PUC Minas Virtual

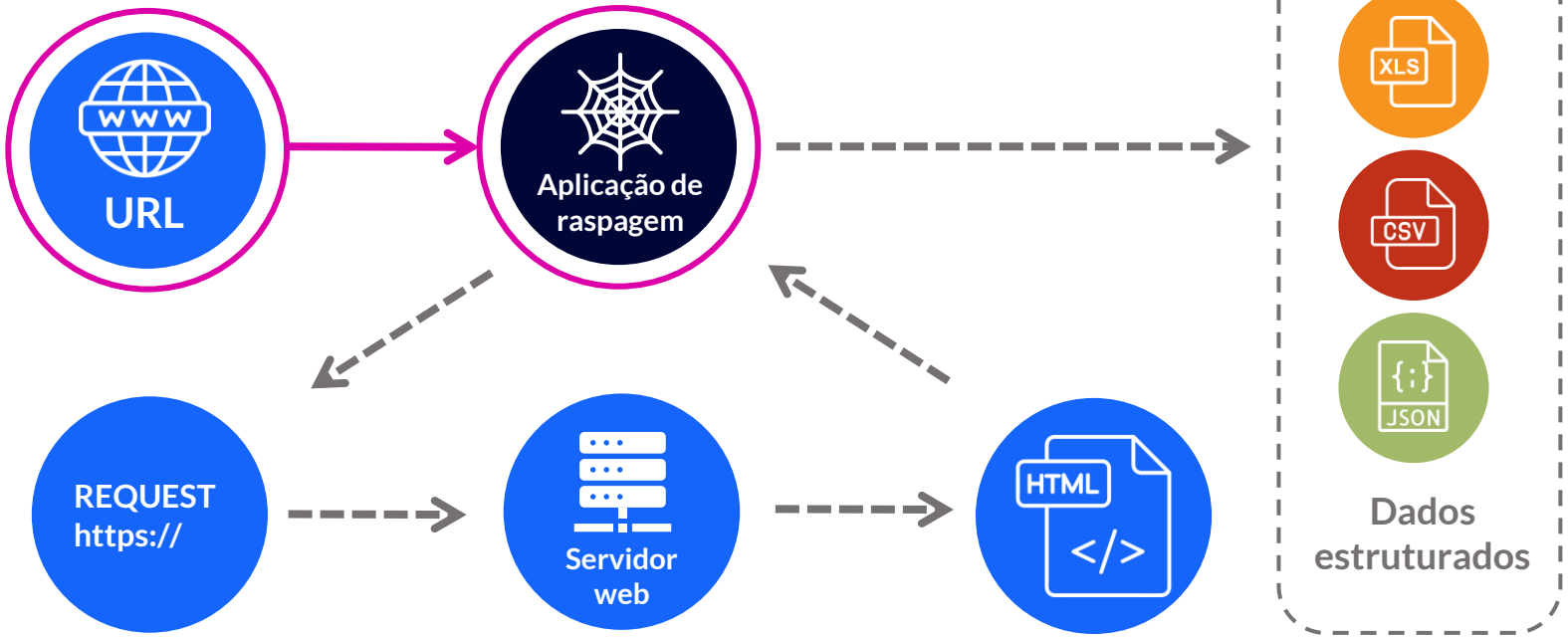
PROCESSAMENTO DE LINGUAGEM NATURAL

RASPAGEM DE DADOS

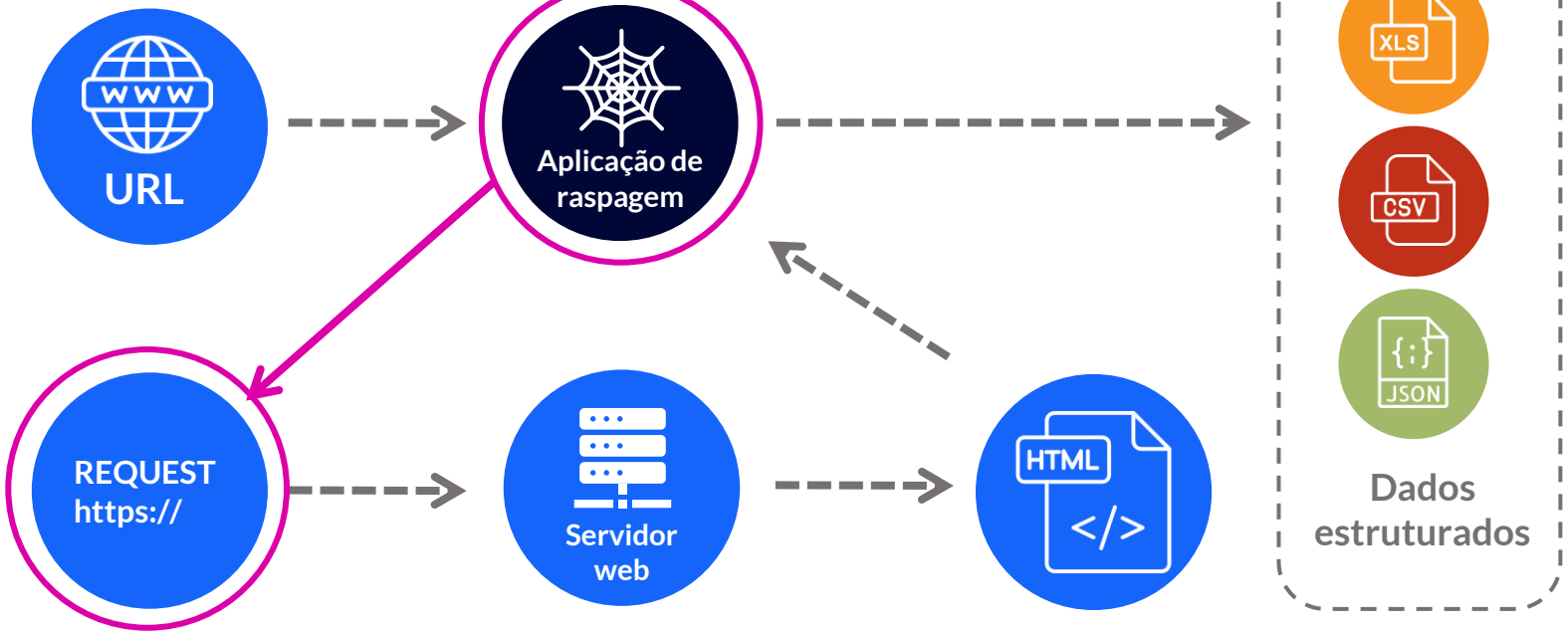
