
This is the **published version** of the bachelor thesis:

Alonso González, Anna; Acuña Torras, David, dir. Eina de traçabilitat de consultes sobre una plataforma de gestió de les dades en una administració pública. 2022. (Enginyeria Informàtica)

This version is available at <https://ddd.uab.cat/record/280717>

under the terms of the  license

Eina de traçabilitat de consultes sobre una plataforma de gestió de dades en una administració pública

Anna Alonso González

Resum—Aquest projecte aporta un valor real ja que ha estat desenvolupat dins de la empresa NTT DATA sobre un projecte amb un client existent. En aquest article s'exposa el desenvolupament fet per tal d'obtenir una eina de traçabilitat de consultes sobre una plataforma de gestió de dades en una Administració Pública. La seva finalitat és que empresaris puguin obtenir, a través del seu portal del empresari, quines dades de les seves empreses han estat consultades, per qui i a quina data. Aquestes consultes sobre les dades queden enregistrades en arxius de tipus log file, per tant, es desenvolupa un procés per tal d'obtenir la informació necessària d'aquest arxiu i del model on es troben les dades. A més, es realitza un report sobre aquestes consultes realitzades per a que el personal tècnic de l'Administració pública client pugui fer estudis sobre aquestes dades. Aquest personal tècnic també es qui realitza les consultes sobre les dades dels empresaris dins del model de dades.

Paraules clau—Informatica, consulta, DEE, empresa, establiment, registre, script, ETL.

Abstract— This project brings real value as it has been developed within the company NTT DATA on a real project with an existing client. This article presents the development carried out to obtain a query traceability tool on a data management platform in a Public Administration. Its purpose is to allow business owners to access, through their business portal, the information on which data from their companies have been queried, by whom, and on what date. These data queries are recorded in log file-type archive, therefore a process is developed to obtain the necessary information from that file and from the data model where the data is located. Additionally, a report is generated on these queries for the technical staff of the client Public Administration to perform studies on this data. This technical staff is also the one who performs the queries on the data of the business owners within the data model.

Index Terms—Informatica, query, DEE, business, establishment, register, script, ETL.



1 INTRODUCCIÓ - CONTEXT DEL TREBALL

A MB l'entrada en vigor el 20 de febrer de 2021 de la llei 18/2020 de facilitació de l'activitat econòmica, la qual la seva finalitat és "facilitar l'activitat econòmica en l'entorn digital" sorgeixen diverses necessitats per part de les administracions públiques per tal de complir aquesta llei.

Com a solució a part d'aquestes necessitats es crea, en concret per la nostra administració pública client, el Directori d'Empreses i Establiments (DEE) que és un repositori on es troba tota la informació relacionada amb les empreses: els seus establiments, contactes, registres i adreces. Això, permet obtenir/proporcionar informació de les dades emmagatzemades en diferents fonts, per tal d'unificar-les i dotar-les de qualitat amb una estructura normalitzada. A més, que els titulars d'activitats econòmiques, és a dir, empresaris, puguin tenir una visió completa de les dades que disposen totes les administracions, a través de la

integració amb l'Àrea Privada o portal de l'empresari de l'empresa, fent possible la consulta i recuperació de les dades. Aquesta Àrea Privada és un portal fet pels empresaris per tal que puguin fer consultes de les seves dades emmagatzemades al DEE.

Com s'ha dit just abans es faran possibles les consultes i recuperació de les dades, i aquí és on entra el propòsit d'aquest projecte. Els titulars d'activitats econòmiques també han de tenir el dret de veure qui ha fet i quan hi ha hagut una consulta sobre les seves dades que es troben en aquest directori.

Llavors, l'objectiu principal d'aquest treball es desenvolupa per una eina que permeti als empresaris rebre la informació sobre quines consultes s'han fet sobre les seves dades, per quin usuari i quan s'han realitzat.

-
- E-mail de contacte: 1562628@uab.cat
 - Menció realitzada: Enginyeria del Software
 - Treball tutoritzat per: David Acuña Torras (Departament de Ciències de la Computació)
 - Curs 2022/23

1.1 Funcionament portal de l'empresari i Informatica Data Director

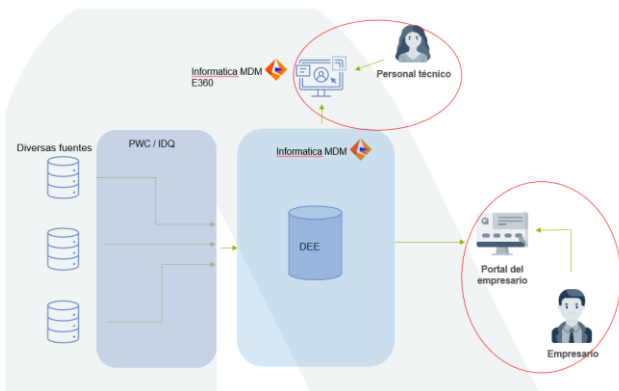


Fig. 1: Accés al DEE.

Com es pot observar en aquest esquema, els empresaris accedeixen a través del portal de l'empresari a les dades que es troben dins del DEE (Directori d'Empreses i Establiments). Per tal d'accedir es necessita un usuari i una contrasenya amb els permisos necessaris per consumir aquestes dades. El portal de l'empresari és una aplicació de Informatica, en concret Informatica Data Director, que permet als usuaris empresarials consumir i supervisar dades principals de forma eficaç. Des d'on accedeix el personal tècnic es una altra aplicació de Informatica, en concret Informatica E360. Aquesta aplicació es troba dins del servidor MDM i també permet realitzar consultes sobre les dades. Per tant, el que es vol aconseguir és que aquest empresari pugui veure si algun usuari de l'administració pública o personal tècnic ha consultat les dades de les seves empreses que es troben al DEE i si és així, quines dades han estat, qui les ha consultat i a quina data.

1.2 Objectius

Aquest treball es basa en desenvolupar una funcionalitat que permeti a l'usuari final, en aquest cas, empresaris, tindre una traçabilitat de les consultes que es fan a les seves dades dins del DEE.

Llavors els principals objectius que ha de complir de forma general per considerar que s'han satisfet les necessitats del client són els següents:

1. Que els empresaris rebin al seu portal de l'empresari quines de les seves dades han estat consultades, per quin usuari i a quina data.
2. Que el personal tècnic poden veure un report de les consultes fetes al directori DEE.

Cal indicar que el primer objectiu és enviar les dades perquè arribin al portal, però en cap cas s'indica que es desenvolupi la interfície per tal de mostrar aquestes dades als empresaris. La creació d'aquesta interfície depèn de l'administració pública client, nosaltres només hem d'enviar-li les dades de les consultes fetes.

