Proyecto ABP – Módulo 8: Integración de Datos (Google Colab + PySpark)

Estudiante: Hans Jorge Contreras Robledo

Fecha: 16/09/2025

1. Objetivo y Alcance

Implementación de dos pipelines con PySpark en Colab:

(a) Batch ETL con consultas SQL y regresión.
(b) Streaming simulado con ventanas y clasificación.

2. Arquitectura de Solución

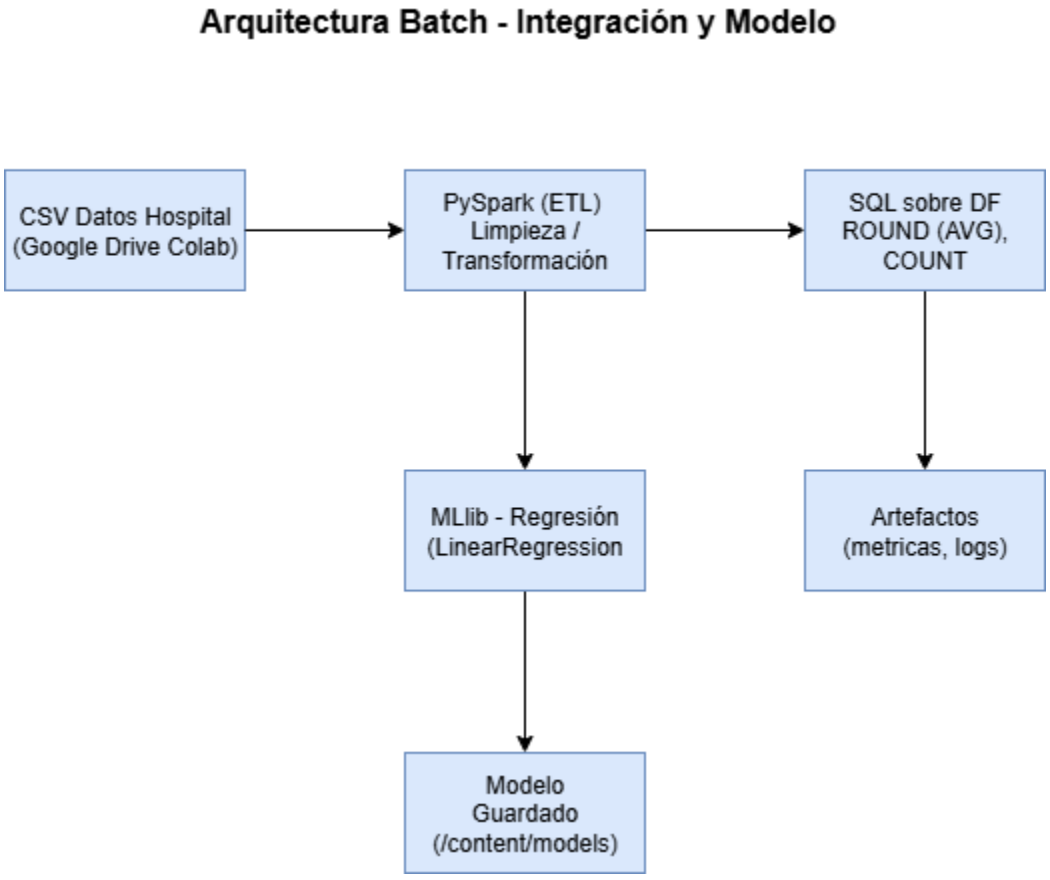

Figura 1. Arquitectura del pipeline Batch en Colab.

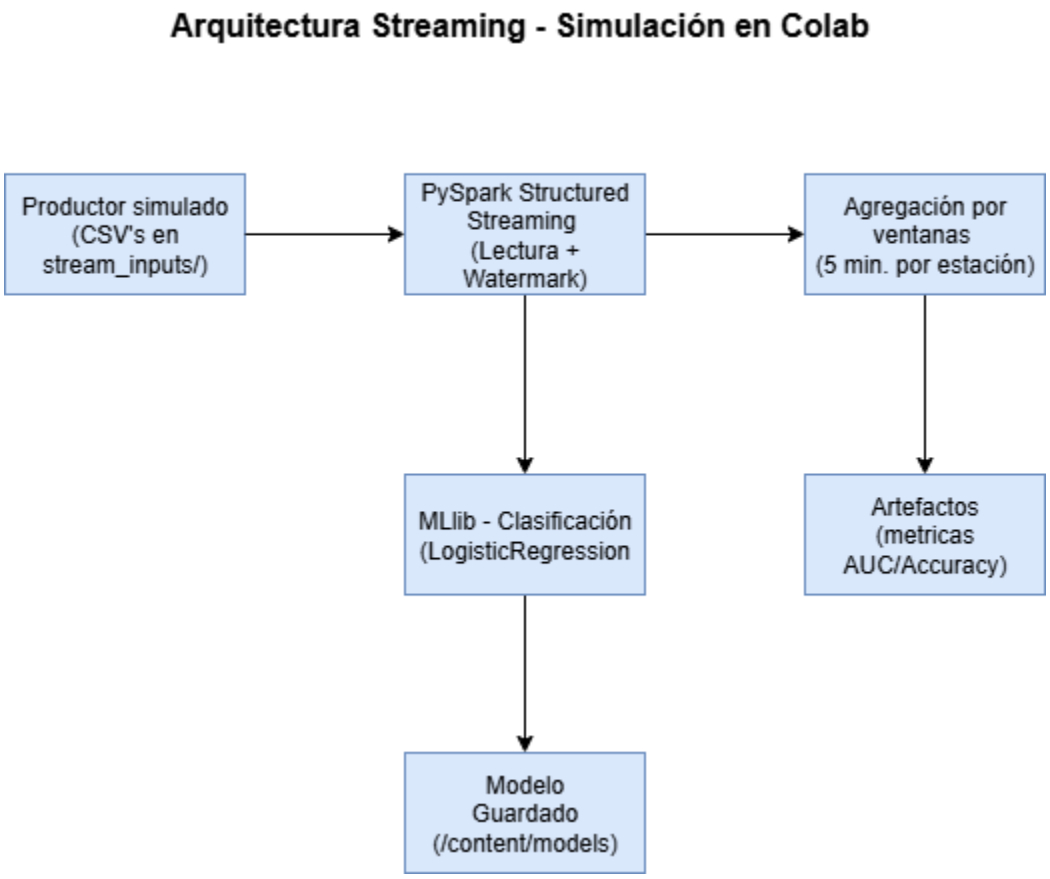

Figura 2. Arquitectura del pipeline Streaming simulado en Colab.

3. Pipeline Batch – Pasos Principales

Instalación de Spark, carga de CSV, limpieza, consultas SQL (ej. ROUND(AVG)), regresión lineal y guardado de métricas/modelo.

4. Pipeline Streaming – Pasos Principales

CSV incremental en stream_inputs/, Structured Streaming con watermark, ventana de 5 min por estación, clasificación y guardado de artefactos.

5. Ejecución en Colab

Abrir y ejecutar en orden: Batch_ETL_PySpark_Colab.ipynb y Streaming_PySpark_Colab.ipynb. Agregar nuevos CSVs mientras corre el streaming para ver actualizaciones.

6. Evidencias

Notebooks .ipynb que al ejecutarlos en entorno Colab, permiten guardar las Métricas en /content/artifacts/*.json y modelos en /content/models/*; Además, se incluyen diagramas en este informe.

7. Reflexión

Colab permite demostrar integración batch + streaming sin instalación local, con trazabilidad y posibilidad de escalar a fuentes reales.