CLOUDERA
Educational Services

Cloudera Essentials for CDP

191204

CLOUDERA

Introduction

Chapter 1

Course Chapters

- **Introduction**
- Introducing the Enterprise Data Cloud
- Cloudera Data Platform Overview
- Workload and Data Management
- Data in Motion
- Data Warehousing and Analytics
- Data Science and Machine Learning
- Security and Data Governance
- Planning for Success
- Conclusion

Trademark Information

- **The names and logos of Apache products mentioned in Cloudera training courses, including those listed below, are trademarks of the Apache Software Foundation**

Apache Accumulo	Apache Hive	Apache Pig
Apache Avro	Apache Impala	Apache Ranger
Apache Ambari	Apache Kafka	Apache Sentry
Apache Atlas	Apache Knox	Apache Solr
Apache Bigtop	Apache Kudu	Apache Spark
Apache Crunch	Apache Lucene	Apache Sqoop
Apache Druid	Apache Mahout	Apache Storm
Apache Flume	Apache NiFi	Apache Tez
Apache Hadoop	Apache Oozie	Apache Tika
Apache HBase	Apache Parquet	Apache Zeppelin
Apache HCatalog	Apache Phoenix	Apache ZooKeeper

- **All other product names, logos, and brands cited herein are the property of their respective owners**

Course Objectives (1)

During this course, you will learn

- **Which characteristics define the Enterprise Data Cloud**
- **What Cloudera Data Platform is and what capabilities it provides**
- **How the Cloudera Data Platform supports both on-premises and cloud-based deployments**
- **How organizations use streaming data and the Internet of Things (IoT) to improve efficiency**
- **How companies are using Cloudera data warehouse tools to better understand their business**

Course Objectives (2)

- How data scientists use Cloudera's products and services to help organizations benefit from machine learning
- How Cloudera helps organizations meet requirements for data security and governance
- What roles and skills organizations should look for when building data teams
- What resources are available to assist with planning, implementing, and supporting a Cloudera solution
- What factors should one consider before moving to the cloud

CLOUDERA

Introducing the Enterprise Data Cloud

Chapter 2

Course Chapters

- Introduction
- **Introducing the Enterprise Data Cloud**
- Cloudera Data Platform Overview
- Workload and Data Management
- Data in Motion
- Data Warehousing and Analytics
- Data Science and Machine Learning
- Security and Data Governance
- Planning for Success
- Conclusion

Chapter Topics

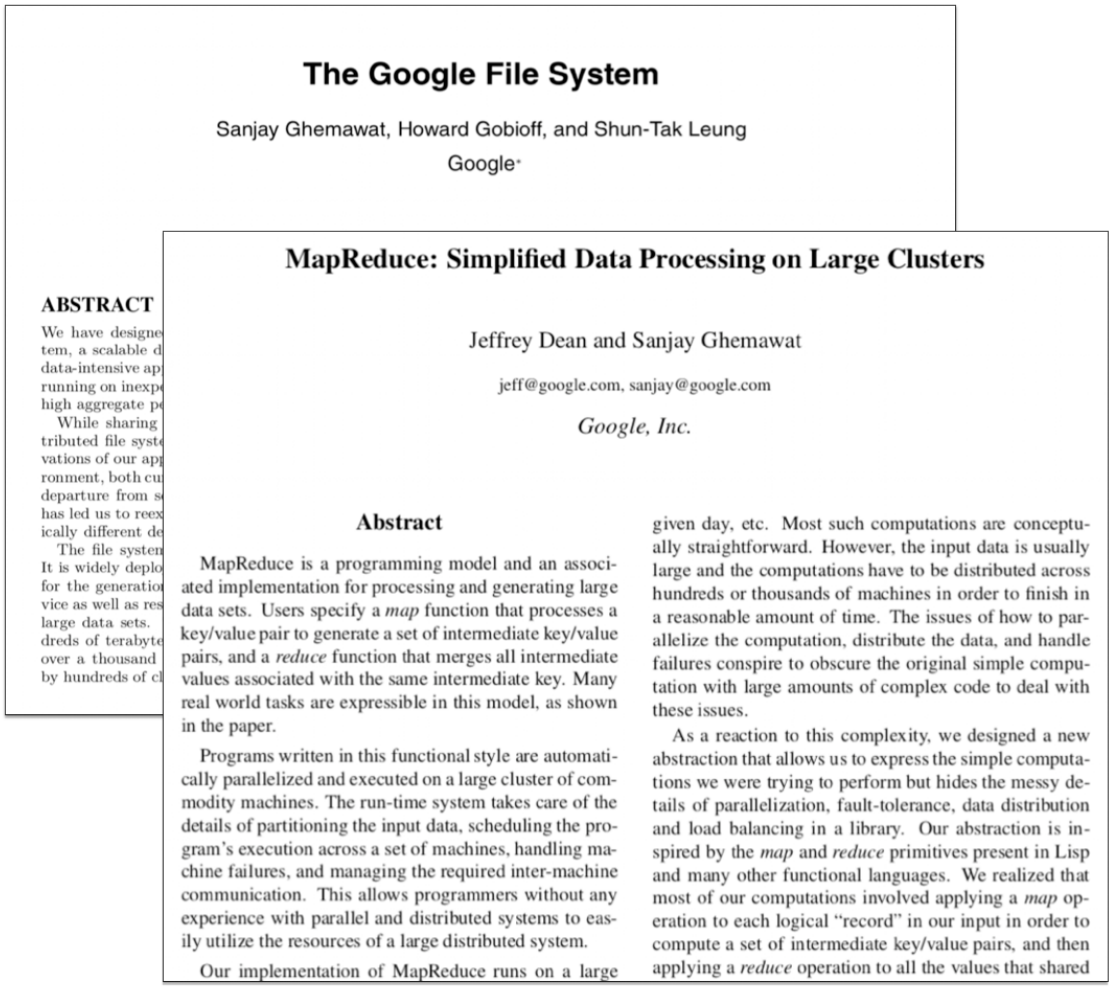
Introducing the Enterprise Data Cloud

- **The Evolution of CDP**
- Characteristics of an Enterprise Data Cloud
- From the Edge to AI: An End-to-End Use Case
- Essential Points

What is CDP?

Cloudera Data Platform

History: New Approaches for Data Storage and Processing



The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung

Google*

ABSTRACT

We have designe
tem, a scalable d
data-intensive ap
running on inexp
high aggregate p
While sharing
tributed file syst
vations of our ap
ronment, both cu
departure from s
has led us to reex
ically different de
The file systen
It is widely deplo
for the generatio
vice as well as res
large data sets.
dreds of terabyte
over a thousand
by hundreds of cl

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.

Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system.

Our implementation of MapReduce runs on a large given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

As a reaction to this complexity, we designed a new abstraction that allows us to express the simple computations we were trying to perform but hides the messy details of parallelization, fault-tolerance, data distribution and load balancing in a library. Our abstraction is inspired by the *map* and *reduce* primitives present in Lisp and many other functional languages. We realized that most of our computations involved applying a *map* operation to each logical "record" in our input in order to compute a set of intermediate key/value pairs, and then applying a *reduce* operation to all the values that shared

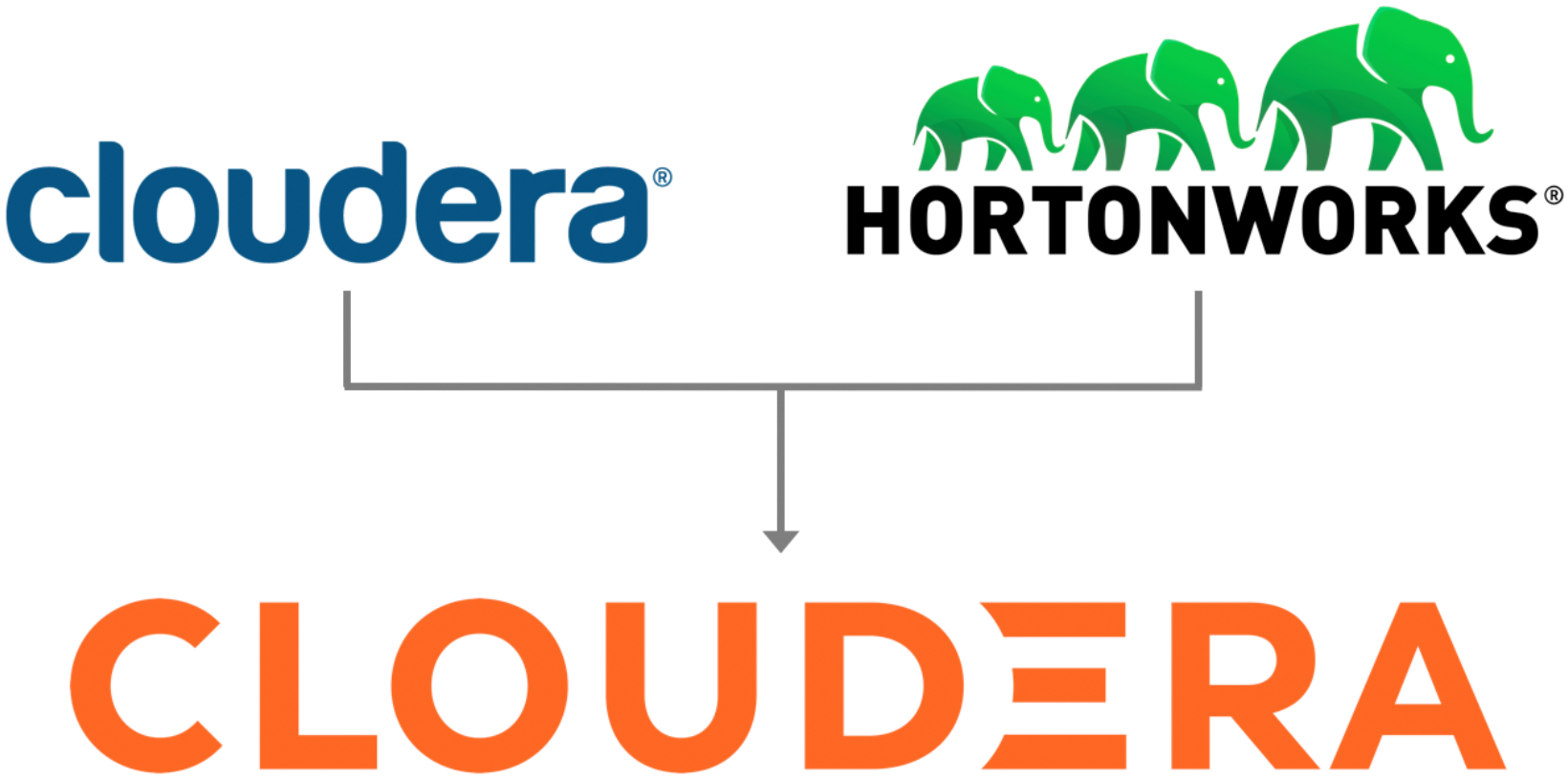
Our Heritage is Open Source

The Original Data Platforms: CDH and HDP

Cloudera CDH	Hortonworks HDP
Apache Hadoop	Apache Hadoop
Apache Hive	Apache Hive
Apache Spark	Apache Spark
Apache Flume	Apache NiFi
Apache Kudu	
	Apache Druid

The Platform Evolves

2019: Cloudera and Hortonworks Merge



Chapter Topics

Introducing the Enterprise Data Cloud

- The Evolution of CDP
- **Characteristics of an Enterprise Data Cloud**
- From the Edge to AI: An End-to-End Use Case
- Essential Points

Characteristics of the Enterprise Data Cloud

Hybrid &
Multi-Cloud

EDGE
2AI

Multi-Function

CLOUDERA
SDX

Secure &
Governed

Open

Hybrid and Multi-Cloud

- **Runs where you do**
- **Helps you to control cost**
- **Avoids cloud vendor lock-in**

Multi-Function

- **Supports many types of workloads**
- **Avoids data duplication**
- **Reduces operational expense**

Secure and Governed

- **Protect sensitive data across all environments**
- **Regulatory compliance**
- **Respond to business opportunities**

Open

- **Open source**
- **Open for integration**
- **Reduces technology and business risk**

Chapter Topics

Introducing the Enterprise Data Cloud

- The Evolution of CDP
- Characteristics of an Enterprise Data Cloud
- **From the Edge to AI: An End-to-End Use Case**
- Essential Points

Traditional Maintenance

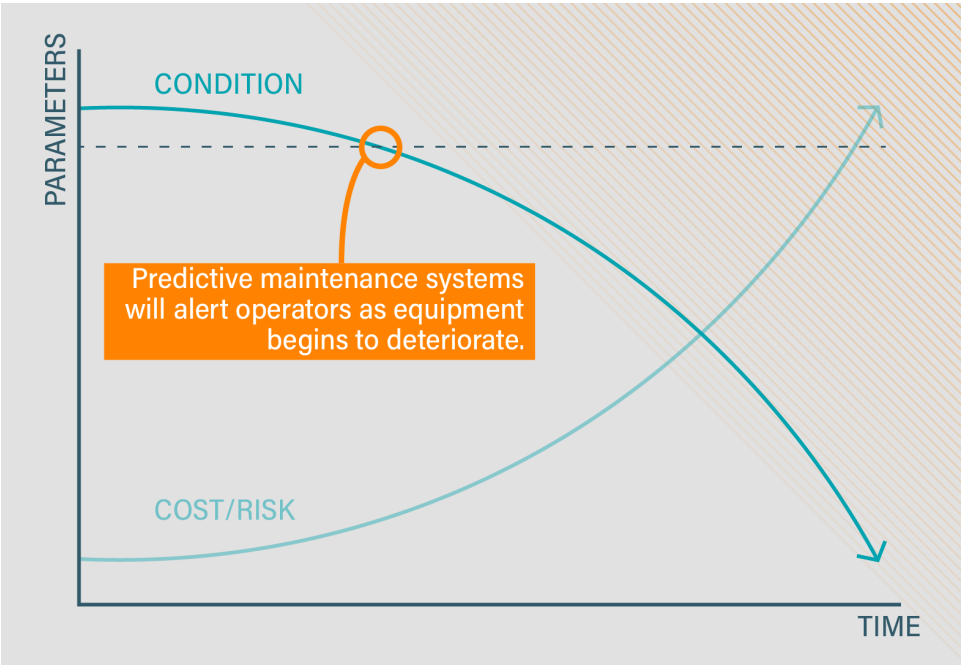
- **Maintenance must be regularly scheduled for factory equipment**
  - During maintenance, equipment is unavailable
  - Equipment downtime means a loss in profit and productivity
- **Maintenance was traditionally either *reactive* or *preventive***
  - *Reactive* maintenance: Respond to equipment failures after they occur
    - *Example: Updating a light fixture whenever a bulb burns out*
  - *Preventive* maintenance: Prevent such failures through routine check-ups
    - *Example: Changing a car's oil every 3000 miles*

