Cloudera AI Release Notes

Date published: 2020-07-16
Date modified: 2025-12-18

https://docs.cloudera.com/

Legal Notice

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

What's New

Major features and updates for the Cloudera AI data service.

December 17, 2025

Release notes and fixed issues for version 2.0.53-b273.

New Features / Improvements

Cloudera AI Workbench

- Added a new and dedicated Cloudera AI Overview page in the Control Plane. This page now serves as the primary landing page, replacing the previous Workspace Administration panel view. (DSE-43180)
- Model endpoints now support the configuration of vetted environment variables using the UI and API. This enables specialized use cases, such as toggling reasoning output for NVIDIA NIMs. (DSE-45592)

Cloudera AI Platform

- Added support for the AWS Bahrain region. (DSE-46039)
- Added support for the Azure Spain region. (DSE-44839)
- Added monitoring for metadata reporting, including the tracking of the workspace version (appVersion). (DSE-32676)

Cloudera AI Inference service

- Added Non-Transparent Proxy (NTP) support. (DSE-39437)

Fixed Issues

Cloudera AI Registry

- Upgraded NIM-CLI to the latest version to incorporate improvements in memory management and error handling when downloading large models. (DSE-46374)
- All critical and high-severity CVEs in Cloudera AI Registry have been addressed.
- The Registered Models UI now loads successfully when using Google Chrome version 142 or above. (DSE-48824)

Cloudera AI Inference service

- The Model Endpoints UI now loads successfully when using Google Chrome version 142 or above. (DSE-48824)

Cloudera AI Workbench

- Resolved a critical race condition in the reconciler that caused successfully completed jobs to be incorrectly marked as failed with a -1 exit code. (DSE-48076 and DSE-44901)
- The Spark UI button is now correctly displayed in the dropdown menu for Jupyter Notebooks within the Cloudera AI/ML environment. (DSE-42562)
- Resolved an issue where users with the Team Admin role were unable to update team descriptions due to a permission mismatch. The system now correctly aligns the required permission levels, allowing team creators to manage their team settings as expected. (DSE-48770)

Behavioral Change

Cloudera AI Workbench

- Increased file transfer timeouts to 10 minutes for uploads and 15 minutes for downloads, ensuring reliable API v2 transfer of large files. (DSE-39671)

Deprecation Notice

Cloudera AI Inference service

- NVIDIA optimized profiles for the following models are no longer supported and have been removed from the Model Hub in this release:
  - Llama 3.2 11B
  - Llama 3.2 90B
  - Mixtral 8x22B

  Although these optimized profiles are no longer available in the Model Hub, the models remain supported and accessible through Hugging Face.

October 31, 2025

Release notes and fixed issues for version 2.0.53-b241.

New Features / Improvements

Model Hub

- Improved model information display by surfacing attributes like MaxTokens, Parameters, and Dimensions for Embedding Models, enabling better decision-making before importing models.

Cloudera AI Inference service

- Added the ability for users to start and stop deployed model endpoints, providing greater control over resource management and cost optimization. This feature allows you to pause inactive endpoints to save resources while maintaining the ability to quickly restart them when needed.
- Improved user experience by enabling customers to open model endpoints in different browser tabs, allowing for better multitasking and simultaneous monitoring of multiple endpoints.
- Enhanced API accessibility by providing customers access to the Swagger UI for invoking AI Inference APIs, enabling interactive API testing and documentation exploration.
- Added visibility to the underlying compute cluster through a direct link in the UI for AI Inference instances, providing seamless navigation to cluster details and infrastructure monitoring.
- Implemented validation for root volume size when adding new nodes to AI Inference deployments, preventing configuration errors and ensuring adequate storage capacity.
- Enhanced node group management by displaying GPU types and enabling filtering based on GPU type when searching through available nodes, improving resource selection efficiency.
- Upgraded KServe from version 0.12 to 0.15, enhancing the underlying model serving infrastructure.

Cloudera AI Platform

- Enhanced support for self-signed certificates in public cloud deployments, resolving installation and update pain points.
- Re-enabled cadence workflow for workbench upgrades, fixing version compatibility issues.

Cloudera AI Workbench

- Optimized the loading speed of the Site Administration overview and the Projects pages by improving cluster-wide resource usage data collection, ensuring quick loading even in environments with 1,000+ user namespaces.
- The Job Retry feature introduces automated recovery for failed, timed out, or skipped jobs. Administrators can now set concurrent retry limits and customize the retry behavior to enhance job resilience and eliminate the need for manual failure intervention.

Fixed Issues

Cloudera AI Registry

- Fixed Events & Logs rendering during the creation of new registries in the UI. (DSE-48101)

Cloudera AI Inference service

- Resolved inconsistencies in the tooltips displayed on the Model Endpoint Details page, ensuring accurate and helpful information is shown to users. (DSE-45042)
- The API service can now inspect the metadata of MLflow models from the Cloudera AI Registry to correctly identify the Hugging Face transformer flavor. This ensures the model is automatically routed to the correct runtime (i.e. HuggingFace runtime with vLLM backend unlike defaulting to Triton + ONNX earlier), enabling successful deployment of MLflow transformer models. (DSE-46462)

Cloudera AI Workbench

- Removed the rigid requirement for every team to have an Administrator, allowing Team Administrators (or Site Administrators if none are designated) to manage team membership and roles. (DSE-41700)
- You can now customize the model build process with two new options: model_root_dir allows setting a custom source directory, and build_script_path enables specifying a custom location for the build script. (DSE-45166)
- Resolved a page layout issue that occurred after horizontal window resizing. The fix ensures the browser view restores correctly without requiring a manual refresh. (DSE-40393).

MLRuntimes

- Fixed an issue where system messages were hidden or suppressed when a Workload (such as a Session, Job, or Application) failed to start up while using a PBJ Workbench Runtime. (DSE-46126)

Behavioral Change

MLRuntimes

- The R kernel output mechanism has been updated to improve visibility of interactive commands. The output of R commands (for example, install.packages("sparklyr")) is now mirrored to both the Session tab and the Logs tab by default, rather than appearing solely in the Logs. This ensures immediate visibility of important messages like installation progress, compilation output, and errors. Users can revert to the previous behavior by setting the environment variable JUPYTER_KERNEL_OUTPUT_FILTER_REGEX to "DISABLED" or "^$". (DSE-35335)
- Clearing a session now permanently removes all content from the session history, ensuring that subsequent PDF exports and email reports contain only current, relevant results and no stale data from previous runs. (DSE-12565)

Cloudera AI Inference service

- MLFlow integration with transformer models is now complete, enabling models saved using the workbench (via MLFlow) in the Hugging Face transformer format to be successfully deployed on Cloudera AI Inference service, including both regular and finetuned models. (DSE-43840)

September 25, 2025

Release notes and fixed issues for version 2.0.52-b<TBA>.

New Features / Improvements

Cloudera AI Registry

- You can now update the visibility of models directly from the UI.
- You can now delete registries in a failed status directly from the UI.
- Added the ability to use pre-downloaded artifacts when importing external models.
- The model card for embedding models now includes MaxTokens, Parameters, and Dimensions.
- Added support for upgrading Azure registries with UDR-enabled subnets.
- Improved error resolution suggestions have been added to the UI for Cloudera AI Registry, empowering users to troubleshoot issues on their own.

Cloudera AI Inference service

- Added support for new models, including NeMo Retriever’s GraphicElements, PageElements, and TableStructure, as well as PaddleOCR, Boltz-2, and GPT-OSS models.
- The Test Model feature is now available for Speech-to-Text models.
- You can now specify vLLM arguments when deploying Hugging Face models.
- Improved error resolution suggestions have been added to the UI for Cloudera AI Inference service empowering users to troubleshoot issues on their own.
- Grouped CPU/GPU nodes under the Cloudera AI Inference service Details page for a more user-friendly experience.

Cloudera AI Platform

- You can now provision multiple CPU/GPU resource groups in the Cloudera AI Workbench. This provides enhanced control over workload scheduling and allows segregation of workloads based on the instance types.
- Added support for EKS 1.32.
- Added support for AKS 1.32.

Fixed Issues

Cloudera AI Inference service

- Addressed critical CVEs in most NGC models by upgrading to the latest NIM versions. (DSE-47435)
- Resolved an issue where code samples were not correctly rendered for riva, reranking, and retrieval models. (DSE-46413)
- Ensured that the UI carries out proper validation of GPU, CPU, and memory during model deployment. (DSE-46250)
- Resolved an issue where Prometheus errors were not surfaced in the Model Endpoints UI. (DSE-46077)
- Addressed an accessibility issue with the 'Deploy external dropdown button'. (DSE-46071)
- Resolved an issue that blocked navigation to other tabs while waiting for a response from the Model Endpoint's Test Model. (DSE-46247)
- Resolved an issue where a scale alert box was not displayed in the UI when the model endpoint was ready and the current replica count was zero. (DSE-46301)

Cloudera AI Registry

- Resolved an issue where the UI did not properly surface failures occurring during NVIDIA model imports from the Model Hub page. (DSE-45971)
- Resolved an issue where the UI was not correctly auto-selecting the running registry. (DSE-46246)
- Resolved an issue that prevented users from registering models from within a workbench. (DSE-47077)

Cloudera AI Workbench

- Resolved an issue where proxy environment variables were being overwritten during model builds. (DSE-46070)
- Resolved an issue where the Project Settings page displayed the project owner's ID instead of their name. (DSE-46572)

Cloudera AI Platform

- Resolved an issue where spaces were incorrectly added to the `ldap_dn` during synchronization of some users, which caused those members of CML groups to be unintentionally removed. (DSE-47591)

Older releases

Overview of new features, enhancements, and changed behavior introduced in earlier releases of Cloudera Machine Learning.

August 22, 2025

Release notes and fixed issues for version 2.0.52-b34.

New Features / Improvements

ML Runtimes

- Released new Spark Runtime Addons versions - 2.4.8, 3.2.3, 3.3.0, and 3.5.1.
- Spark 3.5.1 in Cloudera AI is now certified to work with the Data Lake 7.3.1 version.
- New HadoopCLI Runtime Addon versions 7.2.18 and 7.3.1 are now available. Versions 7.2.16 and 7.2.17 have been removed.

Fixed Issues

Cloudera AI Registry

- Resolved an issue that blocked AI Registry creation in regions other than us-west-2. (DSE-46535)
- Resolved an issue that prevented registries in renew:finished status from being used by Cloudera AI Inference service. Upgrading Cloudera AI Inference service is required to use these registries. (DSE-46815)
- Previously, the UI would flicker when trying to access the Registered Models UI due to an issue. This issue is now resolved. (DSE-46714)
- Resolved a critical bug that caused cdh-client to restart in a workbench, which prevented sessions from starting. (DSE-46989)

Cloudera AI Workbench

- Previously, users were unable to register models from within Cloudera AI Workbenches. This issue is now resolved. Upgrading workbenches is required to register models from within Cloudera AI Workbenches . (DSE-46656)

July 31, 2025

Release notes and fixed issues for version 2.0.52-b27.

New Features / Improvements

Cloudera AI Platform

- When an ML workload is suspended, its ml-infra autoscaling range is now automatically set to 0-0, which allows for better cost optimization.
- Added support for file storage replication in AWS EFS (Elastic File System), enhancing data redundancy and availability. For information, see Configuring File Storage Replication on AWS.
- Improvements have been made to enable retriable in-place upgrades, leading to more robust upgrade processes.
- For information, see Upgrading Cloudera AI Workbenches.
- Added support for Azure regions: Poland Central and Italy North.
- Added support for Istio for Cloudera Control Plane.
- Added support for Customer-Managed Key (CMK) encryption in Cloudera AI Azure workbenches. For information, see Enabling Customer Managed Keys on Microsoft Azure.
- You can now change Persistent Volume Claim (PVC) sizes through both the UI on the Workbench Details page and using the CDP CLI. For information, see Modifying workbench persistent volume size.
- Added a liveness probe to the mlx crud application pod and implemented graceful shutdown, improving the stability and resilience of the application.
- The system will now automatically select the Cloudera AI Registry if only one instance exists within a given tenant.

- Improvements have been made to the UI, allowing you to create Cloudera AI Registries with private clusters and enable User-Defined Routing (UDR) more easily.
- Added user-friendly information to the UI to assist users when utilizing the air-gapped model hub import functionality.

Cloudera AI Registry

- Added CSI driver support for AI Registries, removing previous resource constraints. You can now download any number of models in parallel without encountering resource limitations with Azure.
- User-friendly and informative error messages are now displayed when users are unable to import a model to the AI Registry.
- A new caching mechanism has been introduced in Model Hub, significantly reducing the time it takes for pages to load.

Cloudera AI Inference service

- Added support for the Nemotron Super 49B model.
- Added support for Riva ASR NIM (NVIDIA Inference Microservice), enabling advanced automatic speech recognition. This feature is compatible with the Whisper mode, requiring a 16-bit, mono, 16000 Hz, uncompressed WAV file as input.
- Added support for several new vLLM load formats, including sharded_state, gguf, bitsandbytes, mistral, runai_streamer, and fastsafetensors. This enhances the list of supported vLLM quantization options.
- Nemotron's `thinking mode` is now user-configurable, allowing you to explicitly activate this advanced reasoning capability by including "content": "detailed thinking on" within the system role of your prompt payload, giving you precise control over resource usage.
- Implemented necessary validators for GPU instance types during the deployment of NVIDIA models to prevent misconfigurations.
- Significantly improved the performance of Cloudera AI Inference service by caching tokens to improve UI responsiveness and decrease network load.
- The replica for endpoint logs and events is now automatically selected for any given model endpoint.
- Added a Refresh button to various sub-sections of the model endpoint details page for easier data updates.
- A force fetch button is now available on the Model Hub UI for users to override cached values and ensure the latest data is displayed.
- Replaced generic Failed to Fetch messages with more user-friendly error messages when a user attempts to import a Hugging Face model not present in our Model Hub.
- An alert box is now displayed in the UI to notify users when an Ingress Ready endpoint has a replica count of 0.

ML Runtimes

- Resource requests for several core Cloudera AI services have been increased. This change is designed to boost performance and stability, ensuring a smoother experience without requiring any action on your part.

Fixed Issues

Cloudera AI Platform

- Resolved an issue that prevented Cloudera AI Registries from being visible in the control plane after their certificates were renewed. (DSE-44836)
- Previously, the configuration map of Cloudera AI Inference service was not updating correctly during the upgrade process. This issue is now resolved. (DSE-45417)

Cloudera AI Registry

- Previously, constraints in the UI prevented upgrading an AI Registry already in a Ready status. This issue is now resolved. (DSE-45663)

Cloudera AI Workbench

- Addressed an issue due to which registries whose certificates were renewed were not visible from within the workbench. (DSE-44837)

Cloudera AI Inference service

- Resolved an issue that prevented the configuration map of AI Inference from updating correctly during the upgrade process. (DSE-45417
- Resolved an issue encountered when importing the A10G profile of the Llama 3.2 rerank 1B model. (DSE-45375)
- Previously, copying the base URL of model endpoints from the UI did not work properly, as the wrong link was getting copied. This issue is now resolved.(DSE-45107 and DSE-45534)
- Resolved an issue that prevented rendering of Test Model and code samples for external Hugging Face models, such as Gemma 3. (DSE-45419)
- Addressed inconsistencies observed in the Summary and Details metrics displayed for Model Endpoints. (DSE-45185)
- Fixed an issue where the Cloudera AI Registry upgrade pop-up was not correctly being dismissed. (DSE-46224).
- Previously, when the Test Model under Model Endpoint was executed, the UI blocked navigation to other tabs, such as Metrics, Logs, and so on. This issue is now resolved. (DSE-46247)
- Previously, due to a known KServe issue (kserve/kserve#4471), all newly created model endpoints would initially deploy with a single replica, regardless of the specified configuration. This issue is now resolved. (DSE-45876)

ML Runtimes

- Previously, PBJ Workbench-based workloads kept running when the underlying kernel had stopped or restarted. This is now fixed, and PBJ Workbench-based workloads will terminate if the underlying kernel terminates. (DSE-42964)
- Previously, messages printed from PBJ Workbench-based models did not appear in the model logs. This issue is now resolved. (DSE-42960)
- Previously, in the PBJ Workbench editor where text with special styling (for example, colored fonts) lost its formatting in the console. This issue is now resolved. (DSE-42958)
- Previously, PBJ Workbench Runtimes did not stop when "exit" was executed in a Python kernel or when "quit()" was executed in an R kernel. This issue is now resolved, and now these commands terminate the workload as expected. (DSE-36835)
- Fixed the known issue related to Spark executors in R. Now the environment variable R_LIBS_USER has the same configuration in Spark executors as in other workloads. It is no longer needed to manually configure R_LIBS_USER for Spark executors in R. (DSE-32839)
- Previously, some types of output were not fully shown in the Workbench UI when the workload was running a PBJ Workbench Runtime. This issue is now resolved. (DSE-43865)

June 06, 2025

Release notes and fixed issues for version 2.0.50-b68 and MLX-CRUD-APP version 1.50.0-b139.

New Features / Improvements

Cloudera AI Platform

- Added support for Azure AMD Easv5 instances. (DSE-38566)

May 20, 2025

Release notes and fixed issues for version 2.0.50-b68.

New Features / Improvements

Cloudera AI Workbench

- AI Studios (Technical Preview): Cloudera AI Studios is a comprehensive suite of low-code tools designed to simplify the development, customization, and deployment of generative AI solutions within enterprises. This suite empowers organizations to operationalize AI workflows quickly and efficiently by leveraging real-time enterprise data. For more information, see Managing AI Studios.

- Added APIs within the workbench to list Cloudera AI Inference service applications and their associated model endpoints.

Cloudera AI Platform

- Added Azure UDR support for Cloudera AI Inference service.
- Added Azure NTP support.
- Added API support to retry the creation of Cloudera AI Inference service application upon failure.

Cloudera AI Registry

- Added a new set of models in the Model Hub, including Llama3.3, DeepSeek-R1-Distill-Llama, Starcorder2, Llama-Nemotron-Nano, NeMo-Retriever-Parse, Llama 3.2 Embedding, and Llama 3.2 Encoder models. To access these models, you must upgrade your Cloudera AI Registries.
- Added support for nim-cli in the AI Registry to import the latest offerings from NVIDIA.
- Enhanced troubleshooting by surfacing underlying issues encountered during AI Registry installation in the Event logs.
- Provided the ability to upgrade the AI Registry directly through the UI, eliminating the reliance on the CLI.
- Implemented automatic redirection to the model import status page whenever a new model import is triggered.

Cloudera AI Inference service

- Users must upgrade their Cloudera AI Inference service applications to serve the latest optimized models from NVIDIA, including Llama3.3, DeepSeek-R1-Distill-Llama, Starcorder2, Llama-Nemotron-Nano, NeMo-Retriever-Parse, Llama 3.2 Embedding, and Llama 3.2 Encoder models.
- Optimization profile details for deployed model endpoints are now surfaced in the UI for improved visibility.
- A user-friendly warning message will now be displayed when replicas of a deployed model scale down.
- Added an option in the UI to retry the creation of Cloudera AI Inference service applications.
- Users will be automatically redirected to the model endpoint page upon triggering the deployment of a new endpoint.
- Enhanced the UI with a variety of user-friendly tooltips for better usability.
- The metrics page for model endpoints will now refresh automatically every 15 seconds for real-time updates.
- GPU count is now auto-selected for NIM profiles when deploying a model endpoint.
- Ensured that dangling pods of deleted endpoints are immediately terminated, preventing them from being left for garbage collection cleanup.

Fixed Issues

Cloudera AI Workbench

- Resolved an issue where duplicate machine user CRNs were preventing the catalog page backup from loading. (DSE-43729)
- Fixed an invalid error issue in the Cloudera AI Registry search filter within the workbench. (DSE-44401)

Cloudera AI Platform

- Resolved the issues causing failures during the retry of upgrade operations. (DSE-44761)
- Resolved an issue where team synchronization was removing all collaborators for a project created with LDAP team when a member was removed from the UMS group mapped to that LDAP team. (DSE-43524)

Cloudera AI Inference service

- Resolved an issue causing the deletion of the incorrect node group from the Cloudera AI Inference service UI. (DSE-44981)
- Resolved an issue preventing the import of older models like Mixtral due to compatibility constraints. (DSE-44972)

Apri 28, 2025

Release notes and fixed issues for version 2.0.50-b52.

New Features / Improvements

Cloudera AI Workbench

- Previously, when a job was already running and another job run was triggered by a cron job or an API call, the new run would be skipped and displayed as Failed in the UI. This update introduces a Skipped status, and any skipped job runs will now appear with the Skipped status in the UI.
- The Shared Memory Limit set under Project Settings now applies to both applications and sessions. Previously, it was applied only for sessions.
- Custom Spark settings can now be configured at Cloudera AI Workbench level. When set, the custom Spark configuration provided by the administrator will be merged with the default Spark configuration used in Cloudera AI sessions. These settings will automatically apply to all newly launched Spark sessions within the workbench. The configuration option is available under Site Administration > Runtimes.

Cloudera AI Platform

- Cloudera AI Workbench now utilizes EBS SSD gp3 volumes for newly created or restored Cloudera AI Workbench instances, replacing the previously used EBS SSD gp2 volumes.
- Added support for Sweden Central on Azure.
- Support for manually modified PVC sizes.
- Added support for EKS 1.31.
- Added support for AKS 1.31.

