

Caso práctico: almacén de datos para el análisis de la cobertura de inmunización

Autora: Nerea Sevilla Marchena

Índice

- Introducción
- Contexto
- Enunciado
- Programas
- Bibliografía

Introducción

Este material está creado para practicar el diseño y la implementación del almacén de datos como sistema de almacenamiento en el análisis de datos.

El diseño, el desarrollo y la implantación de un sistema de *data warehouse* (DW) en cualquier organización supone llevar a cabo un proyecto que puede durar meses, o incluso años, en función del alcance del proyecto, de la naturaleza y del grado de madurez de la organización, así como de la participación de los equipos multidisciplinares que van implementando diferentes proyectos en un proceso de mejora continua del almacén.

El objetivo de este caso no es desarrollar un almacén de datos que dé respuesta a todas las necesidades, sino entender y utilizar las metodologías para desarrollar este tipo de proyectos en un contexto real, pasando por todas las fases que comprenden proyectos de esta tipología:

1. **Análisis, diseño e implementación:** consiste en desarrollar e implementar un almacén de datos que permita la gestión de la información disponible.
2. **Carga:** implica diseñar e implementar los procesos de carga de datos necesarios para disponer de información en el almacén de datos implementado.

3. **Explotación:** conlleva la creación de informes y elementos de análisis multidimensional para la explotación de la citada información.

Con el fin de poder desarrollar un proyecto lo más específico posible, el estudiante tendrá que afrontar el reto de desarrollar un almacén de datos que solo describe parte de los servicios que se pueden ofrecer, basándose en datos tratados en el caso y que formarían parte de un sistema real.

A partir de unas necesidades de negocio acotadas, el estudiante deberá adquirir un conocimiento básico del entorno tecnológico, de los procesos de negocio, de las necesidades existentes y definir una propuesta adecuada que responda a ellas.

Mediante el desarrollo del caso, el estudiante se va a encontrar con los problemas, las dudas y las dificultades que se plantean en un proyecto de estas características.

Contexto

Las vacunas constituyen, junto con la potabilización del agua y la depuración de las aguas residuales, el método más eficaz para disminuir el número de personas enfermas y fallecidas por infecciones. Además, permiten mejorar la salud y aumentar la esperanza de vida de la población. Se estima que, gracias a la vacunación, se previenen unos 2,5 millones de fallecimientos cada año.

Los esfuerzos en pos de la inmunización universal han cobrado un nuevo impulso. En mayo de 2012,¹ la Asamblea Mundial de la Salud, el órgano decisorio de la Organización Mundial de la Salud (OMS), aprobó el **Plan de Acción Mundial sobre Vacunas² (GVAP, Global Vaccine Action Plan)** para responder a los retos de la inmunización y combatir las enfermedades prevenibles con vacunas en todo el mundo.

Entre los objetivos estratégicos que propone el Plan de Acción Mundial sobre Vacunas está el cumplimiento del compromiso de cobertura de vacunación. **Concretamente, se pretende:**

- **Meta para 2015:**
 - **Objetivo 1.** Conseguir una cobertura nacional del 90 % para las vacunas contra la difteria, el tétanos y la tos ferina (vacunas DPT).
 - **Objetivo 2.** Conseguir que en cuatro regiones de la OMS, el 100 % de países que han declarado la interrupción de la transmisión endémica del virus del sarampión (cobertura regional del 100 %).
- **Meta para 2020:**
 - **Objetivo 1.** Conseguir una cobertura nacional del 90 % para todas las vacunas en programas nacionales.
 - **Objetivo 2.** Conseguir que en cinco regiones de la OMS, el 100 % de países que han declarado la interrupción de la transmisión endémica del virus del sarampión (cobertura regional del 100%).

¹ Comunicado de prensa: la Asamblea Mundial de la Salud respalda un nuevo plan para aumentar el acceso global a las vacunas
https://www.who.int/immunization/newsroom/press/wha_endorses_gvap/en/

² Plan de Acción Mundial sobre Vacunas: https://www.who.int/immunization/global_vaccine_action_plan/DoV_GVAP_2012_2020/es/

Se estima que, si se cumple con los objetivos de cobertura para la introducción y/o la utilización continua de únicamente diez vacunas³ (contra la hepatitis B, el *Haemophilus influenzae* tipo b, el virus del papiloma humano, la poliomielitis, el sarampión, el meningococo A, el neumococo, el rotavirus, la rubéola y la fiebre amarilla), se podrían evitar de 24 a 26 millones de futuras muertes en 94 países durante el decenio.

