Prueba Técnica Data Engineer - Grupo Creativo Herrera

Departamento de Analítica Avanzada y BI

Agosto, 2023

1. Instrucciones

1. Lee cuidadosamente cada problema y proporciona las respuestas o soluciones en el formato indicado.
2. Lee el documento 'diccionario y tablas', en este documento encontrarás un diccionario de variables y de terminos necesarios para la prueba. Además, encontrarás las estructuras de las tablas que se mencionan en los ejercicios de la prueba.
3. Para los problemas de programación (Python/R), proporciona el código necesario junto con los comentarios explicativos, para el caso de R el **script** y en el caso de Pyhton un **notebook**.
4. Para la parte de SQL proporcionar un archivo **.sql**.
5. Utiliza los recursos de Azure según sea necesario para los problemas que lo requieran.

2. Azure SQL y Modelado de Datos:

Supongamos que tienes acceso a una base de datos en Azure SQL con las tablas 'LEADS', 'Productos', 'Campaña' y 'Ubicaciones'. Además, asume que las campañas publicitarias son segmentas y unicas por región.

1. **Escribe una consulta SQL que identifique el top 3 de campañas que han tenido las mayores 'LEADS' en cada región geográfica durante el último año. La consulta debe incluir:**

   Nombre de la campaña.

   Región geográfica.

   Total de LEADS.

   Consideraciones:

   Utiliza información de la tabla 'Ubicaciones' para determinar la región geográfica de cada campaña.

Utiliza una función analítica avanzada para clasificar las campañas dentro de cada región en función de sus LEADS.

Considera también la diversidad geográfica de las campañas, evitando que una campaña aparezca en el top 3 de más de una región.

2. **Escribe una consulta SQL que muestre las 5 campañas con el mayor aumento porcentual en LEADS en el último trimestre en comparación con el mismo trimestre del año anterior. La consulta debe incluir:**

   Nombre de la campaña.

   Porcentaje de crecimiento de LEADS.

   Trimestre correspondiente.

   Consideraciones:

   El aumento porcentual se calcula como ((LEADS en el último trimestre del año actual - LEADS en el último trimestre del año anterior) / LEADS en el último trimestre del año anterior) * 100.

   Utiliza funciones analíticas de SQL para calcular el trimestre correspondiente y realizar cálculos avanzados.

   Asegúrate de manejar correctamente las situaciones donde no haya LEADS en el último trimestre del año anterior para una campaña.

   Sugerencias:

   Utilizar ventanas de tiempo deslizantes para definir los trimestres.

   Utilizar funciones analíticas avanzadas como LAG o LEAD para acceder a valores de LEADS de trimestres anteriores.

3. Procesos Almacenados

3. **Supongamos que tienes una base de datos en SQL Server con la tabla 'tmp.impression' que contiene información sobre las impresiones logradas por cada campaña realizada. Cada fila representa una campaña y tiene las siguientes columnas: ID_Campaña, Fecha, Producto, Impression, CPM.**

   Escribe un procedimiento almacenado que realice la siguiente tarea:

   Calcule el total de las impresiones y el promedio del CPM por día para cada mes en el último año.

   Inserte los resultados en una nueva tabla llamada 'prod.resumen_impr_mensual'.

4. **Supongamos que estás trabajando con Azure Data Factory para recibir datos cada hora y cargarlos en una tabla 'tmp.leads_hora' en una base de datos SQL Server. Cada fila en la tabla 'tmp.leads_hora' contiene las siguientes columnas: Hora (con marca de tiempo), Campaña y Leads.**

   Escribe un procedimiento almacenado que realice la siguiente tarea:

   Cada vez que se ejecute, el procedimiento debe:

   Recopilar los datos recibidos cada hora en la tabla 'tmp.leads_hora'. Separar la columna 'Campaña' por el delimitador '_' en dos columnas: 'Plataforma' y 'Nombre_Campaña'.

   Borrar los datos acumulados en la tabla 'prod.leads_diaria' correspondientes al día actual. Insertar los datos acumulados por hora, plataforma y campaña del día actual en la tabla 'prod.leads_diaria'.

4. Procesamiento de Datos con Azure Data Factory

Plantea un ejemplo de tu experiencia (adjunta la evidencia necesaria que consideres para comprender la resolucion de los problemas).

5. **Diseña un pipeline en Azure Data Factory que extraiga datos de una fuente de datos externa, como una API REST, y los cargue en una tabla en Azure SQL. Asegúrate de describir las transformaciones necesarias durante el proceso de carga.**

6. **Un proveedor externo te proporciona datos en archivos Excel almacenados en un contenedor de Azure Blob Storage. Diseña un flujo en Azure Data Factory para importar estos datos a una tabla en una base de datos de SQL Server, asegurándote de manejar posibles problemas de formato y calidad de datos.**

5. Estadistica descriptiva y Machine Learning

7. **Para el siguiente caso unicamente enlista los pasos a seguir y el código que utilizarías:**

   Utiliza Azure Databricks para cargar datos desde Azure SQL y realizar un análisis exploratorio utilizando Python . Describe el proceso para realizar las siguiente tarea:

   Carga los datos desde Azure SQL a través de Databricks.

8. **De la base de datos 'prueba.xlsx' realiza un análisis exploratorio y crea índice que recoja el comportamiento de las variables: 'Inversión', 'Inv Estimada', 'Segundos', 'Avisos', categorizar y describir a las marcas en grupos. Esto puedes realizarlo en Python y/o R**