Google Cloud

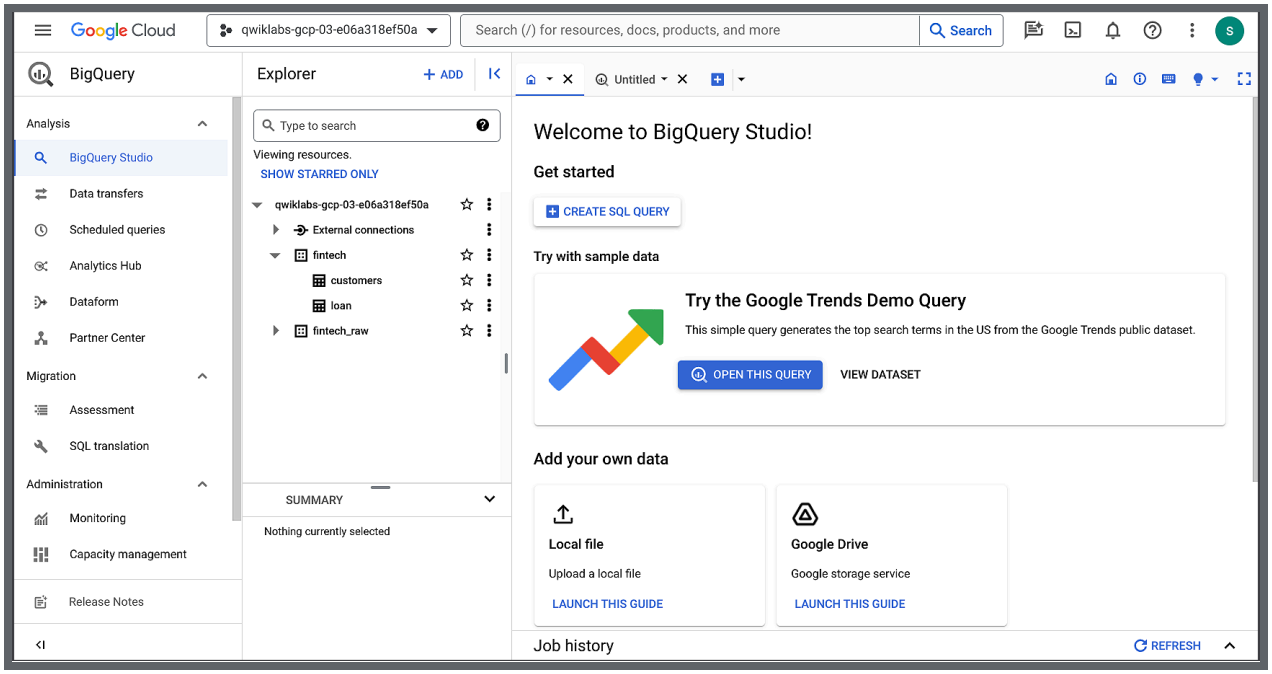
Solution Guide: Collect, process, and store data in BigQuery

The **Collect, process, and store data in Big Query lab** is a portion of the capstone project that puts your data analysis skills to the test; the lab includes a set of tasks and challenges that involve transforming data within a business scenario. Each task in the lab guides you to apply the skills you learned throughout the course, focusing on data collection, processing, and storage within the BigQuery environment. The lab also requires you to tackle two challenges to assess your skills on your own: transform data and create a report.

This solution guide provides the results of each guided task in the lab for you to assess against your own work. It also includes the solution query and results for the two challenges so that you may evaluate your approach, as well as identify potential areas for improvement.

