DATAPRIX

Published on *Dataprix* (http://www.dataprix.com)

By *Dataprix*
Created *26/12/2009 - 17:11*

Artículos de Integración de Datos de Dataprix

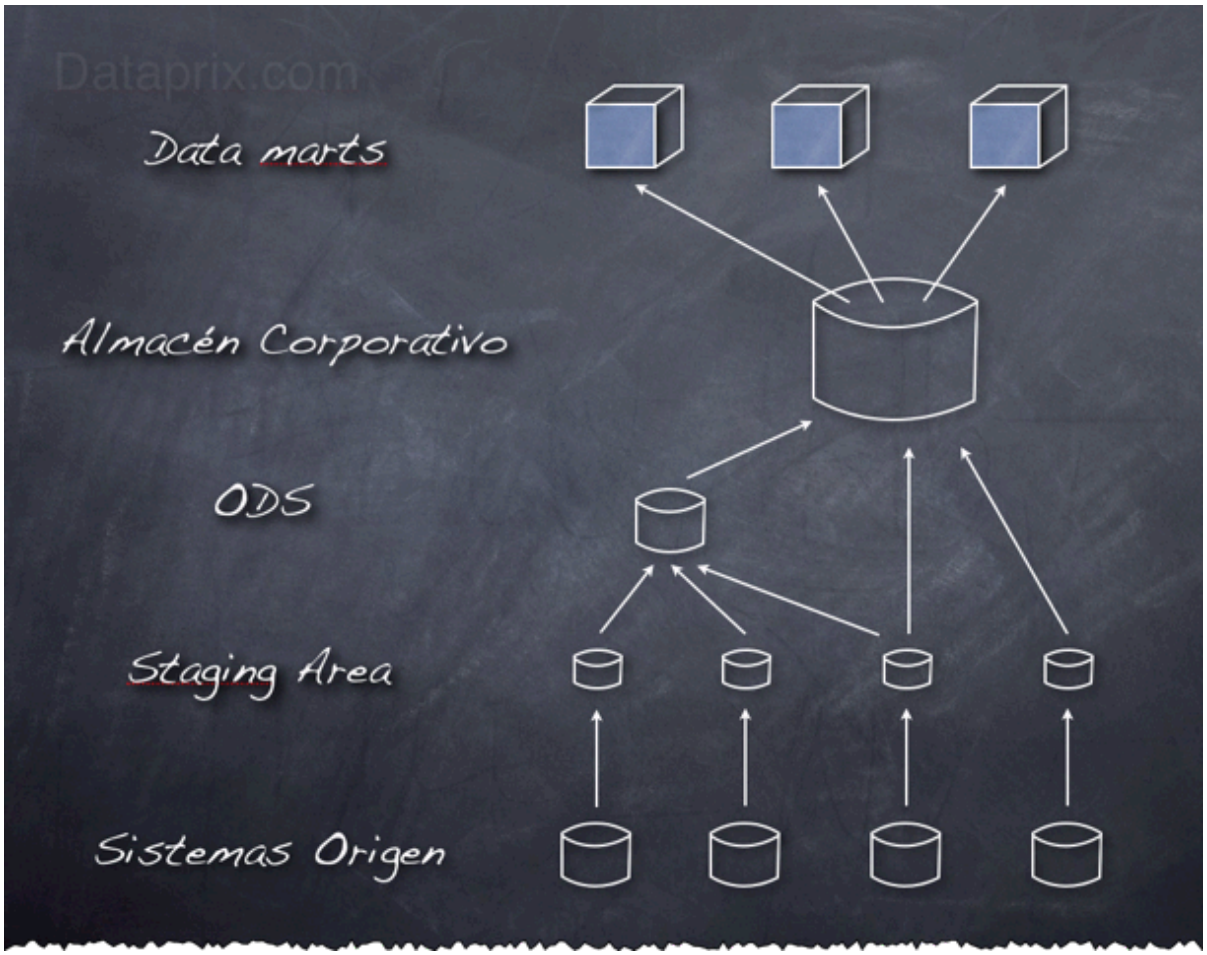
Arquitectura del Data Warehouse: áreas de datos de nuestro Almacén Corporativo

Cuando diseñamos la arquitectura de un sistema de **Data Warehouse** nos hemos de plantear los diferentes **entornos** por los que han de pasar los datos en su camino hacia su *Data mart* o *cubo* de destino. Dada la cantidad de transformaciones que se han de realizar, y que normalmente el *DWH*, además de cumplir su función de soporte a los **requerimientos analíticos**, realiza una función de **integración de datos** que van a conformar el **Almacén Corporativo** y que van a tener que ser consultados también de la manera tradicional por los *sistemas operacionales*, es muy recomendable crear diferentes *áreas de datos* en el camino entre los *sistemas origen* y las *herramientas OLAP*.

Cada una de estas áreas se distinguirá por las funciones que realiza, de qué manera se organizan los datos en la misma, y a qué tipo de necesidad puede dar servicio. El área que se encuentra 'al final del camino' es importante, pero no va a ser la única que almacene los datos que van a explotar las herramientas de *reporting*.

Tampoco hay una convención estandar sobre lo que abarca exactamente cada área, y la obligatoriedad de utilizar cada una de ellas. Cada proyecto es un mundo, e influyen muchos factores como la complejidad, el volumen de información del mismo, si realmente se quiere utilizar el Data Warehouse como almacén corporativo o Sistema Maestro de Datos, o si existen necesidades reales de soporte al reporting operacional.

Visto esto, comentaré a continuación las áreas de datos que se suelen utilizar, e iré perfilando una propuesta de arquitectura que cada uno ha de adaptar a sus necesidades o simplemente a su gusto en función de su experiencia.



Staging Area

Es un área temporal donde se recogen los datos que se necesitan de los sistemas origen. Se recogen los datos estrictamente necesarios para las cargas, y se aplica el mínimo de transformaciones a los mismos. No se aplican restricciones de integridad ni se utilizan claves, los datos se tratan como si las tablas fueran ficheros planos. De esta manera se minimiza la afectación a los sistemas origen, la carga es lo más rápida posible para minimizar la ventana horaria necesaria, y se reduce también al mínimo la posibilidad de error. Una vez que los datos están traspasados, el DWH se independiza de los sistemas origen hasta la siguiente carga. Lo único que se suele añadir es algún campo que almacene la fecha de la carga.

Obviamente estos datos no van a dar servicio a ninguna aplicación de reporting, son datos temporales que una vez hayan cumplido su función serán eliminados, de hecho en el esquema lógico de la arquitectura muchas veces no aparece, ya que su función es meramente operativa.

Hay quien considera que la Staging Area abarca más de lo que he comentado, o incluso que este area engloba todo el entorno donde se realizan los procesos de ETL, yo me decanto por su utilización sólo como área temporal.

ODS (Operational Data Store)

Como su nombre indica, este area es la que va a dar soporte a los sistemas operacionales. El modelo de datos del Almacén de Datos Operacional sigue una

estructura relacional y normalizada, para que cualquier herramienta de reporting o sistema operacional pueda consultar sus datos. Está dentro del Data Warehouse porque se aprovecha el esfuerzo de integración que supone la creación del Almacén de Datos Corporativo para poder atender también a necesidades operacionales, pero no es obligatorio, y ni siquiera es algo específico del Business Intelligence, los ODS ya existían antes de que empezáramos a hablar de BI y de DWH.

No almacena datos históricos, muestra la imagen del momento actual, aunque eso no significa que no se puedan registrar los cambios.

Los datos del ODS se recogen de la Stage Area, y aquí sí que se realizan transformaciones, limpieza de datos y controles de integridad referencial para que los datos estén perfectamente integrados en el modelo relacional normalizado.

Hay que tener en cuenta que la actualización de los datos del ODS no va a ser instantánea, los cambios en los datos de los sistemas origen no se verán reflejados hasta que finalice la carga correspondiente. Es decir, que se irán actualizando los datos cada cierto tiempo, cosa que hay que explicar a los usuarios, porque los informes que se lancen contra el ODS casi nunca podrán estar tan 'al minuto' como los que existan en el sistema origen. Lo que sí se puede hacer es definir una mayor frecuencia de carga para el ODS que para el Almacén Corporativo. Si es necesario, se puede refrescar el ODS cada 15 minutos, y el resto cada día, por ejemplo.

Almacén de Datos Corporativo

El Almacén de Datos Corporativo sí que contiene datos históricos, y está orientado a la explotación analítica de la información que recoge. Las herramientas DSS o de reporting analítico atacarán principalmente a los Data marts, pero también se pueden realizar consultas directamente contra el Almacén de Datos Corporativo, sobretodo cuando sea necesario mostrar a la vez información que se encuentre en diferentes Datamarts.

En él se almacenan datos que pueden provenir tanto de la Staging Area como del ODS. Si ya hemos realizado procesos de transformación e integración en el ODS no los vamos a repetir para pasar los mismos datos al Almacén Corporativo. Lo que no se pueda recoger desde el ODS sí que habrá que ir a buscarlo a la Staging Area.

El esquema se parece al de un modelo relacional normalizado, pero en él ya se aplican técnicas de desnormalización. No debería contener un número excesivo de tablas ni de relaciones ya que, por ejemplo, muchas relaciones jerárquicas que en un modelo normalizado se implementarían con tablas separadas aquí ya deberían crearse en una misma tabla, que después representará una dimensión. Otra particularidad es que la mayoría de las tablas han de incorporar campos de fecha para controlar la fecha de carga, la fecha en que se produce un hecho, o el periodo de validez del registro.

