Mejora del Almacenamiento de Certificaciones Cisco en la Universidad Compensar Una Solución para la Eficiencia y Cumplimiento.

Omar David Romero Velosa, Johan Jair Alexis García, Diego Fajardo, John Jairo Gómez, Diego Cucunubá, Omar David Sepulveda Romero

Especialización Big Data, Fundación Universitaria Compensar

Proyecto fin de grado Facultad de Ingeniería

Wilson Hernando Soto Urrea

23 de noviembre de 2023

Esta obra está bajo una Licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional.

Mejora del Almacenamiento de Certificaciones Cisco en la Fundación Universitaria Compensar Una Solución para la Eficiencia y Cumplimiento.

Omar David Romero Velosa, Diego Fajardo, Diego Cucunuba, Jhon Jairo Gómez, Johan Jair Alexis García, Omar David Sepulveda Romero

Trabajo de grado presentado como requisito parcial para optar al título de:

Especialista en Big Data

Director (a):

Ingeniero magister PH Wilson Hernando Soto Urrea

Fundación Universitaria Compensar

Facultad de Ingeniería, (Especialización en Big Data o Especialización en Seguridad Informática)

Bogotá, Colombia

Año 2023

Resumen

La propuesta de este proyecto contempla la automatización de la base de datos relacionada con las certificaciones Cisco de la Fundación Universitaria Compensar. Todo este proceso se llevará a cabo bajo la metodología CRISP-DM, aplicando las seis fases de la metodología. En cuanto a las herramientas que se utilizarán, se empleará un entorno en la nube para las bases de datos, respaldado por el uso de herramientas ETL para la minería de datos. Finalmente, los reportes se generarán con herramientas de BI. La ejecución de este proyecto tiene un tiempo estimado de ocho meses, teniendo en cuenta todas las bases correspondientes a las certificaciones Cisco que posee la Universidad Compensar. Esto se hace con el fin de generar informes en tiempo récord que permitan un mejor seguimiento de la información, facilitando así la presentación eficiente de los informes necesarios para la acreditación.

Palabras clave: Orquestador, lago de datos, Calidad de datos, Docker, Entorno de almacenamiento.

Abstract

The proposal of this project contemplates the automation of the database related to the Cisco certifications of the Compensar University Foundation. This entire process will be carried out under the CRISP-DM methodology, applying the six phases of the methodology. Regarding the tools that will be used, a cloud environment will be used for the databases, supported by the use of ETL tools for data mining. Finally, the reports will be generated with BI tools. The execution of this project has an estimated time of eight months, taking into account all the bases corresponding to the Cisco certifications held by Compensar University. This is done in order to generate reports in record time that allow better tracking of information, thus facilitating the efficient presentation of the reports necessary for accreditation.

Keywords: Orchestrator, Data lake, Data Quality, Docker, Storage environment

Contenido

PÁG.

Introducción

En el entorno académico de la Fundación Universitaria Compensar, la obtención de certificaciones Cisco por parte de los estudiantes se ha convertido en un aspecto crucial para su desarrollo educativo. A medida que el programa de certificaciones Cisco ha crecido desde su inicio en 2017, con un pequeño número de participantes, hasta alcanzar un promedio de 300 certificaciones por semestre, se ha generado un extenso repositorio de archivos en formato Excel. Estos archivos, aunque valiosos en términos de información, presentan un desafío considerable en cuanto a su organización y manejo efectivo.

La acumulación de una gran cantidad de datos distribuidos en archivos separados ha dificultado la búsqueda y recuperación precisa de la información. Cuando se requiere generar informes agrupados de estudiantes según diversos criterios, como año, semestre, carrera, entre otros, la complejidad se incrementa significativamente. Si esta situación persiste en el tiempo, la gestión de la información se tornará aún más problemática a medida que se acumulen más datos.

