KoDB
한국데이터베이스진흥원

빅데이터 거버넌스

BA BIGDATA ACADEMY
빅 데 이 터 아 카 데 미

목 차

제 1부. 빅데이터 거버넌스 일반

Module-01. 빅데이터 거버넌스 소개
Module-02. 빅데이터 거버넌스 프레임워크
Module-03. 빅데이터 플랫폼
Module-04. 빅데이터 참조아키텍처

제 2부. 빅데이터 거버넌스 기술

Module-05. 빅데이터 프라이버시
Module-06. 빅데이터 품질
Module-07. 마스터 데이터 통합
Module-08. 메타데이터
Module-09. 빅데이터 수명주기 관리

제 3부. 빅데이터 거버넌스 구축

Module-10. 성숙도 측정
Module-11. 로드맵 수립
Module-12. 빅데이터 거버넌스 조직
Module-13. 비즈니스 프로세스 통합

제 4부. 빅데이터 거버넌스 구축 사례

Module-14. 웹과 소셜미디어
Module-15. M2M 데이터
Module-16. 빅 트랜잭션 데이터
Module-17. 생체 데이터
Module-18. 사람이 생성한 데이터
Module-19. 헬스케어
Module-20. 유틸리티 산업
Module-21. 통신 서비스 공급자

Module-01. 빅데이터 거버넌스 소개

□ 빅데이터 정의

▷ 일반적인 데이터베이스 SW가 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터(Mackinsey, 2011)

▷ 다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고 (데이터의) 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처(IDC, 2011)

▷ 대량의 데이터 집합으로부터 유용한 정보를 추출하는 것.(Hand et al. 2001)

▷ 주된 기법 : 데이터마이닝

⊙ 데이터마이닝이란 의미있는 패턴과 규칙을 발견하기 위해서 자동화되거나 반 자동화된 도구를 이용하여 대량의 데이터를 탐색하고 분석하는 과정이다. (Berry and Linoff. 1997, 2000)

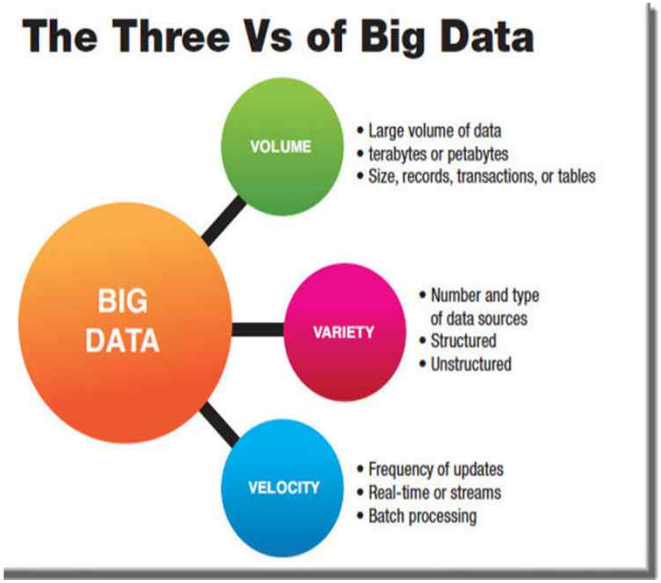
⊙ 데이터마이닝은 통계 및 수학적 기술뿐만 아니라 패턴인식 기술들을 이용하여 데이터 저장소에 저장된 대용량의 데이터를 조사함으로써 의미있는 새로운 상관관계, 패턴, 추세 등을 발견하는 과정이다. (가트너그룹 2004.1)

참고문헌 : 빅데이터 핵심 기술 및 표준화 동향 (ETRI, 2013.2)

Module-01. 빅데이터 거버넌스 소개

□ 빅데이터 특성

▷ 일반적으로 3가지 특징을 가지는 큰 데이터



출처 : 가트너

1. Volume
- 분석하는 데이터의 크기가 일정 수준 이상이어야 의미있는 데이터를 취득하는 것이 가능
- 통계에서 표본 데이터가 많아야 정확도가 올라가는 것과 같음
- 일반적으로 100테라바이트 이상의 데이터를 빅데이터라 칭함

2. Velocity
- 가공되지 않는 원시 데이터에서 가치를 찾는 것
- 데이터가 계속 변하는 경우 변화에 따라 새로운 분석 방법, 새로운 가치 부여가 가능
  예) 소셜 네트워크

3. Variety
- 데이터의 다양성은 데이터가 만들어 내는 정보의 가치를 건강하게 함
- 건강하다 => 사실에 가깝거나, 사람들이 체감적으로 공감하는 내용에 가깝다는 것을 의미
  예) 선거 여론 조사에서 표본의 다양성

참고문헌 : 빅데이터 핵심 기술 및 표준화 동향 (ETRI, 2013.2)

Module-01. 빅데이터 거버넌스 소개

□ 빅데이터 범위

▷ 산업계에서 빅데이터를 바라보는 시선

⊙ 현재 거론되는 빅데이터

- 비정형 데이터를 중심으로 효과적으로 데이터를 처리하는 기술
- 정형 데이터베이스에서 대규모 저장 시스템을 연구

⊙ 광의의 빅데이터

- 기존의 대용량 데이터베이스와 장비 시장을 장악하는 글로벌 기업들을 중심으로 부각되고 있는 기술

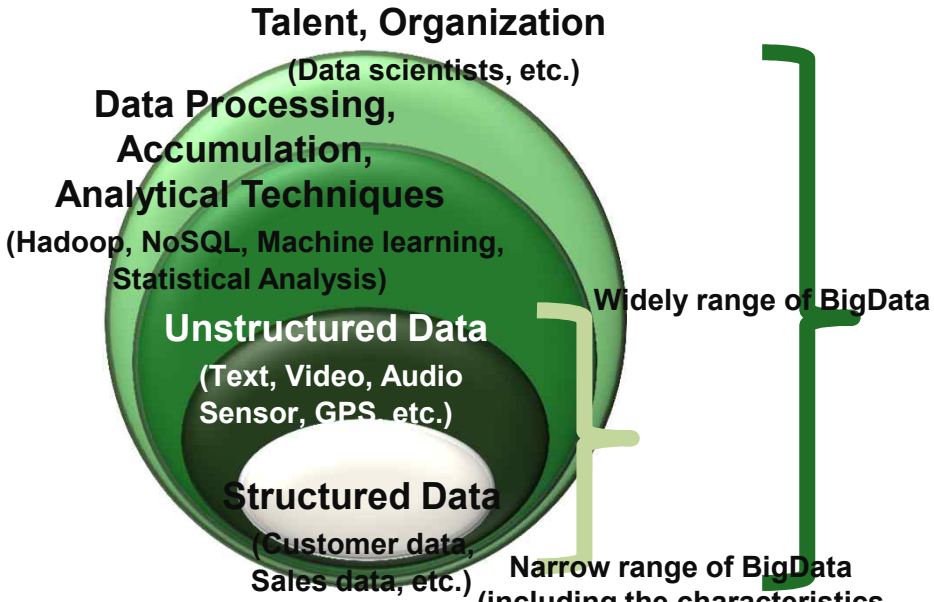
▷ 빅데이터에 대한 데이터베이스 기술

⊙ 버려지고 있는 데이터들에 대한 관심으로 기술의 변화

⊙ 대상 데이터를 클러스터링과 필터링 방식을 통해 효과적으로 걸러내는 기술이 등장

⊙ 적은 비용으로 새로운 가치를 발굴하는 것이 가능

- 데이터 처리 기술의 발전과 컴퓨팅 속도의 발전으로 가능



출처 : 노무라종합연구소

참고문헌 : 빅데이터 핵심 기술 및 표준화 동향 (ETRI, 2013.2)

Module-01. 빅데이터 거버넌스 소개

□ 빅데이터 특성

▷ 변화된 데이터 생산현장

⊙ 조직 내·외부의 다양한 서비스 환경은 새로운 부가가치를 창출하는 데이터를 생산, 관리하는 새로운 환경

항목	전통적 데이터	빅데이터
데이터 소스	전통적 정보 서비스(기업/기관의 업무 관련)	일상화된 정보 서비스(Twitter, Facebook 등 포함)
목적	업무, 효율성	사회적 소통, 자기표현, 사회기반 서비스
생성 주체	정부, 기업 등 조직	개인, 시스템 등
데이터 유형	.정형 데이터 .조직 내부 데이터(고객정보, 거래정보 등) .주로 비공개 데이터	.비정형 데이터(비디오 스트림, 이미지, 오디오, 소셜 네트워크 등의 사용자 데이터, 센서 데이터, 응용프로그램 데이터 등) .조직 외부 데이터 .일부 공개 데이터
데이터 특징	.데이터 증가량 관리 가능 .신뢰성 높은 핵심 데이터	.기하급수적 양적 증가 .Garbage(쓰레기) 데이터 비중 높음 .문맥정보 등 다양한 데이터
데이터 보유	정부, 기업 등 대부분 조직	.인터넷 서비스 기업(구글, 아마존 등) .포털(네이버, 다음 등) .이동통신회사(SKT, KTF 등) .디바이스 생산회사(애플, 삼성전자 등)
데이터 플랫폼	정형 데이터를 생산, 저장, 분석, 처리할 수 있는 전통적 플랫폼 ex) 분산 DBMS, multi processor, 중앙집중처리	비정형의 대량 데이터를 생산, 저장, 분석, 처리할 수 있는 새로운 플랫폼 ex) 대용량 비정형 데이터 분산 병렬 처리(HDFS)

참고문헌 : 빅데이터 시대의 데이터 자원 확보와 품질관리방안 (NIA, 2012.5)

Module-01. 빅데이터 거버넌스 소개

□ 빅데이터 프로젝트 시작 전 알아야 할 사항

▷ 대부분의 조직에서 빅데이터 프로젝트를 수행하는 이유

⊙ 더 나은 분석을 수행하고 분석할 데이터의 양이 크게 증가할 것임을 인식하고 있기 때문

⊙ 고객이 제품을 보다 효율적이고 효과적으로 사용하도록 도와줄 수 있는 분석을 제공(종종 실시간으로)하여 제품을 서비스 계층에서 보완할 수 있다는 것을 인식하기 때문

⊙ 빅데이터를 사용하여 특정 비즈니스 단위 또는 프로세스에 대한 모든 의사 결정을 알려 작업을 보다 빠르고 효율적이며 저렴하게 수행하고 싶기 때문

⊙ 빅데이터가 조직의 모든 비즈니스 단위에 중요하다는 것을 인식하여 데이터 중심 관점을 위한 토대를 마련하려고 하기 때문

⊙ 너무 늦지 않게 빅데이터에 익숙해지기 시작해야 한다는 점을 알고 있지만 실제로 빅데이터로 무엇을 해야 할지 아직 파악하지 못했기 때문에 빅데이터를 학습하고 실험함

참고문헌 : 최고의 빅데이터 워크북, 인포메티카

Module-01. 빅데이터 거버넌스 소개

□ 빅데이터 프로젝트 시작 전 알아야 할 사항(계속)

▷ 빅데이터 프로젝트가 실패하는 이유

⊙ 모호한 목표

- 프로젝트의 '부정확한 범위 ' 때문임. 명확한 목표 없이 의욕적인(한데 합치면 과도하게 모호한) 프로젝트를 목표로 한 다음 중요한 사항을 가려내는 힘든 선택을 해야 할 때 실패

⊙ 잘못 관리되는 기대치

- 빅데이터와 관련된 모든 부풀려진 측면은 프로젝트가 제공할 수 있는 가치에 대한 몇 가지 매우 위험한 가정으로 이어짐. 영향도와 통찰력에 대한 기대가 지나치게 높으면 기한 및 예산이 전혀 타당하지 않게 됨

⊙ 프로젝트의 과도한 비용 지출 및 지연

- 이 분야가 기업에 얼마나 새로운지를 고려하면 대부분의 빅데이터 프로젝트가 과도한 비용으로 종결되거나 시간이 너무 오래 걸림. 희소하고 비싼 Hadoop/Java 개발자를 채용하여 핸드 코딩 구현을 맡긴다면 오류 없이 샌드박스 환경에서 벗어나는 것이 불가능하다는 것을 곧 깨닫게 됨

⊙ 확장 불가능

- 5명의 뛰어난 Hadoop/Java 개발자를 찾는 것도 어려운데, 프로젝트가 성장하여 1년 후 30명의 Java 개발자로 확장해야 한다면 벽에 부딪칠 수 있음

참고문헌 : 최고의 빅데이터 워크북, 인포메티카

Module-01. 빅데이터 거버넌스 소개

□ 빅데이터 프로젝트 시작 전 알아야 할 사항(계속)

▷ 데이터 거버넌스 수립

⊙ 빅데이터에 대한 보다 기초적인 노력을 기울인다면 데이터 거버넌스에 대한 절차적 프레임워크를 수립해야 함

⊙ 실제로 빅데이터 프로젝트가 단일 부서에 가치를 제공하는 것을 목표로 하는 경우에도 소규모 데이터 거버넌스 위원회를 구성하여 이러한 기구에서 겪는 고유한 문제에 대처하는 방법을 습득할 수 있도록 할 것

⊙ 기본적으로 데이터 거버넌스 위원회는 기업의 데이터 접근 방식을 감독하기 위해 임원들로 구성되는 정식 기구이며, 특정 비즈니스 단위의 데이터를 관리하는 직능 또는 부서별 담당자인 데이터 관리자도 포함됨. 실제로 일부 고객은 데이터 도메인에 따라 데이터 관리 역할을 할당하는데, 제품 데이터, 고객 데이터 등에 대한 담당자를 별도로 두고 있음

⊙ 데이터 거버넌스 위원회

- 빅데이터 프로젝트에 참여하는 각 현업 부서 단위의 직능간 고유한 관점 및 요구 사항을 대표할 수 있는 기구
- 직능, 부서 및 도메인 간의 원활한 커뮤니케이션
- 직능 간 프로세스는 자동화 및 예외 보고 규칙을 만들고 협업 툴을 채택하여 커뮤니케이션을 개방적이고 편리하게 유지
- 프로젝트의 주요 목표를 효과적으로 전달하고 데이터 거버넌스 위원회에 참여한 모든 사람이 이러한 목표에 전념
- 의사 결정은 한 부서 단위에서 생각하는 단기적 이익보다 전체 위원회의 장기적 이익을 고려해야 함

참고문헌 : 최고의 빅데이터 워크북, 인포메티카

Module-01. 빅데이터 거버넌스 소개

□ 빅데이터 프로젝트 시작 전 알아야 할 사항(계속)

▷ 데이터 거버넌스 수립(계속)

역할	이미 이 역할을 수행할 수 있는 사람이 있습니까?	이 역할을 위한 채용 필요성	주어진 시간에 따라 채용해야 하는 인원
데이터 과학자	✓또는 ×	✓또는 ×	
도메인 전문가			
비즈니스 분석가			
데이터 분석가			
데이터 엔지니어			
데이터베이스 관리자			
엔터프라이즈 아키텍트			
비즈니스 솔루션 아키텍트			
데이터 아키텍트			
데이터 관리자			
ETL(데이터 통합) 개발자			
애플리케이션 개발자			
대시보드 개발자			
통계 모델러			
기타			
기타			
기타			

통합적 사고의 필요성

새로운 팀 구성원을 찾을 때 적절한 자격을 갖춘 사람으로 제한하지 마십시오. 적절한 자격을 갖춘 사람을 찾는 것 자체가 분명히 하나의 도전 과제입니다. 비즈니스 목표와 기술적 역량을 조합할 의지가 있는 사람도 찾아야 합니다.

빅 데이터 프로젝트에 참여하는 사람들이 비즈니스의 현실을 이해하고 복잡한 데이터 과학을 이행할 수 있는 것이 얼마나 중요한지를 강조하는 고객이 점점 늘어나고 있습니다. 이러한 유형의 통합적 사고는 거대하고 찾기 어렵습니다. 이는 교육하고 보상할 가치가 있습니다.

참고문헌 : 최고의 빅데이터 워크북, 인포메티카

Module-01. 빅데이터 거버넌스 소개

□ 데이터 거버넌스

▷ 데이터 거버넌스 정의

⊙ 전사적인 차원에서 보유하고 있는 모든 데이터에 대해 관리에 대한 정책, 지침, 표준, 전략 및 방향을 수립하고 데이터를 관리할 수 있는 조직 및 서비스를 구축하는 데이터 관점에서의 IT 관리체계를 말함

⊙ 데이터 거버넌스의 궁극적인 목적은 고품질 데이터의 확보와 관리를 통해 기업에 제공할 수 있는 정보 활용을 극대화하여 기업의 다양한 가치 창출에 기여하는 것임. 이것은 데이터로 인한 리스크의 감소, 데이터 관리 절감과 데이터 활용도 증대를 통해 데이터의 가치를 향상시키고, 이러한 고품질 정보가 기업의 비즈니스 목적에 부합하고 최적의 서비스를 제공할 수 있도록 효과적으로 관리되고 진화될 수 있어야 한다는 것을 의미함

▷ 데이터 거버넌스 구현 시 고려사항

⊙ 데이터 거버넌스가 성공적으로 구현되기 위해서 반드시 데이터 아키텍처가 구축되어야 함. 데이터 아키텍처는 기업의 전사적 아키텍처(EA)의 중요한 하부구조로, 기업의 모든 비즈니스를 데이터 측면에서 처음부터 끝까지 조명하여 시스템의 본질인 데이터를 체계적, 구조적으로 관리하고 설계하는 전 과정을 말함

참고문헌 : 빅데이터 성공의 지름길_데이터 거버넌스 구현, 엔코아

Module-01. 빅데이터 거버넌스 소개

□ 데이터 거버넌스

▷ 효과적인 데이터 거버넌스 구축 전략

⊙ 데이터의 품질 관리

- 데이터의 품질 관리는 데이터가 생성된 이후에 데이터 값의 정합성을 관리하는 것으로서, 데이터 검증을 위한 품질 지표를 규정하여 데이터 값의 정합성을 측정하고 오류 원인 파악과 해결, 지속적인 측정 모니터링을 통해 결과를 개선

⊙ 데이터 구조관리

- 데이터 구조 관리는 기존 시스템의 데이터 항목을 수집, 분석하여 이를 기준으로 전사 표준 단어, 도메인, 항목을 정의하고 데이터 모델링 시 이를 적용하여 작성. 또한 DB에 생성 시 절차를 통해 데이터 모델의 반영을 통제하고, 데이터 모델의 표준 준수와 데이터 모델과 DB와의 GAP분석을 통해 설계와 구현간의 차이를 최소화

⊙ 데이터 관리체계 수립

- 데이터 품질과 구조를 관리한 이후에는 데이터 관리 원칙 및 절차를 수립하고, 조직/책임/역할 등을 정의 및 분담하여 지속적이고 체계적인 관리가 수행되어야 함

참고문헌 : 빅데이터 성공의 지름길_데이터 거버넌스 구현, 엔코아

Module-01. 빅데이터 거버넌스 소개

□ 데이터 거버넌스

▷ 빅데이터 시대의 데이터 거버넌스의 중요성

⊙ 최근 빅데이터의 대부분을 형성하는 비정형 데이터는 구조화되지 않은 상태로 데이터 품질이 매우 낮음. 더욱이 정형 데이터의 품질조차 확보되지 않은 상태에서 데이터의 양이 기하급수적으로 늘어나 데이터를 통한 분석 및 예측이 힘들어 지자, 그에 따른 비용이 발생

⊙ 데이터 거버넌스의 출발점은 바로 의미있는 기존 정형 데이터를 분석하고 품질을 확보하는 것에 있음. 트렌드에 휩쓸려 빅데이터에 너무 조급하게 접근하기 보다는 기업 내부에 있는 기존 데이터에 대해 다시 한 번 바라보고 내부 인력과 책임질 조직들을 만들어 데이터 거버넌스부터 구현해야 함

⊙ 데이터 거버넌스는 빅데이터 시대의 중요한 요소로서, 데이터가 증가할수록 더 철저한 거버넌스 정책이 필요함. 그렇지 않으면 거대한 데이터의 홍수에 빠져서 헤어 나오지 못하고 통찰력을 잃을 수 있으며, 데이터의 의미를 잘못 파악해 경쟁력을 악화시키는 결과를 낳을 수도 있음

참고문헌 : 빅데이터 성공의 지름길_데이터 거버넌스 구현, 엔코아

Module-01. 빅데이터 거버넌스 소개

□ 빅데이터 거버넌스 고려사항

▷ 6가지 고려사항

⊙ 빅데이터와 정형데이터 통합 운영 시 : 중복정보에 대한 우선순위(신뢰성↑ or 최신성↓)

⊙ 사용자를 알 수 없는 악의적 빅데이터(abuse)에 대한 필터링 방안 연구

⊙ 빅데이터 처리 단계별 데이터 품질 확보방안 연구

- 데이터 자체 품질에 대한 신뢰도, 추출/통합 데이터에 대한 신뢰도, 분석정보 품질에 대한 신뢰도

⊙ 빅데이터 연계를 통한 타 기관 정보 제공(공유)시 보증방안 연구

- 제공자 측면 데이터 품질 확보 후 제공, 공유정보에 대한 일정 표준 포맷 정의, 품질이 확보(검증)된 데이터라는 공식적인 승인 후 제공(제도 마련), 분석정보 활용에 대한 관리

⊙ 빅데이터 개인정보보호 수준(승인) : 일반 데이터와는 다른 보호강도 조절, 일정기간 유예

- 사용자 정보수집 시 유형별 인증 수준, 사용자 정보 활용 시 유형별 인증수준

⊙ 빅데이터 수집, 통합, 활용을 위한 빅데이터 시스템내의 품질관리체계 개발

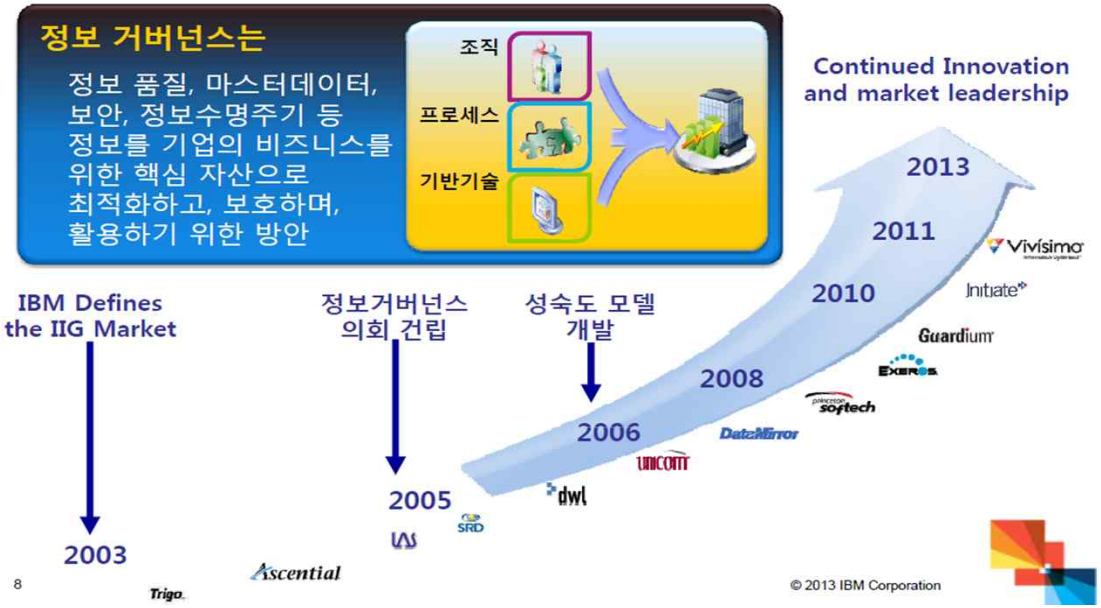
- 품질관리 전문가, 빅데이터 분석가 양성 및 활동 지원, 품질관리 체계 격상 필요

참고문헌 : [특별기고] 기업이 빅데이터 경쟁력을 갖는 방법3 (비투엔, 2013.7)

Module-01. 빅데이터 거버넌스 소개

□ 정보 거버넌스

▷ 정보 품질, 마스터데이터, 보안, 정보수명주기 등 정보를 기업의 비즈니스를 위한 핵심 자산으로 최적화하고, 보호하며, 활용하기 위한 방안



참고문헌 : 빅데이터 시대의 정보 거버넌스 전략 (IBM, 2013)

Module-01. 빅데이터 거버넌스 소개

□ 빅데이터 거버넌스

▷ 빅데이터 거버넌스 정의

: 정보 거버넌스의 일부로써 빅데이터의 최적화, 프라이버시, 가치창출과 관련된 정책을 정하는 것

⊙ 정보 거버넌스의 일부

- 정보 거버넌스의 범위를 확대하여 빅데이터 거버넌스를 포함하도록 함
- 정보 거버넌스 위원으로 데이터 과학자와 같은 빅데이터 전문가 추가
- 빅데이터 카테고리별(예를 들면, 소셜미디어) 데이터 관리책임자(Data Steward)를 지정
- 빅데이터에 필요한 메타데이터, 프라이버시, 데이터 품질, 마스터 데이터 등을 정보 거버넌스 원칙에 맞게 조정

⊙ 빅데이터 거버넌스 정책을 정함

예) Facebook 프로파일은 고객의 동의가 없는 한 마스터 데이터 레코드에 통합할 수 없음

⊙ 빅데이터는 최적화되어야 함

- 메타데이터 관리 : 빅데이터의 종류를 파악하여 목록을 작성
- 데이터 품질관리 : 빅데이터를 정제하여 품질을 관리
- 정보수명주기 관리 : 의미가 없는 데이터는 비용관리 상 폐기하거나 보관

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-01. 빅데이터 거버넌스 소개

□ 빅데이터 거버넌스

▷ 빅데이터 거버넌스 정의(계속)

⊙ 빅데이터의 프라이버시 관리가 중요함

- 소셜 미디어, 위치정보, 생체정보, 그외 개인식별정보를 취급하는 조직들은 프라이버시 침해로 인한 평판 리스크, 규제 리스크, 법적 리스크를 신중하게 고려해야 함

⊙ 빅데이터는 돈으로 전환될 수 있음

- 빅데이터는 제3의 조직에 판매하거나, 새로운 서비스를 개발하는데 사용함으로써 돈으로 전환이 가능해야 함
- 빅데이터는 재정적인 가치를 지닌 회사의 자산으로 인식되어야 함

예) 센서 데이터를 이용하여 장비의 가동시간을 향상시킬 수 있음

⊙ 빅데이터는 기능(부서)들간에 자연스런 긴장을 조성함

- 빅데이터 거버넌스는 조직의 여러 기능들간에 조성되는 경쟁적, 대립적인 목표들을 조정해야 함

예) 무선통신 마케팅 부서는 위치정보를 이용하여 새로운 수입원 발굴. 유선 사업부서는 가입자의 동의없이 위치정보를 사용하는데 있어서 발생할 수 있는 평판 리스크 우려

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-01. 빅데이터 거버넌스 소개

□ 빅데이터 프로젝트 수행 및 활용을 위한 고려사항

▷ 일반적인 정보 거버넌스 프로젝트에 대한 빅데이터 거버넌스 프로젝트의 애로사항

- ⊙ 프로젝트가 초기수용자(early adapters)에 의해 주도됨
- ⊙ 비즈니스 문제가 깊이있게 발견될 필요가 있음
- ⊙ 하둡(Hadoop)과 같은 기술이 대개 전방을 차지하고 있음
- ⊙ 비즈니스 케이스가 아직 개발되지 못하였음
- ⊙ 데이터의 특징이 아직 명확하지 못함

▷ 빅데이터 거버넌스 프로젝트 수행 시 준비사항

- ⊙ 빅데이터 분석 및 처리 기술보다 우선시되지 못하는 현실이지만, 프라이버시, 빅데이터 스튜어드십(stewardship, 관리권), 데이터 품질, 메타데이터, 정보 수명주기관리 등에 대한 대책이 고려되어야 함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-02. 빅데이터 거버넌스 프레임워크

□ 프레임워크와 유사 용어 정의

▷ 프레임워크

⊙ 틀, 특정 기술을 체계적으로 설명하기 위한 수단으로 구성요소(컴포넌트)들을 체계적으로 위치시키고, 그들간의 연관/연계 관계를 표시함

⊙ 편리한 사용환경 제공 , 표준절차 및 프로세스 지원, 재사용성 보장, 유지보수의 용이성 보장, 시스템 공통모듈 제공 및 레이어 독립, 시스템 확장성 보장, 다양한 어플리케이션 지원, 플랫폼 독립성 보장

▷ 플랫폼

⊙ 하드웨어 플랫폼 : 표준 공정을 통해 다양한 제품을 만들어내는 기반이자 도구를 지칭

⊙ 소프트웨어 플랫폼 : 여러가지 기능들을 제공해주는 공통 실행환경

⊙ 빅데이터 플랫폼 : 빅데이터 참조아키텍처의 여러 요소를 포함하는 도구(Tool) 세트를 제공하는 제품. 빅데이터 처리를 위해 통합된 사양(Specification)

▷ 아키텍처

⊙ 프로세스와 전체적인 구조나 논리적 요소들 그리고 컴퓨터와 운영체계, 네트워크 및 기타 다른 개념들 간의 논리적 상호관계 등을 생각해내고 정의하는 등, 모든 곳에 적용되는 개념

⊙ 구성요소들이 연관성을 가진 구조

Module-02. 빅데이터 거버넌스 프레임워크

□ IT 거버넌스 체계 수립 프레임워크

▷ 구성요소 설명

⊙ STEP(수립 단계)

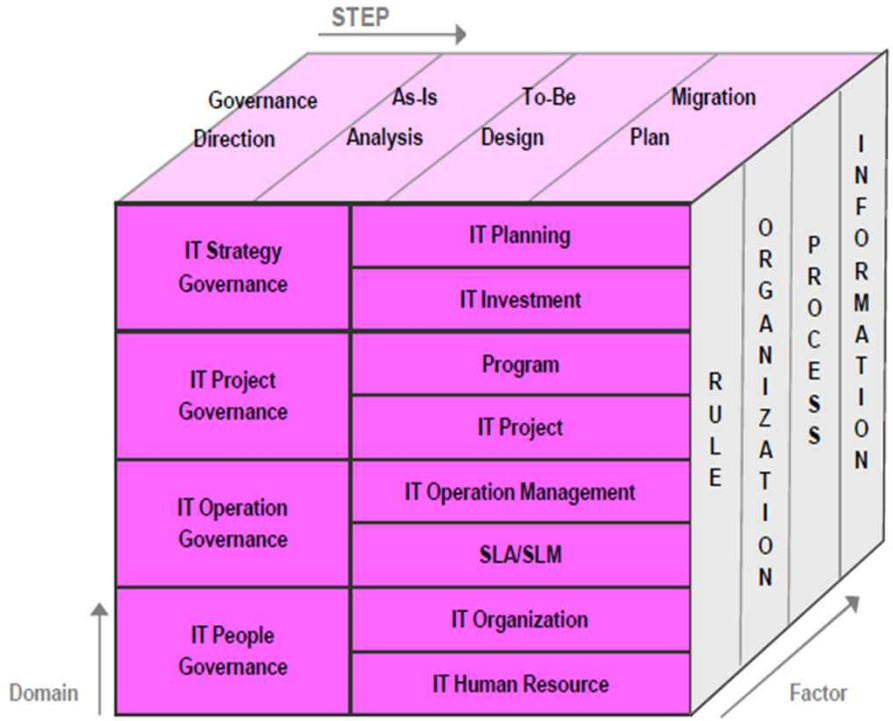
- 방향 설정, 현황 분석, 목표 설계, 이행 계획

⊙ Domain(영역)

- IT전략 거버넌스, IT프로젝트 거버넌스, IT운영 거버넌스, IT인력 거버넌스

⊙ Factor(요소)

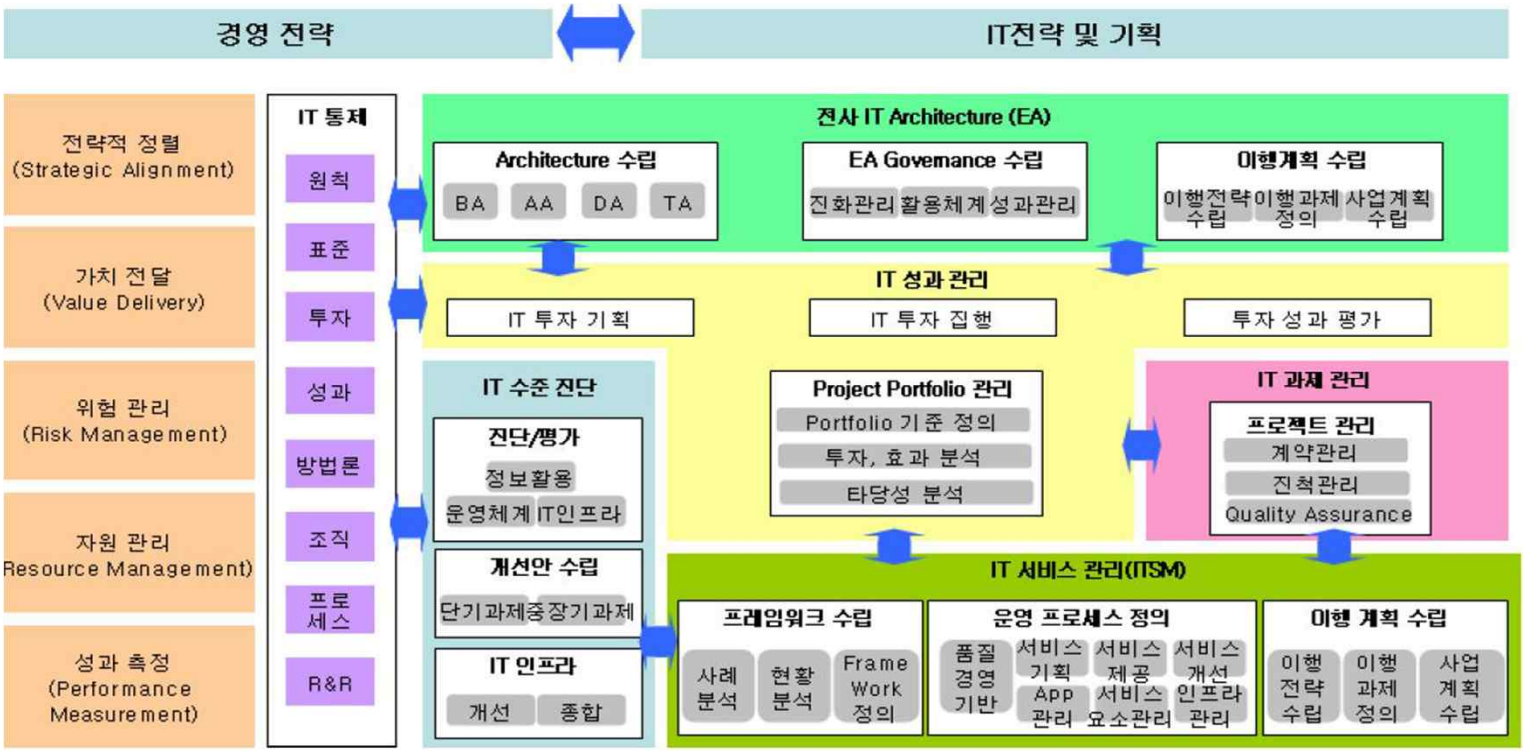
- 원칙, 조직, 프로세스, 정보



참고문헌 : IT Governance 소개 (인포레버, 2009.7)

Module-02. 빅데이터 거버넌스 프레임워크

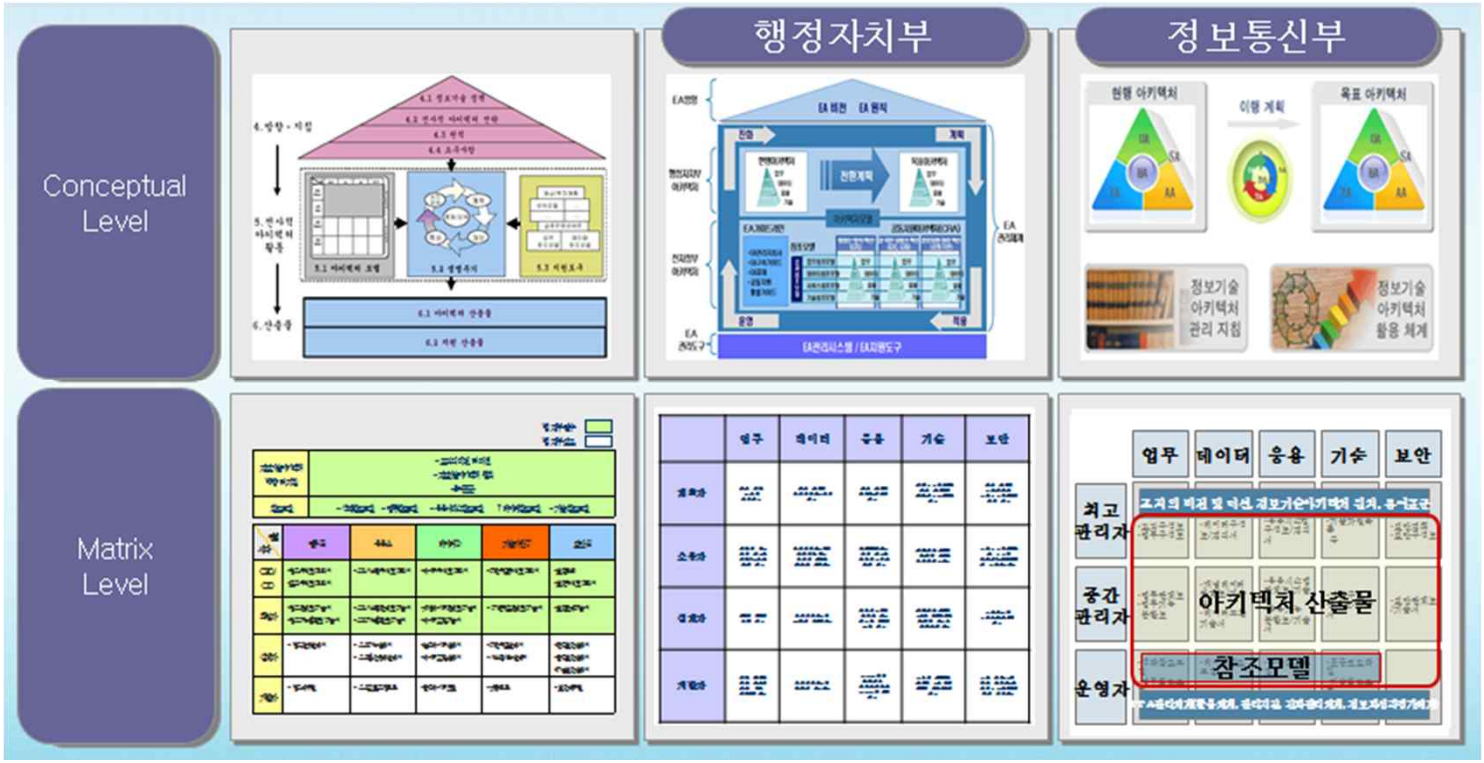
□ 삼성SDS IT 거버넌스 프레임워크



참고문헌 : IT Governance 소개 (인포레버, 2009.7)

Module-02. 빅데이터 거버넌스 프레임워크

□ EA(전사아키텍처) 거버넌스 프레임워크



참고문헌 : EA와 IT거버넌스 (E4Net)