A continuació es redacten els objectius més tècnics que descriuen amb més detall quines són les tasques a realitzar

durant el desenvolupament del treball segons prioritització d'assoliment:

1. Documentar i exposar de forma completa la justificació i context del treball a realitzar.
2. Entendre i explicar les eines de desenvolupament del projecte i fer un estudi sobre quines eines es faran servir per tal d'aconseguir l'acompliment de forma més eficient els objectius principals.
3. Visualitzar i comprendre el model de dades ja existent i fer un estudi sobre el nou model que s'haurà de crear per visualitzar les dades de les consultes.
4. Enviament de l'arxiu log generat al servidor de la eina Informatica que permetrà realitzar el procediment complet del treball.
5. Creació d'un script capaç de llegir un arxiu log on s'ha emmagatzemat l'historial de les consultes i que emmagatzemi la informació necessària per poder ser explotada posteriorment.
6. Creació i configuració dels processos en l'eina Informatica per executar l'script i emmagatzemar les dades al model (creació dels mappings i el workflow).
7. Explotació de les dades mitjançant PowerBI per tal de poder fer un report útil.
8. Enviament correcte i complet mitjançant processos ETL de les dades sol·licitades pel empresari al seu portal en format XML per poder ser visualitzades.
9. Generar l'article final del desenvolupament sencer del treball.

2 METODOLOGIA

La metodologia utilitzada pel desenvolupament d'aquest treball és una àgil amb Kanban.

L'objectiu de les metodologies àgils es basa en repartir el treball en parts més petites que han de completar-se i entregar-se en un període de temps. En aquest tipus de metodologia es realitzen reunions, ja siguin diàries o setmanals amb l'equip per veure quin ha estat l'avanç de les tasques o si hi ha hagut cap imprevist que endarrereix l'entrega. Durant aquest projecte es farà cada setmana dues reunions amb l'empresa per tal de tenir un control sobre el projecte. A més, durant tot el desenvolupament del projecte poden produir-se canvis, cosa que aquesta metodologia permet acceptar aquests canvis.

Per altra part, utilitzaré el mètode Kanban que considero que és molt visual i permet veure amb més claredat l'estat actual del projecte. Kanban es basa en un tauler amb, originalment, tres columnes on a la primera es troben les tasques a realitzar, en la segona les tasques que s'estan executant, i en l'última les tasques ja realitzades. En el meu cas he considerat afegir una columna més entre les que s'estan executant i les realitzades que és en revisió. Aquesta columna em permetrà veure quines de les tasques es troben totalment finalitzades o quines necessiten una revisió o un

testage abans de donar-les per acabades.

Pel control de les tasques amb el mètode Kanban he creat un projecte a l'aplicació web Jira que permet la gestió del desenvolupament de projectes de forma àgil. Aquí es on es crearan les tasques a realitzar i s'aniran canviant de columna en columna. S'ha afegit a les tasques el camp obligatori de data màxima per tal de tindre un control i seguir la planificació prevista. El tauler Kanban resultant es pot veure al *apèndix Kanban* [A1].

3 ESTAT DE L'ART

L'estat de l'art d'un projecte es refereix al nivell més avançat i actual de coneixement, tecnologia o recerca en un camp específic relacionat amb aquest projecte. És una avaluació exhaustiva i actualitzada dels avanços, descobriments, metodologies i pràctiques rellevants en l'àrea temàtica en la qual es troba el projecte.

Com ja s'ha explicat, aquest treball es basa en obtenir una traçabilitat de les consultes fetes sobre una base de dades i enviar la informació d'aquestes consultes al portal de l'empresari per aconseguir que ells puguin tindre un control sobre les seves dades.

Existeixen diverses eines que son capaços d'obtenir les consultes en format SQL que s'han fet sobre una base de dades. Per exemple, per una base de dades Oracle existeix 'Oracle Audit Vault and Database Firewall' que proporciona una solució d'auditoria completa i permet la captura i el registre de consultes. També existeixen extensions per altres base de dades com per a PostgreSQL la extensió 'pgAudit' que permet el mateix.

Però, aquestes eines i/o extensions només aporten les consultes fetes sobre una base de dades i no un procés complet que sigui capaç d'extreure la informació important sobre les consultes com la data en que s'ha realitzat o quin usuari ha realitzat aquesta consulta. Aquest procés es realitza en aquest treball en la creació i l'execució de l'script en Java que extreu quina es la informació realment important d'aquestes consultes. A més, dins d'aquest treball es realitzen processos posteriors de qualitat de les dades, en els qual es 'netejen' les dades i es posen en un format estàndar per tal d'emmagatzemar-les en la base de dades de forma consistent. Finalment, com que l'objectiu és enviar aquestes dades de les consultes, també es realitza un altre procés que és capaç d'entendre les dades que es troben guardades sobre les consultes i extreure més informació sobre el model per tal d'enviar el que realment és necessari al client. Per tant, no existeix cap eina o aplicació que realitzi el cas concret d'aquest treball perquè és per a un cas d'un client en concret amb unes dades en concret.

4 PLANIFICACIÓ

Per la planificació del projecte s'ha elaborat un *Diagrama de Gantt* [A2] a on es mostra les diferents fases i les subtasques amb la seva duració, les dates de començament i fi i els recursos que es necessiten.

Hi han 4 recursos que s'han repartit en les tasques: Cap de Projecte, Dissenyador, Gestor Documental i Programador.

El Cap de Projecte s'encarrega de les tasques que son importants per la definició del projecte com la configuració de l'eina de seguiment Jira. El Gestor Documental és l'encarregat de realitzar tota la documentació dins del projecte. El Programador és qui s'encarrega de realitzar les tasques de desenvolupament del projecte. I finalment, el dissenyador s'encarrega de les tasques en les que s'han de dissenyar tasques. Cal indicar que s'ha establert que un dia de treball equival a 4 hores de treball, per tant son 20 hores setmanals de treball. També s'ha establert que no hi ha hores de treball el cap de setmana, només de dilluns a divendres.

5 EINES UTILITZADES

En el projecte en el que es desenvolupa aquest TFG ja hi han unes eines establertes per l'empresa per tal de fer el millor desenvolupament possible. Les eines principals amb les que s'ha desenvolupat el projecte i, per tant aquest treball, han estat Informatica i Denodo.

5.1 Informatica

Informatica és un software d'integració de dades i gestió de dades empresarials desenvolupada per Informatica Corporation. Permet a les organitzacions extreure, transformar i carregar dades (ETL), així com integrar i gestionar dades en una varietat d'entorns i sistemes. És una de les aplicacions escollides ja que com s'ha dit és líder en el mercat i, a més, a l'empresa on es desenvolupa aquest projecte, porten anys d'experiència treballant amb aquesta aplicació. Això m'ha permet aprendre amb formacions dins de l'empresa i per tant, facilitar l'ús d'aquest software per la realització d'aquest TFG.

Aquest software té diverses eines que abasten diferents aspectes del cicle de vida de les dades, però en aquest document només s'explicaran les utilitzades que son:

- **PowerCenter:** És la principal eina ETL (extracció, transformació i càrrega) de Informatica. Permet als usuaris dissenyar, programar i executar fluxos de treball per a extreure dades de múltiples fonts, transformar-les segons les regles de negoci i carregar-les en el destí desitjat. En el cas d'aquest treball s'utilitza per realitzar el flux de treball que s'encarrega d'executar els processos creats en Informatica Developer. Per tant, un flux de treball (workflow) es refereix a un conjunt de tasques o accions seqüencials que s'executen en un ordre predefinit per aconseguir un objectiu específic d'integració i processament de dades. Ens permet unificar l'execució dels processos.

o **Informatica Developer:** És un entorn de desenvolupament integrat (IDE) utilitzat pel desenvolupament de processos d'integració de dades. És a dir, ens permet crear processos per ser executats per tal d'assolir un objectiu. Ha sigut l'eina que s'ha utilitzat per la creació dels processos que realitzen la funcionalitat d'executar l'script i emmagatzemar a la base de dades la informació necessària.

