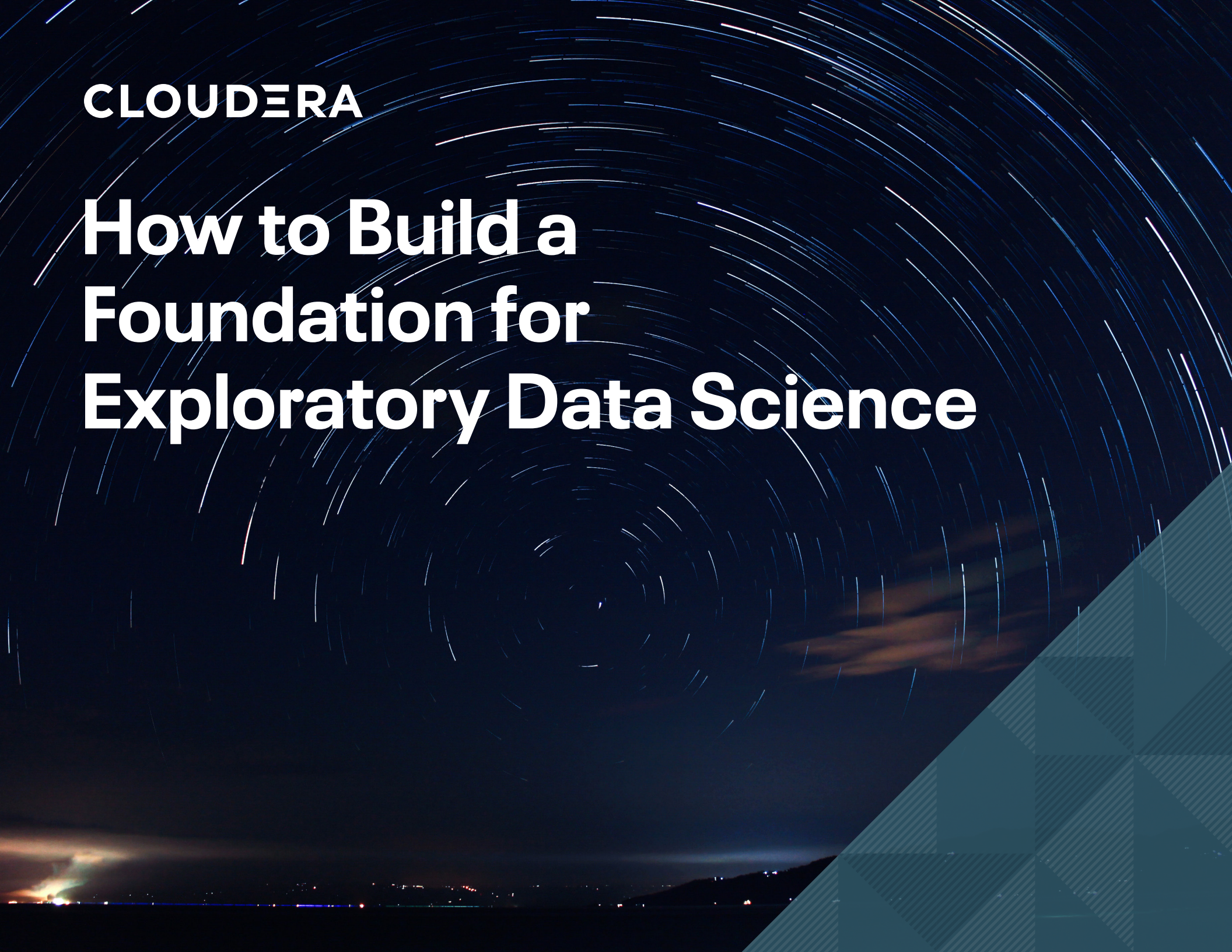

CLOUDERA
How to Build a Foundation for Exploratory Data Science

Table of Contents

Introduction: Emerge from the Data Swamp

You have to start somewhere. Before you can deploy Artificial Intelligence (AI) and Machine Learning (ML) applications across your enterprise, you have to go through a period of data exploration.

