



**THE  
POWER  
TO KNOW.®**

## Denver SAS User Group

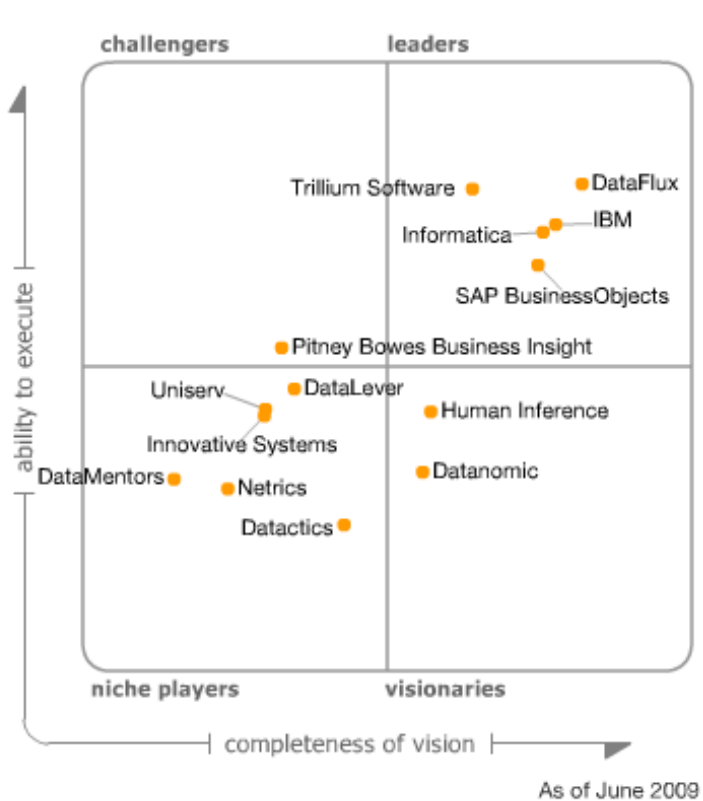
# SAS® Enterprise Data Integration and Data Quality