- **Master Data Management (MDM):** És un component de Informatica que ajuda les organitzacions a gestionar i mantenir dades mestres de manera centralitzada i coherent en tota l'empresa. Les dades mestres són un conjunt de dades

clau i fonamentals que representen entitats comercials o conceptes centrals dins d'una organització, com a dades de clients, productes o proveïdors.

o **Informatica MDM Data Director:** És una interfície d'usuari basada en web que permet als usuaris empresarials i tècnics col·laborar en la gestió i administració de dades mestres en un entorn de MDM. En el nostre treball, s'ha utilitzat per la creació del portal on accedeix el personal tècnic, el qual és el que permet realitzar les consultes en les quals es basa aquest treball, que després queden enregistrades.

o **Informatica MDM Provisioning:** S'utilitza per a gestionar i configurar els models i la configuració de dades mestres en un entorn de MDM. És a dir, ens permet crear un model sobre la interfície creada amb Data Director per mostrar només les dades del nostre model que en son rellevants. Permet la creació de vistes i transformacions de les dades per aconseguir una millor visualització de les dades. En el nostre projecte de l'empresa, s'ha utilitzat per la creació de diferents vistes sobre la plataforma de Informatica E360 que és d'on el personal tècnic realitza les consultes. Per tant, aquestes vistes ens indiquen quines de les dades del model del DEE s'han mostrat realment.

5.2 Denodo

Denodo és una plataforma de virtualització de dades que permet a les empreses accedir i gestionar dades de múltiples fonts i formats de manera unificada i coherent. Amb Denodo, els usuaris poden crear vistes virtuals de les dades que estan emmagatzemats en diferents sistemes i aplicacions. Pel desenvolupament del projecte a l'empresa s'ha utilitzat per la creació de diferents vistes de diferents fonts per tal d'aconseguir que l'administració pública client visualitzi les dades que en son rellevants per ella. Per aquest projecte, ha permet fer una virtualització de les noves taules que es creen dins del model del DEE amb la informació de les consultes per tal de realitzar la connexió amb PowerBI i realitzar el report de les dades.

6 DESENVOLUPAMENT

Un cop s'ha posat en contexte el treball a realitzar comença la fase de desenvolupament d'aquest per tal d'assolir els objectius.

El primer punt d'aquest desenvolupament constarà en la realització del enviament del log, que es troba al servidor de MDM des d'on el personal tècnic realitza les consultes, al servidor de Informatica PowerCenter(IDQ) on ja es podrà treballar amb ell.

Un cop el log ja es troba al servidor de PowerCenter, es realitzarà la creació del script que llegirà aquest arxiu i generarà un arxiu csv amb les dades que son rellevants.

Es crearan unes noves taules al model del DEE per tal d'emmagatzemar aquesta informació resultant del csv generat per l'script.

Després es faran els processos dins de Developer i PowerCenter que executaran aquest script i llegirà aquest csv generat i ho emmagatzemarà a les taules creades.

Finalment, es realitzarà el report amb PowerBI de les dades i es crearà un nou procés ETL complet en PowerCenter que

generarà un XML amb la informació emmagatzemada en les taules sobre les consultes i s'enviarà al portal de l'empresari per ser consumides pels empresaris.

6.1 Enviament log de servidor MDM a IDQ

Per tal de realitzar l'enviament del arxiu log s'havia de pensar un procés que enviés aquest arxiu log d'un servidor a un altre perquè sigui processat. La solució proposta que s'ha implementat ha sigut amb l'utilització del comando *curl* en un script .sh de Linux i el programa *cron* per programar la execució d'aquest script.

El comando *curl* és una eina de línia de comandos utilitzada per transferir dades a través de diversos protocols. En aquest cas, la seva utilització ha sigut per enviar a través del protocol SFTP l'arxiu log del servidor de MDM al de PowerCenter.

El comando *curl* creat s'encarrega de connectar-se a la IP amb el port de la màquina, en concret a la ruta que és on es guardarà el log. Ho fa amb el usuari i la contrasenya per accedir a aquest servidor, en aquest cas al de PowerCenter. Després, se li indica amb la opció -T quin arxiu es vol enviar de la pròpia màquina, la de MDM, i finalment, amb l'opció -ftp-create-dirs en cas que no existeixi ja una carpeta en aquesta ruta, en crea una dins on guardarà el arxiu enviat. Aquest comando es crea dins d'un script *auditoria_accesslog.sh* de Linux.

Després de la creació de l'script que s'encarrega d'enviar el log d'una màquina a una altra, falta crear el procés que farà que s'executi de forma automàtica cada dia a una hora determinada. Cron és un programa d'administració de tasques programades. Per tant, ens permetrà configurar l'execució de l'arxiu *auditoria_accesslog.sh* per a que s'executi quan ens interressi. Per tal de configurar el cron, hem d'utilitzar el comando *crontab -e* per veure els processos cron creats i editar-los. El comando creat té diversos paràmetres, que en ordre son: minuts, hora, dia del mes, mes, dia de la setmana, comando. S'ha programat per a que s'executi al minut 0 de la hora 0 sense cap restricció de dia, mes o dia de la setmana. Per tant, executarà l'script *auditoria_accesslog.sh* cada dia a les 00:00h. És a dir, que cada dia s'enviarà el log desde MDM a PowerCenter a les 00:00h.

6.2 Creació de l'script

Un cop l'arxiu log amb les consultes fetes es troba dins de la màquina de PowerCenter ja podem començar a tractar l'arxiu. Aquest arxiu log guarda totes les consultes realitzades fins a un límit màxim de tamany de 5GB. Un cop arriba a aquest tamany, es restableix a zero.

La principal funció de l'script és emmagatzemar la informació rellevant del log file on es troben les consultes per tal de poder utilitzar-la després. L'arxiu log registra peticions de tipus HTTP, de tipus POST i GET. En el apèndix es mostra un exemple de les *Línies Arxiu Log* [A3].

Les de tipus POST no ens aporten res d'informació rellevant, ja que quan es fa una consulta és de tipus GET.

Dins de les línies de tipus GET, només ens interessen aquelles que seguidament després del GET tenen el patró `"/cmx/cs/emo_dee_web-CMX_ORs/"`, ja que això és el que ens indica que s'està fent una consulta a la base de dades CMX_ORs que és on es troben les taules del nostre

model DEE.

A més, just després del patró de la base de dades, podem veure “/VISTA_XXXX”, que ens indica quina ha estat la vista que s’ha mostrat, és a dir, quines de les taules del nostre model del DEE s’han mostrat: empresa, establiment, registre, contactes i/o adreça. Aquesta vista prové de la eina mencionada anteriorment Informatica MDM Provisioning.

