# PTD—Estàndards tècnics de govern

Finançat per la Unió Europea NextGenerationEU

GOBIERNO DE ESPAÑA

Plan de Recuperación, Transformación y Resiliencia

Next Generation Catalunya

Generalitat de Catalunya

# Estàndards tècnics de govern—INDEX

Index


4. Estructura de Capes
16. Nomenclatures
26. Seguretat
32. Privacitat

# Estàndards tècnics de govern

## Necessitat i avantatges dels estàndards tècnics

| Necessitats | Solució | Avantatges |
| --- | --- | --- |
| Establir un procediment per assegurar que tots els desenvolupaments compleixen el model de govern i de bones pràctiques. | **Capes funcionals**: Múltiples capes de validacions i transformacions abans d'emmagatzemar-se per després fer una anàlisi eficient.<br><br>**Nomenclatura**: conjunt de guies i estàndards per definir els noms dels actius analítics. | • Orientació Data Mesh<br>• Traçabilitat<br>• Auditoria<br>• Dades catalogades<br>• Consistència<br>• Processos reutilitzables |

# 1

# Estructura de Capes

# Estructura de Capes

## Definició i implementació de l'estructura

### Per a què serveix?

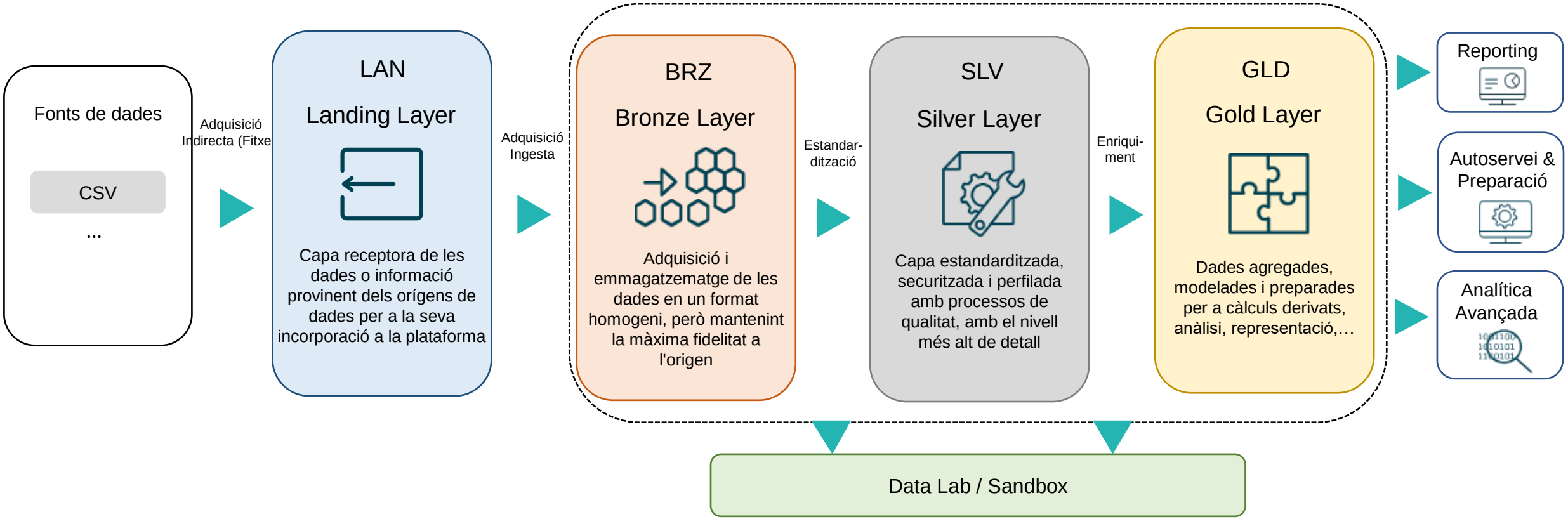
L'estructura de capes en el DataLake permet **separar les dades en funció de la maduresa**, en estat original, de les estandarditzades i de les enriquides (explotables). Aquesta estructura diferenciada permetrà **estandarditzar el flux de dades**, així facilitar tant una correcta definició dels **nivells de seguretat** i aplicació de permisos d'accés a la dada, com facilitar el **govern i descobriment d'aquesta**.

### Capes funcionals

Les capes funcionals són **divisions lògiques** que distingeixen els diferents estats per on ha de passar una dada al llarg del cicle de desenvolupament i el seu processament. Com que és una distribució lògica s'haurà de definir la **metodologia d'implementació** a cada component respectiu en què s'emmagatzemi o gestioni informació.
Aquestes diferents capes serien, **Landing** que es distribuirà per origen de dades, **Bronze** per domini i aplicació origen, **Silver** per subdomini funcional i **Gold** per Cas d'ús.

# Ordenació funcional

## Proposta d'esquema de les diferents capes lògiques en la PTD



# Capa Landing (LAN)

Es tracta d'una capa receptora de les dades o informació (generalment fitxers) provinent dels orígens de dades per a la seva incorporació en la plataforma

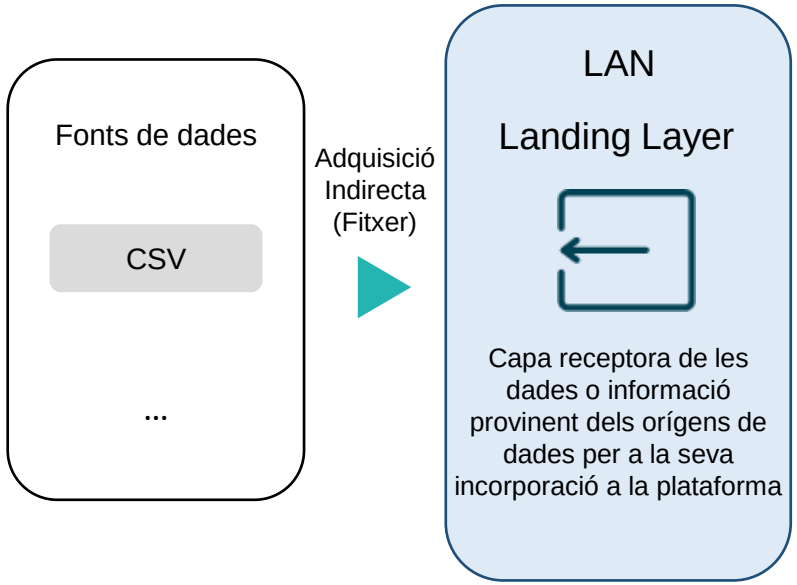
## Característiques principals

- La capa Landing **no forma part del Data Lake**, sinó que s'emmagatzema per separat i és el pas previ a la primera capa lògica (Bronze).
- La capa Landing rep de **manera temporal els fitxers** i els processa per a **enviar-los a la capa Bronze**. En el processament **no es realitzen transformacions.**
- La dada **no es troba estructurada** en aquesta capa, ja que l'emmagatzematge és efímer (**capa no persistent**).
- El propòsit d'aquesta capa és servir com a **punt transitori** de la dada, el qual no persisteix en aquesta capa.
- El **fitxer s'elimina** una vegada és acceptat per al seu moviment.

## Divisió funcional: Aplicació origen

**Propòsit**

Emmagatzemar de forma temporal fitxers a introduir en la plataforma (per exemple, un FTP).



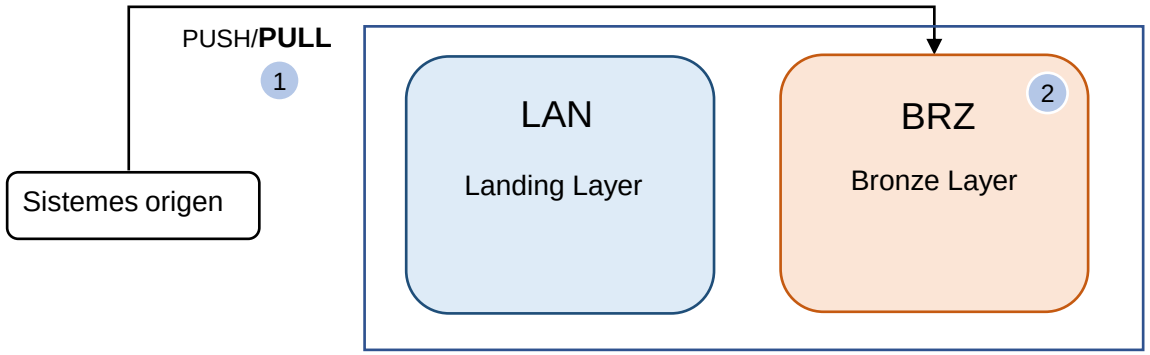
# Capa Landing (LAN)

El procés d'adquisició extreu les dades dels sistemes origen per a la seva posterior ingesta en el repositori comú.

Existeixen dos procediments d'adquisició de dades segons es produeixi directament en la capa de Bronze, o pel contrari, la primera recepció es produeixi en la capa de Landing.

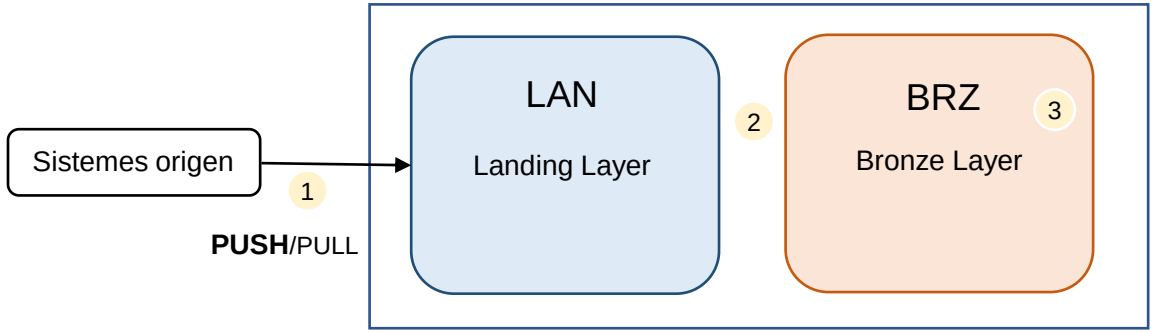
**Procediment d'adquisició directa:**

- **L'extracció i ingesta** d'informació **es produeix directament des dels sistemes origen** fins a la capa de Bronze, ja que les dades s'homogeneïtzen en format. Per exemple el CDC no rep fitxers, una API de REST connectada directament a BRZ, o una BBDD.
- La capa Bronze posseeix la **dada en cru**, amb el **mateix format** i la **mateixa estructura**. Bronze funciona com back-up dels fitxers originals davant una possible auditoria o potencials processes.




