#5 EN INNOVACIÓN

RETO DE CIUDAD 1
DATOS CON HISTORIA

DOCUMENTO DESCRIPTIVO DE LA SOLUCIÓN

Reto de Ciudad 1: Datos con Historia

1. Información del equipo postulante

Título de la solución: **PROSA - V**
Nombre del equipo o razón social: **Publiqly SAS**

Integrantes y roles:

Nombre completo	Rol en el equipo	Formación	Experiencia relevante
Karen Aparicio	Coordinador	Politóloga especializada en economía.	Emprendedora en el sector GOVTECH y CITIZEN TECH. Experta en políticas públicas, investigación cuantitativa y cualitativa. Gerencia y dirección de proyectos en el sector público, cooperación internacional y Banca Multilateral. Ganadora DATAJAM. Transparencia USAID 2022 Ganadora SUMATECH. FHI360 USAID 2023. Ganadora HACK CORRUPTION LATAM. USAID 2023.

Natalia Sanabria	Arquitecta de Solución	Ingeniera de sistemas y computación	Ingeniera de sistemas con experiencia en el diseño de soluciones para aplicaciones basadas en datos. Experiencia con contenerización, ETLs, Cloud computing en Azure y GCP y aplicaciones que hacen uso de inteligencia artificial. Ganadora Hackathon para un Periodismo por la Paz 2022
Jesus Salazar	Ingeniero DevOps	Ingeniero de Sistemas	Profesional DevOps con amplia experiencia en tecnología, infraestructura, automatización de procesos, scripting e inteligencia artificial. Especializado en plataformas como AWS, con conocimientos adicionales en Azure y GCP, y habilidades sólidas en integración y entrega continua (CI/CD), Kubernetes, Docker, redes, seguridad e infraestructura como código (IaC). Destacado por su enfoque orientado a soluciones y su compromiso con el aprendizaje continuo, lo que garantiza la implementación de prácticas innovadoras y efectivas en cada proyecto.
Esteban Ortiz	Científico de datos	Ingeniero de Sistemas y Computación con Maestría en Ingeniería de la Información	Científico de Datos con una sólida experiencia en el desarrollo de soluciones basadas en aprendizaje automático y modelos de lenguaje natural (ML y LLM). En su rol como Ingeniero de Machine Learning, lideró el diseño de algoritmos

			avanzados para la segmentación y cuantificación de isquemias cerebrales, aplicando redes neuronales profundas y procesamiento de imágenes médicas. Adicionalmente, ha trabajado en proyectos integrando modelos de datos en plataformas web y diseñando tableros de control para análisis de desempeño utilizando herramientas como Power BI y DAX. Su dominio de TensorFlow, PyTorch, Python y técnicas modernas de aprendizaje profundo respalda su capacidad para implementar soluciones innovadoras.
Juanpablo Barriga Alvarez	Analista de datos	Ingeniero de Sistemas y Computación	Como Analista de Inteligencia Artificial, ha desarrollado herramientas en Python para automatizar la gestión documental y actualizar bases de datos mediante scrapers, además de integrar APIs para facilitar flujos de datos entre CRMs. Tiene experiencia en innovación y en la implementación de tecnologías basadas en inteligencia artificial para optimizar procesos corporativos. En su rol actual, ha liderado transformaciones digitales que incluyen el uso de modelos de IA y herramientas automatizadas. Como Analista de Datos, ha diseñado dashboards y métricas de desempeño, facilitando el análisis de

			datos y la visualización de resultados clave. He trabajado con equipos de diversas áreas para definir y alinear OKRs, lo que ha permitido establecer indicadores de rendimiento claros y medibles para el seguimiento de objetivos estratégicos. Ganador DATAJAM. Transparencia USAID 2022 Ganador Hackathon para un Periodismo por la Paz 2022 Ganador Code Fest Ad Astra 2022
Cesar Cruz	Desarrollador Front End	Ingeniero Electrónico	Desarrollador y diseñador web enfocado en crear experiencias modernas y únicas para la web. Apasionado por las tareas creativas, gracias a una filosofía de trabajo que prioriza la innovación y la excelencia.

2. Resumen ejecutivo

La Secretaría Distrital de Planeación enfrenta serios desafíos en la recolección, clasificación y análisis de información cualitativa, crucial para cumplir con sus funciones misionales. La gran cantidad de datos generados, las limitaciones de los sistemas de información existentes y la falta de procesos automatizados dificultan el uso eficiente de la información, impactando la formulación y seguimiento de políticas públicas, la eficiencia operativa y la participación ciudadana en la toma de decisiones estratégicas.

Para abordar este problema, se propone **PROSA-V**, una solución tecnológica integrada y escalable que optimiza todo el ciclo de vida de los datos cualitativos. **PROSA-V** consta de dos componentes principales:

- PROSA (Plataforma para la Recolección y Organización Sistematizada de Datos): Un motor ETL basado en tecnologías como Apache Kafka y Apache Spark, diseñado para la ingesta, preprocesamiento, clasificación automática y almacenamiento estructurado de datos en

bases relacionales (PostgreSQL), NoSQL (MongoDB) y vectoriales (Chroma). Utiliza una arquitectura medallion para garantizar la calidad y trazabilidad de los datos, empleando modelos ligeros de inteligencia artificial como LLAMA 3.2 para su enriquecimiento semántico.

- VISTA (Visualización Interactiva y Semántica para Toma de Acción): Una interfaz intuitiva desarrollada en Angular y respaldada por FastAPI que facilita la consulta, visualización y generación de reportes. Incluye dashboards interactivos, búsquedas semánticas avanzadas y la capacidad de automatizar informes estratégicos.

El desarrollo de **PROSA-V** se estructurará en tres etapas principales:

1. Discovery: Identificación de requisitos funcionales y técnicos, modelado de la estructura de datos y definición de los componentes base del sistema.
2. Development: Implementación de las funcionalidades principales de PROSA y VISTA, garantizando una integración robusta de capacidades transversales como la gestión de flujos de datos y el análisis semántico.
3. Deployment: Refinamiento de la solución mediante pruebas exhaustivas de funcionalidad, rendimiento y seguridad, con su posterior despliegue en un entorno de producción.

El stack tecnológico incluye herramientas de última generación como MinIO, PostgreSQL, MongoDB y Chroma para almacenamiento. Además, incorpora Apache Kafka y Spark para la gestión de datos en tiempo real y procesamiento distribuido. Python, como lenguaje principal, asegura la versatilidad y eficiencia en el desarrollo de la solución.

El objetivo general es implementar PROSA-V en un plazo de 3 meses, garantizando trazabilidad, calidad de los datos y generación de insights estratégicos. Esto permitirá mejorar la eficiencia operativa, fortalecer la participación ciudadana y facilitar el diseño de políticas públicas basadas en evidencia.

PROSA-V no solo ofrece una solución tecnológica integral, sino que también responde a las necesidades crecientes de modularidad y escalabilidad de la Secretaría. Su capacidad para adaptarse a volúmenes crecientes de datos y nuevos requerimientos asegura su sostenibilidad y relevancia en el futuro, posicionándose como una herramienta clave para una planificación urbana más eficiente, inclusiva y estratégica.

