

**ACORD MARC PER A LA PRESTACIÓ DEL SERVEI D'ASSESSORAMENT
EN DIVERSOS ÀMBITS DE LA GESTIÓ DE DADES A LA UNIVERSITAT
OBERTA DE CATALUNYA**

EXPEDIENT HSE00003/2023

PLEC DE PRESCRIPCIONS TÈCNIQUES

LOT 1 – GOVERN DE LA DADA I ARQUITECTURA DE LA INFORMACIÓ

1. CONTEXT	3
2. OBJECTE DEL PLEC	3
3. ABAST DEL SERVEI	4
4. ACORDS DE NIVELL DE SERVEI .	5
5. ORGANITZACIÓ DEL SERVEI I EQUIP DE TREBALL	5
5.1. 6	
5.2. 6	
6. LLOC DE REALITZACIÓ	7
7. RELACIÓ ENTRE LA UOC I L'ADJUDICATARI	8
 ANNEX 1 - Model de Dades UOC	 9
ANNEX 2 - Paradigma Data Mesh	10
Data Mesh i Domain Driven Design (DDD)	10
Principis del Data Mesh	11
Aproximació al Domain Driven Design (DDD)	11
ANNEX 3 - Programa de dades	15

1. CONTEXT

La Universitat Oberta de Catalunya (en endavant, “la UOC”) és una universitat innovadora, arrelada a Catalunya i oberta al món, que forma les persones al llarg de la vida contribuint al seu progrés i al de la societat, alhora que duu a terme recerca sobre la societat del coneixement. El seu model educatiu es basa en la personalització i l'acompanyament de l'estudiant mitjançant l'e-learning; en aquest sentit, es tracta de la primera universitat a la xarxa a l'estat espanyol.

La UOC vol ser una universitat que, connectada en xarxa amb la resta d'universitats del món, impulsa la construcció d'un espai global de coneixement i la recerca frontera en societat del coneixement. Innova en el model educatiu propi que se centra en l'estudiant, oferint una formació de qualitat i personalitzable, per fomentar la seva competitivitat i contribuir al progrés de la societat.

Per aconseguir aquests objectius, la UOC ha decidit apostar, entre altres coses, per una nova estratègia per a la gestió i govern de les dades la missió de les quals és la següent:

“Millorar l'experiència i l'èxit acadèmic i professional dels nostres estudiants, eficientar els processos organitzatius, impulsar la recerca i assegurar la presa de decisions informades mitjançant un ús innovador de les dades, que ajudi a habilitar l'estratègia de la UOC des de l'operativa diària.”

L'estratègia té 4 objectius principals:

- **Govern de la dada:** Proporcionar a la Universitat d'un sistema de decisions i responsabilitats al llarg de tot el cicle de gestió de les dades
- **Cultura:** Estendre una cultura de presa de decisions basada en les dades i acostar les dades a qui les pren, amb usos analítics cada vegada més avançats
- **Talent:** Augmentar el talent analític entre directius, tècnics i docents, i aprofitar l'experiència de recerca i innovació de la UOC en aquest àmbit
- **Diferenciació:** Crear productes, serveis i eines d'ajuda per enfortir el model docent i aconseguir nous avantatges competitius

Per aconseguir aquests objectius, s'han identificat 3 palanques habilitadores fonamentals:

- **Comunitat de pràctica de dades:** grup creat per referents de dades (analistes, tècnics, propietaris) a les àrees de negoci que seran els encarregats de promoure i implantar el resultat de la nova estratègia en aquestes.
- **Plataforma de dades:** plataforma tecnològica sobre la qual es crearà el model de dades UOC i que n'ha de permetre l'explotació adequada als diferents perfils.
- **Programa de dades:** assegura l'execució de l'estratègia mitjançant un pressupost, un backlog de projecte i un Comitè que pren decisions i prioritza.

A més, cal ressaltar que per a la implementació d'aquesta estratègia ens estem basant en el paradigma Data mesh, el qual fa èmfasi en la descentralització i l'autonomia del domini de dades. El gran canvi amb Data Mesh està en la gestió de dades com un conjunt de productes de dades, no com una col·lecció de processos i canalitzacions.

2. OBJECTE DEL PLEC

L'objecte del present Plec de Prescripcions Tècniques és establir els requisits, concretar els treballs i definir les obligacions que es deriven de l'Acord marc per a la selecció de diversos proveïdors per a la prestació del servei d'assessorament en l'àmbit del Govern de la Dada i Arquitectura de la Informació per a la Universitat Oberta de Catalunya ("UOC") i les empreses del seu Grup.

Les determinacions establertes en el present Plec de Prescripcions Tècniques de l'Acord Marc, així com les contingudes en el Plec de Clàusules Particulars, constitueixen normes vinculants per als proveïdors homologats, els quals realitzaran les prestacions dels contractes que se'n derivin amb expressa submissió als mateixos.

3. ABAST DEL SERVEI

Els treballs a realitzar consistiran en la prestació d'assessorament i execució sobre el Govern de la Dada i l'Arquitectura de la informació.

L'abast dels treballs que s'inclouen en el present lot comprenen, a títol enunciatiu i sense que sigui limitatiu, les següents àrees d'activitat:

- **Govern de la dada**

Serveis d'acompanyament i execució en la definició de polítiques i procediments consistents tant en l'assessorament basat ens els millors marcs de referència del Govern de la Dada com en l'estratègia de definició, recerca del consens, posada en marxa i publicació d'aquestes polítiques i procediments.

