Data Management with Cloudera Data Platform on Dell Infrastructure

White Paper

Abstract

This white paper provides overview information for the Dell Technologies Validated Design for Data Management with Cloudera Data Platform (CDP) Private Cloud Base, for deployment on Dell PowerEdge servers, PowerSwitch networking, and PowerScale storage.

Dell Technologies Solutions

Dell Technologies
Validated Design

Part Number: H18932.1
March 2022

DELL Technologies

Copyright

Contents

Revision history

Table 1. Document revision history

Date	Document revision	Description of changes
March 2022	1.1	Included references to AMD and Intel versions of the *Data Management with Cloudera Data Platform on Dell Infrastructure Design Guide*. Updated "CDP Private Cloud components".
October 2021	1.0	Initial release. This document contains material from the prior reference architecture document, entitled *ClouderaCDP Private Cloud Base Reference Architecture Guide*.

Introduction

CDP Private Cloud Base, previously known as CDP Data Center, is the on-premises version of Cloudera Data Platform. This product combines the best of the prior Cloudera's and Hortonworks' technologies, along with new features and enhancements. CDP Private Cloud Base is also used with CDP Private Cloud Data Services to form CDP Private Cloud.

Executive summary

CDP Private Cloud delivers powerful analytic, transactional, and machine learning workloads in a hybrid data platform. It combines the agility and flexibility of a public cloud with the control of the data center. With elastic analytics and scalable object storage, CDP Private Cloud modernizes traditional single-cluster deployments into a scalable and efficient end-to-end data platform.

CDP Private Cloud Base is the unification of Cloudera Distribution for Apache Hadoop (CDH) and Hortonworks Data Platform (HDP), giving customers the best of both worlds. This new product combines the best technologies from Cloudera and Hortonworks with new features and enhancements across the stack. CDP Private Cloud Base forms a comprehensive data platform that encompasses the entire data life cycle. This unified distribution is a scalable and customizable platform where you can securely run many types of data analytics workloads.

CDP Private Cloud Base can be a stand-alone data analytics platform. It also supports a hybrid or multicluster solution, where compute tasks can be separated from data storage, and where data can be accessed from remote clusters. In this case, the CDP Private Cloud Base cluster is deployed alongside CDP Private Cloud Data Services, a separate computing cluster running on a container platform that can be deployed with CDP Private Cloud Base. This approach provides a foundation for containerized applications by managing storage, table schema, authentication, authorization, and governance in CDP Private Cloud Base. It consists of various components such as Apache HDFS, Apache Hive 3, Apache HBase, and Apache Impala, along with many other components for specialized workloads. You can select any combination of these services to create clusters that address your business requirements and workloads.

Dell Technologies and Cloudera have designed and validated two separate architectures as the basis for deploying CDP Private Cloud Base; one on AMD-powered Dell infrastructure and one on Intel-powered Dell infrastructure. Both architectures are based on PowerEdge servers, PowerSwitch networking, and PowerScale storage and provide optimized configurations for deploying and operating CDP Private Cloud Base.

In summary, CDP Private Cloud Base is a complete data platform and stand-alone instance of CDP for the on-premises data center that can be deployed on a choice of optimized Dell infrastructure. CDP Private Cloud Base can also be deployed with the CDP Private Cloud Data Services cluster to form the complete CDP Private Cloud. Dell Technologies encourages CDH and HDP customers to upgrade to CDP Private Cloud Base for improved enterprise data management capabilities, new platform innovations, and the ability to add the Data Services cluster when needed.

Document purpose

This document provides an overview of what an enterprise data platform is, along with benefits and typical use cases. It provides a description of CDP, including the component clusters of CDP Private Cloud Base and CDP Private Cloud Data Services. It provides a high-level description of the solution architecture for CDP on Dell infrastructure. It also discusses the journey to CDP, including:

- Upgrades and migrations to CDP Private Cloud Base
- The relation of CDP Private Cloud Base as a foundation for CDP Private Cloud

For more information about the architecture and design of the solution, see the companion design guides for Data Management with Cloudera Data Platform on Dell Infrastructure on the Dell Technologies Info Hub for Data Analytics:

- Data Management with Cloudera Data Platform on AMD-powered Dell Infrastructure Design Guide
- Data Management with Cloudera Data Platform on Intel-powered Dell Infrastructure Design Guide

When CDP Private Cloud Base is being used with CDP Private Cloud Data Services to form CDP Private Cloud, this document should be used with the CDP Private Cloud Data Services documentation on the Cloudera documentation website.

Dell Technologies and Cloudera have been collaborating for over eight years to provide customers with guidance on optimal hardware to streamline the design, planning, and configuration of their Cloudera deployments. This document is based on the collective experience of both companies in deploying and running enterprise production environments.

Audience

This document is intended for data center managers and IT architects who are involved with designing, planning, or operating the hardware and software infrastructure for CDP Private Cloud, for:

- New deployments
- Upgrades or migrations from Cloudera Distribution for Apache Hadoop (CDH) or Hortonworks Data Platform (HDP)

This document assumes some familiarity with CDP capabilities and functions.

Business challenges

The considerations and requirements for data management are constantly evolving.

Market environment

There are new realities for managing data and data-centric workloads across the enterprise in a unified and comprehensive manner:

- Use cases were previously focused on efficiently storing and processing data in batch processes. Now there are increasing needs for integrating the entire data life cycle and for processing in both real time and batch.
- Technology infrastructure formerly demanded the co-location of compute and storage to avoid costly network transfers. Now the needs of high-performance analytics drive a move toward disaggregated compute and storage, where each can be sized and scaled independently.
- From a user experience viewpoint, it used to be acceptable to deploy and run in timeframes of weeks, months, or even quarters. Now the expectation is to be able to spin up services in minutes, give users their own clusters, and get insights quickly.
- From the privacy, security, and governance perspectives, the primary concerns were formerly about network perimeter and physical access controls. Now, with the entire data life cycle being managed, operators need fine-grained authentication and authorization at the workload and data layers.

The emergence of the data platform for end-to-end data management has been one of the most significant developments in the data analytics field.

Data platform overview

Most are very familiar with software applications, especially the plethora of "apps" available for mobile devices. Applications are ready to provide value almost instantly after installation. Think of something like a mapping application with navigation capability. You install the app, turn on location services, enter an address and you are on the way in less than five minutes. Conversely, platforms are tools for application developers. Platforms do almost nothing for end users after installation. Application developers must first configure and build applications using the platform before end users begin to recognize value.

Developers have been using platforms for decades. There are some classes of applications that require core services that are complex to develop but universally useful. In those cases, it makes sense for a group of experienced system developers to build a platform for use by the larger application developer community. Many developers lack the skills to do it on their own. Some of

the first and most successful examples are relational database management systems (RDBMS). These systems include IBM DB2, Oracle, and Microsoft SQL Server.

The RDBMS category has expanded to include many more platforms over the last several decades. Millions of application developers and billions of end users have benefitted from software applications that are developed using these RDBMS platforms.

The most successful data platforms are both robust and flexible. Millions of application developers, who otherwise could not build the scalable foundations that are required to support enterprise class data management, can use them. "Reinventing the wheel" has always been costly and rarely produces superior modes of transportation. Despite that history, many organizations spend many months or years contemplating and prototyping proprietary data platforms.

Enterprise developers may be encouraged that most of the hyperscale Internet companies have developed proprietary data platforms to meet their specific industry and scale challenges. Some of these companies include Airbnb, Facebook, LinkedIn, Lyft, Netflix, Twitter, and Uber.

These organizations differ from most traditional enterprise organizations in several key ways. They were born "cloud native", meaning the platforms that they have developed constitute the business. They can recruit and retain top talent with the backgrounds that are required to build platforms. Also, they are constantly adding the already large initial development investments because their data platform is critical to their main value proposition.

