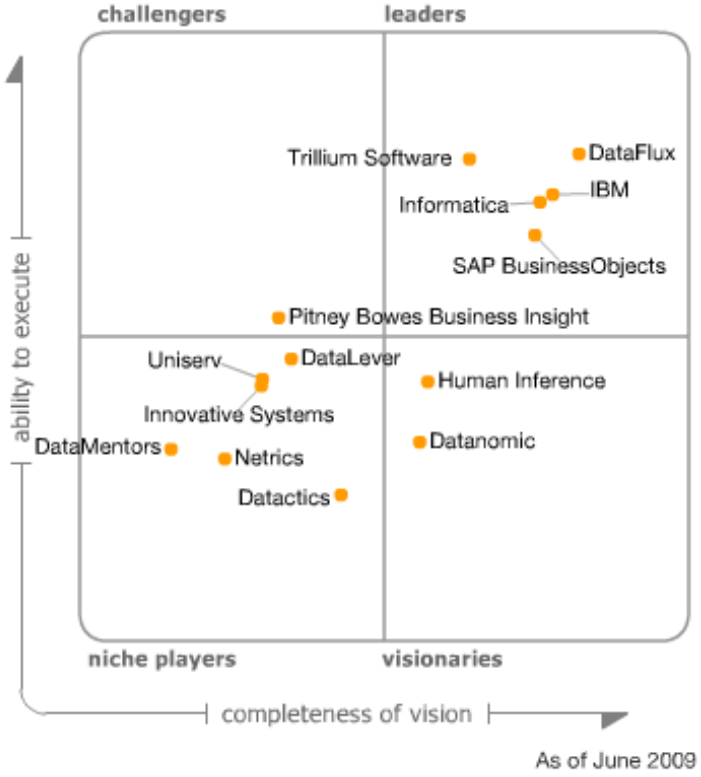
Denver SAS User Group

SAS® Enterprise Data Integration and Data Quality

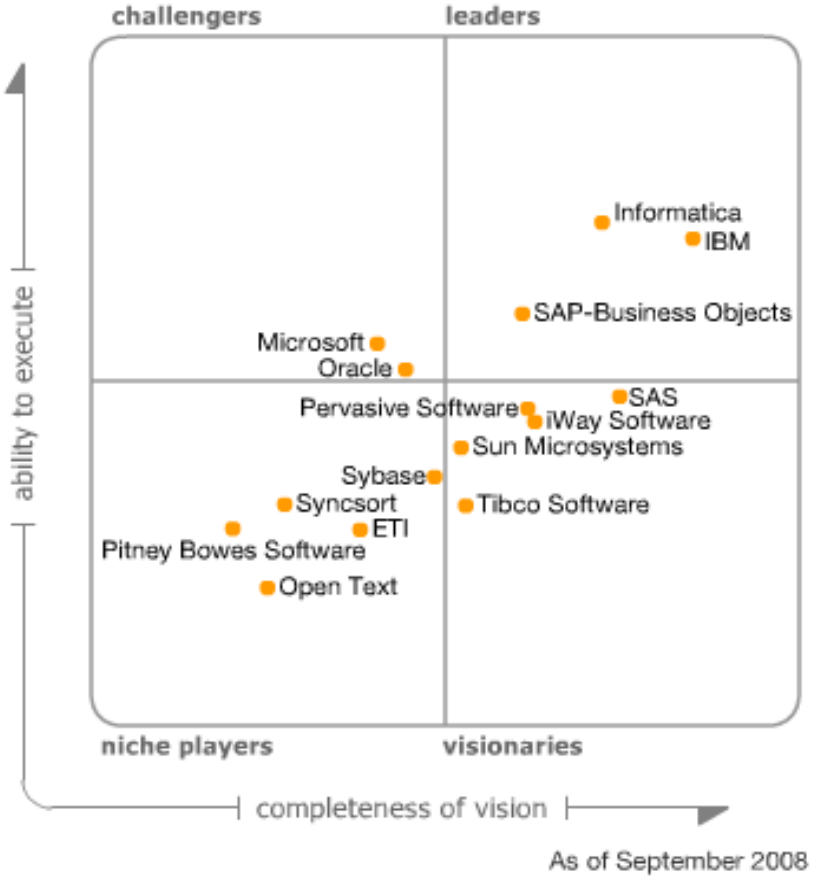
John Motler – Sales Engineer

January 13, 2010

Gartner Market Validation



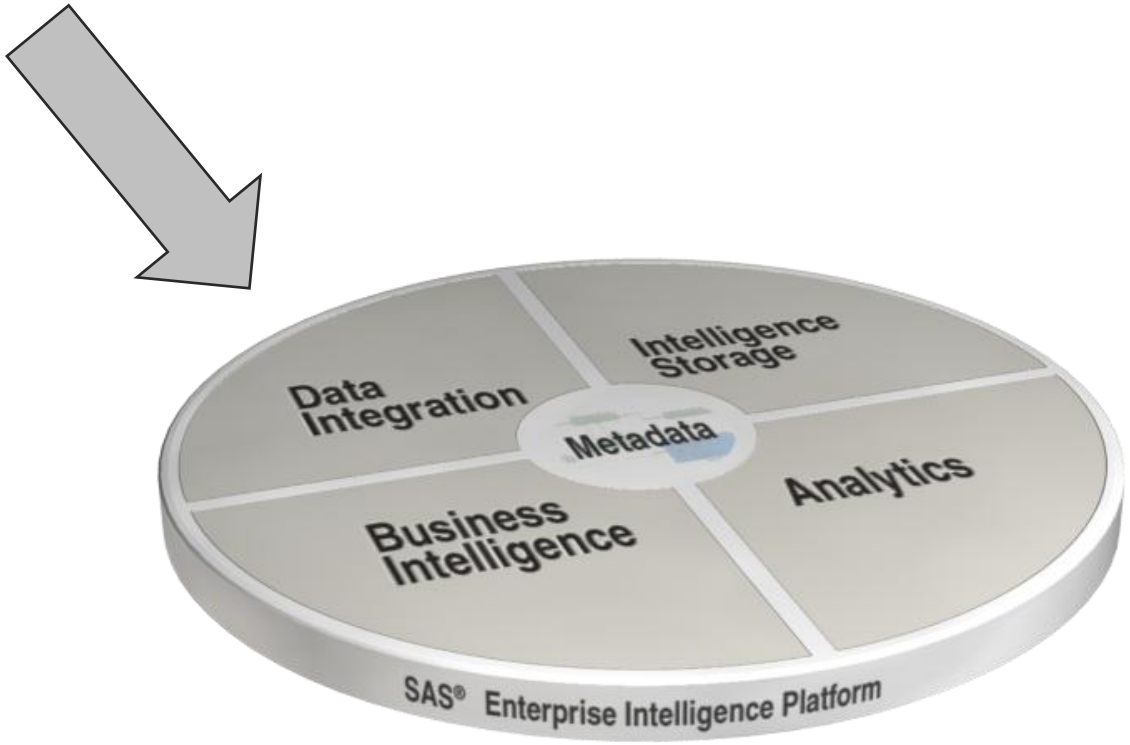
Data Quality Tools
June 2009