Este escenario plantea un riesgo importante, especialmente en lo que respecta a la acreditación de alta calidad de la fundación. Las certificaciones Cisco, siendo actividades extracurriculares fundamentales, pueden afectar de manera significativa el proceso de acreditación. El mantenimiento y la presentación efectiva de estos registros se convierten, por lo tanto, en una cuestión de gran relevancia para la institución. Dado que la pérdida de la acreditación de alta calidad es una posibilidad real si no se abordan adecuadamente los desafíos derivados de la gestión de estos datos. En este contexto, se hace necesario explorar en profundidad la metodología KDD (Knowledge Discovery in Databases) como un enfoque integral para abordar estos desafíos y garantizar la eficiencia en la gestión de las certificaciones Cisco en beneficio de la Universidad Compensar.

Antecedentes y Justificación

De acuerdo con el documento "Aplicación de Metodología CRISP-DM para Segmentación Geográfica de una Base de Datos Pública" de Javier Jesús Espinosa-Zúñiga, que utiliza el análisis de datos y la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), se pueden almacenar grandes cantidades de datos, analizarlos y extraer información de los mismos.

Según el artículo "Análisis Bibliométrico de la Producción Científica de las Universidades del Suroriente Peruano en la Base de Datos SCOPUS" de Edwin Gustavo Estrada-Araoz y Percy Samuel Yabar-Miranda, se presenta el análisis, filtrado y caracterización de las producciones científicas de las universidades según la base de datos Scopus.

En el documento "Análisis de Deserción-Permanencia de Estudiantes Universitarios Utilizando Técnica de Clasificación en Minería de Datos" de Karina B. Eckert y Roberto Suénaga, se aplican técnicas de minería de datos para la selección y depuración de datos, utilizando diferentes criterios de representación y aplicando algoritmos de redes bayesianas para caracterizar las variables que afectan la deserción de los estudiantes de la Universidad Gastón Dachary de Argentina.

En el presente proyecto, nos enfocamos en analizar y presentar la base de datos de las certificaciones de CISCO realizadas por la Fundación Universitaria Compensar. Debido a la gran cantidad de información que se tiene y se tendrá a lo largo de los semestres, este trabajo nos permitirá utilizar las bases de datos diseñadas para el análisis de datos y aplicaciones como Power BI, que se encarga de visualizar los datos de manera más detallada según las necesidades del usuario. Además, permitirá tener acceso a los datos de manera más rápida y eficiente, ya sea por fecha o si se realizó la certificación.

Formulación del Problema

¿Cómo mejorar la eficiencia, seguridad y accesibilidad del almacenamiento de certificaciones Cisco en la Universidad Compensar?

El sistema actual de almacenamiento de certificaciones Cisco en la Universidad Compensar es manual y obsoleto, lo que ocasiona varios problemas: Ineficiencia en el proceso de almacenamiento y gestión de certificaciones, inseguridad al exponerlas a pérdidas, daños o falsificaciones, y dificultad de acceso para estudiantes y egresados al consultar sus certificaciones.

Objetivo General

Optimizar la gestión de las certificaciones Cisco obtenidas por los estudiantes de la Universidad Compensar mediante la implementación de motores de bases de datos avanzados, respaldados por técnicas de análisis de series temporales. Esta iniciativa busca agilizar la generación de informes mediante herramientas de Business Intelligence (BI), permitiendo obtener reportes precisos y oportunos durante el proceso de acreditación, garantizando así la disponibilidad ágil y confiable de la información requerida.

Objetivos Específicos

Realizar un análisis exploratorio de los distintos repositorios donde se almacena la información, identificando posibles duplicados, inconsistencias y problemas de formato.

Generar un sistema de transformación que permita depurar la información extraída de los distintos archivos en Excel, asegurando la integridad y coherencia de los datos.

Construir un sistema de informes en un entorno propicio para el procesamiento y presentación eficiente de la información, integrando herramientas de BI.

