#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: ISIDataInsights

Nombre del equipo o razón social: Inversiones Gutierrez García y Cía. S en C

Integrantes y roles:

Nombre completo	Rol en el equipo	Formación	Experiencia relevante
Luis Fernando Gutierrez García	Científico de Datos	• PhD en Ingeniería Industrial en creatividad computacional. Universidad de los Andes, 2020 • Magíster en Ingeniería Industrial con profundización en estética computacional. Universidad de los Andes, 2012 • Matemáticas Puras Universidad de los Andes, 2005 • Maestro en Artes Plásticas con profundización en Medios Electrónicos. Universidad de los Andes, 2005	**Director de Tecnología e Innovación – desarrollo de:** • Plataforma MiRecarga • Biometría – Código Nacional de Policía • Plataforma ISIPoint: • Central de Procesamiento de Infracciones de Tránsito (CPIT) • Prototipo Detección de Prohibidos Parquear • Plataforma Contacto Móvil • GM3 – ERP • GM3 – iReports
Jhonatan Javier Castellanos Bravo	Desarrollador Fullstack	• Ingeniería de sistemas, Universidad Distrital • Desarrollo de aplicaciones web, Universidad Nacional	• Desarrollo Full Stack. • Implementación de microservicios. • Web Scraping. • Implementación de sistemas de gran transaccionalidad. Automatización de sistemas con modelos NLP. Implementación de sistemas escalables.

#5EN INNOVACIÓN

RETO DE CIUDAD 1

DATOS CON HISTORIA

Antonio María Escobar Torres	Director del proyecto	Ingeniero de Sistemas - UNIVERSIDAD DE LOS ANDES – Bogotá Colombia. (1985) • MÁSTER MBA- Especialidad en Gestión de Proyectos Campus Universitario Europeo (junio 2017-Feb 2018) GERENCIA DE PROYECTOS • PMP – Project Manager – Certifiied PMI – International (junio 2017) • Certificación Fundamentos ITIL 2011 Axelos – (agosto 2017) • DIPLOMADO EN GERENCIA DE PROYECTOS Politécnico Colombiano (agosto 2017) METODOLOGIA DE DESARROLLO SCRUM • Certificado SCRUM MÁSTER Scrum Institute (mayo 2017) • Team Kanban Practitioner - Team Kanban LeanKanban University Certified Kanban Training Program (Marzo 2017) • Kanban System Design - KMP Foundation I • TPM - Managing a Project Using TSM UNIX Certificación en Administración de Sistemas Unix (Job Role) –Brainbench	LENGUAJES DE PROGRAMACIÓN: - SQL Ansi 92 usados en RDBMS (Teradata, Oracle, Unify, MySQL, MS SQL Server) - C, C+, C++ - Acell 4GL BASES DE DATOS - Teradata - MS SQL Server - Oracle - MySQL - Unify 2000 - Unify IDS BUSINESS INTELLIGENCE - SAP BO - Tableu - MicroStrategy SISTEMAS OPERATIVOS - UNIX (Linux, Primo, AIX, HP-UX) - Windows

		certificación en Administración de Sistemas Linux (Job Role) – Brainbench • Administración de Unix (General), Administración de Linux (General), Terminología en Tecnología de Información, Conceptos de Networking, Administración de Servidores, Asistencia a los Clientes, Codificación en Unix Korn Shell, Habilidades de escucha a los clientes, Planeación y Recuperación de desastres Capella University, Chantilly, Virginia nov. 2003	
Juan Pablo Pájaro Hernández	Analista de datos	• **Máster en Ingeniería de Sistemas y Computación Pontificia Universidad Javeriana** Bogotá, 2019 • **Especialista en evaluación social de proyectos Universidad de los Andes** Bogotá, 2009 • **Administrador Industrial Universidad de Cartagena** Cartagena, 2007	**ICBF Oct/16 – Sept/18** **Coordinador de proyectos y analítica** Apoyar la ejecución de las evaluaciones de procesos, resultados e impactos, por medio del desarrollo de modelos estadísticos. **Secretaría de Cultura, Recreación y Deporte Jul/15 – Dic/16** **Coordinador de proyectos** Apoyar la trazabilidad, modelación y pronósticos de los datos de los proyectos de inversión, que permita mejorar la gestión. **OIM Organización Internacional para las Migraciones sept/15 – Ene/16** **Formulador de proyectos** Brindar asistencia técnica en la formulación de proyectos y políticas públicas a través de mesas técnicas y demás mecanismos. **Inversiones Gutiérrez García - Contacto Móvil Ene/13 – Abril/15** **Coordinador de desarrollo de productos en TIC** Evaluar modelos estadísticos y financieros que contribuyan al pronóstico de ingresos, costos y gastos.

			Secretaría de Cultura, Recreación y Deporte May/11 – Dic/12 **Profesional en proyectos** Garantizar el monitoreo efectivo de la gestión y administración de los recursos en los proyectos de inversión. **Programa de las Naciones Unidas para el Desarrollo - Objetivos Del Milenio Jul/09 – Dic/10** **Profesional Apoyo de Gerencias Locales** Formular proyectos estratégicos para la reducción de la pobreza en Cundinamarca. **Departamento de Planeación - Gobernación de Bolívar Ene/08 – Jun/09** **Asesor de proyectos** Coordinar la elaboración de diagnósticos socioeconómicos de las Zonas del Departamento de Bolívar.

2. Resumen ejecutivo

Gran parte de los datos disponibles provienen de fuentes estructuradas y no estructuradas como documentos técnicos, hojas de cálculo y otros archivos clave. En este proyecto, nos centraremos en la extracción de información de fuentes específicas proporcionadas, incluyendo el Plan de Desarrollo Distrital (PDD), el Plan de Ordenamiento Territorial (POT) y los programas, metas e indicadores del PDD contenidos en archivos Excel. Estas fuentes serán la base para los procesos de analítica necesarios para abordar los retos identificados.

Una vez que los datos han sido extraídos, implementaremos procesos avanzados como el reconocimiento de entidades nombradas. Este método permitirá identificar conceptos clave como nombres de personas, entidades, lugares, fechas y otros datos relevantes para los objetivos del proyecto. La información estructurada y sus metadatos serán almacenados en bases de datos especializadas, asegurando su disponibilidad para análisis posteriores.

En relación con los retos planteados, hemos diseñado estrategias que maximizan la eficiencia y los resultados:

- ***Procesamiento de políticas públicas:** Actualmente, la revisión manual de 6 políticas toma entre 1 y 2 semanas para una persona. Nuestra solución reducirá este tiempo en al menos un 60%, lo que liberará recursos para actividades estratégicas.*

- ***Análisis de actas y documentos clave:*** *Automatizamos la extracción y estructuración de información contenida en actas y documentos relevantes, permitiendo su visualización y análisis desde perspectivas más funcionales y orientadas a los objetivos del PDD.*
- ***Alineación con necesidades de los usuarios:*** *Abordaremos las expectativas de los diferentes actores mediante un enfoque analítico robusto que optimice los recursos disponibles, sin incluir inicialmente integraciones complejas como autenticación LDAP, visualización en ArcGIS o PowerBI. Esto asegura que las soluciones entregadas sean viables dentro del alcance establecido.*

Soluciones específicas para las dependencias:

- ***Subsecretaría de Políticas Públicas y Planeación Social y Económica:*** *Reconociendo la importancia de los datos para las políticas públicas de la ciudad y los desafíos asociados a su aprovechamiento, hemos diseñado una solución innovadora que facilita la categorización de aportes en variables estratégicas y la generación automatizada de contribuciones alineadas con el plan de desarrollo y los ODS. Nuestro sistema de analítica permitirá identificar errores de ortografía, información incompleta y duplicidad de aportes, definiendo la categoría correspondiente, evaluando su impacto, relacionándolo con indicadores y resultados, y generando reportes que evidencian alertas, logros y acciones de mejora.*
- ***Oficina de Participación y Diálogo de Ciudad:*** *Entendemos la importancia de la participación ciudadana y el impacto que tiene en la confianza pública. Proponemos optimizar la recolección, trazabilidad e intercambio de información mediante un sistema que depure aportes irrelevantes y evalúe dimensiones clave como relevancia estratégica, impacto, factibilidad, innovación y sostenibilidad. Esto facilitará el seguimiento de los ciudadanos a sus aportes y garantizará su relación con áreas específicas de trabajo, todo a través de una interfaz gráfica que simplifique la visualización categorizada y permita a los usuarios apropiarse del área de su interés.*
- ***Subsecretaría de Planeación de la Inversión:*** *Incluiremos un sistema de análisis de texto, impulsado por inteligencia artificial, para identificar tendencias clave en la información recolectada por las entidades distritales, clasificándose por temáticas estratégicas. Este sistema implementará un control de calidad basado en criterios definidos que validen coherencia, precisión y alineación temática, detectando inconsistencias y generando reportes sobre tendencias, alertas y recomendaciones que aseguren datos útiles y confiables para la toma de decisiones.*

Con nuestro enfoque, ofreceremos una plataforma que organiza y estructura la información mientras optimiza tiempos operativos y mejora la toma de decisiones, todo alineado a las necesidades específicas de cada dependencia.

