Proof Of Technology
IBM InfoSphere Information Analyzer

13 de Diciembre de 2012

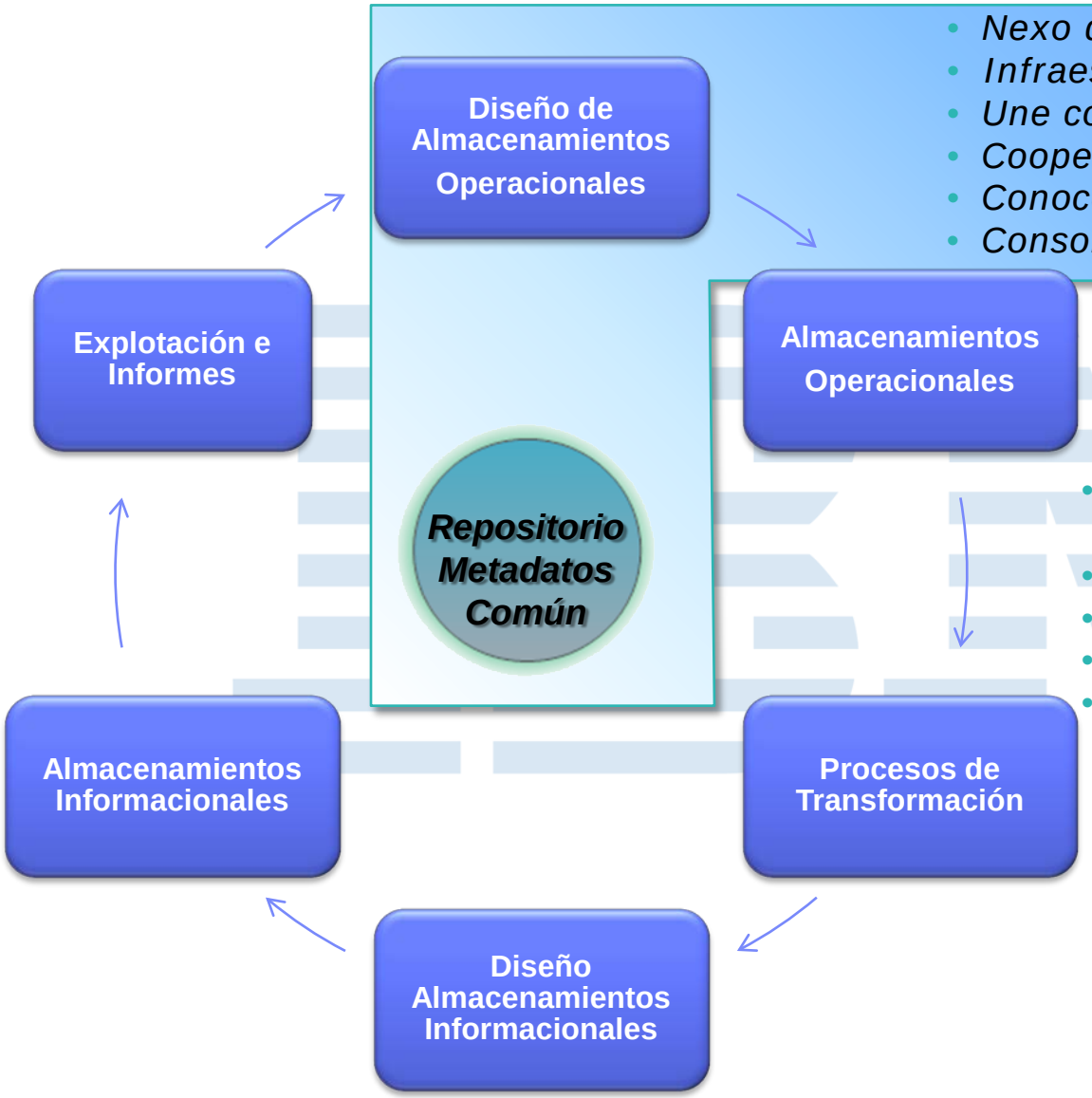
Javier Gómez Kayser
Javier.gkayser@es.ibm.com

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

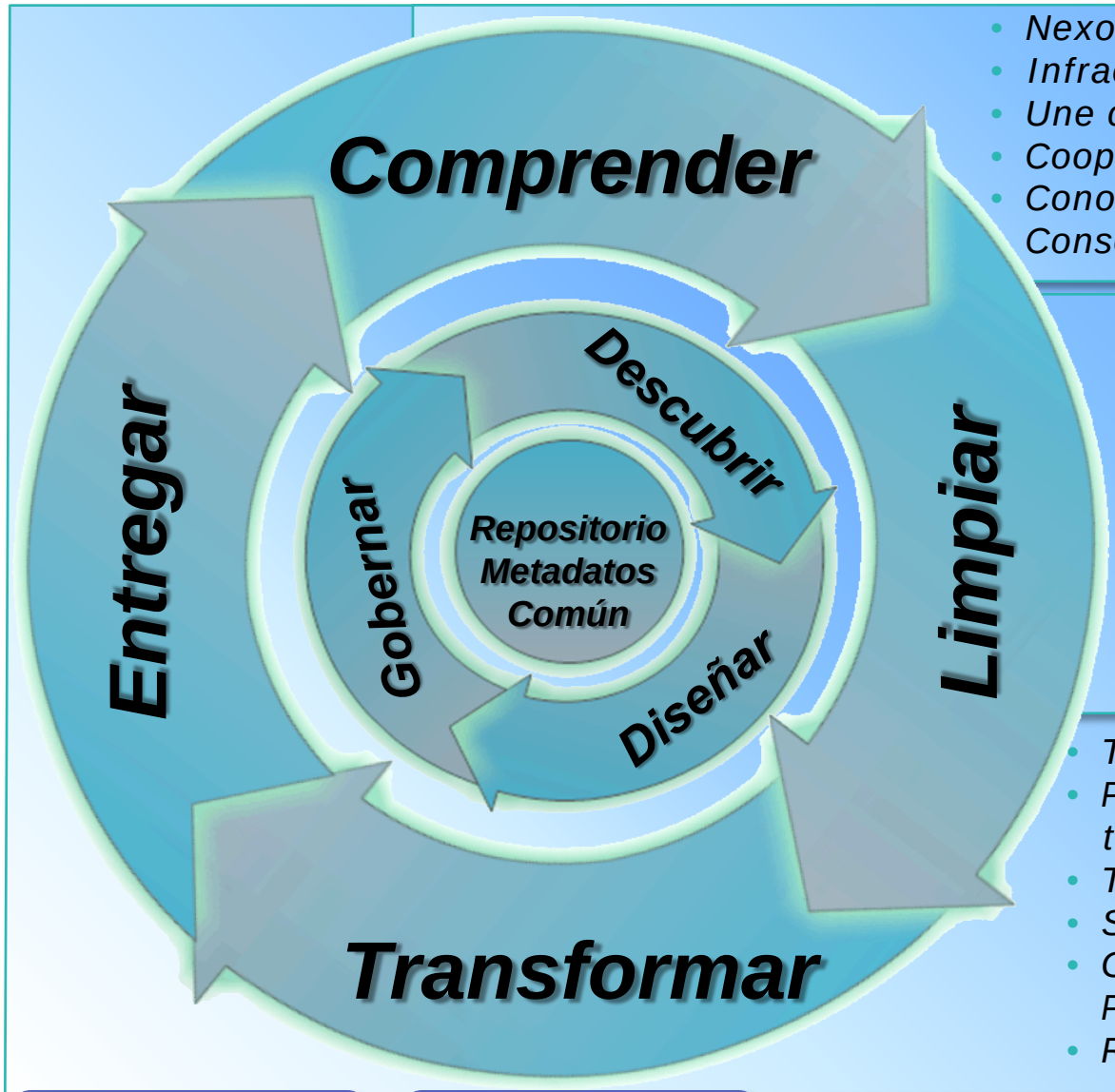
Visión General de IBM InfoSphere Information Server

- **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*

- **Cobertura del ciclo de vida del dato**
- **Visión unificada de Metadatos**
- **Metadatos propios y de terceros**
- **Análisis de impacto inter-fase**
- **Linaje de datos**



- *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*

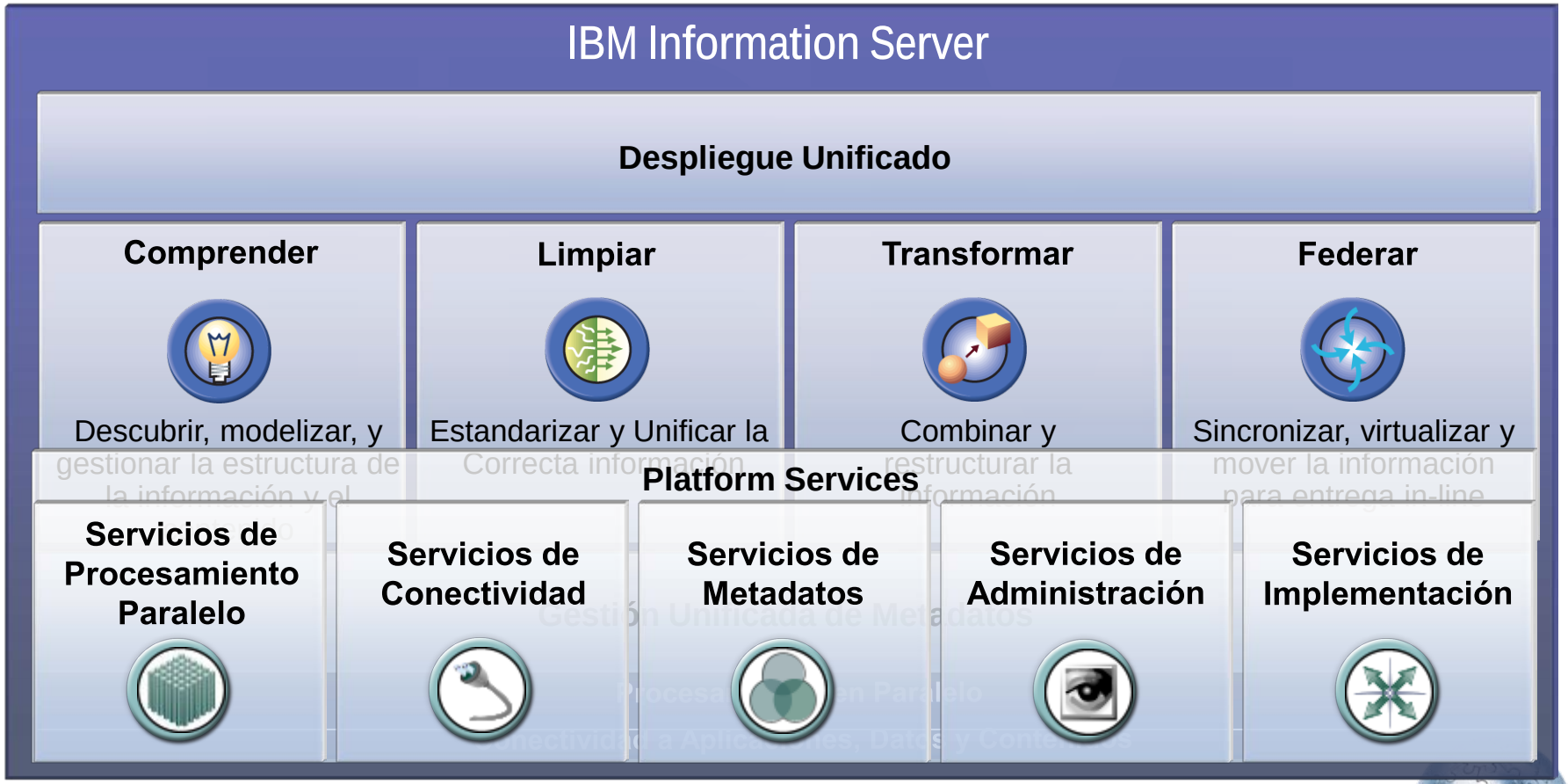
- *Acortar los tiempos de ciclo de desarrollo*
- *Aumentar la eficiencia operativa*
- *Eliminar y simplificar los procesos duplicados*
- *Colaboración del equipo*
- *Respuesta rápida al mercado y al cliente*
- *Minimizar los riesgos del proyecto*

- *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*

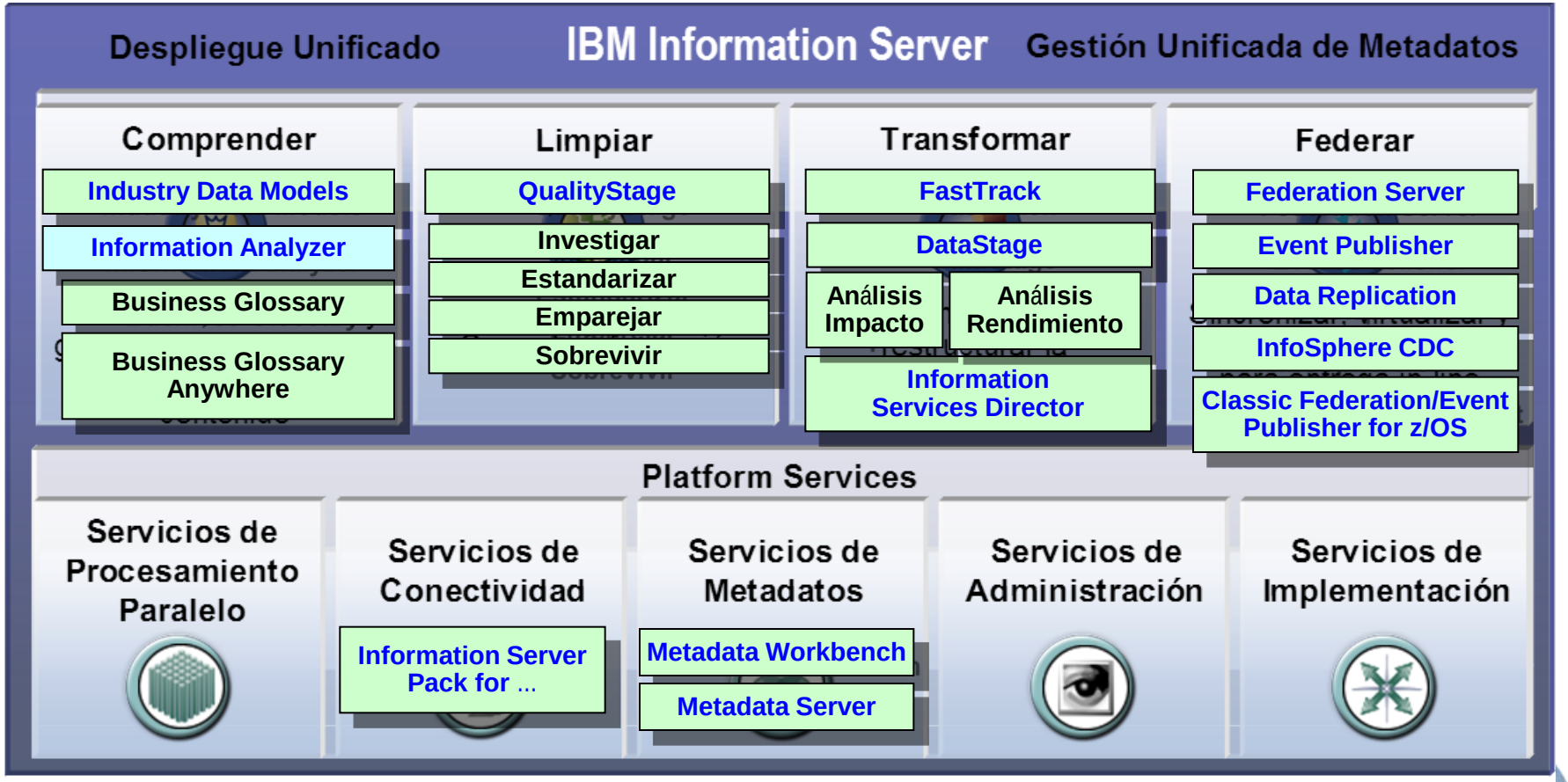
Componentes de Information Server

Confiando en la entrega de la Información



La Solución IBM : IBM Information Server

Confiando en la entrega de la Información

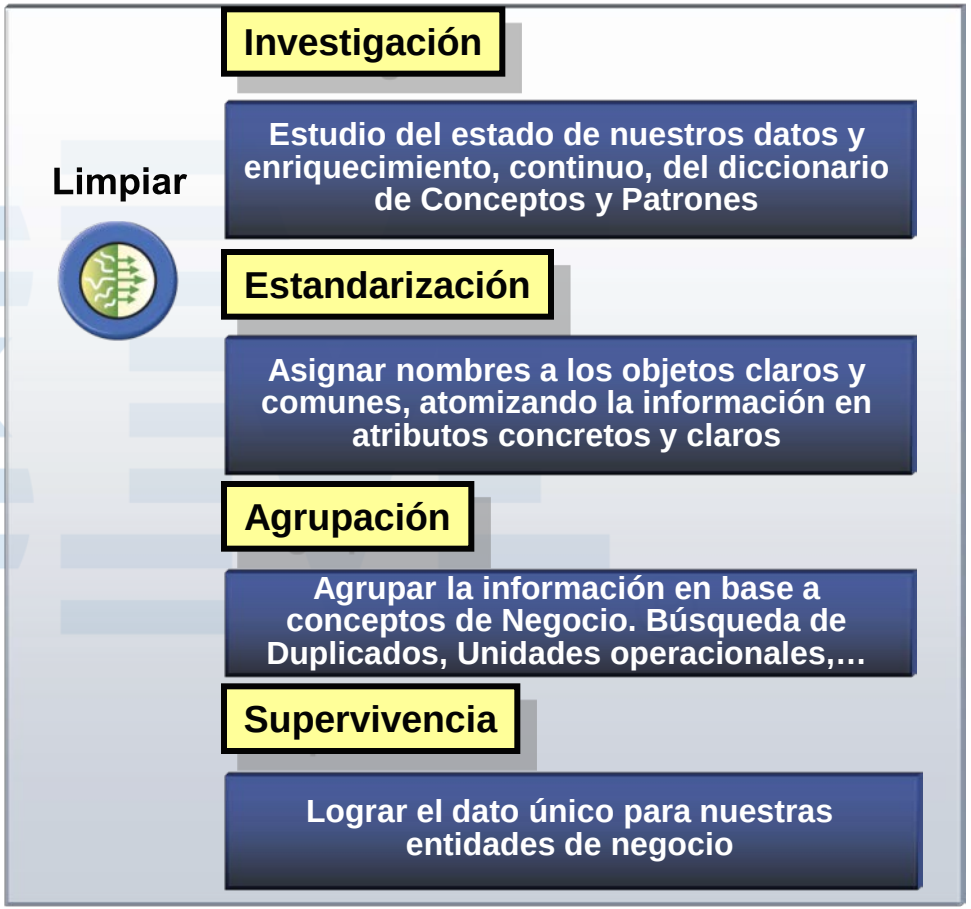


Metadatos físicos: InfoSphere Information Analyzer

- 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
  - Analisis 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

