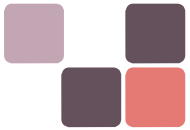
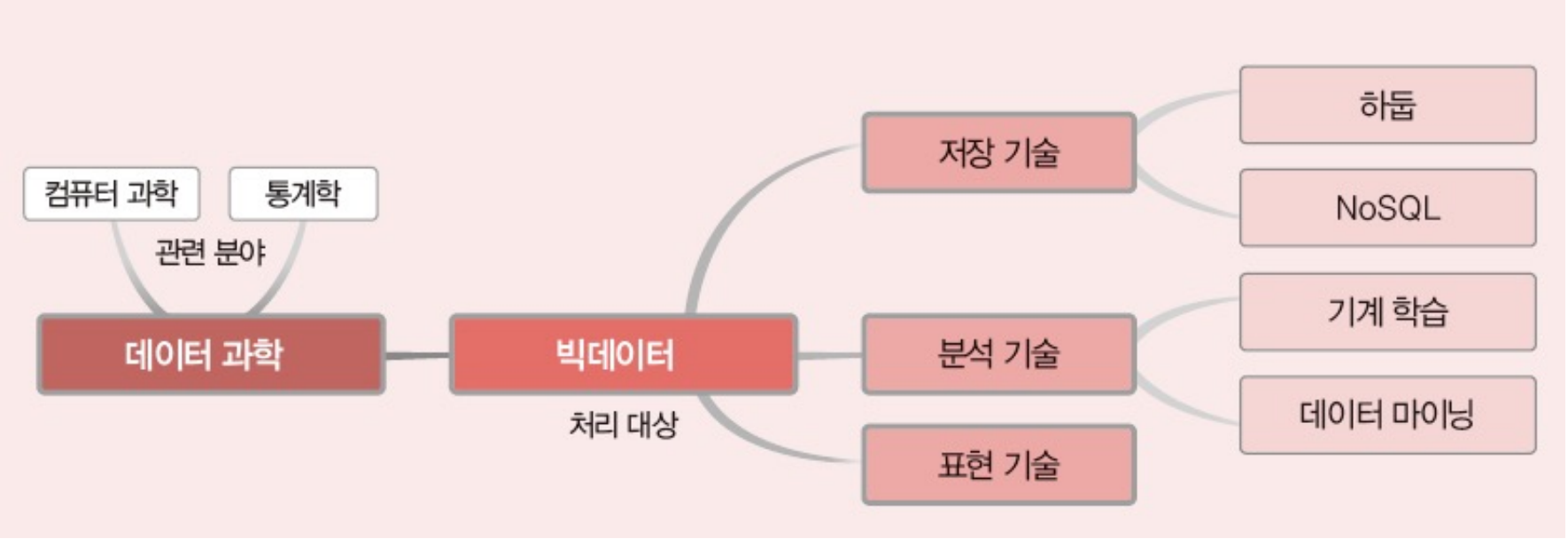


13장. 데이터 과학과 빅데이터

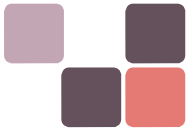
- 데이터 과학
- 빅데이터
- 빅데이터 저장 기술 : NoSQL
- 빅데이터 분석 기술 : 데이터 마이닝

학습목표



- 데이터 과학의 개념과 빅데이터와의 관련성을 이해한다.
- 빅데이터의 개념과 특징을 이해한다.
- 빅데이터의 저장 기술에 대해 살펴본다.
- 빅데이터의 분석 기술에 대해 살펴본다.

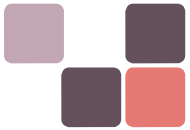
01 데이터 과학



❖ 데이터 과학(Data Science)의 필요성

- 4차 산업혁명 시대로의 진입
  - 빅데이터, 사물 인터넷, 인공지능 등 핵심 기술의 중심에 데이터가 있음
  - 21세기의 원유는 데이터
- 데이터의 방대한 규모와 다양한 형태
  - 전통적인 방식으로 수집하고 저장하는데 한계가 있음
- 다양해진 데이터 활용에 대한 요구
  - 단순히 데이터를 분류하고 검색하는 것을 넘어, 방대한 양의 데이터 속에 숨겨진 규칙과 패턴을 찾아내 문제 해결에 활용하고 미래의 일을 예측하여 미리 준비하기를 원함

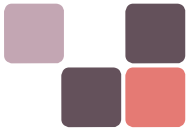
01 데이터 과학



❖ 데이터 과학의 개념과 목표

- 개념
  - 데이터를 수집한 후 분석을 통해 데이터를 정확히 이해함으로써 그 속에 숨겨진 새로운 지식을 발견하고, 이를 문제 해결에 활용하는 모든 과정의 활동을 의미
    - 데이터 생성, 수집, 저장, 분석, 표현의 모든 과정과 연관됨
  - 활동을 지원하는 수단이나 기술도 포함
- 목표
  - 수집된 데이터로부터 가공된 정보를 거쳐 지식과 지혜를 추출하는 것

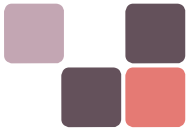
01 데이터 과학



❖ 데이터 과학의 적용 예

- 게임 회사에서 동시 접속자 수, 아이템 구매 정보 등을 분석하여 마케팅과 새로운 게임 개발에 활용
- 선거 전략을 세우고 당선자를 예측
- 선수, 경기장, 날씨 등의 데이터를 분석하여 경기 결과를 예측

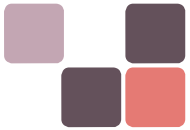
01 데이터 과학



❖ DIKW(Data-Information-Knowledge-Wisdom) 계층 구조

- 데이터(data)
  - 관찰하거나 측정하여 수집한 사실이나 값
  - 예) 출판사에서 3년간 1월부터 12월까지 매달 책의 판매량을 조사한 결과
- 정보(information)
  - 상황에 대한 이해를 바탕으로 데이터를 목적에 맞게 가공한 것
  - 예) 연간 분기별 책 판매량의 합계를 계산한 것

