

실무에 적용 가능한 Big Data 분석 개론

빅데이터 플랫폼 개요



한국기술교육대학교
온라인평생교육원

■ 빅데이터 플랫폼의 개념과 진화

1. 빅데이터 플랫폼의 개념

빅데이터 플랫폼



빅데이터 기술을 잘 사용할 수 있도록 준비된 환경



빅데이터 플랫폼 구축

➤ 하드웨어 인프라 구축



빅데이터 수용 용량 및 처리, 분석 작업에 대한 부하 등을 감안

➤ 소프트웨어 구축



분석에 필요한 수집, 관리, 분석, 이용자 환경 등에 관련한 소프트웨어를 구축



부하



• 시스템에서 원하는 어떤 효과를 얻기 위해 취하는 행동에 필요한 동작이나 자원



• 컴퓨팅 부하



➤ 빅데이터를 처리하려고 CPU 등 자원소비에서 발생함



• 저장 부하



➤ 빅데이터 처리과정의 입력 데이터, 중간 데이터, 출력 데이터 등 저장에서 발생함



• 네트워크 부하



➤ 빅데이터 처리에서 사용하는 컴퓨터 노드간의 통신에서 발생함

부하 제어에 따라 빅데이터 성능이 좌우됨

■ 빅데이터 플랫폼의 개념과 진화

1. 빅데이터 플랫폼의 개념

빅데이터 플랫폼 설계 시 고려사항

컴퓨팅 부하

빅데이터를 처리하기 위해 CPU 등 자원 소비에서 발생

플랫폼 설계

- CPU 성능을 높임
- 클러스터 노드 작업을 처리함

용도에 따른 플랫폼

- 과거 발생 이력을 분석하는 용도,
- 현 상황을 분석하는 용도
- 미래를 예측, 대응하는 용도

컴퓨팅 부하 처리를 효과적으로 수행하는
빅데이터 플랫폼은 용도에 따라 다양하게 설계할 수 있음

■ 빅데이터 플랫폼의 개념과 진화

1. 빅데이터 플랫폼의 개념

빅데이터 플랫폼 설계 시 고려사항

빅데이터 분석을 수행하는 대상 영역에 따른 고려사항

빅데이터
소스

빅데이터
분석 방법

빅데이터
분석 수준

빅데이터
분석 목적

빅데이터 분석 결과의 활용 목적

누적된 데이터

과거 이력을 분석하기 위한 분석 기술 및 플랫폼 구성

현 상황을 분석하기 위한 분석 기술 및 플랫폼 구성

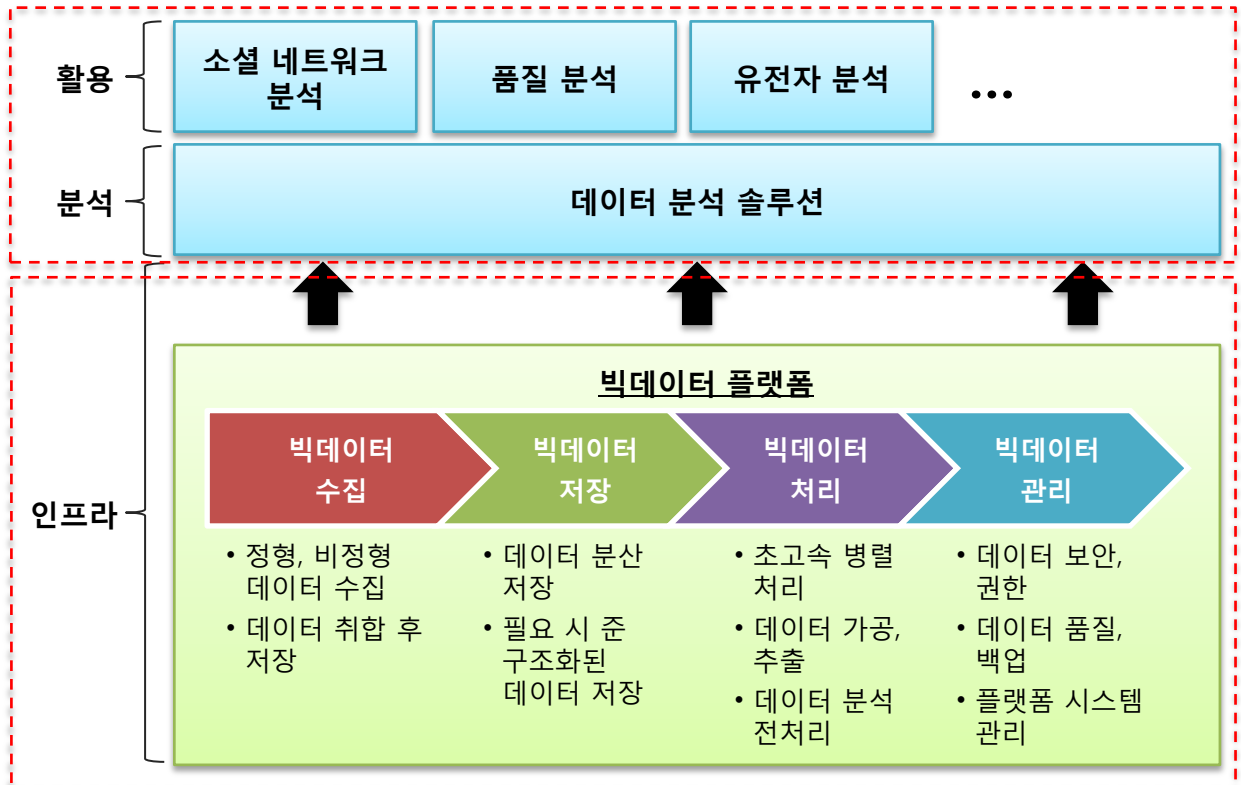
미래 예측 및 대응하기 위한 분석 기술 및 플랫폼 구성

빅데이터 분석 결과의 활용 목적에 따라서 분석 방법과 기술 및 플랫폼 구성은 달라질 수 있음

■ 빅데이터 플랫폼의 개념과 진화

1. 빅데이터 플랫폼의 개념

빅데이터 플랫폼의 역할과 기능



