

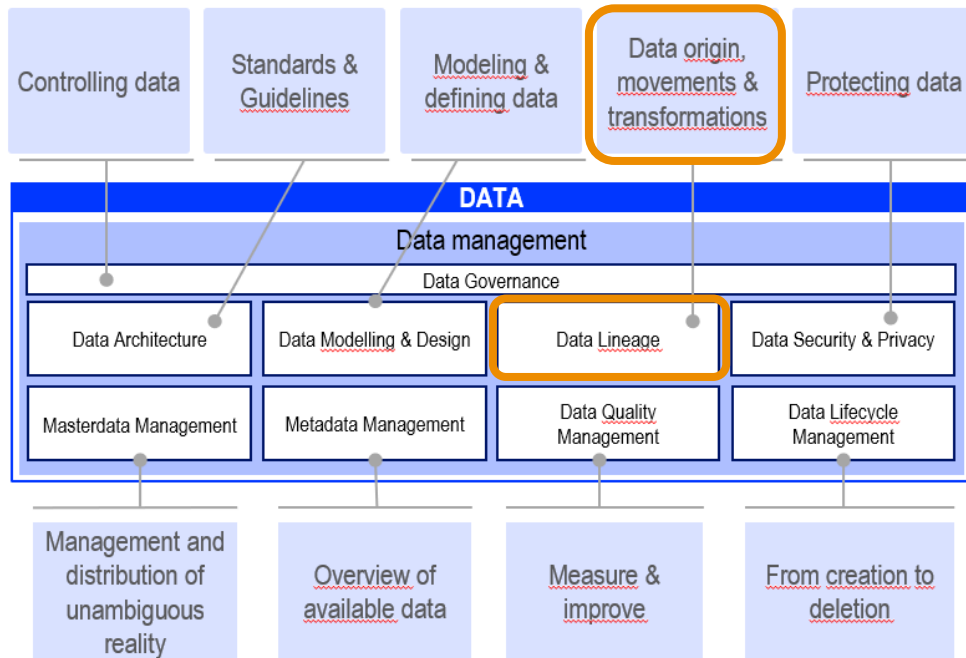


Data Lineage

Verkenning Data Capability v0.1

Data Lineage

Eén van de data capabilities in de Enterprise Architectuur



Capability	Description
Data Governance	Provides direction and oversight for data management by establishing a system of decision rights over data that accounts for the needs of the enterprise.
Data Architecture	Defines the blueprint for managing data assets by aligning with organizational strategy to establish strategic data requirements and designs to meet these requirements.
Data Modeling & Design	The process of discovering, analyzing, representing and communicating data requirements in a precise form called the data model.
Data Lineage	Specifies the data's origins and where it moves over time. It also describes what happens to data as it passes through diverse processes.
Data Lifecycle Management	Includes planning, implementation and control activities used to manage the lifecycle of data and information found in a range of unstructured media.
Security & Privacy	Ensures that data privacy and confidentiality are maintained, that data is not breached and that data is accessed appropriately.
Master Data Management	The ongoing reconciliation and maintenance of critical shared data to enable consistent use across systems of the most accurate, timely and relevant version of truth about essential business entities.
Metadata Management	The planning, implementation and control activities to enable access to high quality, integrated Metadata, including definitions, models, data flows and other information critical to understanding data and the systems through which it is created, maintained and accessed.
Data Quality	The planning and implementation of quality management techniques to measure, assess and improve the fitness of data for use within an organization.

Data Lineage

Een wirwar van gerelateerde termen

Wie in de wereld van Data Lineage duikt komt er al snel achter dat er veel termen rondgaan die op elkaar lijken of in ieder geval sterke raakvlakken met elkaar hebben. Daarnaast worden die termen regelmatig gebruikt alsof ze synoniem zijn.



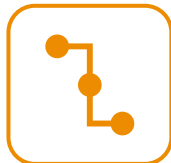
Data Lineage: *Specifies the data's origins and where it moves over time. It also describes what happens to data as it passes through diverse processes.* [Gartner]

In definities uit andere bronnen wordt de proceskant vaak niet zo nadrukkelijk benoemd, of weggelaten. Sommige bronnen benoemen alleen de 'origin' kant van de definitie. De Gartner definitie lijkt de meest complete te zijn.



Data Chain / Data Lineage: *pathway along which it [data] moves from its point of origin to its point of usage.* [DMBOK]

De DMBOK geeft dezelfde definitie aan Data Lineage en stelt deze twee concepten dus gelijk. Hier valt op dat de definitie de relatie met processen niet expliciet benoemt, zoals de Gartner definitie wel doet.



Data Flow:

1. *Data flows are a type of Data Lineage documentation that depicts how data moves through business processes and systems.* [DMBOK]
2. *A movement of data from one point to another, without involving any intermediaries at a specific level of granularity, to transport data.* [EDM Council]

'Data flow' is een breed gebruikt begrip dat regelmatig synoniem gebruikt wordt met Data Lineage. In de eerste definitie wordt de data flow als product op zichzelf gezien. In de tweede definitie wordt verwezen naar het concept data flow zoals je dat ook terugziet in data flow diagrammen. In deze context betekent 'without intermediaries': zonder tussenliggende data stores of processen.

Data Lineage

Een wirwar van gerelateerde termen



Data Provenance: *Documentation of the source of data with sufficient detail to allow reproducibility of a specific dataset. [EDM Council]*

Provenance vertaalt zich letterlijk naar 'herkomst', waar lineage zich vertaalt naar 'afstamming'. Het is dan ook niet vreemd dat Data Provenance vaak synoniem gebruikt wordt met Data Lineage. Er bestaat echter ook een specifiekere definitie, die hierboven genoemd wordt.



Data Lifecycle: *[the data lifecycle] includes processes that create or obtain data, those that move, transform and store it and enable it to be maintained and shared, and those that use or apply it, as well as those that dispose of it. [DMBOK]*

De processen die benoemd worden als onderdeel van Data Lifecycle Management hebben invloed op Data Lineage.



Data Traceability: *The ability to track a data construct back to the construct it was derived from as a more concrete instantiation. [EDM Council]*

Deze definitie duidt op de 'doorvertaling' van conceptuele objecten naar logische objecten & attributen en van logische objecten & attributen naar fysieke objecten & attributen. Deze definitie en uitleg bij Data Traceability zie je vaak ook bij Data Lineage.

Voorstel voor een definitie – in context

- ✓ **Data Lineage** als capability gaat over het in staat zijn om te duiden waar data vandaan komt, waar het naartoe beweegt, in welke processen het verwerkt wordt en welke bewerkingen het daarbij ondergaat. Het gaat dus om de totstandkoming van een bepaald *inzicht*. Dat inzicht kan ontstaan middels het creëren van een product, vaak **Data Flow** genoemd. In die Data Flow zien we processen die ook onderdeel zijn van de **Data Lifecycle**. **Data Traceability** gaat over een ander type inzicht dat niet gaat over beweging van data of processen die data verwerken.

