# Plataforma Transversal de Dades

Qualitat de la dada

Novembre 2024







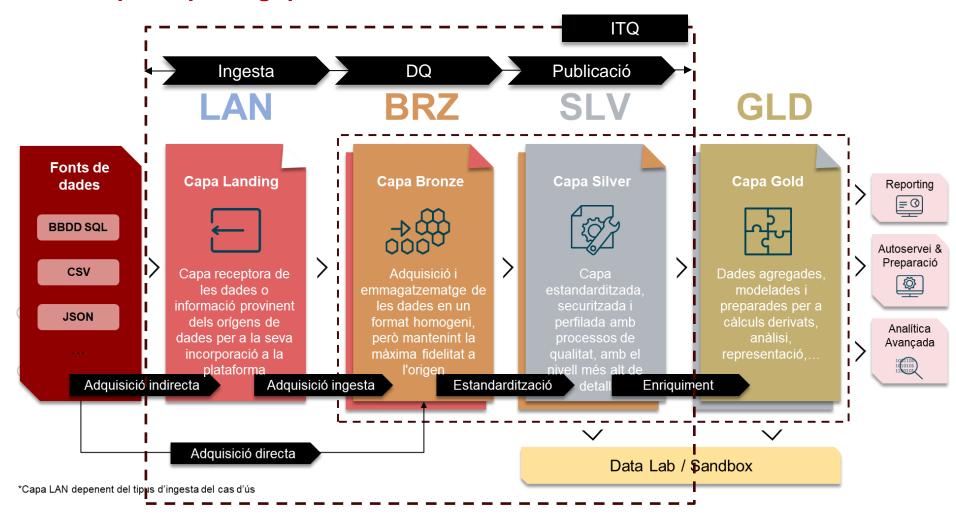






Capes lògiques i estàndards tècnics

Distribució de la PTD per capes lògiques



#### **Nomenclatura**

Per tal de poder utilitzar les dades per desenvolupar casos d'ús, és necessari que aquestes estiguin **estandarditzades** i compleixin una sèrie de requisits de **nomenclatura (noms de taules i camps**):



Noms auto-explicatius



Noms de camps escrits en minúscules



Notacions clares i concises



Utilitzar "\_" com a separador a canvi de l'espai



Tot escrit en català



Utilitzar els prefixos definits per identificar la naturalesa de taules i camps



Noms de taules i camps escrits en singular



Evitar l'ús d'abreviatures comunes a diversos àmbits

Complir amb tots aquests requisits ens ajudarà a obtenir noms representatius, funcionals i estandarditzats que ens facilitaran la comprensió dels nostres actius de dades.

#### **Nomenclatura**

El nom de les **taules** ha d'anar precedit per un dels següents prefixos definits que determinen la tipologia de la taula:

- "DIM": Taula dimensional que serveix per a classificar o categoritzar un fet.
- "FAC": Taula fact o de fets que conté els valors de les mesures o indicadors de negoci.
- "AGG": Taula agregada d'una taula de fets respecte alguna de les seves dimensions d'anàlisi.
- "REL": Taula relacional que s'utilitza per establir el vincle entre dos dimensions.
- "LKP": Taula lookup o diccionari que s'utilitza per a la traducció o descodificació de valors.
- "SUK": Taula per identificar claus subrogades.

- "AUX": Taula auxiliar, per necessitat d'un pas intermedi dins d'un procés tècnic.
- "REF": Taula que conté dades de referència.
- "MES": Taules mestres.
- "OUT": Taules de sortida que contenen els resultats finals d'un procés d'anàlisi de dades o d'aplicació d'un model d'aprenentatge automàtic.
- "FTR": Taules que contenen les característiques (features) que s'utilitzaran com a variables independents en un model d'aprenentatge automàtic.

#### **Nomenclatura**

El nom dels **camps** ha d'anar precedit per un dels següents prefixos definits que determinen la seva tipologia:

- "IDE": Identificador. Es recomana utilitzar identificadors del tipus "integer" excepte en casos on pugui ser necessari tipus "varchar".
- "DAT": Data i hora. Per defecte tindrà una precisió nivell dia i hora.
- "HRS": Data i hora. Per defecte tindrà una precisió a nivell de dia i hora.
- "TMS": "Timestamp" o Marca de temps. Utilitzat per indicar moments precisos amb precisió fins al segon, per exemple, per indicar quan s'ha executat una càrrega.
- "DES": Descriptor.