**Procediment d'adquisició indirecta:**

- Aquest tipus d'extraccions necessiten un **pas intermedi**. La recepció d'informació es produeix en la **capa Landing**. La capa de Landing posseeix **la dada original íntegra**, és a dir, sense cap transformació.
- Les **dades originals íntegres es traslladen a la capa de Bronze**. En aquest procés, es produeix l'homogeneïtzació únicament de format dels fitxers, si fos necessari.
- Bronze posseeix la dada en cru, amb el mateix format i la mateixa estructura que Landing. BRZ funciona com back-up dels fitxers originals davant una possible auditoria o potencials processos.



# Capa Bronze (BRZ)

En aquesta capa es reben els fitxers des de diferents orígens per al seu processament a les capes del Data Lake. Les dades es reben en brut i es realitza una validació de l'estructura i s'emmagatzemen les dades respectant el contingut provinent de l'origen.

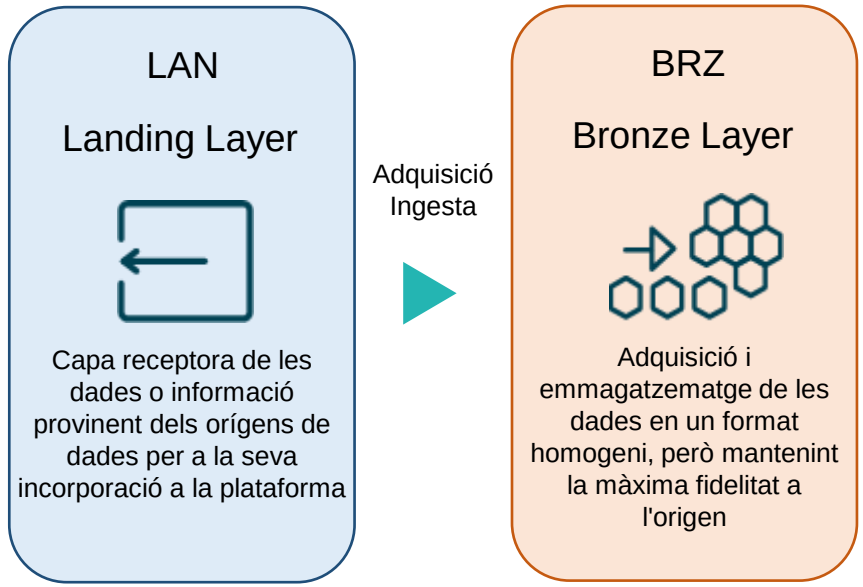
## Característiques principals

- Les diferents fonts origen envien dades a aquesta capa mitjançant **sistemes push o pull** (aquest sistema és previ a la pròpia plataforma)
- Aquestes dades s'incorporen en un format homogeni però mantenint els valors de les dades d'origen, **respectant la naturalesa de la dada** (màxima fidelitat a l'origen)
- És una capa operativa per a permetre el processament posterior a la plataforma, incorporant les dades dels orígens de forma incremental o total per simplificar el procés d'ingesta
- Les dades s'organitzen en funció del seu origen per a **guardar una traça de la seva procedència** i no es realitza **cap adaptació a la seva nomenclatura**
- Les dades s'emmagatzemen en un sistema de fitxers distribuït
- Intervenen **Usuaris de procés** (Gestors de la càrrega), amb accés i capacitat de lectura, creació i esborrat (no edició)

## Divisió funcional: Domini / Sistema Origen

**Propòsit**

Emmagatzemar de forma persistent a la plataforma les dades d'origen amb l'objectiu de independitzar-la de la font. No es realitzen transformacions



# Capa Silver (SLV)

La capa Silver presenta tota la informació disponible amb el major nivell de detall, securitzada i que ha estat validada per processos bàsics de qualitat (format de dades, validació de patrons o transformacions bàsiques, entre d'altres).

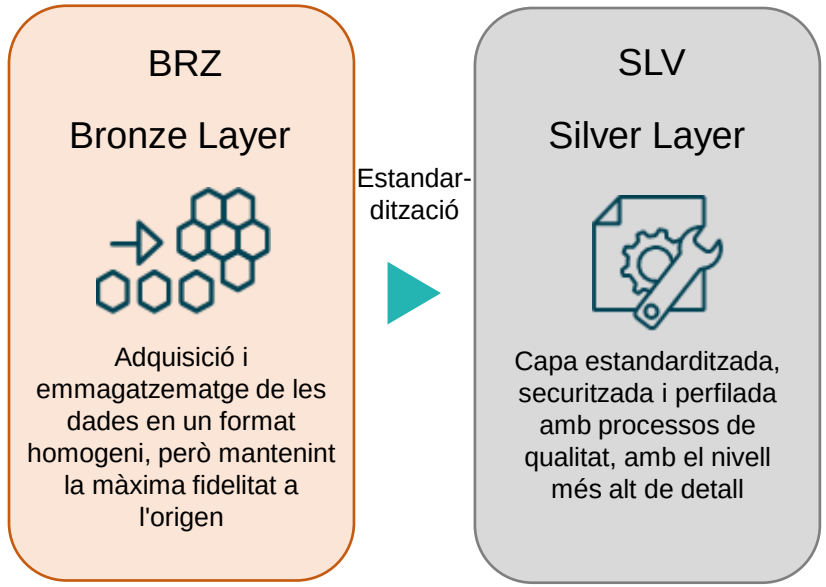
## Característiques principals

- Es processen les dades de la capa Bronze consolidant les ingestes (incremental, deltes, total, ...), fent una **homogeneïtzació i/o estandardització de la dada i aplicant regles de qualitat,**
- La capa estàndard **emmagatzema les dades amb el màxim nivell de detall possible**, aplicant validacions i altres processos de seguretat i qualitat (formats, camps nuls, integritat referencial i unicitat)
- Les dades s'organitzen en agrupacions funcionals que hauran de ser coherents amb les taxonomies i àmbits funcionals definits per a cada departament, tenint també una nomenclatura uniforme i de negoci sigui quin sigui el cas d'ús
- Les dades s'emmagatzemen en sistemes de bases de dades SQL o NO SQL tot i que pot haver casos en que es guardi en fitxers.
- Poden intervenir dos tipus d'usuaris, **Usuaris consumidors** (Analistes & Data Scientists) amb accés i explotació (lectura) i **Usuaris de procés** (gestors de càrrega) amb accés i capacitat de lectura, edició, creació i esborrat

## Divisió funcional: Domini / Subdomini

**Propòsit**

Disposar d'una visió complerta de les dades d'origen amb qualitat, homogènies, amb una ordenació i nomenclatura funcional i comprensible



# Capa Gold (GLD)

Es tracta d'una capa processada amb dades preparades per a l'anàlisi, els càlculs, la visualització i altres explotacions. La capa Gold respon a les necessitats pròpies de cada departament i atorga una visió modelada de dades agregades i calculades

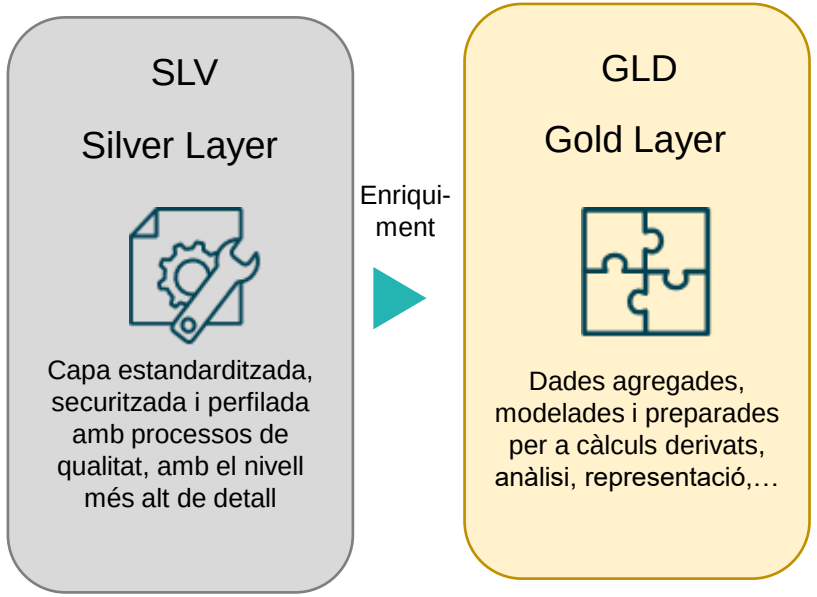
## Característiques principals

- Aquesta capa conté les dades adients per a necessitats de negoci concretes i/o casos d'ús, essent **dades enriquides, processades, agregades**, etc..
- Les dades contingudes en aquesta capa han de ser molt confiables, és a dir, **màxima qualitat** i **ben definides.**
- Conté només l'historial necessari per a la explotació per la que ha estat creada, disposant a la capa estàndard del històric complert.
- Les dades es poden emmagatzemar en un sistema de fitxers distribuït, en sistemes de bases de dades o NO SQL.
- Poden intervenir dos tipus d'usuaris, **Usuaris consumidors** (Analistes & Data Scientists) amb accés i explotació (lectura) i **Usuaris de procés** (gestors de càrrega) amb accés i capacitat de lectura, edició, creació i esborrat.

