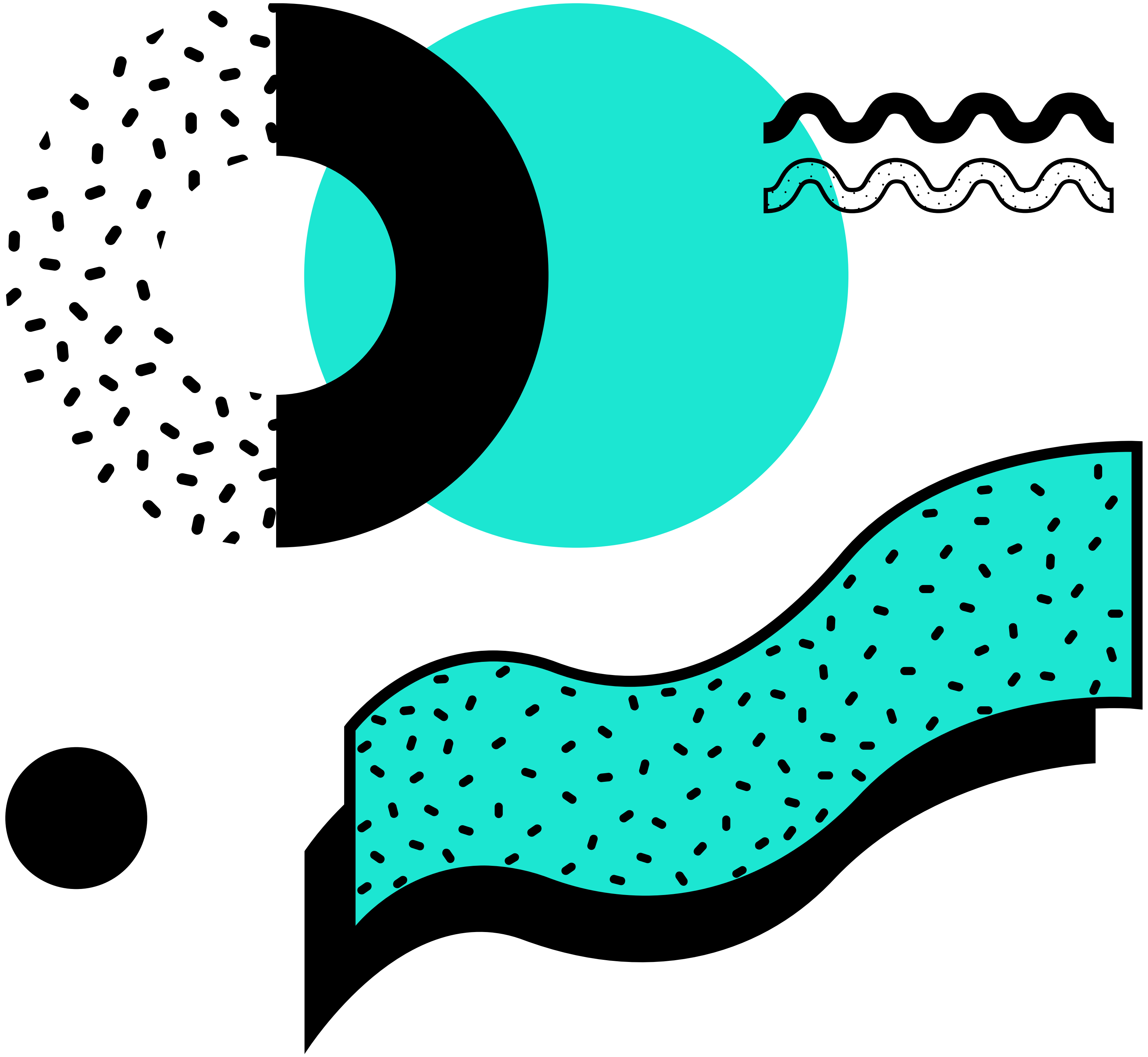


DATA MESH

DATA AS A PRODUCT

A Domain Centric Data Solution

Author: Ajit Dash

INTRODUCTION

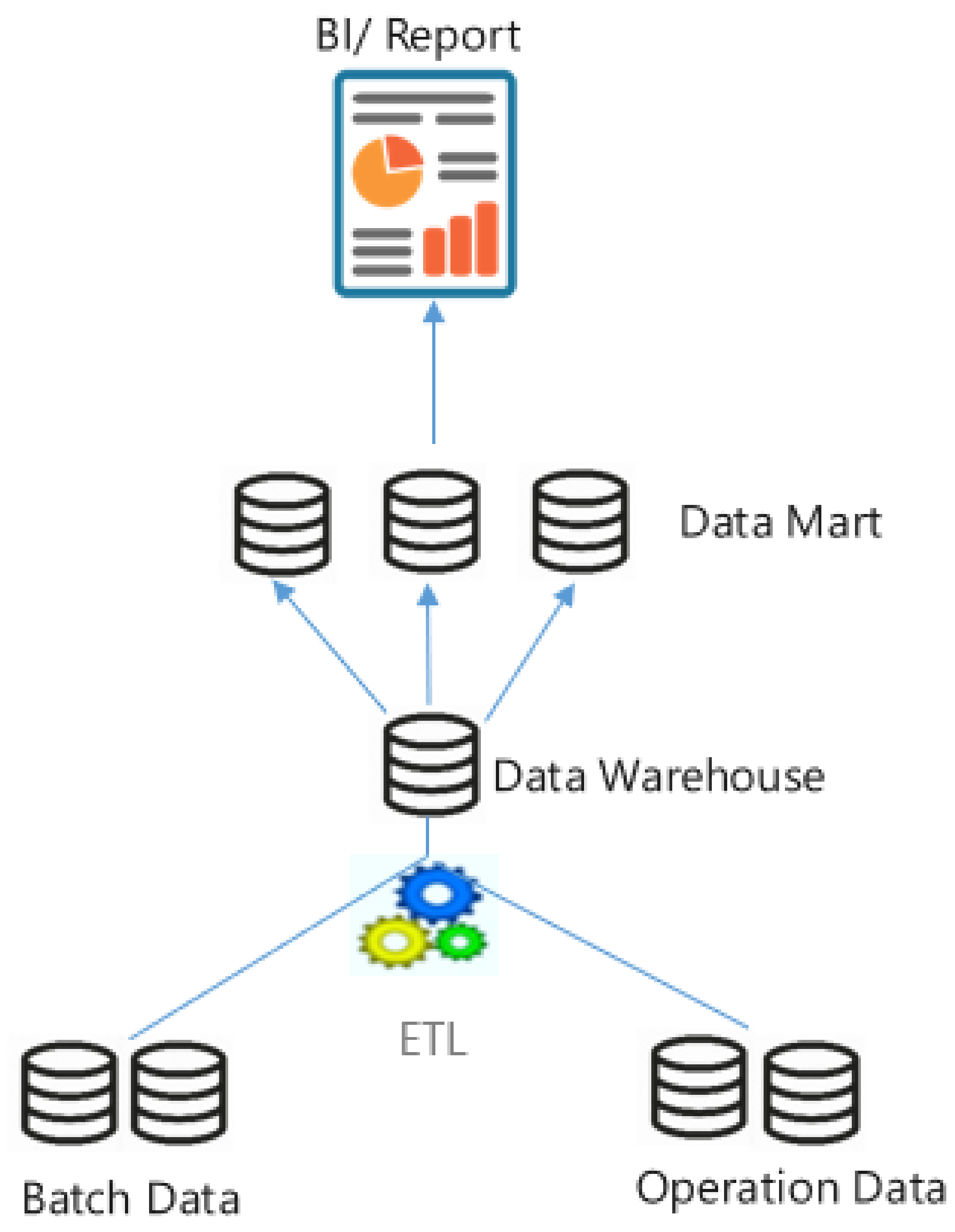
This whitepaper discusses the challenges and limitations with the Data Warehouse(summarized data), Data Lake (a centralized data store) & Data Fabric (common format data) concerning the data quality, real-time data processing, common data format. Provides some insight about how centralized data stores create isolation among source providers, data developers, end-users, etc. Also, discussed in detail how to overcome the limitations and challenges, by a Data Lake / Data Warehouse/Data Fabric by using a domain-centric data solution such as a "Data Mesh " (decentralized data store) which delivers "Data As a Product ". Further, it provides details about the "Date Mesh" team structure, the role of a product owner & how to determine your company needs a Data Mesh or not.

