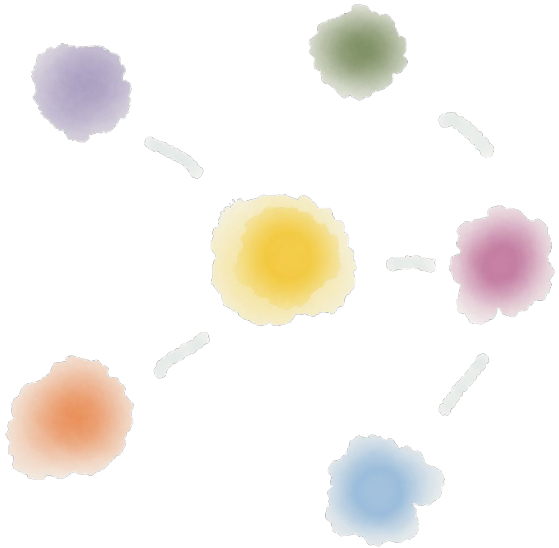
DATA MESH

A Paradigm Shift in Data Management



Zhamak Dehghani
@zhamakd

O'Reilly - Jan 2021

The Great Expectations of Data

"Our mission at Intuit is to power prosperity around the world as an ***AI-driven expert platform*** *company, by addressing the most pressing financial challenges facing our consumer, small business and self-employed customers." - Intuit (finance)*

The Great Expectations of Data

"Our mission at Intuit is to power prosperity around the world as an ***AI-driven expert platform*** *company, by addressing the most pressing financial challenges facing our consumer, small business and self-employed customers." - Intuit (finance)*

By People, For People: We incorporate human oversight into AI. With people at the core, AI can enhance the workforce, expand capability and benefit society as a whole." - AT&T (Telco)

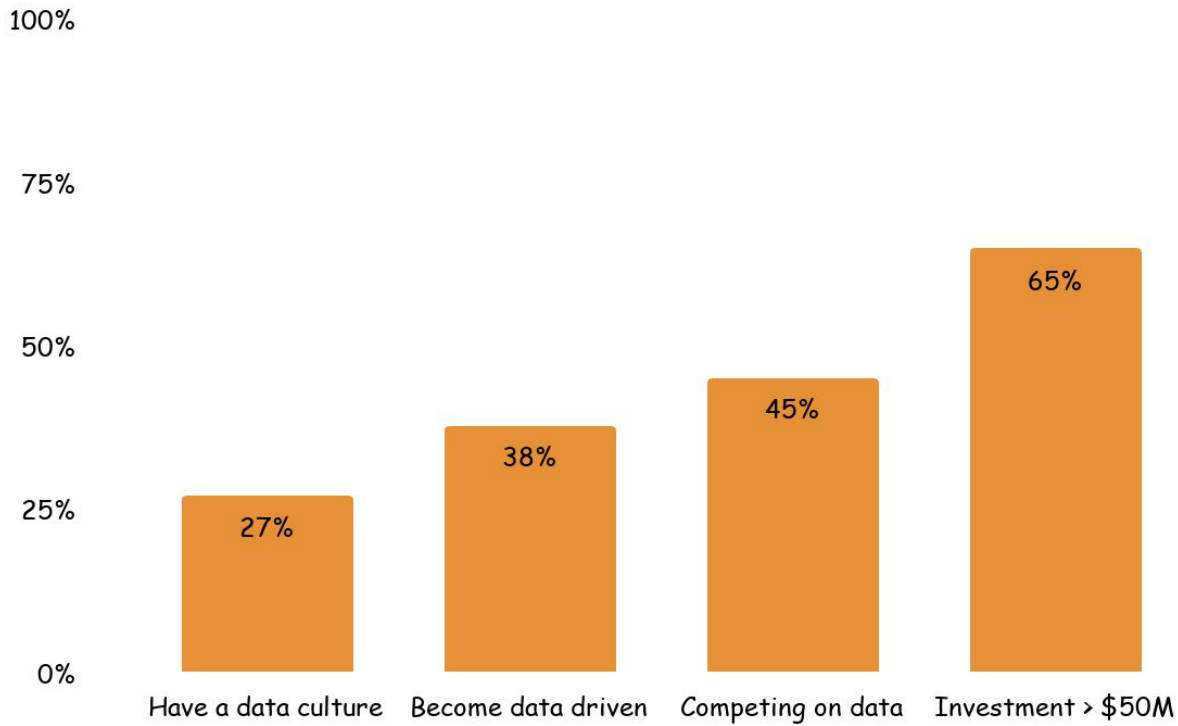
The Great Expectations of Data

"Our mission at Intuit is to power prosperity around the world as an ***AI-driven expert platform*** *company, by addressing the most pressing financial challenges facing our consumer, small business and self-employed customers." - Intuit (finance)*

By People, For People: We incorporate human oversight into AI. With people at the core, AI can enhance the workforce, expand capability and benefit society as a whole." - AT&T (Telco)

"Our mission is to improve every single member's experience at every single touchpoint with our organization through ***data and AI****". - Humana (health insurance)*

The Inconvenient Truth



NewVantage Partners annual reports,

Failure Modes

“Seen in the wild”

FAIL TO BOOTSTRAP

Technology driven initiatives - disconnected from use case

Months of data lake design - no implementation

Months of technology evaluation - no value

Org wide onboarding - no end-to-end solution

FAIL TO SCALE SOURCES

Fail to onboard large sets of ubiquitous data from many operational systems

Fail to respond timely to proliferation of sources

FAIL TO SCALE CONSUMERS

Fail to respond timely to innovation agenda

Built and overfit for narrow use cases

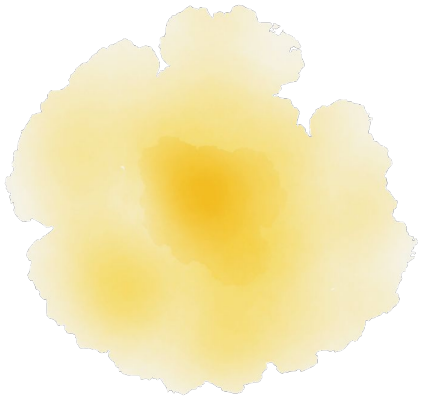
Fail to deliver to an ever growing use cases

FAIL TO MATERIALIZE DATA-DRIVEN VALUE

High development, maintenance and operational cost

Failed to compete based on data

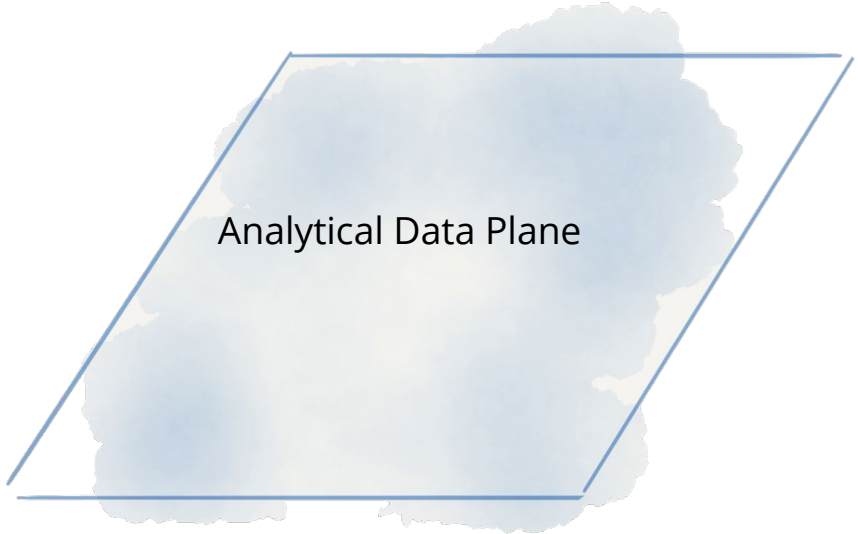
Fail to change culture beyond pockets of R&D



CURRENT DATA ARCHITECTURE LANDSCAPE

The Great Divide of Data

Running the business
Serving the users



Optimizing the business
Improving the user experience

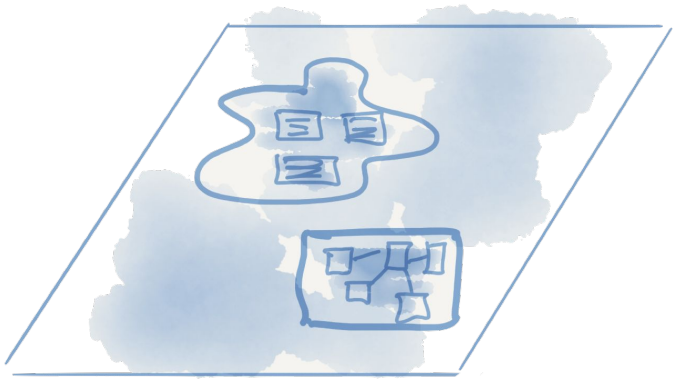
Operational Data Plane

Operational Data Plane

Running the business
Serving the users

- Captures current state of applications (business)
- Optimized for application/services
- Support transactional CRUD operations
- Accessed through APIs - Data on the inside
- Polyglot: graph database, no-sql document store, relational database, etc.

Analytical Data Plane

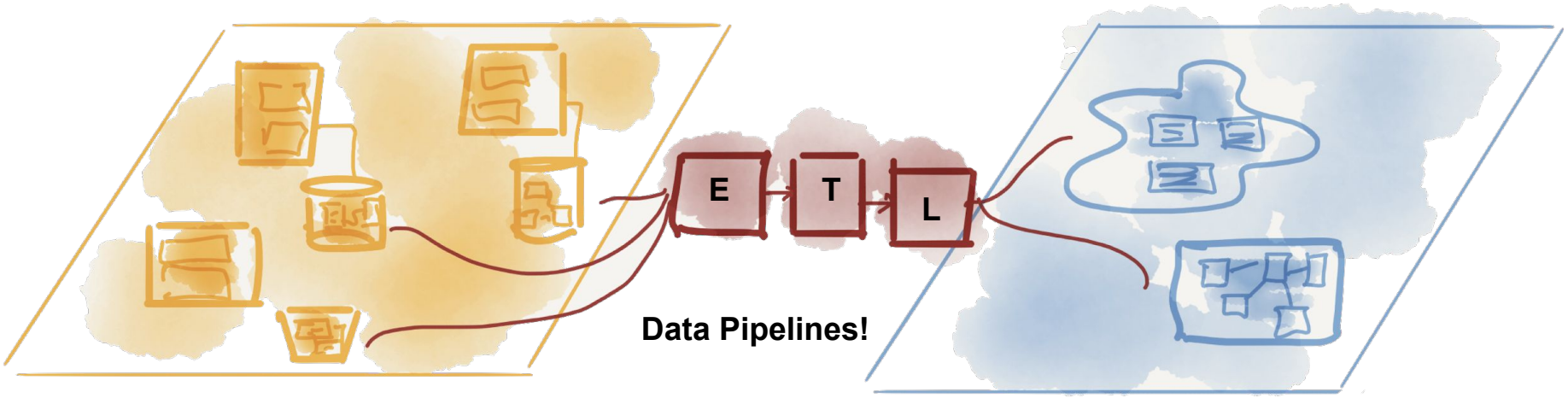


Analytical Data Plane

Optimizing the business
Improving the user experience

- Optimized for analytical logic i.e. create reports
- Optimized for intelligent modeling - i.e. machine learning training
- Temporal (Time-variant)
- Data on the outside - events, files, tables
- Polyglot e.g. object store, big table, streams