This data exploration, while valuable, often begins as a series of loosely organized initiatives. Beyond just facilitating data science, your goal should be to compress the space between data exploration and business action. Only then will your enterprise benefit from data-driven insights that drive decisions.

While most enterprises are now using AI, many find it challenging to fully integrate it into their business. In a recent survey, a vast majority of organizations—92.1%—said they are achieving measurable results from their data and AI investments. But just 27% said they've created a data-driven organization.[1]

If your data initiatives are treated as one-offs and disconnected from one another, you may be in that situation of valuing data but struggling to use it. Your challenge is to build a truly data-driven organization, rather than one mired in a "data swamp."

Your vision for the future should include a data platform that fosters trust and allows users to easily assess the validity of the available corporate data assets. Your data teams are looking for "certified datasets," as well as consistent and robust tooling to make data exploration, ad-hoc data science, and insight generation as fast as possible.

The good news is that getting AI and ML off the ground doesn't have to be challenging. This eBook will help you understand what's ahead and what's possible with exploratory data science.

1 NewVantage Partners, "2022 Big Data and AI Executive Survey"

Chapter 1: Deliver the Tools Data Practitioners Need

It's easy for exploratory data science to tip into a precarious juggling act, as users find themselves toggling among various disconnected tools and environments. Early on in your efforts it pays off to consider how data practitioners can work in ways that support one another and deliver results with impact.

Data scientists

Data scientists need to collaborate while maintaining end-to-end control and visibility. Generating outcomes through unified data workflows ensures the outcomes generated are accurate — and can be trusted across the business.

This means that flexibility in the use of ML tools is important for successful data science work. Data scientists may favor notebook environments for some kinds of work, but SQL for others. A unified data platform lets data scientists perform data discovery work using the tools of their choice.

Essential for these goals are high-quality data pipelines with verified sources of data.

Data engineers

Data engineers need to create reliable data pipelines to support the work of data scientists and others. Improved tooling for data pipelines help data engineers improve reliability, by making sure current data is readily available at scale.

Additionally, improved ETL and data orchestration tools can help connect data seamlessly to data science notebooks, for practitioners who prefer them. Solutions that can be deployed in containerized environments help deliver data quickly and seamlessly to stakeholders who need it.

Must-haves for data practitioners

Data scientists

Unified data workflows

Trusted sources of data

Flexibility in choice of tools

Containerized environments

Data engineers

Data pipeline tools and environments

Enterprise orchestration tools

Containerized environments

Cloudera's solutions

Cloudera Data Platform (CDP) with Cloudera Machine Learning (CML) is a unified solution that can deliver a foundation for exploratory data science quickly and easily. Data scientists benefit from a single UI, from point-and-click data discovery to an SQL editor to visualizations.

Cloudera Data Engineering (CDE) offers a completely containerized and managed Apache Spark service that serves the needs of data engineers with an easy-to-use interface. Fully integrated with CDP, CDE includes Apache Airflow, which enables data engineers to quickly orchestrate and automate complex data pipelines.

Cloudera Data Warehouse (CDW) can support a self-service analytic experience by making it fast and cost-effective to query all sizes and types of data.

Chapter 2: Gain Flexibility through Containerized Runtimes

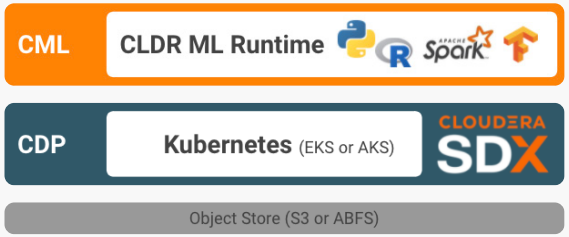
Data science requirements differ from team to team and project to project—so much that any environment that attempts to serve every use case would be bloated and inflexible. A better option is to offer lightweight, isolated environments with granular customization.

