

DATA CLEANSING PROCESS

Unul din cele mai importante active ale unei companii este informația. Abilitatea unei companii de a concura, de a se adapta și de a se dezvolta într-un climat de afaceri cu schimbări rapide, depinde în mare parte de cât de bine utilizează informațiile în luarea deciziilor.

Ce reprezintă datele: un activ construit conform standardelor internaționale (vezi Fig. 1)

- În acest moment diferite articole MRO (Maintenance, Repair and Operations) precum și echipamentele pot fi achiziționate prin specificarea standardelor internaționale relevante în contractele cu furnizorii;
- Datele sunt asemenea oricărui alt activ, sunt acoperite de standardele internaționale, au un cost de achiziție, instalare și întreținere/gestionare.

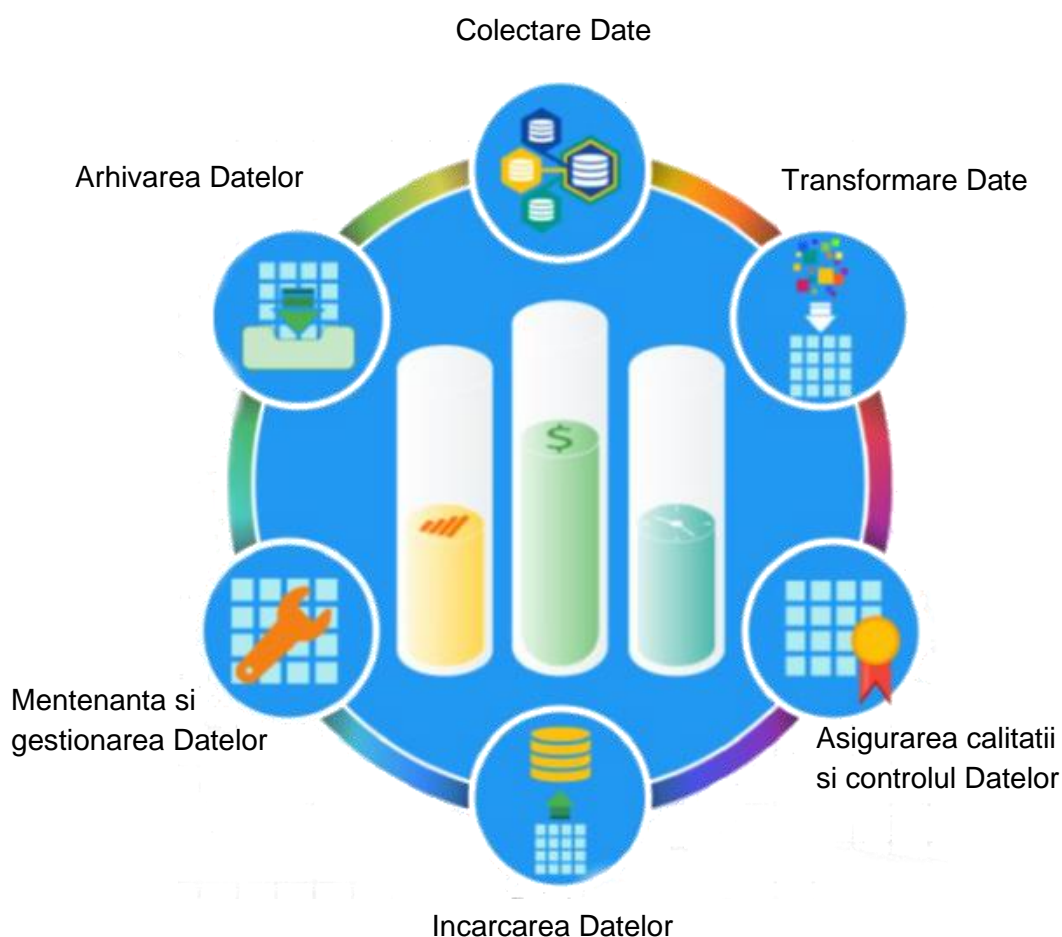


Fig. 1

Indiferent de modul în care sunt colectate datele, întotdeauna va exista un anumit nivel de eroare. Datele “nesigure” sau “murdare” se referă la date care sunt pline de inconsecvențe.

