

# InfoSphere™



## Proof Of Technology IBM InfoSphere Information Analyzer

*13 de Diciembre de 2012*

*Javier Gómez Kayser*  
*Javier.gkayser@es.ibm.com*

# Agenda

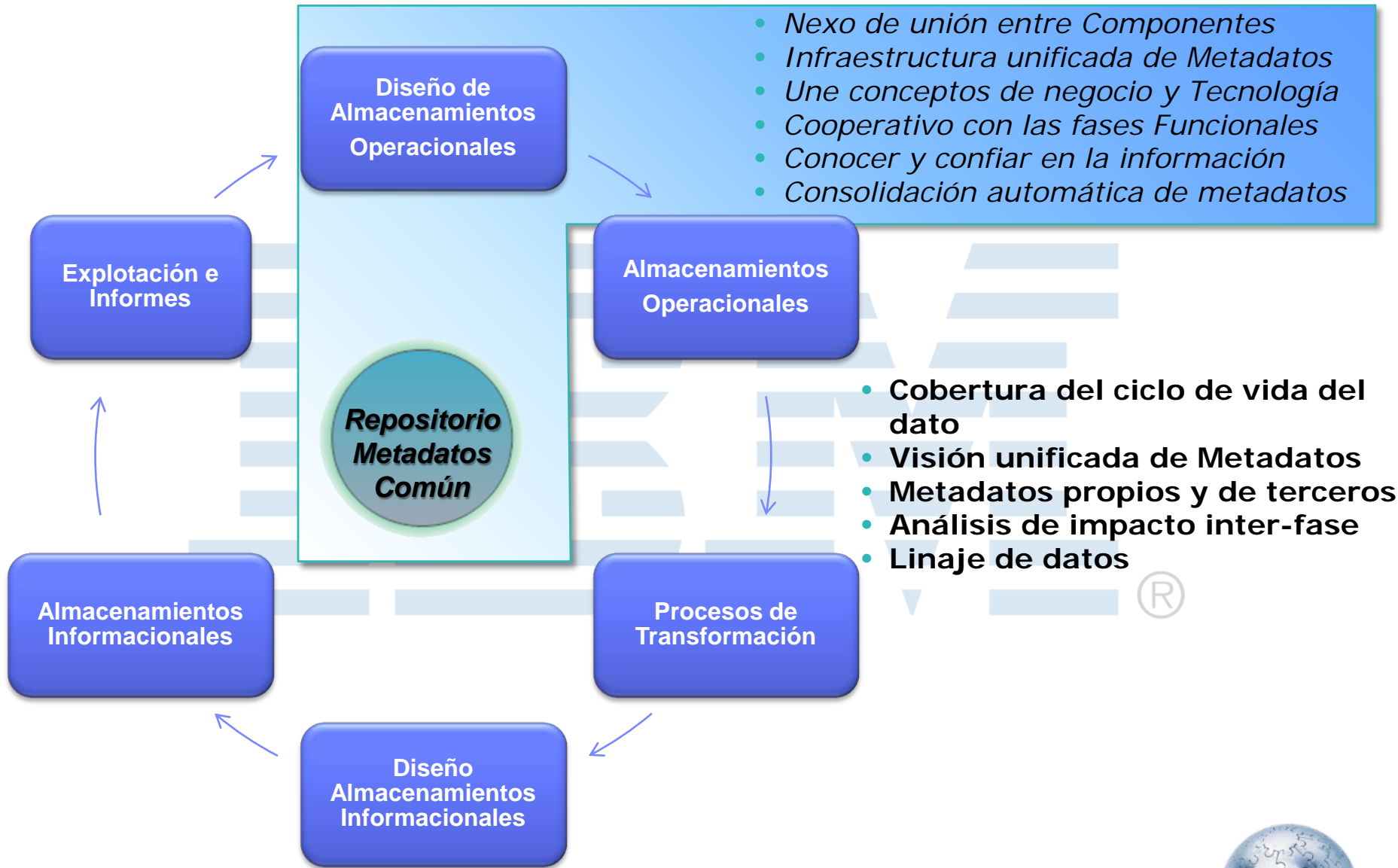
- **Visión General de IBM InfoSphere Information Server**
- **Funcionalidades de Information Analyzer**
- *Descanso – Café*
- **Configuración e importación de los datos**
- **Análisis de columnas**
- *Comida*
- **Identificación de claves y análisis de relaciones**
- **Integridad referencial**
- **Análisis de base**
- **Reglas de negocio y métricas**
- **Generación de informes**



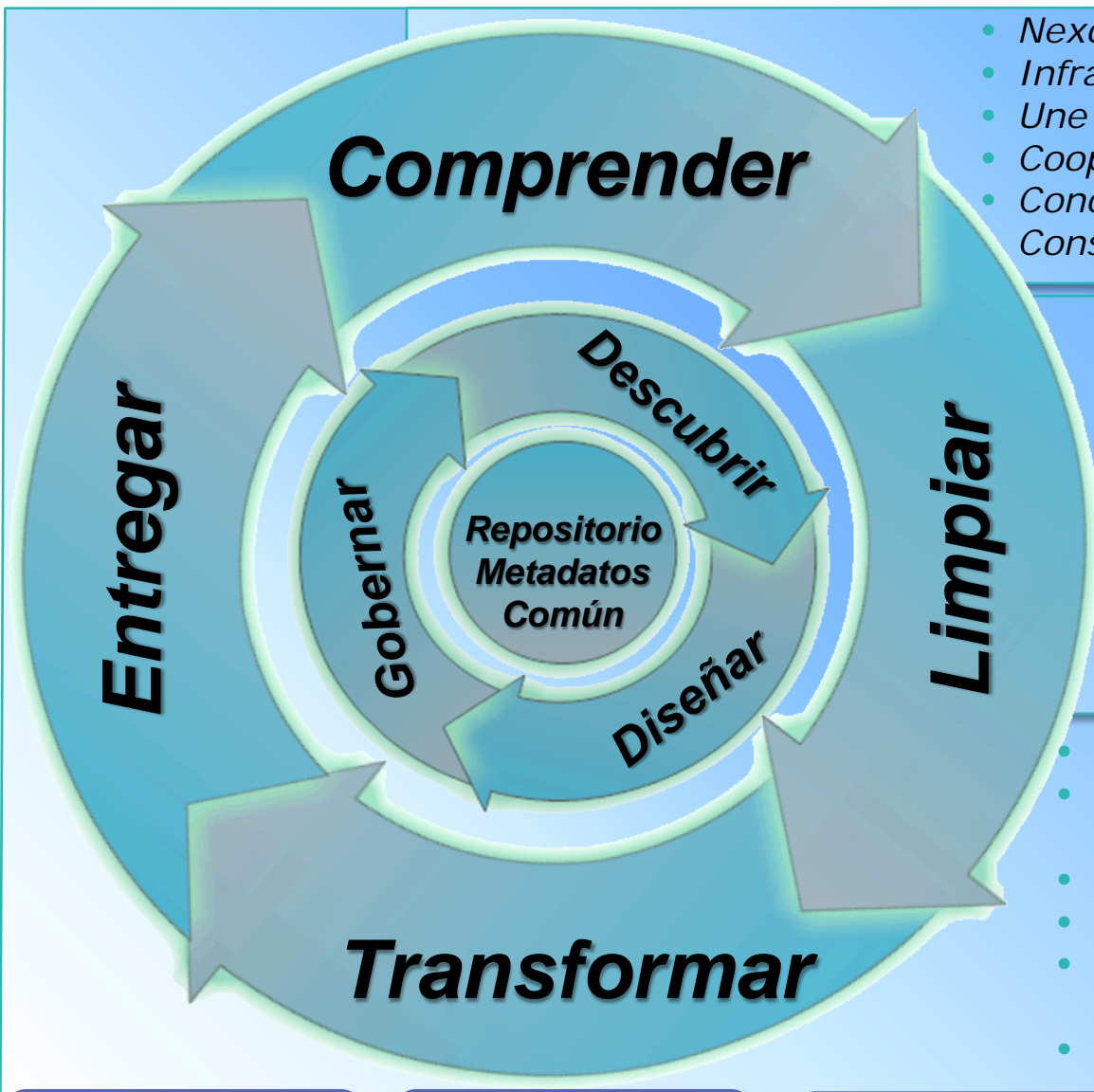
# Visión General de IBM InfoSphere Information Server

- **Plataforma de integración de datos, permite comprender, depurar, transformar y ofrecer información fiable al resto del Negocio**
  - Completa integración de metadatos
  - Unión entre objetivos empresariales e IT
  - Conversión de Datos en Información
  - Linaje de datos
  - Calidad de datos
  - Integración de datos operativos
  - Escalabilidad lineal
  - Optimización de la infraestructura
  - Conectividad para múltiples los orígenes de datos
  - Herramientas de productividad
  - Multi-idioma: francés, coreano, chino, español, portugués de Brasil, alemán, japonés, inglés e italiano









- *Nexo de unión entre Componentes*
- *Infraestructura unificada de Metadatos*
- *Une conceptos de negocio y Tecnología*
- *Cooperativo con las fases Funcionales*
- *Conocer y confiar en la información*
- *Consolidación automática de metadatos*

- *Acortar los tiempos de ciclo de desarrollo*
- *Aumentar la eficiencia operativa*
- *Eliminar y simplificar los procesos duplicados*
- *Colaboración del equipo*
- *Respuesta rápida al mercado y al cliente*
- *Minimizar los riesgos del proyecto*

- *Tratamiento Completo de la calidad*
- *Potente herramienta de procesos de transformación*
- *Tratamiento de On-Line y Batch*
- *Simplificación de la Conectividad*
- *Gran rendimiento en Desarrollo y Producción*
- *Reducido coste de la propiedad*

Procesamiento en Paralelo

Servicios de Conectividad

Servicios de Administración

Servicios de Implementación



## Componentes de Information Server Confiando en la entrega de la Información

### IBM Information Server

#### Despliegue Unificado

**Comprender**



Descubrir, modelizar, y

**Limpiar**



Estandarizar y Unificar la

**Transformar**



Combinar y

**Federar**



Sincronizar, virtualizar y

#### Platform Services

**Servicios de  
Procesamiento  
Paralelo**



**Servicios de  
Conectividad**



**Servicios de  
Metadatos**



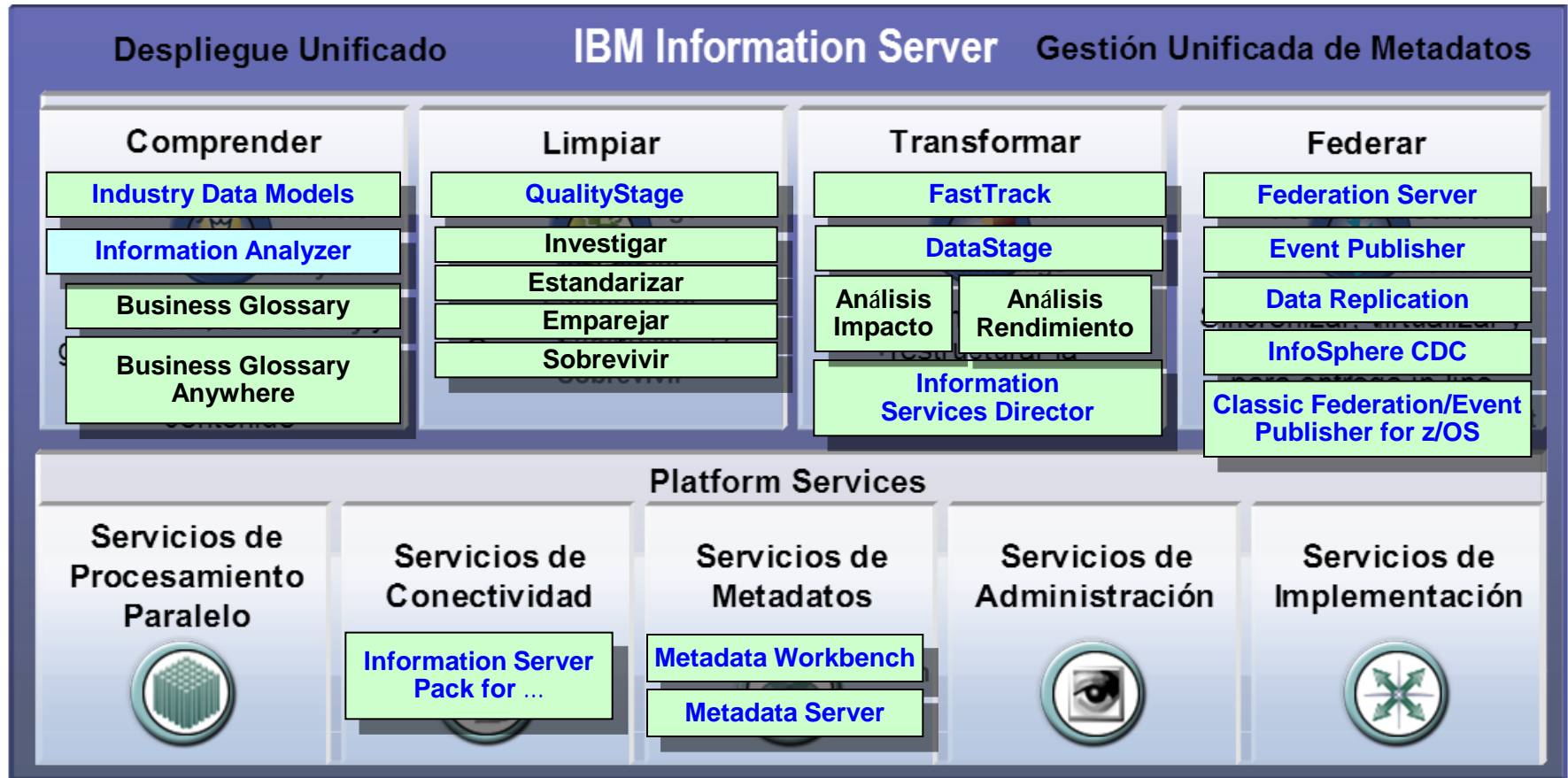
**Servicios de  
Administración**



**Servicios de  
Implementación**



## La Solución IBM : IBM Information Server Confiando en la entrega de la Información



Procesamiento en Paralelo

Conectividad a Aplicaciones, Datos y Contenidos

## Metadatos físicos: InfoSphere Information Analyzer

- Análisis de la estructura lógica de los datos basada en su contenido
- Cartografía detallada del uso de los almacenamientos (campos)
- Relaciones entre campos y entre fuentes
- Creación de metadatos a partir de los resultados del profiling
- Capacidades
  - Análisis de contenidos
    - Columnas
    - Claves primarias
    - Claves Foraneas
    - Datos duplicados
  - Reglas de Negocio
    - Creación y ejecución de Reglas Multi-Nivel (regla, registro y patron)
    - Planificador
    - Informes de resultado

**Analistas de negocio**

**Analistas de datos**

**Comprender**

**InfoSphere Information Analyzer**  
**Analiza estructuras de fuentes de datos, validando reglas de integración y de calidad**

Frequency Distribution | Data Class | Properties | Domain & Consistency | Format | Pattern

Required Review Not Complete | Reviewed

Data Type  
 Original: MarChar | Inferred: MarChar | New: Select...