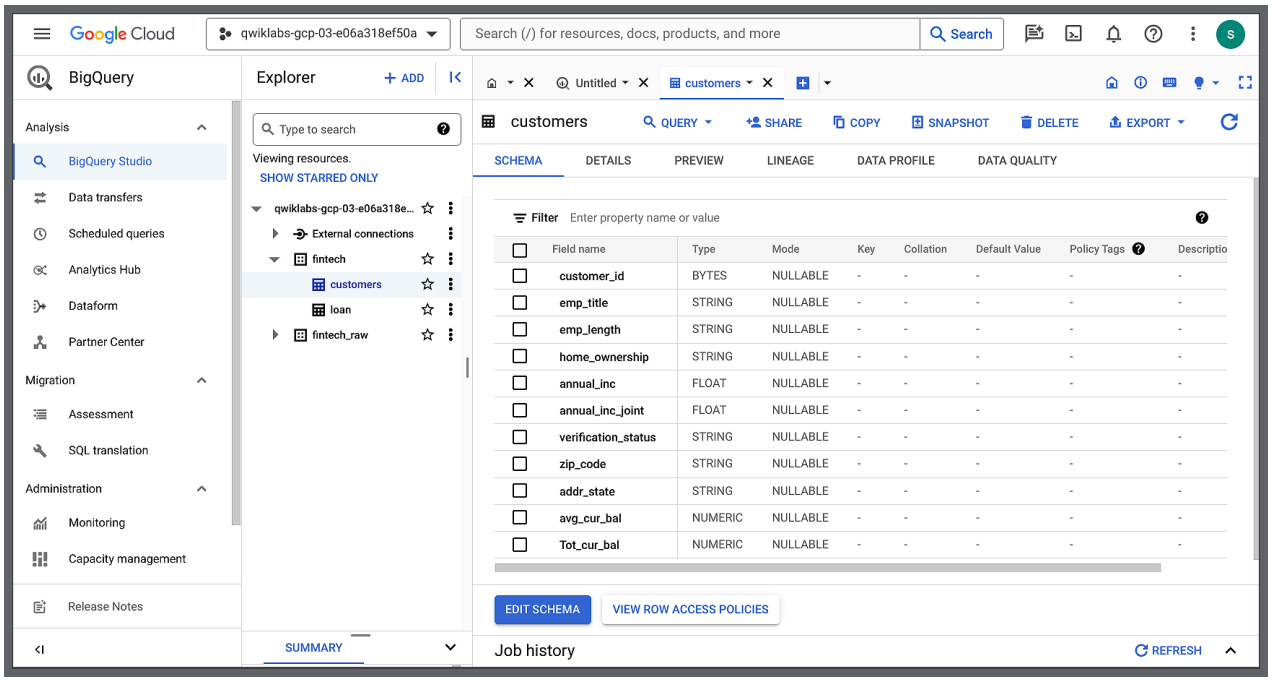
Task 1: Get started with BigQuery

To complete this task, open up the BigQuery environment. Select the project that matches the **Google Cloud project ID** provided during login, and locate the fintech dataset in the **Explorer** pane.



