Hackaton Talento TECH proveedores para Empresarios B2B CCB presencial

Nivel de Implementación: Nivel Básico

B2B Scout

Santiago Palomar Acosta - Líder del proyecto
Albi Steed Sanchez Herrera - Especialista en Scraping de Datos
David Nicolas Charris Chacón - Administrador de Base de Datos
Maryori Henith Santos Bohorquez - Desarrolladora de Visualización

Inicio: 9/11/2024
Entrega: 16/11/2024

Este proyecto desarrolla un sistema de extracción, almacenamiento y visualización de datos para empresas B2B. Mediante técnicas de scraping, se recopila información relevante sobre empresas proveedoras, la cual se almacena en PostgreSQL y se presenta en una interfaz visual con Streamlit.

Introducción

En un mundo cada vez más digital, el acceso y manejo eficiente de la información es clave para las empresas que buscan mantenerse competitivas. Los datos sobre proveedores de productos y servicios, especialmente en el sector B2B (Business to Business), son fundamentales para facilitar relaciones comerciales, optimizar procesos y tomar decisiones informadas.

Este proyecto, desarrollado en el marco de la ***"Hackaton Talento TECH proveedores para Empresarios B2B CCB presencial"***, tiene como objetivo crear una solución que permita extraer, almacenar y visualizar información relevante de empresas proveedoras B2B. Mediante técnicas de scraping de datos, almacenamiento en una base de datos PostgreSQL y visualización interactiva con Streamlit, el sistema proporcionará un recurso centralizado y accesible para la consulta de datos.

A lo largo de esta documentación, se describen las fases de desarrollo, los métodos utilizados, y los resultados obtenidos, destacando el trabajo colaborativo y las herramientas tecnológicas empleadas para alcanzar los objetivos propuestos.

Objetivos

Objetivo General

Desarrollar una solución integral que permita extraer, almacenar y visualizar información relevante de empresas B2B mediante técnicas de scraping de datos, almacenamiento en PostgreSQL y visualización interactiva en Streamlit, proporcionando así un recurso accesible y confiable para la consulta y análisis de información comercial.

Objetivos Específicos

1. Implementar un script de scraping que extraiga información de empresas B2B desde plataformas en línea, recolectando datos como nombre, ubicación, servicios ofrecidos y contacto.
2. Diseñar y configurar una base de datos en PostgreSQL para almacenar de manera estructurada la información extraída, facilitando su acceso y gestión.
3. Desarrollar una interfaz de visualización en Streamlight que permita a los usuarios consultar y explorar los datos almacenados mediante gráficos y listados interactivos.
4. Documentar el proceso de desarrollo, incluyendo la metodología, roles y responsabilidades, y resultados obtenidos, para garantizar la replicabilidad y comprensión del proyecto.

Metodología

Para el desarrollo de este proyecto se ha seguido una metodología ágil, permitiendo iterar sobre cada fase del proceso y realizar ajustes en función de los avances y desafíos encontrados. La metodología comprende las siguientes etapas:

1. **Planificación y análisis de requisitos:** Identificación de las necesidades del proyecto y definición de los datos clave a extraer, almacenar y visualizar.
2. **Scraping de datos:** Implementación de un script de scraping para extraer la información de fuentes en línea de empresas B2B. Esta fase incluye la

selección de las plataformas de las cuales se extraerán los datos y la validación de su disponibilidad.

3. **Diseño y configuración de la base de datos:** Creación de una base de datos en PostgreSQL para almacenar la información de manera organizada y segura.
4. **Desarrollo de la interfaz de visualización:** Construcción de una aplicación en Streamlit que permita a los usuarios interactuar con los datos mediante gráficos y visualizaciones.
5. **Documentación:** Registro detallado de cada fase del proyecto, incluyendo los roles, las responsabilidades y los resultados obtenidos.

Esta metodología permite un desarrollo estructurado y adaptativo, maximizando la eficiencia y la colaboración del equipo.

