



TIBCO® Data Virtualization

Azure DataLake Adapter Guide

Version 8.7.0 | October 2023

Contents

| | |
|---|-----------|
| Contents | 2 |
| TDV Azure Data Lake Adapter | 3 |
| Connecting to Azure Data Lake Data Source | 3 |
| Basic Tab | 4 |
| Advanced Tab | 4 |
| Data Type Mapping | 7 |
| Azure Data Lake Storage to TDV Data Types | 7 |
| TDV to Azure Data Lake Storage Data Types | 11 |
| Supported Functions | 13 |
| Limitations | 14 |
| TIBCO Product Documentation and Support Services | 15 |
| How to Access TIBCO Documentation | 15 |
| How to Contact TIBCO Support | 16 |
| Release Version Support | 16 |
| How to Join TIBCO Community | 17 |
| Legal and Third-Party Notices | 18 |

TDV Azure Data Lake Adapter

TDV supports the following Cloud based File Systems:

- Amazon S3
- Microsoft Azure Data Lake Storage
- Local File Storage
- Google Cloud Storage

This topic describes the configuration of the Azure Data Lake Storage adapter. Refer to the corresponding Adapter guides for details on how to configure those.

The File formats supported for Azure Data Lake adapter are:

- Delimited file format
- Parquet file format

Note: The delimited file format is a flat file format while the Parquet files can store data in a hierarchical format also. If the “Infer Schema” option is selected for the data source, TDV will infer the schema and datatypes of each column based on the data in the file.

In this chapter, the following topics are discussed:

[Connecting to Azure Data Lake Data Source](#)

[Data Type Mapping](#)

[Supported Functions](#)

Connecting to Azure Data Lake Data Source

From the New Data Source dialog window, choose Azure Data Lake Storage data source to connect to it. The following sections explain the different connection parameters.

Basic Tab

To connect to the Azure Data Lake Storage adapter, set the following properties in the Basic tab of the New Data Source connection window:

| Field | Description |
|-----------------------|--|
| Data Source Name | The name of the data source. |
| URL | A URL to connect to the physical data source. |
| Application/Client ID | Application/Client ID is the ID created when registering your application in the Microsoft active directory. You will need this to connect to the Azure Data Lake storage. |
| Client Secret | This is the secret string that the application uses to prove its identity when requesting a token. You will need this to connect to the Azure Data Lake storage. |
| Refresh Url | This is the refresh token that is needed for the client to stay connected to the storage. To get this value, in the Azure portal, go to Azure Active Directory > App registrations > Endpoints*. The OAUTH 2.0 TOKEN ENDPOINT found in the Endpoints region is the value for the RefreshUrl. |

Advanced Tab

To connect to the Azure Data Lake Storage Adapters, set the following properties in the advanced tab of the New Data Source connection window:

| Field | Description |
|--------------------------|--|
| Concurrent Request Limit | This configuration can take a value between 0 to 65536. It specifies the concurrency limits to be imposed on the underlying data source. |

| Field | Description |
|---------------------------------------|---|
| Default String Length | The default VARCHAR length. |
| Detect Partition During Introspection | <p>Include this option to automatically detect partitions that the file might have.</p> <p>Note that if they are not properly detected, both usability and performance will be adversely impacted.</p> |
| CSV Options | |
| Include CSV Files | Check this option to include the delimited files from the storage area. |
| Character Set | The character set used by the datasource. |
| Delimiter | Indicates the file delimiter character. |
| Text Qualifier | Indicates the type of qualifier that is used in the file to enclose a string field. |
| Has Header Row | Indicates whether or not the file has a header row. |
| Infer Schema | <p>Choosing this option enables the parser to infer the schema and datatypes of each column based on the data in the file.</p> <p>Note: If this option is selected, it is recommended to provide a “sampling ratio” while introspecting the data source, where sampling of the data might be used when inferring the schema. Providing the sampling ratio helps reduce the overhead of not having to read all the rows while inferring the schema. Parquet files do not require schema inference as their schema is encoded in their metadata.</p> |
| CSV Escape Character | Indicates the character that should be ignored by the parser in the file. |
| CSV Parser Lib | The libraries used to parse the delimited files. The libraries supported |

