Instituto Tecnológico de Costa Rica

IC-4302 Bases de Datos II

Proyecto Programado #2

Base de Datos de Home Credit Group

Estudiantes:

Joshua Esteban Sancho Burgos / 2021108350

Gabriel Arturo Alfaro Ulate / 2022321442

Fabian Rojas Ugalde / 2022045135

Grupo 50

Campus San Carlos

23 de noviembre del 2023

Tabla de Contenidos

Resumen Ejecutivo

El segundo proyecto programado del curso de Bases de Datos II se centra en el desarrollo de un sistema de bases de datos en colaboración con Home Credit Group para mejorar la experiencia crediticia de personas con historiales crediticios limitados. Desde el diseño hasta la implementación, se prioriza la fusión entre teoría y práctica, abordando la migración de datos, la seguridad, el rendimiento y la visualización de datos a través de un dashboard. Se destaca la colaboración con Home Credit para usar datos alternativos en la predicción crediticia. Aunque el proyecto alcanza sus metas, se recomienda mantener una supervisión constante, actualizar la base de datos, explorar técnicas avanzadas y recopilar feedback para un uso óptimo del sistema.

Objetivo General

El objetivo general del proyecto consiste en desarrollar un sistema de bases de datos integral y eficiente, en colaboración con Home Credit Group, con el propósito de mejorar la experiencia de préstamos para individuos con historiales crediticios insuficientes o inexistentes. Este sistema busca utilizar datos alternativos para optimizar la predicción de la capacidad de reembolso de los clientes, garantizando préstamos con condiciones que empoderen a los clientes para alcanzar el éxito financiero y evitando el rechazo de aquellos capaces de pagar.

Objetivos Específicos

- Diseñar un modelo de base de datos aplicado a situaciones reales que resuelva problemas específicos, priorizando la normalización, optimización y seguridad del Sistema de Gestión de Bases de Datos (SGBD).
- Interactuar con diversas herramientas para diseñar y desarrollar los elementos fundamentales de la base de datos, garantizando su funcionalidad en un entorno de aplicación real.
- Desarrollar modelos E-R y relacionales que sirvan como base para la construcción de la base de datos. Segmentar las tablas en esquemas lógicos para una mejor organización y eficiencia.
- Implementar un servidor de auditoría para registrar transacciones y asegurar la integridad de los datos mediante funciones de inserción con triggers.
- Mejorar la eficiencia del sistema mediante consultas SARGABLES para un entorno web y configurar al menos cinco índices no clúster.
- Establecer roles de usuarios como administrador, usuario normal y respaldo para contribuir a la seguridad y administración del sistema.
- Crear un dashboard visual que presente los datos derivados de la base de datos de manera clara y comprensible.

Introducción

El segundo proyecto correspondiente al segundo semestre del año 2023 para los estudiantes del curso de Bases de Datos II, dirigido por el Profesor Efrén Jiménez Delgado, se centra en la creación de un sistema de bases de datos completo y funcional. Esta iniciativa desafía a los estudiantes a abordar desafíos reales en el ámbito comercial, priorizando la eficiencia y estabilidad del Sistema de Gestión de Bases de Datos (SGBD).

En este contexto, el proyecto busca concebir un modelo de base de datos aplicado a situaciones reales, con un enfoque meticuloso en la resolución de problemas específicos mediante la garantía de la normalización, optimización y seguridad del SGBD. Además, se pretende la implementación práctica de la lógica de negocios del SGBD, explorando la interacción con diversas herramientas para el diseño y desarrollo de la base de datos.

La esencia del proyecto radica en esta fusión de teoría y práctica. Esto incluye la creación de modelos E-R y relacionales, la segmentación de tablas en esquemas lógicos, la implementación de un migrador para la actualización eficiente de la base de datos, y la configuración de un servidor de auditoría para registrar transacciones. La gestión de usuarios con roles específicos como administrador, usuario normal y respaldo es vital para la seguridad y administración del sistema.

Simultáneamente, los estudiantes se embarcan en una colaboración con Home Credit Group, una institución financiera comprometida con la inclusión financiera. En este proyecto conjunto, se enfocan en mejorar la experiencia de préstamos para aquellos con historial crediticio insuficiente o inexistente. La empresa hace uso de una amplia gama de datos alternativos, incluyendo información transaccional y de telecomunicaciones, para predecir la capacidad de reembolso de sus clientes.

