Integration of GCP BigQuery with CipherTrust REST Data Protection (CRDP)

GCP BigQuery [Overview]

This document describes how to configure and integrate CipherTrust Manager with GCP BigQuery. BigQuery is a fully managed enterprise data warehouse that helps you manage and analyze your data with built-in features like machine learning, geospatial analysis, and business intelligence. BigQuery's serverless architecture lets you use SQL queries to answer your organization's biggest questions with zero infrastructure management. Federated queries let you read data from external sources while streaming supports continuous data updates. BigQuery's scalable, distributed analysis engine lets you query terabytes in seconds and petabytes in minutes.

Thales provides a couple of different methods to protect sensitive data in GCP BigQuery.

Bring Your Own Encryption (BYOE)

- Data Ingest with Thales Batch Data Transformation (BDT)
- **Data Access** external remote functions for column level encrypt and decryption using Ciphertrust REST Data Protection (CRDP).

Bring/Hold Your Own Key (BYOK) (HYOK)

GCP BigQuery Customer Managed Keys- with Thales CM CCKM BYOK and HYOK.

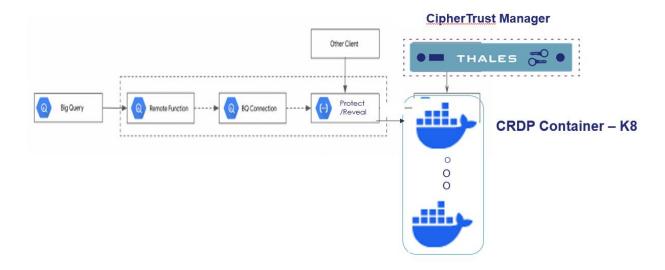
The above methods are NOT mutually exclusive. All methods can be used to build a strong defense in depth strategy to protect sensitive data in the cloud. The focus of this integration will be on Data Access protecting sensitive data in GCP BigQuery columns by using CRDP to create User Defined Functions (UDF) for encryption and decryption of sensitive data.

Architecture

The examples provided in this document use a capability GCP BigQuery called "Remote Function". A BigQuery remote function lets you incorporate GoogleSQL functionality with software outside of BigQuery by providing a direct integration with Cloud Functions and Cloud Run. With BigQuery remote functions, you can deploy your functions in Cloud Functions or

Cloud Run implemented with any supported language, and then invoke them from GoogleSQL queries.

A BigQuery remote function allows you to implement your function in other languages than SQL and Javascript or with the libraries or services which are not allowed in BigQuery user-defined functions. Listed below is a diagram of how this integration works.



Supported Product Versions

- CipherTrust Manager CipherTrust Manager 2.14 and higher
- **CRDP** 1.0 and higher
- GCP BigQuery

This integration is validated using 2^{nd} generation Google Cloud Functions and Java 11 along with CM 2.14.

Prerequisites

Steps performed for this integration were provided by this GCP link: https://cloud.google.com/bigquery/docs/reference/standard-sql/remote-functions

https://cloud.google.com/bigguery/docs/remote-function-tutorial#console

- Ensure that CRDP for is installed and configured. Refer to https://thalesdocs.com/ctp/con/crdp/latest/admin/index.html
- Ensure that the CipherTrust Manager is installed and configured. Refer to the <u>CipherTrust Manager documentation</u> for details.

GCP Cloud function communicates with the CRDP Container using REST API's

Steps for Integration

- [Installing and Configuring Thales CRDP container]
- [Download code from Thales github and compile]
- [Publish jar/zip file to GCP Cloud Function (endpoint)]
- [Create BigQuery Connection and GCP BigQuery Remote Function]
- [Integration with Thales CipherTrust Manager]
- [Environment Variables]

Installing and Configuring CRDP Container

To install and configure **CDRP**, refer to **Quick Start**.

Eclipse development tool was used for these examples. Here is the version used for testing along with the Maven plugin for Eclipse.

eclipse.buildId=4.15.0.I20200305-0155

m2e - Maven Integration for Eclipse

Here is a screenshot in eclipse of the classes used for these examples: The classes that contain CRDP in the names below are the ones to use.

