**Cuestionario Data Engineer**

Los ingenieros de datos tienen muchas responsabilidades. Gestionan los sistemas de fuente de datos, simplifican estructuras de datos compleja y proveen de insumos de datos en diferentes formatos a diferentes usuarios a la par que mantienen calidad e integridad en los datos. Muchas veces también proporcionan ETL, generación de KPIs, transformación de datos y otros procedimientos programables para satisfacer las necesidades de datos de sus usuarios.

Para ayudarnos a conocerle un poco más, por favor responda a las siguientes preguntas:

1. **¿Qué es el modelado de datos?**

Es una etapa siguiente a la recopilación y procesos ETL en la que se analizan los datos obtenidos y se hace la definición de su uso y dependiendo la aplicación que tendrán se realiza el diseño de su estructura y organización y posteriormente la representación visual del flujo de datos para que puedan alimentar un sistema de información.

1. **¿Cuál es la diferencia entre un almacén de datos y una base de datos operativa?**

Los almacenes de datos suelen alimentarse de diversas fuentes de datos, también funcionan como un sistema de Big Data que funciona en una sola ubicación y como una única fuente de datos para una organización, mientras que las bases de datos operan con datos de un único origen, y por consiguiente puede realizar una gran cantidad de transacciones en las tablas de manera rápida

1. **Explique las diferencias entre el esquema de estrella y copo de nieve**

Ambos esquemas comparten una tabla de Hechos (tabla principal del esquema) y tablas Dimensión que se unen a la tabla de Hechos a través de determinados campos

En el esquema de estrella existe únicamente una tabla Dimensión conectada a la tabla Hechos, es decir, no hay jerarquías entre tablas Dimensión, solo existe la jerarquía entre la tabla de Hechos y las tablas Dimensión

En el esquema de copo de nieve existe la relación entre tabla de Hechos y las tablas Dimnesión, sin embargo, las tablas Dmensión también se encuentran relacionadas con otras dimensiones y también puede existir la jerarquía entre tablas Dimensión

1. **Explique la diferencia entre datos estructurados y datos no estructurados**

Los datos estructurados tienen una estructura y organización fácil de identificar y analizar, así como un formato de tipo cuantitativo

Los datos no estructurados tienen una estructura o formato que no puede ser procesado o identificado por un sistema de alimentación de información convencional

**5. ¿Cuándo enfrentó un desafío al tratar con datos no estructurados y cómo lo resolvió?**

Al momento de extraer información se especifica explícitamente la fuente de donde se van a extraer datos. Una vez determinada la alimentación de datos, se realiza un algoritmo que detecte la información a extraer y la almacene en un contenedor. La eficacia del proceso mencionado podía ponerse en duda, así que se hicieron experimentaciones en un conjunto de datos pequeño y luego mediano para finalmente determinar que el algoritmo cumple su función

**6. Explique que es un valor atípico**

Valores cuyo patron de comportamiento o atributos son muy diferentes al resto dentro de un sistema de datos

**7. Enumera algunos problemas comunes a los que se enfrentan al analizar datos y como los soluciona habitualmente**

Valores atípicos: dependiendo el patrón de comportamiento al que está sujeto un conjunto de datos, los valores atípicos se pueden reemplazar por algún valor promedio o pueden eliminarse

Mayor cantidad de valores cualitativos que cuantitativos: reemplazar valores cualitativos por algún valor derivado que pueda representar a esa cualidad, por ejemplo, el resultado de alguna operación estadística

Fuente de datos: si la fuente de datos proviene de un único origen el cual tiene acceso limitado a ciertos usuario entonces se crea una simulación del conjunto de datos que podría extraerse

Visualización de datos: si la fuente de alimentación de información proviene de una base de datos puede suceder que las operaciones de visualización sean mas lentas, así que se extrae la información de la base de datos y se carga a la plataforma para hacer mas rápido el análisis y las operaciones.

**8. ¿Con qué herramientas ETL has trabajado? ¿Cuál es tu favorito y por qué?**

1. Weka permite aplicar gran cantidad de filtros y transformaciones a un conjunto de datos dado y permite su visualización o aplicación a algoritmos de machine learning utilizando una interfaz de software, lo que ahorra tiempo en cuestión del preprocesamiento de datos con programación

Apache Spark, ya que además ser compatible con Python contiene un módulo llamado Apache SQL el cual permite hacer consultas SQL sobre dataframes, lo cual facilita y acelera el acceso a la información evitando los errores de realizar la extracción de información con algoritmos programados manualmente, los cuales pueden fallar al realizar el proceso

**9. ¿Cuáles son las cuatro V del big data?**

a. Volumen: se refiere a garantizar que la información concentrada en un gran volumen de datos se ajuste a las necesidades de un sistema de información

b. Velocidad: análisis de datos en tiempo real

c. Variedad: capacidad para procesar datos estructurados, semiestructurados y no estructurados

d. Valor/Veracidad: los datos recopilados se ajustan a las expectativas proyectadas para su aplicación

**10. ¿Qué frameworks para procesar grandes conjuntos de datos conoces y cuáles ha utilizado?**

Apache Spark, Apache Hadoop, Apache Hive. Solo he utilizado Apache Spark

**11. ¿Qué herramientas de versionamiento de código has utilizado y cuál es tu preferida?**

GitHub

**12. Nombra tu dominio del 1 al 10 para los siguientes lenguajes: SQL, Python, JavaScript**

SQL: 7

Python: 8

JavaScript: 8