Database for Bigdata

MLP LAB 임경태

Limitation of Pandas

- **다음에 대해 생각해보자**

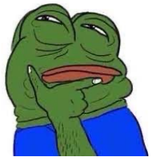
  - 검색: 거대한 Wikipedia의 모든 정보를 Padas로 불러와 검색을 하고싶다..

  - 저장: 매일 뉴스기사를 크롤링해 저장해두고 싶다..

  - 공유: 하나의 dataframe을 다른 사람과 동시에 조작하고 싶다..

  - 무결성: 데이터를 더 정교하게 정리하고 싶다

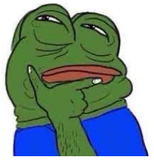
무결성: 데이터를 더 정교하게 정리하고 싶다



이건 뭔 말임?

종목번호	회사명	관련뉴스	뉴스날짜
36570	Ncsoft	시련의 엔씨, 출시예고 신작도 구조조정…"나 떨...05/01	5월 1일

무결성: 데이터를 더 정교하게 정리하고 싶다

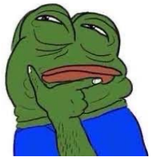


이건 뭔 말임?

종목번호	회사명	관련뉴스	뉴스날짜
36570	Ncsoft	시련의 엔씨, 출시예고 신작도 구조조정…"나 떨...05/01	5월 1일

종목번호	회사명	관련뉴스	뉴스날짜
36570	Ncsoft	시련의 엔씨, 출시예고 신작도 구조조정…"나 떨...05/01	5월 1일
36570	Ncsoft	엔씨소프트, 블소 IP로 중국 시장 ... 관련 3건04/30	4.3
36570	Ncsoft	엔씨소프트 블소2, 1인 콘텐츠 '무원의 탑' 최상...04/30	4.3
36570	Ncsoft	"기술로 장벽 없앤다"… 장애인 의사소통·게임 접...04/30	4.3
36570	Ncsoft	글로벌 무대 오르는 엔씨 "게임 경... 관련 2건	4.29

무결성: 데이터를 더 정교하게 정리하고 싶다



이건 뭔 말임?

종목번호	회사명	관련뉴스	뉴스날짜
36570	Ncsoft	시련의 엔씨, 출시예고 신작도 구조조정…"나 떨...05/01	5월 1일

종목번호	회사명	관련뉴스	뉴스날짜
36570	Ncsoft	시련의 엔씨, 출시예고 신작도 구조조정…"나 떨...05/01	5월 1일
36570	Ncsoft	엔씨소프트, 블소 IP로 중국 시장 ... 관련 3건04/30	4.3
36570	Ncsoft	엔씨소프트 블소2, 1인 콘텐츠 '무원의 탑' 최상...04/30	4.3
36570	Ncsoft	"기술로 장벽 없앤다"… 장애인 의사소통·게임 접...04/30	4.3
36570	Ncsoft	글로벌 무대 오르는 엔씨 "게임 경... 관련 2건	4.29

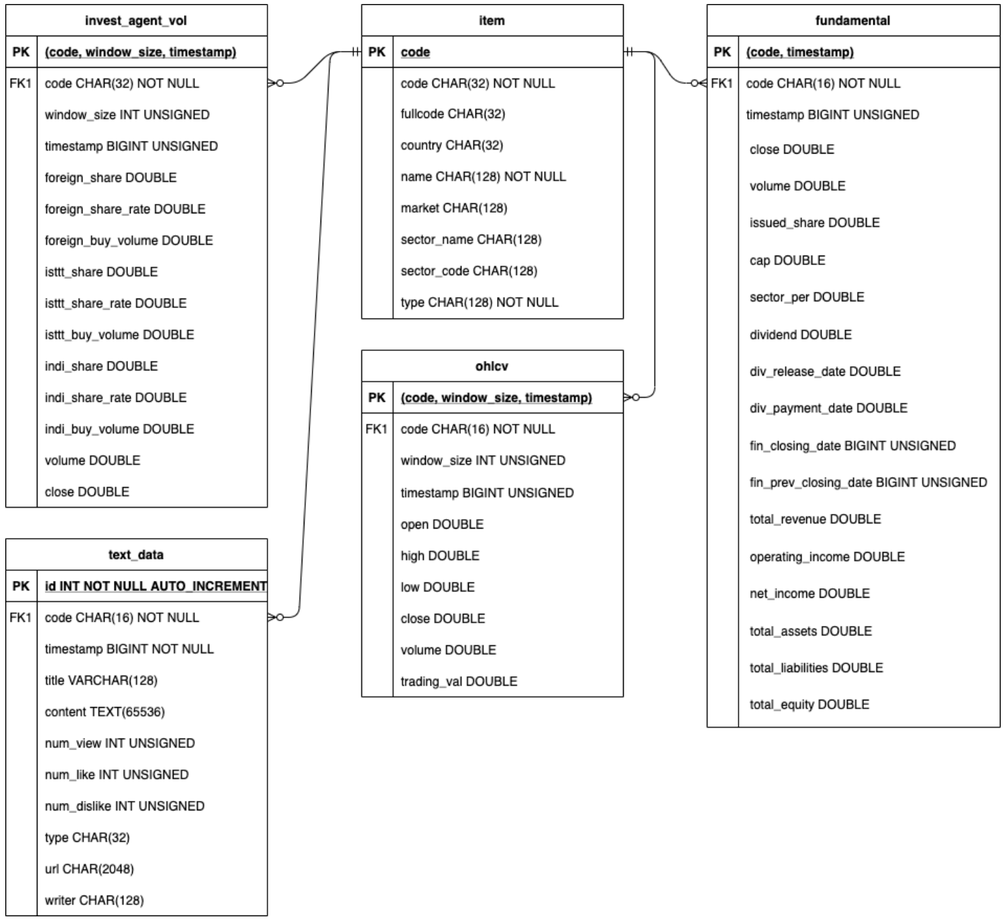
서로 이질적인 기준 (날짜, 시간)의 데이터는 pandas로 어떻게 저장함?

