Project Name:1. **Statistical Entity Extraction from Web**

**Contents:1. Limitation of Previous Work**

When we need big data from from some other platforms. Then we we have to go on that websites and manually get the data and then we will be able to put that data in our application and our application is then work on that data. This is the behavior of old applications. In the new age of developing technology there is need of something that is going to get respective data from third party web applications and store that data in our application and utilize that data when ever we need that data.

The process of manually grabbing data from the other websites and adding it to our application is too much time consuming process and also Lengthy process. While doing this copy and paste there is too much chances of getting something wrong while copy paste.

**2. Proposed work**

We will create a bot application which is going to get all the HTML which we can see on browser and harvest that HTML and get particular HTML content which is useful for our application and output of that harvesting is useful for our application. We can use this harvested data to store in our database and hence useful in future for our application. There are many web based applications where we can see large amount of data and we can only see that large amount of data on browser they don't allow us to download that data. Our application is actually responsible to grab all the HTML and harvest that HTML to grab the data and store that data in our database.

We have created an application where users can get register with the application and can get login into the user panel. In that application after getting logged in users can search baseball players which are listed on <https://www.baseball-reference.com/> . After searching by some name user can see multiple players with similar name. Then user can select any one of it. After selecting that player user will be able to see the statistics of that particular player.  
  
Statistics include their all complete record of playing of that particular playe.

**3. Project Scope**

Registration : registration contains following fields. Anyone can get register with the application and start using the application to check get the record of players.

1. first name
2. last name
3. email
4. phone
5. address
6. username
7. password
8. confirm password

Login: Login of application is normal one. The registered user can get login with the application.

Dashboard : Dashboard of the application shows all the things which have used in application shortly. It is showing the history of sharing the links.

It is also showing your latest history of the search.

Search Page : your search page is a bit complex page. Here user will have to enter name first. Application will find similar type of name on the website and then show the list of same types of users. Users will have to select any user from it and then the application is again fetching the complete data from that website and showing to the page. Also the data is getting store in to the database. We also have provided option to share this search with any of other users who have registered with this application.