01 데이터 과학



❖ DIKW(Data-Information-Knowledge-Wisdom) 계층 구조

- 지식(knowledge)
  - 규칙이나 패턴을 통해 찾아낸 의미 있고 유용한 정보
  - 예) 연간 분기별 책 판매량을 분석하여 3분기에 책의 판매량이 증가하는 규칙과 그 원인을 찾아낸 것
- 지혜(wisdom)
  - 지식에 통찰력을 더해 새롭고 창의적인 아이디어를 도출한 것
  - 예) 찾아낸 지식을 토대로 내년 3분기에 새로 출간할 책의 콘텐츠를 기획하고 적합한 홍보 전략을 세우는 것

01 데이터 과학

❖ DIKW 계층 구조

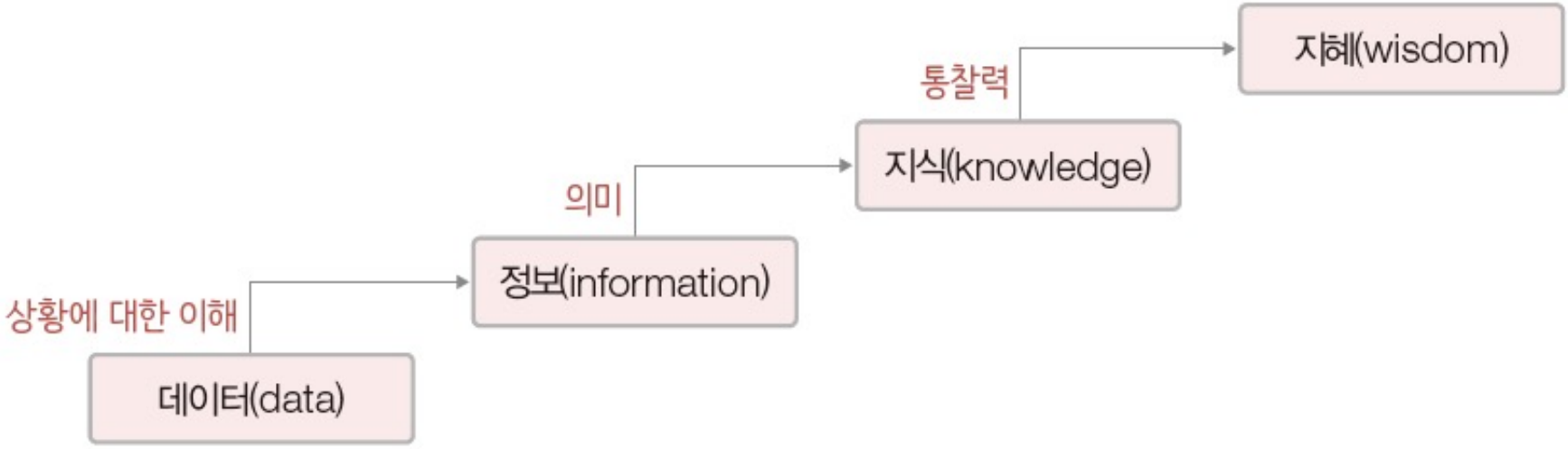


그림 13-1 데이터 과학의 계층 구조

01 데이터 과학

❖ 데이터 과학의 특징

- 컴퓨터 과학, 통계학, 적용 분야에 대한 이해를 필요로 하는 복합적인 기술

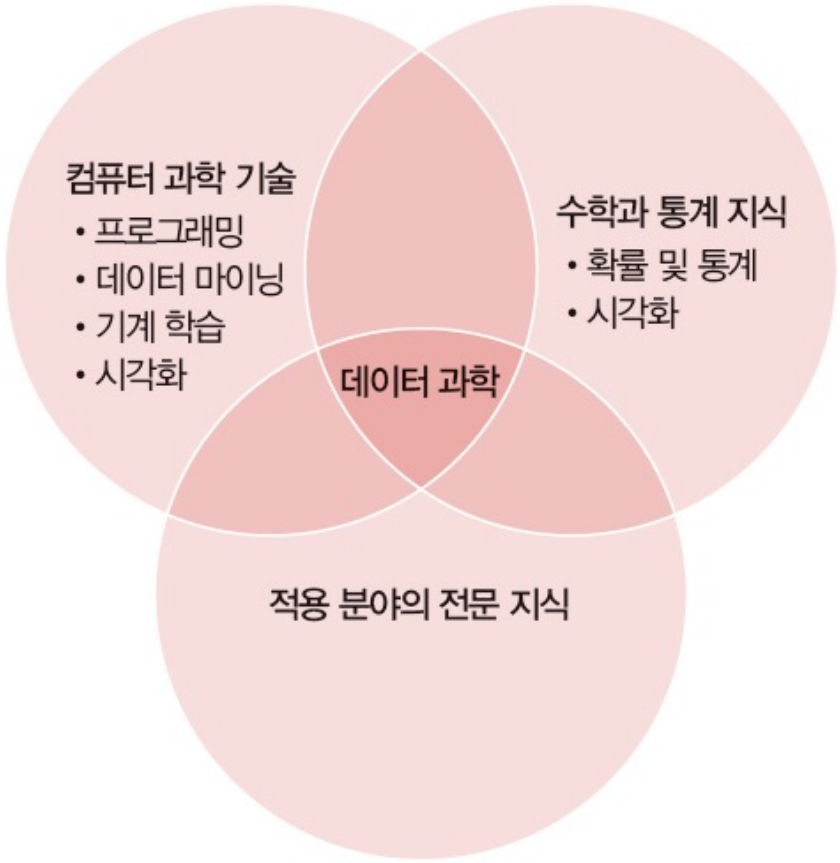
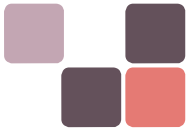