Seguidament de la vista es veu un nombre, per exemple el “10005”. Aquest nombre és el ID únic de empresa/establiment/registre dins del model del DEE (Directori d’Empreses i Establiments). Aquest ID canviarà segons quina vista ha consultat, és a dir, si la vista és d’una empresa, aquest ID serà de l’empresa, si la vista es d’un establiment, el ID també serà de l’establiment. Per tant, amb aquest ID i la vista que s’ha mostrat, ja podem extreure tota la informació que necessitem: quina empresa s’ha consultat i quines dades exactament: registres, establiments, contactes i/o adreces.

Per últim, com es pot veure en totes les línies, es troba la data en la que s’ha realitzat i l’usuari que ha realitzat aquesta consulta, en el cas de l’exemple admin o DEE_OGE.

El llenguatge de programació utilitzat ha estat Java ja que alhora de després insertar aquest script dins de Informatica Developer era més senzill perquè aquest programa treballa sobre Java.

Com a resum general del funcionament del script: llegeix i obre l’arxiu log, genera un patró amb regex per trobar les línies que son rellevants. Després s’escriuen els noms de les columnes del csv. A continuació, comença a llegir línia per línia el log i quan troba una coincidència amb el patró regex, emmagatzema la vista i l’ID únic dins del csv. Seguidament, busca amb altres patrons regex l’usuari i la data. Fa una transformació de la data per tal que es guardi fins als minuts i ho emmagatzema al csv. Finalment tanca els arxius.

El arxiu csv generat per l’script serà el que es llegeixi des de Informatica Developer per tal d’emmagatzemar la informació al model de dades DEE. Es pot veure a l’apèndix el CSV Generat Per l’Script [A4].

6.3 Creació taules al model

Per tal d’emmagatzemar la informació resultant del csv creat per l’script, sorgeix la necessitat de la creació de dues noves taules dins del model de DEE. Una primera taula (LOG_INFO) on es guardarà aquesta informació que es troba en el csv generat per l’script, per tal de tindre-la dins d’una taula de forma més consistent que en un arxiu. Aquesta taula conté els següents camps amb els seus respectius tipus:

123 ROW_ID	1	NUMBER(38,0)
ABC VISTA	2	VARCHAR2(100)
ABC ID_ROWID_OBJECT	3	VARCHAR2(10)
DATA	4	DATE
ABC USUARI	5	VARCHAR2(75)

Fig. 2. Taula LOG_INFO

Després una segona taula (VISTAS_PROVISIONING) on

es troba la relació entre el nom de la vista (camp VISTA de la taula anterior) i la informació que conté aquesta vista, és a dir, per exemple, si només conté informació de la empresa, o de empresa i registre o només del registre, etc. Per tant, per cada fila haurà el nom de totes les vistes que es poden mostrar-se i un booleà indicant si aquesta vista mostra informació de: empresa, establiment, registre i/o contacte. Sent un 1 que sí mostra informació d’aquella taula i un 0 que no. La definició de les columnes es tal que així:

ABC VISTA	1	VARCHAR2(75)
ABC EMPRESA	2	CHAR(1)
ABC ESTABLIMENT	3	CHAR(1)
ABC REGISTRE	4	CHAR(1)
ABC CONTACTE	5	CHAR(1)

Fig. 3. Taula VISTAS_PROVISIONING

Després amb una sèrie de comandes SQL del tipus “INSERT INTO” es van inserir la correspondència entre cada vista i la resta de camps. On el resultat d’aquests inserts es fa la taula completa de correspondència entre la vista que s’ha mostrat i la informació que ha donat del model del DEE.

Aquesta taula s’utilitzarà més endavant alhora de fer l’enviament de les dades al portal de l’usuari. Amb la informació de l’ID únic i la vista de la taula LOG_INFO i la correspondència d’aquesta vista amb les taules del model, es podrà realitzar una consulta al model del DEE amb aquest ID únic a les taules que aquesta vista mostra (Empresa, Establiment, Registre o Contacte). Aconseguint així que se mostri a l’empresari tota la informació real que ha sigut mostrada de la seva empresa.

6.4 Procès ETL d’execució de l’script i emmagatzematge de les dades al model

Un cop creat l’script i la taula on s’emmagatzemarà el resultat, fa falta crear el procés que llegirà l’arxiu csv generat per l’script i ho guardarà a la base de dades. Aquest procés serà implementat a través de Informatica Developer i PowerCenter.

Com s’ha explicat a l’apartat 5.1 Informatica, Developer es un IDE que permet la creació de processos d’integració de dades i PowerCenter permet fer la integració completa d’aquesta integració, és a dir, permet executar o crear, en aquest cas executar, els processos desenvolupats a Developer.

Per tant, amb Developer es crearan, en concret, dos mapatges (mappings) que permetin executar l’script i després guardar el resultat emmagatzemat en el csv en la base de dades. Després, amb PowerCenter, es crearà el flux de treball (workflow) per tal d’executar els processos o mapatges creats amb Developer.

Tant Developer com PowerCenter funcionen amb transformacions alhora de tractar les dades. Una transformació és refereix a una funció o procés específic que s’aplica a les dades durant el flux de treball d’una ETL. S’utilitzen per a manipular, netejar i enriquir les dades abans de carregar-les en un magatzem de dades o en una altra ubicació de destí.

El primer mapatge que es crea és el que ha d'executar l'script. Conté només 3 transformacions, una de tipus Lectura, una de tipus Java i una de tipus Escriptura. Tant la de lectura com la d'escriptura no tenen cap funcionalitat real, només permeten que es pugui executar la transformació de Java. Son necessàries perquè un mapatge sempre ha de tindre una font (source) y un destí (target), si no, no es pot executar. Es pot veure el mapatge en l'apèndix *Mapatge Execució Script Java* [A5]. La transformació que executa la funcionalitat d'executar l'script es la de tipus Java. Aquesta permet inserir un codi en llenguatge Java i compilar-lo i executar-lo. Dins de les propietats, al apartat "Código Java" és on es troba el codi que s'ha explicat en el punt anterior 6.2 Creació del script.

El segon mapatge és el que s'encarrega del processament de les dades generades en el arxiu csv donat per l'script per guardar-les a la base de dades. El funcionament general d'aquest mapping es basa en calcular el MD5 de les dades que arriben del csv, un hash únic per cada combinació dels camps rellevants (VISTA, ID_ROWID_OBJECT, DATA i USER). Després es fa una cerca pels camps rellevants a la taula LOG_INFO i es calcula de nou el MD5 (MD5_TABLE) d'aquestes línies que ja es troben dins de la base de dades. Seguidament, es comprova que el MD5_TABLE sigui null, perquè si ho és vol dir que les línies que provenen del csv NO es troben dins de la taula LOG_INFO i han de ser inserides. Si no fos null, vol dir que ja existeixen aquestes línies dins de la taula, per tant, NO han de ser inserides per no tindre valors duplicats. Aquest procés de comprovar que les línies no es troben ja dins de la taula de la base de dades calculat amb la funció MD5 ve donat pel motiu que com l'arxiu log no es restableix cada cert temps, si no que es fa quan arriba a 5GB, com s'ha mencionat abans, cal comprovar que no estem inserint dades duplicades. Ja que pot ocórrer que s'executin els mappings a l'hora que li toca i l'arxiu log contingui les mateixes línies perquè no s'han realitzat més consultes i, per tant, no ha arribat a tamany màxim i no s'ha restablert. Es pot veure el mapatge a l'apèndix *Mapatge Guardat Dades* [A6].

