

# Подключение систем IBM Cloud Pak for Data к источникам данных Apache Hive с аутентификацией Kerberos и защитой трафика по протоколу TLS

## Компоненты решения

1. Источник данных Apache Hive версии 3.x, в режиме аутентификации Kerberos и (опционально) применением защиты трафика по протоколу TLS.
2. Домен либо лес доменов Microsoft Active Directory (Windows Server 2016) как источник данных аутентификации по протоколу Kerberos.
3. IBM Cloud Pak for Data (версии 4.0), включая модули:
  - a. Watson Knowledge Catalog,
  - b. Db2 Warehouse.

## Описание решения

- В Db2 Warehouse настраивается подключение к сервису Apache Hive путём создания объектов типа «SERVER» и «USER MAPPING».
- Для каждой базы данных Apache Hive создаётся отдельная схема Db2 Warehouse, имена таблиц и схем предпочтительно должны совпадать.
- Для каждой таблицы Apache Hive в Db2 Warehouse создаётся псевдоним («NICKNAME») в соответствующей схеме, позволяющий обращаться к исходным таблицам в клиентских сессиях Db2. Имена таблиц и полей предпочтительно должны совпадать (но при необходимости могут отличаться).
- Watson Knowledge Catalog и другие модули IBM Cloud Pak for Data (например, Watson Studio) получают доступ к описанию структуры таблиц и данным таблиц, размещённым в Apache Hive, через промежуточный слой функции Db2 Federation в составе модуля Db2 Warehouse.
- При обращении к Apache Hive используется специально созданная в Kerberos технологическая учётная запись. Опционально может использоваться вариант доступа с использованием отдельных учётных записей для каждого конечного пользователя.

## Порядок действий по настройке

### 1. Подготовка файла Kerberos keytab для технологического пользователя

Средствами Microsoft Active Directory создаётся новая учётная запись технологического пользователя.

Пользователю необходимо настроить права, достаточные для чтения данных и метаданных Apache Hive, как минимум по тем таблицам, которые планируется сделать доступными в IBM Cloud Pak for Data.

В окне командного процессора, выполняемого с правами администратора домена либо администратора организационной единицы, выполняется команда создания файла Kerberos keytab:

```
ktpass /princ username@DOMAIN.COM /pass *** /ptype KRB5_NT_PRINCIPAL /out  
username.keytab
```

## 2. Подготовка файла PKCS #12 с сертификатом центра сертификации

Информация о настройках TLS для сервиса Hive обычно фиксируется в файле `ssl-client.xml`. Для Cloudera Data Platform соответствующие настройки сервиса Hive on Tez можно просмотреть следующей командой:

```
less /etc/hive/conf.cloudera.hive_on_tez/ssl-client.xml
```

Основными параметрами является расположение файла-хранилища сертификатов и пароль для доступа к нему.

Хранилище сертификатов на кластере Hadoop имеет формат JKS и защищено паролем. Для просмотра списка сертификатов можно использовать команду, аналогичную приведённой ниже:

```
keytool -list -keystore /var/lib/cloudera-scm-agent/agent-cert/cm-auto-global_truststore.jks
```

Из хранилища необходимо выгрузить сертификат доверенного центра сертификации:

```
keytool -export -alias cmrootca-0 \  
-keystore /var/lib/cloudera-scm-agent/agent-cert/cm-auto-global_truststore.jks \  
-rfc -file hiveca.cer
```

Просмотреть детали сертификата можно следующей командой:

```
openssl x509 -in hiveca.cer -text
```

Хранилище сертификатов в формате PKCS #12, защищённое вводимым из командной строки паролем, может быть сформировано приведённой ниже командой:

```
openssl pkcs12 -export -nokeys -in hiveca.cer -out hiveca.p12
```

Введённый на предыдущем шаге пароль необходимо сохранить, так как он потребуется для настройки объекта «SERVER» на стороне Db2 Warehouse (параметр `SSL_KEYSTOREPASSWORD`).

## 3. Настройка экземпляра Db2 Warehouse

При выполнении настройки требуется наличие файлов Kerberos keytab и хранилища сертификатов центров сертификации в формате PKCS #12.

Для выполнения настройки требуется доступ к OpenShift, в котором установлен IBM Cloud Pak for Data, с правами, достаточными для подключения к подам (обычно – администратор проекта OpenShift).

Созданные файлы Kerberos keytab и файл хранилища сертификатов необходимо скопировать в контейнер OpenShift, в котором выполняется Db2 Warehouse:

```
oc cp username.keytab c-db2wh-1627300681559617-db2u-0:/tmp/  
oc cp hiveca.p12 c-db2wh-1627300681559617-db2u-0:/tmp/
```

Дальнейшие действия выполняются на стороне контейнера Db2 Warehouse, как показано на примере последовательности команд ниже.

```
oc rsh c-db2wh-1627300681559617-db2u-0
sudo mkdir /mnt/bludata0/db2/misc
sudo mv /tmp/username.keytab /mnt/bludata0/db2/misc/
sudo mv /tmp/hiveca.p12 /mnt/bludata0/db2/misc/
sudo chown -R db2inst1:db2iadml /mnt/bludata0/db2/misc
sudo chmod 444 /mnt/bludata0/db2/misc/username.keytab
sudo chmod 444 /mnt/bludata0/db2/misc/hiveca.p12
```

В примерах команд выше каталог /mnt/bludata0/db2/misc создаётся в рамках тома постоянного хранения (Persistent Volume), штатно подключенного к контейнеру Db2 Warehouse и используемого для постоянного размещения информации.

