Proyecto #1 Flujos de ejecución

Client: Ministerio de Educación de la Nación

Situación inicial 📍

Somos un equipo de desarrollo y data analytics, que trabajamos para la consultora "MyData" y nuestro líder técnico nos comparte un pedido comercial directamente del Consejo Nacional de Calidad de la Educación (por sus siglas, CNCE).

El CNCE es un grupo deliberante que pertenece al Ministerio de Educación de la Nación Argentina. Este se encuentra analizando opciones universitarias disponibles en los últimos 10 años para comparar datos extraídos de universidades de todo el país, públicas y privadas, con el fin de tener una muestra representativa que facilite el análisis.

Para esto, compartieron a "MyData" información disponible de más de 15 universidades y centros educativos con gran volumen de datos sensibles sobre las inscripciones de alumnos. El CNCE requiere que preparemos el set de datos para que puedan analizar la información relevante y tomar directrices en cuanto a qué carreras universitarias requieren programa de becas, qué planes de estudios tienen adhesión, entre otros.

Tu objetivo 📋

Como parte de un equipo de desarrollo y data analytics de "MyData", deberás analizar y preparar flujos de ejecución del set de datos recibido para obtener las comparaciones y mediciones requeridas por el CNCE.

Requerimientos 🔧

- El Ministerio necesita que ordenemos los datos para obtener un archivo con sólo la información necesaria de cierto periodo de tiempo y de determinados lugares geográficos de una base de datos SQL (las especificaciones serán vistas en la primera reunión de equipo). Será necesario generar un diagrama de base de datos para que se comprenda la estructura.

- Los datos deben ser procesados de manera que se puedan ejecutar consultas a dos universidades del total disponible para hacer análisis parciales. Para esto será necesario realizar DAGs con Airflow que permitan procesar datos con Python y consultas SQL.

- Calcular, evaluar y ajustar formatos de determinados datos como fechas, nombres, códigos postales según requerimientos normalizados que se especifican para cada grupo de universidades, utilizando Pandas.

Assets 🎨

La base de datos con la información que reunió el Ministerio de Educación se encuentra aquí:

- Host: http://training-main.cghe7e6sfljt.us-east-1.rds.amazonaws.com
- Database: training
- Credenciales: pedirla a su mentor/a

El archivo auxiliar de códigos postales se encuentra aquí: https://drive.google.com/file/d/1or8pr7-XRVf5dIbRblSKlRmcP0wiP9QJ/view

---

Proyecto #2 - Big Data

Client: Stack Overflow

Situación inicial 📍

Somos un equipo de desarrollo y data analytics, que trabajamos para la consultora "MyData" y nuestro líder técnico nos comparte un pedido comercial desde el grupo administrador de Stack Overflow.

Si no conocen Stack Overflow, es un sitio web que forma parte de la vida de la mayoría de los programadores. Reúne preguntas y respuestas tanto para programadores profesionales como para aficionados. Fue creado en el 2008 por Jeff Atwood y Joel Spolsky. Funciona como un blog de consultas y respuestas donde se puede votar y categorizar las respuestas de manera que permanezca la información disponible para futuras visitas.

Para elevar su actividad al próximo nivel, los fundadores de Stack Overflow necesitan analizar la performance del sitio y de sus usuarios de manera de reconocer fortalezas y puntos de mejora continua.

Para esto, Stack Overflow envió un set de datos extraído de su sitio que registra el movimiento de los últimos años, que necesitan que nuestro equipo pueda analizar, pulir y reportar de acuerdo a requerimientos específicos.

Tu objetivo 📋

Como parte de un equipo de desarrollo y data analytics, deberás analizar, reportar, y documentar el procesamiento del set de datos recibido para obtener las comparaciones y mediciones requeridas por el grupo administrador de Stack Overflow.

Assets 🎨

El dataset que nos compartió Stack Overflow se encuentra aquí: https://drive.google.com/drive/folders/177pT0qAizXLfw4lpuDBy6u-lFLEVegXx