---

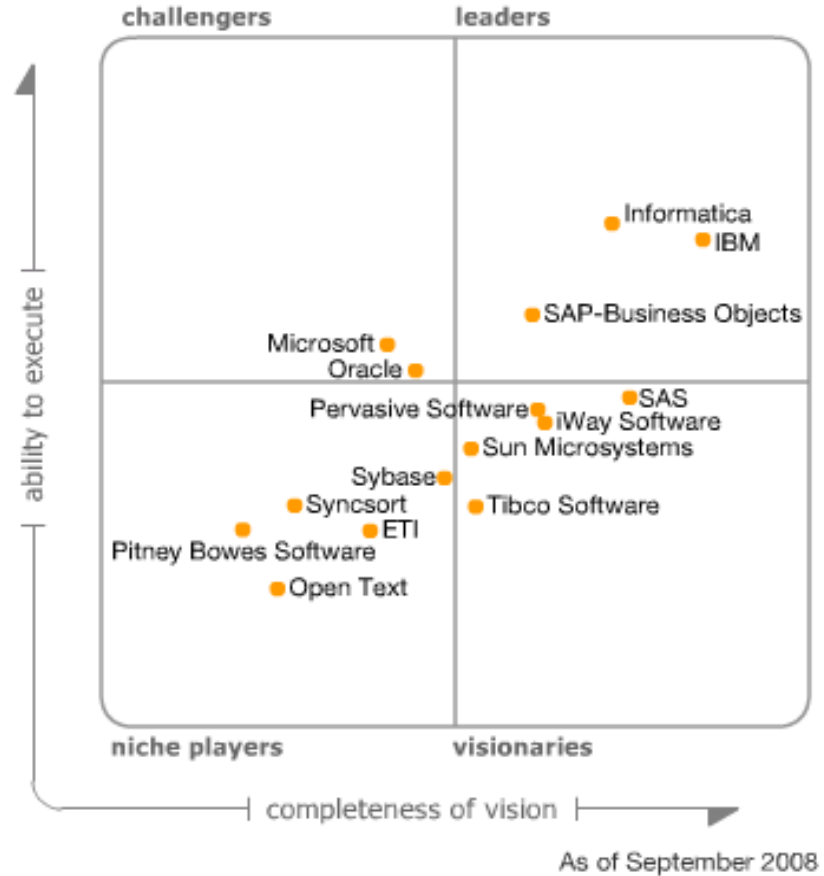
John Motler – Sales Engineer

January 13, 2010

# Gartner Market Validation



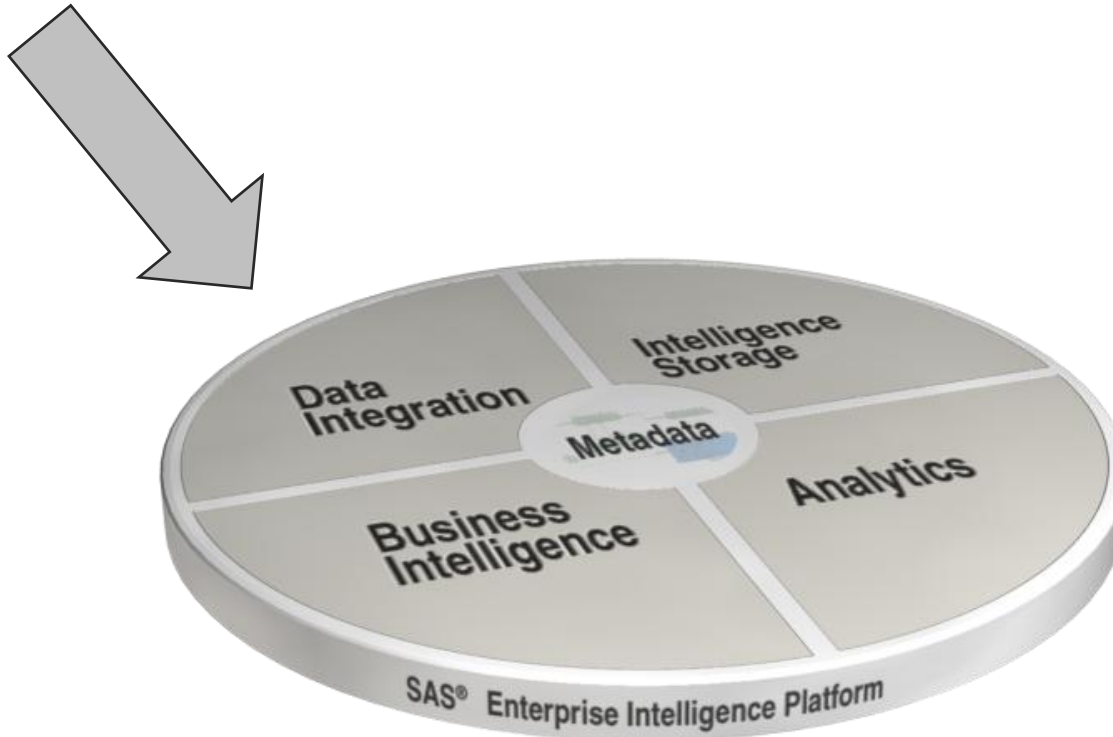
**Data Quality Tools**  
June 2009



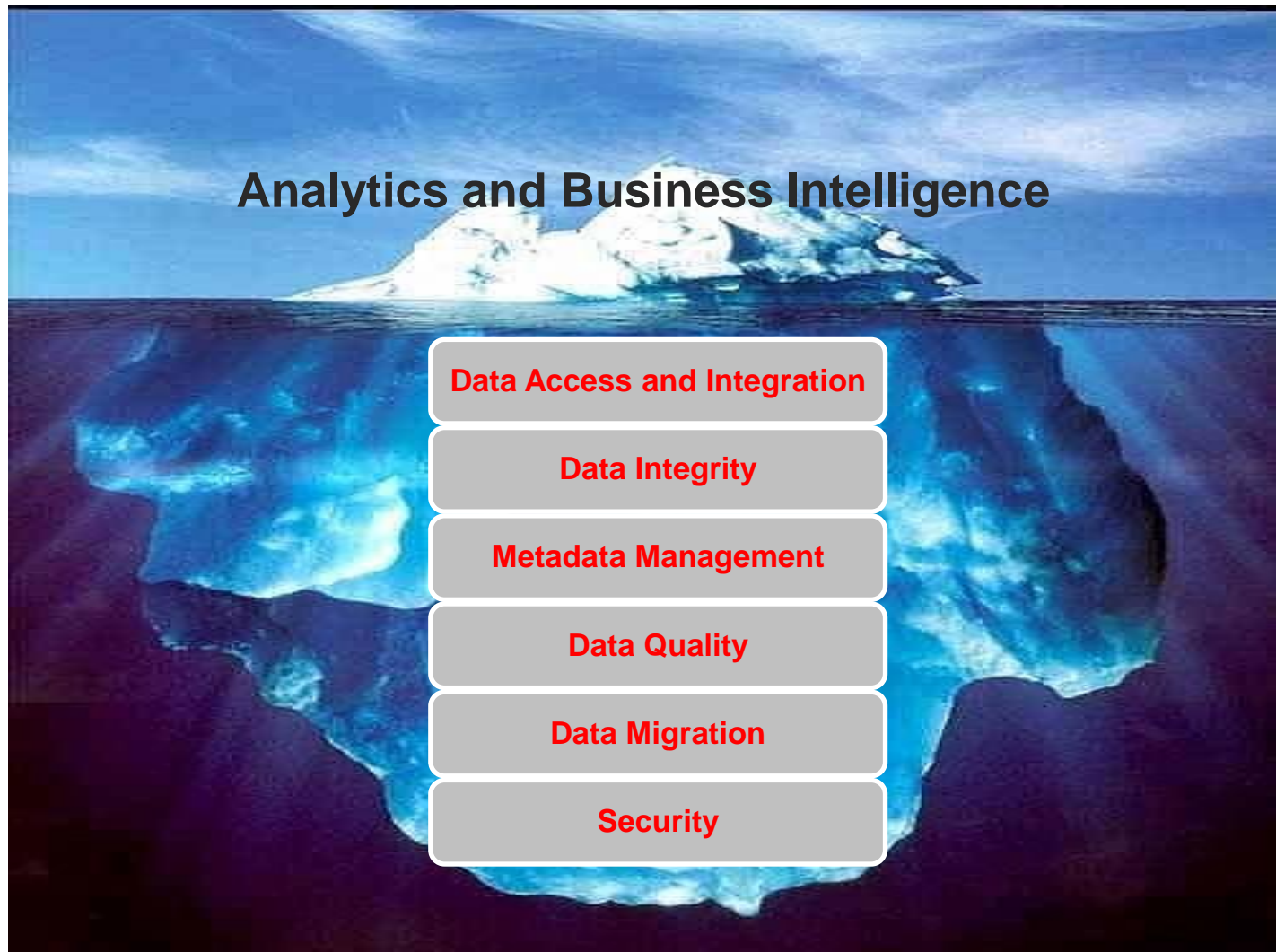
**Data Integration Tools**  
September 2008

# SAS<sup>®</sup> Enterprise Intelligence Platform

## *Data Integration*



# The Hidden Data Challenge



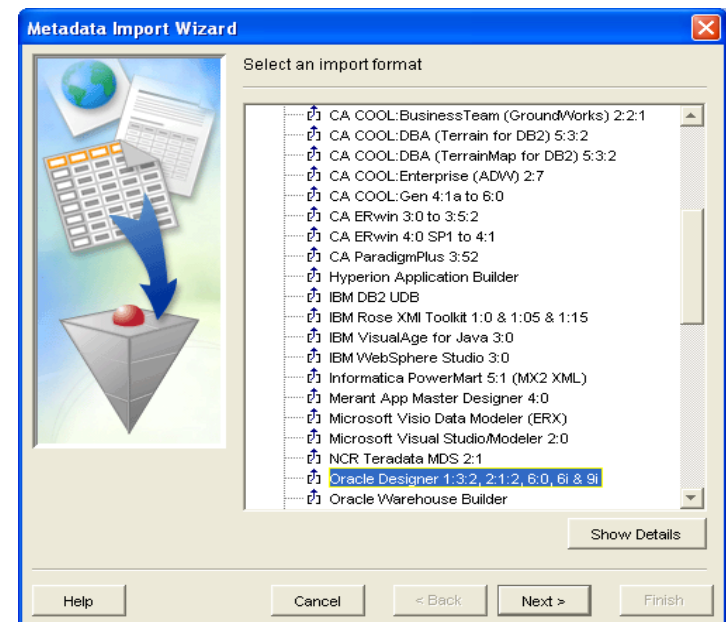
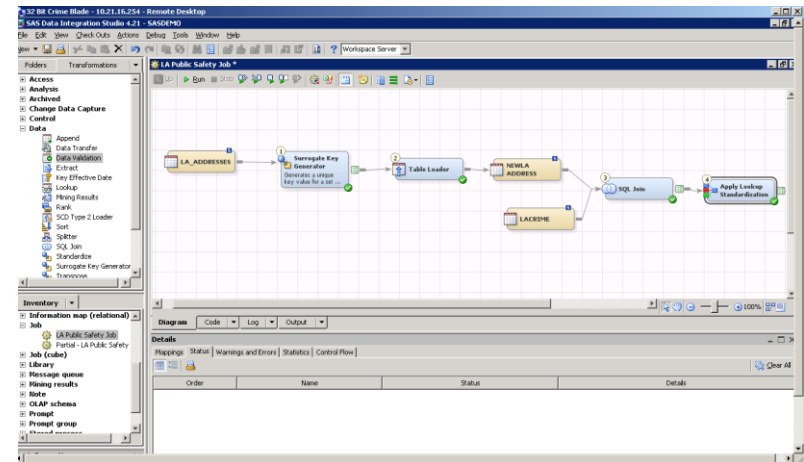
# Enterprise Connectivity

