PI Data Engineer Challenge

Aclaración inicial

- Para la resolución del caso se puede utilizar cualquier lenguaje de programación, herramienta o servicio que considere necesario y pertinente. Nos gustaría ver tu código o implementación.
- Esto NO es un examen de ingreso a la empresa, la resolución de este caso nos permite conocer tus habilidades de una forma más práctica.

El problema

Un proceso ETL toma datos de un archivo y lo deposita en la tabla dbo.Unificado.

Por algún error en estos archivos, aparecieron registros duplicados en la tabla.

Consultando con el cliente, nos cuenta que es posible que estas cosas sucedan como consecuencia de errores en el sistema que genera estos archivos, pero que siempre tomemos el ultimo registro que fue copiado, considerando que un registro será duplicado si los campos [ID], [MUESTRA] y [RESULTADO] son iguales en dos filas distintas.

Consignas

a. Montar el backup de la base de datos (SQL Server - .bak)
b. Descargar programaticamente el archivo csv con el link de abajo. Hacelo teniendo en cuenta que este archivo cambia semana a semana con datos nuevos para integrar a la base de datos.
c. Desarrollar un proceso que inserte las filas del archivo .csv en la tabla Unificado. Tener en cuenta que la columna FECHA_COPIA esta vacia en el archivo, y hay que agregarle la fecha en la cual estas insertando las nuevas filas a la base de datos.
d. Testearlo y verificar que no haya perdida de información. Documentar.
e. Dejar de algún modo programado ese proceso para que se ejecute los lunes de cada semana, a las 5:00 AM.

Santa Rosa 320 Piso 5 A
Teléfono +54(351)5693075
Email: info@piconsulting.com.ar
www.piconsulting.com.ar

f. Mejorar el proceso para que guarde logs con la información que crea necesaria (cantidad de filas afectadas, fecha del proceso, instancia de base de datos, etc).

Restricciones

- No se dispone del SQL Agent para hacer el scheduling, hay que buscar otra forma.
- La implementación de la base de datos no soporta cursores.

Link al archivo CSV:

https://gen2cluster.blob.core.windows.net/challenge/csv/nuevas_filas.csv?sp=r&st=2020-10-30T14:05:08Z&se=2020-11-30T22:05:08Z&spr=https&sv=2019-12-12&sr=b&sig=UCK4aQvPAIHz19h%2By2NNAYdzs2RF9myeVAFQkwP3Iuc%3D

Presentación de resultados

- El entregable principal es un Informe con tus respuestas: explicación de la solución y scripts utilizados.
- Queremos conocer detalles como los criterios utilizados, dificultades encontradas en el camino y resultados parciales.
- ¿Cuántas horas pensaste que te llevaría? ¿Cuántas te llevo realmente?
- ¿Qué sitios web de consulta utilizaste como ayuda?
- Tendrás una presentación de 30 minutos para mostrar los resultados.

Si tenés alguna duda, contactanos en recursos.humanos@piconsulting.com.ar