ARENADATA

Корпоративная платформа хранения
и обработки больших данных

Что такое Data Mesh и примеры реализации

Интро

Александр Ермаков

- 17+ лет в DWH & Data Engineering
- 35+ проектов по миру
- Участник Apache Software Foundation
- Участник Open Source Initiative
- CTO в Arenadata

О чем поговорим?

- Совсем коротко про Arenadata
- Текущие архитектурные подходы обработки данных
- Архитектура следующего поколения платформы данных
- Реализация платформы данных для Data Mesh

Disclaimer

- Материал подготовлен на основе уже ранее описанных методологий и подходов реализации концепции и архитектуры Data Mesh

  - Data Mesh Principles and Logical Architecture (Zhamak Dehghani)
  - How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh (Zhamak Dehghani)
  - Data Mesh Architecture - Data Mesh From an Engineering Perspective (Jochen Christ, Dr. Larysa Visengeriyeva, Dr. Simon Harrer)
  - Evolution of Data Mesh Architecture Can Drive Significant Value in Modern Enterprise (JP Morgan)
  - Deploying Data Products at the speed of the business (Jon Cooke - CTO Dataception)

Совсем коротко про Arenadata

Кто мы

ARENADATA — российский разработчик платформы сбора, хранения и обработки данных

Коммитер ряда Open Source проектов

Решаемые задачи

Корпоративные хранилища больших данных

- Классические КХД и хранилища больших данных с обновлением в реальном и околореальном времени.
- Виртуализация и федерация данных, доступ ко всем данным через единое окно с помощью SQL-запросов.

Разгрузка дорогостоящих инженерных систем (Teradata, Exadata)

- Импортозамещающие ПАК на базе массово-параллельных СУБД с открытым исходным кодом Arenadata DB и Arenadata QuickMarts.
- Прозрачная методика перехода, минимум рисков и сохранение всех преимуществ.

Машинное обучение и искусственный интеллект

- Поддержка R, Python и Madlib.
- Distributed ML на базе Spark.
- Возможность эффективной эксплуатации моделей в SQL-среде с помощью встроенных функций Madlib.

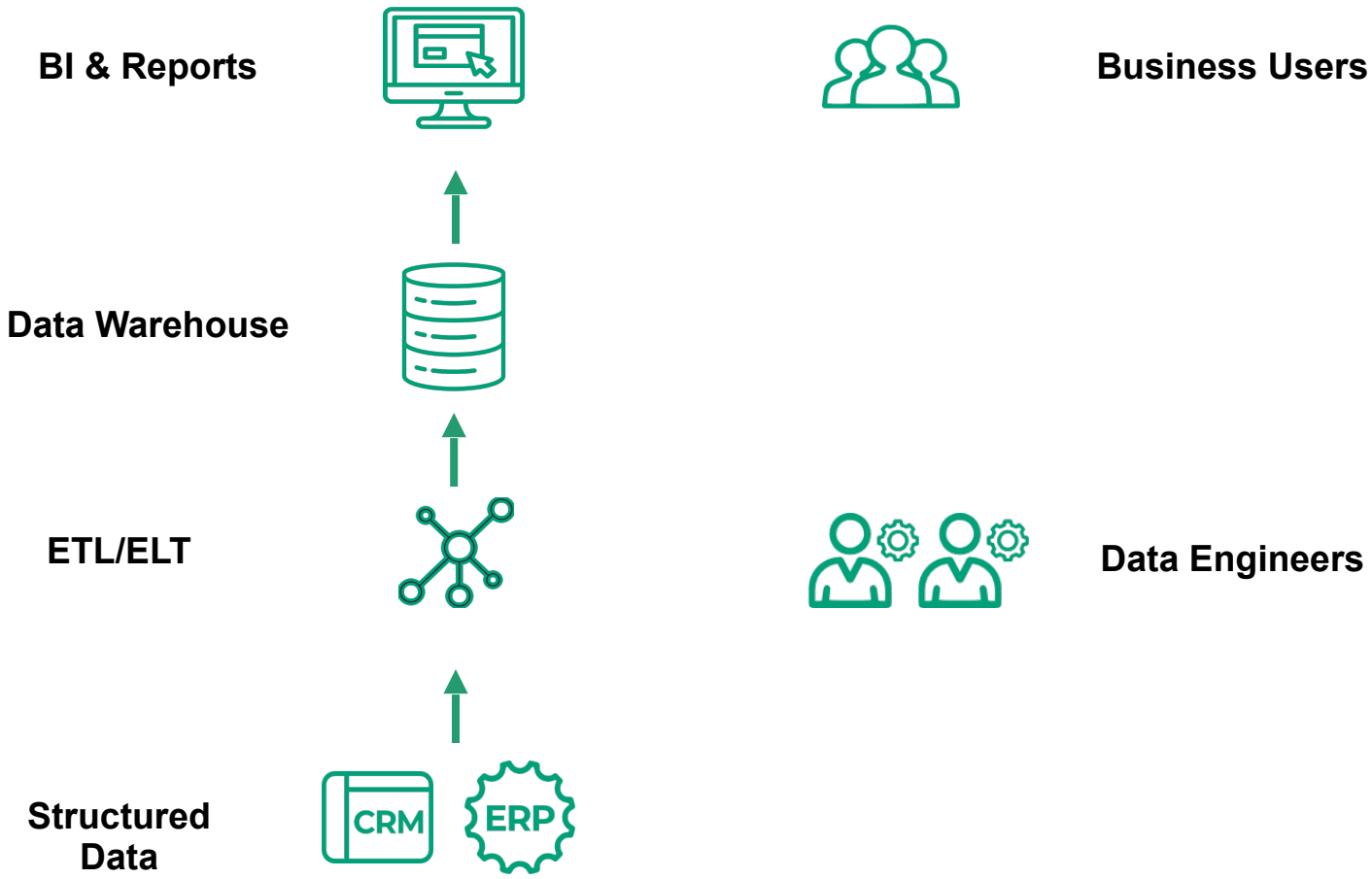
Озёра и фабрики данных

- Единый хаб всех данных компании.
- Возможность быстрого развёртывания и сворачивания зон «песочниц» для пилотных проектов и проверки статистических гипотез.
- Работа со всеми аналитическими инструментами в единой среде.

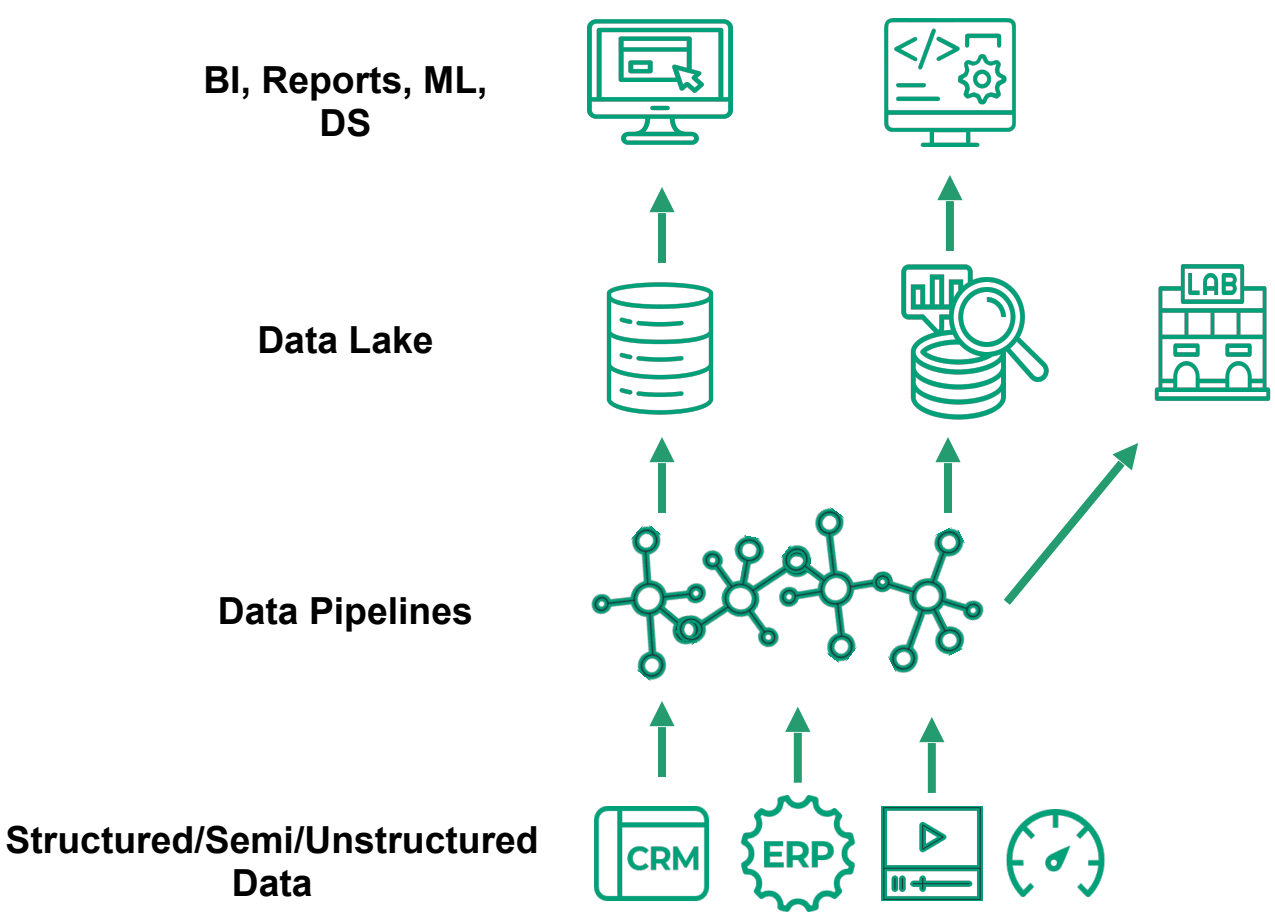
ARENADATA

Текущие архитектурные подходы обработки данных

Data Warehouse



Data Lake



Data Analyst/Scientist's

Data Engineers

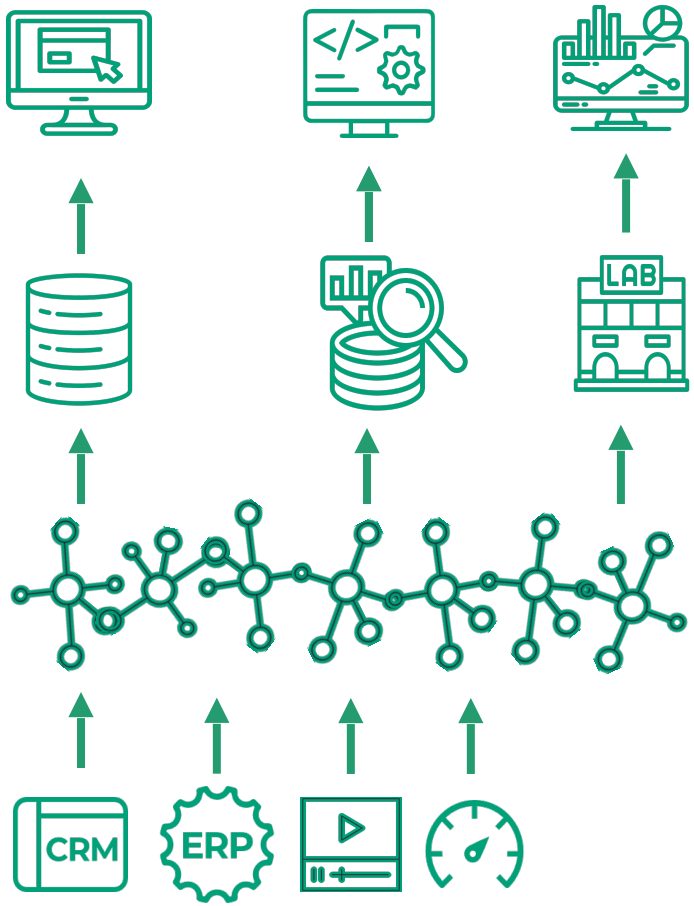
Data Swamp :)

