Documento Técnico

Integración de Salesforce y Pentaho

Fecha creación: 27/09/2011

info@stratebi.com

www.stratebi.com - www.todobi.com

1. Integrando SalesForce Cloud usando Pentaho Data Integration (Kettle)

En tiempos de crisis el mercado se encuentra estancado por regla general, y esa gran “tarta” debe ser repartida cada vez por más compañías (empresas industriales, proveedores de servicio…). Para que una empresa logre sus objetivos, debe ser más competitiva que nunca, y ahí entra lo que se conoce como “toma de decisiones corporativa".

La mayoría de compañías toman decisiones lentas, previsibles y basándose en formulas, en ocasiones intangibles y que no reflejan la realidad de su negocio. Como se puede intuir entonces, uno de los secretos en la toma de decisiones, son los datos en los que se basa, por lo que cuanta más muestra y variedad de datos se tenga, más fina y acertada será nuestra respuesta al mercado.

Será entonces, una buena práctica integrar progresivamente nuevas fuentes en nuestro datawarehouse/datamart. Un proyecto por lo general extrae datos de; BBDD de negocio, ERPs, Redes Sociales, Excels, Ficheros Planos, web logs, web services, CRMs…y en el caso particular que nos atañe, nos detendremos en estas dos últimas, que pueden ser englobadas perfectamente dentro del mayor exponente de CRM en la nube hoy en día, SalesForce.

En el artículo se refleja brevemente algunas experiencias adquiridas durante algunos de nuestros proyectos en el ámbito de la integración de datos con este origen tan peculiar, “SalesForce Cloud”.

2. Breve introducción sobre Cloud Computing.

Cloud Computing, como su propio nombre indica, es una disciplina de computación distribuida que se basa en que la mayor parte del hardware y software este en la nube, es decir, fuera de nuestra compañía.

Todo el hardware y software se ofrece como servicio, de manera que las compañías "alquilan" lo que estrictamente necesitan. A continuación se enumeran algunas de sus ventajas:

- Prestación y acceso a servicios mundialmente.
- Implementación de las aplicaciones más rápida.
- Actualizaciones totalmente transparentes al usuario y a las aplicaciones.
- Uso eficiente de recursos y energía.
- Ahorro de costes en licencias e infraestructuras.

Toda esta oferta y posible demanda de servicios se lleva a cabo por medio del paradigma orientado a servicios SOA (Service Oriented Architecture), que intenta eclipsar a anteriores como el Orientado a Objetos.

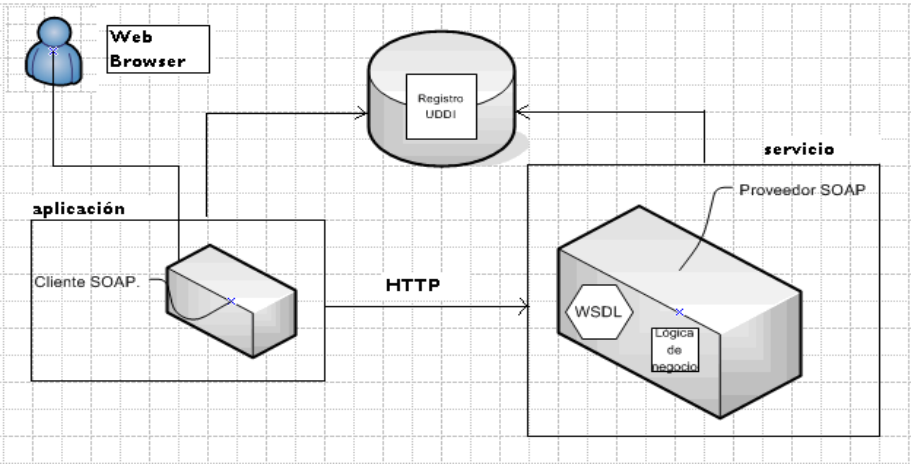
Y la pregunta es; ¿Qué es un servicio? Un servicio es un reflejo de un proceso de negocio, una abstracción para aplicaciones organizativas de gran escala (P. Singh) orientada a interactuar. Este paradigma trata aspectos como que los servicios deben ser dirigidos por eventos, es decir, los procesos de negocio no pueden ser diseñados asumiendo a priori el flujo de eventos sino que deben de ser diseñados dinámicamente, con asincronismo, proveyendo ciertas paradas en el proceso y su posible continuidad. Las características básicas de un servicio son:

- Disponible en Internet y/o Intranets.
- Utiliza un sistema estándar de mensajería.
- Es independiente del lenguaje de programación y del sistema operativo.
- Es Auto-descriptivo vía una gramática XML.
- Descubrirle vía un mecanismo simple de búsqueda.

La pila de protocolos de un servicio web se puede representar como sigue:

Descubrimiento – UDDI: Esquemas XMl cuyo objetivo es describir empresas y servicios web
Descripción – WSDL: Lenguaje común para describir servicios web, permitiendo al cliente localizar un servicio web e invocar sus funciones públicas
Mensajería XML– XML-RPC, XML, SOAP. Nos centraremos en SOAP ya que es

el más extendido, posibilitando el intercambio de información en formato XML. La transmisión de información se basa en la infraestructura de correo electrónico de Internet, por ello, un mensaje SOAP se puede definir como un fichero XML que sigue la RFC 2822 para mensajes de correo electrónico
Transporte – HTTP, SMTP, FTP, BEEP



Los tipos de servicios disponibles en la nube, pueden ser clasificados en las siguientes tres capas:

- IaaS ("Infraestructure as a Service"): Es la capa de más bajo nivel, en ella se encuentran compañías como Amazon, con su AWS (Amazon Web Services), alquilando infraestructuras con un bajo coste y alta disponibilidad. Se pueden alquilar productos como:
    - BBDD relacionales.
    - Almacenamiento.
    - VPNs.
    - Servicios de computación (máquinas)
    - Clustering.
- PaaS ("Platform as a Service"): Esta capa ofrece una encapsulación sistema+entorno de desarrollo, por ejemplo una maquina Linux, con conexión de red al exterior, con una bbdd mysql y con un apache montado, con el objetivo, por ejemplo, de hospedar la web corporativa. Un ejemplo de este caso es Google App Engine, que ofrece un entorno de desarrollo de GWT donde podemos compilar, probar y ejecutar aplicaciones web en la infraestructura de Google, es decir, no necesitas ningún servidor para que tus usuarios puedan comenzar a usar las aplicaciones desarrolladas, tan sólo, sube la aplicación a Google, proporciónale el nombre de

dominio y comparte tu aplicación con el mundo o limítalo a usuarios de tu organización.

