

빅데이터 활용 단계별 업무절차 및 기술 활용 매뉴얼 (Version 1.0)

－ 빅데이터 수집 · 저장 · 분석 · 이용 단계별 업무절차 및 기술 활용 시 고려사항 제공 －



미래창조과학부

NIA



한국정보화진흥원

KBiG 빅데이터
KOREA BUSINESS DATA CENTER 전략센터

작성 및 문의

미래창조과학부 양현철 사무관
한국정보화진흥원 김진철 수석, 신신애 부장
케이론소프트 김배현 박사
문의 (cckim@nia.or.kr, 02-2131-0216)

작성 지원

넥스트 아이엔아이

자문

김화중 교수(강원대학교), 노병희 교수(아주대학교), 김상재 매니저(SKT), 최재영 교수(성균관대),
이순근 과장(국민건강보험공단), 장진웅 대표(카이언스), 최광선 본부장(솔트룩스), 하정역 팀장(사이람),
최성호 부장(KT넥스알), 민영근 책임(KT넥스알)

머리말

빅데이터가 21세기의 디지털 자원으로 부각되면서 세계 각국은 빅데이터를 국가 정책으로 적극 추진하고 있습니다. 우리나라도 2013년부터 창조경제, 정부3.0 국정과제 기조에 따라 빅데이터가 적극 추진되어 국민생활은 물론 사회, 경제, 산업 활성화 등 각계 분야에서 활용 될 전망입니다.

하지만, 미국, 영국 등 선진국에 비하여 빅데이터 도입이 늦은 우리나라에서는 공공 및 민간에서 빅데이터 활용시 업무에 참고할 지침서가 부족한 상황입니다.

이에, 본 매뉴얼은 빅데이터 활용을 위해 데이터 수집, 저장·관리, 분석, 이용 단계별 업무절차 및 활용 기술, 기술 고려사항 등 빅데이터 실무자를 위한 내용을 담고 있습니다.

공공 및 민간에서 데이터 수집 및 개방, 빅데이터 분석 플랫폼 구축·운영, 빅데이터 분석 및 이용 등 다양한 업무에 활용 되었으면 합니다.

앞으로, 매뉴얼 이용자들의 현장 경험을 수렴하고 내용을 보완하여 빅데이터가 활성화 되는데 좋은 길잡이가 될 수 있도록 개선해 나아가겠습니다.

목 차

머리말	04
-----	----

I 개요

1 필요성	08
2 목적 및 활용범위	09

II 빅데이터 업무절차 및 기술활용 방안

1 데이터 수집	13
2 데이터 저장관리	26
3 데이터 분석	56
4 데이터 제공·이용	70



I

개 요

1_ 필요성

2_ 목적 및 활용범위

1 필요성

● (배경) 빅데이터가 ICT 분야의 새로운 패러다임이자 신 성장동력으로 부상

- 스마트폰 · SNS · 사물인터넷(M2M) 확산 등에 따른 데이터 폭증이 배경

* 산업화 시대 철 · 석탄, 정보화 시대 인터넷처럼 스마트 시대에는 빅데이터가 핵심자원

- ICT 인프라 시장 성숙 이후 신규 비즈니스 영역으로 관심 급증

* Davos 포럼에서는 빅데이터를 '12년도의 가장 중요한 기술로 지목

● (현황) 주요국 및 글로벌 기업은 빅데이터 '산업' 육성 및 '활용'에 주력

- 美는 빅데이터 R&D 이니셔티브('12. 3월), 英은 오픈데이터 전략('12. 3월), 日은 액티브 재팬(ICT 전략('12. 5월) 등을 통해 적극 추진 중

– 글로벌 ICT기업 중심으로 빅데이터 핵심기술 및 신규 비즈니스 모델 개발이 활발하고, 일반기업의 창의적 활용사례도 속속 등장

* 英의 기업실사 감정평가사 Duedil, 美의 부동산 정보서비스사 Zillow 등의 스타트업 등

- 반면, 우리나라는 데이터 생산량이 많은 산업(통신 · 조업 등)이 발달하여 잠재력이 크지만 서비스 도입초기 단계로 빅데이터 활용 노하우가 부족한 상황

* 1인당 모바일 트래픽이 월 963MB로 세계평균(89MB)의 열배인 데이터 생산강국이나(시스코, '11년) 빅데이터 활용은 저조("빅데이터 적극 활용" 응답 경영자는 19.4%(SERI, '12년))

- (필요성) 국내 빅데이터 도입 및 활용 시 참고할 수 있도록 빅데이터 업무절차 및 기술활용 매뉴얼 제작

2 목적 및 활용방법

● 목적

공공 · 민간에서 빅데이터 수집, 관리, 분석 및 데이터 제공 · 이용 등 빅데이터 업무처리 시 지침서로 활용

● 활용범위

- (데이터 수집) 다양한 유형의 빅데이터를 수집하는데 필요한 업무절차 및 기술 도입 시 고려사항 제시
- (데이터 관리) 수집된 데이터를 분석할 수 있도록 가공 · 처리하고 관리하는데 필요한 업무절차 및 적용기술 제시
- (데이터 분석) 빅데이터 분석 플랫폼을 구축하고 데이터를 분석하는데 필요한 기술 및 분석기법 등 제시
- (데이터 제공 · 이용) 빅데이터 분석 플랫폼을 이용하여 서비스를 개발(API 등)하고 분석결과 등을 이용하는데 필요한 절차 및 기술 활용 방안 제시