Data Lineage

Wat hebben we eraan? Enkele voorbeelden

“If an organization does not know where data came from and how it has changed as it has moved between systems, then the organization cannot prove that the data represents what they claim it represents” – DMBOK2

- **Data eigenaren** kunnen hun data beter besturen omdat ze weten waar het naartoe stroomt en waar het gebruikt wordt.
- **Agile maakteams** hebben beter inzicht in welke data waar beschikbaar is, wat ermee gebeurd is voordat ze het gaan gebruiken en kunnen beter inschatten welke invloed features en user stories hebben op data afnemers in de keten.
- **BI teams** zijn beter in staat om de juistheid van rapportages te garanderen en hebben zicht op afhankelijkheden die daar invloed op hebben.
- **Data scientists** kunnen beter inschatten of data bruikbaar is voor hun projecten, en kunnen gericht modelleren en bewerken.
- **Privacy Office, Audit en Risk Management** kunnen risico's beter inschatten en compliancy eenvoudiger toetsen.
- **Architectuur**

Data Lineage

Samenhang met de andere data capabilities

- De Data Lineage capability is in veel gevallen ondersteunend aan andere data capabilities.

Data Capability	Relatie met Data Lineage
Data Governance	▪ Data Governance vereist Data Lineage om data effectief te kunnen besturen en alignment tussen data eigenaren te realiseren
Data Modeling & Design	▪ Data Modeling & Design betreft het modelleren en definiëren van de data die door het landschap stroomt
Data Lifecycle Management	▪ Data Lineage geeft inzicht in de plekken in het landschap waar Data Lifecycle Management ingericht zou kunnen of moeten worden
Security & Privacy	▪ Data Lineage geeft inzicht in waar welke data zich bevindt; dit helpt om die data effectief te kunnen beschermen
Data Integration	▪ Realiseert de beweging van data tussen systemen
Master Data Management	▪ Master Data Management realiseert ontwerp en sturing van de Data Lineage van Masterdata
Data Warehousing & Business Intelligence	▪ Data Lineage kan helpen bij het onder controle brengen van de datastromen naar DWH's en juistheid van rapportages te bepalen
Data Analytics	▪ Kennis over de oorsprong van en reeds gedane bewerkingen op data verhoogt de effectiviteit en nauwkeurigheid van Data Analytics projecten
Metadata Management	▪ Lineage metadata is een component van Metadata Management
Data Quality	▪ Data Lineage kan helpen om de oorzaken van Data Quality issues vast te stellen en effectief te kunnen verhelpen

Data Lineage

Positionering i.r.t. Metadata Management



- In de markt wordt Data Lineage vrijwel altijd gezien als onderdeel van Metadata Management.
- Gartner ziet Data Lineage als één van de critical capabilities¹ van metadata management solutions.

Data lineage can help with efforts to analyze how information is used and to track key bits of information that serve a particular purpose. Any good metadata management program will manage this insight and expose it to users.

Data Lineage als metadata capability...

... 'supports the process of examining complex data pipelines for analytics and data science use cases, where scale, unstructured data, long runtime and complex platforms are the key computing challenges. This gives data scientists and analysts a clearer understanding of their data and its veracity'

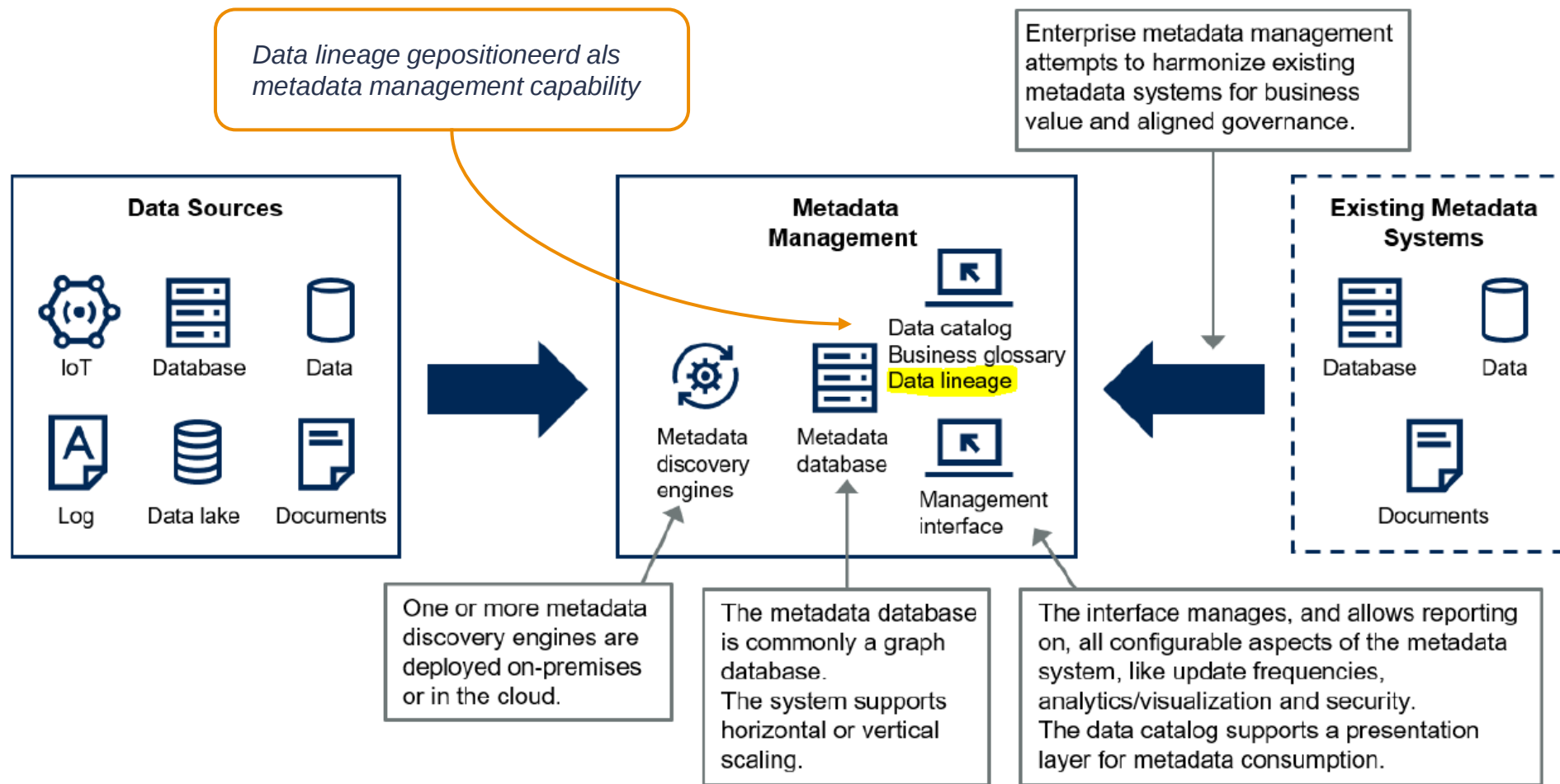
... is een 'enabler for the acceleration of business decision making, offering better understanding of information at hand, along with trust and confidence in information accuracy and consistency'

... 'helps to reduce costs and time by identifying redundancies in data integration processes — different processes updating the same attributes or tables, or unused attribute'