그림 13-2 데이터 과학의 특징

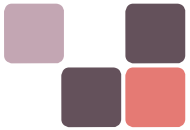
02 빅데이터



❖ 빅데이터(big data)의 개념

- 좁은 정의
  - 기존 데이터베이스가 저장하고 관리할 수 있는 범위를 넘어서는 대규모의 다양한 데이터
- 넓은 정의
  - 대규모 데이터를 저장 및 관리하는 기술과 가치 있는 정보를 만들기 위해 분석하는 기술까지 포함

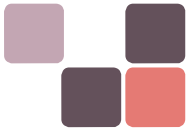
02 빅데이터



❖ 빅데이터 활용 사례

- 아마존 닷컴
  - 빅데이터 기술로 상품 구매 내역을 저장하고 분석하여 고객의 소비 성향을 파악하고 그 정보를 활용해 고객이 관심을 가질 만한 상품의 소개 메일을 전송하거나 로그인 시 자동으로 제시
- 구글
  - 빅데이터 기술을 활용해 사용자의 개인 정보와 사용자가 입력한 검색 조건 등을 분석하여 사용자에게 맞춤형 광고 제시
- 페이스북
  - 빅데이터 기술을 활용해 사용자가 작성한 글과 사진, 동영상 데이터를 분석하여 사용자에게 맞춤형 광고 제시

02 빅데이터



❖ 빅데이터 활용 사례

- 정치 분야
  - 국내에서 여론조사 기관들이 투표 결과를 더 정확히 예측하기 위해 SNS를 통해 생성된 선거 관련 데이터를 빅데이터 기술을 활용해 분석
  - 미국에서 대통령 선거를 위해 다양한 경로로 수집한 유권자의 데이터를 빅데이터 기술을 활용해 분석하여 성향을 파악하고 선거 전략을 수립

02 빅데이터

❖ 빅데이터의 특징 : 3V

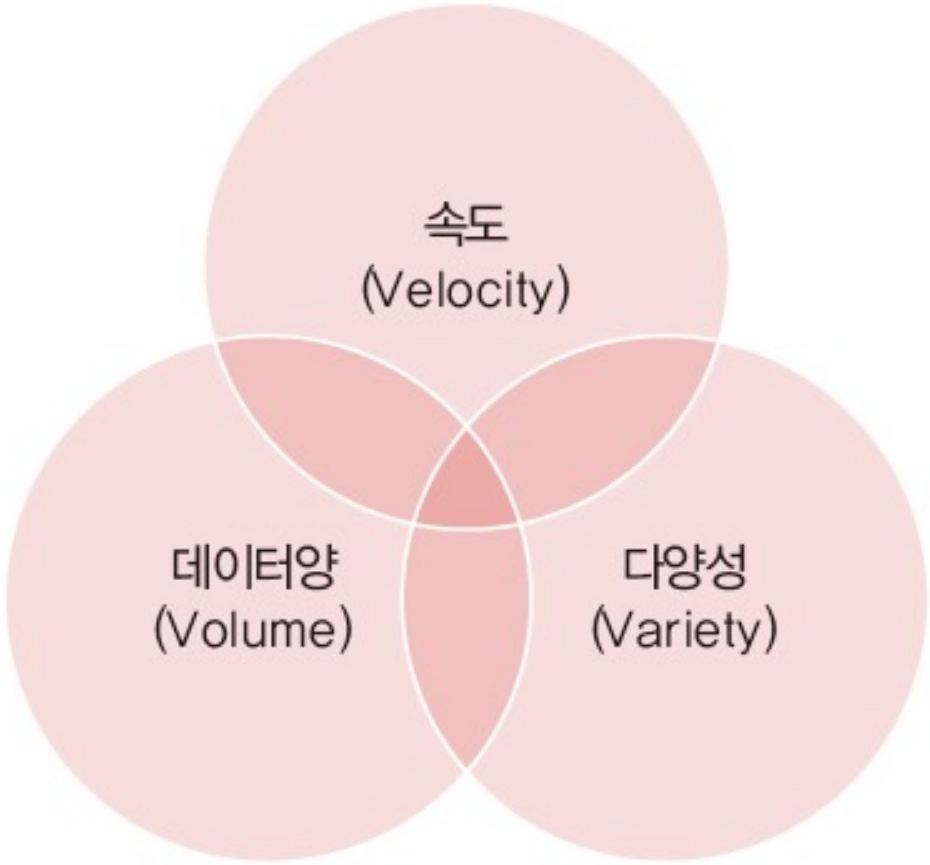
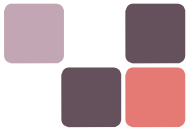


그림 13-3 빅데이터의 특징

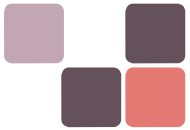
02 빅데이터



❖ 빅데이터의 특징 : 3V

- 데이터양(Volume)
  - 테라바이트(TB) 단위 이상의 대량 데이터
  - 여러 경로를 통해 계속 생성되고 있는 많은 양의 데이터를 의미
- 속도(Velocity)
  - 데이터의 수집과 분석을 정해진 시간 내에 처리해야 함
  - 많은 양의 데이터가 생성되고 전달되는 속도가 빠르므로 수집 및 분석 작업도 실시간으로 진행되어야 함

02 빅데이터



❖ 빅데이터의 특징 : 3V

- 다양성(Variety)
  - 형태의 다양성이 존재
  - 정형, 반정형, 비정형 같은 다양한 형태의 데이터를 모두 포함
    - 정형 데이터 : 관계 데이터베이스와 같이 정형화된 시스템에 저장된 데이터 형태
    - 반정형 데이터 : 정형화된 시스템에 저장되어 있지 않지만 내부적으로 스키마를 어느 정도 포함하고 있는 XML, HTML 등을 의미
    - 비정형 데이터 : 구조가 정해져 있지 않은 데이터
      - » 예) 책, 잡지, 의료 기록, 비디오, 오디오 같은 전통적인 비정형 데이터
      - » 예) 위치 정보, 이메일, SNS 등에서 생성되는 비정형 데이터

