Cloudera AI

Release Notes

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

https://docs.cloudera.com/

Legal Notice

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

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 in Cloudera AI on premises 1.5.5...4

Known issues for Cloudera AI on premises 1.5.5.. 5

Fixed issues in Cloudera AI on premises 1.5.5...10

What's new in Cloudera AI on premises 1.5.5

Understand the functionalities and improvements in Cloudera AIon premises 1.5.5

New features and improvements

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 AI Studios Overview.

Cloudera AI Inference service [Technical Preview]

The Cloudera AI Inference service is a data service available in Technical Preview. 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.

Cloudera Copilot [Technical Preview]

Cloudera Copilot is an AI-powered coding assistant designed for seamless integration within JupyterLab ML Runtimes. With its chat interface and comprehensive code completion features, Cloudera Copilot enhances the development experience for machine learning projects. It offers compatibility with model endpoints deployed in Cloudera AI Inference service model, providing developers with flexibility and efficiency in their workflows.

For details, see Cloudera Copilot Overview.

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. 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

Certification Manager

Cert-manager is an open-source tool for Kubernetes that automates the provisioning, management, and renewal of TLS certificates. Its documentation at https://cert-manager.io/docs/ provides comprehensive guidance on installing, configuring, and using cert-manager to secure workloads with trusted X.509 certificates. Cloudera provides out-of-the-box support for Venafi TPP as part of the Cloudera Embedded Container Service installation. By integrating cert-manager, the Cloudera Data Services on premises achieve secure communication, reduced manual overhead, and compliance with security standards, leveraging its robust automation and flexibility. For more information on setting up Cert-manager using Venafi TPP, see Certification Manager service for increased security.

Workbench-level Spark defaults

Custom Spark settings can now be configured at 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 .

Spark pushdown

The Administrator can set Spark pushdown to be enabled during project creation by default.

Auto-synchronization for user and team is enabled by default

The efficiency and usability of the auto-synchronization features have been enhanced for user and team management. Key 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.
- You can switch on or off the User and Team auto synchronization feature.

These improvements are aimed at optimizing performance and streamlining the synchronization process for users and teams.

Multiple docker registry accounts

Cloudera AI now supports storing multiple Docker credentials for your custom runtimes and provides a dedicated UI and API for managing them. Additionally, Cloudera AI no longer retrieves Docker registry credentials from the regcred secret.

If you previously relied on credentials stored in the regcred secret, you must add these credentials to Cloudera AI under Site Administration Runtime to ensure your workloads continue functioning seamlessly.

UI displays skipped job runs with skipped status tag

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.

Related Information

Known Issues and Limitations

Certification Manager service for increased security

Managing Users

Known issues for Cloudera AI on premises 1.5.5

This section lists known issues that you might run into while using Cloudera AI on premises.

DSE-44699: Provisioning workbench is failing with error pool guaranteed resources larger than parent's available guaranteed resources

With the Quota Management feature enabled, creating Cloudera AI Workbench might fail with the error pool guaranteed resources larger than parent's available guaranteed resources.

DSE-44367: Buildkitd Pod CrashLoopBackOff due to port conflicts

During the creation or upgrade of a Cloudera AI Workbench, buildkitd pods may occasionally enter a CrashLoopBackOff state. This typically happens when the port used by BuildKit is not properly released during pod restarts or is occupied by another process. You may encounter errors such as:

```
buildkitd: listen tcp 0.0.0.0:1234: bind: address already in use
```

Workaround:

If you experience this issue, follow this step to resolve it:

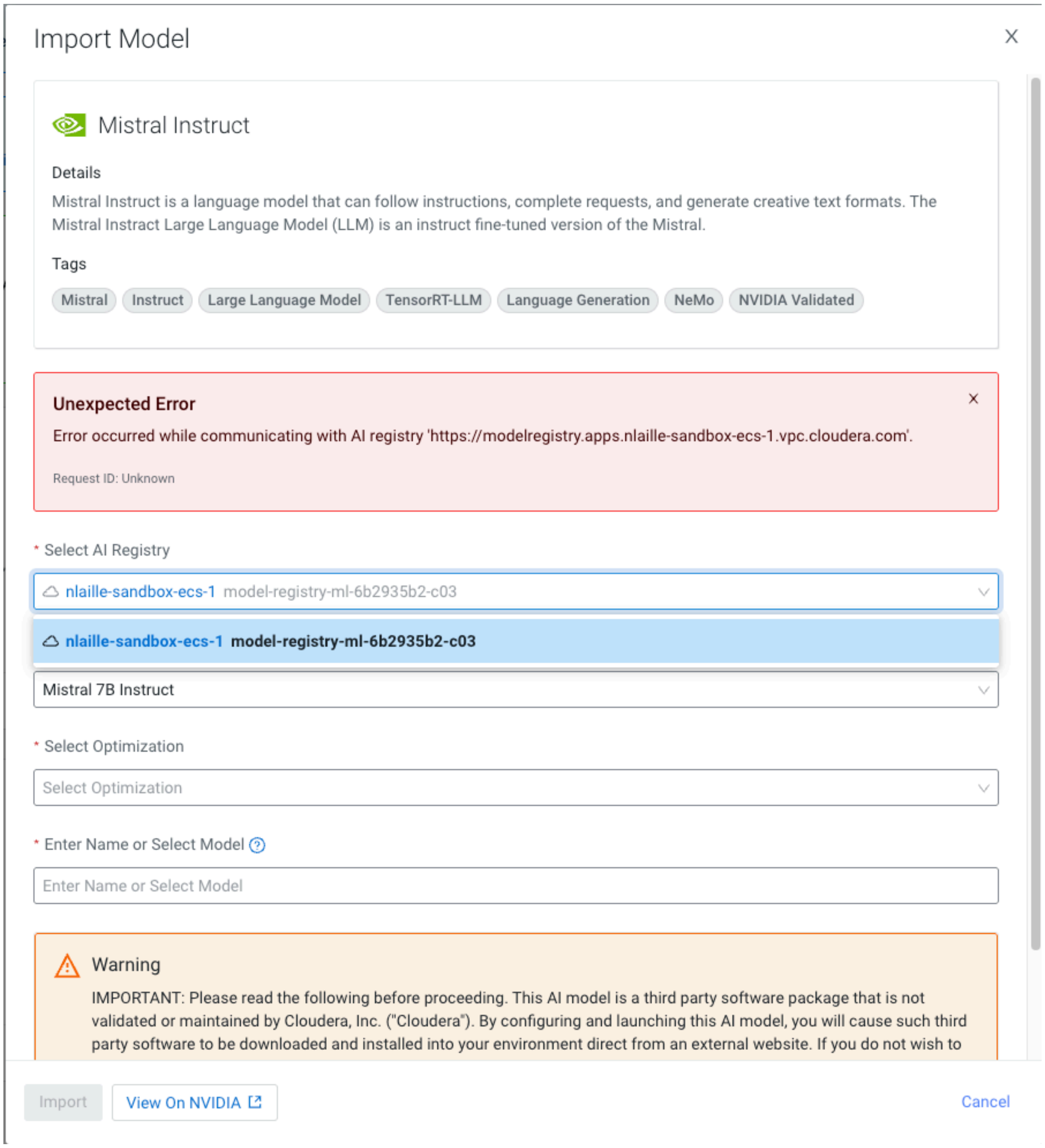
- Perform a rollout restart of the buildkitd pods to ensure they start correctly:

  ```
  kubectl rollout restart daemonset buildkitd -n [***WORKBENCH
  NAMESPACE***]
  ```

DSE-44319 The model import from Model Hub is failing with error to communicate with Cloudera AI Registry

If you have used self-signed certificates or your trust store lacks the certificate required to trust the Cloudera AI Registry domain, the Cloudera AI Registry calls from the UI might fail.

Figure 1: Importing a model fails



Import Model

Mistral Instruct

Details

Mistral Instruct is a language model that can follow instructions, complete requests, and generate creative text formats. The Mistral Instract Large Language Model (LLM) is an instruct fine-tuned version of the Mistral.

Tags

Mistral Instruct Large Language Model TensorRT-LLM Language Generation NeMo NVIDIA Validated

Unexpected Error

Error occurred while communicating with AI registry 'https://modelregistry.apps.nlaille-sandbox-ecs-1.vpc.cloudera.com'.