Data Lineage

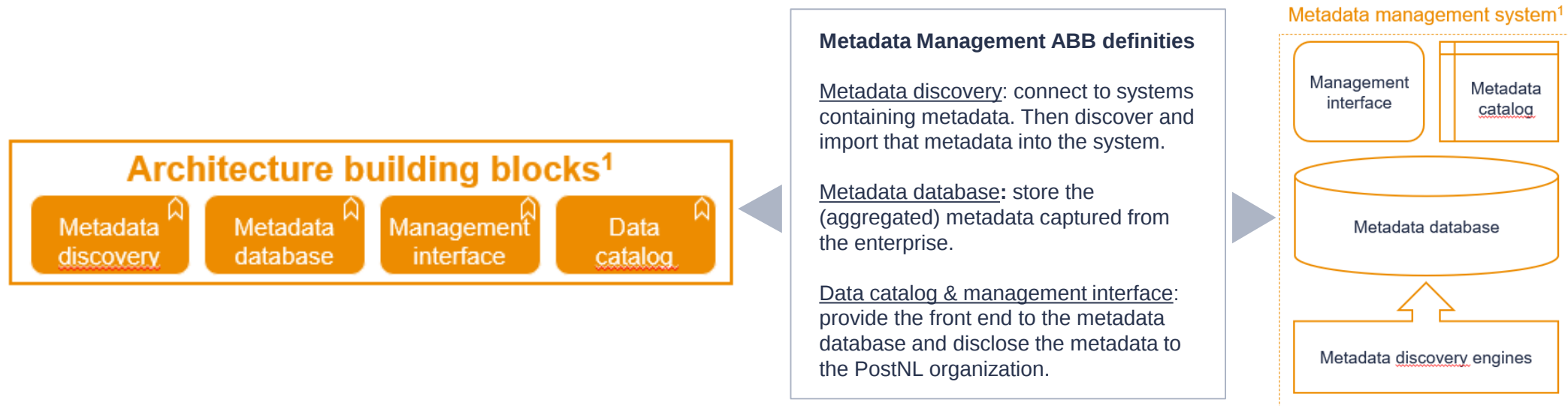
Positioning i.r.t. Metadata Management

Gartner



Data Lineage

Positionering i.r.t. Metadata Management



- In de Referentie Architectuur Metadata Management gelden bovenstaande ABB's, en die gelden dan ook voor Lineage metadata.
- Data Lineage wordt niet uitgebreid uitgewerkt in deze referentie architectuur, maar er wordt hier nog wel een belangrijk verschil benoemt tussen een 'design lineage' (design time data lineage) en 'real time lineage' (runtime data lineage).

Feature	Purpose	Example
<u>Data Lineage</u>	▪ Design lineage displays data flow in a design mode ▪ Real-time lineage based on history of data movement to capture real-time changes for historical reference ▪ Impact analysis of data interactions & prototyping of data movement scenarios	▪ Lineage relates to data movement ▪ Design lineage is used in architecture ▪ Real-time lineage is used in operations, in troubleshooting or for feedback to a design Model ▪ Generated model of data impacts ▪ Lineage map over time

Data Lineage

Positionering i.r.t. Metadata Management

- Een Metadata Management solution hoeft niet randvoorwaardelijk te zijn om te starten met Data Lineage.
- De DMBOK maakt onderscheid tussen Data Lineage 'as designed' en Data Lineage 'as implemented'. **Data Lineage 'as designed'** is handmatig gemodelleerde Data Lineage. **Data Lineage 'as implemented'** is de (aantoonbare) weg die data volgt door het IT landschap op een bepaald moment in de tijd.
- Voor Data Lineage 'as designed' hebben we niet direct een Metadata Management solution nodig. Handmatig modelleren van Data Lineage kan met een grote diversiteit aan tools.
- In de referentie architectuur wordt deze dimensie van Data Lineage niet verder gespecificeerd, waardoor dit aspect onderbelicht blijft. Voor Data Lineage zouden additionele ABB's gedefinieerd kunnen worden voor Data Lineage, bijvoorbeeld:

Data Flow 
Modelling

Handmatig modelleren van data lineage op meerdere niveaus van granulariteit

Data Flow 
Visualization

Gemodelleerde data lineage visualiseren op meerdere niveaus van granulariteit

Toepassingen van Data Lineage as designed

- Via een ontwerp richting geven aan de gewenste verandering
- Data Lineage inzichtelijk maken op met name een business 'niveau'
- Vastleggen of valideren SoE / SoR / SoU
- 'Gaten' dichten in de end-to-end data flow op punten waar geen data lineage 'as implemented' mogelijk is

Data Lineage

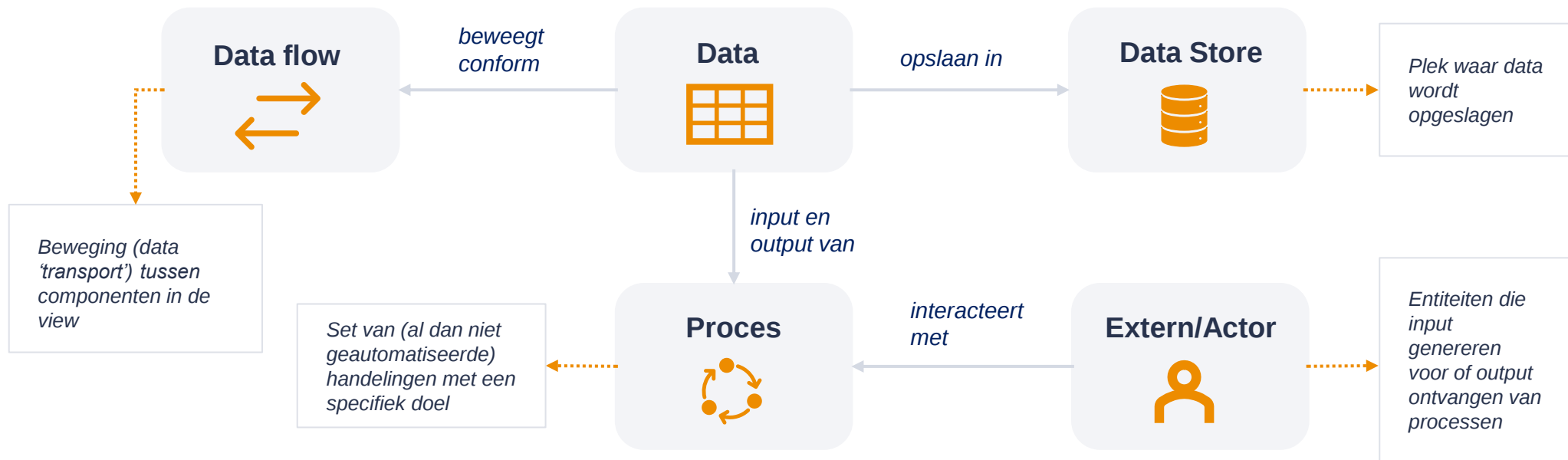
Metamodel

- Om een bruikbaar inzicht in Data Lineage te creëren is een product nodig met een goed gedefinieerde structuur.
- Die structuur wordt een 'metamodel' genoemd: een metadata datamodel dat ontstaat als antwoord op de business requirements.
- Een goed metamodel beschrijft de juiste elementen en relaties, met een specifieke granulariteit en een consistent 'beschrijvingsniveau'.
- Metamodellen kunnen naast elkaar bestaan en elkaar aanvullen.
- Een metamodel in de context van een Metadata Management solution zal naast Data Lineage ook andere metadata elementen beschrijven.