- SaaS ("Software as a Service"): Se corresponde con la capa más alta de los servicios en la nube, y en la que se ofrecen software como servicio, un ejemplo, es Windows, con su Windows 365, ofreciendo todas las funcionalidades de la suite MS Office a través de servicios web. Otro ejemplo que en este articulo no puede dejarse en el tintero es SalesForce con su "SalesForce Cloud" que veremos de manera detallada en el siguiente punto.

Esta disciplina de computación tiene muchas ventajas, pero también se encuentra envuelta en un ambiente de controversia; hay expertos que afirman que se ha reinventado la rueda (The Times) y otros, que con esta disciplina se estan limitando las libertades del usuario (Richard Stallman). Con muchos seguidores y detractores, lo que realmente si se puede llegar a decir de ella, es que es una disciplina que ha dado y que seguirá dando mucho que hablar durante los próximos años.

3. SalesForce Cloud.

Es conocido como el mejor CRM off-site que actualmente existe. Para los usuarios finales, simplemente parece una página web con información de clientes, pero es mucho más que eso, es una aplicación con una arquitectura orientada a servicios que posibilita:

- Tener los datos centralizados en un único servidor.
- Posibilidad de utilización y gestión de estos datos con un simple navegador web desde cualquier dispositivo con conexión a Internet, un usuario y una password.

Esta arquitectura orientada a servicio utiliza el protocolo de transmisión SOAP(Simple Object Access Protocol.

Para conocer detalladamente los servicios que provee la arquitectura de SalesForce, tenemos a nuestra disposición el API de salesForce:

http://www.salesforce.com/us/developer/docs/api/index.htm

En él, podemos encontrar llamadas a servicios web como:

https://www.salesforce.com/services/Soap/u/20.0

La llamada anterior, se corresponde con el servicio web para conexión a Salesforce Cloud, en ella cabe destacar la versión del servicio web del que estamos hablando, en este caso 20.0.

Existen diferentes modalidades de alquiler del servicio:

SalesForce desde Pentaho Data Integration (Kettle)

Ya conocemos Pentaho Data Integration y sabemos de su gran potencia y funcionalidad. Desde la versión 3.2 se encuentra ya integrado por defecto el paso “Salesforce Input” para poder extraer información de cualquier módulo de SalesForce y en versiones actuales como 4.1.0 se encuentran además otros pasos como “Salesforce Delete”, “SalesForce Insert”, “SalesForce Update”, ”SalesForce Upsert”:

Salesforce Input	 Salesforce Input
Salesforce Delete	 Salesforce Delete
SalesForce Insert	 Salesforce Insert
SalesForce Update	 Salesforce Update
SalesForce Upsert	 Salesforce Upsert

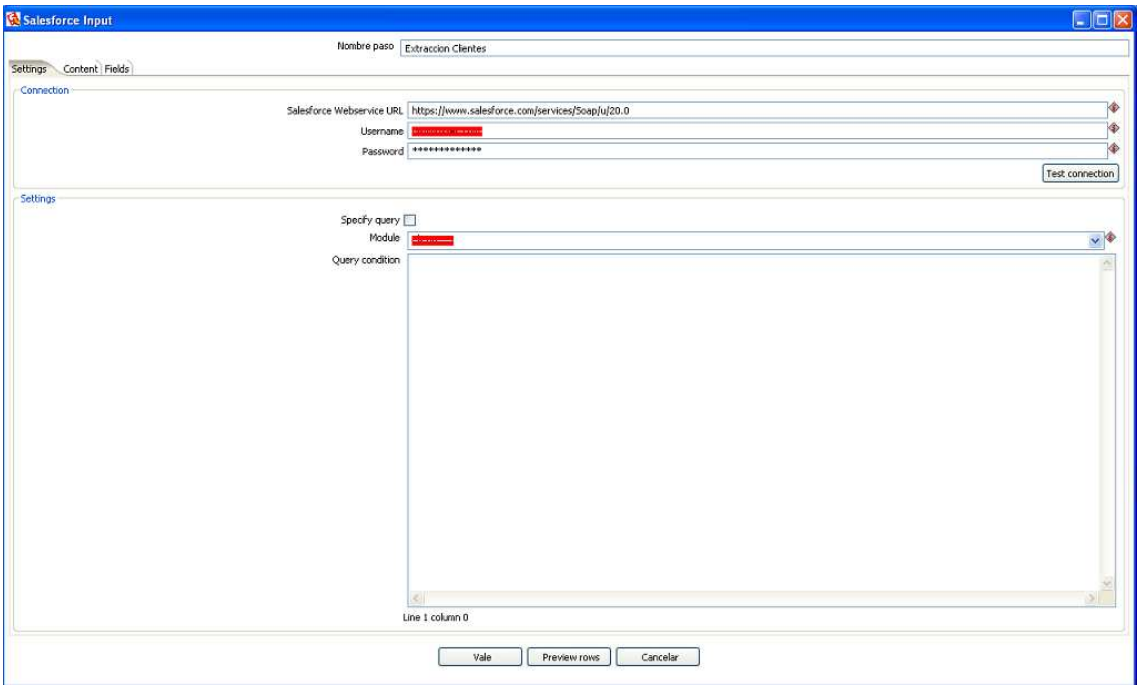
Antes de arrancar PDI, se deben revisar si los jars que se encuentran disponibles en nuestra versión de PDI se corresponder con los necesarios para conectar con la versión de SalesForce que queremos interactuar. Los plugins de salesForce en PDI se encuentran en el directorio:

\data-integration\libext\salesforce

Por ejemplo, en la versión PDI 3.2 tenía por defecto el plugin para Salesforce 10.0 la versión PDI 3.8 la 20 y la versión PDI 4.2.0 GA, tiene incorporada ya la 21.