Alcances y Limitaciones

Alcance: El proyecto se enfoca en la creación de un entorno para el diseño de una base de datos MySQL destinada al almacenamiento de datos relacionados con las certificaciones Cisco. Posteriormente, se establecerá una conexión con la herramienta Power BI para generar informes automáticos que serán presentados durante el proceso de acreditación Cisco. El tiempo de ejecución se estima en un plazo de 8 meses, considerando la cantidad de archivos en Excel manejados en el proyecto. Finalmente, se abordarán los siguientes componentes:

Análisis de los Datasets: Revisión y análisis de los conjuntos de datos correspondientes a las certificaciones Cisco desde el año 2017.

Estructuración de Entorno en la Nube (iCloud) para el Diseño de la Base de Datos: Configuración de un entorno en la nube para facilitar el diseño y la gestión eficiente de la base de datos MySQL.

Estructuración de Canales de Procesamiento mediante ETL: Desarrollo de canales de procesamiento utilizando técnicas ETL (Extract, Transform, Load) para garantizar la calidad y coherencia de la información.

Conexión a Herramienta BI para Automatización de Reportes: Establecimiento de conexión entre la base de datos y la herramienta Power BI para la generación automatizada de informes.

Entregables:

- Documentación detallada de la estructura de los conjuntos de datos.
- Entorno funcional de la base de datos en la nube con datos actualizados.
- Proceso ETL implementado y funcionando correctamente.

- Conexión establecida entre la base de datos y las herramientas de Business Intelligence (BI).
- Informes y paneles de control generados automáticamente y listos para su presentación.

Limitaciones: Descentralización en la gestión de la información, ya que al manejar distintos archivos en Excel, el flujo de datos proviene de múltiples fuentes con una misma estructura. Además, no se cuenta con un entorno especializado previamente diseñado para el almacenamiento de la información. Finalmente, debido al margen de tiempo disponible para la realización de este proyecto, este debe ser preciso.

Marco teórico

1.KDD: "Descubrimiento de conocimiento en bases de datos"

Digitales, S. (2021, abril 29). Breve explicación del proceso KDD: "Es un proceso utilizado para llevar a cabo la extracción automatizada de conocimiento partiendo de grandes volúmenes de datos, el cual es de naturaleza iterativa, por lo tanto, es aplicable tantas veces como sea necesario hasta obtener la información necesaria.

El proceso KDD tiene como motivación la detección de información que permita resolver los problemas o necesidades que surgen en las empresas y es a menudo solicitado por directivos y/o stakeholders. El proceso KDD consta de las siguientes etapas:

• Recopilación de datos

• Selección, limpieza y transformación

• Minería de datos

• Interpretación y evaluación de los modelos obtenidos"

1.1Minería de datos: (S/f-b). Amazon.com. Recuperado el 24 de noviembre de 2023 "La minería de datos es un componente esencial del análisis de datos y una disciplina central en la ciencia de datos. Se enfoca en la extracción de información valiosa de los datos. A nivel más detallado, la minería de datos es una etapa dentro del proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD), una metodología en la ciencia de datos que implica la recopilación, el procesamiento y el análisis de datos".

1.2 Knime Herramienta ETL: ¿Qué es Knime? (2022, diciembre 18). LIS Data Solutions " KNIME pertenece a una nueva generación de herramientas dominadas como Plataformas de Data Science y Machine Learning por Gartner. Estas herramientas permiten a científicos de datos expertos, analistas o usuarios de negocio interactuar con sus datos y crear, desplegar y gestionar sus modelos de analítica avanzada. Las herramientas integran las funcionalidades principales para realizar proyectos de minería de datos: importación de datos, preparación de datos, exploración de datos, modelado, evaluación y despliegue".