## Divisió funcional: Domini / Cas d'ús

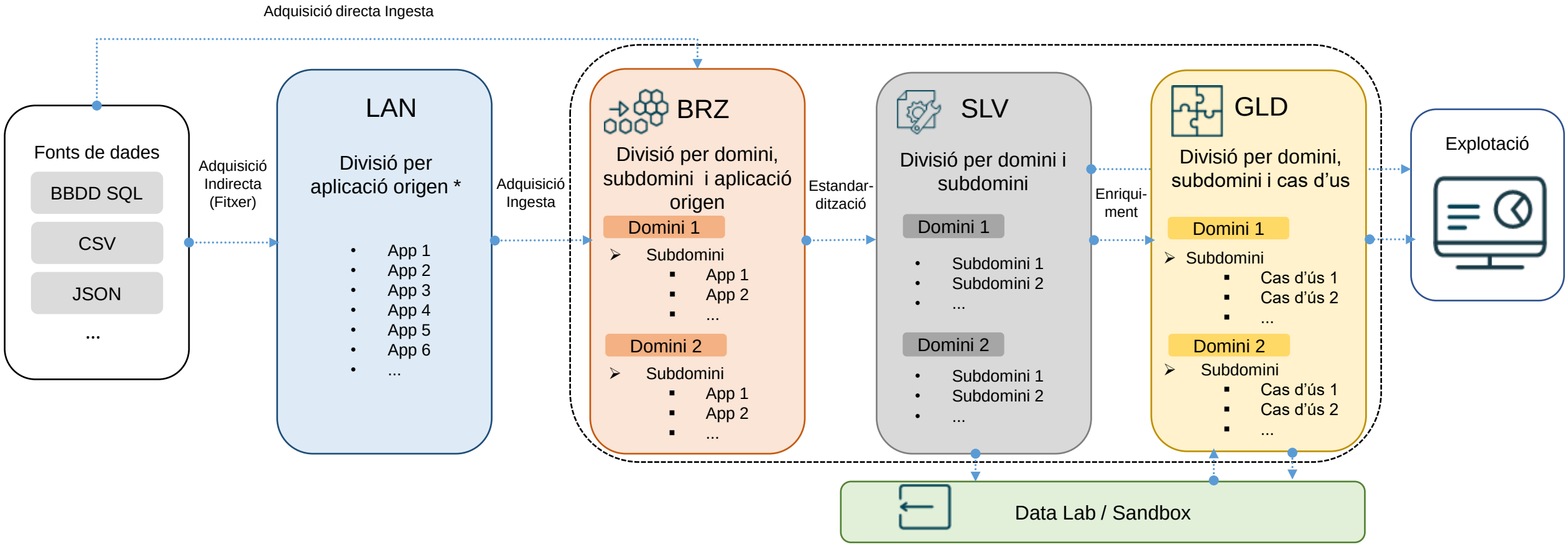
**Propòsit**

És la capa de dades processades i de qualitat per a facilitar la analítica avançada, la creació de quadres de comandament, explotació de la dada, etc.



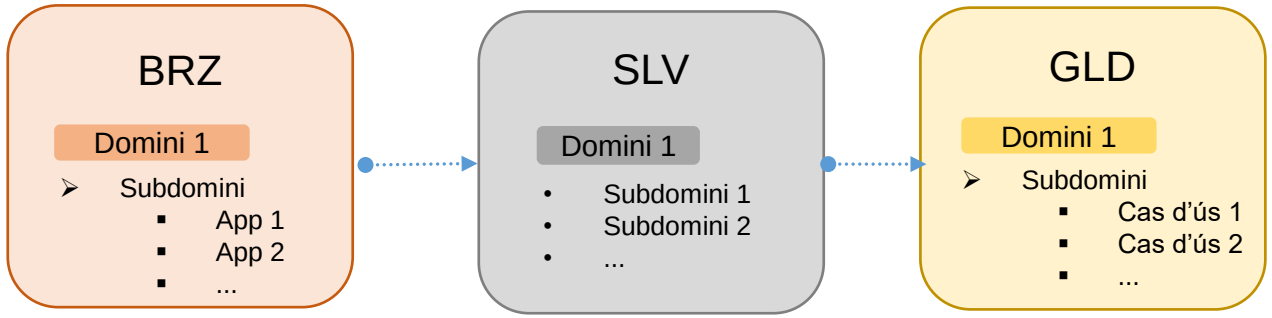
# Distribució funcional per capa dins de la plataforma analítica

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



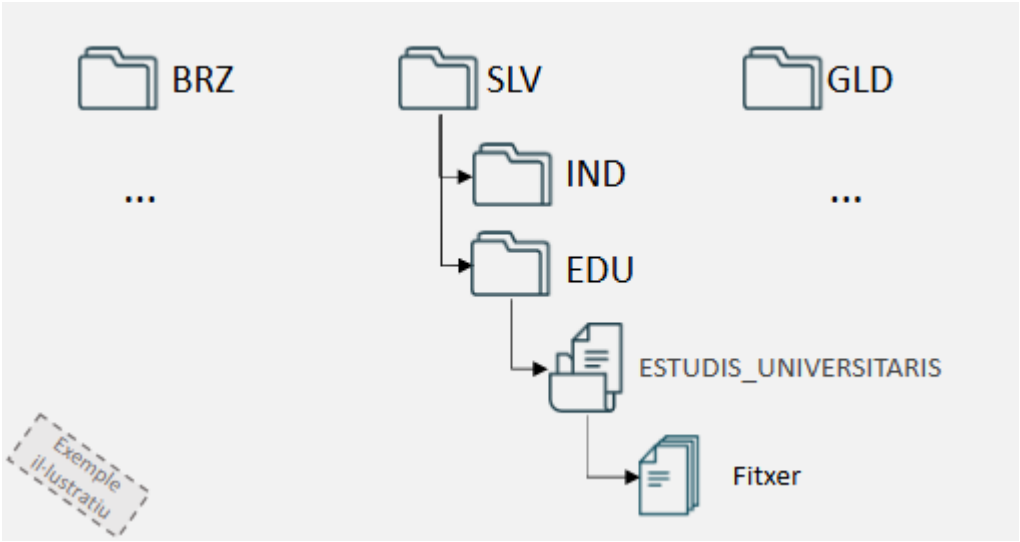
# Implementació de les capes en Components tecnològics

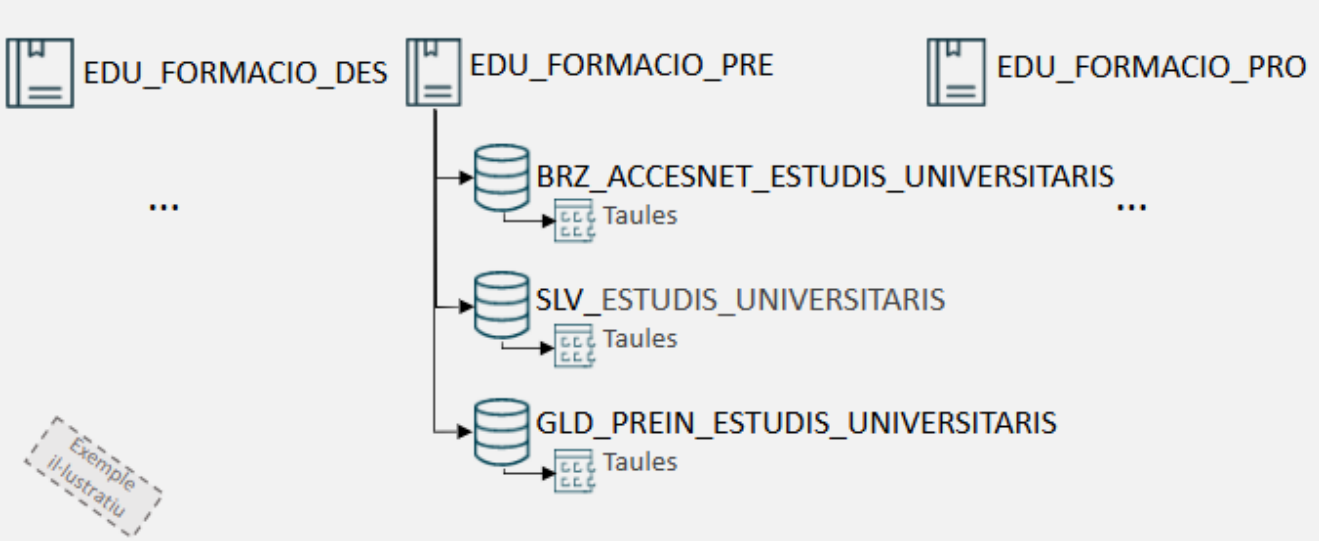
L'estructura de capes és una organització funcional pel que s'ha de definir com implementar-la en els diferents components tecnològics.
Com a exemple, s'exposa el cas d'una distribució per directoris en un data lake i de BBDD:




**Cas Data Lake**: Per al cas de directoris, com existeix la llibertat de crear tots els nivells necessaris, es possible replicar les mateixes divisions que en l'estructura lògica.

**Cas BBDD (Databriks)**: Per al cas de Databricks, l'estructura d'emmagatzematge que ofereix, es basa en 3 nivells (catàleg -> BBDD -> taula). Hi haurà un catàleg per cada domini i entorn (DES, PRE, PRO).

