# Table of Contents

Chapter-1: Introduction to Cloudera Data Platform ........ 3

Introduction ........ 3

Platform as a Service ........ 4

Edition: Cloudera Data Platform: Public Cloud Services ........ 4

Kubernetes and S3 ........ 6

Cloudera Data Platform: Data Center ........ 6

Revisit Cloudera Hadoop Distribution (CDH) and HDP ........ 7

Chapter-2: Cloudera Runtime ........ 9

Cloudera Runtime is a base ........ 11

Simplified Deployment ........ 11

Cloudera Runtime in detail ........ 12

Impala v/s Hive ........ 12

Chapter-3: Cloudera Shared Data Experience ........ 13

Data Security ........ 13

Security move with the data ........ 13

Shared Data Experience features ........ 14

Chapter-4: Cloudera Control Plane ........ 14

Management Console ........ 16

Cluster ........ 16

Chapter-5: Data platform Hybrid Workload ........ 17

Dynamic CDP ........ 18

Chapter-6: Characteristics of the Enterprise Data Cloud ........ 18

Hybrid and Multi-Cloud ........ 19

Multi-function ........ 19

Secure & Governed (Shared Data Services) ........ 19

Open & Open Source ........ 20

Meaning of the Edge2AI ........ 20

Chapter-7: Virtualization vs Containerization ........ 21

Virtualization ........ 21

Containerization ........ 22

Hypervisor ........ 23

Bootstrap process for virtualizations/container ........ 24

Docker container ........ 24

Docker Image ........ 25

Registries ........ 25

Flow creating and deploying Docker container ........ 26

Kubernetes ........ 26

CDP and Containerization ........ 27

Chapter-8: Cloudera Data Platform Deployment ........ 28

Introduction ........ 28

Public cloud deployment ........ 28

Private Cloud deployment ........ 29

Data Center Deployment ........ 29

Chapter-9: Cloudera Data Hub ........ 31

Introduction ........ 31

Chapter-10: Self-Service Experiences ........ 32

Introduction ........ 32

Self-service experiences ........ 32

Chapter-11: CDP Administrators ........ 34

Introduction ........ 34

Operations ........ 34

Backup and Disaster recovery ........ 34

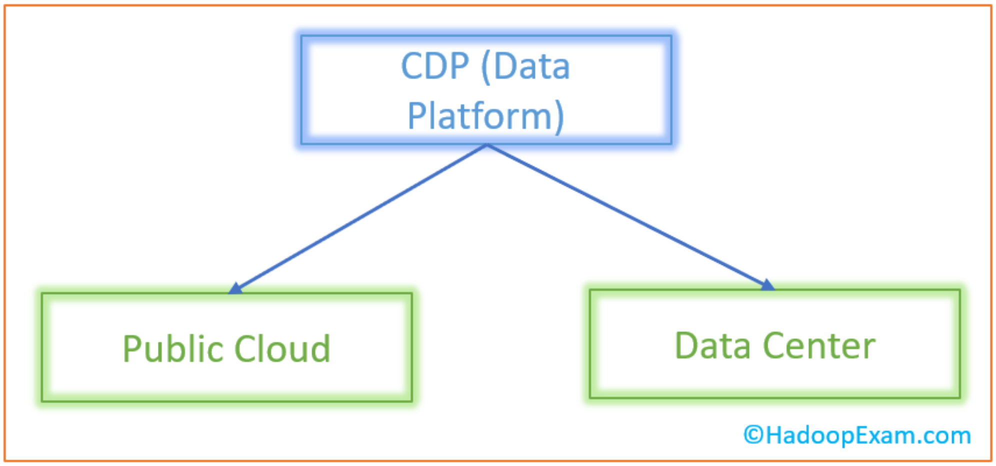
Cluster planning ........ 35

Technical skillsets ........ 35

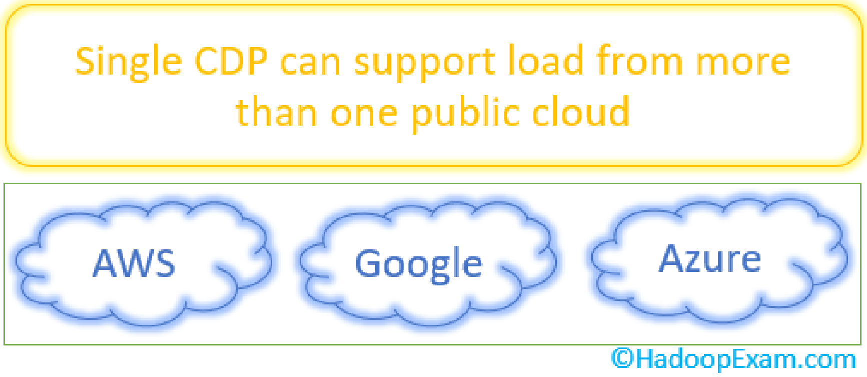
# Chapter-1: Introduction to Cloudera Data Platform

## Introduction

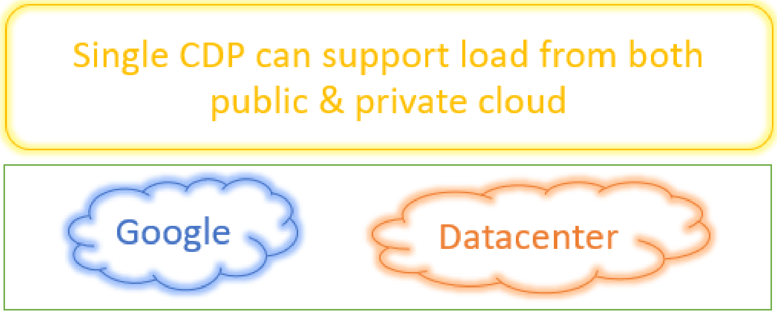
Cloudera Data Platform Data Center, there are multiple version of the Cloudera Data Platform exists, we can broadly divide in two parts as below. These are the two different editions which Cloudera provides



This is a completely new way through which Cloudera supports their platform on the public cloud, or on-premises data center or in the Hybrid env. Hybrid env could be like mixing two Cloud solution providers like Google Cloud, Amazon Cloud, Azure Cloud and your On-premise private data center. As you can see in the below image, you can do the setup of the CDP platform on the public cloud and your Cloudera workload could be on any of the underline public cloud or it could me more than one public cloud.



Similarly, you can see your load can be on the Hybrid env, which can include public cloud and on-premises datacenter.


Single CDP can support load from both public & private cloud
Google
Datacenter

©HadoopExam.com

If you don’t want to use public cloud then you can choose purely CDP-Data Center Edition. Both edition of Data platform has the same underline feature with somewhat different functionalities. Let’s discuss each one them in brief first.

In future it seems to be planned run on the other public cloud like IBM and Oracle.

## Platform as a Service

CDP (Cloudera Data Platform) fall under the platform as service product using this you can securely manage and govern their data with the deployment of the Analytics and Artificial Intelligence workload. CDP invented because of the two big organization merger (Cloudera Inc and Hortonwork Inc.) both were providing almost same solution. There were two major products were existing

- **Cloudera Inc:** Cloudera Distribution of Hadoop (CDH)
- **Hortonworks Inc:** Hortonworks Data Platform (HDP)

As if you were using CDP for public cloud then two major changes you can observe

- **Container Management:** YARN (Yet another resource negotiator) have been replaced with the Kubernetes and container management.
- **Storage**: Similarly to provide support of Big Data load in the public cloud HDFS is being replaced by the Object storage, because all three public cloud Amazon (S3), Google (Cloud Storage) and Microsoft (Data Lake) support the object storage solution.

By implementing and creating CDP platform Cloudera made it quite simple model for the Hadoop complex deployment. Most of the cluster management is done by the software installed on the platform and reduces lot of work overhead.

## Edition: Cloudera Data Platform: Public Cloud Services

This is a managed solution from the Cloudera and you don’t have to install any software your own. Even this is managed by the Cloudera, entire data remain under your VPC (Virtual Private Cloud), you can learn from here. As of this writing it supports the Amazon AWS, Microsoft Azure and Google Cloud Platform. Even it is managed by the Cloudera, control is in your hand for creating new workload and can completely control the cost, security, access control etc. You can even automate your workload like

- As per schedule or on-demand based on the new data, it can spin up the new load and as soon as work is done load can be suspended.

