Práctica Desarrollo:
Modelo de datos multidimensional

David Pacios Vázquez
Pedro López Chaves

ÍNDICE

Dataset: Partidos de fútbol 2018/2019

Contexto de la actividad

Somos un grupo inversionista Arabia Saudí que basamos nuestra fuente de ingresos en la venta de petróleo. Con todos nuestros ingresos queremos ejercer la compra de un equipo de fútbol de primer nivel. Los equipos que más nos interesan son los que disputan las competiciones nacionales de España e Inglaterra, es decir, los equipos de La Liga y La Premier League.
Para ellos queremos crear una base de datos multidimensional que nos permita realizar consultas que nos ayuden a que nuestra decisión de compra sea lo más acertada posible.

Objetivos generales perseguidos

Para analizar los equipos, queremos generar una base de datos, que permita comparar el rendimiento de los equipos en sus partidos. Haciendo foco en su rendimiento deportivo, en los goles que generan, los que conceden y los resultados obtenidos.

Por otro lado, nos interesa conocer como rinde el equipo ante su afición, ya que queremos mantener la insignia y fuerza que le brindan los aficionados al club. Y que deriva directamente en un mayor consumo de los aficionados en los servicios del club. También nos interesa conocer el club, su ciudad y su presupuesto, ya que influirán claramente en el coste de sus adquisiciones.

Por último, la infraestructura del club es importante, ya que un estadio con mucha capacidad puede generar ingresos extra por eventos y un mayor número entradas disponibles para vender.

Fuentes de datos para contextualizar el problema

Con el fin de analizar todos los aspectos anteriores, se ha recurrido a las siguientes fuentes de datos:

- **https://datahub.io/sports-data/spanish-la-liga**
- **https://datahub.io/sports-data/english-premier-league**

Estas fuentes proporcionan información de los partidos de LaLiga y La Premier League de la temporada 2018/2019.

Además, hemos generado una fuente de datos a mayores con información de los equipos, como su entrenador, su estadio y la capacidad del estadio para albergar aficionados entre otros.

Los datos obtenidos de la fuente datahub se han limpiado eliminando todos los datos de estadísticas que no nos interesaron manteniendo únicamente: liga, fecha, hometeam, awayteam, goles_local, goles_visitante.

El campo fecha lo dividimos en dia, mes y año; anotando sus valores en formato numérico para facilitar luego el trabajo con los mismos.

Añadimos un campo calculado Resultado, en el que según los goles anotados por el local y visitante marcaremos el resultado. La 1 marca que ganó el local, el 2 que ganó el visitante y la X que hubo empate.

Listado de consultas analíticas de ejemplo que podrían realizarse

Creamos ahora un listado de consultas analíticas de ejemplo que podrían utilizarse y nos van a ayudar a modelar la base de datos. Dejamos aquí unos ejemplos:

1. ¿Qué equipo gano más partidos en cada liga?
2. ¿Qué equipo perdió más partidos en cada liga?
3. ¿Qué equipo marcó más goles en cada liga?
4. ¿Qué equipo es el más fuerte en casa? ¿Qué equipo sacó más puntos en casa?
5. ¿Qué equipo es el más fuerte fuera de casa? ¿Qué equipo sacó más puntos fuera de casa?
6. ¿Cuál fue la mayor goleada (diferencia de goles) en un partido en cada liga?
7. ¿Cuántos puntos obtuvo el equipo como local con mayor afluencia de público?
8. ¿Cuántos puntos obtuvieron los 5 equipos con mayor presupuesto?
9. ¿Quién fue el mejor equipo en el mes de diciembre en cada liga?
10. ¿En qué liga se marcaron más goles?

Desarrollo del modelo de datos multidimensional

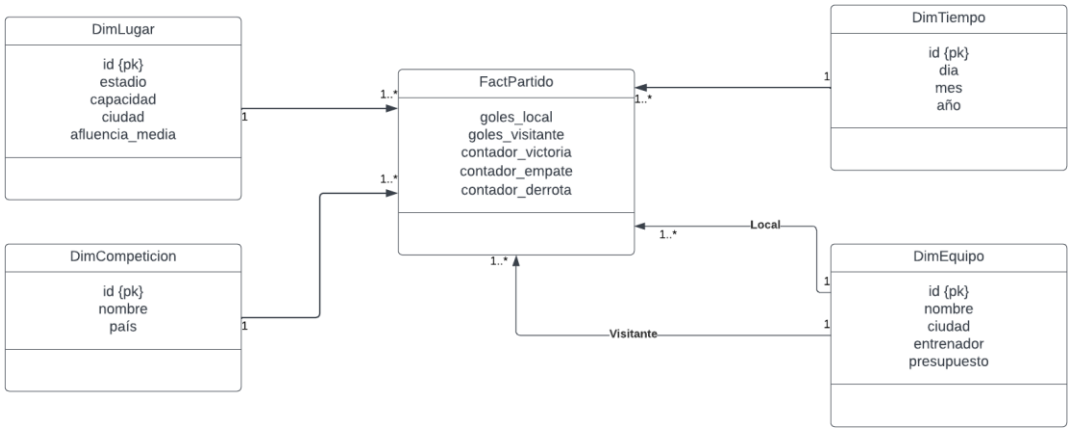
Hemos seleccionado una granuralidad de partido con los datos de los que disponemos y para responder las cuestiones planteadas. En la tabla de hechos tendremos los datos de goles_local, goles_visitante, contador_victoria, contador_empate, contador_derrota.

