#5 EN INNOVACIÓN

RETO DE CIUDAD 1

DATOS CON HISTORIA

DOCUMENTO DESCRIPTIVO DE LA SOLUCIÓN

Reto de Ciudad 1: Datos con Historia

I. Información del equipo postulante

Título de la solución: ***Optimización Inteligente de Datos para la Planeación Urbana: Una Solución Basada en IA para la SDP***

Nombre del equipo o razón social: **CivisAI**

Integrantes y roles:

Nombre completo	Rol en el equipo	Formación	Experiencia relevante
Javier Arturo Forero Ruiz	Coordinador técnico, analista de datos, científico de datos	Estadístico y Científico de Datos	Más de 20 Años trabajando en Analítica de Datos, Transformación Digital y Estadística Más de 7 Años trabajando en Inteligencia Artificial, Manejo de Datos en la Nube y Manejo de proyectos
Jairo José Niño Pérez	Analista de datos, Científico de datos y desarrollador full stack	Politólogo y Científico de Datos.	Más de 10 años desarrollando productos de datos y aplicaciones para toma de decisiones. 7 años de Investigación en Teoría de Información y Deep Learning.

II. Resumen ejecutivo

Descripción del Problema y la Solución Propuesta

La Secretaría Distrital de Planeación (SDP) enfrenta serios desafíos en el manejo eficiente de grandes volúmenes de datos cualitativos provenientes de diversas fuentes y formatos. Las áreas críticas incluyen la gestión de aportes ciudadanos, el seguimiento de políticas públicas y el monitoreo de proyectos de inversión. Los procesos actuales, predominantemente manuales, generan cuellos de botella y afectan la toma de decisiones estratégicas.

CivisAI propone una solución tecnológica basada en Inteligencia Artificial (IA) para automatizar el análisis y categorización de datos cualitativos. Esta plataforma integra herramientas de procesamiento de lenguaje natural (NLP), técnicas de clasificación automática y visualización avanzada para facilitar el acceso, análisis y presentación de la información.

Principales Beneficios e Impacto Esperado

- **Optimización del Análisis**: Reducción significativa de los tiempos operativos mediante la automatización del procesamiento de actas, reportes y propuestas.
- **Mejora en la Toma de Decisiones**: Acceso a resúmenes personalizados con rankings de palabras clave y trazabilidad de los datos utilizados, facilitando decisiones informadas y oportunas.
- **Fortalecimiento de la Participación Ciudadana**: Garantía de que los aportes ciudadanos se integren efectivamente en los procesos de planeación urbana.
- **Integración Eficiente**: Compatibilidad con los sistemas existentes de la SDP, como SEGPLAN y plataformas colaborativas como OneDrive y SharePoint.
- **Escalabilidad y Seguridad**: Uso de infraestructura en la nube y medidas robustas de seguridad, garantizando el crecimiento y la protección de los datos.
- **Basado en las expectativas y necesidades** de las tres subsecretarías, se reconoce la importancia de integrar y vincular la información para una solución colaborativa. Esta **integración** facilita la correlación entre el PDD, POT y los ODS. Una solución de este tipo permite que los aportes cualitativos respalden y complementen la inversión y las políticas públicas, asegurando una trazabilidad clara y efectiva de todas las fuentes de información, así como el involucramiento de ciudadanos y organizaciones en la toma de decisiones

Aspectos Innovadores Destacados

- **Uso de Librerías y Modelos de Código Abierto**: Implementación de herramientas como **SpaCy**, **HuggingFace Transformers**, **LlamaIndex** y **sentence-transformers**. Estas soluciones ofrecen:
    - **Flexibilidad y Adaptabilidad**: Permiten ajustar los modelos a contextos específicos como políticas públicas, PDD, POT y ODS.
    - **Reducción de Costos**: El uso de tecnologías abiertas minimiza costos de licenciamiento.
    - **Transparencia y Reproducibilidad**: Facilita la auditoría y el seguimiento de los procesos, garantizando resultados confiables y verificables.
    - **Innovación Continua**: Acceso a actualizaciones y mejoras constantes de una comunidad global de desarrolladores.
- **Pipeline RAG (Retrieval-Augmented Generation)**: Facilita búsquedas basadas en significado semántico, mejorando la precisión en la recuperación de información.
- **Embeddings Semánticos**: Uso de representaciones vectoriales para capturar el significado de los textos y realizar búsquedas relevantes.
- **Integración de IA y NLP**: Clasificación automática, extracción de entidades y generación de resúmenes adaptados a las necesidades específicas de cada dependencia.

CivisAI no solo mejora la eficiencia operativa de la **SDP**, sino que también asegura una mayor transparencia y participación ciudadana en los procesos de planificación, transformando la forma en que se gestionan los datos cualitativos en la ciudad.

III. Identificación del problema

Contexto y problemática:

El principal desafío identificado por la Secretaría Distrital de Planeación (SDP) es la gestión y análisis de grandes volúmenes de datos cualitativos desestructurados, que dificultan la integración de estos insumos en la toma de decisiones estratégicas y la planificación urbana. Este problema se descompone en tres áreas principales:

a. **Categorización y análisis de datos de participación ciudadana**: La Oficina de Participación y Diálogo Ciudadano (OPDC) enfrenta dificultades para procesar actas, propuestas y aportes ciudadanos recogidos en múltiples formatos y de calidad variable.

b. **Seguimiento y análisis de políticas públicas**: La Subsecretaría de Políticas Públicas y Planeación Social y Económica (SPPPSE) gestiona más de 3,400 productos relacionados con 51 políticas públicas, enfrentando cuellos de botella por procesos manuales de análisis.

c. **Monitoreo de proyectos de inversión**: La Subsecretaría de Planeación de la Inversión (SPI) requiere herramientas más robustas para analizar datos cualitativos y detectar alertas tempranas en proyectos de inversión del Plan Distrital de Desarrollo (PDD).

Estos problemas generan ineficiencias operativas, afectan la calidad de la toma de decisiones y limitan la capacidad de la ciudadanía para participar y supervisar las iniciativas urbanas.

