The Cloud First journey with CDP

A position paper for current Cloudera and Hortonworks customers planning to become Cloud First

accenture

Abstract

COVID-19 has created a new inflection point that requires every company to dramatically accelerate the move to the cloud. It's the foundation required for digital transformation, enabling resilience, new experiences and products, trust, speed and structural cost reduction Given this, a number of enterprise companies have deployed Cloudera CDH or Hortonworks HDP Hadoop clusters in their on-premise data center environments. These clusters are being used for everything ranging from Data Backup to ETL offload to supporting Advanced Analytics to full-blown Data Warehousing/BI/Visualization. Recently, cloud computing has become the de facto platform for fueling digital transformations and modernizing IT portfolios. Clients are leveraging cloud for its ease of use, agility, flexibility and often cost savings. With increasing cloud adoption, the Big Data workloads built on the on-premise Hadoop clusters need to be migrated to the cloud. It will be challenging for our clients to move all this on-premise data infrastructure and solutions to the cloud, given the investments and value locked over several years. They need to adopt a more pragmatic and risk-optimized approach.

This paper provides a point of view and guidance around migration of these workloads to the cloud leveraging Cloudera's offering – Cloudera Data Platform or CDP. CDP provides a next generation multi-function, open and governed data platform that is suitable for deployment in any cloud. Leveraging CDP, businesses can expedite their "cloud first" journey with lower risk.

Situation overview

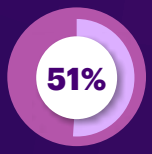
Companies today want speed, predictable outcomes and holistic business value and cloud is absolutely fundamental to their growth and success. Over the past few years, many organizations have been implementing Data Hubs and Data Lakes using Big Data technology such as Hadoop in their quest to become more data-driven and use insights from data for improving their business outcomes. Big Data allowed for acquisition and integration of large scale, structured and unstructured data from diverse sources that could then be used for analytics, BI, visualization and machine learning. Apache Hadoop became the key enabling technology for these Big Data environments. Cloudera and Hortonworks (now integrated company – Cloudera) were the main distributors of Hadoop. Clients today have large deployments of Hadoop clusters (CDH or HDP) in their data centers. Data is ingested into these Hadoop clusters using a variety of techniques – native Hadoop (Sqoop and Flume and NiFi) or third party tools like Informatica. Over the years, clients have developed ingestion pipelines to feed 10s of thousands of datasets from hundreds of sources into these clusters – both in batch and real-time. Data is typically stored in HDFS and then further data curation (cleansing, standardization, integration, etc.) is being performed on the cluster using tools such as MapReduce, Spark, Hive and Impala. Resulting datasets, are often stored in Hive databases and tables providing a standard SQL interface via HiveQL, Impala or LLAP. Numerous workloads have been developed over time to consume this data – Tableau and Qlik visualization and dashboards, reports, ad-hoc queries, machine learning models, etc. All of this is currently occurring on multiple clusters deployed in on-premise datacenters, largely over bare metal hardware (virtualized in some cases).

Growth in data, new workloads and increased consumption often requires additional servers to be procured or more storage to be added to the environment. This adds time, cost and additional burden on internal IT resources. Enter Public Cloud-

Public Cloud, such as Amazon AWS, Microsoft Azure and Google GCP, provides a range of infrastructure and platform services delivered "as-a-service" that frees up internal IT staff and makes provisioning of service much faster. In addition, Public Cloud provides the flexibility and scalability to provision what you need with the ability to modify it quickly as your needs change and pay-as-you-go. Cloud has become an urgent imperative technology for every enterprise. Migration of application workloads to Public Cloud has been well underway, but the migration of data-centric workloads to cloud is just beginning and at the nascent stage. Migrating data-centric workloads with PBs of data is indeed much more complex than migrating an application with a database. Our clients, who have large CDH or HDP clusters, are now looking to migrate this on-premise data-centric infrastructure and solution to Public Cloud to reap the benefits.

What are clients demanding?

Clients want to manage data across multiple deployment platforms: On-premise / Private Cloud, Public Cloud or Multi-cloud.



Clients want to either move these data-centric workloads to cloud or augment and extend these workloads leveraging cloud. Most enterprises procure cloud services from two or more vendors, a trend that is gaining traction in 2020 and hence support for cloud portability is important. A recent Harvard Business Review of 185 global executives across a wide range of industries finds that 51% of respondents plan to leverage multiple cloud providers as part of their data strategy.