La Organización Mundial de la Salud (OMS), para el cumplimiento de los objetivos del Plan de Acción Mundial sobre Vacunas, precisa de la construcción de un almacén de información de la cobertura de la inmunización de los Estados miembros. El análisis de dicha información permitirá generar informes de progreso sobre el GVAP que servirán de guía para las estrategias de inmunización a escala mundial y regional. El sistema permitirá la recogida, la integración y el análisis de la información de alta calidad como apoyo a la toma de decisiones de sus usuarios potenciales, con lo que se contribuirá al buen funcionamiento del plan de inmunización.

El desarrollo del almacén de datos para el análisis de cobertura de inmunización requiere de los siguientes aspectos:

- Diseño y construcción del *data warehouse* que permita la integración de datos de diferentes fuentes.
- Diseño e implementación de los procesos de carga inicial al *data warehouse*.
- Implantación de un sistema analítico que ofrezca apoyo a la toma de decisiones de los usuarios potenciales y permita analizar la información sobre la cobertura de la vacunación.

Con datos de mayor calidad y herramientas nuevas de análisis, los gestores del GVAP pueden utilizar la información para mejorar el funcionamiento del programa de inmunización, repartir los fondos adecuadamente y seguir el avance de manera más eficaz.

Usuarios potenciales

Como fase inicial del diseño del almacén de datos de cobertura de inmunización realizaremos el análisis de requerimientos teniendo en cuenta quiénes serán los usuarios potenciales. Tendremos que tener en cuenta que el sistema responde a sus necesidades y genera información útil.

Los usuarios finales que harán uso del sistema son los siguientes:

- Gestores del GVAP que utilizarán el almacén de datos para seguir de cerca el progreso del plan de vacunas y así disponer de información valiosa para supervisar su desempeño y medir su impacto. Concretamente, necesitarán análisis del cumplimiento de los objetivos definidos en el GVAP.
- Gobiernos de los Estados miembros como principales proveedores de la inmunización, deberán contar con información oportuna cuando la población manifieste sus preocupaciones sobre la cobertura de vacunación para conservar así la confianza pública. Por eso, cualquier análisis evolutivo sobre la cobertura de inmunización será de gran utilidad para los gobiernos.

³ Cobertura de vacunación: <https://www.who.int/es/news-room/fact-sheets/detail/immunization-coverage>

- Organismos mundiales, como la OMS y UNICEF, para promover, compartir y respaldar el proceso de toma de decisiones basado en datos científicos para todas las partes interesadas en el desarrollo, la salud y la inmunización.
- Profesionales de la salud, para poder ofrecer servicios de inmunización de alta calidad basados en la información extraída del almacén de datos de cobertura de inmunización.
- Fabricantes de vacunas, para consultar los datos del sistema sobre índices de cobertura de las vacunas que les sirva de apoyo para seguir desarrollando, produciendo y suministrando vacunas innovadoras de alta calidad.

Fuentes de datos

Uno de los objetivos de este caso de estudio es integrar las fuentes de datos proporcionadas para poder realizar diferentes tipos de análisis. Los ficheros de datos de inmunizaciones se han obtenido del portal de datos de la Organización Mundial de la Salud (<https://www.who.int/es/>) y en el repositorio de datos abiertos de la ONG UNICEF (<https://data.unicef.org/>).

En concreto, disponemos de información anual en materia de inmunización extraída del Observatorio Mundial de la Salud (<https://www.who.int/data/gho/data/themes/theme-details/GHO/immunization>) y del área de estadísticas de UNICEF (<https://data.unicef.org/resources/dataset/immunization/>).

La relación de ficheros que utilizaremos para la carga inicial al *data warehouse* es la que sigue:

- Series cronológicas de coberturas de inmunización por región de la OMS, Estados miembros y vacunas. Años 1966-2018. Nombre de archivo: *coverage_estimates_series.xls*.

Fuente disponible en:

https://www.who.int/immunization/monitoring_surveillance/data/en/
https://apps.who.int/immunization_monitoring/globalsummary/timeseries/tswucoveragedtp1.html

- Datos de la encuesta de cobertura de inmunización, número de niños vacunados y población objetivo por región de la OMS, Estados miembros y vacunas. Años 1997-2018. Nombre de archivo: *wuenic2018rev_data_2019-12-19.cs*, donde 2019-12-19 es la fecha de cuando se realizó la descarga.