Roles y Responsabilidades

El equipo de trabajo está compuesto por cuatro integrantes, cada uno con un rol específico que contribuye a alcanzar los objetivos del proyecto de manera eficaz:

- **Santiago Palomar Acosta - Líder del proyecto:** Responsable de la coordinación general del equipo, asignación de tareas y seguimiento del cronograma para asegurar el cumplimiento de los plazos y objetivos del proyecto.
- **Albi Steed Sanchez Herrera - Especialista en Scraping de Datos:** Encargado de desarrollar el script de scraping para extraer información de las plataformas en línea. También es responsable de optimizar el código para garantizar una extracción de datos eficiente y precisa.
- **David Nicolas Charris Chacón - Administrador de Base de Datos:** Responsable del diseño y la configuración de la base de datos PostgreSQL. Administra el almacenamiento de datos, asegurando su integridad y organización.
- **Maryori Henith Santos Bohorquez - Desarrolladora de Visualización:** Encargada de crear la interfaz visualización en Streamlit. Diseña y desarrolla

gráficos y la presentación de los datos para asegurar una experiencia de usuario intuitiva y funcional.

Cada integrante aporta sus conocimientos específicos para cumplir con los objetivos y garantizar una implementación efectiva del proyecto.

Descripción de la Solución

Scraping de Datos

Objetivo

El propósito del scraping es recopilar datos de proveedores B2B y sus productos, incluyendo nombre del proveedor, categoría, descripción del producto, precios, ubicaciones y enlaces de contacto. Esta información será utilizada para analizar la oferta del mercado y facilitar conexiones comerciales.

Herramientas

Scrapy

- Razón de elección: Framework ligero y eficiente para scraping a gran escala, con soporte integrado para estructuras jerárquicas y pipelines para limpiar los datos.
- Uso: Navegar por las páginas HTML estáticas y extraer la información estructurada.

Playwright

- Razón de elección: Ideal para interactuar con sitios dinámicos o protegidos por JavaScript. Permite emular navegadores y manejar sesiones complejas.
- Uso: Extraer datos de sitios con contenido dinámico.

Pandas

- Razón de elección: Biblioteca potente para limpiar, transformar y analizar los datos.

- Uso: Procesar y estructurar los datos obtenidos antes de almacenarlos en un archivo CSV o una base de datos.

Desafíos y Soluciones

Bloqueos de IP

- Problema: Los portales bloqueaban nuestras solicitudes debido a la cantidad de accesos.
- Solución: Implementamos rotación de proxies y configuración de tiempos aleatorios entre solicitudes.

Detección de Bots

- Problema: Algunos portales usan captchas o detección avanzada de bots.
- Solución: Usamos Playwright para emular la interacción de un usuario real, incluyendo desplazamiento y clics.

Estructuras de Datos Inconsistentes

- Problema: Los datos en diferentes portales tenían formatos variados.
- Solución: Establecimos reglas de limpieza y estandarización con Pandas.

Almacenamiento de Datos

Objetivo del Almacenamiento de Datos

El objetivo principal de almacenar datos en una base de datos es garantizar la disponibilidad, integridad y accesibilidad de la información obtenida durante el proceso de scraping.
Esto permite:

- **Confiabilidad**: Centralizar los datos recolectados para su posterior análisis y visualización.
- **Escalabilidad**: Facilitar el manejo de grandes volúmenes de datos.

- **Integridad**: Mantener un control estructurado que garantice que los datos sean consistentes y útiles para las partes interesadas.

La base de datos servirá como repositorio para información como nombres de empresas, ubicaciones, contactos y sectores económicos, clave para las estrategias B2B.

Herramientas Elegidas:

Se utilizó Aiven como herramienta principal para la creación y gestión de la base de datos. Aiven es una plataforma que permite implementar y gestionar servicios de datos en la nube de forma rápida y eficiente.

Razones para Elegir Aiven

1. Gestión Simplificada: Aiven ofrece una interfaz intuitiva para la configuración y administración de bases de datos.
2. Soporte Multi-nube: Permite implementar servicios en varias nubes (AWS, Google Cloud, Azure, etc.), lo que brinda flexibilidad y alta disponibilidad.
3. Seguridad Incorporada: Incluye cifrado en tránsito y en reposo, así como herramientas avanzadas para el manejo de credenciales y accesos.
4. Escalabilidad: Facilita la ampliación de recursos de la base de datos según las necesidades del proyecto.
5. Integración Fluida: Es compatible con herramientas modernas como psycopg2, utilizada en este proyecto para la conexión desde Python.