BI, Reports, ML, DS

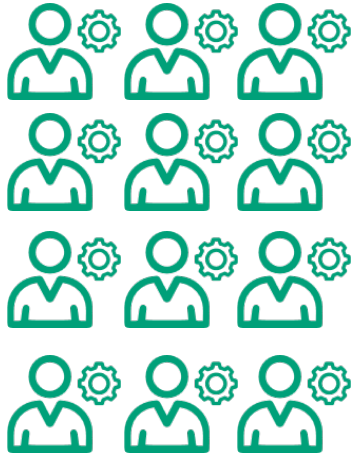
Data Lake

Data Pipelines

Structured/Semi/Unstructured Data



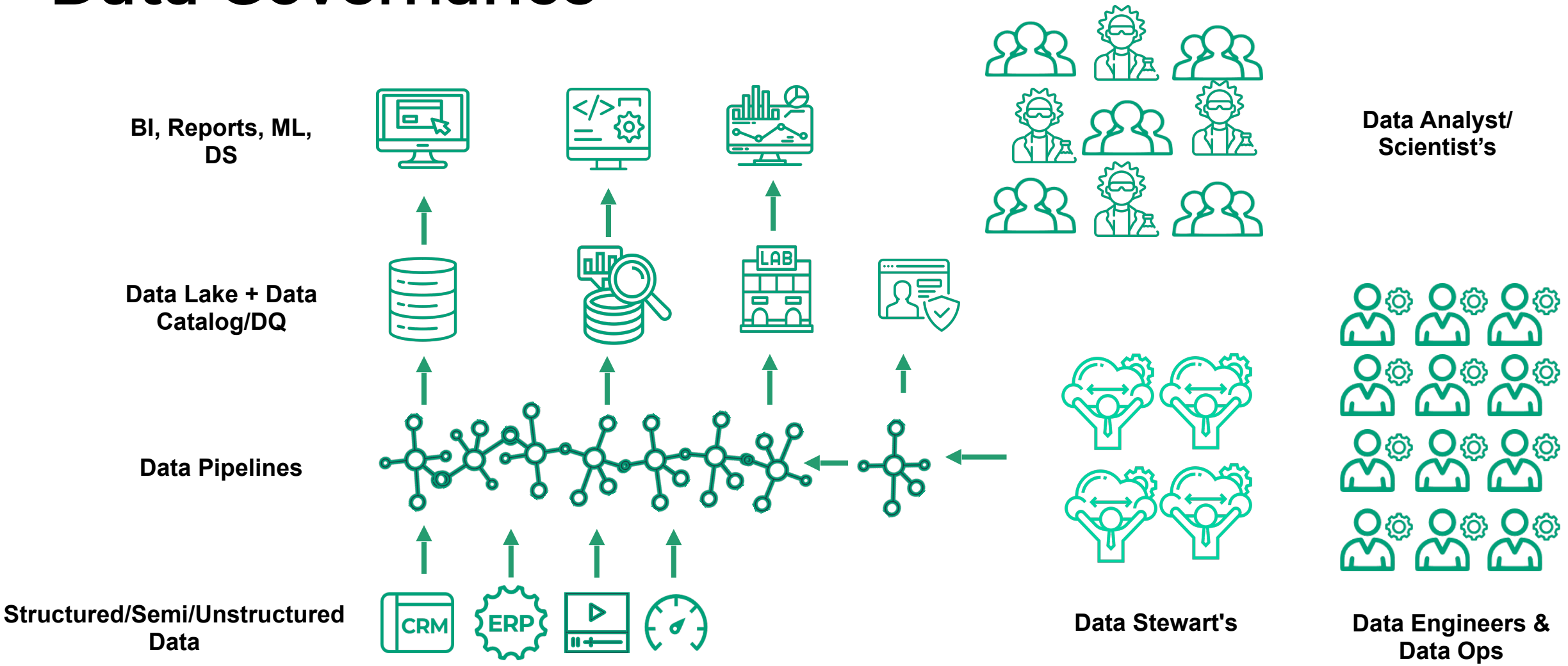
Data Analyst/Scientist's



Data Engineers & Data Ops

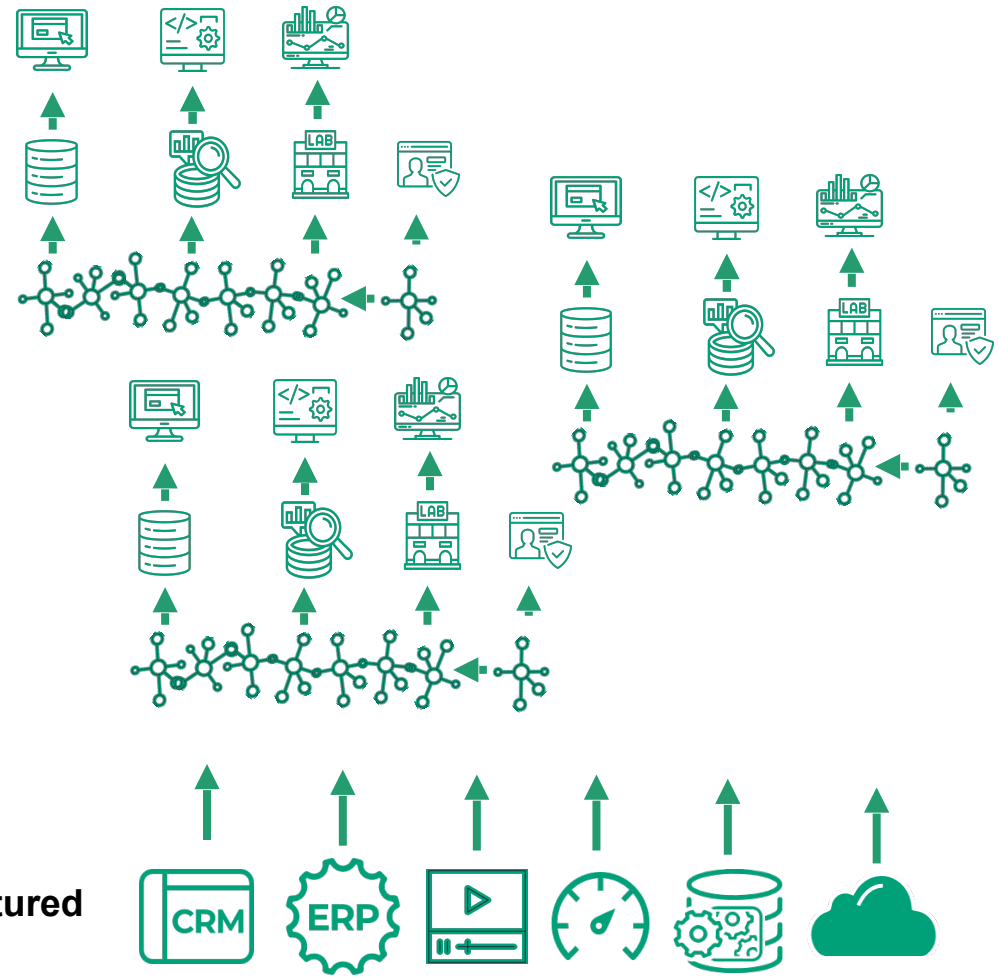
ARENADATA

Data Governance



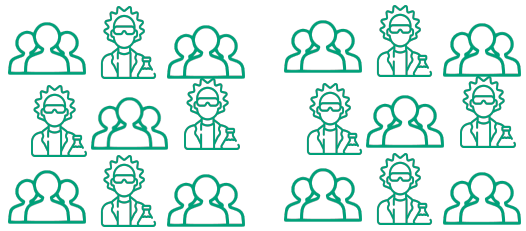
ARENADATA

Data Oceans



Data Lakes

Structured/Semi/Unstructured Data



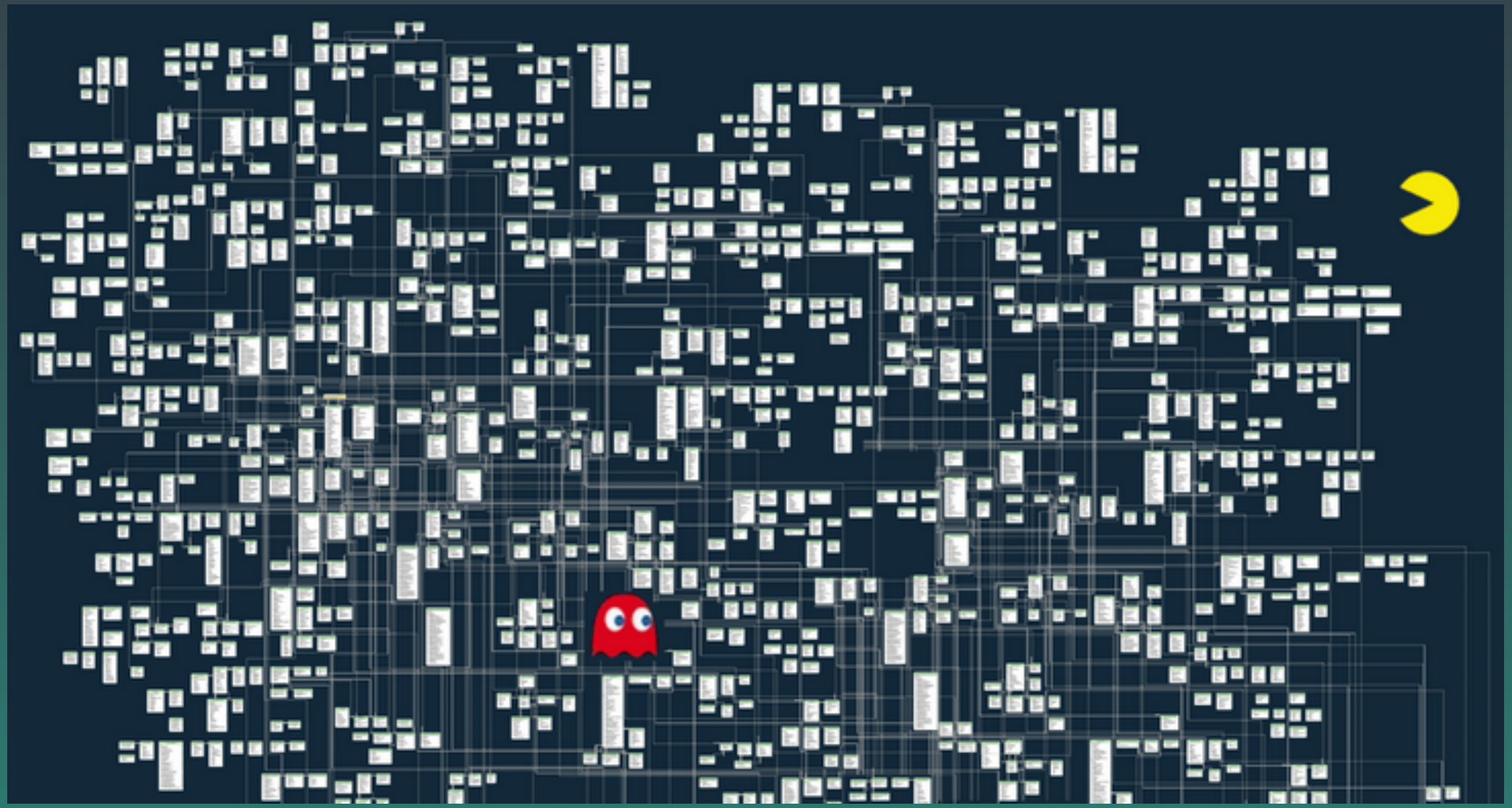
Data Analyst/ Scientist's

Data Stewart's

Data Engineers & Data Ops

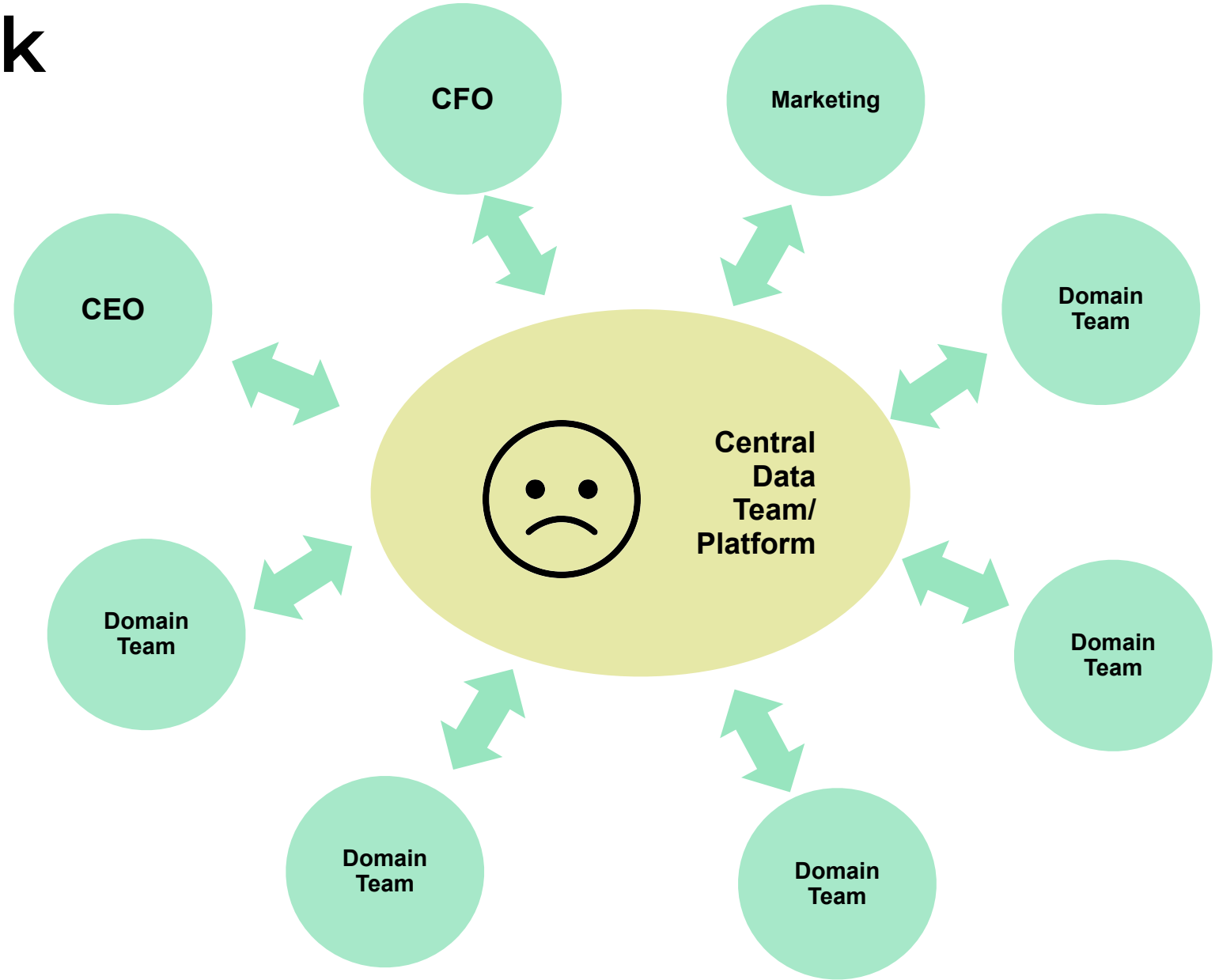
ARENADATA

Модель данных

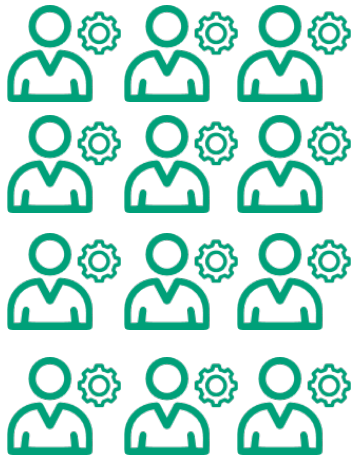


ARENADATA

Bottleneck


CFO
Marketing
CEO
Domain Team
Central Data Team/ Platform
Domain Team
Domain Team
Domain Team