<출처 : '빅데이터' 플랫폼의 미래', Technology Inside LG CNS R&D Journal, 이주열, 2013>

■ 빅데이터 플랫폼의 개념과 진화

2. 빅데이터 플랫폼의 진화

실시간 빅데이터 처리 가능

다양한 분산병렬 처리 방법 제공

관계형 데이터 모델과 대규모 업무 트랜잭션 지원

파일 관리 효율화

실시간 빅데이터 처리 가능

- 기존의 일괄처리 방식이 아님
- 실시간 처리 기술 발전
 - 온라인으로 연결된 상태에서 빅데이터 처리 요청과 응답을 즉각적으로 주고 받을 수 있음

다양한 분산병렬 처리 방법 제공

- 보편적인 분산병렬 프레임워크로 진화
 - 그래프 연산 등 다양한 데이터 처리도 지원함

그래프 연산

- 데이터 간의 관계를 표현할 때 사용함
- 점과 선을 처리함

■ 빅데이터 플랫폼의 개념과 진화

2. 빅데이터 플랫폼의 진화

관계형 데이터 모델과 대규모 업무 트랜잭션을 지원할 수 있도록 진화부여

- 빅데이터의 저장, 트랜잭션 및 SQL처리가 가능하도록 관계형 데이터베이스에 NoSQL(Not only SQL) 확장성과 고성능 기능을 부여함

관계형 데이터베이스

- 일련의 정형화된 테이블로 구성된 데이터 항목들의 집합체
- 관계형 데이터베이스가 테이블(Table), 로우(Row), 칼럼(column)이라는 간단한 구조로 구성되어 있음

트랜잭션 및 SQL처리

- 트랜잭션 : 데이터베이스의 상호작용 단위
- SQL : 사용자와 관계형 데이터베이스를 연결시켜 주는 표준검색 언어

NoSQL 확장성과 고성능 기능

- 관계형 데이터베이스는 여러 장비에 데이터를 나누어 저장하고, 나누어진 데이터에 대한 트랜잭션이 안전하게 수행되는 것을 보장하기 위해서 복잡한 방식을 사용할 수 밖에 없음
 - 결국 성능 저하를 유발함
- 분산 환경에 적합한 형태의 새로운 저장시스템인 NoSQL 으로의 확장과 성능향상 요구

■ 빅데이터 플랫폼의 개념과 진화

2. 빅데이터 플랫폼의 진화

파일 관리 효율화

- 기존 대비 절반 정도의 저장 공간을 절약할 수 있도록 진화
 - 작은 파일도 메모리 공간 낭비 없이 저장할 수 있음
 - 원본 복구가 가능한 알고리즘을 활용하여 데이터 복구 정보를 관리함

실시간 빅데이터 처리 가능

- 기존의 일괄처리 방식이 아님
- 실시간 처리 기술 발전
 - 온라인으로 연결된 상태에서 빅데이터 처리 요청과 응답을 즉각적으로 주고 받을 수 있음

