Almacenes y Minería de Datos

Prácticas

DEMO Apache Hop

Joaquín Ángel Triñanes Fernández
Instituto de Investigacións Tecnolóxicas
Joaquin.Trinanes@usc.es Ext: 16001 Externo: 881816001

Reanálisis

- Seguimiento Prácticas:
  - ¿Principales dificultades?
  - ¿Carga de trabajo?
- ¿Qué hemos aprendido?
  - Importar datos a un entorno de BBDD (copy, \copy, \lo_import,\i, FDW, insert,..)
  - Llevar a cabo agregaciones (GROUP BY, COUNT()), y ordenaciones (ORDER BY)
  - Empezar a trabajar con BBDD columnares a través de cstore_fdw
  - Analizar consultas (EXPLAIN ANALYZE)
  - Importar BBDD OLTP/DW que simula las operaciones de una empresa

Apache Hop

- Hop Orchestration Platform. Necesitamos Java 11 para ejecutarlo.
- Projecto de Apache Sofware Foundation (licencia Apache v2)
- Comunidad muy activa y versiones frecuentes.
- Representa un avance respecto a la integración de datos
- Nos permite organizar los flujos de trabajo de datos y metadatos procedentes de múltiples fuentes, poniéndolos a disposición de herramientas analíticas.
- Es un fork de Kettle (Pentaho DI). Podemos importar proyectos PDI.
- Podemos trabajar visualmente. Funcionalidad adicional a través de plugins.
- Los metadatos tiene extrema importancia.
- https://hop.apache.org/download/

Apache Hop

- Herramientas
  - Hop GUI: Diseñar pipelines, workflows,
  - CLI: Hop Run, Hop Search, Hop Encrypt, Hop Translator, Hop Conf
  - Hop Server
- ¿Para qué lo usaremos?
  - Necesitamos cargar grandes conjuntos de datos en nuestras BBDD
  - Migración de datos a otros entornos y limpieza de datos.
  - Integrar datos en entornos heterogéneos (ej: BBDD NoSQL, relacionales, …)
  - Dentro de los procesos ETL como herramienta para ir añadiendo datos a nuestro DW.

Ejemplo Talend

- Datos el ECDC sobre COVID-19: https://opendata.ecdc.europa.eu/covid19/nationalcasedeath_archive/csv/data.csv
- Vamos importar los datos y guardarlos en PostgreSQL
- Otras herramientas similares para Integrar Datos: Tableau Prep, Stitch (adquirida por Talend), Oracle Dta Integrator, Informatica Powercenter, Pentaho DI, AWS Glue, Xplenty, Skyvia,…
- En un DW tradicional, la integración de datos conlleva el uso de dimensiones con atributos comunes en todas las BBDD, y de hechos sobre los que aplicamos métricas communes que puedan ser comparadas en las BBDD.
- Ejercicio adicional Talend usando los datos anteriores del ECDC:
  - Agrupa los casos por país y año - Agrupa los casos por continente
  - Añade a la salida un nuevo campo de incidencia semanal por millón de hab.
  - Toda la información anterior la guardaremos en PostgreSQL.

Ejemplo Apache Hop

- Datos el ECDC sobre COVID-1. Vamos a usar los datos almacenados en PostgreSQL del ejercicio anterior.
  - Número de muertes por país - Número de muertes por continente
  - País con el mayor número de casos por semana
  - Muertes anuales por países con más de 10 millones de población.

  Las salidas las almacenaremos en un fichero csv.