Si el Data Warehouse no es demasiado grande, o el nivel de exigencia no es muy elevado en cuanto a los requerimientos 'operacionales', para simplificar la estructura se puede optar por prescindir del ODS, y si es necesario adecuar el Almacén de Datos Corporativo para servir a los dos tipos de reporting. En este caso, el área resultante sería el DWH Corporativo, pero a veces también se le llama ODS.

Data marts

Y por fin llegamos a la última área de datos, que es el lugar donde se crean los Data marts. Éstos se obtienen a partir de la información recopilada en el área del Almacén

Corporativo. Cada Data Mart es como un subconjunto de este almacén, pero orientado a un tema de análisis, normalmente asociado a un departamento de la empresa.

Los Data marts se diseñan con estructura multidimensional, cada objeto de análisis es una tabla de hechos enlazada con diversas tablas de dimensiones. Si se diseñan siguiendo el Modelo en Estrella habrá prácticamente una tabla para cada dimensión, es la versión más desnormalizada. Si se sigue un modelo de Copo de Nieve las tablas de dimensiones estarán menos desnormalizadas y para cada dimensión se podrán utilizar varias tablas enlazadas jerárquicamente.

Este área puede residir en la misma base de datos que las demás si la herramienta de explotación es de tipo ROLAP, o también puede crearse ya fuera de la BD, en la estructura de datos propia que generan las aplicaciones de tipo MOLAP, más conocida como los cubos multidimensionales.

El paso del anterior área de datos a esta ha de ser bastante simple, cosa que además proporciona una cierta independencia sobre el software que se utiliza para el reporting analítico. Si por cualquier razón es necesario cambiar la herramienta de OLAP habría que hacer poco más que redefinir los metadatos y regenerar los cubos, y si el cambio es entre dos de tipo ROLAP ni siquiera esto último sería necesario. En cualquier caso, las áreas anteriores no tienen porqué modificarse.

Coméntalo en el foro [1]

Data profiling con SQL Server 2008

Una de las múltiples mejoras que aporta **SQL Server 2008** en la parte de *ETL* con Integration Services es su capacidad para realizar **perfilado de datos** con su nueva **Data Profile Task**.

El *data profiling* es una de las primeras tareas que se suelen abordar en procesos Calidad de Datos, y consiste en realizar un primer análisis sobre los datos de origen, normalmente sobre tablas, con el objetivo de empezar a conocer su estructura, formato y nivel de calidad. Se hacen consultas a nivel de tabla, columna, relaciones entre columnas, e incluso relaciones entre tablas.

La Data Profile Task de *SSIS* funciona seleccionando una tabla de una base de datos *SQLServer 2000* o superior (no sirven otras bases de datos), las opciones de perfilado que se quiera realizar sobre los datos de la tabla, y un fichero XML donde se almacenarán los resultados cuando se ejecute la misma. Es realmente sencillo.

Se pueden seleccionar hasta 8 tipos de perfilado, 5 a nivel de columna y 3 a nivel de varias columnas.

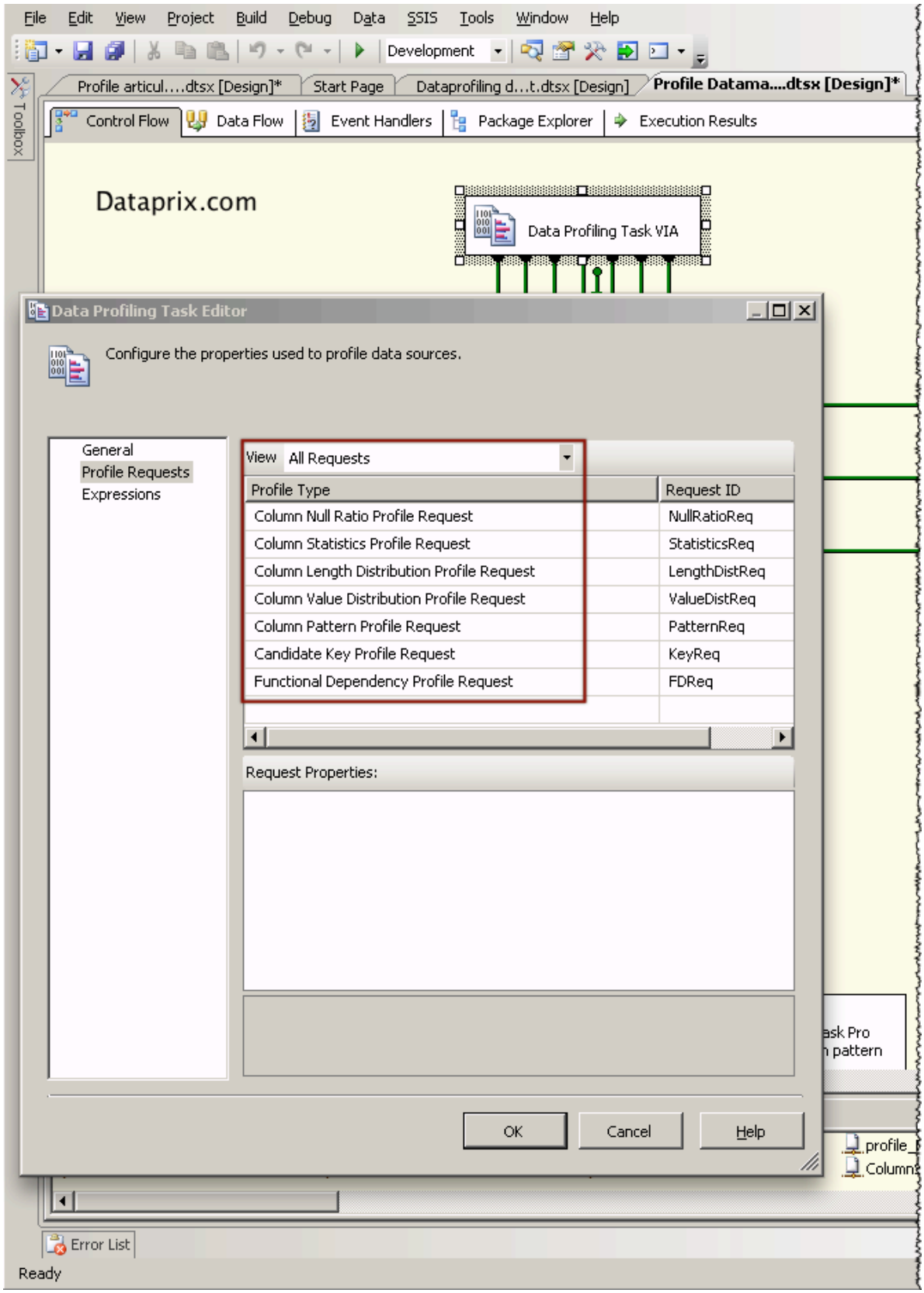
Perfilados a nivel de columna:

- Distribución de la *longitud de los valores*
- Porcentaje de *valores nulos*
- *Patrones*, expresados mediante expresiones regulares

- *Estadísticas* de columna: mínimo, máximo, media o desviación standard
- *Distribución de los valores*, valores diferentes y porcentaje de aparición de cada uno sobre el total de filas

Perfilados a nivel multicolumna:

- Claves candidatas, qué columnas podrían ser clave primaria de la tabla
- Dependencia funcional, los valores de una columna pueden depender de los de otra
- Inclusión de valores, que columnas podrían ser claves foráneas de otras