View Summary

**Inferred Summary:**

- Integer - 25%
- Decimal - 50%
- Char - 12.5%
- Big int - 12.5%

**Inferred Frequency Distribution**

Data Value	Data Type	#	%
efdvce	Char	45	1.00
vfve	Char	364	8.53
evb	Char	769	17.09
afvefng	Char	444	9.87
vfrefvev	Char	252	5.60
dfdf	Char	444	9.87
fdf	Char	252	5.60

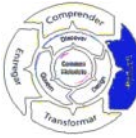
Length  
 Precision  
 Scale  
 Nullability

**Vista física**



**Detalles**





## Limpieza de datos: InfoSphere QualityStage

- Tratamiento de cualquier atributo
- Estudio de Patrones de Escritura
- Validación de datos en base a diccionario
  - Re-alimentación del diccionario
  - Diccionarios pre-construidos
  - Creación de nuevos diccionarios y conceptos
- Normaliza la “escritura” de los datos (Pepe=José, Glez=Gonzalez,...)
- Estandarización de fuentes en cualquier formato de salida
  - Definición a medida de formatos de estandarización
  - Enriquecimiento de la información
  - Tratamiento de la información no contemplada
- Agrupación en base a lógica difusa y probabilística
  - Duplicados
  - Colectivos
  - Casación de registros
- Tablas de Relación
- Obtención del “Mejor Registro”





## Tratamiento de Información: Address Doctor

- Tratamiento, corrección de direcciones postales.
- Tratamiento de direcciones de 240 Países
- Elemento participativo en QualityStage/DataStage
- Tres formas de funcionamiento
  - Sugerencia
  - Corrección
  - Validación





## Transformación y Movimiento: InfoSphere DataStage



- **Diseño visual**, sin código, de flujos de información, con cientos de funciones predefinidas
- **Reutilización** optimizada de los objetos de integración de datos
- Motor orientado al **Rendimiento**
  - Procesamiento paralelo sin cambios en el diseño
  - Soporte de operaciones batch y en tiempo real
- **Reducción de Coste de la Propiedad**
  - Análisis de Impacto
  - Análisis de Rendimiento
  - Generación de Documentación



Developers



Architects

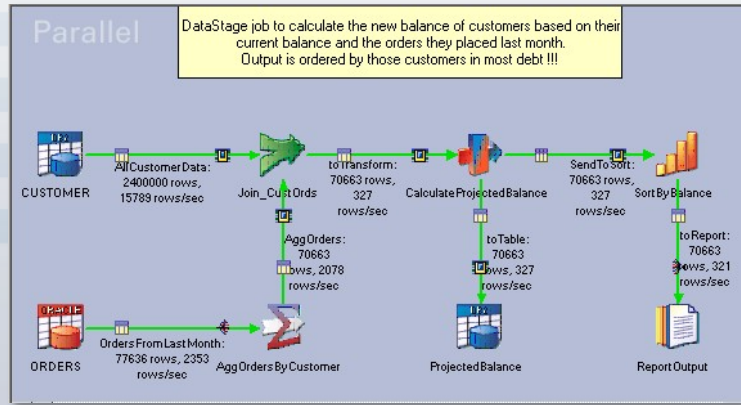
Transformar



Entregar



**InfoSphere DataStage®**  
 Transforma y manipula cualquier volumen de información tanto en batch como en tiempo real mediante una lógica diseñada gráficamente



**Diseño gráfico de un proceso**



Detalles



## Compartir Procesos: Information Services Director



- Automatiza la publicación de Objetos de Information Server como Servicios Web
- Controla la invocación de los Servicios Publicados
- Permite el balanceo de Carga y asegura la posibilidad de invocación
- Se pueden establecer Reglas de Integración tanto para procesos Batch como para tiempo real



Developers



Architects

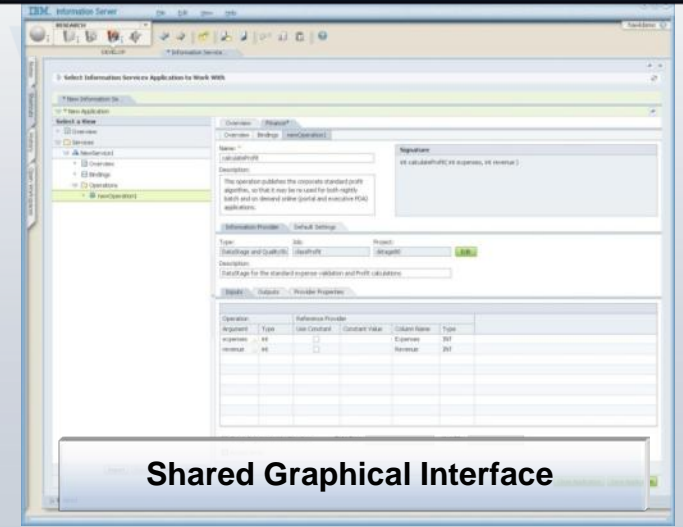
Transformar



Entregar



**InfoSphere Information Services Director®**  
 Universaliza los procesos realizados mediante InformationServer ofreciéndolos como Web services



Shared Graphical Interface



Detalles



## Transformación y movimiento: InfoSphere FastTrack



- Características

- Dirigido a usuarios no informáticos
- Centraliza la gestión y compartición de requerimientos de desarrollo
- Captura reglas del negocio en un formato común para compartición
- Algoritmos de descubrimiento de emparejamientos de atributos y relaciones entre tablas
- Uso de interfaces similares a Excel

- Beneficios

- Línea de Comunicación Negocio/TI
- Minimiza el riesgo de mala interpretación de requerimientos
- Maximiza la productividad del analista mediante un interface conocido



Transformar



Source Columns	Source Business Term	Target Columns	Target Business Term	Business Rule	Function	Annotation
tablah tablah BANK CUSTOMERS NA_ARE	Customer	tablah tablah BANK MASTER FIRST NA, IE, tablah tablah BANK MASTER ANNO, C_NAME, tablah tablah BANK MASTER L_AFT_NAME	High Value Customer	Standarize with USAREA.SET relevant Use first character of last name as blocking column for match. First name and last name are very important for match.		
tablah tablah BANK CUSTOMERS ADD (R) tablah tablah BANK CUSTOMERS ADDRESS	Street_Addr	tablah tablah BANK MASTER HOUSE, ST, RESET, NAME, tablah tablah BANK MASTER STR, STREET, SUITE	Street	Standarize with USADDR.SET relevant Use first character of house number as blocking column in match. Street name very important.		
tablah tablah BANK CUSTOMERS CITY, tablah tablah BANK CUSTOMERS STATE	City	tablah tablah BANK MASTER CITY, State	City	Standarize with USAREA.SET. Standarize with USAREA.SET. Add first		

IBM Information Server FastTrack

Mapping Specifications

name	Target	Created	Modified	Created By	Status	Pre-Load Reviewed	Tools
DDI_Bank2Extract	CUSTOMERS	1/2/2008 9:36	1/2/2008 9:36	dbjordan	In Progress	0%	Open Projects... New Job Generation Status
Bank2Extract	CUSTOMERS	1/2/2008 9:33	1/2/2008 9:33	dbjordan	In Progress	0%	New Mapping Specification Project Run CUIV
Bank2Extract	CUSTOMERS	1/2/2008 9:33	1/2/2008 9:33	dbjordan	In Progress	0%	Open
ClientName Address	MARKETS	1/2/2008 9:34	1/2/2008 9:34	dbjordan	In Progress	0%	Generate Job Delete Public Relationships Support Cardshare Table DDL
Identify High Value Customer	MARKETING_C	1/2/2008 9:35	1/2/2008 9:35	dbjordan	In Progress	0%	

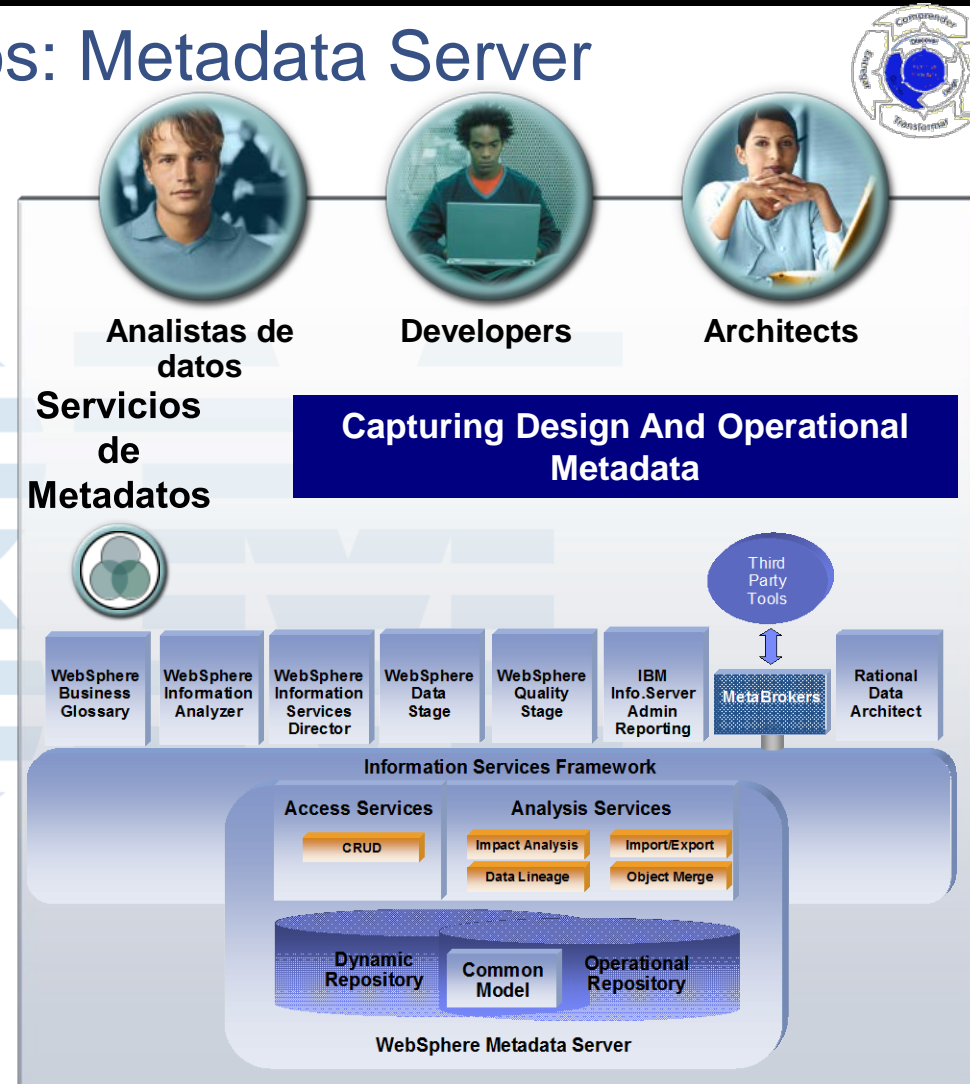
Mapping Editor

Source	Target	Transformation	Reviewed
CUSTOMERS.PHONE_CUSTOM	CUSTOMERS.NAME	To Name (NDC)	
CUSTOMERS.ADDR1	CUSTOMERS.ADDR1	Concat first name and last name	
CUSTOMERS.CITY	CUSTOMERS.CITY		
CUSTOMERS.STATE	CUSTOMERS.STATE		
CUSTOMERS.ZIP	CUSTOMERS.ZIP		
ACCOUNTS.CUSTOMER_ID_A	CUSTOMERS.CUSTOMER_ID_C	Aggregate account balance over se	
CUSTOMERS.BR_NUM	CUSTOMERS.TAX_ID		
CUSTOMERS.YEARS_CLOSED	CUSTOMERS.YEARS_CLOSED		
INVOICES.ONLINE_ACCESS	CUSTOMERS.ONLINE_ACCESS		
INVOICES.ONLINE_ACCESS	CUSTOMERS.LEVELS		



## Conocimiento sobre los datos: Metadata Server

- Infraestructura unificada para facilitar la compartición del conocimiento sobre la forma de los datos
- Unifica los conceptos sobre el negocio, acercando Tecnología y Negocio
- Reduce los plazos de entrega de proyectos (funcional)
- Mejorar el conocimiento y la confianza en la información entregada a los usuarios
- Elimina la carga manual de metadatos