Fuente disponible en:

https://unicef.shinyapps.io/wuenic_analytics/

Aunque se proporciona la fuente de datos junto con el enunciado del caso, si se desea descargar de la web WUENIC Analytics de UNICEF, es necesario realizar unos pasos previos. Concretamente, se debe elegir los años de los datos (*Year*) y el grupo de países (*Group Type*), en la pestaña Charts; la fuente de datos proporcionada corresponde a los años entre 1997 y 2018 y para los

países miembros de la OMS, siguiendo el tipo WHO-Wold Health Organization. Después ya se puede proceder a la descarga de los datos en la pestaña Data, utilizando el botón *download data*, disponible en la parte inferior de la página.

- Cuatro ficheros .xml con datos sobre casos de sarampión (*measles*) de los países miembros de la región de Europa. Años 2008-2018.

Nombre de archivos:

Sarampion_5001.xml: Número de casos detectados.

Sarampion_5002.xml: Número de muertes.

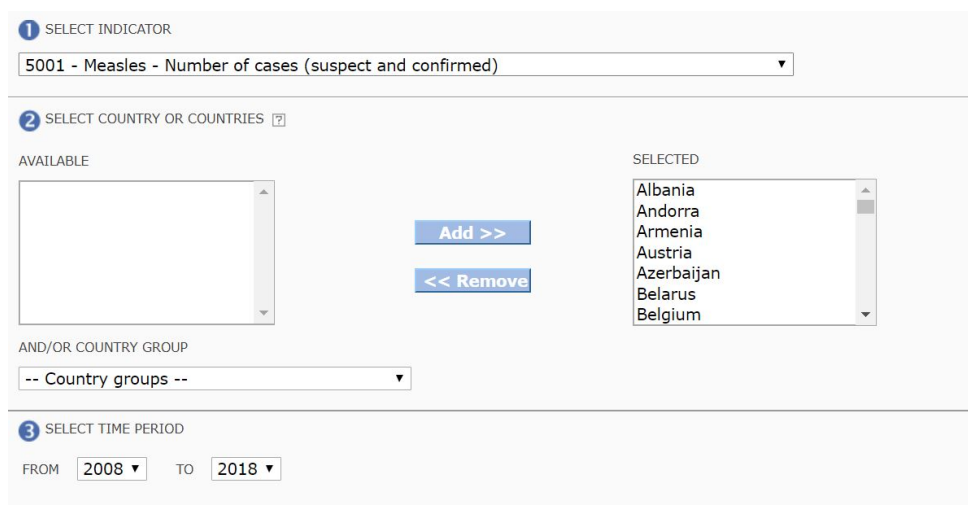
Sarampion_5003.xml: Número de hospitalizaciones.

Sarampion_5005.xml: Número de casos confirmados en laboratorio.

Fuente disponible en:

<http://data.euro.who.int/cisid/?TabID=507227>

A pesar de que se proporciona la fuente de datos junto con el enunciado del caso, si se desea descargar de la web de la región europea de la OMS, es necesario realizar unos pasos previos. Concretamente, como primer paso se debe elegir la base de datos *measles* (sarampión), y como segundo paso elegir el indicador, el grupo de países y los años de los datos y ejecutar la consulta mediante el botón de *submit query*, tal como se muestra en la siguiente imagen.



La descarga se realiza desde la pantalla de visualización de resultados de la consulta, mediante el desplegable *options*

La fuente de datos proporcionada corresponde a cuatro indicadores de la enfermedad del sarampión, del grupo de países miembros según la OMS (WHO european region) y del periodo comprendido entre 2008 y 2018.

- Regiones de la OMS (https://en.wikipedia.org/wiki/WHO_regions). Los Estados miembros de la OMS se agrupan en seis regiones de la OMS: región de África, región de las Américas, región de Asia Sudoriental, región de Europa, región del Mediterráneo Oriental y región del Pacífico Occidental. Nombre de archivo: REGION.json.

Fuente disponible en:

<http://apps.who.int/gho/data/node.metadata.REGION?lang=en>

- Países miembros. Información de los países miembros de la OMS. Nombre de archivo: COUNTRY.json.

Fuentes disponibles en:

<http://apps.who.int/gho/data/node.metadata.COUNTRY?lang=en>

- Causas de mortalidad. Información de los países miembros de la OMS. Nombre de archivo: MORTCAUSE.json.

Fuentes disponibles en:

<http://apps.who.int/gho/data/node.metadata.MORTCAUSE?lang=en>

Se tendrá en cuenta que, con frecuencia anual, los países miembros de cada región de la OMS nos enviarán los datos relativos a los indicadores de la cobertura de vacunación y, por lo tanto, se realizarán cargas incrementales para la integración de esos datos en el *data warehouse* para su análisis en nuestro sistema.

