

InfoSphere™



Calidad y Gestión de datos

IBM Information Server

*Trusted Information.
Confident Decisions.
Better Business Performance.*



Los retos de la gestión de la información



El reto del negocio

¿Cómo hacer frente a las necesidades de todos los usuarios con una vista completa y coherente de la información?

El reto de la información

¿Cómo entregar información de calidad, dispersa en diferentes sistemas, aplicaciones o repositorios, a la velocidad requerida por negocio?

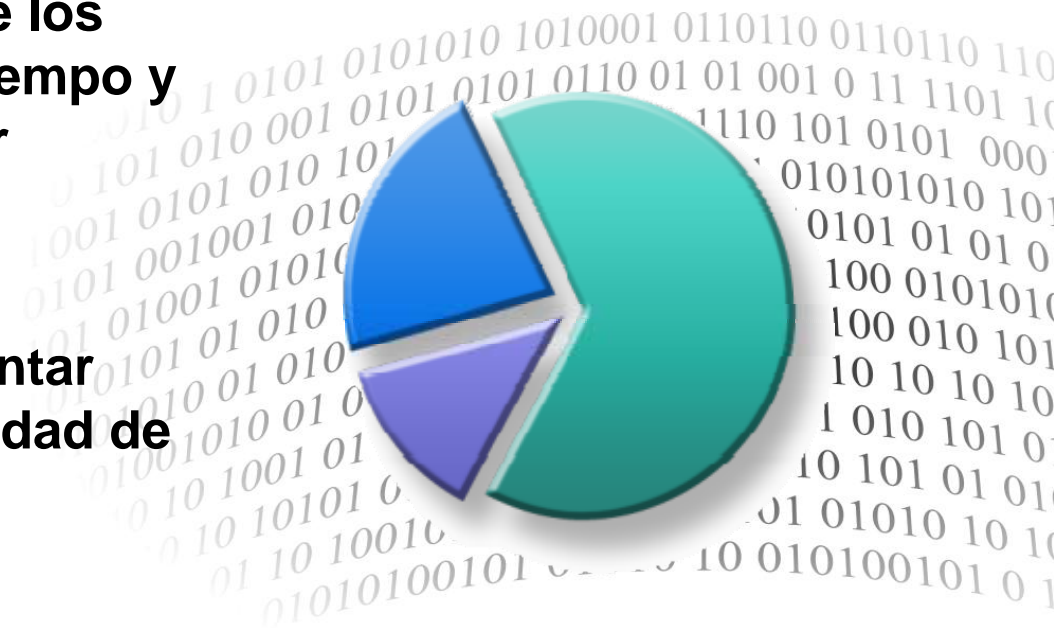
El reto de los procesos

¿Cómo establecer estándares y normas para crear una comunicación entre IT y negocio fluida y eficiente?



Descubriendo la importancia de la calidad de los datos

- **Las empresas empiezan a comprender que los problemas de calidad no sólo les cuesta tiempo y dinero, sino también les impide desarrollar proyectos estratégicos para el negocio**
- **Cada vez más empresas empiezan a implantar procesos de calidad para mejorar la fiabilidad de sus datos**
- **Empresas con una información más precisa y fiable encontrarán más oportunidades para batir a sus competidores**



