# Diseño e Implementación de un Pipeline de Datos ETLT

## I. Objetivo

El objetivo principal de este proyecto es **diseñar, implementar y orquestar un pipeline de datos de tipo ETLT (Extract, Load, Transform, Load)** sobre una arquitectura de Data Lake altamente escalable en Amazon Web Services (AWS). Esta solución está diseñada para integrar datos meteorológicos de diversas fuentes (lotes y streaming), transformarlos en información de valor y ponerlos a disposición de los usuarios de negocio para mejorar la toma de decisiones predictivas y operacionales basadas en datos confiables.

## II. Diagrama de Arquitectura

Para el pipeline se utilizo una arquitectura **Medallion (Bronze, Silver, Gold)**, la cual garantiza la calidad de los datos, la trazabilidad y la reproducibilidad de los resultados analíticos. Todos los servicios de procesamiento y orquestación se ejecutan en entornos **Dockerizados sobre Amazon EC2** para asegurar la portabilidad y la gestión consistente de dependencias.

### Flujo General del Pipeline

|  |  |  |
| --- | --- | --- |
| **Etapa** | **Componentes Clave** | **Descripción de la Tarea** |
| **Extracción (E)** | Airbyte (Batch), Kafka Producer (Streaming) | Obtención de datos desde la API de OpenWeather, tanto en formato histórico (Batch) como en tiempo real (Streaming).  De dos regiones Patagonia y Riohacha |
| **Carga (L1)** | Airbyte, Kafka, Amazon S3 | Almacenamiento inicial de los datos crudos, en formato JSON (historicos) o Parquet, en la **Capa Bronze** del Data Lake (Infraestructura-datos-raw). |
| **Transformación (T)** | Spark (PySpark) en EC2 Docker | Limpieza, normalización, enriquecimiento y validación de datos. |
| **Carga (L2)** | Apache Spark, Amazon S3 | Escritura de los datos curados en la **Capa Silver** (modelo de datos estandarizado) y de los datos modelados en la **Capa Gold** (modelo dimensional). |
| **Orquestación** | Apache Airflow | Programación, monitoreo y gestión de dependencias (DAGs) para garantizar la ejecución secuencial, robusta y automática de los procesos ETLT. |
| **Gobernanza** | AWS IAM |  |

## III. Propósito Detallado de las Capas del Data Lake

La estructura del Data Lake sigue el patrón Medallion, optimizando el costo, la trazabilidad y el rendimiento analítico.

|  |  |  |  |
| --- | --- | --- | --- |
| **Capa** | **Propósito Central** | **Rol en el Flujo** | **Tecnologías Clave** |
| **Bronze / Raw Streaming** | **Ingesta Cruda e Inmutable.** Almacena el dump exacto de los datos fuente, manteniendo su formato original (JSON o archivos Parquet). | **Trazabilidad.** Es el punto de partida para cualquier auditoría o reproceso. Garantiza que siempre se pueda reconstruir el estado de Silver/Gold. | S3, Airbyte, Kafka (Temas), AWS Glue Catalog |
| **Silver / Curated** | **Estandarización y Enriquecimiento.** Contiene datos limpios, sin duplicados, con tipado de datos estricto, y enriquecidos con campos de valor agregado (ej. cálculos de índices solares, columnas de fecha/hora normalizadas). | **Consistencia.** Facilita la construcción de métricas consistentes y reduce la complejidad para poder responder las preguntas de negocio más facilmente. La granularidad es consistente (e.g., por hora). | Spark (PySpark), S3, Docker, EC2 |
| **Gold / Analytics** | **Modelo Dimensional para Consumo.** Almacena datos modelados, optimizados para la velocidad de consulta analítica. | **Consumo Directo.** Fuente principal para Business Intelligence, modelos de Machine Learning y herramientas de consulta. Responde directamente a las preguntas de negocio. | Spark (PySpark), S3, EC2, Docker, Jupyter Notebook |

## IV. Justificación del Stack Tecnológico

La selección tecnológica prioriza la robustez del código abierto (Open Source) para el procesamiento distribuido (Spark, Kafka, Airflow) y la infraestructura escalable y gestionada de AWS (S3, EC2).

