



Denodo Distributed File System Custom Wrapper

Revision 20181121

NOTE

This document is confidential and proprietary of **Denodo Technologies**.
No part of this document may be reproduced in any form by any means without prior written authorization of **Denodo Technologies**.

Copyright © 2018
Denodo Technologies Proprietary and Confidential

CONTENTS

1 INTRODUCTION.....	6
1.1 DELIMITED TEXT FILES.....	6
1.2 SEQUENCEFILES.....	6
1.3 MAPFILES.....	6
1.4 AVRO FILES.....	7
1.5 PARQUET FILES.....	7
2 USAGE.....	8
2.1 IMPORTING THE CUSTOM WRAPPER INTO VDP.....	8
2.2 CREATING A DISTRIBUTED FILE SYSTEM DATA SOURCE.....	10
2.3 CREATING A BASE VIEW.....	10
3 AMAZON S3.....	39
3.1 CONFIGURING S3 AUTHENTICATION PROPERTIES.....	39
3.2 CONFIGURING S3N AUTHENTICATION PROPERTIES.....	39
3.3 CONFIGURING S3A AUTHENTICATION PROPERTIES.....	40
3.4 SIGNATURE VERSION 4 SUPPORT.....	43
4 AZURE DATA LAKE STORE.....	44
4.1 CONFIGURING AUTHENTICATION PROPERTIES.....	44
5 AZURE BLOB STORAGE.....	45
5.1 CONFIGURING AUTHENTICATION PROPERTIES.....	45
6 GOOGLE CLOUD STORAGE.....	46
6.1 CONFIGURING AUTHENTICATION PROPERTIES.....	46
6.2 PERMISSIONS.....	46
7 COMPRESSED FILES.....	47
8 SECURE CLUSTER WITH KERBEROS.....	48
9 TROUBLESHOOTING.....	51
10 APPENDICES.....	53
10.1 HOW TO USE THE HADOOP VENDOR'S CLIENT LIBRARIES.....	53

Warning

Although this wrapper is capable of reading files stores in HDFS, Amazon S3, Azure Blob Storage, Azure Data Lake Storage and Google Cloud Storage, most of the technical artifacts of this wrapper include HDFS in their names for legacy compatibility:

- Jars: denodo-hdfs-custom-wrapper-xxx
- Wrappers: com.denodo.connect.hadoop.hdfs.wrapper.HDFSxxxFileWrapper

1 INTRODUCTION

The Distributed File System Custom Wrapper distribution contains five Virtual DataPort custom wrappers capable of reading several file formats stored in **HDFS**, **Amazon S3**, **Azure Blob Storage**, **Azure Data Lake Storage** and **Google Cloud Storage**.

Supported formats are:

- Delimited text files
- Sequence files
- Map files
- Avro files
- Parquet files

Also, there is a custom wrapper for retrieve information from the distributed file system and display it in a relational way:

- DFSListFilesWrapper

This wrapper allows to inspect distributed folders, retrieve lists of files (in a single folder or recursively) and filter files using any part of its metadata (file name, file size, last modification date, etc.).

1.1 DELIMITED TEXT FILES

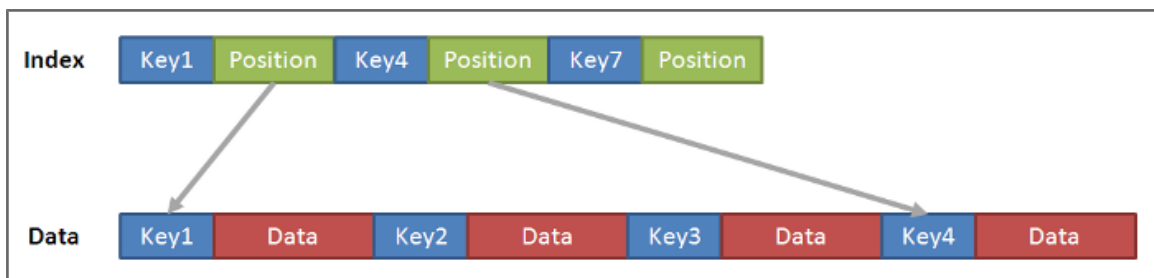
Delimited text files store plain text and each line has values separated by a delimiter, such as tab, space, comma, etc.

1.2 SEQUENCEFILES

Sequence files are binary record-oriented files, where each record has a serialized key and a serialized value.

1.3 MAPFILES

A map is a directory containing two sequence files. The data file (/data) is identical to the sequence file and contains the data stored as binary key/value pairs. The index file (/index), which contains a key/value map with seek positions inside the data file to quickly access the data.



Map file format

1.4 AVRO FILES

Avro data files are self-describing, containing the full schema for the data in the file. An Avro schema is defined using JSON. The schema allows you to define two types of data:

- primitive data types: string, integer, long, float, double, byte, null and boolean.
- complex type definitions: a record, an array, an enum, a map, a union or a fixed type.

```
{ "namespace": "example.avro",  
  "type": "record",  
  "name": "User",  
  "fields":  
  [  
    { "name": "name", "type": "string" },  
    { "name": "favorite_number", "type": [ "int", "null" ] },  
    { "name": "favorite_color", "type": [ "string", "null" ] }  
  ]  
}
```

Avro schema

1.5 PARQUET FILES

Parquet is a column-oriented data store of the Hadoop ecosystem. It provides data compression on a per-column level and encoding schemas.

The data are described by a schema that starts with the word *Message* and contains a group of fields. Each field is defined by a *repetition* (required, optional, or repeated), a *type* and a *name*.

```
Message Customer {  
  required int32 id;  
  required binary firstname (UTF8);  
  required binary lastname (UTF8);  
}
```

Parquet schema

Primitives types in parquet are boolean, int32, int64, int96, float, double, binary and fixed_len_byte_array. There are no String types but there are logical types which allows interpreting binaries as a String, JSON or other types.

Complex types are defined by a group type, which adds a layer of nesting.

2 USAGE

The Distributed File System Custom Wrapper distribution consists of:

- /conf: A folder containing a sample core-site.xml file with properties you might need commented out.
- /dist:
 - denodo-hdfs-customwrapper-\${version}.jar. The custom wrapper.
 - denodo-hdfs-customwrapper-\${version}-jar-with-dependencies.jar. The custom wrapper plus its dependencies. This is the wrapper we recommend to use, as it is easier to install in VDP.
- /doc: A documentation folder containing this user manual
- /lib: All the dependencies required by this wrapper in case you need to use the denodo-hdfs-customwrapper-\${version}.jar

2.1 IMPORTING THE CUSTOM WRAPPER INTO VDP

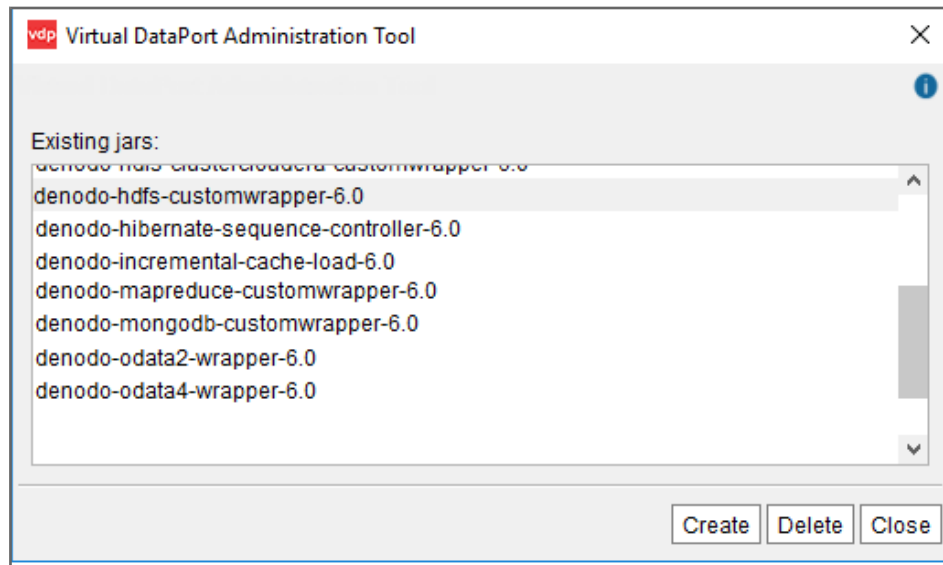
In order to use the Distributed File System Custom Wrapper in VDP, we must configure the Admin Tool to import the extension.

From the Distributed File System Custom Wrapper distribution, we will select the denodo-hdfs-customwrapper-\${version}-**jar-with-dependencies.jar** file and upload it to VDP.

Important

As this wrapper, (the **jar-with-dependencies** version), contains the Hadoop client libraries themselves, increasing the JVM's heap space for VDP Admin Tool is required to avoid a Java heap space when uploading the jar to VDP.

No other jars are required as this one will already contain all the required dependencies.



Distributed File System extension in VDP

2.2 CREATING A DISTRIBUTED FILE SYSTEM DATA SOURCE

Once the custom wrapper jar file has been uploaded to VDP using the Admin Tool, we can create new data sources for this custom wrapper --and their corresponding base views-- as usual.

Go to New → Data Source → Custom and specify one of the possible wrappers:

- `com.denodo.connect.hadoop.hdfs.wrapper.HDFSDelimitedTextFileWrapper`
- `com.denodo.connect.hadoop.hdfs.wrapper.HDFSSequenceFileWrapper`
- `com.denodo.connect.hadoop.hdfs.wrapper.HDFSMapFileWrapper`
- `com.denodo.connect.hadoop.hdfs.wrapper.HDFSAvroFileWrapper`
- `com.denodo.connect.hadoop.hdfs.wrapper.WebHDFSFileWrapper`
(deprecated)
- `com.denodo.connect.hadoop.hdfs.wrapper.HDFSParquetFileWrapper`
- `com.denodo.connect.hadoop.hdfs.wrapper.DFSListFilesWrapper`

Also check 'Select Jars' and choose the jar file of the custom wrapper.

