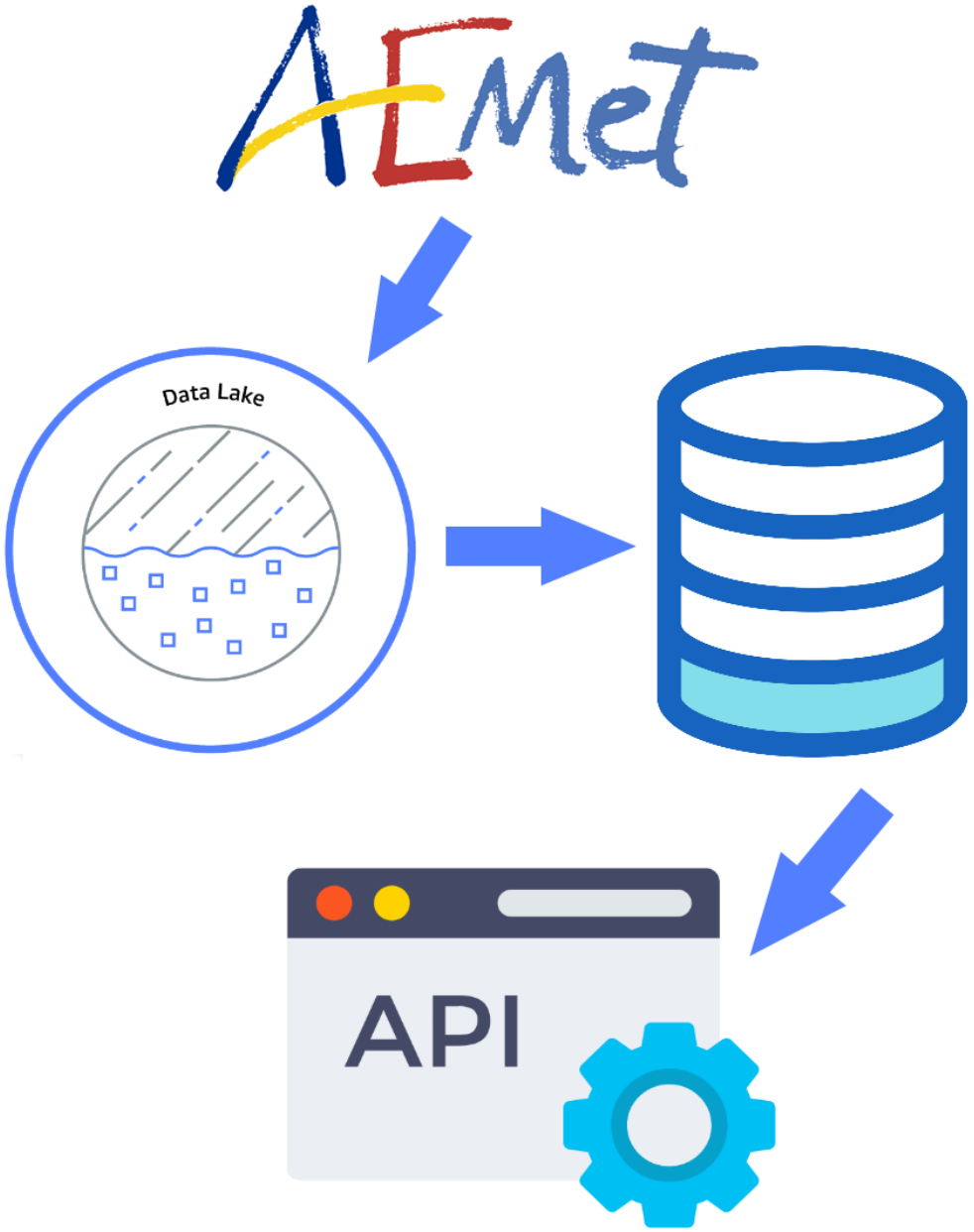
Proyecto datalake to datamart



Asignatura: Desarrollo de Aplicaciones para Ciencia de Datos

Curso: Segundo

Titulación: Grado en Ciencia e Ingeniería de Datos

Escuela: Escuela de Ingeniería informática

Universidad: Universidad de Las Palmas de Gran Canaria

Autor: Jia Hao Yang

Fechas: 13/01/2023

Versiones: versión v1.0

Revisiones totales de la memoria: 1

INDICE

RESUMEN

Este proyecto consistía en la extracción de datos de la webservice que ofrece AEMET para posteriormente, registrar información en formato JSON en unos archivos .events y almacenarlos en un datalake. Con esta información, construiremos un datamart hecho en SQLITE y a partir de aquí alimentaremos a una API REST hecho con el framework de Spark, donde se podrán consultar datos como las temperaturas máximas y mínimas de la isla.

