

IBM Software Expo 2006. Madrid 23 de Mayo

Consiga datos fiables



Philip Little
IBM Information Integration Solutions

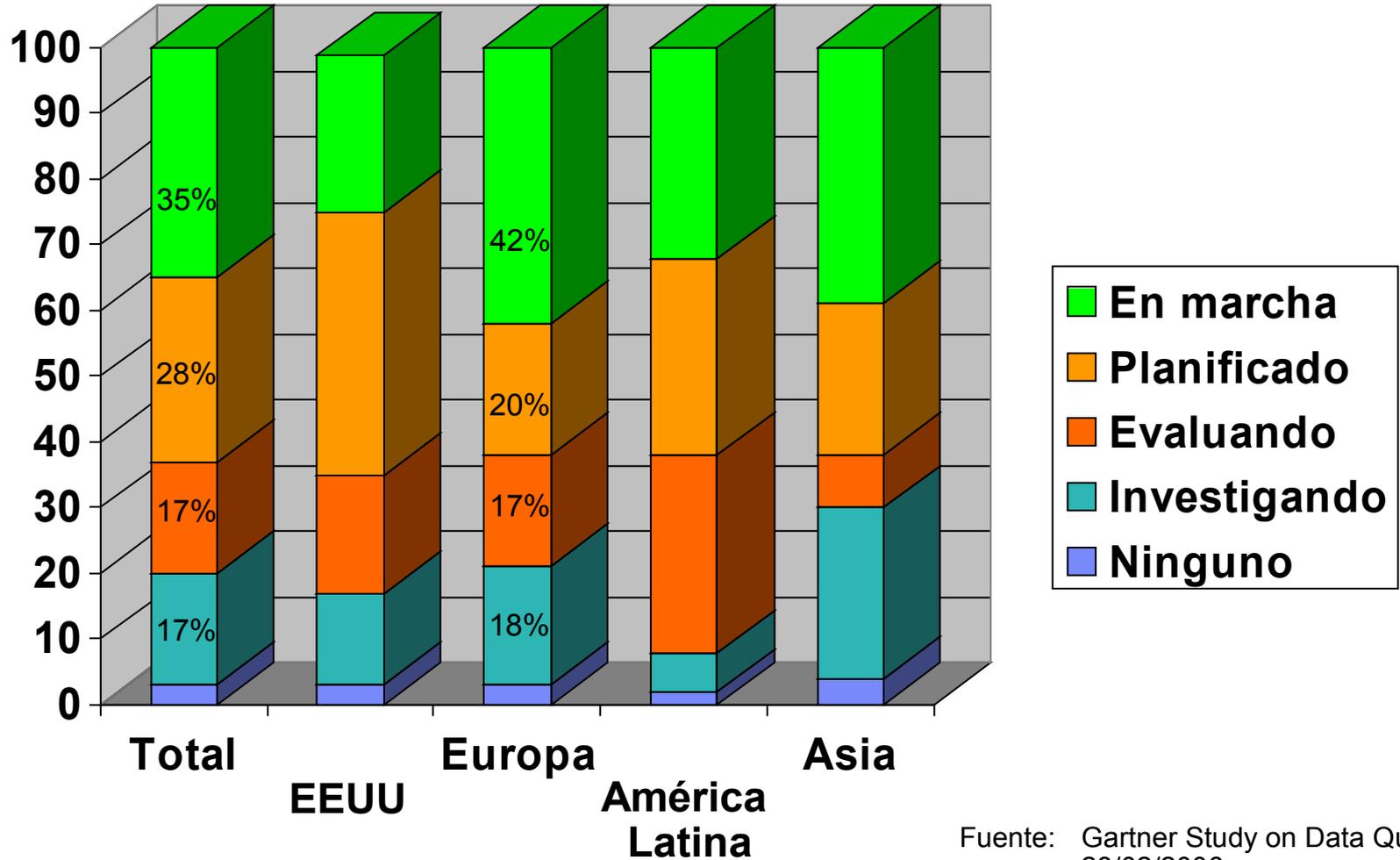


Agenda

- Situación actual
- Los procesos para mejorar la calidad de datos



Un problema reconocido



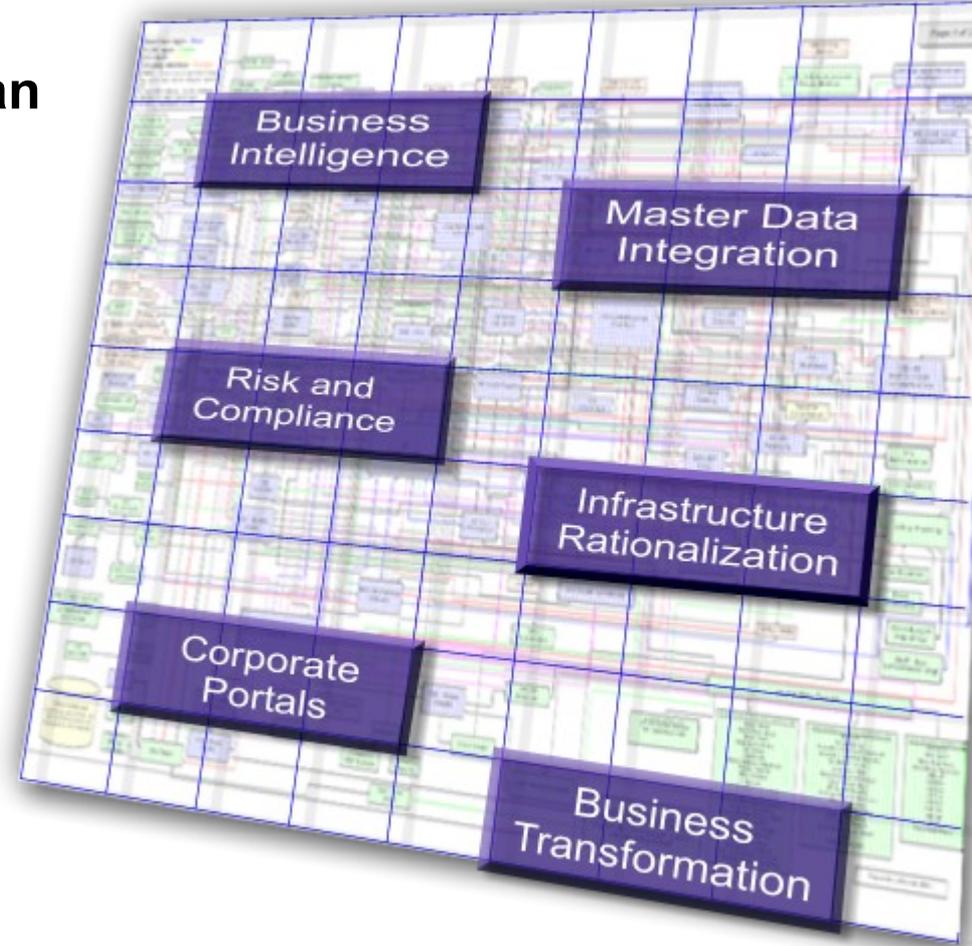
Fuente: Gartner Study on Data Quality
23/02/2006



Iniciativas de negocio críticas que necesiten Información Integrada de Calidad

Todas estas iniciativas necesitan información que sea:

- **Accesible**
- **Fiable / Correcto**
- **Consistente**
- **A Tiempo**
- **Completa**



Calidad de Datos

Los Mitos de la Calidad

- **La Calidad de Información es para TI**
 - ▶ Muy ligada al negocio y sus reglas
- **La Información está bien o mal**
 - ▶ La calidad de información depende del uso / necesidad
- **Lo depuramos el año pasado - ¡todo está bien!**