TOPICS

1. Introduction
2. Data Warehouse
3. Data Lake
4. Data Fabric
5. Data Mesh
6. Compare Data Warehouse,Data Lake, Data Fabric & Data Mesh
7. Pain Points
8. Competitors Approach & Solution
9. Data Mesh in Detail
10. Data Mesh Team & Domain
11. Team Structure & Comparision
12. Evaluate To Data Mesh or Not to Data Mesh
13. Use Cases -Delivery Services & Defence Manufacturer
14. Data Mesh integrated to Lake House

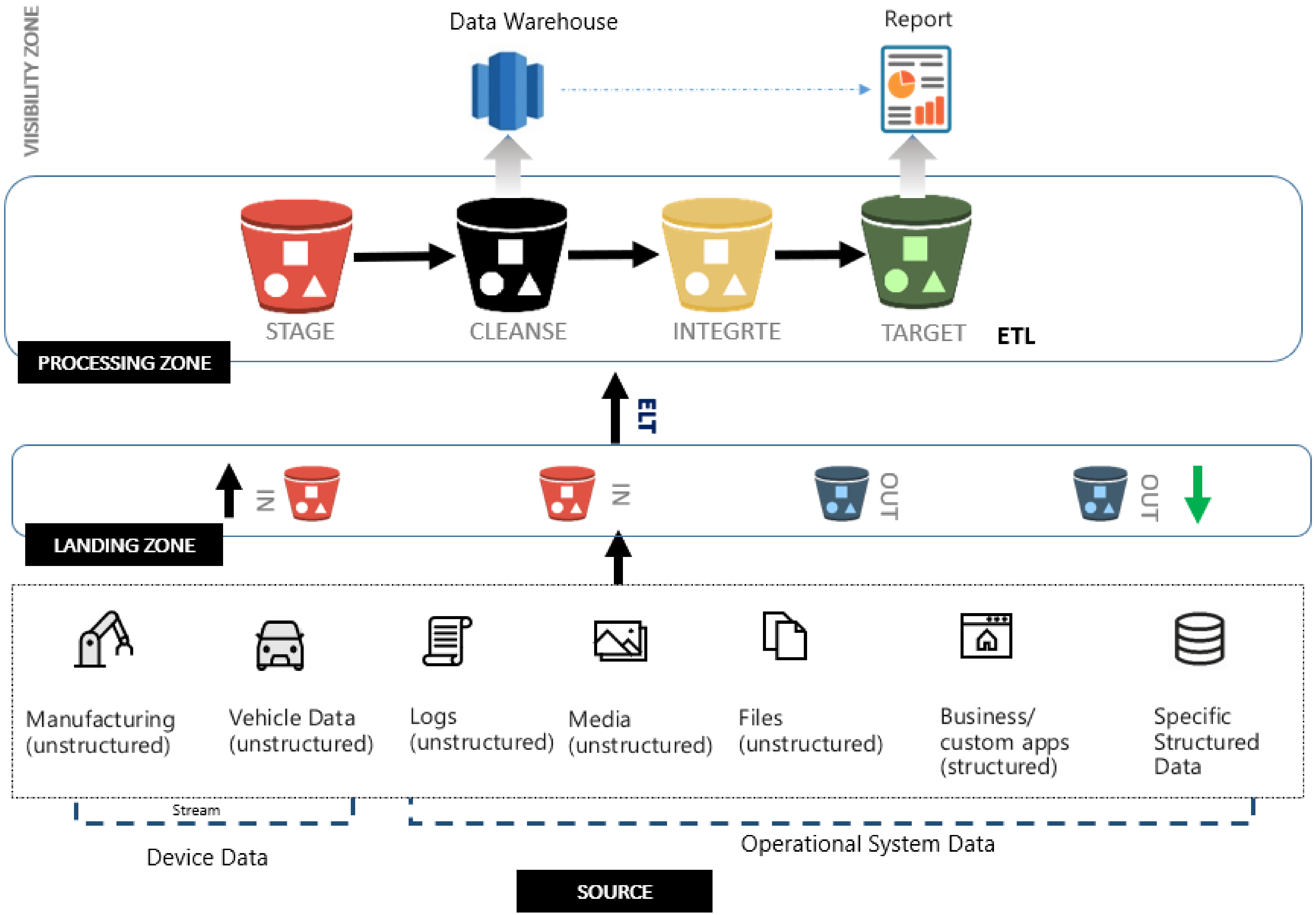
DATA WAREHOUSE (SUMMERIZED DATA)

A Data Warehouse is a storage architecture designed to store the data extracted from the source systems (operational or transaction). Usually, it stores data in summarized (aggregated)format either in a Star, Snowflake, or a Hybrid format as per the business need. The data from a Data Warehouse used for reporting purposes (BI reporting)



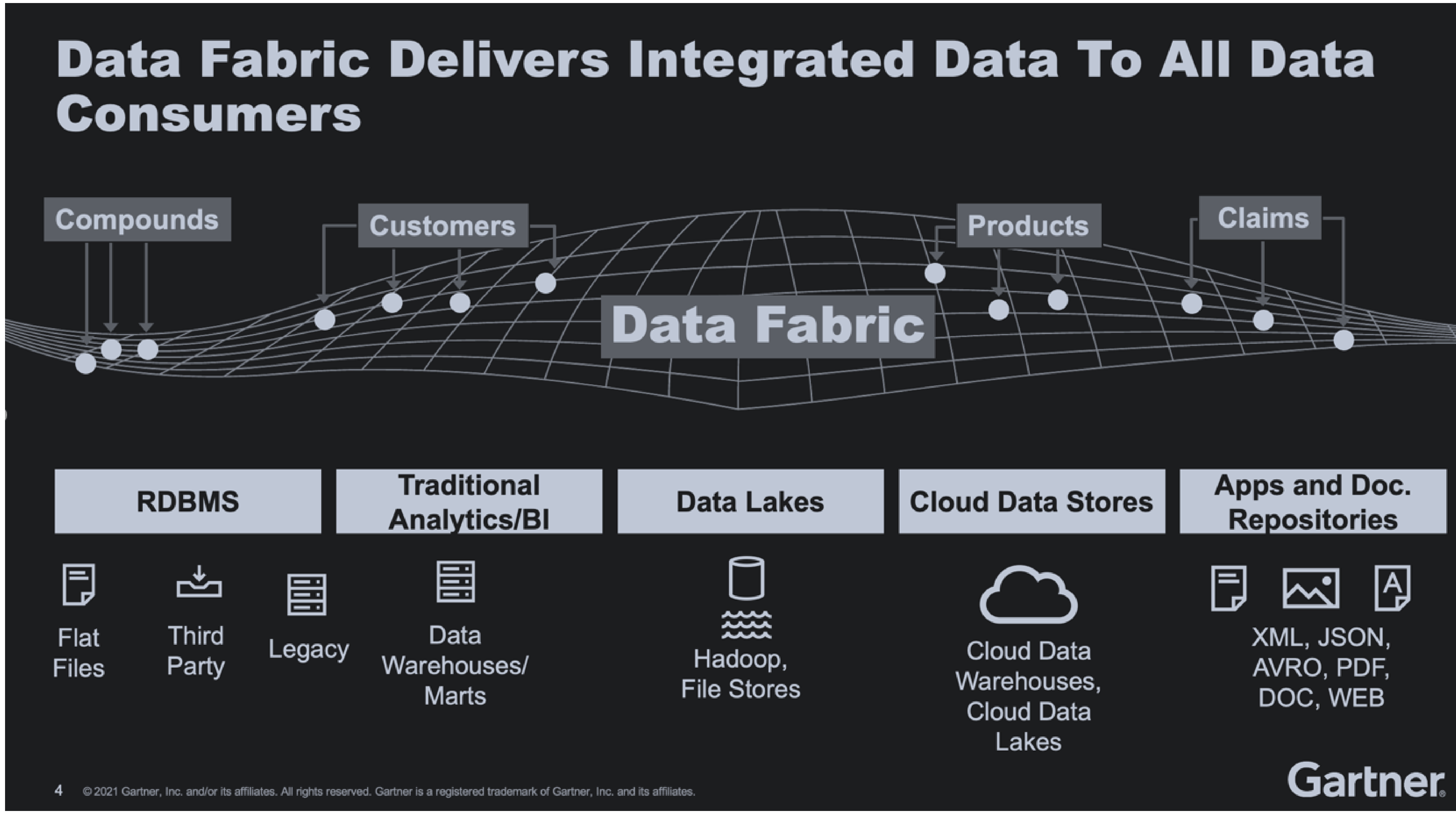
DATA LAKE (CENTRALIZED DATA STORE)