Exemples de polítiques i procediments:

- Qualitat de la informació
- Classificació de la informació
- Glossari i/o diccionari de dades
- Incorporació de nous dominis d'informació
- Incorporació de nous rols/perfils al marc de Govern establert a la Universitat

També es poden incloure treballs de prospecció d'eines tecnològiques que donin suport al Govern de Dades i gestió del cicle de vida de les dades (Glossaris, Traçabilitat de la informació, Qualitat de dades, Gestió dels rols inclosos en l'estratègia de Govern de la Dada de la UOC ...)

- **Arquitectura de la informació**

Serveis d'acompanyament i execució de projectes orientats a la co-creació i manteniment de l'arquitectura d'informació de la UOC basats en el paradigma DataMesh i utilitzant models d'informació com EduMesh, EDUAPI, HERM, DXTERA, HEDA, etc...

Exemple de tipologies de treballs a realitzar:

- Definició de dominis d'informació
- Definició dels rols de gestió del dominis
- Unificació semàntica entre dominis

- **Execució dels plans formatius**

Serveis d'execució de formacions expertes en l'àmbit del Govern de la Dada i Arquitectura de la Informació. Aquesta activitat inclou, no només l'execució de la formació sinó també l'elaboració dels materials formatius de suport corresponents.

Cal tenir en compte que l'abast del lot especificat en aquest Plec conté una relació orientativa de tasques a dur a terme i en cap cas exhaustiva dels serveis a prestar mitjançant les contractacions derivades.

Degut a la complexitat, varietat i peculiaritats inherents a la prestació dels serveis objecte del present Acord marc, i més concretament d'aquest lot, resulta impossible detallar de manera exhaustiva la totalitat dels serveis que han de ser objecte del present lot, així com tampoc el número de serveis a contractar, que es troben sotmesos a les necessitats de la UOC en cada moment determinat.

Per tant, forma part del present Acord marc la prestació de serveis relacionats amb l'objecte del present lot tot i que no es reflecteixin expressament en la descripció, sempre i quan tinguin relació directa amb aquests i siguin susceptibles d'ésser requerits per la UOC.

4. ACORDS DE NIVELL DE SERVEI.

Els terminis per a la prestació del servei són els que s'estableixen a continuació:

4.1. Un cop el contracte basat ha estat formalitzat amb l'empresa adjudicatària, el temps que aquesta està en disposició de poder iniciar els treballs. El termini computarà a partir del dia següent de la data de formalització del contracte.	7 dies laborables
4.2. Substitució dels perfils assignats a l'equip. El termini computarà a partir del dia següent de la constatació de la recepció del requeriment a l'empresa que està realitzant els treballs.	7 dies laborables

Els terminis anteriorment referits podran variar en funció de la tipologia de les tasques a realitzar, l'extensió, la complexitat o que existeixi un període preclusiu per l'acte que dona lloc a la prestació del servei. És per això que la UOC podrà informar del temps en què es requereix la resposta i el mateix podrà diferir dels enunciats en la taula anterior.

Els serveis es prestaran seguint les instruccions de l'Oficina de Dades de la UOC i, en particular, del/a responsable del contracte.

5. ORGANITZACIÓ DEL SERVEI I EQUIP DE TREBALL

Per a la realització dels diferents serveis que la UOC requereixi, el licitador haurà de proposar un equip de treball que asseguri l'assoliment dels objectius, mitjançant professionals amb experiència contrastada en la matèria objecte del plec i que compleixin com a mínim els requeriments establerts al PCAM.

5.1. Model de relació

En cada contracte basat de l'Acord Marc l'adjudicatari haurà de presentar, a nivell general, una estructura de governança necessària per a mantenir una interlocució permanent amb les parts involucrades. Aquesta estructura serà conjunta per a tots els basats de l'Acord Marc assignats a un mateix adjudicatari independentment del context al que pertanyin els contractes basats adjudicats.

A nivell de cada contracte basat, com a mínim, hi haurà d'haver el següent òrgan de govern:

- Comitè de Seguiment: assumirà les funcions de supervisió i control de l'execució del contracte així com la presa de decisions que afectin a l'objectiu i l'abast del contracte.

Les principals funcions del Comitè de Seguiment són:

- seguiment contractual del projecte (inici, aprovació de treballs, finalització, canvis d'abast) i gestió de conflictes.
- seguiment executiu del projectes. Canvis de planificació.
- assegurar l'adequació de l'equip d'acord a les necessitats de cada fase del projecte.
- validació de l'acabament correcte de les fites establertes en la planificació del projecte.
- proposar els canvis estratègics, econòmics, d'abast i de gestió que consideri oportuns per a la bona marxa del projecte i si s'escau, les modificacions contractuals que pertoquin.

El Comitè de Seguiment reunirà com a mínim al responsable de l'empresa i/o el responsable del servei de l'empresa adjudicatària i els representants que la UOC determini, amb la periodicitat que s'estableixi a l'inici del contracte basat, si bé es podrà convocar amb caràcter extraordinari sempre que es consideri necessari.