→ B CADP-GCP-BigQuery-UDF Li CMUserSetHelper.java > In ThalesGCPBigQueryCADPFPE.java > ThalesGCPBigQueryCADPFPERLRetry.java > In ThalesGCPBigQueryCRDPBulkFPE.java > 🕖 ThalesGCPBigQueryCRDPFPE.java > # src/main/resources √

// src/test/java Li CMUserSetHelper.java GCPBigQueryUDFTester.java > 🕖 GCPBigQueryUDFTesterSingle.java > 🕖 TestThalesGCPBigQueryCADPFPERLRetry.java > 🕖 TestThalesGCPBigQueryCRDP.java > All TestThalesGCPBigQueryCRDPBulkFPE.java > 🕖 TestThalesGCPBigQueryCADPFPERLRetry.java

Download code from github and compile.

git clone https://github.com/ThalesGroup/CipherTrust Application Protection.git

The database directory has all the code for GCP BigQuery. You should see the above classes in your project.

CRDP supports a bulk API which allows for CRDP to batch requests before calling protect or reveal. A separate class file for bulk and nonbulk is available for testing with this API.

Assuming you have your CM already configured the TestThalesGCPBigQueryCRDP can be used to test basic connection to your CRDP container to make sure your environment is configured correctly. You will need to modify the keyname to make sure it exists in CM and there are environment variables that must be set as well. If you plan on using UserSets the user must also be in the Application Data Protection Client Group. See section on Environment Variables for more details.

Generate the jar file to upload to the CSP.

To compile and generate the target jar file to be uploaded to the CSP select the project and choose "Run As" "maven install" to generate the target.

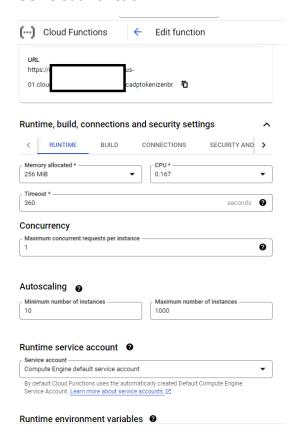
```
[INFO] --- maven-install-plugin:3.0.1:install (default-install) @ CADP-GCP-Function -
[INFO] Installing C:\Users\t0185905\workspace\CADP-GCP-Function\pom.xml to
C:\Users\t0185905\.m2\repository\Thales\CADP-GCP-Function\0.0.1-SNAPSHOT\CADP-GCP-
Function-0.0.1-SNAPSHOT.pom
[INFO] Installing C:\Users\t0185905\workspace\CADP-GCP-Function\target\CADP-GCP-
Function-0.0.1-SNAPSHOT.jar to C:\Users\t0185905\.m2\repository\Thales\CADP-GCP-
Function\0.0.1-SNAPSHOT\CADP-GCP-Function-0.0.1-SNAPSHOT.jar
[INFO] Installing C:\Users\t0185905\workspace\CADP-GCP-Function\target\CADP-GCP-
Function-0.0.1-SNAPSHOT-jar-with-dependencies.jar to
C:\Users\t0185905\.m2\repository\Thales\CADP-GCP-Function\0.0.1-SNAPSHOT\CADP-GCP-
Function-0.0.1-SNAPSHOT-jar-with-dependencies.jar
[INFO] ------
[INFO] BUILD SUCCESS
[INFO] -----
[INFO] Total time: 7.250 s
[INFO] Finished at: 2024-02-28T10:56:43-05:00
[INFO] ------The
```

The code provided has uses Googles userDefinedContext capability. This code accepts a mode and a datatype as a keyvalue pair.

Publish jar/zip file Google Cloud Function. (endpoint)

Once you have generated the jar file to upload you can then create the CSP function. Google requires a zip file so zip up the jar file in the target directory of your eclipse project.

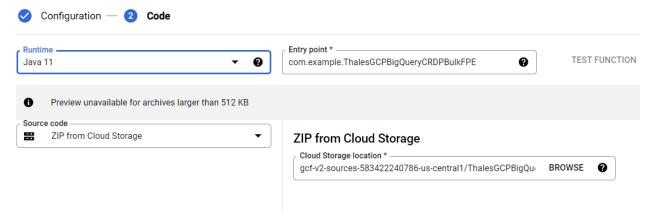
GCP Cloud Function



Set environment variables appropriate values (See section on Environment Variables for details) and then select Next. Upload the zip file on the next screen. Be sure to change the entry point to reflect your class name. *The example below is*

com.example.ThalesGCPBigQueryCRDPBulkFPE This should match your code path.