 - ▶ La calidad de información es dinámica, afectado por personas, procesos y aplicaciones
- **Conocemos nuestros datos**
 - ▶ La documentación existe
 - ▶ Los datos corresponden a sus meta datos
- **Las reglas de negocio no han cambiado**

La realidad de los Datos

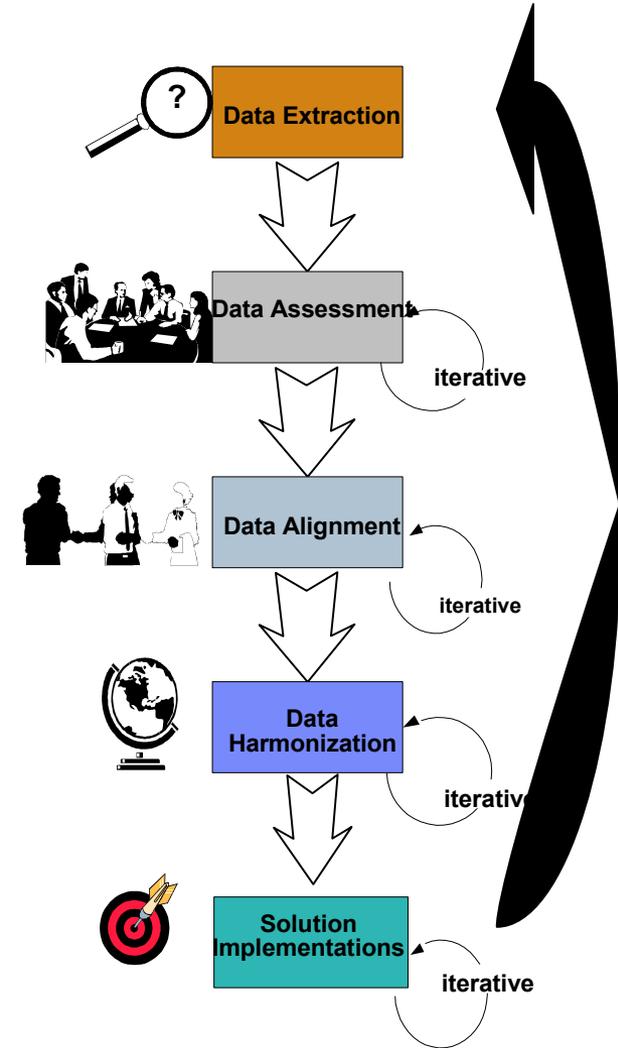
- La mayoría de organizaciones tienen **distintas aplicaciones** para las áreas de ventas, servicios, marketing, manufacturing, y financiero – **cada una con sus propios “maestros”**
- **No existe un consenso sobre un registro común** de sistema
- Hay que **“re-dirigir”** datos **“antiguos”** a las **nuevas aplicaciones**
- No es sólo un problema de entrada de datos, es un **problema de integración / reconciliación**
- Es **demasiado tarde y demasiado caro** arreglar los datos **después de la implantación**



