Chapter. 02

데이터 저장

IHDF 및 기타 저장파일

FASTCAMPUS

ONLINE

금융공학/퀀트I

강사. 서찬웅

이번 시간에 배울 내용

강의내용

1. HDF란 무엇인가?
2. pandas에서 저장하는 방법
3. 다른 저장 방식들

FAST CAMPUS
ONLINE
서찬웅 강사.

I HDF에 대해서 알아보겠습니다.

- HDF란?
  - 계층적 데이터 형식(Hierarchical Data Format)은 HDF Group에 의해 관리되고 있는 대용량의 데이터를 저장하기 위한 파일 형식이다.
- 장점
  - 많은 양의 데이터를 저장 가능하다.
  - 검색 속도가 빠르다.
  - 병렬 입출력을 지원한다.
  - 데이터의 무작위 조회가 가능하다.
  - 20여년 이상 개발되어온 포맷으로 안정적이다.

- wikipedia

- 현재 버전은 HDF5이며, 파이썬에서는 h5py 라이브러리를 사용하여 데이터 저장이 가능합니다.
- 이번 수업에서는 DataFrame을 사용하여 hdf 형태로 저장하는 방법에 대해서 간단히 배워보겠습니다.
  - to_hdf() : hdf5 형태로 저장
  - read_hdf() : hdf5 파일 읽기

I Parquet에 대해서 알아보겠습니다.

- Parquet(파케이)란?
  - 하둡 에코 시스템의 무료 오픈 소스 컬럼 기반 데이터 저장 형식
- 장점
  - 빅 데이터 생태계에서 많이 사용되고 있음
  - 데이터는 컬럼 방식으로 가지고 있어서 훌륭한 표현력을 가짐
  - 데이터 분석에도 뛰어난 성능을 보임
  - 특정 컬럼의 값만 가져오는 쿼리는 전체 로우 데이터를 읽을 필요가 없기 때문에 성능이 향상됨
  - 서로 다른 인코딩 기술을 서로 다른 컬럼에 적용 가능

- pandas에서 Parquet 파일을 지원함
  - to_parquet() : 파케이 형태로 저장
  - read_parquet() : 파케이 형태의 파일을 읽기

I Feather에 대해서 알아보겠습니다.

- Apache Arrow 란?
  - Apache Arrow이라는 인메모리 컬럼 데이터 스토리지와 분석 프로젝트가 공개
  - Arrow는 메모리 상에서 컬럼 구조로 데이터를 구성 및 데이터를 사용할 수 있는 라이브러리 제공
  - Arrow 포맷은 CPU 캐시 로컬리티 특성을 극대화
  - 인텔 CPU의 명령어를 사용하여 벡터화 해서 활용할 수 있는 기능을 제공

- Feather이란?
  - DataFrame을 저장하기 위한 빠르고 가볍고 사용하기 쉬운 이진 파일 형식
  - Feather는 R하고 파이썬 및 다른 언어도 읽고 쓰기가 가능하고 쉽게 사용할 수 있도록 설계
  - 높은 읽기 및 쓰기 성능. 작업은 로컬 디스크 성능에 많은 영향을 받습니다.
- DataFrame에서 저장 및 읽기
  - to_feather() : feather 형식으로 저장합니다.
  - read_feather() : feather 형식의 데이터를 읽습니다.

I 정리

- hdf
- parquet
- feather
- 각각의 데이터 읽고, 쓰기 방법 및 시간

감사합니다

FAST CAMPUS
ONLINE
서찬웅 강사.