

# 在GCP创建AWS RDS实例的远程备份并作为BigQuery外表查询

本文介绍如何在谷歌云配置AWS RDS MySQL实例的同步副本，并在谷歌云BigQuery数据仓库使用Federated Query联合查询来查询来自于RDS的数据，并与数据仓库里的数据做合并查询。这样做可以扩展BigQuery查询的数据来源，丰富使用场景，简化数据导出导入过程。

[更新RDS实例参数](#)

[创建同步复制数据库实例](#)

[在BigQuery中执行联合查询](#)

## 更新RDS实例参数

当在谷歌云创建外部源数据库的同步副本时，对于外部源数据库有如下要求[\[1\]](#)。

- 正在运行 MySQL Community Edition 版本 5.6 或 5.7。
- 已启用二进制日志。 [了解详情](#)。
- 将二进制日志保留足够长的时间，让副本能完成导入。  
一般来说，一周的时间应该足够了。 [详细了解如何设置二进制日志保留政策](#)。
- 使用基于行的二进制日志记录。 [了解详情](#)。
- 启用 GTID 并强制执行 GTID 一致性。  
[详细了解 GTID](#)。请查看[强制执行 GTID 一致性](#)的选项。
- 所有表（系统数据库中的表除外）都使用 InnoDB 存储引擎。  
[详细了解 InnoDB](#)。 [详细了解如何转换为 InnoDB](#)。
- 具有 REPLICATION\_SLAVE 权限的 MySQL 用户帐号。  
将此帐号配置为接受来自任何位置 (host = %) 的连接。您可以在[后面的步骤](#)中限制此用户的访问权限。您不应将此用户帐号用于复制以外的任何其他目的。  
[详细了解权限](#)。 [详细了解用户帐号](#)。
- 可在外部访问的 IPv4 地址和 TCP 端口。

请检查您的RDS实例满足上述要求。下面的步骤作为示例介绍如何更新RDS实例满足启用GTID和创建有REPLICATION\_SLAVE权限的用户账号的要求。

首次要为已有的RDS实例配置开启GTID复制，需要实例为MySQL 5.7.23或以上版本。目前（2020年3月）的默认版本一般为5.7.22，所以大部分已有实例需要进行升级。可以在AWS RDS控制台查看实例的MySQL版本。

---

## Configuration

DB instance id

rds-sin-1

Engine version

5.7.22

在控制台对实例进行修改，选择5.7.23或以上的版本（提前对数据做好备份）。

---

DB engine version

Version number of the database engine to be used for this instance.

MySQL 5.7.23 ▼

RDS默认的Parameter Group参数组没有开启GTID，需要创建单独的参数组来开启（除非已经有了非默认的参数组）。如下图。

RDS > Parameter groups > Create parameter group

## Create parameter group

**Parameter group details**

To create a parameter group, choose a parameter group family, then name and describe your parameter group

**Parameter group family**  
DB family that this DB parameter group will apply to

mysql5.7

**Group name**  
Identifier for the DB parameter group

db-param-group-for-gcp-replica

**Description**  
Description for the DB parameter group

For parameters to enable replica in GCP

Cancel Create

在新创建的参数组里，配置开启两个参数gtid-mode和enforce\_gtid\_consistency。

RDS > Parameter groups > db-param-group-for-gcp-replica

## db-param-group-for-gcp-replica

**Parameters** Cancel editing Preview changes Reset Save changes

Q gtid X < 1 > ⚙

<input type="checkbox"/>	Name	Values	Allowed values	Modifiable	Source	Apply type
<input type="checkbox"/>	binlog_gtid_simple_recovery		0, 1	true	engine-default	static
<input type="checkbox"/>	enforce_gtid_consistency	ON	OFF, WARN, ON	true	user	static
<input type="checkbox"/>	gtid_executed_compression_period		0-4294967295	true	engine-default	dynamic
<input type="checkbox"/>	gtid-mode	ON	OFF, OFF_PERMISSIVE, ON_PERMISSIVE, ON	true	user	static
<input type="checkbox"/>	gtid_purged			false	engine-default	dynamic
<input type="checkbox"/>	session_track_gtids		0, 1	true	engine-default	dynamic