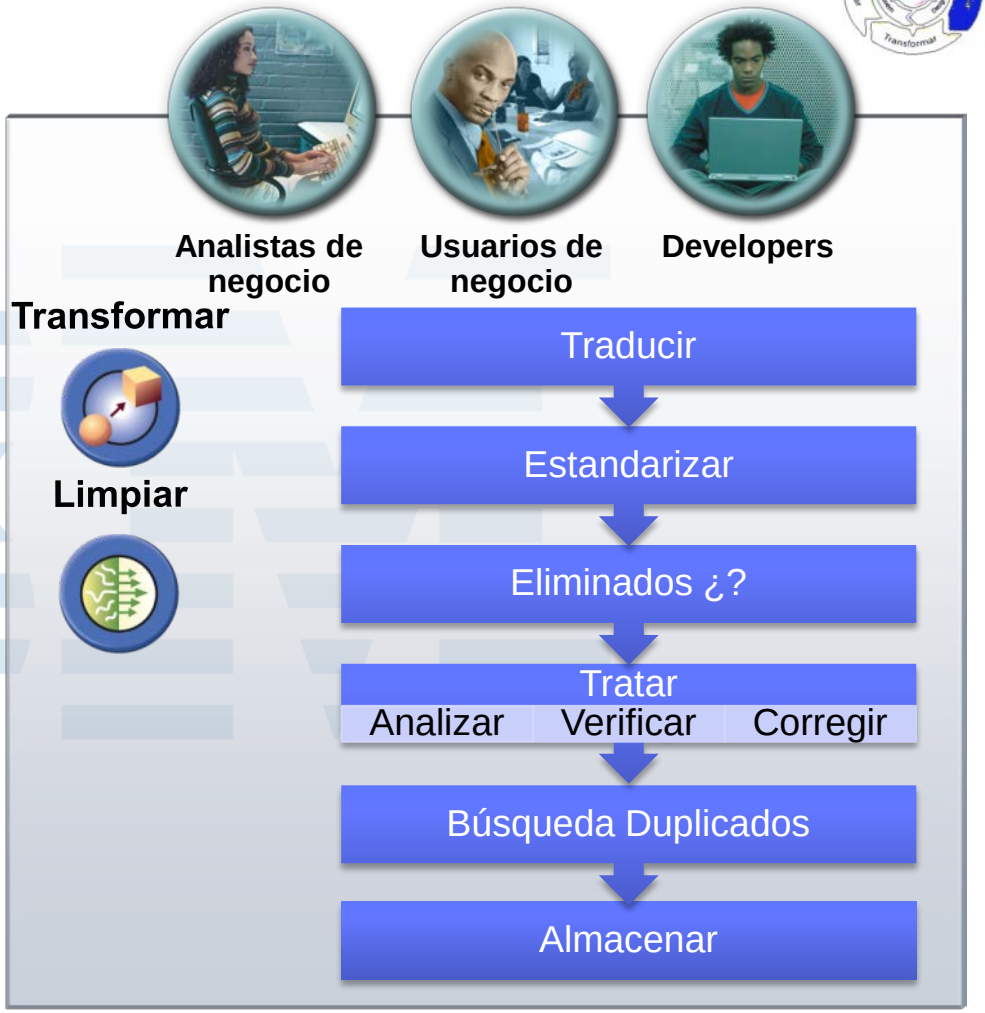
Limpieza de datos: InfoSphere QualityStage

- 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”



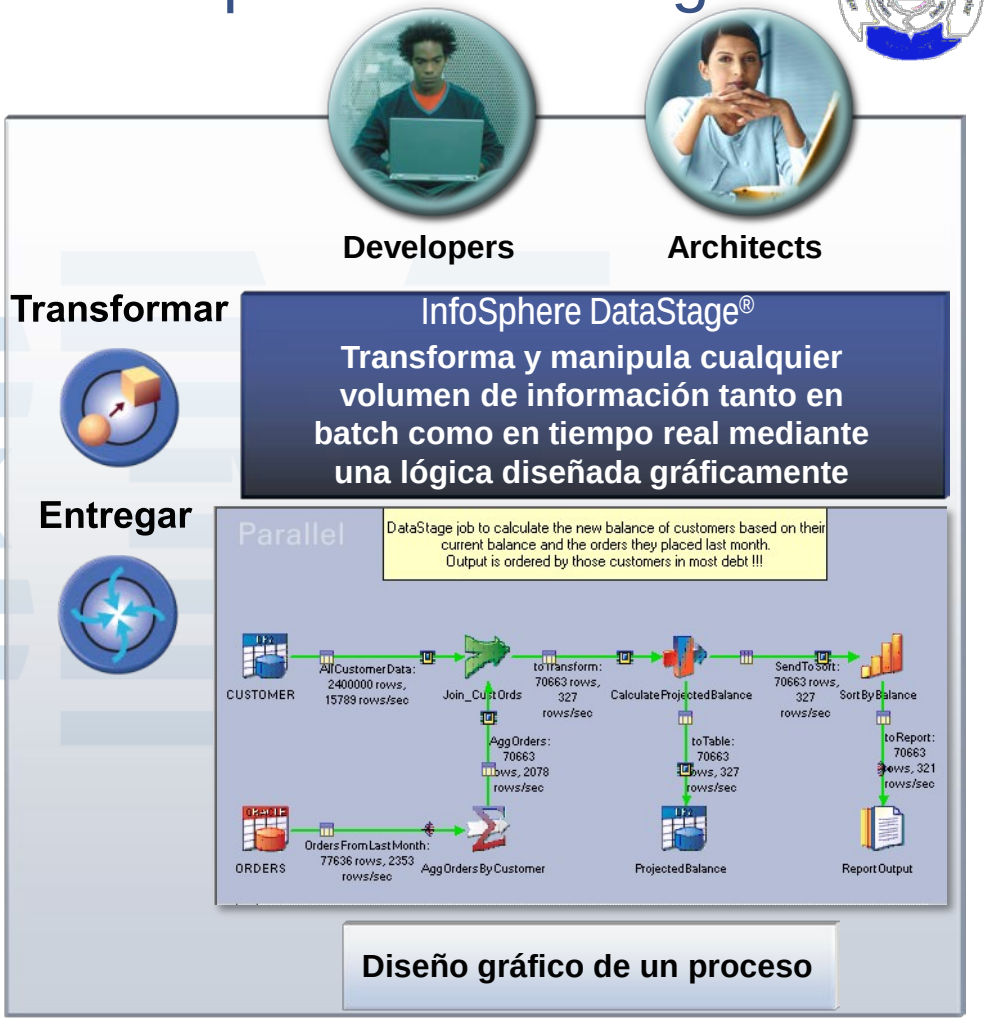
Tratamiento de Información: Address Doctor

- 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



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 Bach como para tiempo real

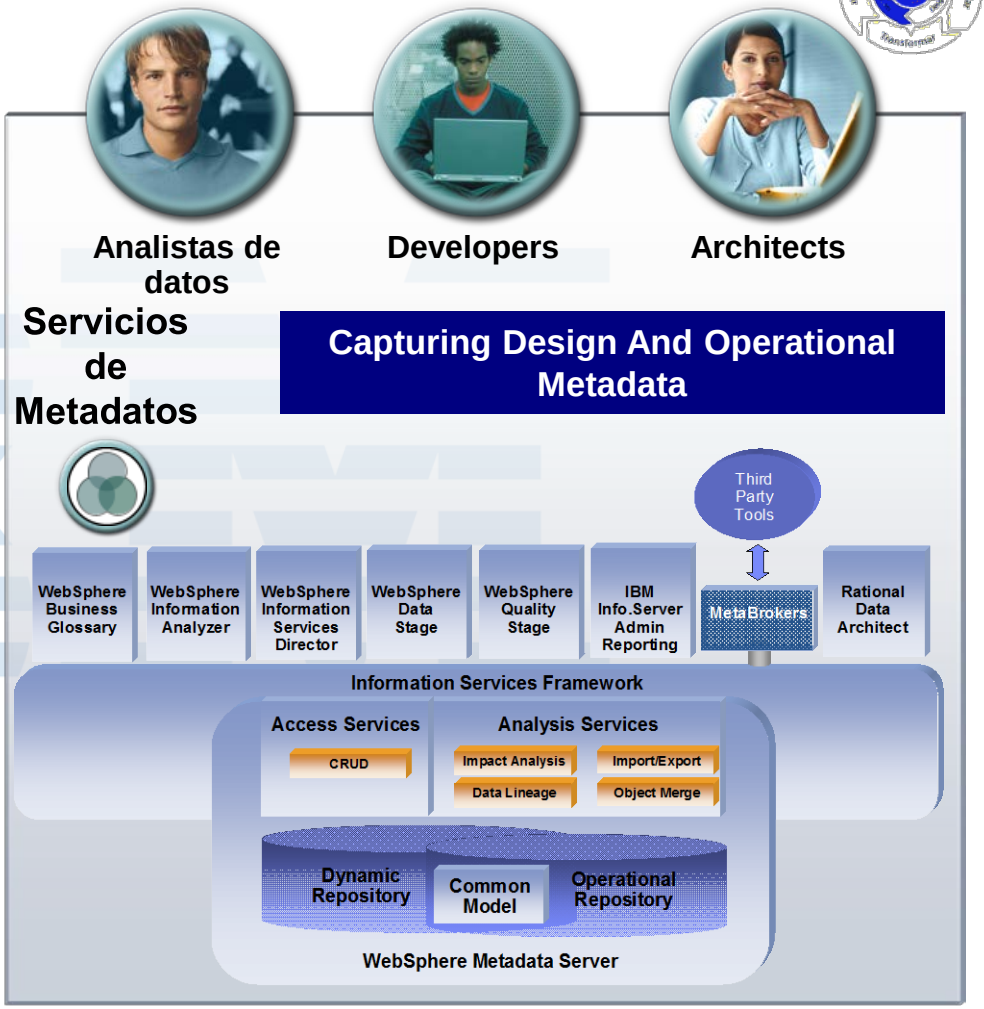
Transformación y movimiento: InfoSphere FastTrack

- 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

Detalles

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-FIIIIVVVV.

Detalles

Business Glossary Anywhere

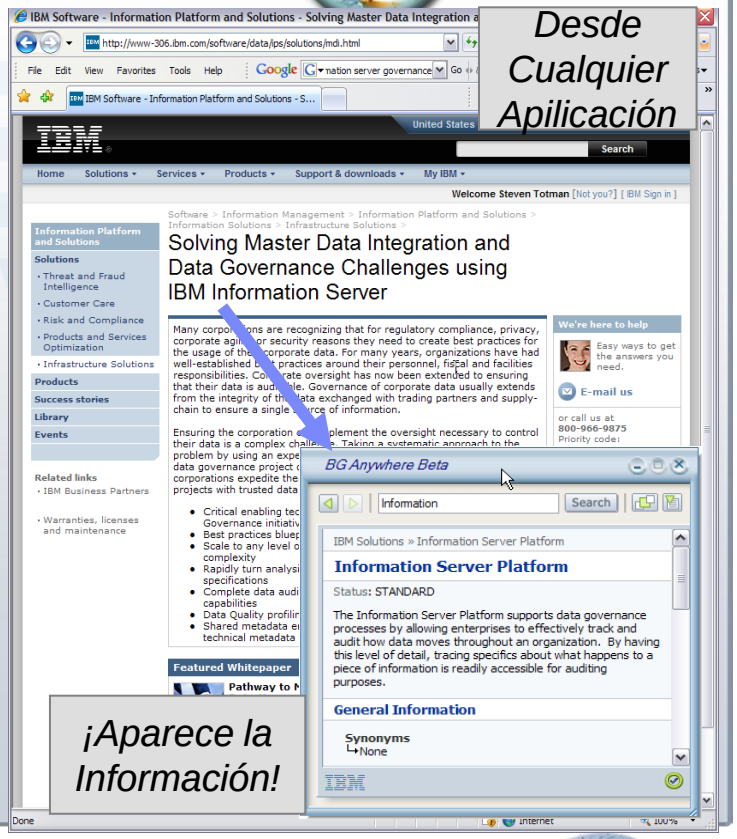
Acceso al "Business Glosary" desde cualquier Aplicación

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

Comprender



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

La Forma de los datos: Metadata Workbench

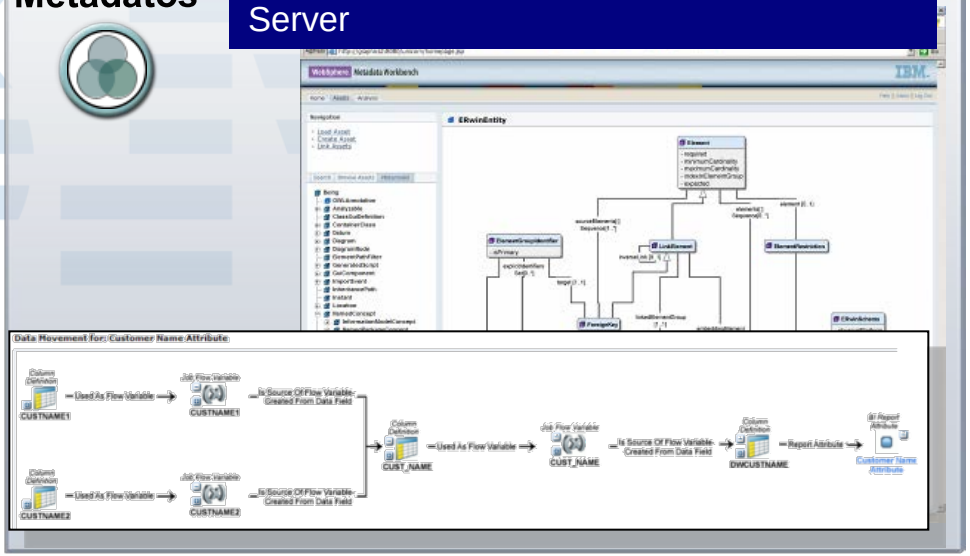
- Módulo para revisar y administrar activos de Information Server
- Explora, analiza y gestiona metadatos, de forma gráfica o textual
- Informes Claros sobre la herencia (lineage), de los datos para verificar el cumplimiento de reglas de Negocio
- Evaluar el impacto ante cambios, tanto sobre los objetos 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

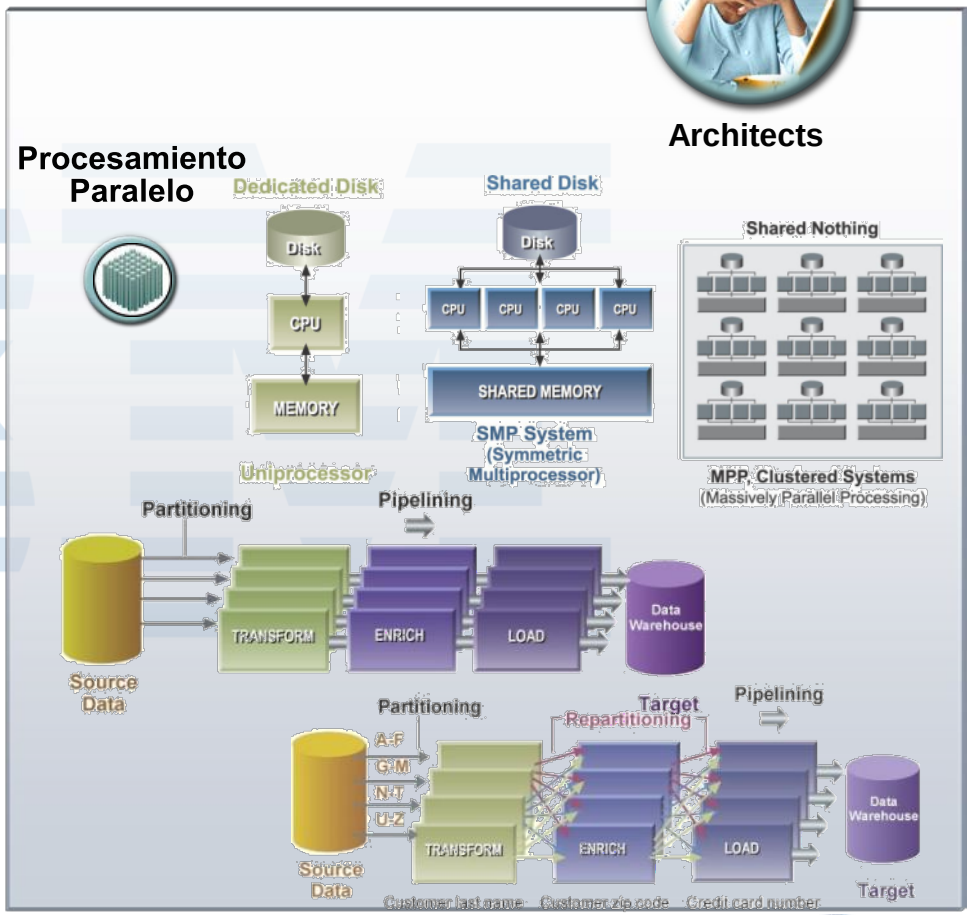
Exploración mediante Web de los activos generados o usados por Information Server



