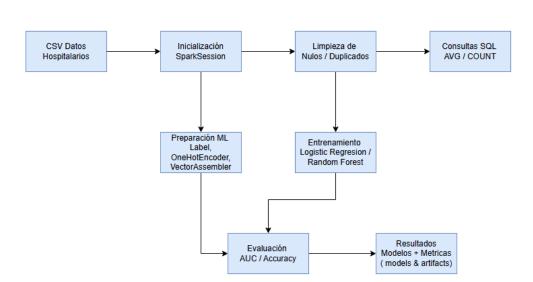
# **Esquemas Casos PySpark**

Hans Jorge Contreras Robledo

Fecha: 15/09/2025

## Caso 1 – Hospital con PySpark

Este caso muestra el procesamiento batch de datos hospitalarios con PySpark. El flujo incluye inicialización de Spark, carga de datos desde CSV, limpieza, consultas SQL, preparación de datos para Machine Learning, entrenamiento de modelos, evaluación y guardado de resultados.



Esquema Caso 1: Dataset Hospital con PySpark

Figura 1. Flujo de procesos para el Caso 1 - Dataset Hospital con PySpark.

## Caso 2 – Metro de Santiago con PySpark

Este caso muestra cómo aplicar PySpark a datos del Metro de Santiago. Incluye carga de un CSV, generación de muestra, consultas con RDD/DataFrame/SQL, aplicación de clustering con KMeans, simulación de streaming y guardado de artefactos y modelo.

#### Limpieza de datos **CSV Datos** Inicialización Consultas Metro Santiago SparkSession RDD / SQL / DF Muestras 20% KMeans VectorAssembler + Evaluación StandardScaler Silhouette Score k=3 clusters Streaming Simulado Resultados Structured Streaming Modelo + Metricas (models & artifacts) (stream inputs/)

Esquema Caso 2: Dataset Metro con PySpark

Figura 2. Flujo de procesos para el Caso 2 - Dataset Metro de Santiago con PySpark.

### **Conclusiones**

Ambos casos ejemplifican cómo PySpark permite manejar datos en distintos escenarios: batch (hospital) y streaming (metro). La representación de los diagramas ayuda a visualizar los pasos principales del pipeline, a manera de facilitar la comprensión del flujo de datos.