Enhanced Image Metadata Extractor

Project Members (Group B)

- Merlin Volkmer
- Bence Balázs
- Florian Johann Hafner

Language

Python & Terraform (for deployment)

Provider

Google Cloud

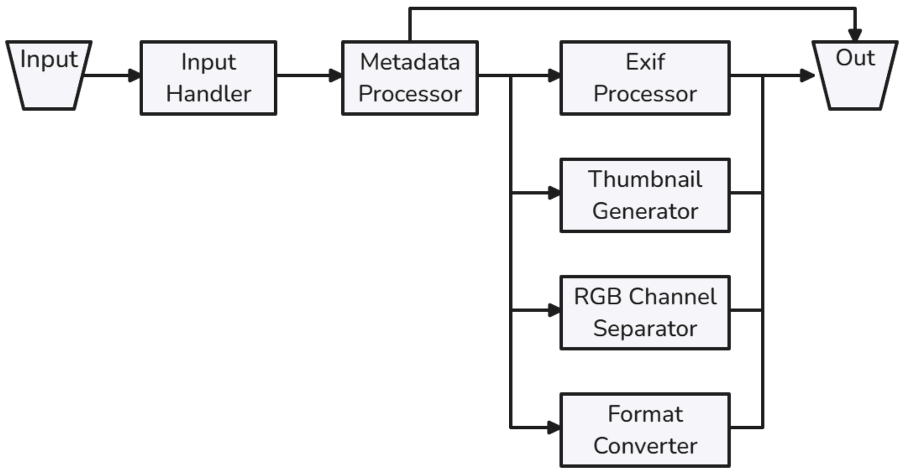
Input

An image file (png, jpg, jpeg)

Output

JSON metadata and converted images

Workflow Design



Connections between the functions

Each function is independent to the other, but it's execution is triggered by the previous step using Pub/Sub (i.e the execution of the Format Conversion is triggered by the Pub/Sub topic "convert-image")

Functions

Input Handler (`input_handler/main.py`)

Trigger: Cloud Storage event (object finalized in input bucket)
Purpose: Initial entry point for processing pipeline
Inputs:

- Cloud Event with file metadata
- File formats: JPG/JPEG/PNG only

Outputs:

- PubSub messages to `extract-metadata` topic
- Error response for invalid formats

Process:

1. Validates file extension
2. Publishes messages to metadata extraction topic
3. Hardcodes output bucket to `serverless-project-24w-output`

Metadata Extractor (`metadata_extractor/main.py`)

Trigger: PubSub message from `extract-metadata` topic
Purpose: Extract basic image metadata
Inputs:

- PubSub message with file location
- Image from input bucket

Outputs:

- JSON metadata file in output bucket (`metadata/{filename}.json`)
- PubSub messages to 4 processing topics:
  - `generate-thumbnail`
  - `convert-format`
  - `separate-rgb-channels`
  - `process-exif`

Collected Metadata:

- Format, dimensions, color mode, timestamp

EXIF Processor (`exif_processor/main.py`)

Trigger: PubSub message from `process-exif` topic
Purpose: Extract EXIF metadata
Inputs:

- Image from input bucket
- Existing metadata JSON

Outputs:

- Updated metadata JSON with EXIF data
- Handles binary EXIF data conversion

Special Handling:

- Converts binary data to strings
- Error tracking for problematic tags

Format Converter (`format_converter/main.py`)

Trigger: PubSub message from `convert-format` topic
Purpose: Convert image to multiple formats
Inputs:

- Source image from input bucket

Outputs:

- Converted images in 3 formats: JPEG, PNG, WEBP
- Metadata JSON with conversion paths
- Files stored in `converted/{filename}/` path

Note: Skips conversion to original format

RGB Channel Separator (`rgb_channel_separator/main.py`)

Trigger: PubSub message from `separate-rgb-channels` topic
Purpose: Split image into RGB components
Inputs:

- Source image from input bucket

Outputs:

- 3 channel images (red, green, blue) as PNGs
- Metadata JSON with channel paths
- Files stored in `channels/{filename}/` path

Preprocessing:

- Converts non-RGB images to RGB mode

Thumbnail Generator (`thumbnail_generator/main.py`)

Trigger: PubSub message from `generate-thumbnail` topic
Purpose: Create resized thumbnails
Inputs:

- Source image from input bucket

Outputs:

- Thumbnails in 2 sizes: 200x200 and 400x400
- Maintains original image format
- Metadata JSON with thumbnail paths
- Files stored in `thumbnails/{filename}/` path

Testing

Load Test

The load Test axis features two modes that differ in the time spent to send images and the amount of concurrent processes running for that duration.