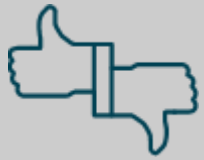


# Alineament del cas RDM

## Taules pròpies d'un cas d'ús, que contenen dades de referència

Es pot donar la situació en què existeixi una taula pròpia d'un cas d'ús que contingui dades que també pertanyen a una taula de referència. En aquest cas es procedirà de la següent manera:



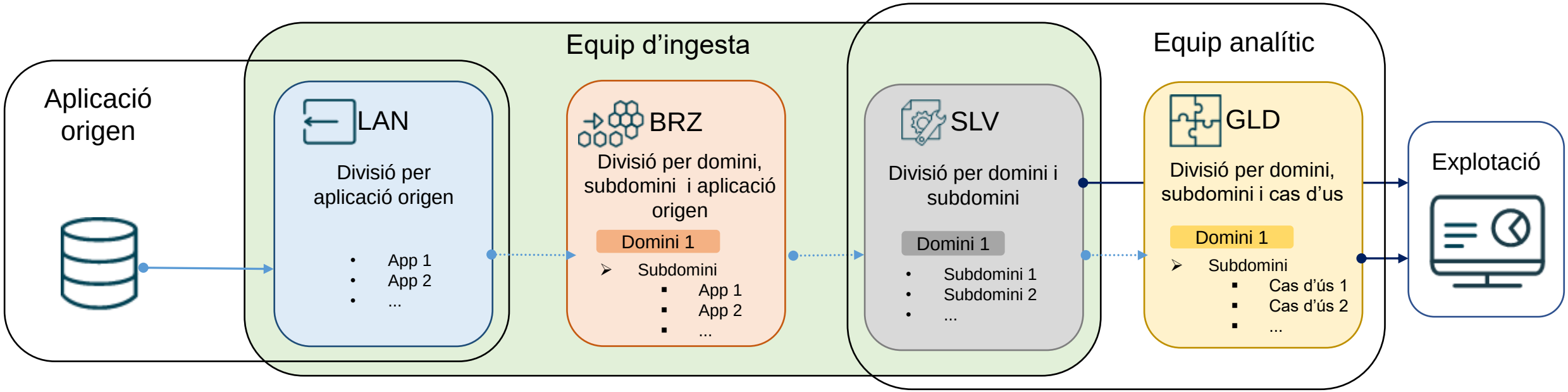
- Primerament, s'haurà de passar integritat referencial sobre aquestes taules, en cas de detectar discrepàncies, s'informarà l'origen perquè realitzi les modificacions pertinents.

- En cas que la taula pròpia del cas d'ús tingui camps addicionals, fet que fa que no es pugui utilitzar directament la de referència, es relacionarà amb el terme de negoci més apropiat, que molt probablement, formarà part d'un domini i subdomini diferent.

- Encara que el terme de negoci relacionat amb aquesta taula no formi part del subdomini on es troba la resta de taules del cas d'ús, s'emmagatzemarà dins de la mateixa BBDD on estiguin les taules del cas d'ús.

- S'especificarà a la descripció de la taula, que el contingut d'aquesta només està sent utilitzat en el present cas d'ús.

# Flux de responsabilitats al llarg de l'estructura de capes




**Aplicació origen**
L'aplicació origen carrega el fitxer en el directori de la capa LAN que s'hagi assignat, en cas que sigui una ingesta Push

**Equip d'ingesta**
L'equip de la PTD estandarditza la dada (MDM, RDM, DQ (qualitat tècnica), nomenclatures…) Avantatge, les dades es trobaran en el mateix format, ja que ha estat estandarditzat pel mateix equip.

**Equip analític**
Els equips de cada cas d'ús treballen amb les dades ingestades generant productes de dades (resultats de models, taules creuades, informes precalculats...)

**Usuaris explotadors**
Usuaris que poden explotar la informació mitjançant eines de reporting

# 2

# Nomenclatures

# Nomenclatura

## Definició i implantació de polítiques de nomenclatura

### Què és?

La política de nomenclatura és **un conjunt de guies i estàndards** que ajuden a definir els noms de tots els actius analítics.

### Per què serveix?

La nomenclatura és el primer pas cap a un **ús correcte de la informació**. Identificar cada idea pel seu **nom adequat i estandarditzat facilita el tractament i l'explotació de la informació** en qualsevol punt del model corporatiu.

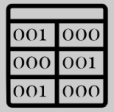
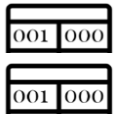
### Avantatges

Una política ha d'assegurar una dada:

Auto-explicativa

Clara i concisa

Estandarditzada

| Entitat | Estàndards | Exemples origen | Exemple nomenclatura |
|---|---|---|---|
| Taules  | •Noms representatius, singulars i amb "_" com separador.<br>•<br>•Amb prefix identificatiu del contingut i tipus de taula. | •CENTRES | •DIM_PAU_CENTRE |
| Camps  | •Noms representatius, singulars i amb "_" com separador.<br>•<br>•Amb prefix identificatiu del contingut i tipus de camp. | •TRAMITS | •IDE_TRAMIT<br>•DES_TRAMIT<br>•IDE_TIPUS_TRAMIT |

# Bones pràctiques

## Estàndards de creació de les entitats de dades a la plataforma

| POLÍTIQUES GENERALS | EXEMPLE |
|---|---|
| Donada la naturalesa i la projecció del negoci, s'utilitza **el català** com a llengua vehicular en la notació de conceptes per taules i camps, entre d'altres. | No aplica |
| Dins de la nomenclatura no s'utilitzaran caràcters especials com accents, dièresis o altres (com "ç" que es substituirà amb "c") | IDE_CIUTADA en lloc de IDE_CIUTADÀ |
| Els **noms funcionals** han de ser **representatius** del seu contingut. | IDE_CIUTADA |
| Els **camps** sempre es definiran en **singular** sempre que la semàntica ho permeti. | IDE_CONTRACTE en lloc de IDE_CONTRACTES |
| No existeixen restriccions en la **longitud** del nom de les taules o camps, però es recomana utilitzar conceptes simples i intuïtius (que no superin 30 caràcters). | NUM_SUPERF_M2_UTIL en lloc de NUM_SUPERFICIE_METRES_QUADRATS_UTILS |
| El nom ha d'estar escrit en majúscules. | IDE_VEHICLE |
| **Evitar l'ús d'abreviatures** en camps o taules que puguin provocar confusions debut a la seva ambigüitat o el seu ús comú a d'altres àmbits. | DES_TIPUS_CONTRACTE en lloc de DES_TP_CTR |
| Les diferents paraules o termes que s'utilitzin sempre es **separen** per "_", mai amb espais. | NUM_SUPERF_M2_UTIL en lloc de NUM SUPERF M2 UTIL |
| A la capa **Bronze** tots els camps han de **mantenir el seu format d'origen**. En cas que no provinguin de fitxers i necessitin adaptacions (per exemple un format de dades no considerat dins la PTD), el format d'adaptació serà tipus VARCHAR. | No aplica |
| A totes les taules hauran d'existir **camps de control: data de càrrega** (TMS_CARREGA) referent a la data en que la Informació va ser carregada a la base de dades; i **la data de les dades** (TMS_MODIFICACIO) referent a la data en què va realitzar la darrera modificació del registre. Es una informació purament tècnica arran de quan es va realitzar el procés d'integració a la taula consultada, necessari per realitzar proves de validació de dades. | TMS_CARREGA i TMS_MODIFICACIO |

# Bones pràctiques

## Proposta d'estàndards de nomenclatures per dominis, subdominis i cas d'ús

| Domini | Subdomini | Cas d'ús |
| --- | --- | --- |
| Es tracta d'acrònims de 3 lletres. Que permeten abreviar el nom actual dels dominis per tal de definir els noms de les BBDD i directoris de la manera més curta possible, basats en les inicials. | Són abreviacions de dues paraules. Que permeten abreviar el nom actual dels subdominis per tal de definir els noms de les BBDD i directoris de la manera més curta possible i sent autoexplicatius. | Al principi de cada cas d'ús es determina una abreviació o un acrònim de com a màxim una paraula, que servirà per identificar el cas d'ús i disposar una versió del nom més curta que s'utilitzarà en les BBDD i directoris. |
| • Medi ambient: **MAM**<br>• Seguretat i emergències: **SEM**<br>• Serveis socials: **SSO**<br>• Educació, formació i universitats: **EDU** | • Defensa de la competència i de consumidors: **COMPET_CONSUM**<br>• Cooperatives i economia social: **COOPE_ECOSOCIAL**<br>• Estudis universitaris: **ESTUDIS_UNIVERSIT** | • Preinscripcions universitàries: **PREIN**<br>• Predicció probabilitat d'ocupació: **PROPOC** |

# Notació dels objectes

## Notació catàlegs Databricks

En l'estructura d'emmagatzematge que existeix a Databricks, es necessària la creació d'agrupacions de bases de dades, els catàlegs. Es crearan a partir del domini i l'entorn tecnològic (DES, PRE i PRO). Tots els catàlegs creats hauran de seguir la següent notació:

[ABREVIACIÓ DOMINI]_[ENTORN TECNOLÒGIC]

| Codi | Acrònim | Nom del domini | Abreviació Domini |
|---|---|---|---|
| D01 | ADM | Administració pública, govern i relacions institucionals | ADMIN_GOVERN |
| D02 | AGR | Agricultura, ramaderia, pesca i alimentació | AGRIC_ALIMENT |
| D03 | CCT | Comerç, consum i turisme | CONSUM_TURISME |
| D04 | CLT | Cultura, llengua i identitat | CULT_LLENGUA |
| D05 | ECO | Economia | ECONOMIA |
| D06 | EDU | Educació, formació i universitats | EDU_FORMACIO |
| D07 | ELL | Esports i lleure | ESPORT_LLEURE |
| D08 | FPH | Finances públiques i hisenda | FINANCES_HISENDA |
| D09 | IND | Indústria i energia | IND_ENERGIA |
| D10 | INF | Informació i comunicació | INFORM_COMUN |
| D11 | JUD | Justícia i dret | JUSTICIA_DRET |
| D12 | MAM | Medi ambient | MEDI_AMBIENT |
| D13 | MOB | Mobilitat i transports | MOBILITAT |
| D14 | SLT | Salut | SALUT |
| D15 | SEM | Seguretat i emergències | SEGUR_EMERG |
| D16 | SSO | Serveis socials | SERVEIS_SOCIALS |
| D17 | SCI | Societat i ciutadania | SOCIETAT |
| D18 | TEC | Tecnologia, recerca i innovació | TECNO_INNOVA |
| D19 | TER | Territori, infraestructures i urbanisme | TERRITORI |
| D20 | TRB | Treball | TREBALL |
| D98 | NCO | No consta | NO_CONSTA |
| D99 | ALT | Altres/Diversos | ALTRES |

# Notació dels objectes

## Notació Capes lògiques i Taules BRZ

Es planteja l'ús d'esquemes per capes lògiques en base a la següent codificació, independent de l'entorn tecnològic (DES, PRE i PRO):

CAPA LÒGICA

Identifica la capa funcional i, per tant, la maduresa de l'estat de les dades. Aquest camp es limita a 3 caràcters i pot prendre els valors següents:

- **BRZ**: Capa Bronze on les dades es carreguen tal qual com ens arriben dels orígens.
- **SLV**: Capa estandarditzada, securitzada i perfilada amb processos de qualitat, amb el major nivell de detall.
- **GLD**: Capa amb les dades agregades, modelades i preparades per a càlculs derivats, anàlisis, representació...

