

Creating a new BigQuery dataset

This is a guide for uploading existing data into BigQuery. If you have any questions, please email patents-public-data@googlegroups.com or see <https://cloud.google.com/bigquery/docs/loading-data>.

Project

You will need to [create a Google Cloud Platform account](#) to upload data into BigQuery. The project name you pick will become part of your BigQuery table name: "**project-name**:dataset-name.table-name".

Content

Type	Example
DOCDB publication number	US-1705778-A
DOCDB application number	US-18159027-A
DOCDB family id	22664921
INCHI key	OKKDEIYWILRZIA-OSZBKLCCSA-N
SMILES	<chem>C=CCC(=O)CC=C</chem>

For publication and application numbers, we have a free API for taking messy numbers and returning DOCDB-format numbers:

<https://patents.google.com/api/match?num=US800000&type=pub>

<https://patents.google.com/api/match?num=800000&type=pub&country=us,ep>

<https://patents.google.com/api/match?num=US800000&num=US7000000&num=US12345678&type=pub>

<https://patents.google.com/api/match?num=US800000&type=app>

For small one-time CSV datasets, we have a web interface for matching:

<https://patents.google.com/import>

Content can be in CSV or newline-separated JSON format. CSV is the most common and easiest to generate. JSON is only required if your table needs nested fields or arrays. There is a per-record (row) size limit of 100 MB.

PostgreSQL

```
COPY <table> TO '/home/me/table.csv' DELIMITER ',' CSV HEADER;
```

For big tables, this may work to stream the result directly into Google Cloud Storage:

```
COPY <table> TO PROGRAM 'split -b4G --filter="gzip-f | gsutil cp -  
gs://my-bigquery-data-bucket/table/$FILE.csv.gz";
```

SQLite

```
$ sqlite database.db  
.mode csv  
.headers on  
.out /home/me/table.csv  
select * from <table>;
```

Uploading

If your table is small (<10 MB), you can upload it directly through the BigQuery interface as part of the table creation step (skip to the table creation section).

If it is under ??GB, you can upload it to Google Drive and then select it in the BigQuery interface as part of the table creation step (skip to the table creation section).

If it's larger, you'll need to upload your data to a Google Cloud Storage bucket. There are two ways: an automatic CSV upload tool that will first load your data into Google Cloud Storage and then into BigQuery, or you can manually upload into Google Cloud Storage and use the BigQuery interface to create your new table.

Upload tool

https://github.com/google/patents-public-data/blob/master/tools/csv_upload.pysh

You will need to install the [BigQuery command-line tool](#) and run `bq init` to get credentials, and you need to install the Python "[sh](#)" package with `pip install sh`.

Single file, single table:

```
python3 csv_upload.pysh --source '~/Downloads/table.csv' --tables=jefferson-1790:dataset.table
```

Multiple files, single table:

```
python3 csv_upload.pysh --source '~/Downloads/table_*.csv'
--tables=jefferson-1790:dataset.table
```

Multiple files per table, multiple tables:

```
python3 csv_upload.pysh --source '~/Downloads/patstat/Data/{}_part*.txt'
--tables=jefferson-1790:epo_patstat.{}
table1_part00.txt, table1_part01.txt, ... -> jefferson-1790:epo_patstat.table1
table2_part00.txt -> jefferson-1790:epo_patstat.table2
```

Manual upload

Create a [new regional bucket](#) in any US region with any name. For this guide we'll pick "my-bigquery-data-bucket".

BigQuery has a per-file size limit of 4 GB (optionally compressed with gzip). Files larger than this must be split per row.

For single larger CSV files you can use the [Google Cloud Storage web interface](#) to [upload](#).

There is also a command-line utility called [gsutil](#) to upload your data.

If files are already < 4 GB:

```
gsutil cp /home/me/table.csv gs://my-bigquery-data-bucket/table/
```

For larger files, this splits a CSV into 4 GB chunks. Each of those chunks is gzipped (compressed) and uploaded to the bucket:

```
split -b4G --filter='gzip -f | gsutil cp - gs://my-bigquery-data-bucket/table/$FILE.csv.gz'
/home/me/table.csv
```

Table creation

GUI: Create a new [dataset](#), then create a new [table](#) inside that dataset. Point it to the data uploaded in the previous step, or select local files.

Tables should be created in the US region, not EU - only tables in the same country can be JOINed together with SQL. All existing public datasets are in the US.

CLI: You can also use the BigQuery [command line tool](#) to automate uploading of tables.

```
$ bq load my-project:dataset.table gs://my-bigquery-data-bucket/table/table.csv.*  
column1_name:string,column2_name:string,column3_name:integer
```

If your table is usually queried with a date restrict and is very large (>50GB), you can reduce the number of rows scanned by creating a partitioned table by date. You can also create a [clustered table](#) with an additional index column if your table is usually queried by date and by a string or number (country = "US").

Table updating

You can [append CSV records](#) onto an existing BigQuery table. [Updating existing records](#) is possible, but it is generally easier to do a full export of your original database and re-create the full table.

For real-time updates, [streaming new records](#) into BigQuery is also possible, but requires using a special API.

Sharing











Note: the sharing UI can only be accessed in the Classic UI. Each dataset can be [shared](#) with a set of email addresses, [Google Groups](#), and/or made accessible to the world ("allAuthenticatedUsers" string constant). This matches the sharing settings you would see in Google Docs.

patents-public-data:cpc

 CREATE TABLE

 SHARE DATASET

 DELETE DATASET

BigQuery Data Viewer (4 members)			
Access to view datasets and all its tables			
Type	Members 	Inherited	
	allAuthenticatedUsers		
	google-patents-public-datasets-partners@googlegroups.com		
	google.com		
	Viewers of project: patents-public-data		

Google Groups is a popular way to control access for a set of people without needing to add their email addresses to every individual dataset.

You would create a new group called "my-bigquery-users", then add the email "my-bigquery-users@googlegroups.com" as a reader of your datasets. Any user you add to that Google Group would have access to all of the datasets that my-bigquery-users@googlegroups.com has access to.

Enterprise environments can also look at [Google Cloud Identity and Access Management \(IAM\) roles for BigQuery](#). This can be combined with Active Directory or LDAP sync with [Google Cloud Directory Sync](#) so users in your existing systems can be given the correct access and managed automatically as the AD/LDAP source of truth updates.

Marketing pages

Once your public or commercial dataset is ready, Google can add a marketing page at <https://console.cloud.google.com/marketplace> so users can find your data.

<https://github.com/google/patents-public-data/tree/master/tables>

We are working on additional tools and reports to document links and join statistics between tables in the ecosystem. These documentation pages can also be generated for your private tables by running the code under your account.

https://github.com/google/patents-public-data/blob/master/tools/dataset_report.pysh