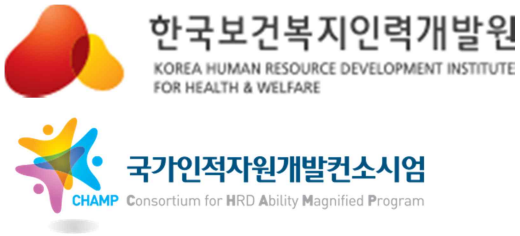


의약품 빅데이터 분석 과정

- 4 차시 -

빅데이터 기술 분류 및 분석

4차시. 빅데이터 기술 분류 및 분석

• 학습목표

1. 빅데이터 분석의 개념과 기술 동향을 파악할 수 있다.
2. 빅데이터 요소기술과 특징을 설명할 수 있다.
2. 빅데이터 분석기술과 특징을 설명할 수 있다.

• 학습하기

1. 빅데이터 분석의 개념 및 기술 동향

■ 빅데이터 분석의 개념

초창기에 "일반적인 데이터베이스 소프트웨어가 저장, 관리, 분석하는 범위를 초과하는 데이터"와 같이 "일정 규모"를 강조하는 관점에서, 최근에는 "다양한 종류의 데이터로부터 저렴한 비용으로 가치를 추출하고 초고속 분석을 지원하는 기술"과 같이 "가치창출 및 활용에 초점을 두고 있다.

■ 빅데이터 기술 동향

정보 과부하, 데이터 홍수는 인간의 본질적 변화를 가져오지는 않지만, 데이터의 양적 변화는 인간 삶의 질적 변화를 만들어 냈다.

- 한국은 데이터 생산대국이다.
- 현재 빅데이터 관리·분석은 취약한 지식 기반과 전통적 경영 현장 환경 등으로 인해 새로운 기술 도입의 필요성
- 빅데이터의 수집경로는 내부 데이터와 외부데이터로 나누어 볼 수 있다.
- 빅데이터는 스마트시대의 경쟁력 향상을 위한 필수 기술로 인식되고 있다.

■ 빅데이터 기술 조건

크기가 큰 데이터만을 처리한다고 모두가 빅데이터 기술이 아니며, 다양한 형태(Variety), 데이터의 양(Volume), 빠른 생성속도(Velocity)의 빅데이터의 세 가지 요소 가운데 두 가지를 충족시킬 수 있으면 빅데이터 기술이라고 정의하고 있다.

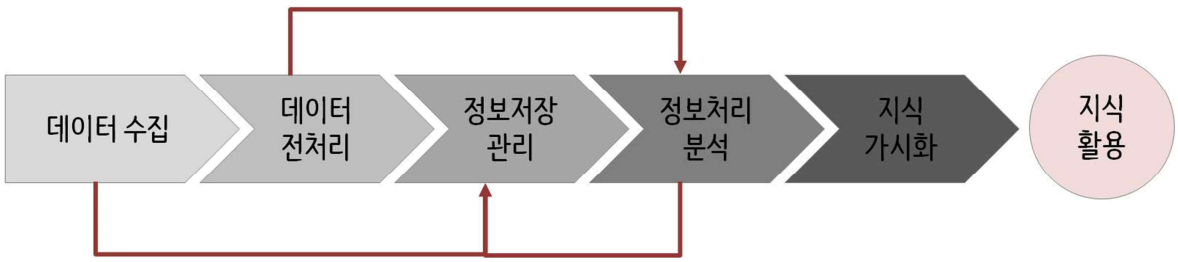
■ 빅데이터의 효용

빅데이터는 단순한 산업이 아니라 인터넷처럼 경제사회 전방에서 혁신을 주도하는 일종의 "플랫폼" 기술(GPT : General Purpose Technology)을 의미함, 빅데이터를 효과적으로 이용 시 기업, 정부, 개인 모두에 막대한 효용을 수반한다.

- (정부) 데이터 기반의 국정혁신 및 사회현안에 선제적 대응
- (기업) 생산성 향상으로 기업의 경쟁력 강화와 새로운 시장 창출
- (개인) 맞춤형 서비스 제공으로 삶의 질 향상