Module-02. 빅데이터 거버넌스 프레임워크

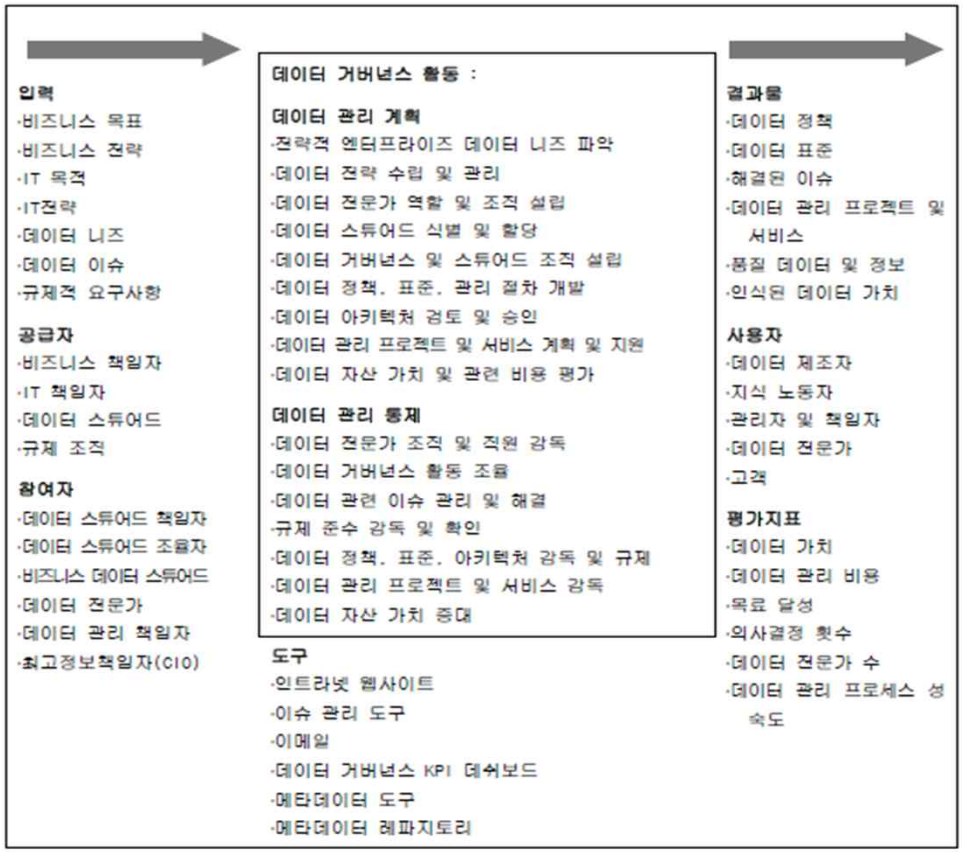
□ 데이터 거버넌스 프레임워크

▷ 프레임워크

⊙ 데이터 거버넌스 활동은 입력물에 대해 데이터 관리계획(프로세스)과, 관리통제(프로세스)로 이루어지는데, 입력물을 공급하는 공급자와, 데이터 거버넌스 활동을 수행하는 참여자가 있음

⊙ 데이터 거버넌스 활동으로 인한 결과물을 평가하는 평가지표가 있고, 결과물에 대한 사용자가 있음

⊙ 데이터 거버넌스 활동을 용이하게 하기 위하여 각종 도구가 존재함



참고문헌 : 데이터거버넌스 포럼운영 연구보고서_(2012.12)

Module-02. 빅데이터 거버넌스 프레임워크

□ 빅데이터 거버넌스 프레임워크

▷ 프레임워크 축

⊙ 빅데이터 유형 축

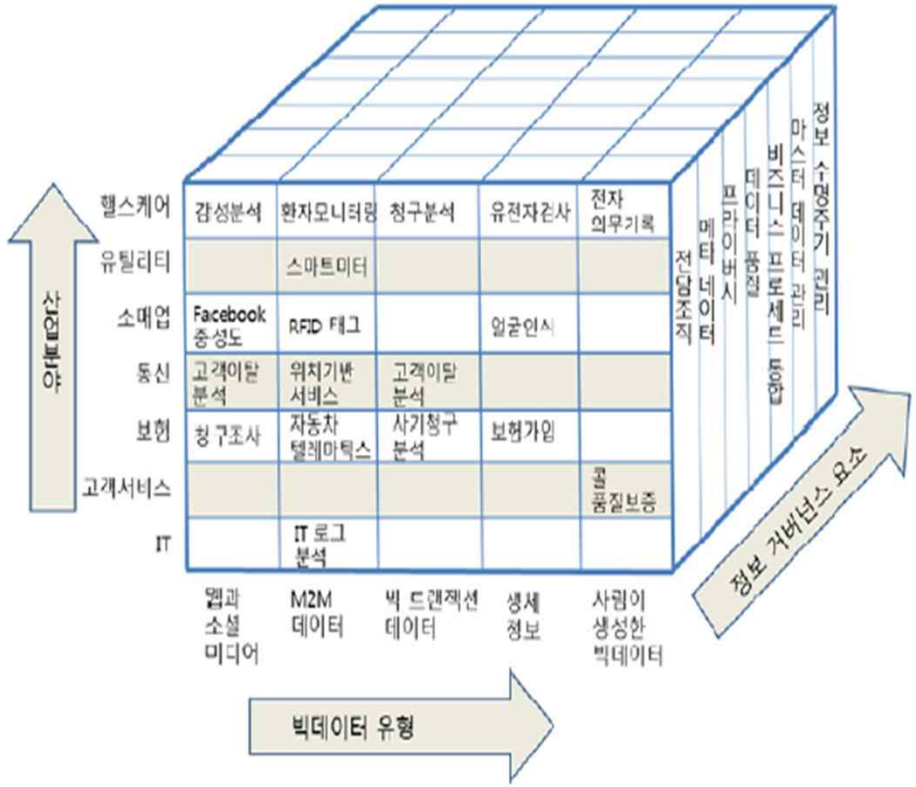
- 웹과 소셜미디어, M2M 데이터, 빅 트랜잭션 데이터, 생체정보, 사람이 생성한 데이터

⊙ 정보 거버넌스 축

- 조직, 메타데이터, 프라이버시, 데이터 품질, 비즈니스 프로세스 통합, 마스터 데이터 관리, 정보 수명주기 관리

⊙ 산업분야 축

- 헬스케어, 유틸리티, 소매업, 통신, 보험, 고객서비스, IT



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ IBM 빅데이터 플랫폼

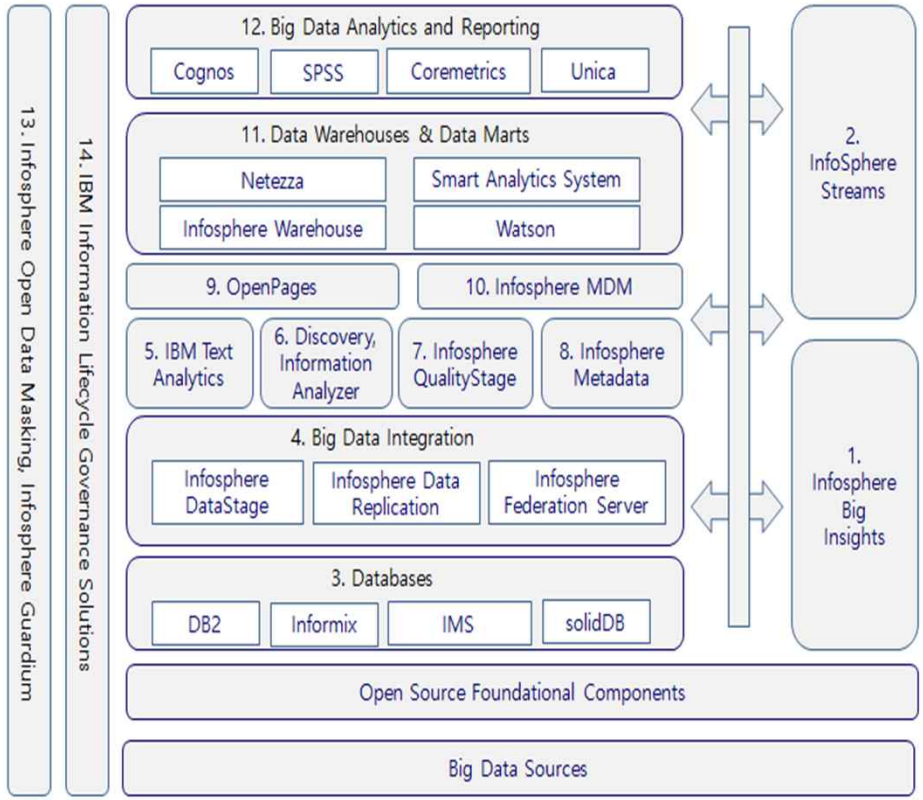
▷ 플랫폼 기능 및 특징

⊙ 자사의 빅데이터 플랫폼에서 Hadoop 지원 기능을 추가

⊙ IBM Infospere BigInsights는 IBM에서 나온 Hadoop 배포판임

⊙ 제공 기능

- 스트리밍 분석, Databases, 빅데이터 통합, 텍스트 분석, 빅데이터 디스커버리, 빅데이터 품질, 빅데이터 메타데이터, 정책관리, 마스터 데이터 관리, 데이터웨어하우스와 데이터 마트, 빅데이터 분석과 리포팅, 빅데이터 보안과 프라이버시, 빅데이터 수명주기 관리



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ IBM 빅데이터 플랫폼

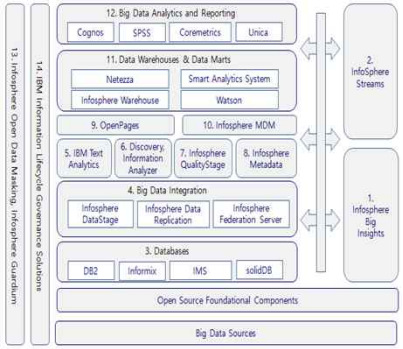
▷ 플랫폼 기능 상세 설명

⊙ 1. Infosphere BigInsights

- IBM에서 나온 Hadoop 배포판
- IBM Infosphere BigInsights는 기업 고객이 Hadoop을 사용할 수 있도록 개선한 제품
- IBM은 최근 이러한 개선된 기능을 IBM의 하둡 배포판이나 CDH(Cloudera's Distribution for Apache Hadoop) 패키지 모두에서 사용할 수 있을 것이라고 발표함
- 개선된 기능은 구조적 데이터와 비구조적 데이터를 모두 처리할 수 있는 기능을 제공하는 오픈소스인 Jaql, 텍스트 분석 엔진, 데이터 저장소 커넥터, 데이터 통합, 접근권한 인증 기능, 권한관리, Hadoop에 대한 Bigindex 기능, 사용자가 IBM Infospere BigInsights에 있는 데이터를 탐색하고 코드 실행없이 쿼리를 만들 수 있게 하는 기능이 대폭 확장되어 있음

⊙ 2. Infosphere Stream

- 생성되고 있는 데이터를 분석하기 위해 대량의 병렬처리 기술을 호라용할 수 있도록 지원
- 분석을 위해 대규모의 구조적, 비구조적, 반구조적 데이터를 디스크에 저장하는 것과 달리 IBM Infosphere Streams는 생성되고 있는 데이터를 디스크에 저장하는 과정없이 바로 분석이 가능



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ IBM 빅데이터 플랫폼

▷ 플랫폼 기능 상세 설명

⊙ 3. Databases

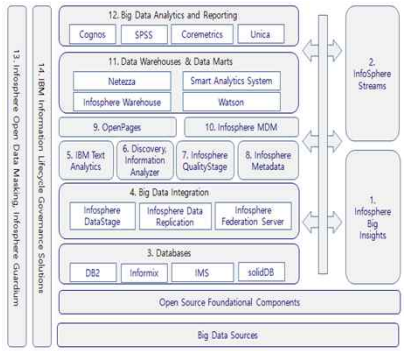
- IBM DB2, Informix, IMS, 메모리 데이터베이스인 solidDB를 포함해서 많은 OLTP(Online Realtime Processing) 데이터베이스를 제공

⊙ 4. Big Data Integration

- IBM Infosphere DataStage 버전 8.7은 Hadoop 분산파일 시스템에서 빅데이터를 접근하는 것을 지원하는 도구
- 새로운 Big Data File 스테이지를 추가하여 Hadoop으로부터, 혹은 Hadoop으로 다수 파일에 병렬로 읽고 쓰는 것을 지원함으로써 데이터가 공통의 변환과정으로 쉽게 병합될 수 있도록 함
- IBM Infosphere Data Replication은 변경되는 데이터의 식별 기능을 포함한 데이터 복제 지원
- IBM Infosphere Federation 서버는 데이터 가시화 지원

⊙ 5. Text Analytics

- IBM 텍스트 분석 기술은 IBM SPSS Text Analytics for Surveys, IBM Infosphere Streams 등 여러 제품과 도구에 내장되어 있음



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ IBM 빅데이터 플랫폼

▷ 플랫폼 기능 상세 설명

⊙ 6. Infosphere Information Analyzer와 Discovery

- 정보 거버넌스 프로젝트의 일부로 데이터 프로파일을 작성하는데 사용

⊙ 7. Infosphere QualityStage

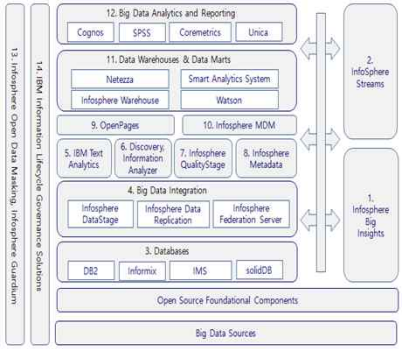
- 빅데이터 거버넌스 프로젝트의 하나로 데이터를 정제하는데 사용

⊙ 8. Infosphere Metadata

- IBM Infosphere Business Glossary, IBM Infosphere Metadata Workbench, IBM Cognos Business Viewpoint는 빅데이터 거버넌스 프로젝트의 일부로 비즈니스 메타데이터, 기술적 메타데이터 운영, 운영 메타데이터, 분석 메타데이터를 관리

⊙ 9. OpenPages

- GRC(governance, risk management and compliance) 플랫폼의 IBM OpenPages를 확장하여 정책정보 관리에 사용



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ IBM 빅데이터 플랫폼

▷ 플랫폼 기능 상세 설명

⊙ 10. Infosphere MDM

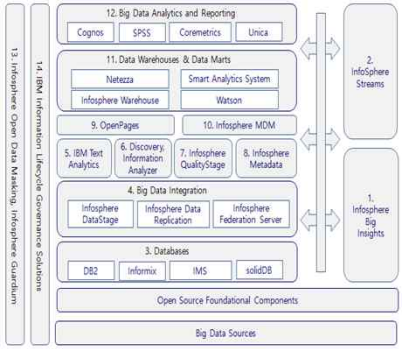
- IBM Infosphere MDM v10은 v10은 협업형, 표준형, 고급형, 엔터프라이즈 형의 4가지 버전으로 제공되어 마스터 데이터 관리를 수행

⊙ 11. Data Warehouses & Data marts

- IBM Netezza는 데이터 웨어하우스 통합지원 패키지로써 매우 큰 데이터 셋에 빠르게 쿼리하기 위해 데이터베이스 내에서 병렬 데이터 분석을 수행
- IBM Smart Analytics System은 Netezza가 인수되기 전부터 있었고, IBM 소프트웨어, 하드웨어, 스토리지 기반의 데이터 웨어하우스 통합지원 패키지
- IBM Infosphere Warehouse는 DB2에 기반을 둔 데이터 웨어하우징을 위한 소프트웨어 패키지

⊙ 12. Big Data Analytics and Reporting

- IBM Cognos는 리포팅, 대시보드, 성과표, 예산, 계획, 실시간 모니터링 위한 IBM의 주력 플랫폼
- IBM SPSS는 더 나은 예측모델을 얻기 위해 인구통계학적 요소, 선호도, 세분화된 정보를 텍스트, 소셜 미디어, 클릭 스트림과 같은 빅데이터와 결합시킬 수 있음
- IBM Coremetrics는 개인이 기업의 웹사이트와 소셜 미디어를 포함한 다른 형태의 디지털 공간과 어떻게 상호작용하고 있는 지에 대한 정보를 제공
- IBM Unica는 여러 채널에서의 캠페인 관리와 마케팅 자원관리를 지원



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ IBM 빅데이터 플랫폼

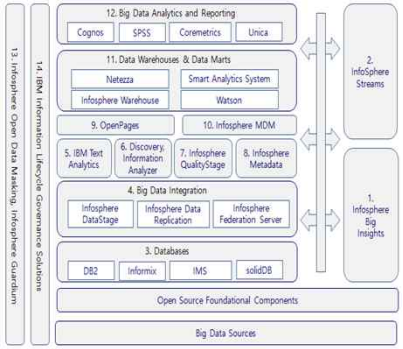
▷ 플랫폼 기능 상세 설명

⊙ 13. Infosphere Open Data Masking, Infosphere Guardium

- IBM은 빅데이터 보안과 프라이버시를 위해 개인정보법과 규제 준수사항 및 절차를 특정 데이터 소스나 데이터 카테고리에 연결시켜 목록으로 만드는 IBM Global retention Policy and Schedule Management, 민감한 데이터를 문맥상으로 맞는 사실적인 정보로 변형하는 다양한 데이터 변형기술을 사용할 수 있게 하는 IBM Infosphere Optim Data masking Solution, 각 처리 작업에서 누가, 무엇을, 언제, 어디서, 어떻게 했는지 등의 데이터베이스 활동을 모니터링하는 IBM Infosphere Guardium, 보안정보와 이벤트를 관리하는 IBM Tivoli Security Operations Manager 등이 있음

⊙ 14. IBM Information Lifecycle Governance Solutions

- 아카이브를 만드는 IBM Smart Archive, 법률 지원팀이 IT, 기록, 사업팀과의 협력과 같은 증거 관련 의무를 정의할 수 있게 해주고, 법률 관련 사건에서 많은 증거를 만드는데 드는 비용을 줄여주는 IBM eDiscovery Solution, 조직이 보유계획에 따라 기록을 관리할 수 있게 해주는 IBM Records and Retention Management, 실제적이고 적당한 크기의 데이터베이스를 만들기 위해 테스트 환경, 데이터 부분집합의 생성과 관리를 효과적으로 만들고, 민감한 데이터를 가리며, 테스트 결과 비교를 자동화하고 여러 복제된 데이터베이스를 유지하는 비용과 노력을 없애주는 IBM Infosphere Optim Test Data management Solution이 있음



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ 오라클 빅데이터 플랫폼

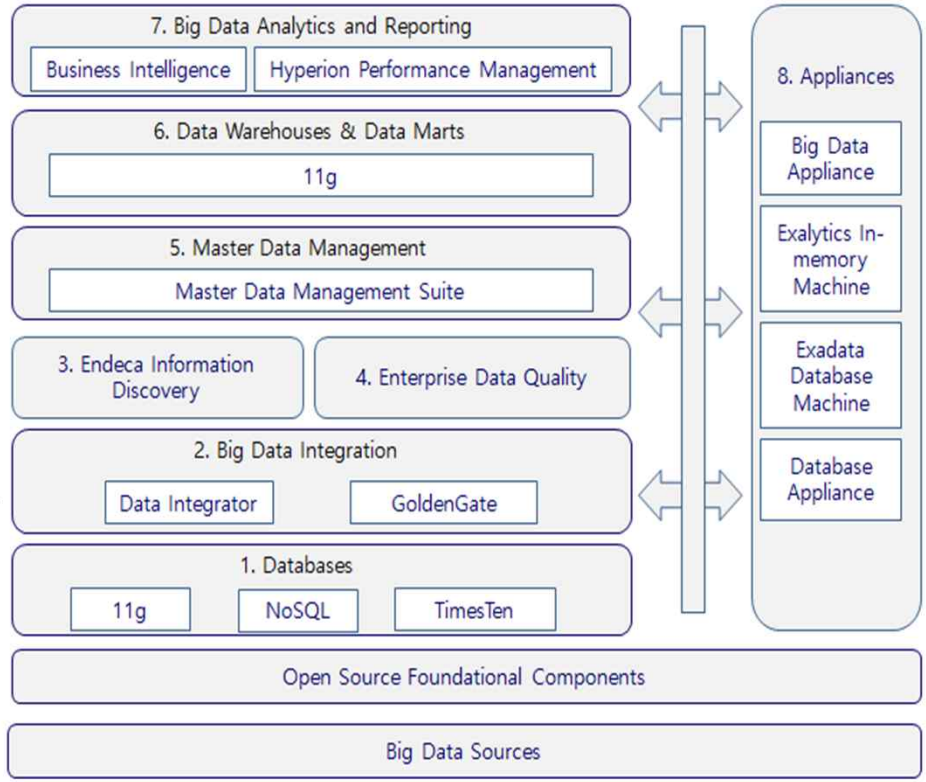
▷ 플랫폼 기능 및 특징

⊙ Oracle Database 11g에 Oracle Big Data Connectors를 이용하여 Hadoop간의 시너지 효과를 향상함

- Oracle Loader for Hadoop(OLH)
- Oracle Direct Connector for HDFS

⊙ 제공 기능

- 빅데이터 통합, 빅데이터 디스커버리, 빅데이터 품질, 마스터 데이터 관리, 데이터 웨어하우스와 데이터 마트, 빅데이터 분석과 리포팅, 통합제품(Oracle Database Appliance, Oracle Exadata Database Machine, Oracle Exalytics In-Memory Machine, Oracle Big Data Appliance

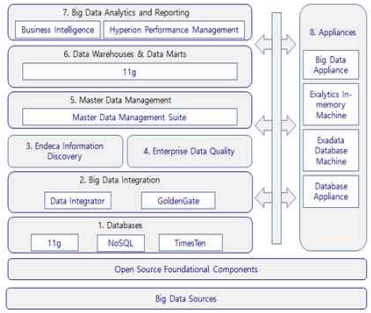


참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ 오라클 빅데이터 플랫폼

▷ 플랫폼 상세 기능 설명



⊙ 1. Databases
- Oracle Database 11g는 Oracle사의 주력 데이터베이스임. Oracle Big Data Connectors는 Oracle Database와 hadoop 간 시너지 효과를 높이는데, 이 커넥터들은 Oracle 빅데이터 제품이나 범용 Hadoop 클러스터에 설치가 가능하며, 맵리듀스 작업을 이용하여 Oracle RDBMS내에서 데이터를 로드하고 분석하는데 최적화된 데이터 셋을 만드는 Oracle Loader for Hadoop(OLH), Oracle 데이터베이스 내에서 HDFS 데이터에 매우 빨리 액세스를 가능하게 해주는 Oracle Direct Connector for HDFS 가 있음

⊙ 2. Big Data Integration
- Oracle Data Integrator Enterprise Edition은 Oracle Exadata Database machine에서 데이터 변형을 할 수 있도록 지원하며 Hadoop용 Oracle Data Integrator Application Adapter는 데이터 통합 개발자들이 Hadoop에 있는 데이터를 Oracle Data Integrator를 사용해서 쉽게 통합하고 변형할 수 있게 함

⊙ 3. Endeca Information Discovery
- 구조적, 비구조적 데이터용 데이터 검색, 디스커버리 플랫폼

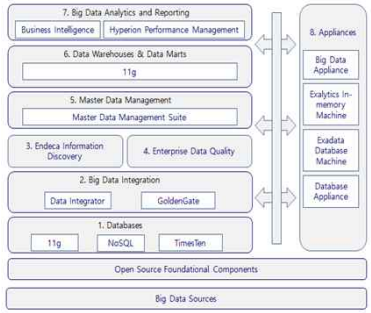
⊙ 4. Enterprise Data Quality
- 목적별로 여러가지 도구를 제공. 또한 Oracle Customer Hub와 Oracle product Hub에 통합되어 있어서 정제되고 표준화된 고객 및 제품정보가 마스터 데이터 관리 환경으로 로드될 수 있도록 지원

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ 오라클 빅데이터 플랫폼

▷ 플랫폼 상세 기능 설명



⊙ 5. Master Data management

- Oracle Master data Management Suite는 여러 모듈들로 구성되어 있는데, 고객에 대한 싱글 뷰를 제공하는 Oracle Customer Hub, 제품에 대해 싱글 뷰를 제공하는 Oracle product Hub, 고객, 경쟁사, 공급자와 관련된 내부, 외부 사이트에 대한 싱글 뷰를 제공하는 Oracle Site Hub, 공급자에 대한 싱글 뷰를 제공하는 Oracle Supplier Hub, 학생, 교직원, 동창, 직원, 다른 구성원에 관한 싱글 뷰를 제공하는 Oracle Higher education Constituent Hub 등이 있음

⊙ 6. Data Warehouse & Data Marts

- Oracle database 11g는 Oracle의 주력 데이터 웨어하우스, 데이터 마트 솔루션임

⊙ 7. Big Data Analytics & Reporting

- Oracle은 분석과 리포팅을 위한 여러가지 사항들을 제공하고 있는데, 리포팅, ad-hoc 쿼리, OLAP, 대시보드, 성과표 작업을 원활하게 해주는 Oracle Business Intelligence Tools, 전략관리, 계획, 예산, 예측, 수익성 관리, 결산, 리포팅을 지원하는 Oracle Hyperion Performance management 제품, HDFS에 저장된 많은 양의 데이터에 대하여 R 통계 모델을 활용하여 맵리듀스 처리를 할 수 있게 하는 Hadoop용 Oracle R Connector 등이 있음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

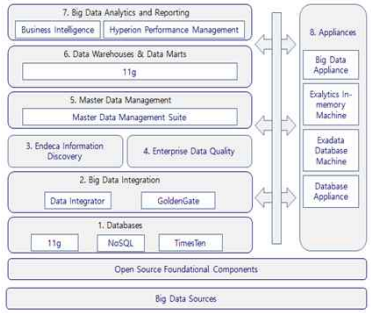
Module-03. 빅데이터 플랫폼

□ 오라클 빅데이터 플랫폼

▷ 플랫폼 상세 기능 설명

⊙ 8. Appliances

- Oracle은 최적화된 하드웨어와 소프트웨어 묶음들을 포함하는 'engineered systems'라는 통합 제품(Appliances)을 출시하였는데, 기업이 자사의 빅데이터를 쉽게 관리하게 하기 위한 Oracle Big Data Appliance, 분석 업무 위주인 기업 성과관리 애플리케이션의 성능을 향상시키기 위하여 Oracle Exalytics In-Memory Machine, OLTP와 데이터 웨어하우징을 위한 통합작업을 고성능으로 지원하기 위해서 제공되는 Oracle Exadata Database Machine, Oracle database, Enterprise 11g, Oracle Real Application Cluster(RAC)와 Oracle Linux가 포함되어 있되 중소-중견 기업을 타깃으로 하는 제품들이 있음



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

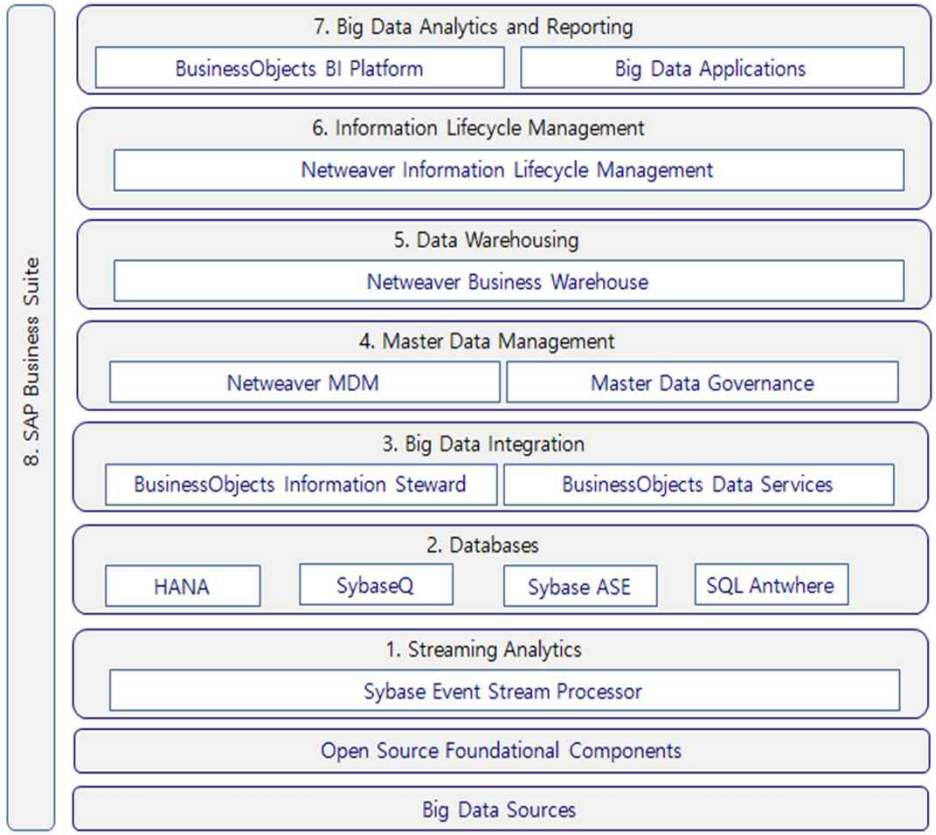
□ SAP 빅데이터 플랫폼

▷ 플랫폼 기능 및 특징

⊙ SAP은 자사의 빅데이터 플랫폼의 일부로 여러 기술들을 개발함

⊙ 제공 기능

- 스트리밍 분석, 데이터베이스, 빅데이터 통합, 마스터 데이터 관리, 데이터 웨어하우징, 정보 수명주기 관리, 분석과 리포팅, 비즈니스 애플리케이션



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ SAP 빅데이터 플랫폼

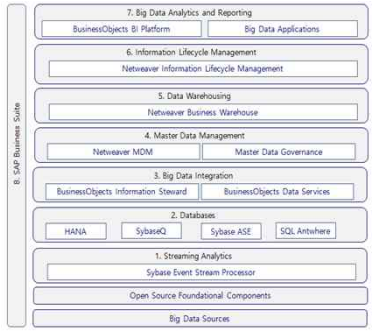
▷ 플랫폼 상세 기능 설명

⊙ 1. Streaming Analytics

- Sybase Event Stream Processor(ESP)는 자본시장과 이와 유사한 환경에서 스트리밍 데이터를 실시간으로 분석하기 위한 복합 이벤트 프로세싱 플랫폼

⊙ 2. Databases

- SAP HANA는 인메모리 상주 데이터베이스로서 기업에 설치된 대형 SAP ERP 환경에서 사용자들의 관심이 증가하고 있음
- HANA의 빅데이터 관련 기능
  . 비구조적, 반구조적 데이터에 대한 텍스트 검색
  . 오픈소스인 R 통계 패키지와의 호환성 제공
  . Hadoop과의 호환 기능
  . Amazon Web Service(AWS) 내에서 SAP HANA를 활용이 가능
- SAP Sybase IQ는 비즈니스 인텔리전스와 분석을 위한 칼럼 데이터베이스이며, 인-데이터베이스 자체 맵리듀스 작업을 지원
- SAP Server Enterprise(ASE)는 Sybase의 주력 관계형 데이터베이스. SAP Sybase SQL Anywhere는 모바일 디바이스를 위한 데이터베이스임



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

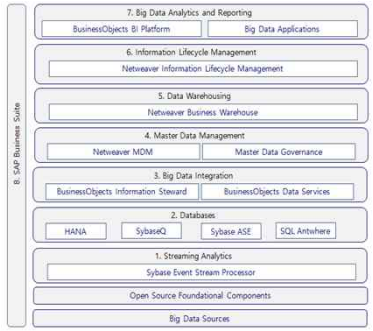
□ SAP 빅데이터 플랫폼

▷ 플랫폼 상세 기능 설명

⊙ 3. Big Data Integration
- SAP BusinessObjects Data services는 ETL, 데이터 품질, 데이터 프로파일링, 메타데이터, 텍스트 분석과 같은 통합된 기능을 가지고 있음. 데이터를 HANA로 옮기는 작업과 Hadoop과의 계획된 통합을 지원
- SAP BusinessObjects Information Steward는 데이터 관리자가 데이터 품질 점검표, 데이터 인증 비즈니스 규칙, 비즈니스 정의, 메타데이터를 관리하고 감독할 수 있게 하는 기능을 제공

⊙ 4. Master Data Management
- SAP NetWeaver Master Data Management는 고객, 제품, 벤터와 같은 핵심 도메인에 대한 싱글 뷰를 제공
- SAP master Data Governance는 SAP Busienss Suite, SAP NetWeaver Master Data Management, SAP NetWeaver Process Orchestration을 통합하여 비즈니스 절차면에서 정보 거버넌스 정책을 강화

⊙ 5. Data Warehousing
- SAP NetWeaver Business Warehouse는 SAP의 주력 데이터 웨어하우징 제품이며 SAP HANA도 해당 제품을 지원



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

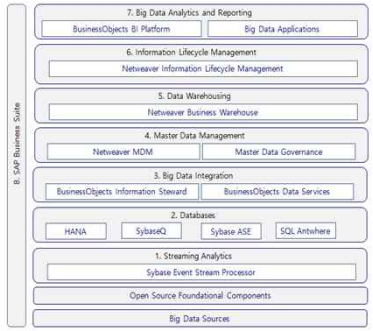
□ SAP 빅데이터 플랫폼

▷ 플랫폼 상세 기능 설명

⊙ 6. Information Lifecycle Management
- SAP NetWeaver Information Lifecycle Management는 조직이 비즈니스 요구나 규제 요건에 따라 데이터를 아카이브로 만들고, 데이터에 대한 여러 유형의 보유 정책을 세울 수 있게 함

⊙ 7. Big Data Analytics and Reporting
- SAP BusinessObjects Business Intelligence 플랫폼은 회사의 주력 비즈니스 인텔리전스 분석 패키지

⊙ 8. SAP Business Suite
- SAP 비즈니스 패키지에는 다양한 기능과 산업을 위한 SAP의 유명한 애플리케이션이 포함됨



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

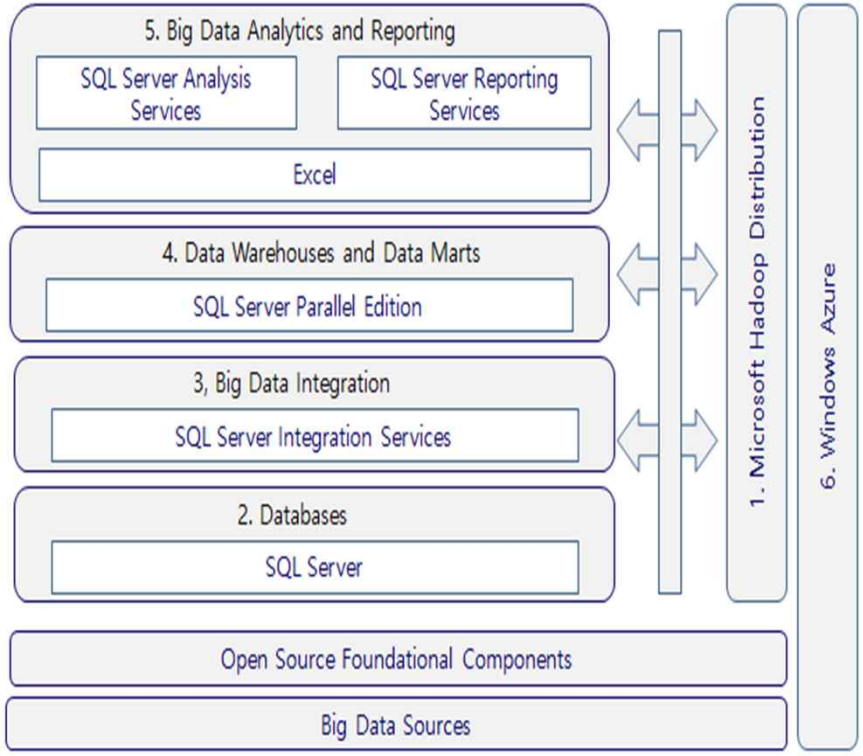
□ Microsoft 빅데이터 플랫폼

▷ 플랫폼 기능 및 특징

⊙ Microsoft사는 윈도우 서버상에 설치하거나 Window Azure의 클라우드 기반 서비스로 제공되는 Hadoop 배포판의 CTP(Community Technology Preview)를 공개 예정

⊙ 제공 기능

- 데이터베이스, 빅데이터 통합, 데이터 웨어하우스와 데이터마트, 빅데이터 분석과 리포팅, 클라우드

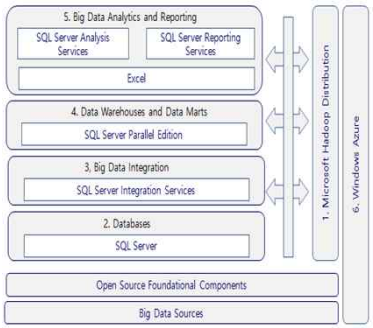


참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ Microsoft 빅데이터 플랫폼

▷ 플랫폼 상세 기능 설명



⊙ 1. Microsoft Hadoop Distribution

- Microsoft사는 Window Server 상에 설치하거나 Windows Azure의 클라우드 기반 서비스로 제공되는 Hadoop 배포판의 CTP(Community Technology Preview)를 공개할 예정이라고 함

⊙ 2. Databases

- Microsoft사는 두 환경간에 데이터를 이동하기 위해 Microsoft SQL Server를 위한 쌍방향 Hadoop 커넥터를 출시

⊙ 3. Big Data Integration

- Microsoft SQL Server Integration Services(SSIS)는 데이터 통합을 위한 Microsoft의 ETL 묶음임.
- SSIS는 Microsoft의 나머지 빅데이터 플랫폼과 밀접하게 연관되어 있음

⊙ 4. Data Warehouses & Data Marts

- Microsoft SQL Server Parallel Data Warehouse는 2008년 DATAllegro 인수를 토대로 한 거대한 병렬처리 데이터 웨어하우스 엔진임. Microsoft는 두 환경간에 데이터를 이동하기 위해 SQL Server Parallel Data warehouse용 쌍방향 Hadoop 커넥터를 발표

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ Microsoft 빅데이터 플랫폼

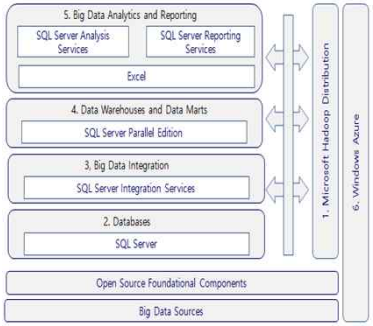
▷ 플랫폼 상세 기능 설명

⊙ 5. Big Data Analytics & reporting

- Microsoft SQL server Reporting Services(SSRS)는 사용자들이 리포트를 작성할 수 있는 기능을 제공
- Microsoft SQL Server Analysis Services(SSAS)는 Microsoft의 주력 온라인 분석처리(OLAP)와 데이터마이닝 도구임

⊙ 6. Windows Azure

- Microsoft Windows Azure는 여러 호스키드 데이터 관리솔루션을 제공하는 클라우드 플랫폼으로 데이터베이스와 리포팅 및 분석을 위한 도구들을 포함. 또한 Microsoft Windows Azure는 Hosted Hadoop Framework를 가지고 있음



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

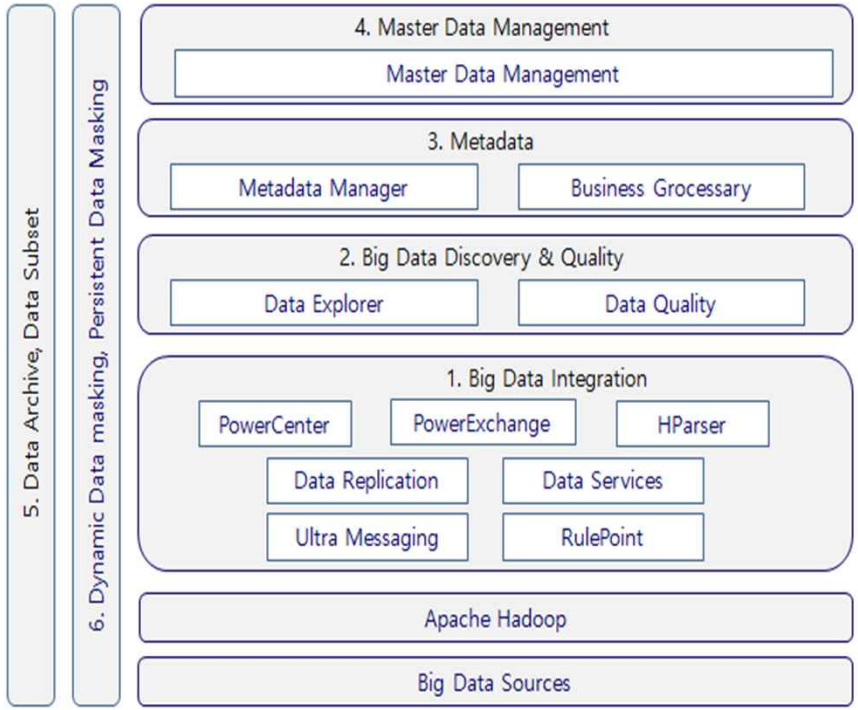
□ Informatica 빅데이터 플랫폼

▷ 플랫폼 기능 및 특징

⊙ Informatica 9.5를 출시하면서 Hadoop과 빅데이터에 대한 적극적인 지지 표시함

⊙ 제공 기능

- 빅데이터 통합, 빅데이터 디스커버리와 품질, 메타데이터, 마스터 데이터 관리, 빅데이터 수명주기 관리, 빅데이터 보안과 개인정보보호, 클라우드



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ Informatica 빅데이터 플랫폼

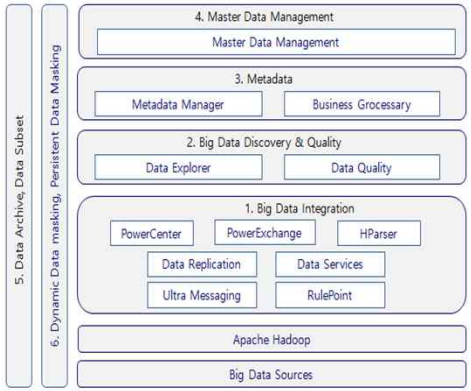
▷ 플랫폼 상세 기능 설명

⊙ 1. Big Data Integration

- Informatica의 빅데이터 통합 포트폴리오에는 Informatica의 주력 ETL 도구로써 Twitter, Facebook, LinkedIn에 대한 커넥터를 가지고 있는 PowerCenter, hdfs로 소스시스템의 데이터를 가져오고 HDFS에서 비즈니스 인텔리전스와 데이터 웨어하우징 환경으로 데이터를 옮기는 Hadoop 어댑터를 위한 PowerExchange, Hadoop에 최적화된 변환 도구로서 여러 비구조적, 반구조적 형식의 데이터 파싱을 지원하는 Informatica Hparser, Hadoop으로의 대량 데이터 복제 기능이 제공되는 Informatica Data replication, 데이터 가상화 도구인 Informatica Data Services, 메시지 미들웨어인 Informatica Ultra Messaging1ID, 실시간으로 비즈니스 분석을 수행하는 복합 이벤트 처리 도구인 Informatica RulePoint가 있음

⊙ 2. Big Data Discovery & Quality

- 빅데이터 디스커버리와 품질 기능을 수행하는 요소에는 데이터 관리자와 비즈니스 분석가가 빅데이터를 포함한 모든 형태의 데이터에서 이상치를 찾을 수 있게 하는 데이터 프로파일링 도구인 Informatica Data Explorer, 자료 대조와 글로벌 주소 정제 기능을 제공하는 Informatica Data Quality가 있음



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-03. 빅데이터 플랫폼

□ Informatica 빅데이터 플랫폼

▷ 플랫폼 상세 기능 설명

⊙ 3. Meta Data

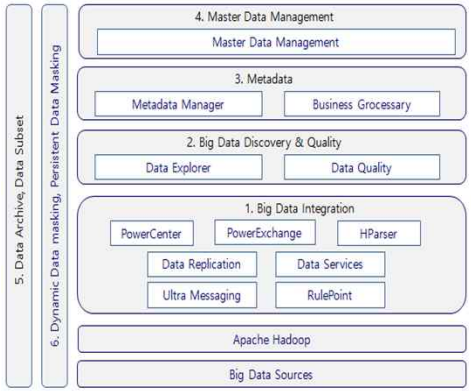
- 메타데이터 기능을 수행하는 요소로는 데이터 계보, 비즈니스 용어관리 기능을 제공하는 Informatica Metadata manager & Business Glossary가 있음

⊙ 4. Master Data management

- 마스터 데이터관리 기능을 수행하는 요소에는 다중 도메인에서의 구현들을 관리하고, 과거에는 기록이 어땠고 누가 그 기록을 수정했는지에 대한 데이터 타임라인 기능을 제공하는 Informatica Master Data management가 있습니다.

