Prueba Técnica Data Engineer

Requerimientos

- Herramientas open source.

Caso de estudio proceso ETL

En USA, los ciudadanos pueden cambiar o ingresar a un plan de salud entre noviembre y diciembre de cada año, este proceso anual se conoce como Open Enrollment. En este proceso, se categorizan tres grupos:

a. personas nuevas,
b. personas que renuevan automáticamente el plan de salud
c. personas que cambian de plan.

Se desea saber cuáles son los *Estados* donde se ha registrado mayor incidencia del COVID-19 desde el inicio de la pandemia, pues información de este tipo es requerida por el equipo de Ciencia de datos para realizar modelos predictivos en aras de anticiparse a la dinámica del Open Enrollment (Temporada de renovación y compra de seguros de salud ACA) del año en curso.

Para dar respuesta al requerimiento de Ciencia de Datos, se debe realizar un proceso de ETL automático o semiautomático el cual se describe a continuación:

1. Extracción de datos.

- Descargue la información del censo US. https://censusreporter.org/
- Descargue la información de personas sin seguro en US. https://aspe.hhs.gov/reports/state-county-local-estimates-uninsured-population-prevalence-key-demographic-features
- Utilizar una base de datos abierta que tenga la información del COVID-19 USA. https://console.cloud.google.com/bigquery?project=bigquery-public-data&page=project

2. Transformación de los datos.

Combine las fuentes de datos para mostrar los Estados en donde hay mayor oportunidad para vender un seguro a personas, teniendo en cuenta el efecto del COVID-19. Se debe mostrar las características demográficas del Estado:

- Población
- Distribución por Género
- Número de personas que No tienen seguro
- Personas entre 19 y 64 años sin seguro
- Número de contagios por COVID-19

3. Carga de los datos.

El resultado debe cargarse en una base de datos (preferiblemente MySQL). Argumentar por qué SQL o NOSQL.

4. Visualización.

Un valor agregado para el equipo de Ciencia de Datos resulta contar con un panel de visualización con la información objeto del proceso de ETL, por lo que se requiere que, en la herramienta de su preferencia, muestre los resultados obtenidos (deseable: un mapa donde se incluya las características demográficas).

¿Qué se debe entregar?

1. Dos videos de 8 minutos cada uno, en donde:
    a. Explique en el primer video, cómo se realiza el proceso de diseño e implementación de una solución ETL; pasos que sigue, frameworks utilizados, buenas prácticas, principales problemas. Además, cuéntenos en 2 minutos el caso más exitoso que haya implementado en una solución ETL.

    b. Explique en un segundo video, el caso de éxito mencionado en el video anterior, en términos de: las tecnologías que utilizó, problemas que se le presentaron y cómo los resolvió. Por otro lado, en no más de 3 minutos coméntenos si ha participado en procesos que requieran almacenamiento y gestión de datos distribuidos. ¿Aplicaría una solución de este tipo en el requerimiento realizado por el área de Ciencia de Datos? ¿Por qué?

2. Resultados del ejercicio Open Enrollment:
    - Archivos de código usado
    - Diagrama de la solución implementada
    - Queries usadas en SQL para combinar los datos.
    - Print screen datos visualizados.

Nota: Tiene 2 días para enviar los resultados.