3. Identificación del problema

Contexto y problemática:

La Secretaría Distrital de Planeación enfrenta dificultades para la recolección, categorización y análisis de información cualitativa que es esencial para cumplir con sus funciones misionales. La gran cantidad de datos generados, las limitaciones de los sistemas de información existentes y la ausencia de procesos automatizados obstaculizan el uso efectivo de la información cualitativa. Estas limitaciones afectan la eficiencia operativa de la Secretaría, limitan el aprovechamiento de los datos en la formulación y seguimiento de políticas públicas y reducen la capacidad institucional para asegurar una participación ciudadana significativa en la toma de decisiones estratégicas.

Tres dependencias de la Secretaría Distrital de Planeación enfrentan problemas relacionados con la gestión de información:

Seguimiento y monitoreo de políticas públicas: La Dirección de Formulación y Seguimiento de Políticas Públicas de la Subsecretaría de Políticas Públicas y Planeación Social y Económica supervisa semestralmente 51 políticas públicas que comprenden más de 3,400 productos, cada uno asociado a indicadores cuantitativos y justificaciones cualitativas. Este proceso implica validar información enviada por múltiples entidades distritales, lo que resulta en cuellos de botella debido al análisis manual. La ausencia de un sistema que automatice la validación y estructuración de esta información compromete la generación de alertas tempranas y dificulta la identificación de patrones en la implementación de políticas.

Seguimiento y monitoreo de proyectos de inversión. La Dirección Distrital de Programación, Seguimiento a la Inversión y Plan de Desarrollo de la Subsecretaría de la Planeación de la Inversión realiza a través del SEGPLAN el seguimiento de 425 metas e indicadores asociados a los proyectos de inversión del PDD. Sin embargo, la falta de estandarización en los reportes cualitativos enviados por las entidades ejecutoras, sumado a la naturaleza no estructurada de los datos dificulta el monitoreo efectivo y la elaboración de informes de seguimiento. Esto afecta la capacidad gerencial de la Secretaría para tomar decisiones estratégicas oportunas, identificar riesgos o ajustar la ejecución de proyectos.

Clasificación y análisis de propuestas ciudadanas. La Oficina de Participación y Diálogo Ciudadano recopila aportes ciudadanos en las diferentes etapas del ciclo de planeación y a través de diferentes mecanismos. Por ejemplo, durante procesos participativos como la formulación del Plan Distrital de Desarrollo (PDD) o el Plan de Ordenamiento Territorial (POT), Chatico registró 148,129

propuestas de 41,832 ciudadanos para el PDD, muchas de ellas en formatos abiertos y de baja calidad. Estas propuestas, provenientes de la ciudadanía en general, suelen carecer de estructura y presentan problemas como ambigüedad, errores de redacción y diversidad en los temas abordados. La falta de herramientas automatizadas para organizar y analizar esta información genera retrasos significativos en su incorporación a los diferentes planes.

Usuarios y necesidades:

Dependencia	Perfil del Usuario	Necesidades específicas identificadas	Casos de uso
Subsecretaría de Políticas Públicas y Planeación Social y Económica	Funcionario de la Dirección de Formulación y Seguimiento de Políticas Públicas	1) Categorización de los reportes cualitativos 2) Generación automatizada de aportes a la ciudad de los productos a los objetivos definidos en el plan de desarrollo y ODS.	**Inicio de sesión** **Consulta Estado de la Información** ● Consulta de repositorios de información ● Cargue de archivos ● Consulta de alertas de calidad de información **Consulta de Alertas de Calidad de la Información** ● Consulta, consolidación y descarga de reporte. **Consulta de reportes estructurados de información cualitativa** ● Consulta y filtra la información por categorías relevantes de acuerdo al usuario. ● Visualización del listado de las fuentes de información. ● Descarga de reporte.
Subsecretaría de la Planeación de la Inversión	Funcionario de la Dirección Distrital de Programación, Seguimiento a la Inversión y Plan de Desarrollo	1) Debería generar las principales tendencias de la información cualitativa brindada por otras áreas de la SDP y las entidades distritales. 2) Control de calidad de información garantizando la coherencia de las temáticas tratadas.	
Oficina de Participación y Diálogo Ciudadano	Funcionario de la Oficina	1) Garantizar la recolección de insumos cualitativos, que vaya más allá de agrupar categorías	

		temáticas. 2) Permitir la trazabilidad de los aportes, para que la ciudadanía sepa qué dijo, y qué efectivamente se pudo incorporar en la toma de decisiones. 3) Facilitar el intercambio de información al interior de la entidad, creando un entorno colaborativo que mejore la toma de decisiones.	**Consulta de chat bot** ● Consulta y descarga de respuesta.

4. Objetivos de la solución

Objetivo general:

Diseñar e implementar **PROSA-V** en un plazo de 3 meses, como una solución tecnológica integrada y escalable, que optimice la recolección, clasificación, análisis y visualización de información cualitativa para las dependencias de la Secretaría Distrital de Planeación. Este sistema garantizará la trazabilidad y calidad de los datos, generando insights estratégicos que faciliten el seguimiento a políticas públicas, mejoren la eficiencia operativa y fortalezcan la incidencia de la participación ciudadana en la toma de decisiones.

Objetivos específicos:

- Identificar y documentar claramente los requisitos funcionales y técnicos, modelar la estructura de datos y definir los componentes base del sistema PROSA-V, asegurando una base sólida para su desarrollo.
- Implementar las funcionalidades principales de los componentes de PROSA - V, incluyendo la recolección, clasificación y análisis de datos, así como la visualización interactiva;

garantizando la integración de capacidades transversales que aseguren la cohesión del sistema.

- Refinar la solución mediante pruebas exhaustivas de funcionalidad, rendimiento y seguridad, para posteriormente desplegar el sistema en un entorno de producción, asegurando su operatividad y efectividad.

5. Descripción de la solución tecnológica

Arquitectura de la solución:

La solución **PROSA -V** está compuesta por dos componentes principales que trabajan de manera integrada para gestionar el ciclo completo de los datos, desde su ingesta y procesamiento hasta su presentación y análisis. Cada componente desempeña un rol específico dentro del sistema, asegurando modularidad y escalabilidad:

1. **PROSA (Plataforma para la Recolección y Organización Sistematizada de Datos):**
   Es el motor ETL que gestiona todo el ciclo de vida de los datos, desde la ingesta y el preprocesamiento hasta su almacenamiento estructurado y enriquecimiento mediante inteligencia artificial.

   PROSA implementa una arquitectura medallion para garantizar la trazabilidad y la calidad de los datos, utilizando tecnologías como Kafka, Spark y MinIO para procesamiento distribuido y almacenamiento.

2. **VISTA (Visualización Interactiva y Semántica para Toma de Acción):**
   Es la interfaz de consumo de datos, diseñada para facilitar la interacción con los usuarios finales. Proporciona una interfaz para consultar, visualizar y generar insights a partir de los datos procesados.

   Permite realizar búsquedas semánticas avanzadas, visualizar información relevante a través de dashboards interactivos y automatizar la generación de informes estratégicos.

Ambos componentes están diseñados bajo principios de modularidad y escalabilidad, permitiendo que la solución se adapte a volúmenes crecientes de datos o a nuevos requerimientos en el futuro. Asimismo, su naturaleza opensource permite su integración con los sistemas actuales de la secretaría (ya sea on-prem o desde la nube) al autenticar a los usuarios con las librerías del

directorio activo y garantiza la capacidad de la solución de resolver los problemas sin limitaciones asociadas a licencias o costos de software.