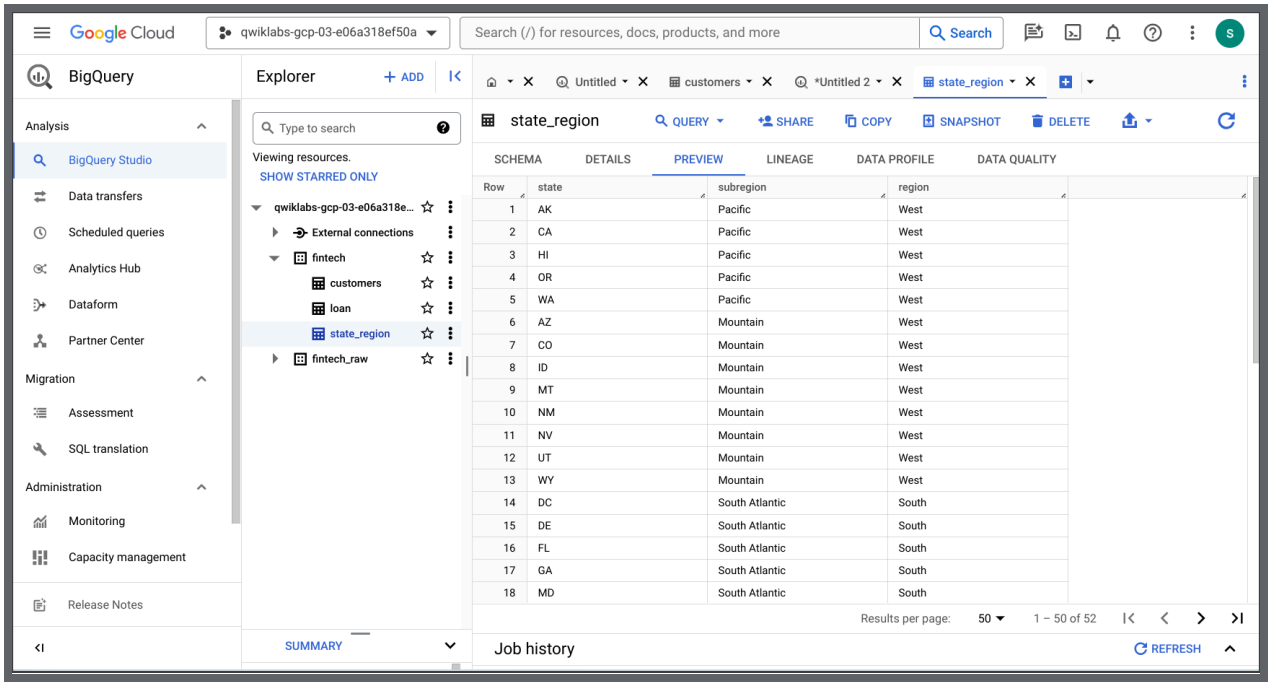
Task 2: Explore the Fintech data

To complete this task, expand the fintech dataset to view the customers and loans table. Then, click on each name to select the table and review the **Details, Preview,** and **Schema** tabs.



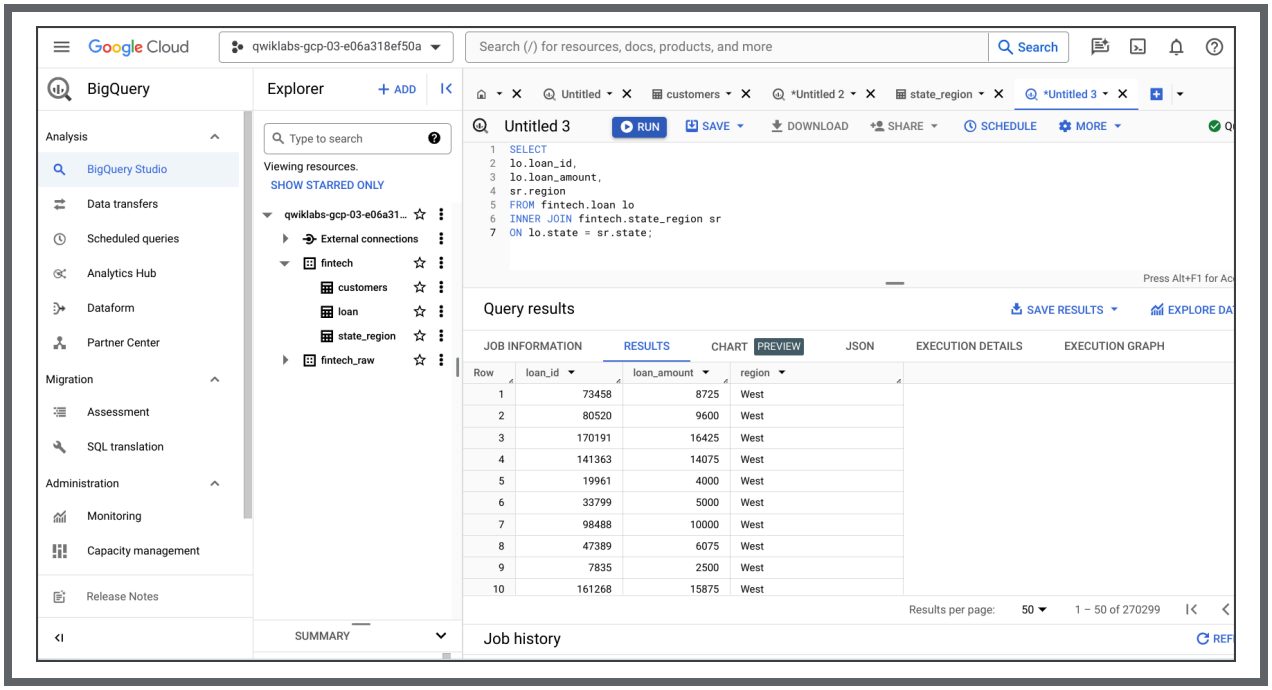
Task 3: Import a CSV file and create a standard table

To complete this task, run the provided code in the query editor to import a CSV file from Cloud Storage. Then, review the table using the preview tab. A new table called state_region will be added to the fintech dataset.



