# Advanced Image Downloader

Naveen Pujar

# Objective

Develop a bulk image extractor job which would be able to download required amount of images. Once the job is completed, an URL would be sent to the user's email address to download specified amount of images in one click.

## Benefits:

- Can download any number of images on one click.
- User can schedule the time at which he/she wishes to get the email with the downloadable link.
- High quality images for training deep learning models.
- No manual download of images.

# Team members & Duration

- 1 Product manager
- 1 Solution architect
- 1 Lead
- 2 Dev – ops engineers
- 2 QA Engineers
- 2 UI Developers
- 3 Python develops
- Project duration was 2 months

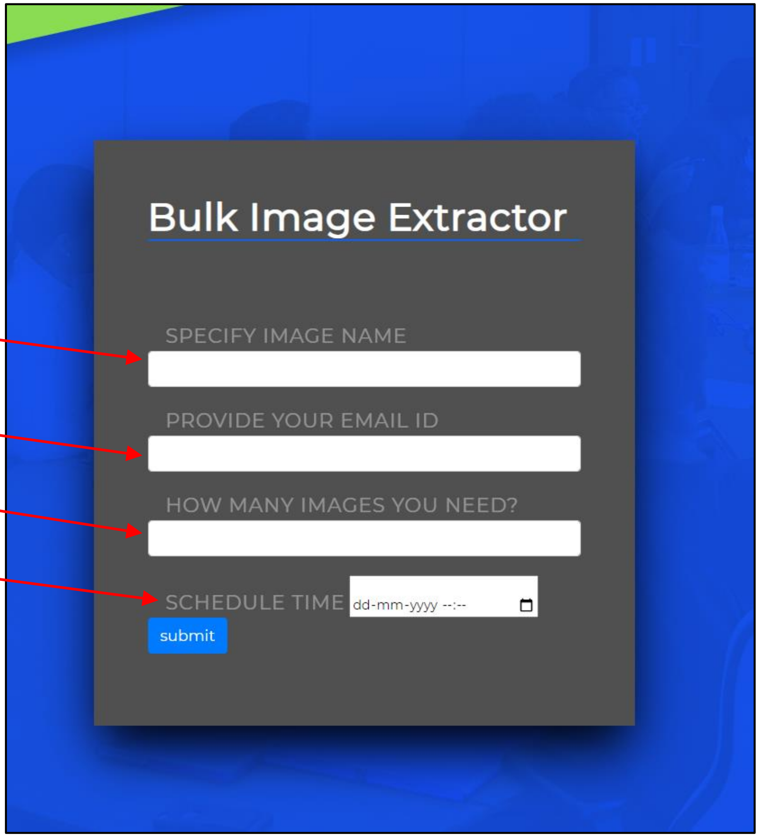
# Development strategy

❑ <u>User request through Front end</u>

Requestor must provide a desired inputs such as

- Search string
- email id
- number of images
- scheduled time

To enable this user interface is developed using HTML and CSS.



Bulk Image Extractor

SPECIFY IMAGE NAME

PROVIDE YOUR EMAIL ID

HOW MANY IMAGES YOU NEED?

SCHEDULE TIME dd-mm-yyyy --:--

submit

# Development strategy

❑ Web scraping

Once the search keyword has been submitted at the front-end, selenium web driver will launch a browser and starts downloading images and would store them in a folder. This folder will be later zipped for further uploading to a cloud storage service.

❑ Upload the Zipped Folder to S3 Bucket and Obtain URL link

The zipped folder containing images would be uploaded to the S3 bucket. S3 Bucket provides feasibility to fetch PRE-Signed URL of an object available in the bucket. This URL can be passed on to the requestor to download the file in their respective local system.

# Development strategy

❑AWS Lambda & Email

The front-end part of this project requests user to specify the time at which they would want to get the URL for downloading the images to their email. AWS Lambda service has been included within the code for sending emails at the instance of time.