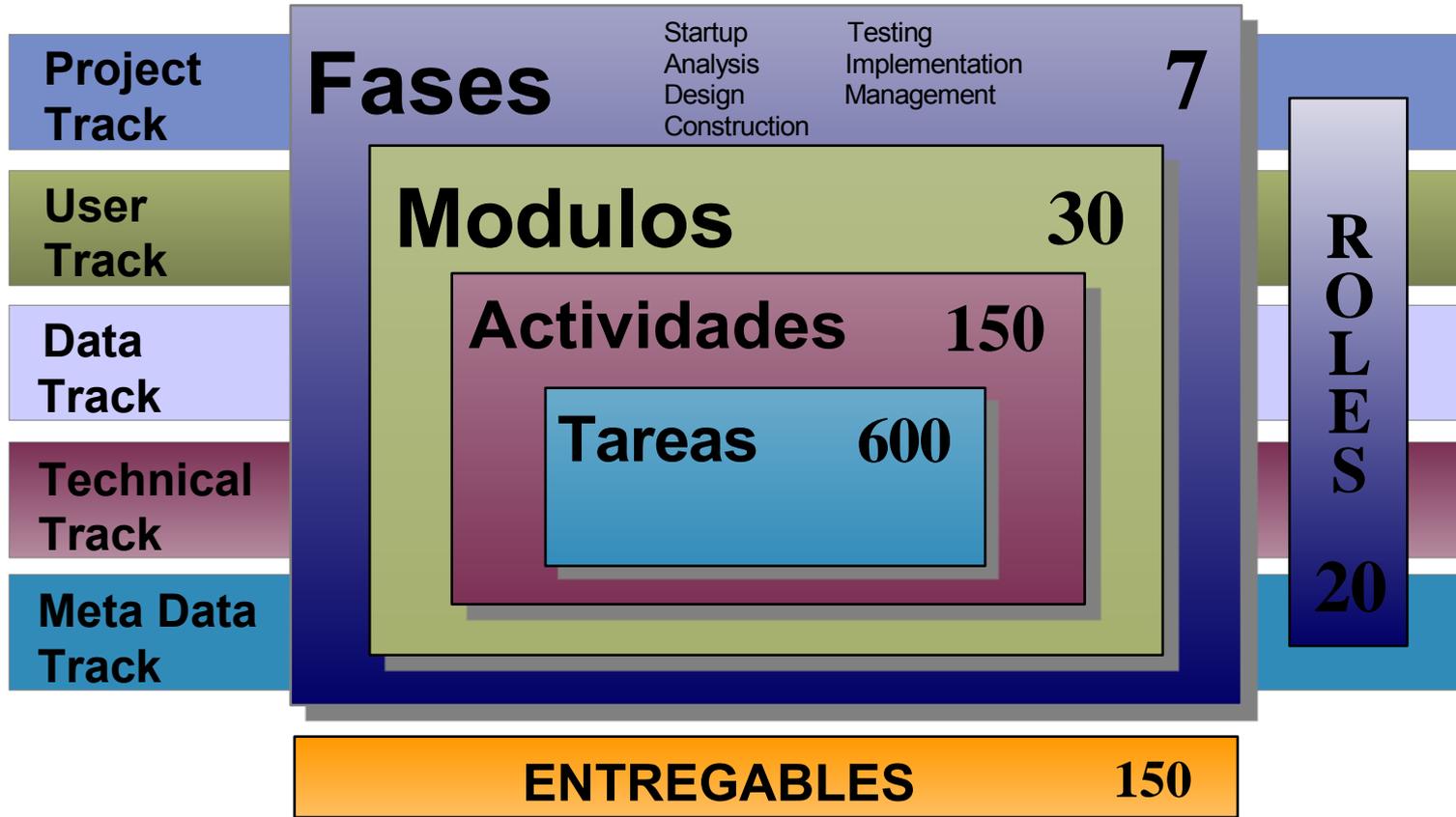
El Control de Calidad necesita

- Estándares para los datos de negocio
- Responsables para los datos de negocio
- Calidad de Información
- Manejo de Meta datos
- Planificación de migración