The screenshot shows the Denodo Admin Tool interface for configuring a new data source. The 'Configuration' tab is selected, and the 'Metadata' sub-tab is active. The 'Name' field contains 'avro_ds'. The 'Class name' dropdown is set to 'com.denodo.connect.hadoop.hdfs.wrapper.HDFSAvroFileWrapper'. The 'Class path (optional)' field is empty, with a 'Browse' button to its right. The 'Select Jars' checkbox is checked. Below it, a list of jars is shown: 'denodo-hdfs-customwrapper-6.0' (selected), 'denodo-mapreduce-customwrapper-6.0', 'denodo-mongodb-customwrapper-6.0', 'denodo-odata2-wrapper-6.0', and 'denodo-odata4-wrapper-6.0'. The interface also includes buttons for 'Configuration', 'VQL', 'Save', 'Create base view', 'Export', 'Drop', and an information icon.

Distributed File System Data Source

2.3 CREATING A BASE VIEW

Once the custom wrapper has been registered, we will be asked by VDP to create a base view for it.

2.3.1 HDFSDelimitedTextFileWrapper

Custom wrapper for reading delimited text files. Its base views need the following parameters:

- **File system URI:** A URI whose scheme and authority identify the file system.

- **HDFS:** `hdfs://<ip>:<port>`.
- **Amazon S3:** `s3a://@<bucket>`. For configuring the credentials see **Amazon S3** section.
- **Azure Data Lake Store:**
`adl://<account name>.azuredatalakestore.net/`
For configuring the credentials see **Azure Data Lake Store** section.
- **Azure Blob Storage:**
`wasb://<container>@<account>.blob.core.windows.net`
For configuring the credentials see **Azure Blob Storage** section.
- **Google Cloud Storage:**
`gs://<bucket>`
For configuring the credentials see **Google Storage** section.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables @, \, ^, {, }, you have to escape these characters with \.

E.g if the URI contains @ you have to enter \@.

- **Path:** input path for the delimited file or the directory containing the files.
- **File name pattern:** If you want this wrapper to only obtain data from some of the files of the directory, you can enter a regular expression that matches the names of these files.
For example, if you want the base view to return the data of all the files with the extension csv set the File name pattern to `(.*)\\\.csv`, (notice that the regular expression is escaped as explained in the note below). Optional.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables @, \, ^, {, }, you have to escape these characters with \.

E.g if the File name pattern contains \ you have to enter \\.

- **Delete after reading:** Requests that the file or directory denoted by the path be deleted when the wrapper terminates.
- **Custom core-site.xml file:** configuration file that overrides the default core parameters. Optional.
- **Custom hdfs-site xml file:** configuration file that overrides the default HDFS parameters. Optional.
- **Separator:** delimiter between the values of a row. Default is the comma (,) and

cannot be a line break (`\n` or `\r`). Optional.

Some “invisible” characters have to be entered in a special way:

Character	Meaning
<code>\t</code>	Tab
<code>\f</code>	Formfeed

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables `@`, `\`, `^`, `{`, `}`, you have to escape these characters with `\`.

E.g if the separator is the tab character `\t` you have to enter `\\t`.

- **Quote:** Character used to encapsulate values containing special characters. Default is quote (`"`). Optional.
- **Comment marker:** Character marking the start of a line comment. Comments are disabled by default. Optional.
- **Escape:** Escape character. Escapes are disabled by default. Optional.
- **Null value:** String used to represent a null value. Default is: none; nulls are not distinguished from empty strings. Optional.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables `@`, `\`, `^`, `{`, `}`, you have to escape these characters with `\`.

E.g if the null value is `\N` you have to enter `\\N`.

- **Ignore spaces:** Whether spaces around values are ignored. False by default.
- **Header:** If selected, the wrapper considers that the first line contains the names of the fields in this file. These names will be the fields' names of the base views created from this wrapper. True by default.
- **Ignore matching errors:** Whether the wrapper will ignore the lines of the file that do not have the expected number of columns. True by default.
If you clear this check box, the wrapper will return an error if there is a row that does not have the expected structure. When you select this check box, you can

check if the wrapper has ignored any row in a query in the execution trace, in the attribute "Number of invalid rows".

File system URI	<input type="text" value="hdfs://quickstart.cloudera:8020"/>		
Path	<input type="text" value="/user/cloudera/df/SearchLog.tsv"/>		
File name pattern	<input type="text" value="(.*)\\.tsv"/>		
	<input type="checkbox"/> Delete after reading		
Custom core-site.xml file	<input type="text" value="None"/>	<input type="button" value="Configure"/>	
Custom hdfs-site.xml file	<input type="text" value="None"/>	<input type="button" value="Configure"/>	
Separator	<input type="text" value="\\t"/>		
Quote	<input type="text"/>		
Comment marker	<input type="text"/>		
Escape	<input type="text"/>		
Null value	<input type="text"/>		
	<input type="checkbox"/> Ignore spaces		
	<input checked="" type="checkbox"/> Header		
	<input type="checkbox"/> Ignore matching errors		
	<input type="checkbox"/> Kerberos enabled		
Kerberos principal name	<input type="text"/>		
Kerberos keytab file	<input type="text" value="None"/>	<input type="button" value="Configure"/>	
Kerberos password	<input type="text"/>		
Kerberos Distribution Center	<input type="text"/>		

HDFSDeimitedTextFileWrapper base view edition

View schema:	Field Name	Field Type
	playerid	text
	yearid	text
	stint	text
	teamid	text
	lgid	text
	pos	text
	g	text
	gs_0	text
	innouts	text
	po	text
	a	text
	e	text
	dp	text
	pb	text
	wp	text
	sb	text
	cs	text
	zr	text

View schema

The execution of the wrapper returns the values contained in the file or group of files, if the Path input parameter denotes a directory.

Total rows received: 167938 (shown 150)							
playerid	yearid	stint	teamid	lgid	pos	g	g
abercda01	1871	1	TRO	NA	SS	1	
addybo01	1871	1	RC1	NA	2B	22	
addybo01	1871	1	RC1	NA	SS	3	
allisar01	1871	1	CL1	NA	2B	2	
allisar01	1871	1	CL1	NA	OF	29	
allisdo01	1871	1	WS3	NA	C	27	
ansonca01	1871	1	RC1	NA	1B	1	
ansonca01	1871	1	RC1	NA	2B	2	
ansonca01	1871	1	RC1	NA	3B	20	
ansonca01	1871	1	RC1	NA	C	5	
ansonca01	1871	1	RC1	NA	OF	1	

View results

2.3.2 HDFSSequenceFileWrapper

Custom wrapper for reading sequence files. Its base views need the following parameters:

- **File system URI:** A URI whose scheme and authority identify the file system.
 - **HDFS:** `hdfs://<ip>:<port>`.
 - **Amazon S3:** `s3a://@<bucket>`. For configuring the credentials see **Amazon S3** section.
 - **Azure Data Lake Store:**
`adl://<account name>.azuredatalakestore.net/`
For configuring the credentials see **Azure Data Lake Store** section.
 - **Azure Blob Storage:**
`wasb://<container>@<account>.blob.core.windows.net`
For configuring the credentials see **Azure Blob Storage** section.
 - **Google Cloud Storage:**
`gs://<bucket>`
For configuring the credentials see **Google Storage** section.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables @, \, ^, {, }, you have to escape these characters with \.

E.g if the URI contains @ you have to enter \@.

- **Path:** input path for the sequence file or the directory containing the files.
- **File name pattern:** If you want this wrapper to only obtain data from some of the files of the directory, you can enter a regular expression that matches the names of these files.
For example, if you want the base view to return the data of all the files with the extension seq set the File name pattern to `(.*)\\.seq`, (notice that the regular expression is escaped as explained in the note below). Optional.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables @, \, ^, {, }, you have to escape these characters with \.

E.g if the File name pattern contains \ you have to enter \\.

- **Delete after reading:** Requests that the file or directory denoted by the path be deleted when the wrapper terminates.
- **Custom core-site.xml file:** configuration file that overrides the default core parameters. Optional.
- **Custom hdfs-site.xml file:** configuration file that overrides the default HDFS parameters. Optional.
- **Key class:** key class name implementing org.apache.hadoop.io.Writable interface.
- **Value class:** value class name implementing org.apache.hadoop.io.Writable interface.

File system URI	<input type="text" value="h3n://AKIAJXVIEYD2Q74CGMTA:SALdz9RIETUxCOlvd4eVOkgGT5FnkmGvt1"/>		
Path	<input type="text" value="/sequencefil/sequence.seq"/>		
File name pattern	<input type="text"/>		
	<input type="checkbox"/> Delete after reading		
Custom core-site.xml file	<input type="text" value="None"/>	<input type="button" value="Configure"/>	
Custom hdfs-site.xml file	<input type="text" value="None"/>	<input type="button" value="Configure"/>	
Key class	<input type="text" value="org.apache.hadoop.io.IntWritable"/>		
Value class	<input type="text" value="org.apache.hadoop.io.Text"/>		
	<input type="checkbox"/> Kerberos enabled		
Kerberos principal name	<input type="text"/>		
Kerberos keytab file	<input type="text" value="None"/>	<input type="button" value="Configure"/>	
Kerberos password	<input type="text"/>		
Kerberos Distribution Center	<input type="text"/>		

HDFSSequenceFileWrapper base view edition

View schema:	Field Name	Field Type
	key	int
	value	text

View schema

The execution of the wrapper returns the key/value pairs contained in the file or group of files, if the Path input parameter denotes a directory.