Note: Initial testing indicate that the CRDP Bulk version is faster and should be used.



Click Deploy to deploy the function.

Once you have created the functions above and if you have already configured and setup CM with the key and all the environment variables you can test the function with the test tab. You will need to provide the appropriate json to test. Please see section below for valid values for the mode. The protection_profile values will be values you have in CM under Application Data Protection Tile/Protection Profiles. Here is an example:

```
"requestId": "124ab1c",
  "caller":
"//bigquery.googleapis.com/projects/myproject/jobs/myproject:US.bquxjob_5b4c112c_17961fafeaf",
  "sessionUser": "test-user@test-company.com",
  "userDefinedContext": {
   "mode": "protectbulk",
   "protection profile": "alpha-external"
  "calls": [
   [
     93309296
    ],
     74705755
     39056597430
   ],
    [
     6621883
     2662402956
     17506289853
    1
```

Create and configure BigQuery connection and GCP BigQuery Remote Function.

Here are some links that provide details.

https://cloud.google.com/bigquery/docs/remote-function-tutorial#console

https://cloud.google.com/bigguery/docs/remote-functions

As noted above the steps are:

- 1. Create the GCP Cloud Function. (should already be done from above)
- 2. Create GCP BigQuery Connection.
- 3. Create remote function object in GCP BigQuery

Examples:

bq mk --connection --display_name='warnerscorner' --connection_type=CLOUD_RESOURCE -- project id=yourprojectid --location=US mw-remote-add-conn

Big Query Function Definitions.

The functions are created using the following format:

data type in db/actual data format/return type

Examples:

```
SELECT `abc-sales-l-app-us-
02.your_demo_dataset_US.thales_crdp_protect_char`
('123382742348293479')
093381171100859737
```

```
CREATE or replace FUNCTION `your-gcp-
project.mw demo dataset US.thales crdp protect char`(x String)
RETURNS String
REMOTE WITH CONNECTION `your-gcp-project.us.mw-remote-add-conn`
OPTIONS (
endpoint = 'https://us-central1-your-gcp-
project.cloudfunctions.net/ThalesGCPBigQueryCRDPFPE',
user defined context = [("mode", "protect"),("protection profile", "plain-
alpha-internal")]
);
CREATE or replace FUNCTION `your-gcp-
project.mw demo dataset US.thales crdp reveal char`(x String)
RETURNS String
REMOTE WITH CONNECTION `your-qcp-project.us.mw-remote-add-conn`
OPTIONS (
endpoint = 'https://us-central1-your-gcp-
project.cloudfunctions.net/ThalesGCPBigQueryCRDPFPE',
user_defined_context = [("mode", "reveal"),("protection profile", "plain-
alpha-internal")]
);
CREATE or replace FUNCTION `your-gcp-
project.mw demo dataset US.thales crdp protect char ext`(x String)
RETURNS String
REMOTE WITH CONNECTION `your-gcp-project.us.mw-remote-add-conn`
OPTIONS (
endpoint = 'https://us-central1-your-gcp-
project.cloudfunctions.net/ThalesGCPBigQueryCRDPFPE',
user defined context = [("mode", "protect"),("protection profile", "alpha-
external")]
);
CREATE or replace FUNCTION `your-gcp-
project.mw demo dataset US.thales crdp reveal char ext`(x String)
RETURNS String
REMOTE WITH CONNECTION `your-gcp-project.us.mw-remote-add-conn`
OPTIONS (
endpoint = 'https://us-central1-your-gcp-
project.cloudfunctions.net/ThalesGCPBigQueryCRDPFPE',
```

```
user defined context = [("mode", "reveal"),("protection profile", "alpha-
external")]
);
CREATE or replace FUNCTION `your-gcp-
project.mw demo dataset US.thales crdp protect char bulk`(x String)
RETURNS String
REMOTE WITH CONNECTION `your-gcp-project.us.mw-remote-add-conn`
OPTIONS (
endpoint = 'https://us-central1-your-gcp-
project.cloudfunctions.net/ThalesGCPBigQueryCRDPBulkFPE',
user defined context = [("mode", "protectbulk"),("protection profile",
"plain-alpha-internal")]
CREATE or replace FUNCTION `your-gcp-
project.mw demo dataset US.thales crdp reveal char bulk`(x String)
RETURNS String
REMOTE WITH CONNECTION `your-gcp-project.us.mw-remote-add-conn`
OPTIONS (
endpoint = 'https://us-central1-your-gcp-
project.cloudfunctions.net/ThalesGCPBigQueryCRDPBulkFPE',
user defined context = [("mode", "revealbulk"),("protection profile", "plain-
alpha-internal")]
);
CREATE or replace FUNCTION `your-gcp-
project.mw demo dataset US.thales crdp protect char ext bulk (x String)
RETURNS String
REMOTE WITH CONNECTION `your-gcp-project.us.mw-remote-add-conn`
OPTIONS (
endpoint = 'https://us-central1-your-gcp-
project.cloudfunctions.net/ThalesGCPBigQueryCRDPBulkFPE',
user defined context = [("mode", "protectbulk"),("protection profile",
"alpha-external")]
CREATE or replace FUNCTION `your-gcp-
project.mw demo dataset US.thales crdp reveal char ext bulk`(x String)
RETURNS String
REMOTE WITH CONNECTION `your-gcp-project.us.mw-remote-add-conn`
endpoint = 'https://us-centrall-your-gcp-
project.cloudfunctions.net/ThalesGCPBigQueryCRDPBulkFPE',
user defined context = [("mode", "revealbulk"),("protection_profile", "alpha-
external") |
);
```