Create different types of workload: based on the need you can create different kind of workload and each workload can be isolated based on

- User Types
- Workload Types
- Priority

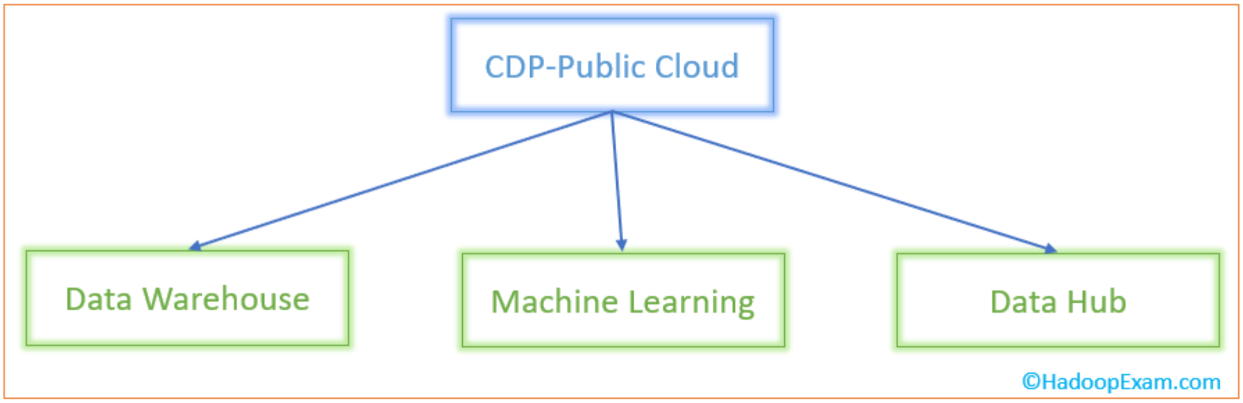
Central managed, whether you are using your Hadoop Cluster on the public cloud, private cloud or on-premises data center, you can control all of them together from a single glass pane view. Even you can create clusters or workload for your customer and you can control the same from the Cloudera Data Center.

There are various benefits of using the Public Cloud Edition, lets discuss few of them as below

- **Auto-scaling:** It is difficult to predict sometime the work load when you work on the new upcoming data. As your setup is already in the public cloud and managed by the Cloudera and controlled by you. You can increase the capacity of the cluster with just few clicks in the management console. Even for your new customer you can create a new workload deployment quite quickly.
- **Migrating from On-premises:** You may have see that your on-premises hardware capacity is not enough to run the workload and you need immediate migration from on-premises to the cloud. This can be done quickly as well as having proper control in place. And you don't have to buy and do the capacity planning for your on-premises data center Hardware resources.

I know its benefit listed only two bullet points, but this represent huge improvement and reduces the operational overhead and setting up the cluster in the cloud env.  As we move ahead we will discuss further on this topic.

There are three major types of workload (functional load) which is defined for the Cloudera Platform public cloud version, which is depicted below.



Let's briefly discuss each one, this you can say that three application supported on the platform as a service.

**Data Warehouse**

As name suggest this is a service available on the Cloudera Data Platform where you get the huge volume of data ingested. Your data format can be structured, unstructured, semi-structured etc. And on this data, you can build an analytics service like generating reports, dashboards, ad-hoc queries etc. And obviously everything in secure manner. This can use some tools under this are Impala, Hive LLAP, Hive on Tez, Hue (Hadoop User Interface), Kudu, Druid, Solr,  and Workload XM. Below are the kind of data this workload can support

- Time-series data
- Server logs
- Clickstream data
- Sensor data

This basically minimally includes the Hive, Impala query engines.

**Machine Learning**

As you all the Machine Learning workload needs the iterative data processing and this can be easily taken care in the public cloud, even you can replicate your on-premises data into the cloud. And once data is replicated in the Cloud then you can deploy new Machine Learning model on them using the Machine Learning Workspaces and this ML workspaces are pre-configured with the data access, controlled computing resources which you were using in your 0n-premises datacenter.

In this it minimally includes the Data Science solution using the Python, R and Spark.

**DataHub**

This is a service which can be used to create new clusters quickly based on the pre-defined templates. And you can do the cluster management as well like adding new nodes to the clusters, stoping, starting and terminating the cluster.

This is more of your traditional YARN-managed environment where you can run Hadoop workloads like MapReduce and Spark.

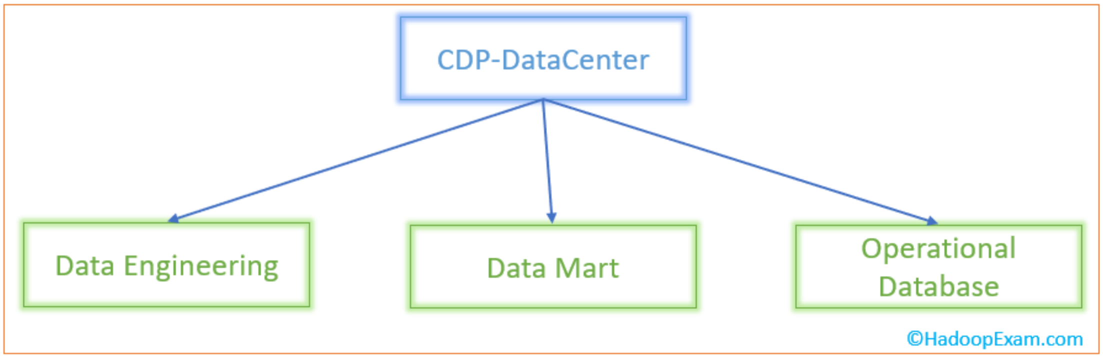
In future it is possible that Cloudera add more application on the public cloud (PasS) platform.

## Kubernetes and S3

Cloudera Data platform in public cloud is based on the Kubernetes and S3 (In case of AWS) and for the Data Center this is on YARN and HDFS. This is not where as a user you have to worry too much. If you have seen trends from the last few years Kubernetes has emerged as a de-facto standard workload orchestrator in the public cloud.

## Cloudera Data Platform: Data Center

As we have discussed and introduced ourselves with the public cloud version now let's see the Data Center version. These products have been introduced after the merging of Cloudera and Hortnworks Inc and feature included here are the combined feature from both the company's products as well some additional new features. This is mainly developed for in-house datacenter and create different kind of workloads based on the need. The kind of workloads you can create are as below



I will briefly introduce all this workload as of now and later on we can discuss the same broadly.

- **Data Engineering:** As regular ETL or ELT work you need to ingest the data and then transform this data in the desired format and saving this data on the HDFS. Using different analytical tool you can analyze this data. This entirely depend on the underline components Hive, Oozie, Spark etc.
- **Data Mart:** As you already have experience working with the scenario where you need to browse your data using the queries, you or your team can apply the analytics using the interactive way. There are again components like HDFS, Hive, Impala etc are used.
- **Operational Database**: If you want to have transactional data load then using the Atlas, HBase, HDFS you can create Low Latency read and write database.

## Revisit Cloudera Hadoop Distribution (CDH) and HDP

In 2008 Cloudera pioneered the Hadoop framework and began working with the Hadoop distribution with other best suitable open source tools and that distribution was known as a CDH (Cloudera Hadoop Distribution). This is more of each integrating all independent component and they all should work in the desired manner. And providing commercial support for all these products together. On the similar fashion in 2011 there is another company founded named Hortonworks and created very similar platform known as HDP (Hortonworks Data Platform). There are certainly some differenced between the components provided by each company and both are competing with each other.

And these both companies were actively involved with the development and enhancement of the open source projects. Both companies were having overlapping open source projects like Hadoop, Hive, Spark etc. And there are some products which were specific like Flume and Kudu was part of CDH and NiFi and Druid was part of HDP.

Since then many things have changed and updated like MapReduce processing engine is replaced by Spark, which has a very good support for faster data processing, stream data processing and Machine Learning capability and HDFS storage layer is being replaced by Kudu etc. Kudu is good fit for the time-series data and data collected from the IOT devices and works well with the Apache Spark. Even some companies were using Cloud storage like S3, Azure Data Lake etc.

Overall if you see the role of the Hadoop is getting reduced and each individual component integrated together are solving the BigData problems much better way then only by the Hadoop framework.

Around 2019 both the Cloudera and Hortonworks got merged and their both the products CDP and HDP are working together. And CDP come as a resultant product which has capability much beyond both the HDP and CDP combined provide for the BigData, Machine Learning and Artificial Intelligence. This is complete revolution for the Enterprise Data Cloud and not simple merge of the CDH and HDP.