Deși unele discrepante sunt legitime deoarece reflectă variații în context, altele ar reflecta probabil o eroare de măsurare sau intrare. Acestea pot varia de la greșeli datorate erorii umane, sisteme de înregistrare prost concepute sau pur și simplu pentru că există un control incomplet asupra formatului și tipului de date

importate din surse externe de date. Asemenea discrepanțe provoacă probleme atunci când se încearcă efectuarea unor analize, rapoarte, achiziții bazate pe acestea. Înainte de prelucrarea datelor pentru analize, rapoarte, achiziții trebuie să se acorde o atenție sporită asigurării exactității, consistenței și coerenței datelor.

Curățarea datelor (data cleansing) este procesul de identificare și corectare (sau eliminarea) a înregistrărilor corupte sau inexacte dintr-un anumit set de înregistrări, tabele sau baze de date. Folosit în principal pentru baze de date, termenul face referire la identificarea părților de date incomplete, incorecte, inexacte, irelevante etc. urmata de înlocuirea, modificarea sau ștergerea acestor date "murdare". Procesul de curățarea datelor este ilustrat în Fig. 2.

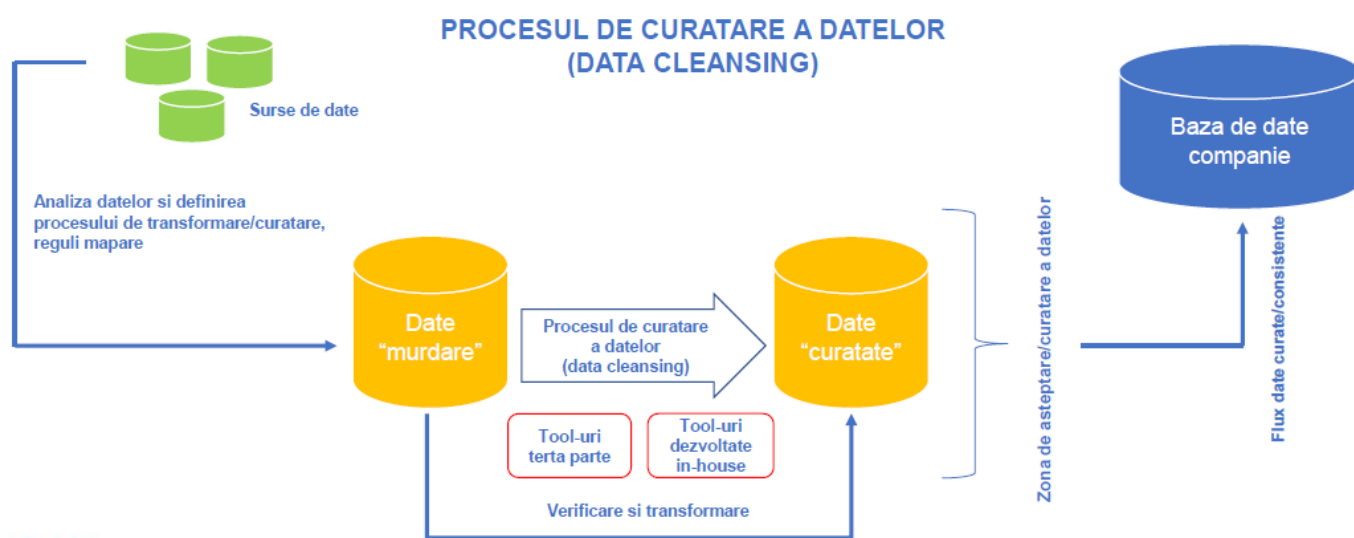


Fig. 2

CEI 4 PASI AI PROCESULUI DE CURATARE A DATELOR (DATA CLEANSING STEPS), sunt reprezentați în Fig. 3

