

# Slalom Data Engineering

Assignment

# Serverless Event Driven Loads



1. CSV File lands in GCS

2. Cloud Function is Triggered

3. Data loaded to BQ

# Snapshot of File load on GCS

Google Cloud Platform slalom-de ▾ Search products and re

Storage Bucket details ←

Browser

Monitoring

Settings

slalom-de

OBJECTS CONFIGURATION PERMISSIONS RETENTION LIFECYCLE

Buckets > slalom-de > csv

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS DELETE

Filter Filter by object or folder name prefix

<input type="checkbox"/> Name	Size	Type	Created time
<input type="checkbox"/> reviews1.csv	4.9 MB	application/vnd.ms-excel	Oct 29, 2020, 3:23:12 PM

# Snapshot of Cloud Function

Google Cloud Platform slalom-de Search products and resources

Cloud Functions Function details EDIT DELETE COPY

load\_csv Version 21, deployed at Oct 31, 2020, 7:03:28 P...

METRICS DETAILS SOURCE VARIABLES TRIGGER PERMISSIONS LOGS TESTING

Runtime: Python 3.7 Entry point: load\_csv\_files

main.py requirements.txt

```
1 from google.cloud import bigquery
2
3 def load_csv_files(event, context):
4     """Triggered by a change to a Cloud Storage bucket.
5     Args:
6         event (dict): Event payload.
7         context (google.cloud.functions.Context): Metadata for the event.
8     """
9
10    file = event
11    uri = "gs://" + file['bucket'] + "/" + file['name']
12    print(f"Processing file: {uri}")
13
14    # Construct a BigQuery client object.
15    client = bigquery.Client()
16
17    # Set the BigQuery Table ID
18    table_id = "slalom-de.slalom.reviews"
19
20
21    job_config = bigquery.LoadJobConfig(
22        schema=[
23            bigquery.SchemaField("rec_id", "INTEGER"),
24            bigquery.SchemaField("BusinessId", "STRING"),
25            bigquery.SchemaField("Review_Date", "STRING"), #Had to Convert this to STRING due to BQ limitation
26            bigquery.SchemaField("Review_Id", "STRING"),
27            bigquery.SchemaField("Review_Stars", "INTEGER"),
28            bigquery.SchemaField("User_Id", "STRING"),
29            bigquery.SchemaField("Review_Votes_Cool", "INTEGER"),
30            bigquery.SchemaField("Review_Votes_Funny", "INTEGER"),
31            bigquery.SchemaField("Review_Votes_Useful", "INTEGER"),
32        ],
33        skip_leading_rows=1,
34
35    )
36
37    # Load the File if CSV
38    if uri.split(".")[-1] == "csv":
39
```

# Snapshot of Cloud Function

```
35     )
36
37     #Load the File if CSV
38     if uri.split(".")[1] == "csv":
39
40         load_job = client.load_table_from_uri(
41             uri, table_id, job_config=job_config
42         )
43         load_job.result() # Get Loaded Results
44         table = client.get_table(table_id)
45         print("Loaded {} rows to table {}".format(table.num_rows, table_id))
46     else:
47         print("File {} is not a CSV file".format(uri))
48     except:
49         print("Something went wrong")
50
```

# Cloud Function Dependency Packages

(...) Cloud Functions    [Edit function](#)

Configuration —  2 Code

Runtime  
Python 3.7

Source code  
[Inline Editor](#) ▾

+  
main.py ...  
requirements.txt  

```
1 #Function dependencies
2 google-cloud==0.34.0
3 google-cloud-bigquery==2.2.0
```

# Data Load in BigQuery

Google Cloud Platform slalom-de ▾ Search products and resources

BigQuery FEATURES & INFO SHORTCUT

Query history

Saved queries

Job history

Transfers

Scheduled queries

Reservations

BI Engine

Resources + ADD DATA

Search for your tables and datasets

slalom-de

+ slalom

reviews

Query editor

```
1 select * from slalom-de.slalom.reviews
```

Run Save query Save view Schedule query More

Query results SAVE RESULTS EXPLORE DATA

Query complete (1.8 sec elapsed, 6.6 MB processed)

Job information Results JSON Execution details

Row	rec_id	BusinessId	Review_Date	Review_Id	Review_Stars	User_Id	Review_Votes_Cool	Review_Votes_Funny	Review_Votes_Usedful
1	200	8_wUsDlOE8Guecq5RZZjDg	3/05/06	C_2bz3VFC9wd5lo53nfzg	1	8w7aq0Lnm0eyVbSXKsr7qw	0	0	0
2	204	9VZ_imQjvgpt3k6W6AQ_A	3/12/06	JMYWP-WXP4zbcdCTCYeubGg	1	e4TQFVfpzHf-hnBsjntg	0	1	0
3	332	g8ewG8o0ubTUTAUDj7UZLQ	7/21/06	bY3YhIKIAhHq_hREDJ-SDw	1	sXwvDlM0spjJ1zzr3cEM7Q	1	1	0
4	634	-98noPZylsH6NndWmdtuDw	12/27/06	rr0E_aFNHS9BB1nXIT58Gw	1	rPGZtaVjRoVI3Gysbs62cg	1	1	0
5	1056	-JL0CWSLkkzFHZXlctnf8A	3/7/07	qez4YU140U1MHCRILHu5w	1	7GC9fVWKa4a1ZmBGLH6Uww	0	0	0
6	1153	AOTOKHGbEorQSWr-KVmHzw	3/15/07	yRevOH81XwWjkkzk0BnbHg	1	7GC9fVWKa4a1ZmBGLH6Uww	0	0	0
7	1194	_tkAI5Q5XQSfbqjDkSDQ	3/20/07	eeFBEPs11KJz9q6v4kQxPg	1	SwnLshf0NJBKR9UiGA80Q	0	1	0
8	1451	r677iLEb5K1nnacFCEQ0jQ	4/29/07	zqxabwhPZRxs8yruoSQbsA	1	x3zv07BzR9cmauLjyp7qXQ	0	3	0
9	1487	uR2aNW75R4oYs9w7aw_-kQ	5/5/07	8zJ_ueMdFYEfDsGQATBBQ	1	NgS8s99qQliHOszKd-m87g	0	2	0
10	1592	7_zL7NX_rDFwhbLp98PwZg	5/24/07	2GaZ2YzsybOOC0ZLX5JMMWA	1	APLIPf1Rf8QyhHHk2uAyA	0	3	0
11	2221	nQOfSLvkN6GrBCYKsC0vYg	8/12/07	vCh7HNz2udKczlGz1myw	1	8RzsGBVrqLWLCE8sOSYQkw	2	1	0
12	2369	zxqU415r_RtZRKDtdbIKQ	8/26/07	tLOErJsSAMYgLugCnzAbwg	1	POHQEtLi_u1Mkysi7Hn0DQ	0	0	0
13	2584	9VZ_imQjvgpt3k6W6AQ_A	9/17/07	YCskBX190g96tUNXbt0cDw	1	Db5IKqe1M2iP5EnZNdroRQ	0	2	0
14	2759	gBVytpilYGLB1wce-Dfbxg	10/6/07	4ASp4Xns5bHzn_wz07kN0g	1	JnQYW_cYENygvYLHiy_MSw	0	0	0
15	2905	wWtUrFcweGXAWsI_lpfjw	10/22/07	w2T9jZtee6EAvivZ1HSNQ	1	5E193w76lho1gkyakW2A	1	1	0

Rows per page: 100 1 - 100 of 57153

# StackDriver/Google Cloud Operations

Google Cloud Platform slalom-de

Search products and resources

Last 1 hour Page Lay

Operations Logging

Logs Explorer OPTIONS REFINE SCOPE Project

New features are available in the Logs Explorer. Dismiss Learn more

Query preview  
resource.type="cloud\_function" resource.labels.function\_name="load\_csv" resource.labels.region="us-central1"

Save Run Query

Log fields

Search fields and values

LOG NAME

- cloudfunctions.googleapis.com/cloud-functions 33
- cloudaudit.googleapis.com/activity 4

FUNCTION\_NAME

- load\_csv

PROJECT\_ID

- slalom-de 37

REGION

- us-central1

RESOURCE TYPE

- Cloud Function

SEVERITY

- Info 16
- Debug 14
- Notice 4
- Error 3

Histogram

Oct 29, 2:23:30 PM 3:00 PM 2:40 PM 2:50 PM 3:00 PM 3:10 PM Oct 29, 3:23:30 PM

Query results

SEVERITY	TIMESTAMP	PDT	SUMMARY
Info	2028-10-29 15:13:04.121	PDT	load_csv n283dbwbas7d Traceback (most recent call last): File "/env/local/lib/python3.7/site-packages/google/cloud/functions/worker_v2.py",
Warning	2028-10-29 15:13:04.126	PDT	load_csv n283dbwbas7d Function execution took 845 ms, finished with status: 'crash'
Info	2028-10-29 15:15:03.486	PDT	cloudfunctions.googleapis.com google.cloud.functions.v1.CloudFunctionsService.UpdateFunction -
Info	2028-10-29 15:16:00.716	PDT	load_csv - This service is instrumented using OpenTelemetry. OpenTelemetry could not be imported; please add opentelemetry-api and opentelemetry-instrumentation to your dependencies.
Info	2028-10-29 15:16:07.335	PDT	cloudfunctions.googleapis.com google.cloud.functions.v1.CloudFunctionsService.UpdateFunction projects/slalom-de/locations/us-central1/functions/load_csv vijay.gopu@gmail.com audit_log, method: "google.cloud.functions.v1.CloudFunctionsService.UpdateFunction", principal_email: "vijay.gopu@gmail.com"
Info	2028-10-29 15:16:42.873	PDT	load_csv om01zg2rthe3 Function execution started
Info	2028-10-29 15:16:42.887	PDT	load_csv om01zg2rthe3 Processing file: gs://slalom-de/csv/reviews_sample.csv
Info	2028-10-29 15:16:45.468	PDT	load_csv om01zg2rthe3 Loaded 1 rows to table slalom-de.slalom.reviews
Info	2028-10-29 15:16:45.469	PDT	load_csv om01zg2rthe3 Function execution took 2599 ms, finished with status: 'ok'
Info	2028-10-29 15:23:13.442	PDT	load_csv om01lg9z1npt Function execution started
Info	2028-10-29 15:23:13.447	PDT	load_csv om01lg9z1npt Processing file: gs://slalom-de/csv/reviews1.csv
Info	2028-10-29 15:23:17.611	PDT	load_csv om01lg9z1npt Loaded 57153 rows to table slalom-de.slalom.reviews
Info	2028-10-29 15:23:17.612	PDT	load_csv om01lg9z1npt Function execution took 4172 ms, finished with status: 'ok'

