CLOUDERA

WHITEPAPER

Data Architecture Series

The Data Mesh Paradigm

Building a modern data mesh with Cloudera

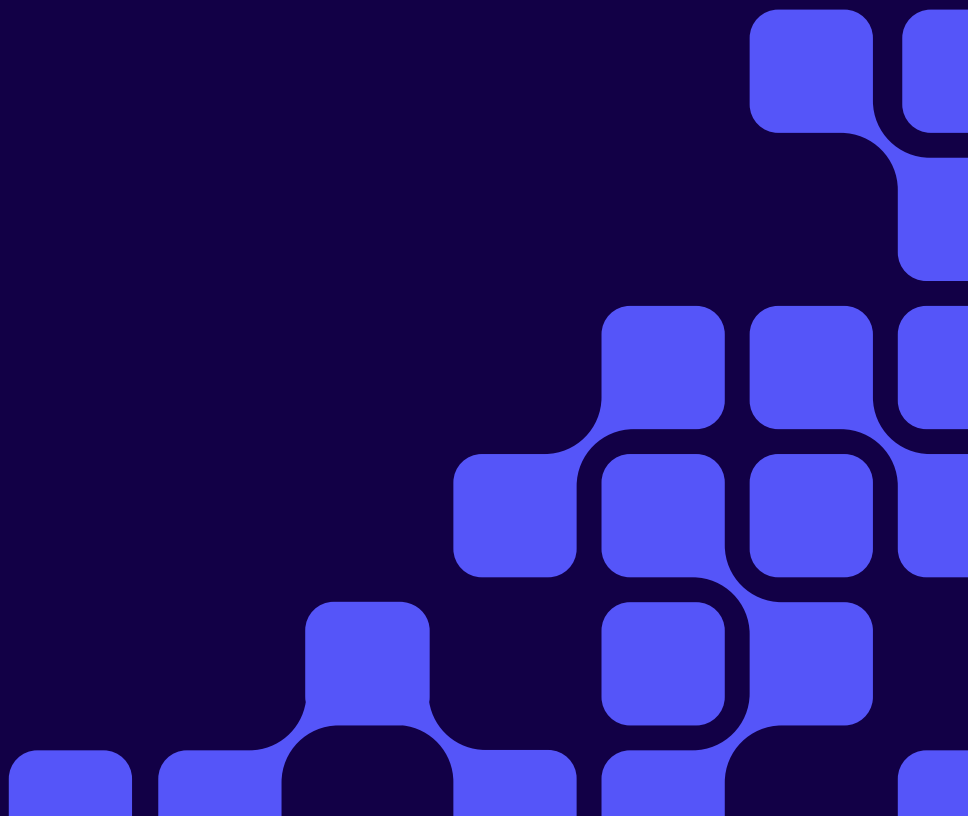


Table of Contents

Abstract 3

Introduction 3

What is a Data Mesh? 4

- Definition 4
- Origin 4
- Properties 4

Section 2: Why is a Data Mesh Paradigm Useful 6

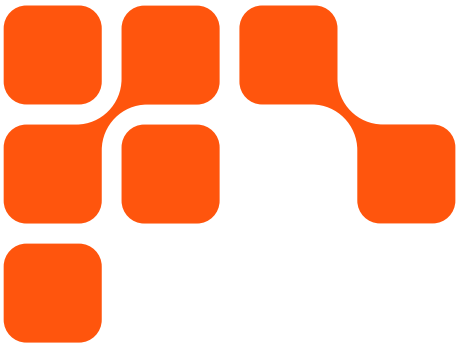
- Gen1: On Premises, Monolithic 6
- Gen2: Public Cloud, Disaggregated 6
- Gen3: Hybrid Cloud, Distributed 6
- The Data Mesh Principles 6

Section 3: How to Align with Data Mesh Using Cloudera 7

- Cloudera 8

Section 4: Beyond the Data Mesh 10

About Cloudera 11



Abstract

This whitepaper provides an introduction to the Data Mesh architecture. It explains what it is, why it was created, especially the challenges it addresses, offers a Cloudera-based reference architecture and highlights two key areas the Mesh can be extended.

