Deloitte.

The Data Lake Journey

Cloudera Data Platform

Executive summary

Data lakes have become a core component of many companies' IT infrastructure, keeping petabytes of raw business data in store. Modern analytics-driven use cases have outgrown existing infrastructure and demand flexibility, agility, and elasticity. Classical data lakes, commonly based on Apache Hadoop technology and deployed in companies' data centres, are struggling to fulfil these demands and thus require a reinvention.

This paper explores the multiple options to advance existing Cloudera and Hortonworks data lakes with the capabilities of the Cloudera Data Platform (CDP). With the release of CDP, Cloudera announced the discontinuation of the legacy Cloudera (CDH) and Hortonworks (HDP) distributions (CDH 5.x and HDP 2.x – end of 2020, CDH 6.x and HDP 3.x - Q1 of 2022). The paper discusses three scenarios which help clients succeed in their move to CDP:

- **Migrate to CDP Public Cloud**
  If the legal and technical requirements for an available public cloud environment are met, CDP Public Cloud is the most most flexible, agile, and elastic option available to users. Data is securely stored in object stores of cloud providers; data science environments and analytical engines (such as Apache Spark) can flexibly scale while SDX provides a common security layer across multiple clouds.

- **Upgrade to CDP Data Center**
  If clients are not able to leverage a cloud environment and would like to re-use their existing on-premises infrastructure, the upgrade to CDP Data Center with the converged Cloudera Runtime software distribution is the preferred way. Cloudera Runtime combines open source projects from both distributions (CDH and HDP), but also replaces and discontinues specific projects. A detailed analysis of the current platform tools in use is required to minimize the downtime of business-critical applications prior to migration.

- **Migrate to CDP Data Center**
  A migration to a new on-premises infrastructure is justified by three reasons: infrastructure is outdated, an appliance is used, or when a minimal impact to existing Big Data & AI workloads running on the client's Hadoop platform is required. This scenario allows clients to enhance their data lake with object storage, a separation of storage and compute, and containerised compute environments.

Deloitte recommends to follow an incremental migration strategy for Big Data & AI workloads – starting with a detailed analysis and prioritisation of use cases. The analysis identifies workloads using deprecated and discontinued open source projects. Deloitte offers standardised migration paths for these workloads and supports clients to prepare for the migration. The detailed migration path to CDP is dependent on the infrastructure and organisational preconditions, as well as the future vision and scope of the respective client. Deloitte supports clients in the development of a tailored vision and roadmap for their future Big Data & AI workloads, thus optimising the data architecture during migration.

“Cloudera enables our clients to move their Big Data & AI workloads to the new containerized world with the ability to execute and scale across multiple environments without any business downtime.”

Sandra C. Bauer
Data & Analytics Modernization Lead, Deloitte Germany

Current Situation and Objectives

With the increase of modern analytics-driven use cases, users are demanding flexibility, agility, and elasticity for Artificial Intelligence workloads. With the Cloudera Data Platform (CDP), Cloudera offers a solution that can serve the end-to-end data lifecycle in a hybrid setup with support for public cloud, multi-cloud, and on-premises deployments.

Data lakes were implemented in IT infrastructures to foster the digital transformation and development of new digital products. These data lakes have evolved over the last decade – from centralised data storages to data analytics platforms hosting critical business applications. The historical centralised storage solutions allowed the generation of new insights from structured and unstructured data being combined out of multiple source systems. Apache Hadoop became the de facto standard technology for data lake implementations. Cloudera (formerly Cloudera and Hortonworks), the main distributor of Apache Hadoop technology, enabled the deployment of large scale Hadoop clusters in data centers to ingest and process Big Data. With the integration of visualisation and machine learning capabilities it was possible to generate new business insights. The close integration formed complex data analytics platforms that were mainly hosted on bare-metal hardware combining compute and storage. These modernised data lakes created two main challenges:

- **Operability & Maintainability**
  Clients underestimated the effort of keeping a data lake up-to-date and well governed. The integration of multiple components and software packages created dependencies between the different sofware versions, and with the increasing number of implemented use cases, the number of specific requirements torwards the data lakes caused interlocks. This complexity caused high efforts in the overall management of data lakes.

- **Agility & Flexibility**
  One of the key advantages of data lakes is the agility and flexibility to transform the required data for new use case requirements. With the concept of combining storage and compute in one platform, use cases are locked to one specific compute environment. This dependency on the existing environments with defined programming languages, packages and resources prevent clients from trying new ways of solving problems and curb the agility and flexibility required for the implementation of new AI use cases.

Modern analytics-driven use cases have outgrown existing infrastructures and demand flexilibility, agility, and elasticity in all aspects. The required data must be accessible on-demand and use cases must be able to consume data in various ways (e.g. real-time or batch) and structures (e.g. raw, cleaned, or curated). The compute environments require the possibility to execute multiple programming languages, the flexibility to integrate various packages, and the elasticity to provide tailored resources at any time. AI workload requires these customised environments to deliver the best results. These requirements towards storage and compute resources led to major architectural and technological shifts in the design of data lakes:

- **Cloud Integration**
  The deployment of data lakes in the cloud reduces the provisioning and scalability efforts drastically. Tedious hardware orders and high capital expenditures are replaced by on-demand resource usage and a consumption-based pricing model. These improvements increase the flexibility and elasticity of data lakes. However, only modern data platforms with containerisation and independent services, as well as a separated compute and storage, can efficiently leverage the advantages of the cloud.

- **Containerisation**
  The containerisation of applications and operating environments is the new standard when designing agile IT solutions. Data lakes are adopting these principles to remove dependencies between different monolithic services. Major services are containerised to allow multiple provisonings of different service versions to improve agility and multi-tenancy capabilities.

