Universidad Panamericana
Maestría en Ciencia de Datos
Datos Masivos

Proyecto Final: *Monitoreo Inteligente de Tráfico Marítimo con Big Data AIS*

Enrique Ulises Báez Gómez Tagle, Luis Alejandro Guillén Alvarez

24 de septiembre de 2025

Índice

1. Resumen ejecutivo

1.1. Introducción

El tráfico marítimo global genera millones de registros de posicionamiento a través del Sistema de Identificación Automática (AIS), que transmite automáticamente la identidad, posición, velocidad y otros datos de los buques. Originalmente implementado para mejorar la seguridad de la navegación, hoy es crucial para la gestión del tráfico marítimo, la conciencia del entorno y operaciones de búsqueda y rescate.

Este proyecto propone aprovechar Big Data y aprendizaje automático (ML) para analizar una base de datos masiva de mensajes AIS y ofrecer una solución innovadora a la pregunta: **¿Cómo detectar y entender comportamientos anómalos en el tráfico marítimo para mejorar la seguridad y la eficiencia?**

Responder esta pregunta implica analizar patrones normales de navegación y descubrir desviaciones significativas. En esencia:

- **¿Qué estamos haciendo?** Diseñamos un sistema de análisis que procesa grandes volúmenes de datos AIS para identificar anomalías en el comportamiento de los buques (posiciones fuera de lugar, velocidades inusuales, maniobras erráticas, etc.) y extraer patrones útiles sobre la operación de diferentes tipos de embarcaciones.
- **¿Para qué lo hacemos?** Para mejorar la seguridad marítima, la gestión del tráfico y la toma de decisiones, ofreciendo alertas tempranas de potenciales riesgos (colisiones, actividades ilícitas o fallos) y conocimiento profundo a autoridades y empresas navieras.
- **¿Cómo lo hacemos?** Empleando herramientas de Big Data de la Suite de Google Cloud (Storage + DataProc + BigQuery) para procesar datos geoespaciales masivos en en poco tiempo, complementado con modelos de ML que aprenden patrones habituales y detectan comportamientos que se apartan de lo normal.
- **¿A quién beneficia?** A organismos de seguridad marítima (marinas, guardacostas), a empresas navieras optimizando rutas y monitoreo de flotas, a aseguradoras evaluando riesgos operativos e incluso a investigadores del medio marino en estudios ecológicos.

Finalmente, los resultados se integraron en un *dashboard* interactivo con visualizaciones (mapas geográficos, gráficos de tendencias y rankings) que facilitan la interpretación y la toma de decisiones.

2. Visión general del desarrollo

2.1. Solución Actual / visión general

La solución actual se implementa de extremo a extremo sobre la nube de Google Cloud. El flujo completo es el siguiente:

- Una máquina virtual en Google Compute Engine se encarga de **scrapear los datos AIS** y cargarlos en Google Cloud Storage.
- En Cloud Storage se organizan dos capas diferenciadas: *raw* y *curated*, con prefijos de carpeta y particiones por mes (`YYYY-MM=`).
- Un clúster de Google Cloud DataProc ejecuta dos jobs:
  - Job **raw**: descomprime los archivos obtenidos y organiza los datos en particiones por año y mes, y finalmente los guarda en formato Parquet en Google Cloud Storage.
  - Job **curated**: lee la capa cruda, aplica transformaciones iniciales y genera la capa refinada lista para análisis.
- Una Cloud Function crea el **dataset y la tabla en BigQuery**, cargando la información desde la capa curated.

- En BigQuery se centraliza la explotación de datos mediante consultas SQL optimizadas, generando los datasets finales que alimentan las visualizaciones.
- El **dashboard de Streamlit**, desplegado en la misma VM, consume los resultados de BigQuery y presenta mapas, gráficos de tendencias y rankings de manera interactiva.

Toda la solución se ejecuta en la región `us-central1`. Para evitar duplicados en el flujo, se utiliza la clave compuesta **MMSI + timestamp**.

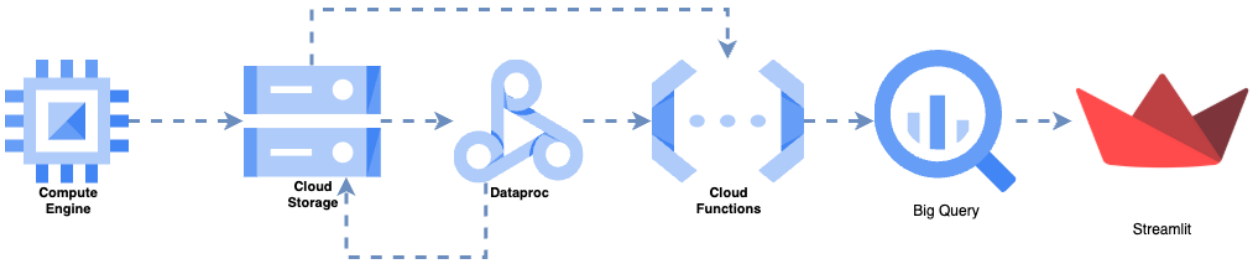

Figura 1: Arquitectura actual de la solución: flujo end-to-end en Google Cloud.

2.2. Limitaciones actuales de la solución

2.3. Propósito, uso y alcance de la herramienta

3. Revisión y uso de datos

3.1. Orígenes y control de datos

3.2. Preparación de datos

3.3. Limpieza y tratamiento de datos

3.4. Integridad de los datos

3.5. Limitaciones de los datos

4. Proceso de desarrollo

4.1. Metodología

4.2. Pruebas

5. Resultados y conclusiones

6. Código utilizado

6.1. Link al repositorio con código fuente y salidas correspondientes

`https://github.com/enriquegomeztagle/MCD-BigData-SmartMaritimeTrafficMonitoring-FinalProject`