AUTOMATIC ANSWERING SERVICE FOR CORONAVIRUS QUESTION

Hello!

SAYANTAN PAL

COLLEGE

HERITAGE INSTITUTE OF TECHNOLOGY

B.TECH - COMPUTER SCIENCE & ENG.

2018-2022

MITACS GLOBALINK RESEARCH INTERN 2021

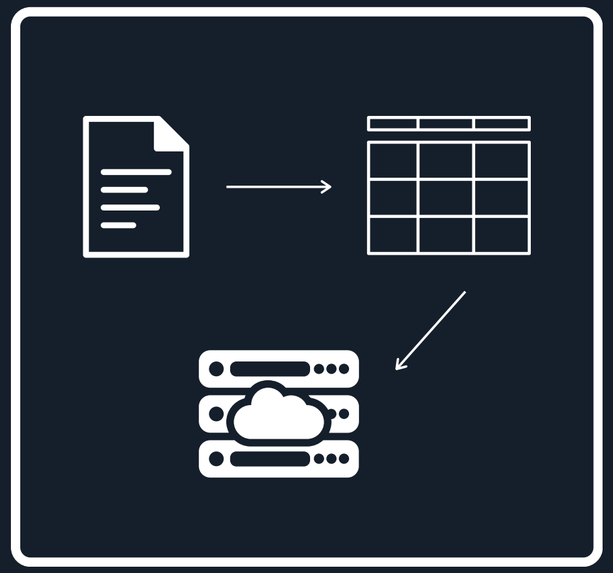
AUTOMATIC ANSWERING SERVICE FOR CORONAVIRUS QUESTION

RESEARCH GOAL

The research goal is to have an **automatic** answering service that correctly identifies the **keys** from a question. It summarizes the associated content that is relevant to the question and makes the user satisfied.