En la mayoría de los proyectos de integración en lo que interviene SalesForce como fuente, usamos todos los pasos de salesforce que nos provee PDI, ya que tras el proceso de integración y data quality se nos suele requerir realimentar SalesForce con información limpia y más "rica". En este caso concreto sólo vamos a indagar en el paso "SalesForce Input", este paso nos permitirá extraer datos de nuestro CRM, para posteriormente integrar datos de SalesForce en nuestro DataWarehouse. A continuación se describe la configuración de este paso y cada una de sus pestañas:

1) Pestaña Settings



En esta pestaña existen dos bloques:

Bloque de connection:

- **SalesForce Webservice URL**: En este campo se debe incluir la URL de conexión de nuestro salesForce. En nuestro caso es:
    - https://www.salesforce.com/services/Soap/u/20.0

      Pero el webservice que se utilice depende de:
        - La versión de la instancia de SalesForce
        - Llamada a SalesForce sin/con seguridad (http, https)
        - La versión del WSDL que tengas contratada (enterprise o partner).

      Otros ejemplos de llamadas al webservice de conexión de SalesForce pueden ser:
        - https://login.salesforce.com/services/Soap/c/22.0
        - http://login.salesforce.com/services/Soap/u/22.0

- **Username and Password**: En estos campos se incluyen en nombre de la cuenta con la que te quieres conectar a salesForce y su password respectivamente.

  NOTA: Lo más conveniente es crear tres variables en el fichero kettle.properties una conteniendo la URL del webservice, otra con el nombre de usuario y la ultima con la password, de esa manera podemos usar esas variables en todos los pasos de SalesForce y asi evitar incluir estos datos cada vez que se use un paso de SalesForce.

Bloque de Settings:

- **Specify query**: Si se marca este campo aparecerá un textarea en la parte de debajo de la pestaña que nos posibilita la inclusión de una query SOQL (Salesforce object query language), muy similar a SQL.

A continuación se puede observar la estructura básica de un select en SOQL:

```
SELECT f1, f2, …
FROM SalesForce_Module_with_read_permission
[WHERE  f1=xxx and f2=xxx…]
[WITH [DATA CATEGORY] filteringExpression]
[GROUP BY  fieldGroupByList]
[HAVING  havingConditionExpression]
[ORDER BY  fieldOrderByList ASC | DESC]
[LIMIT top_rows]
```

Como se puede observar en la estructura de la query este lenguaje de consulta es muy similar a SQL, pero tiene ciertas peculiaridades que salvan la diferencia. Algunas de ellas son:

- No es posible usar * como comodín de campos en la cláusula SELECT.
- No es necesario incluir “;” al final de la query.
- No se puede incluir comentarios tipo “--” en la query
- La cláusula “WITH DATA CATEGORY”, se utiliza para filtrar datos por diferentes jerarquías definidas en salesForce. Suponiendo una jerarquía Geográfica llamada Geo_Spain (PAIS -> CCAA -> PROVINCIA -> MUNICIPIO) se incluyen un par de ejemplos:

    - SELECT Nombre FROM Comerciales_Mod WHERE Actividad='online' WITH DATA CATEGORY Geo_Spain ABOVE Madrid
    - SELECT Nombre FROM Comerciales_Mod WHERE Actividad='online' WITH DATA CATEGORY Geo_Spain BELOW Cataluña
    - SELECT Nombre FROM Comerciales_Mod WHERE Actividad='online' WITH DATA CATEGORY Geo_Spain at (Madrid, Cataluña)
    - SELECT Nombre FROM Comerciales_Mod WHERE Actividad='offline' WITH DATA CATEGORY Geo_Spain at (Barcelona, Tarragona)

- La cláusula GROUP BY”, sólo esta disponible en desde la versión 18.0 de SalesForce, y tiene el mismo uso que en una query SQL regular.

Si quieres saber más detalles sobre el lenguaje de consulta SOQL, accede a su manual completo en:

http://www.salesforce.com/us/developer/docs/api/Content/sforce_api_calls_soql.htm

- **Module**: Si se desmarca el campo “Specify Query”, en el campo Module aparecerán todos los módulos a los que tiene permiso el usuario que se ha conectado, incluso los custom modules que han sido definido por los propios usuarios de SalesForce.

  NOTA: En cada paso “SalesForce Input” sólo es posible extraer datos de un módulo.

- **Query Condition**: En este campo se pueden incluir condiciones sobre la extracción de datos del modulo indicado en al campo “Module”. Se corresponde con el contenido de la cláusula WHERE idéntica a SQL:

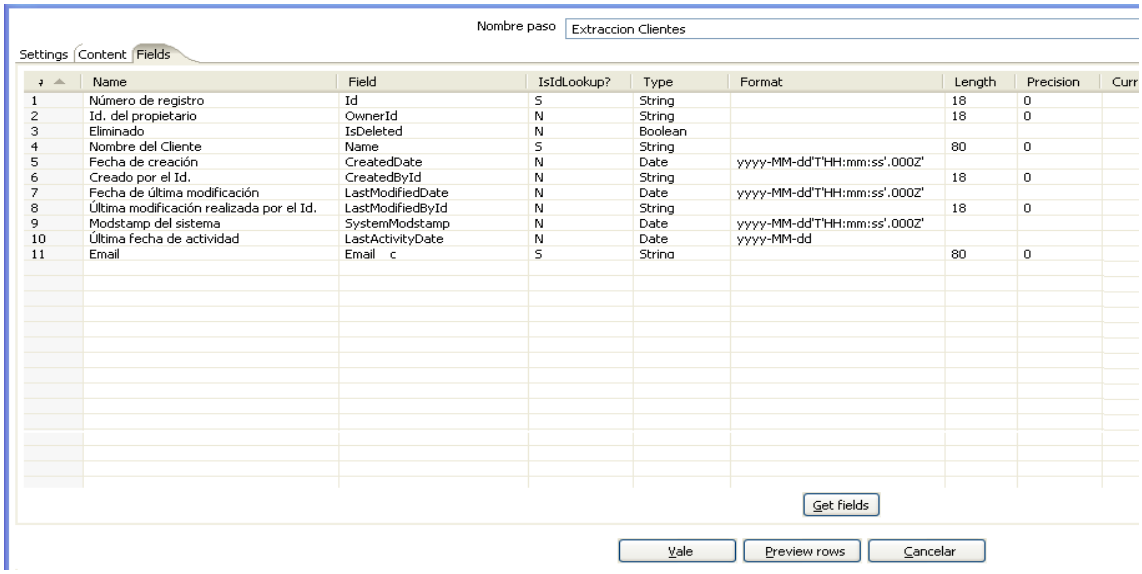
NOTA: No es necesario incluir la palabra reservada WHERE.