The ability to run variants of Python, R, Workbench, JupyterLab, and other common ML environments in a streamlined fashion can help maximize flexibility. Data scientists can also benefit from variants that support GPU acceleration with all the tools and libraries needed to take advantage of those performance gains.

CML's solution is Runtimes, docker container images for these ML environments, along with the necessary infrastructure to create, maintain, and manage them.

CML users access Runtimes through a catalog that lists information about available Runtimes, including the editor and kernel supported by the Runtime along with the edition, version, and a brief description.

CML Runtimes are Containerized Environments for ML



Runtimes reduce the time it takes for data practitioners to deploy environments with direct access to secure data and compute resources. They're customizable and self-service, letting users provision IDEs, libraries, frameworks, algorithms, and more without needing to engage IT. In short, it's easier to get to work and deliver results.

Chapter 3: Turn Data Science into Action with Data Visualization

At the exploratory stage, data visualizations help users not only improve their understanding of data, but also share what they've learned. Whether you're communicating insights to other data practitioners, or to the wider business, you want visual tools available in this development step to eliminate knowledge gaps and foster collaboration.

Today's data visualization capabilities go beyond charts and graphs. They offer access to advanced analytical predictions to users throughout the business. A modern data visualization solution supports:

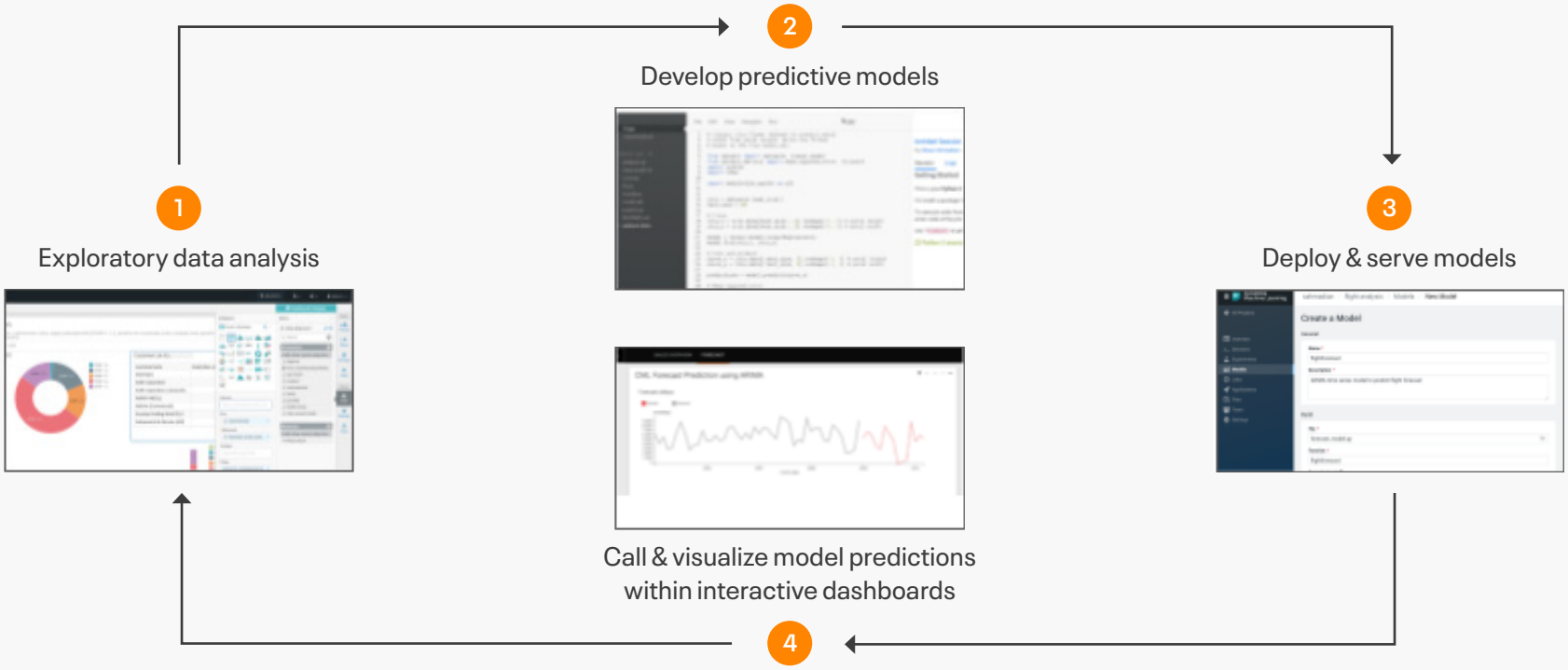
Sharing insights everywhere. The ability to set up and share dashboards easily means data gets to the right people faster.