Total rows received: 100 (shown 100)	
△ ▽	
key	value
100	One, two, buckle my shoe
99	Three, four, shut the door
98	Five, six, pick up sticks
97	Seven, eight, lay them straight
96	Nine, ten, a big fat hen
95	One, two, buckle my shoe
94	Three, four, shut the door
93	Five, six, pick up sticks
92	Seven, eight, lay them straight
91	Nine, ten, a big fat hen
90	One, two, buckle my shoe
89	Three, four, shut the door

View results

2.3.3 HDFSMapFileWrapper

Custom wrapper for reading map files. Its base views need the following parameters:

- **File system URI:** A URI whose scheme and authority identify the file system.
 - **HDFS:** `hdfs://<ip>:<port>`.
 - **Amazon S3:** `s3a://@<bucket>`. For configuring the credentials see **Amazon S3** section.
 - **Azure Data Lake Store:**
`adl://<account name>.azuredatalakestore.net/`
For configuring the credentials see **Azure Data Lake Store** section.
 - **Azure Blob Storage:**
`wasb://<container>@<account>.blob.core.windows.net`
For configuring the credentials see **Azure Blob Storage** section.
 - **Google Cloud Storage:**
`gs://<bucket>`
For configuring the credentials see **Google Storage** section.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables @, \, ^, {, }, you have to escape these characters with \.

E.g if the URI contains @ you have to enter \@.

- **Path:** input path for the directory containing the map file. Also the path to the index or data file could be specified. When using **Amazon S3**, a flat file system where there is no folder concept, **the path to the index or data should be**

used.

- **File name pattern:** If you want this wrapper to only obtain data from some of the files of the directory, you can enter a regular expression that matches the names of these files.
For example, if you want the base view to return the data of all the files with the extension whatever set the File name pattern to `(.*)\\\.whatever`, (notice that the regular expression is escaped as explained in the note below). Optional.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables `@`, `\`, `^`, `{`, `}`, you have to escape these characters with `\`.

E.g if the File name pattern contains `\` you have to enter `\\`.

- **Delete after reading:** Requests that the file or directory denoted by the path be deleted when the wrapper terminates.
- **Custom core-site.xml file:** configuration file that overrides the default core parameters. Optional.
- **Custom hdfs-site xml file:** configuration file that overrides the default HDFS parameters. Optional.
- **Key class:** key class name implementing the `org.apache.hadoop.io.WritableComparable` interface. `WritableComparable` is used because records are sorted in **key order**.
- **Value class:** value class name implementing the `org.apache.hadoop.io.Writable` interface.

File system URI	<input type="text" value="hdfs://quickstart.cloudera:8020"/>		
Path	<input type="text" value="/user/cloudera/MAP"/>		
File name pattern	<input type="text"/>		
	<input type="checkbox"/> Delete after reading		
Custom core-site.xml file	<input type="text" value="None"/>	<input type="button" value="▼"/>	<input type="button" value="Configure"/>
Custom hdfs-site.xml file	<input type="text" value="None"/>	<input type="button" value="▼"/>	<input type="button" value="Configure"/>
Key class	<input type="text" value="org.apache.hadoop.io.IntWritable"/>		
Value class	<input type="text" value="org.apache.hadoop.io.Text"/>		
	<input type="checkbox"/> Kerberos enabled		
Kerberos principal name	<input type="text"/>		
Kerberos keytab file	<input type="text" value="None"/>	<input type="button" value="▼"/>	<input type="button" value="Configure"/>
Kerberos password	<input type="text"/>		
Kerberos Distribution Center	<input type="text"/>		

HDFSMapFileWrapper base view edition

View schema:	Field Name	Field Type
	key	int
	value	text

View schema

The execution of the wrapper returns the key/value pairs contained in the file or group of files, if the Path input parameter denotes a directory.

Total rows received: 1024 (shown 150)	
△ ▽	
key	value
1	One, two, buckle my shoe
2	Three, four, shut the door
3	Five, six, pick up sticks
4	Seven, eight, lay them straight
5	Nine, ten, a big fat hen
6	One, two, buckle my shoe
7	Three, four, shut the door
8	Five, six, pick up sticks
9	Seven, eight, lay them straight
10	Nine, ten, a big fat hen
11	One, two, buckle my shoe
12	Three, four, shut the door

View results

2.3.4 HDFS Avro File Wrapper

Custom wrapper for reading Avro files.

Important

We recommend not to use the HDFS Avro File Wrapper to directly access Avro files, as this is an internal serialization system mainly meant for use by applications running on the Hadoop cluster. Instead, we recommend to use an abstraction layer on top of those files like e.g. Hive, Impala, Spark...

Its base views need the following parameters:

- **File system URI:** A URI whose scheme and authority identify the file system.
 - **HDFS:** `hdfs://<ip>:<port>`.
 - **Amazon S3:** `s3a://<bucket>`. For configuring the credentials see **Amazon S3** section.
 - **Azure Data Lake Store:**
`adl://<account name>.azuredatalakestore.net/`
For configuring the credentials see **Azure Data Lake Store** section.
 - **Azure Blob Storage:**
`wasb://<container>@<account>.blob.core.windows.net`
For configuring the credentials see **Azure Blob Storage** section.
 - **Google Cloud Storage:**
`gs://<bucket>`
For configuring the credentials see **Google Storage** section.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables @, \, ^, {, }, you have to escape these characters with \.

E.g if the URI contains @ you have to enter \@.

- **File name pattern:** If you want this wrapper to only obtain data from some of the files of the directory, you can enter a regular expression that matches the names of these files.
For example, if you want the base view to return the data of all the files with the extension avro set the File name pattern to (.*)\.avro, (notice that the regular expression is escaped as explained in the note below). Optional.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables @, \, ^, {, }, you have to escape these characters with \.

E.g if the File name pattern contains \ you have to enter \.

- **Delete after reading:** Requests that the file denoted by the path be deleted when the wrapper terminates.
- **Custom core-site.xml file:** configuration file that overrides the default core parameters. Optional.
- **Custom hdfs-site.xml file:** configuration file that overrides the default HDFS parameters. Optional.

There is also two parameters that are **mutually exclusive**:

- **Avro schema path:** input path for the Avro schema file or
- **Avro schema JSON:** JSON of the Avro schema.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables @, \, ^, {, } in the **Avro schema JSON** parameter, you have to escape these characters with \. For example:

```
\{
  "type": "map",
  "values": \{
    "type": "record",
    "name": "ATM",
    "fields": [
      \{ "name": "serial_no", "type": "string" \},
      \{ "name": "location", "type": "string" \}
    ]
  }
}
```

File system URI	<input type="text" value="hdfs://quickstart.cloudera:8020"/>		
Avro schema path	<input type="text" value="/user/cloudera/avro/RecordWithAllTypes.avsc"/>		
Avro schema JSON	<input type="text"/>		
File name pattern	<input type="text" value="(*.*)\avro"/>		
	<input type="checkbox"/> Delete after reading		
Custom core-site.xml file	<input type="text" value="None"/>	<input type="button" value="v"/>	<input type="button" value="Configure"/>
Custom hdfs-site.xml file	<input type="text" value="None"/>	<input type="button" value="v"/>	<input type="button" value="Configure"/>
	<input type="checkbox"/> Kerberos enabled		
Kerberos principal name	<input type="text"/>		
Kerberos keytab file	<input type="text" value="None"/>	<input type="button" value="v"/>	<input type="button" value="Configure"/>
Kerberos password	<input type="text"/>		
Kerberos Distribution Center	<input type="text"/>		

HDFSAvroFileWrapper base view edition

```
{
  "type" : "record",
  "name" : "Doc",
  "doc" : "adoc",
  "fields" : [ {
    "name" : "id",
    "type" : "string"
  }, {
    "name" : "user_friends_count",
    "type" : [ "int", "null" ]
  }, {
    "name" : "user_location",
    "type" : [ "string", "null" ]
  }, {
    "name" : "user_description",
    "type" : [ "string", "null" ]
  }, {
    "name" : "user_statuses_count",
    "type" : [ "int", "null" ]
  }, {
    "name" : "user_followers_count",
    "type" : [ "int", "null" ]
  }, {
    "name" : "user_name",
    "type" : [ "string", "null" ]
  }, {
    "name" : "user_screen_name",
    "type" : [ "string", "null" ]
  }, {
    "name" : "created_at",
    "type" : [ "string", "null" ]
  }, {
    "name" : "text",
    "type" : [ "string", "null" ]
  }, {
    "name" : "retweet_count",
    "type" : [ "int", "null" ]
  }, {
    "name" : "retweeted",
    "type" : [ "boolean", "null" ]
  }, {
    "name" : "in_reply_to_user_id",
    "type" : [ "long", "null" ]
  }, {
    "name" : "source",
    "type" : [ "string", "null" ]
  }, {
    "name" : "in_reply_to_status_id",
    "type" : [ "long", "null" ]
  }, {
    "name" : "media_url_https",
    "type" : [ "string", "null" ]
  }, {
    "name" : "expanded_url",
```

```
"type" : [ "string", "null" ]
} ] }
```

Content of the /user/cloudera/schema.avsc file

View schema:	Field Name	Field Type
	avrofilepath	text
	doc	avro_ds_doc
	id	text
	user_friends_count	int
	user_location	text
	user_description	text
	user_statuses_count	int
	user_followers_count	int
	user_name	text
	user_screen_name	text
	created_at	text
	text	text
	retweet_count	int
	retweeted	boolean
	in_reply_to_user_id	long
	source	text
	in_reply_to_status_id	long
	media_url_https	text
	expanded_url	text

View schema

The execution of the view returns the values contained in the Avro file specified in the WHERE clause of the VQL sentence:

```
SELECT * FROM avro_ds_file
WHERE avrofilepath = '/user/cloudera/file.avro'
```

Total rows received: 2104 (shown 150)

avrofilepath	doc
/user/cloudera/file.avro	[Register]...
/user/cloudera/file.avro	[Register]...
/user/cloudera/file.avro	[Register]...



RESU... -> doc

id	user_frien...	user_loca...	user_des...	user_stat...	user_follo...	user_name	user_scre...	created_at	text	re
10000	10000	location1...		1	1	fake user...	fake_user...	1985-05-...	tweet text ...	0

View results

After applying a flattening operation results are as follows.