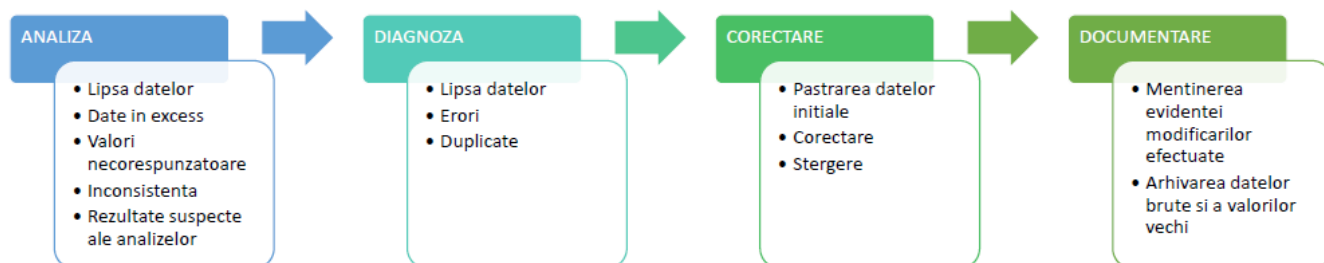


Fig. 3

ANALIZA - presupune cautarea in mod sistematic a elementelor suspecte in bazele de date sau in seturile de date supuse analizei.

DIAGNOZA - (identificarea naturii datelor defectuoase) și **CORECTAREA** (stergerea, editarea sau pastrarea datelor in starea lor initiala) sunt etape ale procesului de curatare a datelor ce necesita o intelegere aprofundata a tuturor tipurilor și surselor de erori posibile din timpul proceselor de colectare si de introducere a datelor.

DOCUMENTAREA - modificarilor presupune inregistrarea tuturor erorilor identificate, a alterarii datelor, completarii si a verificarii erorilor, oferind posibilitatea revenirii la valorile initiale daca este cazul.

Principalele probleme legate de inventarierea materialelor/itemilor intalnite in bazele de date “murdare” sunt urmatoarele:

- Lipsa unui model de specificatie afferent materialelor/itemilor;
- Lipsa impartirii materialelor/itemilor pe categorii – Inregistrari prea multe si irelevante;
- Inexistenta conceptului de catalog de materiale/dictionar itemi;
- Surse separate de creare a inregistrarilor - lipsa unui proces de standardizare sau a unor tool-uri;
- Descrieri de materiale/itemi nestandardizate sau incomplete;
- Metode de ambalare si specificatii diferite in functie de destinatia materialelor/itemilor;
- Alocarea de noi coduri de materile/ietmi pentru mici diferente;
- Materiale neidentificabile.

EFECTELE/COSTURILE REZULTATE DIN SPECIFICATIILE SI DATELE INCONSISTENTE, sunt reprezentate schematic in Fig. 4