Task 4: Join data from two tables

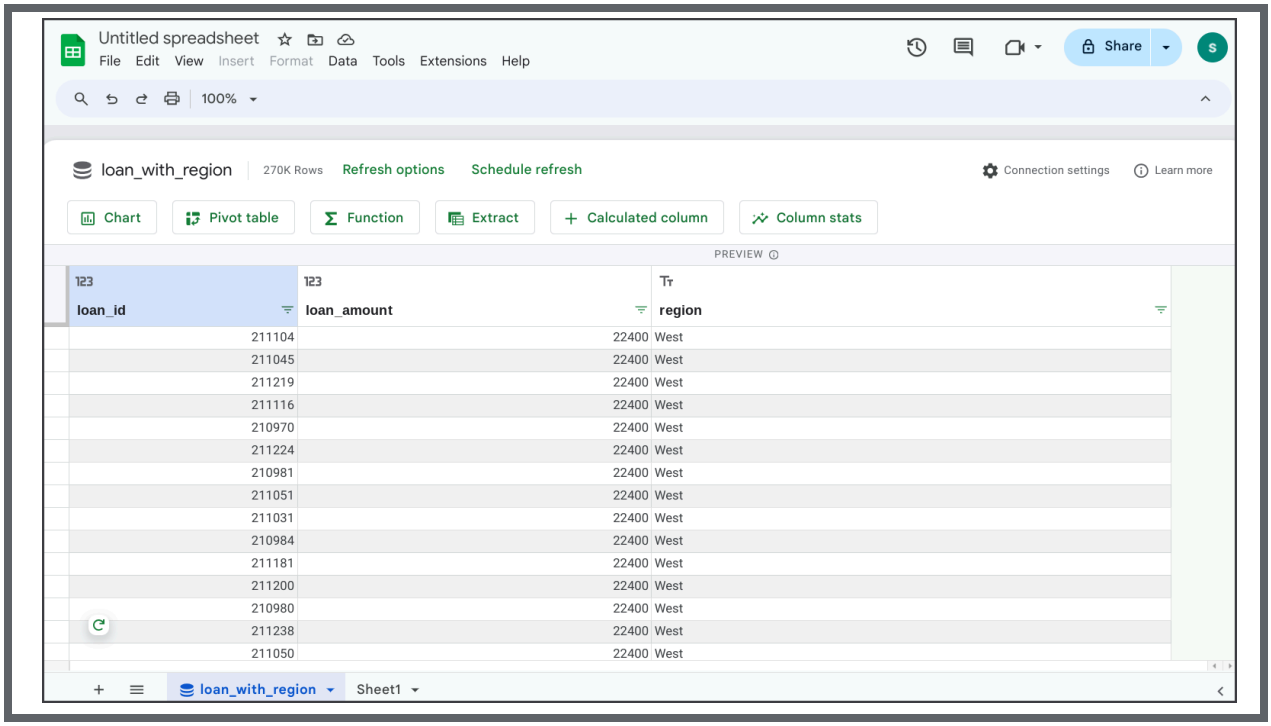
To complete this task, run **Query B** in the **Query Editor** to join the two tables and review the results in the **Query results** panel.



Task 5: Create a table based on the results of a query using CTAS

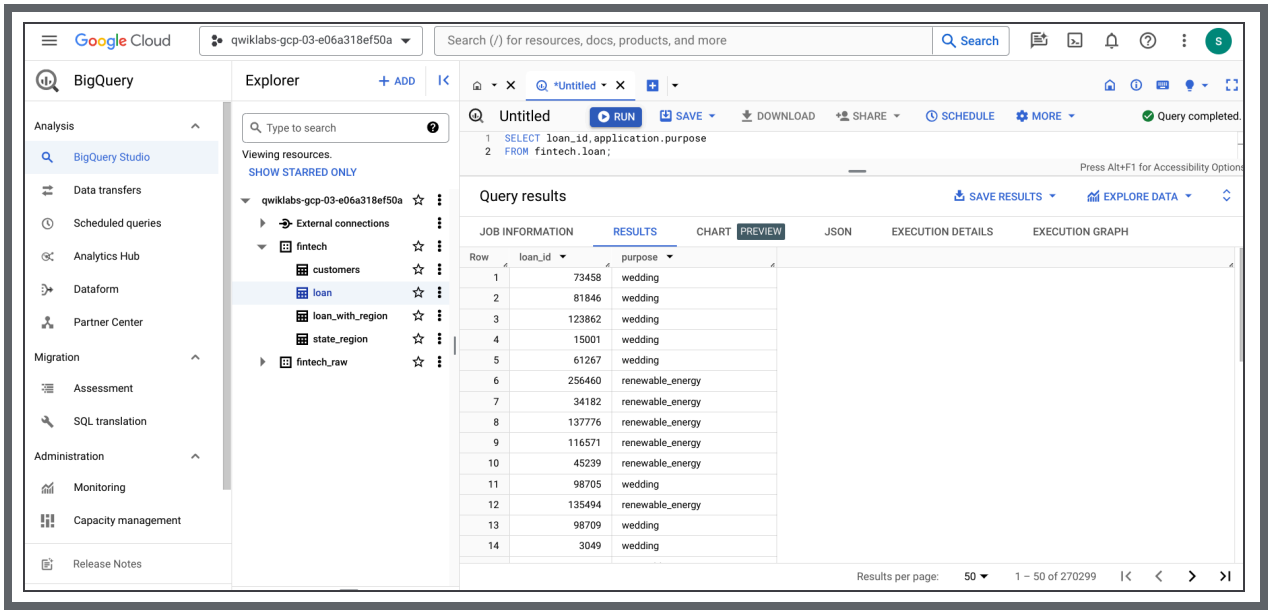
To complete this task, run the provided query in the **Query Editor** to create a new table named loan_with_region. The new table appears in the fintech dataset. Then, export the data to Google Sheets to review the loan_with_region data.

