CLOUDERA

WHITE PAPER

DATA ARCHITECTURE SERIES

THE UNIFIED DATA FABRIC

Building a Modern Data Fabric with Cloudera

Abstract

This whitepaper provides an introduction to the Data Fabric 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 Fabric can be extended.

Version: 1.0
Author: Jean-Philippe Player

Table of Contents

Introduction

In this section we briefly summarise 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 Fabric architecture, specifically as it applies to Cloudera's products and services.

Purpose

The Data Fabric is one of three important emerging data architectures; the other two are Data Mesh 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.

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.

- Enterprise Data Fabric Enables DataOps | Forrester
- How a Big Data Fabric con Transform Your Data Architecture
- Conquering hybrid and multi-cloud with big data fabric
- CDP Data Catalog overview
- Shared Data Experience (SDX) | Cloudera

Why is a Data Fabric Architecture Needed

Enterprises today have to contend with exponentially increasing volumes of batch and streaming data, comprising a variety of structured, unstructured, and semi-structured data types, and originating from an expanding number of disparate sources located on-premises, in the cloud and at the edge.

At the same time, business users demand faster and easier access to reliable, trusted, up-to-date data to make accurate business decisions.

Traditional data approaches require spending a lot of time on manually preparing data, managing ingestion, standardizing data sets and orchestrating data movement between on premises and cloud environments. Finding the right data sets and making them available for analytics is often a convoluted process that further slows down business decisions. That is compounded by regulatory compliance and security controls that must be manually applied at every step of the data lifecycle, from ingestion to analytical applications.

The Data Fabric has emerged as a modern data architecture to overcome these challenges and supports the needs of a hybrid, multi-cloud environment. The architecture focuses on making data readily available to business users wherever it resides, improve collaboration, enable self-service, and leverage automation to simplify data management and enforce the necessary compliance and security requirements.

What is a Data Fabric ?

Definition

The Data Fabric offers a comprehensive approach to centrally manage, own, curate, secure and govern enterprise data across multiple clouds and on premises. Forrester coined the term and their definition of a Data Fabric is as follows:

"A Data Fabric orchestrates disparate data sources intelligently and securely in a self-service manner, leveraging data platforms such as data lakes, data warehouses, NoSQL, translytical, and others to deliver a unified, trusted, and comprehensive view of customer and business data across the enterprise to support applications and insights."

Properties

A modern Data Fabric comprises multiple layers that work together to meet these needs:

1. Data management
2. Data ingestion and streaming
3. Data processing and persistence
4. Data orchestration
5. Data discovery
6. Global data access

