Diplomado Data Engineer

Módulo Sistemas de computación distribuida
Tarea 1
José Araya

¿Qué opinión tiene sobre la calidad de los datos?

Es una tarea sumamente compleja de alinear en todas las áreas de una compañía pero muy necesaria cuando se tiene conciencia del impacto de la calidad de los datos en las decisiones de la compañía, por un lado requiere madurez de los procedimientos y a un profesional con conocimiento del tema, el cual, pueda crear validaciones en el ingreso de estos, realizar correcciones y también al momento de buscar algún insight tenga conciencia del estado actual de la calidad de los datos tomando alguna decisión sobre registros nulos y atípicos.

En relación a mi experiencia, tanto empresas pequeñas como grandes tienen problemas en este ámbito ya que los esfuerzos económicos están enfocados en la obtención de resultados que generen ganancias para la empresa, en donde si bien la calidad de los datos influye en esta, es complejo dimensionar su impacto por lo que muchas empresas no lo toman en cuenta tomando decisiones de forma sesgada, las que si bien no son las optimas les permiten seguir obteniendo ganancias.

Si tuviera que medir la calidad de sus datos ¿qué operaciones básicas realizaría sobre sus datos para calcular alguna métrica?.

Revisaría el porcentaje de registros nulos por columna, porcentaje del total de registros duplicados, porcentaje de registros fuera de los rangos admitidos, porcentaje de registros que no están en el formato correcto, revisión de registros que están con símbolos por problemas de codificación,

Tarea: ordene los conceptos según el tipo de estructura que tienen

Estructurado

- Datos tradicionales
- Almacenable en un RDBMS

Semi - estructurado

- .json
- .csv
- .xml

No estructurado

- .png
- .pdf
- .txt
- .log
- NO tiene key:value
- Señal Sensor IoT
- .wav
- .mov
- NO tiene filas ni columnas
- No tiene esquema que defina los datos
- Almacenable en una planilla
- .xlsx

¿Para qué ocupamos los datos? escriba una reflexión sobre el uso que se le da a los datos e información en la sociedad actual.

En mi experiencia, los datos han cobrado más y más importancia los últimos años, en donde se utilizan para las aplicaciones, electrodomésticos, automóviles, toma de decisión en las empresas, telecomunicaciones, sensores IoT, etc.

Si bien en los últimos años ha crecido de forma exponencial la generación de datos, aún para la mediana y gran empresa no están los procesos maduros para su procesamiento y consumo. En las empresas se siguen tomando muchas decisiones en torno a costumbres o “desde la guata” siendo que al potenciar una empresa data driven de forma automatizada o semiautomatizada podría impulsar la toma de decisiones en torno a criterios, como por ejemplo, podrían realizar simulaciones de resultados para nuevos proyectos, generar KPI en tiempo real para apoyar la toma de decisiones, unir distintos sistemas para seguir el flujo completo de un proceso, etc. Pero en cambio veo que estas están construidas en torno a Excel, en donde se realizan reportes manuales de forma repetitiva ocupando un tiempo excesivo en tareas repetitivas y con una alta probabilidad de error en donde se termina dando información erronea para la toma de decisiones.