El equipo de Ingeniería en Computación tiene la responsabilidad de recopilar, gestionar y optimizar datos provenientes de diversas fuentes para construir una base de datos robusta. Esta base de datos, diseñada para optimizar la predicción de capacidades de reembolso, se propone asegurar que aquellos clientes con la capacidad de pago no sean rechazados y que los préstamos se otorguen con condiciones que empoderen a los clientes para alcanzar el éxito financiero.

Además de la base de datos, se ha de desarrollar una interfaz visual. Esta interfaz, materializada como un dashboard, presentará visualmente los datos derivados de la base de datos, proporcionando una comprensión clara para los stakeholders de Home Credit Group.

Este proyecto desafía a los estudiantes a aplicar habilidades teóricas y prácticas en un contexto real. Busca ofrecer una experiencia de préstamo positiva y segura para una población desatendida, al mismo tiempo que contribuye al desarrollo de soluciones innovadoras en el sector financiero, abordando desafíos reales en el ámbito empresarial y tecnológico actual.

Descripción del problema

En un escenario donde numerosas personas enfrentan dificultades para obtener préstamos debido a historiales crediticios insuficientes o inexistentes, surge la problemática de exclusión financiera. Esta situación conlleva a menudo a que esta población sea objeto de abusos por parte de prestamistas poco confiables.

En respuesta a esta problemática, surge Home Credit Group, una entidad dedicada a ampliar la inclusión financiera para la población no bancarizada. La misión de Home Credit radica en ofrecer una experiencia de préstamo positiva y segura para aquellos que enfrentan dificultades crediticias. Para lograr este objetivo, la empresa hace uso de una amplia gama de datos alternativos, incluyendo información transaccional y de telecomunicaciones, para prever la capacidad de reembolso de sus clientes.

Actualmente, Home Credit emplea diversos métodos estadísticos y de aprendizaje automático para realizar estas predicciones. Sin embargo, la empresa busca desafiar a los Kagglers a desbloquear todo el potencial de sus datos. El propósito es asegurar que los clientes capaces de realizar pagos no sean rechazados y que los préstamos se otorguen con condiciones que permitan el empoderamiento financiero de sus clientes.

El evento en cuestión tuvo lugar con una serie de fechas límite importantes: la fecha de inscripción, la fecha límite para la unión de equipos y la fecha límite final para la presentación de soluciones. Todas estas fechas se rigieron bajo el horario UTC y marcaron el inicio y el final de una competición que desafiaba a los participantes a aprovechar al máximo los datos de Home Credit Group para mejorar el acceso a préstamos y asegurar una experiencia crediticia positiva para una población marginada.

Como parte de su desafío para mejorar la experiencia crediticia, Home Credit Group busca migrar datos clave almacenados en diversos archivos .csv disponibles en Kaggle. Estos archivos contienen información esencial, desde datos estáticos de solicitudes de préstamos hasta balances mensuales de créditos anteriores y aplicaciones pasadas para préstamos con Home Credit.

Los archivos críticos para esta migración incluyen la tabla principal, application_{train|test}.csv, que se divide en dos archivos, uno para entrenamiento y otro

para prueba, detallando datos estáticos de todas las solicitudes de préstamos. Además, se encuentran archivos como bureau.csv, que registra créditos previos de los clientes en otras instituciones financieras, bureau_balance.csv para balances mensuales de esos créditos y otros archivos que describen el historial crediticio detallado de los clientes con Home Credit.

El objetivo principal es migrar esta variada información desde los archivos .csv hacia una base de datos estructurada en SQL. Esta base de datos servirá como un repositorio centralizado y optimizado para almacenar y gestionar los datos, facilitando análisis futuros y la aplicación de soluciones basadas en datos.

Además de la migración, se planea la construcción de la base de datos SQL para alojar la información migrada. Esta base de datos será diseñada de manera que proporcione una estructura sólida y eficiente, facilitando consultas y análisis posteriores.

La base de datos SQL resultante servirá como base fundamental para el desarrollo de una interfaz con un dashboard. Este dashboard será diseñado para presentar visualmente componentes significativos de la base de datos. Permitirá a los stakeholders de Home Credit Group obtener información clave de manera clara y accesible, contribuyendo así a la toma de decisiones fundamentadas y estratégicas dentro de la empresa.