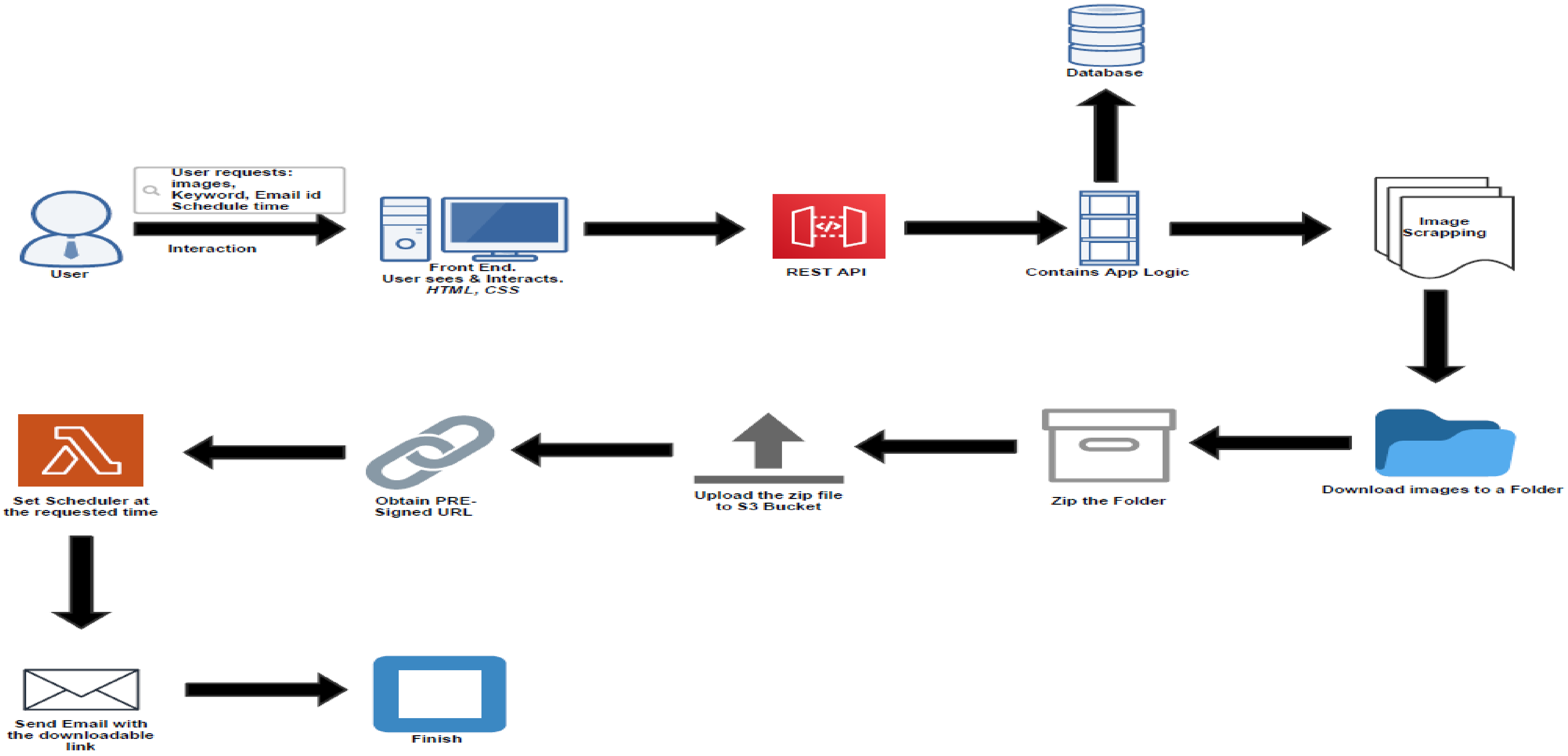
❑Database

Cassandra database will be used to store the entries made by the user at the front end and logging details.

# Development strategy

- ❑User Input Table – Once any user makes a request is at the front end, a table will be created in the Cassandra database (DataStax) with the table name identical to the search string which was inserted in the input field.
- ❑Log Table – for every request a log table (log_table) will be created and the table will be populated with the log entries.
- ❑Deployment- AWS ESC has been used for the deployment of this project.

# Architecture



# Tech Stacks used in the Project

- VS Code is used as IDE.
- AWS is used for deployment of the model.
- Cassandra is used to save inputs from requestor and logging data.
- Front end development is done using HTML/CSS
- Python Flask is used for backend development.
- GitHub is used as version control system.
- AWS Lambda is used to trigger email at the scheduled time.
- S3 to store the images
- Selenium to scrape the image

# FAQ

Q1) Once the images are uploaded to S3 bucket, what will happen to the folder stored locally?

Ans:- The folder will be deleted as it is of no use and consumes storage space.

Q2) Which package has been used for scraping the images?

Ans:- Selenium web driver has been used for scraping the images.

Q3) Who is most benefited by this project?

Ans:- For training deep learning models, a deep learning engineer would need thousands and thousands of images. Since manual extraction of images is laborious, in such cases this project can save lot of time. Since it can download any number of images in one shot.

Q4) How much time it will take to download?

Ans:- Depending upon the number of images to download, it would take 1 to 7 minutes

# FAQ

Q5) Which services has been used to store images in cloud?

Ans:- AWS S3 bucket has been used as it provides additional functionality to obtaining the URL link to download.

Q7) How many images can be downloaded?

Ans:- Currently we have tested it with thousands of images. For sure a 1000 images can be downloaded.

Q8) How are we handling  email addresses?

Ans:- Email addresses are stored only for the purpose of sending the email of the downloadable link. Email record will be stored in the database.

Q8) How logs are managed?

Ans:- For every request, a log table (log_table) will be created in the database and the table will be populated with the log entries.

*END*