A Metric for Production

- ***Overall equipment effectiveness* (OEE) is a key performance indicator**
- **OEE is determined by three factors:**
  - *Availability*: Percentage of time the equipment is running when needed
  - *Performance*: Relative quantity of generated output
  - *Quality*: Relative grade of the output
- **Used to determine which maintenance procedures yield the highest profits**
- **Downtime due to traditional maintenance diminishes OEE significantly**

Predictive Maintenance

- **A *predictive* maintenance system leverages the problems faced by more reactive systems**
  - Incorporates data about *real-time* condition of equipment
  - Also analyzes real-time environmental data, such as weather or acoustic readings
  - Compares data to previous readings and against data from other equipment



- **Used to minimize deterioration of equipment and optimize maintenance procedures**

Predictive Maintenance and the Internet of Things (IoT)

- **The *Internet of Things* (IoT) is the intersection of computers with everyday objects**
- **Predictive maintenance architectures typically use IoT-based data sources**
- **Predictive maintenance often leverages existing or custom-built IoT-based architectures**
- **IoT-based maintenance can result in positive business outcomes**
  - Cut 10-40% of maintenance costs on factory equipment
  - Increase OEE of equipment significantly

A Characterization of Data

- **Predictive maintenance applications support machine learning algorithms and support the data that drives the platform**
- **That data has three important characteristics**
  - *Volume*: IoT generates petabytes of data
  - *Variety*: Different data sources create different data types
  - *Velocity*: New data passes into the streaming data processor at a quick rate
- **The application must be versatile and accept data in motion on top of data at rest**

Predictive Maintenance Workflow

- **Ingest**
  - Takes data from sensors as input
- **Store**
  - Sensor data stored in large database
- **Process**
  - Data compressed into a usable format
- **Analyze**
  - Machine learning predicts when components need replacement

Cloudera Customer Experience: Lufthansa Technik

- **Lufthansa Technik**
  - Uses Cloudera's platform to optimizes aircraft availability and reliability
  - Provides its broad range of predictive services to 800 customers
  - Observed a 40% decrease in component removal
  - Experience a significant reduction in airline operating costs

Cloudera Customer Experience: Navistar

- **Navistar**
  - Manufactures vehicles with predictive maintenance capabilities
  - Collects telematics and sensor data from more than 300,000 connected vehicles
  - Reports that predictive maintenance has reduced maintenance cost and downtime by 40%

Chapter Topics

Introducing the Enterprise Data Cloud

- The Evolution of CDP
- Characteristics of an Enterprise Data Cloud
- From the Edge to AI: An End-to-End Use Case
- **Essential Points**

Essential Points

- **CDP is the Cloudera Data Platform**
- **It's an implementation of the Enterprise Data Cloud**
  - Built on the foundation of earlier Cloudera/Hortonworks platform
- **Characteristics of an Enterprise Data Cloud**
  - Hybrid and multi-cloud
  - Multi-function
  - Secure and governed
  - Open

Bibliography

The following offer more information on topics discussed in this chapter

- **Customer Success Story: Lufthansa Technik**
  - http://tiny.cloudera.com/cdpec02a
- **Customer Success Story: Navistar**
  - http://tiny.cloudera.com/cdpec02b
- **Faurecia: Using IoT Analytics to Reduce Cost and Improve Quality**
  - http://tiny.cloudera.com/cdpec02c
- **Journey to the Enterprise Data Cloud with Doug Cutting**
  - http://tiny.cloudera.com/cdpec02d

CLOUDERA

Cloudera Data Platform Overview

Chapter 3

Course Chapters

- Introduction
- Introducing the Enterprise Data Cloud
- **Cloudera Data Platform Overview**
- Workload and Data Management
- Data in Motion
- Data Warehousing and Analytics
- Data Science and Machine Learning
- Security and Data Governance
- Planning for Success
- Conclusion

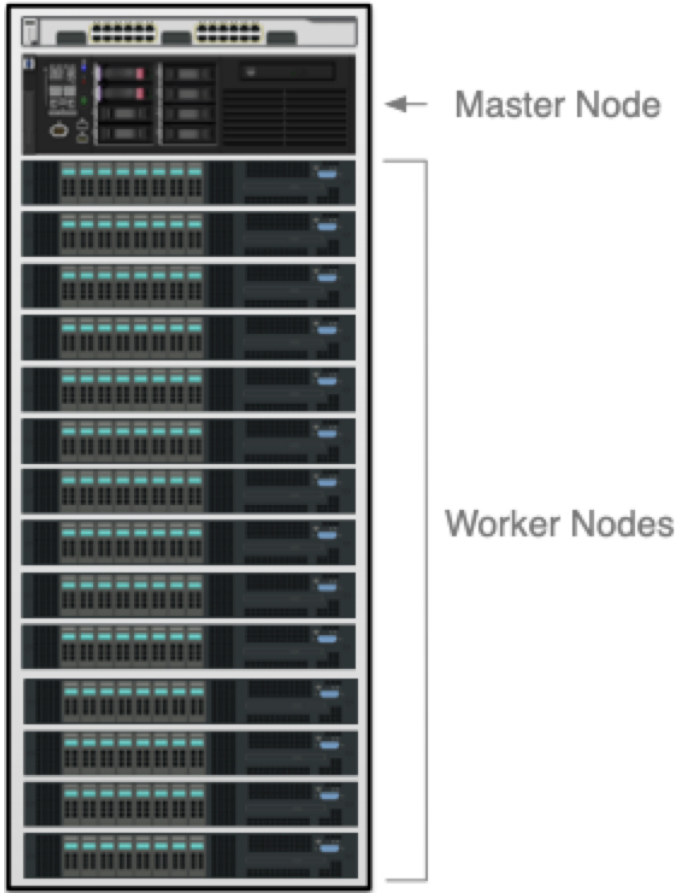
Chapter Topics

Cloudera Data Platform Overview

- **Key Concepts**
- Cloudera Runtime
- Core Services
- Types of CDP Deployments
- Custom Deployments with Data Hub
- Self-Service Experiences
- A Tour of the CDP User Interface
- Essential Points

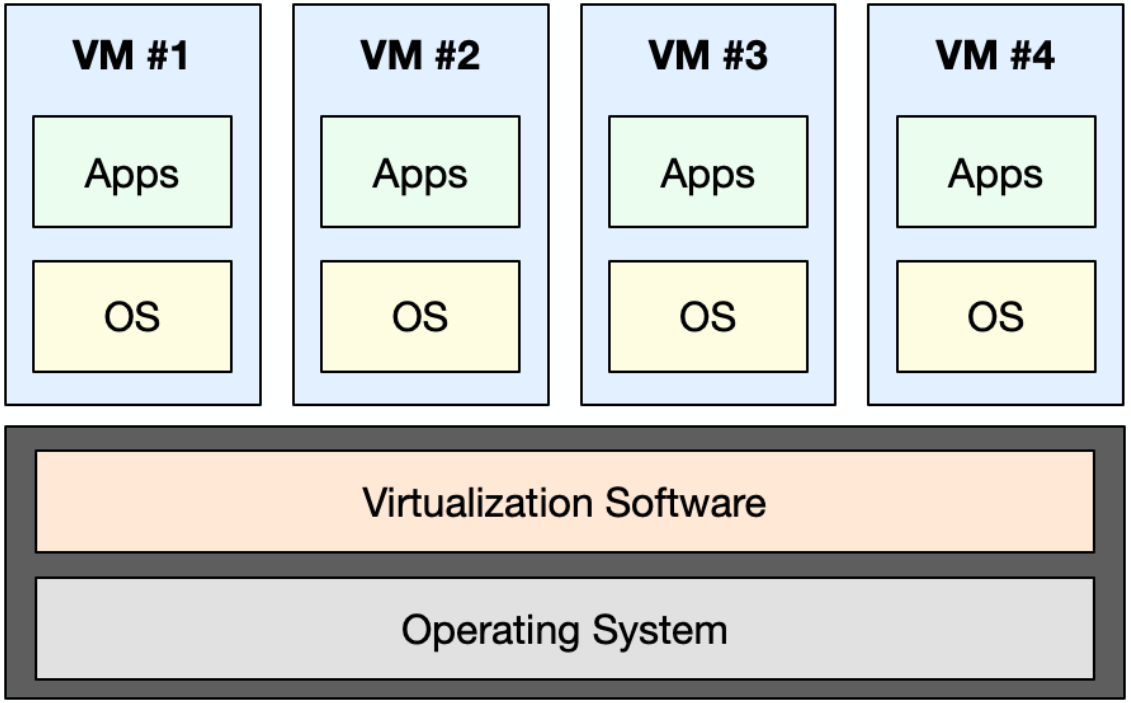
System Architecture: Running on Bare Metal

- **Deployed directly on server hardware**
- **A cluster is a group of related servers**
  - These servers are called *nodes*
  - Clusters may have thousands of nodes
- **There are two main types of nodes**
  - Master (manages resources and status)
  - Worker (performs the actual work)
- **Nodes both store and process data**
  - Hadoop's storage layer is called HDFS



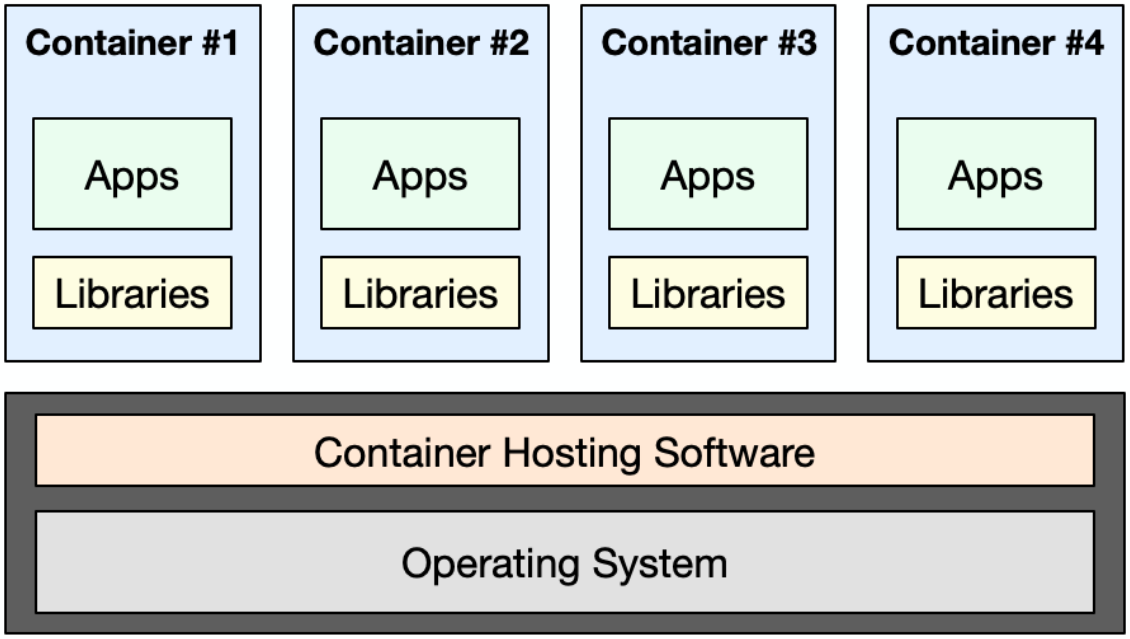
System Architecture: Virtualization

- **Virtualization abstracts away details of server hardware**
  - Makes it possible to run multiple virtual machines (VMs) on a single server
  - Each VM has its own operating system and applications
- **Amazon EC2 and Azure Virtual Machines are cloud services for VMs**



