Automatización desatendida de descarga de datos mediante RPA para nutrir data lake

Héctor Martín Gutiérrez
Escuela de Ingeniería Informática, Oviedo

Cátedra TotalEnergies de Analítica de Datos e Inteligencia Artificial

Introducción

El primer paso que hay que realizar en analitica de datos es la obtención de información, para posteriormente, estructurarla y tratarla.
En este trabajo se ha estudiado el desarollo de un automatismo para la descarga masiva de datos, con los que, alimentar un lago de datos donde tener centralizada toda la información.

Objetivos del estudio

En este contexto, todas las distribuidoras publican cada día, alojadas en sus servidores, los datos horarios de energía bruta validada. Se trata de una cantidad grande de ficheros, por lo que, sería interesante que se pudiesen descargar de manera desatendida y que se ordenarsen por distribuidora para, poder nutrir un lago de datos.

Métodos

Se ha desarrollado un automatismo capaz de descargar todos los ficheros de manera segura.

Para este fin, se han utilizado las colas del orquestador del RPA, donde se han guardado todos los datos de acceso como path, puertos, directorios o credenciales, optando por la encriptación de estas ultimas.

El automatismo hace uso de flujo de decision, dependiendo de si la cola se encuentra con elementos o, por el contrario, está vacía.

Si se encuentra algún tipo de incidencia en el proceso, se realiza un reporte por correo electronico avisando de donde se encuentra el fallo y facilitando un log al departamento.

Si no hay ningún tipo de problema, el proceso de descarga inicia la aplicación necesaria para la descarga, y filtra los ficheros que se necesitan, para, finalmente, descargarlos, ordenarlos y descomprimirlos.

Figuras y Resultados

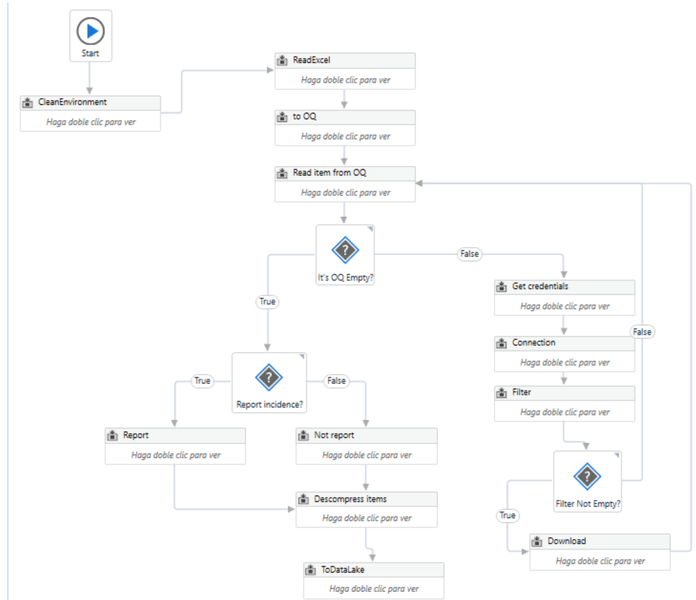

Fig1. *Esquema de funcionamiento*

En la ***fig1*** se puede apreciar de manera esquematica las estructura del automatismo con el que se descargaran todos los datos.

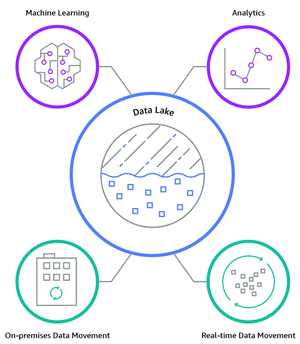

Fig2. *Estructura data lake*

Una vez descargados todos estos datos, se almacenan en un lago de datos. Este almacen es desestructurado, por lo que, ofrece multitud de posibilidades tal y como se aprecia en ***fig2.***

Conclusiones

Automatizar la descarga de gran cantidad de datos de manera manual puede resultar una tarea ardua y llevar a confusiones.
Con el desarrollo de este automatismo completamente desatendido nos aseguramos que, todos los ficheros con los datos necesarios se han descargado.

Alimentar un lago de datos, nos ayudará a tener centralizados todos los datos que necesitamos para nuestros análisis actuales y futuros.

Trabajo futuro

Con todos los datos almacenados en nuestro lago, el camino a seguir sería realizar distintos analisis con estos datos y poder seguir nutriendo el lago. Tambien, el uso de herramientas de reporte conectadas nos daría una visual del estado actual y de años anteriores para realizar analiticas.

TotalEnergies

Universidad de Oviedo