Configure Data Service

Contents:

The Data Service enables the Trifacta® platform to stream metadata and records from JDBC sources for sampling and job execution in the Trifacta Photon running environment. This section describes how to enable and configure the service, including performance tweaks and connection-specific configuration.

Configure Service

The following basic properties enable the service and specify basic location for it.

You can apply this change through the *Admin Settings Page* (recommended) or

`trifacta-conf.json`

. For more information, see *Platform Configuration Methods*.

Property	Description
`"data-service.enabled"`	When `true`, the data service is enabled. **NOTE:** When set to `false`, access to any relational connection is prevented. Default is `true` .
`"data-service.host"`	Hostname for the service. Default is `localhost`.
`"data-service.port"`	Port number used by the service. Default is `41912` . **NOTE:** If you are changing the port number, avoid creating conflicts with existing ports in use. For more information, see *System Ports*.
`"data-service.classpath"`	The Java class path for the data service.
`"data-service.autoRestart"`	When `true`, the data service is automatically restarted if it crashes. Default is `true` .

`"data-service.vendorPath"`	Path to the vendor configuration files for relational connections. Default value: `%(topOfTree)s/services/data-service/build/conf/vendor`

Configure SQL Options

Configure relational read stream limits

You can apply this change through the *Admin Settings Page* (recommended) or

`trifacta-conf.json`

. For more information, see *Platform Configuration Methods*.

The Data Service reads data from relational sources in streams of records. You can modify the following parameters to configure the limits of SQL record streaming during read operations. The size of these streams are defined by the following parameters:

```
"data-service.sqlOptions.maxReadStreamRecords": -1,
"data-service.sqlOptions.limitedReadStreamRecords": 1000000,
"data-service.sqlOptions.initialReadStreamRecords": 25,
"data-service.sqlOptions.hiveReadStreamRecords": 100000000,
```

Property	Description
`"data-service.sqlOptions.maxReadStreamRecords"`	The maximum number of JDBC records pulled in per stream read during batch execution. If this value is set to `-1`, then no limit is applied.
`"data-service.sqlOptions.limitedReadStreamRecords"`	Max number of records read for the initial sample and quick scan sampling. Set ting to -1 means there is no limit.
`"data-service.sqlOptions.initialReadStreamRecords"`	Initial number of records to read for client-side preview and for client-side transform. Set to `-1` to apply no limit.
`"data-service.sqlOptions.hiveReadStreamRecords"`	Max number of records that can be read from Hive, if maxReadStreamRecords is `-1`. **NOTE:** This value cannot be set to `-1`, which results in a Data Service error. Hive reads must be limited.

Configure Caching

The data service maintains a cache of JDBC objects that have been retrieved for use. You can configure the following properties to tune settings of the cache.

You can apply this change through the *Admin Settings Page* (recommended) or

`trifacta-conf.json`

. For more information, see *Platform Configuration Methods*.

Property	Description
`"data-service.cacheOptions.validationDelayMilliseconds"`	Number of milliseconds to wait between checks validating cached pools. Default is `3600000` (1 hour).

`"data-service.cacheOptions.maxSize"`	Maximum number of objects in the cache. Default is `100`. **NOTE:** Set this value to `0` to disable data service caching.
`"data-service.cacheOptions.expirySeconds"`	Objects in the cache that are older than this number of seconds are automatically expired. Default is `86400` (1 day).

Configure for Specific Integrations

Configure Data Service for Hive

The following properties apply to how the platform connects to Hive.

You can apply this change through the *Admin Settings Page* (recommended) or

`trifacta-conf.json`

. For more information, see *Platform Configuration Methods*.

Property	Description
`"data-service.hiveManagedTableFormat"`	Managed table format for your Hive deployment. Default is `PARQUET`.
`"data-service.hiveJdbcJar"`	Path to the JAR to use for JDBC connectivity to Hive. Default path depends on your Hadoop distribution.

Configure Data Service for Tableau Server

The following properties apply to how the platform publishes to Tableau Server.

You can apply this change through the *Admin Settings Page* (recommended) or

`trifacta-conf.json`

. For more information, see *Platform Configuration Methods*.

Property	Description
`"data-service.tableauBufferSizeInBytes"`	Number of bytes of data to include in each HTTP request chunk when publishing to Tableau Server. When the Trifacta application publishes a file to Tableau Server, the file is divided into chunks, and each chunk is attached as part of a HTTP request payload. This flag controls the chunk size in bytes. When the chunk size is large, the number of HTTP requests required to send the whole file to Tableau Server is smaller. However, a large chunk size increases the risk of a RequestTimeoutException, which causes the publishing job to fail. Default is `3000000` bytes.

Additional Configuration

The following aspects of the data service can be configured outside of the application:

- Connection pool size and retry parameters
- Vendor field mappings
- Oracle ciphers for SSL connections
- JDBC fetch size by vendor

For more information, please contact *Trifacta Customer Success Services*.

Logging

For more information on logging for the service, see *Configure Logging for Services*.

Other Topics

- If you are reading large datasets from relational sources, you can enable JDBC ingestion, which reads source data in the background and stages on the backend datastore for execution. For more information, see *Configure JDBC Ingestion*.
- Optionally, SSO authentication can be applied to relational connections. For more information, see *Enable SSO for Relational Connections*.