Data Profiling:
Best Practices by Example

GIAN DI LORETO, PH. D.

IDQ CONFERENCE
NOVEMBER 4, 2013
LITTLE ROCK, AR

Class Overview

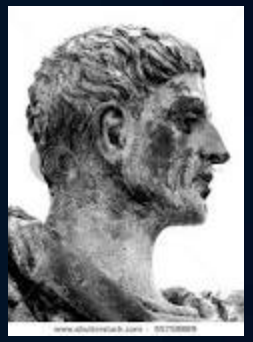
- Meet your instructor, class demographics
- Data profiling: an overview
- Tools
- Where does data profiling fit?
- Modern data profiling techniques
  - Basic Data Profiling
  - Advanced Data Profiling Techniques
  - Subjects and Subject Level Data Profiling
- Reporting
- Tools
- Discussion, Questions, Wrap-Up

Gian Di Loreto (me)

- Ph.D. from Michigan State in 1998 studying particle physics
- Transitioned to industry in 1998 as a programmer
- Got involved in Data Quality in 2000
  - DQ was an emerging discipline then
  - Companies didn't understand it, what it was, why they should care (do they now?)
- Still involved in Data Quality
  - Data Quality Assessments (data profiling is key here)
  - Data Cleansing
  - Ongoing Data Quality
  - Data Stewardship Program Design and Implementation
  - Trainings
- I try to remain faithful to scientific, academic principles

Data Profiling: an Overview

- We hear about *profiling* all the time
  - Profiling is the study of a multidimensional object during which all but 1 (or a small few) dimension is examined while ignoring all others
  - This allows us to more fully understand one degree of freedom, but at the expense of all others
    - Mathematical
    - Geometrical (your instructor has been described as having a 'Roman Profile')



    - Racial
      - Ignore other factors when deciding whom to stop, frisk, scan, etc..
    - Data
      - Simplest example: study one data element regardless of any other attributes

Data Profiling: an Overview

- Data Profiling – definitions:
  - Data Entity – data table, Excel sheet, etc.
  - Data Attribute – data field, column, etc.
  - Subject – the real world object your data describes, aka the thing in your data that you care about
  - Metadata – derived data, data about data
- Simple data profiling involves exhaustively studying one data attribute without regard to the values or behavior of other data attributes in the same entity.
- More complex data profiling will involve studying the relationship between data attributes, the behavior of one data attribute as it relates to one or more others within the same or a different entity
- Even more complex data profiling will involve the definition of a subject type and profiling subject derived metadata
- We will discuss each of these today, with examples, as time permits

Data Profiling: Tools

- Virtually all data profiling performed today employs the use of a tool, a software package, that performs (usually) both canned and custom data profiling
- We will briefly look at three such tools today during the 2nd session
  - Trillium
  - DataFlux
  - Talend
- We will not dwell on the tools, just give you a quick feel for how these techniques have been implemented by the software development community (in the cases where they have)
- We will demo by example using Talend since it's free and you guys can go ahead and download it yourselves if you like

Data Profiling: Tools - continued

- Because data profiling is well understood (relatively) and easily programmed, there are many good tools out there
- The problem is that often management will look at the tool as a solution to your data integrity, data quality issues rather that just that, a tool
- Like any tool, a data profiling tool (such as those we will study today) is only as good as its operator

Where Does Data Profiling Fit?

- Data profiling is a quick way to learn a great deal about any given data set.
- It is usually done at the outset of a data quality investigation, or any data-centric project, such as
  - A data quality assessment
  - A data cleansing
  - The creation of a data warehouse
  - A system upgrade or new implementation
  - Any data migration
- Essentially, anytime you want an overview of what you've got in your data, a data profile is great way to start, however there are caveats:
  - A data profile generates a great deal of reports, charts, metadata
  - We must resist the temptation to focus one of the excellent tools on our data, create a bunch of reports and call it a data profile
  - The analogy is the highlighter in college

What doesn't data profiling do ?

- Data profiling does not improve data quality
- Data profiling does not improve data quality
- Data profiling does not improve data quality
- Data profiling will not simplify your project
- Data profiling will not create a project plan
- Data profiling will not set expectations for time, resources, or cost.

- What it does do is provide a vast amount of metadata that if carefully studied, can render a path toward all of the above.

Modern Data Profiling Techniques

What is considered ‘modern’ of course changes all the time, these techniques are some that have worked for your instructor, have added value to actual projects in real life.

Initial Data Profiling Exercises

- Statistics Gathering
  - Max/Min/Mean/Median/SD/Field Data Type
- Key Constraints
- Frequency Distributions
- Outlier Study
  - Frequent Values, Infrequent Values

Statistics Gathering

- Entity Level (table level)
  - Very useful during data transmission
  - If reports match before and after a data migration, confidence can be high that all data was successfully migrated (like a checksum)
  - The example on the next page is from DataFlux (we'll see more later) and gives an overview of statistics culled from the table PS_PERSONAL_DATA a PeopleSoft table
  - This is the first toe in the pool that most tools provide when data profiling.
  - You can see the conundrum already; it's a lot of information that needs to be examined and filtered before sharing. We'll talk more about reporting later.

Entity Profile Example

DataFlux Data Management Studio 2.1

PS_PERSONAL_DATA Profile

File Edit View Insert Actions Tools Window Help Version: 10/2/2011 1:16:46 PM

Report | Properties

Tables

- PS_PERSONAL_DATA
  - ADDRESS1
  - ADDRESS1_AC
  - ADDRESS1_OTHER
  - ADDRESS2
  - ADDRESS2_AC
  - ADDRESS2_OTHER
  - ADDRESS3
  - ADDRESS3_AC
  - ADDRESS3_OTHER
  - ADDRESS4
  - ADDRESS4_OTHER
  - ADDR_FIELD1
  - ADDR_FIELD1_OTHE
  - ADDR_FIELD2
  - ADDR_FIELD2_OTHE
  - ADDR_FIELD3
  - ADDR_FIELD3_OTHE
  - AGE_STATUS
  - ALTER_EMPLID
  - APPLID
  - APP_DT
  - BARG_UNIT
  - BILINGUALISM_CODE
  - BIRTHCOUNTRY
  - BIRTHDATE
  - BIRTHPLACE
  - BIRTHSTATE
  - CITIZEN_PROOF1
  - CITIZEN_PROOF2
  - CITY
  - CITY_AC
  - CITY_OTHER
  - COUNTRY
  - COUNTRY_CODE
  - COUNTRY_NM_FORM
  - COUNTRY_OTHER
  - COUNTY
  - COUNTY_OTHER
  - CPAMID
  - DISABLED
  - DISABLED_YET
  - DT_OF_DEATH
  - EMPLID
  - EMPLID_INT

Tables | Collections | Redundant Data Analyses

PS_PERSONAL_DATA Schema: TDWI_OWNER Data Source: TDWI_SQL

Standard Metrics | Custom Metrics | Business Rules | Alerts | Visualizations | Notes

Field Name	Collections	rdinal Positio /	Count	Null Count	Percent Null	Blank Count	Minimum Value	Maximum Value	Mode	Pattern Count	Unique Count	Uniqueness	Primary Key Car
SUBJECT_ID		1	46016	0	0	(not applicable)	0	17360	0	(not applicable)	5709	12.50	no
EMPLID		2	46016	0	0	0	100006056	99991362	(no data/ambig.)	7	45983	99.93	no
COUNTRY_NM_FORMA		3	46016	2	0	0	AUT	VEN	USA	1	24	0.05	no
NAME		4	46016	0	0	0	Aadahl,Frank P	von Kalben,Stephan K	Void,Void V	1762	45822	99.58	no
NAME_INITIALS		5	46016	6616	14.4	0	AA	wte	JS	7	839	2.13	no
NAME_PREFIX		6	46016	37769	82.1	0	Dr	Ms	Mr	3	5	0.06	no
NAME_SUFFIX		7	46016	44227	96.1	0	BSN, RN	V	Jr	5	8	0.45	no
NAME_ROYAL_PREFIX		8	46016	46016	100	0			(no data/ambig.)	0	0	0	no
NAME_ROYAL_SUFFIX		9	46016	46016	100	0			(no data/ambig.)	0	0	0	no
NAME_TITLE		10	46016	46016	100	0			(no data/ambig.)	0	0	0	no
LAST_NAME_SRCH		11	46016	91	0.2	0	AADAHL	ZYWICZYNSKI	SMITH	45	21427	46.66	no
FIRST_NAME_SRCH		12	46016	94	0.2	0	A	ZYNTHIA	JOHN	29	6086	13.25	no
LAST_NAME		13	46016	91	0.2	0	Aadahl	von Kalben	Smith	241	21623	47.00	no
FIRST_NAME		14	46016	94	0.2	0	A	the	John	185	6139	13.37	no
MIDDLE_NAME		15	46016	8001	17.4	0	(	x	A	27	279	0.73	no
SECOND_LAST_NAME		16	46016	46016	100	0			(no data/ambig.)	0	0	0	no
SECOND_LAST_SRCH		17	46016	46016	100	0			(no data/ambig.)	0	0	0	no
NAME_AC		18	46016	46016	100	0			(no data/ambig.)	0	0	0	no
PREF_FIRST_NAME		19	46016	46016	100	0			(no data/ambig.)	0	0	0	no
PARTNER_LAST_NAME		20	46016	46016	100	0			(no data/ambig.)	0	0	0	no
PARTNER_ROY_PREFI		21	46016	46016	100	0			(no data/ambig.)	0	0	0	no
LAST_NAME_PREF_NL		22	46016	0	0	0	1	1	1	1	1	0	no
COUNTRY		23	46016	114	0.2	0	AUS	VNM	USA	1	33	0.07	no
ADDRESS1		24	46016	850	1.8	0	1708 Sherman Pl #6	rt 1 box 624	Manpower	11000	44541	98.62	no
ADDRESS2		25	46016	44437	96.6	0	Space 80	unit 201	(no data/ambig.)	628	1382	87.52	no
ADDRESS3		26	46016	46000	100	0	31701 Blagnac Cedex	Vito Ciciretto	(no data/ambig.)	16	16	100	no
ADDRESS4		27	46016	46016	100	0			(no data/ambig.)	0	0	0	no
CITY		28	46016	871	1.9	0	Long Beach	zimmerman	Phoenix	511	6329	14.02	no
NUM1		29	46016	46014	100	0	14	56	(no data/ambig.)	1	2	100	no
NUM2		30	46016	46016	100	0			(no data/ambig.)	0	0	0	no
HOUSE_TYPE		31	46016	46016	100	0			(no data/ambig.)	0	0	0	no
ADDR_FIELD1		32	46016	46016	100	0			(no data/ambig.)	0	0	0	no
ADDR_FIELD2		33	46016	46016	100	0			(no data/ambig.)	0	0	0	no
ADDR_FIELD3		34	46016	46016	100	0			(no data/ambig.)	0	0	0	no
COUNTY		35	46016	43414	94.3	0	001	will	Maricopa	106	559	21.40	no
STATE		36	46016	984	2.1	0	02	ZZ	AZ	3	68	0.15	no
POSTAL		37	46016	738	1.6	0	00000	WF93PX	61032	19	10353	22.87	no
GEO_CODE		38	46016	46016	100	0			(no data/ambig.)	0	0	0	no

Statistics Gathering

- Attribute Level (data row level) profiling
- All Data Types
  - Null Count – Null Percentage: number and/or percentage of records with a null value
  - Mode – Most frequent value
  - Pattern Count – Number of difference distinct patterns observed; mm/dd/yyyy or 999999.99 for example
  - Data type observed always or almost always in the column
  - Length of data in the column (most of the time)
  - Uniqueness
- Numeric Data Types
  - Mean
  - Median
  - Precision
  - Standard Deviation

Attribute Profile Example

ADDRESS1 Table: PS_PERSONAL_DATA Schema: TDWI_OWNER Data Source: TDWI_SQL

Column Profiling | Frequency Distribution | Pattern Frequency Distribution | Percentiles | Outliers | Primary Key/Foreign Key Analysis | Notes

Metric Name	Metric Value
Ordinal Position	24
Count	46016
Null Count	850
Percent Null	1.8
Blank Count	0
Minimum Value	1708 Sherman Pl #6
Maximum Value	rt 1 box 624
Mode	Manpower
Pattern Count	11000
Unique Count	44541
Uniqueness	98.62
Primary Key Candidate	no
Data Type	varchar
Data Length	165 chars
Actual Type	string
Minimum Length	2
Maximum Length	54
Mean	(not applicable)
Median	(not applicable)
Non-null Count	45166
Nullable	YES
Decimal Places	0
Standard Deviation	(not applicable)
Standard Error	(not applicable)

Statistics Gathering

Attribute Level (data row level) profiling - continued

- For fields with non-unique data the frequency distribution (group-by) results can yield very interesting results
- Can be compared with allowed values
- Frequent and in frequent values should be studied

ACTION_DT Table: PS_JOB Schema: TD

Column Profiling | **Frequency Distribution** | Pattern Freq

Value	Count	Percentage
2000-12-23	7171	9.29
2001-12-28	1356	1.76
2005-07-16	1215	1.57
2004-12-17	1204	1.56
2009-08-25	1129	1.46
2007-07-06	1122	1.45
2010-03-18	1043	1.35
1998-01-01	890	1.15
2003-03-19	833	1.08
2004-03-20	814	1.05
2005-03-25	807	1.05
2007-03-24	802	1.04
2008-03-22	786	1.02

CITY Table: PS_PERSONAL_DATA Schema: TDWI_OWNER Data Source: TDWI_SQL

Column Profiling | **Frequency Distribution** | Pattern Frequency Distribution | Percentiles | Outliers | Prin

Value	Count	Percentage
Phoenix	2160	4.69
(null value)	871	1.89
Glendale	692	1.50
Mesa	585	1.27
Tucson	554	1.20
Chandler	516	1.12
Tempe	497	1.08
Scottsdale	470	1.02
South Bend	451	0.98
Freeport	433	0.94
Baltimore	400	0.87
Albuquerque	375	0.81
Torrance	342	0.74
Minneapolis	326	0.71
Peoria	313	0.68
Columbia	277	0.60
Gilbert	253	0.55
Jacksonville	247	0.54
Colorado Springs	234	0.51
Los Angeles	231	0.50
Houston	216	0.47
Greenville	205	0.45
Las Cruces	192	0.42
Chester	169	0.37
San Diego	159	0.35
Springfield	155	0.34
Long Beach	151	0.33
Petersburg	147	0.32
Plymouth	146	0.32

Statistics Gathering

Attribute Level (data row level) profiling - examples

ACTION_DT Table: PS_JOB Schema: TD

Column Profiling | **Frequency Distribution** | Pattern Freq

Value	Count	Percentage
2000-12-23	7171	9.29
2001-12-28	1356	1.76
2005-07-16	1215	1.57
2004-12-17	1204	1.56
2009-08-25	1129	1.46
2007-07-06	1122	1.45
2010-03-18	1043	1.35
1998-01-01	890	1.15
2003-03-19	833	1.08
2004-03-20	814	1.05
2005-03-25	807	1.05
2007-03-24	802	1.04
2008-03-22	786	1.02

CITY Table: PS_PERSONAL_DATA Schema: TDWI_OWNER Data Source: TDWI_SQL

Column Profiling | **Frequency Distribution** | Pattern Frequency Distribution | Percentiles | Outliers | Prin

Value	Count	Percentage
Phoenix	2160	4.69
(null value)	871	1.89
Glendale	692	1.50
Mesa	585	1.27
Tucson	554	1.20
Chandler	516	1.12
Tempe	497	1.08
Scottsdale	470	1.02
South Bend	451	0.98
Freeport	433	0.94
Baltimore	400	0.87
Albuquerque	375	0.81
Torrance	342	0.74
Minneapolis	326	0.71
Peoria	313	0.68
Columbia	277	0.60
Gilbert	253	0.55
Jacksonville	247	0.54
Colorado Springs	234	0.51
Los Angeles	231	0.50
Houston	216	0.47
Greenville	205	0.45
Las Cruces	192	0.42
Chester	169	0.37
San Diego	159	0.35
Springfield	155	0.34
Long Beach	151	0.33
Petersburg	147	0.32
Plymouth	146	0.32