Note: If the export to Google Sheets fails, an error will appear stating that Google Sheets will not open. Try exporting the data again.



Task 6: Work with nested data

To complete this task, use dot notation to query the purpose column, which is nested inside of the application record. The results of the query will be a table with two columns: loan_id and purpose.



Task 7 challenge: Deduplicate

To complete this challenge, write a query to create a table named fintech.loan_purposes that has a single column named purpose. The purpose column should include the unique results found in the nested purpose column in the loan table of the fintech dataset.

Solution

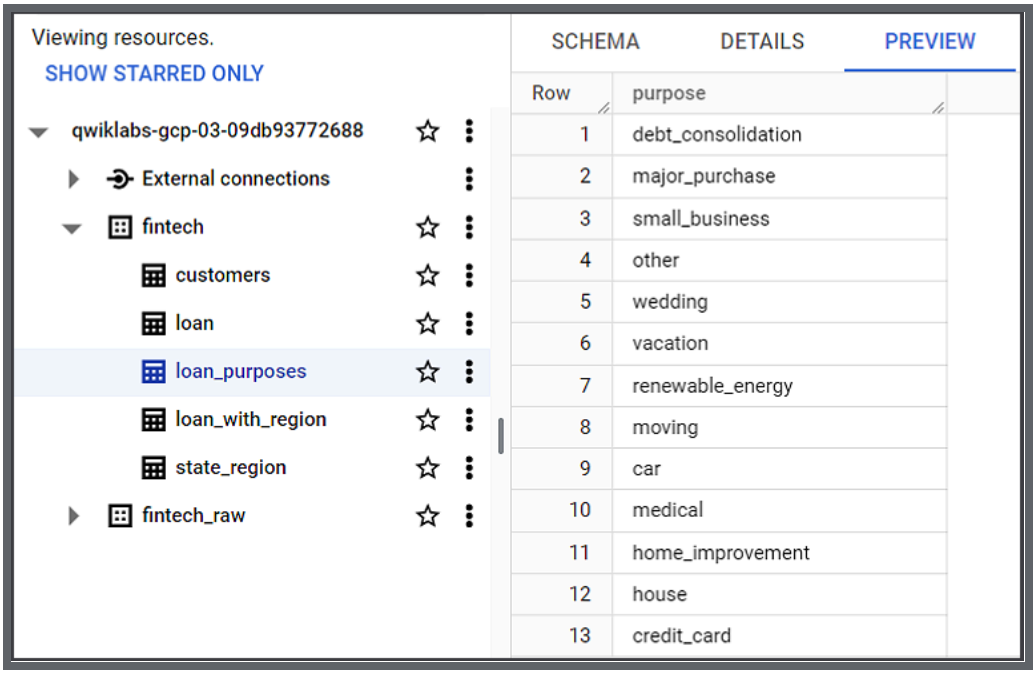
You can use a Create Table as Select (CTAS) statement to create the table and dot notation to select the purpose column that is nested in the application record.