Raspagem de Dados

Processo de coleta de dados estruturados da web de maneira automatizada

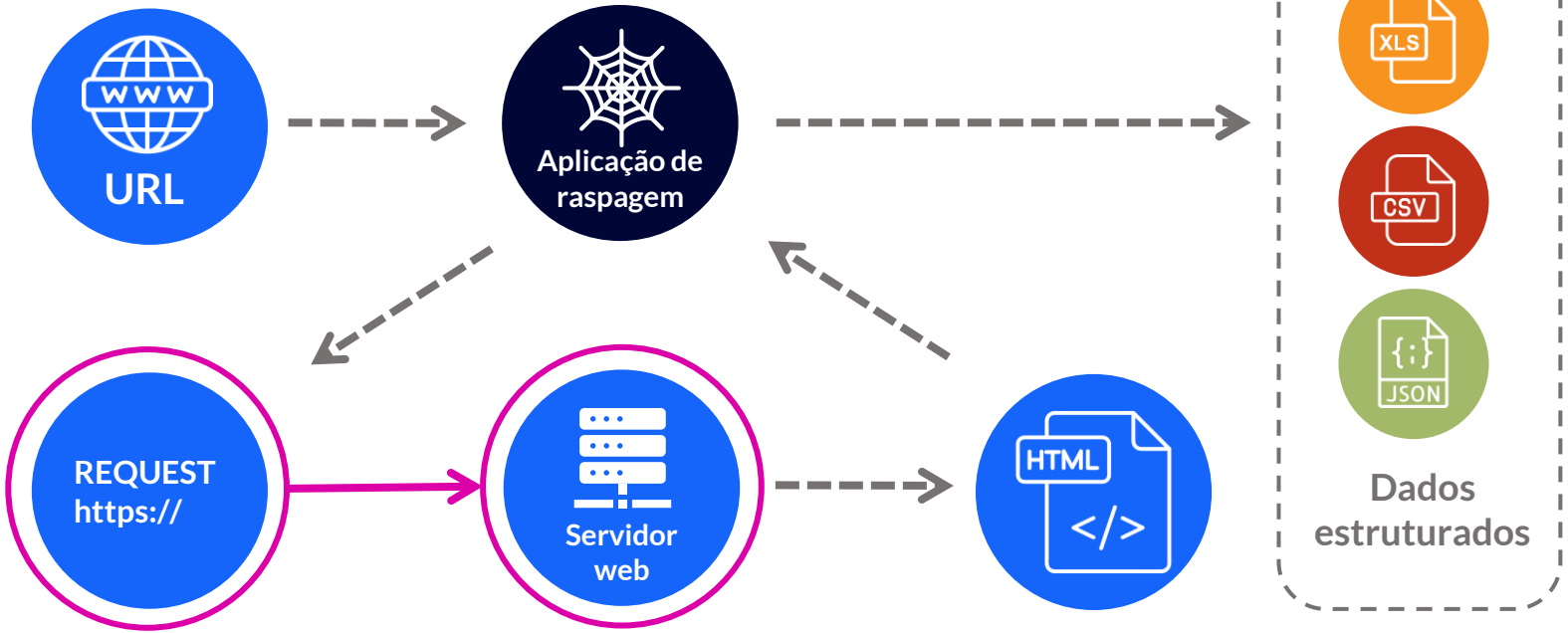
Raspagem de Dados



