Cloudera Data Science Workbench

Monitoring Cloudera Data Science Workbench Activity

Date published: 2020-02-28
Date modified: 2021-11-30

https://docs.cloudera.com/

Legal Notice

Contents

Monitoring Cloudera Data Science Workbench Usage

There are at least three ways that an Administrator can view the cluster resources and utilization:

- The Site Administration Overview tab. This page displays basic information about your deployment, such as the number of users signed up, the number of teams and projects created, memory used, and some average job scheduling and run times.
- The Projects View Resource Usage Details Workspace Resources widget. This widget will show "Used" and "Available" resources for the entire cluster, as well as the logged in user's resources.

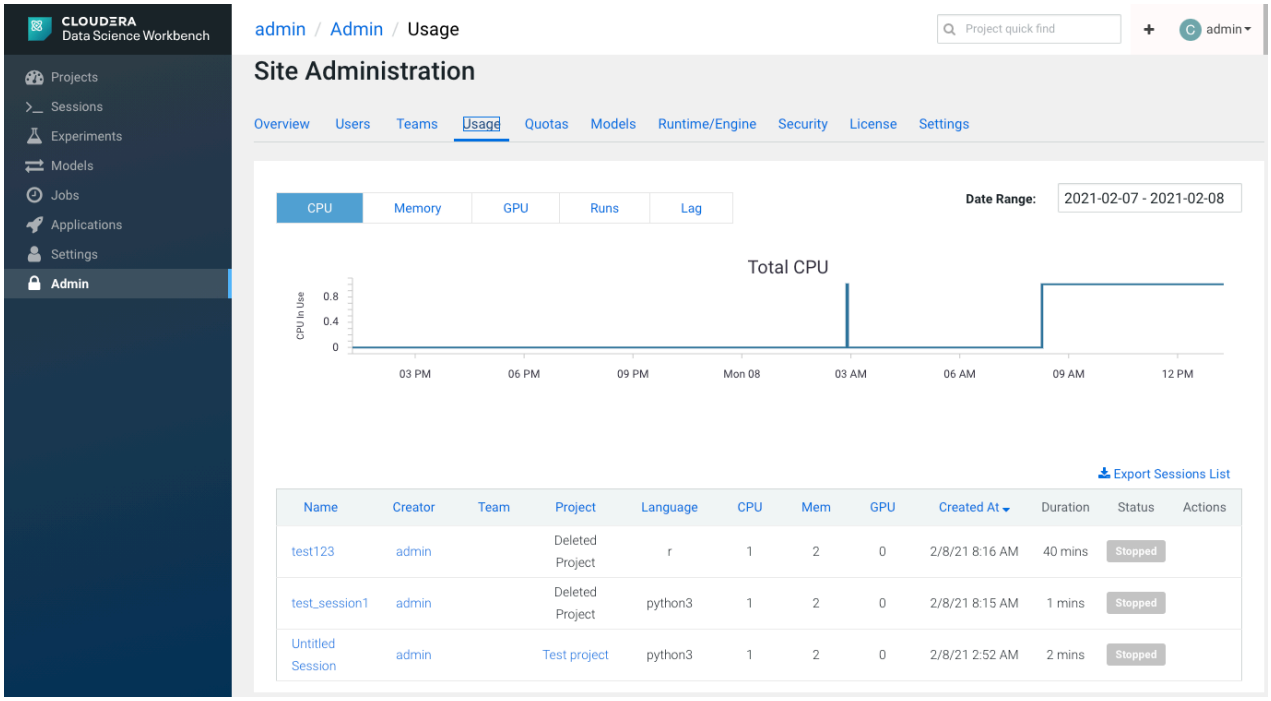
Required Role: Site Administrator

The Admin Users tab displays information about users and their resource use. The tab includes the following columns, among others:

- CPU Time: The amount of time, rounded to one hour increments, that a workspace is utilizing a CPU. If the session uses 100% of CPUs available, and there are two CPUs in the instance, then one hour of use results in two hours of CPU time. Time spent idling also counts towards this metric. This metric is tracked as a 30-day moving average.
- GPU Time: The same as CPU Time, but with GPUs.
- Memory Time: 1 GB of memory allocated to the user's engines per hour. If a 2 GB session is used for one hour, it is counted as two hours of memory time. This metric is also tracked as a 30-day moving average.

