
PROYECTO DE EXTRAER, TRANSFORMAR Y CARGAR (ETL) EN DATABRICKS FREE EDITION:

EMPRESA COSMET S.A.C

AUTOR: NECIOSUP BOLAÑOS BRAYAN RAFAEL FECHA: 19/07/2025

PROYECTO ETL – EMPRESA COSMET S.A.C

A. DESCRIPCIÓN DE LA EMPRESA

COSMET S.A.C es una empresa dedicada a la comercialización y distribución de productos cosméticos en el mercado nacional. Opera bajo una estructura tradicional con un sistema interno que almacena su información en hojas de cálculo Excel. El área de **ventas** genera y maneja datos de forma manual, acumulando archivos para los años 2023, 2024 y parte del 2025.

B. CONTEXTO Y NECESIDAD EMPRESARIAL

A medida que COSMET S.A.C crece, también lo hace el volumen de datos. El uso de archivos Excel como fuente principal ha generado limitaciones en:

- Escalabilidad del análisis
- Acceso concurrente y seguro a los datos
- Control de versiones y gobernanza
- Integración con modelos de analítica avanzada

Por ello, la empresa busca migrar sus fuentes locales hacia una base de datos centralizada y escalable en la nube, que permita tanto almacenamiento estructurado como la futura aplicación de analítica avanzada y machine learning.

C. SOLUCIÓN PROPUESTA

Como parte de mi formación como Ingeniero de Datos, propongo una arquitectura basada en Databricks, por ser una plataforma altamente escalable, orientada a procesamiento de datos en la nube, que además permite una integración fluida con herramientas analíticas, notebooks colaborativos y administración robusta de datos. La solución implementada incluye:

✓ Almacenamiento gobernado con Unity Catalog

Utilización de Unity Catalog, una funcionalidad de Databricks, para manejar:

- Gobernanza de datos (permisos, auditorías)
- Organización lógica de catálogos, esquemas y tablas
- Control de accesos basado en roles

- ✓ Proceso ETL con Databricks Notebooks

  Todo el proceso ETL ha sido desarrollado directamente en Notebooks dentro de la plataforma:

  - Extracción: conversión y carga de archivos Excel a formato CSV optimizado
  - Transformación: limpieza, normalización y enriquecimiento de datos
  - Carga: escritura en tablas Delta Lake organizadas por entidad (clientes, productos, ventas, etc.)

- ✓ Durante el proceso de exploración de los datos crudos provenientes de Excel, se identificaron múltiples inconsistencias en los archivos de clientes, productos y ventas, como ausencia de identificadores únicos, campos concatenados o mal estructurados, entre otros. Estas inconsistencias fueron corregidas mediante el proceso ETL implementado en Databricks, utilizando PySpark y herramientas del entorno Unity Catalog para garantizar un modelo relacional y gobernado.

D. HERRAMIENTAS Y TECNOLOGÍAS UTILIZADAS

- ✓ Databricks (versión Free Edition)
- ✓ Unity Catalog para gobernanza de datos
- ✓ PySpark para procesamiento distribuido
- ✓ Delta Lake como formato de almacenamiento transaccional
- ✓ GitHub para versionamiento y documentación

E. ACCESO AL PROYECTO

Todo el proceso detallado y ejecutado se encuentra documentado paso a paso en mi notebook alojado en Databricks, el cual puede consultarse en el siguiente repositorio de GitHub:

- Repositorio Github

F. ¿PORQUÉ ES IMPORTANTE ESTE PROYECTO?

Este proyecto representa una implementación realista del rol de un ingeniero de datos, donde se abordan desafíos comunes en empresas tradicionales:

- ✓ Migración de datos locales a la nube
- ✓ Automatización y control del pipeline ETL
- ✓ Aplicación de prácticas de gobernanza
- ✓ Uso de herramientas modernas como Databricks en un entorno de producción

Con este proyecto he fortalecido mis competencias técnicas en procesamiento distribuido, modelado de datos, manejo de grandes volúmenes de información y despliegue de soluciones escalables.