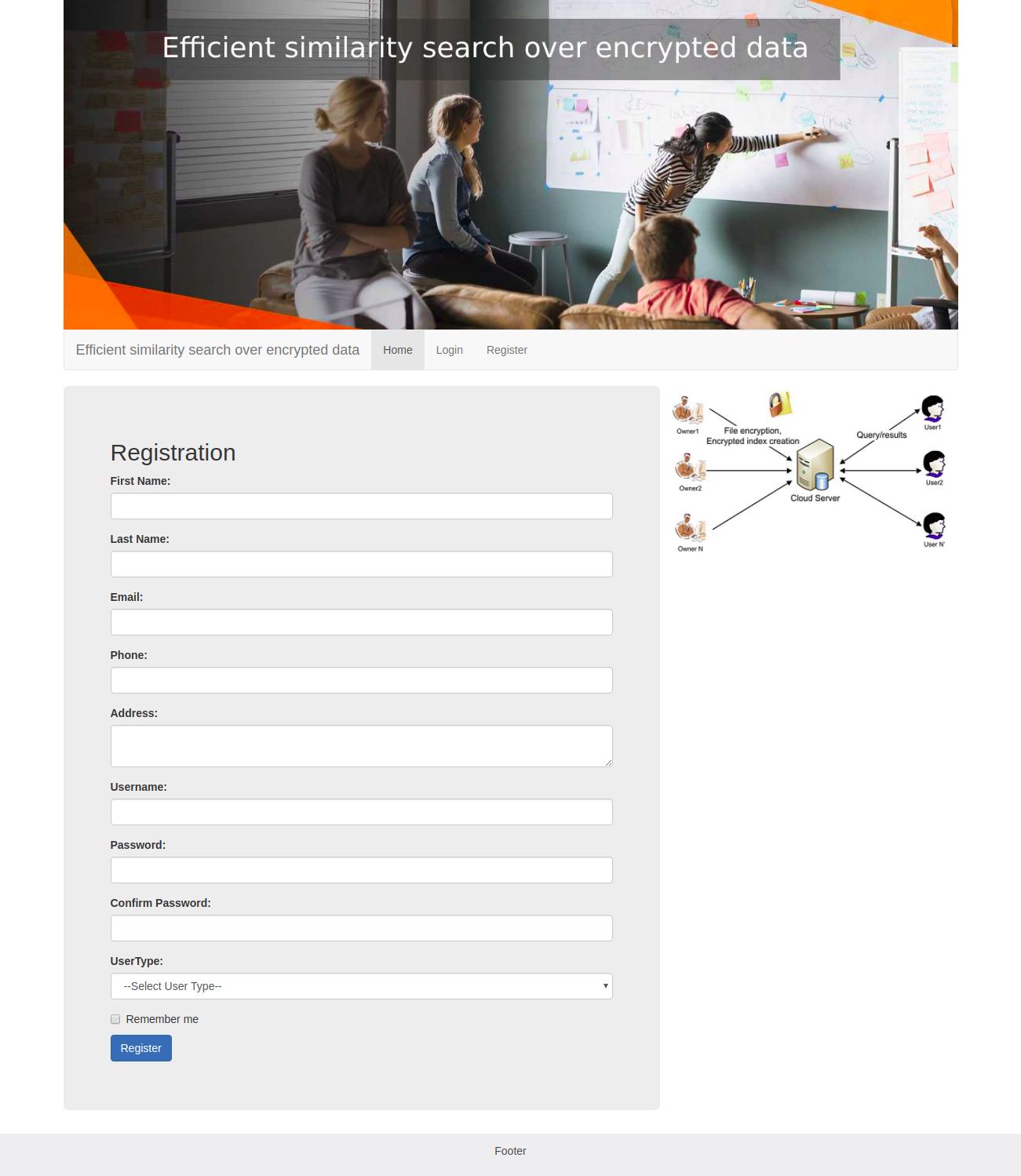
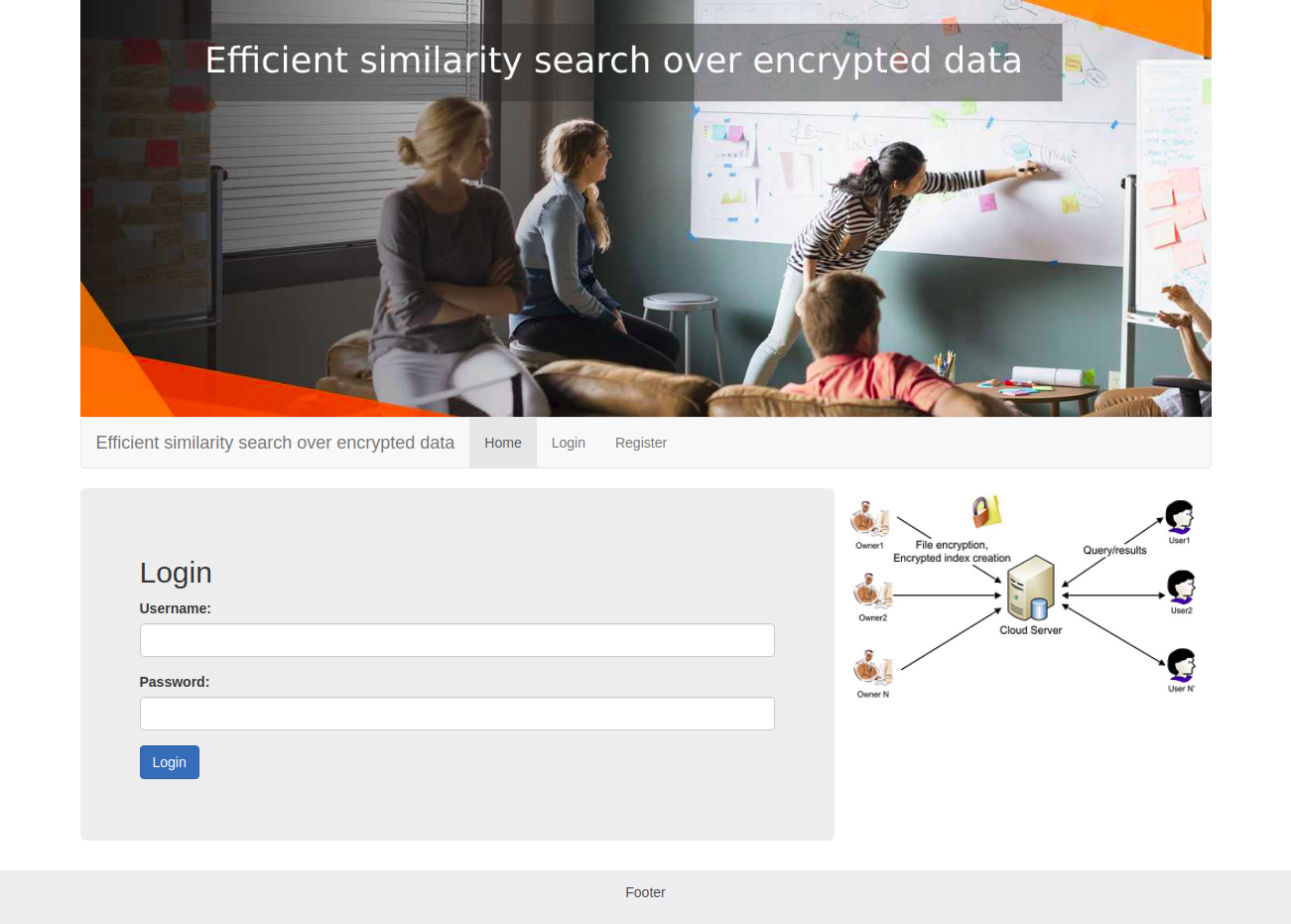
**4.Advantage**

