data Analysis– Analysis Techniques

The AGE command is used to produce aged summaries of the input data file. Numeric fields can be accumulated for each age interval.

Tools → Analyze → Age

Command Log

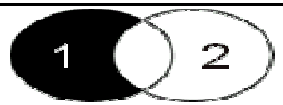
Log File

```
@ AGE ON Paid_Date CUTOFF 20040630 INTERVAL 0,30,60,90,120,10000 ACCUMULATE Amount TO GRAPH
<<< Graphable Data >>>

 Page ...   1                                          07/19/2004  10:46:12
Produced with ACL by: PricewaterhouseCoopers
  <<< AGE over 0-> 10,000 >>>
  >>> Minimum encountered was -8
  >>> Maximum encountered was 38,167

Paid_Date                     COUNT   <-- %   % -->        Amount

< 0                              10   0.51%   0.89%    306,342.02
     0 ->     29                 49   2.52%   3.61%  1,237,969.75
    30 ->     59                 40   2.05%   1.81%    619,061.00
    60 ->     89                 38   1.95%   2.05%    701,363.57
    90 ->    119                 78   4.00%   5.72%  1,959,201.14
   120 -> 10,000               1717  88.14%  85.28%  29,215,176.32
> 10,000                         16   0.82%   0.64%    219,721.82

                                1948 100.00% 100.00%  34,258,835.62
```

Data Analysis– Analysis Techniques

Working with Multiple Data Sets

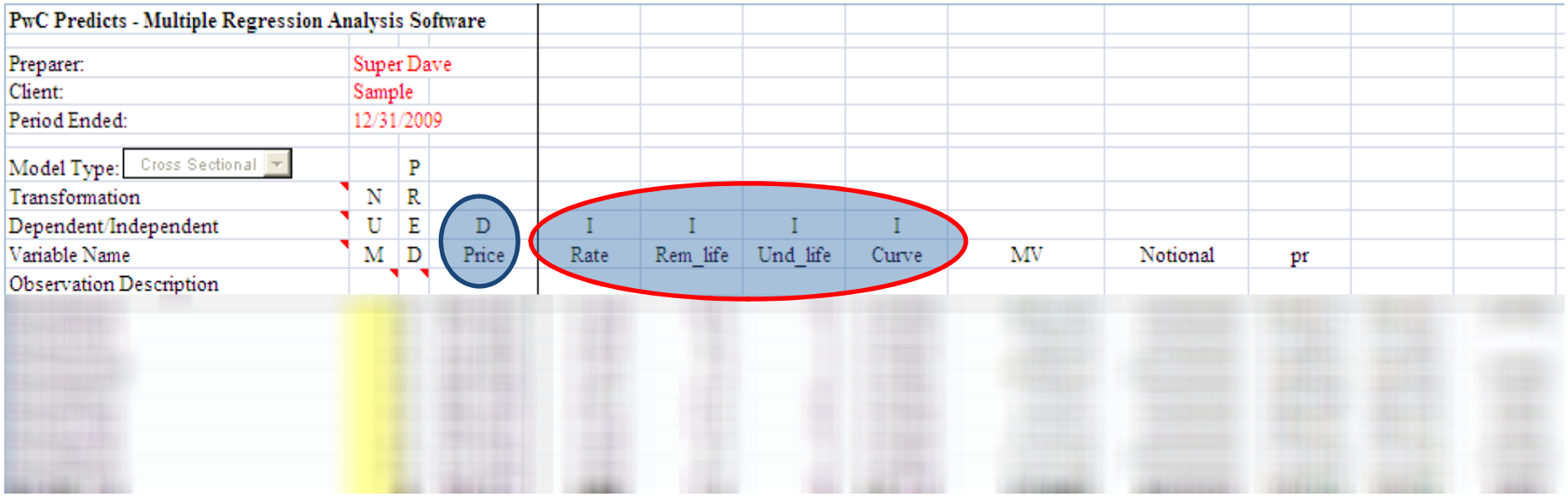
Performing a JOIN between two files in ACL is very similar to performing a JOIN between two tables using a SQL Statement, or a basic vlookup in Microsoft Excel. The ultimate goal is identifying overlapping and missing data between two data sets.

Join	ACL Join Type	SQL Join Type
	MATCH	Inner Join
	PRIMARY	Left Outer Join
	SECONDARY	Right Outer Join
	PRIMARY SECONDARY	Full Outer Join
	UNMATCHED	Left Outer Join (with <> criteria)

Data Analysis– Analysis Techniques

Regression Analysis – Swaptions

In general, it is used to model a response variable (Y) as a function of one or more driver variables. There are two types, which is determined by the number of driver variables. A model using a single variable is called "Simple Linear Regression while more than variable is called "Multiple Linear Regression Analysis"



PwC Predicts - Multiple Regression Analysis Software											
Preparer:	Super Dave										
Client:	Sample										
Period Ended:	12/31/2009										
Model Type:	Cross Sectional		P								
Transformation	N	R									
Dependent/Independent	U	E	D	I	I	I	I				
Variable Name	M	D	Price	Rate	Rem_life	Und_life	Curve	MV	Notional	pr	
Observation Description											

Data Analysis– Analysis Techniques

Regression Analysis

A. THE REGRESSION MODEL

			Number of Observations:	
Cross Sectional - Levels			Base	72
Variable to be explained: Price			Prediction	0
Descriptor Variable	**Transformation**	**Regression Coefficient**	**t-Statistic**	**Confidence Level**
Constant		-919.978	-3.641	99%
Rate		231.492	**33.061**	**99%**
Rem_life		1.889	0.945	65%
Und_life		25.353	**11.285**	**99%**
Curve		-25.352	-0.578	43%

The t-statistic measures the statistical significance of an individual regression coefficient. Benchmark: t-statistic of absolute value of 2 or better.