Domain Team
Domain Team

Разрозненные и узко специализированные команды



Data & Ops Engineers

Data Governance, Data Science & ML Team

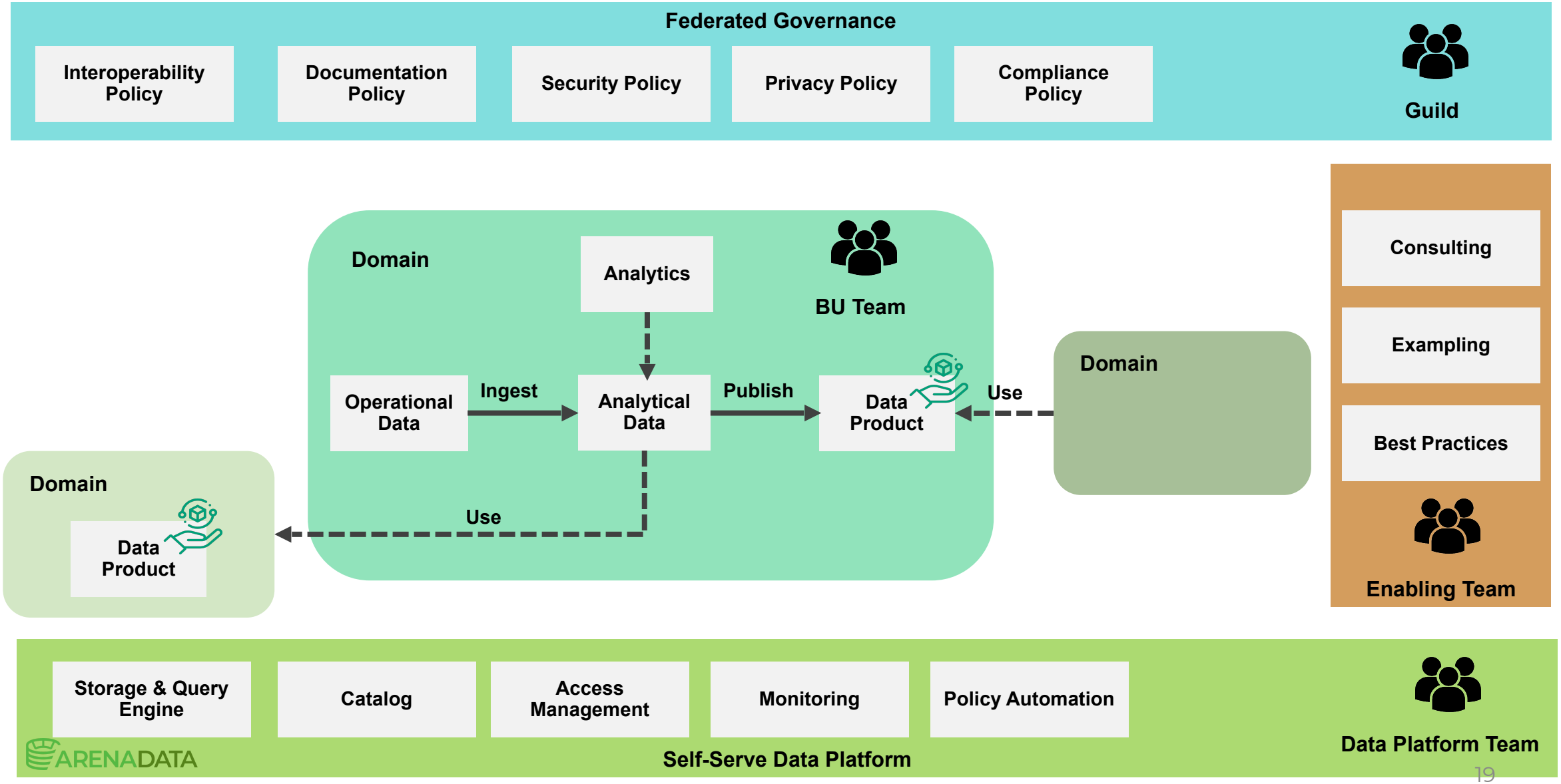
Cross-Functional Domain Oriented Customer/Consumer Team

Архитектура следующего поколения платформы данных

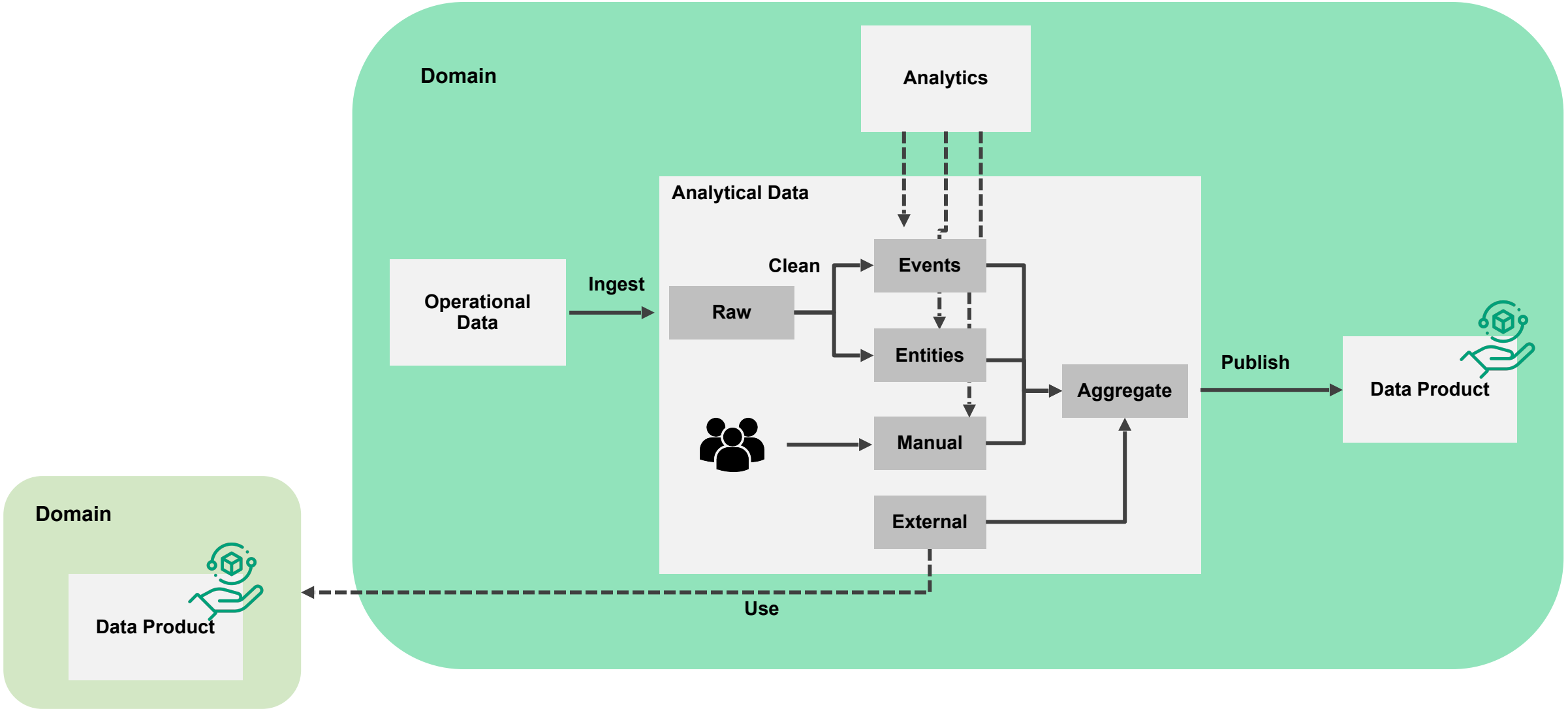
Data Mesh 101

	Domain Ownership	Data as a Product	Self-serve Data Platform	Federated Governance
Strategic Domain-Driven Design	Domain Bounded Context 	Product Thinking 	Domain Agnostic 	Content Mapping 
Socio-Technical Perspective	Domain Teams 	Data Product by Domain Team 	Data Platform Team 	Guild 
Technology	Operational & Analytical Data 	Interoperability Interfaces 	Self-Serve Data Platform 	Data Governance & Automation 