02 빅데이터

❖ 빅데이터의 유형

- 빅데이터를 양적 측면의 대규모 데이터를 넘어서 질적 측면의 다양한 형태를 포함하는 대규모 데이터로 이해해야 함

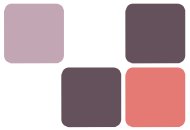
그림 13-4 빅데이터의 유형

02 빅데이터

❖ 빅데이터의 기술

그림 13-5 빅데이터의 기술

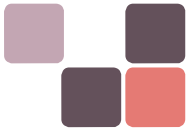
02 빅데이터



❖ 빅데이터의 기술 – 저장 기술

- 하둡(Hadoop)
  - 대용량 데이터를 분산 처리할 수 있는 자바 기반의 오픈 소스 프레임워크
  - 분산 파일 시스템인 HDFS(Hadoop Distributed File System)에 데이터를 저장하고, 분산 처리 시스템인 맵리듀스(MapReduce)를 이용해 데이터를 처리
  - 오픈 소스이기 때문에 기존 데이터베이스 시스템보다 비용이 적게 들고, 여러 대의 서버에 데이터를 분산해서 저장해두기 때문에 처리 속도가 빠름

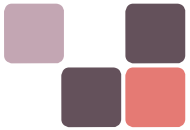
02 빅데이터



❖ 빅데이터의 기술 – 저장 기술

- NoSQL
  - 관계 데이터 모델과 SQL을 사용하지 않는 데이터베이스 시스템
  - 일관성보다는 가용성과 확장성에 중점을 두고 있음
  - 비정형 데이터의 저장을 위해 유연한 데이터 모델을 지원하고, 관계 데이터베이스와 동일한 데이터 처리가 가능하면서도 더 저렴한 비용으로 분산 처리와 병렬 처리가 가능
  - 예) H베이스, 카산드라, 몽고DB, 카우치DB 등

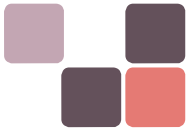
02 빅데이터



❖ 빅데이터의 기술 – 분석 기술

- 텍스트 마이닝(text mining)
  - 반정형 또는 비정형 텍스트에서 자연어 처리 기술로 정보를 추출하고 가공함
- 오피니언 마이닝(opinion mining)
  - SNS, 블로그, 게시판 등에 기록된 사용자들의 의견을 수집하고 분석하여, 제품이나 서비스에 대한 긍정, 부정, 중립 등의 선호도를 추출
- 소셜 네트워크 분석(social network analysis)
  - 소셜 네트워크의 연결 구조나 강도 등을 바탕으로 소셜 네트워크에 나타난 영향력, 관심사, 성향, 행동 패턴 등을 추출
- 군집 분석(cluster analysis)
  - 데이터 간의 유사도를 측정한 후 이를 바탕으로 특성이 비슷한 데이터를 합쳐가면서 최종적으로 유사 특성의 데이터 집합을 추출

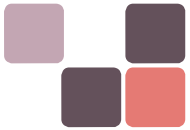
02 빅데이터



❖ 빅데이터의 기술 – 표현 기술

- R 언어
  - 데이터 분석을 통해 추출한 의미와 가치를 시각적으로 표현하기 위해 사용
  - 기본 통계 기법부터 최신 데이터 마이닝 기법까지 구현이 가능
  - 다양한 프로그래밍 언어와 연동 가능하고 다양한 운영체제를 지원하며, 하둡 환경에서 분산 처리를 지원하는 라이브러리를 제공
  - R : 통계 계산과 다양한 시각화를 위한 언어와 개발 환경을 제공하는 오픈 소스

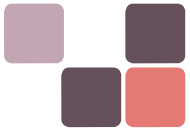
02 빅데이터



❖ 이전 데이터 vs. 빅 데이터

구분	빅데이터 이전의 데이터	빅데이터
데이터 유형	정형화된 문자, 수치 데이터 중심	정형, 반정형, 비정형 데이터 모두 포함
관련 기술	• 관계 데이터베이스 • SAS, SPSS와 같은 통계 패키지 • 데이터 마이닝 • 기계 학습	• 저장 기술 : 하둡, NoSQL • 분석 기술 : 텍스트 마이닝, 오피니언 마이닝, 소셜 네트워크 분석, 군집 분석 • 표현 기술 : R 언어
저장 장치	데이터베이스나 데이터 웨어하우스와 같은 고가의 저장 장치	비용이 저렴한 클라우드 컴퓨팅 장비 활용 가능

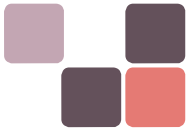
03 빅데이터 저장 기술 : NoSQL



❖ NoSQL의 등장 배경

- 관계 데이터베이스를 대신할 새로운 대안의 필요성
  - 정형화된 데이터를 주로 처리하는 관계 데이터베이스는 빠른 속도로 대량 생산되는 다양한 유형의 비정형 데이터를 저장 및 관리하는데 적합하지 않음
  - 단일 컴퓨터 환경에서 주로 사용되는 관계 데이터베이스는 여러 컴퓨터가 연결되어 하나의 시스템을 구성하는 클러스터 환경에는 확장성 측면에서 비효율적임
- 새로운 대안으로 NoSQL 등장
  - 관계 데이터베이스만 고집하지 말고 필요에 따라 다른 특성을 제공하는 데이터베이스를 사용하는 것이 좋다는 의미로 이해

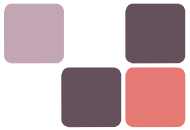
03 빅데이터 저장 기술 : NoSQL



❖ NoSQL(Not Only SQL)