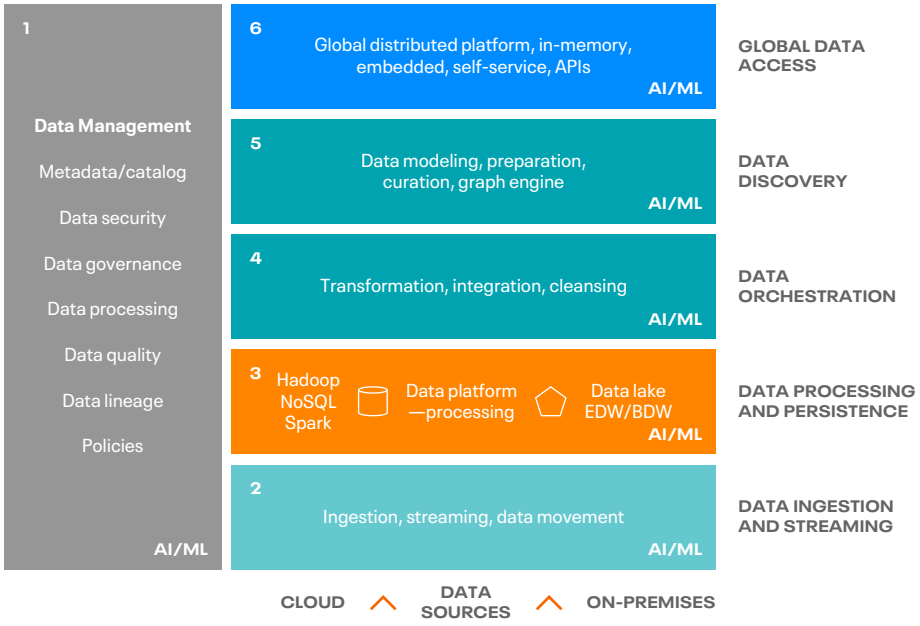


Figure 01—Enterprise Data Fabric Reference Architecture, Enterprise Data Fabric Enables DataOps, Forrester, August 2, 2021

The Data Management layer is the core layer of the architecture and provides the needed tools and interfaces for all the other layers. It is what provides the end-to-end data management capabilities that ensure the reliability, security and governance of data and is the main focus of this paper.

How to build a Data Fabric with Cloudera

Cloudera Data Platform (CDP) has been built from the ground up to support hybrid, multi-cloud data management in support of a Data Fabric architecture. In this section we provide an introduction to CDP, with a focus on the data management capabilities that enable the Data Fabric.

Cloudera Data Platform (CDP)

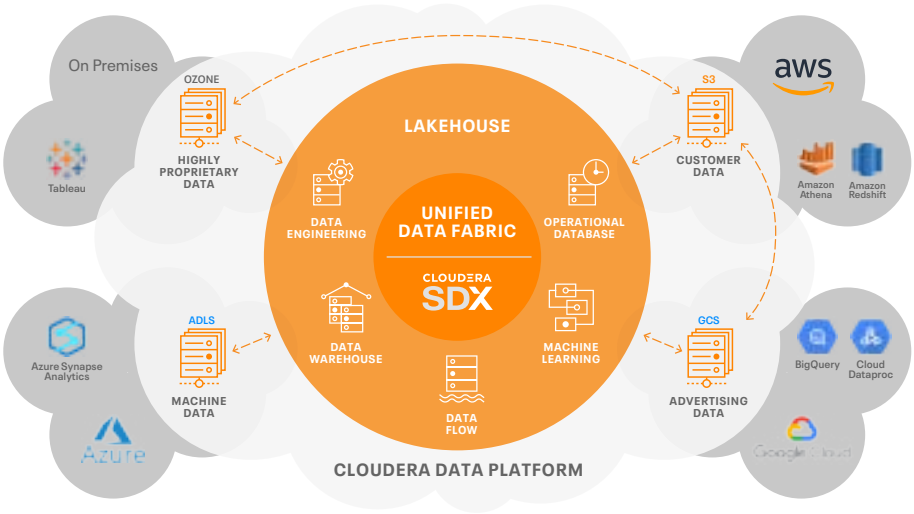
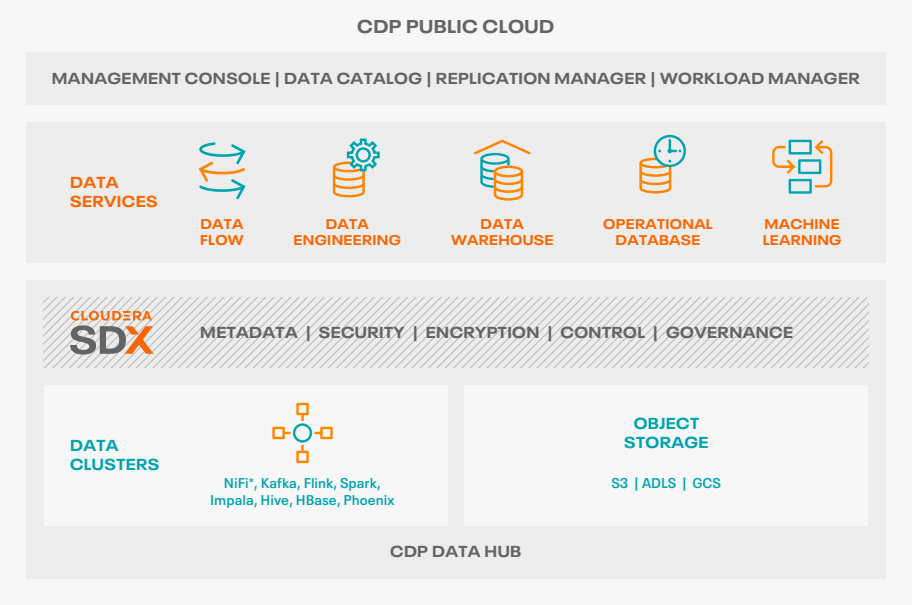