**Recent events** ↺

然后更新RDS实例，采用新的参数组。

RDS > Databases > rds-sin-1

### rds-sin-1

**Modify** **Actions** ▼

#### Summary

DB identifier rds-sin-1	CPU 1.50%	Info Available	Class db.t2.micro
Role Instance	Current activity 0 Connections	Engine MySQL Community	Region & AZ ap-southeast-1a

Connectivity & security | Monitoring | Logs & events | **Configuration** | Maintenance & backups | Tags

#### Instance

#### Database options

**Database port**  
Specify the TCP/IP port that the DB instance will use for application connections. The connection string of any application connecting to the DB instance must specify the port number of the DB instance. Both the security group applied to the DB instance and your company's firewalls must allow connections to the port. [Learn More](#)

3306

**DB parameter group**  
Database parameter group to associate with this DB instance

default:mysql5.7 ▼

db-param-group-for-gcp-replica

default:mysql5.7

default:mysql-5-7 ▼

**IAM DB authentication** [Info](#)

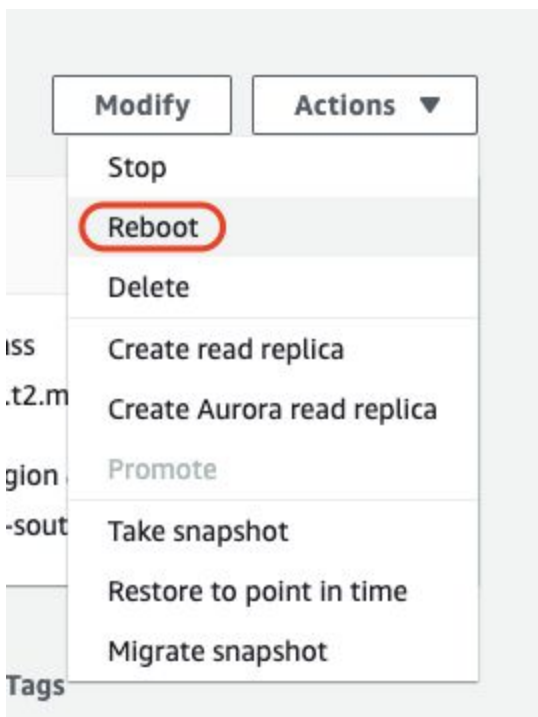
☐ Enable IAM DB authentication  
Manage your database user credentials through AWS IAM users and roles.

☒ Disable

更新完参数组后，实例信息会提醒需要实例重启后才能生效。

Instance		
Configuration	Instance class	Storage
DB instance id rds-sin-1	Instance class db.t2.micro	Encryption Not Enabled
Engine version 5.7.22	vCPU 1	Storage type General Purpose (SSD)
DB name -	RAM 1 GB	IOPS -
License model General Public License	Availability	Storage 20 GiB
Option groups <a href="#">default:mysql-5-7</a>	Master username admin	Storage autoscaling Enabled
ARN arn:aws:rds:ap-southeast-1:881826878679:db:rds-sin-1	IAM db authentication Not Enabled	Maximum storage threshold 1000 GiB
Resource id db-B64CVTJWCGCPTIVR4HH5D5RIDI	Multi AZ No	
Created time Tue Mar 10 2020 10:39:18 GMT+0800 (China Standard Time)	Secondary Zone -	
Parameter group <a href="#">db-param-group-for-gcp-replica</a> (pending-reboot)		
Deletion protection Disabled		

对实例进行重启。



重启后，确认实例状态正常，并且相关参数已经更新。

```
SHOW GLOBAL VARIABLES LIKE 'gtid_mode';
```

正确返回：

Variable_name	Value
gtid_mode	ON

确认enforce\_gtid\_consistency：

```
SHOW GLOBAL VARIABLES LIKE 'enforce_gtid_consistency';
```