- 의미
  - 빠른 속도로 생성되는 대량의 비정형 데이터를 저장하고 처리하기 위해 ACID(원자성, 일관성, 격리성, 지속성)를 위한 트랜잭션 기능을 제공하지 않는 대신 저렴한 비용으로 여러 대의 컴퓨터에 데이터를 분산·저장·처리하는 것이 가능한 데이터베이스
- 특징
  - 관계 모델보다 더 융통성 있는 데이터 모델을 사용
  - 스키마 없이 동작하기 때문에 데이터 구조를 미리 정의할 필요가 없고 수시로 그 구조를 바꿀 수 있어 비정형 데이터를 저장하기에 적합
  - 대부분 오픈 소스로 제공

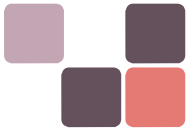
03 빅데이터 저장 기술 : NoSQL



❖ 관계 데이터베이스 vs NoSQL

- 관계 데이터베이스
  - 장점
    - 트랜잭션을 통해 일관성을 유지하고, 외래키로 테이블 간의 관계를 표현함으로써 조인과 같은 복잡한 질의 처리가 가능
  - 단점
    - 빠른 속도로 증가하는 대량의 비정형 데이터를 저장하는데 확장성 측면에서 비효율적
- NoSQL
  - 장점
    - 트랜잭션 기능을 제공하지 않고 정해진 스키마도 없기 때문에 자유롭게 구조를 바꾸며 대량의 비정형 데이터를 빠르게 저장하고 처리할 수 있음
  - 단점
    - SQL 대신 별도의 분석 기술을 이용해 데이터 속에 숨겨진 의미를 찾아내야 함

03 빅데이터 저장 기술 : NoSQL

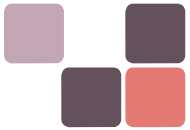


❖ 관계 데이터베이스 vs NoSQL

표 13-1 관계 데이터베이스와 NoSQL의 비교

구분	관계 데이터베이스	NoSQL
처리 데이터	정형 데이터	정형 데이터, 비정형(반정형 포함) 데이터
대용량 데이터	대용량 처리 시 성능 저하	대용량 데이터 처리 지원
스키마	미리 정해진 스키마가 존재	스키마가 없거나 변경이 자유로움
트랜잭션	트랜잭션을 통해 일관성 유지를 보장함	트랜잭션을 지원하지 않아 일관성 유지를 보장하기 어려움
검색 기능	조인 등의 복잡한 검색 기능 제공	단순한 데이터 검색 기능 제공
확장성	클러스터 환경에 적합하지 않음	클러스터 환경에 적합함
라이선스	고가의 라이선스 비용	오픈 소스
대표적 사례	Oracle, MySQL, MS SQL 서버 등	카산드라, 몽고DB, H베이스 등

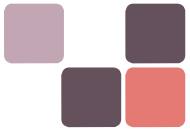
03 빅데이터 저장 기술 : NoSQL



❖ 관계 데이터베이스 vs NoSQL

- NoSQL은 관계 데이터베이스의 경쟁자가 아니다!
  - 관계 데이터베이스가 적합하지 않은 새로운 환경에서 선택의 폭을 넓히기 위한 대안으로 NoSQL을 이해
  - 저장될 데이터의 형태와 처리 목적에 더 적합한 것을 선택
    - 예) 기업의 인사, 회계 자료와 같이 일관성이 중요하고 조인과 같은 복잡한 질의 처리가 필요한 정형화된 데이터를 관리하는 용도 → 관계 데이터베이스를 선택
    - 예) SNS에 게시된 이미지와 텍스트, CCTV 촬영 영상, 센싱 데이터와 같이 빠른 속도로 엄청난 양이 생성되지만 수정보다는 삽입 연산 위주의 데이터를 저장하고 관리하는 용도 → NoSQL을 선택

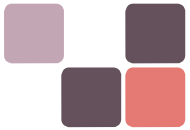
03 빅데이터 저장 기술 : NoSQL



❖ NoSQL의 종류

- 어떤 데이터 모델로 데이터를 저장하는지에 따라 4가지로 분류
  - 키-값(key-value) 데이터베이스
  - 문서 기반(document-based) 데이터베이스
  - 컬럼 기반(column-based) 데이터베이스
  - 그래프 기반(graph-based) 데이터베이스

03 빅데이터 저장 기술 : NoSQL



❖ NoSQL의 종류 : 키-값 데이터베이스

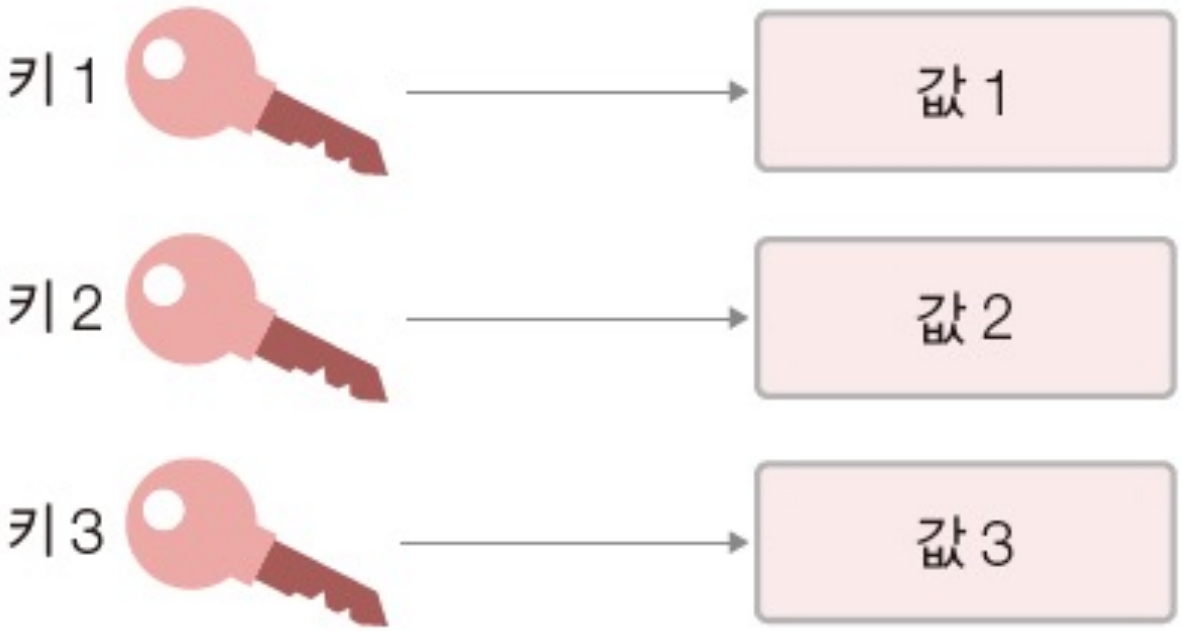
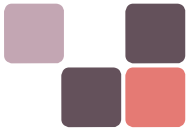