2) Pestaña Fields.

En esta pestaña existe un grid en el que se deben especificar los campos del módulo de SalesForce del que queremos extraer datos, también es necesario incluir su tipo, formato, tamaño y todas las propiedades típicas que nos encontramos en los pasos de PDI.

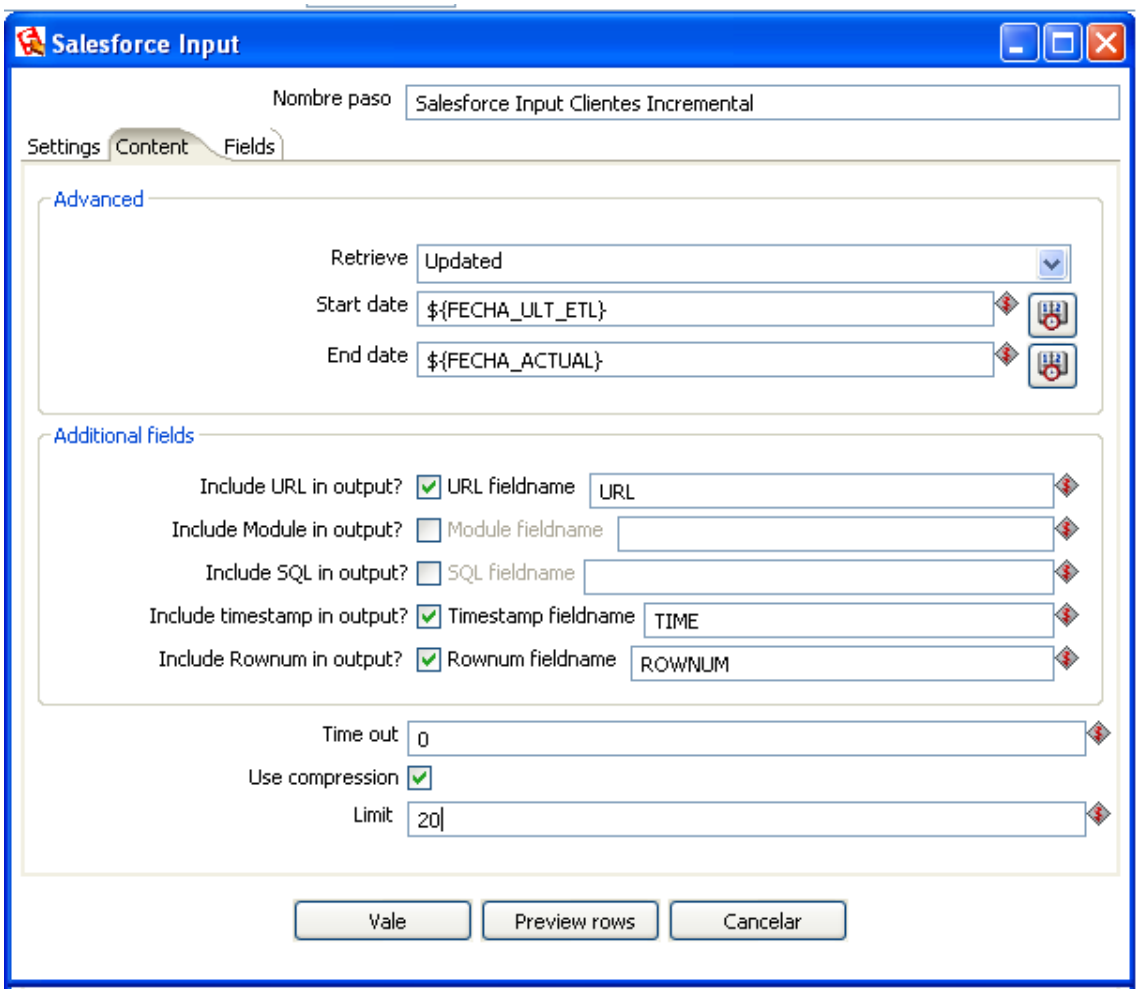
Durante el desarrollo de la ETL posiblemente haya que modificar la query SOQL, debido a nuevos requisitos del cliente, o puede que incluso tengamos que cambiar el módulo de donde se van a extraer los datos, tras esto es muy necesario recordar que es necesario refrescar los campos que se encuentran en el grid, ya que sino lo hacemos y estos campos no se corresponden con lo que estamos extrayendo al intentar lanzar la ETL o hacer "preview" del paso en cuestión se producirá un error.

Para facilitar la tarea de inclusión de campos en el grid, existe la opción "Get Fields" que permitirá refrescar el contenido del grid, con los campos y tipos apropiados.



#	Name	Field	IsIdLookup?	Type	Format	Length	Precision	Curr
1	Número de registro	Id	S	String		18	0	
2	Id. del propietario	OwnerId	N	String		18	0	
3	Eliminado	IsDeleted	N	Boolean				
4	Nombre del Cliente	Name	S	String		80	0	
5	Fecha de creación	CreatedDate	N	Date	yyyy-MM-dd'T'HH:mm:ss'.000Z'			
6	Creado por el Id.	CreatedById	N	String		18	0	
7	Fecha de última modificación	LastModifiedDate	N	Date	yyyy-MM-dd'T'HH:mm:ss'.000Z'			
8	Última modificación realizada por el Id.	LastModifiedById	N	String		18	0	
9	Modstamp del sistema	SystemModstamp	N	Date	yyyy-MM-dd'T'HH:mm:ss'.000Z'			
10	Última fecha de actividad	LastActivityDate	N	Date	yyyy-MM-dd			
11	Email	Email__c	S	String		80	0	

3) Pestaña Content.