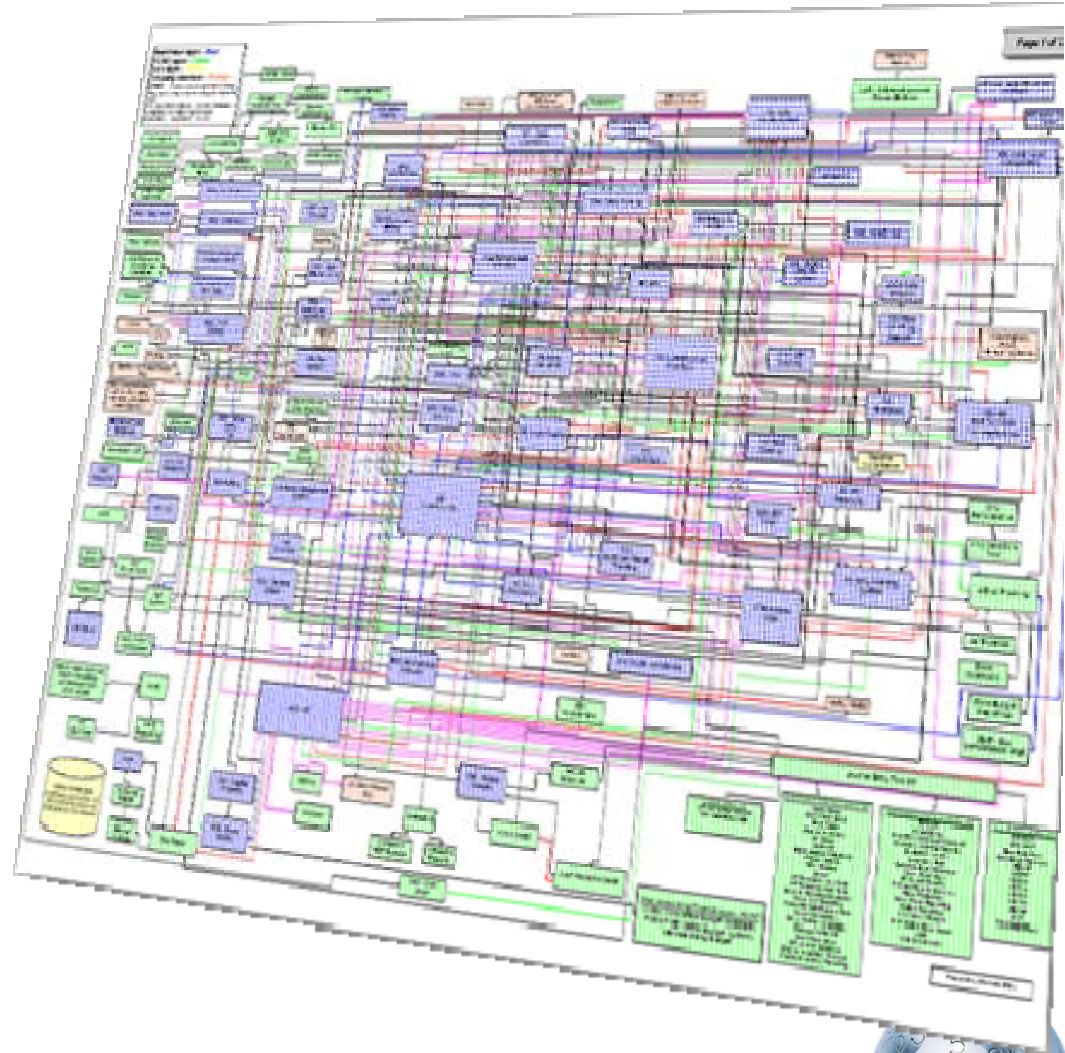
El éxito del negocio depende de la calidad de los datos

- **Facilitando el cumplimiento de normativas legales y el desarrollo de auditorías.**
- **Optimizando las oportunidades de negocio asegurando interacciones eficientes y efectivas con clientes, proveedores y partners.**
- **Permitiendo procesos colaborativos entre diversas áreas con información consistente y fiable.**
- **Reduciendo el coste de la gestión y mantenimiento de información consistente en la organización.**



¿Por qué existe este problema?

- La mayoría de las empresas utilizan diferentes aplicaciones (marketing, ventas, finanzas,...) cada una de ellas con sus propios repositorios.
- No existe un maestro de datos universal a la organización.
- Las aplicaciones no garantizan una vista completa e integrada de los datos, dependen de la calidad de los datos a los que acceden.
- La calidad de los datos se va deteriorando con el tiempo.



Necesidad de calidad de datos

Data Sources	Data Values
	Kentucky Fried Chicken
	KFC
	227G CB&NAT STICK P QUE/MOZZ WRAPP.
	Molly Talber DBA KFC
	Kent Fried Chick
	Kentucky Fried
	Mrs. M. Talber
	227G CB&NATURAL STICK MOZZ WRAPPER
	John & Molly Talber
	Talber, KFC, ATIMA

Problemas Críticos

- Crear y mantener vistas completas de clientes, proveedores, productos, direcciones, eventos...
- Consolidar datos – tomar decisiones fiables, cumplir con normativas, alcanzar acuerdos de servicio

¿Porqué?

