

# (빅데이터 관점에서) RDB의 한계점, 이를 해결하기 위해 어떤 솔루션이 있을까?

Data Engineering

2018/09/03 10:40

<http://blog.naver.com/dgrsgr667/221351349083>

## 빅데이터란?

### 1. 빅데이터의 5가지 출처 → Internet of Events

빅데이터로 촉발된 디지털 경험세계에서 방대한 이벤트(Internet of Events)가 다양한 출처에서 발생하고 있음

이벤트의 예 : 여행사 사이트에서 비행기 티켓 구매, 공항버스 탑승을 위한 교통카드 승인, 휴대폰을 통한 사진촬영, 페이스북에 사진 올리기, 공항에서의 체크인 등등

- ① Internet of Content : 사람이 만든 특정 주제에 관한 방대한 정보 (ex. 위키피디아, 유튜브)
- ② Internet of People : 사람간 관계에 관한 모든 데이터 (ex. 페이스북, 트위터, 카카오톡)
- ③ Internet of Locations : 지리적인 차원을 가지는 모든 데이터, 스마트폰의 확산으로 더욱 많은 이벤트가 위치 속성을 가짐
- ④ Internet of Things : 네트워크에 연결된 모든 물리적 객체에 관한 데이터
- ⑤ Internet of Processes : 기업의 가치사슬을 구성하는 프로세스들에 관한 데이터

위 다섯 가지 출처 관련 데이터는 **서로 중복됨** : ex. 특정 개인이 찍은 사진은 위치 데이터를 포함할 수 있음

### 2. 빅데이터 출현의 중요 동인

무어의 법칙 : 어떤 기준에 따라 판단하건 (실리콘 칩에 들어가는 트랜지스터의 수를 기준으로 하든, 1달러로 구입할 수 있는 저장용량을 기준으로 하든) **컴퓨터의 성능은 18개월마다 2배로 증가한다.**

### 3. 빅데이터 개념

빅데이터 정의

: 기존의 데이터 수집, 저장, 관리, 분석 역량을 넘어서는 대량의 데이터 세트를 의미

: 데이터 규모와 기술 측면에서 출발했으나 가치와 활용, 효과 측면으로 의미가 점차 확대

### 4. 빅데이터 특징

- ① Velocity : 속도
- ② Variety : 다양성
- ③ Value : 가치
- ④ Veracity : 정확성
- ⑤ Volume : 크기

## 5. 빅데이터 활용 효과 : 전략적 가치 창출과 운영 효율 향상

전략적 가치 창출 - 의사결정 적시성 향상, 의사결정 효과성 향상, 선제적 의사결정 기반 마련

운영 효율 향상 - 내부 역량 개선, 업무 자동화 및 중복 제거, 프로세스 안정화

### 1. 빅데이터에 대한 5가지 미신

① 빅데이터 분석에서 가장 중요한 것은 데이터 규모이다.

-> 다양성과 속도가 더 중요

② 빅데이터는 SNS 데이터이고, 기업전략과 의사결정에 중요한 분석기회를 제공한다.

-> 기업 내부데이터가 더 많고 더 중요하다.

③ 빅데이터 분석기법을 적용하면 획기적인 성과를 거둘 수 있다.

-> 분석자체만으로는 성과를 낼 수 없고, 데이터에 근거한 의사결정 문화와 방식이 내재된 프로세스가 있어야 성과를 낼 수 있다.

④ 빅데이터 분석은 데이터 과학자 같은 전문가들의 업무이다.

-> 데이터 분석은 모든 임직원들의 기본 업무이다.

⑤ 분석요건이 정해지지 않으면, 분석을 수행할 수 없다.

-> 분석을 하면서 결과를 보고, 새로운 질문들이 도출된다.

## 6. Structured Data vs. Unstructured Data

: 80% of business-relevant information originates in unstructured form, primarily text.

RDB(Relational Database 관계형 데이터베이스)란?

데이터를 단순한 표(table)형태로 표현하는 데이터베이스로, 계층구조보다 사용자와 프로그래머간의 의사소통을 원활하게 할 수 있는 구조이다.

현재까지 가장 범용적으로 사용되고 있는 DBMS 입니다.

1970년 관계형 데이터베이스에 관련된 논문을 발표(Edger Frank Codd 박사)하여 관계형 데이터모델링을 정의합니다.

- 논문: 'A Relational Model of Data for Large Shared Data Banks'

1979년 첫번째 상업용 관계형DBMS (데이터베이스 Version2)가 RSI사(Relational Software Inc.)를 통해 출시되었습니다.

데이터를 열과 행으로 이루어진 테이블 형태로 관리하며, 테이블 간에 관계(Relation)라는 개념을 통해 연결할 수 있습니다.

테이블은, 레코드(로우, 행)들의 집합으로 구성되며, 레코드는 필드(컬럼)으로 구성됩니다.