5.2. Equip de treball

Amb independència de l'estructura organitzativa proposada en cada contracte basat, l'equip de treball homologat en l'Acord Marc estarà compost pels següents perfils:

Perfil professional
Cap de projecte
Consultor sènior Govern de Dades
Arquitecte d'informació

Arquitecte de dades
Enginyer de dades

Els perfils i número de persones a assignar anirà en funció de la necessitat de cada servei que es contracti mitjançant un contracte basat i de les condicions i requisits que s'inclouen en les seves característiques.

La formació i l'experiència dels professionals proposats pels licitadors s'haurà d'acreditar d'acord amb el que preveu el Plec de Clàusules Particulars.

L'adjudicatari haurà de garantir la continuïtat de l'equip durant tot el termini d'execució dels treballs. Qualsevol canvi en els integrants de l'equip haurà d'ésser autoritzat prèviament per la UOC, que es concedirà sempre i quan el/la nou/va integrant de l'equip compleixi els requisits mínims de formació i d'experiència professional previstos en els plecs.

La UOC es reserva el dret a exigir el canvi de qualsevol membre de l'equip que no compleixi amb els requisits d'aquest plec o que no presti el servei de forma satisfactòria per a la UOC.

L'obligació de destinar o adscriure a l'execució del contracte els mitjans personals previstos en el present apartat tindrà la consideració d'obligació contractual essencial.

En cap cas existirà relació laboral entre els treballadors de l'empresa adjudicatària i la UOC.

6. LLOC DE REALITZACIÓ

Les activitats objecte dels contractes basats es podran desenvolupar, a criteri de la UOC, virtualment o presencialment a les instal·lacions o als edificis que la UOC consideri necessaris en cada cas.

Sens perjudici que les seus i/o centres de la UOC puguin augmentar o disminuir en el moment de la prestació dels serveis, actualment, i a títol informatiu, les seues i centres d'aquesta Fundació són principalment els següents:

- Seus d'Avinguda Tibidabo, 39-43, a la ciutat de Barcelona
- Seus del districte del 22@ de Barcelona (Rambla Poblenou, 156)

Adicionalment, pot ser necessària la prestació del servei a on la UOC té presència amb diferents Seus Territorials a tot el territori de Catalunya i les seues Territorials a nivell estatal que la UOC tingui en cada moment. Aquesta circumstància però, serà especificada prèviament en el contracte basat.

Els costos de qualsevol desplaçament a les dependències de la UOC del personal de l'empresa adjudicatària serà a càrrec de l'adjudicatari i amb mitjans de transport facilitats per ell mateix.

L'equipament i els recursos materials i tecnològics, i en especial aquells que l'adjudicatari disposi a les seves dependències per a l'execució del servei, serà a càrrec de l'adjudicatari. Els equips connectats a la xarxa de la UOC hauran d'ésser configurats d'acord amb els estàndards i les polítiques de la UOC.

Els costos de les comunicacions (veu i dades) i de les intercomunicacions entre els centres de treball de l'adjudicatari i la UOC, ocasionats per la prestació del servei, seran a càrrec de l'adjudicatari.

7. RELACIÓ ENTRE LA UOC I L'ADJUDICATARI

En el marc de les contractacions basades, l'adjudicatari designarà un/a representant, qui representarà a l'adjudicatari front a la UOC i disposarà en tot cas de les atribucions necessàries que li permetin implementar les instruccions rebudes per l'Oficina de Dades de la UOC i, en particular, del/a responsable del contracte.

Les relacions entre la UOC i l'adjudicatari es vehicularan a través del/a representant designat/da per l'adjudicatari com a Director/a del servei.

Periòdicament, i en el marc de les corresponents contractacions basades, s'efectuaran reunions en les quals el/a representant designat per l'adjudicatari aixecarà acta. En dita acta s'inclouran, en el seu cas, les noves tasques a realitzar per l'equip adscrit a l'execució del contracte derivat i reportarà l'avanç de les actuacions pendents.

El responsable del contracte podrà comunicar l'existència d'incoherències, mancances o defectes a la relació entre la UOC i l'adjudicatari. En aquest cas, l'adjudicatari haurà d'adoptar les mesures escaients per tal de donar una adequada solució.

[Jorge Fernández González](#)