Cloudera AI Registry

- The Cloudera AI UI now displays clear error messages for failed Model Imports, enabling quicker troubleshooting.
- Users without the appropriate roles now see actionable error messages in the Model Hub popup.
- Load Balancer Subnet option is added during the AI Registry creation.
- The Cloudera AI UI now supports the force deletion of AI Registry.

> ⚠ **Important:** You must upgrade AI Registry after consuming the latest release to ensure compatibility with the most recent models available in our Model Hub. For information about AI Registry upgrade, see, Upgrading Cloudera AI Registry.

Cloudera AI Inference service

- Informative tooltips have been added to the Create Model Endpoints page to improve the user experience.
- Cloudera AI Inference service can now be created without the need for a node group.

ML Runtimes

- Project template files updated to fully support ML Runtimes. Project template files no longer work with Legacy Engines.

Fixed Issues

Cloudera AI Workbench

- Previously, in the Workbench editor, the search input retained the previous search string when reopened using Ctrl+F or Command+F. This issue has now been resolved. Now, clicking anywhere in the editor text area after a search clears the input, allowing a new search string to be entered. (DSE-40874)
- Increased GRPC operator timeout to ensure reliable handling of more than 150 concurrent session creation attempts. (DSE-36922)
- The Custom Data Connections feature is now properly enabled when the necessary entitlement is present. (DSE-42821)

- Previously, invalid entries in the runtime_addons table prevented the registration of models from Cloudera AI Workbenches. This issue has now been resolved. (DSE-44211)

Cloudera AI Platform

- CVE Fixes – This release includes numerous security fixes for critical and high Common Vulnerability and Exposures (CVE) and includes a CVE on a security vulnerability in the third-party SAML authentication service.
- TLS certificates are now properly deleted during the removal of Cloudera AI Inference service workflow. (DSE-41405)
- Grafana charts have been updated to eliminate references to deprecated metrics. (DSE-42389)
- Cloudera AI Workbench restoration now supports EFS backups larger than 10 TB. (DSE-42986)

Cloudera AI Registry

- Users can now download the kubeconfig file for the AI Registry as soon as the underlying cluster is provisioned. (DSE-42566)
- Hugging Face token is accurately passed by the UI during Hugging Face model imports. (DSE-42578)

Cloudera AI Inference service

- Instance groups for Cloudera AI Inference service can now be successfully deleted from the UI. (DSE-43182)
- Cloudera AI Inference service no longer overwrites underlying error messages and displays the actual error. (DSE-43315)
- Cloudera AI Inference service UI now accurately reflects the correct state when a user cancels the action of adding a new instance group. (DSE-43183)
- Metrics are now properly displayed in the UI for the model endpoint names that include mixed-case characters. (DSE-43339)
- The root volume size of node groups is now displayed in the Cloudera AI Inference service UI, enabling users to make informed decisions. (DSE-40603)

ML Runtimes

- Previously, the HTML code generated to embed an image in PBJ Workbench did not work. Now, you can embed images generated in PBJ Workbench-based sessions similarly to how you can embed images from Workbench-based sessions using the share icon next to the generated images. (DSE-42595)
- Previously, the Interrupt button did not work in Cloudera AI Sessions that ran a PBJ Workbench Runtime. This issue has been fixed. (DSE-42344)
- PBJ Workbench Runtime images now comply with the value set in the MAX_TEXT_LENGTH environmental variable. This limits the maximum number of characters that can be displayed by each command executed. (DSE-42962)
- Sessions using PBJ-based custom Runtimes with a custom editor could not start previously. This issue has been fixed. (DSE-43297)
- When PBJ Workbench R Runtime was used, the tables and help text were not properly displayed. This issue has been fixed. (DSE-42077)
- Previously, when PBJ Workbench R Runtime was used, the editor was not able to show code completion. This issue has been fixed. (DSE-42345)
- Previously, when PBJ Workbench R Runtime was used, it logged the start of each command execution with a DEBUG severity instead of the INFO severity. This issue has been fixed. (DSE-35299)
- Previously, when PBJ Workbench R Runtime was used, the content of the tables was not properly displayed. This issue has been fixed. (DSE-42967)
- Previously, when PBJ Workbench R Runtime was used, the output of the session commands was not properly displayed. This issue has been fixed. (DSE-42966)

Behavioral Changes

Cloudera AI Workbench

- The new Cloudera AI Workbenches will no longer have the Legacy Engine image configured. If you want to configure and use Legacy Engines in your projects, configure them by going to the Site Administration > Runtimes page. (DSE-42593 and DSE-39531)

March 28, 2025

Release notes and fixed issues for version 2.0.47-b365.

Fixed Issues

Cloudera AI Workbench

- The issue of sessions and pods getting stuck in the `Stopping` state has been resolved. (DSE-42144)
- Pods in an `Error` or `Stuck` state within Cloudera AI Workbenches are now being properly garbage-collected. (DSE-43549)
- Reduced the frequency of initialization failures for user workloads that launch immediately after node autoscaling. (DSE-43311)

Cloudera AI Platform

- Previously, users with MLAdmin roles were initially assigned the MLUser role during the first sync, but their permissions are updated correctly in subsequent syncs or when they log in. This issue is now resolved. (DSE-42775)

March 02, 2025

Release notes and fixed issues for version 2.0.47-b360.

Fixed Issues

Cloudera AI Workbench

- Previously, when users try to create a session, the ssh: This private key is passphrase protected error was displayed. This issue is now resolved. (DSE-426980)

February 26, 2025

Release notes and fixed issues for version 2.0.47-b359.

New Features / Improvements

Cloudera AI Platform

- We have improved the synchronization efficiency and ease of use of the user management and team management auto synchronization features. The major updates include:

  - **Auto-synchronization is enabled by default:** Auto synchronization for users and teams is now enabled by default, with a synchronization interval set to 12 hours.
  - **User management service:** User management is now handled by a new service, reducing overhead on the web pod. It now prevents multiple synchronization operations from running in parallel.
  - **Logging:** Detailed logging has been added for the failure cases.
  - **Synchronization trigger sequence:** The team synchronization now internally triggers user synchronization to pull the most recent user details from the Cloudera control plane.

  These improvements are aimed at optimizing performance and streamlining the synchronization process for users and teams. (DSE-37941)

- We have added support to set maximum input/output operations per second (IOPS) and throughput for root volumes attached to worker nodes, using the UI while provisioning a workbench. Note, that this is supported only for AWS. For more details on how to *Maximize IOPS and throughput of the root volumes*, see Provisioning Cloudera AI Workbenches. (DSE-42075)

Cloudera AI Registry

- You can now specify subnets for load balancers when creating the AI Registry. (DSE-42156)
- We have enhanced the security of the AI Registry's search capability. (DSE-41740)

Cloudera AI Inference service

- We have improved the UI usability of the Hugging Face import feature by adding a tooltip example. (DSE-41926)

Fixed Issues

Cloudera AI Workbench

- We have increased Grafana pod's default memory and CPU to prevent from out of memory (OOM) errors. (DSE-39525)
- We have increased the Remote Procedure Call (GRPC) Operator timeout to two minutes to prevent from errors encountered with 150 concurrent sessions. (DSE-36922)
- We have removed unessential calls to the usage API to resolve slowness during new workload creation under heavy load in a workbench. (DSE-42231)

Cloudera AI Platform

- We have optimized the Suspend timeout during periods of high network latency. (DSE-42055)
- Previously, when restoring a workbench with a very large Elastic File System (EFS) drive was failing due to session time out. This issue is now resolved. (DSE-42171)

Cloudera AI Registry

- We have fixed an issue that prevented from model registration to the AI Registry within a workbench. (DSE-42360)
- We have fixed a page token issue that prevented users from viewing AI Registry models on subsequent pages within the workbench. (DSE-42379)
- We have fixed an incorrect error message displayed in the UI when deleting AI Registry models from within a workbench. (DSE-42379)
- Error visibility has been improved during AI Registry backup. (DSE-42163)

Cloudera AI Inference service

- We have fixed an issue that prevented from rendering TPOT (Time per Output Token) and TTFT (Time to First Token) charts for Hugging Face models. (DSE-42192)

ML Runtimes

- Previously, non-administrator users were unable to add new Runtimes to the Runtime Catalog. This issue is now resolved. (DSE-42298)

February 07, 2025

Release notes and fixed issues for version 2.0.47-b345.

New Features / Improvements

Cloudera AI Workbench

- Support is now provided for API keys to invoke applications deployed using Cloudera AI Workbenches. This not only eases the invocation of those applications programmatically but also allows one application to easily invoke another application that they have access to.

- MLFLOW upgrade for Cloudera AI Workbenches now enables making use of the latest offerings and APIs from the MLFLOW community like *evaluateLLM*.

Cloudera AI Platform

- The autoscaling range of Suspend Workflow is now set to the value 0 to ensure that other Kubernetes deployments outside the scope of MLX can deploy their pods on worker nodes.

Cloudera AI Registry

- An enhanced error message is now displayed during model upload failure.
- UI for Registered Models displays the environment name of the registry along with an error message when any user is unable to access any Cloudera AI Registry.
- A checkbox is now added to enable Public Load Balancer for new Cloudera AI Registries on Azure.

Cloudera AI Inference service

- The Hugging Face model server backend has been upgraded, which expands the compatibility with a larger number of model families, such as Llama 3.3 and models derived from it.
- Llama 3.2 Vision Language Model NIM version has been updated to address compatibility with A10G (g5.*) and L40S (g6e.*) GPU instances on AWS.
- You can now upgrade Cloudera AI Inference service using the UI. Previously, the upgrade was supported only using CDPCLI.
- You can now upgrade from Cloudera AI Inference service version 1.2.0-b80 to version 1.3.0-b113 or higher. Note that you cannot upgrade from 1.3.0-b111 to 1.3.0-b113 or higher. For more information on the 1.3.0-b111 upgrade issue and workaround, see the Known Issues section.

Fixed Issues

Cloudera AI Workbench

- Previously, due to an issue, users could stop sessions under projects that they were not authorized to access using the session’s UUID. This issue is now resolved. (DSE-39798)
- Previously, when a Kubernetes object was deleted, and the reconciler was overwhelmed by a large number of events, the Deleted status failed to propagate properly. This issue is now resolved. (DSE-41431)
- Previously, the stopped_at column was not correctly populated when applications were stopped. This issue is now resolved. (DSE-41636)
- Previously, engine pods were stuck in the Init:StartError state and you had to manually delete it. With this fix, pods stuck in Init:StartError in the Garbage Collection will be deleted after a certain grace period. (DSE-41430)
- Previously, Spark environment configurations were not inherited by models running Spark. With this fix, models use the appropriate Spark configurations to run Spark. (DSE-36343)

Cloudera AI Registry

- An issue around how Hugging Face token was being consumed during the import of a model was addressed. (DSE-41714)
- The Cloudera AI Registry deletion flow is improved to take care of race conditions when both creation and deletion are triggered in a short frame of time. (DSE-41634)

Cloudera AI Inference service

- Previously, the GetEndpointLogs failed with an error. With this fix, endpoint logs for the model container do not exceed the gRPC messaging size. (DSE-41765)
- A new field called loadBalancerIPWhitelists is added to display a list of IPs whitelisted for the load balancer and deprecated isPublic and ipAllowlist. (DSE-39397)
- Infrastructure nodes are no longer shown as instances that can be used for deploying a new endpoint. (DSE-41726)

ML Runtimes

- Previously, due to an issue, to ensure the compatibility of AMPs with ML Runtimes 2025.01.1, users had to switch to JupyterLab PBJ Workbench in the AMPs' .project-metadata.yaml file or use jobs instead of sessions for automated tasks. This issue is now resolved. (DSE-41263)
- Resolved issues related to using R interactively in PBJ Runtimes. (DSE-41771)

January 29, 2025

Release notes and fixed issues for version 2.0.47-b302.

New Features / Improvements

- Migrated Cloudera AI Workbench, Cloudera AI Registry, and Cloudera AI Inference service images to chainguard to address CVEs.
- Added APIv2 support for Enhanced Group Sync.
- Added support to create AMPs (Cloudera Accelerators for Machine Learning Projects) using APIv2. Previously, this option was available only using UI.
- Added support for H100 GPU instances for Cloudera AI Inference service on Azure.
- Added support for AKS workload identity.
- Added support for AWS M7a, M7i, C7a, C7i, R7a, R7i instance families.
- Added support for Cloudera AI Inference service on EU Control Plane.
- Added support for EKS 1.30.
- Added support for AKS 1.30.
- Hugging Face support (Technical Preview): You can now import text-generating language models from Hugging Face and deploy them on Cloudera AI Inference service.
- Added profiles for HuggingFace Models and multi-modal models in the Model Hub catalog.
- Updated existing model manifests in the catalog after upgrading the NIM version in Cloudera AI Inference service.
- Enhanced error messages related to model import failure in the Model Hub UI.
- Carried enhancements in AI Registry to ensure that multi-modals can be supported.
- Added runtime support for Llama 3.2 11B and 90B Vision Language Model NIMs to ensure that they can be deployed using AI Inference. Only model profiles optimized for the H100 GPU are supported for these two models in this release.
- Llama 3 NIM is no longer supported since we now have both Llama 3.1 and Llama 3.2.
- Added support for Diagnostic Bundles in Cloudera AI Inference service.
- Upgraded text-generating and embedding NIMs.
- Added Code Sample functionality for endpoints deployed using Cloudera AI Inference service.
- Model endpoint replica events can now be viewed on the Model Endpoint details UI.You can now add numerous docker credentials using UI or API which can be used to enable Cloudera AI to fetch custom ML Runtimes from a secure repository. For more information, see Add Docker registry credentials and certificates.

Fixed Issues

- Previously, some Cloudera AI Inference service clusters did not have the 'creationDate' field. This field is now added.(DSE-38817)
- Previously, the deletion of backup for older workspaces was failing. This issue is now resolved. (DSE-41031)
- Previously, deleting a workbench backup created by a deleted user displayed an error. This issue is now resolved. (DSE-41052)
- Multiple UI improvements are made both in the Create, Read, Update, and Delete operations of Cloudera AI Inference service and while deploying or editing a model endpoint.
- The model_name field is now displayed instead of model_id in the Endpoint Details UI. (DSE-38937)
- Previously, the NIM model profile environment variable was only assigned for LLMs. Now support for Model Profile override is added for Embedding and Reranker NIMs. (DSE-40508)

- Previously, there was an issue with rendering of existing instance type in the "Edit Endpoint" UI. This issue is now resolved. (DSE-40636)
- Validated all node group (instance type) selection from UI. (DSE-40754)
- Previously, NGC manifest components were missing from the download. This issue is now resolved. (DSE-41055)
- The Create ML Serving application now enables the public load balancer. (DSE-41305)
- The Instance Type field in the Edit Model Endpoint UI is no longer mandatory. (DSE-41278)
- Added force delete option to delete the Cloudera AI Inference service using UI. (DSE-41035)
- The Cloudera AI Inference service UI now displays optimization profile details. (DSE-40927)
- You can now create, download, and delete log archives for Cloudera AI Inference service. (DSE-40921)
- The Test Model UI now fails gracefully when the replica is scaled down to zero for a model deployed using Cloudera AI Inference service. (DSE-40957)
- Previously, the Storage initializer had the wrong task values. This issue is now resolved. (DSE-41058)
- Enabled storage initializer to now handle more than two directories for NIM artifacts. (DSE-40986)
- Removed Llama 3 runtimes. (DSE-40956)
- Addressed SQL injection issue in AI Registry that allowed non-authorized but authenticated users to perform Create, Read, Update, and Delete operations on AI Registry’s metadata tables. (DSE-41542)

November 21, 2024

Release notes and fixed issues for version 2.0.46-b238.

New Features / Improvements

- Model Hub Enhancement: The model size is now shown in the user-friendly format both in the Model Hub UI and Cloudera AI Registry UI.
- Cloudera AI Inference service Enhancement: New AI Inference Services menu item is added to the left-navigation pane of the Cloudera AI UI to manage the lifecycle of Cloudera AI Inference service using UI. For more information, see Using Cloudera AI Inference service.
- Added Spark 3.5 ML Runtime Addon
- Product and features named:
  - *Clouder Machine Learning (CML)* is renamed to *Cloudera AI*.
  - *Cloudera Machine Learning Model Registry* is renamed to *Cloudera AI Registry*.
  - *Cloudera Machine Learning Workspace* is renamed to `Cloudera AI Workbench`.
  - *Cloudera Applied Machine Learning Prototypes* and `Accelerators for ML Projects` is renamed to *Cloudera Accelerators for Machine Learning Projects*.

Fixed Issues

- CVE fixes - This release includes numerous security fixes for critical and high Common Vulnerability and Exposures (CVE).
- Previously, the public and private settings did not carry forward after the AI Registry upgrade. This issue is now resolved. (DSE-36799)
- Enhanced the error message that was displayed when importing a model from Model Hub to Registered Models. (DSE-39897)
- Generic (vLLM) NIM profile deployment was returning an empty GPU list in the UI. This issue is now resolved. (DSE-39913)
- Previously, public cloud CDP CLI was not showing the instance type's GPU count. This issue is now resolved. (DSE-39539)
- Cloudera AI v2 API deployed application did not inherit user-level environment variables and site-level environment variables. This issue has been solved, and now an application created using APIv2 does not only

inherit project-level environment variables but also user-level environment variables and site-level environment variables. (DSE-37611)
- Previously, scheduled jobs skipped job runs and did not specify the error. Now, the skipped jobs runs have improved exit code to distinguish them from failed jobs. (DSE-39976)
- Previously, the Next buttons on the Site Administration page did not work. This issue is now resolved. (DSE-34133).
- Previously, restarting the application using the Cloudera AI v2 API did not inherit account application-level environment variables. This issue is now resolved. (DSE-39894)
- Users can now view the existing applications in the Cloudera AI UI even if the creation of a new application is disabled. (DSE-39980)
- Previously, Python logging did not work with PBJ Runtimes. This issue is now resolved. (DSE-39929)
- Previously, reloading the session page would result in an incorrect state where the PBJ session's editor cell could appear green even if it is in a processing state (executing some commands). With this fix, an accurate representation of the processing state is displayed even after a refresh. (DSE-40049)

October 10, 2024

Release notes and fixed issues for version 2.0.46-b210.

New Features / Improvements

- Model Hub: Model Hub is now a fully supported feature. Model Hub is a catalog of top-performing models LLM and generative AI models. You can now easily import the models listed in the Model Hub into the Cloudera AI Registry and then deploy it using the Cloudera AI Inference service.

  For more information, see Using Model Hub.
- Cloudera AI Inference service Enhancements:
  - Added support for NVIDIA's NIM profiles requiring for the L40S GPU models.
  - Made auto-scale configuration which is rendered in UI during the creation of model endpoint user-friendly. (DSE-38845)
  - Optimized Cloudera AI UI service to become more responsive.
  - User actionable error messages are now rendered in Cloudera AI service UI.

    For more information, see Using Cloudera AI Inference service.

Fixed Issues

- Addressed scaling issues with web services to support high active user concurrency (DSE-39597).
- CVE fixes - This release includes numerous security fixes for critical and high Common Vulnerability and Exposures (CVE).
- Fixed CORS issue to ensure that DELETE/PATCH V1 API can be used from within a workbench. (DSE-39357)
- Made the NGC service key used to download Nvidia’s optimized models more restrictive. (DSE-39475)
- Previously, users were unable to copy the model-id from Cloudera AI UI. This issue is now resolved. (DSE-38889)
- Authorization issues related to the listing of Cloudera AI applications have been addressed. (DSE-39386)
- Fixed an issue to ensure that instance type validation is correctly carried out during the creation of a new model endpoint. (DSE-39634)
- Added required validation rules for the creation of a new model endpoint. (DSE-38412)
- Addressed an issue around empty model list during navigation from registry models to deployment of models. (DSE-39634)

October 8, 2024

Release notes and fixed issues for Cloudera AI Inference service version 1.2.0-b73.

New Features / Improvements

- Cloudera AI Inference service: Cloudera AI Inference service is now a fully supported data service. Cloudera AI Inference service is a production-grade serving environment for traditional, generative AI, and Large Language Models. It is designed to handle the challenges of production deployments, such as high availability, fault tolerance, and scalability. The service is now available to carry out inference on the following categories of models:

  - Optimized open-source Large Language Models.
  - Traditional machine learning models like classification, regression, and so on. Models need to be imported to the Cloudera AI Registry to be served using the Cloudera AI Inference service.

  For more information, see Using Cloudera AI Inference service.

September 26, 2024

Release notes and fixed issues for version 2.0.46-b200.

New Features / Improvements

- Model Hub (Technical Preview): Model Hub is a catalog of top-performing LLM and generative AI models. You can now easily import the models listed in the Model Hub into the Cloudera AI Registry and then deploy it using the Cloudera AI Inference service service. This streamlines the workflow of developers working on AI use cases by simplifying the process of discovering, deploying, and testing models.

  For more information, see Using Model Hub.
- Registered Models: Registered Models offers a single view for models stored in Cloudera AI Registry instances across Cloudera Environments and facilitate easy deployment to the Cloudera AI Inference service service. When you import models from Model Hub, the models are listed under Registered Models. This page lists all imported models and associated metadata, such as the model's associated environment, visibility, owner name, and created date. You can click on any model to view details about that model, and its versions, and deploy any specific version of the model to the Cloudera AI Inference service service.