- No existen estándares unificados en la organización
- Valores inesperados en algunos campos
- Información enterrada en campos de formato libre
- Los campos evolucionan – utilizados para varios propósitos
- No existen claves adecuadas para vistas consolidadas
- Datos operacionales degradados un 2% al mes

Planteamientos Alternativos

- Negación – el problema es ignorado o mal entendido hasta que es demasiado tarde
- Hand-coding – procesamiento individual de excepciones; requiere mucho tiempo y muchos recursos
- Aplicaciones de limpieza simples – poco potentes y nada flexibles



El reto de la calidad de datos

- No hay estándares de datos - **diferentes formatos y estructuras en diferentes sistemas**
- Valores inesperados - **datos insertados en campos erróneos**
- Información enterrada en campos de formato libre
- Miopía de datos - **falta de identificadores consistentes que proporcionen vistas consolidadas**
- La pesadilla de la redundancia - **Registros duplicados con diferente información y formatos**

Kate A. Roberts 416 Columbus Ave #2, Boston, Mass 02116

Catherine Roberts Four sixteen Columbus APT2, Boston, MA 02116

Mrs. K. Roberts 416 Columbus Suite #2, Suffolk County 02116

Name	Tax ID	Telephone
J Smith DBA Lime Cons.	228-02-1975	6173380300
Williams & Co. C/O Bill	025-37-1888	415-392-2000
1st Natl Provident	34-2671434	3380321
HP 15 State St.	508-466-1200	Orlando

WING ASSY DRILL 4 HOLE USE 5J868A HEXBOLT 1/4 INCH
 WING ASSEMBY, USE 5J868-A HEX BOLT .25" - DRILL FOUR HOLES
 USE 4 5J868A BOLTS (HEX .25) - DRILL HOLES FOR EA ON WING ASSEM
 RUDER, TAP 6 WHOLES, SECURE W/KL2301 RIVETS (10 CM)