⊙ 5. Data Archive, Sata Subset

- 빅데이터 수명주기 관리 기능을 수행하는 요소로는 정책을 기반으로 아카이브를 만들고, 테스트 데이터를 관리하는 강력한 ILM(Information Lifecycle Management) 플랫폼을 개발했는데, 조직이 빅데이터를 낮에 다시 가져올 것을 대비하여 더 저렴한 스토리지로 옮기게 해주는 Informatica Data archive가 있음



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

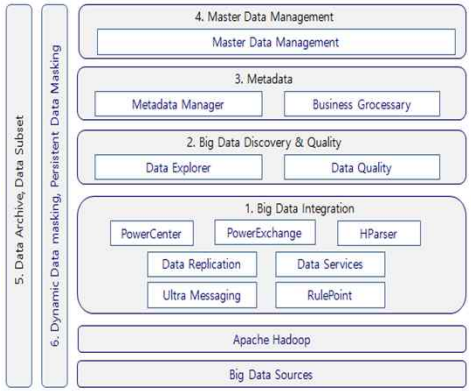
Module-03. 빅데이터 플랫폼

□ Informatica 빅데이터 플랫폼

▷ 플랫폼 상세 기능 설명

⊙ 6. Dynamic Data Masking, persistent Data Masking

- 빅데이터 보안과 개인정보 보호 기능을 수행하는 요소로 Informatica Dynamic Masking은 생산중이나 생산 환경과 비슷한 환경에서의 민감한 데이터를 가립니다. Informatica Persistent Data Masking은 테스팅과 같은 생산 환경이 아닌 곳에서의 민감한 데이터를 가림



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-04. 빅데이터 참조아키텍처

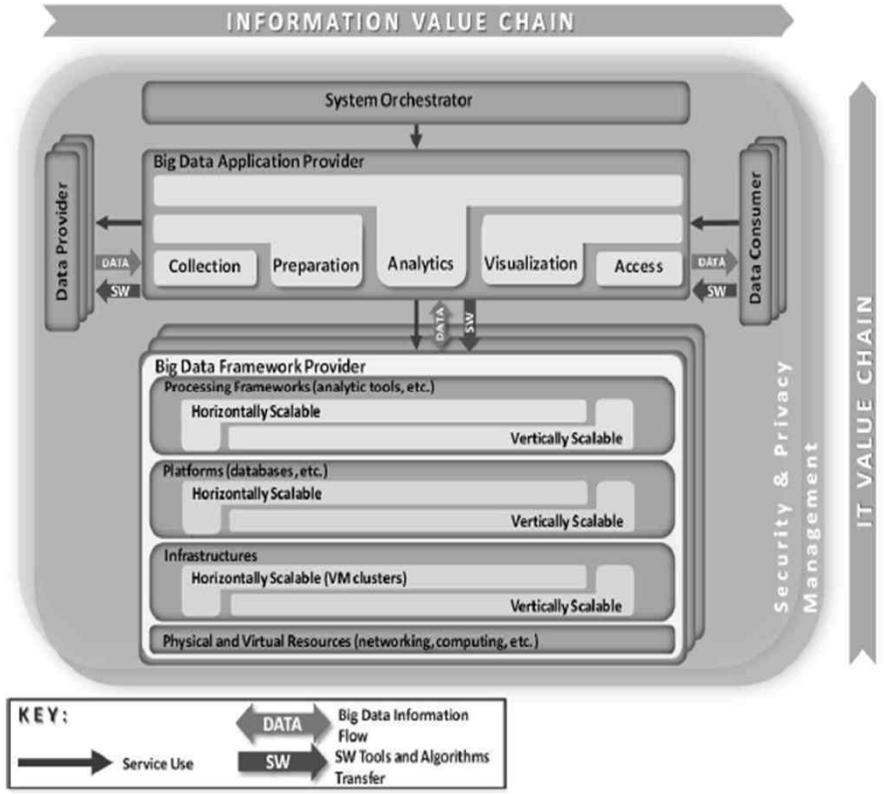
□ NIST 빅데이터 참조아키텍처

▷ 참조아키텍처 기능 및 특징

⊙ 미국 국립표준기술연구소(NIST, National Institute of Standards and Technology)

⊙ 특징

- NIST의 빅데이터 참조아키텍처는 IT와 정보의 2가지 가치 사슬(value chain) 축으로 구성. 수평 축의 정보의 흐름을 통한 가치는 데이터의 수집, 통합, 분석, 그리고 결과의 사용을 통해 생성되며, 수직 축은 네트워크, 인프라, 플랫폼, 응용 서비스 등 IT 서비스 제공 및 활용을 통해 생성됨을 나타내고 있음
- NIST의 빅데이터 참조아키텍처는 데이터 제공자, 빅데이터 어플리케이션 제공자, 빅데이터 프레임워크 제공자, 데이터 소비자, 시스템 오케스트레이터의 5가지 핵심 컴포넌트와 보안 및 개인정보, 관리의 2가지 부가 컴포넌트로 구성



참고문헌 : 국제 표준화 · 시험인증 회의 참가보고 (TTA, 2014.3)

Module-04. 빅데이터 참조아키텍처

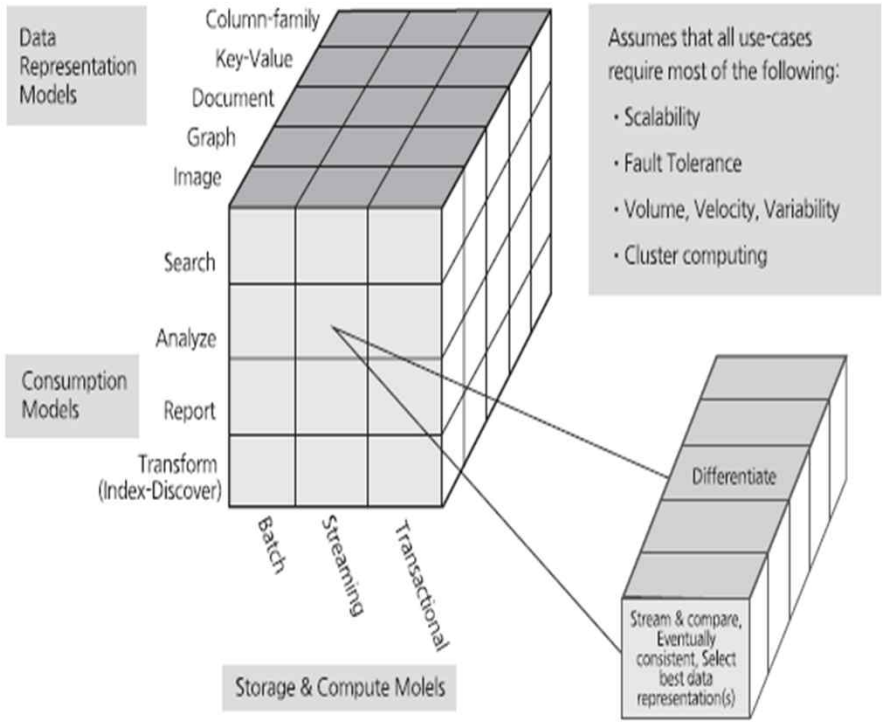
□ HP 빅데이터 참조아키텍처

▷ 참조아키텍처 기능 및 특징

⊙ 데이터 표현 모델 축, 소비 모델 축, 저장과 계산 모델 축

⊙ 제공 기능

- 데이터 표현 모델(column-family, Key-value, Document, Graph, Image), 소비 모델(검색, 분석, 리포트, 변환), 저장과 계산 모델(뱃치, 스트리밍, 트랜잭션성)



참고문헌 : 국제 표준화 · 시험인증 회의 참가보고 (TTA, 2014.3)

Module-04. 빅데이터 참조아키텍처

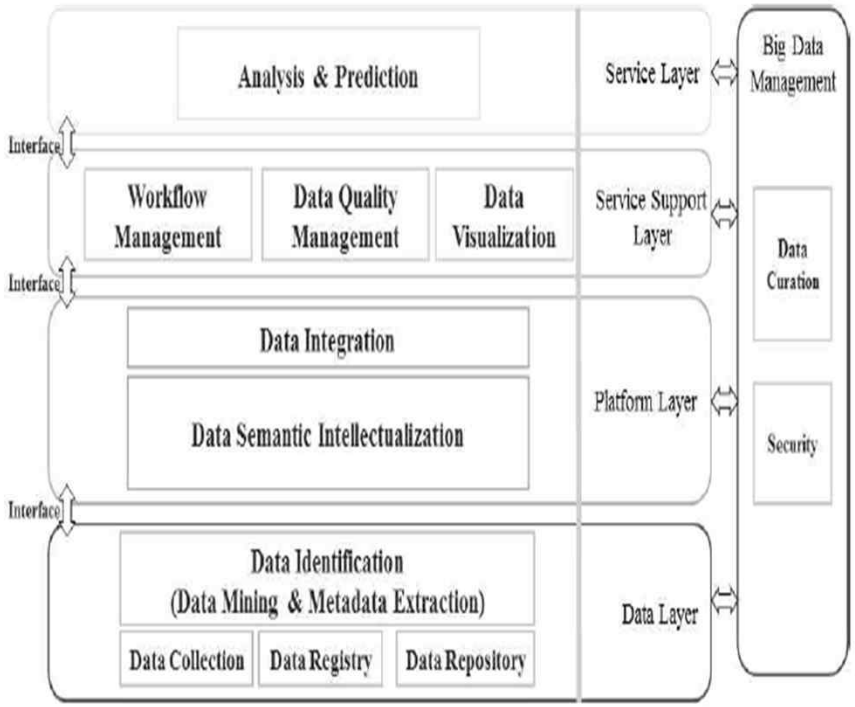
□ JTC 1 SC32 빅데이터 참조아키텍처

▷ 참조아키텍처 기능 및 특징

⊙ JTC 1 SC32는 데이터 관리 및 교환과 관련된 공적 표준을 개발하는 위원회로써 빅데이터 참조아키텍처에 대한 초안을 데이터, 플랫폼, 서비스 지원, 서비스로 구성된 4개의 수평적 레이어와 이들을 지원하는 1개의 빅데이터 관리 레이어로 구성됨

⊙ 아키텍처 설명

- 데이터 레이어(Data Layer)는 데이터를 수집, 저장, 레지스트링 하며, 이들을 식별하는 기능을 제공하며, 플랫폼 레이어 (Platform Layer)는 데이터에 대한 통합과 데이터를 의미를 기반으로 지식화하는 기능 수행. 서비스 레이어(Service Layer)는 분석 및 예측 등의 지식 서비스를 제공하며, 서비스 지원 레이어 (Service Support Layer)는 플랫폼과 서비스 레이어에 대한 워크플로우, 데이터 품질, 데이터 시각화 등의 기능 제공. 빅데이터 관리 레이어는 각 레이어에 대한 데이터 큐레이션 및 보안 기능을 제공하도록 구성



참고문헌 : 빅데이터 에코시스템 기반의 참조아키텍처 개발, ETRI(이강찬 외)

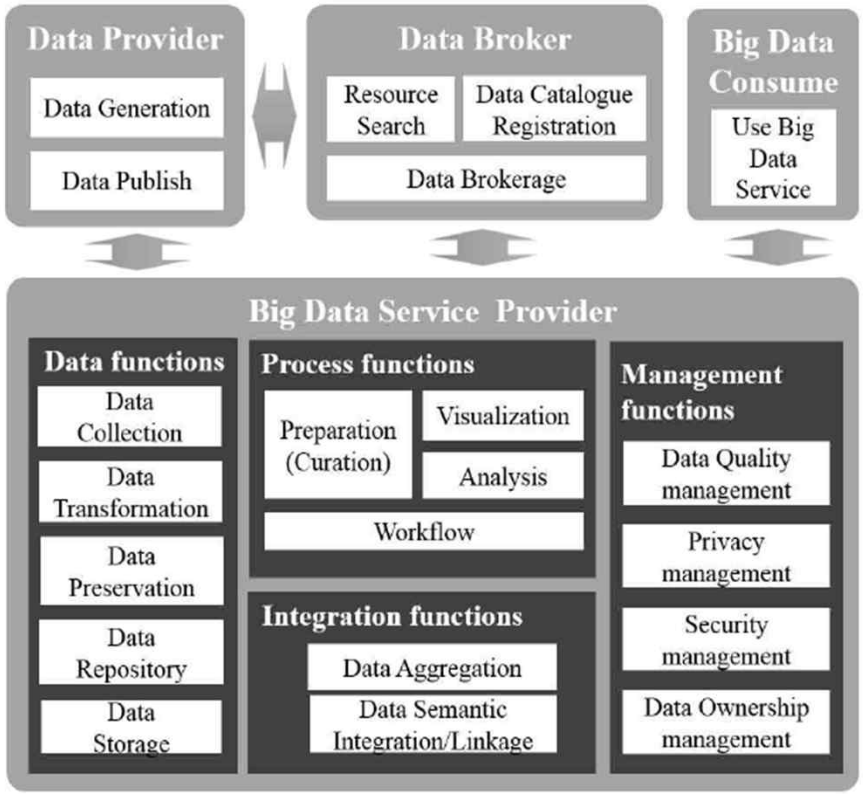
Module-04. 빅데이터 참조아키텍처

□ 빅데이터 에코시스템 기반의 참조아키텍처

▷ 참조아키텍처 기능 및 특징

⊙ 아키텍처 설명

- 빅데이터 제공자는 데이터를 생성하고 이를 데이터 중계자에 등록하는 기능을 수행, 데이터 서비스 제공자에게 데이터에 대한 접근성을 제공
- 데이터 중계자는 시스템 또는 온라인상의 가용한 데이터 자원을 검색하고 이를 카탈로그 형태로 등록함으로써 빅데이터 서비스 제공자가 이를 이용하여 데이터에 접근할 수 있는 정보를 취득
- 빅데이터 소비자는 빅데이터 서비스를 이용하는 주체로써 빅데이터 서비스 제공자를 통해 제공되는 기능을 소비하는 역할을 수행
- 빅데이터 서비스 제공자는 개별 기능들을 데이터 기능, 처리기능, 통합기능, 관리 기능의 4가지 기능 블록으로 구분하였음



참고문헌 : 빅데이터 에코시스템 기반의 참조아키텍처 개발, ETRI(이강찬 외)

Module-04. 빅데이터 참조아키텍처

□ 빅데이터 참조아키텍처

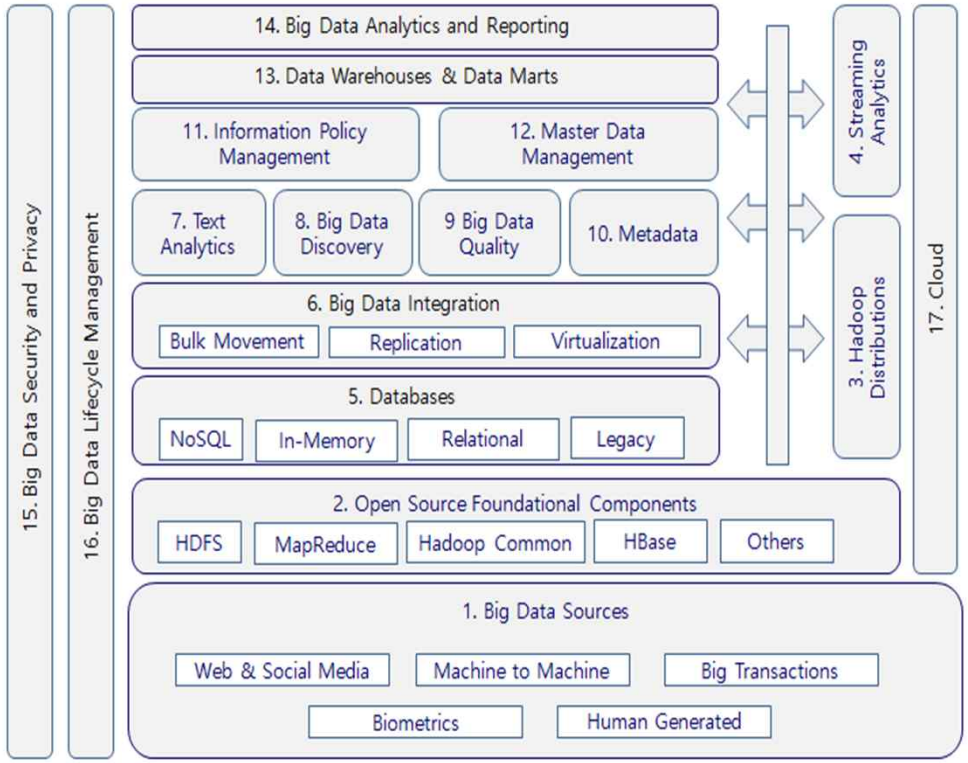
▷ 참조아키텍처 기능 및 특징

⊙ 조직의 할일

- 조직은 빅데이터를 현재 기술인프라에 통합시켜야 함
- 빅데이터 거버넌스는 소프트웨어 도구(툴)을 사용해야 함
- 거버넌스 도구는 넓은 빅데이터 플랫폼 맥락에서 평가
- 빅데이터 거버넌스의 인적, 절차적 차원을 활성화시키기 위해 소프트웨어 도구를 사용해야 함
- 빅데이터 거버넌스를 위해 도구를 도입하는 경우, 넓은 빅데이터 플랫폼 맥락에서 평가되어야 함

⊙ 제공 기능

- 빅데이터 소스, 오픈소스 진영의 컴포넌트들, Hadoop 배포판, 스트리밍 분석, 데이터베이스, 빅데이터 통합, 텍스트 분석, 빅데이터 디스커버리, 빅데이터 품질, 빅데이터 메타데이터, 정보정책관리, 마스터 데이터관리, 데이터 웨어하우스와 데이터마트, 빅데이터 분석과 리포팅, 빅데이터 보안과 프라이버시, 빅데이터 수명주기 관리, 클라우드



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-04. 빅데이터 참조아키텍처

□ 빅데이터 참조아키텍처

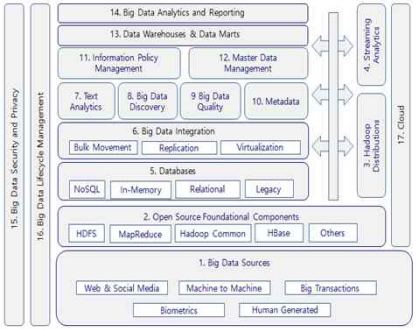
▷ 참조아키텍처 상세 기능 설명

⊙ 1. Big Data Source
- 웹과 소셜 미디어 데이터, M2M(Machine-to-machine) 데이터, 많은 양의 전자상거래 데이터, 생물 측정 통계학, 사람이 생성하는 데이터 등이 있으며 구조적, 반구조적, 비구조적인 형태로 되어 있음

⊙ 2. Open Source Foundational Component
- HDFS(Hadoop 분산 파일시스템), 맵리듀스, Hadoop Common, Hbase, Hive 등이 있음

⊙ 3. Hadoop Distributions
- 사용자들이 쉽게 Hadoop 관련 기술을 사용하도록 벤더들 자체가 자체적으로 개발한 소프트웨어 패키지

⊙ 4. Streaming Analytics
- 스트리밍 분석은 생성되는 데이터를 분석하는 대량 병렬처리 능력을 제공하는 일련의 기술을 말함. 스트리밍 분석을 빅데이터 솔루션에 통합하면 입력되는 데이터를 시간이 많이 걸리는 디스크에 저장하기 전에 필터링하고 연관성을 찾을 수 있으므로 응답시간을 줄일 수 있게 됨

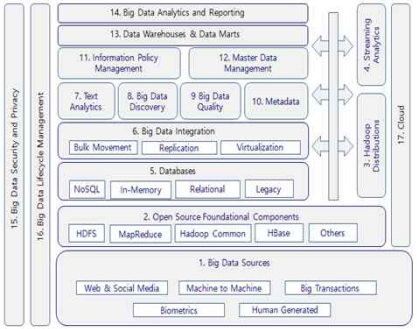


참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-04. 빅데이터 참조아키텍처

□ 빅데이터 참조아키텍처

▷ 참조아키텍처 상세 기능 설명



⊙ 5. Databases
- 고정된 테이블 형식을 필요로 하지 않으며, 조인 연산도 지원하지 않으며 데이터의 일관성보다는 확장성이 보장되는 읽고 쓰는 작업이 최적화되어 있는 NoSQL(Not Only SQL), 주기억장치를 데이터 저장용으로 사용하며 디스크에 데이터를 저장하는 기존의 DBMS와는 달리 인-메모리 데이터베이스는 속도에 최적화되어 있어서 조직이 엄청난 양의 빅데이터를 빠르게 처리하려고 할 때 사용되는 In-Memory 데이터베이스, 분산 컴퓨팅의 핵심 기술을 포함하고 있는 오라클 등의 관계형 데이터베이스, 이미 많은 양의 데이터를 다루고 있는 레거시 데이터베이스로 구성됨

⊙ 6. Big Data Integration
- 다수의 데이터 소스로부터 데이터를 추출, 변형, 타킷 데이터베이스로의 로딩을 담당하는 ETL 도구 등이 포함되는 대규모 데이터 이동, 데이터베이스 일부를 한 환경에서 다른 환경으로 복사하고 복사된 데이터를 원래 소스와 동기화되게 하는 과정인 데이터 복제, 두 개 혹은 그 이상의 물리적으로 다른 위치에 있던 데이터를 연결함으로써 마치 데이터가 동일한 장소에 있는 것처럼 만드는 데이터 가상화의 요소로 구성되어 있음

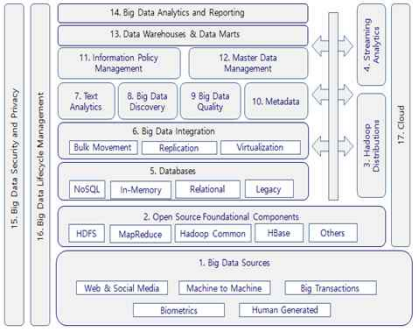
⊙ 7. Text Analytics
- 비구조적 텍스트 데이터로부터 유용한 지식을 뽑아내고 이 지식을 의사결정을 돕는데 사용. 이 과정에는 핵심 개념과 감성 및 트랜드 분석 등이 포함되는데 텍스트 분석 결과는 예측 분석을 위한 모델에 통합될 수 있음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-04. 빅데이터 참조아키텍처

□ 빅데이터 참조아키텍처

▷ 참조아키텍처 상세 기능 설명



⊙ 8. Big Data Discovery

- 데이터 발견과 프로파일링 작업이 이에 포함됨. 여러 소스의 데이터를 대상으로 모든 칼럼들의 상호비교를 수행함으로써 데이터 소스들 사이의 오버랩 기준치를 제정하는 Cross-source column overlap analysis, 동시에 여러 개의 데이터 소스를 대상으로 서로 대응되는 키의 품질에 대한 가설을 세우고 테스트하는 Matching key prototypes, 두 개의 구조적 데이터셋간의 복잡한 비즈니스 법칙을 발견하는 것을 자동화한 Transformation rule discovery, 서로 대응되는 키를 찾고 두 데이터 소스간의 키를 통계적으로 검증하는 Automatic matching key discovery, 동일한 논리적 행에 대하여 여러 데이터 소스간의 데이터 미리보기를 제공함으로써 이를 통하여 분석가들이 비즈니스 규칙에 맞는 값뿐 아니라 맞지않는 이상치를 발견하게 해주는 Cross-source data preview, 고객과 같이 관련된 목표들을 논리적으로 묶어서 완전한 비즈니스 목표를 정의함으로써 데이터 통합, 마스터 데이터 관리, 데이터 웨어하우징, 시범 데이터 관리 그리고 데이터 아카이브 만들기와 같은 정보 중심 사업에서 중요한 역할을 수행하는 Business object creation, 열,키, 소스, 교차-도메인 수준에서 데이터를 포괄적으로 이해하게 해주는 Deep profiling capabilities가 있음

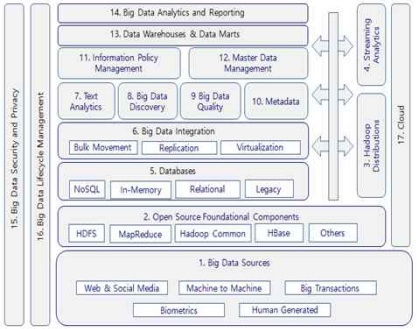
⊙ 9. Big Data Quality

- 데이터 품질관리는 조직 데이터의 품질과 무결성을 검사하고 개선하는 방법에 관한 활동으로써 데이터 표준화, 매칭, 생존, 시간에 따른 품질 모니터링이 있습니다. 빅데이터 품질은 기존 시스템과 달리 구조적, 반구조적, 비구조적 형태의 정보가 실시간적으로 처리되어야 함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

□ 빅데이터 참조아키텍처

▷ 참조아키텍처 상세 기능 설명



⊙ 10. Big Data Metadata
- 조직이 관리하는 데이터의 특징을 기술한 정보로써 데이터의 이름, 위치, 중요성의 정도, 품질, 기업에게 주는 가치, 다른 데이터와의 관련성 등을 포함

⊙ 11. Information Policy management
- 데이터 품질, 메타데이터, 프라이버시, 정보 수명주기관리와 관련된 정책을 기록
- 데이터 관리자, 실제 데이터를 활용하는 데이터 스폰서, 데이터 책임자와 같은 역할과 책임을 할당
- 데이터 정책에 부합하는지를 모니터링
- 데이터 이슈에 대해 허용 가능한 선을 정의
- 오래 지속되고 비즈니스의 여러 기능과 라인에 영향을 미치는 것을 집중 관리

⊙ 12. Master Data Management
- 빅데이터에 대한 이해를 바탕으로 마스터 데이터를 보강한다든가 소셜 미디어 감성분석을 마스터 데이터와 연계하여 어떤 고객들이 회사 제품에 더 긍정적인지를 알 수가 있게 됨

⊙ 13. Data Warehousing & Data Marts
- 조직들이 빅데이터를 받아들임에 따라 점점 Hadoop과 NoSQL 기술을 기존의 데이터 웨어하우징 환경과 통합하는 혼합 접근법을 많이 따르게 됨

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-04. 빅데이터 참조아키텍처

□ 빅데이터 참조아키텍처

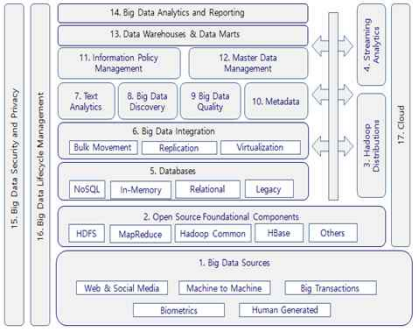
▷ 참조아키텍처 상세 기능 설명

⊙ 14. Big Data Analytics & reporting
- 빅데이터를 가시화하고 리포팅, 빅데이터를 기반으로 예측 모델을 설정하는 것을 가능하게 해줌.
- 고객이 여러 광고 네트워크, 이메일, 비디오, 제휴 사이트, 소셜 미디어를 통해 상호작용하는 모든 정보를 수집해 내는데 관련된 웹 분석 도구, 소셜 듣기 도구의 사용 등 전문화된 분석 도구들을 사용하는 등의 내용을 포함하고 있음

⊙ 15. Big Data Security & Privacy
- Hadoop을 사용하여 여러 분산된 데이터 소스로부터 데이터를 모으기 전에 많은 잠재적인 보안 이슈들에 대하여 프라이버시 의무를 준수하게 하고, 민감한 데이터들은 감추게 해주는 데이터 마스킹, 데이터베이스 암호화, 데이터베이스 모니터링, 보안 정보와 이벤트 관리 등

⊙ 16. Big Data Lifecycle Management
- 정보의 생성부터 폐기까지의 기간 동안의 정보관리 절차와 방법론을 의미하는 것으로 법, 규제, 프라이버시 보호, 정보 아카이빙, 기록 및 유지관리, 법적 영향력과 증거수집, 시범 데이터 관리 등을 포함

⊙ 17. Cloud
- 조직이 클라우드 환경으로 이전함으로써 더 높은 유연성, 더 빠른 처리시간, 비용절감 등을 달성하게 함



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-05. 빅데이터 프라이버시

□ 빅데이터와 프라이버시

▷ 빅데이터 시대의 도래에 따른 안전 및 위험관리 화두

⊙ 대부분의 정보들이 디지털로 저장됨에 따라 데이터가 급격하게 증가하였으며, 2010년부터 빅데이터가 본격적으로 시작됨. Facebook, 구글서치 등 새로운 비즈니스가 시작되었으며, 이러한 비즈니스 모델들은 대부분 개인정보를 바탕으로 운영됨

⊙ 수집되는 개인정보의 양 증가에 따라 프라이버시 및 보안 리스크 관리를 더 잘하기 위해 무엇을 측정해야 하는지, 개인들에게 얼마나 많은 책임을 부여해야 하는지가 중요한 문제로 대두됨

⊙ 특히, 세계가 글로벌화되고, 인터넷이 전세계적으로 연결된 상황에서 프라이버시와 관련된 국가적 혹은 지역적 규제를 하는 접근방식이 유효한 것인지 따져볼 필요가 있음

▷ 기업들의 노력과 기술의 발전으로 안전과 프라이버시를 담보한 빅데이터의 활용이 가능

⊙ 정보보호와 프라이버시 문제는 빅데이터에만 국한된 것이 아니며, 모바일, 소셜, 클라우드 등 다양한 것들(all combination of things)이 조합되어 있음

⊙ 사회 전체적으로 효용창출이 가능한 개인정보에 대해서는 엄격한 보호아래 적극적인 활용이 필요함

⊙ 기업뿐만 아니라 소비자, 시민, 정부 모두 개인정보를 안전하게 활용할 수 있는 최대한의 노력을 기울여야 하며, 개인정보를 얼마나 투명하게 운영하는 지에 대한 정보가 공유된다면, 상회 신뢰구축이 가능

참고문헌 : 디지털 경제의 안전과 위험에 주목 (OECD포럼, 2014.5)

Module-05. 빅데이터 프라이버시

□ 빅데이터와 프라이버시

▷ 기업들의 노력과 기술의 발전으로 안전과 프라이버시를 담보한 빅데이터의 활용이 가능(계속)

⊙ MS의 경우는 제품 개발, 디자인 단계에서부터 보안과 프라이버시를 고려하고 있으며, 익명화(anonymization) 기술 등 보안을 위해 필요한 사안을 필수적으로 고려하고 있음

▷ 빅데이터 시대에서는 개인정보의 보안과 프라이버시는 더욱 엄격하게 보호될 필요가 있음

⊙ 보안 관련 이슈는 점점 더 복잡해지고 있으며, 모든 사람들이 개인정보의 보호와 관련된 기술을 다 이해하고 있지 못함

⊙ 빅데이터 시대에 대응하여 사람들이 신뢰할 수 있도록 올바른 데이터 관리가 필요하며, 이를 위해 합법적인 절차를 제정해야만 함

- 개인데이터를 보호하기 위한 기술적 솔루션도 중요하지만, 이를 기업들이 어떻게 관리하느냐도 중요한데, 특히, 기업들의 관리 소홀로 개인정보가 유출될 경우 그 기업의 CEO가 형사적 처벌을 묻게 하는 것도 필요함

⊙ 정보보안과 프라이버시 문제는 글로벌한 이슈가 되고 있으며 이러한 문제를 해결하기 위해 국제적인 동조가 필요

참고문헌 : 디지털 경제의 안전과 위험에 주목 (OECD포럼, 2014.5)

Module-05. 빅데이터 프라이버시

□ 개인정보보호 기술

단계	필요 기술
데이터 수집 단계	o 데이터 수집 시 동의 기술 - 개인정보 수집 시 동의 지원 o 데이터 수집 시 법률적 위반사항 검토 기술 o 데이터 수집 거부 기술 - 로봇 등 자동수집 배제 표준(권고안) - 대용량 크롤링 제한 기술(허용.불허용 의사표시)
데이터 저장 및 관리 단계	o 데이터 암호화 기술 - 데이터 암호화 기술(공개키 암호화, 대칭키 암호화) - DB 성능 저하가 되지 않는 데이터 암호화 o 데이터 접근통제(제어) 기술 - 임의 접근통제, 강제 접근통제, 역할기반 접근통제(일반적) 등 - 침입 탐지 및 차단 시스템 - VPN(Virtual Private Network) 등 네트워크 기반 기술 - 사용자 인증.권한 부여 등 계정 관리 기술 o 데이터 필터링 및 등급 분류 기술 - 데이터 등급별 분류(필터링).관리기술 - 데이터 자동 필터링(개인정보 자동 비식별 기술)

참고문헌 : DB Issue report, 빅데이터와 개인정보보호, 한국데이터베이스진흥원

Module-05. 빅데이터 프라이버시

□ 개인정보보호 기술(계속)

단계	필요 기술
데이터 처리 및 분석 단계	o 익명화된 데이터 처리 기술 - PPDM(Privacy Preserving Data Mining : 프라이버시 보호 분석 기술) : 데이터 소유자의 프라이버시를 침해하지 않으면서도 데이터에 함축적으로 들어 있는 지식이나 패턴을 찾아내는 기술 - K-anonymity & L-diversity 기술(DB를 분석하여 통계정보만 제공) · K-anonymity(익명성) : K개 이상의 동일한 데이터를 유지하여 특정인이 추론될 확률을 1/k 이하로 낮추는 기술 · L-diversity(다양성) : 민감한 데이터의 종류를 L개 이상 유지하는 기술 o 암호화된 데이터 처리 기술 - 순서보존 암호화 기술 : 암호화된 데이터 검색 - 연산보존 암호 기술 : 암호화된 데이터 연산*(MIT 10대 유망기술)
데이터 분석결과 가시화 및 이용 단계	o 이용자 동의 관련 기술 - 이용자 동의 기술 - 빅데이터 분석 결과의 영향력 사전 예측 기술 - 사전 동의없는 경우 사후 동의 지원 기술 o 분석정보의 이용 모니터링 기술 - 빅데이터 분석 모니터링 기술(이용 내역 고지) - 정보 이용과정 공개 등 모니터링 기술
데이터 폐기 단계	o 데이터 폐기 모니터링 기술 - 데이터 분석, 활용 후 폐기 확인 기술 - 분산 환경에서 완전한 데이터 폐기 기술(디가우징)

참고문헌 : DB Issue report, 빅데이터와 개인정보보호, 한국데이터베이스진흥원

Module-05. 빅데이터 프라이버시

□ 개인정보 제거방법

기법	주요 내용
가명처리 (pseudonymisation)	o 개인정보 중 주요 식별 요소를 다른 값으로 대체하여 개인 식별을 곤란하게 함 (예) 홍길동, 35세, 서울 거주, 한국대 재학 → 임꺽정, 30대 서울 거주, 국제대 재학 o 다른 값으로 대체하는 규칙이 노출되더라도 개인 식별이 불가능해야 함
총계처리 (Aggregation) 또는 평균값 대체 (Replacement)	o 데이터의 총합 값을 보임으로서 개별 데이터의 값을 보이지 않도록 함 (예) 임꺽정 180cm, 홍길동 170cm, 이콩쥐 160cm, 김팥쥐 150cm → 물리학과 학생 키 합 : 660cm, 평균키 165cm o 특정 속성을 지닌 개인으로 구성된 단체의 속성 정보를 공개하는 것은 비식별화에 무의미 (예) 에이즈 환자 집단임을 공개하면서 특정 인물 '갑'이 그 집단에 속함을 알 수 있도록 표시하는 것
데이터 값 삭제 (Data Reduction)	o 데이터셋에 구성된 값 중에 필요없는 값 또는 개인 식별에 중요한 값을 삭제 (예) 홍길동, 35세, 서울 거주, 한국대 졸업 → 35세, 서울 거주, 주민등록번호 901206-1234567 → 90년대 생, 남자 날짜 정보(자격취득일자, 합격일 등)의 연 단위 처리
범주화 (Data Suppression)	o 데이터의 값을 범주의 값으로 변환 (예) 홍길동, 35세 → 홍씨, 30-40세
데이터 마스킹 (Data Masking)	o 공개된 정보 등과 결합하여 개인을 식별하는데 기여할 확률이 높은 경우, 주요 개인 식별자가 보이지 않도록 처리하여 개인을 식별하지 못 하도록 함 (예) 홍길동, 35세, 서울 거주, 한국대 재학 → 홍**, 35세, 서울 거주, **대학 재학 o 남아 있는 정보 그 자체로 개인을 식별할 수 없어야 하며 인터넷 등에 공개되어 있는 정보 등과 결합하였을 경우에도 개인을 식별할 수 없어야 함

참고문헌 : DB Issue report, 빅데이터와 개인정보보호, 한국데이터베이스진흥원

□ 개인정보 규제 및 완화 관련 법·제도 국내외 현황

▷ 미국·유럽 vs 한국

⊙ 유럽 및 미국 등에서 '망각의 권리' 등으로 논의되다 2012년 EU 유럽정보보호지침 개정안에 잊혀질 권리 신설

⊙ 미국 캘리포니아주, 2013년 9월, 18세 이하 미성년자에 한해 인터넷 서비스 업체에 자신 관련 기록물을 지우거나 숨기도록 요청할 수 있는 법안 통과

⊙ 미국은 2014년 5월까지 프라이버시 보호를 위한 빅데이터 정책을 검토, '소비자 프라이버시 권리장전(2012년 오바마 행정부 제안)'의 통과 촉구, 사이버보안 입법안(2011년 제안)의 이행, 전자 커뮤니케이션에 관한 프라이버시법 개정 등을 제안

⊙ 한국은 미국·유럽과는 반대로 안행부에서 '빅데이터 개인정보보호 가이드라인(안)'을 준비하는 등 개인정보보호를 완화하려는 시도

- 2013년 12월 18일 방송통신위원회와 한국인터넷진흥원이 발표 후 지속 수정 중임
  2014.12.23 방송통신위원회 '빅데이터 개인정보보호 가이드라인' 제정
- 개인정보보호위원회에서 '개인정보보호법' 및 '정보통신망 이용촉진 및 정보보호 등에 관한 법률'의 규정과 입법 취지에 부합하지 아니하는 일부 내용을 포함하고 있으므로 재검토하도록 권고

참고문헌 : DB Issue report, 빅데이터와 개인정보보호, 한국데이터베이스진흥원

□ 2014.12.23 방송통신위원회 '빅데이터 개인정보보호 가이드라인'

▷ 빅데이터 개인정보보호 가이드라인 전문

⊙ 제1조(목적) 이 가이드라인은 공개된 개인정보 또는 이용내역정보 등을 전자적으로 설정된 체계에 의해 수집·저장·조합·분석 등 처리하여 새로운 정보를 생성함에 있어서 이용자의 프라이버시 등을 보호하고 안전한 이용환경을 조성하는 것을 목적으로 한다.