Having CDP provides some capabilities like

- Ease of use (Until you understand the platform, it may seems complicated)
- Pay as you go model
- Ad Hoc workload support
- Spin up/down the cluster or workload
- Auto scale
- Auto resume and stop the workload
- Quick deployment

With the CDP basically you need three things, if you are using public cloud as depicted below.

Cloud Account
CDP Subscription
Customer Data

©HadoopExam.com

By having all these three things are not time taking and you can start running analytics on your data. There are the features like multi-cloud which is not possible with any other public cloud provider.

And if there is any regulatory constraint with the data like it should not go outside the datacenter then same solution can be used with the Cloudera Data Platform Data Center edition.

# Chapter-2: Cloudera Runtime

In this chapter we would be discussing about the Cloudera Runtime environment. Cloudera Runtime is a single component having more than 50 different open source technologies included in it. Yes, Cloudera Runtime is a core of the Cloudera Data Platform Data Center on which various types of workloads run. We can discuss each individual service using which this Cloudera Runtime have been created. But rather first understand the based on various use cases, and across the use cases you use the combination of the different services out of all the 50 different services.

- **Storage**: Cloudera Runtime support the data storage using the combined services like HDFS, Apache Kudu.
- **Compute**: Apache YARN (Yet another resource negotiator)
- **Data Access:** Already ingested data can be accessed using the various services like
  - Apache Hive
  - Apache Impala
  - Data Analytics Studio
  - Hadoop User Interface (Hue)
  - Cloudera Search using Apache Solr
  - Apache Phoenix
- **Operation Datastore**: If you need low latency read and write of the data, it can be achieved using the Apache HBase NoSQL databases.
- **Data Science:** Data scientists and Data Analytics team can use the components like Apache Spark, Apache Zeppelin
- **Security**: There are various load and data and all needs to have secure access and that can be achieved using the Cloudera Data Platform security functionality.
- **Data Governance**: There is need sometime you need to have access to the data which may not be part of your cluster and what kind of the structure this data has. Cloudera Runtime provide such capability using the Apache Atlas which work as Common metastore and can be used for exchanging metadata both within or outside the Hadoop stack.
- **Streaming**: We need to have continuous stream of data to be loaded and for the Cloudera Runtime currently usage Apache Kafka for the same.

Following is the list of components which are included in the Cloudera Runtime, list is not provided in a particular order and does not include any kind of versioning information, this is more of the understanding perspective, what all you can do using the Cloudera Runtime env. If you know the components. And when you create any specific workload in your Cloudera Data Platform (Public/Datacenter), it can automatically include the service required for the pre-defined workload. At the same time you can select specific service from below and define your custom workload as well.

- Apache Flume-ng
- Hadoop Distributed File System (HDFS)
- Apache Knox
- Apache Livy
- Apache Parquet
- Apache Ranger

- Cloudera Search
- YARN (Yet another resource negotiator)
- Apache Atlas
- Apache Avro
- Apache Hadoop
- Apache HBase
- Apache Hive
- Apache Hadoop User Interface (Hue)
- Apache Impala
- Apache Kafka
- Apache Kudu
- Apache Oozie
- Apache ORC
- Apache Ozone
- Apache Parquet
- Apache Phoneix
- Apache Ranger
- Apache Solr
- Apache Spark
- Apache Sqoop
- Apache Tez
- Apache Zookeeper
- Apache Zeppline
- DAS
- Hive Warehouse Connector
- Java KeyStore KMS
- Key Trustee KMS
- Key Trustee Server

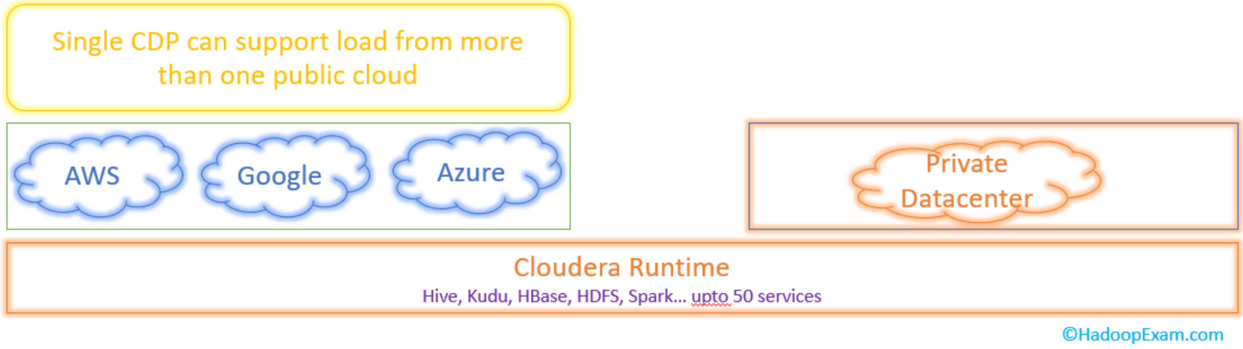
Please note that following components are not included as part of the Cloudera Runtime Environment and based on the which CDP you are using would be included separately.

- **Cloud Data Platform:** Public Cloud Edition
    - Cloudera Data Hub
    - Cloudera Data Warehouse
    - Cloudera Machine Learning
- Cloudera Management tools:
    - Management Console
    - Cloudera Workload Manager
    - Replication Manager
- Data Catalog
- Cloudera Add-on products like
    - Cloudera Data Science workbench
    - Cloudera Data Flow

- Cloudera Metron

## Cloudera Runtime is a base

If you see whether you are using the Cloudera Data Platform (CDP) public cloud edition or the Cloudera Data Platform Data Center edition both are fully depended on the Cloudera Runtime as you can see in the below image.


Single CDP can support load from more than one public cloud
AWS
Google
Azure
Private Datacenter
Cloudera Runtime
Hive, Kudu, HBase, HDFS, Spark… upto 50 services
©HadoopExam.com

## Simplified Deployment

If you have been working on the Hadoop framework from few years then you might be knowing that many of the services built by the Cloudera from above 50+ services and they made it open source. Wiring all this component for the administrator is not that easy task and they need to have lot of knowledge across various 50 open source services and this not as easy tasks. And just for analyzing the big data regularly it is really a big work for the company and Cloudera trying to make it simple using the Cloudera Data Platform solution.

As a Data Scientist or Data Engineer or Data Analytics engineer, I don't have to worry about the entire Hadoop infrastructure but rather need to know the SQL query and how to make JDBC connection for analyzing data and creating the reports. Same thing is being done by the Cloud provided using their structured data solution like BigTable from Google cloud, DynamoDB from AWS and for unstructured data S3 and Data lake kind of solutions. And till now most of the time Hadoop system are preferred as a on-prem solution. There are already some solution exists in the public cloud which are based on the Hadoop as you can see below

- Amazon Elastic MapReduce
- Microsoft Azure HDInsight
- Google Cloud DataProc

Most of the largest companies need such solution and industry is wide growing, and many of the organization started using the public cloud as well then Cloudera Data Platform would be good choice for the BigData Analytics, Data Warehouse solution.

## Cloudera Runtime in detail

Cloudera Runtime is open source and its source code is available free to download and if you want you can create binary from the source code. As of now binary download is not freely available. You need to either use the 60 days free version of the Cloudera Data Platform Data Center version.

Cloudera runtime is again a combination of various open source software and services and on which this new CDP platform was built. And actual functionality like ingesting, storing, processing, managing and analyzing the data happen via the various underline services.

Cloudera Runtime can include new service in the new release or updated version of the service for example as of now they are supporting Spark 2.4 version but in future new release of the Cloudera Runtime version they can include Apache Spark 3.0 and you don't need to update your entire CDP platform. Just updating the underline Cloudera Runtime would provide you the updated version of the Cloudera Data Platform.

As I mentioned there are around 50 different services are included in the Cloudera Runtime and all of them are the open source software component which you already have been using under the CDH and HDP distribution previously.

Each software component in the CDP are considered are part of the specific role, for example we can say that Hive and Impala are the part of following two services

- Cloudera Data Warehouse
- Cloudera Data Mart

Which can be used for running SQL queries in both batch mode as well as interactive mode. Similarly, there is another service for implementing the access control mechanism and data governance feature which is Apache Ranger.

You can say that Cloudera Runtime is merged component of both the CDH and HDP and provide the best features of both the components. People already familiar with the CDH or HDP may not feel that CDP is different thing. Yes, certainly they would find around 30% new feature which was not previously exists or different from CDP/HDP.