La primera versión de **PROSA – V** abordará:

- La carga y preprocesamiento de datos desde diferentes fuentes.
- La clasificación automática y almacenamiento en bases de datos estructuradas y vectoriales.
- La creación de un entorno interactivo para consultas, visualización y generación de reportes.

Stack tecnológico:

El stack tecnológico propuesto se seleccionó en función de su capacidad para cumplir con los objetivos del proyecto, teniendo en cuenta factores como escalabilidad, rendimiento, costo, facilidad de implementación y alineación con las necesidades del cliente.

Categoría	Tecnología(s) empleadas	Justificación
Framework Web	Frontend: Angular Backend: FastAPI	Angular es un framework web robusto y completo. Su enfoque modular facilita la mantenibilidad y la modularidad de la solución. Asimismo, tiene una alta compatibilidad con herramientas de terceros. FastAPI es un framework ligero y rápido para crear APIs RESTful. Tiene alta compatibilidad con Python, el lenguaje principal del proyecto. Cuenta con soporte para validación automática de datos usando Pydantic. Además la documentación es auto-generada (OpenAPI), lo que facilita la comprensión y uso por parte de terceros.
Lenguaje de Programación	Python	Python destaca por su simplicidad, extensa biblioteca y compatibilidad con tecnologías modernas, lo que acelera el desarrollo y facilita la integración en arquitecturas escalables. Su versatilidad, comunidad activa y soporte para múltiples paradigmas lo

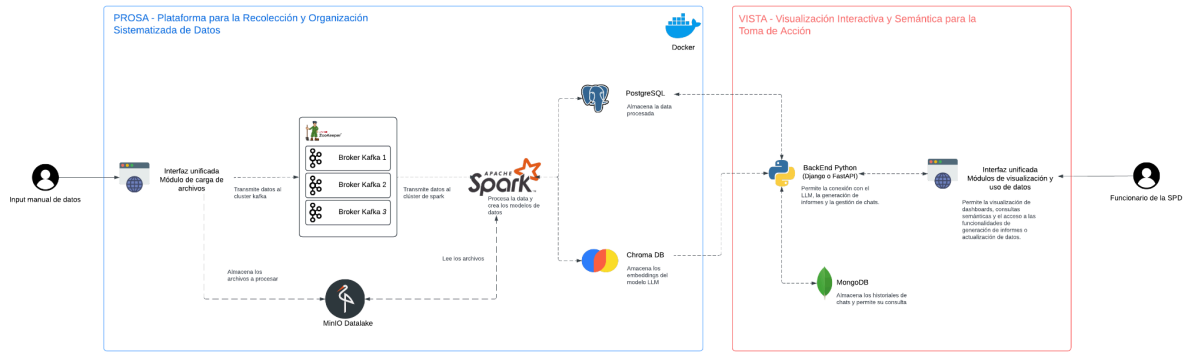
		convierten en una opción robusta para proyectos de alta exigencia técnica.
Base de Datos	Data lake: MinIO Base de datos relacional: PostgreSQL Base de datos NoSQL: MongoDB Base de datos vectorial: Chroma	MinIO es compatible con el protocolo S3, lo que permite futuras migraciones a servicios en la nube. Permite implementar una arquitectura medallion (bronce, plata y oro) para la gestión de datos. Es de código abierto y fácil de integrar en un entorno local. PostgreSQL es una base de datos relacional robusta, confiable y de código abierto. Tiene capacidad para manejar relaciones complejas entre datos estructurados. Cuenta con amplio soporte en herramientas de análisis y visualización como Power BI y Grafana. Chroma está especializado en la gestión de datos vectoriales, lo que permite búsquedas basadas en similitudes. Tiene soporte nativo para integrarse con modelos LLM como LLAMA o GPT. Es una solución ligera y eficiente, con una curva de aprendizaje rápida
Modelos de IA/NLP	LLAMA 3.2 (Modelos ligeros de 3B de parámetros)	Se eligió LLAMA 3.2 por su licencia de código abierto, su capacidad de generar respuestas precisas y su compatibilidad con entornos de recursos limitados. Además, se planifica actualizar continuamente a futuras versiones de LLAMA para mejorar el desempeño del sistema RAG a medida que se lancen nuevas versiones. Tienes flexibilidad para entrenar o ajustar el modelo según necesidades futuras.
Otra(s)	Componente de ETL Batch y Streaming: Apache Kafka Componente de procesamiento de datos: Apache Spark	Kafka favorece una arquitectura basada en eventos que asegura un desacoplamiento eficiente entre componentes. Tiene capacidad de manejar tanto flujos de datos en streaming como en batch. Cuenta con alta escalabilidad y soporte para múltiples productores y

		consumidores. Es una solución madura y ampliamente adoptada en la industria. Spark tiene alta eficiencia en el procesamiento de grandes volúmenes de datos. Cuenta con compatibilidad con múltiples formatos de datos, incluyendo datos no estructurados. Integración con Kafka para procesamiento en tiempo real. Soporte para librerías como MLlib y PySpark, que permiten el uso de modelos de machine learning. Prometheus y Grafana. Prometheus es ideal para recopilar métricas en tiempo real y generar alertas. Mientras que, Grafana permite la creación de dashboards interactivos y fáciles de usar. Ambas herramientas son de código abierto y ampliamente adoptadas en la industria, lo que asegura soporte a largo plazo.

Procesamiento y gestión de datos:

Flujo de datos entre componentes

Figura 1. Diagrama de flujo de datos



El sistema sigue un flujo definido que asegura que los datos se manejen de manera eficiente y organizada. Este flujo se puede resumir en las siguientes etapas:

- Ingesta de Datos a través de una interfaz de carga de archivos
- Procesamiento en PROSA:
    - Los datos son procesados, limpiados y categorizados a través de un proceso ETL que se detalla más abajo, utilizando reglas de negocio y modelos LLM.
- Consumo a través de VISTA:
    - Los datos categorizados son presentados en dashboards interactivos alimentados por PostgreSQL.
    - Las búsquedas semánticas en VISTA utilizan los embeddings almacenados en Chroma.
    - La interfaz también permite generar informes personalizados y editar datos procesados.

Flujo de datos en PROSA

Figura 2. Detalle del ETL en Prosa



PROSA sigue un flujo ETL estructurado en cuatro etapas principales:

1. Ingesta de datos
    - Los usuarios cargan archivos mediante la interfaz web unificada.
    - Los datos se almacenan en la capa Bronce del data lake (MinIO) como están, sin modificaciones.

2. Preprocesamiento
    - Spark consume los datos desde Bronce para:
        - Convertir PDFs a texto plano.
        - Validar formatos y campos básicos.
        - Identificar contenido ambiguo o de baja calidad.
        - Los datos limpios se almacenan en Silver.
3. Procesamiento completo:
    - Spark utiliza LLMs (ej., LLAMA 3.2) para:
        - Categorizar y etiquetar datos cualitativos.
        - Generar embeddings para búsquedas semánticas.
    - Los datos procesados se almacenan en:
        - PostgreSQL (datos relacionales).
        - Chroma (embeddings vectoriales).

Estos datasets estructurados alimentan dashboards, consultas y reportes en VISTA.