Choosing a data management approach

For most organizations that are not committed to developing a proprietary data platform, the approach most likely to succeed is adopting a full-featured commercial or open-source data platform. Focus your internal development efforts on producing rich applications using platform features in unique and creative ways that add business value. A great data platform may even allow experienced developers to design solutions beyond what the core system developers anticipated.

More organizations understand the importance of extracting insights from data. In response, the open-source and commercial software industries have responded with a growing array of products and services that are marketed under the data platform umbrella. These products include:

- Cloud data platforms
- Big data platforms
- Data management platforms
- Data analytics platforms
- And more

Evaluating whether your organization would benefit from investing in a data platform and then choosing an approach, can be a complex endeavor given the variety of overlapping and competing options. Before starting that journey, it is useful to examine the potential benefits that can make the time and cost of evaluation worthwhile.

Data platform applications

The use of pipeline analogies for describing data work is popular. However, generic discussions only go so far in developing strategies for choosing tools and processes appropriate to any specific use case. The first step in deciding the value potential from adopting a data platform for your organization is to develop the most complete data pipelines library possible. Remember that some data sources are important parts of many pipelines, and other sources may be specific to a single analysis task.

The tracking of these details is important, since it impacts the needs for scalability and reliability when looking at the features of a data platform. It may also be useful to:

1. Look for patterns in the type and number of steps that are required in all the pipelines.
2. Group the patterns that have many similarities.

You may find that one platform is not adequate for all the needs of the organization, but in most situations there are many commonalities.

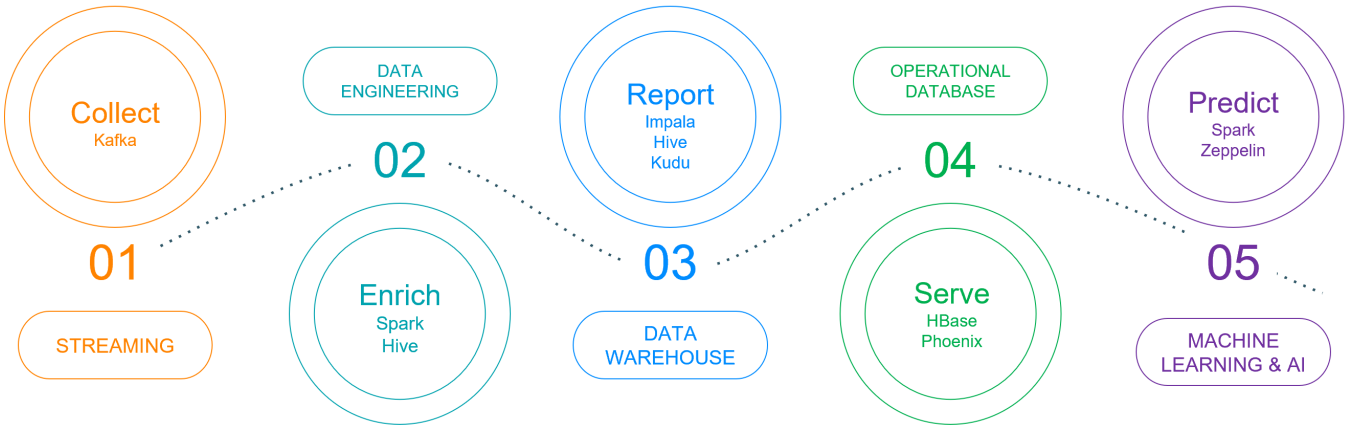

Figure 1. Generic data pipeline

Generic data pipeline is typical of generic data analytics pipeline that shows the end-to-end functional categories that are required for many types of data work. A high-level view like this is not enough for evaluating a data platform investment. The task details for a category like Collect (for example, how many and what types of data sources) significantly impact the features that you need from a data platform. The potential variety and complexity of the Enrich category is often underestimated in tools and storage performance assessments.

Each of the pipeline processing categories from Generic data pipeline is also a market for specialty software that applies only to that category. Different platforms and specialty applications may use different terminology than the Collect, Enrich, Report, Serve, and Predict terms as shown here. However, the concepts and functional requirements are generally the same.

Data platforms that meet all or most of the needs of your data pipelines simplify the process of getting from raw source data to insights. Whenever data in the pipeline must move between platforms there is a real possibility of introducing complexity both in the development phase and in sustaining operations.

Data management

The value of implementing a robust data platform lies in a broad spectrum of data sources and types. This data can contain hidden or latent information that is combined with a common framework for applying a full suite of data analytics techniques. While there are common analytic applications that almost every organization knows about, there are probably as many or more yet to be discovered and developed. Many organizations acknowledge that the backlog of proposed applications that are based in part on analytics insight is overwhelming. Many sources of data in large organizations have yet to be profiled let alone enhanced and merged into an analytics pipeline. Such a pipeline feeds value into a software application or report.

Most digital data has some type of structure or common properties when it is committed to a storage medium. Some examples include:

- Files have a size property and a filetype (application, text, binary).
- Text files have an encoding scheme.
- Images have dimensional size and color depth encoding.
- Audio has bit rate and frequency range.

These characteristics impact the requirements for a data platform. Some file systems are better suited to handling lots of small files while others are better at fewer large files. For audio and other "stream based" data, data engineers have a choice of buffer size and file creation characteristics that must be matched to the capabilities of the platform and may also impact the complexity of using the data for analysis.

If you have more knowledge about the final stages of how your analysis pipelines look, you can build more intelligence into the early stages of data management. If possible, one area that should be resisted is “down sampling” the data because of the capability or preferences of the reporting and modeling requirements. Although storing high-fidelity data when it is not required for analysis may seem wasteful, think of it as an insurance policy to protect against changing analysis requirements. Storing data in a form that matches the data generation process as closely as possible can provide many clues, should questions regarding data reliability or quality arise later. You can always look at using down sampling or other forms of compression that lose information for archive.

Another aspect of data management, that surprises IT professionals, is the storage that is required to manage multiple copies of data being used for analysis. Even the most seasoned data science professionals consume many copies of data that by some appearances are identical. There are several important reasons why this situation is necessary:

1. Both report and model development must be isolated from uncontrolled change. This initial copy is typically a direct copy of source with little or no transformation. This measure ensures that the developers can always return to a ground truth version of the data. That data can be used to compare alternate transformation schemes with repeatability.
2. Managing alternate transformations. One common pattern is grouping and counting events by various factors such as time, geography, market segment, and so on, as well as transformations to clean and regularize the data.
3. Efficiency. Complex data transformation pipelines should get developed in stages. It may be too inefficient to go all the way back to source data for testing an incremental set of tasks late in the pipeline. Data scientists may prefer to stage intermediate steps to reduce the complexity and time investment to run the pipeline from the absolute beginning.

This list is not exhaustive, but should provide some ways to assess the sizing of a data platform. More importantly, it can help you assess the flexibility that is available for expanding and tiering storage that candidate platforms provide. Another requirement that derives in part from the data copy management challenge is tracking metadata that are associated with transformation logic and history. Creating many copies of the same data may seem reasonable in the heat of shipping a project, but it will be difficult to ascertain why six months later.

There is a growing interest in platforms that include "feature stores". The concept is to both better track logic and metadata and to promote a more disaggregated approach to data management. If the only difference between two datasets is how the customer dimension is managed, you should keep two copies of that feature, rather than two copies of the entire dataset. This is a simple example to explain the basic idea. Reusing transformation logic to manage frequently used dimensions - like customers and products - independently from all the other features, and all the other analysis datasets in which they are used, could greatly simplify data management.