Data Lineage

Elementen in een Data Lineage metamodel

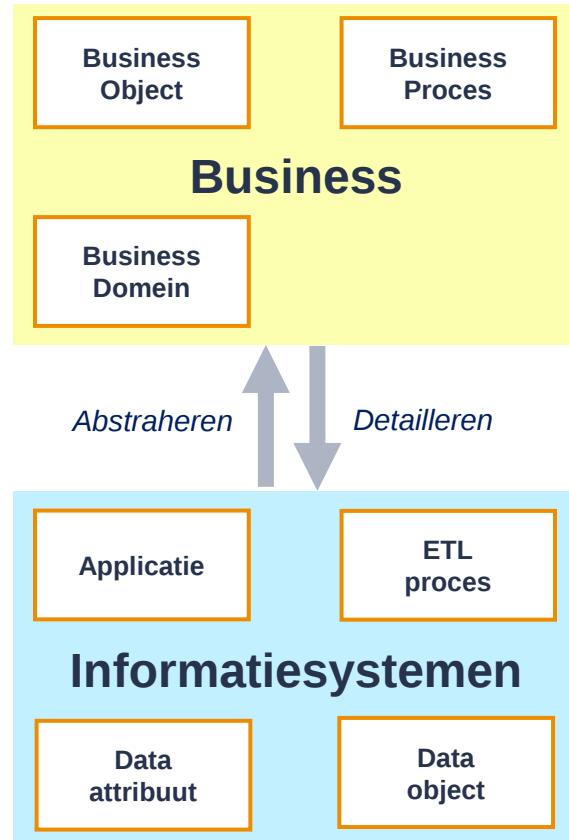
- Een view op Data Lineage is opgebouwd uit een aantal elementen. Het onderwerp van Data Lineage is altijd de component **Data**. De meest volledige view op Data Lineage beschrijft daarnaast **Processen** die de data verwerken, **Externe entiteiten** die interacteren met die Processen en **Data Stores** waar de data wordt opgeslagen.
- Afhankelijk van de informatiebehoefte in een use case kan het voldoende zijn om een view op Data Lineage te creëren bestaande uit een subset van componenten. Bijvoorbeeld Data en Data Stores, of Data en Processen.



Data Lineage

Voorbeeld
element

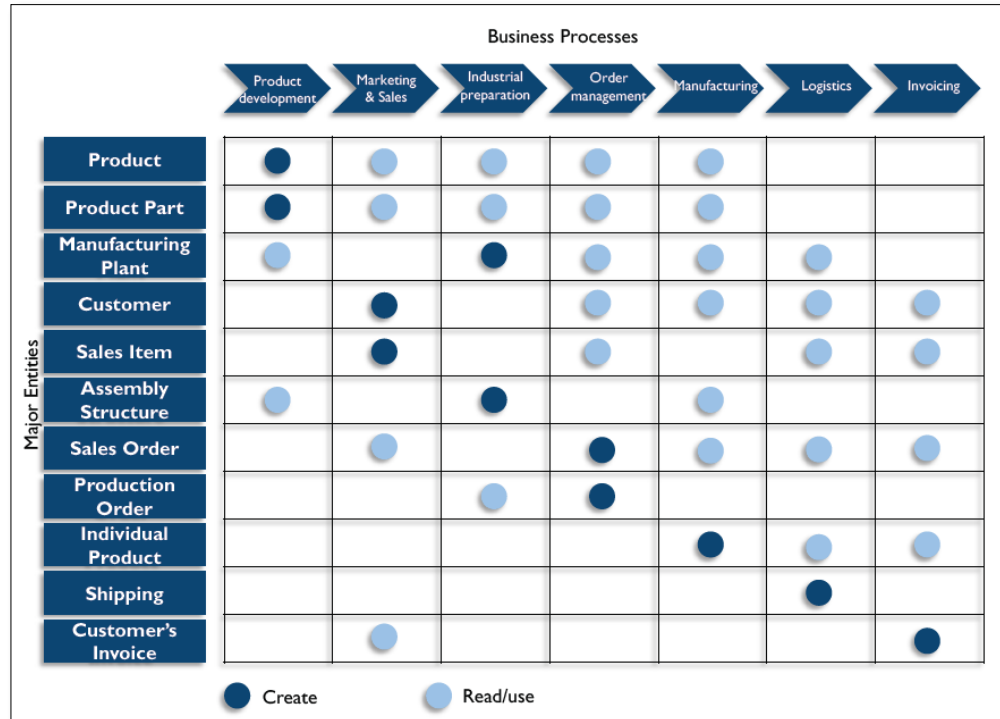
Beschrijvingsniveaus in een Data Lineage metamodel



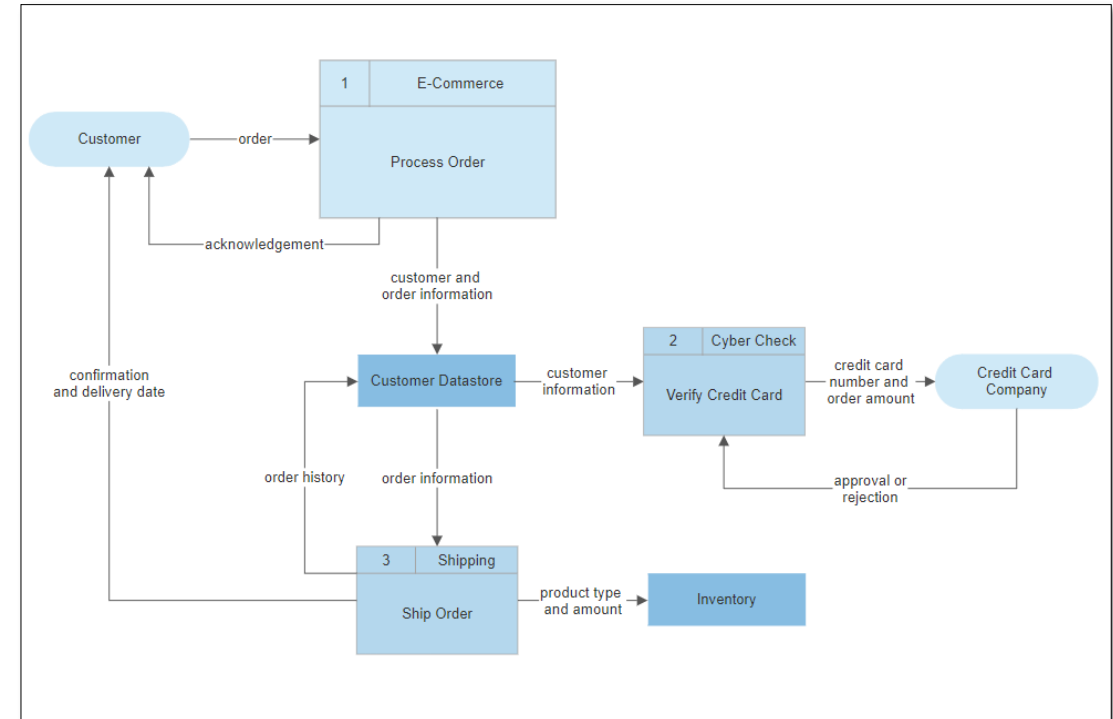
- Het creëren van een view op Data Lineage bestaat naast het selecteren van relevante elementen uit het bepalen van een passend beschrijvingsniveau.
- Een 'business view' op Data Lineage heeft als doel om inzicht te geven aan business stakeholders. Een business view reduceert complexiteit door applicatie-specifieke terminologie te vermijden, zoals tabelnamen en jobs in applicaties. In plaats daarvan wordt bijvoorbeeld eerder gesproken over business processen en business objecten. Een business view op Data Lineage vormt een ander beeld t.o.v. de fysieke werkelijkheid in systemen, maar moet daar wel consistent mee zijn.
- Een 'informatiesysteem view' op Data Lineage heeft als doel om inzicht te geven aan IT gerelateerde functies over de fysiek geïmplementeerde datastromen in informatiesystemen. Deze view ondersteunt bijvoorbeeld change management, audit & compliance en datakwaliteitsmanagement overwegend beter dan een business view.
- Een view op Data Lineage kan meerdere subniveaus kennen, waarbij een lager niveau een detaillering is van een hoger niveau.
- Het belangrijkste is dat een duidelijk beschrijvingsniveau gekozen wordt dat consistent gebruikt wordt en waarbij elementen in die context logisch met elkaar verbonden zijn.