2. 빅데이터 요소기술 및 특징

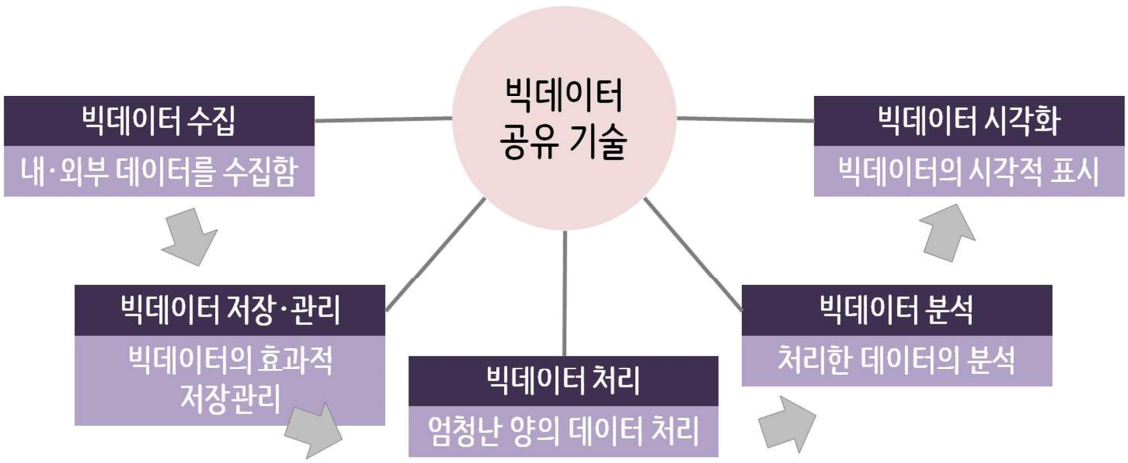
■ 빅데이터의 활용 단계



빅데이터로부터 지식을 발굴해 활용하기까지는 여러 단계를 거쳐 이루어진다.

빅데이터 활용 과정은 데이터 소스, 지식을 활용하는 서비스 분야가 무엇인지에 따라 일부 단계를 건너뛰거나 반복 수행되기도 한다.

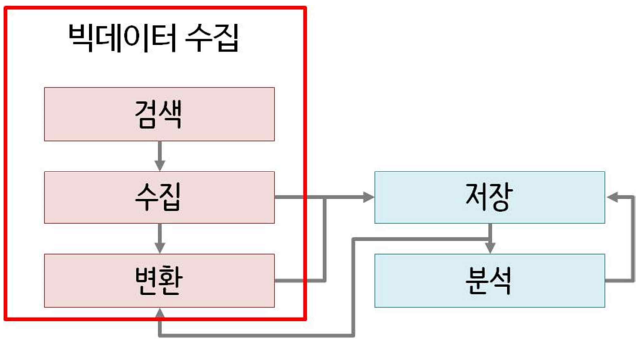
■ 빅데이터의 요소기술의 분류



활용 단계를 토대로 빅데이터 요소기술을 분류할 수 있다. 요소기술은 빅데이터 공유 기

술을 바탕으로 한다. 요소기술에는 내, 외부 데이터를 수집하여 정제되지 않은 데이터를 확보하는 빅데이터 수집, 활용을 위해 빅데이터를 효과적으로 저장 관리하는 빅데이터 저장관리, 엄청난 양의 빅데이터를 처리하는 빅데이터 처리, 빅데이터 분석, 빅데이터 시각화 등이 있다.

■ 빅데이터의 수집기술



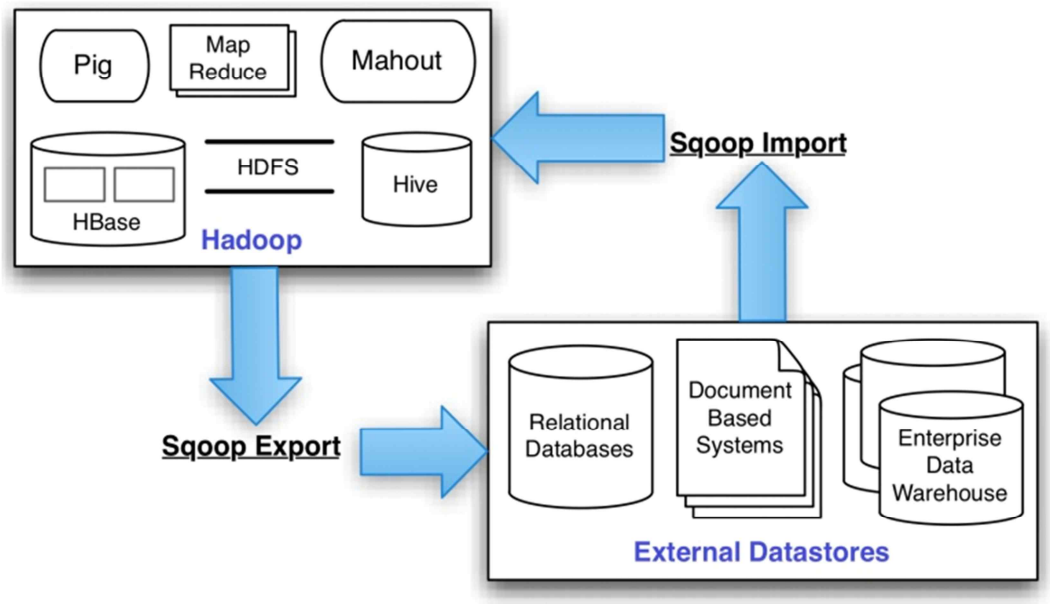
빅데이터 수집기술은 조직내부와 외부의 분산된 여러 데이터 소스로부터 필요로 하는 데이터를 검색하여 수동 또는 자동으로 수집하는 과정과 관련된 기술로 단순 데이터 확보가 아닌 검색/수집/변환을 통해 정제된 데이터를 확보하는 기술을 의미한다.