3. Identificación del problema

#5 EN INNOVACIÓN

RETO DE CIUDAD 1
DATOS CON HISTORIA

Contexto y problemática:

Las dependencias de la Secretaría Distrital de Planeación enfrentan múltiples desafíos relacionados con el manejo de información cualitativa, clave para el cumplimiento de sus funciones misionales. Estas funciones incluyen la elaboración y seguimiento de planes estratégicos como el Plan de Ordenamiento Territorial y el Plan Distrital de Desarrollo, la evaluación de políticas públicas, la medición de impactos y el control de la inversión realizada. Sin embargo, el volumen, la diversidad y las características no estructuradas de la información representan obstáculos importantes para lograr resultados eficientes y efectivos.

Uno de los principales retos es la gran cantidad de documentos y datos no estructurados que los funcionarios deben revisar manualmente, lo que requiere un tiempo significativo y prolonga el cumplimiento de tareas críticas. La falta de estandarización en los formatos de información y la presencia de errores o inconsistencias dificultan su análisis, incrementando la carga operativa de los equipos.

Además, la SDP recibe información de múltiples fuentes y entidades distritales, cada una con niveles de detalle y formatos diferentes. Esta diversidad añade complejidad a la consolidación y análisis integral de los datos, ya que los funcionarios deben validar y depurar la información manualmente para garantizar su coherencia y alineación con los objetivos estratégicos.

Otro desafío crítico es la falta de herramientas tecnológicas adecuadas para automatizar los procesos de validación y análisis de información cualitativa. Los mecanismos actuales, aunque funcionales, no son suficientes para abordar de manera eficiente la identificación de discrepancias, la alineación temática o la evaluación de la relevancia y precisión de los datos. Esto afecta tanto la evaluación de productos como la medición de impactos y procesos participativos, donde la claridad y precisión de la información son esenciales.

En este contexto, la gestión de la información cualitativa enfrenta problemas relacionados con su calidad, extensión y diversidad, los cuales generan una mayor carga de trabajo para los funcionarios, afectando los tiempos y la eficiencia de las tareas misionales de la SDP.

Usuarios y necesidades:

(Completa la tabla identificando a los usuarios clave de cada dependencia, sus perfiles, necesidades específicas y casos de uso asociados. Es necesario identificar al menos un usuario para cada una de las tres dependencias)

Dependencia	Perfil del Usuario	Necesidades específicas identificadas	Casos de uso

Subsecretaría de Políticas Públicas y Planeación Social y Económica / Dirección de Formulación y Seguimiento de Políticas Públicas	Especialista en políticas públicas / analista de datos	La Subsecretaría busca una solución innovadora que permita: ● Clasificar los aportes en variables estratégicas alineadas con los objetivos del plan de desarrollo y los ODS. ● Automatizar la generación de contribuciones estratégicas. ● Corregir errores de ortografía. ● Detectar información incompleta. ● Eliminar duplicidades. ● Relacionar indicadores con resultados concretos. ● Identificar alertas, logros y áreas de mejora. ● Sugerir acciones para optimizar los resultados. ● Evidenciar el impacto de las políticas públicas.	casos de uso: login, inicio de sesión, visita de página inicial – reporte simplificado de tareas realizadas, Carga de Documentos, visualización procesos de negocio – documentos cargados y procesados, visualización de resultados, visualización de resultados detallados, exportar datos, cierre de sesión.

		● Relacionar los resultados con actividades específicas. Todo esto mediante un sistema basado en analítica y tecnologías avanzadas para optimizar el análisis y el seguimiento de los objetivos estratégicos.	
Oficina de Participación y Diálogo de Ciudad	Facilitadores de participación ciudadana / gestores de información	La Oficina necesita un sistema que permita: ● Mejorar la trazabilidad y el intercambio de datos relacionados con la participación ciudadana. ● Ordenar, clasificar y agrupar los diferentes tipos de aportes según criterios definidos. ● Analizar dimensiones como relevancia estratégica, impacto, factibilidad, innovación y sostenibilidad. ● Producir datos categorizados y medibles	casos de uso: login, inicio de sesión, visita de página inicial – reporte simplificado de tareas realizadas, Carga de Documentos, visualización procesos de negocio – documentos cargados y procesados, visualización de resultados, visualización de resultados detallados, exportar datos, cierre de sesión.

		que faciliten el seguimiento ciudadano y su integración en áreas de trabajo específicas. ● Permitir la visualización categorizada de la información. ● Facilitar la apropiación de los datos por parte de los usuarios según sus áreas de interés. Estas funcionalidades estarán respaldadas por análisis avanzados y algoritmos que evitarán sesgos, garantizando datos transparentes y útiles para la toma de decisiones estratégicas.	
Subsecretaría de Planeación de la Inversión Dirección Distrital de Programación, Seguimiento a la Inversión y Plan de Desarrollo	Planificadores de inversión / analista de datos	La Subsecretaría de Planeación de la Inversión requiere una solución basada en inteligencia artificial para analizar y clasificar la información recolectada por las entidades distritales según temáticas	casos de uso: login, inicio de sesión, visita de página inicial – reporte simplificado de tareas realizadas, Carga de Documentos, visualización procesos de negocio – documentos cargados y procesados, visualización de resultados,

		estratégicas. El sistema deberá: ● Identificar tendencias clave en la información y clasificarlas de manera estratégica. ● Garantizar la calidad de los datos mediante controles automatizados que validen coherencia, precisión y alineación temática. ● Detectar y reportar inconsistencias, generando alertas y recomendaciones para la toma de decisiones. ● Producir reportes detallados sobre tendencias, alertas de incoherencia y acciones sugeridas. La solución integrará análisis de datos cualitativos, categorización	visualización de resultados detallados, exportar datos, cierre de sesión.

		estratégica y herramientas de visualización, facilitando la evaluación de propuestas ciudadanas con base en criterios de relevancia estratégica, impacto social y factibilidad. Esto garantizará información confiable y útil para una toma de decisiones informada y efectiva en Bogotá.	

4. Objetivos de la solución

Objetivo general:

Implementar una solución tecnológica integral en la Secretaría Distrital de Planeación (SDP) para mejorar la calidad, consistencia y análisis de la información cualitativa proveniente de diversas fuentes. Esta solución permitirá fortalecer la toma de decisiones estratégicas alineadas con los objetivos de desarrollo urbano, optimizando tiempos de procesamiento, y reduciendo horas dedicadas a tareas manuales.