Version: 1.0
Author: Dr Christopher Royles

Introduction

In this section we briefly summarize why we wrote this whitepaper, who it is intended for, why they should read it, and recommendations for further reading.

Audience

This whitepaper was written for members of Architecture, Operations, Engineering and Business leaders of Enterprise Data Platform teams. It may also provide useful reading for Chief Data Officers (CDO) and Chief Information Officers (CIO) that want to establish or strengthen their understanding of the Data Mesh architecture, specifically as it applies to Cloudera's products and services.

Purpose

The Data Mesh is one of three important emerging data architectures; the other two are Data Fabric and Data Lakehouse. Organizations need to clearly understand what each of them is, why they are important and how to implement them at scale, in a hybrid landscape. That is the goal of this short introductory whitepaper.

Cloudera has been helping organizations scale their data platforms to manage larger volumes and higher data throughputs than ever before. In part we have been engineering and integrating technologies that directly address and continue to improve on the volume and throughput of data. The discussion has therefore moved to how we can help organizations find and build the right skills and apply those skills to the data management problem at Enterprise scale.

We use the Data Mesh Paradigm to describe this.

Recommended Reading

The recommended reading listed below is limited to only those sources that directly support this whitepaper. Reading the official Cloudera blog or subscribing via email will provide access to a stream of useful reading.

- How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh (20/05/2019)
- Data Mesh Principles and Logical Architecture (03/12/2020)
- Products over Projects (20/02/2018)
- Technology Radar — Data Mesh (20/11/2019)
- HSBC Data Mesh for Securities Data (07/07/2020)
- Agile Labs (Cloudera Partner) — Data Mesh in Action
- Agile Labs — How and why successful data-driven companies are adopting Data Mesh (20/11/2019)
- Blog: Hellofresh Journey to the Data Mesh (20/10/2021)

What is a Data Mesh?

Data Mesh Refers to a "Paradigm shift", where the data management challenge is seen from the business perspective and the different domains within the business. Thinking of data as a product, which can be equated to a goods or service in the context of Data as Capital. Further, the connected mesh of a Data Mesh (with each node contributing a domain specific product into the network) provides an architectural pattern for scaling data and information systems, and driving efficiencies in both cost and value.

Definition

The paradigm refers to domains in a number of ways. The first is domains aligned with data sources and data consumers and shared domains that may be common across both. It also outlines how domains become responsible for the ownership, preparation, aggregation and serving of the domain data product. However it is made clear that the ownership of data by a domain is delegated from a central platform.

The paradigm is founded on four principles:

1. **domain-oriented decentralization of data ownership and architecture;**
2. **domain-oriented data served as a product;**
3. **self-serve data infrastructure as a platform to enable autonomous, domain-oriented data teams; and**
4. **federated governance to enable ecosystems and interoperability.**

An item of note in the above principles is they are not opinionated in what defines a domain, or define a solution to meet the principles, that is left as an exercise for the reader.

Origin

Data Mesh is based on four principles as defined by Zhamak Dehghani in her paper How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh.

The paper discusses historical Enterprise approaches to data platform architectures. The description of first (Data Warehouse) and second (Data Lake) generation approaches, which over-promised and under-delivered. The third generation is also described, but the pattern of a data lakehouse is not explored in detail. The failure conditions of each generation are described to set the problem space. This can be best summarized as centralized and monolithic, lacking the qualities to enable scaling to Enterprise and complex system levels. The paper then describes the next data platform architecture and is premised on "ubiquitous data and distributed Data Mesh". This is then further defined as self-service platform design, coupled with product thinking around data.

Items of note

Data Mesh is tracked by the Thoughtworks Technology Radar — Data Mesh which placed Data Mesh in the 'trail' stage on 29/03/2022. Prior to that it was placed in Assess on 05/2020. Data Mesh is still young in its journey with a three to five year horizon.

Properties

At Cloudera we are fortunate to be working with organizations that have built against the Data Mesh principles and are beginning to recognize the advantages of the approach. They are in the process of trialling Data Mesh at a project level.