System Architecture: Cloud-Native Applications

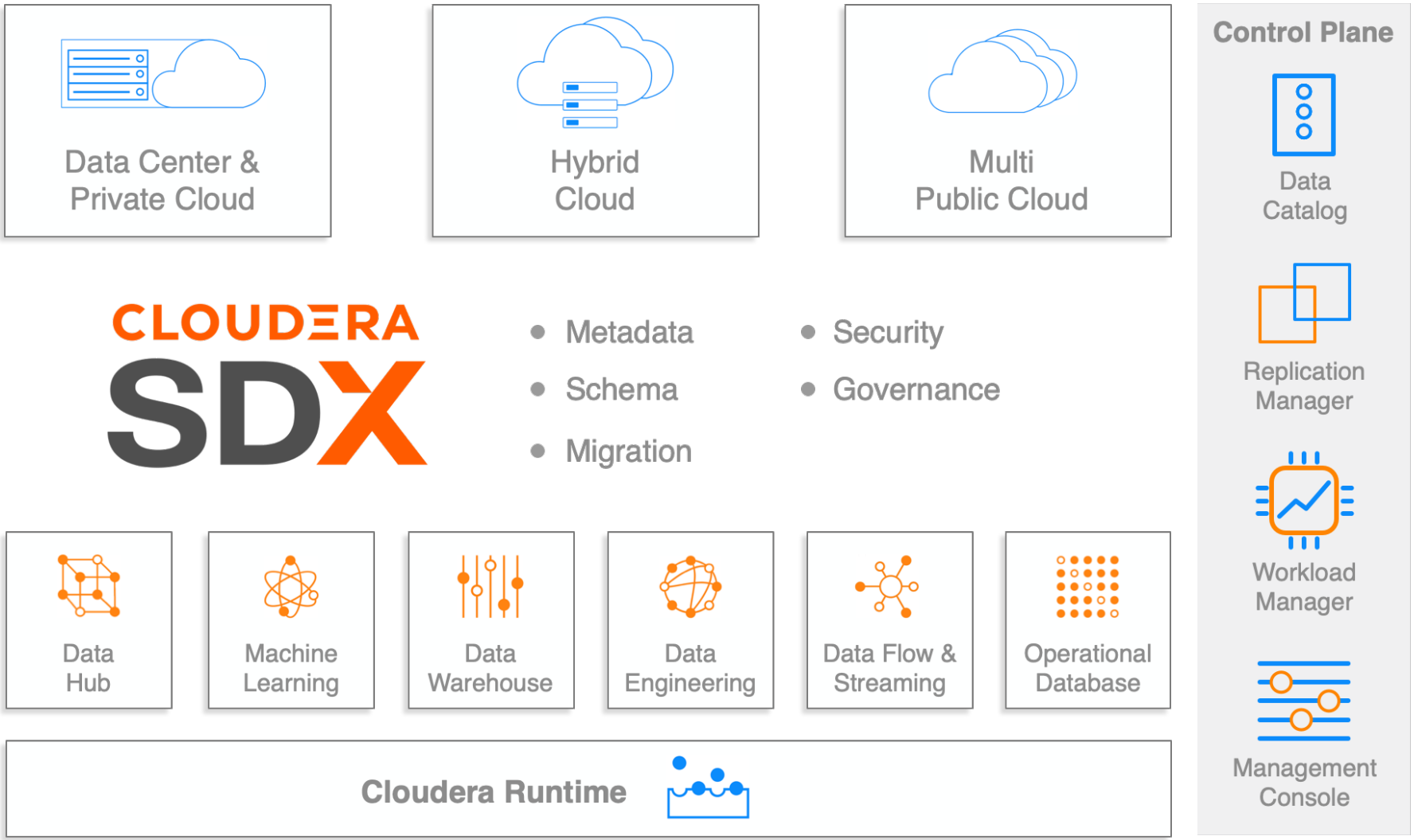
- **Virtual machines work well, but consume a lot of resources**
  - Each VM has its own complete operating system
- **Containerization offers a more efficient alternative to virtualization**
  - Each container shares the underling operating system



What is CDP?

- Not just a new release of an old platform
- Not just a “unity” release of CDH and HDP
- Not just a single product
- It’s the first implementation of the Enterprise Data Cloud

The Cloudera Data Platform

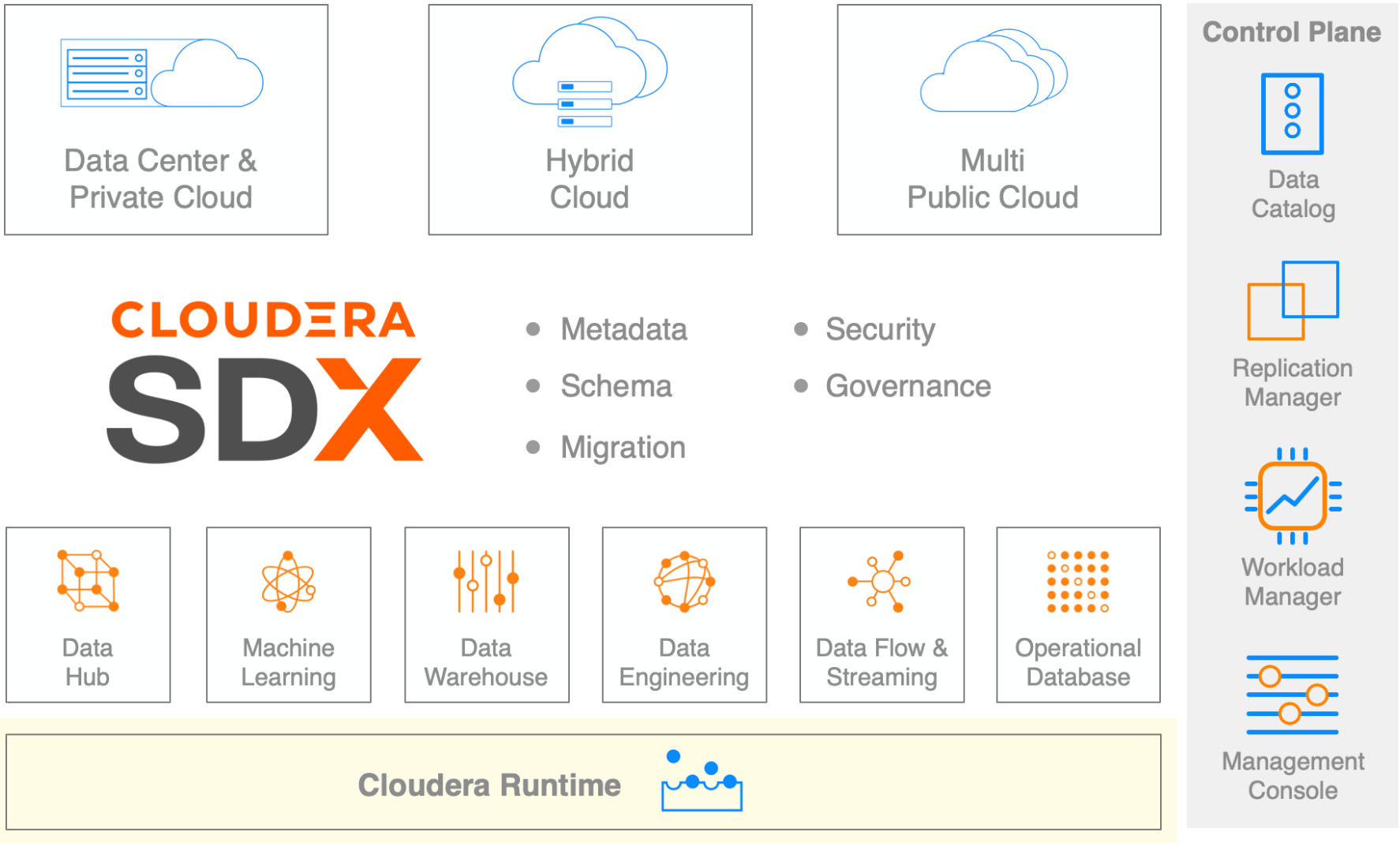


Chapter Topics

Cloudera Data Platform Overview

- Key Concepts
- **Cloudera Runtime**
- Core Services
- Types of CDP Deployments
- Custom Deployments with Data Hub
- Self-Service Experiences
- A Tour of the CDP User Interface
- Essential Points

Cloudera Runtime



The Role of Cloudera Runtime

- Open source distribution of Big Data components
- Result of CDH and HDP convergence
- The foundation of CDP

Changes in Platform Components

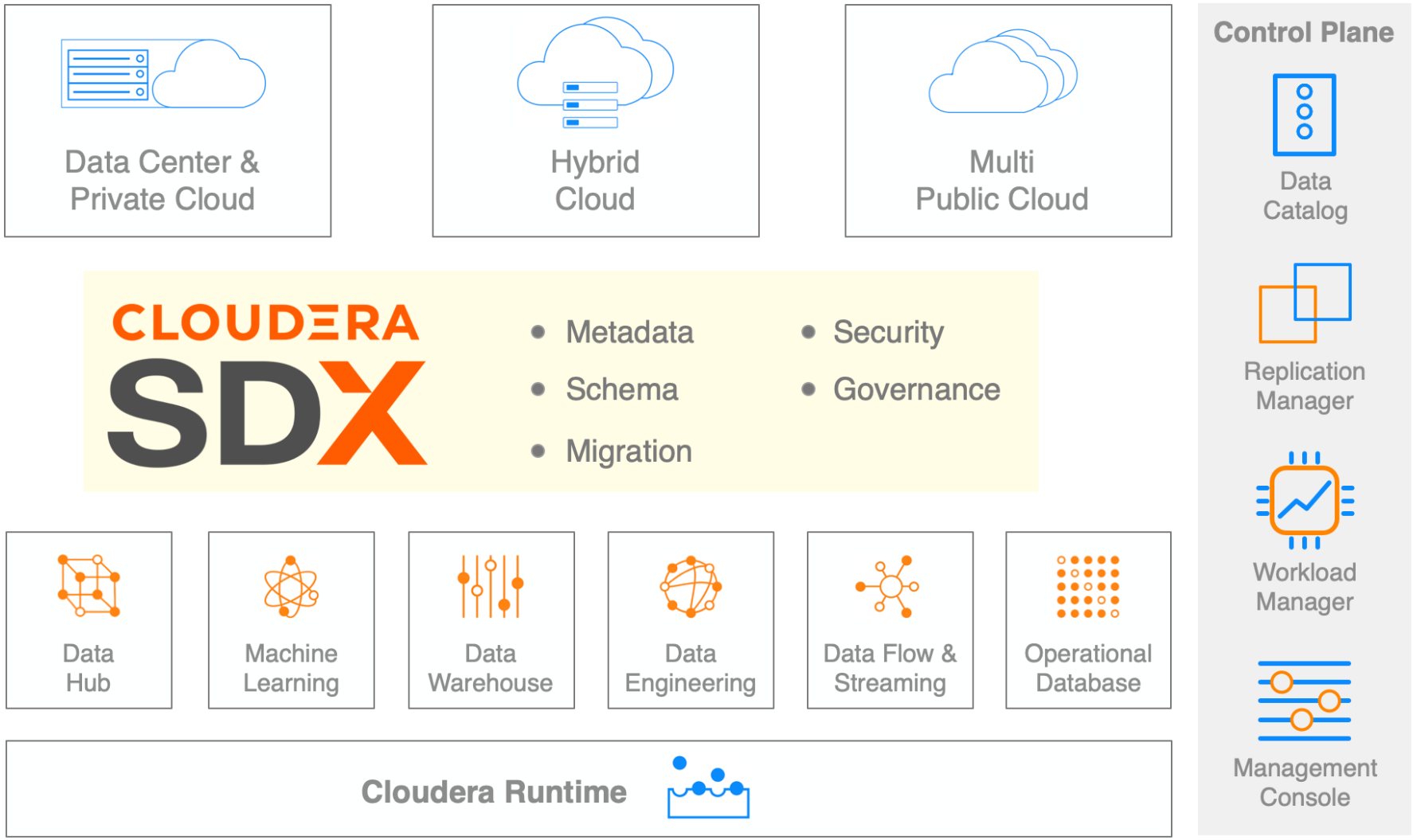
- **Product team primarily used multiple strategies to address overlap**
  - Choose the better of two components
  - Retain both components
- **Newer versions of nearly all components**
- **We will periodically update Cloudera Runtime**
  - Check CDP documentation for a list of components and versions

Chapter Topics

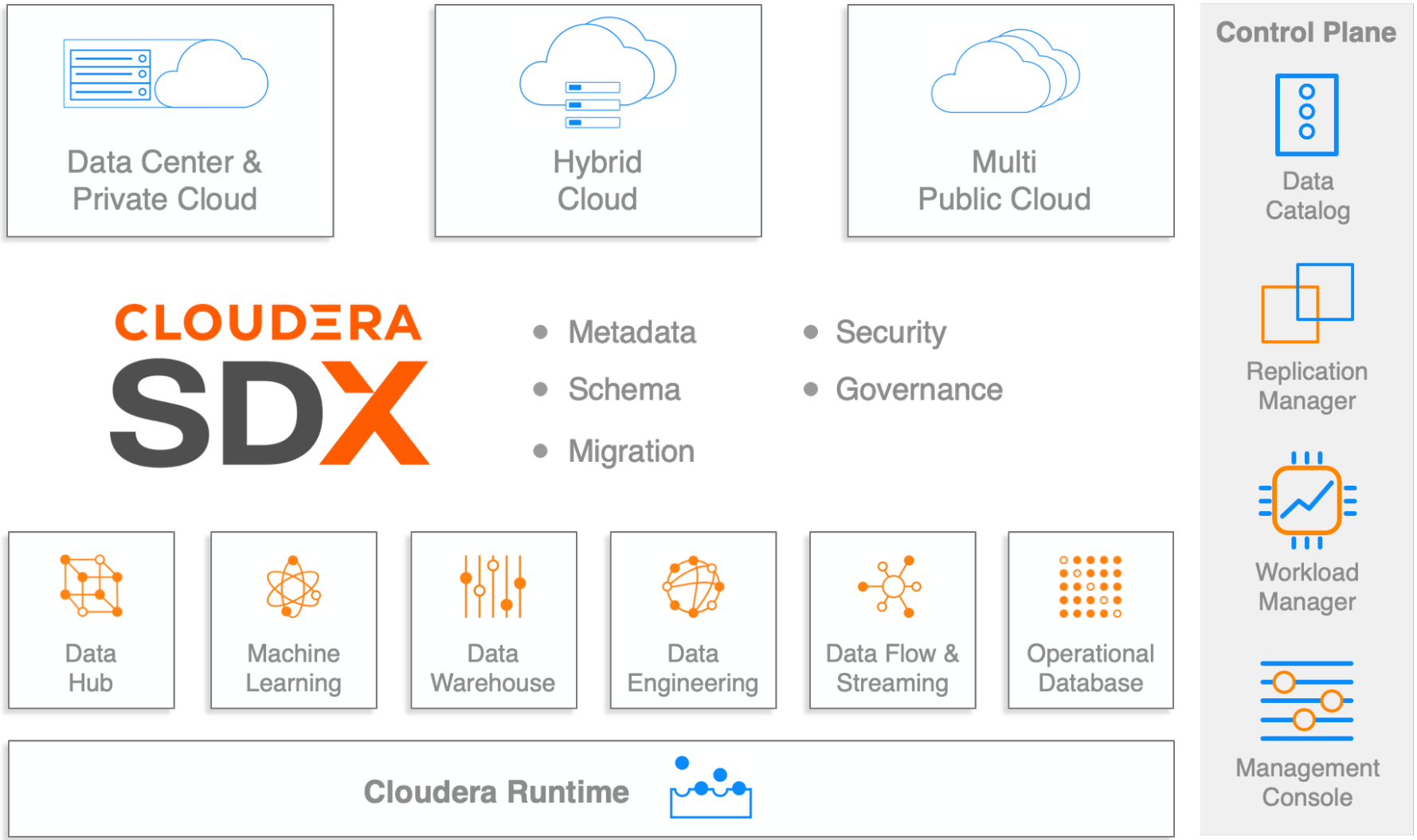
Cloudera Data Platform Overview

- Key Concepts
- Cloudera Runtime
- **Core Services**
- Types of CDP Deployments
- Custom Deployments with Data Hub
- Self-Service Experiences
- A Tour of the CDP User Interface
- Essential Points