正确返回：

Variable_name	Value
enforce_gtid_consistency	ON

用下面命令创建用于同步和复制数据的数据库用户。

```
CREATE USER 'repl'@'%' IDENTIFIED BY 'Replicate4gcp';
GRANT REPLICATION SLAVE ON *.* TO 'repl'@'%';
GRANT SELECT, SHOW VIEW, RELOAD, REPLICATION CLIENT, EVENT, TRIGGER ON *.* TO 'repl'@'%';
```

## 创建同步复制数据库实例

首先创建谷歌云对象存储桶，用来上传存量数据的导出。可以在谷歌云控制台创建一个存储桶，选择与BigQuery数据集相同的区域，比如香港，并命名为youzhi-lab（此名为示例）。

导出数据库存量数据。本例的guestbook数据库为一个只有一个表的示例数据库。可以在谷歌云内的虚拟机，或者Cloud Shell上执行。

```
mysqldump \
-h rds-sin-1.cdmhzspt0jdq.ap-southeast-1.rds.amazonaws.com \
```

```
-P 3306 -u repl -p \  
--databases guestbook \  
--hex-blob --skip-triggers \  
--order-by-primary --no-autocommit \  
--default-character-set=utf8mb4 \  
--single-transaction --set-gtid-purged=on | gzip > entries_dump.sql.gz
```

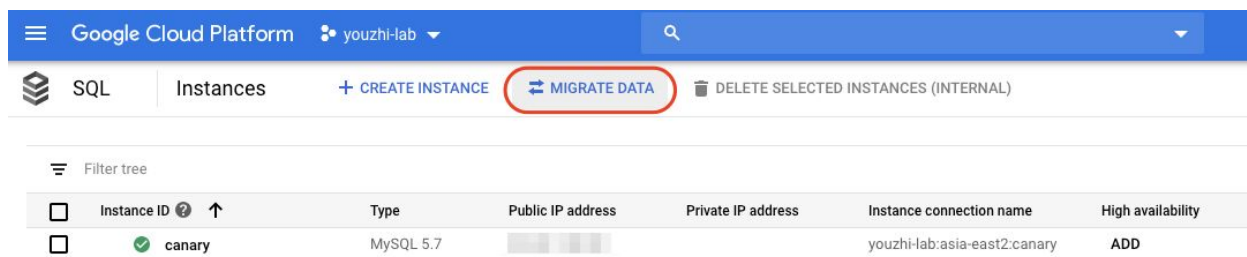
生成的导出文件可以在控制台上传到存储桶youzhi-lab，也可以执行下面命令。

```
gsutil cp entries_dump.sql.gz gs://[YOUR_BUCKET]/
```

另外，如果RDS实例Binlog保存时间很短，可以在RDS实例执行下面命令，让其保存至少24小时。

```
call mysql.rds_set_configuration('binlog retention hours', 24);
```

在谷歌云Cloud SQL控制台，开始创建指向外部源数据库的同步副本。首先点击MIGRATE DATA 打开创建向导。



填写副本的基本信息。注意选择MySQL版本为5.7。区域与BigQuery数据集所在的区域一致。SQL Dump File选择上面步骤上传在存储桶的文件。



To begin migration, proceed through the following steps.

### 1 Data source details



The migration assistant will generate a Cloud SQL external primary instance that maps to your data source details.

Name of data source ?

Public IP address of source ?

Port number of source ?

MySQL replication username ?

MySQL replication user password ?



Database version



SSL/TLS certification

☐ Enable SSL/TLS security

If selected, Cloud SQL will use SSL/TLS encryption for the replication connection between the replica and the data source. Recommended for security.

Next

2 Cloud SQL read replica creation



3 Data synchronization



4 Read replica promotion (optional)



Finish

Cancel





SQL



Migrate to Cloud SQL

**Read replica instance ID**

Cannot be changed later. Use lowercase letters, numbers, and hyphens.  
Start with a letter.

**Location**

For better performance, keep your data close to the services that need it.