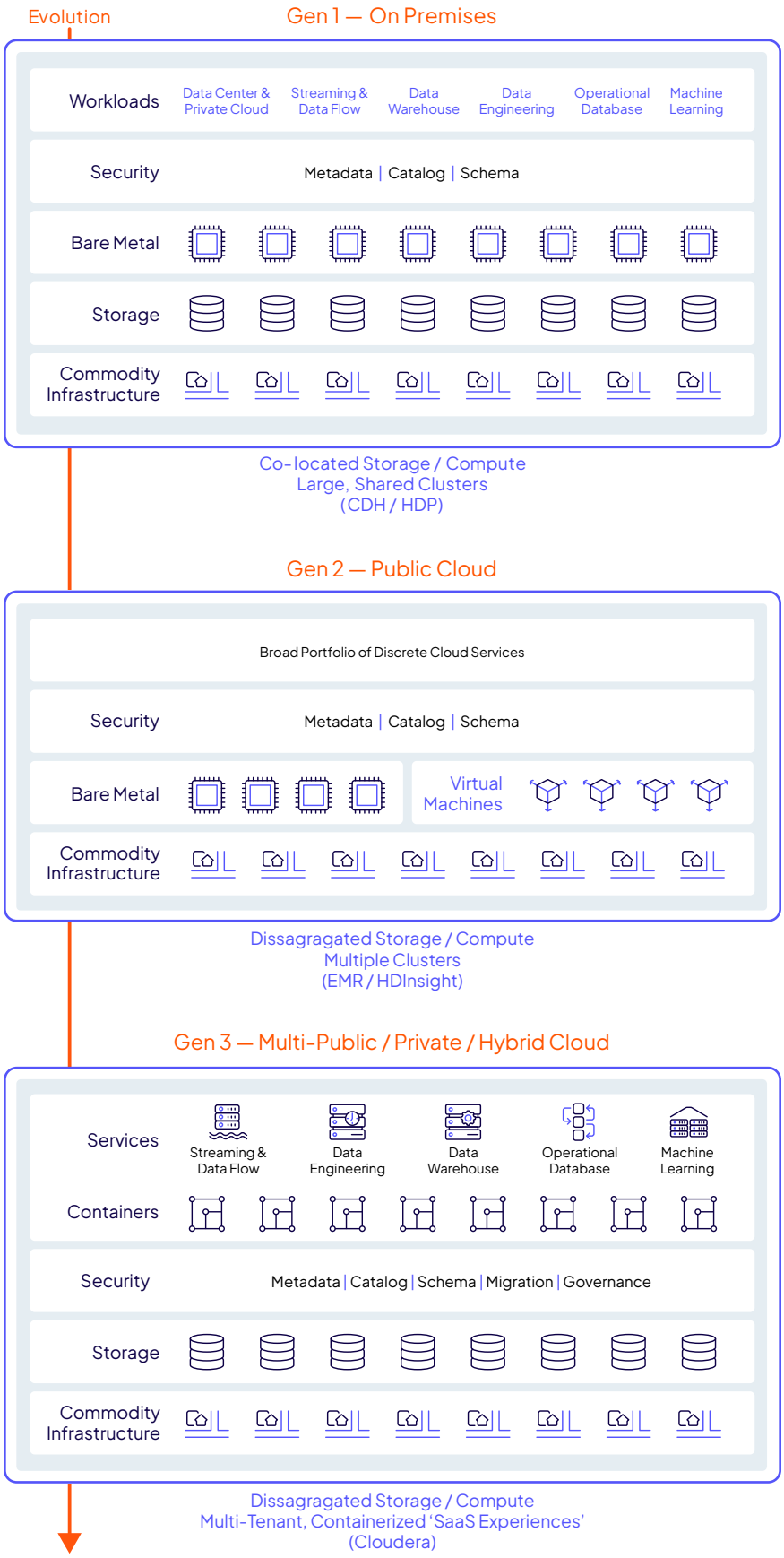
The most common feedback we get from those customers falls into two categories. The first is defining clear ownership of the hosting and serving of data products. This is closely linked with recognizing policies around data as product, policies such as how to find and sustain the team, how to set product SLAs across dimensions such as completeness, timeliness and availability.

When we work with global businesses, the ability to manage the deployment of a data architecture across the globe, across multiple clouds and in data centers is another key foundation for building toward the Data Mesh principles.