Desarrollo

Para construir la solución al problema del proyecto programado, se diseñó una base de datos en SQL Server para almacenar la información de los archivos .csv de la empresa de Home Credit Group. Esta base de datos, denominada "Home_Credit_Default_Risk", se ha diseñado para almacenar información vital relacionada con el historial crediticio y las solicitudes de préstamos de los clientes de Home Credit Group. Está compuesta por varios esquemas: "client", "credit", y "repayment", cada uno dedicado a aspectos específicos del historial crediticio y los detalles de los clientes.

El esquema "client" contiene tablas fundamentales que almacenan información detallada sobre los clientes y sus solicitudes de préstamos. Por ejemplo, la tabla "client.client" almacena datos personales como género, educación, estado familiar, ingresos, entre otros. Las tablas "contact_flags", "heritage_flags", y "document_flags" guardan banderas indicativas sobre el contacto del cliente, la posesión de bienes y documentos presentados, respectivamente. Además, "client.application" y "client.previous_application" almacenan detalles de las aplicaciones actuales y anteriores de los clientes.

El esquema "credit" se enfoca en el historial de crédito y deudas de los clientes. Contiene tablas como "credit.credit_card_balance", "credit.pos_cash_balance", "credit.bureau", y "credit.bureau_balance" que almacenan información sobre saldos de tarjetas de crédito, balances de préstamos en efectivo, detalles de créditos en el Bureau, y saldos mensuales del Bureau, respectivamente.

Por último, el esquema "repayment" se centra en los pagos realizados por los clientes. La tabla "repayment.installment_payment" guarda información sobre los pagos de las cuotas, incluyendo detalles como el número de cuotas, montos de pagos, y días de pago.

Posterior a la creación de la base de datos, se diseñó y programó un migrador de datos para movilizar información. El migrador de datos diseñado para el grupo Home Credit es una herramienta integral desarrollada en Python con el propósito específico de transferir archivos CSV a una base de datos alojada en SQL Server. Este migrador se ha construido cuidadosamente utilizando las librerías pandas y pyodbc para la manipulación eficiente de datos y la conexión a la base de datos, respectivamente.

El objetivo principal es facilitar la migración de múltiples archivos CSV proporcionados por Home Credit hacia una estructura de base de datos en SQL Server, garantizando que los datos se integren correctamente en las tablas correspondientes.

El proceso de migración se ejecuta mediante métodos distintos dentro de la clase DataMigrator. Cada uno de estos métodos está específicamente diseñado para migrar datos desde un archivo CSV determinado hacia tablas concretas dentro de la base de datos SQL Server.

El flujo de trabajo inicia con la conexión al servidor SQL Server utilizando el método connect_to_sql_server, que emplea la cadena de conexión ODBC para establecer la conexión. Esto asegura que la herramienta esté lista para migrar los datos a la base de datos.

Cada método de migración está destinado a manejar un archivo CSV específico y se encarga de iterar sobre las filas del DataFrame generado a partir de dicho archivo. Cada fila se procesa individualmente, extrayendo los valores relevantes y ejecutando consultas SQL para insertar esos datos en las tablas correspondientes de la base de datos.

Además de la inserción de datos, se implementa un sistema de registro de eventos utilizando la librería logging. Este sistema registra información detallada sobre el progreso de la migración, lo que facilita el seguimiento y la resolución de problemas en caso de errores durante el proceso.

Luego de añadir los datos a la base de datos creada en SQL Server, se implementaron procedimientos almacenados y desencadenadores (triggers) para la base de datos "Home_Credit_Default_Risk". Los procedimientos almacenados son bloques de código SQL que se pueden llamar con parámetros para realizar operaciones específicas, como insertar datos en tablas particulares. Los desencadenadores, por otro lado, se ejecutan automáticamente cuando ocurren ciertos eventos, como la inserción de datos en una tabla.

Cada procedimiento almacenado creado, como Insert_Client, Insert_Contact_Flags, Insert_Application, entre otros, recibe parámetros correspondientes a columnas de tablas específicas y utiliza sentencias INSERT INTO para añadir datos a esas tablas.

Por ejemplo, Insert_Client inserta datos en la tabla client.client con los valores proporcionados para cada columna correspondiente. Se definen procedimientos similares

para otras tablas como contact_flags, heritage_flags, document_flags, application, etc., cada uno con sus respectivos campos y valores.