Statistics Gathering

Pattern Frequency Distribution

CITY Table: PS_PERSONAL_DATA Schema: TDWI_OWNER Data Source: TDWI_SQL

Column Profiling | Frequency Distribution | **Pattern Frequency Distribution** | Percentiles | Outliers | Primary Key/Foreign Ke

Pattern	Alternate	Count	Percentage
Aaaaaaa	Aa(6)	7399	16.39
Aaaaaaaa	Aa(7)	6694	14.83
Aaaaaa	Aa(5)	5014	11.11
Aaaaaaaaa	Aa(8)	3722	8.24
Aaaaaaaaaa	Aa(9)	3667	8.12
Aaaaa	Aa(4)	2134	4.73
Aaaaaaaaaaa	Aa(10)	1707	3.78
Aaaa	Aa(3)	1236	2.74
Aaaaaaaaaaaa	Aa(11)	895	1.98
Aaaaa Aaaa	Aa(4) Aa(3)	860	1.90
Aaaaa Aaaaa	Aa(4) Aa(4)	714	1.58
Aaaa Aaaaa	Aa(3) Aa(4)	594	1.32
Aaaa Aaaaaa	Aa(3) Aa(5)	517	1.15
Aaa Aaaaa	Aa(2) Aa(4)	515	1.14
Aaaaaaa Aaaaa	Aa(6) Aa(4)	476	1.05
Aaaaaa Aaaa	Aa(5) Aa(3)	468	1.04
Aaaaaaaa Aaaa	Aa(7) Aa(3)	455	1.01
Aaa Aaaaaa	Aa(2) Aa(5)	438	0.97
Aaa Aaaaaaa	Aa(2) Aa(6)	435	0.96
Aaaaaaaa Aaaaaaa	Aa(7) Aa(6)	405	0.90
Aaa Aaaa	Aa(2) Aa(3)	401	0.89
Aaaaaa Aaaaa	Aa(5) Aa(4)	325	0.72
Aaaaaa Aaaaaa	Aa(5) Aa(5)	316	0.70
Aaaa Aaaa	Aa(3) Aa(3)	278	0.62
Aaaaaaa Aaaa	Aa(6) Aa(3)	258	0.57
Aa Aaaa	Aa Aa(3)	252	0.56
Aaaaa Aaaaaa	Aa(4) Aa(5)	237	0.52
Aaaaaaaa Aaaaa	Aa(7) Aa(4)	227	0.50
Aaaa Aaaaaaa	Aa(3) Aa(6)	198	0.44
Aaaaa Aaaaaaa	Aa(4) Aa(6)	160	0.35

Statistics Gathering

- That will conclude the simple data profiling exercise
- These examples are data type, industry, tool, non-specific
- These quantities can and should be studied in data profiling exercise and any good tool will provide these for you
- Always remember in all examples to look at outliers (outliers?), anything that appears very rarely, or very frequently is usually worth looking into.
- Your job is to figure out where the interesting metadata lies and to filter that out and prepare for consumption by the non-technical crowd (more on reporting later)

Advanced Data Profiling Techniques

- The data profiling techniques we have described so far can be thought of as studying the data 'at rest'. But there is often a time dependence to the data that can provide useful insight.

for example, consider the following:

HR Employee Statuses: Active, Term, Dead, Hire, LOA, RFL

can yield time dependent pairs of statuses some of which are allowed, some of which are not

Active-> Term

Hire -> LOA

LOA -> RFL

- This 'state transition analysis' can be applied to any time dependent data.
- We will show examples during the demo section of our lesson today.

Advanced Data Profiling Techniques

Subject Profiling

- The last type of data profiling we will introduce today is called ‘subject profiling’.
- The subject is the real life entity your data describes.
  - Most of your data can be tied back to a subject
  - Your data can have multiple subject types
  - The subject is best described as the object in your data that you care about
  - In HR, one subject is the employee
- Subject: discussion, what is the subject in your data?

Advanced Data Profiling Techniques

Subject Profiling

- How to ID your subjects, build subject table
- Where in your data do your subjects exist? (from flag analysis, next slide)

Advanced Data Profiling Techniques

From Flag Analysis:

Subject ID	Subject Identifier	system 1 ID	system 2 ID	system 3 ID	from system 1	from system 2	from system 3	SystemString
1	SIMPSON, WILLIAM		13383747	51128015	N	Y	Y	NYY
2	SHAH, SANJEEVKUMAR		97990694	83825324	N	Y	Y	NYY
3	KLECKA, ELIZABETH	57424517		77159268	Y	N	Y	YNY
4	HEYNIS, MARY	50643168	38773091	61705282	Y	Y	Y	YYY
5	FROEHLICH, DEBRA	65474857	4788680	20263455	Y	Y	Y	YYY
6	WHITSURA, FRANK		34081383	9028648	N	Y	Y	NYY
7	JAMES, NELSON	40521824	61221964	88640578	Y	Y	Y	YYY
8	VEGA, HONORIO	96206762		3189667	Y	N	Y	YNY
9	WULF, LONARTA		96796216	17319820	N	Y	Y	NYY
10	BALL, JON		71316335	83662510	N	Y	Y	NYY
11	KOLJACK, MATHIAS			67996093	N	N	Y	NNY
12	WALENGA, JOEL		22949861		N	Y	N	NYN
13	PARK, RODNEY		20408982	82230657	N	Y	Y	NYY
14	WISE, IRENE		54396436	56040041	N	Y	Y	NYY
15	HATAMI, RICHARD	84648416	33360377	44539967	Y	Y	Y	YYY
16	SANDIFER, FREDDIE		26719162		N	Y	N	NYN
17	ELGAR, ALBERT	67369462			Y	N	N	YNN
18	SAKATA, FORD		83613295	94593963	N	Y	Y	NYY
19	PROSEK, SUSAN	30045977	14659663	2629273	Y	Y	Y	YYY
20	SCHREIBER, JOHN			44465024	N	N	Y	NNY
21	SORENSEN, BERNARD	85980590	20965008	77776535	Y	Y	Y	YYY
22	MADSEN, ROBERT	840806	11426695	71071458	Y	Y	Y	YYY
23	SWEENEY, PATRICK			75698270	N	N	Y	NNY
24	MA, DANIEL	7225529		57966111	Y	N	Y	YNY
25	HEDRICK, BRIAN			19034742	N	N	Y	NNY
26	JEDRZEJEK, JOSEPH		28205762	99595414	N	Y	Y	NYY
27	SUSTR, RONALD			41667946	N	N	Y	NNY
28	COSTAGLI, BLAS		8263132	39083443	N	Y	Y	NYY
29	ZINNEN, HERMANN	70126267	13695857	57872540	Y	Y	Y	YYY
30	MARTE, JULIO		10398477	3847652	N	Y	Y	NYY

Tools - summary

- As I've mentioned previously, the marketplace is crowded (you might say overcrowded with data profiling tools)
- I have played around with three, but I can't make any claims regarding which is better than the other, I will talk about the three and compare them, look at plusses and minuses.
- Tools can be stand alone or part of a larger software package
- Tools can operate as desktop versions and/or client server installs which can facilitate collaboration.
- I will demonstrate Talend, and give you some hand outs I have prepared which should give you a flavor of Trillium and DataFlux.
- Bottom line is the tool will provide you a great deal of metadata, the art here is how you arrange and disseminate that metadata.

Reporting

- Reporting is a dying art form
- As a physicist, reporting was everything. Without distilling some very complicated analyses down to a digestible summary, interest would wane quickly. Not to mention funding.
- The situation in the 'real world' is no different. If you cannot take the output of your data profile and create some simple and easy to swallow summaries, your project sponsors will feel lost. This will lead to bad things.
- The best reporting tools start at a very high level and allow drill down so interested parties can dig and see the detail, but the details are not provided until asked for.
- A lot of good work as been done with regard to this sort of drill down reporting, the entire field of BI is essentially (as I understand it) designed around the careful extraction and dissemination of information.
- Anybody can press a button and create a bunch of meta-data, the art of this business is preparing useful, usable, and actionable reports.

Reporting

- Anybody can press a button and create a bunch of meta-data, the art of this business is preparing useful, usable, and actionable reports… how to do this?

- I find the simple approach best. Create a single table (or as few as possible) to hold all your results, this is especially easy at the subject level.

- This table itself can then be profiled to provide summaries and overviews, but since it contains all the metadata can allow drill down to the meta-data and if you're careful, the data itself.

Demonstration using Talend

At this point we will fire up Talend and run through some examples of the data profiling techniques we discussed.

In the handouts I've included examples of data profiling taken as screen shots from DataFlux and Trillium just for your information, read through on your own if you are curious how these products handle the same tasks.

Wrap-Up

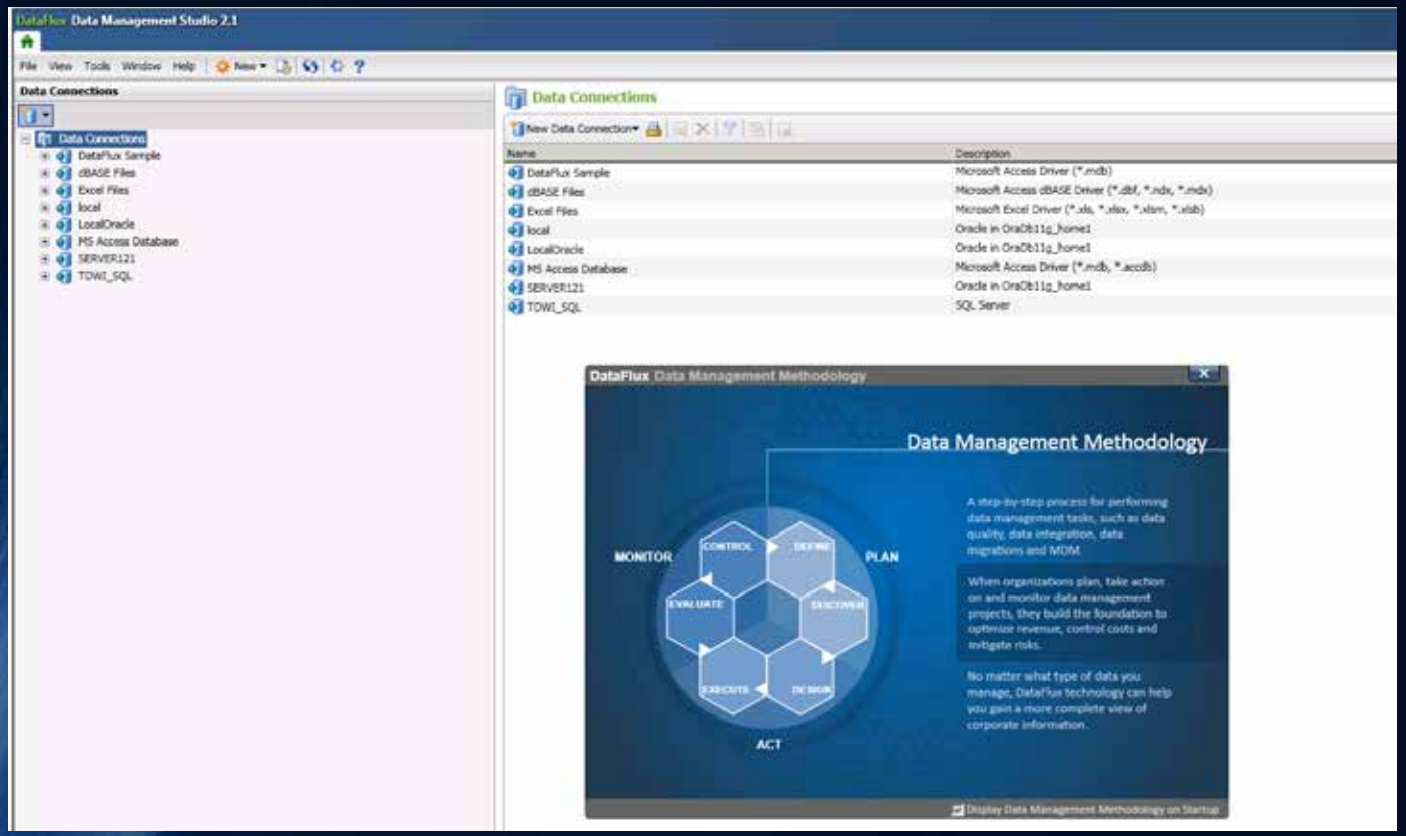
Main points I hoped to cover:

- Data Profiling is a valuable exercise, but it has its place, its limitations
- Biggest risk is overwhelming project sponsors with many reports which, if not carefully disseminated can obfuscate rather than clarify the data and the state of the data
- Questions/feedback?

Dataflux, Talend, Trillium

We’ve attached some screen shots and notes for you to read later at the bar or on the plane back home

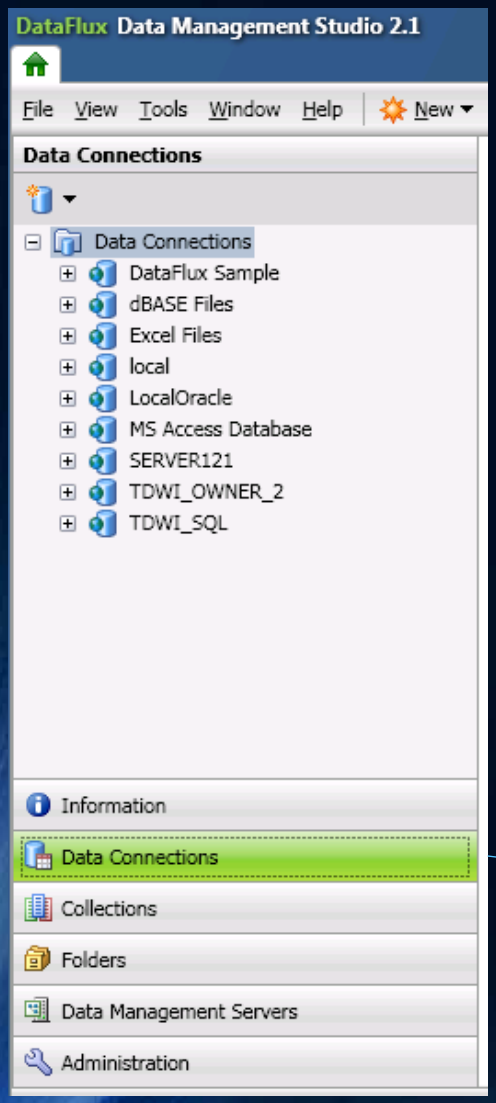
DataFlux Demo



Desktop looks like this, this product here is called the Data Management Studio.

DataFlux Demo

One of the nicer features of DataFlux is that it recognizes your already existing locally defined data connections and allows you to begin work there without spending time defining them.