ETL (Extracción, transformación y carga).
(S/f-b). Amazon.com. Recuperado el 24 de noviembre de 2023, "Es un proceso donde consiste en combinar datos de diferentes origines un gran repositorio central llamado almacenamiento de datos. Esta herramienta mejora la inteligencia comercial y el análisis al hacer que el proceso sea más fiable, preciso, detallado y eficiente. "

1.3 MYSQL
¿Qué Es MySQL? Una Explicación para Principiantes. (2019, mayo 2) " MySQL es un potente motor de base de datos que ofrece una amplia gama de características para que la gestión de sus datos sea fácil y eficiente como una alternativa a otros motores de bases de datos como SQL Server de Microsoft. Mysql es de código abierto, software libre escalable, rápido, fiable y seguro. "

1.4 Power BI

Urrutia, D. (2020, enero 28). *Qué es Power Bi* " Es una herramienta de **Microsoft** con Inteligencia artificial que crea visualizaciones de datos interactivas en poco tiempo para empresas y a través de paneles. Con Power BI se tiene de manera fácil acceso a datos dentro y fuera de la organización. "

Diseño Metodológico

El uso de la metodología KDD en el proyecto de optimización del almacenamiento de certificaciones Cisco en la Fundación Universitaria Compensar se justifica por las siguientes razones:

Los datos históricos de almacenamiento, que cuentan aproximadamente con 5400 registros desde 2017 en la base de estudiantes y 3849 registros en la base consolidada, representan una fuente valiosa de información para identificar los problemas del actual sistema. La minería de datos puede ayudar a identificar patrones y tendencias en estos datos que no serían visibles a simple vista mediante un análisis temporal o detección de anomalías.

Fase 1: Selección de datos

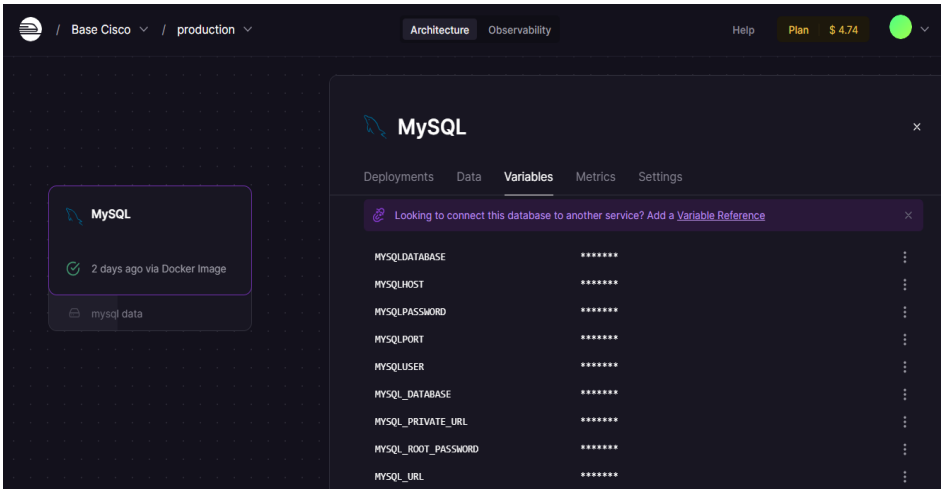
En esta fase se recopilarán los datos necesarios para la minería de datos. Estos datos pueden incluir: Datos históricos de almacenamiento de certificaciones desde el año 2017, datos de encuestas a los estudiantes y egresados, y datos de la revisión documental.

Dentro de esta etapa, ya se identificaron los archivos y se cuenta con un consolidado de la información desde 2017 correspondiente a la base de Certificaciones Cisco. En cuanto a la base estudiantil, esta se encuentra fragmentada, razón por la cual se generó un repositorio para el cargue de todos los archivos en Excel hasta la fecha.

Fase 2: Pre procesamiento de Datos

Para esta fase, se tendrá en cuenta el uso de dos herramientas, como son MySQL como motor de base de datos y Railway. En esta etapa, se prepararán los datos para su análisis. Para ello, se realizarán las siguientes actividades: construcción del entorno MySQL y carga inicial de los datos al lago de datos.