Here's the query:

```
CREATE TABLE fintech.loan_purposes AS
SELECT DISTINCT application.purpose
FROM fintech.loan;
```

1. Copy and paste the above query into the **Query Editor.**
2. Click **Run.**

As a result of this query, a new table named loan_purposes is added to the fintech dataset. This table has one column that selects the distinct values from the purpose column within the application record of the loan table.



Task 8 challenge: Answer business questions with a report

To complete this challenge, write a query to create a table called loan_count_by_year in the fintech dataset that counts loans grouped by issue_year.

Solution

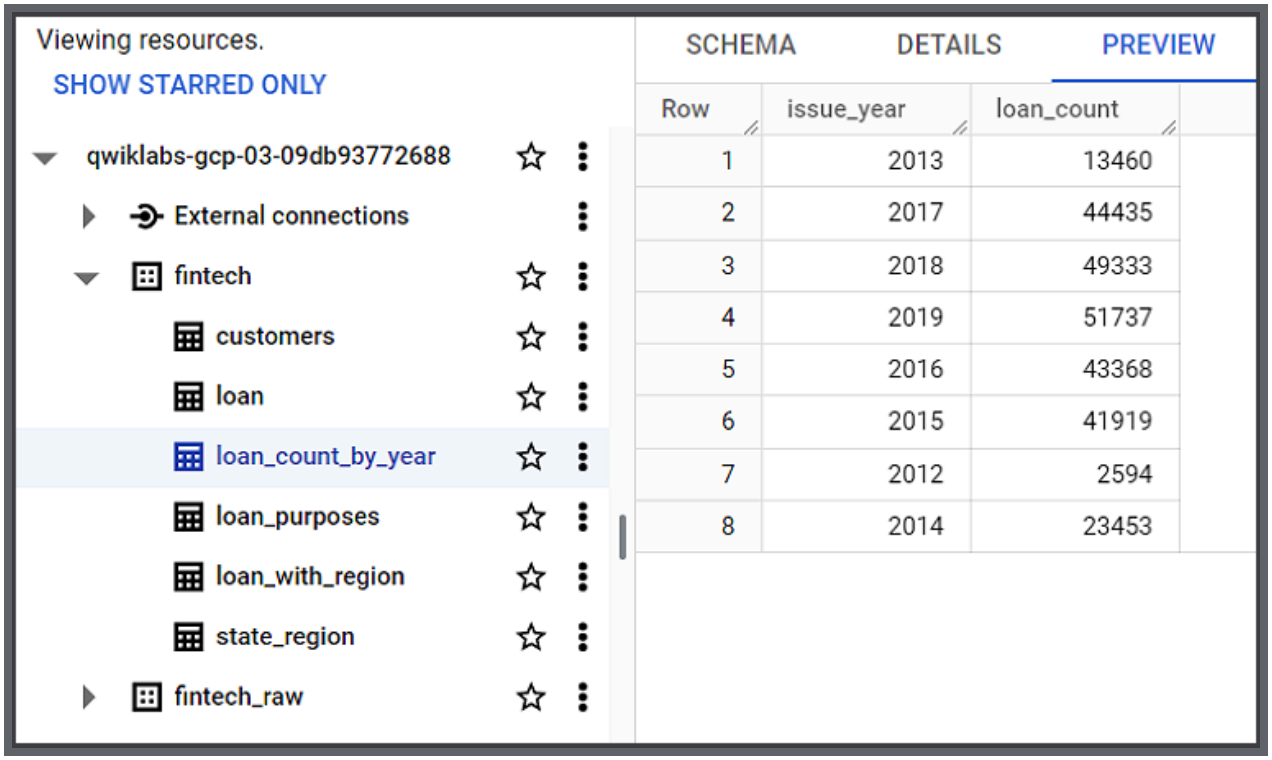
You can use a Create Table as Select (CTAS) statement to create the table and **COUNT** and **GROUP BY** to count the loans and group them by issue_year.

Here's the query:

```
Unset
CREATE TABLE fintech.loan_count_by_year AS
SELECT issue_year, count(loan_id) AS loan_count
FROM fintech.loan
GROUP BY issue_year;
```

1. Copy and paste the above query into the **Query Editor.**
2. Click **Run.**

As a result of this query, a new table named loan_count_by_year is added to the fintech dataset. This table has two columns: issue_year and loan_count. The loan_count column counts the number of loans by issue year.



Resources for more information

Use these readings to help support you as you work through the solution:

- **SQL query terms reading** available in course 1 module 1
- **Guide to BigQuery reading** available in course 2 module 1