Los desencadenadores, como Trigger_Insert_Client, Trigger_Check_Contact_Flags, Trigger_Check_Application_Target, están diseñados para ejecutarse después de que se realice una inserción en las tablas respectivas. Su función principal es validar los datos insertados, asegurándose de que cumplan ciertos criterios. Por ejemplo, Trigger_Insert_Client verifica si el valor de género (gender) insertado en la tabla client.client es válido ('F', 'M' o ' '). En caso contrario, realiza un ROLLBACK, lo que revierte la inserción y evita que se guarden los datos incorrectos.

Los desencadenadores Trigger_Check_Contact_Flags, Trigger_Check_Heritage_Flags, Trigger_Check_Document_Flags y Trigger_Check_Bureau_Credit_Active realizan verificaciones similares para asegurar que los valores booleanos (0 o 1) o bits (0 o 1) sean los únicos aceptados en ciertas columnas específicas, deshaciendo la inserción si los valores no cumplen con esa condición.

Estas estructuras en SQL Server buscan mantener la integridad de los datos, asegurando que los valores ingresados cumplan con ciertos estándares predefinidos para evitar inconsistencias o errores en la base de datos.

Además de estos se incluyeron procesos dedicado a la gestión de usuarios y permisos en la base de datos destinada al almacenamiento de estos datos. Se crearon tres usuarios distintos con roles y privilegios específicos.

En primer lugar, se creó un usuario administrador, AdminUser, dotado con un rol de db_owner. Este rol confirió todos los permisos posibles sobre la base de datos "Home_Credit_Default_Risk", permitiendo operaciones amplias y modificaciones en la estructura de la base de datos.

El segundo usuario, NormalUser, fue creado con permisos más limitados y selectivos. A este usuario se le otorgó la capacidad de ejecutar procedimientos almacenados específicos como Insert_Client, Insert_Contact_Flags, entre otros, y se le concedió acceso de lectura (SELECT) a los esquemas client, credit y repayment. Esta restricción en sus permisos

aseguró que sus acciones estuvieran enfocadas y controladas, sin la posibilidad de realizar modificaciones extensas en los datos.

Por último, se creó un usuario dedicado a las tareas de respaldo, BackupUser. Este usuario recibió el permiso BACKUP DATABASE, lo que le habilitó para llevar a cabo copias de seguridad de la base de datos, asegurando así la disponibilidad de estos datos en caso de ser necesario.

Este enfoque en la gestión de usuarios y permisos dentro del proyecto garantizó un control detallado sobre las acciones permitidas a cada usuario, manteniendo la integridad y seguridad de la base de datos "Home_Credit_Default_Risk" durante el proceso de migración y más allá.

Además, se incorporó un sistema de auditoría para realizar un seguimiento detallado de las operaciones sobre las tablas clave de la base de datos. Dicho sistema de auditoría se pensaba implementar utilizando un servidor de auditoría propio de las herramientas accesibles en SQL Server, pero produjo problemas a la hora de escribir datos, por lo que se optó por otra alternativa.

El componente de auditoría se implementó mediante la creación de un esquema denominado "audit" y la tabla "auditLog". Esta tabla registró información vital, como el nombre de la tabla afectada, la acción realizada (inserción, actualización o eliminación), la fecha y hora de la operación, así como el nombre de usuario asociado a la acción.

Se establecieron diversos desencadenadores (triggers) para cada acción principal (INSERT, UPDATE, DELETE) en las tablas relevantes del esquema. Por ejemplo, se crearon triggers específicos para cada tabla clave como client.application, client.client, client.contact_flags, client.document_flags, client.heritage_flags, client.previous_application, credit.bureau, credit.bureau_balance, credit.credit_card_balance, credit.pos_cash_balance, y repayment.installment_payment.

Cada trigger se configuró para que, después de una acción específica (inserción, actualización o eliminación) en su tabla correspondiente, registre automáticamente los detalles relevantes en la tabla de auditoría audit.auditLog. Esto incluyó información

contextual como el nombre de la tabla afectada, el tipo de acción llevada a cabo, la fecha y hora de la operación, y el nombre de usuario responsable de la acción.

Aunado a lo anterior, se implementaron índices y consultas SARGABLES para mejorar el rendimiento de las consultas. Estos índices están diseñados para optimizar la velocidad de recuperación de datos y reducir la carga del sistema al ejecutar las consultas.

Se crearon varios índices en tablas relevantes para consultas específicas. Por ejemplo, se indexaron columnas que se utilizan frecuentemente en condiciones de búsqueda y filtrado en consultas clave. Además, se implementaron consultas SARGABLES, lo que significa que se realizaron de manera que puedan aprovechar eficientemente los índices existentes para acelerar la recuperación de datos.

