Temas Selectos de Ciencia de Datos: Big Data

Diego Godinez Bravo

Centro de Investigación en Matemáticas

Maestría en Cómputo Estadístico

Septiembre 2024

Tipos de Datos

La gran diversidad de tipos de datos, formatos y fuentes que existen en el entorno y se encuentra disponible para su análisis se denomina *variedad* en big data. Big data abarca una amplia gama de tipos de datos, entre los cuales se encuentran **datos estructurados**, **no estructurados** y **semi-estructurados**, a diferencia de las fuentes de datos tradicionales que generalmente solo contienen datos estructurados. Esta variabilidad supone un gran desafío para la integración, almacenamiento, procesamiento y análisis de datos.

Datos Estructurados

Los datos estructurados, comúnmente provienen de bases de datos relacionales, es la forma convencional y bien definida de organizar los datos. Generalmente estos se organizan en filas y columnas, los cuales definen claramente sus atributos.

Características de los Datos Estructurados

Atributos definibles. Los datos estructurados comparten los mismos atributos para todos los valores en un conjunto de datos (e.g. en una base de datos de empleados cada registro comparte los atributos: nombre del empleado; número de identificación; puesto; fecha de contratación y el salario; etc.).

Atributos definibles. Las tablas de datos estructurados tienen variables (atributos) en común para vincular entre sí diferentes conjuntos de datos.

Almacenamiento. Se pueden almacenar datos estructurados en bases de datos relacionales y administrarlos con el lenguaje de consulta estructurado (SQL, por sus siglas en inglés).

Ejemplos de Datos Estructurados

- Hojas de cálculo (e.g. Microsoft Excel)
- Bases de datos SQL
- Archivos CSV
- Registro de ventas

Datos No Estructurados

Los datos no estructurados son generados principalmente en documentos de texto, correos electrónicos, publicaciones en redes sociales, fotos, etc., tipos de datos que generalmente no tienen una estructura definida.

Ejemplos de Datos No Estructurados

- Documentos de texto
- Archivos de video
- Grabaciones de audio
- Imágenes

Las fuentes de datos no estructurados pueden ofrecer información valiosa a través del uso de técnicas de inteligencia artificial, como el procesamiento del lenguaje natural y el análisis de imágenes.

Diferencias entre datos estructurados y datos no estructurados

Facilidad de análisis. Una de las principales ventajas de los datos estructurados es su capacidad al momento de analizar la información. Por otro lado, es más complicado analizar datos que no tengan un formato predefinido.

Capacidad de búsqueda. Además de la facilidad de análisis, dentro de las ventajas de trabajar con datos estructurados se encuentra la facilidad con la cual se realizan las búsquedas. Los datos no estructurados carecen de un orden, por lo que obtener información mediante técnicas convencionales se vuelve inviable.

Almacenamiento. Los datos no estructrados por lo regular requieren más espacio de almacenamiento, lo que se traduce en mayor cantidad de recursos, y por lo tanto, mayor cantidad de dinero. Caso contrario a los datos estructurados, los cuales tienen un proceso de almacenamiento optimizado.

Datos Semistructurados

Entre estos dos tipos de datos se encuentra una categoría intermedia: los datos semiestructurados. Son datos que poseen cierta organización pero no siguen un formato específico, es decir, carecen de un modelo de datos relacional o tabular. Sin embargo, contienen **metadatos** que se pueden analizar, estos son datos descriptivos que ayudan a entender y gestionar la información principal (e.g. origen, fecha y hora, formato, permisos y derechos, etc.).

Ejemplos de Datos Semiestructurados

- JSON
- XML
- Archivos comprimidos
- Archivos de páginas web

La gran variedad de fuentes de las cuales provienen los datos incrementa aún más su variabilidad. De entre estas fuentes se pueden encontrar redes sociales, plataformas de e-commerce, teléfonos móviles, sensores, archivos de registro, entre otros. Cada fuente de datos aporta características únicas, por lo cual se generan desafíos en cuanto a su gobernanza, integración y calidad. Big data permite la exploraración de cada una de estas fuentes de datos, lo que permite la extracción de información valiosa para el desarrollo de productos, análisis de sentimientos y toma de decisiones en tiempo real.

Referencias

Demirbaga, Ü., Aujla, G. S., Jindal, A., & Kalyon, O. (2024). *Big data analytics: Theory, techniques, platforms, and applications*. Springer Nature.

Amazon Web Services. (2023). *¿Qué son los datos estructurados?* Retrieved from aws.amazon.com/es/what-is/structured-data/.