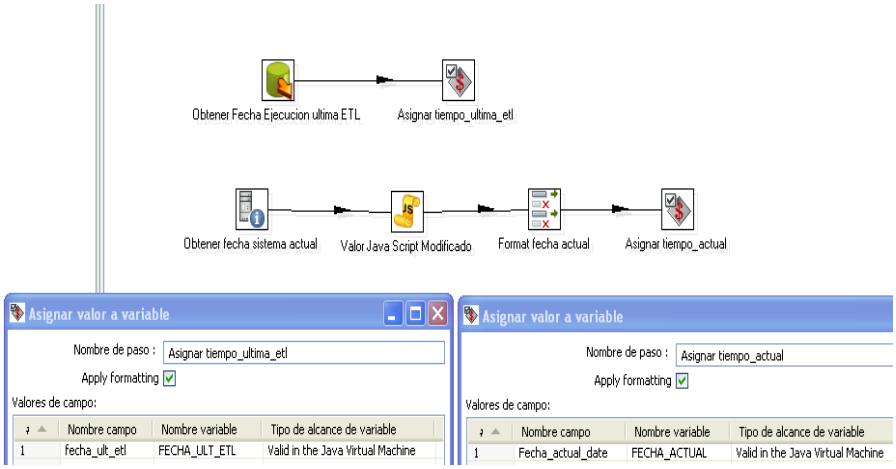
Bloque Advanced:

- El parametros “Retrieve”, “Start date” y “End date” describen que tipo de registros se van a extraer:
    - All: Todos los registros en ese módulo.
    - Deleted: Los registros eliminados de ese módulo, entre las fechas “Start date” y “End date”.
    - Updated: Los registros nuevos o actualizados de ese módulo, entre las fechas “Start date” y “End date”.

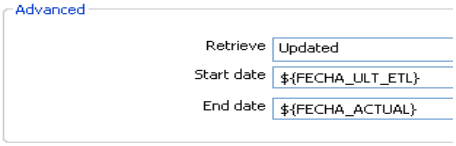
    NOTA: Los parámetros “Start date” y “End date” tienen el formato YYYY-mm-dd HH:MM:SS y pueden ser variables de PDI extraídas del fichero kettle.properties o de una tabla de la base de datos.

Estos parámetros son muy útiles si se quieren hacer extracciones incrementales. Una posible solución para ello sería:

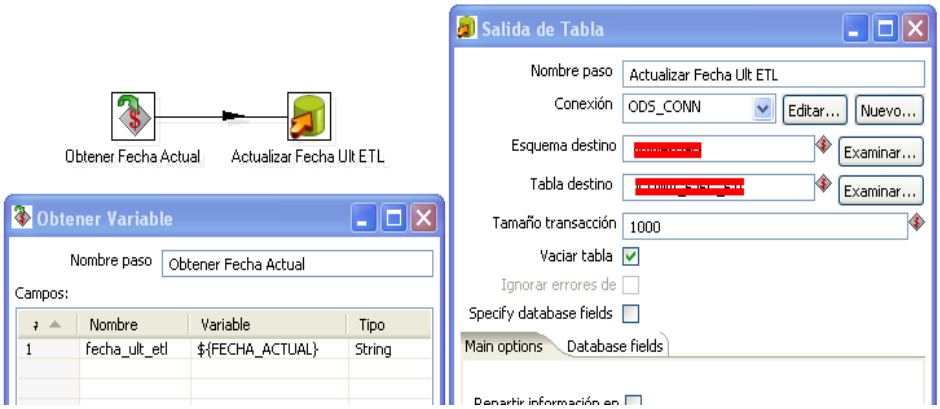
1) Crear una transformación al comenzar la ETL que coga de una tabla de BBDD la fecha de la última ejecución de la ETL y la fecha actual del sistema y las incluya en dos variables:



2) En el paso de “SalesForce Input” y concretamente en el bloque “Retrieve” de la pestaña “Content”, se incluyen las variables creadas en el paso anterior:



3) Antes de finalizar la ETL recuerda crear una transformación que guarde en BBDD la fecha actual, para que en la siguiente ejecución de la ETL se extraigan los datos incrementalmente:



Bloque Additional fields:

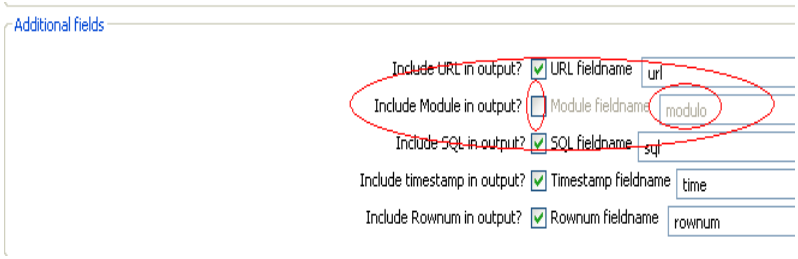
Este bloque nos permite extraer datos adicionales a los del módulo o query SOQL. Estos campos son:

- **URL servicio web**. Será la URL del webservice de conexión de SalesForce.
- **Modulo de SalesForce de donde se estan extrayendo los datos**. Correspondiente al valor seleccionado en el combo "Module". Si no elegimos el módulo mediante el combo y lo hacemos mediante una consulta SOQL, este campo se desmarcará automáticamente.
- **SQL generada**. Sólo la parte del "SELECT" de la query SOQL.
- **Marca de Tiempo** formato YYYY/mm/dd HH:MM:SS.
- **Número de fila extraído**. Comienza en cero.