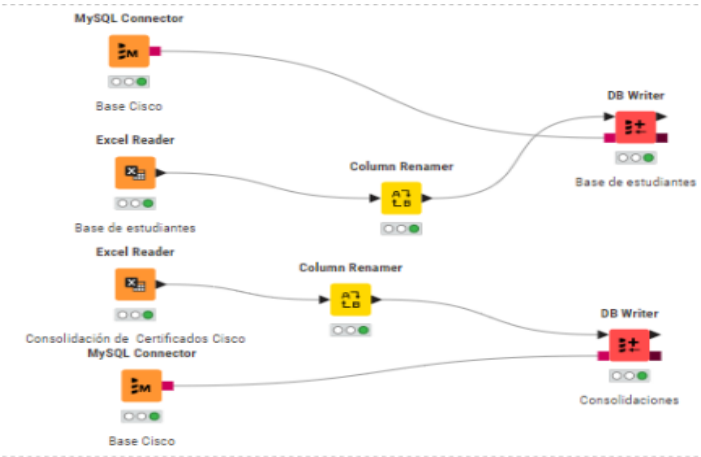
En cuanto a esta etapa, se llevó a cabo la construcción del entorno de datos, además del ingreso de los datos, validando las estructuras planteadas en base al diccionario de datos.



Fase 3: Transformación de datos.

Para la transformación de los datos, se utilizará la herramienta conocida como Knime. Este ETL permitirá procesar los datos para, posteriormente, depositarlos en un entorno de almacenamiento.

En relación al desarrollo de esta etapa, se logró establecer una conexión sólida entre los archivos en Excel y la base de datos. Sin embargo, aún queda pendiente la validación en cuanto a la limpieza de datos.



Fase 4: Minería de datos

En esta etapa, se emplearán técnicas de minería de datos para identificar los problemas del actual sistema de almacenamiento. Esto se llevará a cabo mediante la herramienta Power BI, lo que nos permitirá evaluar alternativas de optimización. Las técnicas de minería de datos que se pueden utilizar incluyen análisis de patrones, análisis de asociación, análisis de línea de tiempo, análisis de clasificación y análisis de clusterización.

Fase 5: Evaluación

Finalmente, en esta fase se evaluarán las alternativas de optimización identificadas en la fase anterior. Para ello, se utilizarán los siguientes criterios: eficiencia, seguridad, accesibilidad y costo.