### Nomenclatura directoris data lake (per emmagatzemar fitxers)

•Bronze: [CAPA_LOGICA]/[DOMINI]/[SUBDOMINI]/[APPORIGEN]
•Silver: [CAPA_LOGICA ]/[DOMINI]/[SUBDOMINI]
•Gold: [CAPA_LOGICA]/ [DOMINI]/[SUBDOMINI]/[CAS_US]

EXEMPLE:
•Bronze: BRZ/EDU/ESTUDIS_UNIVERRSITARIS/ACCESNET/
•Silver: SLV/EDU/ESTUDIS_UNIVERSITARIS
•Gold: GLD/EDU/ESTUDIS_UNIVERSITARIS/PREIN

### Nomenclatura BBDD (per emmagatzemar taules)

•Bronze: [CAPA_LOGICA]_[APPORIGEN]_[SUBDOMINI]
•Silver: [CAPA_LOGICA ]_[SUBDOMINI]
•Gold: [CAPA_LOGICA]_[CAS_US]_[SUBDOMINI]

EXEMPLE:
•Bronze: BRZ_ACCESNET_ESTUDIS_UNIVERSITARIS
•Silver: SLV_ESTUDIS_UNIVERSITARIS
•Gold: GLD_PREIN_ESTUDIS_UNIVERSITARIS

# Notació dels objectes

## Notació taules SLV i GLD

Per determinar el nom d'una taula física de dades s'utilitzen estructures codificades per cada una de les capes lògiques, independentment a l'entorn tecnològic al que pertanyin (DES,PRE o PRO). Totes les taules creades a les capes SLV i GLD, hauran de respectar la següent notació:

[TIPUS TAULA]_[NOM FUNCIONAL]_[FASE DQ]

| CAPA SILVER (SLV) i CAPA GOLD (GLD) | |
|---|---|
| TIPUS TAULA | Determina la tipologia de la taula i tindrà una extensió fixe de 3 caràcters. A la capa SLV i GLD, existiran 11 tipus de taules disponibles:<br>•"DIM": Taula dimensional que serveix per a classificar o categoritzar un fet.<br>•"FAC": Taula *fact o* de fets que conté els valors de les mesures o indicadors de negoci.<br>•"AGG": Taula agregada d'una taula de fets respecte alguna de les seves dimensions d'anàlisi.<br>•"REL": Taula relacional que s'utilitza per establir el vincle entre dos dimensions.<br>•"LKP": Taula lookup o diccionari que s'utilitza per a la traducció o descodificació de valors.<br>•"SUK": Taula per identificar claus subrogades.<br>•"AUX": Taula auxiliar, per necessitat d'un pas intermedi dins d'un procés tècnic.<br>•"REF": Taula que conté dades de referència.<br>•"MES": Taules mestres.<br>•"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.<br>•"FTR": Taules que contenen les característiques (features) que s'utilitzaran com a variables independents en un model d'aprenentatge automàtic. |
| NOM FUNCIONAL | Identifica de forma simple el concepte de negoci que descriu millor el contingut de la taula |
| FASE DQ | Es una notació opcional per a aquelles taules que s'utilitzin en el procés de data quality |
| VISTA | Indicar si es tracta d'una vista amb la lletra "V_". |
| EXEMPLE | DIM_UNIVERSITATS; REF_CODIS_POSTALS; AGG_IMPORT_MENSUAL_SUBVENCIO |

# Notació dels objectes

## Notació taules de caràcter tècnic

En el cas de les taules que s'utilitzen en processos  tècnics de mongo DB i que contenen dades tècniques o de configuració que no són de negoci i són no explotables, es procedirà de la següent manera:

- No es realitzarà govern, es a dir, no és necessari seguir els estàndards tècnics de nomenclatures, ni capes.
- No es catalogaran les dades, no és necessària la presa de metadades, ni l'ús de les plantilles.
- No seran visibles pels usuaris, ja que no estaran a l'eina de govern.

- No es realitzarà govern, però es farà ús dels prefixos per identificar-les:
  - "TEC": Taula tècnica, de configuració interna de la solució.
  - "LOG": Taules de logs o de auditoria i registre de canvis o històrics.
  - "TMP": Taula temporal utilitzada com a pas intermedi per als desenvolupaments.

TEC_MDM_CH_POS

```
CREATE TABLE ptd_dev.mdm.mdm_ch_pos (
  id_census_object BIGINT NOT NULL,
  attribute STRING NOT NULL,
  start_date STRING NOT NULL,
  source STRING,
  execution_id BIGINT,
  id_reg STRING NOT NULL,
  original_id STRING,
  priority INT NOT NULL,
  value STRING,
  end_date STRING,
  CONSTRAINT mdm_ch_pos_pk PRIMARY KEY (`id_census_object`,
`attribute`, `start_date`))
USING delta
--LOCATION 's3://s3-dev-dataprime-
data/HNK/data/lakehouse/core/operational_layer/mdm/mdm_ch_pos'
TBLPROPERTIES ('Type' = 'EXTERNAL');
```

# Notació dels objectes

## Notació dels camps

El nom dels camps d'una taula és un element vital per a poder identificar la informació de forma coherent. Per tant, s'han definit les següents tipologies de camps per tal de facilitar la comprensió dels conceptes. Cada tipus de camp ha d'anar associat a un format de dades predefinit que garanteixi la integritat i homogeneïtat de la informació. Tots els camps de SLV i GLD han de satisfer la notació

[TIPUS CAMP]_[NOM CAMP]

| TIPUS CAMP | DESCRIPCIÓ TIPUS DE CAMP | FORMAT DATA LAKE | FORMAT QUERY ENGINE | EXEMPLE |
|---|---|---|---|---|
| IDE | Identificador. Es recomana utilitzar identificadors del tipus integer excepte en casos on pugui ser necessari tipus varchar. | INT o NVARCHAR | INT, INTEGER, BIGINT, SMALLINT, TINYINT, BYTEINT // VARCHAR, CHAR | IDE_PORT |
| DAT | Data (date). Per defecte tindrà una precisió a nivell dia. | DATE | DATE | DAT_INICI_CONTRACTE |
| HRS | Data i hora. Per defecte tindrà una precisió nivell dia i hora. | DATETIME | DATETIME | HRS_VENCIMENT_VIGENCIA |
| TMS | Timestamp. | TIMESTAMP | TIMESTAMP | TMS_CARREGA |
| DES | Descriptor. Els descriptors podran tenir associat el sufix _{IDIOMA} per a determinar l'idioma del seu contingut que tindrà. Aquest sufix tindrà una extensió fixe de 2 caràcters (ISO 639-1). | NVARCHAR | VARCHAR, CHAR | DES_TIPUS_CONTRACTE |
| FLG | Flag o Booleà (boolean). | INT o NVARCHAR | BOOLEAN | FLG_VIGENCIA |
| 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). | DECIMAL | DECIMAL | IMP_VENDA |
| NUM | Indicador numèric sempre i quant no sigui un import o percentatge. | DECIMAL | DECIMAL | NUM_SUPERF_M2_UTIL |
| PER | Indicador numèric per percentatges. | DECIMAL | DECIMAL | PER_APROVAT |
| URL | Utilitzat per anomenar camps que contenen una direcció web (URL). | NVARCHAR | VARCHAR, CHAR | URL_TRIPADVISOR |

# Notació dels objectes

## Notació dels camps especials

El nom dels camps d'una taula és un element vital per a poder identificar la informació de forma coherent. Per tant, s'han definit les següents tipologies de camps per tal de facilitar la comprensió dels conceptes. Cada tipus de camp ha d'anar associat a un format de dades predefinit que garanteixi la integritat i homogeneïtat de la informació. Tots els camps de SLV i GLD han de satisfer la notació:

[TIPUS CAMP]_[NOM CAMP]

