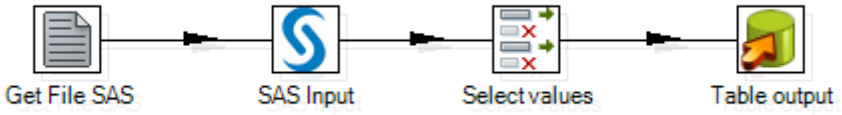
Pentaho y SAS: Obteniendo Datos de SAS y explotándolos con Pentaho

Con este post, queremos mostraros un nuevo Plugin que trae consigo Pentaho Data Integration a partir de su versión 4.4, y que puede ser de utilidad en aquellas organizaciones donde se utilice SAS como herramienta de Business Analytics y quieran explotar la información que proporciona SAS con Pentaho.

Lo comentado en el punto anterior, no es más que una muestra de lo que Pentaho viene haciendo desde sus orígenes dentro de la estrategia de ampliar sus productos con nuevas funcionalidades así como su ecosistema de herramientas.

El objetivo que nos marcamos, es leer con Pentaho Data Integration un fichero de SAS para posteriormente explotar su información utilizando AgileBI desde Pentaho Data Integration.

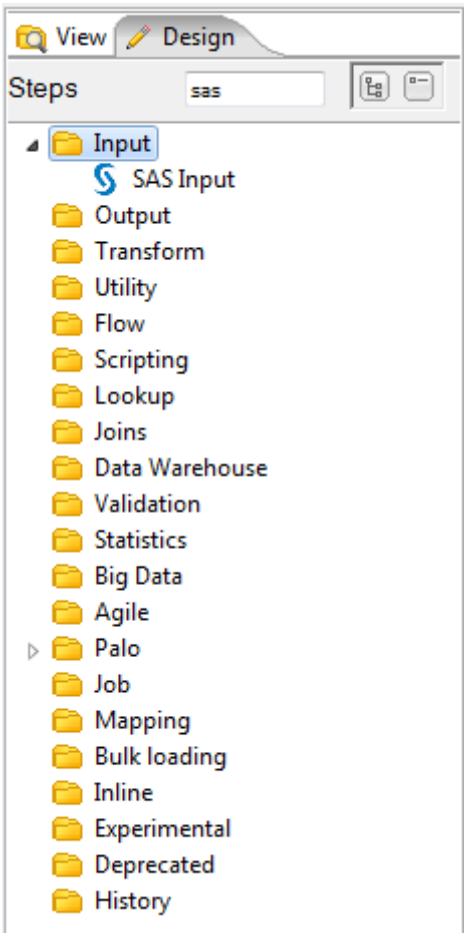


Para el ejemplo, tomamos como referencia unos datos publicados por el investigador Paul Dickman en el año 2004, relativos a la supervivencia de personas con cáncer de colon diagnosticado en Finlandia entre los años 1975 y 1994. Para más información, facilitamos el enlace http://biostat3.net/download/sas/readme.pdf

Sobre los datos se manejan los siguientes conceptos:

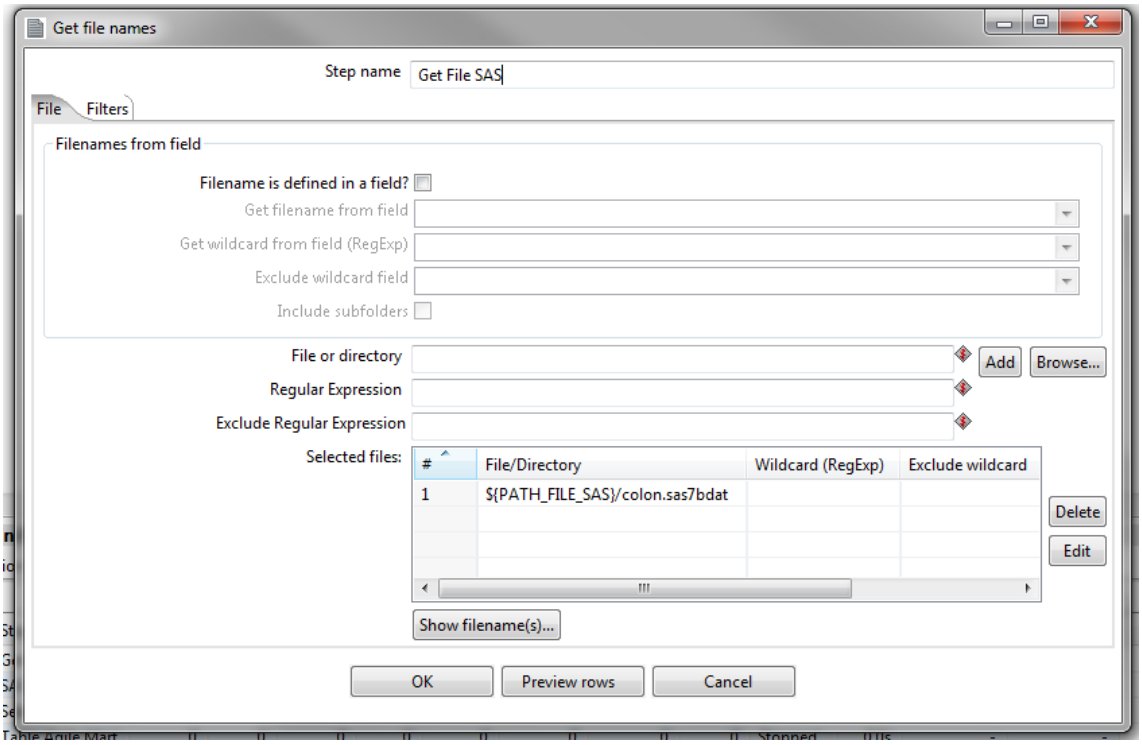
- Edad del individuo
- Fecha de diagnóstico
- Fecha de salida
- Mes de la diagnosis
- Sexo
- Estado clínico de Diagnosis
- Localización del tumor
- Tiempo de supervivencia en meses
- Tiempo de supervivencia en años

Lógicamente, como estamos referenciando a datos, el nuevo plugin está disponible a nivel de transformación de datos tal y como se puede ver en la siguiente imagen.