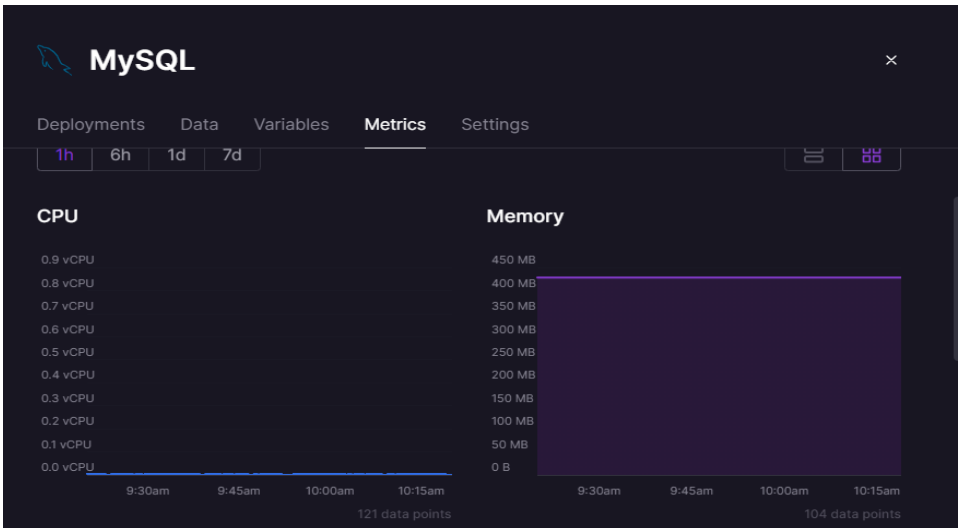
Resultados esperados

Documentación detallada de la estructura de los conjuntos de datos. Esta documentación permitirá comprender mejor la información almacenada en la base de datos, facilitando su uso y análisis. En el progreso actual, se ha estructurado un diccionario de datos que comprende los conjuntos de datos utilizados por la universidad para el control de las certificaciones Cisco.

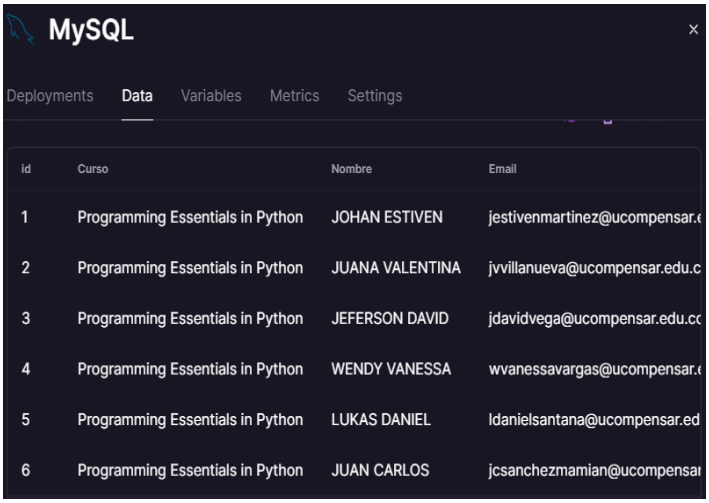
Base de estudiantes			
Nombre del Campo	**Descripción**	**Tipo de dato**	**Tipo de llave**
CODIGO	Codigo_unico_de_estudiante	Varchar(25)	PK
APELLIDOS	Apellido de estudiante	Varchar(100)	
NOMBRES	Nombre de estudiante	Varchar(100)	
CURSO	Codigo_Curso	Varchar(150)	FK
HORAS	Numero de horas	Int(10)	
Hijos	Numero de hijos	Int(10)	
Genero	Genero del estudiante	Varchar(100)	
Salario	Salario del estudiante	Int(10)	
GRUPO	Codigo concatenado del grupo	Varchar(150)	

Consolidado Certificaciones Cisco			
Nombre del campo	**Descripción**	**Tipo de dato**	**Tipo de llave**
Curso	Descripción del curso	Varchar(25)	PK
Nombre	Nombre del estudiante	Varchar(100)	
Apellido	Apellido del estudiante	Varchar(100)	
Email	Correo institucional	Varchar(180)	
Semestre	Numero de semestre	Date	
Cumulative Grade	Calificación acumulada	Int(10)	
Certificacion	Obtuvo la certificación si o no	Varchar(5)	

Desarrollo de un entorno funcional para el diseño de la base de datos. Esto se llevará a cabo a través de un entorno en la nube, como Railway, con el objetivo de crear una base de datos que sea capaz de adaptarse a las necesidades del proyecto. Inicialmente, se espera que pueda manejar hasta 700,000 datos. En cuanto al desarrollo generado, observamos que el límite de datos es el esperado y que el consumo de CPU es mínimo, gracias a los conjuntos de datos utilizados en la implementación.



En relación a los resultados obtenidos en la conexión generada mediante el ETL Knime, observamos un tiempo de carga promedio de 10 minutos, ligeramente superior al tiempo esperado. Sin embargo, es importante señalar que este tiempo puede variar según la máquina virtual en la cual se despliegue finalmente. Además, en cuanto a la compatibilidad de la información, basada en el tipo de cada dato, fue óptima. La única novedad presentada en este flujo es el cambio de nombre de las columnas para que la migración pueda realizarse de manera normal.



