CASO PRÁCTICO 104 LA CIENCIA DE DATOS. TÉCNICAS DE ANÁLISIS. MINERÍA Y VISUALIZACIÓN

LA CIENCIA DE DATOS. TÉCNICAS DE ANÁLISIS. MINERÍA Y VISUALIZACIÓN

Fernando Aleisy González
9 de mayo de 2024

Índice

1. SOFTWARE, LENGUAJES Y LIBRERÍAS PARA EL TRATAMIENTOS DE LOS DATOS

Para el manejo de las bases de se utiliza motores como SQL para las bases de datos relacionados, en donde la información se encuentra distribuidas en tablas que a su vez se relacionan por medio de claves primarias y foráneas.

1.2. Lenguajes para el tratamiento de los datos

Los lenguajes más utilizados para bases de datos estructurados son SQL y SQLite, Tabla 1, Donde, este último es una versión ligera de SQL, permitiendo su uso en dispositivos de menos capacidad de memoria como los celualares inteligentes.

Tabla 1: Características y sugerencias para elegir cada lenguaje que se puede utilizar en las técnicas de análisis, minería y visualización de los datos

Herramienta	Características	Por qué elegirlo
SQL	SQL (Structured Query Language) es el lenguaje estándar ANSI/ISO de definición, manipulación y control de bases de datos relacionales.	Es un lenguaje declarativo: sólo hay que indicar qué se quiere hacer. SQL es un lenguaje muy parecido al lenguaje natural.
SQLite	SQLite es un sistema de gestión de bases de datos relacional, famoso por su pequeño tamaño comparado con SQL.	Por su pequeño tamaño se puede usar en Smartphones ya sea en Android o iOS. Al ser rápido y ligero se ejecuta en muchas plataformas. Es de dominio público y por tanto sin costo

1.2. Módulos para el tratamiento de los datos

Los lenguajes más populares para la ciencia de datos y por lo tanto para el análisis, minería y visualización de los mismo, son R y Python. Estos lenguajes tiene una comunidad activa que constantemente se encuentran desarrollando módulos que expanden sus utilidades. En el caso del análisis, las minería y visualización de los datos, se han desarrollado, entre otros, los módulos que se encuentran en la Tabla 2

Tabla 2: Características y sugerencias para elegir cada módulo o dependencia, de los lenguajes, que se puede utilizar en las técnicas de análisis, minería y visualización de los datos

Herramienta	Características	Por qué elegirlo
Pandas	pandas es una herramienta de análisis y manipulación de datos de código abierto rápida, potente, flexible y fácil de usar, construida sobre el lenguaje de programación Python.	Lectura y escritura de datos (CSV, Excel, SQL, etc.). Hace más amigable el uso de Numpy. Facilita el manejo de series temporales
Data.table	El package data.table es un paquete que lleva la eficiencia al siguiente nivel. Como dije, la sintaxis es algo menos intuitiva que el lenguaje tidyverse.	Los Data Table pueden ser utilizados como data.frame y las librerías que solo usan data.frame no tendrían problemas al usar data.table.
RSQlite	RSQLite incorpora el motor de base de datos SQLite en R, lo que facilita aún más la gestión de permisos de las bases de datos.	Permite unir la versatilidad entre SQLite y la capacidad del análisis de datos de R.
RJDBC	Este paquete de R proporciona acceso a las bases de datos mediante el método Interfaz JDBC (Java DataBase Connectivity) el cual, a su vez, permite comunicar una aplicación en Java (ya sea standalone, GUI, Servlet o JSP) con una base de datos (Postgres, MySQL, SQL Lite, entre otras)	Proporciona un acceso uniforme a una gran variedad de bases de datos relacionales. Proporciona una base común para la construcción de herramientas y utilidades de alto nivel.
pyODBC	pyodbc es un módulo de Python de código abierto que simplifica el acceso a las bases de datos ODBC (Open DataBase Connectivity), el cual, es un estándar de acceso a las bases de datos desarrollado por SQL Access Group (SAG) en 1992.	Sirve para tener la integración de la comunicación con bases de datos de una manera sencilla.

1.3. Herramientas para el tratamiento de los datos

Dentro de las herramientas para el consulta, análisis, minería y visualización de los datos, en la Tabla 3 tenemos, en primer lugar, a los archivos json, los cuales se utilizan para compartir la información en un archivo de texto plano (como los archivos csv), lo que descarta la posibilidad de transmitir comandos maliciosos como si podría suceder con archivos más complejos como los libros, con macros, de Excel. Toad, RapidMiner, Kminer y Pentaho son herramientas que facilitan la minería y el análisis de los datos y/o el desarollo de modelos matemáticos como los de regresión.

Tabla 3: Características y sugerencias para elegir cada herramienta que se puede utilizar en las técnicas de análisis, minería y visualización de los datos

