

1.0.0

Troubleshooting Apache Impala

Date published: 2020-11-30

Date modified: 2024-07-26

CLOUDERA

<https://docs.cloudera.com/>

Legal Notice

© Cloudera Inc. 2024. All rights reserved.

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

- Troubleshooting Impala.....4**
- Access Impala Workload Logs..... 4**
 - Locations of Impala Log Files in S3.....5
 - Locations of Impala Log Files in Azure.....7

Troubleshooting Impala

This topic describes the general troubleshooting procedures to diagnose some of the commonly encountered issues in Impala.

| Symptom | Explanation | Recommendation |
|--|--|---|
| Impala takes a long time to start. | Impala instances with large numbers of tables, partitions, or data files take longer to start because the metadata for these objects is broadcast to all <code>impalad</code> nodes and cached. | Adjust timeout and synchronicity settings. |
| Query rejected with the default pool-defined memory limit settings. | Some complex queries fail because the minimum memory reservation per host is greater than the memory available to the query for buffer reservations. | Increase VW t-shirt size so that there are more hosts in the executor group and less memory is needed per host. |
| Joins fail to complete. | There may be insufficient memory. During a join, data from the second, third, and so on sets to be joined is loaded into memory. If Impala chooses an inefficient join order or join mechanism, the query could exceed the total memory available. | <p>Start by gathering statistics with the <code>COMPUTE STATS</code> statement for each table involved in the join.</p> <p>Consider specifying the <code>[SHUFFLE]</code> hint so that data from the joined tables is split up between nodes rather than broadcast to each node.</p> <p>If tuning at the SQL level is not sufficient, add more memory to your system or join smaller data sets.</p> |
| Queries return incorrect results. | Impala metadata may be outdated after changes are performed in Hive. | After inserting data, adding a partition, or other operation in Hive, refresh the metadata for the table with the <code>REFRESH</code> statement. |
| Attempts to complete Impala tasks such as executing <code>INSERT SELECT</code> statements fail. The Impala logs include notes that files could not be opened due to permission denied. | This can be the result of permissions issues. For example, you could use the Hive shell as the hive user to create a table. After creating this table, you could attempt to complete some action, such as an <code>INSERT SELECT</code> on the table. Because the table was created using one user and the <code>INSERT SELECT</code> is attempted by another, this action may fail due to permissions issues. | Ensure the Impala user has sufficient permissions to the table that the Hive user created. |
| Impala fails to start up, with the <code>impalad</code> logs referring to errors connecting to the statestore service and attempts to re-register. | A large number of databases, tables, partitions, and so on can require metadata synchronization, particularly on startup, that takes longer than the default timeout for the statestore service. | Configure the statestore timeout value and possibly other settings related to the frequency of statestore updates and metadata loading. |

Access Impala Workload Logs

Describes how to locate Impala logs in S3 or Azure to diagnose some of the commonly encountered issues in Impala.

Using Impala Logs

The Impala logs record information about:

- Any errors Impala encountered.
- How Impala is configured.
- Jobs Impala has completed.

However, you can use the logs record information to troubleshoot only if the relevant logs are downloaded and then uploaded to a location where you can access them. To download the logs from S3 or Azure you must first identify the locations.

Locations of Impala Log Files in S3

This topic describes how to identify the Amazon S3 locations of Impala logs for the different Impala components.

About this task

The Cloudera Data Warehouse service collects logs from Impala Virtual Warehouses and uploads them to an Amazon S3 location. This S3 log location is configured under an external warehouse directory so that the logs are preserved even if the Virtual Warehouse they are collected from is destroyed.