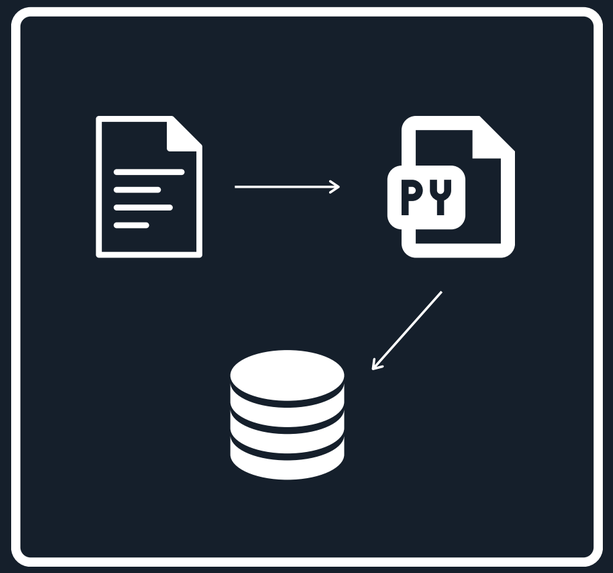
STAGES

1



FILE EXTRACTION
AND VERIFICATION

2



DATA PROCESSING

3

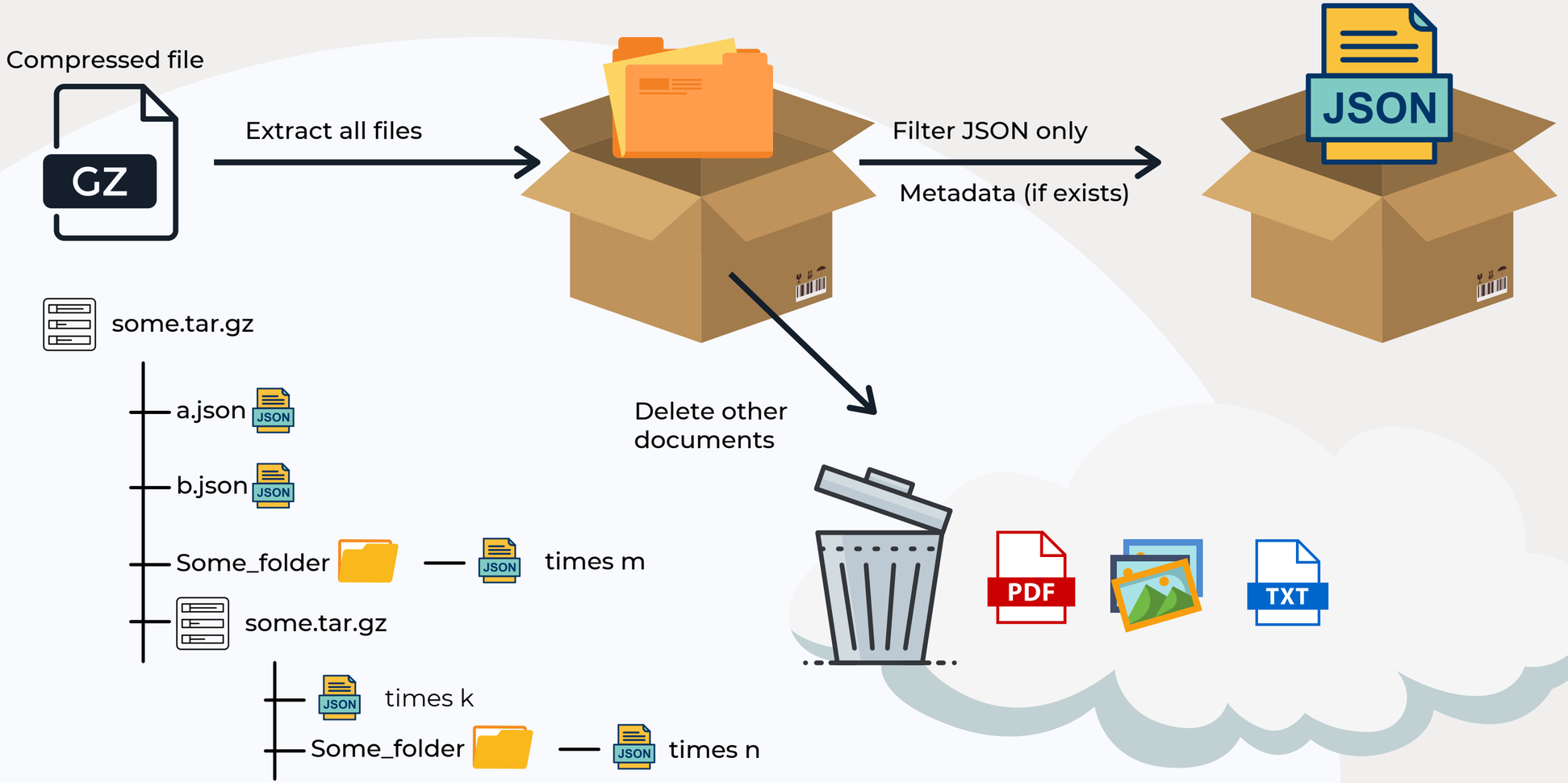
SUMMARY
GENERATION

STAGE 1 - REQUIREMENTS

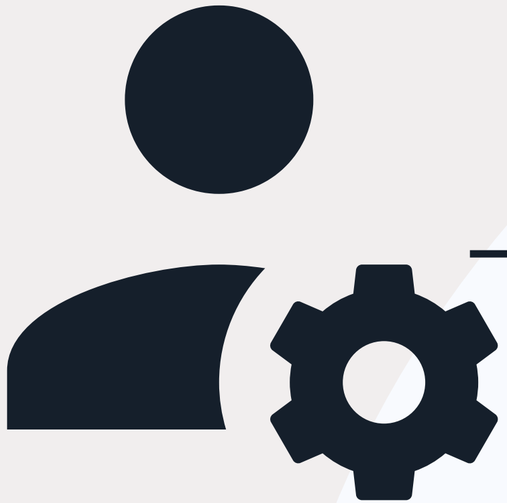
LANGUAGES AND FRAMEWORKS

- HTML 5
- CSS 3
- JAVASCRIPT
- PHP 7.2.24
- BOOTSTRAP FRAMEWORK 4
- MYSQL
- AJAX

STAGE 1 OVERVIEW - FILE EXTRACTION AND VERIFICATION

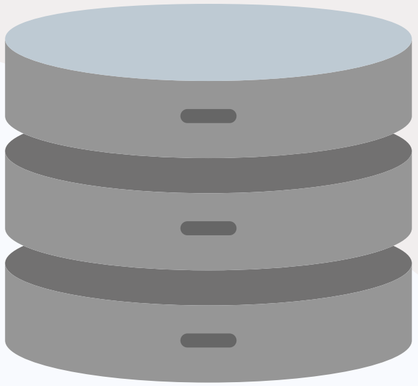


STAGE 1 - MANAGEMENT DASHBOARD



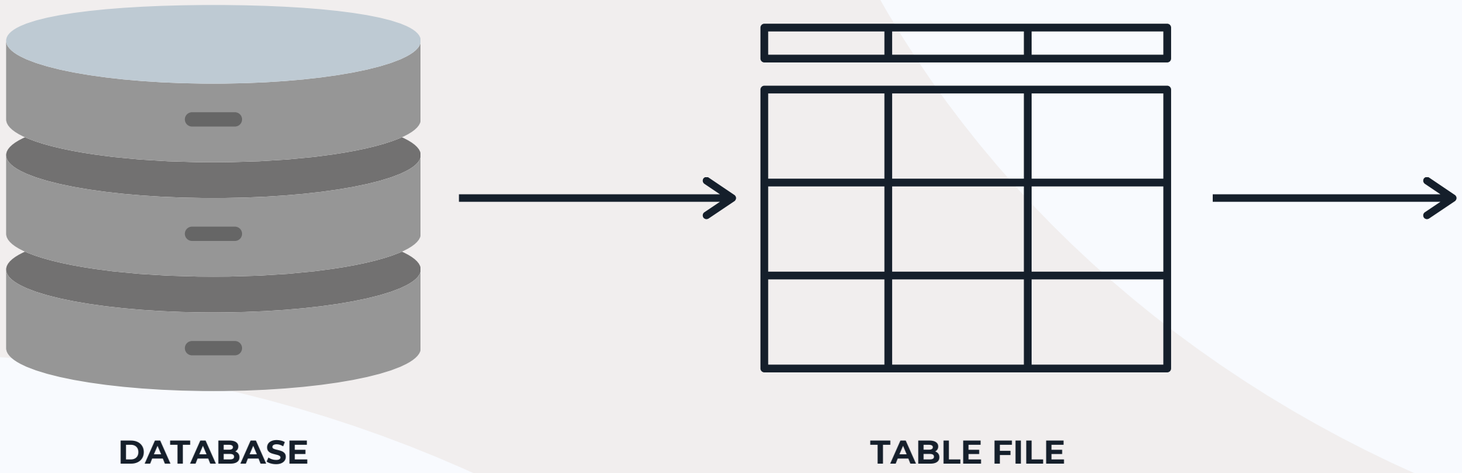
ADMIN

