UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

Trabajo Almacenes de Datos

Integración de datos (ETL) de un almacén de UCI Sanitaria

Realizado por
De Pablo Diego y Soriano Juan

Profesor encargado:
Luque Baena Rafael Marcos

Departamento
Lenguajes y Ciencias de la Computación

MÁLAGA, DICIEMBRE de 2024

UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

ESTUDIANTES DE INGENIERÍA BIOINFORMÁTICA

Integración de datos (ETL) de un almacén de UCI Sanitaria

Almacenes de Datos

Realizado por
De Pablo Diego y Soriano Juan

Profesor encargado:
Luque Baena Rafael Marcos

Departamento
Lenguajes y Ciencias de la Computación

UNIVERSIDAD DE MÁLAGA

MÁLAGA, DICIEMBRE DE 2024

Contents

1 Introducción

El diseño y desarrollo de **almacenes de datos** es crucial para el análisis clínico, mejorando la toma de decisiones médicas y la calidad de atención a los pacientes. Para este proyecto, se utilizará una visión parcial de un almacén de datos basado en la *Base de Datos de Investigación Colaborativa eICU*, la cual contiene una vasta cantidad de información sobre los ingresos en **Unidades de Cuidados Intensivos (UCI)** en diversos hospitales de los Estados Unidos.[1]

El enfoque de este trabajo está dirigido específicamente a los **pacientes con problemas respiratorios**. Estos casos son de especial relevancia en entornos de cuidados críticos, ya que las afecciones respiratorias representan una de las principales causas de ingreso en las UCI. Por lo cual un almacén de datos centrada en estas patologías sigue representando una gran fuente de información y con mucho interés para hacer un análisis de datos. [4, 5].

En este trabajo, el **hecho principal** está constituido por los ingresos en la UCI, complementado por un conjunto de dimensiones que permiten analizar diversas variables. La selección de las tablas y atributos más relevantes de la base de datos eICU será el mayor reto para construir un **almacén de datos eficiente y de interés analítico**, que no solo facilite la consulta de información clave, a su vez soporte y facilité tanto la investigación clínica como la optimización de los protocolos de tratamiento.

2 Objetivos

El objetivo de este trabajo es desarrollar un **diseño conceptual y lógico** de un almacén de datos centrado en los pacientes ingresados en la **UCI** con afecciones respiratorias, seleccionando las partes más relevantes de la base de datos para el análisis de estas patologías.

Se procederá a la restauración de dicha base de datos en *SQL Server*, para realizar una selección de las tablas más relevantes que contribuyan a la construcción del almacén de datos centrado en estos pacientes. En cada una de las tablas seleccionadas, se identificarán los atributos más significativos para el análisis clínico y el seguimiento de los pacientes con problemas respiratorios.

Una vez completada la selección de dimensiones y tablas de hechos, se diseñará el modelo conceptual utilizando la herramienta *draw.io*, y posteriormente se desarrollará el modelo lógico mediante un diagrama de base de datos en el entorno de *SQL Server*.

3 Modificación del almacén de datos

NOTA: Si se ha modificado el almacén de datos de la tarea anterior debéis comentar las modificaciones realizadas (y justificarlas) en una sección de esta tarea.

La base de datos basad*eICU-CRD*, desarrollada por el Philips eICU Research Institute (eRI), contiene datos desidentificados de pacientes ingresados en Unidades de Cuidados Intensivos (UCI). Esta base de datos, que forma parte del programa de telesalud en cuidados críticos de Philips, documenta detalles sobre diagnósticos, tratamientos, pruebas y resultados.

En los cambios realizados al almacén de datos, se ha modificado algunas tablas y relaciones para ajustarlas a las recomendaciones del profesor y para mejorar la estructura del modelo. A continuación, se detallan los ajustes más relevantes:

IngresoUCI: El único cambio que se ha realizado en esta tabla ha sido la adición del atributo *hospitalDischargeOffset*, siguiendo la indicación de almacenar el tiempo de hospitalización del paciente. Este campo permite calcular el tiempo exacto que el paciente estuvo ingresado.

Tiempo: Se ha decidido almacenar únicamente los campos *hospitalDischargeTime24* y *hospitalDischargeYear*, ya que no existe un campo para el año de admisión del hospital (*hospitalAdmitYear*), como sugirió el profesor. Este cambio simplifica el almacenamiento de información temporal sobre la estancia del paciente.