Showing logs for last 1 hour ending at 10/29/20, 3:23 PM. Extend time by: 1 hour Edit time

# StackDriver/Google Cloud Operations

Query results				Jump to Now	Actions	Config
SEVERITY	TIMESTAMP	PDT	SUMMARY			
✓	2020-10-29 15:13:00.501 PDT	load_csv	n283dbwbas7d	Traceback (most recent call last): File "/env/local/lib/python3.7/site-packages/google/cloud/functions/worker_v2.py:1080: CsvAv	Processing file: gs://slalom-de/csv/reviews_sample.csv	
> ⚠	2020-10-29 15:13:04.121 PDT	load_csv	n283dbwbas7d	Traceback (most recent call last): File "/env/local/lib/python3.7/site-packages/google/cloud/functions/worker_v2.py:1080: CsvAv	Function execution took 845 ms, finished with status: 'crash'	
> ⚡	2020-10-29 15:13:04.126 PDT	load_csv	n283dbwbas7d	Traceback (most recent call last): File "/env/local/lib/python3.7/site-packages/google/cloud/functions/worker_v2.py:1080: CsvAv	Function execution took 845 ms, finished with status: 'crash'	
> ⓘ	2020-10-29 15:15:03.486 PDT	cloudfunctions.googleapis.com	google.cloud.functions.v1.CloudFunctionsService.UpdateFunction	-		
> ⓘ	2020-10-29 15:16:00.716 PDT	load_csv	-	This service is instrumented using OpenTelemetry. OpenTelemetry could not be imported; please add opentelemetry-api and opentelemetry		
> ⓘ	2020-10-29 15:16:07.335 PDT	cloudfunctions.googleapis.com	google.cloud.functions.v1.CloudFunctionsService.UpdateFunction	projects/slalom-de/locations/us-central1/functions/load_csv	vijay.gopu@gmail.com audit_log, method: "google.cloud.functions.v1.CloudFunctionsService.UpdateFunction", principal_email: "vijay.gopu@gmail.com"	
> ⚡	2020-10-29 15:16:42.873 PDT	load_csv	om01zg2rthe3	Function execution started		
> ⓘ	2020-10-29 15:16:42.887 PDT	load_csv	om01zg2rthe3	Processing file: gs://slalom-de/csv/reviews_sample.csv		
> ⓘ	2020-10-29 15:16:45.468 PDT	load_csv	om01zg2rthe3	Loaded 1 rows to table slalom-de.slalom.reviews		
> ⚡	2020-10-29 15:16:45.469 PDT	load_csv	om01zg2rthe3	Function execution took 2599 ms, finished with status: 'ok'		
> ⚡	2020-10-29 15:23:13.442 PDT	load_csv	om01lg9z1npt	Function execution started		
> ⓘ	2020-10-29 15:23:13.447 PDT	load_csv	om01lg9z1npt	Processing file: gs://slalom-de/csv/reviews1.csv		
> ⓘ	2020-10-29 15:23:17.611 PDT	load_csv	om01lg9z1npt	Loaded 57153 rows to table slalom-de.slalom.reviews		
> ⚡	2020-10-29 15:23:17.612 PDT	load_csv	om01lg9z1npt	Function execution took 4172 ms, finished with status: 'ok'		

Showing logs for last 1 hour ending at 10/29/20, 3:23 PM.

Extend time by: 1 hour

Edit time

# Serverless Batch Processing using Cloud Dataflow



1. JSON File Pushed to GCS

2. Batch Job is Started

3. Loaded to BigQuery

# File Loaded to Google Cloud Storage

Google Cloud Platform slalom-de Search products and resources

Storage Bucket details slalom-de

Browser Monitoring Settings

OBJECTS CONFIGURATION PERMISSIONS RETENTION LIFECYCLE

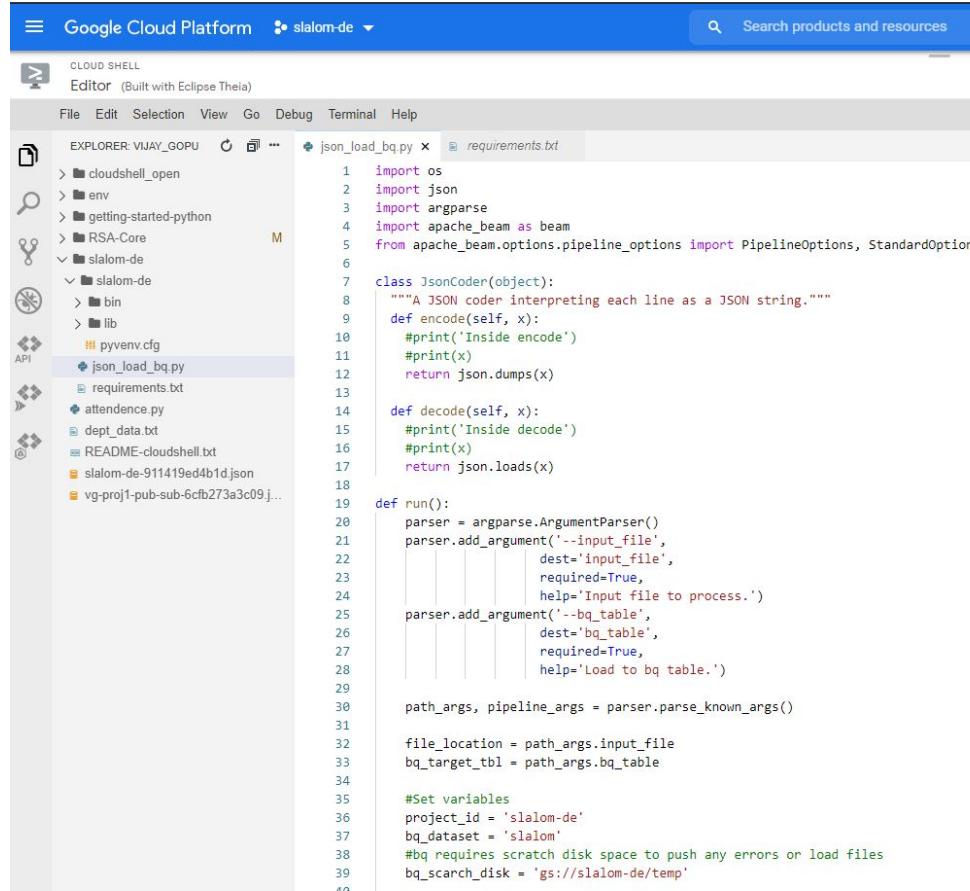
Buckets > slalom-de > json

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS DELETE

Filter Filter by object or folder name prefix

<input type="checkbox"/>	Name	Size	Type	Created time	Storage class
<input type="checkbox"/>	business_composition.json	13 MB	application/json	Oct 31, 2020, 2:54:36 PM	Standard
<input type="checkbox"/>	business_composition_sample.json	3.1 KB	application/json	Oct 31, 2020, 9:34:51 AM	Standard
<input type="checkbox"/>	test_comp.json	6.2 MB	application/json	Oct 31, 2020, 4:22:23 PM	Standard

# Pipeline Code



The screenshot shows the Google Cloud Platform Cloud Shell interface. The title bar indicates "Google Cloud Platform" and "slalom-de". A search bar at the top right says "Search products and resources". The main area is a code editor with tabs for "json\_load\_bq.py" and "requirements.txt". The left sidebar has icons for Cloud Shell, Editor (Built with Eclipse Theia), File, Edit, Selection, View, Go, Debug, Terminal, Help, and a file browser. The file browser shows a project structure: cloudshell\_open, env, getting-started-python, RSA-Core, slalom-de (which contains bin, lib, pyvenv.cfg, json\_load\_bq.py, requirements.txt, attendance.py, dep\_data.txt, README-cloudshell.txt, slalom-de-911419ed4b1d.json, and vg-proj1-pub-sub-6cfb273a3c09.j...).

```
1  import os
2  import json
3  import argparse
4  import apache_beam as beam
5  from apache_beam.options.pipeline_options import PipelineOptions, StandardOptions
6
7  class JsonCoder(object):
8      """A JSON coder interpreting each line as a JSON string."""
9      def encode(self, x):
10          #print('Inside encode')
11          #print(x)
12          return json.dumps(x)
13
14      def decode(self, x):
15          #print('Inside decode')
16          #print(x)
17          return json.loads(x)
18
19  def run():
20      parser = argparse.ArgumentParser()
21      parser.add_argument('--input_file',
22                          dest='input_file',
23                          required=True,
24                          help='Input file to process.')
25      parser.add_argument('--bq_table',
26                          dest='bq_table',
27                          required=True,
28                          help='Load to bq table.')
29
30      path_args, pipeline_args = parser.parse_known_args()
31
32      file_location = path_args.input_file
33      bq_target_tbl = path_args.bq_table
34
35      #Set variables
36      project_id = 'slalom-de'
37      bq_dataset = 'slalom'
38      #bq requires scratch disk space to push any errors or load files
39      bq_search_disk = 'gs://slalom-de/temp'
40
```

