TRUST

acción para la Transparencia, Responsabilidad y Unidad Social mediante Tecnología

Documento de implementación Técnica

Índice

1. Introducción

La contratación pública, específicamente la negociación entre el Estado y los contratistas, es una de las áreas más susceptibles a la corrupción. En este ámbito, se observan irregularidades como la manipulación de licitaciones, la existencia de sobrecostos y la adjudicación de contratos a empresas con vínculos políticos. Con el fin de abordar estos desafíos, se propone el desarrollo de un prototipo con las siguientes características, enfocado en un principio en el análisis de contratos de emergencia de República Dominicana:

1. **Disponibilización de la información existente en la documentación física de los contratos:** Se desarrolló un prototipo enfocado en la digitalización y estructuración de la información en formato físico, la cual es de difícil acceso e interpretación. Para el desarrollo del prototipo en cuestión se partió de la digitalización de los oferentes de los contratos, las características financieras de los contratos, y los componentes, productos o ítems que componen cada contrato.
2. **Detección de irregularidades y datos faltantes:** Se implementaron mecanismos para identificar diferencias entre los documentos de contratación en formato físico y las bases de datos de Compras Dominicana.

Además, se identificaron las siguientes características que representan potenciales mejoras para su desarrollo en un mediano y largo plazo.

3. **Identificación de anomalías:** El sistema estará diseñado para detectar irregularidades en los procesos de contratación, permitiendo una supervisión más efectiva y oportuna.
4. **Consulta en Lenguaje Natural:** Se desarrollará una propuesta que permita la consulta en lenguaje natural de los documentos adjuntos a cada contrato, independientemente de si estos están en formato digital o físico.
5. **Plataforma Centralizada:** El diseño de una plataforma centralizada facilitará el análisis y la visualización detallada de los contratos, ofreciendo una experiencia amigable para los usuarios.

Este prototipo cuenta con las características necesarias para su integración con la arquitectura existente de MapaInversiones, como se explicara a continuación. Además, se priorizó el desarrollo de una solución que altamente escalable, haciendo énfasis en el diseño de componentes desacoplados. Esto permite una mayor flexibilidad y adaptabilidad para satisfacer futuras necesidades tecnológicas.

2. Estructura del prototipo

a. Componente de Ingesta de datos

En cuanto a la ingesta de datos, inicialmente se procesa información estructurada y no estructurada, incluyendo no solo los datos actuales de República Dominicana sobre contratación, sino también el enriquecimiento mediante un proceso de **Web Scraping**. Este proceso permite extraer información detallada de cada proceso dentro del portal de **Compras Dominicana** y la **documentación física adjunta.** Este enfoque proporcionó un enriquecimiento de la información con los siguientes beneficios:

- **Validación de la información**: Asegura la consistencia dentro del portal y permite notificar sobre posibles irregularidades o diferencias entre las fuentes de información para una rápida toma de decisiones.
- **Extracción de información de documentos**: Mediante herramientas de procesamiento de documentos se puede ejecutar una rápida extracción y digitalización de los archivos adjuntos de cada proceso, facilitando su consulta por parte de los diferentes interesados.

Para el desarrollo de la solución, dentro del componente de ingesta de datos se implementó un módulo de minería de datos, el cual cuenta con las siguientes características:

1. **Sistema de Web Scraping:**

   Funcionalidad: Este sistema está diseñado para extraer información disponible en la plataforma de Compras Dominicana. La información obtenida se destina a un análisis detallado y se contrasta con otras fuentes de datos, permitiendo una visión más completa y precisa.

2. **Bot de Web Scraping:**

   Funcionalidad: Este bot automatiza la descarga y organización de documentos provenientes de la plataforma de Compras Dominicana. Como parte de su funcionalidad de control de cambios, el bot incorpora un sistema de metadatos que registra los archivos ya procesados, incluyendo la fecha de su procesamiento. Esto permite

identificar rápidamente nueva información en la plataforma y evitar la duplicación en el procesamiento de documentos.