- **Compute & Storage**
  In modern data architectures compute and storage capabilities are separated to increase the flexibility of both. The separation allows tailoring of both capabilities to the specific requirements and improves the efficient usage of the provisioned resources.

Companies are recognising these major architectural and technological shifts and are demanding these optimisations for their existing Hadoop distributions. Existing data center architectures, based on Cloudera or Hortonworks technology, are not able to deliver these capabilities and therefore require a reinvention and the integration of the new concepts and technologies.

Modern analytics-driven use cases have outgrown existing infrastructures and demand flexibility, agility, and elasticity in all aspects.

Journey to Cloudera Data Platform

Cloudera Data Platform
CDP combines the best of Hortonworks' and Cloudera's technologies and delivers self-service analytics across hybrid and multi-cloud environments, along with granular security and governance policies. Initially delivered as a public cloud service and followed up with a data center version, a data management and analytics platform for on-premises IT environments, CDP includes:

- Data Warehouse (DW) and Machine Learning (ML) services, as well as a Data Hub (DH) service for building custom business applications
- A unified control plane to manage infrastructure, data, and analytic workloads across hybrid and multi-cloud environments
- Consistent data security, governance and control that safeguards data privacy, regulatory compliance, and prevents cybersecurity attacks across environments
- 100 percent open source, to avoid vendor lock-in and accelerate enterprise innovation

A hybrid cloud approach is developing to become the de facto standard way of integrating the cloud into enterprise data architectures. Cloudera is following this way and developed the Cloudera Data Platform – an enterprise data cloud which manages and secures the data lifecycle across the cloud and data center.

In CDP, the hybrid and multi-cloud approach is enabled by Cloudera's Shared Data Experience (SDX) and Control Plane. Cloudera SDX provides an enterprise-wide data security and governance fabric. SDX enables data and metadata security and governance policies to be set once and automatically enforced across the data lifecycle. The Cloudera Control Plane manages, monitors, and orchestrates all CDP services with consistent security and governance. Consisting of Workload Manager, Replication Manager, Data Catalogue, and Management Console, the Control Plane delivers a powerful set of tools that provide data management, workload analysis, data movement, and data discovery capabilities. These capabilities can be used for intelligent workload migration and bursting of workloads to the cloud.

With the release of CDP, Cloudera announced the discontinuation of the legacy Cloudera and Hortonworks distributions. End of service will be approximately end of 2020 for CDH 5.x and HDP 2.x and in the first half of 2022 for CDH 6.x and HDP 3.x. These deadlines require clients to innovate their existing data lakes within the next months. Cloudera has set up their platform with two different form factors: public cloud and on-premises. The on-premises offering consists of CDP Data Center and CDP Private Cloud. CDP Private Cloud is announced for mid-2020 and will enable the clients to execute containerised workloads on-premises. The currently available CDP releases allow clients to innovate their data lakes in three ways:

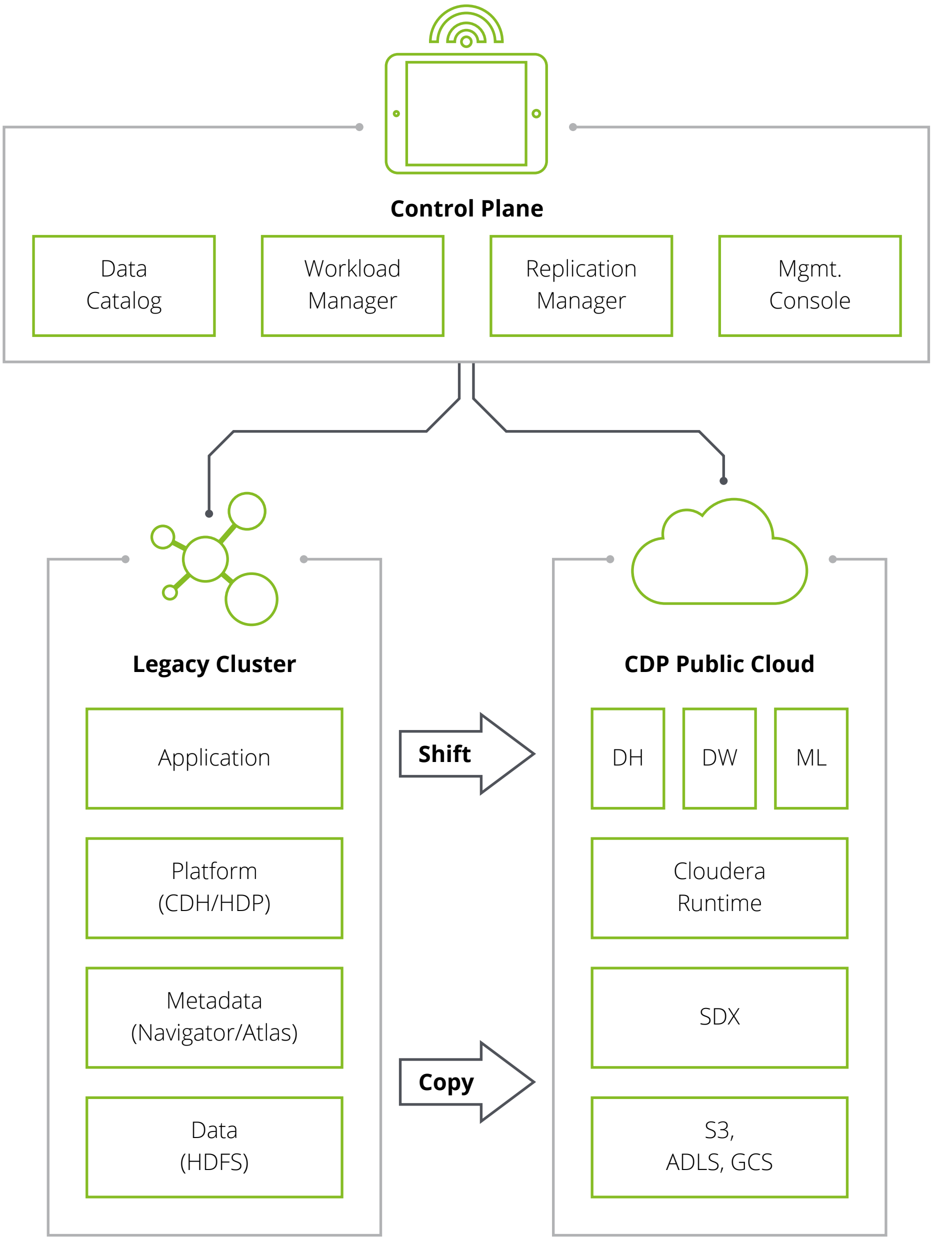
- **Migrate to CDP Public Cloud**
  Copy data and metadata to the public cloud; implement new or migrate existing workloads on CDP Public Cloud.
- **Upgrade to CDP Data Center**
  Upgrade from a classical cluster to CDP Data Center on the same hardware infrastructure.
- **Migrate to CDP Data Center**
  Build a new CDP Data Center cluster on-premises; copy data and metadata from existing classical clusters and migrate existing workloads.

