

Proyecto final: Soluciones para integración y transformación de datos (ETL)

Con Rafael Ortega

Inteligencia y ciencia de datos

crehana

El proyecto final del curso consiste en realizar una extracción, transformación y carga de una base de datos de consumo de agua en la Ciudad de México. A través de los diferentes avances, deberás realizar la planeación para el siguiente problema:

Supongamos que tenemos que extraer datos del agua de la ciudad de México para explicar qué tanta agua se consume por cada delegación y zona de la ciudad. Para ello, necesitaremos obtener la base de la página de Datos Abiertos de la Ciudad de México (que se encuentran en la siguiente liga: <https://datos.cdmx.gob.mx/dataset/consumo-agua>), procesarlos y dejarlos listos para el consumo de la herramienta que se utilizará para crear un dashboard en Python desde una S3 de AWS. Para efectos de este proyecto, sabemos que será un proyecto pro bono (gratuito) como servicio a la comunidad. Adicionalmente, tus colegas del trabajo están ocupados con otras tareas, por lo que el proyecto deberás realizarlo tú solo.

Es importante mencionar también, que los analistas quieren tener la posibilidad de hacer modificaciones a la granularidad en que se entregan los datos, en caso de que les pidan el reporte anual, semestral o mensual.




Adicionalmente, el cliente (la Ciudad de México) ha solicitado que se le entreguen los nombres de las columnas estandarizadas para una homologación dentro de sus sistemas. Esto es, todos los textos en minúsculas, sin espacios y sin caracteres extraños.

El tiempo que te dará la ciudad para el desarrollo será de 2 años, que es tiempo más que suficiente para un desarrollo completo desde cero.

Este proyecto está dividido en 3 partes:

- Parte 1: Planeando un pipeline
- Parte 2: Extracción y carga de datos
- Parte 3: Transformación de datos

Parte 1: Planeando un pipeline

-  En este primer avance, deberás realizar la planeación para un data pipeline, considerando tecnologías, costos y la forma en que se utilizará.
-  Para ello, deberás responder a una serie de preguntas sobre la planeación de un pipeline.
-  En este avance, lo más importante es que desarrolles todo el tren de pensamiento necesario para llegar a tus conclusiones de qué es importante tomar en cuanto a tecnologías o tipo de patrón (ETL o ELT).



Pregunta 1: ¿Cuál es el problema a resolver?

Pregunta 2: ¿Las fuentes de datos son internas o externas?

Pregunta 3: ¿En qué formatos se encuentran los archivos?

Tip: Puedes entrar a la liga proporcionada arriba para ver el formato en que se encuentran almacenados los archivos.

Pregunta 4: ¿Cuánto espacio en memoria ocupan los datos?

Pregunta 5: ¿Quién será el usuario final de los datos?



Pregunta 6: ¿Tienen algún requerimiento que pueda hacerte inclinarte por ETL o ELT?

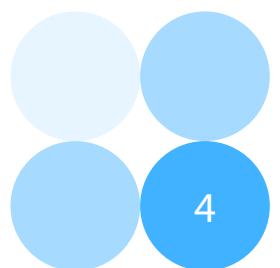
Pregunta 7: ¿Cuál será el presupuesto que utilizarás para el proyecto?

Con esto, hemos concluido la parte inicial con respecto al conocimiento de negocio y de las condiciones para el proyecto. Ahora, necesitamos definir si utilizaremos una herramienta pagada o no.

Tip: El cliente (la Ciudad de México) pagará por el almacenamiento, pero todo el procesamiento lo deberás realizar por tu cuenta en infraestructura propia.

Pregunta 8: ¿Cuál es el tamaño del equipo con el que cuentas para desarrollar el proyecto?

Pregunta 9: ¿Existe un tiempo límite para realizar el proyecto?





Pregunta 10: ¿Sería tiempo suficiente para desarrollarlo con la cantidad de personas con las que cuentas?

Pregunta 11: ¿Necesitarás tecnologías que tengan interoperabilidad con alguna otra tecnología?

Pregunta 12: ¿Cuánto dinero tendrías disponible para invertir en herramientas pagadas?

Pregunta 13: ¿Dónde serán almacenados los datos?

Pregunta 14: Con base en tus respuestas para todas las preguntas anteriores, ¿te convendría construir tu propia solución o pagar por una herramienta no code o low code?

Tip: Considera los gastos en los que tendrías que incurrir en caso de utilizar la herramienta de tu elección.



Pregunta 15: ¿Qué herramienta utilizarás para realizar la ingesta y transformación de los datos?

Pregunta 16: ¿Qué patrón utilizarás para tu pipeline?

Tip: Toma en cuenta las necesidades de los analistas.

Con base en lo aprendido en la clase de “Planeando tu pipeline”, marca en cada una de estas características si te conviene realizar un desarrollo propio o utilizar una herramienta pagada.