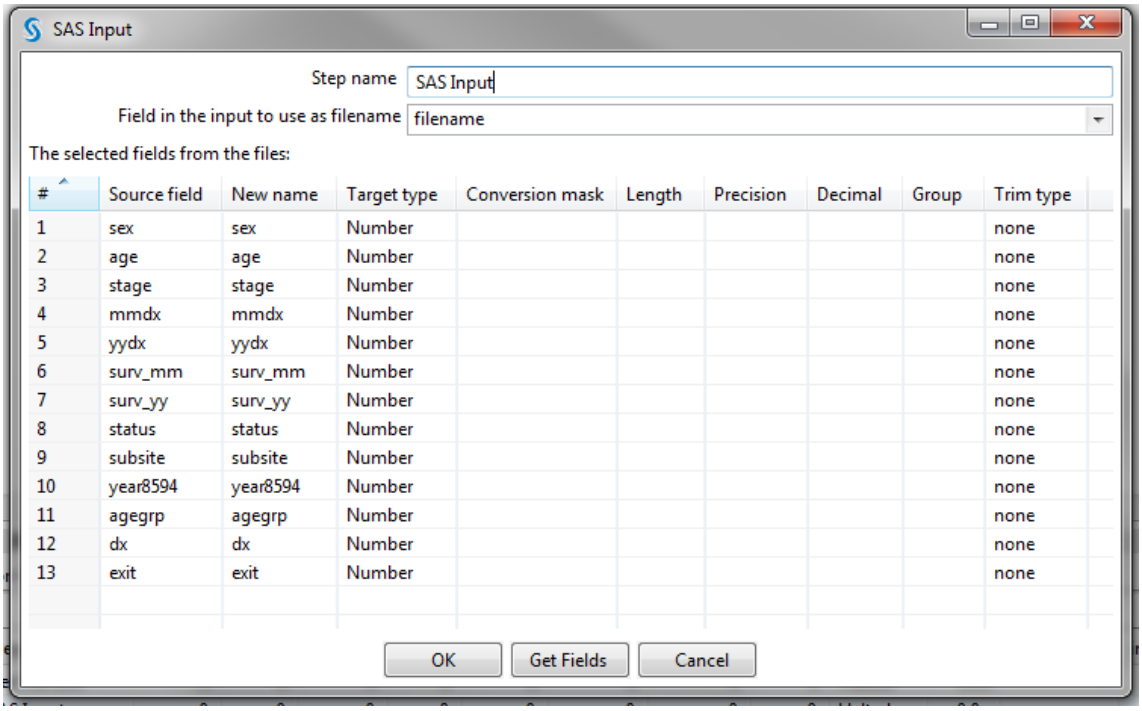


Los componentes de Pentaho Data Integration utilizados han sido los siguientes:

- **Get File Names:** Para indicar la ruta y nombre de fichero SAS que contiene los datos.

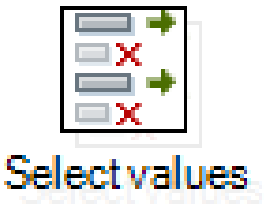


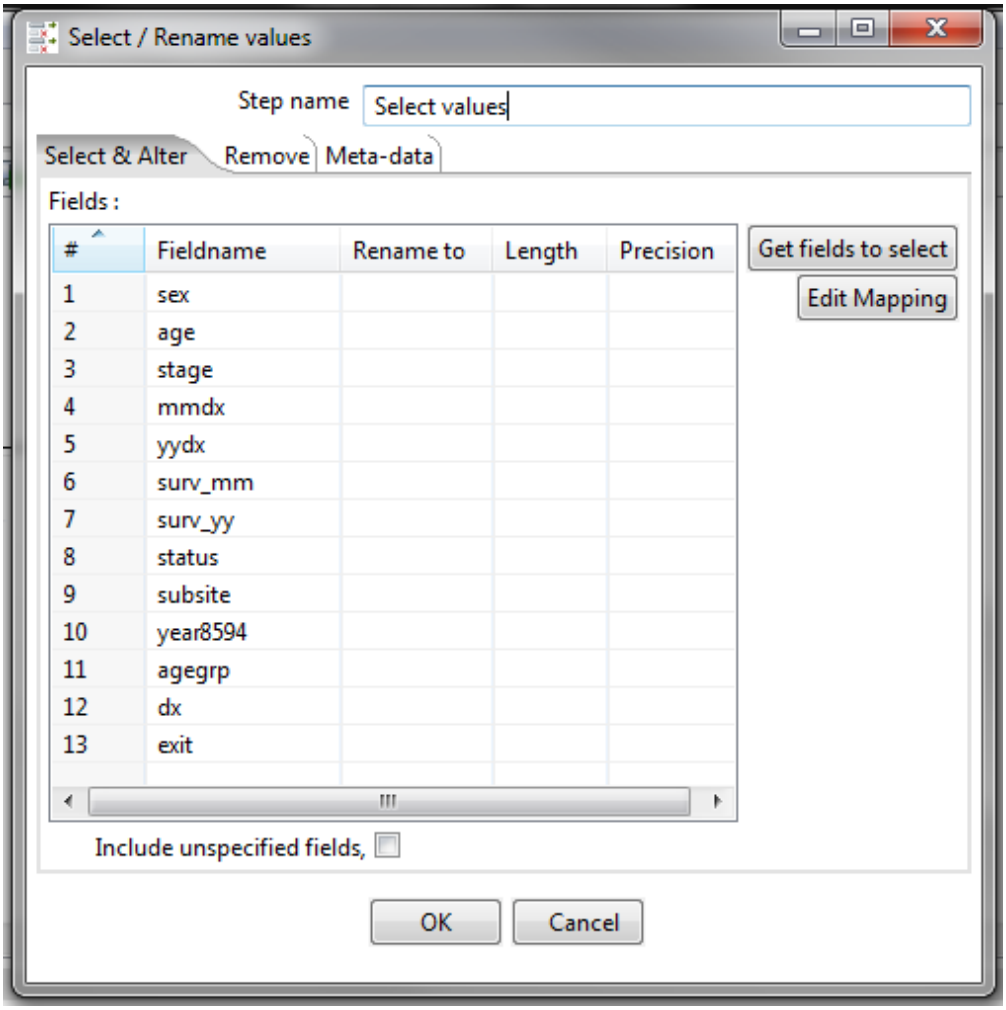
- **SAS Input:** En primer lugar hay que traer los campos, para ello pulsamos el botón "Get Fields", y posteriormente si lo consideramos necesario cambiamos las propiedades que la herramienta ha identificado.