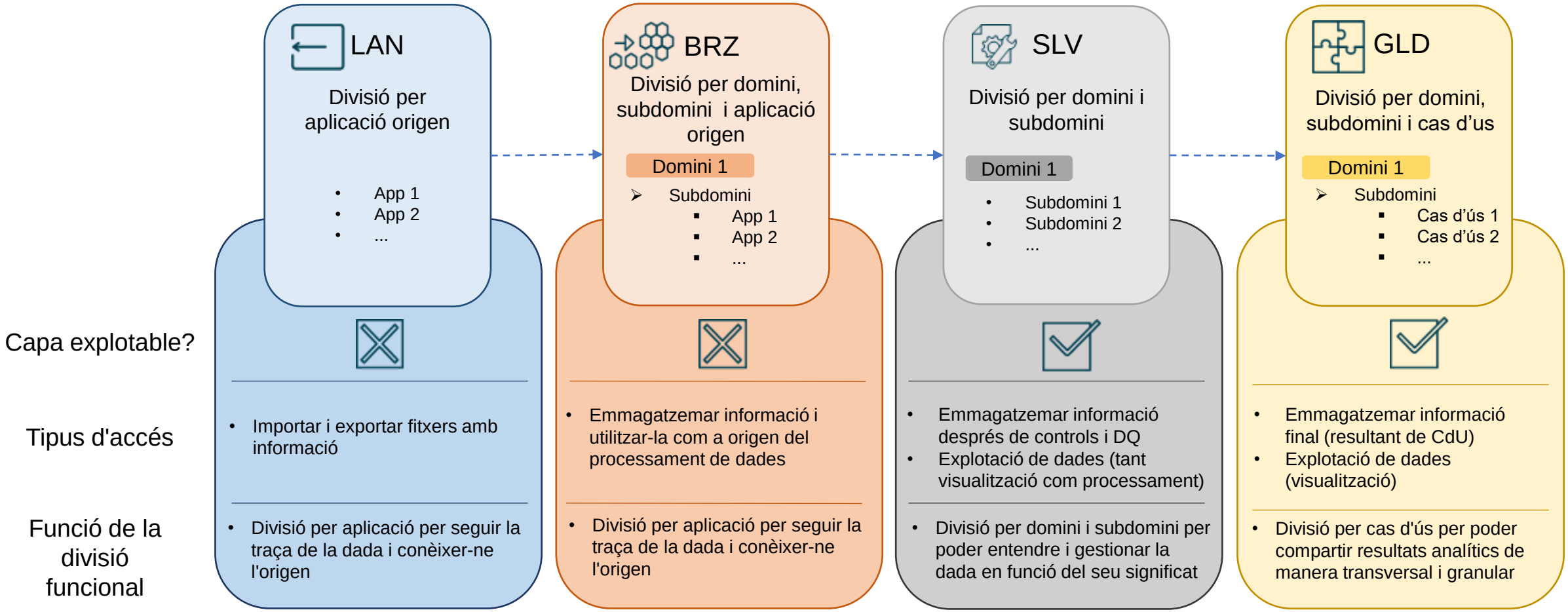
| TIPUS CAMP | DESCRIPCIÓ TIPUS DE CAMP | FORMAT DATA LAKE | EXEMPLE |
|---|---|---|---|
| OBJ | Objecte. Tipus de dada complexa composta de dates amidades | OBJECTE (Mongo db) | OBJ_ALUMNE |
| ARR | Array. Col·lecció d'elements del mateix tipus | ARRAY (Mongo db - databricks) | ARR_CURSOS_MATRICULATS |
| MAP | Map. Diccionari de dates clau-valor | MAP (Data lake - Databricks) | MAP_CODIS_LLICENCIA |
| STU | Struct. Tipologia "objecte" equivalent en Databricks | STRUCT (Data lake - Databricks) | STU_ALUMNE |

# 3

# Seguretat

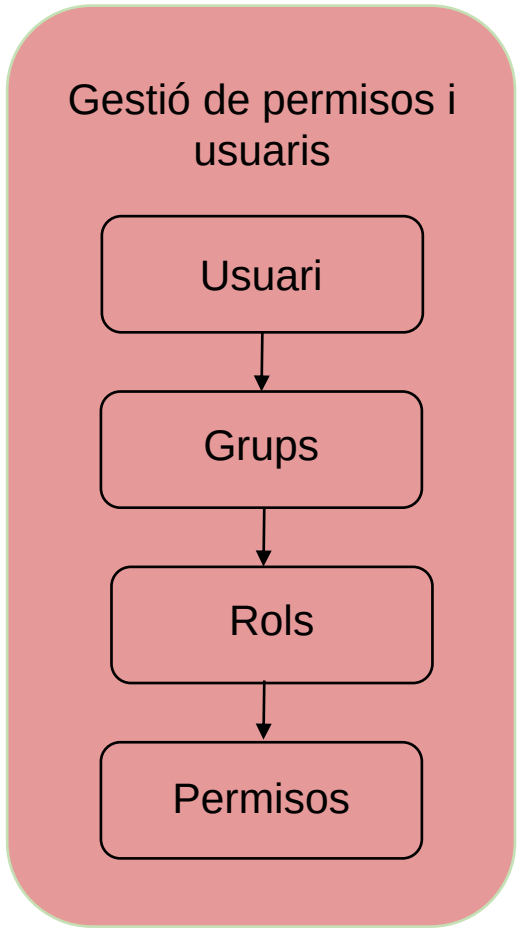
# Accés a les capes en funció de la finalitat

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.



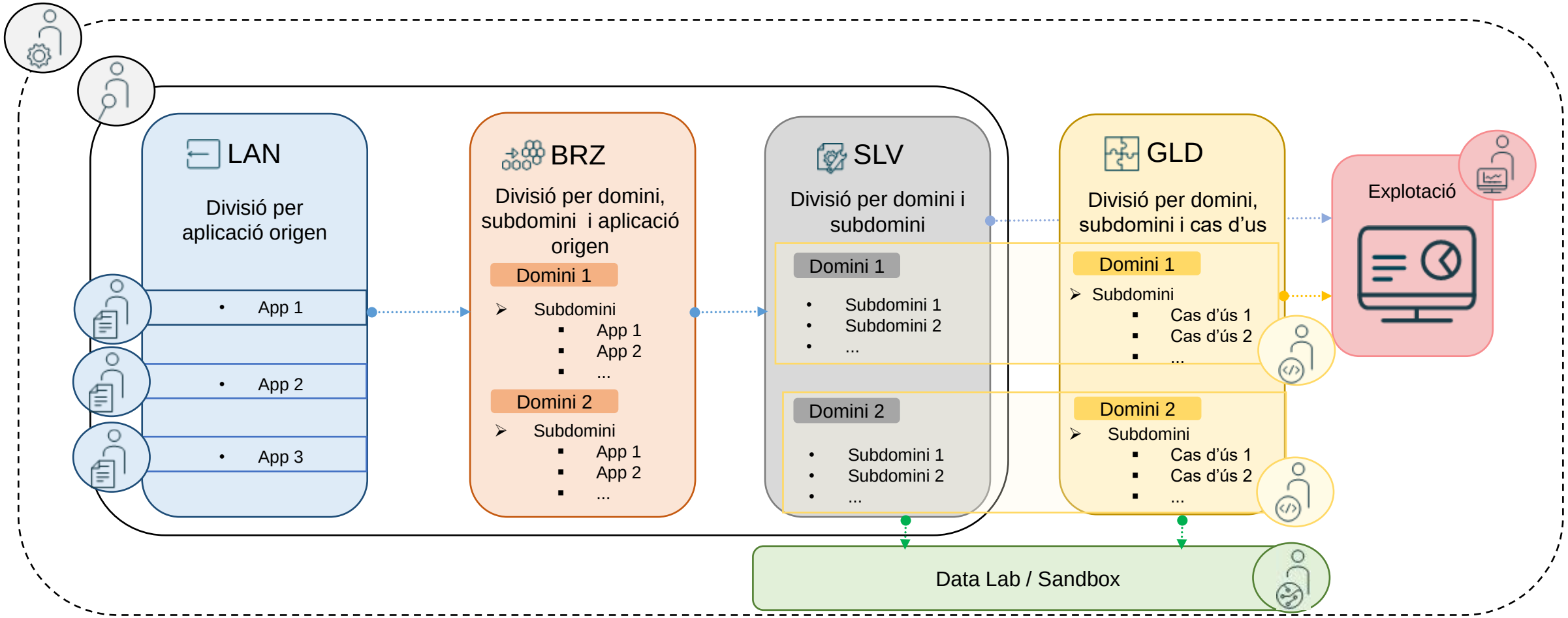
# Usuaris, funcions i permisos dins de la plataforma

La creació d'usuaris i concessió de permisos es realitzarà mitjançant grups i rols per facilitar-ne la gestió de manera àgil.




| Tipus d'usuari | Funció | Capes i permisos | Divisió de permisos |
|---|---|---|---|
| Aplicació | Importació de fitxers a la capa Landing. | •Landing (read & write) | Aplicació origen |
| Procés | Gestió de càrrega des de Landing fins a Silver. | •Landing (read & write)<br>•Bronze (read & write)<br>•Silver (read & write) | Nivell plataforma |
| Explotació projecte | Explotació i transformació/ processament de dades entre Silver i Gold. | •Silver (read)<br>•Gold (read & write) | Subdomini (SLV)<br>Cas d'us (GLD) |
| Lectura projecte | Visualització de dades de Silver i Gold. Particularitzable en funció cas d'ús. | •Silver (read)<br>•Gold (read) | Subdomini (SLV)<br>Cas d'us (GLD) |
| Govern | Control de la metadata i compliment de polítiques. | •Landing (read metadata)<br>•Bronze (read metadata)<br>•Silver (read metadata)<br>•Gold (read metadata) | Nivell plataforma |
| Data Scientist | Creació i gestió de models d'analítica avançada. | •Silver (read)<br>•Gold (read) | Subdomini (SLV)<br>Cas d'us (GLD) |