Un cop creats els dos mapatges que processen les dades, ha de haver-hi un flux de treball que els executi. Aquí es on es crea el workflow dins de PowerCenter. Aquest workflow executa dos ordres (commands) que s'encarreguen d'executar una aplicació que és on es troben els mapatges del projecte. Dins d'aquest comando s'ha d'indicar quin dels mapatges que hi ha s'han d'executar, en aquest cas els dos anteriors. També, dins del workflow es permet configurar el scheduler, és a dir, cada quan s'han d'executar aquests processos. Aquesta configuració s'ha establert per a que es realitzi cada dia a les 00:15h, uns 15 minuts més tard que el procés fet en el apartat 6.1 Enviament log de servidor MDM a IDQ per tal de donar-li temps a que el log s'envii de forma correcta d'un servidor a un altre abans d'executar aquests processos. Es pot veure aquest workflow en l'apèndix *Flux de treball guardat dades* [A7].

6.5 Report de les dades en PowerBI

Un cop s'ha emmagatzemat tota la informació dins de les taules creades (LOG_INFO i VISTAS_PROVISIONING) al model, podem realitzar un report amb aquestes dades que

sigui útil pels funcionaris i personal tècnic de l'administració pública.

Aquest report de les dades es realitzarà amb l'eina de PowerBI. Power BI és una suite d'eines d'anàlisi empresarial desenvolupada per Microsoft. Permet a les organitzacions visualitzar i analitzar dades de manera interactiva i compartir informació de manera més efectiva. Per tant, ens permet realitzar reports amb la informació per tal de permetre que els funcionaris i personal tècnic de l'administració pública puguin realitzar estudis sobre les consultes fetes o prendre decisions sobre les dades.

Abans de començar a crear el report, s'han de definir els indicadors que es volen mostrar, és a dir, quines dades i de quina manera es volen representar. Basant-me en les dades que es trobaven dins de les taules LOG_INFO i VISTAS_PROVISIONING, vaig decidir a realitzar els següents reports:

- Nombre totals d'empreses consultades
- Nombre totals d'usuaris que han fet consultes
- Nombre total de consultes
- Nombre de consultes per cada empresa i per usuari
- Evolució del nombre de consultes per la data
- Les últimes N consultes fetes
- Mapa geogràfic amb nombre de consultes segons l'adreça de l'empresa

Un cop definit què es vol mostrar es comença amb la creació en PowerBI. Primer de tot, es va fer una connexió a la base de dades per tal d'importar les taules dins de l'aplicació per treballar amb les dades. Una vegada ja importades, s'han de fer unes petites transformacions per 'netejar' les dades i tindre el que interessa. La majoria de les transformacions realitzades sobre les dades ha sigut l'eliminació de columnes que no ens aporten gaire informació i la comprovació de que el tipus de la dada (si es string, int, etc) està correcte. Després de les transformacions ve la creació del model de les taules dins del PowerBI, on es defineixen les relacions entre les taules del model per tal de poder obtenir informació de diverses taules.

El resultat del report es mostrarà a l'apartat de resultats, ja que és un dels objectius del treball que s'havia d'assolir.

6.6 Procés ETL de creació XML i enviament al portal de l'empresari

Per tal d'assolir el primer objectiu d'aquest treball, s'ha de desenvolupar un procés que envii les dades emmagatzemades a les taules creades i dades que ja es trobaven dins del model del DEE, com les dades de l'empresa/establiment/registre, en format XML al portal del empresari.

Els camps definits per enviar en aquest arxiu XML han sigut els que s'han considerat més importants quan l'empresari consulti qui ha vist les seves dades. Els camps definits son:

- Data: la data en la que s'ha realitzat la consulta. Es troba a la taula LOG_INFO.
- User: l'usuari que ha fet la consulta. Es troba a la taula LOG_INFO.
- Vista: quina vista de Informatica Provisioning s'ha mostrat al portal. Es troba a la taula LOG_INFO.
- DocumentEmpresa: nif de l'empresa que ha estat

consultada, no pot ser null, sempre ha de tenir un valor. Es troba a la taula C_B_EMPRESA del model del DEE.

- IdUnicEstabliment: identificador únic de l'establiment. Només s'informarà aquest camp si la consulta ha estat a un establiment. Es troba a la taula C_B_ESTABLIMENT del model del DEE.

- CategRegistre: de quin catàleg prové el registre consultat. Només s'informarà aquest camp si la consulta ha estat a un registre. Es troba a la taula C_B_RGSTR del model del DEE.

- NumeroRegistre: identificador únic del registre. Només s'informarà si la consulta ha estat a un registre. Es troba a la taula C_B_RGSTR del model del DEE.

- Empresa: camp que conté un 1 si la empresa ha estat consultada un 0 si no és així. Es troba a la taula VISTAS_PROVISIONING.

- Establiment: camp que conté un 1 si el establiment ha estat consultat un 0 si no és així. Es troba a la taula VISTAS_PROVISIONING.

- Registre: camp que conté un 1 si el registre ha estat consultat un 0 si no és així. Es troba a la taula VISTAS_PROVISIONING.

Tots aquests camps definits son els que veurà l'empresari des del seu portal quan vulgui comprovar quines consultes s'han realitzat a la seva empresa. Aquests camps vindran informats per cada consulta que s'hagi realitzat.

Com s'ha dit, hi ha camps que provenen de diferents taules, LOG_INFO, C_B_EMPRESA, C_B_ESTABLIMENT... Per tant, s'ha de realitzar un procés que uneixi aquestes taules per obtenir la informació definida. Aquest procés s'implementarà des de PowerCenter per crear una ETL completa. Després de definir el procés s'haurà de crear el flux de treball que ho executi i s'encarregui d'enviar aquest fitxer XML generat cap al portal de l'empresari per poder ser consumit.

Aquest procés generat té la funcionalitat de segons les dades que arriben de la taula de LOG_INFO i la seva vista amb la correspondència de la informació que s'ha mostrat per cada consulta sobre l'empresa, establiment i/o registre, guardar al XML tots els camps que s'han mencionat anteriorment. Es realitzen diverses transformacions per tal d'emplenar aquests camps. Resumidament, segons el valor dels booleans que provenen de la taula VISTAS_PROVISIONING per cada fila de la taula LOG_INFO es realitzen diverses cerques a les taules de Empresa, Establiment o Registre del model del DEE per tal d'obtenir els camps que es guarden al XML generat. També conté filtres per eliminar informació que no és rellevant, com si la empresa es troba de baixa, o la consulta ha estat a un establiment sense una empresa associada, ja que no té sentit enviar aquesta informació perquè no hi ha empresari que vulgui consultar-la. Es pot veure el procés a l'apèndix *Mapatge enviament dades* [A8].