## SAS® Data Integration

- Enterprise Applications: Oracle Applications, PeopleSoft, SAP BW, SAP R/3, Siebel and more.
- Database sources: AS400, ODBC, IBM DB2/UDB, Informix, Microsoft Access, Microsoft Excel, Microsoft SQL Server, MySQL, Netezza, DATAlegro, HP NeoView, OLE/DB, Oracle, Sybase, Sybase ASIQ, SAS data sets, SAS Scalable Performance Data Server, SAS Scalable Performance Data Engine, Teradata, and more.
- Mainframe data sources (OS/390 and z/OS): ADABAS, CA-Datcom, CA-IDMS, COBOL, IBM DB2, IMS-DL/I, ISAM files, Oracle, SYSTEM 2000, Teradata, VSAM (KSDS and ESDS), and other file formats.
- File formats: CSV, XLS, Access, WKS, text/flat files, XML, COBOL Copybooks, and FTP and URL-based sources. Reads and writes data representations such as ASCII, Binary, EBCDIC, Hexadecimal and Octal.
- Support for Message-Oriented Middleware, including WebSphere MQ from IBM, MSMQ from Microsoft and Tibco's Rendezvous.
- Support for unstructured and semi-structured data to parse and process files.
- Access to static and streaming data for sending and receiving via Web services.

# SAS Data Integration

- SAS Data Integration Studio is a powerful visual design tool for the construction, execution, and maintenance of data integration
- SAS Data Integration Studio provides key integration functions including:
  - Connectivity
  - Data Cleansing and Enrichment
  - Extraction, Transform and Load (ETL)
  - Data Federation
  - Migration and Synchronization
- SAS Enterprise Data Integration Server is a flexible, reliable and complete data integration solution, designed to meet the comprehensive data integration needs of the enterprise.

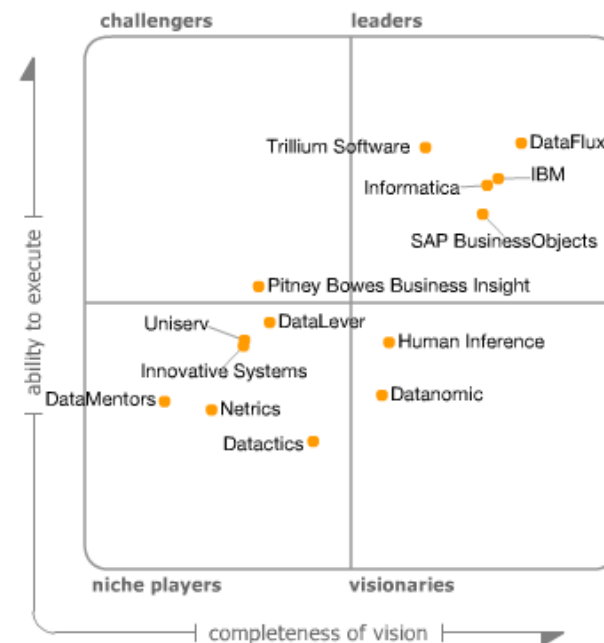
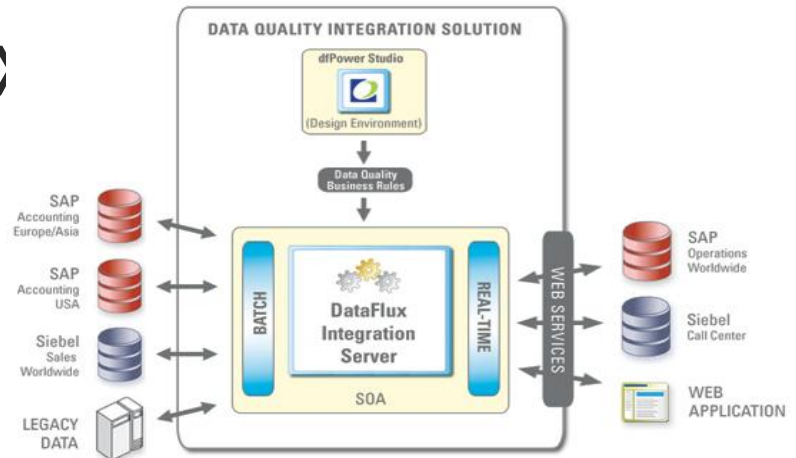


# Why Data Quality Issues Happen...

- Data entry Issues – Typos, character transpositions, inconsistent use of abbreviations, etc.
- Data Processing Issues – system upgrades, system re-designs, inconsistent use of fields, etc.
- Data Migration Issues – Data migration from legacy systems, data migration due to mergers and acquisitions, etc.
- Data Decay Over Time – Some data that was initially appropriate for a specific use is no longer appropriate for that use.

# Data Quality: SAS DataFlux

- Poor data quality is one of the risks that must be addressed with any data warehouse implementation.
- DataFlux is comprised of a suite of applications that solve complex data management issues.
- Uses a Quality Knowledge Base based on locales that has thousands of pre-defined schemes, grammar and vocabularies, these can be supplemented or new ones created.
- Jobs and reports are stored in a central repository that can be shared across users.



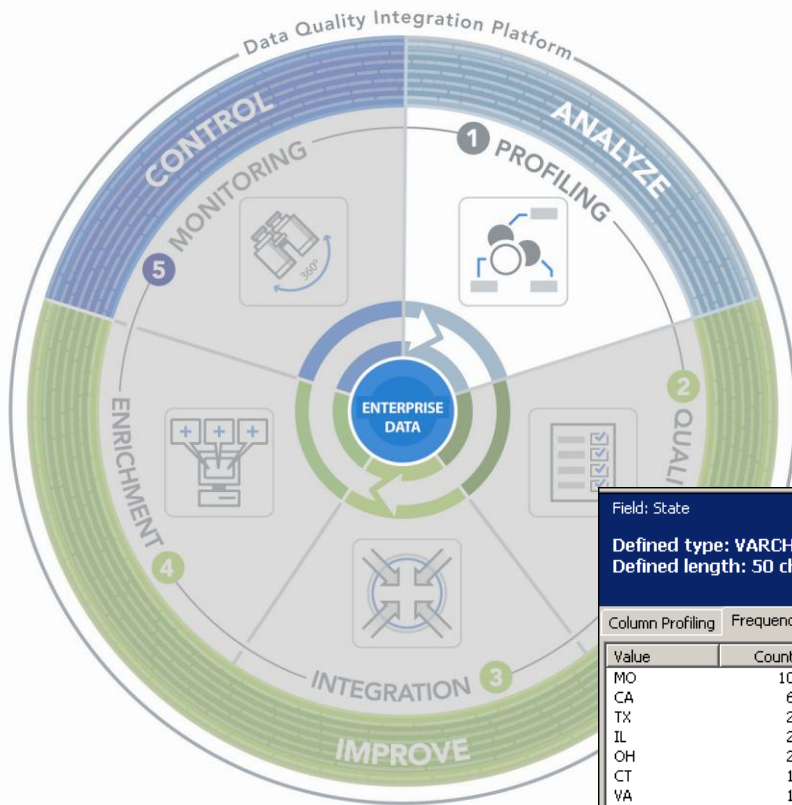
As of June 2009

# DataFlux Platform



# DataFlux Methodology: Analyze

## Metadata Profiling and Data Profiling



dfPower Explorer - DataFlux Sample

File View Reports Export Help

Project: metadata

Table: Contacts

Database: DataFlux Sample  
Matching tables: 10  
% Match: 83

Database	Table	Schema	Matching Columns	Total Columns	% Match
DF_ORA_TEST	CLIENT_INFO	SCOTT	6	9	66
DF_ORA_TEST	CONTACTS	SCOTT	12	13	92
DF_ORA_TEST	EMPLOYEE_DATA	SCOTT	2	9	22
DF_SQL_TEST	CONTACTS	dbo	12	13	92
DataFlux Sample	Client_Merge_Data		5	12	41
DataFlux Sample	CompanyNumeric		1	12	8
DataFlux Sample	Client_Info		5	12	41
DataFlux Sample	Product		1	6	16
DataFlux Sample	Purchase		7	14	50
DataFlux Sample	Sales		8	13	61