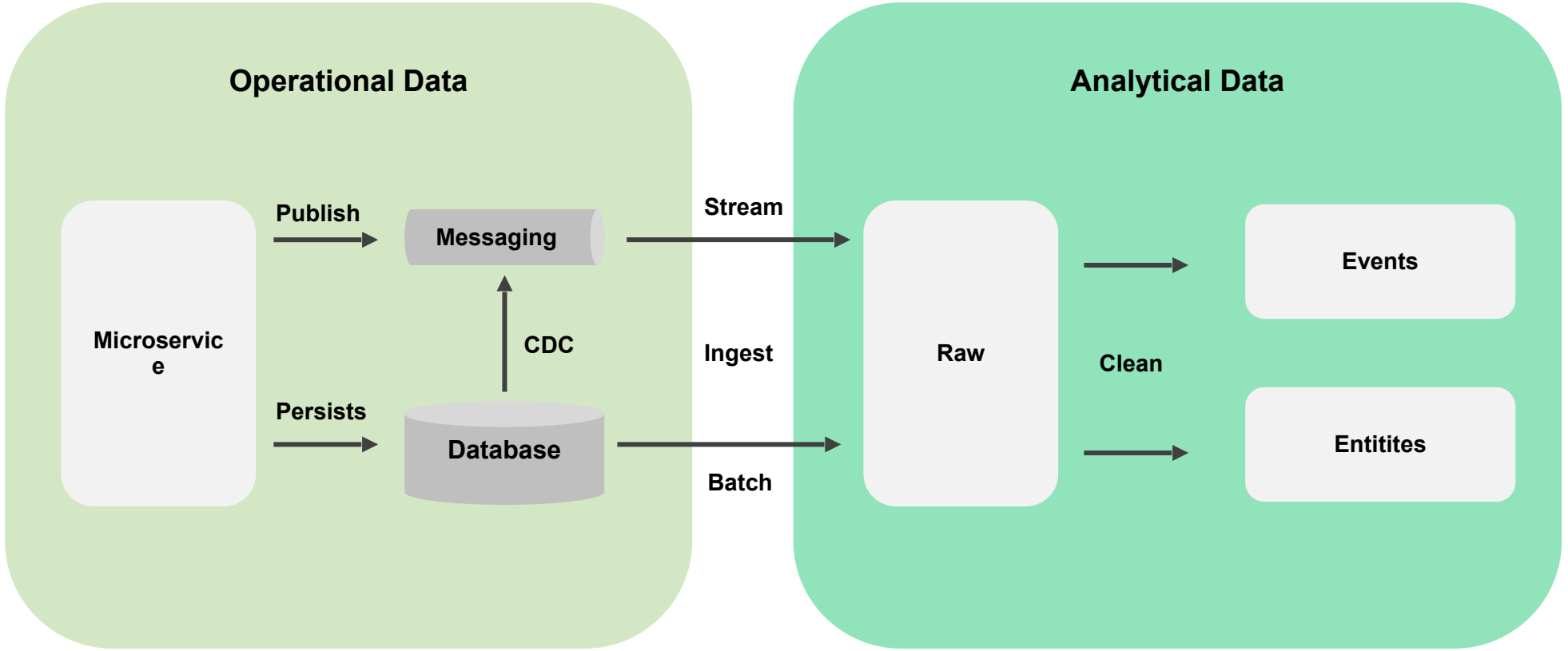
Data Mesh Concept Architecture



Analytical Data

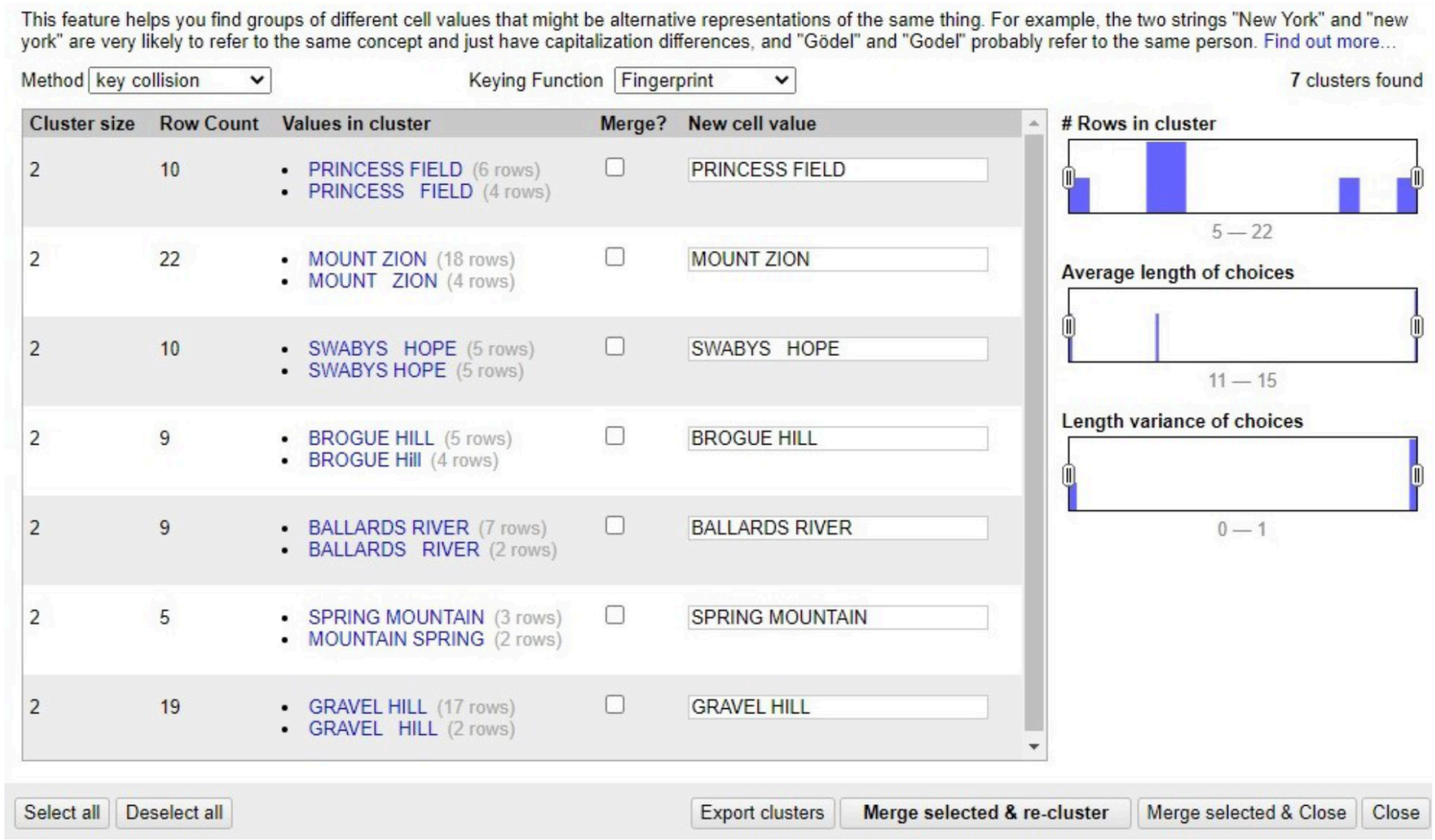


Data Ingesting & Cleaning



Clean Data

- Структурирование
- Дедупликация
- Полнота
- Артефакты



ARENADATA

Federated Governance

- Глобальная политика, которая определяет, как доменные команды должны создавать свои информационные продукты
- Политика интероперабельности
- Централизованный каталог метаданных
- Унифицированный способ безопасного доступа к фактическому продукту данных

Global Policy 1: Data Format

CSV-Files

Standard: RFC 4180

Delimiter: COMMA

Header-Record: Required

Global Policy 2: Location

S3

Bucket is managed by domain team

Global Policy 3: Data Discovery

A common table in Wiki that lists:

- Domain Team
- Data Product Name
- Responsible Product Owner
- S3-Location
- Data Freshness / Update frequency
- Tags
- Data Model

