Datawarehouse PRACTICA 1

Mineria de datos (Universitat Oberta de Catalunya)

M2.863 - Diseño y construcción del data warehouse

PRA1: "Diseño de un almacén de datos para la gestión de información de aeropuertos, tráfico aéreo e indicadores turísticos"

Jesus Martínez Casadiego

1. Diseño del Datawarehouse

En la practica se detalla los posibles usaurios de negocio de esta base de datos (stakeholders):

a) **Sector turismo:** Agentes relacionados con el sector turístico: Ministerios de Turismo Organización Mundial de Turismo Agencias de viajes Cadenas hoteleras Secretarías de comercio internacional de los Ministerios de Economía y Competitividad

b) **Aerolíneas**: las empresas aerolíneas pueden utilizar los datos como apoyo para detectar nuevas posibles rutas, mejora la gestión de la demanda de pasajeros, la ocupación de los vuelos, la estacionalidad, el mercado potencial, identificar destinos turísticos con baja cobertura, optimizar número y tipo de aeronaves, necesidades de personal, etc.

c) **Industria aerocomercial**: Diferentes agentes relacionados con esta industria como pueden ser:

- Tiendas retail aeropuertos
- Empresas gestoras de espacios de parking
- Servicios de transporte relacionados con desplazamiento a aeropuertos
- Empresas que comercialicen soluciones tecnológicas para aeropuertos o aerolíneas
- Empresas que distribuyan combustible para aeronaves
- Servicios auxiliares aeropuertos (seguridad, limpieza, etc.)
- Estudios de investigación universitarios

d) **Administración pública**: Diferentes ministerios y agencias gubernamentales como pueden ser:

- Ministerio de Transporte (del país o países vinculantes a los datos)
- Ministerio de Turismo
- Ministerio de Energía
- Agencia de control aéreo
- Agencia de control de aeropuertos a nivel nacional e internacional

e) **Ciudadanos**: finalmente los ciudadanos pueden estar interesados en tener una mayor trasparencia respecto aerolíneas, aeropuertos, aviones y su servicio.

3.1. Análisis de requerimientos

En la práctica se detalla los posibles usuarios de negocio de esta base de datos (stakeholders),

En este caso, el objetivo es diseñar un almacén de datos que permita analizar las rutas de los diferentes países con los indicadores económicos de casa país. Esto se traduce en que, dentro de las necesidades de información por parte de los diferentes actores, podemos identificar las siguientes.

1) **Conocer la relación del número de rutas de un país con sus indicadores económicos**
2) Conocer las rutas de las diferentes compañías
3) Relacionar los tipos de aviones en cada vuelo
4) Disponer de datos relativos a vuelos relativos a las rutas

3.2. Análisis de fuentes de datos

En el ejercicio práctico dado que no se tiene acceso a las partes interesada del proyectos, se utilizará el sentido común.

Las fuentes se detallan a continuación

Fichero	Descripción	Fuente
airpots.dat	Listado de aeropuertos del mundo con conjunto de características (código IATA1, ICAO2, ciudad, país, coordenadas, zona horaria, etc.)	openflights
airlines.dat	Listado de aerolíneas del mundo con conjunto de características (identificador, código IATA, ICAO, país, atributos actividad, etc.)	openflights
routes.dat	Listado de rutas aéreas del mundo con conjunto de características (aerolínea, aeropuerto origen y destino, paradas, equipamiento,	openflights
Equipamientos.js	Aviones de rutas aéreas con código IATA y descripción	aircraft type codes y IATA aircrafts codes
Paises.xml	Listado de países con códigos estándar. Códigos ISO (International Organization for Standardisation) de 2 y 3 caracteres y coordenadas país	Country codes , listado country codes wikipedia y Public data Google Countries
worldbank_pasajeros_anual_pais.csv	Serie histórica de número de pasajeros en vuelos nacionales e internacionales por país y año	Datos pasajeros banco mundial
worldbank_poblacion_anual_pais.csv	Serie histórica de poblaciones a nivel de país y año	Datos poblaciones banco mundial
worldbank_ing_turimo_anual_pais.csv	Serie histórica de ingresos por turismo por país y año (importe en dólares USA)	Datos turismo banco mundial
worldbank_comercio_ext_anual_pais.csv	Serie histórica de datos comercio exterior (suma de exportaciones e importaciones de bienes y servicios representadas como porcentaje del PIB) por país y año	Datos comercio exterior banco mundial

