Welcome pack

Gener 2025













Context

Punt de partida

Reptes en la gestió de dades



Què és i avantatges

La Plataforma Transversal de Dades (PTD) té com a objectiu integrar, gestionar i explotar dades de diferents departaments de manera eficient i segura, permetent a la Generalitat de Catalunya a convertir-se en una organització data driven, maximitzant els avantatges i beneficis que les dades poden oferir per a la societat.

¿Quines avantatges te adherir-se a la plataforma transversal de dades (PTD)?

Estalvi

En ser una plataforma compartida, el cost d'utilització es redueix considerablement. Per al repartiment del cost d'ús de la plataforma es tenen en compte diferents factors, com el tràfic de dades, l'emmagatzematge utilitzat, la quantitat de missatges dipositats, etc.

Temps de posada en producció

El temps de posada en producció es redueix dràsticament, ja que no s'ha de fer cap aprovisionament d'infraestructura, desplegament del producte ni tot el que comporta una plataforma nova.

Robustesa

La plataforma és altament disponible, tolerant a fallades i redundant.

Servei monitoritzat 24/7

La plataforma està declarada com un sistema crític i per tant monitoritzada tot el dia tots els dies pel Centre del Control del CTTI.

Suport

Equip de suport especialitzat en la tecnologia per ajudar-vos des de la definició de l'arquitectura, millors pràctiques de definició de topics, integració amb la plataforma, resolució de incidències i posada en producció. Compta amb suport de fabricant, Databricks, amb temps de resposta inferior a 1h per a incidents de severitat crítica en producció.

Característiques principals



Ser la plataforma corporativa de dades de la Generalitat que permeti adquirir, governar i explotar les dades per tal d'optimitzar els serveis públics.







Adaptat a la realitat tecnològica de la Generalitat

Funcions principals



Capacitat per gestionar múltiples formats



Aplicació de regles de qualitat a dades massives



Capacitat d'explotació i govern de dades massives



Gestió de dades mestres i manteniment de dades de referència



Consolidació i publicació de dades de forma automàtica



Visió única de les dades i també dels algoritmes que es creïn

Objectius

Dades agregades, i catalogades

Poder agregar dades de diverses fonts facilitant una visió unificada d'elles, tant en temps real com en batch, amb paràmetres d'alta qualitat i confiabilitat, publicades en un catàleg central i fàcilment accessibles per al seu consum autònom, amb diferents nivells d'accés, visibilitat i seguretat. Ha de ser la de base per a realitzar els diferents nivells d'anàlisi de les dades que es posin a disposició dels usuaris.

Explotació de les dades

Disposar de dades per a diferents finalitats, sustentades en un govern de dades robust i fiable, però sobretot que es pugui implementar per a obtenir el valor real que les dades poden oferir. Centrar-se en la dada per a estandarditzar-la, publicar-la, i explotar-la tant descriptivament com amb IA.

Qualitat de les dades

Les dades han de passar un conjunt definit de regles de qualitat i de remeiació automàtica que han de garantir que les dades tenen els estàndards de qualitat desitjats. Quan no compleixin aquests estàndards, es definiran interfícies per remeiar-les.

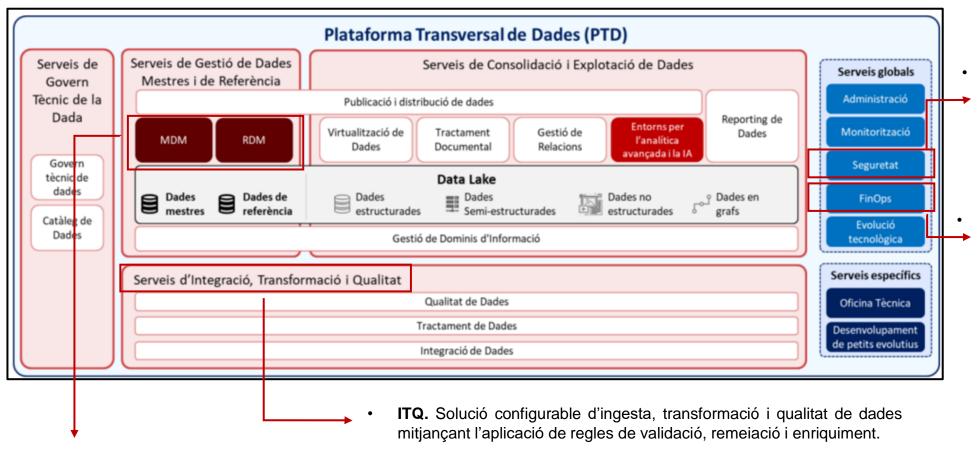
Seguretat de les dades

Establir les directrius bàsiques que han de seguir totes les aplicacions per poder treballar amb la PTD de forma segura, treballant amb l'Agència de Ciberseguretat de Catalunya, seguint l'Esquema Nacional de Seguretat (ENS), juntament amb el marc normatiu de la Generalitat de Catalunya en matèria de seguretat de la informació.

Control de costos

Seguiment de costos mitjançant l'ús d'una estratègia de FinOps ens permeti identificar, mesurar, monitoritzar i optimitzar el costos de cadascun dels projectes desplegats a la PTD.