⊙ 제2조(정의) 이 가이드라인에서 사용하는 용어의 정의는 다음과 같으며, 본 조에서 정의되지 않은 용어는 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」, 「개인정보 보호법」 등 관련 법률에서 정의한 바에 따른다.

1. "공개된 정보"란 이용자 및 정당한 권한이 있는 자에 의해 공개 대상이나 목적의 제한 없이 합법적으로 일반 공중에게 공개된 부호·문자·음성·음향 및 영상 등의 정보를 말한다.
2. "이용내역정보"란 이용자가 정보통신서비스를 이용하는 과정에서 자동으로 발생하는 서비스 이용기록, 인터넷 접속정보, 거래기록 등의 정보를 말한다.
3. "정보 처리시스템"이란 공개된 개인정보 또는 이용내역정보 등을 전자적으로 설정된 체계에 의해 조합·분석 등 처리하여 새로운 정보를 생성하는 시스템을 말한다.
4. "비식별화"란 데이터 값 삭제, 가명처리, 총계처리, 범주화, 데이터 마스킹 등을 통해 개인정보의 일부 또는 전부를 삭제하거나 대체함으로써 다른 정보와 쉽게 결합하여도 특정 개인을 식별할 수 없도록 하는 조치를 말한다.

□ 2014.12.23 방송통신위원회 '빅데이터 개인정보보호 가이드라인'

▷ 빅데이터 개인정보보호 가이드라인 전문

⊙ 제3조(개인정보의 보호)

① 정보통신서비스 제공자가 정보 처리시스템을 통해 공개된 정보, 이용내역정보를 수집.저장.조합·분석 등 처리하고자 하는 경우, 개인정보의 보호를 위해 다음 각 호의 조치를 취하여야 한다.

1. 개인정보가 포함된 공개된 정보 및 이용내역정보는 비식별화 조치를 취한 후 수집·저장·조합·분석 등 처리하여야 한다.
2. 비식별화 조치된 공개된 정보 및 이용내역정보를 조합·분석 등 처리하는 과정에서 개인정보가 생성되지 않도록 하여야 한다. 다만, 개인정보가 생성되는 경우에는 지체없이 파기하거나 비식별화 조치를 취하여야 한다.

② 비식별화 조치된 공개된 정보 및 이용내역정보를 정보 처리시스템에 저장·관리하는 경우 다음 각 호의 보호조치를 취하여야 한다.

1. 불법적인 접근을 차단하기 위한 침입차단시스템 등 접근 통제장치의 설치·운영
2. 접속기록의 위조·변조 방지를 위한 조치
3. 백신 소프트웨어의 설치·운영 등 악성 프로그램에 의한 침해 방지조치
4. 기타 안전성 확보를 위해 필요한 보호조치

□ 2014.12.23 방송통신위원회 '빅데이터 개인정보보호 가이드라인'

▷ 빅데이터 개인정보보호 가이드라인 전문

⊙ 제4조(공개된 정보의 수집.이용)

① 정보통신서비스 제공자가 개인정보가 포함된 공개된 정보를 비식별화 조치한 경우에는 이용자의 동의 없이 수집·이용할 수 있다. 다만, 이용자의 동의를 받거나 법령상 허용하는 경우에는 비식별화 조치를 취하지 아니하고 수집·이용할 수 있다.

② 정보통신서비스 제공자는 제1항에 따라 개인정보가 포함된 공개된 정보를 수집.이용하는 경우 공개된 정보의 수집 출처, 수집·저장·조합·분석 등 처리하는 사실 및 그 목적을 이용자 등이 언제든지 쉽게 확인할 수 있도록 개인정보 취급방침을 통해 공개하여야 한다.

③ 정보통신서비스 제공자는 이용자 등의 요구가 있으면 즉시 다음 각 호의 모든 사항을 이용자 등에게 알려야 한다.

1. 개인정보의 수집 출처
2. 개인정보의 수집·저장·조합·분석 등 처리의 목적
3. 해당 개인정보의 처리 정지를 요구할 권리가 있다는 사실

□ 2014.12.23 방송통신위원회 '빅데이터 개인정보보호 가이드라인'

▷ 빅데이터 개인정보보호 가이드라인 전문

⊙ 제5조(이용내역정보의 수집·이용)

① 정보통신서비스 제공자는 다음 각 호의 경우를 제외하고는 이용자의 동의를 받거나 비식별화 조치를 취한 후 이용내역정보를 수집.이용할 수 있다.

1. 정보통신서비스의 제공에 관한 계약을 이행하기 위하여 필요한 이용내역정보로서 경제적.기술적인 사유로 통상적인 동의를 받는 것이 뚜렷하게 곤란한 경우
2. 정보통신서비스의 제공에 따른 요금정산을 위하여 필요한 경우
3. 다른 법률에 특별한 규정이 있는 경우

② 정보통신서비스 제공자는 제1항에 따라 이용내역정보를 수집하는 경우 해당 정보가 수집·저장·조합·분석 등 처리되는 사실 및 목적을 이용자가 언제든지 쉽게 확인할 수 있도록 개인정보 취급방침을 통해 공개하여야 한다.

③ 정보통신서비스 제공자는 이용내역정보의 수집·저장·조합·분석 등 처리를 거부할 수 있는 방법 및 절차를 마련하여야 한다.

④ 정보통신서비스 제공자는 이용자의 검색프로그램 등에서 이용자 또는 검색프로그램 등 공급자가 설정해 놓은 이용내역정보의 수집 거부 선택을 이용자의 동의 없이 변경해서는 아니 된다.

Module-05. 빅데이터 프라이버시

□ 2014.12.23 방송통신위원회 '빅데이터 개인정보보호 가이드라인'

▷ 빅데이터 개인정보보호 가이드라인 전문

⊙ 제6조(새로운 정보의 생성)

① 정보통신서비스 제공자는 비식별화 조치하여 수집한 공개된 정보 및 이용내역정보를 정보 처리시스템을 통해 조합·분석하여 새로운 정보를 생성할 수 있다. 다만, 새롭게 생성된 정보에 개인정보가 포함되어 있을 경우, 즉시 파기하거나 비식별화 조치를 취하여야 한다.

② 정보통신서비스 제공자는 제1항에 따라 개인정보가 포함된 정보가 생성될 수 있다는 사실 및 그 처리 방법을 이용자가 언제든지 쉽게 확인할 수 있도록 개인정보 취급방침을 통해 공개하여야 한다.

⊙ 제7조(민감정보 생성의 금지) 특정한 개인의 사상·신념, 노동조합·정당의 가입·탈퇴, 정치적 견해, 건강, 성생활 등에 관한 정보, 그 밖에 이용자의 사생활을 현저히 침해할 우려가 있는 정보의 생성을 목적으로 공개된 개인정보 등을 수집·저장·조합·분석 등 처리하여서는 아니 된다. 다만, 이용자의 사전 동의를 받거나 법률에 따라 허용된 경우에는 그러하지 아니하다.

⊙ 제8조(통신 내용의 조합, 분석 또는 처리 금지) 정보통신서비스 제공자는 전송중인 이메일, 문자메시지 등의 통신 내용에 대하여 양 당사자의 동의를 얻은 경우를 제외하고, 통신 내용의 전부 또는 일부를 조합, 분석 또는 처리하여서는 아니 된다

Module-05. 빅데이터 프라이버시

□ 2014.12.23 방송통신위원회 '빅데이터 개인정보보호 가이드라인'

▷ 빅데이터 개인정보보호 가이드라인 전문

⊙ 제9조(공개된 정보 및 이용내역정보의 이용)

① 정보통신서비스 제공자는 비식별화 처리된 공개된 정보 및 이용내역정보를 자신의 서비스 제공업무 수행을 위해 내부에서 이용할 수 있다. 다만, 이용자가 거부 의사를 표시한 때에는 그러하지 아니하다.

② 정보통신서비스 제공자는 제1항에 따라 공개된 정보 및 이용내역정보를 이용하는 경우 해당 정보가 이용된다는 사실 및 그 목적을 이용자가 언제든지 쉽게 확인할 수 있도록 개인정보 취급방침을 통해 공개하여야 한다.

⊙ 제10조(제3자 제공) 정보통신서비스 제공자는 개인정보가 포함된 공개된 정보, 이용내역정보, 생성 정보의 경우, 이용자의 동의를 얻어 제3자에게 제공할 수 있다. 다만, 비식별화 처리된 공개된 정보, 이용내역정보, 생성 정보는 이용자 동의 없이 제3자 제공이 가능하다.

⊙ 제11조(적용범위) 이 가이드라인에서 규정하지 않은 사항은 「정보통신망 이용촉진 및 정보보호 등에 관한 법률」, 「개인정보 보호법」 등 관련 법률에 따른다.

Module-05. 빅데이터 프라이버시

□ 개인정보 활용을 위한 새로운 대안 시도

▷ 개인 데이터 개념의 진화

⊙ 전통적으로 개인의 프라이버시를 보호하는 한편, 데이터를 익명화하면서 사회에 가치를 창출하는 다양한 기술을 사용

⊙ 개인 데이터의 정의는 개인의 선호, 새로운 애플리케이션, 사용의 맥락, 문화적·사회적 규범의 변화에 따라 진화

▷ 개인 데이터 사용의 새로운 접근방법 요구

⊙ 전통적인 접근방법은 데이터가 수집된 이후, 이에 대한 새로운 가치를 창출할 수 있는 가능성을 설명하는데 제약

- 데이터 주체로부터 그리고 그에 대해 생성되는 수많은 데이터는 개인적인 데이터 주체에게 부당한 인지적 부담을 지움
- 전통적인 접근방법으로 개인의 동의를 구하는 것은 더 이상 실용적이거나 효과적이지 못함

⊙ 개인적 데이터 생태계의 복잡성, 변화의 속도, 잠재적 가치와 변화하는 개인의 역할을 고려하여 유연한 접근방법 필요

- 데이터의 흐름과 결합의 허용을 통해, 얻을 수 있는 잠재적 가치와 그로 인해 야기되는 위험과 균형을 이룰 필요가 있음
- 데이터는 자체적으로 가치를 창출하거나, 문제를 초래하는 것이 아니라 데이터의 사용이 가치를 창출하거나 문제를 초래
- 개인 데이터 자체에 대한 보호의 초점으로부터 데이터 사용 권한의 가능성으로 초점을 바꾸는 사고의 전환 필요
- 개인 데이터 수집을 통제하는 것으로부터 데이터 사용에 초점을 맞추도록 요구

참고문헌 : 빅데이터 시대의 개인 데이터 보호와활용 (NIA, 2013.6)

□ 개인정보 활용을 위한 새로운 대안 시도

▷ 개인 데이터 사용의 새로운 접근방법 요구(계속)

⊙ 데이터를 통해 가치를 창출할 수 있으나 이를 침해하거나 손상시키는 것을 예방할 수 있는 응용을 가능하게 하는 허용, 통제, 신용할 수 있는 데이터 처리 관행의 확립 필요

▷ 신뢰할 수 있는 개인 데이터 생태계 조성을 위한 원칙의 모색

⊙ OECD는 개인정보보호 8원칙(보호화 보안, 책임, 데이터 사용에 대한 권한과 의무(개인의 참여, 데이터 품질, 개방성, 수집 제한, 목적 명시, 사용 제한))을 천명하여 세계 개인 데이터 및 프라이버시 보호 관련 법제도의 근간으로 역할

- 인터넷 등 정보기술의 급격한 변화와 이로 인한 경제·사회적 환경의 변화에 맞추어 OECD 원칙의 개정 필요성 제기

⊙ 세계경제포럼(WEF)은 OECD 원칙을 세가지 범주로 구분하고 신뢰할 수 있는 데이터 생태계 조성을 위한 새로운 이슈를 식별

- 의도적이든 비의도적이든 보안 침해와 남용으로부터 어떻게 개인 데이터를 보호하고 지킬 것인가?
- 이해관계자들 간의 조화를 유지하면서 개인 데이터가 흐를 수 있도록 어떻게 상호간의 권한과 의무를 확립할 것인가?
- 확립된 권한과 의무에 따라 개인 데이터를 보호하고 지키기 위해 책임과 법 집행을 강제할 것인가?

참고문헌 : 빅데이터 시대의 개인 데이터 보호와활용 (NIA, 2013.6)

□ 개인정보 활용을 위한 새로운 대안 시도

▷ 신뢰할 수 있는 개인 데이터 생태계 조성을 위한 원칙의 모색(계속)

⊙ 개인 데이터의 보호와 보안과 관련하여 사이버 안전을 위한 기존의 노력들과 긴밀한 연계 필요

- 개인 데이터 보호는 데이터 관리와 저장의 상호의존적인 본질에 따라 다양한 이해관계가 얽혀 있으므로 협력적인 접근방법 필요

⊙ 조직에게 인센티브와 함께 책임을 부과할 수 있는 법·규제적 집행 메커니즘 확립 필요

- 모든 조직이 신뢰를 구축하여 데이터의 침해와 오남용으로부터 안전을 보장하기 위하여 규제자로서 정부의 역할 수행

⊙ 개인에 관한 데이터가 어떻게 수집되고 사용되는지 개인에게 고지하여 그들의 이해를 돕는 새로운 방법 필요

- 개인 데이터 수집 및 활용의 투명성을 보장하기 위해 조직이 공개하는 프라이버시 정책의 복잡성으로 인해 실효성 미흡
- 개인 데이터의 공정한 활용을 위해서는 데이터가 수집·처리·사용되는 방식을 보다 쉽고 명확하게 이해할 수 있는 수단의 제공 필요

⊙ 개인이 자신의 데이터 수집 및 활용에 대해 보다 효과적으로 선택하고 통제할 수 있는 새로운 방식 필요

- 소극적인 데이터 주체에서 적극적인 이해관계자 및 데이터 생산자로서 개인의 역할 변화에 적법한 참여와 권한 강화
- 개인 데이터 활용을 통해 발생하는 경제·사회적 편익에 대하여 이해관계자 간 공정한 가치 배분의 메커니즘 확립 필요

참고문헌 : 빅데이터 시대의 개인 데이터 보호와활용 (NIA, 2013.6)

□ 개인정보 활용을 위한 새로운 대안 시도

▷ 신뢰할 수 있는 개인 데이터 생태계 조성을 위한 원칙의 모색(계속)

⊙ 모든 상황에 적용되는 개인 데이터 규제에서 탈피하여 개인 데이터 사용의 상황 및 맥락을 고려한 유연한 접근방법 필요

- 데이터의 수집 시점에서 개인의 동의를 구하는 개인 참여의 초점을 활용 시점으로 전환
- 서로 다른 문화 규범, 실행을 위한 서로 다른 일정, 잠재적 해결방안에 대한 상이한 경로 등을 가진 이해관계자들의 맥락 고려

참고문헌 : 빅데이터 시대의 개인 데이터 보호와활용 (NIA, 2013.6)

Module-05. 빅데이터 프라이버시

□ 빅데이터 프라이버시 베스트 프렉티스

▷ 민감한 빅데이터의 식별

⊙ 개인과 연결되는 빅데이터는 개인식별정보로 간주되어야 함

⊙ 미국 연방통상위원회(FTC)가 제정한 프라이버시 프레임워크 가이드

- 조직은 개인 식별자를 제거하는 단계를 취해야 함. 민감한 데이터 필드에 대한 수정, 제거, 노이즈 추가, 집계 합성 등의 과정을 거침
- 변형된 민감한 데이터를 다시 식별자로 만드는 행위의 금지
- 조직은 민감한 데이터를 사용한 서비스 제공자나 이를 관리하는 제3자에게 다시 식별자로 만드는 행위를 금지시켜야 함

▷ 메타데이터 리포지토리에 민감한 빅데이터를 표시

⊙ 민감한 데이터는 조직내에서 별도의 영역에 저장하여 관리함

⊙ 개인식별정보를 포함한 특정 데이터 필드들이 특정 필드내에 포함되어 있든가 하여 프라이버시 통제 밖으로 벗어날 수 있음. 이경우 데이터 디스커버리 도구들을 이용하여 개인식별정보와 연관이 있는 필드들을 조사함

▷ 국가별로 프라이버시 법률과 규제 연구

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-05. 빅데이터 프라이버시

□ 빅데이터 프라이버시 베스트 프렉티스

▷ 국경을 넘는 개인 데이터의 관리

⊙ 미국 상무성(1999.4), 'U.S.-EU Safe Harbor Principles' : 미국과 유럽간의 데이터 이동 규칙

- 고지 : 개인정보의 수집, 활용, 재사용 목적에 관하여 당사자에게 고지
- 선택 : 개인에게 opt-out의 기회를 제공해야 한다. 민감한 정보에 대하여 개인에게 동의 혹은 명시적인 opt-in 선택권이 주어져야 함
- 데이터 이동 : 조직은 고지와 선택의 원칙에 부합한 상황에서만 제3자에게 개인정보를 제공할 수 있음
- 보안 : 조직은 개인정보에 대한 분실, 오용, 비인가, 접근, 노출, 변조 및 파괴 등으로부터 보호할 조치를 취해야 함
- 데이터 무결성 : 조직은 데이터 수집 시 명시한 목적에 맞게 데이터를 사용해야 함. 또한 데이터가 정확하고 완전하며 현재 것이라는 것을 보장해야 함
- 접근 : 개인은 자신의 개인정보에 접근하여 오류를 수정하거나 보충자료를 제공할 수 있어야 함
- 통제 강화 : Safe Harbor Principles를 준수하는 지를 확인해서 그렇지 않을 때 개인에 구제를 보증할 효과적인 장치가 있어야 함

▷ 인가받은 사용자에 의한 민감한 빅데이터 접근 관리

⊙ 빅데이터 거버넌스 프로그램은 민감한 데이터를 정의하고 이를 접근하는 인가된 사용자의 접근도 모니터링

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-06. 빅데이터 품질

□ 빅데이터 품질관리

▷ 빅데이터 품질 요소와 품질 전략

⊙ 빅데이터 품질관리는 데이터의 사용목적, 데이터의 재사용 여부, 일관성 유지 여부에 따라 정확성, 완전성, 적시성, 일관성에 대한 별도의 품질 전략을 수립

⊙ 빅데이터 품질 관리는 'exact(정확성)' 보다는 'good enough(충분성)' 개념 하에서 조직의 비즈니스 영역 및 목적에 따라 수행하는 것이 바람직함

데이터 품질 요소	데이터 품질 전략
정확성 (Accuracy)	데이터 사용 목적에 따라 데이터 정확성의 기준을 다르게 적용 ex) 사용자가 접속한 사이트와 이동 지점을 분석하는 클릭스트림 분석과 부정이나 사기를 탐지하는 경우 데이터의 품질 수준은 다름
완전성 (completeness)	필요한 데이터의 완전한 확보 보다는 필요한 데이터를 식별하는 수준으로 적용 가능
적시성 (Timeliness)	소멸성이 강한 데이터에 대해 어느 정도의 품질 기준을 적용할 것인지 결정 ex) 웹 로그 데이터, 트윗 데이터, 위치 데이터 등은 하루, 몇 시간, 몇 분 동안만 타당성을 가짐
일관성 (Consistency)	동일한 데이터라 할지라도 사용 목적에 따라 달라지는 데이터 수집 기준 때문에 데이터 의미가 달라질 수 있음

참고문헌 : 빅데이터 시대의 데이터 자원 확보와 품질관리방안 (NIA, 2012.5)

Module-06. 빅데이터 품질

□ 빅데이터 품질관리

▷ 빅데이터 활용 결과의 정확성 및 신뢰성 향상을 위해 빅데이터 품질 관리 체계 구축 필요

⊙ 기업의 정형 데이터, 공공기관이 보유한 공공정보 등은 개별 정보에 대한 품질 관리를 통해 데이터의 중복성, 불일치성 등을 관리

⊙ 다양한 데이터 소스를 활용하는 빅데이터는 각 데이터의 특성을 고려한 종합적인 빅데이터 품질 관리 가이드라인 마련 필요

- 기존 품질 관리 지침에 따라 관리되는 고품질 데이터는 빅데이터 자원으로 그대로 활용 가능하나
- 3V 데이터는 그 특성상 기존 데이터와 다른 품질 기준, 품질관리 프로세스, 품질 전략의 수립이 필요

⊙ 빅데이터 자원의 품질을 보장하고 활용을 극대화하기 위한 빅데이터 자원 품질 인증 방안 연구 필요

- 빅데이터 품질 관리 가이드라인을 준수한 데이터에 대한 '빅데이터 품질 라이센스'와 데이터 공개 가이드라인을 준수한 공개 데이터에 대한 '데이터 공유 라이센스' 부여 방안 검토 필요

  참조 :「지방자치단체 공공데이터 이용 활성화를 위한」 공공데이터 제공 방안 및 가이드라인 연구

참고문헌 : 빅데이터 시대의 데이터 자원 확보와 품질관리방안 (NIA, 2012.5)

Module-06. 빅데이터 품질

□ 빅데이터 품질관리

▷ 공공관리 품질관리 제도 현황

⊙ 공공정보 품질관리 관련 근거법

- 국가정보화 기본법 제18조(지식・정보의 공유・유통), 제25조(지식정보자원의 관리 등), 제26조(지식정보자원의 표준화)
- 국가정보화 기본법 시행령 제20조(지식정보자원의 관리), 제23조(지식정보자원의 활용 촉진)

⊙ 공공기관의 데이터베이스 품질관리 지침

- 데이터베이스 품질관리계획의 수립, 품질오류 신고접수 및 처리, 데이터베이스 표준화, 연계데이터 품질관리, 품질관리지원센터 등을 규정

⊙ 공공정보 품질 관리 매뉴얼

- 국가 및 기관 차원의 데이터 품질 관리 체계, 데이터 품질 진단 및 개선 절차와 단계별 주요 활동을 기술하고, 데이터 품질 관리 수준 체크리스트, 데이터 품질 지표별 체크리스트, 지표별 품질기준 및 진단방법 등을 제공

참고문헌 : 빅데이터 시대의 데이터 자원 확보와 품질관리방안 (NIA, 2012.5)

Module-06. 빅데이터 품질

□ 빅데이터 품질관리 베스트 프랙티스

▷ 기존 데이터와 빅데이터의 품질관리 비교

차원	기존 데이터 품질관리	빅데이터 품질관리
프로세싱 방식	배치 방식의 프로세싱	실시간 혹은 배치 방식의 프로세싱
데이터의 다양성	주로 구조적 데이터	구조적 데이터, 반구조적 데이터, 비구조적 데이터
신뢰수준	데이터 웨어하우스에 정제된 상태로 데이터가 저장되어 있어야 올바른 분석이 가능함	Noise는 미리 필터링되어야 하지만 데이터는 'good enough'로 충분할 수 있음(분석 인사이트가 데이터 품질에 제한적으로만 영향을 받는 경우도 있음)
정제시간	데이터 웨어하우스로 로드하기 전에 정제되어 있어야 함	중요한 데이터와 그들사이의 관계를 완벽히 알 수 없기에 우선 있는 그대로 적재함 데이터의 엄청남 양과 빠른 속도로 인하여 스트리밍 방식이나 인메모리 방식의 분석을 통한 정제가 요구되며, 결과적으로 저장공간을 줄일 수 있음
핵심 데이터의 요소	데이터 품질은 고객 데이터와 같은 핵심 데이터에 대하여 평가됨	빅데이터는 불분명하고, 때론 더 깊은 탐구가 필요하기 때문에 기준이 되는 핵심 데이터 요소들이 반복적으로 바뀔 수 있음
분석장소	분석할 데이터가 데이터 품질 및 분석 담당 엔진으로 이동함	반대로 품질관리 및 분석 서버가 데이터 쪽으로 이동함(데이터를 이동하는 비용과 시간이 크기 때문에)
관리주체	관리주체는 대부분의 데이터를 관리할 수 있음	데이터의 빠른 생성속도와 많은 양 때문에 관리 주체는 데이터의 일부분만을 담당할 수 있을 것임

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-06. 빅데이터 품질

□ 빅데이터 품질관리 베스트 프랙티스

▷ 빅데이터 품질 프로그램

⊙ 비즈니스 당사자들과 함께 빅데이터 품질에 대한 신뢰구간을 설정

- 데이터 품질이 먼저 해결되어야 하는 기존 Business Intelligence 사업과 달리, 빅데이터 사업에는 데이터 품질 문제를 사정에 맞추어 해결해야 함
- 핵심 데이터 요소를 구분하여 잘 관리해야 함
- Twitter 데이터에 대한 신뢰구간

  원문) Best day ever. Went Mountain biking today. Will get one for my husband too.

  속성 추정) 성별 : 여성, 결혼유무 : 기혼, 스포츠 : 자전거 타기, 나이 : 25~55(결혼하였으며, 활발하게 스포츠를 즐김)

카테고리	신뢰구간
Very High	90% ~ 99%
High	80% ~ 89%
Medium	70% ~ 79%
Low	< 69%

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-06. 빅데이터 품질

□ 빅데이터 품질관리 베스트 프랙티스

▷ 빅데이터 품질 프로그램

⊙ 비즈니스 당사자들과 함께 빅데이터 품질에 대한 신뢰구간을 설정(계속)

- Twitter 데이터에 대한 품질 요구 행렬

핵심 데이터	데이터 품질 문제점	데이터 품질 비즈니스 규칙
Twitter의 타임스탬프	표준형식과 다른 Twitter의 타임스탬프의 형식은 외부 데이터와 조인을 하는데 있어서 오류를 발생시킬 수 있다.	모든 양식은 YYYY-MM-DD HH:MM:SS로 재구성한다.
사용자 이름	프로필에 기재된 사용자 이름 중 40~50%가 실제 이름과 다르지만, 이러한 이름은 MDM(Master Data Management)에서 실제 고객정보와 연결할 수 있는 유용한 요소이다. Tweet handle(screen_name)보다 사용자 프로필로부터 이름을 추출하는 것이 중요하다.	o 만약 숫자나 부호와 같은 문자가 사용자 이름에 포함되어 있다면 신뢰 수준은 : 0% o 만약 성이나 이름에 해당하는 한 단어만 포함되어 있다면 신뢰수준은 : 25% o 만약 2~3개의 단어가 포함되어 있다면 신뢰수준은 : 50% o 만약 사용자 이름이 사전에 등록된 것이라면 신뢰수준은 : 99%
회사 언급	Acme Corporation에 대한 트윗인가 혹은 필터링되어야 할 노이즈인가?	o Twitter에 "@Acme" 단어가 포함되어 있다면 신뢰수준은 : 99% o Twitter에 "Acme"과 Acme사의 제품 이름이 포함되어 있다면 신뢰수준은 : 75% o Twitter에 "Acme"에 대해 아무런 언급이 없으면 신뢰수준은 : 0%
장소	o Twitter에서 장소 데이터는 매우 중요한 역할을 한다. 예) 미국 동남부에 위치한 지역 사람들이 다른 지역에 위치한 사람들보다 불만스러워하는 이유는 무엇인가? o 사용자 프로파일에 사용자 거주 도시와 주 정보가 포함될 수는 있으나 입증된 것은 아니다.	o Tweet 메타데이터로부터 Tweet.user.location 문자열을 추출하여 사용자의 거주 도시와 주 이름을 확인하라. o 만약 위치를 알아내기가 어렵다면, "알 수 없음"으로 표시해 둔다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-06. 빅데이터 품질

□ 빅데이터 품질관리

▷ 품질관리에 영향을 미치는 빅데이터 특징과 품질관리 접근방법

⊙ 빅데이터는 '대량의 데이터', '세밀한 수준의 데이터', '소유자가 불분명한 데이터' 특성으로 인해 다른 접근 방식의 품질관리가 필요. 모든 개별 데이터에 대한 타당성 보장보다는 빅데이터 개념 및 특성 측면에서 관리되어야 할 항목과 수준에 대해 품질을 정의

빅데이터 특징			품질관리 접근방법
대량의 데이터	수작업으로 수집되기 보다는 기계, 프로그램 등에 의해 수집되는 대량의 데이터	→	.혹시 발생할지 모르는 데이터 사용자의 오류는 무시 .데이터 수집 과정의 타당성을 방해하는 예외상황을 탐지하는 수준으로 품질 기준 정의(ex. 장치 고장으로 인한 데이터 손실, 장치의 비정상적 상황으로 인한 비정상적 수치 등)
미세하고 정밀한 데이터	클릭 스트림, 미터 값 등 기계, 센서, 프로그램 등에서 생산되는 데이터로 기존 데이터 보다 훨씬 미세한 데이터	→	.개별 데이터에 대한 타당성 검증은 경우에 따라 불필요 .개별 레코드에 대한 의미보다 데이터 전체가 나타내는 의미를 중심으로 품질 기준 정의
데이터 소유자 불분명	누가 언제 어디서 데이터를 생산한 것인지에 대한 관리.감독이 불가능한 조직 외부의 데이터	→	.목적이나 통제없이 생산된 데이터에 대한 데이터 품질 기준을 정의하기 위한 다른 방법 필요

참고문헌 : 빅데이터 시대의 데이터 자원 확보와 품질관리방안 (NIA, 2012.5)

Module-06. 빅데이터 품질

□ 빅데이터 품질관리 베스트 프랙티스

▷ 빅데이터 품질 프로그램

⊙ 반구조적, 비구조적 데이터 활용으로 구조적 데이터의 품질 향상

- 구조적 데이터의 품질을 높이기 위해 반구조적 데이터를 이용할 수 있음

예) 병원시스템에서 사용하는 비구조적 데이터(전자의료기록, 환자 신체검사 결과, 의사의 주의사항, 퇴원 요약정보 등)를 사용하여 금연상태, 마약과 알코올 남용 등에 관한 정확한 통찰력의 확보가 가능함

⊙ 디스크 비적재 방식의 스트리밍 데이터 분석을 활용한 데이터 품질 향상

- 스트리밍 분석에서는 대규모 데이터를 디스크에 저장하지 않고, 실시간으로 직접 분석하는데 스트리밍 애플리케이션에 입력되는 데이터인 소스(Source)와 스트림 애플리케이션에서 생성된 데이터인 싱크 혹은 목적지를 상호 연계함

예) 스트리밍 애플리케이션 예 : 소켓 연결, 데이터베이스 질의, 자바 메시지 서비스 topic/queue, 혹은 파일 등

- 스트리밍 데이터 프로파일링

. 시간적인 조정 : 스트리밍 애플리케이션은 다른 소스들로부터 데이터를 받아서 조인하고, 연관시키며, 매칭 작업을 수행할 때 시간적인 차이점을 이해하고 조정해야 함

. 도착 비율 : 데이터가 연속적으로 유입되는가, 폭증하는 현상이 있는가, 데이터 도착에서 차이가 있는가를 분석함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-06. 빅데이터 품질

□ 빅데이터 품질관리 베스트 프랙티스

▷ 빅데이터 품질 프로그램

⊙ 정보 거버넌스 위원회에 데이터 관리 주체 지정

- 국가 및 기관 차원의 데이터 빅데이터 품질에 대한 비즈니스 규칙과 신뢰구간을 설정하고 개선해 나감
- 빅데이터 품질에 대한 이슈를 다룸 : 기존 데이터보다 더욱 다양하고 다른 형태의 빅데이터에 대하여 관리주체는 빅데이터의 일부분만 품질관리를 할 수 있음. 따라서 고속으로 입력되는 빅데이터에 대하여 적절한 기준이 설정되어 관리자가 관심을 가져야 하는 상황 등을 자동으로 탐지해 내야 함
- 빅데이터 품질의 트랜드를 보고함 : 데이터 품질을 향상하는데 있어서 도움을 주는 리포트를 생성하여 일정기간 동안 데이터 품질이 어떻게 바뀌고 있는지, 데이터 품질의 동향에 관한 정보를 정보 거버넌스 위원회에 시각화된 형태로 보고함

  예) 핵심 데이터 항목에 대한 데이터 품질의 동향, 관리권이 요구되는 데이터 품질에 관한 이슈들의 동향

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-07. 마스터 데이터 통합

□ 마스터 데이터 관리 개요

▷ 마스터 데이터 관리 필요성

⊙ 부서 간 정보 교환이 어려우며, 이 중 대부분은 기업 내 정보 교환 프로세스가 존재하지 않음

⊙ 시스템의 기능이나 데이터들이 여러 시스템에 걸쳐 중복되어 있으며, 관리도 되지 않음

⊙ 사용자들은 정보의 정확성이 그들이 원하는 수준만큼 정확하지 못한다고 생각함

⊙ 데이터의 품질이 만족스럽지 못함. 즉, 기업 내 데이터의 정의가 부서마다 다르고, 동일한 데이터가 여러 시스템에 서로 다른 포맷으로 정의가 되어 있음

⊙ 현업 부서 간 정보에 대한 공감대 형성이나 협력이 부족함

⊙ 데이터의 재사용률이 낮고, 데이터를 이해하기가 어려움

⊙ 전사적인 데이터에 대한 정의가 없음

⊙ 부서 간 전사 차원의 정보를 통합하고 표준화하는 것에 대해 공감대가 형성되지 않음

⊙ 정보에 대한 통합이나 데이터에 대한 표준이 개별 프로젝트마다 필요에 따라 만들어짐

▷ 근본 원인

⊙ 데이터 관리 전략 및 정책의 부재, 체계화된 프로세스 및 조직 구성 미비, 개별 시스템 단위의 데이터 표준화 수행, 데이터 오류/사고 발생 후 사후 조치에 의존, 프로젝트 완료 후 변화 관리 미수행, 모니터링 방법 부재

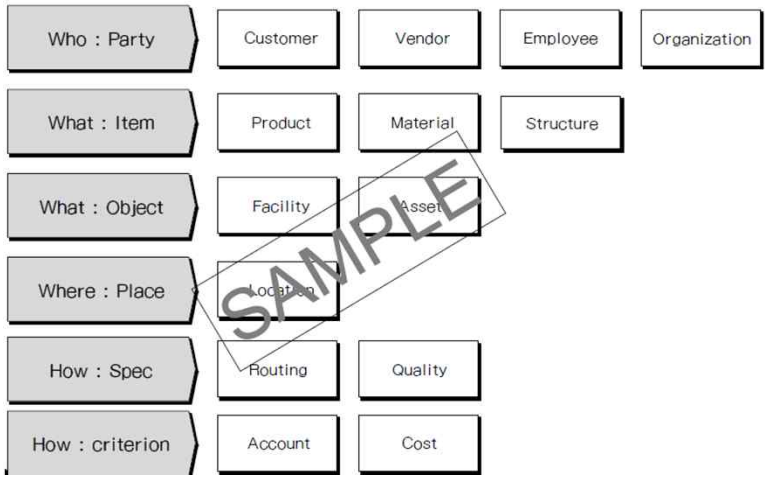
참고문헌 : 기업의 핵심데이터 관리 효율화 방안 Master Data (IBM, 2009.2)

Module-07. 마스터 데이터 통합

□ 마스터 데이터 관리 개요

▷ 마스터 데이터 정의

- ⊙ 일반적으로 시간이 경과함에 따라 지속적으로 발생하는 Event Data와 이력으로 구성된 Transaction Data의 상대적 개념
- ⊙ 기업이 비즈니스를 수행하는데 있어 매우 중요한 의미를 가지는 정보
- ⊙ 자주 변하지 않고 자료 처리 운용에 기본 자료로 제공
- ⊙ 여러 시스템에서 활용 되어지고 중앙에서 통제, 관리되어야 하는 표준 참조 데이터



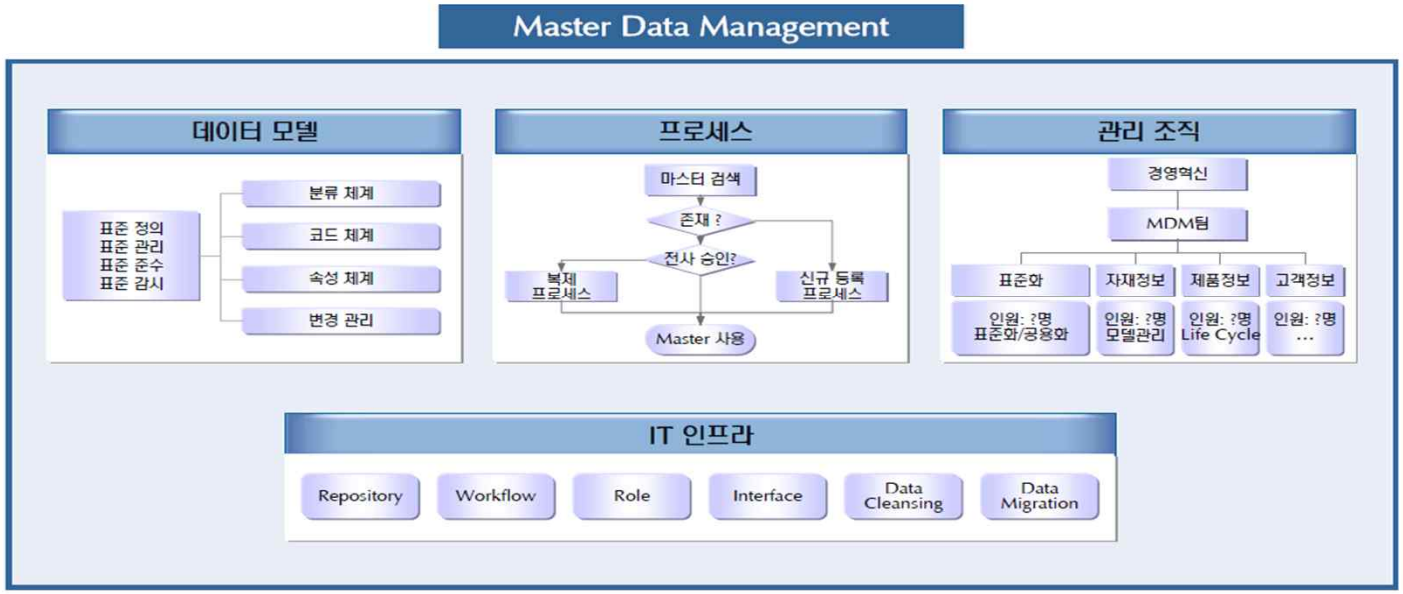
참고문헌 : 기업의 핵심데이터 관리 효율화 방안 Master Data (IBM, 2009.2)