A Data Lake stores the data instances from various data sources in centralized storage. Source data could be of any of these formats unstructured, semi-structured, or structured. Data in a Data Lake is usually stored in the raw formats as it's (being pushed by or pulled) from the source systems.



DATA FABRIC (COMMON FORMAT DATA)

Data fabric provides a unified data architecture. It stores the source data in its storage in a common consumable format irrespective of the source systems. It supports both the operational as well as analytical data. All the common format data are available to the consumers as per their needs (operation or analytical AI/ML etc) for downstream consumptions. All the formatted data follow the standard governance and privacy policy.



Data Fabric Delivers Integrated Data To All Data Consumers

Compounds

Customers

Products

Claims

Data Fabric

RDBMS

Traditional Analytics/BI

Data Lakes

Cloud Data Stores

Apps and Doc. Repositories

Flat Files

Third Party

Legacy

Data Warehouses/ Marts

Hadoop, File Stores

Cloud Data Warehouses, Cloud Data Lakes

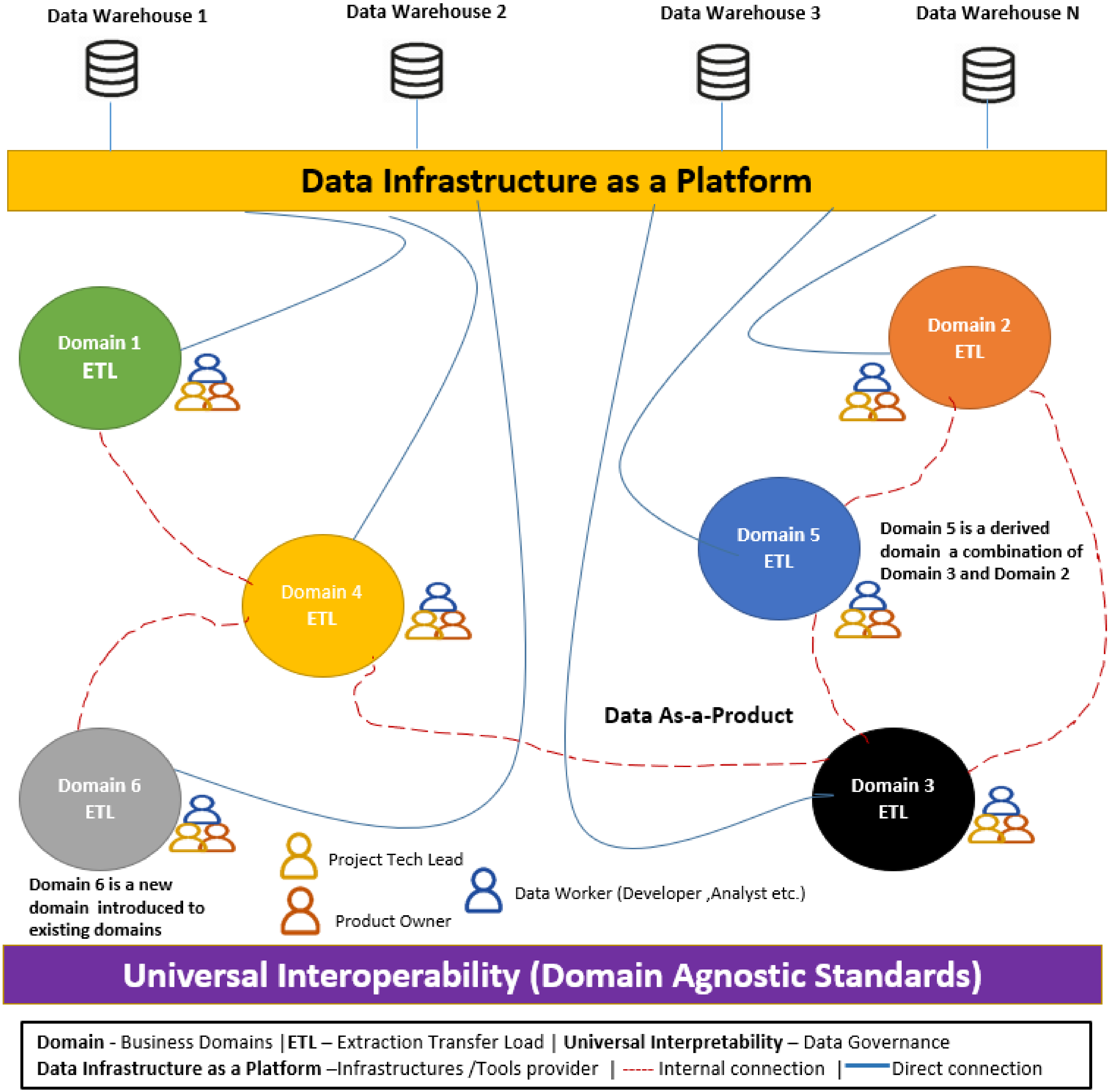
XML, JSON, AVRO, PDF, DOC, WEB

4

Gartner

DATA MESH (DECENTRALIZED DATA)

Data mesh is a domain-centric distributed architecture it provides domain-specific data to each consumer. It represents "Data As a Product".Unlike the monolithic traditional data systems consume data from various sources using the ETL process and store it in a centralized location (Like a Data Lake). But in Data mesh each domain owner handles their data (Product Owner, Developer, and Consumer) specific to their domain. Each specific domain is governed through the domain agnostic universal data governance standards.


Data Warehouse 1
Data Warehouse 2
Data Warehouse 3
Data Warehouse N
Data Infrastructure as a Platform
Domain 1 ETL
Domain 2 ETL
Domain 4 ETL
Domain 5 ETL
Domain 5 is a derived domain a combination of Domain 3 and Domain 2
Data As-a-Product
Domain 6 ETL
Domain 3 ETL
Domain 6 is a new domain introduced to existing domains
Project Tech Lead
Product Owner
Data Worker (Developer ,Analyst etc.)
Universal Interoperability (Domain Agnostic Standards)

Domain - Business Domains |**ETL** – Extraction Transfer Load | **Universal Interpretability** – Data Governance
Data Infrastructure as a Platform –Infrastructures /Tools provider | ----- Internal connection | ▬ Direct connection

COMPARE

DATA WAREHOUSE
DATA LAKE
DATA FABRIC
&
DATA MESH

PROS
Data Warehouse

- Good for summarized data
- Usually used for BI Reporting
- Stored data in Star, Snowflake, or Hybrid format
- Mostly dimensional modeling
- The major process involves ETL

PROS
Data Lake

- Stores Raw data from various sources
- Centralized storage system
- Keeps Structured, Unstructured, and Semi-structured data
- Complex data query support, across structured and unstructured data

PROS
Data Fabric

- Full SQL Support
- Distributed data stores support linear scalability
- Support high concurrency with real-time performance
- Work with Data Mesh
- Semi Collaboration among Product owners, developer & consumer

PROS
Data Mesh

- Decentralized storage and domain-centric approach
- Provide "Data-As Product "
- A universal standard for data governance
- Reliable data quality
- Well collaboration among product owner, developer & consumer