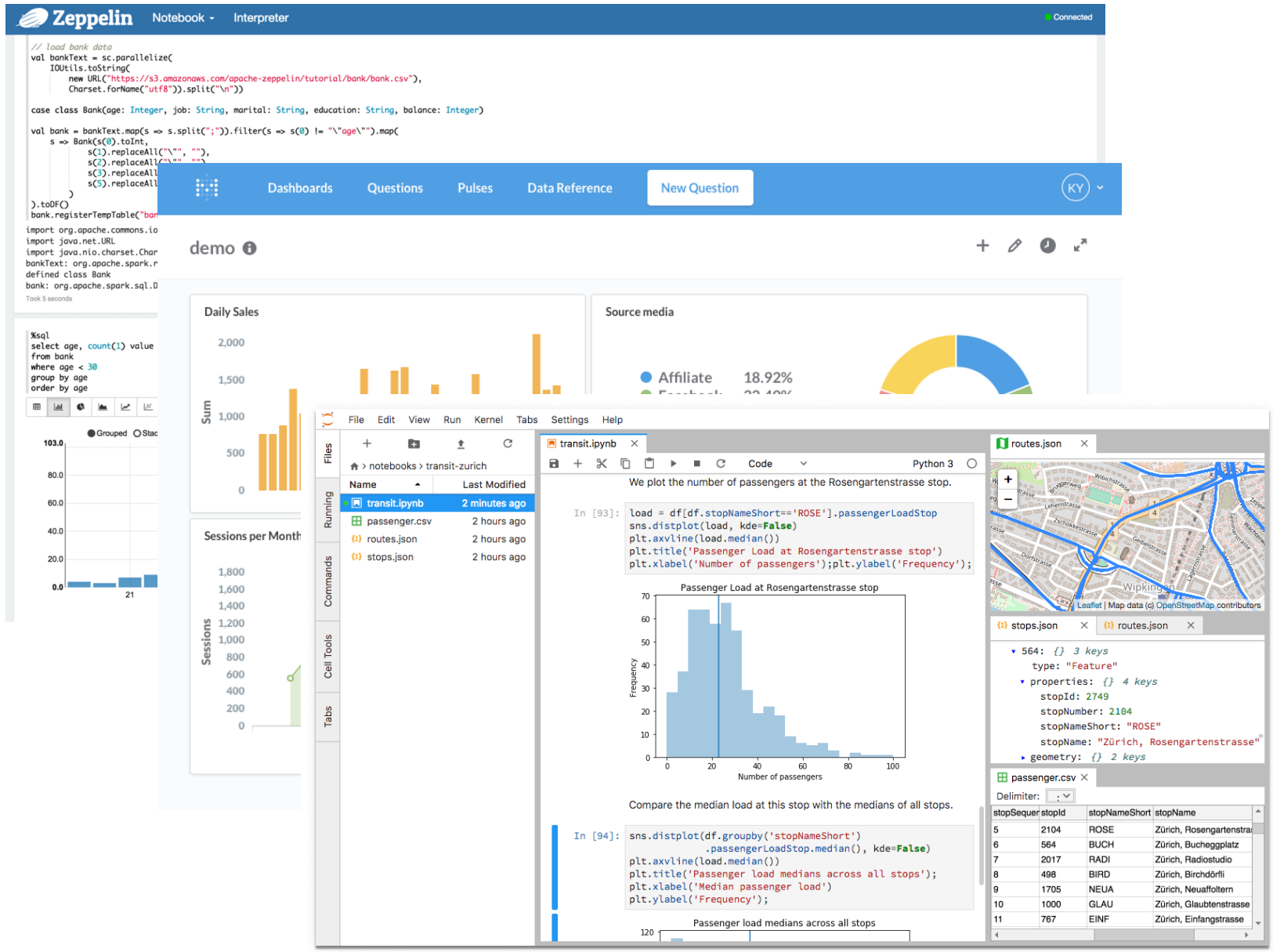
Global Policy 4: Access Control

AWS IAM Role based ACL

ARENADATA

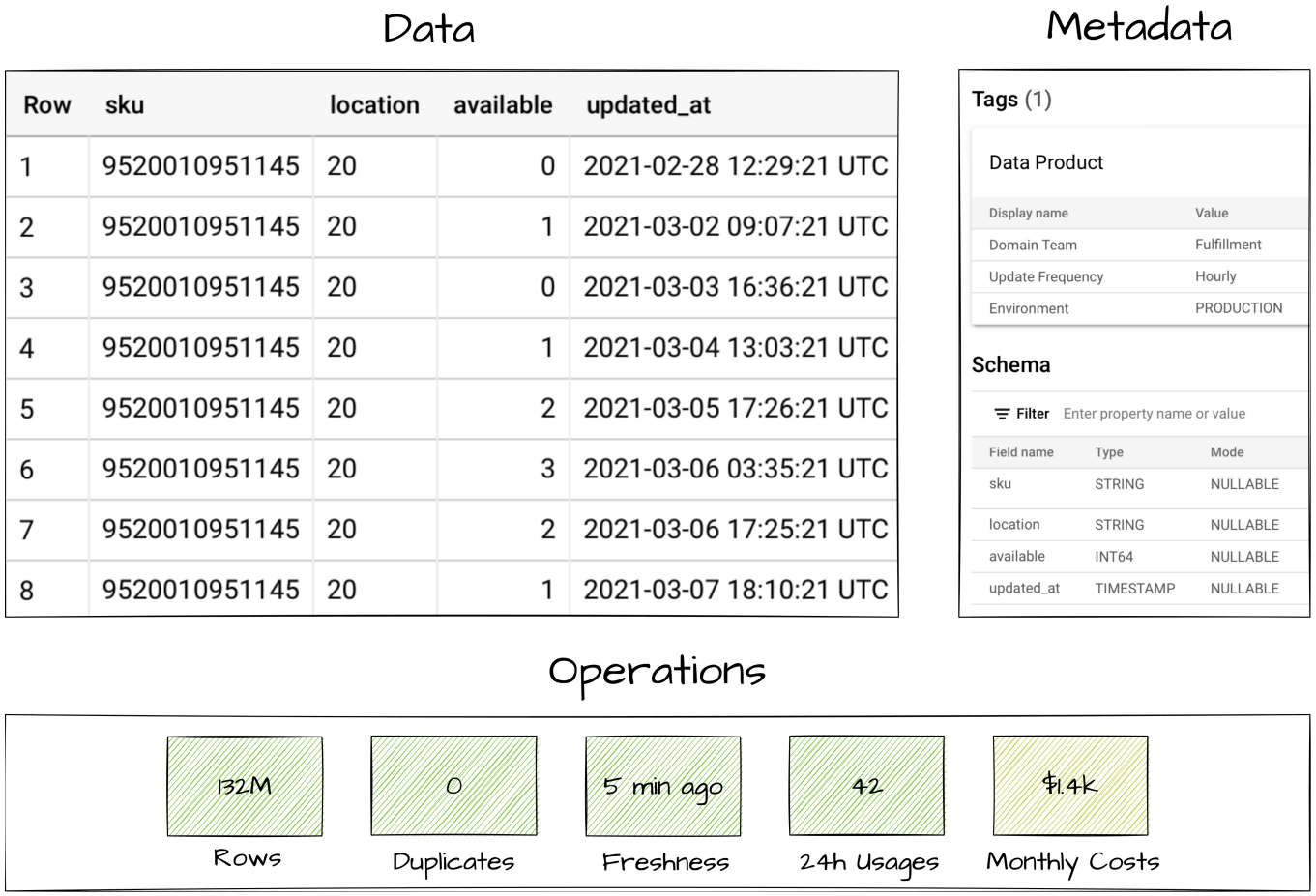
Analytics

- Поддержка интеграции с различными источниками
- Поддержка синтаксиса SQL
- Поддержка in-database и filter pushdown функционала



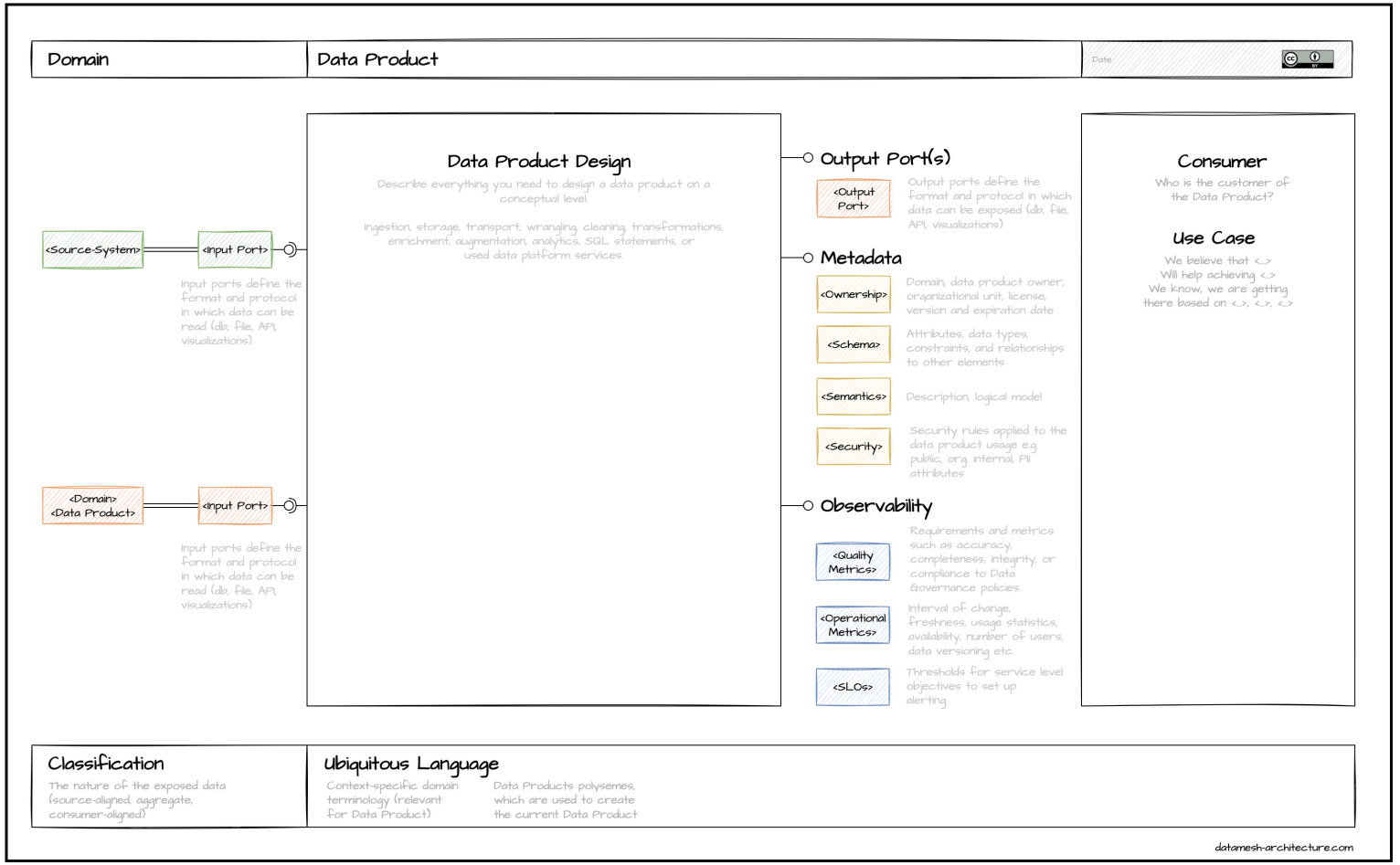
Data Product

- Продукт данных обычно представляет собой опубликованный набор данных, к которому могут получить доступ другие домены, аналогично API.
- Продукт данных описывается метаданными, включая информацию о владельце и контактную информацию, местоположение данных и доступ к ним, частоту обновления и спецификацию модели данных.
- Команда домена отвечает за работу продукта данных в течение всего его жизненного цикла.