Combinations of the three scenarios are not separately discussed. A hybrid approach with an on-premises and public cloud platform would require a migration to CDP Public Cloud and an upgrade (or migration) to CDP Data Center. The following chapters analyse the three different scenarios in more detail, highlighting the critical success factors and providing the best practices on how to move forward.

Migrate to CDP Public Cloud

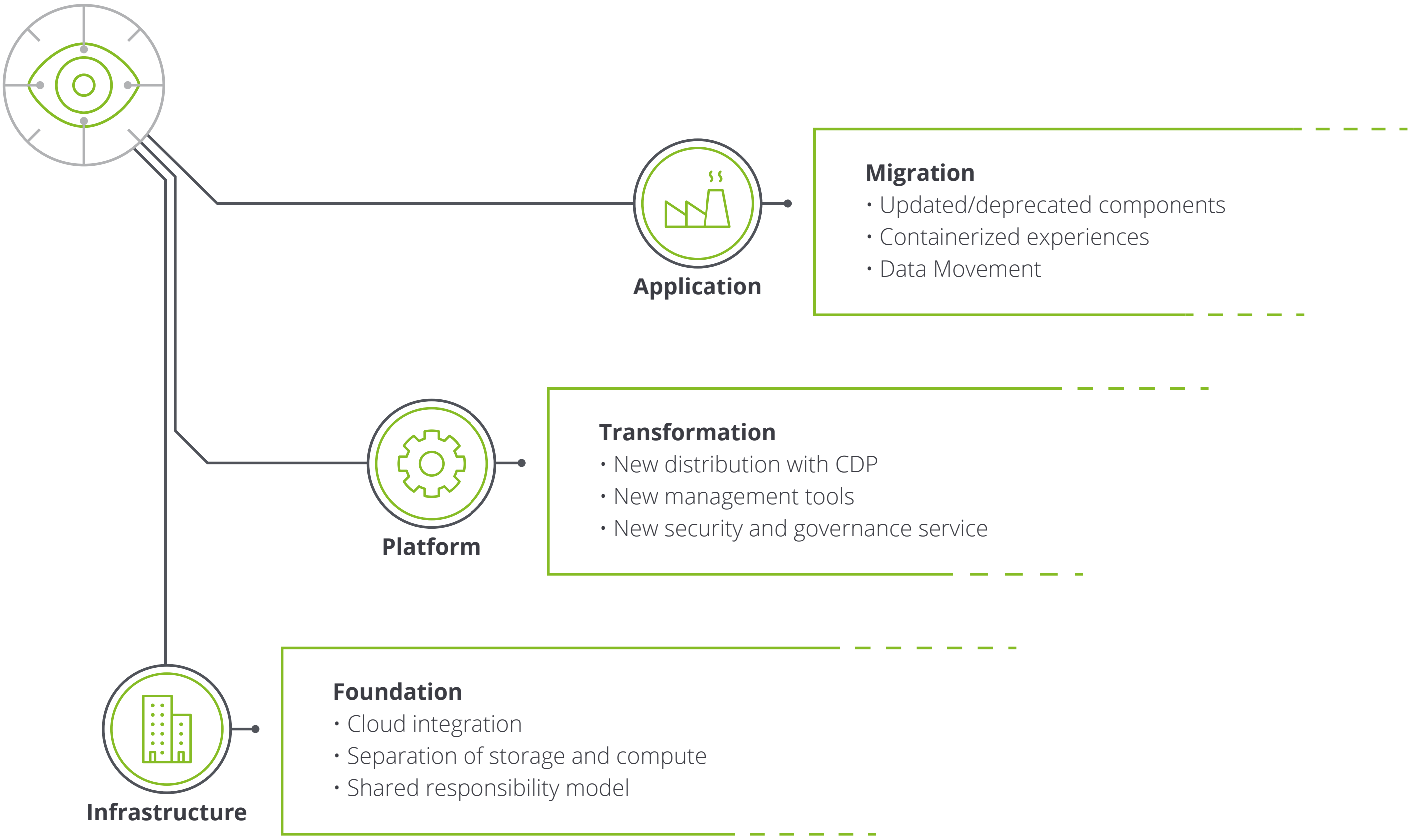
The migration to CDP Public Cloud enables Cloudera clients to create and manage secure data lakes, self-service analytics, and machine learning services in the Cloud without installing and managing the data platform software. CDP Public Cloud services are managed by Cloudera. Cloudera offers a Control Plane that is connected to the client cloud environment and provisions the required resources. The data itself always remains under the control of the client and is not transferred out of the cloud environment. This migration path implements CDP in the cloud environment of the client and allows clients to shift workloads and copy data incrementally to the cloud.