Mis-Integration



Operational Data Plane

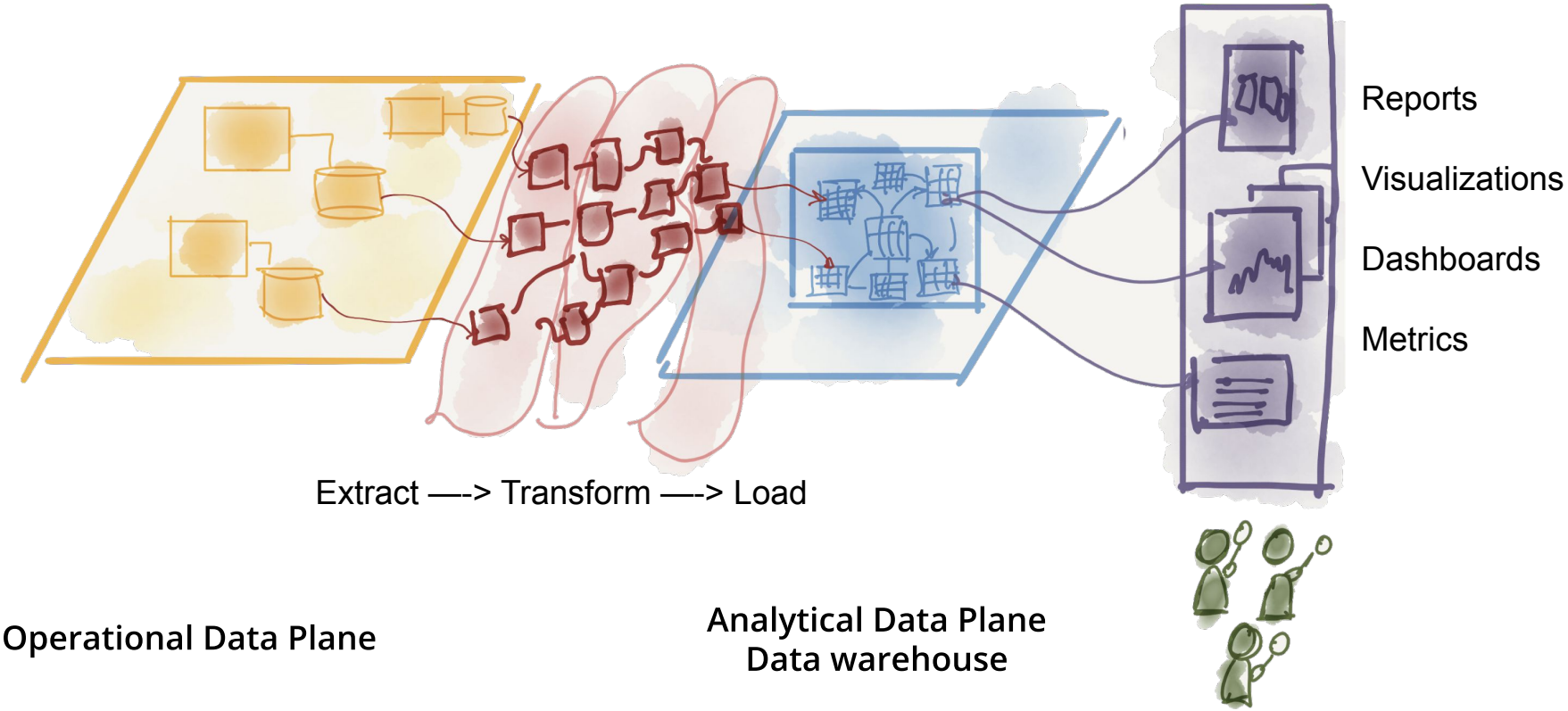
Running the business
Serving the users

Analytical Data Plane

Optimizing the business
Improving the user experience

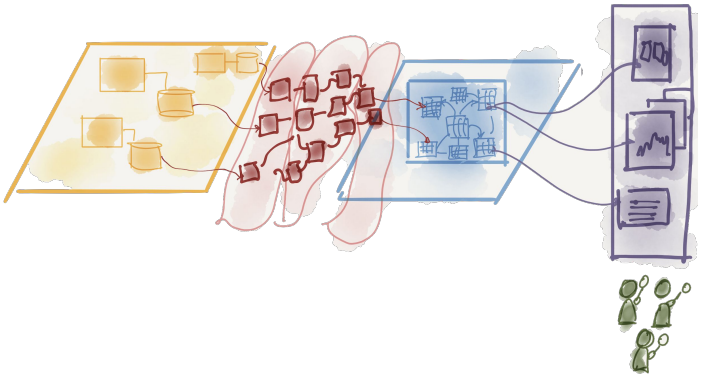
Analytical Data Architecture: Data Warehouse

"First generation 1960s"

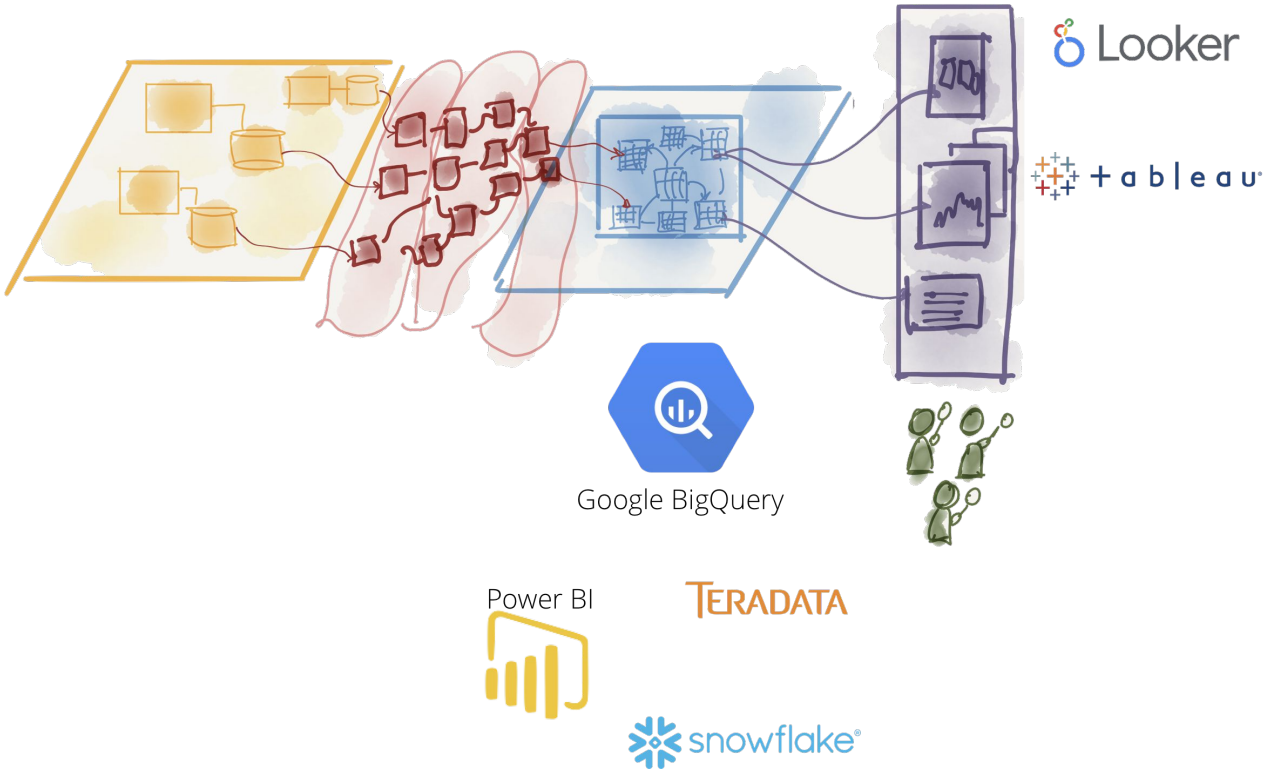


Data Warehouse Architecture Characteristics

- Data extracted from many sources - pathological coupling
- Data is transformed to single multi-dimensional schema
- Data is loaded into warehouse tables
- Warehouse is a single stack technology - monolithic
- Warehouse is used by analysis and visualization layer
- Data analysts use the analysis layer
- Data analysts create business intelligence reports and dashboards
- Data in warehouse tables is accessed with SQL-ish interface

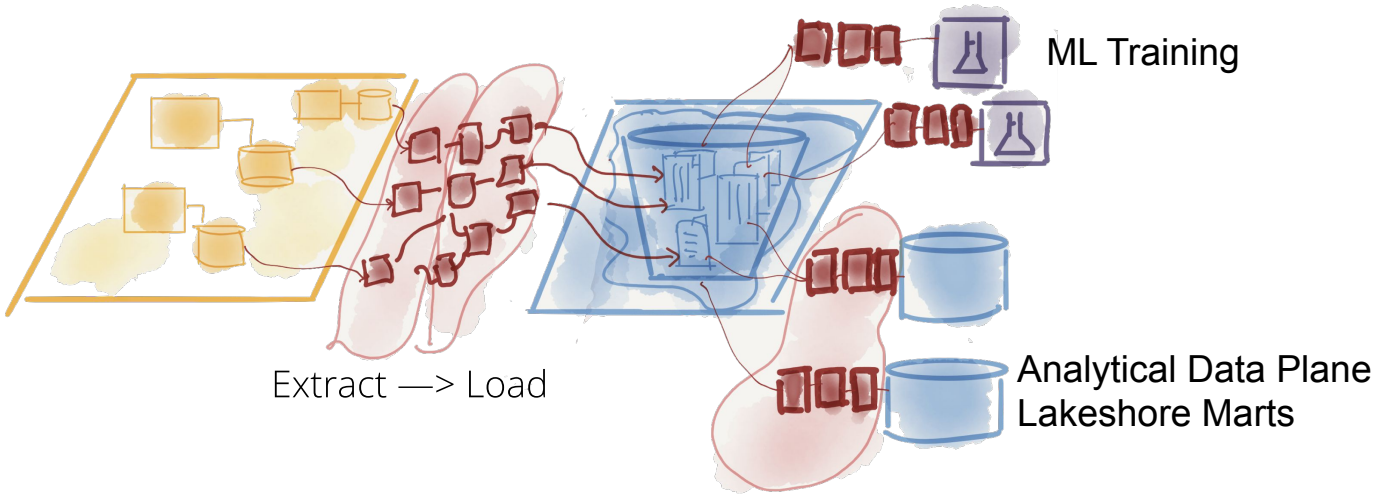


Data Warehouse Architecture: Technologies



Analytical Data Architecture: Data Lake

"Second generation 2010"

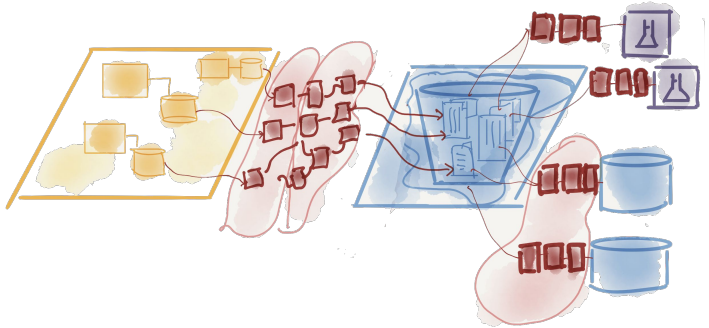


Operational Data Plane

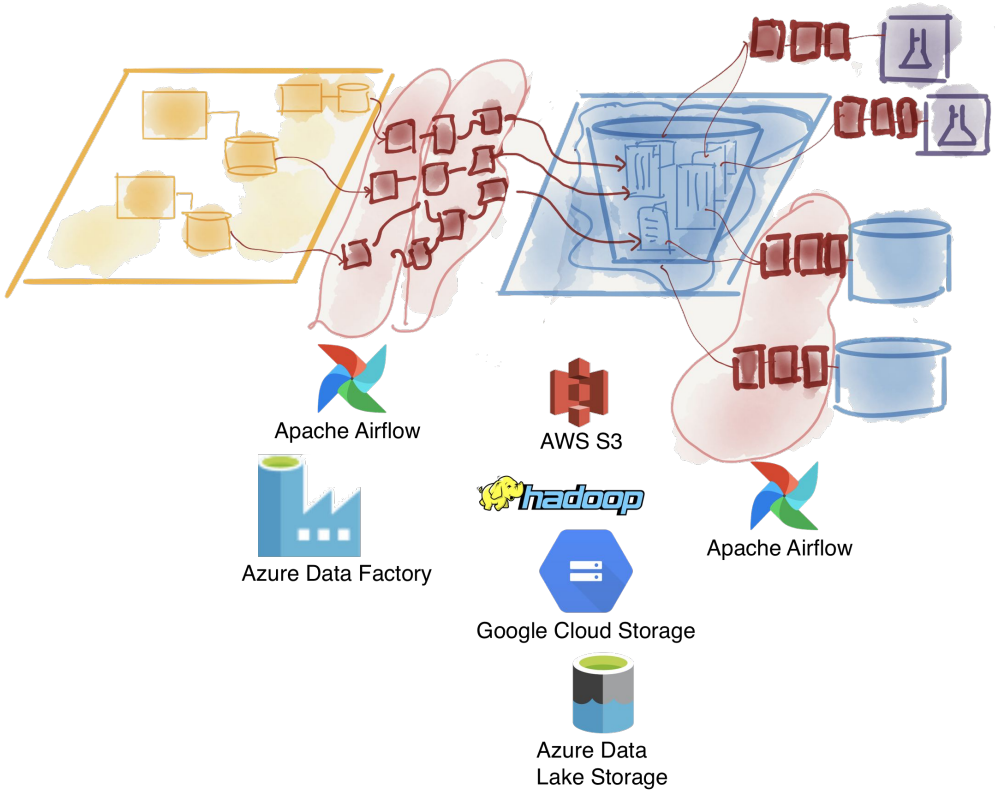
Analytical Data Plane
Data warehouse

Data Lake Architecture Characteristics

- Data is extracted from many sources
- Data is loaded into the lake in raw format
- Raw format includes semi-structured (e.g. JSON) files - monolithic
- Data scientists consume data from lake and transform for their usage
- Analysts and other data consumers transform data into their lakeshore databases for fit-for-purpose usages e.g. reports
- ML model training, data-driven apps both consume data from lake