En función de como esta estructurada la información podemos agrupar en los siguientes fuentes:

Tipo de información/ Conector	Formato	Ficheros incluidos
Datos opendata	Ficheros en formato .dat	airpots.dat airlines.dat routes.dat
Datos Paises	Fichero en formato xml	Paises.xml

Datos Equipamientos	Fichero en formato js (JSONI	Equipamientos.js
Datos del banco de datos	Ficheros en formato CSV	worldbank_pasajeros_anual_pais.csv worldbank_poblacion_anual_pais.csv worldbank_ing_turimo_anual_pais.csv worldbank_comercio_ext_anual_pais.csv

A continuación para cada uno de los ficheros analizamos cada uno de los campos y

A. Airpot.dat

Atributo	Descripcion	Tipo de Dato	Comentarios
Airport ID	Unique OpenFlights identifier for this airport.	Numérico	(Obligatoria) Identificador que utilizaremos para identificar el aeropuesto
Name	Name of airport. May or may not contain the **City** name.	Texto	(Obligatoria) Nombre del aeropueto
City	Main city served by airport. May be spelled differently from **Name**.	Texto	(Opcional) Nombre de la ciudad, todo el analisis de indicadores del banco de datos va por pais, por tanto es um valor opcional
Country	Country or territory where airport is located. See countries.dat to cross-reference to ISO 3166-1 codes.	Texto	(Obligatorio) Código del pais
IATA	3-letter IATA code. Null if not assigned/unknown.	Texto (Lontigud 3)	(Opcional) Código IATA (no lo cargaremos ya que tomaremos como maestro Airport ID)
ICAO	4-letter ICAO code. Null if not assigned.	Texto (Longitud 4)	(Opcional) Código IATA (no lo cargaremos ya que tomaremos como maestro Airport ID)

Latitude	Decimal degrees, usually to six significant digits. Negative is South, positive is North.	Número Decimal	(Opcional) No lo cargaremos ya que la información económica se relaciona por country
Longitude	Decimal degrees, usually to six significant digits. Negative is West, positive is East.	Número Decimal	(Opcional) No lo cargaremos ya que la información económica se relaciona por country
Altitude	In feet.	Float	(Opcional) No lo cargaremos ya que la información económica se relaciona por country
Timezone	Hours offset from UTC. Fractional hours are expressed as decimals, eg. India is 5.5.	Float	(Opcional) No lo cargaremos ya que la información económica se relaciona por country
DST	Daylight savings time. One of E (Europe), A (US/Canada), S (South America), O (Australia), Z (New Zealand), N (None) or U (Unknown). *See also: Help: Time*	Texto	(Opcional) No lo cargaremos ya que la información económica se relaciona por country
Tz database time zone	Timezone in "tz" (Olson) format, eg. "America/Los_Angeles".	Texto	No necessário ya que se utilizará el campo Timezone
Type	Type of the airport. Value "airport" for air terminals, "station" for train stations, "port" for ferry terminals and "unknown" if not known.*In airports.csv, only type=airport is included.*	Texto	No necessário ya que siempre tendra el mismo valor

B. Airlines.dat

Atributo	Descripcion	Tipo de Dato	Nivel de Información
Airline ID	Unique OpenFlights identifier for this airline.	Numerico	(Obligatorio)
Name	Name of the airline.	Texto	(Opcional)