Mòduls i detall dels principals

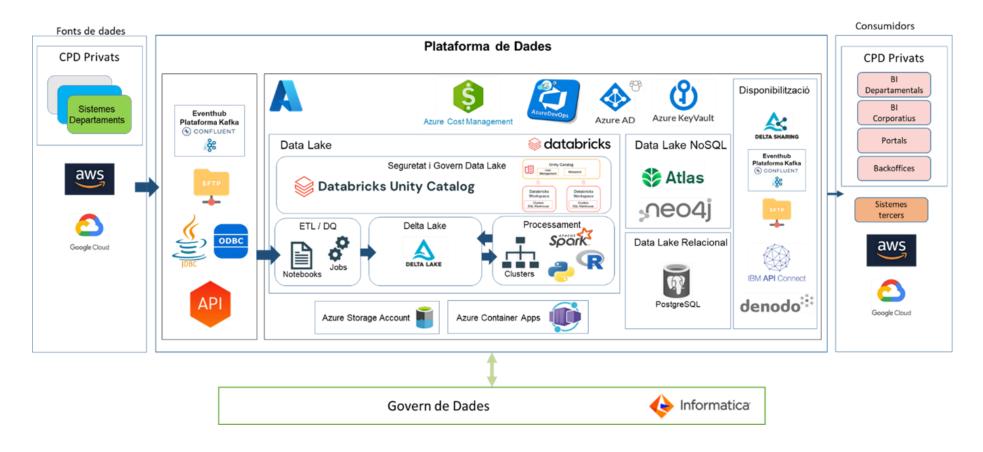


Seguretat. Establiment de les directrius bàsiques que han de seguir totes les aplicacions per poder treballar amb la PTD de forma segura.

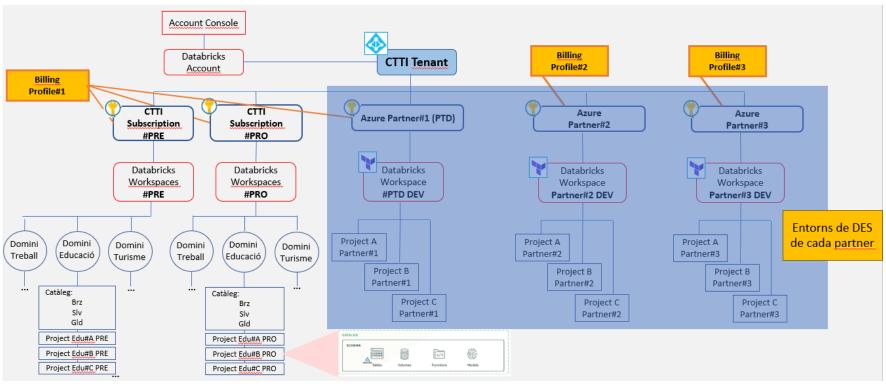
FinOps. Permet mesurar i monitorar l'aprovisionament de recursos al núvol en una organització, per tal d'optimitzar-ne l'ús de manera eficient.

- MDM. Ofereix una visió única i fiable de les dades clau dels departaments (master de dades).
- **RDM**. Dades de referència per proveir dades que seran comunes a diferents aplicacions i departaments.

Diagrama funcional d'arquitectura



Arquitectura integrada

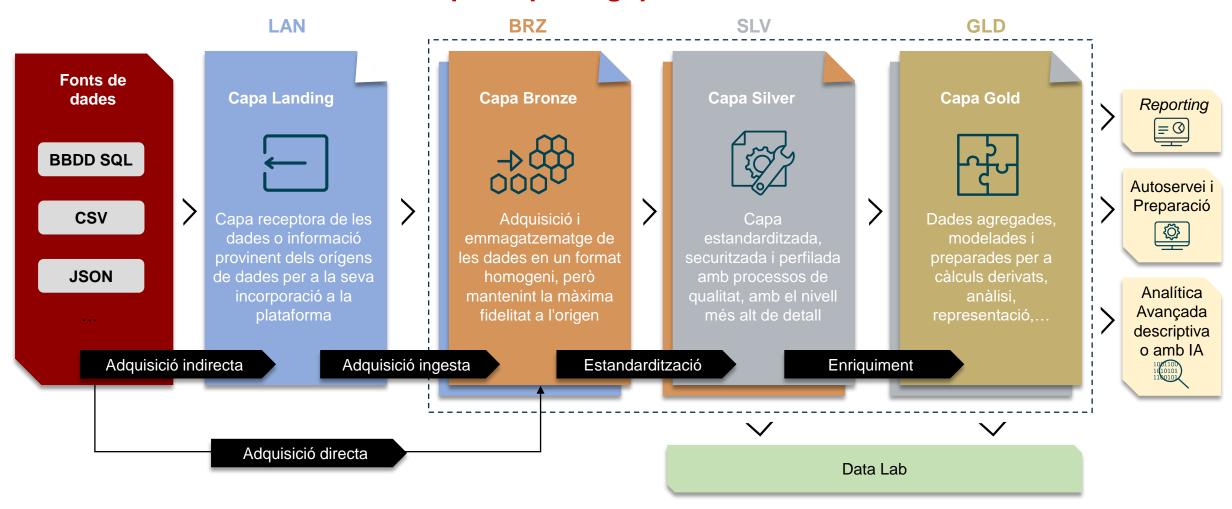


- El compte de Databricks, el Unity Catalog, va pel tenant de CTTI.
- La classificació dels entorns és fa per subscripció Azure.
- Els proveïdors s'integren a la PTD sol·licitant una subscripció on es muntarà l'entorn de desenvolupament. Aquest entorn va a compte de cada proveïdor.
- Cada subscripció té associat un BillingProfile que és el mitjà de facturació.
- En els entorns preproductius existeix un workspace de Databricks per domini per Entorn, un domini és un catàleg.
- A l'entorn del proveïdor (DES), els catàlegs a treballar estaran sota el mateix workspace assignat: [proveïdor]_[Domini]_[environment].