# Pipeline Code

Google Cloud Platform slalom-de Search products and resources

CLOUD SHELL Editor (Built with Eclipse Theia)

File Edit Selection View Go Debug Terminal Help

EXPLORER: VIJAY\_GOPU json\_load\_bq.py x requirements.txt

```
41     table_schema = {
42         'fields': [
43             {'name': 'Longitude', 'type': 'STRING', 'mode': 'NULLABLE'},
44             {'name': 'Latitude', 'type': 'STRING', 'mode': 'NULLABLE'},
45             {'name': 'Business_State', 'type': 'STRING', 'mode': 'NULLABLE'},
46             {'name': 'Business_City', 'type': 'STRING', 'mode': 'NULLABLE'},
47             {'name': 'Business_Address', 'type': 'STRING', 'mode': 'NULLABLE'},
48             {'name': 'Business_Id', 'type': 'STRING', 'mode': 'NULLABLE'},
49             {'name': 'Business_Name', 'type': 'STRING', 'mode': 'NULLABLE'},
50             {'name': 'alias_id', 'type': 'STRING', 'mode': 'NULLABLE'},
51             {'name': 'day_of_the_week', 'type': 'STRING', 'mode': 'NULLABLE'},
52             {'name': 'close', 'type': 'STRING', 'mode': 'NULLABLE'},
53             {'name': 'open', 'type': 'STRING', 'mode': 'NULLABLE'},
54         ]
55     }
56     options = PipelineOptions(pipeline_args)
57     p = beam.Pipeline(options=options)
58     #with beam.Pipeline() as p1:
59     composition = (
60         p
61         | 'Read txt file' >> beam.io.ReadFromText(file_location,coder=JsonCoder())
62         | 'load to bq' >> beam.io.WriteToBigQuery(bq_target_tbl,dataset=bq_dataset, project=project_id, schema=table_schema,
63             | | | | create_disposition='CREATE_IF_NEEDED', write_disposition='WRITE_APPEND', custom_gcs_temp_location=bq_scarch_disk)
64     )
65
66     if __name__ == '__main__':
67         run()
```

# requirements.txt

The screenshot shows the Google Cloud Platform Cloud Shell interface. The top navigation bar includes the Google Cloud logo, the project name "slalom-de", and a search bar. Below the navigation bar is a toolbar with icons for Cloud Shell, Editor (Built with Eclipse Theia), File, Edit, Selection, View, Go, Debug, Terminal, and Help. The main area is divided into two panes. The left pane is the File Explorer, showing a directory structure under "EXPLORER: VIJAY GOPU". The right pane is the code editor, currently displaying the "requirements.txt" file. The content of the "requirements.txt" file is:

```
apache-beam==2.25
```

# Prerequisites

1. Create Service Account and Download the JSON API Key
2. Export the file location as GOOGLE\_APPLICATION\_CREDENTIALS

```
(slalom-de)~ vijay_gopu@cloudshell:~ (slalom-de)$ export GOOGLE_APPLICATION_CREDENTIALS="/home/vijay_gopu/slalom-de-911419ed4b1d.json"  
(slalom-de)~ vijay_gopu@cloudshell:~ (slalom-de)$ []
```

3. Create a virtual environment and perform installation of the requirements.txt
4. The JSON file need to be a file in the format such as each line represents a json object.

# Command-line execution

```
(slalom-de) vijay_gopu@cloudshell:~/slalom-de (slalom-de)$ python json_load_bq.py \
> --input_file gs://slalom-de/json/test_comp.json \
> --bq_table composition \
> --runner DataflowRunner \
> --project slalom-de \
> --temp_location gs://slalom-de/temp \
> --region us-central1
/home/vijay_gopu/slalom-de/slalom-de/lib/python3.7/site-packages/apache_beam/io/gcp/bigquery.py:1677: BeamDeprecationWarning: options is deprecated since First stable release. References to <pipeline>.options will not be supported
    experiments = p.options.view_as(DebugOptions).experiments or []
/home/vijay_gopu/slalom-de/slalom-de/lib/python3.7/site-packages/apache_beam/io/gcp/bigquery_file_loads.py:900: BeamDeprecationWarning: options is deprecated since First stable release. References to <pipeline>.options will not be supported
    temp_location = p.options.view_as(GoogleCloudOptions).temp_location
WARNING: You are using pip version 20.2.3; however, version 20.2.4 is available.
You should consider upgrading via the '/home/vijay_gopu/slalom-de/slalom-de/bin/python -m pip install --upgrade pip' command.
WARNING: You are using pip version 20.2.3; however, version 20.2.4 is available.
You should consider upgrading via the '/home/vijay_gopu/slalom-de/slalom-de/bin/python -m pip install --upgrade pip' command.
WARNING:root:Make sure that locally built Python SDK docker image has Python 3.7 interpreter.
(slalom-de) vijay_gopu@cloudshell:~/slalom-de (slalom-de)$ []
```

# Dataflow Job Run

Google Cloud Platform slalom-de ▾ Search products and resources

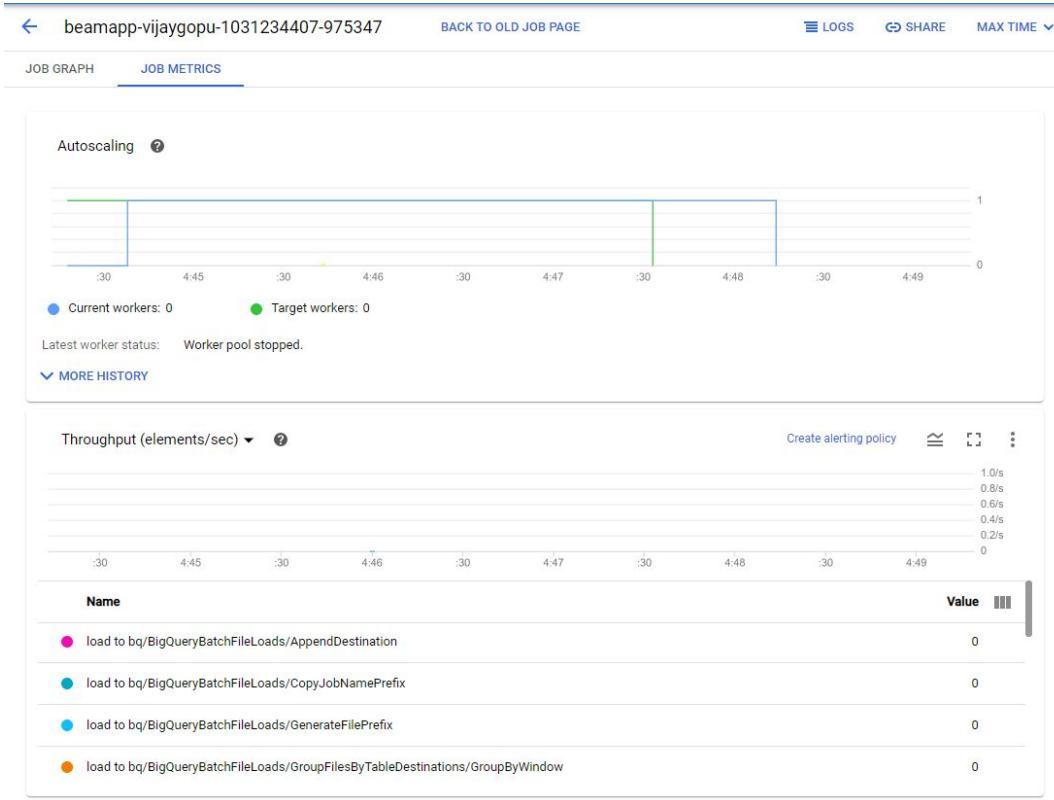
☰ Dataflow	Jobs	+ CREATE JOB FROM TEMPLATE	+ CREATE JOB FROM SQL			
☰ Jobs	<input checked="" type="checkbox"/> Running Filter jobs					
☰ Notebooks						
Name	Type	End time	Elapsed time	Start time ↓	Status	SDK version
beamapp-vijaygopu-1031234407-975347	Batch		2 min 10 sec	Oct 31, 2020, 4:44:12 PM	Running	2.25.0

Google Cloud Platform slalom-de ▾ Search products and resources

☰ Dataflow	◀ beamapp-vijaygopu-1031234407-975347	BACK TO OLD JOB PAGE	STOP
☰ Jobs	JOB GRAPH	JOB METRICS	
☰ Notebooks			

```
graph TD; A[Read txt file  
Succeeded  
2 sec  
1 of 1 stage succeeded] --> B[load to bq  
Running  
8 sec  
7 of 7 stages succeeded]
```

# Dataflow Job Metrics



### Job info

Job name	beamapp-vijaygopu-1031234407-975347
Job ID	2020-10-31_16_44_11-11151559422184315932
Job type	Batch
Job status	<span>✓ Succeeded</span>
SDK version	Apache Beam Python 3.7 SDK 2.25.0
Job region	us-central1
Worker location	us-central1-b
Current workers	0
Latest worker status	Worker pool stopped.
Start time	October 31, 2020 at 4:44:12 PM GMT-7
Elapsed time	4 min 6 sec
Encryption type	Google-managed key

### Resource metrics

Current vCPUs	1
Total vCPU time	0.049 vCPU hr
Current memory	3.75 GB
Total memory time	0.185 GB hr
Current HDD PD	250 GB
Total HDD PD time	12.354 GB hr
Current SSD PD	0 B
Total SSD PD time	0 GB hr

