**Documentación Unificada del Sistema SIRE (Sistema de Integración de Registros Estadísticos)**

1. **Introducción**

El Sistema de Integración de Registros Estadísticos (SIRE) es una plataforma de procesamiento de datos diseñada para integrar múltiples fuentes de información administrativa y empresarial.   
Su propósito es proporcionar un marco seguro, eficiente y escalable para el manejo de datos sensibles, asegurando la protección de la identidad de las personas y las organizaciones mediante la pseudonimización y la trazabilidad de cada registro.

El sistema no solo garantiza el cumplimiento de normativas nacionales e internacionales sobre protección de datos, sino que además facilita el análisis estadístico al integrar datos de diferentes dominios en un Data Warehouse centralizado.

Ruta anexo detallado: **/docs/technical/architecture.md**

1. **Arquitectura del Sistema**

La arquitectura de SIRE se basa en principios de modularidad y escalabilidad.   
Cada componente cumple una función específica, orquestada a través de contenedores Docker para facilitar la portabilidad y el despliegue.

Componentes principales:

* Apache Airflow para la orquestación de flujos ETL complejos.
* Apache Spark (PySpark) para el procesamiento paralelo de grandes volúmenes de datos.
* FastAPI como interfaz de servicios REST para la generación y validación de IDs estadísticos.
* PostgreSQL (desarrollo) y Oracle (producción) como motores de base de datos.
* Docker y Docker Compose para la administración de infraestructura.

La arquitectura también contempla un esquema de auditoría para asegurar la trazabilidad de cada operación.

Ruta anexo detallado: **/docs/technical/architecture.md**

1. **Gobernanza de Datos**

SIRE implementa políticas de gobernanza de datos alineadas con estándares internacionales,   
enfocadas en garantizar calidad, seguridad y transparencia.

Dimensiones principales:

* **Calidad:** mecanismos de validación automáticos en cada etapa del pipeline ETL.
* **Seguridad:** cifrado en tránsito y en reposo, además de pseudonimización de identificadores.
* **Cumplimiento:** alineación con normativas como GDPR y leyes nacionales de protección de datos.
* **Transparencia:** registro detallado de logs y generación de reportes de auditoría.

Cada dato procesado en el sistema queda asociado a un registro de auditoría que permite reconstruir el flujo completo de su transformación.

Ruta anexo detallado: **/docs/methodology/data-governance.md**

1. **Metodología Data Vault 2.0**

La metodología Data Vault 2.0 es el núcleo de modelado de datos en SIRE.   
Se utiliza para manejar la variabilidad de las fuentes y mantener un historial completo   
de los cambios en los datos.

Estructura implementada:

* **Hubs:** definen claves de negocio únicas como identificadores de personas o empresas.
* **Satélites:** almacenan atributos descriptivos e históricos, permitiendo auditar cambios.
* **Links:** modelan relaciones entre diferentes hubs, como las asociaciones entre personas y empresas.

Beneficios:

* **Escalabilidad:** se adapta a incrementos en el volumen y variedad de los datos.
* **Flexibilidad:** soporta cambios en las estructuras de datos fuente sin comprometer la arquitectura central.
* **Auditoría completa:** mantiene historial de cada modificación realizada a los datos.

Ruta anexo detallado: **/docs/methodology/data-vault-2.0.md**

1. **Procesos ETL**

Los procesos ETL de SIRE están altamente automatizados y orquestados mediante DAGs de Airflow.

Fases:

* **Extracción:** ingesta de archivos CSV provenientes de BDUA y RUES.
* **Validación:** control de estructura, tipos de datos, duplicados y consistencia de claves.
* **Transformación:** cálculo de hashes criptográficos, normalización y pseudonimización.
* **Carga:** almacenamiento incremental en las capas staging, obtenidos y Data Vault.
* **Auditoría:** registro detallado de cada ejecución y generación de reportes de control.

Cada proceso ETL cuenta con mecanismos de control de errores y reintentos para garantizar resiliencia.

Ruta anexo detallado: **/docs/methodology/etl-processes.md**

1. **API FastAPI**

La API expone servicios REST que permiten generar identificadores estadísticos únicos.

Endpoints:

* **POST /generar-id-personas:** registra una nueva persona en el sistema con un ID estadístico único.
* **POST /generar-id-empresas:** registra empresas y asigna IDs únicos.
* **POST /generar-id-empresas-cc:** gestiona casos especiales de persona natural registrada como empresa.