Detalles

Escalabilidad Ilimitada: Paralelismo

- Atributos
  - Paralelismo mediante Pipeline. Clonación automática de los procesos y sus sub-tareas por nodo de ejcución
  - 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 3rd 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

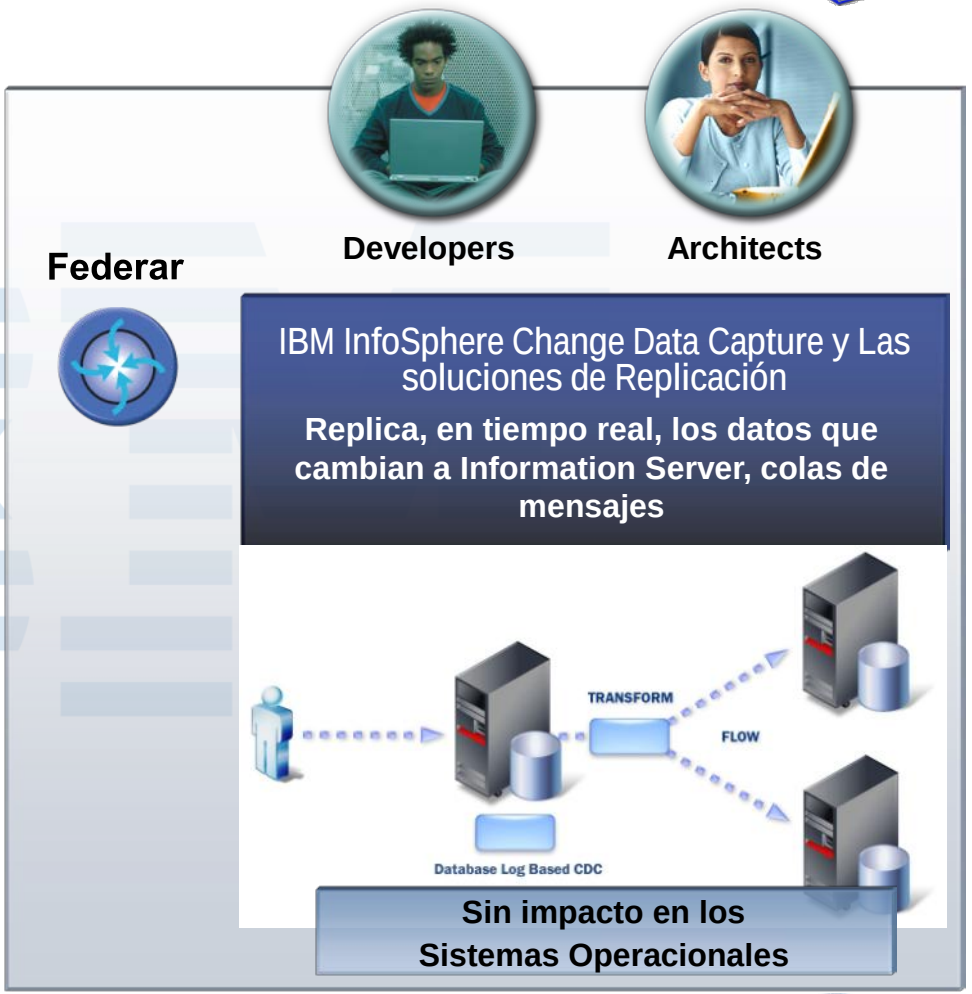
DB2 (on Z, I, P, X series)
Oracle
SQL Server
Sybase
Informix
IMS
VSAM
ADABAS …

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

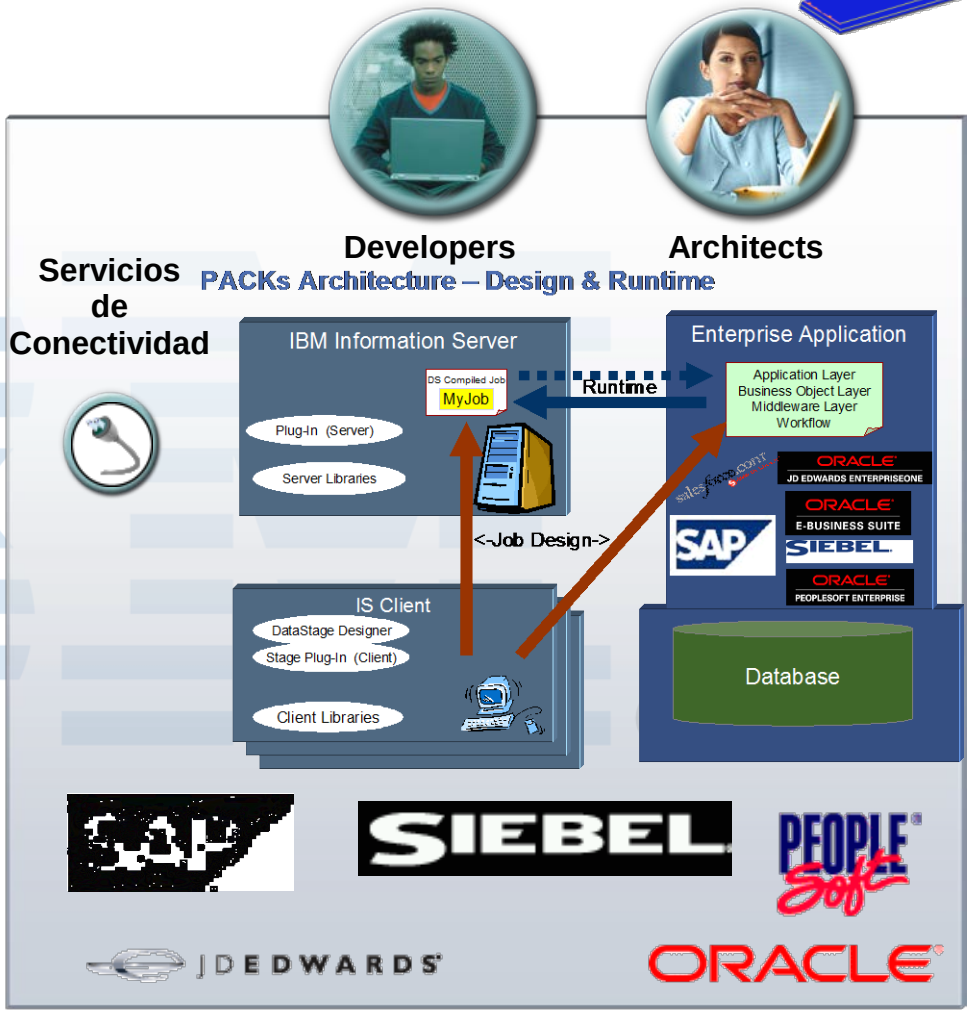
- Captura de los cambios en los datos en tiempo real
  - Integración Dinámica
  - Sincronización
  - Replicación
- Mínimo impacto en los sistemas Operacionales
- Gran escalabilidad y rendimiento
- Garantía de integridad en los datos



Detalles

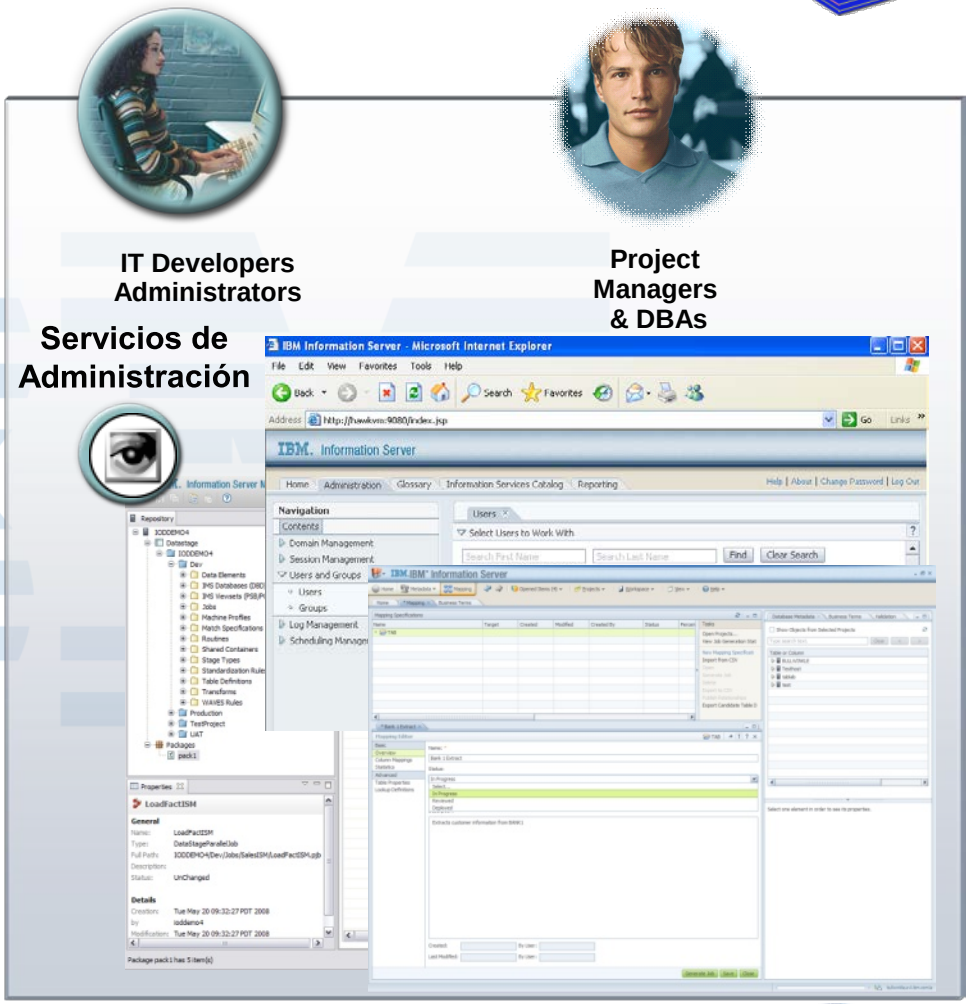
Servicios de Conectividad: Information Server Pack for …

- 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…

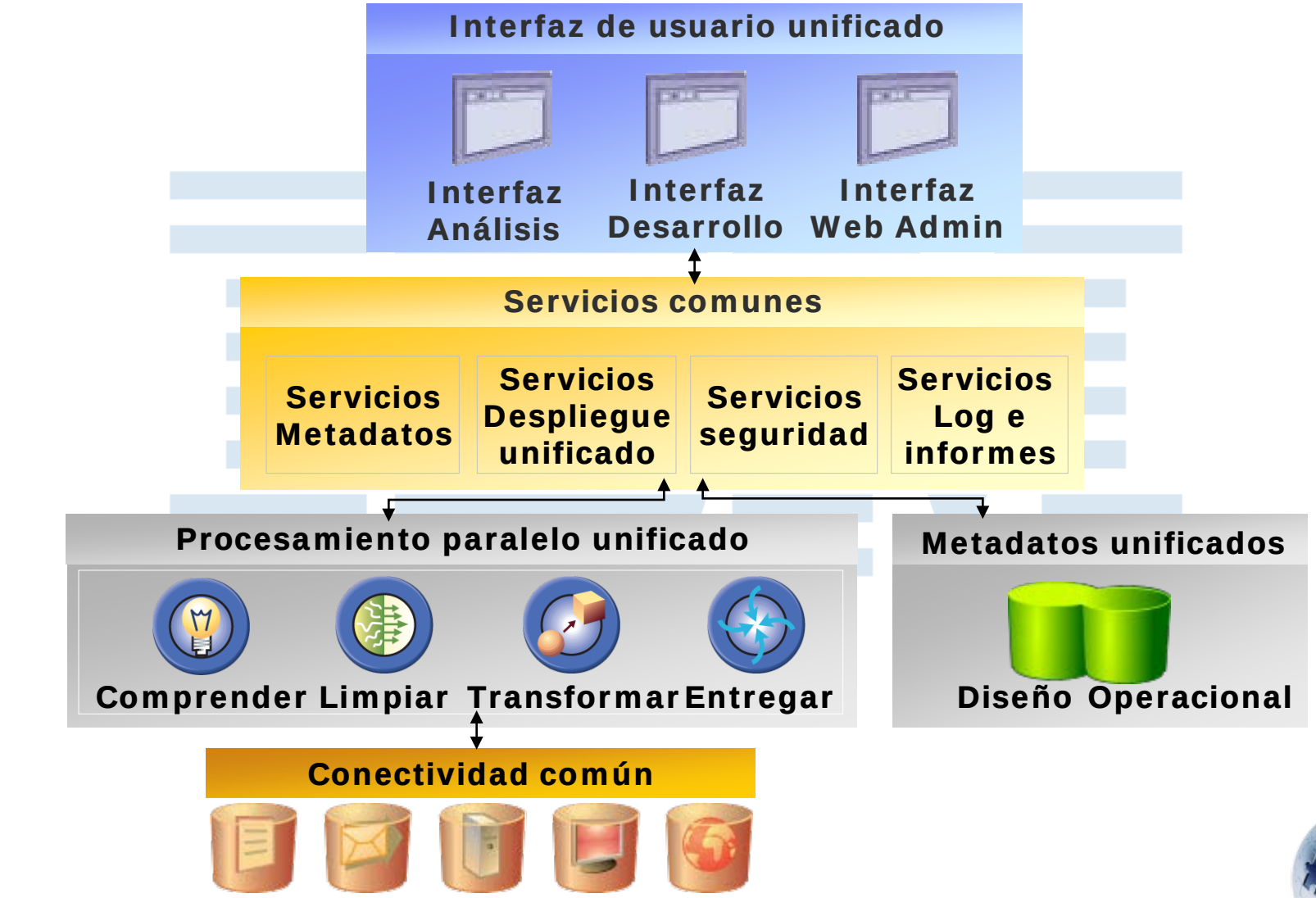


Punto único de Control: Servicios de Administración

- 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)



Arquitectura de IBM Information Server

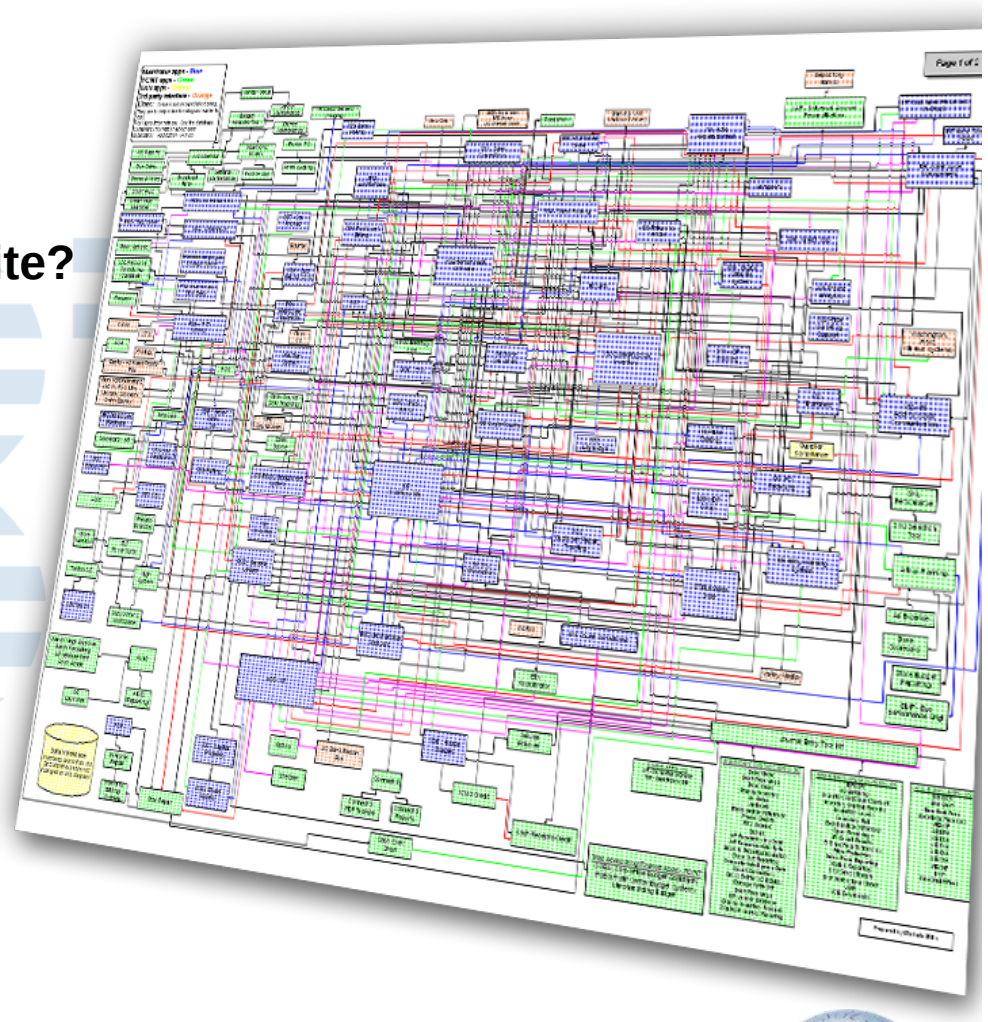


Agenda

- **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

- **¿Donde 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

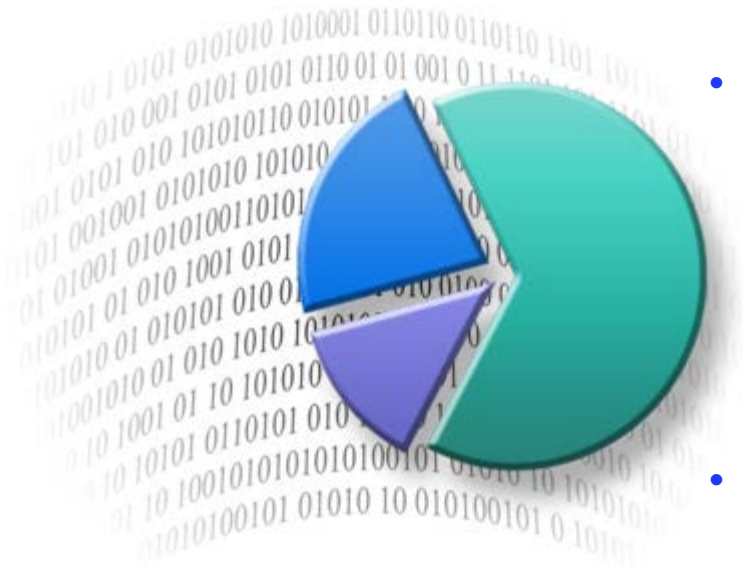
Arquitecto

Admin de datos

IT Admin

Arquitecto de SW

¿Por qué Data Profiling?

