

Web Phishing Project (ENG version)

Database presentation: the database to be used contains URL data from web pages to determine which fields of a URL suggest that the URL may be phishing. The goal of the project is to optimize the performance of the database by cleaning the data, creating a new fact and dimension model, export a view for later use in creating visualizations in Power BI.

The final report should answer the business questions provided by the client. The client is only interested in analyzing the first 1000 URLs containing more than 150 characters, so the database must be filtered to meet these conditions. Preferably a view will be developed to be exported to Power BI.

The project will consist of:

- Data cleansing.
- Data exploration.
- Data modeling.
- Creation of visualization with descriptive analysis.
- Final dashboard and step-by-step explanatory document.

Project development:

1- Import table to PostgreSQL through pgadmin 8.4v:

Create table with each of the columns and their data types, then use pgadmin's Import/Export Wizard tool to import the data, and check that everything is working correctly.

Table name: webphisingraw.

Columns:

1. url_length: the length of the URL.
2. n_dots: The count of '.' characters in the URL.
3. n_hyphens: The count of '-' characters in the URL.
4. n_underline: The '_' character count in the URL.
5. n_slash: The count of '/' characters in the URL.
6. n_questionmark: The '?' character count in the URL.
7. n_equal: The '=' character count in the URL.
8. n_at: The '@' character count in the URL.
9. n_and: The '&' character count in the URL.
10. n_exclamation: The '!' character count in the URL.
11. n_space: The ' ' character count in the URL.
12. n_tilde: The count of '~' characters in the URL.
13. n_comma: The ',' character count in the URL.
14. n_plus: The '+' character count in the URL.
15. n_asterisk: The '*' character count in the URL.
16. n_hastag: The '#' character count in the URL.
17. n_dollar: The '\$' character count in the URL.
18. n_percent: The '%' character count in the URL.
19. n_redirection: The redirection count in the URL.
20. phishing: The labels in the URL. 1 is phishing and 0 is legitimate.

SELECT * FROM webphisingraw;

2- Normalize data model (create a fact table and a dimension table):

Split and reorganize the webphisingraw table into two new tables, in order to optimize the data exploration experience, and make a normalized model for use.

Name of the fact table: web_page_phising. Columns:

- 1- unique_id (SERIAL, PK),
- 2- url_length (INT),
- 3- n_redirection (INT),
- 4- phising (INT).

Dimension table name: phishing_dataset. Columns:

- 1. unique_id, (FK, INT).
- 2. n_dots (INT)
- 3. n_hyphens (INT)
- 4. n_underline (INT)
- 5. n_slash (INT)
- 6. n_questionmark (INT)
- 7. n_equal (INT)
- 8. n_at (INT)
- 9. n_and (INT)
- 10. n_exclamation (INT)
- 11. n_space (INT)

12. n_tilde (INT)

13. n_comma (INT)

14. n_plus (INT)

15. n_asterisk (INT)

16. n_hashtag (INT)

17. n_dollar (INT)

18. n_percent (INT)

3- Connecting PostgreSQL database to Microsoft Power BI

Open Microsoft Power BI, and connect to the database through Start/Data/Get Data/PostgreSQL database.

4- Business questions

● Which field(s) has the strongest correlation with the "phishing" field? Which field(s) has the weakest correlation with the "phishing" field?

1. Strongest correlation: url_length, n_slash, n_equal.

2. Weakest correlation: n_redirection, n_space, n_plus.

● **Would you say that URL length is a strong indicator of whether the URL is phishing? Why yes or why not? What metrics do you have to support your answer?**

Yes. The longer the number of characters, the more frequent phishing is. Conversely, when the length is shorter, there tend to be fewer instances of phishing. This is tested by running the following PostgreSQL query:

SELECT phishing, CASE

WHEN url_length BETWEEN 0 AND 299 THEN '1-299' WHEN url_length
BETWEEN 0 AND 299 THEN '1-299'.

WHEN url_length BETWEEN 300 AND 599 THEN '300-599' WHEN url_length
BETWEEN 300 AND 599 THEN '300-599

WHEN url_length BETWEEN 600 AND 899 THEN '600-899' WHEN url_length
BETWEEN 600 AND 899 THEN '600-899

WHEN url_length BETWEEN 900 AND 1199 THEN '900-1199'

ELSE '1200+'









END AS range_group, COUNT(phishing)

FROM web_page_phishing

GROUP BY phishing, range_group

ORDER BY range_group ASC;

The purpose of this SELECT is to group the data into ranges, and with these ranges count the number of times a URL is legitimate (0) or phishing (1). The results obtained reflect: the longer the URL, the more common it is to find phishing.