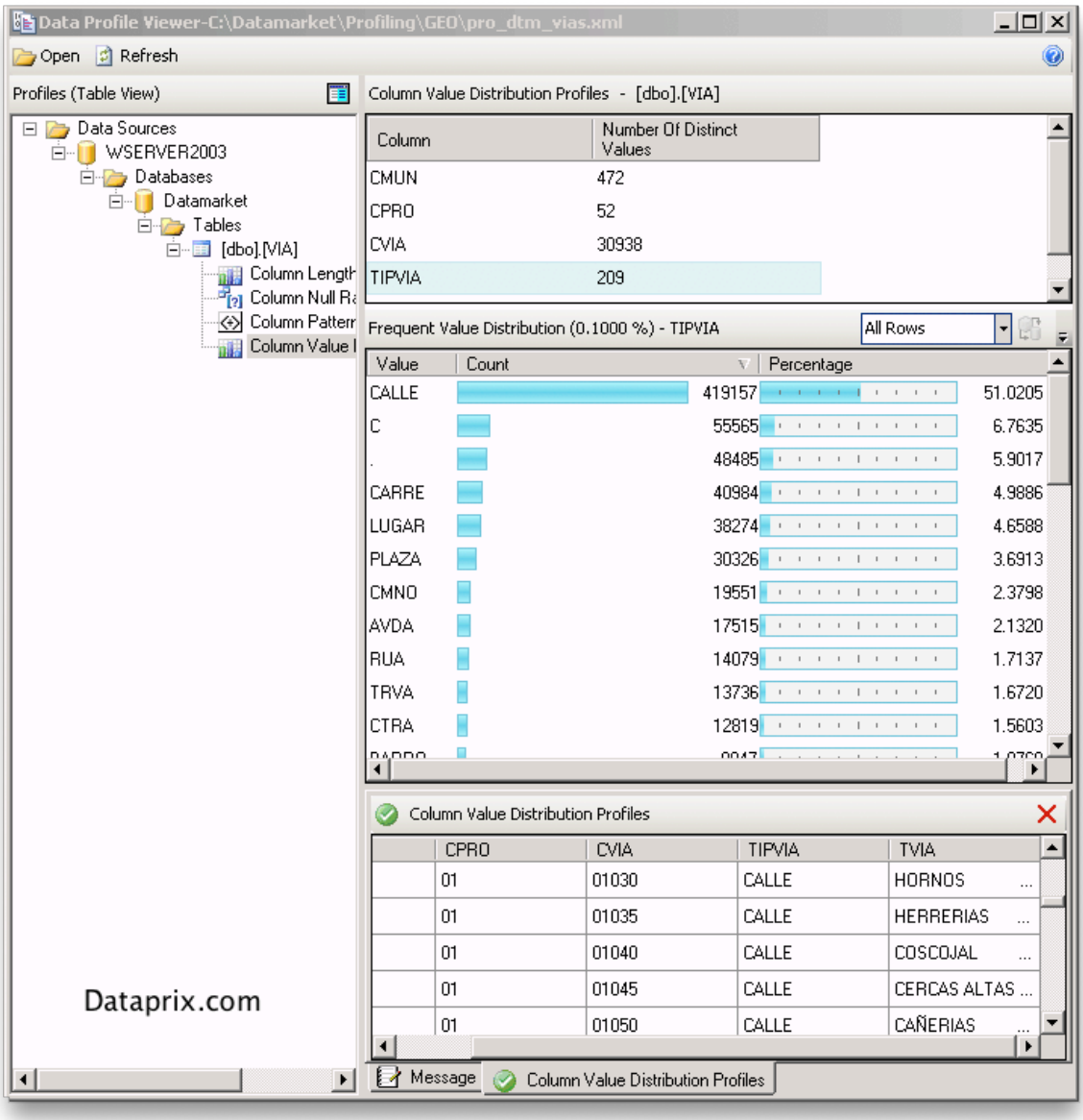


Tras ejecutar la tarea se genera un fichero XML en la ubicación elegida donde se almacena toda la información resultado del análisis. Para poder examinar estos resultados SQL Server proporciona la aplicación Data Profile Viewer que en una

instalación normal sobre la unidad C debería encontrarse en este directorio:

C:\Archivos de programa\Microsoft SQL Server\100\DTS\Binn\DataProfileViewer.exe

Sólo hay que seleccionar el XML generado por la tarea de SSIS y comenzar a explorar los resultados:



Para obtener información más detallada se puede consultar el apartado Tarea de generación de perfiles de datos [2] de la documentación en linea de *Microsoft Technet.*

También está muy bien comentada esta tarea en los artículos de *SQLServer Performance* SSIS New Features in SQL Server 2008 - Part 3 [3] y Using The Data Profiler Task and FTP Task in SQL Server 2008 Integration Services [4]

Coméntalo en el foro [5]

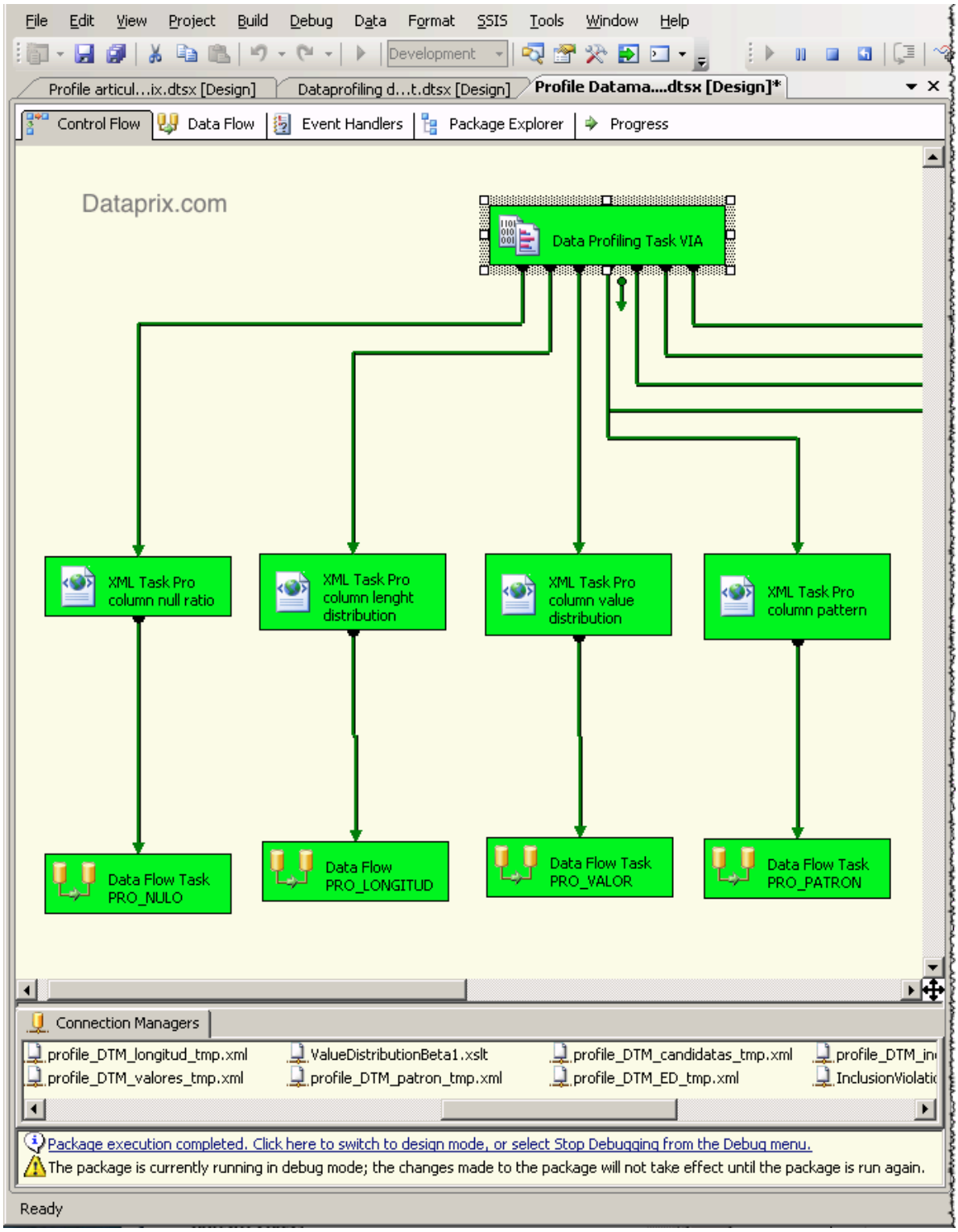
Data profiles de SQL Server IS almacenados en tablas

La tarea de **Data Profile de SQL Server Information Services** almacena los resultados del perfilado en un *documento XML* que se puede examinar con el *Data Profile Viewer*. En el artículo Dataprofiling con SQL Server 2008 [6] explico cómo se utiliza esta nueva Task de *SSIS*.

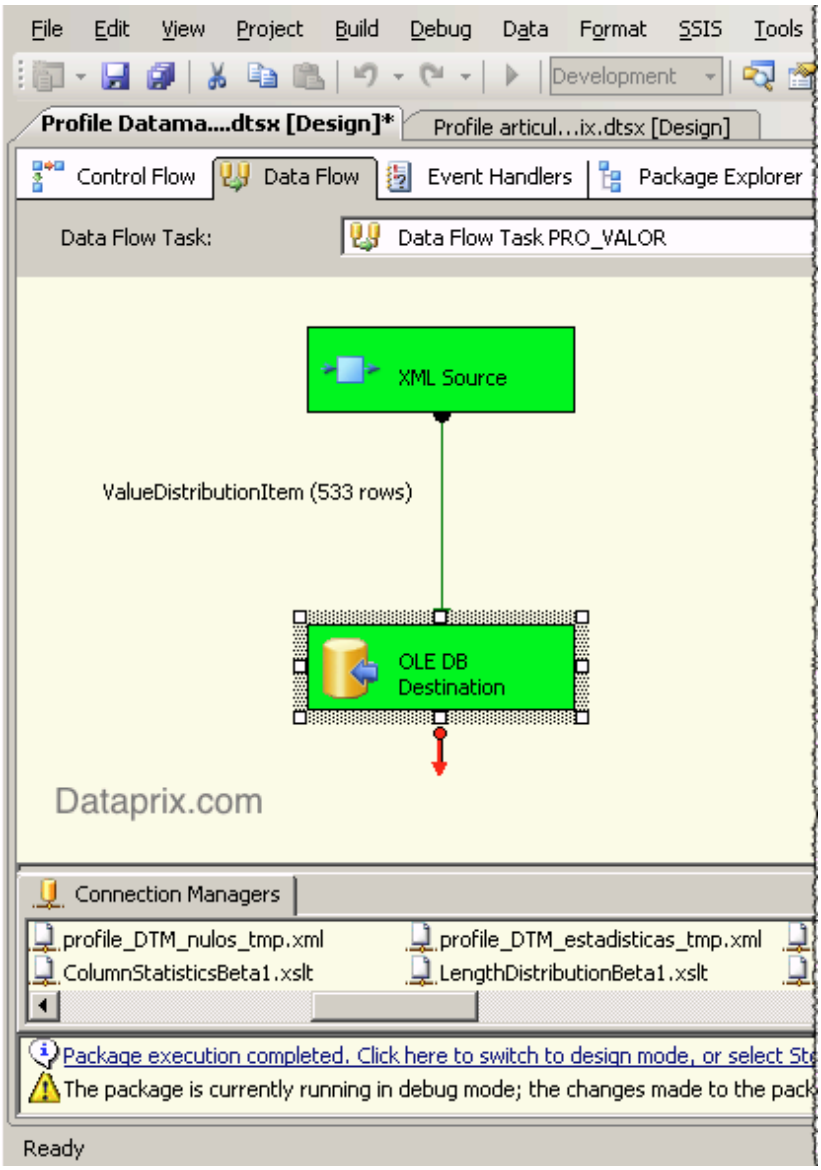
Aunque este método sea muy sencillo, a veces puede no resultar suficiente. Si se aborda un proyecto de calidad de datos puede interesar, por ejemplo, almacenar un histórico de los *perfilados* para poder evaluar cómo ha ido mejorando la *calidad de los datos* tratados.