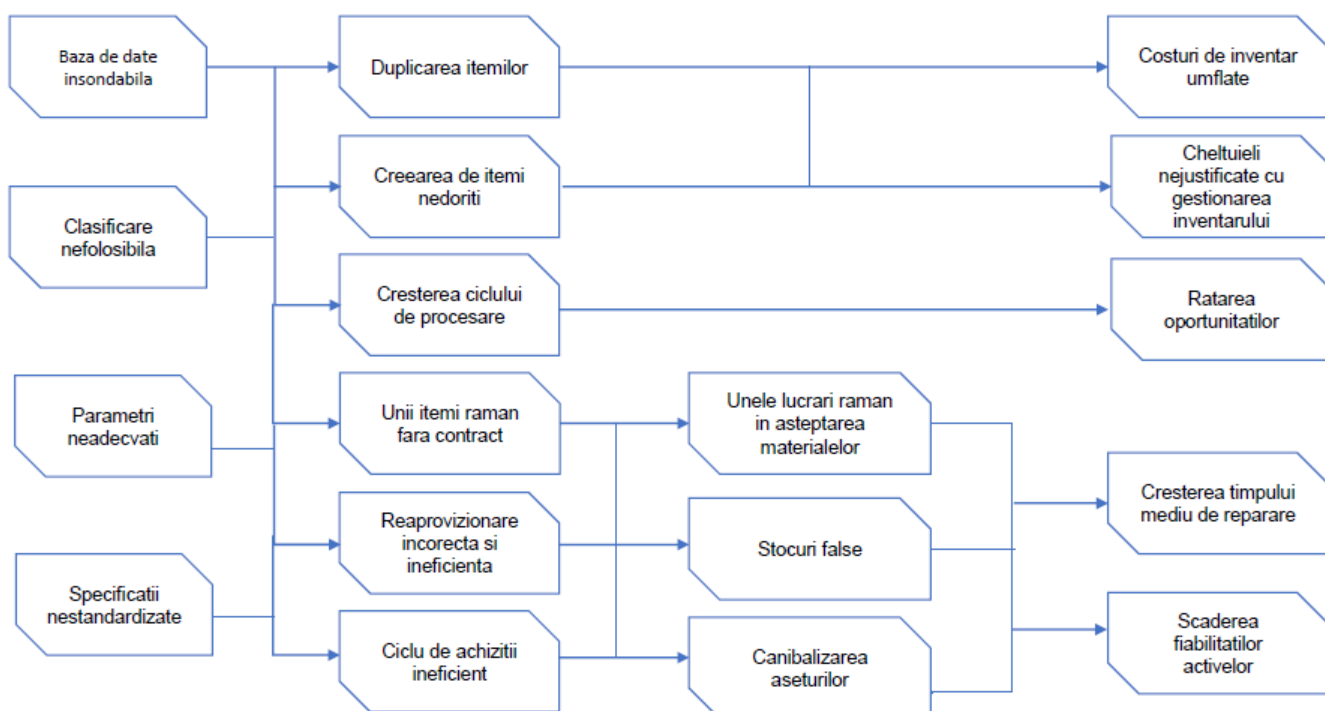


Fig. 4

CLASIFICAREA SI IDENTIFICAREA

Clasificare – gruparea itemilor comuni in vederea analizarii acestora (de ex. Clasificare dupa: codul UNSPSC (United Nations Standard Products and Services Code), sisteme internationale armonizate de codificare si descriere a marfurilor, etc.). O prezentare schematica a nivelului de detaliere este in Fig. 5.



Fig. 5

Identificare – descrierea unui singur item. (de ex. folosind un catalog de descrieri conform ISO 8000) Procesul de identificare este prezentat in Fig. 6.

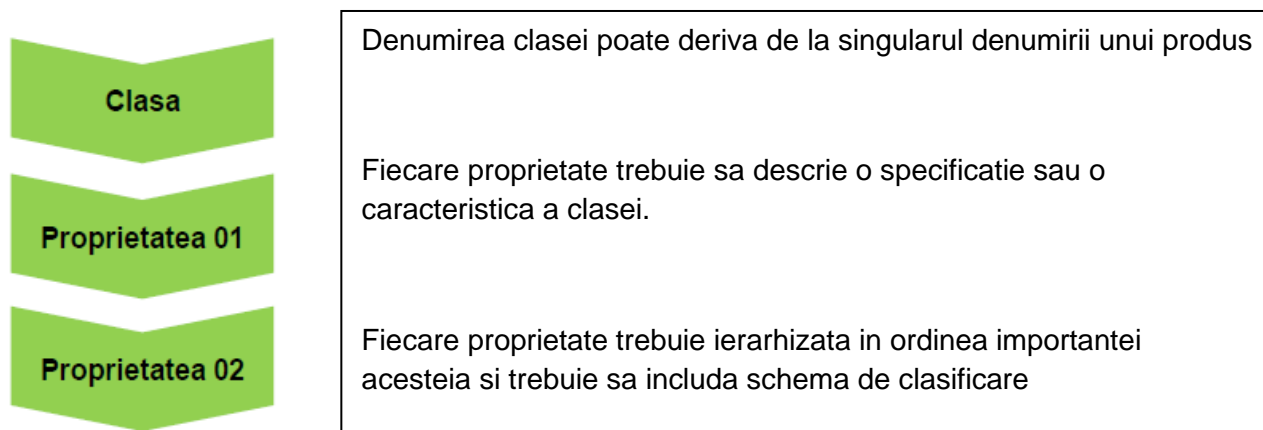


Fig. 6

Apare cu siguranta intrebarea:

CE ESTE DE FACUT? / DIRECTII DE URMAT

In primul rand trebuie definite o serie de principii de clasificare si o metodologie a acestui proces.

Principiile sunt prezentate in Fig. 6.

Se vor seta principiile de baza, se va defini un dictionar/catalog de materiale in baza caruia se vor reface descrierile itemurilor/materialelor.