The Admin Usage tab of the dashboard displays the following time series charts. These graphs should help site administrators identify basic usage patterns, understand how cluster resources are being utilized over time, and how they are being distributed among teams and users.

Note: Opening the Usage page can cause a brief period of slowness in the rest of the CDSW UI if the CDSW cluster has been in use for a long time. You can avoid or reduce this slow down by purging the cluster's old data. Otherwise, for clusters with many users and projects, use the Usage page sparingly.



Important: The graphs and numbers on the Admin Usage page do not account for any resources used by active models on the deployment. For that information, go to Admin Models page.

- CPU - Total number of CPUs requested by sessions running at this time.

  Note that code running inside an n-CPU session, job, experiment or model replica can access at least n CPUs worth of CPU time. Each user pod can utilize all of its host's CPU resources except the amount requested by other user workloads or Cloudera Data Science Workbench application components. For example, a 1-core Python session can use more than 1 core if other cores have not been requested by other user workloads or CDSW application components.
- Memory - Total memory (in GiB) requested by sessions running at this time.
- GPU - Total number of GPUs requested by sessions running at this time.
- Runs - Total number of sessions and jobs running at this time.
- Lag - Depicts session scheduling and startup times.
  - Scheduling Duration: The amount of time it took for a session pod to be scheduled on the cluster.
  - Starting Duration: The amount of time it took for a session to be ready for user input. This is the amount of time since a pod was scheduled on the cluster until code could be executed.

The Export Sessions List provides a CSV export file of the columns listed in the table. It is important to note that the exported duration column is in seconds for a more detailed output.

The Projects View Resource Usage Details widget shows a bar chart of the available CPU and Memory. This widget is further divided into User Resources and Workspace Resources. This provides a quick "at-a-glance" view of the resources being used by your user account and the entire cluster. If Quotas are used, the quota will be reflected in the Available Resources in the User section.

Note: When viewing the Workspace Resources chart, this will automatically include any internal resources used by the CML/CDSW system itself. This means that if you stop all user workloads, you would still expect to see some System Resources being used. This is expected.

Related Resources

This topic contains related resources for monitoring workbench activity.

- Models - Monitoring All Active Models.
- Tracking Disk Usage - Tracking Disk Usage on the Application Block Device

Monitoring User Events

You can query the PostgresSQL database that is embedded within the Cloudera Data Science Workbench deployment to monitor or audit user events. This requires root access to the Cloudera Data Science Workbench Master host.

Procedure

1. SSH to the Cloudera Data Science Workbench Master host and log in as root.

   ```
   ssh root@<cdsw_master_host_domain_name>
   ```
2. Get the name of the database pod:

   ```
   kubectl get pods -l role=db
   ```

   The command returns information similar to the following example:

   ```
   NAME                  READY   STATUS    RESTARTS   AGE
   db-86bbb69b54-d5q88   1/1     Running   0          4h46m
   ```

3. Enter the following command to log into the database as the sense user:

   ```
   kubectl exec <database pod> -ti -- psql -U sense
   ```

   For example, the following command logs in to the database on pod db-86bbb69b54-d5q88:

   ```
   kubectl exec db-86bbb69b54-d5q88 -ti -- psql -U sense
   ```

   You are logged into the database as the sense user.
4. Run queries against the user_events table.

   For example, run the following query to view the most recent user event:

   ```
   select * from user_events order by created_at DESC LIMIT 1
   ```

   The command returns information similar to the following:

   ```
   id          | 3658
   user_id     | 273
   ipaddr      | ::ffff:127.0.0.1
   user_agent  | node-superagent/2.3.0
   event_name  | model created
   description | {"model":"Simple Model 1559154287-ex5yn","modelId":"50","
   userType":"NORMAL","username":"LucyMilton"}
   created_at  | 2019-05-29 18:24:47.65449
   ```
5. (Optional) Export the user events to a CSV file for further analysis:
   a) While still logged into the database shell, copy the user_events table to a CSV file:

      ```
      copy user_events to '/tmp/user_events.csv' DELIMITER ',' CSV HEADER;
      ```

   b) Exit the PostgeSQL shell. Type \q and press ENTER.
   c) Find the Docker container that the database runs in:

      ```
      docker ps | grep db-86bbb
      ```

      The command returns output similar to the following:

      ```
      8c56d04bbd58 c230b2f564da "docker-entrypoint..." 7 days ago Up 7 days k8
      s_db_db-86bbb69b54-fcfm6_default_8b2dd23d-88b9-11e9-bc34-0245eb679f96_0
      ```

      The first entry in bold is the container ID.
   d) Copy the user_events.csv file out of the container into a temporary directory on the Master host:

      ```
      docker cp <container_ID>:/tmp/user_events.csv /tmp/user_events.csv
      ```

      For example:

      ```
      docker cp 8c56d04bbd58:/tmp/user_events.csv /tmp/user_events.csv
      ```

   e) Use SCP to copy /tmp/user_events.csv from the Cloudera Data Science Workbench Master host to a destination of your choice.

      For example, run the following command on your local machine to copy user_events.csv to a local directory:

      ```
      scp root@<cdsw_master_host_domain_name>:/tmp/user_events.csv /path/to/l
      ocal/directory/
      ```

What to do next

For information about the different user events, see Tracked User Events.

Tracked User Events

The tables on this page describe the user events that are logged by Cloudera Data Science Workbench.

Table 1: Database Columns

When you query the user_events table, the following information can be returned:

Information	Description
id	The ID assigned to the event.
user_id	The UUID of the user who triggered the event.
ipaddr	The IP address of the user or component that triggered the event. 127.0.0.1 indicates an internal component.
user agent	The user agent for this action, such as the web browser. For example: `Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36`
event_name	The event that was logged. The tables on this page list possible events.
description	This field contains the model name and ID, the user type (NORMAL or ADMIN), and the username.
created_at	The date (YYYY-MM-DD format) and time (24-hour clock) the event occurred .

Table 2: Events Related to Engines

Event	Description
engine environment vars updated	-
engine mount created	-
engine mount deleted	-
engine mount updated	-
engine profile created	-
engine profile deleted	-
engine profile updated	-

Table 3: Events Related to Experiments

Event	Description
experiment run created	-Starting with version 1.6.1, this event also captures resource usage (cpu, memory, gpu) per experiment.
experiment run repeated	-Starting with version 1.6.1, this event also captures resource usage (cpu, memory, gpu) per experiment.
experiment run cancelled	-

Table 4: Events Related to Files

Event	Description
file downloaded	-
file updated	-
file deleted	-
file copied	-
file renamed	-
file linked	The logged event indicates when a symlink is created for a file or directory.
directory uploaded	-

Table 5: Events Related to Models

Event	Description
model created	-Starting with version 1.6.1, this event also captures resource usage (cpu, memory, gpu) per model created.
model deleted	-

Table 6: Events Related to Jobs

Event	Description
job created	-
job started	-Starting with version 1.6.1, this event also captures resource usage (cpu, memory, gpu) per job started.
stopped all runs for job	-
job shared with user	-
job unshared with user	-
job sharing updated	The logged event indicates when the sharing status for a job is changed from one of the following options to another: • All anonymous users with the link • All authenticated users with the link • Specific users and teams

Table 7: Events Related to Licenses

Event	Description
license created	-
license deleted	-

Table 8: Events Related to Projects

Event	Description
project created	-
project updated	-
project deleted	-
collaborator added	-
collaborator removed	-

Event	Description
collaborator invited	-

Table 9: Events Related to Sessions

Event	Description
session launched	-Starting with version 1.6.1, this event also captures resource usage (cpu, memory, gpu, custom session metadata (if enabled in the settings)) per session launched.
session terminated	-
session stopped	-
session shared with user	-
session unshared with user	-
update session sharing status	The logged event indicates when the sharing status for a session is changed from one of the following options to another: • All anonymous users with the link • All authenticated users with the link • Specific users and teams

Table 10: Events Related to Admin Settings

site config updated	The logged event indicates when a setting on the Admin Settings page is changed.

Table 11: Events Related to Teams

Event	Description
add member to team	-
delete team member	-
update team member	-

Table 12: Events Related to Users

Event	Description
forgot password	-
password reset	-
update user	If the logged event shows that a user is banned, that means that the user account has been deactivated and does not count toward the license.
user signup	-
user login	The logged event includes the authorization method, LDAP/SAML or local.
user logout	-
ldap/saml user creation	The logged event indicates when a user is created with LDAP or SAML.