La mejor manera de trabajar con datos históricos es utilizando una base de datos y almacenando estos datos en tablas, sobre las que se podrán hacer las consultas, informes y comparativas que haga falta. Para conseguirlo lo único que haría falta es pasar a tablas los *metadatos* que la tarea de perfilado ha almacenado en el fichero *XML*.

Pues alguien ya se ha dedicado a buscar una manera sencilla de hacerlo. Thomas Frisendal, desde su web Information quality solutions [7] explica cómo ha creado un archivo ***XSLT*** para cada tipo de perfilado que sirve para extraer del XML que genera la *Data Profile Task de SSIS* uno o más ficheros *XML* en un formato que puede ser directamente importado a **tablas**.

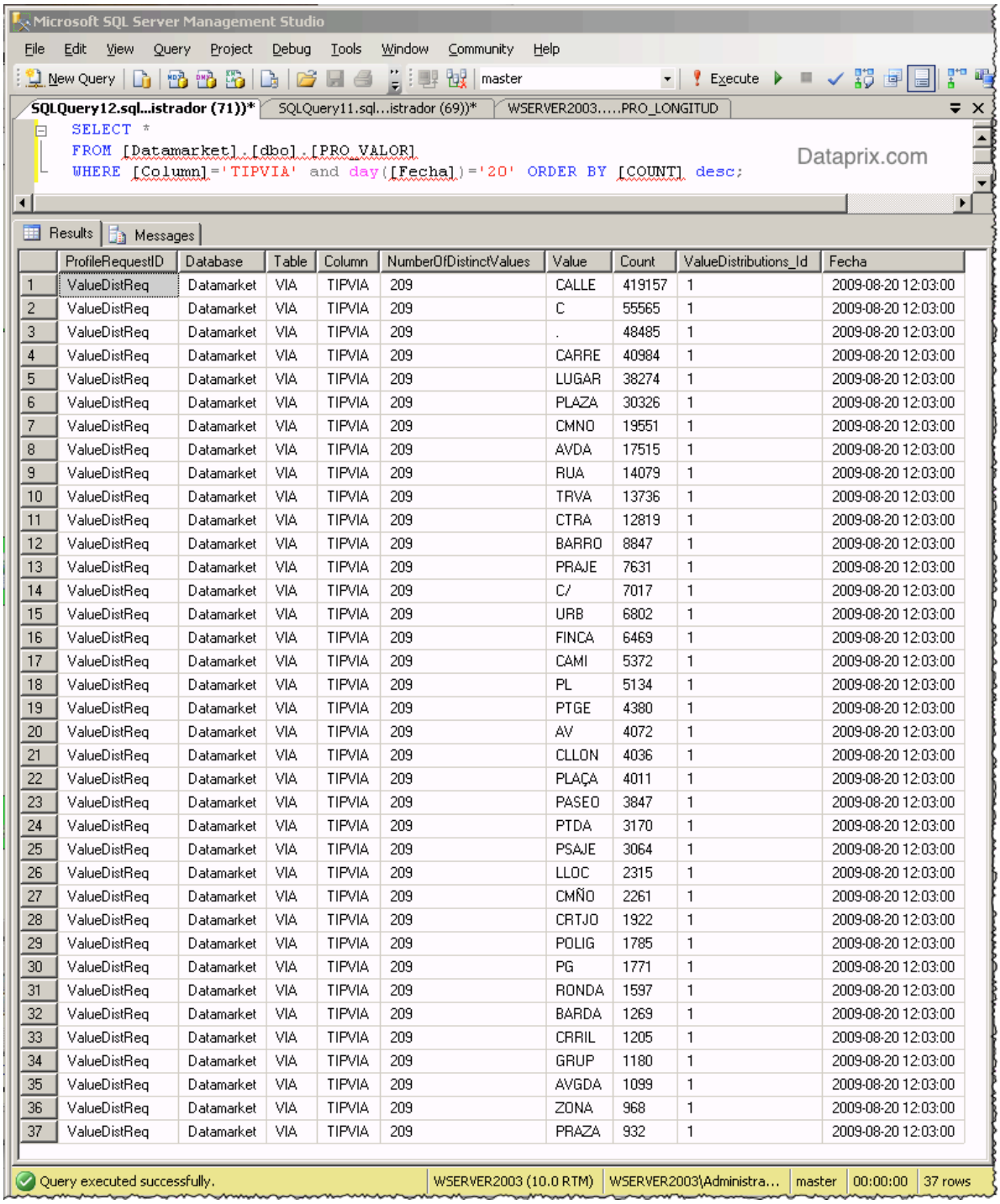


Así, con sólo crear un proceso que aplique un *XSLT* por cada tipo de perfilado, y después cargue cada fichero *XML* resultante en una tabla ya se pueden almacenar los datos de perfilado en tablas. Como además en todos los ficheros se incluye un campo que informa del nombre de la tabla origen, con una sola tabla para cada tipo de perfilado ya se pueden almacenar los profiles de todas las tablas que se traten.



Yo he añadido además a la tabla un campo de *fecha* que almacena la fecha en que se realiza el proceso, y de momento el resultado ha sido bueno.

En el gráfico podemos ver la distribución de *tipos de vía* diferentes para el *callejero español* según datos del INE y darnos cuenta, por ejemplo, de que los **identificadores** no están demasiado bien tipificados, ya que podemos encontrar cosas como más de un identificador para el mismo tipo de vía (CALLE, C, C/), o bastantes vías con un punto como identificador.



	ProfileRequestID	Database	Table	Column	NumberOfDistinctValues	Value	Count	ValueDistributions_Id	Fecha
1	ValueDistReq	Datamarket	VIA	TIPVIA	209	CALLE	419157	1	2009-08-20 12:03:00
2	ValueDistReq	Datamarket	VIA	TIPVIA	209	C	55565	1	2009-08-20 12:03:00
3	ValueDistReq	Datamarket	VIA	TIPVIA	209	.	48485	1	2009-08-20 12:03:00
4	ValueDistReq	Datamarket	VIA	TIPVIA	209	CARRE	40984	1	2009-08-20 12:03:00
5	ValueDistReq	Datamarket	VIA	TIPVIA	209	LUGAR	38274	1	2009-08-20 12:03:00
6	ValueDistReq	Datamarket	VIA	TIPVIA	209	PLAZA	30326	1	2009-08-20 12:03:00
7	ValueDistReq	Datamarket	VIA	TIPVIA	209	CMNO	19551	1	2009-08-20 12:03:00
8	ValueDistReq	Datamarket	VIA	TIPVIA	209	AVDA	17515	1	2009-08-20 12:03:00
9	ValueDistReq	Datamarket	VIA	TIPVIA	209	RUA	14079	1	2009-08-20 12:03:00
10	ValueDistReq	Datamarket	VIA	TIPVIA	209	TRVA	13736	1	2009-08-20 12:03:00
11	ValueDistReq	Datamarket	VIA	TIPVIA	209	CTRA	12819	1	2009-08-20 12:03:00
12	ValueDistReq	Datamarket	VIA	TIPVIA	209	BARRO	8847	1	2009-08-20 12:03:00
13	ValueDistReq	Datamarket	VIA	TIPVIA	209	PRAJE	7631	1	2009-08-20 12:03:00
14	ValueDistReq	Datamarket	VIA	TIPVIA	209	C/	7017	1	2009-08-20 12:03:00
15	ValueDistReq	Datamarket	VIA	TIPVIA	209	URB	6802	1	2009-08-20 12:03:00
16	ValueDistReq	Datamarket	VIA	TIPVIA	209	FINCA	6469	1	2009-08-20 12:03:00
17	ValueDistReq	Datamarket	VIA	TIPVIA	209	CAMI	5372	1	2009-08-20 12:03:00
18	ValueDistReq	Datamarket	VIA	TIPVIA	209	PL	5134	1	2009-08-20 12:03:00
19	ValueDistReq	Datamarket	VIA	TIPVIA	209	PTGE	4380	1	2009-08-20 12:03:00
20	ValueDistReq	Datamarket	VIA	TIPVIA	209	AV	4072	1	2009-08-20 12:03:00
21	ValueDistReq	Datamarket	VIA	TIPVIA	209	CLLON	4036	1	2009-08-20 12:03:00
22	ValueDistReq	Datamarket	VIA	TIPVIA	209	PLAÇA	4011	1	2009-08-20 12:03:00
23	ValueDistReq	Datamarket	VIA	TIPVIA	209	PASEO	3847	1	2009-08-20 12:03:00
24	ValueDistReq	Datamarket	VIA	TIPVIA	209	PTDA	3170	1	2009-08-20 12:03:00
25	ValueDistReq	Datamarket	VIA	TIPVIA	209	PSAJE	3064	1	2009-08-20 12:03:00
26	ValueDistReq	Datamarket	VIA	TIPVIA	209	LLOC	2315	1	2009-08-20 12:03:00
27	ValueDistReq	Datamarket	VIA	TIPVIA	209	CMÑO	2261	1	2009-08-20 12:03:00
28	ValueDistReq	Datamarket	VIA	TIPVIA	209	CRTJO	1922	1	2009-08-20 12:03:00
29	ValueDistReq	Datamarket	VIA	TIPVIA	209	POLIG	1785	1	2009-08-20 12:03:00
30	ValueDistReq	Datamarket	VIA	TIPVIA	209	PG	1771	1	2009-08-20 12:03:00
31	ValueDistReq	Datamarket	VIA	TIPVIA	209	RONDA	1597	1	2009-08-20 12:03:00
32	ValueDistReq	Datamarket	VIA	TIPVIA	209	BARDA	1269	1	2009-08-20 12:03:00
33	ValueDistReq	Datamarket	VIA	TIPVIA	209	CRRIL	1205	1	2009-08-20 12:03:00
34	ValueDistReq	Datamarket	VIA	TIPVIA	209	GRUP	1180	1	2009-08-20 12:03:00
35	ValueDistReq	Datamarket	VIA	TIPVIA	209	AVGDA	1099	1	2009-08-20 12:03:00
36	ValueDistReq	Datamarket	VIA	TIPVIA	209	ZONA	968	1	2009-08-20 12:03:00
37	ValueDistReq	Datamarket	VIA	TIPVIA	209	PRAZA	932	1	2009-08-20 12:03:00

En Free tool for automation of SQL Server [8] el autor comenta cómo funciona esta solución y cómo obtener las hojas de estilo, y en Usage recommandations for the ProfileToSQL stylesheets [9] explica más en detalle cómo utilizar los XSLT, e incluye un disclaimer dejando claro que este software es una versión de **test**.