Example use cases

The potential list of use cases that a full-featured data platform can address is nearly limitless. Looking at the intersection of industry type, data sources, business function, and value alone is too large a list to document. The following list gives some sense of the common use cases that Dell Technologies sees most often:

- Customer 360 analytics
- Retail inventory and sales analysis
- Manufacturing operational analysis
- eCommerce fraud prevention
- Network security intelligence
- Data warehouse consolidation
- Discount pricing optimization
- Financial services
- Insurance industry predictive analytics
- Recommendation engines
- Social media analysis and engagement

A good business practice is to maintain an active list of potential use cases where the availability of a data platform can enhance development. Assess the list so that you do not tackle too many use cases that are high-priority and high-investment too early.

This document describes two of the use cases in more detail: financial services and manufacturing.

Financial services

Financial services encompass a wide range of business models, including:

- Consumer and commercial banking
- Individual wealth management
- Primary or secondary capital markets

The importance of relationship management is shared across all these businesses and therefore has been a key focus area for analysis. Virtually all midsized and larger financial services organizations have one or more data platforms. The intense pressure to compete with other players makes finding, securing, keeping, and nurturing relationships with customers a priority that drives profit. There is also a requirement to manage investment risk and assure compliance with all regulatory requirements, which often involve multiple, overlapping jurisdictions.

While personal relationships still matter, data driven modeling and reporting across multiple channels including mobile, online, phone, or branch agent are a must-have for these organizations. Organizations that build trust by arming the organization with data-driven information increase the confidence of their customers, along with wallet share and lifetime value. To achieve that on a global scale, you must leverage big data and predictive analytics using a proven and modern hybrid data platform.

Manufacturing

Industry 4.0 is an emerging term that means smart manufacturing. Advanced technologies are combined with traditional manufacturing and industrial practices to improve operational efficiency across the board. The innovations and documented successes of Industry 4.0 initiatives are encouraging more manufacturing companies to adopt Industrial IoT (IIoT) concepts and technology. Such adoptions transform product development, supply chains, and manufacturing operations.

Many recent case studies show that connecting analysis of smart products, design engineering, factory floor operations, and customer experience enable faster time-to-market, improved product quality, and scaling production output while reducing waste and operating costs. Connected products are a key initiative of Industry 4.0. The connectivity these products provide drives customer satisfaction and revenue while reshaping the relationship between people and products.

Achieving these benefits requires the abilities to ingest, process, and analyze sometimes massive volumes of IoT data. This data processing scale enables manufacturers the access to near real-time customer feedback to identify product quality issues. Another growing area of Industry 4.0 is intelligent supply chain management. Disruptions and delays in a critical supply chain will ripple through an organization from sales to operations.

Many manufacturers are using near real-time data, analytics, and machine learning to ensure that supply chains are functioning well while risk is managed end-to-end. Combined with a modern data platform that supports advanced analytics, including machine learning capabilities, required investments to take advantage of these latest innovations in manufacturing include:

- Special purpose sensors
- GPS
- RFID
- Production stream data

Apache Hadoop overview

During the startup incubation of Google, the founders realized that to revolutionize the efficiency and relevance of web search, they had to develop new computing tools.

Google needed both a new scale-out file system and a new scale-out computing platform to deal with:

- The number of URLs that existed on the Internet in the early 2000s
- The complexity of analyzing the interpage linking relationships

The first publicly available descriptions of one method for overcoming those two challenges were published as public white papers in 2003 to 2004. The Yahoo researchers who developed the first versions of the Hadoop Distributed File System (HDFS) and the Hadoop MapReduce computing platform credit those early Google papers for the architecture foundations that started the Hadoop open-source initiative.

Cloudera and Hortonworks

In 2018, Cloudera and Hortonworks announced they would merge to form a single company. This merger completed in January 2019. Its stated goal was to produce the first enterprise data cloud, with a platform to support hybrid and multicloud deployments, and contain 100% open-source components. Originally released as CDP Data Center and described in the following chapter, CDP Private Cloud Base is the first release from the combined company. It integrates the best of Cloudera and Hortonworks technologies into an on-premises offering.

Cloudera Data Platform

CDP is an integrated data platform that is easy to deploy, manage, and use for a wide range of Data Analytics capabilities. By simplifying operations, CDP reduces the time to onboard new use cases across the organization. CDP can be deployed in a public cloud, in an on-premises data center, and as an on-premises private cloud.

The focus of this document is CDP Private Cloud Base, which is the first on-premises release to combine CDH, most recently known as Cloudera Enterprise Data Hub, and HDP.

NOTE: This document generally uses "CDH" and "HDP" when referencing prior versions from Cloudera and Hortonworks, respectively.

CDP Private Cloud

The complete CDP Private Cloud offering is the next step in the CDP journey. CDP Private Cloud Base is a mandatory component of, and is the base for, CDP Private Cloud. It is the storage and data lake cluster, and contains the SDX layer. It is therefore important to have some understanding of the full CDP Private Cloud as you plan your new deployment or upgrade to CDP Private Cloud Base.

CDP Private Cloud overview

The complete CDP Private Cloud offering delivers a cloud-like experience in customer data center environments. CDP Private Cloud is a new approach to data management and analytics that delivers powerful self-service analytics across hybrid and multicloud environments. CDP Private Cloud Data Services leverages disaggregated compute and storage models to provide:

- Simpler multitenancy and isolation
- Better infrastructure utilization
- Containerization
- Cloud native architecture

As shown in CDP Private Cloud high-level architecture, CDP Private Cloud builds on the storage and services that are established in CDP Private Cloud Base and delivers what are known as "Data Services", as containerized workloads. These workloads will, in time, include:

- Data Flow and Streaming
- Data Engineering
- Data Warehouse
- Operational Database
- Machine Learning

NOTE: Data Engineering, Data Warehouse, and Machine Learning are the first three workloads that are delivered in this release.

The Cloudera Shared Data Experience supports CDP Private Cloud and CDP Private Cloud Base with all the capabilities of security, metadata, and governance.

Spanning across the platform is a management console that delivers a unified control plane which operates across multiple deployments.