The properties of a Data Mesh are encoded in the four principles, and also the above observations, of how ownership of data is assigned, how data is hosted and served as a data product.

Data products are the foundation of the approach; they need to be discoverable and consumable. When consuming data, think in terms of consuming it as events distributed across the mesh and finally making sure central governance can be applied at an Enterprise level, with full delegation down to domains in the analytical tools and applications they choose to apply, and enabling those tools and applications to be self-serviced.

Figure 01 — The evolution from monolithic, to disaggregated, to hybrid cloud



CLOUDERA

Section 2: Why is a Data Mesh Paradigm Useful

In this section we consider why the Data Mesh Paradigm is useful. We consider the limitations of traditional monolithic approaches to Data Architectures, and explore how the principles help address these limitations.

We start by observing the architectural trends over the last 10 years. There are three discrete generations of data architecture. What is listed below, differs slightly from that defined by Thoughtworks in How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh.

Gen1: On Premises, Monolithic

While built on Open Source software and using commodity hardware to provide cost efficiencies, it represents a monolithic tightly coupled stack between applications, compute and storage. To expand storage the architecture requires an expansion of compute. Organizations would look to limit the number of instances deployed, only duplicating for resilience, or development policies such as pre-production. It was common to see entire clusters built to support a single use-case. Shared workloads, or multi-tenant deployments were hard to operationalize, taking great care in turning and resource isolation with each service having to be built on common underlying frameworks to manage resourcing within the monolithic cluster.

Gen2: Public Cloud, Disaggregated

The cloud providers evolved their platform services with compute and storage clearly separated and independently provisionable. Relying more on their network fabrics to take on the load and throughput of data across the integrated services. This separation of storage and compute brings a number of what we can term cloud qualities to the architecture. As an example, it is possible to isolate resources from one another by standing up discrete service instances; these service instances can share storage. It is possible to bring auto-scaling and scale to zero into the architecture, so compute is only used when there are workloads to process.

Gen3: Hybrid Cloud, Distributed

This is where we deviate from the paper referenced above, in that we move directly from Gen2 to Gen3 and we redefine Gen 3 to be Hybrid, in using "Cloud based managed services", but also in bringing Containerization into the Architecture for workload and compute portability. This also enables qualities of the cloud to be brought into the data center. So now we have disaggregated storage and compute, full resource isolation and elastic scaling in whichever form-factor of deployment is required.

The Data Mesh Principles

However: just providing a Hybrid and Distributed Data Architecture is not enough, there is also a need to bring additional shared services into the logical architecture to align and address the principles.

Domain Ownership

In order to provide ownership of data to people close to the domain and business problems, it is important to be able to authenticate and authorize users in a common way across all instances. Once authorized, then being able to assign ownership to a data object is foundational to addressing this principle.

An object may be at any level of abstraction, from data-lake, catalog, table, column or cell. This means ownership can be assigned at any level of granularity and with ownership comes the delegated responsibilities of managing metadata, access and use policies and service endpoints across both operational and analytical end-points.

Data as Product

Data as product helps the domain combine the elements of data, metadata, code, and infrastructure to provide an atomic unit. This unit can then be maintained efficiently within the Data Mesh. It helps the domain owner think more critically about the use of the data product, the SLAs, qualities and constraints that will be useful to the data product consumers. It also helps in organizational approaches such as how a data product team is organized and funded. Data as product supports wider thinking about the F.A.I.R qualities: Findable, Accessible, Interoperable and Reusable. The data should be easy to discover, self-describing and accessible for use.

Self-Serve Data Platform

So we have organized and funded teams around data products, these teams are delegated ownership to manage and build their own data products and serve them to their customers across the Data Mesh. We need to be able to bring the analytical tools required by the team in order to build high quality, reusable data products. Data Owners will want to instantiate CI/CD services, discover new data products, integrate those products, enrich and profile data for quality and then serve the data as a new composited Data Product. Being able to choose which data infrastructure services are deployed and at what scale, as well as managing the roles that use those tools within the team should be easy and manageable by the domain owners.

