Trabajo de Modelado Multidimensional

Nombre: Freddy Leonardo Abad León

1. **Descripción del problema**

Control de Calidad de Encuesta

Descripción:

Generar un modelo multidimensional de una encuesta realizada a la población de ciertos sectores de una ciudad, la encuesta tiene 6 secciones y consta de dos formularios a llenar, uno de levantamiento de datos y otro de verificación de horarios de atención y horarios de trabajo en días entre semana. Esta encuesta se realizó a la población cuencana.

Esta modelo se realizará para verificar la consistencia de datos de las encuestas tipo 2 de horarios de atención y labor, verificar que no tengan anormalidades entre secciones (evitar encuestas inventadas) y estimar un corte que cuantifique los resultados de esta.

2. Identificación de dimensiones y hechos
    - El modelo debe describir una relación M:N

Se identifican 5 dimensiones y una tabla de hechos. *Ver imagen 1*. Estas dimensiones son:

1. DimFormulario1
2. DimHorarioAtencionLV
3. DimHorarioLaboralLV
4. DimHorarios
5. DimTipoHorario
6. FactFormulario2

Existen 2 relaciones n a m, los cuales se visualizan en la *figura 2*. Esta relación contempla la posibilidad de que un horario de atención pueda ser uno de los 98 contemplados (de 05:00 a 13:00, de 07:00 a 10:00, etc.) y estos a su vez pueden pertenecer a 4 tipos de horarios (jornada única, jornada doble, medio tiempo, sin horario fijo). Así mismo sucede con la relación con la dimensión horario laboral. La solución para este caso fue la creación de una dimensión puente, y el código identificador se coloca en la tabla hecho. *[1]*