Herramienta	Características	Por qué elegirlo
JSON	JSON es la sigla de JavaScript Object Notation. Es un estándar para enviar y recibir datos entre un servidor y el navegador.	Es sencillo de leer por las personas y fácil de manejar en distintoa ámbitos. Se puede trabajar con lenguajes de manipulación de datos. Para editarlos solo se requiere un bloc de notas.
Toad	Toad de Quest es un conjunto de herramientas de bases de datos que se utiliza para simplificar flujos de trabajo, crear códigos de alta calidad sin defectos, automatizar procesos frecuentes o repetitivos y minimizar riesgos.	Toad permite que los equipos se centren en iniciativas de mayor valor estratégico y permitan avanzar a la empresa en la economía impulsada por los datos de la actualidad.
RapidMiner	RapidMiner, una aplicación de software libre, con una interfaz de usuario muy sencilla y que ofrece muchas ventajas con respecto a otras herramientas.	Su sistema de programación visual (Drag&Drop) requiere de una menor curva de aprendizaje logrando mayor productividad en menos tiempo. RapidMiner es una plataforma de análisis que permite acelerar la creación, entrega y mantenimiento de analíticas predictivas de alto valor.
Kminer	KNIME (Konstanz Information Miner) 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.	Permite el desarrollo de un proyecto analítico completo, siendo posible desarrollar cualquier fase del proyecto, desde ingestas y transformaciones, hasta modelos analíticos, predicciones y visualizaciones.
Pentaho	Pentaho es una plataforma de Business Intelligence (BI) orientada a la solución y centrada en procesos que incluye los componentes requeridos para implementar soluciones basadas en procesos como minería de datos, ETL y generación de informes.	Genera informes programáticos sobre la base de un archivo de definición XML. Proporcionar información sobre los datos, donde se pueden ver informes, gráficos, etc. Usa estrategias de aprendizaje de máquina, automático y minería de datos. Facilita el acceso a grandes volúmenes de datos.

2. EJEMPLOS DE USOS DE BIG DATA

A continuación, en la Tabla 4 se muestra dos ejemplos, de éxito, de aplicación del Big Data, Avis Budget Group y Bristol Myers Squibb. La primera es una empresa que optimiza los servicios de alquiler de vehículos con ayuda de los datos en tiempo real. Bristol Myers Squibb es una biofarmacéutica global que está aprovechando los datos que han almacenado durante décadas para descubrir, desarrollar y ofrecer medicamentos innovadores que ayudan a los pacientes a superar enfermedades graves.

Tabla 4: Ejemplos de projectos de aplicación de Big Data

Información	Avis Budget Group	Bristol Myers Squibb
Objetivos	Conectar una enorme flota de 650 000 vehículos en tiempo real y con total visibilidad global para mejorar la eficiencia, reducir los costes y aumentar los ingresos. Reducir el riesgo empresarial mediante el perfilado y el gobierno de los datos telemáticos procedentes de los sistemas GPS y de navegación del vehículo y detectar los problemas de calidad de datos en fases muy tempranas. Documentar los activos principales, como los datos de flotas y telemáticos y capturar, al mismo tiempo, contexto de negocio de expertos en la materia.	Llevar una terapia exitosa y segura a los pacientes más rápido. Mejorar la atención al paciente mediante una visión amplia de lo que está sucediendo en la enfermedad y cómo se maneja y descubrir cómo mejorar el viaje del paciente. Agregar valor y ser un participante activo en la atención médica de una manera impactante y adecuada
Soluciones	Implantar soluciones de Informática en AWS para poner en funcionamiento los datos y realizar análisis en tiempo real como parte de una plataforma de próxima generación. Aprovechar informática Big Data Management para agilizar, flexibilizar y poder repetir los procesos de ingestión e incorporación de Big Data. Organizar los datos de flotas y telemáticos mediante el plugin informática Enterprise Data Catalog para proporcionar visibilidad de la ubicación, el linaje y el contexto de negocio de los datos.	Simulaciones de ensayos clínicos mediante la implementación de un entorno de red alojado internamente en los sistemas cloud de AWS (Amazon Web Services). La compañía está aprovechando los aprendizajes pasados en combinación con conjuntos de datos novedosos y sólidos de fuentes internas y externas para amplificar el poder predictivo en todas las etapas de la línea de investigación y desarrollo (I&D) Todo el trabajo es verdaderamente multifuncional, y todos aportan su experiencia en la materia
Resultados	Uso del análisis global de vehículos con procesos de datos de extremo a extremo, lo que permite a los gestores de flotas acceder más rápidamente a los sistemas de seguimiento de vehículos en tiempo real. Mitiga el riesgo mediante la mejora de la calidad y el gobierno de datos, lo que ayuda a garantizar que los datos de la flota estén completos y en el formato adecuado. Aumenta la productividad, puesto que permite a los usuarios de negocio buscar, localizar y comprender los activos de datos por sí mismos, con una línea de visión en el linaje de datos.	Hacer llegar las terapias a los pacientes con mayor rapidez. Producir una visión de 360 grados de enfermedades específicas y tipos de pacientes específicos y para responder a preguntas clave desde la investigación temprana hasta la disponibilidad del tratamiento.
Fuente	https://www.avisbudgetgroup.com/	https://www.bms.com/es

REFERENCIAS

Aguirre, S. (2020). *JSON-vol. 1: Primeros pasos-sintaxis-tipos de datos* (Vol. 1). RedUsers.

Castro Quintero, N. J. (2017). *Introducción a data science con python.*

Cortés Domínguez, E. (2017). *Sistema de consulta para la búsqueda y visualización de auditorias en estructuras jerárquicas de datos.*

Escofet, C. M. (2002). *El lenguaje SQL.* UOC, la universidad virtual.

García Bermejo, P. (2022). *Data science y KNIME, combinación perfecta para el éxito en la toma de decisiones.*

Hernández Baez, I. (n.d.). *K-medias en RapidMiner.*

Sarmiento Ponce, H. (2018). *Inteligencia de negocios usando pentaho para la gestión académica en la UNAMBA-2016.*

Torres, S. L. (2021). Componente de revisión de estándar de arquitectura de datos para el gestor de bases de datos SQLite. *Innovación y Software*, *2*(1), 20–32.

Vera Briceño, J. J. (2023). *Base de datos unificada de datos experimentales geomecánicos para ecoage web.*