2

Capes lògiques de la PTD i estàndards tècnics

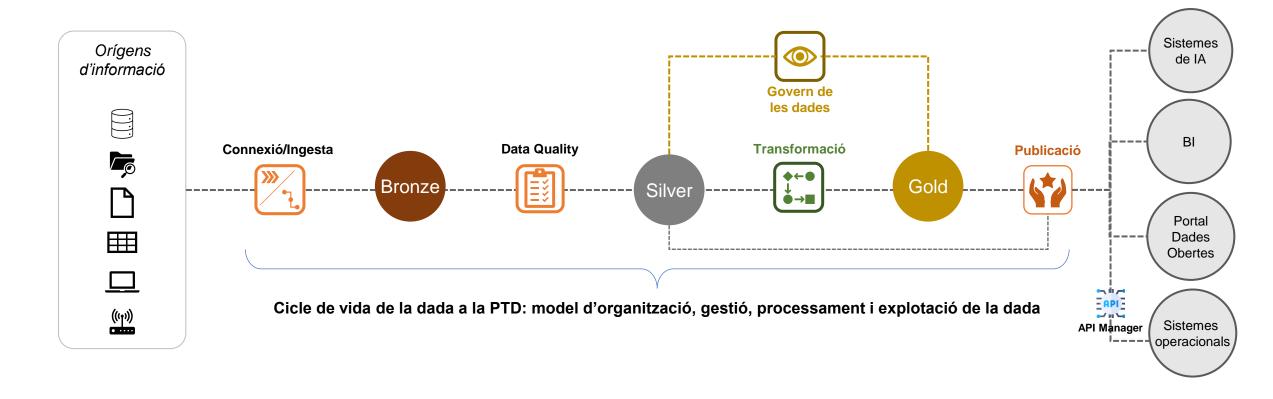
Distribució funcional de la PTD per capes lògiques



3

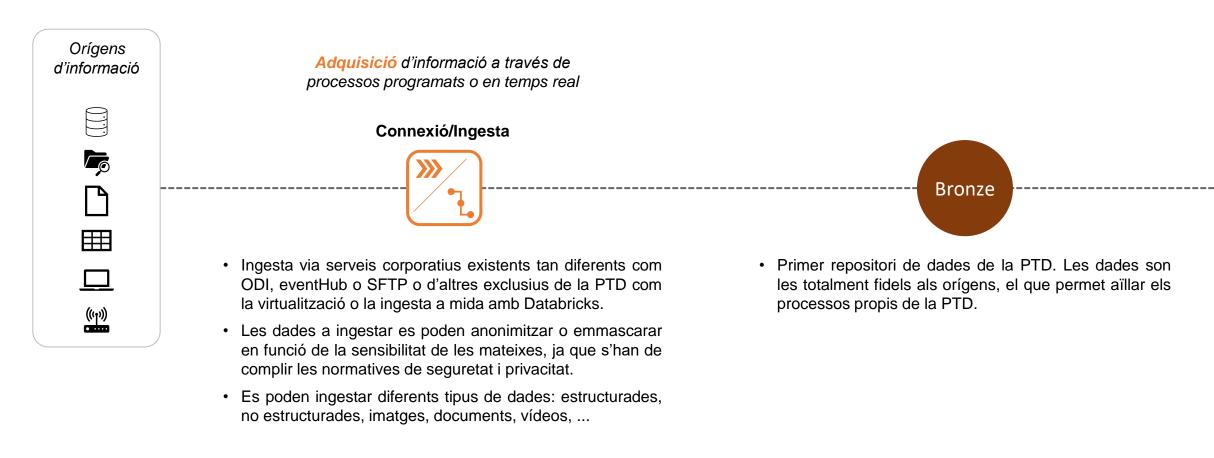
Cicle de la dada

Esquema integral



Ingesta

Comença el viatge a la PTD. Les dades entren al núvol de la PTD (Azure) en un entorn segur i confiable:



Qualitat

Primers serveis que ofereix la PTD per a la gestió de les dades: validar, millorar i enriquir les dades:

Aplicació de regles de qualitat i d'homogeneïtzació perquè les dades entrin estandarditzades

Bronze Silver

- Assegurar la qualitat de les dades és objectiu fonamental de la PTD.
- Hi ha diferents tipus de regles de qualitat, que s'hauran de definir específicament per a cada conjunt de dades, tot i que hi haurà algunes establertes com a bàsiques (homogeneïtzació, etc.).
- Més enllà dels processos de comprovació de la qualitat, es podrà remeiar dades (de forma manual o automàtica) o definir regles d'enriquiment de les dades.
- Els diferents processos de qualitat es poden consumir com a servei (via API)

Transformació

La transformació de les dades és el procés on es preparen les dades per a ser explotades o per a l'objectiu que s'hagi definit:

Transformació de dades mitjançant els diferents mètodes disponibles a la PTD (Databriks: Python, SQL, Scala, ...)