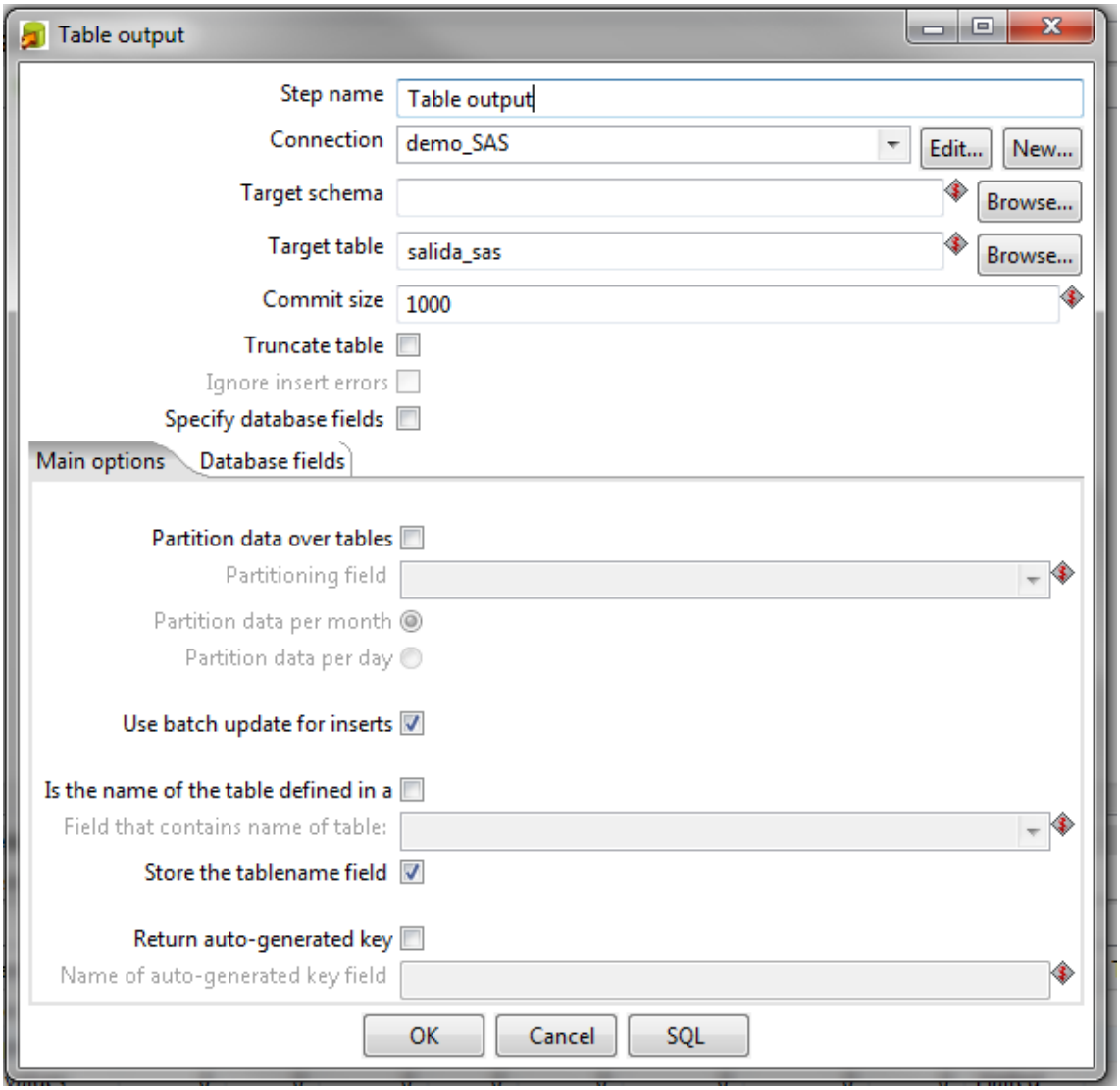
#	Source field	New name	Target type	Conversion mask	Length	Precision	Decimal	Group	Trim type
1	sex	sex	Number						none
2	age	age	Number						none
3	stage	stage	Number						none
4	mmdx	mmdx	Number						none
5	yydx	yydx	Number						none
6	surv_mm	surv_mm	Number						none
7	surv_yy	surv_yy	Number						none
8	status	status	Number						none
9	subsite	subsite	Number						none
10	year8594	year8594	Number						none
11	agegrp	agegrp	Number						none
12	dx	dx	Number						none
13	exit	exit	Number						none

- **Select values:** Para indicar que campos del flujo queremos seleccionar con la intención de volcarlos a una tabla.

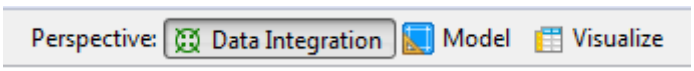




- **Table output:** Finalmente y previo a su análisis, vamos a volcar los datos a una tabla. En este caso vamos a usar MySQL. En caso de que no exista la tabla, tenemos la opción de crearla automáticamente seleccionando el botón SQL.