Estas consultas están destinadas a extraer datos de manera óptima de la base de datos, utilizando índices específicos para las condiciones de búsqueda. Esta estrategia ayuda a mejorar el rendimiento de las consultas al minimizar el tiempo de búsqueda y optimizar la eficiencia del sistema al acceder a los datos necesarios de manera rápida y precisa.

Finalmente, se creó el dashboard empleando la base de datos resultante de la migración de datos como fundamento primordial. Este proceso incluye la creación de cinco vistas específicas, diseñadas estratégicamente para reunir datos esenciales que serán fundamentales en la implementación del dashboard. Cada vista se estructura con el propósito de proporcionar información relevante y específica para la visualización final.

La siguiente etapa se desarrolla en Power BI, estableciendo la conexión con la base de datos mencionada anteriormente. Las vistas previamente creadas se importan a la interfaz de Power BI para iniciar la configuración del entorno de trabajo. Es de suma importancia garantizar la correcta vinculación de las vistas con la base de datos para mantener la integridad de los datos presentados en el dashboard.

Una vez que las vistas están disponibles en Power BI, se procede a la creación de los elementos fundamentales que compondrán el dashboard. Estos elementos pueden adoptar diversas formas, como gráficos de pastel, de barras, entre otros. Cada uno se conecta directamente a una de las cinco vistas, estableciendo así la fuente de datos que nutrirá la visualización.

La personalización de cada elemento es crucial para asegurar la claridad y relevancia de la información presentada. Se seleccionan cuidadosamente los datos específicos para cada elemento, ajustando configuraciones como colores, títulos y líneas divisorias para mejorar la legibilidad y comprensión de la información.

La fase final comprende el refinamiento visual, donde se ajustan aspectos estéticos y funcionales de cada elemento. Los colores se adaptan para transmitir la información de manera efectiva, se añaden encabezados claros y concisos, y se incorporan elementos interactivos para permitir a los usuarios explorar y analizar los datos con mayor profundidad. Este proceso garantiza que el dashboard ofrezca una experiencia visual y analítica óptima.

Conclusiones

El proyecto ha alcanzado con éxito su objetivo general al desarrollar un sistema de bases de datos integral y eficiente en colaboración con Home Credit Group. La mejora en la experiencia crediticia para individuos con historiales crediticios insuficientes o inexistentes, ofreciendo préstamos con condiciones favorables, refleja el logro de este objetivo primordial.

La fusión efectiva entre la teoría y la práctica ha sido un componente fundamental en el desarrollo del proyecto. Desde la conceptualización y diseño de la base de datos hasta su implementación, se ha evidenciado una aplicación efectiva de conocimientos teóricos en un contexto real, permitiendo la resolución de desafíos concretos.

La colaboración estrecha con Home Credit Group ha sido esencial para comprender las necesidades reales del sector y aplicar soluciones efectivas. La utilización de datos alternativos para mejorar la predicción crediticia ha demostrado ser una estrategia eficaz y pertinente en el contexto actual.

La implementación de índices, consultas SARGABLES y la gestión de usuarios con roles específicos ha destacado un enfoque orientado a la mejora continua del sistema. Estas medidas han contribuido significativamente al rendimiento y la seguridad del proyecto.

La experiencia práctica obtenida por los estudiantes en la construcción completa de una base de datos, desde la migración de datos hasta la creación de procedimientos almacenados y desencadenadores, ha sido valiosa para su desarrollo profesional.

Recomendaciones

Se sugiere mantener una supervisión constante del sistema implementado para detectar posibles mejoras y evitar posibles problemas a medida que el sistema esté en funcionamiento.

Es recomendable considerar la actualización periódica de la base de datos para incluir nuevas fuentes de datos y expandir las funcionalidades del sistema, adaptándolo a las necesidades cambiantes del entorno.

Explorar la implementación de técnicas analíticas más avanzadas, como el machine learning, puede ser beneficioso para mejorar aún más la precisión en la predicción crediticia y potenciar las capacidades del sistema.

Se aconseja recopilar feedback de los usuarios finales del sistema, tanto de Home Credit Group como de los individuos beneficiarios de los préstamos, para ajustar y mejorar continuamente la experiencia ofrecida.

Finalmente, proporcionar capacitación y documentación adecuadas a los usuarios finales será fundamental para asegurar un uso óptimo y comprensión completa del sistema desarrollado.