Hay algunos puntos a destacar, referentes al uso de este bloque:

- Si marcas la inclusión de un campo adicional, incluye en el campo correspondiente el nombre de dicho campo.
- Si tras incluir algún campo adicional, vamos a la "Pestaña Fields" y pulsamos en "Get Fields" no esperemos que los campos adicionales aparezcan en el grid central, "esto no va a ocurrir", pero no caer en preocupación por que aunque no aparezcan en dicho grid, si aparecerán en la salida del paso.
- No es posible incluir condiciones que hagan referencia a alguno de estos campos en el textarea "Query condition" de la "Pestaña Settings".
- Antes de desmarcar un campo concreto, borra el nombre de dicho campo, por que puede producir una salida incorrecta, por ejemplo, imaginemos que no

queremos ver el campo del módulo y que lo dejamos así:

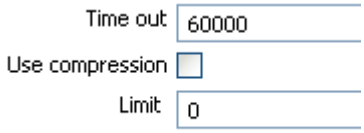


La salida que veremos será algo como:

url	sql	time	rownum
https://www.salesforce.com/services/Soap/u/20.0	Cliente__c		

Sólo se devolverá el campo URL correctamente, ya se produce un desplazamiento de valores, es decir, el campo "SQL" tendrá el contenido del campo "Módulo", en el campo "Timestamp" se intentará meter el contenido del campo "SQL" y fallará por que no es de tipo fecha sino cadena de texto, y por último dentro del campo "Rownum" se intentará meter el campo de "TimeStamp" que también fallará por incongruencia de tipos, ya que el campo "Rownum" es de tipo número.

Opciones Time Out, Limit y Compression:



- Parámetro "**Time out**", es útil para limitar la ejecución de la extracción de tiempo en milisegundos, ya que como sabemos la ejecución de un webservice es asíncrona y en momentos puntuales pueden aparecer tiempos de espera grandes. Si se pone 0 en este parámetro el tiempo será ilimitado.

- Parámetro "**Use Compression**". Permite comprimir el envió de mensajes SOAP a formato gzip, este parámetro mejorará el performance de la extracción de datos. Usa el Standard

HTTP 1.1 para compression de datos. Cabe destacar que no todas las herramientas o sistemas admiten compresión, en el caso que nos atañe PDI, si, pero si quieres asegurarte consulta:

http://wiki.developerforce.com/index.php/Tools

- Parámetro **"Limit"**, permite limitar el número de filas a devolver. Imagine que sólo le interesan para el análisis, un conjunto concreto de sucursales de su empresa, por ejemplo el "top 10 en número de personal". Para ello se puede introducir una query SOQL en la "Pestaña Settings" que agrupe por sucursal y cuente el número diferente de personas, se ordene el resultado por dicho número de personas diferentes y se limite el resultado a diez filas.

  NOTA: Este parámetro puede ser también incluido en la query SOQL. Si se incluye en los dos sitios a la vez, pueden existir incongruencias en el número de filas a extraer.

A. Información Stratebi

Stratebi es una empresa española, radicada en Madrid y Barcelona, líderes en España en soluciones Business Intelligence Open Source.

En Stratebi nos planteamos como **objetivo** dotar a las compañías e instituciones, de herramientas escalables y adaptadas a sus necesidades, que conformen una estrategia Business Intelligence capaz de rentabilizar la información disponible. Para ello, nos basamos en el desarrollo de soluciones de Inteligencia de Negocio, mediante tecnología Open Source.

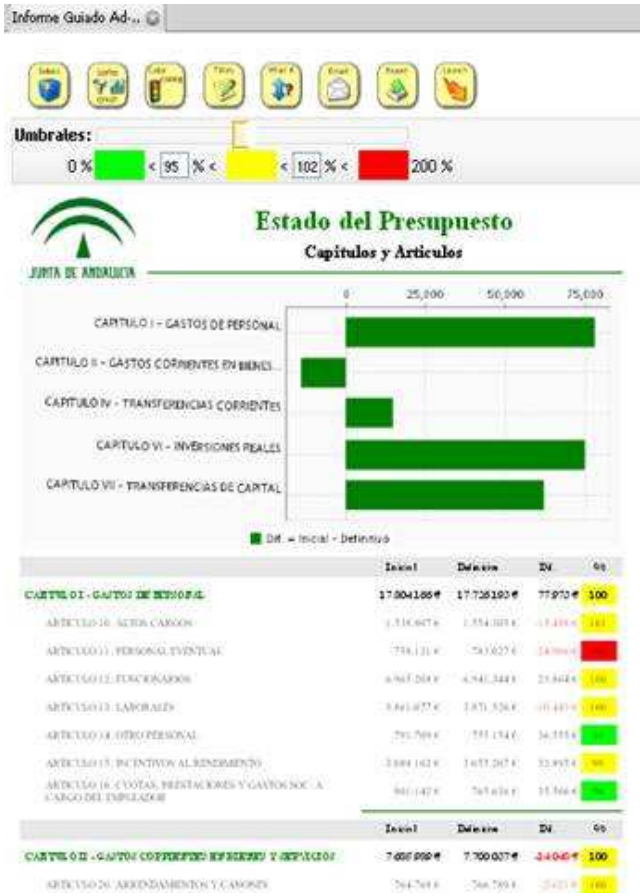
Stratebi **son** profesores y responsables de proyectos **del Master en Business Intelligence de la Universidad UOC.**

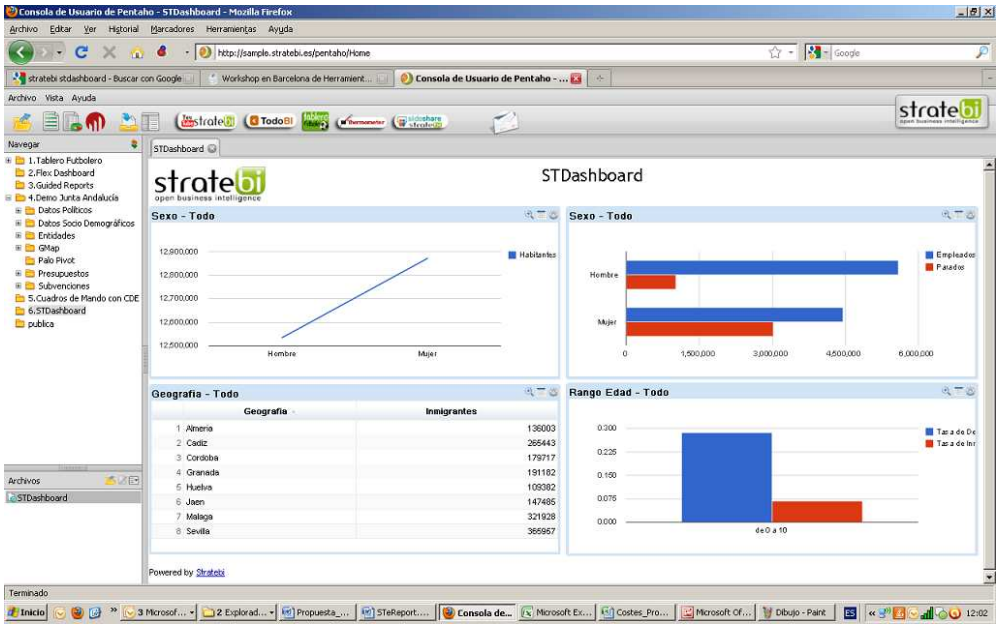
Los profesionales de Stratebi son los creadores y autores del primer weblog en español sobre el mundo del Business Intelligence, Data Warehouse, CRM, Dashboards, Scorecard y Open Source.