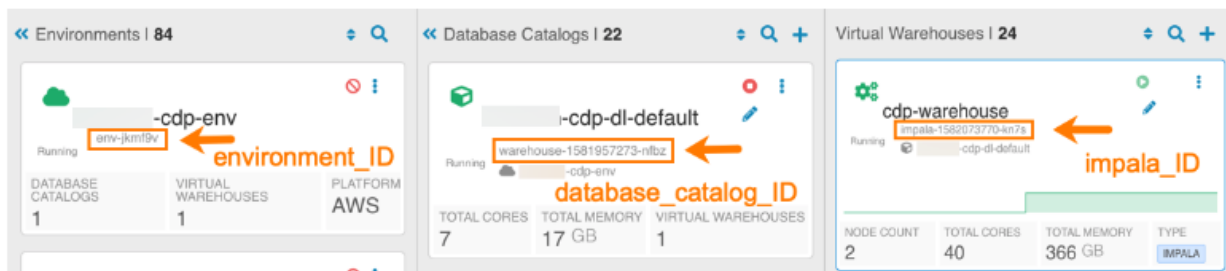
Before you begin

To identify the location of the logs in S3, you must have the `environment_ID`, `database_catalog_ID`, `impala_ID` identifiers, and S3 bucket name.

Procedure

Finding the `environment_ID`, `database_catalog_ID`, and `impala_ID` identifiers

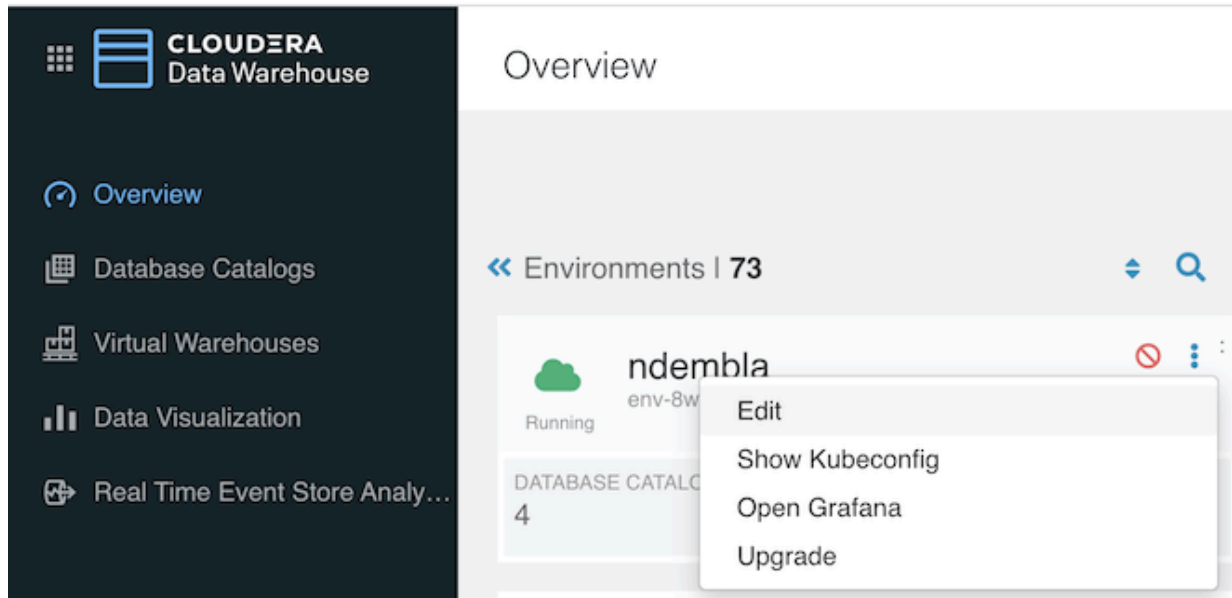
1. In the Data Warehouse service, expand the Environments column by clicking More....
2. From the Overview page, note down the `environment_ID`, `database_catalog_ID`, and `impala_ID` identifiers.