Transformació



 Dades homogènies i de qualitat per a poder executar qualsevol tipus de transformació



- Es poden crear models d'explotació BI (Data Marts, o models d'autoservei dels usuaris), taulells analítics (per fer córrer models analítics, com ML, etc.) o preparar dades per aplicar-hi Intel·ligència Artificial.
- Capacitat per processar grans volums de dades, fent qualsevol tipus de transformació (agregacions, càlculs, creuaments, etc.).
- També permet compartir dades entre departaments (sempre sota la aprovació de totes les parts implicades) per tal de crear productes de dades més complerts.
- Es pot aprofitar el RDM corporatiu, així com crear MDM necessaris per a qualsevol Departament.



- En aquesta capa hi hauran dades processades i preparades per a usos concrets (els que hagin demanat els usuaris).
- Les dades seran de qualitat i amb les mesures de seguretat que siguin necessàries en cada cas.

Govern

El govern de la dada serà un factor clau en la gestió, comprensió i fiabilitat de les dades a la PTD



Inclusió al catàleg de dades corporatiu per tal de facilitar la seva localització i comprensió

Homogeneïtzació de les dades, sigui quin sigui l'origen

· Catalogació d'informes i KPI per facilitar el seu ús i entesa

Explotació

Finalment, les dades serviran per a l'ús que necessiti el peticionari, sent la PTD polivalent en les seves capacitats d'explotació:

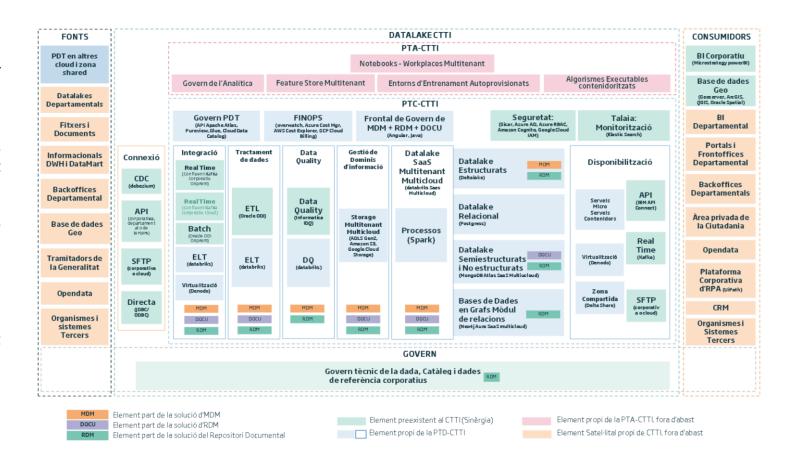


4

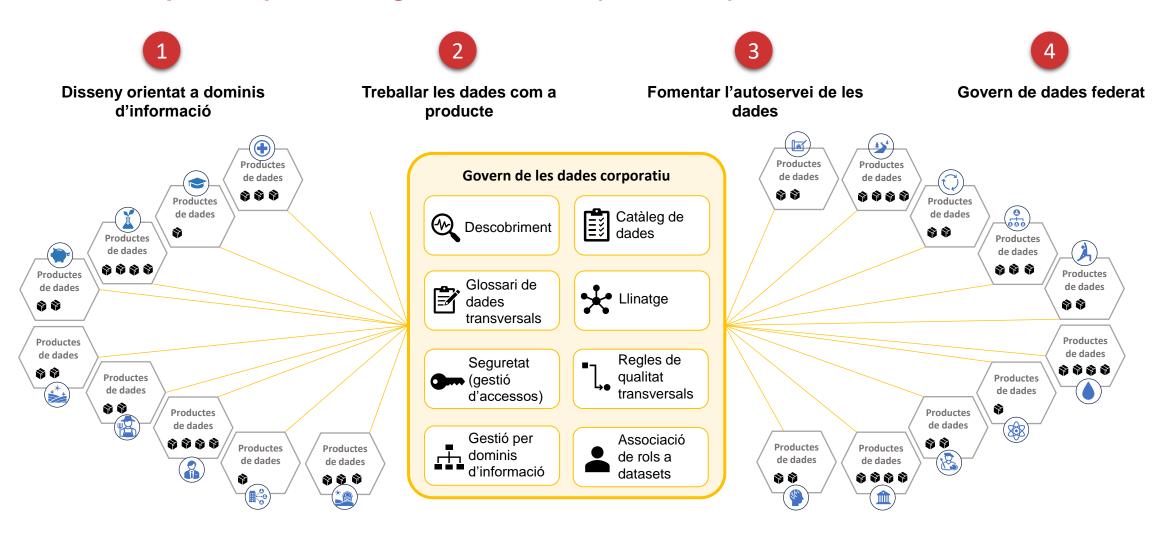
Mòduls i capacitats de la PTD

Característiques claus

- Solució estratégica, integral i gobernada per gestionar i compartir dades a tota l'administració catalana.
- Arquitectura multicloud basada en tecnologies open source com Databricks i Spark, garantint escalabilitat i flexibilitat.
- Proporciona mòduls especialitzats d'Integració de dades, data quality, dades de referencia i mestres, a l'hora de garantir una governança i qualitat de les dades.
- Comparteix una arquitectura federada, descentralitzada i oberta de Gaia-X assegurant la sobirania i seguretat de les dades.