Enunciado

Análisis y diseño del *data warehouse*

A partir del análisis del contexto del caso y de las fuentes de datos disponibles, el estudiante deberá diseñar y proponer un almacén de datos que ofrezca soporte al funcionamiento del almacén de datos para el análisis de la cobertura de la inmunización.

Mediante la metodología de diseño de un *data warehouse* propuesta en la asignatura, el estudiante debe llevar a cabo:

- **El análisis de requerimientos:** como resultado se generará un documento que describirá las preguntas a las que el sistema dará respuesta para los usuarios potenciales del mismo.
- **El análisis de fuentes de datos:** se deben revisar las fuentes de datos proporcionadas, qué tipo de información contienen, cuál es su formato, y qué cantidad representan para la carga inicial.
- **El análisis funcional:** se debe proponer el tipo de arquitectura para la factoría de información que mejor se adecue al proyecto (por ejemplo, si es necesario un *data mart* operacional o una estructura de carga intermedia).
- **Diseño del modelo conceptual, lógico y físico del almacén de datos:** se deben identificar, diseñar e implementar las tablas de hechos, las dimensiones y los atributos que describen la información.

Para este apartado, el estudiante debe preparar un documento en el que se expliquen las secciones anteriores.

Se deberá tener en cuenta que, para el desarrollo del DW, es preciso definir correctamente los hechos (*facts*), las dimensiones de análisis (*dimensions*) y los atributos que nos permitan tener el nivel de granularidad suficiente para medir y presentar los objetivos que se definan en el análisis de requerimientos.

Carga y explotación de datos

A partir de la solución oficial de la primera práctica (PRA1), el estudiante deberá diseñar, implementar y ejecutar los procesos de extracción, transformación y carga de los datos de las fuentes de datos proporcionadas. Tras la carga efectiva de los datos en el almacén, se debe implementar un cubo multidimensional para la explotación de la información como apoyo a la toma de decisiones de los usuarios potenciales.

El estudiantado, por lo tanto, deberá de proceder a lo siguiente:

- Identificar los procesos ETL (extracción, transformación y carga de datos) hacia el almacén de datos.
- Diseñar y desarrollar los procesos ETL mediante las herramientas de diseño proporcionadas.
- Implementar con trabajos los procesos ETL para su carga efectiva planificada.
- Diseñar un modelo OLAP (*multidimensional on line analytical processing*) para el análisis multidimensional de la información disponible en el almacén de datos.

Programas

Para el presente caso, la UOC proporciona una infraestructura de escritorio virtual (VDI, *Virtual Desktop Infrastructure*) con todo el software preconfigurado con las siguientes características:

- Sistema operativo: Windows 10.
- Base de datos: base de datos remota Microsoft SQL Server 2016 accesible desde cliente mediante SQL Server Management Studio 17.
- Herramienta para la creación de cubos OLAP: Visual Studio 2017.
- Herramienta de diseño de ETL: Spoon – Pentaho Data Integration 8.0.
- Herramienta de creación de informes: PowerBI Desktop.

Bibliografía

Material de la asignatura Data Warehouse de la UOC.

Kimball, R. (2013). *The Data Warehouse Toolkit*. (3.^a edición). Nueva York: John Wiley & Sons Inc.

Inmon, W. H., Imhoff, C. y Sousa, R. (1998). *Corporate Information Factory*. EE. UU.: John Wiley & Sons Inc.

Inmon W.H. (1996) . *Building the Data Warehouse*. (2.^a edición). EE. UU.: John Wiley & Sons Inc.

Inmon, W. H., Strauss, D., Neushloss, G. (2008). *DW 2.0: The Architecture for next generation of Data Warehousing*. EEUU: Morgan Kaufman Series.

Krishnan, K. (2013). *Data Warehousing in the Age of Big Data. The Morgan Kaufmann Series on Business Intelligence*.

Enlaces a internet:

Getting Started with SQL Server Analysis Services:

<http://www.mssqltips.com/sqlservertip/1167/getting-started-with-sql-server-analysis-services/>

MSDN Analysis Services tutorial:

<https://docs.microsoft.com/es-es/analysis-services/analysis-services-tutorials-ssas?view=asallproducts-allversions>

Tutorial Pentaho Data Integration:

<http://wiki.pentaho.com/display/EAI/Latest+Pentaho+Data+Integration+%28aka+Kettle%29+Documentation>