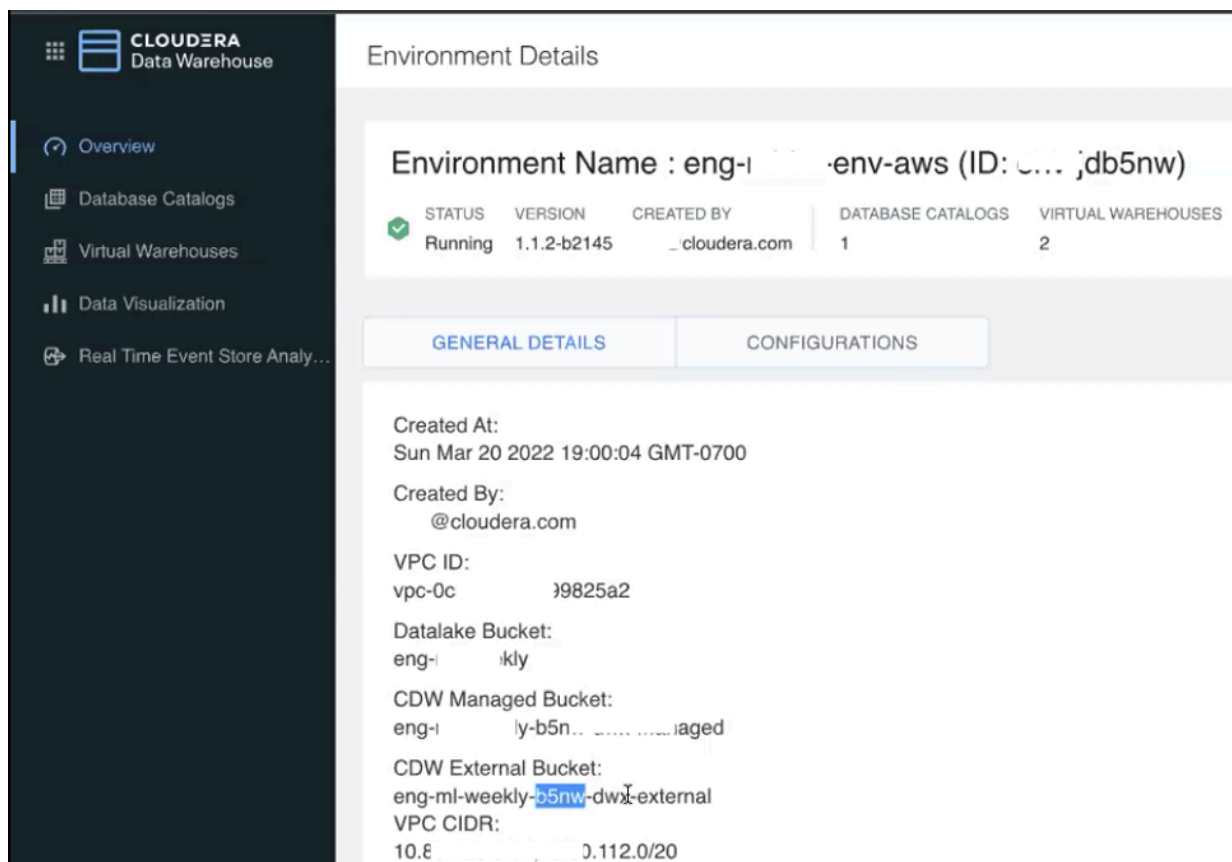
Identifying the external bucket name

3. On the Overview page, locate the environment for which you want to find the external bucket name.

4. In the Environment tile, click the Options menu and select Edit.



5. A dialog opens that shows the general details of the environment including the CDW External Bucket name. This name is required to identify the S3 location of the logs.



Log locations in S3

6. Now that you have identified the S3 bucket name, `environment_ID`, `database_catalog_ID`, and `impala_ID` identifiers, use the following prefix to find the logs generated by specific components in the following directories. Use the different directories listed here to view Impala/Hue logs.

```
PREFIX =
s3://<s3_bucket_name>/clusters/<environment_ID>/<database_catalog_ID>/warehouse/tablespace/external/hive/sys.db/logs/dt=<date_stamp>/ns=<impala_ID>
```

| Impala component | S3 directory location |
|--------------------------|--|
| impalad | PREFIX + "app=impala-executor-log" |
| catalogd | PREFIX + "app=catalogd-log" |
| coordinator | PREFIX + "app=coordinator-log" |
| auto-scaler | PREFIX + "app=impala-autoscaler-log" |
| Hue | PREFIX + "app=huebackend-log" PREFIX + "app=hue-huedb-create-job-log" PREFIX + "app=huefrontend-log" |
| statestored | PREFIX + "app=statestored-log" |
| hs2 (applies only to UA) | PREFIX + "app=hiveserver2" |

The impalad logs for 8 March 2020 are located in the following S3 location:

```
s3://<s3_bucket_name>/clusters/<environment_ID>/<database_catalog_ID>/warehouse/tablespace/external/hive/sys.db/logs/dt=2020-03-08/ns=<impala_ID>/app=impala-executor-log/
```

In the above location, you can find multiple logs that were generated on the specified day.

