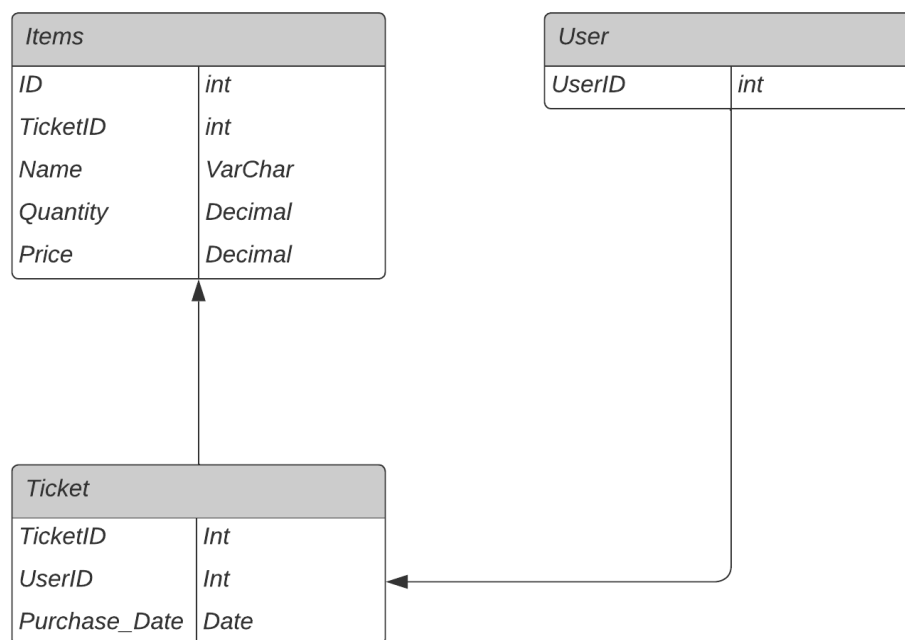There are four steps to the process:

1. We run the tickets through a script that consists mainly of this code ⬚1 that uses tesseract. This script is based on pre-processing the image in various ways, such as converting the colors of the photos to grayscale so that the character recognition will be more accurate later.

2. The script returns a total of 4 OCR versions with different pre-processing and manually chooses the best one to send later to the AI, in this case ChatGPT 4.0, which interprets the OCR text to return the sql code used to insert the ticket information into the database. The prompt for this is ⬚2

3. The database looks like this:

Items	
ID	int
TicketID	int
Name	VarChar
Quantity	Decimal
Price	Decimal

User	
UserID	int

Ticket	
TicketID	Int
UserID	Int
Purchase_Date	Date

4. Once all the data is uploaded to the database, the recommendation is made. To do this, a prompt ⬚3 is given to the AI along with the first report and the complete list of items of the corresponding user taken from the database.

As for related papers, in paper [4] they use a very similar procedure but they use an HP Capshare to scan the tickets and they use database [5] to get information about nutrition.

On the other hand in paper [6] They use office scanners to scan the tickets which allows them to have better OCR results because the photos of the tickets are consistent (optimized photo resolution at 300 dpi, less noise produced...). They use fuzzy matching and regular expressions to solve string recognition problems. Very focused on Human in the loop (HuIL) because if any kind of error occurs, the user is notified to correct it. They mention food dataset with categories although they do not mention it in the references.

In 7 they try to make the application as simple as possible since their target audience is people over 70 years old. They also mention UNESCO Thesaurus [8] together with OCR.

In [9] They use Azure Form Recognizer service to recognize the text of the tickets, Open food facts for the products and a German database for those products bought by weight.

[1] https://nanonets.com/blog/ocr-with-tesseract/#ocr-with-pytesseract-and-opencv

[2] Prompt: "Having a database like this: "-- Create the 'User' table
```sql
CREATE TABLE User (
    UserID INT PRIMARY KEY
);

-- Create the 'Ticket' table
CREATE TABLE Ticket (
    TicketID INT PRIMARY KEY,
    UserID INT,
    Purchase_date DATE,
    FOREIGN KEY (UserID) REFERENCES User(UserID)
);

-- Create the 'Items' table
CREATE TABLE Items (
    ID INT PRIMARY KEY,
    TicketID INT,
    Name VARCHAR(255),
    Quantity DECIMAL(10,4),
    Price DECIMAL(10, 2),
    FOREIGN KEY (TicketID) REFERENCES Ticket(TicketID)
)",
```
you will receive the OCR data of scanned tickets. You will respond with the sql code to insert the data into the database. The User will be User1 until stated otherwise. If the name of an item is not found you should fill it with related information in the ticket and enumerate them if there are several of them. The ID of items ID will start at 1 and go increasing for each item added. The ticketID will start at 1 and go increasing for each ticket added.
"

[3] Prompt: Based on the initial list of products I shared with you, the report associated with these tickets, and the new list of food items purchased. "Make a descriptive and prospective report on the nutrition and sustainability impact of the food items provided per user with special emphasis on the recommendations to enhance sustainability and nutrion". Be sure that you connect the report to specific products purchased in the list I provide to you." The sections I want you to provide are:
*Nutritional Overview: * This section highlights the variety and type of food items purchased, such as fresh produce, whole grains, lean proteins, legumes, and processed items. It assesses the balance of the diet and points out areas for improvement.
* Sustainability Insights: * This part evaluates the environmental impact of the consumer's choices, considering factors like packaging waste, processing, and the sourcing of items, especially organic or locally sourced products and seafood.
* Recommendations: * Based on the nutritional overview and sustainability insights, this section provides suggestions to improve diet quality and promote more sustainable shopping practices. Recommendations focus on incorporating more whole foods, reducing processed items, and choosing eco-friendly packaging and sourcing options, to cite some.
* Conclusions: * A summary that reflects the overall dietary habits and shopping practices of the consumer, acknowledging the positive aspects and suggesting future directions for enhancing both nutritional quality and sustainability.""

[4] https://faculty.washington.edu/garyhs/docs/mankoff-ubicomp02-shoppersAid.pdf

[5] Nutrient Data Laboratory. Usda nutrient database for standard reference, release 14 Nutrient Data Laboratory Web Page https://fdc.nal.usda.gov/download-datasets.html.

[6] https://dl.acm.org/doi/pdf/10.1145/3377325.3377489

7 https://www.researchgate.net/profile/Beatriz-Sainz-De-Abajo/publication/340496024_FoodScan_A_New_mHealth_App_for_Food_Monitoring_through_the_Analysis_of_Purchase_Groceries_Receipts_Preprint/links/603f36d7299bf1e078513f36/FoodScan-A-New-mHealth-App-for-Food-Monitoring-through-the-Analysis-of-Purchase-Groceries-Receipts-Preprint.pdf

[8] https://vocabularies.unesco.org/browser/thesaurus/es/?clang=en

[9] https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/6421307e-2412-4199-9a91-c6b5e90424d2/content