- "FLG": "Flag" o Booleà (boolean).
- "IMP": Import. Els imports podran tenir associats el sufix \_{TIPO MONEDA} on s'especifica si el valor de la moneda es local ("LC") o l'identificador propi de la moneda d'extensió fixe 3 caràcters (ISO 4217).
- "NUM": Indicador numèric sempre i quant no sigui un import o percentatge.
- "PER": Indicador numèric per percentatges.
- "URL": Utilitzat per anomenar camps que contenen una direcció web (URL).

### Eina de govern de les dades Cloud de la Generalitat (IDMC)

#### Disponibilitat i entorns

Entorn únic de prova que passarà a ser l'entorn productiu. **No esta disponible** pels usuaris.

#### Ús actual

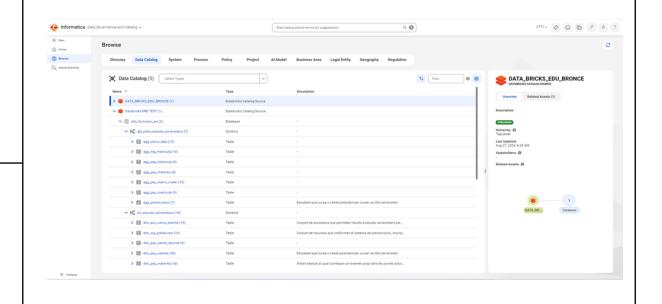
El glossaris de termes de negoci s'introdueixen a l'eina mitjançant plantilles (de la mateixa manera que amb l'eina On-Premise), mentre que els actius tècnics s'introdueixen a partir d'establir una connexió amb la BBDD i dur a terme escanejos per tal d'extreure les metadades.

#### **Contingut**

- Dominis i subdominis (Glossari)
- Cas d'ús de REU, sobre el que s'estan fent totes les proves:
  - Taules i camps dels esquemes de BRZ, SLV i GLD
  - Glossari de termes
  - Perfilats de les taules (Distinct, Non-Distinct i Null)

#### Proves en curs

- Rols, permisos i responsabilitats
- Càrrega d'actius tècnics mitjançant escanejos a la BBDD
- Càrrega d'actius tècnics mitjançant plantilles excel
- Perfilats i qualitat de la dada
- Representació del llinatge tècnic