- Seguridad y Protección de Datos
- Procesos estándares en el uso de datos en el desarrollo de aplicaciones

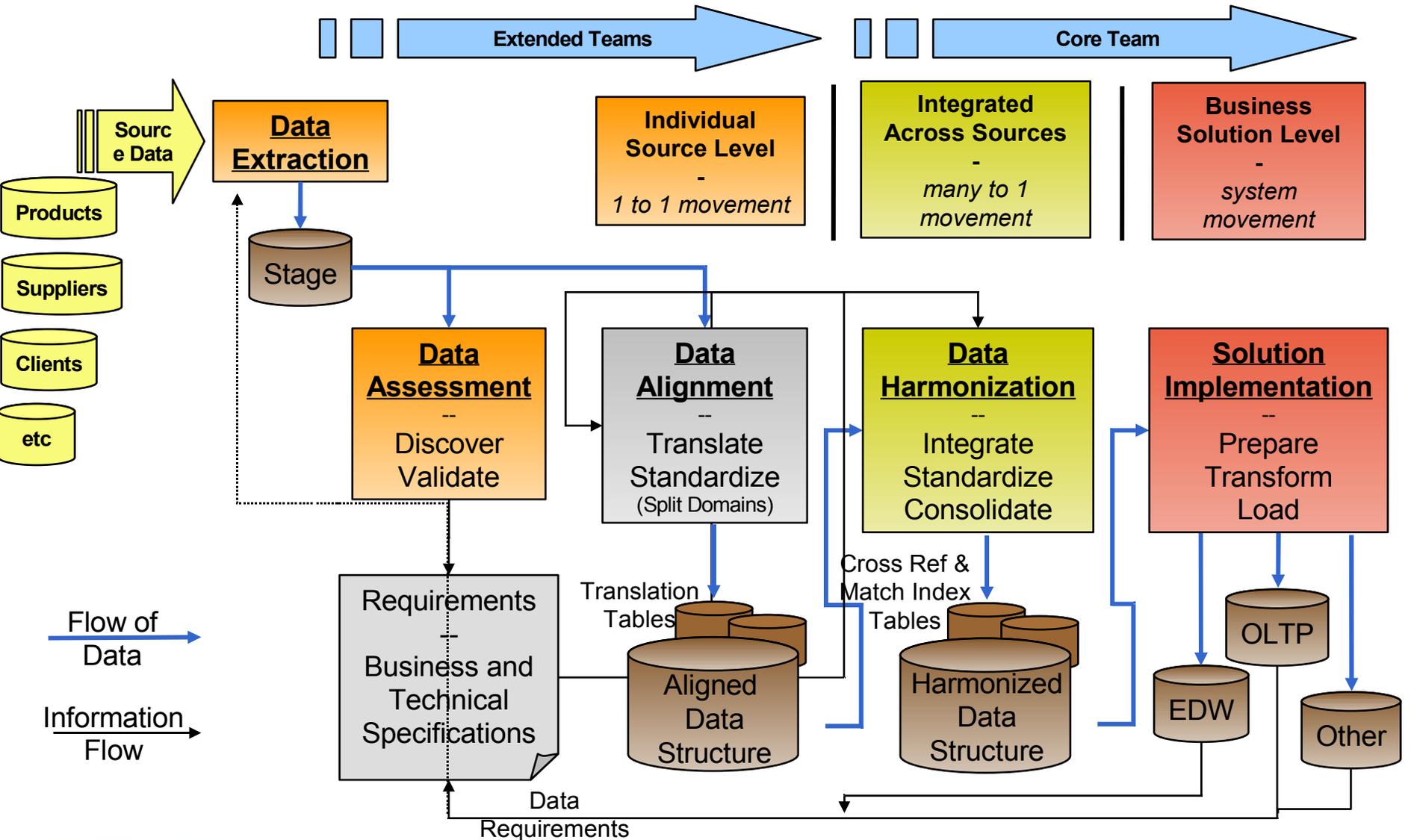
Productos + Metodología



Metodología: Iterations 2



Integración de Datos – el proceso en acción !