- NAME OF THE DATABASE
- MOST RECENT UPLOAD TIMESTAMP



- TOTAL NUMBER OF FILES
- TOTAL NUMBER OF FILES PROCESSED
- THE CURRENT PROGRESS

STAGE 1 - INFORMATION IN DATABASE

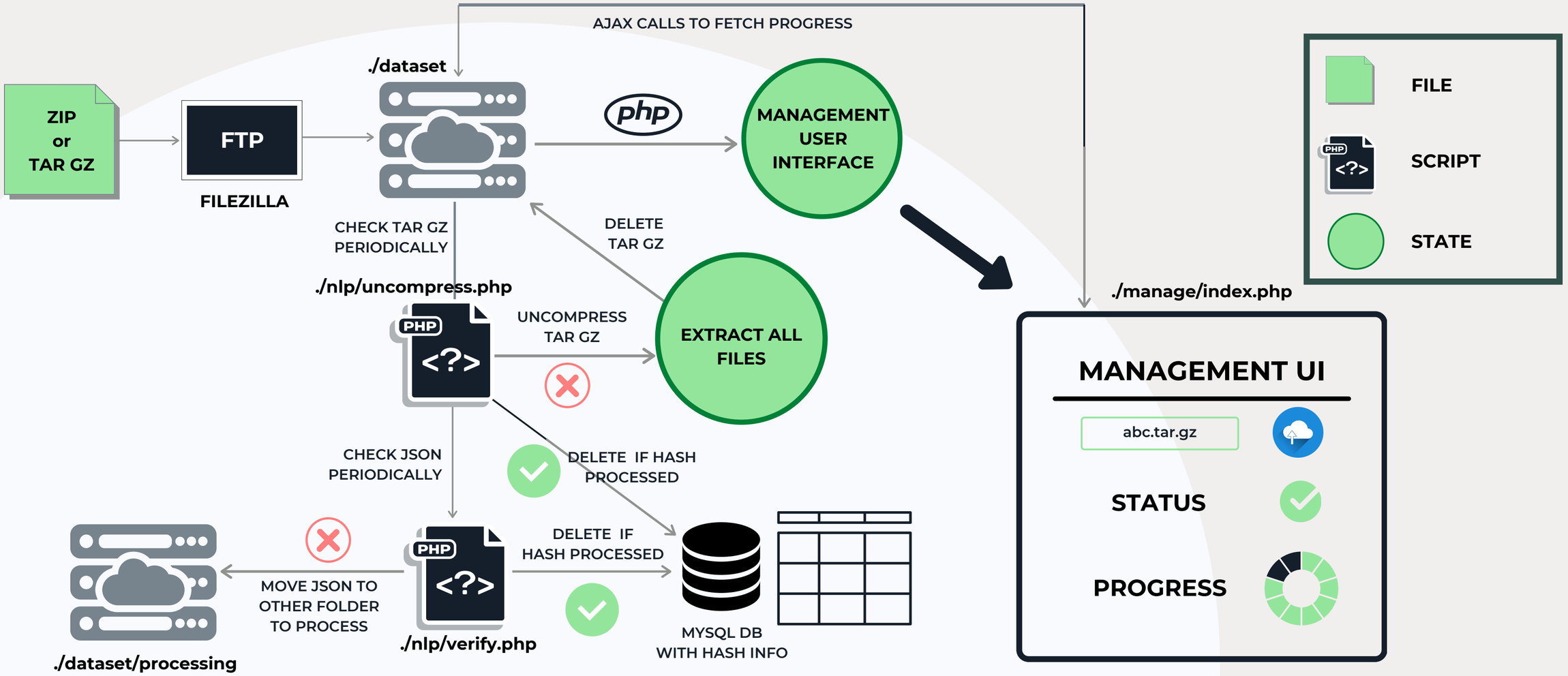


DATABASE

TABLE FILE

- File Hash - Ensures no file duplication
- Start Time - Starting of the extraction and verification process
- End time - End time of the extraction and verification process
- File size - Size of the compressed files
- File amount - Amount of files in each of the compressed files