CONS
Data Warehouse

- Complex process
- Dependency of the ETL jobs success
- Data quality is a bigger issue
- High maintenance
- Isolation among source provider, Developer, and Consumer

CONS
Data Lake

- Not optimized for single entity queries
- Data quality is a bigger issue
- Live data is not supported, continuous updating data is either unreliable or increase delivery time
- Isolation among source provider, Developer, and Consumer

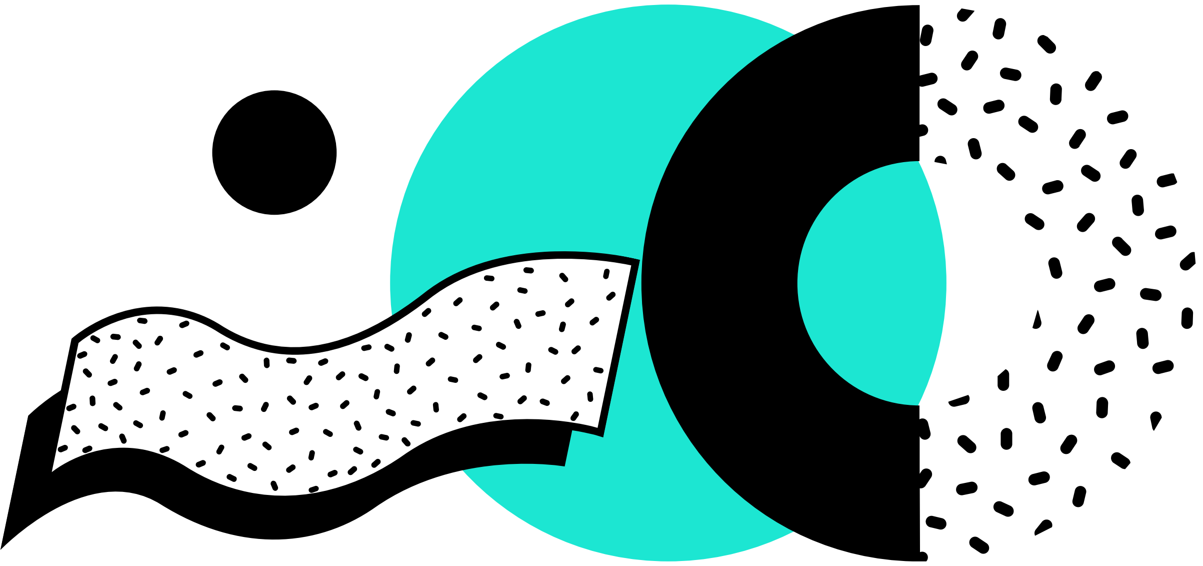
CONS
Data Fabric

- Challange with converting data sources to a common format
- High maintenance
- Isolation among source provider, Developer, and Consumer

CONS
Data Mesh

- Thorough knowledge of the domain helps in the successful delivery
- Team collaboration is very important
- Data governance policy need to be well defined

PAIN POINTS



Problem 1

Data Warehouse designed for the summarized data

Data Lake is centralized storage for all kinds of source data such as structured, unstructured, or semi-structured. Data lake stores the raw data from the source system. Data Lake and DWH are not domain-centric.

Problem 2

In a Data Warehouse or a Data Lake, Source providers, Developers, and consumers are not well connected they work most of the time independently. They have minimal interaction. There is no proper well defined product ownership.

Problem 3

Data Warehouse or Data Lake have limitations with data quality as the don't have universal data standards or governance

COMPETITOR APPROACH & SOLUTION

DATA MESH

Approach 1

In a domain-centric solution where Product Owner, Developer, Source providers, and consumer are well connected to deliver a Domain centric data solution "Data As a Product". The product owner is responsible for communication among the team members as well to the data consumer and responsible for the overall quality delivered

Approach 2

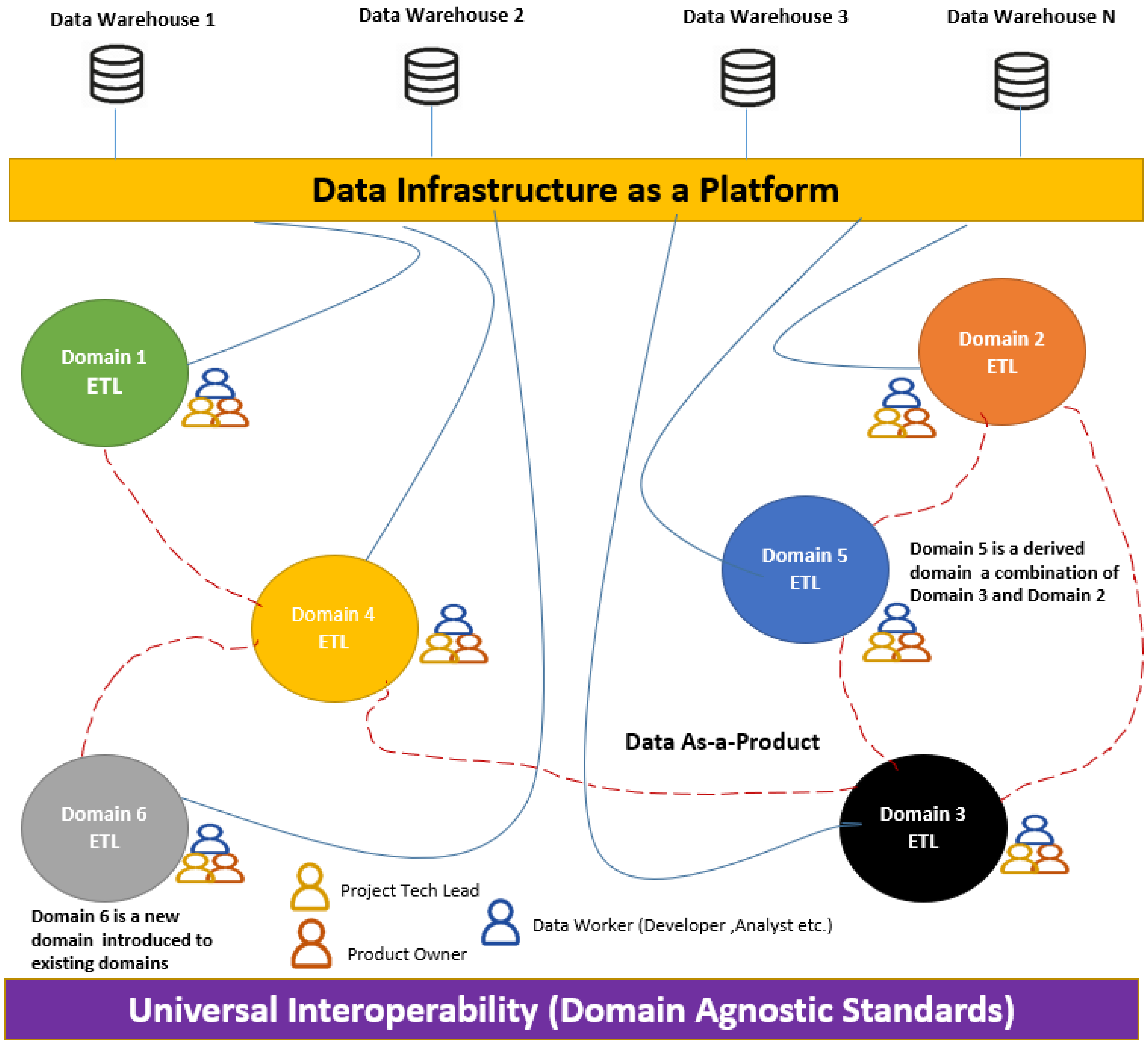
Implement Universal data standards for data governance and quality irrespective of types of domain - Universal Interoperability - Domain Agnostic

Approach 3

Easy addition of new domains to the existing framework and communication among the other domains for data exchange and data interoperability.All the above approaches are possible through a domain-centric solution "DATA MESH"