**Region**

Choice is permanent

**Zone**

Can be changed at any time

**Machine type**

Select a machine type for your read replica instance. For best results, select similar or higher specifications to your source database, instance, or machine.



db-n1-standard-1

vCPUs

Memory

1

3.75 GB

[Change](#)**Network throughput (MB/s)**

250 of 2,000

**Storage type**

Choice is permanent.

☒ **SSD (Recommended)**

Most popular choice. Lower latency than HDD with higher QPS and data throughput.

☐ **HDD**

Lower performance than SSD with lower storage rates.

**Storage capacity**

10 – 30720 GB. Higher capacity improves performance, up to the limits set by the machine type. Capacity can't be decreased later.

 ☒ **Enable automatic storage increases**

If enabled, whenever you're nearing capacity, storage will be incrementally (and permanently) increased. [Learn more](#)

**Import SQL dump from Google Cloud Storage**

Choose the SQL file to import into your read replica that will seed your instance

**SQL Dump File** ☒ youzhi-lab/entries\_dump.sql.gz[Browse](#)

Once you select the SQL file, a Cloud SQL service account will receive read-only access to your Cloud Storage file and the bucket that contains it. Your bucket and file permissions will reflect this access.

[Show advanced options](#)[Create](#)

最后点击Finish按钮。等待一段时间可以看到副本创建成功。界面上可以看到实例的连接参数和状态。

### 3 Data synchronization

The migration assistant has begun the process of creating your read replica and importing the SQL dump file, and will synchronize with your source instance.

Next, make sure your primary instance will accept connections from the Cloud SQL read replica's IP address.

**i** Not sure if your replica is able to connect? Check the [replication setup logs](#).

#### Primary

Name	rds-sin-1
Host name and port	[redacted]:3306
Username	repl
Password	..... 
Database version	MySQL 5.7
Location	asia-east2

#### Replica

Name	rds-sin-1-cloudsql-replica
Outgoing IP Address	[redacted]
Location	asia-east2
SSL	No SSL certs specified
Tier	db-n1-standard-1

Click Next to confirm that your primary instance will accept connections from the Cloud SQL read replica's IP address.

Next

如果实例加载数据成功，在Cloud SQL实例列表里应该有绿色图标前缀。

Google Cloud Platform youzhi-lab						
SQL Instances + CREATE INSTANCE MIGRATE DATA DELETE SELECTED INSTANCES (INTERNAL)						
Filter tree						
Instance ID	Type	Public IP address	Private IP address	Instance connection name	High availability	
<input type="checkbox"/> rds-sin-1	MySQL external primary	[redacted]:3306			N/A	
<input checked="" type="checkbox"/> rds-sin-1-cloudsql-replica	MySQL read replica	[redacted] ?		youzhi-lab:asia-east...	N/A	

在实例概况界面，选择Connect using Cloud Shell，可以在下部展开Cloud Shell命令行，进行快速实例登录测试。

The screenshot displays the Google Cloud SQL console interface. On the left, a sidebar lists navigation options: Overview, Connections, Users, Databases, Operations, and Corp Access. The main content area is titled 'Overview' and includes buttons for EDIT, EXPORT, RESTART, and DELETE. A graph shows CPU utilization for the instance 'rds-sin-1-cloudsql-replica'. Below the graph, the 'Connect to this instance' section provides fields for Public IP address, Outgoing IP address, and Instance connection name. The Instance connection name field is populated with 'youzhi-lab:asia-east2:rds-sin-1-cloudsql-replica'. Three connection methods are listed: 'Connect using Cloud Shell', 'Connect from a Compute Engine VM instance', and 'See all connection methods'. At the bottom, a 'CLOUD SHELL' terminal window is open, showing the command 'gcloud sql connect rds-sin-1-cloudsql-replica --user=root --quiet' being executed.

```
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to youzhi-lab.
Use "gcloud config set project [PROJECT ID]" to change to a different project.
eugeneyu@cloudshell:~ (youzhi-lab)$ gcloud sql connect rds-sin-1-cloudsql-replica --user=root --quiet
Whitelisting your IP for incoming connection for 5 minutes...#
```

