

Google Cloud Platform

Exporting Data

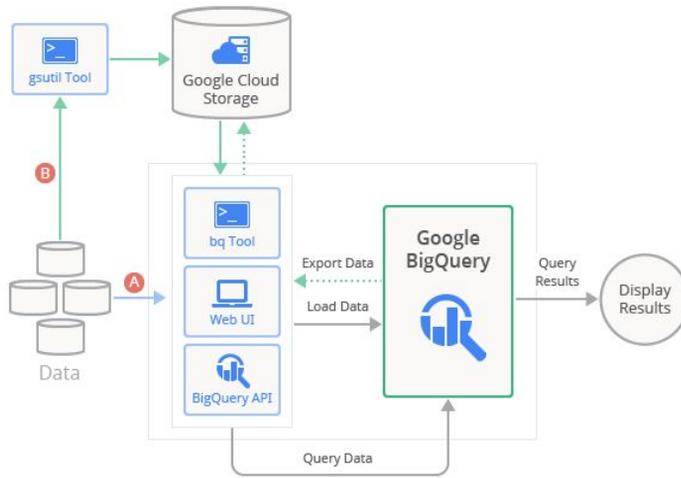
BigQuery for Data Analysts
V1.2

Approximate timing: 30 minutes

Agenda

- 1 Exporting Data
- 2 Running Export Jobs
- 3 Wildcard Exports
- 4 Quiz & Lab

Exporting Data (1 of 2)



Why export data?

- Using data with third-party tools
- Snapshots
- Backups

Export using:

- Web UI
- CLI
- REST API

Exporting Data (2 of 2)

- Limitations
 - Export up to 1 GB of data per file (multiple file export supported)
 - Daily limit: 1,000 exports per day, up to 10 TB
- ACL requirements for exporting data:

Product	Access
BigQuery	Dataset-level READER access
Google Cloud Storage	WRITE access to Google Cloud Storage bucket(s)

Notes:

BigQuery uses ACLs to manage permissions on [projects](#) and [datasets](#). ACLs are not directly supported on [tables](#). A table inherits its ACL from the dataset that contains it.

Export Configuration Options

Two aspects: Format and compression

- **destinationFormat**
 - **CSV**
 - JSON
 - Avro
- **compression**
 - GZIP
 - **NONE**

- **Notes:**

- AVRO cannot be used with GZIP compression
- Nested and repeated data cannot be exported in CSV format
- Defaults: CSV with no compression

Notes:

Default options are in bold.

Avro: Serialization format for persistent storage, from Apache. Primary use is Hadoop.

AVRO Export Format

- Exported files are **Avro container files**
- Each **row** is represented as an Avro record
 - Nested data is represented by nested record objects
- **REQUIRED** fields represented as corresponding Avro types
 - For example: An **INTEGER** type maps to an Avro **LONG** type
- **NULLABLE** fields represented as Avro **Union** of corresponding type and "null"
- **REPEATED** fields are represented as Avro **arrays**
- **TIMESTAMP** data types represented as Avro **LONG** types

Agenda

- 1 Exporting Data
- 2 Running Export Jobs
- 3 Wildcard Exports
- 4 Quiz & Lab

CLI

- `bq extract` - Perform an extract operation against `source_table` into `destination_uris`
 - `bq extract <source_table> <destination_uris>`

```
bq extract my_dataset.my_table  
gs://mybucket/myfilename.csv  
gsutil cp gs://mybucket/myfilename.csv .
```

Wildcard example

Web UI

Export to Google Storage ✕

Export format

Compression None GZIP ?

Google Cloud Storage URI ?

[View Files](#)

Configuration Example

JSON

```
jobData = {
  'projectId': projectId,
  'configuration': {
    'extract': {
      'sourceTable': {
        'projectId': projectId,
        'datasetId': datasetId,
        'tableId': tableId
      },
      'destinationUri': ['gs://<bucket>/<file>'],
      'destinationFormat': 'NEWLINE_DELIMITED_JSON'
    } ...
  }
}
```