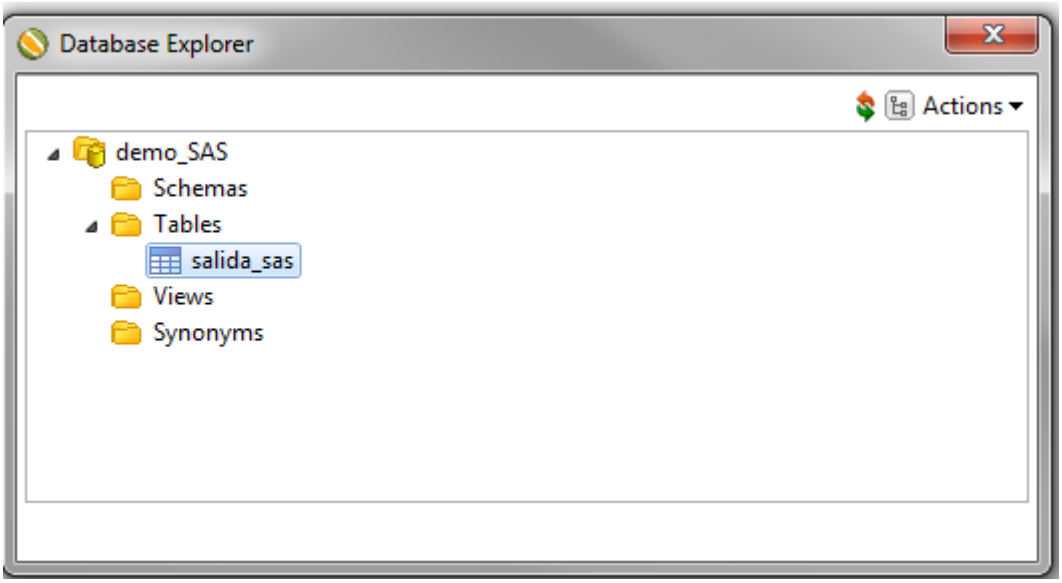


Una vez que tenemos los datos en una tabla, vamos a visualizarlos usando la propia herramienta y de manera muy sencilla. En la parte superior derecha, tenemos la opción de cambiar de perspectiva. Hasta ahora, hemos estado construyendo nuestra transformación y por tanto estábamos trabajando con la vista Data Integration.



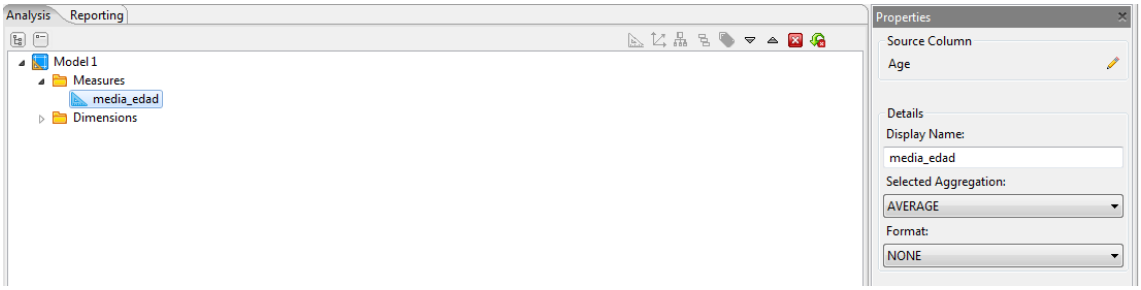
Al seleccionar la opción Model podremos visualizar de manera rápida los datos de nuestro modelo, en una tabla dinámica o construirnos un informe usando un asistente. Para ello, tenemos las pestañas Analysis y Reporting que inicialmente presentan el siguiente aspecto.

Lo primero, será asignar un nombre al modelo en Model Name y definir un Datasource. Nuestro origen de datos será la tabla MySQL que hemos definido anteriormente.

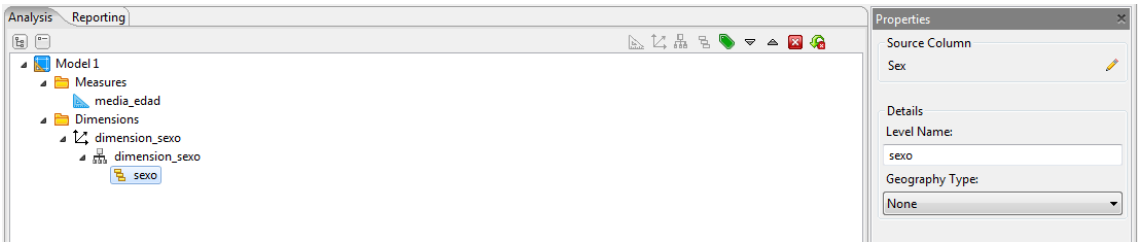


Seguidamente, empezamos con la sección Analysis que nos permitirá definir una estructura para explotarla mediante una tabla dinámica. Es necesario definir al menos una métrica y una dimensión.

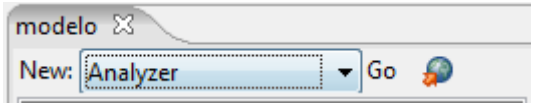
Definimos una métrica que llamamos media_edad y cuya agregación es la media. Decimos que esta métrica se corresponde con el campo de BBDD Age. De esta manera podremos analizar las edades de los individuos por sexo.