대표적인 RDBMS

- Oracle, DB2, MS-SQL, Sybase, MySQL 등

SQL (Structured Query Language)라는 직관적이고 언어를 통해 데이터를 처리합니다.  
기존에 어플리케이션에서 처리하던 많은 기능들을 DBMS 및 SQL을 통해 지원가능합니다.  
일반적으로 높은 성능을 보여주고 있으며, 데이터의 일관성을 보장합니다.  
가장 오랜기간 사용되어 검증되고 보장된 기술입니다.

단점 : 복합 애플리케이션(CAD/CAM, GIS 등) 및 실시간 분석에는 적합하지 않으며, 테이블 구조 변경시 영향도 확인후 작업해야 합니다.(즉각적인 변경에 한계)

( 1970년 ~ 단점 내용의 출처 : [http://playdata.io/etcs/sqltutorial/level0\\_008.html](http://playdata.io/etcs/sqltutorial/level0_008.html) )

## 1. 데이터베이스 시스템

‘정보’와 ‘데이터’는 서로 다름

데이터베이스 정의 : 조직체의 응용 시스템들이 공유해서 사용하는 운영 데이터들이 구조적으로 통합된 모임이다. 데이터 베이스의 구조는 사용되는 데이터 모델에 의해 결정된다.

### 1-1. 데이터베이스 특징

- ① 데이터의 대규모 저장소, 여러 부서에 속하는 여러 사용자에게 의해 동시에 사용됨
- ② 모든 데이터가 중복을 최소화하면서 통합됨
- ③ 데이터베이스는 한 조직체의 운영 데이터뿐만 아니라 그 데이터에 관한 설명(데이터베이스 스키마 또는 메타데이터)까지 포함
- ④ 프로그램과 데이터 간의 독립성이 제공됨
- ⑤ 효율적으로 접근이 가능하고 질의할 수 있음

### 1-2. 데이터베이스 관리 시스템 DBMS

: 데이터베이스를 정의하고, 질의어를 지원하고, 리포트를 생성하는 등의 작업을 수행하는 소프트웨어

### 1-3. 데이터베이스 시스템 개요

- 데이터베이스 스키마

: 전체적인 데이터베이스 구조를 의미, 자주 변경되지는 않음

: 데이터베이스의 모든 가능한 상태를 미리 정의

- 데이터베이스 상태

: 특정 시점의 데이터베이스의 내용을 의미, 시간의 경과에 따라 계속해서 바뀜

#### 데이터베이스 스키마

DEPARTMENT (DEPTNO, DEPTNAME, FLOOR)

EMPLOYEE (EMPNO, EMPNAME, TITLE, DNO, SALARY)

#### 데이터베이스 상태

##### DEPARTMENT

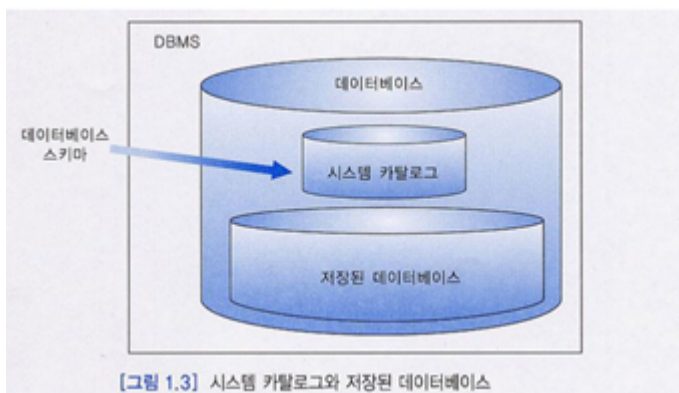
DEPTNO	DEPTNAME	FLOOR
1	영업	8
2	기획	10
3	개발	9

##### EMPLOYEE

EMPNO	EMPNAME	TITLE	DNO	SALARY
2106	김창섭	대리	2	2000000
3426	박영권	과장	3	2500000
3011	이수민	부장	1	3000000
1003	조민희	대리	1	2000000
3427	최종철	사원	3	1500000

【그림 1.1】 데이터베이스 스키마와 데이터베이스 상태

#### - 데이터베이스 시스템(DBS)의 구성요소



【그림 1.3】 시스템 카탈로그와 저장된 데이터베이스

#### - DBMS

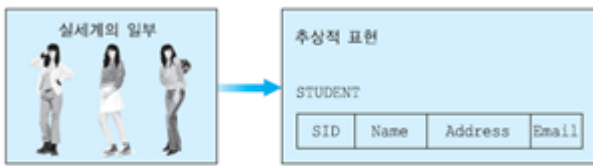
- : 사용자가 새로운 데이터베이스를 생성하고
- : 데이터베이스의 구조를 명시할 수 있게 하고
- : 사용자가 데이터를 효율적으로 질의&수정할 수 있도록 하며
- : 시스템의 고장이나 권한이 없는 사용자로부터 데이터를 안전하게 보호하며
- : 동시에 여러 사용자가 데이터베이스를 접근하는 것을 제어하는 소프트웨어 패키지
- : 데이터베이스 언어라고 부르는 특별한 프로그래밍 언어를 한 개 이상 제공
- : SQL은 여러 DBMS에서 제공되는 사실상의 표준 데이터베이스 언어