  For more information, see Using Registered Models.
- Cloudera AI Inference service (Technical Preview): Cloudera AI Inference service service is a production-grade serving environment for traditional, generative AI, and LLM models. It is designed to handle the challenges of production deployments, such as high availability, fault tolerance, and scalability. The service is now available for users to carry out inference on the following three categories of models:

  - TRT-LLMs: LLMs that are optimized to TRT engine and available in NVIDIA GPU Cloud catalog, also known as NGC catalog.
  - LLMs available through Hugging Face Hub.
  - Traditional machine learning models like classification, regression, and so on. Models need to be imported to the Cloudera AI Registry to be served using the Cloudera AI Inference service Service.
- Cloudera AI Registry Standalone API: Cloudera AI Registry Standalone API is now fully supported. This new API is available from the Cloudera AI Registry service to import, get, update and delete models without relying on the Cloudera AI Workbenchservice.

  For more information, see Cloudera AI Registry Standalone API.
- New Amazon S3 Data Connection: A new Amazon S3 object store connection is automatically created for Cloudera AI Workbenches to make it easier to connect to the data stored within the same environment. Other Data Connections can be configured to other S3 locations manually.

  For more information, see Setting up Amazon S3 data connection.

- Enhancements to Synced Team: Team administrators and Site administrators can now add multiple groups to a synced team, view members of a group, delete a group within a team, update roles for a group within a team, and update a custom role for a member within a group.

  For more information, see Managing a Synced Team.
- Auto synchronization of Cloudera AI Registry with a Cloudera AI Workbench: If you deploy a Cloudera AI Registry in an environment that contains one or more Cloudera AI Workbenches, the Model Registry is now auto-discovered and periodically synchronized by Cloudera AI Inference service service and Cloudera AI Workbenches and no manual synchronization is required. Cloudera AI Workbench is auto-synchronized every five minutes and Cloudera Cloudera AI Inference service service is auto-synchronized every 30 seconds.

  For more information, see Synchronizing the Cloudera AI Registry with a Cloudera AI Workbench.
- Environment: Support for Environment V2 is added for Cloudera AI Workbenches.
- Kubernetes: Support for AKS 1.29 and EKS 1.29 was added.
- Metering: Support for Metering V2 is added for new Cloudera AI Workbenches.

Fixed Issues

- DSE-35779: Fixed the issue related to a race condition between writing the JWT file by kinit container and reading by the engine container in the workload pod.
- DSE-37065: Previously, API V2 did not allow collaborators to be added as admin. This issue is now resolved.
- DSE-33647: Previously, workbench instances reset to default when upgraded. This issue is now resolved.

July 17, 2024

Release notes and fixed issues for version 2.0.45-b86.

Fixed Issues

- Previously, Cloudera teams with the MLBusinessUser role were not available for Synced Teams in Cloudera AI workbenches. This issue is now resolved.

June 20, 2024

Release notes and fixed issues for version 2.0.45-b82.

New Features / Improvements

- You can now delete the Model Endpoint under AI Inference from the Cloudera Machine Learning UI.

Fixed Issues

- Previously, the Model Registry service was broken due to a change in the certificates. This issue is now resolved.

June 11, 2024

Release notes and fixed issues for CDSW version 2.0.45-b81 and MLX-CRUD-APP version 1.45.0-b71.

Fixed Issues

- Previously, users were unable to upgrade their model registry or create new model registries if their cluster was resized. This issue is now resolved.

May 29, 2024

Release notes and fixed issues for CDSW version 2.0.45-b81 and MLX-CRUD-APP version 1.45.0-b68.

New Features / Improvements

- The NodeSelector label can now be added for inference services. The label can be specified in the `instance_type` field of the deploy or update requests. This would enable you to direct inference service pods to specific nodes.
- Enhancements to the Export API to support the Observability APIs.

Fixed Issues

- Added support to create model registries in AWS with a public load balancer from the UI. Previously, this option was available only from the backend.

May 16, 2024

Release notes and fixed issues for version 2.0.45-b80.

Fixed Issues

Cloudera AI Workbench

- Previously, Cron jobs failed to run after upgrading from CML 2.0.43 to CML 2.0.45. This issue is now resolved.

May 15, 2024

Release notes and fixed issues for version 2.0.45-b76.

New Features / Improvements

- Cloudera AI Inference Service (Technical Preview): AI Inference service is a production-grade serving environment for traditional, generative AI, and LLM models. It is designed to handle the challenges of production deployments, such as high availability, fault tolerance, and scalability. The service is now available for users to carry out inference on the following three categories of models:

  - TRT-LLMs: LLMs that are optimized to TRT engine and available in NVIDIA GPU Cloud catalog, also known as NGC catalog.
  - LLMs available through Hugging Face Hub.
  - Traditional machine learning models like classification, regression, and so on.

  Models need to be imported to the model registry to be served using the Cloudera AI Inference Service.

  For more information, see Using Cloudera AI Inference service.
- Cloudera Copilot (Technical Preview): Cloudera Copilot is a highly configurable AI coding assistant integrated with the JupyterLab editor. The Copilot improves developer productivity by debugging code, answering questions and generate notebooks.

  For more information, see Cloudera Copilot.
- Model Registry API (Technical Preview): New API is available from the Model Registry service to import, get, update and delete models without relying on the CML Workspace service.

  For more information, see Model Registry API.
- Ephemeral storage limit: The default ephemeral storage limit for CML Projects has been increased from 10 GB to 30 GB.

Fixed Issues

- Fixed an error that occurs while sorting public projects on the Jobs column.
- Fixed a bug that was uploading files to the root directory of a project instead to the specified subfolder.

April 25, 2024

Release notes and fixed issues for version 2.0.45-b54.

New Features / Improvements

- Model Registry - A model registry can now be upgraded with the Upgrade feature. Learn more: Upgrade model registry
- AMPs - various infrastructure improvements.
- API v2 - Modified model creation, build, and deployment endpoints more consistent with API v1, and support for CDV application creation was added.
- API v2 - Improvements made to List Projects endpoint, and other changes to support file uploads.
- Models - Model metrics are now supported with models that are deployed from a model registry.
- Kubernetes - Support for EKS 1.28 and AKS 1.28 was added.
- Workspace - The Create Workspace flow was improved to add validation of endpoint access and provide diagnostic responses.
- Azure - New Azure GPU instance types are supported: NC*A100, D16x v5, and D8s v5.
- AWS - New AWS GPU SKU is supported: p5
- Azure - The Middle East (Qatar Central) region is now supported.
- Runtimes - New Runtime Addons are released:
  - HadoopCLI 7.2.16.600
  - HadoopCLI 7.2.17.300
  - Spark 2.4.8
  - Spark 3.2.3
  - Spark 3.3.0

Fixed Issues

- DSE-28892 - Improved readability of Spark3 session startup WARN logs.
- DSE-35080 - Fixed an issue where Spark 3.2.3 configuration files were empty.
- DSE-35633 - Fixed an error that occurred when performing Create Project to create an AMP from a zip file.

February 20, 2024

Release notes and fixed issues for version 2.0.43-b229.

Fixed Issues

- DSE-33078 - Upgraded Go version to address CVE-2023-45285 and CVE-2023-39325.
- DSE-34663, DSE-34635 - Upgraded Grafana version and `kube-state-metrics:v1.9.3` image to address critical CVEs.
- DSE-32001 - Upgraded thunderhead components (`tgtgenerator`, `tgtloader` and `configtemplate`) to include CVE fixes that were carried out in `thunderhead`.
- DSE-34595 - Upgraded zookeeper version to address CVE-2023-44981.

March 6, 2024

Release notes and fixed issues for version 2.0.43-b233.

Fixed Issues

- DSE-34127, DSE-34692 - Allow MLUsers to create team projects. Project owner is added as a collaborator.
- DSE-35258 - Ensured that manual entries that were added to AMPs catalog sources in the past do not prevent upgrade of workbenches.
- DSE-35211 - Upgraded Traefik version to address CVEs: CVE-2023-44487.
- DSE-35270, DSE-34148 - Fixed the mlinfra scale-up issue during Resume and Upgrade Workflow.

February 8, 2024

Release notes and fixed issues for version 2.0.43-b220.

New Features / Improvements

- AMPs - The AMPs page has been upgraded to render images, make the UI more reactive and improve the overall experience.
- Azure - Added Azure Qatar Central as a supported region.

Fixed Issues

- DSE-33041 - Fixed a bug in API v2 to ensure that cron jobs created via API v2 are now adding the schedule correctly and not resulting in multiple instances of jobs running at any given moment.
- DSE-33793 - Fixed an issue in the UI to ensure that users are able to select the ‘Scala’ kernel while creating a new runtime.
- DSE-34166 - Made a change to ensure that auto-generated CML model sample code shown in the UI is not missing a quote.
- DSE-33082 - Upgraded the Go version to address CVEs: CVE-2023-45285 and CVE-2023-39325.
- DSE-34080 - Upgraded the UBI base images in CDSW to address multiple CVEs.
- DSE-33598 - Ensure that the error message about a missing workload password when JWT authorization is disabled is shown in the UI.
- DSE-34088 - Made a change to the renderSpark UI Executor Logs.

January 23, 2024

Release notes and fixed issues for version 2.0.43-b208.

New Features / Improvements

- Cloudera Data Warehouse - Automatic JWT-based authentication enables passwordless connectivity to CDW. Users do not need to use their workload password to query data from CML. This feature depends on Data Lake 7.2.18, please upgrade your environment when the new version is available.
- Redesigned AMP Catalog - The AMPs pane is redesigned to improve navigation and search capabilities.
- HuggingFace Spaces - A curated list of HuggingFace Spaces is available in the AMPs Catalog.
- Community AMPs - A selected list of community-created AMPs is available to run in CML in the AMPs pane.
- Azure - Support for new GPU instances: NVadsA10 v5-series (non-fractional)
- Azure - Certificate based authentication using Managed Identity to provision in AKS.
- AWS - Support for new GPU instances: g5
- AWS - Deprecated support for P2 instance types.

- AWS - Added support for CML workspaces in af-south-1, Africa (Cape Town) region.
- Kubernetes - Kubernetes version 1.27 is supported on both Azure and AWS.
- Restore workflow - Improved reliability of the workspace restore workflow.
- Private DNS Zone - CML now certified to work with private DNS zones.
- Project Migration tool - A command line argument is added to check if source and destination files are the same, covering job, app, model, project data and metadata files.
- Runtimes - The R version of cmladdon is upgraded to version 4.3.1.
- Runtimes - The HadoopCLI 7.2.17.100 Runtime Addon is released for the Public Cloud.
- Runtimes - ML Users can now register custom ML Runtimes.

  A new site-level configuration option has been introduced on the Site Administration page to enhance runtime registration capabilities:

  - Admins can enable a configuration option, granting users the capability to register runtimes.
  - The new option is disabled by default, ensuring that existing permissions remain unchanged.
  - With this update, we've implemented a mechanism to record the names of the users who register ML Runtimes, displayed on the Runtime Catalog Page.
  - Users now can add runtimes however, it's important to note that they are not permitted to deprecate/disable the added runtimes.
- Runtimes - We have introduced a new button in the Site Administration Runtime tab, where users are now able to update the Runtimes catalog at any time based on runtime repos.
- Runtimes - On the Project creation page:

  - Renamed the Runtime setup section to Runtime.
  - Updated the UI for Basic/Advanced option selection.
  - Modified basic configuration settings.
  - Added GPU-enabled runtime variant by default.
- Security - When adding project collaborators or team members, non-admins can be prevented from seeing the entire user list. This functionality can be restricted to Site Admins in Site Administration Security by selecting Allow all authenticated users to access /api/v1/users endpoint.
- Teams - The ownership structure for team creation is changed. Previously, the creator of a team was assigned the role of Owner. Now, the creator is assigned the role of Admin by default. Admins have the authority to add team members, including other admins. Each team must have at least one admin to manage the team.

Fixed Issues

- DSE-33545 - Fixed an issue where workspace shows as Ready even if workspace URL is returning 404.
- DSE-24423 - In Network Settings of a Workspace, removing last remaining allowed source IP range for Load Balancer throws error
- DSE-27910 - Fixed an issue where evicted pods caused CML backup to fail.
- DSE-23954 - Fixed a problem where load balancer subnet preflight validation is skipped if no load balancer is specified during workspace creation.
- DSE-23953 - Fixed a problem where worker subnet validation is skipped if no load balancer is specified during workspace creation.

December 15, 2023

Release notes and fixed issues for version 2.0.41-b238.

Fixed Issues

- ENGSEC-22894 - Fixed an issue with non-admin users being unable to add or remove collaborators on team projects.

- DSE-32894 - Fixed a problem with Model Registry causing the PATH variable to become too long and fail. See *Cloudera Customer Advisory (2023-725)* for more information.
- ENGSEC-23097 - Fixed an issue with CML upgrades failing when the control plane scales down nodes.

November 15, 2023

Release notes and fixed issues for version 2.0.41-b236.

Fixed Issues

- DSE-32250 - Spark Executors now report the correct create time and resource usage in the Site Administration Usage Export Usage List . Note that previously incorrect entries for Spark executors are reset to 0.
- DSE-32222 - On AWS, upgrading an NTP-enabled workspace to Kubernetes 1.25 AWS now incorporates various internal improvments.

October 19, 2023

Release notes and fixed issues for version 2.0.41-b225.

New Features / Improvements

- Model Registry - Model Registry is now GA. Model Registry is the core enabler for MLOps, or DevOps for machine learning. For more information, see Using Model Registry.
- Experiments - Experiments v2 is now GA. The Experiments feature now integrates with MLflow for managing the model lifecycle. For more information, see Experiments.
- Service Accounts - Service Accounts, which allow automated processes to run with their own user account, is now GA. For more information, see Service Accounts.
- Usage log tracking - Usage log records all workloads: sessions, jobs, models, applications and distributed compute, enabling administrators to export and analyze workload statistics on-demand.
- Kubernetes - Kubernetes 1.25 is now supported for EKS.
- Azure - New Azure instance types are supported: D4asv4, D16asv5 and D8asv5.
- Azure - On new installations, the nfs-csi-driver is now enabled.
- Azure - Cross-environment backup and restore of workspaces is now supported.
- Applications - Users can now see pod logs for applications. In Application Details, go to the Container Logs tab, and the pod logs are shown. Application and pod logs can be downloaded from the respective pages.
- Runtime Addons - CML now includes HadoopCLI Runtime Addon 7.2.15, and HadoopCLI 7.2.14 Runtime Addon is removed.

Fixed Issues

- DSE-30784 - Added more logging and event logging in Resume Workflow
- DSE-20734 - Fixed an issue where a machine user was shown as Unknown in events and log pages.
- DSE-30229 - Fixed an issue that caused CML upgrades to fail on AWS.
- DSE-29365 - Fixed an issue where the YQ version checker script was not working for latest versions of `yq` in the CML application.
- DSE-28187 - Fixed an issue so that Retry Workspace Installation is disabled if Liftie Provisioning fails.
- DSE-27000 - Upgraded cdp-js to version 4.2.5 in the CML application.
- DSE-26069 - Fixed an issue so that deleted and failed Spark addons do not appear as selectable addons in the New or Modify Workload pages.
- DSE-26966 - Added Datalake preflight checks to the upgrade API.

August 31, 2023

Release notes and fixed issues for version 2.0.40-b157.

Fixed Issues

- AWS : Non-Transparent Proxy Support

  Fixed issues with supporting non-transparent proxy related to web pod connectivity to control plane, image-puller, pod to pod connectivity and api pods missing configuration.
- Upgrades
  - Fixed an issue where workspace upgrade fails due to not deleting orphan pods.
  - Fixed an issue where workspace upgrade fails due to incorrect error handling.
- Fixed an issue where the Team Sync tab was disabled.
- Azure
  - Added a preflight check when creating NTP clusters: Azure does not support NTP.
  - The NCv2 series is no longer supported due to its upcoming end of support on September 6, 2023. See TSB-692 for details. The default GPU for creating Azure workspaces will now be the `Standard_NC6s_v3`. This new GPU offers nearly identical configurations to the previous default GPU. If you are currently using NCv2 series GPUs, it is recommended to make the necessary adjustments before the end of support date.

July 25, 2023

Release notes and fixed issues for version 2.0.40.

New Features / Improvements

- CML Home - A new landing experience that helps you to jump to your most recent Projects, walks you through the key capabilities of the platform and keeps you up-to-date with the latest developments.
- 3rd-party editor support - The PBJ architecture adds support for 3rd party editors, enabling building custom ML Runtimes from scratch with JupyterLab, RStudio, and other editors of your choice.
- Models - Models can now be deployed from the model registry to CML workspaces using API v2.
- Kubernetes - Kubernetes 1.26 support is available for Azure.
- Model Registry (Preview) - Model Registries can now be deployed on an Azure private cluster.
- CML Scalability - CML Control Plane flows have been verified for 100-node clusters and high volume workloads are now enabled by default.
- Project Migration - Support for project export and import from CDSW to CML public or private cloud environments, or for migration between development and production environments.
- Retry Install Workspace - Workflow-based support for retry of CML workspace creation, in the event workspace installation fails.
- Preflight Checks for Instance Groups - Pre-flight checks are run when an instance group is being modified to ensure the requested configuration is valid."
- Preflight Checks for Update Workspace - Pre-flight checks are run when a workspace update is being requested to ensure the requested change is valid.
- SDX - SDX 7.2.17 has been integrated and verified with CML.
- Runtimes - On the New Project page, CML code now defaults to the Python 3.9 ML Runtime Edition.
- HadoopCLI - The DL versions of HadoopCLI 7.2.8, 7.2.10 and 7.2.11 Runtime-Addon versions on Public Cloud reached End of Support and have been removed.
- Job notifications - All email-related control will be hidden by job creation or job settings if the SMTP host is not configured. If email recipients have been added previously to the job but the SMTP host is not configured, the Job Notifications section will be displayed as a warning message, informing the user of the problem.
- Project - Site administrators can now restrict project creation for users and/or teams.

- Environment variables - Users can now hide the values of sensitive environment variables on the Account/Project/Workload/Workspace level.

Fixed Issues

- Upgrade workspace (DSE-28827) - Fixed an issue where after upgrading a workspace, the workspace (such as instance type and autoscale range) could not be modified.
- Pod (DSE-28771) - Fixed an issue where the health-poller pod was crashlooping due to a missing UMS_HOST environment variable.
- Workspaces (DSE-28595) - Fixed an issue where workspaces with static subdomains would not register freeIPA DNS entries were not added correctly, for both created and restored workspaces.
- Installation (DSE-28160) - Fixed an issue where an installation is not marked 'Fail' when it has timed out for 3 hours. This happened on Azure private cluster (UDR) installations.
- Installation (DSE-28047) - Fixed an issue where when a failed installation on Azure is retried, pre-install validations are not performed again, which can lead the installation to eventually time out.
- Environmental variables (DSE-28005) - Fixed an issue where environment variables were not visible or modifiable on an application's settings page.
- CML Service Link (DSE-28072) - This change enables users to use the CML Service Link on the project session pages started either with Workbench or with the remote editor.
- Workspace details (DSE-27891) - Fixed an issue where Workspace Details shows some incorrect information (Creator, User, Workspace tags) when the workspace is created via a machine user.
- Project (DSE- 27516) - Fixed an issue where project creation in a workspace on Azure was failing due to an NFS error.
- Retry Install Workspace (DSE-26944) - Improved the language in the Retry Install Workspace modal.
- Preflight (DSE-26752) - Improved the instance preflight failure message to mention that validation is not skippable.
- Spark addons (DSE-26269) - Fixed an issue where deleted and failed Spark Addons appeared on New/Modify Workload pages as selectable Addon.
- CDP CML (DSE-26012) - Prefix and update all workload usage events with CDPCML.
- CML consumption (DSE-25951) - Made improvements to issues that caused discrepancies in measurement of CML consumption.
- Status icon on Runtime Catalog Page (DSE-24589) - This change enables users to find out if there is at least one runtime among the runtime variants with the enabled status.
- PBJ Workbench (DSE-21770) - In PBJ Workbench Sessions, history navigation (with UP key) is working now, and users are able to navigate to previously executed commands with this key.
- GPU (DSE-29159) - Fixed an issue where `P4d.24xlarge` GPU is displayed as a CPU, not GPU.

July 12, 2023

Release notes and fixed issues for version 2.0.38-H3.

Fixed Issues

- Team Sync (DSE-29009) - Fixed an issue where Team Sync did not work in AWS regions other than us-west-1.

May 31, 2023

Release notes and fixed issues for version 2.0.38-H2.

New Features / Improvements

- Embedding applications - CML applications can now be embedded in i-frames in another domain. See Embed a CML application in an external website for more information.

- Team collaborator - It is now possible to add a team account as a collaborator on a project. This feature is supported via UI as well as API. For more information, see Adding project collaborators.
- Project owner - It is now possible for an administrator to change the owner of a project. This feature is supported via UI as well as API. For more information, see Modifying project settings
- Load Balancer - You can now select from the environment's endpoint access gateway subnets in the Subnets for Load Balancer field.

Fixed Issues

- Projects (DSE-27795) - Fixed an issue where certain public projects owned by a user or where the user is a collaborator were not displayed in the user's My Projects view.
- Jobs (DSE-27535) - Fixed an issue so that a job can be created in disabled or paused state.
- API v2 (DSE-27463) - Added missing fields (timezone, paused) to createJobRequest endpoint.
- API v2 (DSE-27442) - Fixed an issue so that files in subfolders are uploaded to their respective subfolders, and not to a root directory, when using the uploadFile endpoint.
- API v2 (DSE-27391) - When creating a project via API v2, the ephemeral storage value is now set to NULL (instead of zero). This fixed an issue that was blocking certain CDSW to CML migration workflows.