当Cloud Shell中提示要求输入root的密码时，输入回车，因为root用户初始密码为空。登录数据库后，执行SQL命令，确认可以看到从RDS实例导入的数据。



CLOUD SHELL

Terminal

(youzhi-lab) x + ▾

affiliates. Other names may be trademarks of their respective owners.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

```
mysql> SHOW DATABASES;
```

Database
information_schema
guestbook
mysql
performance_schema
sys

5 rows in set (0.02 sec)

```
mysql> USE guestbook;
```

Reading table information for completion of table and column names  
You can turn off this feature to get a quicker startup with -A

Database changed

```
mysql> SELECT * FROM entries;
```

guestName	content	entryID
first guest	I got here!	1
second guest	Me too!	2

2 rows in set (0.01 sec)

```
mysql> █
```

为了查看数据同步正常，可以回到RDS实例执行下面数据插入命令。

```
INSERT INTO entries (guestName, content) values ("Eugene", "Nice place!");
```

之后在Cloud Shell里与谷歌云数据库副本的会话中查询数据，确认新增数据同步成功。

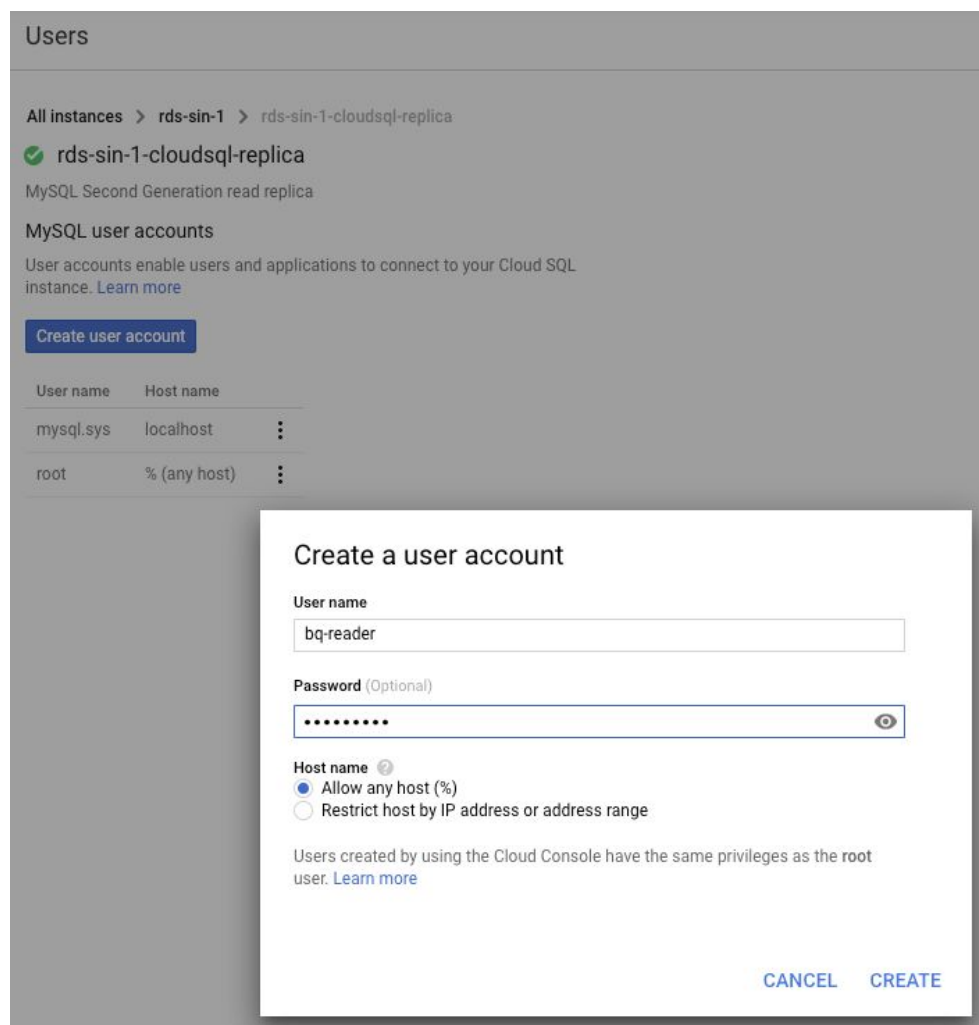
```
mysql> SELECT * FROM entries;
```

guestName	content	entryID
first guest	I got here!	1
second guest	Me too!	2
Eugene	Nice place!	3

