Enable AWS Glue Access

Contents:

- *Supported Deployment*
  - *Required Glue table properties*
- *Limitations*
- *Enable*
- *Use*

> (i) **NOTE:** This feature is in Beta release.

If you have integrated with an EMR cluster version 5.8.0 or later, you can configure your Hive instance to use AWS Glue Data Catalog for storage and access to Hive metadata.

> ⊘ **Tip:** For metastores that are used across a set of services, accounts, and applications, AWS Glue is the recommended method of access.

For more information on AWS Glue, see *https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hive-metastore-glue.html*.

This section describes how to enable integration with your AWS Glue deployment.

Supported Deployment

AWS Glue tables can be read under the following conditions:

- The Trifacta platform is integrated with an EMR cluster:
  - EMR version 5.8.0 or later
  - EMR cluster has been configured with HiveServer2
- The Hive deployment must be integrated with AWS Glue.

  > (i) **NOTE:** Hive connections are supported when S3 is the backend datastore.

- For HiveServer2 connectivity, the Trifacta node has direct access to the Master node of the EMR cluster.
- Hive metastore must be configured to use AWS Glue
- For Hive on AWS EMR to access AWS Glue, EMR roles assigned to the cluster should have the AWS Glue functions in their IAM roles. For more information, see *https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-iam-roles-defaultroles.html#emr-iam-contents-ec2role*   .

Required Glue table properties

Each Glue table must be created with the following properties specified:

- `InputFormat`
- `OutputFormat`
- `Serde`

These properties must be specified for the Hive JDBC driver to read the Glue tables.

For additional limitations on access Hive tables through Glue, see
https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hive-metastore-glue.html#emr-hive-glue-considerations-hive .

Limitations

- Access is read-only. Publishing to Hive hosted on EMR is not supported.
- You cannot select datasets through the Database browser in the Trifacta application.

> (i) **NOTE:** Use of this integration requires the development of custom SQL queries against the AWS Glue metadata store.

Enable

Please verify the following have been enabled and configured.

1. Your deployment has been configured to meet the Supported Deployment guidelines above.
2. You must integrate the platform with Hive.

   > (i) **NOTE:** For the Hive hostname and port number, use the Master public DNS values. For more information, see
   > *https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hive-metastore-glue.html*.

   For more information, see *Configure for Hive*.
3. AWS Glue tables can be accessed only through custom SQL statements. The custom SQL query feature must be enabled. For more information, see *Enable Custom SQL Query*.

Use

After the integration has been made between the platform and AWS Glue, you can import datasets using custom SQL queries. For more information, see *Create Dataset with SQL*.