■ 빅데이터의 수집 방법

빅데이터를 수집하는 방법은 몇 가지가 있는데 좀 더 자세하게 알아보겠습니다. 먼저 로그수집기가 있습니다. 이 방법은 조직의 내부에 존재하는 웹서버의 로그를 수집하고 웹

로드와 트랜잭션 로그, 클릭 로그, 유 로그 데이터 등을 수집하는 방법입니다. 크롤링은 웹로봇을 이용하는 방법으로 조직의 외부에 존재하는 소셜 데이터와 인터넷에 공개되어 있는 자료를 수집하는 방법입니다. 센싱은 각종 센서를 통해 데이터를 수집하는 방법이고 RSS와 오픈 API는 데이터의 생산과 공유, 참여 환경인 웹 2.0을 구현하는 기술을 의미하는데 필요한 데이터를 프로그래밍을 통해서 수집할 수 있습니다.

➢ Sqoop

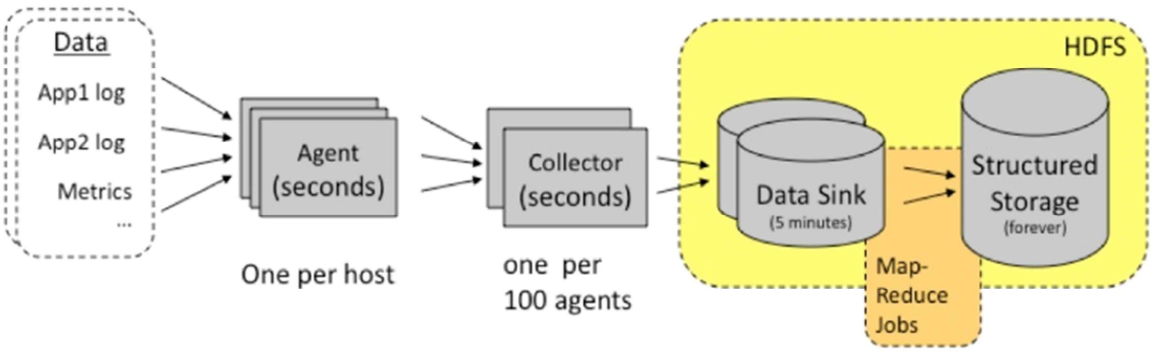


- Hadoop과 관계형 데이터베이스 간의 데이터 전송을 지원하는 기술이다.
- MySQL 같은 관계형 데이터베이스로부터 하둡 분산 파일 시스템으로 데이터를 전송하는데 사용될 수 있다.
- Sqoop은 이러한 프로세스를 자동화하여 처리하는 편의성이 있다.
- Hadoop으의 데이터 전송 시 맵리듀스를 지원하여 분산 및 병렬 처리 등으로 보다 빠른 처리가 가능하다.

➢ Flume

- Cloudera의 Flume은 분산 관리와 안정성으로 만들어진 대용량 데이터를 효율적으로 서버 간 혹은 노드 간 전송할 수 있는 시스템이다.
- Flume은 신뢰성, 가용성, 관리성, 확장성을 설계 목표로 설정하고 간단하고 유연한 구조로서 설계해야 한다.
- Flume은 물리적 노드와 논리적 노드를 모두 마스터가 제어한다.