Federated Computational Governance

Data Mesh is a distributed system architecture, However a modern Enterprise will have global and regulatory constraints and policies that will need to be encoded into each domain. By delegating ownership it may also be required to apply constraints to the activities a domain can undertake, to protect and reduce risks to the system as a whole. Therefore governance becomes a federated service which can be applied at all domains and object levels, bringing global consistency in policies, security, auditing and observability.

The Data Mesh Paradigm is useful as it helps explain the organizational benefits of distributed system thinking. It helps an organization to think about how they are structured and can bring those valuable and limited skills to bear on complex data problems, rather than taking a technology, or feature and function first approach.

Section 3: How to Align with Data Mesh Using Cloudera

Cloudera helped bring the Data Lake to the Enterprise with the Cloudera Data Hub and helped package the Open Source Software components of Apache Hadoop (published by Google, donated by Yahoo (2008)). It was recognized quickly the value of structure and SQL as a first line language for business on Datalakes and brought in Apache Hive donated by Facebook (2010).

From 2010 it was possible to run a SQL data warehouse on a data lake and bring structure to the underlying unstructured data store. This is the foundation of the modern data lakehouse architecture. Qualities such as wider ANSI SQL support, ACID compliance and faster processing engines such as Tez and LLAP brought the warehouse performance on par with modern RDBMS systems.

Between 2015 and now the focus has been on bringing the platform to where customers and their data is. This drove an imperative for cloud native support and a recognition that vast data volumes and workloads continue to reside on 1000+ node clusters in the data center.

This is the foundation on which Cloudera was built. It represents the best of the open source components, organized in a model that decomposes the monolithic platform into separate storage and compute. It brings in a common security and governance framework that provides consistency across instances, and has form factors for deployment into the data center, or public cloud.

We refer to this as the hybrid platform. It is recognized by Analysts as a Data Fabric when implemented at scale across an Enterprise. We built Cloudera as the only true hybrid platform for data, analytics, and AI, in order to bring cloud qualities into the data center. We also needed to be where the customer and data was being generated, more-so in the cloud, and we needed to align with our customers initiatives, including their journey to their choice of cloud deployment. This meant we had to build in flexibility, not just in where the data cloud is deployed, but also for our customers in terms of where they deploy now, and into the future.