Total rows received: 2104 (shown 150)

avrofilepath	id	user_frien...	user_locati...	user_desc...	user_statu...	user_follo...	user_name	user_scre...	cre
/user/cloud...	10000	10000	location10...		1	1	fake user1...	fake_user1...	198
/user/cloud...	10001	10001	location10...		1	1	fake user1...	fake_user1...	198

Flattened results

2.3.4.1 Field Projection

The recommended way for dealing with **projections** in HDFSAvroFileWrapper is by means of the JSON schema parameters:

- Avro schema path or
- Avro schema JSON

By giving to the wrapper a JSON schema containing exclusively the fields we are interested in, the reader used by the HDFSAvroFileWrapper will return to VDP only these fields, making the select operation faster.

If we configure the parameter Avro schema JSON with only some of the fields of the /user/cloudera/schema.avsc file used in the previous example, like in the example below (notice the escaped characters):

```
\{
  "type" : "record",
  "name" : "Doc",
  "doc" : "adoc",
  "fields" : [ \{
    "name" : "id",
    "type" : "string"
  \}, \{
    "name" : "user_friends_count",
    "type" : [ "int", "null" ]
  \}, \{
    "name" : "user_location",
    "type" : [ "string", "null" ]
  \}, \{
    "name" : "user_followers_count",
    "type" : [ "int", "null" ]
  \}, \{
    "name" : "user_name",
    "type" : [ "string", "null" ]
  \}, \{
    "name" : "created_at",
    "type" : [ "string", "null" ]
  \} ]
\}
```

Schema with the selected fields

the base view in VDP will contain a subset of the previous base view of the example: the ones matching the new JSON schema provided to the wrapper.

View schema:	Field Name	Field Type
	avroFilepath	text
Doc	avro_ds_Doc	
	id	text
	user_friends_count	int
	user_location	text
	user_followers_count	int
	user_name	text
	created_at	text

Base view with the selected fields

RESU... -> Doc					
id	user_friends_count	user_location	user_followers_count	user_name	created_at
10000	10000	location10000	1	fake user10000	1985-05-11T10:09:19Z

View results with the selected fields

2.3.5 WebHDFSFileWrapper

Warning

WebHDFSFileWrapper is **deprecated**.

- For XML, JSON and Delimited files the best alternative is using the **VDP standard data sources**, using the HTTP Client in its Data route parameter. These data sources offers a better solution for HTTP/HTTPs access as they include proxy access, SPNEGO authentication, OAuth2 etc.
- For Avro, Sequence, Map and Parquet files the best alternative is using the specific custom wrapper type: **HDFSAvroFileWrapper**, **HDFSSequenceFileWrapper**, **HDFSMapFileWrapper** or **HDFSParquetFileWrapper** with webhdfs scheme in their File system URI parameter. And placing their credentials in the xml configuration files.

Custom wrapper for reading delimited text files using the **WebHDFS**.

2.3.5.1 About WebHDFS

WebHDFS provides HTTP REST access to HDFS. It supports all HDFS user operations including reading files, writing to files, making directories, changing permissions and renaming.

The advantage of WebHDFS are:

- **Version-independent** REST-based protocol which means that can be read and written to/from Hadoop clusters no matter their version.
- Read and write data in a cluster behind a firewall. A proxy WebHDFS (for example: HttpFS) could be use, it acts as a gateway and is the only system that is allowed to send and receive data through the firewall.
The only difference between using or not the proxy will be in the host:port pair where the HTTP requests are issued:
 - Default port for WebHDFS is 50070.
 - Default port for HttpFS is 14000.

2.3.5.2 Custom wrapper

The base views created from the WebHDFSFileWrapper need the following parameters:

- **Host IP:** IP or <bucket>.s3.amazonaws.com for Amazon S3.
- **Host port:** HTTP port. Default port for WebHDFS is 50070. For HttpFS is 14000. For Amazon S3 is 80.

- **User:** The name of the the authenticated user when security is off. If is not set, the server may either set the authenticated user to a default web user, if there is any, or return an error response.
When using Amazon S3 <id>:<secret> should be indicated.
- **Path:** input path for the delimited file.
- **Separator:** delimiter between values. Default is the comma.
- **Quote:** Character used to encapsulate values containing special characters. Default is quote.
- **Comment marker:** Character marking the start of a line comment. Comments are disable by default.
- **Escape:** Escape character. Escapes are disabled by default.
- **Null value:** String used to represent a null value. Default is: none; nulls are not distinguished from empty strings. Optional.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables @, \, ^, {, }, you have to escape these characters with \.

E.g if the null value is \N you have to enter \\N.

- **Ignore spaces:** Whether spaces around values are ignored. False by default.
- **Header:** Whether the file has a header or not. True by default.
- **Delete after reading:** Requests that the file or directory denoted by the path be deleted when the wrapper terminates.

Host IP	<input type="text" value="melkus.denodo.com"/>
Host port	<input type="text" value="50070"/>
User	<input type="text"/>
Path	<input type="text" value="/user/cloudera/csv/Master.csv"/>
Separator	<input type="text" value=","/>
Quote	<input type="text"/>
Comment marker	<input type="text"/>
Escape	<input type="text"/>
Null value	<input type="text"/>
	<input type="checkbox"/> Ignore spaces <input checked="" type="checkbox"/> Header <input type="checkbox"/> Delete after reading

WebHDFSFileWrapper base view edition

View schema:

Field Name	Field Type
playerid	text
birthyear	text
birthmonth	text
birthday	text
birthcountry	text
birthstate	text
birthcity	text
deathyear	text
deathmonth	text
deathday	text
deathcountry	text
deathstate	text
deathcity	text
namefirst	text
namelast	text
namegiven	text
weight	text
height	text
bats	text
throws	text

View schema

The execution of the wrapper returns the values contained in the file.

Total rows received: 18589 (shown 150)

playerid	birthyear	birthmonth	birthday	birthcountry	birthstate	birthcity	deathyear	deathmonth	deathday	de
aardsda01	1981	12	27	USA	CO	Denver				
aaronha01	1934	2	5	USA	AL	Mobile				
aaronto01	1939	8	5	USA	AL	Mobile	1984	8	16	US
aasedo01	1954	9	8	USA	CA	Orange				
abadan01	1972	8	25	USA	FL	Palm Beach				
abadfe01	1985	12	17	D.R.	La Romana	La Romana				
abadijo01	1854	11	4	USA	PA	Philadelphia	1905	5	17	US
abbated01	1877	4	15	USA	PA	Latrobe	1957	1	6	US
abbeybe01	1869	11	11	USA	VT	Essex	1962	6	11	US
abbeych01	1866	10	14	USA	NE	Falls City	1926	4	27	US
abbetde01	1862	2	16	USA	OH	Portage	1920	2	12	US

View results

2.3.6 HDFSParquetFileWrapper

Custom wrapper for reading Parquet files.

Important

We recommend not to use the HDFSParquetFileWrapper to directly access Parquet files, as this is an internal columnar data representation mainly meant for use by applications running on the Hadoop cluster. Instead, we recommend to use an abstraction layer on top of those files like e.g. Hive, Impala, Spark...

Its base views need the following parameters:

- **File system URI:** A URI whose scheme and authority identify the file system.
 - **HDFS:** hdfs://<ip>:<port>.
 - **Amazon S3:** s3a://@<bucket>. For configuring the credentials see **Amazon S3** section.
 - **Azure Data Lake Store:**
adl://<account name>.azuredatalakestore.net/
For configuring the credentials see **Azure Data Lake Store** section.
 - **Azure Blob Storage:**
wasb://<container>@<account>.blob.core.windows.net
For configuring the credentials see **Azure Blob Storage** section.
 - **Google Cloud Storage:**
gs://<bucket>
For configuring the credentials see **Google Storage** section.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables @, \, ^, {, }, you have to escape these characters with \.

E.g if the URI contains @ you have to enter \@.

- **Parquet File Path:** path of the file that we want to read.
- **File name pattern:** If you want this wrapper to only obtain data from some of the files of the directory, you can enter a regular expression that matches the names of these files.
For example, if you want the base view to return the data of all the files with the extension parquet set the File name pattern to (.*)\\.parquet, (notice that the regular expression is escaped as explained in the note below). Optional.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables @, \, ^, {, }, you have to escape these characters with \.

E.g if the File name pattern contains \ you have to enter \\\.

- **Custom core-site.xml file:** configuration file that overrides the default core parameters. Optional.
- **Custom hdfs-site.xml file:** configuration file that overrides the default HDFS parameters. Optional.

File system URI	<input type="text" value="hdfs://quickstart.cloudera:8020"/>		
Parquet File path	<input type="text" value="/user/cloudera/parquet/complex_types.snappy.parquet"/>		
File name pattern	<input type="text"/>		
Custom core-site.xml file	<input type="text" value="None"/>	<input type="button" value="Configure"/>	
Custom hdfs-site.xml file	<input type="text" value="None"/>	<input type="button" value="Configure"/>	
	<input type="checkbox"/> Kerberos enabled		
Kerberos principal name	<input type="text"/>		
Kerberos keytab file	<input type="text" value="None"/>	<input type="button" value="Configure"/>	
Kerberos password	<input type="text"/>		
Kerberos Distribution Center	<input type="text"/>		

HDFSParquetWrapper base view edition

View schema Metadata

View name: parquet

PK	Field Name	Field Type	Description
<input type="checkbox"/>	statecode	text	
<input type="checkbox"/>	countrycode	text	
<input type="checkbox"/>	sitenum	text	
<input type="checkbox"/>	paramcode	text	
<input type="checkbox"/>	poc	text	
<input type="checkbox"/>	latitude	text	
<input type="checkbox"/>	longitude	text	
<input type="checkbox"/>	datum	text	
<input type="checkbox"/>	param	text	
<input type="checkbox"/>	datelocal	text	
<input type="checkbox"/>	timelocal	text	
<input type="checkbox"/>	dategmt	text	
<input type="checkbox"/>	timegmt	text	
<input type="checkbox"/>	degrees	double	
<input type="checkbox"/>	uom	text	
<input type="checkbox"/>	mdl	text	
<input type="checkbox"/>	uncert	text	
<input type="checkbox"/>	qual	text	

Set selected as PK

View schema

The execution of the wrapper returns the values contained in the file.