➢ Chukwa



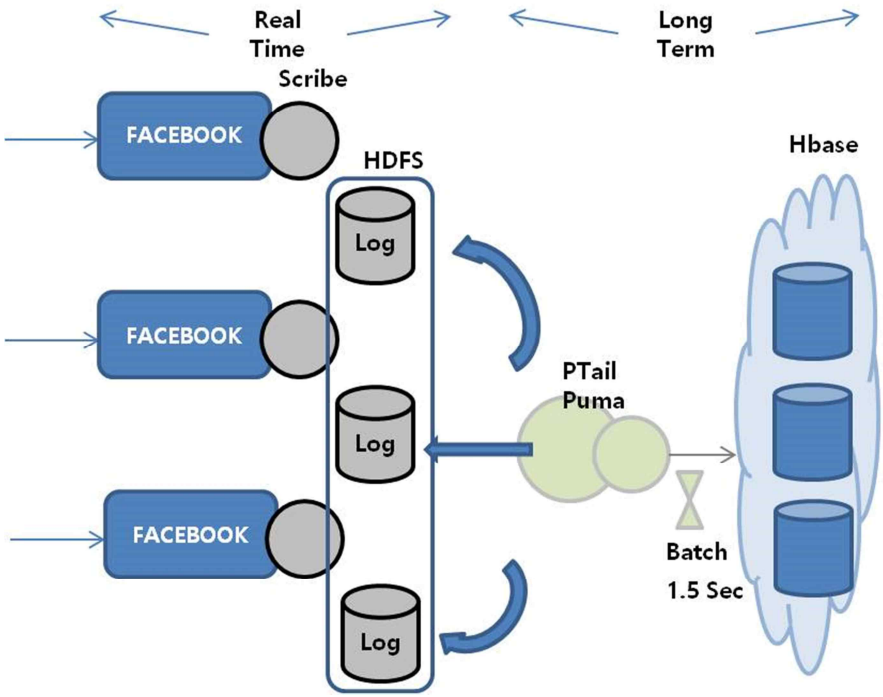
- Apache의 Chukwa는 분산 서버로부터 로그 데이터를 수집하고 이를 저장, 분석하는 것을 목적으로 한다.
- 하둡 클러스터의 로그나 서버의 상태 정보를 관리하며, 응용 프로그램의 로그저장 모듈을 수정하지 않고도 필요한 로그를 수집하여 하둡 분산 파일 시스템에 저장 및 분

석이 가능하다.

➢ Splunk

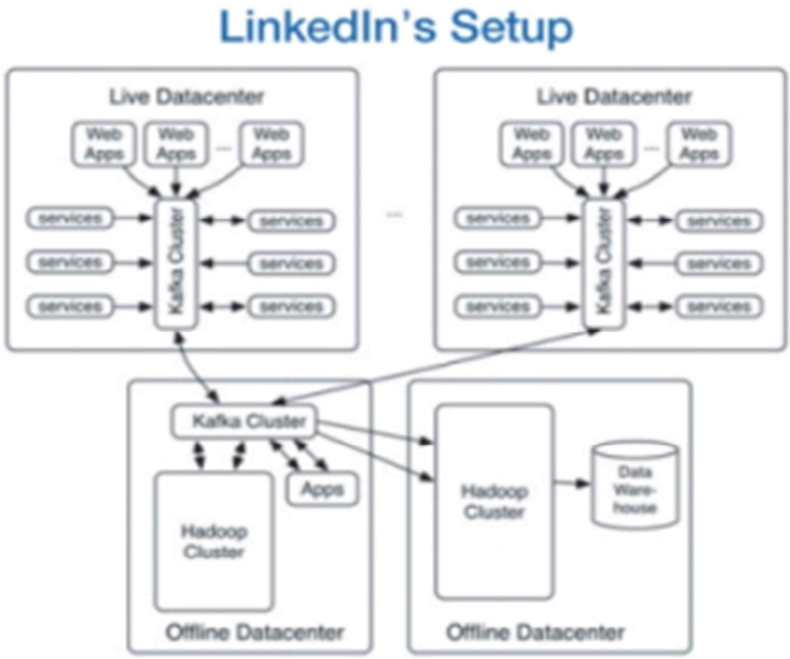
Splunk는 페타비트급의 기록 데이터와 실시간 기계 데이터를 모니터링, 보고, 분석하는데 사용되는 소프트웨어로 기계 데이터 처리의 엔진 역할을 한다.

➢ Scribe



Scribe는 Facebook이 개발하여 공개한 로그 수집 기술로 대량의 서버로부터 실시간으로 흘러오는 로그 데이터를 집약하여, 하둡 분산 파일 시스템에 로그를 저장할 수 있다.

➢ Kafka



- Linkedin에서 최초로 만들어진 Kalka는 분산 메시징 시스템으로써 단일 Kalka 브로커만으로 수천 개의 클라이언트로부터 초당 수백 메가바이트의 읽기와 쓰기를 처리할 수 있다.
- 주요 특징은 아래와 같으며 단순한 로그 수집 기능을 넘어서서 향후 다양한 기술에 활용 가능할 것으로 보인다.

■ 빅데이터 공유기술

- 기업 내 운영 환경에서 한 데이터베이스 시스템에서 발생하거나 변경된 데이터를 다른 시스템에 적용하려는 분산 및 복제 환경은 보편화 되어 있다.
- 데이터 공유를 위한 가장 일반적인 형태로서는 운영 시스템의 데이터 복제(Replication) 기술과 정보계 시스템을 위한 데이터 웨어하우스의 ETL(Extract, Transformation, Load)의 프로세스가 대표적이다.