B: MEASURES OF EXPANATORY POWER AND PRECISION

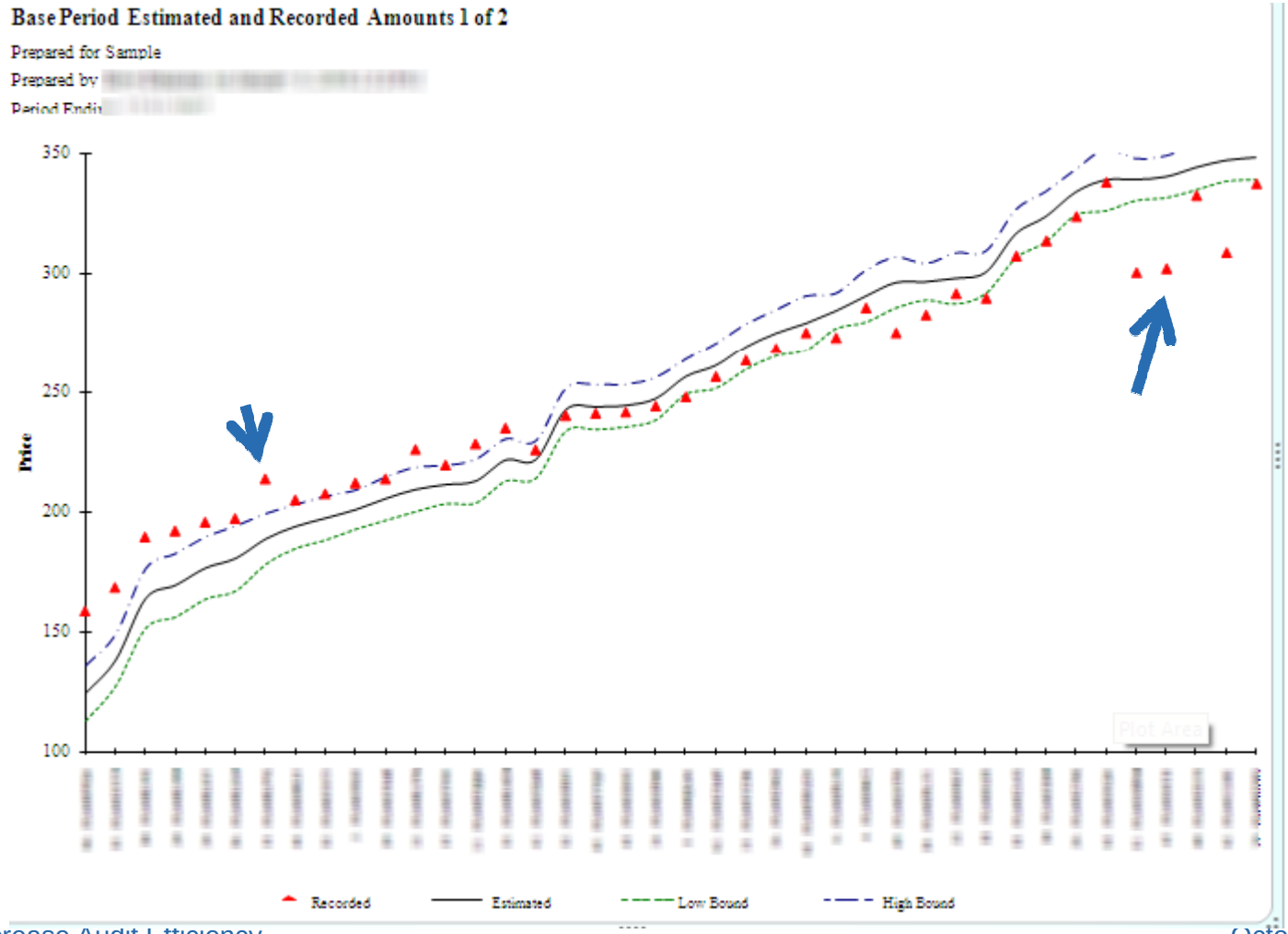
R-squared value	98%
Adjusted R-squared value	98%
F-statistic at 99.99% Confidence level	718.34

As a benchmark, the adjusted R-squared should be the principal focus, and the closer the value is to 1, the better the model explains the variable of interest. The F-statistic tests the overall descriptive power of the regression model. Models with less than 95% level of confidence for the F-Statistic should be reconsidered.