DATA MESH-IN DETAIL

- Domain 1 to Domain N - Individual domains or line of business
- ETL -Extraction -Transfer -Load -This load data from the source system and provides fine-tuned data for downstream consumptions
- Data Infrastructure as a platform is a gateway for the downstream data consumption such as Data Warehouse (Data Warehouse 1 to Data Warehouse N). This plays a major role in data communication and it provides desired tools, technologies, etc irrespective of domains.
- Gray solid lines are connected paths between the data platform and domain this could be an ETL /APIs or Connectors
- Red dotted lines are the communication path among the Domain
- Universal interoperability provides data standards and data governance respective of the domain. Its final product to consumers **Data-As-A-PRODUCT**



Domain 5 is a derived domain a combination of Domain 3 and Domain 2

Domain 6 is a new domain introduced to existing domains

Domain - Business Domains |**ETL** – Extraction Transfer Load | **Universal Interpretability** – Data Governance
Data Infrastructure as a Platform –Infrastructures /Tools provider | ----- Internal connection | ▬ Direct connection

DATA MESH TEAM & DOMAIN

DATA MESH TEAM

Product Owner

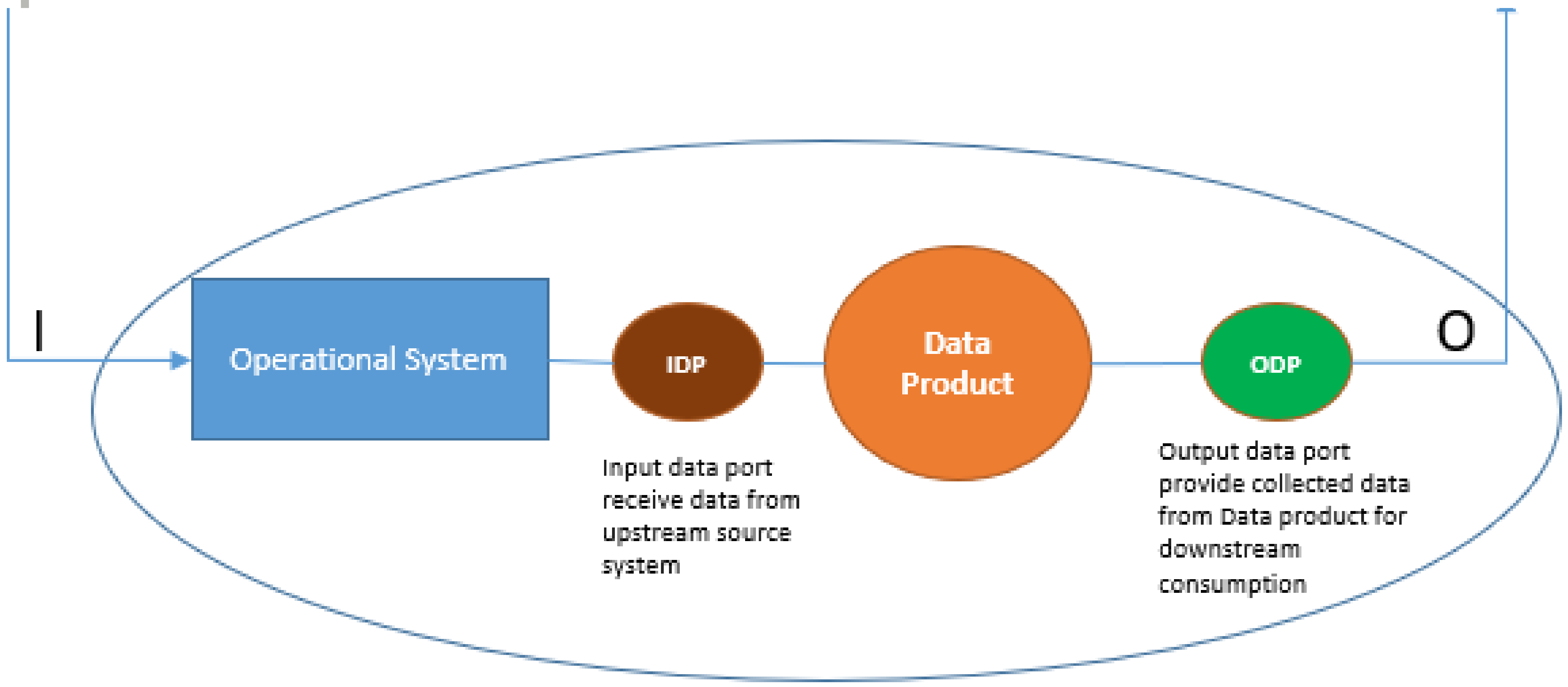
- A domain data product owner is a lead person whose main focus is to deliver "Data As a Product". Also, make sure the product is meeting all the quality standards and have decreased the lead time of the source consumptions
- The product owner must have a deep understanding of the end-users requirements and be well-aligned with the usage of the data by the user
- A product owner responsible for communication within the team as well as the end-user

Team

- A Data mesh team is a combination of Product Owner, Developer, Data, and Business Analyst, Data Scientist, etc

DOMAIN

- A domain consisting of the Input data port, Data Product, and Output data product
- Input Data Port: Receive the data from upstream data port and it processes these data and provides to the Data Product
- Output Data Port: Collect data from Data Product and make it available for downstream consumption

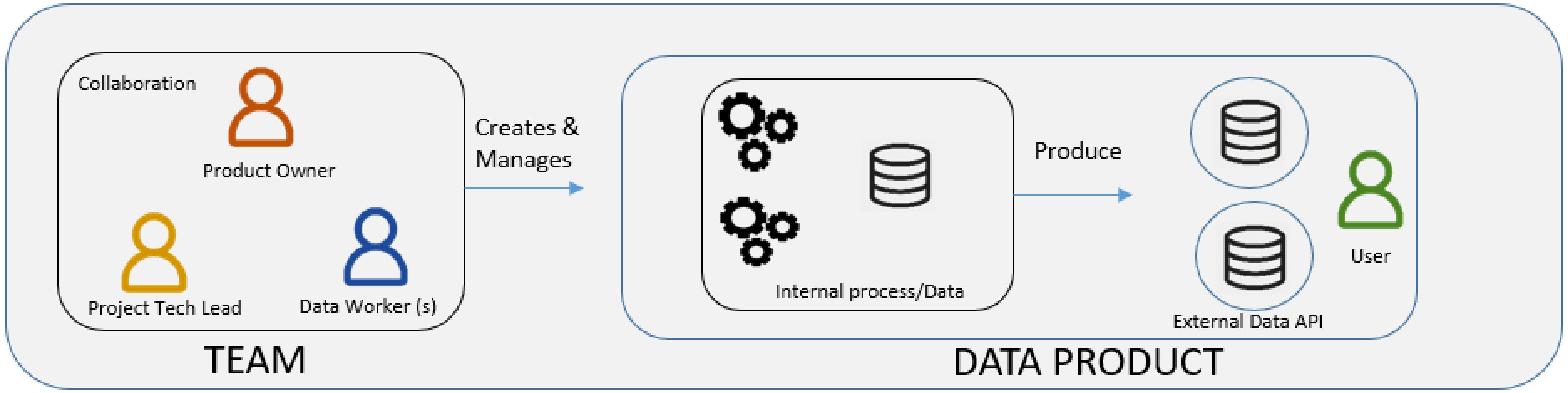

I
Operational System
IDP
Data Product
ODP
O
Input data port receive data from upstream source system
Output data port provide collected data from Data product for downstream consumption