Detalles





# Incorporación de Metadatos de terceros...



## **Bridge**

CA ERwin Data Modeler 4.x  
 CA ERwin Data Modeler 7.x  
 categories and Term Metabroker  
 DBM File Metabroker  
 Embarcadero ER/Studio Business Architect 1.5  
 ER/Studio Data Architect 5.1 to 8.5  
 IBM Cognos BI Reporting - Content Manager RN  
 v1.1, C8 v8.0 to 8.4, and v10  
 IBM Cognos BI Reporting - Content Manager  
 Packages RN v1.1, C8 v8.0 to 8.4, and v10  
 IBM Cognos BI Reporting - Content Manager  
 ReportStudio RN v1.1, C8 v8.0 to 8.4, and v10  
 IBM DB2 Warehouse Manager 7.2  
 IBM InfoSphere Data Architect Metabroker  
 IBM InfoSphere Warehouse / Cubing Services for  
 OLAP 9.x  
 Microsoft SQL Server Data Source View (DSV) 9.0  
 (2005) to 10.0 (2008)  
 MicroStrategy Intelligence Server 7.0 to 9.0  
 Object Management Group Common Warehouse  
 Metamodel (CWM) XMI 1.0 to 1.1

## **Bridge**

ODBC Metrabroker  
 Oracle Hyperion Application Builder N/A  
 Oracle Warehouse Builder (OWB) 9i  
 SAP BusinessObjects (BO) Designer 5.1.5 to 12.x  
 SAP BusinessObjects (BO) Desktop Intelligence  
 5.1.5 to 12.x  
 SAP BusinessObjects (BO) Repository 11.x to 12.x  
 SAP BusinessObjects (BO) Web Intelligence 11.x to  
 12.x  
 SAS Data Integration Studio 9.x  
 Sybase PowerDesigner PDM (Physical Data  
 Modeling) 6.1.x  
 PowerDesigner PDM (Physical Data Modeling) 7.5 to  
 15.x  
 User Defined Metabroker  
 Altova XML Spy 2004 to 2007  
 Informatica Metadata Manager (MM) 8.x  
 Meta Integration Technology, Inc. Meta Integration  
 Repository (MIR) XMI file 6.x  
 World Wide Web Consortium XML Schema 1.0  
 (<http://www.w3.org/XML/Schema>)



## Metadatos de negocio: InfoSphere Business Glossary



- Creación, gestión e intercambio de metadatos de negocio
- Alinea los esfuerzos de TI con los objetivos de negocio
- Proporciona contexto de negocio a las áreas técnicas
- Completa API "REST"
  - Expone el contenido glosario como recursos Web (URLs)

Database = DB2

Schema = NAACCT

Table = DLYTRANS

Column = ACCT\_NO

data type = char(11)



Technical



Business

GL Account Number

The ten digit account number. Sometimes referred to as the account ID. This value is of the form L-FIIIVVVV.



Analistas de negocio



Usuarios de negocio

Comprender



InfoSphere Business Glossary

Crea y gestiona vocabulario de negocio y relaciones, enlazándolo con los metadatos físicos



Detalles







## Business Glossary Anywhere

Acceso al “Business Glosary” desde cualquier Aplicación



ANY User

### Características

- Desde cualquier aplicación, Click en un termino, y verá su definición de negocio, y sin perder el contexto ni el foco
- Las búsquedas Inteligente mostrarán los mejores candidatos en un intento
- Motor de búsqueda de términos y categorías
- Acceso guiado a la información de contacto
- Seguridad mantenida a través de Information Server

### Beneficios

- Aumenta la confianza y aceptación de la información con el uso de definiciones en el contexto
- Expande la adopción del “Business Glosary” fuera de Tecnología

**Comprender**

Desde Cualquier Aplicación

¡Aparece la Información!





## La Forma de los datos: Metadata Workbench

- Módulo para revisar y administrar activos de Information Server
- Explora, analiza y gestiona metadatos, de forma gráfica o textual
- Informes Claros sobre la herencia (lineage), de los datos para verificar el cumplimiento de reglas de Negocio
- Evaluar el impacto ante cambios, tanto sobre los objetos de Information Server como herramientas de terceros
- Ayuda al cumplimiento de normativas Internacionales (Basilea II, Solvencia II, arbanes-Oxley)



IT Developers  
Administrators



Project Managers  
& DBAs

Servicios  
de  
Metadatos



Exploración mediante Web de los activos generados o usados por Information Server

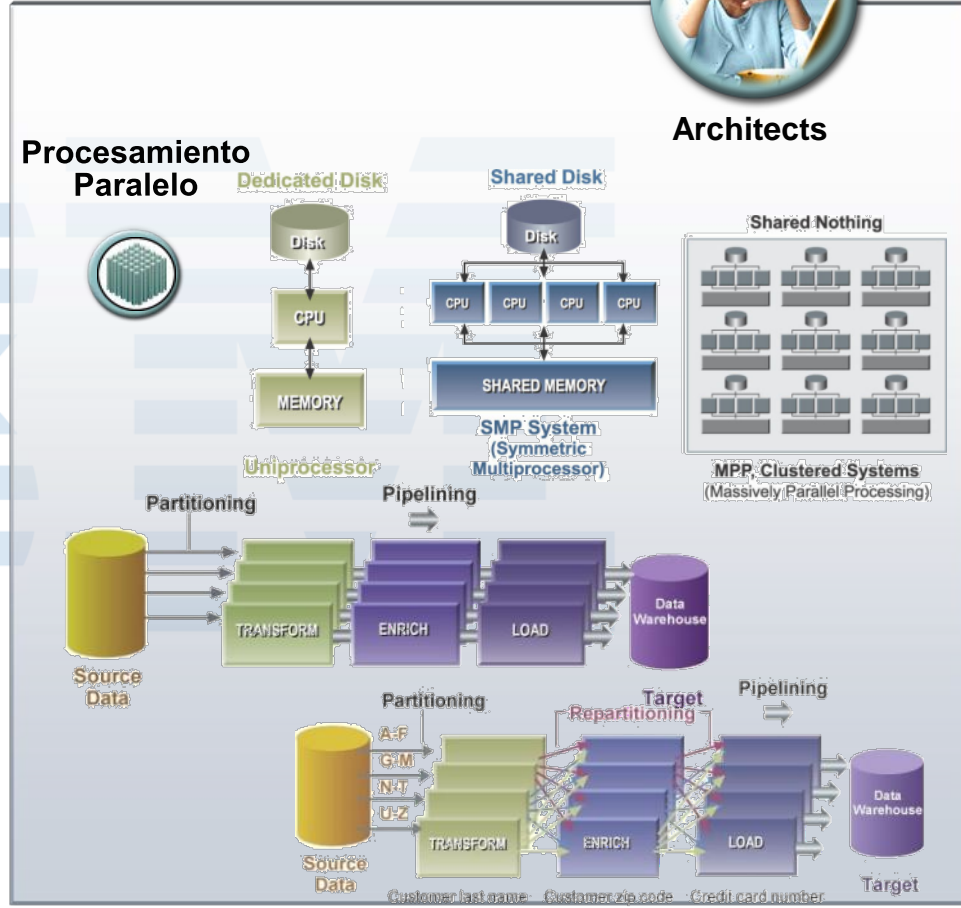


Detalles

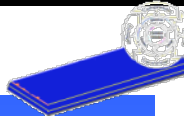


## Escalabilidad Ilimitada: Paralelismo

- Atributos
  - Paralelismo mediante Pipeline. Clonación automática de los procesos y sus sub-tareas por nodo de ejecución
  - Paralelismo Particionado. Repartición de las tareas de un proceso a través de los nodos
  - Paralelismo Dinámico “re-partitioning” (on the fly). Clonación automática de las tareas de los procesos por hebra de nodo y asignación de los datos en base a la carga de trabajo de cada tarea
  - Soporte de paralelismo en “grid”. Control de los distintos escenarios en máquinas remotas



## Conectividad a fuentes y destinos de Datos



**RDBMS**  
DB2 (on Z, I, P or X series)  
Oracle  
Informix (IDS and XPS)  
Ingres  
Netezza  
Progress  
RDB  
RedBrick  
SQL/DS  
SQL Server  
Sybase (ASE & IQ)  
Teradata  
Universe  
UniData  
NonStop SQL  
InfoSphere Federation Server  
InfoSphere Classic Federation  
And more.....

**General Access**  
Sequential File  
Complex Flat File  
File Set  
Data Set  
Named Pipe  
iWay  
FTP  
SFTP  
Compressed / Encoded Data  
External Command Call  
Parallel/wrapped 3<sup>rd</sup> party apps  
EMC InfoMover  
Web logs  
Email

**Standards & Real Time**  
InfoSphere MQ  
Java Messaging Services (JMS)  
Java  
XML & XSL-T  
EBXML  
Web Services (SOAP)  
EJB (Enterprise Java Beans)  
EDI  
...

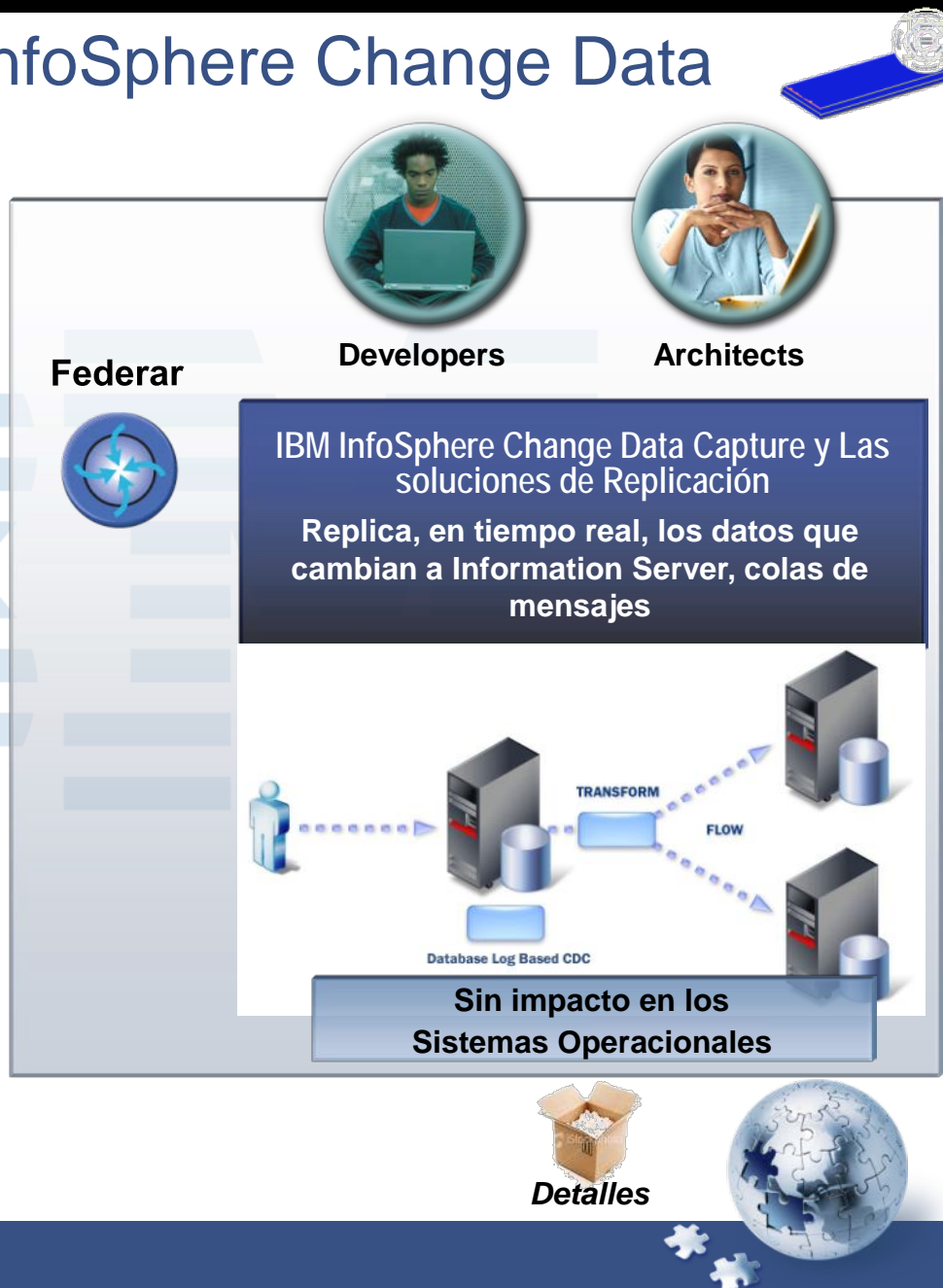
**CDC**  
DB2 (on Z, I, P, X series)  
Oracle  
SQL Server  
Sybase  
Informix  
IMS  
VSAM  
ADABAS ...

**Legacy**  
ADABAS  
C-ISAM  
D-ISAM  
Datacom/DB  
DS Mumps  
Enscribe  
Essbase  
FOCUS  
IDMS/SQL  
ImageSQL  
Infoman  
KSAM  
M204  
MS Analysis  
RMS S2000  
Supra  
TOTAL  
TurboImage  
Unify  
Y ¡Muchos más!....