그림 13-6 키-값 데이터베이스 저장 구조

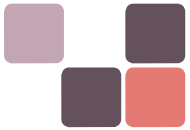
03 빅데이터 저장 기술 : NoSQL



❖ NoSQL의 종류 : 키-값 데이터베이스

- 특징
  - 키와 값의 쌍으로 데이터가 저장됨
  - 가장 단순한 형태
  - 이미지와 동영상은 물론 어떠한 형태의 값도 저장 가능
  - 질의 처리 속도 빠름
  - 키를 이용해 값 전체를 검색할 수는 있지만, 값의 일부를 검색하거나 값의 내용을 이용한 질의는 할 수 없고 별도의 처리가 필요함
- 대표적인 예
  - 아마존의 다이나모DB(DynamoDB), 트위터 등에서 사용되는 레디스(Redis) 등

03 빅데이터 저장 기술 : NoSQL



❖ NoSQL의 종류 : 문서 기반 데이터베이스

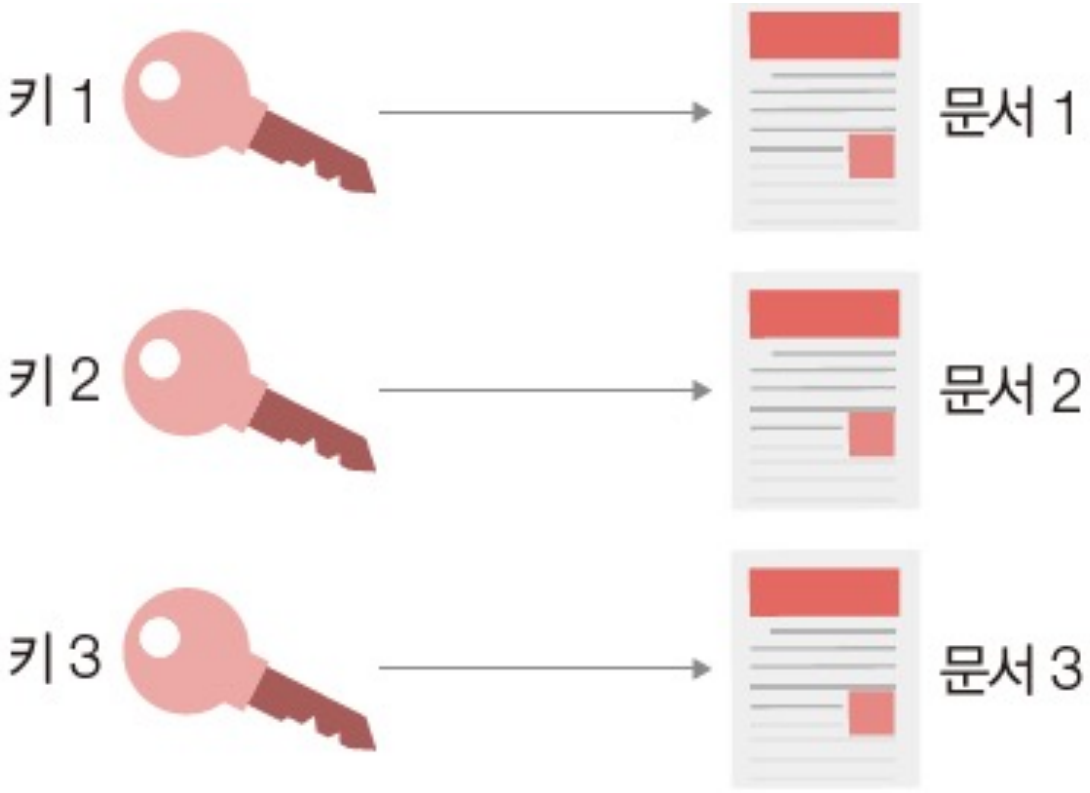
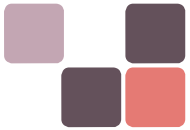


그림 13-7 문서 기반 데이터베이스 저장 구조

03 빅데이터 저장 기술 : NoSQL



❖ NoSQL의 종류 : 문서 기반 데이터베이스

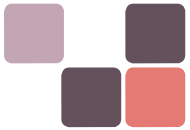
- 특징
  - 키와 문서의 쌍으로 데이터를 저장
    - 트리 형태의 계층적 구조가 존재하는 JSON, XML 등과 같은 반정형 형태의 문서로 데이터를 저장
    - 문서는 객체지향에서 객체의 개념과 유사
  - 키- 값 데이터 모델이 확장된 형태
  - 문서 전체를 검색하는 것도 가능하지만, XQuery와 같은 특별한 문서 대상 질의 언어를 이용하면 문서 내의 일부를 검색할 수도 있음
- 대표적인 예
  - 몽고DB(MongoDB), 카우치DB(CouchDB) 등