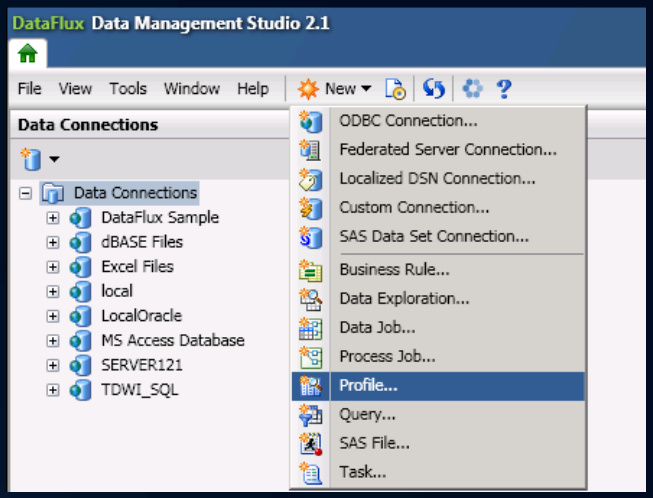


These risers are part of the navigation through DataFlux

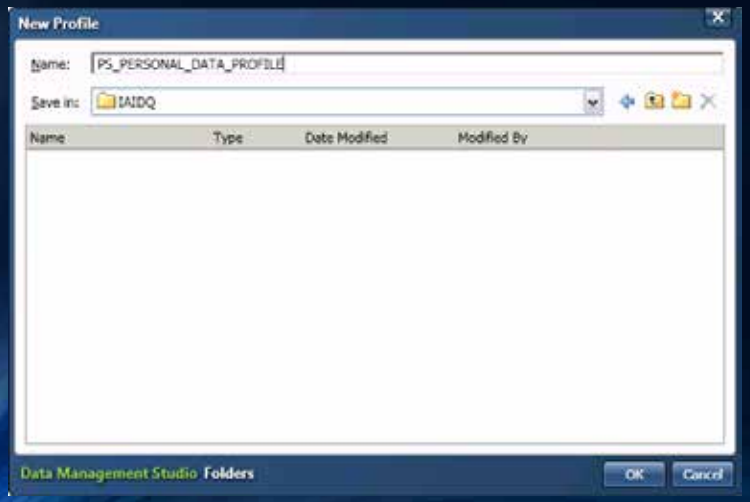
DataFlux Demo

Let's Profile a data table

Let's use the PS_PERSONAL_DATA table since it has a lot of recognizable data fields. From the main screen, select new-> profile



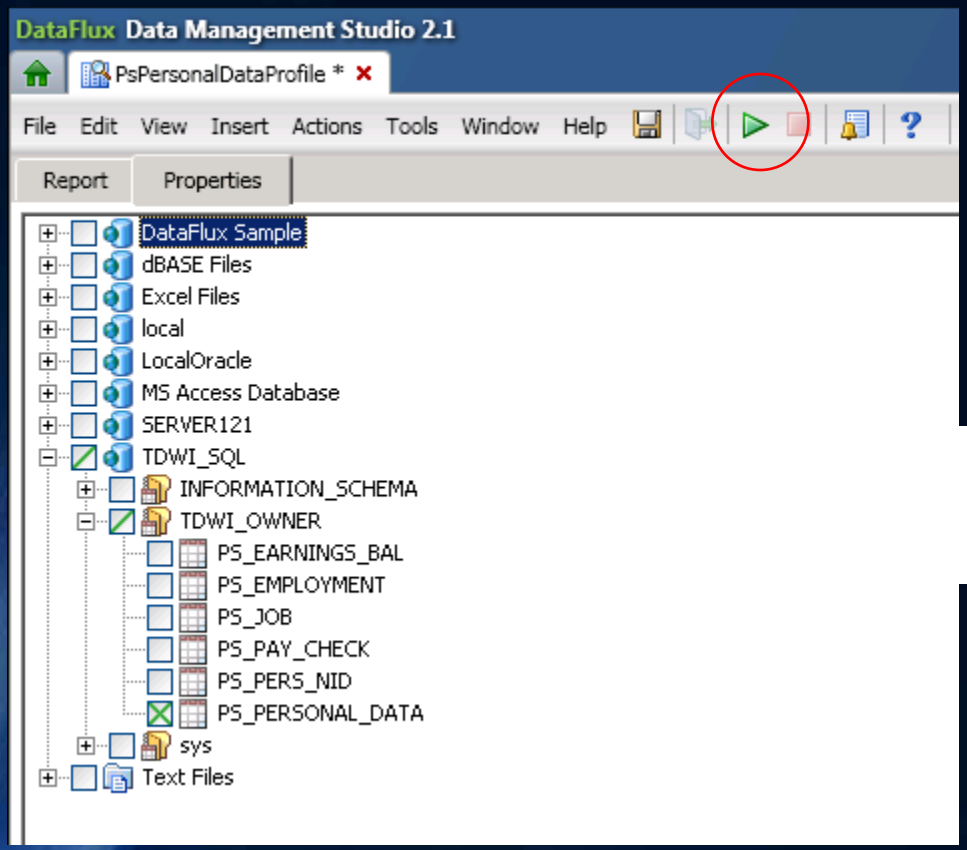
First you'll be prompted for a name and a folder to save it in.

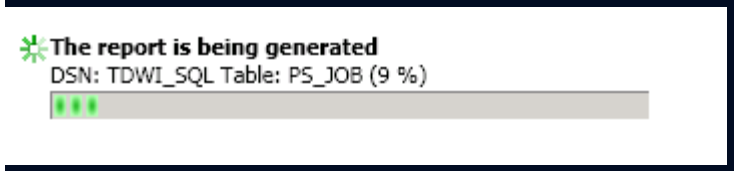


DataFlux Demo

Drill down to the table you want to profile

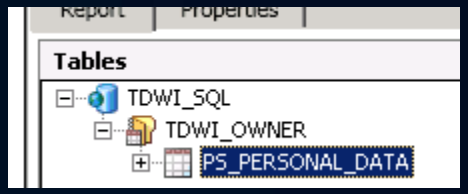
- Check the box next to the table we want to profile
- Press the green right arrow to run
- For some reason you'll be prompted for a description, enter something.
- Job will run in a few minutes





DataFlux Demo

Looking at the job profile results:
When the job is done, click the table name on the left, PS_PERSONAL_DATA in our example



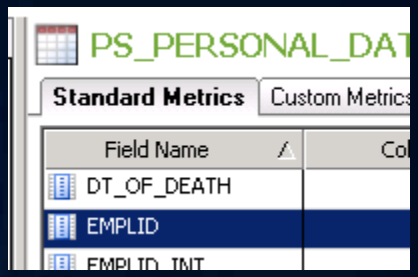
Let's click around here and see what jumps out
Below, we've picture the overview of the table, select a column on the left for more detail about that particular column.

PS_PERSONAL_DATA Schema: TDWI_OWNER Data Source: TDWI_SQL

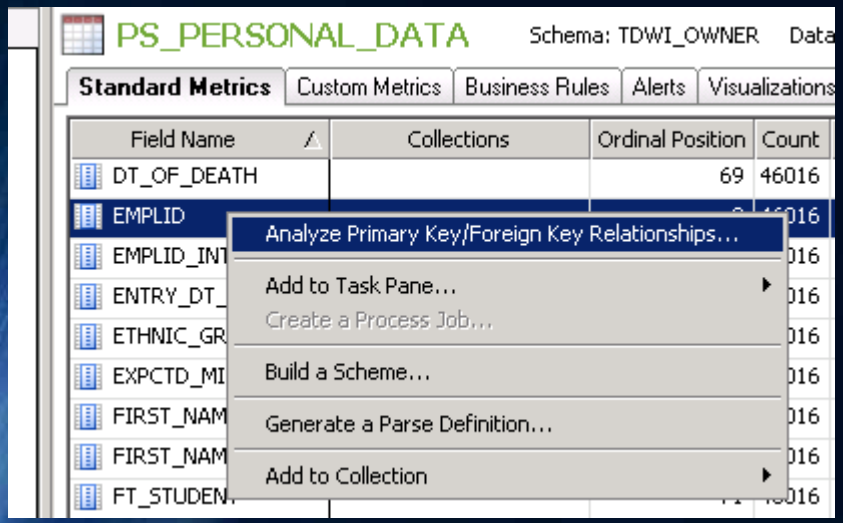
Field Name	Collections	Ordinal Position	Count	Null Count	Percent Null	Blank Count	Minimum Value
ADDRESS1		24	46016	850	1.8	0	1708 Sherman Pl #6
ADDRESS1_AC		100	46016	46016	100	0	
ADDRESS1_OTHER		41	46016	45896	99.7	0	10211 North 105th. Drive
ADDRESS2		25	46016	44437	96.6	0	Space 80
ADDRESS2_AC		101	46016	46016	100	0	
ADDRESS2_OTHER		42	46016	46001	100	0	#1016
ADDRESS3		26	46016	46000	100	0	31701 Blagnac Cedex
ADDRESS3_AC		102	46016	46016	100	0	
ADDRESS3_OTHER		43	46016	46014	100	0	BUILDING 7615
ADDRESS4		27	46016	46016	100	0	
ADDRESS4_OTHER		44	46016	46016	100	0	

DataFlux Demo

We can also profile the relationships between two or more fields in a table.
Start by selecting a field name from the list under Standard Metrics (we're checking EMPLID vs EMPLID_INT here)

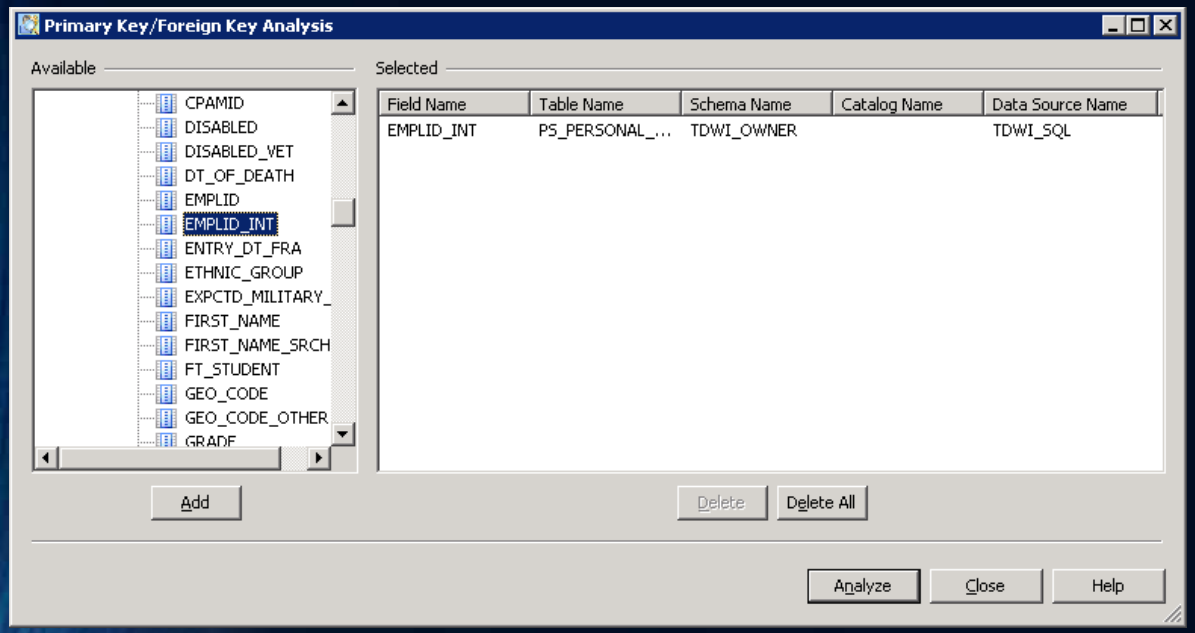


Right click and select Analyze Primary Key/Foreign Key Relationships...

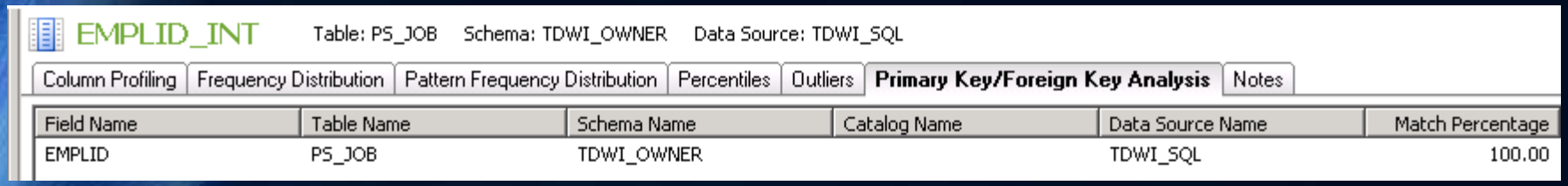


DataFlux Demo

Add the EMPLID_INT field to those to which we will compare EMPLID.



The results validate that EMPLID is always the same as EMPLID_INT



DataFlux Demo

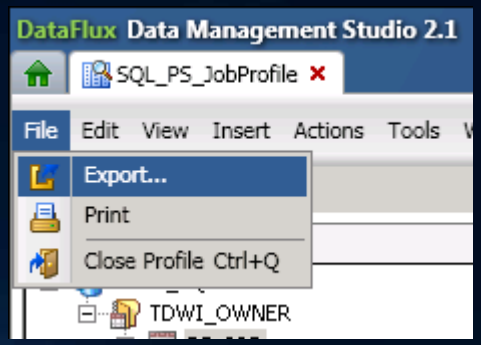
EMPLID_INT Table: PS_JOB Schema: TDWI_OWNER Data Source: TDWI_SQL

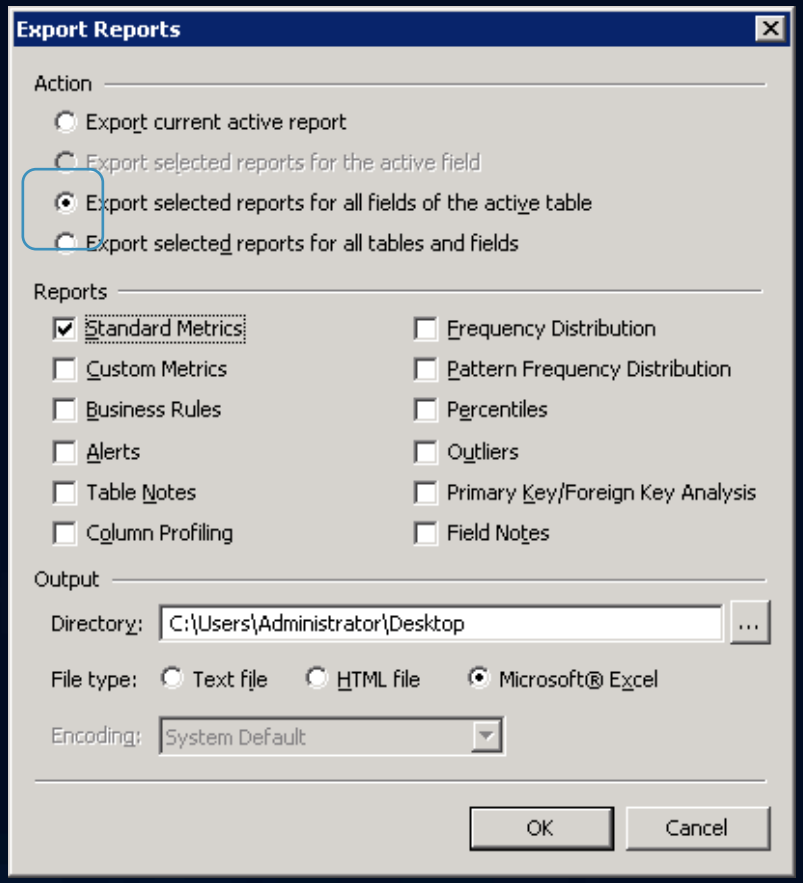
Column Profiling | Frequency Distribution | Pattern Frequency Distribution | Percentiles | Outliers | **Primary Key/Foreign Key Analysis** | Notes

Field Name	Table Name	Schema Name	Catalog Name	Data Source Name	Match Percentage
EMPLID	PS_JOB	TDWI_OWNER		TDWI_SQL	100.00

DataFlux Demo

- Finally, we can save each and all of these reports to Excel for easy distribution.
- In fact, you can schedule a job to run this profile on a schedule, create the excel report and email it around.
- Start by selecting Export… from the file menu.
- Configure the next menu like this. Careful, if you select all tables and fields you'll get 1500 excel reports.