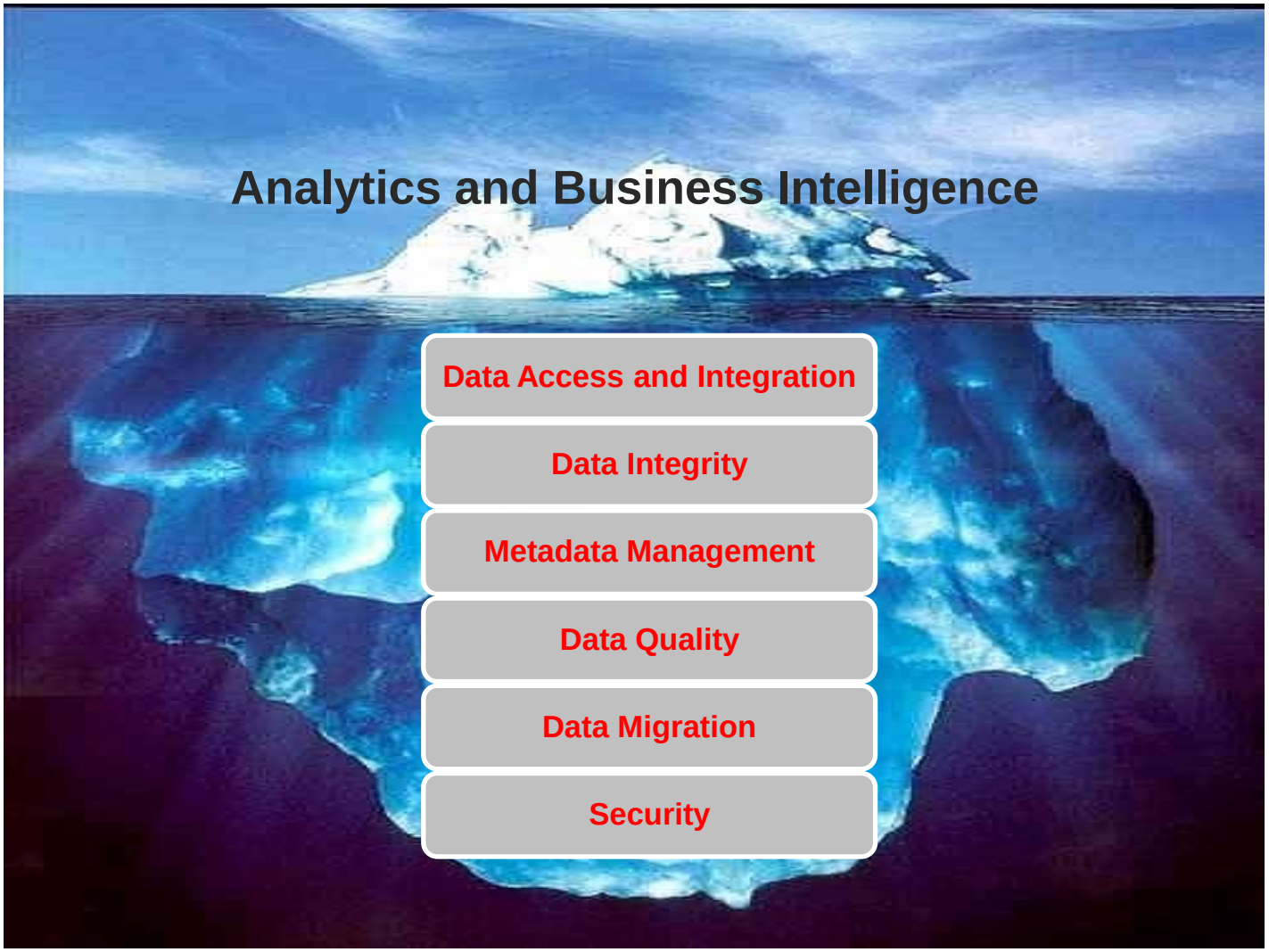
Data Integration Tools
September 2008

SAS® Enterprise Intelligence Platform

Data Integration



The Hidden Data Challenge

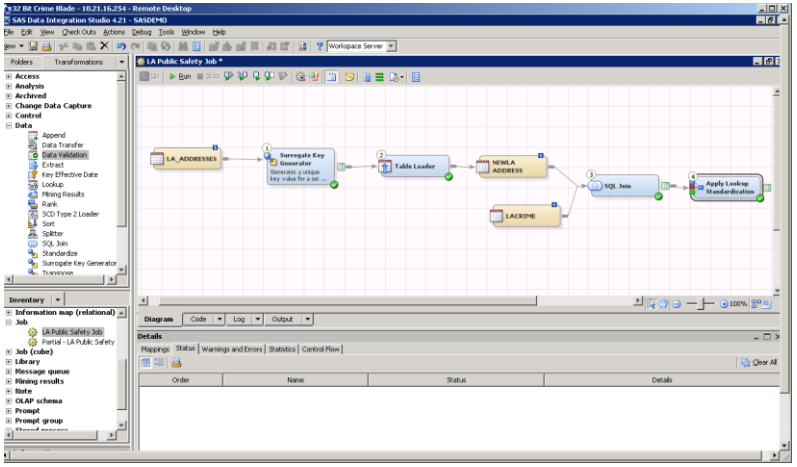


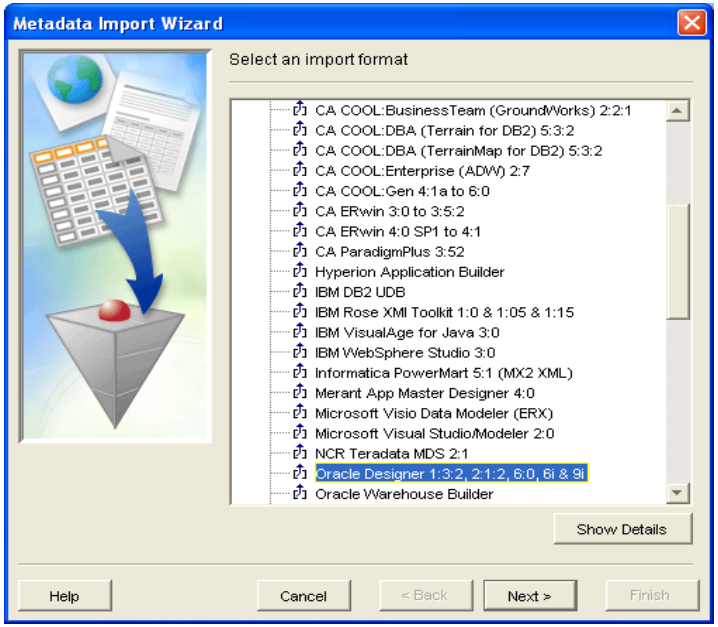
Enterprise Connectivity
SAS® Data Integration