TEAM STRUCTURE COMPARISION

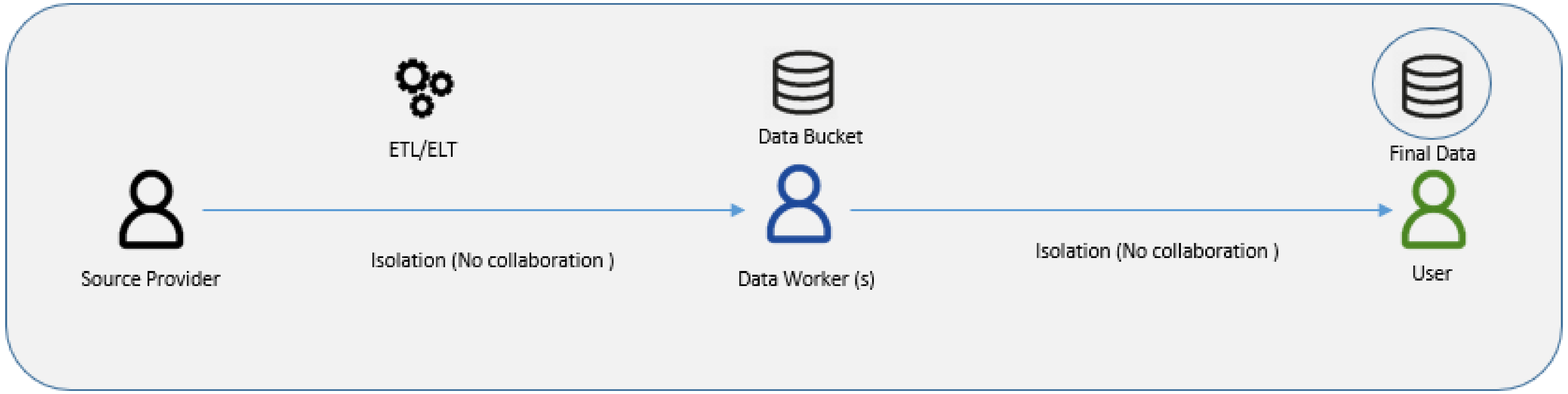
Data Mesh team consists of :

- Project lead / Team Leader: Lead all team activities
- Product Manager or Product owner delivers "Data As a Product "
- Data Worker(s): Developers, Analyst, etc,

The product owner communicates the information within the team as well as with the source provider and with end-users



DATA MESH - BUSINESS DOMAIN



DATA LAKE TEAM

- Stores Raw data from various sources
- Centralized storage system
- Keeps Structured, Unstructured, and Semi-structured data
- Complex data query support, across structured and unstructured data
- No Collorabotation among source provide,Data Workers and end-users

HOW TO EVALUATE YOUR COMPANY NEED A DATA MESH : TO MESH OR NOT TO MESH

Number of Data Sources (Score 1 to 10)

Total numbers of data sources your company have ?

Data Team Size (Score 1 to 10)

Total number (s) of Product manage(s) , Developer(s) and Analyst(s) in your team

Total Number (s) of Data Domain (Score 1 to 10)

Total number (s) functional teams (marketing, sales, finance etc. total numbers of products your company have and utilize for decision making

Data Engineer Dependency (Score 1 to 10)

How many times data engineers are bottle neck for the new implementation or any change

Data Governance and Quality (Score 1 to 10)

How important for your company to have quality data or data standards

Score 1 Low 10 High

Here is a simple calculation add scores of each category score it 1 to 10 and calculate total

- 30 or above strongly recommend a Data Mesh
- 20 to 30 it is recommended
- 1 to 20 may think in future

DELIVERY SERVICE -USE CASE 1

Business Model: ABC delivery service provides on-demand auto service, users request an auto service either going through the app or website. The driver receives the request and avails the service to the customer.

New Business: ABC delivery service now expanding its business for food and drugs (medication will be implemented after food) delivery.

Future Business: It has a plan to add in few years Air taxi service where the customer can order an Airtaxi to travel by Air

Pain Points :

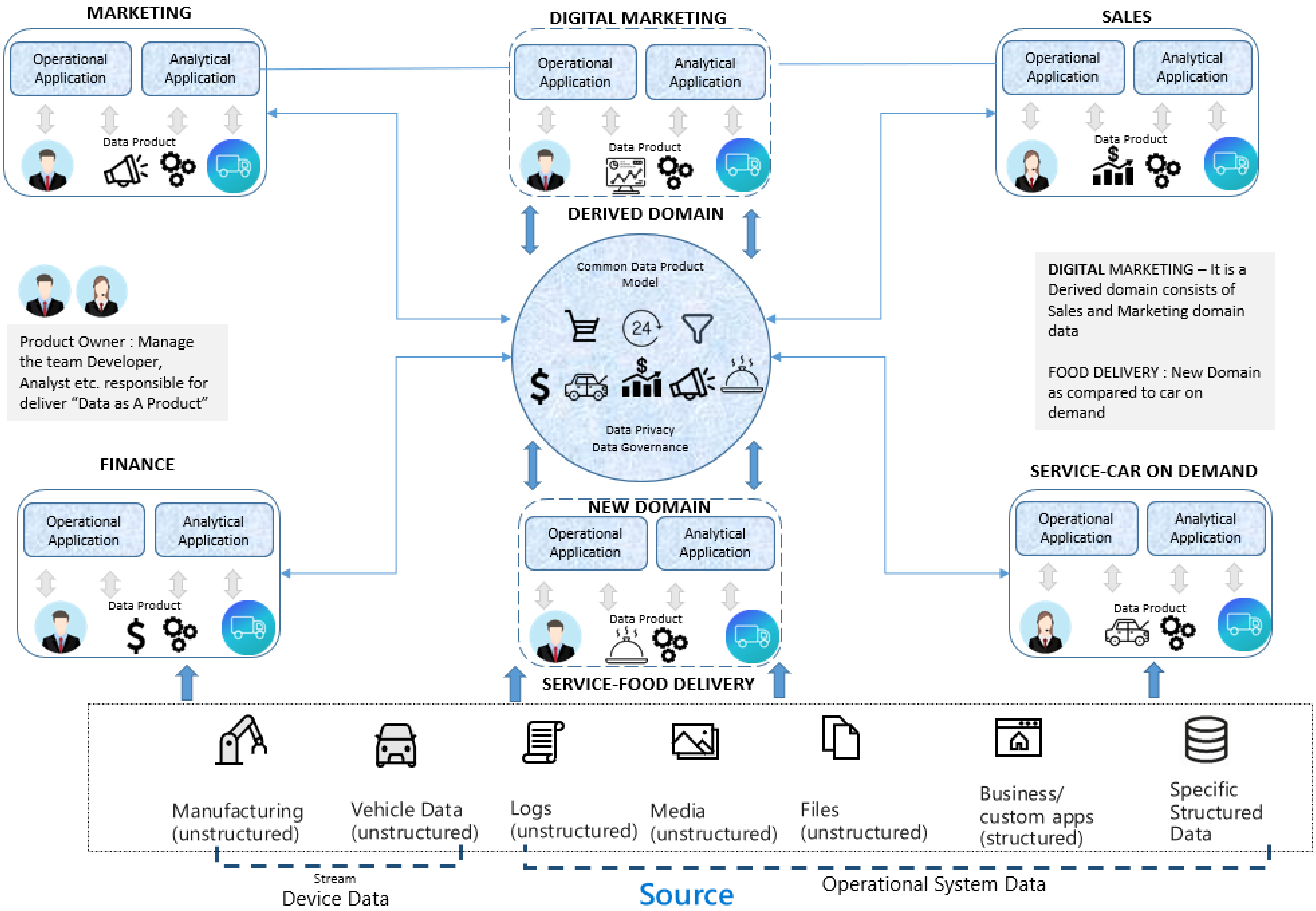
- Existing car service able to meet the new business need or not
- Existing car services have issues with data quality and delivery issues through a data lake
- Time for delivery and product ownership is not well defined and data delivered are not well enough to meet the data quality standards
- The issue with meeting the Time to Market, Time to Customer, and Time to Business
- Addition of new business line Food and Drugs a challenge as these are a different product line
- Food delivery has different guidelines and challenges
- Drugs delivery have different guidelines involving HIPPA and privacy guidelines

Solution Proposed :

- Understand the business goal
- Separate all the business domain
- Assigned a product owner for data quality control and to meet customer requirements
- Find out the common model and create a newly derived model
- New domain for Food delivery (Later a new domain for Drugs delivery will be added)

Business Domain: Marketing, Sales, Finance, Derived Domain Digital marketing, New Domain On-Demand Food Delivery

ON-DEMAND DELIVERY SERVICES



Product Owner : Manage the team Developer, Analyst etc. responsible for deliver “Data as A Product”

DIGITAL MARKETING – It is a Derived domain consists of Sales and Marketing domain data

FOOD DELIVERY : New Domain as compared to car on demand

DEFENCE MAINFACTURER SERVICE -USE CASE 2

Business Model: XYZ is a defense manufacturer and has multiple product lines such as Assult Tank, Anti mine vehicle, etc.

