Diseño y uso de bases de datos analíticas

Material teórico-práctico

Índice

Introducción 2

1. Caso de uso. Actividad volcánica mundial 4

1.1. Contexto 4

2. Reto 1. Las bases de datos analíticas: de los datos al conocimiento 5

2.1. *Data warehouse*: concepto y elementos 5

3. Reto 2. La arquitectura de la FIC, importa y mucho 9

4. Reto 3. Dimensión vs. hecho: ¿cuál es la diferencia? 14

4.1. Parte I. Diseño conceptual 16

4.1.1. Tablas de dimensión frente a tablas de hecho 16

4.2. Parte II. Diseño lógico 18

4.3. Parte III. Diseño físico 20

5. Reto 4. Los ETL son procesos clave en los sistemas analíticos 23

6. Reto 5. ¡Y llegó la magia! Cubos multidimensionales 30

Resumen 38

Introducción

Este material, como su nombre indica, pretende acercar la teoría a la práctica del curso «Diseño y uso de bases de datos analíticas». Ha sido creado para ayudar al estudiantado a mejorar la comprensión de los contenidos teóricos expuestos en el material didáctico de la asignatura y facilitar, así, su puesta en práctica al resolver las actividades evaluables del curso.

En este documento se presenta un caso de uso completo dividido en cinco retos, cuyo contenido se detalla a continuación:

- En el primer reto, se plantean las necesidades informacionales que tienen las organizaciones a la hora de tomar decisiones estratégicas. Introducimos los conceptos teóricos elementales relacionados con las bases de datos analíticas para entender que dicho análisis de datos permite la generación de conocimiento empresarial.

- En el siguiente reto, se atienden las necesidades de tipo informacional, mostrando los problemas con los que se enfrentan las personas responsables de los sistemas analíticos y la importancia de elegir un buen enfoque de arquitectura según las diferentes necesidades y perspectivas que se deben cubrir.

- El tercer reto se centra en mostrar la importancia de diferenciar los conceptos de hechos y dimensiones en las tres fases de diseño de una base de datos para definir el modelo multidimensional, y así responder a las necesidades analíticas.

    - **Parte I. Diseño conceptual.** Corresponde al mayor nivel de abstracción del modelo donde se definen los hechos y las dimensiones según el contexto, los requerimientos de los usuarios y el análisis del caso de uso. Se realiza un paralelismo entre los conceptos teóricos del diseño conceptual multidimensional estudiado en el material didáctico y la definición del diseño conceptual del caso práctico.
    - **Parte II. Diseño lógico.** En la segunda parte de este reto realizamos el diseño lógico como una etapa fundamental para avanzar en el diseño multidimensional, y relacionamos conceptos teóricos con su implementación práctica en nuestro caso de uso.
    - **Parte III. Diseño físico.** Tercer y último nivel del diseño multidimensional, el cual nos permite obtener las tablas de dimensiones y de hechos del modelo físico del almacén de datos.

- Diseñado el modelo físico multidimensional, llegamos al cuarto reto planteado, que consiste en poblar de datos el almacén siguiendo el modelo definido. Vemos que para ello es clave el diseño e implementación de los procesos de integración y transformación.

- Y finalmente, con el quinto reto se consigue la magia, implementando el cubo multidimensional y realizando análisis mediante operaciones de tratamiento de datos para proporcionar valor a las organizaciones.

1. Caso de uso. Actividad volcánica mundial

Como se ha comentado en la introducción, en este material se proponen una serie de retos cuyo hilo conductor será un caso de estudio real basado en la actividad volcánica mundial.

1.1. Contexto

En septiembre de 2021, una erupción del volcán de La Palma consternó a los vecinos de esta isla de las Canarias de 704 km2 y más de 84.000 habitantes, en Montaña Rajada, en la zona forestal de Cabeza de Vaca, en El Paso.

Este suceso natural nos recuerda que la Tierra es un planeta vivo en constante transformación y que los volcanes levantan montañas en cada erupción, modelando el paisaje y arrasando con lo que se encuentran a lo largo de su camino. El rápido incremento de la población y la mayor ocupación del territorio por edificaciones o instalaciones humanas implican un inevitable aumento del peligro.