| Field | Description |
|------------------------|---|
| | <p>currently are commons (default) and uniVocity. For more information, refer:</p> <ul style="list-style-type: none"> • http://commons.apache.org/proper/commons-csv/ • https://www.univocity.com/ |
| CSV Parsing Mode | The various parsing modes used by the data source. Allowed values are “PERMISSIVE (include a malformed row), DROPMALFORMED (Drop bad rows), FAILFAST (Fail the introspection when a bad row is encountered). |
| CSV Comment Character | Indicates the character that is used as comment in the file. |
| CSV Null Value | Indicates what is considered a Null value in a row. |
| CSV File Name Filters | Indicates the file name extensions that are valid. |
| Parquet Options | |
| Include Parquet Files | Check this option to include the parquet files from the storage area. |
| Binary as String | Check this option to read binary value as string. |
| INT96 as Timestamp | Check this option to read INT96 value as Timestamp. |
| Compression Codec | <p>Parquet files are typically compressed. This setting controls the compression algorithm used to process them. For more information about the different options, refer</p> <p>https://spark.apache.org/docs/2.4.3/sql-data-sources-parquet.html</p> |
| Filter Push-Down | Controls whether a predicate specified in a WHERE clause in a SQL query will be pushed down to the Cloud File System data source. |
| Convert | Controls whether to use the built-in Parquet reader and writer for Hive |

| Field | Description |
|---------------------------|--|
| Metastore | tables with the parquet storage format. By default, this is set to True. |
| Merge Schema | In case of partitioned files, choosing this option merges the data and creates a single schema that includes columns from all partitions. |
| Write Legacy Format | Use this option to choose how the data should be written. If true, data will be written in a way of Spark 1.4 and earlier. For example, decimal values will be written in Apache Parquet's fixed-length byte array format, which other systems such as Apache Hive and Apache Impala use. If false, the newer format in Parquet will be used. For example, decimals will be written in int-based format. If Parquet output is intended for use with systems that do not support this newer format, set this to true. |
| Parquet File Name Filters | Indicates the file name extensions that are valid. |

Data Type Mapping

Azure Data Lake Storage to TDV Data Types

Mapped Azure Data Lake data types have the following restrictions:

- Maximum VARBINARY length is 2000.
- Maximum CHAR length is 10485760.
- Maximum VARCHAR length is 10485760.

The following table shows the mapping from Azure Data Lake data types to TDV data types.

| Azure Data Lake Storage Data Type | TDV Data Type |
|--|----------------------|
| BYTEA | BLOB |
| CHAR | CHAR |
| CHARACTER | CHAR |
| CHARACTER_VARYING | VARCHAR |
| DECIMAL | DECIMAL |
| BPCHAR | CHAR |
| TEXT | CLOB |
| FLOAT | FLOAT |
| LONG | CLOB |
| NUMBER | DECIMAL |
| RAW | BYTEA |
| ROWID | VARCHAR |
| UROWID | VARCHAR |
| VARCHAR | VARCHAR |
| VARCHAR2 | VARCHAR |
| DATETIME | TIMESTAMP |
| TIMESTAMP | TIMESTAMP |
| TIMESTAMPZ | TIMESTAMP |

| Azure Data Lake Storage Data Type | TDV Data Type |
|-----------------------------------|---------------|
| SMALLDATETIME | TIMESTAMP |
| TIMETZ | TIME |
| DOUBLE | DOUBLE |
| FLOAT | FLOAT |
| FLOAT4 | REAL |
| FLOAT8 | DOUBLE |
| REAL | REAL |
| INTEGER | INTEGER |
| INT | INTEGER |
| BIGINT | BIGINT |
| INT8 | BIGINT |
| INT4 | INTEGER |
| INT2 | SMALLINT |
| SMALLINT | SMALLINT |
| BOOL | BOOLEAN |
| BOOLEAN | BOOLEAN |
| BIT | CHAR |
| VARBIT | VARCHAR |

| Azure Data Lake Storage Data Type | TDV Data Type |
|-----------------------------------|---------------|
| TINYINT | SMALLINT |
| NUMERIC | NUMERIC |
| UUID | CHAR |
| XID | INTEGER |
| XML | XML |
| BOX | VARCHAR |
| OID | BLOB |
| BINARY_DOUBLE | DOUBLE |
| DOUBLE_PRECISION | DOUBLE |
| CIDR | VARCHAR |
| INET | VARCHAR |
| LINE | VARCHAR |
| LSEG | VARCHAR |
| MACADDR | VARCHAR |
| MONEY | DECIMAL |
| SERIAL | INTEGER |
| BIGSERIAL | BIGINT |
| CIRCLE | VARCHAR |

| Azure Data Lake Storage Data Type | TDV Data Type |
|-----------------------------------|---------------|
| PATH | VARCHAR |
| POINT | CHAR |
| POLYGON | VARCHAR |
| BINARY_FLOAT | REAL |