Director de l'Oficina de Dades

ANNEX 1 - Model de Dades UOC

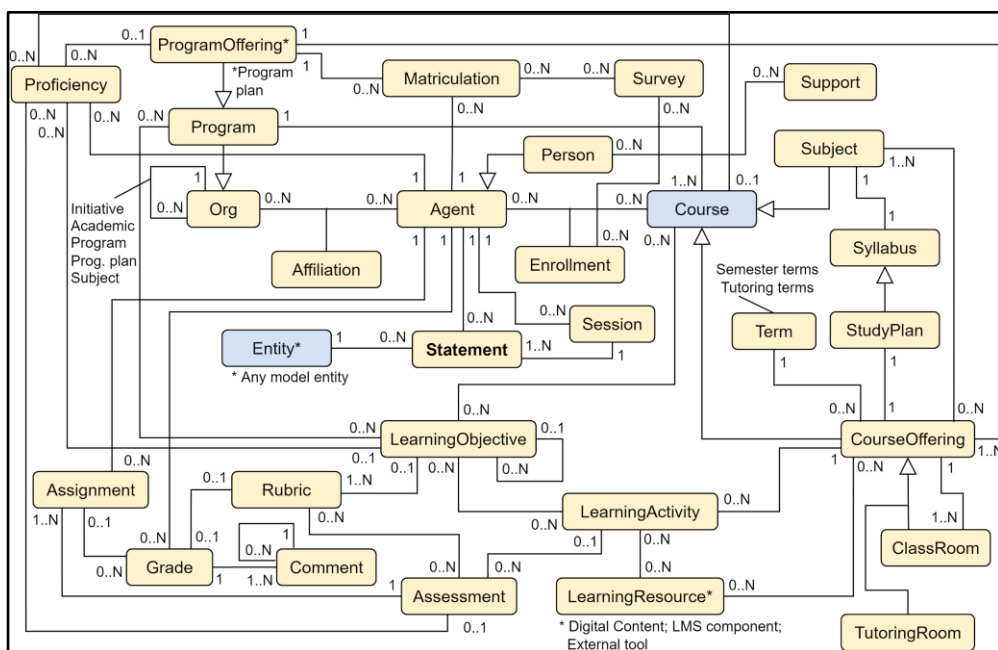
El model de dades UOC, té un enfoc cap a la interacció que tenen els usuaris amb la plataforma de Dades.

Això vol dir que és un model que s'obté en gran mesura en temps real (aprox.) i que conté les dades de la interacció dels usuaris (Estudiants, professors i gestió) amb el sistema i totes aquelles entitats que aporten dades complementàries a aquesta interacció. En tot cas, el model contemple un atribut freqüència i un atribut lastUpdate a totes les entitats que permeten conèixer la freqüència de refresc de les dades i la darrera data d'actualització.

La darrera versió del model publicada és la 1.0.3:

- v1.0.1: model que recull els requisits del vertical d'assignacions.
- v1.0.2: model ampliat amb els requisits del vertical tutoria.
- v1.0.3: model ampliat amb els requisits de seguiment de l'activitat a les Aules.

El següent diagrama defineix l'abast del model versió 1.0.3:



ANNEX 2 - Paradigma Data Mesh

La plataforma de dades de la UOC s'ha definit en base al paradigma **Data Mesh** i **Domain Driven Design (DDD)** i gestionat amb un model de **Data Governance (DG)** tracten les **Dades com a Producte (DaaP)**.

A continuació farem un repàs d'aquestes **conceptes**:

Data Mesh i Domain Driven Design (DDD)

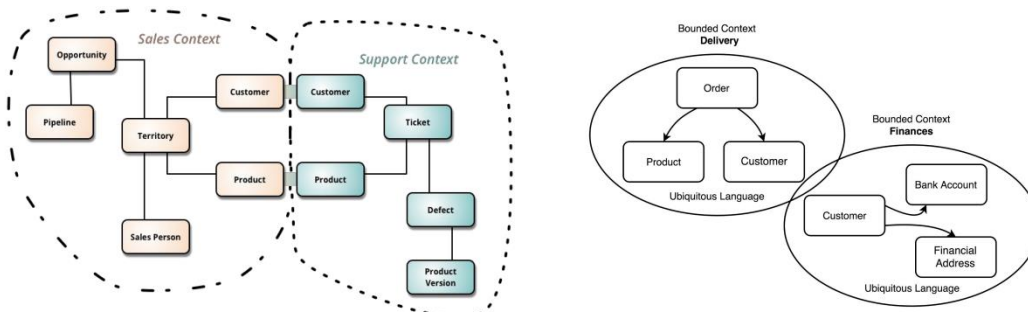
Moltes empreses estan invertint en el seu **data lake** de pròxima generació, amb l'esperança de

- **democratitzar les dades** a escala per proporcionar informació sobre el negoci i, finalment,
- **prendre decisions intel·ligents automatitzades**.

Les plataformes de dades basades en l'arquitectura del **data lake** tenen modes de fallada comuns que condueixen a promeses incomplertes a escala. Per abordar aquests modes de fallada hem de canviar del paradigma centralitzat d'un **data lake** o el seu predecessor **data warehouse**.

Aquest paradigma que anomenem **Data Mesh** és un **marc d'arquitectura** d'un model operatiu per al desenvolupament de **productes i aplicacions de dades basat en dominis** que inclou la ubicuïtat de les dades a l'empresa aprofitant:

- un disseny **d'autoservei orientat al domini (en una perspectiva de desenvolupament de programari) i**
- **pren prestada la teoria d' Eric Evans del domain-driven design:**
 - **Alinear el disseny amb el domini, les necessitats i l'estratègia de l'empresa.**
 - **Propietat descentralitzada** de les dades **orientada al domini i l'arquitectura descentralitzada**.
 - Descentralitzar la propietat de compartir dades analítiques als dominis empresarials més propers a les dades, **publicats** per la **font de les dades** o els seus principals **consumidors**.



La proposta principal és que els **arquitectes de dades** haurien de **crear conjunts de dades distribuïts**, o **data meshes**, que siguin gestionades per un grup particular d'experts humans.

Aquests equips de domini saben com dissenyar les seves dades en tipus de dades i formats de fitxers que:

- **satisfan les necessitats** dels **consumidors** de dades de tota una organització i
- **alineats amb la estratègia de l'organització**.