19-84-103 RS232 Cable 6' M-F Cands

CS-89641 6 ft. Cable Male-F, RS232 #87951

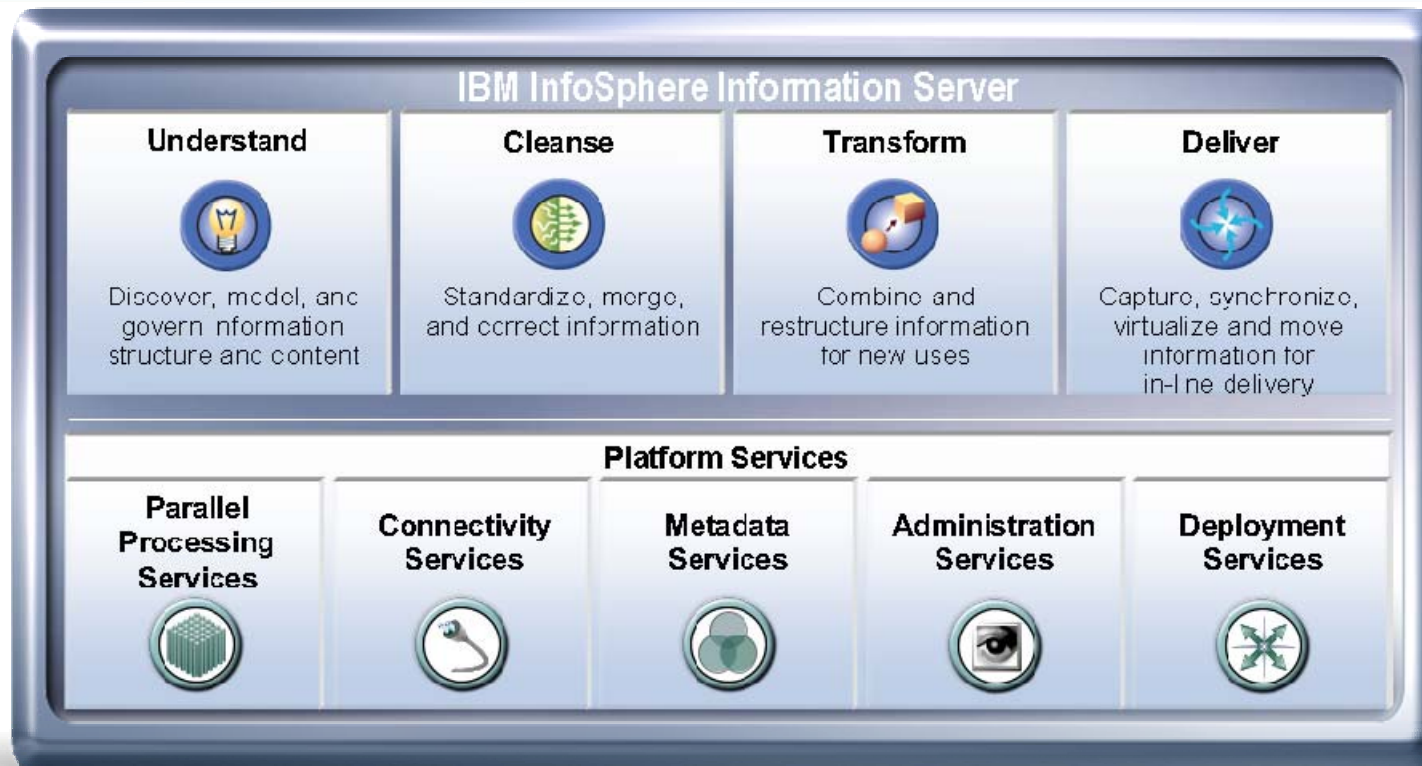
C&SUCH6 Male/Female 25 PIN 6 Foot Cable

90328574	IBM	187 N.Pk. Str. Salem NH 01456
90328575	I.B.M. Inc.	187 N.Pk. St. Salem NH 01456
90238495	Int. Bus. Machines	187 No. Park St Salem NH 04156
90233479	International Bus. M.	187 Park Ave Salem NH 04156
90233489	Inter-Nation Consults	15 Main Street Andover MA 02341
90345672	I.B. Manufacturing	Park Blvd. Bostno MA 04106



Plataforma para la calidad de datos

InfoSphere Information Server



El proceso de la calidad de datos

Establish Data Quality Ownership & Sponsorship

Analizar las fuentes de datos

Medir la evolución de la calidad

Analizando la calidad de los datos

Estandarizar

Certificar y enriquecer

Buscar duplicados

Enlazar o sobrevivir

Limpiando los datos

Evaluar

Informar

Monitorizar la calidad



Analizar las fuentes de datos

- **Da un análisis detallado de la información en los sistemas existentes**
 - Análisis centrado en datos de los fuentes de la aplicación, sean bases de datos o ficheros para ver el contenido, la calidad y la estructura de los mismos
 - Perfilado detallado de todos los campos y relaciones a través de tablas, campos y fuentes de datos.
- **Proporciona un completo conocimiento de la información, mostrando formatos, estructuras, patrones y tendencias, para contribuir al éxito de nuestros proyectos de integración.**
- **Permite la creación de baselines para la comparación entre sucesivos análisis de información.**
- **Creación de umbrales para auditoria**