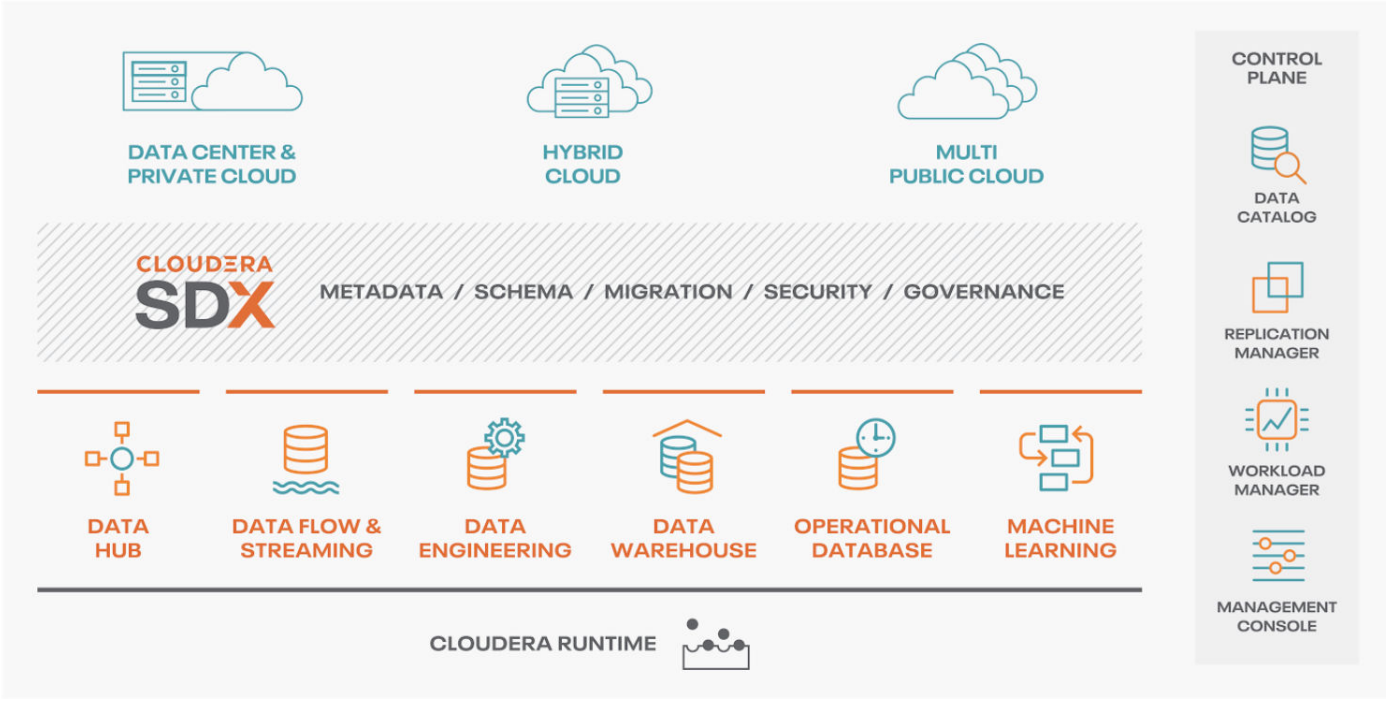

Figure 2. CDP Private Cloud high-level architecture

CDP Private Cloud architecture

Two clusters are deployed with CDP Private Cloud:

- The CDP Private Cloud Base cluster, which runs on Red Hat Enterprise Linux Server
- The CDP Private Cloud Data Services cluster, which runs on a container platform

These two clusters are separate, and are independent tracks from an architecture and deployment planning perspective. CDP Private Cloud Base and CDP Private Cloud Data Services clusters illustrates these primary components in a complete deployment of CDP Private Cloud.

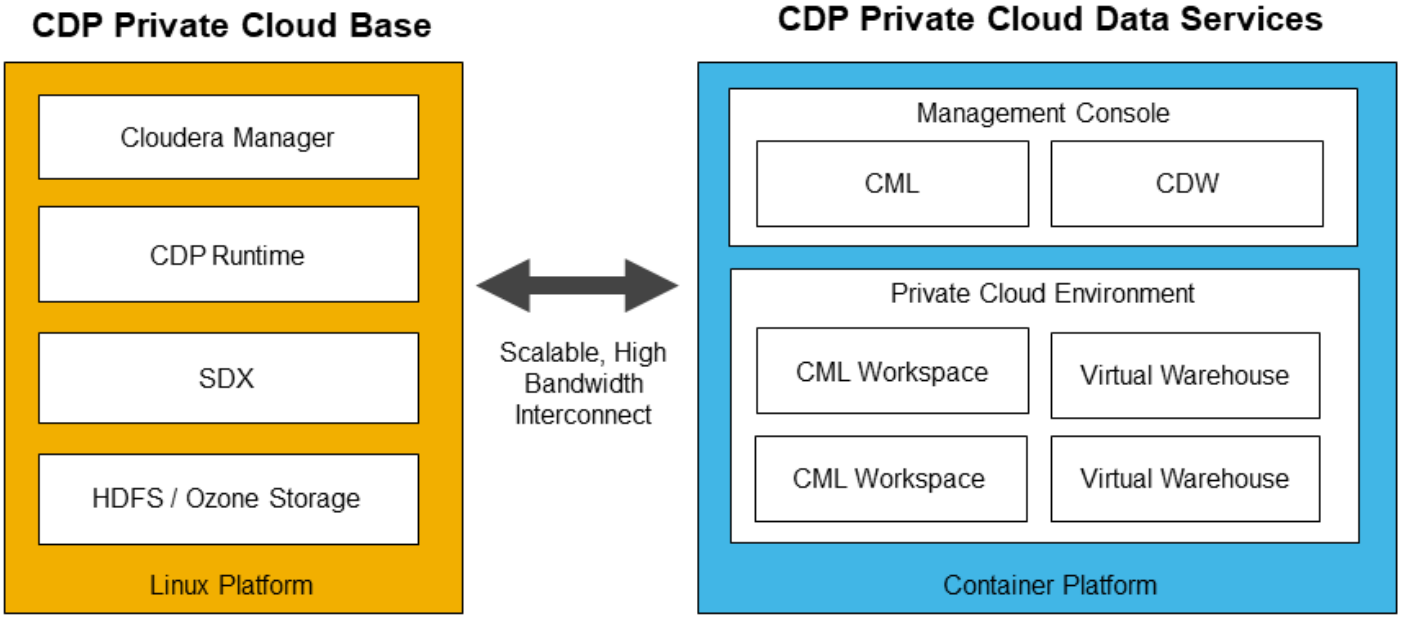

Figure 3. CDP Private Cloud Base and CDP Private Cloud Data Services clusters

An instance of CDP Private Cloud Base remains the base data lake cluster when you upgrade to CDP Private Cloud. When planning your CDP Private Cloud Base installation and potential hardware refresh, you should review:

- The listed Data Management with Cloudera Data Platform on Dell Infrastructure Design Guides
- The CDP Private Cloud Data Services documentation on the Cloudera documentation website

CDP Private Cloud Base

CDP Private Cloud Base is a comprehensive, on-premises platform for integrated data analytics. CDP Private Cloud Base encompasses ingest, processing, analysis, experimentation, and deployment. It integrates the best of CDH and HDP to deliver the latest and best open-source data management and analytics technologies. CDP Private Cloud Base is optimized for deployment within the data center, and ready for private cloud.

A core layer of CDP Private Cloud Base is Cloudera Shared Data Experience (SDX), with uniform capabilities of Data, Schema, Replication, Security, and Governance. Cloudera SDX Shared Data Experience includes the following capabilities:

Schema	Automatic capture and storage of all schema and metadata definitions as platform workloads use and create them.
Replication	Deliver data copies and data policies that the enterprise requires to work, with complete consistency and security.
Security	Role-based access control applied consistently across the platform, including full stack encryption and key management.
Governance	Enterprise-grade auditing, lineage, and governance capabilities applied across the platform with rich extensibility for partner integrations.

CDP Private Cloud components shows a high-level view of CDP Private Cloud Base in relation to CDP Private Cloud Data Services. Cloudera Runtime consists of a large set of software components including Apache Hadoop, Apache Hive 3, Apache HBase, and Apache Impala, and many other components for specialized workloads. The full list is shown in CDP Private Cloud Base software components.

Several preconfigured packages of services, sometimes known as cluster shapes, are available for common workloads on CDP Private Cloud Base. These services include:

Data Engineering	Provides the abilities to ingest, transform, and analyze data. Services include: HDFS, YARN, YARN Queue Manager, Ranger, Atlas, Hive, Hive on Tez, Spark, Oozie, Hue, and Data Analytics Studio.
Data Mart	Enables you to browse, query, and explore your data in an interactive way. Services include: HDFS, Ranger, Atlas, Hive, Impala, and Hue.
Operational Database	Provides low-latency writes, reads, and persistent access to data for Online Transactional Processing (OLTP) use cases and real-time insights. Services include: HDFS, Ranger, Atlas, and HBase.