SDX

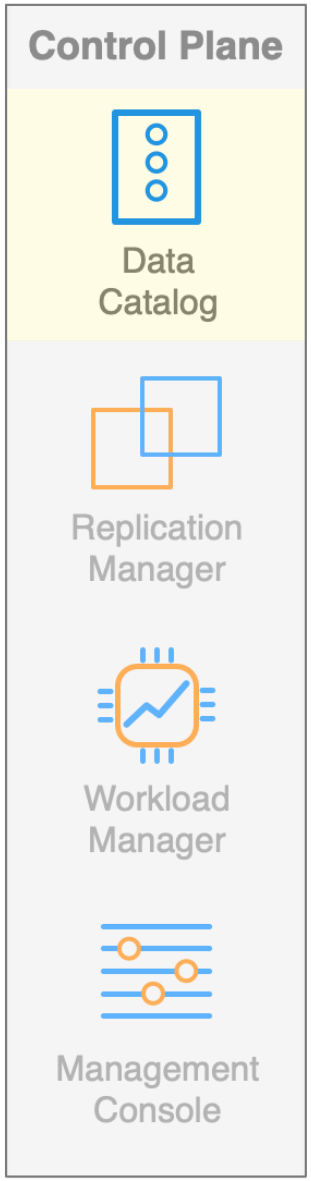


Control Plane



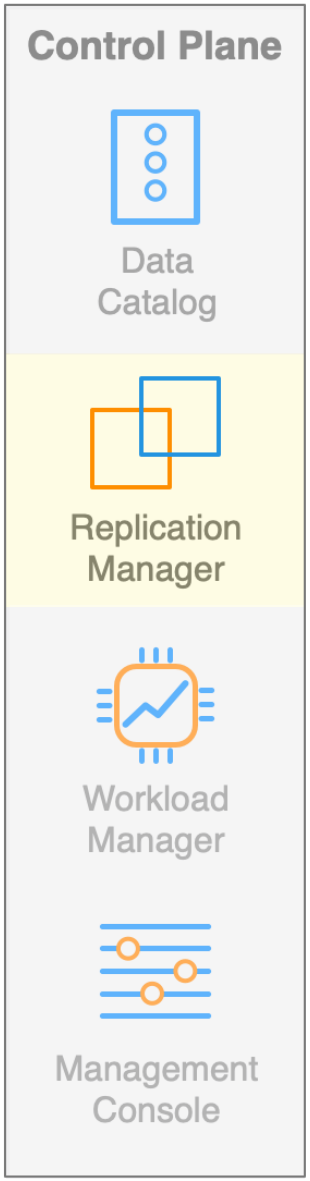
Data Catalog

- **Primarily intended for data stewards**
  - Helps them to understand, manage, and secure data assets



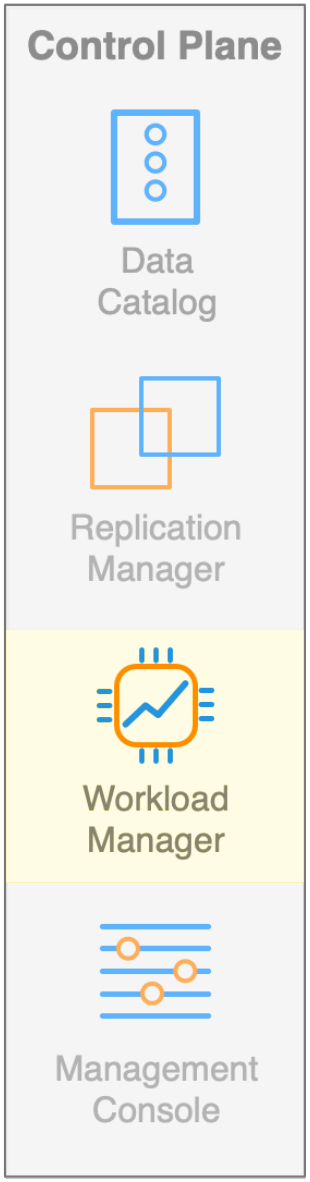
Replication Manager

- **Primarily intended for administrators**
  - Used to replicate migrate data and metadata between environments



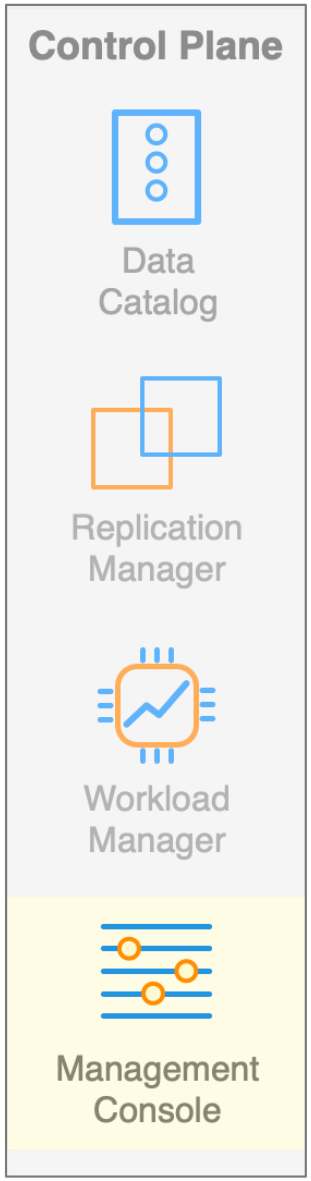
Workload Manager

- **Primarily intended for database administrators**
  - Used to analyze, troubleshoot, and optimize workloads



Management Console

- **Primarily intended for administrators**
  - Provides a single pane of glass for managing all clusters



Chapter Topics

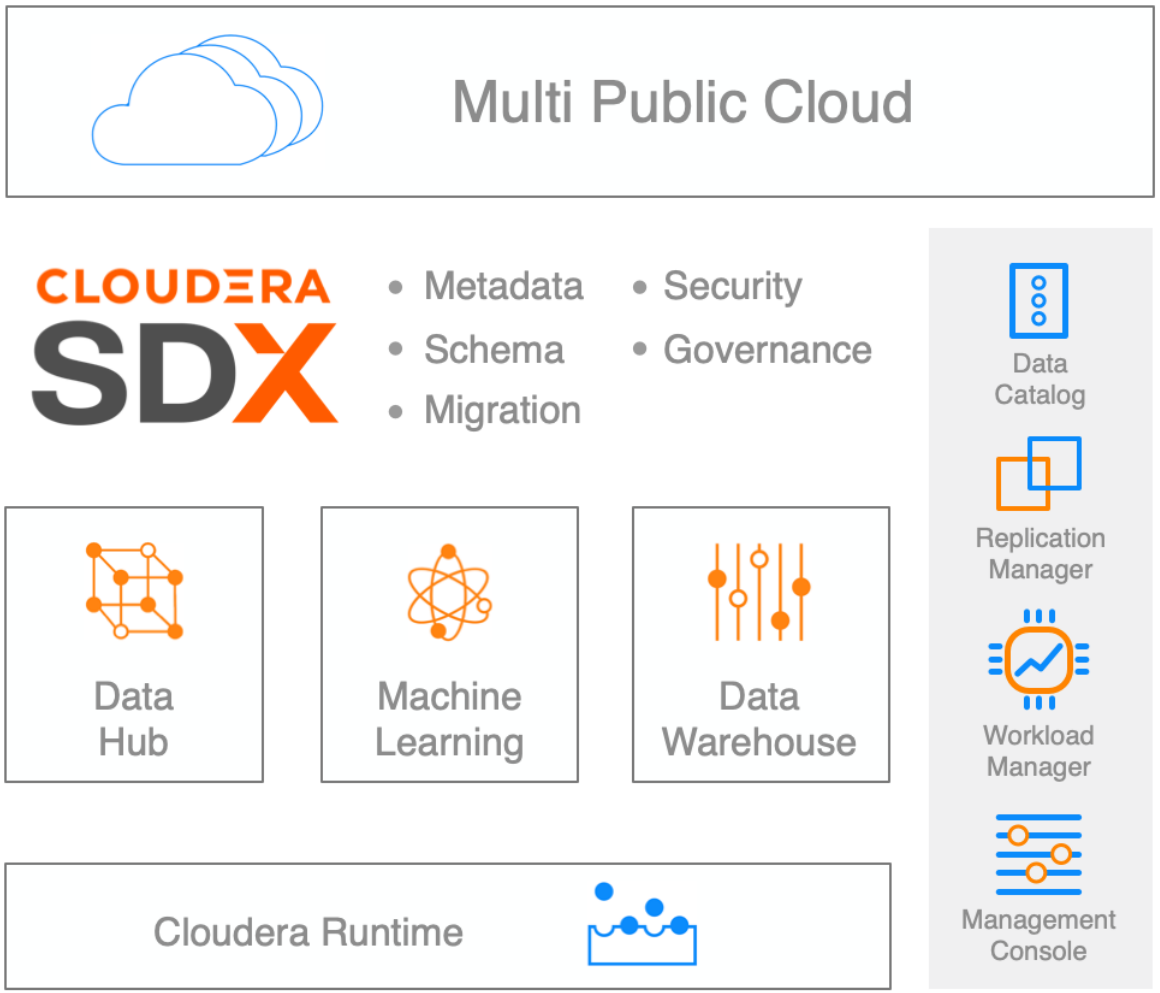
Cloudera Data Platform Overview

- Key Concepts
- Cloudera Runtime
- Core Services
- **Types of CDP Deployments**
- Custom Deployments with Data Hub
- Self-Service Experiences
- A Tour of the CDP User Interface
- Essential Points

One Platform for Three Types of Deployments

- **CDP will be available in different formats**
  - CDP Public Cloud
  - CDP Private Cloud
  - CDP Data Center

CDP Public Cloud



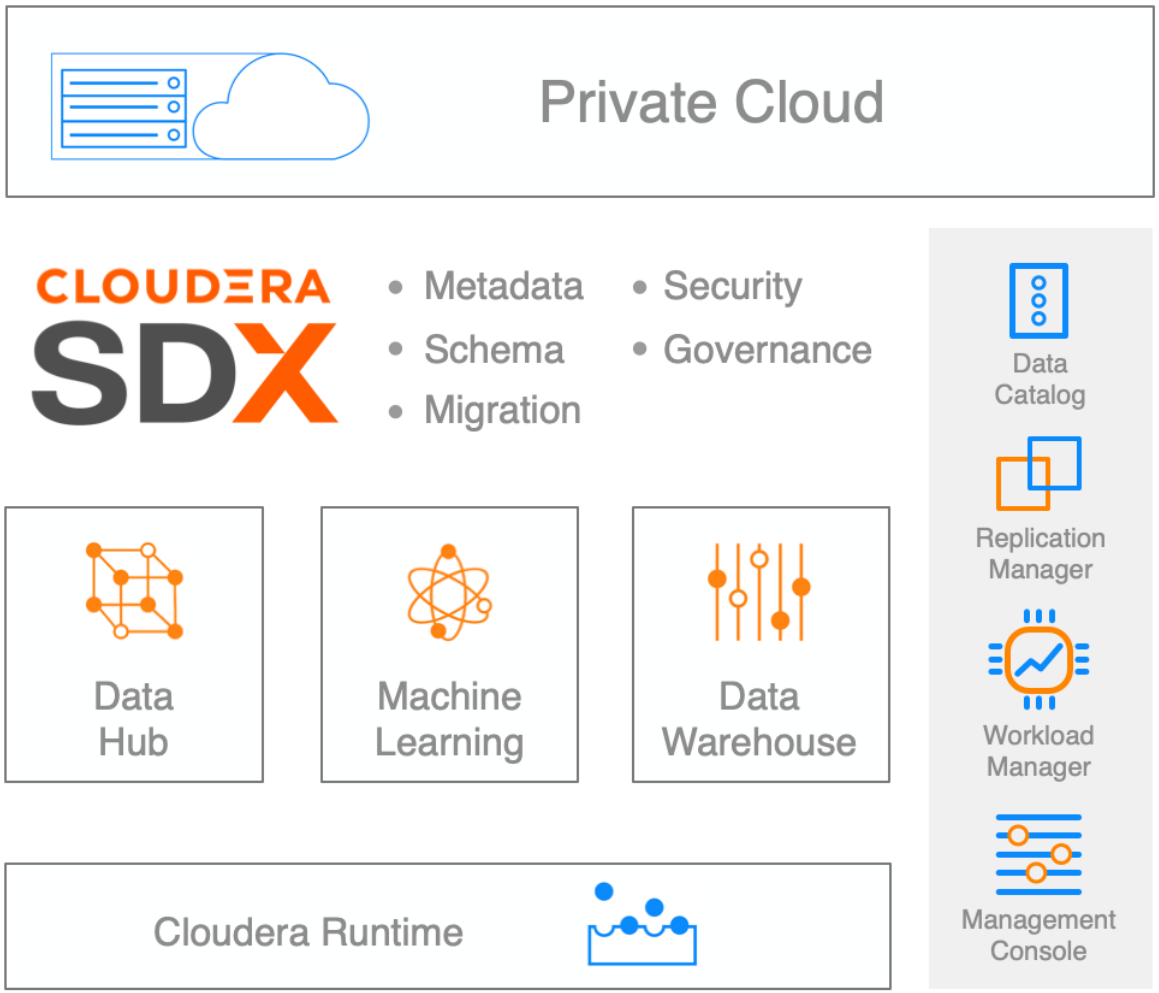
Comparing CDP Public Cloud to CDH and HDP