|  |  |  |
| --- | --- | --- |
| **Componente** | **Tecnología** | **Justificación y Ventaja Competitiva** |
| **Ingesta Batch** | Airbyte (Conector HTTP) | **Flexibilidad y Velocidad.** Airbyte, como conector open source, permite una configuración rápida y modular. Su conector HTTP se adapta fácilmente a la API de OpenWeather, gestionando la extracción de datos históricos de manera eficiente sin código de scripting pesado. |
| **Ingesta Streaming** | Apache Kafka | **Desacoplamiento y Durabilidad.** Actúa como un buffer de alta disponibilidad entre el productor (API/Script) y el consumidor (Spark Structured Streaming), garantizando que ningún dato se pierda durante picos de tráfico o fallos del consumidor. |
| **Almacenamiento** | Amazon S3 | **Escalabilidad y Economía.** Almacenamiento de objetos los archivos .parquet de las diferentes capas. La separación en múltiples buckets (Bronze, Silver, Gold) |
| **Gobierno** | AWS IAM | Define a los usuarios, grupos y roles. Establece las credenciales que se usan para el acceso inicial a la plataforma AWS, incluyendo permisos para interactuar con los diferentes servicios |
| **Procesamiento** | Apache Spark (PySpark) | **Procesamiento Distribuido.** Permite manejar el gran volumen de datos con transformaciones complejas PySpark ofrece la velocidad de Spark con la simplicidad de Python. |
| **Orquestación** | Apache Airflow | **Fiabilidad Industrial.** Airflow gestiona los DAGs (Grafo Acíclico Dirigido), asegurando que las transformaciones se ejecuten solo después de que la ingesta haya finalizado, gestionando reintentos, dependencias y alertas de fallo. |
| **Infraestructura** | Amazon EC2 + Docker | **Consistencia del Entorno.** Docker garantiza que Airflow, Spark y Kafka siempre se ejecuten en el mismo entorno preconfigurado, eliminando problemas de "funciona en mi máquina" y facilitando la migración o el scale-out. |

## V. Síntesis de Resultados Analíticos (Capa Gold)

El modelo dimensional Gold permite responder a preguntas clave de negocio, con un enfoque en la eficiencia energética de Patagonia y Riohacha.

Preguntas:

1 ¿Cuáles son los días con mayor potencial eólico para la generación de energía?

2 ¿Cuál es el promedio de potencial eólico en una ubicación y día dados?

3 ¿Cómo afecta la nubosidad a la eficiencia de los paneles solares?

4 ¿Cuál es el potencial solar máximo registrado en el día para la optimización de la red?

5 ¿Cuáles fueron las temperaturas mínimas y máximas del día en la ciudad Patagonia/Riohacha'?

6 ¿Cuál es la humedad promedio?

|  |  |  |
| --- | --- | --- |
| **Métrica Clave** | **Hallazgo Analítico** | **Implicación de Negocio** |
| **Eólico Máximo** | **Máximo:** 29.36 WPI (Patagonia, 2025-05-03). **Hora Pico:** 19:00 (Promedio 2.49 WPI). | La Patagonia presenta momentos de potencial eólico excepcional. La máxima generación eólica debe planificarse para el atardecer (19:00h). |
| **Solar Máximo** | **Máximo:** 13.0 SPI (Patagonia, 2024-12-26). **Hora Pico:** 18:00 (Promedio 5.04 SPI). | Existe una alta exposición solar, siendo la hora pico para la generación fotovoltaica justo antes del anochecer. |
| **Correlación Lluvia** | Día más lluvioso: 497.63 mm (Riohacha, 204-11-01). Potencial solar promedio ese día: 1.87 SPI. | La fuerte precipitación en Riohacha tiene un impacto negativo en el potencial solar, confirmando que la Capa Gold debe incorporar métricas de precipitación para la planificación de generación. |
| **Temperaturas Extremas** | **Máxima:** 34.69 °C (Riohacha, 2024-08-20). **Mínima:** -15.01 °C (Patagonia, 2025-06-29). | El pipeline maneja rangos extremos. La Capa Gold debe usarse para generar alertas de mantenimiento por calor (Riohacha) y gestión de infraestructura por frío (Patagonia). |

## VI. Diagrama arquitectura