Considerant això, **Data Mesh no és només tecnologia, també és una cultura.**

Principis del Data Mesh

Data mesh es defineix per 4 principis:

- **Arquitectura i propietat de dades descentralitzada orientada al domini:** Perquè l'ecosistema que creï i consumeixi dades pugui escalar a mesura que augmenta:
 - la quantitat de fonts de dades
 - la quantitat de casos d'ús
 - la diversitat de models d'accés a les dadesi simplement augmenti els nodes autònoms a la malla (Mesh).
- **Dades com a producte:** Perquè els usuaris de dades puguin descobrir, comprendre i utilitzar de manera segura dades d'alta qualitat amb una experiència agradable; dades que es distribueixen a molts dominis.
- **Infraestructura de dades d'autoservei com a plataforma:** Perquè els equips de domini puguin crear i consumir productes de dades de manera autònoma utilitzant les abstraccions de la plataforma, ocultant la complexitat de construir, executar i mantenir productes de dades segurs i interoperables. El **Data mesh** es basa en la gestió i l'operació de dades mitjançant una plataforma compartida i un conjunt d'eines i serveis que els equips dels dominis de negoci puguin utilitzar.
- **Governança computacional federada:** Perquè els usuaris de dades puguin obtenir valor de l'agregació i la correlació de productes de dades independents: la malla (**Mesh**) es comporta com un ecosistema seguint els estàndards d'interoperabilitat global, que compleixi les normes organitzatives i les regulacions del sector

Aproximació al Domain Driven Design (DDD)

Ara veurem els criteris que podem seguir a l'hora de definir un Domini o un altre. Domain Driven Design Approaches:

Source oriented domain data

Alguns dominis s'alineen de manera natural amb la font, d'on s'originen les dades.

Els conjunts de dades del domini font representen els fets i la realitat del negoci. **Els conjunts de dades del domini d'origen capturen les dades que es mapegen molt a prop del que generen els sistemes operatius del seu origen, els sistemes de la realitat.**

En una situació madura i ideal, un sistema operacional i el seu equip o unitat organitzativa, no només són responsables de:

- proporcionar capacitats empresarials, sinó que també són responsables de
- proporcionar les veritats del seu domini empresarial com a conjunts de dades del domini font.

A escala empresarial, **mai no hi ha un mapa un a un entre un concepte de domini i un sistema font. Sovint hi ha molts sistemes que poden servir parts de les dades que pertanyen a un domini**, alguns heretats i altres fàcils de canviar. Per tant, pot haver-hi molts conjunts de dades alineats a la font, també coneguts com conjunts de dades de realitat que finalment s'han d'agregar a un conjunt de dades cohesionat alineat amb domini.

S'ha de tenir en compte que els conjunts de dades del domini font representen de prop les dades en brut en el punt de creació i no s'ajusten ni es modelen per a un consumidor concret.

Aggregate data domains

Un agregat és un conjunt d'objectes de domini que poden ser tractats com una sola unitat. Un objecte de domini és qualsevol objecte que defineix un concepte del nostre domini de negoci.

Suposem que volem modelar motocicletes, per a nosaltres un objecte de domini podria ser el Motor o una Roda de la nostra motocicleta: l'objecte Motocicleta en si seria el nostre agregat.

Això ens permet definir certes invariants del nostre model com:

- Una motocicleta ha de tenir exactament dues rodes.
- Una motocicleta ha de tindre un motor.
- El motor pot ser dièsel o benzina.

Qualsevol referència de fora de l'agregat només hauria d'anar a l'arrel de l'agregat. Així, l'arrel pot assegurar la integritat de l'agregat en el seu conjunt.

Fent referència a l'exemple motocicleta, implicaria que si volem comunicar-nos amb l'objecte Motor ho hem de fer usant l'objecte Motocicleta com a intermediari. Sota cap circumstància podrem accedir directament al Motor. L'agregat actua com a interfície dels altres objectes de domini, exposant el comportament permès per als objectes interns (Llei de Demeter). Això permet treballar amb la Motocicleta com un "tot" en comptes de tractar els objectes Motocicleta i Motor com a entitats aïllades, mantenint així la integritat del conjunt d'objectes. Per tant, si el nostre agregat Motocicleta no exposa un mètode per desmuntar una Roda, seria impossible fer-ho, cosa que garantiria que la nostra Motocicleta sempre tindrà dues rodes.

Els agregats en tot el seu conjunt són els elements que s'emmagatzemen al sistema de persistència.

Quan volem carregar o desar una dada, hem de fer servir l'agregat complet. És a dir, quan volem persistir el nostre objecte Motocicleta ho farem persistint l'objecte sencer, no persistirem mai les rodes, el motor i els altres objectes de la motocicleta de manera separada.

No crear ambiciosos dominis de dades agregades: dominis de dades agregats que intenten capturar totes les facetes d'un concepte particular com a 'client 360'. Aquest agregat pot tornar massa complex i difícil de manejar per administrar i difícil d'entendre i utilitzar per a qualsevol cas d'ús en particular.

Els agregats modelen els conceptes principals del nostre domini, salvaguarda la integritat dels altres objectes directament relacionats amb ell i serveix d'interfície a l'hora d'interactuar-hi. Són un dels pilars del DDD, definint-los correctament podem modelar i encapsular la lògica del nostre negoci de manera efectiva.

Consumer oriented and shared domain data

Alguns dominis **s'alineen estretament amb el consum**. Els conjunts de dades del domini del consumidor i els equips que els posseeixen tenen com a objectiu satisfer un grup de casos d'ús estretament relacionat.