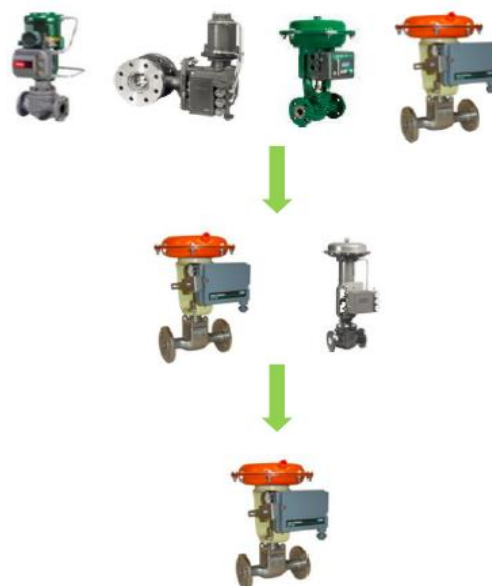
Clasificarea materialelor

Aranjarea sistematica a materialelor in grupe sau categorii intr-o ierarhie conform unui anumit criteriu. Este o clasificare (sau) separare a materialelor in clase sau nume si a atributelor acestora intr-o ierarhie.

Noun (Clasa) Denumirea de baza a unui tip de produs

Modifier (Sub-Clasa) Atribut sau calificativ pentru substantiv (Noun)

Template (Model) O scurta fisa tehnica ce continue attribute cheie care diferentiaza itemii din cadrul aceleiasi familii de itemi



Metodologie

Exemplu

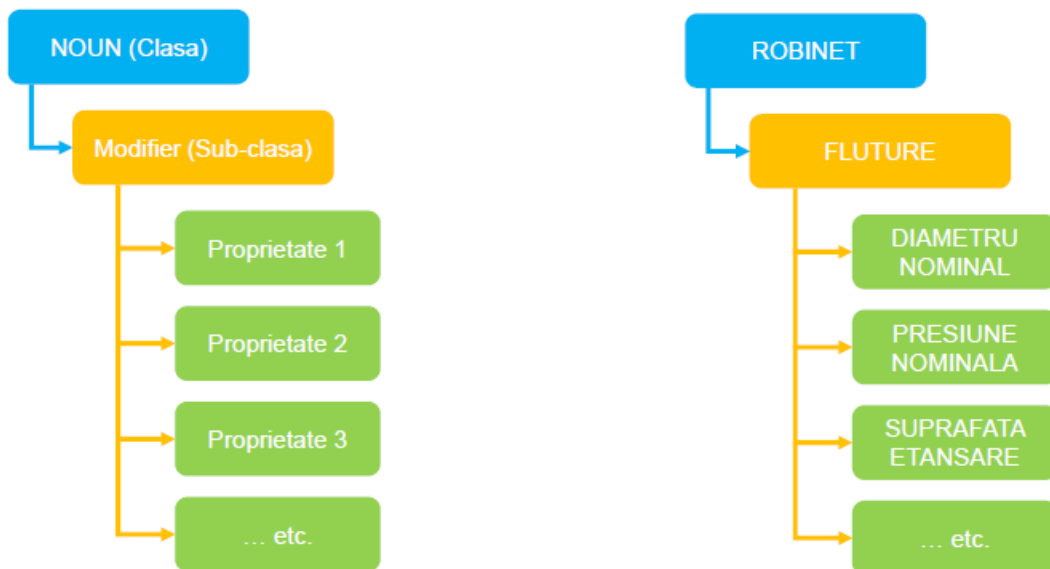


Fig. 6

Deasemenea trebuie definit un sablon pentru clasificarea datelor (ilustrat in Fig. 7).