Figure 02—Cloudera Data Platform (CDP)

Cloudera Data Platform (CDP) is a hybrid data platform designed to provide the freedom to choose any cloud, any analytics, any data. CDP delivers faster and easier data management and data analytics for data anywhere, with optimal performance, scalability, and security. With CDP you get the value of CDP Private Cloud and CDP Public Cloud for faster time to value and increased IT control.

Cloudera Data Platform 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. As a result, the platform is perfectly placed to implement modern data architectures:

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



*Flow Management rate card

Figure 03—Services Components of CDP Public Cloud

Figure 03 provides a summary of the logical components that make up CDP in the Public Cloud. We'll now explore how each of these components support the Data Fabric.

Common Control Plane

The Common Control Plane in CDP provides a ubiquitous service that is consistent and spans an organization's 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 central, but federated service. The fundamental building blocks are based 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.

Data Catalog

The CDP Data Catalog sits within the Common Control Plane. This global catalog provides a searchable inventory of all the assets that are part of the Data Fabric, making data assets easily discoverable.

- Comprehensive—Support for all entities that make up the hybrid cloud ecosystem: Hive tables, Kafka topics, Nifi flow, HBase tables, Machine Learning Models, etc. Each asset will be displayed alongside its contextual metadata, such as schema, security policies, tags and classifications, profile, governance rules and business annotations.
- Discoverability—Single location to discover and search for data from all nodes of the Fabric.
- Governance—Built in profiling to give insights into data quality and sensitivity, built in classification engine that assigns security, compliance and policy related attributes such as PII.
- Lineage—Automatic capture of lineage information helps understand where the data came from, how it is being used, what impact changes would have. It can further be extended to propagate security policies across the entire Data Fabric, making it safer and easier to share data.
- Policy—Security, Compliance and Governance policies can be assigned to any data asset directly from the Catalog.

- Security—Complete audit log of all access and modifications made to data sets locate anywhere in the fabric.
- Collaboration—Supports business annotations and metadata, curation and collaboration.

This addresses the requirements of the Data Management layer (1) of the Data Fabric, when deployed in conjunction with the Shared Data Experience (SDX).

Shared Data Experience (SDX)

Cloudera SDX combines enterprise-grade security, governance and management capabilities with shared metadata that is deployed locally in each node of the Data Fabric, and federated via the Control Plane. It provides a governance layer that is truly global—spanning control planes and deployment instances to assign ownership, capture audit and apply global policies across on premises deployments and public clouds.

- Metadata—establish information assets for increased usability, trust and value leveraging all metadata (structural, operational, business and social).
- Security—granular, dynamic, role- and attribute-based security policies. Prevent and audit unauthorized access to sensitive or restricted data across platform.
- Encryption—strong cryptography for data in motion and rest, centralized authentication with single-sign on.
- Control—move data and workloads between deployments for optimum performance, cost and resilience, meeting ever changing business needs.
- Governance—enterprise-grade auditing, lineage, and governance capabilities applied across the platform with rich extensibility for partner integrations.

Replication Manager

The Replication Manager is designed to serve a number of use cases around cross-fabric data orchestration and replication: workload migration, cloud bursting, backup and disaster recovery, and replication in support of development and test systems. It supports full and incremental replication for all data storage types available in the fabric.