Algunas Tipos de Anomolías

Información Incorrecta	La definición o contenido no está alineada con lo que necesita el proceso de negocio
Campos Inconsistentes	El contenido del campo es inconsistente entre registro y registro. El mismo valor en un campo significa cosas distintas
Sentidos Múltiples	El mismo campo contiene múltiples tipos de información dependiendo del registro
Campos Duplicados	Dos campos distintas con nombres distintas contienen la misma información y sentido
Sentidos Compuestos	El campo tiene códigos múltiples compuestos
Correlación Física Errónea	La características físicas del fuente no corresponden con el destino
Información no Disponible	Falta información obligatoria según las reglas de negocio
Formatos Inconsistentes	El contenido tiene múltiples formatos, dificultando su entendimiento y proceso
Valore Permitidos Inconsistentes	El contenido puede conformar a distintos valores permitidos
Valores Duplicados	No deben existir valores duplicados, pero existen

Análisis de Columnas

ProfileStage

File Set Up Analysis Integration Reports Data Repository Tools Help

1) Select a database...
 Database1
 Tables in Database - 12
 Columns in Database - 94

2) Select a process...
 Column Analysis

Analysis status for this database

Tables Excluded - 8% (1)	Columns Excluded - 22% (21)						
Tables Reviewed - (see below)	Columns Reviewed - 31% (29)						
E	COLUMN ANALYSIS	TABLE ANALYSIS	PRIMARY KEY ANALYSIS	ACCEPT PRIMARY KEYS	CROSS-TABLE ANALYSIS	RELATIONSHIP ANALYSIS	ACCEPT RELATIONSHIPS
X	33%	25%	25%	25%	25%	25%	25%
C	(4)	(3)	(3)	(3)	(3)	(3)	(3)

3) Drag and drop tables or columns to an Analysis Server...

- View Progress
- Confirm Property Changes
- Copy Existing Keys
- Create Package:

EDFOLEYW2K

Database1

- Categories.bt
- Cust_Info.bt
- Division.bt
- Employees.bt
- EmployeeTerritories.bt
- OrderDetails.bt
- Orders.bt
- Organization.bt
- Products.bt
- Shippers.txt
- Suppliers.bt
- Territories.bt

Column Analysis for Database1!Employees.txt

Column Name	Value	Percent	Chosen
EMPLOYEEID	Smallint	97.8	Smallint
SSN			
LASTNAME			
FIRSTNAME			
TITLE			
TITLEOFCOURTESY			
BIRTHDATE			
HIREDATE			
ADDRESS			
CITY			
REGION			
POSTALCODE			
COUNTRY			
HOMEPHONE			
EXTENSION			

Exclude Column from Target Database
 Exclude Column from Analysis

Review Complete for Employees.bt

Analysis Results View Source View Sample

Close Help

Database: Database1 Column Analysis

Análisis de Tablas

ProfileStage
 File Set Up Analysis Integration Reports Data Repository Tools Help

1) Select a database...
 Database1
 Tables in Database - 12
 Columns in Database - 94

2) Select a process...
 Table Analysis

Analysis status for this database
 Tables Excluded - 8% (1) Columns Excluded - 22% (21)
 Tables Reviewed - (see below) Columns Reviewed - 31% (29)

	COLUMN ANALYSIS	TABLE ANALYSIS	PRIMARY KEY ANALYSIS	ACCEPT PRIMARY KEYS	CROSS-TABLE ANALYSIS	RELATIONSHIP ANALYSIS	ACCEPT RELATIONSHIPS
E	33%	25%	25%	25%	25%	25%	25%
X	(4)	(3)	(3)	(3)	(3)	(3)	(3)
C							

3) Drag and drop tables or columns to an Analysis Server...
 View Progress
 Confirm Property Changes
 Copy Existing Keys
 Create Package:

Database1

- Categories.bt
- Cust_Info.bt
- Division.bt
- Employees.bt
- EmployeeTerritories.bt
- OrderDetails.bt
- Orders.bt
- Organization.bt
- Products.bt
- Shippers.txt
- Suppliers.bt
- Territories.bt