3. **Extracción de Información del Portal de Mapa Inversiones:**

   Funcionalidad: Este sistema está orientado a la extracción de datos desde el portal de Mapa Inversiones. La información extraída se utiliza para su validación y contraste con las demás fuentes de información, garantizando así la coherencia de los datos procesados.

Finalmente, toda esta información del modúlo se almacena dentro del componente de almacenamiento, en el esquema **Bronce**, sobre el cual se detallará más adelante.

b. Componente de Procesamiento

El componente de procesamiento se compone de dos módulos principales, ambos enfocados en el tratamiento de la información extraída en el módulo anterior. Los módulos son los siguientes:

1. Módulo de Procesamiento de Documentos (OCR):

Este módulo está diseñado para utilizar herramientas Open Source en la ejecución de procesos de Reconocimiento Óptico de Caracteres (OCR). Específicamente, se implementó el OCR de Tesseract para extraer información de los documentos, la cual luego es estructurada por páginas e indexada utilizando el código de contrato correspondiente para su posterior análisis. Toda la información procesada se almacena en el esquema de **Plata**, ya que, aunque ha sido procesada, aún no está lista para su consumo final.

2. Módulo de Procesamiento de Textos (NLP):

Este módulo se encarga de aplicar técnicas de Procesamiento de Lenguaje Natural (NLP) para detectar y estructurar la información de interés dentro de los documentos, utilizando modelos de Lenguaje de Gran Escala (LLM) públicos, de menor tamaño y de ejecución local, además de expresiones regulares. Además, identifica los riesgos relevantes que serán visualizados en el siguiente componente. En este caso, dado que la información ya ha sido limpiada y estructurada, se almacena en el esquema de **Oro** dentro del componente de almacenamiento.

c. Componente de Visualización

Como parte final del prototipo, se desarrolló un informe en Power BI con el objetivo de integrarlo dentro de la página de MapaInversiones, permitiendo una consulta detallada de la información previamente recolectada. Este informe consta de un dashboard que facilita un análisis comprensivo y detallado en varias dimensiones:

1. **Análisis General de Contratos de Emergencia:**El dashboard ofrece una visión global de todos los contratos de emergencia, permitiendo a los usuarios contar con la información necesaria para comprender como se está desarrollando el proceso de contratación de emergencia del país.
2. **Análisis Detallado de Cada Contrato y Producto:** Además de la vista general, el informe proporciona un desglose de cada contrato. Esto incluye detalles específicos sobre los componentes, ítems o productos involucrados en cada acuerdo, proporcionando una comprensión granular de los mismos.
3. **Estadísticas sobre Indicadores de Riesgo e Irregularidades:** El dashboard también incluye métricas e indicadores clave relacionados con posibles riesgos e irregularidades. Entre las irregularidades se destaca la existencia de discrepancias entre la información disponible en diferentes fuentes. Como factor de riesgo, se monitorea la coherencia de la información mencionada entre los documentos asociados a cada contrato.

Este enfoque permite a los funcionarios públicos identificar rápidamente situaciones potencialmente riesgosas, facilitando la investigación oportuna y la toma de decisiones informadas en la gestión de contratos públicos.

d. Componente de Almacenamiento

El componente de almacenamiento está organizado en tres esquemas: Bronce, Plata y Oro, cada uno diseñado para cumplir con propósitos específicos:

1. Esquema de Bronce:

Este esquema almacena la información tal como se encuentra en su origen, sin aplicar ningún procesamiento inicial. Su principal función es centralizar los datos, preparándolos para su posterior consumo y procesamiento.