Paciente: Se ha creado una nueva tabla llamada *Paciente*, donde el identificador principal (PK) es *uniquePID*. También se ha incluido el campo de la *edad*. Además, se ha desarrollado una jerarquía paralela para los atributos de *género* y *etnia*, basándose en el trabajo de los compañeros Gonzalo y Carmen, ya que esta estructura resulta adecuada para la representación de estos datos.

Cambios en las relaciones

1. **Paciente**: Se ha establecido una relación de **1-n** entre Paciente e IngresoUCI, donde cada paciente puede tener múltiples ingresos a la UCI en diferentes momentos.

2. **Diagnosis**: El profesor ya había indicado que la relación entre IngresoUCI y Diagnosis es **n-m**. Un paciente puede tener varios diagnósticos durante un ingreso, y un diagnóstico específico puede estar asociado a varios ingresos de diferentes pacientes.

3. **Medicamentos**: La relación aquí también es **n-m**, ya que cada ingreso puede requerir varios medicamentos, y un medicamento puede administrarse en varios ingresos.

4. **Alergia**: La relación entre Paciente y Alergia es **n-m**, dado que un paciente puede tener múltiples alergias, y cada alergia puede estar relacionada con varios ingresos, debido a su impacto en el tratamiento.

5. **RespiratoryCharting**: Se ha mantenido una relación de **1-n** entre IngresoUCI y RespiratoryCharting, ya que cada ingreso puede tener varios registros de datos respiratorios (como la saturación de oxígeno).

6. **RespiratoryCare**: En este caso, la relación también es de **1-n**. Cada ingreso puede tener múltiples intervenciones de cuidado respiratorio, como la administración de oxígeno o ventilación mecánica.

7. **apacheApsVar**: La relación entre IngresoUCI y *apacheApsVar* puede ser de **1-1** o de **1-n**, dependiendo de si se desea almacenar un solo resumen de evaluación APS por ingreso o varios componentes evaluados.

8. **Admissiondx**: La relación es **1-n**, ya que cada ingreso tiene un diagnóstico primario de admisión, aunque los diagnósticos secundarios se almacenan en otra tabla.

9. **Hospital**: La relación es de **1-n**. Un hospital puede tener múltiples ingresos a la UCI asociados, dado que cada ingreso pertenece a un único hospital.

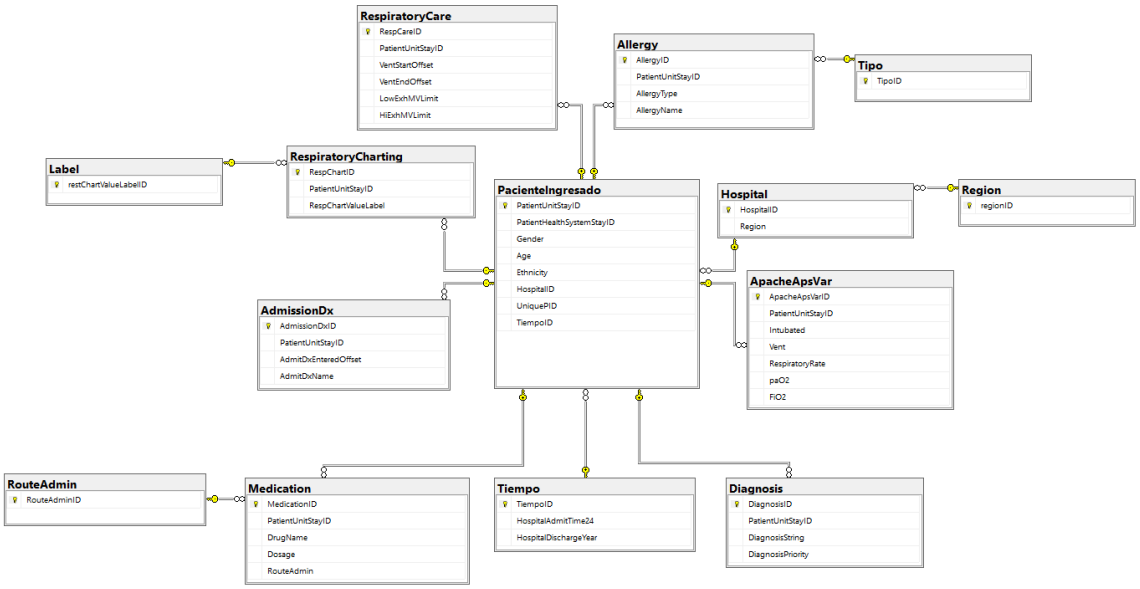

Figure 1: Diseño lógico

4 Dificultades encontradas

Una de las principales dificultades fue la complejidad de la base de datos eICU, que incluye una gran cantidad de tablas y atributos. Esto exigió un análisis detallado para identificar las tablas y campos clave en un modelo centrado en pacientes con enfermedades respiratorias. Además, enfrentamos problemas de permisos al intentar visualizar el modelo relacional en SQL Server, lo que requirió modificar las autorizaciones del propietario de la base de datos para acceder a los diagramas de relación.