May 16, 2023

Release notes and fixed issues for version 2.0.38-H1.

New Features / Improvements

- Private Cluster - Private clusters can now be created on AWS. For more information, see Private cluster support.
- Spark - INSERT events are now fired via HMS after a DML action from Spark

Fixed Issues

- Workspace (DSE-27291) - Fixed an issue where Instance Group Modification failed with an older version workspace.
- Workspace (DSE-27406) - Fixed an issue where SkipValidation was not observed when restoring a workspace.
- UI (DSE-27259) - Fixed an issue where the incorrect resource details were displayed for an instance type.
- Workspace (DSE-9552) - Fixed an issue to allow provisioning of a workspace with more than 30 nodes.

April 26, 2023

Release notes and fixed issues for version 2.0.38.

New Features / Improvements

- Custom Data Connections - Site administrators can now configure access to external data sources with the new Custom Data Connections feature. Data Scientists can access the external data via the `cml.data` library and its 2-liner abstractions.
- Add Data - Data Scientists can now upload files to Hive and Impala Virtual Warehouse tables from the Data tab of any CML Project.
- Model Replicas - Site Administrators can now configure the maximum number of model replicas that users can select for their models via the Maximum Model Replicas field on the Administration # Settings # Model Deployment Settings page.
- Models - Users can now deploy large ML models. Model size is not limited to 50 MB anymore.
- Model Registry - Administrators can now find the Machine User Workload User Name that is needed for configuring their model registries to access their S3 or ADLS Gen2 bucket on the Workspace Details page.
- Install Workspaces (Tech Preview) - Auto-retriable workflow for Install Workspace for Azure Private Cluster. See the Feature Preview doc for more information.

- Usage Monitoring - Usage data for Spark Executors is now recorded in the CSV file that Site Administrators can download from the Administration Usage tab.
- API Keys - Improved security by storing Legacy API keys as hashes in the database. Existing Legacy API keys are automatically rotated as part of the upgrade process to ensure that previous keys cannot be used. API v1 keys will not be usable after the upgrade. This does not impact the Models API, and Models authentication is not affected. To manually rotate a Legacy API key, do the following:
  - In User Settings API Keys , click Rotate to generate a new Legacy ApiKey and ApiKeyHash pair.
  - Copy the Api Key that is shown after rotation and use it in future requests.
  - Note - The API key will not be visible on the UI once you refresh the page. Make sure to copy it before leaving the page.
- Modify Instance Group Type - Administrators can easily change the CPU or GPU instance types of node groups for a CML workspace, without having to re-provision the workspace.

Fixed Issues

- Model Registry (DSE-25909) - Removed the 19 character limit for model names in Model Registry.
- Model Registry (DSE-25641) - Fixed an issue where redeploying models from model registry to a CML workspace may fail.
- Model Registry (DSE-24683) - Added workload name to Model Registry service account list.
- Model Registry (DSE-25906) - Model Registry is no longer limited to listing 10 models at a time.

April 5, 2023

Release notes and fixed issues for version 2.0.36-H4.

Fixed Issues

- Azure (DSE-26517) - Fixed an issue that caused a failure when upgrading CML Azure workspaces to AKS 1.25.

March 27, 2023

Release notes and fixed issues for version 2.0.36-H3.

New Features / Improvements

- Kubernetes - EKS 1.24 (AWS) and AKS 1.25 (Azure) are now supported.
- AWS - The eu-south-1 region is now supported.

Fixed Issues

- Entitlements (DSE-25941) - Fixed an issue where regional control planes could not access entitlements.

February 14, 2023

Release notes and fixed issues for version 2.0.36-H2.

Fixed Issues

- Model Registry (DSE-25331) - Fixed an issue where users may see an error message x509: Certificate signed by unknown authority on the Model Registry page in the CML workspace application. Note that if you are seeing this error, or if you create a model registry after you create one or more CML workspaces in an environment, you must synchronize the model registry with the workspaces.

February 10, 2023

Release notes and fixed issues for version 2.0.36-H1.

New Features / Improvements

- Model Registry (Tech Preview) - The Model Registry is a new CML Product to store and manage machine learning models and associated metadata, such as the model's version, dependencies, and performance. The registry enables MLOps and facilitates the development, deployment, and maintenance of machine learning models in a production environment.

  Note: Model Registry is not supported with R models.

February 7, 2023

Release notes and fixed issues for version 2.0.36.

New Features / Improvements

- Custom Runtime Addons - This feature enables administrators to mount shared dependencies like connection drivers or configuration files to all CML workloads. You can start by following the documentation.
- Cancellable CML Workspace Backup - Administrators now can cancel in-progress backups and resume the CML Workspace.
- AWS Support - The AWS Jakarta and Hong Kong regions are now supported.
- Kubernetes - EKS 1.23 (AWS) and AKS 1.24 (Azure) are now supported.

Fixed Issues

- Applications (DSE-22857) - Fixed an issue where unauthenticated users could access the terminal session in a running public application if the application was configured to allow anonymous access to the terminal endpoint. Now anonymous access to the terminal endpoint is redirected to an authentication page.
- Applications (DSE-24105) - Fixed an issue where the Application List page may not load correctly when there are a large amount of Applications accessible by the user.

November 29, 2022

Release notes and fixed issues for version 2.0.34.

New Features / Improvements

- Suspend/Resume Workspace - Administrators can optimize cloud costs by suspending CML Workspaces for weekends and resume operation for business hours. See Suspend and resume ML workspaces for details.
- PBJ Runtimes - PBJ Workbench Runtimes are now GA, providing a more consistent experience with the Jupyter ecosystem. With the elimination of proprietary code you can now build a new Runtime from scratch, on your custom base image with your selected kernel, with no need to build on top of a Cloudera image anymore.
- Experiment tracking - A new Experiments feature built on MLFlow is now available. You can now track your experiments in CML sessions with the preinstalled mlflow library and visualize and compare them in the CML application. This feature is enabled by default.
- Customizable Scratch Space - You can now configure the amount of ephemeral storage space (also known as scratch space) a CML session, job, application or model can use. This feature helps in better scheduling of CML pods, and provides a safety valve to ensure runaway computations do not consume all available scratch space on the node. By default, each user pod in CML is allowed to use up to 10 GB of scratch space.
- Iceberg support - Iceberg v2 is supported via Spark Runtime Addon, based on the CDE 1.17-h1 Runtime Addon version.

- Jobs UI - The Jobs List page and Job Details page now display the job ID and Created At time.
- Projects UI - When you select a runtime kernel on the project creation page, only the latest standard version of the selected kernel is added to the project.
- Data tab - The Cloudera Data Visualization application that provides the Data tab experience got upgraded to v7.0.2. You can review the changes here.
- Applications - Custom polling endpoints for Applications can now be specified in the UI.
- Register new runtimes via APIv2 - Site administrators can register new ML Runtimes using the APIv2.
- Runtime addon management - Site administrators can now also disable or deprecate specific Runtime Addons.
- Environmental variables - The new project environmental variable PROJECT_OWNER holds the username of the project owner.

Fixed Issues

- Python packages (DSE-21313) - Fixed an issue where Python packages installed via pip install may not be imported correctly until a new session is started.
- Jobs (DSE-21771) - Fixed an issue where Jobs scheduled with PBJ Runtimes may terminate with a Success state even when there were errors reported in job runs.
- PBJ Runtimes (DSE-19971) - Fixed an issue where comment blocks may not be rendered correctly in sessions or jobs launched with PBJ Runtimes.
- Models (DSE-22527) - Fixed an issue where the Model Monitoring Chart was displaying data from other projects or deployments.

October 19, 2022

Release notes and fixed issues for version 2.0.32-H4.

New Features / Improvements

- In-place Upgrades - In-place upgrades are now available for CML versions 2.0.29 and 2.0.30.
- Azure backups - When creating snapshots for an Azure backup, the resource group name defined in the environment can be used (if present) as part of the snapshot name.

Fixed Issues

- Environmental Variables (DSE-23098) - The environment variable WORKLOAD_PASSWORD is redacted in logs.
- Environmental Variables (DSE-23499) - Environment variables are redacted from grpc request logging.
- CVE (DSE-23534) - Upgraded several libraries to fix critical CVE vulnerabilities.
- UI (DSE-22128) - Improved behavior of New Project and New Session buttons.
- Secret Generator (DSE-22696) - Fixed an issue that was causing secret generator slowness.

September 27, 2022

Release notes and fixed issues for version 2.0.32-H3.

New Features / Improvements

- Iceberg support - CML Snippets now fully support the Iceberg v1 table format for all Spark, Hive, and Impala data connections. See Create an Iceberg data connection for details.
- Data connections - Data Hubs are now automatically discovered and connections are created for them in newly-created CML Workspaces.
- Backup / Restore Workspace - The default timeout is now 12 hours, and the estimated time to complete a backup (from the cloud provider) is now periodically added to the event logs.

August 30, 2022

Release notes and fixed issues for version 2.0.32-H2.

New Features / Improvements

- Private Cluster on Azure - This option is temporarily disabled. (CDPAM-3279)
- Backup / Restore Jobs - Timeouts can now be customized in both the Backup Workspace and Restore Workspace UI.

Fixed Issues

- Workspace Installation (DSE-22545) - Increased the workspace installation timeout limit to 1 hour from 30 min
- CML Workspace (DSE-22638) - Fixed an issue where CML workspace creation using a machine user caused an error.

July 21, 2022

Release notes and fixed issues for version 2.0.32.

New Features / Improvements

- Garbage collection for deleted projects - This feature allows you to trigger cleanup of deleted projects. A separate feature allows older orphaned projects to be marked for cleanup. For more information, see Project Garbage Collection.
- Disable Runtimes - It is now possible to disable and enable runtimes. For more information, see Disabling Runtimes.
- Monitoring for Applications - This feature allows you to monitor the technical health of deployed Applications, including statistics and visualizations of CPU and memory usage. For more information, see Monitoring Applications.
- Custom polling endpoints for applications - This feature allows the application creator to define what application endpoint servers poll to detect if the application is running, that avoids problems some applications have with polling the root endpoint. For more information, see Application polling endpoint.
- PBJ Workbench Runtimes (Tech Preview) now work with Sessions, Experiments, Jobs and Applications - This feature enables the classic workbench UI backed by the open-source Jupyter protocol. This architectural change improves consistency, stability, and ease of customization while eliminating the dependency on proprietary CML code. For more information, see *PBJ Workbench* in Preview Features.
- Kubernetes - Kubernetes 1.22 is now supported for both AWS and Azure.

Fixed Issues

- Job quotas (DSE-12664) - Fixed an issue where subsequent jobs in a job pipeline can fail if quota is enabled.
- HDFS via Python (DSE-19775) - Fixed an issue where accessing HDFS via Python libraries that connect natively to HDFS, such as Tensorflow or PyArrow, may fail due to an error that the libhdfs.so file cannot be found.

May 31, 2022

Release notes and fixed issues for version 2.0.30.

New Features / Improvements

- ML Discovery and Exploration, SQL and Visualization (GA) - This feature enables Data Scientists to understand their data using a SQL editor and drag-and-drop Visual Dashboards within CML. Users can start with their

pre-configured Data Connections and create Datasets that they can rely on for model development. For more information, see ML Discovery and Exploration.
- CML Endpoint Stability - This feature adds the ability for CML admin to define the prefix for the URL for the CML Workspace. This enables a new CML Workspace to be created and leverage the endpoint of a previously deleted CML Workspace. This ensures that deployed models and applications deployed in the new Workspace will have the same endpoint as the same models and applications deployed in the old Workspace. For more information, see *CML Static Subdomain* in Provisioning ML Workspaces.
- Add/Delete GPU Nodes - This feature enables MLAdmins to reconfigure CML Workspaces by adding or removing GPU Worker groups for existing deployments.

April 21, 2022

Release notes and fixed issues for version 2.0.29.

New Features / Improvements

- ML Discovery and Exploration, SQL and Visualization (Preview) - This feature enables Data Scientists to understand their data using a SQL editor and drag-and-drop Visual Dashboards within CML. Users can start with their pre-configured Data Connections and create Datasets that they can rely on for model development. For more information, see *ML Discovery and Exploration*.
- Model metrics visualization - This feature allows Data Scientists and Machine Learning Engineers to monitor technical metrics relating to their running models, such as resource consumption and request throughput, within Cloudera Machine Learning.

Control Plane Application

- Azure Files - Support for Azure Files NFS - TP
- Public Load Balancer - This feature allows MLAdmins to configure a Public Load Balancer for a CML workspace but with fully private (Private EKS API Server endpoint) EKS deployment (AWS only).

Fixed Issues

- Pagination widget (DSE-19937) - Fixed an issue where the pagination widget on the Session list page may not function as expected.
- Notification emails (DSE-20085) - Fixed a bug where Job report recipients who subscribed to notification emails when their jobs terminated, may receive notification emails for termination statuses that they did not subscribe to.
- Project list page (DSE-1975) - Fixed a bug where projects may not be sorted correctly on the project list page when users use the Created By field for sorting.

Related Information

Provisioning ML Workspaces

March 14, 2022

Release notes and fixed issues

New Features / Improvements

- Site Administration - The Project Creation Timeout (minutes) settings on the Site Administration Settings page now applies to project creation via both git clone and fork. Consider increasing this value in case you need to create a large project. If you have tried forking projects before doing this and they get "stuck," the only way to delete these projects is to manually navigate to the <user>/<project-name> URL, and select Project Settings Delete .

- Load Balancer subnet - Support specifying subnets for the Load Balancer created by CML on AWS. The Workspace Provisioning page now has a field for specifying a subnet for the Load Balancer. See *Provisioning a Workspace* for more information.

Fixed Issues

- Editor (DSE-11423) - Fixed an issue so that the editor, session, or workbench windows resize when the browser window resizes.

February 10, 2022

Release notes and fixed issues

New Features / Improvements

- ML Discovery and Exploration - Data Connections and Snippets are now Generally Available. CML workspaces now automatically discover data connections within the CDP environment and offer connection snippets for users. For more information, see ML Discovery and Exploration.
- ML Runtimes - You can now filter the list of ML Runtimes that can be used in a given project.
- Model Technical Metrics visualization is now available in CML (Technical Preview).
- API v2 - You can now specify an input data example when you create a model build.
- Backup and Restore - CLI-based Backup and Restore of CML workspaces is now available as a Preview Feature. AWS only.
- Kubernetes - Kubernetes 1.21 is now supported on Azure.

Fixed Issues

- API v2 (DSE-18782) - You can change the Runtimes and Spark versions (specified via the runtime addon setting) associated with a job, by using the Update Job API endpoint.

January 12, 2022

Release notes and fixed issues

New Features / Improvements

- AWS - Resolved an issue with the way certain AWS regions (such as eu-south-1) identify region specific endpoints, which caused workspace creation to fail.

December 15, 2021

Release notes and fixed issues

New Features / Improvements

- CML - Released remediation for CVE-2021-44228, the Apache Log4j2 vulnerability. See CVE-2021-44228 Remediation for CML for details.
- Engines / Runtime Addons - Released versions containing fixed version of Apache Log4j2.

December 13, 2021

Release notes and fixed issues.

New Features / Improvements

- What's New - CML Users can now discover and read about new features, blog posts, FFL research reports from within CML without leaving the product.
- Non-Transparent Proxy (Preview) - Non-transparent proxy support is now available for AWS.
- API v2 - API v2 now supports Spark 2 and 3 via ML Runtime Addons.
- Spark - Spark Dynamic Allocation Lite is now supported in CML.
- Projects - Job lineage/dependencies are now maintained when forking a project.
- User Management - CML Admins can now create CML Teams that have their membership synchronized with CDP Groups.
- User Management - Users who no longer have access to a Workspace are deactivated within CML after a user sync.
- Unified Diagnostics - Unified diagnostics are now integrated in CML support bundles.

Fixed issues

- Username display - Fixed an issue where the user’s full name was left blank if it was not available from the identity provider. Now, if it is not available, the username is displayed next to the avatar.
- Terminal access button spins forever (DSE-18456) - Fixed an issue where the terminal access button on an Azure cluster failed to become ready.

October 27, 2021

Release notes and fixed issues.

New Features / Improvements

- ML Flow - Cloudera Machine Learning now supports experiment tracking using open MLflow standards. For more information, see *CML Experiment Tracking through MLflow (Preview)* in Preview Features.
- Spark 3 - Users can now use Spark 3 in Sessions, Jobs, Models or Applications in their projects configured to use ML Runtimes.
- cdswctl - Users can now use the cdswctl CLI client to create Sessions with Spark.
- Root volume size - On existing CML workspaces based on AWS, there is a known issue where the provisioned root volume capacity is not enough to accomodate CML deployments with K8s versions >= 1.21. To mitigate this issue, when upgrading existing workspaces with a K8s version older than 1.19 on AWS, the root volume size is automatically increased/updated to 128 GB. Please note that this is not an issue on Azure where the pre-provisioned capacity for workspaces is already 128GB
- Workbench Editor - The Workbench now supports wrapping lines in the editor window. Users can enable this by selecting the Line Wrapping Enabled option under the View menu item in the Workbench.
- Private AWS clusters - Support for Private AWS clusters is now available in Preview mode. With Private AWS cluster support, CML workspaces are provisioned with a Private K8s API server making the deployments more secure.
- Pre-flight checks - NFS Server pre-flight checks have been added for Azure where validations are done to make sure that the specified BYONFS server directory meet the following criteria NFS server directory specified should be:
    - Reachable
    - Mountable
    - The directory must be owned by CML user 8536.
    - The directory must be empty.

Fixed issues

- Atlas (DSE-16706) - Fixed an issue with reporting data lineage of CML model deployments to Atlas.

October 18, 2021

Release notes and fixed issues.

New Features / Improvements

- Kubernetes 1.20 - Kubernetes 1.20 is now supported on both AWS and Azure.

Fixed issues

- Workspace email (DSE-17615) - Fixed an issue where sending email from a CML workspace failed.
- Sessions (DSE-17776) - Fixed an issue where forking a project with a job failed.
- Data lake upgrade (DSE-17687) - Fixed an issue where upgrading the data lake failed if it had been replaced after it was initially installed.

August 31, 2021

Release notes and fixed issues.

New Features / Improvements

- API v2 - A new API for operations on projects, jobs, models, and applications is now generally available.
- ML Workflow Discovery and Exploration (Preview) - This feature accelerates the ML development workflow with preconfigured data connections and readily available snippets. See ML Discovery and Exploration for more information.
- Kubernetes 1.19 - Kubernetes 1.19 support has been added on AWS.

Fixed issues

- RStudio editor issue (DSE-16843) - Fixed an issue where users may see a blank page when launching Sessions with RStudio editor using legacy engines.
- Spark (DSE-17279) - Fixed an issue that could cause sessions to fail to start if the Enable Spark option was not selected.
- SSH tunnels (DSE-17406) - Fixed an issue where users could not create a new SSH tunnel on the Project Settings > SSH tunnels page.
- GPU node (DSE-17114) - The default GPU node was changed to a p3 generation instance.

August 23, 2021

Release notes and fixed issues.

New Features / Improvements

- AMPs and Runtimes - Runtimes requirements for tasks can now be specified in AMPs.
- API v2 - A new API endpoint for creating models is available.
- API v2 - An ephemeral API key, which is valid during the lifespan of Sessions/Jobs/Applications/Experiments, for accessing APIv2 endpoints is now available in the environmental variable CDSW_APIV2_KEY.
- Customer Master Keys - CMK encryption is now supported for AWS environments. For more information, see CMK Encryption on AWS (Preview) in Preview Features.
- User-level environmental variables - Users can now set environmental variables at the user level, which are available in all Sessions, Jobs, Applications, Experiments, and Models launched by the user.
- Session launch UX improvement - The console output page now shows autoscaling information of the CML workspaces when scheduling new Sessions, Jobs, Experiments or Applications

Fixed issues

- ML Runtimes (DSE-16429) - Fixed an issue so that warnings are displayed for models during runtime migration in the case of an engine type mismatch.
- ML Runtimes (DSE-14834) - Fixed an issue so that when workers are added with ML Runtimes, they are initialized with the correct runtime
- ML Runtimes (DSE-16163) - Fixed an issue where the New Session > Configure button opened the Runtime/ Engine page, even if the Runtimes feature was disabled.
- ML Runtimes (DSE-16790) - Fixed an issue where the Add Runtime button on the Runtimes Catalog page is displayed for regular users.
- Spark (DSE-16714) - Enabled/disabled state of a Spark addon is now remembered across sessions.
- Grafana dashboards (DSE-16615) - Fixed an issue where Grafana dashboards did not appear on public cloud workspaces.

July 8, 2021

Release notes and fixed issues.

July 8, 2021

Fixed issues

- Large files (DSE-16884) - Fixed a bug that could truncate large file uploads or lose changes to large files.
- Custom editors (DSE-16570) - Fixed a bug that could cause unexpected UI refreshes when using custom editors such as JupyterLab.

June 24, 2021

Release notes and fixed issues.