Un cop realitzat el procés de la creació del XML, queda el flux de treball que executi aquest procés i envii aquest XML generat al portal de l'empresari.

Primer s'utilitza una tasca de tipus Assignment que calcula la data del dia actual. Aquesta data s'utilitzarà per enviar el fitxer XML amb la data del enviament en el nom, per seguir un control dels fitxers generats. Finalment es crea la

tasca que executa el mapping realitzat anteriorment i es configura de manera que s'envia a les màquines de la interfície del portal de l'empresari via SFTP. També s'ha configurat per tal que la seva execució quedi programada un cop per setmana.

Per tant, un cop a la setmana serà enviat un XML amb les consultes que s'han realitzat durant tota la setmana dins del model del DEE.

El XML resultant, es mostrarà a l'apartat de resultats ja que és un dels objectius del treball que s'havia d'assolir.

7 PRESENTACIÓ I DISCUSSIÓ DELS RESULTATS

Com a resultats d'aquest treball son els que mostren l'assoliment dels objectius d'aquest. Per tant, serà el gràfic del report fet amb PowerBI i el XML generat que s'envia al portal de l'empresari.

Començant amb el report es pot veure a l'apèndix *Report PowerBI* [A9] on es mostren els gràfics amb els indicadors que es van definir. Es poden observar 3 targes amb el nombre total d'empreses consultades, el nombre total d'usuaris i el total de consultes que s'han fet al model. Seguidament, es mostra el gràfic on es troba la relació entre el nom de l'empresa consultada amb la quantitat de consultes fetes a aquesta empresa per cada usuari. Juntament amb el mapa que mostra el nombre de consultes per cada regió d'Espanya o Catalunya. Per finalitzar, es troba la taula que mostra les consultes més recents per data i un gràfic amb l'evolució del nombre de consultes per data també. L'objectiu d'aquest report és que el personal tècnic de l'administració pública client pugui fer estudis sobre les empreses que han estat consultades dins del DEE.

Com a següent, i últim, resultat obtenim l'XML amb les dades de la traçabilitat de les consultes. Es pot veure a l'apèndix *XML Resultant* [A10]. Aquí, s'observen com s'han emplenat els camps definits anteriorment. Cada <Auditoria> conté la informació sobre una consulta feta. Es pot observar també els diferents casos de les consultes que es poden fer: només a una empresa, a una empresa amb el seu establiment, empresa amb registre, etc. Això influeix en el camps que venen o no informats dins de cada consulta. Finalment, cal comentar que s'ha hagut de tancar el camp de DocumentEmpresa que es el NIF de l'empresa consultada per la protecció de dades. Encara que no es vegi molt, es pot veure que sí ve informat.

Com s'acaba d'exposar, els resultats d'aquest treball han sigut complets i han aconseguit complir amb els objectius establerts al començament.

Una opció ampliable a aquest treball seria la realització de la interfície dins del portal de l'usuari per tal de poder visualitzar aquestes dades que s'envien. També es podria arribar a realitzar una interfície dins del portal que permeti accedir al report que s'ha desenvolupat en PowerBI. Així s'aconseguiria que totes les dades es trobin en un mateix lloc, al portal de l'empresari, facilitant així la interacció del usuari amb les dades.

8 CONCLUSIÓ

Aquest treball ha estat basat en un problema real sobre un projecte que una administració pública client ha sol·licitat, per tant, el seu ús serà aplicat en el món real en el moment en que l'administració desenvolupi l'interfície per consumir les dades enviades en l'XML. Encara no es sap quan serà la data en la que es desenvoluparà aquesta interfície, però això ja no depèn de l'empresa ni dels objectius aconseguits d'aquest treball.

Com a aspecte a millorar seria que alhora de definir els camps que al final s'envien al portal no ha hagut cap especificació per part de l'administració pública client, és a dir, s'han definits aquells que han semblat més rellevants. Per tant, és possible que en un futur s'hagin de redefinir aquest camps per complir amb el que vol el client. També la part de la planificació d'aquest treball ja que s'ha hagut de realitzar canvis i reajustaments constants sobre les tasques a realitzar.

Com a reflexió personal aquest treball m'ha ajudat a aprendre a gestionar un projecte des de zero i com les prediccions sobre el temps de realització d'una tasca mai son encertades del tot. També a analitzar i entendre millor quines son exactament les necessitats del client, ja que per aconseguir un bon treball es requereix entendre molt bé el que el client sol·licita. A més, he après sobre la llei 18/2020 i com ha afectat a les administracions públiques. Tanmateix, sobre lo important que és la protecció de dades i quines dades poden fer-se públiques i quines no.

Per finalitzar, part d'aquest projecte té una utilització futura per qualsevol client que vulgui aconseguir una traçabilitat sobre les consultes que s'han fet a unes determinades dades. El script desenvolupat amb Java a la primera part d'aquest treball permet la seva utilització en altres eines i entorns que no sigui amb l'eina Informatica, facilitant gran part dels objectius d'aconseguir aquesta traçabilitat.

AGRAÏMENTS

En primer lloc voldria agrair a la empresa NTT DATA per confiar en mi i donar-me la oportunitat de realitzar un projecte real que és funcional per l'empresa. En concret al meu company de projecte, Miguel, que m'ha ajudat i guiat durant tot el desenvolupament d'aquest TFG.

També a la meua família i parella per sempre confiar en mi i en el meu treball.

Finalment al tutor de l'escola David Acuña per tutoritzar-me tot el procés de desenvolupament i ajudar-me a millorar aquest treball.

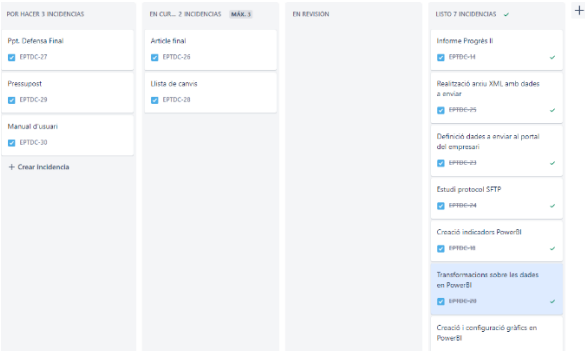
BIBLIOGRAFIA

- [1] Atlassian. (s. f.). Jira | Software de seguimiento de proyectos e incidencias. <https://www.atlassian.com/es/software/jira>
- [2] BOE-A-2021-1662 Ley 18/2020, de 28 de diciembre, de facilitación de la actividad económica. (s.f.). BOE.es - Agencia Estatal Boletín Oficial del Estado. <https://www.boe.es/buscar/doc.php?id=BOE-A-2021-1662>
- [3] Tablado, F. (2023). RGPD (Reglamento general de protección de datos): Guía 2023. Grupo Atico34. <https://protecciondatos-lopd.com/empresas/rgpd-reglamento-generalproteccion-datos/>