### Pipeline options

runner	DataflowRunner
project	slalom-de
job_name	beamapp-vijaygopu-1031234407-975347
staging_location	gs://slalom-de/temp/beamapp-vijaygopu-1031234407-975347.160
temp_location	gs://slalom-de/temp/beamapp-vijaygopu-1031234407-975347.160
region	us-central1

# Data Landed in BigQuery

Google Cloud Platform slalom-de ▾ Search products and resources

BigQuery FEATURES & INFO SHORTCUT

Query history Saved queries Job history Transfers Scheduled queries Reservations BI Engine Resources + ADD DATA

Query results [SAVE RESULTS](#) [EXPLORE DATA](#)

Query complete (2.4 sec elapsed, 26.4 MB processed)

Row	Longitude	Latitude	Business_State	Business_City	Business_Address	Business_Id	Business_Name	alias_id	day_of_the_week	close	open
1	-111.8397048	33.4364193	AZ	Mesa	1211 N Country Club Dr Mesa, AZ 85201	8q6a5_rWBL8BNimvtZ0a-g	Krazy Sub-Steve's	8q6a5_rWBL8BNimvtZ0a-g	friday		
2	-111.8397048	33.4364193	AZ	Mesa	1211 N Country Club Dr Mesa, AZ 85201	8q6a5_rWBL8BNimvtZ0a-g	Krazy Sub-Steve's	8q6a5_rWBL8BNimvtZ0a-g	monday		
3	-111.8397048	33.4364193	AZ	Mesa	1211 N Country Club Dr Mesa, AZ 85201	8q6a5_rWBL8BNimvtZ0a-g	Krazy Sub-Steve's	8q6a5_rWBL8BNimvtZ0a-g	saturday		
4	-111.8397048	33.4364193	AZ	Mesa	1211 N Country Club Dr Mesa, AZ 85201	8q6a5_rWBL8BNimvtZ0a-g	Krazy Sub-Steve's	8q6a5_rWBL8BNimvtZ0a-g	sunday		
5	-111.8397048	33.4364193	AZ	Mesa	1211 N Country Club Dr Mesa, AZ 85201	8q6a5_rWBL8BNimvtZ0a-g	Krazy Sub-Steve's	8q6a5_rWBL8BNimvtZ0a-g	thursday		
6	-111.8397048	33.4364193	AZ	Mesa	1211 N Country Club Dr Mesa, AZ 85201	8q6a5_rWBL8BNimvtZ0a-g	Krazy Sub-Steve's	8q6a5_rWBL8BNimvtZ0a-g	tuesday		
7	-111.8397048	33.4364193	AZ	Mesa	1211 N Country Club Dr Mesa, AZ 85201	8q6a5_rWBL8BNimvtZ0a-g	Krazy Sub-Steve's	8q6a5_rWBL8BNimvtZ0a-g	wednesday		
8	-111.8627477	33.3934584	AZ	Mesa	1440 W Southern Ave Mesa, AZ 85202	9B0eSWUvvh4rRhpQNgrdqA	Carl's Jr.	9B0eSWUvvh4rRhpQNgrdqA	friday		
9	-111.8627477	33.3934584	AZ	Mesa	1440 W Southern Ave Mesa, AZ 85202	9B0eSWUvvh4rRhpQNgrdqA	Carl's Jr.	9B0eSWUvvh4rRhpQNgrdqA	monday		
10	-111.8627477	33.3934584	AZ	Mesa	1440 W Southern Ave Mesa, AZ 85202	9B0eSWUvvh4rRhpQNgrdqA	Carl's Jr.	9B0eSWUvvh4rRhpQNgrdqA	saturday		
11	-111.8627477	33.3934584	AZ	Mesa	1440 W Southern Ave Mesa, AZ 85202	9B0eSWUvvh4rRhpQNgrdqA	Carl's Jr.	9B0eSWUvvh4rRhpQNgrdqA	sunday		

# StackDriver Logs

Google Cloud Platform slalom-de ▾ Search products and resources LAST 1 HOUR PAGE LAYOUT

Operations Logging Logs Explorer OPTIONS REFINE SCOPE Project

New features are available in the Logs Explorer.

Query preview resource.type="dataflow\_step"

Log fields

Search fields and values

Histogram

Oct 31, 4:05:30 PM 4:20 PM 4:30 PM 4:40 PM 4:50 PM Oct 31, 5:05:30 PM

Query results

SEVERITY TIMESTAMP PDT SUMMARY

2020-10-31 16:55:39.491 PDT Executing operation load to bq/bigquerybatchfileloads/removetemptables/deduplicatetables/read+load to bq/bigquerybatchfileloads/removetemptables/ued...

2020-10-31 16:55:40.373 PDT "Executing workitem <BatchWorkItem s22 steps=load to bq/BigQueryBatchFileLoads/RemoveTempTables/DeduplicateTables/Read+load to bq/BigQueryBatchFile...

2020-10-31 16:55:40.373 PDT "Memory usage of worker beamapp-vijaygopu-1031235-10311652-lnr0-harness-ij9t is 195 MB"

2020-10-31 16:55:40.389 PDT "I1031 23:55:40.389378 146 global\_config.custom.cc:114] GlobalConfig(grpc\_poll\_strategy) is set by environment variable. Please use flag(--grpc\_poll...

2020-10-31 16:55:40.389 PDT "I1031 23:55:40.389882 146 global\_config.custom.cc:79] GlobalConfig(grpc\_client\_channel\_backup\_poll\_interval\_ms) is set by environment variable. Ple...

2020-10-31 16:55:41.450 PDT "Refusing to split <dataflow\_worker.shuffle.GroupedShuffleRangeTracker object at 0x7fbac1b9c50> at b'\x00\x00\xff\x00\x00\xff\x00\x01': unstart...

2020-10-31 16:55:41.450 PDT "Refusing to split GroupedShuffleReader <dataflow\_worker.shuffle.GroupedShuffleReader object at 0x7fbac1b9ed0> at gAD\_APBA\_wAB"

2020-10-31 16:55:41.450 PDT "Finished processing workitem 5370482457148866474 successfully. Reporting status to Dataflow service."

2020-10-31 16:55:42.477 PDT "Completed workitem: 5370482457148866474 in 2.104264736 seconds"

2020-10-31 16:55:42.584 PDT Finished operation load to bq/BigQueryBatchFileLoads/RemoveTempTables/DeduplicateTables/Read+load to bq/BigQueryBatchFileLoads/RemoveTempTables/Dedu...

2020-10-31 16:55:42.569 PDT Executing success step success44

2020-10-31 16:55:42.658 PDT Cleaning up.

2020-10-31 16:55:42.708 PDT Starting worker pool teardown.

2020-10-31 16:55:42.742 PDT Stopping worker pool...

2020-10-31 16:56:27.395 PDT Autoscaling: Resized worker pool from 1 to 0.

2020-10-31 16:56:27.435 PDT Worker pool stopped.

2020-10-31 16:56:27.474 PDT Tearing down pending resources...

Showing logs for last 1 hour ending at 10/31/20, 5:05 PM Extend time by 1 hour Edit time

# Loading sqlite file to BigQuery

Using Pandas

# requirements.txt

The screenshot shows the Google Cloud Platform Cloud Shell interface. The top navigation bar includes the Google Cloud logo, the project name "slalom-de", and a search bar. Below the header is a toolbar with icons for Cloud Shell, Editor (Built with Eclipse Theia), File, Edit, Selection, View, Go, Debug, Terminal, and Help.

The left sidebar is an Explorer view showing the directory structure:

- cloudshell\_open
- env
- getting-started-python
- RSA-Core
- slalom-de
  - data
  - bin
  - lib
    - pyvenv.cfg
  - json\_load\_bq.py
  - load\_users\_ba.py
  - requirements.txt
  - requirements2.txt
  - attendence.py
  - dept\_data.txt
  - README-cloudshell.txt
  - slalom-de-911419ed4b1d.json
  - vg-proj1-pub-sub-6cfb273a3c09.j...

The "requirements2.txt" file is selected in the sidebar and is displayed in the main editor area. The content of the file is:

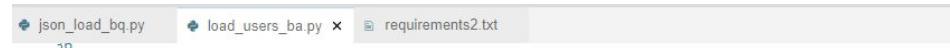
```
1 pandas-gbq==0.14.0
2 pandas==1.1.4
```

# Load Job Code

```
json_load_bq.py  load_users_ba.py x  requirements2.txt

 1  import os
 2  import sqlite3
 3  import pandas as pd
 4  import argparse
 5  #import pandas_gbq
 6  from google.oauth2 import service_account
 7
 8  def load_users(conn,credentials):
 9      #using pandas read the USERS2 table using connection object
10      df = pd.read_sql_query("SELECT * FROM USERS2", conn)
11
12      #rename the df column name because bq will not accept spaces
13      users_df = df.rename(columns={'User - Years Elite':'User_Years_Elite',
14                               'Business - Id':'Business_Id',
15                               'User - Compliments Cool':'User_Compliments_Cool',
16                               'User - Compliments Cute':'User_Compliments_Cute',
17                               'User - Compliments Funny':'User_Compliments_Funny',
18                               'User - Compliments Hot':'User_Compliments_Hot',
19                               'User - Compliments List':'User_Compliments_List',
20                               'User - Compliments More':'User_Compliments_More',
21                               'User - Compliments Note':'User_Compliments_Note',
22                               'User - Compliments Photos':'User_Compliments_Photos',
23                               'User - Compliments Plain':'User_Compliments_Plain',
24                               'User - Compliments Profile':'User_Compliments_Profile',
25                               'User - Compliments Writer':'User_Compliments_Writer',
26                               'User - Fans':'User_Fans','User - Name':'User_Name',
27                               'Review - Id':'Review_Id','User - Id':'User_Id',
28                               'User - Votes Cool':'User_Votes_Cool',
29                               'User - Votes Funny':'User_Votes_Funny',
30                               'User - Votes Useful':'User_Votes_Usedful',
31                               'User - Yelping Since':'User_Yelping_Since'})}
32
33      #Load the data to bq using the dataframe method to_gbq
34      #Provide table name along with the dataset name, project where the dataset resides, path to the service account key as credentials
35      users_df.to_gbq(destination_table='slalom.users',project_id='slalom-de',credentials=credentials,if_exists='append')
36
```