STAGE 1 DETAILED - STRUCTURAL OUTLINE



DEMO TIME - STAGE 1

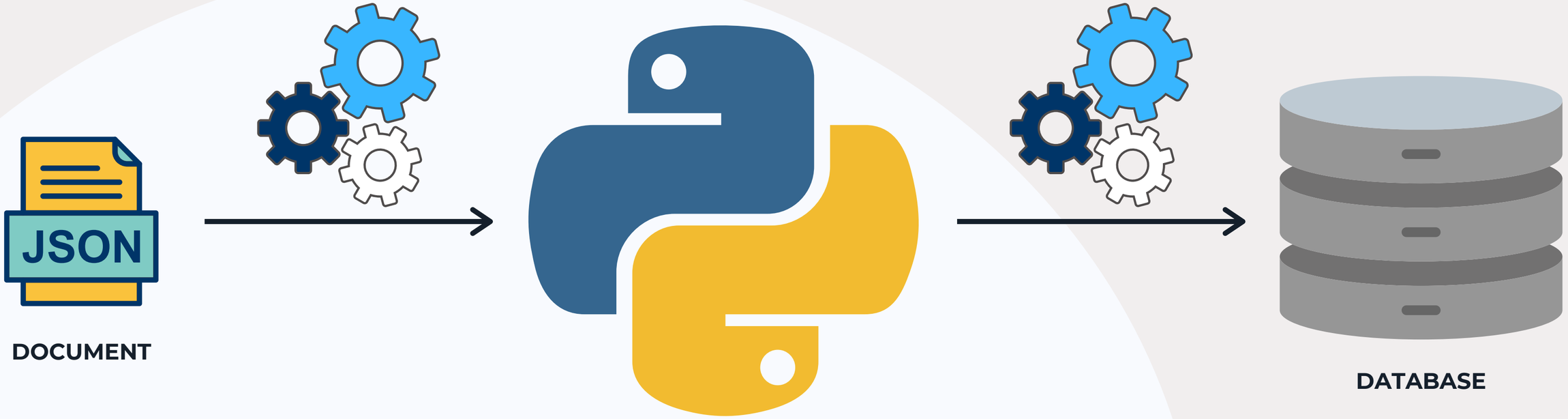
- LANDING PAGE
- MANAGEMENT DASHBOARD
- VIDEO - CRON DEMO

STAGE 2 - REQUIREMENTS

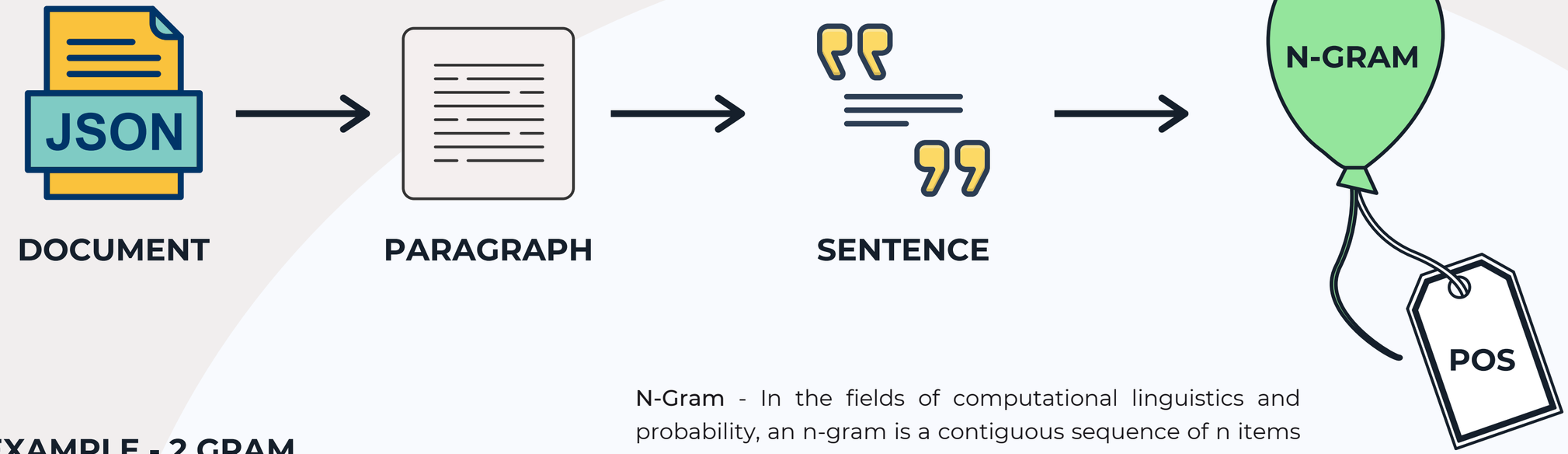
LANGUAGES AND FRAMEWORKS

- PYTHON 3
- NLTK - NATURAL LANGUAGE TOOLKIT 3.6.2
- MYSQL CONNECTOR PYTHON 8.0
- PHP 7.2.24

STAGE 2 OVERVIEW - DATA PROCESSING



STAGE 2 - UNDERSTANDING THE KEYWORDS



N-Gram - In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sample of text or speech.

EXAMPLE - 2 GRAM

Sentence - "I am a boy"

2-gram list : ['I am', 'am a', 'a boy']

POS list : ['PRP-VBP', 'VBP-DT', 'DT-NN']

Pos-Tagging - In corpus linguistics, part-of-speech tagging, also called grammatical tagging is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context.

STAGE 2 - INFORMATION IN DATABASE

1

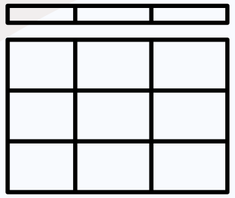


TABLE DOCUMENT

STORES INFORMATION ABOUT THE JSON DOCUMENT

Example attributes:

- SHA1 (UUID)
- Start timestamp
- End timestamp
- Paper details
- Author Details, etc

2

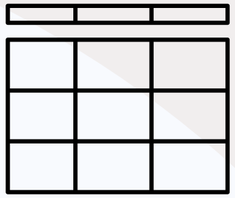


TABLE SENTENCE

STORES INFORMATION ABOUT THE SENTENCES IN THE DOCUMENT

Example attributes:

- SHA1 (UUID)
- Sentence ID
- Paragraph Number
- Section Number
- Sentence, etc

3

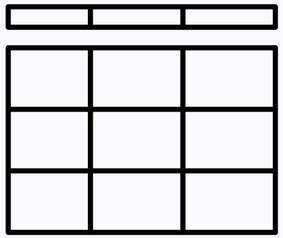


TABLE POS

STORES INFORMATION ABOUT THE ELIGIBLE POS TAGS

Attributes:

- POS ID
- POS TAG
- N-Gram Length

STAGE 2 - INFORMATION IN DATABASE

4

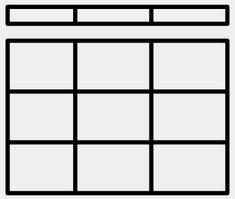


TABLE
NGRAM

STORES INFORMATION ABOUT THE N-GRAMS

Attributes:

- N-Gram ID
- N-gram
- POS ID

5

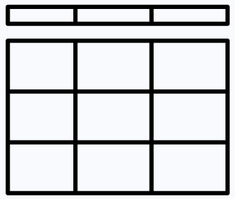


TABLE
NGRAM COUNT

STORES INFORMATION ABOUT THE COUNT OF N-GRAM IN A DOCUMENT

Attributes:

- SHA1 (UUID)
- N-Gram ID
- N-Gram Count

6

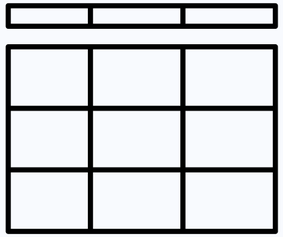


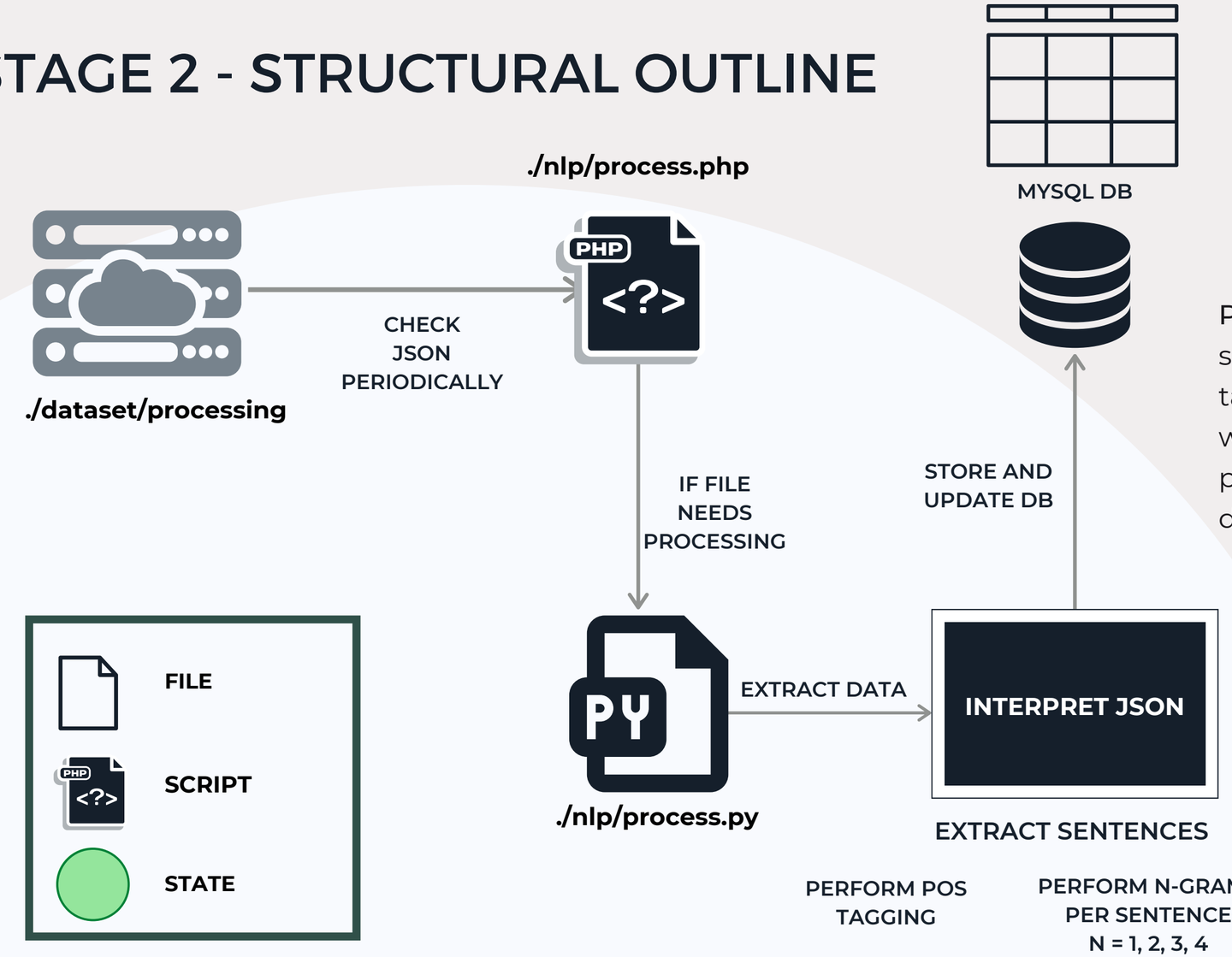
TABLE
NGRAM IN SENTENCE

STORES INFORMATION ABOUT THE N-GRAMS PER SENTENCE

Attributes:

- N-Gram ID
- Sentence ID
- Order Number

STAGE 2 - STRUCTURAL OUTLINE



Pos-Tagging - In corpus linguistics, part-of-speech tagging, also called grammatical tagging is the process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context.