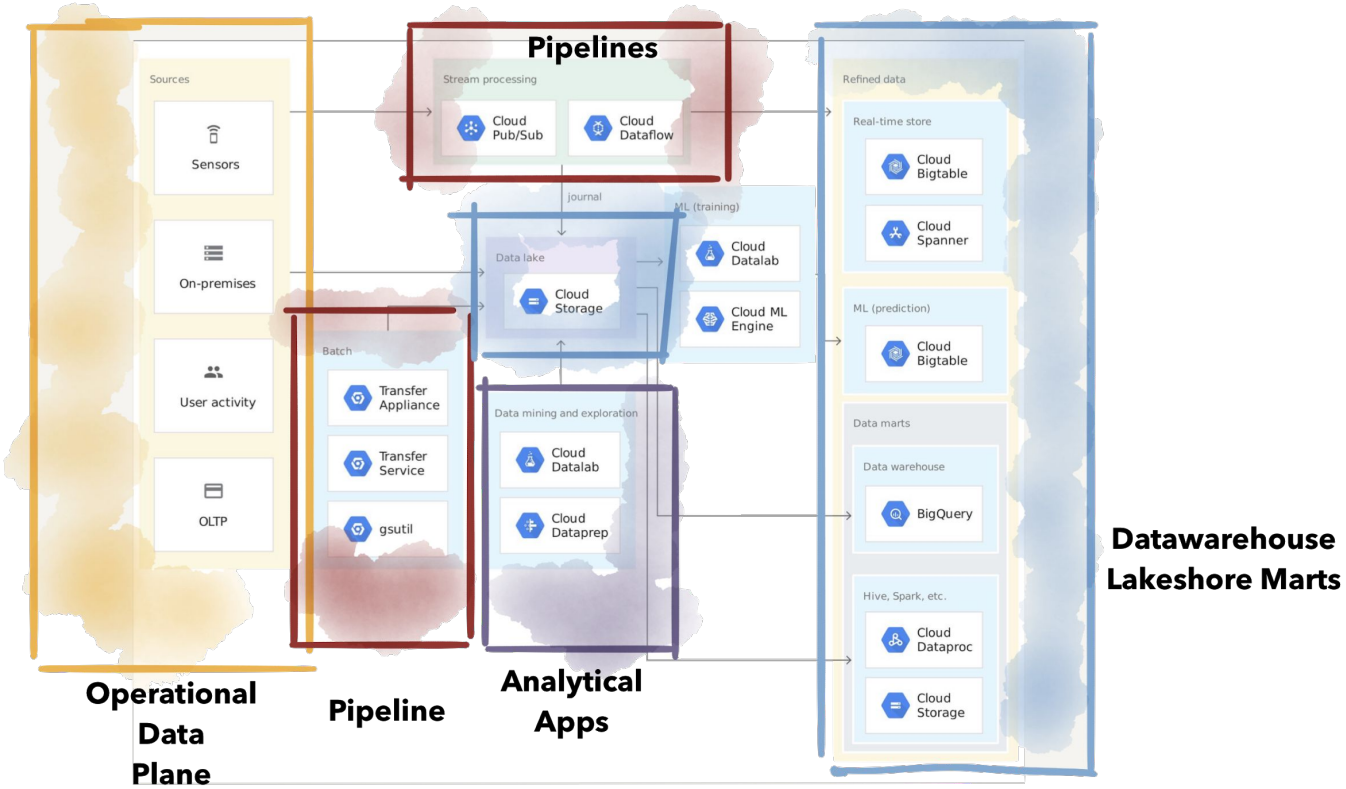


Data Lake Architecture: Technologies

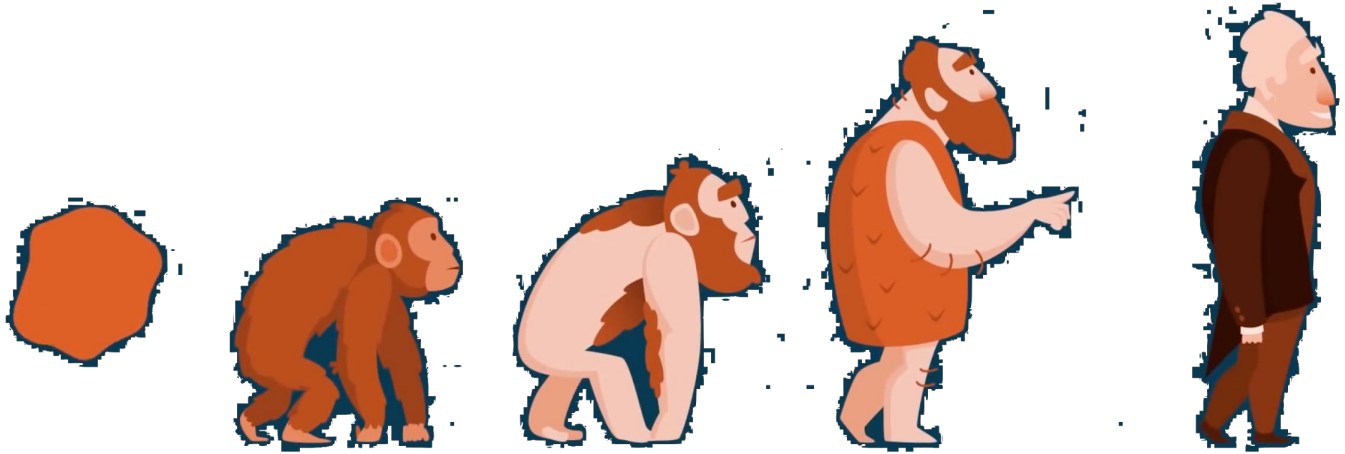


Multi-modal Data Architecture on Cloud

"Third generation now"

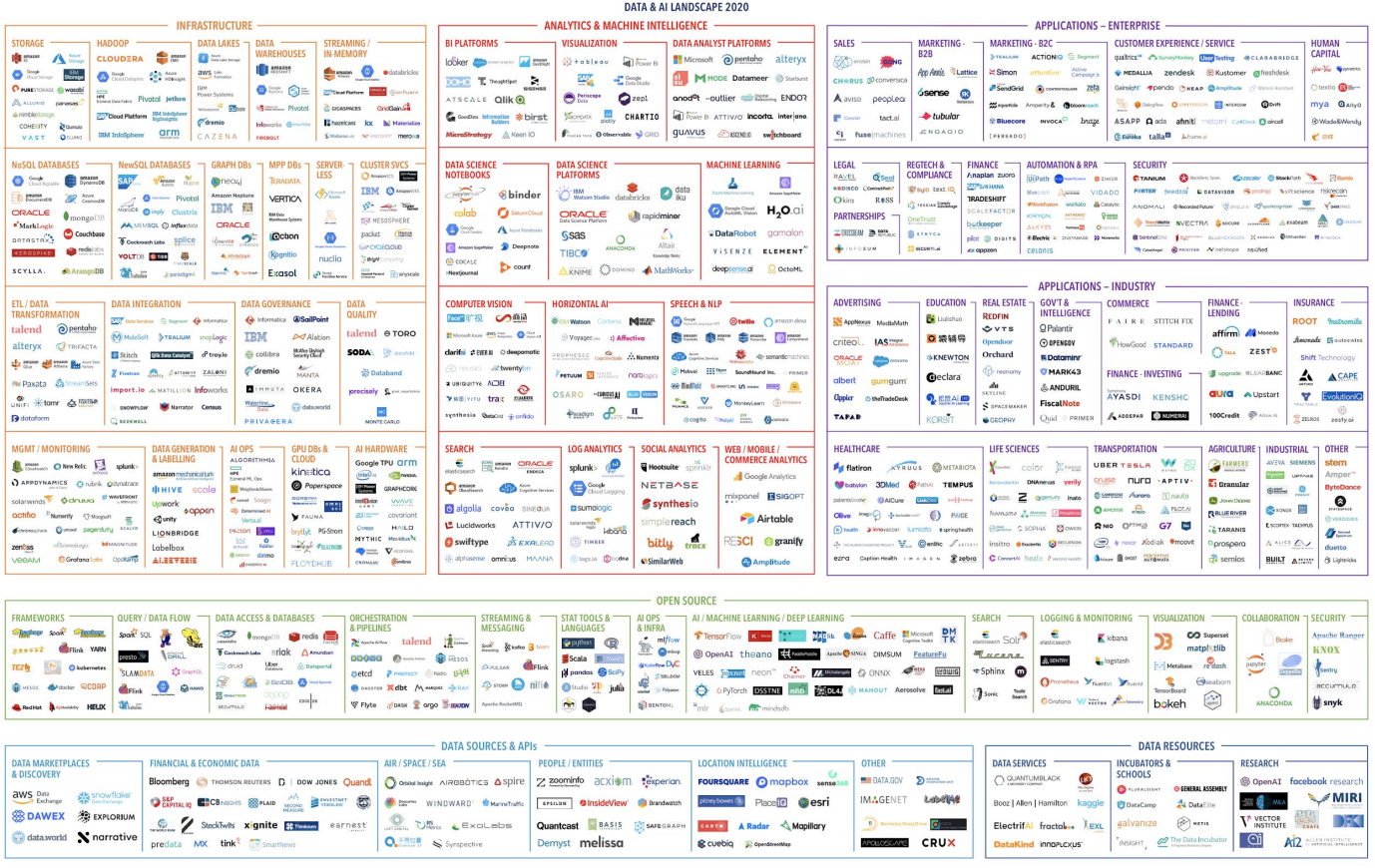


Evolutionary Improvement



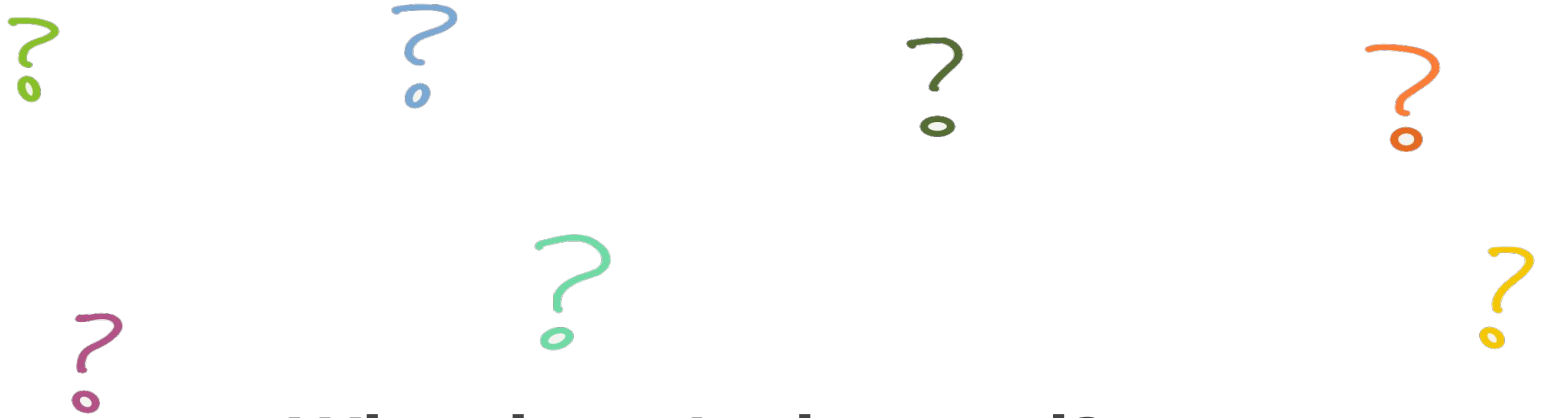
WAREHOUSE LAKE MULTI-MODAL CLOUD

Cambrian Explosion



http://mattturck.com/wp-content/uploads/2020/09/2020-Data-and-AI-Landscape-Matt-Turck-at-FirstMark-v1.pdf