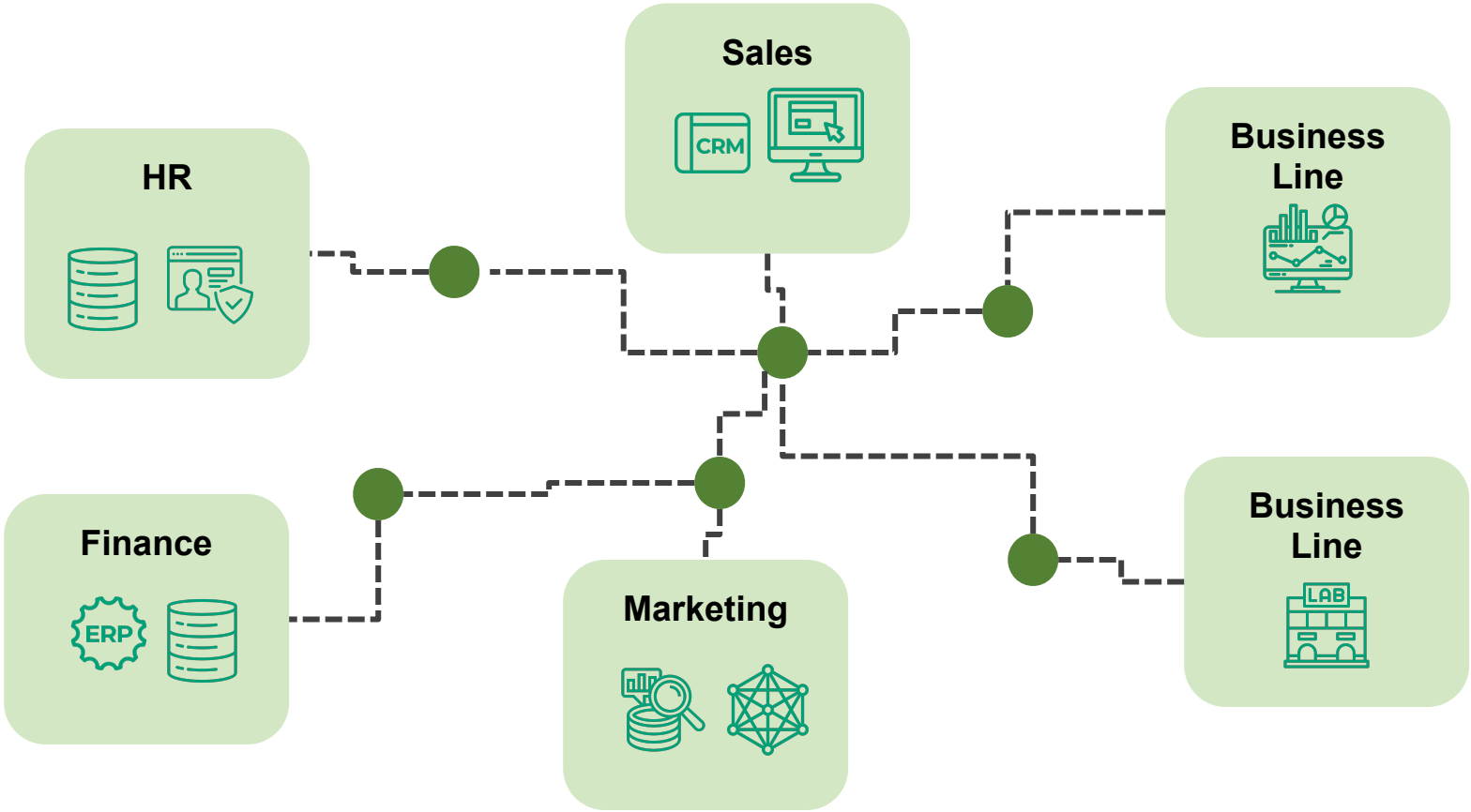


Data Product Canvas

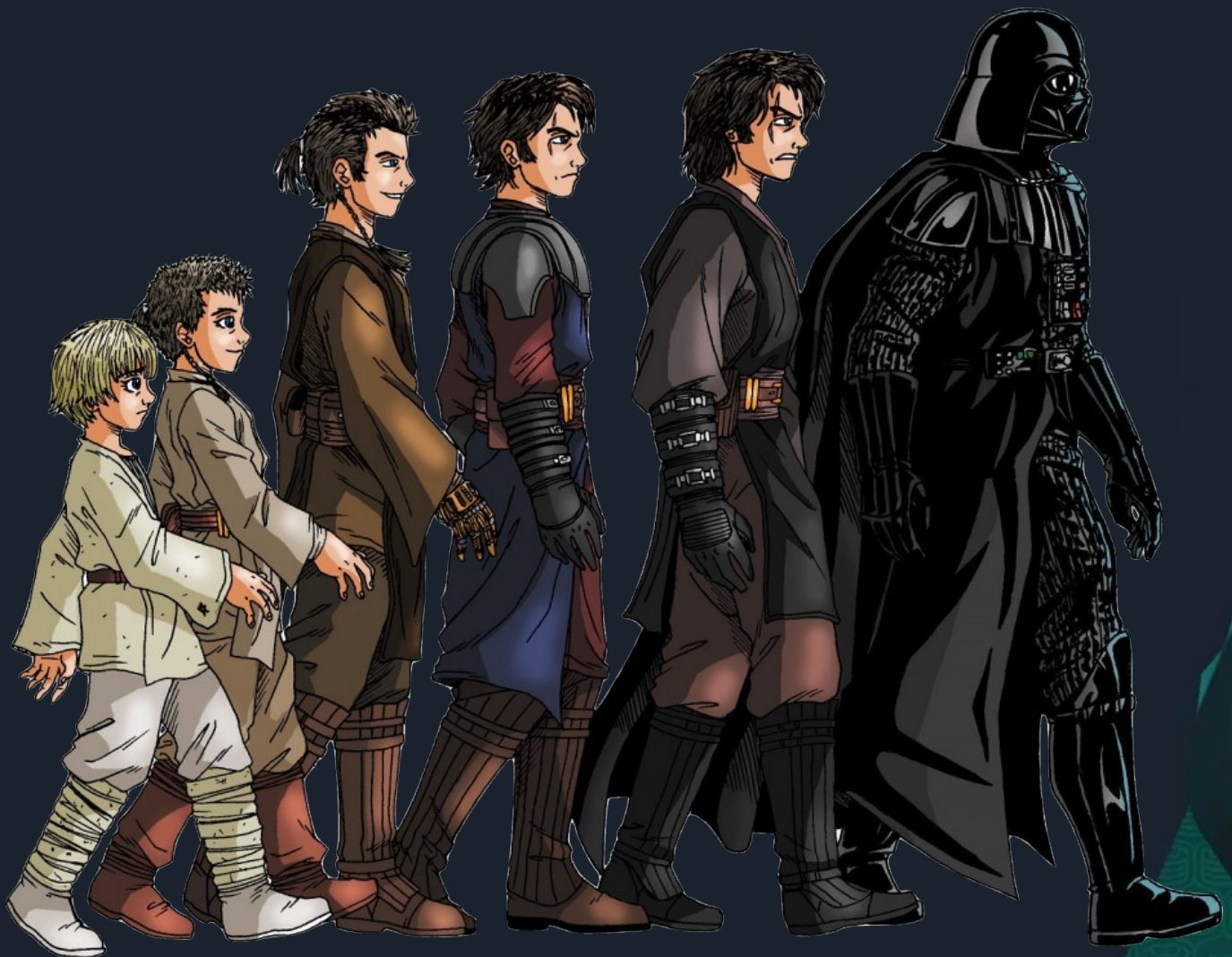
- Домен
- Название продукта
- Потребитель и пример использования
- Выходной порт
- Метаданные
- Входные порты
- Описание продукта данных
- Состояние продукта данных
  - Показатели качества
  - Операционные показатели
  - SLO
- Синтаксис взаимодействия
- Классификация продукта



Domain Driven Design

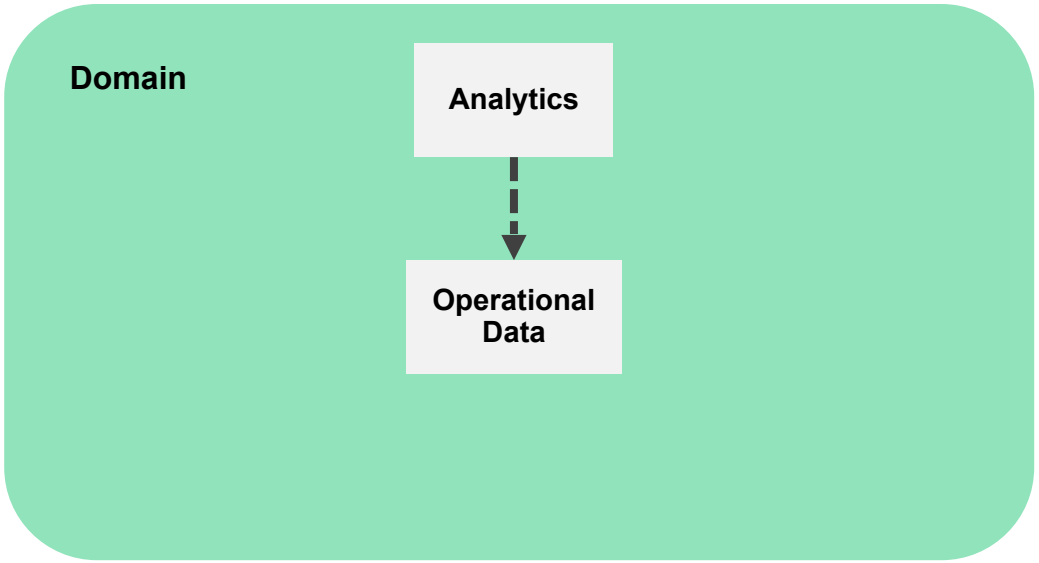


Enabling Team

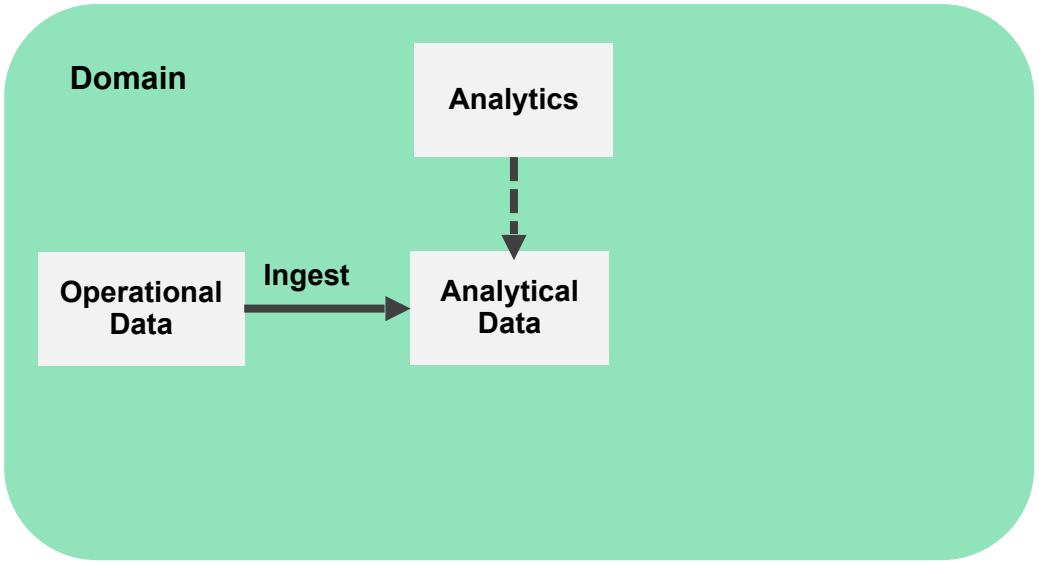


No Data Analytics (lvl 0)

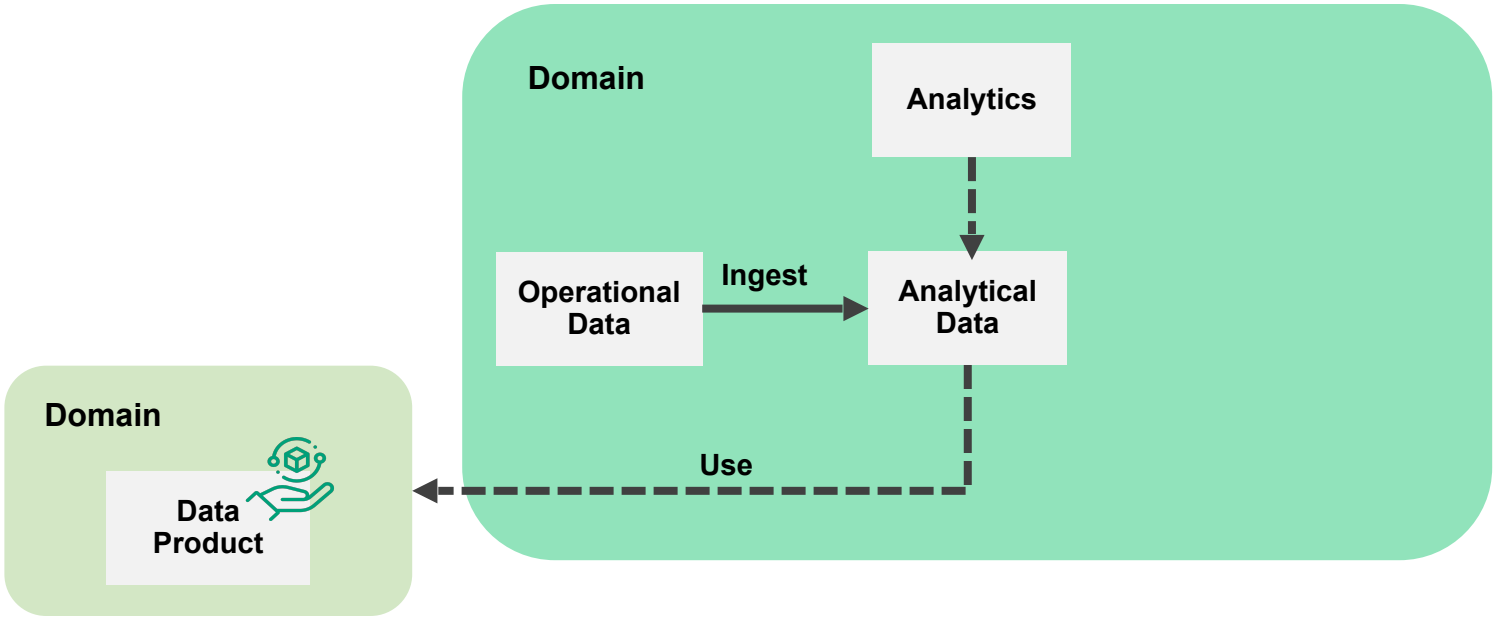
Operational Database Queries (lvl 1)



