Universitat de Girona
Escola Politècnica Superior

Projecte fi de grau

Estudi: Grau en Enginyeria Informàtica

Títol: Data pipeline de les comunicacions de Som Energia

Document: Resum

Alumne: Pau Boix Tura

Tutor:Esteve del Acebo Peña
Departament: Informàtica, matemàtica aplicada i estadística
Àrea: Llenguatges I Sistemes Informàtics

Convocatòria (mes/any): Setembre 2022

Resum

El projecte tracta sobre el desenvolupament del data pipeline de les comunicacions de Som Energia.

Som Energia és la comercialitzadora més ben valorada pels seus clients i té una gran preocupació per l'atenció al client. En el moment que es va iniciar el projecte però no existia cap eina per poder avaluar mínimament en quin estat es trobava l'atenció al client.

A part d'aquesta necessitat també hi havia la de l'equip de comunicació de poder tenir un històric de les comunicacions i conèixer quin era l'impacte de les seves comunicacions.

Davant d'aquesta necessitat es va decidir realitzar aquest projecte, que tot i que en un principi es volien integrar més fonts de dades, finalment ha consistit en el data pipeline de les comunicacions per correu.

A partir del requeriments inicials es va decidir posar en marxa tota la infraestructura de dades i al prioritzar les necessitats els correus electrònics eren les comunicacions que ens proporcionaven més informació més fàcilment.

Seguint la metodologia SCRUM s'han realitzat 7 sprints on s'han dut a terme les tasques de recull de requisits, anàlisi, desenvolupament i testeig.

Per poder dur a terme aquest pipeline s'ha posat en marxa una infraestructura que permet ser reutilitzada per altres projectes, i que actualment ja s'està fent servir fora del projecte.

Aquesta infraestructura es basa en Apache Airflow com a orquestrador de pipelines, però per aconseguir els objectius del projecte d'obtenir una infraestructura comuna per tots els pipelines de Som Energia no era suficient i s'ha desenvolupat una infraestructura que combina Apache Airflow i Portainer.

Aquesta infraestructura assoleix un gran rendiment al executar el codi en un servidor separat al d'Airflow, al executar-se el codi en contenidors de Docker orquestrats gestionats per Portainer també permet executar codi en diferents entorns virtuals i diferents versions de Python, Aquests contenidors es construeixen a partir d'una imatge que s'ha intentat que fos el més lleugera possible, i sempre comptant amb tots els paquets de Python que requereix l'última versió del projecte instal·lats.

Per controlar tot això s'ha dissenyat una part comuna que tenen tots els DAGs que permet aquesta interacció entre Airflow i Portainer, a més d'automatitzar el desplegament del nou codi, cada DAG té tasques comunes que s'encarreguen d'interactuar amb Github i Portainer per tal de poder executar sempre l'última versió del codi que exiseixi en un contenidor amb tots els requeriments necessaris. Aquest codi es troba en el servidor on s'està executant Airflow mitjançant

docker-compose i carregant-se mitjançant NFS a dins de cada contenidor on s'executa.

Tota aquesta infraestructura s'ha posat a prova amb el pipeline del projecte, que ha implicat capturar tots els correus que han arribat a Som Energia des del març del 2020, i mitjançant un procés ELT amb un data lake i un data warehouse, dues bases Postgres, el data lake on es guarden la informació original que ens retorna l'api i el data lake que s'ha modelat utilitzant SQLAlchemy ORM.

Tot orquestrat per Airflow, s'han transformat totes les dades obtingudes a través de l'API de Helpscout per poder fer-ne un anàlisi.

Un cop ja es disponia de les dades netes en el data warehouse es va procedir a realitzar una visualització de les dades mitjançant Apache Superset, s'ha fet un dashboard seguint els requeriments del equip tècnic.

En aquest dashboard es realitza un anàlisi de les principals bústies d'atenció al client, acompanyat de gràfics que donen informació com el temps de resposta als correus, el nombre de correus contestats, el nombre de correus rebuts. També s'analitza l'arribada dels correus i es relacionen amb els diferents events que hi ha hagut darrerament en el mercat elèctric.

També s'analitzen els correus que han causat més impacte durant els últims anys, i s'explora la teoria de la factura mensual com la comunicació més important que fa Som Energia, i una petita ullada a les compres col·lectives.

En aquest dashboard es pot veure l'evolució històrica dels correus que han arribat, també com ha evolucionat el seu temps de resposta,

Al haver fer tot el pipeline en Airflow s'actualitzen cada hora amb noves dades automàticament tots els gràfics i per tant també servirà per a part d'avaluar la feina passada i preveure les accions futures, també tenir una eina per saber a temps real l'impacte de les comunicacions que es fan sobre l'atenció al client i si aquest s'ha saturat.

Al acabar el dashboard ja es va fer una presentació dels resultats a persones de l'equip tècnic que van estar molt satisfetes amb el resultat obtingut i van acabar d'ajudar a donar explicacions a alguns dels fenòmens que es poden veure a les gràfiques.

També cal destacar que la infraestructura fins ara només ha fallat en una ocasió i gràcies a com estan programats els DAGs no es va perdre cap dada. La infraestructura també s'està utilitzant per altres projectes amb èxit i l'equip està molt satisfet amb els avantatges que aporta utilitzar airflow en comptes de fitxers cron com s'estava utilitzant fins ara.