Analytics with CDP Private Cloud

Budapest Data Forum - September 2020

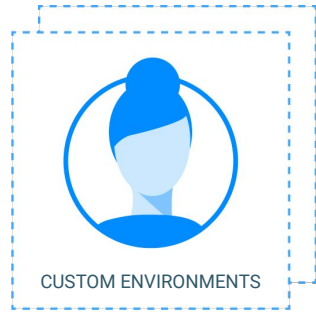Hemanth Yamijala, Director of Engineering
Cloudera

CLOUDERA

Agenda

- Introduction to CDP Private Cloud
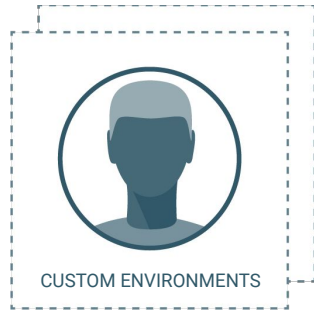- Architecture
- Benefits
- What Next

A NEW DATA PLATFORM FOR A TRANSFORMED CUSTOMER EXPERIENCE

Change from a Distribution of Hadoop to a Collection of Analytics 'Experiences'
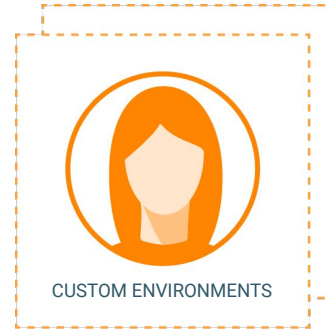
CUSTOM ENVIRONMENTS
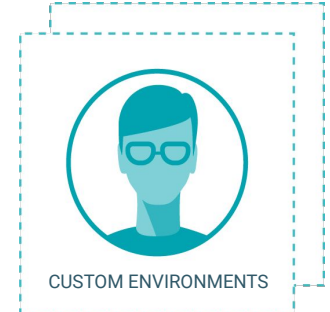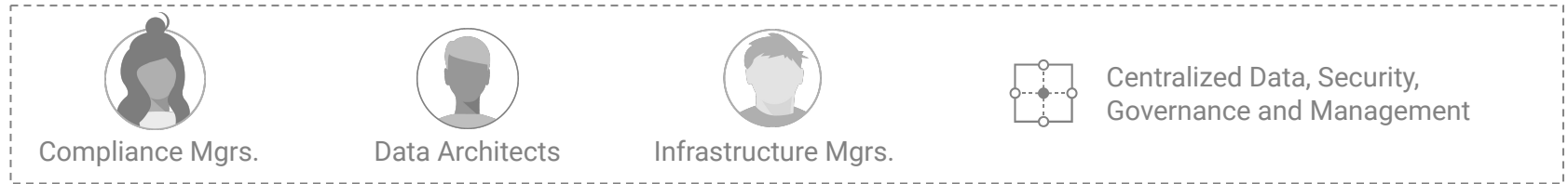
App Developers

CUSTOM ENVIRONMENTS

Data Engineers

CUSTOM ENVIRONMENTS

BI Analysts

CUSTOM ENVIRONMENTS

Data Scientists

Compliance Mgrs.

Data Architects

Infrastructure Mgrs.

Centralized Data, Security, Governance and Management

Components of CDP Private Cloud
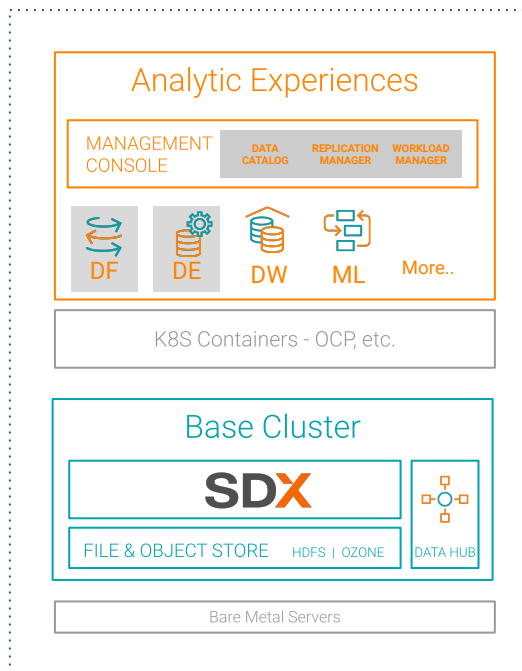
CDP Private Cloud Base

New name for the conventional Hadoop Distribution (CDH, HDP) **on bare metal hardware**

Combines the **best of CDH and HDP** services into one offering

Runs **metadata**, **security** and **governance** services that we call SDX

Provides **data storage** on HDFS and **Ozone**

Supports **conventional compute workloads** (Yarn, Spark in **Data Hub**)

Analytic Experiences

MANAGEMENT CONSOLE	DATA CATALOG	REPLICATION MANAGER	WORKLOAD MANAGER

DF   DE   DW   ML   More..

K8S Containers - OCP, etc.