〈 매뉴얼 세부적 활용 범위 〉

주체		활용 범위
데이터 제공자	공공기관	– 개방 데이터에 대한 범위, 수준, 시기, 방법 등 계획 수립 시 참고
	연구기관	– 개인정보보호, 데이터 보안, 데이터 품질관리 등 정책 수립 및 관리 시 참고
	민간기업	– 데이터 제공에 필요한 업무절차, 기술 고려사항 확인
플랫폼 운영자	자체운영	– 빅데이터 플랫폼 구축을 위한 설계 및 RFP 작성에 활용
	외부제공	– 빅데이터 플랫폼 운영(수집, 저장관리, 분석 · 활용 등) 시 참고 – 빅데이터 서비스 개발 참고
서비스 이용자	개발자·인 청조기업 개인 등	– 빅데이터 플랫폼에 접근하여 제공 데이터 이용 및 신규 서비스 개발 등에 활용

II

빅데이터 업무절차 및 기술활용 방안

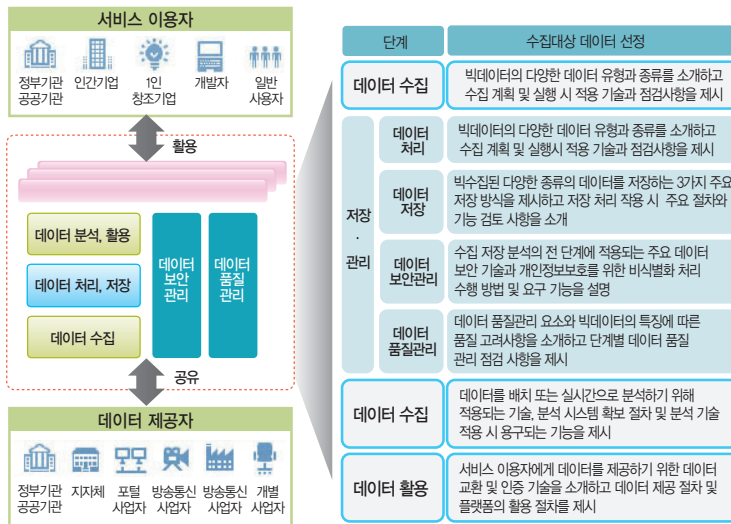
- 1_ 데이터 수집
- 2_ 데이터 저장관리
- 3_ 데이터 분석
- 4_ 데이터 제공 · 이용

매뉴얼 구성 및 주요 내용

본 매뉴얼에서는 빅데이터 수집, 저장·관리, 분석, 이용 등 업무 단계별 개념, 업무절차, 활용 기술, 기술 활용 시 고려사항(관련 Tip 등) 제공

- (개념 정의) 단계별 업무 정의
- (활용 기술) 단계별 업무 처리 시 필요 한 기술 소개
- (업무 절차) 단계별 업무 처리절차 제시
- (기술 고려사항) 기술 활용 시 고려사항 제시

〈 매뉴얼 구성 및 세부 내용 〉



※ 데이터 수명주기 관리는 각 기관의 목적과 운영 지침 등에 따라 관리되어야 함

1 데이터 수집

1.1 개요

정의

조직 내부 및 외부에 있는 다양한 데이터를 일괄(주기) 혹은 실시간으로 수집하는 과정

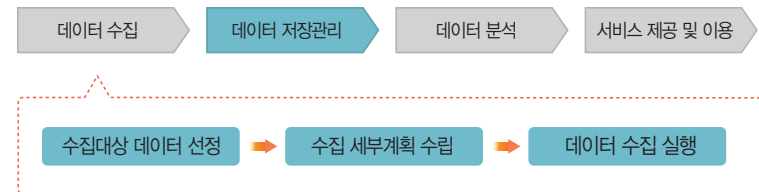
〈 참고 사항 〉

- 데이터 수집활동은 빅데이터 서비스 제공시 서비스 품질을 결정하는 중요한 업무로 수집가능 여부, 개인정보 및 보안, 데이터 품질수준, 수집주기, 비용 등 관련 사항들을 사전에 검토하여 데이터 제공기관과 상호 협의
- 데이터 수집 시에는 데이터 유형에 따라 실시간 크롤링, open API 등 적절한 기술의 선택적 활용이 필요