Usuarios y necesidades:

Dependencia	Perfil del Usuario	Necesidades específicas identificadas	Casos de uso
SPPPSE (Subsecretaría de Políticas Públicas y Planeación Social y Económica)	Los funcionarios responsables de las **(6)** principales politicas: Acción Climática, Movilidad, Pobreza, Calidad de Educación, Acceso a Salud y Seguridad.	Categorización de los reportes cualitativos en variables estratégicas de ciudad	Ofrecemos resúmenes personalizados para cada política, con flexibilidad en la longitud de la respuesta (según el número de palabras), tanto en el componente **cualitativo** o de **enfoque**, y con la opción de seleccionar el **período** de análisis trimestral. La respuesta incluirá: Resumen, Ranking de **Top 10** Palabras Claves utilizadas y Trazabilidad con cuales registros **(2720)** fueron utilizados para obtener dicha información.
SPPPSE (Subsecretaría de Políticas Públicas y	Restantes funcionarios de la Subsecretaría **(≈4)**	Generación automatizada de aportes a la ciudad de los	Producción de resúmenes personalizados para el **PDD/ODS** o preguntas abiertas (*ej.: ¿Que hizo el*

Planeación Social y Económica)	que requieren análisis más generales por Productos, Indicadores o Sectores	productos a los objetivos definidos en el plan de desarrollo (PDD) y ODS (Objetivos de Desarrollo Sostenible).	*distrito para mejorar la seguridad de las mujeres en la ciudad?*), con flexibilidad en la longitud (según el número de palabras), tanto en el componente **cualitativo** o de **enfoque**, y con la opción de seleccionar el **período** de análisis trimestral. La respuesta incluirá: Resumen, Ranking de **Top 10** Palabras Claves utilizadas y Trazabilidad con cuales registros **(2720)** fueron utilizados para obtener dicha información.
Subsecretaría de Planeación de la Inversión **(SPI)**	Usuarios de la Subsecretaría que deben revisar la información entregada por las diferentes entidades	Control de calidad de información garantizando la coherencia de las temáticas tratadas	Podrán acceder a resúmenes de la información presentada por las **Entidades**, con la opción de filtrar por **Objetivo, Programa, Meta o Indicador**. con flexibilidad en la longitud de la respuesta (según el número de palabras), La visualización estará disponible en dos niveles: **'Retrasos y Soluciones'** o **'Avances y Logros'**. La respuesta incluirá: un resumen, un ranking de las **10** palabras clave más utilizadas y la trazabilidad de los indicadores **(512)** empleados para obtener dicha información
Subsecretaría de Planeación de la Inversión **(SPI)**	Funcionarios de la Subsecretaría que deben analizar la rendición de cuentas, definir planes de acción o en general hacer el diagnóstico del PDD	Generar las principales tendencias de la información cualitativa brindada por otras áreas de la SDP y las entidades distritales	Podrán acceder a resúmenes con la opción de filtrar por **Objetivo, Programa, Meta o Indicador**. Asimismo, tendrán la posibilidad de responder **preguntas abiertas** (por ejemplo: *'¿Qué hizo el plan por los adultos mayores?*), con flexibilidad en la longitud de la respuesta según el **número de palabras**. La información se visualizará en dos niveles: **'Retrasos y Soluciones'** o **'Avances y Logros'.** Cada respuesta incluirá un resumen, un ranking de las **10 palabras** clave más relevantes y la trazabilidad de cuales de los indicadores **(512)** fueron utilizados para generar dicha información.

Oficina de Participación y Diálogo Ciudadano **(OPDC)**	Funcionarios de la Subsecretaria a cargo de las relatorías o los informes técnicos de las estrategias de participación	Facilitar el intercambio de información al interior de la entidad, creando un entorno colaborativo que mejore la toma de decisiones, permitiendo la trazabilidad de los aportes ciudadanos para que se corrobore efectivamente donde fueron incorporados o pudieran ser considerados	Generación de resúmenes personalizados sobre la participación ciudadana, relacionados con los **PDD, POT, ODS** o con preguntas abiertas específicas (*por ejemplo: '¿Qué ideas o propuestas existen para mejorar la seguridad en la ciudad?*'). Se ofrece flexibilidad en la longitud del resumen según el número de palabras y la posibilidad de segmentar por variables demográficas disponibles en los aportes. La respuesta incluirá (3): un resumen, un ranking de las **10 palabras** clave más relevantes y la **trazabilidad de los aportes** específicos de ciudadanos, entidades u organizaciones utilizados para generar dicha información.
Oficina de Participación y Diálogo Ciudadano **(OPDC)**	**Funcionarios** de la Subsecretaría encargados de los productos pedagógicos o visores de participación, con **usuarios secundarios** como ciudadanos, entidades u organizaciones que consumen dichos reportes	Garantizar la recolección de insumos cualitativos, que vaya más allá de agrupar categorías temáticas.	Elaboración de resúmenes personalizados sobre **participación ciudadana** para identificar aportes provenientes de ciudadanos o entidades/organizaciones relacionados con aspectos específicos de los **PDD, POT, ODS**, o en respuesta a preguntas abiertas (por ejemplo: *'¿Qué ideas o propuestas sobre movilidad aborda el PDD?'*). Se ofrece flexibilidad en la extensión del resumen según el número de palabras y la opción de segmentar por variables sociodemográficas disponibles en los aportes. La respuesta incluirá (3): un resumen, un ranking de las 10 palabras clave más relevantes y la trazabilidad de los aportes específicos de ciudadanos, entidades u organizaciones utilizados para generar dicha información

IV. Objetivos de la solución

Objetivo general:

Desarrollar una solución tecnológica innovadora para automatizar el análisis y categorización de información cualitativa en la Secretaría Distrital de Planeación (SDP). Esta solución permitirá procesar de manera eficiente, rápida y precisa estos insumos valiosos, facilitando su integración en los procesos de planificación urbana y toma de decisiones estratégicas, a partir de los aportes ciudadanos y reportes de las entidades.

Objetivos específicos:

1. Automatizar el análisis de datos cualitativos provenientes de diversos formatos (PDF, matrices, propuestas abiertas)
   - **Meta Concreta y Medible**: Implementar una solución tecnológica que permita procesar automáticamente distintos tipos de archivos y formatos.
   - **Contribución al Objetivo General**: Facilita el procesamiento eficiente, rápido y preciso de información cualitativa, alineado con la necesidad de automatización y optimización.

2. Estandarizar y categorizar la información cualitativa, garantizando su coherencia y calidad.
   - **Meta Concreta y Medible**: Desarrollar procesos automatizados más modernos y de punta que aseguren la clasificación coherente de datos y reduzcan inconsistencias. Por ejemplo, clasificar y mapear los **512** indicadores del Plan Distrital de Desarrollo (PDD) con los **17** Objetivos de Desarrollo Sostenible (ODS), permitiendo una categorización estructurada y alineada con estándares internacionales.
   - **Contribución al Objetivo General**: Permite asegurar la calidad de los insumos procesados, lo que mejora la toma de decisiones estratégicas y la planificación urbana.

