Data Engineer

Prueba técnica

Bienvenido/a

¡Bienvenido/a al proceso de selección de Experimentality!

Estamos muy contentos de que hagas parte de este proceso de hacer parte de nuestra compañía, somos una compañía innovadora que apuesta por soluciones tecnológicas escalables, diseñadas y desarrolladas con las mejores prácticas disponibles en el estado del arte del desarrollo de software y de la ingeniería. Si te surge alguna duda en el transcurso de la prueba, escribe un correo de contacto con el asunto ***[TT] Duda en prueba técnica*** a alejocas@experimentality.co, se te responderá lo más pronto posible.

Contexto

Una compañía emergente llamada *Otaku LATAM* situada en Colombia ha recolectado datos de los ingresos anuales de personas en diferentes ciudades de Estados Unidos, además de recolectar información sobre sus enfermedades; esto lo ha hecho con el fin de establecer una base de conocimiento al interior de la compañía acerca de cómo puede ser la demografía estadounidense y así construir una estrategia de mercadeo interesante para ingresar al mercado de dicho país. El producto con el que quieren ingresar a Estados Unidos es con contenido japonés digital por suscripción (animes, mangas y productos para elaboración de cosplay hentai). Hubiesen querido recolectar otras características, pero desconocen los límites de la tecnología. Dicha compañía nos ha contactado porque quiere utilizar esos datos en una nueva herramienta de visualización que acaban de adquirir.

Problema a resolver

La herramienta de visualización de datos que han adquirido tiene limitaciones considerables, pero no se quieren deshacer de ella ya que han pagado mucho dinero. Uno de los limitantes es que la herramienta solo es capaz de hacer lecturas sobre archivos en formato JSON y Parquet, además de esto, quieren también almacenar la información en una base de datos propia.

Por simplicidad se tendrán varias suposiciones sobre la información recolectada:

1. Toda la información recolectada está en inglés.
2. La información recolectada en el campo *Income* está dada en dólares estadounidenses (USD).
3. La información recolectada en el campo *Illness* solo dirá si la persona tiene enfermedades o no.
4. La edad está dada en años.

El reto a resolver es entonces:

- Diseñar e implementar un ETL que transforme los datos del formato en que se encuentren a formato JSON y Parquet.
    - Las personas de negocio de dicha compañía al estar ubicadas en Colombia, prefieren el español como su primera lengua, es por esto que quieren que los datos de *Gender* y de *Illness* se muestren en español.

    - A su vez, el personal de Colombia quiere poder ver los datos de *Income* en pesos colombianos (COP) a una TRM fija (Por efectos prácticos puede ser la TRM del día que se ejecute el ETL) y almacenar la TRM a la que fue convertida.
    - Por alguna extraña razón, las personas de negocio quieren que la edad de las personas sean mostradas en lustros.
- Diseñar un modelo de almacenamiento de datos para los datos recolectados (SQL o NoSQL).
- Diseñar e implementar un ETL que transforme los datos del formato en que se encuentren a un script de inserción en la base de datos escogida (SQL o NoSQL)
- Utilizar un servicio de almacenamiento de archivos para publicar el resultado de la ETL en formato JSON y hacer accesible la información desde la nube.

Entregables requeridos

- Diagrama de arquitectura.
- Lista de requerimientos mínimos para ejecutar la ETL.
- Prueba de la ETL funcionando en vivo.

Entregables opcionales

- Lista de decisiones técnicas ¿Por qué hizo la prueba de la manera en que la hizo?

Retos opcionales

Puedes realizar cualquiera de estos retos opcionales si consideras que tienes tiempo

- Realizar análisis de escalabilidad de ETL.
    - ¿Qué pasaría si 4 procesos concurrentes ejecutan la ETL que construiste?¿Se generarían condiciones de carrera?
    - ¿Cómo se comportaría la ETL que construiste si la ejecutamos con 100 millones de registros?¿Qué recursos necesitaría para ejecutarse de manera fluida?
    - ¿La ETL construida soporta el procesamiento en streaming? Explique por qué sí lo hace o qué necesitaría para soportar procesamiento en streaming o si se requeriría realizar un ETL diferente
- Proposiciones de valor para negocio
    - ¿Qué datos le propondrías al personal de negocio que se deben recolectar para agregar más valor a su estrategia de mercadeo?
        - ¿Con qué herramienta propondrías recolectarlos?
    - Suponiendo que están dispuestos a cambiar su herramienta de visualización y tienen un presupuesto casi-ilimitado ¿qué herramienta de visualización le propondrías?
    - Suponiendo que están dispuestos a cambiar su herramienta de visualización y tienen un presupuesto increíblemente limitado ¿qué herramienta de visualización le propondrías?
    - Realizar un informe con los datos disponibilizados en una herramienta de visualización.

#Gracias

EXPERIMENTALITY