Universidad Nacional Autónoma de México

Facultad de Ciencias, 2024 - 1

Almacenes y Minería de Datos

Práctica 03. Integración de Datos

PostgresandoesoSQLazos

Integrantes

Adrian Aguilera Moreno	421005200
Marco Antonio Rivera Silva	318183583
Sebastían Alejandro Gutiérrez Medina	318287021
Israel Hernández Dorantes	318206604
Alejandra Ortega García	420002495

1. Resumen.

En la actualidad existe una gran generación de datos, grandes empresas como google y facebook generan en un minuto lo que hace unos años se generaba en meses o años. Los datos provienen de distintas fuentes y en general debemos poder integrarlos en grandes bases de datos incluso si los datos a integrar vienen de fuentes relacionales y no relacionales, ¿Esto causa problemas? Sí, muchos y de hecho es una de las partes en un procesamiento de datos que conlleva más tiempo.

Las organizaciones siempre están en busca de crear y adquirir los mejores procedimientos para la integración de datos, en el Big Data nos interesa realizar esta integración respetando las 5 Vs, que a continuación se resumen:

- Volumen: los grandes cúmulos de datos que se generan a diario en empresas cómo google o facebook resultan en cada vez requerir más espacios para almacenar la información generada.
- Velocidad: la función de transmitir datos de un lugar a otro es de suma importancia, en especial nos interesa la rapidez con la que estos datos son transmitidos, es por esto que requerimos un menor tiempo entre el viaje de nuestros datos.
- Variedad: datos generados desde distintos trabajos y fuentes deben poder ser integrados, entonces podemos tener una gran variedad en maneras y formas en que se estructuran los datos.
- Veracidad: depende únicamente de los datos y su fuente. Pues, si tenemos fuentes con falacias nuestro proceso de integración no será vraz o al menos no del todo.
- Valor: el ejemplo más común está en las redes, pues los datos asumen valores de; fotos, texto, HTML, videos, $\cdots$

Las 5 Vs juegan un gran papel para determinar la eficiencia de las organizaciones que realizan la integración de los datos en cada uno de sus niveles.

Los procesos ETL tradicionales asumen la tarea de integrar los grandes cúmulos de datos accediendo a ellos, cargando, extrayendo y transformando los datos. Sin embargo esto se vuelve tedioso y complejo cuando existe una gran variedad de datos, los procesos Hadoop prometen integrar datos con gran variedad de manera más amistosa, este nuevo marco en el Big Data procesa enormes conjuntos de datos de diferentes fuentes. Algunos de los líderes del mercado están trabajando para integrar Hadoop con los sistemas heredados para procesar sus datos para uso comercial según la tendencia actual del mercado. Los datos de diferentes fuentes vienen en forma de etiquetas y códigos donde las organizaciones estaban rezagadas en tecnologías para comprender e interpretar estos datos. Hoy la tecnología Hadoop nos proporciona la flexibilidad para trabajar con estas formas de datos y utilizarlos para decisiones comerciales.

El principal problema de la integración de los datos es el tiempo en el que se realiza, otros retos son:

- Adaptar el alcance de los datos.
- Inconsistencia de datos.
- Optimización de consultas.
- Recursos e implementación inadecuados sistema de apoyo.

- Escalabilidad.

En la actualidad han surgido diferentes herramientas que apoyan al proceso de integración de los datos y logran respetar las 5 Vs y apoyan a solucionar los problemas antes nmencionados. Una tecnología empleada en las bases de Groupón por ejemplo es Talend y ha mostrado ser una gran promesa.

2. Desarrollo de la practica

Para iniciar colocamos los archivos csv en la carpeta compartida de la MV y después creamos un nuevo proyecto de Integración en SQL Data Tools.

Una vez en el proyecto, creamos una tarea de flujo de datos y dentro de ella utilizamos el origen de archivo plano para acceder a los datos en los csv, el siguiente paso fue unir los csv con nombres similares con la herramienta unión de todo.

Lo siguiente fue hacer la ”mezcla” de los dos archivos csv resultantes, y para ello escogimos un formato en el que se pudieran unir ambos archivos, el cual fue en colocar primero los datos (columnas) de los viajeros y después lo de los viajes, donde se colocaron de manera ordenada en base a los CURP’s.

Después de haber realizado la ”mezcla” de los archivos, se guardaron los datos en una tabla **IntegracionViaje** creada en una base de datos en SQL Server llamada **BDPractica3**, de tal manera que las columnas de la tabla coincidieran con el formato del archivo csv que se definió en la mezcla de archivos. Y aquí ocurrió un problema ya que nos empezó a marcar un error en la codificación de caracteres de los archivos, ya que detectaba que un archivo tenía una codificación de caracteres distinta al otro archivo de la mezcla pero ambos **sí** tenían la misma codificación `UTF-8`.

Luego, a partir de la tabla **IntegracionViaje** con los datos ya cargados en la base de datos, se creó y obtuvo un archivo que guarda los datos *integrados*, que lo nombramos como `integracion.csv`.

3. Análisis final

Consideramos que el proceso que se llevó a cabo para la **integración de datos** entre los cuatro archivos no fue tan laborioso como pensábamos antes de iniciar la práctica, pues con la ayuda de **SSIS** notamos que se facilita mucho esta tarea y se puede dar una buena organización en la estructura en el flujo de las tareas que íbamos definiendo, siendo más fácil de entender su funcionamiento. Aunque hubo ocasiones en las que la herramienta **SSIS** tardaba un poco en responder los cambios que hacíamos.