3. Integrar los datos procesados en herramientas existentes para facilitar el seguimiento y la toma de decisiones estratégicas.
   - **Meta Concreta y Medible**: Lograr una integración efectiva con sistemas como **SEGPLAN** y otras plataformas de la **SDP**. Por ejemplo, asegurar que la información cualitativa procesada permita correlacionar el **PDD**, el **POT** y los **ODS**. Esta integración facilita una solución colaborativa que respalda la inversión y las políticas públicas, garantiza una trazabilidad clara de los aportes y fomenta el involucramiento ciudadano en la toma de decisiones.
   - **Contribución al Objetivo General**: Optimiza el uso de la información procesada al conectarla con las plataformas utilizadas en los procesos de planificación y monitoreo.

4. Reducir tiempos operativos asociados a la recolección, análisis y reporte de datos cualitativos.
   - **Meta Concreta y Medible**: Disminuir significativamente el tiempo necesario para procesar grandes volúmenes de datos cualitativos. Por ejemplo:

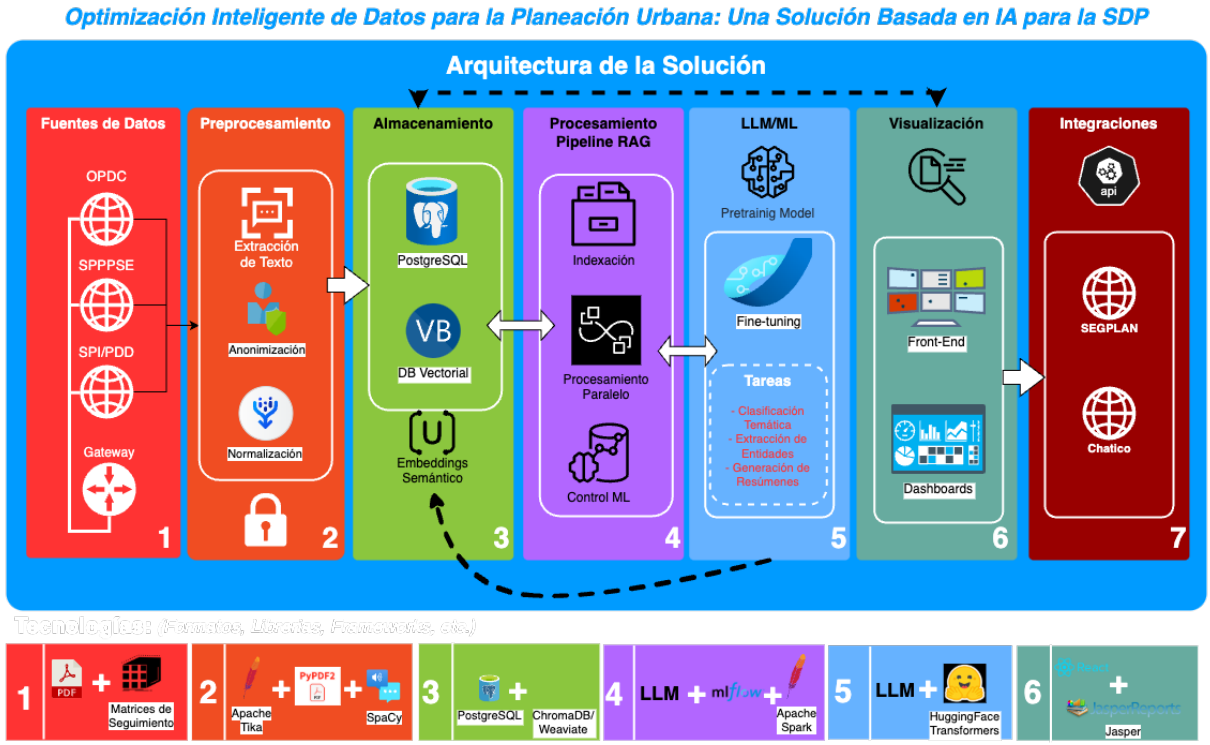
- Al equipo de políticas públicas de la **SPPPSE**, actualmente le toma **2 semanas** analizar información cualitativa de **6 políticas** con un equipo de **10 a 12 personas**. Con la automatización, este análisis podría reducirse a **horas**, permitiéndoles enfocarse en cómo los hallazgos aportan a cada política y en la formulación de estrategias concretas.
- Para el equipo de inversión de la **SPI**, el procesamiento parcial de aportes cualitativos de las entidades toma entre **3 a 4 días**. Con una solución tecnológica, este proceso podría reducirse a **horas**, facilitando una revisión más rápida y mejorando la eficiencia en la rendición de cuentas.

- **Contribución al Objetivo General**: Atiende la necesidad de procesar información rápidamente para cumplir con los plazos estrictos en la toma de decisiones.

5. Fortalecer la participación ciudadana, asegurando que los aportes realizados sean procesados y utilizados de manera efectiva.
   - **Meta Concreta y Medible**: Mejorar el procesamiento y aprovechamiento de los aportes ciudadanos para garantizar su impacto en decisiones de planificación.
   - **Contribución al Objetivo General**: Responde a la necesidad de asegurar que la participación ciudadana influya efectivamente en la toma de decisiones urbanas.

V. Descripción de la solución tecnológica

Arquitectura de la solución:



📊 Descripción de los Principales Módulos de la Arquitectura de CivisAI

🟥 Módulo 1: Fuentes de Datos

Descripción: Este módulo agrupa las diferentes fuentes de información que alimentan el sistema, como:

- OPDC, SPPPSE, SPI, PDD, POT, ODS *(Mayor descripción en la sección posterior)*

Tecnologías: PDF y Matrices de Seguimiento como formatos de datos de entrada.

🟧 Módulo 2: Preprocesamiento

Descripción: Realiza la extracción y limpieza de datos para su posterior análisis. Incluye procesos como:

- **Extracción de Texto:** Extraer contenido de documentos en PDF de ser necesario
- **Anonimización:** Ocultar o eliminar información sensible.
- **Normalización:** Estandarizar los datos para garantizar coherencia.
- **Preprocesamiento de texto**, incluyendo lematización y eliminación de stopwords (SpaCy)
- **Facilita el procesamiento** y limpieza de textos en español para mejorar el análisis (SpaCy)
- **División de textos** (expresiones regulares) en oraciones para realizar chunking (re)

Tecnologías: Apache Tika, PyPDF2, Librerías Python (SpaCy, re)

🟩 Módulo 3: Almacenamiento

Descripción: Almacena los datos preprocesados y los embeddings semánticos generados por modelos de lenguaje.