June 24, 2021

- Engine Deprecation - Cloudera ML Runtimes are the default and recommended solution to run user workloads. New projects will be created with ML Runtimes configured by default and we recommend migrating existing projects to use ML Runtimes. Legacy Engines are deprecated and will be removed in a future release but workloads running on them remain fully supported.
- Register customized Runtime - Administrators can register an externally built Runtime to provide Data Scientists with a customized environment.
- Support user API key rotation - Administrators can rotate keys for all users, or users can rotate their own keys, just by clicking on a button.
- AMP specification for Runtimes - You can specify which Runtimes to use in AMPs.
- New base engine released - Engine:14-cml2021.05-1 is now available.
- Configurable engine image for Jobs - You can specify which engine to use for a Job. Jobs use the Project engine by default.
- Web session timeouts - Timeout limits for User and Admin User web sessions were changed to be a period of inactivity, instead of a set time limit.

Fixed issues

- Spark configuration (DSE-16422) - Fixed an issue where user-added Spark configurations in the ~/spark-defaults.conf file may not be populated to /etc/spark/conf/spark-defaults.conf correctly.
- Hive table (DSE-16108) - Fixed an issue where saving data to a managed Hive table using Hive Warehouse Connector via CDSW from CML may fail with a StreamCorruptedException.
- API keys (DSE-15678) - Fixed an issue where user API keys may be accessible through non-private projects.
- Web sessions (DSE-11394) - Fixed an issue where user web session may not be refreshed even when the user is actively using CML.

Related Information

Managing ML Runtimes

Pre-installed Packages in Engines

May 13, 2021

Release notes and fixed issues

May 13, 2021

New features

- Business User Experience - A new user role, ML BusinessUser, provides restricted access to view Applications created in CML.
- ML Runtimes - Runtimes provide a lightweight alternative to Engines.
- View All Applications - Admins can view all applications on the application list page.
- Jobs improvements - Job Scheduling UI now supports cron expressions, job notification email subject line now includes the project name, and the Job History page now shows detailed start and end timestamps.

Fixed issues

- DSE-15059 - Fixed an issue where project creators may not be able to authenticate to models created by project collaborators using model API key.

March 22, 2021

Release notes and fixed issues

March 22, 2021

New features

- CML Model DNS resolution - For CML on Azure and AWS, DataHub and CDE connectivity to CML services such as workspaces and model endpoints is now possible. This is supported by enabling internal DNS resolution for CML workspace endpoints and models on private IPs. However, any existing workspaces must be upgraded to the current release to obtain this functionality.
- RAZ integration - CML now integrates with RAZ for S3.
- Cdswctl support - ML Runtimes is now supported in `cdswctl`.
- External engine repositories - Admins can now add basic Docker credentials for external engine repositories.
- ML workspace provisioning - In case of validation errors, it is possible to skip preflight validation for ML workspace provisioning. See Provisioning ML Workspaces for more information.
- Cloud permissions - CML now supports configuring AWS environment cross-account roles with fine-grained cloud permissions and policies for provisioning and running workspaces. See *AWS restricted policies* for more information.

Fixed issues

- DSE-14481 - CML no longer needs Network Security group ports to be open for EFS on AWS (port 2049).
- DSE-14685 - Clarified impact of changing Load Balancer Source Range settings on Security Group settings. See Provisioning ML Workspaces for details.

February 3, 2021

Release notes and fixed issues

February 3, 2021

New features

- Applied ML Prototypes - Applied ML Prototypes (AMPs) are now generally available inside the product, which provide end-to-end prototypes to help kickstart real customer use cases. AMPs effectively package pre-built models for data scientists to tailor for their own needs, as well as enable data scientists to learn how the different parts of CML work together.
- Kubernetes Support - CML AWS now supports K8s 1.18, keeping up with updated EKS support.
- NFS version picker - In Provision Workspace, the NFS version field was changed to a pick list with the default version set to 3. Make sure that this is the correct version to use, otherwise the workspace installation will fail.

Fixed issues

- DSE-13866 - Export Session List includes duration information.
- DSE-13254 - Python 2 kernel is enabled or disabled on Job Update and Create/Update Application pages based on Admin setting.
- DSE-14251 - Fixed an issue where a workspace could not be upgraded if its configuration was modified.
- DSE-13928 - Fixed an issue where revoking a user's access to an Application may not take effect immediately.
- DSE-14170 - Fixed an issue where TLS certificate renewal may not succeed for ML workspaces that have been upgraded.

December 21, 2020

Release notes and fixed issues

December 21, 2020

New features

- Added support for re-configuring ML workspaces:
  - Autoscaling Group (ASG) limits
  - Authorized IP ranges for k8s api server
  - Load Balancer Source Ranges and allowed list
- AWS and Azure instance types expanded - Support for additional AWS and Azure instance types

Fixed issues

- DSE-13904 - Fixed an issue where CML workspace installation may take up to 10 minutes longer when the Autoscaling Group minimum is set to 0.
- DSE-13898 - Updated TGT image to fix an issue so that the freeIPA HA release retries on failure.
- CDPSDX-2673 - Added a Retry step to the login context to reduce the chance of PRE_AUTH failures.

November 23, 2020

Release notes and fixed issues

New features

- Updated Projects Dashboard - Upon logging in to CML, users have a new streamlined page to be able to access their Projects. Projects can be displayed in a Summary card-based view for direct access, or in a Detail table-based view to be able to jump directly to different workloads (Sessions, Experiments, Models, etc) within a particular Project. In addition, the Resource Usage Details can be displayed or hidden as needed.
- Scaling Workers on Azure - On Azure, the number of worker nodes can now be scaled to zero when the workspace is idle.
- ML Runtimes - As an alternative to the existing Engines in CML, ML Runtimes are more lightweight than the current monolithic Engines. By specifying the desired Editor, Kernel, Edition, and Version, a streamlined Runtime will be used to run the user's code in Sessions, Jobs, Experiments, Models, or Applications.
- Asynchronous Project Creation - New Project creation, in particular via Git or forking from an existing project, now executes in the background so that the user does not have to wait.

- Sessions List Export - The date format on the Sessions List Export has been updated to include the complete date plus hours, minutes, seconds and a decimal fraction of a second.
- Custom root volume size - The root volume size of CML workspaces can now be specified on the Provision Workspace UI. Also, the default root volume size of CML workspaces on AWS has been increased from 40 GiB to 96 GiB.

Fixed issues

- DSE-13246 - Fixed an issue where cdswctl version always returns "unknown".
- DSE-11685 - The Export PDF function is improved and will display charts, tables, and maps. However, some content still might not display well, such as tables with many columns and dynamic content, video, or iframe content containing dynamic content.

October 29, 2020

Release notes and fixed issues.

New features

- Scaling workers on AWS - On AWS, the number of worker nodes can now be scaled to zero when the workspace is idle.
- Kubernetes - Access to the Kubernetes API server of the ML workspaces can now be restricted to certain authorized IP ranges.
- ADLS Gen2 - Access to ADLS-Gen2 from CML can now be managed with Ranger policies.
- Subnets - Subnets used for ML workspaces can now be reviewed in the Workspace Details page.
- ML workspace upgrade - You can now retry ML workspace upgrade if the previous attempt fails.
- Deprecated Kubernetes versions - ML workspaces with deprecated Kubernetes versions will now be highlighted on the CDP control plane. It is recommended to upgrade your ML workspaces with the deprecated Kubernetes versions at the earliest opportunity.
- Enable Unauthenticated Access To Applications - Users can now create Analytical Applications in CML that can be accessed by unauthenticated users. This functionality can also be wholly disabled for an ML workspace by an administrator.
- Applications header - CML now supports authentication using REMOTE_USER HTTP header for the Applications created on CML.
- Audited events - Quota changes and “Download All” files are now audited in the user_events table.
- Sessions page redesign - The Sessions page has been improved, making it easier to take actions such as filtering down to running sessions and performing bulk deletion of old sessions.

Fixed issues

- DSE-12164 - Fixed a bug where applications may not use the engine image configured with the project upon restart.
- DSE-12033 - Fixed an issue where the Engine profile on the Job Overview page did not update when the profile was changed on the Job Settings page.
- DSE-12026 - Grafana deployments now include 8536 in the supplemental groups list.
- DSE-11988 - Fixed a UI issue when creating a new session, where the user had to delete Untitled Session before typing in the session name.
- DSE-11945 - Fixed an issue where a poor network connection could cause an “Engine Exited” error message.
- DSE-11911 - Fixed an issue where updating the Project name caused navigation to other tabs to fail.
- DSE-11891 - Fixed an issue where attempting to open a file with Chinese characters in the name fails.
- DSE-10408 - Fixed an issue where quota checks may cause sessions or jobs to fail.

August 04, 2020

Release notes and fixed issues

New features

- Session Start UX - Sessions start page is now displayed with a fresh design.
- Engine Schedule - Improve the page loading performance for the “Lag” view of the Site Administration Activity tab.
- Subnet Selection - Subnets (one per workspace on Azure, multiple subnets per workspace on Amazon) can now be, optionally, manually selected when creating a new workspace.
- Model Metrics Development Workflow - You can develop and test model functions that track model metrics in a workbench session without actually deploying the model. This method makes it easy for you to develop and test code that tracks metrics and the code which consumes the tracked metrics without having to rebuild and redeploy a model to test every change.
- Default Model API Key Expiration - Administrators can now set a default expiration for Model API keys. If the user sets a longer expiration date, an error is returned. The default expiration duration is set at Admin Security Default Model API keys expiration in days .

Fixed issues

- DSE-11770 - Fixed an issue where the SHM size configuration was not applied for sessions started via `cdswctl`.
- DSE-11436 - Fixed an issue where SSH connections to CML terminal via the `cdswctl ssh-endpoint` command may time out after a few minutes.
- DSE-11321 - Fixed an issue where uploading files that have Chinese characters in the file names may fail.
- DSE-11089 - Fixed an issue where accessing Applications in the browser directly results in a 401 error and a blank page. Now users are redirected to a login page, and then directed back to the Application after successful authentication.
- DSE-10977 - Fixed an issue where the Remote Editing link on the Teams page may display a 500 error.
- DSE-10784 - Fixed an issue where Applications hosted on CML may be stopped after 7 days.
- DSE-10616 - Fixed an issue where the Jobs dependency pipeline only shows the first job.
- DSE-10031 - Fixed an issue where Windows line endings were converted automatically to Linux line endings in the Workbench. This behavior is now configurable.
- DSE-9986 - Fixed an issue so that users cannot create public projects via API when public project creation is restricted.
- DSE-8698 - Fixed redirections for CDSW links when TLS is enabled.
- DSE-6068 - Fixed an issue where sessions, jobs. experiments, and models may fail to be created if the engine tag includes an underscore character.

June 30, 2020

Release notes and fixed issues

New features

- CML Test Drive - CML now includes guided interactive modules for setting up workspaces, running data science workflows, and getting models to production. These workflows are designed with real-world business use case prototypes.

- Base Engine v12 - The default base engine is now version 12 (12-cml-2020.06-2).

  Key changes:

  - Python 3.6.10 (was 3.6.9)
  - R 3.6.3 (was 3.6.2)
  - CDPD 7.1.0 parts (Spark, Hadoop, Hive, Avro, Parquet)
  - Some bundled Python and R data science library upgrades (see Pre-installed Packages in Engines).

  Bug fixes and improvements:

  - DSE-10250 - Fixed a bug that could cause an job or experiment to exit with status 7 if used to run a very short script.
  - DSE-9872 - Fixed a bug that prevented users from using the `locate` command.
  - DSE-8601 - The Python launch_workers command now defaults the kernel argument match the kernel of the current session (previously it used "python3" in Python 2 sessions).
  - Reduced the engine image size by approximately 400 MB

Fixed issues

- DSE-11395 - Shared memory size configuration is now applied to sessions, jobs, and experiments.
- DSE-11433 - Fixed an issue where files opened in the Workbench editor are deselected when a new session starts.
- DSE-11421 - Fixed an issue where ML workspace provisioning can fail when a GPU instance that is not available in the AWS region is selected.

June 9, 2020

Release notes and fixed issues

New features

- Production ML - Production ML functionality is available to all accounts by default.
- Firefox support - CML now supports Firefox on newly provisioned CML workspaces.
- Monitoring - Grafana dashboard now shows the user ID to help administrators monitor usage and resources.
- Session UI - Added a new pop-up window to the session start UI.

Fixed issues

- DSE-10985 - Fixed a bug where non-admin users were not able to start or list experiments.
- DSE-10893 - Fixed a bug where upgrading a monitoring-enabled CML workspace may flush the imported Grafana dashboards.
- DSE-10671 - Fixed a link on the Team page to go to Team settings.
- DSE-10618 - Fixed a bug where Engine Profile is not displayed on the Jobs page.
- DSE-10407 - Fixed a bug where selecting the Python2 kernel causes problems when creating an experiment or job.
- DSE-10405 - Fixed a bug where forking a CML project caused the project to be copied repeatedly until interrupted.
- DSE-10185 - Fixed a bug where CML Applications can get stuck in the “Starting” state due to user-specified subdomains that are not handled by CML correctly.
- DSE-10054 - (SAML 2.0 SSO) Fixed a bug to prevent arbitrary redirection of browsers for SSO to third-party applications.

May 5, 2020

Release notes and fixed issues

- General Availability of Production ML features - including:
  - Model Metrics and Monitoring
  - Model Technical Metrics
  - Model Governance
  - CLI for Models
  - Model API Key for securing Models
- CML on Azure Updates
  - Custom GPUs are now supported.
  - Availability Zones are no longer required to provision an ML Workspace.
- Enable Monitoring - Monitoring can be enabled when provisioning ML workspaces by CLI.

Fixed issues

- DSE-10631 - Fixed Default Quota initialization issue.
- DSE-10441 - Fixed a bug where NFS input field is not marked as a required field on the Provision Workspace page.
- DSE-10399 - Fixed a UI bug on the CML Models pages on Firefox.
- DSE-8836 - Fixed a bug where Jobs & Applications usage are not metered correctly.
- DSE-8779 - Fixed tooltip typo on UI which mixed up CML and CDSW.
- DSE-7944 - Upgraded Kubernetes to defend against CVE 2019-11253.

April 14, 2020

Release notes and fixed issues

- Command line arguments - The Create a Job UI now has an Arguments field where you can enter command line arguments for the executed script. This feature is available for R and Python scripts.
- User information required - Administrators can now require users to enter information when starting sessions, for audit purposes.
- UI update - In Site Administration, the Activity tab is now renamed to Usage.
- Accessibility - Improvements for W3C WCAG 2-AA Compliance.
- Grafana dashboards - A Grafana dashboard is available in Models.
- Cross-Origin Resource Sharing (CORS) - CORS is now disabled by default. Enable CORS if you have web applications on different domains that need cross-domain communication with the Cloudera Machine Learning API.

Fixed Issues

- DSE-9063 - Fixed an issue where a custom quota takes effect even if the quota feature is turned off.
- DSE-9469 - Fixed an issue where email attachments greater than 4 MB caused problems.
- DSE-6375 - Fixed an issue where creating a private project from a Git repo fails, even with an authorized team SSH key.
- DSE-9041 - Improved the SMTP settings for test email.
- DSE-9694 - Fixed an issue where the order of columns in the Workspace instances table was inconsistent.
- DSE-8601 - The kernel argument launch_workers should default to kernel of current session.
- DSE-9810 - Fixed an issue where users could not download files that have non-ASCII characters in the file name.

March 16, 2020

Release notes and fixed issues

- Cloudera Machine Learning on Azure - Cloudera Machine Learning workspaces can now be created in environments hosted in Azure. See the documentation for information on Azure network planning, limitations, and setup.

- Base Engine v11 - This release adds Hive Warehouse Connector (HWC) to the base engine (11-cml1.4-hwc-v1). Additionally, the base engine now has the ability to connect to DWX. No changes are made to the included library versions.
- Navigation improvement - Cloudera Machine Learning users can click the nine-dot icon to navigate to the Cloudera Manager console.
- Changing user names and addresses - In Cloudera Machine Learning instances where user authentication is performed through single sign-on (SSO), users cannot change their names and addresses in CML workspaces.
- Jupyter enabled by default - Fixed an issue where Jupyter was not enabled by default in the default base engine.

Fixed Issues

- DSE-8791 - Fixed an issue where Job or Model containers might not start or restart correctly.
- DSE-9264 - Fixed an issue where the Regenerate Access Key button on the Model Settings page does not work for non-admin users.

February 13, 2020

Release notes and fixed issues

- Kerberos Authentication Improvements - Previously, users needed to manually authenticate themselves by entering their CDP FreeIPA credentials into their CML workspaces (on the Account settings Hadoop Authentication page). This is no longer required.

  Users will now automatically receive the Kerberos credentials required for any CML workloads such as sessions, jobs, models, etc. Existing workspaces can be upgraded to take advantage of this improvement.

  Additionally, the Hadoop Authentication tab has been removed from the workspace UI.
- Workspace Monitoring Enabled - When you provision a Machine Learning workspace, Monitoring is enabled by default under Advanced Options.

January 30, 2020

Release notes and fixed issues

- Base Engine v11 - The default base engine is now version 11. The only change to the included libraries is the R interpreter, which is updated to version 3.6.2.
- Python 2 checkbox disabled - Python 2 sessions are now disabled by default on new clusters, but can be re-enabled by admin users.
- Load Balancer Source Ranges - When provisioning a Machine Learning workspace, in Advanced Settings, you can enter the CIDR range of IP addresses allowed to connect to the workspace. You must include the entire IP pool in the allowlist for your VPC to ensure that terminal sessions can connect to the workspace.
- Updated Open Workbench button name - When launching a session, the Open Workbench button has been renamed to New Session.
- Login error message - Fixed a bug where users might see the following error message upon login to a CML workspace: "Email already associated with an account."

December 19, 2019

Release notes and fixed issues

- Monitoring Workspaces with Grafana - CML now leverages Prometheus and Grafana to provide a dashboard that allows you to monitor how CPU, memory, storage, and other resources are being consumed by ML workspaces.
- Custom Quotas - CML workspace site administrators can now enable custom quotas to set resource usage per user.

- Tags - There are three new AWS resource tracking tags:
  - Cloudera-Resource-Name: <The CRN of the associated CML Workspace for which the resource was provisioned.>
  - Cloudera-Environment-Resource-Name: <The CRN of the Environment in which the resource was created.>
  - Cloudera-Creator-Resource-Name: <The CRN of the CDP user who requested creation of the resource.>

  AWS resource tags are set by default. They can be searched and viewed through the AWS console or CLI. These tags are helpful for tracking resource usage and cost.
- Granting remote access - The procedure to grant and revoke remote access to ML workspaces is improved. You can easily add new users. You can also see which users currently have access, and then quickly revoke access to specific users.
- CML instance type cost reduced - The node used to run the CML application was downsized to a more economical AWS instance type. The instance type was changed from m5.12xlarge to m5.4xlarge, which should result in a noticeable reduction in cloud costs.
- Base Engine v10-cml1.3 - The default engine is now v10. See the package listing for the updated versions of included libraries.
  - Python 3 is version 3.6.9 (was 3.6.8 in Engine v8).
  - Python 2 is version 2.7.17 (was 2.7.11 in Engine v8).

November 1, 2019

Release notes and fixed issues

- Analytical Applications - CML now gives data scientists a way to create long-running standalone ML web applications/dashboards that can easily be shared with other business stakeholders.
- Quotas - CML workspace site administrators can now enable CPU, GPU, and memory usage quotas per user. Quotas must be enabled separately for each workspace.

  Note: The Quotas feature is in Technical Preview.
- Diagnostic Bundles - CML now allows site administrators to download diagnostic bundles from the Site Admin panel.
- UI Improvements and Changes
  - You can now display the Details page by clicking on the Workspace Name in the ML Workspaces page.
  - The ML Workspace Details page now contains a Events tab.

    The Events tab displays high-level events for your workspace. You can click View Logs to display additional log information about the action.
  - The ML Workspace Details page now displays AWS workspace tags.

    The tag information displays both default tags and any tags you have specified. You can specify these workspace tags in the provision workspace page, under the Advanced options. The default workspace tags include:
    - Creator
    - Environment
    - Owner
    - WorkspaceName

September 23, 2019

Release notes and fixed issues

- Single Sign-on (SSO) Changes - You no longer need to create separate authorization groups to grant users SSO access to workspaces. Authorization groups are now managed per-environment using the MLAdmin and MLUser resource roles.

- Resource Tags - You can now add resource tags to all the cloud infrastructure, compute, and storage resources used by an ML workspace you provision.
- Remove Workspaces - New options added that allow you to retain project storage (in EFS) and force delete workspaces from CDP.
- View Workspace Details - Each workspace now has an associated details page where you can access links to the workspace itself, the environment where the workspace was created, and the underlying Kubernetes cluster (links to cloud service provider). A link to this page is available under the Actions menu.
- Search and Filter workspaces - New filter that allows you to display workspaces in a specific environment. You can also search for a workspace by name.

August 22, 2019

Release notes and fixed issues

This marks the General Availability (GA) release of Cloudera Machine Learning.

Compatibility for Cloudera AI and Runtime components

Learn about Cloudera AI and compatibility for Runtime components across different versions.