Execute Query Results

Results Execution Trace Stop Refresh Save Query: SELECT * FROM parquet CONTEXT ('i18n'=us_pst, 'cache_wait_for_load'=true) TRACE

Total rows received: 7027695 (shown 150)

statecode	countrycode	sitenum	paramcode	poc	latitude	longitude	datum	param	datelocal	timelocal	dategmt	timegmt	degrees
"01"	"003"	"0010"	"44201"	1	30.497478	-87.880258	"NAD83"	"Ozone"	"2016-03-01"	"15:00"	"2016-03-01"	"21:00"	0.041
"01"	"003"	"0010"	"44201"	1	30.497478	-87.880258	"NAD83"	"Ozone"	"2016-03-01"	"16:00"	"2016-03-01"	"22:00"	0.041
"01"	"003"	"0010"	"44201"	1	30.497478	-87.880258	"NAD83"	"Ozone"	"2016-03-01"	"17:00"	"2016-03-01"	"23:00"	0.042
"01"	"003"	"0010"	"44201"	1	30.497478	-87.880258	"NAD83"	"Ozone"	"2016-03-01"	"18:00"	"2016-03-02"	"00:00"	0.041
"01"	"003"	"0010"	"44201"	1	30.497478	-87.880258	"NAD83"	"Ozone"	"2016-03-01"	"19:00"	"2016-03-02"	"01:00"	0.038

View results

2.3.7 DFSListFilesWrapper

Custom wrapper to retrieve file information from a distributed file system.

Its base views need the following parameters:

- **File system URI:** A URI whose scheme and authority identify the file system.
 - **HDFS:** hdfs://<ip>:<port>.
 - **Amazon S3:** s3a://@<bucket>. For configuring the credentials see **Amazon S3** section.
 - **Azure Data Lake Store:** adl://<account name>.azuredatalakestore.net/ For configuring the credentials see **Azure Data Lake Store** section.
 - **Azure Blob Storage:** wasb://<container>@<account>.blob.core.windows.net For configuring the credentials see **Azure Blob Storage** section.

- **Google Cloud Storage:**
gs://<bucket>
For configuring the credentials see **Google Cloud Storage** section.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables @, \, ^, {, }, you have to escape these characters with \.

E.g if the URI contains @ you have to enter \@.

- **Custom hdfs-site xml file:** configuration file that overrides the default HDFS parameters. Optional.

Enter values for the following wrapper parameters:	
File system URI	hdfs://quickstart.cloudera:8020
Custom hdfs-site.xml file	None
	<input checked="" type="checkbox"/> Kerberos enabled
Kerberos principal name	cloudera-scm/admin\@CLOUDERA
Kerberos keytab file	None
Kerberos password	• • • • • • • •
Kerberos Distribution Center	quickstart.cloudera

DFSListFilesFilesWrapper base view edition

The entry point for querying the wrapper is the parameter **parentfolder**. The wrapper will list the files that are located in this supplied directory. It is possible to do this in a recursive way, retrieving also the contents of the subfolders, setting the parameter **recursive** to true.

Current sentence:

```
SELECT * FROM listing_dfs WHERE parentfolder = '/user/cloudera' and recursive = true CONTEXT ('il8n'='us_pst', 'cache_wait_for_load'='true')
```

☐ Do not use cache ☐ Invalidate existing results ☒ Limit rows 150 ☒ Execute with TRACE
☐ Store results in cache ☐ Do not use swap ☐ Stop query when the limit is reached ☐ Open results in new tab

Conditions +

parentfolder	=	'/user/cloudera'
recursive	=	true

Execution panel

The schema of the custom wrapper contains the following columns:



- **parentfolder**: the path of the parent directory. The wrapper will list all the files located in this directory
 - This parameter is **mandatory** in a SELECT operation
- **relativepath**: the location of the file respect of the parentfolder. Useful when executing a recursive query.
- **filename**: the name of the file or the folder, including the extension for files.
- **extension**: the extension of the file. It will be null if the file is a directory.
- **fullpath**: the full path of the file **with** the scheme information.
- **pathwithouscheme**: the full path of the file **without** the scheme information.
- **filetype**: either 'file' or 'directory'.
- **encrypted**: true if the file is encrypted, false otherwise.
- **datemodified**: the modification time of the file in milliseconds since January 1, 1970 UTC.
- **owner**: the owner of the file.
- **group**: the group associated with the file.
- **permissions**: the permissions of the file, using the symbolic notation (rwxr-xr-x).
- **size**: the size of the file in bytes. It will be null for folders.
- **recursive**: if false the search for files will be limited to the files that are direct children of the **parentfolder**. If true, the search will be done recursively, including subfolders of parentfolder.
 - This parameter is **mandatory** in a SELECT operation

View schema:	Field Name	Field Type
	parentfolder	text
	relativepath	text
	filename	text
	extension	text
	fullpath	text
	pathwithoutscheme	text
	filetype	text
	encrypted	boolean
	datemodified	timestamp
	owner	text
	group	text
	permissions	text
	size	int
	recursive	boolean

View schema

The following VQL sentence returns the files in the '/user/cloudera' hdfs directory, recursively:

```
SELECT * FROM listing_dfs
WHERE parentfolder = '/user/cloudera' AND recursive = true
```

Total rows received: 62 (shown 62)													
 													
parentfol...	relative...	filename	extensi...	fullpath	pathwithoutscheme	filetype	encrypted	datemod...	owner	group	permissi...	size	recursive
/user/do...		Trash	<null>	hdfs://quickstart.cloudera:802...	/user/cloudera/.Trash	directory	false	2018-09-...	cloudera	cloudera	rw-r-xr-x	<null>	true
/user/do...		MAP	<null>	hdfs://quickstart.cloudera:802...	/user/cloudera/MAP	directory	false	2018-07-...	cloudera	cloudera	rw-r-xr-x	<null>	true
/user/do...		SEQUEN...	<null>	hdfs://quickstart.cloudera:802...	/user/cloudera/SEQ...	directory	false	2018-09-...	cloudera	cloudera	rw-r-xr-x	<null>	true
/user/do...		avro	<null>	hdfs://quickstart.cloudera:802...	/user/cloudera/avro	directory	false	2018-09-...	cloudera	cloudera	rw-r-xr-x	<null>	true
/user/do...		df	<null>	hdfs://quickstart.cloudera:802...	/user/cloudera/df	directory	false	2018-09-...	cloudera	cloudera	rw-r-xr-x	<null>	true
/user/do...		emptydir	<null>	hdfs://quickstart.cloudera:802...	/user/cloudera/empt...	directory	false	2018-11-...	cloudera	cloudera	rw-r-xr-x	<null>	true
/user/do...		parquet	<null>	hdfs://quickstart.cloudera:802...	/user/cloudera/parqu...	directory	false	2018-09-...	cloudera	cloudera	rw-r-xr-x	<null>	true
/user/do...	/MAP	data		hdfs://quickstart.cloudera:802...	/user/cloudera/MAP/...	file	false	2018-07-...	cloudera	cloudera	rw-r--r--	47898	true
/user/do...	/MAP	index		hdfs://quickstart.cloudera:802...	/user/cloudera/MAP/...	file	false	2018-07-...	cloudera	cloudera	rw-r--r--	251	true
/user/do...	/SEQUE...	InputFile...	seq	hdfs://quickstart.cloudera:802...	/user/cloudera/SEQ...	file	false	2018-09-...	cloudera	cloudera	rw-r--r--	55293871	true
/user/do...	/SEQUE...	sequenc...	seq	hdfs://quickstart.cloudera:802...	/user/cloudera/SEQ...	file	false	2018-07-...	cloudera	cloudera	rw-r--r--	4788	true
/user/do...	/avro	Array.avro	avro	hdfs://quickstart.cloudera:802...	/user/cloudera/avro/...	file	false	2018-07-...	cloudera	cloudera	rw-r--r--	89	true

View results

We can filter our query a bit more and retrieve only those files that were modified after '2018-09-01':

```
SELECT * FROM listing_dfs
WHERE parentfolder = '/user/cloudera' AND recursive = true
AND datemodified > DATE '2018-09-01'
```

Total rows received: 26 (shown 26)

parentfolder	relativepath	filename	extension	fullpath	pathwithou...	filetype	encrypted	datemodified	owner	group	permissi...	size	recursi...
/user/cloud...		.Trash	<null>	hdfs://quick...	/user/cloud...	directory	false	2018-09-09 23:29:49.268	cloudera	cloudera	rw-r--r--	<null>	true
/user/cloud...		SEQUENCE	<null>	hdfs://quick...	/user/cloud...	directory	false	2018-09-06 09:13:38.443	cloudera	cloudera	rw-r--r--	<null>	true
/user/cloud...		avro	<null>	hdfs://quick...	/user/cloud...	directory	false	2018-09-11 04:55:57.695	cloudera	cloudera	rw-r--r--	<null>	true
/user/cloud...		df	<null>	hdfs://quick...	/user/cloud...	directory	false	2018-09-12 08:14:40.328	cloudera	cloudera	rw-r--r--	<null>	true
/user/cloud...		emptydir	<null>	hdfs://quick...	/user/cloud...	directory	false	2018-11-08 07:38:34.764	cloudera	cloudera	rw-r--r--	<null>	true
/user/cloud...		parquet	<null>	hdfs://quick...	/user/cloud...	directory	false	2018-09-05 08:56:09.683	cloudera	cloudera	rw-r--r--	<null>	true
/user/cloud...	/SEQUENCE	InputFile.seq	seq	hdfs://quick...	/user/cloud...	file	false	2018-09-06 04:22:23.956	cloudera	cloudera	rw-r--r--	552938...	true
/user/cloud...	/avro	sample-sta...	avro	hdfs://quick...	/user/cloud...	file	false	2018-09-11 04:55:57.651	cloudera	cloudera	rw-r--r--	249540	true
/user/cloud...	/avro	sample-sta...	avsc	hdfs://quick...	/user/cloud...	file	false	2018-09-11 04:55:57.658	cloudera	cloudera	rw-r--r--	1258	true
/user/cloud...	/df	2017	<null>	hdfs://quick...	/user/cloud...	directory	false	2018-09-07 04:26:18.623	cloudera	cloudera	rw-r--r--	<null>	true
/user/cloud...	/df	2018	<null>	hdfs://quick...	/user/cloud...	directory	false	2018-09-07 04:27:03.564	cloudera	cloudera	rw-r--r--	<null>	true
/user/cloud...	/df	SearchLog...	tsv	hdfs://quick...	/user/cloud...	file	false	2018-09-12 08:14:40.291	cloudera	cloudera	rw-r--r--	3158	true







[View results](#)

2.3.8 Extending capabilities with the DFSListFilesWrapper

The wrappers of this distribution that reads file formats like Parquet, Avro, Delimited Files, Sequence or Map, can increase their capabilities when combined with the DFSListFilesWrapper.