Request ID: Unknown

* Select AI Registry

nlaille-sandbox-ecs-1 model-registry-ml-6b2935b2-c03

nlaille-sandbox-ecs-1 **model-registry-ml-6b2935b2-c03**

Mistral 7B Instruct

* Select Optimization

Select Optimization

* Enter Name or Select Model

Enter Name or Select Model

Warning

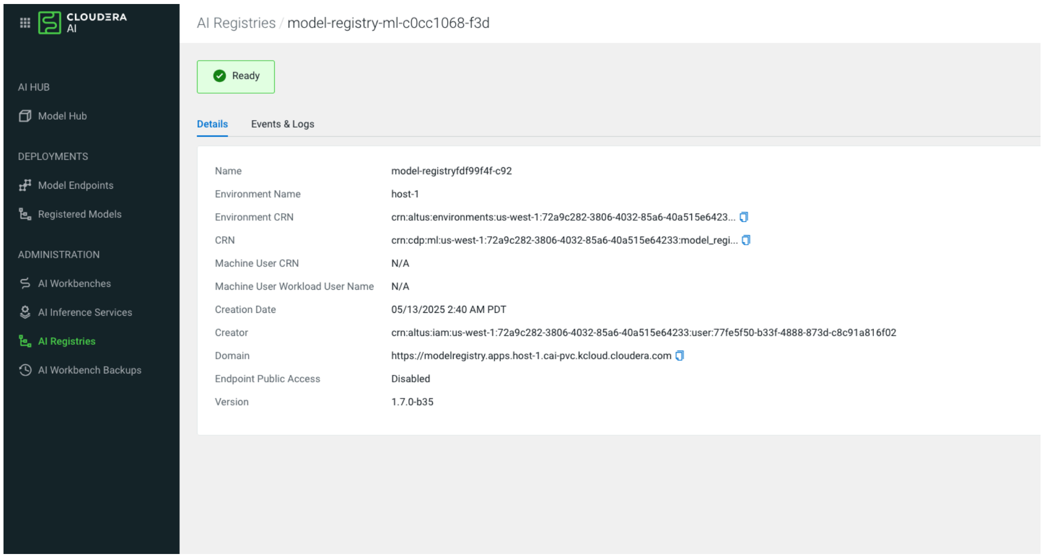
IMPORTANT: Please read the following before proceeding. This AI model is a third party software package that is not validated or maintained by Cloudera, Inc. ("Cloudera"). By configuring and launching this AI model, you will cause such third party software to be downloaded and installed into your environment direct from an external website. If you do not wish to

Workaround:

When facing issues with an untrusted certificate:

- Select Cloudera AI Registry from the left navigation pane. The Cloudera AI Registry page is displayed.
- Open up the Cloudera AI Registry **Details** page.

- Copy the domain and open it in a new browser.



Name	model-registryfdf99f4f-c92
Environment Name	host-1
Environment CRN	crn:altus:environments:us-west-1:72a9c282-3806-4032-85a6-40a515e6423...
CRN	crn:cdp:ml:us-west-1:72a9c282-3806-4032-85a6-40a515e64233:model_regi...
Machine User CRN	N/A
Machine User Workload User Name	N/A
Creation Date	05/13/2025 2:40 AM PDT
Creator	crn:altus:iam:us-west-1:72a9c282-3806-4032-85a6-40a515e64233:user:77fe5f50-b33f-4888-873d-c8c91a816f02
Domain	https://modelregistry.apps.host-1.cai-pvc.kcloud.cloudera.com
Endpoint Public Access	Disabled
Version	1.7.0-b35

- Add the certificate permanently to your device's trust store to avoid the risk with the current session.

For permanent trust, export the certificate from your browser and import it into your operating system's certificate manager.

DSE-43704: Rename custom tee binary to cml-tee

Certain vulnerability scanners may incorrectly flag Cloudera AI as using a vulnerable version of coreutils and tee.

Cloudera AI services include a custom tee binary, developed entirely in-house by Cloudera, which is not based on the open-source coreutils library. The current version of Cloudera's custom tee command is 0.9, which may be mistakenly identified as the tee command from coreutils that contains known vulnerabilities.

DSE-44827: The model is failing with unknown status

Model deployments may fail to start if the model build relies on an add-on that was hotfixed in the release. This issue occurs when the model deployment restarts and the add-on hotfixing process overlap, leading to conflicts.

Workaround:

To resolve this issue, create a new build and deploy it for the affected model deployments.

DSE-44682: Model deployment is failing at building stage due to TLS issues

TLS-related issues may occasionally occur during the Cloudera AI model build process in Cloudera Embedded Container Service clusters, specifically when pulling images from the container registry. These issues are typically caused by missing registry certificates on the worker nodes, which should be located at the following path: /etc/docker/certs.d/.