Aggregate Regression Estimate (Equals Recorded Amount) for 72 Base Model Observations	25,338
Aggregate Achieved Precision for 72 Base Model Observations:	354

Data Analysis– Analysis Techniques

Regression Analysis



Tools of the Trade

Data Mining Tools and Software for IA

Tool	Type	Learning Curve	Scalable	CCM Suite	Use (H/M/L)
Microsoft Excel	Spreadsheet	Varies	Low	NA	H
ACL	Analysis Tool	Beginner\Inter	Medium	Yes	M
IDEA	Analysis Tool	Intermediate	Medium	Yes	L
SAS	Analysis Tool	High	Low	Yes	L
MS Access	Database	Beginner\Inter	Low	NA	H
RDBS (Oracle, MSSQL, etc)	Database	Intermediate	High	NA	H
Crystal Reports	Reporting	Beginner	NA	NA	M
4th Generation Programming Language	Programming Language	High	Low	NA	L
Clementine\ Enterprise Miner	Analysis Tool	High	Medium	NA	L

Case Studies and Recent Examples

Cases \ Examples

Social Network Fraud Vote Testing

Overview

A major Social Network Website , donates $25 million dollars to charities who receive the most votes from the user base. Prizes range between $2.5 mn dollars and $50,000 dollars Larger charities were excluded (i.e. American Cancer Society) since this was meant to help small charities grow and gain access to capital that they would normally have trouble attaining. Any charity could be nominated, and so long as they received enough vote count.

As a side effect, organized members of society (globally) began attempting to influence the voting process. This included:

1) Organizing voters in third world countries to create profiles and vote for a certain charity
2) Electronic bots to automate voter registration and voting
3) [Censored]

Key Topics

a) What would be your approach?

Cases \ Examples

Social Network Fraud Vote Testing

Social Network _ID	DATE	CHARITY_HASH	CHARITY_EIN	SN_RATING	SN_IP	VN_USERR	BROWSER_STRING
702328739	2010-06-14 21:00:12 -0700	54997A8f7Zabdb08fbZ96ee688e5	91A041609	High	773.11.777.73	384751	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0; GTB6.3;
7368857404	2010-06-14 21:00:14 -0700	fa09d78AAeeZ550A4116Z5fbf1f46	311805306	High	73.118.99.76	262494	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; Trident/4.0; SLCC1;
502657548	2010-06-14 21:00:16 -0700	899d59fdfa5f55a4e89Z0A67Z0a15	75A577687	Medium	13.133.797.111	1938494	Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.1.9) Gecko/2010(
7857379089	2010-06-14 21:00:18 -0700	f458Ae8ff01695680AZ84Z871078	10856715	High	105.788.776.107	259879	Mozilla/4.0 (compatible; MSIE 7.0; AOL 9.0; Windows NT 6.0; WOW64; T
749826858	2010-06-14 21:00:19 -0700	e8Z607444dZf10a6a50A881Zdf5A	A05477A00	High	773.731.38.775	839306	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; Trident/4.0; .NET C
7646845777	2010-06-14 21:00:20 -0700	70e6b65fAeZfb5444e6bA06da184	A00385930	High	66.56.33.68	310773	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; Trident/4.(
7.00001E+14	2010-06-14 21:00:21 -0700	8a84ea557Abaade97Z1eaeb01AAd8	141377504	High	63.177.755.137	2	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; Trident/4.0; GTB6.5
7326932682	2010-06-14 21:00:28 -0700	b1795adZ485b8d50aAf86ed38f5Z	5A1A60470	Medium	96.117.53.91	456562	Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.6; en-US; rv:1.9.2.3) Gecko/2
7625688238	2010-06-14 21:00:28 -0700	164914a0141effe74bd03beeZA1Aa1	A08468493	High	68.73.50.77	545823	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; Trident/4.0; GTB6.5
375744	2010-06-14 21:00:40 -0700	a977adfZ1e0ff5Ae67d115A79ZAbA2	A64356A55	High	37.736.178.791	2113145	Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_5_8; en-us) AppleWebKit/5
520253779	2010-06-14 21:00:46 -0700	Af00ea71ffd0Zf80e969bbfZ969bd	A60A67341	High	63.775.793.770	1115696	Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_5_8; en-US) AppleWebKit/5
7033700028	2010-06-14 21:00:50 -0700	0b968a9d61eZAa7e7ZdZ6948Z94	341789597	Medium	75.178.77.770	1303235	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; Trident/4.(