			Lo cargaremos para no tener que conocer los código de la aerolinea
Alias	Alias of the airline. For example, All Nippon Airways is commonly known as "ANA".	Texto	(Opcional) Lo cargaremos por si alguien necesita hacer un análisis por ese atributo
IATA	2-letter IATA code, if available.		(Fuera Carga)
ICAO	3-letter ICAO code, if available.		(Fuera Carga)
Callsign	Airline callsign.		(Fuera Carga)
Country	Country or territory where airline is incorporated.		(Obligatorio) Para poder hacer en alaisis
Active	"Y" if the airline is or has until recently been operational, "N" if it is defunct. This field is not reliable: in particular, major airlines that stopped flying long ago, but have not had their IATA code reassigned (eg. Ansett/AN), will incorrectly show as "Y".		(Obligatorio) Para conocer si esta activa (en un buen análisis mejor sería si tuviéramos la información desde cuando estuvo activa y fecha fin)

C. Routes.dat

Atributo	Descripcion	Tipo de Dato	Comentarios
Airline	2-letter (IATA) or 3-letter (ICAO) code of the airline.	Numérico	(Obligatorio)
Airline ID	Unique OpenFlights identifier for airline (see Airline).	Numérico	(Obligatorio)
Source airport	3-letter (IATA) or 4-letter (ICAO) code of the source airport.	Texto	(Fuera de carga) Para evitar duplicidad solo cargaremos el ID
Source airport ID	Unique OpenFlights identifier for source airport (see Airport)	Numérico	(Obligatorio)
Destination airport	3-letter (IATA) or 4-letter (ICAO) code of the destination airport.	Texto	(Fuera de carga) Para evitar duplicidad solo cargaremos el ID

Destination airport ID	Unique OpenFlights identifier for destination airport (see Airport)	Numérico	(Obligatorio)
Codeshare	"Y" if this flight is a codeshare (that is, not operated by Airline, but another carrier), empty otherwise.	Texto	(Opcional) En una análisis inicial no necesario para calcular ningún indicador pero se deja para necesidades futuras
Stops	Number of stops on this flight ("0" for direct)	Numerico	(Opcional) En una análisis inicial no necesario para calcular ningún indicador pero se deja para necesidades futuras
Equipment	3-letter codes for plane type(s) generally used on this flight, separated by spaces	Texto	(Obligatorio)

D. Equipamientos.js

Atributo	Descripcion	Tipo de Dato	Nivel de Información
Equipament	IATA Code	Numerico	
Manufaturer	Fabricante del avión	Texto	
Type/Model	Modelo de Avion		
Wake	Wake category (L-Light, M-Medium, H-Heavy)		

E. Paises.xml

Atributo	Descripcion	Tipo de Dato	Nivel de Información
Cod_pais	Código de Pais	Numerico	
cod_pais2	Codigo de PAis (3)	Texto	
desc_pais	Descripción Pais	Texto	
cod_continente	Código Continente	Texto(2)	
desc_continente	Descripción Continente	Texto	

Latitude	Latitud Pais	Decimal	No lo cargamos ya que con la información geospacial del aeropuesto es el dato que queremos explotar
longitude	Longitud Pais	Decimal	No lo cargamos ya que con la información geospacial del aeropuesto es el dato que queremos explotar

F. worldbank_pasajeros_anual_pais.csv

Atributo	Descripcion	Tipo de Dato	Comentarios
Country	Código de Pais	Numerico	No lo cargaremos ya que en nuestra base datos utilizamos el identificador de dos
Indicator Name	Nombre del indicador	Texto	Air transport passengers carried
Indicator Code	Codigo de PAis (3)	Texto	IS.AIR.PSGR
Tantas columnas como años	56 Columnas (una por año entre 1960 y 2016) con los valores	Decimal	