Base Cluster

SDX

FILE & OBJECT STORE   HDFS | OZONE   DATA HUB

Bare Metal Servers

Grey boxes indicate services in the Roadmap

Analytical Experiences

New compute capabilities running on **K8S Containers**.

Support Analytics via **Cloudera Data Warehouse (CDW)** and Machine Learning workloads via **Cloudera Machine Learning (CML)**

A **Management Console** (also running on K8S) for managing the lifecycle of these compute experiences

Any number of these workloads can be launched for supporting **new apps**, **bursty workloads**

K8S (**currently OpenShift**) based infrastructure implies **better utilisation of resources**

Cloudera Data Warehouse

- Run Analytics services like **Apache Hive, Apache Impala, and Apache Hue** on Kubernetes
- **Dynamically provision** compute clusters with these services from the Management Console
- Compute clusters can be used to model **'data warehouses'** or **'data marts'**
- **Automatically connected to the Base cluster** for metadata, security & governance from there.
- **Auto-scale and auto-suspend** capabilities

Cloudera Machine Learning

- Run ML tools like **Apache Spark, Python and R** on Kubernetes
- **Dynamically provision** compute clusters with these services from the Management Console
- Compute clusters can be used for **running Notebooks (e.g. Jupyter), train and build models, run experiments and even schedule them** with a cron.
- **Automatically connected to the Base cluster** for metadata, security & governance from there.
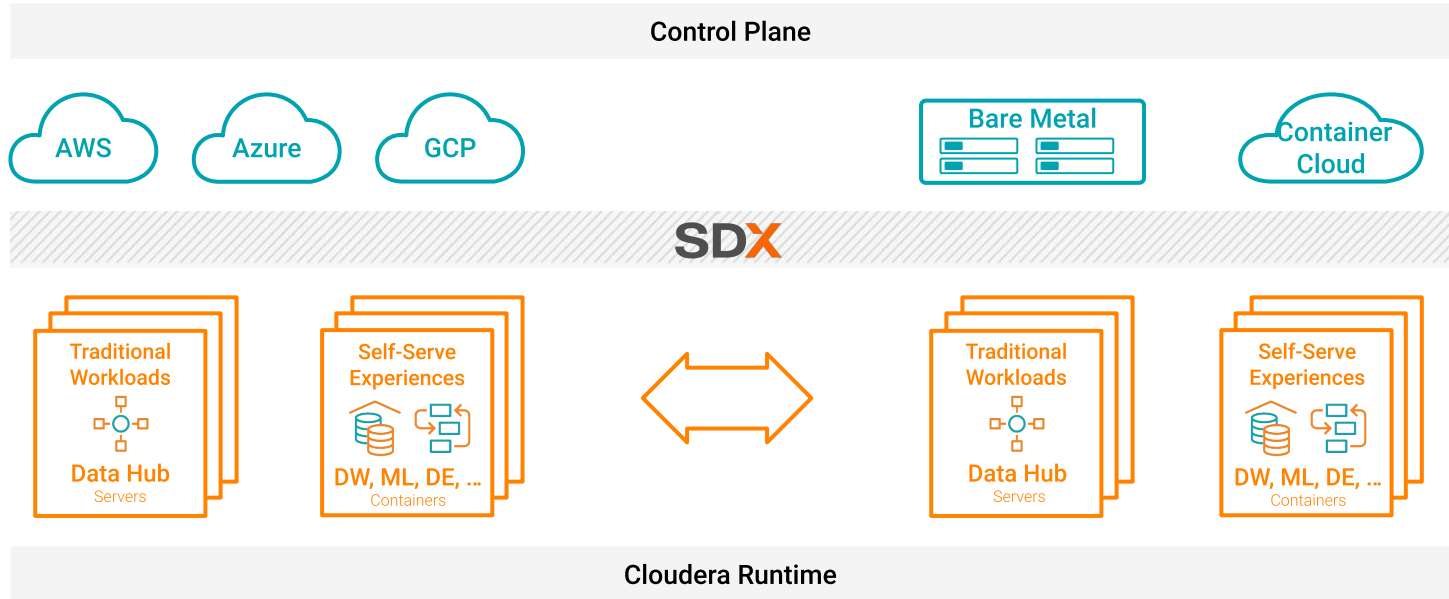- **Auto-scale and auto-suspend** capabilities