Workloads running on on-premise clusters are either multi-function (data engineering, data warehousing, machine learning, operational data) today or if not, then clients want an integrated data fabric in the cloud to support these workloads.

Clients want the data platforms in the cloud to be secure and governed. Sensitive data may be kept in multiple environments: On-premise, Private Cloud or one or more Public Clouds. This data needs to be managed and controlled easily and consistently. Capabilities such as access control and security, metadata management, catalog and lineage are paramount.

Clients want to store their data on cloud in open source formats (such as ORC and Parquet) so that a variety of cloud native and third-party tools can be leveraged to process and analyze the data.

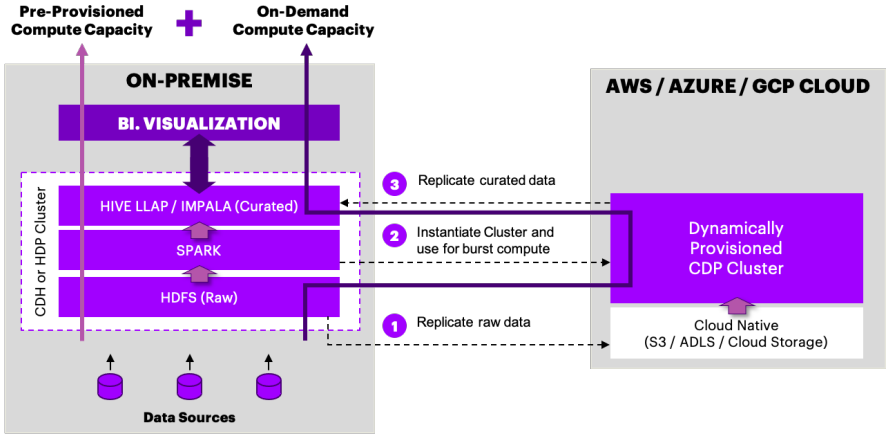
Clients are not going to migrate their entire data platform to cloud overnight. Data pipelines, data, schemas, workloads and security policies all have to be migrated. For a significant period of time they would operate in a hybrid mode. They need an approach which expedites this migration and yet reduces the risk.