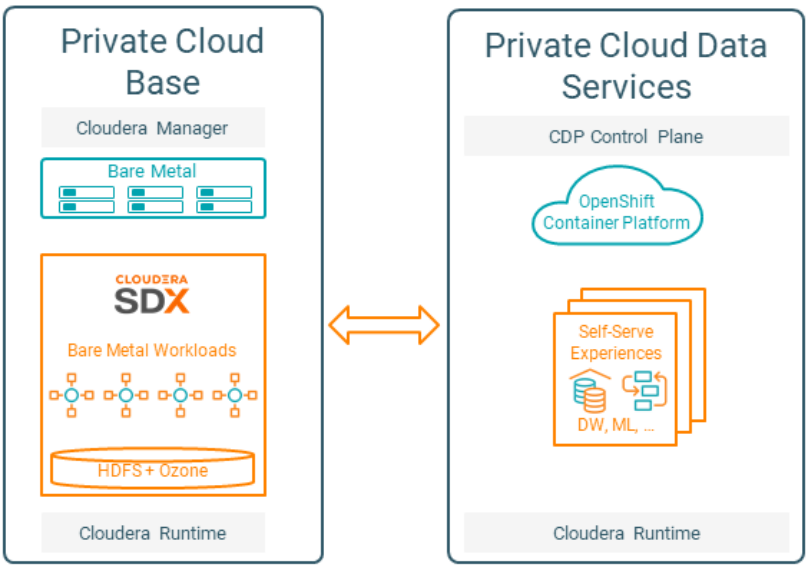

Figure 4. CDP Private Cloud components

You can also create custom services and clusters from Cloudera Manager, which deploys any combination of supported services that you select from all available services in the Cloudera Runtime distribution.

CDP Private Cloud Base benefits

Key features, improvements, and benefits of CDP Private Cloud Base7.1.7 include:

Streams Messaging	Complete and comprehensive Kafka streaming experience improving operational efficiency, business continuity, and scalability.
Data Engineering	Improved performance and interoperability for Apache Spark and management of data engineering workflows and pipeline creation.
Data Warehouse	Faster SQL analytics on larger datasets, deeper understanding from unstructured data sources, and easier visualizations of business insights.
Machine Learning	Data Science Workbench is now available on CDP Private Cloud Base with advanced control over experiments and model deployment.
Operational Database	Improved performance, policy management, and availability.
SDX	Enhanced security, compliance, and consistency across CDP.
Support for in-place upgrades and migrations	From CDH 5.13+ and CDH 6.1+, and from HDP 2.6.5 and HDP 3.1.5, to CDP Private Cloud Base.

The features and capabilities that are new to users migrating or upgrading from CDH or HDP are described in CDP Private Cloud Base components.

CDP Private Cloud Base components

Cloudera Runtime is the core open-source software distribution within CDP that Cloudera maintains, supports, versions, and packages as a single entity. Cloudera Runtime includes multiple open-source projects, including Apache components, connectors and encryption components, and other components from Cloudera. These components constitute the core distribution of data management tools within CDP.

Cloudera Manager is a web application that administrators and others can use to configure, manage, and monitor CDP clusters and Cloudera Runtime services. You can also use the Cloudera Manager API to programmatically perform management tasks.

CDP Private Cloud Base software components shows the major Apache software components that constitute Cloudera Runtime7.1.7 for CDP Private Cloud Base, along with a brief description of each. For more information about all included components, including versions, see Cloudera Runtime Component Versions on the Cloudera documentation website.

The associated Data Management with Cloudera Data Platform on Dell Infrastructure Design Guides describe where these components are deployed across the various nodes.

Table 2. CDP Private Cloud Base software components

Component	Description
Apache Arrow	Arrow is a cross-language development platform for in-memory data.
Apache Atlas	Atlas provides data governance capabilities for Hadoop. Atlas is also a common metadata store, which is designed to exchange metadata within and outside of the Hadoop stack.
Apache Avro	Avro is a row-oriented remote procedure call and data serialization framework for Apache Hadoop.
Apache Calcite	Calcite is a framework for building databases and data management systems. It includes a SQL parser, an API for building expressions in relational algebra, and a query planning engine.
Apache Hadoop	Apache Hadoop is a framework that enables distributed processing of large datasets across clusters of systems, using simple programming models. Apache Hadoop is designed to scale out from single servers to thousands of servers. Hadoop also includes

Table 2. CDP Private Cloud Base software components (continued)

Component	Description
	YARN for resource management and job scheduling and HDFS, the Hadoop Distributed File System.
Apache HBase	HBase provides random, persistent access to data as a natively nonrelational database. HBase is ideal for scenarios that require real-time analysis and tabular data for end-user applications.
Apache Hive	Hive is a data warehouse system for summarizing, querying, and analyzing huge, disparate datasets.
Apache Impala	Impala provides high-performance, low-latency SQL queries on data stored in Apache Hadoop file formats.
Apache Kafka	Kafka is a distributed and highly available event streaming platform. It is used for high-performance data pipelines, streaming analytics, data integration, and mission-critical applications.
Apache Knox	Knox is an application gateway for interacting securely with the REST APIs and user interfaces of one or more Hadoop clusters.
Apache Kudu	Kudu combines fast inserts and updates, and efficient columnar scans, to enable multiple real-time analytic workloads across a single storage layer. Kudu provides fast analytics on fast data.
Apache Livy	Livy is a service that enables easy interaction with a Spark cluster over a REST interface.
Apache MapReduce	MapReduce is a software framework for writing applications that process vast amounts of data in-parallel on large clusters in a reliable, fault-tolerant manner.
Apache Oozie	Oozie is a workflow and coordination service for managing Apache Hadoop jobs.
Apache ORC	Optimized Row Columnar (ORC) is a self-describing, type-aware columnar file format designed for Hadoop workloads.
Apache Ozone	Ozone is a scalable, redundant, and distributed object store optimized for big data workloads.
Apache Parquet	Parquet is a columnar storage format available to any project in the Hadoop ecosystem, regardless of the data processing framework, data model, or programming language.
Apache Phoenix	Phoenix is an add-on for Apache HBase that provides a programmatic ANSI SQL interface.
Apache Ranger	Ranger is a CDP security component that enables you to control access to CDP services. Ranger also provides access auditing and reporting.
Apache Solr	Solr provides natural language access to data stored in, or ingested into, Hadoop, HBase, or cloud storage.
Apache Spark	Spark is a distributed, in-memory data processing engine designed for large-scale data processing and analytics.
Apache Sqoop	Sqoop is a CLI-based tool for bulk transfers of data between relational databases and HDFS or cloud object stores.
Apache Tez	Tez is an extensible framework for building high-performance batch and interactive data processing applications, which YARN coordinates in Apache Hadoop.
Apache YARN	YARN is the processing layer for managing distributed applications that run on multiple machines in a network.
Apache Zeppelin	Zeppelin is a multipurpose, web-based notebook that enables data-driven, interactive data analytics and collaborative documents with SQL, Scala, Python, R and more.
Apache ZooKeeper	ZooKeeper is a centralized service that enables highly reliable, distributed coordination, including maintaining configuration information, naming, and providing distributed synchronization and group services.

New features

It is useful to understand what features and components are net new in CDP Private Cloud Base. It is also helpful to know what is new and changed when coming from the legacy CDH or HDP releases.

While this topic describes what is in the release, more information about the journey to CDP Private Cloud Base, including the upgrade and migration pathways, is described in the chapter, Journey to CDP Private Cloud Base.

Net new features

There are several new features and capabilities that have been included for the first time in CDP as compared to the former CDH and HDP products. These features are over and above what was in prior CDH and HDP releases, and therefore are new to all users who deploy CDP Private Cloud Base. These new features include:

Atlas 2.0	Includes advanced data discovery, metadata catalog and search, data lineage and chain of custody, metadata audit, and support for improved security. Also includes support for Spark.
Enhanced security capabilities	Includes encryption with Ranger KMS-Key Trustee integration, and Navigator Encrypt (Navencrypt) for secure data at rest.
Streaming Services	Introduced with the addition of Kafka and related components, includes cluster management and replication for Kafka clusters, storage and schema through the schema registry service, and the ability to rebalance clusters with Cruise Control. Also includes support for Kafka Connect, which enables you to connect HDFS, Amazon S3, and Kafka Streams.
Ozone object storage	Ozone is a next-generation file system that bridges object store and HDFS, and supports billions of objects.