Govern corporatiu però amb gestió federada (datamesh)



Govern corporatiu però amb gestió federada (datamesh)

Connexió/Ingesta



La PTD integra dades
de múltiples orígens
i formats, amb
capacitat d'ingestar les
dades en temps real
o batch, mitjançant
l'automatització amb
Databricks o la
virtualització amb
Denodo.

Data Quality



La PTD proporciona
una solució
configurable de
qualitat de dades
basada en Databricks,
amb regles de
validació,
remediació i
enriquiment preexistents.

Emmagatzemament i dominis d'informació



Gestió centralitzada i
col·laborativa de
dades a la PTD amb
Databricks, mitjançant
un enfocament Data
Lakehouse i
l'estandardització de
fluxos en capes
Bronze, Silver i Gold.

Transformació



La PTD amb Databricks
ofereix una plataforma
integrada per a
l'execució de
processos Spark que
faciliten la
transformació,
l'enriquiment i la
gestió del cicle de
vida de les dades.

Publicació



La PTD ofereix la

disponibilització de

dades segura amb

Databricks, l'API

Corporativa i amb la

virtualització amb

Denodo, que permet

l'accés en temps real a

les dades sense tenir

que duplicar-les al delta

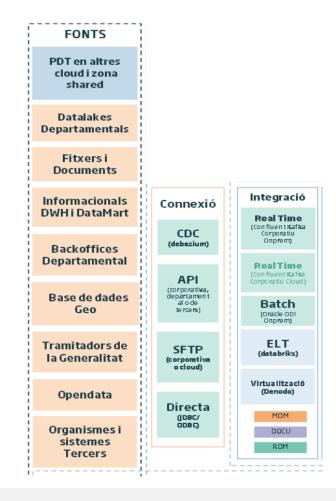
Lake.

Detall de la ingesta

Connexió/Ingesta



- Integració amb múltiples orígens de dades en diversos formats, tant on-premise com en cloud.
- Capacitat d'ingesta en batch o temps real, adaptant-se a les necessitats de processament.
- Capacitat integració amb webservices, API Rest i sistemes d'esdeveniments com Confluent Kafka.
- Utilització de Databricks per a processos adhoc d'ingesta i automatització amb el Mòdul ITQ.
- Virtualització de dades amb Denodo, desacoblant les fonts de dades dels processos de càrrega.

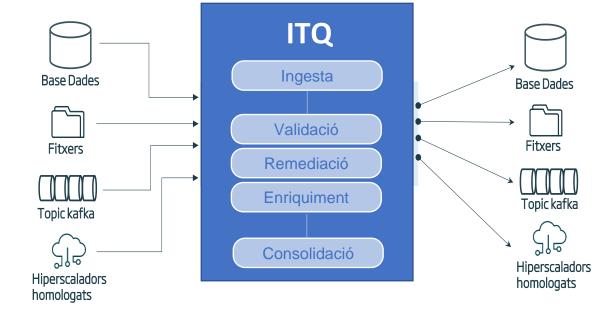


Detall de qualitat de les dades

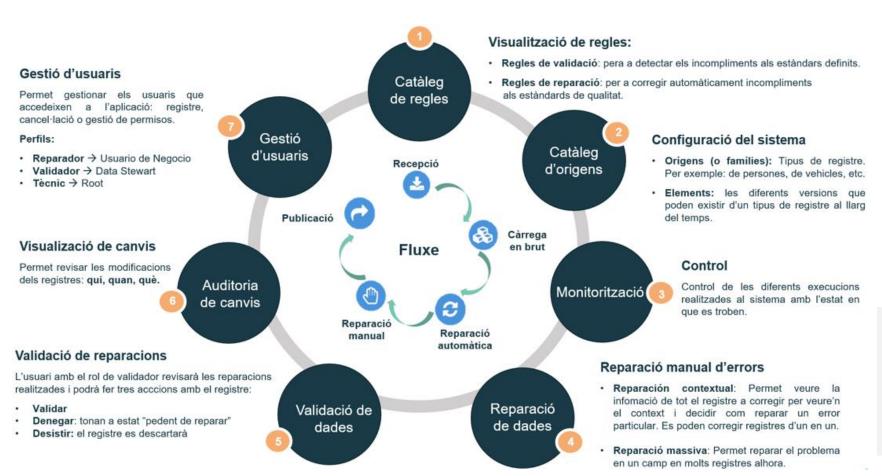
Data Quality



- Solució configurable de qualitat de dades basada en Spark i Databricks amb taules Delta Lake.
- Catàleg de regles per garantir la qualitat de les dades, incloent regles de reparació automàtica i enriquiment d'informació.
- Capacitats de remediació manual i consolidació de dades per garantir informació precisa i fiable.
- Sistema integrat amb GICAR per a l'autenticació, multitenant i control d'accés basat en RBAC.
- Persistencia en el Data Lake i **explotable via l'API corporativa** per a tercers (p.e. Normalitzador d'adreces).



Detall de qualitat de les dades



Regles de validació

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



Regles de remeiació

Processos disenyats per a 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

Mostra de regles de data quality

Exemple de Regles de validació



Regla	Descripció
VALIDATE_NI F	La regla VALIDATE_NIF ens permet comprovar que el valor del camp té la longitud i el format correcte d'un NIF.
VALIDATE_CI F_NIF	La regla VALIDATE_CIF_NIF valida que el valor del camp on apliquem la regla tingui la longitud i el format correcte per ser un CIF o un NIF segons el tipus de persona, per això en els paràmetres passem el valor del camp del tipus de persona (física/jurídica).
VALIDATE_E MAIL	La regla VALIDATE_EMAIL comprova que el camp conté el format correcte d'un correu electrònic.
VALIDATE_CU RRENT_DATE	La regla VALIDATE_CURRENT_DATE valida que la data del camp on apliquem la regla no sigui superior a la data actual.
VALIDATE_BI RTHDATE	La regla VALIDATE_BIRTHDATE valida que el valor del camp de la data de naixement no sigui anterior a 1850.