El Programa de Vulcanismo Global (en inglés, GVP) del Instituto Smithsoniano alberga la base de datos de los volcanes de la Tierra y su historia eruptiva durante los últimos diez mil años, un esfuerzo visionario en funcionamiento desde 1968 que documenta de manera única y completa las características físicas y las historias eruptivas del vulcanismo activo del planeta.

Con el fin de facilitar la investigación y divulgación en el campo del vulcanismo global, el Smithsoniano quiere aprovechar la información recopilada a lo largo de los años para lograr una nueva y audaz visión: la integración completa de estos activos informacionales para contribuir a un mejor conocimiento.

El caso de uso se basa en un conjunto de datos que está a vuestra disposición en el archivo comprimido Fuentes.zip.

Animamos al estudiantado a descargar los datos para poder alcanzar los retos propuestos.

2. Reto 1. Las bases de datos analíticas: de los datos al conocimiento

Ante la relevancia que ha alcanzado la información en nuestros días, transformar los datos recopilados en conocimiento es el gran reto para las organizaciones. Ser capaz de recopilar y manejar datos relevantes para poder tomar mejores decisiones es una de las claves para el éxito de las compañías. El reto es convertir el dato en información útil para, posteriormente, poder transformar esa información en conocimiento.

Para llevar a cabo este proceso, existen dos pilares fundamentales: las personas, con nuevos perfiles formados, y la tecnología, que nos ayudará a acelerar el proceso y alcanzar los objetivos.

Una parte importante de la tecnología que posibilita que sea efectiva la conversión de datos en conocimiento son las **bases de datos analíticas**. El almacén de datos o *data warehouse* (DW) es una base de datos orientada al análisis, es el corazón de la inteligencia empresarial.

2.1. *Data warehouse*: concepto y elementos

Bill Inmon, ingeniero especializado en bases de datos, acuñó por primera vez el término *data warehouse* para hacer referencia a un almacén de información temática orientado a cubrir las necesidades de los Sistemas de Soporte de Decisiones (DSS), que permitiese a los usuarios acceder a la información corporativa para la gestión, el control y el apoyo a la toma de decisiones[1].

Tal y como se explica en el material teórico del módulo «Introducción a las bases de datos analíticas», dentro del subapartado 1.2, los almacenes de datos según Bill Inmon están destinados a ser un reflejo del mundo real a través de los datos y se caracteriza por ser:

Orientado al tema: los datos están almacenados por materias o temas. Solo se integran los datos necesarios para el proceso de generación del conocimiento del negocio. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.

> En el contexto de nuestro caso de estudio de la actividad volcánica mundial, todos los datos relacionados con la actividad volcánica pueden ser consolidados en un almacén de datos. De esta forma, las peticiones de información sobre esta temática serán más fáciles de responder dado que toda la información reside en el mismo lugar.

[1] «Introducción a las bases de datos analíticas»; «La construcción de la factoría de información corporativa».

Integrado: los datos almacenados en el *data warehouse* deben integrarse en una estructura consistente para eliminar las inconsistencias entre los diversos sistemas operacionales.

Además del Instituto Smithsoniano, existen diferentes instituciones científicas que también disponen de información relevante relacionada con la actividad volcánica mundial. Para que toda esa información esté disponible para su análisis y facilitar la investigación es necesario integrarla en un almacén de datos.

No volátil: no hay actualización de datos en el *data warehouse*, se van acumulando datos de diferentes períodos de tiempo. El almacén de datos existe para ser leído, pero no modificado ni borrado. La información es por tanto permanente.

Los datos de la actividad volcánica se reciben con una frecuencia anual, incrementándose año a año la información disponible para su análisis.