TEMPLATE CLASIFICARE

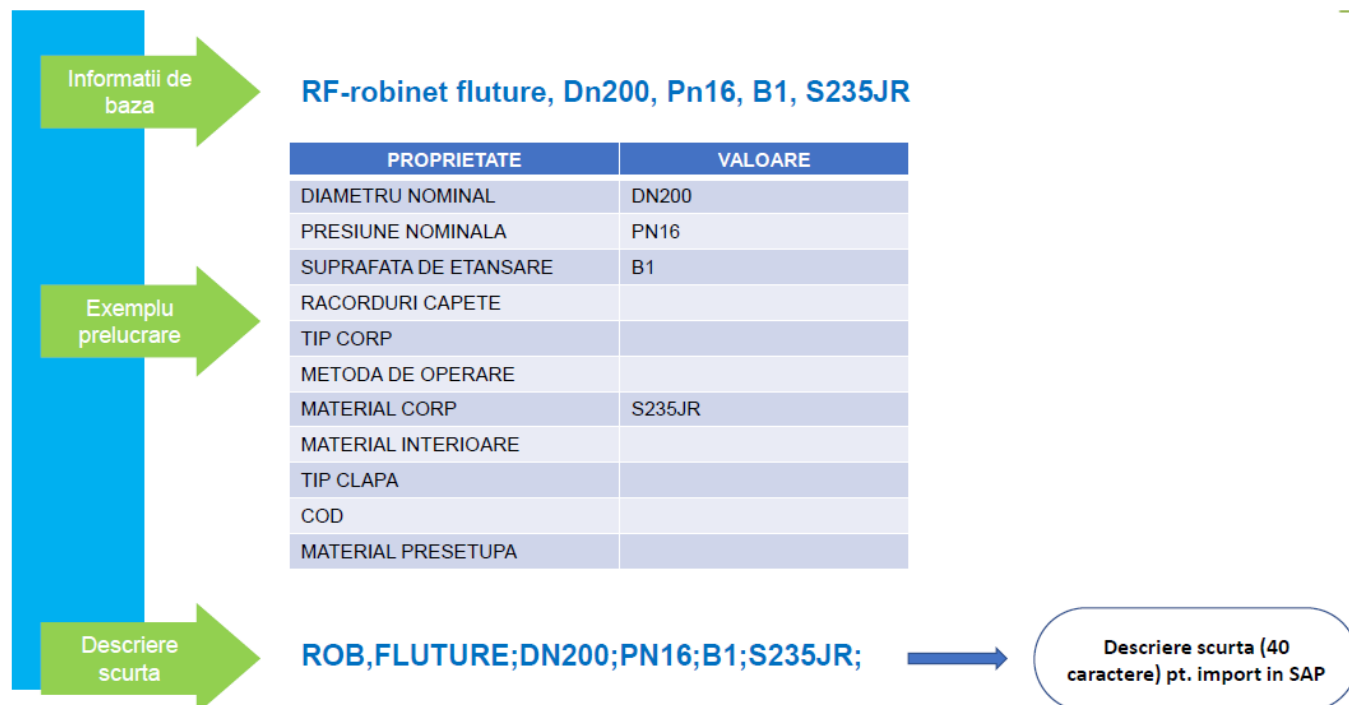


Fig. 7

Ce urmarim prin procesul de curatare a datelor? Un rezumat al scopului este prezentat in Fig. 8.

SCOPUL PROCESULUI DE CURATARE A DATELOR (DATA CLEANSING)