# Load Job Code



```
json_load_bq.py      load_users_ba.py x  requirements2.txt

30
37  def load_business_attributes(conn,credentials):
38      df = pd.read_sql_query("SELECT * FROM BUSINESS_ATTRIBUTES", conn)
39      ba = df.rename(columns={'Business - Restaurant':'Business_Restaurant',
40                           'Business - Accepts Credit Cards':'Business_Acccepts_Credit_Cards',
41                           'Business - Accepts Insurance':'Business_Acccepts_Insurance',
42                           'Business - Ages Allowed':'Business_Ages_Allowed',
43                           'Business - Alcohol':'Business_Alcohol',
44                           'Business - Attire':'Business_Attire',
45                           'Business - BYOB/Corkage':'Business_BYOB_Corkage',
46                           'Business - BYOB':'Business_BYOB',
47                           'Business - By Appointment Only':'Business_By_Appointment_Only',
48                           'Business - Caters':'Business_Caters',
49                           'Business - Coat Check':'Business_Coat_Check',
50                           'Business - Corkage':'Business_Corkage',
51                           'Business - Delivery':'Business_Delivery',
52                           'Business - Dietary Restrictions':'Business_Dietary_Restrictions',
53                           'Business - Dogs Allowed':'Business_Dogs_Allowed',
54                           'Business - Drive-Thru':'Business_Drive_Thru',
55                           'Business - Good For Dancing':'Business_Good_For_Dancing',
56                           'Business - Good For Groups':'Business_Good_For_Groups',
57                           'Business - Good For Kids':'Business_Good_For_Kids',
58                           'Business - Good for Kids2':'Business_Good_for_Kids2',
59                           'Business - Happy Hour':'Business_Happy_Hour',
60                           'Business - Has TV':'Business_Has_TV',
61                           'Business - Noise Level':'Business_Noise_Level',
62                           'Business - Open 24 Hours':'Business_Open_24_Hours',
63                           'Business - Orden at Counter':'Business_Order_at_Counter',
64                           'Business - Outdoor Seating':'Business_Outdoor_Seating',
65                           'Business - Parking':'Business_Parking',
66                           'Business - Payment Types':'Business_Payment_Types',
67                           'Business - Price Range':'Business_Price_Range',
68                           'Business - Smoking':'Business_Smoking',
69                           'Business - Take-out':'Business_Take_out',
70                           'Business - Takes Reservations':'Business_Takes_Reservations',
71                           'Business - Waiter Service':'Business_Waiter_Service',
72                           'Business - Wheelchair Accessible':'Business_Wheelchair_Accessible',
73                           'Business - Wi-Fi':'Business_Wi_Fi',
74                           'Business - Id':'Business_Id',
75                           'Business - Categories':'Business_Categories',
```

# Load Job Code

```
json_load_bq.py load_users_ba.py requirements2.txt
75     'Business - Categories':'Business_Categories',
76     'Business - Name':'Business_Name',
77     'Business - Neighborhoods':'Business_Neighborhoods',
78     'Business - Open?':'Business_Open'})
79 #Load the data to bq using the dataframe method to_gbq
80 #Provide table name along with the dataset name, project where the dataset resides, path to the service account key as credentials
81 ba.to_gbq(destination_table='slalom.business_attributes',project_id='slalom-de',credentials=credentials,if_exists='append')
82
83 def create_connection(sqlite_file):
84     conn = sqlite3.connect(sqlite_file)
85     return conn
86
87 def close_connection(conn):
88     conn.close()
89
90 def set_credentials(credentials_file):
91     #Credentials required to load data to project using the service account
92     credentials = service_account.Credentials.from_service_account_file(
93         credentials_file,
94     )
95     return credentials
96
97 def run():
98     parser = argparse.ArgumentParser()
99     parser.add_argument('--sqlite_file',
100                         dest='sqlite_file',
101                         required=True,
102                         help='Provide Sqlite file path')
103     parser.add_argument('--credentials_file',
104                         dest='credentials_file',
105                         required=True,
106                         help='Service Account Credentials file path')
107     args = parser.parse_args()
108
109     #Setting variables from parser
110     sqlite_file = args.sqlite_file
111     credentials_file = args.credentials_file
```

# Load Job Code

```
❸ json_load_bq.py ❸ load_users_ba.py x ❸ requirements2.txt
 96
 97     def run():
 98         parser = argparse.ArgumentParser()
 99         parser.add_argument('--sqlite_file',
100                             dest='sqlite_file',
101                             required=True,
102                             help='Provide Sqlite file path')
103         parser.add_argument('--credentials_file',
104                             dest='credentials_file',
105                             required=True,
106                             help='Service Account Credentials file path')
107     args = parser.parse_args()
108
109     #Setting variables from parser
110     sqlite_file = args.sqlite_file
111     credentials_file = args.credentials_file
112
113     #set credentials
114     credentials = set_credentials(credentials_file)
115
116     #create connection
117     conn = create_connection(sqlite_file)
118
119     #load_users_table_to_bq
120     load_users(conn,credentials)
121
122     #load_business_attributes_to_bq
123     load_business_attributes(conn,credentials)
124
125     #close connection
126     close_connection(conn)
127
128 if __name__ == '__main__':
129     run()
```

# Command line execution

```
(slalom-de) vijay_gopu@cloudshell:~/slalom-de (slalom-de)$ python load_users.py \
> --sqlite_file /home/vijay_gopu/slalom-de/data/user.sqlite \
> --credentials_file /home/vijay_gopu/slalom-de-911419ed4b1d.json
(slalom-de) vijay_gopu@cloudshell:~/slalom-de (slalom-de)$ []
```

# Users In BigQuery

Google Cloud Platform slalom-de ▾

Search products and resources

BigQuery FEATURES & INFO SHORTCUT

Query history Unsaved query Edited

+ COMPOSE NEW QUERY HIDE EDITOR FULL SCREEN

1 select \* from slalom-de.slalom.users

Run Save query Save view Schedule query More This query will process 32.5 MB when run. ✓

slalom-de

slalom

business\_attributes

composition

reviews

users

USERS

QUERY TABLE SHARE TABLE COPY TABLE DELETE TABLE EXPORT

Row	User_Years_Elite	Business_Id	User_Compliments_Cool	User_Compliments_Cute	User_Compliments_Funny	User_Compliments_Hot	User_Compliments_List	User_Compliments_More	User_Compliments_Note	User_Com
285701	2008 2009 2010 2011 2012	KBG28p3lGx17hOPoHhq5PQ	3755	212	1866	2596	122	211	1094	583
285702	2008 2009 2010 2011 2012	Oc-F6091v0e3T8lOsCBB8Q	3755	212	1866	2596	122	211	1094	583
285703	2008 2009 2010 2011 2012	cBpJlOrVXotDI0XAZH_k0g	3755	212	1866	2596	122	211	1094	583
285704	2008 2009 2010 2011 2012	jf4RUa9EQ037hqxRCxbEXQ	3755	212	1866	2596	122	211	1094	583
285705	2008 2009 2010 2011 2012	jf4RUa9EQ037hqxRCxbEXQ	3755	212	1866	2596	122	211	1094	583
285706	2008 2009 2010 2011 2012	gUt-pUpOVVhaCFC8-E4yQ	3755	212	1866	2596	122	211	1094	583
285707	2008 2009 2010 2011 2012	AAG0391o2Ke0fgr6BhH-yQ	3755	212	1866	2596	122	211	1094	583
285708	2008 2009 2010 2011 2012	JILK9QPjd5pOBEEaY83lw	3755	212	1866	2596	122	211	1094	583
285709	2008 2009 2010 2011 2012	IC8no-tldDWgkyzb5d-Nvw	3755	212	1866	2596	122	211	1094	583
285710	2008 2009 2010 2011 2012	3QtWG5sN3HJNBVKn_wHQEw	3755	212	1866	2596	122	211	1094	583
285711	2008 2009 2010 2011 2012	byhwHi0lhYdyY5kSpuqoaQ	3755	212	1866	2596	122	211	1094	583
285712	2008 2009 2010 2011 2012	-JpZiiGPkOUCEiODGNyow	3755	212	1866	2596	122	211	1094	583

Rows per page: 100 285701 - 285764 of 285764 First page < > >> Last page

# Business\_Attributes In BigQuery

Google Cloud Platform slalom-de

Search products and resources

BigQuery FEATURES & INFO SHORTCUT

Query history

Saved queries

Job history

Transfers

Scheduled queries

Reservations

BI Engine

Resources +

Search for your tables and d...

slalom-de

+ slalom

business\_attributes

Run Save query Save view Schedule query More

This query will process 80.9 MB when run.