Para cumplir los objetivos de este proyecto, se ha dividido el trabajo en tres módulos.

El primer modulo que se llama "feeder", tiene como objetivo principal recoger los datos de la API de AEMET, crear el directorio donde se alojara el datalake, crear los ficheros .events, escribir sobre esos archivos .events los eventos que vaya registrando cada hora y sensorizarlos.

Un aspecto a destacar de este módulo, es que al inicializar el programa por primera vez, se registran dos archivos .events. Uno del día de hoy y otro del día de ayer ya que la API de AEMET ofrece información de la ultimas 24 horas y por lo tanto, hay que filtrar esos datos y recogerlos. A partir de ahí, se reescribe sobre el mismo fichero del día de hoy los datos que vaya recogiendo, a cada hora, se sobrescribe para no generar eventos duplicados.

El segundo modulo se llama "datamart", este tiene como objetivo leer los archivos .events previamente escritos, crear la base de datos, crear las tablas de las base de datos, y escribir sobre las tablas de la base de datos.

En este módulo también tenemos que destacar otro aspecto importante. Al estar trabajando con tablas y bases de datos y no ficheros comunes, sobrescribir sobre una base de datos no es tan trivial como lo es con un archivo normal. Por lo tanto, el módulo contiene métodos para comprobar que no se repitan los eventos registrados. La manera en la que se realiza no es la más eficiente y a la larga, conforme vaya creciendo nuestra base de datos es mas y más insostenible, la manera en la que lo hace es que lee el nuevo fichero reescrito del feeder, lo añade, por lo tanto, ya tendríamos una tabla con información duplicada pero luego se eliminan los registros repetidos. También cabe destacar que este se ejecutará cada hora en sincronización con el feeder, a medida que el feeder sensorice nuevos datos, el datamart los incluirá a la base de datos.

Y finalmente, el ultimo modulo que se llama "webservice", este tiene como objetivo seleccionar información en especifica de la base de datos creada anteriormente y generar una respuesta que más adelante servirá para mostrarlo a través de una API. Para crear nuestra API REST, el programa hace uso de un framework en Java que se llama Spark. Este está constantemente en ejecución, por lo tanto, no necesitaría un TimerTask como los dos anteriores módulos.

Un detalle importante cuando se vaya a ejecutar el programa es que a pesar de que cada modulo tenga su propio Main, donde se tiene que ejecutar realmente es en otro modulo que se llama "application" que engloba a los tres anteriores y pone en ejecución todos los módulos a la vez.

RECURSOS UTILIZADOS

Entornos de desarrollo: IntelliJ Idea EDU hecho por la empresa JetBrains

Herramientas de control de versiones: Git

Herramientas de documentación: Microsoft Office Word hecho por la empresa Microsoft

DISEÑO

Para realizar este proyecto, se pidió que siguiéramos la estructura de una arquitectura Lambda. En una arquitectura Lambda podemos diferenciar tres partes. Una Batch Layer que en nuestro caso seria el modulo "feeder", una Serving Layer que seria el modulo "datamart" y una Speed Layer que seria en nuestro caso el modulo "webservice".

CONCLUSIONES

Este proyecto ha sido como un “endpoint” que recogía todos los conocimientos que hemos desarrollado a lo largo del curso. Desde como crear una base de datos que fue uno de nuestros primeros proyectos con el de Spotify hasta la construcción de una API REST que fue con el proyecto opcional del Scraper de las cadenas hoteleras.

He aprendido mucho a lo largo de este primer cuatrimestre en el desarrollo de aplicaciones en el lenguaje Java y sé que todavía me queda mucho por mejorar.

En el caso de que, si se empezara de nuevo, sinceramente, creo que lo hubiera hecho de la misma manera o muy similar o como está ahora.

LINEAS FUTURAS

Implementaría técnicas de manejo de bases de datos mas avanzadas y eficientes para su comercialización y alojarla en un servidor con mas capacidad de cómputo que un ordenador doméstico.

BIBLIOGRAFIA

https://www.sqlitetutorial.net/

https://stackoverflow.com/

https://www.javatpoint.com/