Histórico: el tiempo es parte implícita de la información contenida en un *data warehouse*. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada en un almacén de datos sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, se cargan con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

En nuestro almacén de datos de la actividad volcánica mundial, al incrementarse su información con frecuencia anual, se dispone de un histórico de datos desde el año 1991 hasta el 2021. Al construir un almacén de datos, se dispone de histórico a lo largo de los años que posibilita el diseño de análisis evolutivos para conocer, por ejemplo, el incremento de explosividad y peligrosidad del volcán Etna durante el tiempo que estuvo activo. La historicidad de los datos, sin posibilidad de actualización, es una característica de los sistemas de *data warehouse*, no de los sistemas operacionales en los que sí realizan cambios sobre los datos, y solo disponen de la situación de los estos en un momento puntual, no a lo largo del tiempo.

A continuación, en la figura 1, se muestra con un ejemplo operacional que muestra la última foto de la actividad. Y otro de historificación de la actividad volcánica en un almacén de datos donde guardaremos un histórico que nos permitirá estudiar la evolución a lo largo del tiempo. Así, los científicos disponen de gran riqueza de datos integrados para generar un conocimiento social sobre el efecto de los volcanes en el interior de la Tierra.

Figura 1. Ejemplo de historicidad de almacén de datos

Base de datos operacional

Volcán	Índice explosividad	Días de erupción
Etna	5	472

Almacén de datos

Id	Volcán	Índice explosividad	Días de erupción	Inicio_actividad	Fin_actividad
5	Etna	2	23	2001/07/17	2002/10/30
6	Etna	4	318	2002/10/26	2003/01/28

Para interiorizar las diferencias entre un *data warehouse* y un sistema operacional, es recomendable revisar el apartado 3 del módulo «Introducción a las bases de datos analíticas», donde se tratan los diferentes aspectos que distinguen a estos dos tipos de sistemas.

A continuación, se relacionan estas características con nuestro caso de estudio.

Aspecto	*Data warehouse*	Actividad volcánica
Objetivo	La actividad más importante es el análisis y la decisión estratégica.	Las administraciones, al conocer las poblaciones con mayor actividad volcánica, podrán establecer políticas para evitar la construcción de viviendas.
Granularidad	Datos en diferentes niveles de detalle y agregación.	• Los cinco países con volcanes con el mayor número de días en erupción en 2020. • Promedio de días por erupción de los volcanes.
Validez de datos	Importancia del dato histórico.	Este sistema integrado contiene datos de los volcanes de la Tierra y su historia eruptiva durante los últimos diez mil años. Información sobre peligros volcánicos globales, eventos históricos, exposición de la población, vulnerabilidad e impacto, etc.
Tipo de operaciones	Predomina la consulta.	Para el análisis de la actividad volcánica mundial, el sistema permite responder a múltiples preguntas para cubrir las necesidades de sus usuarios potenciales.
Procesamiento	Predomina el proceso masivo.	La carga de datos tiene una primera fase en la que se realiza una carga inicial y, *a posteriori*, una segunda fase para realizar las cargas incrementales de los datos nuevos que van llegando.
Volumen de respuesta	Importancia de la respuesta masiva.	Los procesos automatizados permiten realizar análisis *ad hoc* logrando aumentar la productividad, ya que el tiempo de respuesta se acorta a la hora de hacer consultas y acceder a la información.
Estructura	Visión multidimensional.	La actividad volcánica mundial se puede analizar desde tres perspectivas: desde una visión temporal, desde una visión geográfica y desde una visión del tipo de volcán que genera dicha actividad volcánica.

Aspecto	*Data warehouse*	Actividad volcánica
Usuarios	Usuarios de perfiles analíticos.	Los usuarios finales que harán uso del sistema son responsables de las administraciones públicas y gobiernos, directivos de empresas turísticas y científicos.
Orientación	Explotación de toda la información interna y externa relacionada con el negocio.	La integración de las diversas fuentes de datos (y formatos) disponibles con información de estadísticas de volcanes, información sobre peligros volcánicos globales, eventos históricos, exposición de la población y vulnerabilidad e impacto servirán para el análisis de datos y la posterior toma de decisiones.

Antes de finalizar este apartado insistir en que los principales objetivos de un almacén de datos son: ser un repositorio central e integrado de información empresarial y ser un repositorio base para procesos de análisis y *reporting*.

Es importante profundizar en el estudio de los objetivos de las bases de datos analíticas en el apartado 2 del módulo «Introducción a las bases de datos analíticas».