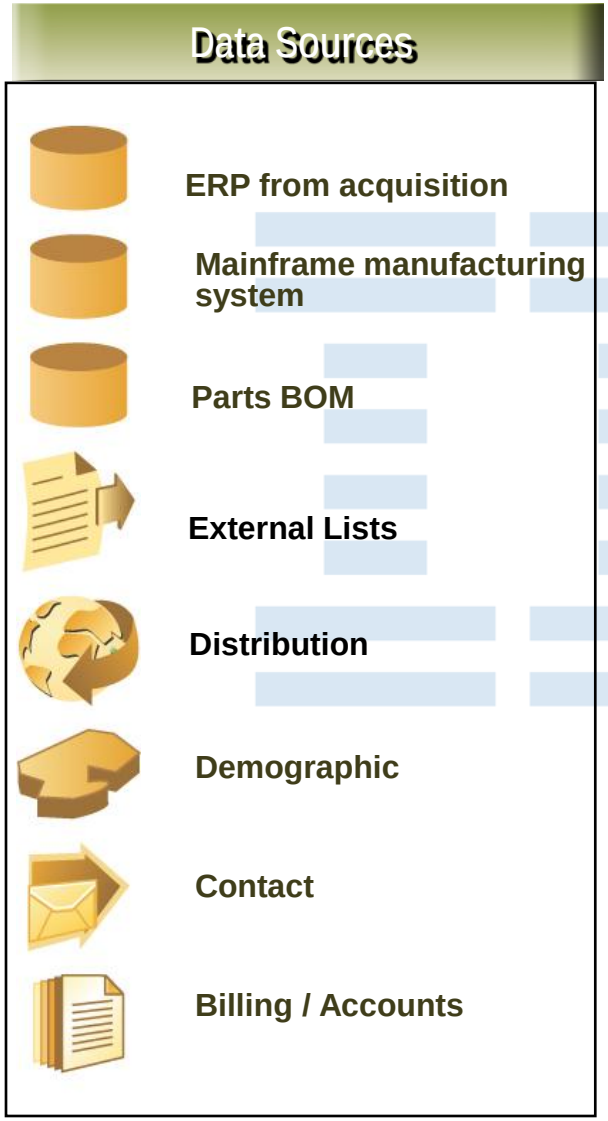


- Problemas críticos:
  - No saber que datos están realmente en los sistemas
  - Las fuentes han cambiado, o son nuevas o desconocidas
- ¿Por qué?
  - Los valores de los datos y sus relaciones son diferentes de las reglas documentadas
  - Documentación incompleta o desaparecida
  - Las fuentes de datos 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,…..

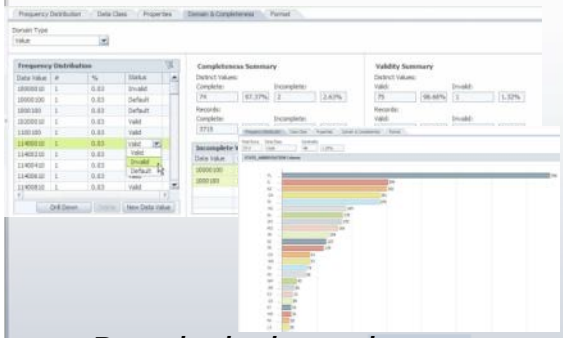
Data Profiling: IBM Information Analyzer



- 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%

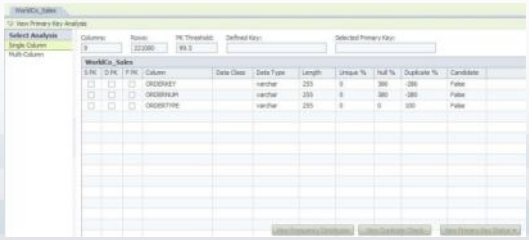
Funcionalidades de Information Analyzer

Análisis de Columnas



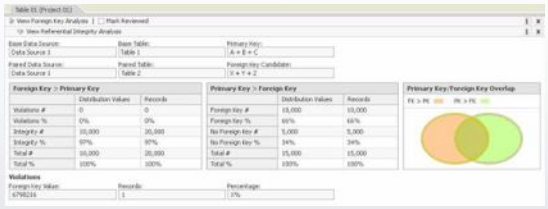
Propiedades, clases, formatos, Dominios, completitud…

Análisis de Tablas



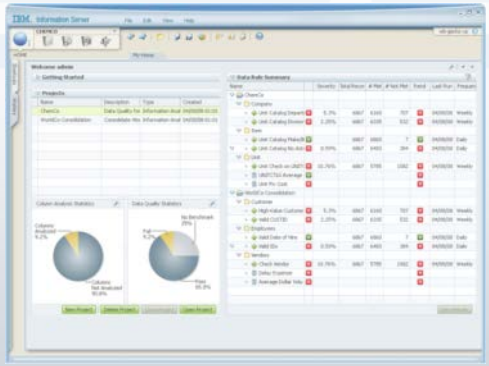
Análisis de claves Primarias Simples o Múltiples

Análisis Cruzados



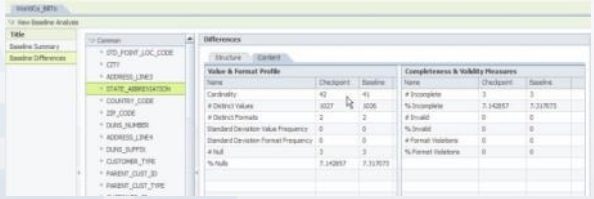
Foreign Key & Análisis multitabla

Reglas de Negocio

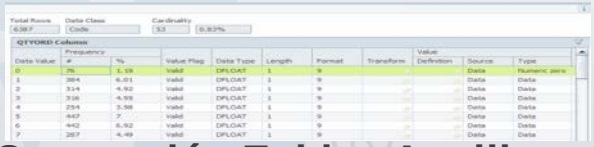


Verifica cumplimiento de reglas criticas

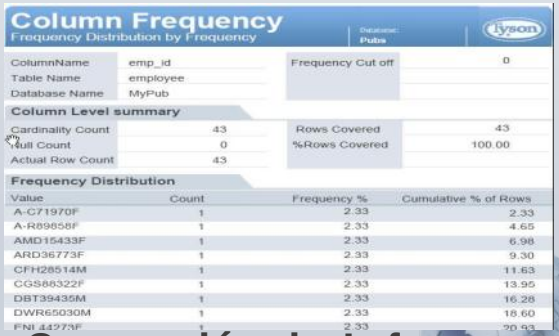
Anotaciones, marcas para revisión,…



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

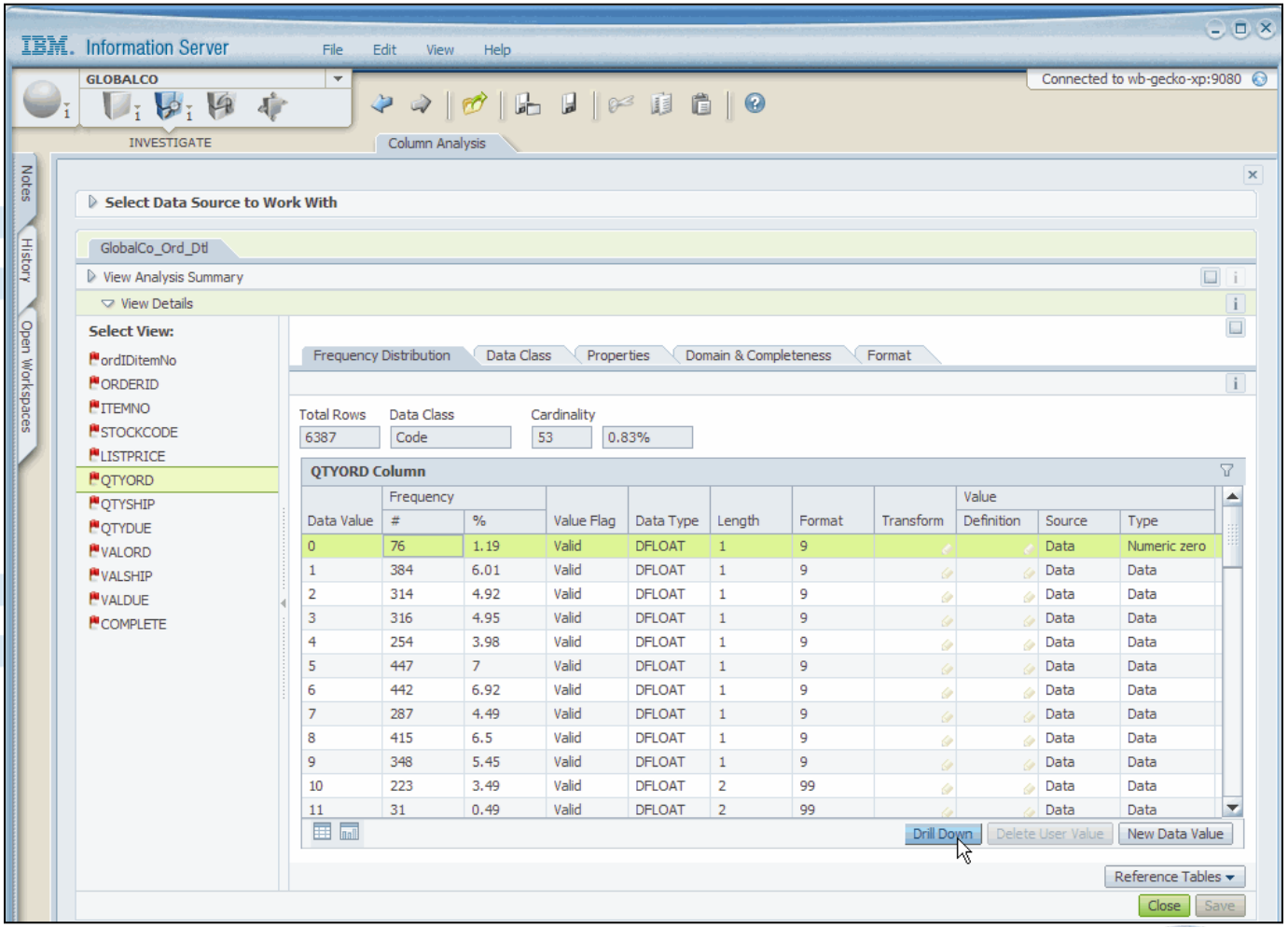
- Herramienta
- Configuración
- Importación de metadatos
- Crear proyecto y asociar fuentes de datos

Agenda

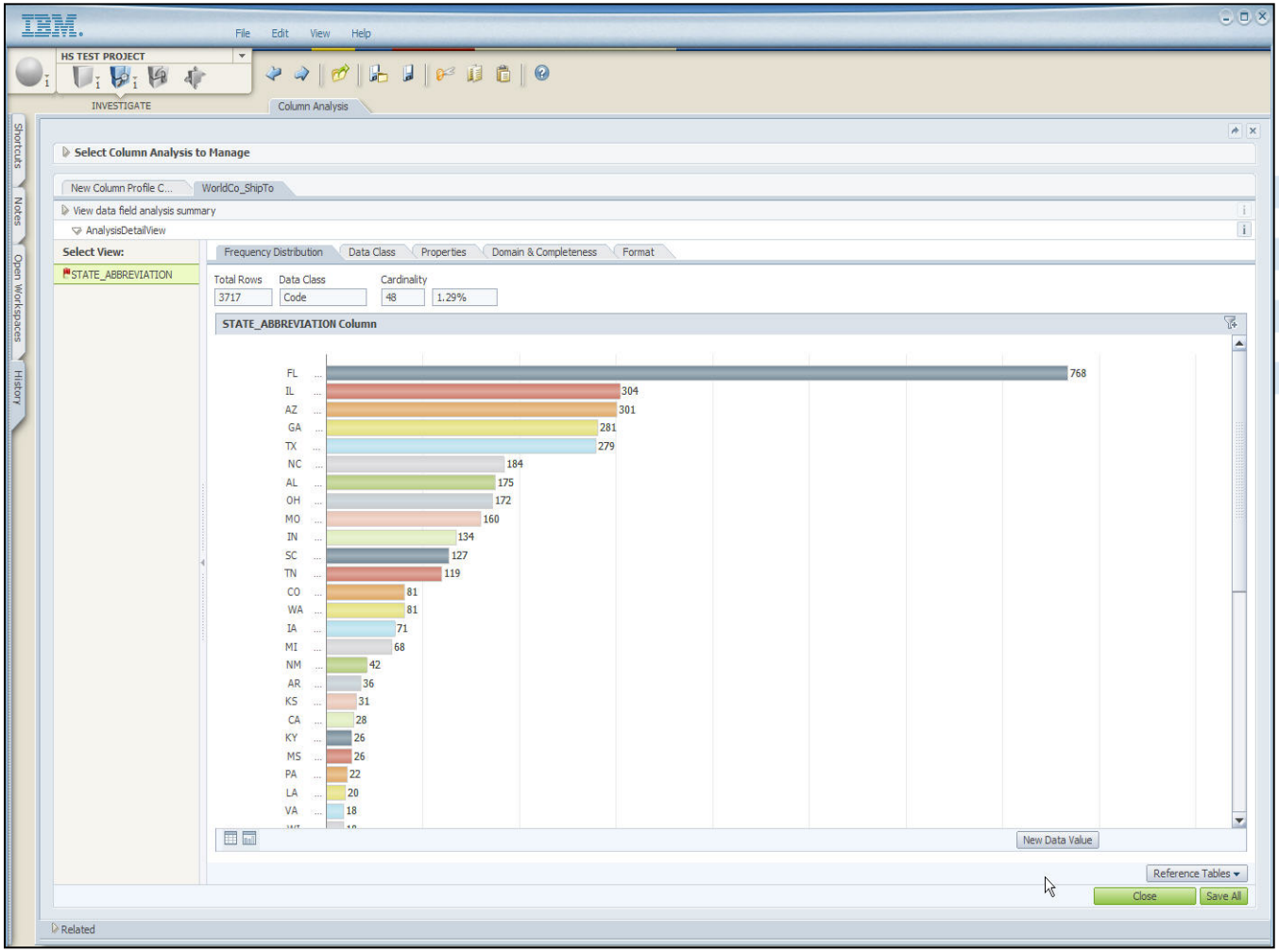
- **Visión General de IBM InfoSphere Information Server**
- **Funcionalidades de Information Analyzer**
- ***Descanso – Café***
- **Configuración e importación de los datos**
- **Análisis de columnas**
- ***Comida***
- **Identificación de claves y análisis de relaciones**
- **Integridad referencial**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