7.00008E+14	2010-06-14 21:01:08 -0700	Zd01A8800a43fcA195f7a583e479	30013613A	Low	13.736.0.173	872	Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/2010(
7566300730	2010-06-14 21:01:11 -0700	8989846Af965aaAa784Z50A5fc0	71078AA55	High	71.735.101.33	1071648	Mozilla/5.0 (Macintosh; U; PPC Mac OS X 10.5; en-US; rv:1.9.2.3) Gecko/2
20007928	2010-06-14 21:01:11 -0700	b19Zf16d11e58Ab0e6a678A0Af67	A0804579A	High	97.65.105.33	2523081	Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.6; en-US; rv:1.9.2.3) Gecko/2
69600478	2010-06-14 21:01:17 -0700	dZ761b550bZffea1474907beZ999d	39131A503	High	63.733.766.139	2446996	Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.5; en-US; rv:1.9.2) Gecko/20
7293979254	2010-06-14 21:01:26 -0700	bee64fdb7190d9AAa1AA5Z6Af1e7	A61768860	High	66.75.133.701	802689	Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.2.3) Gecko/2010
595930468	2010-06-14 21:01:29 -0700	d95a8d5571614ff7b7d08884a1b0A	760700153	High	13.90.73.71	1388015	Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.3) Gecko/2010(
508278748	2010-06-14 21:01:31 -0700	ba5Z2551Z94fA97eZ4aZ4dfe98754	133358148	High	770.33.0.133	1595751	Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_3; en-us) AppleWebKit/5
7.00001E+14	2010-06-14 21:01:37 -0700	54997A8f7Zabdb08fbZ96ee688e5	91A041609	Low	98.77.66.153	93951	Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US) AppleWebKit/533.16
7374900256	1278113237.439940	7Ad7baAAZd6d6db094Z79659e0	A0859944A	High	738.770.75.716	2059428	Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.4; en-US; rv:1.9.2.4) Gecko/2
728655607	1278113237.476390	eb044ef6fbf8AZ0fd91751ZAA1Zaa	A7004800A	High	13.137.179.75	1778542	Mozilla/5.0 (Windows; U; Windows NT 6.0; en-US) AppleWebKit/533.4 (
7E+14	1278113242.157060	Ab6fbZ6b1467410085abe4A057ea	41183813A	Medium	108.87.68.703	361483	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322)
7373634976	1278113242.431730	0A0b79Z4efdaa5bb1Ae7Z4fbf8A3	4A1607378	High	766.105.739.133	455653	Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_4; en-us) AppleWebKit/5
7760737344	1278113243.808630	Z0f04a86ff0870Z0a08ad4f180886	954AA6AA3	High	766.105.739.138	433024	Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10_6_4; en-us) AppleWebKit/5
7084670690	1278113245.312610	07e46ZAA7f6AAaf6f467af9fd1eZ1	650870575	High	63.71.776.65	376675	Mozilla/4.0 (compatible; MSIE 7.0; AOL 9.5; AOLBuild 4337.155; Window:
7.00001E+14	1278113247.241060	ZbAdd1bAb0a8de6de5409afAaf5c	A03410498	Low	76.779.730.136	260890	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0; GTB6.5;
546585594	1278113247.338860	Ab88aabZ568A56afZ78b6e0e0fd4	651306978	High	760.39.757.53	1245027	Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2.6) Gecko/2010(
573400334	1278113248.526790	d4Aad50ff9a5Z096f18795043771A	A04801654	High	77.763.767.706	1881865	Mozilla/4.0 (compatible; MSIE 7.0; AOL 9.0; Windows NT 5.1; (R1 1.5); .NE
7243873044	1278113249.687940	Z5154a8Zfde7561Z07aa9A9e0f51bb	810664911	Medium	96.118.763.730	739157	Mozilla/5.0 (iPad; U; CPU OS 3_2 like Mac OS X; en-us) AppleWebKit/53
507675307	1278113249.902860	48db768A9ZfbA30Z71977f4a987c	640897384	High	68.777.135.87	1974229	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; WOW64; Trident/4.
7543907093	1278113251.210220	8fdaA81fda68fbfe4e6eA4Z5f486f	6A1835463	High	75.770.50.33	879237	Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.0; Trident/4.0; FunWel
7430755544	1278113251.502030	A8Z550a9dZd5fb5e5bA440090a	A60851887	High	75.103.739.19	680021	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0; GTB6.5; SLCC1; .NE
7E+14	1278113251.930610	0ffaaZ9651a1b4b1eddAf4e0da1d86	A01461577	Medium	63.778.130.133	335689	Mozilla/5.0 (iPad; U; CPU OS 3_2 like Mac OS X; en-us) AppleWebKit/53
7E+14	1278116608.264820	Z5154a8Zfde7561Z07aa9A9e0f51bb	810664911	Medium	63.775.3.107	318532	Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; Trident/4.0; .NET CL