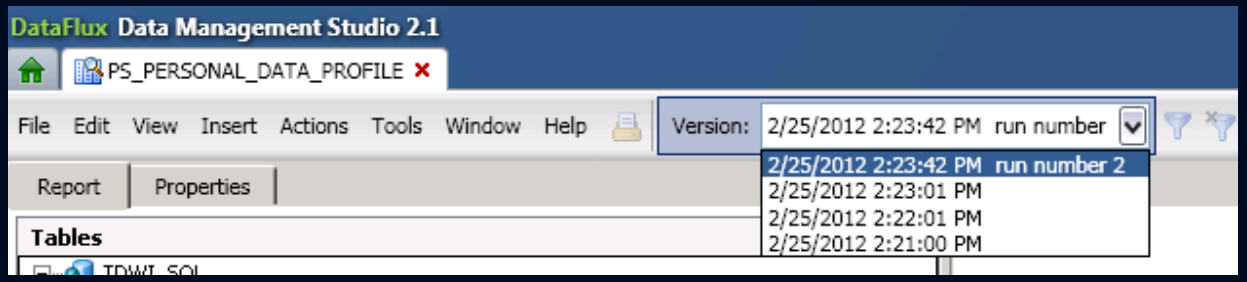
DataFlux Demo

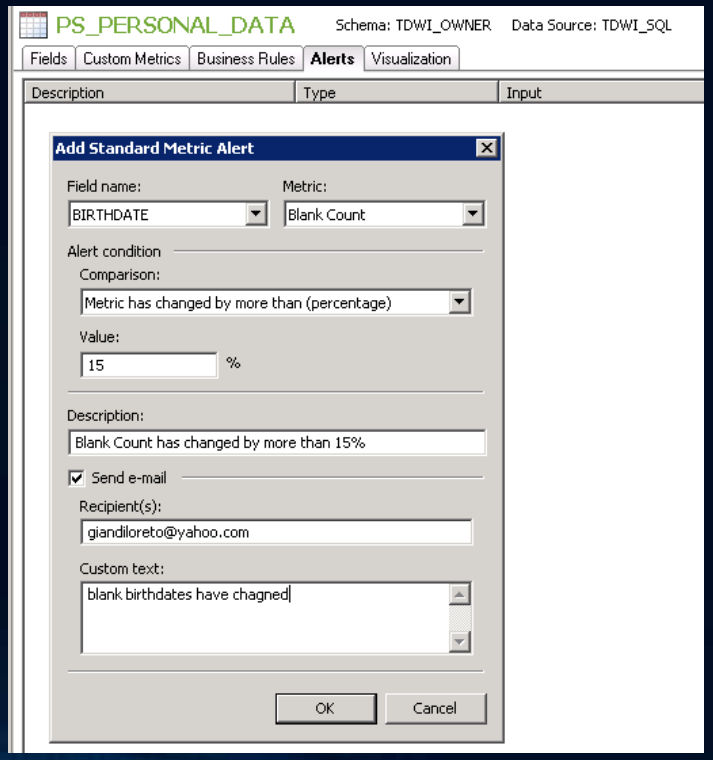
- It takes a few seconds, but you'll get a nice excel report with some useful stats.
- I find this report useful for a file delivery, it provides a good overview of the structure of the data, max's and min's number of unique and null values.

Field Name	Collections	Ordinal Position	Count	Null Count	Percent Null	Blank Count	Minimum Value	Maximum Value	Mode	Pattern Count	Unique Count	Uniqueness	Prim
ACCDNT_CD_FRA		85	77185	77185	100	0			(no data/ambig.)	0	0	0	no
ACCT_CD		50	77185	308	0.4	0	590	Z07Z07DISAB	107	79	7123	9.27	no
ACTION		12	77185	0	0	0	CNV	XFR	PAY	1	31	0.04	no
ACTION_DT		13	77185	0	0	0	0002-09-27	1/13/2011	12/23/2000	1	5532	7.17	no
ACTION_REASON		14	77185	0	0	0	ACQ	XYP	MER	2	211	0.27	no
ADDS_TO_FTE_ACTUAL		199	77185	0	0	0	N	N	N	1	1	0	no
ANNL_BENEF_BASE_RT		64	77185	0	0	(not applicable)	0	890384	0	(not applicable)	16972	21.99	no
ANNL_BEN_BASE_OVRD		164	77185	0	0	0	N	N	N	1	1	0	no
ANNUAL_RT		60	77185	0	0	(not applicable)	0	410000	22880	(not applicable)	15266	19.78	no
AS_ALTERNATE_SCHED		224	77185	76649	99.3	0	3D	S4	S4	4	17	3.17	no
AS_ATSC_PAY_CD		229	77185	72423	93.8	0	1	NA	1	4	19	0.4	no
AS_GL_PAY_TYPE		231	77185	0	0	0	1	ZW	4K	4	381	0.49	no
AS_INCENTIVE_TYPE		223	77185	0	0	0	100	T27	NON	4	109	0.14	no
AS_MATRIX		220	77185	68222	88.4	0	0	X-00	N-00	16	126	1.41	no
AS_MATRIX_ENTRY_DT		221	77185	72235	93.6	0	1/1/1900	10/5/2009	6/5/1992	1	225	4.55	no
AS_PAY_CODE		222	77185	49	0.1	0	D	I	D	1	2	0	no
AS_PAY_DISTRIBUTN		230	77185	67273	87.2	0	0	H370230	3440000	15	713	7.19	no
AS_PLANT		219	77185	0	0	0	1	WA	80	5	373	0.48	no
AS_REC_CHG_DTTM		235	77185	0	0	0	8/27/1921	1/13/2011	12/23/2000	1	5175	6.7	no
AS_REC_CREATE_DTTM		236	77185	0	0	0	8/27/1921	1/13/2011	12/23/2000	1	5522	7.15	no
AS_SITE_LOC_ID		232	77185	0	0	0	AF01	XX999	XX999	4	555	0.72	no
AS_UNION_DUE		227	77185	76571	99.2	0	A	Y	A	1	4	0.65	no
AS_UNION_ENTRY_DT		225	77185	2660	3.4	0	7/26/1933	1/18/2011	12/18/1997	1	3302	4.43	no
AS_UNION_TIME		226	77185	77185	100	0			(no data/ambig.)	0	0	0	no
BARG_UNIT		104	77185	61889	80.2	0	Z	Z	Z	1	1	0.01	no
BAS_ACTION		36	77185	0	0	0	CNV	XFR	PAY	1	32	0.04	no

DataFlux Demo

- DataFlux supports time dependent data profiling.
- You can program a job to profile a table for example every 24 hours and even to send you an alert if a metric changes based on your input.

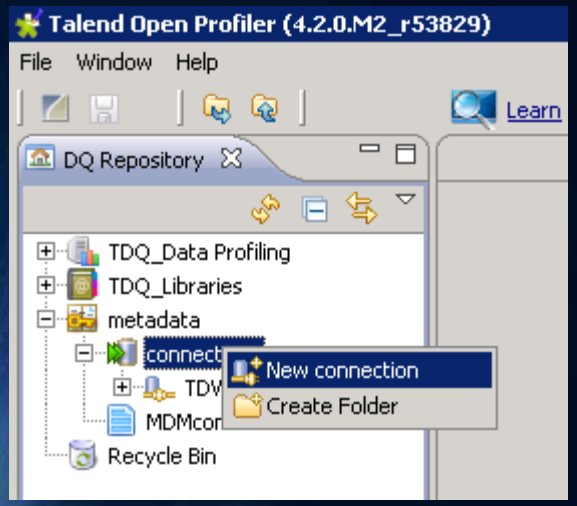




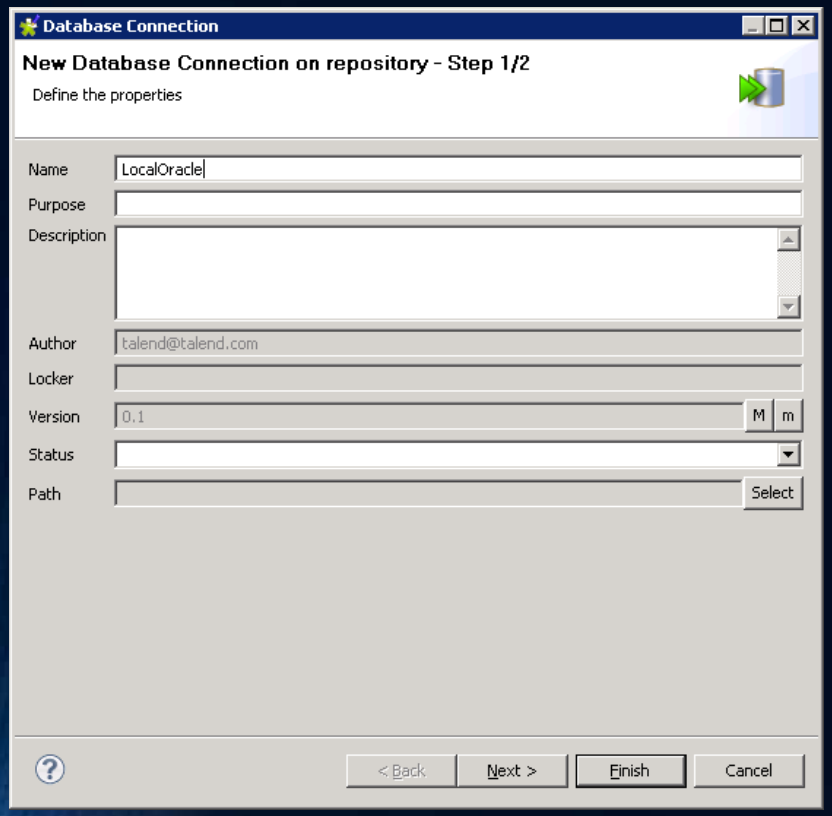
Talend

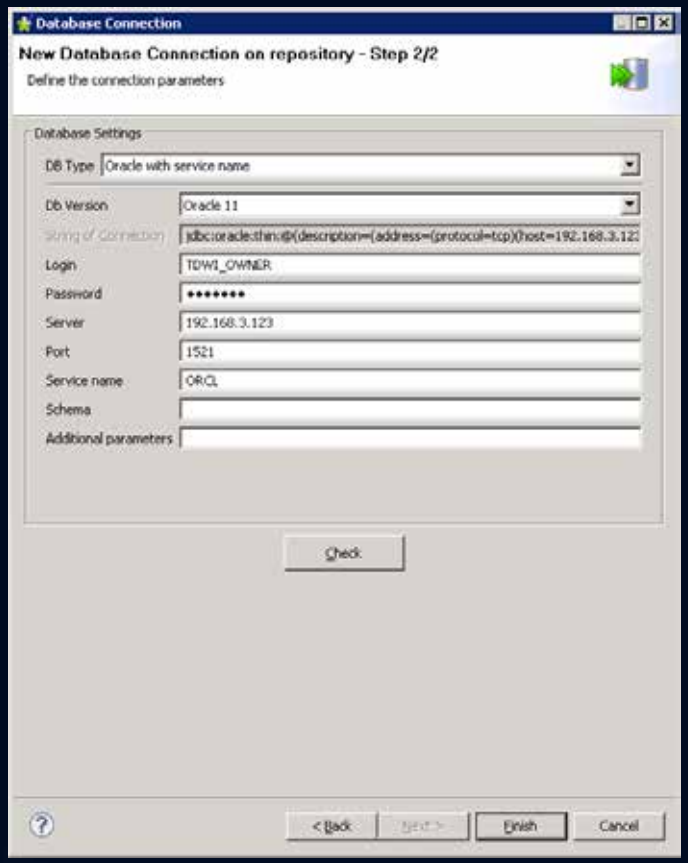
The freeware product we will demo today is called Talend Open Studio for Data Quality (TOS-DQ)

- We'll start with a simple data profile
- We need to point the product to our external data sources, in this case we'll use Oracle Data
- Create a New connection



TOS-DQ Profile Example

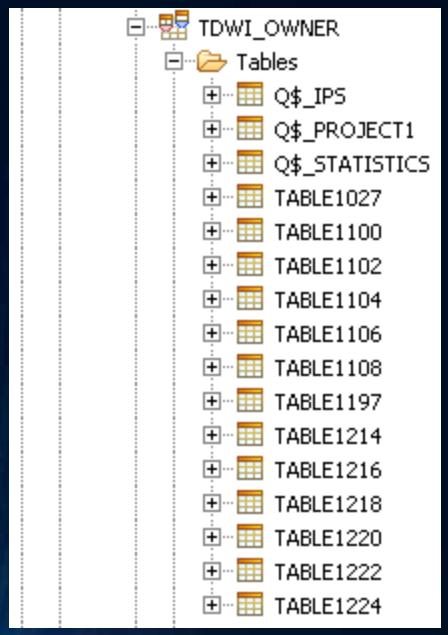




We will be connecting to our small server's Oracle instance. IP Address = 192.168.3.123

Password is 'forward'
Hit check to test connection

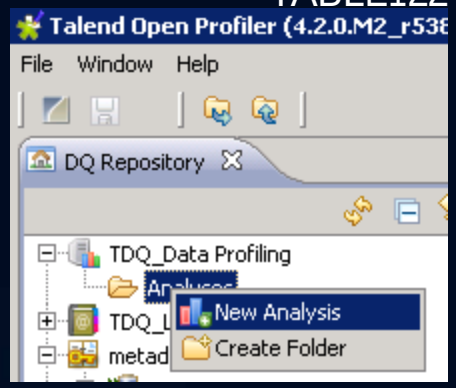
TOS-DQ Profile Example



Expand to see list of tables. We will want to analyze table 1224, the PERSONAL_DATA table

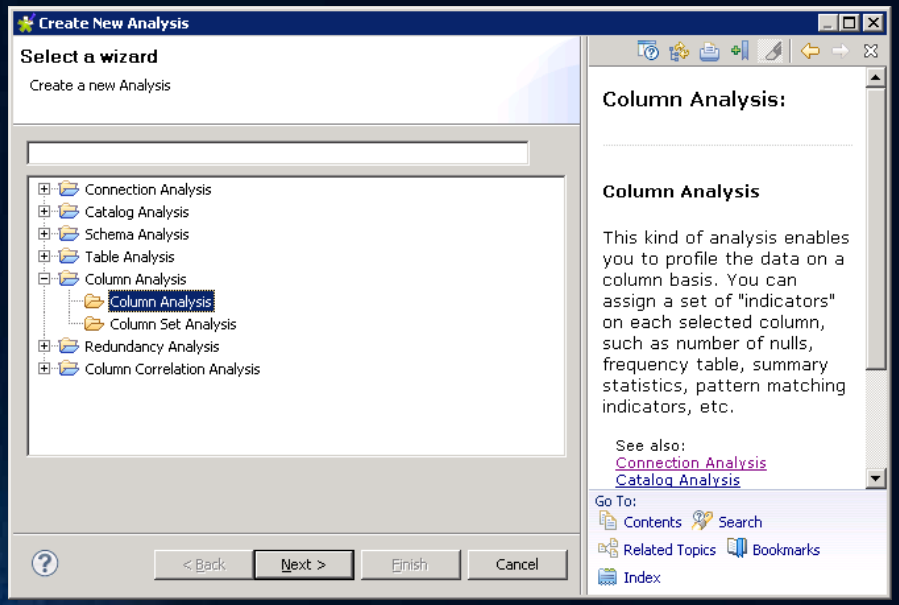
Oracle Tables:

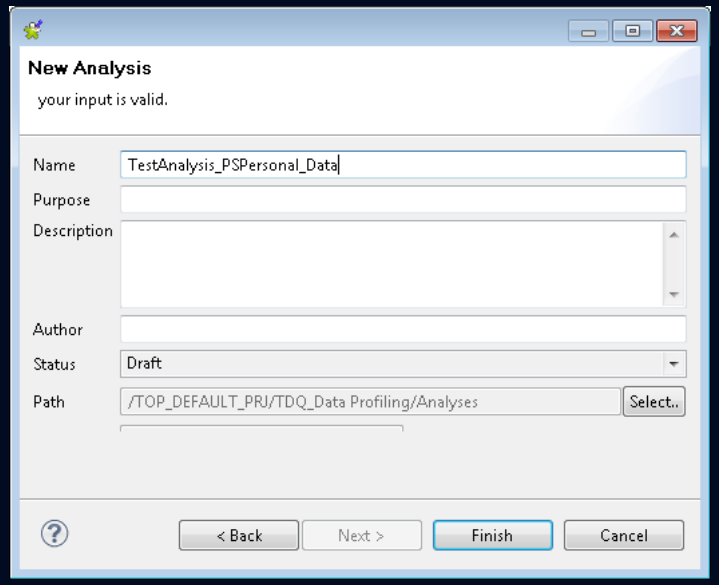
PS_JOB - TABLE1214
PS_EMPLOYMENT – TABLE1216
PS_EARNINGS_BAL – TABLE1218
PS_PAY_CHECK – TABLE1220
PS_PERS_NID – TABLE1222
PS_PERSONAL_DATA – TABLE1224