2. Esquema de Plata:

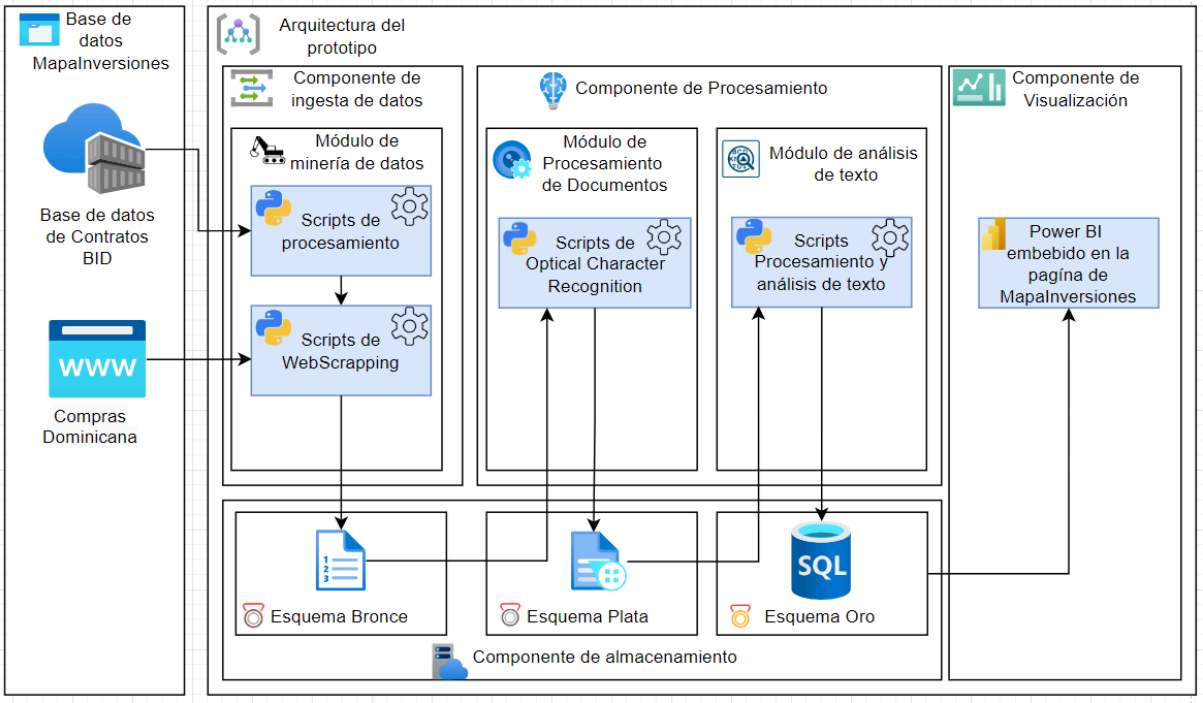
   El esquema de Plata actúa como un paso intermedio. Aquí se almacenan los documentos digitalizados, aunque aún no han sido estructurados ni procesados completamente. Sin embargo, se disponibilizan para consulta por parte de los actores interesados, permitiendo acceso anticipado a la información.

3. Esquema de Oro:

   En el esquema de Oro se almacena la información final, que ha sido completamente analizada y estructurada de manera relacional. Esta información está lista para ser consumida por herramientas de visualización y análisis.

En conjunto, el componente de almacenamiento permite desacoplar cada parte del sistema, actuando como el único intermediario. Esto proporciona flexibilidad, permitiendo que cada componente escale según sus necesidades específicas sin requerir un aumento innecesario en la capacidad de almacenamiento o la incorporación de componentes adicionales durante su implementación.

3. Diagrama de Arquitectura



4. Características del prototipo

a. **Disponibilidad de la Información**: El desarrollo del prototipo manejando esquemas independientes, facilita la integración de APIs para cada parte del flujo de datos, garantizando la disponibilidad continua de la información en cada etapa del proceso.

b. **Escalabilidad**: La implementación de componentes desacoplados permite una fácil escalabilidad y ajustada a las necesidades específicas de cada componente, optimizando así los recursos y evitando el escalado innecesario de componentes.

c. **Flexibilidad**: Al utilizar componentes desacoplados, se puede sustituir cualquier sección del proceso según las necesidades tecnológicas sin afectar el flujo general, proporcionando una gran flexibilidad en el diseño, implementación y adaptación del prototipo.

d. **Interoperabilidad**: Un único punto de acceso, dado por el componente de ingesta, permite, tanto el desarrollo de componentes especializados para el procesamiento de datos, como la fácil integración de la información en los pasos subsecuentes del flujo de datos.