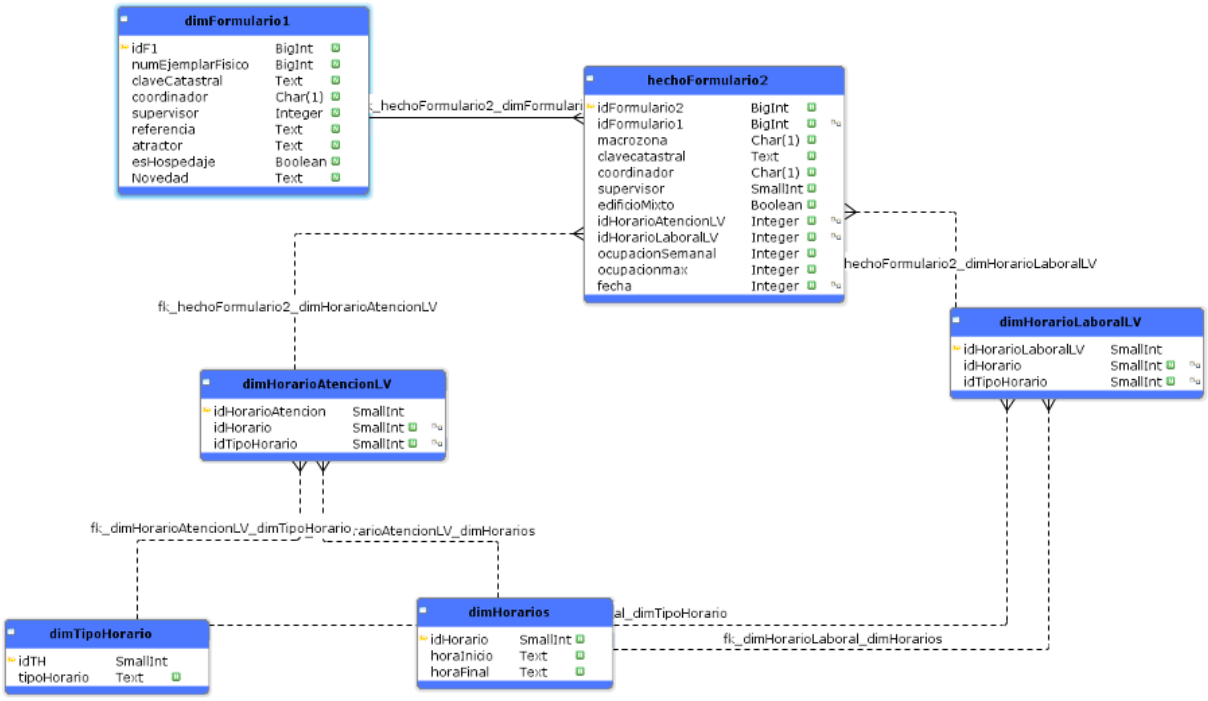

Imagen 1: Modelo Multidimensional General

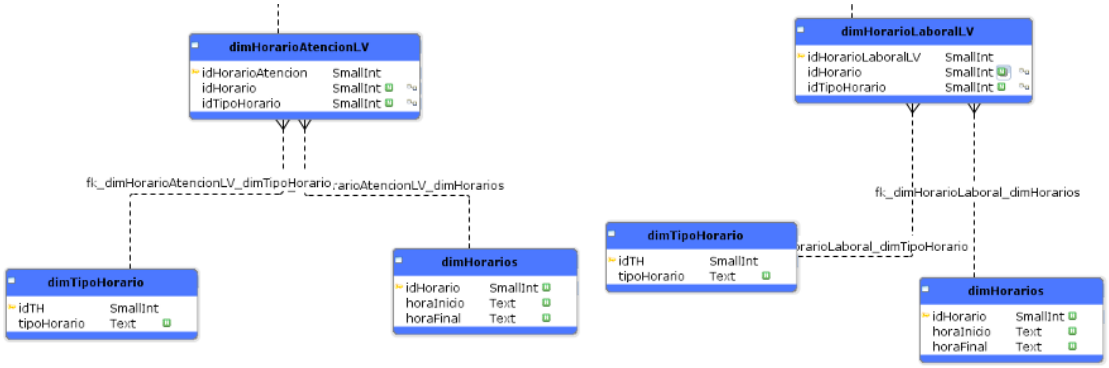

Imagen 2: Relacion N:M en las tablas de horario de atencion y horario laboral

3. Creación del modelo multidimensional en la base de datos

Para la creación del modelo multidimensional en la base de datos se realizó un proceso previo ETL en Pentaho Data Integration, esto con la finalidad de unir los archivos iniciales de los formularios. Entre los procesos que se realizaron en el ETL, fue la limpieza de los datos, el cambio de mayúscula – minúscula, la creación de las tablas dimensión, la filtración de datos incorrectos que restringirán los procesos de visualización en Pentaho Business Inteligencie. *Ver imagen 3*

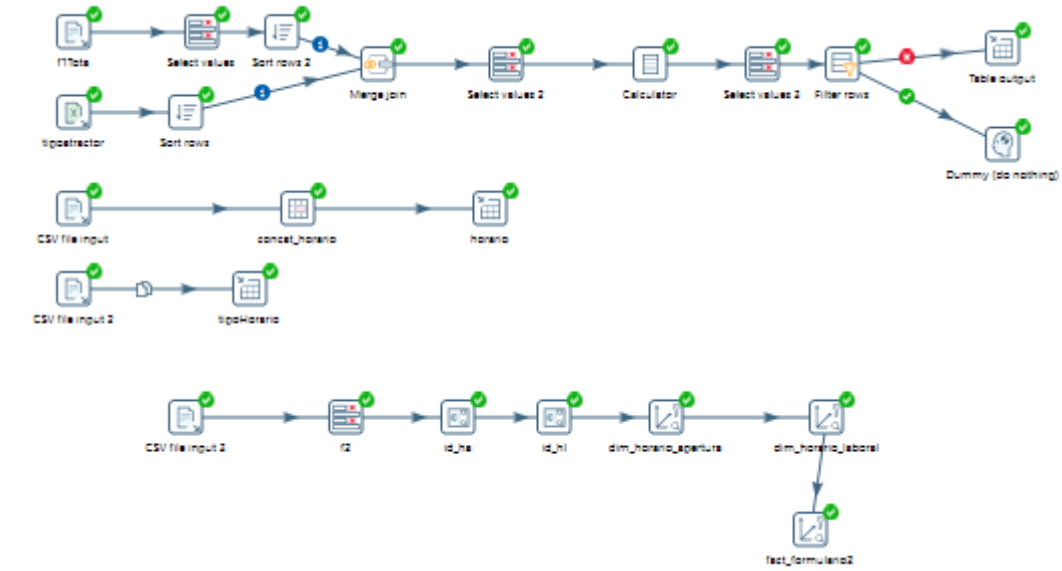

Imagen 3: Proceso ETL de creacion del Modelo Multidimensional

El resultado de estos se almacena en una base de datos en Postgres, para posteriormente llamarla desde Pentaho BI. *Ver imagen 4 y 5*