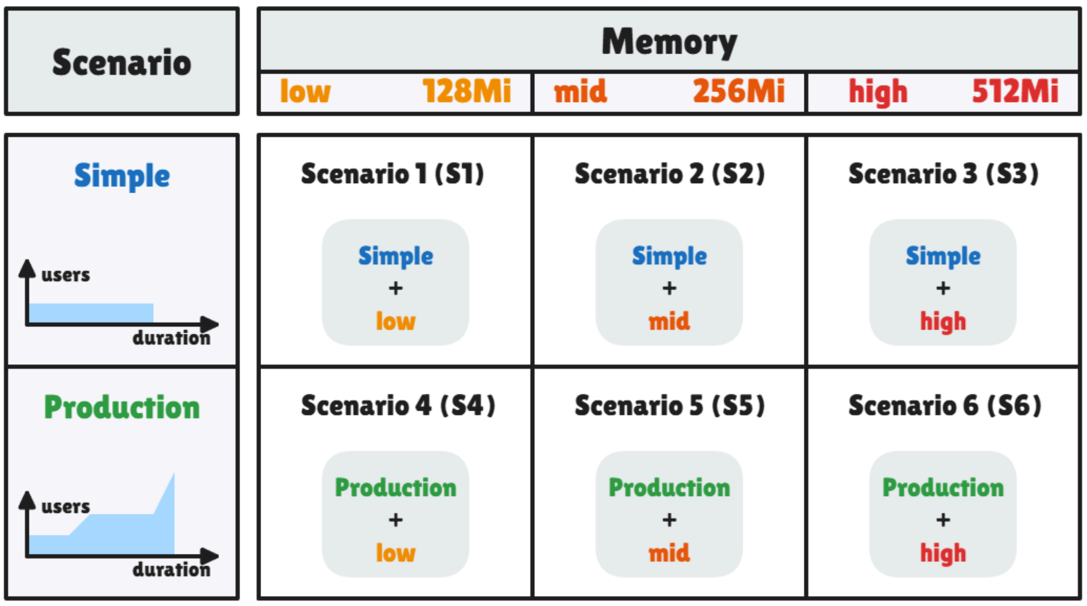
- simple
  - 5 concurrent users uploading images for 10 seconds
- production
  - baseline: 5 concurrent users uploading images for 60 seconds
  - medium: 10 concurrent users uploading images for 100 seconds
  - stress: 100 concurrent users uploading images for 3 seconds

Terraform configuration

The Terraform configuration axis features three modes that differ by the available memory allocated to each function.

- Low: 128 **MiB (Mebibytes)**
- Mid: 256 **MiB**
- High: 512 **MiB**

This results in the following six scenarios.



Scenario	Memory		
	low 128Mi	mid 256Mi	high 512Mi
Simple (users / duration)	Scenario 1 (S1) Simple + low	Scenario 2 (S2) Simple + mid	Scenario 3 (S3) Simple + high
Production (users / duration)	Scenario 4 (S4) Production + low	Scenario 5 (S5) Production + mid	Scenario 6 (S6) Production + high

Results

All data collected during the tests can be found in the results directory of the submitted archive.

Observations

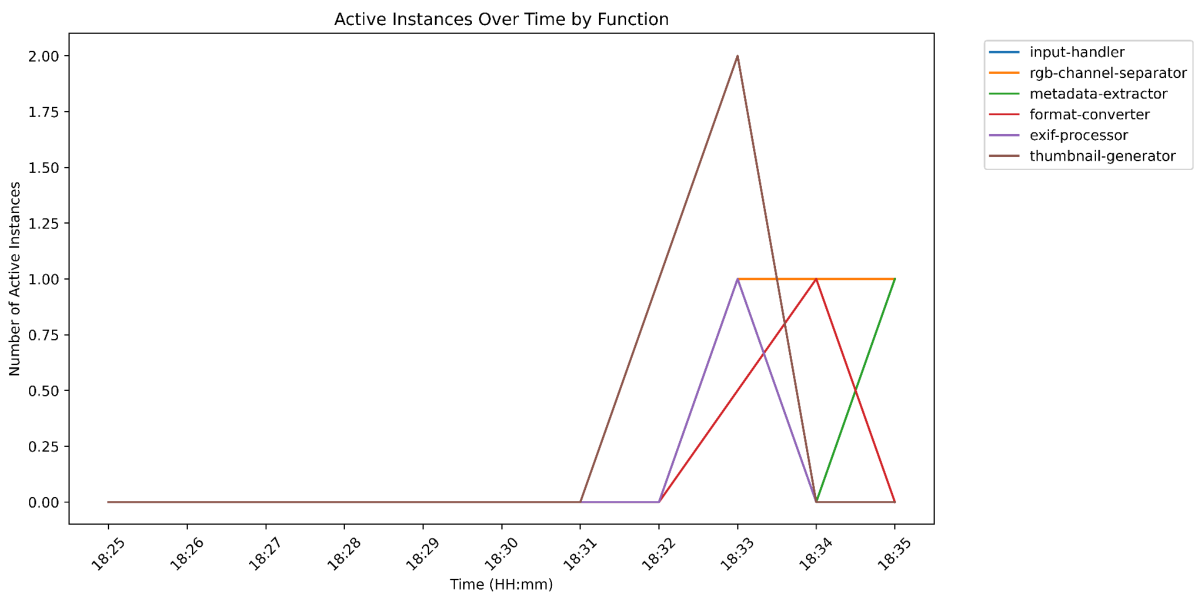
- `rgb-channel-seperator` and `format-converter` need significantly more resources than the other functions
- functions crash due to being out of memory if they are allocated less than 128 Mi