Integration with CipherTrust Manager.

Here is a link to a demo that shows all the steps required to setup CRDP in CM.

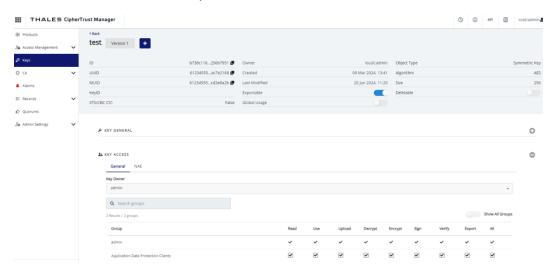
thales.navattic.com/thalesprotectreveal

Here is a list of the steps.

1. Create an Encryption Key

- 2. Create a CRDP Application
- 3. Create a User Set
- 4. Create an Access Policy
- 5. Create a Protection Policy
- 6. Perform Integration

When creating a key to be used by the protection policy make sure that it is in the Application Data Protection Clients Group.



If userset lookups are desired, then when creating a user it should have the ability to export the key and also be in the Application Data Protection Clients Group. This user will also be needed to be provided as an environment variable for the function along with the password. The examples provided have the key as a hardcoded value, but this can be easily altered to be provided as an environment variable, obtained from a secrets manager or in the userDefinedContext of the create function statement.

As noted above there is a test class that can be used to test connectivity with CM without having to publish the Function.

When all the above steps are performed you should see your UDF's in GCP BigQuery under Routines in the UI. Here is a sample query using one of the UDF's.

Sample Results:

```
select `abc-sales-l-app-us-02.your_demo_dataset_US.thales_crdp_protect_char` (email)
as encemail, email from `abc-sales-l-app-us-02.your_demo_dataset_US.plaintext` limit
10
```

Row	encemail ▼	email ▼
1	QUsk4ch@T5VJn.RqR	lkemmer@yahoo.com
2	R1pr5@09aOhf.FoX	kim46@bednar.com
3	G04W.CRNIb@riDnQ.DHT	batz.geary@gmail.com
4	jR4fSE.odmST@bfQ.eFb	shelbi.mertz@kub.com
5	ZmxHvXyVc@GpAjRfq0.j0D	mathias47@hanebode.com
6	1IOEk@RPDqC.GHd	ikris@gmail.com
7	hf69RzWcQs.gMyslj@EEJJEnf	rutherford.maxine@casperwilki
8	yBiX8AFovMk0@CBIt93M.oKV	charlottie26@hotmail.com
9	70Hpfz.PXI0PoJJaWIQ@AvTP	huston.christiansen@gmail.com
10	dhXYofz@X1YVe.9s9	edson67@yahoo.com

Environment Variables

Listed below are the environment variables currently required for the Cloud Function.