Exemple de Regles de remeiació



Regia	Descripcio
REM_REG_ REPLACE	La regla REM_REG_REPLACE ens permet reemplaçar els valors que compleixen amb certa expressió regular per un valor constant que configurem.
REM_COU NTY_INVAL ID	La regla REM_COUNTY_INVALID ens permet assignar al camp al que s'aplica la regla el valor que configurem com codi de municipi erroni (o null) sempre i quan el camp sobre el que aplica la regla no tingui un codi INE de municipi vàlid.
REM_PROV INCE	La regla REM_PROVINCE ens permet assignar al camp al que s'aplica la regla el valor que configurem com codi de província erroni (o null) sempre i quan el camp sobre el que aplica la regla no tingui un codi INE de província vàlid
REM_EMAI L	La regla REM_EMAIL ens permet corregir el valor del camp configurat quan aquest camp és una adreça de correu electrònic i en cas de que no sigui reparable li assigna el valor que configurem a la regla com a valor de e-mail erroni.
REM_LPAD	La regla REM_LPAD ens permet reparar la longitud del camp afegint per la esquerra el caràcter que configurem fins arribar a la longitud configurada.
REM_RPAD	La regla REM_RPAD ens permet reparar la longitud del camp afegint per la dreta el caràcter que configurem fins a arribar a la longitud configurada.
REM_REJE CT_DNI	La regla REM_REJECT_DNI ens permet descartar registres on el camp amb el valor del DNI no sigui vàlid.

Exemple de Regles d'enriquiment



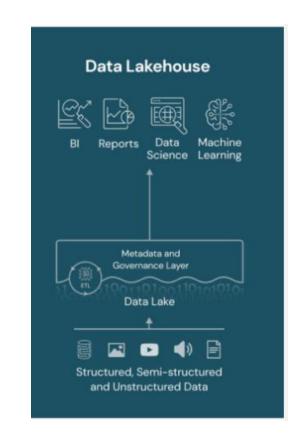
Regla	Descripció
REM_ENRICHMEN T_NORMALIZED_ ADDRESS	La regla REM_ENRICHMENT_NORMALIZED_ADDRESS és una regla d'enriquiment que ens permet normalitzar les adreces de l'origen/versió on s'aplica retornant cada adreça normalitzada amb els codis INE corresponents i els descriptius associats a la província, municipi, localitat i tipus via així com el nom oficial de la via
REM_ENRICHMEN T_SPLIT	La regla REM_ENRICHMENT_SPLIT és una regla d'enriquiment que ens permet trossejar el camp en partícules que es guarden sobre els camps de la pròpia taula que s'especifiquen al paràmetre "cols_to_split" de la regla

Emmagatzematge i analítica avaçada

Emmatzegament i dominis d'informació



- Basat en el concepte de *Data Lakehouse*, conforma una plataforma d'emmagatzematge i d'analítica unificada, mullti-cloud i independent del format o estructura de les dades.
- Dades estructurades: empra el Delta Lake de Databricks per a dades estructurades, oferint una capa d'emmagatzematge optimitzat, compatible amb Spark per a processaments batch i streaming, i estandaritzat amb fluxos en capes Bronze, Silver i Gold.
- Dades Semi/No-estructurades: utilitza MongoDB Atlas en versió SAS i Multicloud com a solució per a dades semiestructurades i no estructurades, oferint funcionalitats avançades que permet redimensionar la infraestructura segons l'ús, amb nodes en diversos proveïdors cloud, optimitzant costos i rendiment.
- Base de Dades Relacionals i graf: la PTD gestiona dades relacionals amb PostgreSQL per projectes que ho requereixin i Neo4j AuraDB per a dades en graf.

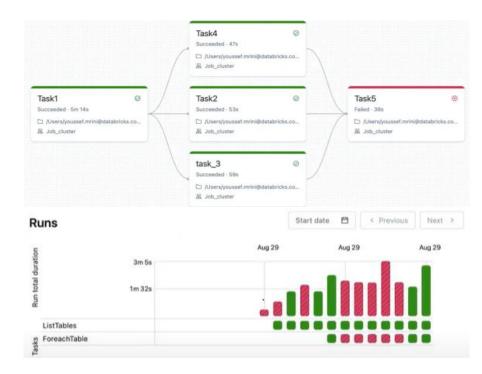


Execució de workflows i notebooks en processos de transformació i enriquiment

Transformació



- La PTD amb Databricks ofereix una plataforma integrada que utilitza processos Spark per gestionar el cicle de vida de les dades de manera eficient, permetent als usuaris executar, monitoritzar i ajustar processos de forma aïllada, i que facilita la gestió de tasques, modificació de paràmetres i visualització de logs.
- Capacitat d'orquestració avançada de tasques mitjançant workflows visuals, dotant a l'usuari de flexibilitat i autonomia en els seus procesos de transformació i enriquiment de les dades.
- Capacitat d'executar notebooks a la PTD, que permet als usuaris desenvolupar, executar i gestionar processos de dades en temps real de manera autònoma i col·laborativa, tot això sense la necessitat de desplegaments tradicionals.