Estrategia de análisis de datos:

Procesamiento de Datos Estructurados y No Estructurados

La solución propuesta utilizará modelos de la familia **Llama 3.2**, que ofrecen capacidades avanzadas tanto para el procesamiento de datos textuales como para la extracción de información desde documentos en formatos no estructurados como PDF. Estos modelos serán la base para automatizar el flujo de trabajo.

1. Preprocesamiento y Limpieza de Datos:

    - Normalización del Texto: Una vez extraído el texto, se aplicarán técnicas de NLP como:
        - Tokenización y Lematización para estructurar y simplificar el texto.
        - Eliminación de Stopwords y limpieza de caracteres especiales para optimizar el análisis posterior.

2. Categorización y Análisis:

    - Clasificación Automática de Aportes: Usaremos modelos como Llama-3.2-1B-Instruct para categorizar automáticamente los datos en temáticas específicas (e.g., medio ambiente, movilidad, infraestructura).
    - Análisis Temático: Los embeddings generados por Llama ayudarán a identificar patrones y temas recurrentes a través de técnicas de clustering.

- Extracción de Entidades Clave: Mediante NER, se identificarán y clasificarán entidades como categorías, nombres de proyectos, lugares y eventos clave.

3. Generación de Consultas y Reportes:

- Los modelos se utilizarán para resumir grandes volúmenes de datos, generando resúmenes automáticos de aportes ciudadanos y reportes ejecutivos listos para la toma de decisiones.

Capacidades transversales en el ETL:

En esta solución hemos implementado tres capacidades transversales principales:

1. Autenticación y Autorización:
   - Se asegura de que solo las personas autorizadas puedan acceder al sistema.
   - Controla qué permisos tiene cada usuario (por ejemplo, cargar datos o editar reportes).
2. Monitoreo:
   - Permite supervisar constantemente el desempeño del sistema.
   - Detecta problemas como lentitud, errores o un uso excesivo de recursos.
3. Seguridad:
   - Protege las comunicaciones y datos sensibles contra accesos no autorizados o ataques.

Tabla de Capacidades Transversales

Capacidad	Componente Recomendado	Propósito
Autenticación y Autorización	Entra ID (Azure Active Directory) / LDAP Active Directory	Gestiona accesos seguros a través de OAuth 2.0.
Monitoreo	Prometheus	Recopila métricas sobre el desempeño del sistema.
	Grafana	Visualiza métricas y genera alertas configurables.
Seguridad	TLS/SASL_SSL	Cifra comunicaciones y protege conexiones.

Cómo ayudan estas capacidades en el proyecto

1. Entra ID o LDAP Active Directory:
   - Permite que las personas usen su usuario institucional para entrar al sistema.

- Evita que usuarios no autorizados carguen o vean datos sensibles.

2. Prometheus y Grafana:
    - Detectan problemas como:
        - Servicios lentos.
        - Errores en Kafka o PostgreSQL.
    - Muestran datos en dashboards interactivos para tomar decisiones rápidas.
3. TLS/SASL_SSL:
    - Protege la información al moverla entre componentes del sistema.
    - Evita accesos no autorizados a datos sensibles o comunicaciones internas.

Monitoreo y Retroalimentación:

Monitoreo del sistema. Para garantizar la estabilidad y el desempeño del sistema, se implementará un conjunto de herramientas de monitoreo que supervisen el estado de los componentes y las métricas clave en tiempo real.

Herramientas Planeadas:

1. Prometheus:
    - Se utilizará para recopilar métricas como uso de CPU y memoria en cada servicio; latencia de respuestas de FastAPI; y estado de los topics y consumidores en Kafka.
    - Permitirá generar alertas configurables para notificar posibles problemas.
2. Grafana:
    - Se conectará a Prometheus para visualizar métricas en dashboards interactivos
    - Dashboards configurados para: Estado de Kafka (lags, consumidores); rendimiento de Spark (tareas activas, tiempo de ejecución); y Estado de PostgreSQL (consultas lentas, uso de disco).

Retroalimentación del usuario:

Se planea incluir mecanismos para garantizar que las observaciones de los usuarios sean incorporadas en las futuras iteraciones del sistema.

Mecanismos Planeados:

1. Edición de Categorías y Etiquetas (VISTA):
    - Los usuarios podrán ajustar etiquetas asignadas por el LLM.
    - Los cambios serán registrados con metadatos como el usuario que los realizó y la fecha de modificación.
2. Historias de Uso:

- Cada interacción con el sistema de chat (ej., búsquedas, consultas semánticas) será registrada en MongoDB.
- Estas historias permitirán identificar patrones de uso y detectar necesidades adicionales.

3. Formularios de Retroalimentación:
   - VISTA incluirá un formulario donde los usuarios puedan reportar errores o sugerir mejoras.
   - Los reportes se almacenarán junto con logs relevantes en PostgreSQL para análisis posterior.

Consideraciones finales:

Limitaciones actuales:

1. Restricciones de Tiempo:

   - El tiempo disponible para el desarrollo (3 meses) limitará la implementación inicial a las capacidades más críticas.

2. Restricciones de Recursos:

   - La solución se diseñará para ejecutarse en contenedores Docker, optimizando los recursos, pero limitando el alcance de procesamiento en comparación con arquitecturas más avanzadas.

Escalabilidad y futuro: Aunque esta versión inicial se enfocará en las necesidades más inmediatas, el sistema está diseñado para escalar en el futuro. Algunas funcionalidades o ampliaciones que se pueden considerar, si resultan ser necesarias en el futuro, son:

1. Orquestación con Kubernetes: Permitiría un manejo dinámico de recursos y mayor capacidad de respuesta ante picos de actividad.
2. Capacidades en Streaming: Se evaluará extender Kafka para manejar flujos de datos en tiempo real si las necesidades de actualización cambian.
3. Integración con Nuevos Modelos LLM: Se adoptarán modelos más avanzados según las necesidades de la SDP.
4. Ampliación de Funcionalidades en VISTA: Se considera incluir análisis predictivo y visualizaciones más avanzadas en los dashboards.

Arquitectura detallada:
https://docs.google.com/document/d/18BIIoU-Nunr93lJ5-dnVeXhAZ3igDW4p/edit?usp=share_link&ouid=116769659083836602513&rtpof=true&sd=true

6. Metodología

El proyecto se dividirá en tres etapas principales: Discovery, Development, y Deployment. Cada etapa tiene actividades específicas que progresivamente construyen la solución final.