DEMO TIME - STAGE 2

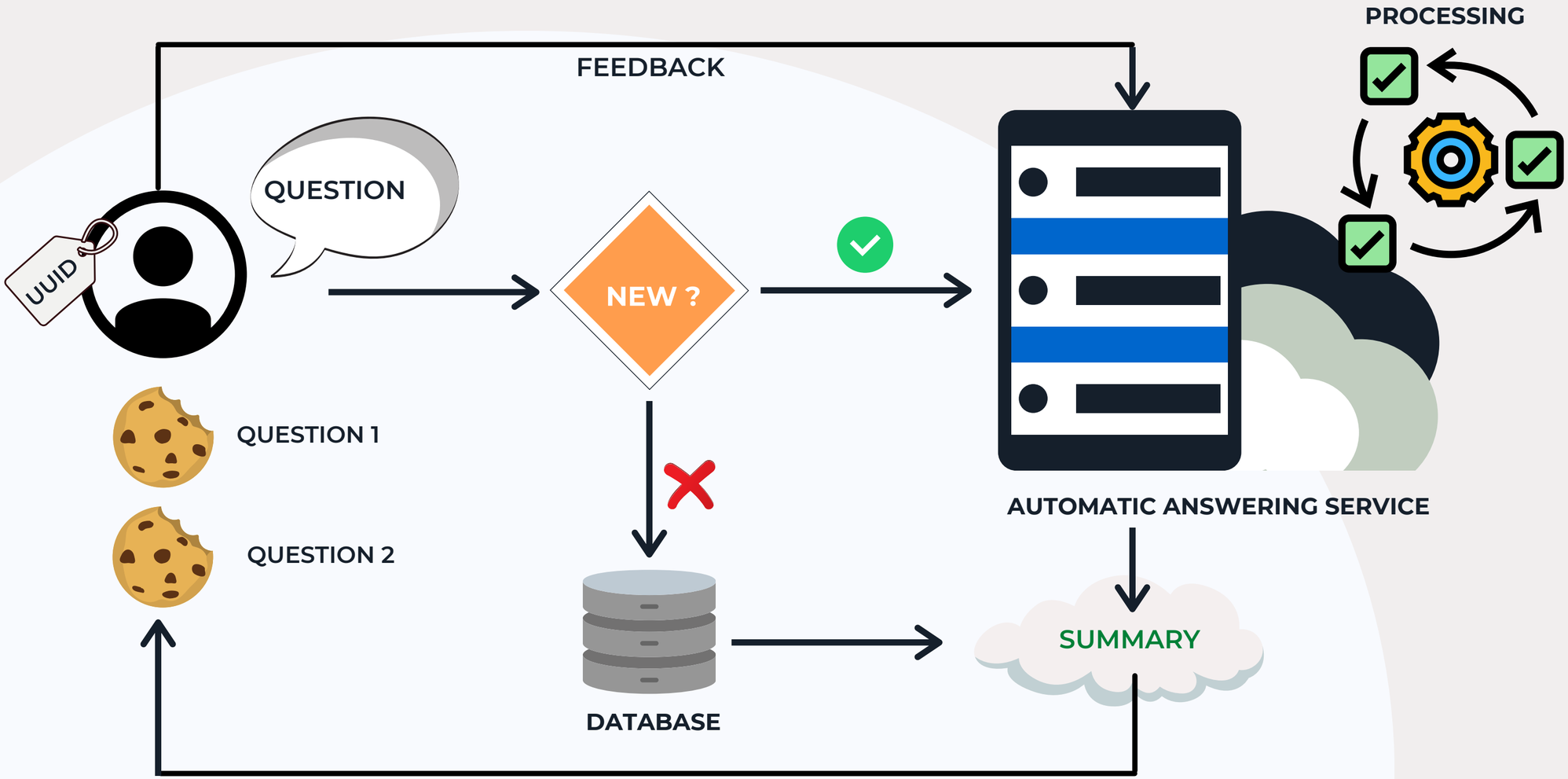
- **VIDEO - PROCESSING CRON DEMO**

STAGE 3 - REQUIREMENTS

LANGUAGES AND FRAMEWORKS

- Python 3
- NLTK - Natural Language Toolkit 3.6.2
- MySQL Connector Python 8.0
- HTML 5
- CSS 3
- JAVASCRIPT
- PHP 7.4
- BOOTSTRAP FRAMEWORK 4
- MySQL
- AJAX

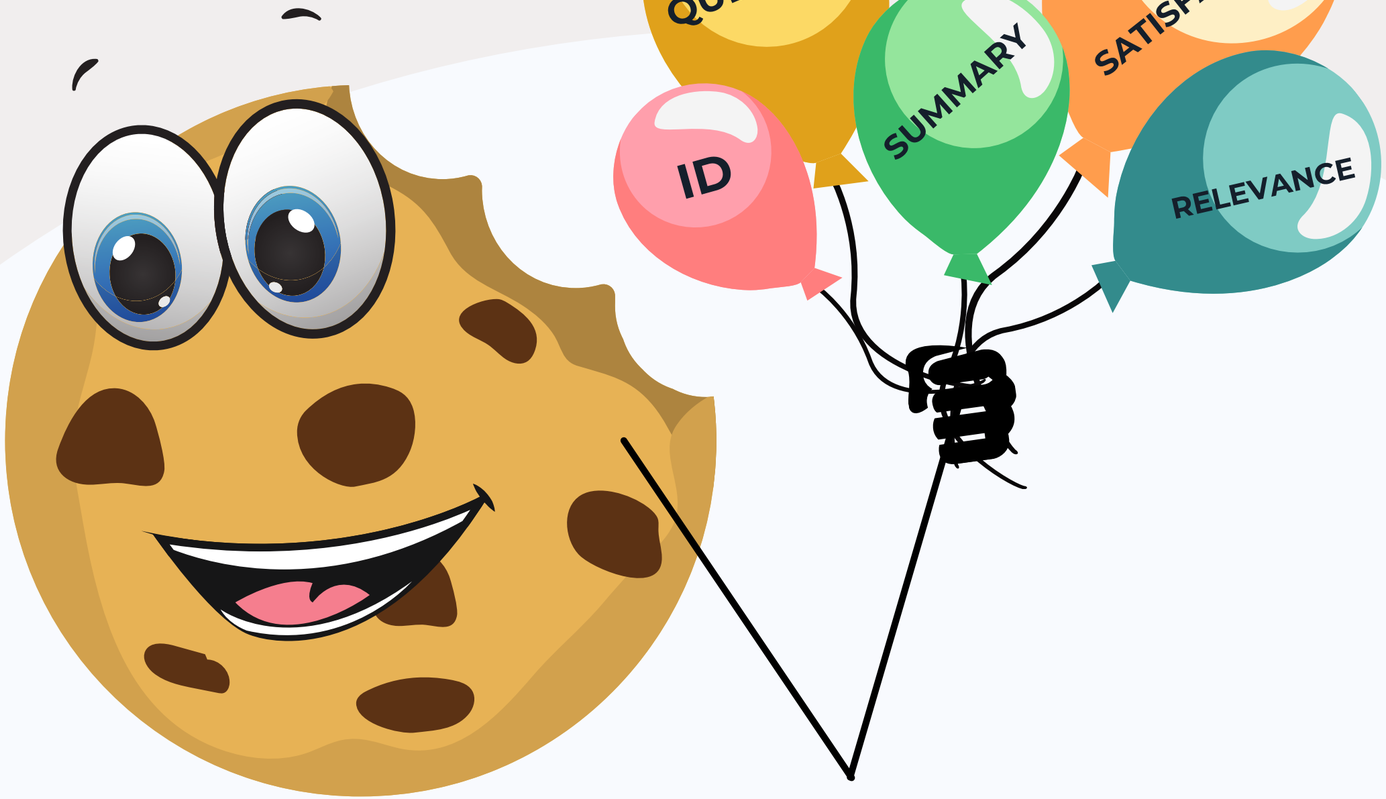
STAGE 3 OVERVIEW - SUMMARY GENERATION



STAGE 3 - USER COOKIES

QUESTION 1

QUESTION 2



QUESTION COOKIE

Each and every question asked is stored in a cookie so that users can take a look at them at any time. They can also update their feedback at any time.

STAGE 3 - INFORMATION IN DATABASE

1

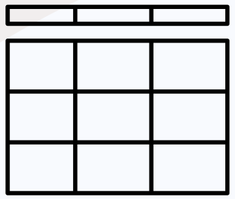


TABLE USER CLIENT

STORES THE UUID OF THE USER

Attribute:

- User UUID - Unique identification number of the user

2

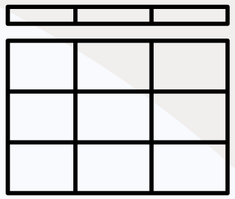


TABLE QUESTION

STORES INFORMATION ABOUT THE QUESTION

Attributes:

- Question ID
- Question
- Summary

3

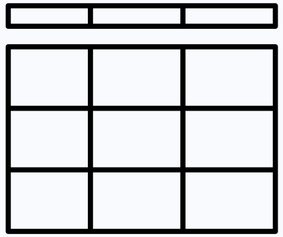


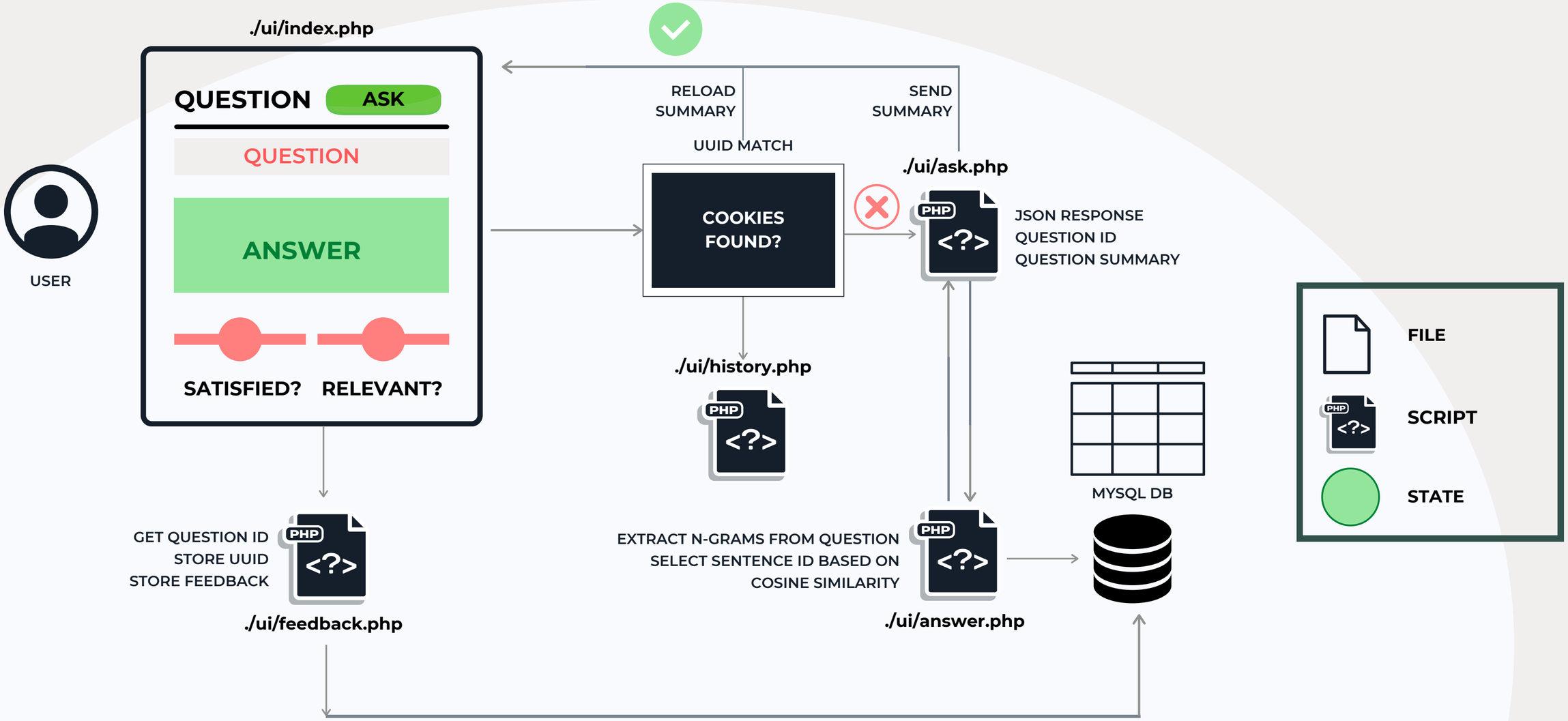
TABLE USER PERCEPTION

STORES INFORMATION ABOUT THE USER FEEDBACK

Attributes:

- Question ID
- User UUID
- Satisfaction
- Relevance

STAGE 3 - STRUCTURAL OUTLINE



DEMO TIME - STAGE 3

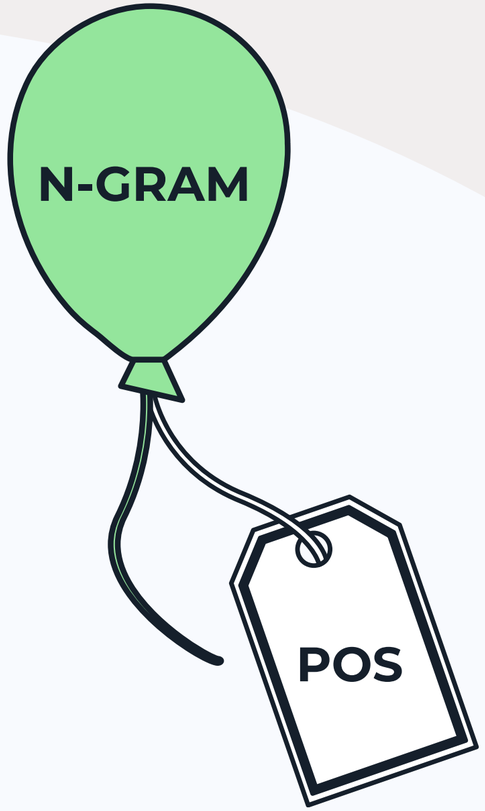
- USER INTERFACE
- LIVE SUMMARY GENERATION
- COOKIES USAGE

STAGE 3 - UNDERSTANDING HOW TO IDENTIFY KEYWORDS

Question: Is covid-19 deadly, will it ever vanish?

Eligible N-Gram: ['covid-19', 'deadly', 'covid-19 deadly', 'vanish']

Example Tags: NN ⟵ Noun ⟶ NN



POS TO BE CONSIDERED

- 1-GRAM: IN, NN, NNP, JJ, DT (5)
- 2-GRAM: NN-IN, JJ-NN, NNP-NNP, DT-NN (4)
- 3-GRAM: NNP-NNP-NNP, DT-JJ-NN, JJ-NN-IN, IN-DT-NN (4)
- 4-GRAM: NNP-NNP-NNP-NNP, DT-JJ-NN-IN, NN-IN-DT-NN (3)

ANALYSIS OF JSON DOCS - OBSERVATION

Why do we need to Analyze? Let's take a look at the POS tags

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunctio
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker

10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb

21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle

30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

N = 1 --> COMBINATIONS = 36

N = 2 --> COMBINATIONS = 36 X 36

N = 3 --> COMBINATIONS = 36 X 36 X 36

N = 4 --> COMBINATIONS = 36 X 36 X 36 X 36

Total rows in table_pos would become -> 1,727,604 where there would be many such POS patterns that would not be much useful.