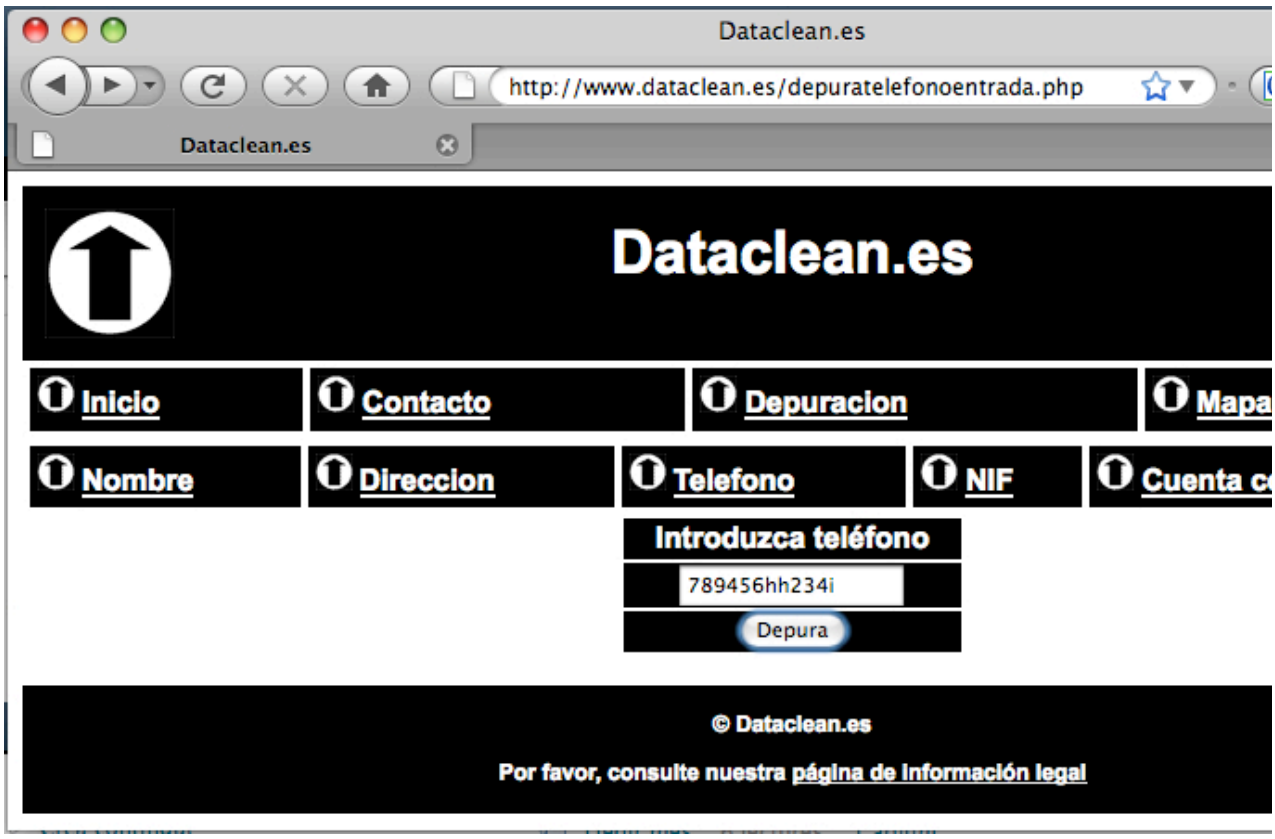
Coméntalo en el foro [10]

Dataclean.es: un proyecto de servicios de limpieza de datos

Hace ya bastante tiempo me planteé la posibilidad de arrancar un proyecto para ofrecer servicios de limpieza de datos online. Si hablamos en términos de lo que ara se oye más, podríamos interpretarlo como un nuevo significado de las siglas DAAS: Datacleansing As A Service.

En aquel momento escogí el nombre de Dataclean.es, entre otras cosas porque el dominio estaba libre. Lo registré a mi nombre e hice una aproximación a un plan de empresa. Hasta comencé a preparar una web donde quería crear una primera versión sencilla de la idea. Este prototipo se quedó en prácticamente una simple estructura, pero pienso que puede servir para ilustrar la intención que tenía.

Como al final no me decidí a dar el gran paso y desarrollar el proyecto, y es una pena que el esfuerzo que dediqué a hacer el planteamiento se quede en un documento de mi portátil, he decidido compartir el plan de empresa [11], adjunto en este post. Tambe he puesto online el prototipo web que comencé. Aviso que está tal como lo dejé, no funciona casi nada.



Lo he activado en Dataclean.es [12]. Como podréis comprobar, mi intención era comenzar con servicios de depuración de Nombres, Direcciones, Teléfonos, NIFs y Cuentas Corrientes. De estas opciones, la única con la que se puede hacer algo (y es bien poco) es la de teléfonos. Obviamente la intención era desarrollar servicios mucho más sofisticados, utilizando bases de datos, y retornando datos depurados, enriquecidos y normalizados. También quería ofrecer servicios de matching y

deduplicación, primero entre tipos de datos simples como los comentados, y después de registros o uniones de registros completos, con componentes o campos de diferentes tipos.

Bueno, espero que este post sirva al menos para animar el tema de Calidad de datos [13] de Dataprix, que reconozco que últimamente tengo un poco olvidado, y espero ver en el tema DAAS: Datacleansing As A Service [14] que he creado en el foro vuestras críticas y comentarios sobre la viabilidad del proyecto, sobre el Datacleansing As A Service en general, sobre otros enfoques posibles, o sobre cualquier aspecto relacionado con la temática o el documento del plan de empresa [15].

Adjunto	Tamaño
PlanEmpresaDataclean.pdf [16]	222.14 KB

En qué consiste el data cleansing

En el artículo del archivo adjunto los autores realizan una exposición bastante completa sobre en qué consiste el data cleansing, o limpieza de datos, las principales maneras en que se suele abordar, e incluso qué técnicas utilizan las principales compañías comerciales que ofrecen este servicio. (Bueno, las que lo ofrecían el año 2000, pero las técnicas principales no han variado mucho desde entonces).

Adjunto	Tamaño
IQ2000.pdf [17]	42.93 KB

Datacleansing con Power*MatchMaker/ DQGuru

DQGuru (antes **Power MatchMaker**) es una herramienta de *Data Cleansing* que **SQLPower** ha liberado convirtiendo la licencia en *Open Source*, junto con la de **Power Architect** (herramienta para modelización de datos).

Como no es que haya demasiadas herramientas Open Source en el campo de la *limpieza de datos*, me ha podido la curiosidad y la he instalado para ver que tal funciona.

La instalación ha sido muy sencilla, el software se descarga desde Descarga de DQGuru [18], en diferentes versiones según el SO. Yo he probado la de *windows*, que se instala a golpe de botón en 2 minutos. Importante no olvidarse del requerimiento del *Java Runtime 5*. Una vez instalado, para ver como funciona lo mejor es seguir el tutorial que se encuentra en la misma ayuda de la herramienta. También recomiendo ver la demo accesible desde la misma página de DQGuru [19].

El funcionamiento del software es muy sencillo, se crea un repositorio sobre una de las diferentes BBDD sobre las que puede trabajar, y con las que conecta por JDBC, y se pueden crear proyectos de 3 tipos diferentes: ***Deduplicación, Datacleansing*** y ***Referencias cruzadas***.
Eso en teoría, porque la funcionalidad de *referencias Cruzadas* aún no está implementada y no se puede utilizar. El proyecto de *Datacleansing* tampoco aporta nada nuevo, ya que toda la funcionalidad que utiliza es un subconjunto de la que ofrece uno de *Deduplicación*, con lo que con crear un proyecto de este último tipo ya lo vemos todo.