As all of these wrappers need an input path for the file or the directory that is going to be read, we can use the DFSListFilesWrapper for retrieving the file paths that we are interested in, according with some attribute value of their metadata, e.g. modification time.

For example, suppose that we want to retrieve the files in the /user/cloudera/df/awards directory that were modified in November.

Home	/ user / cloudera / df / awards						Trash
<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date	
<input type="checkbox"/>	 ↑		cloudera	cloudera	drwxr-xr-x	September 12, 2018 08:14 AM	
<input type="checkbox"/>	 .		cloudera	cloudera	drwxr-xr-x	November 15, 2018 05:37 AM	
<input type="checkbox"/>	 AwardsShareManagers.csv	16.3 KB	cloudera	cloudera	-rw-r--r--	November 15, 2018 05:37 AM	
<input type="checkbox"/>	 AwardsSharePlayers.csv	208.0 KB	cloudera	cloudera	-rw-r--r--	November 15, 2018 05:37 AM	
<input type="checkbox"/>	 AwardsManagers.csv	7.6 KB	cloudera	cloudera	-rw-r--r--	August 20, 2018 04:35 AM	
<input type="checkbox"/>	 AwardsPlayers.csv	239.1 KB	cloudera	cloudera	-rw-r--r--	August 20, 2018 04:35 AM	

The following steps explain how to configure this scenario:

1. Create a DFSListFilesWrapper base view that will list the files of the /user/cloudera/df/awards directory.
2. Create an HDFSDelimitedTextFileWrapper base view that will read the content of the csv files.

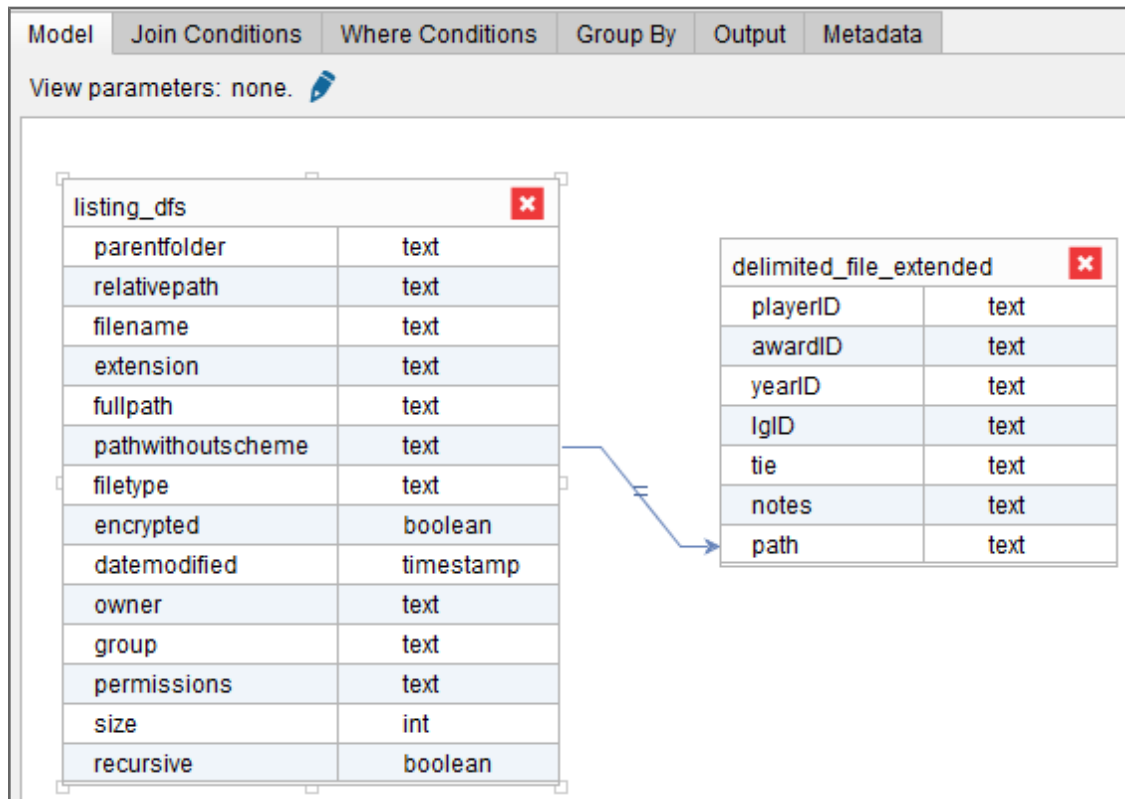
Parameterize the Path of the base view by adding an interpolation variable to its value, e.g. @path, (@ is the prefix that identifies a value parameter as an interpolation variable).

By using the variable @path, you do not have to provide the final path value when creating the base view. Instead, the values of the Path parameter will be provided at runtime by the DFSListFilesWrapper view through the join operation (configured in the next step).

File system URI	hdfs://quickstart.cloudera:8020	
Path	@path	
File name pattern		
	<input type="checkbox"/> Delete after reading	
Custom core-site.xml file	None	<input type="button" value="Configure"/>
Custom hdfs-site.xml file	None	<input type="button" value="Configure"/>
Separator		
Quote		
Comment marker		
Escape		
Null value		
	<input type="checkbox"/> Ignore spaces	
	<input checked="" type="checkbox"/> Header	
	<input checked="" type="checkbox"/> Ignore matching errors	
	<input type="checkbox"/> Kerberos enabled	
Kerberos principal name		
Kerberos keytab file	None	<input type="button" value="Configure"/>
Kerberos password		
Kerberos Distribution Center		

3. Create a derived view joining the two previously created views. The join condition will be:

DFSListFilesWrapper.pathwithoutscheme =
HDFSDelimitedTextFileWrapper.path



4. By executing the join view with these conditions:

```
SELECT * FROM join:view
WHERE recursive = true
      AND parentfolder = '/user/cloudera/df/awards'
      AND datemodified > DATE '2018-11-1'
```

we obtain data only from the delimited files that were modified in November.

Total rows received: 7106 (shown 7106)

parentfolder	r...	filename	pathwithout...	datemodified	rec...	playerID	awardID	yearID	lgID	tie	notes	path
/user/cloude...		AwardsShareManagers.csv	/user/cloude...	2018-11-15 05:3...	true	Mgr of the Y...	2014	NL	matlido01	12	150	/user/cloudera/d...
/user/cloude...		AwardsShareManagers.csv	/user/cloude...	2018-11-15 05:3...	true	Mgr of the Y...	2014	NL	roeniro01	3	150	/user/cloudera/d...
/user/cloude...		AwardsShareManagers.csv	/user/cloude...	2018-11-15 05:3...	true	Mgr of the Y...	2014	NL	blackbu02	1	150	/user/cloudera/d...
/user/cloude...		AwardsShareManagers.csv	/user/cloude...	2018-11-15 05:3...	true	Mgr of the Y...	2014	NL	collite99	1	150	/user/cloudera/d...
/user/cloude...		AwardsSharePlayers.csv	/user/cloude...	2018-11-15 05:3...	true	Cy Young	1956	ML	fordwh01	1	16	/user/cloudera/d...
/user/cloude...		AwardsSharePlayers.csv	/user/cloude...	2018-11-15 05:3...	true	Cy Young	1956	ML	maglisa01	4	16	/user/cloudera/d...
/user/cloude...		AwardsSharePlayers.csv	/user/cloude...	2018-11-15 05:3...	true	Cy Young	1956	ML	newcodo01	10	16	/user/cloudera/d...
/user/cloude...		AwardsSharePlayers.csv	/user/cloude...	2018-11-15 05:3...	true	Cy Young	1956	ML	spahnwa01	1	16	/user/cloudera/d...

3 AMAZON S3

The Distributed File System Custom Wrapper can access data stored in Amazon S3 with the following Hadoop FileSystem clients:

- S3.
It is deprecated and it is not supported by the new version of this custom wrapper, version 7.0, as it was deleted from Hadoop 3.x versions.
- S3N.
Use S3A instead, as S3A client can read all files created by S3N.
S3N is not supported by the new version of this custom wrapper, version 7.0, as it was deleted from Hadoop 3.x versions.
- S3A.
S3A client can read all files created by S3N. **It should be used wherever possible.**

3.1 CONFIGURING S3 AUTHENTICATION PROPERTIES

Place the credentials in the wrapper configuration file `Custom core-site.xml`. You can use the `core-site.xml`, located in the `conf` folder of the distribution, as a guide.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

<property>
  <name>fs.s3.awsAccessKeyId</name>
  <description>AWS access key ID</description>
  <value>YOUR ACCESS KEY ID</value>
</property>

<property>
  <name>fs.s3.awsSecretAccessKey</name>
  <description>AWS secret key</description>
  <value>YOUR SECRET ACCESS KEY</value>
</property>

</configuration>
```

Alternatively, you could place the credentials in the URI `s3://ID:SECRET@BUCKET/` but this method is discouraged as they will end up in logs and error messages that untrusted people could read.

3.2 CONFIGURING S3N AUTHENTICATION PROPERTIES

Place the credentials in the wrapper configuration file `Custom core-site.xml`. You can use the `core-site.xml`, located in the `conf` folder of the distribution, as a guide.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

<property>
  <name>fs.s3n.awsAccessKeyId</name>
  <description>AWS access key ID</description>
  <value>YOUR ACCESS KEY ID</value>
</property>

<property>
  <name>fs.s3n.awsSecretAccessKey</name>
  <description>AWS secret key</description>
  <value>YOUR SECRET ACCESS KEY</value>
</property>

</configuration>
```

Alternatively, you could place the credentials in the URI `s3n://ID:SECRET@BUCKET/` but this method is discouraged as they will end up in logs and error messages that untrusted people could read.