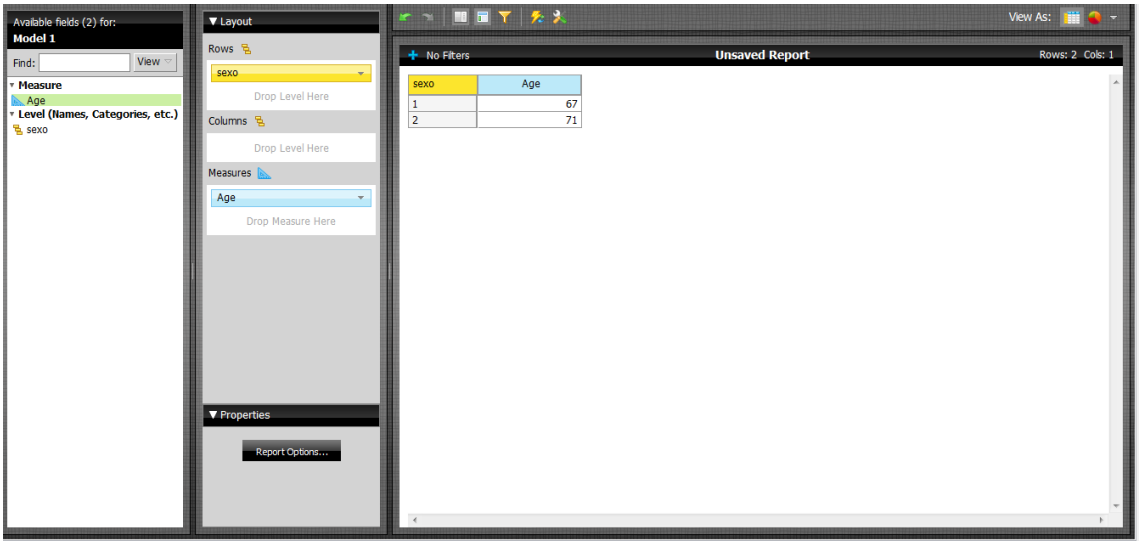
De la misma manera definimos una dimensión. Tendrá una jerarquía y nivel que llamaremos sexo. El campo origen de la BBDD es Sex.

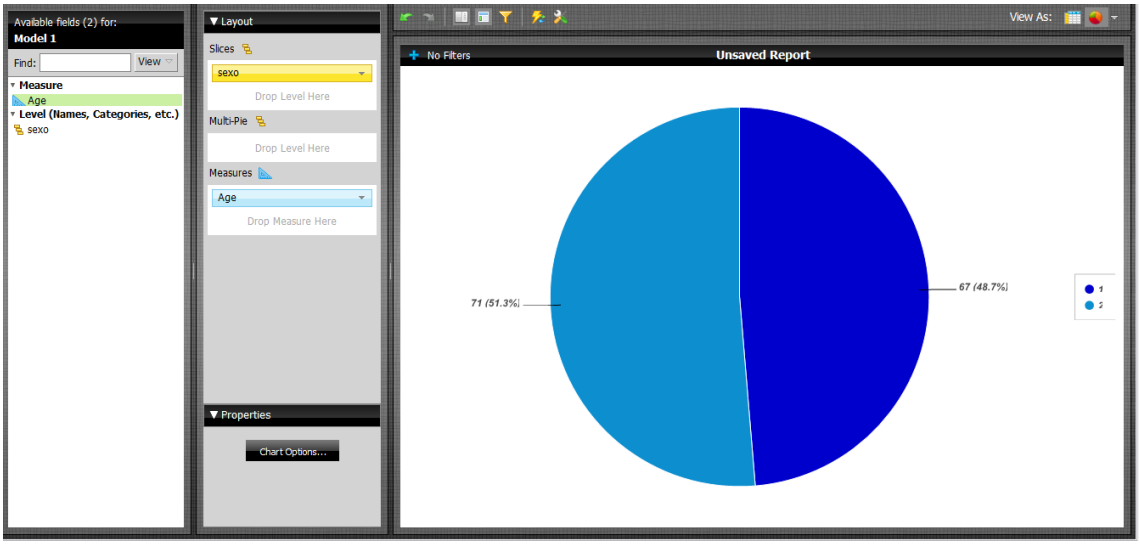


Para visualizar lo que hemos definido le damos a Go y automáticamente pasaremos a la perspectiva Visualize.

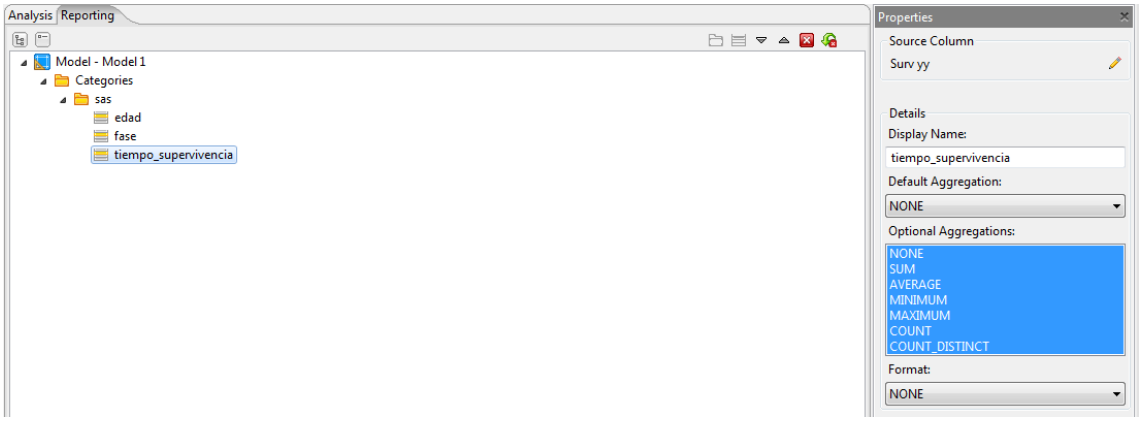


Mediante drag & drop vamos colocando los elementos y podemos visualizarlos tanto en tabla como gráfico (seleccionando la opción de la parte superior derecha).