Características adicionales:

* Validación automática de datos con Pydantic.
* Documentación interactiva con Swagger UI y ReDoc.
* Manejo robusto de errores y mensajes claros para el usuario.

La API es clave para integrar SIRE con otros sistemas externos que requieran pseudonimización de datos.

Ruta anexo detallado: **/docs/technical/fastapi-api.md**

1. **PySpark Jobs**

Los jobs implementados en PySpark son el motor de procesamiento masivo de datos.

Componentes:

* **Spark Session Manager:** configuración optimizada de sesiones Spark.
* **Carga de Hubs y Satélites** en Data Vault con jobs dedicados.
* **Hash Calculator:** implementa algoritmos SHA-256 para generar claves de negocio y diferenciales.
* **File Validator:** asegura que los archivos fuente cumplen requisitos antes del procesamiento.

Estos jobs se integran con los DAGs de Airflow, permitiendo escalar el procesamiento en clústeres Spark.

Ruta anexo detallado: **/docs/technical/pyspark-jobs.md**

1. **Scripts SQL**

Los scripts SQL permiten crear y mantener la estructura de la base de datos tanto en PostgreSQL   
como en Oracle.

Incluyen:

* Creación de esquemas sire\_sta, sire\_obt, sire\_dv, sire\_audit.
* Definición de tablas hub\_persona, hub\_empresa, sat\_persona, sat\_empresa.
* Tablas de control como control\_ids\_generados para asegurar la unicidad de IDs.
* Creación de índices en claves hash para optimizar consultas.

Estos scripts permiten una inicialización rápida y segura del sistema en nuevos entornos.

Ruta anexo detallado: **/docs/technical/sql-scripts.md**

1. **Docker y Contenedores**

La infraestructura de SIRE se despliega mediante Docker.

Contenedores principales:

* Airflow (webserver, scheduler, workers).
* FastAPI con versiones específicas para PostgreSQL u Oracle.
* PostgreSQL para desarrollo y pruebas locales.
* Oracle para ambientes productivos.

Docker Compose permite levantar todo el ecosistema con un solo comando,   
facilitando tanto el desarrollo como el despliegue en producción.

Ruta anexo detallado: **/docs/technical/docker-containers.md**

1. **Guía de Instalación**

El sistema está diseñado para instalarse fácilmente en cualquier máquina con Docker.

Pasos principales:

1. Instalar Docker, Docker Compose y Git.
2. Clonar el repositorio SIRE.
3. Configurar variables de entorno en archivo .env.
4. Ejecutar docker-compose para levantar servicios de BD, Airflow y FastAPI.
5. Verificar que los servicios estén activos.

Se recomienda hardware con al menos 8 cores y 32 GB de RAM para entornos productivos.

Ruta anexo detallado: **/docs/user/installation-guide.md**

1. **Guía de Usuario**

El usuario puede interactuar con el sistema de dos maneras principales:

1. **Interfaz web de Airflow:**

* Monitorear ejecución de DAGs.
* Lanzar ejecuciones manuales de procesos ETL.
* Revisar logs y métricas.

1. **API FastAPI:**

* Consumir endpoints para registrar personas y empresas.
* Validar la salud del sistema a través del endpoint /health.

Los usuarios también pueden cargar archivos CSV directamente en las rutas configuradas para ser procesados por los DAGs.

Ruta anexo detallado: **/docs/user/user-guide.md**

1. **Troubleshooting**

Existen guías de resolución para los problemas más frecuentes:

* **Docker no encontrado:** instalar y agregar al PATH.
* **Puerto ocupado:** cambiar configuración en docker-compose.yml.
* **Memoria insuficiente:** aumentar recursos en Docker Desktop.
* **Problemas de conexión a BD:** revisar logs y reiniciar servicios.
* **DAGs que no aparecen en Airflow:** verificar sintaxis y reiniciar scheduler.

Incluye scripts de diagnóstico y de backup para garantizar continuidad operativa.

Ruta anexo detallado: **/docs/technical/troubleshooting.md**

1. **Conclusión**

SIRE consolida en un único sistema el procesamiento seguro y eficiente de datos estadísticos,   
asegurando escalabilidad y cumplimiento normativo. Su diseño modular y containerizado permite   
responder rápidamente a necesidades de negocio y regulatorias.