3.3 CONFIGURING S3A AUTHENTICATION PROPERTIES

S3A supports several authentication mechanisms. By default the custom wrapper will **search for credentials in the following order**:

1. In the Hadoop configuration files.

For using this authentication method, declare the credentials in the wrapper configuration file `Custom core-site.xml`. You can use the `core-site.xml`, located in the `conf` folder of the distribution, as a guide.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

<property>
  <name>fs.s3a.access.key</name>
  <description>AWS access key ID.</description>
  <value>YOUR ACCESS KEY ID</value>
</property>
```

```
<property>
  <name>fs.s3a.secret.key</name>
  <description>AWS secret key.</description>
  <value>YOUR SECRET ACCESS KEY</value>
</property>

</configuration>
```

2. Then, the environment variables named `AWS_ACCESS_KEY_ID` and `AWS_SECRET_ACCESS_KEY` are looked for.
3. Otherwise, an attempt is made to query the Amazon EC2 Instance Metadata Service to retrieve credentials published to EC2 VMs. **This mechanism is available only when running your application on an Amazon EC2 instance, but provides the greatest ease of use and best security** when working with Amazon EC2 instances.

Alternatively, you could place the credentials in the URI `s3a://ID:SECRET@BUCKET/` but this method is discouraged as they will end up in logs and error messages that untrusted people could read.

3.3.1 Using IAM Assumed Roles

To use assumed roles, the wrapper must be configured to use the Assumed Role Credential Provider, `org.apache.hadoop.fs.s3a.auth.AssumedRoleCredentialProvider`, in the configuration option `fs.s3a.aws.credentials.provider` **in the wrapper configuration file `Custom core-site.xml`**.

This Assumed Role Credential provider will read in the `fs.s3a.assumed.role.*` options needed to connect to the Session Token Service Assumed Role API:

1. First authenticating with the full credentials. This means the normal `fs.s3a.access.key` and `fs.s3a.secret.key` pair, environment variables, or some other supplier of long-lived secrets.

If you wish to use a different authentication mechanism, other than `org.apache.hadoop.fs.s3a.SimpleAWSCredentialsProvider`, set it in the property `fs.s3a.assumed.role.credentials.provider`.

2. Then assuming the specific role specified in `fs.s3a.assumed.role.arn`
3. It will then refresh this login at the configured rate in `fs.s3a.assumed.role.session.duration`.

Below you can see the properties required for configuring IAM Assumed Roles in this custom wrapper, using its configuration file, `Custom core-site.xml`. You can use the `core-site.xml`, located in the `conf` folder of the distribution, as a guide.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

<property>
  <name>fs.s3a.aws.credentials.provider</name>
  <value>org.apache.hadoop.fs.s3a.AssumedRoleCredentialProvider</value>
  <value>org.apache.hadoop.fs.s3a.auth.AssumedRoleCredentialProvider</value>
</property>

<property>
  <name>fs.s3a.assumed.role.arn</name>
  <description>
    AWS ARN for the role to be assumed. Required if the
    fs.s3a.aws.credentials.provider contains
    org.apache.hadoop.fs.s3a.AssumedRoleCredentialProvider
  </description>
  <value>YOUR AWS ROLE</value>
</property>

<property>
  <name>fs.s3a.assumed.role.credentials.provider</name>
  <description>
    List of credential providers to authenticate with the
    STS endpoint and retrieve short-lived role credentials.
    Only used if AssumedRoleCredentialProvider is the AWS credential
    Provider. If unset, uses
    "org.apache.hadoop.fs.s3a.SimpleAWSCredentialsProvider".
  </description>
  <value>org.apache.hadoop.fs.s3a.SimpleAWSCredentialsProvider</value>
</property>

<property>
  <name>fs.s3a.assumed.role.session.duration</name>
  <value>30m</value>
  <description>
    Duration of assumed roles before a refresh is attempted.
    Only used if AssumedRoleCredentialProvider is the AWS credential
    Provider.
    Range: 15m to 1h
  </description>
</property>

<property>
  <name>fs.s3a.access.key</name>
  <description>AWS access key ID.</description>
  <value>YOUR ACCESS KEY ID</value>
</property>

<property>
  <name>fs.s3a.secret.key</name>
  <description>AWS secret key.</description>
```



```
<value>YOUR SECRET ACCESS KEY</value>
</property>

</configuration>
```

3.4 SIGNATURE VERSION 4 SUPPORT

When the V4 signing protocol is used, AWS requires the explicit region endpoint to be used —hence **S3A** must be configured to use the specific endpoint. This is done in the configuration option `fs.s3a.endpoint` in the `Custom core-site.xml` of the wrapper. You can use the `core-site.xml`, located in the `conf` folder of the distribution, as a guide. Otherwise a Bad Request exception could be thrown.

As an example of configuration, the endpoint for S3 Frankfurt is `S3.eu-central-1.amazonaws.com`:

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

<property>
  <name>fs.s3a.endpoint</name>
  <description>AWS S3 endpoint to connect to. An up-to-date list is
    provided in the AWS Documentation: regions and endpoints. Without
    this property, the standard region (s3.amazonaws.com) is assumed.
  </description>
  <value>s3.eu-central-1.amazonaws.com</value>
</property>

</configuration>
```

You can find the full list of supported versions for AWS Regions in their website: [Amazon Simple Storage Service \(Amazon S3\)](https://aws.amazon.com/s3/).

4 AZURE DATA LAKE STORE

The Distributed File System Custom Wrapper can access data stored in Azure Data Lake Store.

4.1 CONFIGURING AUTHENTICATION PROPERTIES

Place the credentials in the wrapper configuration file `Custom core-site.xml`. You can use the `core-site.xml`, located in the `conf` folder of the distribution, as a guide.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

<property>
  <name>fs.adl.oauth2.access.token.provider.type</name>
  <value>ClientCredential</value>
</property>
<property>
  <name>fs.adl.oauth2.refresh.url</name>
  <value>YOUR TOKEN ENDPOINT</value>
</property>
<property>
  <name>fs.adl.oauth2.client.id</name>
  <value>YOUR CLIENT ID</value>
</property>
<property>
  <name>fs.adl.oauth2.credential</name>
  <value>YOUR CLIENT SECRET</value>
</property>

</configuration>
```

5 AZURE BLOB STORAGE

The Distributed File System Custom Wrapper can access data stored in Azure Blob Storage.

5.1 CONFIGURING AUTHENTICATION PROPERTIES

Place the credentials in the wrapper configuration file `Custom core-site.xml`. You can use the `core-site.xml`, located in the `conf` folder of the distribution, as a guide.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

  <property>
    <name>fs.azure.account.key.<account>.blob.core.windows.net</name>
    <value>YOUR ACCESS KEY</value>
  </property>

</configuration>
```

6 GOOGLE CLOUD STORAGE

Since the Distributed File System Custom Wrapper for **Denodo 7.0**, (as this functionality requires Java 8), this wrapper can access data stored in Google Cloud Storage.

6.1 CONFIGURING AUTHENTICATION PROPERTIES

Place the credentials in the wrapper configuration file `Custom core-site.xml`. You can use the `core-site.xml`, located in the `conf` folder of the distribution, as a guide.

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

  <property>
    <name>google.cloud.auth.service.account.enable</name>
    <value>true</value>
    <description>Whether to use a service account for GCS authorization.
    If an email and keyfile are provided then that service account
    will be used. Otherwise the connector will look to see if it running
    On a GCE VM with some level of GCS access in its service account
    scope, and use that service account.</description>
  </property>

  <property>
    <name>google.cloud.auth.service.account.json.keyfile</name>
    <value>/PATH/TO/KEYFILE</value>
    <description>The JSON key file of the service account used for GCS
    access when google.cloud.auth.service.account.enable is
    true.</description>
  </property>

</configuration>
```

6.2 PERMISSIONS

Wrappers that read files content from Google Cloud Storage, like `HDFSDelimitedTextFileWrapper`, `HDFSAvroFileWrapper`, etc. requires accessing with a member with `storage.objects.get` permissions.

The `DFSListFilesWrapper`, as it list files from buckets, requires accessing with a member with `storage.buckets.get` permissions.

For more information on roles and permissions see <https://cloud.google.com/storage/docs/access-control/iam-roles>.

7 COMPRESSED FILES

Hadoop is intended for storing large data volumes, so compression becomes a mandatory requirement here. There are different compression formats available like zlib, bzip2, snappy, LZO and LZ4.

Hadoop has native implementations of compression libraries for performance reasons and for non-availability of Java implementations:

Compression format	Java implementation	Native implementation
DEFLATE	Yes	Yes
gzip	Yes	Yes
bzip2	Yes	No
LZO	No	Yes
Snappy	No	Yes

Compression library implementations

For reading compressed files using the Distributed File System Custom Wrapper there are two options:

- Use the Java implementation. In this case the wrapper handles compressed files transparently.
- Use the native implementation:
 - for performance reasons or
 - for non-availability of Java implementationIn this case the wrapper must have Hadoop the native libraries in the `java.library.path`.

8 SECURE CLUSTER WITH KERBEROS

The configuration required for accessing a Hadoop cluster with Kerberos enabled is the same as the one needed to access the distributed file system and, additionally, the user must supply the Kerberos credentials.

The Kerberos parameters are:

- **Kerberos enabled:** Check it when accessing a Hadoop cluster with Kerberos enabled. Required.
- **Kerberos principal name:** Kerberos v5 Principal name, e.g. `primary/instance@realm`. Optional.

! Note

If you enter a literal that contains one of the special characters used to indicate interpolation variables `@`, `\`, `^`, `{`, `}`, you have to escape these characters with `\`.