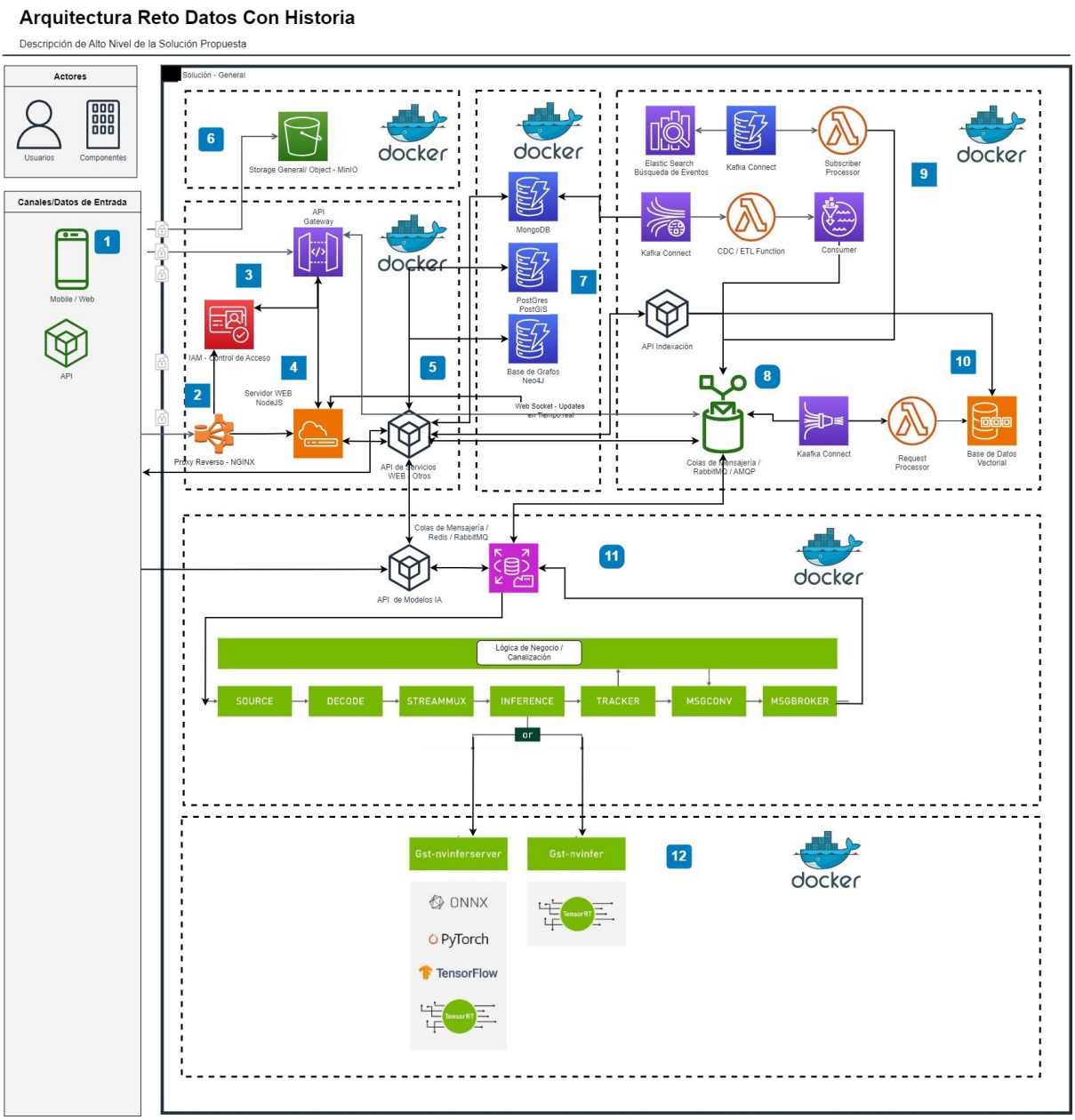
Objetivos específicos:

1. *Desarrollar un sistema que clasifique los reportes cualitativos en variables estratégicas como Acción Climática, Movilidad, Pobreza, Calidad de Educación, y Seguridad, alineados con el plan de desarrollo y los ODS.*
2. *Implementar herramientas que transforman los insumos cualitativos en contribuciones automáticas y alineadas con los objetivos estratégicos de la ciudad.*
3. *Diseñar un sistema que permita a la ciudadanía rastrear sus aportes y verificar cómo se integraron en la toma de decisiones estratégicas.*
4. *Incorporar funcionalidades de minería de datos y análisis de texto para identificar patrones clave en la información cualitativa, apoyando decisiones basadas en evidencia.*
5. *Garantizar la coherencia, precisión y claridad de los datos mediante algoritmos que detectan errores e inconsistencias, optimizando los reportes para su análisis estratégico.*

5. Descripción de la solución tecnológica

Arquitectura de la solución:

*(Describe la **arquitectura general del sistema utilizando uno o más diagramas** que representen los módulos principales, su integración con los sistemas existentes de la SDP y las medidas de seguridad implementadas para proteger los datos. Acompaña los diagramas con **una descripción complementaria de máximo 1000 palabras**)*



La arquitectura de la solución propuesta está basada en módulos. Cada módulo está compuesto por uno o más contenedores basados en Docker. Esta propuesta de solución utiliza tecnologías modernas y permite correr la solución on-premise o en cualquier proveedor de nube basado en contenedores. Esta arquitectura permite escalar de manera independiente cada uno de los módulos y/o de sus diferentes componentes de manera independiente. Entre la mayoría de componentes principales se implementan colas de mensajería, lo que permite realizar comunicación asíncrona y procesamiento basado en eventos. Los diferentes módulos se describen a continuación.

Módulo Front-End y Capa Media:

1. Acceso por un servidor web local que tiene un proxy reverso basado en Apache Nginx. Todos los otros dispositivos se comunican por medio de una API y/o un componente del contenedor del servidor web y/o del contenedor que tiene el servicio del Colas de mensajería abre una conexión remota.
2. Proxy reverso que permite escalar varios de los componentes de la solución planteada utilizando Apache Nginx. Control de Acceso a las diferentes aplicaciones / OAuth2 - JWT. Los datos de autenticación estarán en la base de datos Mongo y en caso de necesidad se utilizará KeyCloak.
3. Componente de API Gateway para lograr balanceo de carga, control de throughput y diferentes de temas de seguridad. Este API Gateway puede funcionar como un pequeño bus de integración para diferentes protocolos: por ejemplo, websocket y/o basados en REST.
4. Servicio WEB que tiene una aplicación basada en node.js que tendrá todos los casos de uso como autenticación, un CRUD básico de procesos de carga y transformación de archivos basado en formato PDF y/o EXCEL. También se implementarán todos los otros casos de uso necesarios para las 3 áreas de la Secretaría Distrital de Planeación y todos sus diferentes roles.
5. API de servicios implementada usando SpringBoot y java para atender todos los servicios de capa media necesarios por el servidor web, extraer los datos desde algún otro componente de la solución en caso de necesidad.

Módulo Almacenamiento:

6. Servicios de almacenamiento de objetos basado en un protocolo compatible con S3 y/o similar. Se propone la herramienta de código abierto MinIO. En caso de que en el futuro se desee migrar a una solución de nube, se puede cambiar este servicio sin problema. Este servicio permite la replicación.

Módulo de Bases de Datos:

7. Servicios de base de datos. La solución contempla una base de datos no estructurada basada en MongoDB. También se contempla una base de datos postgres para persistir información de tipo geo-espacial (extensión postgis) y cualquier otro tipo de información estructurada y relacional necesaria. También se contempla una base de datos basada en grafos como Neo4J o similar.

Módulo de Indexación de Documentos:

8. Servidor de colas de mensajería para poder implementar toda la lógica de comunicación asíncrona propuesta en la solución. Se propone utilizar RabbitMQ y/o un servidor basado en el protocolo AMQP. Este protocolo permite escalar todo el proceso de indexación con uno o más nodos respectivamente para realizar todos los trabajos de indexación de los diferentes documentos y/o casos de uso en batch y/o streaming necesarios. Este servicio es fundamental para recibir todos los eventos generados por los pipelines y/o canalizaciones de procesamiento de modelos de Inteligencia Artificial.
9. Componentes para poder hacer procesamiento en streaming de eventos. Se utiliza para implementar el patrón de captura (Change Data Capture) y así poder realizar la persistencia de los textos de los diferentes documentos y persistirlos en un servidor de ElasticSearch. Este componente permitirá realizar búsquedas híbridas. Se fundamenta en el protocolo de Kafka, Kafka Connect y todos los conectores necesarios. Este componente permitirá generar alarmas y eventos de interés frente a reglas de negocio como geocercas, objetos en una región de interés, entre otros.
10. En esta sección se tiene el servicio de la base de datos vectorial basado en una tecnología como Milvus y/o similar. Este motor se encargará de persistir los sumergimientos (embeddings - por su nombre en inglés). Esto le permitirá a la solución realizar búsquedas semánticas, implementar varios temas de similitud semántica, entre otros. Este servicio está disponible por medio de la API visualizada en el contenedor y/o conectado directamente a las colas de mensajería.

Módulo de Modelos de Inteligencia Artificial:

11. Contenedor Docker que implementa una solución basada en la tecnología de nvidia, python y otros servicios para exponer uno o más modelos de Inteligencia Artificial de una manera sincrónica y/o asincrónica. Este módulo cuanta con un servicio de cola de mensajería para hacer replicación y/o atender solicitudes en batch de manera asíncrona. Servicios de canalización y flujos de trabajo para encadenar uno o más modelos en un flujo de trabajo con capacidad para escalar cada componente del módulo de manera independiente. Integración con mensajería basadas en MQTT y/o KAFKA en caso de necesidad.
12. Contenedor Docker dedicado a correr triton-server y/o tensro_rt y/o servidor de inferencia con tecnologías similares, para exponer modelos de inteligencia artificial, procesamiento de lenguaje natural en múltiples formatos como ONNX y pytorch.