#### 4. Настройка подключения Db2 Warehouse к Apache Hive

Для подключения Db2 Warehouse к Apache Hive используется поставляемый с Db2 Warehouse ODBC-драйвер для Apache Hive, входящий в состав комплект драйверов DataDirect.

Приводимые ниже команды можно выполнять в любом клиенте Db2, включая Web-консоль Db2 Warehouse, Db2 Command Line Processor либо произвольный совместимый JDBC-клиент.

Для включения поддержки источников данных ODBC Data Direct в Db2 Warehouse необходимо создать объект типа «WRAPPER» с помощью следующей команды:

```
CREATE WRAPPER odbc LIBRARY 'libdb2rcodbc.so' OPTIONS(
  DB2_FENCED 'Y',
  MODULE '/mnt/blumeta0/home/db2inst1/sqlllib/federation/odbc/lib/libodbc.so');
```

Выше и далее в командах используется имя объекта «ODBC», но оно может быть произвольным.

Для настройки подключений к Apache Hive используются команды по созданию объектов типа «SERVER» с указанием параметров подключения, а именно:

- имени сервера,
- номера порта,
- наименования принципа Kerveros для сервиса Hive,
- пути к файлу хранилища сертификатов,
- пароля для доступа к хранилищу сертификатов.

```
CREATE SERVER hive1 TYPE hive VERSION 6.0 WRAPPER odbc
OPTIONS(HOST 'cldr66.ibmcc.ru', PORT '10000', DBNAME 'default',
  SERVER_PRINCIPAL_NAME 'hive/cldr66.ibmcc.ru@DOMAIN.COM',
  SSL_KEYSTORE '/mnt/bludata0/db2/misc/hiveca.p12',
  SSL_KEYSTOREPASSWORD 'passw0rd',
  PUSHDOWN 'Y');
```

Вместо имени объекта «hive1» в примере команды выше можно использовать произвольное имя.

Указание имени принципа Kerberos является обязательным при использовании аутентификации Kerberos.

Указание пути к хранилищу сертификатов и пароля доступа к нему является обязательным при использовании защиты трафика по протоколу TLS.

Для указания параметров аутентификации при доступе со стороны Db2 Warehouse к Apache Hive используются объекты типа «USER MAPPING».

```
CREATE USER MAPPING FOR PUBLIC SERVER hive1 OPTIONS (  
  REMOTE_AUTHID 'username@DOMAIN.COM',  
  CLIENT_PRINCIPAL_NAME 'username@DOMAIN.COM',  
  KERBEROS_KEYTAB '/mnt/bludata0/db2/misc/username.keytab');
```

Объект типа «USER MAPPING» в варианте «FOR PUBLIC» определяет параметры аутентификации по умолчанию, когда для конкретного пользователя IBM Cloud Pak for Data не настроен собственный объект «USER MAPPING».

Альтернативно для всех или части пользователей могут быть созданы отдельные объекты «USER MAPPING», что позволяет управлять используемыми для доступа к Apache Hive учётными записями Kerberos.

## 5. Создание схем и псевдонимов таблиц в Db2 Warehouse

Создание схем в Db2 Warehouse выполняется с помощью команды CREATE SCHEMA:

```
CREATE SCHEMA tpcds_10t;
```

Для создания псевдонимов таблиц Apache Hive в Db2 Warehouse используется команда CREATE NICKNAME:

```
CREATE NICKNAME tpcds_10t.call_center FOR hive1."tpcds_1t"."call_center";
```


Схема для размещения псевдонима указывается перед именем создаваемого псевдонима.

Объект, на который ссылается псевдоним, определяется через имя ранее созданного объекта типа «SERVER», имя базы данных Apache Hive и имя таблицы Apache Hive. Имена базы данных и таблицы Apache Hive обычно требуется указывать в двойных кавычках с соблюдением регистра символов.

При необходимости может быть разработан достаточно простой скрипт для периодической автоматической синхронизации состава схем и псевдонимов таблиц в экземпляре Db2 Warehouse в соответствии с изменениями структур данных на стороне Apache Hive.

## 6. Доступ к данным Apache Hive в среде IBM Cloud Pak for Data

На скриншоте ниже приведён пример списка таблиц в базе данных Apache Hive, отображаемый в интерфейсе Hue.



# Table Browser

Hive ▾

Databases

>

tpcds\_10t

Refresh

TABLES

View

Query

Drop

+ New

Filter...