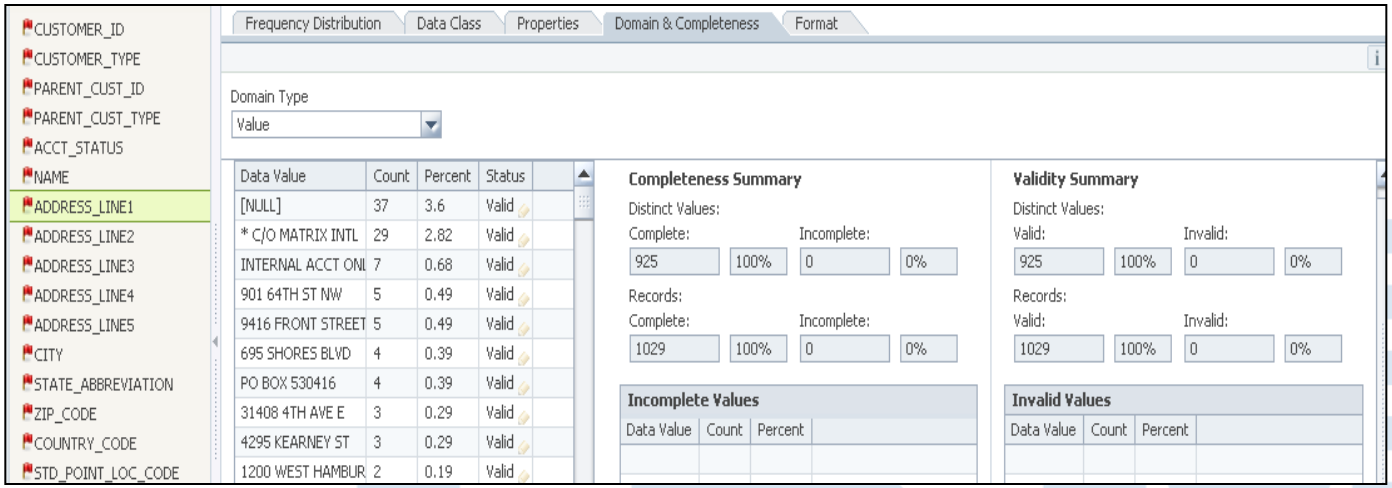


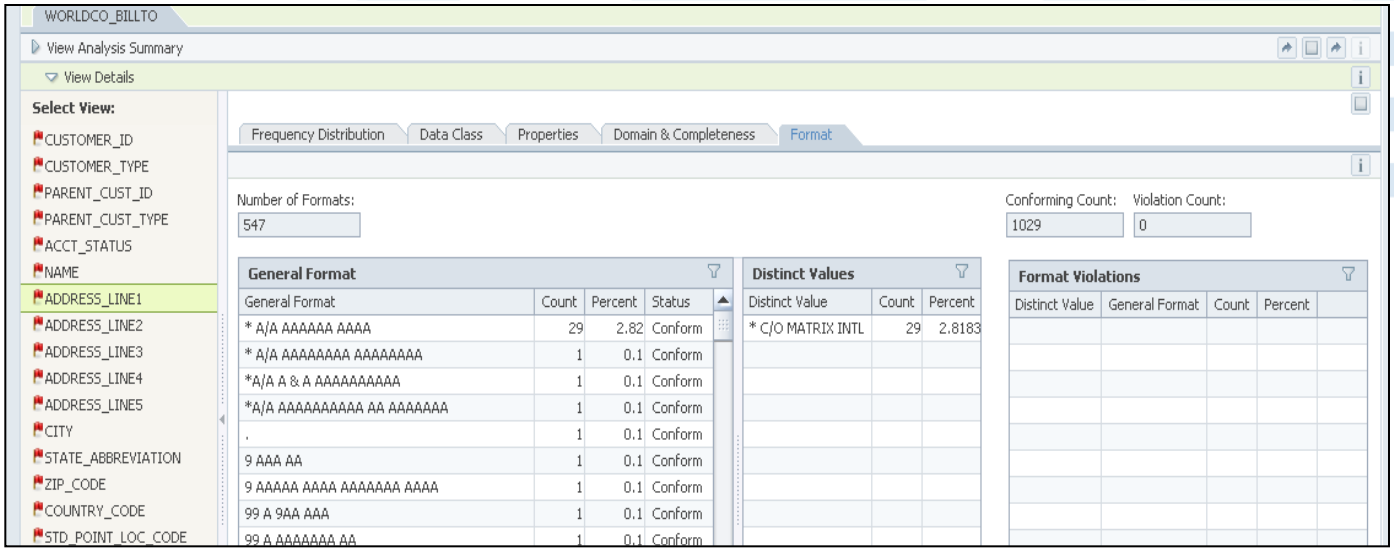
Resumen y detalle de Columna



- 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

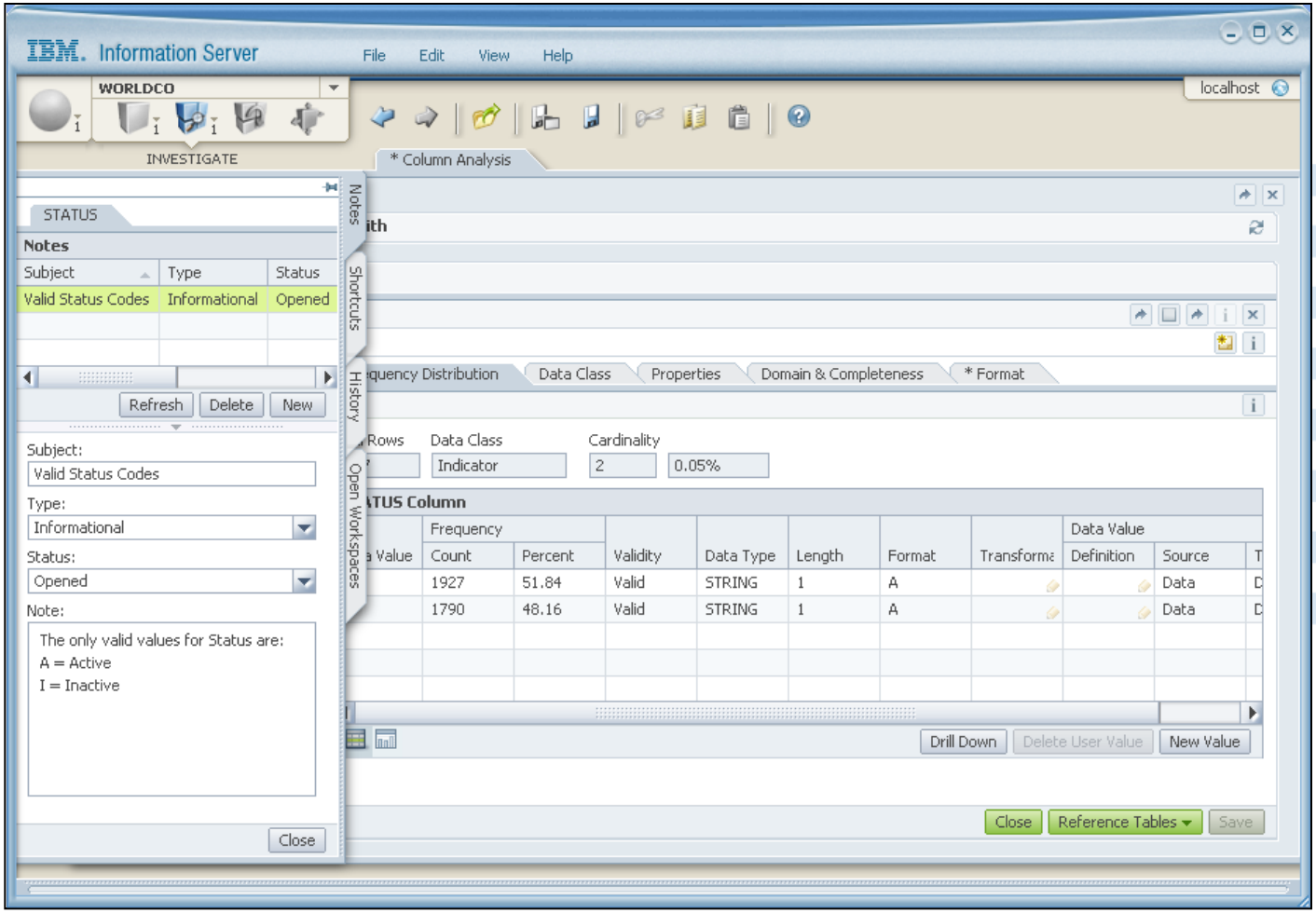
Evaluación de la calidad del dato





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

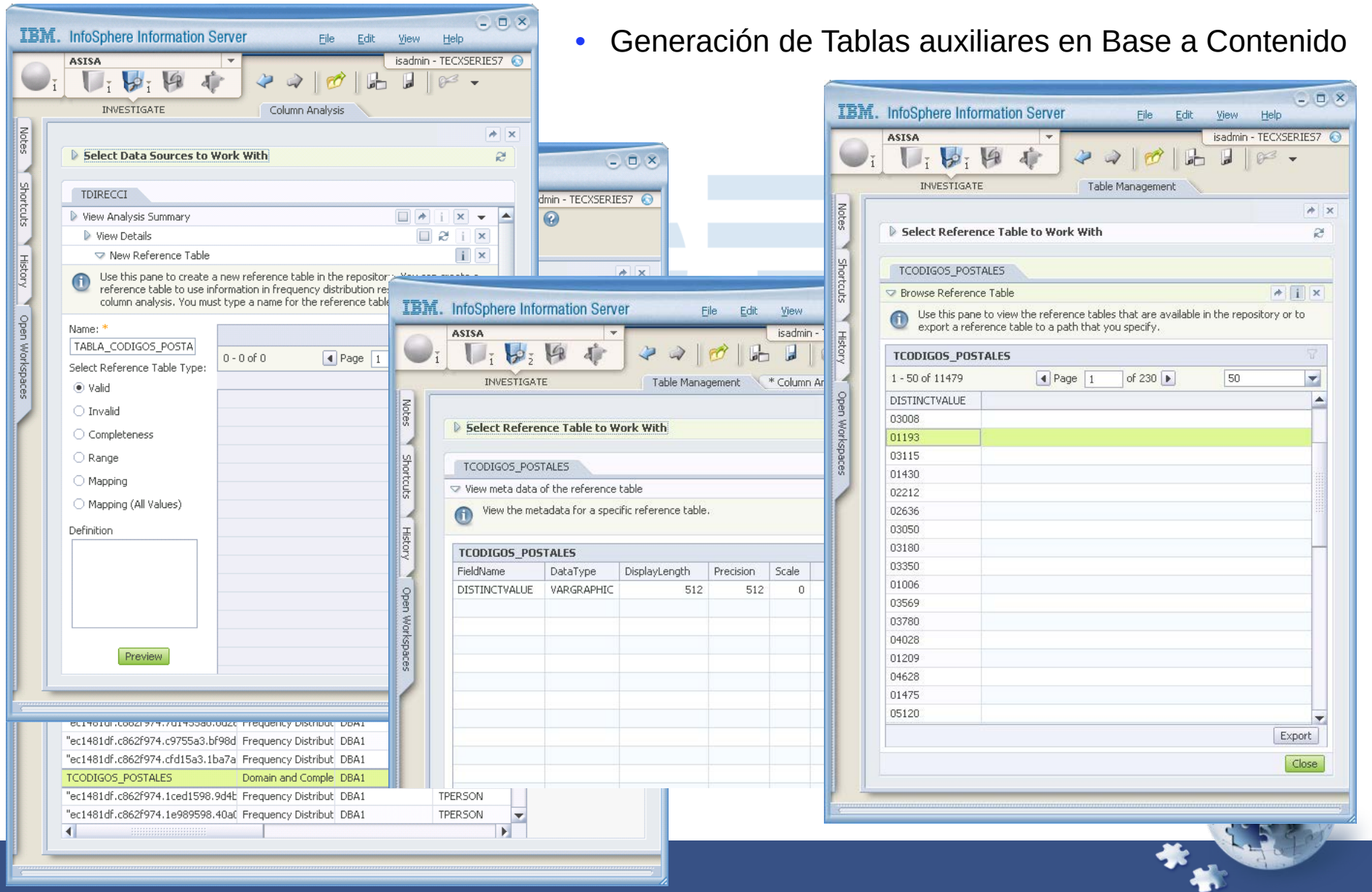
Anotaciones



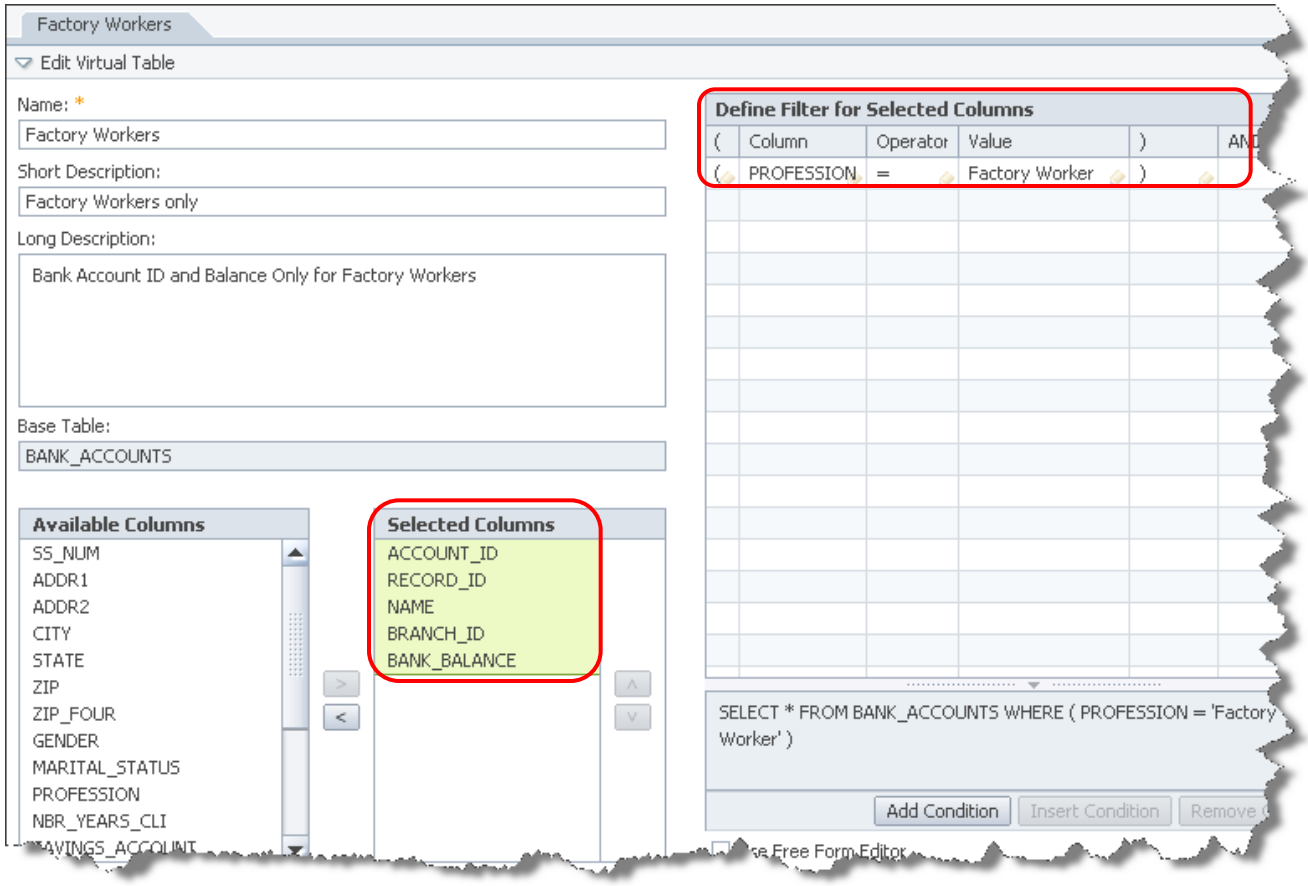
- Anotaciones analíticas en cualquier objeto
- Las anotaciones con los valores descritos son presentadas en los informes

Generación de Tablas de Referencia

- Generación de Tablas auxiliares en Base a Contenido



Selección de Datos Flexible – Tablas Virtuales



- 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

Ejercicio 2: Análisis de Columnas

LAB 2: ANÁLISIS DE COLUMNAS CON INFORMATION ANALYZER

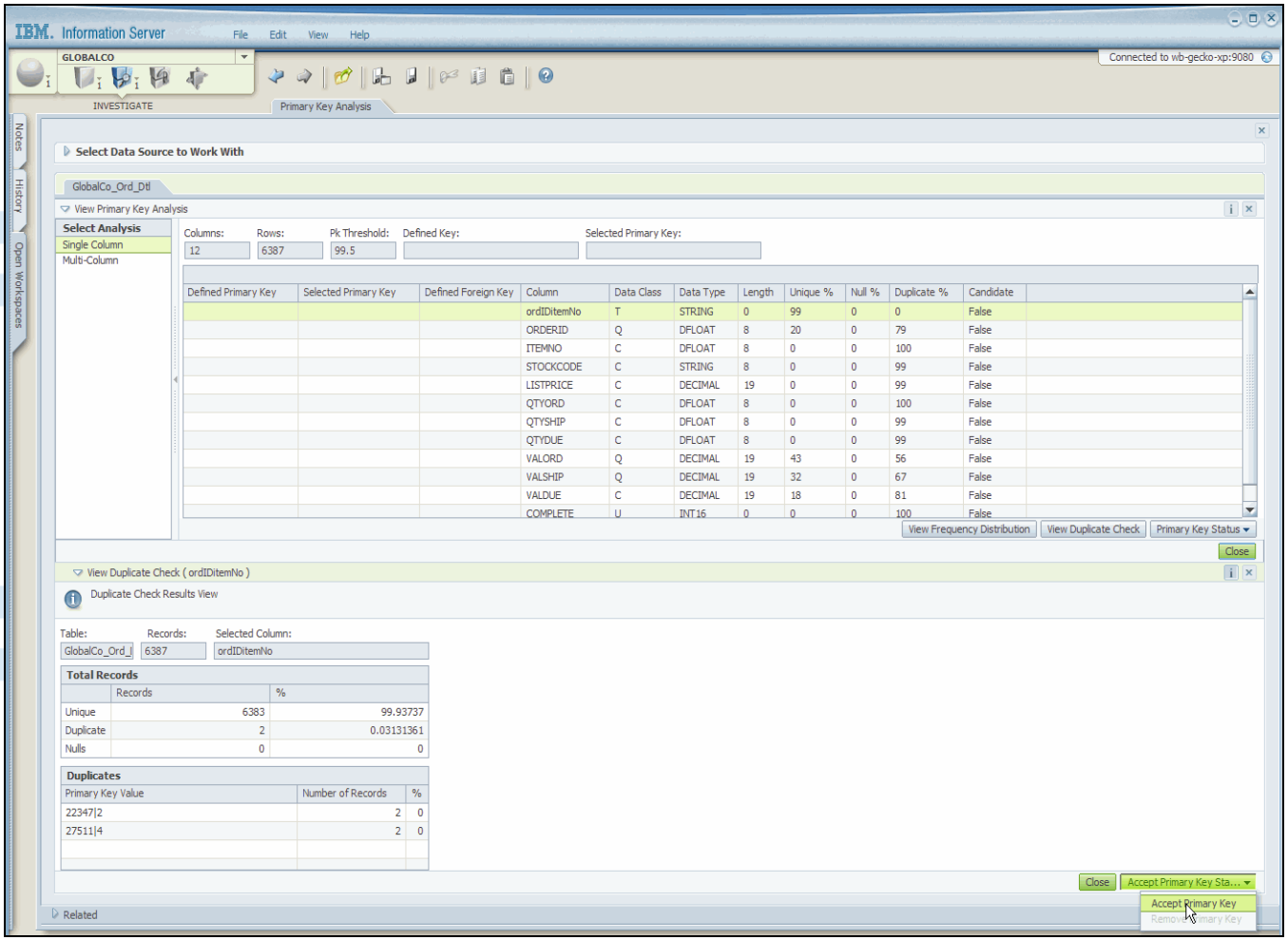
- Ejecutar análisis de columnas
- Abrir análisis de columnas
  - Ver detalles
  - Añadir notas
  - Tablas de referencia
- Tablas de gestión
- Tablas virtuales
- Columnas virtuales