Module-07. 마스터 데이터 통합

□ 마스터 데이터 관리 개요

▷ 마스터 데이터 관리(MDM, Master Data Management) 정의

⊙ MDM은 전사에서 활용되는 마스터데이터를 통제하기 위한 관리 체계로, 데이터 모델, 프로세스, 관리 조직, 이를 운영하기 위한 시스템 인프라를 포함



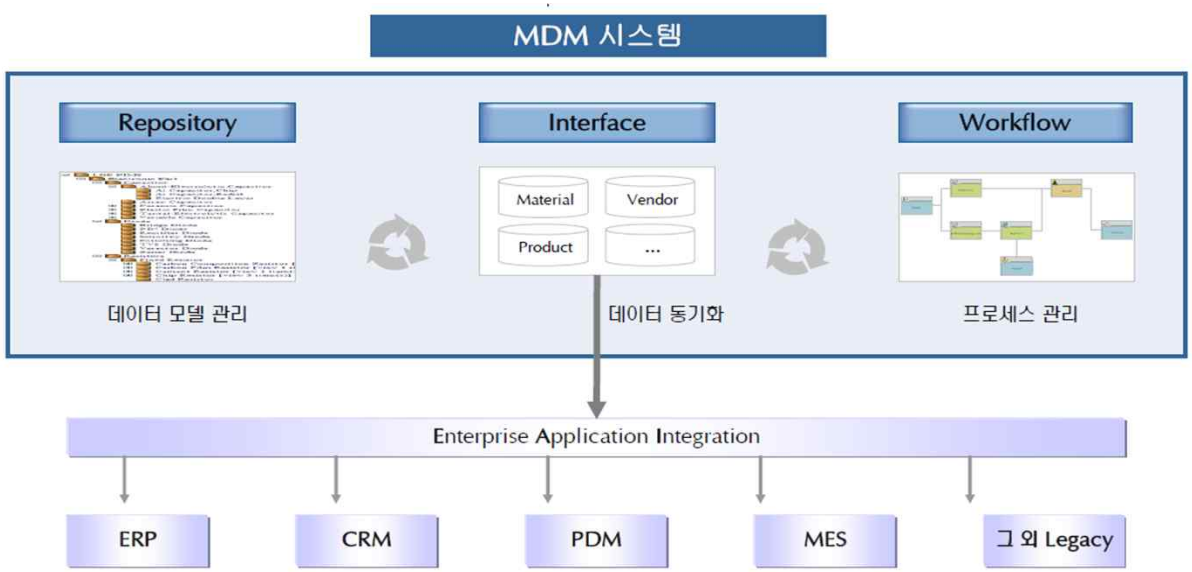
참고문헌 : 기업의 핵심데이터 관리 효율화 방안 Master Data (IBM, 2009.2)

Module-07. 마스터 데이터 통합

□ 마스터 데이터 관리 개요

▷ 마스터 데이터 관리(MDM, Master Data Management) 시스템 정의

⊙ MDM 시스템은 데이터 모델 관리를 위한 Repository, 프로세스 관리를 위한 Workflow, 타 시스템으로의 데이터 전송을 위한 Interface 영역으로 구성



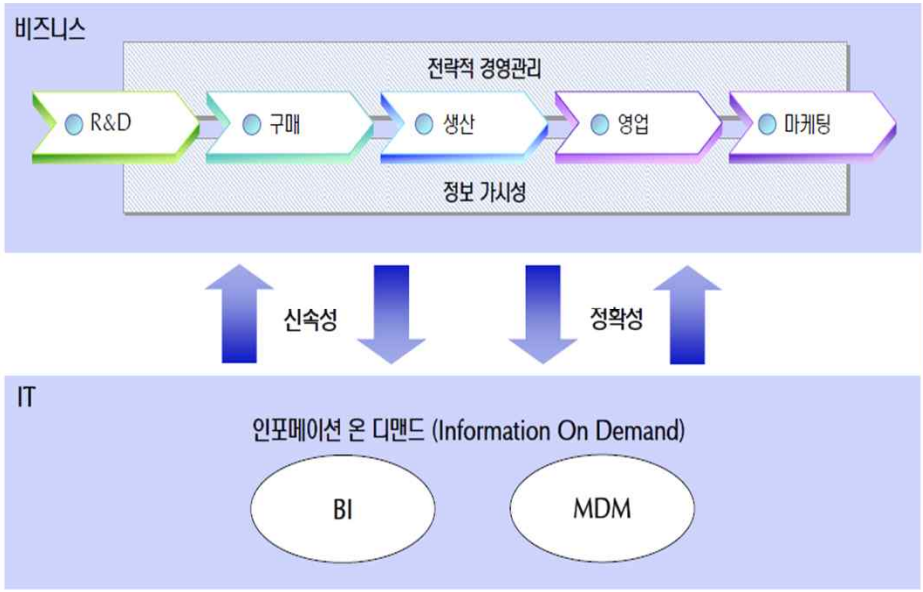
참고문헌 : 기업의 핵심데이터 관리 효율화 방안 Master Data (IBM, 2009.2)

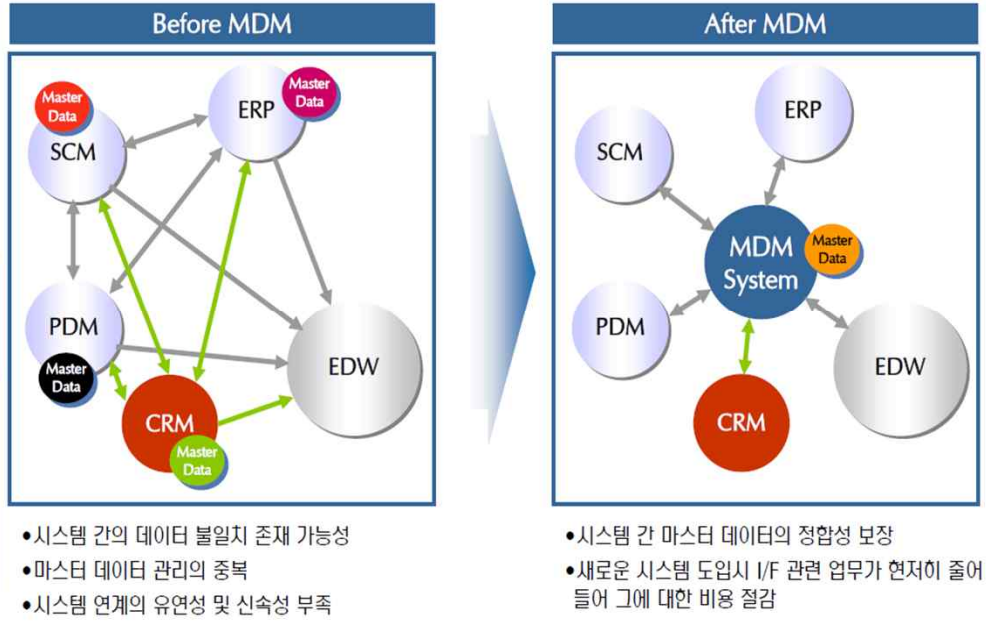
Module-07. 마스터 데이터 통합

□ 마스터 데이터 관리 개요

▷ 마스터 데이터 관리(MDM, Master Data Management) 역할

⊙ MDM은 단순한 시스템이 아니라 지속적인 비즈니스 혁신을 지원하기 위하여 혁신도구,혁신의 인프라 역할을 수행해야 하며, 실시간 의사결정 지원을 위해 즉각적으로 정보를 제공





- 시스템 간의 데이터 불일치 존재 가능성
- 마스터 데이터 관리의 중복
- 시스템 연계의 유연성 및 신속성 부족

- 시스템 간 마스터 데이터의 정합성 보장
- 새로운 시스템 도입시 I/F 관련 업무가 현저히 줄어들어 그에 대한 비용 절감

참고문헌 : 기업의 핵심데이터 관리 효율화 방안 Master Data (IBM, 2009.2)

Module-07. 마스터 데이터 통합

□ 마스터 데이터 관리 개요

▷ 마스터 데이터 관리(MDM, Master Data Management) 기대효과

⊙ MDM을 통하여 전사 정보 View를 단일화하여 원활한 소통을 추진할 수 있고, 글로벌 운영의 기반을 구축할 수 있으며, 분석정보의 신뢰성과 신속성을 제고

참고문헌 : 기업의 핵심데이터 관리 효율화 방안 Master Data (IBM, 2009.2)

Module-07. 마스터 데이터 통합

□ 마스터 데이터 관리 베스트 프랙티스

▷ 빅데이터와 MDM 통합 필요성

⊙ 이탈고객관리 : Twitter, Facebook, 음성기록 등의 빅데이터와 고객 마스터 데이터의 통합으로 얻어진 통찰력을 활용하여 고객 이탈 모델의 예측력을 강화할 수 있음

⊙ 리스크 관리 : 비구조화된 재정정보를 활용하여 회사의 소유구조의 변화와 MDM 계층구조를 갱신

⊙ 고객 분류 : 소셜 미디어 등의 정보를 이용하여 고객 분류와 고객 행동을 모델링

⊙ 차선책 제안 : 고객이 회사와 인터랙션한 정보를 바탕으로 cross-sell 혹은 up-sell을 제안

⊙ 콜 센터의 운용비용 절감 : 고객의 인구통계학적 정보와 그들의 질문에 대한 사항을 이해하여 미리 대처함으로써 콜센터에서 고객 응대시간과 빈도를 줄일 수 있음

⊙ 중복 방지 : 마스터 데이터의 중복성을 줄이는 것은 MDM의 주요 목적 중 하나임. 마스터 데이터 중복방지 기능은 빅데이터 플랫폼 안에 포함되어야 함

⊙ 선호도 관리 : 고객 선호도는 MDM내에서 잘 관리되어야 함. 빅데이터 분석 플랫폼에서는 캠페인 시 MDM내의 선호도를 기반으로 고객을 선정하게 됨

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-07. 마스터 데이터 통합

□ 마스터 데이터 관리 베스트 프랙티스

▷ 빅데이터 분석 지원을 위한 마스터 데이터 품질 향상

⊙ 빅데이터와 MDM과의 관계는 MDM과 데이터 웨어하우징의 관계와 유사함. MDM을 사용하여 데이터 웨어하우스로 입력되는 데이터를 정제함

⊙ 빅데이터 분석 프로젝트에서 고품질의 마스터 데이터가 필요한 이유

- 재료(materials) 사례 : 소비재 제품 기업들은 소매상들의 POS 로그를 활용하여 어느 제품이 어느 상점에서 많이 팔리고 있는지를 분석하는데 일관성있는 제품 정보(MDM)를 활용할 수 있어야 함
- 자산(assets) 사례 : 사전 예방관리 모델을 세우기 위해 실시간 센서 데이터를 사용하는데, 예방관리 모델에는 수백만 건의 이벤트 레코드들이 수백 개의 시설에 설치된 수천 개의 센서들로부터 유입됨. 특정 자산관리 시스템이 동일한 펌프에 대하여 서로 다른 이름을 가질 때 해당 시스템의 장비들로부터 해당 이벤트를 추출하는데 어려움이 있으므로 모델의 고장예측능력을 저하시키게 됨
- 고객(customer) 사례 : 빅데이터 분석에 의해 고객이탈에 관한 예측이 가능한데, 고객의 성별, 수입, 나이, 거주지 등에 의존하는 바가 큼

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-07. 마스터 데이터 통합

□ 마스터 데이터 관리 베스트 프랙티스

▷ 빅데이터 활용을 위한 마스터 데이터 품질 향상

⊙ 빅데이터를 활용하면 마스터 데이터의 품질을 향상시킬 수 있음

- 고객 마스터 데이터 품질 향상에 CDRs(Call Detail Recordings, 통화내역기록) 혹은 위치정보를 활용해 품질을 향상시킴

예) 통신운영사는 고객에 대하여 이름, 주소, 기타 인구통계학적인 정보를 가짐. 모바일 사업자는 가입자에 대한 대략적인 정보만 가지고 있음. 빅데이터 분석 부서에서는 통화 패턴 즉 전화발신자, 수신자, 통화시간, 통화빈도를 분석하여 가입자간의 관계를 알 수 있음

▷ 핵심 참조 데이터의 일치성과 품질 향상을 통한 빅데이터 거버넌스 프로그램 지원

⊙ 참조 데이터는 회사내 다른 응용들이 참조할 수 있도록 일종의 룩업(lookup) 테이블에 저장되며, 정적인 특성을 가진다는 면에서 마스터 데이터와 차이가 있음

예) 참조 데이터의 예 : 국가/주/지방 코드, 통화코드, 업체코드 등

⊙ 빅데이터 거버넌스에서 데이터 로드전에 핵심 참조 데이터를 체크하는 것이 필요함. 입력되는 데이터의 편차를 표시하여 이후 검증하는 과정을 거치도록 함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-07. 마스터 데이터 통합

□ 마스터 데이터 관리 베스트 프랙티스

▷ 마스터 데이터 관리와의 통합 수준을 결정하기 위한 소셜 미디어 플랫폼 정책 고려

⊙ 소셜 미디어 데이터를 고객 마스터 데이터와 통합하기 전에 프라이버시 정책과 소셜 미디어 플랫폼의 규제나 정책을 먼저 고려해야 함

순서	Facebook 플랫폼 정책(2012년3월6일)	MDM에의 영향
1	당신의 앱상에서 어떤 사용자의 친구에 관한 데이터는 그 사용자와의 경험 콘테스트(교류)에서만 사용될 수 있다.	조직은 Facebook app에서 어떤 사람의 친구에 관한 데이터를 다른 용도로 사용할 수 없다.
2	사용자가 당신의 앱에 연결할 때 일정한 제한(예를 들어 전송)을 조건으로 자신의 기본적인 계정정보를 제공한다. Facebook API를 통하여 얻어진 그 밖의 모든 데이터는 해당 고객의 명시적인 동의를 얻어야 한다.	조직은 기본적인 계정정보(이름, 이메일, 성별, 생일, 현재 도시 및 프로필 사진 URL) 이외의 정보를 사용하기 전에 사용자의 명시적인 동의를 얻어야 한다.
3	당신은 Facebook 사용자 ID를 당신의 앱 이외에 어떠한 용도로도 사용할 수 없다. Facebook 사용자 ID는 당신의 응용을 구축하고 실행하는 외부 서비스에 이용될 수 있으나 이러한 서비스들이 당신의 응용을 실행하는데 필요하고, 또 이러한 서비스들이 Facebook 사용자 ID보호에 대한 의무를 가져야 한다.	조직은 MDM에 Facebook 사용자 ID를 저장할 수 있는지를 확인해야 한다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-07. 마스터 데이터 통합

□ 마스터 데이터 관리 베스트 프랙티스

▷ 마스터 데이터 관리와의 통합 수준을 결정하기 위한 소셜 미디어 플랫폼 정책 고려(계속)

순서	Facebook 플랫폼 정책(2012년3월6일)	MDM에의 영향
4	당신이 플랫폼 사용을 중지하였거나, 우리가 어떤 이유로 중단시켰다면 당신이 Facebook API의 사용을 통해 수집한 모든 데이터를 삭제해야 한다. 다만, (a) 기본적인 계정정보이거나, 또는 (b) 당사자의 명시적인 동의를 받아서 보관한 데이터의 경우는 예외이다.	조직은 MDM 데이터와 Facebook 데이터를 병합할 때 매우 신중해야 한다. 조직이 통합된 데이터를 중요 자료로 제작하여 전파한 경우에 당사자로부터 명시적인 동의를 받지 않았을 경우에는 더 큰 문제가 발생할 수 있다.
5	사용자의 친구 리스트를 한 명이 사용에 동의하였다고 해도 당신의 응용 밖에서 사용하면 안된다. 다만, 두 사람 모두 당신의 응용에 연결된 경우에는 이 연결정보를 사용할 수 있다.	조직은 Facebook app에서 어떤 사람의 친구에 관한 데이터를 다른 용도로 사용할 수 없다.
6	어떤 사용자가 자신에 관한 모든 정보를 당신으로부터 지울 것을 요청한다면 Facebook은 그렇게 할 것이며, 이를 효과적으로 수행하기 위한 방안을 제공할 것이다. 만일 당신이 Facebook에서 제시한 항목을 위반한다면 당신이 Facebook API로부터 받은 데이터를 모두 삭제하도록 요청할 것이다.	조직은 MDM 데이터와 Facebook 데이터를 병합할 때 매우 신중해야 한다. 조직이 통합된 데이터를 중요 자료로 제작하여 전파한 경우에 당사자로부터 명시적인 동의를 받지 않았을 경우에는 더 큰 문제가 발생할 수 있다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-07. 마스터 데이터 통합

□ 마스터 데이터 관리 베스트 프랙티스

▷ 마스터 데이터 확충을 위한 비구조적 텍스트 의미 추출

⊙ 전통적인 MDM 시스템은 고객, 공급업체, 자재, 자산, 기타 단체 등 수많은 구조적 데이터 소스로부터 구조적 데이터를 수집하는데, 빅데이터의 출현으로 MDM 프로젝트는 점차적으로 소셜 미디어, 이메일, 콜센터 보이스 전사(transcripts), 에이전트 로그, 스캔된 텍스트 등과 같은 대량의 비구조적 텍스트로부터 숨겨진 유용한 가치를 이끌어낼 수 있음

예) 고객 MDM 및 전자메일 통합

⊙ 비구조화된 텍스트로부터 마스터 데이터 보완

- 1단계. 관리되어야 할 각 항목에 대하여 속성 정의

  예) 고객 = 이름, 회사, 거주도시, 거주지역, 이메일

- 2단계. MDM 저장소 및 기타 소스로부터 유입된 각 속성에 대하여 사전 파일 생성

  예) 이름, 회사, 거주도시, 거주지역, 이메일 항목의 값(value) 모음

- 3단계. 퍼지 매칭과 비즈니스 규칙을 기반으로 해당 용어에 주석 처리

  예) 텍스트 주석 기술과 사전을 기반으로 강조하는 단어를 찾아냄

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-07. 마스터 데이터 통합

□ 마스터 데이터 관리 베스트 프랙티스

▷ 마스터 데이터 확충을 위한 비구조적 텍스트 의미 추출

⊙ 비구조화된 텍스트로부터 마스터 데이터 보완(계속)

- 4단계. 비구조적 텍스트로부터 추출한 주석들로 구성된 MDM 시스템에의 질의

  예) MDM 쿼리 : 이름 = "영희" 또는 "철수" 또는 "호동" 또는 "길동"

- 5단계. 비구조적 엔티티에 대하여 레코드를 생성

  예) 이름 = "영희", 고용주 = "호동", 거주지역 = "서울", 이메일 = "fit@hanmail.net"

- 6단계. 기존의 MDM 레코드와 새롭게 생성된 레코드의 연계

  . 새롭게 생성된 레코드가 신뢰도가 높은 경우 자동으로 MDM 시스템에 입력되는데, 신뢰도가 낮을 경우에는 자동으로 거부되고 중간 정도의 신뢰도를 가진 경우에는 관리자가 수동으로 확인한 다음에 결정함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-08. 메타데이터

□ 메타데이터 개요

▷ 메타데이터 정의

⊙ '데이터에 관한 데이터' : 정보자원의 속성을 기술하는 데이터
예) 테이블의 칼럼, 도메인(자릿수, 타입, 초기값, 허용값 등) 내역
예) 옷에 있어서의 색깔, 칫수, 유형(원/투피스), 재질 등

⊙ 메타데이터란 실제로 저장하고자 하는 데이터는 아니지만, 이 데이터와 직접적으로 혹은 간접적으로 연관된 정보를 제공하는 데이터를 나타냄

▷ 메타데이터 필요성

⊙ 메타데이터를 사용하면 사용자가 원하는 데이터가 맞는가를 확인할 수 있고, 쉽고 빠르게 원하는 데이터를 찾아낼 수 있음

⊙ 데이터를 소유하고 있는 측면에서는 관리의 용이성을, 데이터를 사용하고 있는 측면에서는 검색의 용이성을 보장받을 수 있기 때문에 메타데이터의 필요성이 더욱 높아지고 있음

참고문헌 : 데이터 통합 및 거버넌스_메타데이터 (데이터스트림즈)

Module-08. 메타데이터

□ 메타데이터 개요

▷ 메타데이터 활용

⊙ 동시다발적인 정보시스템의 개발과 전사 데이터 관리의 마인드 결여, 데이터 관리 인력의 부재 등은 기업의 데이터 품질을 저해하는 요소임. 이러한 결과로 데이터의 중복 및 조직, 업무, 시스템 별 데이터 불일치가 발생하며, 차세대등 시스템 개발 시 데이터에 대한 의미 파악 및 지연으로 데이터 통합 및 시스템의 개발에 막대한 지장을 초래함. 시스템의 개발 후 정보시스템의 변경 및 유지보수에 상당한 애를 먹고 있음

⊙ 데이터 표준화 및 규격화를 위한 데이터 표준화 정책 및 지침, 관리프로세스와 조직을 기반으로 한 표준용어, 도메인, 모델, 데이터베이스의 라이프 사이클(life Cycle)을 관리하며, 각종 툴(Case Tool, ETL, BI, 형상관리 등)과의 연계를 통하여 표준화의 변경에 대한 영향분석을 수행

⊙ 이로 인해 명칭의 통일로 인한 명확한 의사소통이 증대하며, 필요한 데이터의 소재 파악에 소요되는 시간 및 노력이 감소, 일관된 데이터 형식 및 규칙의 적용으로 인한 데이터 품질향상, 정보시스템 간 데이터 인터페이스 시 데이터 변환 등의 비용이 감소 등의 효과를 볼 수 있음

참고문헌 : 데이터 통합 및 거버넌스_메타데이터 (데이터스트림즈)

Module-08. 메타데이터

□ 메타데이터 개요

▷ 메타데이터 발전방향

⊙ 메타데이터를 기반으로 한 데이터 품질관리

- 메타데이터를 활용하여 용어의 속성(도메인)을 정의하고, Case Tool(모델관리)과의 연동을 통한 데이터간의 참조 관계 등의 정보를 정의하여 데이터 품질 활동에 활용

⊙ 메타데이터를 기반으로 한 마스터데이터 관리

- 조직 전반적인 기업 정보관리의 일환으로 Enterprise Information Management 전략은 메타데이터 기반의 마스터 데이터 관리 전략으로 연결되며, 고객 데이터 통합 및 상품정보 통합 등 마스터 데이터 통합으로 구현됨
- 기존의 메타데이터 관리시스템은 통합코드 관리를 수행하며, 코드의 변경 및 주요 코드 정보의 Sync를 담당
- 메타데이터 관리는 회사의 정보 인프라에 중요한 부분인 재사용, 일관성, 무결성 및 공유성을 지원하도록 하는 메타데이터 리포지토리와 함께 애플리케이션 개발 프로젝트에 SOA(서비스지향 아키텍처)로 확장하게 됨

⊙ 메타데이터를 기반으로 한 비즈니스 프로세스 관리

- BPM EAI 등 대부분의 솔루션들은 나름대로의 워크플로우(Workflow) 기능을 가지고 있음. 이는 비즈니스 프로세스의 설계 및 운영 그리고 관리에 대한 중복의 문제를 가지게 됨. 이에 메타데이터 중심의 비즈니스 프로세스와 데이터를 다양한 관점에서 편집 가능한 형태로 잡아내고 관리할 수 있게함

참고문헌 : 데이터 통합 및 거버넌스_메타데이터 (데이터스트림즈)

Module-08. 메타데이터

□ 빅데이터와 메타데이터

▷ 빅데이터에서의 메타데이터 활용방향

⊙ 핵심 빅데이터 용어에 대한 비즈니스 측면의 정의를 담은 사전 제작

⊙ Apache Hadoop 내에서 진행중인 메타데이터에 대한 지원 이해

- Hadoop 메타데이터와 관련된 핵심사항

예) HDFS의 기술적인 아키텍처 이해 : 1개의 namenode와 여러 개의 datanode를 이용한 데이터 분실 방지, 확장성

⊙ 비즈니스 용어사전에서 민감한 빅데이터 표시

⊙ 빅데이터 저장소로부터 기술적인 메타데이터를 가져옴

⊙ 비즈니스 용어사전에서 용어에 관련된 데이터 소스를 링크함

⊙ 운영 메타데이터를 활용하여 빅데이터의 이동을 모니터링함

⊙ 데이터 흐름과 영향성 분석을 위해 기술적인 메타데이터를 유지함

⊙ 전사적 검색을 지원하기 위해 비구조적 문서로부터 메타데이터를 수집함

⊙ 빅데이터를 감안하여 기존 메타데이터의 역할을 확장함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-09. 빅데이터 수명주기 관리

□ 정보수명주기 관리(ILM) 개요

▷ 정보수명주기 관리 필요성

⊙ 정보관리에 대한 전체적인 TCO(총기회비용, Total Cost Opportunity) 절감

⊙ 가치에 대한 데이터 단편화(Segmentation)

⊙ 정보의 이동, 유지, 삭제에 대한 의사결정 지원 필요

⊙ 콘텐트 중심의 정보관리 추세

⊙ 포괄적인 전사적 ECM(전사콘텐트관리, Enterprise Contents Management) 프레임워크 제공

▷ 정보수명주기 관리 목적

⊙ 기업이 정보활용 및 보관을 위해서 기존 시스템의 교체 및 업그레이드 기준 제공

⊙ 포화상태의 데이터를 경제적으로 활용하도록 사용빈도에 따라서 데이터 저장매체를 생명주기에 따라 저장

⊙ 종이문서의 보관규정(회계문서, 이메일자료, 공증자료 등 법적근거가 되는 자료 등)에 따른 법적 규제 준수

Module-09. 빅데이터 수명주기 관리

□ 정보수명주기 거버넌스

▷ 정보수명주기 거버넌스 정의

⊙ 수명주기 동안 정보를 통제하여 비용 절감 및 규제 준수를 보장하는 것

▷ 정보수명주기 거버넌스 솔루션

⊙ IT의 스마트 아카이브

- 오피스 아카이브 및 협업
- ERP 및 정형 정보 아카이브

⊙ 법·제도를 준수하는 eDiscovery(전자적 디스커버리) 관리

- eDiscovery 프로세스 관리
- 사례관리 및 분석

⊙ RIM(Reference Information Model)을 위한 기록과 폐기관리

- 보관주기 정책 및 스케줄 관리
- 기업 기록물 관리

⊙ CIO를 위한 정보 폐기 및 거버넌스

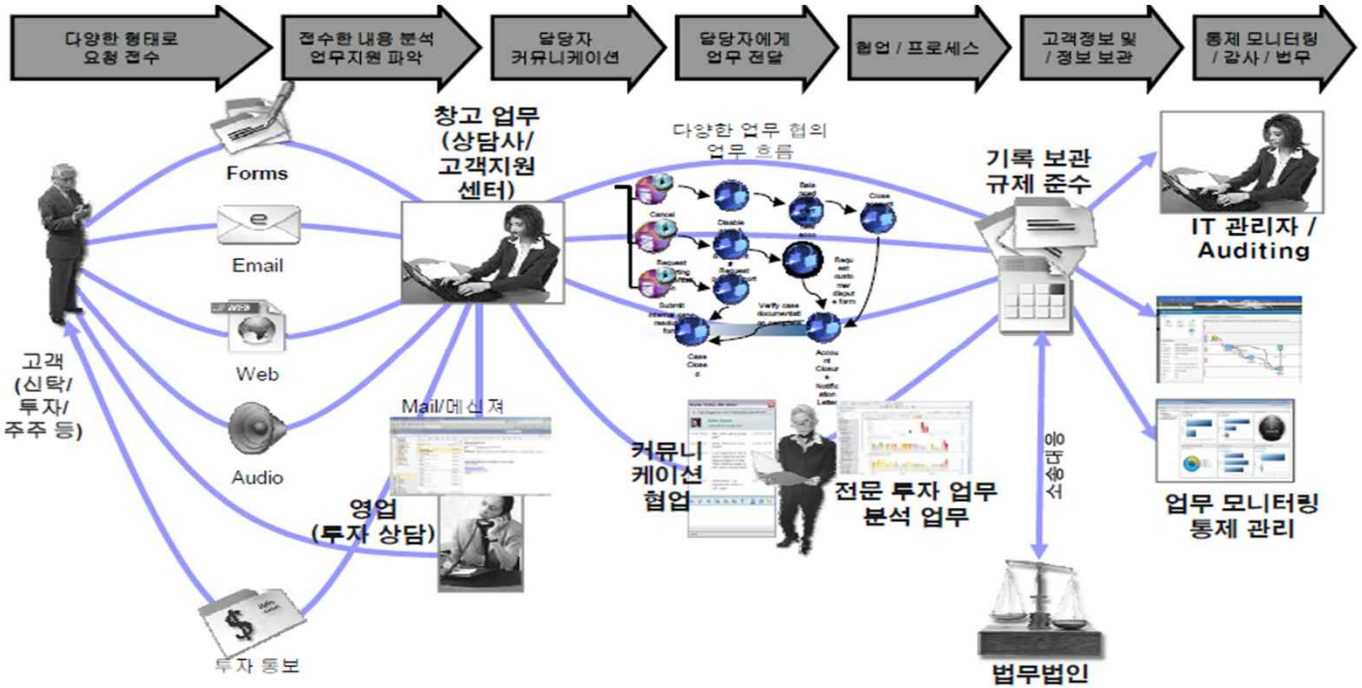
- 정보 재배치 및 폐기
- 규제 준수 및 리스크 관리

참고문헌 : 빅데이터 시대의 정보 주기 관리를 통한 스마트 워크 구현전략 (IBM, 2012)

Module-09. 빅데이터 수명주기 관리

□ 정보수명주기 거버넌스

▷ 정보수명주기 거버넌스 처리 흐름도



참고문헌 : 빅데이터 시대의 정보 주기 관리를 통한 스마트 워크 구현전략 (IBM, 2012)

Module-09. 빅데이터 수명주기 관리

□ 빅데이터 수명주기 관리 베스트 프랙티스

▷ 해당 지역의 규제와 비즈니스 요구에 따라 빅데이터 보관기간을 결정

⊙ 통신데이터는 대다수 나라와 지역에서 규제의 대상이 됨

국가	규정
영국	영국 내무부는 테러방지, 범죄와 보안관련 법률 등 국가안보와 관련된 장치를 위해 통신 데이터 보관에 대한 자율적 규정을 발표했다. - 이름, 생년월일, 설치 및 과금용 주소, 신용카드번호, 전화번호, 글로벌 휴대폰 식별번호(IMEI), 글로벌 모바일 가입자 식별번호(IMSI) 등을 12개월간 보관한다. - 전화 관련 데이터(발신자번호, 수신자번호, 통화기간, 위치정보)는 12개월간 보관한다. - SMS, EMS, MMS 데이터 관련 발신자번호, 수신자번호, 위도와 경도로 구성되는 위치정보는 6개월간 보관한다. - Email 데이터(사용자 이름, 로그인/로그아웃 시간, IP주소, from/to 이메일주소)는 6개월간 보관한다. - 웹활동, 사용시간과 데이터, IP 주소, URL 등의 정보는 4일간 보관한다.
유럽연합	유럽연합의 회원 국가들은 최소 6~24개월 동안 데이터를 보관해야 한다. 그러나 해당 규정은 나라에 따라 달리 적용된다.
이탈리아	이탈리아는 다음과 같이 통신 데이터 유지법을 시행하고 있다. 1. 이동 및 유선통신 데이터는 29개월 동안 보관되어야 한다. 2. 인터넷 사업자는 최저 6개월간 보관해야 하며, 이 기간은 6개월 연장 가능하다. 3. 모든 휴대폰 사업자와 선불 폰 판매자는 가입 시 제시한 신분증 사본을 보관해야 한다.(이탈리아 법은 휴대폰 이용자의 개별인식을 요구하기 때문이다.)

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-09. 빅데이터 수명주기 관리

□ 빅데이터 수명주기 관리 베스트 프랙티스

▷ 해당 지역의 규제와 비즈니스 요구에 따라 빅데이터 보관기간을 결정

⊙ 통신데이터는 대다수 나라와 지역에서 규제의 대상이 됨

국가	규정
독일	독일 의회는 통신 데이터를 6개월 보관할 것을 요구하는 법안을 통과시켰으나 연방 대법원이 개인의 프라이버시에 대한 과도한 침해라는 사유로 위헌판결을 내렸다.
프랑스	인터넷과 통신사업자에게 1년간 데이터를 보관할 것을 요구한다.
미국	미국은 유럽과 비슷한 형태의 데이터 보유규정을 가지고 있지 않다.
태국	모든 통신사업자는 90일간의 트래픽 데이터를 보유해야 한다. 트래픽 데이터는 위치, 트래픽 양, IP주소, URL 등을 포함해야 한다.
인도	인도정부는 ISP 사업자들에게 사업을 시작하기전, 인터넷 서비스의 규정에 대한 인허가 계약에 사인을 하게 하였다. 계약은 네트워크상에서 이루어진 의사소통에 대한 기록을 1년간 보유할 것을 명시한다. 2009년 정보기술법이 제정되었지만 아직 구체적인 규정이 발효되지 않았다. 그러나 관련 산업계는 이미 관련 데이터(SMS Message, 전화통화 로그, 이메일 헤더, 웹 요청 등)를 3~12개월간 보유하고 있다.
호주	호주는 통신 데이터에 대한 어떠한 규정도 가지고 있지 않았지만, 정부는 시행 가능성을 모색하고 있다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-09. 빅데이터 수명주기 관리

□ 빅데이터 수명주기 관리 베스트 프랙티스

▷ 법적으로 민감한 정보를 문서화해서 보관하고 정보요구에 대응

⊙ 거의 모든 기업과 조직들은 소송이나 정부의 조사에 대비하여 많은 양의 잠재적 증거를 보관할 것을 요구받음

⊙ 정보 거버넌스 프로그램은 법적 리스크와 비용을 관리하는 동시에, 의무사항에 대해 정보 관리인과 의사소통하고, 증거를 수집하며, 결과를 분석해야 함

⊙ 기업 내부에서 빅데이터가 보편화되면서 법적 문제도 커지는 추세임
예) 오일 유출로 인해 고소당한 회사는 그들이 적절한 조치를 취했다는 것을 증명할 센서 데이터의 보관

▷ 빅데이터의 압축보관을 통해 IT비용 절감과 응용의 성능 개선을 동시에 달성

⊙ 스마트 미터기 데이터, 센서 데이터, RFID 데이터, 웹로그 등은 관계형 데이터베이스, 파일시스템, NoSQL 데이터베이스, Hadoop 등에 저장됨
예) 중간 규모 회사에서 Hadoop 활용 아카이빙의 경제성 분석

경우의 수	금액(달러)
기존 스토리지의 테라바이트당 데이터 저장비용	20,000
Hadoop을 통한 저장 비용	3,000
소요되는 복사 회수(Hadoop 사용 시 3군데 복제)	3
Hadoop을 통한 전체 데이터 저장 비용	9,000
비용 절감액	11,000

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-09. 빅데이터 수명주기 관리

□ 빅데이터 수명주기 관리 베스트 프랙티스

▷ 실시간 스트리밍 데이터의 수명주기 관리 방법과 효과

⊙ 데이터가 매우 빠른 속도로 유입되는 경우, 팀은 그 데이터가 가치가 있는지, 저장될 필요가 있는지를 구분해야 함

⊙ 센서로부터 읽은 것에 이상현상(이벤트)이 발생하면 스트리밍 분석 응용은 그 이벤트 전후의 데이터를 저장하여 분석할 수 있음

▷ 규제준수를 위한 소셜미디어 기록 보관과 전자정보 공개 요구에 대응이 가능

⊙ 미국의 수집통신 법안(U.S Stored Communications Act, 1986) : 서비스 제공자가 사용자의 커뮤니케이션 정보를 누설하는 것을 금하며, 이 법안은 소셜 미디어에도 동일하게 적용됨

⊙ 미국의 금융산업 규제기구(U.S Financial Industry Regulatory Authority, FIRNA) : 금융기관이 소셜 미디어 사이트를 통해 고객과 커뮤니케이션한 기록을 보관할 것을 요구

▷ 규제 혹은 사업상 필요하지 않은 정보의 방어적인 폐기

⊙ 빅데이터 거버넌스 프로그램은 데이터 보유 스케쥴에 근거하여 법적으로 필요하지 않은 빅데이터를 삭제하는 정책을 수립하여 시행해야 함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-10. 성숙도 측정

□ 국내 빅데이터 역량진단모델

▷ NIA(한국정보화진흥원) 빅데이터 분석 활용센터에서 제시한 모델

▷ 개발배경

⊙ 빅데이터 역량이 새로운 미래 경쟁력으로 부각되고 있는 시대적인 변화에 부응하여, 조직의 현재 상태를 점검하고 기술/조직/경영 전반에 걸친 도전과제들을 해결

⊙ 빅데이터 역량을 구성하는 하위 역량 차원에 대한 체계적인 검토와 진단을 통해 빅데이터 환경의 급격한 변화에 의한 불확실성을 줄임

⊙ 빅데이터 우량 조직으로 거듭날 수 있는 전략적인 기준 제시

▷ 기대효과

⊙ 전략 가이드 및 비교 점검 기준 : 조직들에게 전략적인 가이드 라인 역할을 하고, 이미 빅데이터 전략을 수립한 조직들에게는 비교 점검할 수 있는 기준이 됨

⊙ 투자 우선순위 결정 기준 : 계층별 성숙도 점수의 민감도 분석(sensitivity analysis)을 통해 투자우선순위 결정에 도움을 줄 수 있음

⊙ 빅데이터 핵심 성공 요인 점검 및 관리 : 비즈니스 목표와 연계하여 기술, 조직, 경영 전반에 걸친 빅데이터 핵심 성공 요인(critical success factor)을 점검 및 관리할 수 있음

참고문헌 : 빅데이터 역량진단모델 개발 및 시범적용 배포 (NIA, 2013.12)

Module-10. 성숙도 측정

□ 국내 빅데이터 역량진단모델

▷ 기대효과(계속)

⊙ 빅데이터 비즈니스 모델 활성화 : 다양한 빅데이터를 비즈니스 가치를 창출할 수 있는 자원으로 활용할 수 있는 비즈니스 모델 활성화

⊙ 지속적 경쟁우위 달성 : 빅데이터에 내재되어 있는 유용한 정보와 지식을 효과적으로 활용하여 비즈니스 인텔리전스 역량을 통한 지속적인 경쟁적 우위 달성

⊙ 글로벌 빅데이터 경쟁력 제고 : 국내 조직들의 빅데이터 역량을 강화함으로써 해외 기업들에 대한 빅데이터 기술 종속을 막고 한국의 총체적인 글로벌 빅데이터 경쟁력을 제고

참고문헌 : 빅데이터 역량진단모델 개발 및 시범적용 배포 (NIA, 2013.12)

Module-10. 성숙도 측정

□ 국내 빅데이터 역량진단모델

▷ 빅데이터 역량진단모델의 대항목 개념 및 하위 중항목

대항목	설명	중항목
1. 전략수립역량	현재상황(as-is)에서 어떤 빅데이터를 어떻게 분석하여 어디에 활용함으로써, 비즈니스 목표(to-be)를 효과적/효율적으로 달성할 수 있는지에 대한 전략 및 계획 수립 역량	1.1 빅데이터 이해도 1.2 활용방안 수립 정도 1.3 실행계획의 구체화 정도
2. 추진역량	빅데이터 활용모델을 구현하여 비즈니스 가치를 실현할 수 있도록 지원하는 전반적인 경영 능력 •내/외부 빅데이터를 획득하고 관리하며, 분석 및 활용 역량을 강화하기 위해 조직의 자원을 배분하고 조직화하며, 전반적인 지원 업무 프로세스를 정립하는 역량으로 구성됨	2.1 데이터 확보 및 관리 2.2 조직적 지원 2.3 프로세스 정립
3.분석역량	빅데이터의 특성(3V+1C)에 부합하는 처리 및 분석 기술을 적용하여, 비즈니스 시나리오에 맞는 정보를 추출하고 가공할 수 있는 전문 역량	3.1 빅데이터 분석 인프라 3.2 빅데이터 분석 전문인력 3.3 시스템 운영 및 관리
4. 활용역량	빅데이터 분석을 통해 획득한 정보 및 지식이 조직 전체에 확산되도록 지원하고 실질적으로 활용하는 역량	4.1 활용 지원 4.2 활용 범위
5. 혁신역량	비즈니스 목표 대비 달성된 수준을 평가하여 차이(Gap)의 원인을 규명하고 개선방안을 정립하여 다음 전략 및 계획에 반영하는 학습 능력 •일상적인 관리수준을 넘어 조직 내외부의 변화를 감지하고, 혁신적인 비즈니스 아이디어를 창출하여 비즈니스 성과의 최적화를 실현하는 역량	5.1 평가 및 반영 체계 5.2 변화와 혁신

