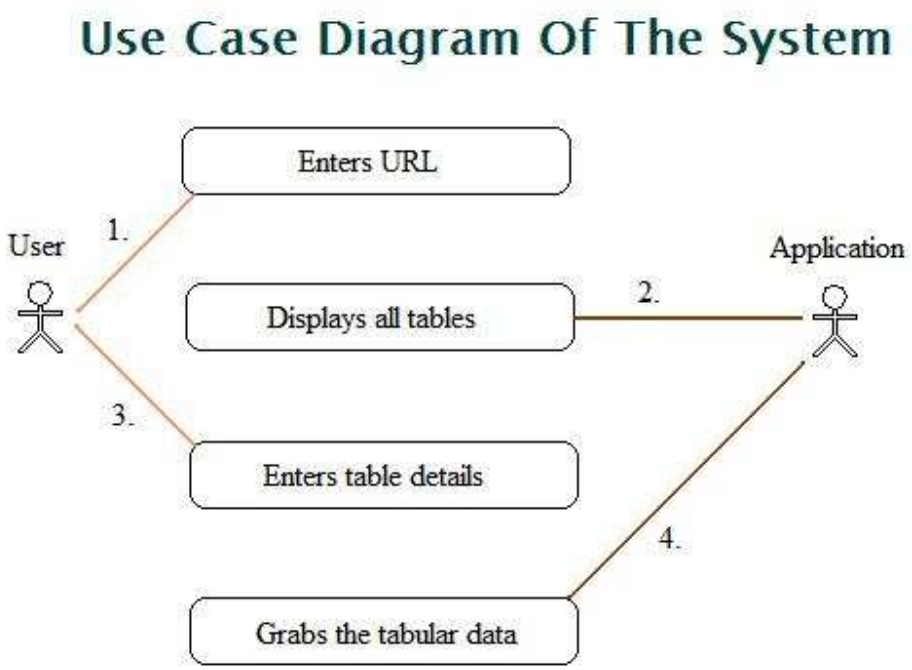
Overview

The diagram shown below demonstrates the simple model of the application;



As depicted in the above diagram user first enters the URL whose table needs to be extracted. Then, the application displays all tables present in the page. By this point, user needs to provide the table number as per the requirement and the header pattern as well. The tabular data is extracted according to the specification. By telling tabular data, we here mean the data hidden inside <table> tags in the website.

That was the non-technical aspect of our application. Now moving to the technical detail; pattern array is the pivot of the whole schema. Using this, the job of extraction is carried out. This array suggests whether or not the header & column names are present and their corresponding depth [in fact first set and last set of Row matrix is unnecessary].

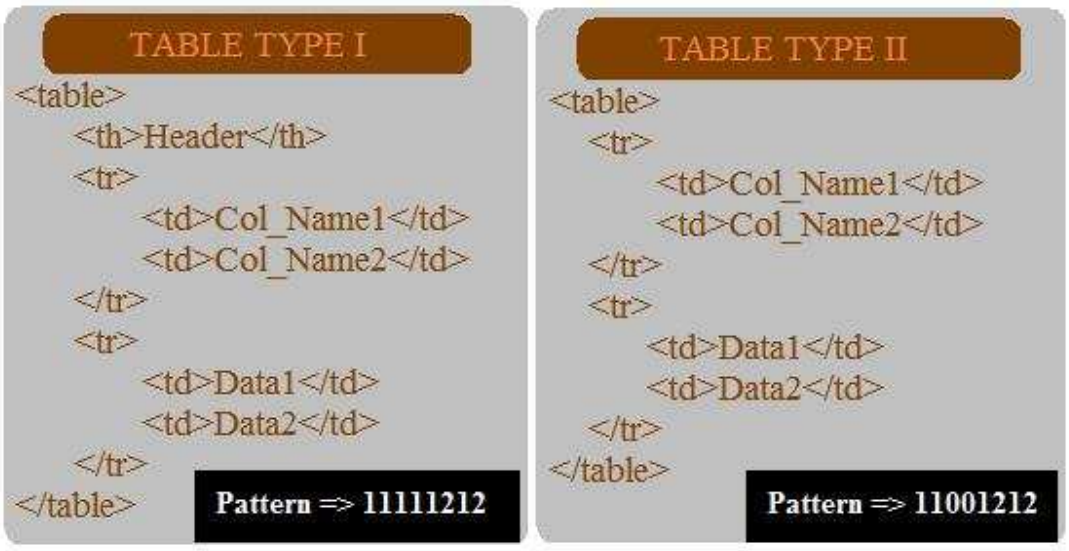
It can be depicted as a 4*4 matrix as shown below;