## Integración en tiempo real: InfoSphere Change Data Capture

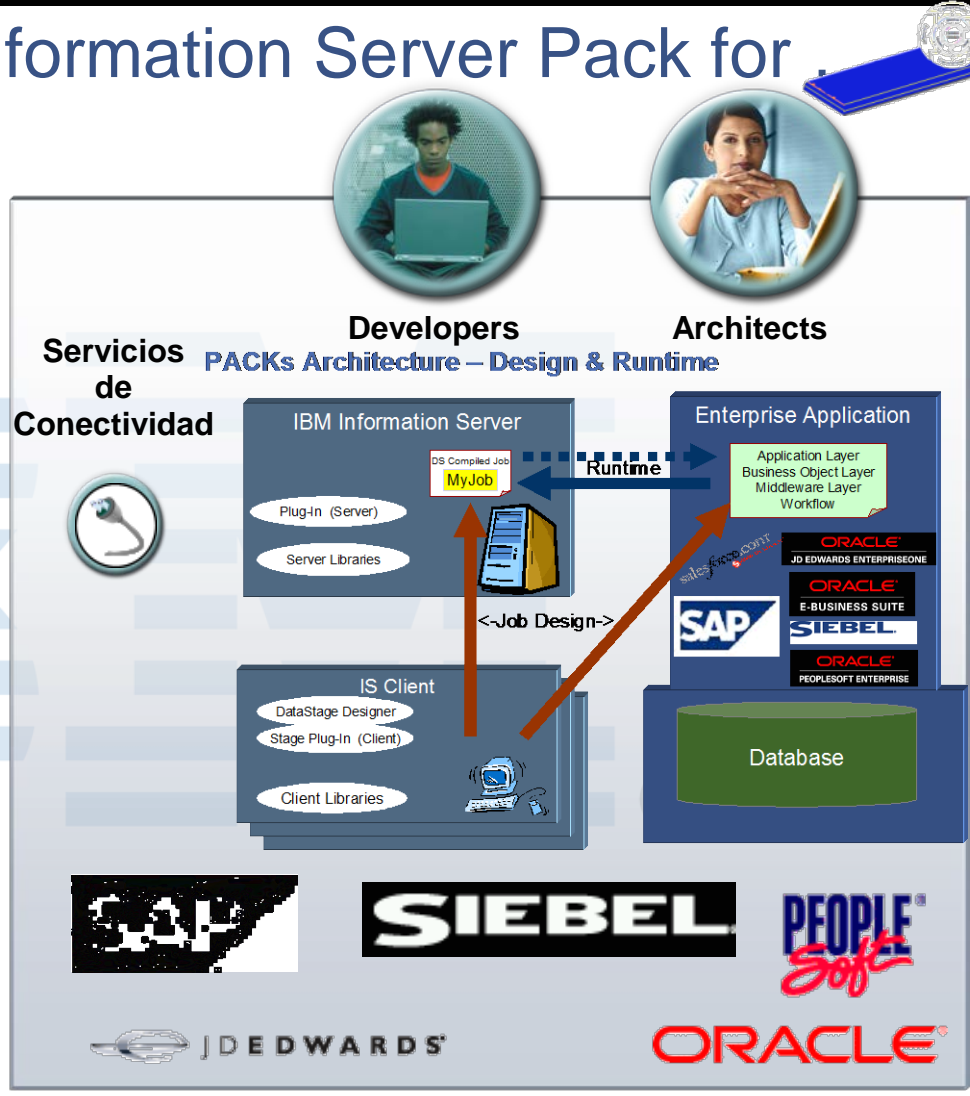
- Captura de los cambios en los datos en tiempo real
- Integración Dinámica
- Sincronización
- Replicación
- Mínimo impacto en los sistemas Operacionales
- Gran escalabilidad y rendimiento
- Garantía de integridad en los datos





## Servicios de Conectividad: Information Server Pack for

- SAP
  - Load PACK for SAP BW
  - Extract PACK for SAP R/3
- Siebel
  - Siebel 1999-20xx
- JDE/PeopleSoft OneWorld
- Oracle Applications
  - Oracle E-Business Suite 11i
- Salesforce.com
- SAS
- Ariba
- Manugistics
- I2
- Etc...





## Punto único de Control: Servicios de Administración

- Punto común Administrador de licencias y Usuarios (IBM Information Server Web Console)
- Control sobre licencias activas
- Consulta a “logs” de producto
- Roles de Usuario por Módulo de Information Server, y/o a nivel de proyecto
- Usuarios proporcionados por LDAP, Active Directory, Was o Sistema Operativo
- Entorno de trabajo Común (IBM Information Server Console)

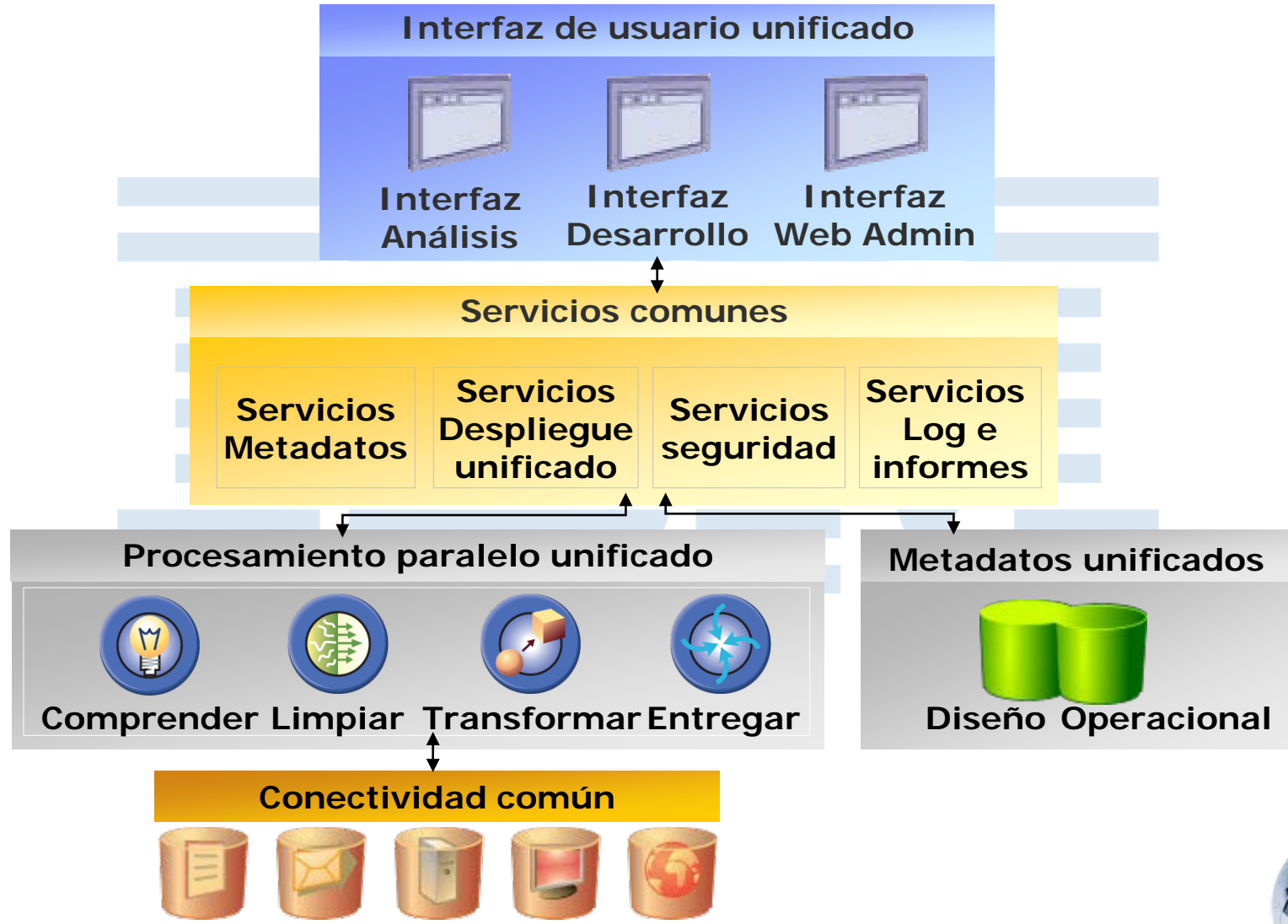
**IT Developers Administrators**

**Project Managers & DBAs**

**Servicios de Administración**



## Arquitectura de IBM Information Server



Estructurado, Desestructurado, Aplicaciones, Mainframe



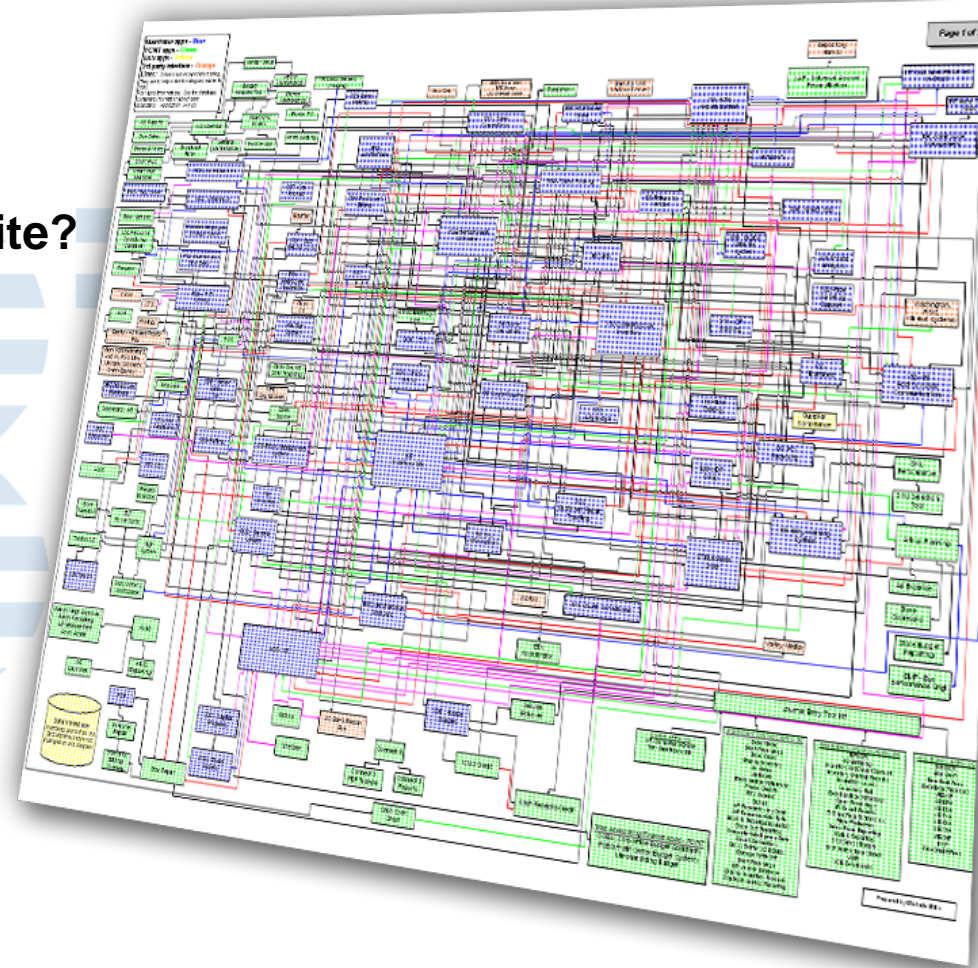
## Agenda

- **Visión General de IBM InfoSphere Information Server**
- **Funcionalidades de Information Analyzer**
- *Descanso – Café*
- **Configuración e importación de los datos**
- **Análisis de columnas**
- *Comida*
- **Identificación de claves y análisis de relaciones**
- **Integridad referencial**
- **Análisis de base**
- **Reglas de negocio y métricas**
- **Generación de informes**



# El Reto de Información

- ¿Dónde está mi información?
- ¿Cómo la consigo cuando la necesite?
- ¿Qué quiere decir?
- ¿Es fiable?
- ¿Cómo la obtengo en el formato requerido?
- ¿Cómo la envío?
- ¿Cómo la controlo?



## Problemas Comunes de los Datos

- Falta de Información estándar
  - Diferentes formatos y estructuras en los diferentes sistemas
- Datos imprevistos en silos aislados
  - Datos perdidos en las bases de datos
- Información enterrada en formularios
- Miopía de Datos
  - La falta de identificadores coherentes impiden tener una única vista
- La pesadilla de la redundancia
  - Registros duplicados y falta de estándares

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





## El dominio del conocimiento cambia constantemente

- El Conocimiento es crítico para comprender los datos, el problema e interpretar los resultados
  - “Reinicia a 0 si el Contador de Llamadas supera N”.
  - “Los valores ocultos se representan con 0, pero el importe facturado por defecto es 0”
- Falta de conocimiento y comprensión es la causa primaria de la pobre calidad de los datos – los datos se vuelven inutilizables
- La Información está en las cabezas de la gente – raramente documentado
- Fragmentado entre organizaciones
- Perdido en cambios de proyectos y de personal
- Si no se documenta, con el tiempo el conocimiento se deteriora y se confunde



Desarrollador



Usuario final



Analista de negocio



Arquitecto



Arquitecto de SW



IT Admin



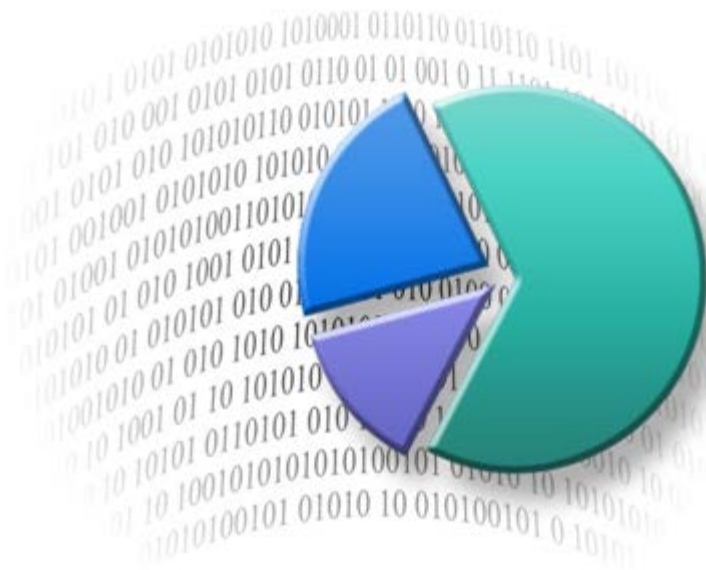
Admin de datos





## ¿Por qué Data Profiling?

- Problemas críticos:
  - No saber que datos están realmente en los sistemas
  - Las fuentes han cambiado, o son nuevas o desconocidas
- ¿Por qué?
  - Los valores de los datos y sus relaciones son diferentes de las reglas documentadas
  - Documentación incompleta o desaparecida
  - Las fuentes de datos nunca son estáticas y frecuentemente cambian sin previo aviso
- Enfoque alternativo
  - Procesos intensivos que requieren muchos recursos
  - Nunca revisa el 100% de los elementos de los datos
  - Sin infraestructura para soportar el mantenimiento
  - No hay enfoques estandarizados para los proyectos
  - La primera generación de herramientas no se ocupaban de la resolución de problemas



# IBM InfoSphere Information Analyzer

- ¿Qué es?
  - Herramienta de análisis y perfilado de datos para fuentes de datos heterogéneas.
    - Integra capacidades de 'profiling' obteniendo tres productos diferentes
      - Tablas Auxiliares
      - Punto de referencia de Acciones
      - Documentación y Metadatos Detallados
- ¿Qué puede hacer?
  - Analiza fuentes de datos para descubrir su estructura, contenido y calidad de la información
  - Deduce la realidad del dato, no únicamente su definición
  - Encuentra e informa sobre los datos desaparecidos, incorrectos e inconsistentes
- ¿Quién lo utiliza?
  - Analistas de datos y de negocio, Especialistas en Calidad de Datos, Arquitectos,.....