business\_attributes

QUERY TABLE SHARE TABLE COPY TABLE DELETE TABLE EXPORT

Schema Details Preview

Row	Business_Restaurant	Business_Accepts_Credit_Cards	Business_Accepts_Insurance	Business_Ages_Allowed	Business_Alcohol	Business_Attire	Business_BYOB_Corkage	Business_BYOB	Business_By_Appointment_Only	Business_Caters	Business_Coat_Check	Business_Drinking_Water	Business_Early_Morning	Business_Evening	Business_Family_Oriented	Business_Holiday_Special	Business_Late_Night	Business_Liquor	Business_Music	Business_Nightlife	Business_Outdoor	Business_Pet_Friendly	Business_Restaurant	Business_Smoking	Business_Takeout	Business_Vegan	Business_Walk_In			
285701	True	True	null	null	full_bar	casual	null	null	null	False	False	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	
285702	True	True	null	null	full_bar	casual	null	null	null	False	False	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	
285703	True	True	null	null	full_bar	casual	null	null	null	False	False	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	
285704	True	True	null	null	full_bar	casual	null	null	null	False	False	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	
285705	True	True	null	null	full_bar	casual	null	null	null	False	False	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	null	
285706	True	True	null	null	null	null	null	null	True	null	null	null	null	null	null	null	True	null	null	null	null	null	null	null	null	null	null	null	null	null
285707	True	True	null	null	null	null	null	null	True	null	null	null	null	null	null	null	True	null	null	null	null	null	null	null	null	null	null	null	null	null
285708	True	True	null	null	null	null	null	null	True	null	null	null	null	null	null	null	True	null	null	null	null	null	null	null	null	null	null	null	null	null
285709	True	True	null	null	null	null	null	null	True	null	null	null	null	null	null	null	True	null	null	null	null	null	null	null	null	null	null	null	null	null
285710	True	True	null	null	null	null	null	null	True	null	null	null	null	null	null	null	True	null	null	null	null	null	null	null	null	null	null	null	null	null
285711	True	True	null	null	null	null	null	null	True	null	null	null	null	null	null	null	True	null	null	null	null	null	null	null	null	null	null	null	null	null
285712	True	True	null	null	null	null	null	null	True	null	null	null	null	null	null	null	True	null	null	null	null	null	null	null	null	null	null	null	null	null

Rows per page: 100 First page < > Last page

# SQL Questions

# Mean Reviews of every Business

Google Cloud Platform slalom-de ▾ Search products and resources

BigQuery FEATURES & INFO SHORTCUT

Query history Saved queries Job history Transfers Scheduled queries Reservations BI Engine Resources + ADD DATA

Search for your tables and datasets

slalom-de slalom business\_attributes composition reviews users

WITH BUSINESS\_ATTR AS (  
 2 SELECT DISTINCT BUSINESS\_ID, BUSINESS\_NAME  
 3 FROM `slalom-de.slalom.business\_attributes`  
4 )  
5 SELECT  
6 ba.BUSINESS\_ID,  
7 ba.BUSINESS\_NAME,  
8 ROUND(AVG(rev.REVIEW\_STARS),2) AS REVIEW\_RATING  
9 FROM BUSINESS\_ATTR ba,  
10 `slalom-de.slalom.reviews` rev  
11 WHERE ba.BUSINESS\_ID = rev.BUSINESSID  
12 GROUP BY BUSINESS\_ID, BUSINESS\_NAME;  
13

Valid.

Run Save query Save view Schedule query More

Query results SAVE RESULTS EXPLORE DATA

Query complete (1.5 sec elapsed, 20.5 MB processed)

Job information Results JSON Execution details

Row	BUSINESS_ID	BUSINESS_NAME	REVIEW_RATING
1	RtXNB9vscyXKQL0hAvuEHw	Don Pedro's Mexican Food	3.05
2	-WZIxGXJHMGidZXRhKxP3w	Mimi's Cafe	3.02
3	YHGpemLe7cbnPSubG-cRRg	Burger King	1.78
4	k6Si433-EJrY4J7SZxsnjA	Grazie Pizzeria & Wine Bar	3.78
5	G5SASWuL_CVxpgwXXGC4DA	Stingray Sushi	3.49
6	bfDQai9X59uWK-XgP0t6rA	Dos Gringos	3.15
7	Exx5ffvnmk4MrTyCkPRuug	Los Olivos	3.14
8	b5cEoKR8iQliq-yT2_O0LQ	Carlsbad Tavern	3.78
9	hh2IP4_2N-tk_OxmaTf_qA	J & G Steakhouse	3.94
10	uVbt9dOe6qJgBBI4_XsXRA	Fleming's Prime Steakhouse & Wine Bar	3.9
11	YeDYa6tYL16CyqYvUV0tLw	Margaritaville	2.61

# Mean Reviews By Top 5 Zip Codes

The screenshot shows the Google Cloud Platform BigQuery interface. The left sidebar displays navigation links: Query history, Saved queries, Job history, Transfers, Scheduled queries, Reservations, BI Engine, and Resources. Under Resources, '+ ADD DATA' is visible. The main area shows an 'Unsaved query' tab with the following SQL code:

```
1 WITH TOP_5 AS (
2     SELECT ZIP_CODE, NUM_OF_BIZ,
3            RANK() OVER (ORDER BY NUM_OF_BIZ DESC) AS RNK
4     FROM ( SELECT
5             SUBSTRING(BUSINESS_ADDRESS, LENGTH(BUSINESS_ADDRESS)-5, LENGTH(BUSINESS_ADDRESS)) AS ZIP_CODE,
6             COUNT(*) AS NUM_OF_BIZ
7            FROM `slalom-de.slalom.composition` WHERE UPPER(DAY_OF_THE_WEEK)= 'SATURDAY'--need only a single record as there are rows for each week
8            GROUP BY ZIP_CODE)
9     )SELECT ZIP_CODE, ROUND(AVG(REVIEW_STARS),2) AS REVIEW_AVG,  FROM `slalom-de.slalom.reviews` rev, `slalom-de.slalom.composition` comp, TOP_5
10 WHERE rev.BUSINESSID = comp.BUSINESS_ID
11 AND SUBSTRING(comp.BUSINESS_ADDRESS, LENGTH(BUSINESS_ADDRESS)-5, LENGTH(BUSINESS_ADDRESS)) = TOP_5.ZIP_CODE
12 AND TOP_5.RNK < 6
13 GROUP BY ZIP_CODE
```

The status bar indicates 'Valid.' Below the code are buttons for Run, Save query, Save view, Schedule query, and More. The 'Query results' section shows the completed query output:

Query complete (3.4 sec elapsed, 11.6 MB processed)

Job information    Results    JSON    Execution details

Row	ZIP_CODE	REVIEW_AVG
1	85251	3.82
2	85308	3.66
3	85004	3.75
4	85016	3.8
5	85281	3.71
6	85032	3.67

# Top 10 Active Reviewers

Google Cloud Platform slalom-de

BigQuery FEATURES & INFO SHORTCUT

Query history Saved queries Job history Transfers Scheduled queries Reservations BI Engine

+ ADD DATA

Search for your tables and datasets

WITH USER\_NAMES AS ( SELECT DISTINCT USER\_ID, USER\_NAME FROM `slalom-de.slalom.users` ) SELECT \* FROM ( SELECT USER\_NAME, TOT\_REVIEWS, RANK() OVER (ORDER BY TOT\_REVIEWS DESC) AS RANK\_ FROM ( SELECT USER\_NAME, COUNT(\*) AS TOT\_REVIEWS FROM `slalom-de.slalom.reviews` rev, USER\_NAMES WHERE rev.User\_Id = USER\_NAMES.User\_Id GROUP BY USER\_NAME ) ) AS TOP\_10 WHERE RANK\_ < 11

Valid.

Run Save query Save view Schedule query More

Query results SAVE RESULTS EXPLORE DATA

Query complete (1.7 sec elapsed, 15.1 MB processed)

Job information Results JSON Execution details

Row	USER_NAME	TOT_REVIEWS	RANK_
1	Michael	3272	1
2	Chris	2896	2
3	Mike	2863	3
4	Jennifer	2856	4
5	John	2728	5
6	David	2435	6
7	Jason	2053	7
8	Brian	2005	8
9	Jessica	1897	9
10	Scott	1881	10

# Loading To Cloud Storage

# Mean Reviews By Business

Cloud Shell Editor [Preview](#)

File Edit Selection View Go Debug Terminal Help

EXPLORER: VIJAY\_GOPU

- cloudshell\_open
- env
- getting-started-python
- RSA-Core
- slalom-de
  - data
  - slalom-de
    - bin
    - lib
      - pyvenv.cfg
    - json\_load\_bq.py
    - load\_mean\_rev\_by\_biz.py
    - load\_users\_ba.py
    - requirements\_load2gcs.txt
    - requirements.txt
    - requirements2.txt
    - attendance.py
    - dept\_data.txt
    - README-cloudshell.txt
    - slalom-de-911419ed4b1d.json
    - vg-proj1-pub-sub-6cfb273a3c09.json

```
1 import pandas as pd
2 import argparse
3 #import pandas_gbq
4 from google.cloud import storage
5 from google.oauth2 import service_account
6
7 """
8 | Must export GOOGLE_APPLICATION_CREDENTIALS
9 ...
10
11 def set_credentials(credentials_file):
12     #Credentials required to query data
13     credentials = service_account.Credentials.from_service_account_file(
14         credentials_file,
15     )
16     return credentials
17
18 def get_mean_reviews_by_biz(credentials,bucket_name):
19     q_mean_reviews_by_biz = "WITH BUSINESS_ATTR AS ( SELECT DISTINCT BUSINESS_ID, BUSINESS_NAME FROM `slalom-de.slalom.business_attributes` ) "
20     q_mean_reviews_by_biz += "SELECT ba.BUSINESS_ID, ba.BUSINESS_NAME, ROUND(AVG(rev.REVIEW_STARS),2) AS REVIEW_RATING"
21     q_mean_reviews_by_biz += " FROM BUSINESS_ATTR ba, `slalom-de.slalom.reviews` rev"
22     q_mean_reviews_by_biz += " WHERE ba.BUSINESS_ID = rev.BUSINESSID"
23     q_mean_reviews_by_biz += " GROUP BY BUSINESS_ID, BUSINESS_NAME"
24
25     #read data from bq and return dataframe
26     df = pd.read_gbq(q_mean_reviews_by_biz,project_id=bucket_name,credentials=credentials)
27
28     return df
29
30 def load_file_gcs(bucket_name,file_name,data_frame):
31     client = storage.Client()
32     bucket = client.get_bucket(bucket_name)
33
34     #create the file in the bucket
35     bucket.blob(file_name).upload_from_string(data_frame.to_csv(sep="|",index=False), 'text/csv')
36
```