Stack tecnológico:

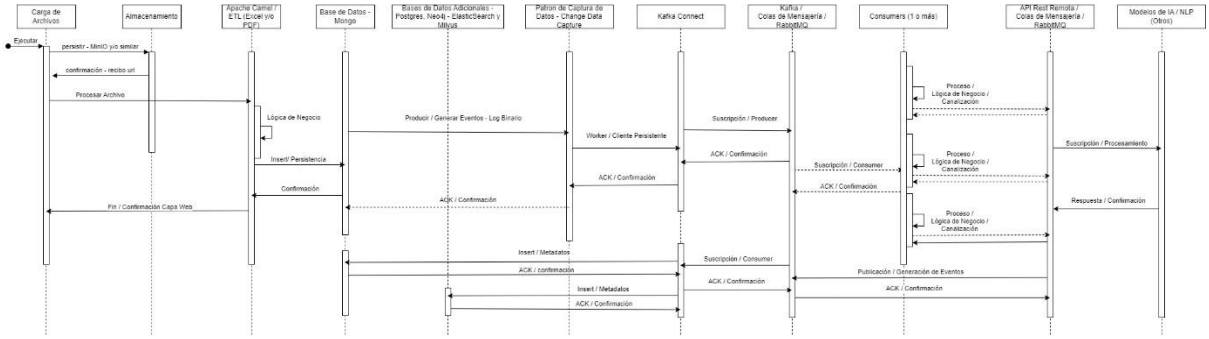
(Completa la tabla identificando las tecnologías empleadas y proporciona una breve justificación)

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	Angular y/o Vue (Nuxtjs)	Frame works para implementación de frontend modernos. Hay muchos componentes necesarios para acelerar el desarrollo de una aplicación web responsiva y con buenos principios de diseño visual.
Lenguaje de Programación	TypeScript, Java, Python y/o C, C++	typeScript es un lenguaje moderno que permite implementar frontend y backend acelerando los tiempos de implementación. Java es un lenguaje empresarial seguro y con muchas librerías para diferentes usos en el backend. Python es el lenguaje por excelencia para implementar temas de Inteligencia Artificial. C y C++ se usan para lograr el mejor desempeño para realizar inferencia de modelos profundos.
Base de Datos	Mongo, Postgres, Neo4J, ElasticSearch, Milvus	Mongo es una base de datos no relacional que permite modelos documentos y sus diversas relaciones. Hay múltiples drivers para su uso en todos los lenguajes. Postgres es una base de datos relacional con varias extensiones entre las que se encuentran PostGIS (Datos Geográficos). Tiene la opción de almacenar documentos en formato JSON (java script object notation – por su nombre en inglés). Neo4J es una base de datos para persistir grafos que pueden representar bases de conocimiento. ElasticSearch es un servidor que permite almacenar documentos en texto crudo permitiendo indexación y búsqueda basada en palabras clave con oportunidad de integración con servicios de búsqueda semántica y basada en vectores. Milvus es uno de los motores de base de datos vectoriales mas utilizados en la actualidad. Puede operar en nube u on-premise. Implementa el almacenamiento de vectores que representan información de datos no estructurados con

		una oportunidad de escalar a billones de vectores en caso de necesidad
Modelos de IA/NLP	Modelos Entrenados por Inversiones Gutiérrez Garcia en reconocimiento de entidades nombradas, clasificación basada en vectores, extracción de relaciones, preguntas extractivas y generativas y modelos de lenguaje basados en transformadores como BERT, FLAN T5 (varias de sus variantes), Gliner, Modelos de Lenguaje de Gran Escala basados en Llama 3 y sus variantes.	Modelos de procesamiento de lenguaje que implementan el estado del arte en tareas como reconocimiento de entidades nombradas, extracción de relaciones, desambiguación de entidades, entre otros. Estas tareas son fundamentales para la solución propuesta. Modelos de clasificación basados en sumergimientos y/o embeddings (por su nombre en inglés). Estos vectores representan características de una palabra, frase, párrafo o documento según sea necesario. Implementación personalizada basada en afinamiento teniendo ejemplos de textos y documentos en Colombia. Modelos de preguntas extractivas y generativas fundamentales para resolver parte de los retos identificados. Modelos de Lenguaje de gran escala afinados para seguir instrucciones en data sets customizados para retos similares.
Otra(s)	Kafka, Kafka Connect, Redis, Colas de Mensajería, Apache Camel	Tecnologías modernas que implementan protocolo de suscripción y producción. Permiten implementar comunicación asíncrona entre microservicios. Kafka Connect tiene una gran cantidad de conectores de fuente y destino para implementar cualquier comunicación entre diferentes fuentes de datos. Redis implementa cachés en memoria para acelerar procesamiento basado en eventos y también permite suscripción y producción asíncrona. Apache Camel es una librería para realizar integraciones e implementación de ETL's (extracción, transformación y Carga, por sus siglas en inglés)

Procesamiento y gestión de datos:

*(Este apartado debe describir el **flujo integral de datos en la solución**, desde su recolección inicial hasta su análisis, categorización, generación de reportes y almacenamiento final. Se deben describir los métodos y modelos utilizados para la recolección, limpieza y preparación de los datos, así como el pipeline de ingesta y clasificación automática, el proceso de generación de reportes automatizados (si se contempla) y el almacenamiento de la información. En este apartado es **obligatorio incluir al menos un esquema o diagrama** que ilustre el flujo completo del proceso. **El texto descriptivo no debe superar las 500 palabras**).*



El flujo de información inicia con la carga de un archivo en formato PDF o EXCEL en el que se encuentre la información en la capa web. Seguidamente se guarda el archivo en el componente de almacenamiento (contenedor Docker – MinIO). Luego se le notifica al usuario quien puede generar el evento para iniciar el procesamiento del archivo. Este procesamiento se implementa en un flujo en Apache Camel que tiene la funcionalidad de procesar los archivos en EXCEL, validar su estructura y aplicar una o más reglas de negocio según corresponda para los diferentes tipos de formatos. Luego de que las validaciones de los diferentes archivos se cumplen, se persiste en batch en la base de datos de Mongo principalmente. Se implementará el patrón de captura de Datos (Change Data Capture) asociado al log binario de la base de datos en Mongo para extraer, analizar y procesar los eventos de inserción, actualización y eliminación de una o mas de las colecciones que representan los diferentes modelos de documentos propuestos en la solución. Se utiliza un conector especial para este fin dentro del componente de Kafka Connect. Este conector persiste en memoria en un servidor Kafka y/o un servidor RabbitMQ y/o otra cola de mensajería los diferentes eventos en formato JSON (Java Script Object Notation). Posteriormente estos diferentes eventos presentes en una o mas colas diferentes, serán consumidos por uno o mas clientes que continuarán el flujo de procesamiento. Cada uno de esos consumidores tiene implementada una lógica de negocio para uno o mas flujos de trabajo de manera respectiva. Algunos ejemplos de estos flujos de trabajo corresponden a realizar la generación de “chunks” o “fragmentación de textos” (en caso de que aplique), generación de sumergimientos utilizando uno o mas modelos de procesamiento de lenguaje natural, extracción de entidades nombradas, extracción de relaciones, extracción de preguntas generativas y/o extractivas, clasificación de textos, construcción de una o mas bases de conocimiento, entre otras tareas. Para poder independizar las inferencias de todos los modelos, nuevamente se utilizan colas de mensajería para poder realizar el mejor procesamiento en batch para cada uno de los diferentes modelos. Al terminar todas las inferencias, se generan las respuestas y/o eventos respectivos en otra cola. Posteriormente y nuevamente utilizando todos los componentes de la arquitectura de Kafka Connect, se consumen dichos eventos y se persisten en una o mas de las colecciones finales con todos los metadatos necesarios para dar solución a todos los retos de las diferentes áreas de la Secretaría de

Planeación. En caso de necesidad, algunos de los conectores también persistirán parte de los datos en el servidor de Postgres y/o Neo4J, ElasticSearch y/o Milvus. Todos estos servicios serán consumidos de manera posterior cuando el usuario realice la visualización de los diferentes análisis implementados en los diferentes flujos de trabajo.

6. Metodología

(Describe la metodología propuesta, completando la información en la tabla. Indica las fases principales y las actividades correspondientes. Asegúrate de que la descripción sea clara, concisa y permita comprender el enfoque y las acciones específicas a desarrollar.)

La metodología propuesta para el desarrollo será de la SCRUM. Se contemplarán sprints de una (1) semana por el tiempo tan corto del proyecto. En la parte de modelos y cuando haya lugar a afinamientos y/o ajustes, se utilizará la metodología CRISP-.DM (Cross-Industry Standard Process for Data Mining, por su nombre en inglés).