# Data Profiling: IBM Information Analyzer

## Data Sources



ERP from acquisition



Mainframe manufacturing system



Parts BOM



External Lists



Distribution



Demographic



Contact



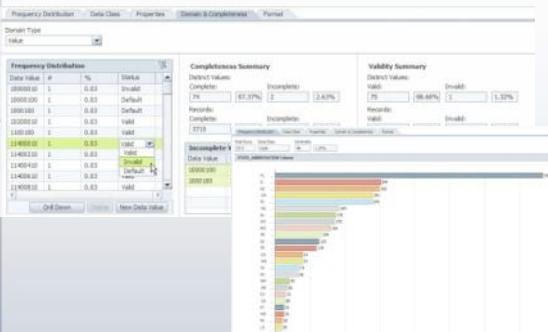
Billing / Accounts

- Automatiza los procesos de descubrimiento de datos
- Permite comprender los datos antes de comenzar a desarrollar
- Elimina el riesgo y la incertidumbre de usar datos no válidos
- Útil para cualquier tipo de proyectos de migración de datos
- Analiza cada atributo del dato y descubrir su metadato real
- Reduce el tiempo de análisis de datos más del 50%



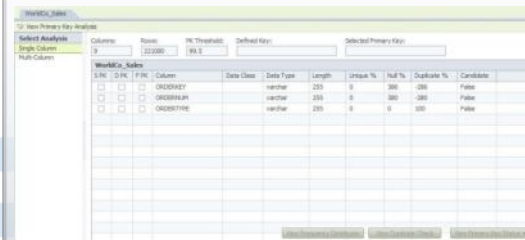
## Funcionalidades de Information Analyzer

### Análisis de Columnas



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

### Análisis de Tablas



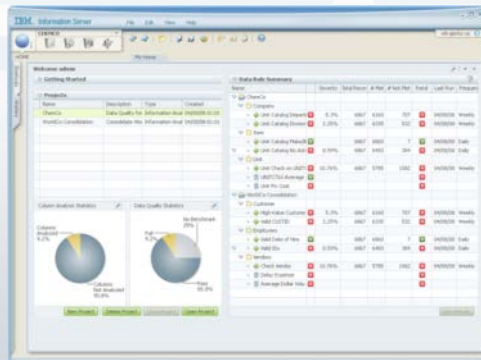
Análisis de claves Primarias Simples o Múltiples

### Análisis Cruzados

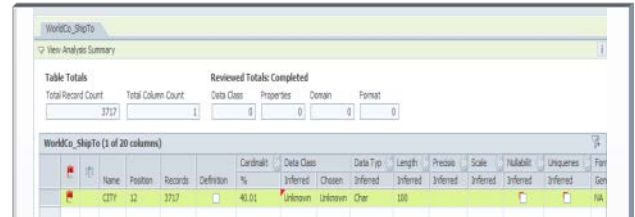


Foreign Key & Análisis multitabla

### Reglas de Negocio



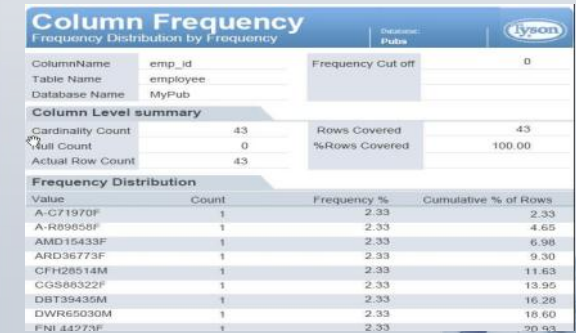
Verifica cumplimiento de reglas criticas



Anotaciones, marcas para revisión,...

Comparativas en el tiempo

Generación Tablas Auxiliares



Generación de Informes

## Agenda

- **Visión General de IBM InfoSphere Information Server**
- **Funcionalidades de Information Analyzer**
- *Descanso – Café*
- **Configuración e importación de los datos**
- **Análisis de columnas**
- *Comida*
- **Identificación de claves y análisis de relaciones**
- **Integridad referencial**
- **Análisis de base**
- **Reglas de negocio y métricas**
- **Generación de informes**

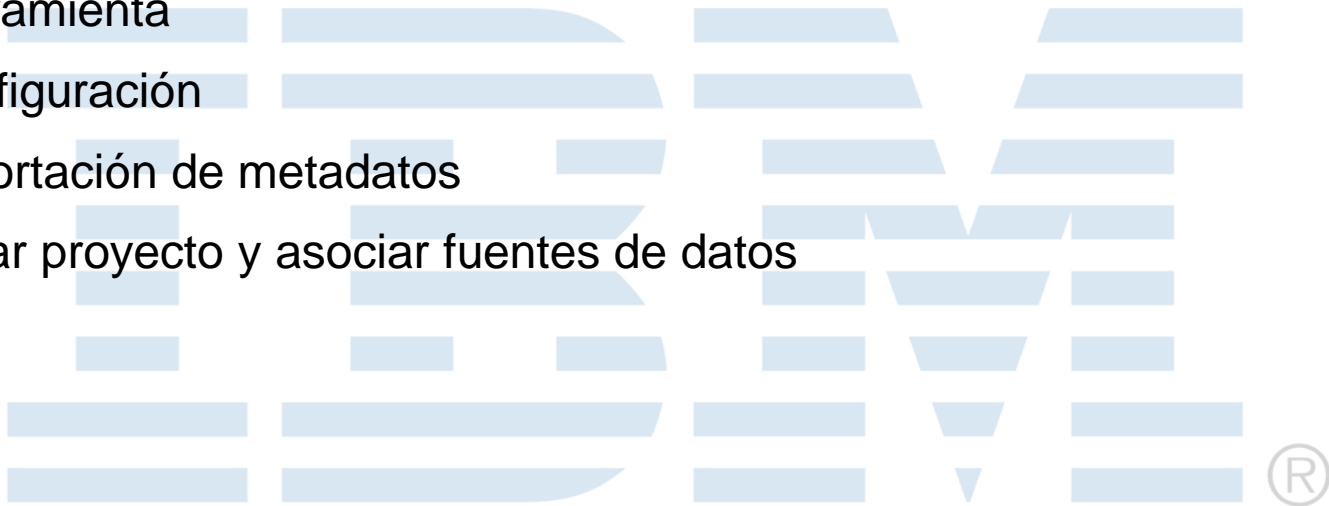




# Ejercicio 1: Configuración e importación de los datos

## LAB 1: PREPARAR EL ENTORNO DE INFORMATION ANALYZER

- Herramienta
- Configuración
- Importación de metadatos
- Crear proyecto y asociar fuentes de datos



## Agenda

- **Visión General de IBM InfoSphere Information Server**
- **Funcionalidades de Information Analyzer**
- *Descanso – Café*
- **Configuración e importación de los datos**
- **Análisis de columnas**
- *Comida*
- **Identificación de claves y análisis de relaciones**
- **Integridad referencial**
- **Análisis de base**
- **Reglas de negocio y métricas**
- **Generación de informes**



## Data Profiling: Análisis de Columnas



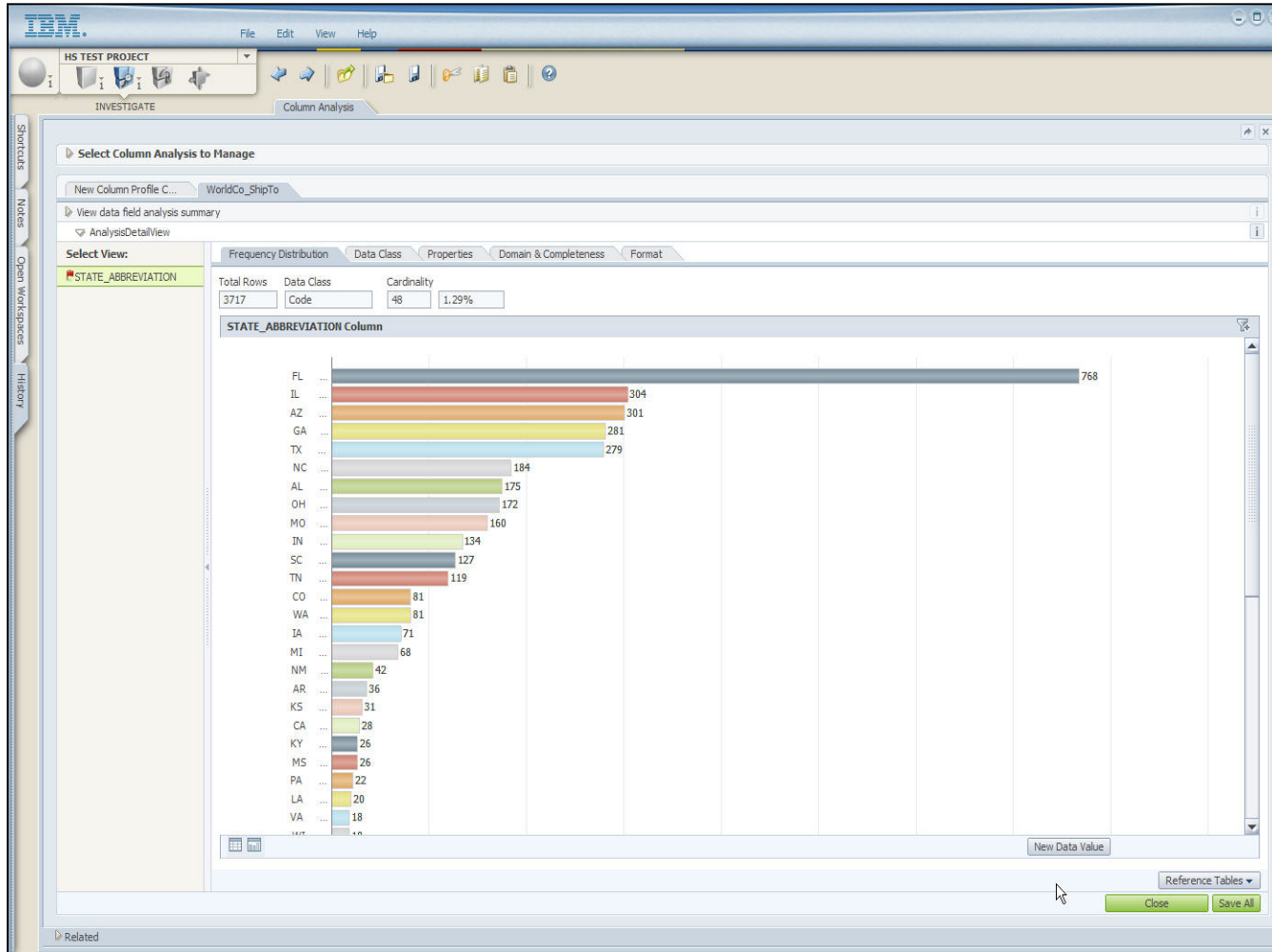
- Valores de dominio y Validación
- Clasificación de datos
- Propiedades de los datos
- Formatos

The screenshot shows the IBM Information Server interface for Column Analysis. The main window displays the 'QTYORD' column analysis for the 'GlobalCo\_Ord\_Dtl' table. The 'View Details' section is active, showing a table with columns for Data Value, Frequency, Value Flag, Data Type, Length, Format, Transform, and Value. The table lists 11 data values from 0 to 11, with their respective frequencies and percentages. The 'Value' column shows the definition and source of each value.