- 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, DATAllegro, 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-Datacom, 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.



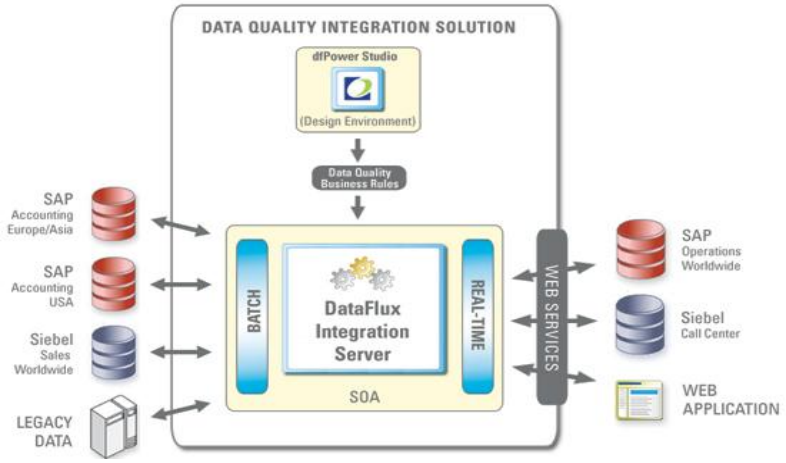


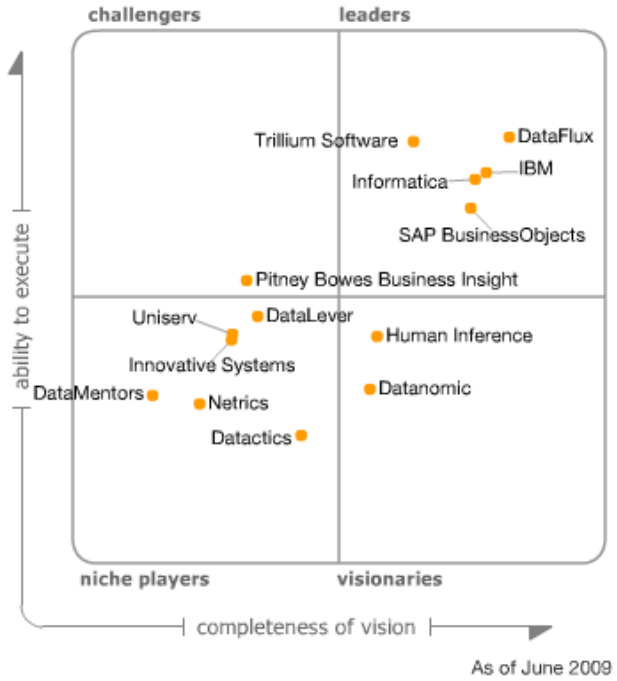
Why Data Quality Issues Happen…

- 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.