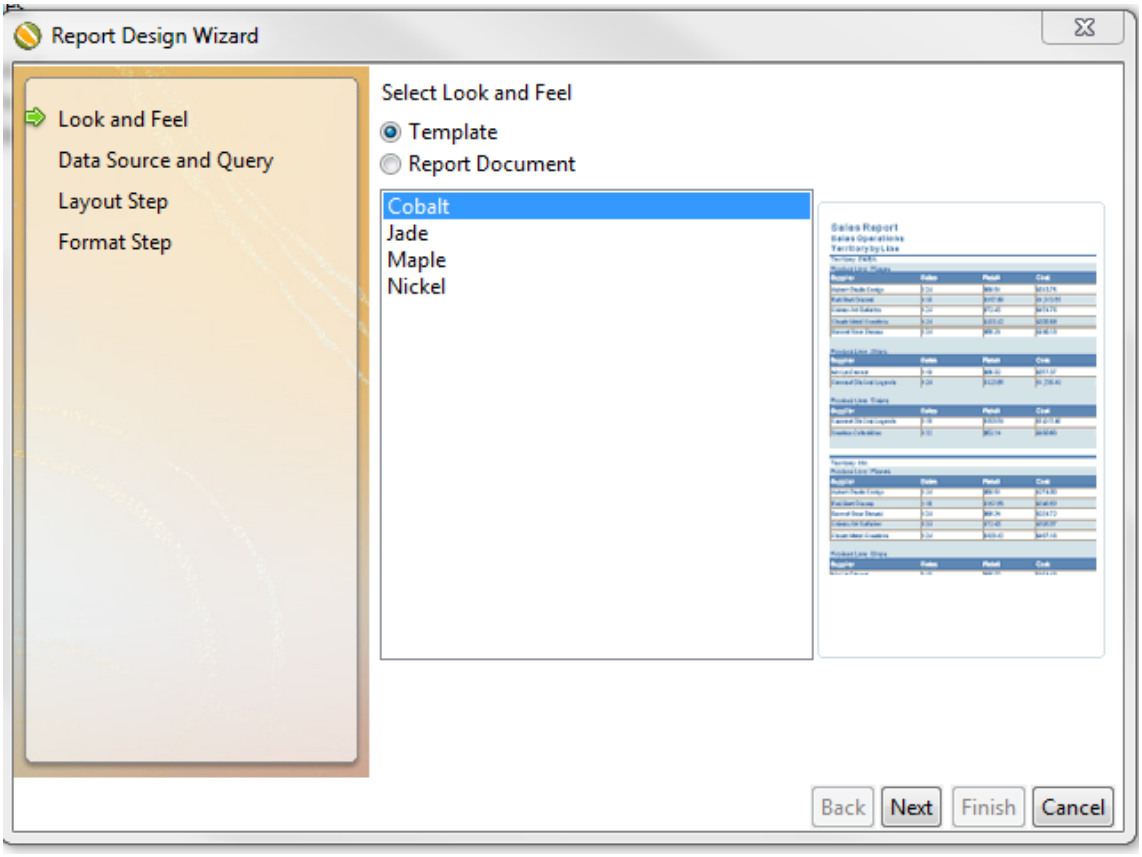




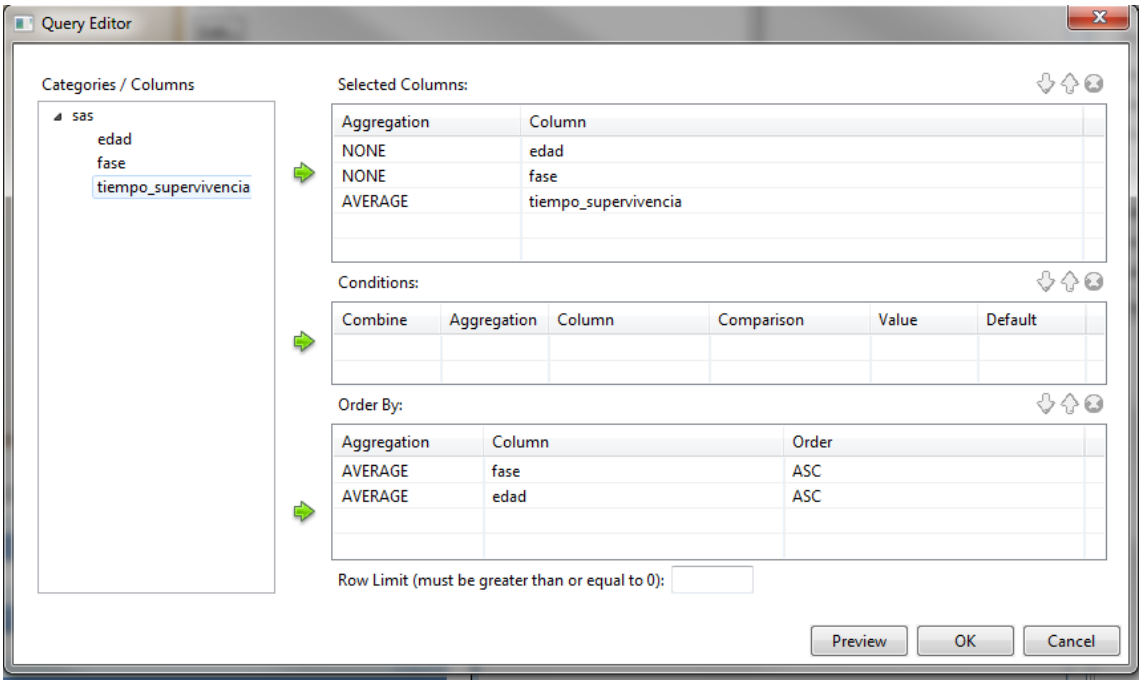
A continuación, vamos a definir la sección de Reporting. Aquí debemos definir una categoría e indicar los campos que queremos tener disponibles. Cogeremos los campos edad, que se corresponde con el campo Age de BBDD, fase de la enfermedad que es el campo de BBDD Stage y tiempo de supervivencia en años, que es el campo de BBDD surv_yy. Para cada campo, podemos definir un tipo de agregación e incluso aplicar una máscara de formato.