# Mean Reviews By Business

```
load_mean_rev_by_biz.py x
 30 def load_file_gcs(bucket_name,file_name,data_frame):
 31     client = storage.Client()
 32     bucket = client.get_bucket(bucket_name)
 33
 34     #create the file in the bucket
 35     bucket.blob(file_name).upload_from_string(data_frame.to_csv(sep="|",index=False), 'text/csv')
 36
 37 def run():
 38     parser = argparse.ArgumentParser()
 39     parser.add_argument('--output_file',
 40                         dest='output_file',
 41                         required=True,
 42                         help='Provide output file path')
 43     parser.add_argument('--bucket_name',
 44                         dest='bucket_name',
 45                         required=True,
 46                         help='bucket name for output file')
 47     parser.add_argument('--credentials_file',
 48                         dest='credentials_file',
 49                         required=True,
 50                         help='Service Account Credentials file path')
 51     args = parser.parse_args()
 52
 53     bucket_name = args.bucket_name
 54     output_file = args.output_file
 55     credentials_file = args.credentials_file
 56
 57     #set credentials using the json file
 58     credentials = set_credentials(credentials_file)
 59
 60     #execute the query to extract data from bq
 61     df_mean_reviews_by_biz = get_mean_reviews_by_biz(credentials,bucket_name)
 62
 63     #create the file in gcs
 64     load_file_gcs(bucket_name,output_file,df_mean_reviews_by_biz)
 65
 66 if __name__ == '__main__':
 67     run()
 68
```

# Mean Reviews By Business - Prerequisites

## 1. requirements.txt

```
load_mean_rev_by_biz.py requirements_load2gcs.txt
1 google-cloud-storage==1.32.0
2 #pandas-gbq==0.14.0 This is required if already not in your environment
```

## 2. Export Google Credentials

```
~/slalom-de (slalom-de)$ export GOOGLE_APPLICATION_CREDENTIALS="/home/vijay_gopu/slalom-de-911419ed4b1d.json"
```

# Mean Reviews By Business

## Command Line Execution

```
(slalom-de) vijay_gopu@cloudshell:~/slalom-de (slalom-de)$ python load_mean_rev_by_biz.py \
> --output_file output/csv/mean_reviews_by_biz.csv \
> --bucket_name slalom-de \
> --credentials_file /home/vijay_gopu/slalom-de-911419ed4b1d.json
(slalom-de) vijay_gopu@cloudshell:~/slalom-de (slalom-de)$ []
```

# Mean Reviews By Business

≡ Google Cloud Platform slalom-de ▾ Search products and resources

Storage Bucket details ← Bucket details

Browser Monitoring Settings

## slalom-de

OBJECTS CONFIGURATION PERMISSIONS RETENTION LIFECYCLE

Buckets > slalom-de > output > csv

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS DELETE

Filter Filter by object or folder name prefix

<input type="checkbox"/>	Name	Size	Type	Created time	Storage clas:
<input type="checkbox"/>	mean_reviews_by_biz.csv	255.4 KB	text/csv	Nov 5, 2020, 8:37:48 AM	Standard

# Mean Reviews By Business

AutoSave  OFF

File Home Insert Page Layout Formulas Data Review View

Cut Copy Format Painter

Font Alignment

A1 BUSINESS\_ID|BUSINESS\_NAME|REVIEW\_RATING

	A	B	C	D	E	F	G	H
1	BUSINESS_ID	BUSINESS_NAME	REVIEW_RATING					
2	RTXNB9vscyXKQL0hAVuEHw	Don Pedro's Mexican Food	3.05					
3	-WZIxGXJHMGidZXRhKxP3w	Mimi's Cafe	3.02					
4	YHGpemLe7cbnPSubG-cRg	Burger King	1.78					
5	k6Si433-EJrY4J7SzxsnjA	Grazie Pizzeria & Wine Bar	3.78					
6	G5SASWuL_CVxpgwXXGC4DA	Stingray Sushi	3.49					
7	bFDQai9X59uWK-XgP0t6rA	Dos Gringos	3.15					
8	Exx5ffvnmk4MrTyCkPruug	Los Olivos	3.14					
9	b5cEoKR8iQliq-yT2_O0LQ	Carlsbad Tavern	3.78					
10	hh2lP4_2N-tk_OxmaTf_qA	J & G Steakhouse	3.94					
11	uvBt9dOe6qjgBBI4_XsKRA	Fleming's Prime Steakhouse & Wine Bar	3.9					
12	YeDYa6tYL16CyqYvUVotLw	Margaritaville	2.61					
13	sbsFamEj5wDxNAjUKrMcSw	RnR	3.35					
14	_wrA7CMbMqcWEDCNgHi5cg	Mimis Cafe	2.44					
15	gPCUlvpMCXVRDqszO8v69w	Spoonz Cafe	3.17					
16	a0eg1AZESJbsUL6i3h_VsA	Kokopeli Deli	1.62					
17	a29C6sg39V2HP_tq3m8nEw	The Dirty Drummer	2.17					
18	0bQKW7e23wTJxgv_p2HjQ	Shish Kabobs & Gyros	3.79					
19	t_0576Esa58pcLHCepj31A	Steve's Grill	2.79					
20	r-a-Cn9hxdEnYTtVTB5bMQ	The Grind	3.57					
21	9Y3aQAVITKEJYe5vLzr13w	Breakfast Club	3.67					
22	TftIOE6h9E9o34dEkw9L_w	The Rose and Crown	3.59					
23	9wWLF9R90BCWr6Cnj5WzQ	Olive Garden Italian Restaurant	2.95					
24	P1cCjELzSI3SqX7mPF5cCw	SakeBomber Sushi & Grill	3.79					

# Top 5 Mean Reviews By Zip

```
load_mean_rev_by_zip.py x
 1  import pandas as pd
 2  import argparse
 3  #import pandas_gbq
 4  from google.cloud import storage
 5  from google.oauth2 import service_account
 6
 7  ...
 8  |  Must export GOOGLE_APPLICATION_CREDENTIALS
 9  ...
10
11 def set_credentials(credentials_file):
12     #Credentials required to query data
13     credentials = service_account.Credentials.from_service_account_file(
14         credentials_file,
15     )
16     return credentials
17
18 def get_mean_reviews_by_zip(credentials,bucket_name):
19     q_mean_reviews_by_zip = "WITH TOP_5 AS (
20     q_mean_reviews_by_zip += " SELECT ZIP_CODE, NUM_OF_BIZ, RANK() OVER (ORDER BY NUM_OF_BIZ DESC) AS RNK"
21     q_mean_reviews_by_zip += " FROM ( SELECT "
22     q_mean_reviews_by_zip += "           SAFE.SUBSTRING(BUSINESS_ADDRESS,LENGTH(BUSINESS_ADDRESS)-5,LENGTH(BUSINESS_ADDRESS)) as ZIP_CODE, COUNT(*) AS NUM_OF_BIZ"
23     q_mean_reviews_by_zip += "           FROM `slalom-de.slalom.composition` WHERE UPPER(DAY_OF_THE_WEEK)='SATURDAY'"
24     q_mean_reviews_by_zip += "           GROUP BY ZIP_CODE)"
25     q_mean_reviews_by_zip += " )SELECT ZIP_CODE, ROUND( AVG(REVIEW_STARS),2) AS REVIEW_AVG,  FROM `slalom-de.slalom.reviews` rev, `slalom-de.slalom.composition` comp, TOP_5 "
26     q_mean_reviews_by_zip += " WHERE rev.BUSINESSID = comp.BUSINESS_ID "
27     q_mean_reviews_by_zip += " AND SAFE.SUBSTRING(comp.BUSINESS_ADDRESS,LENGTH(BUSINESS_ADDRESS)-5,LENGTH(BUSINESS_ADDRESS)) = TOP_5.ZIP_CODE "
28     q_mean_reviews_by_zip += " AND TOP_5.RNK < 6 "
29     q_mean_reviews_by_zip += " GROUP BY ZIP_CODE;"
30     #read data from bq and return dataframe
31     df = pd.read_gbq(q_mean_reviews_by_zip,project_id=bucket_name,credentials=credentials)
32
33     return df
34
35 def load_file_gcs(bucket_name,file_name,data_frame):
36     client = storage.Client()
37     bucket = client.get_bucket(bucket_name)
38
39     #create the file in the bucket
40     bucket.blob(file_name).upload_from_string(data_frame.to_csv(sep="|",index=False), 'text/csv')
41
```