Fig. 1 – Migration to CDP Public Cloud Overview



The journey to CDP Public Cloud requires the consideration of three different layers: infrastructure, platform and application.

Fig. 2 – Migration to CDP Public Cloud Layers



Application
Migration
• Updated/deprecated components
• Containerized experiences
• Data Movement
Platform
Transformation
• New distribution with CDP
• New management tools
• New security and governance service
Infrastructure
Foundation
• Cloud integration
• Separation of storage and compute
• Shared responsibility model

The infrastructure builds the foundation of CDP and requires the complete rethinking of existing concepts and guidelines. Cloud infrastructure is available on-demand and is able to scale up or down according to the use case requirements. To enable these capabilities, the Cloudera management console, which is used to manage environments, users, and services, requires an integration with the client's cloud environment. The specific integration steps differ per cloud provider, but have in common that in the Cloudera management console a cloud environment must be registered. This requires an account which must have adequate permissions to configure resources and services. With the successful registration, a data lake is automatically provisioned. Through the separation of storage and compute resources, Cloudera allows the provisioning of data lakes without a detailed sizing exercise in the cloud. All provisioned environments can thus start small and grow based on number of users and use case requirements. In consequence of the integration of multiple components (Cloudera CDP and cloud resources), it is required to analyse and review the different responsibilities between Cloudera, the cloud provider, and the client. The development and understanding of the shared responsibility model help to address and resolve upcoming issues properly.

The platform layer is redeveloped by Cloudera and integrates new concepts and components. The following concepts are key to understand the platform:

- An environment is a logical subset of your cloud provider account, including a specific virtual network.
- Data lakes provide a mechanism for storing, accessing, organising, securing, and managing data in cloud object storage (e.g. Amazon S3, Azure Data Lake Storage, Google Cloud Storage) or HDFS.
- Data hub clusters are pre-defined blueprint clusters (e.g. data flow, streaming and operational database) that are provisioned on virtual machines. The clusters are easy to launch and their lifecycle can be fully automated. With data hubs, each team can now have their own private cluster.
- Experiences are containerized compute environments providing specific functionalities. Cloudera released a Data Engineering, Data Warehouse, and Machine Learning experiences.

Apart from these new concepts, Cloudera published a converged distribution of CDH and HDP (Cloudera Runtime 7) with new management services. Former Cloudera clients need to get familiar with Apache Ranger to define, administer, and manage security policies consistently across components. Former Hortonworks clients require knowledge in Cloudera Manager to administer data lakes and data hub clusters. Additionally, Cloudera offers a data catalogue (searching, organizing, securing, and governing data), a workload manager (for analysing and optimising workloads), and a replication manager (for copying, migrating, snapshotting, and restoring data between environments). SDX and Control Plane are the collection of these new tools and provide clients with consistent security, governance, and data lineage across all environments. Existing on-premises concepts and usage guidelines must be reviewed and transformed to leverage these new capabilities of CDP.

The migration of applications requires a detailed analysis of the involved on-premises components and data sets. The Cloudera Runtime upgrades components, removes deprecated components, and adds new components. Existing applications using retired components require a redesign. This is required for workloads using Apache Flume (migration to Apache NiFi), Apache Pig (migration to Apache Spark), Sqoop2 (migration to Sqoop1) and Apache Storm (migration to Apache Flink). Besides these changes, a detailed application analysis helps to enhance existing workloads with available improvements of the new compute resources (data hub clusters and Cloudera Experiences). The data hub clusters allow to migrate applications to CDP without major changes. Cloudera Experiences enable clients to leverage the benefits of containerised compute resources which can be provisioned, scaled, and terminated within seconds.

The migration to CDP Public Cloud should be based on an incremental plan with a parallel operation of the legacy cluster and CDP Public Cloud. Following this incremental migration, clients are able to migrate their workloads and data sets by using Control Plane capabilities ("Burst to Cloud" with replication manager).

Shared Data Experience
SDX provides consistent data security, governance and control. Policies defined once and consistently are applied across the platform. SDX also provides the migration and replication capabilities so that, should data need to move, associated security and governance policies stay connected. This, in and of itself, delivers unique and advanced platform capabilities such as intelligent migration between infrastructures and bursting workloads to the cloud from on-premises. SDX combines functionalities from Ranger, Atlas, Knox, Hive Metastore, Data Catalog, Replication Manager, and Workload Manager. The newest version extends SDX governance capabilities to support models.