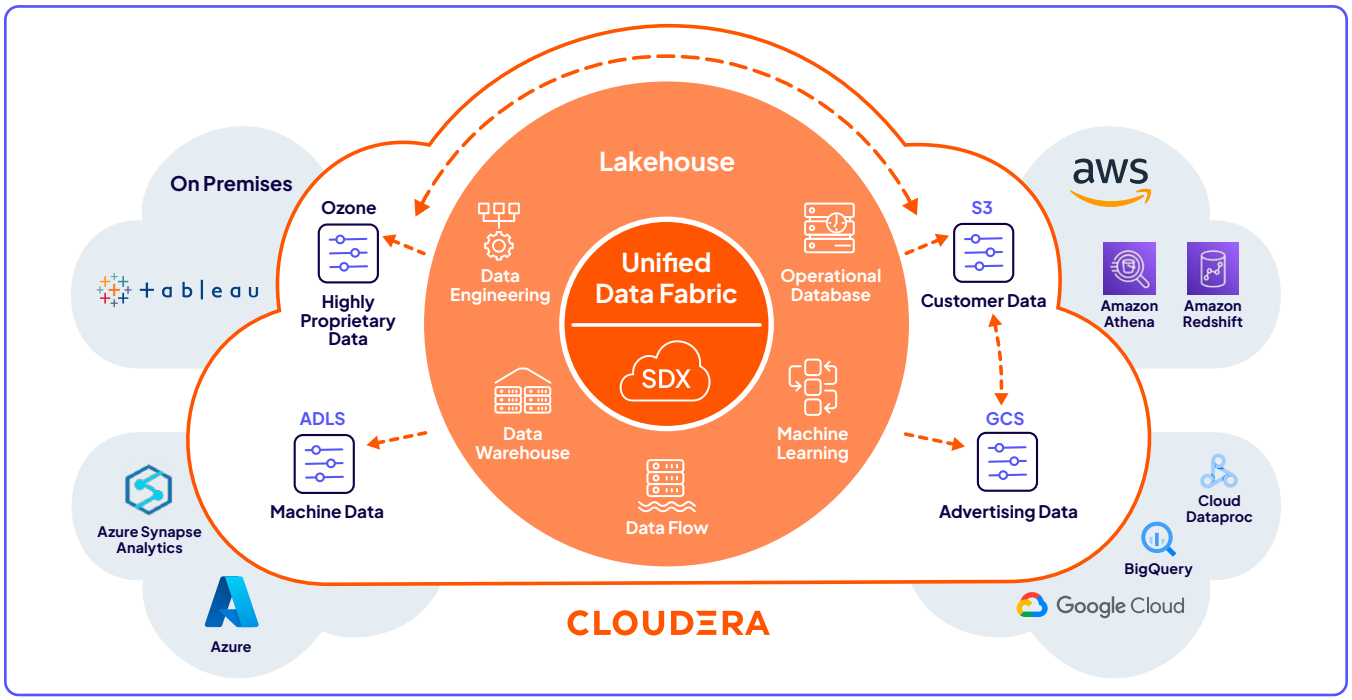


Figure 02 — Cloudera

Cloudera

In this section we provide an introduction to Cloudera, with a focus on Cloudera on cloud. We then summarize the key logical services components that support Data Mesh. We conclude by looking beyond the Data Mesh as we know it today and share how Cloudera is aligned with the four principles.

Cloudera is a true hybrid platform designed to provide the freedom to choose any cloud, any analytics, any data. Cloudera delivers faster and easier data management and data analytics for data anywhere, with optimal performance, scalability, and security. With Cloudera you get the value of Cloudera on premises and Cloudera on cloud for faster time to value and increased IT control.

Cloudera provides the freedom to securely move applications, data, and users bi-directionally between the data center and multiple data clouds, regardless of where your data lives. All thanks to modern data architectures:

- A unified Data Fabric centrally orchestrates disparate data sources intelligently and securely across multiple clouds and on premises
- An open Data Lakehouse enables multi-function analytics on both streaming and stored data in a cloud-native object store across hybrid multi-cloud
- A scalable Data Mesh helps eliminate data silos by distributing ownership to cross-functional teams while maintaining a common data infrastructure

Figure 03 provides a summary of the logical components that make up Cloudera on cloud. We'll now explore how each of these components supports the Data Mesh paradigm.