Data Lineage

Views – Business View



Voorbeeld: high level data flow matrix met business objecten en business processen (geen actoren of data stores)

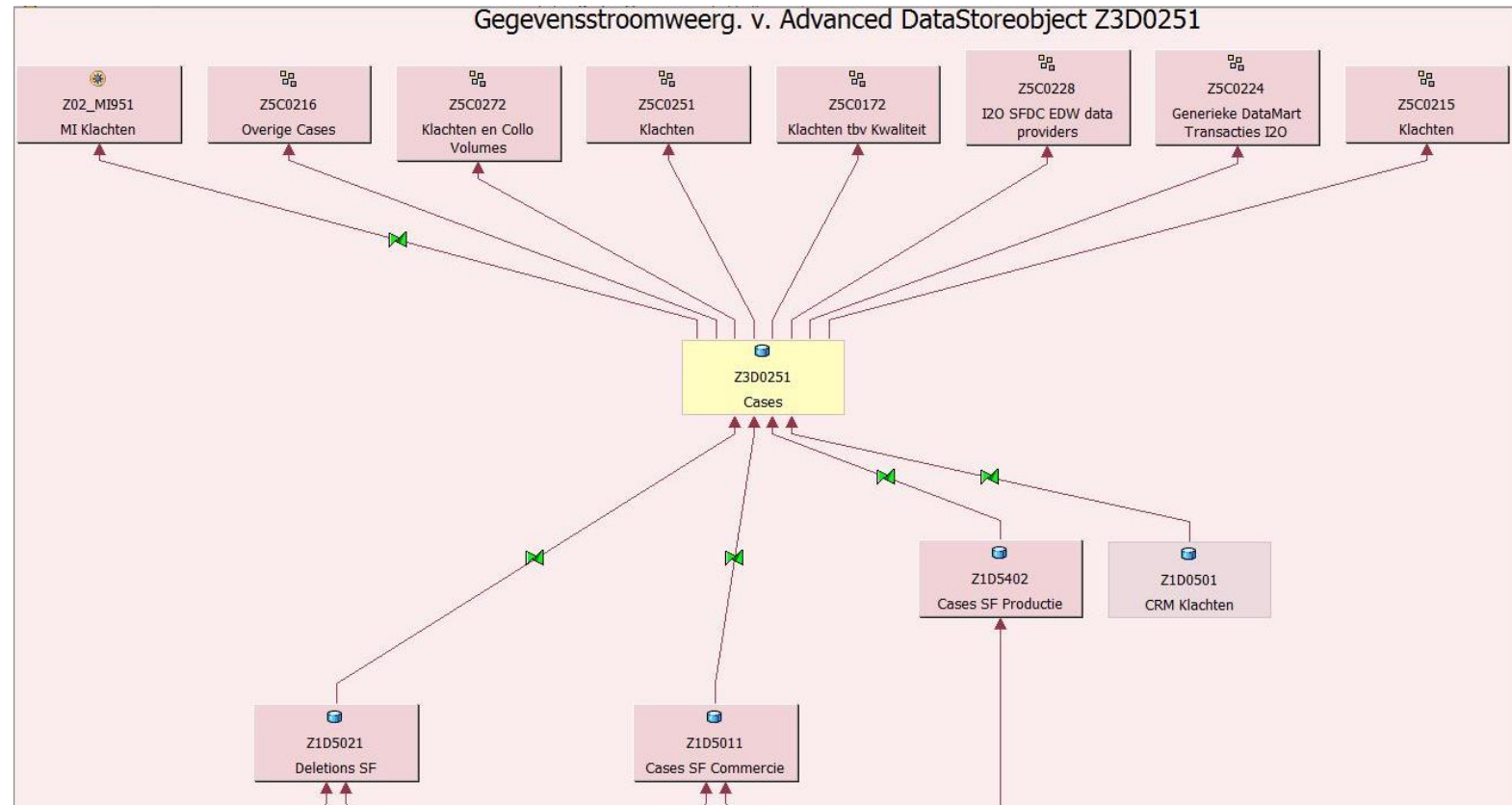


Voorbeeld: data flow diagram met processen, data stores en actoren

Data Lineage

Views – Informatiesysteem View

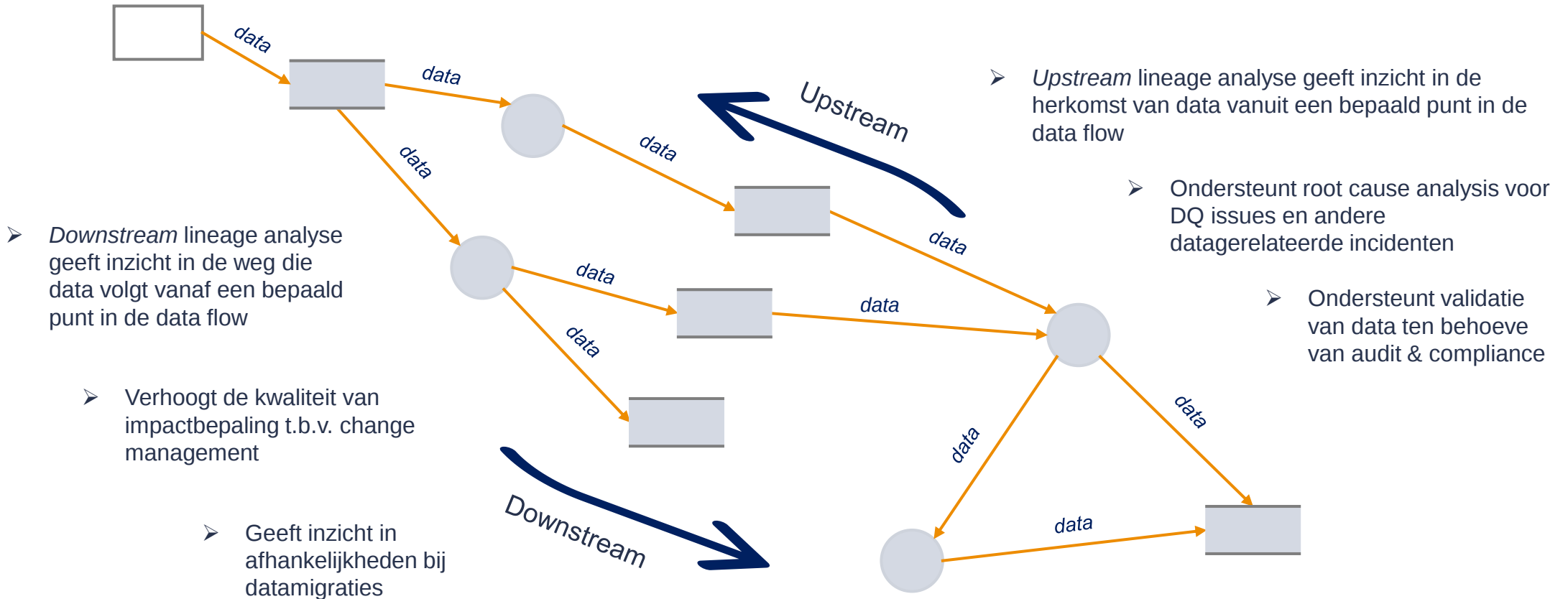
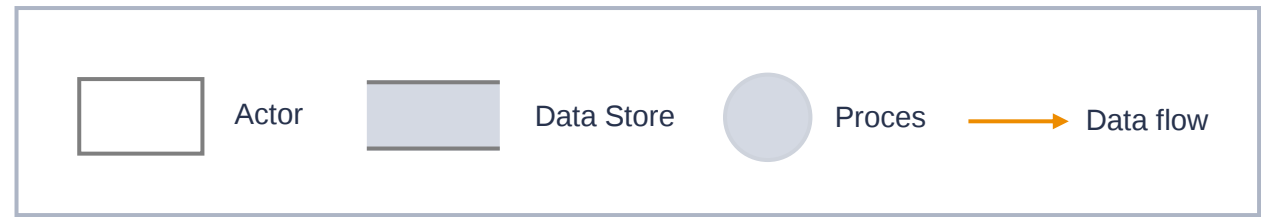
- In het voorbeeld rechts zien we hoe data vanuit de ODS laag (Z1x) zich beweegt naar de EDW laag ('Cases' - Z3D0251) en van daaruit nogmaals beweegt naar virtuele objecten in de volgende laag Z5x).