G. worldbank_poblacion_anual_pais.csv

Atributo	Descripcion	Tipo de Dato	Nivel de Información
Country	Código de Pais	Numerico	No lo cargaremos ya que en nuestra base datos utilizamos el identificador de dos
Indicator Name	Nombre del indicador	Texto	Population, total
Indicator Code	Codigo de PAis (3)	Texto	SP.POP.TOTL
Tantas columnas como años	56 Columnas (una por año entre 1960 y 2016) con los valores	Decimal	

H. worldbank_ing_turimo_anual_pais.csv

Atributo	Descripcion	Tipo de Dato	Nivel de Información
Country	Código de Pais	Numerico	No lo cargaremos ya que en nuestra base datos utilizamos el identificador de dos
Indicator Name	Nombre del indicador	Texto	International tourism, receipts (current US$)
Indicator Code	Codigo de PAis (3)	Texto	ST.INT.RCPT.CD
Tantas columnas como años	56 Columnas (una por año entre 1960 y 2016) con los valores	Decimal	

I. worldbank_comercio_ext_anual_pais.csv

Atributo	Descripcion	Tipo de Dato	Nivel de Información
Country	Código de Pais	Numerico	No lo cargaremos ya que en nuestra base datos utilizamos el identificador de dos
Indicator Name	Nombre del indicador	Texto	Trade (% of GDP)
Indicator Code	Codigo de Pais (3)	Texto	NE.TRD.GNFS.ZS
Tantas columnas como años	56 Columnas (una por año entre 1960 y 2016) con los valores	Decimal	

Estas tablas las tendremos que transformar en una tabla única donde cada valor de cada tabla lo transformamos en un columna y el año lo pasamos a un atributo:

Atributo	Descripcion	Tipo de Dato	Comentarios
Country	Código de Pais	Numerico	(Obligatorio)
Year	Identificador de año	Fecha	(Opcional) Solo cargaremos los datos del ultimo año ya que

			no disponemos información hisotórica de los vuelos
IS.AIR.PSGR	Air transport passengers carried	Decimal	(Obligatorio)
SP.POP.TOTL	Population, total	Decimal	(Obligatorio)
ST.INT.RCPT.CD	International tourism, receipts (current US$)	Decimal	(Obligatorio)
NE.TRD.GNFS.ZS	Trade (% of GDP)	Decimal	(Obligatorio)

3.3. Análisis funcional

A continuación se describe los requerimentos funcionales

ID	Descripcion	Tipo Requisito	Prioridad	Exigible/Deseable
1	Se extrarera la información del openflights	Funcional	1	E
2	Se extraean los datos maestros de countries y de equipments	Funcional	1	E
3	Se extraerá los datos del banco de datos	Funcional	1	E
4	Se creara un almacen de datos con la información de vuelos y datos económicos de los países. Así como las bases datos auxiliares necesarias para su interpretacion	Funcional	1	E
5	Se cargara la información en el staging área	Funcional	1	E
6	Se realizaran las transformación des datos para cargar el Datamart	Funcional	2	E
7	Se creará un modelo OLAP para consultas automáticas de los usuarios.	Funcional	2	E

NOTA:

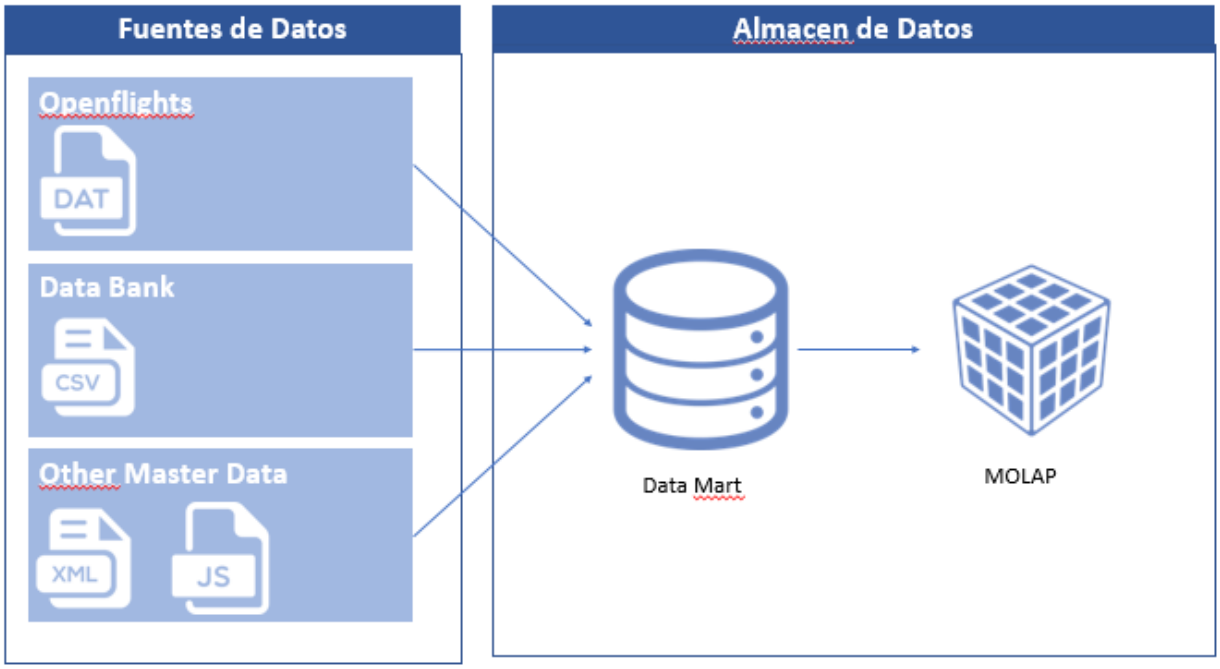
*1 . Los valores de prioridade serán de 1 a 4, siendo 1 completamente prioritario para la actividad y 4, no prioritario para la actividad.
*2 Tipo de requisito (F-Funcional, A-Arquitectura, S-Seguridad

En términos de la arquitectura funcional, tenemos los elementos siguientes.

- Las fuentes de datos están compuestas por un único fichero Excel.
- La arquitectura de la factoría de información puede estar formada por varios elementos que estarán alojados en la misma máquina.
  – *Staging area* (opcional): al tener múltiples ficheros de fuentes diferentes (openfligths, banco de datos mundial, otros) , sería conveniente consolidarlos en una estructura de carga intermedia. En nuestro caso particular, sería creada en la misma base de datos.

  – *Datamart* vuelos: al centrarnos en una única área, es más correcto considerar que se está creando un *datamart* en lugar de un almacén de datos corporativo.
  – MOLAP: a partir de la información en el *datamart* de vuelos, se creará cubo multidimensional.

El siguiente gráfico resume los elementos de la arquitectura para esta actividad.



3.4. Diseño del modelo conceptual, lógico y físico

Diseño Conceptual

A partir del análisis de requisitos se han identificado una tabla de hechos

Tabla de hecho	Descripción
rutas	Recoge las rutas existentes

Las dimensiones sobre las queremos analizar los datos recogidos

Dimensiones	Descripción
Pais Origen	Recoge los países origen
Pais Destino	Recoge los países destino (apartir del aeropuerto)
Aerolinea	La áerolinea que opera la ruta
Tipo Avión	El tipo de avión que opera la ruta