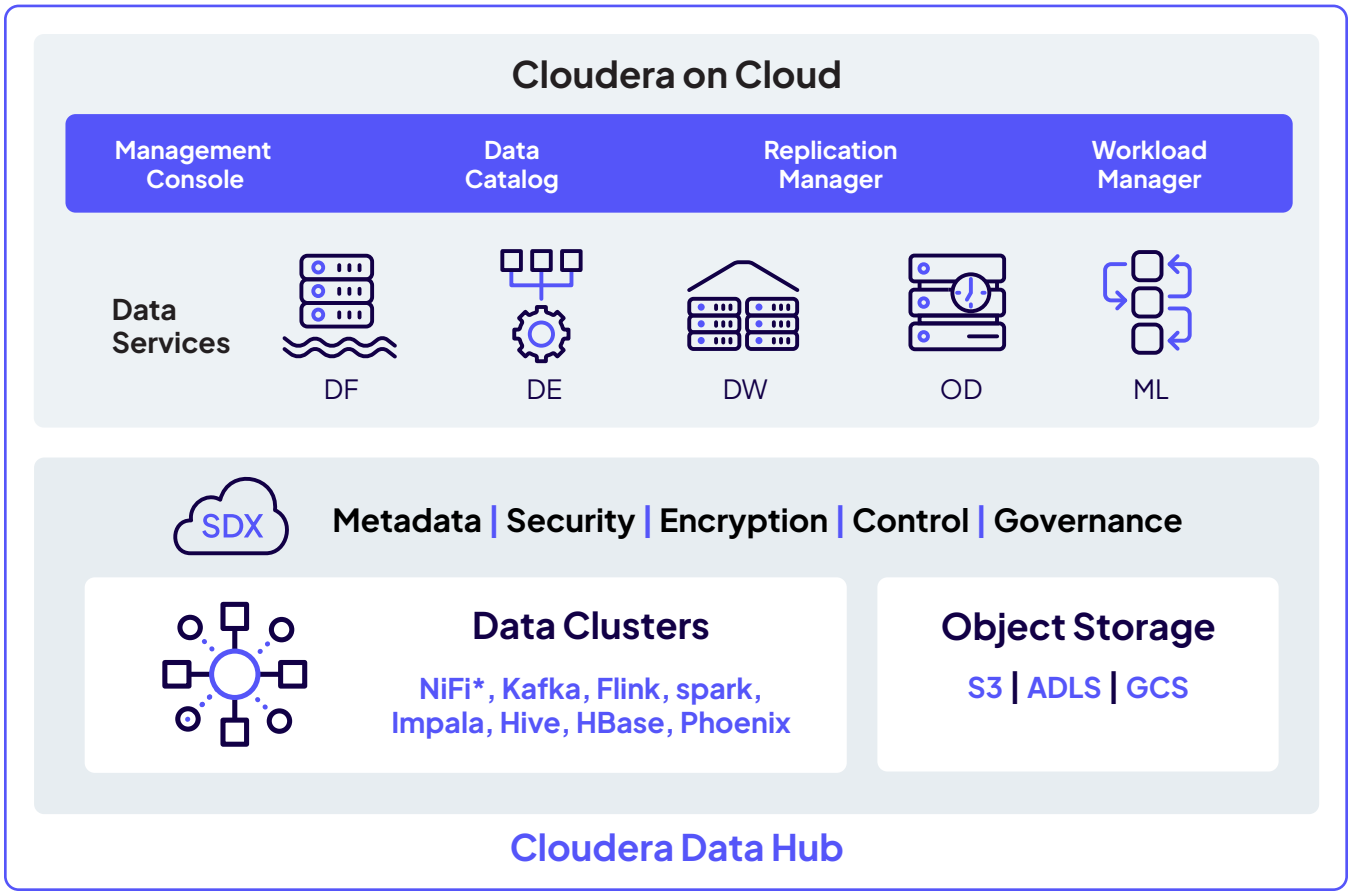


Figure 03 — Services Components of Cloudera on cloud

*Flow Management rate card

Common Control Plane

The Common Controlplane in Cloudera provides a ubiquitous service that is consistent and spans an organizations deployment instances. In the diagram above this shows how a public cloud instance shares services such as governance with the private cloud instance. It goes further in supporting multiple cloud and multiple private cloud deployments. The control-plane is a federated service which enables the metadata, security, encryption and governance to be managed as a centrally, but federated service. The fundamental building blocks are built on Open Source components and have an Open and Accessible API which provides integration to a wider ecosystem of services and supports open standards and Interoperability.

(4) federated governance to enable ecosystems and interoperability.

Cloudera Data Catalog

The Cloudera Data Catalog sits within the Common Control Plane. It addresses many of the foundational qualities of building toward a Data as Product approach. It makes data discoverable across the nodes of the mesh, it provides access to the end-points to make the data addressable, it captures user curated metadata to determine a data products trustworthiness. As well as use-curated metadata, technical and business metadata can be comprehen-sively captured to self-describe the data products. Finally all data is secured at rest and in transit by FIP 140-2 level encryption, and is stored in open formats and open standards. This enables Domain Data as a Product approach.

(2) domain-oriented data served as a product;

Shared Data Experience (SDX)

Cloudera SDX combines enterprise-grade centralized security, governance and management capabilities with shared metadata and a data catalog. It provides a governance layer that spans control planes and deployment instances to assign ownership, capture audit and apply global policies. Cloudera SDX is federated and managed as a shared service on a global and ubiquitously accessible control-plane.

(4) federated governance to enable ecosystems and interoperability;

Data Hub

Cloudera Data Hub allows users to deploy analytical clusters across the entire data lifecycle as elastic IaaS experiences. It provides the greatest control over cluster configurations, including hardware and individual service components installed. Its cloud native design supports separation of compute and storage with the unit of compute being a virtual machine. It provides support for auto scaling of resources based on environmental triggers.

