+ 빅데이터

1. 정의
2. 서버 한대로 처리할 수 없는 규모의 데이터
3. 기존 소프트웨어로는 처리할 수 없는 규모의 데이터

- 기존 소프트웨어 : 오라클, Mysql같은 관계형 데이터베이스

-- 스케일업(Scale-up) : 기존 소프트웨어들이 데이터처리 용량을 늘리기 위해 메모리를 추가,

CPU 디스크를 더 장착하는 방식으로 서버리소스를 더 추가하는 방법  
 -- 스케일아웃(Scale-out) : 서버 자체를 추가하여 전체 시스템의 용량을 추가하는 방식

: 대표적 Nosql, 하둡과 같은 분산환경 시스템🡺저가장비 다수 사용

1. 3V (Value, Velocity, Variety) – 주로 컨설팅 회사가 쓰는 개념
2. Value(규모) : 데이터 크기가 얼마나 대용량인가
3. Velocity(속도) : 데이터가 얼마나 빠르게 생성되는가
4. Variety(다양성) : 데이터가 구조화/비구조화된 데이터를 다 포함하는지 여부
5. 빅데이터 예
6. 웹 검색엔진 데이터
7. 웹페이지 데이터

: 구글 등과 같은 웹검색엔진이 수 조개의 웹페이지를 크롤해서 인덱스를 만들어 냄.

: 중요한 것은 이 웹페이지가 고정되어 있지 않으므로 대용량의 데이터가 만들어 질 수 밖에 없음

: 페이지 랭크 알고리즘

1. 검색어 로그와 클릭 로그 데이터

: 현재 사용자들이 어떤 주제에 관심이 있는지 (검색어로부터 도출)

각 주제에 대한 의미있는 결과 (클릭되는 문서로부터 도출)

=> 이 모든 것들은 분석할 수 있는 중요 데이터

1. 디바이스에서 생성되는 데이터

ex) 스마트폰, 스마트tv, 보잉 제트기, 스마트미터(전기량, 풍량, 강우량…)

1. 소셜미디어 데이터

: 페이스북, 인스타, 트위터, 링크드인, 포스퀘어, …

1. 빅데이터 시스템 구성
2. 데이터 수집

- flume

- chukwq

- kafka

1. 데이터 저장/처리 모듈

- 대표적 : hadoop

가. HDFS 분산 파일 시스템

나. MapReduce 분산 처리 시스템(기법)

=> 데이터 수집 모듈을 통해 저장된 데이터를 HDFS에 저장하고 MapReduce를 통해 처리

참고. Hadoop 자체는 자바 언어로 만듦. MapReduce를 처리 언어 : Java, Javascript, C++ …

1. 처리 데이터 엑세스 모듈

참고. 처리된 데이터(결과)를 외부에 실시간으로 엑세스 해야한다면 하둡 자체는 적합하지 X

1. 기존의 관계형 데이터베이스

: 처리된 결과 데이터 크기가 상대적으로 작거나 별다른 검색이 필요하지 않을 때

- Sqoop

: 하둡과 관계형 데이터 베이스의 구조화된 데이터 소스간 송수신 가능하게하는 오픈 소스

1. NoSQL

참고. 데이터가 상대적으로 크면 스키마가 고정되어있지 않고, 데이터 트래픽이 아주 클때씀

RDBMS NoSQL

테이블 정의 시 스키마 필요 동적으로 새로운 필드를 추가, 삭제 가능

(필드 이름과 타입이 먼저 정의)

기본적으로 한 대의 서버에서 동작 처음부터 분산환경을 염두에 두고 설계

(처리량, 트래픽 처리 한계 有) 분산환경이므로 데이터 중복 저장 가능

서버가 몇대 고장나더라도 전체 서비스 다운X

> NoSQL 종류

- 행 기반(row-oriented) NoSQL : 데이터가 레코드별로 저장 (MongoDB)

- 열 기반(column-oriented) NoSQL : 레코드가 컬럼별로 저장 (HBase, Casandra)

1. 검색엔진

: 하둡으로 처리한 데이터를 key/value 보다 더 복잡한 형태로 엑세스 해야 할 경우가 있음

(Lucene, Solr, ElasticSearch..)

1. 워크플로우 관리 모듈

- Oozie : 야후 오픈 소스 프로젝트 (XML 설정)

- Cascading : JAVA API

- Azkaban : 링크드인 내부에서 사용하는 워크플로우 관리 오픈소스 툴

- Ambrose : 트위터 내부에서 사용하는 워크플로우 관리 오픈소스 툴

1. 데이터 시각화 도구

: Matlab, R

머신러닝에 벡터 데이터를 줘야 함

<https://www.anaconda.com/open-source> 에 쓰이는 오픈소스 쭉 있음

Anaconda Prompt

C:\python\_workspace에 bigdata 경로 생성

(base) C:\WINDOWS\system32>cd \python\_workspace\bigdata

(base) C:\python\_workspace\bigdata>conda info --envs

# conda environments:

#

base \* C:\Users\goott7\anaconda3

goott7 C:\Users\goott7\anaconda3\envs\goott7

(base) C:\python\_workspace\bigdata>conda activate goott7

(goott7) C:\python\_workspace\bigdata>jupyter notebook