Impala Minidumps

7. Impala minidumps can be found under the 'debug-artifacts/impala' directory

```
/clusters/{environment_ID}/{database_catalog_ID}/warehouse/debug-artifacts/impala/{impala_ID}/minidump/$POD_NAME/$file
```

Impala Query Profiles

8. Impala query profiles are written in thrift encoded format in this location:

| Impala component | S3 directory location |
|-----------------------|--------------------------------|
| Impala query profiles | PREFIX + "app=impala-profiles" |

Use the binary tool to decode thrift to text. This binary tool is provided with the upstream runtime Impala 4.0 as a docker image. Run the following command to use this tool.

```
docker run -i apache/impala:4.0.0-impala_profile_tool < name of the thrift encoded file to decode
```

You can use the docker image available [here](#) to use this decoding tool.

Locations of Impala Log Files in Azure

This topic describes how to identify the Azure locations of Impala logs for the different Impala components.

About this task

The Cloudera Data Warehouse service collects logs from Impala Virtual Warehouses and uploads them to the Azure storage account that was provided while registering the Environment. This ABFS log location is configured under an external warehouse directory so that the logs are preserved even if the Virtual Warehouse they are collected from is destroyed.

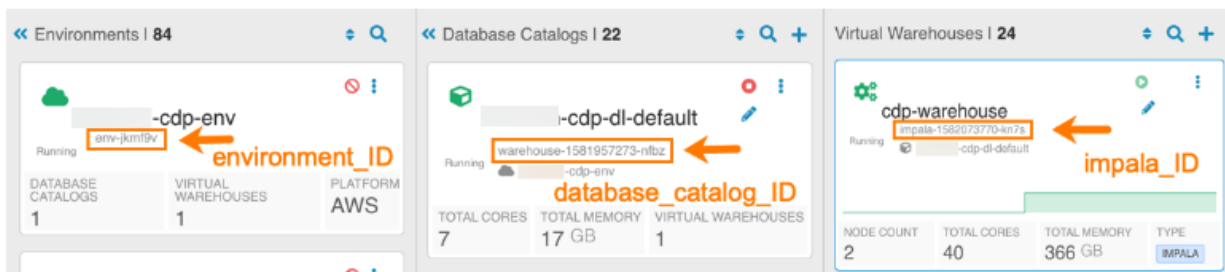
Before you begin

To identify the location of the logs in Azure, you must have the environment_ID, database_catalog_ID, and impala_ID identifiers and to access the logs from the Azure Portal you must know your storage account name.

Procedure

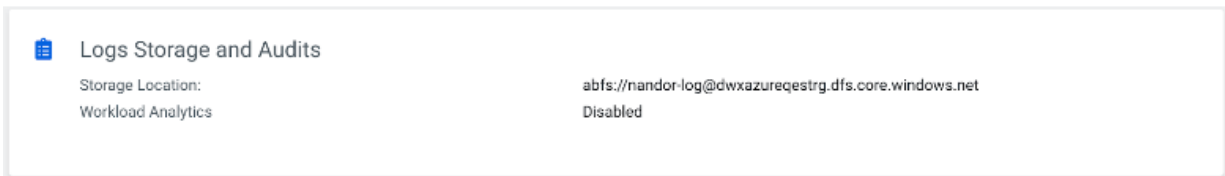
Finding the environment_ID, database_catalog_ID, and impala_ID identifiers

1. In the Data Warehouse service, expand the Environments column by clicking More....
2. From the Overview page, note down the environment_ID, database_catalog_ID, and impala_ID identifiers.



Retrieving your storage account name

3. In the Management Console navigate to the Environments page.
4. On the Environments page, click on your Environment and click on the Summary tab.
5. Scroll down to the Logs Storage and Audits section.



Note down your storage account name.