Diseño Lógico

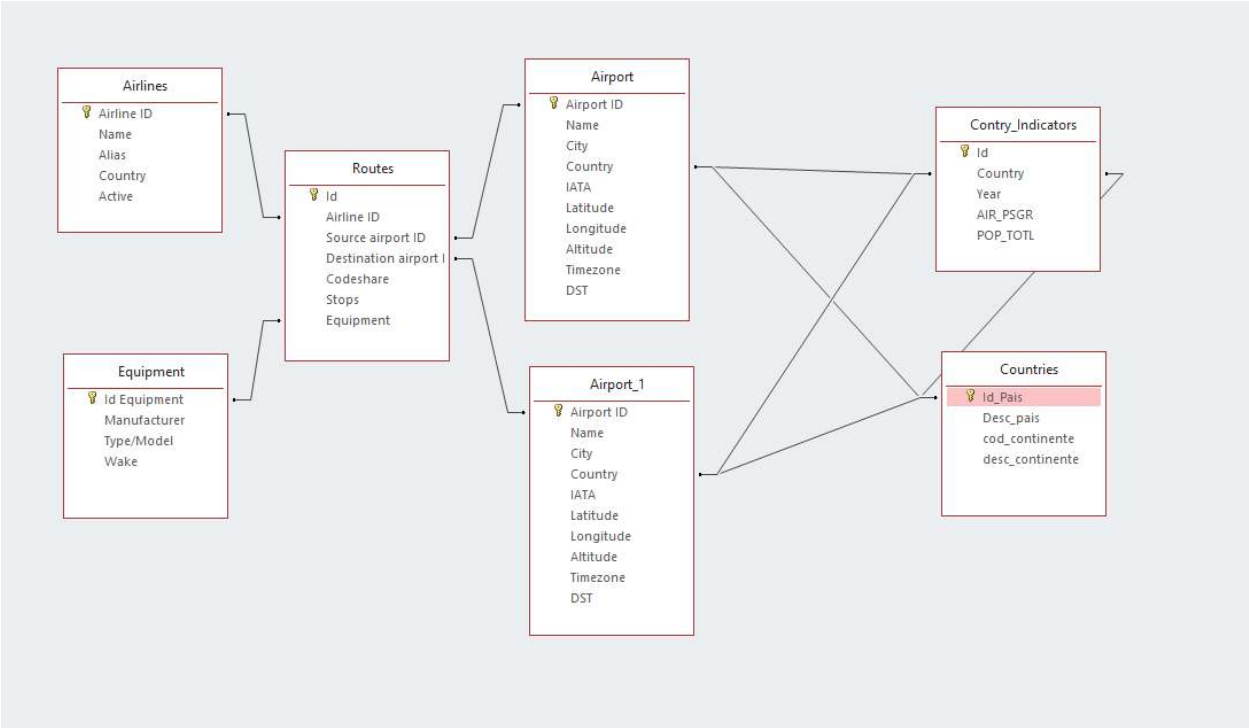
En este apartado detallamos los atributos de las dimensiones y las métricas de las tablas de hechos. A partir del análisis de requisitos se han identificado una tabla de hechos

Tabla de hecho	Métricas
rutas	valor

Y las dimensiones tienen los siguientes atributos:

Dimensiones	Descripción
Pais Origen	• Continente • Ciudad • Air transport passengers carried • Population, total • International tourism, receipts (current US$) • Trade (% of GDP)
Pais Destino	• Continente • Ciudad • Air transport passengers carried • Population, total • International tourism, receipts (current US$) • Trade (% of GDP)
Aerolinea	• Pais • Activa
Tipo Avión	• Fabricante • Wake

Diseño Fisico



Optimización de la factoría de información

Este ejemplo es pequeño por lo que parte de lo que vamos a comentar en esta sección es opcional. La optimización de la factoría de información es un proceso continuo, es necesario tener en cuenta:

- La previsión de crecimiento de datos
- La previsión del tipo de consultas que se realizarán

Con este tipo de información, se implementarían:

- Propuesta de tablespace para Oracle.
- Índices extras para mejorar las consultas.
- Vistas materializadas.

Una vez que el data warehouse está diseñado e implementado en la base de datos, el siguiente paso es la creación de los procesos de carga de datos.