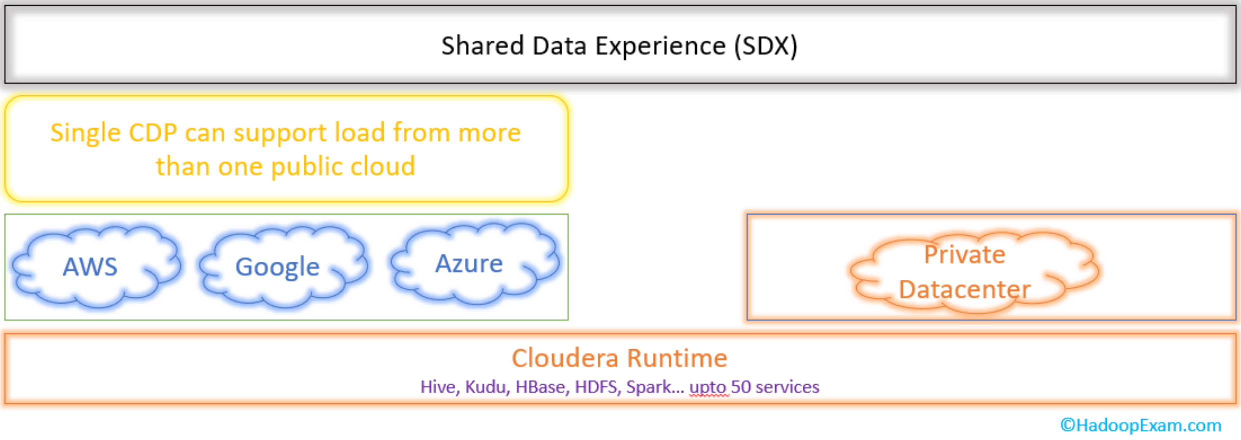
In both the edition CDP (in public cloud) or Data Center version the underline runtime is the same and this work as a foundation for the Cloudera Data platform. Few of the old components have been replaced as well for example for providing role based access Apache Sentry would not be used anymore and for role based access control Apache Ranger would be used which has more features like attribute based access control, dynamic row filtering, and dynamic column masking for the sensitive data.

## Impala v/s Hive

In Cloudera Runtime both Impala and Hive present because both solves the different problems like Hive for the Data Warehouse solution which involves the large data joins and ETL activity. And Impala is better fit for the interactive queries and good fit for Data mart kind of solution. That's the reason both the solution kept in the Cloudera Runtime.

# Chapter-3: Cloudera Shared Data Experience

On the top of the Cloudera application whether this is under the Cloudera Data Platform (public cloud) or Cloudera Data platform (Data Center) edition both usage the Cloudera Shared Data Experience (SDX) and the main responsibility of this component is to provide security, governance, and lineage data to the stored data across both the editions.


Shared Data Experience (SDX)
Single CDP can support load from more than one public cloud
AWS
Google
Azure
Private Datacenter
Cloudera Runtime
Hive, Kudu, HBase, HDFS, Spark... upto 50 services
©HadoopExam.com

## Data Security

As you can see Cloudera Shared Data Experience usage the metadata to track the lineage , enforces authentication and access policies across cloud and on-prem clusters. You are confident enough about the security of your data as well as authorization for accessing the data and which is consistent across the public cloud or Hybrid deployment. Again, this thing can not be implemented by the Cloud providers and CDP lead this security requirement and that is where all of the Cloud provider lacks and even not possible.

## Security move with the data

When you plan to move the data from Cloudera Data Platform Data Center (which is based on the underline HDFS based storage) to the Cloudera Data Platform (Public Cloud) then not only the plain data would be copied but also the metadata, security policies, governance and lineage is also copied. And once all are copied that is in ready to use format only and with the same security and access control in place until this is modified.

### Shared Data Experience features

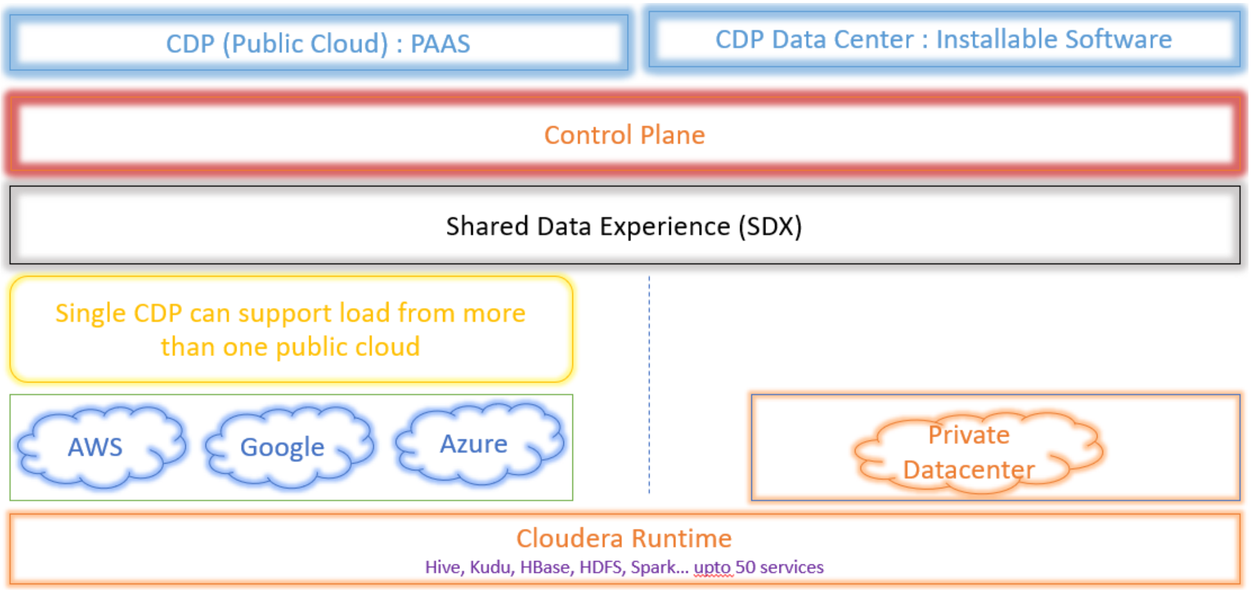
When we talk about the shared data service then generally it talks about the following features

- Metadata
- Data Schema
- Security
- Governance
- Migration of the data

This provides the data access control layer and provides the consistency for security and governance within the CDP and it does not matter whether you are using Cloudera Data platform (In public cloud) or On-premises.

# Chapter-4: Cloudera Control Plane

Cloudera control plane which is mainly used by the administrators and that component works as a “single pane of glass” and would be used for the spin up and spin down the clusters/workloads and in both the editions. And support the cloud, on-prem and hybrid scenarios. By this time you could have completed the overall architectural component understanding for the Cloudera Data Platform as shown below

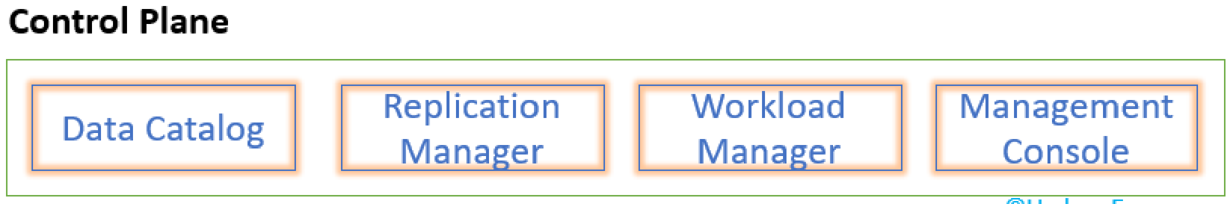

CDP (Public Cloud) : PAAS
CDP Data Center : Installable Software
Control Plane
Shared Data Experience (SDX)
Single CDP can support load from more than one public cloud
AWS
Google
Azure
Private Datacenter
Cloudera Runtime
Hive, Kudu, HBase, HDFS, Spark… upto 50 services

©HadoopExam.com

Control panel is foundation for the operations related to managing and deploying services and mainly used by the administrators. And helps in

- Managing data
- Resource provisioning in CDP
- Managing access control
- Doing Hybrid deployments
- Data replications

Control panel also depend on the Shared Data Experience (SDX). Control plane is an integrated component or service which has following components in it. And all these components together provide the access to the right data, to the right people and at the right location having security in place and not impacting the performance.