Recommendations

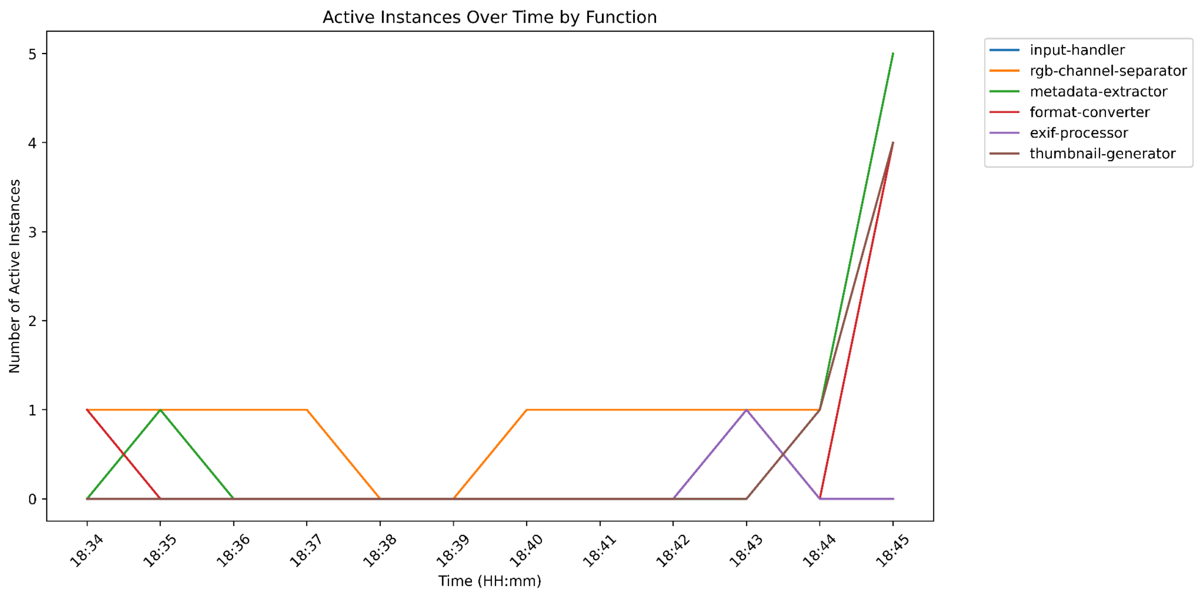
- all functions operate on images and thus need at least 128 Mi
- `rgb-channel-seperator` and `format-converter` should be allocated more resources than other functions