Esta serie de contadores los desarrollamos durante el etl y nos servirán para facilitar las consultas en las que queramos obtener el número de victorias, derrotas y empates, reduciendolo a un simple sumatorio de estos campos. Por poner un ejemplo, en el caso de que se diera un empate, los contadores tendrían los siguientes valores:

- Contador_victoria: 0.
- Contador_empate: 1.
- Contador_derrota: 0.

Es por esto que con un simple sumatorio ya obtendríamos el número de victorias, empates y derrotas.

Creamos además 4 dimensiones que son las de lugar, competición, tiempo y equipo. Esta última con la particularidad de que tiene una relación doble con la tabla de hechos, ya que una relación será para corresponder con el equipo local y la otra con el visitante.

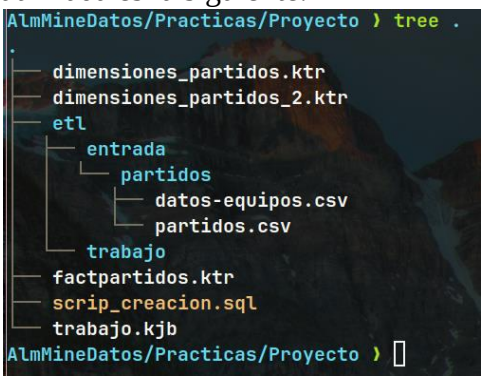


Creación de la base de datos sobre el modelo de datos multidimensional

La base de datos sigue el diseño que realizamos en el apartado anterior y lo entregamos en formato pg_dump junto con los demás archivos de la práctica. Las claves primarias de cada dimensión se generan automáticamente de forma secuencial en la base de datos, y la clave primaria de cada elemento de la tabla de hechos se obtiene a partir de la combinación de todas las claves primarias de sus dimensiones.

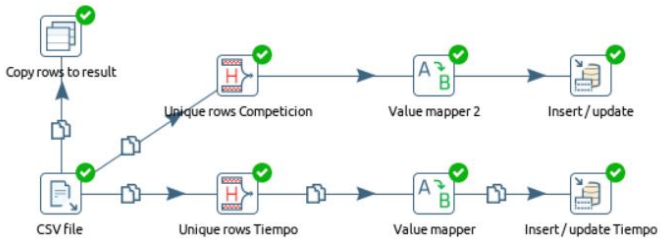
Creación del ETL sobre el modelo de datos multidimensional

La estructura de ficheros utilizada es la siguiente:



De este modo todos los .csv utilizados para las transformaciones, están en una ruta relativa dentro de la carpeta del Proyecto, en concreto en etl > entrada > partidos

Transformación 1: dimensiones_partidos

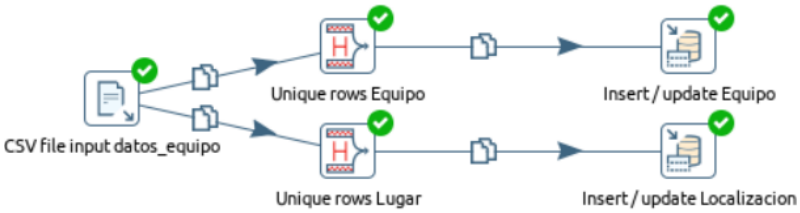


En esta primera transformación cogemos los datos de partidos.csv. Primeramente, extraemos las fechas para la dimensión tiempo. Creamos también, a través del value mapper el campo texto_mes donde almacenamos el nombre del propio mes, por ejemplo, 1 lo cambia por enero. Y añadimos finalmente dia, mes, ano y texto_mes a la base de datos.

Por otro lado, cogemos la información de que competición se trata (liga española o premier), añadiendo otro value mapper para que le añada el país a la competición, si es la liga le añade España y si es Premier le añade Inglaterra.

Además, con el operador copy rows to result, se enviarán los datos necesarios para poder completar la tabla de hechos en la transformación factpartidos.

Transformación 2: dimensiones_partidos_2

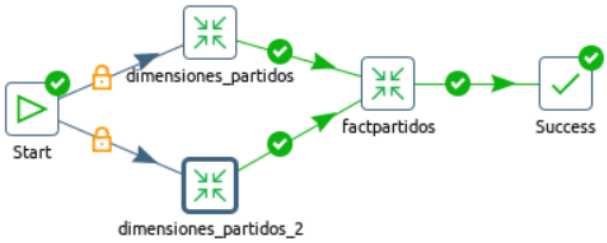


En esta transformación obtenemos los datos del otro dataset datos-equipos.csv. Aquí lo que hacemos es coger los campos que se refieren a equipo (nombre, ciudad, presupuesto...) y los que se refieren a localización (ciudad, estadio, afluencia media...) para introducirlos en sus dimensiones correspondientes.

Transformación 3: factpartidos

En esta transformación obtenemos los datos de dimensiones_partidos en el get rows from result. Los database lookup cogen los ids de la base de datos de cada dimensión para introducirlos en la tabla de hechos. Además, en el último operando añadimos expresiones de java para, a partir del campo resultados, obtener si es victoria del local, si es un empate o si es una victoria del visitante. Añadimos estos 3 campos y serán 0 si no se dio ese resultado y 1 si sí que se dio.

Trabajo



Este job es donde ejecutamos todas las transformaciones juntas. Las de dimensiones partidos se ejecutan paralelamente, dando lugar a una ejecución más rápida, ya que no existe dependencia entre estas, y posteriormente factpartidos rellenando así toda nuestra base de datos.