업무 절차

- 수집 대상 데이터 목록을 선정하고 수집을 위한 세부 수집계획을 작성한 후 업무 진행

〈 데이터 수집 절차 〉



- ① **(수집 대상 데이터 선정)** 분석에 필요한 수집 대상 데이터를 선정하되 수집 가능성 여부 등을 파악하고 세부 목록 및 항목 작성
- ② **(수집 세부계획 작성)** 수집 데이터 유형을 분류하고 관련 수집기술 및 수집주기, 주요 업무 등을 담은 세부 계획 작성
- ③ **(데이터 수집 실행)** 수집계획서에 따라 사전 테스트를 진행하여 관련 시스템을 점검한 후 수집활동 진행

1.2 활용 기술

데이터의 유형 및 특성에 따라서 Crawling, ftp, open API 등 다양한 기술들이 활용중

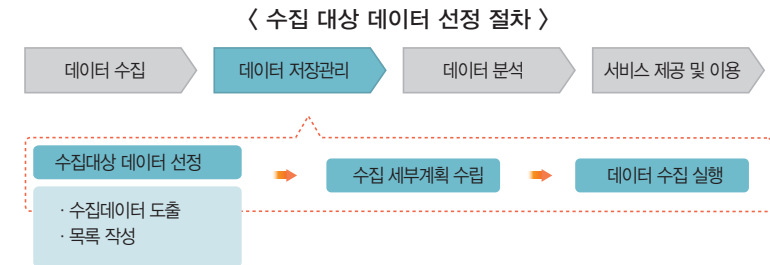
〈 주요 데이터 수집 기술 〉

구 분	특징	비고
Crawling	- SNS, 뉴스, 웹 정보 등 인터넷상에서 제공되는 웹문서 · 정보 수집	웹문서 수집
FTP	- TCP/IP 프로토콜을 활용하는 인터넷 서버로부터 각종 파일들을 송수신 - 보안을 강화하기 위해 SFTP 사용 고려 - 서버간 연동시에는 전용 네트워크 구축 고려	File 수집
Open API	- 서비스, 정보, 데이터 등을 어디서나 쉽게 이용할 수 있도록 개방된 API로 데이터 수집방식 제공 - 다양한 어플리케이션을 개발할 수 있도록 개발자와 사용자에게 공개	실시간 데이터 수집
RSS	- RSS(Really Simple Syndication)는 Web기반 최신의 정보를 공유하기 위한 XML 기반 콘텐츠 배급 프로토콜	콘텐츠 수집
Streaming	- 인터넷에서 음성, 오디오, 비디오 데이터를 실시간으로 수집할 수 있는 기술	실시간 데이터 수집
Log Aggregator	- 웹서버 로그, 웹 로그, 트랜잭션 로그, 클릭 로그, DB의 로그 등 각종 로그 데이터를 수집하는 오픈 소스 기술 - 종류 : Chukwa, Flume, Scribe 등	로그수집
RDB Aggregator	- 관계형 데이터베이스에서 정형 데이터를 수집하여 HDFS(하둡 분산파일시스템)이나 HBase와 같은 NoSQL에 저장하는 오픈 소스 기술 - 종류 : Sqoop, Direct JDBC/ODBC 등	RDB 기반 데이터 수집

1.3 세부 업무처리 절차

수집 대상 데이터 및 범위 선정, 수집 세부계획서 작성, 데이터 수집 진행

수집 대상 데이터 선정



- **(수집 데이터 도출)** 데이터 수집 활동은 빅데이터 서비스 제공 시 서비스 품질을 결정하는 매우 중요한 핵심 업무로,
 - 데이터 도메인의 분석 노하우가 있는 내외부 전문가의 의견을 수렴하여 분석 목적에 맞는 데이터를 도출
- **(목록 작성)** 수집 가능성 여부, 보안 문제, 세부 데이터 항목(품질) 및 비용 등을 검토하여 데이터 수집 목록을 작성
 - **(가능성)** 해당 데이터가 사용 가능하고 수집 가능한가?
 - **(보안)** 수집 시 개인정보 포함여부 및 유출 문제는 없는가?
 - **(정확성)** 활용 목적에 따른 세부 항목들이 적절히 포함되었는가?
 - **(수집비용)** 데이터 수집에 드는 비용은 얼마인가?

● 수집 세부계획 수립



- **(데이터 소유기관 파악 및 협의)** 데이터 소유자의 데이터 개발 현황/조건, 적용 기술, 보안사항 등을 파악하고 필요한 협의 진행

- (내부 데이터) 내부 시스템 간 데이터 연계 가능여부 등 파악
- (외부 데이터) 개방 데이터 종류, 데이터 양, 수집 시스템 연계방식 · 절차, 수집주기 등 관련 기술 · 정책을 파악하고 협의

