

Ch12 데이터 구조

Ch12 Data Structure

원저자: Amy J. Ko, Brett Wortzman

발췌 + 한국어 번역: Alm Chung

[... 전략 ...]

데이터베이스

위에서 이야기한 데이터 구조(data structure)가 컴퓨터 프로그램 여러 곳에서 쓰이고 있고, 소프트웨어가 저장하고 연산하는 다양한 타입의 데이터들을 처리하는데 필수이긴 하지만, 이는 현대 소프트웨어가 대량의 데이터를 저장하는 방식은 아닙니다. 예를 들어, 몇 억 명의 유저에 대한 기록을 저장해야 하는 거대 소셜 미디어 플랫폼을 생각해봅시다. 이렇게 많은 양의 데이터를 저장해 두고, 필요한 연산을 위해 신속하게 데이터를 불러오기 위해서는 한 대의 컴퓨터와 그의 메모리를 뛰어넘는, 새로운 방식이 필요했습니다.

이런 이유로 컴퓨터 과학자들이 발명한 것이 데이터베이스입니다. 일반적으로 말해서 데이터베이스는, 어떤 형태가 되었든, 검색이 가능하게끔 데이터 묶음을 구조화하여 저장하는 컴퓨터 프로그램을 말합니다. 가장 흔한 타입의 데이터베이스는 관계형 데이터베이스(relational database)로 알려져 있는데요, 열과 행으로 이루어진 스프레드시트[역주: 엑셀이나 넥셀, 한셀에서 보는 형태를 떠올려볼 수 있습니다]를 묶어 둔 것이라고 생각해볼 수 있습니다. 하지만 데이터베이스는 스프레드시트과 많은 면에서 차이점을 보입니다:

- 스프레드시트가 ‘딱히 어떤 특정한 구조나 뜻이 부여되지 않은 셀(cell, 빈칸)들이 모여있는 표’의 형태를 가지고 있는 것과 달리, 데이터베이스는 데이터를 다른 여러 종류의 데이터를 담은 별개의 테이블들로 정리합니다.
- 데이터베이스의 테이블이 스프레드시트보다 더 엄격하게 구조화되어 있습니다: 데이터베이스 각각의 테이블은 1개 혹은 그 이상의 열과 0개 혹은 그 이상의 행을 가지며, 각각의 열은 데이터의 특정 속성을 담으며, 각각의 행은 특정 개체(entity)를 담습니다.
- 셀안에 있는 데이터를 읽어오거나 간단한 수식을 써 (합계 sum나 평균값 average과 같은) 행과 열에 대한 연산을 하는 엑셀과는 다르게 데이터베이스는 쿼리(query)를 이용해 데이터를 읽습니다.

[...]

다른 데이터 구조들에 비해 데이터베이스는 여러 가지 확실한 장점들을 가지고 있습니다. 예를 들어, 페이스북[메타]의 모든 데이터를 독립 변수나 거대한 리스트(list) 혹은 딕셔너리(dictionary) 자료 구조로 표현한다고 상상해보세요: 어떤 방법으로도 이런 리스트가 컴퓨터 한 대의 메모리에 다 들어갈 수 없을 것입니다. 데이터베이스를 사용하면 데이터를 여러 대의 컴퓨터에 분산해 저장할 수 있고, 또 여러 곳에서 쿼리를 할 수 있어, 수 억 명의 사람들이 웹 애플리케이션을 빠르고 안정적으로 사용할 수 있습니다.