- In deze visualisatie zijn de transformatieprocessen verstoppt 'in' de pijlen, door erop te klikken zie je mapping van velden, toegepaste berekeningen etc.
- Deze methodiek geeft een hoge mate van inzicht in afhankelijkheden en risico's bij changes, en maakt inzichtelijk hoe een bepaalde uitkomst in een rapportage tot stand is gekomen.



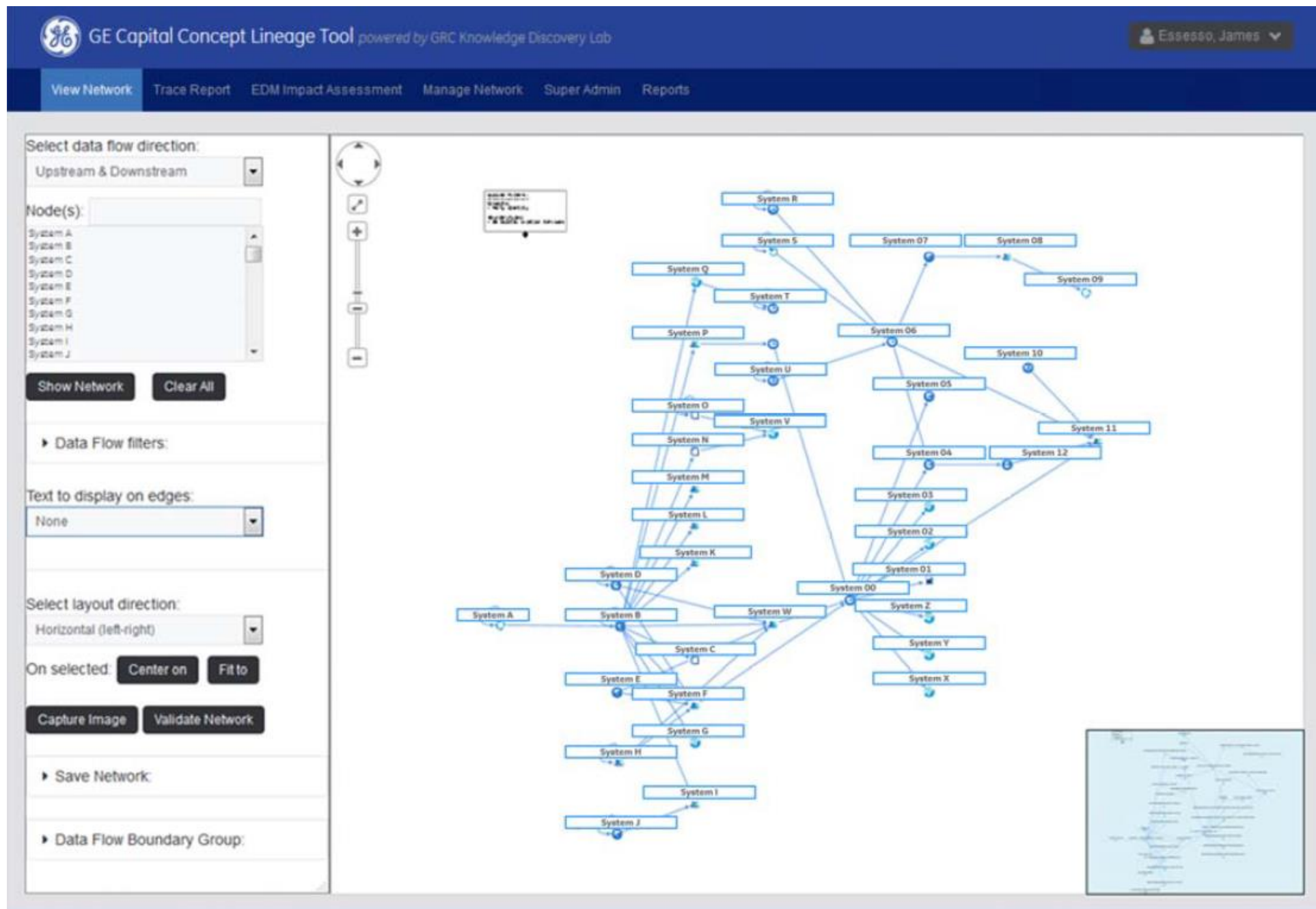
Voorbeeld: Productie Data Flow in Sap BW

Data Lineage

Upstream & downstream analyse

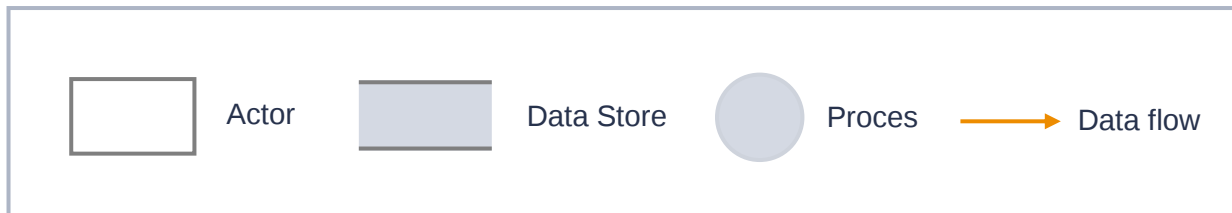


Voorbeeld van een maatwerk Data Lineage applicatie met nodes (data stores) en edges (data bewegingen). Een dergelijk (ambitieuw) ontwerp maakt nieuwe vormen van analyse mogelijk, zoals het valideren van het netwerk.