Table 1: Cloudera AI compatibility with Runtime component details

Cloudera AI	Compatible Spark Versions	Compatible Python Versions	Compatible Data Lake Version
2.0.53-b273	2.4.8	2.7, 3.4 - 3.7	Date Lake 7.2.16 or higher
	3.2.3	3.6 - 3.9	Date Lake 7.2.16 or higher
	3.3.0	3.7 - 3.10	Date Lake 7.2.16 or higher
	3.5.1	3.8 - 3.11	Date Lake 7.2.18 and 7.3.1
2.0.53-b241	2.4.8	2.7, 3.4 - 3.7	Date Lake 7.2.16 or higher
	3.2.3	3.6 - 3.9	Date Lake 7.2.16 or higher
	3.3.0	3.7 - 3.10	Date Lake 7.2.16 or higher
	3.5.1	3.8 - 3.11	Date Lake 7.2.18 and 7.3.1
2.0.52-b<TBA>	2.4.8	2.7, 3.4 - 3.7	Date Lake 7.2.16 or higher
	3.2.3	3.6 - 3.9	Date Lake 7.2.16 or higher
	3.3.0	3.7 - 3.10	Date Lake 7.2.16 or higher
	3.5.1	3.8 - 3.11	Date Lake 7.2.18 and 7.3.1

Table 2: Cloudera AI compatibility with Runtime component details

Cloudera AI	Spark 2.4.8	Spark 3.2.3	Spark 3.3.0	Spark 3.5.1
2.0.52-b34	• Date Lake 7.2.16 to higher • Python 3.7	• Date Lake 7.2.16 to higher • Python 3.9	• Date Lake 7.2.16 to higher • Python 3.10	• Date Lake 7.2.18 and 7.3.1 • Python 3.11

Table 3: Cloudera AI compatibility with Runtime component details

Cloudera AI	Data Lake	Supported Spark Version
2.0.52-b27	7.2.16 to higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0 • Spark 3.5.1
2.0.50-b68	7.2.16 to higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0 • Spark 3.5.1
2.0.50-b52	7.2.16 to higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0 • Spark 3.5.1
2.0.47-b365	7.2.16 to higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0 • Spark 3.5.1
2.0.47-b360	7.2.16 to higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0 • Spark 3.5.1
2.0.47-b359	7.2.16 to higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0 • Spark 3.5.1
2.0.47-b345	7.2.16 to higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0 • Spark 3.5.1
2.0.47-b302	7.2.16 or higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0 • Spark 3.5.1
2.0.46-b238	7.2.16 or higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0 • Spark 3.5.1
2.0.46-b210	7.2.16 or higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0
2.0.46-b200	7.2.16 or higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0
2.0.45-b86	7.2.16 or higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0

Cloudera AI	Data Lake	Supported Spark Version
2.0.45-b82	7.2.16 or higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0
2.0.45-b81	7.2.16 or higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0
2.0.45-b76	7.2.16 or higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0
2.0.45-b54	7.2.16 or higher	• Spark 2.4.8 • Spark 3.2.3 • Spark 3.3.0
2.0.43-b233	7.2.16 or higher	
2.0.43-b229	7.2.16 or higher	
2.0.43-b220	7.2.16 or higher	
2.0.43-b208	7.2.16 or higher	
2.0.41-b225	7.2.15 or higher	
2.0.40-b157	7.2.14 or higher	
2.0.40-b150	7.2.14 or higher	
2.0.38-b126	7.2.14 or higher	
2.0.38-b125	7.2.14 or higher	
2.0.38-b121	7.2.14 or higher	
2.0.38-b101	7.2.14 or higher	
2.0.36-b121	7.2.14 or higher	
2.0.36-b118	7.2.14 or higher	
2.0.34-b116	7.2.14 or higher	
2.0.32-b123	7.2.14 or higher	
2.0.30-b114	7.2.14 or higher	
2.0.29-b61	7.2.11 or higher	
2.0.27-b64	7.2.11 or higher	
2.0.26-b180	7.2.11 or higher	
2.0.25-b110	7.2.11 or higher	
2.0.24-b100	7.2.11 or higher	

Upgraded Cloudera AI deployments keep multiple Hadoop CLI addon versions that administrators can configure to maintain compatibility between Cloudera AI and the Data Lake.

Spark support in Cloudera AI

The end of support (EoS) policy for Spark is the same in both Cloudera Data Engineering and Cloudera AI. For information about end of support for Spark, see *Cloudera Data Engineering Runtime end of support* .

Related Information

Cloudera Data Engineering Runtime end of support

Apache Parquet CVE-2025-30065

On April 1, 2025, a critical vulnerability in the parquet-avro module of Apache Parquet (CVE-2025-30065, CVSS score 10.0) was announced.

Cloudera has determined the list of affected products, and is issuing this TSB to provide details of remediation for affected versions.

Upgraded versions are being released for all currently affected supported releases of Cloudera products. Customers using older versions are advised to upgrade to a supported release that has the remediation, once it becomes available.

Vulnerability Details

Exploiting this vulnerability is only possible by modifying the accepted schema used for translating Parquet files and subsequently submitting a specifically crafted malicious file.

CVE-2025-30065 | **Schema parsing in the parquet-avro module of Apache Parquet 1.15.0 and previous versions allows bad actors to execute arbitrary code.**

CVE: NVD - CVE-2025-30065

Severity (Critical): CVSS:4.0/AV:N/AC:L/AT:N/PR:N/UI:N/VC:H/VI:H/VA:H/SC:H/SI:H/SA:H

Impact

Schema parsing in the parquet-avro module of Apache Parquet 1.15.0 and previous versions allows bad actors to execute arbitrary code. Attackers may be able to modify unexpected objects or data that was assumed to be safe from modification. Deserialized data or code could be modified without using the provided accessor functions, or unexpected functions could be invoked.

Deserialization vulnerabilities most commonly lead to undefined behavior, such as memory modification or remote code execution.

Releases affected

Cloudera AI on cloud

- All versions

Mitigation

Until Cloudera has released product version with the Apache Parquet vulnerability fix, please continue to use the the mitigations listed below:

Customers with their own FIM Solution:

1. Utilize a File Integrity Monitoring (FIM) solution. This allows administrators to monitor files at the filesystem level and receive alerts on any unexpected or suspicious activity in the schema configuration.

General advisory:

1. Use network segmentation and traffic monitoring with a device capable of deep packet inspection, such as a network firewall or web application firewall, to inspect all traffic sent to the affected endpoints.
2. Configure alerts for any suspicious or unexpected activity. You may also configure sample analysis parameters to include:
    - Parquet file format “magic bytes” = PAR1
    - Connections from sending hosts that are not expected source IP ranges.

3. Be cautious with Parquet files from unknown or untrusted sources. If possible, do not process files with uncertain origins or that can be ingested from outside the organization.
4. Ensure that only authorized users have access to endpoints that ingest Parquet files.

For the latest update on this issue, see the corresponding Knowledge article: Cloudera Customer Advisory 2025-847: Cloudera's remediation actions for Apache Parquet CVE-2025-30065

CVE-2021-44228 Remediation for Cloudera AI Data Service

The procedure to remediate CVE-2021-44228 on Cloudera AI Data Service is described in this document.

On December 15 2021, the ML team released version 2.0.25-b110 of Cloudera AI Data Service for Cloudera on cloud. It addresses CVE-2021-44228 which affects Apache Log4j2 versions 2.0 through 2.14.1. We urge all customers to upgrade their workbenches to the latest version.

The Cloudera AI service code itself is not written in Java and hence is not vulnerable. However, the `log4j2` jar file does exist in the (now deprecated) engine as well as the Hadoop CLI and Spark ML Runtimes addons. As a result, the `log4j2` jar file is available in sessions, jobs, models, and applications. Because of this, there is no direct threat, but a data scientist could inadvertently use this `log4j2` jar file and expose themselves to this vulnerability.

As a result, we are releasing a new version of the engine and ML Runtimes addons that remove this vulnerability.

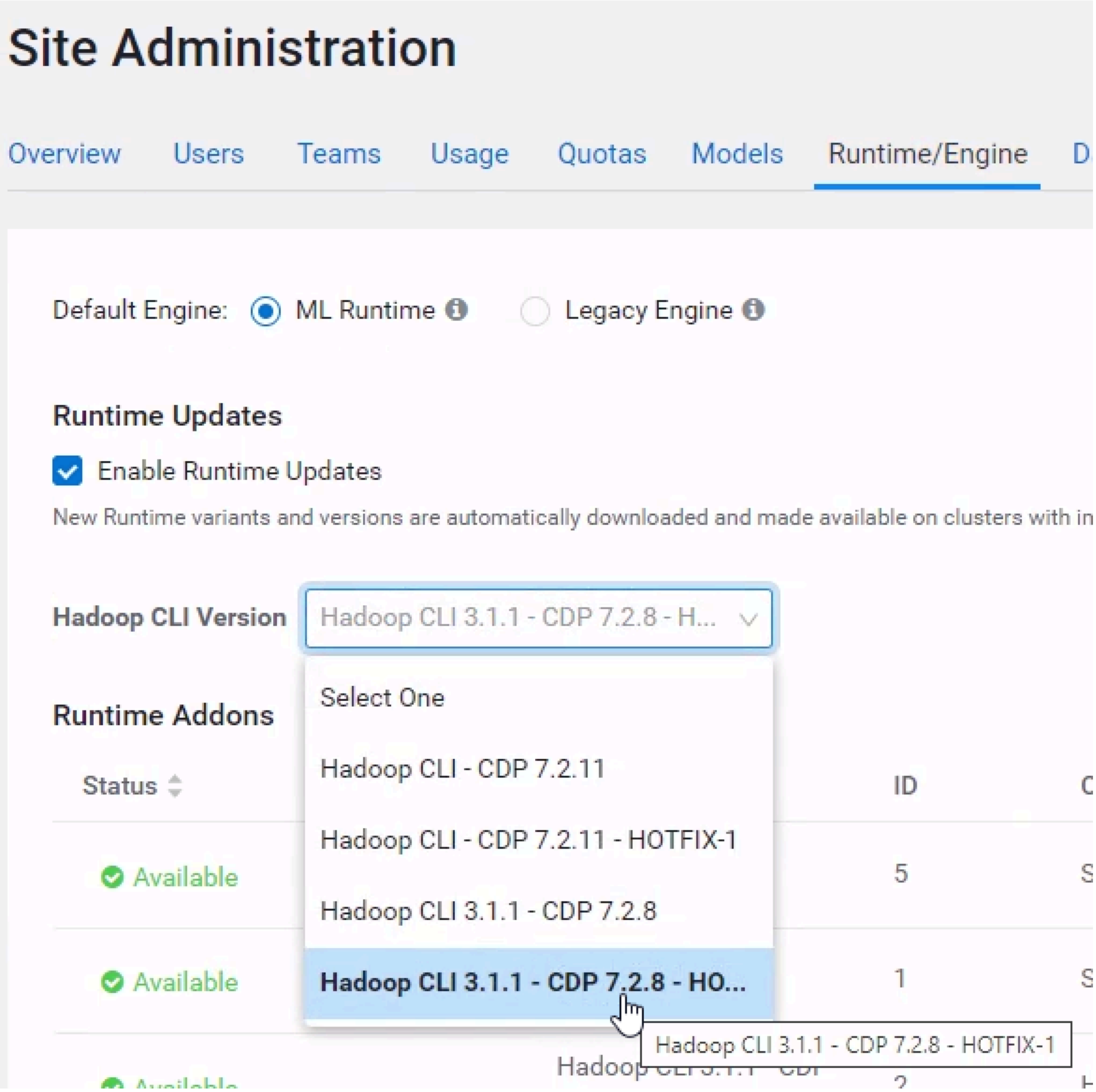
Upgrade Cloudera AI Workbench to the new version

Administrators can now upgrade the Cloudera AI Workbenches to version 2.0.25-b110. To upgrade the workbench, select the Actions icon and select Upgrade Workbench.

Once the workbench is upgraded, you can follow the steps below to ensure the appropriate engine and ML Runtimes addons are used.

ML Runtimes Add-ons

The first step involves the use of runtime addons. These addons consist of Hadoop CLI code and Spark code that is added to ML Runtimes. An administrator must go to the Site Administration -> Runtime/Engine page and ensure that in the “Hadoop CLI Version” drop down box, the chosen selection ends with “HOTFIX-1”.



Site Administration

Overview Users Teams Usage Quotas Models Runtime/Engine D

Default Engine: ● ML Runtime ○ Legacy Engine

Runtime Updates

☑ Enable Runtime Updates

New Runtime variants and versions are automatically downloaded and made available on clusters with in

Hadoop CLI Version: Hadoop CLI 3.1.1 - CDP 7.2.8 - H...

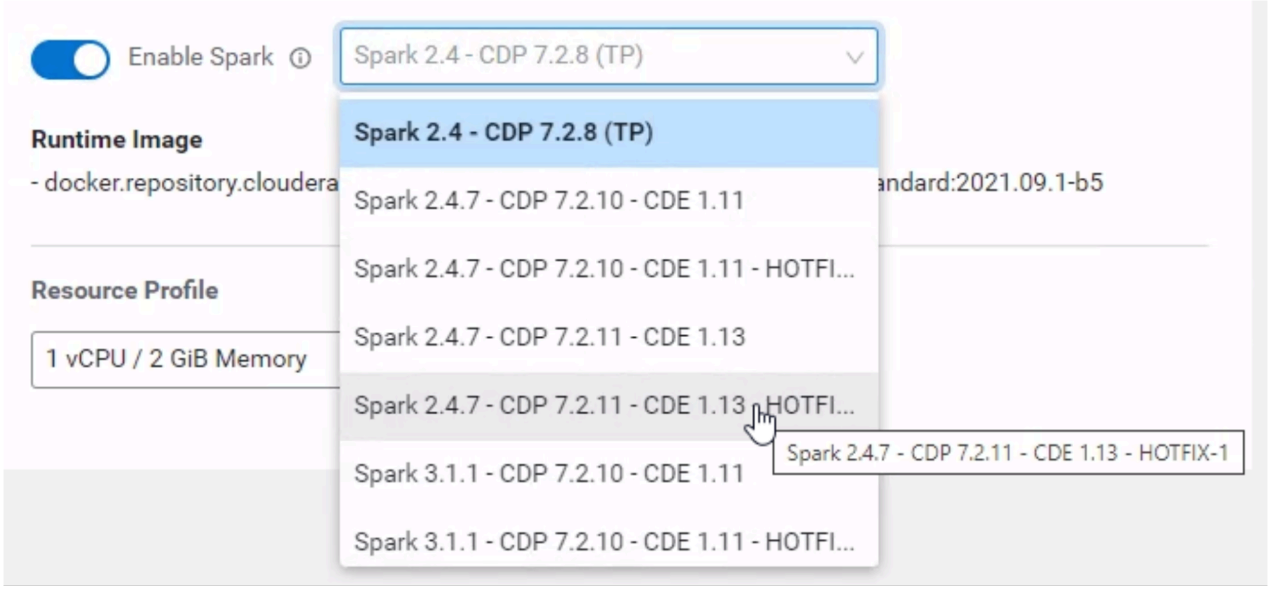
- Select One
- Hadoop CLI - CDP 7.2.11
- Hadoop CLI - CDP 7.2.11 - HOTFIX-1
- Hadoop CLI 3.1.1 - CDP 7.2.8
- **Hadoop CLI 3.1.1 - CDP 7.2.8 - HO...**

Hadoop CLI 3.1.1 - CDP 7.2.8 - HOTFIX-1

Runtime Addons

Status	ID
Available	5
Available	1
Available	2

When starting a session, if “Enable Spark” is turned on, users must use a version of Spark ending with “HOTFIX-1”.


Enable Spark
Spark 2.4 - CDP 7.2.8 (TP)
Spark 2.4 - CDP 7.2.8 (TP)
Spark 2.4.7 - CDP 7.2.10 - CDE 1.11
Spark 2.4.7 - CDP 7.2.10 - CDE 1.11 - HOTFI...
Spark 2.4.7 - CDP 7.2.11 - CDE 1.13
Spark 2.4.7 - CDP 7.2.11 - CDE 1.13 - HOTFI...
Spark 2.4.7 - CDP 7.2.11 - CDE 1.13 - HOTFIX-1
Spark 3.1.1 - CDP 7.2.10 - CDE 1.11
Spark 3.1.1 - CDP 7.2.10 - CDE 1.11 - HOTFI...

Runtime Image

- docker.repository.cloudera andard:2021.09.1-b5

Resource Profile

1 vCPU / 2 GiB Memory

Jobs and applications that use Spark in projects using runtimes, ensure the job or application uses a correct version of the Spark addon by going to the job/application Settings page and selecting a version of Spark that ends with “HOTFIX-1”. Your applications will require a restart.

Models and Experiments that use Spark runtimes must be rebuilt and redeployed with a fixed version of Spark. The models and experiments do not need to be deleted.

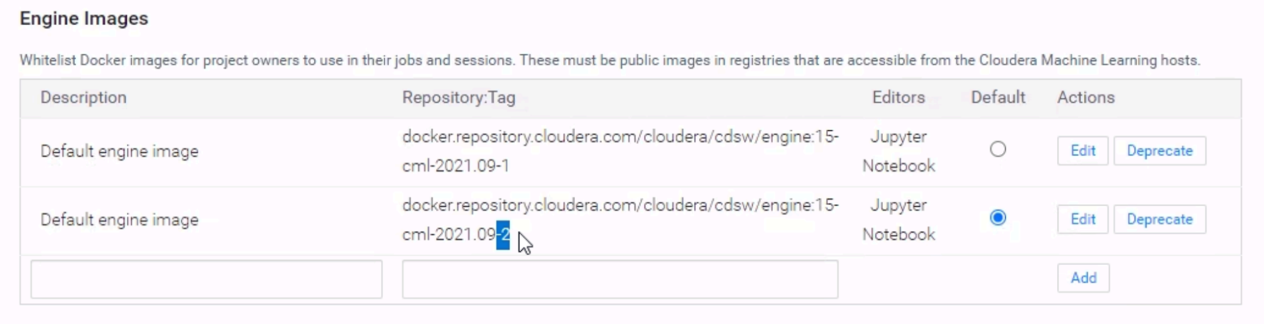
Engines

For customers using projects with (deprecated) engines we have released a new engine version which fixes the vulnerability:

```
docker.repository.cloudera.com/cloudera/cdsw/engine:15-cml-2021.09-2
```

The “-2” at the end is important. It is identical to the engine ending in “-1”, except the vulnerability has been removed. This engine is now the default for new workbenches. To ensure this engine is used:

- An administrator should go to Site Administration Runtime/Engine and scroll down to Engine Images. Ensure the above version is selected as the default.
- For all projects using engines, go to Project Settings Runtime/Engine and verify the engine is set to the version above.
- For all applications and jobs in projects that use engines, go to the application or job settings page and ensure that under Select Job Engine the above version is selected.
- Deploy a new build for all models that use engines.



Engine Images

Whitelist Docker images for project owners to use in their jobs and sessions. These must be public images in registries that are accessible from the Cloudera Machine Learning hosts.

Description	Repository:Tag	Editors	Default	Actions
Default engine image	docker.repository.cloudera.com/cloudera/cdsw/engine:15-cml-2021.09-1	Jupyter Notebook	○	Edit Deprecate
Default engine image	docker.repository.cloudera.com/cloudera/cdsw/engine:15-cml-2021.09-2	Jupyter Notebook	●	Edit Deprecate
				Add

Technical Details of the Fix

The `log4j2` jar file exists in several places and is also packaged inside other jar files. Instead of upgrading the `log4j2` jar file, we have chosen to remove the vulnerable java class file. This is one of the mitigations proposed in the CVE text:

"it can be mitigated ... by removing the JndiLookup class from the classpath (example: zip -q -d log4j-core-*.jar */JndiLookup.class)."

As a result, older versions of the `log4j2` jar file still exist in sessions after the fix, but the offending class file has been removed. To verify this, you can run:

```
grep JndiLookup.class <jar file>
```

Known Issues and Limitations

There are some known issues you might run into while using Cloudera AI.

Limited GPU support for Boltz2 Runtime (DSE-48889)

The new Boltz2 Runtime, upgraded to version 1.3.0, currently only supports deployments configured with a single NVIDIA L40S GPU (1xL40S). Attempting to deploy this runtime on other GPU types (for example, A100, V100) or configurations with multiple L40S GPUs will result in deployment or execution failures.

Additionally, to ensure proper function with the new runtime, the environment variable NIM_MAX_POLYMER_LENGTH has been set to the value of 1536.

AI Registry model import failure behind proxy (NTP Setup) (DSE-48642)

When Cloudera AI Registry is deployed within an NTP (Non-Transparent Proxy) setup, attempts to import models from Model Hub (such as NVIDIA or Hugging Face) may fail with a 401 Unauthorized error. This occurs because the Knox service, which gates the Cloudera AI Registry, does not honor the HTTPS_PROXY environment variable, preventing successful communication with the Control Plane API over the proxy.

Workaround: Manually configure the Knox deployment to honor the necessary proxy settings. This requires access to the cluster's kubeconfig file and the proxy server details.

1. Retrieve proxy details and construct the string:
   a. Go to your environment page where the AI Registry is deployed.
   b. In Summary, go to the Proxy section.
   c. Click on the Proxy value to view the Server Host and Server Port. Take note of these two values.
   d. Use the collected proxy details to create the following string:

      ```
      -Dhttps.proxyHost=<Server Host> -Dhttps.proxyPort=<Server Port>
      ```

      For Example:

      ```
      Dhttps.proxyHost=10.80.123.45 -Dhttps.proxyPort=6789
      ```

2. Edit the knox deployment and append the Proxy string:

a. Get the kubeconfig of your AI Registry.

b. Run the following command to edit the Knox deployment:

```
kubectl edit deployment knox -n knox
```

c. Search for the environmentVariable KNOX_GATEWAY_DBG_OPTS in that deployments file.

d. Append the string you created to the value of KNOX_GATEWAY_DBG_OPTS. It is highly recommended to enclose the entire value in double quotes.