E.g if the Kerberos principal name contains `@` you have to enter `\@`.

- **Kerberos keytab file:** Keytab file containing the key of the Kerberos principal. Optional.
- **Kerberos password:** Password associated with the principal. Optional.
- **Kerberos Distribution Center:** Kerberos Key Distribution Center. Optional.

The Distributed File System Custom Wrapper provides **three ways** for accessing a kerberized Hadoop cluster:

1. The client has a valid Kerberos ticket in the **ticket cache** obtained, for example, using the `kinit` command in the Kerberos Client.
In this case only the `Kerberos enabled` parameter should be checked. The wrapper would use the Kerberos ticket to authenticate itself against the Hadoop cluster.
2. The client does not have a valid Kerberos ticket in the ticket cache. In this case you should provide the `Kerberos principal name` parameter and
 - 2.1. `Kerberos keytab file` parameter or
 - 2.2. `Kerberos password` parameter.

In all these **three scenarios** the **`krb5.conf` file should be present in the file system**. Below there is an example of the Kerberos configuration file:

```
[libdefaults]
renew_lifetime = 7d
forwardable = true
default_realm = EXAMPLE.COM
ticket_lifetime = 24h
dns_lookup_realm = false
dns_lookup_kdc = false

[domain_realm]
sandbox.hortonworks.com = EXAMPLE.COM
cloudera = CLOUDERA

[realms]
EXAMPLE.COM = {
    admin_server = sandbox.hortonworks.com
    kdc = sandbox.hortonworks.com
}

CLOUDERA = {
    kdc = quickstart.cloudera
    admin_server = quickstart.cloudera
    max_renewable_life = 7d 0h 0m 0s
    default_principal_flags = +renewable
}

[logging]
default = FILE:/var/log/krb5kdc.log
admin_server = FILE:/var/log/kadmind.log
kdc = FILE:/var/log/krb5kdc.log
```

The algorithm to locate the `krb5.conf` file is the following:

- If the system property `java.security.krb5.conf` is set, its value is assumed to specify the path and file name.
- If that system property value is not set, then the configuration file is looked for in the directory
 - `<java-home>\lib\security` (Windows)
 - `<java-home>/lib/security` (Solaris and Linux)
- If the file is still not found, then an attempt is made to locate it as follows:
 - `/etc/krb5/krb5.conf` (Solaris)
 - `c:\winnt\krb5.ini` (Windows)
 - `/etc/krb5.conf` (Linux)

There is an **exception**. If you are planning to create VDP views that use the **same Key Distribution Center and the same realm** the Kerberos Distribution Center parameter can be provided instead of having the krb5.conf file in the file system.

File system URI	hdfs://kerberos.cloudera:8020
Path	/user/cloudera/SearchLog.tsv
File name pattern	(.*)\\.tsv
	<input type="checkbox"/> Delete after reading
Custom core-site.xml file	None ▼
Custom hdfs-site.xml file	None ▼
Separator	\\t
Quote	
Comment marker	
Escape	
Null value	
	<input type="checkbox"/> Ignore spaces
	<input type="checkbox"/> Header
	<input checked="" type="checkbox"/> Ignore matching errors
	<input checked="" type="checkbox"/> Kerberos enabled
Kerberos principal name	cloudera-scm/admin\\@CLOUDERA
Kerberos keytab file	None ▼
Kerberos password	• • • • • • • •
Kerberos Distribution Center	kerberos.cloudera

View edition

9 TROUBLESHOOTING

Symptom

Error message: "SIMPLE authentication is not enabled. Available:[TOKEN, KERBEROS]".

Resolution

You are trying to connect to a Kerberos-enabled Hadoop cluster. You should configure the custom wrapper accordingly. See [Secure cluster with Kerberos section](#) for **configuring Kerberos** on this custom wrapper.

Symptom

Error message: "Cannot get Kerberos service ticket: KrbException: Server not found in Kerberos database (7) ".

Resolution

Check that nslookup is returning the fully qualified hostname of the KDC. If not, modify the /etc/hosts of the client machine for the KDC entry to be of the form "IP address fully.qualified.hostname alias".

Symptom

Error message: "Invalid hostname in URI s3n://<id>:<secret>@<bucket>".

Resolution

Check your bucket name: underscores are not permitted.

Also check your secret key, if it contains "/" and "+" symbols they need to be encoded in the URL. This method of **placing credentials in the URL is discouraged**. Configure the credentials on the core-site.xml instead (see **Amazon S3 support** section).

Symptom

Error message: "Error accessing Parquet file: Could not read footer: java.io.IOException: Could not read footer for file FileStatus{path=hdfs://serverhdfs/apps/hive/warehouse/parquet/.hive-staging_hive_2017-03-06_08-/-ext-10000; isDirectory=true; modification_time=1488790684826; access_time=0; owner=hive; group=hdfs; permission=rwxr-xr-x; isSymlink=false}"

Resolution

Hive could store metadata into a parquet file folder. You can check in the error message, if the custom wrapper is trying to access to any metadata. In the error of the example you can see that it is accessing a folder called `.hive-staging*`. The solution is to configure Hive to store metadata in other location.

Symptom

Error message: "Could not initialize class org.xerial.snappy.Snappy"

Resolution

On Linux platforms, an error may occur when Snappy compression/decompression is enabled although its library is available from the classpath.

The native library `snappy-<version>-libsnappyjava.so` for Snappy compression is included in the `snappy-java-<version>.jar` file. When the JVM initializes the JAR, the library is added to the default temp directory. If the default temp directory is mounted with a `noexec` option, it results in the above exception.

One solution is to specify a different temp directory that has already been mounted without the `noexec` option, as follows:

```
-Dorg.xerial.snappy.tmpdir=/path/to/newtmp
```

10 APPENDICES

10.1 HOW TO USE THE HADOOP VENDOR'S CLIENT LIBRARIES

In some cases, it is advisable to use the libraries of the Hadoop vendor you are connecting to (Cloudera, Hortonworks, ...), instead of the Apache Hadoop libraries distributed in this custom wrapper.

In order to use the Hadoop vendor libraries there is **no need to import the Distributed File System Custom Wrapper** as an extension as it is explained in the **Importing the custom wrapper into VDP** section.

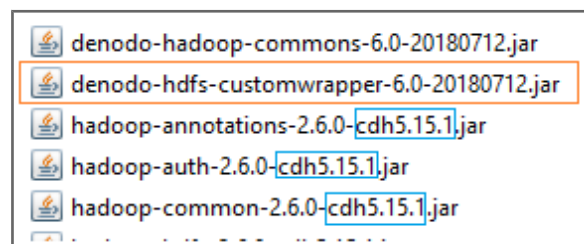
You have to create the custom data sources using the **'Classpath' parameter** instead of the 'Select Jars' option.

Click Browse to select the directory containing the required dependencies for this custom wrapper, that is:

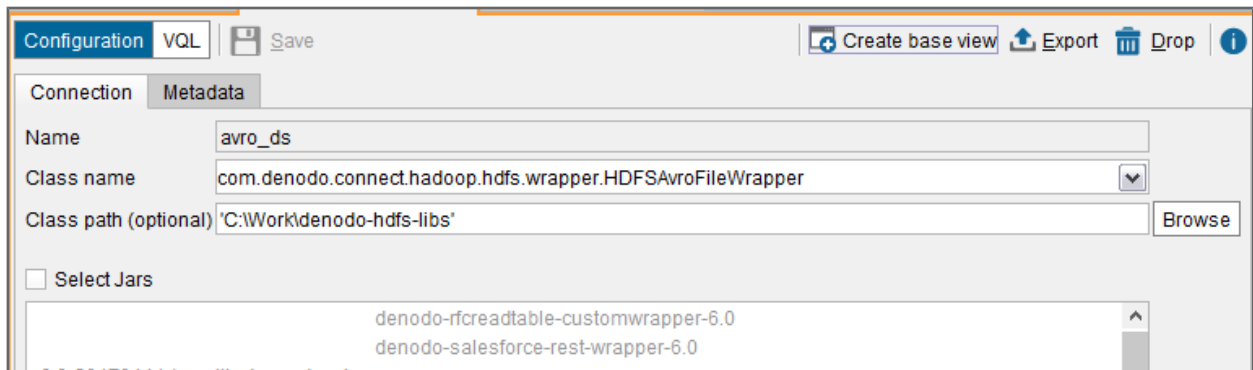
- The denodo-hdfs-customwrapper-\${version}.jar file of the dist directory of this custom wrapper distribution (highlighted in orange in the image below).
- The contents of the lib directory of this custom wrapper distribution, replacing the Apache Hadoop libraries with the vendor specific ones (highlighted in blue in the image below, the suffix indicating that they are Cloudera jars).

Here you can find the libraries for Cloudera and Hortonworks Hadoop distributions:

- Cloudera repository: <https://repository.cloudera.com/artifactory/cloudera-repos/org/apache/hadoop/>
- Hortonworks repository: <http://repo.hortonworks.com/content/repositories/releases/org/apache/hadoop/>



C:\Work\denodo-hdfs-libs directory



The screenshot shows the 'Configuration' tab of the Denodo Administration Tool. The 'Connection' tab is selected, and the 'Metadata' sub-tab is active. The configuration fields are as follows:

- Name:** avro_ds
- Class name:** com.denodo.connect.hadoop.hdfs.wrapper.HDFSAvroFileWrapper
- Class path (optional):** 'C:\Work\denodo-hdfs-libs' (with a 'Browse' button)
- Select Jars:** A checkbox that is currently unchecked.
- Jars list:** A scrollable list containing two entries: 'denodo-rfcreadtable-customwrapper-6.0' and 'denodo-salesforce-rest-wrapper-6.0'.

At the top of the window, there are tabs for 'Configuration' and 'VQL', a 'Save' button, and a toolbar with 'Create base view', 'Export', 'Drop', and an information icon.

Distributed File System Data Source

! Note

When clicking **Browse**, you will browse the file system of the host where the Server is running and not where the Administration Tool is running.