■ 빅데이터 공유기술 유형의 특징

➢ 데이터 복제(Replication)

분산 환경에 있어서 데이터베이스에 발생한 변경된 정보를 다른 데이터베이스에 반영하여 무장애 시스템을 구현하기 위한 솔루션으로 예전에 스탭 샷으로 사용되었던 기능이다.

➢ 시멘틱 기술

- 시멘틱 웹은 정보의 표현을 넘어 인간 지식을 명시적으로 표현, 공유, 재활용 할 수 있는 웹으로 정의할 수 있음
- 데이터 상호 운용 및 데이터 모델 관점에서 시맨틱 웹은 새로운 산업적 가능성을 제시하고 있음

➢ 멀티 테넌트 공유기술

- 멀티 테넌트환경의 데이터 관리 기술로써 데이터와 데이터 스키마를 분리 혹은 공유하여 멀티 테넌트 환경의 데이터 공유를 가능케 하고, 공유에 의한 보안 요소 검증 기술이 포함한다.
- 테넌트 별 분산 저장된 데이터를 공유하여 각각의 테넌트가 가지고 있는 데이터의 양이 줄어들고 효율적인 데이터 관리가 이루어진다.
- 데이터 공유기술을 통해 기존의 비공유 데이터 관리에 비해 초기 비용은 더 크지만 장기적인 비용으로는 멀티 테넌트 환경에서 데이터 공유 기술을 사용하는 것이 효율적이라고 할 수 있다.

➢ 빅데이터 저장기술

- 빅데이터 기술은 작은 데이터라도 모두 저장하여 실시간으로 저렴하게 데이터를 처리하고 처리된 데이터를 더 빠르고 쉽게 분석하여 비즈니스 의사 결정에 이용한다.
- 구글, 애플, 야후에 의해 요소기술로서 상당한 완성도에 도달한다.
- HDFS/Hbase, Cassandra, MongoDB 등이 대표적이다.
- 병렬 DBMS와 NoSQL은 모두 대량의 데이터를 저장하기 위해 수평 확장 접근 방식을 취하고 있다는 점에서는 동일한 저장기술과 Hadoop이 있다.

■ 빅데이터 저장기술의 유형 및 특징

➢ 병렬 DBMS

- 기존의 관계형 데이터베이스 기술은 하나의 시스템이 모든 영역에 맞춰 사용될 수 있도록 만들어져 있다.
- 병렬 DBMS는 전통적인 RDBMS에서 발전한 형태이며, MPP 구조를 취하고 있는 경우가 많다.

➢ 하둡(Hadoop)

- 하둡은 대용량분산저장과 처리를 위한 프레임워크로, 크게 HDFS와 맵리듀스(MapReduce)로 구분한다.
- 오픈소스인 아파치 하둡은 구글의 GFS를 대처할 수 있도록 분산 파일 시스템(HDFS)과 맵리듀스(MapReduce)를 구현한 빅데어터 처리 기술의 대표적인 프레임워크이다.
- 하둡의 분산 파일시스템은 파일을 블록단위로 나누어 각 노드 클러스터에 저장을 하며, 데이터 유실을 막고 부하처리를 위해 각 블록의 복사본(Replication)을 생성한다.
- 하둡은 적은 비용으로 빅데이터의 처리가 가능할 뿐만 아니라 높은 사용 편의성을 제공한다.

➢ HDFS 클러스터

- HDFS 클러스터는 우선 크게 마스터와 슬레이브 노드를 가진다.
- 마스터 노드는 하나의 네임노드와 데이터노드로 구성되며 네임노드는 파일시스템 트리와 그에 속한 모든 파일, 디렉토리 구조 등이 저장되어 있다.
- 나누어진 블록의 위치 정보를 관리하여 파일이 어떻게 블록으로 분할되고 어느 데이터 노드에 있는지를 관리한다.
- 데이터노드는 실제 데이터를 저장하는 서버를 말하며 고정 크기의 64BM 혹은 128MB(메가바이트)의 블록단위로 파일을 나누어 관리한다.
- 보조네임노드는 네임노드에서 관리하는 파일시스템의 이미지 정보를 백업하는 등의 기능을 수행한다.