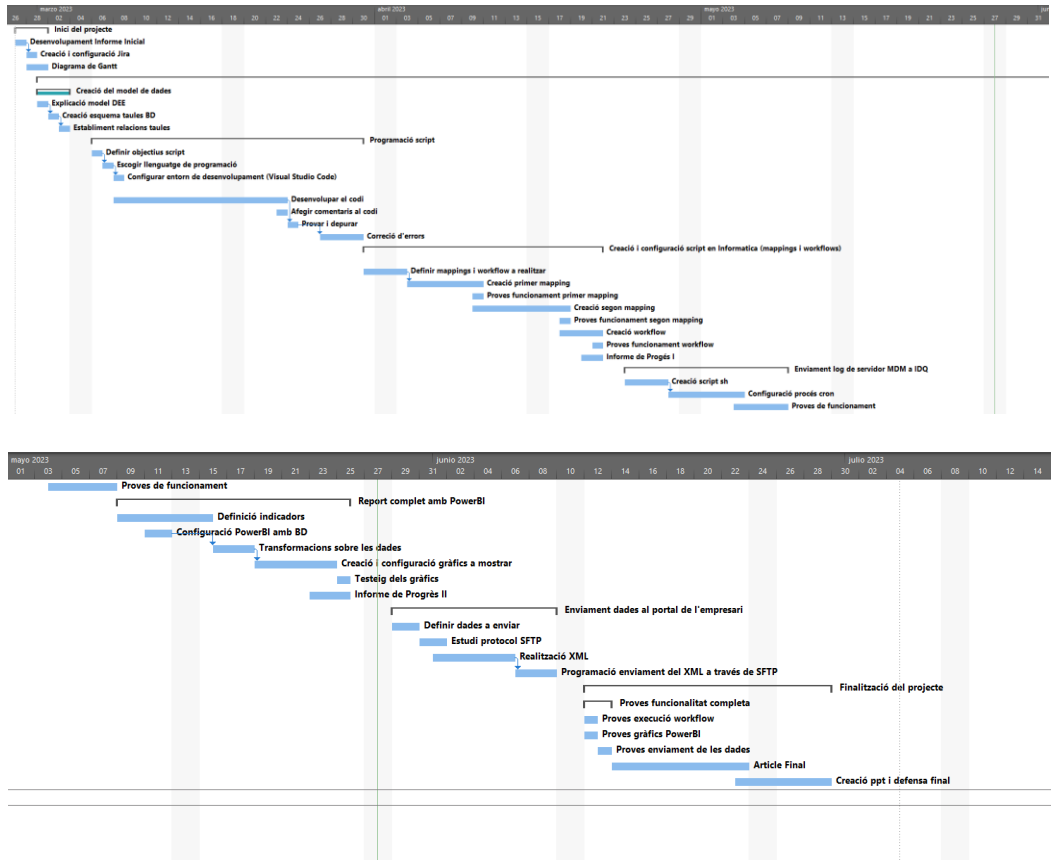
- [4] Informatica. (s. f.). Intelligent Data Management Cloud | Informatica España. <https://www.informatica.com/es/platform.html>
- [5] General Overview of Virtual DataPort — VQL Guide. (s. f.). https://community.denodo.com/docs/html/browse/6.0/vdp/vql/general_overview_of_virtual_dataport/general_overview_of_virtual_data-port
- [6] Getting Started with Java in Visual Studio Code. (2021, 3 novembre). <https://code.visualstudio.com/docs/java/java-tutorial>
- [7] Manual Java. (2023, 19 abril). Manual Web. <https://www.manualweb.net/java/>
- [8] Dib, F. (s. f.). regex101: build, test, and debug regex. <https://regex101.com/>
- [9] B, G., & B, G. (2023). ¿Qué es un comando cURL y cómo usarlo? Tutoriales Hostinger. <https://www.hostinger.es/tutoriales/comando-curl>
- [10] De Luz, S. (2023, 15 mayo). Utiliza Cron y Crontab para programar tareas en tu servidor. RedesZone. <https://www.redes-zone.net/tutoriales/servidores/cron-crontablinux-programar-tareas/>
- [11] XML Tutorial. (s. f.). <https://www.w3schools.com/xml/>
- [12] Transformation types. (s. f.). <https://docs.informatica.com/integration-cloud/dataintegration/current-version/transformations/transformations/transformationtypes.html>
- [13] Working with Mappings. (s. f.). <https://docs.informatica.com/dataintegration/powercenter/10-5-3/powerexchange-interfaces-for-powercenter/part-2--powerexchange-client-for-powercenter--pwxpc-/working-with-mappings.html>

APÈNDIX

A1. KANBAN



A2. DIAGRAMA DE GANTT



A3. LÍNIES ARXIU LOG

```

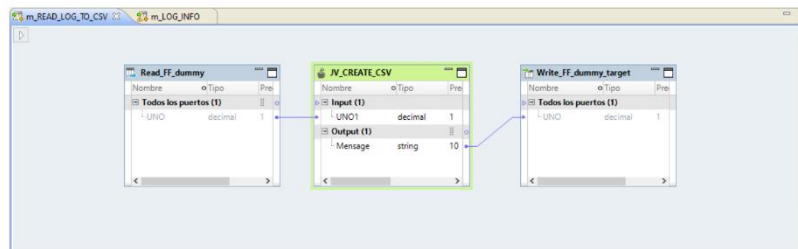
10.51.228.23 - - [19/Apr/2023:09:36:04 +0200] "POST /cleanse/ HTTP/1.1" 200 202
10.51.225.200 - admin [19/Apr/2023:09:36:03 +0200] "GET /cmx/cs/emo_dee_web-CMX_ORIS/VISTA_EMPRESA_SOLO/10005?compact=y&readFileMetadata=true&recordStates=ACTIVE&#37;2CPENDING&readSystemFields=
10.51.225.200 - - [19/Apr/2023:09:36:04 +0200] "GET /e360/mdm/ui/res/mdmstyle/images/ajax-loader-16x16.gif HTTP/1.1" 200 1002
10.51.225.200 - DEE_OGE [19/Apr/2023:09:36:39 +0200] "GET /cmx/cs/emo_dee_web-CMX_ORIS/VISTA_BE_ESTABLIMENT/100021?compact=y&readFileMetadata=true&recordStates=ACTIVE&#37;2CPENDING&readSystemFi
10.51.225.200 - - [19/Apr/2023:09:36:39 +0200] "GET /e360/mdm/ui/res/mdmstyle/images/ajax-loader-16x16.gif HTTP/1.1" 200 1002