For example, if original value was -Dcom.sun.jndi.ldap.object.disableEndpointIdentification=true, then the string would be:

```
"Dhttps.proxyHost=10.80.123.45 -Dhttps.proxyPort=6789 -Dcom.sun.jndi.lda
p.object.disableEndpointIdentification=true"
```

e. Save the changes to the deployment. This action automatically restarts the Knox pod.

f. Verify if the new pod is running:

```
kubectl get pods -n knox
```

g. Restart model-registry-v2 pods:

```
kubectl rollout restart deployment model-registry-v2 -n mlx
```

h. Once the new pod is up, attempt to import the model from NVIDIA or Hugging Face again. The import should now succeed.

Note: Whitelisting URLs: Ensure that the necessary NVIDIA and Hugging Face URLs are whitelisted in your NTP environment's firewall configuration (example, Squid ACL rules). Below are sample access control list (ACL) rules for common services:

```
acl allowed_http_sites dstdom_regex -i (^|\.)nvidia\.com
acl allowed_http_sites dstdom_regex -i (^|\.)huggingface\.co
acl allowed_http_sites dstdom_regex -i (^|\.)hf\.co
```

User or Team Sync status in UI does not update in real-time after sync trigger (DSE-43688)

After triggering a user or team sync, the sync status in the UI does not update instantly. Instead, it is refreshed based on a 30-second polling interval, causing a delay before the updated sync status is displayed in the UI.

Resource Group selection not visible without GPU profile (DSE-48683)

When editing or redeploying an existing workload (Job, Application, or Model), the Resource Group that was previously selected is not displayed. This display issue affects all workloads when the workbench does not have a GPU Resource Group configured in the Control Plane. The selected values are still active but are not loaded or visible in the UI edit screens.

Workaround: An Administrator must configure at least one GPU Resource Group for the workbench on the Control Plane. Once a GPU profile exists, the correct Resource Group values will be loaded and displayed in the workload edit screens.

Session Timeout (SESSION_MAXIMUM_MINUTES) does not stop specific Runtime sessions (DSE-47697)

The environment variable `SESSION_MAXIMUM_MINUTES`, which is designed to enforce a maximum duration for Workbench sessions, currently does not correctly terminate sessions under the following conditions:

- JupyterLab Editor: The session is running with the JupyterLab Editor and if the Runtime is from the Runtime release 2025.01.1 or later.

- PBJ Custom Runtime: The session is running with a PBJ Custom Runtime that utilizes a custom editor.

Events and Logs unavailable for Cloudera AI Registry (DSE-47205)

Events and Logs information is not displayed in the UI for a fresh installation of Cloudera AI Registry. However, logs for other events, such as Upgrade and Renew, are displayed correctly.

JWT token not refreshing before expiry in Cloudera AI Workbench (DSE-41395)

In the Cloudera AI Workbench, the JSON Web Token (JWT) does not automatically refresh before the existing token expires. This can lead to unexpected session interruptions and connection failures when using the workbench.

Workaround: An administrator can manually extend the lifetime of the JWT token and prevent premature expiry by modifying the following configuration setting in the Knox service.

1. Navigate to the environment where your workbench is provisioned.
   a. Go to Environment Details # Datalake # Cloudera Manager UI # knox.
2. Find the setting named `knox_token_kerberos_ttl_ms`.
3. It is recommended to set the value of this setting to a number larger than 20 minutes (1,200,000 milliseconds).
4. After setting the new value, restart the Knox service for the change to take effect.

Note: The value is set in milliseconds (ms). For example, setting the value to 30 minutes would be 1,800,000 ms.

Resource profiles are not automatically updated after an instance type change (DSE-47568)

If an administrator modifies the instance type of a Resource Group, the associated Resource Profiles will not automatically reflect the new instance capacity.

Modifying the Persistent Volume size fails in private cluster (DSE-46334)

When attempting to modify the volume size for a Persistent Volume Claim (PVC) in private clusters, the modify operation fails with the Helm upgrade failure: another operation is in progress error, even if the UI appears updated.

Note: The feature is currently disabled for private clusters.

Sessions stuck in Scheduling state (DSE-48087)

After launching a new Session, the browser interface may intermittently get stuck in the Scheduling state. This prevents the Session from starting automatically.

Workaround:

- (Immediate) Refresh the page in your browser to immediately unstick the Session and allow it to finish launching.
- (Permanent) An administrator can permanently resolve this issue by restarting the web pods in the Kubernetes cluster.

Note: This issue is fixed in release 2.0.52-b27.

Cloudera AI Workbench upgrade intermittently fails (DSE-46502)

In certain edge cases, Cloudera AI Workbench upgrades may fail with the following error:

```
Status: Upgrade Failed
```

```
failed to execute post-upgrade processes: failed to scale up nodes: failed
 to scale up nodes: rpc error: code = Internal desc = A workbench update is
 currently in progress please try again later
```

Workaround: If you encounter this issue, use the Retry Upgrade Workbench option in the Cloudera AI UI to restart the upgrade workflow.

Upgrade of Cloudera AI Registry in Azure with UDR-enabled subnet fails with subnet error (DSE-46225)

Cloudera AI is not selecting the correct subnet when upgrading the Cloudera AI Registry, which causes the upgrade to fail.

Workaround: Migrate your Cloudera AI Registry database. The following steps are required to migrate your Cloudera AI Registry database, which includes generating a database dump from an existing instance and restoring it to a new AI Registry.

1. Obtain the kubeconfig file for your Cloudera AI Registry using the following steps:
   a. In the Cloudera console, click the Cloudera AI tile.

      The **Cloudera AI Workbenches** page displays.
   b. Click AI Registries under Administration on the left navigation menu.

      The AI Registries page displays.
   c. In the AI Registries page, click ⋮ in the Actions column.
   d. Click Download Kubeconfig.
2. Access the database pod of your existing Cloudera AI Registry instance.

   ```
   kubectl exec -it model-registry-db-0 -n mlx -- bash
   ```
3. Connect to the PostgreSQL sense database and verify the schema migration status.

   ```
   psql -U sense
   \c sense
   \dt
   ```

   A successful schema migration will display a list of relations similar to the following:

   ```
                   List of relations
    Schema |         Name          | Type  | Owner
   --------+-----------------------+-------+-------
    public | config                | table | sense
    public | model_permissions     | table | sense
    public | model_registry_users  | table | sense
    public | model_versions        | table | sense
    public | models                | table | sense
    public | schema_migrations     | table | sense
    public | tags                  | table | sense
   ```
4. Generate database dump. The type of database dump you generate depends on whether the schema migration is complete.
   - If the schema migration is confirmed (output matches the example above), generate a data-only dump of the sense database:

     ```
     pg_dump -U sense --data-only sense > /tmp/dump.sql
     ```
   - If the output differs (schema migration is not complete), generate a full database dump:

     ```
     pg_dump -U sense sense > /tmp/dump.sql
     ```

5. Copy the generated dump file from your local machine to the new Cloudera AI Registry database pod:

   ```
   kubectl cp mlx/model-registry-db-0:/tmp/dump1.sql ./dump1.sql
   ```

6. Set up the new environment for data restoration.

   a. Delete the old Cloudera AI Registry, as you can only have one Cloudera AI Registry in any given environment.

   b. Create a new Cloudera AI Registry instance. For information on creating an AI Registry in an Azure subnet with UDR, see Creating a Cloudera AI Registry on an Azure UDR Private Cluster.

   c. Configure kubectl to point to the kubeconfig of the newly created Cloudera AI Registry.

7. Once the new Cloudera AI Registry is set up, copy the dump.sql file from your local machine to the new Cloudera AI Registry database pod:

   ```
   kubectl cp <abs path of dump.sql> model-registry-db-0:/tmp/ -n mlx
   ```

8. Restore the database within the new Cloudera AI Registry pod.

   a. Execute into the new Cloudera AI Registry database pod:

   ```
   kubectl exec -it model-registry-db-0 -n mlx -- bash
   ```

   b. Verify the presence of the dump file inside the /tmp/ directory of the new pod:

   ```
   ls /tmp/dump.sql
   ```

   c. Run the command to restore the database by applying the dump to the sense database:

   ```
   psql -U sense sense < /tmp/dump.sql
   ```

When opening a Cloudera AI Workbench from the Control Plane for the first time, it might take five or more minutes to load. (DSE-41691)

This is caused by a known issue where the code for Cloudera AI Workbench checks to see that the Data Lake is available and online. If the Data Lake is not in the *READY* (green) status, then this code will prevent the Cloudera AI Workbench from loading.

Workaround: Ensure that every node in the Data Lake is running completely and upgrade to the latest version of Cloudera AI.

Unable to deploy ONNX optimization profiles for embedding and ranking NIMs on GPUs with optimization profiles (DSE-40509)

Deploying ONNX profiles for embedding and ranking NIMs on GPUs where compatible GPU profiles exist will lead to deployment failure. Before deploying an ONNX optimization profile for embedding or ranking NIMs from the Model Hub, ensure that the NIM does not have a supported profile for the target GPU.

Cloudera AI does not support non-transparent proxy with authentication (DSE-36512)

Cloudera AI on cloud does not support non-transparent proxy with authentication. While configuring proxy using the Cloudera console, do not specify your username and password.

Unable to paginate in Cloudera AI Workbench backup table (DSE-41406)

When selecting a backup in the Workbench backup table, pagination to the next page does not work as expected. This is due to an issue in the cuix library.

Workaround: To avoid the issue, do not try to navigate to the next page after selecting a backup. Instead, refresh the page to view details of other Workbench backups.

Too many Federated Identity Credentials have been assigned to the Managed Identity. error is displayed in the Event Log file (DSE-41063)

Each Azure managed identity, that is, Logger Identity in a Cloudera environment can have a maximum of 20 federated identity credentials. Because each cluster requires 1–2 federated identities, there is a maximum number of clusters that can be created per environment. If you exceed the maximum number of clusters, Too many Federated Identity Credentials have been assigned to the Managed Identity." error is displayed.

Workaround: Perform the following:

1. Log into the Microsoft Azure portal.
2. Go to Managed Identity and select the logger identity you defined of your resource group.
3. Click Setting > Federated Credentials .
4. In the Federated Credentials page, delete the unused federated identity credentials (with no AKS cluster associated with it).

Cloudera AI automatic JWT authorization to Cloudera Data Warehouse is failing due to a wrong KNOX URL (DSE-39855)

Due to an issue, there is a mismatch of the Data Lake name between the actual Data Lake name in the environment and the one parsed by the CML 2.0.43-b208 version or later.

Workaround:

1. Obtain the correct Data Lake version by running the following command using CDP CLI:

   ```
   cdp datalake describe-datalake
   ```
2. Override the KNOX URL in the environment variable by performing the following:
   a. Run the following command to save the deployment status to a file:

      ```
      kubectl get deployment ds-cdh-client -o json -n mlx > /tmp/rs.json
      ```
   b. Edit the /tmp/rs.json file and add the below object for ds-cdh-client environment under the spec.template.spec.containers.env section.

      ```
      {
           "name": "FIXED_KNOX_URL",
           "value": "https://[***ENVIRONMENT-VARIABLE***]/value"
      }
      ```
   c. Apply the configuration.

      ```
      kubectl apply -f /tmp/rs.json
      ```

Python workloads running multiple-line comments or strings fail (DSE-41757)

Python workloads running multiple-line comments or strings might fail to run when using the Workbench Editor..

Workaround: Run the code using the PBJ Workbench Editor.

Web pod crashes if a project forking takes more than 60 minutes (DSE-35251)

The web pod crashes if a project forking takes more than 60 minutes. This is because the timeout is set to 60 minutes using the `grpc_git_clone_timeout_minutes` property. The following error is displayed after the web pod crash:

```
2024-04-23 22:52:36.384   1737    ERROR      AppServer.VFS.grpc
       crossCopy grpc error    data = [{"error":"1"},{"code":4,"details":"2"
,"metadata":"3"},"Deadline exceeded",{}]
```

```
["Error: 4 DEADLINE_EXCEEDED: Deadline exceeded\n    at callErrorFromStatus
 (/home/cdswint/services/web/node_modules/@grpc/grpc-js/build/src/call.js:31
:19)\n    at Object.onReceiveStatus (/home/cdswint/services/web/node_modules
/@grpc/grpc-js/build/src/client.js:192:76)\n    at Object.onReceiveStatus (/
home/cdswint/services/web/node_modules/@grpc/grpc-js/build/src/client-interc
eptors.js:360:141)\n    at Object.onReceiveStatus (/home/cdswint/services/we
b/node_modules/@grpc/grpc-js/build/src/client-interceptors.js:323:181)\n    
at /home/cdswint/services/web/node_modules/@grpc/grpc-js/build/src/resolving
-call.js:94:78\n    at process.processTicksAndRejections (node:internal/proc
ess/task_queues:77:11)\nfor call at\n    at ServiceClientImpl.makeUnaryReque
st (/home/cdswint/services/web/node_modules/@grpc/grpc-js/build/src/client.j
s:160:34)\n    at ServiceClientImpl.crossCopy (/home/cdswint/services/web/no
de_modules/@grpc/grpc-js/build/src/make-client.js:105:19)\n    at /home/cdsw
int/services/web/server-dist/grpc/vfs-client.js:235:19\n    at new Promise (
<anonymous>)\n    at Object.crossCopy (/home/cdswint/services/web/server-dis
t/grpc/vfs-client.js:234:12)\n    at Object.crossCopy (/home/cdswint/service
s/web/server-dist/models/vfs.js:280:38)\n    at projectForkAsyncWrapper (/ho
me/cdswint/services/web/server-dist/models/projects/projects-create.js:229:1
9)"]
node:internal/process/promises:288
            triggerUncaughtException(err, true /* fromPromise */);
            ^Error: 4 DEADLINE_EXCEEDED: Deadline exceeded
    at callErrorFromStatus (/home/cdswint/services/web/node_modules/@grpc/
grpc-js/build/src/call.js:31:19)
    at Object.onReceiveStatus (/home/cdswint/services/web/node_modules/@grp
c/grpc-js/build/src/client.js:192:76)
    at Object.onReceiveStatus (/home/cdswint/services/web/node_modules/@gr
pc/grpc-js/build/src/client-interceptors.js:360:141)
    at Object.onReceiveStatus (/home/cdswint/services/web/node_modules/@grp
c/grpc-js/build/src/client-interceptors.js:323:181)
    at /home/cdswint/services/web/node_modules/@grpc/grpc-js/build/src/resol
ving-call.js:94:78
    at process.processTicksAndRejections (node:internal/process/task_queu
es:77:11)
for call at
    at ServiceClientImpl.makeUnaryRequest (/home/cdswint/services/web/node
_modules/@grpc/grpc-js/build/src/client.js:160:34)
    at ServiceClientImpl.crossCopy (/home/cdswint/services/web/node_modules/
@grpc/grpc-js/build/src/make-client.js:105:19)
    at /home/cdswint/services/web/server-dist/grpc/vfs-client.js:235:19
    at new Promise (<anonymous>)
    at Object.crossCopy (/home/cdswint/services/web/server-dist/grpc/vfs-
client.js:234:12)
    at Object.crossCopy (/home/cdswint/services/web/server-dist/models/vfs
.js:280:38)
    at projectForkAsyncWrapper (/home/cdswint/services/web/server-dist/model
s/projects/projects-create.js:229:19) {
  code: 4,
  details: 'Deadline exceeded',
  metadata: Metadata { internalRepr: Map(0) {}, options: {} }
}
```

Workaround: Increase the timeout limit using the `grpc_git_clone_timeout_minutes` property. For example, 120 minutes.

```
UPDATE site_config SET grpc_git_clone_timeout_minutes = <NEW VALUE>;
```

Enabling Service Accounts (DSE-32943)

Teams in the Cloudera AI Workbench can only run workloads within team projects with the Run as option for service accounts if they have previously manually added service accounts as a collaborator to the team.

Working with files larger than 1 MB in Jupyter causes error (OPSAPS-61524)

While working on files or saving files of size larger than 1 MB, Jupyter Notebook may display an error message such as 413 Request Entity Too Large.

Workaround:

Clean up the notebook cell results often to keep the notebook below 1 MB. Use the kubectl CLI to add the following annotation to the ingress corresponding to the session.

```
annotations:
        nginx.ingress.kubernetes.io/proxy-body-size: "0"
```

1. Get the session ID (the alphanumeric suffix in the URL) from the web UI.
2. Get the corresponding namespace:

   ```
   kubectl get pods -A | grep <session ID>
   ```
3. List the ingress in the namespace

   ```
   kubectl get ingress -n <user-namespace> | grep <session ID>
   ```
4. In the metadata, add the annotation.

   ```
   kubectl edit ingress <ingress corresponding to the session> -n <user-nam
   espace>
   ```

Terminal does not stop after time-out (DSE-12064)

After a web session times out, the terminal should stop running, but it remaings functional.

Cloudera AI Workbench upgrades disabled with NTP

Upgrades are disabled for Cloudera AI Workbench configured with non-transparent proxy (NTP). This issue is anticipated to be fixed in a subsequent hotfix release.

Using dollar character in environment variables in Cloudera AI

Environment variables with the dollar ($) character are not parsed correctly by Cloudera AI. For example, if you set PASSWORD="pass$123" in the project environment variables, and then try to read it using the echo command, the following output will be displayed: pass23

Workaround: Use one of the following commands to print the $ sign:

```
echo 24 | xxd -r -p
or
echo JAo= | base64 -d
```

Insert the value of the environment variable by wrapping it in the command substitution using $() or ``. For example, if you want to set the environment variable to ABC$123, specify:

```
ABC$(echo 24 | xxd -r -p)123
or
ABC`echo 24 | xxd -r -p`123
```

Models: Some API endpoints not fully supported

In the create_run api, run_name is not supported.

Also, search_experiments only supports pagination.

When a team added as a collaborator, it does not appear in the UI. (DSE-31570)

Run Job as displays even if the job is enabled on a service account. (DSE-31573)

If the job is enabled on a service account, the Run Job as option should not display. Even if me is selected at this point, the job still runs in the service account.

AMP archive upload fails if Project does not contain metadata YAML file

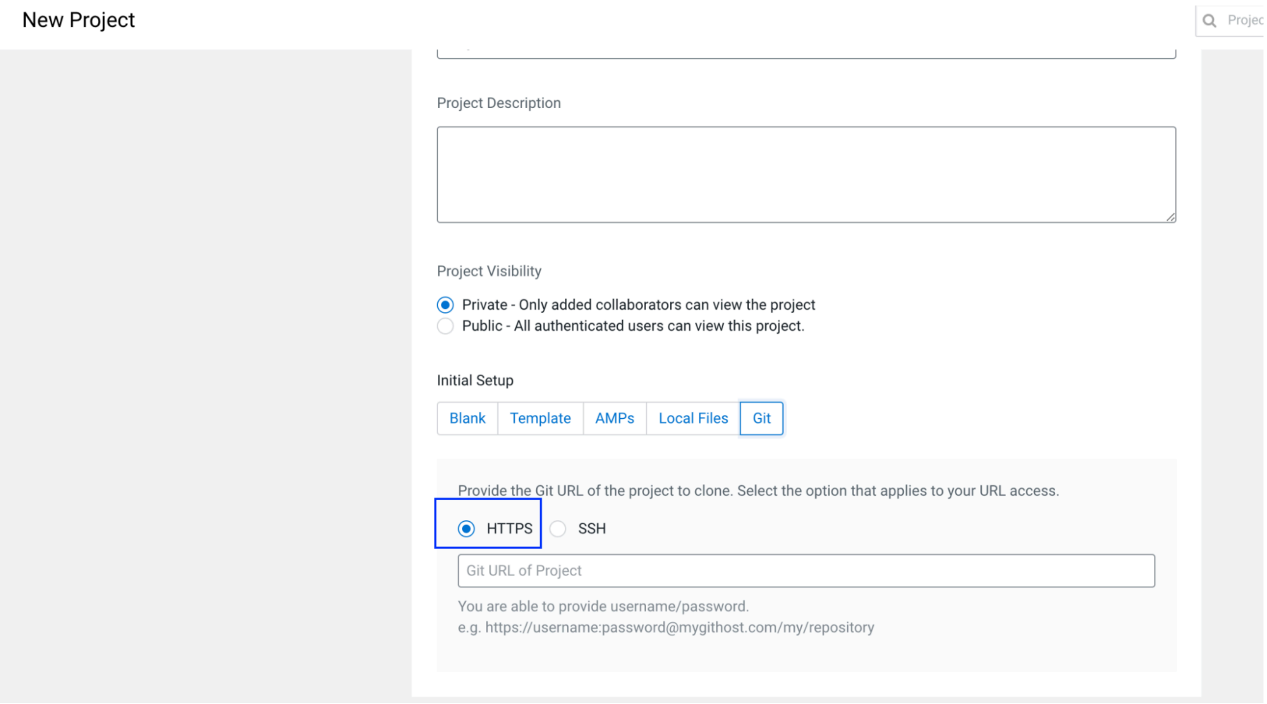
To create a zip file to deploy AMPs in Cloudera AI, do the following:

1. Download the AMP zip file from GitHub
2. Unzip it to a temp directory
3. From the command line navigate to the root directory of the zip
4. Run this command to create the new zip file: zip -r amp.zip .

Make sure you see the .project-metadata/yaml in the root of the zip file.

Cloning from Git using SSH is not supported via HTTP proxy

Workaround: Cloudera AI Projects support HTTPS for cloning git projects. It is suggested to use this as the workaround.