©HadoopExam.com

- **Data Catalog:** Using this service user can search, understand, organizes and govern the data assets across all the CDP environments.
- **Replication Manager:** As name suggest it is a management tool which is primarily used by the administrator for replicating, migrating data, metadata, security policies between various CDP environments.
- **Workload Manager:** Administrator uses this component for the workload management and for the performance management as well. Administrator analyze, troubleshoot and optimize the workloads across the environments also helps in reducing the cost.

- **Management Console**: This is known as single pane of glass for managing all the clusters across the environments. It has three mode by which clusters can be managed as below
  - **Web Interface**
  - **Command Line Interface**
  - **API (Good for automating things)**

## Management Console

Management console is not just managing a single Hadoop cluster but rather a all the cluster which are created by the organization using the single pane of glass. And there are multiple region for which so many clusters are being created. Cluster can be created for few hours and some cluster for long running jobs etc.

Using the Management console first you go in a particular env and then from there you can select specific workload or cluster. Let me give you the example of environments. You could have created three different env based on the cloud provider and CDP data center.

- Environment-1:  All clusters are on AWS (9 workloads or cluster in this env.)
- Environment-2: All clusters are on Azure (3 workloads or cluster in this env.)
- Environment-3: All clusters are on local data center (2 workloads or cluster in this env.)

So, as per above configuration we can say that there are total 14 different workloads across three environments.

## Cluster

When you open a specific cluster then you can find the following details on the console

- What all are the services this cluster is running e.g. Hive, Hue, Oozie etc.
- You can find the URL for the Cloudera Manager for that cluster
- Information about the environment on which this cluster or workload is created.
- Which template have been used for creating this cluster?
- Hardware detail about the each node in the cluster.
- You can find the storage detail for the data and metadata, in case of AWS. You can see all the S3 urls in this tab. For example location of the Hive Warehouse directory etc.

Start/Stop and terminate the cluster

- If you need cluster only for specific use case then once done you can terminate that cluster.
- Similarly, if you want this cluster during the specific time more than once then it is better to stop the cluster and whenever needed back then start the cluster again.

Increasing and decreasing the size of cluster

- You can add additional node to the cluster from the cluster resize action.

# Chapter-5: Data platform Hybrid Workload

There are already services exists in the public cloud for the BigData solution then what different thing is provided by the Cloudera Data Platform, why we should go for this instead of the below existing services in the public cloud.

- AWS Elastic MapReduce
- Azure HDInsight
- Google Cloud DataProc

If you see all above three are purely Hadoop distribution on the Cloud and Cloudera Data Platform is much-much more than that. Cloudera Data platform can have 3 different deployment flavor which is not possible with the Hadoop distribution on the Cloud platform as below

- Cloudera Data Platfiorm (On single Public Cloud)
- Cloudera Data Platfiorm (On multiple Public Cloud platform together)
- Cloudera Data Platfiorm (On-prem , private Data Center)

This multi-cloud, Hybrid Cloud and On-premise data platform can not be provided by the public cloud and CDP is designed for solving all these different deployment use cases.

### Dynamic CDP

This is not just about the deployment modes but there are many more things which can be done in the Cloudera Data Platform which is not straight forward in the public cloud.

#### *Data Movement*

If you are using the CDP then you would be able to move data and workloads from on-prem to the cloud and other way round as well. Even you can transfer the data between different Cloud providers with the just few clicks.

## Chapter-6: Characteristics of the Enterprise Data Cloud

There are basically four characteristics on which this Enterprise Data Cloud was built upon.

1. Hybrid & Multi-Cloud
2. Multi-Function a.k.a. Edge2AI
3. Secure & Governed
4. Open Source

## Hybrid and Multi-Cloud

As we have already discussed previously that CDP (Public Cloud) will support the multiple public cloud providers like AWS, Azure and Google cloud. And in future at least couple of more cloud provider like IBM and Oracle would be added.

Using the Hybrid model you can divide your workload in Local Data center as well as public cloud. As there are various reason of having data like legacy data in the on-premises data center and new data is getting collected in the Cloud. And without moving or replicating load should be able to run. This makes CDP highly flexible platform and even there is no-vendor lock in because CDP support multiple public cloud providers.

## Multi-function

CDP support the multiple types of workloads which include the processing and analyzing the streaming data, machine learning, data warehouse etc. And it is expected that the CDP should support any kind of work load with any kind of data format or types from any possible source.

Same data can be used for multiple purpose like online ecommerce transactions data can be used for generating the reports and the data scientists can use the same for the experiment, recommendations, product pricing etc.

And CDP is made for the multi-function and covering variety of use cases and new capabilities continuously being added. And you don't need very strong system admin or database admin team to handle such use cases. Not even separate vendors are needed.

Same data can be used for different kind of workload which can reduce the data duplication and reduce storage as well as operational cost. No multiple teams or resources for maintaining the same data at multiple places.

## Secure & Governed (Shared Data Services)

There are some regulatory requirement as well as you need proper control in place who can access the data and that can be achieved using the shared data services from the CDP platform. Especially financial and healthcare organization are much more worried about the data access because there are stringent regulations and compliance requirements for data security.

Using the SDX everybody can comply with the data privacy, governance, migration, and metadata management.  At the time of auditing historical data should be available for investigation.

To understand the data, we need good metadata information which can help in creating new business opportunity.

## Open & Open Source

The source code and underline Cloudera Runtime services source code is 100% open source and anybody can build services using the same source code which is being used by the Cloudera. (Is it really that easy, obviously not so simple). But it certainly helps the developer to know and check the underline source code if there is any issue detected and needs to be resolved or understood. And same can be used with the people who are working in the open source world.

CDP system can be easily integrated with the other system as well. There is no cloud vendor lock-in, you can choose any public cloud provider for your requirement. '

## Meaning of the Edge2AI

Recently IOT/IOAT (Internet of Any/Thing) has recently grown a lot and machine sensors are generating lot of data continuously and that needs be brought into the system for further analysis. Cloudera has the solution for this problem and that is known as Cloudera Data Flow which can bring data from the edge (sensor) to the CDP and their data scientists can play with the data by applying the artificial intelligence.

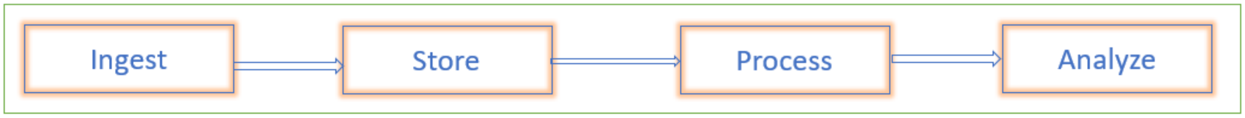
It means bring the data from the sensor using CDF (Cloudera Data Flow) and using this data make prediction for having business decision done and all using the single platform. Scenarios where this can be used

- Automobile industry to continuously monitoring the vehicle components.
- In factories continuously monitoring the equipment
- Predicting the vehicle maintenance
- Predicting equipment maintenance

With this you can always be proactive and predict any failure of the equipment, appliance, vehicle etc. IOT data can come from various sources like thermal sensors, vibration sensors and this data then needs to be transmitted to the Cloud, should be stored in the data lake and finally should be processed in the data warehouse. Using the machine learning continuously you can analyze this data and need to detect any discrepancies or issue with the data. This is the data which has following three BigData characteristics

- **High Volume:** Obviously traditional systems are not capable of handling such a high volume of the data.
- **High Velocity:** Data is generated with the micro-second level.
- And different variety: It can generate structured, unstructured and semi-structured kind of data.