Els conjunts de dades de dominis alineats amb el **consumidor** tenen una naturalesa diferent en comparació amb els conjunts de dades de dominis d'origen. **Passen estructuralment per més canvis i transformen els esdeveniments del domini d'origen per agregar vistes i estructures que s'ajusten a un model d'accés concret.**

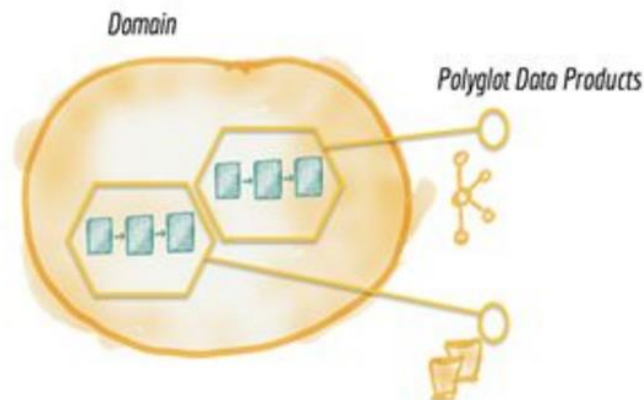
Producte de Dades (Data as a Product)

Tal i com hem vista abans, un dels principis del paradigma de la **Data Mesh** és considerar les **dades com un producte**. Si ens submergim en el món del **Data Mesh**, aquesta cita de l'article original de Zhamak Dehghani és clau per entendre la definició de **Data as a product**: *"Els equips de dades de cada domini han d'aplicar el pensament del producte [...] als conjunts de dades que proporcionen; considerant els seus actius de dades com els seus productes i la resta de científics de dades de l'organització, ML i enginyers de dades com els seus clients"*.

Cada font té el seu propi gestor/propietari de productes de dades (que formen part d'un equip multifuncional d'enginyers de dades) i és el seu propi domini clarament enfocat que té una oferta autònoma, convertint-se en els components fonamentals d'una malla (**Mesh**), donant lloc a un **Arquitectura distribuïda basada en dominis**. S'ha de tenir en compte que per raons de rendiment, es pot tenir un domini que agrupi dades de diverses fonts. Cada domini ha de ser:

- **Detectable** (Data Catalog Centralitzat i amb llenguatge ubic),
- **Adreçable** (minimitzar la fricció entre cerca y accés),
- **Autodescriptiu** (semàntica homogènia),
- **segur** (governat pel control d'accés global),
- **fiable** i
- **interoperable** (governat per un estàndard obert: dominis correlacionats)

Cada domini emmagatzemarà les seves dades en un llac de dades i, en molts casos, també tindrà una còpia d'algunes de les dades en una base de dades relacional.

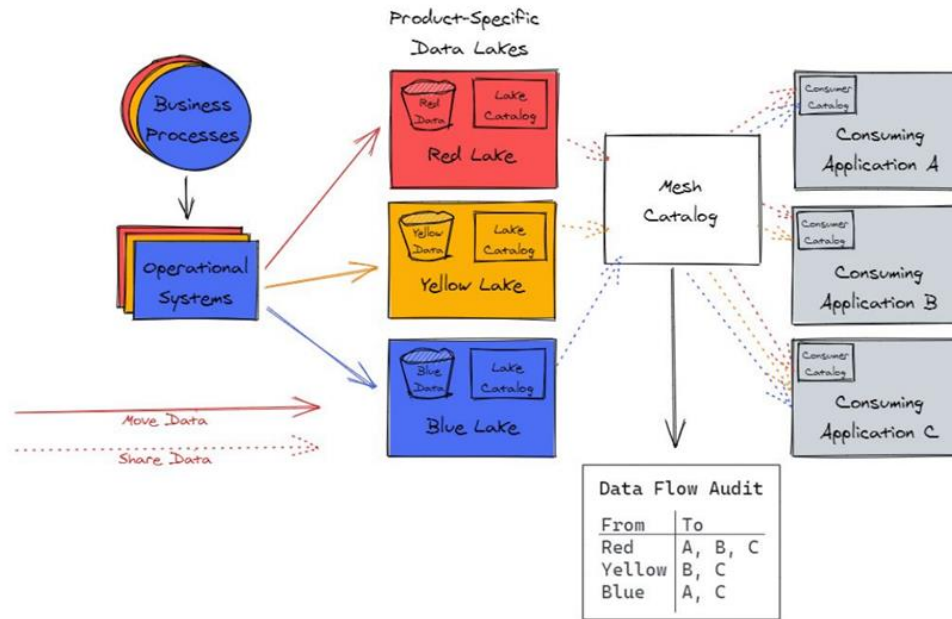


Els productes de dades **han de tenir**:

- **Xifratge** de dades en repòs i en moviment
- **Versions** de productes de dades
- **Esquema** del producte de dades
- **Governança i estandardització** de dades
- **Llinatge** de producció de dades
- **Supervisió, alertes i registre** de productes de dades
- **Mètriques de qualitat** del producte de dades

del Pull al Push

Per descentralitzar la plataforma de dades monolítica, necessitem revertir la forma en què pensem sobre les dades, la seva localitat i propietat. En lloc de fer fluir les dades dels dominis a una plataforma o Data Lake de propietat central, els dominis han d'allotjar i servir els seus conjunts de dades de domini d'una manera fàcilment consumible.