Automating intelligent reporting. Everyone can have access to the latest insights with scheduled updates, emailed reports, and dynamic alerts.

Building predictive applications. Visualizations built on machine learning models can enable everyone to access the value of predictive analytics.

Cloudera Data Visualization (CDV), which is fully integrated with CDP, let's data practitioners quickly create and share visual objects with easy-to-use web-based tools. It allows data scientists to share, and business users to interact with the predictions from the models via interactive dashboards. Users can share the same visualization tools across the platform, connecting the dots between different teams and ideas. Even at the exploratory stage, getting insights into view can justify the value of data science programs and lead to ongoing support for more.

Visualizing ML



Chapter 4: Manage the Risks of Security or Governance

Data security and governance must be properly enforced in exploratory data science. Data regulations require enterprises to control how data is processed and who has access to it. Compliance failures can lead to fines, penalties, and reputational damage.

If users are forced to manage several disconnected solutions for exploration, or go rogue and spin up their own environments unknown to IT security, you can end up with security and compliance blind spots. Risks also go up when data practitioners are moving data from place for the purposes of exploratory analysis.

To keep compliance and security risks to a minimum, data users need the ability to analyze datasets in place, using data of known lineage. As a fully integrated part of Cloudera Data Platform, the Shared Data Experience (SDX) provides automatic model cataloging and lineage, along with governed and secure production workflows.

SDX runs independently of compute and storage layers, and offers an integrated set of security and governance technologies built on metadata. This means consistent data context through automatic model cataloging and lineage, along with governed and secure production workflows.

CML Success Story: IQVIA

IQVIA, a global life sciences technology solutions provider, relies on CDP for collaborative analytics on sensitive clinical trial data. A R&D data science team collaborates with a pharmaceutical organization on ML algorithms to predict site capacity for supporting clinical trials. Both sets of teams can collaborate in a confidential, secure environment using highly protected data in a private cloud managed by IQVIA. With SDX, security, privacy, governance, and encryption are built into the platform.

Read more here

Conclusion: Get the Benefits of a Single Platform

In exploratory data science, data practitioners might gravitate toward several different disconnected tools. But using a patchwork of solutions makes it harder to work efficiently, foster trust in the data, share results visually, and maintain security and compliance.

Data exploration is far more effective with the fast, easy-to-use, and unified tools offered through CDP and CML and a data warehouse like CDW. Instead of toggling between different frameworks, trusted data and the optimal tools for analysis are in a single platform. Agility and self-service access help meet user needs.

CML offers integrated data connections, making it simple to turn connections on and off and pull in data from anywhere and use it in any project. An easy-to-use UI allows users to connect to data no matter where it is, complete with code snippets for data practitioners to use. CML includes a data tab that allows users to work within one file, where they can save queries, manipulate data, and visualize data.

All these tools work together to help exploratory data science practitioners work efficiently, deliver results quickly, and drive business outcomes.

Take Your Next Step

Discover how Cloudera Data Platform can enable data scientists and engineers to discover valuable new uses for AI and ML, while collaborating on a single trusted platform.