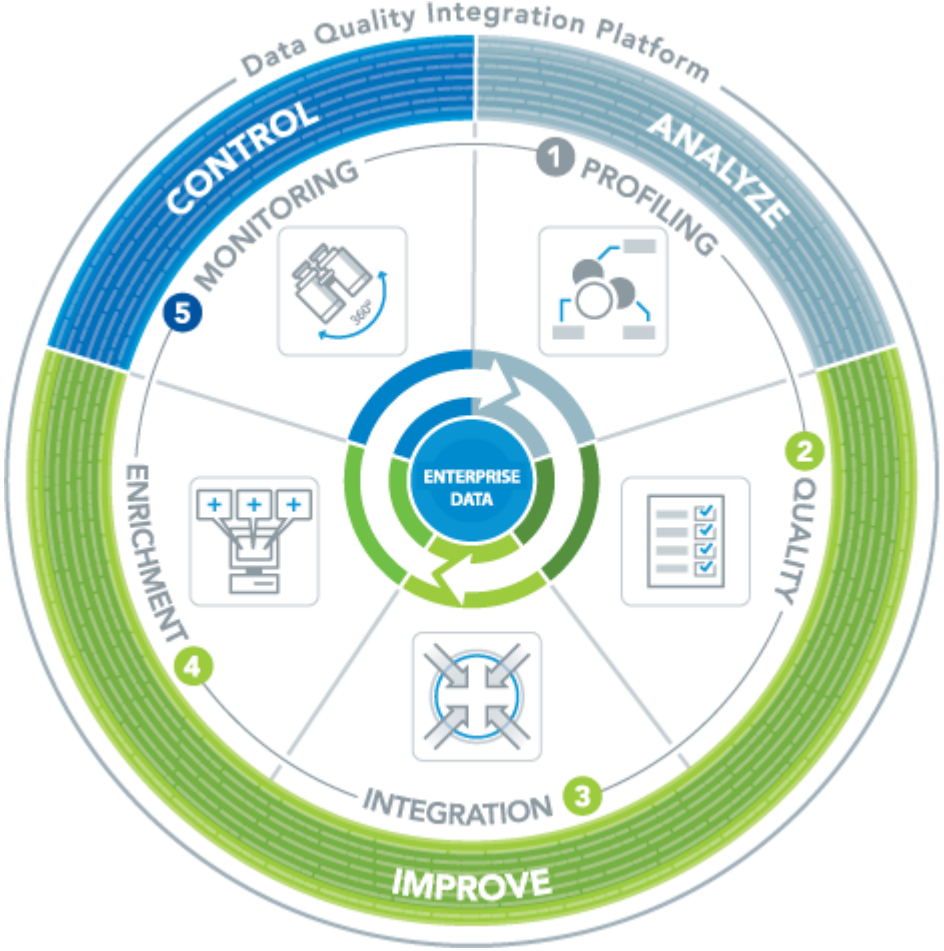
Data Quality: SAS DataFlux

- 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.