TDV to Azure Data Lake Storage Data Types

| TDV Data Types | Azure Data Lake Storage Data Types |
|----------------|------------------------------------|
| DECIMAL_FLOAT | VARCHAR(128) |
| BIGINT | BIGINT |
| BINARY | BYTEA |
| BINARY_PROMOTE | BYTEA |
| BIT | SMALLINT |
| BLOB | BYTEA |
| BOOLEAN | BOOLEAN |
| CHAR | CHAR(&1) |
| CHAR_PROMOTE | TEXT |
| CLOB | TEXT |

| TDV Data Types | Azure Data Lake Storage Data Types |
|-----------------------|---|
| DATE | VARCHAR |
| DECIMAL | DECIMAL(&p, &s) |
| DECIMAL_PROMOTE | TEXT |
| DOUBLE | DOUBLE |
| FLOAT | FLOAT |
| INTEGER | INTEGER |
| NUMERIC | DECIMAL(&p, &s) |
| NUMERIC_PROMOTE | TEXT |
| REAL | REAL |
| SMALLINT | SMALLINT |
| TIME | VARCHAR |
| TIMESTAMP | TIMESTAMP |
| TINYINT | SMALLINT |
| VARBINARY | BYTEA |
| VARBINARY_PROMOTE | BYTEA |
| VARCHAR | VARCHAR(&1) |
| VARCHAR_PROMOTE | TEXT |
| XML | XML |

| TDV Data Types | Azure Data Lake Storage Data Types |
|---------------------------|---|
| INTERVAL_YEAR | INTERVAL YEAR |
| INTERVAL_MONTH | INTERVAL MONTH |
| INTERVAL_DAY | INTERVAL DAY |
| INTERVAL_HOUR | INTERVAL HOUR |
| INTERVAL_MINUTE | INTERVAL MINUTE |
| INTERVAL_SECOND | INTERVAL SECOND |
| INTERVAL_YEAR_TO_MONTH | INTERVAL_YEAR_TO_MONTH |
| INTERVAL_DAY_TO_HOUR | INTERVAL_DAY_TO_HOUR |
| INTERVAL_DAY_TO_MINUTE | INTERVAL DAY TO MINUTE |
| INTERVAL_DAY_TO_SECOND | INTERVAL DAY TO SECOND |
| INTERVAL_HOUR_TO_MINUTE | INTERVAL HOUR TO MINUTE |
| INTERVAL_HOUR_TO_SECOND | INTERVAL HOUR TO SECOND |
| INTERVAL_MINUTE_TO_SECOND | INTERVAL MINUTE TO SECOND |

Supported Functions

Following functions are supported within TDV for the Azure Data Lake Storage Cloud File System adapter:

- CAST
- COUNT
- EXTRACT
- LOWER

- MAX
- MIN
- NTILE
- SUM
- UPPER

Limitations

CSV file names with the special characters {,} , [,] are not supported.

Cloud file system adapters are currently not supported on Windows OS.

TIBCO Product Documentation and Support Services

For information about this product, you can read the documentation, contact TIBCO Support, and join the TIBCO Community.

How to Access TIBCO Documentation

Documentation for TIBCO products is available on the [TIBCO Product Documentation](#) website, mainly in HTML and PDF formats.

The [TIBCO Product Documentation](#) website is updated frequently and is more current than any other documentation included with the product.

Product-Specific Documentation

The following documentation for this product is available on the [TIBCO® Data Virtualization](#) page.

- **Users**
 - TDV Getting Started Guide
 - TDV User Guide
 - TDV Web UI User Guide
 - TDV Client Interfaces Guide
 - TDV Tutorial Guide
 - TDV Northbay Example
- **Administration**
 - TDV Installation and Upgrade Guide
 - TDV Administration Guide
 - TDV Active Cluster Guide
 - TDV Security Features Guide
- **Data Sources**

TDV Adapter Guides

TDV Data Source Toolkit Guide (Formerly Extensibility Guide)

- **References**

TDV Reference Guide

TDV Application Programming Interface Guide

- **Other**

TDV Business Directory Guide

TDV Discovery Guide

- *TIBCO TDV and Business Directory Release Notes* Read the release notes for a list of new and changed features. This document also contains lists of known issues and closed issues for this release.

How to Contact TIBCO Support

Get an overview of [TIBCO Support](#). You can contact TIBCO Support in the following ways:

- For accessing the Support Knowledge Base and getting personalized content about products you are interested in, visit the [TIBCO Support](#) website.
- For creating a Support case, you must have a valid maintenance or support contract with TIBCO. You also need a user name and password to log in to [TIBCO Support](#) website. If you do not have a user name, you can request one by clicking **Register** on the website.