Además el comienzo del trabajo podría ser lo más angustioso, al tener tanta información y opciones llega a ser un poco abrumador, desde la selección de una población concreta y modelar un almacén para dicha población termina dejando muchas dudas sobre cuantas tablas es esperable eliminar, si se esta simplificando de más o se esta tomando una decisión que afectará los siguientes apartados.

5 Conclusiones

Durante el desarrollo del almacén, se tomó la decisión de integrar únicamente los datos más relevantes de pacientes con patologías respiratorias, lo que implicó reducir significativamente el número de tablas y atributos presentes en la base de datos original. Este proceso permitió centrarse en la información esencial para el análisis, facilitando la optimización del sistema y garantizando que los datos fueran manejables y claros para los futuros pasos del proyecto. A lo largo de esta fase, se pudo aprender mucho sobre el contenido y la estructura de cada tabla, lo que ayudó a realizar una selección cuidadosa de las que mejor se alineaban con los requerimientos del modelo de datos, mejorando así la calidad del análisis final.

El diagrama conceptual fue una herramienta crucial en este proceso, ya que permitió plasmar de forma clara las ideas y objetivos principales del almacén de datos. Gracias a este enfoque, se logró una mejor organización y se tomaron de-

cisiones informadas en cuanto a las entidades y relaciones a incluir. Este esquema conceptual facilitó no solo una representación más clara de la estructura, sino que también permitió anticipar ajustes necesarios y optar por decisiones estratégicas que beneficiarían el rendimiento y escalabilidad del sistema a largo plazo.

Por otro lado, el diagrama lógico implementado, a través del modelo en copo de nieve, ofreció una representación detallada de la base de datos. Este modelo permitió una alta normalización, reduciendo la redundancia de datos sin comprometer la consistencia de la información. A pesar de la mayor complejidad en el diseño, este enfoque aseguró una estructura sólida y flexible para futuros análisis.

Finalmente, habiendo completado todas las etapas de análisis, diseño conceptual y lógico, el sistema se encuentra preparado para proceder con la fase de integración de datos.

6 Github y conjunto de instrucciones para su correcto despliegue en SQL Server.

Todo el proyecto está accesible en github [3] donde se detalla más específicamente como desplegar en SQL.

References

[1] Chamsi Bah, Sultan Alharthi, and Ashraf El Metwally. "Clinical data warehousing: A review of current systems and future directions". In: *Journal of Healthcare Engineering* 2017 (2017), pp. 1–11. DOI: 10.1155/2017/8326740.

[2] eICU Collaborative Research Database. *eICU Collaborative Research Database*. https://eicu-crd.mit.edu/about/eicu/. Accessed: 2024-11-14. 2024. URL: https://eicu-crd.mit.edu/about/eicu/.

[3] Diegodepab. *Almacén UCI Sanitaria*. https://github.com/Diegodepab/almacen_UCI_Sanitaria. Accessed: 2024-11-14. 2024. URL: https://github.com/Diegodepab/almacen_UCI_Sanitaria.

[4] Christopher Pekar, Sven Gordan, and Ewan Goligher. "Epidemiology of respiratory failure in the intensive care unit: A review". In: *Critical Care* 25.1 (2021), pp. 1–9. DOI: 10.1186/s13054-021-03772-y.

[5] Jean-Louis Vincent, Yasser Sakr, and V Marco Ranieri. "The epidemiology of respiratory failure in the ICU". In: *Chest* 129.1 (2006), pp. 90–99. DOI: 10.1378/chest.129.1.90.

UNIVERSIDAD DE MÁLAGA | uma.es

E.T.S de Ingeniería Informática
Bulevar Louis Pasteur, 35
Campus de Teatinos
29071 Málaga

E.T.S. DE INGENIERÍA INFORMÁTICA