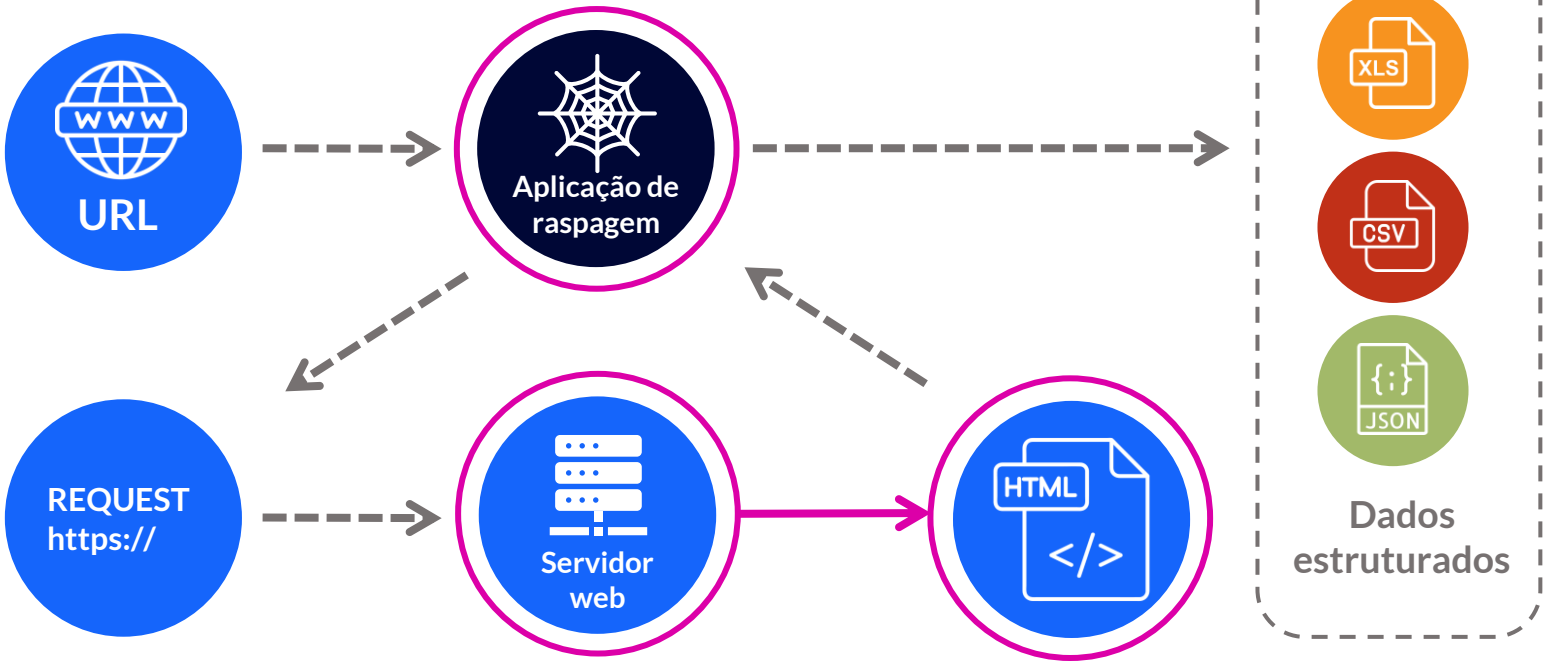
Raspagem de Dados



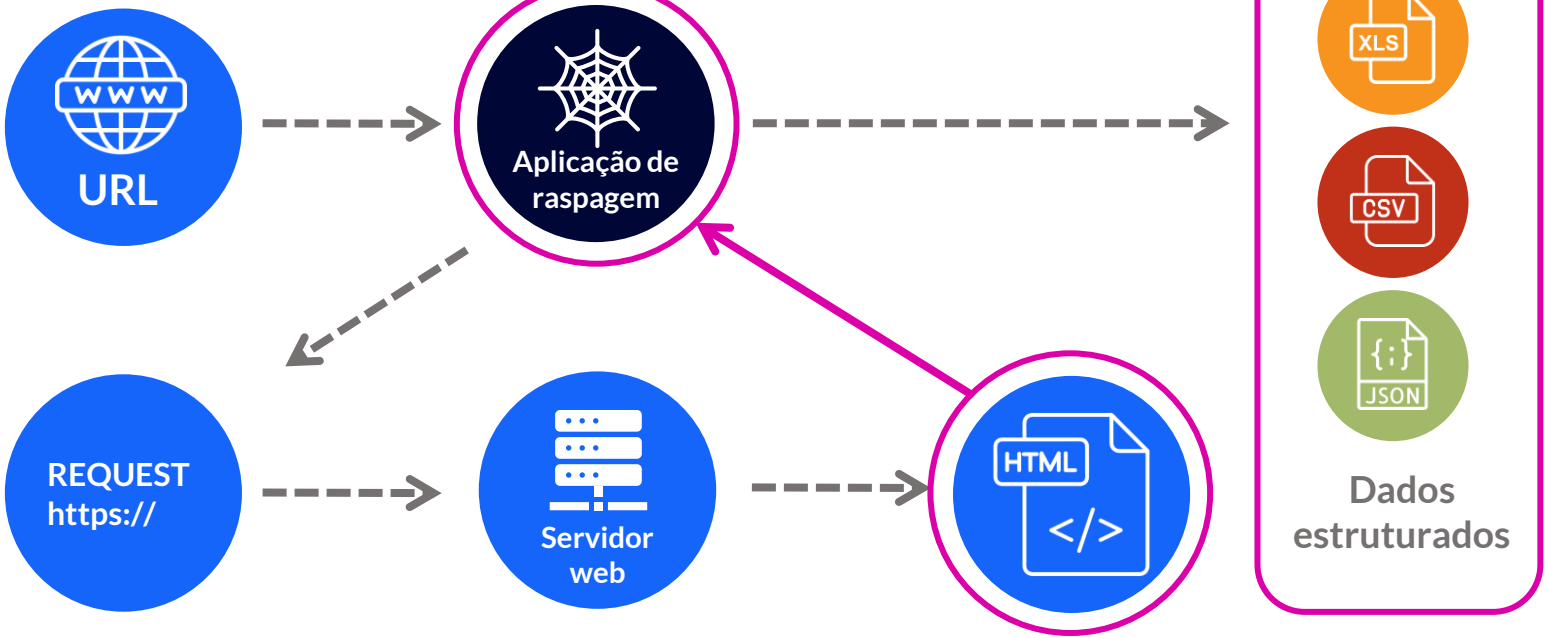
Raspagem de Dados



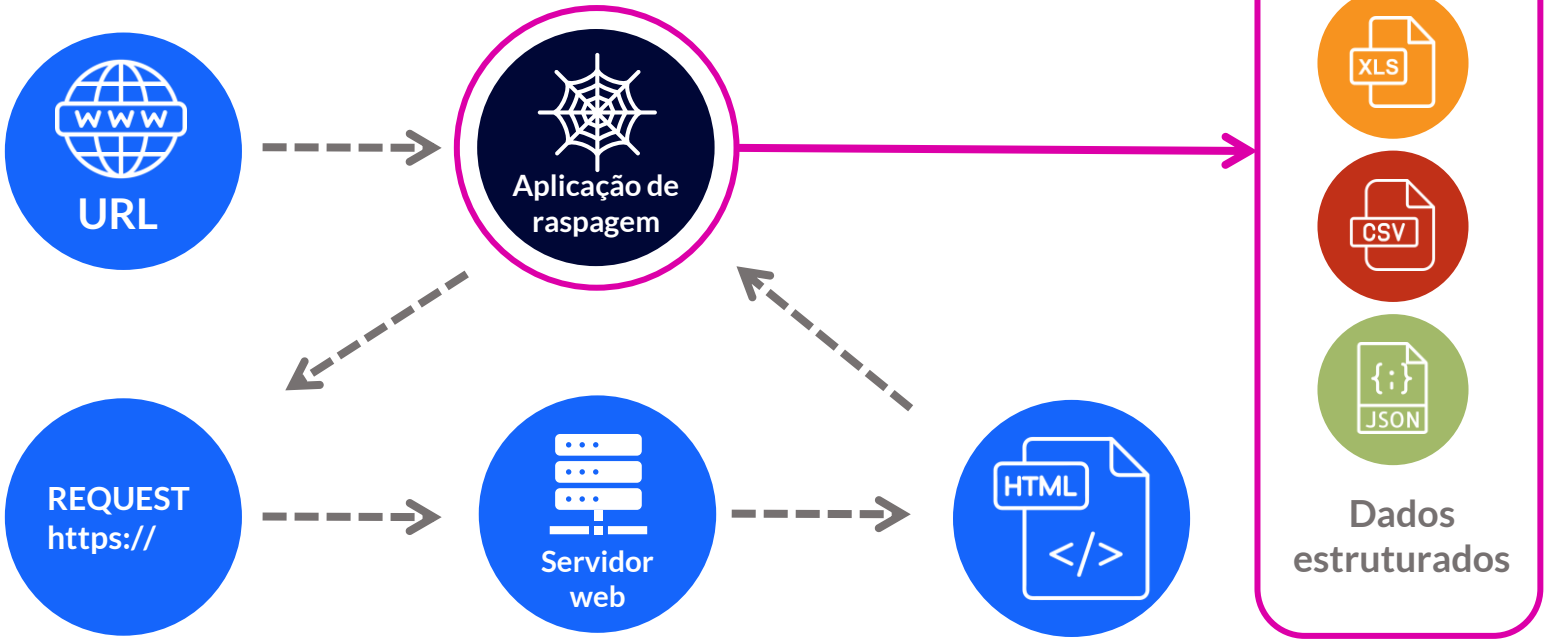
Raspagem de Dados



Raspagem de Dados



Raspagem de Dados



BeautifulSoup

BeautifulSoup

```
$   pip install beautifulsoup4
```

```
01  from bs4 import BeautifulSoup

02  p = requests.get(url, params)

03  s = BeautifulSoup(
        p.content,"html.parser")

04  r = s.find(“table”)
```