Data Value	#	%	Value Flag	Data Type	Length	Format	Transform	Value
0	76	1.19	Valid	DFLOAT	1	9		Data Numeric zero
1	384	6.01	Valid	DFLOAT	1	9		Data Data
2	314	4.92	Valid	DFLOAT	1	9		Data Data
3	316	4.95	Valid	DFLOAT	1	9		Data Data
4	254	3.98	Valid	DFLOAT	1	9		Data Data
5	447	7	Valid	DFLOAT	1	9		Data Data
6	442	6.92	Valid	DFLOAT	1	9		Data Data
7	287	4.49	Valid	DFLOAT	1	9		Data Data
8	415	6.5	Valid	DFLOAT	1	9		Data Data
9	348	5.45	Valid	DFLOAT	1	9		Data Data
10	223	3.49	Valid	DFLOAT	2	99		Data Data
11	31	0.49	Valid	DFLOAT	2	99		Data Data



# Resumen y detalle de Columna



- Completo análisis de volumen de fuentes de datos al nivel de columna
- Frecuencia de valores y modelos de formato
- Propiedades de datos basada en el contenido actual de los datos
- Soporte a columnas virtuales para extender todos los análisis
- Todos los procesos analíticos pueden ser programados



## Evaluación de la calidad del dato

Frequency Distribution | Data Class | Properties | Domain & Completeness | Format

Domain Type: Value

Data Value	Count	Percent	Status
[NULL]	37	3.6	Valid
* C/O MATRIX INTL	29	2.82	Valid
INTERNAL ACCT ONL	7	0.68	Valid
901 64TH ST NW	5	0.49	Valid
9416 FRONT STREET	5	0.49	Valid
695 SHORES BLVD	4	0.39	Valid
PO BOX 530416	4	0.39	Valid
31408 4TH AVE E	3	0.29	Valid
4295 KEARNEY ST	3	0.29	Valid
1200 WEST HAMBUR	2	0.19	Valid

**Completeness Summary**

Distinct Values: Complete: 925 (100%) Incomplete: 0 (0%)

Records: Complete: 1029 (100%) Incomplete: 0 (0%)

**Incomplete Values**

Data Value	Count	Percent

**Validity Summary**

Distinct Values: Valid: 925 (100%) Invalid: 0 (0%)

Records: Valid: 1029 (100%) Invalid: 0 (0%)

**Invalid Values**

Data Value	Count	Percent

WORLDCO\_BILLTO

View Analysis Summary

View Details

Select View:

Frequency Distribution | Data Class | Properties | Domain & Completeness | Format

Number of Formats: 547

Conforming Count: 1029 | Violation Count: 0

General Format	Count	Percent	Status
* A/A AAAAAA AAAA	29	2.82	Conform
* A/A AAAAAAAAA AAAAAAAAA	1	0.1	Conform
*A/A A & A AAAAAAAAAA	1	0.1	Conform
*A/A AAAAAAAAAA AA AAAAAAAAA	1	0.1	Conform
.	1	0.1	Conform
9 AAA AA	1	0.1	Conform
9 AAAAA AAAA AAAAAAAAA AAAA	1	0.1	Conform
99 A 9AA AAA	1	0.1	Conform
99 A AAAAAA AA	1	0.1	Conform

Distinct Value	Count	Percent
* C/O MATRIX INTL	29	2.8183

Distinct Value	General Format	Count	Percent

- Controles en la validez de los valores
- Formatos de datos y de violaciones de formato
- Genera fácilmente tablas de referencia u omisión, datos validos o inválidos, así como los valores de transformación
- Es aprovechado por DataStage o QualityStage





## Anotaciones

The screenshot shows the IBM Information Server interface. The main window displays a column analysis for the 'STATUS' column. A 'Notes' pane is open on the left, showing a note for 'Valid Status Codes' with the type 'Informational' and status 'Opened'. The note text reads: 'The only valid values for Status are: A = Active, I = Inactive'. The main window shows a table with columns for 'Data Value', 'Frequency', 'Percent', 'Validity', 'Data Type', 'Length', 'Format', 'Transform', 'Definition', 'Source', and 'Type'. The table contains two rows of data for the 'STATUS' column.

Data Value	Count	Percent	Validity	Data Type	Length	Format	Transform	Definition	Source	Type
A	1927	51.84	Valid	STRING	1	A		Data	Data	C
I	1790	48.16	Valid	STRING	1	A		Data	Data	C

- Anotaciones analíticas en cualquier objeto
- Las anotaciones con los valores descritos son presentadas en los informes



## Generación de Tablas de Referencia

- Generación de Tablas auxiliares en Base a Contenido

The screenshot shows the 'New Reference Table' dialog box in the IBM InfoSphere Information Server. The 'Name' field is set to 'TABLA\_CODIGOS\_POSTA'. Under 'Select Reference Table Type', the 'Valid' radio button is selected. The 'Definition' field is empty. A 'Preview' button is visible at the bottom.

Field Name	Data Type	Display Length	Precision	Scale
DISTINCTVALUE	VARGRAPHIC	512	512	0

The screenshot shows the 'Browse Reference Table' dialog box in the IBM InfoSphere Information Server. The 'TCODIGOS\_POSTALES' table is selected. The table metadata is displayed below:

Field Name	Data Type	Display Length	Precision	Scale
DISTINCTVALUE	VARGRAPHIC	512	512	0

The dialog also shows a list of distinct values for the 'DISTINCTVALUE' field, with '01193' highlighted. The list includes values: 03008, 01193, 03115, 01430, 02212, 02636, 03050, 03180, 03350, 01006, 03569, 03780, 04028, 01209, 04628, 01475, and 05120.

## Selección de Datos Flexible – Tablas Virtuales

**Factory Workers**

▼ Edit Virtual Table

Name: \*  
Factory Workers

Short Description:  
Factory Workers only

Long Description:  
Bank Account ID and Balance Only for Factory Workers

Base Table:  
BANK\_ACCOUNTS

**Available Columns**

- SS\_NUM
- ADDR1
- ADDR2
- CITY
- STATE
- ZIP
- ZIP\_FOUR
- GENDER
- MARITAL\_STATUS
- PROFESSION
- NBR\_YEARS\_CLI
- SAVINGS\_ACCOUNT

**Selected Columns**

- ACCOUNT\_ID
- RECORD\_ID
- NAME
- BRANCH\_ID
- BANK\_BALANCE

**Define Filter for Selected Columns**

(	Column	Operator	Value	)	AND
(	PROFESSION	=	Factory Worker	)	

SELECT \* FROM BANK\_ACCOUNTS WHERE ( PROFESSION = 'Factory Worker' )

Add Condition Insert Condition Remove Condition

Use Free Form Editor

- Filtro de la fuente de datos para especificar el análisis de necesidades
  - Muestra sobre el original
  - Columnas seleccionadas por el usuario
  - Análisis sobre la tabla virtual <sup>®</sup>



## Ejercicio 2: Análisis de Columnas

### LAB 2: ANÁLISIS DE COLUMNAS CON INFORMATION ANALYZER

- Ejecutar análisis de columnas
- Abrir análisis de columnas
- Ver detalles
- Añadir notas
- Tablas de referencia
- Tablas de gestión
- Tablas virtuales
- Columnas virtuales



## Agenda

- **Visión General de IBM InfoSphere Information Server**
- **Funcionalidades de Information Analyzer**
- *Descanso – Café*
- **Configuración e importación de los datos**
- **Análisis de columnas**
- **Comida**
- **Identificación de claves y análisis de relaciones**
- **Integridad referencial**
- **Análisis de base**
- **Reglas de negocio y métricas**
- **Generación de informes**





## Data Profiling: Análisis de Tablas



- Claves primarias (simples o multicolumnas)
- Claves Duplicadas

The screenshot shows the 'Primary Key Analysis' window in IBM Information Server. It displays a table of candidate primary keys for the 'GlobalCo\_Ord\_Dtl' table. The table includes columns for 'Defined Primary Key', 'Selected Primary Key', 'Defined Foreign Key', 'Column', 'Data Class', 'Data Type', 'Length', 'Unique %', 'Null %', 'Duplicate %', and 'Candidate'. The 'ordIDitemNo' column is highlighted as the selected primary key.

Defined Primary Key	Selected Primary Key	Defined Foreign Key	Column	Data Class	Data Type	Length	Unique %	Null %	Duplicate %	Candidate
			ordIDitemNo	T	STRING	0	99	0	0	False
			ORDERID	Q	DFLOAT	8	20	0	79	False
			ITEMNO	C	DFLOAT	8	0	0	100	False
			STOCKCODE	C	STRING	8	0	0	99	False
			LISTPRICE	C	DECIMAL	19	0	0	99	False
			QTYORD	C	DFLOAT	8	0	0	100	False
			QTYSHIP	C	DFLOAT	8	0	0	99	False
			QTYDUE	C	DFLOAT	8	0	0	99	False
			VALORD	Q	DECIMAL	19	43	0	56	False
			VALSHIP	Q	DECIMAL	19	32	0	67	False
			VALDUE	C	DECIMAL	19	18	0	81	False
			COMPLETE	U	INT 16	0	0	0	100	False

Below the table, the 'Duplicate Check (ordIDitemNo)' window is open, showing the following summary:

Total Records			
Records	%		
Unique	6383	99.93737	
Duplicate	2	0.03131361	
Nulls	0	0	

The 'Duplicates' section shows:

Primary Key Value	Number of Records	%
22347 2	2	0
27511 4	2	0



## Evaluación de clave primaria

Single Column		Multi-Column					
Flag Percentage Minimum:	Data Sample Status:	Sample Size:	Sampling Method:				
99 <input type="button" value="Apply"/>	Complete	5000	Row Count				
<b>GlobalCo_Ord_Dtl</b>							
Number c	Key Columns			Analysis Statistics		Primary Key	
	1	2	3	Unique %	Duplicate %	Candidate	Selected
3	ITEMNO	ORDERID	STOCKCODE	100.00	0.00	✓	✓
2	ITEMNO	ORDERID		99.94	0.06	✓	
2	ORDERID	STOCKCODE		99.81	0.19	✓	
2	ITEMNO	STOCKCODE		24.17	75.83		

- Evaluación de la clave primaria candidata
- Columnas únicas automáticamente evaluadas durante el análisis de columnas
- Soporte a la evaluación de clave multicolumna contra muestra de datos o volumen completo
- Detalles de violaciones de clave duplicada



## Agenda

- **Visión General de IBM InfoSphere Information Server**
- **Funcionalidades de Information Analyzer**
- *Descanso – Café*
- **Configuración e importación de los datos**
- **Análisis de columnas**
- **Comida**
- **Identificación de claves y análisis de relaciones**
- **Integridad referencial**
- **Análisis de base**
- **Reglas de negocio y métricas**
- **Generación de informes**



## Data Profiling: Análisis a través de tablas / Foráneas



INVESTIGATE Foreign Key Analysis

Select Data Source to Work With

WorldCo\_BillTo WorldCo\_ShipTo

View Foreign Key Analysis

ViewDetailsView

Select View: CUSTOMER\_ID

Frequency Values Analysis Details

Foreign Key Candidate Pair		
	Base Column	Paired Column
Column	CUSTOMER_ID	PARENT_CUST_ID
Table	WorldCo_BillTo	WorldCo_ShipTo
Source	GlobalCo	GlobalCo
Primary Ke	Yes	No
Foreign Ke	No	No
Data Class	Identifier	Code
Data Type	INT32	INT32
Length	0	0
Precision	0	0
Scale	0	0
Cardinality	1030	3717
Unique	No	No
Constant	No	No
Definition	No	No

Paired to Base:

#: 1021 %: 99 Common Domain:

Base to Paired:

#: 1021 %: 99 Common Domain:

Common Domain #:

- Relaciones de claves extrañas
- Integridad referencial
- Relaciones a través de dominios
- Redundancia de datos



## Relaciones entre claves e Integridad referencial

The screenshot shows the 'Foreign Key Analysis' tool in IBM Information Server. It displays the configuration for a Foreign Key analysis between 'WorldCo\_BillTo' (Base Table) and 'WorldCo\_ShipTo' (Paired Table). The Primary Key is 'BILLTO\_CUSTOMER' and the Foreign Key Candidate is also 'BILLTO\_CUSTOMER'. The tool provides a summary of violations, integrity percentages, and a Venn diagram showing the overlap between the Primary Key and Foreign Key sets.

Foreign Key > Primary Key	
	Total Records
Violations	2
Violations %	0.05380683
Integrity	3715
Integrity %	99.9462
Total Values	3717
Total %	100%

Primary Key > Foreign Key	
	Total Records
Values that	3715
Match %	100
Values that	0
% of Values	0
Total Values	3715
Total %	100%

**Violations**

Foreign Key Value	Records	Percentage
C1013	1	0.027
C7651	1	0.027

The Venn diagram shows two overlapping circles: a larger light blue circle representing the Primary Key (PK) with 3715 unique values, and a smaller light green circle representing the Foreign Key (FK) with 2 unique values. The intersection of the two circles is shaded light blue, indicating that the 2 FK values are also present in the PK set. The legend below the diagram shows a green square for 'FK > PK' and a blue square for 'PK > FK'.

- Relaciones de clave primaria y clave Foránea
- Resumen de claves Foráneas candidatas e integridad referencial
- Detalles de violaciones de clave incluyendo valores duplicados y huérfanos
- Soporte para la evaluación de clave única y multicolumna contra muestras de datos o volumen completo





## Ejercicio 3: Análisis de Claves e Integridad Referencial

### LAB 3: IDENTIFICACIÓN DE CLAVES Y ANÁLISIS DE RELACIONES

- Análisis de clave primaria
  - Clave primaria de columna única
  - Clave primaria de varias columnas
- Análisis de claves naturales
- Análisis de claves foráneas
- Ubicar elementos solapados entre dominios
- Integridad Referencial
- Asignar clave foránea



## Agenda

- **Visión General de IBM InfoSphere Information Server**
- **Funcionalidades de Information Analyzer**
- *Descanso – Café*
- **Configuración e importación de los datos**
- **Análisis de columnas**
- **Comida**
- **Identificación de claves y análisis de relaciones**
- **Integridad referencial**
- **Análisis de base**
- **Reglas de negocio y métricas**
- **Generación de informes**



## Análisis de Base

**Ejemplo: El campo 'State Abbreviation' tiene más registros que el último test, pero no hay valores o formatos incompletos o inválidos**

- Comparación Actual-Real
- Cambios estructurales
  - Definiciones
  - Deducciones
- Cambios de contenidos
- Diferentes Valores/Formatos
  - Integridad de dominio

**Select Data Source to Work With**

WorldCo\_BillTo

**View Baseline Analysis**

**Title**

- Baseline Summary
- Baseline Differences

**Common**

- STD\_POINT\_LOC\_CODE
- CITY
- ADDRESS\_LINE3
- STATE\_ABBREVIATION**
- COUNTRY\_CODE
- ZIP\_CODE
- DUNS\_NUMBER
- ADDRESS\_LINE4
- DUNS\_SUFFIX
- CUSTOMER\_TYPE
- PARENT\_CUST\_ID
- PARENT\_CUST\_TYPE
- CUSTOMER\_ID
- ACCT\_STATUS
- CUST\_AGN\_IBP\_ID
- ADDRESS\_LINE2
- ADDRESS\_LINE5
- STORE\_ID
- NAME
- ADDRESS\_LINE1
- Current Analysis Only
- Base Only

**Differences**

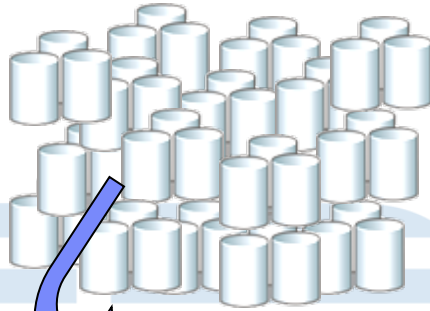
Structure Content

Value & Format Profile			Completeness & Validity Measures		
Name	Checkpoint	Baseline	Name	Checkpoint	Baseline
Cardinality	42	41	# Incomplete	3	3
# Distinct Values	1027	1026	% Incomplete	7.142857	7.317073
# Distinct Formats	2	2	# Invalid	0	0
Standard Deviation Value Frequency	0	0	% Invalid	0	0
Standard Deviation Format Frequency	0	0	# Format Violations	0	0
# Null	3	3	% Format Violations	0	0
% Nulls	7.142857	7.317073			

Close



## Data Profiling: Monitor de Cambios



**Database = DB2**  
**Schema = NewGlobalCo**  
**Table = Ord\_Dtl**  
**Column = QTYORD**  
**Total Records = full volume**



Run Column Analysis

Select the columns that you want to analyze, type a name for the analysis job, and schedule or run the job. On the Options tab, you can choose to retain OSH scripts on the server or retain earlier results when you run a column analysis job.