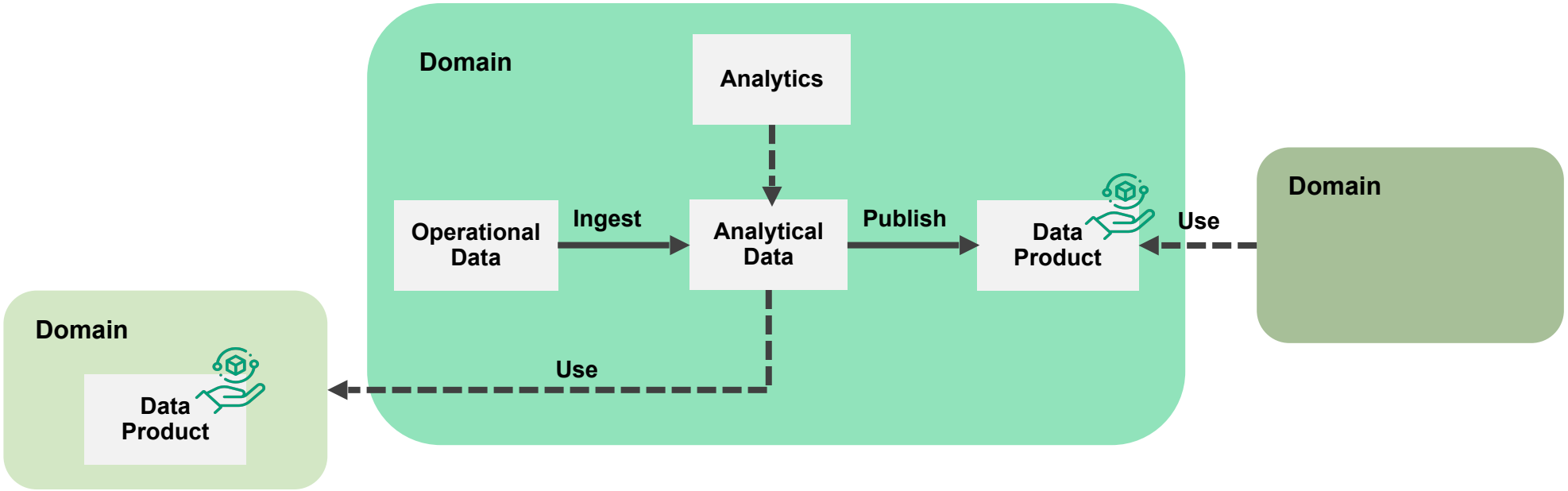
Analyze Own Data (lvl 2)



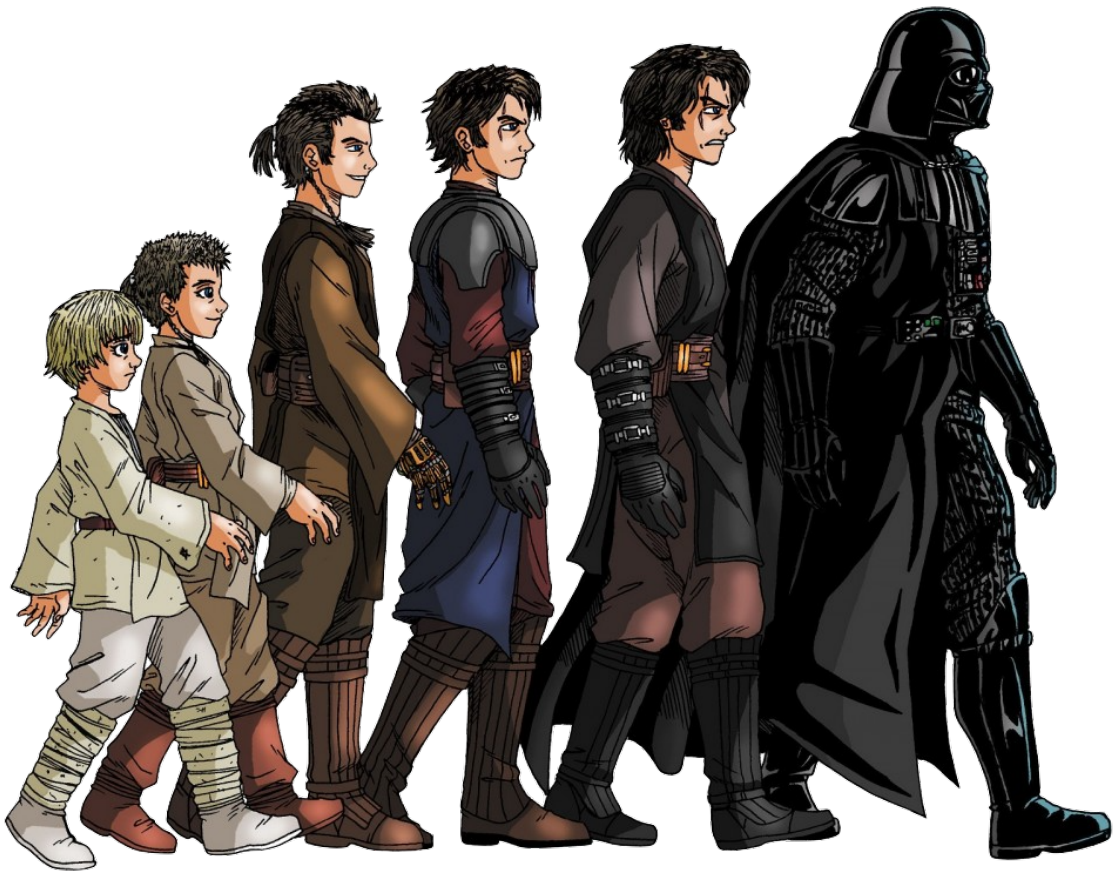
Analyze Cross Domain Data (lvl 3)



Publish Data as a Product (lvl 4)



Data Team Journey

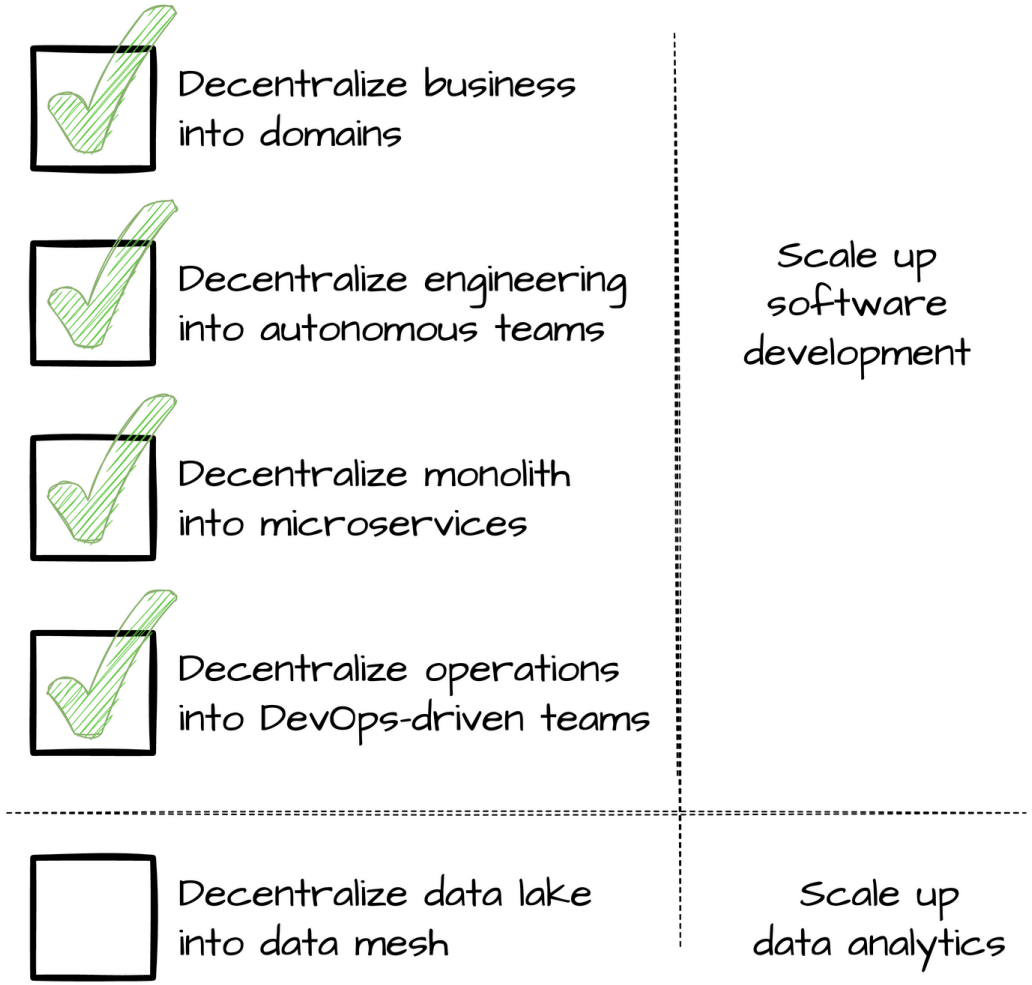


Level 4	**Publish Data as a Product**
Level 3	**Analyze Cross Domain Data**
Level 2	**Analyze Own Data**
Level 1	**Operational Database Queries**
Level 0	**No Data Analytics**

ARENADATA

Реализация платформы данных для Data Mesh

А надо ли?


Decentralize business into domains
Decentralize engineering into autonomous teams
Decentralize monolith into microservices
Decentralize operations into DevOps-driven teams
Decentralize data lake into data mesh
Scale up software development
Scale up data analytics

Data Platform

Single Point Management

- Catalogue
- User Management
- Security
- Policy
- Monitoring
- Platform Management
- Workspaces
- Governance

Data Products (Microservices)

Data Access Layer (Virtualization, Meta-Data, Clean, Aggregation)

Open Metadata · OpenRefine · trino

Domain Product

Data Processing & Orchestration

Apache Spark · jupyter · scikit learn · Apache Airflow · ClickHouse · Apache nifi