Active Instances

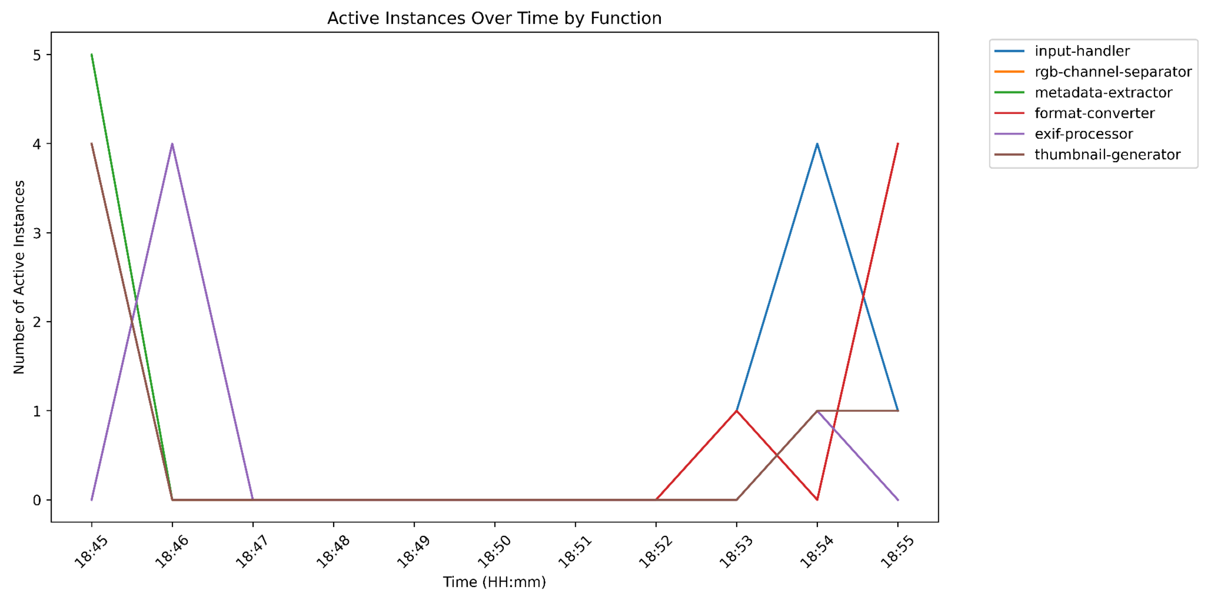
R1



R2

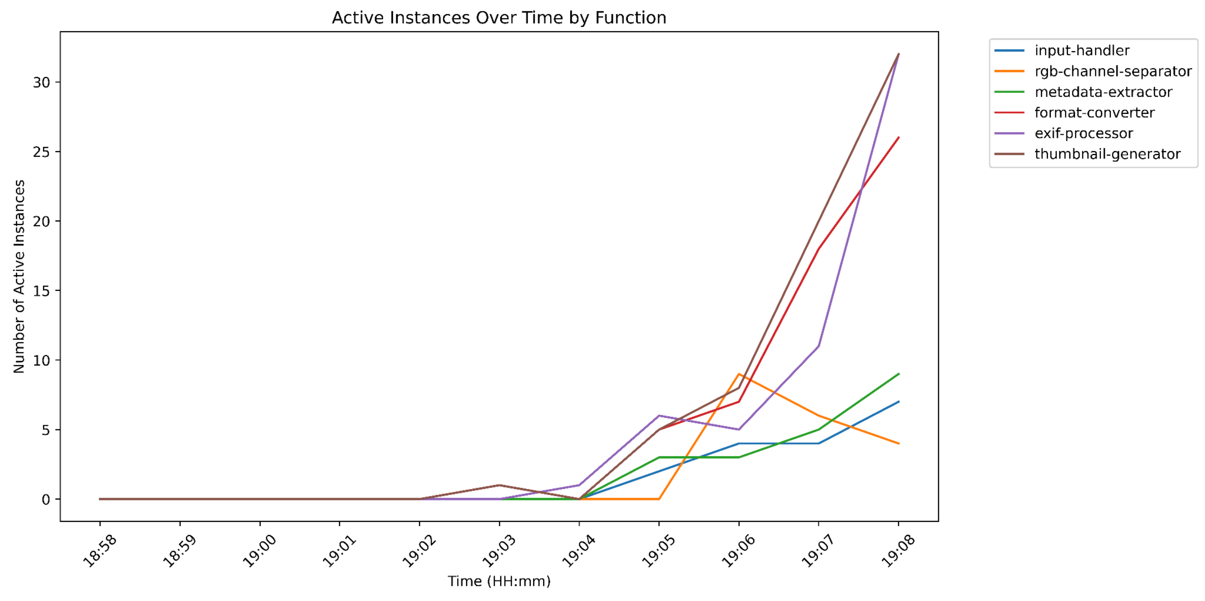


R3



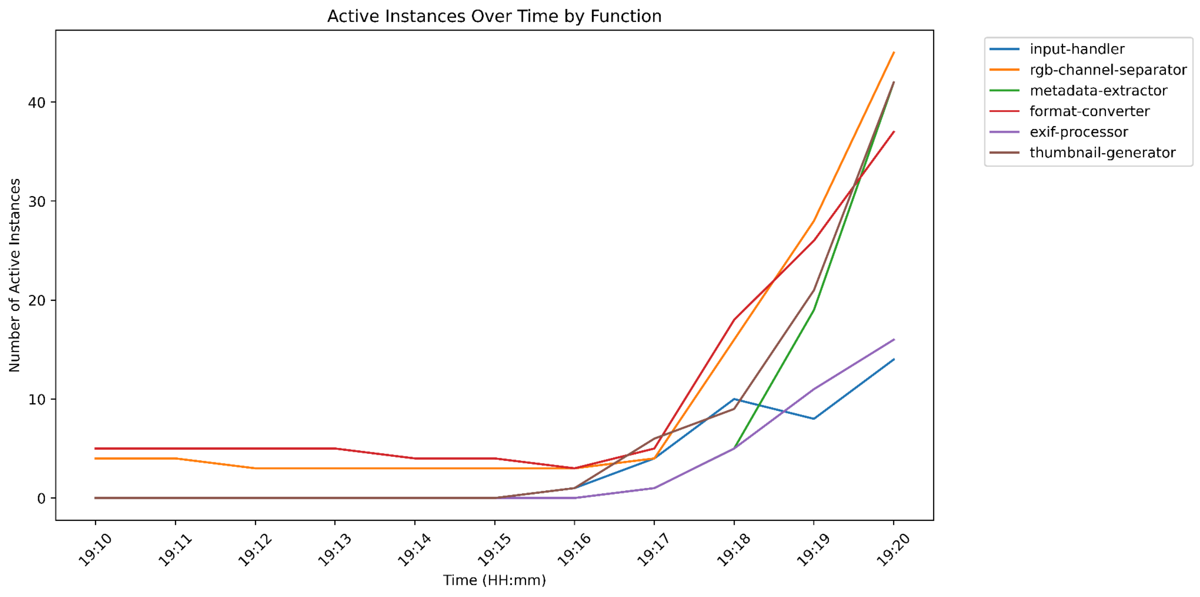
Active Instances Over Time by Function
Number of Active Instances
Time (HH:mm)
18:45
18:46
18:47
18:48
18:49
18:50
18:51
18:52
18:53
18:54
18:55
input-handler
rgb-channel-separator
metadata-extractor
format-converter
exif-processor
thumbnail-generator