Fase	Actividades
Discovery Objetivo: Definir claramente los requisitos, modelar los datos y estructurar los componentes base del sistema.	1. Recolección y Análisis de Requisitos: ○ Entender las necesidades específicas de las dependencias de la SDP. ○ Identificar limitaciones de datos, políticas de seguridad y expectativas. 2. Modelamiento de Datos: ○ Diseñar modelos de datos para PostgreSQL (estructurados) y MongoDB (chats). ○ Definir el esquema de capas (Bronce, Silver, Gold) en el datalake MinIO. 3. Estructuración del Proyecto: ○ Configurar todos los componentes base en Docker Compose: ■ Kafka, MinIO, Spark, PostgreSQL, Chroma, Prometheus, Grafana, FastAPI. ○ Levantar servicios básicos y garantizar conectividad inicial. ○ Implementar integración como Entra ID para autenticación. 4. Prácticas de DevOps: ○ Configurar el repositorio con integración continua (CI/CD). ○ Documentar cómo desplegar el proyecto y contribuir al desarrollo. **Entregables al Final de la Fase:** ● Modelo de datos aprobado. ● Arquitectura técnica funcional con todos los componentes básicos levantados. ● Configuración inicial de prácticas DevOps.
Development Objetivo. Implementar las funcionalidades principales de PROSA y	1. Implementación de PROSA: ○ Módulo de ingesta: ■ Interfaz para carga de datos. ■ Publicación de jobs en Kafka y almacenamiento en Bronce. ○ Módulo de procesamiento:

VISTA, integrando capacidades transversales.	■ Spark limpia datos y pasa a Silver. ■ Spark utiliza LLMs para categorizar información y generar embeddings. ■ Almacenamiento final en Gold, PostgreSQL y Chroma. 2. Implementación de VISTA: ○ Frontend: ■ Dashboards interactivos. ■ Interfaz de búsquedas semánticas y generación de informes. ○ Backend: ■ Gestión de consultas semánticas con Chroma. ■ Generación de informes personalizados. ■ Persistencia de interacciones en MongoDB. 3. Pruebas Iniciales: ○ Pruebas unitarias y funcionales básicas de los módulos implementados. ○ Validación de integración entre componentes. **Entregables al Final de la Fase:** ● PROSA y VISTA funcionales, con capacidades básicas implementadas. ● Documentación inicial del funcionamiento de los módulos.
Deployment Objetivo. Pulir la solución, realizar pruebas exhaustivas, y desplegar el sistema en producción.	1. Pruebas Exhaustivas: ○ Pruebas funcionales completas de todos los módulos. ○ Validación de métricas de rendimiento en Prometheus. ○ Revisión de logs en Grafana para identificar errores. 2. Optimización del Sistema: ○ Ajustes según resultados de pruebas. ○ Validación final de la calidad de datos y generación de informes. 3. Despliegue: ○ Migración a un entorno de producción. ○ Documentación completa para usuarios y administradores. **Entregables al Final de la Fase:** ● Sistema desplegado en producción. ● Documentación final y guía de usuario.

7. Plan de implementación

Cronograma general de implementación:

Actividad	Mes		
	Mes 1	Mes 2	Mes 3
Fase: Discovery			
1. Recolección y Análisis de Requisitos	x		
2. Modelamiento de Datos	x		
3. Estructuración del Proyecto	x		
4. Prácticas DevOps	x		
5. Documentación y presentación de avances	x		
Fase: Development			
1. Implementación PROSA	x	x	
2. Implementación VISTA	x	x	
3. Pruebas Iniciales		x	
4. Documentación y presentación de avances		x	
Fase: Deployment			
1. Pruebas Funcionales (exhaustivas)			x
2. Optimización del Sistema			x
3. Despliegue Final			x

Presupuesto general de implementación:

El presupuesto se distribuye en un 67% en equipo, 23% en impuestos y 10% en gastos administrativos y pólizas.

Categoría	Item	Cantidad	Costo unitario	Costo total
Project Manager	Personas	1	-	$10.000.000
Equipo Desarrollador	Personas	6	-	$50.000.000
Gastos Administrativos y pólizas			-	$8.850.000
Impuestos (retefuente, reteica, iva)				$21.150.000

8. Guía de usuario y requerimientos técnicos

Principales funcionalidades:

Funcionalidad	Descripción	Usuarios objetivo	Valor agregado
Autenticación	Verificar la identidad del usuario para proporcionar acceso a la herramienta.	Todos los usuarios	Se asegura de que solo las personas autorizadas puedan acceder al sistema.
Cargue de archivos	Recolección e Ingesta de Datos	Usuarios autorizados de SPPPSE, SPI, OPDC.	Facilita la actualización de la herramienta en las condiciones actuales.
Consulta de alertas de calidad de la información	Generación de alertas de calidad y recomendaciones automáticas para la mejora de los datos.	Usuarios autorizados de SPPPSE, SPI, OPDC.	Busca incentivar una cultura alrededor de la calidad de la información en aras de mejorar los datos ingestados en la herramienta y mejorar la calidad de los reportes.
Consulta y reportes	Dashboards de consulta de información personalizados según las necesidades de los usuarios y la información disponible.	Todos los usuarios. Sin embargo el acceso a consultas puede estar segmentado según las competencias de las dependencias.	Desarrollo de informes por diferentes categorías en minutos. Facilitar el seguimiento y monitoreo de planes y compromisos.

Descargue de reportes	Descarga de las consultas realizadas en archivos editables.	Todos los usuarios	Apoya el proceso de elaboración de Informes de los usuarios y la incorporación de insights en los procesos de planeación.
ChatBot	Herramienta de consulta para preguntas abiertas.	Todos los usuarios	Facilita la búsqueda integrada en las diferentes fuentes de información con las que cuenta la herramienta.

Descripción de la interfaz:

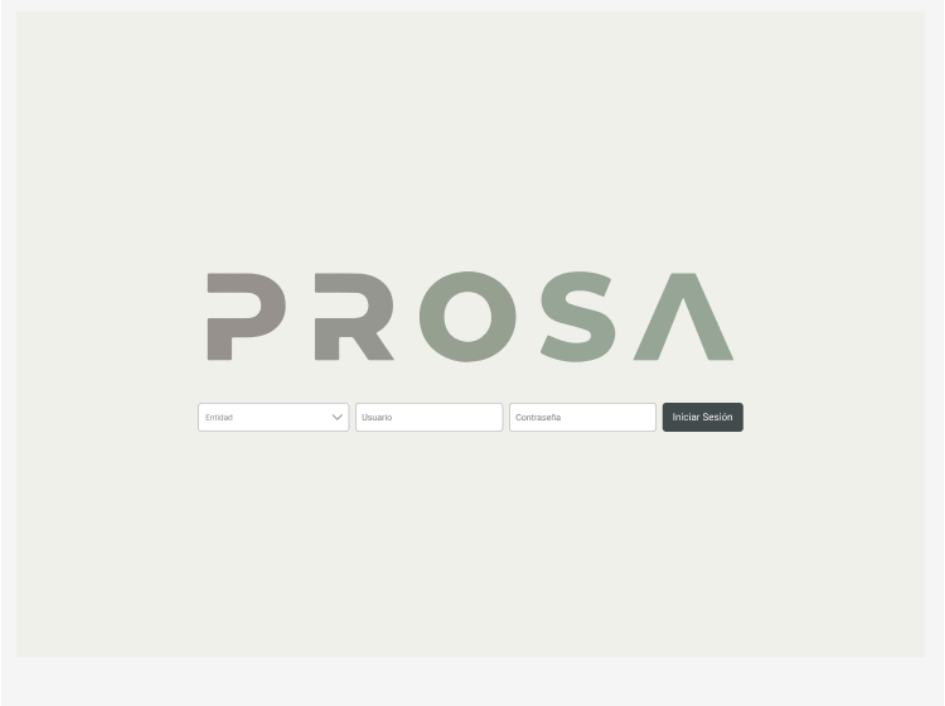
Interfaz	Objetivo	Contenido	Acciones
1. Login	Verificar la identidad del usuario para proporcionarle acceso a la herramienta	Entidad, correo electrónico y contraseña	Autenticarse en la plataforma
2. Inicio	Informar al usuario sobre el Estado de la Información a consultar y poner a su disposición las últimas consultas realizadas. Dar acceso rápido a las funcionalidades de la herramienta.	Acceso a diferentes funcionalidades Segmento 1. Última fecha de actualización de la información, número de archivos cargados, usuario que lo cargo, evaluación de calidad de la información cargada. Segmento 2. Listado de las últimas consultas realizadas.	Informarse sobre la información disponible a consultar y su calidad (esto variará dependiendo la configuración de permisos) Podrá acceder a todas las funcionalidades de la herramienta.
3. Configuración - Super admnistrador	Administrar los permisos de los usuarios en la herramienta.	Listado de usuarios con el perfil de permisos asignado. Agregar usuarios. Acceso a diferentes funcionalidades.	El super administrador podrá agregar usuarios para configurar permisos (perfil 1: cargue de documentos + acceso a alertas de calidad + consulta / Perfil 2: consulta).