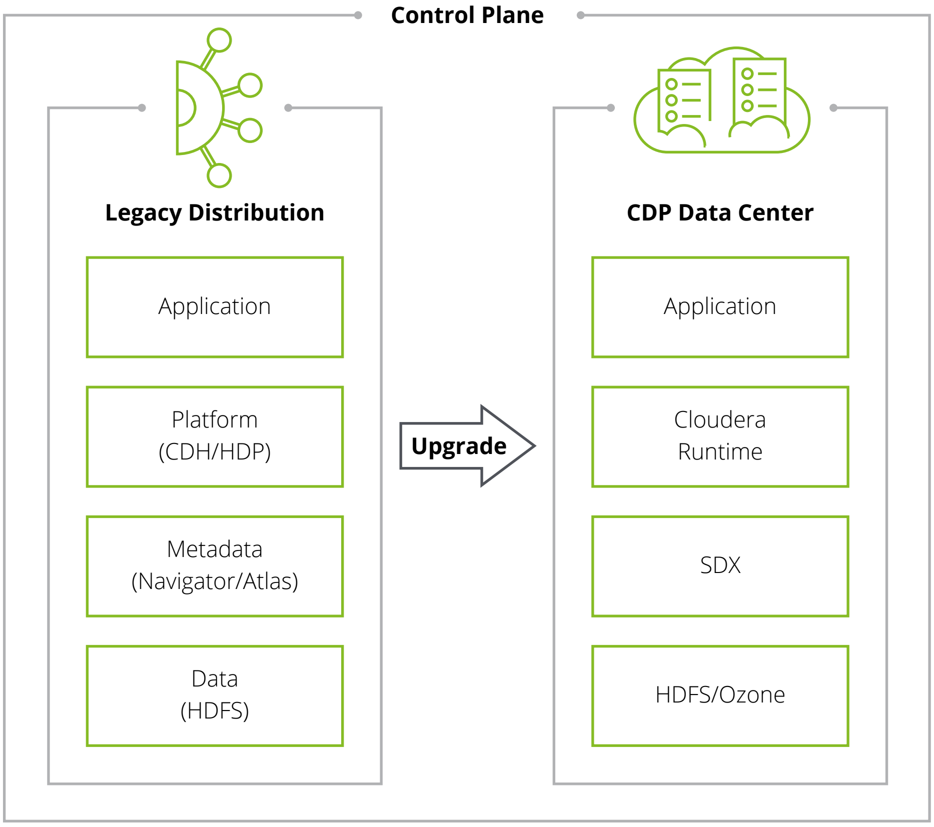
Upgrade to CDP Data Center

The CDP Data Center upgrade path is ideal for clients currently operating existing CDH or HDP on-premises clusters and who wish to leverage the capabilities of CDP Data Center on the same existing infrastructure. In comparison with the other scenarios mentioned in this whitepaper, the upgrade path is preferred for situations where clients are not capable of leveraging cloud technologies (e.g. legal requirements), face restrictions on the expansion of their on-premises infrastructure, or do not require any additional infrastructure resources to fulfil their business goals.

CDP Data Center – Upgrade Timeline

It is recommended for clients with either CDH or HDP clusters to wait for the release of CDP Data Center 7.1 (released in the first half of 2020) and upgrade directly to this version. In addition to platform functionality and stability improvements, the release will natively support upgrading CDH 5.14-16 and HDP 2.6.5 clusters to CDP Data Center 7.1. This will be the only officially supported direct upgrade path. Further in the second half of the year, CDP Data Center 7.2 is expected to include a direct upgrade path for CDH 6 and HDP 3 clusters.

Fig. 3 – Upgrade to CDP Data Center



The upgrade to CDP Data Center also requires taking into consideration the infrastructure, platform, and application layers. From an infrastructure perspective, one of the most significant changes introduced for CDP Data Center is that the platform only supports servers with CentOS or Red Hat Enterprise Linux (RHEL) 7.6 as operating system. CDH and HDP clusters on SUSE Linux Enterprise Server (SLES), Ubuntu, Debian, or Oracle Linux operating systems are currently not eligible for a direct upgrade to CDP Data Center. Clients with CDH or HDP clusters running on operating systems other than CentOS or Red Hat Enterprise Linux (RHEL) are recommended to opt for a migration to CDP Data Center instead. Additionally, system owners must pay attention to the platform's database and Java requirements when planning for an upgrade, as these may differ from the older versions previously supported. Although hardware requirements remain practically the same for CDP Data Center, companies are nonetheless recommended to evaluate their individual setup as part of the upgrade preparation procedure.

The underlying platform of CDP Data Center is the Cloudera Runtime 7 distribution – the same converged stack of the Cloudera and Hortonworks Hadoop distributions which is used in the CDP Public Cloud. As a result of this convergence, however, some individual features from both platforms are no longer supported or have been replaced by new ones. In comparison to a one-by-one application migration to the CDP Public Cloud, clients that are upgrading their on-premises platform to CDP Data Center need to analyse these component changes in detail. Clients with CDH clusters will need to get familiar with Apache Ranger, as Apache Sentry is no longer available in the new distribution. In order to support with this transition, Cloudera includes a Sentry-to-Ranger policy migration tool as part of CDP Data Center 7.1. Cloudera Navigator has also been replaced by Atlas on CDP Data Center. Cloudera Navigator lineage data is transferred to Atlas as part of the CDH to CDP Data Center upgrade process. Audit data is, however, not transferred from Cloudera Navigator to Atlas. One of the major changes for HDP clusters is that next to Apache Ambari now Cloudera Manager can be used. Clients might need to adapt their cluster administration practices within their IT teams as part of this process. Another major change in CDP Data Center is that clients will be able to leverage Ozone as the storage layer of their applications. Compared to HDFS, Ozone shows improvements mainly in the handling of small files and cluster storage scaling. Ozone is currently available as a Tech Preview in CDP Data Center and will be fully available in the upcoming releases. The new version of HDFS, HDFS 3, supports erasure coding when dealing with storage fault tolerance. HDFS 3 will be able to reduce the 200 percent storage overhead present on HDFS in CDH and HDP clusters to approximately 50 percent overhead – this while keeping same level of fault tolerance.