Key	Value	Desc	
BATCHSIZE	200	Nbr of rows to chunk when using batch mode	
CMPWD	Yourpwd!	CM PWD if using CADP	
CMUSER	apiuser	CM USERID if using CADP	
CRDPIP	20.221.216.666	CRDP Container IP if using CRDP	
datatype	charint	datatype of column (char or charint)	
keymetadata	1001000	policy and key version if using CRDP	
keymetadatalocation	external	location of metadata if using CRDP (internal,external)	
mode	revealbulk	mode of operation(protect,reveal,protectbulk,revealbulk) CRDP	
protection_profile	plain-nbr-ext	protection profile in CM for CRDP	
returnciphertextforuserwithnokeyacc ess	yes	if user in CM not exist should UDF error out or retur ciphertext	
usersetidincm	716f01a6-5cab-4799-925a- 6dc2d8712fc1	userset in cm if user lookup is done	
usersetlookup	no	should uselookup be done (yes,no)	
usersetlookupip	20.241.70.666	userset lookup	
showrevealinternalkey	yes	show keymetadata when issuing a protect call (CRDP)	

Advanced Topics.

Google also provides the ability to setup a secure perimeter with VPC Service Controls. See link below for more information.

https://cloud.google.com/bigquery/docs/remote-functions#using vpc service controls

Application Data Protection UserSets

Application Data Protection UserSets are currently used for DPG and CRDP to control how the data will be revealed to users. These UserSets can also be independent of any Access Policy. Most cloud databases have some way to capture who is running the query and this information can be passed to CM to be verified in a UserSet to ensure the person running the query has been granted proper access. In github there is a sample class file called CMUserSetHelper that can be used to load a userset with values from an external identity provider such as LDAP. The name of this method is addAUserToUserSetFromFile. Once users have been loaded into this userset the usersetid must be captured and used as an environment variable to the Function. The function has a number of environment variables that must be provided for the function to work. Please review the section on Environment Variables for more details.

Options for handling null values.

Since it is not possible to encrypt a column that contains null values or any column that has 1 byte it is necessary to skip those to avoid getting an error when running the query. There are a couple of ways to handle this use case.

Option 1. Modify the queries.

Many times, simply adding a where clause to exclude values that have nulls or less than 2 bytes can avoid query errors. For example: select * from FROM

mw_demo_dataset_US.plaintext50cadpemailprotected_nulltest where email is not null and length(email) > 1; For those scenarios where that is not suffice some other examples are listed below. Here is an example of a select statement that can be modified to handle null values:

```
SELECT

name,

CASE

WHEN email IS NULL THEN 'null'

WHEN length(email) < 2 then email

WHEN email = 'null' then email

ELSE `your-gcp-project.mw_demo_dataset_US.thales_crdp_protect_char`(email)

END AS email

FROM

mw_demo_dataset_US.plaintext50cadpemailprotected_nulltest;
```

The above use case was for situations where the column contained both null and the word 'null'.

Here is an example of a select statement that can be modified to handle null values in the where clause.

```
SELECT name, email, email_enc
FROM
mw_demo_dataset_US.plaintext50cadpemailprotected_nulltest
where CASE
```

```
WHEN email_enc IS NULL THEN 'null'
WHEN length(email_enc) < 2 then email_enc
WHEN email_enc = 'null' then email_enc
ELSE `your-gcp-project.mw_demo_dataset_US.thales_crdp_reveal_char`(email_enc)
END like "%gmail%"</pre>
```

Row	name ▼	email ▼	email_enc ▼
1	Dr. Lemmie Zboncak	ikris@gmail.com	1IOEk@RPDqC.GHd
2	Zillah Leuschke	scronin@gmail.com	53mWY7Y@4bzb6.2D4
3	Troy Gaylord	devon49@gmail.com	EsTMziF@ISZg2.yN6
4	Dr. Spurgeon Wintheiser	hilah17@gmail.com	PIM4vAd@qAIV9.oJC
5	Tatyana Bernhard	denny56@gmail.com	yVQ7ITA@3idYW.GqV
6	Renata Hilpert	tdickens@gmail.com	plzg3zLb@oetcj.Opi
7	Deliah Douglas	sconnelly@gmail.com	jpXAUZWPX@koaH2.yS8
8	Amare Feeney	batz.geary@gmail.com	G04W.CRNIb@riDnQ.DHT
9	Dr. Cristy Schinner	schulist.garfield@gmail.com	HII90wCU.LPtU3p6o@F9Nb7.G
10	Vena Douglas	huston.christiansen@gmail.com	70Hpfz.PXI0PoJJaWIQ@AvTP

Option 2. Modify the Cloud Function to skip encrypting these values.

