

RDB 와 GDB 의 CRUD 및 분석 질의 처리에 대한 성능 비교 연구

윤현호¹, 박영호², 임선영³

¹배재대학교 컴퓨터공학과 학부생

²숙명여자대학교 인공지능공학부 교수

³배재대학교 컴퓨터공학과 조교수, 교신저자

dktlfhr123@naver.com, yhpark@sm.ac.kr, sunnyihm@pcu.ac.kr

Comparison of CRUD and Complex Query Processing Performance between RDB and GDB

Hyeon-Ho Yoon¹, Young-Ho Park², Sun-Young Ihm¹

¹Dept. of Computer Engineering, Pai Chai University

¹Dept. of AI Engineering, Sookmyung Women's University

요 약

빅데이터 시대가 오면서 정형화된 데이터뿐만 아니라 비정형 데이터들도 즐비하게 되었다. 데이터의 많은 비율을 비정형 데이터가 차지하게 됨으로써 관계형 데이터베이스로는 여러 이유로 인해 한계에 봉착하게 되었고 NoSQL 계열의 데이터베이스를 채택하기 시작했다. 본 연구에서는 그래프 데이터베이스는 노드와 관계를 이용하여 데이터를 조회하고 분석한다. 관계형 DB 인 MySQL 과 그래프 DB 인 Neo4j 를 사용하여 CRUD 및 분석 성능을 비교한다.

1. 서론

전세계가 발전함에 따라 데이터의 양은 늘어가고, 그 형태 또한 여러 형태로 나타나고 있다. 그에 따라 데이터베이스도 끊임없이 발전하고 있다. 그렇지만 데이터의 형태가 다양해지고 그 비율을 비정형 데이터가 상당수 차지함에 따라 관계형 데이터베이스로 수용할 수 있는 비정형 데이터에 대한 한계를 겪게 되었다. 비정형 데이터를 수용하려고 노력해왔지만 결국 NoSQL 계열의 데이터베이스가 나타나게 되었다. 관계형 데이터베이스라는 이름에 맞지 않게 관계에 대해 유연하게 대처하지 못하는 양상을 보여주었다. 관계형 데이터베이스에 비해 스키마로부터 자유롭고 구조적으로 유연하기 때문에 확장성이 뛰어나다. 뿐만 아니라 대용량 데이터를 처리하는데 효율적이다. NoSQL 계열의 데이터베이스는 Key-Value, Document, Graph 등이 존재하는데 본 연구에서는 그 중 그래프 데이터베이스인 Neo4j 를 선택해 관계형 데이터베이스 MySQL 과 CRUD 및 분석 질의 성능을 비교하고자 한다.

2. 그래프 데이터베이스

그래프 데이터베이스는 정형화된 규격에 따른 테이블을 가진 관계형 데이터베이스와 다르게 노드와 엣지로 이루어진 그래프 데이터 모델을 따르고 있다[1].

각 노드와 엣지에 데이터를 저장하면서 속성값을 부여하고 연결된 데이터를 기반으로 대량의 정보를 효율적으로 조회할 수 있다. 또한 관계형 DB 와는 다르게 데이터를 처리하기 때문에 직관적인 데이터 모델을 가진 그래프 DB 는 Index-free adjacency 라는 특징을 가지고 있어 복잡한 관계의 데이터를 처리하는데 효율적이다. 특히 재귀적인 관계를 보이는 모델에서 우수한 성능을 보인다. Index-free adjacency 는 서로 연결된 노드가 데이터베이스 내에서 서로 물리적으로 관계로 가리키고 있다는 것을 의미한다. 관계형 DB 와 다르게 더 적게 인덱스를 활용하고 데이터의 중복을 줄인 모습을 볼 수 있다. 그래프 DB 가 효율적이지만 새로운 데이터 모델링 방법이 나타났다고 해서 기존의 DB 를 대체할 수 있다는 정당성을 부여할 수 없기 때문에 확실한 이점이 존재해야 한다[1]. 그렇기에 기본적인 CRUD 와 다소 복잡한 질의를 사용해 성능 비교를 한다.