			Podrá acceder a todas las funcionalidades de la herramienta. Podrá acceder a todas las funcionalidades de la herramienta.
4. Consulta de carpetas de documentos	Organizar la visualización y acceso al repositorio de archivos de consulta de la herramienta.	Carpetas de archivos. Acceso a diferentes funcionalidades.	El usuario con permisos podrá acceder a las carpetas del repositorio de información y visualizar el histórico de archivos agregado. Podrá acceder a todas las funcionalidades de la herramienta.
5. Cargue de documentos	Facilitar el cargue de los archivos de consulta de la herramienta.	Acceso al cargue de documentos. Listado de documentos recientemente cargados. Acceso a diferentes funcionalidades.	Cargar archivos. Consultar archivos. Podrá acceder a todas las funcionalidades de la herramienta.
6. Alertas calidad de la información	Visualizar la evaluación y recomendaciones de calidad realizada sobre los archivos que contienen reportes cualitativos.	Dashboard con estadísticas de calidad de los archivos y por entidades. Desempeño de calidad de las entidades en los reportes cualitativos	Consultar y consolidar las recomendaciones de mejoras de calidad por Entidad en un archivo descargable.
7. Consulta y Reportes SPPPSE	Facilitar la consulta cualitativa de los reportes de productos de las políticas públicas realizados por las Entidades.	Filtros para ajustar las consulta y segmentos de análisis de información de acuerdo a los criterios de visualización establecidos (resultados, atrasos, etc)	Consultar y descargar archivos. Podrá acceder a todas las funcionalidades de la herramienta.
8. Consulta y Reportes SPI	Facilitar la consulta cualitativa de los reportes de avance en la implementación del PDD realizados por las Entidades.	Filtros para ajustar las consulta y segmentos de análisis de información de acuerdo a los criterios de visualización	Consultar y descargar archivos

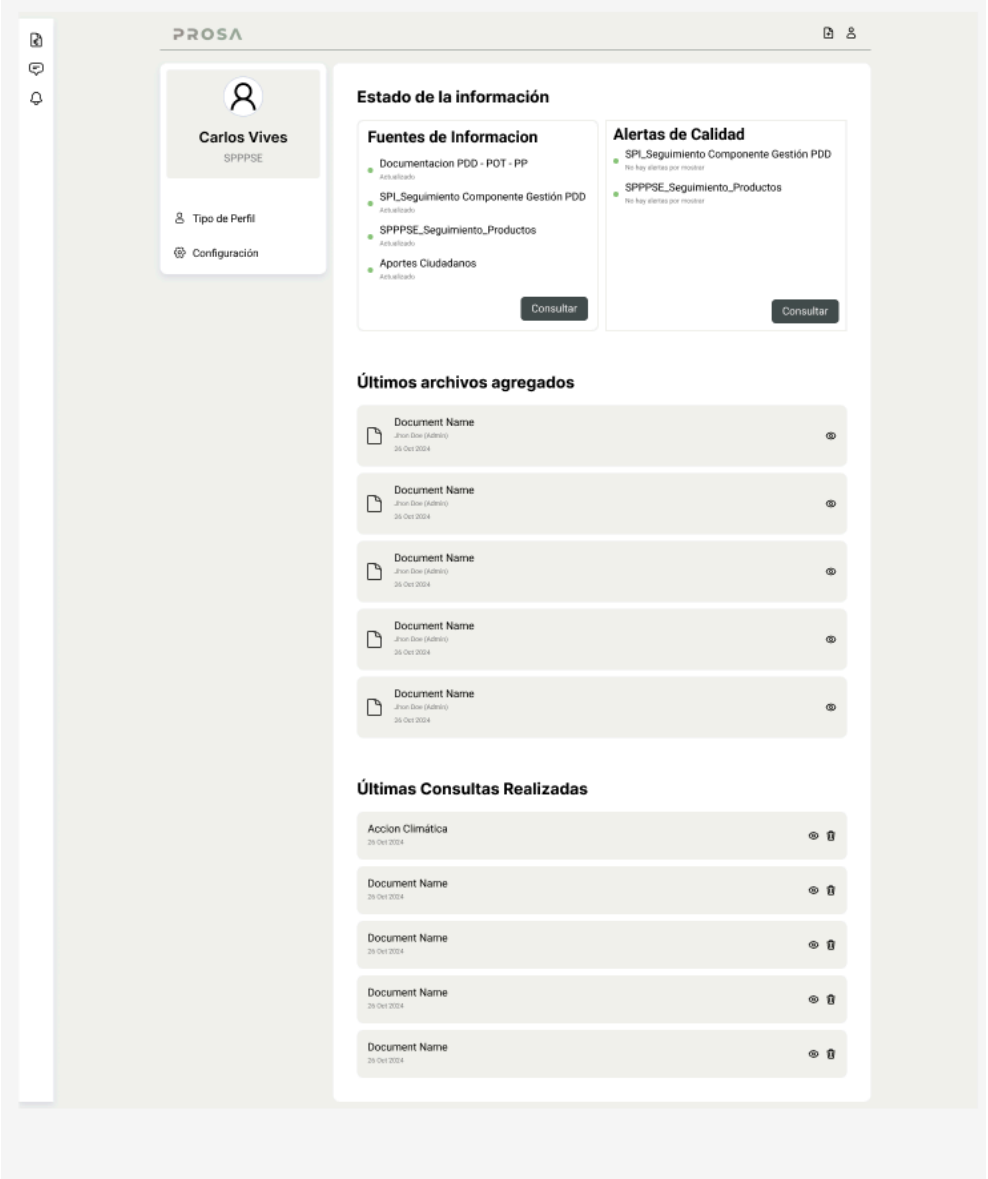
		establecidos (resultados, atrasos, etc)	Podrá acceder a todas las funcionalidades de la herramienta.
9. Consulta y Reportes OPDC (1) - Chatico	Facilitar la consulta y visualización de los aportes de Chatico (prioridades y soluciones)	Dashboard con información cuantitativa y cualitativa.	Consultar y descargar archivos Podrá acceder a todas las funcionalidades de la herramienta.
9. Consulta y Reportes OPDC (2) - POT	Facilitar la consulta y visualización de los aportes al POT (prioridades y soluciones)	Dashboard con información cuantitativa y cualitativa.	Consultar y descargar archivos Podrá acceder a todas las funcionalidades de la herramienta.
10. Consulta y Reportes OPDC (3) - Actas	Facilitar la consolidación, y consulta y visualización de los aportes en espacio de participación.	Dashboard con información cuantitativa y cualitativa.	Consultar y descargar archivos Podrá acceder a todas las funcionalidades de la herramienta.
11. Chatbot	Facilitar la búsqueda integrada en las diferentes fuentes de información con las que cuenta la herramienta.	Un buscador con un robot que le dará consejos de cómo hacer los mejores prompts de consulta.	Consultar y descargar archivos Podrá acceder a todas las funcionalidades de la herramienta.