Transformation scenarios

Scenario 1

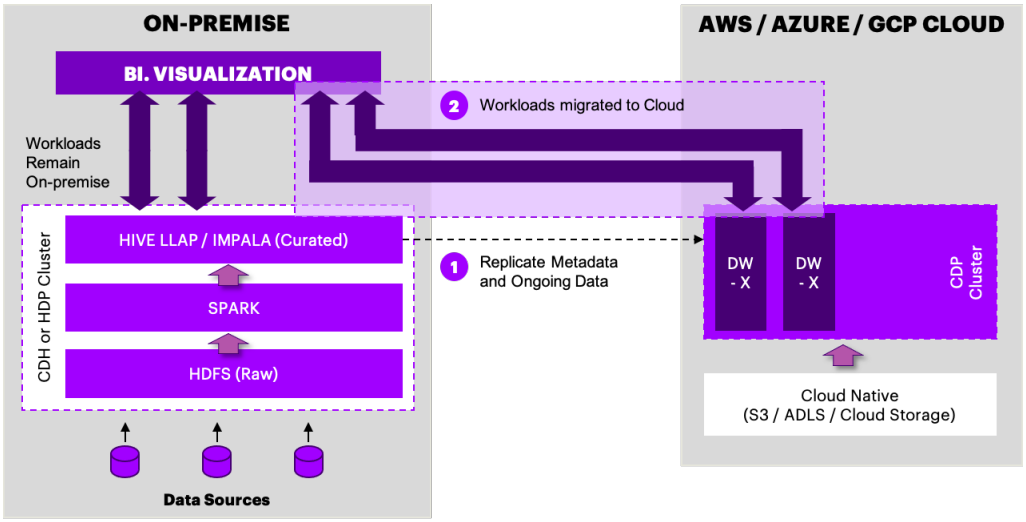
Burst current workloads to cloud

In this scenario, data engineering and data consumption workloads that are constrained by capacity on the on-premise clusters can temporarily take advantage of additional compute capacity in the cloud on an as needed basis. Let's assume that multiple data sources are being ingested into an on-premise CDH cluster in a batch mode on a nightly basis. CDC is implemented and hence only incremental changes (add, update, delete) transactions are being ingested into the cluster in a batch mode at 10 PM. This means that all add, update, deletes from the previous 24 hrs. are extracted from the source and ingested into the cluster. As part of curation activities in the cluster (Spark jobs), this change data is processed – cleansed, standardized, integrated – and a data mart that supports BI reporting is updated every day. Typical daily volume from these sources takes 8 hrs. to process and hence the updated reports from the data mart are refreshed every morning at 6 AM. The cluster has been sized to ensure the SLA, i.e. enough capacity is provisioned on the cluster such that processing can complete in 8 hrs. However, one source – say an Enterprise Campaign Management system – generates 10x the normal volume on certain days of the month when email campaigns are executed. Given the fixed size of the cluster, processing this data on the same cluster may now take 36 hrs. causing the 8 hr. SLAs to be violated. If the cluster is designed for peak load on these days, then tremendous capacity will be wasted. To temporarily burst and provision additional compute cycles requires spare servers. A natural benefit of Public Cloud is the infinite capacity available instantaneously and on demand. However, in order to burst the compute to the cloud, data associated with the compute also needs to be available in cloud. Cloudera CDP will enable this easily. CDP Replication Manager can be setup to constantly replicate select data to the cloud. In our example, daily data from the campaign management system can be automatically replicated to the cloud in a timely manner. When additional compute cycles are needed (on those certain days), CDP will instantiate a CDP cluster in public cloud and copy over all configuration and metadata (e.g., Hive tables definitions and access policies) to cloud cluster. Now additional Spark jobs will execute in the CDP cluster on the cloud, leveraging the additional compute capacity that was provisioned just-in-time. Data resulting from the Spark curation jobs can be replicated back to on-premise environment to support the BI / dashboards. Leveraging this CDP capability will allow the data platform to meet the business SLAs without expanding on-premise infrastructure.