※ 유의사항 : 데이터 수집관련 보안사항, 개인정보보호 관련 문제 등 점검 필수

- **(데이터 유형 분류 및 확인)** 수집 대상 데이터 유형을 분류하고 데이터 포맷 등 확인

〈 데이터 유형 분류 내용 〉

유형	특징	데이터 종류
정형 데이터 (Structured)	- RDBMS의 고정된 필드에 저장 - 데이터 스키마 지원	RDB, 스프레드 시트
반정형 데이터 (Semi-structured)	- 데이터 속성인 메타데이터를 가지며, 일반적으로 스토리지에 저장되는 데이터 파일 - XML 형태의 데이터로 값과 형식이 다소 일관성이 없음	HTML, XML, JSON, 웹문서, 웹로그, 센서 데이터
비정형 데이터 (Unstructured)	- 언어 분석이 가능한 텍스트 데이터 - 형태와 구조가 복잡한 이미지, 동영상 같은 멀티미디어 데이터	소셜 데이터, 문서, 이미지, 오디오, 비디오

※ 참조 : Gartner, Credit Suisse 데이터 분류체계를 재구성

- **(수집기술 선정)** 데이터 유형 및 포맷 등에 맞는 수집 기술 선정

- 수집기술은 데이터 소스로부터 다양한 유형의 데이터를 수집하기 위해 확장성, 안정성, 실시간성 및 유연성 확보 필요
- 수집기술 선정 시, 1)정제 · 변환 과정, 2)전처리 및 저장 프로세스의 필요성 유무를 우선 점검

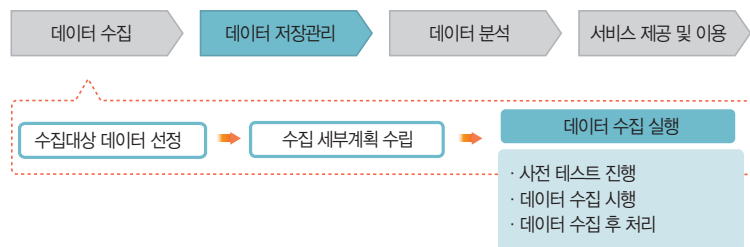
〈 데이터 유형에 따른 수집 기술 〉

데이터 유형	데이터 종류	수집 기술
정형 데이터	RDB, 스프레드 시트	ETL, FTP, Open API
반정형 데이터	HTML, XML, JSON, 웹문서, 웹로그, 센서 데이터	Crawling, RSS, Open API, FTP
비정형 데이터	소셜 데이터, 문서(워드, 한글), 이미지, 오디오, 비디오, IoT	Crawling, RSS, Open API, Streaming, FTP

- **(수집주기 결정)** 데이터 유형에 따라 배치(batch), 실시간 방식 적용
 - **(기본 고려사항)** 수집주기는 데이터 종류와 사이즈, 데이터 발생 빈도 · 주기, 분석주기, 시스템 및 네트워크 부하 수준, 적용 가능 수집기술 등을 고려
 - 분석시점에서 필요한 충분한 데이터양을 확보할 수 있도록 최소한의 수집기간을 설정할 것을 권고
 - 실시간 스트림은 수집과 분석이 동시에 이루어지므로 이벤트 기반 분석 체계 구축이 요구 됨
 - 동적 수집주기에 대한 설정은 사전에 일정 기간 샘플 데이터를 수집한 후 데이터 양 등을 점검한 후 적절하게 주기 조정
- **(수집계획서 작성)** 앞서 소개 된 수집대상 ‘데이터출처, 수집기술, 수집주기 및 수집 담당자의 주요 업무’ 등을 반영하여 계획서를 작성하여 활용

데이터 수집 진행

< 데이터 수집 실행 절차 >



- **(사전 테스트 진행)** 수집계획에 따라서 수집주기, 적용기술 등 관련 수집환경에 대한 사전 테스트 진행
 - **(점검 사항)** 네트워크 트래픽 문제, 데이터 누락여부, 정확성(원본 데이터와 샘플 데이터 비교), 보안성(개인정보 포함여부 등) 등
 - 테스트 수행 결과에 따라서 필요시 수집방법 보완 또는 변경
- **(데이터 수집 시행)** 데이터 수집을 진행하되 향후 장애점검 등을 위하여 관련 로그 기록을 확보할 것을 권고
 - ※ 수집 데이터의 출처, 수집 방식, 장애발생, 로그, 시간 등 수집 당시 상황 등을 시스템적으로 기록
- **(데이터 수집 후 처리)** 데이터 수집 후 저장된 데이터에 대한 외부인의 접근 방지 및 유출 시 대처방안 등 관련 업무지침 마련 필요

Tip

- ① 빅데이터 서비스 활용을 위한 데이터 수집 시 데이터의 질, 수집 기술, 데이터 보안 및 개인정보보호 문제 등 다양한 부분을 고려해야 하므로 가급적 관련 전문가의 조언을 받을 것을 권장합니다.
- ② 또한, 데이터 수집 활동은 분석결과와 질을 좌우하는 중요한 과정으로 분석에 필요한 데이터 항목들이 반드시 포함될 수 있도록 사전 점검하십시오.
- ③ 수집기술은 다양한 데이터 소스로부터 다양한 유형의 데이터를 수집하기 위해 확장성, 안정성, 실시간성 및 유연성 확보가 중요합니다.