R4

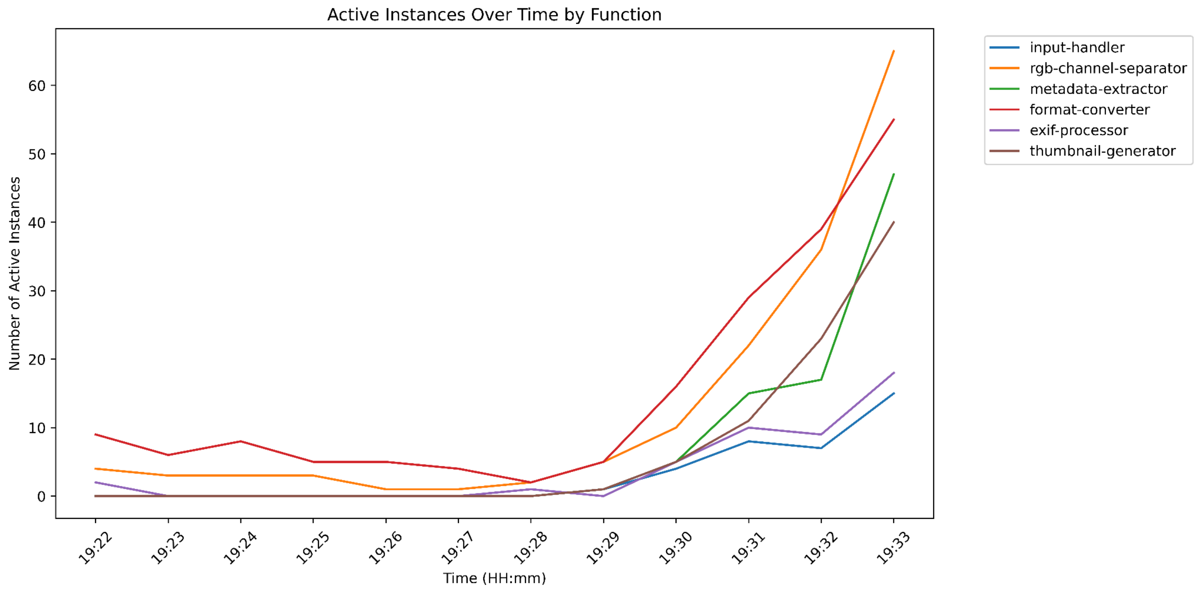


Active Instances Over Time by Function
Number of Active Instances
Time (HH:mm)
18:58
18:59
19:00
19:01
19:02
19:03
19:04
19:05
19:06
19:07
19:08
input-handler
rgb-channel-separator
metadata-extractor
format-converter
exif-processor
thumbnail-generator