Fase	Actividad
General	Análisis y Diseño final de los Modelos de Datos
General	Implementación del Modelo de Datos – Pruebas de Escritorio
General	Configuración y Aprovisionamiento de los Entornos de Desarrollo - Servicios
General	Pruebas de Conectividad y/o Comunicación de Componentes
Front-End	Desarrollo de Mockups
Front-End	Implementación de todas las vistas y/o páginas de la solución
Front-End	Integración con Api de Servicios / Api Gateway
Front-End	Integraciones de Servicios Adicionales
Front-End	Integración de Autenticación y/o Privilegios, Roles, entre otros temas de seguridad
Front-End	Pruebas de Usuario / Aseguramiento de Calidad / Criterios de Aceptación de Componentes
Capa Media	Implementación del Api Gateway – Seguridad – Flujos de Comunicación necesarios
Capa Media	Implementación de la API de Servicios Capa Web y Otros
Capa Media	Implementación de Integración con Otros Servicios
Capa Media	Pruebas de Casos de Uso – Operaciones
ETLs	Implementación de todos los flujos necesarios en Apache Camel para implementar lógica de negocio requerida para diferentes casos de uso
ETLs	Pruebas de ETLs
ETLs	Pruebas de Carga Simple

Canalización de Datos	Análisis y Diseño de las diferentes canalizaciones de datos necesarias
Canalización de Datos	Implementación de las diferentes canalizaciones de datos - conectores
Canalización de Datos	Pruebas de Flujos de Canalizaciones
Canalización de Datos	Pruebas de Carga
Modelos de Inteligencia Artificial	Etiquetado, Pruebas de Latencia, Criterios de Aceptación, Pruebas integrales según metodología CRISP-DM (Cross-Industry Standard Process for Data Mining, por su nombre en inglés)
Modelos de Inteligencia Artificial	Construcción de Conjunto de Datos de Prueba
Modelos de Inteligencia Artificial	Construcción de Conjunto de Datos de Evaluación
Modelos de Inteligencia Artificial	Evaluación de Flujos de Trabajo
Modelos de Inteligencia Artificial	Pruebas de Carga
General	Pruebas Integrales
General	Validación de Criterios de Aceptación

7. Plan de implementación

Cronograma general de implementación:

(Enumera las actividades requeridas para implementar la solución y marca con una "X" los meses correspondientes a su ejecución. Incluye todas las actividades necesarias para llevar a cabo la implementación en un tiempo no mayor al 10 de marzo de 2025, asegurando que estén dispuestas de manera lógica y alineadas con los objetivos del proyecto.)

Actividad	Mes		
	Mes 1	Mes 2	Mes 3
General - Análisis y Diseño final de los Modelos de Datos	X		
General - Implementación del Modelo de Datos – Pruebas de Escritorio	X		
General -Configuración y Aprovisionamiento de los Entornos de Desarrollo - Servicios	X		

General - Pruebas de Conectividad y/o Comunicación de Componentes	X	X	
Front-End - Desarrollo de Mockups	X		
Front-End - Implementación de todas las vistas y/o páginas de la solución	X		
Front-End - Integración con Api de Servicios / Api Gateway	X	X	
Front-End - Integraciones de Servicios Adicionales	X	X	
Front-End - Integración de Autenticación y/o Privilegios, Roles, entre otros temas de seguridad	X		
Front-End - Pruebas de Usuario / Aseguramiento de Calidad / Criterios de Aceptación de Componentes	X	X	X
Capa Media - Implementación del Api Gateway – Seguridad – Flujos de Comunicación necesarios	X	X	
Capa Media - Implementación de la API de Servicios Capa Web y Otros	X	X	X
Capa Media - Implementación de Integración con Otros Servicios	X	X	
Capa Media - Pruebas de Casos de Uso – Operaciones	X	X	X
ETLs - Implementación de todos los flujos necesarios en Apache Camel para implementar lógica de negocio requerida para diferentes casos de uso	X	X	X
ETLs - Pruebas de ETLs	X	X	X
ETLs - Pruebas de Carga Simple		X	X
Canalización de Datos - Análisis y Diseño de las diferentes canalizaciones de datos necesarias	X		
Canalización de Datos - Implementación de las diferentes canalizaciones de datos - conectores	X	X	X
Canalización de Datos - Pruebas de Flujos de Canalizaciones	X	X	X
Canalización de Datos - Pruebas de Carga		X	X

Modelos de Inteligencia Artificial - Etiquetado, Pruebas de Latencia, Criterios de Aceptación, Pruebas integrales según metodología CRISP-DM (Cross-Industry Standard Process for Data Mining, por su nombre en inglés)	X	X	
Modelos de Inteligencia Artificial -Construcción de Conjunto de Datos de Prueba	X		
Modelos de Inteligencia Artificial -Construcción de Conjunto de Datos de Evaluación	X		
Modelos de Inteligencia Artificial -Evaluación de Flujos de Trabajo	X	X	X
Modelos de Inteligencia Artificial -Pruebas de Carga		X	X
General - Pruebas Integrales		X	X
General - Validación de Criterios de Aceptación			X

Presupuesto general de implementación:

(Con base en un presupuesto total de 90 millones de pesos colombianos, completa la tabla detallando los ítems y costos presupuestados para la implementación de la solución. Agrega tantas filas como sea necesario)

Categoría	Ítem	Cantidad	Costo unitario	Costo total
General	*Implementación de Actividades a nivel general ya mencionadas previamente (6 – Actividades específicas)*	*1*	*5.000.000*	*5.000.000*
Front-End	*Implementación de Actividades a nivel de front-end ya mencionadas previamente (6 – Actividades específicas)*	*1*	*17.000.000*	*17.000.000*
Capa Media	*Implementación de Actividades a nivel de capa media ya mencionadas previamente (4 – Actividades específicas)*	*1*	*10.000.000*	*10.000.000*

ETLs	*Implementación de Actividades a nivel de ETLs (Extracción, Transformación y Carga) ya mencionadas previamente (3 – Actividades específicas)*	*1*	*10.000.000*	*10.000.000*
Canalización de Datos	*Implementación de Actividades a nivel de canalizaciones de datos ya mencionadas previamente (4 – Actividades específicas)*	*1*	*20.000.000*	*20.000.000*
Modelos de Inteligencia Artificial	*Implementación de Actividades a nivel de todo lo relacionado con uno o más modelos de inteligencia artificial ya mencionadas previamente (5 – Actividades específicas)*	*1*	*28.000.000*	*28.000.000*
TOTAL				***90.000.000***

8. Guía de usuario y requerimientos técnicos

Principales funcionalidades:

(Completa la tabla identificando las principales funcionalidades contempladas en la solución)

Funcionalidad	Descripción	Usuarios objetivo	Valor agregado
Login	*Inicio de Sesión*	*Todos los usuarios que van a utilizar la solución*	*Oportunidad de uso de KeyCloak con posibilidad de integrar LDAP y/o directorio activo de la entidad posteriormente.*
Inicio	*Página de Inicio con resumen de las tareas más importantes*	*Todos los usuarios que van a utilizar la solución*	*Pequeño tablero de control que muestra lo que ha realizado el usuario que ha abierto sesión y/o el grupo de usuarios del área a la que pertenece el usuario.*

Carga de Documentos	*Operaciones Necesarias para realizar carga de archivos en formato Excel y/o PDF para realizar extracción de información, procesamiento, entre otros.*	*Todos los usuarios que van a utilizar la solución*	*Implementación de una arquitectura para realizar el indexado y extracción de información de los diferentes documentos basada en eventos. Esto permite escalar cada uno de los componentes de la solución de manera independiente.*
Visualización Procesos de Negocio	*Operaciones para revisar el estado de los diferentes procesos de extracción, análisis y generación de metadatos*	*Todos los usuarios que van a utilizar la solución*	*Visualización y búsqueda simple de todos los procesos en batch necesarios para dar solución a los diferentes retos.*
Visualización de Resultados	*Operaciones para revisar los resultados de los diferentes análisis realizados por uno o más modelos de Inteligencia Artificial. Oportunidad de realizar filtros y selección de registros. Visualización de datos que permite exploración de los resultados de una manera intuitiva y que permite análisis jerárquico y detallado según interés de cada usuario.*	*Todos los usuarios que van a utilizar la solución*	*Análisis de uno o mas modelos de procesamiento de lenguaje, generación de sumergimientos, flujos de trabajo modulares. Oportunidad de realizar búsquedas semánticas, de similitud, utilizando conceptos relevantes según cada uno de los diferentes retos, visualización del contexto de cada registro de los diferentes documentos. Visualización de la información que permite capturar patrones a nivel general. Oportunidad de ordenar los registros según algunos criterios relevantes.*
Visualización de Resultados Detallados	*Visualización detallada de los diferentes criterios utilizados para resolver cada uno de*	*Todos los usuarios que van a utilizar la solución*	*Visualización detallada y avanzada de los diferentes criterios utilizados para resolver cada uno de los retos*