To do this, goto TDQ_Data Profiling and select New Analysis

TOS-DQ Profile Example

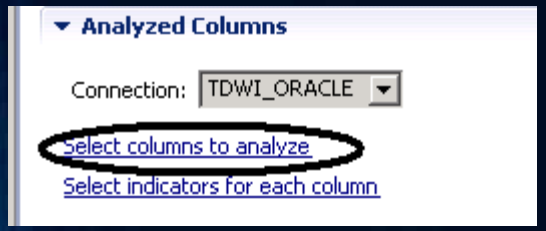




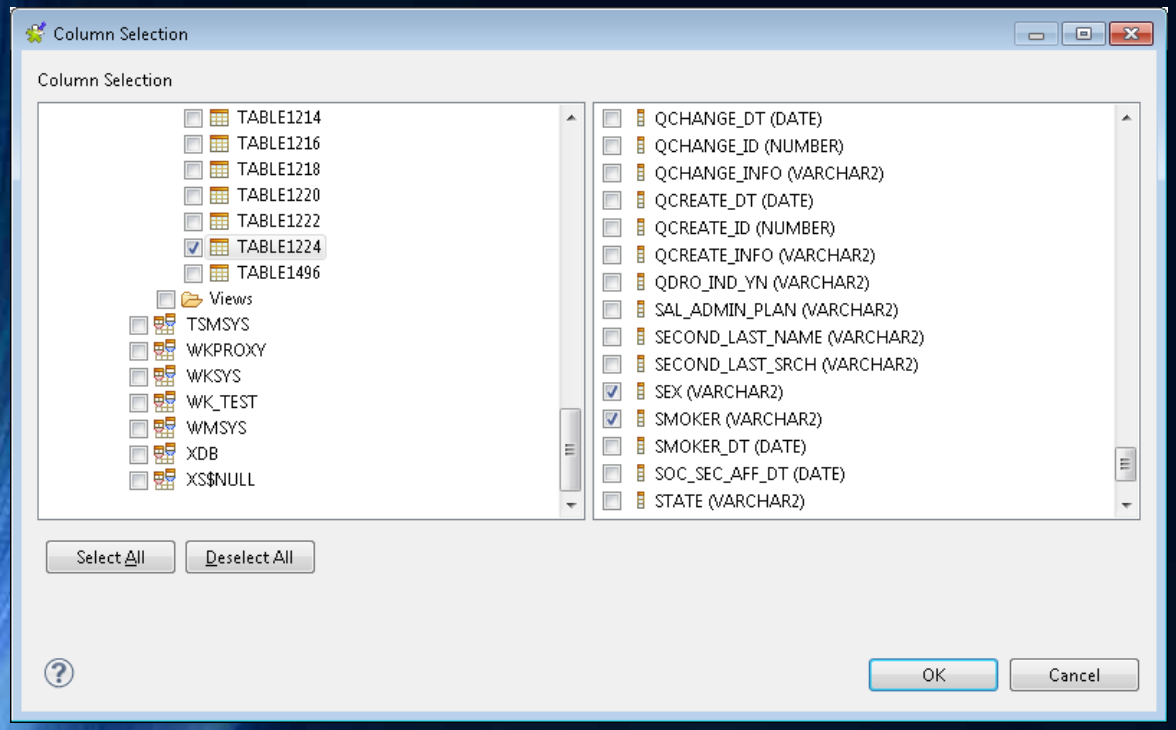
Select ‘Column Analysis’

Give it any name you like

TOS-DQ Profile Example



Select columns to analyze

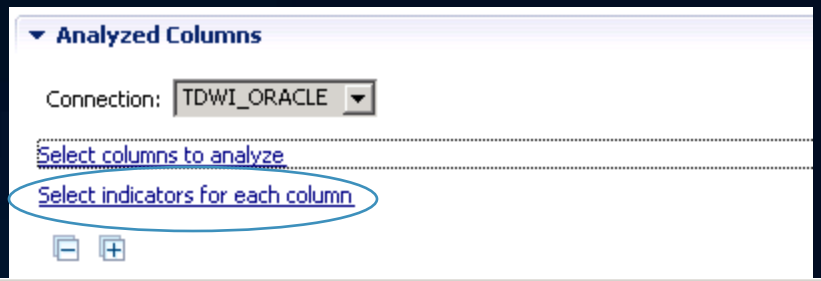


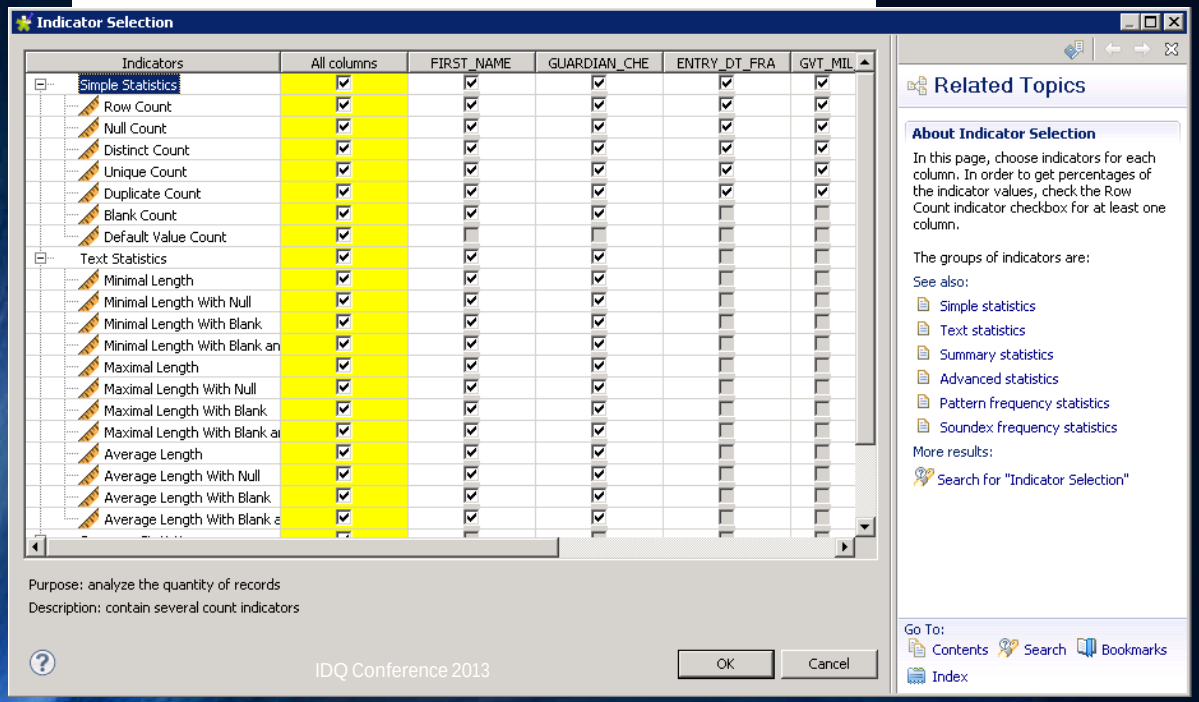
Select TABLE1224
This table has very many columns; select 15 or 20 or so.

TOS-DQ Profile Example

Now we need to tell Talend what to analyze, by default it will do nothing and complain that you didn't set the 'indicators'.

Select the hyperlink (click on) 'Select indicators for each column'. This will allow you to analyze specific things for each column you selected in the prior step.

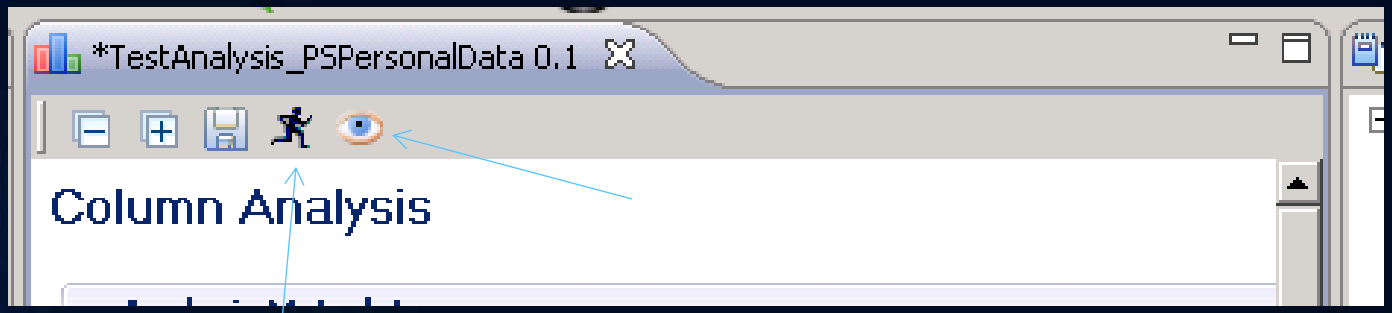




You can select all columns here for simplicity.

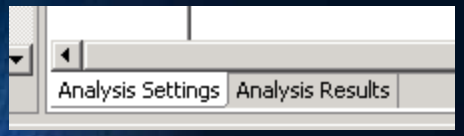
The help window on the right will tell you about the different kinds of analyses available.

TOS-DQ Profile Example



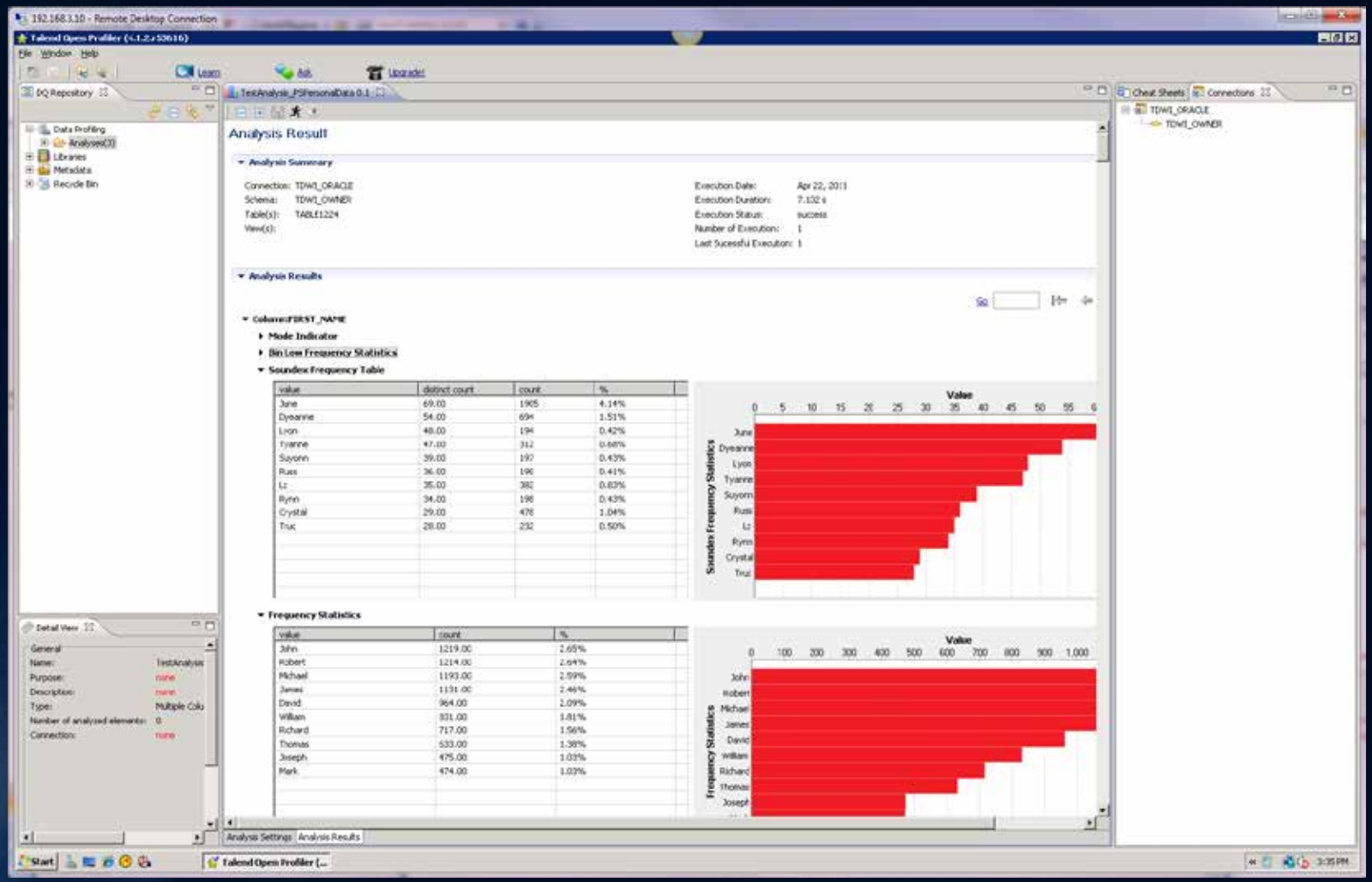
Run Job

The running man runs the job and the eye brings up the results.



Results are also visible from the Analysis Results tab.

TOS-DQ Profile Example



The results screen is quite dense, let's go through it together

TOS-DQ Profile Example

- Just for fun, let's profile another table that we derived from the personal data table.
- We created a table that has two columns, one for SSN and one for FIRST_NAME | LAST_NAME | BIRTHDATE
- We can use Talend to look for duplicate entries in this table

Column:NAME_DOB_GIAN.EMPLID

Simple Statistics

Label	Count	%
Row Count	46016.00	100.00%
Null Count	0.00	0.00%
Distinct Count	45983.00	99.93%
Unique Count	45950.00	99.86%
Duplicate Count	33.00	0.07%
Blank Count		0.00%

View rows
View values

Column:NAME_DOB_GIAN.NAME_DOB

Simple Statistics

Label	Count	%
Row Count	46016.00	100.00%
Null Count	0.00	0.00%
Distinct Count	46014.00	99.99%
Unique Count	46012.00	99.99%
Duplicate Count	2.00	4.346E-3%
Blank Count	0.00	0.00%

One of the nicer features of Talend is that you can drill down to the data in question from the profile report and furthermore see the generic SQL query that produced that report.

TOS-DQ Profile Example

LocalOracle 0.1 | ProfileNameDOB 0.1 | SQL Editor (LocalOracle.Duplicate Count).sql

LocalOracle/TDWI_OWNER Limit Rows: 100

```
1 --Analysis: ProfileNameDOB ;
2 --Type of Analysis: Multiple Column Analysis ;
3 --Purpose:  ;
4 --Description:  ;
5 --AnalyzedElement: EMPLID ;
6 --Indicator: Duplicate Count ;
7 --Showing: View rows ;
8  SELECT * FROM "TDWI_OWNER"."NAME_DOB_GIAN" WHERE "EMPLID" IN  (SELECT
```

1 [SELECT * FROM "TDWI_OW...] | Messages

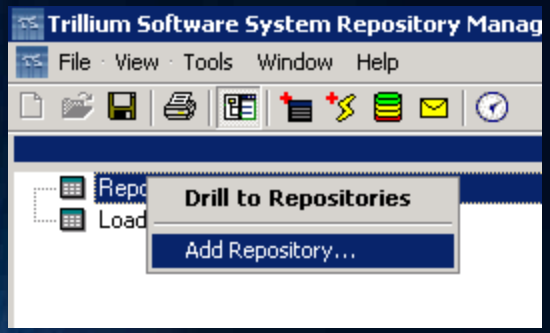
E...	NAME_DOB
103274706	HarrisMeryvalSat Nov 20 00:00:00 EST 1948
103274706	StollChristopherMon Aug 31 00:00:00 EDT 1959
105525276	MachiNicoloFri Apr 09 00:00:00 EST 1965
105525276	MachiNiconull
117325800	IyerBaskaranSat May 19 00:00:00 EDT 1962
117325800	IyerBaskarnull
126021024	BargerAndrewFri Sep 05 00:00:00 EDT 1986
126021024	BuckleyLeonFri Aug 14 00:00:00 EDT 1942
14012952	BeanMichaelSat Mar 10 00:00:00 EST 1973
14012952	BeanMichaelnull
142679574	BustamanteBernardTue Jul 04 00:00:00 EDT 1961
142679574	BustamanteBernienull
16449384	KerneyTimothySun Sep 20 00:00:00 EDT 1953
16449384	TherrienRaymondMon Jun 03 00:00:00 EDT 1946
178814436	YenCheng LungThu Feb 11 00:00:00 EST 1960
178814436	YenEricnull
181655934	JohnsonDeniseSat Mar 15 00:00:00 EST 1958
181655934	ZupanJasonWed Sep 10 00:00:00 EDT 1980
188344398	AcostaDarleneThu Dec 29 00:00:00 EST 1966
188344398	AcostaDarlenenull

Here we can see we have some duplicate SSN's which actually is probably due to my scrambling algorithm, in addition to naturally occurring doubles.