BeautifulSoup

```
$   pip install beautifulsoup4

01  from bs4 import BeautifulSoup

02  p = requests.get(url, params)

03  s = BeautifulSoup(
        p.content,"html.parser")

04  r = s.find(“table”)
```

BeautifulSoup

```
$   pip install beautifulsoup4

01  from bs4 import BeautifulSoup

02  p = requests.get(url, params)

03  s = BeautifulSoup(
        p.content,"html.parser")

04  r = s.find(“table”)
```

BeautifulSoup

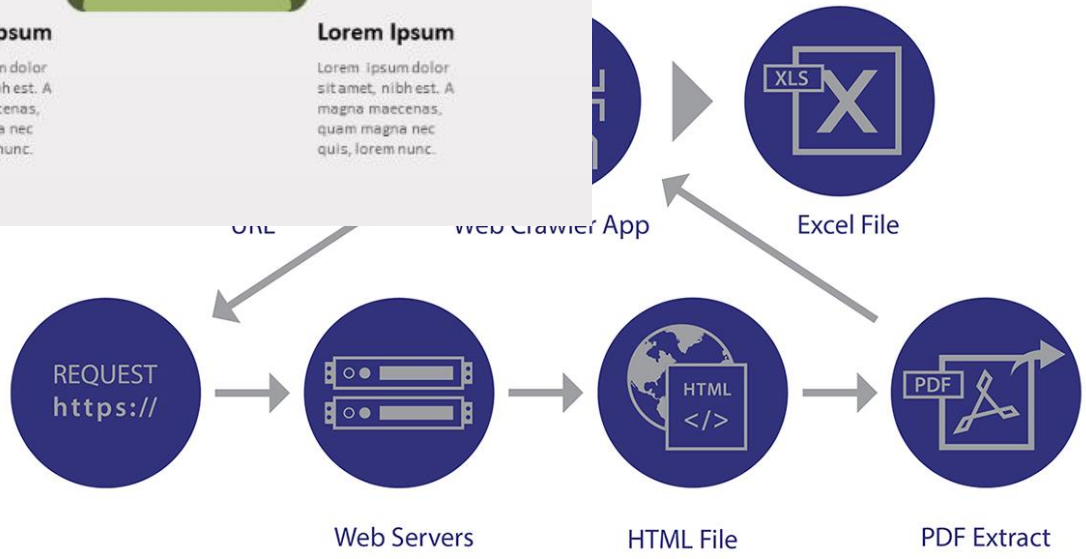
```
$   pip install beautifulsoup4

01  from bs4 import BeautifulSoup

02  p = requests.get(url, params)

03  s = BeautifulSoup(
        p.content,"html.parser")

04  r = s.find(“table”)
```

BeautifulSoup

```
$    pip install beautifulsoup4

01   from bs4 import BeautifulSoup

02   p = requests.get(url, params)

03   s = BeautifulSoup(
        p.content,"html.parser")

04   r = s.find(“table”)
```

PUC Minas
Virtual

Lorem Ipsum
Lorem ipsum dolor sit amet, nibh est. A magna maecenas, quam magna nec quis, lorem nunc.
Lorem Ipsum
Lorem ipsum dolor sit amet, nibh est. A magna maecenas, quam magna nec quis, lorem nunc.
LOREM IPSUM
LOREM IPSUM
LOREM IPSUM
LOREM IPSUM
LOREM IPSUM
Lorem Ipsum
Lorem ipsum dolor sit amet, nibh est. A magna maecenas, quam magna nec quis, lorem nunc.
Lorem Ipsum
Lorem ipsum dolor sit amet, nibh est. A magna maecenas, quam magna nec quis, lorem nunc.
Lorem Ipsum
Lorem ipsum dolor sit amet, nibh est. A magna maecenas, quam magna nec quis, lorem nunc.


XLS
Excel File
REQUEST
https://
HTML
</>
PDF
Web Servers
HTML File
PDF Extract