	los retos propuestos. Según corresponda, se tendrá clasificación, extracción de entidades nombradas, extracción de relaciones, emparejamiento de entidades relevantes, identificación del contexto de cada registro según un marco conceptual relevante. Visualización de métricas de calidad y diferentes criterios propuestos para cada reto. Oportunidad de realizar realimentación de los diferentes resultados de los modelos de procesamiento.		*propuestos. Identificación de la clasificación de cada registro de manera simple. Medición de calidad, ranking detallado de cada registro. Extracción de entidades nombradas, extracción de relaciones, emparejamiento de entidades relevantes, identificación del contexto de cada registro según un marco conceptual relevante., en caso de que sea relevante. Visualización de métricas de calidad y diferentes criterios propuestos para cada reto de manera muy intuitiva. Mecanismo para visualizar relevancia, coherencia e importancia de cada aporte y/o texto, en caso de que aplique. Oportunidad de realizar realimentación de los diferentes resultados de los modelos de procesamiento.*
Exportar Datos	*Oportunidad de Exportar información relevante en formato CSV y/o EXCEL*	*Todos los usuarios que van a utilizar la solución*	*Oportunidad de exportar los datos analizados y varios de sus metadatos a formatos como CVS y/o Excel que pueden ser importados en otras soluciones de la entidad. Oportunidad de integrarse posteriormente con los sistemas de reportes de la entidad.*

Descripción de la interfaz:

(Describe brevemente cada pantalla o sección principal de la interfaz, incluyendo su propósito, elementos clave, acciones permitidas y resultados esperados. Incluye un esquema o captura de pantalla que represente visualmente la pantalla con los elementos destacados)

Login: Pantalla en la que el usuario inicia sesión con su correo y una clave. Permite recuperar la clave enviando un correo al correo institucional. Ver figura 1 - Login.

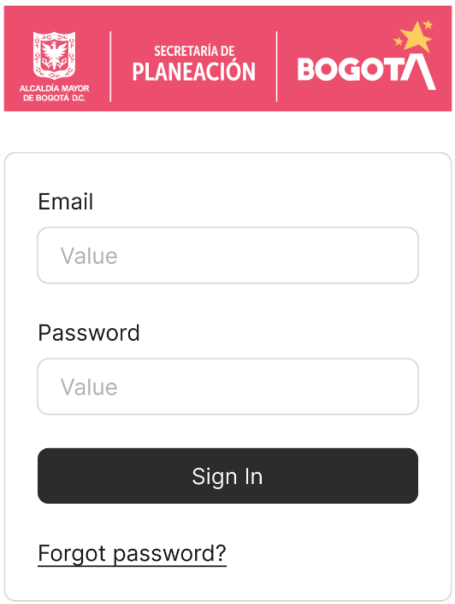


Figura 1 – Login

Inicio – Página Principal: Página de Inicio con resumen de las tareas más importantes. Componente simple que muestra lo que ha realizado el usuario que ha abierto sesión y/o el grupo de usuarios del área a la que pertenece el usuario. Ver Figura 2 – Inicio.

Carga de Documentos: Operaciones Necesarias para realizar carga de archivos en formato Excel y/o PDF para realizar extracción de información, procesamiento, entre otros. Ver Figura 3 – Carga de Documentos y Figura 4 – Carga De Documentos – Detalle.

Visualización de Procesos de Negocio: Operaciones para revisar el estado de los diferentes procesos de extracción, análisis y generación de metadatos. Visualización y búsqueda simple de todos los procesos en batch necesarios para dar solución a los diferentes retos. Ver Figura 5 – Visualización de Procesos de Negocio.

Visualización de Resultados: Operaciones para revisar los resultados de los diferentes análisis realizados por uno o más modelos de Inteligencia Artificial. Oportunidad de realizar filtros y selección de registros. Visualización de datos que permite exploración de los resultados de una manera intuitiva y que permite análisis jerárquico y detallado según interés de cada usuario. En la primera sección de la pantalla propuesta, se pueden aplicar múltiples criterios para filtrar la información procesada y analizada. En la segunda sección, se proponen varios tipos de visualización que el usuario podrá utilizar para obtener una visión general de los patrones de los registros seleccionados. Estas visualizaciones dependen en gran medida de los sumergimientos (o embeddings – por su nombre en inglés) generados por uno o más de los modelos de Inteligencia Artificial utilizados. Esta interfaz podrá ser utilizada por el usuario para filtrar aún más los registros y realizar una exploración que le permitirá identificar registros anómalos, entre otros (Ver Figura 7 – Detalle Visualización). Ver Figura 6 – Visualización de Resultados.