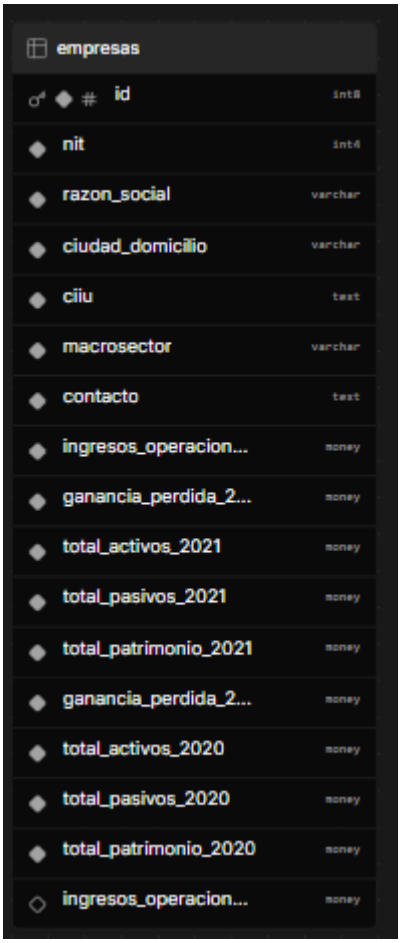
Base de Datos Seleccionada

Dentro de Aiven, se utilizó PostgreSQL debido a su robustez y características avanzadas como:

- Soporte para JSON para manejar datos semiestructurados.
- Transacciones confiables y consultas complejas.
- Escalabilidad para manejar grandes volúmenes de datos.

Con Aiven, fue posible configurar rápidamente el servicio de base de datos, establecer políticas de seguridad y asegurar la accesibilidad desde la aplicación Python.

Diseño de la Base de Datos



Descripción del Diseño

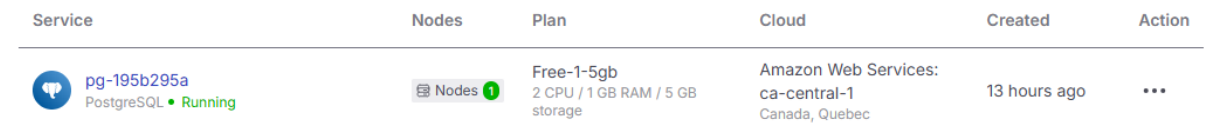
- id: bigint, no nulo, clave primaria.
- nit: integer, no nulo.
- razon_social: varchar, no nulo.
- ciudad_domicilio: varchar, no nulo.
- ciiu: text, no nulo.
- macrosector: varchar, no nulo.
- contacto: text, no nulo.
- ingresos_operacionales_2021: money, no nulo.
- ganancia_perdida_2021: money, no nulo.
- total_activos_2021: money, no nulo.
- total_pasivos_2021: money, no nulo.
- total_patrimonio_2021: money, no nulo.

- ganancia_perdida_2020: money, no nulo.
- total_activos_2020: money, no nulo.
- total_pasivos_2020: money, no nulo.
- total_patrimonio_2020: money, no nulo.
- ingresos_operacionales_2020: money, nulo.

Esta tabla almacena información financiera y de identificación de empresas, incluyendo datos de ingresos, ganancias, activos, pasivos y patrimonio para los años 2020 y 2021.

Proceso de Conexión

Explicar cómo conectaron el script de scraping con la base de datos para insertar datos.




```python
import os

def get_db_config():
    return {
        'host': 'pg-195b295a-nicolasdavid2409-dd5b.d.aivencloud.com',
        'port': '27225',
        'database': 'defaultdb',
        'user': 'avnadmin',
        'password': 'AVNS_nEfx8XN1vdiYhATDyS2'
    }
```

Desafíos y Soluciones

El error TypeError: unsupported operand type(s) for /: 'str' and 'str' indica que estás intentando realizar una operación de división (/) entre dos valores que son cadenas de texto (strings)

El error KeyError: '2021' que estás viendo indica que la columna '2021' no existe en el DataFrame filtered_df.
El error KeyError: 'TOTAL_PASIVOS_2021' indica que el DataFrame filtered_df no contiene una columna con el nombre 'TOTAL_PASIVOS_2021'.
El error que estás viendo, OperationalError: connection to server at "pg-195b295a-nicolasdavid2409-dd5b.d.aivencloud.com" (35.183.184.210), port 27225 failed: FATAL: no pg_hba.conf entry for host "186.113.69.19", user "avnadmin", database "defaultdb", no encryption, indica que tu conexión a la base de datos PostgreSQL está siendo rechazada porque no hay una entrada correspondiente en el archivo pg_hba.conf para permitir esa conexión.