New Project

Project Description

Project Visibility

● Private - Only added collaborators can view the project
○ Public - All authenticated users can view this project.

Initial Setup

Blank | Template | AMPs | Local Files | [✓ Git]

Provide the Git URL of the project to clone. Select the option that applies to your URL access.

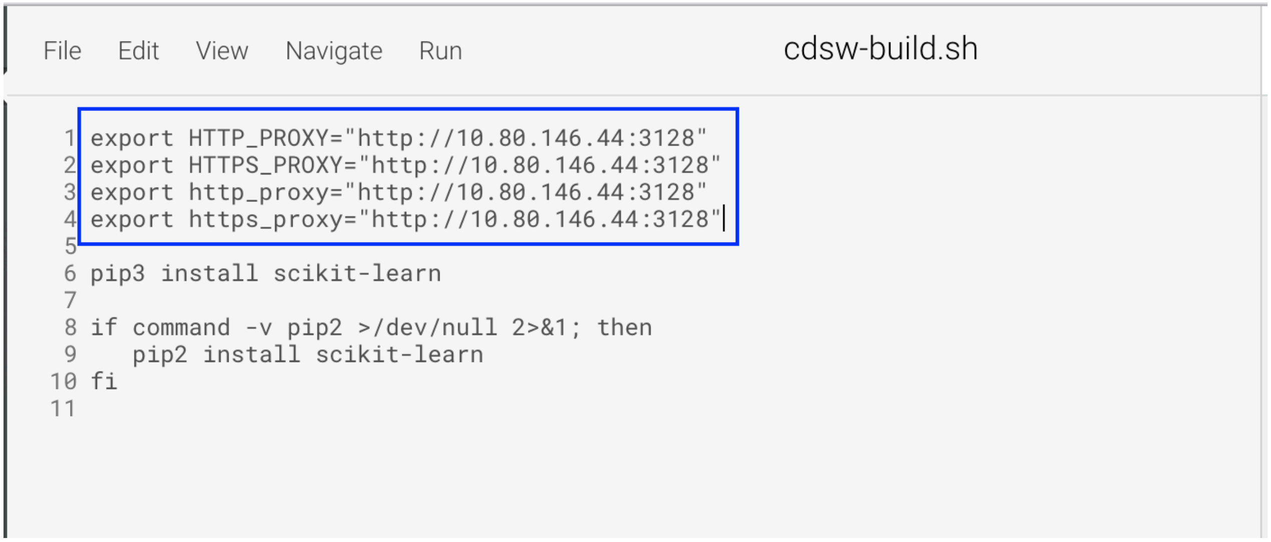
● HTTPS ○ SSH

Git URL of Project

You are able to provide username/password.
e.g. https://username:password@mygithost.com/my/repository

Model deployments requiring outbound access via proxy do not honor HTTP_PROXY, HTTPS_PROXY environment variables

Workaround: Add the `HTTP_PROXY`, `HTTPS_PROXY`, `http_proxy` and `https_proxy` environment variables to the cdsw-build.sh file of the Project Repository.



```
export HTTP_PROXY="http://10.80.146.44:3128"
export HTTPS_PROXY="http://10.80.146.44:3128"
export http_proxy="http://10.80.146.44:3128"
export https_proxy="http://10.80.146.44:3128"

pip3 install scikit-learn

if command -v pip2 >/dev/null 2>&1; then
   pip2 install scikit-learn
fi

```

Application does not restart after upgrade or migration

An application may fail to automatically restart after a workbench upgrade or migration. In this case, manually restart the application.

Do not use backtick characters in environment variable names

Avoid using backtick characters (`) in environment variable names, as this will cause sessions to fail with exit code 2.

Cloudera AI Registry is not supported on R models

Cloudera AI Registry is not supported on R models.

The mlflow.log_model registered model files might not be available on NFS Server (DSE-27709)

When using mlflow.log_model, registered model files might not be available on the NFS server due to NFS server settings or network connections. This could cause the model to remain in the registering status.

Workaround:

- Re-register the model. It will register as an additional version, but it should correct the problem.
- Add the ARTIFACT_SYNC_PERIOD environment variable to hdfscli-server Kubernetes deployment and set it to an integer value. This will set the Cloudera AI Registry retry operation to twice the number of seconds specified by the artifact sync period integer value. If the ARTIFACT_SYNC_PERIOD is set to 30 seconds then Cloudera AI Registry will retry for 60 seconds. The default value is 10 and Cloudera AI Registry retries for 20 seconds. For example: -name: ARTIFACT_SYNC_PERIOD value: “30”.

Applications appear in failed state after upgrade (DSE-23330)

After upgrading Cloudera AI from version 1.29.0 on AWS, some applications may be in a Failed state. The workaround is to restart the application.

Cannot use hashtag character in JDBC connection string

The special character # (hashtag) cannot be used in a password that is then used in a JDBC connection string. Avoid using this special character, or use '%23' instead.

Cloudera AI Workbench installation fails

Cloudera AI Workbench installation with Azure NetApp Files on NFS v4.1 fails. The workaround is to use NFS v3.

Spark executors fail due to insufficient disk space

Generally, the administrator should estimate the shuffle data set size before provisioning the workbench, and then specify the root volume size of the compute node that is appropriate given that estimate. For more specific guidelines, see the following resources.

- How do I avoid the "No space left on device" error where my disk is running out of space?
- How can I prevent a Hadoop or Spark job's user cache from consuming too much disk space in Amazon EMR?

Runtime Addon fails to load (DSE-16200)

A Spark runtime add-on may fail when upgrading a workbench.

Solution: To resolve this problem, try to reload the add-on. In Site Administration Runtime/Engine , in the option menu next to the failed add-on, select Reload.

Cloudera AI Workbench provisioning times out

When provisioning a Cloudera AI Workbench, the process may time out with an error similar to Warning FailedMount or Failed to sync secret cache:timed out waiting for the condition. This can happen on AWS or Azure.

Solution: Delete the workbench and retry provisioning.

Cloudera AI endpoint connectivity from Cloudera Data Hub and Cloudera Data Engineering (DSE-14882)

When Cloudera services connect to Cloudera AI services, if the Cloudera AI Workbench is provisioned on a public subnet, traffic is routed out of the VPC first, and then routed back in. On Cloudera on premises Cloudera AI, traffic is not routed externally.

NFS performance issues on AWS EFS (DSE-12404)

Cloudera AI uses NFS as the filesystem for storing application and user data. NFS performance may be much slower than expected in situations where a data scientist writes a very large number (typically in the thousands) of small files. Example tasks include: using `git clone` to clone a very large source repository (such as TensorFlow), or using `pip` to install a Python package that includes JavaScript code (such as `plotly`). Reduced performance is particularly common with Cloudera AI on AWS (which uses EFS), but it may be seen in other environments.

Disable file upload and download (DSE-12065)

You cannot disable file upload and download when using the Jupyter Notebook.

Remove Workbench operation fails (DSE-8834)

Remove Workbench operation fails if workbench creation is still in progress.

API does not enforce a maximum number of nodes for Cloudera AI Workbench

When the API is used to provision new Cloudera AI Workbench, it does not enforce an upper limit on the autoscale range.

Downscaling Cloudera AI Workbench nodes does not work as expected (MLX-637, MLX-638)

Downscaling nodes does not work as seamlessly as expected due to a lack of Bin Packing on the Spark default scheduler, and because dynamic allocation is not currently enabled. As a result, currently infrastructure pods, Spark driver/executor pods, and session pods are tagged as non-evictable using the cluster-autoscaler.kubernetes.io/safe-to-evict: "false" annotation.

First time user synchronization adds ML user as the default role to the already synced users (DSE-42775)

When you perform user synchronization, by default, the newly added users are designated as ML User. This behavior applies to all users who are newly added and synced to CML from CML2.0.47-b359 or in CML2.0.47-b360 releases.

Workaround: You must perform user synchronization once again to display the actual or designated roles.

Disable auto synchronization feature for users and teams (DSE-36718)

The automated team and user synchronization feature is disabled. Newly installed or upgraded workbenches do not have the automatic synchronization option in the Cloudera AI UI.

Cloudera AI workload sessions will not be created using password-protected SSH key (DSE-42698)

Impacted users will not be able to start workloads or clone private github accounts. Cloudera recommends upgrading the Workbench to version the CML2.0.47-b360 on priority.

This issue impacts following set of users:

- Newly added users synced to a freshly created Cloudera AI Workbench version CML2.0.47-b359.
- Users added and synced to Cloudera AI Workbench version CML2.0.47-b359, later upgraded to version CML2.0.47-b360 (using in-place or backup-restore upgrades).
- Users added after upgrading Cloudera AI Workbench to version CML2.0.47-b359 from an earlier version.

Workaround: Involves addressing these issues in the following ways:

1. Users can rotate their own SSH keys individually by logging into Cloudera AI Workbench UI > select User Settings # select the Outbound SSH tab # click Reset SSH Key button.
2. If the Workbench is upgraded to CML 2.0.47-b360 from CML 2.0.47-b359, run the mitigation.sh script only once.

Note: Cloudera recommends upgrading the Workbench to CML 2.0.47-b360.

While using CML 2.0.47-b359 version, the administrator has to run the mitigation.sh script after synchronizing each user or team operation.

You must run the mitigation.sh script providing the namespace, kubeconfig, and db-reset-ssh-pass-keys.sh script as arguments.

Run the scripts are as follows:

1. On your local system, copy and paste both of the code snippets to mitigation.sh and db-reset-ssh-pass-keys.sh respectively.

   Copy the following code to db-reset-ssh-pass-keys.sh script:

```
#!/bin/bash
# This script must be run inside of web pod in mlx namespace by getting a
 k8s exec shell into that pod
# Detects if a private key is inappropriately password protected and rot
ates it for a new one
# This should be executed as often as user sync is performed to fix ssh ke
ys for any newly synced users

export PGPASSWORD=$POSTGRESQL_PASS
query="psql -h db.mlx.svc.cluster.local -U $POSTGRESQL_USER -qAtX -c"
USERS_ID_LIST=$($query "select id from users;")
touch /tmp/existing-priv-key
chmod 0700 /tmp/existing-priv-key

while IFS= read -r user_id
do
```

```
    $query "select private_key from public.ssh_keys where id=$user_id;" >
/tmp/existing-priv-key
    ssh-keygen -y -P "" -f /tmp/existing-priv-key &> /dev/null
    if [ $? -ne 0 ]; then
        echo "Detected passphrase protected SSH key for User $user_id"
        rm -f /tmp/new-priv-key
        rm -f /tmp/new-priv-key.pub
        ssh-keygen -f /tmp/new-priv-key -b 2048 -C cdsw -q -N ""
        priv_key=`cat /tmp/new-priv-key`
        pub_key=`cat /tmp/new-priv-key.pub`
        $query "update public.ssh_keys set private_key='$priv_key', pub
lic_key='$pub_key' where id=$user_id"
        echo "Rotated SSH key for User $user_id"
    fi
done <<< $USERS_ID_LIST

rm -f /tmp/new-priv-key
rm -f /tmp/new-priv-key.pub
rm -f /tmp/existing-priv-key
```

Copy the following code to mitigation.sh script:

```
#!/bin/bash
# ### Usage: ./mitigation.sh <namespace> <kubeconfig> <file_to_copy_and
_run>
# ex: ./mitigation.sh mlx /root/configs/ums.conf db-reset-only-pass-ssh.sh

# Input arguments
NAMESPACE=$1
KUBECONFIG=$2
FILE_TO_COPY_AND_RUN=$3

# Check if required arguments are provided
if [ -z "$NAMESPACE" ] || [ -z "$KUBECONFIG" ] || [ -z "$FILE_TO_COPY_AND
_RUN" ]; then
  echo "# Error: Missing required arguments."
  echo "Usage: $0 <namespace> <kubeconfig> <file_to_copy_and_run>"
  exit 1
fi

# Ensure kubectl is using the provided kubeconfig
export KUBECONFIG=$KUBECONFIG
# Get the name of the pod under the 'web' deployment in the specified n
amespace
POD_NAME=$(kubectl get pods -n $NAMESPACE -l app=web -o jsonpath='{.ite
ms[0].metadata.name}' --kubeconfig=$KUBECONFIG)
if [ -z "$POD_NAME" ]; then
  echo "# Error: No pod found for the 'web' deployment in the namespace
 $NAMESPACE."
  exit 1
fi

echo "# Found pod: $POD_NAME"

# Copy the file to the pod
kubectl cp $FILE_TO_COPY_AND_RUN $NAMESPACE/$POD_NAME:/tmp/$(basename $
FILE_TO_COPY_AND_RUN) --kubeconfig=$KUBECONFIG

# Check if the file copy was successful
if [ $? -ne 0 ]; then
  echo "# Error: Failed to copy file $FILE_TO_COPY_AND_RUN to pod $POD_NAM
E."
  exit 1
```

```
fi
echo "# File copied to pod successfully."

# Run the script inside the pod
kubectl exec -n $NAMESPACE $POD_NAME -- /bin/bash /tmp/$(basename $FILE_
TO_COPY_AND_RUN) --kubeconfig=$KUBECONFIG

# Check if the script execution was successful
if [ $? -ne 0 ]; then
  echo "# Error: Failed to execute the script inside the pod."
  exit 1
fi

echo "# Script executed successfully inside the pod."
```

2. Download the Cloudera AI cluster's kubeconfig. For more details see Granting remote access to Cloudera AI Workbenches.
3. Run the ./mitigation.sh script as follows:

   ./mitigation.sh <cml-namespace> <kubeconfig-path>   db-reset-ssh-pass-keys.sh

Cloudera AI Inference service Known issues

- DSE-48604: When the Access Control feature is enabled on the Cloudera AI Inference service Details page, you can encounter two related issues:

  - The list of currently deployed endpoints is not visible.
  - The UI slider/toggle does not immediately refresh its state after enabling or disabling the feature.

  Workaround: To successfully view the list of deployed endpoints, the Access Control feature must be disabled. If the UI toggle does not reflect the change immediately after adjusting the setting, manually reload the page to see the current status.
- DSE-48399: Clicking the Swagger UI link on the Cloudera AI Inference service Details page incorrectly results in an HTTP ERROR 401   Unauthorized error instead of displaying the documentation. This is caused by a missing trailing slash in the URL generated by the link.

  Workaround: After clicking the link and seeing the error, manually append a trailing slash (/) to the end of the URL in your browser's address bar and reload the page. The Swagger UI should then load correctly.
- DSE-48145: The Boltz2 NIM model server does not publish Prometheus metrics starting in Cloudera AI version 2.0.52-b60. Consequently, the charts in the Metrics tab on the Model Endpoint Details page will not display any data for Boltz2 models.
- The following compute instance types are not supported by Cloudera AI Inference service:

  - Azure: NVadsA10_v5 series.
  - AWS: p4d.24xlarge
- DSE-39826: Running the modify-ml-serving-app command on Cloudera AI Inference service Azure clusters in the us-west-2 workload region fails. When this occurs, the status of the application is incorrectly displayed as modify:failed.

  Workaround: You must first delete the instance group you want to modify using the `delete-instance-group-ml-serving-app` API. Then, recreate the instance group, modify the configuration based on your requirements, and add the instance group using the `add-instance-groups-ml-serving-app` API.
- Unclean deletion of Cloudera AI Inference service version 1.2.0 and older. If you delete Cloudera AI Inference service version 1.2.0 or older, some Kubernetes resources are left in the cluster and causes a subsequent creation of Cloudera AI Inference service on the same cluster to fail. It is recommended that you delete the Compute cluster and recreate it to deploy Cloudera AI Inference service on it.

- Graceful deletion of Cloudera AI Inference service version older than 1.3.0-b111 fails. A new feature introduced in version 1.3.0-b111 has caused a regression where graceful deletion of an existing Cloudera AI Inference service version 1.2.0 fails.

  Workaround: Use the CDP CLI version 0.9.131 or higher to forcefully delete Cloudera AI Inference service.

  ```
  cdp ml delete-ml-serving-app --app-crn [***APP_CRN***] --force
  ```

  Cloudera recommends that after a forceful deletion of Cloudera AI Inference service, you delete the underlying compute cluster as well to ensure that all resources are cleaned up properly.
- Updating the description after a model has been added to a model endpoint will lead to a UI mismatch in the model builder for models listed by the model builder and the models deployed.
- When you create a model endpoint from the **Create Endpoint** page, even though the instance type selection is not mandatory, the endpoint creation fails if the instance type is not selected.
- DSE-39626: If no worker node can be found within 60 minutes to schedule a model endpoint that is either newly created or is scaling up from 0, the system will give up trying to create and schedule the replica. A common reason for this behavior is insufficient cloud quota, or capacity constraints on the cloud service provider’s side. You could either ask for increased quota, or try to use an instance type that is more readily available.
- When updating an Model Endpoint with a specific GPU requirement, the instance type must be explicitly set again even if there is no change.
  - To bring up the endpoint after a revision is failed, the endpoint configuration needs to be updated. This can be achieved currently in the form of an autoscale range change, or resource requirements change.
- When updating an endpoint with a specific GPU requirement, the instance type must be explicitly set again even if there is no change.
- Embedding models function in two modes: query or passage. This has to be specified when interacting with the models. There are two ways to do this:
  - suffix the model id in the payload by either -query or -passage or
  - specify the input_type parameter in the request payload.

    For more information, see NVIDIA documentation.
- Embedding models only accept strings as input. Token stream input is currently not supported.
- Llama 3.2 Vision models are not supported on AWS on A10G and L40S GPUs.
- Llama 3.1 70B Instruct model L40S profile needs 8 GPUs to deploy successfully, while Nvidia documentation lists this model profile as needing only 4 L40S GPUs.
- Model Runtimes have been changed in a non-backward compatible way between Cloudera AI Inference service version 1.2 and 1.3. Therefore, NIM model endpoints deployed in version 1.2 need to be redeployed by downloading their profiles again through Model Hub and creating a new endpoint from the most recent version of the model in the Cloudera AI Registry.
- You cannot upgrade from Cloudera AI Inference service version 1.3.0-b111 to higher. You must first delete the service and recreate it to deploy version 1.3.0-b113 or higher.
- Hugging Face model deployment fails in Cloudera AI Inference service 1.3.0-b114.
- DSE-42519: Specifying subnets for load balancer from the UI when creating Cloudera AI Inference service does not work. The specified subnets are accepted by the UI, but these settings are not actually applied to the load balancer service created in the cluster.

  Workaround: Use **CDP CLI** to specify subnets for the load balancer.

Limitations

Cloudera AI Inference service and Cloudera AI Registry are not supported on Azure East Asia and Qatar Central regions

Cloudera AI Inference service and Cloudera AI Registry are not supported on Microsoft Azure East Asia and Qatar Central regions due to lack of support for Workload Identity by Microsoft Azure.

Technical Service Bulletins

TSB 2025-844: Garbage collection for pods in Error or Stuck states in Cloudera AI on cloud is not working in Cloudera AI

In certain (older) versions of Cloudera AI on cloud, garbage collection of pods in states such as Error and Init:Unknown inside Cloudera AI Workbenches is not occurring. This can prevent the deployment of new pods and lead to unnecessary cloud costs for stale workload pods no longer serving any purpose.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2025-844: Garbage collection for pods in Error or Stuck states in Cloudera AI on cloud is not working .

TSB 2025-826: Non-authorized users can perform CRUD operations in Cloudera AI

Non-authorized but authenticated users can carry out CRUD operations on Cloudera AI Registry metadata tables. This includes the ability to update/delete existing metadata of models, model-versions, and tags tables. This can allow them to update permissions in the metadata to in turn access existing model artifacts that are stored in S3 or Azure Blob Storage.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2025-826: Non-authorized users can perform CRUD operations in Cloudera AI

TSB 2024-761: Orphan EBS Volumes in Cloudera AI Workbench

Cloudera AI provisions Elastic Block Store (EBS) volumes during provisioning of a workbench. Due to missing labels on Cloudera AI Workbench, delete operations on previously restored Cloudera AI Workbench didn’t clean up a subset of the provisioned block volumes.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB 2024-761: Orphan EBS Volumes in Cloudera AI Workbench

TSB 2023-628: Sensitive user data getting collected in Cloudera AI Workbench or CDSW workbench diagnostic bundles

When using Cloudera Data Science Workbench (CDSW), Cloudera recommends users to store sensitive information, such as passwords or access keys, in environment variables rather than in the code. See Engine Environment Variables in the official Cloudera documentation for details. Cloudera recently learned that all session environment variables in the affected releases of CDSW and Cloudera AI are logged in web pod logs, which may be included in support diagnostic bundles sent to Cloudera as part of support tickets.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB-2023-628: Sensitive user data getting collected in Cloudera AI Workbench or in CDSW workbench diagnostic bundles

TSB 2022-588: Kubeconfig and new version of aws-iam-authenticator

Regenerate Kubeconfig and in conjunction use a newer version of aws-iam-authenticator on AWS. Kubeconfig in Cloudera Cloudera on cloud Data Services needs to be regenerated because the Kubeconfig generated before June 15, 2022 uses an old APIVersion (client.authentication.k8s.io/v1alpha1) which is no longer supported. This causes compatibility issues with aws-iam-authenticator starting from v0.5.7. To be able to use the new aws-iam-authenticator, the Kubeconfig needs to be regenerated.

Knowledge article

For the latest update on this issue see the corresponding Knowledge article: TSB-2022-588: Kubeconfig and new version of aws-iam-authenticator

Related Information

AWS Limitations

Azure Limitations