Agenda

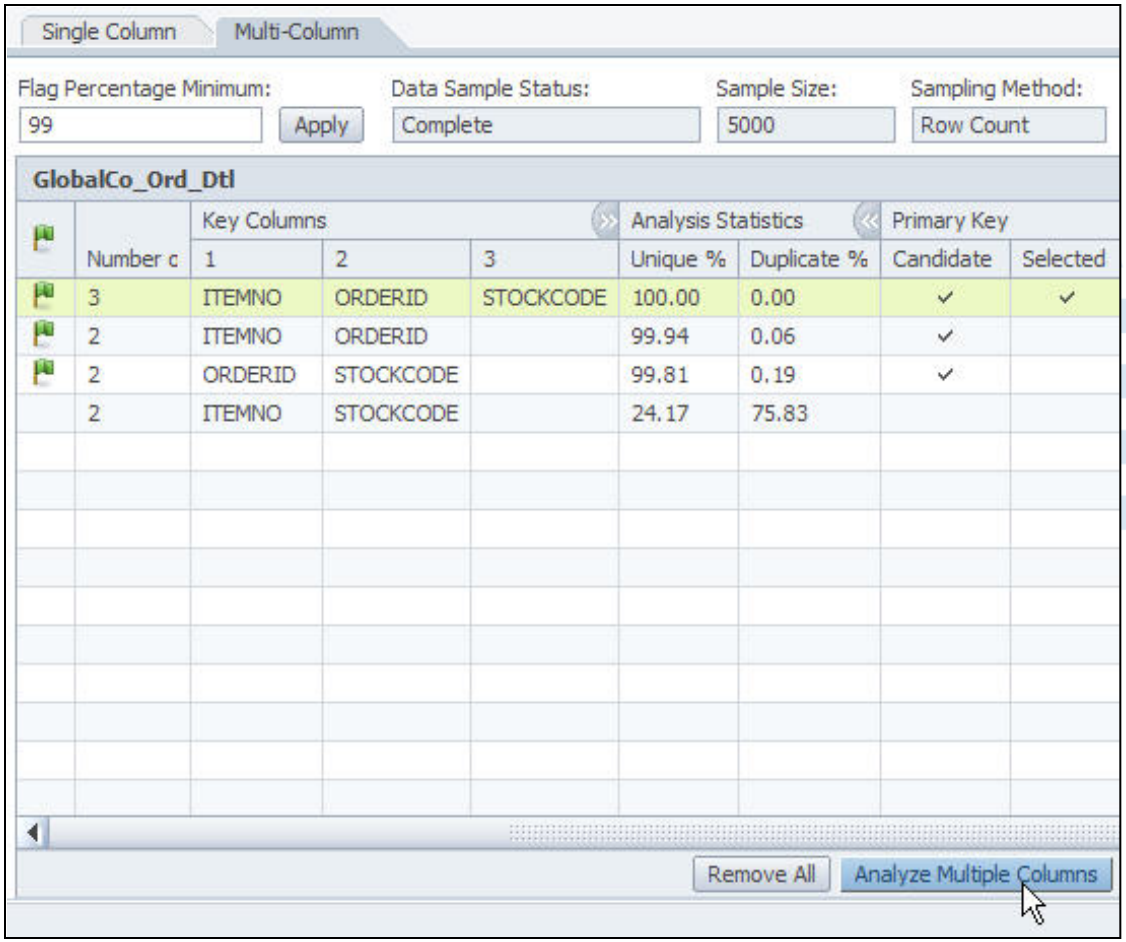
- **Visión General de IBM InfoSphere Information Server**
- **Funcionalidades de Information Analyzer**
- ***Descanso – Café***
- **Configuración e importación de los datos**
- **Análisis de columnas**
- ***Comida***
- **Identificación de claves y análisis de relaciones**
- **Integridad referencial**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

Data Profiling: Análisis de Tablas

- Claves primarias (simples o multicolumnas)
- Claves Duplicadas



Evaluación de clave primaria



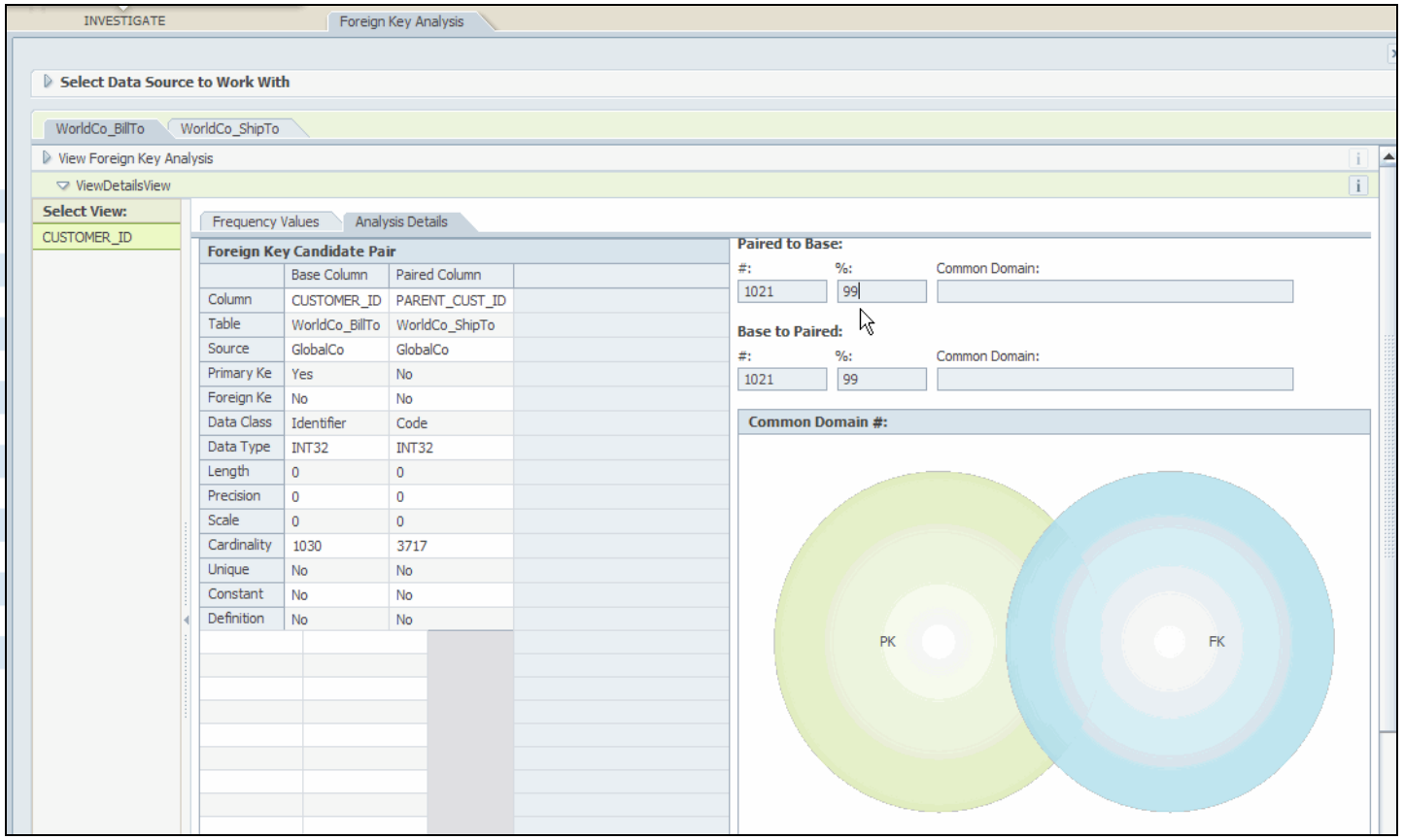
- Evaluación de la clave primaria candidata
- Columnas únicas automáticamente evaluadas durante el análisis de columnas
- Soporte a la evaluación de clave multicolumna contra muestra de datos o volumen completo
- Detalles de violaciones de clave duplicada

Agenda

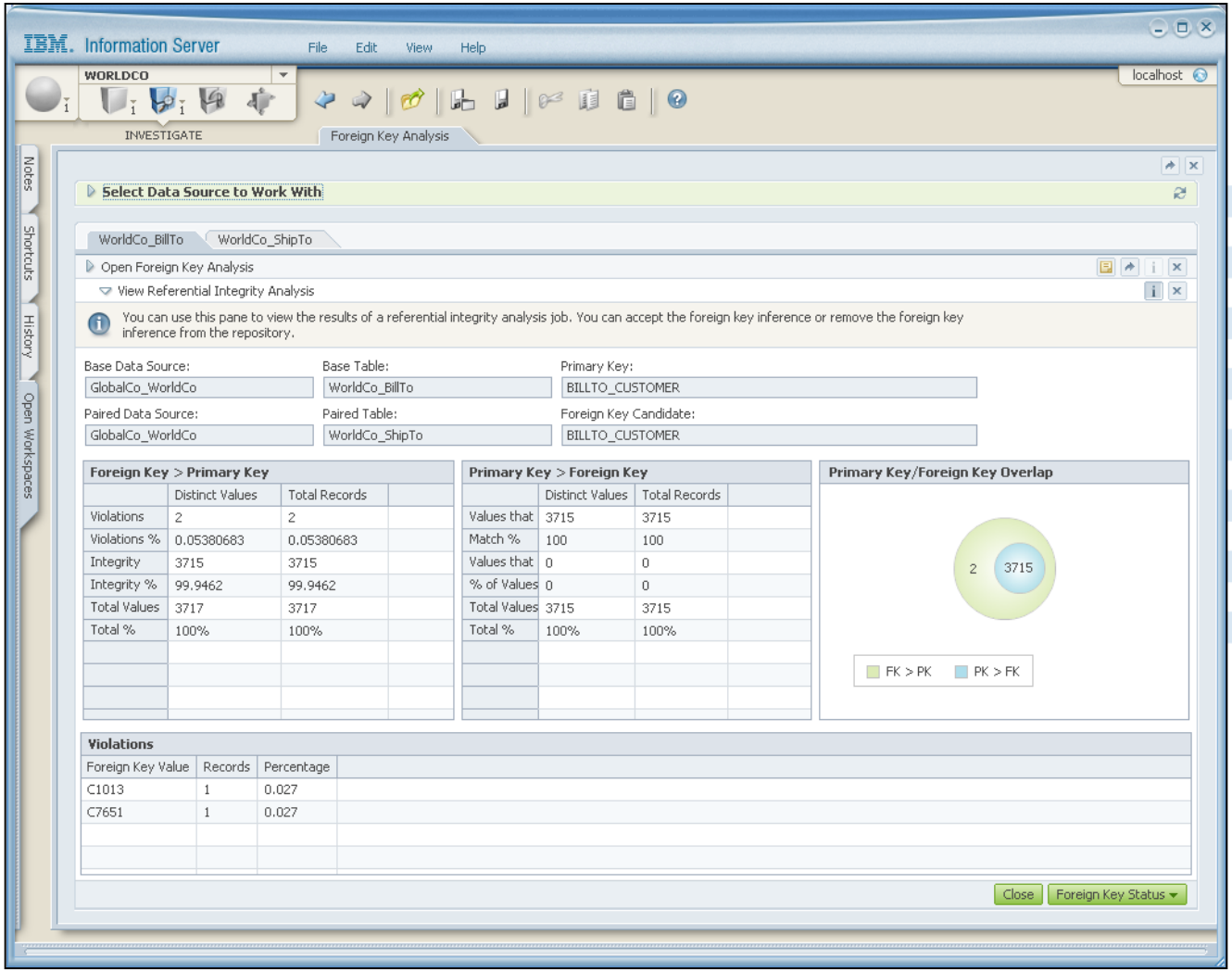
- **Visión General de IBM InfoSphere Information Server**
- **Funcionalidades de Information Analyzer**
- ***Descanso – Café***
- **Configuración e importación de los datos**
- **Análisis de columnas**
- ***Comida***
- **Identificación de claves y análisis de relaciones**
- **Integridad referencial**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

Data Profiling: Análisis a través de tablas / Foráneas

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



Relaciones entre claves e Integridad referencial



- 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

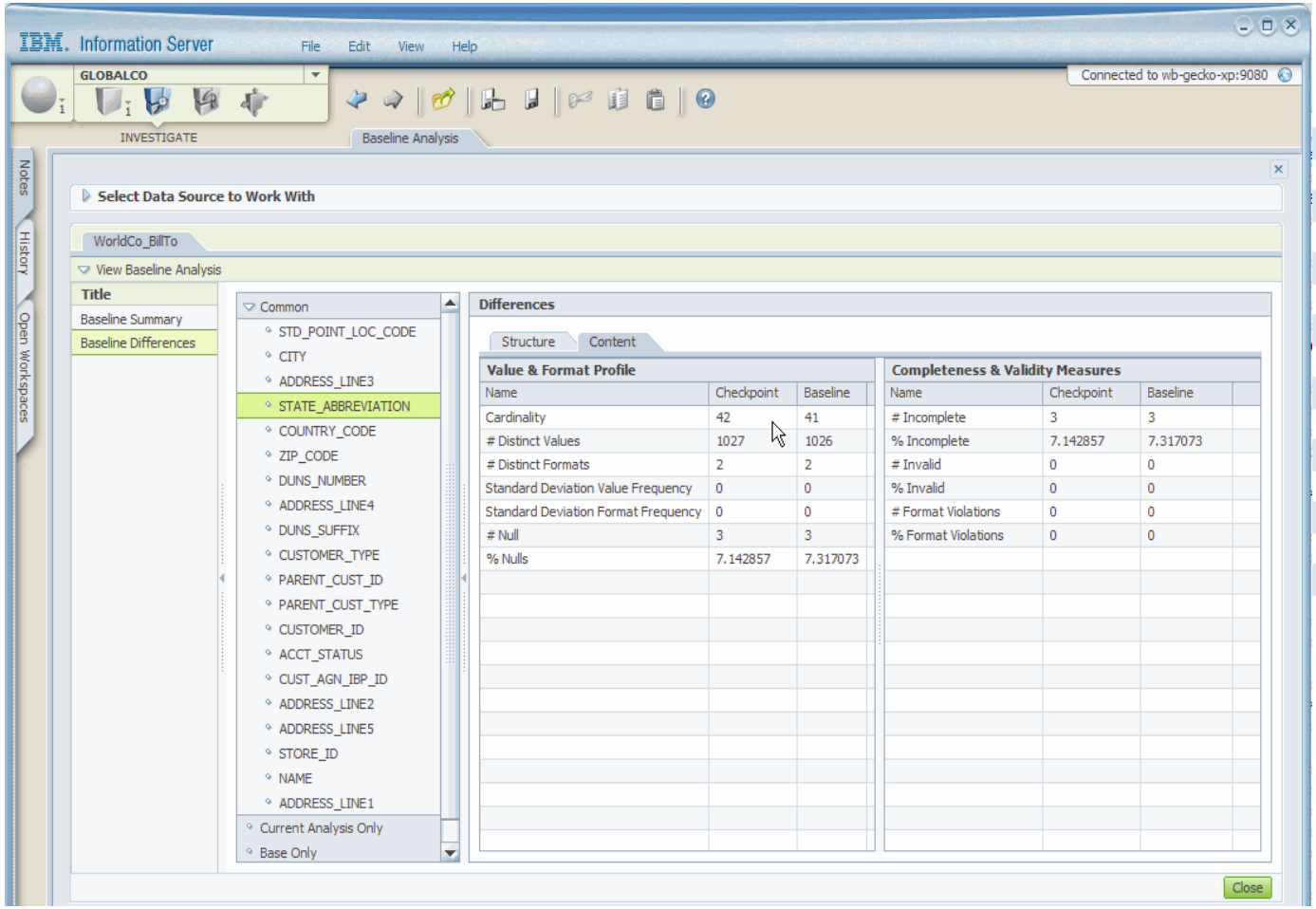
- Análisis de clave primaria
  - Clave primaria de columna única
  - Clave primaria de varias columnas
- Análisis de claves naturales
- Análisis de claves foráneas
- Ubicar elementos solapados entre dominios
- Integridad Referencial
- Asignar clave foránea

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**

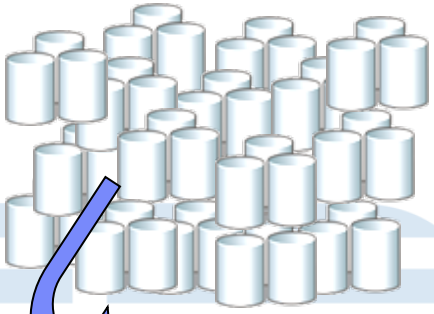
Análisis de Base

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

Data Profiling: Monitor de Cambios



Database = DB2
Schema = NewGlobalCo
Table = Ord_Dtl
Column = QTYORD
Total Records = full volume

- Ejecución programada
- Comparaciones Actual-Anterior

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: * Column Analysis Task

Job Description:

Sources

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 : 8/31/2006 15:04:57

End Date : 8/31/2006 15:04:57

Occurrence : (Not set) Edit

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

Ejercicio 4: Análisis de Base

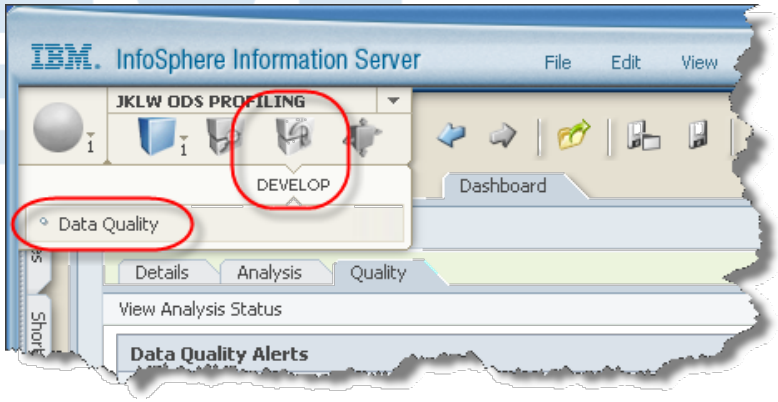
LAB 4: IDENTIFICIACIÓN DE LOS CAMBIOS SUFRIDOS POR LOS DATOS EN EL TIEMPO

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 criticas 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*