■ NoSQL의 등장 배경

- 빅데이터를 효과적으로 저장, 관리하는데 여러 가지 문제가 발생하여 이 문제를 개선, 보완하기 위해 새로운 데이터 저장 기술이 요구되는데 이것을 NoSQL이라고 한다.
- NoSQL은 비관계형 데이터베이스를 지칭하는 분산 환경의 데이터 저장소를 의미하며, Not Only SQL로 불리기도 한다.
- 데이터의 폭발적인 증가로 기존의 관계형 DB로 대용량 데이터를 저장하는 데 한계에 이르렀고, 페이스북, 트위터 같은 새로운 유형의 서비스와 애플리케이션이 출현함에 따라 기존과는 다른 데이터 관리 방식 및 정책이 필요하게 되었다.

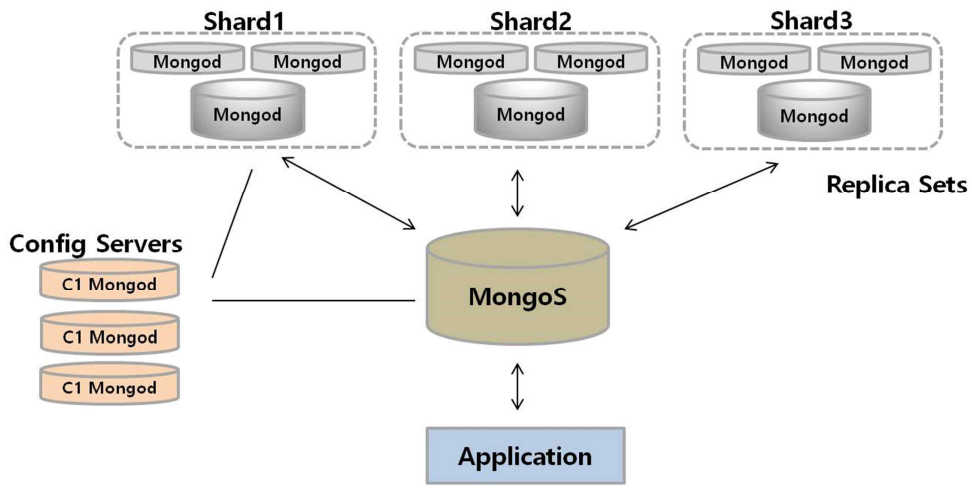
■ NoSQL의 특징

- NoSQL에서 데이터 저장은 해쉬맵(HashMap)처럼 키와 값의 형식으로 다수의 서버에 분산해서 저장하며, 특정 키에 대한 값을 빠르게 조회할 수 있다.
- 동시에 다수의 클라이언트가 접속해서 사용하더라도 트랜잭션이 클러스터를 구성하는 전체 서버에 분산되기 때문에 속도가 빠르며 Throughput도 매우 높다.
- NoSQL 영역에는 수많은 제품이 있기 때문에 어떤 제품을 고를지는 특성이 유사하다 하더라도 교육은 쉽고, 기술 지원 가능한지, 오픈소스일 경우 지속적인 유지 보수 등은 가능한지를 판단하는 것이 중요하다.
- NoSQL은 빅데이터를 지탱하는 기반 기술로서 하둡과 함께 관심이 높아지고 있는 기술이다.

■ MongoDB

- MongoDB는 Replica Sets과 Auto-Sharding으로 구성된 데이터베이스로써, Replica Sets으로 안정성과 가용성을 확보하고 Auto-Sharding으로 분산 확장을 하는 구조이다.
- Config Server에 메타데이터를 저장하고 실제 데이터는 Chunk단위로 나뉘어 Shard서버에 저장하며, Auto-Sharding으로 분산 확장을 하며 데이터의 유실을 막고 가용성을 높이기 위해 각 서버의 Replica를 설정한다.
- JSON의 데이터 저장 구조를 제공하고, Sharding(분산)/Replica(복제) 기능 제공하며, MapReduce(분산/병렬처리) 기능 제공한다.
- CRUD(Create, Read, Update, Delete) 위주의 다중 트랜잭션 처리 가능하다.
- Memory Mapping 기술을 기반으로 빅데이터 처리에 탁월한 성능을 가지고 있다.