On-demand Cloud Infrastructure

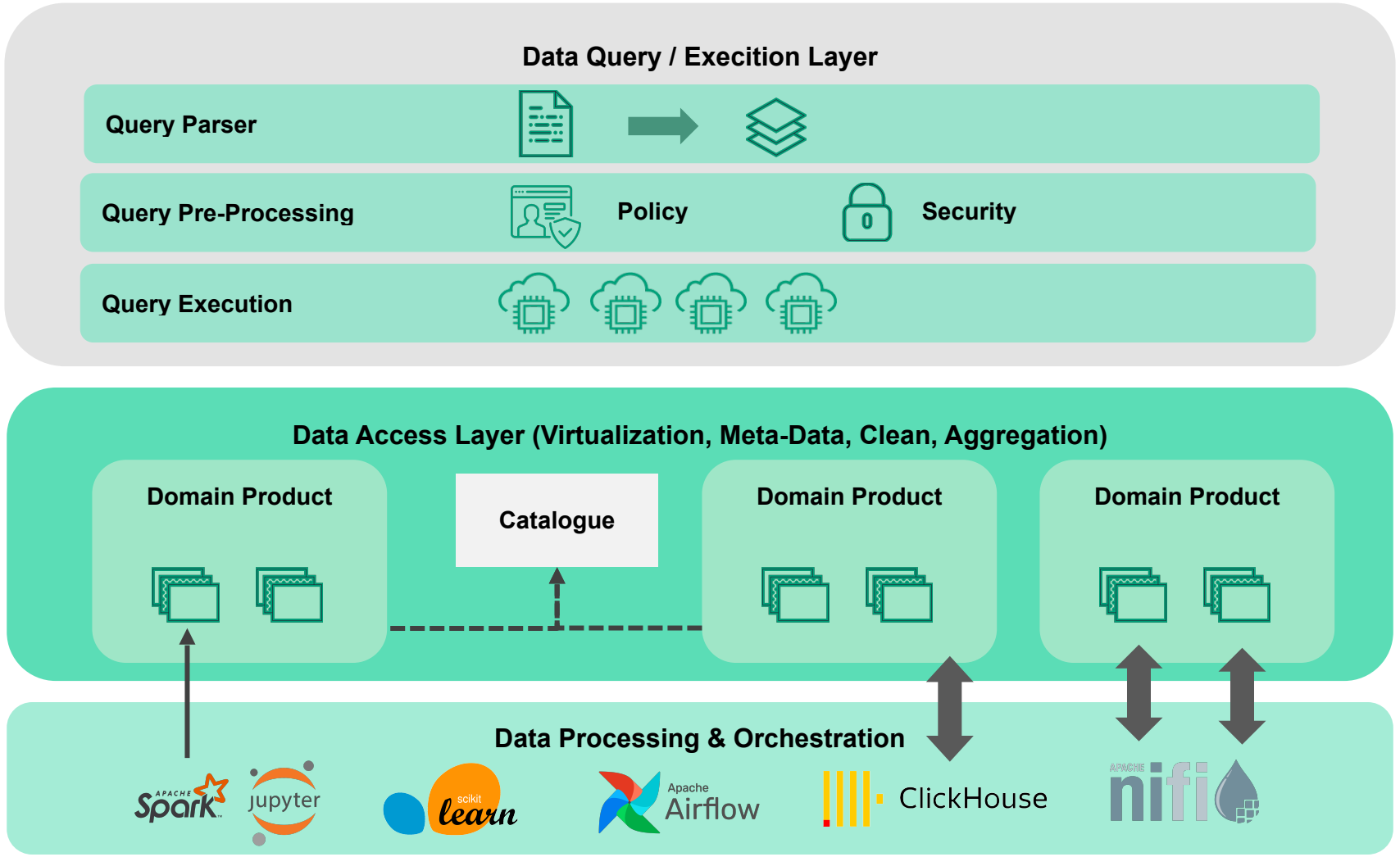
kubernetes · aws · OPENSHIFT · Google Cloud Platform · CLOUD FOUNDRY · openstack.

Source Systems / Infra

kafka · MySQL · S3 · hadoop HDFS · ORACLE

ARENADATA

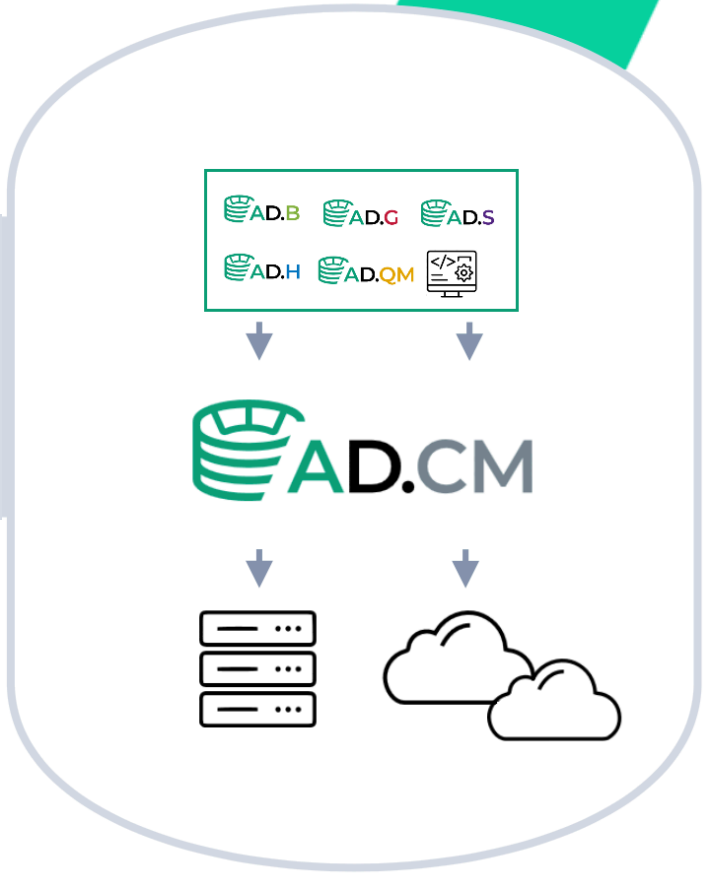
Data Virtualization



ARENADATA

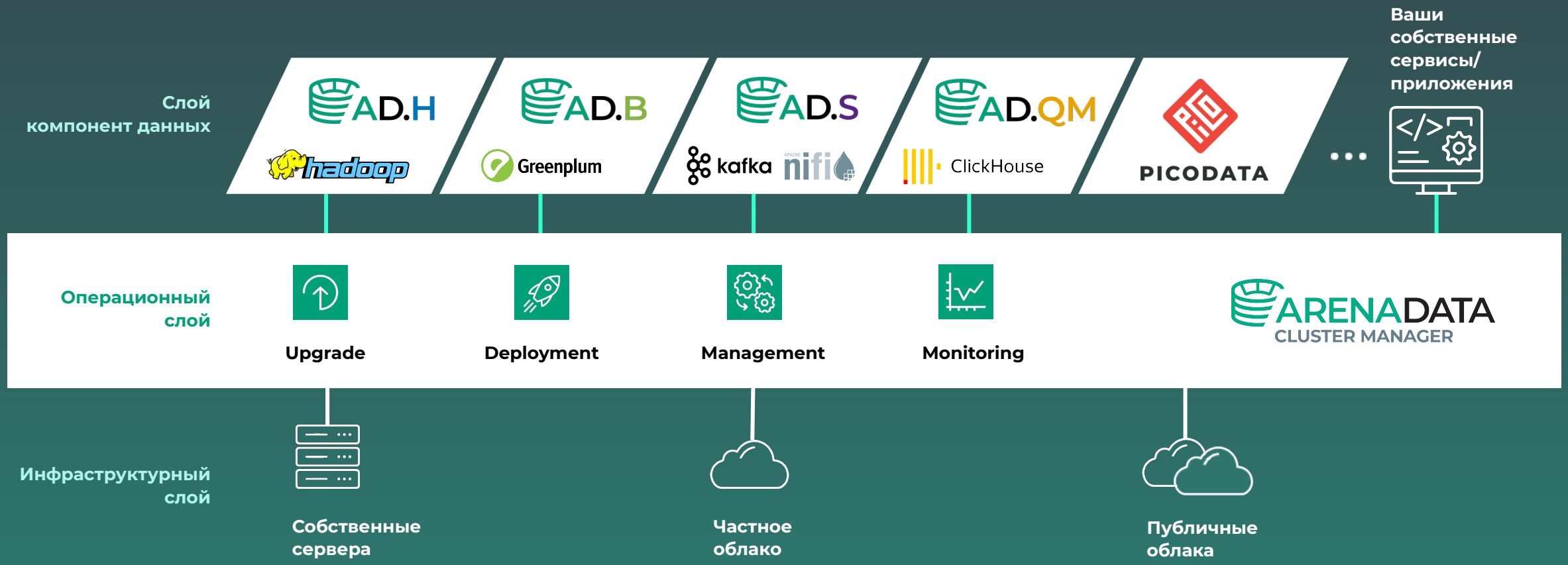
ARENADATA
CLUSTER MANAGER

Arenadata Cluster Manager (ADCM) - универсальный оркестратор гибридного ландшафта. Позволяет быстро устанавливать, настраивать все data-сервисы компании и управлять ими независимо от инфраструктуры.



- В Arenadata Cluster Manager установка, настройка и обновление кластеров в оркестраторе производятся по нажатию кнопки в графическом интерфейсе или по запросу в API.
- При этом все настройки ОС, сервисов, сети, монтирование дисков происходят автоматически.
- Наиболее ярко преимущества ADCM раскрываются при работе с гетерогенной инфраструктурой, при которой появляется возможность размещать data-сервисы на различных типах инфраструктур: в облаке, on-premise или в качестве PaaS-сервисов.

Платформа данных Arenadata

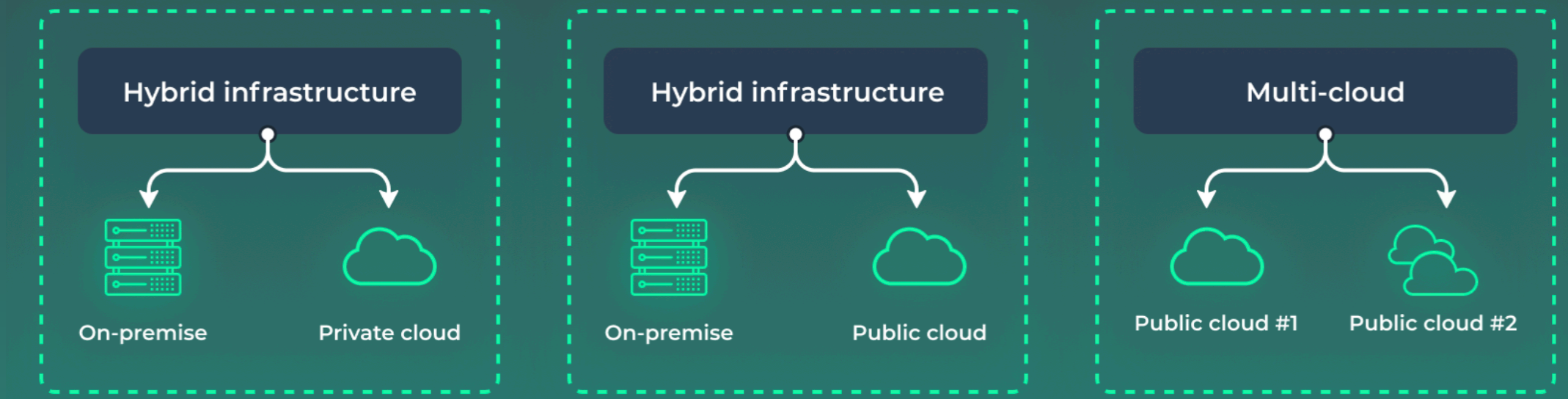


ARENADATA

Объединение разных облачных провайдеров

Мы заложили в Arenadata Cluster Manager возможность поддерживать любые ИТ-инфраструктуры и виртуальные машины в самых разных облаках и на bare-metal оборудовании. В настоящий момент у нас есть коннекторы для Google, Яндекс.Облако, Mail.ru Cloud Solutions, VMware. При необходимости мы можем добавить коннекторы и к другим облакам.

Благодаря ADCM установка наших продуктов в гибридной ИТ-архитектуре происходит в несколько кликов из интерфейса оркестратора.



ARENADATA

ARENADATA
CLUSTER MANAGER

https://github.com/arenadata/adcm

docker pull hub.arenadata.io/adcm/adcm:latest

https://t.me/arenadata_cm

https://store.arenadata.io/#products/arenadata_cluster_manager

ARENADATA

Корпоративная платформа хранения
и обработки больших данных

Оцените потенциал Big Data и Open Source вместе с Arenadata Enterprise Data Platform

- Узнайте больше на **arenadata.tech**
- Скачайте бесплатно на **store.arenadata.io**
- Получите консультацию по почте **info@arenadata.io**

ARENADATA