Técnicas de gestión de bases de datos masivos

Agregación

Método en el que los datos sin procesar se recopilan y se expresan en forma de resumen para el análisis estadístico. Una vez que los datos se agregan y escriben como un informe, se pueden analizar los datos agregados para obtener información sobre recursos o grupos de recursos particulares. Hay dos tipos de agregación de datos: temporal y espacial.

Integración

Es un paso crucial en el preprocesamiento de datos que implica combinar datos que residen en diferentes fuentes y proporcionar a los usuarios una vista unificada de estos. Incluye múltiples bases de datos, cubos de datos o archivos planos y funciona fusionando los datos de varias fuentes. Existen dos enfoques principales para la integración de datos: el de acoplamiento estrecho y el de acoplamiento flexible.

Construcción de atributos

Este método ayuda a crear un proceso de minería de datos eficiente. En la construcción de atributos o la construcción de características de la transformación de datos, se construyen y agregan nuevos atributos a partir del conjunto de atributos dado para ayudar al proceso de minería.

La transformación de datos es una técnica de mapeo y conversión de datos de un formato a otro. Las herramientas y técnicas utilizadas para la transformación de datos dependen del formato, la complejidad, la estructura y el volumen de estos.

Manipulación

Es el proceso de cambiar o alterar datos para hacerlos más legibles y organizados. Las herramientas de manipulación de datos ayudan a identificar patrones en los datos y transformarlos en una forma utilizable para generar información sobre aspectos como datos financieros o comportamientos de un cliente o consumidor.

Discretización

Es el proceso de convertir valores de atributos de datos continuos en un conjunto finito de intervalos y asociar con cada intervalo algún valor de datos específico. Existe una amplia variedad de métodos de discretización que comienzan con métodos ingenuos como el de ancho igual y frecuencia igual y pueden llegar a métodos mucho más sofisticados como el Principio de longitud de descripción mínima (MDLP, por su sigla en inglés).

Normalización

Método para convertir los datos de origen a otro formato a fin de garantizar un procesamiento eficaz. El objetivo principal de la normalización de datos es minimizar o incluso excluir los datos duplicados. Ofrece varias ventajas, como hacer que los algoritmos de minería de datos sean más efectivos o permitir una extracción de datos más rápida.

Generalización

Método que permite generar capas sucesivas de datos de resumen en una base de datos de evaluación para obtener una visión más completa de un problema o situación. La generalización de datos puede ayudar en el procesamiento analítico en línea (OLAP, por su sigla en inglés), que se utiliza principalmente para proporcionar respuestas rápidas a las consultas analíticas que son multidimensionales.

Suavizado

Es una técnica que permite detectar tendencias en datos ruidosos de los cuales se desconoce la forma de su tendencia. El suavizado puede ayudar a identificar tendencias en la economía, tendencia de acciones, sentimientos de los consumidores, entre otros aspectos.