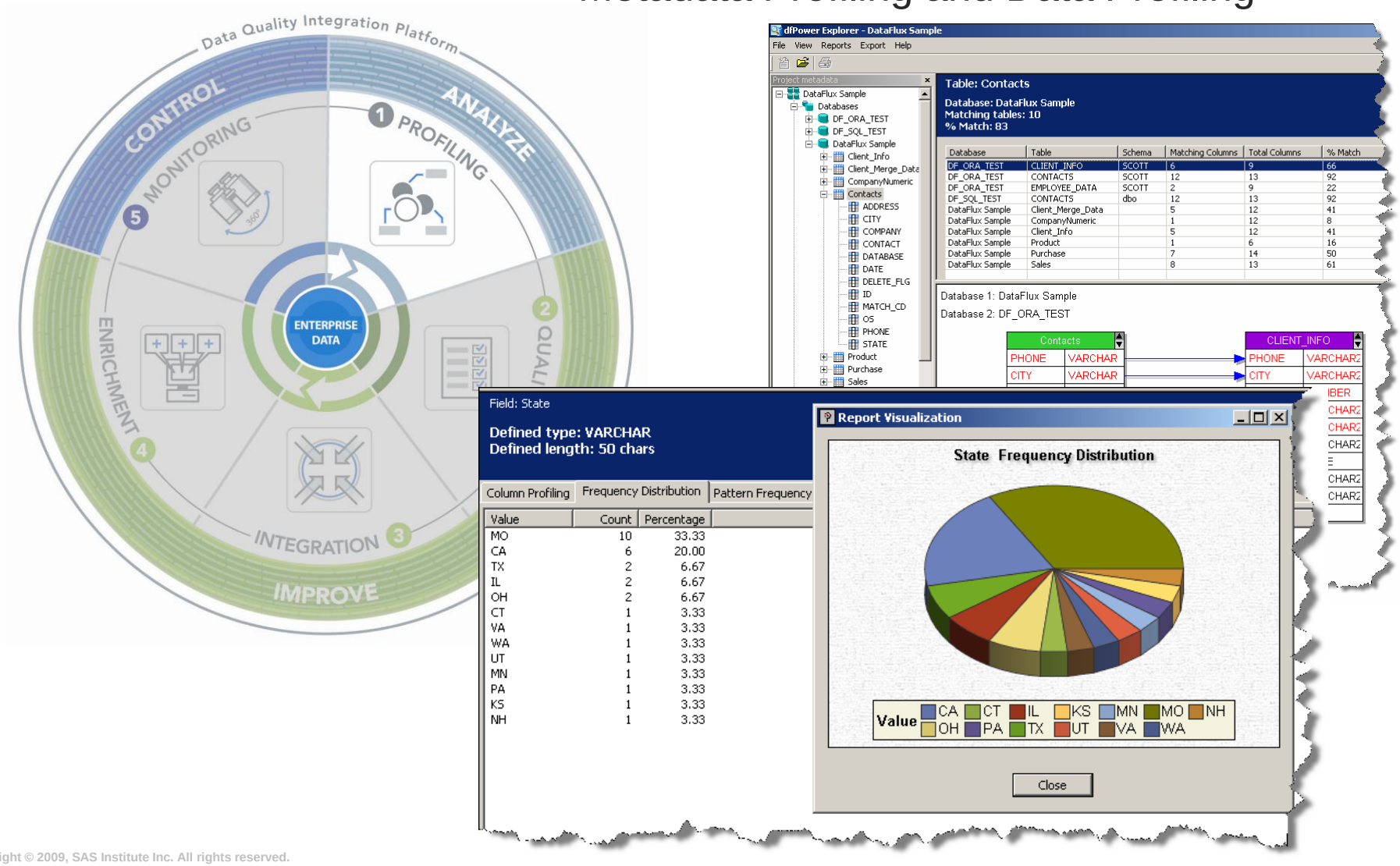


DataFlux Platform



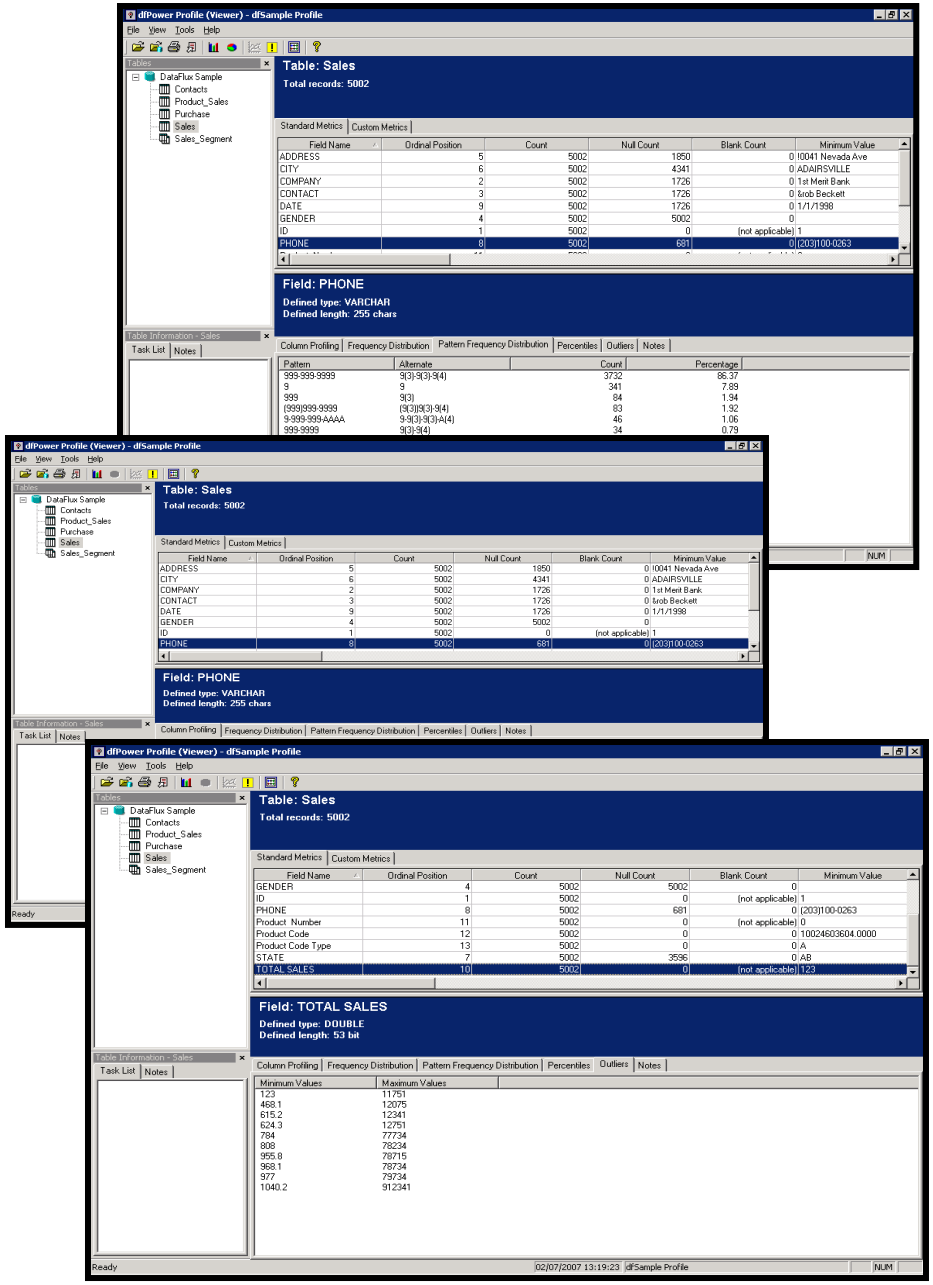
DataFlux Methodology: Analyze

Metadata Profiling and Data Profiling



What is Data Profiling?

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



DataFlux Methodology: Improve

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

Three components:

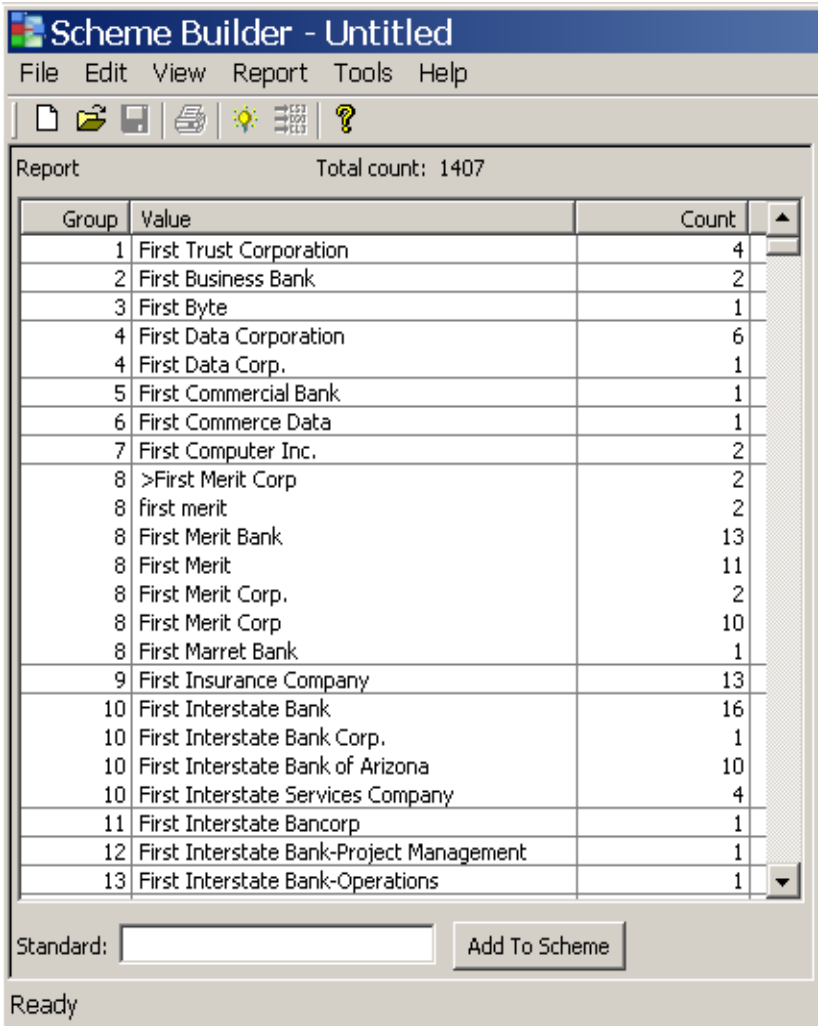
1. Quality
2. Integration
3. Enrichment

Quality: Data Standardization

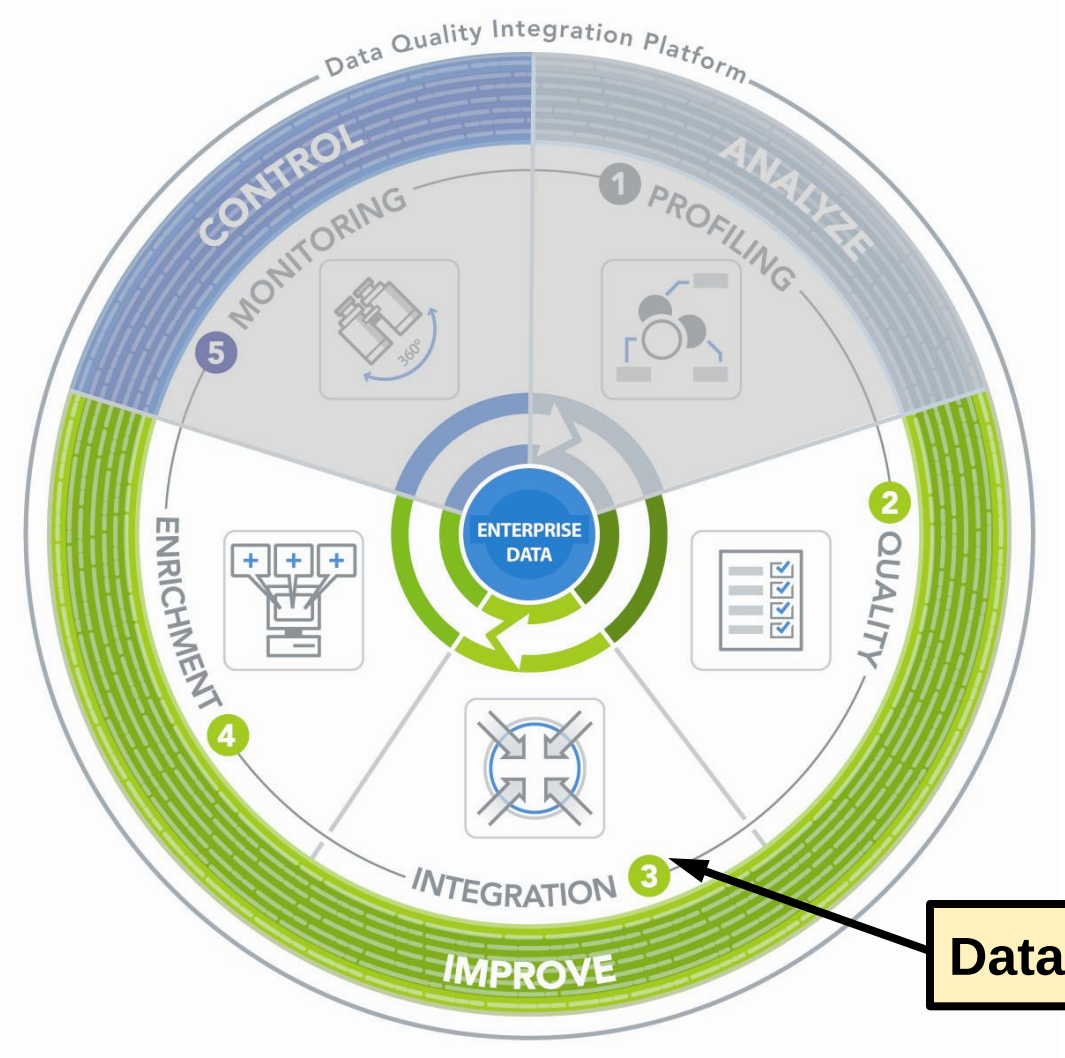
- Ensures uniform representation of a data value
- Used to correct spellings, inconsistent use of nicknames and abbreviations
- Manage child/parent relationships in data
- Establishing consistency in operational data through time
- Allows merging of multiple operational systems

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



DataFlux Methodology: Improve



Data Matching / Integration

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	**GHWS$$EWT$**	**GHWS$$EWT$**	GHWS$$WWI$