CDH to CDP Private Cloud Base changes

Capabilities that are new to prior users of CDH include:

Ranger security	Provides full dynamic capability to set up policies and authorizations, with fine-grained access control, dynamic row filtering, dynamic column masking, and attribute-based access control. With Impala as part of the distribution, Impala-Ranger integration is available, so any policy can be propagated to Impala, Hive, and Kudu.
Hive 3 data warehouse software	Includes Atomicity, Consistency, Isolation, and Durability (ACID) support for better ETL performance, and comprehensive ANSI SQL 2016 coverage.
Hive on Tez	Integrates Hive with Tez, an extensible framework for building high-performance batch and interactive data processing applications, providing better ETL performance at petabyte scale.

HDP to CDP Private Cloud Base changes

Capabilities that are new to prior users of HDP include:

Virtual Private Clusters	Virtual Private Clusters simplify deployment of applications and enable workloads running in different clusters to securely and flexibly share data.
Hue	Hue is a web-based interactive query editor for interacting with databases and data warehouses. It delivers an integrated SQL editor with autocompletions, visualizations, and connection to Hive and Impala to run SQL queries seamlessly.
Kudu	Kudu is a columnar storage manager for fast analytics on fast data. It supports variable character field (varchar) and datatype column, Ranger Authz integration, and fast changing of updatable data for better performance.

Impala	Impala is a SQL query engine for massively parallel processing (MPP) queries. It is ideal for Data Mart migration, interactive SQL, and Business Intelligence (BI) style queries like access reports or dashboards, through Tableau or other BI tools.
Cloudera Manager	Cloudera Manager is a web application for managing multiple clusters. It is a change from Apache Ambari in HDP, and includes automated wire encryption setup, fine-grained role-based access control (RBAC) for administrators, and streamlined maintenance workflows.

Journey to CDP Private Cloud Base

This chapter describes the general upgrade pathways to CDP Private Cloud Base. It provides a high-level summary of how to get from CDH and HDP to CDP Private Cloud Base, including migrating your data or upgrading the platform. Further details of upgrades and migrations, including variations that are known as side-car and rolling side-car migrations, can be found in this Cloudera blog.

NOTE: This document provides a description of the possible pathways and some of the considerations, but is not intended to explain all the steps required.

Paths to CDP Private Cloud Base

NOTE: Deployment, migration, and upgrade projects can be complex. They require careful planning, with assistance from infrastructure and software teams, to ensure success. Dell Technologies highly recommends that you engage with both Cloudera and Dell Technologies for project planning and execution assistance.

There are several paths to get to CDP Private Cloud Base, aside from a new, or "greenfield", installation. For existing installations there are two approaches:

Migration	With this approach, as shown on the left of Migration vs. upgrading, you: 1. Deploy a new CDP Private Cloud Base cluster on-premises and on new hardware infrastructure. 2. Copy the data and metadata from the existing cluster. 3. Migrate the existing workloads.
In-place upgrade for supported upgrade pathways	With this approach, as shown on the right of Migration vs. upgrading, you: 1. Perform required preparations to upgrade from the legacy cluster to CDP Private Cloud Base. 2. Perform an in-place upgrade on the same hardware infrastructure.

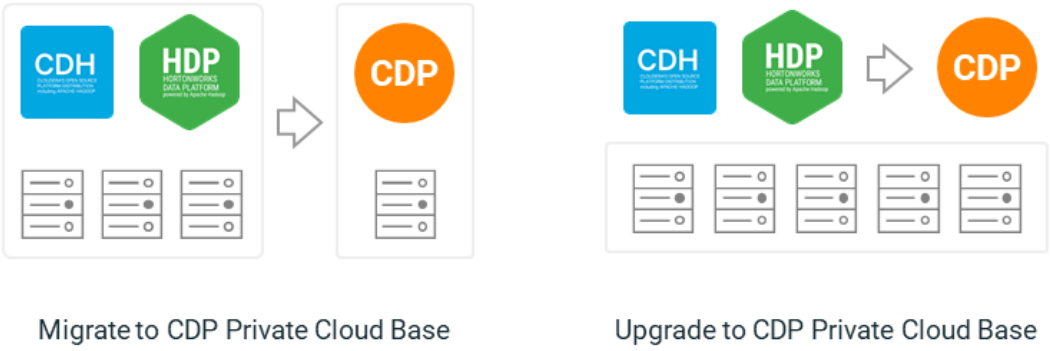

Figure 5. Migration vs. upgrading

Further details on both these approaches are described in:

- Migrate to CDP Private Cloud Base
- Upgrade to CDP Private Cloud Base

Migrate to CDP Private Cloud Base

There are multiple scenarios where migration is the best or most feasible approach. For example, you:

- Have capacity on a new cluster.
- Are doing a hardware refresh such as for greater capacity or performance.
- Do not want to disturb existing workloads.
- Can move workloads one at a time.
- Do not want any downtime.
- Have an existing instance of CDH or HDP that does not have a direct upgrade supported.

There are several tools available, including:

Workload XM	Enables you to migrate or shift workloads after analyzing them, and helps you move workloads one by one.
Replication Manager	Helps you replicate and copy the data and metadata.

The migration process is to:

1. Configure the new cluster.
2. Identify the candidate workloads.
3. Copy the data, metadata, and policy.
4. Migrate and test the workloads.
5. Promote the new cluster and workloads into production.
6. Decommission the legacy cluster.
7. Depending on hardware compatibility, add the nodes to the new CDP Private Cloud Base cluster for additional capacity.

This process enables you to perform a rolling migration of the cluster, by both:

- Deploying new hardware.
- Repurposing existing hardware by gradually migrating data and workloads to the new cluster.

An overview of the migration process is shown in Migration to CDP Private Cloud Base.

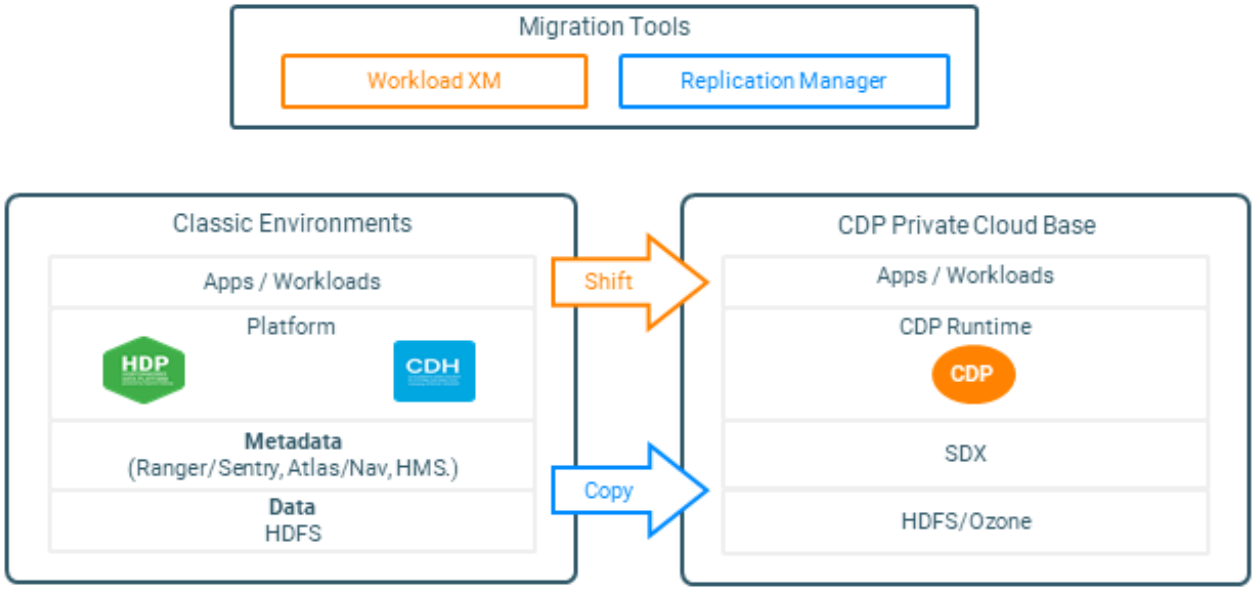

Figure 6. Migration to CDP Private Cloud Base

Upgrade to CDP Private Cloud Base

Sometimes, an in-place upgrade may be preferable to a migration, or it may be that a migration cannot be done, such as when:

- There is no additional hardware capacity available.
- There are multiple clusters where the upgrade can be tested in lower-priority environments.
- The workloads may be more tolerant of downtime, such as for a single tenant cluster.
- There are not multiple types of jobs running in the cluster.