참고문헌 : 빅데이터 역량진단모델 개발 및 시범적용 배포 (NIA, 2013.12)

Module-10. 성숙도 측정

□ 국내 빅데이터 역량진단모델

▷ 빅데이터 분석 및 활용역량수준

역량수준	1단계 초기(Initial)	2단계 인식기(Repeatable)	3단계 정립기(Defined)	4단계 관리기(Managed)	5단계 혁신기(Innovative)
전반적 특성	빅데이터에 대한 관심은 있지만 이해가 부족하고 모든 영역에서 준비가 미흡한 단계	빅데이터의 활용가치를 인식하고 도입전략을 추진하고 있지만 추진 환경 및 제반 역량이 미흡한 단계	빅데이터 지식가치 사슬이 정립되어 가고 있지만 빅데이터 분석결과의 실제적인 활용은 다소 미흡한 단계	빅데이터 분석 결과를 활용하고 있지만 활용성과에 대한 분석과 피드백이 미흡하고 혁신적 활용 역량이 부족한 단계	빅데이터를 통해 조직 의사결정의 질이 제고되고, 도출된 지식을 활용한 혁신적 서비스가 창출되어 새로운 사업기회를 제공하는 단계
조직/경영/기술/구체적특성	o 비즈니스 이슈에 대한 실험적 시도 o 구체적인 빅데이터 활용 전략 부족 o 경영층으로부터 특별한 지원을 받지 못함 o 기존 기술에 의존하면서 비용절감을 위해 오픈소스 또는 클라우드 기술을 조금 시험해보는 수준	o 이전 파일럿 프로젝트 수행 경험으로 교훈을 얻고 이를 새로운 비즈니스 요구사항에 적용 o 새로운 프로젝트를 위한 예산/중간 관리자의 지원도 받을 수 있게 됨 o 전사적 지원은 받지 못하는 상태 성과에 대한 객관적 측정기준 부재	o 예산지원이 되는 중소규모의 빅데이터분석 프로젝트를 반복적으로 실행 o 사업본부 수준의 지원(데이터 수집/통합/관리 프로세스가 사업본부 단위에서 발생) o 전사적인 데이터 거버넌스나 보안정책 부족 o 내부직원 외에 외부 전문가 서비스 활용	o 빅데이터 분석 프로세스가 정립됨 o 여러 사업본부를 포괄하는 수준의 빅데이터 전략이 수립됨 o 고위 임원으로부터의 지원을 받게 됨 o 전사적 성과측정 도구와 방법론이 정립되어 투자의사결정에 반영됨	o 지속적이고 사전 조율된 전사적 빅데이터 분석 프로세스 개선작업과 가치실현이 진행됨 o 전사적으로 절차를 정의하는 전략지침이 만들어지고 강력한 경영진의 지원을 받음

참고문헌 : 빅데이터 역량진단모델 개발 및 시범적용 배포 (NIA, 2013.12)

Module-10. 성숙도 측정

□ 국내 빅데이터 역량진단모델

▷ 역량진단모델 유형 개요

⊙ 빅데이터 역량진단 모델은 활용형태에 따라 내부 분석환경 구축형, 외부 분석서비스 활용형, 내부구축/외부서비스 혼합형 등 세 가지 유형으로 구분

유형 및 활용형태	설명
내부 분석환경 구축형	빅데이터 분석 시스템과 운영 환경을 자체적으로 구축하여 확보한 또는 확보하기를 원하는 유형임. 조직 내부 데이터의 보안 및 관리 문제로 외부 서비스를 활용하기 어렵거나, 분석 요구 사항을 외부 서비스 업체에서 지원하지 못하거나, 또는 자체 구축된 시스템을 기반으로 다른 조직 및 대국민 서비스를 수행하는 경우 등에 해당함
외부 분석서비스 활용형	빅데이터 분석 환경을 자체적으로 구축하지 않고 외부 서비스 업체의 분석 환경을 활용하거나 활용하려는 유형임. 다섯 가지 빅데이터 역량 차원 중 분석역량이 외부 서비스 업체 선정 및 관리, 데이터의 안전한 전송 등에 대한 평가 항목으로 대체되어 추진역량에서 측정됨
내부구축/외부서비스 혼합형	빅데이터 프로젝트의 분석 목표 및 데이터의 보안 요구 수준에 따라 내부 분석환경과 외부 서비스를 혼합적으로 활용하거나 활용하려는 유형임. 내부 분석환경 구축형과 외부 분석서비스 활용형에 대한 평가 항목들이 종합적으로 측정됨

참고문헌 : 빅데이터 역량진단모델 개발 및 시범적용 배포 (NIA, 2013.12)

Module-10. 성숙도 측정

□ 국내 빅데이터 역량진단모델

▷ 역량진단모델 항목 및 배점

⊙ 빅데이터 역량진단 모델 유형별 대항목 배점은 다음 표와 같음

⊙ 빅데이터 분석역량을 확보하고 관리하는 방법만 다를 뿐 빅데이터 지식 가치 사슬의 선순환을 촉진하기 위해 필요한 역량 요소 및 기본적인 역량진단 프레임워크가 서로 비슷함

대항목	내부 분석환경 구축형	외부 분석서비스 활용형	혼합(hybrid)형
1. 전략수립 역량	20점	25점	20점
2. 추진 역량	20점	25점 '2.1 데이터 확보 및 관리' 1문항 '2.3 프로세스 정립' 3문항 소항목 추가	25점 '2.1 데이터 확보 및 관리' 1문항 '2.3 프로세스 정립' 3문항 소항목 추가
3. 분석 역량	20점	N/A	15점
4. 활용 역량	20점	25점	20점
5. 혁신 역량	20점	25점	20점

참고문헌 : 빅데이터 역량진단모델 개발 및 시범적용 배포 (NIA, 2013.12)

Module-10. 성숙도 측정

□ 국내 빅데이터 역량진단모델

▷ 역량진단모델 항목 및 배점

대항목(L)	중항목(M)	소항목(S)
1. 전략수립역량 (20)	1.1 빅데이터 이해도(8)	1.1.1 조직이 당면한 현안을 빅데이터를 활용하여 어떻게 해결할 수 있는지 이해하고 있습니까? (3) 1.1.2 빅데이터 활용에 대한 최고경영자와 경영진의 관심과 추진 의지가 있습니까? (3) 1.1.3 빅데이터 분석 및 활용에 대한 전략을 수립할 때 전체 조직 구성원들의 아이디어와 의견을 반영하고 있습니까? (2)
	1.2 활용방안 수립 정도(8)	1.2.1 빅데이터 도입 또는 활용을 위하여 조직 고유의 상황에 맞는 전략을 수립하고 있습니까? (예를 들어 내부에 데이터 분석환경을 구축한다거나 외부업체의 분석서비스를 활용하는 등) (2) 1.2.2 빅데이터 도입 또는 활용과 관련하여 단계적이고 구체적인 사업계획이 수립되어 있습니까? (2) 1.2.3 빅데이터 분석을 통하여 비즈니스 문제를 어떻게 해결할 수 있는지 논리적으로 이해가능하면서 구체적인 활용 방안(또는 활용 시나리오)이 도출되어 있습니까? (2) 1.2.4 수립한 빅데이터 분석 및 활용 방안은 기술적 그리고 경제적으로 충분히 실현가능합니까? (2)
	1.3 실행계획의 구체화 정도(4)	1.3.1 빅데이터 분석 및 활용을 위한 예산 확보 방안을 마련하고 있습니까? (1) 1.3.2 빅데이터 관련 전문인력 확보를 위한 구체적인 방안을 수립하고 있습니까? (1) 1.3.3 분석 목표 달성을 위하여 필요한 정보 항목과 성과 목표치를 정의하고 있습니까? (1) 1.3.4 내부 분석환경 구축을 위한 시스템 및 공급업체 선정 등 구체적인 빅데이터 시스템 확보 방안을 수립하고 있습니까? (1)

참고문헌 : 빅데이터 역량진단모델 개발 및 시범적용 배포 (NIA, 2013.12)

Module-10. 성숙도 측정

□ 국내 빅데이터 역량진단모델

▷ 빅데이터 분석 및 활용역량수준 진단

⊙ 전반적인 빅데이터 분석 및 활용 역량수준은 설문에 의해 산출된 역량점수와 역량수준 정의 규칙을 함께 고려하여 다음과 같이 결정됨. 빅데이터 역량점수는 충분하더라도 역량수준 정의 규칙에 명시된 조건을 만족하지 못하면, 하위 수준으로 하향 조정됨

역량수준	역량점수	역량수준 정의 규칙
1단계 초기(Initial)	0점 ~ 40점	추가 조건 없음
2단계 인식기(Repeatable)	0점 ~ 40점	전략수립역량, 추진역량 40% 이상
3단계 정립기(Defined)	40점 ~ 60점	전략수립역량, 추진역량, 분석역량* 40% 이상
4단계 관리기(Managed)	60점 ~ 80점	전략수립역량, 추진역량, 분석역량*, 활용역량 40% 이상
5단계 혁신기(Innovative)	80점 ~ 100점	모든 대항목 역량 40% 이상

참고문헌 : 빅데이터 역량진단모델 개발 및 시범적용 배포 (NIA, 2013.12)

Module-10. 성숙도 측정

□ 국내 빅데이터 역량진단모델

▷ 빅데이터 역량진단 결과 유형

⊙ 빅데이터 분석 및 활용 역량수준과 함께 보완이 필요한 핵심적인 취약유형을 3개 이내로 제시. 해당 취약유형이 3개를 초과할 때에는 취득 점수가 낮은 것부터 3개까지 선택

역량수준	가능한 취약유형	진단 결과 유형 예시
1단계 초기(Initial)	기본이해 부족형, 활용방안 미비형 데이터 미비형, 지원조직 미비형 분석인프라 미비형*	[역량수준] 1 단계 초기 [세부유형] 기본이해 부족 및 활용방안 미비형
2단계 인식기(Repeatable)	활용방안 미비형, 데이터 미비형 지원조직 미비형 분석인프라 미비형*	[역량수준] 2 단계 인식기 [세부유형] 활용방안 및 데이터 미비형
3단계 정립기(Defined)	데이터 미비형, 지원조직 미비형 분석인프라 미비형* 분석능력 부족형* 활용범위 미흡형	[역량수준] 3 단계 정립기 [세부유형] 분석능력 부족 및 활용범위 미흡형
4단계 관리기(Managed)	분석능력 부족형* 활용범위 미흡형, 혁신능력 부족형	[역량수준] 4 단계 관리기 [세부유형] 혁신능력 부족형
5단계 혁신기(Innovative)	취약유형 없음	[역량수준] 5 단계 혁신기
* '분석인프라 미비형'과 '분석능력 부족형'은 외부 분석서비스 활용형에서는 제외됨		

참고문헌 : 빅데이터 역량진단모델 개발 및 시범적용 배포 (NIA, 2013.12)

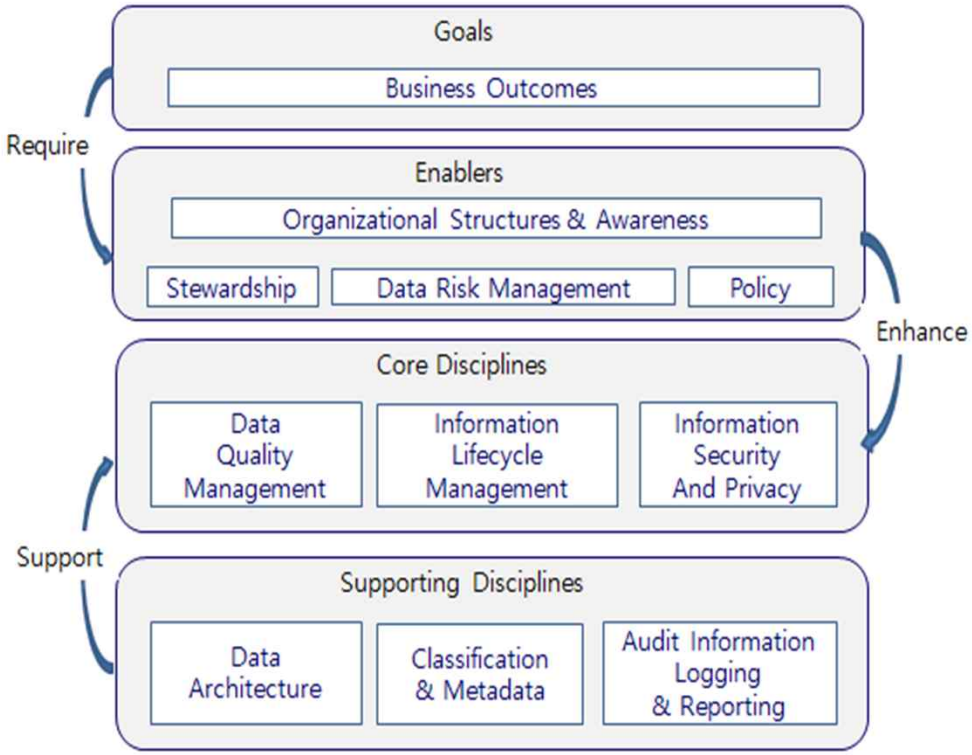
Module-10. 성숙도 측정

□ 해외 빅데이터 역량진단 모델

▷ IBM의 정보 거버넌스 위원회 성숙도 모델

⊙ 정보 거버넌스 성숙도의 4개 그룹과 11개 카테고리를 기술

- 목표(Goals)
  비즈니스 결과
- 조력자(Enablers)
  조직 구조와 인식, 관리권, 데이터 위험관리, 정책
- 핵심 원칙(Core Principles)
  데이터 품질관리, 정보 수명주기 관리, 정보보호와 프라이버시
- 지원 원칙(Supporting Principles)
  데이터 아키텍처, 분류와 메타데이터, 정보 로깅과 보고서의 감사



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-10. 성숙도 측정

□ 해외 빅데이터 역량진단 모델

▷ IBM의 정보 거버넌스 위원회 성숙도 모델

⊙ 빅데이터 성숙도 체크리스트(사례)

그룹	카테고리	정의	체크 항목
I. 목표(Goals)	1. 비즈니스 성과	정보 거버넌스 프로그램의 목적	o 빅데이터 거버넌스 프로그램과 관련된 핵심 비즈니스 부문의 확인 - 소셜 미디어 거버넌스에 대한 마케팅 업무 - RFID 거버넌스를 위한 공급망 관리 - 데이터 유지정책에 대한 법적 문제 - 채용후보자를 미리 선별하기 위한 소셜 미디어 거버넌스와 인적자원 관리 o 빅데이터 거버넌스로부터 얻게 되는 금전적 이익의 정량화 - 데이터 위반으로 벌금 혹은 소송의 위험을 줄일 수 있는가 - 데이터 오용으로 나쁜 이미지가 브랜드에 나쁜 영향을 주는 것을 피할 수 있는가? - 명명법의 불일치 등으로 동일한 데이터를 두 번 구매할 가능성을 줄이는가? - 소셜 미디어 데이터를 마스터 데이터와 통합하여 교차판매와 상향판매의 기회를 확대할 수 있는가? - 센서 데이터를 일관성있고 품질이 높은 자산 데이터와 연계하여 예측 유지관리 프로그램을 수행함으로써 장비의 고장시간을 줄일 수 있는가?

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-10. 성숙도 측정

□ 해외 빅데이터 역량진단 모델

▷ IBM의 정보 거버넌스 위원회 성숙도 모델

⊙ 빅데이터 성숙도 체크리스트(사례)(계속)

그룹	카테고리	정의	체크 항목
II. 조력자(Enablers)	2. 조직구조와 중요성에 대한 인식	비즈니스와 IT간 상호 책임성의 정도	o 일반사항 - 조직에서 다루는 빅데이터의 유형을 인식하고 있는가? 웹과 소셜미디어, M2M 데이터, 빅 트랜잭션 데이터, 생체정보, 사람이 생성한 데이터(이메일 등)가 전형적인 빅데이터의 유형들이다. - 데이터 거버넌스가 필요한 빅데이터의 유형에 대하여 우선순위를 매겼는가? - 기존 정보 거버넌스에 빅데이터로 인한 확장의 필요성을 인식하고 반영하였는가? - 빅데이터 거버넌스가 CDF(Chief Data Officer)와 IGF(Information Governance Officer)의 핵심 직무에 포함되었는가? - 조직에 데이터 과학자가 있으며, 그들이 정보 거버넌스 위원회에서 적절한 활동을 하고 있는가?
	3. 관리권	데이터에 대한 자산 강화, 조직의 제어 등을 위해 제정된 품질관리 규율	o 일반사항 - 빅데이터의 관리권을 어떻게 다룰 것인가? (기존 데이터 관리권으로부터 직무를 확장하여 빅데이터를 포함시키는 방식) - 데이터 관리권이 합법적인 빅데이터 수집, 마케팅, 빅데이터의 적절한 활용과 관련된 부서까지 책임을 져야 하는가?(예를 들어, 소셜 미디어를 마스터 데이터 관리와 통합하는 경우) - 빅데이터의 핵심 속성에 대해 역할과 책임을 정의한 책임할당 정렬(RACI, Responsibility Assignment Matrix)을 생성해 보았는가? -인적자원 관리에서 빅데이터 관리권의 역할이 공식화 되었는가?

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-10. 성숙도 측정

□ 해외 빅데이터 역량진단 모델

▷ IBM의 정보 거버넌스 위원회 성숙도 모델

⊙ 빅데이터 성숙도 체크리스트(사례)(계속)

그룹	카테고리	정의	체크 항목
II. 조력자 (Enablers)	4. 데이터 리스크 관리	리스크를 식별하고, 정성적·정량적으로 측정하며, 가능한 피하고, 그렇지 않으면 수용하거나 완화하며, 밖으로 내보내기 위한 방법론	o 일반사항 - 빅데이터 거버넌스에서 리스크 관리가 핵심 이해관계자인가? - 빅데이터 거버넌스와 리스크 관리 사이에 연결이 있는가?
	5. 정책	조직의 바람직한 행위를 기술한 규정	o 일반사항 - 빅데이터 거버넌스에 관한 정책들이 문서화되어 있는가? - 이러한 정책들이 실행 매뉴얼로 구현되어 있는가? - 거버넌스, 리스크, 규제 프레임워크를 사용하여 실행 매뉴얼의 준수여부가 모니터링되고 있는가? - 빅데이터 거버넌스 정책들이 빅데이터 플랫폼에서 지원되는가?

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-10. 성숙도 측정

□ 해외 빅데이터 역량진단 모델

▷ IBM의 정보 거버넌스 위원회 성숙도 모델

⊙ 빅데이터 성숙도 체크리스트(사례)(계속)

그룹	카테고리	정의	체크 항목
III. 핵심 원칙 (Core Principles)	6. 데이터 품질관리	데이터의 품질과 무결성에 대한 측정, 개선, 인증에 관한 방법들	o 일반사항 - 빅데이터의 경우 데이터의 가치가 높지 않거나 불분명한 경우가 있으며, 이러한 경우에 데이터 품질에 대한 공감대를 형성하고 있는가? - 빅데이터의 저품질로 인한 재정적인 영향에 관해 공감대를 형성하고 있는가? - 조직이 빅데이터 품질문제를 실시간으로 다루는가(data streaming) 혹은 배치 스타일로 다루는가? - 마스터 데이터의 저품질이 빅데이터 분석에 어떤 영향을 미치는가?
	7. 정보 수명주기 관리	정보를 수집하고, 사용하고, 유지하며, 폐기하는 전체 과정	o 일반사항 - 빅데이터를 수용하기 위한 스토리지의 크기는 얼마나 되는가? 빅데이터는 년간 얼마나 증가하는가? - 빅데이터를 수용하기 위한 저장공간의 비용은 얼마나 되는가? 연간 비용은 얼마나 되는가? - 빅데이터의 유지와 관련된 규제 요구사항을 인식하고 있는가? - 빅데이터를 유지하고 싶어하는 비즈니스 요구사항을 이해하고 있는가?
	8. 정보보안과 프라이버시	조직이 리스크를 완화하고, 데이터 자산을 보호하기 위한 정책 및 실행과 제어 등	o 일반사항 - 정보보호 책임자가 빅데이터 거버넌스의 핵심 멤버인가? - 프러이버시 보호 책임자가 빅데이터 거버넌스의 핵심 멤버인가? - Facebook, Twitter 등과 같은 소셜 미디어의 이용약관을 이해하고 있는가? - 소셜 미디어 상의 고객 데이터 사용에 대한 가이드라인을 정했는가? - 고객의 위치정보 사용에 관한 가이드라인을 정했는가?

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-10. 성숙도 측정

□ 해외 빅데이터 역량진단 모델

▷ IBM의 정보 거버넌스 위원회 성숙도 모델

⊙ 빅데이터 성숙도 체크리스트(사례)(계속)

그룹	카테고리	정의	체크 항목
IV. 지원 원칙 (Supporting Principles)	9. 데이터 아키텍처	구조화 또는 비구조화된 데이터와 데이터의 가용성과 분배성을 가능하게 하는 애플리케이션들의 아키텍처	o 일반사항 - 현재 IT인프라와 빅데이터 기술인 Hadoop, NoSQL 등과 같은 새로운 빅데이터 기술들을 어떻게 공존시킬 것인가? - 빅데이터 플랫폼으로 이전되어야 할 응용과 그렇지 않아야 할 응용들을 어떻게 구분하여 이전시킬 것인가? - 기존의 ETL 도구들이 데이터를 빅데이터 플랫폼에 어떻게 적재하고 추출할 것인가? - 빅데이터 플랫폼에서 데이터의 압축과 아카이빙을 어떻게 할 것인가?
	10. 분류와 메타데이터	비즈니스와 IT 용어, 데이터 모델, 데이터 저장소의 의미 등을 정의하는 방법과 도구	비즈니스 사전에 빅데이터 관련 핵심 정의를 포함하고 있는가? 예를 들어 클릭 스트림 데이터에서 'unique visitor'의 의미가 정의되어 있는가? - 빅데이터에 대한 핵심 정의들을 관리하기 위해 관리권을 지정하고 있는가? - 빅데이터 아키텍처 안에서 빅데이터 출처를 다루고 있는가? - 빅데이터 아키텍처 안에서 빅데이터 영향분석을 다루고 있는가? - 빅데이터가 적재되지 않을 경우에 대한 상황을 대처하기 위하여 핵심 운영 메타데이터를 가지고 있는가?
	11. 정보로깅과 보고에 대한 감사	데이터 값과 리스크 및 정보 거버넌스의 효과성을 측정하고 모니터링하는 조직의 프로세스를 의미	o 일반사항 - 민감한 데이터(개인 위치정보, 전화 콜 데이터, 스마트미터 기록, 헬스 클레임 등)에 대하여 암호화되지 않은 상태로 접근할 수 있는 데이터베이스관리자, 계약자, 제3자 등이 있는가? - 빅데이터에 대한 인가된 사용자들의 접근을 어떻게 모니터링하는가?

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-11. 로드맵 수립

□ 빅데이터를 포함하는 정보 거버넌스 로드맵 수립 베스트 프랙티스

▷ 대형 약국의 정보 거버넌스 로드맵 수립 사례

⊙ 로드맵이란 사람, 프로세스, 기술 계획에 대한 단기, 중기, 장기 계획을 의미

⊙ 정보 거버넌스 로드맵이란 향후 18~24개월의 계획을 커버하는데, 정보 거버넌스 프로그램은 그 로드맵안에서 빅데이터 관련 내용을 포함함

⊙ 로드맵 내용

- 1개월~11개월 : 정보 거버넌스 프로그램은 환자, 처방자(의사), 의약품에 대한 마스터 데이터에 집중하고, 프로그램에서는 각 비즈니스 부서에서 데이터 관리인을 교육하여 마스터 데이터 개체를 관리하도록 함
- 12개월~30개월 : 정보 거버넌스 프로그램은 중요한 진단과 처리 코드에 대한 참조 데이터에 집중함
- 31개월~48개월 : 정보 거버넌스 팀은 전체 로드맵에서 빅데이터를 포함하기로 했으며, 두 가지 타입의 빅데이터를 선택했음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-11. 로드맵 수립

□ 빅데이터를 포함하는 정보 거버넌스 로드맵 수립 베스트 프랙티스

▷ 대형 금융기관의 재무부서에서 빅데이터 거버넌스 로드맵 수립 사례

⊙ 대형 금융기관의 재무부서는 대기업 혹은 중견기업에게 통합된 현금관리와 유동성 관리 서비스를 제시

⊙ 전사적 데이터 아키텍트는 빅데이터 프로그램의 초기 후원자였음

⊙ 로드맵 내용

- 1개월~6개월 : 기업의 데이터 아키텍트는 첫 6개월 동안에 빅데이터를 핸들링할 수 있는 기술적 인프라를 정착시킴. 빅데이터는 너무 생소한 것이라 비즈니스 관점에서 성공 가능한 사용예와 재정적인 검증을 하는데 시간을 보내기로 함.
- 7개월~12개월 : 설계자는 상세한 트랜잭션 데이터를 가지고 와서 하루 단위의 상태 분석을 수행하려고 함, 과거에는 높은 인프라 비용때문에 상세한 트랜잭션 수준의 분석을 수행할 수 없었음
- 13개월~24개월 : 설계자는 소셜 미디어 및 다른 비구조화된 데이터를 Hadoop 환경으로 옮기려고 함. 금융기관의 고객들은 대부분 대기업이며, 설계자는 고객의 비구조적 데이터를 탐색하려고 함
- 25개월~36개월 : 설계자는 24개월 후면 금융기관이 빅데이터를 제어하는 상태가 될 것이라고 짐작함. 재무부서는 대형 기업 고객들과 관련된 마스터 데이터에 맞춰진 정보 거버넌스 프로그램을 이미 가지고 있음. 운영그룹은 이 사업의 후원자였음. 빅데이터 계획이 유효한 결과를 내기 시작하는데 수 년이 걸리지만 , 비즈니스의 관심을 끌 수 있었음. 고위 경영진은 빅데이터 관련 파일럿 프로젝트가 주요 데이터 내에서 높은 가시성을 확보했을 때 비로소 투자를 결심할 것이고, 그 결과 정보 거버넌스 팀은 빅데이터에 관련된 사람과 프로세스 및 기술을 로드맵에 포함시켜 변경이 필요함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-12. 빅데이터 거버넌스 조직

□ 빅데이터 거버넌스 조직 구축 절차 베스트 프랙티스

▷ 핵심 프로세스를 밝혀내고, 책임할당(RACI) 차트를 만들어 빅데이터 거버넌스의 이해 당사자 명확히 함

⊙ 건강보험에서 클레임(청구) 처리 과정

- 이해관계자

. 회원 : 건강보험 수급자(건강보험 가입자, 가족)

. 공급자 : 의사, 치료사, 병원

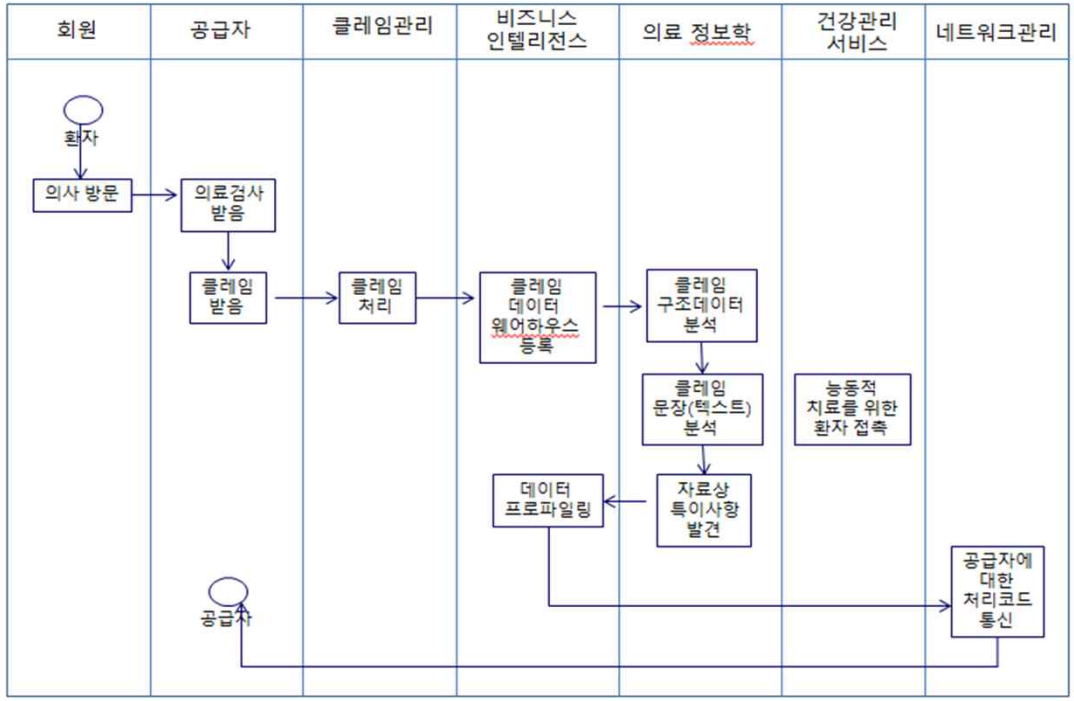
. 클레임관리 : 공급자가 제출한 클레임 처리부서

. 비즈니스 인텔리전스 : 데이터 웨어하우스 분석 부서

. 의료 정보학 : 임상문제 해결 IT사용 담당 부서

. 건강관리 서비스 : 임상 측면을 다루는 부서

. 네트워크관리 : 공급자의 네트워크 담당 부서



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-12. 빅데이터 거버넌스 조직

□ 빅데이터 거버넌스 조직 구축 절차 베스트 프랙티스

▷ 핵심 프로세스를 밝혀내고, 책임할당(RACI) 차트를 만들어 빅데이터 거버넌스의 이해 당사자 명확히 함 (계속)

⊙ 건강보험에서 큰 청구 건을 다루기 위한 빅데이터 거버넌스 정책 들

순서	활동	빅데이터 거버넌스 정책
4	청구건 처리	미국의 의료정보보호 관련 법률인 (Health Insurance Portability and Accountability Act)는 건강 관련 데이터에서 보호가 필요한 개인 프라이버시 정보의 안전장치를 규정한 것이다. 정보보호팀에서는 데이터베이스 모니터링을 통해서 권한을 가진 사람만이 청구 레코드에 접근할 수 있도록 해야 한다. 예를 들어, 보험업자들은 청구 데이터가 데이터베이스 관리자에 의하여 임의로 접근되는 것을 원하지 않을 것이다.
7	청구건 텍스트 분석	전술한 바와 같이 건강보험은 텍스트 분석을 통하여 불일치성을 찾아낸다. 예를 들어, "독감주사" 처치와 건강검진에 사용되는 CPT 코드 ""99214"는 불일치의 예이다. 건강보험에서는 참조 데이터인 ICD-9와 CPT 코드를 사용하여 이 분석을 지원한다.
9	데이터 사이의 이상현상 발견	의료정보학 팀에서는 처치 코드 필드의 많은 항목들이 표준 ICD-9 코드가 아님을 발견하였다.
10	데이터 프로파일링	비즈니스 인텔리전스 팀은 데이터 프로파일링을 통하여 진단 필드가 ICD-9 코드로 기재하도록 되어 있음에도 불구하고, ICD-9와 CPT 코드의 두 가지 형태로 부정확하게 기재되었음을 확인한다. 그 결과 청구보고서는 정확하지 않은 데이터를 보여주게 된다.
11	공급자에 대한 절차코드의 소통	빅데이터 거버넌스 프로그램에서는 네트워크 관리팀에서 공급자에게 청구문서를 작성할 때, ICD-9 코드만을 사용하도록 요청하는 정책을 수립한다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-12. 빅데이터 거버넌스 조직

□ 빅데이터 거버넌스 조직 구축 절차 베스트 프랙티스

▷ 핵심 프로세스를 밝혀내고, 책임할당(RACI) 차트를 만들어 빅데이터 거버넌스의 이해 당사자를 명확히 함(계속)

⊙ 책임할당(RACI) 차트는 조직내 여러 부서들이 빅데이터 거버넌스에 어떻게 동참할 것인지를 보여줌

- 책임(Responsibility) : 어떤 속성을 수행(관리)하는데 있어서의 대표 책임을 말함. 하나의 속성에 대하여 다수의 사람(조직)에게 책임을 부과할 수 있음
- 책무(Accountable) : 데이터 속성에 대하여 최종 책임을 가진 사람으로 승인자(approver) 혹은 승인당국(approving authority)라고 함. 책무를 받은 사람은 어떤 속성을 관리하는 책무를 당사자에게 위임할 수 있음
- 컨설팅(Consulted) : 양방향 소통을 통하여 컨설팅을 받는 사람
- 공지(Informed) : 단방향 소통을 통하여 정보가 제공되는 사람

⊙ 건강보험 청구 데이터에 대한 RACI 차트 일부분 사례

속성들의 카테고리	책임성(R)	책무성(A)	컨설팅(C)	공지(I)
ICD-9과 CPT코드	네트워크 관리자	비즈니스 인텔리전스 담당자	의료정보학, 건강서비스, 청구건 관리자	N/A

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-12. 빅데이터 거버넌스 조직

□ 빅데이터 거버넌스 조직 구축 절차 베스트 프랙티스

▷ 기존 역할과 새로운 역할을 적절히 통합함

⊙ 정보 거버넌스 프로그램이 성숙 단계로 진입한 후에는 CDO(Chief Data Officer)와 IGO(Information Governance Officer), 데이터 스튜어드 등의 역할을 담당하는 주체가 명확해짐

⊙ 기존의 역할 담당자가 빅데이터와 관련된 새로운 역할을 맡을지 혹은 빅데이터와 관련하여 신규로 역할 담당자를 지정할 지를 결정해야 함

▷ 빅데이터에 대한 적절한 관리권 지정

⊙ 빅데이터 관리 주체의 임명 또는 기존 책임자에게 빅데이터 관리권을 추가로 지정해야 함

⊙ 빅데이터 유형별 관리권자의 역할을 지정하는 것이 좋음

▷ 빅데이터의 책임성을 기존 정보 거버넌스의 역할에 추가함

⊙ 기존 정보 거버넌스의 역할에 추가되어야 할 책무

- 최고 데이터 책임 임원(Chief Data Officer, CDO), 정보 거버넌스 책임자, 정보 거버넌스 위원회, 정보 거버넌스 실무반, 고객 데이터 관리자, 자재 데이터 관리자, 자산 데이터 관리자, 최고 데이터 관리자(전사 수준의 빅데이터 등의 정보 거버넌스 프로그램 감독), 빅데이터 실무 관리인

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

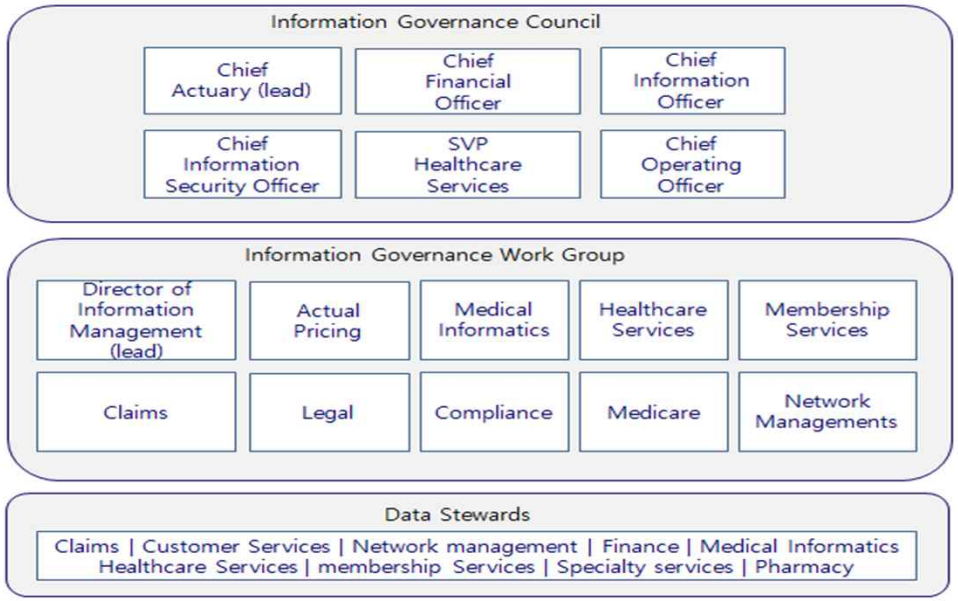
Module-12. 빅데이터 거버넌스 조직

□ 빅데이터 거버넌스 조직 구축 절차 베스트 프랙티스

▷ 빅데이터의 책임성을 포함하는 정보 거버넌스 조직을 생성함

⊙ 기존 데이터와 빅데이터 모두를 잘 관리할 수 있는 체제를 갖추도록 함

⊙ 건강보험에서의 정보 거버넌스 조직



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-13. 비즈니스 프로세스 통합

□ 빅데이터 비즈니스 프로세스 통합 베스트 프랙티스

▷ 빅데이터 거버넌스에 영향을 받는 핵심 프로세스의 구별

⊙ 빅데이터 거버넌스와 BPM(Business Process Management)과의 공생관계 인식

⊙ 핵심프로세스와 빅데이터 거버넌스 통합 사례

- '보험청구처리' 시 일관성없는 데이터 입력은 공통된 문제인데, 자유형식의 텍스트 항목을 사용하여 상처원인을 기록함. 보험회사는 자유형식때문에 청구를 분석하는데 어려움을 가짐. 상해의 원인과 같은 중요한 정보를 기록하기 위한 표준코드를 개발하여 프로세스를 개선해야 함. 청구양식의 표준화를 통해 데이터의 일관성을 향상시킬 수 있음
- 소비재 제조업체가 동일한 제품에 대하여 서로 다른 이름을 사용하는 소매업자들의 POS 트랜잭션 로그를 사용할 때 수요예측, 제품 개발, 마케팅 세분화에 영향을 미침
- 예방적 유지보수 프로그램은 많은 양의 센서 데이터를 사용하는 예측모델에 의존함. 센서 이벤트 1234(기계고장)의 90%는 센서 이벤트 4567의 발생이 선행된다는 사실을 발견했을 때 유지관리 부서는 예방적 유지보수를 시작할 수 있음
- 기업의 인사부서는 입사 지원자를 사전에 선별하기 위해 구직자의 SNS 이용에 관한 지침을 수립해야 함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-13. 비즈니스 프로세스 통합

□ 빅데이터 비즈니스 프로세스 통합 베스트 프랙티스

▷ 프로세스 맵의 설계와 주요 활동

⊙ 가상의 소매업자가 MDM을 사용하여 어떻게 Facebook 앱을 강력하게 만드는지에 대한 사례

- 소매업체(A편의점)의 소셜미디어의 활용 시 핵심 활동들

1) 홍길동은 김삿갓으로부터 Facebook 앱 요청을 받음
2) 홍길동은 A편의점으로부터 할인, 친구 인식경보, 프리미엄 제품과 같은 혜택을 보고 Facebook 앱 요청을 수락
3) 홍길동은 동의를 하고 A편의점가 그녀의 기본 정보와 그녀의 친구에 관한 기본 정보에 접근하는 것을 허락
4) A편의점은 MDM에 있는 그녀의 내부 기록과 홍길동의 Facebook 프로필을 대조함
5) A편의점은 매력적인 상품 제안을 하기 위해 홍길동의 MDM 프로필과 지난 구매정보를 대조함
6) A편의점은 또한 홍보된 상품을 홍길동이 직접 확인할 수 있는 가장 최적의 지점을 제안하고 홍길동의 쇼핑 담당자를 할당함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-13. 비즈니스 프로세스 통합

