Bộ dữ liệu Mũi tên

Mũi tên là gì?

Mũi tên cho phép xử lý và di chuyển một lượng lớn dữ liệu một cách nhanh chóng. Đó là số liệu cụ thể định dạng lưu trữ dữ liệu theo bố cục bộ nhớ cột. Điều này cung cấp một số ý nghĩa thuận lợi:

- Định dạng tiêu chuẩn của Arrow cho phép đọc không bản sao, loại bỏ hầu như tất cả quá trình tuần tự h trên không.
- Arrow không phụ thuộc vào ngôn ngữ nên nó hỗ trợ nhiều ngôn ngữ lập trình khác nhau.
- Mũi tên hướng theo cột nên truy vấn và xử lý các lát hoặc cột của dữ liệu.
- Arrow cho phép chuyển giao không cần sao chép sang các công cụ học máy tiêu chuẩn như NumPy, Gấu trúc, PyTorch và TensorFlow.
- Mũi tên hỗ trợ nhiều loại cột, có thể lồng nhau.

Ánh xạ bộ nhớ

Bộ dữ liệu sử dụng Arrow cho hệ thống bộ đệm cục bộ của nó. Nó cho phép các bộ dữ liệu được hỗ trợ bởi r bộ đệm đĩa, được ánh xạ bộ nhớ để tra cứu nhanh.
Kiến trúc này cho phép sử dụng các bộ dữ liệu lớn trên các máy có thiết bị tương đối nhỏ ký ức.

Ví dụ: tải toàn bộ tập dữ liệu Wikipedia tiếng Anh chỉ mất vài MB RAM:

```
>>> import os; import psutil; import timeit
>>> from datasets import load_dataset

# Process.memory_info được biểu thị bằng byte, vì vậy hãy chuyển đổi thành megabyte
>>> mem_before = psutil.Process(os.getpid()).memory_info().rss / (1024 * 1024)
>>> wiki = load_dataset("wikimedia/wikipedia", "20220301.en", split="train")
>>> mem_after = psutil.Process(os.getpid()).memory_info().rss / (1024 * 1024)

>>> print(f"RAM memory used: {(mem_after - mem_before)} MB")
Bộ nhớ RAM đã sử dụng: 50 MB
```

Điều này có thể thực hiện được vì dữ liệu Mũi tên thực sự được ánh xạ bộ nhớ từ đĩa và không được tải trong bộ nhớ.
Ánh xạ bộ nhớ cho phép truy cập dữ liệu trên đĩa và tận dụng khả năng bộ nhớ ảo cho tra cứu nhanh chóng.

Hiệu suất

Lặp lại tập dữ liệu được ánh xạ bộ nhớ bằng Arrow rất nhanh. Lặp lại Wikipedia trên một máy tính xách tay cung cấp cho bạn tốc độ 1-3 Gbit/s:

```
>>> s = """batch_size = 1000
... for batch in wiki.iter(batch_size):
... ...
None

>>> elapsed_time = timeit.timeit(stmt=s, number=1, globals=globals())
>>> print(f"Time to iterate over the {wiki.dataset_size >> 30} GB dataset: {elapsed_time:.1f} sec,
...f"ie. {float(wiki.dataset_size >> 27)/elapsed_time:.1f} Gb/s")
Thời gian lặp lại trên tập dữ liệu 18 GB: tức là 31,8 giây. 4,8 Gb/giây
```