Cases \ Examples

Independent Price Testing

Overview

To independently validate tier 1 equity prices with prices from a real time feed system and a batch pricing system on an EOD basis. Three sources are available:

(1) **Reuters EIJ** – Request-response pricing available from Reuters

(2) **Bloomberg** - End of day batch file of all equities at close

(3) **Positions and Balances System** - File containing firm positions and prices.

Key Topics

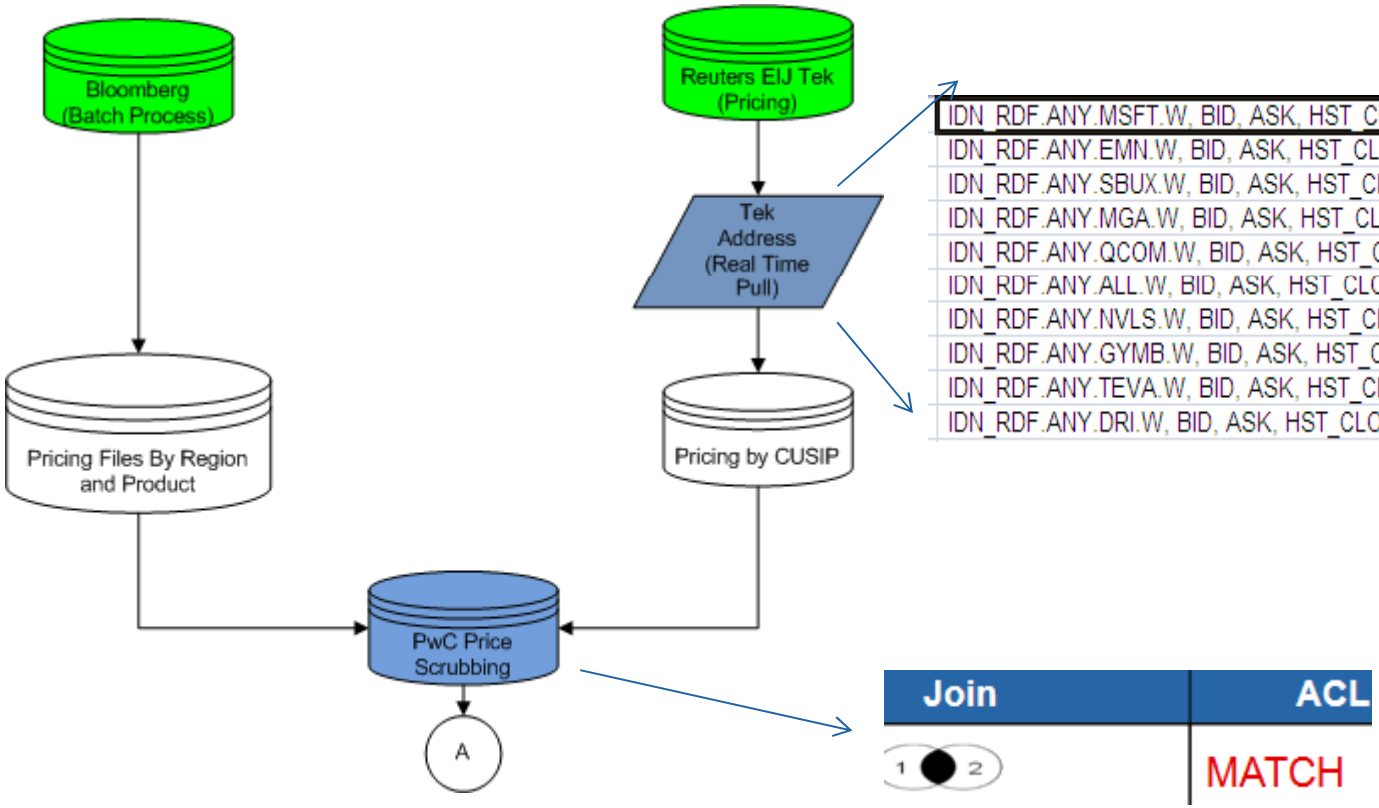
a) How would a request-response mechanism affect your analytics?

b) What are the general steps you would need to execute this?

c) What tool would be ideal for this? How would you do it?

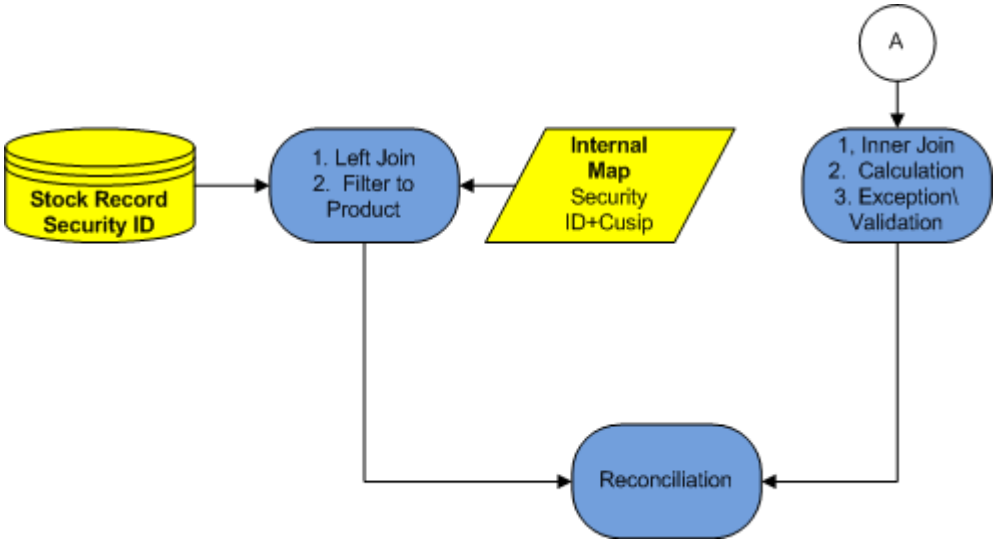
Cases \ Examples

Independent Price Testing



Cases \ Examples

Independent Price Testing



	Client Data							Bloomberg Data				
5	Short Vs Long	ISIN	SEDOL	CUSIP	Price	Position	Market value	securities_1	bbprice	bbcountry	BBcount	Markey Value Difference
6	L	US2107953083	2220527	210795308	17.92	7,524	134,830	CAL	17.92	US	0	0
7	L	US2473617023	NULL	247361702	11.38	17,479	198,911	DAL	11.38	US	0	0

Cases \ Examples

DTCC Custodial Position Testing

Overview

To independently validate firm positions with balances at DTCC at EOD. Perform this on a real-time basis, on demand, without knowledge or impacting operations and technology teams.