A detailed analysis of the client's applications is a mandatory step during the planning phase of any upgrade to CDP Data Center. As the Cloudera Runtime 7 distribution incorporates changes to the Hadoop tool stack compared to CDH or HDP clusters, existing applications, which are dependent on the changed components, may require adaptations or even a complete redesign. Most notably, existing workloads using the services listed below will require considerable changes in order to keep operating:

- Apache Flume ❯ applications will need to be adapted to Cloudera Flow Management (CFM) or another solution.
- Apache Pig ❯ applications will need to be adapted to Hive, Spark, or another solution.
- Spark 1.6 ❯ application code will need to be adapted to Spark 2.
- Apache Storm ❯ depending on the use case, applications will need to be migrated to Flink, Spark Streaming, or Kafka streams.
- Sqoop 2 ❯ applications will need to be adapted to Sqoop 1.

Clients should furthermore analyse how their existing applications and workloads can be improved using the features introduced by CDP Data Center. Traditional workloads, which are fully supported by CDP Data Center, stand to benefit in terms of functionality and performance when adequately re-architected to take full advantage of the Cloudera Runtime 7 distribution tool stack. For example, clients upgrading a CDH cluster may use Hive 3 to improve on enterprise data warehouse (EDW) use cases and leverage ACID support; Hive on Tez can be used for better ETL performance. Similarly, clients upgrading an HDP cluster may start using Apache Impala to enable rapid interactive querying on the platform, or begin using Apache Kudu for low-latency time series data ingestion and analytics with ACID semantics.

As for the actual upgrade procedure, it is recommended as a best practice to first perform the upgrade on a development environment. This will allow the involved use case teams to test the overall functionality of the cluster, of the integrations and applications, and to evaluate the security of the new installation. Once all necessary refactoring, enhancements, and optimisations have been identified and worked upon, then the upgrade procedure can take place sequentially on further test, pre-production, and finally production clusters. By learning from the issues detected on clusters with non-critical workloads, this upgrade plan ensures that the least possible downtime and business impact supervene the upgrade to CDP Data Center.

Apache Ozone

HDFS has been the de facto file system for Big Data. HDFS works best when the files are large, but suffers from small files. Ozone is a distributed key-value store that can manage both small and large files alike. While HDFS provides POSIX-like semantics, Ozone looks and behaves like an Object Store. Ozone is made up of volumes, buckets and keys. Volumes are similar to accounts and are created/deleted by administrators. A volume can contain zero or more buckets. Buckets are comparable with cloud buckets. Keys are unique within a given bucket and are similar to cloud objects.

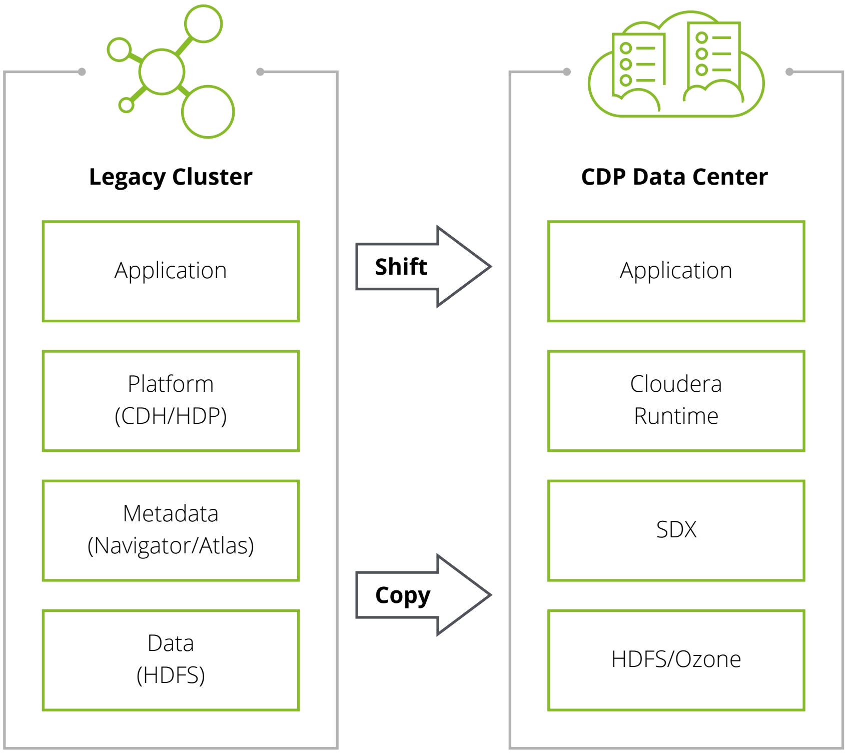
Migrate to CDP Data Center

The migration to CDP Data Center implies the setup of a new on-premises CDP Data Center cluster next to the existing Cloudera or Hortonworks platform. The upgrade to CDP Data Center is the preferred solution to move an on-premises cluster to CDP, but is not applicable to all situations. The following situations can justify a new setup:

- The existing server infrastructure is outdated and requires a renewal.
- Oracle Big Data Appliance or Teradata Appliance for Hadoop is being used and the appliance model is not compatible with CDP Data Center.
- A minimal impact to the existing cluster and workloads is required.

For clients facing one of these circumstances, a new setup of the cluster must be validated with a cost-benefit analysis. It is recommended to compare the occurring infrastructure costs with the newly available platform features and future use case revenues.

Fig. 4 – Migration to CDP Data Center



Flexibility and speed to adopt changes are two key success criteria in today's business world.

The setup of a new CDP Data Center requires the same consideration of three layers: infrastructure, platform and application. The infrastructure setup of CDP Data Center requires the same approach as previously provisioned CDH and HDP clusters. Available performance upgrades in the Cloudera Runtime 7 distribution improve the resource utilisation, but the new available capabilities will enable the extension of existing use cases to include additional features. So, if clients are migrating the full platform with all existing applications, the best approach is to provision the same infrastructure resources for CDP Data Center. Additionally, the storage design must be validated and take into account the availability of Apache Ozone. To leverage the full performance of Ozone, Solid-State-Drives (SSDs) are recommended and must be considered in the hardware sizing.

