This page describes how to prepare Avro files that you exported from non-Cloud Spanner databases and then import those files into Cloud Spanner. If you want to import a Cloud Spanner database that you previously exported, see Importing Cloud Spanner Avro files (/spanner/docs/import).

The process uses Dataflow (/dataflow/); it imports data from a Cloud Storage (/storage/) bucket that contains a JSON manifest file and a set of Avro files (https://en.wikipedia.org/wiki/Apache_Avro).

To import a Cloud Spanner database, first you need to enable the Cloud Spanner, Cloud Storage, Compute Engine, and Dataflow APIs:

Enable the APIs (https://console.cloud.google.com/flows/enableapi?apiid=spanner.googleapis.com,storage_component,compute,dataflow)

You also need enough quota and the required Cloud IAM permissions.

The quota requirements for import jobs, by Google Cloud service, are as follows:

- **Cloud Spanner**: You must have enough nodes to support the amount of data that you are importing. No additional nodes are required to import a database, though you might need to add more nodes so that your job finishes in a reasonable amount of time. See Optimizing jobs (#optimize-slow) for more details.
- **Cloud Storage**: To import, you must have a bucket containing your previously exported files. You do not need to set a size for your bucket.
- **Dataflow**: Import jobs are subject to the same CPU, disk usage, and IP address Compute Engine quotas (/dataflow/quotas#compute-engine-quotas) as other Dataflow jobs.
- **Compute Engine**: Before running your import job, you must set up initial quotas (https://support.google.com/cloud/answer/6075746) for Compute Engine, which Dataflow uses. These quotas represent the *maximum* number of resources that you allow Dataflow to use for your job. Recommended starting values are:
  - **CPUs**: 200
  - **In-use IP addresses**: 200
  - **Standard persistent disk**: 50 TB

  Generally, you do not have to make any other adjustments. Dataflow provides autoscaling so that you only pay for the actual resources used during the import. If your job can make use of more resources, the Dataflow UI displays a warning icon. The job should finish even if there is a warning icon.

To import a database, you also need to have Cloud IAM roles with sufficient permissions to use all of the services involved in an import job. For information on granting roles and permissions, see Applying IAM roles (/spanner/docs/grant-permissions).

To import a database, you need the following roles:

- At the Google Cloud project level:
  - Cloud Spanner Viewer
  - Dataflow Admin
  - Storage Admin
- At the Cloud Spanner database or instance level, or at the Google Cloud project level:
  - Cloud Spanner Reader
  - Cloud Spanner Database Admin (required only for import jobs)

The import process brings data in from Avro files located in a Cloud Storage bucket. You can export data in Avro format from any source and can use any available method to do so.

Keep the following things in mind when exporting your data:

- You can export using any of the Avro primitive types (https://avro.apache.org/docs/current/spec.html#schema_primitive) as well as the array (https://avro.apache.org/docs/current/spec.html#Arrays) complex type.
- Each column in your Avro files must use one of the following column types:
  - `ARRAY`
  - `BOOL`
  - `BYTES`
  - `DOUBLE`
  - `FLOAT`
  - `INT`
  - `LONG`*
  - `STRING`*

  *You can import a `LONG` storing a timestamp or a `STRING` storing a timestamp as a Cloud Spanner `TIMESTAMP`; see recommended mappings (#recommended-map) below for details.
- You do not have to include or generate any metadata when you export the Avro files.
- You do not have to follow any particular naming convention for your files.

If you do not export your files directly to Cloud Storage, you must upload the Avro files to a Cloud Storage bucket. For detailed instructions, see Uploading objects (/storage/docs/uploading-objects).

Before you run your import, you must create the target table in Cloud Spanner and define its schema.

You must create a schema that uses the appropriate column type for each column in the Avro files.

Avro column type	Cloud Spanner column type
`ARRAY`	`ARRAY`
`BOOL`	`BOOL`
`BYTES`	`BYTES`
`DOUBLE`	`FLOAT64`
`FLOAT`	`FLOAT64`
`INT`	`INT64`
`LONG`	`INT64` `TIMESTAMP` (when `LONG` represents a timestamp of the number of microseconds since 1970-01-01 00:00:00 UTC)
`STRING`	`STRING` `TIMESTAMP` (when `STRING` represents a timestamp in the canonical format for SQL queries (/spanner/docs/data-types#canonical-format_1))

If a column in your Avro data contains `NULL` values, you must ensure that you make the corresponding column in your Cloud Spanner table nul

We recommend that you create secondary indexes after you have imported your data into Cloud Spanner instead of when you initially define the schema for the table.

You must also create a file named `spanner-export.json` in your Cloud Storage bucket. This file contains a `tables` array that lists the name and data file locations for each table.

The contents of the file have the following format:

Wildcards are not supported; you must write out all filenames in full.

To start your import job, follow the instructions for using the `gcloud` command-line tool to run a job with the Avro to Cloud Spanner template (/dataflow/docs/guides/templates/provided-batch#gcsavrotocloudspanner).

After you have started an import job, you can see details about the job (#view-dataflow-ui) in the Cloud Console.

After the import job is finished, add any necessary secondary indexes (/spanner/docs/secondary-indexes).