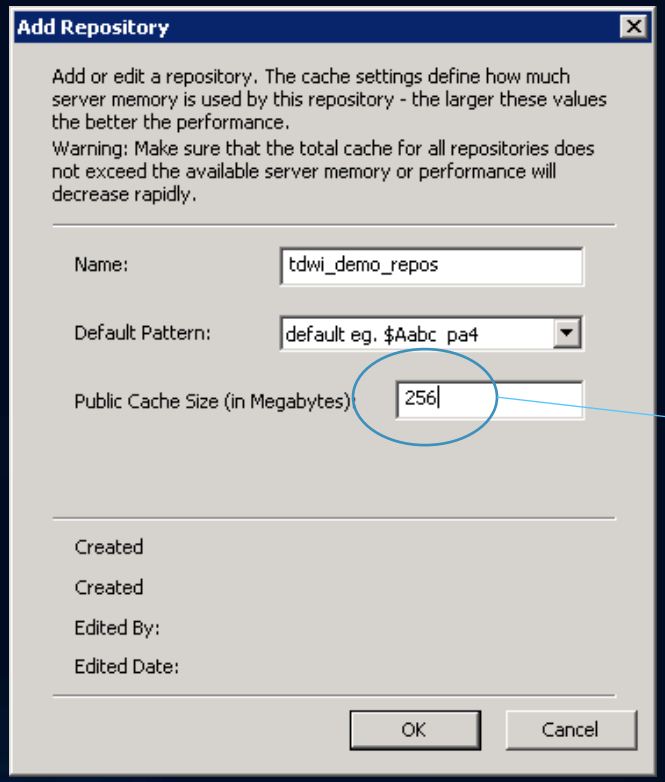
Trillium Product Demo

Trillium operates as a client/server, but for today's exercise, we have the source databases, the server and the client running on the same box

- Start by finding the Repository Manager and starting it.
- Once you're there, right click on Repositories and select Add Repository….
- User/pass is your username/username



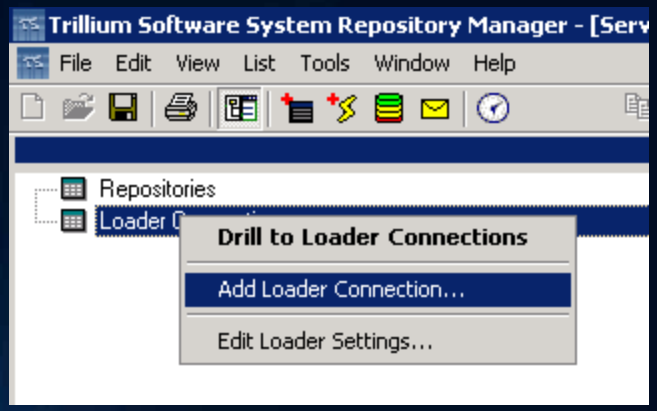
- Set it up like this:



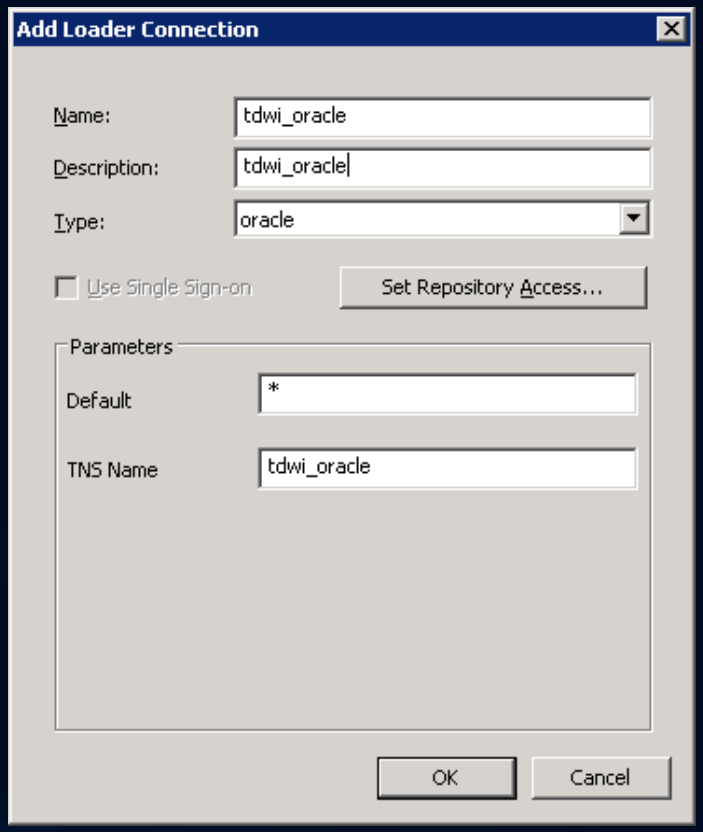
This public cache is interesting, it's the amount of memory allocated by the server to each connected client.

Trillium Product Demo

- Next, we will have to define the database connections we will use, you can also define flat file connections here.
- Right click on Loader Connections and select Add Loader Connection…

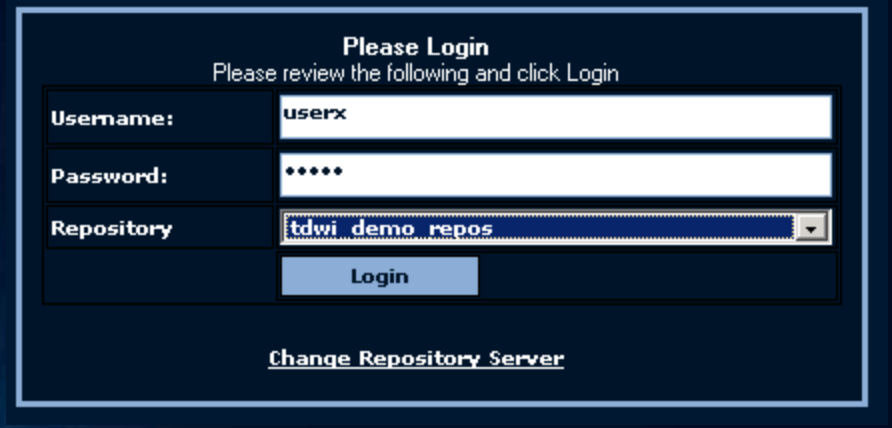


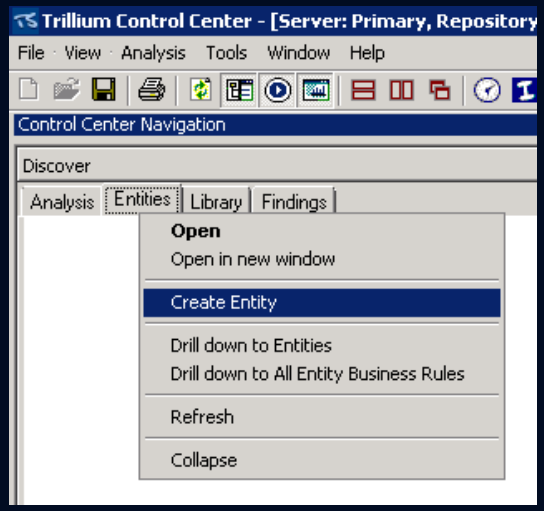
- And set it up like this:
- We will test in the next step



Trillium Product Demo

- Now we will exit out of the Repository Manager and start the TSS-13 Control Center.
- Find the repository and connect

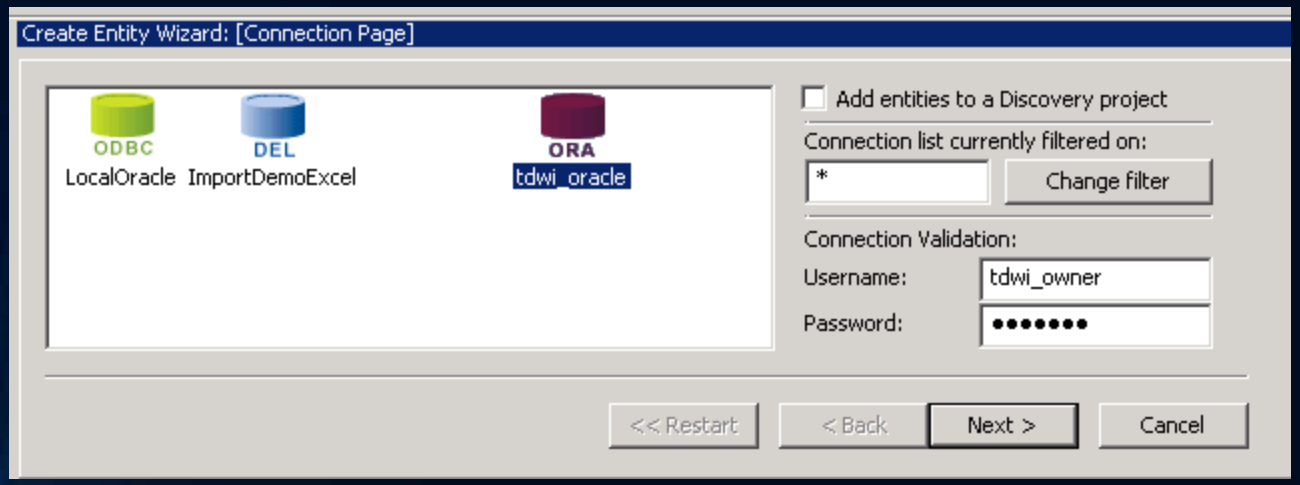




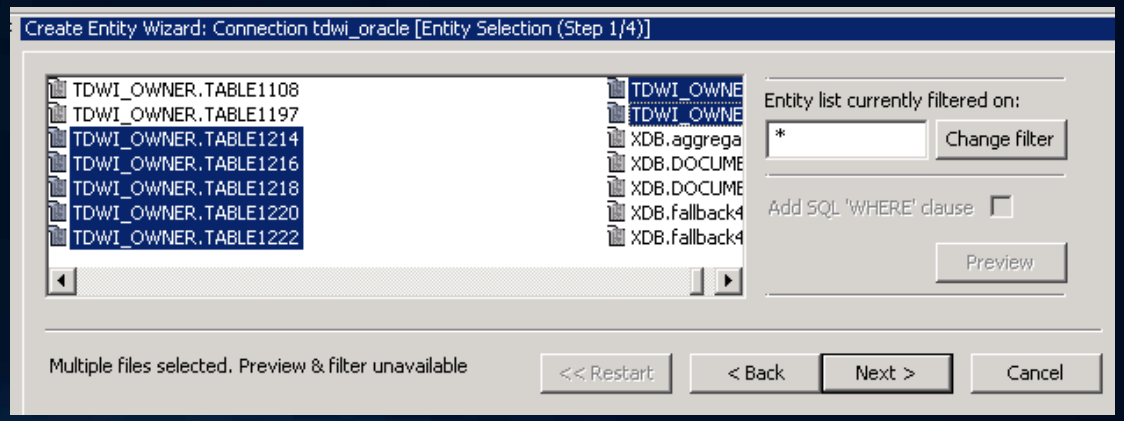
- We will now load some data. Trillium performs its analysis on the data as it is loaded. The parameters of this analysis are set to default values that can be adjusted based on your particular situation
- Select the 'Entities' tab, right click and select 'Create Entity'

Trillium Product Demo

- Select the relevant connection icon, type in credentials
- press Next >

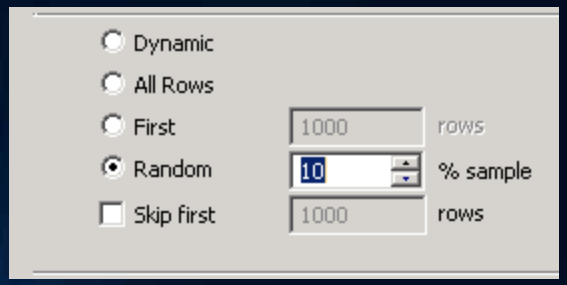


- Select tables TDWI_OWNER.TABLE1214-TABLE1224 (use your control key to select multiple tables)



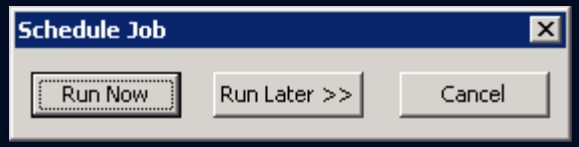
Trillium Product Demo

- IMPORTANT! To save time and space, we selected a 10% sample of the records from the source table.

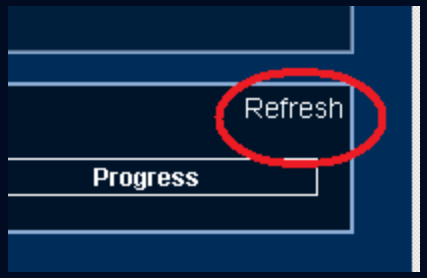


Trillium Product Demo

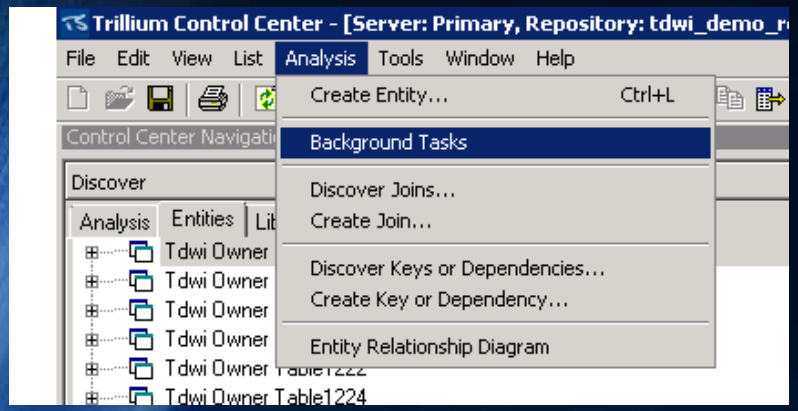
- Select finish on the next screen.
- Select Run Now



- Refresh the main screen and you'll see your jobs.



- This will take few minutes to load and analyze the tables, we'll jump to another repository already created and with its data analyzed.
- You can track the status if you select Analysis -> Background Tasks.



Or, you can press the clock icon

Trillium Product Demo

- When the jobs are done, there is a ton of metadata collected, getting through it can seem daunting at first.
- Start by getting back to the main screen and selecting ‘Entities’, pick a table (for example Tdwi Owner Table1216 and click on it.
- Select ‘Relationships’ and you’ll find the results of Trillium key analysis
- Here, the software has correctly identified *Emplid Int* and *Emplid* as table keys.

