Documento del Proyecto SIRE

1. Introducción

El proyecto SIRE (Sistema de Integración de Registros Estadísticos) es una plataforma integral para la ejecución de procesos ETL (Extract, Transform, Load) para un Data Warehouse sobre archivos estructurados como CSV. Está orientado a entornos de trazabilidad, validación de datos y auditoría. Su arquitectura se basa en herramientas modernas como Apache Spark, Apache Airflow, Docker, y Poetry, que permiten portabilidad, eficiencia y mantenibilidad en entornos locales o en la nube y además la integración de bases de datos en PostgreSQL y Oracle.

2. Objetivo del Sistema

Automatizar el procesamiento, validación, auditoría y carga de grandes volúmenes de datos hacia bases relacionales (PostgreSQL y Oracle), garantizando control de calidad, monitoreo y escalabilidad. El objetivo está orientado a alimentar un data warehouse que incluye las capas de raw, data vault y datamart, permitiendo una estructura organizada para el análisis de datos y la construcción de modelos analíticos.

3. Arquitectura Tecnológica

- **Apache Spark (PySpark):** Procesamiento distribuido, ideal para grandes volúmenes de datos.
- **Apache Airflow:** Orquestador de flujos ETL con trazabilidad y control.
- **Docker & Docker Compose:** Contenedores reproducibles para el entorno de ejecución.
- **Poetry:** Gestor de dependencias y entornos para Python.
- **Bases de datos:** Compatibilidad con PostgreSQL y Oracle.

4. Estructura del Proyecto

La estructura del proyecto `SIRE` está diseñada para garantizar claridad, escalabilidad y mantenibilidad. Su organización modular permite separar de forma efectiva la lógica de negocio, configuración, entornos de ejecución, scripts, orquestación y auditoría, facilitando el desarrollo colaborativo y el despliegue en distintos entornos.

Esta arquitectura favorece la trazabilidad de los procesos, el crecimiento progresivo del sistema y su integración con nuevas funcionalidades o tecnologías, convirtiéndo en una base sólida para soluciones de procesamiento y análisis de datos.

5. Justificación de Tecnologías

- Airflow: Ideal para flujos complejos y programación de ejecuciones.
- Spark: Procesamiento eficiente para grandes volúmenes de datos.
- Docker Compose: Despliegue automatizado de todos los servicios.
- **.**env: Configuraciones sensibles y portabilidad.
- Poetry: Entorno limpio, manejable y reproducible.

6. Auditoría y Registro

Una de las características clave del sistema SIRE es su capacidad para registrar y auditar cada ejecución del proceso ETL. Esto permite no solo el cumplimiento normativo, sino también una trazabilidad completa para análisis posterior, control de calidad y logs, además permite auditar fácilmente cada carga de datos y detectar errores o duplicados de forma precisa.

Propósito y utilidad

- Facilita la auditoría externa.
- Mejora la trazabilidad de errores.
- Permite analizar cuellos de botella o problemas de rendimiento.
- Útil para generar informes de control y gestión.

7. Infraestructura Docker

El entorno Docker del proyecto SIRE está diseñado para facilitar la ejecución automatizada y replicable de todos los servicios del sistema dentro de una arquitectura basada en contenedores. Esta infraestructura elimina la necesidad de configuraciones manuales complejas, permitiendo que el sistema se despliegue de forma rápida y consistente en diferentes entornos, ya sea en desarrollo, pruebas o producción. Gracias a esta estrategia, es posible garantizar que el comportamiento del sistema se mantenga uniforme sin importar la máquina o servidor donde se ejecute, lo que mejora la portabilidad, la eficiencia operativa y la confiabilidad del despliegue.

Ventajas Clave

- Entornos idénticos: Garantiza consistencia entre desarrollo, pruebas y producción.
- Escalabilidad horizontal: Permite agregar workers adicionales según la demanda.
- Portabilidad: El proyecto puede desplegarse en cualquier máquina que soporte Docker, desde una laptop hasta un clúster en la nube.

8. Escalabilidad y Etapas Futuras

El proyecto SIRE está diseñado con Apache Spark como motor de procesamiento, lo que le permite escalar fácilmente para manejar grandes volúmenes de datos. Spark distribuye la

carga entre múltiples workers, acelerando los procesos ETL mediante ejecución paralela. Esto brinda una base técnica sólida que permite al sistema crecer sin comprometer el rendimiento.

La arquitectura desarrollada hasta el momento constituye una base estable y extensible sobre la cual se puede escalar el proyecto de forma progresiva. En futuras etapas, los Spark workers pueden ser orquestados mediante Kubernetes, lo que facilita la gestión dinámica de recursos, el escalado automático y la alta disponibilidad.

9. Conclusiones

SIRE representa una arquitectura moderna, modular y portable para la integración y procesamiento de datos estructurados. Su diseño responde a las necesidades actuales de trazabilidad, control de calidad y auditoría, integrando tecnologías ampliamente adoptadas como Apache Spark, Airflow y Docker. Gracias a esta base tecnológica, el sistema ofrece una solución robusta, confiable y fácilmente escalable.

La estructura del proyecto permite una clara separación de responsabilidades, lo que facilita tanto su mantenimiento como su evolución. Además, su capacidad de adaptación a diferentes motores de bases de datos (PostgreSQL y Oracle) lo convierte en una herramienta flexible y compatible con infraestructuras institucionales diversas.

La integración de contenedores asegura que el entorno de desarrollo y producción se mantenga uniforme, eliminando errores derivados de inconsistencias de entorno. Al mismo tiempo, el registro detallado de ejecuciones proporciona una trazabilidad completa del ciclo de vida de los datos, lo cual es esencial para entornos regulados o críticos.

En resumen, SIRE no solo cumple con los objetivos actuales del procesamiento y auditoría de datos, sino que sienta las bases para su expansión hacia ecosistemas de análisis avanzado y business intelligence.