Múltiples capacitats de publicació

Publicació

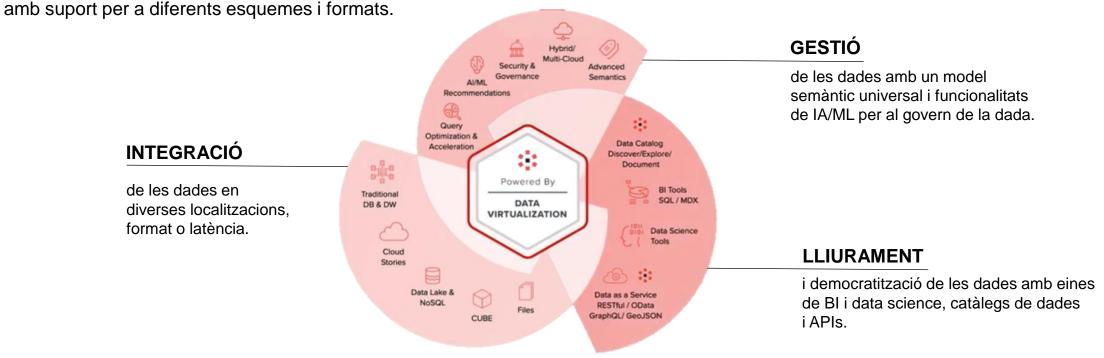


- **SFTP Corporatiu:** Serveis d'FTP segurs per compartir fitxers entre departaments i tercers, on-premise o en el cloud.
- Virtualització de dades: Denodo ofereix accés unificat a fonts heterogènies (SQL, NoSQL), amb centralització d'auditoria i seguretat.
- API Manager: Integració amb l'API Manager Corporatiu per a la publicació i consum de dades per tercers.
- Microserveis multicloud: Compatibles amb Kubernetes (AKS, EKS, GKS) per desplegar serveis en entorns multicloud.
- Connectors JDBC natius: Databricks proporciona conectors JDBC per a una àmplia varietat de bases de dades i aplicacions de tercers.
- **Data Sharing:** Delta Sharing de Databricks facilita la compartició segura de dades sense moure-les, compatible amb múltiples formats.



Virtualització

Denodo és una plataforma de virtualització de la dada escalable que permet connectar a qualsevol font de dades (bases de dades, fitxers, APIs, etc.), i combinar aquestes fonts per crear vistes unificades per anàlisis, informes i altres aplicacions. Denodo s'encarrega d'exposar en temps real les vistes i les dades a les eines externes mitjançant diverses interfícies com JDBC/ODBC o APIs



Virtualització – casos d'ús

Transició de dades BI cap a la PTD

Amb el virtualitzador, les explotacions de dades apunten cap a Denodo i pas cap a BBDD/DWH de l'aplicació. D'aquesta manera es pot iniciar la migració dels diferents repositoris de dades. sense haver de modificar les interfícies (MSTR, d'usuari PowerBI, ...), de manera que sigui Denodo qui vagi canviant les connexions de BBDD sense migrar segons es vagin tenint resultats, i això sigui transparent pels usuaris finals.

Visió unificada de diferents orígens

Denodo ofereix possibilitat de modelar i unificar de manera homogènia, en un únic lloc, l'accés a totes les dades disperses en les diferents bases de dades, sense importar on són els orígens. En un únic pas d'accés segur, Denodo permet a l'usuari realitzar les operacions típiques d' alta, baixa o modificació de les dades disperses en els diferents backoffice dels diferents orígens com si fos una única base de dades.

Historificació de dades

Utilitzant el virtualitzador de dades Denodo, la ubicació exacta de la dada a mostrar a l'usuari es indiferent per l'usuari, ja que és Denodo l'encarregat d'anar a buscar aquesta dada allà on es trobi (a la BBDD històrica o a l'actual).

Arquitectura alternativa de BI per aplicacions origen de dades amb molts canvis

Utilitzant de Denodo com a virtualitzador de dades, permet exposar el conjunt de dades que es vol explotar amb un baix cost, permetent fins i tot persistència de dades en cache, i que qualsevol canvi en els sistemes origen sigui fàcilment molt incorporat al repositori de dades a explotar i amb un cost i impacte molt menor.

Publicació de dades de la PTD

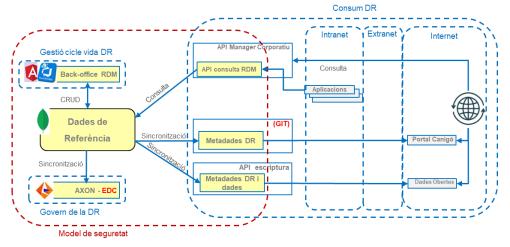
La publicació de dades a Denodo implica de vistes creació virtuals que representen les dades de manera lògica, la **configuració** de serveis d'accés per permetre als usuaris accedir a aquestes dades i l'aplicació de polítiques de seguretat per protegir integritat confidencialitat de dades. És possible posar a cache les dades, per tant, si utilitzem una API de pagament, amb una única crida ja tindríem les dades disponibles.