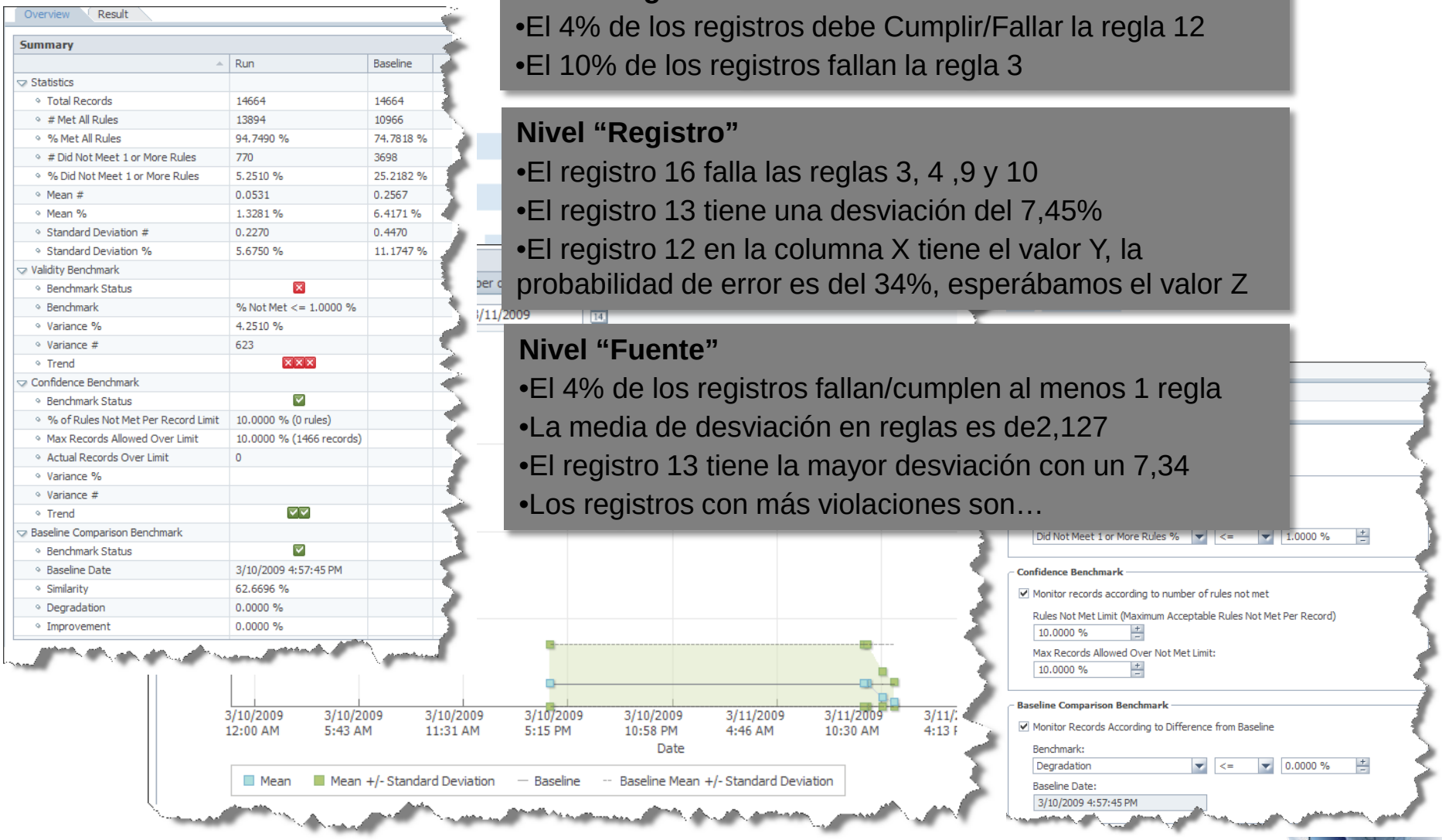
Ejemplos de reglas:
•El “Género” y “Población” deben estar en las tablas de validación
•El Número de “Seguridad Social” debe tener el formato 99/9999999/99
•Si la “Fecha de nacimiento” no está vacía y es mayor que 1900-01-01 y menor que < “HOY” el tipo de cliente es 'P'.
• Si la cuenta bancaria es válida la identificación de la sucursal es la Subdivisión de la cuenta

Soporte de Reglas Multi-nivel

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

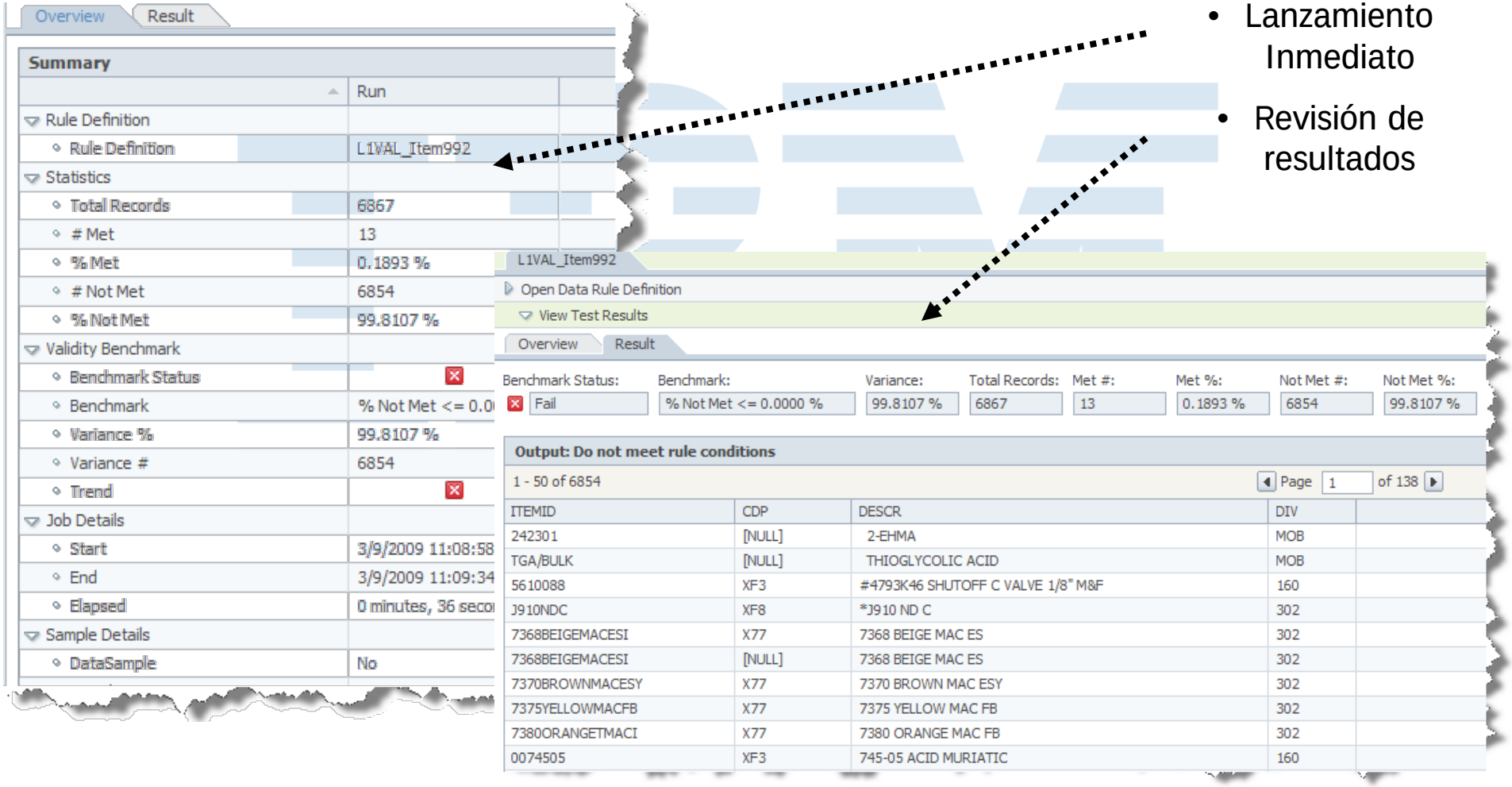
Nivel "Registro"
- El registro 16 falla las reglas 3, 4 ,9 y 10
- El registro 13 tiene una desviación del 7,45%
- El registro 12 en la columna X tiene el valor Y, la probabilidad de error es del 34%, esperábamos el valor Z

Nivel "Fuente"
- El 4% de los registros fallan/cumplen al menos 1 regla
- La media de desviación en reglas es de2,127
- El registro 13 tiene la mayor desviación con un 7,34
- Los registros con más violaciones son…



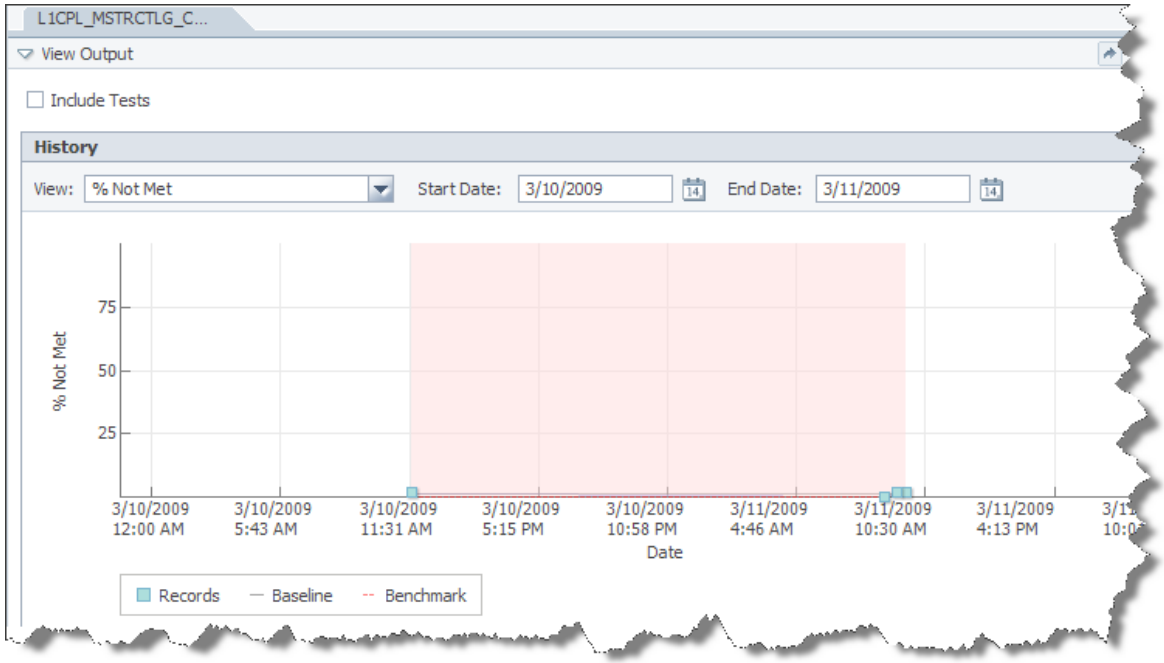
Evaluación de reglas Interactivo

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

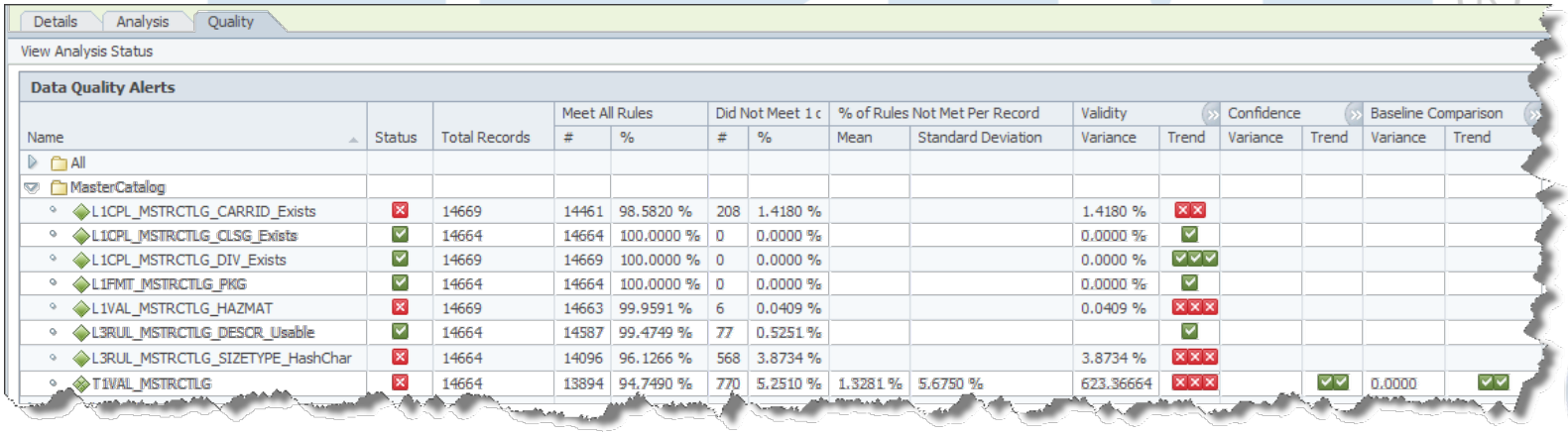


- Lanzamiento Inmediato
- Revisión de resultados

Resultados

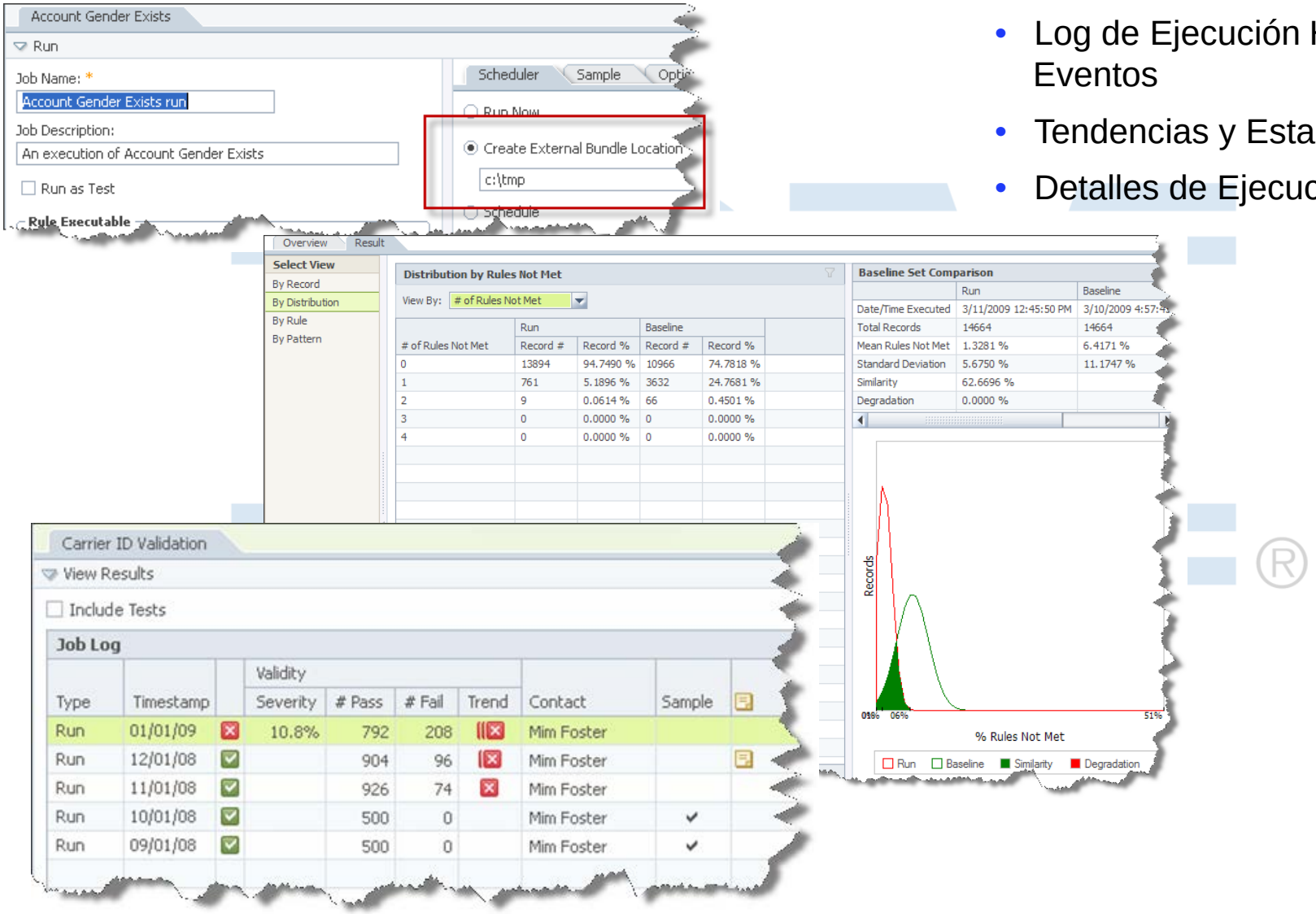


- Establecer parámetros de seguimiento por Diferencia
- Métricas de una o varias Reglas
- Organización de Métricas y normas en carpetas definidas por el usuario



Name	Status	Total Records	Meet All Rules #	Meet All Rules %	Did Not Meet 1 c #	Did Not Meet 1 c %	% of Rules Not Met Per Record Mean	% of Rules Not Met Per Record Standard Deviation	Validity Variance	Validity Trend	Confidence Variance	Confidence Trend	Baseline Comparison Variance	Baseline Comparison Trend
All														
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_SIZETYPE_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