하지만 컴퓨터 과학이 변수, 리스트(list), 레코드(record), 트리(tree), 데이터베이스, 그리고 다른 데이터 구조들에 대해 기술적으로 깊은 이해를 축적하는 와중, 이 노력은 순수하게 기술적인 면에만 집중되어 있었습니다 (예를 들어 각각의 데이터 구조가 어느 정도의 메모리를 차지하는지, 그리고 이런 각각의 차이가 성능에 어떤 영향을 미치는지, 등). 하지만 데이터 구조에는, 얼마나 빠르지에 대한 질문 외에도 고려해봐야 할 많은 것들이 있습니다. 윗 섹션에서 다뤘듯이, 현실 세계를 표현하기 위해 어떤 데이터 조직 방법을 선택하느냐에 따라 사람들을 위험에 노출시킬 수 있는 민감한 정보를 직접적으로 담게 될 수도 있고, 무엇이 가능하고 무엇이 바람직한지에 대한 정치적 견해들을 반영하게 될 수도 있습니다. 이러한 고려 사항은 기본적으로 [1] 현실 세계에 대한 충실도, [2] 현실 세계에 존재하는 예외적이거나 주변적이거나 예상치 못한 현상에 대한 탄력성, 그리고 [3]

이 정보가 현실 세계에서 어떤 뜻을 의미하고 이 정보가 어떻게 변화하여 어떻게 데이터 구조가 바뀌어야 하는지에 대한 프로그램과 그 개발자의 이해도가 되겠습니다.

2020년도에 출판한 *Data Feminism*¹라는 책을 통해 D'Ignazio와 Klein은 데이터 및 데이터 구조에 대한 반억압적 관점들을 소개해줄 수 있는 몇 가지 원칙들을 뽑아냈습니다. 이 원칙들은 다음과 같이 데이터 구조에도 적용 가능합니다 (페이스북[메타]의 유저 프로필 데이터베이스를 예로 들어보겠습니다):

- 세계를 지배하는 권력의 도구로써 데이터 구조가 쓰이고 있는지 검토한다 (예: 어떻게 페이스북[메타]의 데이터베이스가 광고를 타깃 하는데 쓰였는가 검토하는 것).
- 데이터 구조가 강화하는 불공평한 권력 구조에 도전한다 (예: 페이스북의 데이터베이스 테이블에서 쓰이고 있는 정체성 표현에 의문을 가져보는 것).
- 정보가 표상하려 하는 구체적인 형태와 데이터 구조를 연관 짓는다 (예: 페이스북 데이터베이스 프로필이, 비록 불안전하고 문제적인 방식으로 우리를 어느 정도 묘사하고 있긴 하지만, 결국 우리 자신이 아님을 인식하는 것).
- 컴퓨터 과학의 기본(foundational) 데이터 구조에 내재되어있는 카테고리화 위계에 의문을 가져본다 (예: 페이스북 데이터베이스 모델이 그렇게 주장하고 있지만, 친구관계는 2진수가 아님을 생각해 보는 것).
- 데이터에 대한 다원적 관점을 수용한다 (예: 다른 사람들이 다른 방식으로 자신을 표현할 수 있는 것을 인정하는 것, 따라서 하나로 통일된 데이터베이스 표현법은 충분하지 않다는 것).
- 데이터와 데이터 구조가 쓰이는 맥락을 고려해본다 (예: 페이스북 프로필의 주목적은 타깃 광고이지 우리의 친구 네트워크와 소통하는 것이 아니라는 것을 인지하는 것).
- 노동을 가시화한다 (e.g., 페이스북 프로필에 우리가 만들고 제공하는 모든 데이터 하나하나에 어떤 가치가 있으며, 유저들은 그 가치를 자신과 관련성 있는 광고 형태로 보답받고 있음을 인지하는 것).

¹ Catherine D'Ignazio & Lauren F. Klein (2020). *Data Feminism*. MIT Press

그러므로 데이터베이스와 어떤 종류 든 간의 데이터 구조를 디자인하는 사람들은 자신들에게 주어진 선택들이 제한되어 있는 상황에 대해 대처하고, 작업 중 데이터에 관해 드는 가정(assumptions)에 겸손하게 접근하고, 존재하는 억압 시스템을 강화하는 것을 피하거나 새로운 억압 시스템을 만드는 것을 경계할 책임이 있습니다. 그리고 데이터를 쓰는 사회 모든 구성원들을 통틀어 데이터를 쓰는 사람들은, 컴퓨터가 현실을 완벽하게 재현할 수 없다는 것, 그리고 이 불완벽함이 위험한 상황을 초래할 수도 있다는 사실을 인식할 필요가 있습니다.

[... 후략 ...]