WE COULD REDUCE 1.7 MILLION COMBINATIONS TO 16 POS TAGS, WE CAN ANALYZE A BIT MORE TO GET BETTER ACCURACY.

Link to Notebook: https://jovian.ai/sayantan-world98/json-analyse-v4

Link to the Report: https://jovian.ai/sayantan-world98/json-analyse-v4/v/2/files?filename=Report.pdf

ONGOING RESEARCH

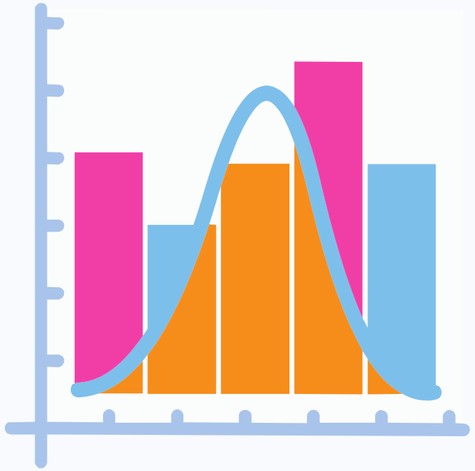
RESEARCH: VALID N-GRAM LEARNING AND VERIFICATION

It aims to extract and store N-grams and tag the POS (Part-of-Speech) from a huge dataset (DBpedia). It uses these stored N-Grams and their corresponding POS to create a service, that identifies valid N-Grams from any given source.

Research by Mr. Bhavesh Gandhi (End of August 2021)

Using statistical methods, to identify valid POS tags for the N-Grams extracted from CORD-19 datasets.

Research by Mr. Dan Boonstra



STAGE 3 - UNDERSTANDING SUMMARY GENERATION

Question: Is covid-19 deadly, will it ever vanish?

Eligible N-Gram: ['covid-19', 'deadly', 'covid-19 deadly', 'vanish']

Target Vector: ['85', '58', '20', '25'] ⟶ Frequency of the eligible N-Grams in the database

DOCUMENT VECTORS 

DOCUMENT - 1: ['35', '25', '10', '12'] ✓

DOCUMENT - 2: ['5', '10', '10', '1'] ✗

Frequency of the eligible N-Grams in the document

SENTENCE VECTORS 

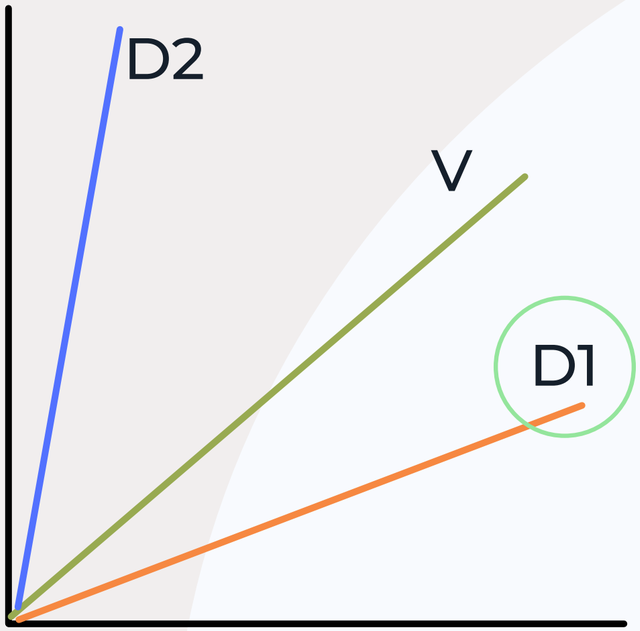
SENTENCE - 1: ['20', '15', '5', '6'] ✓

SENTENCE - 2: ['10', '8', '2', '3] ✓

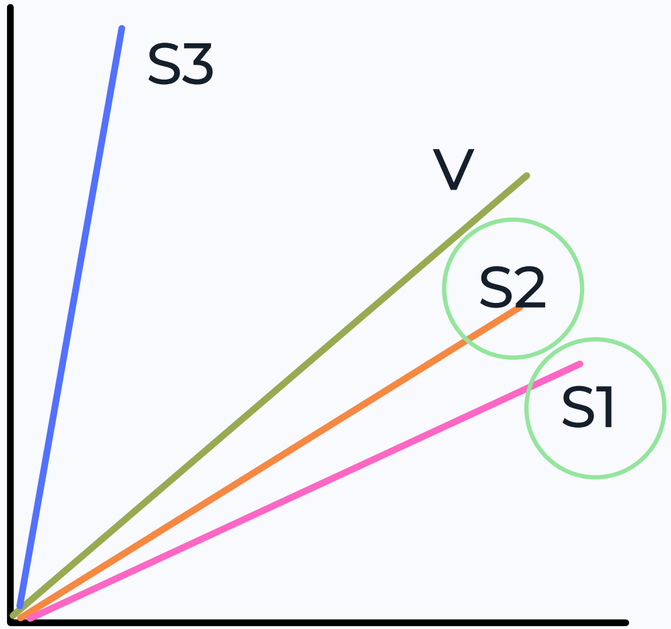
SENTENCE - 3: ['5', '1', '3', '1'] ✗

Frequency of the eligible N-Grams in the sentence

STAGE 3 - UNDERSTANDING COSINE SIMILARITY



1 DOCUMENT VECTORS



2 SENTENCE VECTORS

V - TARGET VECTOR

D - DOCUMENT VECTOR

S - SENTENCE VECTOR

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space. It is defined to equal the cosine of the angle between them, which is also the same as the inner product of the same vectors normalized to both have length 1.

FUTURE RESEARCH

- USING A SERVICE (ONGOING RESEARCH) TO GET A LIST OF ELIGIBLE N-GRAM POS TAGS
- REDUCING THE AMOUNT OF TIME TO GENERATE THE SUMMARY
- GENERATING A BETTER SUMMARY

Questions?
Comments?

Let us know!

THANK YOU