Zhamak Dehghani : @zhamakd : 2021

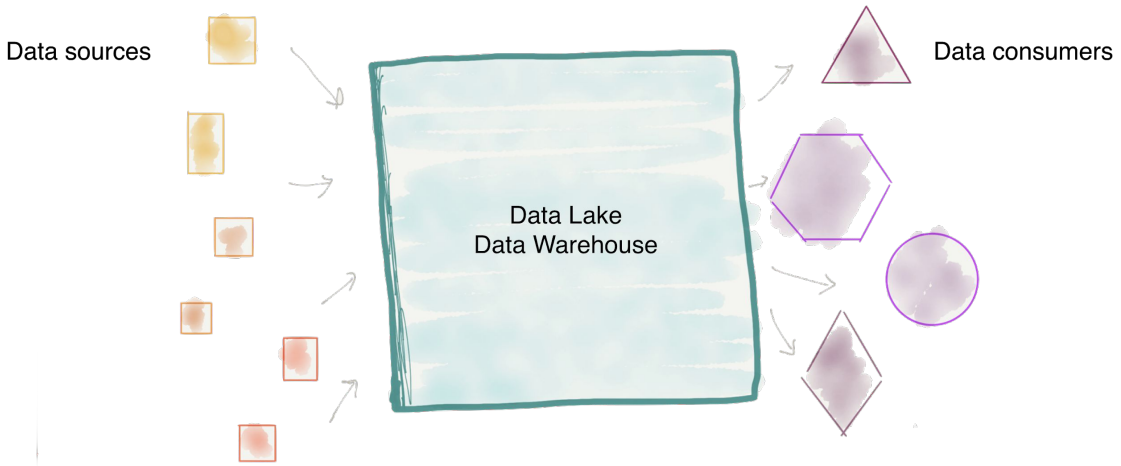


What hasn’t changed?

What assumptions to challenge?

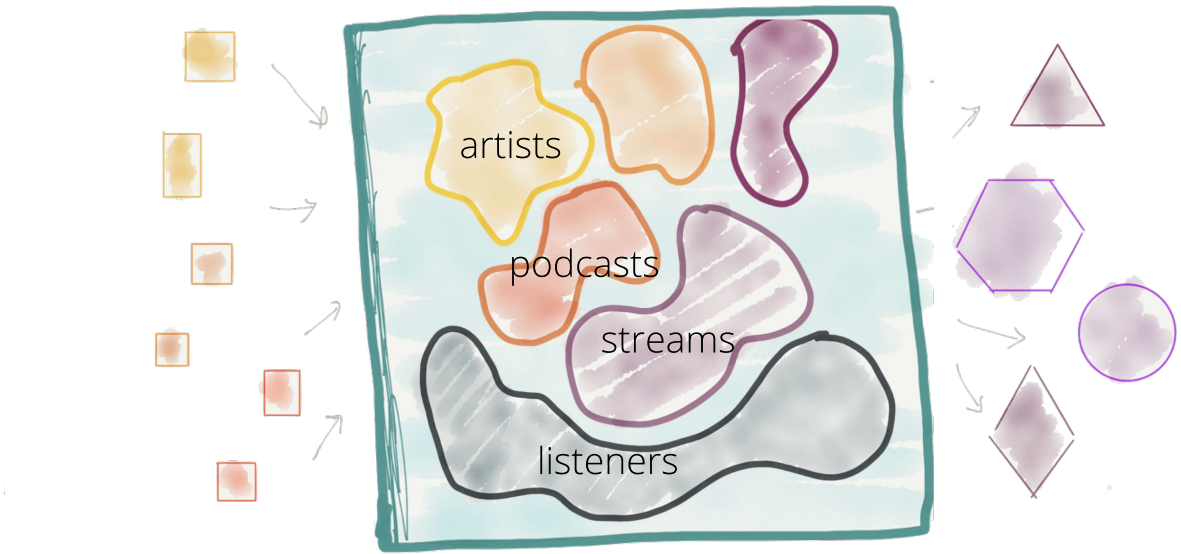
Unchallenged Assumption 1

“Data Management Solution Architecture is Monolithic”



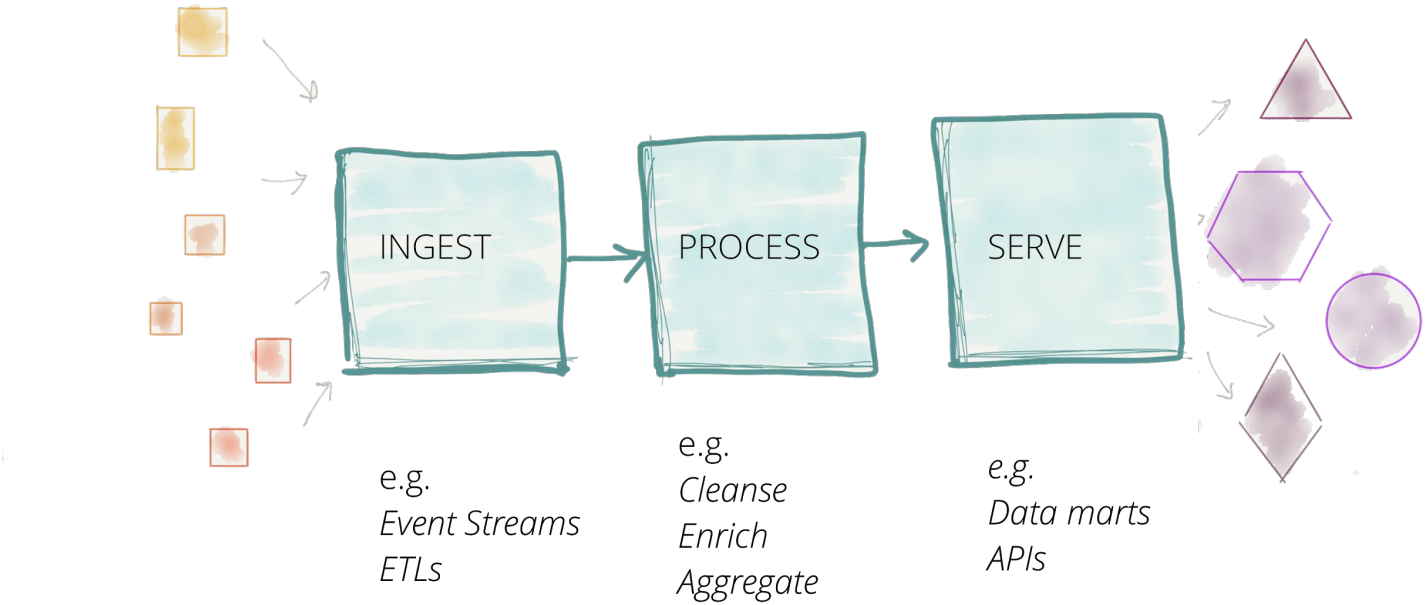
Unchallenged Assumption 2

“Data must be centralized to be useful”



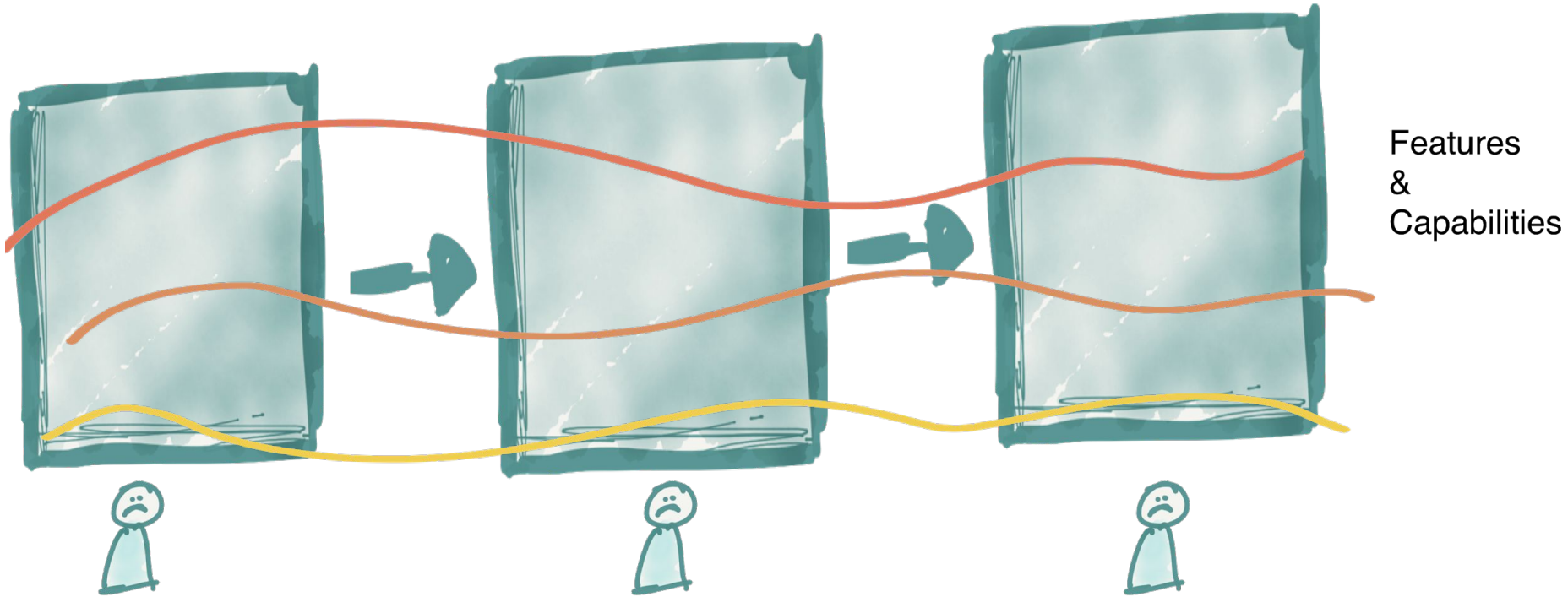
Unchallenged Assumption 3

"Scale Architecture with top-level technical partitioning"



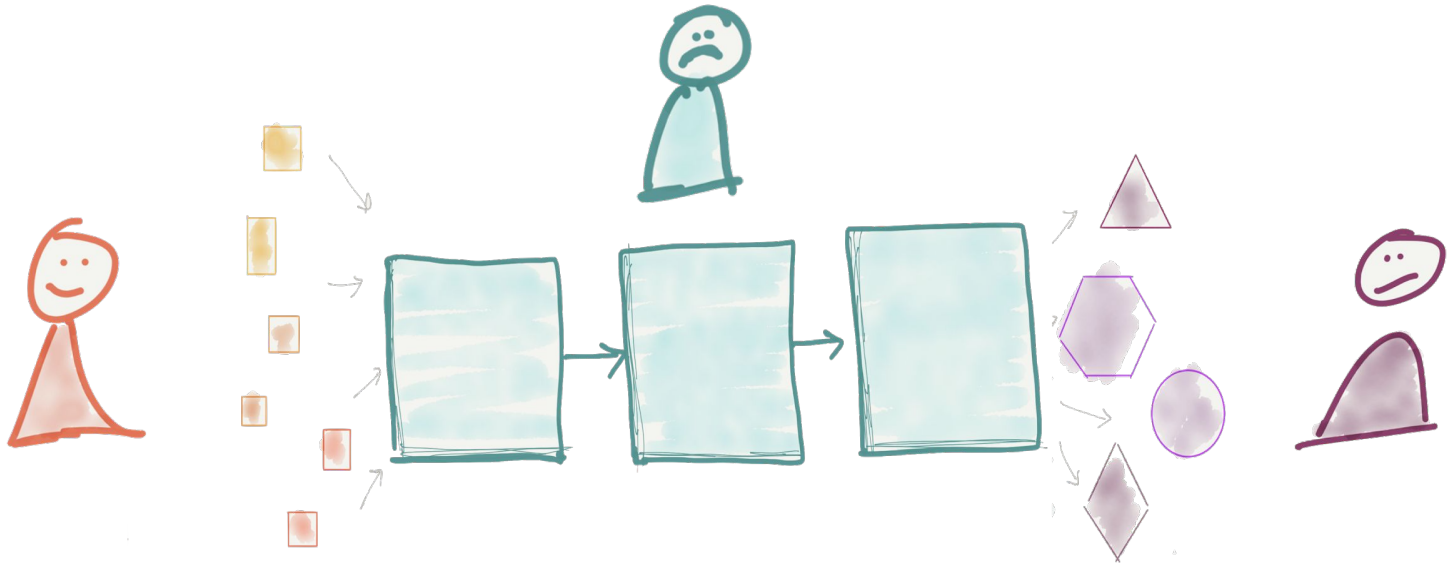
Unchallenged Characteristic 5

“Architecture decomposition orthogonal to change”



Unchallenged Characteristic 6

“Activity-Oriented Team Decomposition”

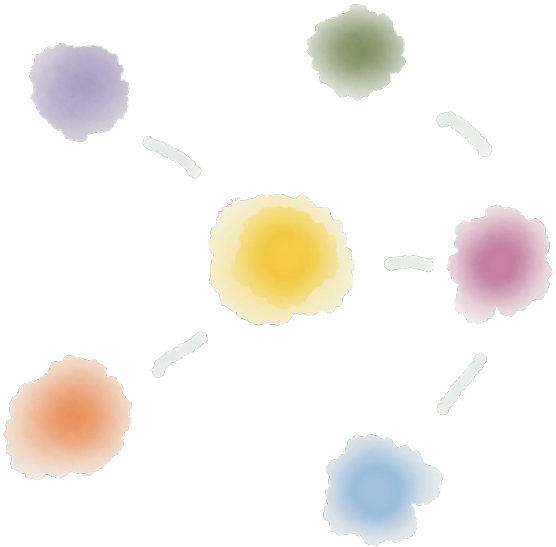


Domains’ Operational systems
(plane) Teams

Data Scientists, BI teams, ...

