



p y t h o n

Data

p

분석을 위한

y

t

강사 정훈희

파이썬

h

o

n



1.개요 및 intro



이번시간에 학습할 내용은..

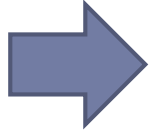
데이터 분석을 위한 파이썬 프로그램과 정
개요와 학습 커리큘럼 소개하는 시간이에요.

python은?

- ◆ 셸(shell)에서 명령을 내리고 엔터를 치면 곧바로 컴퓨터가 알아듣고 실행하여 결과를 보여주는 “인터랙티브(대화형) 모드”
- ◆ 프로그램 초보자도 쉽게 배울수 있는 언어
- ◆ 다른 저수준 언어와의 호환성이 좋은 언어



python은 인터프리터
형 언어라서 속도가
느리다고 들었는데 과
연 빅데이터 분석에도
파이썬이 유리할까?

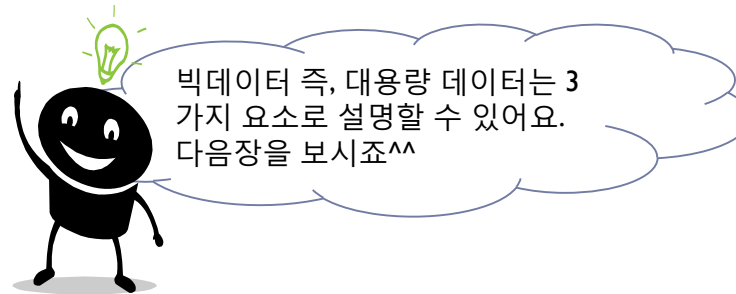


맞아요. 파이썬이 동시다발적 멀티스레드에 적합한 언어는 아니죠. 중요한건 빅데이터를 처리하는데 있어서 꼭 실시간 응답만을 필요로 하는건 아니라는점이에요. 빅데이터 이야기를 하면서 자세히 알아보죠^^



빅데이터란?

- ▶ 데이터의 규모에 초점을 맞춘 정의
 - 기존 데이터베이스 관리도구의 데이터 수집, 저장, 관리, 분석 역량을 넘어서는 데이터
- ▶ 업무 수행 방식에 초점을 맞춘 정의
 - 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고, 데이터의 빠른 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처



크기(Volume)

- ▶ 일반적으로 수십 테라바이트(terabyte) 혹은 수십 페타바이트(petabyte) 이상이 빅데이터의 범위에 해당
- ▶ 빅데이터는 기존 파일 시스템에 저장하기 어려울 뿐 아니라 데이터 분석을 위해 사용하는 기존 DW 같은 솔루션에서 소화하기 어려울 정도로 급격하게 데이터의 양이 증가하고 있음
- ▶ 이러한 문제를 극복하려면 확장 가능한 방식으로 데이터를 저장하고 분석하는 분산 컴퓨팅 방식으로 접근해야 함

속도(Velocity)

