Obtenção de Dados e Dados Colunares

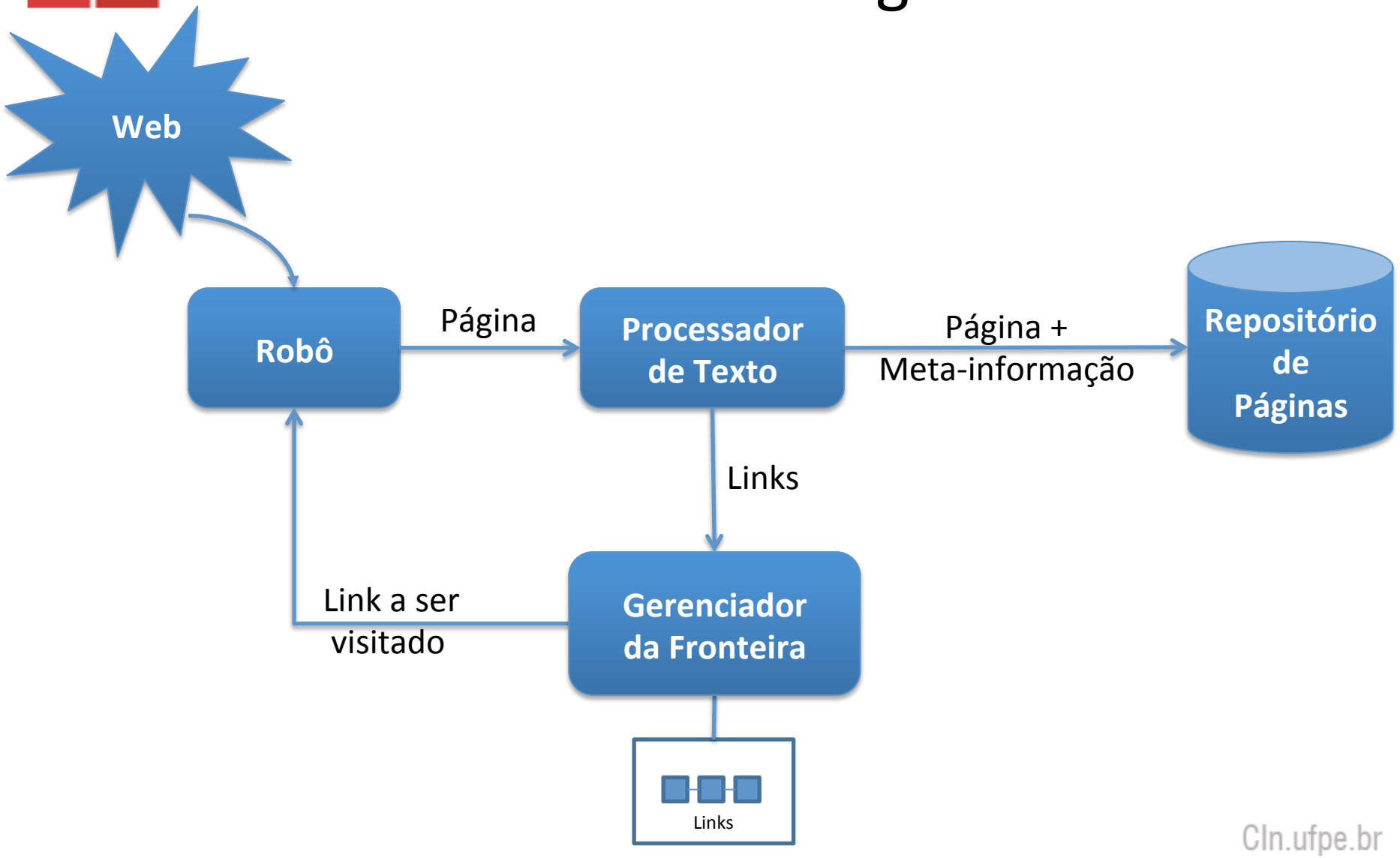
Luciano Barbosa

UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

Dados na Web

- APIs (https://www.pythonforbeginners.com/api/list-of-python-apis)
- Dados abertos
  - http://dados.recife.pe.gov.br/
  - https://opendata.cityofnewyork.us/
- Listas de datasets:
  - https://github.com/awesomedata/awesome-public-datasets
  - https://www.forbes.com/sites/bernardmarr/2016/02/12/big-data-35-brilliant-and-free-data-sources-for-2016/#525c0c58b54d

Coletando Dados de Páginas HTML



Crawlers e Parsers Disponíveis

- Crawlers
  - Scrapy: https://doc.scrapy.org/en/latest/
  - Pyspyder: https://github.com/binux/pyspider
  - MechanicalSoup: https://mechanicalsoup.readthedocs.io/en/stable/
- Parsers
  - BeautifulSoup: https://pypi.org/project/beautifulsoup4/
  - HTMLParser: https://docs.python.org/2/library/htmlparser.html
- Headless browsers
  - SeleniumHQ: https://www.seleniumhq.org/
  - PhantomJS: http://phantomjs.org/

Dados Colunares

- Dados da mesma coluna são armazenados continuamente

```
Hi Bob. How are you?,1508423069,238476,true
This is Alex.,1508423226,238476,true
Hi Alex. I am fine. How are you?,1508423238,9837498,false
```

Dado em linhas

```
Hi Bob. How are you?,This is Alex.,Hi Alex. I am fine. How are you?
1508423069,1508423226,1508423238
238476,238476,9837498
true,true,false
```

Dados colunares

Dados Colunares

- Permite consultas mais rápidas para a realização de análises
  - Orientada à coluna: varre rapidamente valores de uma coluna
  - Orientada à linha: varre todas as linhas e seleciona o valor da coluna desejada
- Facilita compressão: melhor estratégia de compressão para o tipo da coluna
- Exemplo de formato colunar: Apache Parquet (http://parquet.apache.org/)

Dados Colunares: Exemplo

date	price	size
2011-01-20	10.1	10
2011-01-21	10.3	20
2011-01-22	10.5	40
2011-01-23	10.4	5
2011-01-24	11.2	55
2011-01-25	11.4	66
...	...	...
2013-03-31	17.3	100

Dado orientado a linhas

date	price	size
2011-01-20	10.1	10
2011-01-21	10.3	20
2011-01-22	10.5	40
2011-01-23	10.4	5
2011-01-24	11.2	55
2011-01-25	11.4	66
...	...	...
2013-03-31	17.3	100

Dados colunares

- Operações: média, soma em colunas numéricas