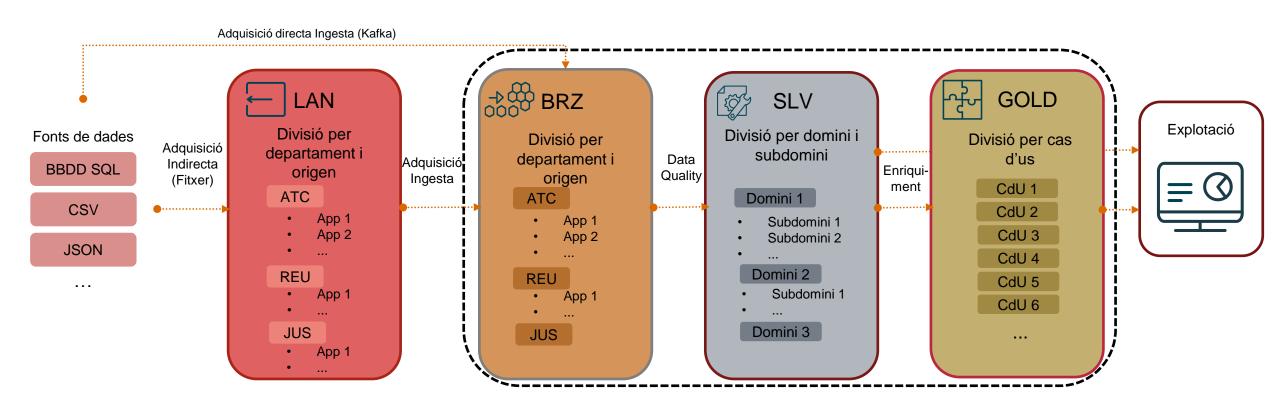


Aspectes de seguretat

## Seguretat

### Distribució funcional de la PTD per capes lògiques

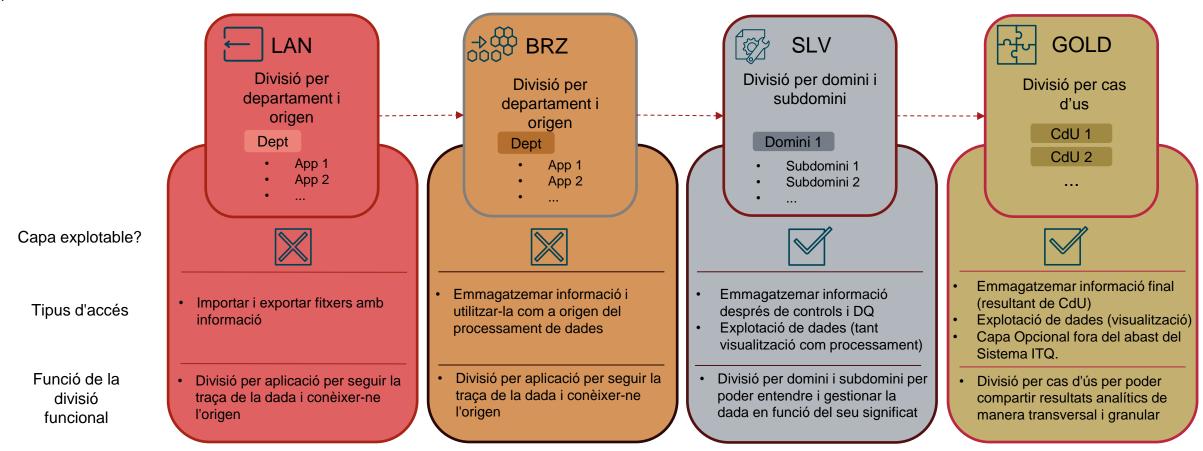
Cada capa de la plataforma tindrà una separació funcional depenent de la seva finalitat i l'estat de la dada. Landing i Bronze es distribuiran per origen de dades, Silver per subdomini funcional i Gold per Cas d'ús:



## Seguretat

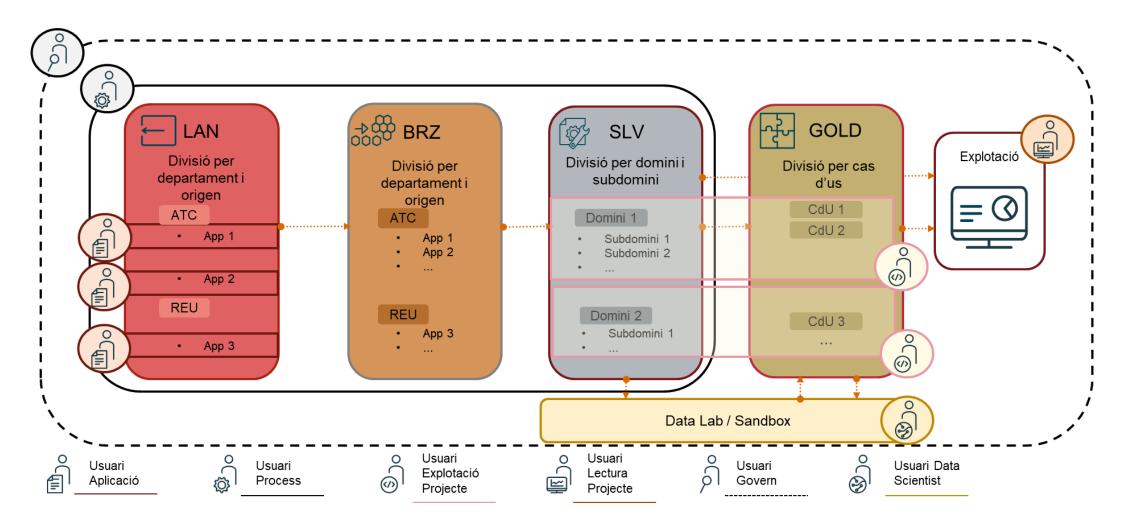
### Distribució funcional de la PTD per capes lògiques

Cada capa tindrà una política d'accés depenent de la finalitat amb què ha estat definida i les subdivisions funcionals dins d'elles permetran la gestió granular de permisos:



## Seguretat

### Esquema de permisos d'usuaris al llarg de la plataforma



Aspectes de privacitat

### **Privacitat**

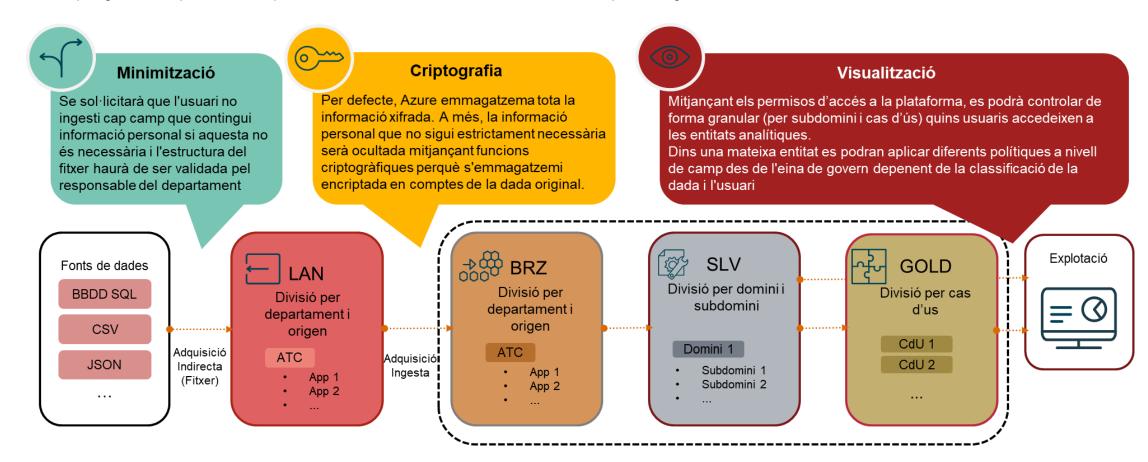
### Els aspectes de privacitat es basen en 3 grans pilars:



### **Privacitat**

### Protecció a l'accés a les dades sensibles de tota la plataforma

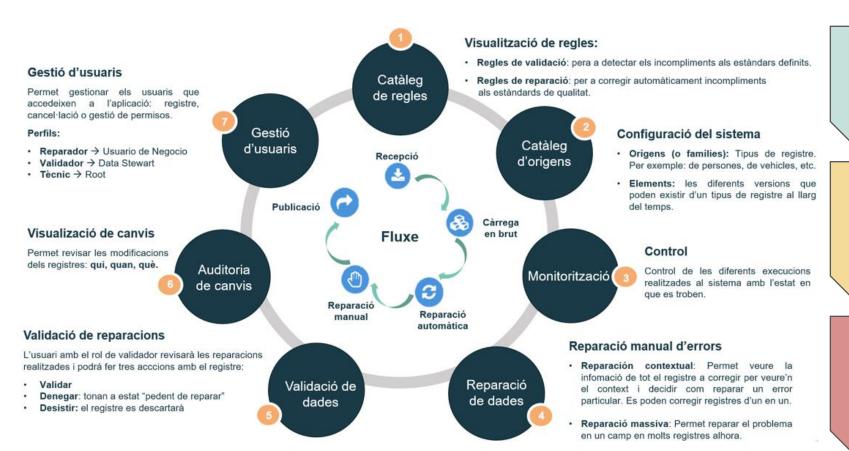
Cada mesura per garantir la privadesa s'aplicarà en una fase diferent del cicle de la dada per assegurar-ne l'ús correcte en tot moment:



Mòdul de qualitat

## Mòdul de qualitat

### Solució configurable mitjançant l'aplicació de regles de validació, remeiació i enriquiment





#### Regles de Validació

Criteris aplicats a les dades per assegurar que compleixen uns estàndards abans de ser processades.

#### Regles de remeiació



Processos dissenyats per corregir problemes en les dades identificades durant les validacions.

#### **Regles d'enriquiment**

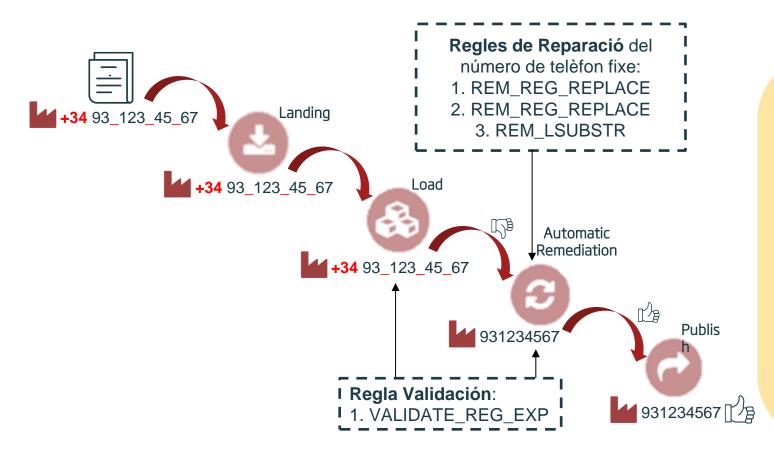


Processos utilitzats per a millorar les dades, afegint informació addicional o context per a fer-les més útils i completes

## Mòdul de qualitat

#### Exemple de funcionament a alt nivell del mòdul de Qualitat del sistema ITQ

Validem el format d'un camp telèfon que ha de complir aquest format 9XYNNNNNN / 8XYNNNNNN on X és sempre diferent de zero:



L'algorisme de la **Regla de Validació** verifica què un número de telèfon fixe es vàlid si i només si:

- 1. VALIDATE\_REG\_EXP:
  - Té longitud de 9 dígits.
  - Està format per caràcters només numèrics.
  - Comença per 9 o 8, el següent dígit no és 0 i a continuació venen 6 dígits més.