Getting Started | Keys

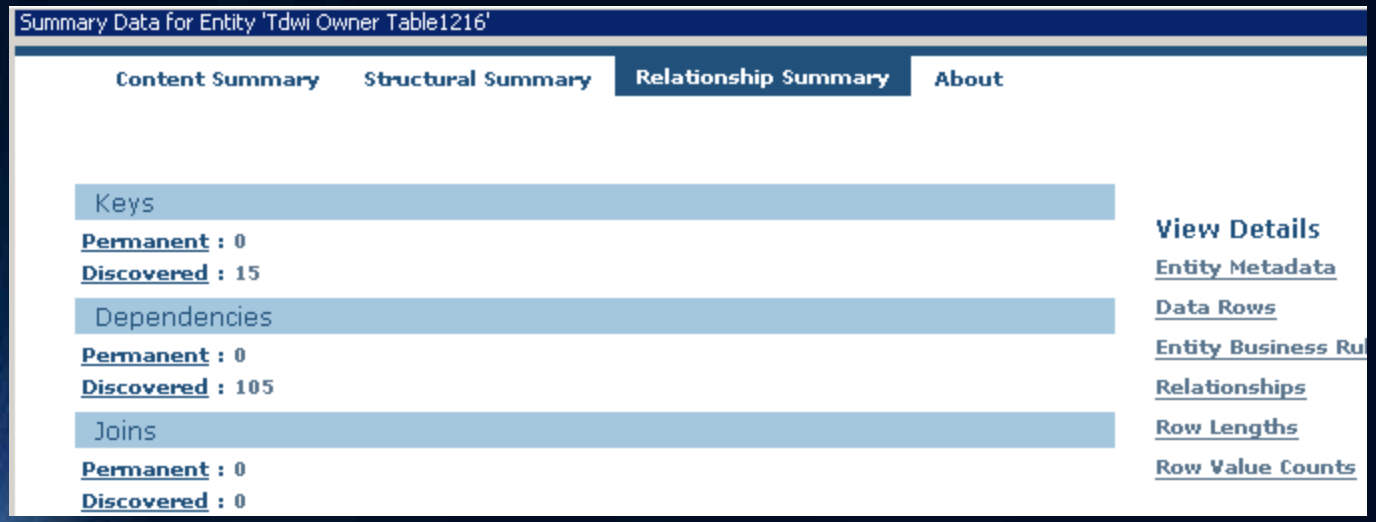
Keys

Entity = Tdwi Owner Table1216

Lh Attrs	Status	Verified	Ref	Quality %	Keys	Duplicate Keys	Duplicate Rows	Verified Date	Verified By
Emplid Int	Discovered	Yes	2	100.000	1211			2011/02/07 17:50:20	wstratton
Emplid	Discovered	Yes	2	100.000	1211			2011/02/07 17:50:20	wstratton
Hire Dt,As Rec Chg Dttm	Discovered	Yes	2	98.348	1179	12	32	2011/02/07 17:50:20	wstratton
Hire Dt,As Univ Id	Discovered	Yes	2	98.679	1181	14	30	2011/02/07 17:50:20	wstratton
Cmpny Seniority Dt,As Rec Chg Dttm	Discovered	Yes	2	99.670	1203	4	8	2011/02/07 17:50:20	wstratton
Cmpny Seniority Dt,Termination Dt	Discovered	Yes	2	99.174	1191	10	20	2011/02/07 17:50:20	wstratton
Cmpny Seniority Dt,Last Increase Dt	Discovered	Yes	2	98.431	1177	15	34	2011/02/07 17:50:20	wstratton
Cmpny Seniority Dt,As Badge Id	Discovered	Yes	2	98.431	1178	14	33	2011/02/07 17:50:20	wstratton

Trillium Product Demo

- One of the more interesting things Trillium uncovers is the relationships between different pairs of data elements, it picks up correlations.
- Like most of its results, Trillium over-shoots and some of what it picks up can be thrown away, but in my example it did uncover some non-obvious relationships between data elements.
- Go to Relationship Summary and select Discovered under Dependencies.



Trillium Product Demo

- Discovered Dependencies from table1216 (this is the PS personal data table)
- I clicked on the heading of the 'Quality' column to sort by this column.

Getting Started | Dependencies

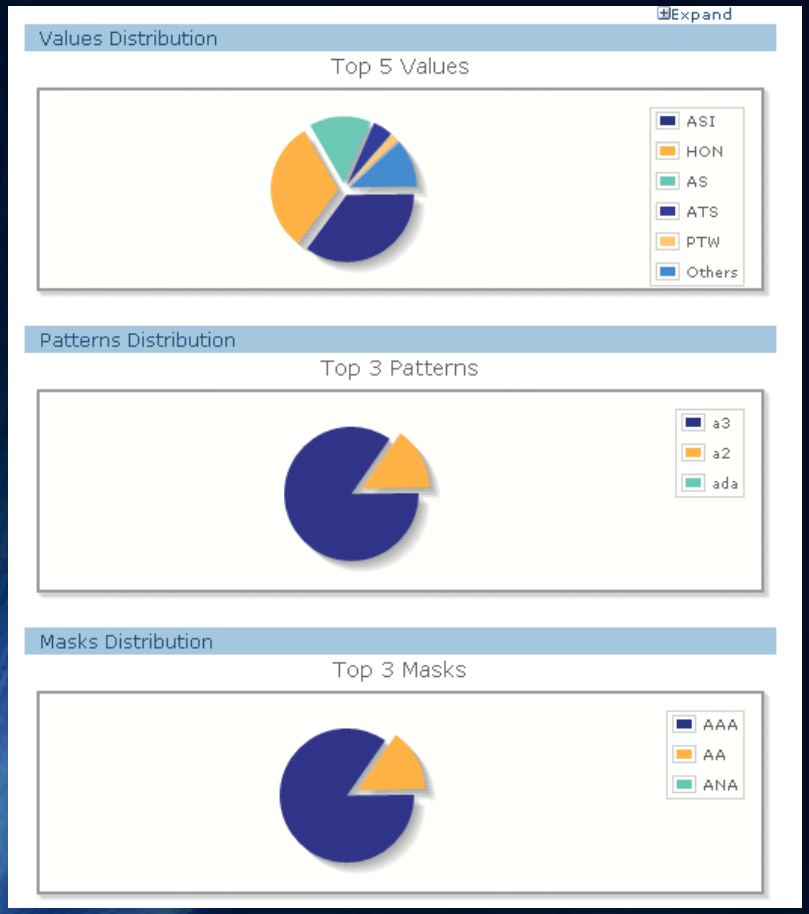
Dependencies
Entity = Tdwi Owner Table1216

Lh Attrs	Rh Attr	Status	Verified	Job	Q...	Confirming LR Values	Conflicting LH Values	Resolved %	Conflicting Rows	Verified Date	Verified By
As Badge Id,As Univ Id	As Room Mailstop	Discovered	Yes	2	100.000	1211				2011/02/07 17:50:21	wstratton
As Rec Create Dttm	Gvt Clnce Stat Dt	Discovered	Yes	2	100.000	1211				2011/02/07 17:50:21	wstratton
As Univ Id	H Mip Eligible Dt	Discovered	Yes	2	100.000	1211				2011/02/07 17:50:21	wstratton
As Univ Id	H Mip Elig End Dt	Discovered	Yes	2	100.000	1211				2011/02/07 17:50:21	wstratton
As Univ Id	Gvt Clnce Stat Dt	Discovered	Yes	2	100.000	1211				2011/02/07 17:50:21	wstratton
As Badge Id	Expected Return Dt	Discovered	Yes	2	99.917	1210	1		2	2011/02/07 17:50:21	wstratton
As Badge Id,As Univ Id	As Rec Create Dttm	Discovered	Yes	2	99.917	1210	1		2	2011/02/07 17:50:21	wstratton
As Badge Id,As Univ Id	As Building	Discovered	Yes	2	99.917	1210	1		2	2011/02/07 17:50:21	wstratton
As Univ Id	As Vac Elig Dt	Discovered	Yes	2	99.917	1210	1		2	2011/02/07 17:50:21	wstratton
Hire Dt,Service Dt	Cmpny Seniority Dt	Discovered	Yes	2	99.917	1210	1		2	2011/02/07 17:50:21	wstratton
Termination Dt,As Univ Id	As Rec Create Dttm	Discovered	Yes	2	99.917	1210	1		2	2011/02/07 17:50:21	wstratton
Termination Dt,As Univ Id	Last Date Worked	Discovered	Yes	2	99.917	1210	1		2	2011/02/07 17:50:21	wstratton
Phone	H Mip Elig End Dt	Discovered	Yes	2	99.835	1209	1		3	2011/02/07 17:50:21	wstratton
Termination Dt,As Univ Id	Supervisor Id	Discovered	Yes	2	99.835	1209	2		4	2011/02/07 17:50:21	wstratton
Hire Dt,As Badge Id	Rehire Dt	Discovered	Yes	2	99.752	1208	3		6	2011/02/07 17:50:21	wstratton
Hire Dt,As Badge Id	As Building	Discovered	Yes	2	99.752	1208	3		6	2011/02/07 17:50:21	wstratton

- Go ahead and click around and see what else you can find, you can do no damage here.

Trillium Product Demo

- Click a field name on the left and wait a few seconds and you'll get some interesting breakdowns.



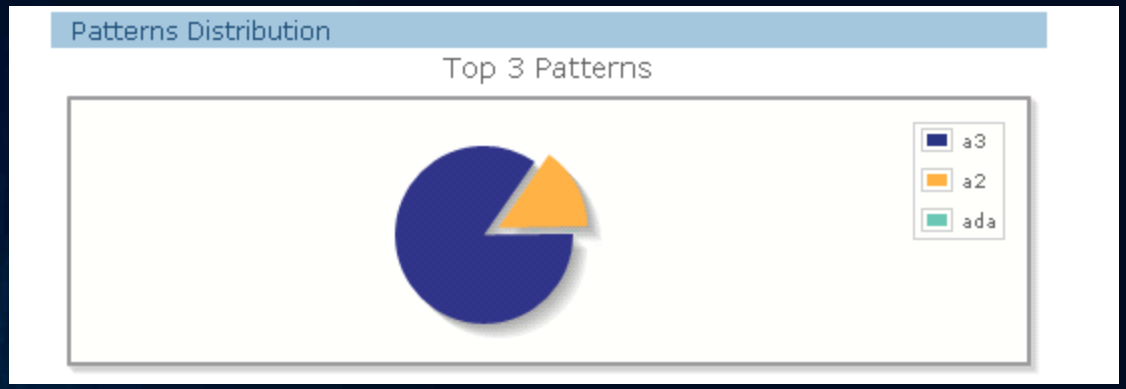
Breakdown of frequent values

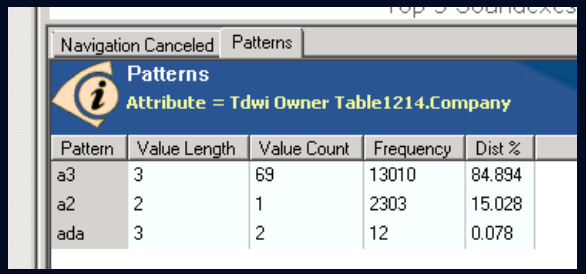
Breakdown of the pattern

Breakdown of the masks

Trillium Product Demo

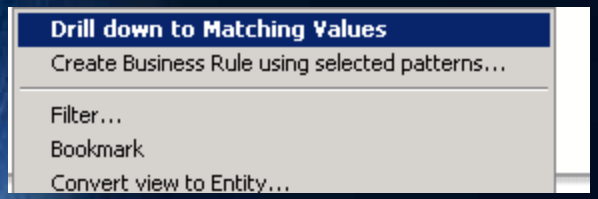
- Let's say you want to know what one of these breakdowns is tell you. Click on the diagram





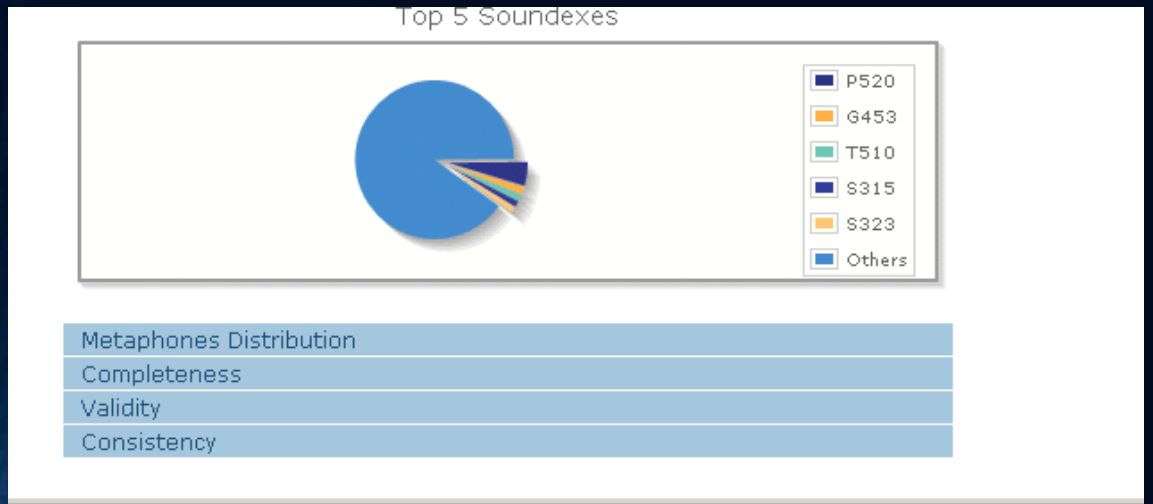
Pattern	Value Length	Value Count	Frequency	Dist %
a3	3	69	13010	84.894
a2	2	1	2303	15.028
ada	3	2	12	0.078

- Now, if you right click on the row listed, you can drill down to the data.



Trillium Product Demo

- The Soundex* analysis is interesting also. If you find a column with, say, a city name, like in our Table1224, you can see how the algorithm is used during the match address analysis.



Navigation Canceled | Values

Values
Attribute = Tdwi Owner Table1224.City

Value	Frequency	Dist %	Length	Soundex	Metaphone	Pattern	Mask
Phoenix	431	4.704	7	P520	FNKS	a7	AAAAAAA
PHOENIX	1	0.011	7	P520	FNKS	a7	AAAAAAA
phoenix	1	0.011	7	P520	FNKS	a7	AAAAAAA
Phoenix,	1	0.011	8	P520	FNKS	a7p	AAAAAAA,
Ponce	1	0.011	5	P520	PNS	a5	AAAAA

**Soundex* is a phonetic algorithm for indexing names by sound, as pronounced in English

Similar to soundex, *metaphone* creates the same key for similar sounding words. It's more accurate than soundex as it knows the basic rules of English

Trillium Product Demo

- Let's explore what Trillium refers to as business rules
- These can be defined at the entity (table) level or at the attribute (data element) level.
- While browsing I noticed some strange annual rate entries in the compensation table, attribute.