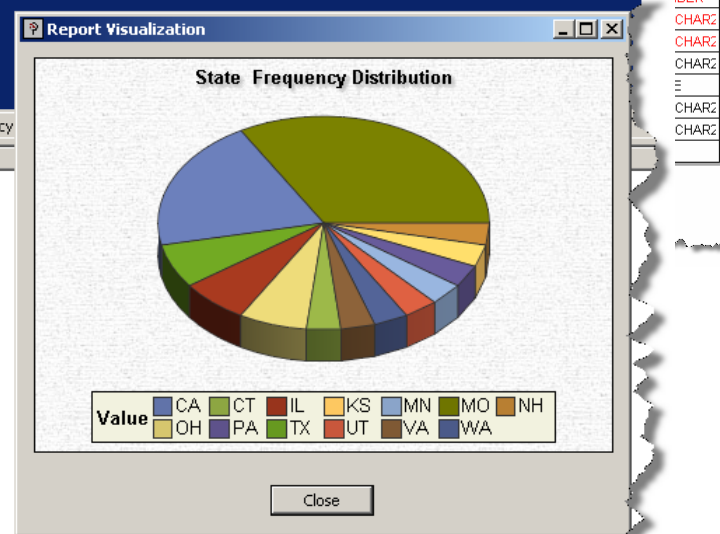
Database 1: DataFlux Sample  
Database 2: DF\_ORA\_TEST

Contacts		CLIENT_INFO	
PHONE	VARCHAR	PHONE	VARCHAR2
CITY	VARCHAR	CITY	VARCHAR2

Field: State

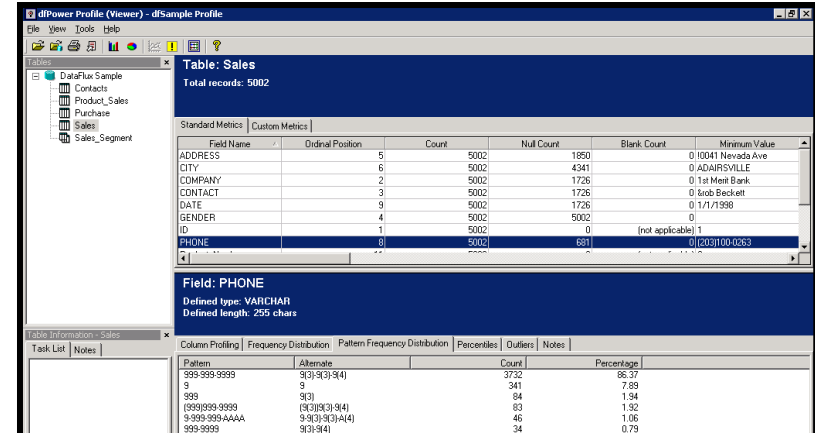
Defined type: VARCHAR  
Defined length: 50 chars

Column Profiling	Frequency Distribution	Pattern Frequency
Value	Count	Percentage
MO	10	33.33
CA	6	20.00
TX	2	6.67
IL	2	6.67
OH	2	6.67
CT	1	3.33
VA	1	3.33
WA	1	3.33
UT	1	3.33
MN	1	3.33
PA	1	3.33
KS	1	3.33
NH	1	3.33



# What is Data Profiling?

- Pattern recognition
- Data scarcity
- Frequency reports
- Calculating basic statistics
- Identifying outliers
- Metadata validation
- Data relationship validation
- Visualization

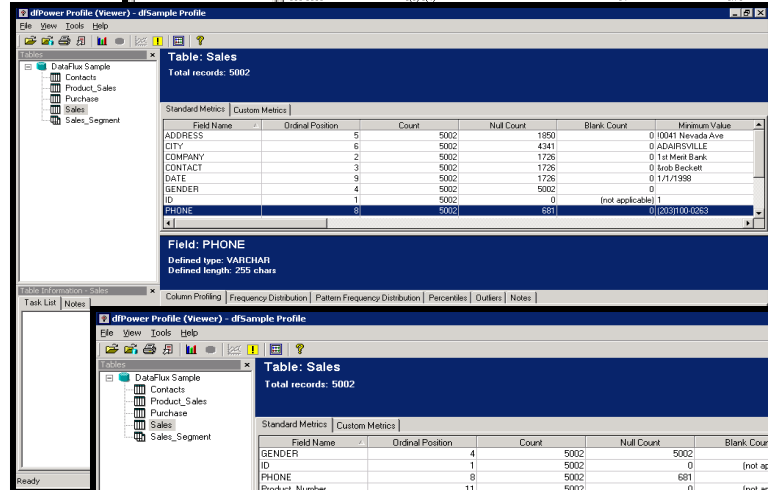


**Table: Sales**  
Total records: 5002

Field Name	Ordinal Position	Count	Null Count	Blank Count	Minimum Value
ADDRESS	5	5002	1950	0	0 10041 Nevada Ave
CITY	6	5002	4341	0	0 ADAIRSVILLE
COMPANY	2	5002	1726	0	0 1st Merit Bank
CONTACT	3	5002	1726	0	0 Irob Beckett
DATE	9	5002	1726	0	0 1/1/1998
GENDER	4	5002	5002	0	0
ID	1	5002	0	0	(not applicable) 1
PHONE	8	5002	681	0	0 (203)100-0263