3 rows in set (0.02 sec)

# 在BigQuery中执行联合查询

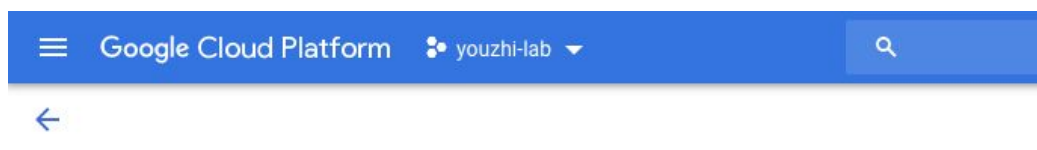
首先在Cloud SQL控制台给数据库副本创建一个代表BigQuery进行查询的用户。



然后点击下面的链接，到谷歌云控制台API Dashboard里开启BigQuery Connection API。

[BigQuery connection API](#)

点击ENABLE按钮开启。等待几分钟开启完成。



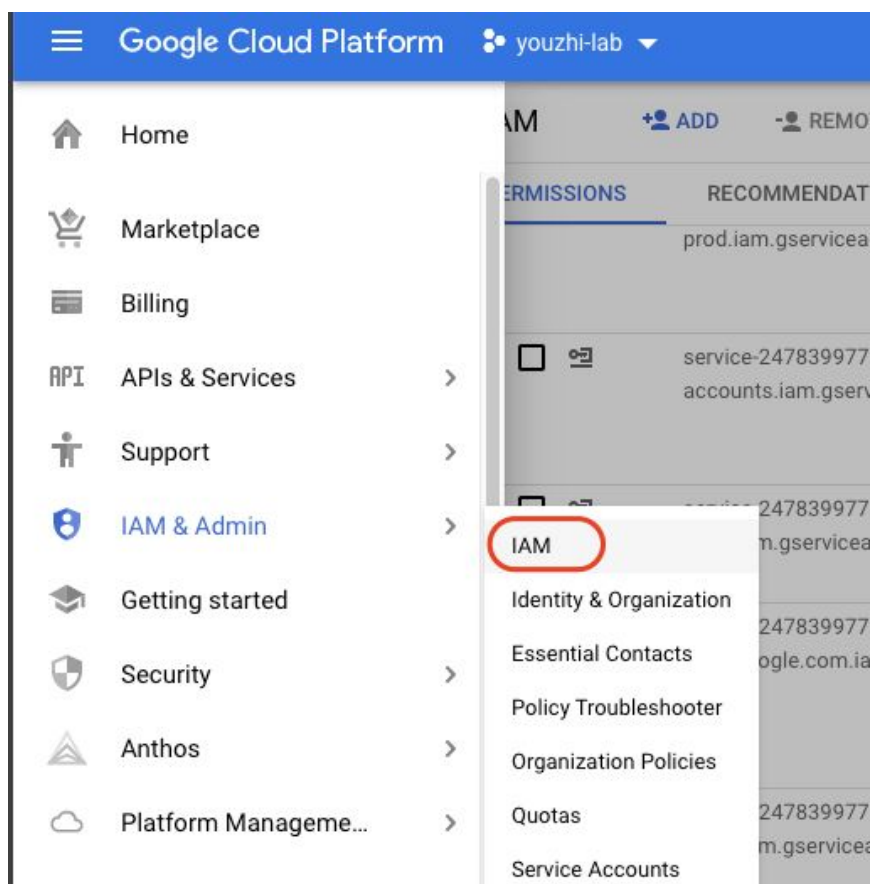
## BigQuery Connection API

Google

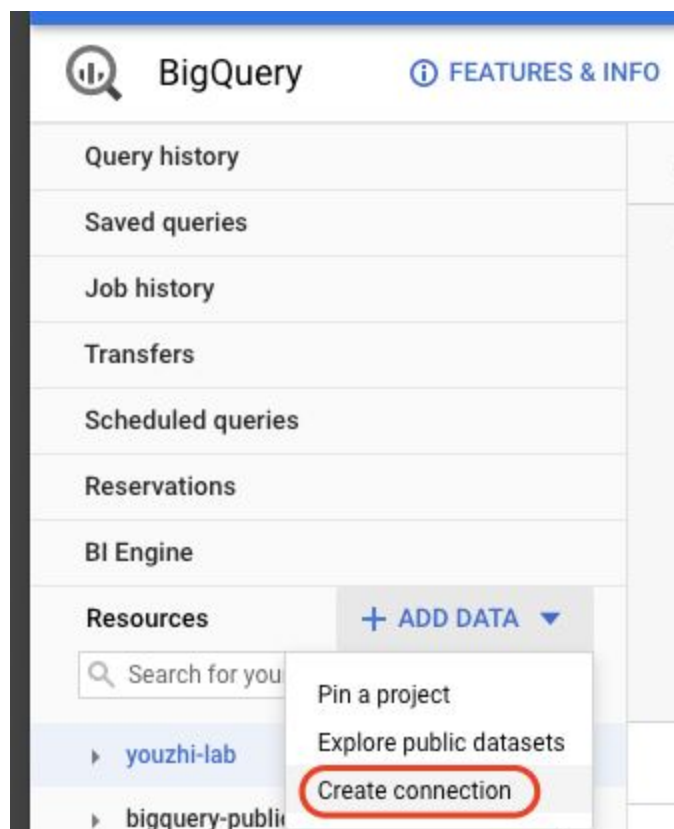
Allows users to manage BigQuery connections to external data sources.



确保你的控制台用户有`bigquery.admin`角色。如果没有，则到IAM配置页面配置一下。如果已经是Project Owner则不需要配置。



现在需要给BigQuery创建一个指向Cloud SQL副本的连接。在BigQuery控制台，点击ADD DATA里面的Create connection。



填写连接信息。选择与BigQuery数据集相同的区域。Cloud SQL Instance ID可以在数据库副本的详情页面获取。


## Create connection

Connection type

Cloud SQL - MySQL

Connection ID

bq-conn-rds-sin-1-replica

Connection location 

Hong Kong (asia-east2)

Friendly name (Optional)

Description (Optional)

Cloud SQL instance ID 

youzhi-lab:asia-east2:rds-sin-1-cloudsql-replica