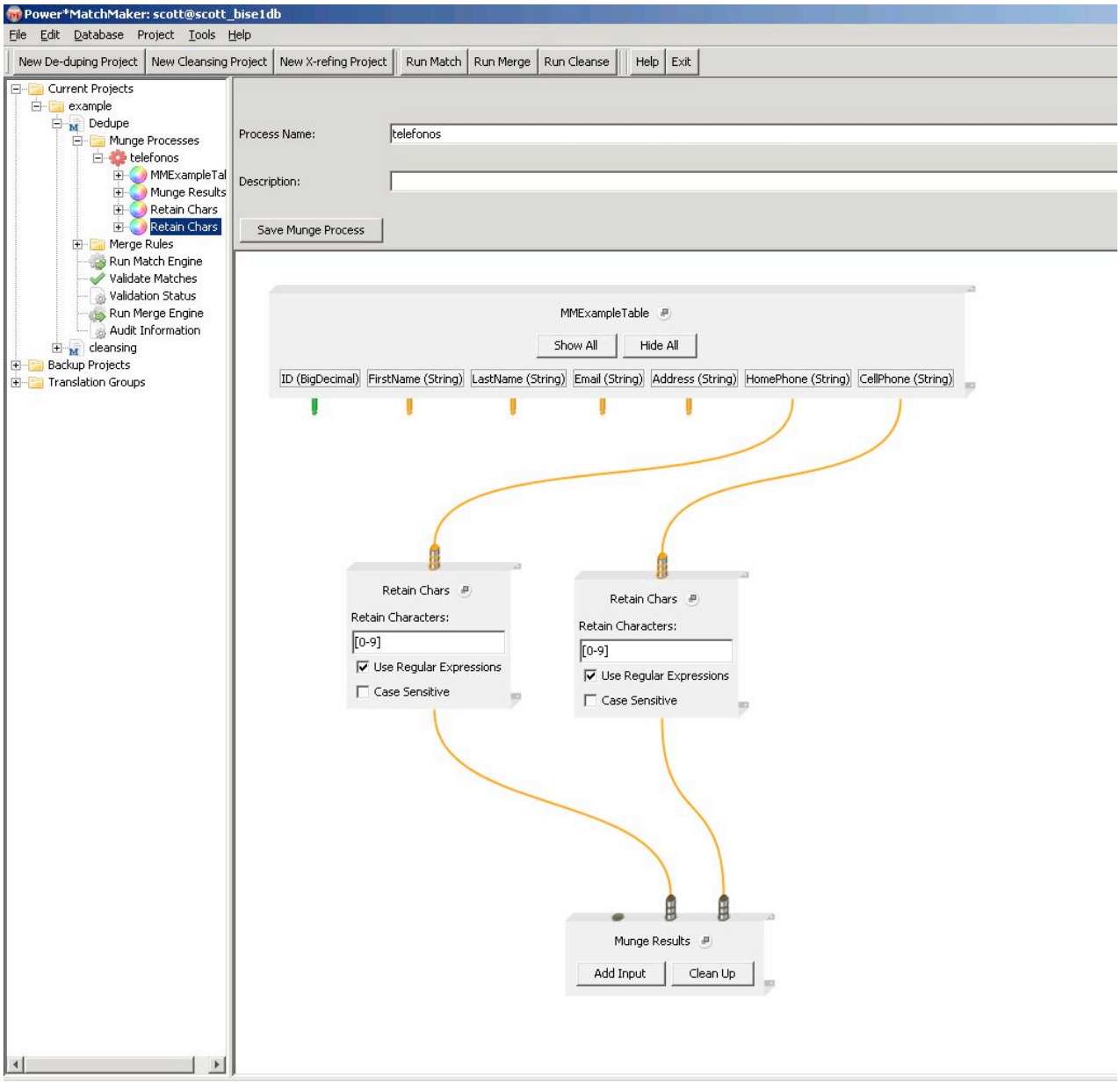
En cuanto a la *deduplicación*, se organiza el proceso en varios pasos:

1. Definición de procesos de transformación de los campos origen y comparación entre ellos

Se pueden definir varios procesos de comparación, aplicando diferentes operadores a los datos originales para obtener datos más significativos de cara a la comparación, y se define también exactamente lo que se quiere comparar.

La interfaz para realizar estas acciones es muy intuitiva y visual, y en un momento se define todo. La pega es que los operadores de comparación son bastante simples. Aunque hay operadores como los de *comparación fonética*, se echan de menos funciones de *fuzzy logic* para comparar palabras parecidas, o que se trabaje un porcentaje de similitud por campo y por registro.

Los resultados finales son que coincida o no todo lo que se ha definido. Lo único que se puede hacer es asignar una prioridad y un color para después distinguirlo visualmente a cada proceso de comparación.
Se echan también de menos funciones específicas de direcciones u otro tipo de datos 'estandar', aunque hay un operador que realiza una validación de la dirección con *Google Maps*. Yo no he conseguido que me funcione, pero es algo a explorar con más calma. También se pueden definir *diccionarios de traducción de palabras*, cosa muy útil cuando se comparan nombres o direcciones, por ejemplo.



2. Ejecución de la comparación

Nada que destacar, con pocos registros funciona bien, habría que probar con tablas grandes y valorar el rendimiento.

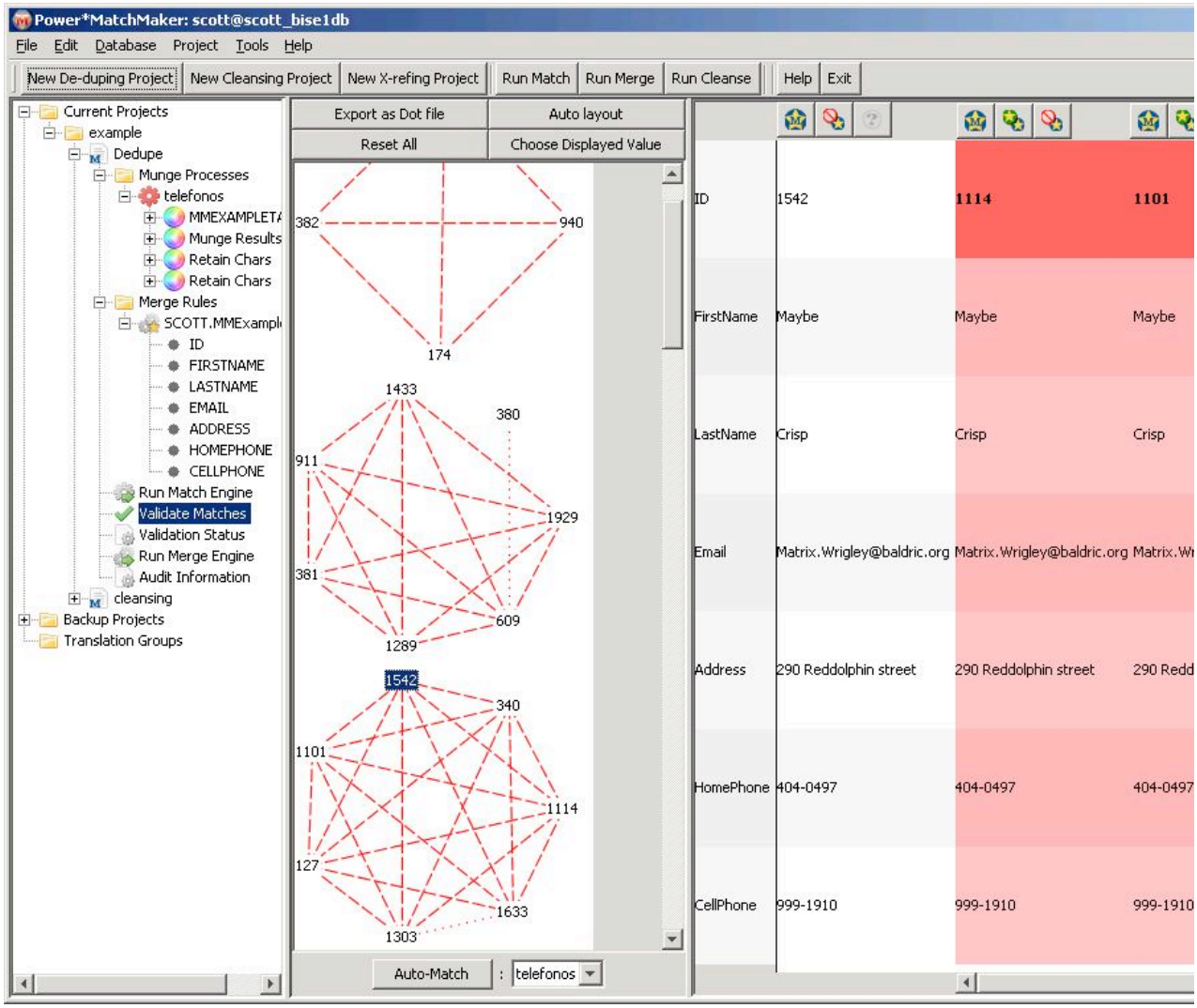
3. Validación de las coincidencias

Esta parte también está muy bien. La herramienta muestra de una manera muy visual las **coincidencias** encontradas, con un color para cada proceso definido, y permite ver las diferencias entre registros, y descartar coincidencias, decidir cuál es el registro maestro (el que va a conservar los datos tras la fución), y qué es lo que se va a fusionar y cómo.

Por defecto se escogen los datos del registro maestro, a menos que el campo sea nulo,

y también se pueden concatenar los datos, o coger el máximo, el mínimo o la suma de todos. Si se quiere se puede dejar escoger a la herramienta automáticamente el registro que queda como *maestro*, y hacer que se fusionen todos los registros en los que se ha encontrado coincidencia.

La herramienta está muy bien para trabajar con un número limitado de registros, que puedan ser revisados por una persona antes de la **fusión**, pero le falta un poco de 'inteligencia' para poder tratar con un número elevado de registros, y realizar fusiones sin necesitar intervención manual. También debería poderse escoger el dato maestro a nivel de campo, y no a nivel de registro, y con los mejores datos de cada campo crear el mejor registro maestro.



4. Proceso de fusión de registros

Funciona correctamente, deja un log y guarda los identificadores de lo que se *fusiona* en una tabla de resultados. Sólo hay que tener cuidado porque trabaja directamente sobre la tabla origen, y borra los registros que se han marcado como **duplicados**.