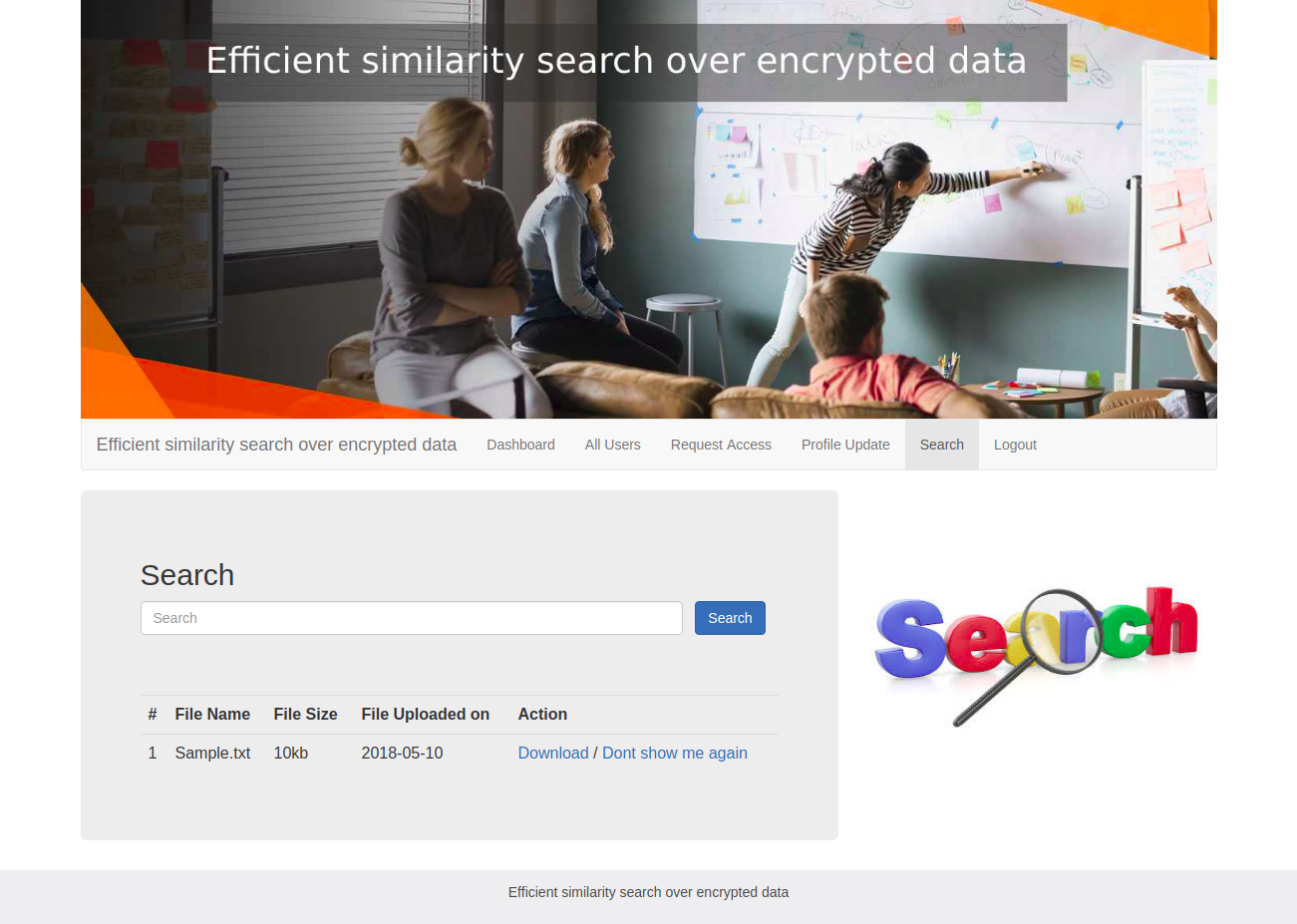
The Web harvesting is actually grabbing the data which is getting displayed on the web browser and that data then can be converted in string format and hence we can grab particular contents from it. For example Just dial is providing the list of vendors in the city. We can hit the link using internal software and can grab all the details which is getting displayed on the particular page when we will hit the link manually.