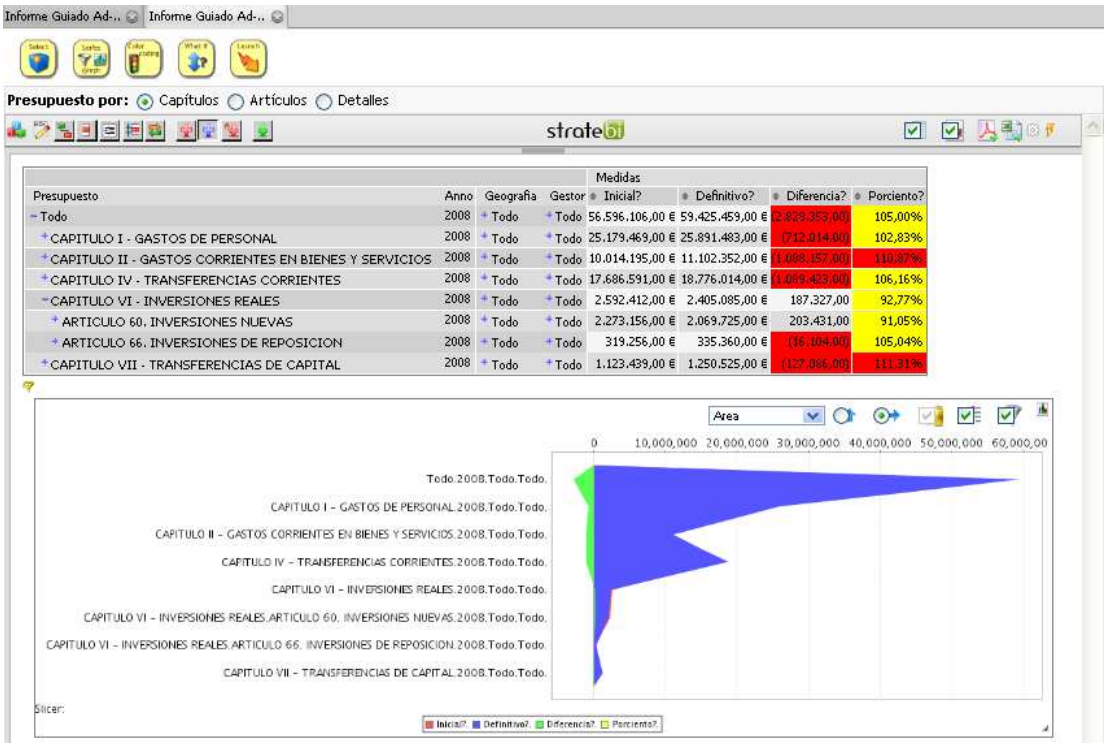
Todo Bi, se ha convertido en una referencia para el conocimiento y divulgación del Business Intelligence en español.

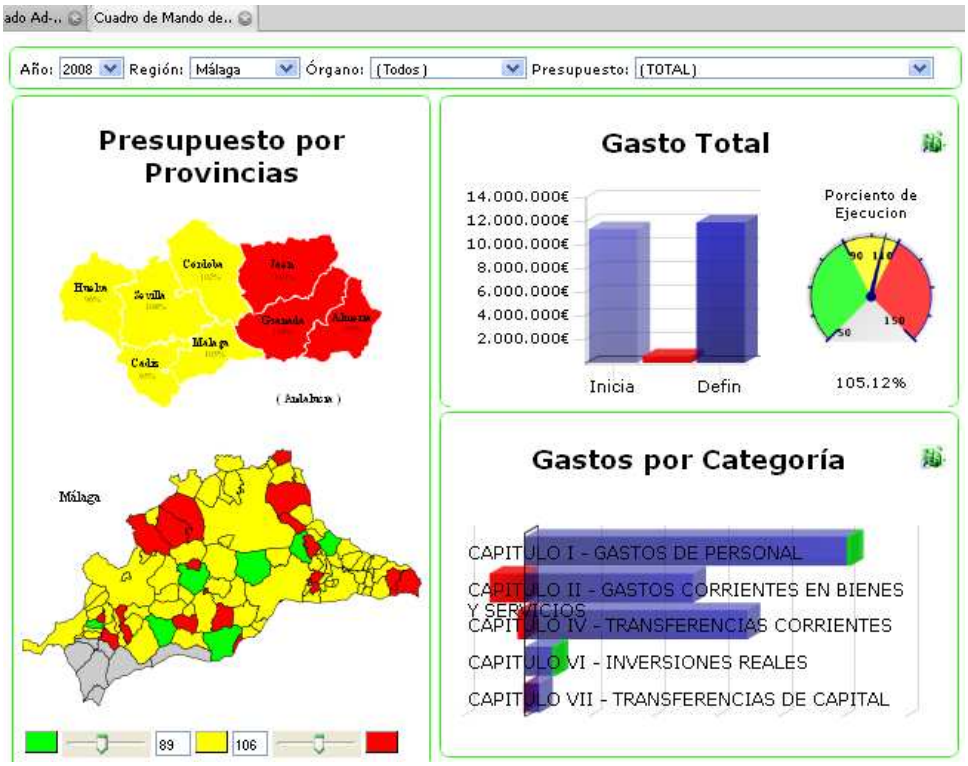
Stratebi ha sido elegida como **Caso Éxito del Observatorio de Fuentes Abiertas de Cenatic**.