Workaround:

To address this issue, ensure that the required registry certificates are present on all worker nodes. Follow these steps to recover:

1. **Identify a reference worker node.**

   Select a worker node where the model build process completes successfully without TLS errors. This node is expected to have the correct registry certificates in place. In general, any worker node running s2i-builder pods is likely to have the necessary certificates on the nodes of the cluster.

2. **Locate the registry certificates.**

   On the reference worker node, navigate to /etc/docker/certs.d/[***REGISTRY NAME***]/ and verify that the registry.crt file exists.
3. **Distribute certificates to affected nodes.**

   Copy the certificate files (registry.crt) from the reference node to the same path (/etc/docker/certs.d/[***REGISTRY NAME***]/) on the affected worker nodes that lack the required certificates. Make sure that the certificates for both the image pull and push registries are present and are correctly placed on all worker nodes.
4. Perform rollout restart of the buildkitd daemonset so that the certificates are applied properly to the buildkit pods.

DSE-44913: Spark executor fluentd init container fails

A bug introduced in a recent change to handle dynamic volume association with Spark pods causes the Fluentbit executor, responsible for log collection, to crash. As a result, logs from affected pods are not included in debug bundles, and the Spark executor pod logs will not appear in the session's **Logs** tab in the UI.

Despite this, the Spark executors will continue to function as expected—the engine container will still start, and the script will execute.

Cloudera AI Inference service Known issues

DSE-44238: Cannot create Cloudera AI Inference service application deployment via CDP CLI when ozone credentials are passed

Cloudera AI Inference service cannot be created via CDP CLI. Create the Cloudera AI Inference service only via the UI.

DSE-44141: Failed to delete deployment in executing DeleteMLServingApp

Cloudera AI Inference service fails to remove all namespaces if the Cloudera AI Inference service is deleted post installation failure.

Workaround:

Manually delete the below namespaces from the cluster:

- knative-serving
- kserve
- cml-serving
- knox-caii
- serving-default

Note:

If you encounter errors related to missing namespaces, you can safely ignore them. These namespaces may have already been deleted as part of the Cloudera AI Inference service removal process.

Further known issues with Cloudera AI Inference service:

- 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.
- 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.
- Mistral 7B models for NIM version 1.1.2 require the max_tokens parameter in the request payload. This API regression is known to affect the Test Model UI functionality for this specific NIM version.
- NIM endpoints will reply with a 307 temporary redirect if the URL ends with a trailing /. Make sure not to have a trailing slash character at the end of NIM endpoint URLs.

Fixed issues in Cloudera AI on premises 1.5.5

This section lists the issues that have been fixed since the last release of Cloudera AI on premises.

Cloudera AI Workbench

DSE-38715: Improve the Last Modified column for Files in the UI

This update standardises the use of absolute timestamps across the entire application, replacing relative timestamps such as 1 week ago with precise date and time information. The **Files** section in the Project Overview page now displays exact Last Modified dates, while the **Jobs** section in the **Job Overview** page shows accurate Latest Run times for improved clarity and consistency.

DSE-41733 - Spark logs are not cleaned up by livelog cleaner

The Livelog cleaner previously cleaned up input and output log topics but did not delete the content of container log topics. It now includes engine containers and Spark executor log topics in the cleanup process.

DSE-42499: ZSTD compression codec support in Cloudera AI Spark while reading Hive tables

Previously, Spark workloads were unable to read the ZSTD-encoded Hive tables when running in the default Spark on Kubernetes mode.

This issue has now been resolved, enabling Spark to successfully read ZSTD-encoded data from Hive tables.

DSE-42183: Shared Memory Not Set on Applications

The Shared Memory configuration set in Project Settings Advanced is now applied to applications in addition to sessions.

This configuration is retrieved from the Projects table when a user creates a new application.

DSE-41429 The timezone for duration for experiment details is not correct

The **Experiment** page on the Details tab occasionally displayed a negative value for the duration of time, due to an incorrect timezone configuration in certain setups. This issue has been solved.

DSE-42858 - [Cloudera Embedded Container Service Restart Stability] db-0 Pod CrashLoops with error after Cloudera Embedded Container Service restart

The following error was intermittently observed in the db-0 pod during restarts or upgrades:

- CDSW_PG FATAL: lock file "postmaster.pid" already exists

-
  ```
  CDSW_PG HINT: Is another postmaster (PID 7) running in data
          directory
  ```

This issue occurred because the db-migrate-xxx job was still in the Running state and was attempting to access the embedded database.

The issue has now been resolved. The db-0 pod no longer enters a crash loop due to a stale postmast er.pid file.