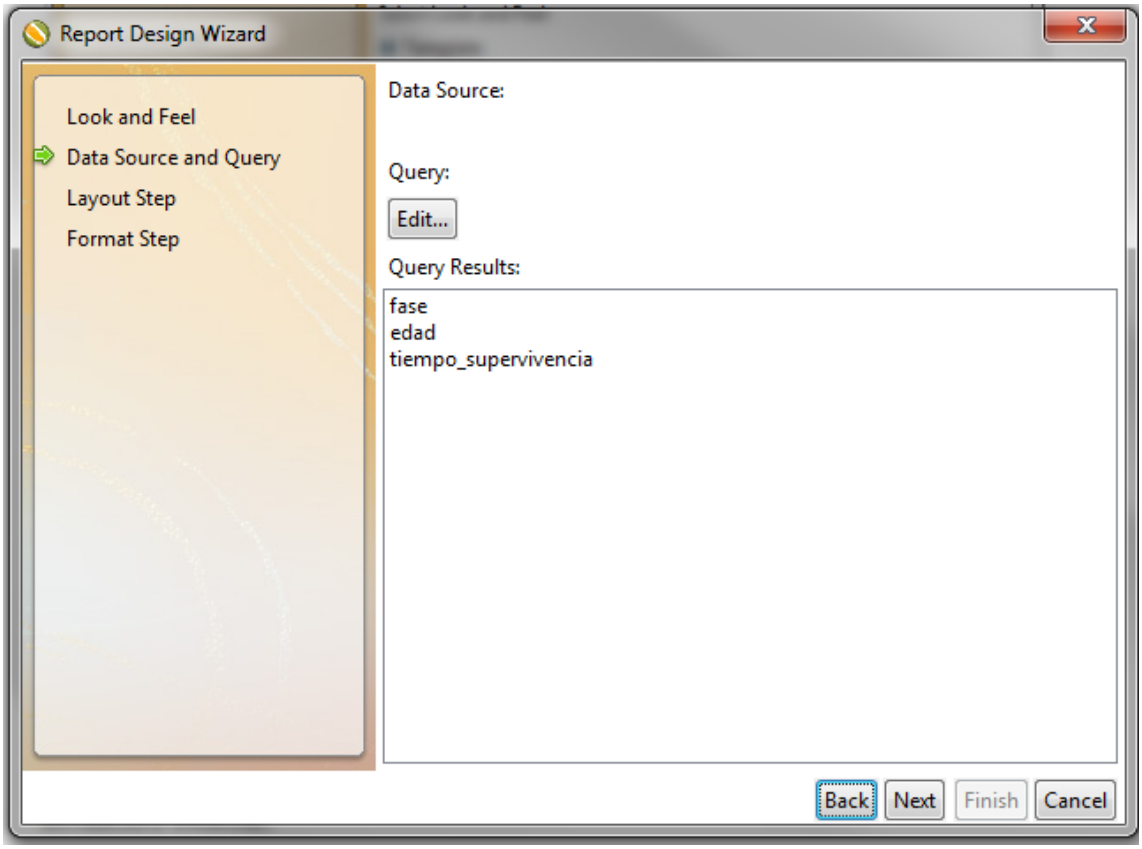


Para visualizar lo que hemos definido, seleccionamos Report Wizard y le damos a Go. Automáticamente pasaremos a la perspectiva Visualize. Lo primero que tenemos que hacer es seleccionar una plantilla.

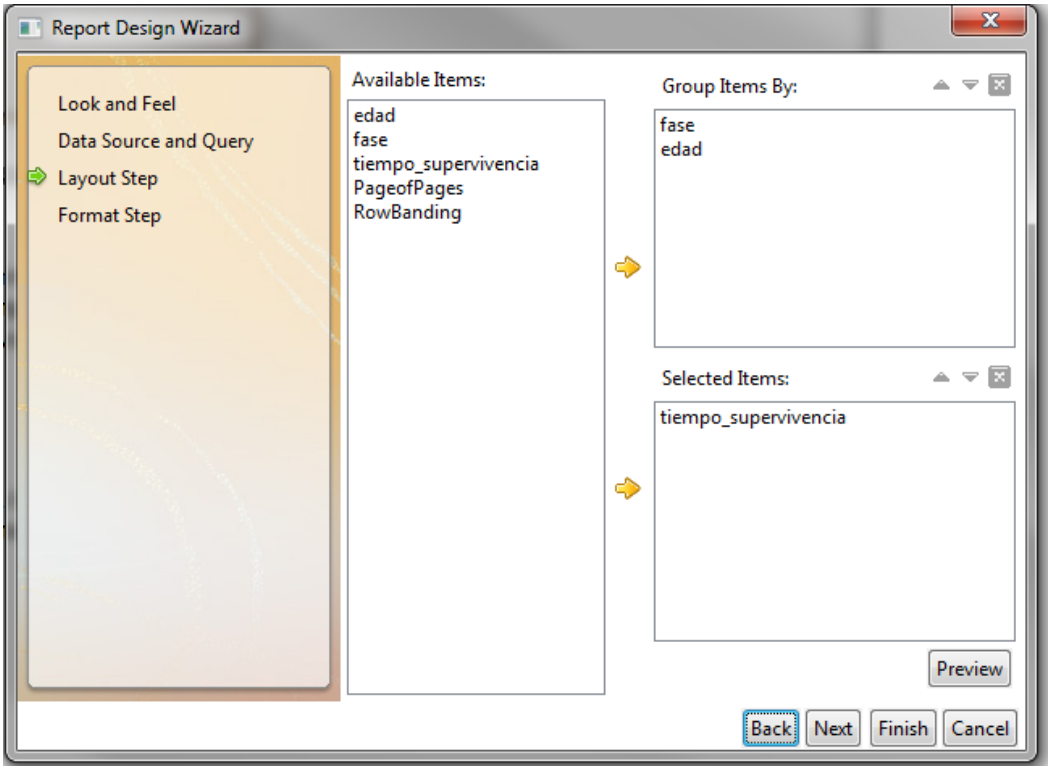


Seguidamente, indicamos nuestro origen de datos y los campos que queremos utilizar en nuestro informe. Vamos a obtener la edad, fase y media de años de supervivencia ordenados por fase de la enfermedad y edad del individuo.