**Field: PHONE**  
Defined type: VARCHAR  
Defined length: 255 chars

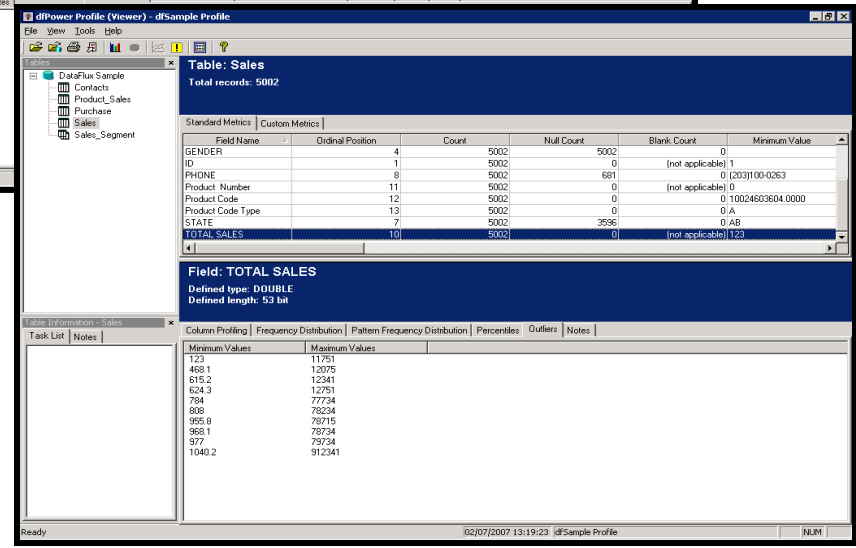
Column Profiling	Frequency Distribution	Pattern Frequency Distribution	Percentiles	Outliers	Notes
Pattern	Alternate	Count	Percentage		
999-999-9999	9(3)9(3)9(4)	9732	86.37		
9	9	341	7.89		
999	9(3)	84	1.94		
(999)999-9999	(9(3)9(3)9(4)	63	1.52		
9-999-999-AAAA	9-9(3)9(3)A(4)	46	1.06		
999-9999	9(3)9(4)	34	0.79		



**Table: Sales**  
Total records: 5002

Field Name	Ordinal Position	Count	Null Count	Blank Count	Minimum Value
ADDRESS	5	5002	1950	0	0 10041 Nevada Ave
CITY	6	5002	4341	0	0 ADAIRSVILLE
COMPANY	2	5002	1726	0	0 1st Merit Bank
CONTACT	3	5002	1726	0	0 Irob Beckett
DATE	9	5002	1726	0	0 1/1/1998
GENDER	4	5002	5002	0	0
ID	1	5002	0	0	(not applicable) 1
PHONE	8	5002	681	0	0 (203)100-0263

**Field: PHONE**  
Defined type: VARCHAR  
Defined length: 255 chars



**Table: Sales**  
Total records: 5002

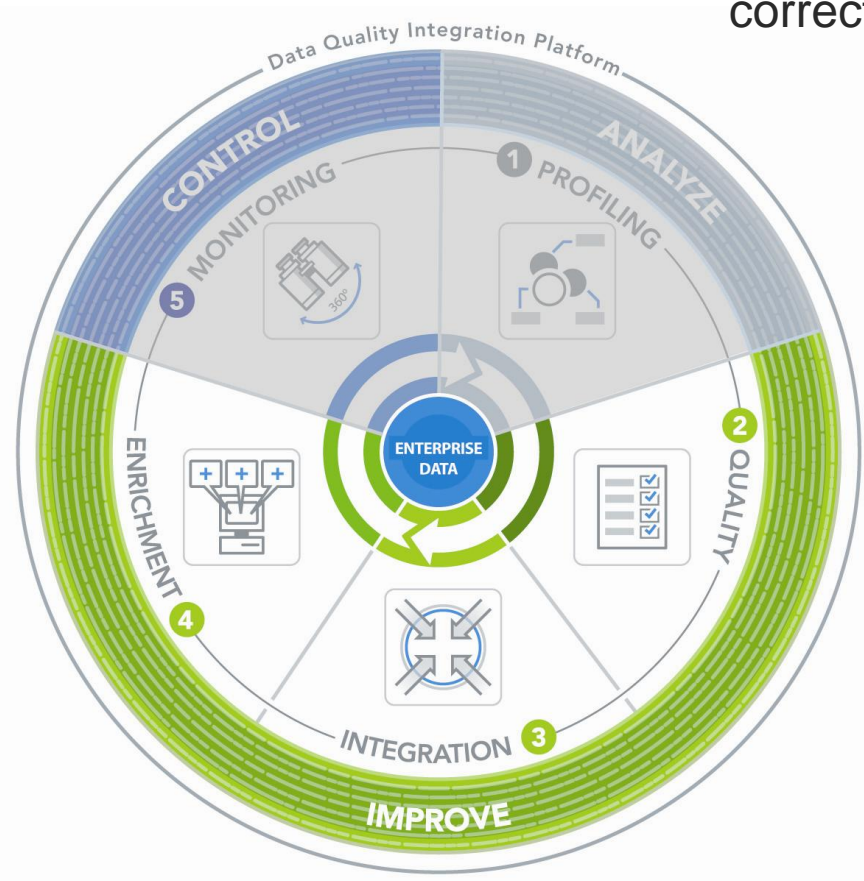
Field Name	Ordinal Position	Count	Null Count	Blank Count	Minimum Value
GENDER	4	5002	5002	0	0
PHONE	1	5002	681	0	(not applicable) 1
ID	8	5002	0	0	0 (203)100-0263
Product Number	11	5002	0	0	(not applicable) 0
Product Code	12	5002	0	0	0 10024603604.0000
Product Code Type	13	5002	0	0	0 A
STATE	7	5002	3536	0	0 AB
TOTAL SALES	10	5002	0	0	(not applicable) 123

**Field: TOTAL SALES**  
Defined type: DOUBLE  
Defined length: 53 bits

Column Profiling	Frequency Distribution	Pattern Frequency Distribution	Percentiles	Outliers	Notes
Minimum Values	Maximum Values				
123	11251				
468 1	12075				
815 2	12341				
624 3	12751				
784	77734				
988	78254				
955 8	78715				
968 1	78734				
977	79734				
1040 2	912341				

# DataFlux Methodology: Improve

The improvement phase includes processes for correcting, consolidating and enriching data.



Three components:

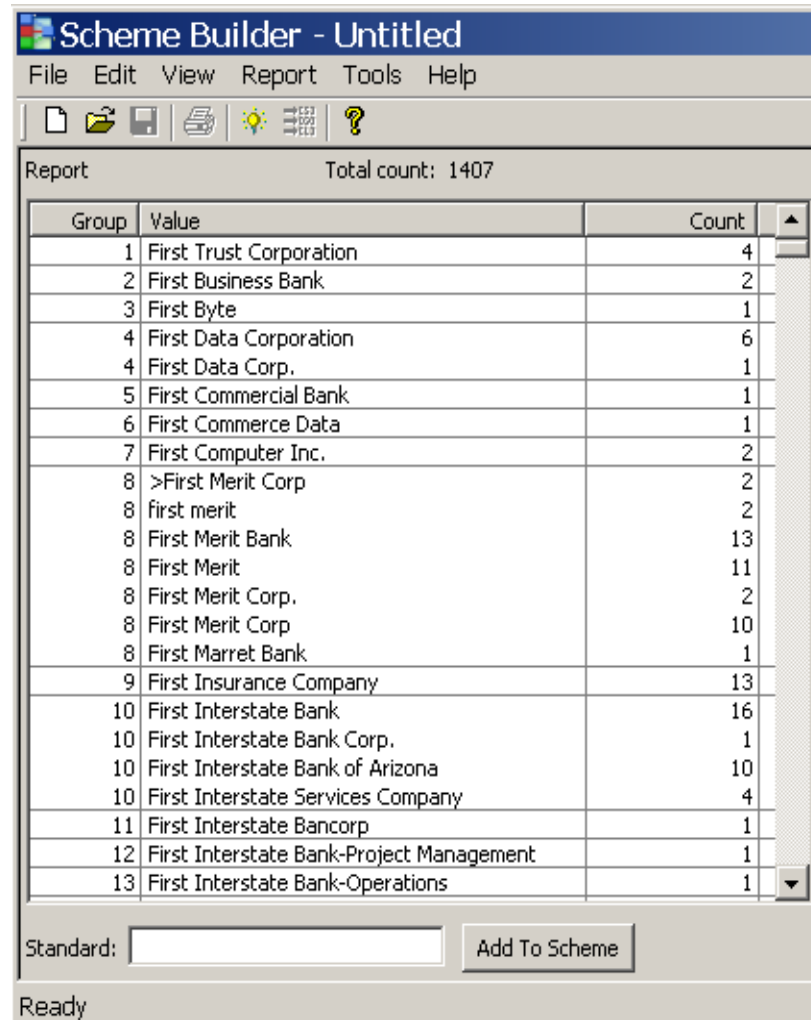
1. Quality
2. Integration
3. Enrichment

# Quality: Data Standardization

- Ensures uniform representation of a data value
- Used to correct spellings, inconsistent use of nicknames and abbreviations
- Manage child/parent relationships in data
- Establishing consistency in operational data through time
- Allows merging of multiple operational systems

# Scheme Definitions and Smart Clustering

- Ability to create match schemes from any data source
- Make inconsistent data consistent
- Leverage for mapping to new codes or reference data in data migration projects



File Edit View Report Tools Help

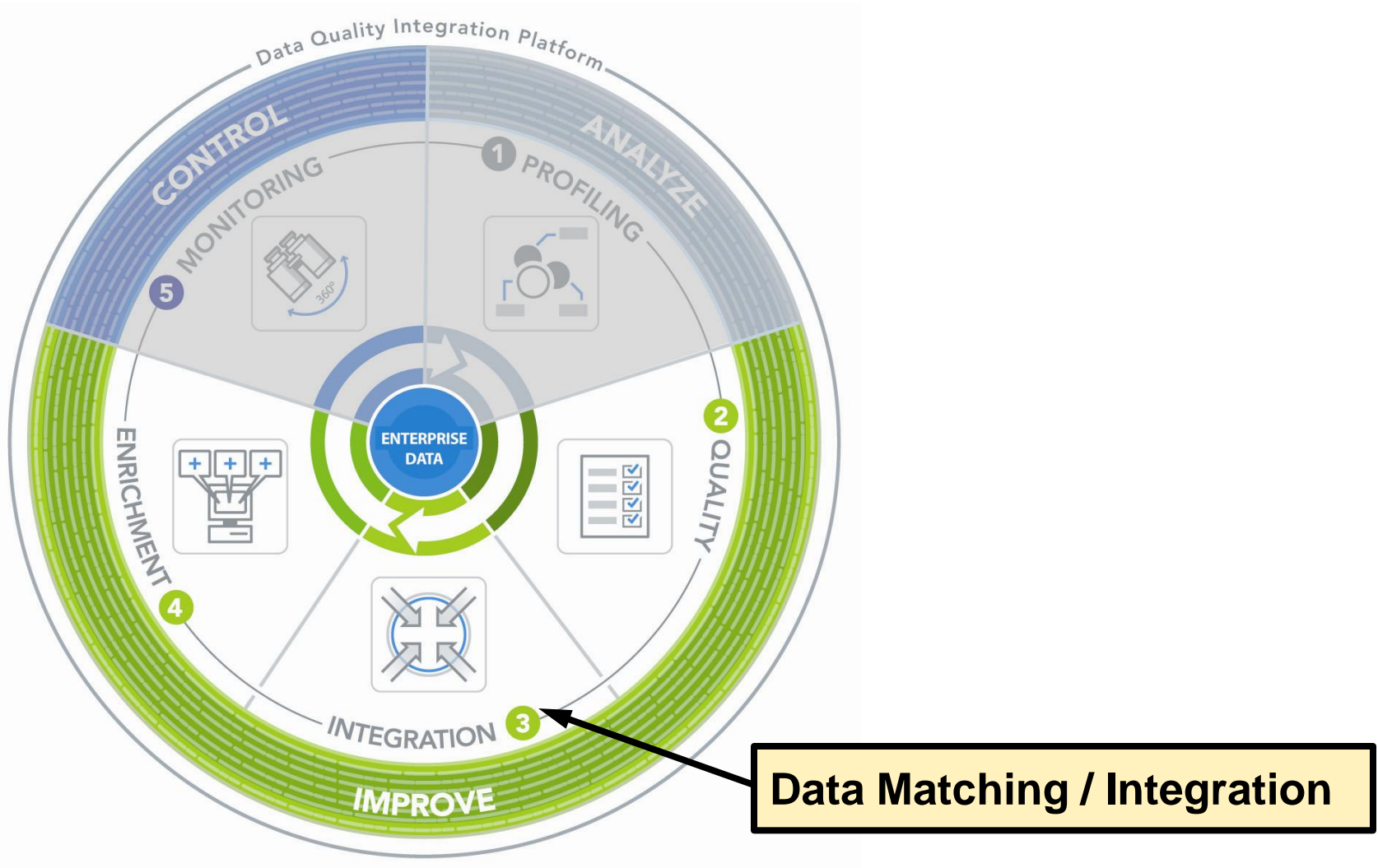
Report Total count: 1407

Group	Value	Count
1	First Trust Corporation	4
2	First Business Bank	2
3	First Byte	1
4	First Data Corporation	6
4	First Data Corp.	1
5	First Commercial Bank	1
6	First Commerce Data	1
7	First Computer Inc.	2
8	>First Merit Corp	2
8	first merit	2
8	First Merit Bank	13
8	First Merit	11
8	First Merit Corp.	2
8	First Merit Corp	10
8	First Marret Bank	1
9	First Insurance Company	13
10	First Interstate Bank	16
10	First Interstate Bank Corp.	1
10	First Interstate Bank of Arizona	10
10	First Interstate Services Company	4
11	First Interstate Bancorp	1
12	First Interstate Bank-Project Management	1
13	First Interstate Bank-Operations	1

Standard:

Ready

# DataFlux Methodology: Improve



# Data Matching

Field	Record 1	Record 2	Record 3
Name	Robert Smith	Bob Smith	Rob Smith
Address	100 Main St	100 Main	100 Main St.
City	Phoenix	Phoenix	Raleigh
Match Code	<b>GHWS\$\$EWT\$</b>	<b>GHWS\$\$EWT\$</b>	GHWS\$\$WWI\$

\*\* The Match Code was generated off of the Name and City fields.