ANNEX 3 - Programa de dades

La UOC ha posat en marxa un **Programa de Dades**, el qual està conformat per la **Oficina de Dades i Arquitectura i Gestió de Dades**.

Els equips del programa de dades queden integrats dins de la Plataforma Informacional. El qual, també integra al departament d'**Arquitectura Tecnològica i Seguretat**.

Funcions i Responsabilitats

Dividim el Programa de Dades en tres línees: Estratègica, Gestió i Operativa.

Estratègic - Comité de programa de dades

Establir les prioritats i l'estratègia de la Universitat dins del programa de Dades. Aquest Comitè és l'sponsor principal i el responsable de l'èxit del programa.

El Comitè de Programa de Dades és responsable de proporcionar orientació a nivell executiu al programa, així com de promoure la governança de dades a tota la Universitat. També és el responsable de determinar la prioritat i el pressupost dels principals projectes relacionats amb les dades.

Gestió - Comité de Govern de Dades

El Comitè de Govern de Dades és l'encarregat del lideratge entre els diferents departaments i els àmbits de dades governats, ofereix una direcció i supervisió quotidiana de la funció de govern de dades.

El Comitè, establirà les definicions clares dels àmbits de dades, aprovarà les polítiques de govern de dades i supervisarà la documentació i processos que les unitats de negoci utilitzaran dins del seu àmbit d'actuació.

Les decisions de dades entre àmbits seran revisades i aprovades dins del propi Comitè. CGD resoldrà els problemes que no es puguin resoldre a nivells inferiors.

Operatiu - Comunitat de pràctica de dades

La Comunitat de Pràctica de Dades és un grup de treball col·laboratiu dels Tècnics de Dades. Cada Tècnic de Dades representa un àmbit de dades en concret dins del departament corresponent.

Aquesta comunitat compartirà els mètodes de treball per tal d'obtenir un catàleg de bones pràctiques, revisió dels processos de treball i confirmar la correcta execució de les polítiques definides dins del Govern de Dades

Adicionalment, han de vetllar perquè l'esforç de govern de dades s'ajusti als objectius del negoci. A més, la comunitat ha de col·laborar amb totes les parts interessades per assegurar-se que els problemes de metadades i dades consideren les necessitats de tots els usuaris de dades.

Rols i funcions

Els rols identificats i funcions són els següents.

Analista de dades

Dissenyar, implementar i recolzar les transformacions i visualitzacions necessàries per transmetre d'una manera òptima i comprensible la informació que s'obté de les diferents fonts de dades disponibles.

Científic de dades

El científic de dades és l'encarregat d'assessorar al negoci sobre el potencial de les dades, de proporcionar informació nova mitjançant l'ús d'anàlisis estadístiques avançades, mineria de

dades i tècniques de visualització de dades, per crear solucions que permetin millorar el rendiment.

Arquitecte d'informació

Defineixen els dominis semàntics i la seva relació per satisfer les necessitats d'informació en el context del negoci.

Asseguren la coherència de la informació entre el producte de dades conceptual, els diccionaris de negoci, el reporting definit i el conjunt d'indicadors necessaris per al seguiment de l'activitat del negoci, des d'una perspectiva transversal a les diferents àrees.

Recolzaran a les àrees de negoci i els equips de tecnologia.

Data Product Architect

Disseny i manté l'arquitectura tècnica de dades.

Referència en el disseny i la implementació tècnica dels productes de dades.

Asseguren que l'arquitectura tècnica definida dóna suport als requeriments funcionals de dades de la Universitat i asseguren la custòdia, el transport, la trazabilidad, la interoperabilitat i l'emmagatzematge segur de les dades del data product.

Recomanen les eines adequades per millorar la fiabilitat, eficiència i qualitat de les dades.

Tècnic de Govern i Qualitat

Serà l'expert encarregat de determinar i implantar controls de qualitat de dades que garanteixin la seva fiabilitat, així com la definició dels plans de millora més adequats. Treballaran molt estretament amb tots els àmbits de dades definides i hauran de vetllar l'aplicació dels requeriments normatius.

Tècnic de dades

Té la responsabilitat de la gestió diària de les dades. Són experts en l'àmbit de dades assignat i recolzen a la resta de consumidors d'informació així com les activitats de reporting.

Gestionen la demanda de dades de la seva àrea o grup de recerca. Són els responsables de treballar amb els altres Tècnics de Dades de tota la universitat i realitzen tasques de representació dels propietaris de Dades.



