**PROIECT ETL**

**STRUCTURA ȘI PRELUCRAREA SURSELOR RAW**

IULIANA-RALUCA BUTNARIU

POLINA CIUMAC

RARES GABRIEL MARGARINT  
ALEXANDRU DUMITRU IONITA

LOCATIE: IASI

**Confluence – Fișierele de absențe (Absences)**

Una dintre sursele principale analizate în acest proiect este reprezentată de fișierele de absențe extrase din platforma Confluence. Datele din aceste fișiere descriu cererile de concediu, lipsa temporară a angajaților sau alte forme de absență. Fișierele au fost exportate în format .csv și importate în zona raw a proiectului.

La prima analiză a datasetului, am identificat o serie de nereguli care ar fi împiedicat o procesare corectă în pașii următori. Una dintre problemele cele mai evidente a fost o desincronizare orară: valorile din coloanele SHOUR și EHOUR prezentau un decalaj de +3 ore față de ora reală înregistrată în platformă. Am decis să corectez această diferență direct în scriptul de preprocesare, folosind TIME disponibila in Excel.

O altă zona care a necesitat modificari, a fost in cadrul câmpului NAME, unde, în mai multe cazuri, apăreau două sau mai multe persoane trecute într-un singur rând, separate prin virgulă sau alte delimitatoare. Acest lucru încalcă principii fundamentale de organizare a datelor, mai exact prima formă normală (1NF), care sugerează că un atribut trebuie să conțină o singură valoare atomică. Pentru rezolvarea acestei probleme, am aplicat o transformare prin care am separat numele multiple într-o listă, pe care ulterior am desfășurat-o (explode) în mai multe rânduri, astfel încât fiecare persoană să fie reprezentată într-o înregistrare proprie, asociată absenței respective.

Prin aceste transformări, fișierul de absențe a fost adus într-o formă normalizată,pregătită pentru a fi încărcată în zona de staging.

**SharePoint – Participare la sesiuni de training (DavaX Academy)**

O altă sursă de date inclusă în proiect a fost reprezentată de fișierele referitoare la participarea angajaților la sesiunile de training organizate de DavaX Academy România. Aceste fișiere au fost colectate dintr-un spațiu SharePoint intern, unde sunt centralizate înregistrările pentru fiecare sesiune în parte, în funcție de tematică și dată.

Scopul integrării acestei surse a fost de a completa imaginea generală a implicării angajaților în activități de formare continuă. Informațiile disponibile au fost preluate ca atare în zona raw, urmând a fi procesate în pașii următori pentru a extrage indicatori de participare și prezență per sesiune sau per angajat.

**Timesheet – Pontaj manual cu date reale**

În cadrul proiectului am integrat și o sursă de tip **timesheet**, care conține înregistrări reale de lucru pentru mai mulți angajați, organizate sub formă de fișier structurat. Acest fișier a fost completat **manual**, dar conține date preluate din activitatea zilnică reală a participanților.

Prin includerea acestui tabel în proiect, ne-am propus să simulam un caz apropiat de realitatea dintr-un mediu de lucru, în care datele nu sunt întotdeauna disponibile prin API sau exporturi automate, ci pot proveni și din surse introduse de utilizatori. Fișierul timesheet a fost folosit pentru a construi o componentă fact în cadrul modelului dimensional, având un rol important în generarea rapoartelor privind orele lucrate, tipul activităților desfășurate și distribuția temporală a muncii.

**Employees – Date construite manual**

Tabelul cu angajați a fost creat în mod **manual**, pornind de la informații reale, introduse individual. Fiecare înregistrare continue detalii precum numele angajatului, orașul în care lucrează și diverse legături cu entități din modelul de date (mentori, coordonatori, echipe de lucru).

Deși fișierul a fost generat de la zero, conținutul său reflectă structura reală. Această componentă a fost utila pentru a putea face legătura între prezența angajaților, activitățile înregistrate în pontaj și participarea la sesiunile de training.

Fișierul a fost construit manual, întrucât echipa este formată din participanți proveniți din mai multe orașe. Pentru a putea completa informațiile necesare în mod corect, s-a accesat fiecare grup de lucru în parte, astfel încât să se preia numele fiecărui angajat, orașul de proveniență și celelalte informații relevante pentru structurarea tabelei.