Ejemplos de Interfaz:

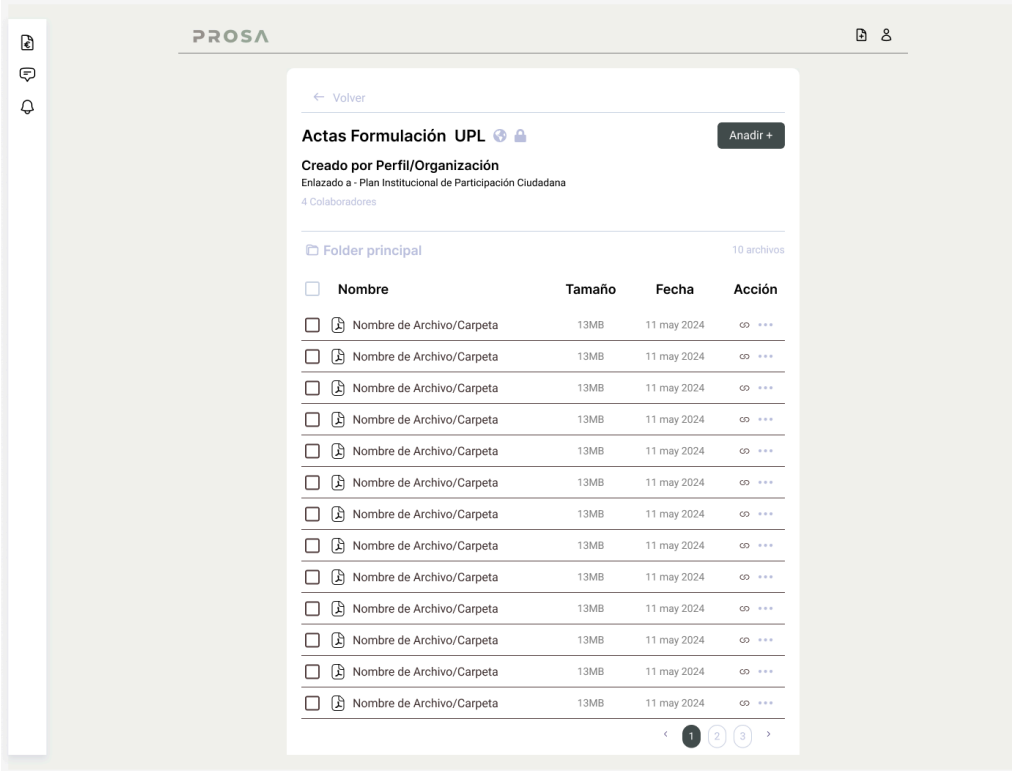
1. Login



2. Inicio

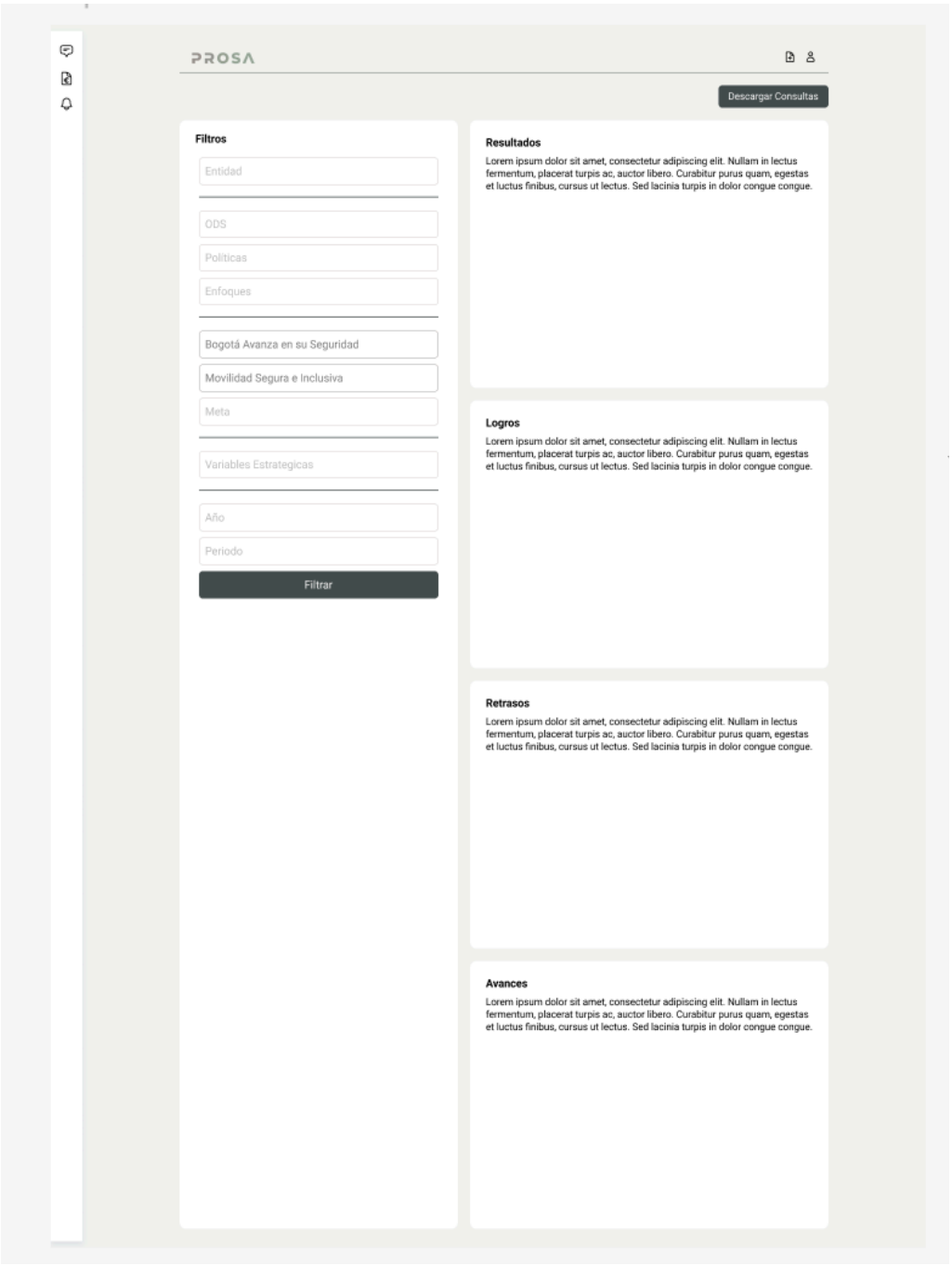


3. Cargue de documentos

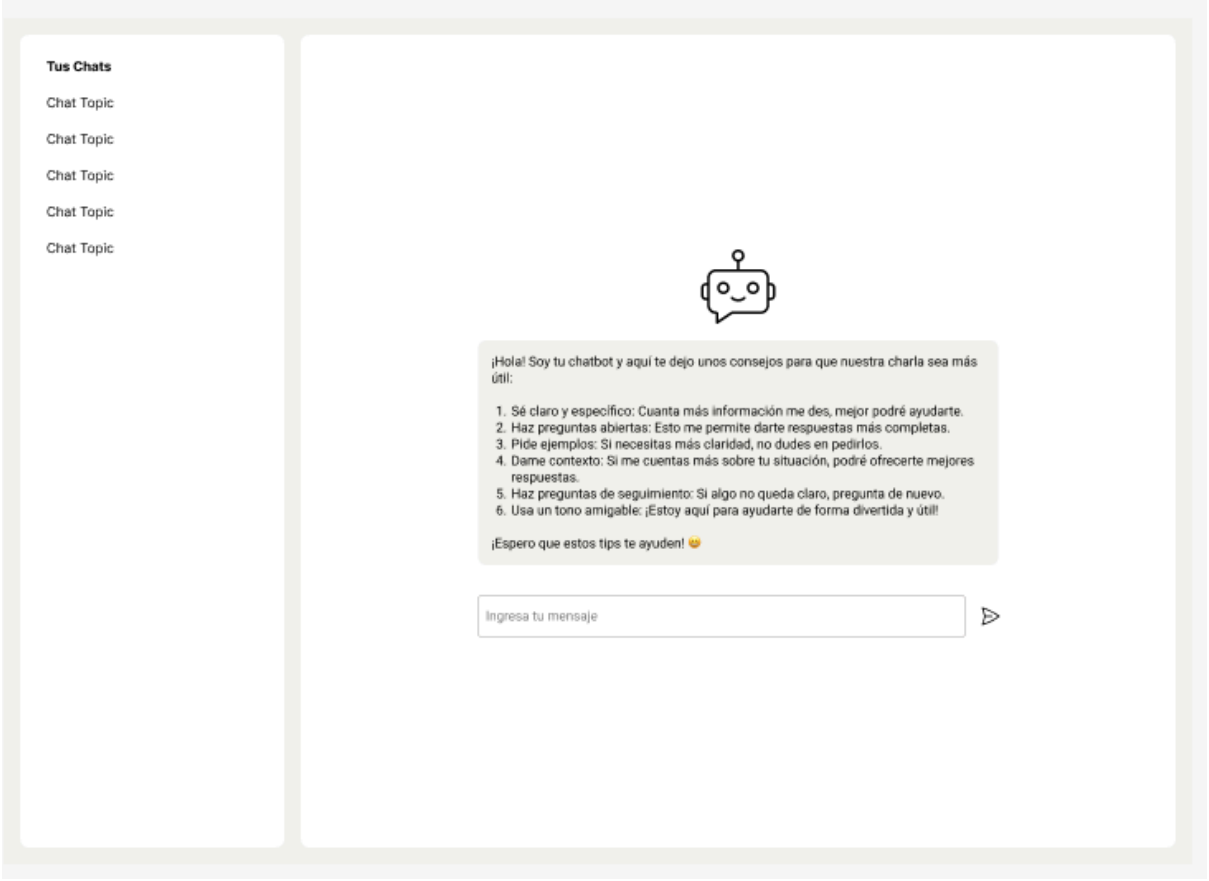


4. Consulta y reportes

La interfaz de consulta y reporte tiene variaciones en los filtros y el número de cajas de información que entrega depende de la fuente y del usuario de consulta. Para la presentación se mostrará con mayor detalle estas interfaces diseñadas por necesidad y usuario.


PROSA
Descargar Consultas
Filtros
Entidad
ODS
Políticas
Enfoques
Bogotá Avanza en su Seguridad
Movilidad Segura e Inclusiva
Meta
Variables Estrategicas
Año
Periodo
Filtrar
Resultados
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam in lectus fermentum, placerat turpis ac, auctor libero. Curabitur purus quam, egestas et luctus finibus, cursus ut lectus. Sed lacinia turpis in dolor congue congue.
Logros
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam in lectus fermentum, placerat turpis ac, auctor libero. Curabitur purus quam, egestas et luctus finibus, cursus ut lectus. Sed lacinia turpis in dolor congue congue.
Retrasos
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam in lectus fermentum, placerat turpis ac, auctor libero. Curabitur purus quam, egestas et luctus finibus, cursus ut lectus. Sed lacinia turpis in dolor congue congue.
Avances
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam in lectus fermentum, placerat turpis ac, auctor libero. Curabitur purus quam, egestas et luctus finibus, cursus ut lectus. Sed lacinia turpis in dolor congue congue.

5. ChatBot



Requerimientos técnicos:

Categoría	Requerimientos identificados
Hardware	- **Servidores**: • Procesador: 8 núcleos (mínimo). • Memoria RAM: 32 GB (mínimo). • Almacenamiento: 1 TB SSD (mínimo) para procesamiento intensivo y almacenamiento inicial. - **Estaciones de trabajo**: • Procesador: 4 núcleos. • Memoria RAM: 16 GB. • Espacio en disco: 256 GB. - **Cluster Spark**:

	• Servidor dedicado o solución cloud con capacidad de escalabilidad horizontal.
Software	- **Backend y procesamiento**: • Python 3.10 (FastAPI y PySpark). • Kafka (v3.2.0 o superior). • PostgreSQL 13. • Chroma para bases de datos vectoriales. • MinIO como datalake. - **Frontend**: • Angular 15. - **Procesamiento con LLMs**: • Ollama con modelos LLAMA3 ligeros (llama 3.2 en adelante). • Langchain (agente de búsqueda semántica) • transformers (para generación de embeddings) • sentence-transformers (para embeddings optimizados) • langchain-community (complementos adicionales para LangChain). • langchain-huggingface (integración con modelos de Hugging Face). • langchain-chroma (conexión con bases de datos vectoriales como Chroma). • chromadb (base de datos vectorial para almacenamiento y recuperación). • langchain-nvidia-ai-endpoints (integración con endpoints de IA de NVIDIA). - **Monitoreo**: • Prometheus y Grafana. - **DevOps**: • Docker y Docker Compose.
Recursos humanos para el desarrollo	- **Desarrolladores**: • 1 Backend Developer (Python, Spark). • 1 Frontend Developer (Angular). • 1 Data Engineer especializado en Kafka y PostgreSQL. - **Especialistas**: • 1 Experto en IA/LLMs para configuración y ajustes de modelos. - **DevOps**: • 1 Ingeniero DevOps para infraestructura y CI/CD. - **Analista de Calidad**: • 1 persona para validar datos y funcionalidad de la plataforma.
Otros	- **Infraestructura**: • Conexión a internet de alta velocidad (mínimo 100 Mbps). • Red interna segura para el entorno de desarrollo y producción. - **Capacitación**: • Entrenamiento básico para el personal en el uso de PROSA - V. • Capacitación técnica para manejo de LLMs y monitoreo con Grafana/Prometheus.

#5 EN INNOVACIÓN

RETO DE CIUDAD 1

DATOS CON HISTORIA