Agenda

- 1 Exporting Data
- 2 Running Export Jobs
- 3 Wildcard Exports
- 4 Quiz & Lab

Wildcard Exports

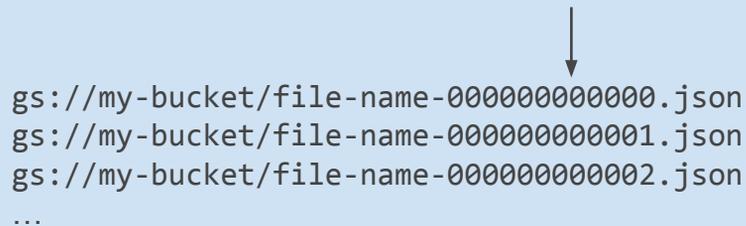
- If export is larger than 1GB, use a wildcard to partition the output into multiple files
- Include a glob character (*) in export file name
 - Glob is replaced by shard value of 12 digits
 - Starts with 000000000000 and increments by 1 for each file
- Wildcard exports are written in parallel
 - Target files are smaller and parallel writers work on separate patterns immediately
- Wildcard exports are subject to quota limitations

Single Wildcard URI

- **destinationUris** property indicates export location(s) and file name(s)
- Data sharded into multiple files based on the pattern

```
'destinationUris': ['gs://my-bucket/file-name-*.json']
```

Single wildcard



The diagram illustrates how a single wildcard URI pattern is expanded into multiple specific file paths. A downward-pointing arrow indicates the expansion from the wildcard pattern to the following list of files:

```
gs://my-bucket/file-name-000000000000.json  
gs://my-bucket/file-name-000000000001.json  
gs://my-bucket/file-name-000000000002.json  
...
```

Multiple Wildcard URIs

- **destinationUris** property indicates export locations and file names
- Data sharded into multiple files based on the pattern

```
'destinationUris': ['gs://my-bucket/file-name-1-*.json',  
'gs://my-bucket/file-name-2-*.json']
```

Multiple wildcards

↓

```
gs://my-bucket/file-name-1-000000000000.json  
gs://my-bucket/file-name-1-000000000001.json  
...  
gs://my-bucket/file-name-1-000000000080.json  
gs://my-bucket/file-name-2-000000000000.json  
gs://my-bucket/file-name-2-000000000001.json  
...  
gs://my-bucket/file-name-2-000000000080.json
```

Notes:

Use multiple wildcard URIs if you want to partition the export output. You would use this option if you're running a parallel processing job with a service like Hadoop on Google Cloud Platform. Determine how many workers that are available to process the job, and create one URI per worker. BigQuery treats each URI location as a partition, and uses parallel processing to shard your data into multiple files in each location. You can use whatever pattern you'd like in your file name, assuming there is a single wildcard operator in each URI, each URI is unique, and the number of URIs does not exceed the quota policy.

When you pass more than one wildcard URI, BigQuery creates a special file at the end of each partition that indicates the "final" file in the set. This file name indicates how many shards BigQuery created.

For example, if your wildcard URI is `gs://my-bucket/file-name-<worker number>*.json`, and BigQuery creates 80 sharded files, the zero record file name is `gs://my-bucket/file-name-<worker number>-000000000080.json`. You can use this file name to determine that BigQuery created 80 sharded files (named `000000000000-000000000079`).

Note that a zero record file might contain more than 0 bytes depending on the data format, such as when exporting data in CSV format with a column header.

Agenda

- 1 Exporting Data
- 2 Running Export Jobs
- 3 Wildcard Exports
- 4 Quiz & Lab

Module Review (1 of 2)

Which of the following are supported data formats when exporting data from BigQuery?

(select 3 of the available options)

- XML
- Avro
- PDF
- CSV
- JSON

Module Review (2 of 2)

Which one of the following is *not* true regarding wildcard exports?