Database name

guestbook

Username

bq-reader

Password

.....

☐ Show password

连接建立好后，可以立即执行BigQuery查询，确认连接工作正常。可以在BigQuery的查询窗口执行类似下面指令的查询。

```
SELECT * FROM
EXTERNAL_QUERY("youzhi-lab:asia-east2.bq-conn-rds-sin-1-replica", "SELECT *
FROM entries;");
```



Query editor

+ COMPOSE NEW QUERY

1

```
SELECT * FROM EXTERNAL_QUERY("youzhi-lab.asia-east2.bq-conn-rds-sin-1-replica", "SELECT * FROM entries;");
```

Processing location: asia-east2

Run

Save query

Save view

Schedule query

More

Query results

SAVE RESULTS

EXPLORE DATA

Query complete (0.7 sec elapsed, 61 B processed)

Job information

Results

JSON

Execution details

Row	guestName	content	entryID
1	first guest	I got here!	1
2	second guest	Me too!	2
3	Eugene	Nice place!	3

如果执行成功，返回正确数据，则说明连接工作正常。然后可以创建更复杂的查询，利用BigQuery的强大功能合并查询不同来源的数据。一些示例可以参考附录[\[2\]](#)。

## 附录

[1] [Replicating from an external server \(external master\)](#)

[2] [Cloud SQL federated queries](#)