(1) **DTCC API BAL**– Daily positions file provided by DTCC, through FTP

(2) **Positions and Balances System** - A batch end of day file that provides prices per ticker

Key Topics

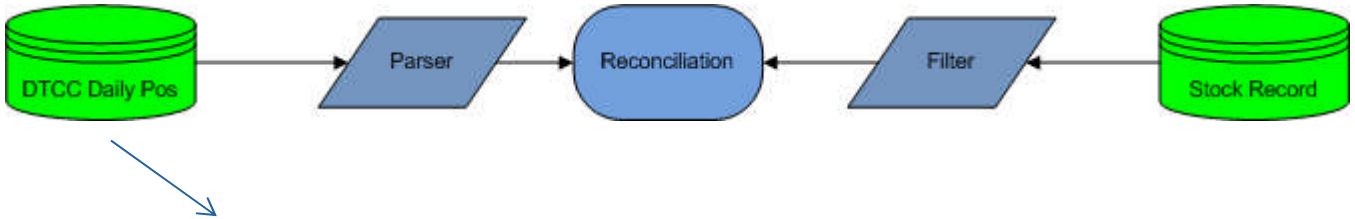
a) The DTCC API BAL is a standard output file, and is not a data file. What techniques could you use to parse this file?

b) How would you automate the retrieval? Would ACL be able to do this?

c) Would you expect the total positions to tie?

Cases \ Examples

DTCC Custodian Testing



PARTICIPANT: XXX THE DEPOSITORY TRUST COMPANY PAGE 1
XXX PARTICIPANT POSITION STATEMENT -- 12/31/09 XXX

G8259U 10 7 GPFDREGSSOLSTIC BE-
CHILLS - COD W/T DEPOSIT
LAST ACTIVE 10/15/08

ACCOUNT SUMMARY	- SDFS EQUITY	- CLOSING BALANCES
(010)	**TOTAL**	7,000

16117M 13 1 WTSCHTER113014 BE+
CHILLS - COD W/T DEPOSIT
LAST ACTIVE 12/09/09

ACCOUNT SUMMARY	- SDFS EQUITY	- CLOSING BALANCES
(010)	**TOTAL**	46,736

05873K 11 6 WTSBALLY083114 BE+
CHILLS - COD W/T DEPOSIT
LAST ACTIVE 12/18/09

ACCOUNT SUMMARY	- SDFS EQUITY	- CLOSING BALANCES
(010)	**TOTAL**	4,667

16117M 30 5 CHARTERCOMMUNIC BE+
CHILLS - COD W/T DEPOSIT
LAST ACTIVE 12/18/09

ACCOUNT SUMMARY	- SDFS EQUITY	- CLOSING BALANCES
(010)	**TOTAL**	166,447

05873K 20 7 BALLY TOTAL FIT BE+
CHILLS - COD W/T DEPOSIT
LAST ACTIVE 12/18/09

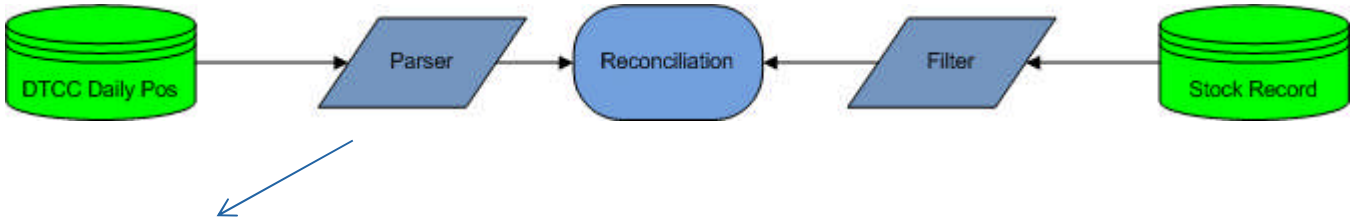
ACCOUNT SUMMARY	- SDFS EQUITY	- CLOSING BALANCES
(010)	**TOTAL**	2,046

218681 10 4 CORE-MARK HLDG CO.
LAST ACTIVE 12/31/09

ACCOUNT SUMMARY	- SDFS EQUITY	- CLOSING BALANCES
(010)	**TOTAL**	723

Cases \ Examples

DTCC Custodian Testing




```
#!C:\oracle\product\10.2.0\client_1\perl\5.8.3\bin

# @title: parse_DTC[illegible]l
# @desc:  parse DTC[illegible] for CUSIP and TOTAL and write to db
# @auth:  Tom Daniels
# @date:  02/02/10

use DBI;
use DBD::ODBC;

#initialize db vars
$server = 'UXXXXXX';
$database = 'XXXX_09';
$user = 'NAM\XXXX';

#connect to db
$dsn = "DBI:ODBC:Driver={SQL Server};Server=$server;Database=$database;Uid=$uid;";
$dbh = DBI->connect($dsn) or die("Error connecting to database: $database");

#create table
$tbl_name = 'DTC_187';
$table_desc = "CREATE TABLE $tbl_name (CUSIP varchar(9), AMOUNT float)";
$sth = $dbh->prepare($table_desc);
```