Conclusiones

En definitiva, es una herramienta muy útil para realizar procesos de *limpieza*, sobretodo si la cantidad de *datos* a revisar no es muy grande. Sin grandes complicaciones permite realizar todo el proceso y facilita mucho la comparación entre candidatos y la elección de los registros que quedarán como maestros.

Tiene varios aspectos a mejorar, pero seguro que no se va a quedar en esta versión, sobretodo sabiendo que ahora es de *código abierto*.

Herramientas de ETL

En la publicación electrónica MCData.ti se puede encontrar una clasificación bastante completa de diferentes herramientas relacionadas con el business intelligence y la gestión de datos. Esta es la descripción que se realiza de las herramientas de ETL (Extraction, Transformation and Load).

Empresa: Cognos Producto: DecisionStream

Descripción: Herramienta específica para sistemas SAP y Siebel que permite implantar rápidamente una serie de data marts relacionados para componer un sistema integrado de BI. Asegura que todos los usuarios tengan acceso a los datos para coordinar de forma eficiente el análisis, los informes y la toma de decisiones.

Empresa: Computer Associates Producto: Advantage Data Transformer - Enterprise Metadata Edition

Descripción: Herramienta de transformación y movimiento de datos que permite crear datamarts y almacenes de datos de forma fácil, transformaciones complejas o una gestión robusta de metadatos.

Producto: Advantage Data Transformer

Descripción: Herramienta de transformación, replicación e integración de datos que cuenta con un entorno de desarrollo de aplicaciones flexible y fácil de usar. Su rico lenguaje de programación permite definir fácilmente tareas de movimientos de datos simples o complejos, juntar datos de Fuentes distintas, limpiar datos, sintetizar nuevos datos y sincronizar varias bases de datos en plataformas mixtas. Producto: Advantage InfoRefiner

Descripción: Herramienta para la migración de datos. Sirve para replicar fuentes de datos en otras estructuras de datos, para difundir cambios hechos en los datos originales en otros datos, y para auditar los cambios hechos en los datos originales a lo largo del tiempo. Producto: Advantage InfoTransport

Descripción: Herramienta de movimiento de datos de alta velocidad que distribuye y carga datos del mainframe en entornos cliente/servidor heterogéneos.

Empresa: Information Builders Producto: WebFocus ETL Manager

Descripción: Herramienta que simplifica la presentación de los datos para proporcionar

información detallada en tiempo real a través de la intranet. De esta forma, los usuarios pueden crear y visualizar informes o mover los datos hasta aplicaciones de sobremesa como Excel.

Empresa: MIS Producto: MIS Import Master

Descripción: Herramienta para la extracción, transformación y carga de datos desde cualquier fuente de información transaccional. Asimismo permite acceder directamente a las tablas funcionales de sistemas ERP tan extendidos como SAP R/3 o Navision.
Empresa:

Oracle Producto: Oracle 9i Warehouse Builder

Descripción: Herramienta perteneciente a Oracle 9i Development Suite que permite al usuario diseñar e implantar data warehouses corporativos, data marts y aplicaciones de negocio electrónico inteligentes. Ofrece integración con entornos CRM, ERP y SEM.
Empresa:

PowerData Ibérica Producto: Informática PowerCenter

Descripción: Plataforma de integración de datos que aúna las funciones de transporte, limpieza y migración de información. Aunque un porcentaje muy alto de su uso se debe al diseño de data warehouses y web houses, su utilización se ha extendido a otras áreas. Así, por ejemplo, es común la integración de los CRM, los sitios web y las plataformas de comercio electrónico con los sistemas operacionales.

Empresa: SAS Producto: SAS Warehouse Administrator

Descripción: Solución de extracción, transformación, carga y limpieza de datos que facilita la definición visual de los procesos corporativos y su documentación.

Empresa: Teradata Producto: Teradata Warehouse Builder

Descripción: Herramienta de carga y descarga que permite al usuario generar un job o secuencia de comandos para acceder a datos heterogéneos, comprobar la integridad de los mismos o fusionarlos con otros. Incorpora puntos de control para el relanzamiento del proceso en caso de caída del sistema

Informatica World 2008 en Las Vegas

Finalmente he podido asistir al Informatica World 2008 [20] y qué menos que explicar un poco lo que me encontré por allí. La conferencia se celebró del 3 al 5 de junio bajo el lema '*Gain the Edge*', una expresión con mucha fuerza en inglés pero difícil de traducir al castellano, a ver si alguien se anima y nos da una traducción válida.

El día 3 comenzó, después del desayuno, con una **sesión general** que llevaba por título *Vision. Strategy. Technology Announcements. Industry leadership.* En la misma, tanto Sohaib Abbasi, CEO y Presidente de Informática [21], como Chris Boorman, Ivan Chong y Girish Pancha, Vicepresidentes en las àreas de Márqueting, Calidad de Datos e Integración de Datos, respectivamente, nos mostraron su visión actual del mercado, cómo están evolucionando la tecnología y los negocios, y qué papel juegan en este marco los datos y las aplicaciones que los gestionan.

Mucho de lo que comentaron ya había podido escucharlo en las presentaciones del Powerday 2008 de Barcelona [22], cosa que muestra que la compañia mantiene una estrategia bien definida, y la comparte con sus partners.

Se hizo especial hincapié, cada ponente bajo la perspectiva de su àrea, del valor que representan los datos, y lo importante que es la habilidad de cada organización para gestionarlos, mantener su coherencia y calidad, garantizar su accesibilidad en el momento oportuno, protegerlos, sincronizarlos y poder intercambiarlos con otras organizaciones.

Nos hicieron notar que ahora ya no sólo se trata de crear un almacén de datos corporativo que nos porporcione información actualizada cada cierto intervalo de tiempo. La evolución tecnológica y de Internet, la globalización y la competencia nos hacen plantearnos que con el Data Warehouse no es suficiente. Se oyeron mucho los términos SaaS, Real Time y Data Quality, cosa que nos da pistas sobre hacia donde van encaminadas las nuevas funcionalidades de las herramientas de la compañia.

También realizaron una interesante **demo** sobre cómo una aplicación como Salesforce.com puede sincronizarse en tiempo real, y a través de Internet, con una hoja de cálculo de Google Docs. Este ejemplo de cloud to cloud computing lo prepararon mostrando en la pantalla de la izquierda Salesforce.com y en la de la

derecha una spreadsheet de Google Docs, cada aplicación 'controlada' desde un portatil. En el portatil de Salesforce realizaron un cambio, y pudimos ver cómo se actualizaba al momento la hoja de cálculo. Después hicieron otra modificación en la hoja de cálculo, y la aplicación de Salesforce también se actualizó, todo a través de Internet. Para *poner la guindilla* después hicieron lo mismo, pero con un iPod touch, no hay que olvidar las posibilidades que nos brindan los nuevos dispositivos móviles cuando se conectan a la web.

También pudimos asistir a una animada presentación de Royce Bell, CEO de Accenture Information Management Services, que supo cómo mantener la atención de todo el mundo.

La sesión general de este Informatica World 2008 daba paso a las **Breakout Sessions**, cada una de ellas clasificada en una de las siguientes categorías:

- Productos y Tecnología
- Arquitectura
- Gestión de Datos Empresarial
- Soluciones
- Presentación Técnica
- Impacto sobre el negocio

Además se catalogaban según nivel de experiencia y rol del público al que iban dirigidas.

Los niveles eran Beginner, Intermediate y Advanced, y los roles Architect, Business and

IT Influencer y Practitioner.

Así cada uno podía seleccionar las sesiones que más le interesaran y mejor se adaptaran a su perfil profesional.

En total había 56 sesiones, de las cuales había que elegir como mucho 8. Como son tantas, listaré a continuación sólo el título de cada una, dentro de cada categoría, todo en el idioma original, y subrayo las que yo seleccioné:

Products and Technology

1. What's New in PowerCenter
2. Data Quality with Identity Resolution: A Leap Forward for Data Quality in the Enterprise
3. How to Get More from Informatica Metadata Manager
4. The Informatica Roadmap: Vision for V9
5. Informatica B2B Data Exchange: Building a Data Exchange
6. What's New in Informatica Data Explorer and Informatica Data Quality 8.6
7. Protecting Private Data Using PowerCenter Data Masking
8. Real-Time Data Integration

Architecture

1. Customer Panel: Real-Time Integration Architectures for Right-Time Business Value
2. Informatica Architecture: Where to Start?
3. A Practical Approach to Building Data Services with PowerCenter 8.5
4. Informatica Orchestration and Human Workflow: Process-Enabled Data Integration and Data
5. Maximizing Operational Uptime: Real-Time Data Integration with Informatica

6. On Demand Data Integration: Overview and Demonstration
7. Deploying PowerCenter on Grid Computing Architectures
8. PowerCenter Data Federation Option: A Unified Platform for Data Integration Flexibility

Enterprise Data Management

1. Data Quality, The First Step on the Path to Master Data Management
2. Where Real-Time Data Integration Meets Real-Time Data Warehousing
3. IMS Health: Global Data Integration for Financial Information Management
4. Customer Master Data Management at Major Telecommunications Company KPN, Netherlands
5. Measuring and Improving Data Governance Maturity: A Practical Approach
6. Information Management: An Implementer's Perspective
7. Measuring Data Quality in Philips Consumer Lifestyle
8. Lowering Cost and Risk with the Data Migration Factory
9. Data Profiling and Data Quality Improvement: A Practitioner's Approach
10. Velocity Methodology: Best Practices