Dades de referència

Mòdul dades de referència



- Una dada de referència és informació estàndard compartida entre diferents aplicacions i departaments, i que assegura la consistència i la uniformitat en l'ús de dades dins l'administració, facilitant així la interoperabilitat i la coherència dels anàlisis transversals.
- El mòdul de RDM a la PTD, permetrà gestionar l'end-to-end de les dades de referència, des de la descoberta fins la publicació. Això s'aconsegueix industrialitzant el consum de dades de referència via l'API Manager corporatiu.
- El mòdul RDM permetrà recuperar el llistat de dades de referència i les seves diferents versions, així com el conjunt de canvis que s'han produïts en aquestes, i també es disposarà de la capacitat de registrar o anular la subscripció a una dada de referencia per les aplicacions.
- Les dades de referencia es publiquen al portal Canigó del CTTI per ús intern, i al portal Open Data a l'hora de donar visibilitat pública.



Les dades de referencia (DR) es composen de:

- Metadades: domini, subdomini, estat, propietari, persones de contacte, idiomes, frequència d'actualització, històric de versions, etc.
- **Contingut:** Ilista de valors amb diferents atributs definits per les metadades, com províncies, municipis, comarques, dades econòmiques, etc.
- Flux de gestió i governança: actors, rols, responsabilitats i tasques associades a cada etapa del cicle de vida de les DR (creació, modificació, publicació, autorització d'ús, subscripció, etc.).

Dades mestres

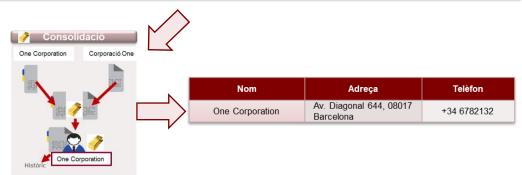
Mòdul dades mestres

MDM

- La PTD compta amb un sistema de gestió de Dades Mestres (MDM) que ofereix una visió única i fiable de les dades clau dels departaments, assegurant la coherència i qualitat mitjançant un sistema de validació abans de la seva integració. L'arquitectura de consolidació del MDM permet alimentar sistemes informacionals sense necessitat de sincronització ni afectació als processos operacionals.
- Etapa de puntuació (Scoring): Assigna un grau de confiança a cada dada segons la seva fiabilitat.
- Etapa de preparació (Staging): Els dades s'incorporen al sistema, preparant-se per a futures operacions.
- Etapa d'identificació: Es dona un identificador únic als registres per mantenir la coherència.
- Etapa de consolidació: Combina dades noves amb les existents per millorar la qualitat.
- Etapa de publicació: El Golden Record s'actualitza i elimina redundàncies.
- Etapa de disponibilitat: Facilita l'accés al Golden Record, incloent informació històrica.



Nom	Fiabilitat	Adreça	Fiabilitat	Telèfon	Fiabilitat	Sistema orgien
Corporació Una	70	Av. Diagonal 644, 08017 Barcelona	80	93 2803337	70	ERP
One Corporation	80	Av. Diagonal 644, Barcelona	75	+34 6782132	80	CRM
Corporació One	65	Diagonal 644	70	+34 6782132	75	Marketing



Aprovisionament de recursos

Monitorització de recursos

Finops

- FinOps permet mesurar i monitorar l'aprovisionament de recursos en una organització per a optimitzar l'ús dels mateixos de manera eficient facilitant la presa de decisions en col·laboració amb els equips d'enginyeria, finances i negoci.
- És possible monitorar qualsevol Cloud Service Provider (CSP), així com Platform as a Service (PaaS) homologats o no, a estàndard de FinOps Open Cost and Usage Specification (FOCUS). Alguns serveis que s'estan monitorant en el FinOps de la PTD actualment són: Azure, MongoDB, Databricks o Denodo. Amb la incorporació de nous serveis es crearà una línia de desenvolupament per a la seva incorporació en el flux de FinOps.
- Amb l'objectiu d'analitzar la informació, realitzar comparatives i evolucions històriques o analitzar els costos segons les dimensions disponibles (unitat organitzativa, projecte, servei, recurs, etc...) l'explotació de les dades es realitza mitjançant un dashboard en Microstrategy.



Resum de costos diari



Visió per centre de cost



Visió per projecte o servei



Visió històrica

5

Servei d'acompanyament (ACOPTD)

Servei d'acompanyament (ACOPTD)

Fases

Nou proveïdor



https://cstd-ctti.atlassian.net/browse/ACOPTD

Sessió de presentació de la PTD

- > Arquitectura de la PTD (mòduls i serveis)
- > Exemples de cas d'ús
- > Cicle de vida d'acompanyament durant un projecte
- > Dades que s'han de proporcionar per part del proveïdor
- > Perfils recomanats / coneixements necessaris

Suport a la configuració de l'entorn DES

> Plantilles de terraform + requisits repositori de codi

Nou projecte

Plantejament del projecte

> Requisits

Resum i aprovació de projecte

> Llum verda al projecte

Directius d'integració

> Específiques pel projecte

Productivitzacio del cas d'ús

> Evidències positives de les proves a pre-producció

Validació dels criteris d'acceptació per a desplegament a entorn pre-productiu

> Evidències i documentació requerides

Coordinació de govern de la dada

> Estàndards tècnics i necessitats particulars del cas



> Repàs del projecte



Departament de Polítiques Digitals i Administració Pública Centre de Telecomunicacions i Tecnologies de la Informació (CTTI)

www.gencat.cat

ptd.ctti@gencat.cat