

컴퓨터학콜로키움 (COSE-405)

:: 2주차 정연돈 교수님 데이터베이스 연구실 소속 최달수 선배님

강의 : 2022-09-14 / 작성 날짜 : 2022-09-14

고려대학교 컴퓨터학과 2017320108

고재영

우리는 이 세상을 살아가면서 수도 없이 많은 내용을 보고 듣고 경험하고 느낀다. 수천가지 물체로부터 다양한 자극들을 느끼며 살아가는데 이 모든 것이 정보이며 데이터로 표현되기도 한다. 그러한 정보가 본인에게 있어서 사소한지 중요한지 경중에 상관없이 범람하는 데이터가 쏟아짐에 따라, 이러한 데이터를 간과하지 않고 여러 방면에 있어서 응용하고 효율적이고 이득이 되는 방향으로 활용할 수 있는 능력이 중요시되는 것도 그리 놀라운 일은 아니다. 이번 세미나에서는, 정연돈 교수님의 고려대학교 데이터 베이스 연구실 소속 최달수 선배님의 강연으로, 데이터 베이스 분야에서의 몇 가지 키워드와 현재 수행 중이신 연구주제에 대해 설명해 주셨다.

첫 번째로 데이터베이스 분야에 현재 화두가 되고 있는 주제에 대해 소개하셨다. 모던 하드웨어나 블록체인과 같은 주제도 있지만, 크게 2가지를 꼽자면 클라우드 데이터베이스와 기계학습을 통한 데이터베이스 향상이 화두라고 말씀하셨다. 기존 데이터베이스의 경우, 사용자가 하드웨어 구성, 클러스터 관리, 확장성과 같이 하나부터 열까지 직접 처리해야 하는 on-premise 방식으로 고객들이 전적으로 모든 부담을 지기 때문에 어려움이 존재했다. 그러나 현재는 이 어려움을 기업들이 일부 경감시키는 동시에 고객은 사용한만큼의 비용만 지불할 수 있도록 클라우드 데이터베이스 방식으로 Amazon, Google, Snowflake, DataBricks, SAP, Oracle 등의 기업들이 서비스하고 있다. 여기에 심지어 기계학습을 도입하는 새로운 패러다임의 전환으로, 기존에 DBA와 데이터베이스 이론 기반으로 Rule-based model의 한계점을 극복하여 work query에 대한 예측과 server를 조절하는 auto-scaling, 학습을 통한 index 활용 등에 기계학습을 응용한 발전된 형태임을 설명하셨다.

두 번째이자 마지막으로, 현재 연구하고 계신 'Waffle'이란 인덱싱 시스템에 대한 전반적인 내용과 모델 내용, 그리고 작동 흐름을 설명하셨다. 해당 시스템은 강화 학습 기반의 동체에 대한 in-memory grid index의 튜닝 시스템이다. In-memory라는 키워드가 생소했는데 기존에 필요한 데이터를 디스크에서 메모리에 가져와 서비스를 수행한 반면, 메인 메모리에 전부 가져와서 처리하는 개념이라는 점이 인상깊었다. Waffle의 중추적인 개념으로는 동체 특성 상 위치의 지속적인 업데이트가 중요하기 때문에, 효율적인 knob setting을 정의하여 이를 통한 grid indexing을 수행하는 것이다. 그 이외에 여러가지 디테일을 설명하셨지만, 개인적으로 재밌었던 부분을 꼽자면, knob setting을 새로 설정하여 기존 index를 새로운 index로 변환하는 과정을 칭하는 Regrid 작업에서 누적되는 query에 대해 block되어 지연시간이 발생하는 문제였다. 처음에 생각하기에 물론 학부생인 필자 입장에선 다른 기타 조건들에 대해서 깊이 알고 있지 않지만, 학부에서 배운 바로 blocking이 일어난다면 역시 해당 transaction 간에 lock을 설정하여 해결하는 방법을 떠올렸는데, 마침 실제로 해결하신 방법이 일련의 과정을 하나의 transfer transaction으로 명명하여 locking을 통해 concurrency control이었다는 점이 굉장히 흥미로웠다.

해당 강의를 듣고 나서, 첫 번째로 받은 인상은 생각보다 학부에서 배운 여러가지 기초 과목들이 긴밀하게 연결되어 있다는 점이였다. 학부 과정 중에서 재미를 느껴서 열의를 가지고 데이터베이스 강의를 즐겁게 수강했던 기억이 있으나 AI와 머신러닝, 컴퓨터 비전 쪽과는 어느 정도 거리감이 있는 분야라고 생각을 했었다. 하지만 오늘 강의를 들어보니, 필자가 단순하게 생각했던 것과 달리, 데이터베이스는 고전적인 틀을 벗어나서 분산 클라우드 환경에서, query processing과 indexing에 있어서 machine learning과 접목되어 발전하기까지 했다. 처음엔 단순히 machine learning이란 큰 물결이기 때문에, 이것이 여러 방면에 연계된 것일까 싶었지만, 그보다는 바로 '데이터'라는 현대 사회에서 어마어마한 가치를 지닌 자원의 특성으로부터 생각치 못한 곳에서 연계가 일어난 것이 아닐까 싶다. 데이터란 자원을 기존에 존재하던 패러다임에서 어떤 식으로 새로이 처리할지, 언제 처리의 과정이 필요할지, 발생 예상되는 오류나 문제를 어떻게 해결할지 등 소프트웨어 종합적인 사고의 필요성도 느낀다.