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

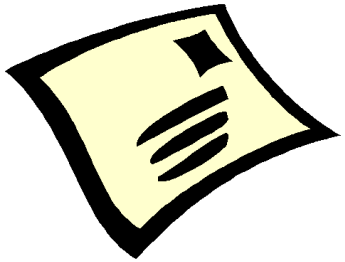
Business Uses for Data Matching

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

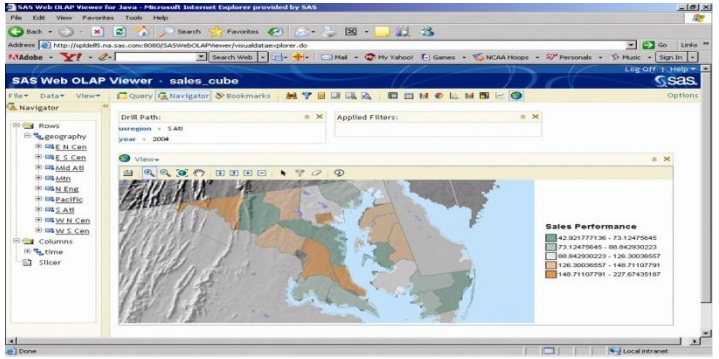
Given Name	Family Name	Address	Phone	MC1	MC2	MC3	HH
Susan	Allen	2208 Vandemere Ct	832-8239	$AV	#V8	%A8	1
Barbara	Sweet	2208 Vandemere Ct	832-8239	$SV	#V8	%S8	1
Richard	Sweet	2208 Vandemere Ct	616-1504	$SV	#V6	%S6	1
Jason	Cheeks	1530 Hidden Cove Dr	688-2856	$GH	#H6	%G6	2
Becker	Ruth	1530 Hidden Cove Dr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

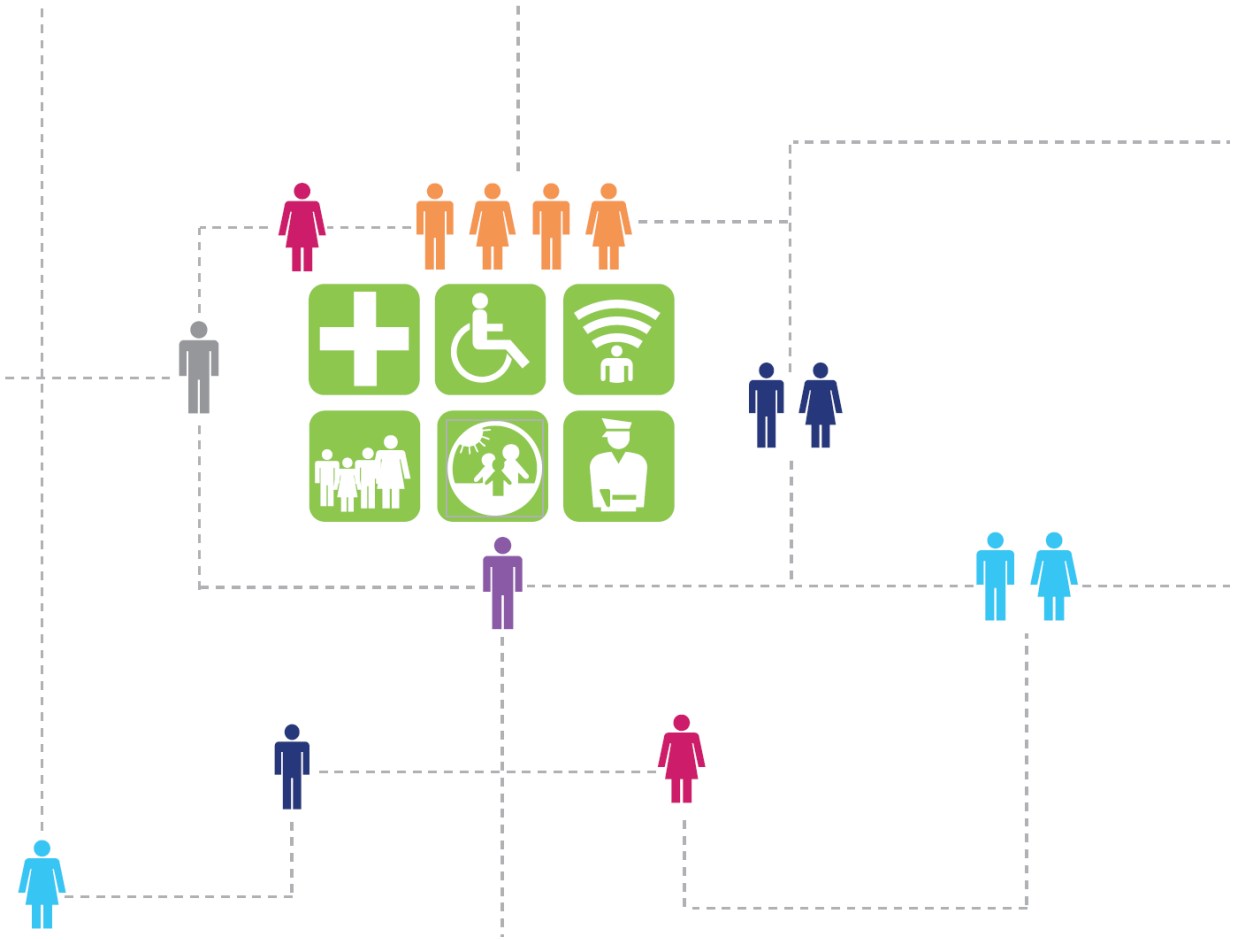


- Geo-Coding addresses for GIS integration



- 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

Standardize

"Anonymization"