R5

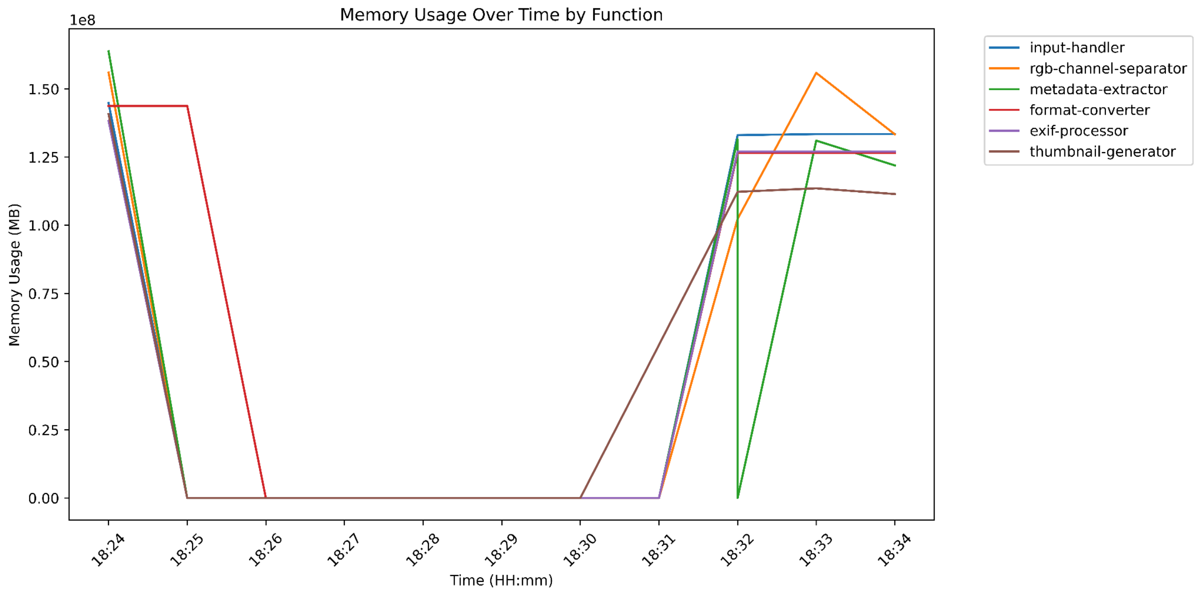

Active Instances Over Time by Function
Number of Active Instances
Time (HH:mm)
19:10
19:11
19:12
19:13
19:14
19:15
19:16
19:17
19:18
19:19
19:20
0
10
20
30
40
input-handler
rgb-channel-separator
metadata-extractor
format-converter
exif-processor
thumbnail-generator

R6

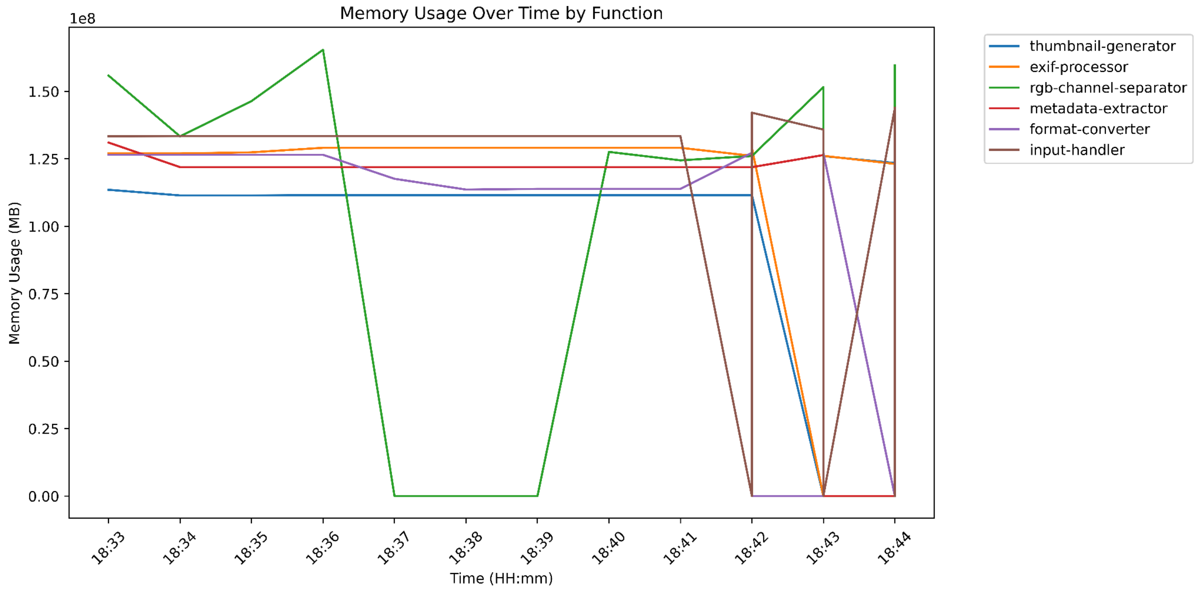

Active Instances Over Time by Function
Number of Active Instances
Time (HH:mm)
19:22
19:23
19:24
19:25
19:26
19:27
19:28
19:29
19:30
19:31
19:32
19:33
0
10
20
30
40
50
60
input-handler
rgb-channel-separator
metadata-extractor
format-converter
exif-processor
thumbnail-generator