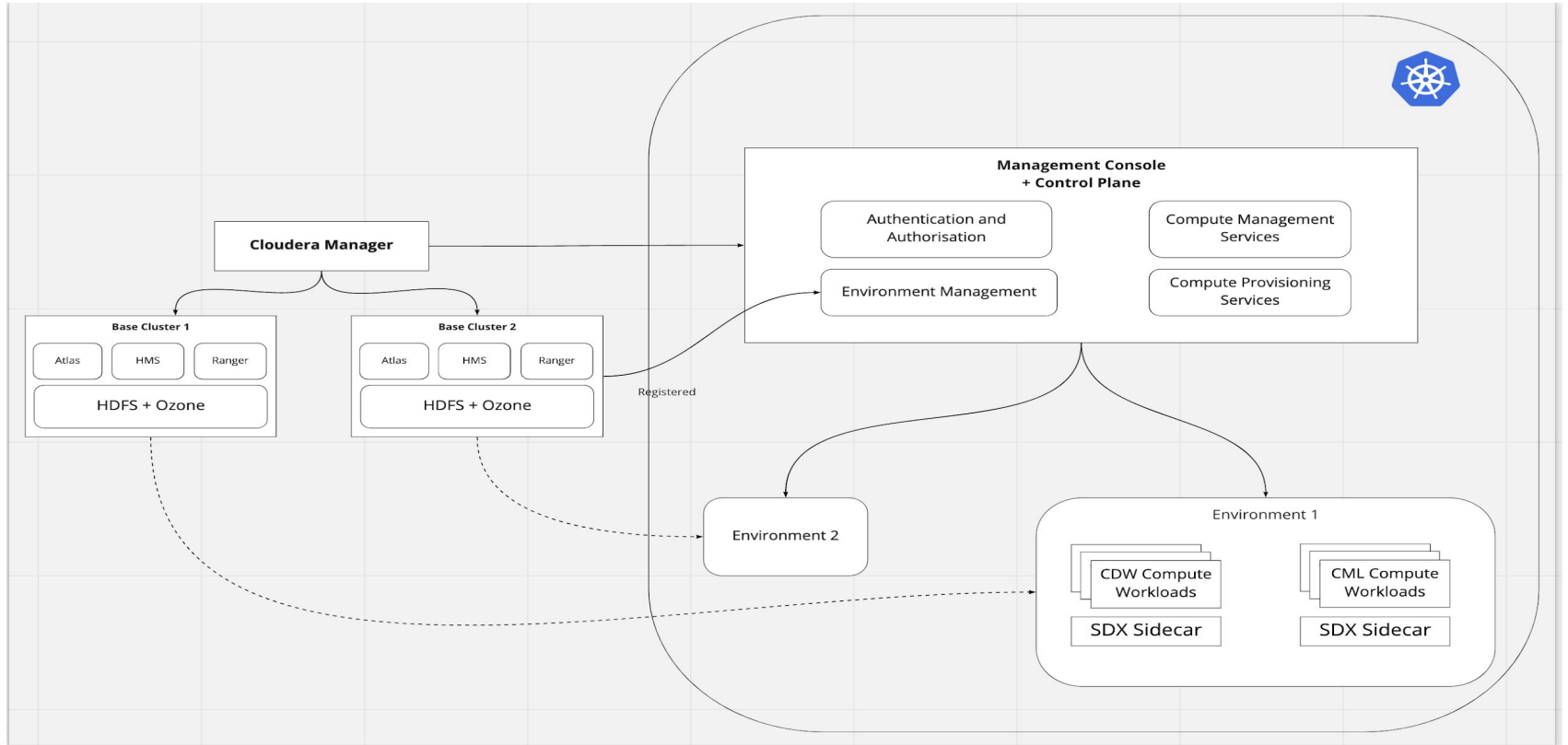**CLOUDERA**

ONE PLATFORM – TWO FORM FACTORS

CDP Public Cloud
(platform-as-a-service)

CDP Private Cloud
(platform as installable software)

Control Plane

AWS   Azure   GCP

Bare Metal

Container Cloud

SDX

Traditional Workloads

Data Hub
Servers

Self-Serve Experiences

DW, ML, DE, …
Containers

Traditional Workloads

Data Hub
Servers

Self-Serve Experiences

DW, ML, DE, …
Containers

Cloudera Runtime

Architecture of Private Cloud

Benefits of CDP Private Cloud

- Avoid noisy neighbors
  - *Launch **dedicated** analytics or machine learning compute workloads for each use case*
  - ***Contained workspaces and tools** for managing your projects distinctly from others.*
- Independent Upgrades
  - *Need the latest version of Spark? Or your own favorite Notebook? No need to upgrade the whole base cluster - just **recreate a new workload with what you need**.*
- Shared Security and Governance
  - ***Centrally administered** via SDX, **Enterprise LDAP support**, **Kerberos support** for standard Hadoop Data Security*
- Efficient utilization of resources
  - *Support for **scaling up and down of compute resources** automatically and on demand.*
  - ***Shared Kubernetes clusters** for all workloads.*

Next steps - Roadmap

- Expanded platform support
  - *Support for more recent versions of OpenShift, other K8s distributions, Databases, etc.*
- More experiences on Private Cloud
  - *Data Engineering, Data Flow, Data Catalog*
- Better administrative control
  - *Quotas for resource management, RBAC support, improvements to monitoring & supportability*
- Continuing improvements to existing experiences.

Resources

- [CDP Private Cloud Overview] https://www.cloudera.com/products/cloudera-data-platform/cdp-private-cloud.html
- [Cloudera Private Cloud Base] https://docs.cloudera.com/cdp-private-cloud-base/7.1.3/index.html
- [Private Cloud Management Console 1.0] https://docs.cloudera.com/management-console/1.0/index.html
- [Cloudera Data Warehouse] https://docs.cloudera.com/data-warehouse/1.0/index.html
- [Cloudera Machine Learning] https://docs.cloudera.com/machine-learning/1.0/index.html

CLOUDERA

Thank you!