3. Neo4j

Neo4j 는 자바 기반의 네이티브 그래프 데이터베이스이며 그 중에도 가장 많이 사용되고 있다. Neo4j 는 Cypher 라는 선언적인 질의 언어를 사용한다. Cypher 는 화이트보드 친화적이며 그래프를 모델링하면서 Cypher 문이 자연스럽게 완성될 수 있다. 그리고 SQL

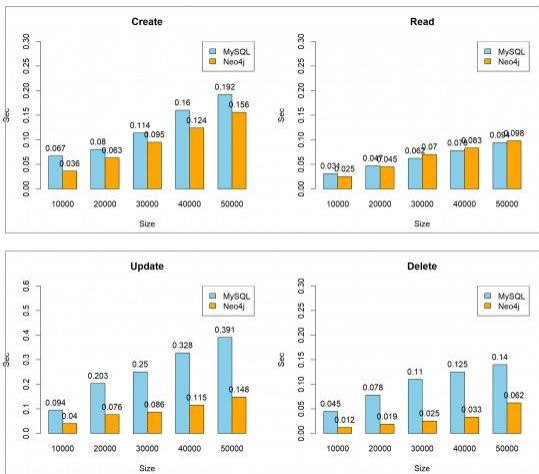
언어와 유사하지만, 그래프 기반 질의에 최적화되어 있으며 직관적이고 효율적이다[2]. 또한 Neo4j 는 ACID 트랜잭션을 지원하기 때문에 데이터를 사용하는데 안전한 환경을 제공한다.

4. 실험

성능 비교를 위한 데이터는 복잡한 관계에서 우수한 성능을 보여줄 수 있는 소셜 네트워크 데이터셋으로 성능 비교하였다. 데이터셋은 Twitter 의 사용자들과 사용자가 게시한 게시물, 게시물의 주소, 게시물에 태그된 해시태그 등 그것들의 관계를 나타내고 있는 엣지로 구성되어 있다. CRUD 및 분석 성능 비교 결과는 다음과 같다.

4.1. CRUD 성능 실험 결과

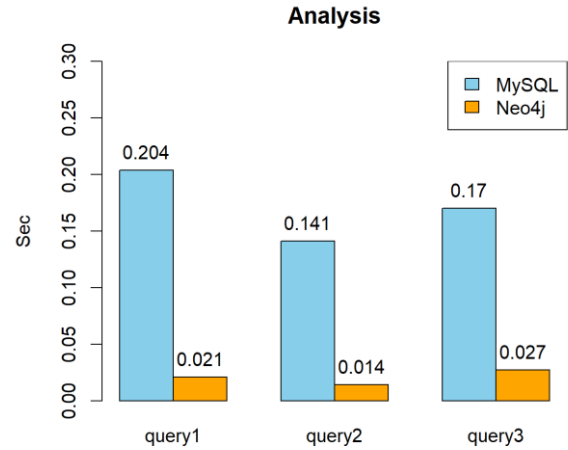
기본적인 CRUD 성능을 비교한 결과는 그림 1 과 같다. 데이터 조회 성능을 제외한 다른 성능들은 그래프 DB 가 우수한 양상을 보였으며, 두 DB 의 데이터 조회 성능이 비슷하다하더라도 단순한 수준의 쿼리문은 Cypher 문이 더욱 간결한 모습을 확인할 수 있다.



(그림 1) Neo4j 와 MySQL CRUD 성능 비교

4.2. 분석 질의 성능 실험 결과

그림 2 는 팔로워가 1 만이상인 유저가 게시한 게시물에 태그된 해시태그의 개수를 구하는 질의와 유저가 게시한 게시물이 재게시된 게시물의 수를 구하는 질의 등을 수행을 비교한 결과로 관계형 데이터베이스는 join 을 사용하며 복잡한 SQL 문으로 소셜 네트워크 데이터를 조회하였다. 반면에, 그래프 데이터베이스는 관계형 DB 와 질의가 같음에도 불구하고 간단하게 구성된 Cypher 문으로 상당한 차이를 보여줌을 확인할 수 있다.



(그림 2) Neo4j 와 MySQL 의 분석 질의 성능 비교

5. 결론

본 연구의 실험 결과 단순 CRUD 수준의 질의문에서는 데이터 삽입, 갱신, 삭제에서 성능 차이를 나타냈지만 데이터 조회에서 다소 비등한 모습을 나타냈다. 하지만 앞서 예상한 것과 같이, 관계형 DB 와 그래프 DB 의 성능 차이는 복잡한 관계를 지닌 데이터셋에서 큰 차이를 보여줌을 알 수 있었다. 소셜 네트워크 데이터 이외에도 그래프 모델링에 특화된 데이터셋을 이용해 데이터베이스 성능 및 기능을 테스트하는 연구를 진행할 예정이다.

사사

이 (성과)는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.NRF-2022R1F1A1074065). 본 과제(결과물)는 2024 년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과입니다. (2021RIS-004)

참고문헌

- [1] Ian Robinson, Jim Webber, Emil Eifrem, "Graph Databases: New Opportunities for Connected Data", O'Reilly Media, Inc., 2015.
- [2] <https://neo4j.com/>