<input type="checkbox"/>	Table	Description
<input type="checkbox"/>	<b>i</b> call_center	
<input type="checkbox"/>	<b>i</b> catalog_page	
<input type="checkbox"/>	<b>i</b> catalog_returns	
<input type="checkbox"/>	<b>i</b> catalog_sales	
<input type="checkbox"/>	<b>i</b> customer	
<input type="checkbox"/>	<b>i</b> customer_address	

Следующий скриншот показывает пример исполнения SQL-команды над таблицей Hive в интерфейсе консоли Db2 Warehouse.

Ниже приведён внешний вид отчёта о работе задания обнаружения данных Watson Knowledge Catalog над таблицами Apache Hive, подключенными через Db2 Warehouse.

IBM Cloud Pak for Data

Все

Поиск

Обнаружение данных /

Активы

Job: qs\_1628263697158, Project: DwhQualityDemo

Тип актива

☐ Файл

☐ Схема

☒ Таблица

☐ Столбец

Фильтры

Схемы

☐ PGDEMO (2)

☐ TPCDS\_10T (24)

Состояние

☐ Опубликовано (26)

Оценка качества (0-100)

0

100

МинимумДопустимый

Опубликовано в

☐ Data Warehouse Quali... (26)

Обнаружен таблицы (26)

Найти таблицу

<input type="checkbox"/>	Имя таблицы	Контекст	Качество	Состояние	Опубликовано в	Имя соединения	Подробности
<input type="checkbox"/>	CALL_CENTER	TPCDS_10T	97,8 %	Опубликовано	Data Warehouse Quality ...	Db2wh1	Просмотреть подробности
<input type="checkbox"/>	CATALOG_PAGE	TPCDS_10T	98,9 %	Опубликовано	Data Warehouse Quality ...	Db2wh1	Просмотреть подробности
<input type="checkbox"/>	CATALOG_RETURNS	TPCDS_10T	95,6 %	Опубликовано	Data Warehouse Quality ...	Db2wh1	Просмотреть подробности
<input type="checkbox"/>	CATALOG_SALES	TPCDS_10T	97,4 %	Опубликовано	Data Warehouse Quality ...	Db2wh1	Просмотреть подробности
<input type="checkbox"/>	CUSTOMER	TPCDS_10T	95,3 %	Опубликовано	Data Warehouse Quality ...	Db2wh1	Просмотреть подробности
<input type="checkbox"/>	CUSTOMER_ADDRESS	TPCDS_10T	96,1 %	Опубликовано	Data Warehouse Quality ...	Db2wh1	Просмотреть подробности
<input type="checkbox"/>	CUSTOMER_DEMOGRAPHICS	TPCDS_10T	100 %	Опубликовано	Data Warehouse Quality ...	Db2wh1	Просмотреть подробности
<input type="checkbox"/>	DATE_DIM	TPCDS_10T	99,9 %	Опубликовано	Data Warehouse Quality ...	Db2wh1	Просмотреть подробности
<input type="checkbox"/>	FHRW	PGDEMO	97,6 %	Опубликовано	Data Warehouse Quality ...	Db2wh1	Просмотреть подробности
<input type="checkbox"/>	HOUSEHOLD_DEMOGRAPHICS	TPCDS_10T	100 %	Опубликовано	Data Warehouse Quality ...	Db2wh1	Просмотреть подробности

Число элементов на странице

10

1–10 из 26 элементов

1 из 3 страниц

← Активы

Актив данных: CALL\_CENTER

Обнаружен columns (30)

Опубликовать актив данных

Найти столбец

Имя столбца	Контекст	Качество	Назначенный класс данных	Предложенный класс данных	Назначенный бизнес-термин	Предложенный бизнес-термин	Действия
CC_CALL_CENTER...	helper.publicocp.ib...	100 %	—	—	—	—	🔗
CC_CITY	helper.publicocp.ib...	78 %	City 69 %	US State Ca... 13 % + больше 2	—	—	🔗
CC_CLASS	helper.publicocp.ib...	100 %	Last Name 61 %	Code 100 %	—	—	🔗
CC_CLOSED_DATE...	helper.publicocp.ib...	100 %	—	—	—	—	🔗
CC_COMPANY	helper.publicocp.ib...	100 %	Code 100 %	Boolean 13 %	—	—	🔗
CC_COMPANY_NAME	helper.publicocp.ib...	98 %	Last Name 89 %	Name Suffix 20 % + больше 2	—	—	🔗
CC_COUNTRY	helper.publicocp.ib...	100 %	Country Name 100 %	Person Na... 100 % + больше 2	—	—	🔗
CC_COUNTY	helper.publicocp.ib...	96 %	Person Name 100 %	Text 100 %	—	—	🔗
CC_DIVISION	helper.publicocp.ib...	100 %	Code 100 %	Boolean 13 %	—	—	🔗
CC_DIVISION_NAME	helper.publicocp.ib...	100 %	Last Name 85 %	Name Suffix 20 % + больше 2	—	—	🔗

Число элементов на странице

10

1–10 из 30 элементов

1 из 3 страниц