Memory Consumed

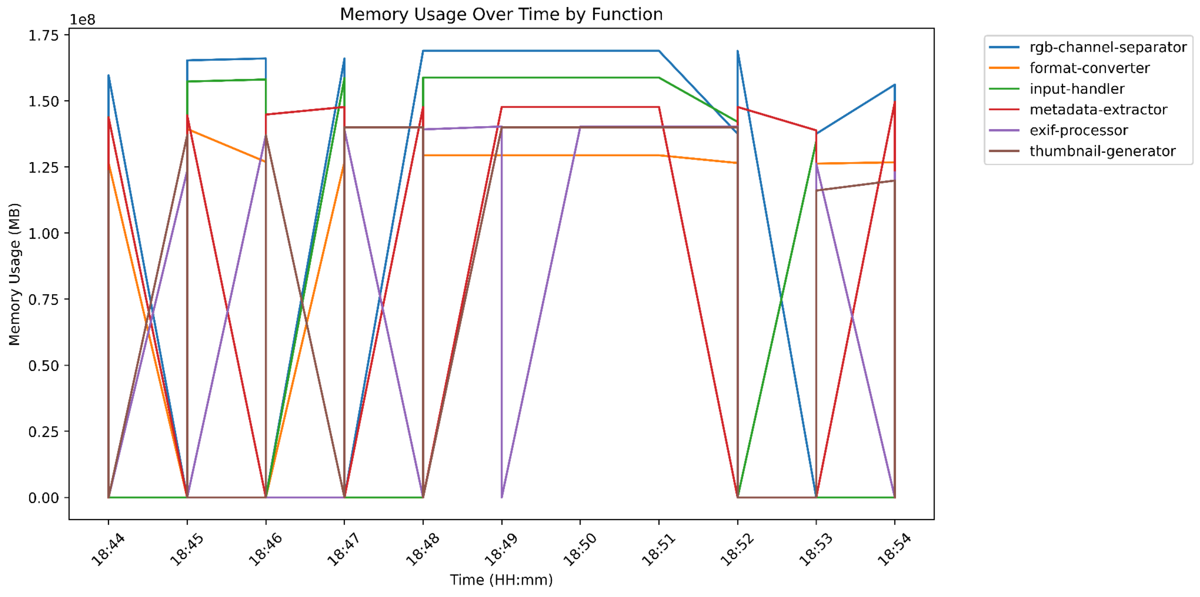
R1



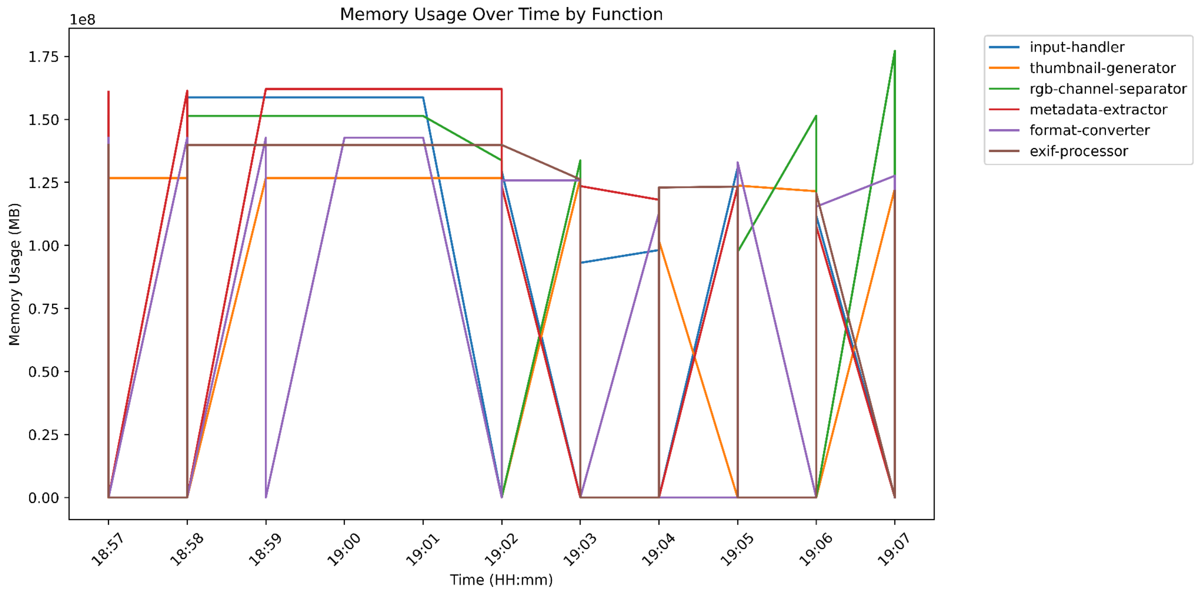
R2



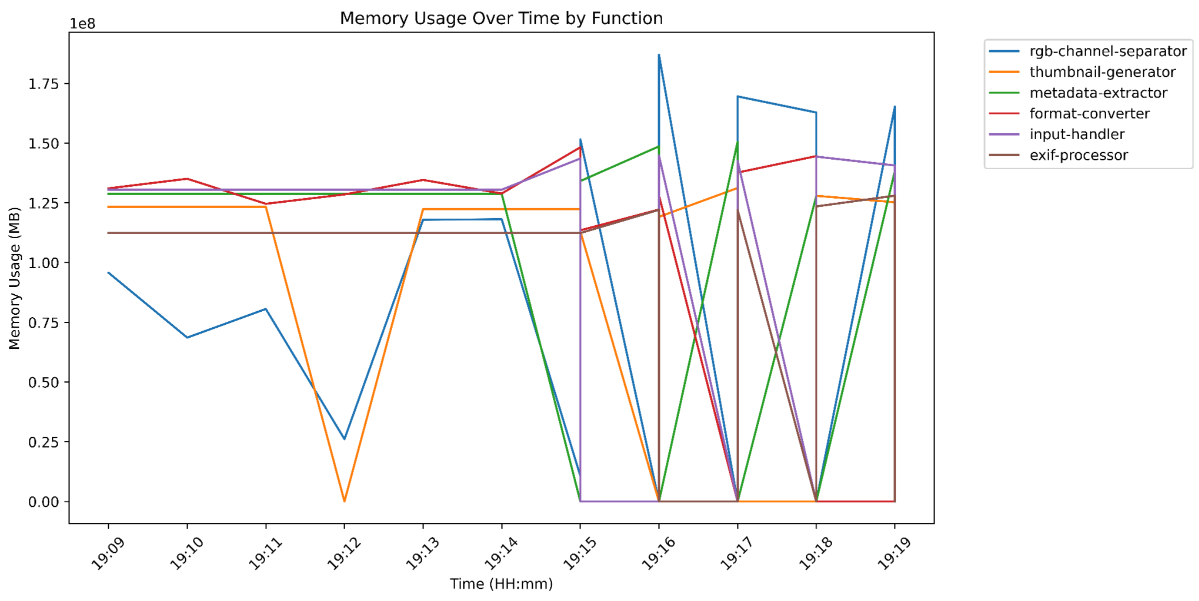
R3


Memory Usage Over Time by Function
1e8
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Memory Usage (MB)
18:44
18:45
18:46
18:47
18:48
18:49
18:50
18:51
18:52
18:53
18:54
Time (HH:mm)
rgb-channel-separator
format-converter
input-handler
metadata-extractor
exif-processor
thumbnail-generator

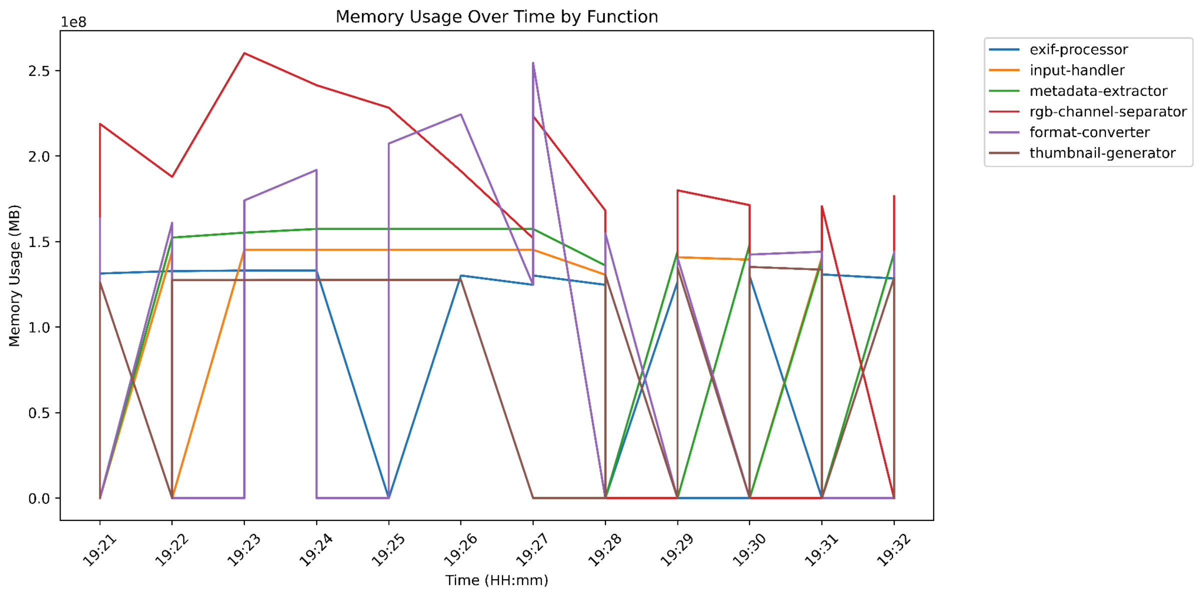
R4


Memory Usage Over Time by Function
1e8
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Memory Usage (MB)
18:57
18:58
18:59
19:00
19:01
19:02
19:03
19:04
19:05
19:06
19:07
Time (HH:mm)
input-handler
thumbnail-generator
rgb-channel-separator
metadata-extractor
format-converter
exif-processor

R5



R6



Challenges, Considerations and Closing Words

The primary challenge encountered during the project solely related to GCPs IAM configurations, specifically in establishing the necessary permissions of the cloud user accounts for communication between the different triggers, pub/sub topics, storage events and cloud functions.

Despite these early setbacks, the project provided valuable insights for all of us into the advantages and challenges associated with serverless architectures and cloud-based computing.