New Business: XYZ is adding a product line assault helicopter

Future Business: It has a plan to add in few years new product line related to air services

Pain Points :

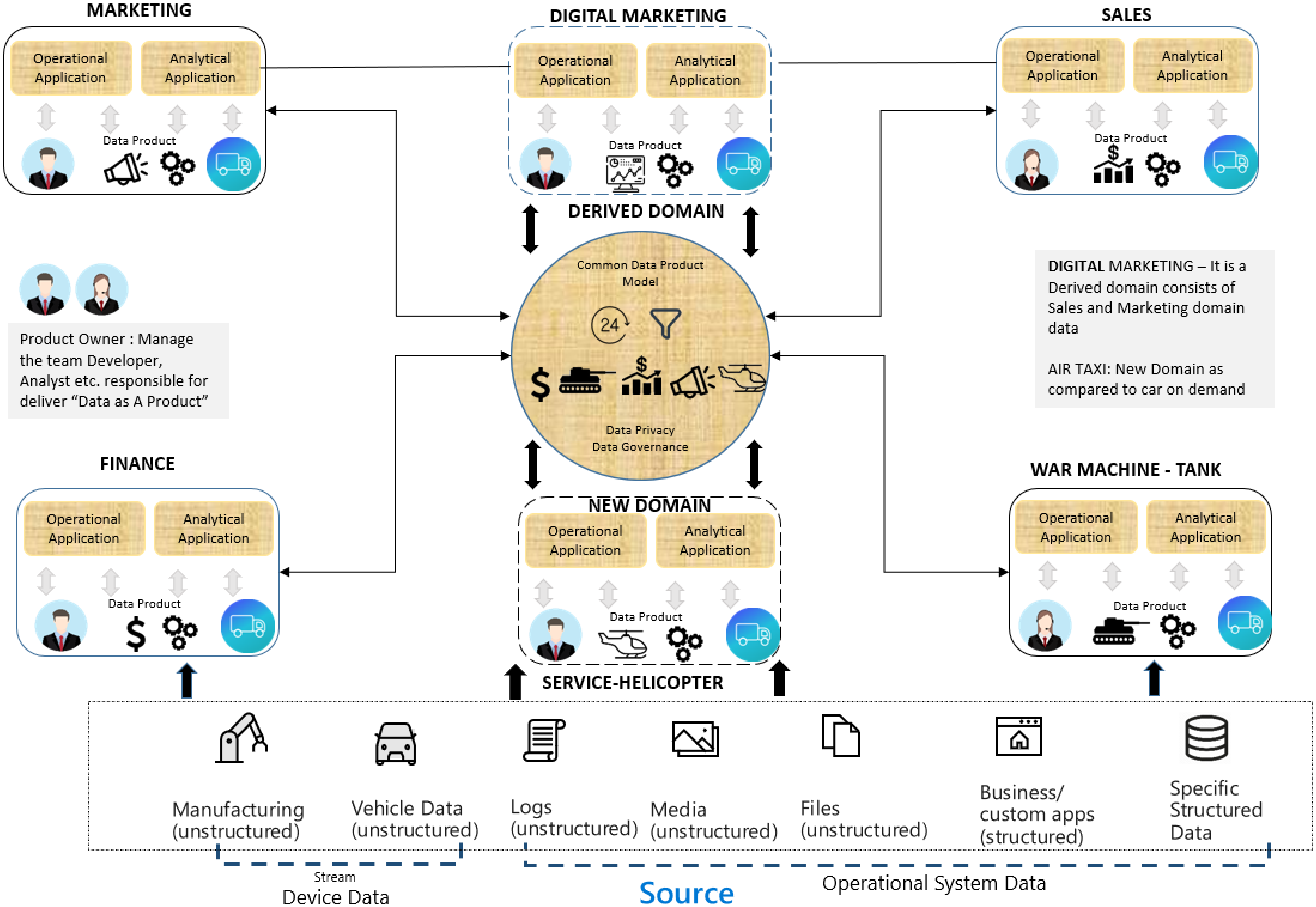
- Existing product line able to meet the business need or not
- The existing business line has issues with data quality and delivery issues through a data lake
- Time for delivery and product ownership is not well defined and data delivered are not well enough to meet the data quality standards
- The issue with meeting the Time to Market, Time to Customer, and Time to Business
- The addition of new business is a challenge as these are a different product lines (air services)
- Air product line manufacture has different guidelines and challenges

Solution Proposed :

- Understand the business goal
- Separate all the business domain
- Assigned a product owner for data quality control and to meet customer requirements
- Find out the common model and create a newly derived model
- New domain for air manufacturer

Business Domain: Marketing, Sales, Finance, Derived Domain Digital marketing, New Domain air product manufacturer

DEFENCE PRODUCT MAINFACTURER



MARKETING

Operational Application

Analytical Application

Data Product

DIGITAL MARKETING

Operational Application

Analytical Application

Data Product

SALES

Operational Application

Analytical Application

Data Product

DERIVED DOMAIN

Common Data Product Model

24

Data Privacy
Data Governance

Product Owner : Manage the team Developer, Analyst etc. responsible for deliver “Data as A Product”

DIGITAL MARKETING – It is a Derived domain consists of Sales and Marketing domain data

AIR TAXI: New Domain as compared to car on demand

FINANCE

Operational Application

Analytical Application

Data Product

NEW DOMAIN

Operational Application

Analytical Application

Data Product

SERVICE-HELICOPTER

WAR MACHINE - TANK

Operational Application

Analytical Application

Data Product

Manufacturing (unstructured)

Vehicle Data (unstructured)

Logs (unstructured)

Media (unstructured)

Files (unstructured)

Business/ custom apps (structured)