다양한 분산병렬 처리 방법 제공

- 보편적인 분산병렬 프레임워크로 진화
 - 그래프 연산 등 다양한 데이터 처리도 지원함

관계형 데이터 모델과 대규모 업무 트랜잭션 지원

- 관계형 데이터베이스에 NoSQL 확장성과 고성능 기능을 부여하여 빅데이터 저장, 트랜잭션 및 SQL처리가 가능하도록 함

파일 관리 효율화

- 기존 대비 절반 정도의 저장 공간을 절약할 수 있도록 진화
 - 작은 파일도 메모리 공간 낭비 없이 저장할 수 있음
 - 원본 복구가 가능한 알고리즘을 활용하여 데이터 복구 정보를 관리함

실무에 적용 가능한 Big Data 분석 개론

빅데이터 플랫폼 개요



한국기술교육대학교
온라인평생교육원

■ 빅데이터 플랫폼의 구조



■ 빅데이터 플랫폼의 구조

1. 소프트웨어 계층

모듈	설명
데이터 처리 및 분석 모듈	▪ 데이터 처리, 분석
데이터 수집 및 정제 모듈	▪ 데이터 추출, 데이터 변환, 데이터 표현
서비스 관리 모듈	▪ 소프트웨어 계층에서 제공하는 서비스 관리
사용자 관리 모듈	▪ 사용자 인증 및 접속관리 ▪ 사용자별 서비스 관리 및 서비스 수준 협약 관리
모니터링 모듈	▪ 플랫폼 및 인프라 스트럭처 서비스 가용성 및 성능 모니터링
보안 모듈	▪ 소프트웨어 계층 보안 관리



빅데이터 처리 및 분석과 이를 위한 데이터 수집 및 정제 등을 수행



■ 빅데이터 플랫폼의 구조

2. 플랫폼 계층

모듈	설명
작업 스케줄링 모듈	• 사용자의 응용을 실행하는 작업을 스케줄링
데이터 및 자원 할당 모듈	• 사용자 데이터의 초기 데이터 및 자원 할당 • 동적 상황을 고려한 데이터 및 자원의 재할당을 수행
프로파일링 모듈	• 인프라 스트럭처 자원을 할당하는 자원 • 사용자의 응용 프로파일링 등을 수행
데이터 관리 모듈	• 사용자 데이터 관리
자원 관리 모듈	• 인프라 스트럭처 자원 관리
서비스 관리 모듈	• 플랫폼 계층에서 제공하는 서비스 관리
사용자 관리 모듈	• 사용자별 인증 및 접속 관리 • 사용자별 서비스 관리
모니터링 모듈	• 인프라 스트럭처 서비스 가용성 및 성능 모니터링
보안 모듈	• 소프트웨어 계층 보안관리

빅데이터를 응용하는 기반을 제공하며
데이터 처리 및 분석과 이를 위한 데이터 수집 및 정제 등을 수행함

■ 빅데이터 플랫폼의 구조

3. 인프라 스트럭처 계층

데이터베이스 관리 시스템의 기능

노드

- 통신망 회선에 사용되는 컴퓨터
- 병렬 및 분산 처리를 지원하는 하드웨어와 스케줄링 정책 등을 포함하고 있어야 함

스토리지

- 대용량의 데이터 저장소
- 파일을 분할하여 분산 저장하는 등 병렬 및 분산 처리에 알맞은 스토리지를 제공해야 함

네트워크

- 컴퓨터 상호 간의 정보 교환과 정보 처리를 위한 데이터 통신망
- 네트워크의 구성, 경로 등을 효과적으로 관리할 수 있어야 함

모듈	설명
자원 배치 모듈	• 사용자에게 제공할 자원 배치
노드 관리 모듈	• 인프라 스트럭처 내의 노드 관리
데이터 관리 모듈	• 인프라 스트럭처 내의 스토리지 관리
자원 관리 모듈	• 인프라 스트럭처 내의 네트워크 관리
서비스 관리 모듈	• 인프라 스트럭처 계층에서 제공하는 서비스 관리
사용자 관리 모듈	• 사용자별 인증 및 접속관리 • 사용자별 서비스 관리 등
모니터링 모듈	• 노드, 스토리지, 네트워크 등 자원 가용성 및 성능 모니터링
보안 모듈	• 소프트웨어 계층 보안 관리



빅데이터 처리 및 분석에 필요한 자원을 제공함