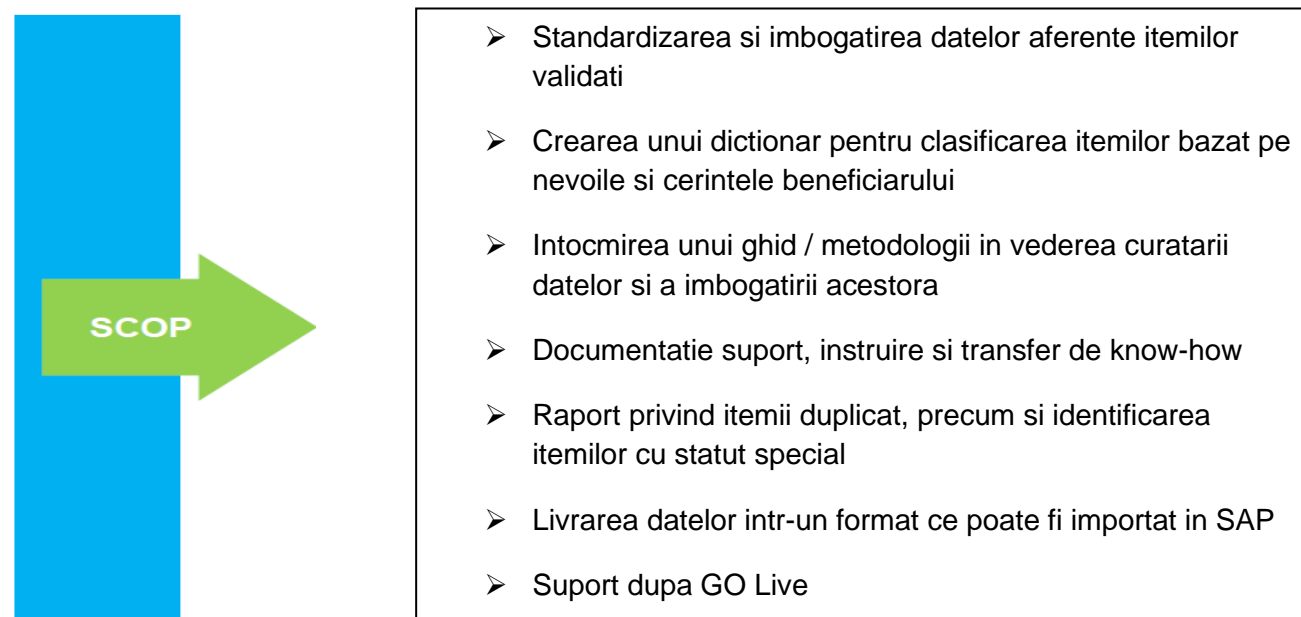


Fig. 8

Cum ar arata o metodologie a procesului de curatare date? Este prezentata in Fig. 9

METODOLOGIA PROCESULUI DE CURATARE A DATELOR (DATA CLEANSING)



Fig. 9

Si desigur end user-ul ar dori sa vada intr-un singur tablou, intreg procesul de curatare date. Este prezentat ilustrativ in Fig. 10.

PROCESUL DE CURATARE A DATELOR (DATA CLEANSING) PAS CU PAS

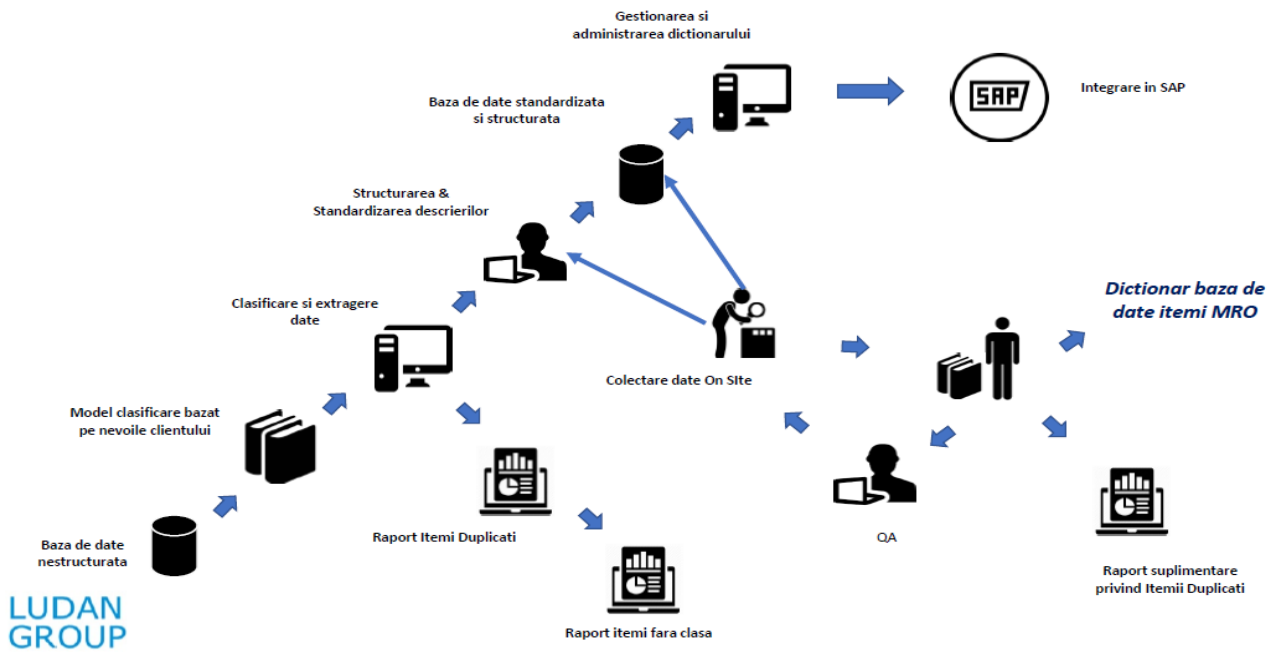


Fig. 10

Bibliografie: www.mroinsyte.com