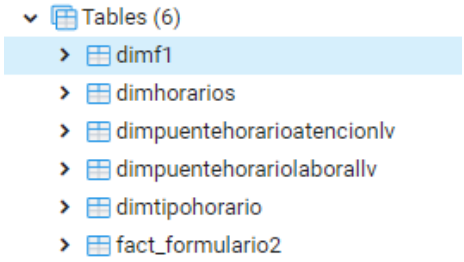

Imagen 4: Tablas del modelo multidimensiaonl en Postgres.

	id bigint	idf1 bigint	coordinador bigint	supervisor bigint	edificiomixto boolean	idhorarioatencion double precision	idhoariolaboral double precision	aforomaxsemana bigint	capacidadsemanamax bigint
1	1	1	1	1	false	1	1	30	70
2	2	68	1	1	false	2	2	100	120
3	3	2	1	1	false	3	3	500	1000

	id bigint	id_horario_atencion double precision	id_horario bigint	id_tipo bigint
1	1	1	91	1
2	2	2	17	2
3	3	3	18	2
4	4	4	18	4
5	5	5	17	1
6	6	6	45	4

Imagen 5: Ejemplo de las tablas de Hechos y la dimension puente de horarios atencion.

4. **Visualización del cubo en el BI-Server**

Para la visualización del cubo en BI Server se necesita crear un DataSource que establezca la conexión con el servidor de base de datos, y cree el cubo en BI. *Ver imagen 6 y 7.*

