**Personal Email Financial Analysis**

1. Introduction
   1. Purpose

This project is intended to extract email data based on the subject, originator email, body to have more specific data pattern in a report manner so that can be used onward decision tasks in regular basis

* 1. Scope

Using personal email especially in domain GMAIL this project is initiated, with limitations in categorising in extracting and analysing which is about billing statements or related expenses email.

1. Background

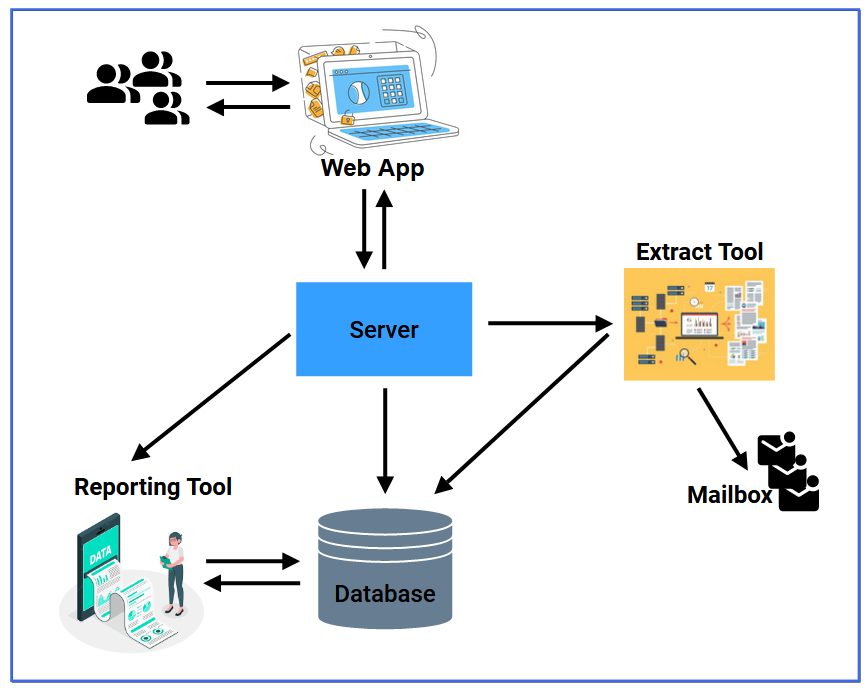
In daily operational tasks having unorganised email needs some allocated times to categorise each email manually one by one which is not efficient. Moreover, it will still lead to pending and abandoned requests or solutions. To make it easier this approach will help to auto-process emails with some predefined categories.

As a part of this prototype the main goal is solely how to give the basic knowledge and understanding how the process is being chosen. Though there are many options in choosing the study case for data science exercise, still using this old and mediocre type of working in email opens up many ideas. As this project will narrow down to extract and analyse email monthly basis expenses only thus opening others to explore more in this research.

1. System Architecture

Objectively, the concept of this research is to build a prototype of a smart personalised financial report system from a mailbox. Users can analyse through a web portal to generate a financial report of total expenses reported in their personal mailbox. As seen in the diagram below, the user will provide the account name and application password generated by the domain server so that can access the tool. By having the correct credential the extraction tool will collect the clean and sufficient data to be stored in the database in a period time depending on the volume.

Next process is when the user needs to access the report, some logic from the reporting tool will display the report in the web portal.



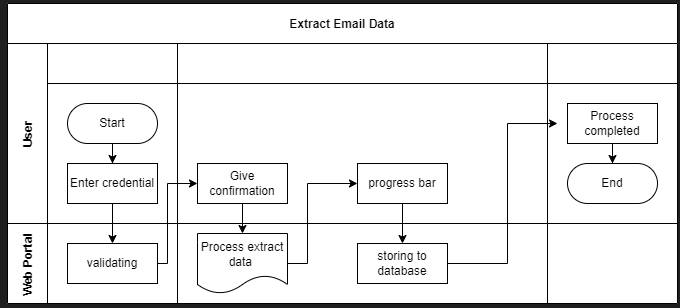
1. System Requirements
   1. Technology Requirements
      1. The project is using Python as the main code development

(TBA)