# Top 5 Mean Reviews By Zip

```
load_mean_rev_by_zip.py x

42  def run():
43      parser = argparse.ArgumentParser()
44      parser.add_argument('--output_file',
45                          dest='output_file',
46                          required=True,
47                          help='Provide output file path')
48      parser.add_argument('--bucket_name',
49                          dest='bucket_name',
50                          required=True,
51                          help='bucket name for output file')
52      parser.add_argument('--credentials_file',
53                          dest='credentials_file',
54                          required=True,
55                          help='Service Account Credentials file path')
56      args = parser.parse_args()
57
58      bucket_name = args.bucket_name
59      output_file = args.output_file
60      credentials_file = args.credentials_file
61
62      #set credentials using the json file
63      credentials = set_credentials(credentials_file)
64
65      #execute the query to extract data from bq
66      df_mean_reviews_by_zip = get_mean_reviews_by_zip(credentials,bucket_name)
67
68      #create the file in gcs
69      load_file_gcs(bucket_name,output_file,df_mean_reviews_by_zip)
70
71  if __name__ == '__main__':
72      run()
73
```

# Top 5 Mean Reviews By Zip

```
(slalom-de) vijay_gopu@cloudshell:~/slalom-de (slalom-de)$ python load_mean_rev_by_zip.py \
> --output_file output/csv/mean_reviews_by_zip.csv \
> --bucket_name slalom-de \
> --credentials_file /home/vijay_gopu/slalom-de-911419ed4b1d.json
```

# Top 5 Mean Reviews By Zip

≡ Google Cloud Platform slalom-de ▾

Storage Bucket details ←

Browser Monitoring Settings

slalom-de

OBJECTS CONFIGURATION PERMISSIONS

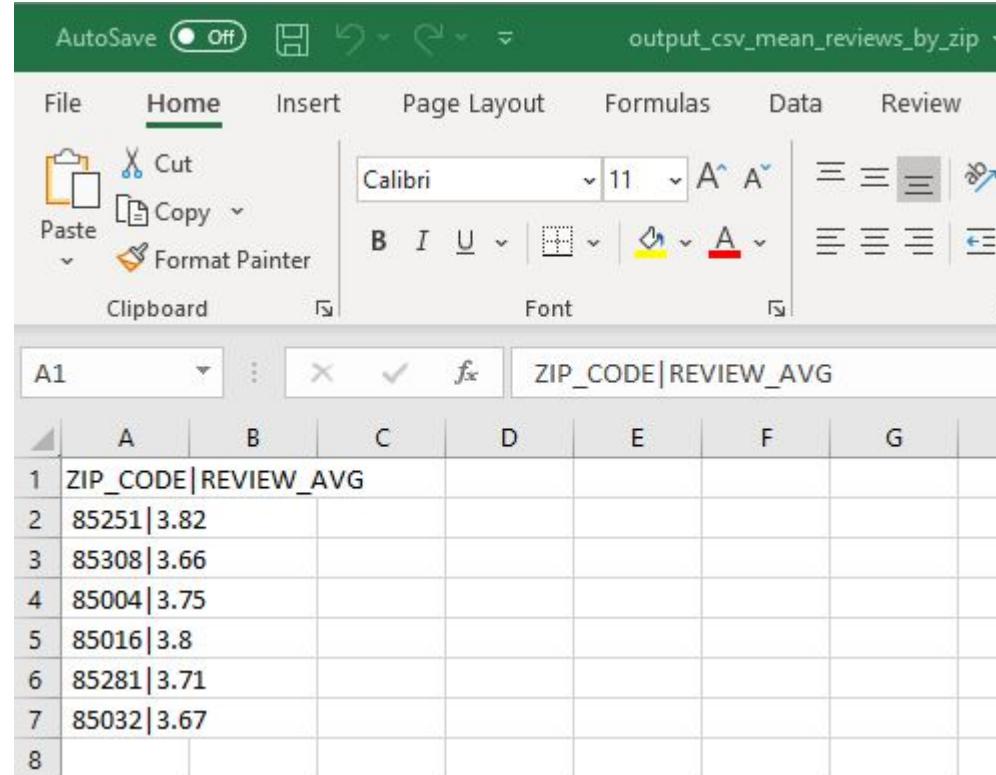
Buckets > slalom-de > output > csv

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER

Filter Filter by object or folder name prefix

<input type="checkbox"/> Name	Size
mean_reviews_by_biz.csv	255.4 KB
mean_reviews_by_zip.csv	91 B

# Top 5 Mean Reviews By Zip



A screenshot of a Microsoft Excel spreadsheet titled "output\_csv\_mean\_reviews\_by\_zip". The spreadsheet displays the top 5 zip codes with their corresponding mean review scores. The columns are labeled "ZIP\_CODE" and "REVIEW\_AVG". The data is as follows:

	ZIP_CODE	REVIEW_AVG
1	85251	3.82
2	85308	3.66
3	85004	3.75
4	85016	3.8
5	85281	3.71
6	85032	3.67
7		
8		

# Top 10 Active Reviewers

```
❶ load_top5_active_reviewers.py x
1  import pandas as pd
2  import argparse
3  #import pandas_gbq
4  from google.cloud import storage
5  from google.oauth2 import service_account
6
7  ...
8  |  Must export GOOGLE_APPLICATION_CREDENTIALS
9  ...
10
11 def set_credentials(credentials_file):
12     #Credentials required to query data
13     credentials = service_account.Credentials.from_service_account_file(
14         credentials_file,
15     )
16     return credentials
17
18 def get_top10_active_reviewers(credentials,bucket_name):
19     q_top_10_reviewers = "WITH USER_NAMES AS ("
20     q_top_10_reviewers += "SELECT"
21     q_top_10_reviewers += " DISTINCT USER_ID, USER_NAME "
22     q_top_10_reviewers += "FROM `slalom-de.slalom.users`)"
23     q_top_10_reviewers += "SELECT * FROM ("
24     q_top_10_reviewers += " SELECT USER_NAME, TOT_REVIEWS, RANK() OVER (ORDER BY TOT_REVIEWS DESC) AS RANK_ FROM ("
25     q_top_10_reviewers += "     SELECT USER_NAME, COUNT(*) AS TOT_REVIEWS FROM `slalom-de.slalom.reviews` rev, USER_NAMES"
26     q_top_10_reviewers += "     WHERE rev.User_Id = USER_NAMES.User_Id"
27     q_top_10_reviewers += "     GROUP BY USER_NAME) ) AS TOP_10 WHERE RANK_ < 11"
28     #read data from bq and return dataframe
29     df = pd.read_gbq(q_top_10_reviewers,project_id=bucket_name,credentials=credentials)
30
31     return df
32
33 def load_file_gcs(bucket_name,file_name,data_frame):
34     client = storage.Client()
35     bucket = client.get_bucket(bucket_name)
36
37     #create the file in the bucket
38     bucket.blob(file_name).upload_from_string(data_frame.to_csv(sep="|",index=False), 'text/csv')
39
```

# Top 10 Active Reviewers

```
load_top5_active_reviewers.py x
40  def run():
41      parser = argparse.ArgumentParser()
42      parser.add_argument('--output_file',
43                          dest='output_file',
44                          required=True,
45                          help='Provide output file path')
46      parser.add_argument('--bucket_name',
47                          dest='bucket_name',
48                          required=True,
49                          help='bucket name for output file')
50      parser.add_argument('--credentials_file',
51                          dest='credentials_file',
52                          required=True,
53                          help='Service Account Credentials file path')
54      args = parser.parse_args()
55
56      bucket_name = args.bucket_name
57      output_file = args.output_file
58      credentials_file = args.credentials_file
59
60      #set credentials using the json file
61      credentials = set_credentials(credentials_file)
62
63      #execute the query to extract data from bq
64      df_top10_active_reviewers = get_top10_active_reviewers(credentials,bucket_name)
65
66      #create the file in gcs
67      load_file_gcs(bucket_name,output_file,df_top10_active_reviewers)
68
69  if __name__ == '__main__':
70      run()
71  --
```

# Top 10 Active Reviewers

```
(slalom-de) vijay_gopu@cloudshell:~/slalom-de (slalom-de)$ python load_top5_active_reviewers.py \
> --output_file output/csv/top10_active_reviewers.csv \
> --bucket_name slalom-de \
> --credentials_file /home/vijay_gopu/slalom-de-911419ed4b1d.json
```

# Top 10 Active Reviewers

slalom-de

OBJECTS    CONFIGURATION    PERMISSIONS    RETENTION    LIFECYCLE

Buckets > slalom-de > output > csv

UPLOAD FILES    UPLOAD FOLDER    CREATE FOLDER    MANAGE HOLDS    DELETE

Filter Filter by object or folder name prefix

<input type="checkbox"/>	Name	Size	Type	Created time
<input type="checkbox"/>	mean_reviews_by_biz.csv	255.4 KB	text/csv	Nov 5, 2020, 8:37:48 AM
<input type="checkbox"/>	mean_reviews_by_zip.csv	91 B	text/csv	Nov 5, 2020, 10:45:37 AM
<input type="checkbox"/>	top10_active_reviewers.csv	164 B	text/csv	Nov 5, 2020, 10:54:52 AM

# Top 10 Active Reviewers

The screenshot shows a Microsoft Excel spreadsheet titled "output\_csv\_top10\_active\_reviewers". The ribbon menu is visible at the top, with "Home" selected. The "Clipboard" group contains icons for Cut, Copy, Paste, and Format Painter. The "Font" group includes Calibri, font size 11, bold, italic, underline, and alignment options. The "Font" dropdown shows "A1" as the active cell. The data starts at row 1, column A, with columns labeled "USER\_NAME|TOT\_REVIEWS|RANK\_". The data rows are as follows:

	A	B	C	D	E
1	USER_NAME TOT_REVIEWS RANK_				
2	Michael 3272 1				
3	Chris 2896 2				
4	Mike 2863 3				
5	Jennifer 2856 4				
6	John 2728 5				
7	David 2435 6				
8	Jason 2053 7				
9	Brian 2005 8				
10	Jessica 1897 9				
11	Scott 1881 10				
12					
13					