Cloudera AI

Cloudera AI Product Overview

Date published: 2020-07-16
Date modified: 2025-06-06

CLOUDERA

Legal Notice

Contents

Cloudera AI overview

Machine learning has become one of the most critical capabilities for modern businesses to grow and stay competitive today. From automating internal processes to optimizing the design, creation, and marketing processes behind virtually every product consumed, machine learning models have permeated almost every aspect of our work and personal lives.

Machine learning development is iterative and complex, made even harder because most machine learning tools are not built for the entire machine learning lifecycle. Cloudera AI accelerates time-to-value by enabling data scientists to collaborate in a single unified platform that is all inclusive for powering any AI use case. Purpose-built for agile experimentation and production machine learning workflows, Cloudera AI manages everything from data preparation to MLOps, to predictive reporting. You can solve mission critical machine learning challenges along the entire lifecycle with greater speed and agility to discover opportunities which can mean the difference for your business.

Each Cloudera AI Workbench enables teams of data scientists to develop, test, train, and ultimately deploy machine learning models for building predictive applications all on the data under management within the enterprise data cloud. Cloudera AI Workbench support fully-containerized execution of Python, R, Scala, and Spark workloads through flexible and extensible engines.

AI applications

Analytical Applications provide a place to host long running applications within a Cloudera AI project.

While Cloudera AI offers a place for Data Scientists to perform advanced analytics and models into production, Analytical Applications provides a place to host long running applications within a Cloudera AI project. This opens up a larger group of users to the insights discovered in Cloudera AI. Applications can be built with a variety of frameworks like Flask and Streamlit. They run within their own isolated compute instance which keeps them from timing out and they take advantage of ML Runtimes. Applications are accessible to users through the web. Applications can be for a variety of use cases like hosting interactive data visualizations or providing a UI frontend for a deployed mode in Cloudera AI.

Exploratory Data Science

Cloudera AI enables data practitioners to discover, query, and easily visualize their data sets all within a single user interface.

The beginning of every data science project begins with finding and understanding the data you need. Cloudera AI brings all the tools you need for exploratory data analysis together in a single UI so that data practitioners don't have to jump between applications, and IT does not have to support multiple tools. Cloudera AI provides users with a view of available data assets that they can connect to, a SQL editor to query those data sources, and an easy-to-use drag-and-drop visualization tool to understand your data and communicate insights.

ML Ops

Cloudera AI enables users to deploy machine learning and other models into production.

Cloudera AI enables users to deploy machine learning and other models into production, either as a batch process through the Jobs functionality, or as near-real-time REST APIs using the Models functionality. In addition, Cloudera AI provides a number of features to help maintain, monitor and govern these models in production. The Model Governance feature ensures that every deployed Model is tracked in the Cloudera Data Catalog, and allows the user to specify which data tables were used to train the model in order to provide model-data lineage. Deployed Models have a built-in dashboard for monitoring technical metrics relating to deployed Cloudera AI Models, such as request throughput, latency, and resource consumption. Additionally, users can track arbitrary business metrics relating to each inference event, and match the results with delayed metrics from a data warehouse or from other source using an

automatically generated UUID. By analyzing these metrics, the user can assess the model for aggregated metrics such as accuracy on an ongoing basis.

Cloudera AI Inference service provides a production-grade serving environment for hosting predictive and generative AI. It is designed to handle the challenges of production deployments, such as high availability, performance, fault tolerance, and scalability. The Cloudera AI Inference service allows data scientists and machine learning engineers to deploy their models quickly, without worrying about the infrastructure and maintenance.

Core capabilities

This section details the core capabilities for Cloudera AI.

Cloudera AI covers the end-to-end machine learning workflow, enabling fully isolated and containerized workloads - including Python, R, and Spark-on-Kubernetes - for scale-out data engineering and machine learning with seamless distributed dependency management.