Las soluciones a los problemas se resolvieron cambiando el tipo de las variables,

Visualización de Datos

Objetivo

El propósito de la visualización es proporcionar un resumen claro y conciso de los datos de proveedores B2B. La aplicación permite a los empresarios explorar información relevante sobre proveedores, incluyendo métricas financieras y de riesgo, facilitando la toma de decisiones informadas. Los gráficos interactivos ofrecen una comparación visual de indicadores clave, lo que ayuda a los usuarios a identificar rápidamente proveedores que cumplen con sus requisitos.

Herramientas

Se eligió **Streamlit** como herramienta para la visualización debido a su simplicidad y capacidad para crear aplicaciones web interactivas de manera rápida. Streamlit permite a los desarrolladores centrarse en el código y la lógica de negocio, mientras que proporciona una interfaz de usuario atractiva y fácil de usar. Además, su integración con bibliotecas de visualización como Plotly hace que sea ideal para crear gráficos interactivos y dinámicos.

Visualización de Datos

Se utilizaron gráficos de barras para comparar métricas clave entre proveedores y a lo largo de diferentes años. Los gráficos de barras son útiles para visualizar comparaciones entre categorías, permitiendo a los usuarios identificar rápidamente las diferencias en el rendimiento de los proveedores.

Estructura de la Interfaz

La interfaz de la aplicación se estructuró en dos secciones principales:

Listado de Proveedores: Permite a los usuarios buscar y filtrar proveedores según su razón social. Se utiliza un cuadro de texto para la búsqueda y una tabla para mostrar los resultados filtrados.

Gráficos de Comparación: Se presentan gráficos comparativos de métricas clave como la razón de endeudamiento, rentabilidad y solvencia para los proveedores seleccionados.

Desafíos y Soluciones

Durante el desarrollo de la visualización, se presentaron varios desafíos

Dificultades para Generar Gráficos Específicos: Al principio, hubo problemas al intentar mostrar gráficos que comparan múltiples métricas en un solo gráfico. Esto se soluciona utilizando la biblioteca Plotly, que permite crear gráficos más complejos y personalizados.

Problemas de Geocodificación: La geocodificación de las ciudades para mostrar ubicaciones en el mapa no siempre devolvía resultados. Se implementó un manejo de errores que muestra un mensaje de advertencia cuando no se pueden geocodificar las ciudades.

Conclusiones

Logros del Proyecto:

Desarrollo de una Aplicación Interactiva

Se creó una aplicación web utilizando Streamlit que permite a los usuarios buscar y comparar proveedores de manera efectiva. La interfaz es intuitiva y permite filtrar proveedores por nombre y macrosector, facilitando la toma de decisiones informadas.

Análisis de Datos Financieros

Se implementaron métricas clave como la razón de endeudamiento, rentabilidad y solvencia para los años 2020 y 2021. Esto proporciona a los usuarios una visión clara del desempeño financiero de los proveedores.

Visualización de Datos

Se utilizaron gráficos interactivos de Plotly para comparar visualmente las métricas de los proveedores, lo que ayuda a los usuarios a identificar rápidamente tendencias y diferencias significativas.

Geocodificación y Visualización en Mapa

Se integró la funcionalidad de geocodificación para mostrar la ubicación de los proveedores en un mapa, lo que añade una dimensión geográfica a la información presentada y ayuda en la logística y planificación de la cadena de suministro.

Recomendaciones Personalizadas

Se generaron recomendaciones basadas en el análisis de datos, lo que añade valor a la aplicación al ofrecer consejos prácticos a los usuarios sobre cómo mejorar su relación con los proveedores.

Cumplimiento de Requisitos de la Hackathon

El proyecto cumplió con los requisitos de la hackathon al abordar un problema real en la gestión de proveedores, utilizando tecnologías modernas de análisis de datos y visualización. La aplicación es escalable y puede adaptarse a diferentes industrias, lo que la hace relevante para un amplio público.

Reflexión del Equipo

La experiencia en la hackathon fue enriquecedora y desafiante. Trabajar en equipo nos permitió combinar nuestras habilidades y conocimientos para crear una solución que aborda un problema real en la gestión de proveedores. La colaboración y el intercambio de ideas fueron fundamentales para el desarrollo del proyecto.