Table Analysis for Database1\Employees.txt

Determinant	[Key Coverage %]	Dependent Column	Dependency %
<input checked="" type="checkbox"/> EMPLOYEEID	[100%]	ADDRESS	100
<input checked="" type="checkbox"/> SSN	[100%]	BIRTHDATE	100
<input type="checkbox"/> ADDRESS,CITY	[100%]	CITY	100
<input type="checkbox"/> ADDRESS,FIRSTNAME	[100%]	COUNTRY	100
<input type="checkbox"/> ADDRESS,LASTNAME	[100%]	DIVISIONID	100
<input type="checkbox"/> ADDRESS,POSTALCODE	[100%]	EXTENSION	100
<input type="checkbox"/> EXTENSION	[100%]	FIRSTNAME	100
<input type="checkbox"/> FIRSTNAME,LASTNAME	[100%]	HIREDATE	100
<input type="checkbox"/> HOMEPHONE	[100%]	HOMEPHONE	100
<input type="checkbox"/> HOMEPHONE	[100%]	LASTNAME	100
<input type="checkbox"/> BIRTHDATE,LASTNAME	[40%]	NOTES	100
<input type="checkbox"/> HIREDATE,LASTNAME	[40%]	POSTALCODE	100
<input type="checkbox"/> ADDRESS,BIRTHDATE	[33.33%]	REGION	100
<input type="checkbox"/> BIRTHDATE,FIRSTNAME	[33.33%]	SSN	100
<input type="checkbox"/> FIRSTNAME,HIREDATE	[33.33%]	TITLE	100
<input type="checkbox"/> ADDRESS,DIVISIONID	[20%]	TITLEOFCOURTESY	100

Review Complete for Employees.bt

View Dependent Values

Add Notes Test New Close Help

EMPLOYEEID

Database: Database1 Table Analysis

Análisis de Tablas Cruzadas

The screenshot shows the ProfileStage software interface. The main window displays the 'Analysis status for this database' and a table of analysis results. The 'Employees.txt' table is selected for cross-table analysis, with 25% of columns reviewed and 31% of columns excluded. The analysis results are shown in a table below.

Analysis status for this database

Tables Excluded - 8% (1) Columns Excluded - 22% (21)
 Tables Reviewed - (see below) Columns Reviewed - 31% (29)

	COLUMN ANALYSIS	TABLE ANALYSIS	PRIMARY KEY ANALYSIS	ACCEPT PRIMARY KEYS	CROSS-TABLE ANALYSIS	RELATIONSHIP ANALYSIS	ACCEPT RELATIONSHIPS
E	33% (4)	25% (3)	25% (3)	25% (3)	25% (3)	25% (3)	25% (3)
X							
C							

Employees.txt Cross-Table Analysis Review

Drag a column header here to group by that column.

BaseDbName	BaseTableName	PairedDbName	PairedTableName	BaseColumnName	PairedColumnName	Percentage	ExcludedFlag
Database1	Employees.txt	Database1	EmployeeTerritories.txt	FIRSTNAME	EMP_FIRSTNAME	100	N
		Database1	EmployeeTerritories.txt	EMPLOYEEID	EMPLOYEEID	100	N
		Database1	EmployeeTerritories.txt	LASTNAME	EMP_LASTNAME	100	N
	EmployeeTerritories.txt	Database1	Employees.txt	EMP_FIRSTNAME	FIRSTNAME	100	N
		Database1	Employees.txt	EMPLOYEEID	EMPLOYEEID	100	N
		Database1	Employees.txt	EMP_LASTNAME	LASTNAME	100	N

Informes

Column Analysis: Candidate Data Types - Ambiguous Columns Only

1 of 1 Total: 15 100% 15 of 15

Column Analysis: Candidate Data Types - Ambiguous Columns Only
Database: ProfileStage

Table Name	Column Name	Data Type Date Format	Summary of Types	
			Percent	Count
dataMart.txt	CUSTID	Char	0.50%	1
		Integer	99.50%	199
Orders.txt	ORDERAMT	Integer	9.52%	2
		Smallint	66.67%	14
		Tinyint	23.81%	5
		Smallint	42.86%	9
register.txt	SHIPCST	Tinyint	57.14%	12
		Smallint	1.50%	3
weborder.txt	ZIP4	Tinyint	98.50%	197
		Integer	34.50%	69
weborder.txt	APPSSN	Tinyint	65.50%	131
		Integer	96.50%	193
		Smallint	3.50%	7
		Smallint	53.50%	107
weborder.txt	POLICY	Integer	96.50%	193
		Smallint	3.50%	7
weborder.txt	ZIP4	Smallint	53.50%	107
		Tinyint	46.50%	93

start Kelly.doc - Microsoft ... ProfileStage Column Analysis: Can... 2:20 PM

Tendencias en Calidad



¿Cómo se puede obtener una visión del negocio consolidada y acertada?



1: Investigación - Palabras

123 St. Virginia St.

Parseo:

123 | St. | Virginia | St.

Separación de campos con múltiples valores en piezas individuales

Análisis Léxico:

Number	Street Type	Alpha	Street Type
123	St.	Virginia	St.

Determinar el significado de negocio de piezas individuales

Sensitivo al contexto:

House Number	Street Name	Street Type
123	St. Virginia	St.