- **CDH and HDP favored "bare metal" installations**
  - CDP Public Cloud is a cloud service
- **CDH and HDP collocated storage and compute on each node**
  - CDP Public Cloud favors separating storage from compute
- **CDP Public Cloud is a great choice for intermittent or transient workloads**

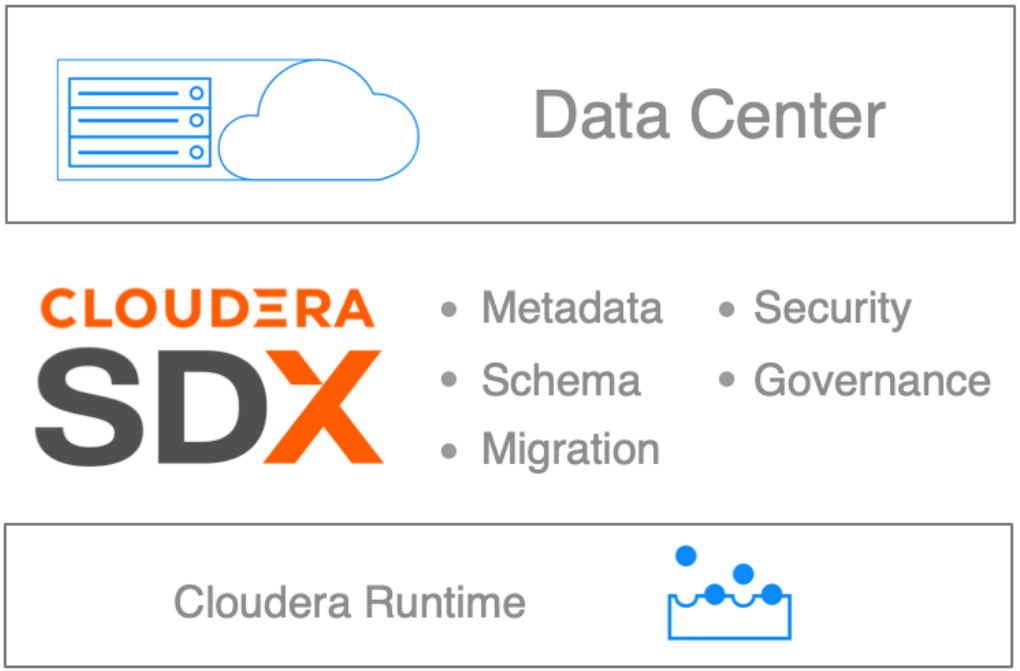
Environments

- **Defines where CDP will create and access cloud resources**
- **Registering an environment is a prerequisite for using CDP**
  - It's among the administrator's first tasks
  - Can register as many environments as needed

CDP Private Cloud



CDP Data Center



Chapter Topics

Cloudera Data Platform Overview

- Key Concepts
- Cloudera Runtime
- Core Services
- Types of CDP Deployments
- **Custom Deployments with Data Hub**
- Self-Service Experiences
- A Tour of the CDP User Interface
- Essential Points

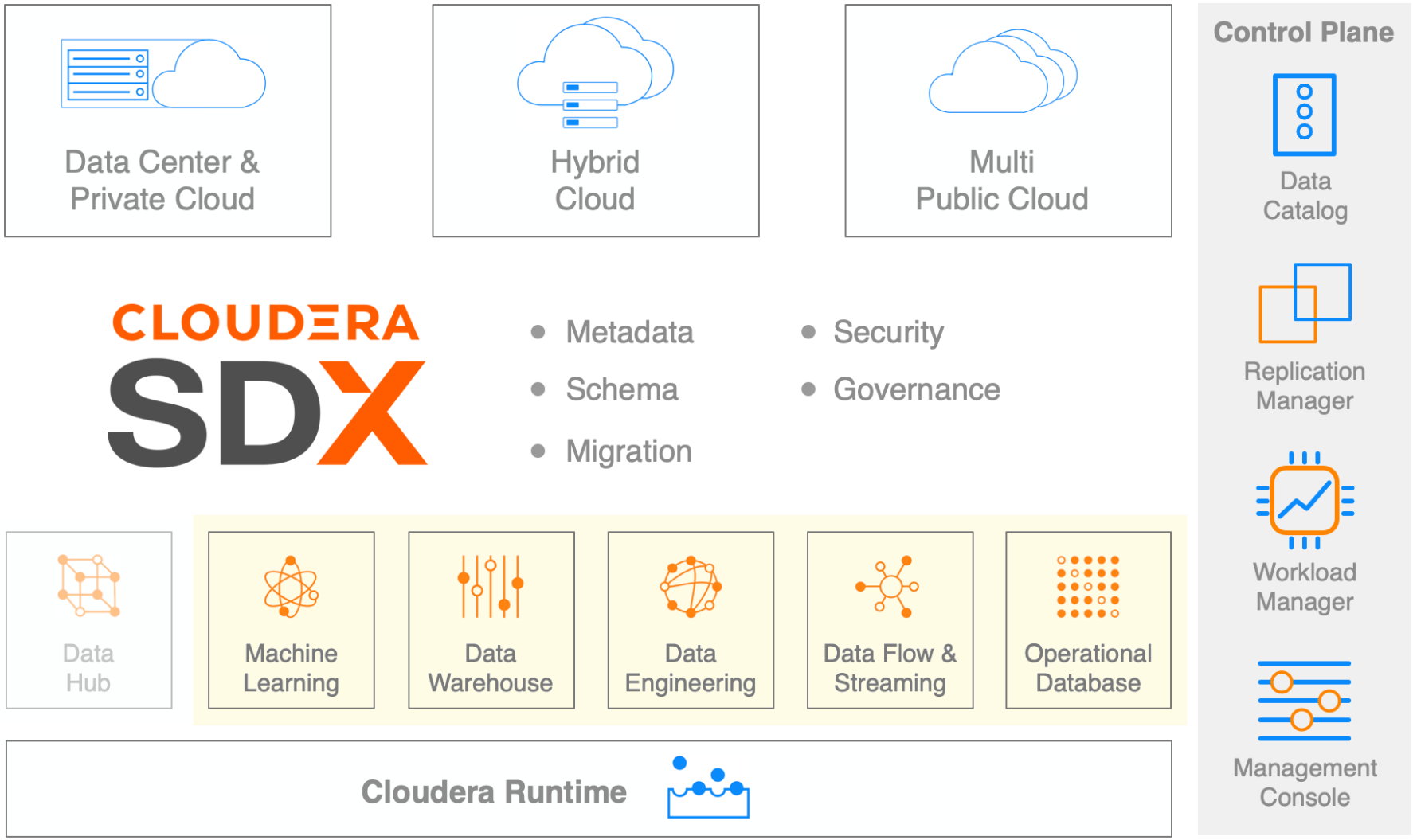
Data Hub

Chapter Topics

Cloudera Data Platform Overview

- Key Concepts
- Cloudera Runtime
- Core Services
- Types of CDP Deployments
- Custom Deployments with Data Hub
- **Self-Service Experiences**
- A Tour of the CDP User Interface
- Essential Points

Self-Service Experiences



Chapter Topics

Cloudera Data Platform Overview

- Key Concepts
- Cloudera Runtime
- Core Services
- Types of CDP Deployments
- Custom Deployments with Data Hub
- Self-Service Experiences
- **A Tour of the CDP User Interface**
- Essential Points

Tour of the User Interface

Your instructor will now take you on a tour of the CDP user interface

Chapter Topics

Cloudera Data Platform Overview

- Key Concepts
- Cloudera Runtime
- Core Services
- Types of CDP Deployments
- Custom Deployments with Data Hub
- Self-Service Experiences
- A Tour of the CDP User Interface
- **Essential Points**

Essential Points

- **CDP is an entirely new platform**
  - Released in multiple formats for different types of deployments
  - Cloudera Runtime is the universal foundation for CDP
  - SDX provides consistent and security and governance
- **CDP Public Cloud and CDP Private Cloud share a similar architecture**
  - Cloud-native: Built on containerization and virtualization technology
  - Both offer self-service experiences optimized for specific workloads
  - Favor cloud storage for data, and separation of storage and compute
- **CDP Data Center is the format most similar to CDH and HDP**
  - Designed for “bare metal” installation, directly on your servers

Bibliography

The following offer more information on topics discussed in this chapter

- **How Cloudera Data Platform Helps Data-Centric Enterprise IT**
  - http://tiny.cloudera.com/cdpec03a
- **Cloudera Runtime Documentation**
  - http://tiny.cloudera.com/cdpec03b
- **Cloudera Data Hub: Where Agility Meets Control**
  - http://tiny.cloudera.com/cdpec03c

CLOUDERA

Workload and Data Management

Chapter 4

Course Chapters

- Introduction
- Introducing the Enterprise Data Cloud
- Cloudera Data Platform Overview
- **Workload and Data Management**
- Data in Motion
- Data Warehousing and Analytics
- Data Science and Machine Learning
- Security and Data Governance
- Planning for Success
- Conclusion

Chapter Topics

Workload and Data Management

- **The Role of an Administrator**
- SDX
- Managing Resources and Costs
- Workload Isolation
- Data Migration and Replication
- Essential Points

The Administrator's Role in Operations

- **Key duties of an administrator**
  - Cluster installation
  - Cluster monitoring
  - Cluster management
- **Duties may also include managing things that clusters require**
  - Hardware
  - Network
  - Security
- **Typically also responsible for backup and disaster recovery**

The Administrator's Role in Planning

- **Often involved in planning deployments**
  - Cluster sizing
  - Designing for scalability and performance
- **Must understand both technical and business requirements**
  - Necessary to identify risks and opportunities

Tips for Finding Qualified Administrators

- **Look for experience with key skills**
  - Linux system administration
  - Performance tuning
  - Networking
  - Security
  - Configuration management and scripting
- **Consider relevant training and certifications when screening candidates**
- **Cloud experience is now an essential skill**

Chapter Topics

Workload and Data Management

- The Role of an Administrator
- **SDX**
- Managing Resources and Costs
- Workload Isolation
- Data Migration and Replication
- Essential Points

What is SDX?

- **Fundamental part of CDP**
- **Delivers consistent data security and governance**
- **Centralized policy enforcement across all environments**

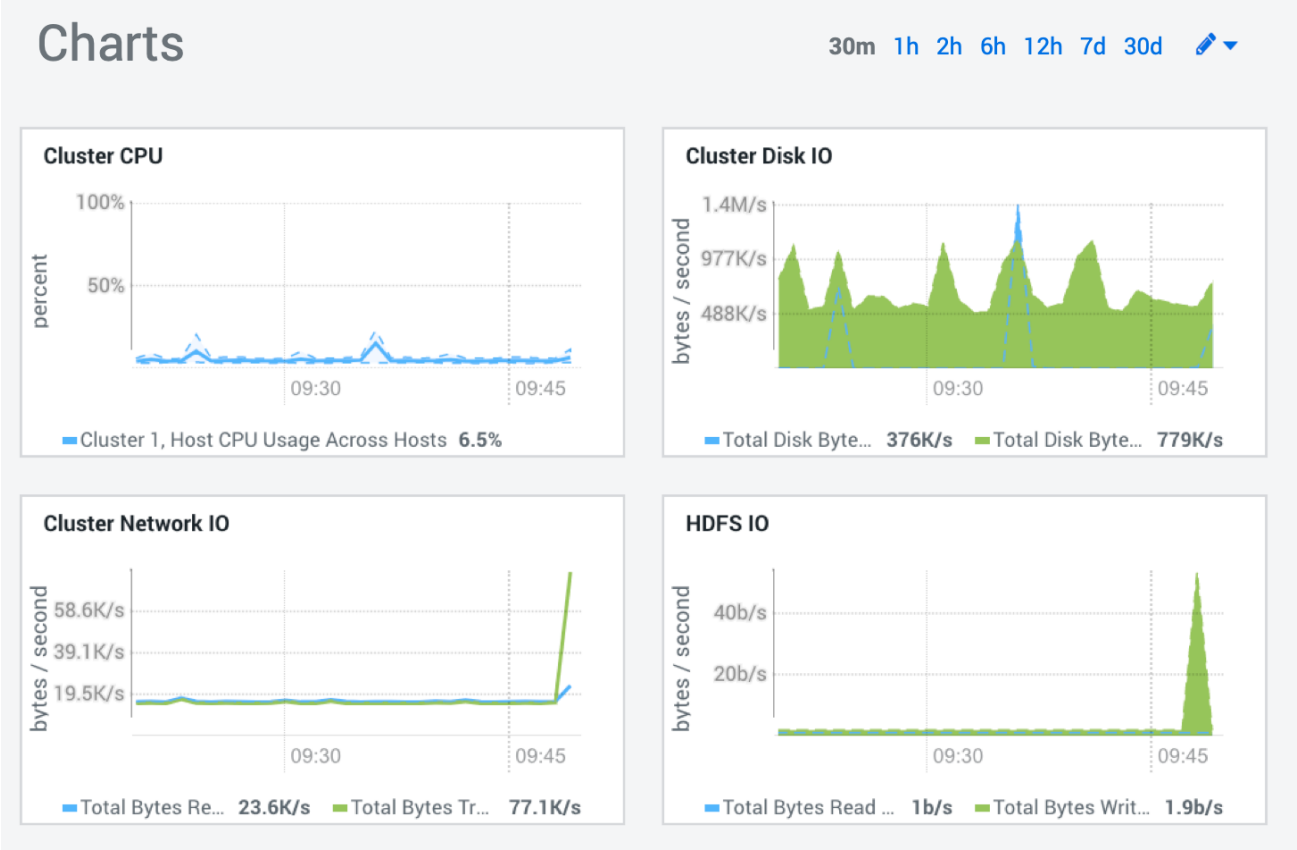
Chapter Topics

Workload and Data Management

- The Role of an Administrator
- SDX
- **Managing Resources and Costs**
- Workload Isolation
- Data Migration and Replication
- Essential Points

Cloudera Manager

- **Cloudera Manager offers helpful charts, metrics, and alerts**
  - Primarily focused on nodes and services
  - Provides information about resource usage



Workload Manager

- **Gives administrators insight into *workloads***
  - Summarizes history
  - Identifies performance problems
  - Provides prescriptive guidance
- **Helps administrators understand how resource usage affect workloads**
  - Enables them to see opportunities for cost savings

Chapter Topics

Workload and Data Management

- The Role of an Administrator
- SDX
- Managing Resources and Costs
- **Workload Isolation**
- Data Migration and Replication
- Essential Points

Why Separate Storage from Compute?