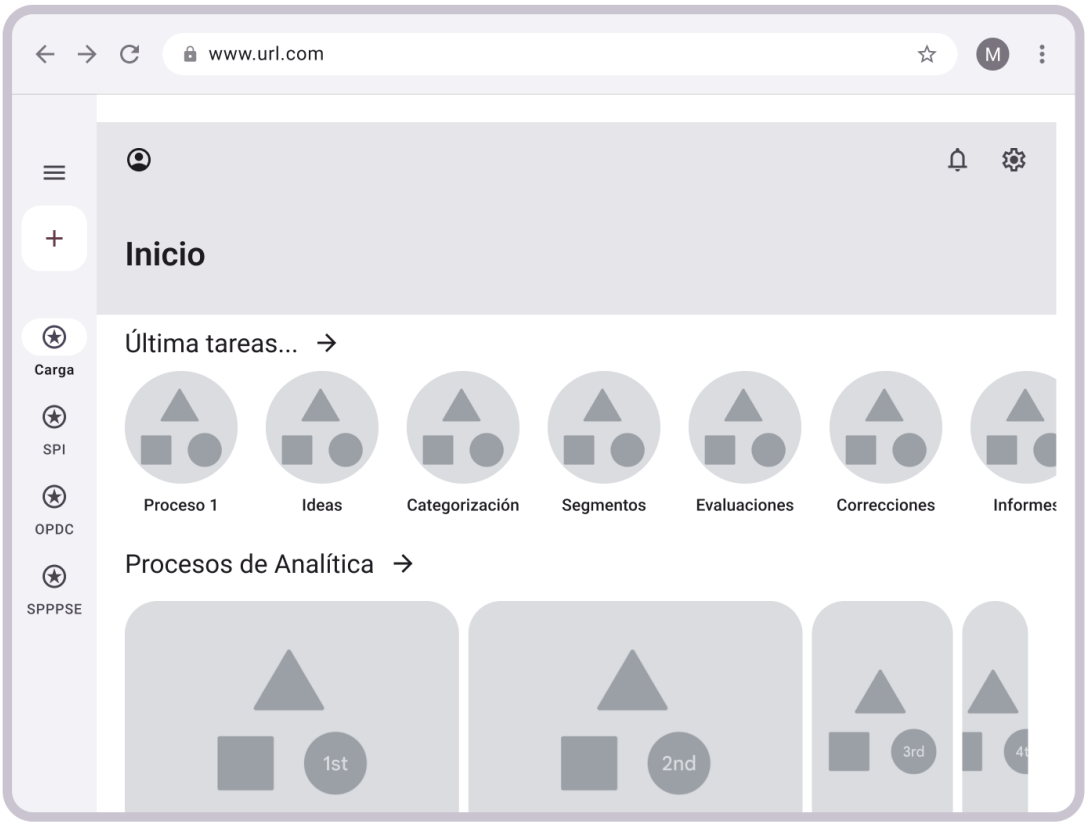


Figura 2 - Inicio



Figura 3 – Carga de Documentos



Figura 4 - Carga De Documentos – Detalle



Figura 5 - Visualización de Procesos de Negocio

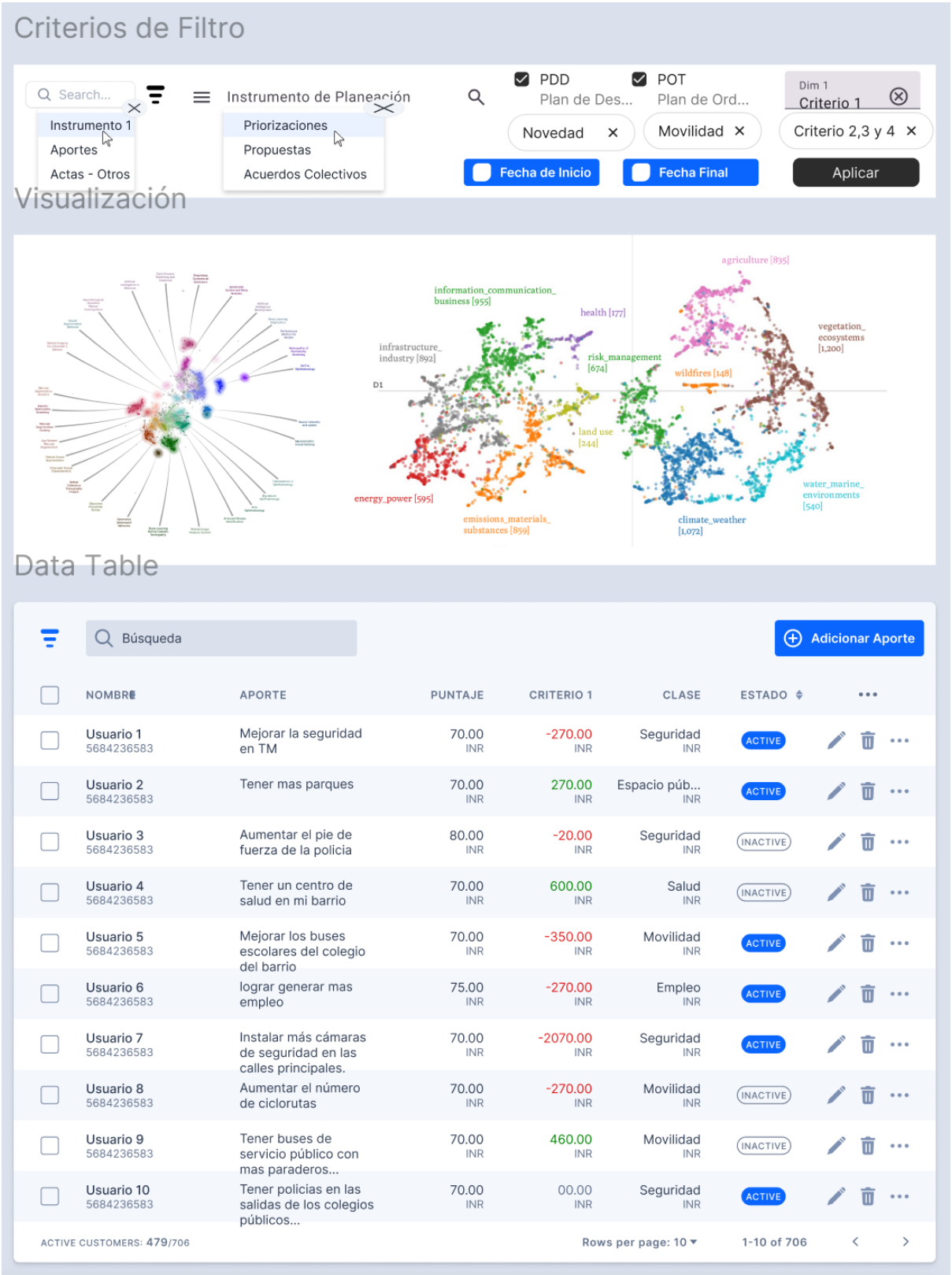


Figura 6 - Visualización de Resultados

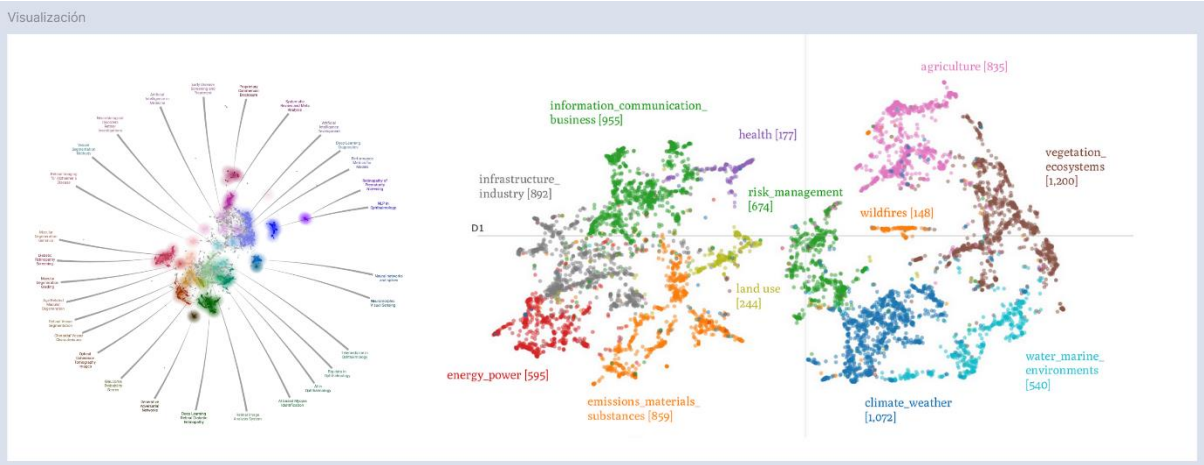


Figura 7 - Detalle Visualización. Identificación de grandes clases de registros, anomalías entre otras.

Visualización de Resultados Detallado: Visualización detallada de los diferentes criterios utilizados para resolver cada uno de los retos propuestos. Según corresponda, se tendrá clasificación, extracción de entidades nombradas, extracción de relaciones, emparejamiento de entidades relevantes, identificación del contexto de cada registro según un marco conceptual relevante. Visualización de métricas de calidad y diferentes criterios propuestos para cada reto. Oportunidad de realizar realimentación de los diferentes resultados de los modelos de procesamiento. Ver Figura 8 y 9 – Visualización de Resultados Detallados.

Exportar Datos: Oportunidad de Exportar información relevante en formato CSV y/o EXCEL. Ver Figura 10 – Exportar Datos.

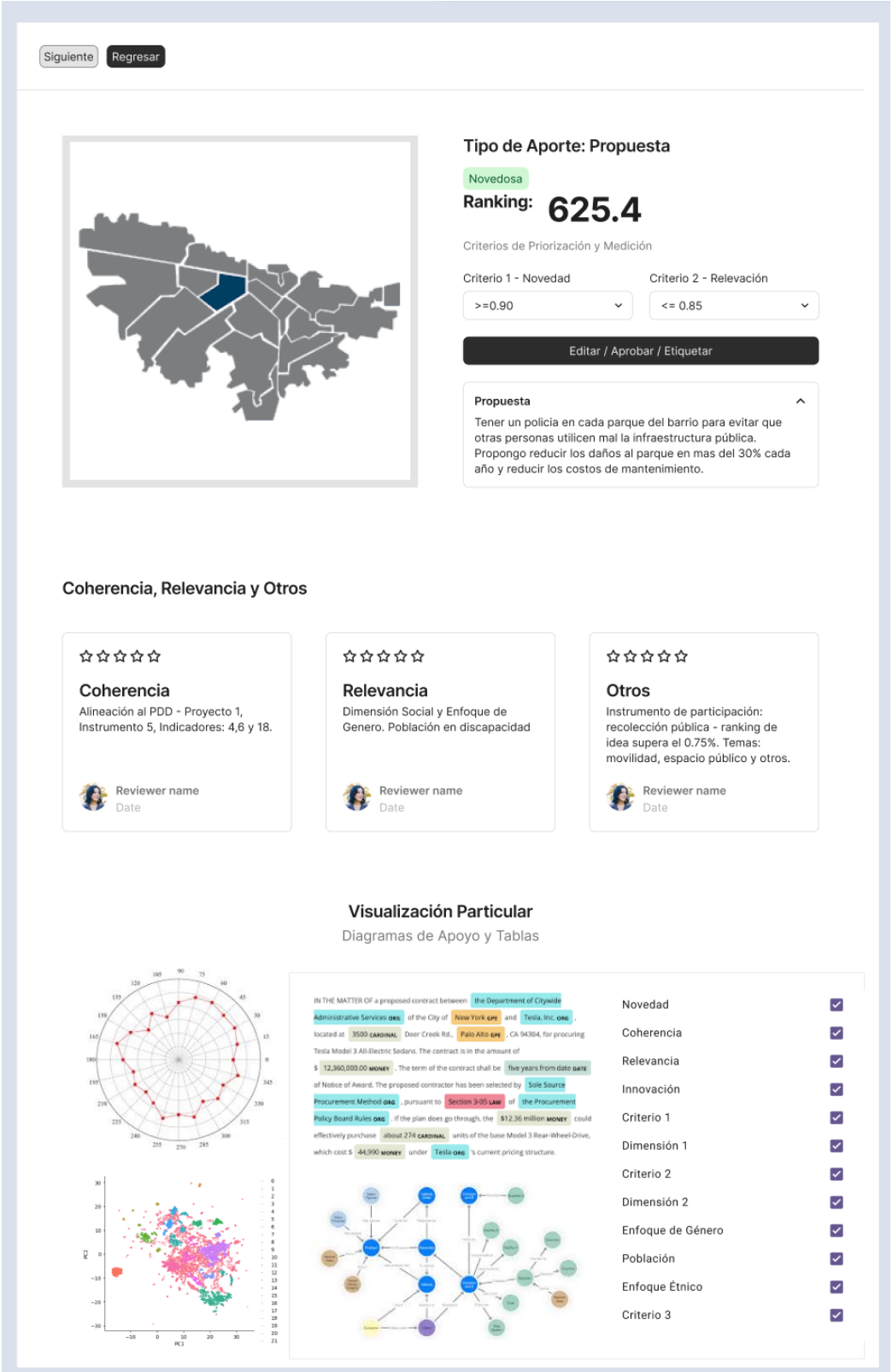


Figura 8 – Visualización Detallada de Resultados.