Table Elements	Flag [1 or 0]	Depth [1 or 2 or 3]
<table> tag	-	-
Header Name	-	-
Column Name	-	-
Data	-	-

But for our convenience we have converted that into Row Matrix as shown below.

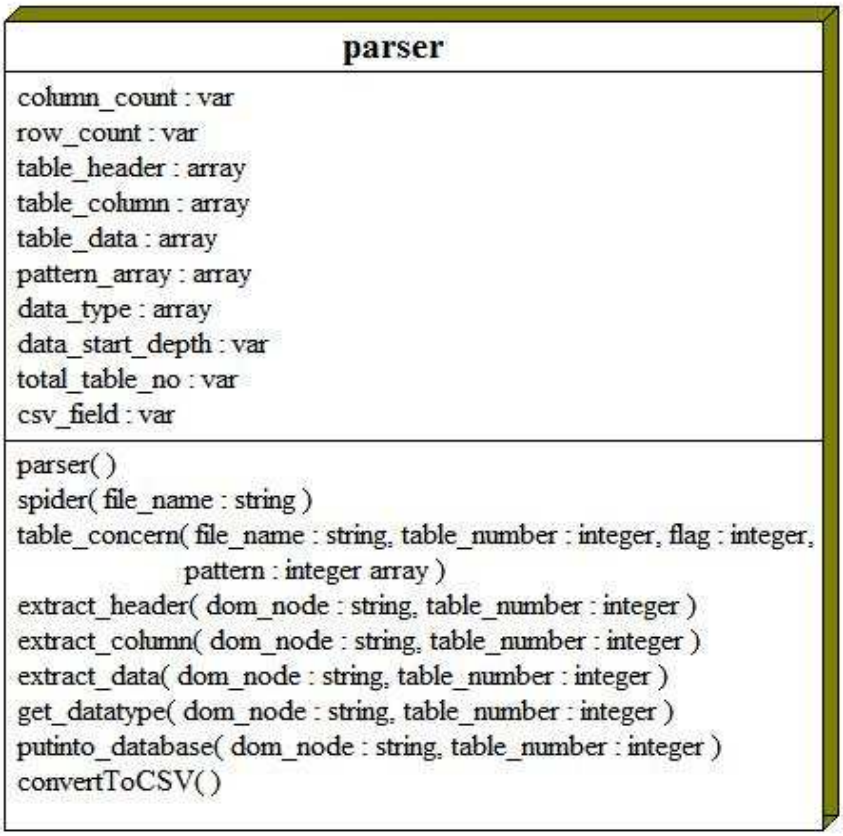
[{table info}{header info}{column info}{data info}]

Depth is basically to what extent the DOM has to penetrate in order to extract the data. The figure presented below should make this idea even more clear.



PHP being loosely coupled with OOP, it was really difficult to produce an accurate UML diagram. Entire Logic behind extracting data resides in the parser class whose UML diagram should clarify the superficial details.

UML Diagram of 'parser' Class



Prerequisite

- A database named 'parsedData' should be created beforehand
- And Definitely a good internet connection is also customary too ...

Limitations

- Cannot extract data from tables with non uniform structures
- Only extracts data from tables with column names & data of 2 depth
- Sometimes problem with spaces and indentation inside table structure [couldn't figured but may be the problem with DOM][check this by selecting the table number 6 of http://nepalstock.com/datanepse, with proper indentation it worked well while testing]
- Error Handlings are not dealt with that sort of care
- CSV conversion is not correct always as it should be

P.S. Consider editing the showTables.php page for proxy authentication.