- **Clusters traditionally collocated storage and compute**
  - Fine for single-tenant clusters with predictable workloads
- **Can lead to the *noisy neighbor* problem on multi-tenant clusters**
  - As utilization increases, performance decreases
  - Adding more nodes increases capacity, not isolation
  - Splitting the cluster adds isolation, but complicates administration
- **SDX makes it possible to have shared data across multiple compute clusters**

Chapter Topics

Workload and Data Management

- The Role of an Administrator
- SDX
- Managing Resources and Costs
- Workload Isolation
- **Data Migration and Replication**
- Essential Points

Replication Manager

- **Management tool for copying data between a source and destination**
- **Replication manager supports many use cases**
  - Backup and disaster recovery
  - Migrating data from on-premises to the cloud
  - Copying data for development and testing

Chapter Topics

Workload and Data Management

- The Role of an Administrator
- SDX
- Managing Resources and Costs
- Workload Isolation
- Data Migration and Replication
- **Essential Points**

Essential Points

- **Administrators are responsible for ensuring reliable access to systems**
  - Typically includes cluster planning, monitoring, and management
  - Specific duties vary by organization and deployment strategy
- **SDX delivers consistent data security and governance across CDP**
- **CDP provides tools that help administrators monitor and manage performance**
  - Cloudera Manager mainly focuses on health of individual nodes and services
  - Workload Manager mainly focuses on application performance
- **CDP favors separation of storage and compute**
  - Provides the isolation needed to avoid the “noisy neighbor” problem
- **Replication Manager is used for backup, recovery, and data migration**

Bibliography

The following offer more information on topics discussed in this chapter

- **Solving the Pain Points of Big Data Management**
  - http://tiny.cloudera.com/cdpec04a
- **Improving Multi-Tenancy with Virtual Private Clusters**
  - http://tiny.cloudera.com/cdpec04b
- **Cloudera Training Courses for Administrators**
  - http://tiny.cloudera.com/cdpec04c
- **Cloudera Training Courses for Administrators**
  - http://tiny.cloudera.com/cdpec04d

CLOUDERA

Data in Motion

Chapter 5

Course Chapters

Chapter Topics

Data in Motion

- **The Role of a Data Engineer**
- Data in Motion Use Cases
- Cloudera DataFlow
- Essential Points

The Role of a Data Engineer

- **Essential part of a data team**
  - Rely on systems provided by administrators
  - Their output enables data analysts and data scientists
- **Design, create, and maintain data pipelines**
  - Increasingly relies on data gathered from streaming sources
- **Build and integrate applications running on the platform**
- **Diagnose and solve problems with application performance**

Tips for Finding Qualified Data Engineers

- **Software engineering**
  - Programming languages (especially Python, Java, and Scala)
  - Distributed computing concepts
  - Data storage
- **Relational databases**
  - Experience with SQL
  - Schema design
- **Basic knowledge of system administration is helpful**

Chapter Topics

Data in Motion

- The Role of a Data Engineer
- **Data in Motion Use Cases**
- Cloudera DataFlow
- Essential Points

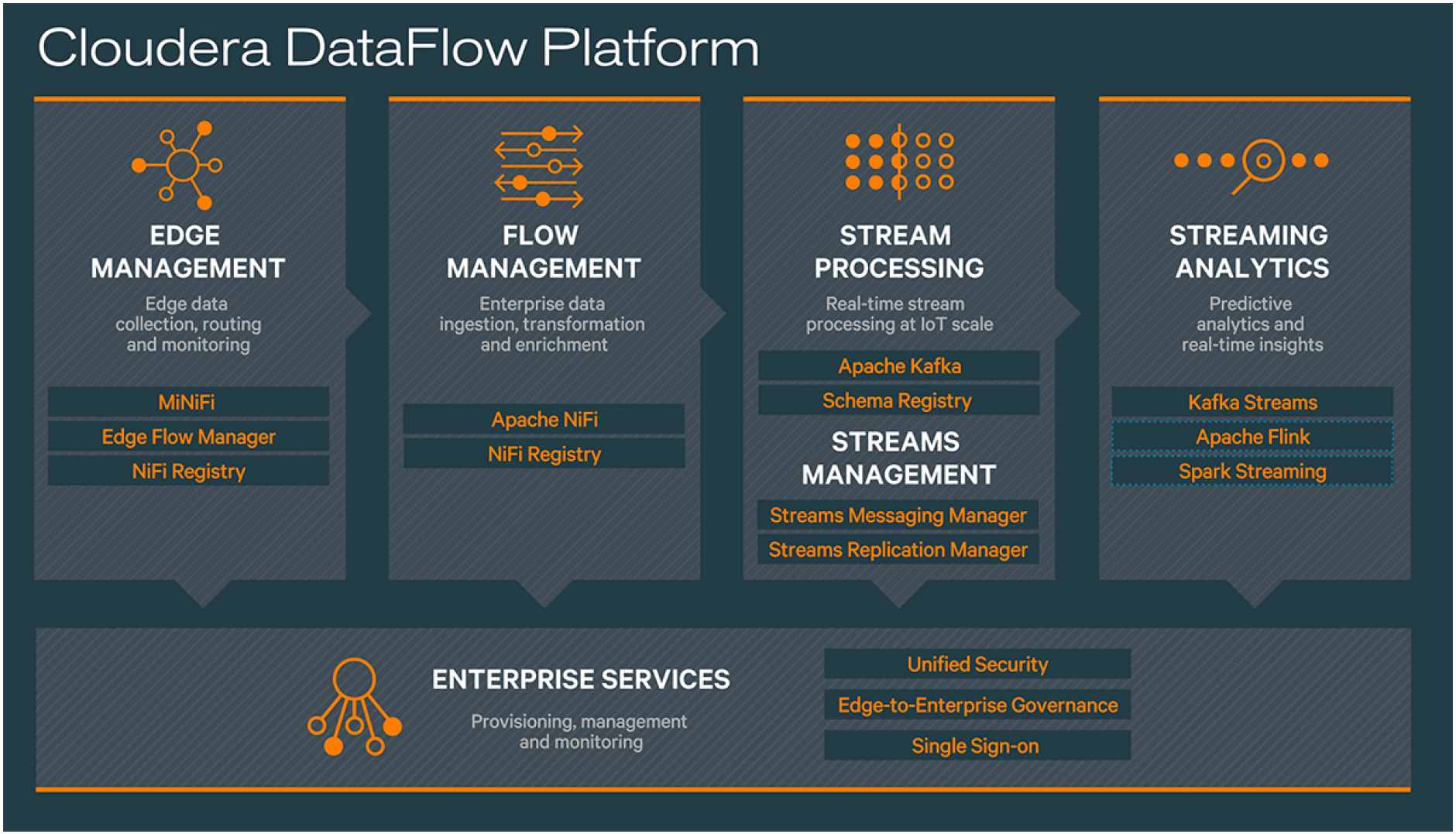
There are no slides in this section of the course.
Please refer to the video in OnDemand.

Chapter Topics

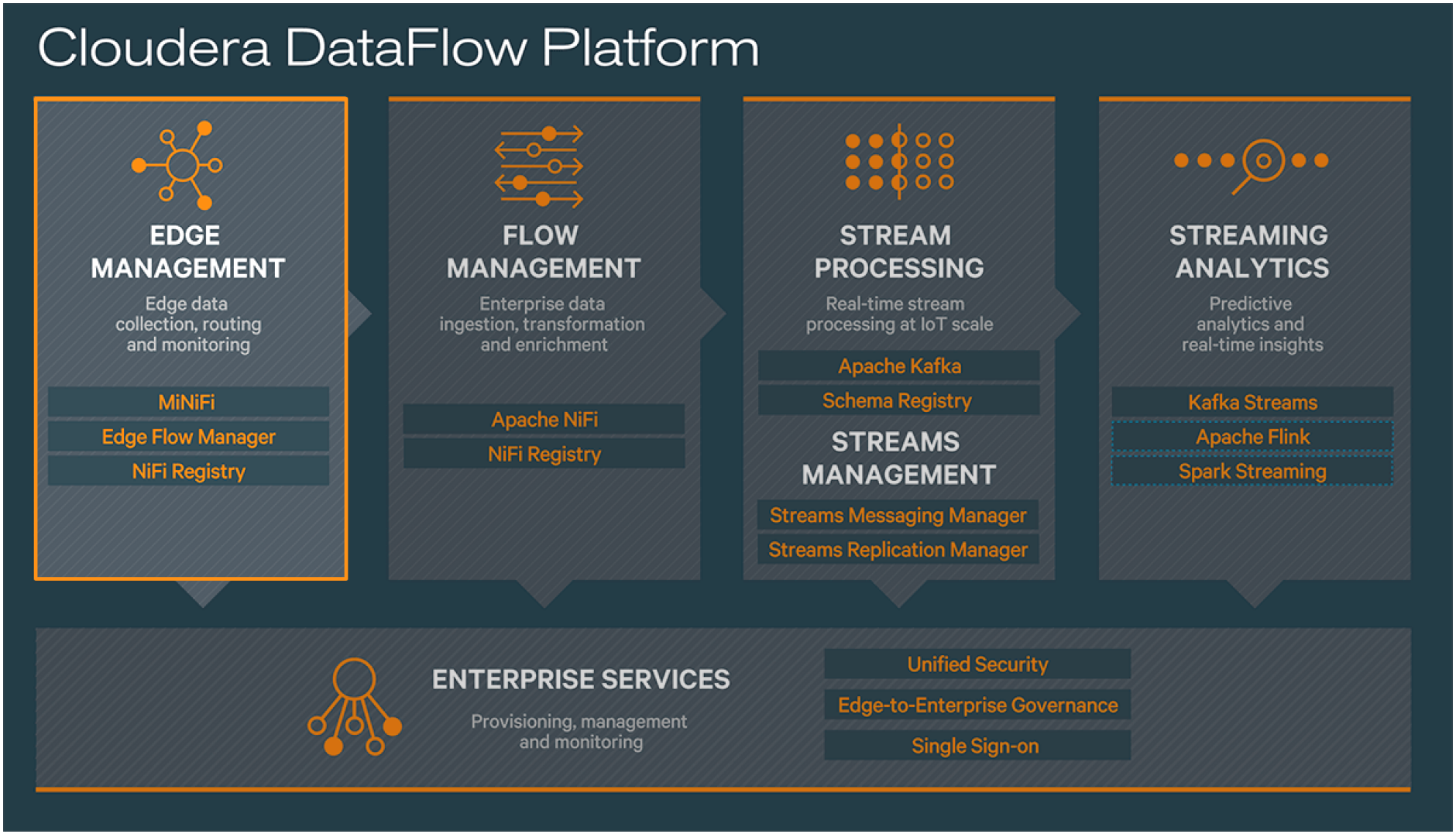
Data in Motion

- The Role of a Data Engineer
- Data in Motion Use Cases
- **Cloudera DataFlow**
- Essential Points