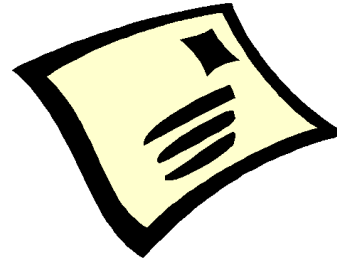
# Business Uses for Data Matching

- Clustering like records together
- De-duplicating data tables
- Fuzzy logic joins
- House holding

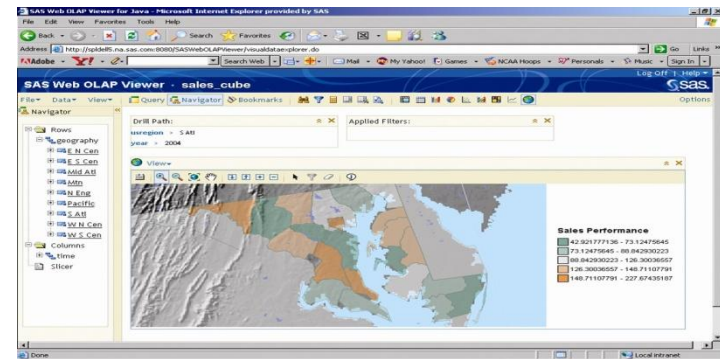
Given Name	Family Name	Address	Phone	MC1	MC2	MC3	HH
Susan	Allen	2208 Vandemere Ct	832-8239	\$AV	#V8	%A8	1
Barbara	Sweet	2208 Vandemere Ct	832-8239	\$SV	#V8	%S8	1
Richard	Sweet	2208 Vandemere Ct	616-1504	\$SV	#V6	%S6	1
Jason	Cheeks	1530 Hidden Cove Dr	688-2856	\$GH	#H6	%G6	2
Becker	Ruth	1530 Hidden Cove Dr	688-2856	\$RH	#H6	%R6	2
Michael	Becker	1530 Hidden Cove Dr	688-2856	\$BH	#H6	%B6	2

# Enrichment

- Address Standardization for mailing lists, census track etc.



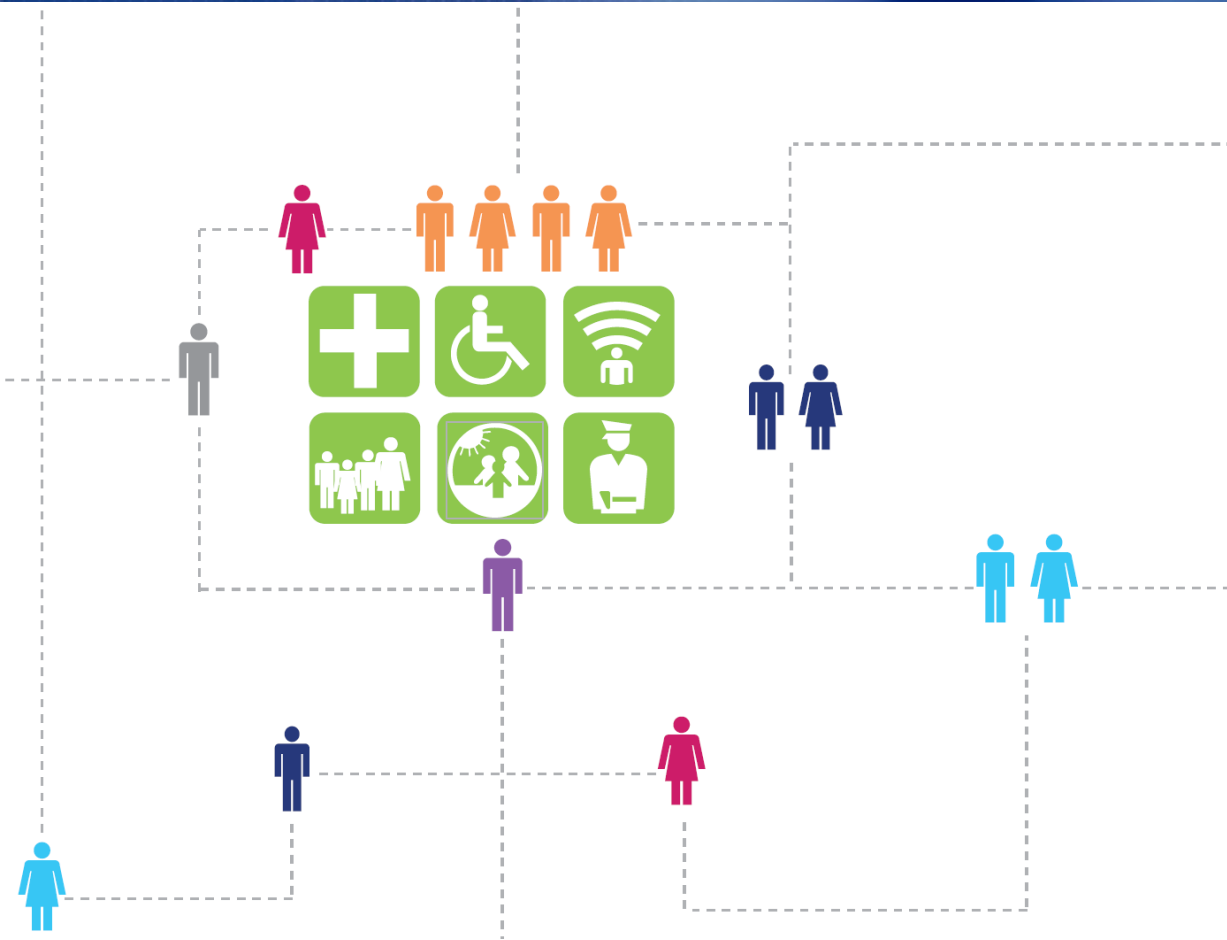
- Geo-Coding addresses for GIS integration



- PhonePlus



# DataFlux Demonstration



# LA County Adult Linkages Project

*Manuel Moreno*

# Objective

The overall objective of the Adult Linkages Project [ALP] is to provide policymakers with empirical information that can support the **enhancement of existing programs for indigent adults and advance social policy making** in Los Angeles County.

# ALP – How it works

- ALP integrates administrative records on indigent adults receiving social and human services from multiple agencies in Los Angeles County.
- ALP makes use of an analytical data warehouse containing the integrated service data.
- ALP has the capacity to produce impact information on service utilization patterns, service gaps, the extent of service engagement, and costs before, during and after participation in Los Angeles County's General Relief (GR) Program for indigent adults.

**ALP Agency Data Sources**



Health Services



Community & Senior Services



Sheriff



Probation



Mental Health



Public Health

**External Data Source**



Employment Development Dept.

**ALP Foundation Data Source**



Public Social Services

**Data Quality Process**



**Data Integration Process**



Agency Service Tables



Linked Services Table



Cohort Table

**Utilize Linked Information**



- Types of services delivered
- Length and timing of services
- Costs of service delivery
- Spatial distribution of service delivery
- Diagnostic codes
- Patterns of multiple service utilizations



**LA County Service Integration Branch Data Warehouse**



Adult Linkages Mart



# Integrating Agency Data Sources

ALP links the administrative files of multiple service agencies: Health Services, Mental Health, Public Health, Children and Family Services, Community and Senior services, Probation, the Sheriff, and Public Social Services.