The platform migration to CDP Data Center relies on the migration to Cloudera Runtime and the same aspects as in the previous explained paths must be considered. Cloudera clients need to consider the change from Apache Sentry to Apache Ranger and Hortonworks clients the integration of Cloudera Manager. However, there are also other possibilities for improvements with the new deployment of Cloudera Data Center. Existing multitenancy platforms based on CDH or HDP might be able to leverage the new concept of virtual private clusters to separate storage and compute resources. An administrator can create multiple, isolated, compute-only clusters that each point to one data repository, one data catalogue, and one set of security policies. To achieve this, Cloudera virtual private clusters rely on the logical separation of compute services from base services. Multiple compute clusters can access the services in a base cluster through a new concept called the Data Context. A Data Context is the grouping of pointers to the base cluster services. To further improve this separation of storage and compute, Cloudera announced the release of CDP Private Cloud. The CDP Private Cloud is fully integrated with CDP Data Center and SDX provides consistent data security, governance, and control across both on-premises offerings. CDP Data Center is the building block of a CDP Private Cloud deployment. It contains all the stateful aspects of the deployment: the data, the metadata, the security, the governance, and the infrastructure. A direct upgrade or migration to CDP Private Cloud without CDP Data Center is not possible, because CDP Private Cloud requires CDP Data Center as storage layer. CDP Private Cloud allows the client to execute their machine learning workloads in a compute-optimised and containerised environment based on Red Hat OpenShift.

Clients applying platform changes with virtual private clusters and in the future with CDP Private Cloud, must adapt their applications to the new concepts and services. The virtual private clusters can be separated by user/team, workload type, or workload priority. The application assigned to a specific virtual private cluster needs to be encapsulated from others and needs to be stateless. The same applies for an application running on the future CDP Private Cloud. CDP Data Center provides the stateful elements for the new containerised applications: storage, table schema, authentication & authorization, and governance. Existing applications can be adapted to this containerised design to gain agility, flexibility, and elasticity. Beside these optimisations, workloads must be adapted to the converged Cloudera Runtime. As previously described, applications using Apache Flume, Apache Pig, Spark 1.6, Apache Storm, or Sqoop2 must be redesigned.

In comparison to the other scenarios, the migration to CDP Data Center requires the largest investment in infrastructure. If the legacy CDH or HDP cluster nodes can be provisioned in the new CDP Data Center cluster, a rolling migration with just a couple of new nodes is possible. Starting with a gradual migration of data and workloads to a small CDP Data Center cluster. As workloads move out from the legacy cluster, servers can be decommissioned and then added to the new CDP Data Center cluster. Following this approach allows the repurposing of the old hardware and to minimize the risk of downtimes for business-critical applications.

CDP Private Cloud

CDP Private Cloud is the second form factor of Cloudera's on-premises offering. CDP Private Cloud enables clients to execute the Cloudera Experiences, originally developed for the Public Cloud, and is developed as the side-car to the base CDP Data Center. SDX provides the consistent security, governance, and control across both environments and allows clients to seamlessly execute containerised workloads on-premises. With the first release in mid-2020, Cloudera plans to support the data warehouse and machine learning experiences.

Getting started the right way

The first step in the journey to Cloudera Data Platform should be to decide on one of the three scenarios discussed in this whitepaper. Whether that is Migrate to CDP Public Cloud, Upgrade to CDP Data Center, or Migrate to CDP Data Center, the chosen scenario should fit best to your organisation's current data lake setup and future requirements. For large organisations with varied workloads, and who therefore seek to implement a hybrid strategy for their platforms, a combination of the three approaches may need to be planned and undertaken. The decision must be taken tailored to every organisation based on its individual strategy, goals, situtation and requirements. With the approach at hand, it is then possible to begin planning the next steps with a clear target state in sight. Revolving around the technology, financial, organisational, and vision perspectives, Deloitte recommends having the following considerations in mind to establish a successful foundation for the migration.

Technical

Financial

Organisational

Vision & Scope

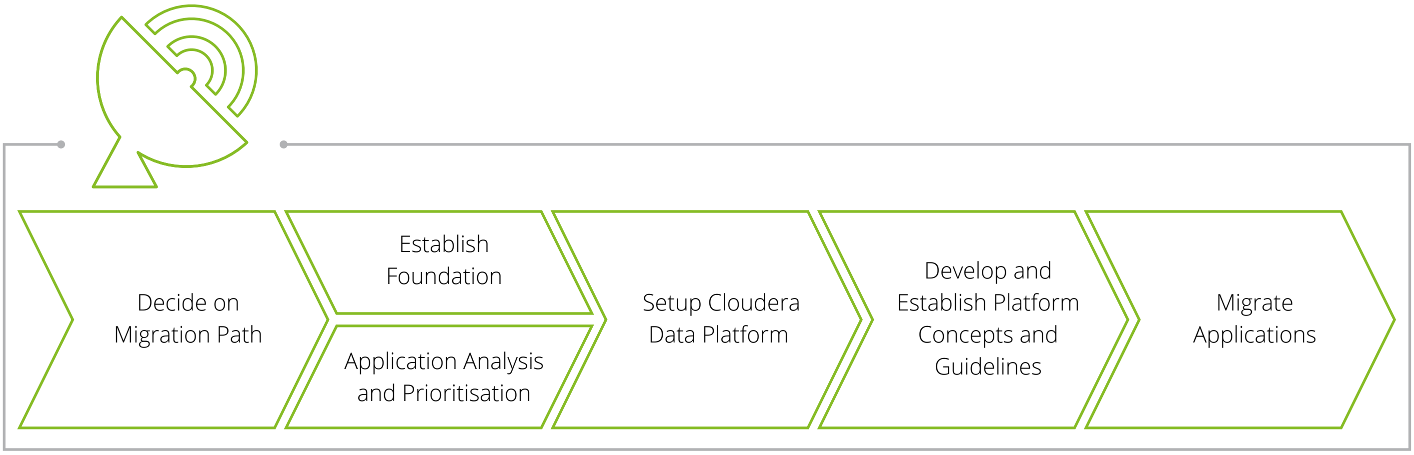
Before any upgrade or migration can take place, your team needs to have a full technical understanding of the existing data platform. Given the complexity behind all the tools in the Hadoop ecosystem and their interactions, the effort behind a complete technical analysis should not be underestimated. The pain points of the existing platform should be evaluated and quantified. This with the goal to detect possible improvements to these pain points, which can be introduced as part of the upgrade or migration process. Dependency analysis should also not be limited to the applications' interactions with the platform. Other data platforms in the organisation could depend on the data lake as well as part of their data lifecycle. Key data assets, configurations, and policies should also be backed up as a standard procedure before any major platform and/or infrastructure change.