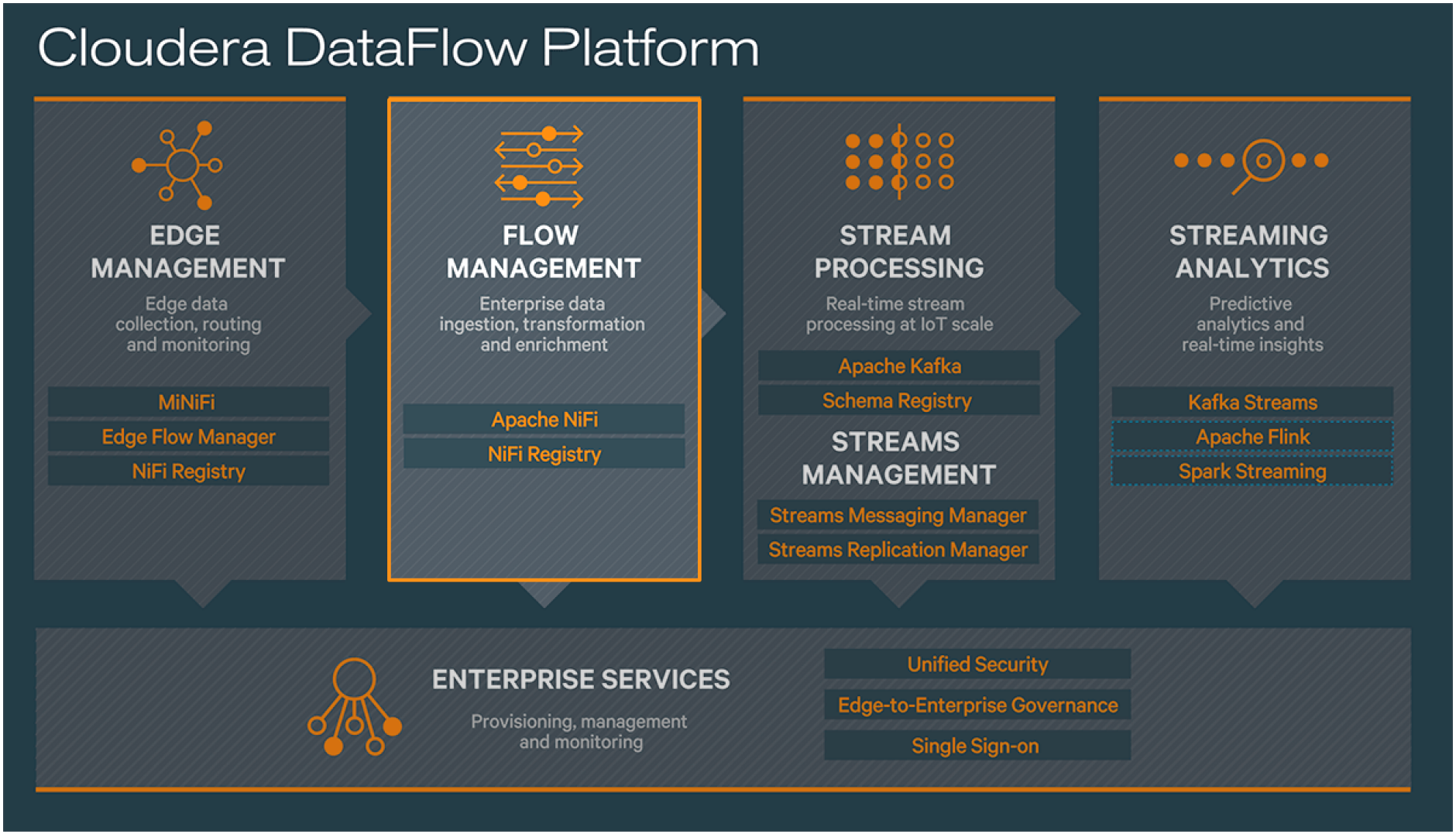
Cloudera DataFlow Platform



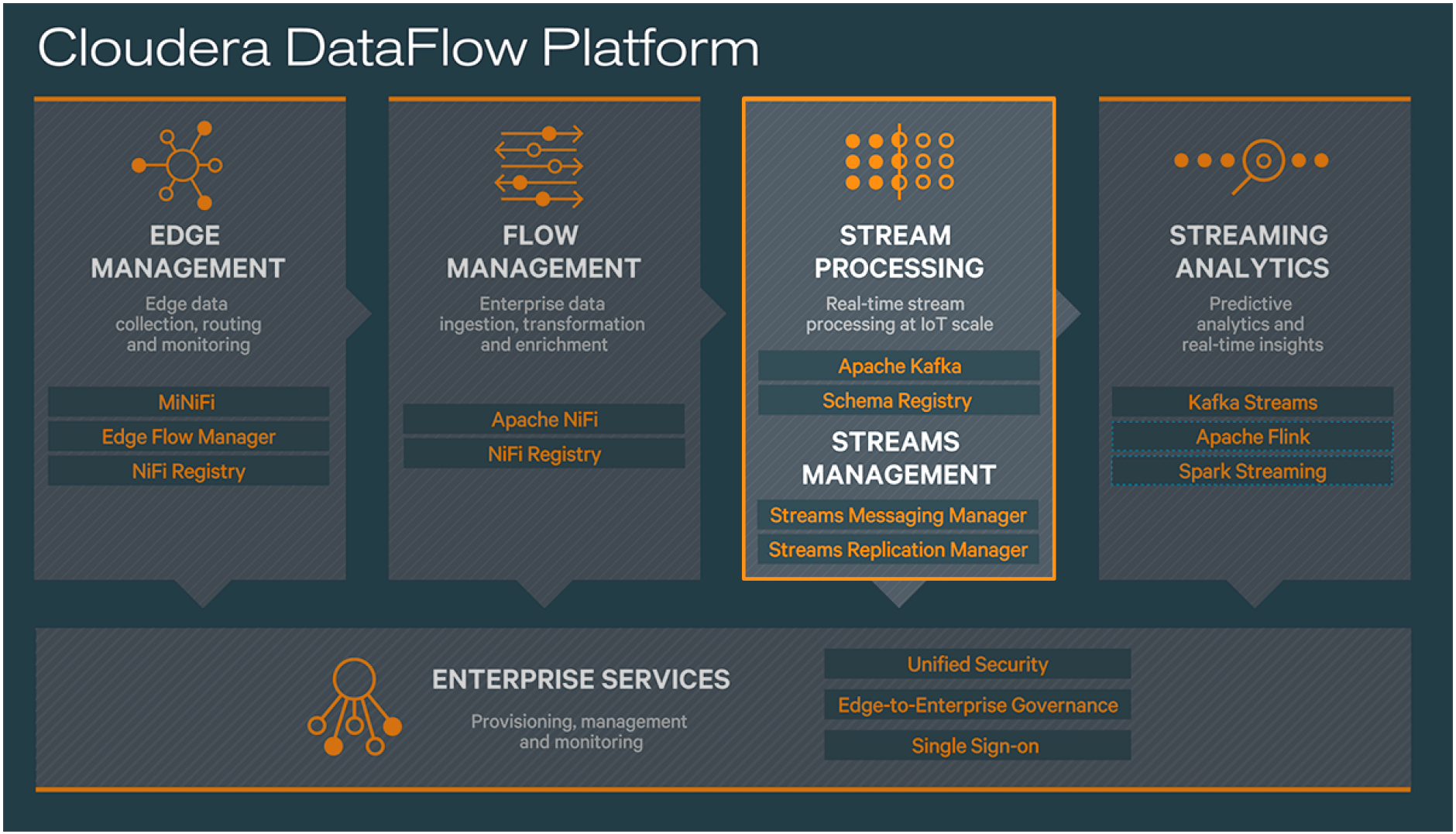
Cloudera Edge Management



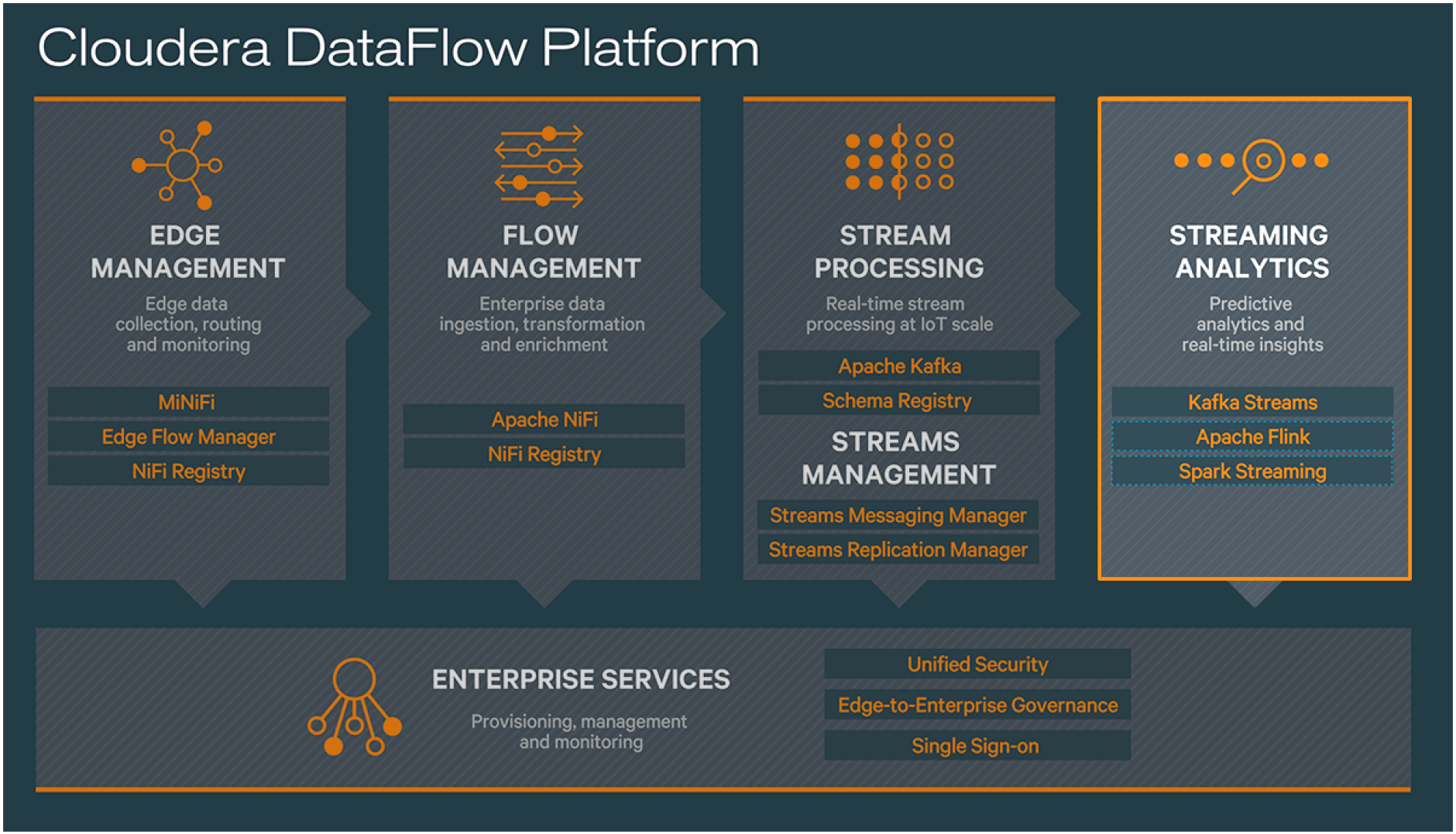
Cloudera Flow Management



Cloudera Stream Processing and Streams Management



Cloudera Streaming Analytics



Chapter Topics

Data in Motion

- The Role of a Data Engineer
- Data in Motion Use Cases
- Cloudera DataFlow
- **Essential Points**

Essential Points

- **Data engineers design, create, and maintain data pipelines**
  - Data analysts, data scientists, and others depend on data from these pipelines
  - These pipelines increasingly involve data gathered from streaming sources
- **Cloudera DataFlow Platform provides several capabilities to work with data**
  - Edge Management: Collect, route, and monitor data from IoT devices
  - Flow Management: Ingest, transform, and enrich data from nearly any source
  - Stream Processing: Scalable messaging for streaming data use cases
  - Streams Management: Monitor, manage, and replicate stream processing systems
  - Streaming Analytics: Enable predictive analytics on streaming data

Bibliography

The following offer more information on topics discussed in this chapter

- **Evolution of Data Management: The Role of Streaming Data and IoT Data Architecture**
  - http://tiny.cloudera.com/cdpec05a
- **Adding NiFi and Kafka to Cloudera Data Platform**
  - http://tiny.cloudera.com/cdpec05b
- **A 5D Model to Assess Your IoT Readiness**
  - http://tiny.cloudera.com/cdpec05c
- **Cloudera Training Courses for Data Engineers and Developers**
  - http://tiny.cloudera.com/cdpec05d

CLOUDERA

Data Warehousing and Analytics

Chapter 6

Course Chapters

- Introduction
- Introducing the Enterprise Data Cloud
- Cloudera Data Platform Overview
- Workload and Data Management
- Data in Motion
- **Data Warehousing and Analytics**
- Data Science and Machine Learning
- Security and Data Governance
- Planning for Success
- Conclusion

Chapter Topics

Data Warehousing and Analytics

- **The Role of a Data Analyst**
- Data Warehouse and Analytics Use Cases
- CDP Data Warehouse Experience
- Operational and Analytic Database Capabilities
- Essential Points

The Role of a Data Analyst

- Explore and query data
- Interpret and communicate results
- Produce reports and visualizations
- Identify interesting patterns for further exploration

Tips for Finding Qualified Data Analysts

- **Background in business, mathematics, or economics**
- **Strong communication skills**
- **Analytic mindset**
- **Experience with tools for querying data**

Chapter Topics

Data Warehousing and Analytics

- The Role of a Data Analyst
- **Data Warehouse and Analytics Use Cases**
- CDP Data Warehouse Experience
- Operational and Analytic Database Capabilities
- Essential Points

There are no slides in this section of the course.
Please refer to the video in OnDemand.