*“The definition of **insanity** is doing the same thing over and over again, but expecting different results.”*

- Albert Einstein



INTRODUCTION TO DATA MESH PARADIGM SHIFT

Principles and Logical Architecture

Zhamak Dehghani : @zhamakd : 2021

Data Mesh Objective

Data mesh objective is to create a foundation for getting value from analytical data and historical facts at ***scale*** - scale being applied to:

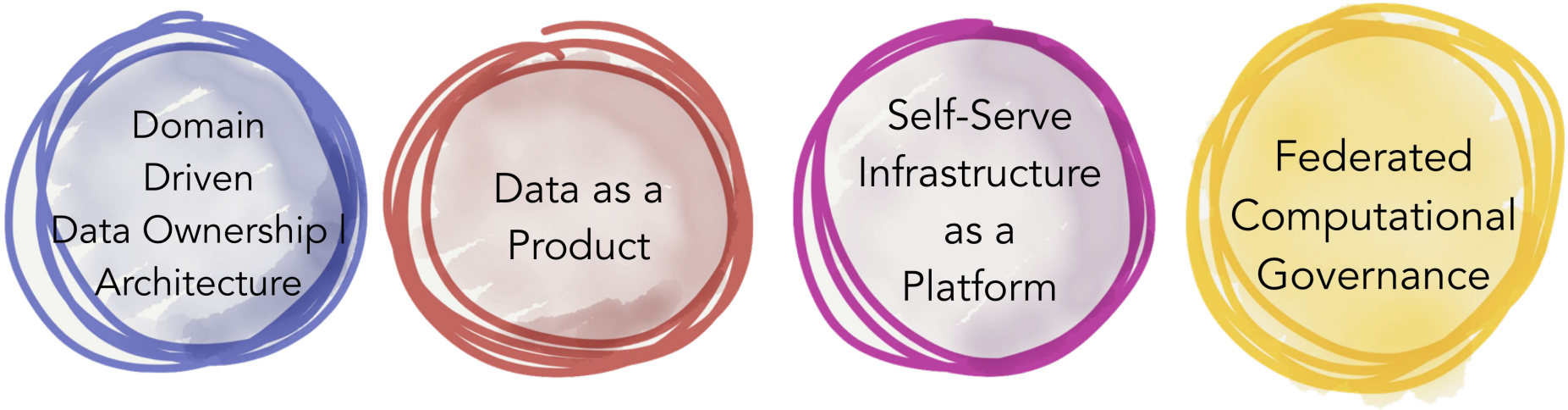
- constant change of data landscape,
- proliferation of both sources of data and consumers,
- diversity of transformation and processing that use cases require,
- speed of response to change.

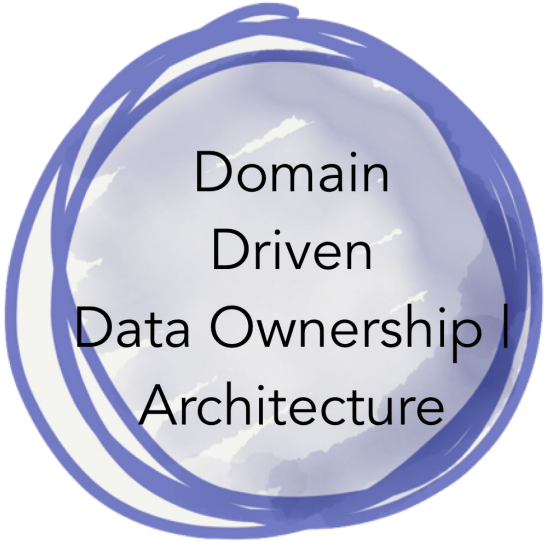
“To reject one paradigm without simultaneously substituting another is to reject science itself.”

— Thomas S. Kuhn, The Structure of Scientific Revolutions

Data Mesh 4 Principles

Underpinning the architecture and organization structure



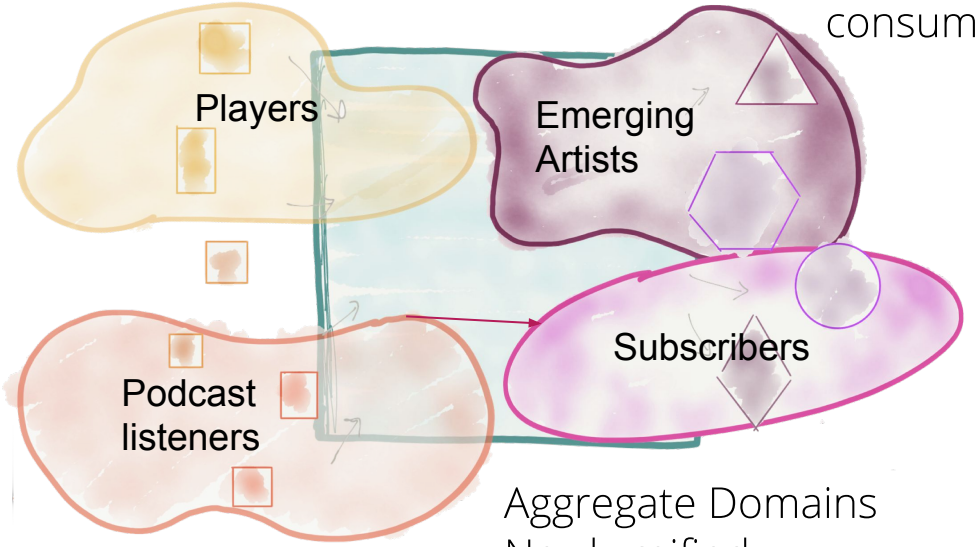


DATA
+
DOMAIN DRIVEN DESIGN

Domain Oriented Data Ownership & Architecture

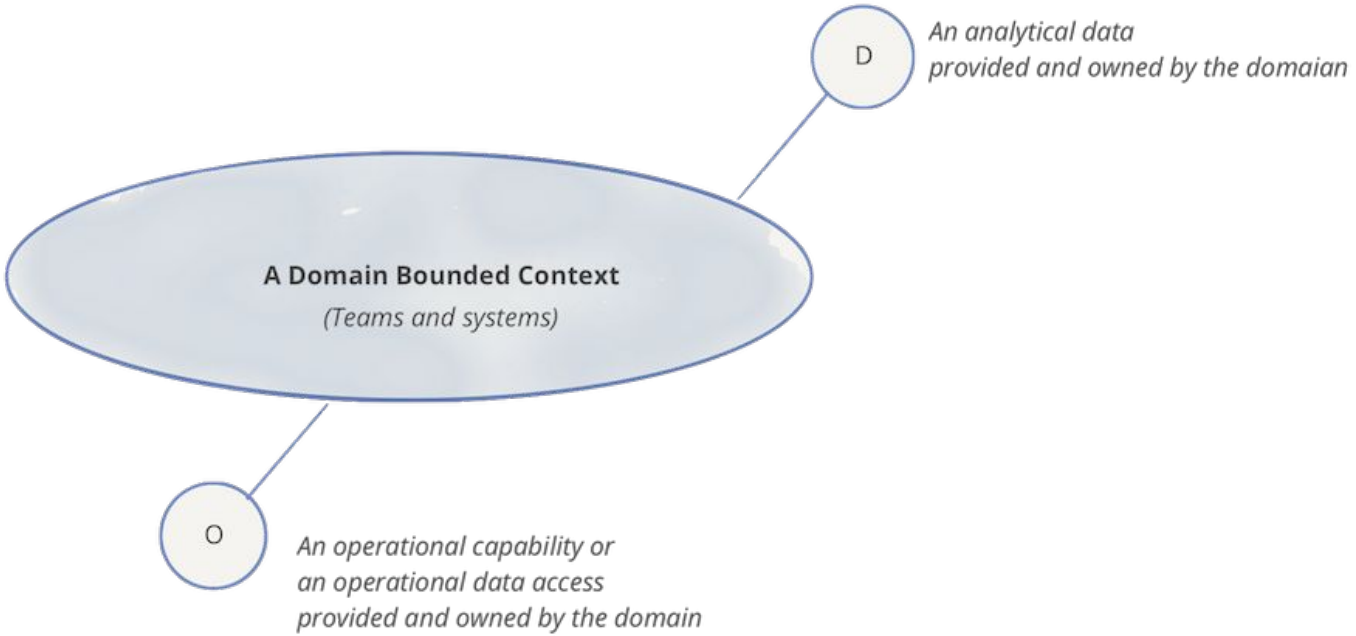
Domains aligned with specific use cases - consumption scenarios

Domains aligned with the source

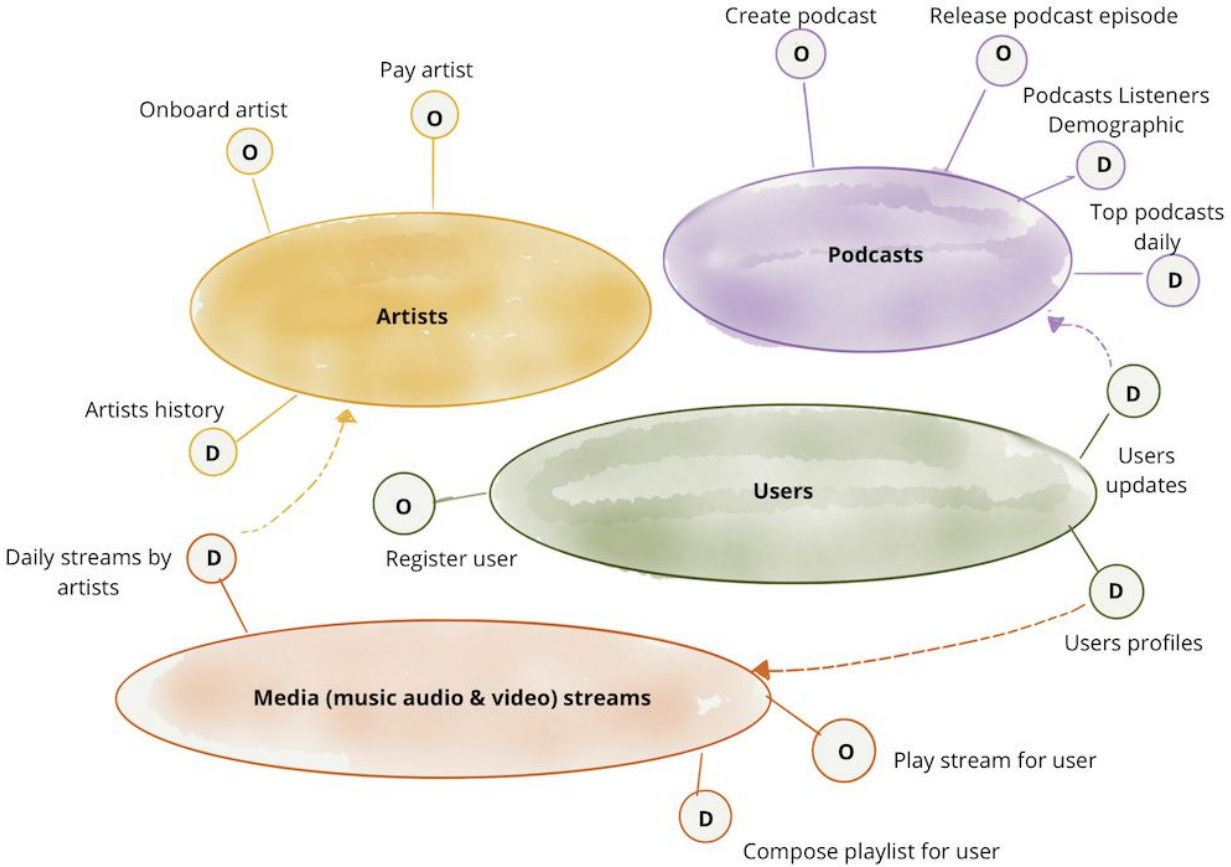


Aggregate Domains
Newly reified

Logical Architecture: Domain-oriented Data and Compute



Example


Create podcast
O
Release podcast episode
O
Pay artist
O
Onboard artist
O
Podcasts Listeners Demographic
D
Top podcasts daily
D
Podcasts
Artists
D
Artists history
D
Users updates
Users
O
Register user
Daily streams by artists
D
D
Users profiles
Media (music audio & video) streams
O
Play stream for user
D
Compose playlist for user