Característica	Desarrollo Propio	Herramienta Pagada
Tamaño del equipo y capacidades	<input type="checkbox"/>	<input type="checkbox"/>
Speed To Market	<input type="checkbox"/>	<input type="checkbox"/>
Interoperabilidad	<input type="checkbox"/>	<input type="checkbox"/>
Optimización de costos y valor de negocio	<input type="checkbox"/>	<input type="checkbox"/>
Inmutabilidad y Tecnologías transitorias	<input type="checkbox"/>	<input type="checkbox"/>
Ubicación: ¿dónde vas a guardar los datos?	<input type="checkbox"/>	<input type="checkbox"/>

Parte 2: Extracción y carga de datos

- En esta segunda parte del proyecto, seguiremos trabajando en base al ejemplo utilizado en la primera parte. Para ello, es importante contestar primero las preguntas que encontrarás en este documento y que te ayudarán a tomar las decisiones pertinentes. Recuerda que los datos podrás encontrarlos en la siguiente liga: <https://datos.cdmx.gob.mx/dataset/consumo-agua>

Pregunta 1: ¿Tu fuente de datos, ¿es interna o externa?

Pregunta 2: ¿Qué tipo de archivo necesitarás extraer?

Pregunta 3: Basándote en el ejemplo visto en la clase práctica de ingestión de datos. ¿Desde qué tipo de fuente necesitarás realizar la extracción?

Pregunta 4: Menciona los 3 esquemas o etapas a considerar durante el proceso de ETL o ELT.

Tip: Nos referimos al guardado de datos entre cada paso.



Pregunta 5: ¿Cuál será el tipo de sistema de almacenamiento que utilizarás?

Pregunta 6: Si estuvieras extrayendo información desde una API, ¿qué tipo de código te indicaría que la conexión y extracción se dio de forma correcta?

Habiendo respondido estas preguntas, realiza ahora los siguientes pasos:

1. Realiza la extracción de los datos desde la siguiente liga: <https://datos.cdmx.gob.mx/dataset/consumo-agua>
2. Imprime el resultado de la extracción para verificar que fué exitosa.
3. Prepara la conexión a tu instancia S3.
4. Ejecuta la carga en formato .csv a tu instancia S3 de AWS tal como lo vimos en clase.
5. Imprime una captura de pantalla de tu .csv en la instancia S3 y colócala aquí debajo.



6. Entrega tu script como adjunto.

🎯 Parte 3: Transformación de datos

- En esta etapa del proyecto, tendrás que realizar una exploración a los datos que hemos venido utilizando en proyectos anteriores, para después ocupar los hallazgos al momento de transformar tu dataset.
 - Para desarrollar este avance, deberás completar las preguntas que encontrarás a continuación, así como finalizar las transformaciones requeridas sobre los datos.
 - Finalmente, tus datos transformados deberán ser cargados en tu instancia S3.
 - En este avance, lo más importante es que apliques lo aprendido en las clases sobre EDA y sobre las transformaciones, para que puedas aplicar las transformaciones necesarias para entregar los datos limpios al área de negocio.
 - Al finalizar, necesitarás enviar tu código, así como una copia del .csv de los datos para su validación.
1. En un notebook de Jupyter, extrae la información cargada durante el segundo proyecto.

Pregunta 1: ¿Cuántas variables tenemos?

Pregunta 2: ¿Cuántas observaciones tenemos?

Pregunta 3: ¿Cuántas observaciones únicas tenemos por variable?



Pregunta 4: ¿Cuántas variables numéricas tenemos?

Pregunta 5: ¿Cuántas variables de fecha tenemos?

Pregunta 6: ¿Cuántas variables categóricas tenemos?

Pregunta 7: ¿Cuántas variables de texto tenemos?

Pregunta 8: ¿Cuántas alcaldías tienes? ¿Cuántos nomgeo tienes? ¿Identificas algún error?

Pregunta 9: ¿Qué conocemos ahora de este set de datos por variable?

1. Ahora, realiza las siguientes transformaciones a las variables:
 - Si encontraste algún error en las columnas de alcaldía y nomgeo, arréglalo.
 - Transforma las variables a formato estándar: minúsculas, sin espacios en blanco, sin signos de puntuación.
 - Agrega la variable latitud y longitud.
 - Pasa la variable latitud y longitud a numérica -si no la tomó como numérica-.
 - Elimina la columna geo_point -una vez que creaste la variable latitud y longitud.
 - Elimina la columna geo_shape.
2. Una vez concluidas las transformaciones, sube tus resultados a tu instancia S3 con el nombre "transformed.csv".
3. Imprime una captura de pantalla de tu .csv en la instancia S3 y colócala aquí debajo.



4. Entrega tu archivo transformado como adjunto.
5. Entrega tu script como adjunto.