Advance Image Downloader/Extractor

Project submitted by –

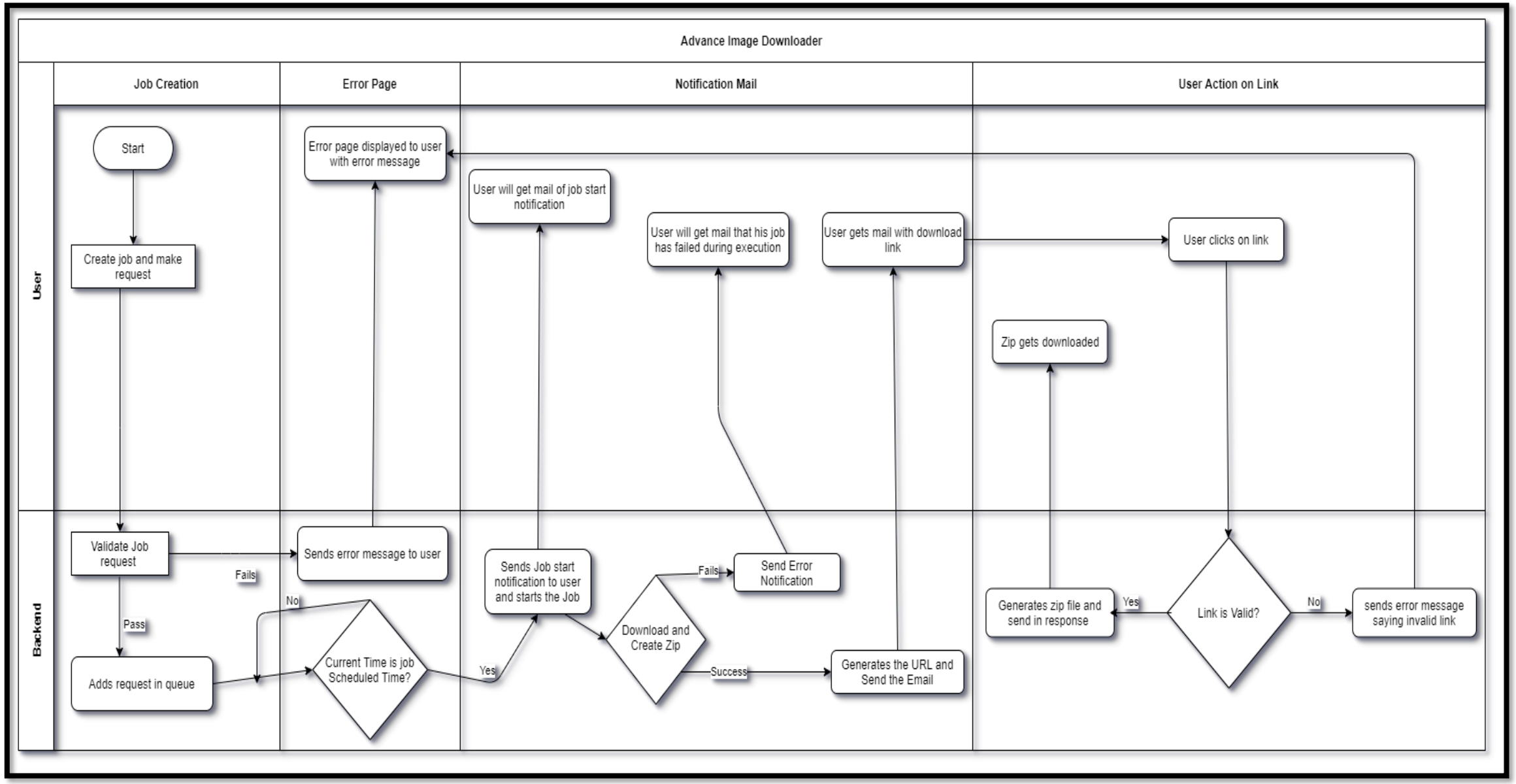
Shreyas Parab

Harshad Kadam

Objectives

- In this time, the images are the most important data source. Be it a training Computer vision model on this, Finding the appealing wallpaper images, going through hundreds of crafts and arts varieties on single click or finding the news related to specific company for market analysis images are crucial in this scenario.
- Advance Image Downloader/Extractor(Job) does exactly what it says, it can download up to 500 images of any kind at any date and time. User just have to submit the query and the download link will be ready to download the images once process is completed.

Architecture



User Input

- In the user input process, user will be treated with the html page for inserting the search query which will contain inputs like Search query term, Date, Time and the Number of images which the user wants to download.
- Once this is filled in, user can click on submit button to submit the query.

Validating Input

- Validation is done both at the front end and backend level.
- Email address and Date validation is carried out at the frontend.
- For handling scenarios where the user inserted the past date and time, we are treating such situations at the backend site by validating the date-time provided by the suer with the current date-time.

Job Scheduling

- If the requested job is validated for both frontend and backend validation, then the job is queued to be scheduled.
- Once the current date-time is equal to the posted job date-time then the process will start its execution.

Web Scrapping

- For getting the images data, web scraping is carried out using selenium.
- Chrome drivers are responsible for getting the data from Google's search engine and then storing that image data into the databases.

Database

- Database creation and connection – Create the database with the keyspace name passed (If it's not already present). Connect with the database. Create the table in that keyspace (If it's not already present).
- Database Insertion – Selenium will fetch the data from the internet and the links of the images will then be stored into the Database for further processing..

Email

Users can get an email notification a maximum of three times

- a. Job Start – Once the current time is equal to the scheduled time, the user will get a notification about the job starts.
- b. Job End – Once the file gets downloaded from the internet user will get the download link for downloading the images.
- c. Job Error – If some error occurs at the backend side, the user will get notified by this by sending him an email notification about the job failure.

Files and Records

- Image files will get downloaded and store on the server side. Since to reduce the size of the images, zip functionality has been added.
- After some predefined time interval, the files which are created at the backend and the database records will get deleted from the system upon which if the user tries to click on the same link, users will get ‘Link has expired’ error.

Download Link

- Once the zip file is created, then a download link will be sent to the user.
- This link will be active for some interval of time.
- Users can download within this time interval.
- After crossing the time limit, the user will get a ‘Link has expired’ error from the frontend.

Q&A

1. Do users have to wait until images get downloaded?

Ans :- No. User can simply submit the job and then close the browser and continue doing other tasks. Once the task starts user will get an email about the process activities.

2. How much images can be downloaded at single go?

Ans :- Up to 500 images can be downloaded at single time. If more required we can simply submit the query again.

3. How will I know if my job suddenly stops in the background?

Ans :- If some error occurs at the backend the user will get email notification informing about the issue and also the link to again retry scheduling the job again.

4. How are we handling email addresses?

Ans :- Email addresses are store only for the purpose of sending the email of the downloadable link. After some interval of time, all the records of the request will get deleted from the databases.