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




Usuari Aplicació

Usuari Procés

Usuari Explotació Projecte

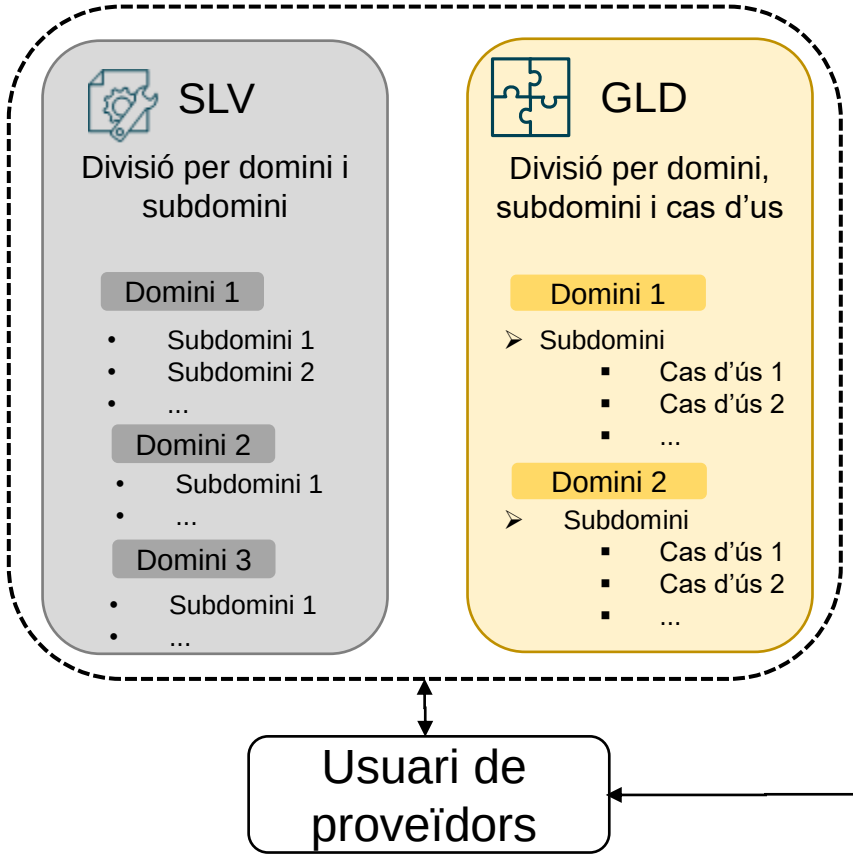
Usuari Lectura Projecte

Usuari Govern

Usuari Data Scientist

# Sistema de permisos d'usuaris multiproveïdor gestionat per subdominis

La distribució funcional de la plataforma permet gestionar de manera granular l'accessibilitat de diferents usuaris i proveïdors de serveis

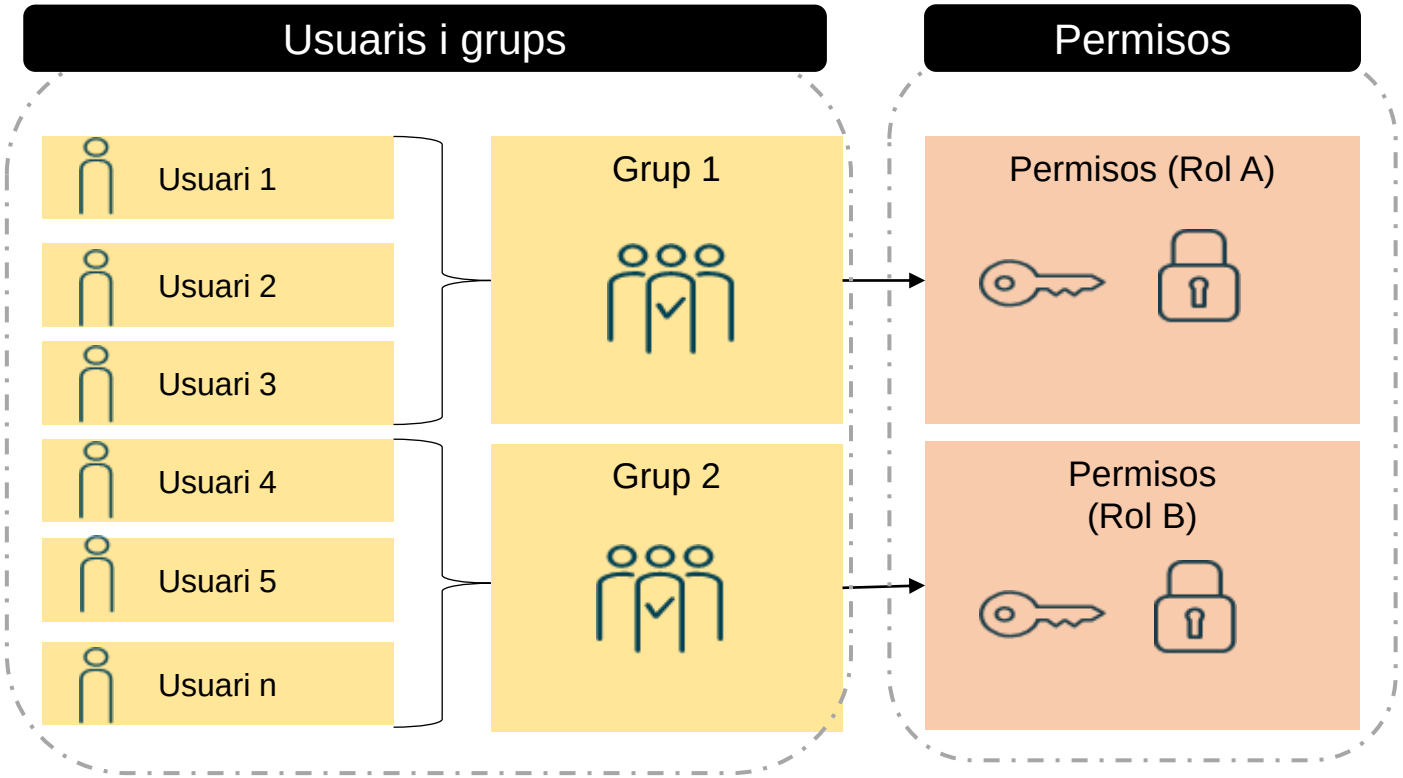



- A la capa Silver, la informació està dividida en **dominis i subdominis**, la qual cosa permet donar **permisos de forma granular** als usuaris de proveïdors, dotant la plataforma de la capacitat d'aïllar i separar les dades visibles a cada perfil (els permisos seran només de lectura),
- A la capa Gold els usuaris proveïdors tindran **permisos de lectura** (usuari lectura projecte) o de **lectura i escriptura** (usuari explotació projecte). Aquests permisos es podran **gestionar a nivell de cas d'ús**, permetent controlar de forma granular la informació a què té accés cada proveïdor i la potestat de poder editar-la o només consultar-la.
- Els rols permeten agrupar les **combinacions de permisos** de Subdominis i casos d'ús necessàries perquè cada perfil pugui exercir les funcions. Per exemple, un usuari de lectura projecte podrà tenir accés de lectura simultàniament a diversos subdominis de dominis diferents i a diversos casos d'ús.

| Tipus d'usuari | Funció | Capes i permisos | Divisió de permisos |
|---|---|---|---|
| Explotació Projecte | Explotació i transformació/ processament de dades entre Silver i Gold iched | •Silver (read)<br>•Gold (read & write) | Subdomini (SLV)<br>Cas d'us (GLD) |
| Lectura Projecte | Visualització de dades de Silver i Gold | •Silver (read)<br>•Gold (read) | Subdomini (SLV)<br>Cas d'us (GLD) |

# Gestió de la seguretat

**Es crearan grups d'usuaris als quals se'ls assignaran rols específics, atorgant-los així determinats permisos.**




Els usuaris* són agregats en grups d'usuaris i des de la plataforma es concediran rols aquests grups, considerant un rol com un conjunt de permisos.

Aquest sistema permet separar la gestió d'usuaris de la de permisos, quan es creï un usuari i s'assigni a un grup, aquest ja comptarà amb els permisos necessaris per operar.

*A excepció d'usuaris màquina

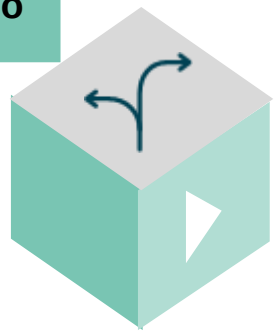
# 4

# Privacitat

# La gestió de la privacitat es basa en 4 accions:

**Minimització**

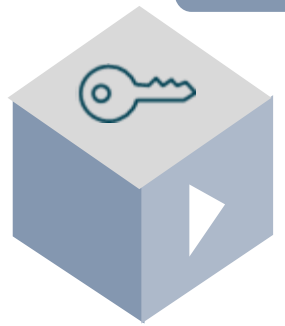
1



Es minimitzarà la informació personal ingestada a la plataforma fent una anàlisi previ dels camps o entitats que puguin estar subjectes a polítiques o regulacions i el seu ús potencial. En altres paraules**, s'evitarà carregar informació personal si aquesta no és necessària.**

**Criptografia**

2



La informació personal que sigui ingestada però no s'exploti directament (dades per creuar taules, futurs casos d'ús**...) s'emmagatzemaran utilitzant funcions criptogràfiques.**

3

**Agrupació estadística**

L'agrupació estadística de dades **implica la supressió d'elements identificatius entre les dades i els individus específics, així les dades no estan associades directament amb les persones.**

**Control d'accés**

4

**Es gestionarà de forma dinàmica i mitjançant permisos qui pot visualitzar cada informació** utilitzant el control d'accés.

