“Análisis de la produccion de gas e hidrocarburos CNH”

Objetivo del Proyecto

Este documento resume los hallazgos clave del proyecto de ingeniería de datos realizado sobre la producción de petróleo y gas en México, utilizando los datasets abiertos de la Comisión Nacional de Hidrocarburos (CNH). El procesamiento fue realizado con PySpark en Databricks Community Edition y la información se estructuró siguiendo las capas **Bronze**, **Silver** y **Gold**, bajo un enfoque moderno de arquitectura de datos.

El objetivo principal de este trabajo es demostrar que la información pública puede ser transformada en valor mediante herramientas tecnológicas accesibles como **Python**, **PySpark** y servicios en la nube, mejorando significativamente los procesos de explotación del dato. Asimismo, se valida que esta metodología puede ser aplicada tanto en ambientes locales como en plataformas cloud como **AWS** y **GCP**, garantizando escalabilidad, seguridad y eficiencia.

Metodología

Descarga de los datasets desde el portal público de la CNH.

1. Separación de datos por tema: aceite (petróleo) y gas.
2. Carga en formato CSV dentro del entorno de Databricks Community Edition.
3. Transformación a formato Parquet bajo las capas:
    - Bronze: datos crudos tal cual fueron descargados.
    - Silver: limpieza, estandarización de columnas, remoción de registros con producción cero.
    - Gold: generación de insights analíticos y agregados.

Por qué Parquet: Aunque el volumen de datos en este ejercicio es moderado, Parquet es el formato óptimo para grandes volúmenes por su compresión eficiente, lectura columnar y compatibilidad con herramientas de big data.

Hallazgos principales

- El campo Akal es el líder histórico en producción de petróleo.
- Los operadores que destacan en petróleo son PEMEX Exploración y Producción, Hokchi Energy y ENI México.
- En gas, PEMEX también encabeza la lista, seguido por campos como Jose Colomo y Reynosa.
- La producción anual muestra una tendencia descendente en los últimos años.
- La presencia de nuevos operadores privados no ha logrado revertir la caída.

Reflexión

Los datos reflejan una dependencia crítica de campos tradicionales, y una limitada explotación efectiva por parte de nuevos operadores. Esto plantea la necesidad de decisiones estratégicas fundamentadas en datos que ayuden a mejorar el rendimiento operativo del sector.

Próximos pasos con Machine Learning

A partir de los datos transformados:

- Entrenamiento de modelos para **predicción de producción futura**.
- Modelos de **mantenimiento predictivo** según patrones de caída por operador o cuenca.
- Clasificación de campos con potencial de mejora mediante clustering.

Este análisis será replicado localmente con librerías como Pandas y Scikit-Learn, y posteriormente escalado a **AWS** y **GCP**, integrando servicios como **Glue**, **Redshift**, **BigQuery** y **Vertex AI**.

Conclusión

Este ejercicio demuestra que el uso estratégico de tecnologías abiertas y cloud permite extraer valor real de datos públicos. Con este enfoque, es posible optimizar la toma de decisiones en el sector energético, reducir costos de infraestructura y construir bases para una futura inteligencia predictiva con Machine Learning.