Upgrade tools include:

Cloudera Manager7.5.4	For CDH users, Cloudera Manager7.5.4 facilitates: • The upgrade from previous versions of Cloudera Manager • The upgrade from the prior runtime to the current runtime with all the components
Apache Ambari	Manages upgrades for HDP users

In order to avoid potential compatibility issues, an upgrade is typically accomplished with a professional services engagement. In the current release, CDP Private Cloud Base7.1.7, upgrades are supported from:

- CDP Private Cloud Base 7.0
- CDH 5.13 to 5.16
- CDH 6.1 to 6.3
- HDP 2.6.5 or 3.1.5

An overview of the migration process is shown in Upgrading to CDP Private Cloud Base.

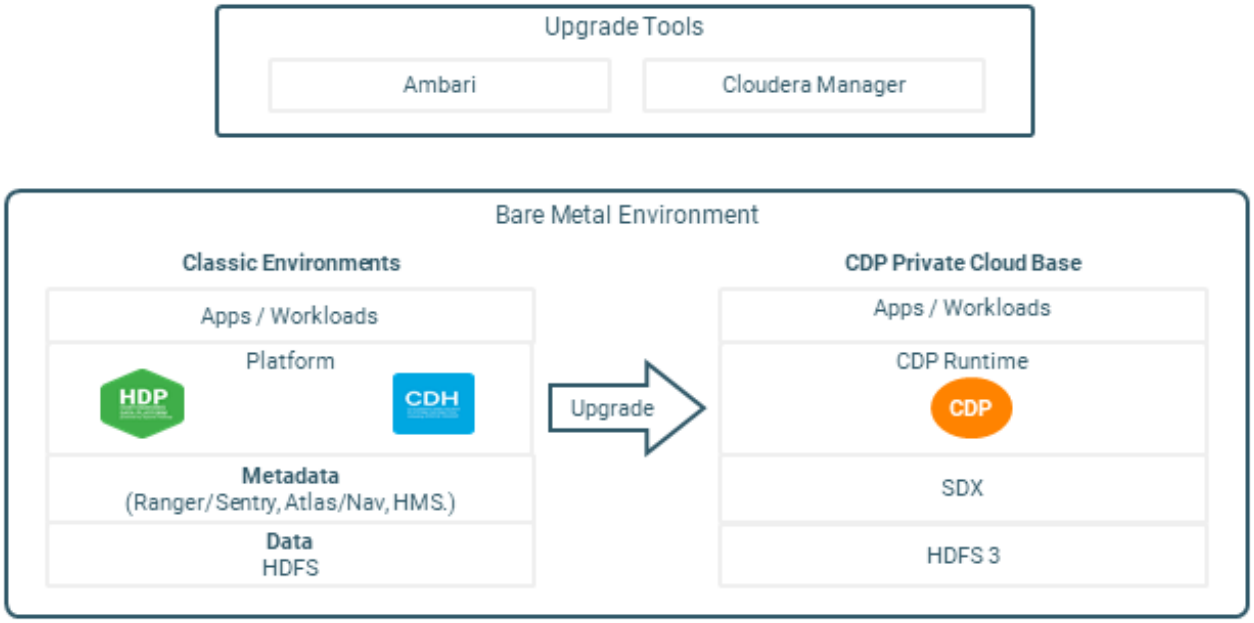

Figure 7. Upgrading to CDP Private Cloud Base

Considerations

Legacy environments can be complex and varied. To reduce risk in the upgrade or migration, there are some factors that you must consider during the planning process.

Upgrades are complex and have many prerequisites. Examples include:

- Upgrading the existing versions of individual components to a supported, upgradable version
- Converting to different components before the overall platform upgrade

There are also differences whether you are upgrading from CDH or HDP. The HDP process requires intermediate steps, such as upgrading and using Apache Ambari with HDP before converting to Cloudera Manager for CDP.

NOTE: Planning is essential. For assistance, you can engage with Cloudera for a Journey Workshop to assist with the planning.

In order to mitigate risk in the upgrade or migration, you must consider data replication and protection before beginning the process. Planning for workload testing and validation is also important for derisking the move to CDP. If a plan is not already in place, you can set up a multicluster replication scenario for the legacy cluster before the upgrade, and for the new cluster.

For full details on upgrade and migration pathways, including checklists, tools, and complete instructions, see the Cloudera documentation about CDP Private Cloud Base upgrades.

Hardware refresh

The upgrade or migration planning period is the ideal time to consider if you need, or could benefit from, a hardware upgrade. Plan in terms of both capacity and performance by considering both:

- The listed Data Management with Cloudera Data Platform on Dell Infrastructure Design Guides
- The hardware requirements for CDP Private Cloud Data Services, which can be found in the CDP Private Cloud Data Services documentation on the Cloudera documentation website

Solution architecture

This topic presents a high-level description of the solution architecture for CDP on Dell Infrastructure.

For full details about the architecture and components of the solution, see the listed Data Management with Cloudera Data Platform on Dell Infrastructure Design Guides. These details include:

- The use of Dell PowerScale storage
- High availability features
- Solution sizing and scaling
- The network architecture
- The precise configuration of all the Infrastructure nodes and Worker nodes
- The validation methods that were used on this Dell Technologies Validated Design

Architecture introduction

CDP Private Cloud Base provides data management, enterprise analytics, and management tools for big data. The data management services include HDFS file storage and Ozone object storage. The Cloudera Runtime provides the analytics services, which include components like Hive, HBase, MapReduce, and Spark. The management tools include:

- Cloudera Manager for cluster management, monitoring, and configuration
- Cloudera SDX for security, governance, and metadata

Successful deployment and operation of Cloudera CDP Private Cloud Base depends on a well-designed infrastructure, with an architecture that provides high performance, scalability, reliability, and manageability. A Dell Technologies Validated Design does exactly that.

Architecture design

CDP Private Cloud Base is deployed on a cluster of multiple physical server nodes. Each node has a particular configuration that is designed for its role in the cluster. These nodes are further specialized through the software services that are assigned to them.

There are two basic types of nodes: Infrastructure nodes and Worker nodes. Infrastructure nodes run on a common server configuration based on PowerEdge servers.

Infrastructure nodes include three specific node types: Master node, Utility node, and Edge node. While these nodes use the same physical configuration, they take on different roles based on the software services deployed on each of them. There are three Master nodes, and together they run all the services that are required to manage the compute services and cluster storage. The Utility node runs Cloudera Manager and the Cloudera Management Services. The Edge node contains the client-facing configurations and services.

The table below defines the various cluster nodes and their physical configuration.