- PostgreSQL para datos estructurados, DB Vectorial (como ChromaDB) para almacenar embeddings semánticos.
- Generación de embeddings (representaciones vectoriales) de textos, que facilita el uso de modelos preentrenados para convertir textos en vectores numéricos, útiles para búsquedas semánticas (sentence_transformers)
- Su beneficio se da en permitir medir la similitud entre textos y realizar búsquedas basadas en el **significado**, no solo en coincidencias exactas de palabras.

Tecnologías:

- PostgreSQL, ChromaDB, Weaviate. Librerias Python (sentence_transformers)

🟪 Módulo 4: Procesamiento Pipeline RAG (Retrieval-Augmented Generation)

Descripción: Realiza el procesamiento avanzado de los datos mediante:

- Indexación: Construcción de índices de documentos y realización de consultas, facilita la búsqueda eficiente y recuperación de información relevante desde un conjunto de textos. (llama_index)
- Procesamiento Paralelo: Optimización del procesamiento con tareas concurrentes.
- Control ML: Gestión del flujo de modelos de Machine Learning.

Tecnologías: Librerias Python (llama_Index), MLFlow, Apache Spark.

🔵 Módulo 5: LLM/ML

Descripción: Utiliza modelos de lenguaje amplio (LLM) y Machine Learning para tareas como:

- Clasificación Temática
- Extracción de Entidades
- Generación de Resúmenes
- El **fine-tuning** de modelos **LLM** permite adaptar modelos preentrenados al vocabulario y contexto específicos del PDD, POT y ODS, mejorando la precisión en tareas de búsqueda y análisis. Este proceso requiere una **GPU** solo durante el entrenamiento inicial, mientras que la inferencia puede ejecutarse eficientemente en **CPU**.

Tecnologías: HuggingFace Transformers, llama_index, hemos probado y seguiremos iterando con varios modelos **LLM** (daremos prioridad aquellos de código abierto) para seleccionar aquel que nos otorgue los mejores resultados para los diferentes casos

🟦 Módulo 6: Visualización

Descripción: Facilita la visualización de datos y resultados mediante:

- Front-End para la interacción con el usuario.
- Dashboards para visualizar métricas, reportes y análisis.

Tecnologías: React, Reflex, JasperReports *(adaptándonos a la infraestructura ya existente en la SDP)*

🟥 Módulo 7: Integraciones

Descripción: Permite la conexión e integración con sistemas existentes de la SDP como:

- SEGPLAN (Sistema de Planeación).
- Chatico para asistencia conversacional.
- APIs para facilitar la comunicación entre sistemas.

Tecnologías: APIs REST para integraciones.

Recomendaciones de Infraestructura Alineadas con el Stack Tecnológico de la Entidad

Teniendo en cuenta la infraestructura de la SDP (Secretaría Distrital de Planeación) detallada en el documento proporcionado, aquí están las recomendaciones específicas para alinear la arquitectura de **CivisAI** con los servicios existentes:

Almacenamiento

- Infraestructura Actual:
  - MySQL, PostgreSQL, Oracle en servidores locales y nube OCI (Oracle Cloud Infrastructure).
  - Servidores Físicos y Virtuales.
- Recomendación:
  - Utilizar PostgreSQL para bases de datos estructuradas y ChromaDB para almacenamiento vectorial, asegurando compatibilidad con la infraestructura actual.
  - Compatible con Python y fácil de integrar con bibliotecas como LlamaIndex y LangChain.
  - Implementar almacenamiento distribuido en la nube OCI para asegurar redundancia y escalabilidad.

Autenticación y Seguridad

- Infraestructura Actual:
  - LDAP Active Directory y LDAP Open DJ para autenticación.
  - Antivirus y medidas de seguridad perimetral.
- Recomendación:
  - Integrar la autenticación en **CivisAI** con LDAP Active Directory para un acceso centralizado y seguro.
  - Implementar cifrado TLS/SSL para proteger las conexiones y datos en tránsito.
  - Asegurar que el procesamiento y visualización respeten los permisos basados en roles (RBAC).

Servidores de Aplicaciones

- Infraestructura Actual:
  - NGINX y Apache para balanceo de carga y despliegue de aplicaciones.
- Recomendación:
  - Utilizar NGINX como proxy inverso y balanceador de carga para las aplicaciones web y servicios de IA.
  - Desplegar la solución en contenedores con **Docker** y orquestar con **Kubernetes** para una mayor resiliencia y alojar **ChromaDB**

Servicios en la Nube

- Infraestructura Actual:
  - Uso de Oracle Cloud Infrastructure (OCI) para servicios en la nube.
  - OneDrive y SharePoint para colaboración.
- Recomendación:
  - Asegurar que los módulos de almacenamiento y procesamiento estén integrados con OCI para aprovechar la infraestructura existente.
  - Permitir exportación de resultados y reportes directamente a OneDrive o SharePoint.

Visualización y Reportes

- Infraestructura Actual:
  - Herramientas como JasperReports para reportes.
  - Bases de datos relacionales PostgreSQL y MySQL.
- Recomendación:
  - Usaremos algunos frameworks con **React/ Reflex** para los reportes
  - Exportar reportes personalizados desde **CivisAI** utilizando JasperReports.

Bases de Datos

- Infraestructura Actual:
  - PostgreSQL, MySQL, Oracle.
- Recomendación:
  - Utilizar PostgreSQL como la base de datos principal para almacenamiento de textos y metadatos.
  - Enlazar los resultados de búsqueda en **VectorDB** con metadatos almacenados en PostgreSQL para proporcionar respuestas más completas.
  - Mantener la compatibilidad con MySQL y Oracle para integraciones con sistemas heredados.

Seguridad y Balanceo

- Infraestructura Actual:
  - Seguridad perimetral y balanceo de aplicaciones con NGINX.
- Recomendación:
  - Implementar WAF (Web Application Firewall) con NGINX para proteger los endpoints.
  - Usar balanceo de carga para distribuir el tráfico entre los microservicios de la arquitectura.