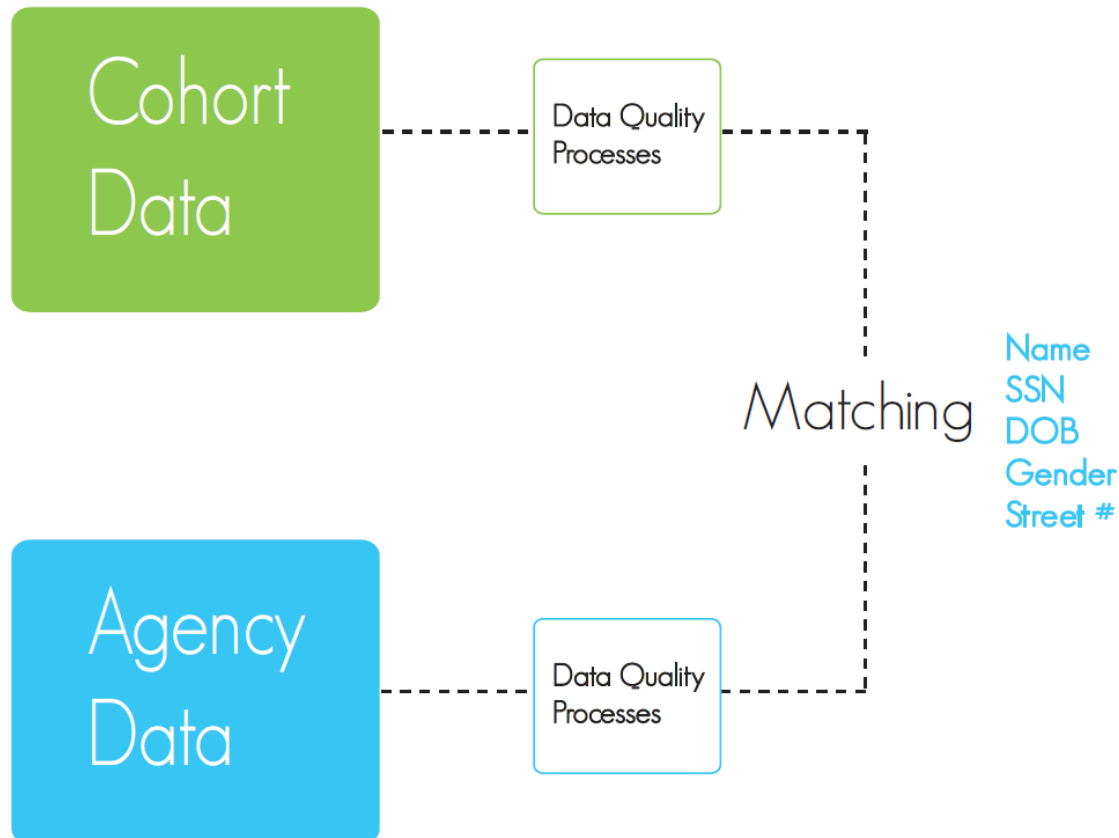


# Methods

- ALP has established a baseline database for a cohort of GR participants who receive County services from multiple service delivery systems.
- ALP uses an anonymous record linkage (Statistical Linkage Key) method to integrate data across departments.
- ALP's record linkage method de-identifies personal information provided in administrative data.

# Data Quality Process

The GR Cohort Data and Agency Data are matched against variables such as Name, Social Security Number, Date of Birth, Gender, and Street #



# Data Integration

- The linked data set contains information such:
  - Random project IDs for each participant. These markers do not identify any client personally
  - Analysis table for each participating Agency and the difference service types offered by each

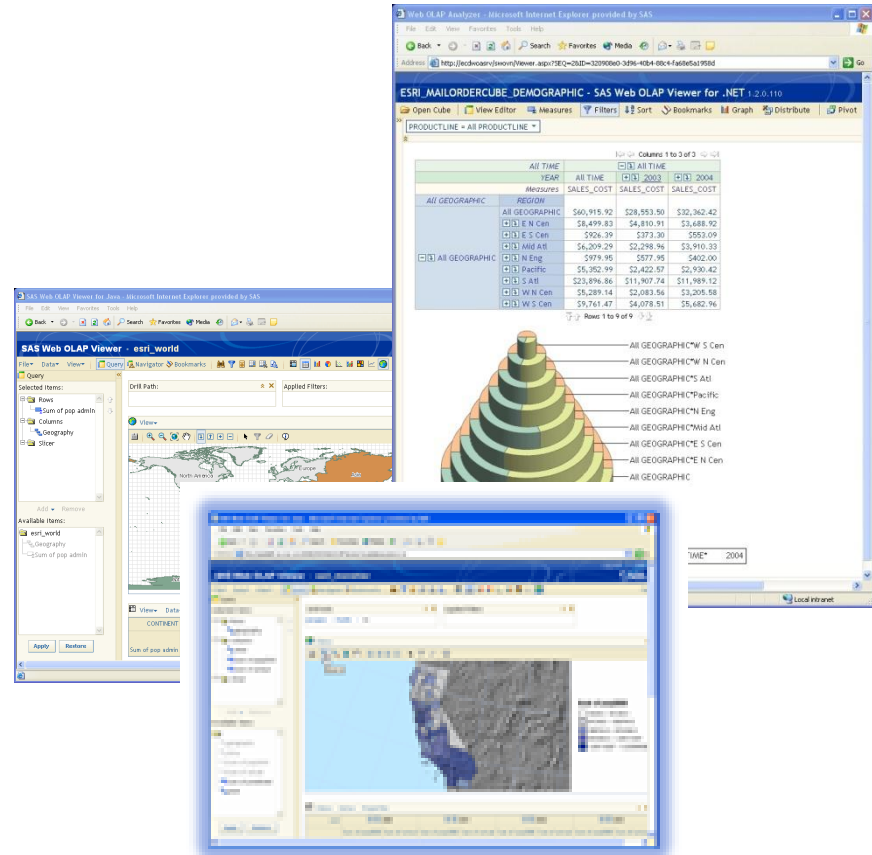


# Utilizing the Information

- The data warehouse can be used to determine patterns and trends in:
  - Types of services delivered
  - Length and timing of services
  - Costs of service delivery
  - Spatial distribution of service delivery
  - Diagnostic Codes
  - Patterns of multiple service utilizations

# Utilizing the Information

- Data visualization with maps, charts, plots
- Extensive suite of graphic data presentation options for business and scientific use
- County departments will have web-based access



# Utilizing the Information

- The ALP supports County agencies in launching pilot projects to better achieve **service coordination**. For example, the ALP methodology will be used to show which categories of GR participants should be targeted for special programs, which agencies should be targeted, and which geographical regions within the County may require the most attention.
- The ALP supports county agencies in **identifying duplicative services** provided by multiple independent service delivery systems, thereby **producing cost avoidance, cost savings, and service enhancements**.



**THE  
POWER  
TO KNOW®**