#### 1-4. DBMS의 발전과정

##### - 데이터 모델

- : 데이터베이스의 구조를 기술하는데 사용되는 개념들의 집합인 구조(데이터 타입과 관계)
- : 이 구조 위에서 동작하는 연산자들, 무결성 제약조건들
- : 사용자에게 내부 저장 방식의 세세한 사항은 숨기면서 데이터에 대한 직관적인 뷰를 제공
- : 그와 동시에 이들 간의 사상을 제공
- : (추가설명 NAVER 지식백과) - DB를 구축할 때 체계화된 구조를 갖추는 것이 필요 -> 이때 DB구조를 명시하기 위한 개

념들의 집합을 data model이라고 한다. data model은 [data의 구조+구조에서 허용되는 연산+이런 구조와 연산에 대한 제약조건]을 포함하는 개념이다.



[그림 1.9] 관계 데이터 모델에서의 실세계 표현

### - 데이터 모델(DB의 구조결정, 데이터 모델링의 결과물을 표현하는 도구)의 분류

(추가설명) 데이터 모델링 : 현실세계의 데이터를 컴퓨터 세계의 DB로 옮기는 과정

EX. 코끼리에 대한 중요 데이터 추출 ⇨ [개념세계로 옮기는 과정 = 개념적 데이터 모델링] ⇨ [개념세계의 data에 대해 DB에 저장할 구조를 결정하고 이 구조로 표현하는 작업 = 논리적 모델링]

#### 1) 고수준 또는 개념적 데이터 모델

: (추가설명) 데이터 구조 = 추상화 했을 때 어떤 요소로 이루어져 있는지 표현하는 개념적 구조

: (추가설명) 현실세계 ⇨ (추상화)개념적 데이터 모델링 ⇨ DB의 개념적 구조로 표현하는 도구

: 사람이 인식하는 것과 유사, DB의 전체적인 논리적 구조를 명시

EX. ER(엔티티-관계)데이터 모델과 객체 지향 데이터 모델

#### 2) 표현(구현)데이터 모델

: (추가설명) 데이터 구조 = 어떤 모습으로 데이터를 저장할 것인가 표현하는 논리적 구조

: (추가설명) 개념적 구조 ⇨ 논리적 데이터 모델링 ⇨ DB의 논리적 구조로 표현하는 도구

: 최종 사용자가 이해하는 개념, 컴퓨터 내에서 데이터가 조작되는 방식과 멀리 떨어져 있지 X

EX. 관계데이터 모델(에 기반하여 표현된 DB = 관계형 DB), 네트워크 데이터 모델

#### 3) 저수준 또는 물리적 데이터 모델

DB에 데이터가 어떻게 저장되는가를 기술

EX. Unifying, ISAM, VSAM ...

### - 관계DBMS

: 모델이 간단, 이해하기 쉬움

: 사용자는 자신이 원하는 것만 명시 ⇨ 데이터의 위치와 접근방식은 DBMS가 결정

EX. Oracle, MySQL Server ...

### 빅데이터 관점에서 RDB의 한계점은?

스마트폰을 시작으로 테블릿 PC까지 다양한 기기가 확산됨에 따라 데이터 양이 급증하게 되면서 개인화된 맞춤형 광고, 추천 상품 등 로그 분석이 기업 전략으로 활용되고 있다. 이처럼 빅데이터가 비즈니스에 활용되다 보니 새로운 솔루션에 대한 수요가 급증하고 있다. 그러나 기존 RDBMS는 대용량/비정형 데이터에 대한 수집, 저장, 처리 속도, 비용 효율성, 확장 측면에서 기존 관리 역량의 한계가 노출되기 시작했다. 이러한 이유로 꼭 필요하면 RDB를, 그렇지 않으면 용도에 맞는 데이터 스토어를 사용하자는 NoSQL이 등장하기 시작했다.

NoSQL은 NOT ONLY SQL의 약어로 RDBMS의 관계형 특징을 제거하고 만들어진 모든 DBMS를 말한다.

기존의 RDBMS를 대신해 효과적으로 빅데이터 분석을 할 수 있는 기술로 최근 각광받고 있다. NoSQL은 RDBMS와는 다르게 비관계형 DBMS로 대규모 데이터를 유연하게 처리할 수 있는 것이 장점이다.

(해당 단락 출처 : <http://www.dator.co.kr/bmonthly/1333146>)