03 빅데이터 저장 기술 : NoSQL

❖ NoSQL의 종류 : 컬럼 기반 데이터베이스

그림 13-8 컬럼 기반 데이터베이스 저장 구조

03 빅데이터 저장 기술 : NoSQL



❖ NoSQL의 종류 : 컬럼 기반 데이터베이스

- 특징
  - 컬럼 패밀리(column family)와 키의 쌍으로 데이터를 저장
    - 컬럼 패밀리는 관련 있는 컬럼 값들을 모아서 구성함
  - 관계 데이터 모델의 테이블과의 유사성
    - 컬럼 패밀리는 테이블에서 한 개의 투플(행)을 구성하는 속성들의 모임으로 볼 수 있음
    - 키가 각 투플을 구분하는 것처럼 키로 각 컬럼 패밀리를 식별함
  - 관계 데이터 모델의 테이블과의 차별성
    - 다양한 형태의 데이터를 값으로 저장할 수 있음
    - 컬럼 패밀리마다 컬럼의 구성을 다르게 할 수 있음
- 대표적인 예
  - 구글의 빅테이블(BigTable), H베이스(HBase), 카산드라(Cassandra) 등

03 빅데이터 저장 기술 : NoSQL

❖ NoSQL의 종류 : 그래프 기반 데이터베이스

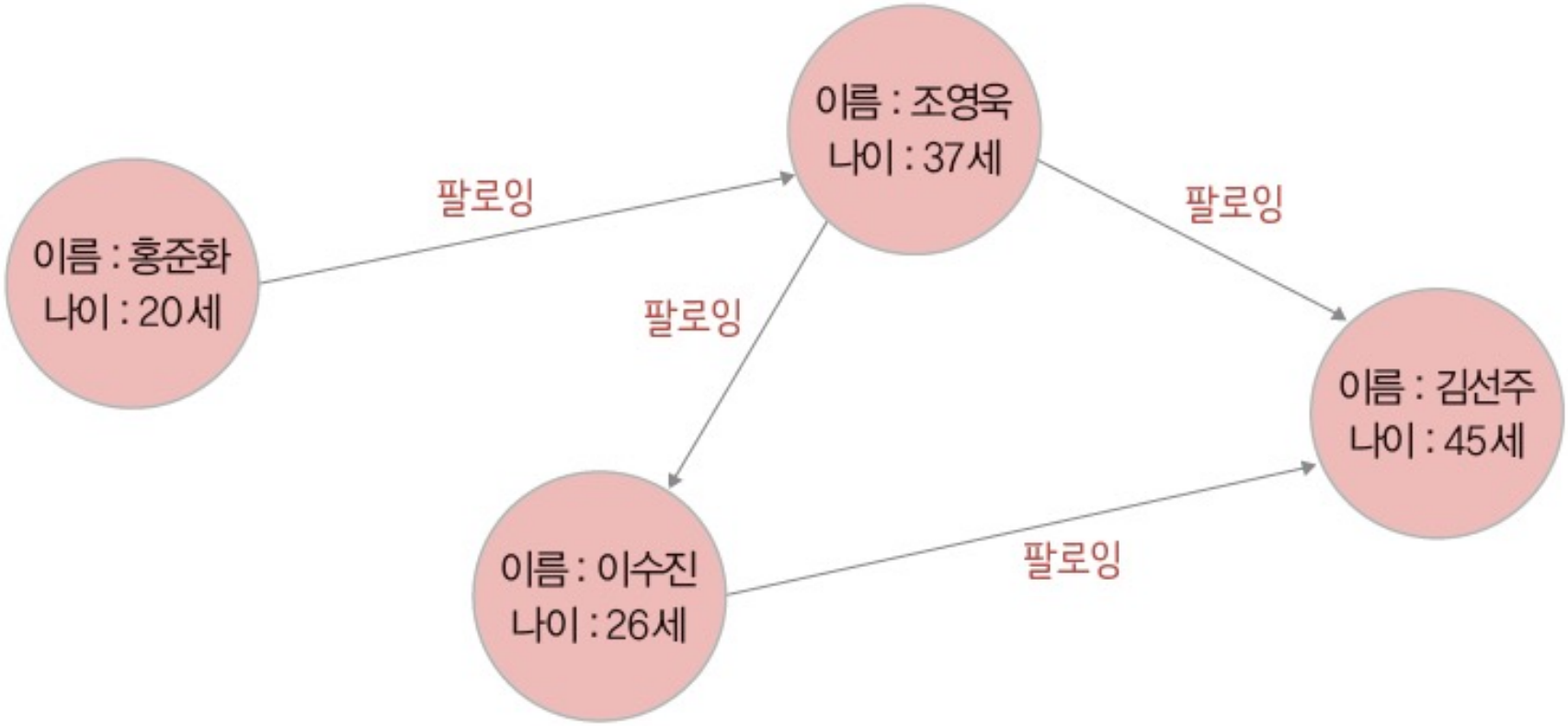
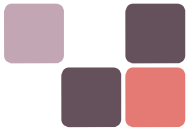


그림 13-9 그래프 기반 데이터베이스 저장 구조

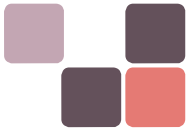
03 빅데이터 저장 기술 : NoSQL



❖ NoSQL의 종류 : 그래프 기반 데이터베이스

- 특징
  - 노드에 데이터를 저장하고 간선으로 데이터 간의 관계를 표현하는 그래프 형태
  - 질의는 그래프 순회 과정을 통해 처리
  - 연관 데이터를 추천하거나 소셜 네트워크에서 친구 찾기를 수행하는데 적합
  - 트랜잭션을 통해 ACID를 지원하며, 클러스터 환경에는 적합하지 않음
    - 다른 NoSQL 데이터 모델과의 차이점
- 대표적인 예
  - 네오포제이(Neo4J), 오리엔트DB(OrientDB) 등

04 빅데이터 분석 기술 : 데이터 마이닝



❖ 데이터 분석 기술

- 데이터 안에 숨겨진 유용한 정보, 즉 지식을 찾아내기 위해 데이터를 가공하는 역할을 담당
- 데이터베이스에서 SQL 문을 통해 자신이 원하는 데이터를 추출해서 분석하는 것도 해당됨