En relación a la conexión con Power BI, se espera que los tiempos de sincronización sean cada 10 minutos, permitiendo una actualización previa de la base. Además, se espera que los indicadores sean compatibles con las plantillas establecidas para la generación de informes. En caso de no serlo, será necesario validar una posible reestructuración.

Conclusiones y recomendaciones

Conclusiones

En lo que respecta a la adopción de la metodología CRISP-DM, esta resulta de gran utilidad para optimizar los procesos vinculados con la gestión de las certificaciones Cisco. Esto se debe a que permite una comprensión detallada y una preparación meticulosa de los datos, facilitando la selección y evaluación de modelos que aporten posibles soluciones en pro de la innovación y la adaptabilidad, elementos esenciales para alcanzar la acreditación de alta calidad.

Con respecto a la ejecución general del proyecto, podemos determinar que la estructura planteada ha sido una de las más óptimas, considerando tanto los costos como las herramientas actualmente disponibles. Validamos que en un solo proyecto del ETL podemos implementar todas las conexiones a las fuentes de información sin afectar el rendimiento. También consideramos la opción de generar un consolidado de todas las certificaciones, lo que nos permitiría aplicar una única migración a la base de datos. En relación al entorno, concluimos que el consumo de CPU y el espacio necesario para el almacenamiento de la información no son elevados, por lo cual establecemos un límite inicial de 10 GB para el consumo de CPU. Este es menor al 5% debido a la inactividad de la base de datos, y nos queda pendiente determinar estas métricas cuando ya se esté generando un consumo por Power BI. En cuanto a la sincronización, se estableció un tiempo de 30 minutos con el fin de disminuir la carga operativa sobre la base de datos.

Recomendaciones

Para el desarrollo de este proyecto y con base en los resultados obtenidos en esta primera etapa, se generan las siguientes recomendaciones:

- Desarrollar un plan de mantenimiento para la base de datos. Esto es crucial para garantizar que los datos sean actualizados y precisos.

- Capacitar al personal en el uso de las herramientas y tecnologías del proyecto con el fin de asegurar un uso efectivo de las mismas.

- Generar un canal de comunicación directa con las partes interesadas. Es importante comunicar los resultados del proyecto a las partes interesadas para que puedan beneficiarse de los mismos.

- Estructurar un archivo batch con el fin de automatizar la ejecución de los modelos ETL.

- Establecer repositorios para almacenar el uso de información no estructurada.

Bibliografía

Blanco, J. M. (2021, septiembre 20). *Power BI: Qué es y para qué sirve esta herramienta de análisis de datos*. Plain Concepts. https://www.plainconcepts.com/es/power-bi-que-es/

Digitales, S. (2021, abril 29). *Breve explicación del proceso KDD*. Laboratorio de Certificación. https://www.laboratoriodecertificacion.es/breve-explicacion-del-proceso-kdd/

Mauricio. (2022, septiembre 18). *Características de Mysql*. Tutoriales Dongee. https://www.dongee.com/tutoriales/caracteristicas-de-mysql/

¿Qué Es MySQL? Una Explicación para Principiantes. (2019, mayo 2). Kinsta®; Kinsta. https://kinsta.com/es/base-de-conocimiento/que-es-mysql/

Urrutia, D. (2020, enero 28). *Qué es Power Bi - Definición, significado y ejemplos*. Arimetrics. https://www.arimetrics.com/glosario-digital/power-bi

(S/f). Amazon.com. Recuperado el 23 de noviembre de 2023, de https://aws.amazon.com/es/whatis/etl/#:~:text=Las%20herramientas%20ETL%20automatizan%20el,como%20mover%20y%20formatear%20datos.

¿Qué es Knime? (2022, diciembre 18). LIS Data Solutions. https://www.lisdatasolutions.com/es/que-es-knime/