LESSON 1

빅데이터(BigData)란?

빅데이터(BigData)란?

"BigData"의 의미

무엇이 "빅데이터"인가?

현재는 상업적/마케팅 용어로 변화

큰 데이터를 대상으로 하면 다 빅데이터?

빅데이터(BigData)란?

"BigData"의 의미

크기 **BIG DATA** 처리방식

모두 만족

HDFS와 같은 분산파일시스템과 MapReduce와
같은 병렬처리 프레임워크를 가지고 있어야
빅데이터 처리 시스템이라고 할 수 있음

대표적인 시스템 : 하둡(Hadoop)

빅데이터(BigData)란?

"Data"의 의미

- 대표적인 데이터를 다루는 기술 : DBMS(주로 RDBMS)
- 데이터를 저장(생성)하고 검색/수정/삭제
- CRUD(Create, Retrieve, Update, Delete)
  - 주로 테이블/레코드 형태로 관리

LESSON 2

빅데이터와 DBMS/NoSQL과의 차이점

빅데이터와 DBMS/NoSQL과의 차이점

RDBMS와 NoSQL과의 비교

- 보통 일반적인 DBMS는 빠른 읽기에 최적화(R >> CUD)
  - CUD가 일어나면 인덱스(Index) 수정
- 데이터가 빈번히 생성/수정/삭제되는 시스템에 적합하지 않음

NoSQL의 등장 → 빠른 쓰기에 최적화

- 하지만 양쪽 다 데이터가 커지는 경우에는 특수한 처리 필요
  - 수 십 테라가 넘어가는 경우
- 빅데이터는 일반적인 파일로 관리(주로 키/밸류 방식)

빅데이터와 DBMS/NoSQL과의 차이점

“Big”의 의미 – 크기

- 일반 DBMS(주로 RDBMS)
  - "Large" 데이터를 다룸
- "Very Large" DB(VLDB)
  - 사용되는 기술이 다름
  - 파티셔닝, 샤딩, 복제
- "Big" Data(데이터)
  - 일반 DBMS로 처리하려면 엄청난 비용
  - 처리 가능하지만 여러 가지 제약조건
    * 외래키 제약/조인-정규화 문제
  - 일반적인 DBMS로 처리하지 못할 정도로 큰 양
    - 통상 수 십 테라 이상의 크기를 가지는 경우

빅데이터와 DBMS/NoSQL과의 차이점

"Big"의 의미 – 크기

정의 1.
서버 한 대로 처리할 수 없는 규모의 데이터
ex) 10TB 소팅 in 1 서버

정의 2.
기존 소프트웨어로는 처리할 수 없는 규모의 데이터
Scale Up vs. Scale Out

정의 3.
3V(Volume, Velocity, Variety)
cf. 20TB in 액센츄어

빅데이터와 DBMS/NoSQL과의 차이점

"Big"의 의미 – 처리방식

큰 데이터를 어떻게 RDBMS와 다른 방식으로 처리할 수 있는가?

- 스케일러블(Scalable)한 방식으로 처리
  - 시스템 수를 추가하면 계속 처리용량이 커지는 방식
- 스케일 아웃(Scale Out)
  - 하둡의 경우 (하둡v1) 4,000대, (하둡v2) 10,000대까지 연결 가능
  - 처리 가능한 데이터량 수 십 PB까지 보통 DBMS는 수 테라바이트(TB) 정도의 크기까지 다룰 수 있음
- 시스템 추가 처리는?
  - 간단한 설정으로 가능(masters/slaves 파일에 주소 등록)

빅데이터와 DBMS/NoSQL과의 차이점

"Big"의 의미 - 처리방식

- 데이터를 나누어 저장하면서 생기는 문제
  - 여러 개의 시스템 중 어디에 원하는 내용이 있나
  - 일부 시스템/네트워크에 장애가 일어난다면?

- 해결책

1 분산파일시스템을 만들자

  - 여러 대의 시스템을 묶는 큰 파일시스템
  - 고가용성(HA : High Availability)를 제공
    - 동일한 정보를 여러 군데에 중복해서 저장
    - 중복성/다중화(Redundancy)

2 병렬처리방식으로 처리성능을 높이자

  - 작업을 나눠 동시에 처리하는 방식
  - 분업화

빅데이터와 DBMS/NoSQL과의 차이점

"Big"의 의미 – 처리방식

"[결 론]"

- 빅데이터는 데이터를 처리하지만
  일반적인 RDBMS나 NoSQL과 다른 방식으로 처리

  테이블/레코드 >>> 파일(키/밸류)

- 처리용량은 수 백 TB에서 수 십 PB까지

- 스케일러블하게 확장 가능한 구조
  - 분산파일시스템과 병렬처리 프레임워크를 통해

LESSON 3

빅데이터 성공사례

빅데이터 성공사례

- 주로 데이터마이닝(Data Mining)분야에서 많이 거론됨
  - 넷플릭스 사례, 페이스북/트위터 사례, 월마트 사례
- 제일 문제는 데이터! 테이터가 있어야 분석함
  - 서울시 공공데이터(http://data.seoul.go.kr)등
  - 데이터는 주로 소유가 있음 → 자유롭게 유통이 되지 않음
- 현재는 주로 내부 데이터를 가지고 기존에 처리하는 데 많은 비용이 들어 포기한 것을 대상으로 진행되는 경우가 많음
- 오픈소스라 별다른 비용이 들지 않음