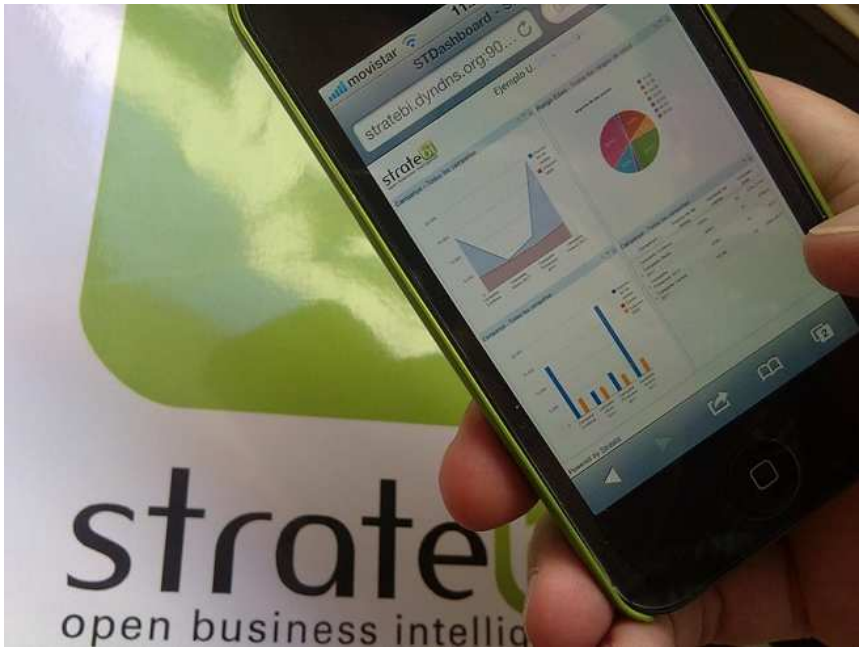
B. Ejemplos


Informe Guiado Ad-...
Umbrales:
0 %
95 %
102 %
200 %
Estado del Presupuesto
Capitulos y Articulos
JUNTA DE ANDALUCIA
CAPITULO I - GASTOS DE PERSONAL
CAPITULO II - GASTOS CORRIENTES EN BIENES...
CAPITULO IV - TRANSFERENCIAS CORRIENTES
CAPITULO VI - INVERSIONES REALES
CAPITULO VII - TRANSFERENCIAS DE CAPITAL
Dif. = Inicial - Definitivo


Consola de Usuario de Pentaho - STDashboard - Mozilla Firefox
STDashboard
Sexo - Todo
Sexo - Todo
Geografia - Todo
Rango Edad - Todo
Powered by Stratebi


Informe Guiado Ad-..
Informe Guiado Ad-..
Presupuesto por: Capítulos Artículos Detalles
Medidas
Presupuesto | Anno | Geografía | Gestor | Inicial? | Definitivo? | Diferencia? | Porciento?
Todo | 2008 | Todo | Todo | 56.596.106,00 € | 59.425.459,00 € | (2.829.353,00) | 105,00%
CAPITULO I - GASTOS DE PERSONAL | 2008 | Todo | Todo | 25.179.469,00 € | 25.891.483,00 € | (712.014,00) | 102,83%
CAPITULO II - GASTOS CORRIENTES EN BIENES Y SERVICIOS | 2008 | Todo | Todo | 10.014.195,00 € | 11.102.352,00 € | (1.088.157,00) | 110,87%
CAPITULO IV - TRANSFERENCIAS CORRIENTES | 2008 | Todo | Todo | 17.686.591,00 € | 18.776.014,00 € | (1.089.423,00) | 106,16%
CAPITULO VI - INVERSIONES REALES | 2008 | Todo | Todo | 2.592.412,00 € | 2.405.085,00 € | 187.327,00 | 92,77%
ARTICULO 60. INVERSIONES NUEVAS | 2008 | Todo | Todo | 2.273.156,00 € | 2.069.725,00 € | 203.431,00 | 91,05%
ARTICULO 66. INVERSIONES DE REPOSICION | 2008 | Todo | Todo | 319.256,00 € | 335.360,00 € | (16.104,00) | 105,04%
CAPITULO VII - TRANSFERENCIAS DE CAPITAL | 2008 | Todo | Todo | 1.123.439,00 € | 1.250.525,00 € | (127.086,00) | 111,31%
Area
0 10,000,000 20,000,000 30,000,000 40,000,000 50,000,000 60,000,000
Todo.2008.Todo.Todo.
CAPITULO I - GASTOS DE PERSONAL.2008.Todo.Todo.
CAPITULO II - GASTOS CORRIENTES EN BIENES Y SERVICIOS.2008.Todo.Todo.
CAPITULO IV - TRANSFERENCIAS CORRIENTES.2008.Todo.Todo.
CAPITULO VI - INVERSIONES REALES.2008.Todo.Todo.
CAPITULO VI - INVERSIONES REALES.ARTICULO 60. INVERSIONES NUEVAS.2008.Todo.Todo.
CAPITULO VI - INVERSIONES REALES.ARTICULO 66. INVERSIONES DE REPOSICION.2008.Todo.Todo.
CAPITULO VII - TRANSFERENCIAS DE CAPITAL.2008.Todo.Todo.
Slicer:
Inicial? Definitivo? Diferencia? Porciento?


ado Ad-..
Cuadro de Mando de..
Año: 2008 Región: Málaga Órgano: (Todos) Presupuesto: (TOTAL)
Presupuesto por Provincias
Córdoba
Jaén
Huelva
Sevilla
Granada
Almería
Málaga
Cádiz
(Andalucía)
Málaga
89
106
Gasto Total
14.000.000€
12.000.000€
10.000.000€
8.000.000€
6.000.000€
4.000.000€
2.000.000€
Inicia
Defin
Porciento de Ejecucion
90
110
50
150
105.12%
Gastos por Categoría
CAPITULO I - GASTOS DE PERSONAL
CAPITULO II - GASTOS CORRIENTES EN BIENES Y SERVICIOS
CAPITULO IV - TRANSFERENCIAS CORRIENTES
CAPITULO VI - INVERSIONES REALES
CAPITULO VII - TRANSFERENCIAS DE CAPITAL


movistar
STDashboard
stratebi.dyndns.org:90...
open business intelli