Componente	Tecnología Recomendada	Infraestructura Existente
Almacenamiento Relacional	PostgreSQL, MySQL, Oracle	PostgreSQL, MySQL, Oracle
Base de Datos Vectorial	ChromaDB, FAISS, Weaviate	***Nueva Adición***
Servidores de Aplicaciones	NGINX, Apache	NGINX, Apache
Contenedores	Docker, Kubernetes	Servidores Virtuales y Físicos
Autenticación	LDAP Active Directory	LDAP Active Directory, LDAP Open DJ
Nube	Oracle Cloud Infrastructure (OCI)	OCI
Seguridad	TLS/SSL, WAF (NGINX), RBAC	Antivirus, Seguridad Perimetral
Reportes y Visualización	JasperReports, React, Reflex	Jasper
Colaboración	OneDrive, SharePoint	OneDrive, SharePoint
Procesamiento de IA	SentenceTransformers, LlamaIndex	***Nueva Adición***
Balanceo de Carga	NGINX	NGINX
Orquestación de Contenedores	Kubernetes	***Nueva Adición***

Stack tecnológico:

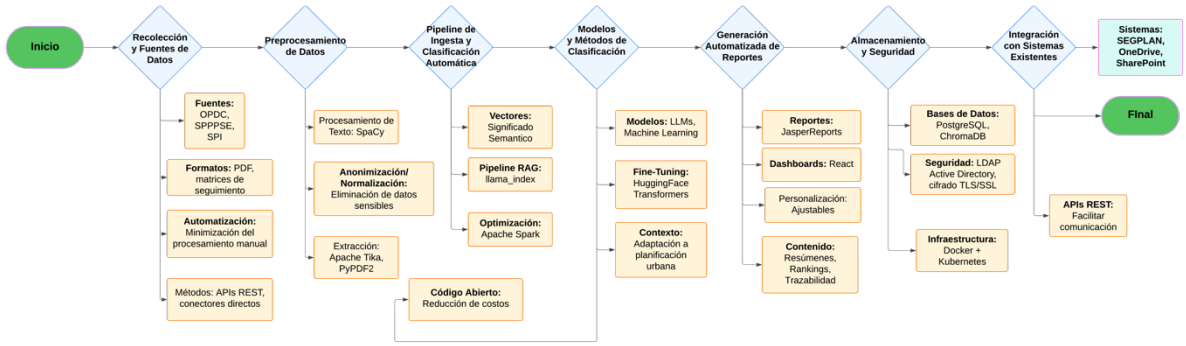
Categoría	**Tecnología(s) empleadas**	**Justificación** (Explica en una o dos frases por qué se eligió esa tecnología, destacando su alineación con los objetivos del proyecto y ventajas técnicas o funcionales)
Framework Web	NGINX, Apache, React/Reflex	Se eligieron por su capacidad de balancear carga y servir aplicaciones de manera eficiente y segura.
Lenguaje de Programación	Python	Python es versátil, cuenta con bibliotecas avanzadas para IA/NLP (como SpaCy y LlamaIndex) y se integra fácilmente con la infraestructura actual.
Base de Datos	PostgreSQL, MySQL, Oracle, ChromaDB	PostgreSQL y MySQL para datos estructurados; ChromaDB para búsqueda semántica y almacenamiento de embeddings.
Modelos de IA/NLP	SentenceTransformers, LlamaIndex, HuggingFace Transformers y modelos LLM de código abierto	Se seleccionaron por su eficiencia en tareas de clasificación, extracción de entidades y generación de resúmenes. Se recomienda que, una vez evaluados y seleccionado el modelo **LLM** más adecuado, se realice una instalación local del mismo en los servidores de la entidad. Esta implementación local permitirá optimizar los tiempos de respuesta en los llamados a los modelos, garantizar mayor seguridad y confidencialidad de los datos, y reducir costos operativos asociados a consultas recurrentes en la nube.
Otra(s)	Docker, Kubernetes, OCI, LDAP Active Directory	Docker y Kubernetes permiten despliegue escalable; OCI para almacenamiento en la nube; LDAP para autenticación centralizada.

#5 EN INNOVACIÓN

RETO DE CIUDAD 1

DATOS CON HISTORIA

Procesamiento y Gestión de Datos en CivisAI:



La solución **CivisAI** presenta un flujo integral de datos diseñado para optimizar el análisis cualitativo en la SDP. Se parte de la recolección inicial hasta el almacenamiento final, garantizando eficiencia, automatización y escalabilidad. A continuación, se detalla el proceso y los métodos utilizados:

1. **Recolección y Fuentes de Datos:**
   - Se integran fuentes de datos provenientes de múltiples dependencias como OPDC, SPPPSE y SPI. Los formatos incluyen documentos PDF, matrices de seguimiento y propuestas abiertas. Se prioriza la automatización en la ingesta de datos mediante APIs REST y conectores directos para minimizar el procesamiento manual.
2. **Preprocesamiento de Datos:**
   - En esta fase, se aplican técnicas avanzadas de extracción y limpieza de datos:
   - **Extracción de Texto:** Herramientas como **Apache Tika** y **PyPDF2** extraen contenido de documentos PDF.
   - **Anonimización y Normalización:** Se eliminan datos sensibles y se estandariza la información para garantizar coherencia.
   - **Procesamiento de Texto:** Se realiza lematización y eliminación de palabras vacías con **SpaCy**, lo que facilita un análisis semántico preciso. La segmentación por oraciones permite dividir textos extensos para mejorar la eficiencia del análisis.
3. **Pipeline de Ingesta y Clasificación Automática:**
   - Los textos preprocesados se convierten en representaciones vectoriales utilizando **sentence_transformers** para capturar significado semántico.
   - Se implementa un pipeline de **Retrieval-Augmented Generation (RAG)** con **llama_index**, que permite indexar documentos y realizar búsquedas basadas en significado, no solo en palabras clave.
   - Se optimiza el procesamiento con **Apache Spark** para manejar grandes volúmenes de datos y garantizar escalabilidad.
4. **Modelos y Métodos de Clasificación:**
   - Se aplican modelos de lenguaje amplio (**LLM**) y técnicas de Machine Learning para la clasificación temática y extracción de entidades.

- El **fine-tuning** de modelos como **HuggingFace Transformers** adapta las soluciones al contexto específico de la planificación urbana, el PDD, POD y los ODS.
- Se priorizan modelos de código abierto que faciliten la implementación y reduzcan costos.