Identificando estructuras de datos y contenidos variados

“Las instrucciones para manejar los datos son inherentes a los datos mismos”

2: Normalización - Direcciones

Fichero Entrada:

Dirección Línea 1

639 N MILLS AVENUE
 306 W MAIN STR, CUMMING, GA 30130
 3142 WEST CENTRAL AV
 843 HEARD AVE
 1139 GREENE ST ACCT #1234
 4275 OWENS ROAD SUITE 536 EVANS
 1775 RUSSELL CIRCLE MILLIS MASSACH

Dirección Línea 2

ORLANDO, FLA 32803

 TOLEDO OH 43606
 AUGUSTA-GA-30904
 AUGUSTA GEORGIA 30901
 GA 30809
 USETTS 02038

Fichero Resultado:

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
4275		OWENS	RD	STE	536	ON	EVANS	E152	GA	30809	
1775		RUSSELL	CIR			RASAL	MILLIS	L260	MA		02038

2: Normalización - Materiales

Fichero de Entrada:

Operation Work Instruction

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 EACH ON WING ASSEM
 RUDER, TAP 6 WHOLES, SECURE W/KL2301 RIVETS (10 CM)

Fichero resultado:

Assembly	Instruction	QTY	Type	Part	Size	Unit Measure	SKU
WING	DRILL	4	HOLES	HEXBOLT	.25	INCH	5J868A
WING	DRILL	4	HOLES	HEXBOLT	.25	INCH	5J868A
WING	DRILL	4	HOLES	HEXBOLT	.25		5J868A
RUDDER	TAP	6	HOLES	RIVET	10	CM	KL2301

3: Emparejamiento

Match Designer - Specification: NameAndAddress

Compose | Total Statistics

Match Type: Unduplicate

Data Link → Standardized Data → NameAndStreet → NameAndPOBox

Match Pass Holding Area

This area holds Match Passes as part of the Match job. To add key and drag the Pass from the

Match Pass - NameAndStreet

Overview | Statistics

Blocking Columns

Descriptions

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

Match Commands

Descriptions

- Last Name
- Match First Name
- House Number
- Street Name
- Street Suffix Type
- Tax ID
- Gender
- Middle Name
- Generational

Test Results - 213 Records (Displayed as: Match Sets, ascending by master SetID, without Residuals)

SetID	Record Type	Weight	DataID	MATCHPRIMARYNAM	MATCHFIRSTNAME	HOUSENUMBER	STREETNAME	STREETSUFFIX
3	XA	40.08	3	COGBORN	JAMES	3	NOTCH	ST
3	DA	27.52	599	COGBORN	JAMES	3	NOTCH	ST
3	DA	33.01	242	COGBORN	JAMES	3	NOTCH	ST
3	DA	34.59	48	COGBORN	JAMES	3	NOTCH	ST
3	DA	40.08	38	COGBORN	JAMES	3	NOTCH	ST
9	XA	39.75	9	MCKINNEY	KATHRYN	2615	OXFORD	DR
9	DA	39.75	204	MCKINNEY	KATHRYN	2615	OXFORD	DR
23	XA	38.92	23	BELL	GEORGIA	4030	OLD PIKE	RD
23	DA	26.65	258	BELL	GEORGE	4030	OLD PIKE	RD
36	XA	41.64	36	SEGARS	LUCINDA	104	TATE	ST
36	DA	37.57	282	SEGARS	LUCINDA	104	TATE	ST
39	XA	40.91	39	FREDERICK	ELNA	309	WOODLAND	CIR
39	DA	40.91	139	FREDERICK	ELNA	309	WOODLAND	CIR
46	XA	32.96	46	JOHNSON	CHARLES	3	JUDKINS	RD
46	DA	32.96	188	JOHNSON	CHARLES	3	JUDKINS	RD
59	XA	39.59	59	KING	LEONA	312	WOODLAND	CIR
59	DA	39.59	207	KING	LEONA	312	WOODLAND	CIR

4: Supervivencia - Piezas

Entrada para supervivencia (Salida del Match)

Group	Record	Operation Work Instruction
1	1	WING ASSY DRILL 4 HOLE USE 5J868A HEXBOLT 1/4 INCH
1	2	WING ASSEMBY, USE 5J868-A HEX BOLT .25" - DRILL FOUR HOLES
1	3	USE 4 5J868A BOLTS (HEX .25) - DRILL HOLES FOR EACH ON WING ASSEM
2	4	RUDER, TAP 6 WHOLES, SECURE W/KL2301 RIVETS (10 CM)