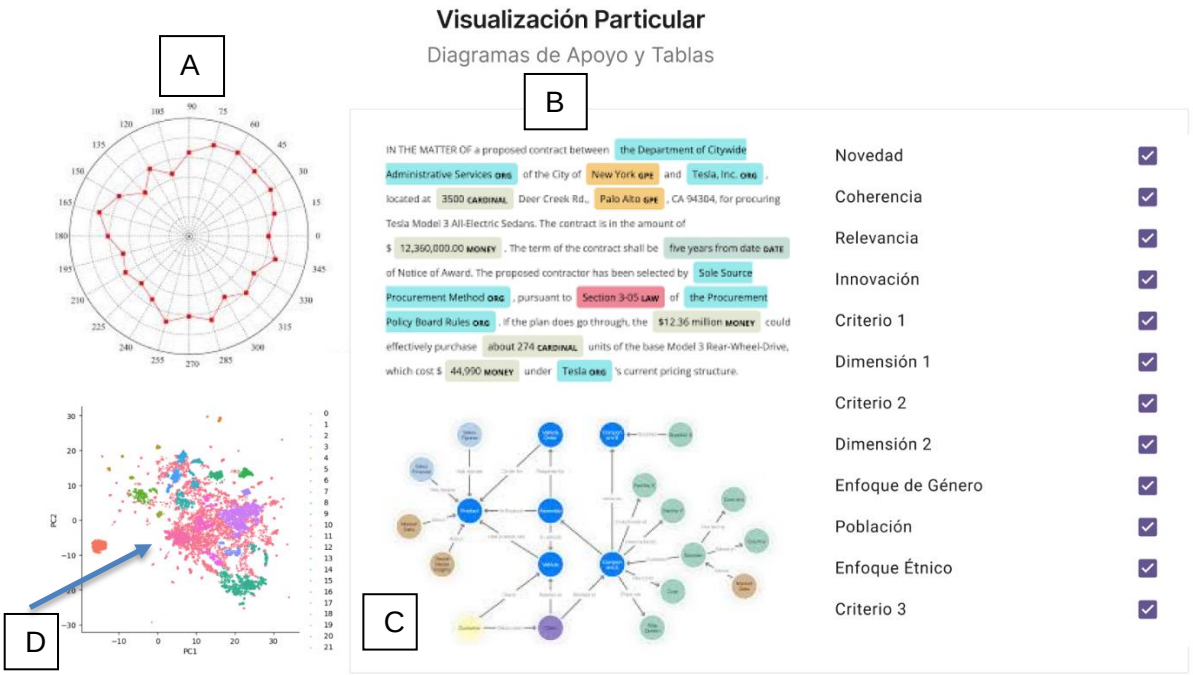


Figura 9 – Visualización Detallada de Resultados. Criterios en forma de gráfico polar (A). Reconocimiento de Entidades Nombradas (Conceptos relevantes para cada reto) (B), Construcción de Contexto basado en grafos (C). Visualización de la idea versus registros similares (D – flecha apunta a registro particular).

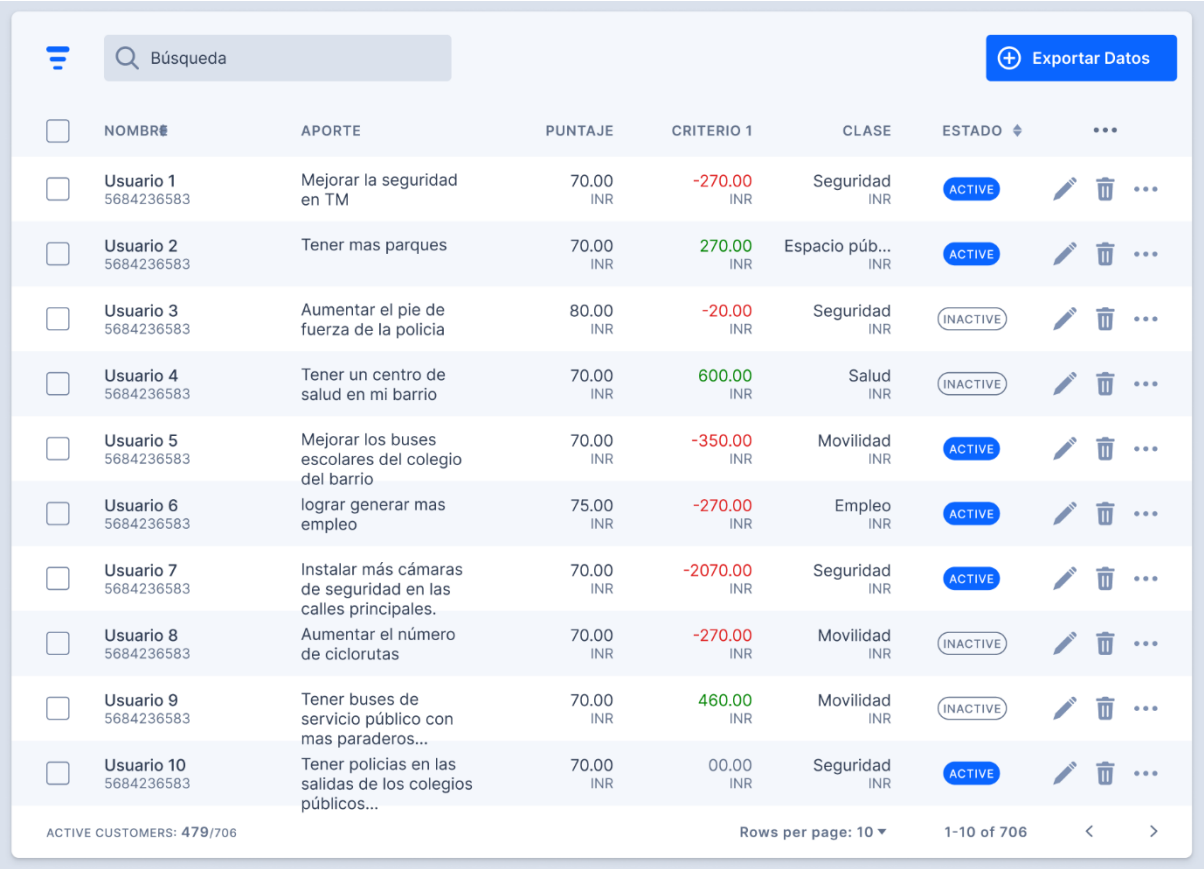


Figura 10 - Exportar Datos

Requerimientos técnicos:

(Completa la tabla indicando los requerimientos técnicos identificados para la implementación de la solución)

Categoría	Requerimientos identificados
Hardware (Especificaciones mínimas de equipos y servidores)	Se requieren máquinas virtuales suficientes para el despliegue de todos los diferentes contenedores mencionados en la arquitectura. En principio se estiman 3 máquinas virtuales cada una con 4-8 núcleos y con 12 y/o 16 GB de memoria RAM respectivamente. El almacenamiento mínimo por cada máquina virtual es el recomendado por el fabricante del sistema operativo. El componente de almacenamiento de archivos y/o de las diferentes bases de datos debe ser acordado entre las partes, pero se estima que mínimo 40 GB en cada servidor debería ser necesario.
Software (Herramientas, plataformas y librerías necesarias)	Servicio de Docker en todas las máquinas virtuales. Imágenes de Docker generadas por Inversiones Gutiérrez Garcia, Integración con APIs de la empresa Inversiones Gutiérrez Garcia

Recursos humanos	Personal de la DTI para realizar las diferentes configuraciones a nivel de seguridad (en caso de necesidad, parametrización de VPN's, entre otras)
Otros (Cualquier otro recurso relevante (ej.., infraestructura, conectividad, capacitación).	Integración con Servicios Externos (en caso de necesidad), Integración con Servicios de Inversiones Gutiérrez Garcia. Capacitación a los usuarios finales para el uso de la solución. Capacitación para el personal de la DTI en caso de necesidad.