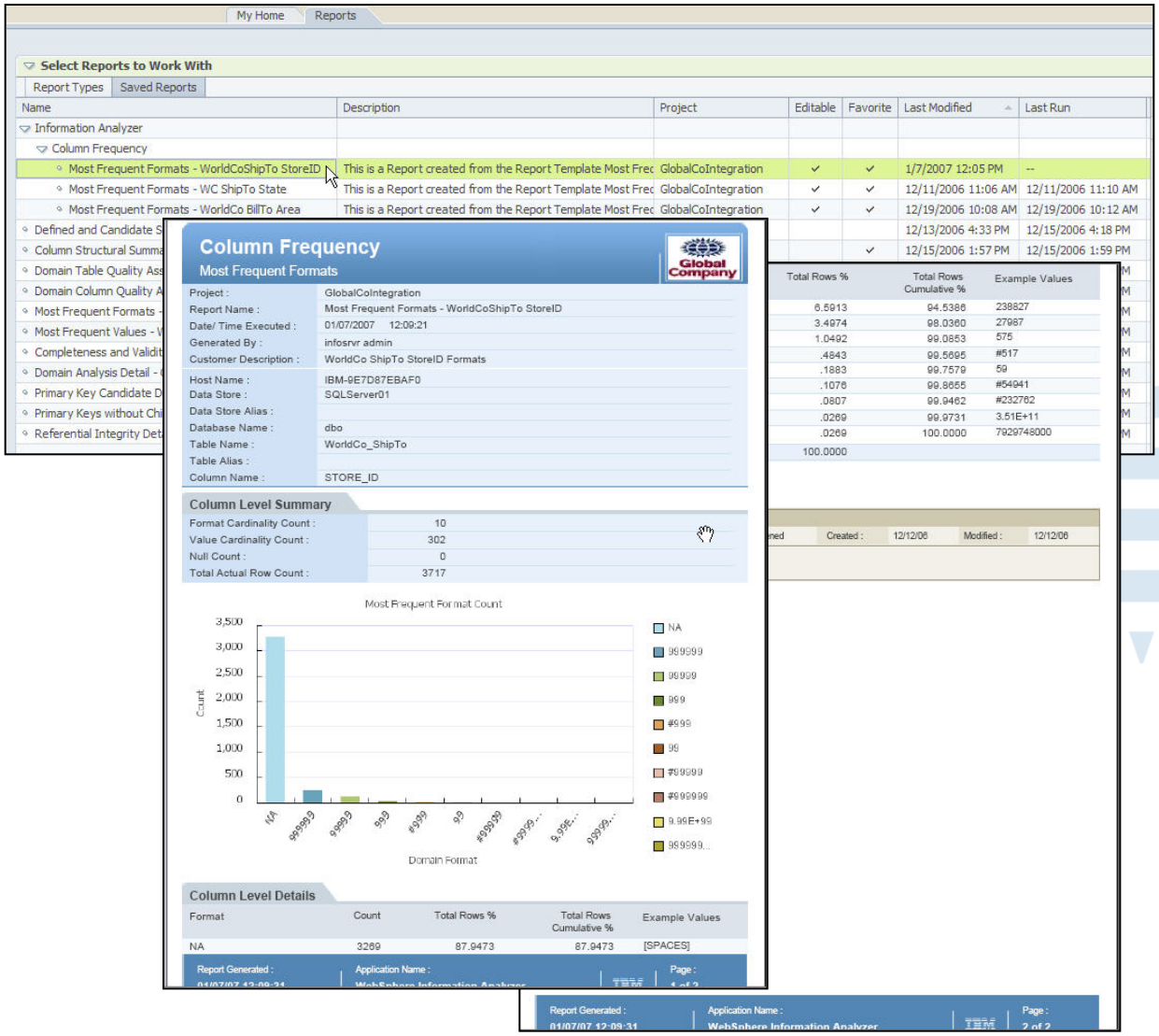


- Log de Ejecución Historico Eventos
- Tendencias y Estadísticas
- Detalles de Ejecuciones

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 multiples 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



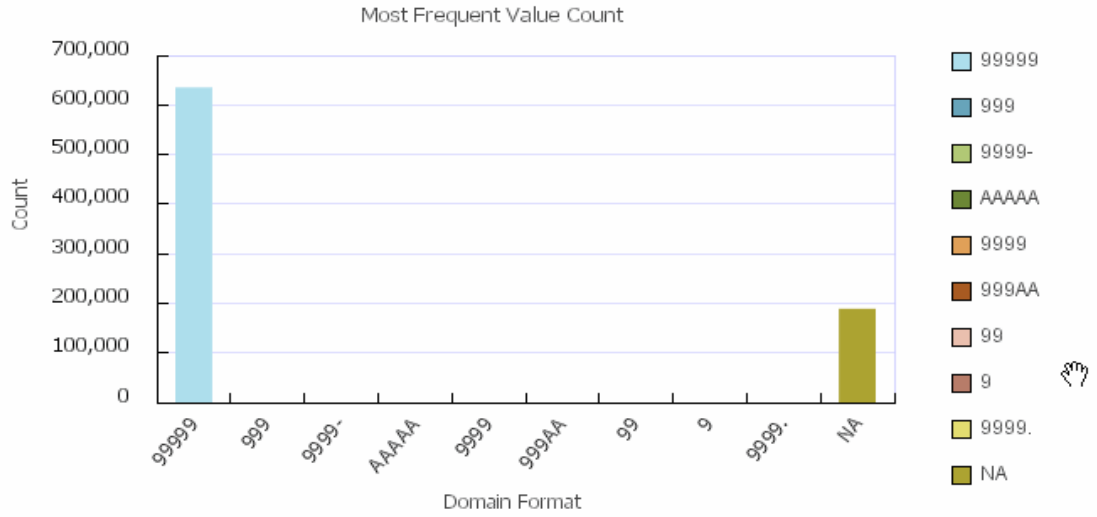
- Más de 40 plantillas de informes Ya predefinidas
- Salva informes con logos particularizados, nombres de informes y parámetros relevantes
- Notas incrustadas
- Entrega informes en UI o consola de reporting basada en navegadoror
- Soporta programación de informes

Data Profiling: Entrega de la Información

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

Column Level Summary

Format Cardinality Count	10
Value Cardinality Count	4738
Null Count	112567
Total Actual Row Count	824634

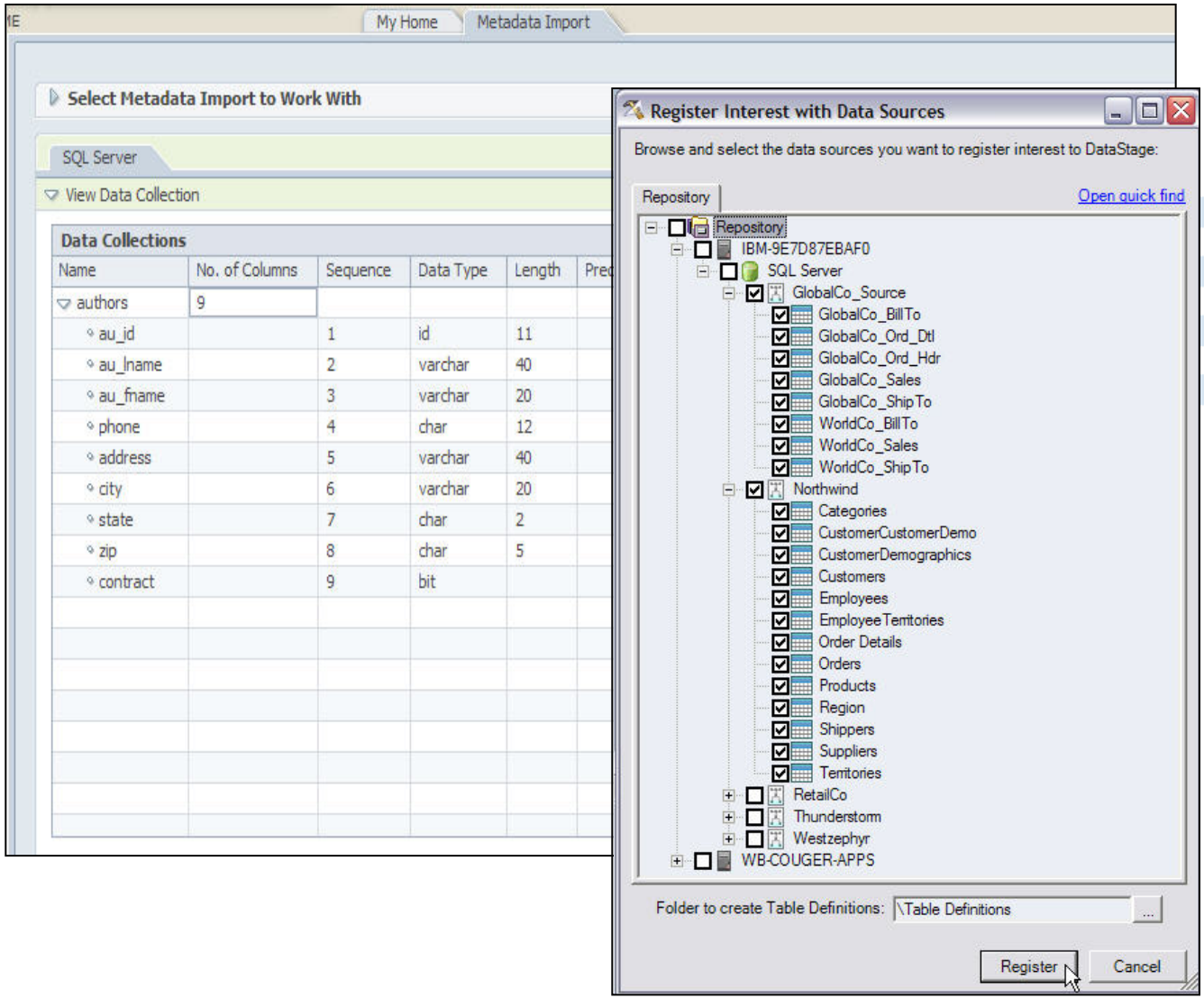


Column Level Details

Format	Count	Total Rows %	Total Rows Cumulative %	Example Value
99999	635085	77.01	77.01	
999	77	.01	77.02	

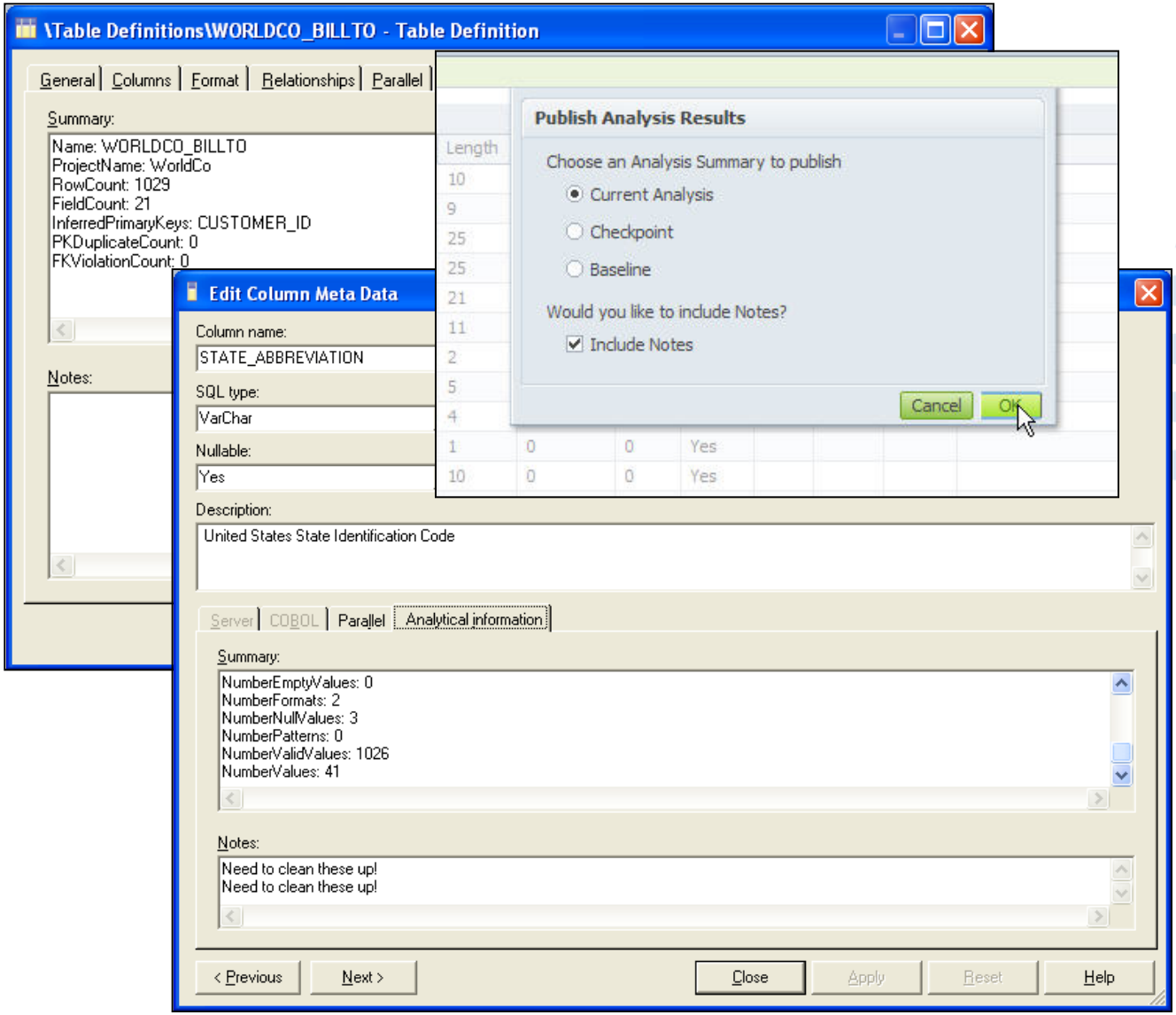
Report Generated : 08/17/06 10:02:34 | Application Name : WebSphere Information Analyzer | IBM | Page : 1 of 2

Compartición de Metadatos – Fuentes de Datos



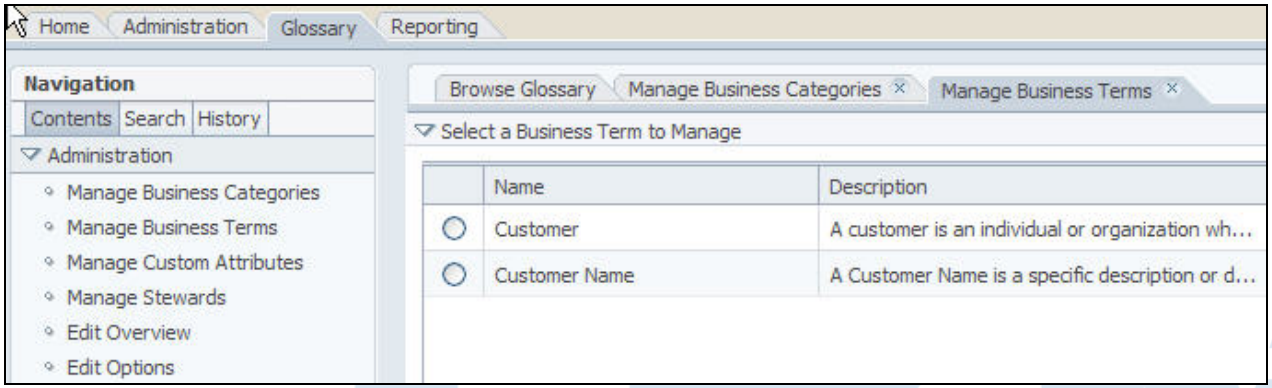
- Respositorio común para todos los metadatos
- Soporta Bases de Datos Estandar
- Los Metadatos descubiertos son compartidos por toda la Suite

Metadatos compartidos – Resultados analíticos

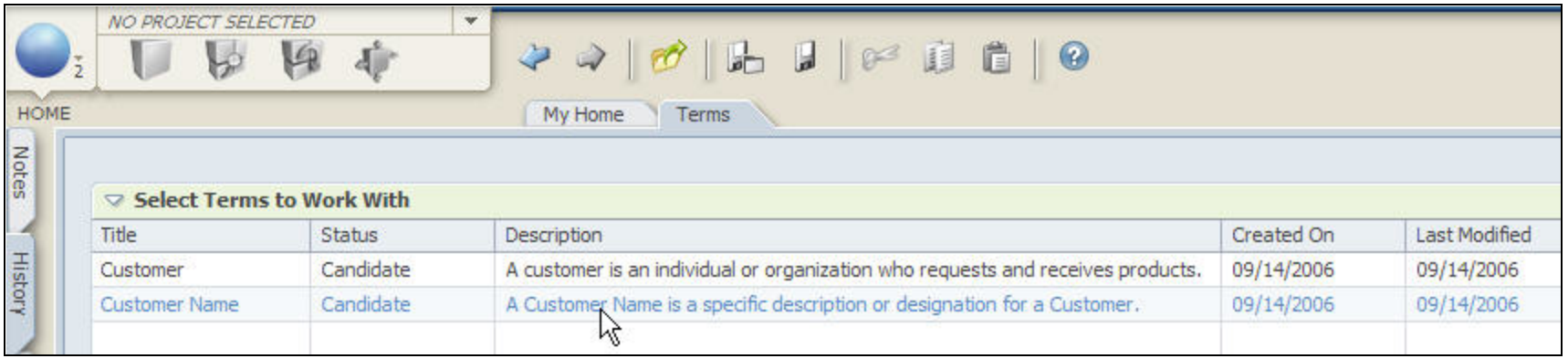


- Resultados analíticos compartidos directamente con DataStage y QualityStage
- Incluye resúmenes de tablas y columnas más anotaciones (si se selecciona)

Metadatos compartidos – Semántica de Negocio



- 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)



Ejercicio 5: Reglas de Negocio

LAB 5: DATA QUALITY

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

¡Gracias por su atención!