2019 Database System

Project #3

이름 : 이승철

학과 : 컴퓨터공학과

학번 : 20141561

1. **MongoDB 질의문 및 결과**
2. 전체 기사의 개수를 구하시오.



1. 연합뉴스 사에 실린 기사들의 개수를 구하시오.



1. 2016년 6월 1일부터 2016년 6월 12일까지 실린 기사 개수를 구하시오.



1. 전체 기사를 날짜 별로 정렬해서 가장 빠른 날짜와 늦은 날짜의 기사 제목을 출력하시오.
   1. 가장 빠른 날짜 제목



* 1. 가장 늦은 날짜 제목



1. 연합뉴스 사에 실린 기사 혹은 2016년 6월 5일부터 2016년 6월 20일까지 실린 기사의 개수를 구하시오.



1. **RDB vs. NoSQL DB에 대한 비교**

* RDB는 Relational Database의 약자로 관계형 데이터베이스라고 불리운다. 대표적인 예시는 MySql, Oracle, PostgreSql 등이 가장 많이 알려져있다. 관계를 기반으로 한 DB개체의 내부 관계가 아닌 외부의 관계를 적용한 것으로 테이블과 테이블 사이의 관계가 맺어지는 DB형태이다. RDB를 사용하기 위해서는 테이블 마다 스키마를 정의해야하고 이 테이블에 SQL질의문을 통해서 요청을 처리를 하게 된다. 데이터 타입과 제약을 통해 데이터의 정확성을 보장을 해준다. 이런 데이터 타입과 제약이 있기 때문에 데이터베이스에 정보를 추가하기 전에 유효성 검사가 진행되므로 데이터 품질을 향상시킨다. 그리고 RDB의 성능을 높이기 위해서는 하드웨어적인 측면으로 접근을 해야되므로 고성능으로 교체를 하는 것이 필요하다. 따라서 고성능 하드웨어는 가격이 비싸기 때문에 RDB의 성능을 높이거나 확장하기 어렵기 때문에 확장성에는 좋지 않다. 그리고 이 RDB모델을 쓰기 위해서는 ER모델링이 개발 전에 완료 되어있어야 하므로 이는 개발을 느리게 만든다. 또한 테이블에 다양한 가변성이 있는 데이터를 저장하기가 어렵다.
* NoSQL의 대표적인 예는 mongoDB와 hBase 등이 있으며 mongoDB의 경우에는 문서형 데이터베이스이다. RDB에서 확장성의 문제가 있었는데 이 확장성 이슈를 해결하기 위해 나온 데이터베이스 모델이다. 분산 컴퓨팅 활용이 목적으로 이를 통해 비교적 저렴한 가격으로 DB성능을 RDB보다 더 잘 확장할 수 있다. 가장 많이 쓰이는 방식은 key-value방식으로 key-value를 통해 데이터를 관리한다. 구조 변경이 용이하고 데이터 형식이 다양하며 바꾸기 쉽고(Schema-less) 정확성 보다는 데이터의 양이 중요한 빅데이터와 실시간 웹 애플리케이션 같은 상업적 이용에 주로 된다. 하지만 정확성이 RDB에 비해서는 떨어지므로 금용이나 결제, 회원정보 같은 트랜잭션이 필요한 경우에는 부적합하다. Memory mapped file로 파일 엔진 DB이기 때문에 메모리 관리를 OS에 위임을 하게 된다. 이는 메모리에 의존적이라는 의미이고 메모리 크기가 NoSQL의 성능을 좌우하게 된다.
* 텍스트 마이닝에 있어서는 RDB보다는 NoSQL DB를 사용하는 것이 더 효율적이다. 이번 세번째 프로젝트는 전체 100개의 기사에 대해서 그 기사에 대한 텍스트를 읽은 다음에 단어 별로 데이터 베이스를 구축을 하고 이 데이터베이스를 통해서 기사의 내용을 형태소 분석 같은 전처리 하는 것을 목적으로 한다. 100개의 기사와 그 기사의 내용은 미리 정의된 데이터 모델이 없고 미리 정의된 방식으로 정리되지 않은 데이터이므로 이는 비정형 데이터로 볼 수 있다. 그리고 이러한 기사의 내용을 데이터 타입이나 제약을 통해서 정의내릴수도 없다. 따라서 이 데이터를 처리하는 데에는 데이터의 타입과 제약을 걸기에는 데이터의 양이 상당히 많고 매우 복잡해질 수 있다. 그리고 데이터를 저장을 하게 될 때 그 데이터의 양이 기하급수적으로 늘 때 이 데이터를 저장하고 읽는데 있어서 RDB는 앞에서 설명했듯이 확장성이 뛰어나지 않으므로 효율적으로 이 데이터를 처리하기 힘들다. 이러한 이유들 때문에 RDB보다 확장성이 뛰어나고 많은 양의 비정형 데이터를 처리하는데 편리한 NoSQL DB를 사용하는 것이 더 효율적이다.