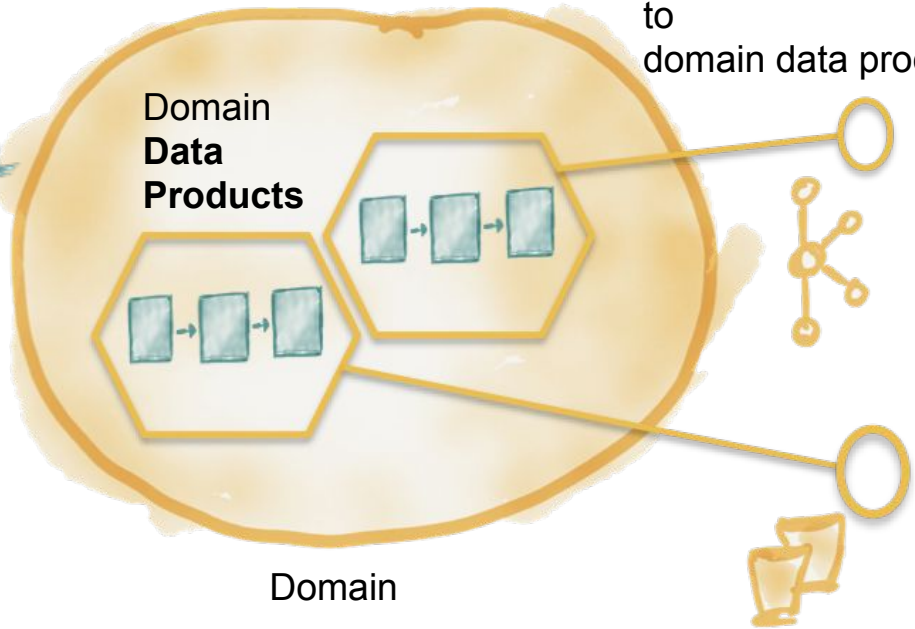
DATA
+
PRODUCT THINKING

Serve Data as a Product

delight the data consumer

Cross-functional Domain Team:

Data product owner
Data product developers
App Developers



Multi-modal access
to
domain data products

Spectrum of
data consumer
Personas

Data Analyst

Data Scientists

Characteristics of Data as a Product

SHARED | DISCOVERABLE

SELF-DESCRIBING

ADDRESSABLE

INTER OPERABLE
(GOVERENED
BY GLOBAL STANDARDS)

TRUSTWORTHY
(DEFINED & MONITORED SLOs)

SECURE
(GOVERENED
BY GLOBAL ACCESS CO
NTROL)

Data Product Owner Role

SUCCESS CRITERIA

Customer satisfaction :

Net Promoter Score

Active Product Users:

#downstream consuming data products, #downstream applications, #users

Ease of discovery and use: Decreased lead time from discovery to consumption

Data Product Owner Role

Responsibilities and Capabilities

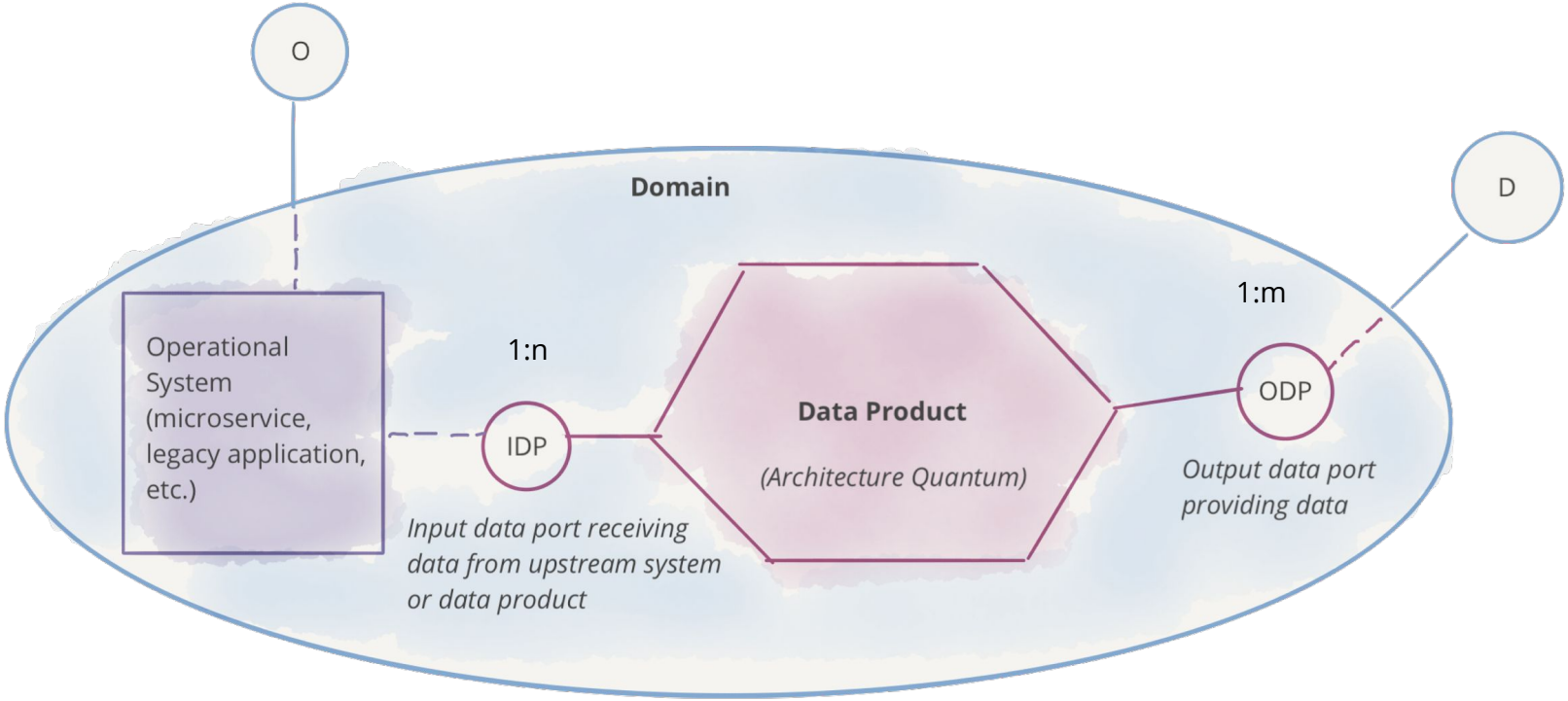
Deep understanding of the business domain and data semantic, syntax, regulations

Understanding of data use cases and applications

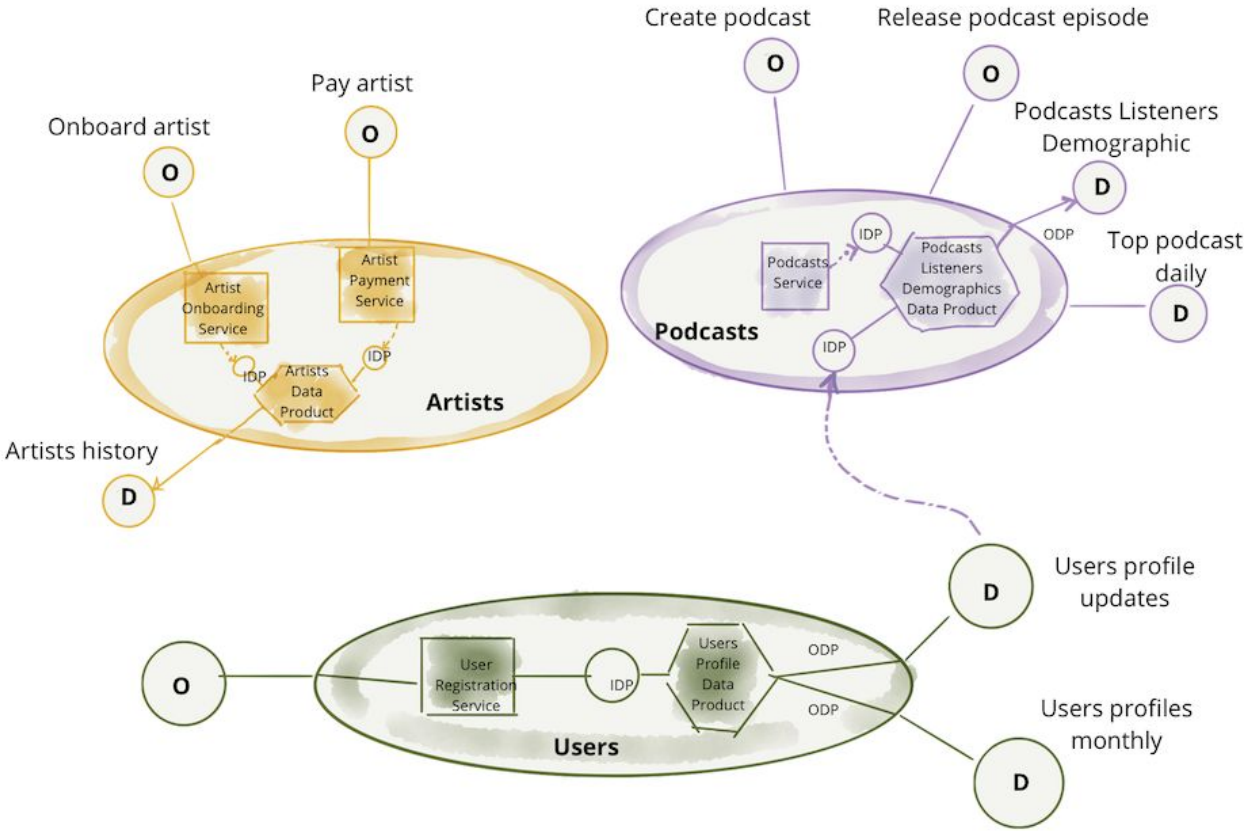
Know the customers

Long term ownership

Logical Architecture
Data Product Quantum of Architecture



Data Product Logical Architecture Example



Logical Architecture: Architectural Quantum

Architectural quantum, as defined by Evolutionary Architecture, is the smallest unit of architecture that can be independently deployed with high functional cohesion, and includes all the **structural elements** required for its function.

https://evolutionaryarchitecture.com/

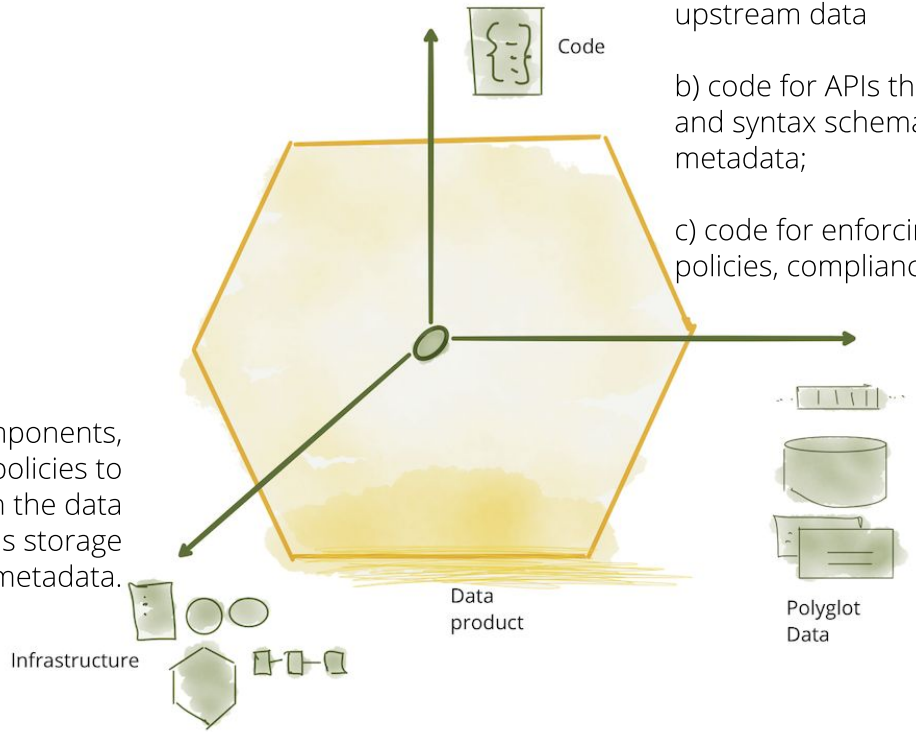
Data Product's Structural Elements

"One Unit of Architecture"

a) code for data pipelines consuming, transforming upstream data

b) code for APIs that provide access to data, semantic and syntax schema, observability metrics and other metadata;

c) code for enforcing traits such as access control policies, compliance, provenance, etc.



