Memoria del Proyecto: Procesamiento de Datos Masivos en Sostenibilidad Energética

1. Introducción

Este documento presenta la memoria del proyecto final de la materia de procesamiento de datos masivos en el máster de Análisis de datos masivos e inteligencias de negocio. El objetivo del proyecto fue realizar un proceso ETL, en el cual teníamos como entrada tres distintos data sets relacionados a empresas, proyectos energéticos de algunas de esas empresas y regulaciones ambientales de distintos países. A partir de esto se debían generar dos tablas de eficiencia en el uso de energía renovable y otra con los beneficios económicos de los proyectos energéticos, después de debidamente analizar y depurar los data sets iniciales.

El trabajo se estructuró en tres fases principales:

- **Exploración de Datos (EDA)**
- **Transformaciones para generación de tabla Eficiencia Energética**
- **Transformaciones para generación de tabla Beneficios de Proyectos Energéticos**

Cada fase se documentó en un notebook independiente, los cuales contienen observaciones específicas y detalladas de cada paso en los procesos, decisiones y dificultades encontradas.

---

2. Exploración de Datos (EDA)

En la primera fase, se realizó un Análisis Exploratorio de Datos (EDA) para entender la estructura y calidad de los datos.

2.1 Decisiones de Implementación

- **Carga de Datos:** Se leyeron los archivos iniciales que estaban en formato CSV, y se trabajaron en Pyspark como Dataframes.
- **Limpieza de Datos:** Se valido que los distintos Dataframes no tuvieran presencia de valores nulos, registros duplicados, y que en los registros tuviéramos una estandarización de los valores. Se verifico también que las variables correspondieran al tipo de dato adecuado (por ejemplo, String para variables categóricas y Double para variables numéricas), según correspondiese.

- **Transformaciones:** Se eliminaron caracteres especiales y se aseguró la consistencia y estandarización en las columnas clave.
- **Visualizaciones:** Se analizaron los principales estadísticos descriptivos de las variables numéricas y se generaron histogramas para observar las distribuciones de los datos, validar gráficamente si puede existir algún tipo de sesgo y detectar si existen valores atípicos.
- **Exportación Data sets Depurados:** Una vez realizada todos los pasos anteriores se exporto en formato parquet, los dataframes resultantes. Estos serán la entrada de los dos procesos siguientes.

2.2 Dificultades Encontradas

- **Problemas con caracteres especiales en nombres de empresas y países.** Se solucionó utilizando Unidecode.

---

3. Transformaciones para tabla Eficiencia Energética

En la segunda fase, a partir de los archivos limpiados en la primera fase, se obtuvieron distintas variables referentes a las energías renovables, sostenibilidad empresarial, entre otras. Las cuales darán forma a la primera tabla de salida.

3.1 Decisiones de Implementación

- **Consumo Energía Total:** El valor ya viene dado en el dataset de empresas.
- **Energía Renovable:** Se agrupó por empresas en la tabla de proyectos y se suma la capacidad de generación.
- **Eficiencia Energética:** Se creó la métrica eficiencia_energetica que es el índice entre la energia_renovable y las emisiones_co2.
- **Porcentaje Renovable:** Es la tasa entre las Energía renovable y el consumo de energía total, multiplicado por 100, para tenerlo como porcentaje entre 0 y 100%.
- **Clasificación de Empresas Sostenibles:** Se definieron empresas como sostenibles si su porcentaje renovable era mayor al 50% y su índice de eficiencia energética era mayor a 0.5.

3.2 Dificultades Encontradas

- **Definir el tipo de cruce:** Como se explica en el notebook correspondiente, se tuvo que decidir realizar un cruce tipo Inner entre proyectos y empresas, para tener únicamente las empresas que tenían proyectos renovables y excluir las que no, ya que no tendría sentido tenerlas llenas de valores nulos en los campos calculados previamente. El detalle de estas empresas que no registran proyectos, al poder ser de interés se exporto en un archivo aparte.

---

4. Evaluación de Beneficios de Proyectos Energéticos

En la tercera fase, a partir de los archivos limpiados en la primera fase, se analizaron distintas variables referentes a los beneficios económicos de los proyectos energéticos de las empresas, considerando costos, subsidios y ahorros.

4.1 Decisiones de Implementación

- **Inversión Total Sostenible:** Suma de los costos de los proyectos por empresa, multiplicado por las emisiones de co2 de la empresa.
- **Subsidios Recibidos:** Se obtuvo del dataframe de regulaciones a través de un cruce por el campo país con el dataframe de proyectos, de esta manera obtenemos todas las regulaciones que aplican a cada proyecto y teniendo esto se multiplico el costo de cada proyecto por el subsidio (0 o 1), si aplicaba o no, dando lugar al valor del subsidio de un proyecto bajo cada regulación que correspondiese y al final se agrupo por empresa y se sumó el cálculo mencionado.
- **Impuestos totales:** Para poder calcular los impuestos que paga una empresa tenemos que calcular sus emisiones netas, ya que el pago del impuesto va ligado a esto. Una empresa pagara en caso de que las emisiones netas sean mayores al límite de una regulación. El monto será esa diferencia por el precio del impuesto de esa regulación. Las emisiones netas a su vez son la diferencia entre las emisiones totales de una empresa y la suma de todas las reducciones de emisiones que lograron con sus proyectos renovables.
- **Ahorro Total:** Es la suma de impuestos totales más subsidios recibidos
- **Balance de Sostenibilidad:** Es la resta entre ahorro e inversión totales.

4.2 Dificultades Encontradas

- **Múltiples cruces y agrupaciones:** Se debió tener muchísima precaución para distinguir cuando se debía realizar un cruce antes que una agrupación y viceversa. Para el cálculo de subsidios e impuestos se trabajó con los tres dataframes, ya que todos aportaban información para el cálculo final.

---

5. Conclusiones

Este proyecto permitió aplicar técnicas de procesamiento de datos masivos en PySpark para aplicar un proceso de ETL. Las principales conclusiones son:

1. El análisis exploratorio reveló datos con caracteres especiales y la necesidad de estandarización. Los histogramas de las distribuciones de las variables numéricas no mostraban sesgos, los tipos de variables estaban bien asignados desde la lectura inicial, no se tenían nulos ni duplicados.
2. Obtuvimos cuales empresas son sostenibles de acuerdo con los criterios trabajados.
3. Obtuvimos el balance de sostenibilidad de las empresas de acuerdo con los criterios trabajados

---