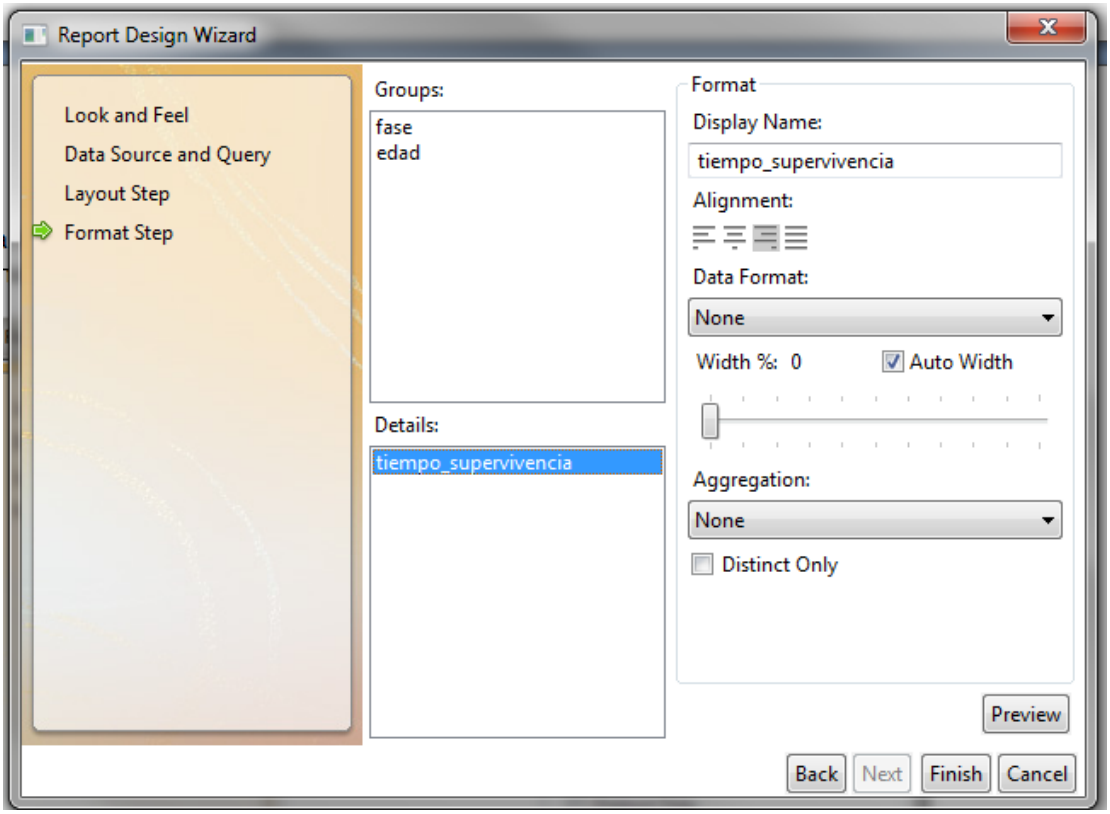




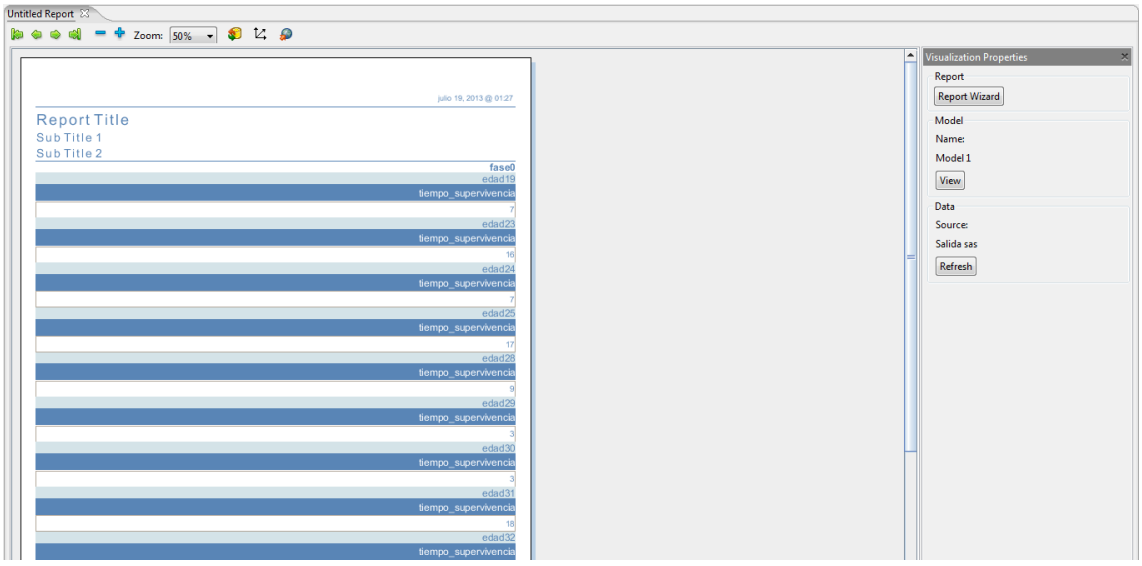
A continuación, indicamos la disposición de los campos en el informe. Queremos mostrar los resultados agrupados por fase y edad del individuo.



Para terminar, podemos aplicar opciones de alineación de elementos en el informe o formateo de datos.



Finalmente, ya tenemos nuestro informe creado de manera automática con unos sencillos pasos. Con el botón Report Wizard situado en la paleta de la derecha, podemos volver a editar el informe.



Este modelo, al igual que la transformación creada, podemos guardarlo con extensión xmi y reutilizarlo tantas veces como queramos.