Cases \ Examples

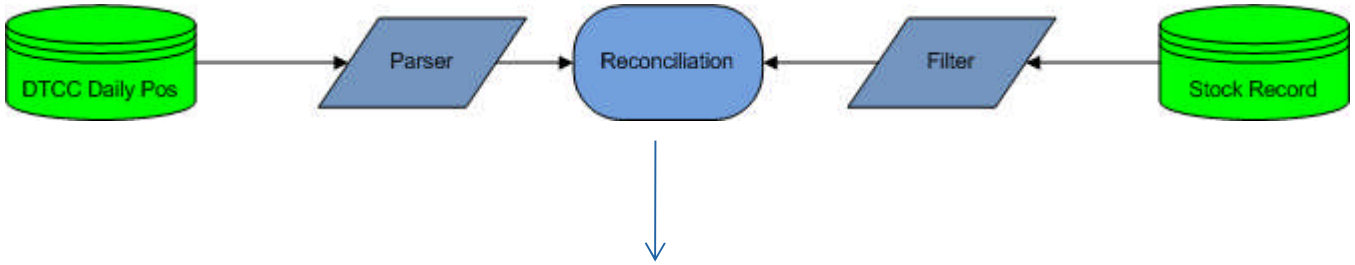
DTCC Custodian Testing


```
#begin parsing DTC xxx report records
$datafile = "dtcxxx_20091231.txt";
$num_recs = 0;
$num_lines = 0;
$state = 0;
$cusip = NULL;
open(DTC, $datafile) or die ("Could not open file: $datafile\n");
while (<DTC>) {
   $num_lines++;
   @vars = split(/ /, $_);
   #identify new record in report by cusip
   if($state eq 0) {
      if(iscusip(@vars)) {
           $cusip = $vars[2].$vars[3].$vars[4];
           $state = 1;
      }
   #we found a cusip and now are looking for a total
   } elsif ($state eq 1) {
      if($totalIndex = isTotal(@vars)) {
           $total = 0;
           for($i=$totalIndex+1;$i<60;$i++) {
             if($vars[$i] =~ (m/(?=.*\d)/)) {
                  $total = $vars[$i]
             }
           }
           #remove commas to insert as int into db
           $total =~ s/\,//g;
           $state = 0;
           $num_recs++;
           #print output"CUSIP = $cusip\t\tTOTAL = $total\n\n";

           #insert values into db
           $sql = "INSERT INTO $tbl_name VALUES ('" . $cusip . "', $total)";
           $sth = $dbh->prepare($sql);
           $sth->execute;
```

From an effort perspective, this program took about 3 hours to write. Monarch is an alternative to code, which provides a GUI

Cases \ Examples

DTCC Custodian Testing \ Continuous Monitoring




```
/* roll up dtc quantities obtained from xxx */
drop table #DTC_XXX
select CUSIP_Number, SUM(Share_Quantity) as QTY
into #DTC_XXX
from dbo.DTC_XXX_APIBAL
group by CUSIP_Number

/* compare rolled xxx quantities to xxx report from DTC and return breaks */
SELECT d1.CUSIP, dj.CUSIP_Number, d1.AMOUNT, dj.QTY, (dj.QTY-d1.AMOUNT) as DIFF
FROM #DTC_XXX dj
INNER JOIN dbo.DTC_xxx d1
ON dj.CUSIP_Number = d1.CUSIP
WHERE dj.QTY - d1.AMOUNT <> 0

/* confirm all cusips listed in the xxx file are in the xxx report */
select * from #DTC_XXX
where CUSIP_Number not in (select CUSIP from dbo.DTC_xxx)
```

Cases \ Examples

Other Than Temporary Impairment\Underwater Analysis

Overview

To identify securities that are considered impaired. Impairment is defined when a security has been underwater (below book) by a threshold percentage for more 25% than a consecutive 9 months period. This is a simple concept, but actually difficult for many RDBMS developers. You have a single source file

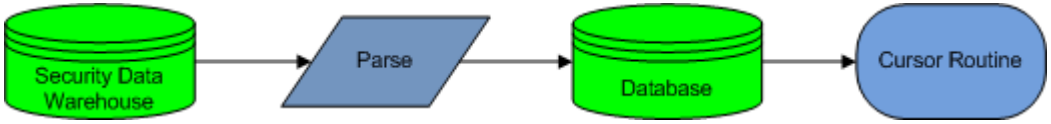
(1) **Security Position with Book Price and Market Value**– Daily positions file provided by DTCC, through FTP

Key Topics

a) For RDBMS experienced individuals how would you code this?

Cases \ Examples

Other Than Temporary Impairment\Underwater Analysis



OTTI – in this client, stated that from an accounting standpoint that a security, if it was under book value by more than 25% for at least 9 months consecutive, then it should be booked as impairment.