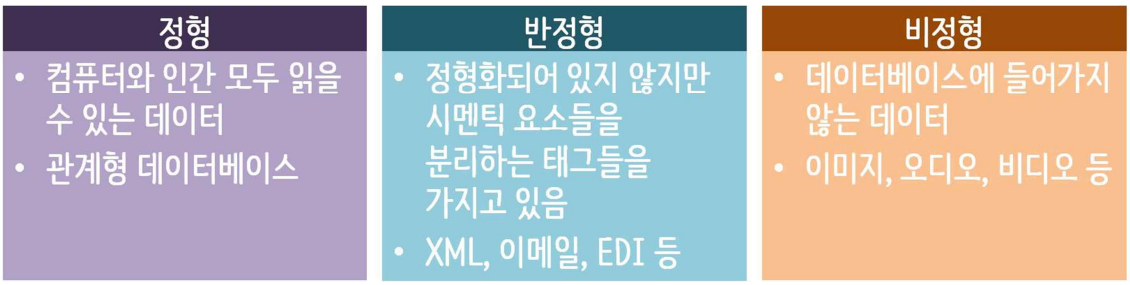
■ 몽고디비의 데이터처리구조



- Shard키로 설정된 칼럼의 범위를 기반으로 각각의 값에 맞는 Shard에 저장된다.
- 필요 시 Shard를 추가하여 migration하여 확장이 가능하다.
- 사용하는 애플리케이션단에서 MongoS라는 라우팅 프로세스로만 연결하기 때문에 Shard의 구조에 대해서는 알 필요도, 수정도 없다.

■ 빅데이터 처리기술 도입 시 고려사항

① 처리할 데이터의 종류



정형 데이터는 컴퓨터와 인간 모두 읽을 수 있는 데이터로 관계형 데이터베이스가 여기에 해당된다.

반정형 데이터는 정형화되어 있지는 않지만 시멘틱 요소들을 분리하는 태그들을 가지고 있는 것으로 XML, 이메일, EDI 등이 여기에 해당된다.

비정형 데이터는 데이터베이스에 들어가지 않는 데이터이다. 이미지, 오디오, 비디오 등이 있다.

② 데이터를 처리하기 위한 조건

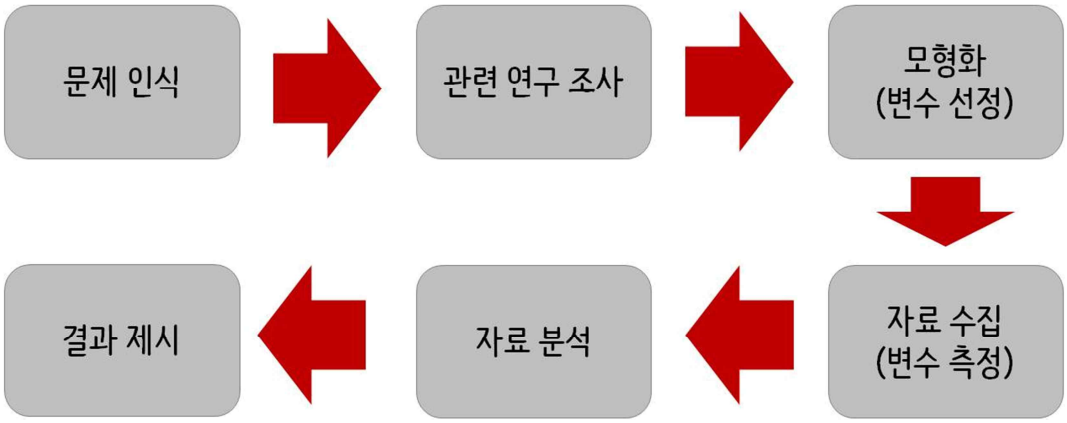
데이터의 접근 제한을 풀고 접속 권한을 제공해 데이터를 저장하고 사용할 수 있게 해야 한다. 또, 정보를 미가공 형태로 남겨야 분석 시스템으로 실시간 스트리밍 되면서 분석하고 보고 가능하다. 정형 데이터에서는 데이터 처리 과정이 직관적으로 일어나지만, 비정형 데이터는 고급 알고리즘과 강력한 엔진을 반드시 거친 후 들어오는 데이터만을 처리할 수 있기 때문에 이에 주의해야 한다.

3. 빅데이터 분석기술 및 특징

■ 빅데이터 분석기술

- 빅데이터 분석은 대량의 데이터로부터 숨겨진 패턴과 알려지지 않은 정보간의 관계를 찾아내기 위한 과정이다.
- 비즈니스 영역에서 주로 수행되는 빅데이터 분석의 목적은 데이터 과학자들에 의해 분석된 정보를 토대로 기업의 의사결정을 수행한다.
- 빅데이터 분석을 위해 크게 데이터마이닝과 예측 분석 등이 고려되며, NoSQL 데이터베이스, 하둡과 맵리듀스 등의 관련 기술이 있다.
- 빅데이터 분석은 더 짧은 시간 안에 보다 더 많은 정보를 빅데이터로부터 추출하는 것을 목표로 한다.

■ 빅데이터 분석 과정



빅데이터를 분석하는 과정은 단계적으로 진행된다. 먼저 문제를 인식한 다음 그 문제에 대한 관련 연구를 조사한다. 그리고 모형화를 하여 변수를 선정하고 변수를 측정할 수 있도록 자료를 수집한다. 수집한 자료를 분석하고 결과를 제시한다.