- Sessions enable Data Scientists to directly leverage the CPU, memory, and GPU compute available across the Cloudera AI Workbenches, while also being directly connected to the data in the data lake.
- Experiments enable Data Scientists to run multiple variations of model training workloads, tracking the results of each Experiment in order to train the best possible Model.
- Models can be deployed in a matter of clicks, removing any roadblocks to production. They are served as REST endpoints in a high availability manner, with automated lineage building and metric tracking for MLOps purposes.
- Jobs can be used to orchestrate an entire end-to-end automated pipeline, including monitoring for model drift and automatically kicking off model re-training and re-deployment as needed.
- Applications deliver interactive experiences for business users in a matter of clicks. Frameworks such as Flask and Shiny can be used in development of these Applications, while Cloudera Data Visualization is also available as a point-and-click interface for building these experiences.
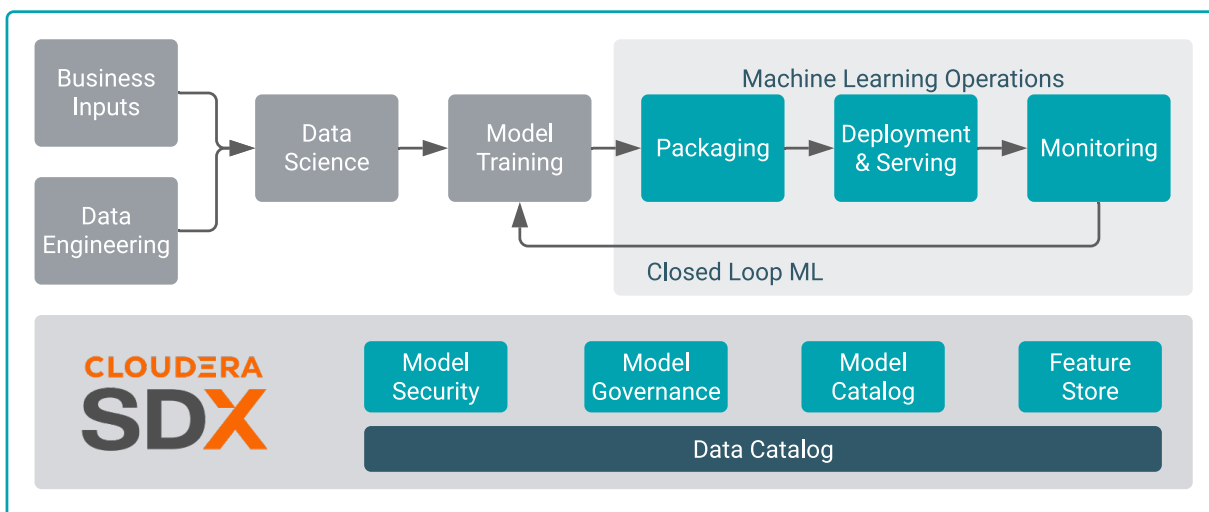
Cloudera AI benefits

This section details the Cloudera AI benefits for each type of user.

Cloudera AI is built for the agility and power of cloud computing, but is not limited to any one provider or data source. It is a comprehensive platform to collaboratively build and deploy machine learning capabilities at scale.

Cloudera AI provides benefits for each type of user.

Data Scientists

- Enable Data Service teams to collaborate and speed up model development and delivery with transparent, secure, and governed workflows
- Expand AI use cases with automated Cloudera AI pipelines and an integrated and complete production Cloudera AI toolkit
- Empower faster decision making and trust with end-to-end visibility and auditability of data, processes, models, and dashboards

IT

- Increase data service productivity with visibility, security, and governance of the complete machine learning lifecycle
- Eliminates silos, blindspots, and the need to move/duplicate data with a fully integrated platform across the data lifecycle
- Accelerate AI with self-service access and containerized Cloudera AI Workbench that remove the heavy lifting and get models to production faster

Business Users

- Access interactive Applications built and deployed by Cloudera teams
- Are empowered with predictive insights to support better business decisions

Key differences between Cloudera AI (Cloudera on cloud) and Cloudera Data Science Workbench

This topic highlights some key differences between Cloudera Data Science Workbench and its cloud-native counterpart, Cloudera AI.

How is Cloudera AI related to Cloudera Data Science Workbench (CDSW)?

Cloudera AI expands the end-to-end workflow of Cloudera Data Science Workbench (CDSW) with cloud-native benefits like rapid provisioning, elastic autoscaling, distributed dependency isolation, and distributed GPU training.

It can run its own native distributed computing workloads without requiring a separate CDH cluster for scale-out compute. It is designed to run on Cloudera in existing Kubernetes environments, reducing operational costs for some customers while delivering multi-cloud portability. On Cloudera on cloud, managed cloud Kubernetes services include EKS, AKS, or GKE, and on Cloudera on premises, they include Red Hat OpenShift Container Platform or Cloudera Embedded Container Service.

Both products help data engineers and data science teams be more productive on shared data and compute, with strong security and governance. They share extensive code.

There is one primary difference:

- CDSW extends an existing CDH cluster, by running on gateway nodes and pushing distributed compute workloads to the cluster. CDSW requires and supports a single CDH cluster for its distributed compute, including Apache Spark.
- In contrast, Cloudera AI is self-contained and manages its own distributed compute, natively running workloads - including but not limited to Apache Spark - in containers on Kubernetes.

  Note: It can still connect to an existing cluster to leverage its distributed compute, data, or metadata (SDX).