Read more

About Cloudera

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises.

Learn more at cloudera.com | US: +1 888 789 1488 | Outside the US: +1 650 362 0488

Sources

1 NewVantage Partners, "2022 Big Data and AI Executive Survey"

CLOUDERA

For the most optimized experience leverage AMD CPUs on Dell hardware

Cloudera Data Platform Private Cloud Base

Pod Network:
PowerSwith S5248F-ON series switch

Cluster Aggregation Network:
PowerSwitch Z9432F-ON series switch

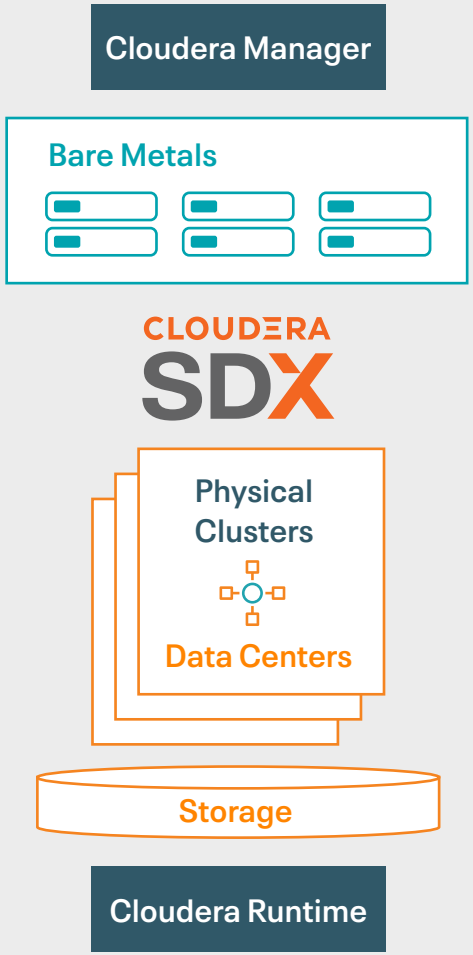
Infrastructure Nodes:
PowerEdge R6515
(3) Master nodes
(1) Utility node
(1) Edge node

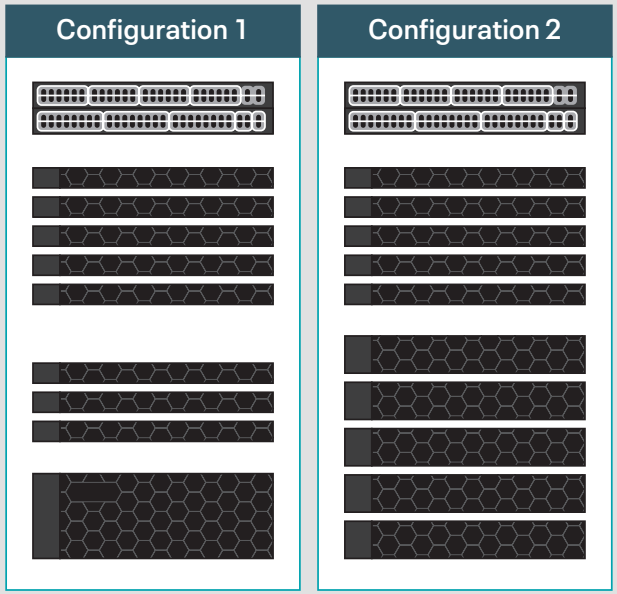
(3+) Worker Nodes:
PowerEdge R6515 (Configuration 1)
or PowerEdge R7515 (Configuration 2)

GPU Accelerated Worker Node Option:
PowerEdge R7525

HDFS:
Powerscale H5600 (Configuration 1)
or Additional Worker Nodes (Configuration 2)

CDP Data Center
Installable Software





Independent Compute & Storage

Combined Compute & Storage

RECOMMENDED

AMD

DELL