Entity Member Attribute | All Library Attributes

Entity Member Attribute

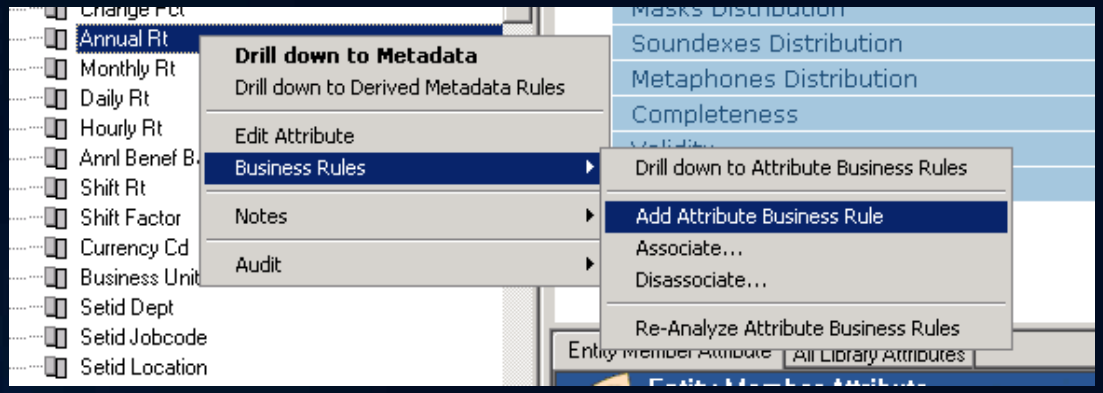
Attribute = Tdwi Owner Table1214.Annual Rt

Metadata	Value	Description
Strings Dist %	0	The percentage of string values.
Decimals	1585	The count of decimal values
Dec Dist %	32.924	The percentage of decimal values
Decimal Min	.01	The minimum decimal value
Decimal Max	213800.04	The maximum decimal value
Integers	3377	The count of integer values
Integer Dist %	67.076	The Percentage of integer values
Integer Min	0	The minimum integer value
Integer Max	410000	The maximum integer value

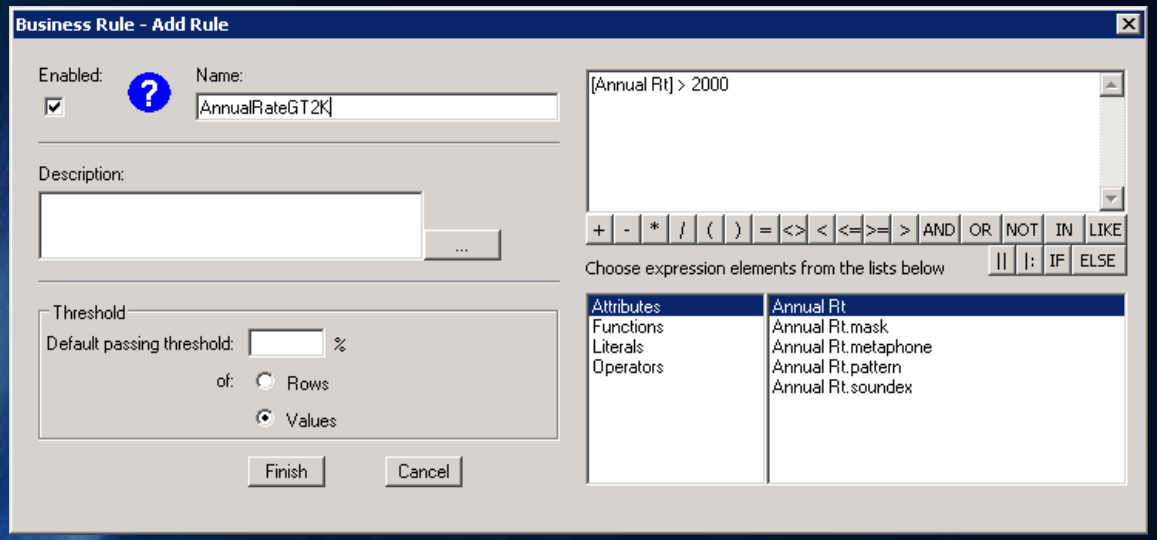
- The Min is $0.01, which even during our current economic situation, seems low for an annual rate. I assume this is 'rate' not actual compensation received.
- So I'd like to create a business rule to see how many tiny annual rates we have.

Trillium Product Demo

- To do so, right click on AnnualRt and Add Attribute Business Rule



- Configure thusly:

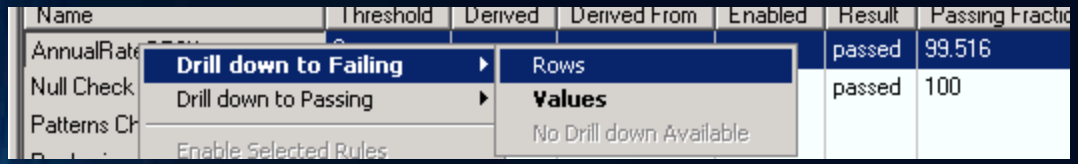


Trillium Product Demo

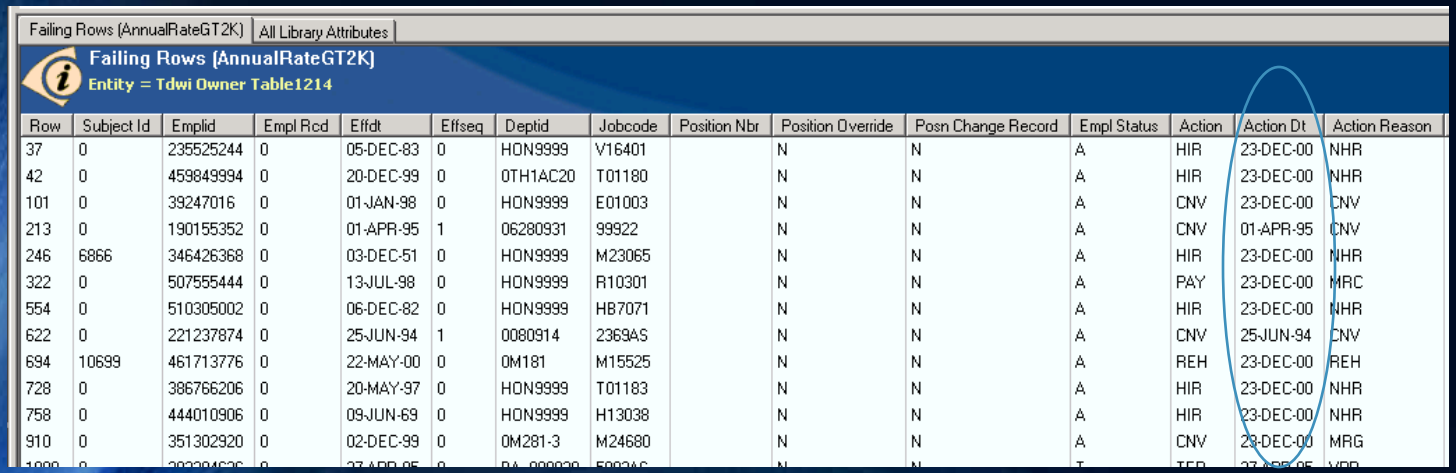
- The results suggest it is a good business rule

Name	Threshold	Derived	Derived From	Enabled	Result	Passing Fraction	Status	Created By	Date Created	Edited By	Date Changed
AnnualRateGT2K	0	no		yes	passed	99.516	analyzed	wstratton	2012/02/09 12:02:55	wstratton	2012/02/09 12:03:08

- Now the cool thing about Trillium is that you can quickly drill down and see the records that passed or failed this business rule. This allows you to research the 'bad' data, tweak your rule and so on.



- Notice anything?

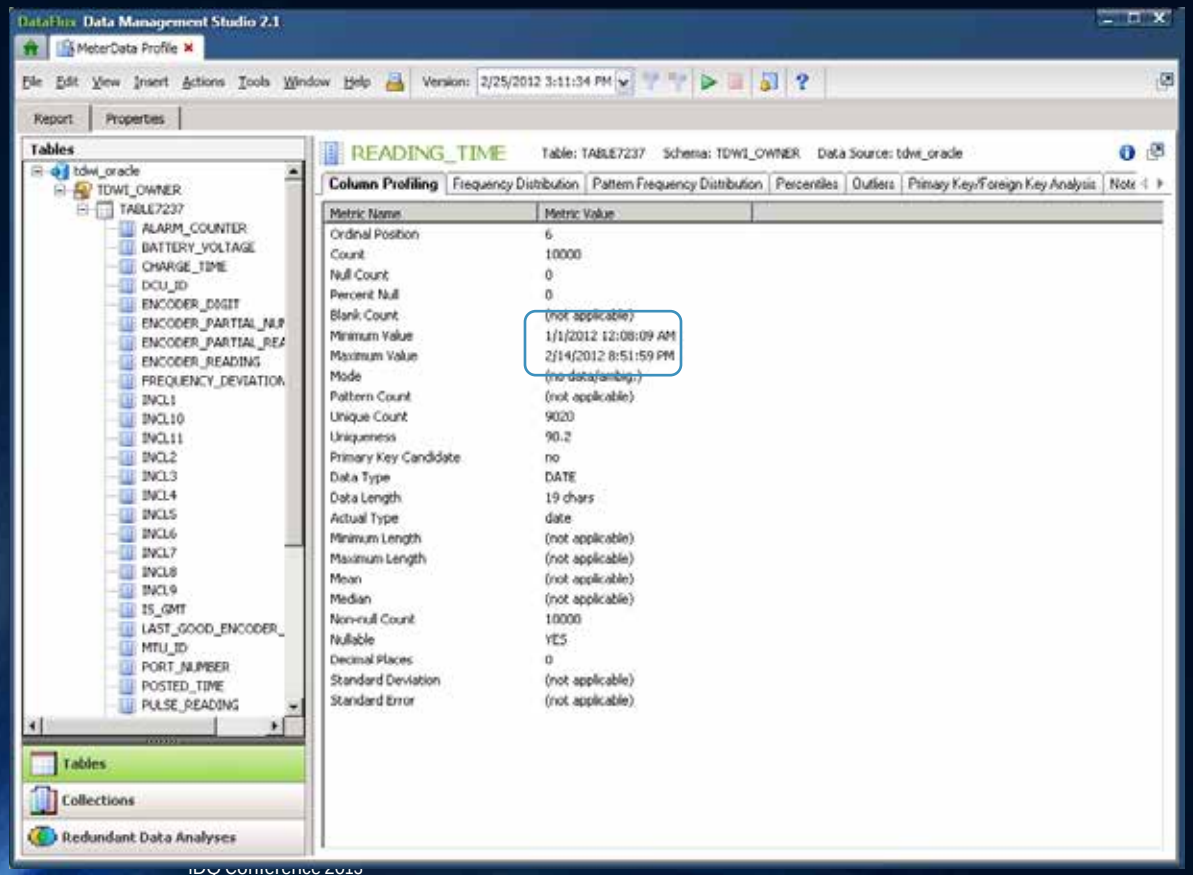


Row	Subject Id	Emplid	Empl Rcd	Effdt	Effseq	Deptid	Jobcode	Position Nbr	Position Override	Posn Change Record	Empl Status	Action	Action Dt	Action Reason
37	0	235525244	0	05-DEC-83	0	HON9999	V16401		N	N	A	HIR	23-DEC-00	NHR
42	0	459849994	0	20-DEC-99	0	0TH1AC20	T01180		N	N	A	HIR	23-DEC-00	NHR
101	0	39247016	0	01-JAN-98	0	HON9999	E01003		N	N	A	CNV	23-DEC-00	CNV
213	0	190155352	0	01-APR-95	1	06280931	99922		N	N	A	CNV	01-APR-95	CNV
246	6866	346426368	0	03-DEC-51	0	HON9999	M23065		N	N	A	HIR	23-DEC-00	NHR
322	0	507555444	0	13-JUL-98	0	HON9999	R10301		N	N	A	PAY	23-DEC-00	MRC
554	0	510305002	0	06-DEC-82	0	HON9999	HB7071		N	N	A	HIR	23-DEC-00	NHR
622	0	221237874	0	25-JUN-94	1	0080914	2369AS		N	N	A	CNV	25-JUN-94	CNV
694	10699	461713776	0	22-MAY-00	0	0M181	M15525		N	N	A	REH	23-DEC-00	REH
728	0	386766206	0	20-MAY-97	0	HON9999	T01183		N	N	A	HIR	23-DEC-00	NHR
758	0	444010906	0	09-JUN-69	0	HON9999	H13038		N	N	A	HIR	23-DEC-00	NHR
910	0	351302920	0	02-DEC-99	0	0M281-3	M24680		N	N	A	CNV	23-DEC-00	MRG

Subject level example

- What's missing from all these tools is a mechanism to process the data at the *subject* level rather than (or in addition to) the *record* level.
- The subject is the real life entity described by the data.

- As an example, we have a file of data containing electric usage data. The data contains a meter ID and a number of readings for each meter.
- 1st we profile this file and record the profiles of each data attribute (AKA column, field).
- In doing so we see that the data spans 6 weeks or so



Subject level example

- Now we used a different product to do the subject level profile and determine the breakdown of number of reads per meter (here meter is the subject)

Book1 - Microsoft Excel

	Frequency	Reads per meter
2	8	86
3	8	81
4	7	78
5	7	80
6	7	79
7	6	1
8	6	2
9	5	85
10	5	76
11	4	72
12	4	77
13	3	61
14	3	60
15	3	64
16	3	83
17	3	88
18	3	84
19	3	73
20	2	70
21	2	75
22	2	58
23	2	69
24	2	62

We see these outlier values and we were troubled.

We looked at the voltage of each of these meters with 1 or 2 reads only.

3.147, 2.714, 2.761, 2.762, and similar values

Looking now to the frequency distribution of the voltages, we can see these are all very low (with in the bottom 0.1%), so we discovered why these meters aren't giving out reads as often as they should. They have dead batteries.

BATTERY_VOLTAGE Table: TABLE7237 Schema: TDWI_OWNER Data Source: tdwi_oracle

Column Profiling | **Frequency Distribution** | Pattern Frequency Distribution | Percentiles | Outliers | Primary Key/Foreign Ke

Value	Count	Percentage
3.422	3450	34.50
3.499	2630	26.30
3.349	1843	18.43
3.579	1002	10.02
3.279	841	8.41
3.211	203	2.03
3.147	14	0.14
2.714	12	0.12
2.761	4	0.04
3.085	1	0.01

Subject level example – state transition analysis

- In order to study the time component we are required to look at subsequent records in the data ordered by time, for a specific subject. (subsequent records that cross into a different subject are meaningless)
- I do not know how to do this with any of the tools we have here, but I suspect it is possible, I am investigating. Meanwhile, I have a created a metadata table with a simple script and then profiled that to understand the state transition behavior of the JOB table.

```
sub StateTransistion()
  dim vwJob, vwJobStateTrans

  set vwJob = QSubject.GetView("PS_JOB","PS_JOB")
  set vwJobStateTrans = QSubject.GetView("STATE_TRANS_JOB","STATE_TRANS_JOB")

  vwJobStateTrans.DeleteAll

  vwJob.Sort("EFFDT")
  vwJob.MoveFirst

  dim PrevAction, PrevReason, Action, Reason
  PrevAction = Space(0)
  PrevReason = Space(0)


  while not vwJob.EOF
    Action = vwJob.Field("ACTION")
    Reason = vwJob.Field("ACTION_REASON")
    if len(PrevAction) > 0 then
      vwJobStateTrans.New
      vwJobStateTrans.Field("ACT1ACT2") = PrevAction & "-" & Action
      vwJobStateTrans.Field("REA1REA2") = PrevReason & "-" & Reason
      vwJobStateTrans.Field("ACTREA1ACTREA2") = PrevAction & "-" & Action & ":" & PrevReason & "-" & Reason
      vwJobStateTrans.Update
    end if
    vwJob.MoveNext
    PrevAction = Action
    PrevReason = Reason
  wend


end sub
```

Subject level example – state transition analysis

- Profiling this table of metadata yields the time ordered pairs present in the data. Often these pairs are different than the 'allowed' values given to the DQ analyst by the IT guys, working through any discrepancies between the actual vs. the expected values is a useful exercise that can yield a few business rules.

- Again, we look for very frequent and very infrequent values, here we have no standouts.

ACT1ACT2 Table: TABLE7300 Schema: TDWI_OWNER Data Source: tdw

Column Profiling | **Frequency Distribution** | Pattern Frequency Distribution | Percentiles | D

Value	Count	Percentage
CNV-CNV	372	0.15
CNV-DTA	934	0.39
CNV-PAY	4111	1.71
CNV-PNP	707	0.29
CNV-PWP	375	0.16
CNV-RGN	894	0.37
CNV-TER	483	0.20
CNV-XFR	849	0.35
DTA-CNV	286	0.12
DTA-DTA	17919	7.44
DTA-LOA	672	0.28
DTA-PAY	20390	8.47
DTA-PLA	907	0.38
DTA-PNP	2455	1.02
DTA-PWP	1279	0.53
DTA-RFL	280	0.12
DTA-RGN	7001	2.91
DTA-RIF	330	0.14
DTA-TER	2065	0.86
DTA-TWP	473	0.20
DTA-XFR	2541	1.06
HIR-CNV	3017	1.25
HIR-DTA	2463	1.02
HIR-PAY	1230	0.51
HIR-PNP	319	0.13
HIR-RGN	968	0.40
HIR-TER	2018	0.84
HIR-XFR	297	0.12
LOA-DTA	341	0.14
LOA-LOA	2195	0.91
LOA-PAY	272	0.11
LOA-PLA	1198	0.50
LOA-RFL	2409	1.00