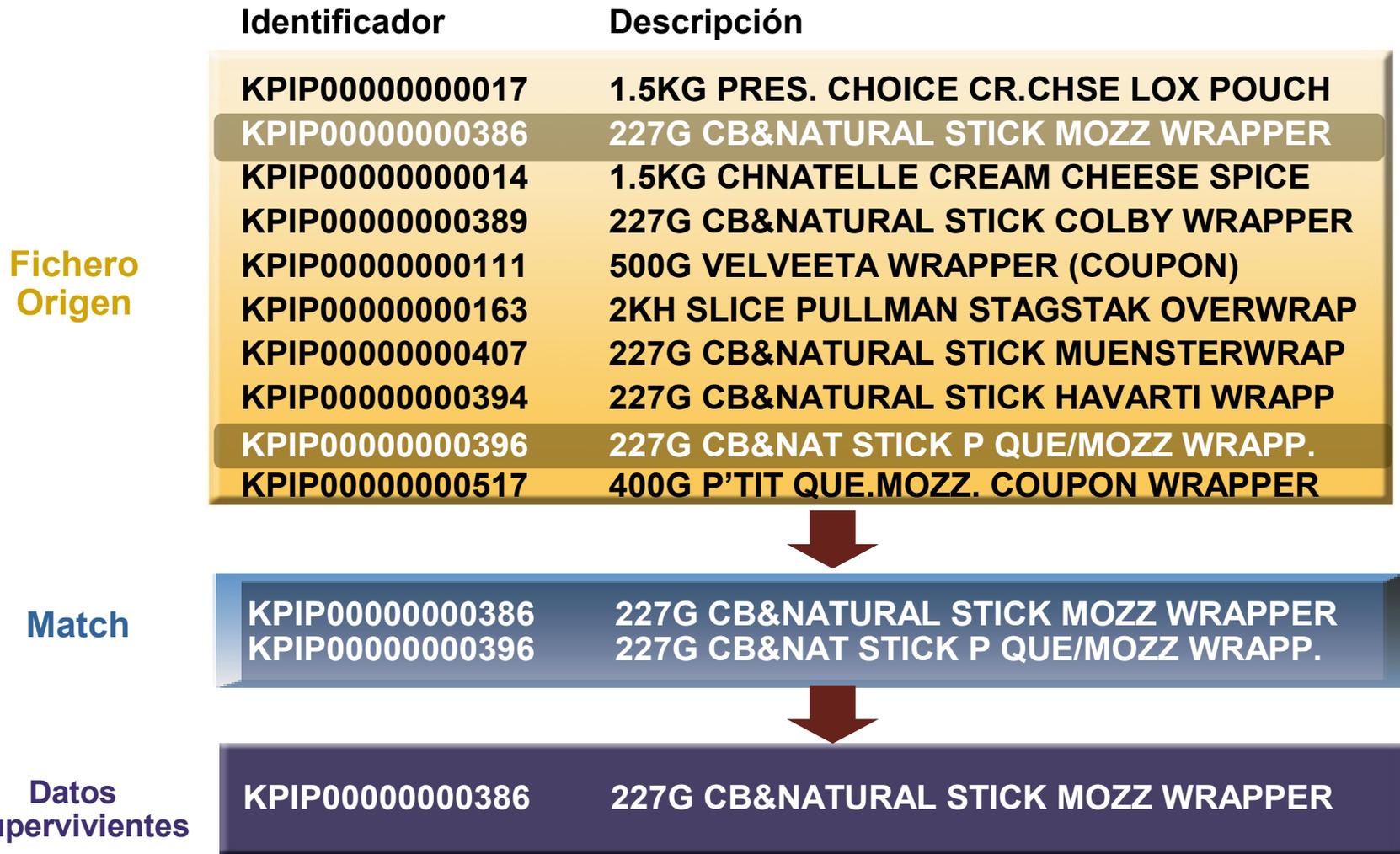
Registros consolidados

Assembly	Instruction	QTY	Type	Part	Size	Unit Measure	SKU
WING	DRILL	4	HOLES	HEXBOLT	.25	INCH	5J868A
RUDDER	TAP	6	HOLES	RIVET	10	CM	KL2301

Clave de referencia cruzada

Group	Assembly	Instruction	QTY	Type	Part	Size	Unit Measure	SKU	Record	Group
									1	1
1	WING	DRILL	4	HOLES	HEXBOLT	.25	INCH	5J868A	2	1
2	RUDDER	TAP	6	HOLES	RIVET	10	CM	KL2301	3	1
									4	2

4: Supervivencia - Datos de Productos



The IBM WebSphere Information Integration Platform

Service-Oriented Architecture

Understand



Discover, define, model and govern information quality and structure

Cleanse



Standardize, merge, and correct information

Transform



Transform and enrich information

Federate



Virtualize access to disparate information

Integrated Metadata Management

Parallel Processing

Data



Connect



Content

Access, publish, and replicate information



Thank
YOU