```

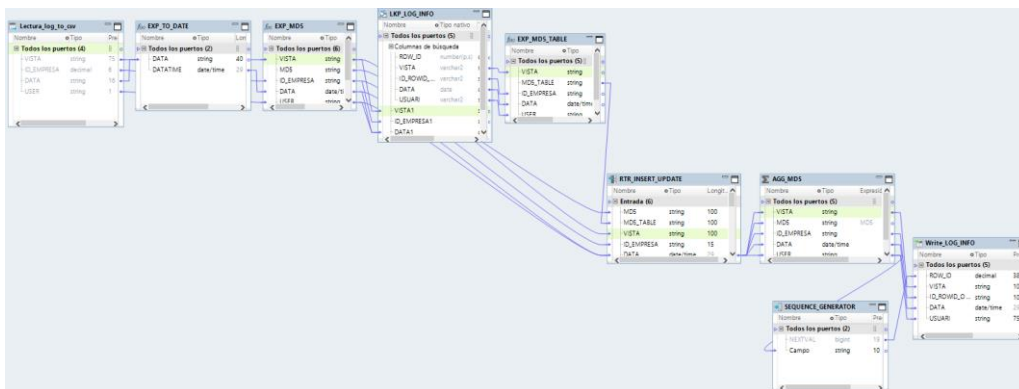
A4. CSV GENERAT PER L'SCRIPT

VISTA	ID_ROWID_OBJEC	DATA	USER
VISTA_REGISTRE	10	06/04/2023 11:49	DEE_ATIC
VISTA_REGISTRE	10	06/04/2023 11:49	admin
VISTA_REGISTRE	10	06/04/2023 11:50	admin
VISTA_EMPRESA_REG_ACT	10000	06/04/2023 12:01	admin
VISTA_EMPRESA_REG_ACT	10000	06/04/2023 12:02	admin
VISTA_REL_REG_EMP	240021	06/04/2023 12:02	DEE_OGE
VISTA_REL_REG_EMP	240021	06/04/2023 12:02	admin
VISTA_REL_REG_EMP	240021	06/04/2023 12:02	admin
VISTA_REL_REG_EMP	240021	06/04/2023 12:02	admin
VISTA_EMPRESA_SOLO	10	06/04/2023 12:38	DEE_ATIC
VISTA_EMPRESA_SOLO	10	06/04/2023 12:38	admin
VISTA_EMPRESA_SOLO	1	06/04/2023 12:39	admin
VISTA_EMPRESA_SOLO	1	06/04/2023 12:39	admin
VISTA_EMPRESA_SOLO	10	06/04/2023 12:39	admin
VISTA_EMPRESA_SOLO	10	06/04/2023 12:39	admin
VISTA_BE_ESTABLIMENT	100005	06/04/2023 12:40	admin
VISTA_BE_ESTABLIMENT	100005	06/04/2023 12:40	DEE_ATIC
VISTA_EMPRESA_REG_ACT	10000	06/04/2023 12:41	admin
VISTA_EMPRESA_REG_ACT	10000	06/04/2023 12:41	admin

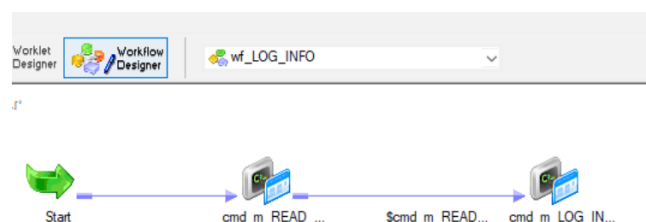
A5. MAPATGE EXECUCIÓ SCRIPT JAVA



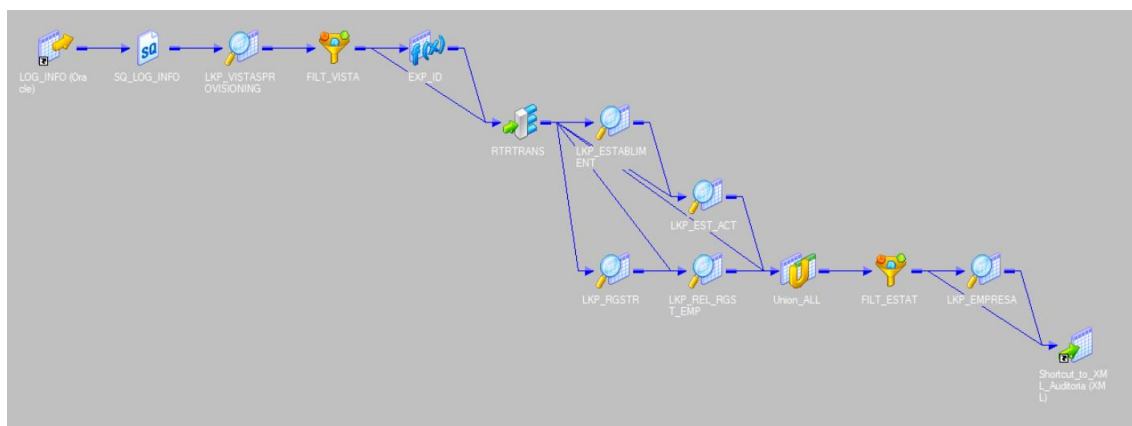
A6. MAPATGE GUARDAT DADES



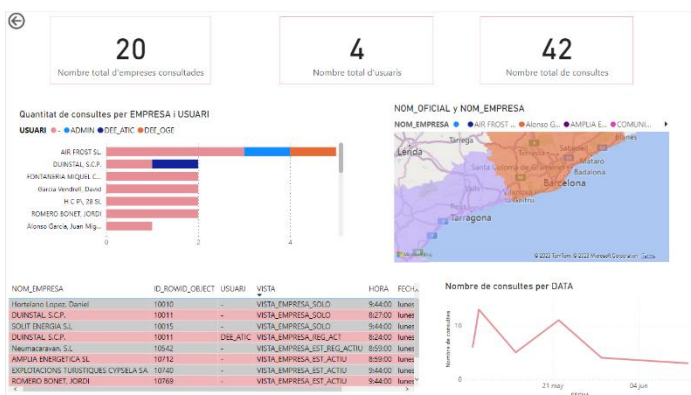
A7. FLUX DE TREBALL GUARDAT DADES



A8. MAPATGE ENVIAMENT DADES



A9. REPORT POWERBI



A10. XML RESULTANT

```

<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<CarregaRegistres>
  <Auditories>
    <Auditoria>
      <Data>2023-05-22</Data>
      <User>-</User>
      <Vista>VISTA_BE_ESTABLIMENT</Vista>
      <DocumentEmpresa>E...</DocumentEmpresa>
      <IdUnicEstabliment>013691888BY</IdUnicEstabliment>
      <Empresa>0</Empresa>
      <Establiment>1</Establiment>
      <Registre>0</Registre>
      <Contacte>0</Contacte>
    </Auditoria>
    <Auditoria>
      <Data>2023-05-22</Data>
      <User>-</User>
      <Vista>VISTA_REL_REG_EMP</Vista>
      <DocumentEmpresa>F...</DocumentEmpresa>
      <CatalogRegistre>REE</CatalogRegistre>
      <NumeroRegistre>CIF-14-1004493-Q</NumeroRegistre>
      <Empresa>0</Empresa>
      <Establiment>0</Establiment>
      <Registre>1</Registre>
      <Contacte>0</Contacte>
    </Auditoria>
    <Auditoria>
      <Data>2023-05-09</Data>
      <User>-</User>
      <Vista>VISTA_REGISTRE</Vista>
      <DocumentEmpresa>C...</DocumentEmpresa>
      <CatalogRegistre>APR</CatalogRegistre>
      <NumeroRegistre>REG_108</NumeroRegistre>
      <Empresa>0</Empresa>
      <Establiment>0</Establiment>
      <Registre>1</Registre>
      <Contacte>0</Contacte>
    </Auditoria>
  </Auditories>
</CarregaRegistres>
  
```