1.0.0

Troubleshooting Apache Impala

Date published: 2020-11-30
Date modified: 2025-10-22

CLOUDERA

Legal Notice

Contents

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 `impalad` 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.	Start by gathering statistics with the COMPUTE STATS statement for each table involved in the join. Consider specifying the [SHUFFLE] hint so that data from the joined tables is split up between nodes rather than broadcast to each node. If tuning at the SQL level is not sufficient, add more memory to your system or join smaller data sets.
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	Impala metadata may be outdated after changes are performed in Hive.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 REFRESH statement.
Attempts to complete Impala tasks such as executing INSE RT SELECT 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 INSERT SELECT on the table. Because the table was created using one user and the INSERT SELECT 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 `impalad` logs referring to errors connecting to the statestored 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 stat estore 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 name, impala ID identifiers, and S3 bucket name.

Procedure

Finding the environment name, and impala ID identifiers

1. In the Data Warehouse service, expand the Environment column by clicking ⋮ .
2. From the Overview page, note down the environment name, and impala ID identifiers.
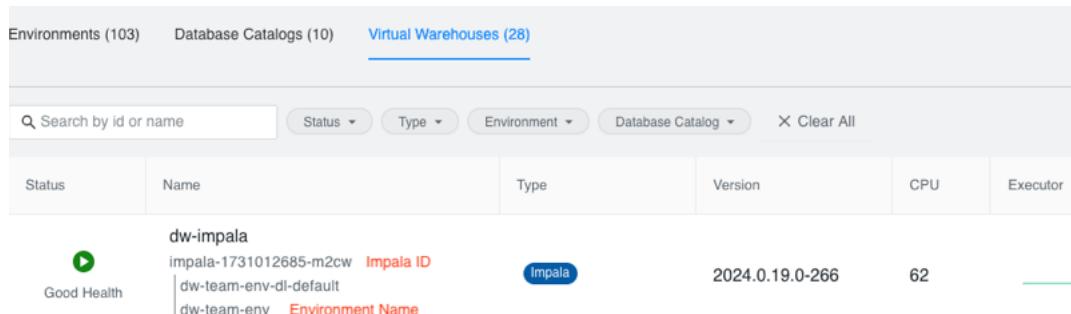
Identifying the bucket name

3. In the Management console navigate to the Environment page.
4. On the Environment page, click on your Environment and then on the Summary tab.
5. On the Logs Storage and Audits section. You can find your log bucket and prefix under the Storage Location. This name is required to identify the S3 location of the logs.
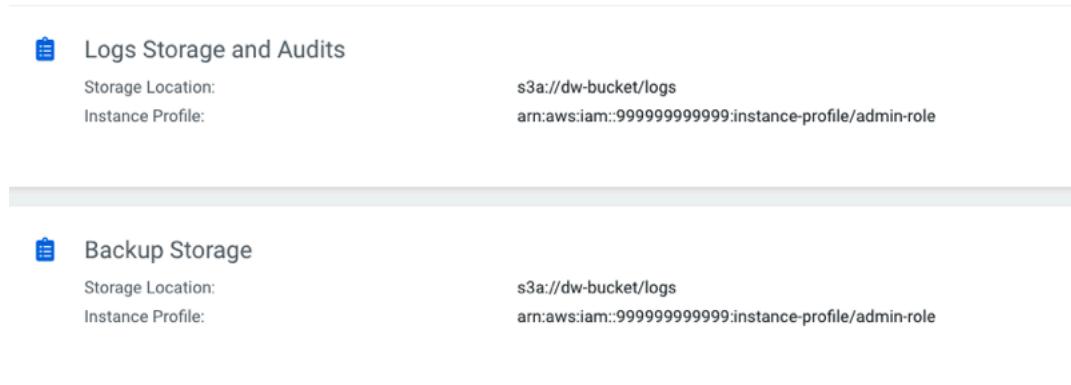
Log locations in S3

6. Now that you have identified the S3 bucket name, environment name, 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>/<prefix>/cdw/clusters/<environment_name>/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"

The impalad executor logs for 25 August 2024 are located in the following S3 location:

```
PREFIX = s3://<s3_bucket_name>/<prefix>/cdw/clusters/<environment_name>/
logs/dt=2024-08-25/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

```
/cdw/clusters/<environment_name>/logs/debug-artifacts/impala/<impala_ID>/
minidump/<pod_name>/
```

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 name and impala ID identifiers and to access the logs from the Azure Portal you must know your storage account name.

Procedure

Finding the environment name, impala ID identifiers

1. In the Cloudera Data Warehouse service, go to the Virtual Warehouse tab.
2. From the Overview page, note down the environment name 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 name, 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 =
/cdw/clusters/<environment_name>/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 executor logs for 25 August 2024 are located in the following ABFS location:

```
PREFIX = /cdw/clusters/<environment_name>/logs/dt=2024-08-25/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

```
/cdw/clusters/<environment_name>/logs/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	ABFS 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.