이러한 데이터가 많아지면?
= Bigdata
그럼 bigdata는 pandas로 관리 가능??

종목번호	회사명	관련뉴스	뉴스날짜	시간	주가
36570	Ncsoft	시련의 엔씨, 출시예고 신작도 구조조정…"나 떨...05/01	5월 1일	9시32분	174300
36570	Ncsoft	엔씨소프트, 블소 IP로 중국 시장 ... 관련 3건04/30	4.3		
36570	Ncsoft	엔씨소프트 블소2, 1인 콘텐츠 '무원의 탑' 최상...04/30	4.3		
36570	Ncsoft	"기술로 장벽 없앤다"… 장애인 의사소통·게임 접...04/30	4.3		
36570	Ncsoft	글로벌 무대 오르는 엔씨 "게임 경... 관련 2건	4.29		

Database system

- 빅데이터를 여러 사람이 동시에 무결성을 지키며 검색, 수정하는 시스템



그림참조 https://hanlue.tistory.com/8

Database system

- 빅데이터를 여러 사람이 동시에 무결성을 지키며 검색, 수정하는 시스템

Item

종목에 대한 정보를 저장한다. 캔들차트, 수급, 뉴스 등의 다른 데이터를 저장하기 전에 해당 종목의 정보를 이 테이블에 저장하는 작업이 선행되어야 한다.

column name	description
code	종목코드
fullcode	종목코드보다 길이가 긴 full code 가 있는 경우 사용. e.g.) 한국거래소 종목코드
country	종목이 상장되어 있는 나라
name	종목명
market	종목이 상장된 시장 (KOSPI, KOSDAQ, NYSE, NASDAQ 등)
sector_name	섹터명
sector_code	섹터코드
type	항목타입 (개별종목, 지수, 매크로, 선물 등)

그림참조 https://hanlue.tistory.com/8

Database system

- 빅데이터를 여러 사람이 동시에 무결성을 지키며 검색, 수정하는 시스템

text_data

뉴스, 토론글, SNS 등의 자연어 데이터를 저장

column name	description
code	item table 에 있는 종목코드
timestamp	글이 발행된 시간 (close price의 시점). 표기는 unix time 으로 한다.
title	제목
content	텍스트
num_view	조회수
num_like	좋아요 수
num_dislike	싫어요 수
type	글 유형 (뉴스, 자유게시판, SNS 등)
url	원문 url
writer	글쓴이 닉네임

그림참조 https://hanlue.tistory.com/8

Database system

- 빅데이터를 여러 사람이 동시에 무결성을 지키며 검색, 수정하는 시스템

ohlcv

각 항목들의 캔들차트를 저장한다. 데이터가 ohlc 형식이 아니라 단순 값 하나로 구성된 경우에는 close 에 가격을 저장한다.