Following is the basic workflow for the predictive maintenance of the data



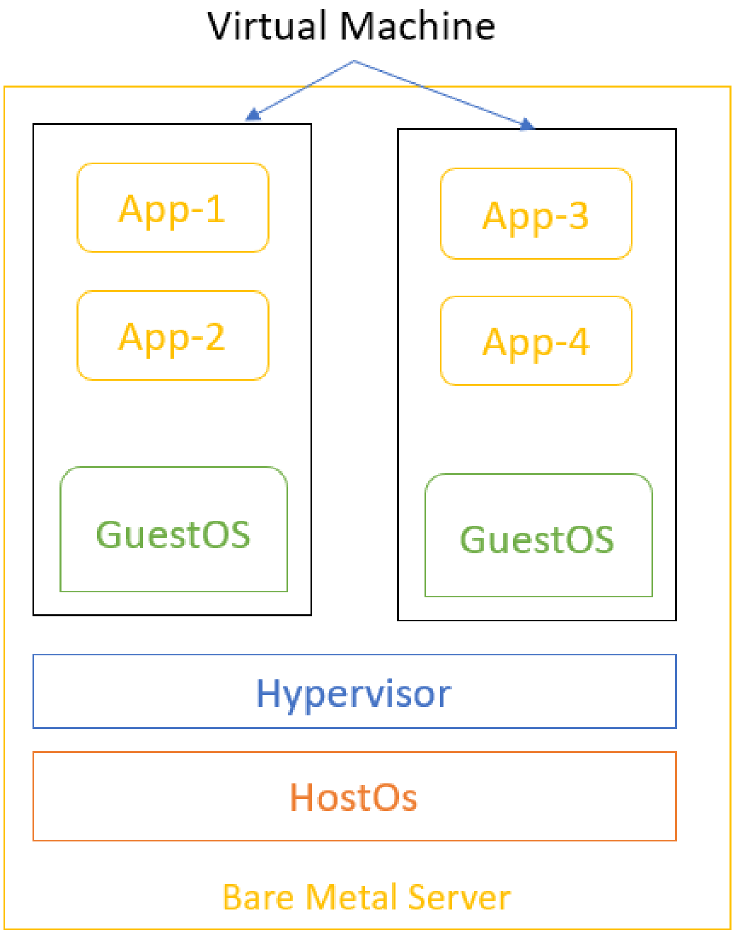
©HadoopExam.com

- **Data ingestion:** Using either CDF (Cloudera Data Flow) or Kafka Streaming engine ingest the data from thermal or vibration sensor. Data may include the source information like sensor ID, payload (temperature) and any other contextual and environment data.
- **Store Data:** And this sensor data needs to be stored in the large databases like HBase, DynamoDB etc. Data would be stored in the initial raw format.
- **Process and transformation:** You need to convert this data in the desired format. This may include removing undesired information from the data and compressing it into the format which we need for applying further processing.
- **Analyze:** You can apply the machine learning or this data and do the predictions.

# Chapter-7: Virtualization vs Containerization

## Virtualization

Since many years and till now virtualization is dominated to divide single bare metal server into the more than one virtual machine. As you can see in the below image



Even, if you have used the Amazon EC2 instances or Azure Virtual Machine they use the same technology to provide you the on-demand virtual machines. Using virtualization provides the isolation between virtual machines. However, there is one major disadvantage of this technology that it consume lot of system resources. Because it has its own operating system in each instance as well as one operating system on the bare metal server. (If you have previously used the VMWare Player to launch Cloudera QuickVM instance then you can understand the same).

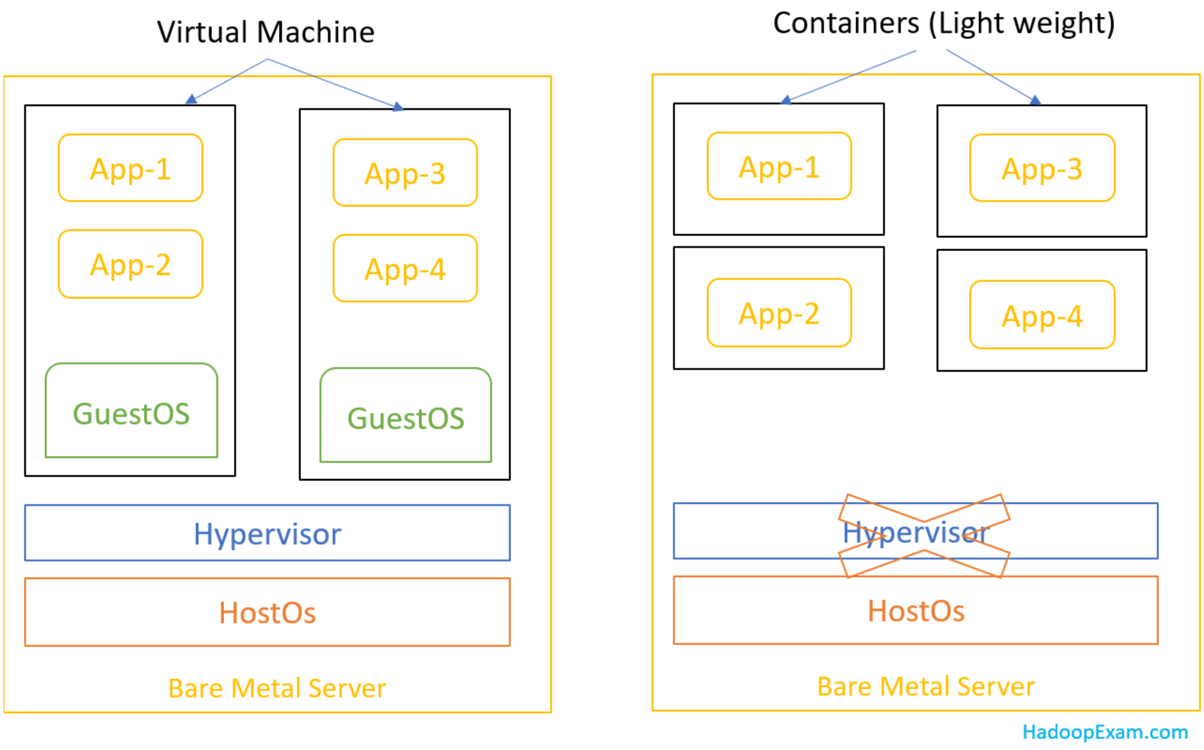
Each time you need a new instance it needs sometime to launch new virtual machine and this is a one of the major issue for the virtual machine. And a newer technology came in recent years to solve this problem and that is known as containerization.

## Containerization

Containers are quite similar to virtual machine. They allow us to run multiple services on the same physical host, while exposing different environments to each of them and also isolating them from each other, but they are having much less overhead then virtual machines. A process running in a container runs inside the host operating system very similar to other processes but in case of a virtual machine each process runs on a separate operating system.  However, the process runs on the single operating system in case of container they are completely isolated with each other.

Containers are lightweight and allows us to run more components on the same hardware, VM generally have their own components on the same hardware and require additional resources.

A container is more of a single isolated process running on the host and consuming only the resources that app consumes and does not have any additional overhead processes which virtual machine has and make it heavyweight compare to container. You can refer below image for understanding where no guest os and Hypervisor for virtual machines.



## Hypervisor

As you can see in the above image there are three separate operating systems are working in case of virtual machine

1. Host operating system
2. Hypervisor which divides physical hardware resources into smaller set of virtual resources which can be used inside operating system for each individual virtual machine
3. Guest operating system: Application running inside the virtual machine makes a system calls two operating system kernel inside the virtual machine then kernel performs x86 instruction on the host physical CPU through the hypervisor. Increase of container application directly communicate Kernels running in the host OS and perform x86 instruction on the host CPU.