Specific Structured Data

Stream
Device Data

Source

Operational System Data

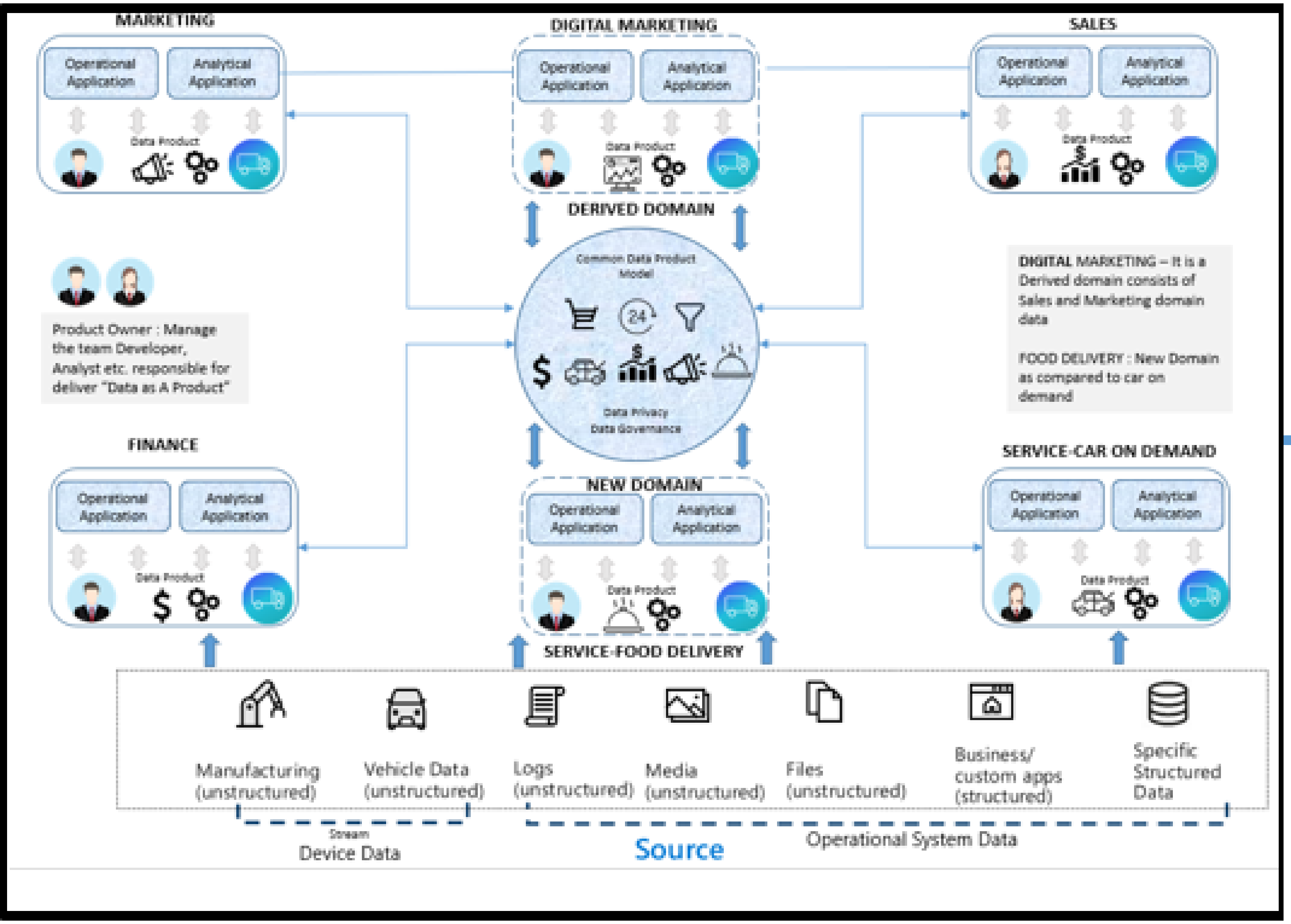
Integration data mesh data to data lake/lake house/ delta lake /data fabric

- Domain centric data which is already processed with business rules for downstream consumptions redly available in the Data lAKE/Lake House
- No need for more quality check
- It could be integrated with the other data source in the centralized data storage like Data for the better data insight
- More information could be derived by integrating domain data with other data source's data
- Information from Data Mesh is either pushed or pulled by using the ETL tool to the Data Lake
- As all the data part of the universal interoperability, governance process it's easy to manage and maintain in a Data Lake

Domain Data as a source for the Data Lake/ Lake House /Delta Lake
Domain Data: Operational and Analytical Data

ARCHITECTURE SAMPLE
DATA MESH WITH A DATA LAKE

Write an intriguing summary
of what your company does.

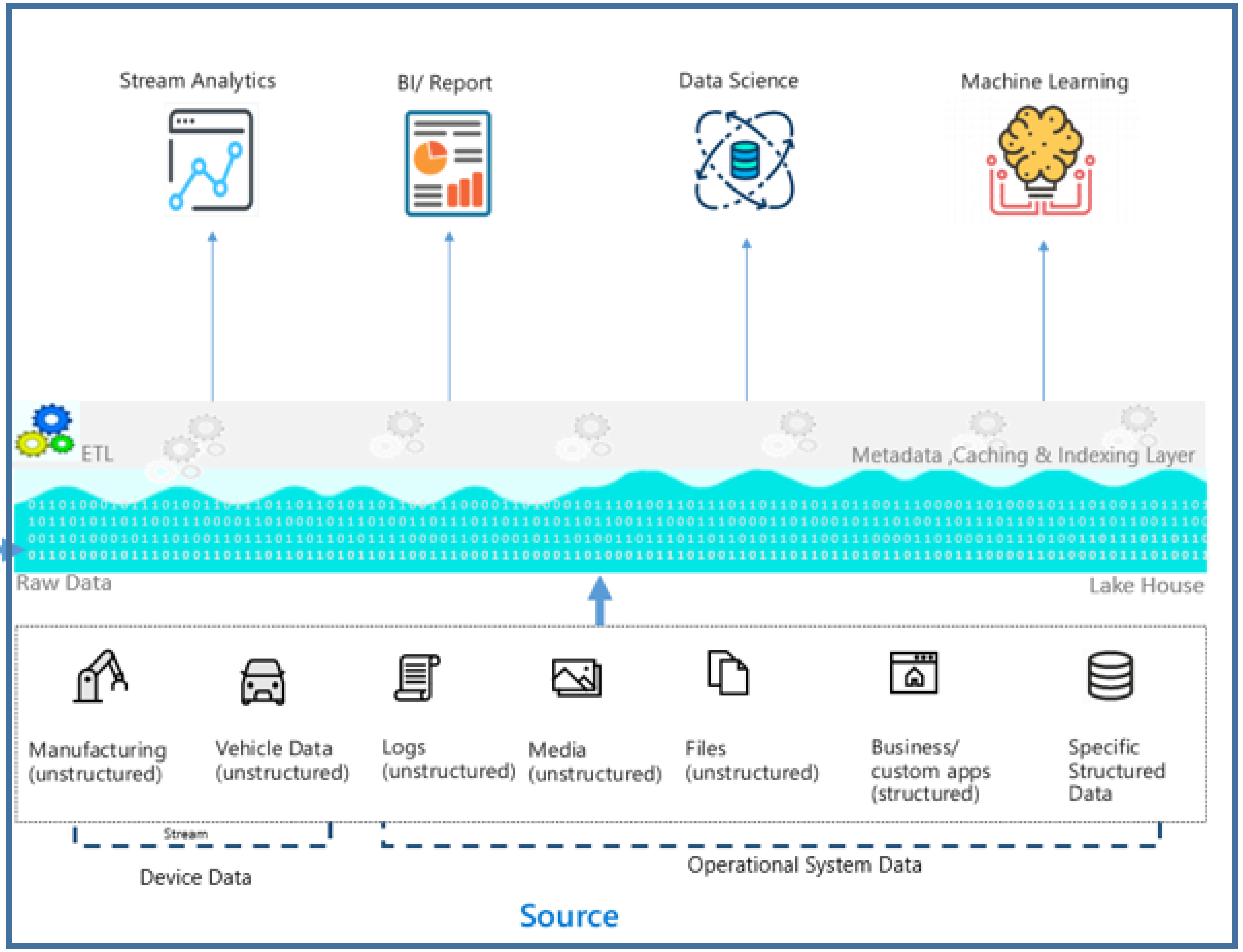


Product Owner : Manage the team Developer, Analyst etc. responsible for deliver "Data as A Product"

DIGITAL MARKETING – It is a Derived domain consists of Sales and Marketing domain data

FOOD DELIVERY : New Domain as compared to car on demand

FOOD & CAR ONDEMAND DATA MESH



LAKE HOUSE

APPENDIX

Companies Using Data Mesh

1. Intuit
2. Hello Fresh

ABOUT THE AUTHOR

Ajit Dash has spent more than 24+ years in data and analytics in various capacities, led various projects as a Sr. Director Data / Cloud Advisor/ Solution Architect / Cloudy data strategy manager/Advance Analytics/Data Scientist Lead, Pre Sales Lead providing Enterprise and Cross-platform integration solutions to various corporations.

His expertise includes strong hands-on analysis and design of Enterprise Solution Architecture, Cloud Advisor, Data Lake, Bigdata, Data Sc, Data warehouse and Data mining, Database Management/Integration, BI Reporting, and Cross-Platform

Domain Expertise: Telecommunication, Biotech, Finance, Banking, Media, Aerospace, Insurance and Technology, etc.: (Clients: Fox, Oshkosh, Otis, Travelers, Apple, Qualcomm, IBM, LPL Fin. etc.)

Education:

Ajit Dash holds a master's degree in General Management from Harvard University

Master's in Computer Information Systems from University of Phoenix

Masters in business Analytics (continuing) Georgia Tech

Bachelor's Degree in Electrical Engineering from India

Blog: http: //www.thedataworld.org