The infrastructure components, configuration and policies to build, deploy and run the data product's code, as well as storage and access to big data and metadata.

Polyglot Data and metadata (documentation, metrics,...)

Summary

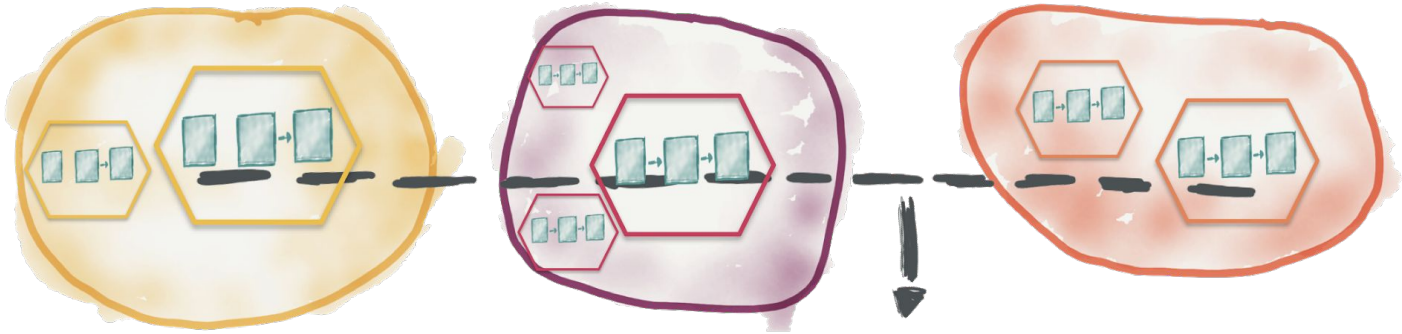
Data as a product

So that data users can easily discover, understand and securely use high quality data with a delightful experience; data that is distributed across many domains.

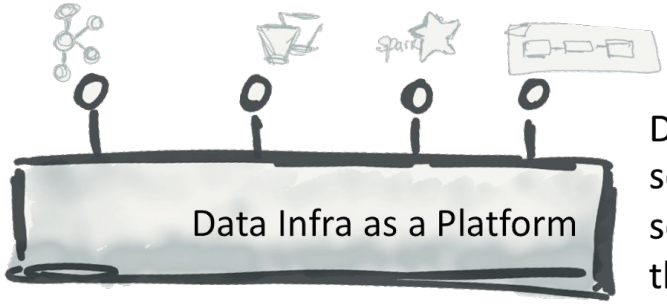
DATA
+
PLATFORM THINKING

Enable Autonomy

Abstract technical complexity by a self-serve data infrastructure



Data platform engineers



Domain agnostic shared/centralized self-serve tooling / infrastructure to create secure data products quickly, discover them, execute their pipelines, etc.

Data Platform Product Owner(s)

Data Platform Team Responsibilities

Success Criteria

Customer - **Data Product Developer** - satisfaction :

Net Promoter Score

Reduced friction for users:

Reduced lead time to create new secure , discoverable data products

Growth of active users:

Number of data products built and hosted by platform

Data Platform Team Responsibilities

Characteristics

Domain-agnostic capabilities

Understanding of data product developer journey

Know the customers

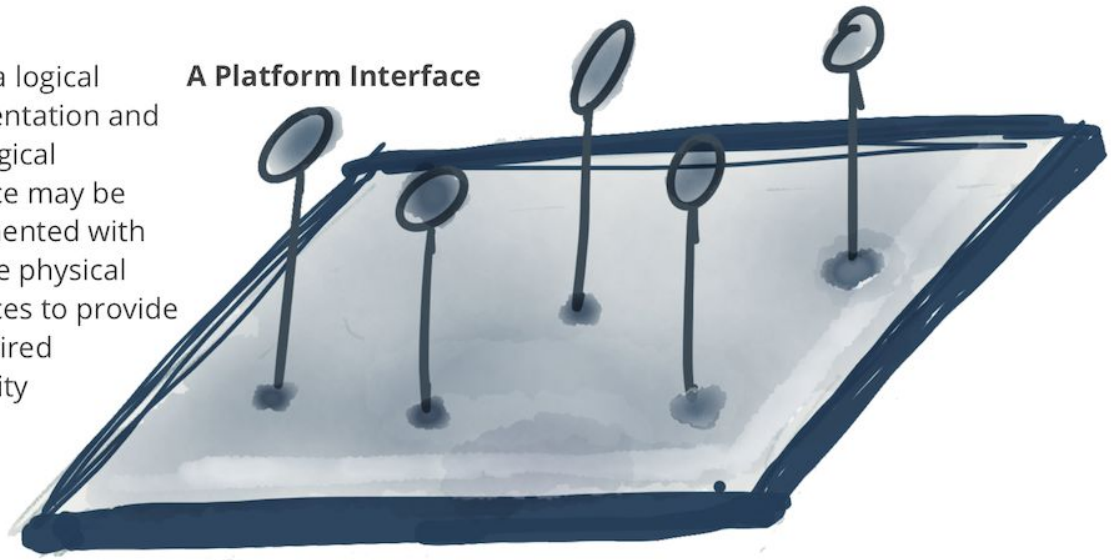
Be opinionated at incubation and target clear segment of customers

Logical Architecture

A multi-plane data platform

This is a logical representation and each logical interface may be implemented with multiple physical interfaces to provide the desired capability

A Platform Interface

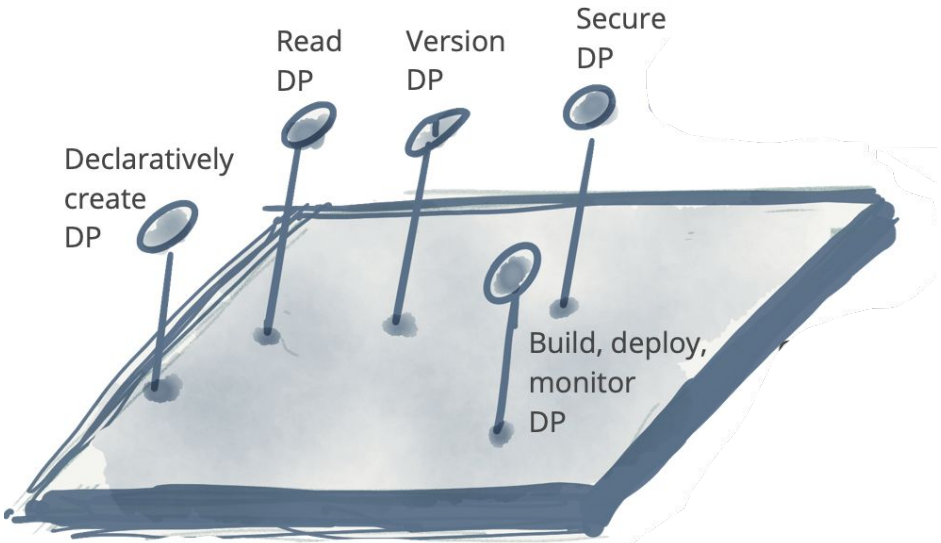


A Self-serve Platform Plane

A group of related capabilities provided to other planes of the platform

Logical Architecture:

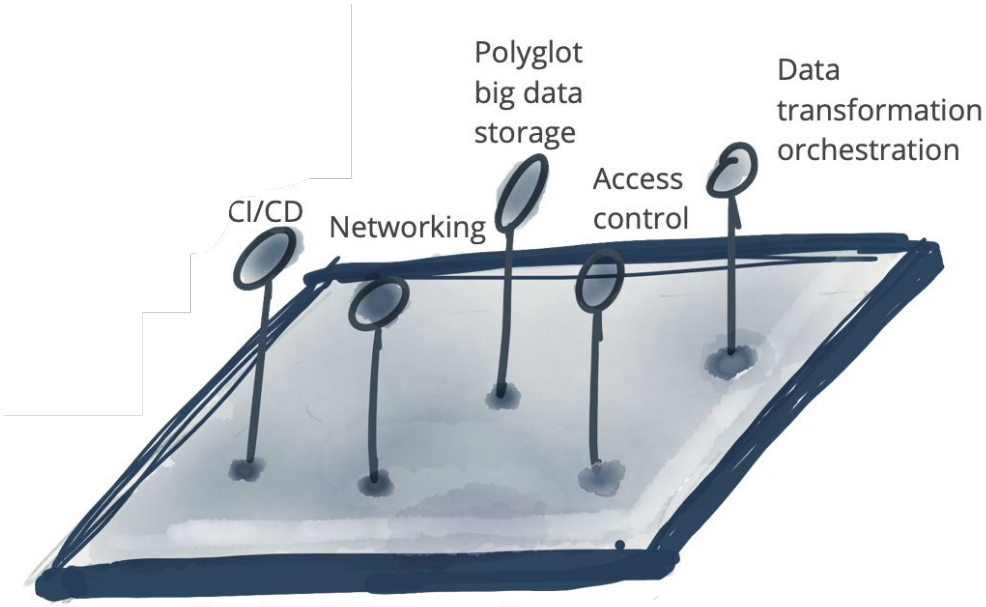
Data Product Developer Experience Plane



Data Product Developer Experience Plane

The higher level abstraction of data infrastructure designed to support the common data product developer journey

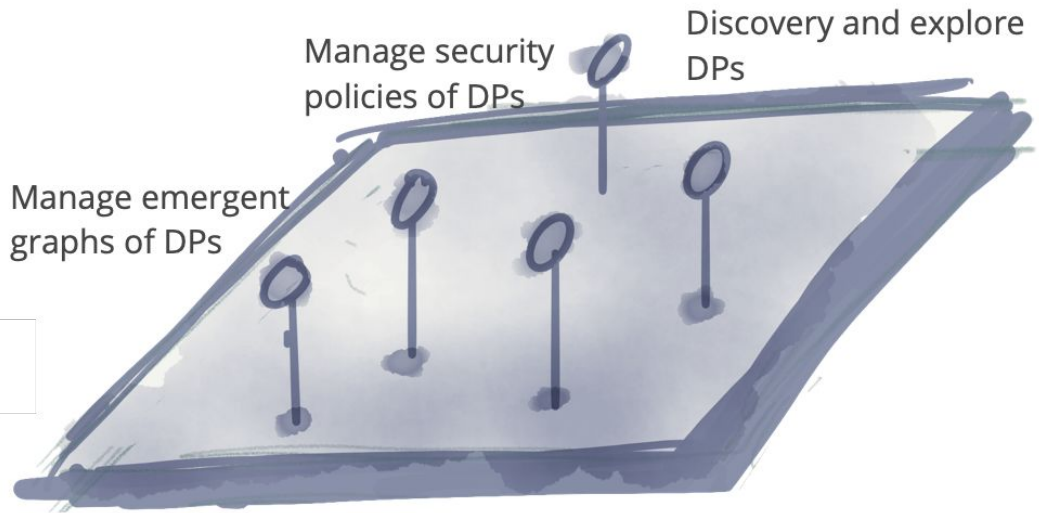
Logical Architecture:
Data Infrastructure Plane



Data Infrastructure Plane
Providing the underlying infrastructure required to build, run, monitor data products

Logical Architecture: Mesh Supervision Plane

Mesh Supervision Plane
Capabilities that are accessible more conveniently at mesh level



Summary

Self-serve data infrastructure as a platform

So that the domain teams can create and consume data products autonomously using the platform abstractions, hiding the complexity of building, executing and maintaining secure and interoperable data products.