ANALISTA DE DADES

CIENTÍFIC DE DADES



ARQUITECTE D'INFORMACIÓ

DATA PRODUCT ARCHITECT



DATA OWNER

TÈCNIC DE GOVERN I
QUALITAT

TÈCNIC DE DADES

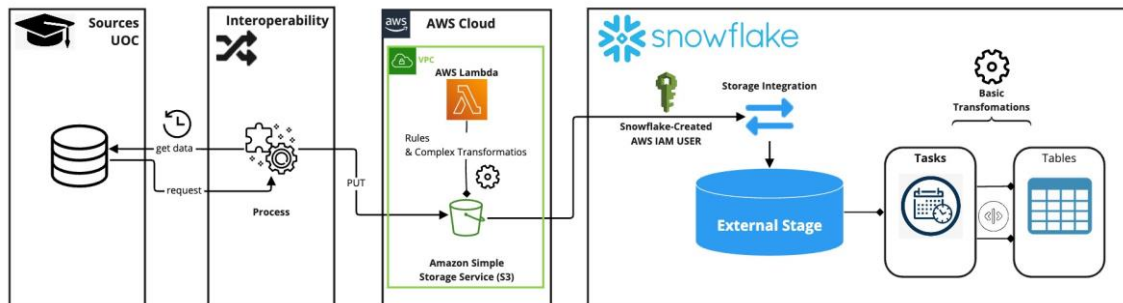
Metodologia

Arquitectures

A continuació proposem una arquitectura que implementi un pipeline de processament de dades. Basada en una solució serverless i que escales automàticament sota demanda.

Es planteja una arquitectura modular, que ens permet anar afegint components a mesura que siguin necessaris.

Els nous desenvolupaments deurien adaptar-se a l'arquitectura aquí descrita, tenint en compte que l'estratègia per origen pot canviar, una proposta seria aquesta.



Les transformacions es plantegen en dos punts de la ELT:

- Regles de transformació complexes: Un cop la informació és al bucket S3 destí, implementant Lambdes que realitzin les transformacions necessàries i tornin a deixar la info al bucket. Segons el tipus de transformacions que es requereixin, s'haurien de valorar altres serveis AWS diferents a les Lambdes com alternativa.
- Transformacions Bàsiques: En les tasques de càrrega d'Snowflake -COPY INTO- <https://docs.snowflake.com/en/user-guide/data-load-transform.html>
- Les transformacions a partir de informació ja carregada a Snowflake, s'implementen en el propi Snowflake amb Tasks o Stored Procedures en funció de les decisions preses en el disseny de cada aplicació.

Aquesta arquitectura no implementa real time, es tracta d'una arquitectura pensada per càrregues de dades programades cada cert temps.

Abans de plantejar una arquitectura en Real Time és molt important entendre que tindrà uns costos d'infra elevats i que no té sentit si no hi han actors involucrats en consumir aquestes dades constantment.

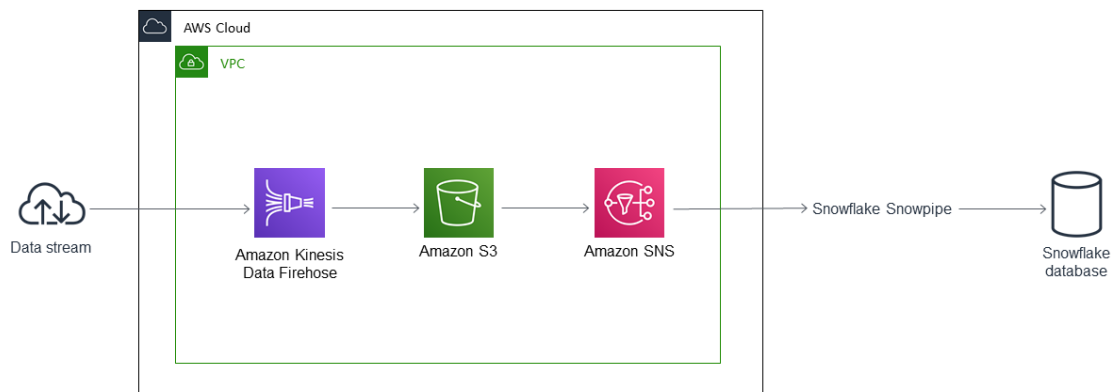
Real Time

Un dels requeriments comuns a l'hora de crear una ETL es el real time. A continuació detallarem com podem utilitzar els serveis al núvol d'Amazon Web Services (**AWS**) per processar un flux continu de dades i carregar-los en una base de dades de **Snowflake**.

La UOC ja disposa d'un compte d'aws (UOCIT_DADES) i snowflake actius, per tant tindrem que implementar:

- Un origen de dades que envii dades continuament (near real time)
- Un Kinesis Data Firehose per rebre aquestes dades
- Un bucket d'S3 que rebí les dades de la seqüència de rebuda de Kinesis Data Firehose

Snowflake no es connecta directament amb Kinesis Data Firehose. Per tant l'arquitectura quedaria així:



El patró utilitza Amazon Kinesis Data Firehose per enviar les dades a Amazon Simple Storage Service (Amazon S3), Amazon Simple Notification Service (Amazon SNS) per enviar notificacions quan es reben noves dades i Snowflake Snowpipe per carregar les dades en una base de dades de Snowflake.

Kinesis Data Firehose és un servei completament administrat per a l'enviament de dades de streaming en temps real a destinacions com ara Amazon S3, Amazon Redshift, Amazon OpenSearch Service, Splunk i qualsevol punt d'enllaç HTTP o punts d'enllaç HTTP personalitzats propietat de proveïdors de serveis de tercers proveïdors de serveis compatibles.

Amazon Simple Notification Service (Amazon SNS) coordina i administra el lliurament o l'enviament de missatges als punts d'enllaç o als clients subscrits.

Aquesta és l'arquitectura que recomana Snowflake per integrar-se amb AWS quan és necessari Real time. Però es té que tenir molt en compte que els PUTs en S3 no són la opció més econòmica i que, de fet, **no ho recomanem pel seu elevat cost**.