Augment

Match

Integrate

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

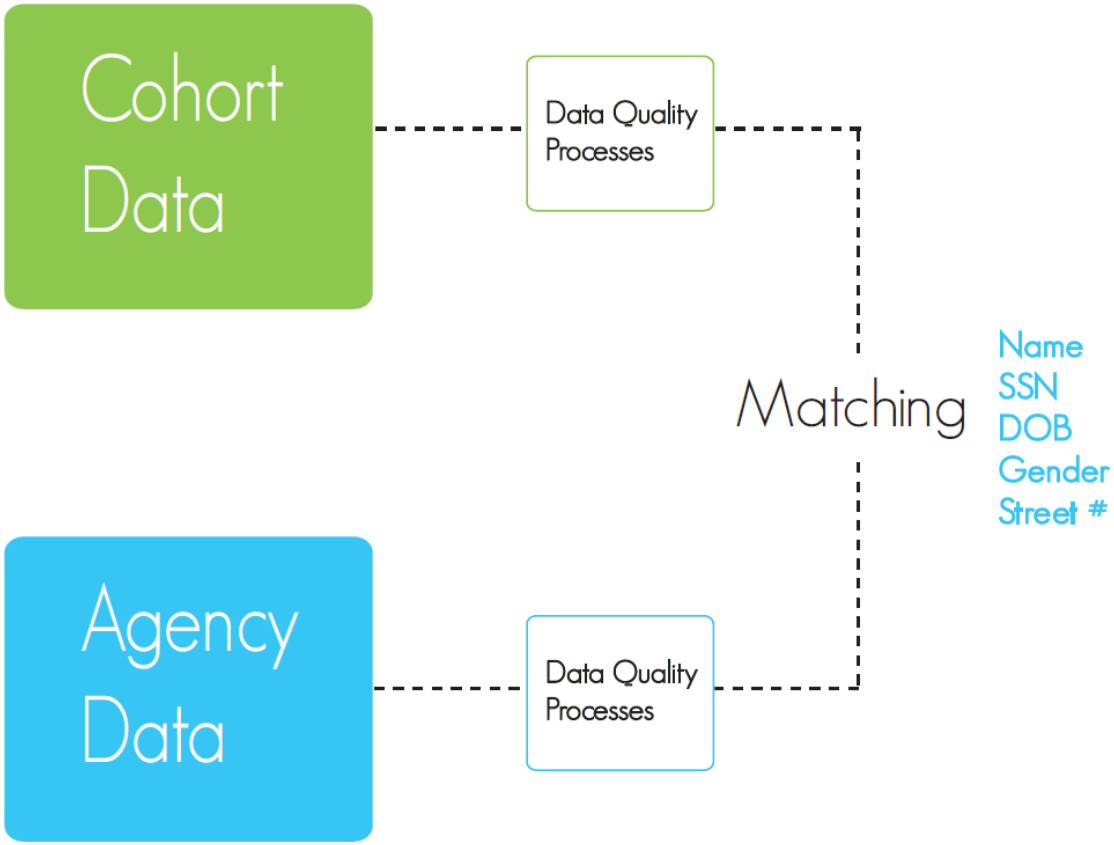
- ALP has established a baseline database for a cohort of GR participants who receive County services from multiple service delivery systems.

- 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

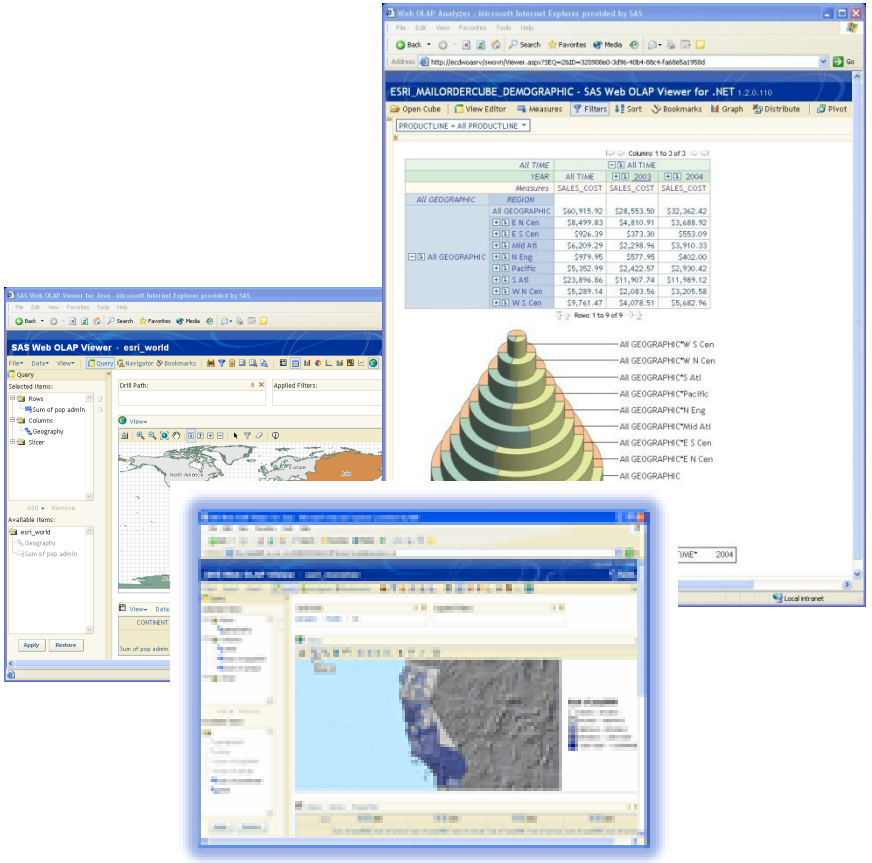
- The linked data set contains information such:
  - Random project IDs for each participant. These markers do not identify any client personally
  - Analysis table for each participating Agency and the difference service types offered by each

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.


sas
THE
POWER
TO KNOW.