## Bootstrap process for virtualizations/container

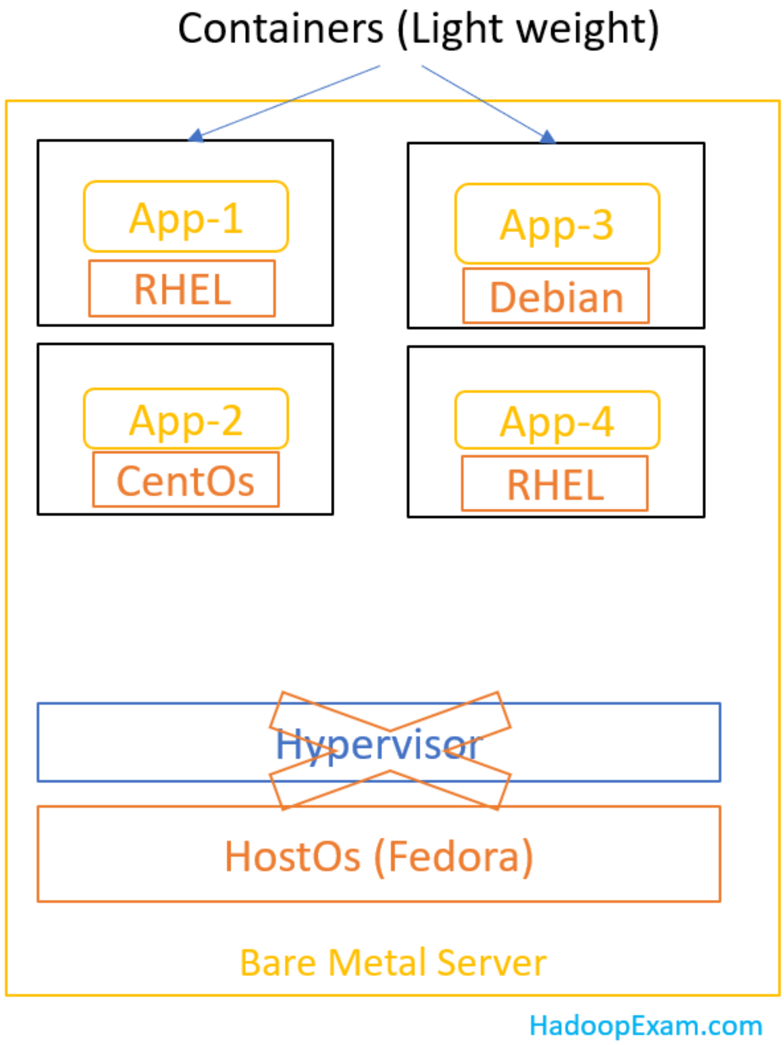
The main benefit of virtual machine is that it is fully isolated and each VM has its own Linux kernel, while containers use the same kernel. If you have high number of processes then you should go for containerization and if you are low number of processes then virtualizations a good choice. Containerization gives low overhead of process bootstrapping because the directly work with the host operating system which is not the case with the virtual machine. So, any process running in container starts up immediately.

Container just need the library and dependencies to run your application and require much less memory than the virtual machines. Booting is much faster than the virtual machines.

## Docker container

When you run an application package with docker, it sees the exact file system content that you have bundle with it. It sees the same files whether it's running on your development machine or a production machine, even if the production server is running a completely different Linux OS. The application won't see anything from the server it's running on, so it doesn't matter if the server has a completely different set of installed libraries compared to your development machine.

Suppose you have completely bundled a new operating system like REHL with your application then wherever you run it whether on Fedora or other Linux distribution, it would always see that it is running on RHEL. The only difference is the Kernel. As you can see in below image HostOs is Fedora but all other application has their own OS files like RHEL, Debian, CentOs etc.



Even after using the operating system this makes your application much lighter with docker container.

## Docker Image

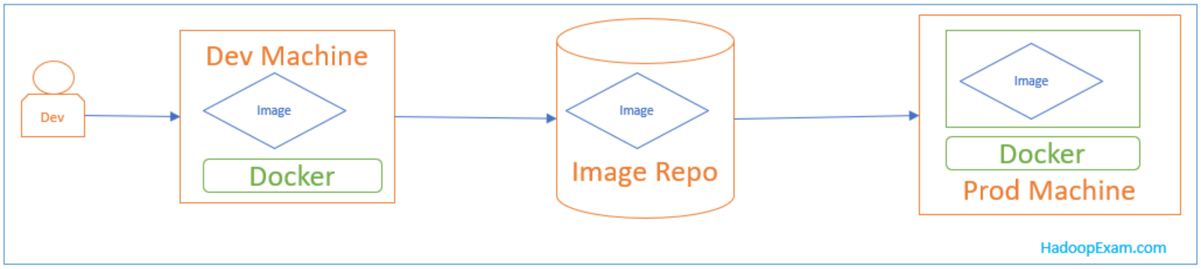
A Docker container images something which you package your application and its environment into. It contains the file system that will be available to the application and other metadata for example path for your executable that should be executed when the image run.

## Registries

It is a General concept same as with other technologies that you might have learned, this is a repository which stores the docker images and from there this can be shared among the different computers. If you have created a Docker image you can run it on your own computer or just upload that image to registry and then download on another computer.

## Flow creating and deploying Docker container

You can follow the below image for better understanding



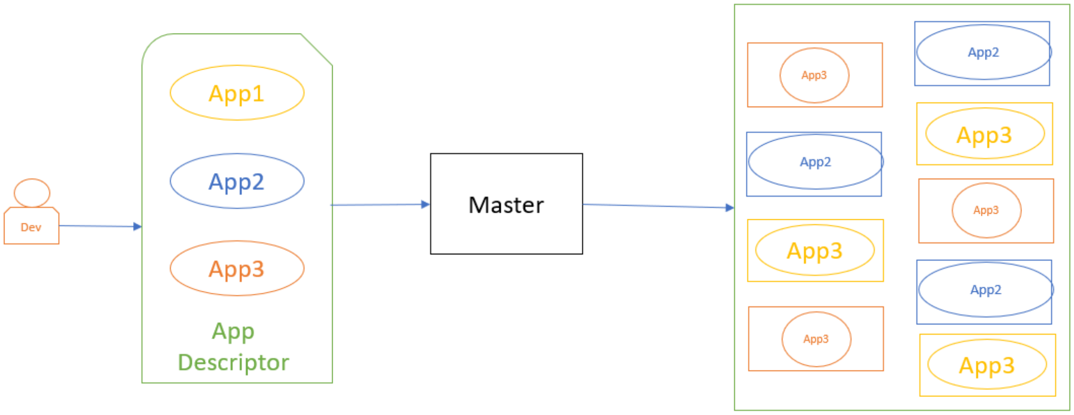
1. Being developer would request on docker to build and push (upload) image.
2. Then docker build the image based on your configuration and all the files.
3. Then docker push that image to repository.
4. On production machine you request to download new docker image
5. Docker would pull the image from repository or registry.
6. Docker runs that image in the container

## Kubernetes

Kubernetes is a container orchestration system. It is a solution for deploying and managing containerized application on top of it. Applications running in the containers don't affect other application on the same hardware or server.

Kubernetes' enable you to run a software application on thousands of computer nodes as if all those nodes a single enormous computer. It totally abstracts away online infrastructure.

The Kubernetes system is composed of master node and worker nodes. As soon as you submit a list of applications to master node then Kubernetes deploy this application on worker nodes.  Even it does not matter what node your component/application is landed even being a developer or system administrator.



You can specify like a certain application must Run together on Kubernetes' and will be deployed on the same worker node. Even the application on different worker node can communicate with each other. So, you can think of Kubernetes as an operating system for cluster and get the benefits like

1. Service Discovery
2. Auto scaling
3. Load balancing
4. Cluster self-healing
5. Leader election

## CDP and Containerization

The public cloud edition of the CDP uses the containerization as well as the virtualization. As we have discussed CDP is not a single product and it is more of a combination of the software products and services and provide same and consistent behavior between the Cloudera Data Platform in public cloud, private cloud or in the on-premises data center. This is also known as implementation of the Enterprise Data Cloud which is a unified data management framework to safely provide the access to the data which you need, when and where you need it.

# Chapter-8:  Cloudera Data Platform Deployment

## Introduction

When you subscribe the product from the Cloudera for CDP you need to think what deployment mode you are going to use. There are basically three deployment modes available and based on that pricing of the product also varies. And also, there are various software components which varies based on this deployment.

- Public Cloud deployment
- Private Cloud deployment
- On-premises deployment (CDP-DC)

This is quite different from the previous CDH and HDP model. And different type of editions or deployments are optimized for that particular deployment mode.

## Public cloud deployment

As name suggest this is for the public cloud only currently there are AWS and Azure supported and soon Google cloud would be supported. As per the Cloudera Website, in future they are planning to have support for the Oracle and IBM cloud as well. As we have discussed previously this is PaaS (platform as a service model). And for using this mode you don't have to provision any hardware and quickly you can productionize it and not even you have to install any software for that.

This is kind of self-service experience and based on your requirement you can quickly spin up the different types of possible workloads for example

- Data Warehouse
- Data Hub
- Machine Learning

And business users can quickly start working. There are some administrative tools available to manage the cluster and controlling the cloud cost, and administrator can control the resources.  This mode is very different till now you have been using through either CDP or HDP, which you were generally directly installing the Bare Metal servers. Your CDP-public cloud are constrained in particular VPC (Virtual Private Cloud) which represent logically isolated network in any particular public cloud. This mode uses the containerization as well as virtualization (In public cloud server instances are launched using the virtual machine).

