ETL BigData – Prueba Técnica para GeoDB

Un proceso ETL (Extract/Transform/Load) es empleado para extraer datos, moverlos de unas localizaciones a otras (por ejemplo, desde el código donde se extraen a una base de datos más permanente) y reformatearlos en el proceso, para analizarlos o almacenarlos.

**Extracción y Pre-procesado**

Después de extraer los datos, el siguiente paso es llevar a cabo un pre-procesado. En este paso se ejecutan tareas que transforman los datos para hacerlos más coherentes, normalizados y, en general, facilitan su análisis. Por ejemplo, muchos algoritmos de Aprendizaje Automático requieren codificar valores categóricos y transformarlos a valores numéricos. Otras tareas importantes son la eliminación de valores incoherentes, ausentes, o duplicados; agregación, sampleado de características, reducción de dimensionalidad para minimizar el tiempo de procesado y el espacio requerido para el almacenamiento, etc.

En Python existen muchas herramientas para pre-procesar datos. En este caso, podríamos usar la librería Pandas, y el formato CSV (comma-separated-values), pues es fácil de manipular.

**Transformación y carga**

En nuestro caso, una opción popular y efectiva sería convertir los datos extraídos por el código en Python a formato XML, para después procesarlos y cargarlos en una base de datos en MySQL.

Definiríamos una clase específica para la conversión de los datos extraídos a formato XML. Otra clase distinta serviría para establecer la conexión con el servidor SQL que usemos y, finalmente, una clase para parsear los archivos XML e insertarlos en la base de datos.

Cualquier transformación de datos que suponga algún cálculo (por ejemplo, para obtener el porcentaje de tickets usados por evento), se realizaría en este paso.

Finalmente se cargan los datos en el sistema de bases de datos deseado.