A key tenet of a Unified Data Fabric is having consistent security and governance controls across all fabric endpoints. Tightly integrated with SDX, Replication Manager supports that function[1] by moving policies with the data, replicating all associated metadata, classification tags, security policies, compliance rules and lineage information.

Global Unified Security with SDX

SDX supports attribute-based policies through the use of tags, such as "PII", that can be assigned to any data asset including individual columns of a table. The data access policy for PII data can be specified by a centralized team responsible for enterprise-wide rules, while the tag itself can be assigned by the creator of the data set, either manually or via automatic classification. The automatic capture of lineage information through the data pipeline enables tag inheritance, and as such propagation of the relevant policies within a local node in the fabric autonomously.

Replication Manager is aware of these tags. As data is moved between environments, classification tags are also automatically propagated and assigned to the data. This enforces the appropriate policies globally across the nodes of the fabric, and provides unified policy management and compliance across all of the organization's environments, while allowing business users self-service access to trusted data.

Source
[1] Available from second half of 2022

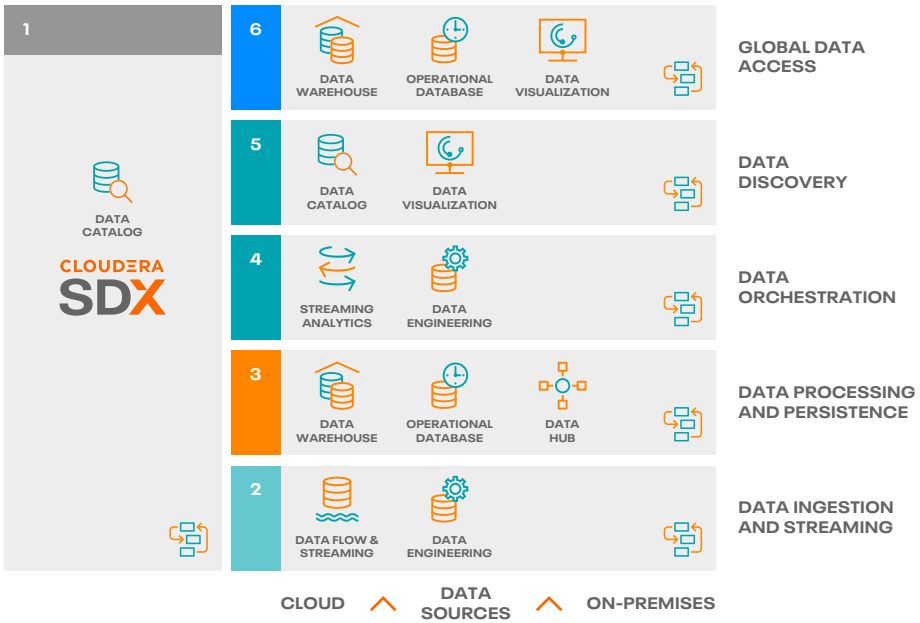


Figure 04—Key CDP Components that make up the Data Fabric architecture

Data Services

The Layers 2 through 6 of the Data Fabric are addressed using Cloudera Data Services, see figure 04:

- Data ingestion and streaming—provided by Cloudera Data Flow (CDF) and Cloudera Data Engineering (CDE)
- Data processing and persistence—provided by Cloudera Data Hub (CDH), Cloudera Data Warehouse (CDW) and Cloudera Operational Database (COD)
- Data orchestration—provided by components embedded in Cloudera Data Engineering (CDE) and Cloudera Streaming Analytics (CSA)
- Data discovery—provided by Cloudera Data Visualization (CDV) and the Cloudera Data Catalog
- Global data access—provided by Cloudera Data Warehouse (CDW), Cloudera Operational Database (COD) and Cloudera Machine LEarning (CML)

Beyond the Data Fabric

The Data Fabric takes a centralized approach to all aspects of data management. As organizations scale, moving to a distributed model for managing the data domains can be beneficial and is encapsulated in the concept of the Data Mesh, which is the subject of a separate white paper in the series.

CLOUDERA

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