❖ 빅데이터 분석 기술

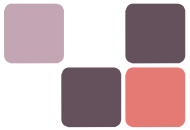
- 기존 데이터 분석 기술 + 빅데이터의 특징
  - 다양한 형태의 비정형 데이터를 기반으로 엄청난 양의 데이터를 처리
- 대표적인 기술
  - 데이터 마이닝(data mining), 기계 학습(machine learning)

04 빅데이터 분석 기술 : 데이터 마이닝

❖ 데이터 마이닝 vs 기계 학습

그림 13-10 데이터 마이닝과 기계 학습

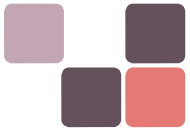
04 빅데이터 분석 기술 : 데이터 마이닝



❖ 데이터 마이닝 vs 기계 학습

- 분석 목적이 발견 → 데이터 마이닝
  - 수집된 데이터에서 숨겨진 규칙과 패턴을 찾아 가치 있는 유용한 정보인 지식을 발견하는 것
- 분석 목적이 예측 → 기계 학습
  - 수집된 데이터로 프로그램을 학습시켜서 유사한 상황의 새로운 데이터가 입력되었을 때 결과를 예측하는 것
- 각자의 목적을 위해 서로의 기법을 활용

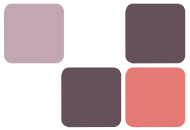
04 빅데이터 분석 기술 : 데이터 마이닝



❖ 데이터 마이닝

- 개념
  - 대량의 데이터 안에 숨겨진 지식을 발견하기 위해 규칙과 패턴을 찾아내는 기술
  - 고객의 성향을 파악해 판매 전략을 세우거나, 개인의 신용 등급을 판단하거나, 불량품이 발생하는 원인을 파악하고 개선하는 등 다양한 분야에서 활용됨
  - '데이터베이스에서의 지식 발견' 으로 시작했지만 빅데이터를 대상으로 하는 데이터 분석 기술로 영역을 넓히고 있음
- 대표적인 분석 기법
  - 분류 분석(classification analysis)
  - 군집 분석(cluster analysis)
  - 연관 분석(association analysis)

04 빅데이터 분석 기술 : 데이터 마이닝



❖ 대표적인 분석 기법 : 분류 분석

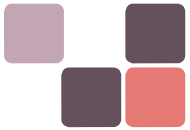
- 새로운 데이터가 어떤 그룹 또는 등급에 속하는지 예측하는데 주로 사용
- 미리 정의된 기준에 따라 기존 데이터의 그룹이 나뉘어 있음
  - 군집 분석과의 차이점
- 예) 의사가 기존 환자들의 데이터를 토대로 새로운 환자의 증상을 듣고 병명을 진단하는 것
- 주로 사용되는 방법
  - 로지스틱 회귀모형, 의사결정나무, K-최근접 이웃모형, 베이즈분류모형, 인공신경망, 지지벡터기계, 유전 알고리즘 등

04 빅데이터 분석 기술 : 데이터 마이닝

❖ 대표적인 분석 기법 : 분류 분석

그림 13-11 분류 분석의 의미

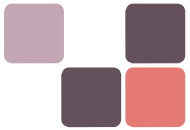
04 빅데이터 분석 기술 : 데이터 마이닝



❖ 대표적인 분석 기법 : 군집 분석

- 미리 정해진 기준이 없는 상태에서 유사한 특성을 공유하는 데이터들을 여러 개의 독립적인 군집으로 나누는 것
  - 군집의 개수나 형태를 미리 가정하지 않은 상태에서 데이터간의 유사성에 기반을 두고 거리가 가까운 데이터들을 하나의 군집으로 모음
  - 형성된 군집들의 특성을 파악하여 군집들 사이의 관계를 분석
- 예) 성격적 특징에 따라 심리학적으로 유사한 사람들의 집단을 나누는 것
- 계층적 군집 분석과 비계층적 군집 분석이 있음

04 빅데이터 분석 기술 : 데이터 마이닝



❖ 대표적인 분석 기법 : 군집 분석

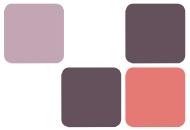
- 계층적 군집 분석
  - 가장 유사한 데이터를 묶어 나가는 과정을 반복하면서 원하는 개수의 군집을 형성하는 방법
  - 거리를 정의하는 방법에 따라 최단 연결법, 최장 연결법, 평균 연결법, 중심 연결법, 와드 연결법 등으로 세분화 됨
- 비계층적 군집 분석
  - 데이터를 군집으로 나눌 수 있는 모든 방법을 생각한 후 가장 최적화된 군집을 형성하는 방법
  - 대표적으로 K-중심 군집이 사용됨

04 빅데이터 분석 기술 : 데이터 마이닝

❖ 대표적인 분석 기법 : 군집 분석

그림 13-12 군집 분석의 의미

04 빅데이터 분석 기술 : 데이터 마이닝



❖ 대표적인 분석 기법 : 연관 분석

- 데이터 간의 발생 빈도를 분석하여 그 속에 숨겨진 연관 규칙(association rule)을 파악하는 방법
  - 연관 규칙을 평가하기 위해 지지도, 상품 신뢰도, 향상도 지표를 이용
- 상품이나 서비스 간의 연관 관계를 분석하여 마케팅에 주로 활용
- 장바구니 분석(market basket analysis)이라고도 함
- 예) 동시 구매가 자주 발생하는 상품들을 파악하여 해당 상품들을 묶음으로 판매하거나 인접한 진열대에 두어 매출을 올림
- Apriori 알고리즘이 대표적임

04 빅데이터 분석 기술 : 데이터 마이닝

❖ 대표적인 분석 기법 : 연관 분석

그림 13-13 연관 분석의 의미

Thank You