The evaluation of the financial ramifications of any upgrade or migration is a necessary prerequisite for any successful CDP implementation. While most organisations focus mainly on getting the technical aspects right, the new and varied cost structure behind CDP takes many IT departments by surprise. Platform owner should look out for new and previously unaccounted cost sources on all levels. For example, a public cloud migration will introduce a cost structure previously not present in an on-premises setup. The migration to CDP Public Cloud will change it from a capital expenditure to an operational expenditure. This will enable platform owners to attach costs to single workloads and allow them to bill users individually. On the other hand, the complex pricing structure of public cloud providers could detonate the platform's operational costs if they are not kept fully under control. Cloudera has also introduced a set of changes in its licensing structure, which might be new for some clients.

The transition to CDP is unlike any other CDH or HDP upgrade performed in the past. Existing IT teams will need to develop a solid understanding of the platform in order to ensure a successful implementation. Even existing teams, with extensive Hadoop experience, might not be fully prepared for the transition to CDP. Additional time and budget need to be taken into account to cover for this learning and adaptation period.

Finally, in order to make the most out of the journey to CDP, we recommend a clear vision for the platform. Its defined scope within the organisation should be fully in line with the company's objectives. The best way to achieve these goals is by establishing an active collaboration between the business and IT departments. The new functional possibilities which CDP provides can pave the path into unlocking new areas of untapped business value. Conserving this collaboration beyond the platform's migration period is a key characteristic of a data-driven enterprise.

Fig.5 – High-Level Migration Plan


Decide on Migration Path
Establish Foundation
Application Analysis and Prioritisation
Setup Cloudera Data Platform
Develop and Establish Platform Concepts and Guidelines
Migrate Applications

Every migration plan should involve the following high-level phases:

1. Decide on the Migration Path: select and develop your customized CDP migration path.

2. Establish the Foundation: establish a common understanding and future state of technical aspects, financial aspects, organisational aspects, vision and scope of the future data platform.

3. Application Analysis and Prioritisation: understand the current established use cases and analyse required changes. Prioritize the use cases based on their business value and evaluate the migration sequence.

4. Set Up Cloudera Data Platform: install and configure CDP On-premises or Public Coud.

5. Develop and Establish Platform Concepts and Guidelines: validate existing concepts and usage guidelines. Adjust them to new available capabilities and tools.

6. Migrate Applications: Incrementally migrate applications from the legacy cluster to the new CDP.

Deloitte and Cloudera work together to jointly enhance the client's data lakes. Cloudera is offering guidance to clients with the CDP Journey Advisor, giving a general overview of the new changes, and with "MyCloudera.com" providing clients with cluster-specific guidance. Additionally, Cloudera offers workshops to develop a customized migration plan. Deloitte is supporting clients with the end-to-end knowledge, offering guidance ranging from the platform strategy to the implementation of application migrations.

Contacts

Please feel free to contact us if you need any further information and guidance.

Sandra C. Bauer
Partner | Analytics & Cognitive
Market Offering Lead Data Modernization
Tel: +49 (0)173 3443105
sabauer@deloitte.de

Fabian Hefner
Senior Manager | Analytics & Cognitive
Tel: +49 (0)151 58004595
fhefner@deloitte.de

Christoph Kolberg
Manager | Analytics & Cognitive
Tel: +49 (0)151 58073740
ckolberg@deloitte.de

Fernando Trejo
Technical Expert | Analytics & Cognitive
Tel: +49 (0)151 58074369
fetrejo@deloitte.de

Deloitte.

This communication contains general information only not suitable for addressing the particular circumstances of any individual case and is not intended to be used as a basis for commercial decisions or decisions of any other kind. None of Deloitte Consulting GmbH or Deloitte Touche Tohmatsu Limited, its member firms, or their related entities (collectively, the "Deloitte network") is, by means of this communication, rendering professional advice or services. No entity in the Deloitte network shall be responsible for any loss whatsoever sustained by any person who relies on this communication.

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. Please see www.deloitte.com/de/UeberUns for a more detailed description of DTTL and its member firms.

Deloitte provides audit, risk advisory, tax, financial advisory and consulting services to public and private clients spanning multiple industries; legal advisory services in Germany are provided by Deloitte Legal. With a globally connected network of member firms in more than 150 countries, Deloitte brings world-class capabilities and high-quality service to clients, delivering the insights they need to address their most complex business challenges. Deloitte's approximately 312,000 professionals are committed to making an impact that matters.

Issue 08/2020