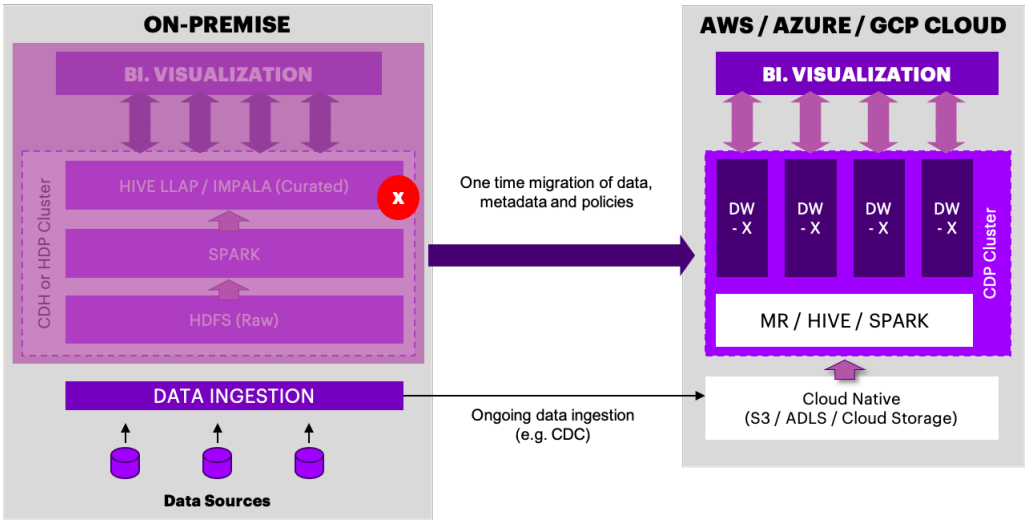
Migrate current workloads to cloud

In this scenario, select data consumption workloads (BI reports / dashboards / warehouse / marts) can be migrated to the cloud, reducing the footprint of the on-premise environment and starting the journey of migration to cloud. CDP Workload XM can help in identifying workloads and associated data that can be migrated. Workload XM will analyze the data from the on-premise cluster and for each workload provide two major outputs: a Capacity Plan (how much compute capacity is required in the cloud) and a Data Replication plan (what data and metadata needs to be replicated and when). The capacity plan can be used to launch the right-sized CDP cluster in the cloud. The replication plan is used to replicate all the metadata and policies (Hive databases, tables, security policies, etc.) as well as set up ongoing data replication from on-premise to the cloud cluster. For the example in the figure below: two workloads have been identified for migration to cloud. These workloads are BI/Visualization workloads and use two Hive databases that house the curated data. A CDP cluster is launched in the cloud with two separate DW-X (data warehouse experiences) – think two separate containers, one for each workload based on the capacity plan. Hive table metadata and security policies associated with the Hive databases is migrated over to the cloud CDP environment first. Next, the entire Hive database is replicated to the cloud. In addition, ongoing data replication is setup to automatically replicate the "select" data from the on-premise environment to CDP in cloud. This ensures that data acquisition and data curation activities can still be performed in the on-premise cluster, but the output data is replicated to the cloud CDP cluster. The BI and visualization reports / dashboards that were using the on-premise cluster can now be re-pointed to the cloud environment. This can be repeated for other workloads and slowly all consumption can happen from the cloud environment. This approach also allows different workloads to be migrated to different target cloud environments. Using CDP capabilities, we can migrate one workload to AWS and the other to GCP if desired and with the same ease. This scenario leveraging the capabilities of CDP will require no rewrite of code and provide an expedited migration path to cloud with low risk.



Migrate entire data ecosystem to cloud

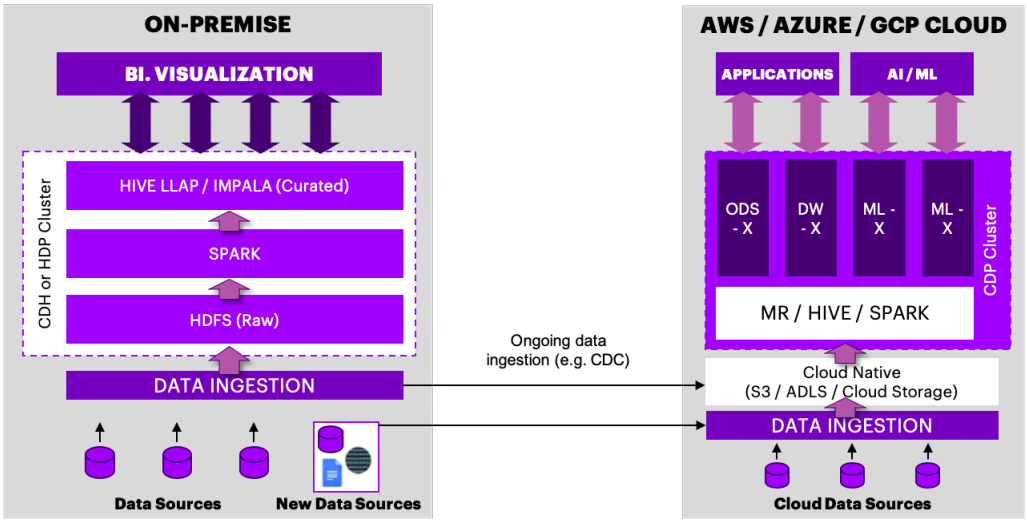
This scenario involves migrating the entire data supply chain of Capture, Curate and Consume to the cloud, so that the on-premise environment can be decommissioned. Using the approach in scenario 2, we can identify the capacity requirements using the CDP Workload XM for all on-premise workloads. Based on that plan, provision the appropriate sized CDP environment in the cloud. Next, migrate all historical data as well as metadata, policies and jobs (MapReduce, Hive, Spark, etc.) to the target CDP cloud environment. In this scenario, data ingestion jobs (like Sqoop or Informatica or Attunity) will have to be re-pointed to the target cloud storage environment. For example, if an Informatica PowerCenter job is currently acquiring data from an Oracle DB and writing to HDFS on the on-premise cluster, it will have to be reconfigured such that the target could now be S3 (if AWS is the chosen). Once the ingestion pipeline is re-configured, the entire data supply chain can be migrated to the cloud CDP environment quickly with minimal re-configuration. Note that it is best to move the BI / Visualization environments also to the cloud since all data will be resident in the cloud. This prevents lot of data movement out of cloud.