- The filename must contain a glob character (*)
- Wildcards must be used if the data being exported is larger than 1GB
- Wildcard exports are not subject to quotas
- Data is exported in parallel
- Filenames are incremented

Lab

Exporting data from BigQuery

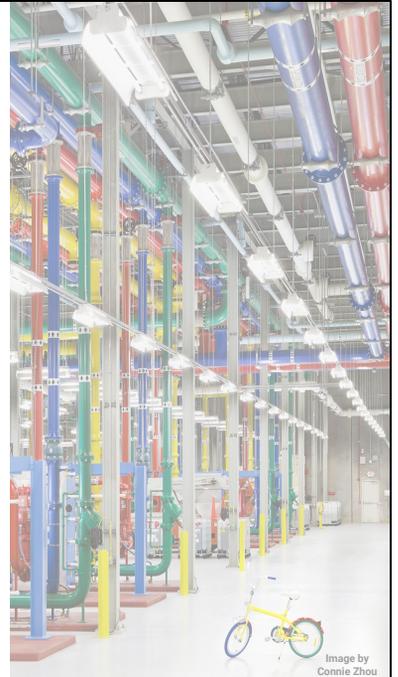


Image by
Cennie Zhou

Resources

- Exporting data from BigQuery

<https://cloud.google.com/bigquery/exporting-data-from-bigquery>

Module Review Answers (1 of 2)

Which of the following are supported data formats when exporting data from BigQuery?

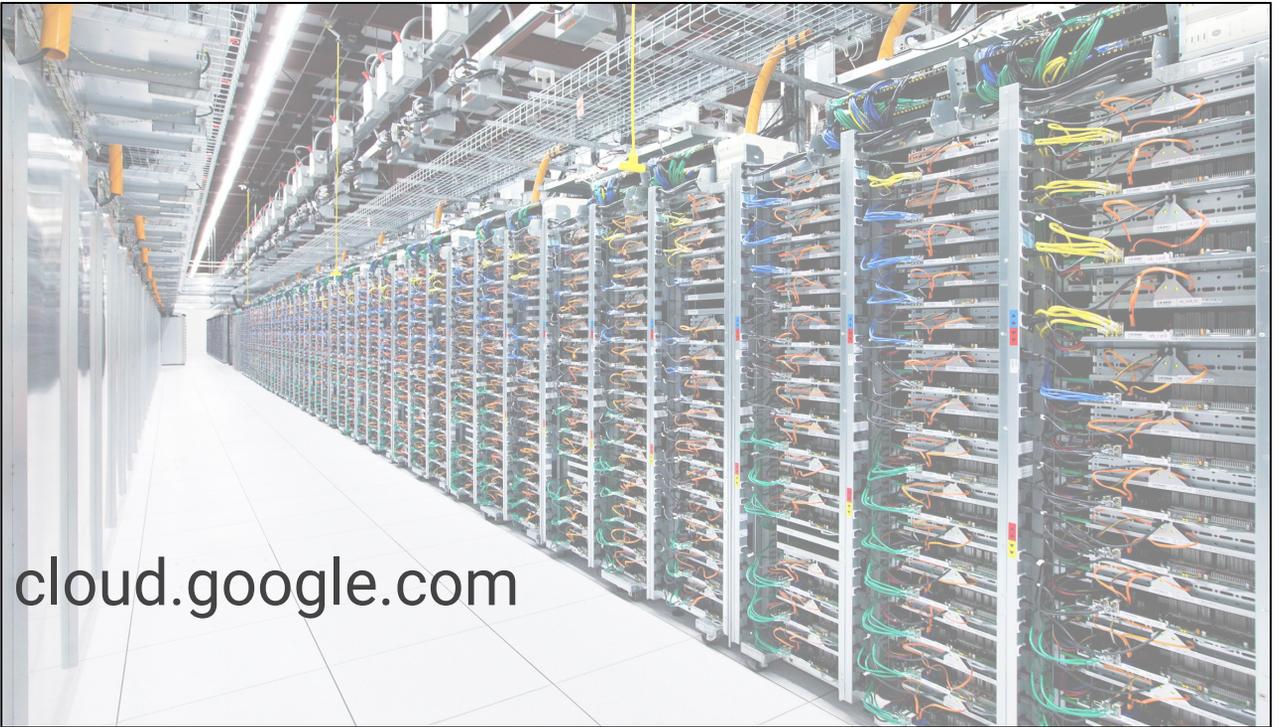
(select 3 of the available options)

- XML
- Avro
- PDF
- CSV
- JSON

Module Review Answers (2 of 2)

Which one of the following is *not* true regarding wildcard exports?

- The filename must contain a glob character (*)
- Wildcards must be used if the data being exported is larger than 1GB
- Wildcard exports are not subject to quotas
- Data is exported in parallel
- Filenames are incremented



cloud.google.com