Job Name: \*  Job Description:

Sources	Profile Status	Last Profile
WB-GECKO-XP		
Bank Demo		
BankDemo		
Savings	100 %	
SOURCE_ID	Analyzed	08/30/2006 08:38:33 PM
SS_NUM	Analyzed	08/30/2006 08:38:33 PM
NAME	Analyzed	08/30/2006 08:38:34 PM
ADDR1	Analyzed	08/30/2006 08:38:35 PM
ADDR2	Analyzed	08/30/2006 08:38:36 PM
CITY	Analyzed	08/30/2006 08:38:36 PM
STATE	Analyzed	08/30/2006 08:38:37 PM
ZIP	Analyzed	08/30/2006 08:38:37 PM
ZIP_FOUR	Analyzed	08/30/2006 08:38:38 PM
ACCOUNT_ID	Analyzed	08/30/2006 08:38:38 PM
ACCOUNT_HOLDER_ID	Analyzed	08/30/2006 08:38:39 PM
ACCOUNT_BALANCE	Analyzed	08/30/2006 08:38:40 PM
JOINT_ACCOUNT_HOLDER	Analyzed	08/30/2006 08:38:40 PM
BANKCARD	Analyzed	08/30/2006 08:38:41 PM
ONLINE_ACCESS	Analyzed	08/30/2006 08:38:41 PM

Scheduling Options

Run Now

Schedule

Schedule Description :

Start Date :

End Date :

Occurrence :  Edit

- (Not set)
- Hourly
- Every day
- Every week
- Every month
- Every year
- Custom

- Ejecución programada
- Comparaciones Actual-Anterior



## Ejercicio 4: Análisis de Base

### LAB 4: IDENTIFICACIÓN DE LOS CAMBIOS SUFRIDOS POR LOS DATOS EN EL TIEMPO





## Agenda

- **Visión General de IBM InfoSphere Information Server**
- **Funcionalidades de Information Analyzer**
- *Descanso – Café*
- **Configuración e importación de los datos**
- **Análisis de columnas**
- **Comida**
- **Identificación de claves y análisis de relaciones**
- **Integridad referencial**
- **Análisis de base**
- **Reglas de negocio y métricas**
- **Generación de informes**

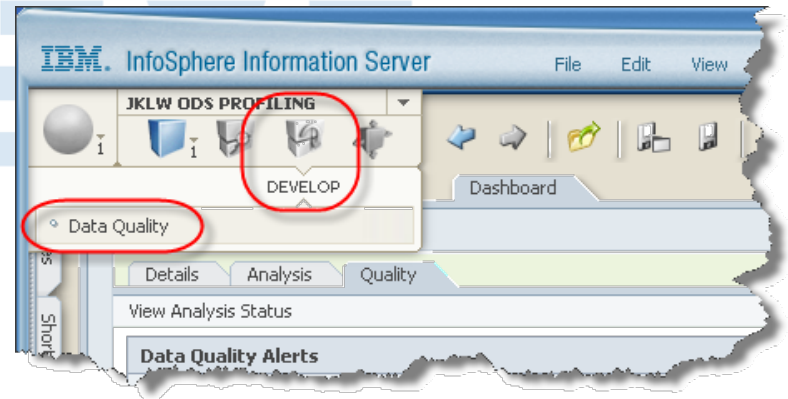


## Evaluación de Reglas de Negocio

- **Análisis integrado de reglas** para verificar el cumplimiento de reglas críticas del negocio y prever actuaciones/resultados sobre el tratamiento de la información
- **Reglas de Negocio reutilizables** Definición y reutilización a través de múltiples fuentes de datos
- **Varios niveles de evaluación**, evaluación de cumplimiento, registro, fuente significativamente los niveles de prestación de un mayor conocimiento de los posibles problemas de calidad.
- **Presentación de Informes** en apoyo de las normas de análisis

### Ejemplos de reglas:

- El "Género" y "Población" deben estar en las tablas de validación
- El Número de "Seguridad Social" debe tener el formato 99/99999999/99
- Si la "Fecha de nacimiento" no está vacía y es mayor que 1900-01-01 y menor que < "HOY" el tipo de cliente es 'P'.
- Si la cuenta bancaria es válida la identificación de la sucursal es la Subdivisión de la cuenta



## Soporte de Reglas Multi-nivel

Summary		Run	Baseline
<b>Statistics</b>			
◦ Total Records		14664	14664
◦ # Met All Rules		13894	10966
◦ % Met All Rules		94.7490 %	74.7818 %
◦ # Did Not Meet 1 or More Rules		770	3698
◦ % Did Not Meet 1 or More Rules		5.2510 %	25.2182 %
◦ Mean #		0.0531	0.2567
◦ Mean %		1.3281 %	6.4171 %
◦ Standard Deviation #		0.2270	0.4470
◦ Standard Deviation %		5.6750 %	11.1747 %
<b>Validity Benchmark</b>			
◦ Benchmark Status		<span style="color:red">✘</span>	
◦ Benchmark		% Not Met <= 1.0000 %	
◦ Variance %		4.2510 %	
◦ Variance #		623	
◦ Trend		<span style="color:red">✘✘✘</span>	
<b>Confidence Benchmark</b>			
◦ Benchmark Status		<span style="color:green">✔</span>	
◦ % of Rules Not Met Per Record Limit		10.0000 % (0 rules)	
◦ Max Records Allowed Over Limit		10.0000 % (1466 records)	
◦ Actual Records Over Limit		0	
◦ Variance %			
◦ Variance #			
◦ Trend		<span style="color:green">✔✔</span>	
<b>Baseline Comparison Benchmark</b>			
◦ Benchmark Status		<span style="color:green">✔</span>	
◦ Baseline Date		3/10/2009 4:57:45 PM	
◦ Similarity		62.6696 %	
◦ Degradation		0.0000 %	
◦ Improvement		0.0000 %	

### Nivel "Regla"

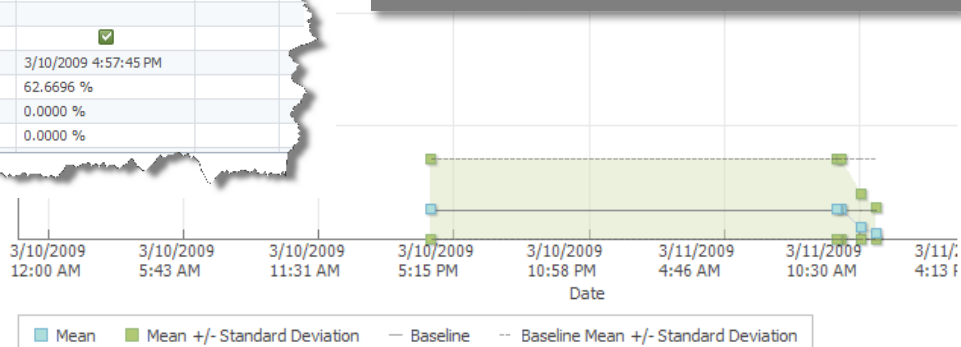
- El 4% de los registros debe Cumplir/Fallar la regla 12
- El 10% de los registros fallan la regla 3

### Nivel "Registro"

- El registro 16 falla las reglas 3, 4 ,9 y 10
- El registro 13 tiene una desviación del 7,45%
- El registro 12 en la columna X tiene el valor Y, la probabilidad de error es del 34%, esperábamos el valor Z

### Nivel "Fuente"

- El 4% de los registros fallan/cumplen al menos 1 regla
- La media de desviación en reglas es de 2,127
- El registro 13 tiene la mayor desviación con un 7,34
- Los registros con más violaciones son...



Did Not Meet 1 or More Rules %

---

**Confidence Benchmark**

Monitor records according to number of rules not met

Rules Not Met Limit (Maximum Acceptable Rules Not Met Per Record)

Max Records Allowed Over Not Met Limit:

---

**Baseline Comparison Benchmark**

Monitor Records According to Difference from Baseline

Benchmark:

Baseline Date:



## Evaluación de reglas Interactivo

- Pruebas “al-vuelo”; Informe de resultados; Salvar/Planificar la Regla evaluada

- Lanzamiento Inmediato
- Revisión de resultados

The screenshot displays the 'Result' tab of a rule evaluation interface. It includes a summary table, a benchmark status section, and a table of failed records.

Summary	
Run	
Rule Definition	L1VAL_Item992
Statistics	
Total Records	6867
# Met	13
% Met	0.1893 %
# Not Met	6854
% Not Met	99.8107 %
Validity Benchmark	
Benchmark Status	Fail
Benchmark	% Not Met <= 0.0000 %
Variance %	99.8107 %
Variance #	6854
Trend	
Job Details	
Start	3/9/2009 11:08:58
End	3/9/2009 11:09:34
Elapsed	0 minutes, 36 seconds
Sample Details	
DataSample	No

Benchmark Status:	Benchmark:	Variance:	Total Records:	Met #:	Met %:	Not Met #:	Not Met %:
Fail	% Not Met <= 0.0000 %	99.8107 %	6867	13	0.1893 %	6854	99.8107 %

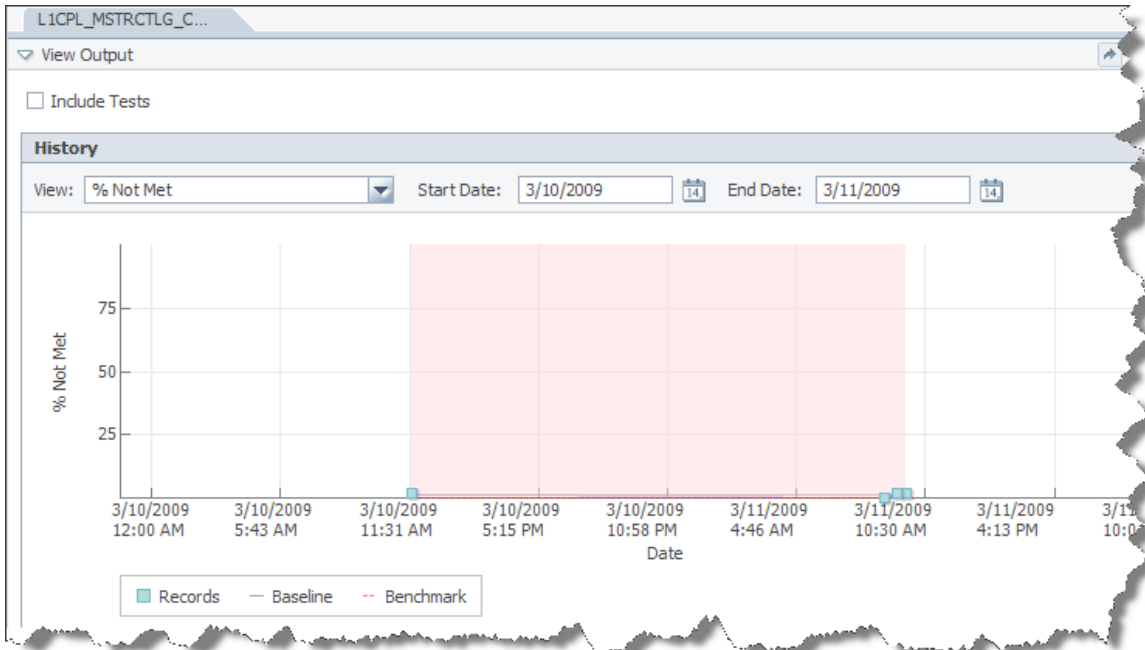
  

**Output: Do not meet rule conditions**

1 - 50 of 6854 Page 1 of 138

ITEMID	CDP	DESCR	DIV
242301	[NULL]	2-EHMA	MOB
TGA/BULK	[NULL]	THIOGLYCOLIC ACID	MOB
5610088	XF3	#4793K46 SHUTOFF C VALVE 1/8" M&F	160
J910NDC	XF8	*J910 ND C	302
7368BEIGEMACESI	X77	7368 BEIGE MAC ES	302
7368BEIGEMACESI	[NULL]	7368 BEIGE MAC ES	302
7370BROWNMACESY	X77	7370 BROWN MAC ESY	302
7375YELLOWMACFB	X77	7375 YELLOW MAC FB	302
7380ORANGETMACI	X77	7380 ORANGE MAC FB	302
0074505	XF3	745-05 ACID MURIATIC	160

## Resultados



- Establecer parámetros de seguimiento por Diferencia
- Métricas de una o varias Reglas
- Organización de Métricas y normas en carpetas definidas por el usuario