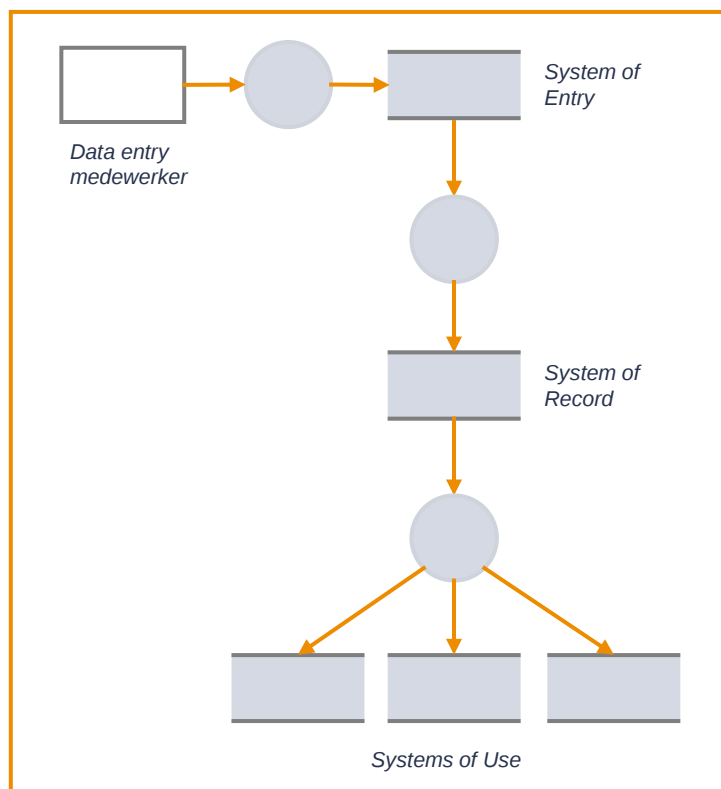


Data Lineage

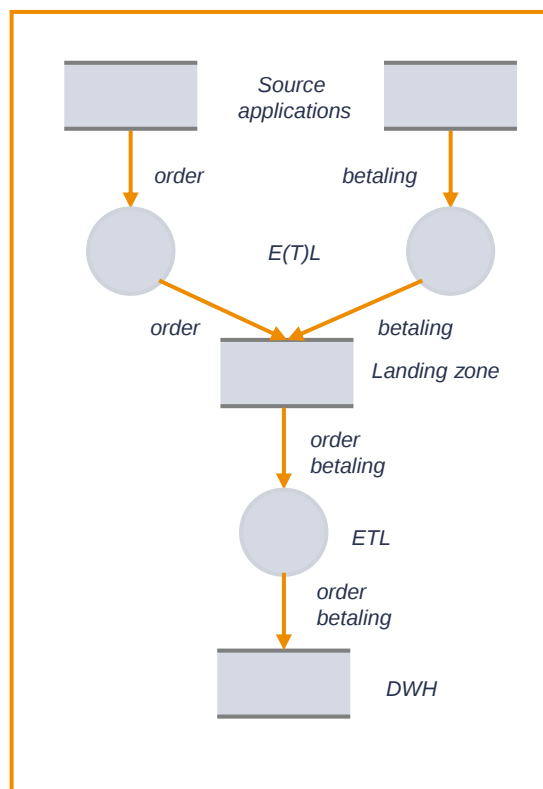
Scoping



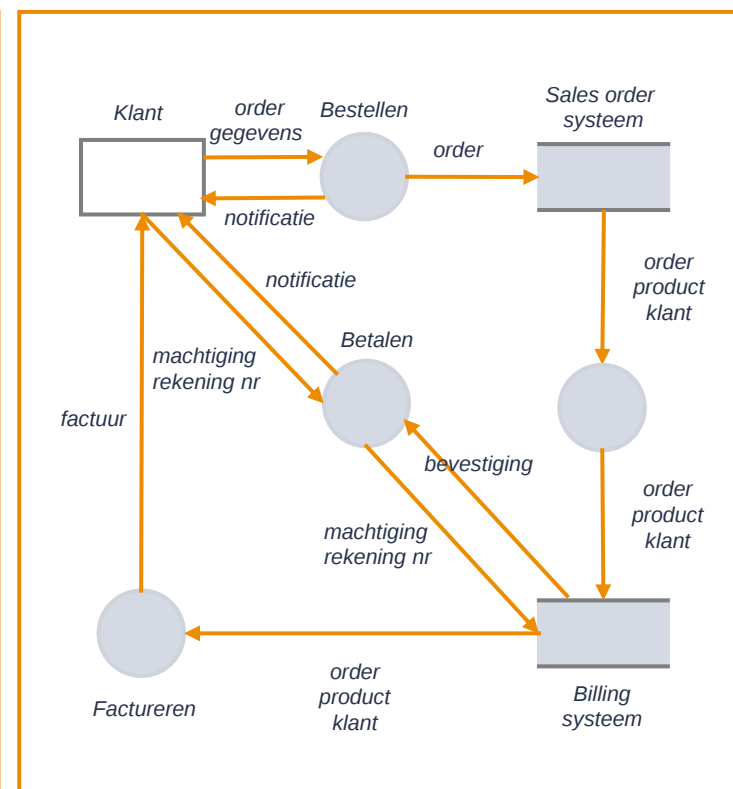
Toegevoegde waarde van Data Lineage is het best af te bakenen per use case*. Belangrijk om af te wegen is dat niet alle potentiële bronsystemen per definitie beschikken over (voldoende) lineage metadata. Gaps in de flow moeten dan handmatig gedicht worden ('stitching'). Naarmate de scope van Data Lineage toeneemt geldt dit ook al snel voor de diversiteit aan betrokken platformen, integratietools en leveranciers. Dit compliceert het verkrijgen van een compleet beeld op Data Lineage voor de gekozen use case. Daarnaast helpt het om te focussen op lineage van bedrijfskritische data.