Release Version Support

TDV 8.5 is designated as a Long Term Support (LTS) version. Some release versions of TIBCO® Data Virtualization products are selected to be long-term support (LTS) versions. Defect corrections will typically be delivered in a new release version and as hotfixes or service packs to one or more LTS versions. See also

https://docs.tibco.com/pub/tdv/general/LTS/tdv_LTS_releases.htm.

How to Join TIBCO Community

TIBCO Community is the official channel for TIBCO customers, partners, and employee subject matter experts to share and access their collective experience. TIBCO Community offers access to Q&A forums, product wikis, and best practices. It also offers access to extensions, adapters, solution accelerators, and tools that extend and enable customers to gain full value from TIBCO products. In addition, users can submit and vote on feature requests from within the [TIBCO Ideas Portal](#). For a free registration, visit [TIBCO Community](#).

Legal and Third-Party Notices

SOME TIBCO SOFTWARE EMBEDS OR BUNDLES OTHER TIBCO SOFTWARE. USE OF SUCH EMBEDDED OR BUNDLED TIBCO SOFTWARE IS SOLELY TO ENABLE THE FUNCTIONALITY (OR PROVIDE LIMITED ADD-ON FUNCTIONALITY) OF THE LICENSED TIBCO SOFTWARE. THE EMBEDDED OR BUNDLED SOFTWARE IS NOT LICENSED TO BE USED OR ACCESSED BY ANY OTHER TIBCO SOFTWARE OR FOR ANY OTHER PURPOSE.

USE OF TIBCO SOFTWARE AND THIS DOCUMENT IS SUBJECT TO THE TERMS AND CONDITIONS OF A LICENSE AGREEMENT FOUND IN EITHER A SEPARATELY EXECUTED SOFTWARE LICENSE AGREEMENT, OR, IF THERE IS NO SUCH SEPARATE AGREEMENT, THE CLICKWRAP END USER LICENSE AGREEMENT WHICH IS DISPLAYED DURING DOWNLOAD OR INSTALLATION OF THE SOFTWARE (AND WHICH IS DUPLICATED IN THE LICENSE FILE) OR IF THERE IS NO SUCH SOFTWARE LICENSE AGREEMENT OR CLICKWRAP END USER LICENSE AGREEMENT, THE LICENSE(S) LOCATED IN THE “LICENSE” FILE(S) OF THE SOFTWARE. USE OF THIS DOCUMENT IS SUBJECT TO THOSE TERMS AND CONDITIONS, AND YOUR USE HEREOF SHALL CONSTITUTE ACCEPTANCE OF AND AN AGREEMENT TO BE BOUND BY THE SAME.

This document is subject to U.S. and international copyright laws and treaties. No part of this document may be reproduced in any form without the written authorization of TIBCO Software Inc.

TIBCO, TIBCO logo, TIBCO O logo, ActiveSpaces, Enterprise Messaging Service, Spotfire, TERR, S-PLUS, and S+ are either registered trademarks or trademarks of TIBCO Software Inc. in the United States and/or other countries.

Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle Corporation and/or its affiliates.

All other product and company names and marks mentioned in this document are the property of their respective owners and are mentioned for identification purposes only.

This software may be available on multiple operating systems. However, not all operating system platforms for a specific software version are released at the same time. See the

readme file for the availability of this software version on a specific operating system platform.

THIS DOCUMENT IS PROVIDED “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NON-INFRINGEMENT.

THIS DOCUMENT COULD INCLUDE TECHNICAL INACCURACIES OR TYPOGRAPHICAL ERRORS. CHANGES ARE PERIODICALLY ADDED TO THE INFORMATION HEREIN; THESE CHANGES WILL BE INCORPORATED IN NEW EDITIONS OF THIS DOCUMENT. TIBCO SOFTWARE INC. MAY MAKE IMPROVEMENTS AND/OR CHANGES IN THE PRODUCT(S) AND/OR THE PROGRAM(S) DESCRIBED IN THIS DOCUMENT AT ANY TIME.

THE CONTENTS OF THIS DOCUMENT MAY BE MODIFIED AND/OR QUALIFIED, DIRECTLY OR INDIRECTLY, BY OTHER DOCUMENTATION WHICH ACCOMPANIES THIS SOFTWARE, INCLUDING BUT NOT LIMITED TO ANY RELEASE NOTES AND "READ ME" FILES.

This and other products of TIBCO Software Inc. may be covered by registered patents. Please refer to TIBCO's Virtual Patent Marking document (<https://www.tibco.com/patents>) for details.

Copyright © 2002-2023 Cloud Software Group, Inc All Rights Reserved.