Table 3. Node definitions

Node	Definition	Physical configuration
Master node	This node runs all the services that are required to manage the cluster storage and compute services.	Infrastructure node
Utility node	This node runs Cloudera Manager and the Cloudera Management Services.	Infrastructure node
Edge node	This node contains all client-facing configurations and services, including gateway configurations.	Infrastructure node
Worker node	This node runs all the services that are required to store blocks of data on the local hard drives, and run processing tasks against that data.	General purpose Worker node
		All flash Worker node[a]
		GPU accelerated Worker node
		PowerScale Worker node

a. Intel version only

The minimum supported configuration for CDP Private Cloud Base is eight cluster nodes, which include three Master nodes, one Utility node, one Edge node, and three Worker nodes. Dell Technologies recommends a ten-node cluster with five Worker nodes as a starting point.

The node-level architecture diagram below shows the node types and assignments, and the primary software services, that are deployed on each node.

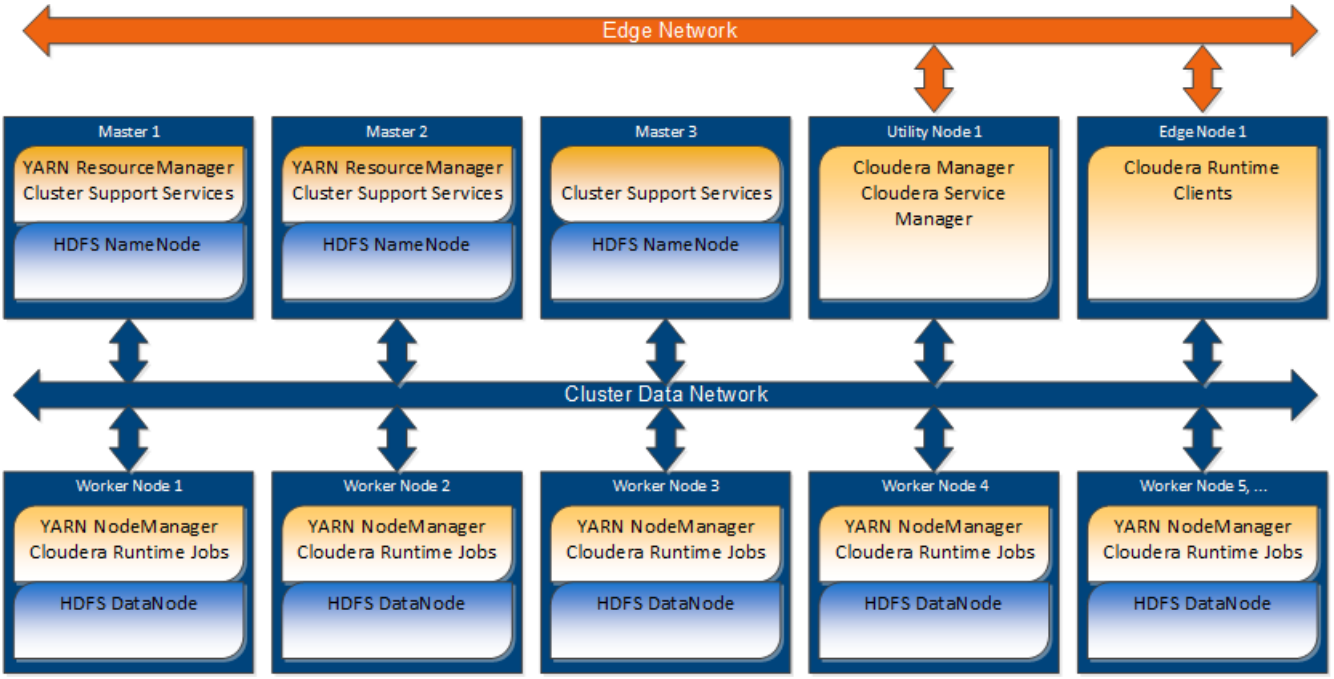

Figure 8. Node architecture

Network design

A high-performance network fabric connects the cluster nodes in a Cluster Data network. An additional Edge network provides an interface between the cluster and external systems and applications. There is also a Management network that connects the iDRAC management ports of the PowerEdge servers for hardware provisioning and management.

The network is designed to meet the needs of a high performance and scalable cluster, while providing redundancy and access to management capabilities. The architecture is a leaf and spine model that is based on 25 GbE networking technologies. It uses PowerSwitch S5248F-ON switches for the leaves and PowerSwitch Z9432F-ON switches for the spine.

PowerScale storage

As an option, this architecture supports the use of Dell PowerScale storage, a highly flexible scale-out network-attached storage solution that can be used as the primary HDFS storage.

Compute and storage can be scaled independently using this alternative architecture. The PowerScale Isilon storage nodes provide the HDFS NameNode and DataNode services instead of the services being assigned to the Master nodes and Worker nodes. The Worker nodes only include enough storage for runtime operations like shuffle-sort spill files and cache.

This alternative architecture reduces the HDFS bandwidth requirements for the Cluster Data network. PowerScaleOneFS implements data durability internally and uses a private back-end network for internal operations. A single copy of the data is transferred to the Isilon storage nodes when Worker nodes write to HDFS. No replication traffic occurs on the Cluster Data network. Also, HDFS recovery traffic for failed drives or nodes does not occur on the Cluster Data network.

In this architecture, Dell Technologies recommends the Isilon H5600 hybrid configuration for storage in clusters using PowerScale storage for their primary HDFS storage. See the associated Data Management with Cloudera Data Platform on Dell Infrastructure Design Guides for more information about this configuration.

Software components

The software components and versions that are validated for CDP Private Cloud Base are listed in the "Software components" table within the listed Data Management with Cloudera Data Platform on Dell Infrastructure Design Guides.

Architecture summary

This chapter provides an overview of the solution architecture for CDP Private Cloud Base on Dell infrastructure. For full details about the server node configurations and the network and storage designs, see the listed Data Management with Cloudera Data Platform on Dell Infrastructure Design Guides.

Conclusion

CDP Private Cloud Base is the first on-premises release of CDP that combines the former Cloudera and Hortonworks software into a single, comprehensive data platform. CDP Private Cloud Base is also used with CDP Private Cloud Data Services to form CDP Private Cloud.

Document summary

The purpose of this document is to provide important background and information for data analytics infrastructure managers and architects who want to run CDP Private Cloud Base on Dell hardware infrastructure. Topics that are discussed include:

- What a data platform is
- The wide range of use cases for a data platform
- The details of CDP Private Cloud Base
- The relationship of CDP Private Cloud Base to CDP Private Cloud Data Services
- A high-level description of the solution architecture for CDP on Dell Infrastructure
- The journey to CDP Private Cloud Base, including upgrade and migration strategies

Dell Technologies and Cloudera have been collaborating for over eight years to provide customers with guidance on optimal hardware to streamline the design, planning, and configuration of their Cloudera deployments. Dell Technologies is a Platinum member of the Cloudera IHV Program, the highest level of partnership that indicates ongoing commitments to both Cloudera and customers. This document is based on the collective experience of both companies in deploying and running enterprise production environments for Cloudera software on Dell hardware infrastructure.

For further information about this solution, see the companion Design Guide for Data Management with Cloudera Data Platform on Dell Infrastructure, which may be found on the Dell Technologies Info Hub for Data Analytics.

We value your feedback

Dell Technologies and the authors of this document welcome your feedback on the solution and the solution documentation. Contact the Dell Technologies Solutions team by email.

Authors: Dell Technologies Integrated Solutions Engineering, Technical Marketing, and Information Design & Development teams

NOTE: For links to additional documentation for this solution, see the Dell Technologies Info Hub for Data Analytics.

This document may contain language from third-party content that is not under Dell's control and is not consistent with Dell's current guidelines for Dell's own content. When such third-party content is updated by the relevant third parties, this document will be revised accordingly.