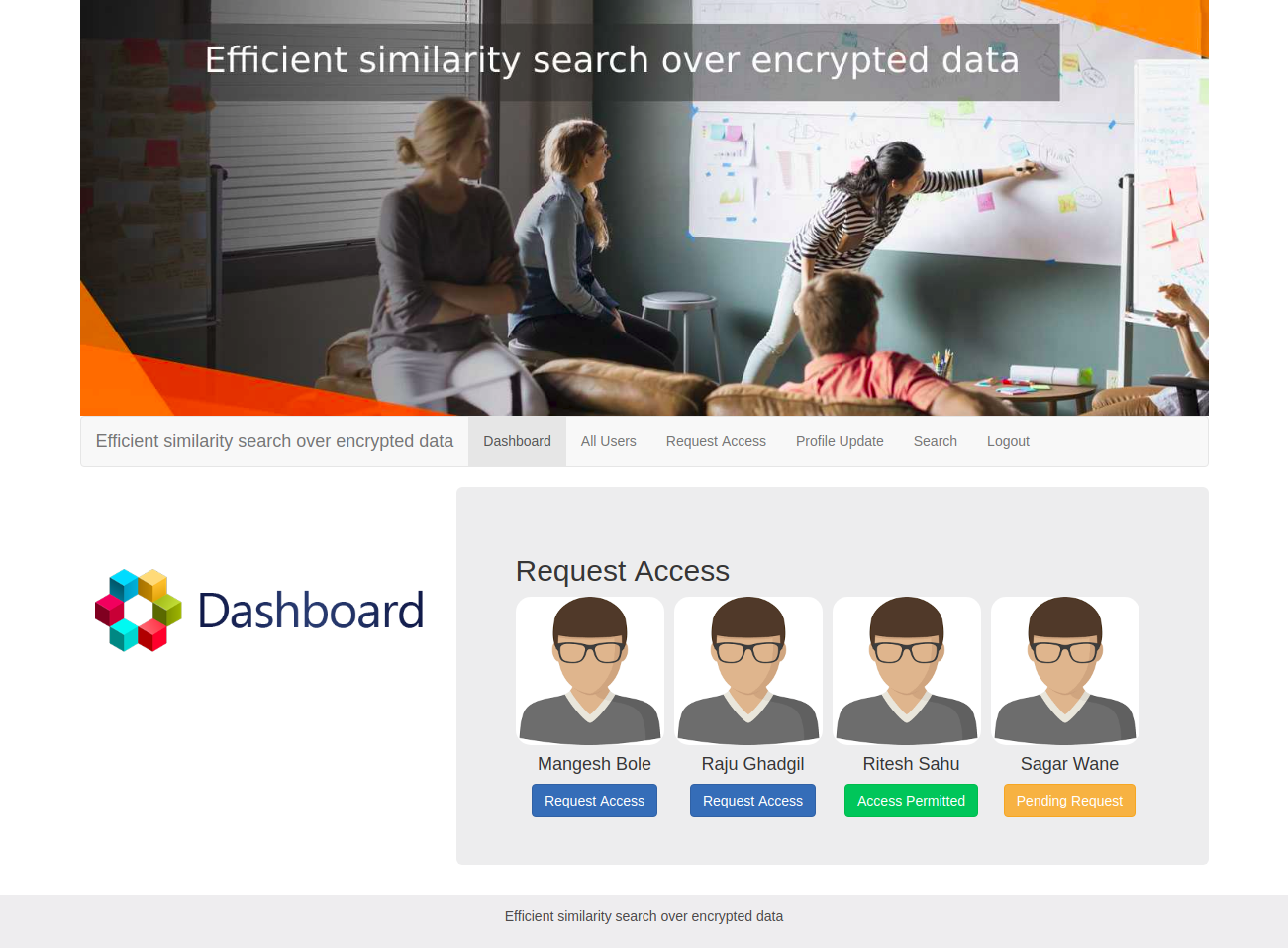
Methodology will be used in the project will be actually to just grab the web content in the web link and grab the particular content from that string and insert those into the expected database columns.