Solutions

1. Campaign Marketing and Customer Relationship Management at Daimler AG
2. A Trip to Better and Faster Corporation Travel Management: A B2B Data Transformation Success
3. Informatica B2B Data Transformation: Success with LOGTEC for the Defense Logistics Agency
4. Assuring Success When Integrating Salesforce CRM with the Rest of Your Business: A Partner Profile with Case Studies from Ellie Mae and Millennium Pharma
5. Data Migration Success at G&K Services
6. Leveraging HP and Informatica for Large-Scale Data Migration Efforts: A Case Study at CVS Caremark
7. Strategy to Implementation: How to Get Started on your Data Quality Initiative
8. Identity Resolution: What It Is and Why It Is Important

Tech Talk

1. Extreme Automation: Traceability of Requirements through Testing, Governance and Compliance
2. Planning and Tuning Informatica for Large Loads
3. Tips to Improve Productivity Using Self-Service Support Tools
4. Command and Control: Using Informatica Workflows to Regulate Complex Business Processes
5. Informatica Developer: Tips and Tricks for Architecture and Development
6. Upgrading to the Latest PowerCenter Release: Tips and Tricks, Testing and Pitfalls to Avoid
7. Using Team-Based Development: A Practical Exposé
8. High-Volume Data Processing (>150GB) Using Informatica
9. Informatica Developer Tips for Troubleshooting Common Issues
10. Power of Informatica PowerCenter at Verizon Wireless

Business Impact

1. Driving Business Value with Integration Competency Centers: Customer Presentations, a Two-Part Series (Part 1 of 2)
2. Integration Competency Centers: Panel Discussion, a Two-Part Series (Part 2 of 2)

3. Anti-Money Laundering Compliance: Stopping Financial Crime - a Data Quality Approach
4. Quantifying Business Value with Informatica: Best Practices and Techniques for Funding Enterprise Data Integration and Data Quality Projects
5. Informatica B2B Data Exchange: Success with Paramount Pictures
6. Integration Competency Center at Duke Energy
7. Building a Business Case for B2B Data Exchange at a Major HMO
8. Data Governance in a Global Enterprise
9. Enterprise Data Warehouse at a Medical Device Manufacturing Company
10. Informatica B2B Data Transformation: Success with GfK Group

Como se puede apreciar, la categoría que más me interesó fue la de Gestión de Datos en la Empresa, seguida de la de Soluciones. De todas maneras debo aclarar que actualmente no utilizo productos de Informática, por lo que las categorías relacionadas con desarrollo o temas específicos del software no me resultaban tan atractivas.

Encontré la mayoría de las sesiones muy enriquecedoras, nadie mejor que los expertos de Informática para asentar conceptos sobre las últimas tendencias en gestión y calidad de datos, en Data Warehousing, o para recomendarte *best practices*, o pasos a seguir para abordar un proyecto de este tipo.

De todas maneras siempre lo mejor es la presentación de alguien que ha vivido en su empresa una implantación o una experiencia, y que la cuenta bajo una perspectiva más imparcial. En este sentido creo que la mejor sesión a la que asistí fue la Customer Data Management en KPN, presentada por Thomas Reichel (KPN) y Chris Phillips (Informatica)

Tras estos días de Breakout Sessions llegó el jueves 5 en que se celebró la **sesión general** que marcaba el **final del evento**. El título de la misma era *Gaining the Edge. In Real Time*

Después de haber mostrado en la sesión inicial la necesidad de las organizaciones de gestionar sus datos con la mayor eficiencia, y adaptándose al progreso tecnológico, esta sesión se enfocó más a cómo conseguirlo con la ayuda del software y el soporte de Informática, se mostraron las nuevas funcionalidades que ofrece la versión 9 del producto, y cómo aprovecharlas.

Me gustó la demo que realizó Ivan Chong sobre cómo gestionar y realizar procesos de Data Quality con esta nueva versión, pero lo que más me impresionó fue la presentación que hizo Ron Swift, vicepresidente de Teradata, sobre la importancia de gestionar datos en tiempo real para poder reaccionar a tiempo ante determinadas situaciones. Puso el acertado ejemplo de un casino que había implementado un sistema que analizaba en tiempo real el comportamiento de sus clientes mientras jugaban y que, si detectaba que alguno estaba perdiendo demasiado dinero, para no acabar perdiéndolo hacía saltar una alarma que avisaba para que el personal pudiera persuadirlo de seguir jugando.

Para finalizar sólo agradecer a Powerdata [23]la invitación para poder asistir a esta edición del Informatica World, y el amable trato que me han brindado durante todo el viaje.

Integracion y calidad de datos en el PowerDay 2008

En marzo-abril se celebró la séptima edición de **Powerday**, un evento anual que organiza PowerData [23], y que este año tenía por objetivo proporcionar a los asistentes una visión global de la estrategia adecuada para sacar el máximo partido a los datos. Yo tuve la oportunidad de asistir al de **Barcelona**, y disfrutar con las interesantes ponencias que se realizaron en el mismo.

Fueron presentaciones de una media hora, en las que se habló sobre la importancia de la **calidad de datos** y los **procesos de integración**, sobre la situación tecnológica y de mercado actual y, por supuesto, sobre cómo facilitar las cosas con la utilización de herramientas de Informática [24]como PowerCenter [25].

Estos son los títulos de las presentaciones:

- *El valor de los datos correctos trasciende el departamento TI*
- *Principios prácticos para garantizar una buena calidad de los datos dentro de la organización*
- *Enmascaramiento de datos: una respuesta efectiva a demandas de confidencialidad*
- *Integración de datos corporativos en Caprabo*
- *Importancia de contar con buenos datos en entornos analíticos*
- *El modelo de organización en tiempo real impone nuevas exigencias en la gestión de la información*
- *Tendencias del mercado español de gestión de datos*

Encontré especialmente interesante la de Caprabo [26], realizada por Sergio Champel, el Jefe del Area de Arquitectura e Integración de esta empresa. Sergio explicó cómo se habían organizado tanto a nivel de gestión como de arquitectura para llevar a cabo con éxito un ambicioso proyecto de integración y remodelación del sistema de Business Intelligence de Caprabo, con el que han conseguido mejorar importantes procesos de negocio, y 'estrechar los lazos' entre los sistemas operacionales y el Data Warehouse.

Me llamó mucho la atención la frase *Aprendemos a utilizar un martillo y todo nos parece un clavo,* que Sergio mencionó para dejar claro lo que querían evitar cuando definieron la arquitectura. Me pareció un frase muy acertada, y aplicable a múltiples situaciones.

Destacar también que la presentación *Importancia de contar con buenos datos en entornos analíticos* la realizó Jorge Zaera, director general de Microstrategy [27].

Las demás fueron presentadas por expertos y directivos de Powerdata, que supieron mostrar los diferentes aspectos a tener en cuenta en todo lo relacionado con la integración y calidad de los datos, y qué papel juegan estas materias en las últimas tendencias tecnológicas del mercado, cada vez más orientadas al **proceso** y al **servicio**, como SaaS (Software as a Service), SOA (Service Oriented Architecture), BPM (Business Process Management), CPM (Corporate Process Management) o EIM (Enterprise Information Management)

Para el que prefiera hablar de cosas más tangibles, también se proporcionó

una clasificación de tipos de proyectos que nos podemos encontrar en cuanto a la gestión de los datos:

- Data warehouse
- Migración de datos
- Consolidación de datos
- Master Data Management
- Sincronización de datos
- Intercambio de datos B2B

Tras las presentaciones se realizó un sorteo de un viaje a Las Vegas para asistir a Informatica World 2008. Resulta que el afortunado ganador del sorteo fui yo, por lo que en unas semanas espero estar publicando un nuevo artículo sobre mis experiencias en este evento [28] al otro lado del charco.

Source URL: http://www.dataprix.com/art-culos-integraci-n-datos-dataprix

Links:
[1] http://www.dataprix.com/forum/2009/07/datawarehouse-tri-capa
[2] http://technet.microsoft.com/es-es/library/bb895263.aspx
[3] http://www.sql-server-performance.com/articles
/biz/SSIS_New_Features_in_SQL_Server_2008_Part3_p1.aspx
[4] http://www.sql-server-performance.com/articles/biz/data_profiler_ftp_task_ssis_p1.aspx
[5] http://www.dataprix.com/forums/herramientas/sql-server-integration-services
[6] http://www.dataprix.com/data-profiling-sql-server-2008
[7] http://informationqualitysolutions.com
[8] http://www.informationqualitysolutions.com/page2/page11/page13/page13.html
[9] http://www.informationqualitysolutions.com/page2/page11/page13/page14/page14.html
[10] http://www.dataprix.com/../../../../../../forums/herramientas/sql-server-integration-services
[11] http://www.dataprix.com/system/files/PlanEmpresaDataclean.pdf
[12] http://www.dataclean.es
[13] http://www.dataprix.com/data-quality
[14] http://www.dataprix.com/es/projecte-daas-datacleansing-as-a-service
[15] http://www.dataprix.com/es/system/files/PlanEmpresaDataclean.pdf
[16] http://www.dataprix.com/files/PlanEmpresaDataclean.pdf
[17] http://www.dataprix.com/files/IQ2000.pdf
[18] http://download.sqlpower.ca/dqguru/current.html
[19] http://www.sqlpower.ca/page/dqguru
[20] http://www.informatica.com/events/customer_conference/default.htm
[21] http://www.informatica.com
[22] http://www.dataprix.com/es/integraci-n-y-calidad-datos-el-powerday-2008
[23] http://www.powerdataib.com/
[24] http://www.informatica.com/
[25] http://www.informatica.com/products/powercenter/default.htm
[26] http://www.caprabo.es/
[27] http://microstrategy.es/
[28] http://www.dataprix.com/es/informatica-world-2008-las-vegas