Quan un camp no passa una regla de validació l'algorisme executa una o més **Regles de Reparació** automàtica. En aquest cas:

- 1. REM\_REG\_REPLACE: Eliminem qualsevol caràcter que no sigui numèric del número de telèfon fixe.
- 2. REM\_REG\_REPLACE: Eliminem el prefix 34 del número de telèfon fixe si escau.
- 3. REM\_LSUBSTR: Ens quedem amb els nou primers dígits com a número de telèfon fixe vàlid.

Estudi d'un fitxer d'exemple

## Exemple

### Data Quality sobre les dades d'Entitats Esportives del Consell Català de l'esport

TOTAL DE	REGISTRES PROCESSATS:	20.728					
-				Malida 1	P> PF-/		
	Percentatge de camps que	arribenvuits on uli:		Validacio	ns Errò neas PRE/P	OST Remediació:	
	Camp	Percentatge vults onuls			Camp	M Davis ValidadA T	De scripdó Regia Validadó
	Va contacto	91,4994%			cif_consolidat		Validem el NIF/CF de l'entitat.
	desc_via_contacto	\$1,4994%			telefon consolidat		Validem el format del teléfon fixe (SXYNNNN NN / SXYNN NNNN) on Xés semone diferent de zero.
	adeca_contacto	90.1582%			mobil_consolidat		Validem el format del teléfon mób I ( Ø NNNNN NN, 71NNN NNNN, 72NN NNNNN , 73N NNNNN N, 74NNNNN NN).
	and e_postal_contacto	SE MINOR.			e mail_consolidat		Validem la correcte sa del e mail informat.
		87.0513%					Validem que el codi postal registral sigui un codi vàlid.
	municipi_contacto	97.0369%					Validem que el codi de la provincia registral e stigui informat.
	person a _con tacte _co mo lid at				cif_consolidat		Validem que el cod de la provincia registra e enga informac. Validem que al el Número de Registre no comença per Allavon; el CIF de l'entitat ha de començarper O, G.o.A.
	ed #	72.501%					Validem que el anys de la junta estiguin entre 2 i 10 anys
	inici mandat	65276%					Validem que el nom de l'entitat vingui tot en majús que s.
	mobil_con so lidat	57.1449%			code_postal_contacto		Validem que el codi postal de contact e sigui un codi válid.
		50,8346%					Validem que en el nom de l'entitat les paraules estiguin se paradas per un únic espai en blanc
		45.4409%			- Internation		
		42.7924%					
	email_consolidat	42.3582%					
	telefon consolidat	41.7551%					
	va rodal	50.851%					
		20.851%	Peo	ristres corre	ctes vs. Incorrectes	s després de OA-	
	code_postal_social	0.0289%		and a contact	Maria Maria Maria Maria	January Com.	
		0.0%				Registres	Descripció
	Committee in	0.0%			Incorrectes		Els registres in correctes son aque lis que un cop feta la remediació de les dades encara incompleixen un a o més validacions.
		0.0%			Come de s		Els regist res correctes son aque ils que un cop feta la remediació de les dade scomple laent otes les validadons.
	ultim tranit	0.0%					TOTAL
		0.0%					
		0.0%					
		0.0%					
		0.0%					
		0.0%					
	r territorial	0.0%					
	est at_actual	0.0%					
Red	stres incorrectes un cop	remediats laue cont	louen fallac	nt on alguna	regla de validació:		
	and the second second second	Territoria de la composición della composición d	100	The state of the s	CENTRAL PROPERTY.		
				legies -	Registre s que falle n		
				3	567		
				13 2	5		
					6		
				1	5s		
					E		
				113	2		
				1	St		
				5 13	5		
					526		
				9 2	5		
				914	Y		



Centre de Telecomunicacions i Tecnologies de la Informació (CTTI)

www.gencat.cat

ptd.ctti@gencat.cat