5. **Generación Automatizada de Reportes:**
   - Los resultados del análisis se transforman en reportes personalizados con **JasperReports** y dashboards interactivos desarrollados con **React/Reflex**.
   - Los reportes incluyen resúmenes, rankings de palabras clave y trazabilidad de los datos utilizados.
   - Se ofrece flexibilidad para seleccionar períodos de análisis y niveles de detalle según las necesidades de cada dependencia.
6. **Almacenamiento y Seguridad:**
   - **PostgreSQL** almacena datos estructurados, mientras que **ChromaDB** gestiona los embeddings semánticos para facilitar búsquedas eficientes.
   - Se garantiza la seguridad mediante autenticación con **LDAP Active Directory** y cifrado **TLS/SSL**.
   - La solución se despliega en contenedores **Docker** orquestados con **Kubernetes** para asegurar alta disponibilidad y escalabilidad.
7. **Integración con Sistemas Existentes:**
   - **CivisAI** se integra con plataformas como **SEGPLAN**, **OneDrive** y **SharePoint** para facilitar el acceso y la colaboración.
   - APIs REST permiten una comunicación fluida entre **CivisAI** y otros sistemas de la SDP.

Este flujo de procesamiento de datos automatiza tareas repetitivas, reduce tiempos operativos y asegura un análisis cualitativo robusto, apoyando una toma de decisiones informada y eficiente.

VI. Metodología

Fase	Actividad
Recolección de Datos	Integrar datos desde OPDC, SPPPSE y SPI en formatos como PDF, matrices y propuestas abiertas. **Beneficio:** Centraliza la información para facilitar su acceso y análisis eficiente.
Preprocesamiento	Extraer y limpiar datos (Apache Tika, PyPDF2); anonimizar y normalizar textos (SpaCy, expresiones regulares). **Beneficio:** Garantiza datos limpios y consistentes para mejorar la calidad del análisis.
Clasificación Automática	Convertir textos en vectores usando sentence_transformers para búsquedas semánticas. **Beneficio:** Facilita encontrar información relevante rápidamente y con mayor precisión.
Pipeline RAG	Implementar RAG (llama_index) para indexar documentos y recuperar información relevante. **Beneficio:** Optimiza la recuperación de información para una toma de decisiones más ágil.
Procesamiento con Modelos LLM/ML	Aplicar modelos de lenguaje (HuggingFace Transformers) para clasificación, extracción y resúmenes. **Beneficio:** Automatiza el análisis temático y la generación de resúmenes detallados.
Almacenamiento	Almacenar datos estructurados en **PostgreSQL** y embeddings en **ChromaDB** para búsqueda semántica. **Beneficio:** Facilita el acceso rápido y organizado a grandes volúmenes de datos.
Seguridad	Implementar autenticación con LDAP, cifrado TLS/SSL y permisos basados en roles (RBAC). **Beneficio:** Garantiza la confidencialidad y el acceso seguro a la información.
Visualización	Desarrollar dashboards interactivos con React/Reflex y reportes personalizados con JasperReports. **Beneficio:** Permite una comprensión visual rápida de los resultados del análisis y nos adaptamos a la infraestructura actual en la entidad
Generación de Reportes	Crear resúmenes automáticos con ranking de palabras clave y trazabilidad de los datos utilizados. **Beneficio:** Facilita la toma de decisiones informadas con información sintetizada.
Integración con Sistemas Existentes	Conectar con SEGPLAN, OneDrive y SharePoint mediante APIs REST para facilitar colaboración. **Beneficio:** Mejora el flujo de información entre plataformas y áreas de la entidad. *Es fundamental que la integración de datos **cualitativos** se complemente con datos **cuantitativos** para proporcionar análisis más robustos y completos. La inclusión de datos numéricos permitirá generar resúmenes que no solo presenten insights cualitativos, sino también resultados concretos y medibles. Esta combinación fortalece la toma de decisiones al dimensionar el impacto real de las políticas y proyectos con cifras y métricas específicas.*

VII. Plan de implementación

Cronograma general de implementación:

Actividad	Mes		
	Mes 1	Mes 2	Mes 3
1. Sesiones individuales profundas con cada subsecretaría y sus usuarios	X		
2. Creación de Historias de Usuario para personalizar soluciones	X		
3. Recolección e integración de datos de OPDC, SPPPSE y SPI	X		
4. Consolidación del PDD, POT y ODS para crear embeddings semánticos	X		
5. Preprocesamiento de datos (extracción, limpieza y normalización)	X	X	
6. Desarrollo del pipeline RAG para indexación y búsqueda semántica	X	X	
7. Fine-tuning de los modelos LLM para adaptación específica		X	
8. Implementación de modelos LLM/ML para clasificación y resúmenes		X	X
9. Generación de embeddings semánticos y almacenamiento en ChromaDB		X	
10. Validación de mockups visuales de dashboards con usuarios		X	
11. Desarrollo de visualizaciones y dashboards interactivos		X	X
12. Integración con sistemas existentes (SEGPLAN, OneDrive, APIs)			X
13. Implementación de seguridad (LDAP, TLS/SSL, RBAC)			X
14. Pruebas y ajustes del sistema en cada módulo			X
15. Capacitación y entrega final de la solución **CivisAI**			X

Presupuesto general de implementación:

Categoría	Item	Cantidad	Costo unitario	Costo total
Análisis y Preparación Estratégica	Fase de Alineación, Integración y Preparación de Datos	1	$ 15,000,000	$ 15,000,000
	Gestión de Proyecto: Coordinación, seguimiento y entrega del proyecto	1	$ 5,000,000	$ 5,000,000
Desarrollo Tecnológico	Implementación de LLM, fine-tuning, embeddings semánticos y RAG	1	$ 20,500,000	$ 20,500,000
	Desarrollo de dashboards y visualizaciones	1	$ 8,200,000	$ 8,200,000
	Pruebas funcionales, de integración y de usuario	1	$ 2,500,000	$ 2,500,000
Integración	Integración con sistemas existentes	1	$ 4,500,000	$ 4,500,000
Seguridad	Autenticación y cifrado (LDAP, TLS/SSL)	1	$ 2,600,000	$ 2,600,000
Capacitación	Documentación Técnica y de Usuario	1	$ 3,000,000	$ 3,000,000
	Sesiones de capacitación y entrega	1	$ 3,000,000	$ 3,000,000
Otros Gastos	Gastos administrativos y operativos	1	$ 2,096,460	$ 2,096,460
	Contingencia para Riesgos	1	$ 1,500,000	$ 1,500,000
Deducciones Legales	Retención en la fuente (4%)	1	$ 3,600,000	$ 3,600,000
	Impuesto sobre la Renta	1	$ 8,243,000	$ 8,243,000
	Aportes a salud	3 meses	$1,500,000	$ 4,500,000
	Aportes a pensión	3 meses	$1,920,000	$ 5,760,000
	Pólizas (6 Meses) y seguros (20%)	1	$540.000	$ 540.000