This mode is good if you have your data in the public cloud or comfortable migrating data to the public cloud like in S3 bucket or Azure Lakes. If you are already running on-premises workload using CDP and HDP and don't want to provision any new hardware then consider this solution and plan to migrate your existing workload in the public cloud. This can hugely reduce your operational cost. Yes, you need good resources who has the knowledge of Cloud administration. Please check below career guide to understand the role of the Cloud Administrator (Only premium subscriber can access this).

Traditional deployment CDH and HDP always prefer that data should be locally available for the better performance. But this is not the case with the CDP, because in the CDP it is always preferred to keep both the compute resources and storage node separately.

Having separate compute and storage node gives the flexibility, creating isolation of the workloads, and having more elastic load. This is more important then the performance requirement.

If you have the ad-hoc loads like running for few minutes or certain times during the day, then you should consider the public cloud deployment which can reduce the cost and only when needed then only cluster or workload would be spin-up.

### *CDP Environments*

In the CDP public cloud there is concept of the environment which you need to understand. Environment defines where CDP will create and access the resources in your cloud provider account. Your CDP administrator always first looks for the environment and register the particular environment. Most of your CDP tasks are depend on the environment, administrator can register more than one environment this is the very general use case like in case of AWS registering environment from different AWS regions. (Wanted to learn AWS Concepts, use this training and book) this is also included in our premium subscription.

In different environment you can create clusters or workloads. Similarly, an environment can represent different cloud providers. For example, if you are working with all three public cloud providers like Google, Amazon and Microsoft Azure then you can have three different environment and your administrator would register all three environments.

## Private Cloud deployment

If you have your own private data center which is based on the virtualization (VM) then you can use the CDP-DC version which can be used and installed directly on that virtual machines or bare metal servers. Private cloud deployment is quite similar to public cloud deployment and uses the cloud native architecture only and your or your administrator can use the tools like control plane etc. This deployment mode as of now not available, so we would not discuss in detail as of now.

## Data Center Deployment

Similar to the private cloud you can use the CDP-DC for this deployment mode. And combining both the CDP-DC and CDP public cloud you can get the Hybrid deployment as well. This is the one where you would be directly installing on the bare metal servers which is very same with the CDH and HDP deployment you have been following till now. Even you can keep both the compute and storage node together and get the highest performance.

In this case you have download the software and install it directly on the servers which can be bare-metal server or virtual machines and this is not at all cloud-native architecture and you don't find the self-service experience or control plane services.

One important things you need to remember is that you can register this CDP-DC in your CDP-public cloud as well and once you register that you can replicate the data between this public cloud and data center environment as well and create a hybrid cloud solution using the "shared data experience"

**Cloudera Manager**: Cloudera manager is the tool for administrating the Data center as well as Data Hub cluster.

# Chapter-9: Cloudera Data Hub

## Introduction

Cloudera Data Hub is not a new component but re-created for quickly creating clusters and this cluster are known as virtual private cluster. As you are doing this activity in the public cloud you are basically provisioning and deploying new clusters.

Nodes created using the Data Hub in the Cloudera Cluster are the Virtual Machines e.g. AWS EC2 instances. and also, the nodes in the cluster are choses and configured specifically for using the Cloud Storage. If you have any existing load which is on-premise Hadoop Cluster and wanted to migrate on the public cloud then Data Hub is the best suitable option.

If you or your administrator already been using the Cloudera Hadoop then he certainly would be aware about the Cloudera Manager. So, similarly cluster created using the Cloudera Data Hub are managed by the Cloudera Manager only. And it feels like that administrator are setting up new Cloudera Hadoop cluster on the bare metal server only (Actually not, because it is being done on the virtual machine and reading data from the cloud storage).

In the Cloudera Data Hub there are already some pre-defined templates are available for creating the cluster this are known as blue-prints (terminology taken from public cloud), this helps you create cluster quickly. Cluster created using the Blue print can also be scaled based on the requirement.

# Chapter-10: Self-Service Experiences

## Introduction

Most of the time in the BigData world you work there are some pre-defined and standard patterns are followed and teams are using the same underline data for solving different problems. Let me give you few examples of such patterns.

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

When you or your team is working for the Machine Learning workload and you need the cluster for that requirement quickly. In the CDH world you have to use the same cluster and same resources which are shared by the other team and this can be a big issue. And even for that you need to regularly get contacted to the Hadoop Administrator for controlling the cluster resources or reserving cluster resources for your Machine Learning team and you don't feel comfortable for all this stuff.

## Self-service experiences

That's the reason Cloudera has come up with the concept of the Self-Service-Experience in this you can quickly spin up new cluster based on the type of work you are going to do. As of writing this book Cloudera was supporting below two self-service experiences

- Machine Learning
- Data Warehouse

And below are the planned for coming months

- Data Engineering
- DataFlow and Streaming
- Operational Database

As these three are not available as of now, it does not mean you can create workload for this requirement. You or your administrator team can leverage the Cloudera Data Hub solution for creating a cluster either using the pre-defined template or customizing the template. This even give you maximum control and flexibility in the CDP public/private cloud.

You can achieve this Self-Service-experiences very quickly because the Cloudera Data Platform (CDP-public cloud) heavily uses the disruptive technology known as containerization and orchestration with the Docker and Kubernetes and represent PaaS (Platform as a service) model.

Self-service experience helps in making your team productive quickly by providing access to the tools and data they need.

Even the control is in the Administrator’s hand for making sure that proper security and resource limit in place for reducing unnecessary cost and risk with the data.

Based on the service you select you or your team would also get the tools which they need, because Cloudera’s expert product design team have pre-selected the tools which are required for the type of workload you select. And your team only focus on the business logic rather then creating a new cluster based on their need and installing each component separately. This workload would also auto-scale if demand increased and if demand reduces then resources would be released to save the cost.

# Chapter-11: CDP Administrators

## Introduction

In case of CDP or CDH Hadoop cluster these the responsibility which needs to be taken care by the Administrators.

- Installing the Cluster
- Monitoring the Cluster
- Managing the Cluster

Now, having a CDP's different editions (deployment mode) administrator should know various things related to the CDP. Even they need to understand that both the editions underline nodes differ in case of public cloud these are virtual machines and in case of Data Center these are the bare metal servers. Following are the additional things which he has to do

- Managing and monitoring Hardware (Only in case of CDP Data Center)
- Managing the networking infrastructure (Only in case of CDP Data Center)
- Data Security and Access control (In both public cloud and CDP Data Center edition)

In case of public cloud there are no physical servers which administrator has to manage. And they more or less look for the accessing the resources and maintaining the connectivity with the cloud providers (If your company is using Direct Connect, VPN etc.)

We have till now talk about the cluster management but there is other part as well which is the data (asset for the company) and the administrator has to create backup for this data continuously.

## Operations

As part of administrator role, the main reasonability is the operations. Which include the following things

- Installing the Cluster
- Monitoring the Cluster
- Managing the Cluster

## Backup and Disaster recovery

Administrator certainly responsible for creating back of the data and in case of any disaster, he should be able to maintain the business continuity.

## Cluster planning

Administrator also involved in planning the cluster sizing and its performance in the defined cost involving IT and business users team. Even the person is an administrator he really need to understand the business requirement and technical requirement as well.  Proactively finding any possible risk with the data and finding the opportunities to reduce the cost of workloads. What all are cluster which are idle and unnecessary incurring the cost or any cluster which reaching its pick performance and may require more resources for continuously increasing the workloads.

- Whether he should move particular in-house workload to the cloud for increasing the performance and reducing the cost. Because it runs only for the few hours in a week.

## Technical skillsets

Administrator should have following technical skills who work with the CDP. You can check the presentation and a career guide created by HadoopExam Team for the System Admin Role at this link (premium subscription required)

- Linux System Administration
- Understanding the performance parameters for the underline Hardware. Choosing best configuration for the servers. It can also involve the knowing the detail about the instances or server provided by the public cloud and for in-house data center he needs to understand the physical hardware as well.
- Computer networking knowledge which includes TCP/IP, VPC, Subnets, IPV4, NACL, Firewalls, VPN, Switches, routers, route table entry creation, virtual appliances etc.
- Need to know how to secure the entire infrastructure
- Software based security using authentication and authorization.
- Configuration management and knowledge of the scripting languages like Bash, python etc. to create template-based deployment models.

To prove all this it is highly recommended that you do the Cloudera Administrator certification (CCA-131)

- It is also highly advised now that Cloudera CDP administrator have a good knowledge of at least one of the Cloud providers. (Get the training for AWS from this link)
- Administrator should know the cost aspect with the cloud provider services.
-