PRA2

Revisión diseños realizados en la PRA1

Diseño conceptual

En la PRA1 la solución que se propuso no correspondía al completo con el requerimiento real. El diseño consistió en describir dos tablas de hechos que contendrían el histórico de los indicadores y la *actividad aeroportuaria.*

Según se describe en la solución el diseño conceptual consistia en describir la interrelación entre las distintas dimensiones, métricas, y hechos. Se puede observar que se trataba de un diseño con forma de “estrella” donde todas los componentes están interrelacionados, siendo la dimensión geográfica el nexo de unión entre los hechos relativos a las rutas y a los indicadores socio-económicos.

Diseño lógico

El diseño conceptual propuesto incorrectamente correspondía con el diseño lógico. Se propusieron dos tablas de hechos;

* *h\_actividad\_aeroportuaria* equivalente a la tabla *hc\_rutas* propuesta en la solución
* h\_KPI\_socio económico equivalente a la tabla *hc\_indicadores de la solución*

Los atributos propuestos son similares recogiendo las dimensiones que definen las rutas y los indicadores históricos.

La mayor diferencia la representa que que en la respuesta que se entregó se definió una dimensión adicional, los indicadores que permiten flexibilizar la inclusión de nuevas métricas en caso de que las fuentes se amplíen, cambien o desaparezcan.

Diseño físico

Los diseños físicos coinciden en gran parte. La nomenclatura que se propuso en la respuesta no sigue el estándar que propone la solución propuesta, requiriendo por tanto un cambio de *d\_dim\_nombredimension…* a *dm\_nombredimension*

La mayor diferencia de nuevo viene en que en el diseño que se propuso existe una tabla adicional con los indicadores y su tipo. Este diseño permite crear, eliminar, o modificar los KPIs de una manera más sencilla y flexible.

Otras diferencias que encontramos son:

* en la solución se incluye una dimensión con el tipo de ruta a emplear
* en la tabla de rutas se incluyen todos los identificadores de los países, aeropuertos, etc.

Diseño procesos ETL

Identificación procesos ETL

En primer lugar vamos a definir qué tipos de procesos ETL tenemos en nuestro caso de uso:

* Carga de dimensiones “estáticas”: alguna de las dimensiones son relativamente estáticas, los países, los tipos de rutas o los continentes no son valores que vayan a cambiar. En el caso de algunas de estas dimensiones, como los países, debido a su cardinalidad quizá no podemos introducirlos a mano, aunque si corregir los errores que puedan existir. Las otras dos dimensiones son claras candidatas a ser cargadas manualmente.
* Carga de dimensiones “dinámicas”: otras dimensiones aunque en nuestro caso son estáticas porque se trata de una carga inicial, como las aerolíneas, las fechas o los equipamientos son dimensiones sujetas a frecuentes cambios (aerolineas que cesan su actividad, equipamientos nuevos que surjen y otros que se descatalogan, etc.).
* Carga de hechos a espacio temporal. Una vez cargadas las dimensiones podemos proceder a cargar los hechos.
* Consolidación de los datos. En la carga inicial vamos a considerar que los datos (tras una inspección de las dimensiones manual) son válidos. Sin embargo, tenemos que dotar al sistema de carga de validadores que ante nuevos datos validan la calidad y veracidad de las dimensiones o hechos. Una vez los datos han pasado estos filtros podremos escribir los datos en las tablas definitivas.

Describimos a continuación todas las acciones que vamos a realizar dentro de estos 4 grupos identificados. El orden con el que se exponen es el orden con el que deberán ejecutarse.

|  |  |  |
| --- | --- | --- |
| **PROCESO ETL** | **Descripcion** | **Subtareas** |
| Carga dimensiones estaticas “manuales" | Existen varias dimensiones que por su naturaleza o frecuencia de actualizacion pueden ser cargadas manualmente, es decir, fijandose los valores en el codigo o introduciendolos en una UI | - *Hardcodear* e insertar los valores de los **DM\_TIPO\_RUTA** |
| Carga resto dimensiones estaticas | No todas las dimensiones estaticas pueden ser *hardcodeadas* ya que serian muchos valores a insertar manualmente. Sin embargo, si las vamos a considerar dimensiones estaticas que no van a tener un proceso ETL de refresco frecuente. Por ello, este paso consistira en la carga masiva de esta dimension desde las fuentes de datos. Esta carga se realizara a una tabla temporal donde manualmente podremos revisar que los valores tienen sentido y corregir aquellos errores que encontremos. | - Extraer los valores de las fuentes de datos (csvs, jsons, etc.)  - Cargar a tablas temporales las siguientes dimensiones: **DM\_ANYO**, **DM\_PAIS**  - Revision de los datos a cargar, validado y correccion  - Carga a las tablas definitivas |
| Carga dimensiones dinamicas | Las dimensiones dinamicas tendran dos procesos ETL asociados, el de *carga inicial* y el *actualizacion*. En el primero se realizara un truncado de la tabla existente y se inserta de forma masiva todos los datos validados. En el segundo se realiza unicamente un *append* de los datos, respetando aquellas claves que ya existan. | - Extraer los valores de las fuentes de datos (csvs, jsons, etc.)  - Validado de los datos: integridad referencial, elimindo de duplicados, etc.  - Insertado de los datos empleando el metodo que corresponda por el tipo de carga  - Las dimensiones que se cargaran en este paso son: **DM\_AEROLINEAS, DM\_AEROPUERTOS, DM\_EQUIPAMIENTOS** |
| Carga de hechos | Una vez las dimensiones estan cargadas se puede proceder a cargar la informacion en las tablas de hechos. Para ello validaremos la integridad referencial de las dimensiones, eliminaremos los duplicados, extraeremos algunas metricas e insertaremos los hechos que sean correctos. Al igual que en el caso de las dimensiones dinamicas, este proceso tiene dos sub-ETL asociadas dependiendo de si es una carga inicial o una actualizacion. | - Extraccion de los datos desde las fuentes  - Validado de la informacion  - Insertado en las tablas finales: **HC\_RUTAS, HC\_INDICADORES** |
| Carga tablas de relacion | El ultimo paso es cargar las tablas de relacion entre los hechos y las dimensiones. | - Extraccion de los datos  - Validado de la integridad referencial de los datos  - Insertado en las tablas finales: **CO\_RUTA\_EQUIPAMIENTOS** |