Cusip	Date (month)	Book	Market	% of Book Value
00139#118	1	100	72	0.72
00139#118	2	100	71	0.71
00139#118	3	100	70	0.7
00139#118	4	100	60	0.6
00139#118	5	100	54	0.54
00139#118	6	100	52	0.52
00139#118	7	100	30	0.3
00139#118	8	100	50	0.5
00139#118	9	100	40	0.4
00139#118	10	100	10	0.1
00139#118	11	100	10	0.1
00139#118	12	100	20	0.2

No easy way to compare that two months are both underwater (<75% and consecutive)

Cases \ Examples

Other Than Temporary Impairment\Underwater Analysis

Cusip	Date (month)	Book	Market	% change
00139#118	1	100	72	0.72
00139#118	2	100	71	0.71
00139#118	3	100	70	0.7
00139#118	4	100	60	0.6
00139#118	5	100	54	0.54
00139#118	6	100	52	0.52
00139#118	7	100	30	0.3
00139#118	8	100	50	0.5
00139#118	9	100	40	0.4
00139#118	10	100	10	0.1
00139#118	11	100	10	0.1
00139#118	12	100	20	0.2

Use code to go row by row to create the below field. This is inefficient and slow, however.

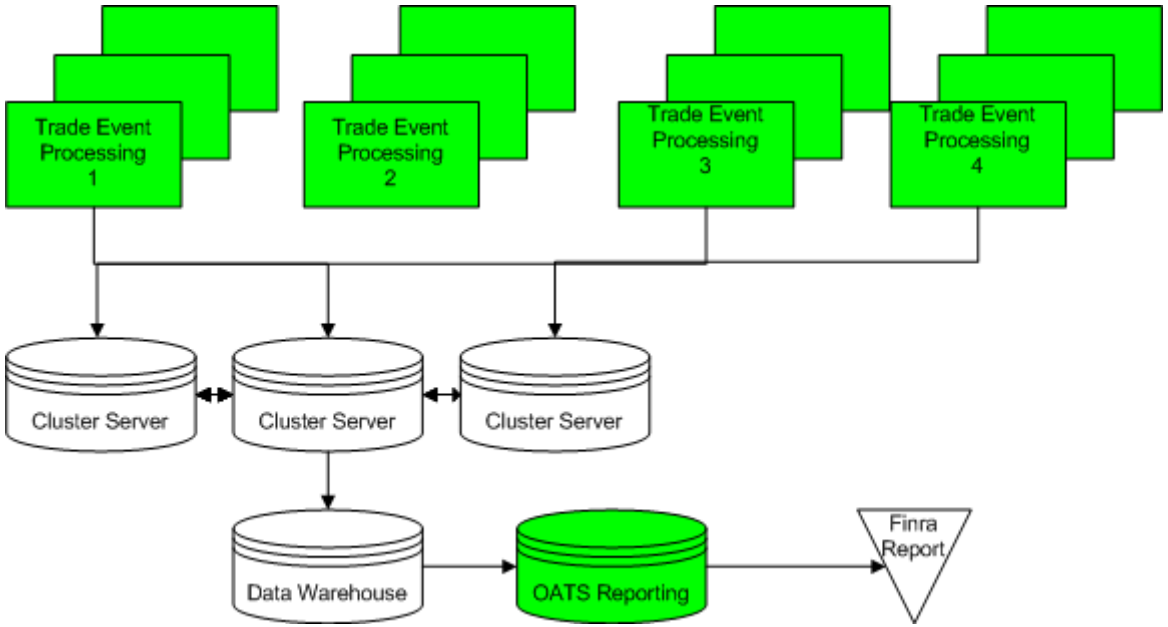
Under Water Sequence
111111111111111111111111
111101110000000000011111

A sequence of 9 consecutive 1's means an OTTI.

Cases \ Examples

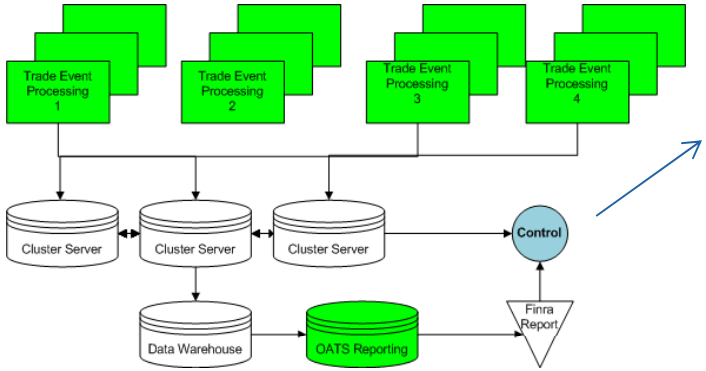
FINRA OATS Reporting

"FINRA member firms are required to develop a means for electronically capturing and reporting to OATS specific data elements related to the handling or execution of orders, including recording all times of these events in hours, minutes, and seconds, and to synchronize their business clocks." – FINRA Order Audit Trail System



Cases \ Examples

FINRA OATS Reporting – Continuous Auditing



Results Overview for 9/9/2010		
Category	Trade Count	%
Total Cluster Server Population	6,988,022	
Total FINRA Reported Population	6,426,552	
(1) Total Matched	6,340,060	90.7%
(2) Total in Cluster, not in FINRA	2,818	0.0%
(3) Total in FORE, not in Cluster	-	0.0%

Category	Description	Trade Count (C)
Total in Cluster, not in FINRA		2,818
Total Explained		2,818
[redacted]		1,409
[redacted]		1,020
[redacted]		-

Contact Information

Glenn Cheng Director Data Assurance	(646) 471-8211 glenn.cheng@us.pwc.com

Dave Dauksas Partner Data Assurance	(703) 918-3859 dave.dauksas@us.pwc.com

PRICEWATERHOUSECOOPERS