ML Runtimes

DSE-41772: Opening session opens up a wrong file intermittently

This update addressed an issue where the Workbench editor occasionally opened the wrong file. The problem stemmed from timing conflicts in DocumentManager.js, where a rapid refresh could retrieve an outdated file from local storage instead of the file that was just clicked.

The fix ensures that the DocumentManager.js consistently selects the most recently clicked file during the refresh process, even in scenarios involving quick refreshes.

DSE-40874: Workbench brackets editor search retains the same string until there is a click on the editor

There was an issue in the Workbench editor where the Find input box (Ctrl+F or Command+F) retained the previous search string upon reopening. The input box now clears automatically when clicking anywhere in the editor text area, enabling you to enter a new search string seamlessly.

DSE-42595: Sharing generated images does not work with PBJ Runtimes

Earlier, 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-42962: PBJ log is not truncated (default value and MAX_TEXT_LENGTH value)

PBJ Workbench Runtime images now comply with the value set in the environmental variable MAX_TEXT_LENGTH. This limits the maximum number of characters that can be displayed by each command executed.

DSE-42344: The Interrupt button is not working with PBJ Workbench Runtime

Earlier, the Interrupt button did not work in Cloudera AI sessions that ran a PBJ Workbench Runtime. This issue has been fixed.

DSE-43297: Non-Python custom editor custom Runtimes fail to start

Sessions using PBJ-based custom Runtimes with a custom editor could not start previously. This issue has been fixed.

DSE-42077: PBJ Workbench not pretty-printing outputs

When PBJ Workbench R Runtime was used, the tables and help text were not properly displayed. This issue has been fixed.

DSE-42966: R PBJ Session commands' documentation is not displayed correctly

When PBJ Workbench R Runtime was used, the help text, required as a session command's output, was not properly displayed. This issue has been fixed.

DSE-42967: R PBJ does not have rich visualization for tables

When PBJ Workbench R Runtime was used, the content of the tables was not properly displayed. This issue has been fixed.

DSE-35299: Stop logging every livelog entry in PBJ Runtimes

When PBJ Workbench R Runtime was used, the start of each command execution was logged with INFO severity, and was consequently displayed. Now, the command execution is logged with DEBUG severity and is not displayed.

DSE-42345: Using PBJ variants code completion in Editor is not working

When PBJ Workbench R Runtime was used, the editor was not able to show code completion. This issue has been fixed.

DSE-42498: Workloads are not starting up: /api/kernels: context deadline exceeded

A previously unauthenticated endpoint is now securely authenticated through a reverse proxy, using a mechanism called Browser Accessible Service.

DSE-43192: Jupyter Kernelgateway in PBJ Workloads is exposed publicly without authentication

A previously unauthenticated endpoint is now securely authenticated via a reverse proxy, using a mechanism known as Browser Accessible Service.

DSE-41438: JOB_MAXIMUM_MINUTES does not affect Jobs using PBJ workbench

PBJ Workbench jobs now respect the JOB_MAXIMUM_MINUTES environment variable. PBJ Workbench jobs are automatically terminated once they exceed the time limit specified by JOB_MAXIMUM_MINUTES.

DSE-42473: Jupyter sessions do not follow SESSION_MAXIMUM_MINUTES

When SESSION_MAXIMUM_MINUTES is set, sessions are expected to expire after the specified duration, regardless of activity (default is 7 days). However, unlike PBJ Sessions and Remote Editor Sessions, Jupyter Sessions previously continued running beyond this time and did not time out. This issue has now been resolved.

DSE-42814: Enhanced Job Arguments UX via JOB_ARGUMENTS environment variable for PBJ Workbench ML Runtimes

This update improves the UI on both the Job Creation page and the Job Settings page to clarify that the PBJ-based ML Runtime transforms job arguments into the JOB_ARGUMENTS environment variable. It introduces a dedicated label and a help block, including a link to a code snippet for user guidance.

Openshift container platform

DSE-42509: Project creation using private repository with SSH key and SSH URL is not working on NTP/Airgap Openshift Container Platform

Project creation using a private repository with SSH key and SSH URL did not work on NTP setups. This issue has been fixed.

DSE-42510: Fetching Cloudera or Huggingface AMP catalog is failing on NTP/Airgap Proxy Openshift Container Platform

Previously there has been an error fetching the AMP catalog on NTP/Airgap Proxy Openshift setup when selecting Site Administration AMP tab.

This issue has been solved.