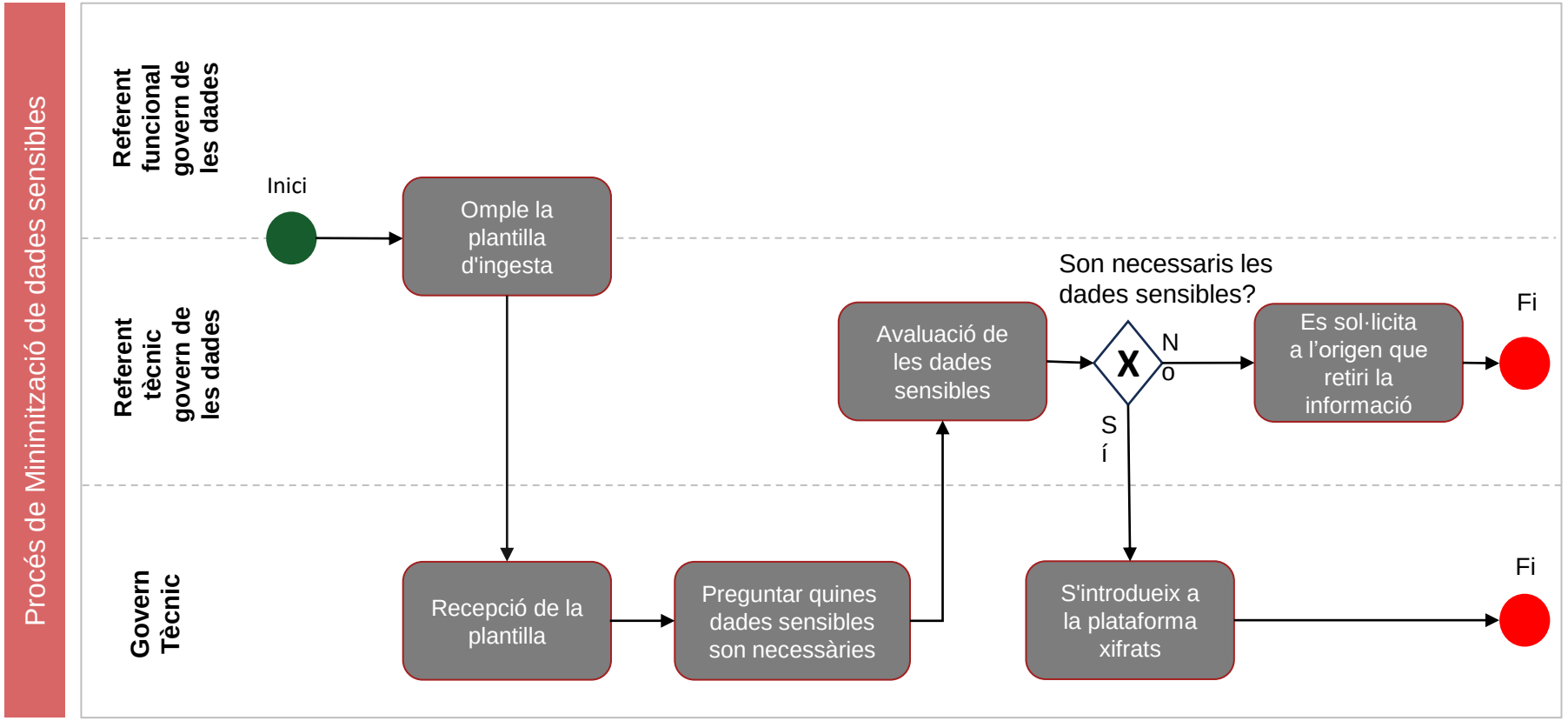
# Què és la Minimització?

La Minimització consta d'ingestar a la  plataforma únicament les dades sensibles que calgui.

## Quines problemàtiques volem evitar?

**La presència de dades sensibles implica riscos** addicionals, que poden posar en perill la seguretat de l'organització.

## Com ho farem?



# Què és la Criptografia?

La criptografia és la codificació de la informació de forma que es converteix intel·ligible per mantenint-la segura i privada a la vista de persones que no tenen els permisos per visualitzar-la.

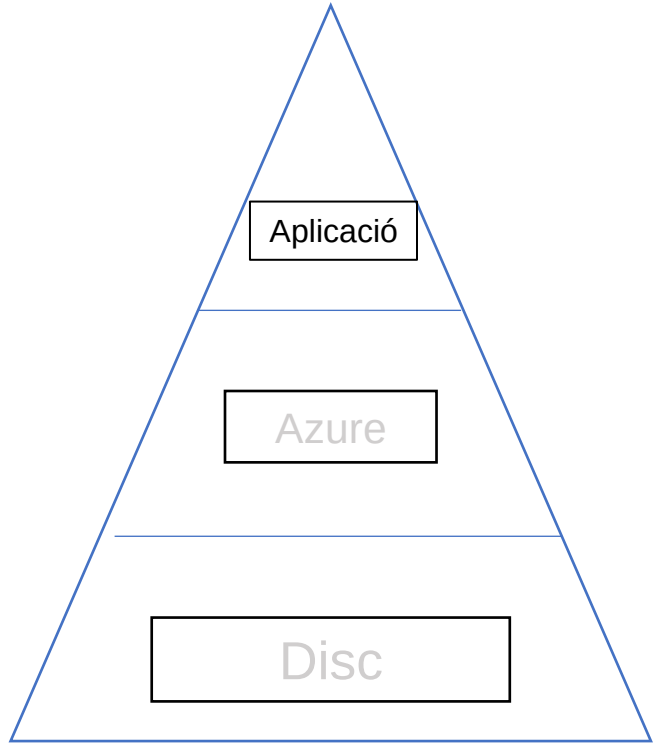
## Quines problemàtiques volem evitar?

La manca d'encriptació deixaria les dades exposades i vulnerables a accessos no autoritzats.

## 3 nivells d'encriptació

A l'aplicació, la informació sensible es xifrarà abans de ser emmagatzemada, adaptant-se segons les necessitats del negoci.

De forma automàtica  Azure encripta sempre tota la informació. A més a més, també s'encripta automàticament la informació que s'emmagatzema al hardware, es a dir, el disc.



# Què és la agrupació estadística?

La agrupació estadística és una estratègia de desvinculació de dades sensibles relacionades amb persones físiques, per visualitzar les dades de forma d'agrupació. La seva finalitat rau en permetre que els usuaris autoritzats puguin accedir i utilitzar aquestes dades a la capa Gold, on es visualitzaran de manera agregada i estadística.

## Quines problemàtiques volem evitar?

No fer una agrupació estadística pot conduir a la divulgació d'informació personal i sensible, vulnerant la privacitat dels individus.

## Exemple pràctic:

| Nom | Salari |
|---|---|
| Ramón Garcia Guardiola | 40 000 |
| Judith Carrasco Huertas | 35 000 |
| Pilar Muñoz Fernández | 20 000 |
| Francisco Pérez Domenech | 14 000 |

| Nombre de persones | Interval salarial |
|---|---|
| 2 | 30 000 – 40 000 |
| 2 | 10 000 – 20 000 |

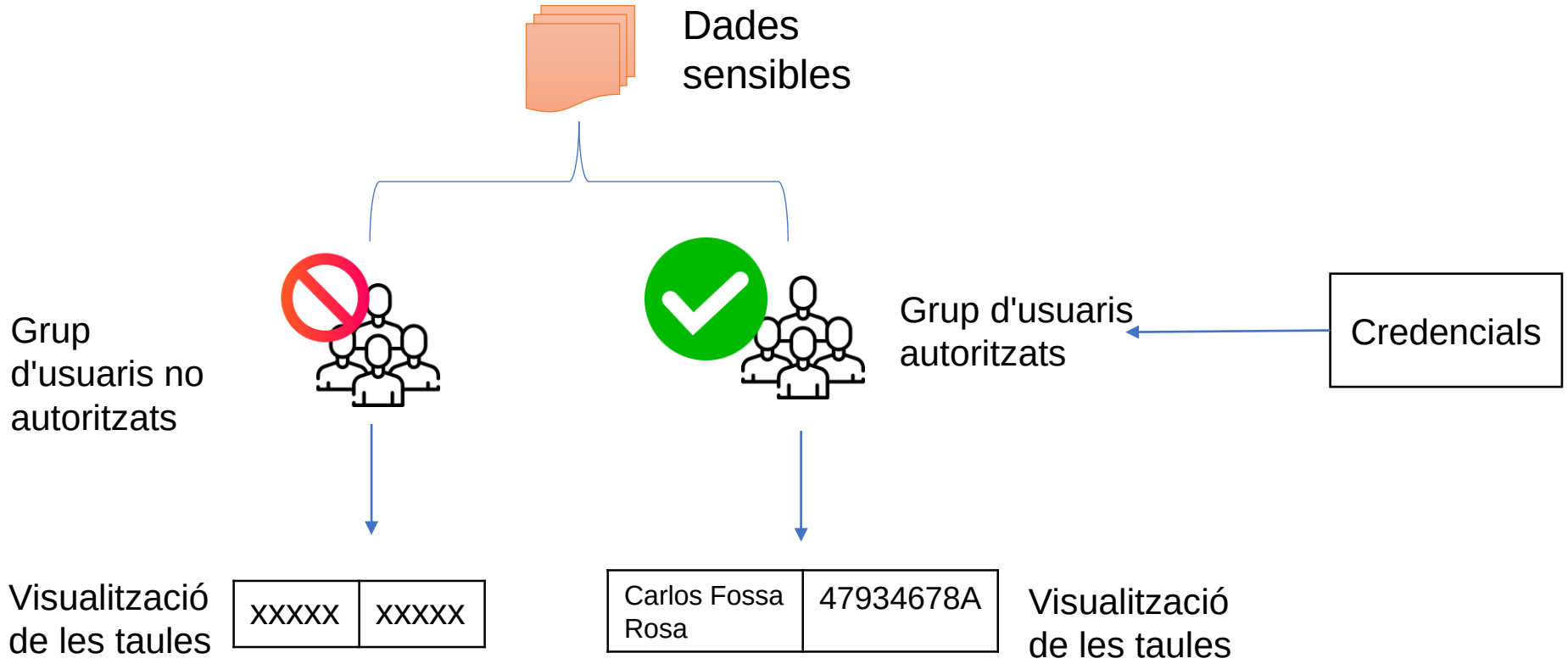
Exemple il·lustratiu

# Què és la agrupació estadística?

Mitjançant els permisos d'accés a la plataforma, es podrà controlar de forma granular (per subdomini i cas d'ús) quins usuaris accedeixen a les entitats analítiques.
Dins una mateixa entitat es podran aplicar diferents polítiques a nivell de camp des de l'eina de govern depenent de la classificació de la dada i l'usuari.

## Quines problemàtiques volem evitar?

El control d'accés ajuda a prevenir que persones no autoritzades obtinguin accés.



# Generalitat de Catalunya


Departament de Polítiques Digitals i Administració Pública

Centre de Telecomunicacions i Tecnologies de la Informació (CTTI)

www.gencat.cat

ptd.ctti@gencat.cat