Data Lineage met MDM scope



Data Lineage met BI scope



Data Lineage met business value stream scope

Voorlopige conclusies & overwegingen

Dit is een subtitel

- Data Lineage wordt in de markt gezien als behorend tot de scope van Metadata Management. Data Lineage metadata kan specifieke eisen met zich meebrengen die niet op dezelfde manier opgaan als voor andere vormen van metadata, zoals visualisatie in data flows.
- Er zijn allerlei verschillende views op Data Lineage mogelijk. Het opstellen van één of meerdere metamodellen, waarbij elk model gericht is op een specifieke set requirements, helpt om een passende oplossing te creëren. Onderdeel van een metamodel is ten minste een duiding van elementen, de keuze voor een beschrijvingsniveau (business of informatiesysteem view) en een mate van granulariteit.
- Data architectuur zou rondom Data Lineage toegevoegde waarde kunnen bieden op bijvoorbeeld de volgende punten:
 - Het opstellen van passende Data Lineage metamodellen
 - Het opstellen van kaders en richtlijnen voor beheer van data lineage metadata in de (metadata) lifecycle
 - Evalueren van Data Lineage & duiding geven in termen van system of entry, system of record, system of use.
- Het tijdig herkennen van een passende oplossing voor het creëren van een view op Data Lineage kan veel tijd en energie besparen. Een Data Lineage view t.b.v. audit doeleinden is bijvoorbeeld gebaat bij een hoge mate van automatisering op basis van runtime metadata, met een fijne granulariteit. Een handmatig gecreëerde view zou in dit geval leiden tot veel beheerwerk om de actualiteit te borgen en mist daarnaast de 'bewijskracht' van een geautomatiseerde oplossing.
- Scoping van een Data Lineage project is een cruciale voorwaarde voor succes. 'End-to-end' Data Lineage met een brede scope is een nobel doel, maar kan moeilijk in te richten en potentieel zeer kostbaar zijn.

BIJLAGE

Data Lineage

Introductie - aan de hand van een metafoor

- Het doel van het domein data, je zou kunnen zeggen het 'product', is kwalitatieve data. Dat betekent: data die *fit for use* is. Data kan dan gebruikt worden voor de doelen die we als organisatie gesteld hebben.
- En dat zijn niet de minste doelen. Data wordt gebruikt om ons business model te innoveren, om diensten te verbeteren en onze logistiek te optimaliseren. Op die manier bekeken is het een logisch gevolg dat we data als een asset zien.
- Maar kwalitatieve data ontstaat niet zomaar. Dat moet je maken. Net als wanneer je een fysiek product maakt heb je grondstoffen of onderdelen nodig, leveranciers die dat kunnen leveren en een betrouwbaar 'logistiek' netwerk waarin je opslag en transport regelt. En je zorgt ervoor dat je die onderdelen kunt transformeren tot een kwalitatief eindproduct.
- Voor fysieke producten vinden we het vanzelfsprekend dat organisaties een supply chain management capability hebben die inzicht geeft in de keten en het geheel bestuurbaar maakt. Maar als het gaat om de totstandkoming van een kwalitatief data 'product', dan is die vanzelfsprekendheid er niet.
- Daar komt Data Lineage om de hoek kijken. Data Lineage geeft namelijk inzicht in de 'supply chain' van data. Leveranciers zijn in dit geval de databronnen, onderdelen/grondstoffen zijn de data zelf, opslag doen we in databases, 'transport' realiseren we met integratie, en met data transformaties creëren we een data 'product'.
- Natuurlijk gaat deze metafoor niet in alle opzichten op. Maar deze manier van denken kan wel helpen om de toegevoegde waarde van Data Lineage te begrijpen.



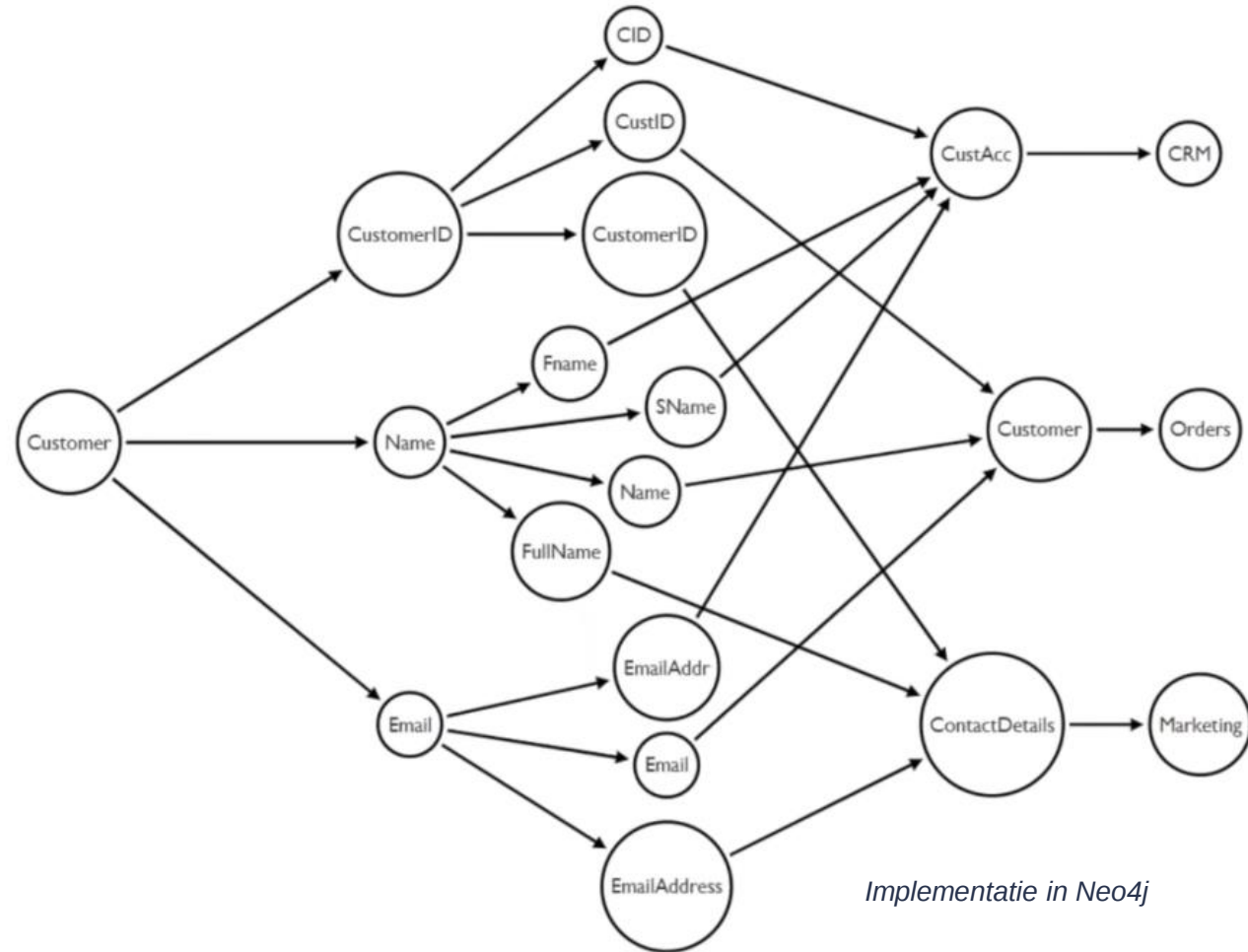
Data Lineage

Data Traceability



Metamodel voor deze Graph metadata database

- In het voorbeeld rechts worden business objecten gemapt naar fysieke objecten en tabellen in databases.
- Dit model is zeer geschikt om queries te beantwoorden zoals 'in welke databases komt business entiteit x voor'.
- Het metamodel in dit voorbeeld geeft geen duiding aan de beweging van data en voldoet niet aan de voorgestelde definitie van data lineage zoals gehanteerd in dit document, maar levert wel een mogelijk zeer bruikbaar perspectief op.



Implementatie in Neo4j