Please review the method checkvalid for the logic of handling these use cases.

Note: When using this logic it is important to know that for any column that is null will return 'null'. This should be fine for use cases where the query is not updating the column. For use cases where the source system is expecting null vs the word 'null' additional testing should be conducted on the systems that rely on this data type as being null vs the word 'null'.

Sample CRDP docker commands.

As noted in the link above there are number of ways to deploy the CRDP container. For demo purposes here are the steps that were performed for a standalone use case.

```
sudo apt update
sudo apt install apt-transport-https ca-certificates curl
software-properties-common
curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo
apt-key add -
sudo add-apt-repository "deb [arch=amd64]
https://download.docker.com/linux/ubuntu bionic stable"
sudo apt update
apt-cache policy docker-ce
sudo apt install docker-ce
```

```
sudo systemctl status docker
sudo usermod -aG docker crdpuser
su - crdpuser
docker ps
docker pull thalesciphertrust/ciphertrust-restful-data-
protection: latest
exit
//Back on as root
docker run -e KEY MANAGER HOST=192.168.159.134 -e
REGISTRATION TOKEN=vw886rxzMolgtL1Gz950Ta9kVbDDvo0qgawGHthOc7iKn
m8b1VP6R2ps3qifVDTN -p 8090:8090 -e SERVER MODE=no-tls
thalesciphertrust/ciphertrust-restful-data-protection
{"level":"info","time":"Thu, 20 Jun 2024 14:56:26 +0000","msg":"going to initialize shield"}
{"level":"info","time":"Thu, 20 Jun 2024 14:56:26 +0000","msg":"registerClient: Going to
register the client"}
{"level":"info","time":"Thu, 20 Jun 2024 14:56:26 +0000","msg":"registerClient: Register the
client successfully"}
{"level":"info","time":"Thu, 20 Jun 2024 14:56:32
+0000", "msg": "", "app name": "crdpapp1", "audit": true, "client id": "be458360-e656-4faf-b7ee-
0d210626b214","endpoint":"/v1/protect","jwt_username":"","key_name":"test","key_version":
"1","method":"POST","protection policy name":"plain-alpha-
internal", "protection_policy_version": "4", "source_ip": "192.168.159.1", "status": "Success"}
{"level":"info","time":"Thu, 20 Jun 2024 14:56:45
+0000", "msg": "", "app name": "crdpapp1", "audit": true, "client id": "be458360-e656-4faf-b7ee-
0d210626b214","endpoint":"/v1/protect","jwt_username":"","key_name":"test","key_version":
"1","method":"POST","protection policy name":"plain-alpha-
internal", "protection_policy_version": "4", "source_ip": "192.168.159.1", "status": "Success"}
{"level":"info","time":"Thu, 20 Jun 2024 14:56:51
+0000", "msg": "", "app name": "crdpapp1", "audit": true, "client id": "be458360-e656-4faf-b7ee-
0d210626b214","endpoint":"/v1/protect","jwt_username":"","key_name":"test","key_version":
"1","method":"POST","protection_policy_name":"plain-alpha-
internal", "protection_policy_version": "4", "source_ip": "192.168.159.1", "status": "Success"}
{"level":"error","time":"Thu, 20 Jun 2024 14:56:57 +0000","msg":"error while encrypting the
data, err: Input buffer is too short (len=1), it has to be at least 2 characters long."}
```

```
{"level":"error","time":"Thu, 20 Jun 2024 14:56:57 +0000","msg":"Input buffer is too short (len=1), it has to be at least 2 characters long.","app_name":"crdpapp1","audit":true,"client_id":"be458360-e656-4faf-b7ee-0d210626b214","endpoint":"/v1/protect","jwt_username":"","key_name":"test","key_version": "1","method":"POST","protection_policy_name":"plain-alpha-internal","protection_policy_version":"4","source_ip":"192.168.159.1","status":"Error"}
```