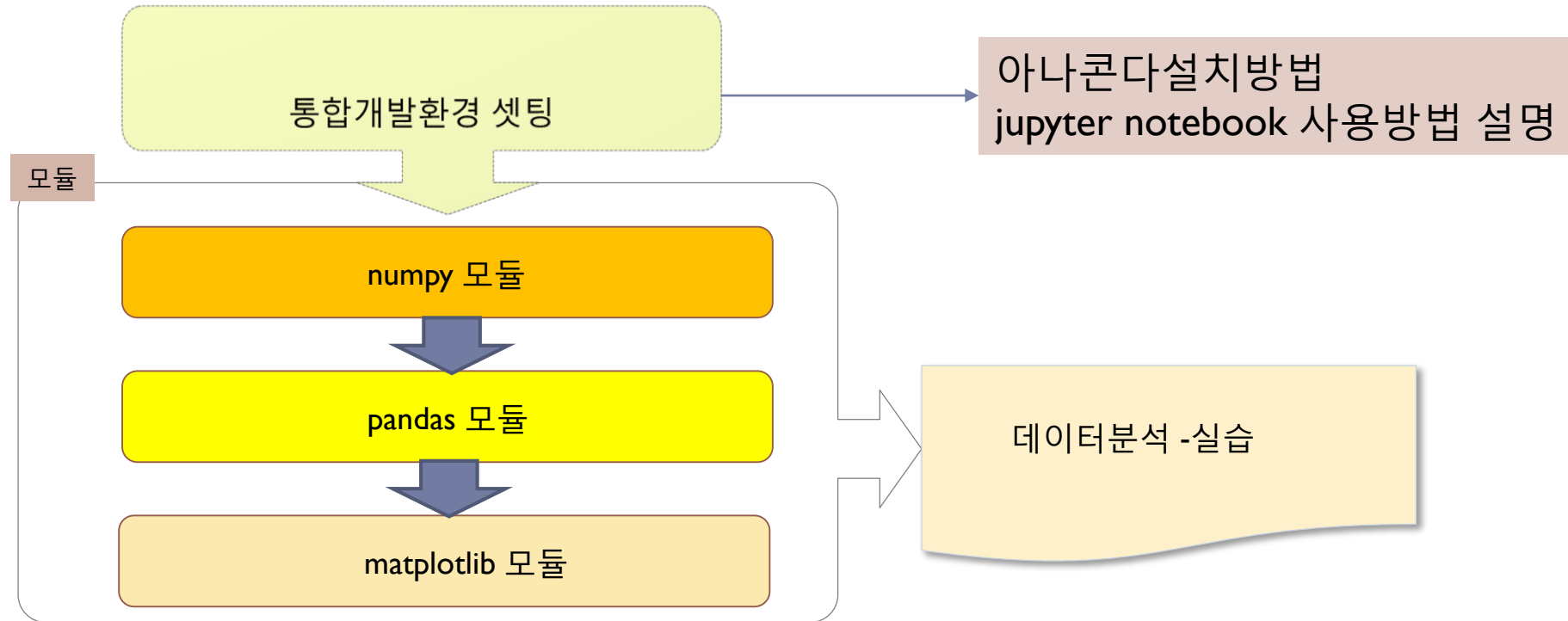
- ▶ 데이터처리를 크게 실시간 처리와 장기적인 접근으로 나눌 수 있음
- ▶ 오늘날 디지털 데이터는 매우 빠른 속도로 생성되기 때문에 데이터의 수집, 저장, 분석 등이 실시간으로 처리돼야 함
- ▶ 모든 데이터가 실시간 처리만을 요구하는 것은 아님
- ▶ 수집된 대량의 데이터를 다양한 분석 기법과 표현 기술로 분석해야 하는데, 이는 장기적이고 전략적인 차원에서 접근할 필요가 있음
- ▶ 통계학과 전산학에서 사용하는 데이터 마이닝, 기계 학습, 자연어 처리, 패턴 인식 등이 분석 기법에 해당함



다양성(Variety)

- ▶ 데이터는 정형화 정도에 따라 정형(Structured), 반정형(Semi-Structured), 비정형(Unstructured)으로 구분
- ▶ 정형 데이터는 고정된 필드에 저장되는 데이터를 의미하며, 일정한 형식을 갖추고 있음
- ▶ 반정형 데이터는 고정된 필드로 저장되지는 않지만, XML 이나 HTML 같이 메타데이터나 스키마 등을 포함
- ▶ 비정형 데이터는 고정된 필드에 저장되지 않는 데이터를 의미함. 사진, 동영상, 메신저로 주고 받은 대화 내용, 스마트폰에 기록되는 위치 정보, 통화 내용 등이 이에 해당
- ▶ 빅데이터는 비정형 데이터도 처리할 수 있어야 함

데이터 분석을 위한 파이썬 프로그램 과정 안내



- numpy 모듈 – numpy배열객체 생성방법 , 난수 생성, 산술계산방법 설명
- pandas 모듈 – 데이터 가공을 위한 핵심 패키지.

Series, DataFrame 자료구조 생성방법 및 관련 기본메소드,데이터를 읽어오는 방법 및 연산작업

- matplotlib모듈 – 데이터 시각화를 위한 패키지. jupyter notebook에서 interactive한 그래픽을 표현