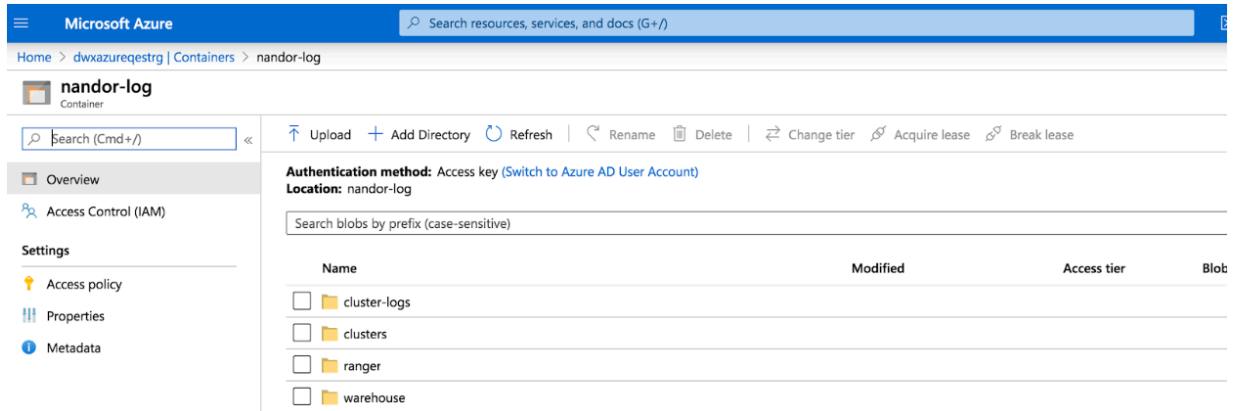
Accessing the different directories in the Azure Portal

6. Log in to the Azure Portal and search for your storage account name using the Search bar.
7. On the Overview page of your storage account, click on the Containers menu.

8. Click on the file system you used during the Environment registration.



Note: You need to enable the firewall rules, click on the Firewalls and virtual networks menu, and set Allow access to “All networks”, then save the changes to access the file system.



Log locations in ABFS

9. Use the environment_ID, database_catalog_ID, and impala_ID identifiers, in the following prefix to find the logs generated by specific components in the following directories. Use the different directories listed here to view Impala/Hue logs

```
PREFIX =
/clusters/<environment_ID>/<database_catalog_ID>/warehouse/tablespace/external/hive/sys.db/logs/dt=<date_stamp>/ns=<impala_ID>/
```

| Impala component | ABFS directory location |
|------------------|--|
| impalad | PREFIX + “app=impala-executor-log” |
| catalogd | PREFIX + “app=catalogd-log” |
| coordinator | PREFIX + “app=coordinator-log” |
| auto-scaler | PREFIX + “app=impala-autoscaler-log” |
| Hue | PREFIX + “app=huebackend-log” PREFIX + “app=hue-huedb-create-job-log” PREFIX + “app=huefrontend-log” |
| statestored | PREFIX + “app=statestored-log” |

The impalad logs for 8 March 2020 are located in the following ABFS location:

```
/clusters/<environment_ID>/<database_catalog_ID>/warehouse/tablespace/external/hive/sys.db/logs/dt=2020-03-08/ns=<impala_ID>/app=impala-executor-log/
```

In the above location, you can find multiple logs that were generated on the specified day.

Impala Minidumps

10. Impala minidumps can be found under the ‘debug-artifacts/impala’ directory

```
/clusters/<environment_ID>/<database_catalog_ID>/warehouse/debug-artifacts/impala/<impala_ID>/minidump/<pod_name>/
```

Impala Query Profiles

11. Impala query profiles are written in thrift encoded format in this location:

| | |
|-----------------------|--------------------------------|
| Impala component | S3 directory location |
| Impala query profiles | PREFIX + "app=impala-profiles" |

Use the binary tool to decode thrift to text. This binary tool is provided with the upstream runtime Impala 4.0 as a docker image. Run the following command to use this tool.

```
docker run -i apache/impala:4.0.0-impala_profile_tool < name of the thrift  
encoded file to decode
```

You can use the docker image available [here](#) to use this decoding tool.