■ 데이터 마이닝

가트너에 따르면 데이터 마이닝은 통계 및 수학적 기술뿐만 아니라 패턴인식 기술들을 이용하여 데이터 저장소에 저장된 대용량의 데이터를 조사함으로써 의미 있는 새로운 상관관계, 패턴, 추세 등을 발견하는 과정이다.

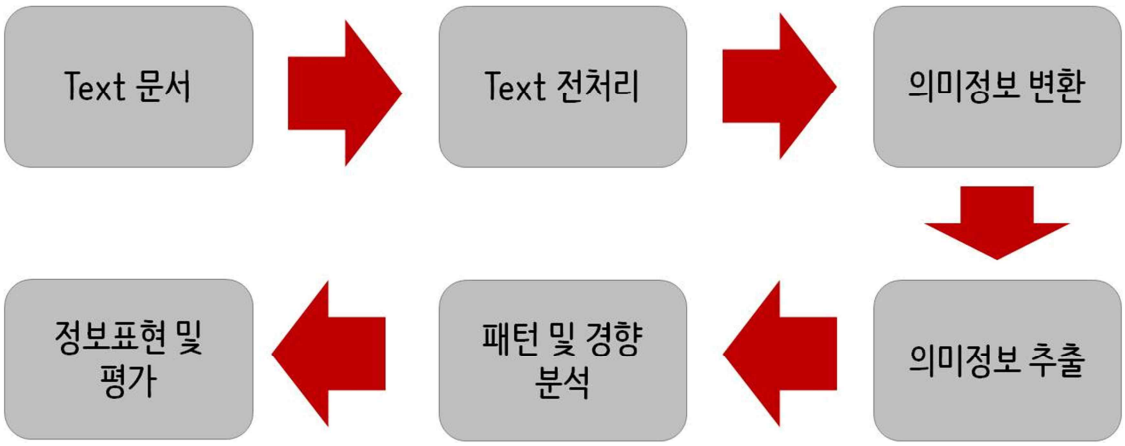
- 데이터 마이닝은 다양한 분야에서 활용될 수 있으며, KDD, 기계학습, 패턴인식, 통계학, 신경망 컴퓨팅 등과 관련하여 빅데이터 분석에 있어서 가장 기본적인 분석 기술이다.
- 분류(Classification), 추정(Estimation), 예측(Prediction), 데이터 축소(Data Reduction), 데이터 탐색(Data Exploration)

■ 텍스트 마이닝

· 좁은 의미로는 불명확하고 찾기 힘든 텍스트 기반의 데이터(문서)로부터 새로운 정보를 발견할 수 있도록 관련 방법을 제공하는 기술이며, 넓은 의미로는 이와 관련된 정보검색, 정보추출, 정보체계화, 정보 분석을 모두 아우르는 Text-Processing 기술 및 처리과정이다.

· 텍스트 마이닝이란 구조화되지 않은 대규모의 텍스트 집합으로부터 새로운 지식을 발견하는 과정으로 텍스트 문서 전처리 및 패턴 분석 등의 단계를 가지며, 순환구조로써 계속적인 피드백을 수행한다.

■ 텍스트 마이닝의 과정



텍스트 마이닝의 수행 단계는 먼저 텍스트 문서를 수집하여 전처리를 한 후 의미 있는 정보로 변환한다. 그 다음 텍스트의 패턴과 경향을 분석하여 정보로 표현하고 평가한다.

■ 예측 분석

· 과거 자료와 변수 간의 관계를 이용하여 관심이 되는 변수를 추정한다.

· 통계분석, 데이터 마이닝 및 텍스트 마이닝 기술들을 기반으로 예측 분석을 수행한다.

· 예측분석은 데이터 마이닝의 기법 중 하나이지만 빅데이터를 분석하고 활용하기 위해 비즈니스적 필요성에 의해 많이 연구 및 개발되고 있다.

· 시계열 분석 기법은 대표적인 비즈니스 예측 분석 기법 중 하나로, 시가의 흐름에 따라 순서대로 관측되어 시간의 영향을 받게 되는 자료를 분석하여 예측하는 기술이다.

·

■ 예측 분석의 목표

예측 분석의 목표는 과거의 데이터나 사건으로부터 미래에 발생 가능한 상황이나 사건을 예측하여 선제적인 의사결정을 지원한다.

■ 빅데이터 활용의 기대효과

- 예산 절감
- 정책 의제 선정 근거 확보
- 정책 수행 최적화
- 대국민 서비스 만족도 제고
- 총체적 사회 비용 절감
- 공공 기관과 국가 투명성 제고
- 위기 대응 능력 향상
- 생산성 향상
- 신 비즈니스 창출
- 업무효율의 제고