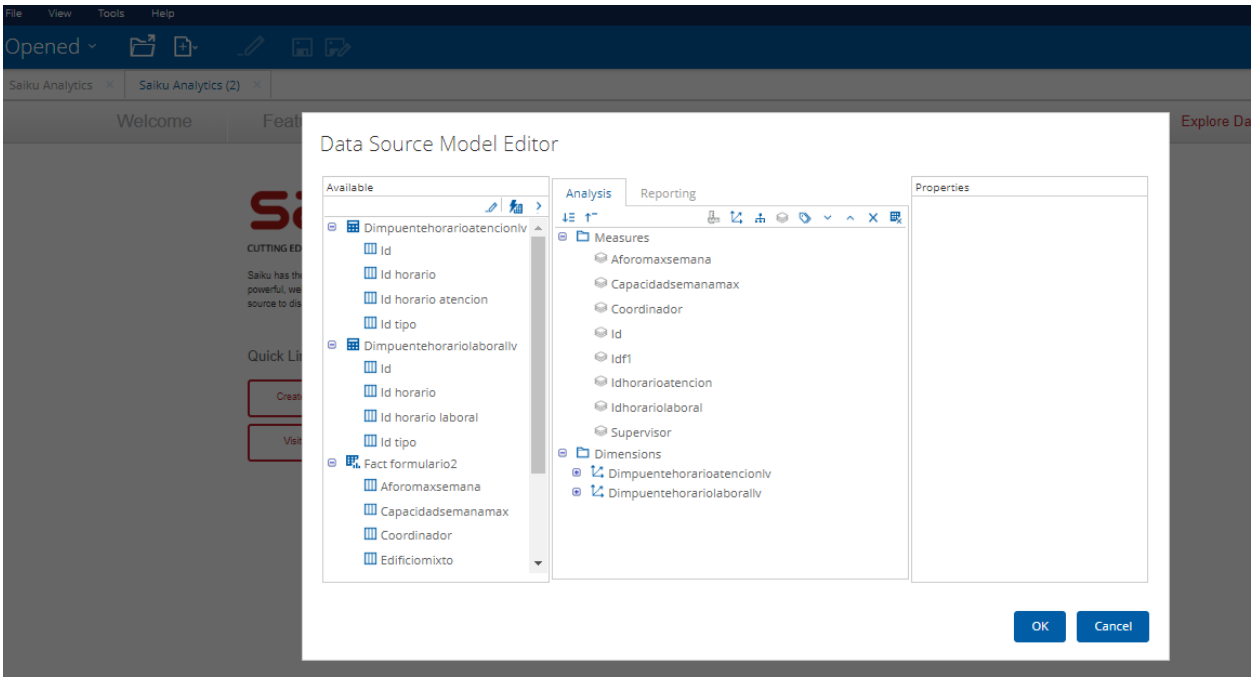

Imagen 6: Visualización del cubo en BI Server

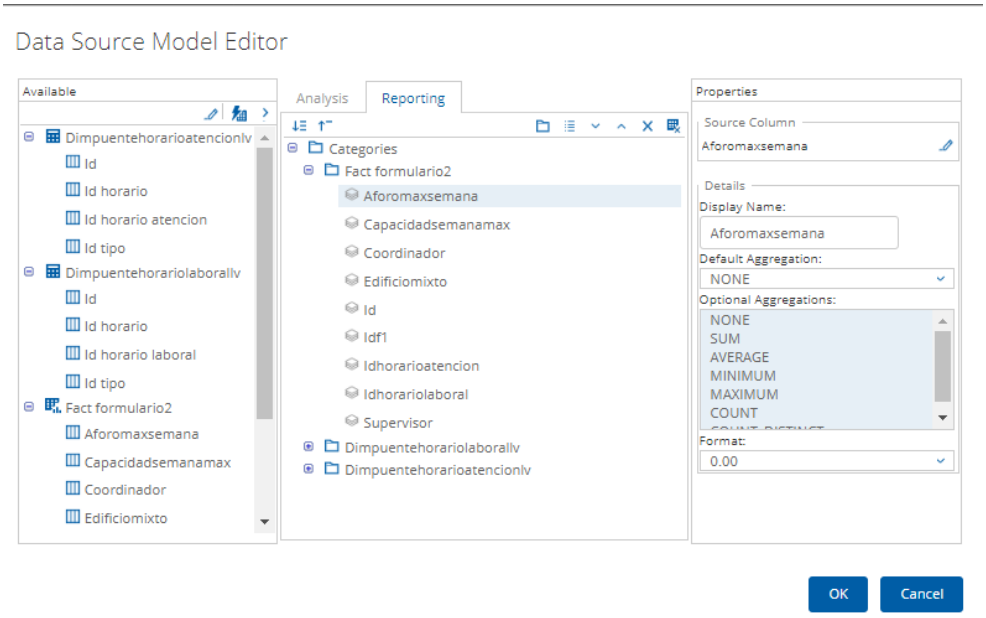

Imagen 7: Visualización del cubo y la reportería en BI Server

5. ¿Qué hacer con datos incompletos / valores null en las dimensiones y en los hechos? ¿Qué pasa en el saiku?

El problema de tener valores nulos en dimensiones y hechos es la inconsistencia de datos, el tener datos incompletos en la tabla de hecho equivale a decir que no apunta a ninguna fuente de datos o tabla, lo cual no permite realizar correctamente las operaciones multidimensionales o análisis multivariable, esto en un reporte tipo saiku o una visualización en JPivot. Con este fin, ante los problemas que surgen por tener datos inconscientes, coloque en el proceso de ETL,

filtrando los valores nulos y asignándoles un valor equivalente, por tal se crea en la tabla de DimHorarios, un código para colocar formularios que no tienen un horario definido o que tengan un horario incompleto, es decir una hora inicio, pero no una hora final, o viceversa. Esto en un reportería identifica que formularios se deben verificar por inconsistencia. Ver imagen 8.

	id [PK] bigint	horarioin_fin text
87	87	De 21:00 a 13:00
88	88	De 21:00 a 15:00
89	89	De 21:00 a 17:00
90	90	De 21:00 a 19:00
91	91	no tiene

Imagen 8: Caso tabla dimensión para registros nulos

	id [PK] bigint	idf1 bigint	coordinador bigint	supervisor bigint	edificiomixto boolean	idhorarioatencion double precision	idhorariolaboral double precision	aforomaxsemana bigint
1	91	89	1	1	true	91	91	15

Imagen 9: Caso tabla hecho para registros nulos

Conclusiones

El uso de los paquetes de Pentaho, acelera el análisis de datos a gran escala, esto debido a que su arquitectura permite manejar grandes cantidades de datos. Un caso que sucedió en esta práctica, fue que, con alrededor de 15.000 registros, Microsoft Excel limitaba su performance, sin embargo, Pentaho Data Integration no sufrió mayores inconvenientes.

El modelamiento de un cubo multidimensional, debe prever el manejo de valores nulos, estos valores deben ser referidos a claves que los identifiquen, no se debe dejarlos "sueltos" por los inconvenientes que surgen ya sea en el proceso ETL, o en su análisis BI.

Referencias Bibliográficas

Bernabéu R. Dario & García Mattío Mariano, (2015) "Relación Muchos a Muchos". Disponible en: http://troyanx.com/Hefesto/relacin_muchos_a_muchos.html