1.4 기술 고려사항

Crawling, FTP, Open API, 실시간 Streaming, Log Aggregator, RDB Aggregator 등 기술 활용 고려사항

Crawling 수집기술

- Crawling 활용에 필요한 정보설정, Agent관리, Ranking 등 주요 기능에 대한 기능 고려사항

〈 Crawling 수집 기술 고려사항 〉

기능	고려사항
정보설정 기능	· 수집할 사이트의 URL 목록을 관리(추가/변경/삭제)하는 기능을 제공해야 한다.
	· 수집주기(시간, 분, 초)를 설정하는 기능을 제공해야 한다.
	· URL, 수집주기 등 설정 값을 에이전트에 전달하는 기능을 가져야 한다.
	· 각종 설정이 자동으로 반영되거나 설정을 자유로이 입력 할 수 있는 관리자 기능이 제공 되어야 한다.
	· SNS 수집의 경우 다중 메일 계정을 적용하여 수집 대상을 지정 할 수 있어야 한다.
수집 에이전트 (웹 로봇) 기능	· 에이전트 관리(추가/변경/삭제) 기능을 제공해야 한다.
	· 에이전트 운영(기동, 중지 등) 기능을 제공해야 한다.
	· 수집할 URL 리스트에서 웹문서 등 콘텐츠를 수집하는 기능을 제공해야 한다.
	· 수집한 웹문서에 포함된 URL을 추출하는 기능을 제공해야 한다.
	· 새로운 URL 리스트를 추가하는 기능을 제공해야 한다.
	· 규칙, 확률 또는 학습 기반으로 문서를 분류하는 기능과 분류 승인 모듈 기능을 제공해야 한다.
	· 분류된 정보를 사용자에게 메일링하거나 DB 테이블에 Upload하는 기능을 제공해야 한다.
	· 데이터 수집 시 불필요한 수집이 일어나지 않도록 사전에 대상을 등록하여 선별적 수집이 되도록 하는 기능을 제공해야 한다.
	· 다수의 웹 페이지를 동시에 수집 할 수 있도록 병렬 웹 크롤링 기능을 제공해야 한다.
	· 수집 된 데이터를 로컬 또는 원격에 있는 DB에 저장할 수 있는 기능을 가져야 한다.
Ranking 등 기타 기능	· 로봇이 웹 사이트에 접근하는 것을 방지하기 위한 로봇배제표준 규약 권고안(robots.txt에 기술)을 준수해야 한다.

FTP 수집기술

- 대용량 파일을 수집하기 위한 정보설정, 클라이언트와 서버 간 연결, 파일 전송, FTP 보안 등 기능 고려사항

〈 FTP 수집 기술 고려사항 〉

기능	고려사항
정보설정 기능	· 수집할 대상 파일의 정보(URI, 파일명)를 설정하는 기능을 제공해야 한다.
	· 수집주기(시간, 분, 초)를 설정하는 기능을 제공해야 한다.
	· Active FTP, Passive FTP 연결에 사용한 통신 포트 번호를 설정하는 기능을 제공해야 한다.
클라이언트와 서버 연결 기능	· 클라이언트와 서버간 통신포트를 사용하여 신호, 명령 제어 연결 기능을 제공해야 한다.
	· 클라이언트와 서버간 통신포트를 사용하여 데이터(파일)을 전송 연결 기능을 제공해야 한다.
	· 클라이언트 프로그램은 여러 파일을 연속으로 송수신하기 위하여 서버와의 지속적인 응답 메시지 전송을 통해 연결 상태(세션)을 유지하는 기능을 제공해야 한다.
	· Active FTP(서버가 클라이언트가 지정한 지점으로의 데이터 연결 설정) 연결 기능을 제공해야 한다.
	· Passive FTP(클라이언트가 서버가 지정한 서버 포트로 연결 설정) 연결 기능을 제공해야 한다.
파일 전송 기능	· 클라이언트가 방화벽, NAT(IP 마스킹) 등을 사용하는 환경에서 Active 모드에서 Passive 모드로 전환하는 기능을 제공해야 한다.
	· ASCII파일, EBCDIC파일, 이진파일을 전송하는 기능을 제공해야 한다.
	· 스트림 모드(연속된 바이트의 흐름), 블록 모드(TCP로 블록 전송), 압축 모드 등의 전송 모드를 선택할 수 있는 기능을 제공해야 한다.
	· get, mget(여러 파일 받는 명령), put, mput(여러 파일을 보내는 명령) 등의 명령 기능을 제공해야 한다.
	· 수집파일에 대한 무결성 확인 기능을 제공해야 한다. - 생성중인 파일을 수집기가 수집하거나 수집중인 파일을 분석에 활용할 경우 활용 불가하므로 무결성 제공 필요
FTP 보안 기능	· 사용자 인증(이름, 암호) 기능을 제공해야 한다.
	· Secure FTP(22번 포트 사용)를 사용하여 계정 정보 등을 암호화하는 기능을 제공해야 한다.