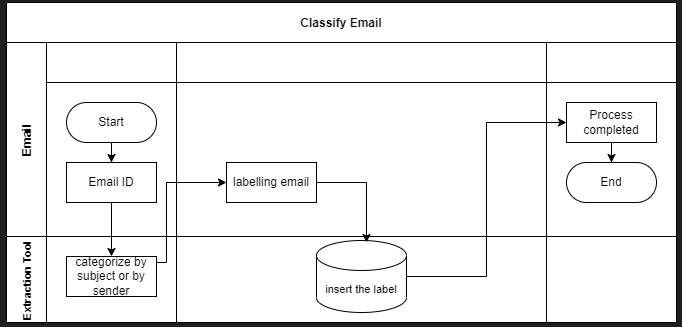
* + 1. The project using MySQL as the main database

(TBA)

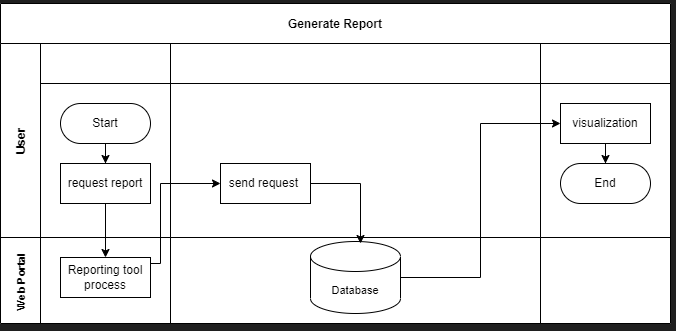
* 1. Functional Requirements
     1. The project will be able to extract the unstructured email data into database



* + 1. The project will be able to classify email into different categories



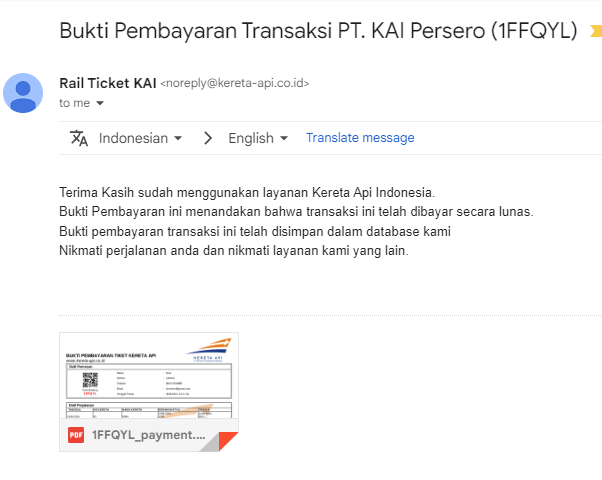
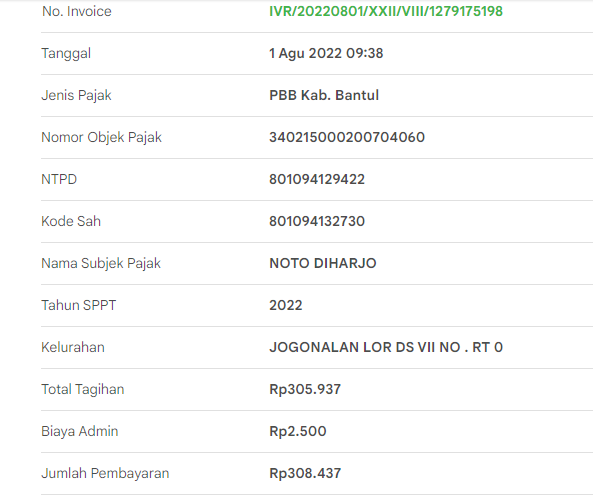
* + 1. The project will be able to count total monthly payment for each category



1. Approach
   1. Email classification

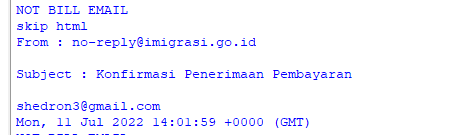
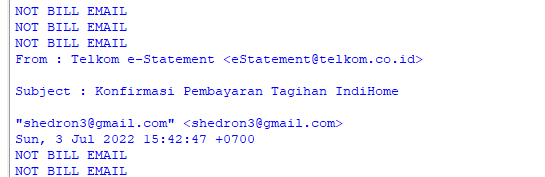
All kinds of email are being retrieved in the mailbox at once. To make it a simplified model this category is needed. Especially In the personal mailbox which is being used as the sample of this project, many types of categories exist. There are job vacancies, promotions, entertainment, hobbies and many more. Usually manual and time-consuming activities are being carried out to handle this process by labelling or grouping differentiated by email id. Such logic will be implemented to separate in classifying email focused on financial related email as billing /expenses email health payment, internet payment, shopping payment and other types shown as below:





* 1. Extract unstructured email data

Using small code Python the emails are being extracted by categorising the Subject by word ‘Payment’ or ‘Billing’ or ‘Tagihan’ to get below data; originator email, subject, destination email, date and body. Those data will be used for analysis. Especially in extracting the body to find out the format in each email to be as below, only data in the body is not shown as part of collecting data. Specifically, from body of each email, total paid amount and other information can be extracted



Below expected structured data

From : no-reply@imigrasi.go.id

Subject : Konfirmasi Penerimaan Pembayaran

To:shedron3@gmail.com

Date :Mon, 11 Jul 2022 14:01:59 +0000 (GMT)

Total amount paid : Rp.500.000

Category : government

From : Telkom e-Statement <eStatement@telkom.co.id>

Subject : Konfirmasi Pembayaran Tagihan IndiHome

To: "shedron3@gmail.com" <shedron3@gmail.com>

Date:Wed, 3 Aug 2022 06:51:47 +0700

Total amount paid : Rp.300.000

Category : communication

From : NOTIF UM UGM <noreply-simaster@ugm.ac.id>

Subject : [Pembayaran]

To:shedron3@gmail.com

Date:Mon, 14 Mar 2022 06:47:38 -0700 (PDT)

Total amount paid : Rp.500.000

Category : education

From : "Shopee" <info@shopee.co.id>

Subject : Pembayaranmu Berhasil Dikonfirmasi

To :<resmiatisari03@gmail.com>

Date: Wed, 22 Jul 2022 10:38:00 +0800

Total amount paid : Rp.300.000

Category : shopping

From : ZALORA <order@updates.zalora.co.id>

Subject : Pembayaran untuk order 236366957 telah kami terima

To:resmiatisari03@gmail.com

Date:Tue, 12 Jul 2022 09:01:19 +0000 (UTC)

Total amount paid : Rp.400.000

Category : shopping

1. Data Analysis Method
   1. Qualitative Analysis

This method is being used to categorise mailboxes into more specific groups based on their similar source. Below table will show results of a classification process used in this project.

|  |  |  |
| --- | --- | --- |
| **Subject** | **Sender** | **Category** |
| Konfirmasi Penerimaan Pembayaran | [no-reply@imigrasi.go.id](mailto:no-reply@imigrasi.go.id) | Government |
| [Pembayaran] | NOTIF UM UGM <noreply-simaster@ugm.ac.id> | Education |
| Konfirmasi Pembayaran Tagihan IndiHome | Telkom e-Statement <eStatement@telkom.co.id> | Communication |
| Pembayaran untuk order 236366957 telah kami terima | ZALORA <order@updates.zalora.co.id> | Shopping |

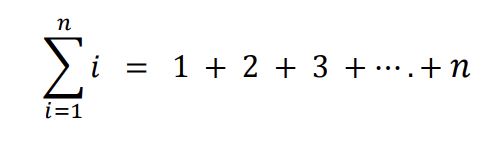
Government category is defined by sender email containing “go.id”, meanwhile Education category uses “ac.id” identifier. Other categories will be classified using master data.

* 1. Quantitative Analysis

This method is being used to count total monthly expenses based on billing notified to the mailboxes. By using the same categorization, the total amount will be calculated shown as below table:

|  |  |  |  |
| --- | --- | --- | --- |
| **To** | **Date** | **Total**  **Amount Paid** | **Category** |
| shedron3@gmail.com | Mon, 11 Jul 2022 14:01:59 +0000 (GMT) | Rp.500.000 | Government |
| shedron3@gmail.com | Mon, 14 Mar 2022 06:47:38 -0700 (PDT) | Rp.500.000 | Education |
| shedron3@gmail.com | Wed, 3 Aug 2022 06:51:47 +0700 | Rp.300.000 | Communication |
| resmiatisari03@gmail.com | Tue, 12 Jul 2022 09:01:19 +0000 (UTC) | Rp.400.000 | Shopping |
| resmiatisari03@gmail.com | Wed, 22 Jul 2022 10:38:00 +0800 | Rp.300.000 | Shopping |

The sum of whole numbers can be represented in the following manner:



For email resmiatisari03@gmail.com, total expenses along July 2022 is

Rp.400.000 + Rp.300.000 = Rp.700.000