□ 빅데이터 비즈니스 프로세스 통합 베스트 프랙티스

▷ 빅데이터 거버넌스 정책을 프로세스의 키 스텝에 맵핑

⊙ 빅데이터 거버넌스 프로그램은 정책을 중요 업무프로세스 내의 핵심활동들과 맵핑하는데 기여

⊙ 빅데이터 거버넌스의 핵심 변화와 구체적인 활동 맵핑

순서	활동	빅데이터 거버넌스 정책
1	홍길동은 김삿갓으로부터 Facebook 앱 요청을 받음	A편의점은 자사 제품에 대하여 소셜 미디어 상에서 최고의 영향력을 행사하는 사람을 확인한다.
2	홍길동은 A편의점으로부터 할인, 친구 인식경보, 프리미엄 제품과 같은 혜택을 보고 Facebook 앱 요청을 수락	A편의점은 인센티브를 미세 조정하기 위해 인구학과 속성들에 기반을 둔 A/B 테스팅을 실시한다. A편의점은 A/B 테스팅을 위해 인구통계, 행동학, 제품선호도 등에 관한 고품질 데이터가 필요하다.
3	홍길동은 동의를 하고 A편의점가 그녀의 기본 정보와 그녀의 친구에 관한 기본 정보에 접근하는 것을 허락	마케팅은 Facebook 플랫폼 정책을 고수한다. 예를 들어, 고객의 친구에 대한 데이터를 목적 외에는 사용하지 않는다.
4	A편의점은 MDM에 있는 그녀의 내부 기록과 홍길동의 Facebook 프로필을 대조함	마케팅 데이터 관리자는 고객의 MDM 기록에 고객의 Facebook 프로필을 연결하는 속성을 확인한다. A편의점은 홍길동의 MDM 기록에 홍길동의 Facebook 사용자 ID를 추가한다.
5	A편의점은 매력적인 상품 제안을 하기 위해 홍길동의 MDM 프로필과 지난 구매정보를 대조함	판촉데이터 관리자는 제품비교를 가능하게 하는 제품 계층을 설정한다. 간단한 예로, A편의점은 홍길동이 이미 '월풀 GX5F-HDXVY'를 구입한 적이 있기 때문에 '냉장고' 계층의 제품을 이미 가지고 있다는 것을 안다.
6	A편의점은 또한 홍보된 상품을 홍길동이 직접 확인할 수 있는 가장 최적의 지점을 제안하고 홍길동의 쇼핑 담당자를 할당함	부동산 데이터 관리자는 고객들이 선호하는 물리적 매장을 확인하기 위하여 위치 마스터 데이터를 구축했다. 점포 영업활동 관리자는 관련 마스터 데이터를 생성하고 고객을 지원하기 위한 프로세스를 정의한다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-14. 웹과 소셜미디어

□ 빅데이터 거버넌스 구축사례_웹과 소셜미디어 데이터

▷ 기업에 있어서의 소셜 미디어 특징

⊙ 소셜 미디어는 조직이 그들의 고객과 관계를 맺는 방식을 브랜드->상품->서비스로 발전시키는 방식을 바꾸고 있음

⊙ 소셜 미디어는 회사가 그들의 고객에 대한 영향력과 친밀감을 주면서 회사에게 고객 데이터를 쉽게 모으고 분류할 수 있는 능력을 보유하게 함

⊙ 소셜 미디어는 마케팅 담당자가 더 빠르게, 더 광범위한 소비자에게 동적으로 더욱 정확하게 접근할 수 있도록 지원함. Twitter와 Facebook에서 즉시 캠페인을 진행할 수 있음

⊙ 소셜 미디어는 매우 저렴한 비용으로 정확히 타깃된 고객들에게 광고를 할 수 있음

▷ 진화하는 규제와 관습을 고려한 소셜 미디어 고객 데이터 활용 정책 수립

⊙ 보험사

- 보험 청구관리

  예) 자동차 보험금 지급 심사에 Facebook 정보를 이용하여 거짓을 밝혀낸 보험조사원

- 보험 판매

  예) 보험사들이 보험가입자의 신용정보를 범죄 가능성 예측에 활용함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

□ 빅데이터 거버넌스 구축사례_웹과 소셜미디어 데이터

▷ 진화하는 규제와 관습을 고려한 소셜 미디어 고객 데이터 활용 정책 수립

⊙ 의료보험(health Plan)

- 소셜 미디어가 의료보험의 비즈니스 프로세스를 변경시켰음

예) Twitter, Facebook, YouTube 등이 잠재적 고객과의 채널로 부상함에 따라 의료보험사는 소셜 미디어에 언급된 자사에 관한 글을 모니터링하고 적절히 응답 글도 작성하며 즉각적인 대응을 수행함

⊙ 생명과학 관련 회사

- 제약 및 의료기기 회사는 소셜 미디어를 활용하여 그들의 고객을 더 많이 이해하려고 하나, 고도의 규제 환경에 처해 있으므로 관련 규제를 준수해야 함

예) 미국 FDA에서 제시한 생명과학 관련 소셜 미디어의 활용 가이드라인

1) 회사가 승인되지 않은 요청에 대하여 답을 할 때 회사는 그 요청이 분명하게 자신의 제품인 경우에 한하여 답변해야 한다.
2) 회사 제품에 관한 질문에 대하여 소셜 미디어 상에서 접할 때 회사의 연락처 정도를 제시하는 것으로 한정하며, 승인되지 않은 정보를 포함해서는 안 된다.
3) 회사의 공식 답변에는 담당부서와 담당자에 대한 연락처를 제공하여 개인이 비공개적인 1:1 방식으로 원하는 정보를 얻을 수 있게 해야 한다.
4) 개인이 사적으로 회사에 접촉하여 회사 제품에 대한 추가 정보를 원하면, 회사는 구체적인 답변과 함께 그 기록을 유지할 수 있다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

□ 빅데이터 거버넌스 구축사례_웹과 소셜미디어 데이터

▷ 진화하는 규제와 관습을 고려한 소셜 미디어 고객 데이터 활용 정책 수립

⊙ 헬스케어 제공자

- 의사가 온라인상에서 활동할 때 지켜야 할 핵심 가이드라인 제시
  예) HIPPA 규제, 환자-의사 기밀 유지 의무, 정보의 분리 보관

⊙ 은행

- 신용 위험관리
  은행들은 대출에 따르는 신용 위험을 평가하기 위해 소셜 미디어 정보를 활용하고 있음. 그러나 개인정보를 수집하는 행위는 FCRA(공정신용보고법)에 금지되고 있지만 완화되는 추세임
- 채권 회수 업무
  채권 회수부서는 Facebook, Twitter, Monster.com 등의 사이트를 통해 연체자의 최신 연락처를 추적하려고 함.
  채권추심자는 Facebook에 친구로 거짓 등록하거나 개인 채무에 대해 Twitting하는 것을 금지하고 있음

⊙ 법조계

- 신용 위험관리이혼 변호사는 소셜 미디어의 정보를 마이닝하여 상대방의 부정, 숨겨진 자산, 진실성, 숨겨진 정보 등을 찾아내는 데 이용. 블로그, 소셜 미디어, 사진공유 사이트 등에서 모여진 정보는 법정에서 핵심 증거가 될 수 있음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-14. 웹과 소셜미디어

□ 빅데이터 거버넌스 구축사례_웹과 소셜미디어 데이터

▷ 종업원과 구직자에 관한 소셜 미디어 데이터의 적절한 사용정책 수립

⊙ 고용주들이 소셜 네트워크를 통해 구직자를 조사

- Facebook 계정과 비밀번호를 요구하는 고용주 사례
- 법률, 규제, 판례법에서 채용 시 소셜 미디어의 활용을 통제하는 것은 나라마다, 주마다, 지역마다 다를 수 있고, 급속히 진화하고 있음. 고용주는 법률 자문팀과 상의하여 각자의 상황에 적합한 정책을 만들어야 함

▷ 신뢰구간을 활용한 소셜 미디어 품질의 측정

⊙ 소셜 미디어 활용 시 데이터 품질 문제 이슈

- Twitter를 통해 우리 회사의 명성을 분석해보고 싶음. 그러나 Twitter 샘플의 모집단 반영유무, Twitter를 하는 사람들의 편향성, 연령층, 재산 많고 적음 등
- 소셜 미디어 데이터를 조직의 내부 자료와 합쳐서 비교함. 그러나 소셜 미디어 데이터는 정제되지 않은 데이터가 다수임
- 소셜 미디어 데이터는 단편적이며 완전하지 않음
  Twitter가 드러내는 사용자의 이름은 5~60%만이 정확하고, 성별에 대해서도 명확하지 않음

⊙ 빅데이터 거버넌스의 관점에서는 신뢰구간으로 데이터의 질을 평가하는 것이 필요함

- 자연어 처리 과정을 통해 사용자의 핵심 속성을 식별함
- Facebook 사는 가짜 데이터를 제거하는 자동화된 도구 사용

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

□ 빅데이터 거버넌스 구축사례_웹과 소셜미디어 데이터

▷ 쿠키 및 다른 형태의 웹 추적 데이터의 활용에 관한 정책 수립

⊙ 쿠키의 기능

- 세션관리 기능
- 개인화 : 사이트를 방문한 유저의 정보를 기억하고 관련 콘텐츠를 보여줌
- 웹 분석 : 웹의 효율성을 측정하거나 마케팅 혹은 인터넷 사용자의 행동을 분석하는데 이용됨

⊙ 쿠키의 유형

- 세션 쿠키 : 사용자가 웹사이트에 연결되어 있는 동안만 지속됨
- 지속 혹은 추적 쿠키 : 웹사이트에 방문한 사용자를 익명으로 식별하는데 사용됨. 어느 웹페이지를 경유하여 현재 웹페이지로 들어왔는지에 대한 추가적인 정보를 저장하는데도 사용
- First-party 쿠키 : 웹브라우저의 주소바에 나타난 도메인으로 설정
- Third-party 쿠키 : 웹브라우저의 주소바에 나타난 도메인과는 다른 도메인이 설정됨. 예를 들어, 사용자가 A사이트를 방문하였고, 이 사이트에는 B로부터의 배너광고를 한다고 했을 때 사용자가 배너광고를 클릭하면 third-party 쿠키를 자동으로 받아 사용자 컴퓨터에 저장함
- Flash 쿠키 : 로컬 공유 객체로 알려진 이 쿠키는 Adobe Flash를 이용하는 사이트가 유저의 컴퓨터에 저장시키는 작은 데이터. 유저들에게 잘 알려져 있지 않으며, 브라우저에 있는 쿠키 프라이버시 제어를 통해 컨트롤할 수 없음. 사용자가 브라우저를 통해 쿠키를 삭제했어도 대부분의 경우 남아있음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-14. 웹과 소셜미디어

□ 빅데이터 거버넌스 구축사례_웹과 소셜미디어 데이터

▷ 쿠키 및 다른 형태의 웹 추적 데이터의 활용에 관한 정책 수립

⊙ 쿠키의 유형(계속)

- web beacon : 웹페이지나 이메일에 포함되어 있는 객체로써 사용자들에게는 보이지 않지만 그 사용자가 메일 혹은 페이지를 봤는지 확인할 수 있음
- Browser fingerprinting : 웹사이트에 접속한 고객의 브라우저에서 확인이 가능한 시스템 폰트, 소프트웨어, 혹은 설치된 플러그인 등의 정보를 조합하여 유일한 컴퓨터 혹은 개인기를 식별해내는 기술

⊙ 웹추적에 대한 논쟁의 초점

- 온라인 행위분석 기반의 광고 : 고객의 온라인 관심사에 관한 정보를 수집하여 관심있는 상품을 광고함
- 관련된 광고시스템은 광고 네트워크를 회전하면서 third-party 쿠키를 사용하여 서로 다른 웹사이트 상의 개인고객(적어도 디바이스)을 트래킹할 수 있음. 이 데이터를 유일 식별자를 기준으로 구성하면 인터넷상에서 개인에 관한 광범위한 정보를 볼 수 있음
- 개인의 행동 프로파일은 고아고자에게 개인의 관심사에 기반을 둔 타킷 광고를 할 수 있도록 지원함
- 많은 고객과 프라이버시 옹호론자들은 개인 행동 트래킹에 기반을 둔 광고는 프라이버시를 침해하는 것이라고 보고 있음. 이러한 걱정에도 불구하고 쿠키와 웹 트래킹 기술에 관한 소비자의 지식은 매우 제한되어 있음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

□ 빅데이터 거버넌스 구축사례_웹과 소셜미디어 데이터

▷ 쿠키 및 다른 형태의 웹 추적 데이터의 활용에 관한 정책 수립

⊙ 사례1. 유럽연합과 영국의 쿠키 이용에 대한 규제환경

- 2003년 영국에서 제정된 '프라이버시와 전자통신 규제(Privacy and Electronic Communications Regulations)는 개인용 컴퓨터나 모바일 디바이스에서 쿠키 및 유사 기술에 대한 저장과 사용에 관한 것을 포함. 이 규제는 유럽명령 2002/58/EC로 제정됨
- 2009년 개정된 명령 2009/136/EC는 가입자의 단말기에 저장된 정보(쿠키 혹은 유사 기술들)를 액세스하거나 저장하는데 동의를 얻도록 하고 있음. 유럽정부는 2011년 5월까지 법으로 제정하려고 함
- 명령 2009/136/EC의 5조 3항
  '회원국들은 가입자의 단말기에서 정보를 저장하거나 기 저장된 정보를 액세스할 때 명호가하고 이해하기 쉬운 형태로 제시된 설명과 명령 95/46/EC에 따라 가입자 동의를 얻어야 한다.'

⊙ 사례2. 애플의 Safari 웹브라우저에서 구글이 프라이버시 세팅을 우회하다.

- 'The Wall Street Journal'은 구글 및 광고회사들이 애플사의 아이폰 기계에서 Safari 웹브라우저를 사용하는 수백만 사용자의 프라이버시 설정 기능을 우회하여 사용자의 웹브라우징 습관을 추적하고 있다고 보도함
- 이런 정보는 분명히 블로킹되어 있었어야 했고, Safari 브라우저는 Third-party 쿠키들을 막아 놓도록 설계되었으나, 구글은 이를 브라우저가 First-party 쿠키로 인식하도록 했음. 구글은 이후에는 해당 기능을 비활성화하였음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-14. 웹과 소셜미디어

□ 빅데이터 거버넌스 구축사례_웹과 소셜미디어 데이터

▷ 쿠키 및 다른 형태의 웹 추적 데이터의 활용에 관한 정책 수립

⊙ 사례3. 온라인 행동 기반의 광고에 대한 스스로 제정한 규제

- 온라인 행동분석 기반의 광고는 규제 대상이지만, 웹페이지의 콘텐츠와 검색질의, 웹사이트에서 동시에 발생하는 사용자 행동 등을 기반으로 이루어지는 콘텍스트 기반의 광고(예를 들면 gmail에서 이메일을 읽고 있을 때 구글의 자가출판 책에 대한 광고를 수행함)는 규제 대상에 포함되지 않음
- 규제 원칙 사례 (미국 광고사 연합, 광고인 연합의 온라인 행동분석 기반의 광고에 대한 규제 원칙)

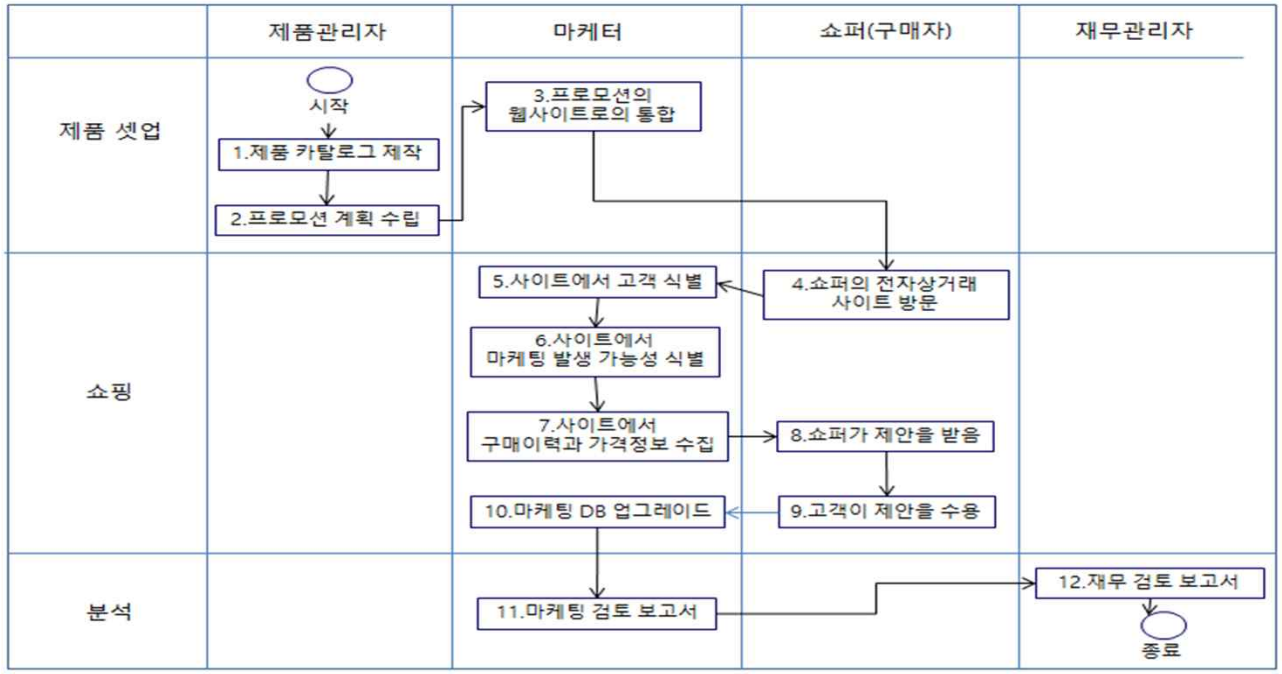
원칙	내용
교육 필요성	소비자와 회사에 온라인 행동분석 기반의 광고에 대해 교육한다.
투명성	온라인 행동기반 광고 관련 데이터 수집에 사용되는 다양한 메커니즘에 대해 소비자에게 확실하게 알린다.
소비자 제어권	소비자들은 데이터가 수집되는 지 여부와 특정 목적을 위해 제3자에게 전송할 수 있는지에 관한 사항을 결정할 수 있다.
데이터 보안	수집된 데이터에 대하여 보안과 유지기간이 합당해야 한다.
중요사항의 변경	온라인 행동기반 광고에서 데이터 수집과 관련한 정책을 수정하기 전에 소비자의 동의를 얻어야 한다.
민감한 데이터 관리	민감한 데이터는 다른 데이터와 다른 적용을 받는다. 특히 관련 법에 따라 보호되는 데이터의 경우 더욱 그러하다. 이러한 데이터에는 온라인 이동 프라이버시 보호법과 관련된 아동 정보가 있다. 유사하게 특정 개인의 금융 계좌번호, 사회보장번호, 약 처방전, 의료기록 등을 온라인 광고에 활용할 때 주의가 요망된다.
책임성	온라인 행동기반 광고와 관련된 조직은 관련 정책을 수립하고 추진하여, 관련 규제를 준수하는데 광범위한 책임을 진다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

□ 빅데이터 거버넌스 구축사례_웹과 소셜미디어 데이터

▷ 프라이버시와 규제를 위반하지 않는 방식으로 온라인과 오프라인의 데이터를 연동시키는 정책 수립

⊙ 사례4. 소프트웨어 회사의 웹 상거래 프로세스



참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-14. 웹과 소셜미디어

□ 빅데이터 거버넌스 구축사례_웹과 소셜미디어 데이터

▷ 프라이버시와 규제를 위반하지 않는 방식으로 온라인과 오프라인의 데이터를 연동시키는 정책 수립

⊙ 사례4. 소프트웨어 회사의 웹 상거래 프로세스(계속)

활동	빅데이터 거버넌스 정책
1. 적절한 제품 카탈로그 관리	제품관리부서는 적절한 계층과 마스터 데이터 속성을 갖춘 제품 카탈로그를 제작함
5. 사이트가 고객을 식별함	사이트는 고객을 식별함 처음에는 쿠키에 대한 분석으로 나중에는 로그인을 통해서
6. 사이트가 마케팅 활동이 가능한 고객인지 확인함	B2B 고객들은 이미 온라인 프로파일이 구축되어 있을 것이다. B2B 고객은 이 프로파일에서 그들이 웹사이트를 방문했을 때 프로모션 오퍼를 제시 받을지에 관한 opt-in을 결정한다. 웹사이트는 로그인과 같은 정보를 이용하여 고객 데이터베이스를 조회하여 그 구매자가 opt-in 동의를 했는지를 확인한다. 고객 데이터 관리자는 이 데이터의 무결성을 책임지고 있다. 웹사이트는 마케팅 제안에 대한 opt-out 신청을 존중한다. 이 과정은 B2C 구매자에게는 적용하지 않는다.
8. 구매자가 제안을 받음	웹사이트는 예측모델을 사용하여 구매자에게 상품과 가격을 제시한다. 이 예측모델은 다음과 같은 정보를 입력으로 사용한다. - 고객의 인구통계학적인 정보(로그인 시 정보를 제공한 경우) - 회사에 관한 정보(제안에 대한 수용율 등) - 방문자의 인터넷 도메인 주소 - 구매자가 기 보유한 상품 - 구매자의 이전 브라우징 히스토리 - 가격 탄력성 - 제품 친밀감(기 보유 제품을 기반으로 산정함) - 계약 금액(B2B 고객인 경우) 이 예측모델은 고객, 제품, 가격 등에 관한 고품질의 마스터 데이터를 기반으로 구축된다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

□ 빅데이터 거버넌스 구축사례_웹과 소셜미디어 데이터

▷ 프라이버시와 규제를 위반하지 않는 방식으로 온라인과 오프라인의 데이터를 연동시키는 정책 수립

⊙ 사례4. 소프트웨어 회사의 웹 상거래 프로세스(계속)

활동	빅데이터 거버넌스 정책
10. 마케팅 데이터베이스가 업데이트됨	구매자의 상호작용 히스토리는 최신의 정보로 갱신되어 향후 웹사이트를 방문 시 적절한 지원뿐 아니라 받아들여지지 않은 제안들을 이해하는데 사용된다.
11. 마케팅 부서에서 보고서 검토	마케팅 보고서는 고객의 트래픽, 구매자의 동선, 전환율, A/B 테스트, 장바구니 분석 등을 담고 있다. 미케팅 부서는 'unique visitor', 'conversion event', 'purchase event' 등과 같은 용어에 대해 비즈니스 관점에서 일관성있는 정의를 필요로 한다.
12. 재무관점에서 보고서 검토	이 보고서는 판매와 수익에 관해 보고하며, 'net sales' 등과 같은 용어에 대하여 비즈니스 측면에서 일관성있는 정의를 요구한다. 또한, 보고서의 데이터에 대하여 데이터 웨어하우스 및 데이터 소스에 이르는 출처를 필요로 한다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-14. 웹과 소셜미디어

□ 빅데이터 거버넌스 구축사례_웹과 소셜미디어 데이터

▷ 웹 측정 방법의 일관성 유지

⊙ 고객들은 서로다른 웹사이트들과 이메일 및 마케팅 캠페인과 같은 다양한 형태로 조직의 특성과 관련한 상호작용을 수행함

⊙ 조직들은 고객의 부라우징과 탐색의 특성이나 마케팅 캠페인에 대한 응답 등과 같은 방대한 양의 온라인 데이터 보유하게 됨

⊙ 데이터 유형 예

항목	사례/설명
KPI로서 최고위층의 비즈니스 목표에 대한 사이트의 성과 측정 데이터	판매량, 광고예산, 고객 셀프-서비스 성공 등
트랜잭션 지표	판매량, 주문항목들, 평균 주문량
변환 지표	변환율, 쇼핑카트 세션 들, 쇼핑카트 변환율, 쇼핑카트 폐기율 등
세션 트래픽	평균 세션 길이, 사이트에 들어와서 한페이지만 보고 나가는 비율, 세션당 보는 페이지 수, 세션당 보는 제품의 개수 등
수명가치 지표	반복 방문자, RFM-Recency, frequency, Monetary-데이터, 2x 구매자, 3x-5x 구매자
모바일 디바이스 지표	판매량중에서 모바일 비중, 사이트 트래픽 중에서 모바일 비중, 모바일 바운스 비율(한 번 보고 가버리는 비율), 안드로이드 트래픽, 아이폰 트래픽, iPad 트래픽 등
소셜 미디어 소개 지표	판매에서 소셜 미디어 비율, 트래픽 중에서 소셜 미디어 비중, Facebook 소개 트래픽, Twitter 소개 트래픽, Pinterest 소개 트래픽 등

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

□ 빅데이터 거버넌스 구축사례_웹과 소셜미디어 데이터

▷ 웹 측정 방법의 일관성 유지

⊙ 빅데이터 거버넌스 프로그램 측면의 이슈 : 웹사이트들간의 절대값의 불일치성

- 온라인 세상에서 비일관적인 용어
  예) 웹사이트들간의 세션 측정시간의 차이
- 오프라인 세상에서 비일관적인 용어
  예) 웹분석 데이터와 오프라인 데이터 통합 시 'high value customer', 'prospect', 'tenuard customer'의 정의의 혼란
- 오류로 인해 페이지 태그가 로드되지 않음
  예) 페이지 태그는 웹페이지내의 코드 조작으로 다른 코드를 실행하는 라이브러리를 가리키고 있으며, 어떤 사용자가 어떤 페이지를 방문한 사실도 있음
- 추적기술
  예) 한 웹사이트는 로그파일을 이용하여 웹분석용 데이터를 수집하고, 다른 사이트는 페이지 태그를 사용함. 어떤 사이트는 first-party 쿠키를 사용하고 다른 사이트는 third-party 쿠키를 사용함으로써 측정된 수치가 다를 수 있음
- 분석기술의 차이
  예) 어떤 두 웹분석 도구도 동일한 데이터 집합에 대해 동일한 결과를 산출한다고 보장할 수 없음. 이러한 불일치는 다양한 소스로부터 기인함(예를 들어, 각 도구가 클릭들을 어떻게 세션으로 그룹핑하는지 등). 한 도구는 고객이 검색을 위해 웹사이트를 떠났다가 돌아오면 다른 세션이라고 간주하기도 함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-15. M2M 데이터

□ 빅데이터 거버넌스 구축사례_M2M 데이터

▷ M2M 데이터 정의

⊙ M2M(Machine-to-Machine)은 다른 디바이스와 통신하는 무선 혹은 유선시스템

⊙ M2M은 센서, 미터기 등의 디바이스를 사용하여 스피드, 온도, 압력, 흐름, 염도 등을 측정하는데 이러한 장치들은 무선, 유선, 하이브리드 네트워크를 경유하면서 수집한 데이터를 유용한 정보로 변환하는 애플리케이션으로 전송

▷ 현재 활용 가능한 위치정보 데이터의 유형 조사

⊙ 위치정보(geolocation)란 개인 혹은 개체의 지리적인 위치에 대한 식별정보. 스마트폰 등 RFID 장착 디바이스들로부터 발생되는 신호에 기반을 두고 있음

⊙ 이 장치들은 전화번호, 현재 위치, MAC Address 등의 정보를 디바이스가 사용되지 않는 상태에서도 지속적으로 서버에 정보를 전달함

⊙ 위치정보 유형(Type)

- 기지국 데이터
- GPS(Global Positioning System) 기술
- WIFI 기술
- RFID(Radio Frequency IDentification) 기술

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-15. M2M 데이터

□ 빅데이터 거버넌스 구축사례_M2M 데이터

▷ 고객의 위치정보에 대한 활용 정책을 수립

⊙ 개인의 모바일 디바이스는 이메일, 사적인 사진, 브라우징 히스토리, 주소록 등 개인적인 정보를 담고 있음

⊙ 위치정보 기반 서비스 제공자는 이러한 모든 정보를 바탕으로 개인에 관한 프로파일을 만드는 것이 가능함. 고객의 야간 행동 패턴, 잠자는 곳에 대한 정보, 아침에 이동 패턴, 직장의 위치 등의 추정이 가능

⊙ 소셜 네트워크 사이트에서 개인의 친구 관계나 행동 패턴에 대한 정보를 수집할 수도 있음. 행동 패턴 정보에는 병원이나 종교기관 방문은 물론, 시위 현장에의 참여도 알 수 있음. 이러한 정보는 개인에게 심각한 영향을 주는 정보로 악용될 수 있음

⊙ 미국의 위치 프라이버시 법(s, 1223, 2011) 가이드라인 사례

- 고객의 위치정보를 수집하기 전에 고객의 분명한 동의를 얻을 것
- 고객의 위치정보를 제3자에게 공유하기 전에 고객의 분명한 동의를 얻을 것

⊙ 유럽연합 데이터 보호 분과 제 29조 – 스마트 모바일 디바이스 상의 위치정보 서비스에 관한 가이드라인 사례

- 스마트 모바일 장치의 위치정보가 개인의 사적인 데이터이므로 통신 사업자는 데이터에 대한 사전동의를 얻어야 한다
- 동의는 구체적인 용어와 조건으로 이루어져야 한다
- 동의는 목적에 맞게 구체적으로 받아야 한다. 데이터에 대한 처리 목적이 구체적으로 변경된다면 조직은 동의를 갱신해야 한다
- 기본적으로 위치정보 서비스는 사용자가 언제든지 끌(off) 수 있어야 한다

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-15. M2M 데이터

□ 빅데이터 거버넌스 구축사례_M2M 데이터

▷ 직원의 위치정보에 대한 활용정책을 수립

⊙ 빅데이터 거버넌스 프로그램은 인사관리 팀과 공조하여 직원에 대한 위치정보의 사용과 관련된 정책을 확고하게 수립해야 함

⊙ 사례1. 유럽연합의 직원 관련 위치정보 사용통제 가이드라인(유럽연합 데이터 보호 분과 제 29조)

- 유럽연합의 법적 프레임워크는 직원이 채용 시 위치 모니터링에 대한 동의를 했더라도 이를 인정하지 않는다
- 그러나 고용주는 합법적인 목적으로 명백히 필요한 경우에 한하여 덜 거슬리는 방식으로 그 목적을 달성할 수 없는 경우에만 제한적으로 위치정보 기술을 채택할 수 있다
- 직원의 동의를 구하기에 앞서 고용주는 먼저 합법적인 목적을 위해 직원의 위치를 감시해야 할 필요성이 존재하는 지와 그 필요성이 직원의 기초적 권리와 자유보다 더 중요한 것인지를 조사해야 한다
- 고용주는 직원들에게 최대한 피해를 주지 않는 방식을 채택해야 하고, 계속적인 모니터링을 하지 않아야 한다.
- 직원은 일하는 시간 외에는 모니터링 장치를 끌(off) 수 있어야 한다
- 자동차 추적장치는 운전하는 직원을 추적하는 장치가 아니며, 단지 자동차의 위치를 모니터링할 뿐이다. 고용주는 이 장치를 운전사나 직원의 위치 내 행동패턴을 추적하거나 모니터링하는데 사용하면 안 된다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-15. M2M 데이터

□ 빅데이터 거버넌스 구축사례_M2M 데이터

▷ RFID 데이터에 대한 프라이버시 인증

⊙ 특정인에게 부착된 RFID 데이터는 사람을 식별하는 정보(PII, Personally Identifiable Information)로 취급되어야 함

⊙ 조직들은 이러한 RFID 데이터를 다른 PII와 동일한 방법으로 다루어야 함

⊙ 2009년 5월 유럽연합 집행기관은 RFID 애플리케이션들을 프라이버시 영향 평가(PIA, Privacy and data protection Impact Assessment) 대상으로 지정

⊙ 2011년 2월 유럽연합 집행기관의 데이터 보호 작업분과 제29조에는 PIA 프레임워크 추가

▷ 다른 M2M 데이터에 대한 프라이버시 정책 수립

⊙ PIA 프로세스의 일부로 RFID 운영자는 개인 데이터를 위협하는 위험성을 인식할 필요가 있음

⊙ 소매업 사례

- RFID 태그는 개인을 추적하거나 프로필을 작성하는데 활용될 수 있음
- 소매점에서 신발을 산 고객을 가정할 때, 그 신발에는 RFID 태그가 붙어 있어서 해당 고객이 다른 양말 상점을 방문할 때 RFID 태그 정보를 읽어서 양말을 20% 싼 가격으로 제공할 수 있다는 메시지를 보낼 수 있음
- 유럽연합 가이드라인에 따르면, 소매업자는 상품을 판매하는 시점에서 고객이 태그의 활성화를 원하지 않는 한 RFID 태그를 비활성화하거나 제거하도록 하고 있음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-15. M2M 데이터

□ 빅데이터 거버넌스 구축사례_M2M 데이터

▷ 다른 M2M 데이터에 대한 프라이버시 정책 수립

⊙ 건강관리사례

- 병원들은 자산과 환자 모니터링을 위해 지속적으로 RFID 사용을 확대하고 있음
- RFID를 사용하여 자산을 관리하면 재고수준을 적절하게 유지할 수 있고, 사기나 낭비 및 남용을 방지할 수 있음
- 많은 병원들은 최근에 RFID 팔찌를 만들어 환자를 모니터링하고 있음. 이 경우 다음과 같은 프라이버시 이슈가 있음
  . 환자의 추적 프로세스가 RFID 데이터에만 의존하는가? RFID 이전에 존재하던 프로세스에 덧붙여 이루어지고 있는가?
  . 환자가 말을 할 수 없거나 혹은 다른 방식으로 통신이 불가능하다면 환자 추적 정책이 어떻게 바뀔 것인가?
  . 환자가 의료용 태그에서 민감한 개인정보를 읽는 것에 대하여 참여하지 않겠다고 선언할 수 있는가?
  . 조직내에서 누가 환자의 개인정보를 액세스할 수 있는가? 조직 밖에 있는 가족들이나 변호인 혹은 다른 사람들이 이 정보에 접근할 수 있는가?
- 미국에서는 이러한 문제에 대하여 어떤 구체적인 법률도 통과되지 않았고, 정부와 의료계에서는 환자의 프라이버시 간 균형을 찾는 연구가 진행중임

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-15. M2M 데이터

□ 빅데이터 거버넌스 구축사례_M2M 데이터

▷ M2M 데이터의 품질과 메타데이터

⊙ RFID

- RFID 리더는 중복이나 결측치와 같은 오류를 포함하는 방대한 양의 데이터를 생산함
- RFID 데이터는 ALE(Application Level Events)와 같은 표준을 따름. RFID 데이터 중 어떤 제품이 분실된 것으로 읽혀진 경우에 대하여, RFID 태그가 어떤 각도에서 읽혀지지 않을 수도 있고, 습기가 많아서 읽을 수 없는 경우도 있고, RFID 태그 자체가 변질되어 오류를 발생시키는 경우도 있음

⊙ 텔레매틱스

- 보험회사는 자동차에 속도 측정 센서를 설치하는데 대하여 고객이 동의하면 보험계약자에게 낮은 보험율을 적용. 센서의 오류때문에 텔레매틱스 애플리케이션이 높은 속도를 기록하게 되면 여러 가지 문제를 초래함

⊙ 전압 모니터링

- 전기 전압 모니터링 장치의 오류로 실제보다 높은 수치의 측정값을 기록함

⊙ 통신 네트워크 교환기

- 사용자의 폭발적인 전화사용에 대하여 살펴본 결과 교환기의 오류로 2,000만분 통화했다고 기록이 됨

⊙ 케이블 TV 셋톱박스

- 케이블 TV 사업자의 마케팅 팀은 고객이 채널을 찾아 다니는 정보를 활용하여 가입자가 시청한 프로그램과 상업 광고 동안에 채널을 바꾼 기록을 분석해낼 수 있음. 그런데 해당 데이터는 다른 TV 사업자의 장치와는 호환이 되지 않았음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

□ 빅데이터 거버넌스 구축사례_M2M 데이터

▷ M2M 데이터의 유지 기간 정책의 설정

⊙ 유럽연합의 데이터 보호위원회 제 29조는 위치정보 애플리케이션 혹은 서비스 제공자는 지리정보 데이터를 정당한 보유기간 이후에 폐기한다는 정책을 수립하여 실행해야 하는 것을 담고 있음

▷ M2M 데이터 지원을 위한 마스터 데이터의 품질 향상

⊙ M2M 데이터 활용성과는 양질의 마스터 데이터에도 의존하고 있음

⊙ 사례2. 철도에서의 고도의 상황 모니터링

- 철도에서의 고도의 상황 모니터링 절차

1) 센서의 설치 : 1,000개 이상의 서로 다른 기계적, 전기적 이벤트를 기록하고 있음. '문열림', '브레이크 작동중' 등과 같은 운행 이벤트, '라인 전압이상', '압축기 X에서 압축 정도 저하'와 같은 경고 이벤트, '집전기 고장', '인버터 록아웃' 등과 같은 고장 이벤트로 분류됨

2) 데이터의 수집과 분석 : 이전 이벤트와 매우 연관된 이벤트의 결정. 예를 들어 고장 이벤트 1245는 90%의 확률로 경고 이벤트 2389 다음에 발생함. 이 경우 운영팀에서는 시스템 로그에서 경고 이벤트 2389가 발견되면 예방정비를 위해 작업팀에게 지시를 내리게 됨

3) 유지관리의 수행 : 빅데이터 거버넌스 이슈인 센서 이벤트들의 일관성이 결여된 명명법(예를 들면 동일한 이벤트들에 대한 서로 다른 코드값), 열차가 수리점에 있을 때 열차들은 false positive를 생성함. 분석팀은 열차로부터 생성된 GPS 데이터를 철도 수리점의 위치정보와 결합하여 false positive 데이터를 제거함, 예방정비는 자산에 대하여 일관성이 결여된 명명법 사용으로 어려움이 가중되고 있음(예를 들면 동일한 부품명의 서로다른 값을 가짐)

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-15. M2M 데이터

□ 빅데이터 거버넌스 구축사례_M2M 데이터

▷ 사이버 공격에 대한 약점을 이기는 SCADA 인프라 강화

⊙ SCADA(Supervisory Control and Data Acquisition)는 산업 프로세스, 기반시설 프로세스, 설비기반 프로세스를 모니터링하고 제어하는 컴퓨터 시스템

⊙ 사용 분야

- 산업 프로세스 : 제조, 생산, 전력 생산, 부품 조립, 정제 등의 산업에서 필요한 과정임. 이들은 연속적으로, 배치 방식으로, 반복적인 방식으로, 이산 모드 등으로 실행됨
- 기반시설 프로세스는 공공 혹은 사적인 것으로 수자원 관리와 분배, 오수의 수거와 처리, 오일 및 가스 파이프라인, 전력 전송과 분배, 풍력 발전, 시민 대피 사이렌 시스템, 거대 통신 시스템 등
- 설비 프로세스는 공공 혹은 사적인 영역에서 발생하는 것으로 빌딩, 공항, 배, 우주 정류장 등을 포함. 이들에 대한 난방, 환기, 공기 정화, 에너지 소비 등을 모니터링하고 제어함

⊙ 사례3. 지멘스 SCADA 시스템을 공격한 Stuxnet 웜

- 2010년 6월 이란의 14개 발전소를 관리하는 지멘스사의 SCADA 시스템에 Stuxnet이라 불리는 웜이 발견됨
- 웜은 산업 기밀을 빼내고 운행을 중단시키도록 설계되었음. 알려지지 않은 윈도우의 취약점을 이용하였음

⊙ 사례4. 스마트 그리드 보안 개선방안

- 스마트 그리드가 인터넷 프로토콜(IP)과 표준 기법들을 사용하므로 여러가지 측면에서의 보안의 강화가 필요함
- 보안 관련 빅데이터 거버넌스 프로그램은 핵심 자산들을 식별하여 모니터링하고, 외부 공격으로부터 보호하며, 특정 문서와 계획들에 대한 기록의 유지기한을 설정함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-15. M2M 데이터

□ 빅데이터 거버넌스 구축사례_M2M 데이터

▷ 사이버 공격에 대한 약점을 이기는 SCADA 인프라 강화

⊙ 사례5. 오일과 가스 산업에서의 센서 데이터 대한 거버넌스

- 오일 산업에서의 센서 데이터를 관리하는 프로세스(핵심활동/마일스톤)