(3) self-serve data infrastructure as a platform to enable autonomous, domain-oriented data teams;

Self-service Data Services

Cloudera Data Services provide containerized compute analytic applications that scale dynamically and can be upgraded independently. Through the use of containers deployed on cloud managed Kubernetes services such as Amazon EKS, Microsoft Azure AKS and Google GKE, users are able to deploy similar clusters to what is possible in Data Hub but with the added advantage of them being delivered as a PaaS experience. Data Flow, Data Engineering, Data Warehousing, Operational Database and Machine Learning are all available as Data Services on Cloudera on cloud.

(3) self-serve data infrastructure as a platform to enable autonomous, domain-oriented data teams;

Section 4: Beyond the Data Mesh

Data Mesh has become a cross industry term used by organizations to explore new innovations in Data Architecture. Thoughtworks coined the term and brought Data Mesh forward in people's thinking and vocabulary. If you have or are building toward a Data Mesh, then you are ahead of the curve.

We would recommend a broader system-wide thinking and we see this in our combined customer base at Cloudera. Let me provide a high level example.

We work with a manufacturing company Scania, helping them drive efficiencies in building vehicles. Their focus is on driving improved yield at a given quality and they approach Data Mesh as a means to integrate a data architecture across their global business.

> **This created a much more efficient and sustainable model of analysis, development and improvement."**

Scania do not operate in isolation, when their trucks are delivered they are then managed in operation by organizations such as Navistar. Navistar monitors the health of the truck and its cargo to ensure it is well maintained, and improves the uptime and availability of trucks in a fleet. Navistar Brings other organizations into its data architecture such as fleet managers to help them understand and make timely decisions on fleet maintenance.

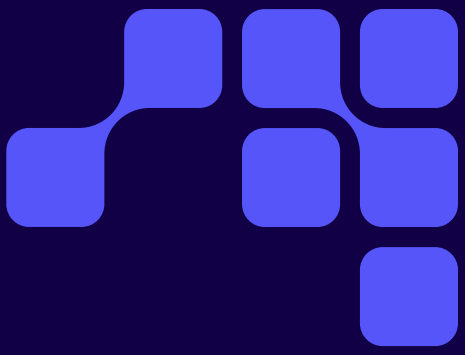
> **With OnCommand Connection, Navistar has helped fleet and vehicle owners reduce maintenance costs by more than 30 percent."**

Now think in terms of a business such as Hello Fresh, who have to connect produce suppliers to customers, and manage the last mile logistics, to make sure the produce is delivered in the right quantities and in a fresh condition. Hello Fresh have recently described how taking a Data Mesh approach has helped them rapidly scale their business and reduce the technical debt of their data architecture. If we step back, each of these organizations and their business units can be thought of as a node on a much wider mesh, with data moving between those organizations as data, each organization having obvious interests in the ownership and governance of their data and data products.

> **Cloudera is at the heart of our data-driven decision making and all internal stakeholders use it to analyze campaign and operations performance at a granular level, and steer the business."**

Moving beyond the Data Mesh is also recognizing that several organizations become dependent on each other for the supply of goods and services, and more-so the use of data and insights as a service. Data Mesh will help your organization think more broadly about your suppliers and customers and how your data products will have a positive impact on them going forward.

About Cloudera



Cloudera is the only true hybrid platform for data, analytics, and AI. With 100× more data under management than other cloud-only vendors, Cloudera empowers global enterprises to transform data of all types, on any public or private cloud, into valuable, trusted insights. Our open data lakehouse delivers scalable and secure data management with portable cloud-native analytics, enabling customers to bring GenAI models to their data while maintaining privacy and ensuring responsible, reliable AI deployments. The world's largest brands in financial services, insurance, media, manufacturing, and government rely on Cloudera to be able to use their data to solve the impossible — today and in the future.

To learn more, visit Cloudera.com and follow us on LinkedIn and X. Cloudera and associated marks are trademarks or registered trademarks of Cloudera, Inc. All other company and product names may be trademarks of their respective owners.

CLOUDERA

Cloudera, Inc. | 5470 Great America Pkwy, Santa Clara, CA 95054 USA | cloudera.com