Open API 수집기술

- Open API를 이용하여 데이터를 수집하기 위한 정보설정, 수집 Agent, RDB 테이블과의 매핑 등 기술 고려사항

〈 Open API 수집 기술 고려사항 〉

기능	고려사항
정보설정 기능	· 수집할 대상 서버의 정보(URL)를 설정하는 기능을 제공해야 한다.
	· 수집주기(시간, 분, 초)와 반복횟수(무제한, 횟수지정)를 설정하는 기능을 제공해야 한다.
	· URL, 수집주기를 에이전트로 배포하는 기능을 제공해야 한다.
수집 에이전트 기능	· 에이전트 관리(추가/변경/삭제) 기능을 제공해야 한다.
	· 에이전트 운영(기동, 중지 등) 기능을 제공해야 한다.
	· 에이전트가 Mash-up이 용이하도록 RESTful 방식의 Open API를 제공해야 한다.
	· 제공하려는 웹 사이트의 콘텐츠 자원(리소스, text, image, audio, video)에 유일한 URI를 부여하는 기능을 제공해야 한다.
	· POST(create), GET(Read), PUT(update), DELETE(delete) Method를 제공해야 한다.
	· XML, JSON, RSS 정보 제공방식을 지원하여야 한다.
	· 에이전트 통신오류, 이상 동작에 대한 에이전트 감시(monitoring) 및 복구 기능이 제공되어야 한다. (자동/수동)
RDB 테이블과 매핑 기능	· XML, JSON 데이터의 Element와 테이블 Column 정보를 Mapping하는 기능을 제공해야 한다.

M2M Aggregator 수집 기술

- 센싱, 장치로그, 사용자 정보 등 M2M(Machine to Machine) 데이터를 실시간으로 수집하기 위해 필요한 기술 고려사항

〈 M2M Aggregator 수집 기술 고려사항 〉

기능	고려사항
정보설정 기능	· 수집할 대상 서버의 주소(URL/IP)정보를 설정하는 기능을 제공해야 한다.
	· 수집주기(시간, 분, 초)와 반복횟수(무제한, 횟수지정)를 설정하는 기능을 제공해야 한다.
	· 데이터 수집주기를 Agent로 배포하는 기능을 제공해야 한다.
M2M Agent 기능	· 에이전트 관리(추가/변경/삭제) 기능을 제공해야 한다.
	· 에이전트 운영(기동, 중지 등) 기능을 제공해야 한다.
	· 수집 서버의 실시간 콘텐츠 자원(M2M RTDB)에 접속가능한 표준 API 기반의 서비스 기능을 제공해야 한다.
	· M2M 데이터를 일정량으로 분할하여 실시간으로 TCP/IP 프로토콜을 이용하여 다양한 외부 어플리케이션으로 전송하는 기능을 제공해야 한다.
	· 센서 데이터 콘텐츠(M2M RTDB)검색 Method를 제공 해야 한다.
M2M Collector 기능	· 네트워크 트래픽을 최소화하기 위해 최적 라우팅 경로를 탐색하는 기능을 제공해야 한다.
	· Agent 모듈에서 수집한 하위 설비정보 및 센서 데이터를 실시간으로 수집서버로 전송하는 기능을 제공해야 한다



Log Aggregator 수집기술

- Log Aggregator를 이용하여 로그 파일을 수집하기 위한 정보설정, 수집 Agent, Collector 등 기술 고려사항

〈 Log Aggregator 수집기술 고려사항 〉

기능	고려사항
정보설정 기능	· 수신 에이전트 정보(URL, 디렉토리, 파일명명규칙, 확장자 종류)를 설정하는 기능을 제공해야 한다.
	· 수집주기(시간, 분, 초)와 반복횟수(무제한, 횟수지정)를 설정하는 기능을 제공해야 한다.
	· URL, 수집주기를 에이전트로 배포하는 기능을 제공해야 한다.
수집 에이전트 기능	· 에이전트 관리(추가/변경/삭제) 기능을 제공해야 한다.
	· 에이전트 운영(기동, 중지 등) 기능을 제공해야 한다.
	· 파일을 수집하고 청크(chunk, 메타데이터 포함) 단위로 전송하는 기능을 제공해야 한다.
	· 수집 시 사전 정의된 저장 단위로 처리 하여 파일을 생성하거나 일정한 데이터를 토큰단위(주기/블럭)로 잘라서 전송 할 수 있는 기능을 제공해야 한다
	· 압축 가능한 파일의 경우, 압축 전송하는 기능을 제공해야 한다.
	· 수집 대상 파일의 변경(추가/수정/삭제) 여부를 체크하는 기능을 제공해야 한다.
Collector 기능	· 초 단위로 다수의 에이전트로부터 로그 정보를 수신하고 분 단위로 직렬화된 청크에 대한 분산파일시스템 시퀀스 파일로 전송하는 기능을 제공해야 한다.
	· 수집 수행 발생 시 병목을 피하기 위하여 손쉬운 확장기능 제공(수집노드 확장).
	· 수집기의 트래픽 밸런싱을 자동 조정하거나 관리자에 의하여 일정한 형식으로 동작 정의 할 수 있는 기능을 제공 해야 한다.
	· 로그 파일 전송을 모니터링하는 기능을 제공해야 한다.