Useful to have accuracy and large amount of data within short period of time with us.

**5. Snapshot Output**





****

**6.Coding**

<!DOCTYPE html>

<html lang="en">

<head>

<title>Statistical entity extraction from web</title>

<meta charset="utf-8">

<meta name="viewport" content="width=device-width, initial-scale=1">

<link rel="stylesheet" href="css/bootstrap.min.css">

<link rel="stylesheet" href="css/style.css">

</head>

<body>

<div class="jumbotron text-center" style="margin-bottom:0">

<h1>Statistical entity extraction from web</h1>

</div>

<nav class="navbar navbar-expand-sm bg-dark navbar-dark">

<a class="navbar-brand" href="#">Navbar</a>

<button class="navbar-toggler" type="button" data-toggle="collapse" data-target="#collapsibleNavbar">

<span class="navbar-toggler-icon"></span>

</button>

<div class="collapse navbar-collapse" id="collapsibleNavbar">

<ul class="navbar-nav">

<li class="nav-item">

<a class="nav-link" href="index.html">Home</a>

</li>

<li class="nav-item">

<a class="nav-link" href="#">About Project</a>

</li>

<li class="nav-item">

<a class="nav-link" href="#">Abstract</a>

</li>

<li class="nav-item">

<a class="nav-link" href="register.html">Register</a>

</li>

<li class="nav-item">

<a class="nav-link" href="login.html">Login</a>

</li>

</ul>

</div>

</nav>

<div class="container" style="margin-top:30px">

<div class="row">

<div class="col-sm-4">

<h2>Abstract</h2>

<h5></h5>

<p>Web harvesting is important sometimes, when we need to grab too much data from some website. Web scraping a web page

involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when you

view the page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing.

Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data

copied into a spreadsheet, and so on. Web scrapers typically take something out of a page, to make use of it for

another purpose somewhere else. An example would be to find and copy names and phone numbers, or companies and

their URLs, to a list (contact scraping). In this project we are going to take one website having some dynamic

data. Like user data or some usefull information. We will write code in such a way that it will grab data from

that website and put the data in our database. This will be helpful to have data in our database.</p>

<h3>Web Extraction Done from </h3>

<p>Some Links.</p>

<ul class="nav nav-pills flex-column">

<li class="nav-item">

<a class="nav-link" target="\_blank" href="https://www.baseball-reference.com/">https://www.baseball-reference.com/</a>

</li>

<li class="nav-item">

<a class="nav-link" target="\_blank" href="https://www.baseballamerica.com/">https://www.baseballamerica.com/</a>

</li>

</ul>

<hr class="d-sm-none">

</div>

<div class="col-sm-8">

<h2>Flow Chart</h2>

<h5>Flow chart to do development of the concept</h5>

<div class="fakeimg">

<img src="images/flowchart.png" class="second-image">

</div>

<p>Some text..</p>

<p>Sunt in culpa qui officia deserunt mollit anim id est laborum consectetur adipiscing elit, sed do eiusmod tempor

incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco.</p>

<br>

<h2>Extraction From </h2>

<h5>Extracted data from </h5>

<div class="fakeimg">

<img src="images/banner2.jpeg" class="second-image">

</div>

<p>Some text..</p>

<p>Sunt in culpa qui officia deserunt mollit anim id est laborum consectetur adipiscing elit, sed do eiusmod tempor

incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco.</p>

</div>

</div>

</div>

<div class="jumbotron text-center" style="margin-bottom:0">

<p>Statistical entity extraction from web @ 2018</p>

</div>

</body>

</html>