Note: Although this scenario talks about migrating the entire data ecosystem to cloud, this should not be done is one flash cut. This should be done in phases where we take pieces / slices of the end-to-end data pipeline (ingestion to curation to consumption) and migrate them to cloud. Eventually and over time, all pipelines will be migrated and the on-premise environment can be decommissioned.

Build new data ecosystem in cloud

This scenario involves building a new parallel data ecosystem in the cloud using CDP while existing workloads continue to run on the on-premise environment. CDP is instantiated in the cloud and data from existing sources, new data sources (databases, files and streams) as well as cloud data sources are ingested into CDP. Additional workloads, such as Machine Learning and ODSs can be built in this cloud environment while BI / Visualization can be driven off the on-premise clusters. This allows our client to easily build net new data capabilities in the Cloud. CDP provides the platform and services to build new data-driven applications in the cloud with the right governance and security controls.



Counter argument

A counter argument to the point-of-view presented in this paper is that cloud native technologies can be used to achieve the desired outcomes described in the scenarios above. One could argue that using cloud native services we can develop a data platform that would support all desired functions and workloads. However, we would expect to encounter the below challenges:

- To use the on-premise data in the cloud, the data from the HDFS environment has to be replicated using some replication technology (as opposed to Cloudera Replication Manager).
- Data curation jobs (Spark, Hive, etc.) built on the on-premise Hadoop cluster have to be tweaked and modified to run on cloud native services (as opposed to migrating them to CDP in Cloud with no re-write).
- Data needs to be copied "multiple" times into different data stores to handle different workloads (as opposed to shared data for all workloads in CDP).
- Data often gets locked-in to proprietary formats and has to be exported out as CSVs for sharing with other applications and would require additional steps to convert into open formats like ORC and Parquet (as opposed to CDP which keeps all data in open format and identically shared across all engines).
- Metadata such as Hive databases and table definitions have to be migrated using custom written scripts (as opposed to inherently copied over with a single click in CDP).
- Security policies have to be redeveloped and applied in the cloud environment across various services, with differing capabilities and tools (as opposed to inherently copied over with a single click in CDP). This problem is exacerbated when we go across multiple cloud providers.
- Consumption workloads have to be modified to run against the cloud services (as opposed to no change needed with CDP).

Yes, it is possible to realize these scenarios using native cloud services, but it will require more effort, take longer and be more risk prone to mitigate these challenges, than using CDP in cloud.

Conclusion

The evolution to cloud needs both a rotation of existing capabilities as well as creation of newer innovation journeys, specific to each industry, that are born and scaled in the cloud. Cloudera's CDP platform can be used by existing Hortonworks and Cloudera customers to quickly migrate workloads to cloud without risk. With new features and functions, and robust governance and controls to go along, CDP can serve as the next generation data platform in the cloud for our clients.

Introducing Cloudera Data Platform

Cloudera Data Platform (CDP) combines the best of Hortonworks' and Cloudera's technologies to deliver the industry's first enterprise data cloud. CDP delivers powerful self-service analytics across hybrid and multi-cloud environments, along with sophisticated and granular security and governance policies that IT and data leaders demand.

Initially delivered as a public cloud service, CDP includes:

- **Data Warehouse** and **Machine Learning** services as well as a **Data Hub** service for building custom business applications powered by our Cloudera Runtime open source distribution.
- 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 threats across environments.
- 100 percent open source, supporting your objectives to avoid vendor lock-in and accelerate enterprise innovation.
- A clear path for extending your existing CDH and HDP investment to the cloud.

Author

Sharad Kumar
Managing Director and CTO, Data & AI
Accenture Technology
sharad.h.kumar@accenture.com

About Accenture

Accenture is a global professional services company with leading capabilities in digital, cloud and security. Combining unmatched experience and specialized skills across more than 40 industries, we offer Strategy and Consulting, Interactive, Technology and Operations services—all powered by the world's largest network of Advanced Technology and Intelligent Operations centers. Our 506,000 people deliver on the promise of technology and human ingenuity every day, serving clients in more than 120 countries. We embrace the power of change to create value and shared success for our clients, people, shareholders, partners and communities.
Visit us at **www.accenture.com**.

@Accenture

Disclaimer

accenture