Data Output		Messages	Notifications
			
			
	phishing integer	range_group text	count bigint
1	0	1-299	63708
2	1	1-299	36043
3	1	1200+	8
4	0	300-599	7
5	1	300-599	270
6	1	600-899	33
7	1	900-1199	8

● **Would you say that the number of redirects is a solid indicator of whether the URL is phishing? Why yes or why not? What metrics do you have to support your answer?**

The number of redirects is **not a strong** indicator of whether the URL is phishing. There is no pattern in the data that indicates that the higher the number of redirects, the higher the frequency of phishing. In fact, the data yields similar numbers when the data is grouped into Legit (0) and Phishing (1), even having a slight difference in favor of the Legit group. This reflects that, although there are many redirects, it is not directly related to Phishing.

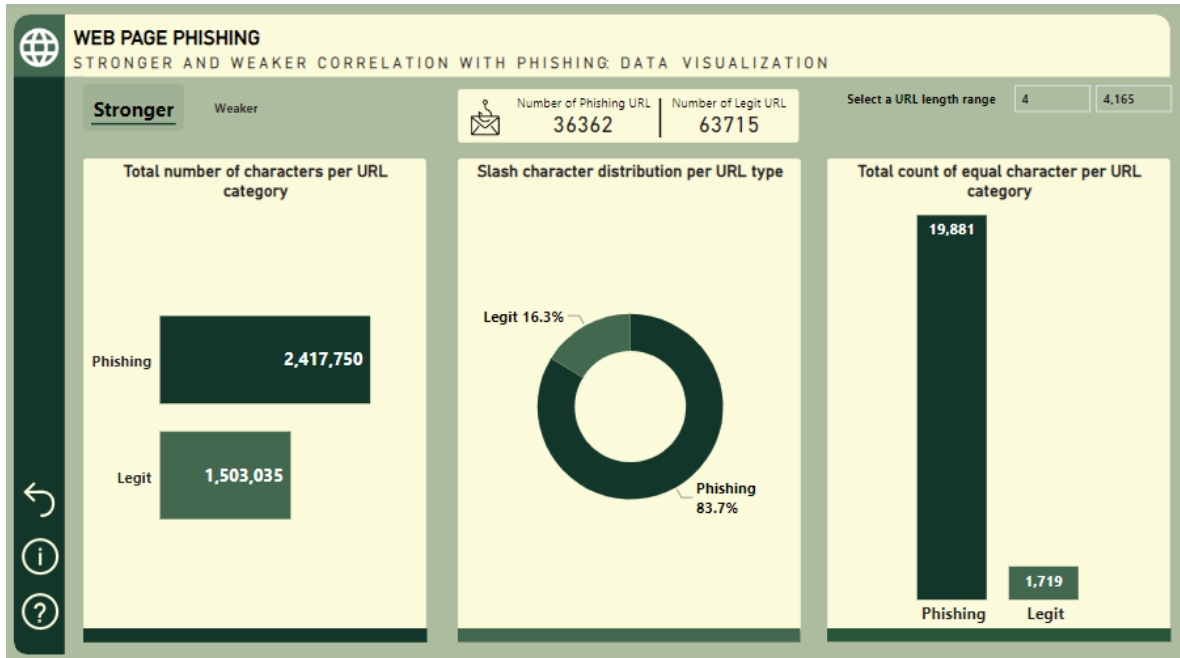
```
SELECT  
  
CASE WHEN phishing = 0 THEN 'Legit' ELSE 'Phishing' END AS phishing,  
  
n_redirection, COUNT(n_redirection) AS redirection_phishing_count  
  
FROM web_page_phishing  
  
GROUP BY phishing, n_redirection  
  
HAVING COUNT(n_redirection) >= 100  
  
ORDER BY n_redirection DESC, redirection_phishing_count DESC;
```

● **Based on your analysis, what advice would you give to others to decipher whether a URL is phishing?**

Any URL with an exaggerated number of characters will in the vast majority of cases be phishing, so it is always advisable to start the analysis of the data from the length of the URL. It is important to group the data into ranges to facilitate the primary analysis when exploring the data, and gradually increase the granularity of detail to deepen the analysis of the fields, especially when the type of character appears rarely (such as at or asterisk).

5- Create visualizations that answer the questions posed by the client (Microsoft Power BI)

Stronger correlations



Weaker correlations