Chapter Topics

Data Warehousing and Analytics

- The Role of a Data Analyst
- Data Warehouse and Analytics Use Cases
- **CDP Data Warehouse Experience**
- Operational and Analytic Database Capabilities
- Essential Points

Cloudera Data Warehouse

- **Cloudera Data Warehouse provides self-service experience**
  - Optimized for data warehouse and data mart workloads
  - Available in CDP Public Cloud and CDP Private Cloud
- **Virtual warehouses maintain high performance through isolation**
  - SDX allows safe access to shared data with consistent security and governance
  - Manages cost and performance through auto-scaling and auto-suspend features

Chapter Topics

Data Warehousing and Analytics

- The Role of a Data Analyst
- Data Warehouse and Analytics Use Cases
- CDP Data Warehouse Experience
- **Operational and Analytic Database Capabilities**
- Essential Points

Flexibility for Data Storage and Processing

- **Traditional data warehouses tightly couple data storage and processing**
  - Once loaded, data is typically stored in a proprietary format
  - Data is not directly accessible to other applications or workload types
- **CDP is an open system with broad data compatibility**
  - Data storage is decoupled from data processing
  - This is a major advantage over traditional data warehouse systems
- **Query engines evaluate schemas to understand structure of data in tables**
  - Applications can have customized views of the same underlying data
  - Data is directly available for other workload types, such as machine learning

Operational Database

- **Cloudera's platform has strong support for data warehousing**
  - However, it also supports other tools for working with data
- **Apache HBase is a high-performance distributed data store**
  - Available in all deployments of CDP as well as CDH and HDP
  - Low-latency access for reading and writing data
  - Its tables can have millions of columns and billions of rows
- **HBase is a key component in Cloudera's Operational Database solution**
  - Used to store and serve data from applications in real time
  - Use cases often relate to event monitoring, detection, and prediction

Cloudera Search

- **Significant growth in the volume of unstructured data**
  - Much of this is freeform text, which doesn't fit the relational model
- **Such text is difficult to query with SQL-based tools**
  - These tools are optimized for exact matches on structured data
  - They have limited ability to understand complexities of natural language
- **Cloudera Search is designed for full-text interactive search**
  - Leverages Apache Solr, which powers many of the world's leading websites
  - Can parse, index, and search text in more than 30 human languages
  - Ideal for data discovery and text analytics

Chapter Topics

Data Warehousing and Analytics

- The Role of a Data Analyst
- Data Warehouse and Analytics Use Cases
- CDP Data Warehouse Experience
- Operational and Analytic Database Capabilities
- **Essential Points**

Essential Points

- **Data analysts query the data warehouse to produce business-oriented reports**
  - This relies on the work of system administrators and data engineers
- **Cloudera Data Warehouse is a self-service experience**
  - Available in CDP Public Cloud and CDP Private Cloud
  - Optimized for data warehouse and data mart workloads
- **CDP makes data processing and analysis more flexible**
  - You can use many different tools and techniques to work with your data
- **Cloudera Search provides a full-text interactive search capability for your data**

Bibliography

The following offer more information on topics discussed in this chapter

- **Cloudera Data Warehouse**
  - http://tiny.cloudera.com/cdpec06a
- **Cloudera Training Courses for Data Analysts**
  - http://tiny.cloudera.com/cdpec06b
- **Extend Data Warehousing to Hybrid and Multi-Cloud with Cloudera Data Platform**
  - http://tiny.cloudera.com/cdpec06c

CLOUDERA

Data Science and Machine Learning

Chapter 7

Course Chapters

Chapter Topics

Data Science and Machine Learning

- **The Role of a Data Scientist**
- Machine Learning Use Cases
- Cloudera Data Science Workbench (CDSW)
- Cloudera Machine Learning
- Cloudera Fast Forward Labs
- Essential Points

The Role of a Data Scientist

- **Data scientists extract value from data**
  - Often involves building and training machine learning models
- **Designing and conducting experiments is an important part of data science**
- **Data scientists also provide value by creating data products**
- **Our products, such as CDSW and Cloudera Machine Learning, aid their work**

Tips for Finding Qualified Data Scientists

- **Data scientists bring a variety of skills to the role**
  - Strong background with statistics
  - Experience with software engineering
  - Ability to design and conduct experiments
  - Domain-specific knowledge
  - Effective communication skills
- **Some of these skills may overlap with those required for other roles**
  - While skills may overlap, the data scientist's *role* is unique

Chapter Topics

Data Science and Machine Learning

- The Role of a Data Scientist
- **Machine Learning Use Cases**
- Cloudera Data Science Workbench (CDSW)
- Cloudera Machine Learning
- Cloudera Fast Forward Labs
- Essential Points

There are no slides in this section of the course.
Please refer to the video in OnDemand.

Chapter Topics

Data Science and Machine Learning

- The Role of a Data Scientist
- Machine Learning Use Cases
- **Cloudera Data Science Workbench (CDSW)**
- Cloudera Machine Learning
- Cloudera Fast Forward Labs
- Essential Points

Cloudera Data Science Workbench

- **CDSW enables collaborative data science at scale**
- **Provides data science teams with web-based access to a secure cluster**
  - End users don't need to install software
  - Multi-tenant system provides efficient use of resources
  - Isolation maintains good performance for user sessions

Chapter Topics

Data Science and Machine Learning

- The Role of a Data Scientist
- Machine Learning Use Cases
- Cloudera Data Science Workbench (CDSW)
- **Cloudera Machine Learning**
- Cloudera Fast Forward Labs
- Essential Points

Cloudera Machine Learning

- **Provides data scientists with self-service access to governed data**
- **Fosters collaboration among data scientists**
- **Delivers a consistent experience**

Chapter Topics

Data Science and Machine Learning

- The Role of a Data Scientist
- Machine Learning Use Cases
- Cloudera Data Science Workbench (CDSW)
- Cloudera Machine Learning
- **Cloudera Fast Forward Labs**
- Essential Points

Cloudera Fast Forward Labs

- **Cloudera Fast Forward Labs is an advisory and research service**
  - Founded by noted industry expert Hilary Mason
  - Acquired by Cloudera in 2017
- **Applies emerging machine learning techniques to business problems**
  - Helps customers bridge the gap between research and industry
  - Offers advisory services to help customers with strategy
  - Publishes research reports and prototypes

Chapter Topics

Data Science and Machine Learning

- The Role of a Data Scientist
- Machine Learning Use Cases
- Cloudera Data Science Workbench (CDSW)
- Cloudera Machine Learning
- Cloudera Fast Forward Labs
- **Essential Points**

Essential Points

- **Data scientists combine a scientific approach with skills in software engineering and statistics to extract value from data**
  - They share many skills with data engineers and data analysts, but have a different focus
  - One way they provide value is by creating data products
- **Data scientists tend to be heavily involved with machine learning**
- **Cloudera Data Science Workbench (CDSW) and Cloudera Machine Learning support their work**
  - CDSW enables collaborative data science for on-premises CDH and HDP clusters
  - Cloudera Machine Learning provides a similar experience for CDP in the cloud

Bibliography

The following offer more information on topics discussed in this chapter

- **Cloudera Data Science Workbench**
  - http://tiny.cloudera.com/cdpec07a
- **Cloudera Machine Learning**
  - http://tiny.cloudera.com/cdpec07b
- **How the Rise of Data and AI Have Redefined the Data-Driven Enterprise**
  - http://tiny.cloudera.com/cdpec07c
- **Cloudera Fast Forward Labs**
  - http://tiny.cloudera.com/cdpec07d
- **Cloudera Training Courses for Data Scientists**
  - http://tiny.cloudera.com/cdpec07e

CLOUDERA

Security and Data Governance

Chapter 8

Course Chapters

Chapter Topics

Security and Data Governance

- **The Role of a Data Steward**
- Data Catalog
- Controlling and Auditing Data Access
- Essential Points

The Role of a Data Steward

- **Data stewards look after an organization's data**
  - This is a business-oriented role that defines data management policies
  - They curate datasets and ensure they're used effectively
  - They work with Information Security to ensure compliance
- **This is a cross-functional role**
  - Must ensure effective use of data across departments
  - Requires tools that provide visibility into how data is modified and accessed

Tips for Finding Qualified Data Stewards

- **Candidate should have a background in business**
- **Look for extensive data management experience**
  - Particularly experience in your industry
  - Understanding of relevant formats, data quality issues, and regulations
- **Solid understanding of SQL and key security concepts**
- **Excellent communication skills**

Chapter Topics

Security and Data Governance

- The Role of a Data Steward
- **Data Catalog**
- Controlling and Auditing Data Access
- Essential Points

CDP Data Catalog

- **Helps data stewards understand, organize, manage, and govern data assets**
  - A Hive table used in a data warehouse query is an example of a data asset
- **Data Catalog dashboard page lists all assets associated with a data lake**
  - Quickly find an asset by filtering on name, tags, date, or other criteria
- **Clicking an asset on the dashboard opens its detail page**
  - Summarizes the asset type, owner, creation date, modification date, and other properties
  - Provides access to view the asset's schema, policy, and audit information
- **Data assets can be organized into collections**
  - Typically done based on age, ownership, or content of metadata tags

Chapter Topics

Security and Data Governance

- The Role of a Data Steward
- Data Catalog
- **Controlling and Auditing Data Access**
- Essential Points

There are no slides in this section of the course.
Please refer to the video in OnDemand.

Chapter Topics

Security and Data Governance

- The Role of a Data Steward
- Data Catalog
- Controlling and Auditing Data Access
- **Essential Points**

Essential Points

- **Data stewards look after an organization's data**
  - This cross-functional role is heavily involved with data governance
- **Data catalogs in CDP help data stewards manage and govern data assets**
  - Enables them to quickly locate and understand these assets
  - Provides easy access to schema, policy, and audit information
- **Apache Ranger provides the ability to define fine-grained access control policies for data**
- **Apache Atlas provides support for metadata management, data classification, and lineage**

Bibliography

The following offer more information on topics discussed in this chapter

- **Data Governance and Security: How to Put Data at the Heart of Your Business Strategy**
  - http://tiny.cloudera.com/cdpec08a
- **Governing for Digital Transformation and Growth**
  - http://tiny.cloudera.com/cdpec08b

CLOUDERA

Planning for Success

Chapter 9

Course Chapters

- Introduction
- Introducing the Enterprise Data Cloud
- Cloudera Data Platform Overview
- Workload and Data Management
- Data in Motion
- Data Warehousing and Analytics
- Data Science and Machine Learning
- Security and Data Governance
- **Planning for Success**
- Conclusion

Chapter Topics

Planning for Success

- **Challenges and Opportunities**
- How Cloudera Can Help
- Essential Points

Benefits of Cloud Migrations

- Provides elasticity for compute and storage
- Reduces capital expense
- Lowers administration costs
- Consumption-based pricing is attractive for certain workloads
- Provides geographic diversity of infrastructure

Challenges with Cloud Migrations

- **Cost**
- **Vendor lock-in**
- **Security**
- **Data governance**
- **Performance**

Chapter Topics

Planning for Success

- Challenges and Opportunities
- **How Cloudera Can Help**
- Essential Points

Recommendations for Success

- **Set realistic goals from the start**
- **Develop a strategy to reaching your goals**
- **Engage with Cloudera for help throughout your journey**
  - Connect with our industry experts to discuss your use cases
  - Enroll in Cloudera training to learn how to use the platform effectively
  - Consult with Professional Services for help with design and implementation
  - Partner with the Cloudera Support team to resolve technical issues

Staffing Your Project

- System administrators
- Data engineers
- Data analysts
- Data scientists
- Data stewards

Cloudera Training

- **Instructor-led**
  - Classroom-based
  - Virtual
- **OnDemand**
- **Blended learning**
- **Customized training plans**
- **Certification**

Cloudera Professional Services

- **Professional Services can provide a wide range of assistance**
  - Architecture
  - Design
  - Deployment
  - Security
  - Operations
  - Production readiness assessment
- **Resident staff are available to augment your team**

Cloudera Support

- **Experienced global support team**
- **Predictive and proactive**
- **Innovative tools that enhance the support experience**
- **Specialized services available from Cloudera Government Support**

Chapter Topics

Planning for Success

- Challenges and Opportunities
- How Cloudera Can Help
- **Essential Points**

Essential Points

- **Migrating to the cloud can reduce capital and administrative expenses**
  - Can also reduce operational expense for certain types of workloads
- **Cloud migrations can also present challenges with costs, security, and performance**
  - CDP provides many features that help customers address these challenges
- **Cloudera provides many services to help customers learn and use the platform successfully**
  - Industry experts can offer guidance on which use cases to pursue
  - We offer training to help your staff gain the skills they need
  - Professional Services can help with design, implementation, and even staff augmentation
  - Our global support team can help you quickly resolve technical issues

Bibliography

The following offer more information on topics discussed in this chapter

- **Harvard Business Review: Critical Success Factors in a Multi-Cloud Environment**
  - http://tiny.cloudera.com/cdpec09a
- **Presentation by Cloudera's Tristan Stevens: What Does Success Look Like?**
  - http://tiny.cloudera.com/cdpec09b
- **Cloudera Educational Services (Training)**
  - http://tiny.cloudera.com/cdpec09c
- **Cloudera Professional Services (Consulting)**
  - http://tiny.cloudera.com/cdpec09d
- **Cloudera Support**
  - http://tiny.cloudera.com/cdpec09e
- **Why Your Public Sector Project Needs Cloudera Secure Support**
  - http://tiny.cloudera.com/cdpec09f

CLOUDERA

Conclusion

Chapter 10

Course Chapters

- Introduction
- Introducing the Enterprise Data Cloud
- Cloudera Data Platform Overview
- Workload and Data Management
- Data in Motion
- Data Warehousing and Analytics
- Data Science and Machine Learning
- Security and Data Governance
- Planning for Success
- **Conclusion**

Course Objectives (1)

During this course, you have learned

- Which characteristics define the Enterprise Data Cloud
- What Cloudera Data Platform is and what capabilities it provides
- How the Cloudera Data Platform supports both on-premises and cloud-based deployments
- How organizations use streaming data and the Internet of Things (IoT) to improve efficiency
- How companies are using Cloudera data warehouse tools to better understand their business

Course Objectives (2)

- How data scientists use Cloudera's products and services to help organizations benefit from machine learning
- How Cloudera helps organizations meet requirements for data security and governance
- What roles and skills organizations should look for when building data teams
- What resources are available to assist with planning, implementing, and supporting a Cloudera solution
- What factors should one consider before moving to the cloud

Which Course to Take Next

- **For developers**
  - *Developer Training for Spark and Hadoop*
  - *Cloudera Search Training*
  - *Cloudera Training for Apache HBase*
- **For system administrators**
  - *Cloudera Administrator Training for Apache Hadoop*
  - *Cloudera Security Training*
- **For data analysts and data scientists**
  - *Cloudera Data Analyst Training*
  - *Cloudera Data Scientist Training*