Subject Matter Experts



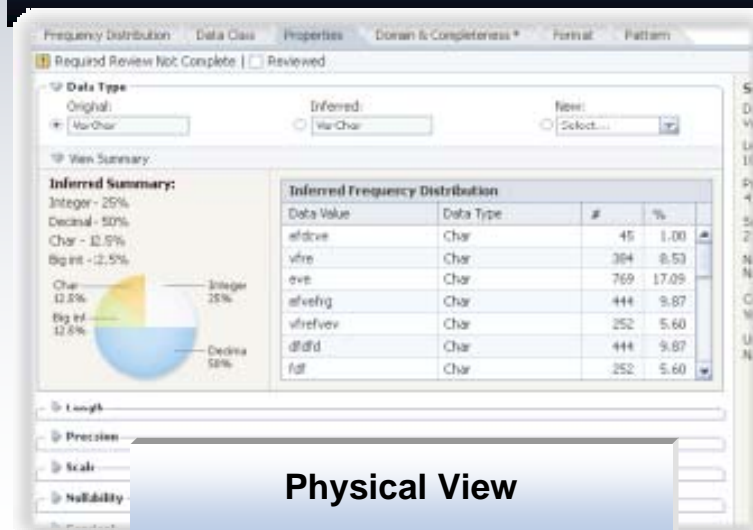
Data Analysts

Understand



IBM Information Analyzer

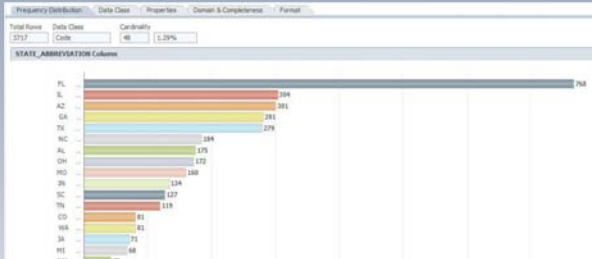
Analyze source data structures, and monitor adherence to integration and quality rules



Physical View

Proceso de análisis

Análisis de Columnas



Distribución de frecuencias

Complete	Incomplete
97.37%	2.63%

Valid	Invalid
99.97%	0.03%

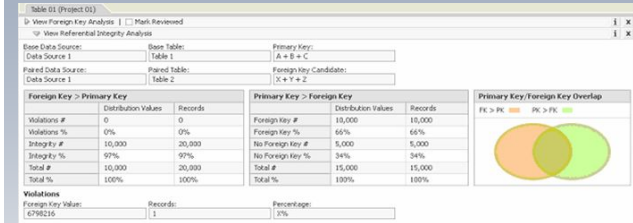
Propiedades, clases, formatos, Dominios, completitud...

Análisis de Tablas

Column	Data Class	Data Type	Length	Unique %	Null %	Duplicate %	Candidate
ORDERSHIP	varchar	255	0	396	-286	False	<input checked="" type="checkbox"/>
ORDERSHIP	varchar	255	0	380	-280	False	<input type="checkbox"/>
ORDERSHIP	varchar	255	0	8	330	False	<input type="checkbox"/>

Análisis de claves primarias

Análisis Cruzados



Foreign Key & Análisis multitabla

Name	Position	Records	Definition	Cardinalit	Data Class	Data Typ	Length	Precisio	Scale	Nullabil	Uniques	For
CTTY	12	3717		40.01	Unknown	Unknown	Char	100				NA

Anotaciones, marcas para revisión,...

Name	Checkpoint	Baseline	Completeness & Validity Measures
STATE_ABBREVIATION	42	41	
Cardinality	1027	1026	
# Distinct Values	2	2	
# Distinct Formats	0	0	
Standard Deviation Value	0	0	
Standard Deviation Percent	0	0	
# Null	3	3	
% Nulls	7.142857	7.317073	

Comparativas en el tiempo

ColumnName	Database Name	Frequency Count	Frequency Cut off
emp_id	MyPub	43	0

Value	Count	Frequency %	Cumulative % of Rows
A-G71970F	1	2.33	2.33
A-R89858F	1	2.33	4.65
AMD15433F	1	2.33	6.98
ARD36773F	1	2.33	9.30
CFH28514M	1	2.33	11.63
CGS88322F	1	2.33	13.95
DBT39435M	1	2.33	16.28
DWR65030M	1	2.33	18.60
FNI44273F	1	2.33	20.93

Generación de Informes



Limpiar los datos

- Proporciona procesamiento específico de calidad de datos
 - Asegura información limpia, estandarizada y sin duplicados
 - Proporciona una visión única de la verdad
 - Soporta verificación postal (Worldwide)
- Proporciona herramientas gráficas para el diseño de reglas de calidad y matching
 - Completamente integrado en la plataforma de IS (un motor, un modelo de metadatos,...)
 - Permite refinar cíclicamente los procesos de calidad
- Permite utilizar la lógica de calidad directamente con ETL o como un servicio distribuido