A NEW WORLD ORDER

BUILD AN ECOSYSTEM INTELLIGENCE

With a federated computational governance

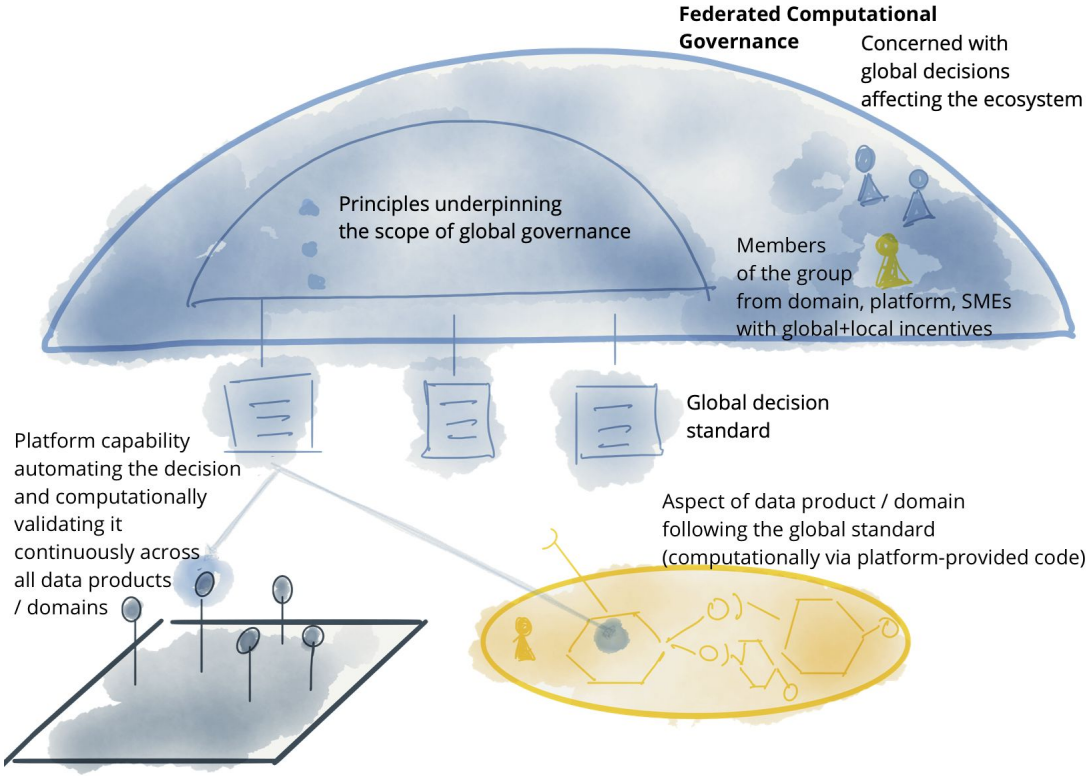
Embrace:

- **Decentralization** and domain self-sovereignty,
- **Interoperability** through global standardization,
- A dynamic topology
- **Automated execution** of decisions by the platform.

“Placing a system in a straitjacket of constancy can cause fragility to evolve.”

-- C.S. Holling, ecologist

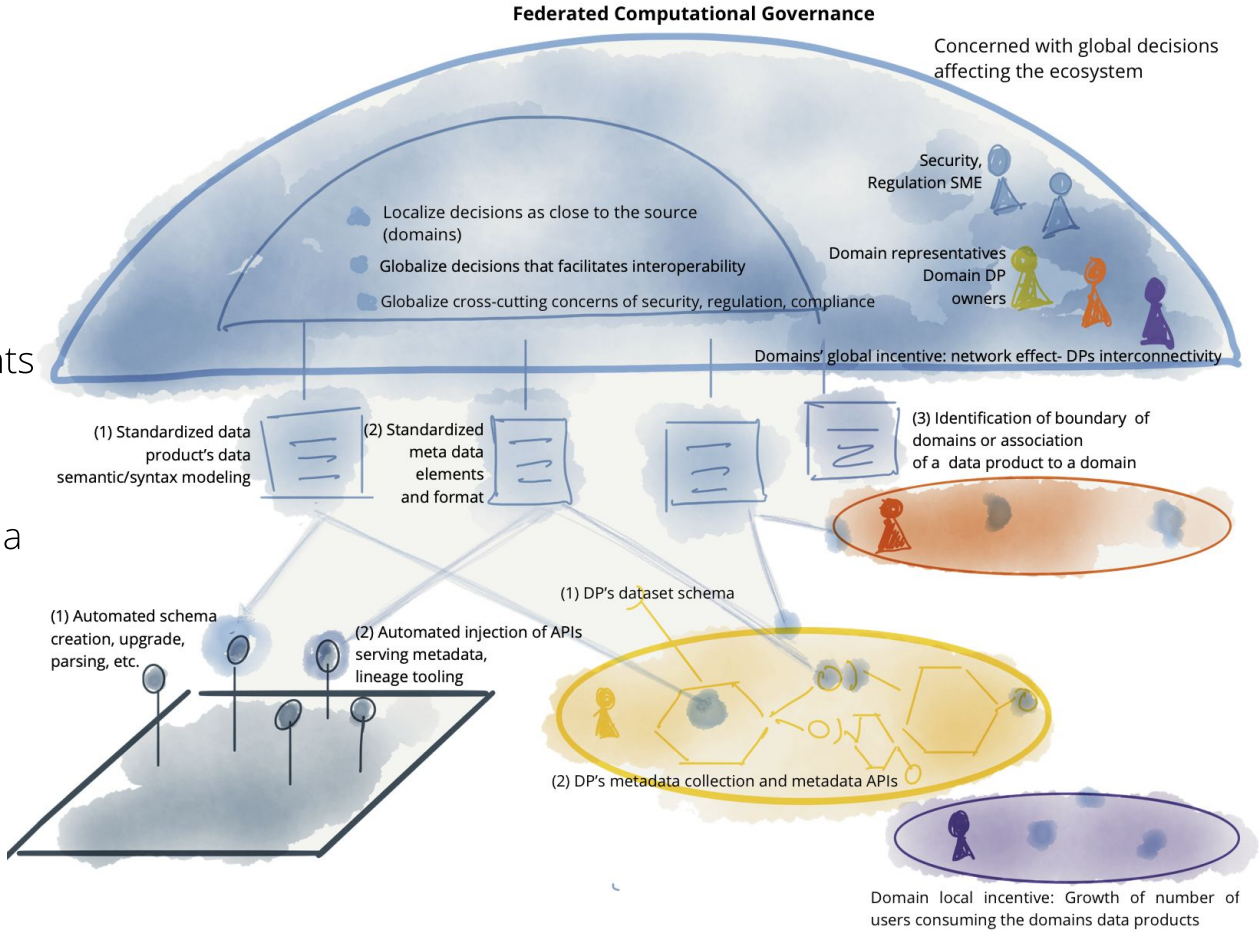
Logical Architecture
computational policies embedded in the mesh



Example

Concerns

- Standardize data product's semantic /syntax modeling
- Standardize metadata elements and format
- Identification of domains and allocation of data products to a domain
- Security, compliance

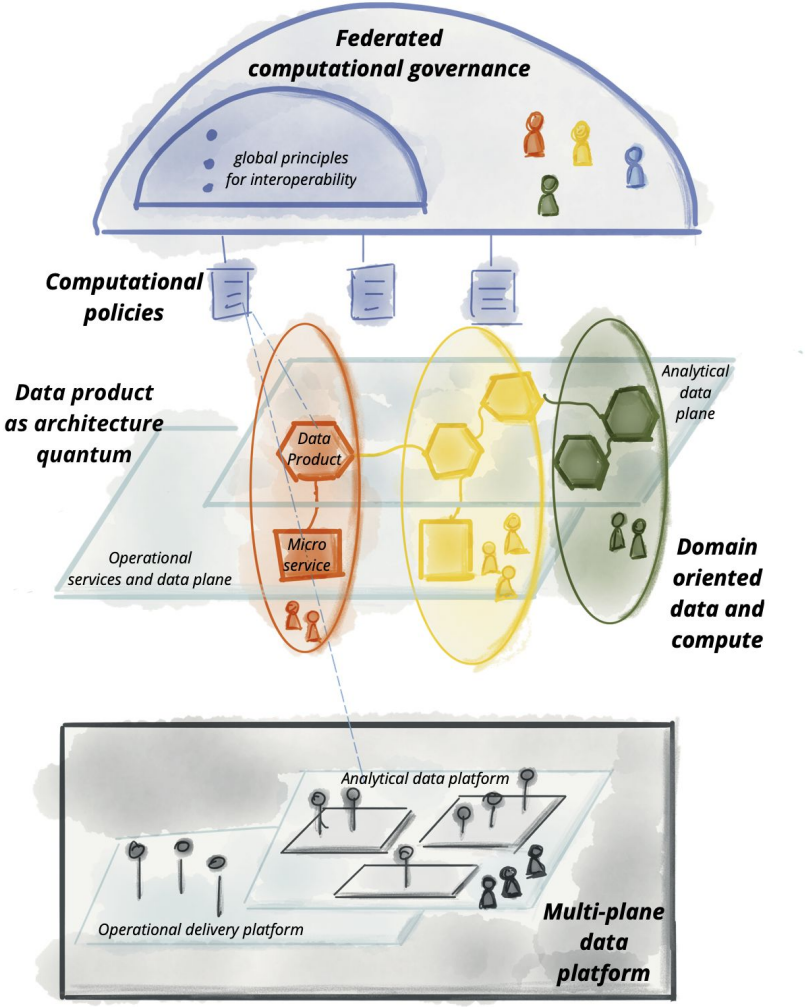


PRE DATA MESH GOVERNANCE	DATA MESH GOVERNANCE
Centralized team of data experts	Federated team of domain data owners and SMEs
Responsible for data quality	Responsible for defining how to model what constitutes quality
Responsible for data security	Responsible for defining aspects of data security i.e. data sensitivity levels for the platform to build in and monitor automatically
Responsible for complying with regulation	Responsible for defining the regulation requirements for the platform to build in and monitor automatically
Responsible for global canonical data modeling	Responsible for modeling polysemes - data elements that cross the boundaries of multiple domains
Measure success based on number or volume of governed data (tables)	Measure success based on the network effect - the connections representing the consumption of data on the mesh

Summary

Federated computational governance

So that data users can get value from aggregation and correlation of independent data products - the mesh is behaving as an ecosystem following global interoperability standards; standards that are baked computationally into the platform.



Paradigm Shift in

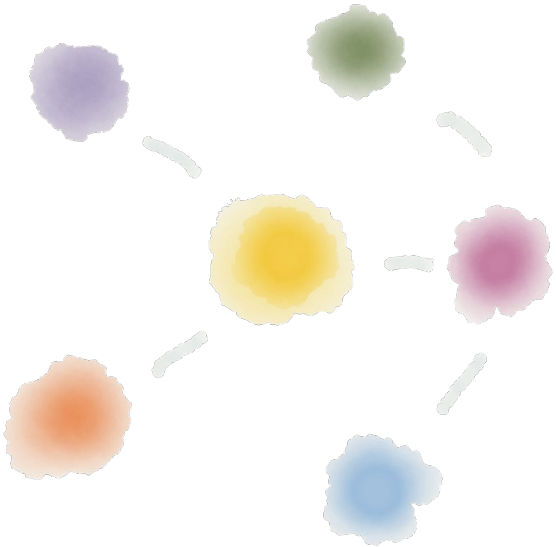
Architecture
Organization
Technology

“A different language is a different vision of life”

- Federico Fellini

A New Language

FROM	TO
IGESTING	SERVING
EXTRACT | LOAD | ONBOARD	DISCOVER | CONSUME | LINK
FLOW DATA THROUGH PIPELINES	PUBLISH DATA VIA DATA PRODUCT'S PORTS
CENTRALIZED LAKE | WAREHOUSE | PLATFORM	ECOSYSTEM OF DATA AS PRODUCTS



THANK YOU!

Zhamak Dehghani
@zhamakd
zdehghan@thoughtworks.com