Data Quality Alerts															
Name	Status	Total Records	Meet All Rules		Did Not Meet 1 c		% of Rules Not Met Per Record		Validity		Confidence		Baseline Comparison		
			#	%	#	%	Mean	Standard Deviation	Variance	Trend	Variance	Trend	Variance	Trend	
MasterCatalog															
L1CPL_MSTRCTLG_CARRID_Exists	✖	14669	14461	98.5820 %	208	1.4180 %			1.4180 %	✖✖					
L1CPL_MSTRCTLG_CLSG_Exists	✔	14664	14664	100.0000 %	0	0.0000 %			0.0000 %	✔					
L1CPL_MSTRCTLG_DIV_Exists	✔	14669	14669	100.0000 %	0	0.0000 %			0.0000 %	✔✔✔					
L1FMT_MSTRCTLG_PKG	✔	14664	14664	100.0000 %	0	0.0000 %			0.0000 %	✔					
L1VAL_MSTRCTLG_HAZMAT	✖	14669	14663	99.9591 %	6	0.0409 %			0.0409 %	✖✖✖					
L3RUL_MSTRCTLG_DESCR_Usable	✔	14664	14587	99.4749 %	77	0.5251 %				✔					
L3RUL_MSTRCTLG_SIZEYPE_HashChar	✖	14664	14096	96.1266 %	568	3.8734 %			3.8734 %	✖✖✖					
T1VAL_MSTRCTLG	✖	14664	13894	94.7490 %	770	5.2510 %	1.3281 %	5.6750 %	623.36664	✖✖✖	✔✔		0.0000	✔✔	





## Monitor de Resultados

- Log de Ejecución Historico Eventos
- Tendencias y Estadísticas
- Detalles de Ejecuciones

Account Gender Exists

Run

Job Name: \*  
Account Gender Exists run

Job Description:  
An execution of Account Gender Exists

Run as Test

Rule Executable

Scheduler Sample Options

Run Now

Create External Bundle Location  
c:\tmp

Schedule

Overview Result

Select View

- By Record
- By Distribution
- By Rule
- By Pattern

Distribution by Rules Not Met

View By: # of Rules Not Met

# of Rules Not Met	Run		Baseline	
	Record #	Record %	Record #	Record %
0	13894	94.7490 %	10966	74.7818 %
1	761	5.1896 %	3632	24.7681 %
2	9	0.0614 %	66	0.4501 %
3	0	0.0000 %	0	0.0000 %
4	0	0.0000 %	0	0.0000 %

Baseline Set Comparison

	Run	Baseline
Date/Time Executed	3/11/2009 12:45:50 PM	3/10/2009 4:57:50 PM
Total Records	14664	14664
Mean Rules Not Met	1.3281 %	6.4171 %
Standard Deviation	5.6750 %	11.1747 %
Similarity	62.6696 %	
Degradation	0.0000 %	

Carrier ID Validation

View Results

Include Tests

Job Log

Type	Timestamp	Validity				Trend	Contact	Sample
		Severity	# Pass	# Fail				
Run	01/01/09	10.8%	792	208	⊘	Mim Foster		
Run	12/01/08		904	96	⊘	Mim Foster		
Run	11/01/08		926	74	⊘	Mim Foster		
Run	10/01/08		500	0		Mim Foster	✓	
Run	09/01/08		500	0		Mim Foster	✓	



## Data Rules Stage – Integración Information Analyzer vs QualityStage/DataStage



- **QualityStage/DataStage** aporta un “stage” que hace uso de la normas/reglas publicadas por Information Analyzer integrando los controles de calidad en los flujos de trabajo QualityStage / DataStage
- El stage de “Data Rules” hace uso de las Reglas de Calidad publicadas por Information Analyzer Data, permitiendo añadir estas comprobaciones de calidad a flujos de QualityStage / DataStage
- Este “stage” permite evaluar múltiples reglas en una sola “pasada”
- De este objeto podremos obtener 3 salidas:
  - Registros que incumplen la regla
  - Reglas incumplidas por registro
  - Relación de registros que cumplen la regla



## Reporting exhaustivo

The screenshot displays the 'Reports' section of the InfoSphere interface. A table lists various reports, with 'Most Frequent Formats - WorldCoShipTo StoreID' selected. A detailed view of this report is shown, including a 'Column Frequency' section with a bar chart and a 'Column Level Summary' table.

Format Cardinality Count :	10
Value Cardinality Count :	302
Null Count :	0
Total Actual Row Count :	3717

Format	Count	Total Rows %	Total Rows Cumulative %	Example Values
NA	3289	87.9473	87.9473	[SPACES]

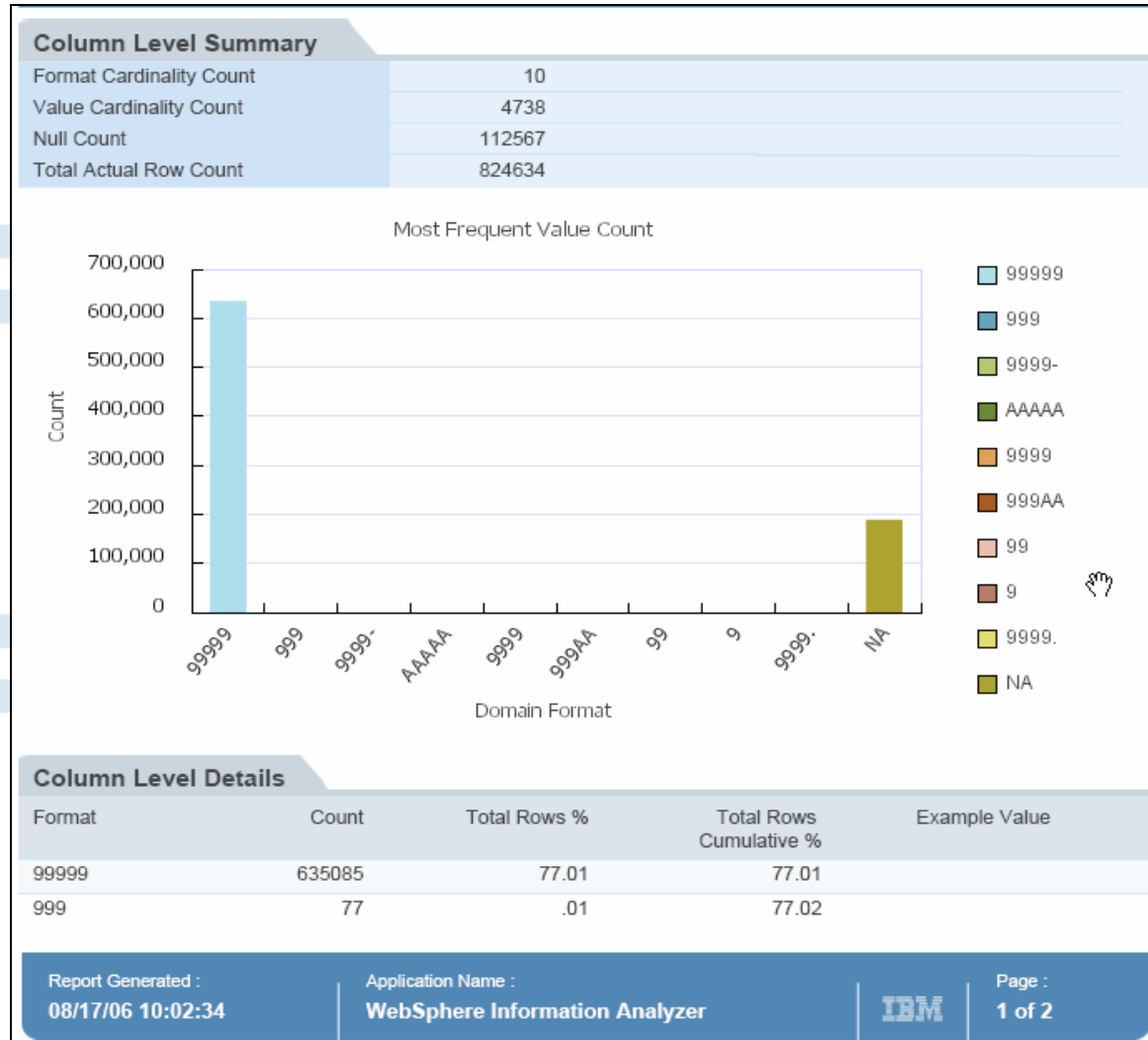
- Más de 40 plantillas de informes Ya predefinidas
- Salva informes con logos particularizados, nombres de informes y parámetros relevantes
- Notas incrustadas
- Entrega informes en UI o consola de reporting basada en navegador
- Soporta programación de informes



## Data Profiling: Entrega de la Información



- Visualiza resultados
- Entrega en UI o vía navegador
- Colaboración con los resultados de la empresa



## Compartición de Metadatos – Fuentes de Datos

The screenshot shows the 'Metadata Import' interface. On the left, a table titled 'Data Collections' lists metadata for an 'authors' collection. On the right, a dialog box titled 'Register Interest with Data Sources' allows selecting data sources to register interest to DataStage.

Name	No. of Columns	Sequence	Data Type	Length	Pre
authors	9				
au_id	1	1	id	11	
au_lname	2	2	varchar	40	
au_fname	3	3	varchar	20	
phone	4	4	char	12	
address	5	5	varchar	40	
city	6	6	varchar	20	
state	7	7	char	2	
zip	8	8	char	5	
contract	9	9	bit		

The 'Register Interest with Data Sources' dialog box shows a tree view of data sources. The 'GlobalCo\_Source' folder is expanded, showing several data sources with checkboxes: GlobalCo\_BillTo, GlobalCo\_Ord\_Dtl, GlobalCo\_Ord\_Hdr, GlobalCo\_Sales, GlobalCo\_ShipTo, WorldCo\_BillTo, WorldCo\_Sales, and WorldCo\_ShipTo. The 'Northwind' folder is also expanded, showing a list of tables: Categories, CustomerCustomerDemo, CustomerDemographics, Customers, Employees, EmployeeTerritories, Order Details, Orders, Products, Region, Shippers, Suppliers, and Territories. Other folders like RetailCo, Thunderstorm, Westzephyr, and WB-COUGER-APPS are visible but not expanded.

- Respositorio común para todos los metadatos
- Soporta Bases de Datos Estandar
- Los Metadatos descubiertos son compartidos por toda la Suite



## Metadatos compartidos – Resultados analíticos

The screenshot shows the 'Table Definitions' window for 'WORLDSCO\_BILLTO'. The 'Columns' tab is active, displaying a table with columns for 'Length', 'NumberEmptyValues', 'NumberFormats', 'NumberNullValues', 'NumberPatterns', 'NumberValidValues', and 'NumberValues'. A 'Publish Analysis Results' dialog box is overlaid on the table, allowing the user to choose an analysis summary to publish (Current Analysis, Checkpoint, or Baseline) and to include notes. The 'Include Notes' checkbox is checked. The 'Edit Column Meta Data' dialog box is also visible, showing details for the 'STATE\_ABBREVIATION' column, including its SQL type (VarChar) and nullable status (Yes). The 'Analytical information' tab is selected, showing a summary of the column's data characteristics and a note: 'Need to clean these up!'. The interface includes standard navigation buttons like 'Previous', 'Next', 'Close', 'Apply', 'Reset', and 'Help'.

- Resultados analíticos compartidos directamente con DataStage y QualityStage
- Incluye resúmenes de tablas y columnas más anotaciones (si se selecciona)





## Metadatos compartidos – Semántica de Negocio

Name	Description
<input type="radio"/> Customer	A customer is an individual or organization wh...
<input type="radio"/> Customer Name	A Customer Name is a specific description or d...

- Categorías de negocio creadas en Business Glossary
- Condiciones de Negocio creadas con Business Glossary o Information Analyzer
- Condiciones vinculadas al repositorio de objetos (columnas, tablas)

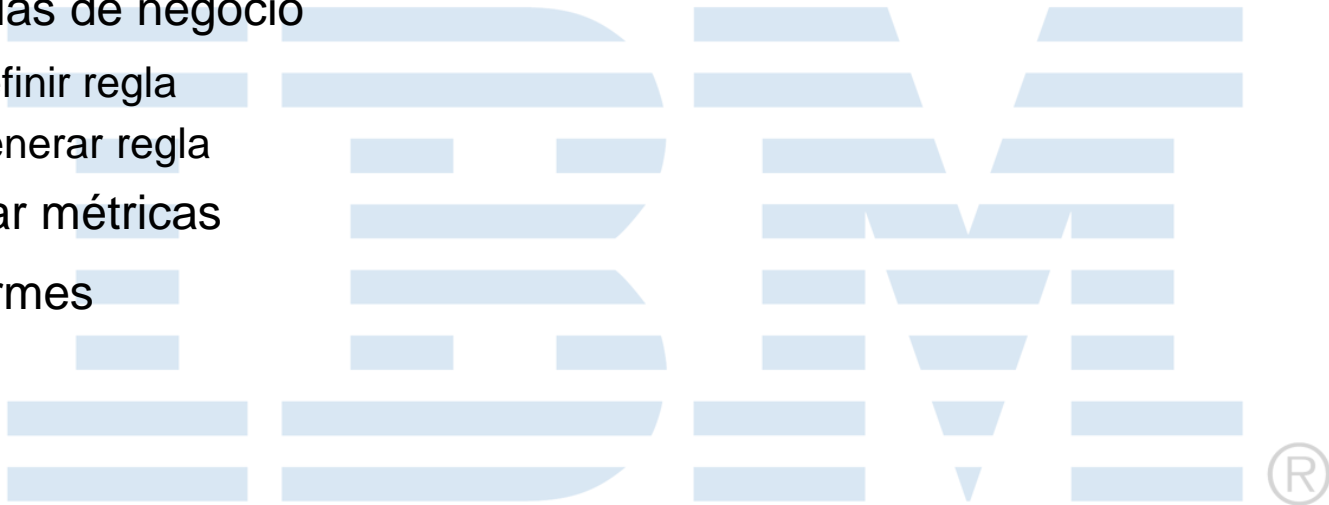
Title	Status	Description	Created On	Last Modified
Customer	Candidate	A customer is an individual or organization who requests and receives products.	09/14/2006	09/14/2006
Customer Name	Candidate	A Customer Name is a specific description or designation for a Customer.	09/14/2006	09/14/2006



## Ejercicio 5: Reglas de Negocio

### LAB 5: DATA QUALITY

- Reglas de negocio
- Definir regla
- Generar regla
- Crear métricas
- Informes



# InfoSphere™



¡Gracias por su atención!