Subject Matter Experts



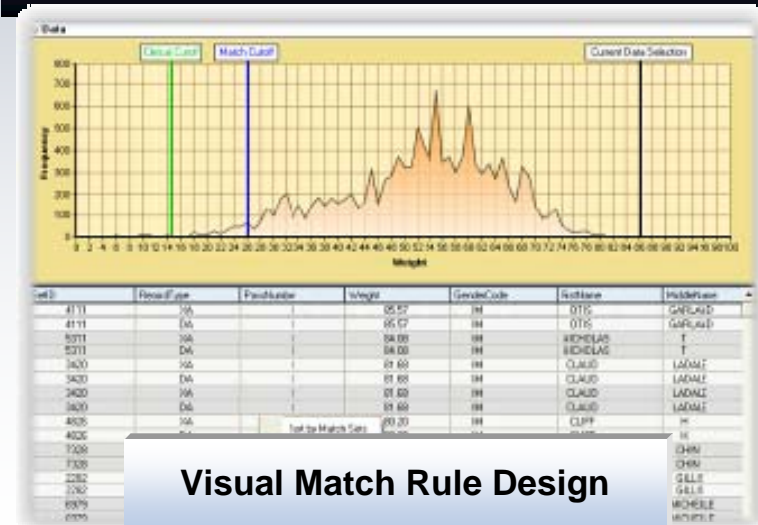
Data Analysts

Cleanse



WebSphere QualityStage™

Standardize and correct source data fields, and match records together across sources to create a single view



Proceso de limpieza

Investigación

123 St. Virginia St.

Parsing: 123 | St. | Virginia | St.

Separar campos multievaluados en piezas individuales

Análisis Léxico: 123 | St. | Virginia | St.

Identificar el significado de dichas piezas

Contexto Sensitivo: 123 | St. Virginia | St.

Descubrir estructuras de datos y su información

“Las instrucciones para la manipulación de los datos son inherentes dentro de los propios datos.”

Investigación en su contexto

Estandarización

Input File:

Address Line 1	Address Line 2
639 N MILLS AVENUE	ORLANDO, FLA 32803
306 W MAIN STR, CUMMING, GA 30130	
3142 WEST CENTRAL AV	TOLEDO OH 43606
843 HEARD AVE	AUGUSTA-GA-30904
1139 GREENE ST ACCT #1234	AUGUSTA GEORGIA 30901
4275 OWENS ROAD SUITE 536 EVANS	GA 30809
1775 RUSSELL CIRCLE MILLIS MASSACH	USETTS 02038

Result File:

House #	Dir	Str. Name	Type	Unit	No.	NYSIIS	City	SOUNDEX	State	Zip	ACCT#
639	N	MILLS	AVE			MAL	ORLANDO	O645	FL	32803	
306	W	MAIN	ST			MAN	CUMMING	C552	GA	30130	
3142	W	CENTRAL	AVE			CANTRAL	TOLEDO	T430	OH	43606	
843		HEARD	AVE			HAD	AUGUSTA	A223	GA	30904	
1139		GREENE	ST			GRAN	AUGUSTA	A223	GA	30901	1234

Datos estandarizados y organizados

Pass Definition
Pass Statistics

Save Pass Test Pass

Blocking Columns:

- Phonetic Last Name
- Phonetic Street Name
- First Character of Match First Name
- Full Postal Code
- First Character of House Number

Test Results:

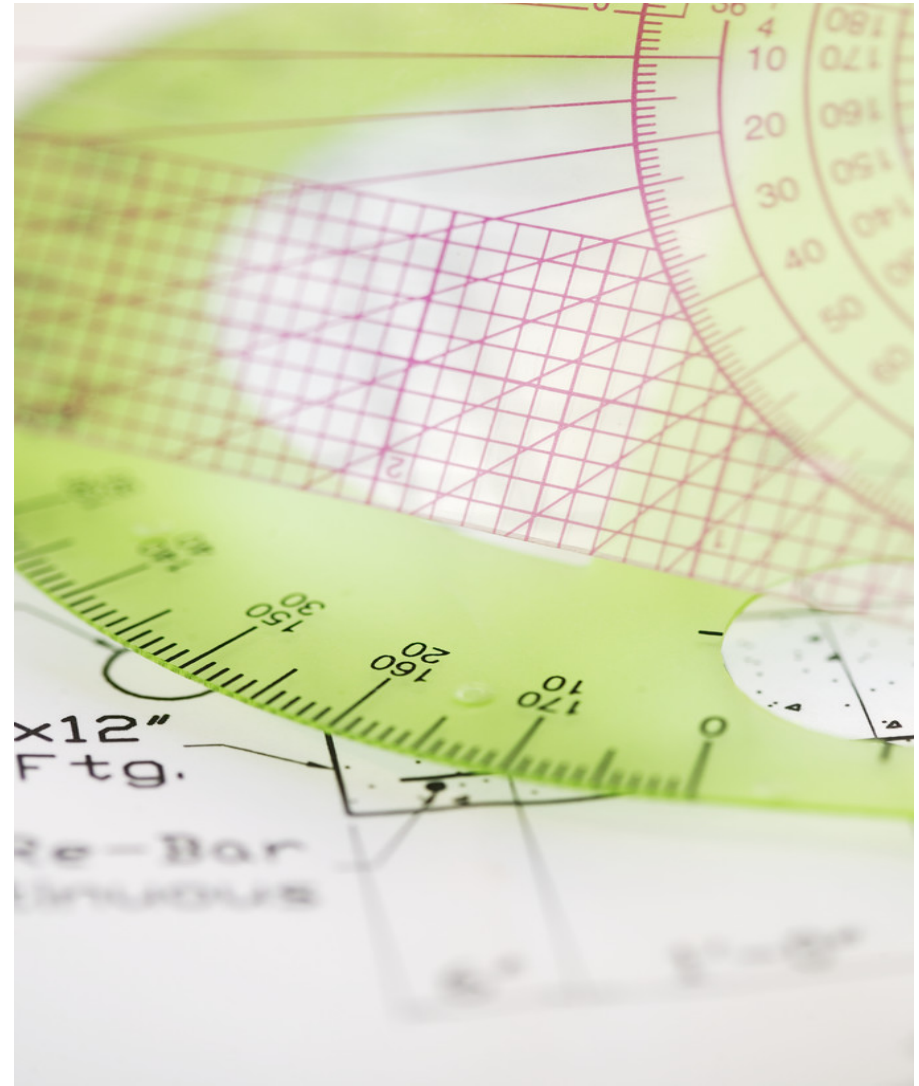
SetID	Record Type	Weight	DataID	Last Name	Match First Name	House Number	Street Name
4	XA	40.08	4	COGBORN	JAMES	3	NOTCH
4	DA	27.52	600	COGBORN	JAMES	3	NOTCH
4	DA	33.01	243	COGBORN	JAMES	3	NOTCH
							NOTCH
							NOTCH
						15	OXFORD
						15	OXFORD

Matching probabilístico y supervivencia



Resumen

- La calidad de los datos se está convirtiendo en una cuestión cada vez más importante dentro de la organización
- Mejorar la calidad de los datos y garantizar la entrega de información fiable requiere programas específicos y con diferentes enfoques
- El objetivo de cualquier programa de calidad de datos es el de proporcionar servicios auditables capaces de mejorar la calidad de nuestros datos.
- IBM InfoSphere Information Server y Cognos 8 BI nos permiten comprender nuestros datos y su fiabilidad.



¿Cómo puede ayudar IBM?

- Con la plataforma más completa para la evaluación de la calidad, limpieza y monitorización de los datos.
- Con su conocimiento sectorial específico y su experiencia en la implantación de procesos de calidad de los datos.
- Con la propuesta de evaluación de la calidad para comprobar el valor comercial de los procesos de limpieza
- Consulte las demostraciones de integración de InfoSphere con Cognos 8 BI en el “Solution Center”
- Contacte con su comercial de IBM o visite: www.ibm.com/infosphere
- Gracias por su tiempo