마일스톤/활동	설명
1. 센서 설치	오일과 가스 회사들은 시설에 센서를 설치하여 생산과 시설의 상태, 안전성, 환경 규제에의 적합성 등을 모니터링하고자 한다. 센서 제어 시스템은 서로 다른 제조사에서 만든 SCADA 시스템들 사이의 실시간 통신을 지원하는 표준 OPC 프로토콜을 사용하고 있다.
1.1 시설에 센서를 설치함	최신의 오일 설비에는 30,000개 이상의 센서가 부착되어 다양한 유형의 실시간 데이터(흐름, 분당 회전수, 전압, 와트수, 온도, 압력 등)를 받아들인다.
1.2 해저에 센서를 설치함	회사는 센서를 해저에도 설치하여 환경 상황(물의 흐름, 온도, 혼탁의 정도 등)을 모니터링해야 한다. 탁도는 수질에 관한 측도로써 육안으로는 볼 수 없는 개체들에 의해 발생하는 혼탁함의 정도를 의미한다.
2. 생산 모니터링	회사는 오일과 가스의 생산과정을 모니터링해야 한다. 오일 회사는 설비주인들에게 할당될 생산량을 계산해야 한다.
2.1 설비에서 생산 모니터링	운영자는 각 설비에 센서를 부착하여 오일과 가스의 생산을 모니터링한다.
2.2 생산 대시보드 운영	오일과 가스회사는 대시보드를 만들어서 시설들에서 에너지 생산을 모니터링한다. 오일과 가스회사는 공동의 운영센터를 만들어서 중앙에서 생산을 모니터링할 수 있도록 한다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-15. M2M 데이터

□ 빅데이터 거버넌스 구축사례_M2M 데이터

▷ 사이버 공격에 대한 약점을 이기는 SCADA 인프라 강화

⊙ 사례5. 오일과 가스 산업에서의 센서 데이터 대한 거버넌스

- 오일 산업에서의 센서 데이터를 관리하는 프로세스(핵심활동과 마일스톤)(계속)

마일스톤/활동	설명
3. 장비 모니터링	시설에 설치된 장비들을 모니터링하기 위해 센서를 활용한다.
3.1 시설에 설치된 장비 모니터링	운영부서는 각 시추공에 있는 펌프와 밸브와 같은 장비들을 모니터링하여 다음과 같은 전형적인 질문에 답할 수 있게 한다. - 현재 발견된 방식으로 장비의 떨림 현상이 시작된다면, 이 브랜드의 엔진은 언제 고장이 날 것인가? - 유정에 알람이 울릴 때 (그 유정의 과거 행위에 기반을 둔 분석으로) 적합한 조치를 취하는데 얼마나 많은 시간이 필요한가? - 관측된 데이터로부터 날씨 이벤트를 어떻게 감지하는가? - 어느 센서가 주어진 위치로부터 반경 100마일 이내에 발생한 눈보라를 감지했는가?
3.2 예방정비를 수행함	만약, 예방 모델이 장비의 특정 일부가 고장날 가능성이 있음을 나타낸다면, 운영자들은 예방 정비를 수행한다.
4. 환경 모니터링	오일과 가스 회사는 환경을 모니터링하기 위해 센서를 사용한다.
4.1 시설 주변의 해저 환경 모니터링	환경 센서들은 platform의 운전(동작) 이전에, 도중에, 그리고 이후에 작동할 것이다.
4.2 시간이 지남에 따른 환경오염 모니터링	기업들은 "시설 주변물의 염도와 탁도가 석유 유출을 나타내는가"와 같은 질문에 답변할 수 있어야 한다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-15. M2M 데이터

□ 빅데이터 거버넌스 구축사례_M2M 데이터

▷ 사이버 공격에 대한 약점을 이기는 SCADA 인프라 강화

⊙ 사례5. 오일과 가스 산업에서의 센서 데이터 대한 거버넌스

- 유전 센서 장치들과 관련된 주요 빅데이터 거버넌스 정책들

마일스톤/활동	빅데이터 거버넌스 정책
1.1 시설에 센서를 설치함	빅데이터 거버넌스 프로그램은 SCADA 시스템들이 혹시 일어날 사이버 공격에 대해 적절히 보호되는지 보장해야 한다.
2.2 생산 계기판 만들기	빅데이터 거버넌스 프로그램은 생산 리포트내에서의 비즈니스 언어들이 일관성 있음을 확신해야 한다.
3.1 시설에 설치된 장비 모니터링	과거에는 한 굴착장치 당 1,000개의 센서밖에 없었을 것이다. 그 중에서도 약 10개만이 용량제한으로 인하여 2주마다 제거되는 데이터베이스에 저장되었다. 요즘의 오일과 가스 회사들은 더 많은 센서 데이터를 오랫동안 보유해야 한다.예를 들어, HSE(보건, 안전, 환경) 부서는 왜 그 분야에서 특정한 결정이 이루어졌는지를, 정당화하기 위해 3개월이나 된 오래된 정보를 사용하여 분석해야 할 것이다. 빅데이터 거버넌스 프로그램은 오일, 가스 생산시설 그리고 관련된 definition을 위한 ISO 15926 같은 표준 모델을 활용해야 한다. 또한 빅데이터 거버넌스 프로그램은 얼마나 많은 정보가 보유될 필요가 있는지 그리고 내적인 요구와 규제를 충족시키는 데에 얼마나 많은 데이터를 얼마동안 보유해야 하는지를 결정하는데 중요한 역할을 담당해야 한다. 여기서 그 굴착기가 비디오, 오디오, 그림, 소리와 같은 다수의 비구조화된 정보를 생산할 수도 있다는 것을 알아차리는 것도 중요하다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-15. M2M 데이터

□ 빅데이터 거버넌스 구축사례_M2M 데이터

▷ 사이버 공격에 대한 약점을 이기는 SCADA 인프라 강화

⊙ 사례5. 오일과 가스 산업에서의 센서 데이터 대한 거버넌스

- 유전 센서 장치들과 관련된 주요 빅데이터 거버넌스 정책들(계속)

마일스톤/활동	빅데이터 거버넌스 정책
3.2 예방적 유지보수 실행하기	만약 한 굴착기에서 특정 타입의 장비가 고장난 경우, 오일 회사는 또 다른 곳에 동일한 장비가 배치되어 있는지 재빨리 확인하여 점검하는 것이 중요하다. 그러나 만약 동일한 장비가 다른 굴착기에서 다른 이름으로 명명된 경우, 그 장비를 적시에 찾아내기 어려워질 것이다. 결과적으로, 빅데이터 거버넌스는 자산 데이터에 관한 일관성있는 명명법을 보장하는 중요한 역할을 담당해야 한다. 자산관리 연구소와 영국 표준 연구소는 핵심 비즈니스 자산에 대한 리스크를 줄이기 위한 전략들을 개발하기 위해 함께 연구해왔다. 연구결과로 PAS(Publicity Available Specification) 55를 발표하였으며, 이 프로젝트는 자산관리 시스템들 안에서의 최고의 실제 사례(관례)에 관하여 최신의 생각을 담고 있는 (PAS) 55를 낳았다. 오일과 가스회사들은 점점 더 PAS 55를 자산관리를 위한 산업기준으로 채택하고 있는 추세이다.
4.1 시설 주변의 해저 환경 모니터링	조금 전에 논의되었듯이, 석유탐사와 생산활동은 다수의 구조적이거나 비구조적인 환경적 정보를 만들어낸다. 이 정보는 환경적 규제를 고수했는지를 증명하기 위하여 그 시설 자체의 수명이 다한 다음에도 계속 온전히 유지될 필요가 있다. 이러한 정보는 50년에서 70년, 심지어 몇몇 경우에는 100년까지도 저장되어야 할 필요가 있을 수 있다. 저장은 값쌀 수는 있지만 무료는 아니다. 빅데이터 거버넌스 프로그램은 특정 유형의 정보를 위한 유지계획(retention schedules)을 정립하고 또 가능하다면 정보를 상대적으로 더 싼 보관함으로 옮기기 위한 적절한 보관정책을 세울 필요가 있다.

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-16. 빅 트랜잭션 데이터

□ 빅데이터 거버넌스 구축사례_빅 트랜잭션 데이터

▷ 빅 트랜잭션 데이터의 예

산업	빅 트랜잭션 데이터	데이터의 특징
통신 및 서비스 제공업자	통화 상세기록 데이터(CDRs, Call Detail Records)	반구조화, 빠른 속도
의료보험	청구 데이터	비구조화 데이터
보험산업	청구 데이터	구조화된 데이터, 소셜 미디어와 통합 필요성 증대
시설산업	스마트 미터기에 기반을 둔 요금청구 데이터	반구조화, 빠른 속도
제약 분배업자	약국 청구기록 데이터	대부분 구조화된 데이터, 소셜 미디어와 통합 필요성 증대
은행	금융거래 데이터	비용절감을 위해 Hadoop 처리
소비재 제품산업	POS 트랜잭션 로그 데이터	반구조화, 소셜 미디어와 통합 필요

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-16. 빅 트랜잭션 데이터

□ 빅데이터 거버넌스 구축사례_빅 트랜잭션 데이터

▷ 사례1. 통신사의 CDRs 중복 데이터 제거를 위한 스트리밍 기술의 사용

⊙ 핸드폰 교환기는 전화발신, 이메일, 웹 브라우징 세션, 텍스트 메시지 등에 대하여 CDRs 데이터를 생성함. 통신사는 초당 최고 100,000 CDRs를 처리함

- 데이터 손실을 막기 위하여, 통신사의 네트워크 내의 스위치들은 각 트랜잭션마다 2개의 CDRs를 생성
- 한 스위치가 리부팅되면 CDRs를 재송신

⊙ 통신사는 중복된 CDRs를 실시간으로 제거하여 요금계산과 같은 후속업무를 지원해야 함. CDRs의 볼륨이 늘어나면, 적절한 시기에 CDRs를 조정하는 것이 어려워짐. 따라서 통신사는 CDRs를 준실시간으로 처리할 수 있어야 하는 동시에 이탈고객에 대한 예측도 수행해야 함

⊙ 통신사는 스트리밍 분석을 활용하여 각 CDR을 실시간으로 기존의 수십억 개의 CDRs와 비교하고 있고, 이 작업은 디스크에 데이터를 넣지 않고 진행했는데, 중복된 CDRs을 실시간으로 제거하고, 데이터베이스내에 저장하는 데이터의 양도 절반으로 줄임

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-16. 빅 트랜잭션 데이터

□ 빅데이터 거버넌스 구축사례_빅 트랜잭션 데이터

▷ 사례2. 유럽국가의 중앙집권화된 보험금 청구 데이터베이스

⊙ 보험회사는 청구손실과 조정비용이 많은 비중을 차지하므로 청구와 관련된 빅 트랜잭션 데이터를 관리할 필요가 있음

- 특정 보험업자의 청구조사관은 상대적으로 저소득 우편구역에서 발생한 마세라티 승용차 절도에 대한 단독 청구를 조사할 가치가 없다고 느꼈으나, 동일한 우편구역내에서 30명의 각기 다른 사람들이 다른 보험업자에게 지난 2달간 고급승용차들에 대해 도난 청구를 제출했다는 것을 미리 알았다면 조사 필요성을 느꼈을 것
- 유럽국가 내 보험회사들이 동일 목적을 가지고 청구 정보를 데이터베이스화했음. 동일한 지역, 동일하거나 유사한 생일을 가진 사람, 동일한 은행의 계좌를 가진 사람, 비슷한 이름을 가진 사람들을 퍼지매칭함

⊙ 현재는 청구 조사관들은 다수의 보험업자사이에서 청구인들의 이름이나 우편번호, 은행계좌 등을 공유하고 있는 동일한 지역의 청구건들을 비교할 수 있게 되었음

⊙ 데이터 관리당국의 빅데이터 거버넌스를 통한 전반적인 접근의 가치 상승

- 효율적인 분류, 데이터 기반 이력, 청구 정보의 수명주기 관리

: 당국의 거대 통합 데이터베이스 구축에 참여하는 보험업자에게 그들의 데이터는 안전하고 경쟁사에게 공개되지 않을 것이고, 단골 고객과의 관계에 해가 되는 일이 없을 것임을 안심시켜야 함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-16. 빅 트랜잭션 데이터

□ 빅데이터 거버넌스 구축사례_빅 트랜잭션 데이터

▷ 사례 2. 유럽국가의 중앙집권화된 보험금 청구 데이터베이스

⊙ 데이터 관리당국의 빅데이터 거버넌스를 통한 전반적인 접근의 가치 상승(계속)

- 청구 분석의 보안성과 비밀성

  : 보험 사기꾼으로 하여금 이러한 사업에 대한 결과 내용의 절대 비밀 유지

- 실시간 혹은 준실시간으로 이력 데이터에 접근

  : 보험업자들은 청구들에 대하여 보험금이 지불되기 전후에 조사함

- 청구들을 검토하기 위해 과거 정보의 이용

  : 통합 데이터베이스가 구축되기 전에는 노동집약적인 특성 때문에 보험업자들이 세밀한 조사없이 대부분의 청구에 대해 보험금을 지불하였음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-17. 생체 데이터

□ 빅데이터 거버넌스 구축사례_생체 데이터

▷ 생체 데이터 정의 및 특징

⊙ 생체인식은 사람의 해부학적 또는 행동적인 특징, 습관을 근거로 그 사람에 대한 자동적인 식별을 말함

⊙ 해부학적 데이터는 지문, 홍채, 망막, 얼굴, 손의 모양, 귀 모양, 음성 패턴, DNA, 체취를 포함한 사람의 신체적 특징으로부터 만들어짐. 행동적 데이터는 필체나 키보드 조작과 관련된 데이터임

⊙ 생체 데이터는 상업용으로 확산되고 있으며, 소셜 미디어와 같은 다른 데이터와 융합되는 추세임

▷ 생체 데이터의 적합한 사용과 관련된 프라이버시 영향 평가

⊙ 사례1. 얼굴 인식 기술과 소셜 미디어의 결합으로 인한 프라이버시 영향 평가

- Facebook이나 LinkedIn과 같은 소셜 네트워크들은 Yelp와 Amazon과 같은 사이트와 더불어 사용자로 하여금 프로필 포토를 업로드하고 이러한 사진들을 공개하도록 장려하고 있음. 최근 카네기 맬론 대학의 연구자들은 사람의 사진으로부터 당사자의 이름, 지리적 위치, 관심사, 그리고 심지어 당사자의 주민등록번호 첫 번째 5자리까지 알아내는 것이 대부분 가능함을 보여줌. 얼굴인식 소프트웨어와 데이터마이닝 알고리즘 및 통계적 인식기술을 결합하여 사용
- 얼굴 데이터 사용에 대한 향후의 법적인 사항 고려

. 조직들은 얼굴 정보의 보유시간을 줄이고, 적절한 보안 측정을 수행하며, 고객이 떠나간 후 해당 데이터 즉각 삭제

. 조직은 개인이 공급하는 얼굴 데이터가 제3자나 공공으로 사용가능한 소스에 연결될 수 있음을 당사자에게 공지

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

□ 빅데이터 거버넌스 구축사례_생체 데이터

▷ 고객과 종업원에 대한 생체 데이터 사용 시 규제 영향 분석을 위한 법률 자문

⊙ 사례2. 유전 데이터 사용을 관리하는 법규의 요약

- 미국

2008년도의 유전자정보 차별금지법은 의료보험과 고용에 있어서 유전적 정보에 근거한 차별을 금지하고 있음

- 영국

영국 보험연합은 생명보험 이외의 다른 종류의 보험에 유전자 검사를 수행하는 것에 대한 중단 선언. 보험회사들은 정부가 특히 그 검사를 인정하지 않는 한 예측적인 유전자 검사를 사용하지 않을 것임

- 유럽연합

2012년 1월 25일, 유럽위원회는 개인 데이터의 처리와 관련한 개인 보호와 그러한 데이터의 자유로운 이동에 관한 범용 데이터 보호규제 초안을 발표. 유전적 데이터란 태생적인 또는 태아 성장기에 습득된 개인의 특성에 대한 모든 종류의 데이터로 규정. 이 규제에서는 유전적 정보를 사용하는 조직들이 데이터 보호 영향평가를 수립하도록 요구함

- 호주

프라이버시 보호법에서 고지 및 개인 동의에 근거하여 유전적 정보의 활용을 관리함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-18. 사람이 생성한 데이터

□ 빅데이터 거버넌스 구축사례_사람이 생성한 데이터

▷ 사람이 생성한 데이터 예

산업	사람이 생성한 데이터	빅데이터 거버넌스 문제점
산업간	음성녹음	음성녹음에서 민감한 정보를 마스킹함
이메일 메시지	MDM 데이터 보강을 위한 항목 추출	
건강보험 제공자	전자 의료 기록	비구조화된 내용을 활용하여 구조화된 데이터를 보강하는 동안에 메타데이터, 프라이버시, 마스터 데이터 문제를 고심함
건강보험	콜센터 요원과 간호원의 노트	건강 및 웰니스 프로그램의 예측 모델내에서 데이터 품질을 향상시킴

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-18. 사람이 생성한 데이터

□ 빅데이터 거버넌스 구축사례_사람이 생성한 데이터

▷ 사람이 생성한 민감한 데이터를 감추기 위한 정책 수립

⊙ 사람이 생성한 데이터는 개인을 인식할 수 있는 정보를 담고 있으므로 마스킹하는 것이 필요함

⊙ 사례1. 음성 데이터 내의 민감한 정보를 감추기 위한 빅데이터 거버넌스 정책

- 많은 콜센터들은 다양한 목적을 위하여 통화의 전부 혹은 일부를 녹음해서 남겨 놓고, 사후에 또는 즉시 통화를 분석함
  . 운영 효율성 제고(음성통화 후 처리) : 고객들의 콜센터 전화 이유를 파악하여 셀프 서비스 옵션을 사용하도록 유도
  . 품질 보증(음성통화후 처리) : 품질관리부에서는 CSR들이 통화를 공손히 하고 정책에 따라서 잘하고 있는지를 점검
  . 교차판매와 상향판매(음성통화의 실시간 처리) : 몇몇 콜센터는 CSR들이 고객들과 통화중에 새로운 제의를 할 수 있도록 실시간 음성분석을 사용함
- 콜센터들은 음성 녹음에 민감한 정보가 들어 있을 때 전화를 건 사람의 프라이버시를 보호해야 할 필요가 있음
  . 민감한 데이터의 녹음을 막기 위한 기술을 제시
  . 인가된 통화녹음 안에 포함된 민감한 데이터를 안전하게 삭제할 수 있는 기술을 제시
  . 왜 민감한 정보들이 제거될 수 없는지 정당한 이유를 서류로 입증하기, 위험평가 실시하기, 통화 녹음 데이터가 문의될 수 없고 PCI DSS에 따라 잘 보호됨을 문서로 확인

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-18. 사람이 생성한 데이터

□ 빅데이터 거버넌스 구축사례_사람이 생성한 데이터

▷ 구조적 데이터의 품질향상을 위해 사람이 생성한 비구조적 데이터 사용하기

⊙ 사람이 생성한 데이터는 구조화된 데이터 내에서는 얻을 수 없는 통찰력을 제공함

⊙ 사례2. 건강보험에서 헬스와 웰니스 프로그램을 위해 빅데이터를 들여오기

- 비즈니스 인텔리전스 관리자는 콜센터 요원과 간호사에 의해 만들어진 노트에서 "울혈성 심부전"과 같은 키워드를 골라내기 위해 텍스트 분석기술을 도입함
- 건강보험은 이 데이터를 예측 모델에 넣어서 만약 특정인이 고위험군에 있어 향상된 케어 관리를 받아야 한다는 결정을 하는데 도움을 제공함. 가입자들이 의사의 지시를 따르지 않아서 전반적인 비용을 높이고 있음을 주지시킴
- 이러한 통찰력은 "나는 내 의사가 싫어"와 "나는 내 약이 싫어"라는 핵심 문구에 기반하여 얻을 수 있었음
- 비즈니스 인텔리전스 관리자는 비즈니스로 부터 호응을 받는 것이 가능해짐

▷ 사람이 생성한 데이터의 수명주기관리를 통한 비용절감 및 규제 준수

⊙ 조직들은 이메일이나 음성녹음과 같이 사람이 생성한 데이터에 관한 규제 요구사항을 준수해야 하는데, 비용 측면을 고려하여 보관 기간이 지난 데이터들은 폐기, 삭제, 보관을 하는 등 음성 데이터도 수명주기 관리 대상이 됨

▷ 사람이 생성한 데이터 활용 MDM 보완

⊙ 사람이 생성한 비구조적 데이터가 고객관계 등 MDM 데이터를 풍부하게 하는데 기여할 수 있음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-19. 헬스케어

□ 빅데이터 거버넌스 구축사례_헬스케어 데이터

▷ 사례1. 울혈심부전증 환자의 30일 이내 재입원 비율을 줄이기 위한 빅데이터 거버넌스 활용

⊙ 빅데이터를 이용한 파일럿 테스트

- 병원시스템은 응급서비스를 포함해 다양한 서비스를 제공하는 15개의 시설로 이루어져 있음
- 병원은 울혈심부전증으로 재입원하는 환자의 비율을 낮추기 위해 빅데이터 분석을 활용
- 파일럿 연구 목적

1) 보험으로 처리되지 않는 비용 줄임

2) 병의 진행을 막기 위해 선제적으로 병을 발견해서 의료의 질을 높임

30일 이내 재입원할 가능성이 높은 환자에게 금연 또는 가정에서의 건강관리 프로그램 참여 높임

- 분석 부서에서는 150개의 변수와 5년 동안 축적된 2만 명의 환자 기록을 바탕으로 예측 모델을 제작함
- 사용된 데이터의 소스들은 전자의료기록 패키지, 입원시스템, 원가계산 데이터베이스를 포함한 다양한 애플리케이션

⊙ 비구조적 데이터 활용으로 구조적 데이터의 품질 개선

- 병원시스템의 분석 팀은 흡연 여부, 약물, 알코올 중독 등을 포함한 많은 변수들이 환자의 재입원율을 결정하는 핵심요인임을 밝혀냄

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

□ 빅데이터 거버넌스 구축사례_헬스케어 데이터

▷ 사례1. 울혈심부전증 환자의 30일 이내 재입원 비율을 줄이기 위한 빅데이터 거버넌스 활용

⊙ 비구조적 데이터 활용으로 구조적 데이터의 품질 개선(계속)

- 흡연여부는 심장병의 핵심 요인임에도 불구하고 병원들은 환자의 흡연기간과 빈도 등에 관한 완전한 이력 데이터를 가지고 있지 못했음. 흡연여부와 관련해서는 예/아니오를 25%만 가지고 있었는데, 이를 85%까지 높였으며 흡연기간과 흡연빈도 관련 정보도 찾을 수 있었음
- 의료팀은 약물과 알코올 남용이 병원 재입원율의 핵심지표라는 것을 경험상으로 알고 있었는데, 환자의 20%만이 입원신청서에 해당 칸을 체크했는데, 분석팀은 비구조적 데이터를 이용하여 이 수치를 76%까지 올렸음

⊙ 구조적 데이터로부터 얻을 수 없는 추가 의료요인들의 추출

- 구조적 데이터에서는 알 수 없는 요양시설 및 약물학 같은 의료지표들을 발견함
- 의료팀은 요양시설의 환자들이 혼자 사는 환자보다 약을 더 잘 챙겨먹게 된다는 것을 발견함. 병원시스템은 이 정보를 정형화된 형태로 찾아내지 못했는데, 비즈니스 인텔리전스 팀은 퇴원기록, 초음파 심장 진단도, 환자기록, 의사소견서, 건강진단서의 텍스트를 분석하여 25%의 환자가 요양시설에 있다는 것을 알아냄
- 약물학 규제 관련 지표는 환자가 그들의 치료계획에 따라 약을 어느 정도까지 먹어야 하는지를 알려주기 때문에 의사나 사례 관리자에게 매우 중요. 비즈니스 인텔리전스 팀은 의사소견서와 전자의료기록을 분석하여 이 데이터를 추가 수집

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

□ 빅데이터 거버넌스 구축사례_헬스케어 데이터

▷ 사례1. 울혈심부전증 환자의 30일 이내 재입원 비율을 줄이기 위한 빅데이터 거버넌스 활용

⊙ 핵심 비즈니스 용어들에 대한 일관된 정의

- '재입원'이라는 용어는 세가지 다른 정의를 가짐

. 의료관점 : 30일 이내, 모든 요인(울혈심부전증 관련 증상 여부에 무관)

. 의료관점 : 30일 이내, 동일한 진단

. 재정적 관점 : 분기별, 연별. 재정적 관점에서는 재입원을 6개월에서 9개월의 기간을 포함해 더 긴 기간 입원한 것으로 정의함

⊙ 모든 시설에서 환자 마스터 데이터의 일치성 보장

- 병원 내부 각기 다른 시설에 저장되어 있는 동일한 환자의 의료기록을 추적하는 것은 사실상 불가능
- 병원시스템은 동일한 환자와 관련된 의료기록들을 통합할 수 있도록 함. 그러나 환자가 매우 짧은 기간 동안에 여러 시설들로 재입원하는 경우 그 환자의 의료기록을 통합하는데 상당한 시간을 낭비함

⊙ 미국 HIPAA 규정에 의거 의료정보에 대한 프라이버시 보호

- 비싼 긴급 의료서비스 대신에 병원으로의 무료 픽업 서비스를 사용할 수 있은 환자를 찾아내는 과정에서, 환자의 주소사용 동의를 얻는 방식으로 개인정보보호 규정을 준수함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-19. 헬스케어

□ 빅데이터 거버넌스 구축사례_헬스케어 데이터

▷ 사례1. 울혈심부전증 환자의 30일 이내 재입원 비율을 줄이기 위한 빅데이터 거버넌스 활용

⊙ 참조 데이터의 창조적 관리를 통한 추가적인 의료 통찰력 확보

- ICD-9 참조 데이터는 세밀한 부분까지 아주 잘 정의된 데이터베이스임. 예를 들어 ICD-9은 심장병에 428 코드를 부여함. 428.1은 심장 왼쪽의 문제, 428.2는 심장 수축의 문제를 나타냄
- ICD-9 코드로 유사한 병을 분류하는 것이 가능해지면서 이러한 연구 결과를 더욱 효율적으로 진행할 수 있게 함
- 분석 팀은 2만 1천 개 이상의 ICD-9 코드로 20개 질병을 분류하기 위해 의사들과 협력함. 이 과정을 통해서 분석 팀은 분석 과정에서의 방해물을 줄였고, 더 나은 의료 통찰력을 얻을 수 있었음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-20. 유틸리티 산업

□ 빅데이터 거버넌스 구축사례_유틸리티 산업 데이터

▷ 사례1. 유틸리티 산업(수도, 가스, 전기 등)에서 발생하는 대용량 스마트 계량기 데이터

⊙ 스마트 계량기 기능 및 특징

- 전력, 가술, 수도 등의 소비량을 측정하는데 사용됨
- 무선기능과 결합하여 자동 계량이 가능함. 또한 실시간 센서를 가지고 있어서 정점과 전력의 질을 모니터링 할 수 있음. 예전에는 매달, 매 분기별로 사용량 검침이 가능했지만 지금은 어느 때라도 검침이 가능함
- 스마트 계량기로 인해 사용자 별 가격책정의 차별화, 실제 사용량에 의한 청구 가능해졌으며, 유틸리티 산업에는 설비투자 감소, 더 낮은 계량비용, 개선된 사용자 분석이 가능해짐

⊙ 스마트 계량기 프로그램의 솔루션 구조

- 1. 계량기 송신장치(MTUs, Meter Transmission Units)
  건물 하나당 1개의 MTU가 있으며 6시간마다 무선으로 간단한 정보를 송신함. 이 정보는 MTU 펌웨어에 포함된 위치 식별자, 타임 스탬프, 검침된 계량기 등으로 구성됨. 이 시스템은 하루에 300만개 이상의 고유한 계량치를 생산하고 계량치를 3일 분량을 보유할 수도 있음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

□ 빅데이터 거버넌스 구축사례_유틸리티 산업 데이터

▷ 사례1. 유틸리티 산업(수도, 가스, 전기 등)에서 발생하는 대용량 스마트 계량기 데이터

⊙ 스마트 계량기 프로그램의 솔루션 구조(계속)

- 2. 수집 MTUs

  MTU가 읽은 계량치는 380개의 수집 MTUs로 보냄. 중복기능을 가지고 있어서 각 MTU가 읽은 계량치는 서버 개의 수집 MTUs로 중복되게 저장함

- 3. 데이터 수집 센터(Data Collection Center)

  각 수집 MTU는 10, 50 또는 300개의 계량치가 담긴 파일을 데이터 수집과 차후 처리를 위해 데이터 수집 센터에 보냄. 데이터 센터는 8개의 서버로 구성되어 있으며 파일을 파싱하고 MTU 식별자를 읽으며, 읽은 계량치를 해당 계좌번호에 저장하기 위해 조회 테이블을 사용. 24시간 동안 1백만 개 이상의 원시 데이터 파일을 처리함

- 4. 데이터 분석 센터(Data Analysis Center)

  데이터는 새벽 1시부터 6시까지 야간 배치 모드 분석환경으로 옮겨짐

- 5. 레거시(기존의) 가격 책정 애플리케이션

  메인프레임상에서 운영되는 가격책정 애플리케이션은 계좌번호, 이름, 주소와 같은 고객 마스터 데이터를 가지고 있음. 주소 데이터는 온라인 주소 표준화 시스템에 의해 유효한 것으로 인증됨

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-20. 유틸리티 산업

□ 빅데이터 거버넌스 구축사례_유틸리티 산업 데이터

▷ 사례1. 유틸리티 산업(수도, 가스, 전기 등)에서 발생하는 대용량 스마트 계량기 데이터

⊙ 미터기 계량치의 중복 문제

- 문제점

각 MTU가 읽은 계량치를 서너 개의 저장소로 중복시키므로 적절한 관리가 없는 경우, 데이터 품질 문제 발생 가능

- 해결책

수도 사업체는 중복 전송되는 계량치를 지속적으로 모니터링하여 특정 건물이나 계량기에 대한 각 타임스템프에 하나의 고유한 계량치 만이 기록되도록 하는데 애플리케이션이 1천만 개 이상의 계량치 데이터중에서 단지 3백만 개 정도만 중복되지 않기 때문에 상당한 시간이 소요됨. 따라서 이를 개선하기 위해 데이터가 실시간에 동적으로 처리될 수 있도록 스트리밍 분석을 도입함. 그 결과 중복되지 않는 계량치만을 데이터베이스로 입력시키는 것이 가능해짐

⊙ 기본키의 참조무결성

- 문제점

데이터베이스의 몇몇 테이블은 기본키로 계량기의 타임스탬프를 사용하고 있음. 타임스탬프는 특정 계량기에만 있는 것이 아니기 때문에 참조무결성을 지키지 못하게 됨. 오류비용이 상당히 많이 들어감

- 해결책

MTU 위치식별자와 타임스탬프로 구성된 복합 키의 사용으로 참조무결성 문제 해결

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-20. 유틸리티 산업

□ 빅데이터 거버넌스 구축사례_유틸리티 산업 데이터

▷ 사례1. 유틸리티 산업(수도, 가스, 전기 등)에서 발생하는 대용량 스마트 계량기 데이터

⊙ 이상치 문제

- 문제점

  자동화된 환경이라도 3백만 개의 기록들 중 아주 일부는 틀리 수 있음

- 해결책

  스마트 계량기 애플리케이션은 누계와 각 계좌에서 정상분포 패턴을 보이는 MTU의 일일 평균소비량을 계산하여 오류나 이상치 기록을 점검하여 이상치를 버리고 이전 기록을 사용하거나 소비량 히스토라인을 근거로 추정치를 사용함. 애플리케이션은 이 값이 추정치라는 것을 표시함

⊙ 고객 주소의 데이터 품질

- 문제점

  시스템은 동일한 고객임에도 가격책정 애플리케이션(메인프레임 상에 위치), 데이터 저장소, 분석센터(분산 환경) 등에서 주소가 상이한 경우를 찾아야 함

- 해결책

  메인프레임의 가격책정 시스템은 고객주소를 표준화하기 위해 라이브 애플리케이션을 활용함. 메인프레임과 분산환경에서 고객주소가 상이할 때 가격책정 애플리케이션의 주소를 정확한 것으로 간주함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-20. 유틸리티 산업

□ 빅데이터 거버넌스 구축사례_유틸리티 산업 데이터

▷ 사례1. 유틸리티 산업(수도, 가스, 전기 등)에서 발생하는 대용량 스마트 계량기 데이터

⊙ 정보 수명주기 관리

- 문제점

스마트 미터기의 계량치를 적절히 보관소로 이동하여 보관하지 않으면, 새로운 계량치를 입력하거나 쿼리를 수행할 때 성능이 저하됨

- 해결책

데이터 수집 센터는 90일이 지난 데이터는 삭제. 분석 센터는 2009년부터 2011년까지의 데이터를 월별로 분할하여 보관함. 2년이 넘은 데이터는 아카이브로 보관하여 필요 시 데이터를 통합하여 사용함. 이경우 통합 쿼리를 개발해야 함

⊙ 데이터베이스 모니터링

- 문제점

스마트 계량기로 인한 잠재적 사생활 침해문제가 신문기사에 보도됨.(스마트 계량기 데이터로부터 전자레인지 음식 조리 빈도, 타월 세탁 빈도, 어느 브랜드의 세탁기를 사용하는 지 등)

- 해결책

데이터베이스 관리자 등 특권을 가진 사용자가 스마트 계량기 데이터에 접근하는 걸 모니터링하는 정책을 전체 로드맵의 일부로 수립함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-21. 통신서비스 공급자

□ 빅데이터 거버넌스 구축사례_통신서비스 공급자 데이터

▷ 통신서비스 공급자(CSP, Communication Service Provider)와 빅데이터 거버넌스

⊙ CSP 정의 : 전화, 무선통신, 인터넷, 케이블, 위성서비스 공급자를 포함하는 넓은 범위의 회사 지칭

⊙ 빅데이터 거버넌스 부문에서 떠오르고 있는 과제

- 1. 고객에 대한 싱글 뷰

  데이터 분석을 통해 고객이 자사 제품을 커뮤니케이션, 콘텐츠, 상업적 필요에 따라 어떻게 사용하는지를 완전하게 파악이 가능. 새롭게 얻은 데이터를 기존의 축적했던 다른 데이터와 통합하여 고객에 대한 포괄적인 이해를 하는데 사용해야 할 필요성이 존재함

- 2. 빅데이터 품질

  고객 데이터는 데이터 품질 수준이 상이한 여러 샘플에서 얻어지는데 데이터를 어떻게 통합하여 그 데이터가 신뢰성을 가지게 할 것인가가 중요함

- 3. 정보 수명주기 관리

  새로운 데이터는 CSP가 지금까지 경험했던 것보다 훨씬 더 많은 데이터임. 현재의 분석시스템은 분석이 불가능. 따라서 이러한 규모의 데이터를 어떻게 실시간으로 혹은 준실시간으로 저장하고 분석하며 사용하는가가 중요함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-21. 통신서비스 공급자

□ 빅데이터 거버넌스 구축사례_통신서비스 공급자 데이터

▷ CSP 빅데이터 유형

⊙ 네트워크 이벤트 : 네트워크 장비로부터 수집된 데이터
예) 통화중 연결이 끊어지는 것과 같은 데이터를 통해 서비스 품질과 사용현황에 대한 정보 제공

⊙ 통화, 사용 세부 기록
예) 발신, 종료, 음성 지속시간 등

⊙ 위치 정보
예) 무선단말기의 위치, 최종적인 전화 끊은 위치, GPS 기술 활용가능

⊙ 웹 트래픽
예) 장치에 남은 웹 쿠키 저장

⊙ 채널 클릭
예) 케이블 TV의 셋톱박스내 저장 정보(채널 선택 정보 등)

⊙ 소셜 미디어
예) 신상품 권고나 고객서비스에 대해 Twitter, YouTube에 올리는 의견

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-21. 통신서비스 공급자

□ 빅데이터 거버넌스 구축사례_통신서비스 공급자 데이터

▷ 빅데이터를 마스터 데이터와 통합하기

⊙ 여러 소스에 흩어져있는 고객 데이터의 매칭

예) 제품 사용과 고장기록 같은 네트워크 데이터는 특정 고객의 상세한 정보만 보강하면 가치있는 통찰력을 얻어낼 수 있음

⊙ T-Mobile에서의 빅데이터

예) T-Mobile은 하루에 통화와 택스트 메시지 기록을 포함하여 170억 개의 이벤트를 처리하는 1.2 페타바이트 규모의 데이터 웨어하우스 구축. 프로젝트 초기 단계에 이 정보를 자사 네트워크 자산의 성능을 개선하는데 이용. 그러나 한 곳에 모여진 이 정보를 재정, 판매, 마케팅 부문 사용자들이 접근하여 고객과 개별적인 상호작용에 활용함

⊙ 네트워크 데이터는 고객 사용과 고장 정보에 관하여 가장 좋은 정보를 제공함. 이 데이터가 전략적인 자산으로 조직의 다른 구성원에게 제공되어 활용하면 고객에 대하여 더 깊은 이해가 가능하게 됨

▷ 빅데이터 개인정보보호

⊙ CSP사들은 자사의 위치 데이터를 제3자에게 팔거나, 새로운 서비스를 개발하는데 활용해서 수익을 창출하는 방법을 적극적으로 알아보고 있음. 그러나 위치 데이터를 사용하는 빅데이터 프로그램은 증가하는 법적 규제, 고객의 프라이버시 관심, 이와 연관된 위험부담 등에 민감해야 함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-21. 통신서비스 공급자

□ 빅데이터 거버넌스 구축사례_통신서비스 공급자 데이터

▷ 빅데이터 품질

⊙ 빅데이터는 데이터 품질관리에 새로운 도전을 제시함. CSP가 외부 데이터에는 영향력이 없지만, 이 데이터의 가치와 품질을 평가해야 함. 내부 데이터와 외부 데이터의 통합은 외부 데이터의 품질을 이해하고 통합된 데이터가 어떻게 사용될 지를 제대로 알아본 후에 신중하게 결정해야 함

예) 제품 사용과 고장기록 같은 네트워크 데이터는 특정 고객의 상세한 정보만 보강하면 가치있는 통찰력을 얻어낼 수 있음

⊙ 사례1. Twitter 데이터의 대표성

특정 CSP사는 국내에서 신제품을 출시하고 제품판매, 장애보고서, 네트워크 사용, Twitter와 관련된 데이터를 수집. 많은 Twitter 사용자들이 제품에 대해 부정적으로 평가하고 있었음. 심도있는 분석을 한 결과 높은 연령층에서는 비교적 제품에 만족했고 피드백을 주기 위해 설문조사나 장애보고서를 이용했는데, 젊은 연령층은 제품에 대해 만족하지 않았고 기존의 설문조사나 장애보고서 같은 피드백을 이용하지 않았음. 소셜 미디어 정보는 자발적으로 올리기 때문에 정보가 편향될 위험이 높음

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저

Module-21. 통신서비스 공급자

□ 빅데이터 거버넌스 구축사례_통신서비스 공급자 데이터

▷ 빅데이터 수명주기 관리

⊙ 빅데이터는 모든 데이터가 저장되어야 한다는 것을 감안할 때 큰 저장 용량을 의미할 수 있음.

⊙ 기존의 데이터 웨어하우징 기술과 비교할 때 CSP는 데이터 수집과 동시에 빅데이터 분석을 수행할 수 있음

⊙ CSP는 1층 스토리지에 표본, 필터, 묶음과 같은 비교적 작은 데이터 집합을 저장할 필요가 있음

⊙ 빅데이터는 2층의 스토리지 환경도 제공하는데, 많은 양의 데이터는 나중에 의미있는 결과를 도출하기 위해 맵리듀스 작업을 거치도록 Hadoop에 저장할 수 있음

⊙ 현재 많은 쿼리 도구들에서 Hadoop에 저장된 빅데이터에 대하여 큰 규모의 쿼리를 수행하는 것이 가능함

참고문헌 : 빅데이터 거버넌스, 홍릉과학출판사, 조완섭, 김상하 외 공저