To avoid network egress charges (/storage/pricing#network-pricing), choose a region that overlaps with your Cloud Storage bucket's location. ing a region (#choose-region) below for more information.

You might want to choose a different region based on whether your Cloud Storage bucket uses a regional or multi-regional configuration. To avoid network egress charges (/storage/pricing#network-pricing), choose a region that overlaps with your Cloud Storage bucket's location.

If your Cloud Storage bucket location is regional (/storage/docs/bucket-locations#location-r), choose the same region for your import job if that region is available to take advantage of free network usage (/storage/pricing#network-buckets).

If the same region is not available, egress charges will apply. Refer to the Cloud Storage network egress (/storage/pricing#network-pricing) pricing to choose a region that will incur the lowest network egress charges.

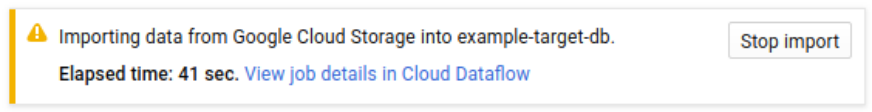
If your Cloud Storage bucket location is multi-regional (/storage/docs/bucket-locations#location-mr), choose one of the regions that make up the multi-regional location to take advantage of free network usage (/storage/pricing#network-buckets).

If an overlapping region is not available, egress charges will apply. Refer to the Cloud Storage network egress (/storage/pricing#network-pricing) pricing to choose a region that will incur the lowest network egress charges.

After you start an import job, you can view details of the job, including logs, in the Dataflow section of the Cloud Console.

To see details for a currently running job:

1. Navigate to the **Database details** page for the database.
2. Click **View job details in Dataflow** in the job status message, which looks similar to the following:

   The Cloud Console displays details of the Dataflow job.

To view a job that you ran recently:

1. Navigate to the **Database details** page for the database.
2. Click the **Import/Export** tab.
3. Click your job's name in the list.

   The Cloud Console displays details of the Dataflow job.

To view a job that you ran more than one week ago:

1. Go to the Dataflow jobs page in the Cloud Console.

   Go to the jobs page (https://console.cloud.google.com/dataflow)

2. Find your job in the list, then click its name.

   The Cloud Console displays details of the Dataflow job.

Jobs of the same type for the same database have the same name. You can tell jobs apart by the values in their **Start time** or **End time** column

To view a Dataflow job's logs, navigate to the job's details page as described above, then click **Logs** to the right of the job's name.

If a job fails, look for errors in the logs. If there are errors, the error count displays next to **Logs**:

To view job errors:

1. Click on the error count next to **Logs**.

   The Cloud Console displays the job's logs. You may need to scroll to see the errors.

2. Locate entries with the error icon
   .

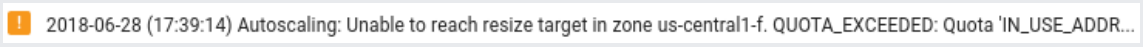
3. Click on an individual log entry to expand its contents.

For more information about troubleshooting Dataflow jobs, see Troubleshooting your pipeline (/dataflow/pipelines/troubleshooting-your-pipeline#basic-troubleshooting-workflow).

If you have followed the suggestions in initial settings (#quota), you should generally not have to make any other adjustments. If your job is running slowly, there are a few other optimizations you can try:

- **Optimize the job and data location**: Run your Dataflow job in the same region (#choose-region) where your Cloud Spanner instance and Cloud Storage bucket are located.
- **Ensure sufficient Dataflow resources**: If the relevant Compute Engine quotas (/dataflow/quotas#compute-engine-quotas) limit your Dataflow job's resources, the job's Dataflow page (#dataflow-job-details) in the Google Cloud Console displays a warning icon
  

  and log messages:

  In this situation, increasing the quotas (https://support.google.com/cloud/answer/6075746) for CPUs, in-use IP addresses, and standard persistent disk might shorten the run time of the job, but you might incur more Compute Engine charges.
- **Check the Cloud Spanner CPU utilization**: If you see that the CPU utilization for the instance is over 65%, you can increase the number of nodes in that instance. The extra nodes add more Cloud Spanner resources and the job should speed up, but you incur more Cloud Spanner charges.

Several factors influence the time it takes to complete an import job.

- **Cloud Spanner database size**: Processing more data takes more time and resources.
- **Cloud Spanner database schema (including indexes)**: The number of tables, the size of the rows, and the number of secondary indexes influence the time it takes to run an import job.
- **Data location**: Data is transferred between Cloud Spanner and Cloud Storage using Dataflow. Ideally all three components are located in the same region. If the components are not in the same region, moving the data across regions slows the job down.
- **Number of Dataflow workers**: By using autoscaling, Dataflow chooses the number of workers for the job depending on the amount of work that needs to be done. The number of workers will, however, be capped by the quotas for CPUs, in-use IP addresses, and standard persistent disk. The Dataflow UI displays a warning icon if it encounters quota caps. In this situation, progress is slower, but the job should still complete.
- **Existing load on Cloud Spanner**: An import job adds significant CPU load on a Cloud Spanner instance. If the instance already has a substantial existing load, then the job runs more slowly.
- **Number of Cloud Spanner nodes**: If the CPU utilization for the instance is over 65%, then the job runs more slowly.