VIII. Guía de usuario y requerimientos técnicos

Principales funcionalidades:

Funcionalidad	Descripción	Usuarios objetivo	Valor agregado
Análisis automatizado de datos cualitativos	Procesa y clasifica información cualitativa proveniente de diversos formatos (PDF, matrices, propuestas abiertas).	Funcionarios de SPPPSE, SPI y OPDC	Ahorra tiempo, reduce procesos manuales y mejora la precisión del análisis.
Generación de resúmenes personalizados	Elabora resúmenes ajustados a políticas, indicadores o preguntas específicas, con ranking de palabras clave y trazabilidad.	Funcionarios de SPPPSE, SPI, OPDC	Facilita el acceso rápido a información relevante para la toma de decisiones.
Clasificación temática y extracción de entidades	Aplica modelos de IA para categorizar información y extraer datos clave de manera automática.	Analistas de datos y coordinadores técnicos	Optimiza el análisis de grandes volúmenes de datos y asegura coherencia en los resultados.
Búsqueda semántica con RAG	Permite búsquedas basadas en significado y no solo en palabras clave exactas, mejorando la recuperación de información.	Funcionarios de todas las subsecretarías	Facilita encontrar información relevante de manera más eficiente y precisa.
Visualización de resultados	Dashboards interactivos y reportes personalizados para visualizar métricas y análisis cualitativos.	Funcionarios y directivos de SDP	Mejora la interpretación de datos y facilita el seguimiento del progreso de políticas y proyectos.
Integración con sistemas existentes	Se conecta con plataformas como SEGPLAN, OneDrive y SharePoint para facilitar la colaboración y gestión de datos.	Usuarios de SDP que requieren interoperabilidad entre sistemas	Asegura un flujo de información continuo y evita la duplicidad de procesos.
Seguridad y autenticación centralizada	Implementa autenticación con LDAP y cifrado TLS/SSL para proteger el acceso y garantizar	Todos los usuarios autorizados de SDP	Protege la información sensible y asegura que solo usuarios autorizados accedan a los datos.

	seguridad de la información.		
Soporte en almacenamiento e infraestructura	Validación y mantenimiento de la infraestructura de almacenamiento (PostgreSQL y ChromaDB) y despliegue en contenedores.	Equipo de IT de la entidad	Garantiza el correcto funcionamiento, escalabilidad y disponibilidad de la solución a largo plazo.
Identificación de riesgos y estrategias de mitigación	Detección de inconsistencias en los datos, problemas de interoperabilidad y limitaciones en aportes ciudadanos. Se proponen estrategias como estandarización de procesos, capacitaciones, evaluación de infraestructura y controles de calidad automatizados.	Equipo técnico, coordinadores y analistas de datos	Mejora la calidad y fiabilidad del análisis, previene interrupciones y garantiza una solución robusta y eficiente.

#5 EN INNOVACIÓN

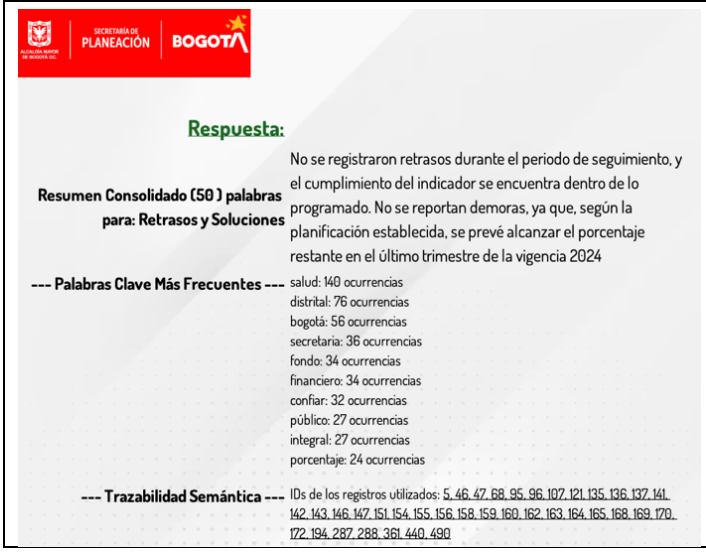
RETO DE CIUDAD 1

DATOS CON HISTORIA

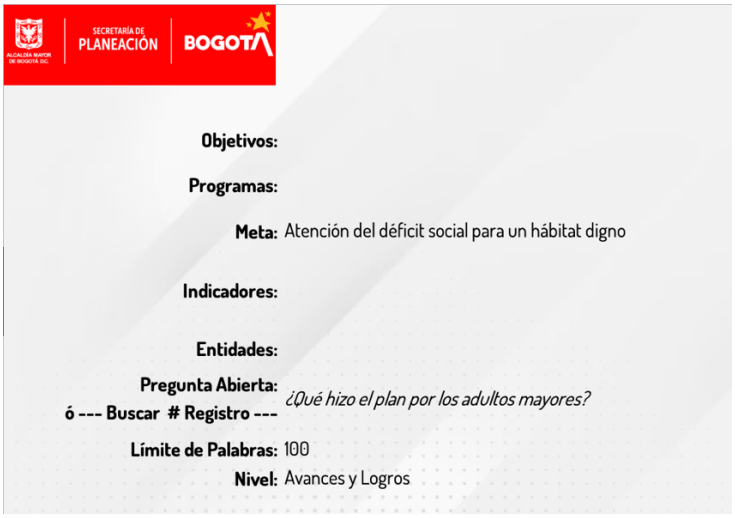
Descripción de la interfaz:

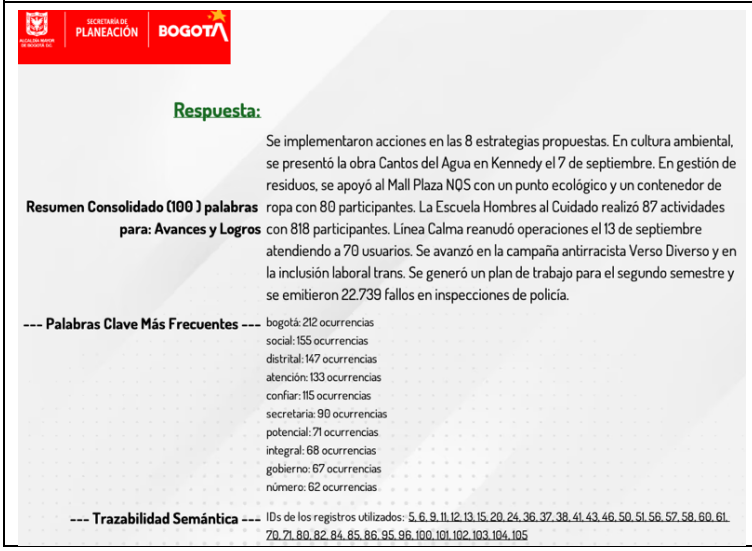
Mostraremos un ejemplo para el caso de la Subsecretaria de Inversión (SPI) para mostrar lo mencionado en la sección Usuarios y necesidades:

1. Caso de Validación de un Objetivo específico:

	**Propósito:** Permitir que el usuario use cualquiera de los filtros para encontrar el resumen según su búsqueda **Elementos Claves:** Puede filtrar por Objetivos, Programas, Metas, Indicadores o Entidades. Ya para el resumen especifico se puede escoger 2 Niveles **Acciones Permitidas:** Es posible seleccionar uno o varios de los elementos claves, solo en el nivel inferior se permite escoger entre "Retrasos y Soluciones" o "Avances y Logros". De igual forma el usuario puede limitar el tamaño de resumen por número de palabras **Acciones Realizadas:** **Objetivo Seleccionado:** *Bogotá confía en su bien-estar.* **Límite de Palabras:** 50 palabras. **Nivel de Visualización:** Retrasos y Soluciones
	**Resultados Esperados:** 1. Se generará un **resumen conciso** (máximo 50 palabras) que destacará los retrasos relacionados con el objetivo de bienestar. 2. Además, se incluirá un **ranking de las 10 palabras clave más relevantes** **3.** Una **trazabilidad de los indicadores utilizados** para la consulta con la funcionalidad de ir a cada uno de ellos desde la interface.

2. Caso de Pregunta Abierta combinada con filtros:

<table>
<tr>
<td>
</td>
<td>Propósito: Permitir que el usuario pese a seleccionar un filtro particular pueda hacer una pregunta especifica respecto a esos registros
Elementos Claves: Puede filtrar por Objetivos, Programas, Metas, Indicadores o Entidades. Y para el resumen especifico podrá escoger 2 Niveles
Acciones Permitidas: Es posible seleccionar uno opción entre Objetivos, Programas, Metas, Indicadores o Entidades, y en nivel inferior solo se permite escoger entre “Retrasos y Soluciones” o “Avances y Logros”. De igual forma el usuario puede limitar el tamaño de resumen por número de palabras

Acciones Realizadas:
Meta Seleccionada:
Atención del déficit social para un hábitat digno
Límite de Palabras:
100 palabras.
Nivel de Visualización:
Avances y Logros</td>
</tr>
<tr>
<td>
</td>
<td>Resultados Esperados:
1. Se generará un resumen conciso (máximo 100 palabras) que destacará los Avances y Logros relacionados con la Meta seleccionada.

2. Además, se incluye un ranking de las 10 palabras clave más relevantes

3. Una trazabilidad de los indicadores utilizados para la consulta con la funcionalidad de ir a cada uno de ellos desde la interface</td>
</tr>
</table>

3. Caso consultas de registros para trazabilidad:

<table>
<tr>
<td>
</td>
<td>Propósito: Permitir que el usuario pueda buscar registros específicos solo con el número de consecutivo
Elementos Claves: Solo debe incluir el número respectivo en el campo de pregunta abierta
Acciones Permitidas: Si recibe números realizara esta búsqueda de registros, pero si se escribe texto realizara el resumen global de dicha pregunta genérica

Acciones Realizadas:.
Selección de Registro: 1</td>
</tr>
<tr>
<td>
</td>
<td>Resultados Esperados:
1. Indicara cual registro se revisa

2. Proporcionará todos los componentes principales de dicho registro como: Objetivos, Programas, Metas, Indicadores o Entidades

3. Incluirá los dos niveles de respuesta de las entidades: “Retrasos y Soluciones” o “Avances y Logros”

Resultados Esperados:
Esta es la misma pantalla de respuesta que se recibirá cuando en los casos anteriores se seleccione dentro de la trazabilidad “semántica” un registro especifico de la lista</td>
</tr>
</table>

Requerimientos técnicos:

Categoría	Requerimientos identificados
Hardware (Especificaciones mínimas de equipos y servidores)	- Servidores: - Procesador: Intel Xeon o equivalente (mínimo 8 núcleos) - Memoria RAM: 32 GB (recomendado 64 GB) - Almacenamiento: 1 TB SSD para bases de datos y 500 GB adicionales para embeddings - GPU para entrenamiento de modelos (NVIDIA RTX 3090 o equivalente) - Red: Conexión Gigabit Ethernet (mínimo 1 Gbps)
Software (Herramientas, plataformas y librerías necesarias)	- Bases de datos: PostgreSQL, ChromaDB - Frameworks de IA/NLP: HuggingFace Transformers, SpaCy, SentenceTransformers, LlamaIndex - Procesamiento: Apache Tika, PyPDF2, Apache Spark - Despliegue: Docker, Kubernetes - Visualización: React, Reflex, JasperReports, - Integración: APIs REST, SEGPLAN, OneDrive, SharePoint - Seguridad: LDAP Active Directory, TLS/SSL
Recursos humanos	- Científico de datos para desarrollo de modelos de IA y análisis cualitativo. - Desarrollador full-stack para implementar visualizaciones y dashboards. - Administrador de sistemas/DevOps para gestión de servidores, contenedores y despliegues en la nube. - Equipo de IT para soporte en almacenamiento, infraestructura y seguridad. - Capacitadores para formar a los usuarios finales en el uso de la solución.
Otros (Infraestructura, conectividad, capacitación)	- Infraestructura en la nube: Oracle Cloud Infrastructure (OCI) para almacenamiento distribuido y escalabilidad. - Conectividad: Red de alta velocidad para garantizar el acceso y procesamiento eficiente. - Seguridad: Implementación de RBAC (control de acceso basado en roles) y cifrado TLS/SSL. - Capacitación: Sesiones de formación y documentación técnica para usuarios finales y equipo de soporte.