RDB Aggregator 수집기술

- RDB 데이터를 수집하기 위하여 필요한 정보설정, 수집 Agent에 대한 기술 고려사항

〈 RDB Aggregator 수집 기술 고려사항 〉

기능	고려사항
정보설정 기능	· 수집할 대상 RDB 서버의 정보(예, URL, SID, DB 연동 드라이버)를 설정하는 기능을 제공해야 한다.
	· 수집주기(시간, 분, 초)와 반복횟수(무제한, 횟수지정)를 설정하는 기능을 제공해야 한다.
	· URL, 수집주기를 에이전트로 배포하는 기능을 제공해야 한다.
수집 에이전트 기능	· 에이전트 관리(추가/변경/삭제) 기능을 제공해야 한다.
	· 에이전트 운영(기동, 중지 등) 기능을 제공해야 한다.
	· RDB 메타정보에서 테이블을 선택하는 기능을 제공해야 한다.
	· RDB 메타정보에서 Column, 유형, 크기를 선택하는 기능을 제공해야 한다.
	· RDB 데이터를 레코드 단위로 수집하고 분산파일시스템으로 Import하는 기능을 제공해야 한다.
	· RDB 데이터 수집 중 오류가 있을 경우 오류 경고 처리와 함께 수행 중단 경고를 출력 하는 기능을 제공해야 한다.

참고

데이터 수집 기술 참조 사이트

- Nutch(웹 크롤링/웹 검색) : <http://nutch.apache.org>
- Chukwa(로그수집기) : <http://incubator.apache.org/chukwa>
- Flume(로그수집기 등) : <http://flume.apache.org>
- Sqoop(RDB 데이터 수집) : <http://sqoop.apache.org>

2 데이터 저장관리

데이터 처리(전·후)→저장→보안·품질관리 등 데이터를 안전하게 활용하기 위해 저장 및 관리 업무를 수행하는 과정

〈 참고 사항 〉

- 데이터 분석을 위하여 불필요한 항목을 제거하고 데이터 품질을 향상시킬 수 있도록 데이터 전후처리 수행
- 대용량의 데이터를 유연하게 저장하고 관리 활용할 수 있도록 확장성 등을 고려하여 수집 데이터 유형에 맞는 DB 구축
- 무엇보다도 데이터의 안전한 활용을 위하여 수집된 데이터에 대한 개인정보 처리 및 데이터 접근 및 보안관리 등 철저활용이 필요

2.1 데이터 (전·후)처리

2.1.1 개요

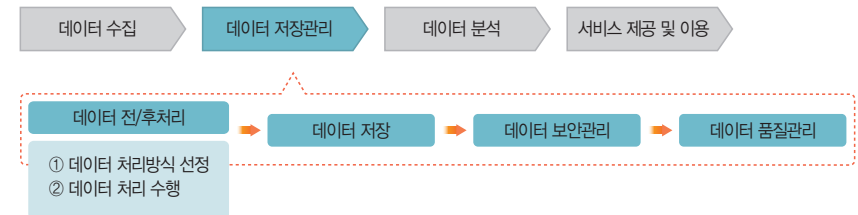
정의

데이터 오류를 필터링하거나 변환, 통합 처리하는 일련의 작업

업무 절차

- 데이터 유형에 따라서 적절한 데이터 처리방식을 선정하여 전후처리 수행

〈 데이터 처리 절차 〉



- ① **(데이터 처리방식 선정)** 데이터 유형과 분석 목적 등을 고려하여 데이터 저장 전·후처리 기법 선정
- ② **(데이터 처리 수행)** 데이터 필터링, 변환, 정제, 통합, 축소 등 선정된 데이터 전후처리 방식에 따라서 데이터 처리



2.1.2 활용 기술

- 데이터 전처리 기술로 필터링, 변환, 정제, 통합 및 축소 등 기술이 활용 중

〈 데이터 처리 기술 소개 〉

방식	설 명
데이터 여과 (Filtering)	- 오류 발견, 보정, 삭제 및 중복성 확인 등의 과정을 통해 데이터 품질을 향상 시키는 기술
데이터 변환 (Transformation)	- 데이터 유형 변환 등 데이터 분석이 용이한 형태로 변환하는 기술 - 정규화(normalization), 집합화(Aggregation), 요약(summarization), 계층 생성 등의 방법 활용 - ETL(extraction/transformation/loading) 도구 제공중
데이터 정제 (Cleansing)	- 결측치들을 채워 넣고, 이상치를 식별 또는 제거하고, 잡음 섞인 데이터를 평활화하여 데이터의 불일치성을 교정하는 기술 ※ 일반적으로 데이터는 불완전하고, 잡음이 섞여있고, 일관성이 없기 때문에 데이터 정제가 필요
데이터 통합 (Integration)	- 데이터 분석이 용이하도록 유사 데이터 및 연계가 필요한 데이터(또는 DB)들을 통합하는 기술
데이터 축소 (Reduction)	- 분석 컴퓨팅 시간을 단축할 수 있도록 데이터 분석에 활용되지 않는 항목 등을 제거하는 기술