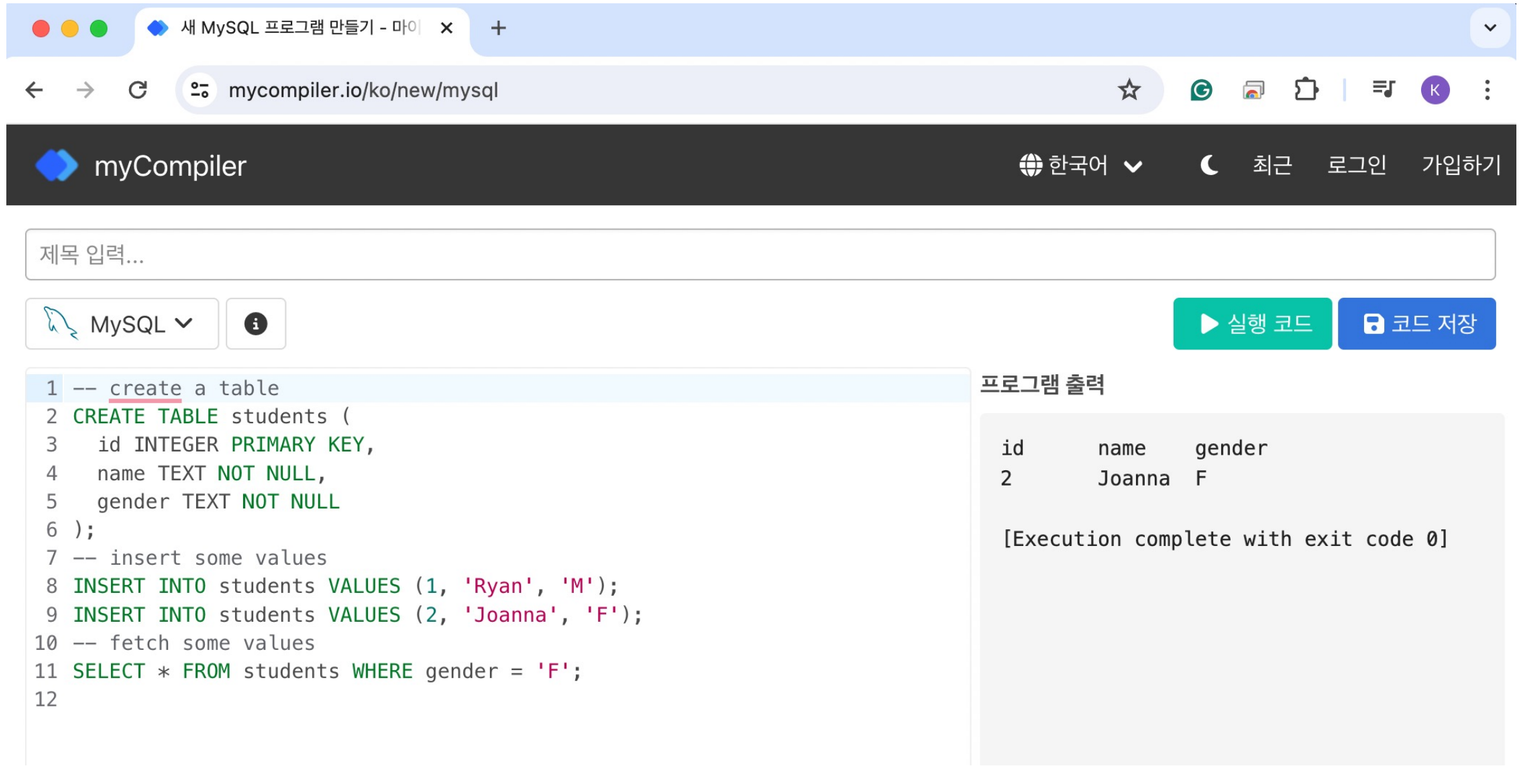
colume name	description
code	item table 에 있는 종목코드
window_size	캔들의 시간 크기 (1분, 30분, 1시간 등). 단위는 분으로 표기한다
timestamp	캔들차트가 완성된 시간 (close price의 시점). 표기는 unix time 으로 한다.
open	시가
high	고가
low	저가
close	종가 혹은 값
volume	거래량
trading_val	거래대금 (거래량 * 종가)

그림참조 https://hanlue.tistory.com/8

Database system

- 빅데이터를 여러 사람이 동시에 무결성을 지키며 검색, 수정하는 시스템

  - SQL 언어를 기반으로 검색, 수정, 삭제, 입력 등이 가능하며, Mysql, MariaDB, MongoDB, SQLight 등이 있음



그림참조 https://www.mycompiler.io/ko/new/mysql

감사합니다.