2.1.3 세부 업무처리 절차

● 데이터 처리방식 선정

- 우선, 데이터 가공 처리가 필요한 데이터를 분류하여 Table 작성
- 수집된 데이터를 저장하기 위한 전 처리 단계와 저장 된 데이터를 분석하기 전 후 처리하는 단계로 구분
 - (전 처리) 수집한 데이터를 저장소에 적재하기 위한 작업으로 데이터 필터링, 유형 변환, 정제 등 기술 활용
 - (후 처리) 저장된 데이터를 분석이 용이하도록 가공하는 작업으로 변환, 통합, 축소 등 기술 활용

- 데이터 유형과 분석 목적 등을 검토하여 전·후 처리 기술 선택

- 분석에 소요되는 시간과 노력을 절약할 수 있도록 일관성 있는 데이터 형태로 통합
- 분석 효율을 높일 수 있도록 데이터로부터 의미 있는 정보만 추출
- 의미 파악이 어려운 비정형 데이터는 분석이 가능한 형태로 변환

● 데이터 처리 수행

- 데이터 필터링

- 데이터 중복성, 오류 제거들을 위한 데이터 필터링 기준 설정
- 실제 사전 테스트를 통하여 오류 발견, 보정, 삭제 및 중복성 검사 등 필터링 과정을 거쳐 필터링 기준을 최적화 하여 활용
- 비정형 데이터는 데이터 마이닝을 통해 오류, 중복, 저품질 데이터를 처리할 수 있도록 자연어처리 및 기계학습과 같은 추가기술 적용 필요
- 분석을 위하여 단위 저장소에 파일형태로 저장 할 경우, 데이터 활용 목적에 맞지 않는 정보는 필터링하여 제거해야 분석시간을 단축하고 저장 공간의 효율적 활용 가능

● 데이터 변환

- 다양한 형식으로 수집 된 데이터를 분석에 용이하도록 일관성 있는 형식으로 변환
 - 평활화(Smoothing) : 데이터로부터 잡음을 제거하기 위해 데이터 추세에 벗어나는 값들을 변환
 - 집계(Aggregation) : 다양한 차원의 방법으로 데이터를 요약
 - 일반화(Generalization) : 특정 구간에 분포하는 값으로 스케일을 변화
 - 정규화(Normalization) : 데이터에 대한 최소-최대 정규화, z-스코어 정규화, 소수 스케일링 등 통계적 기법 적용
 - 속성 생성(Attribute/feature construction) : 데이터 통합을 위해 새로운 속성이나 특징을 만드는 방법으로 주어진 여러 데이터 분포를 대표할 수 있는 새로운 속성/특징 활용

● 데이터 정제

- 수집된 데이터의 불일치성을 교정하기 위한 방식으로 결측치(Missing Value) 처리, 잡음(Noise) 처리 기술 활용

〈 결측치(Missing Value) 처리 방법 〉

방법	설명
해당 레코드 무시	- 분류에서 클래스 구분 라벨이 빠진 경우 레코드 무시 - 결측치가 자주 발생하는 환경에서는 적용시 비효율적
자동으로 채우기	- 결측치에 대한 값을 별도로 정의(예) "unknown" - 통계값 적용 : 전체 평균값, 중앙값, 해당 레코드와 같은 클래스에 속한 데이터의 평균값 - 추정치 적용 : 베이지안 확률 추론, 결정 트리
담당자(전문가)가 수작업 입력	- 담당자가 직접 확인하고 적절한 값으로 수정 - 신뢰성은 높을 수 있으나 많은 작업 시간이 소요 됨

〈 잡음(Noisy) 처리 방법 〉

방법	설명
구간화(Bining)	- 정렬한 데이터를 여러 개의 구간으로 배분한 후 구간 안에 있는 값들을 대표값으로 대체 - 구간 단위별로 잡음 제거 및 데이터 축약 효과 - 사용되는 대표값 : 평균, Median 등
회귀값 적용 (Regression)	- 데이터를 가장 잘 표현하는 추세 함수를 찾아서 이 함수의 값을 사용
군집화 (clustering)	- 비슷한 성격을 가진 클러스터 단위로 묶은 다음 outlier 제거

※ 잡음 : 랜덤 에러나 측정된 변수의 변형된 값

※ 잡음 발생원인 : 센서의 작동 실패, 데이터 엔트리(기입, 표기) 문제, 데이터 전송 문제, 기술적인 한계, 데이터 속성 값의 부정확성 등

● 데이터 통합

- 출처가 다른 상호 연관성이 있는 데이터들을 하나로 결합하는 기술로 다음 사항들을 고려
 - 데이터 통합 시 동일한 데이터가 입력 될 수 있으므로 연관관계 분석 등을 통해 중복 데이터를 검출 할 것
 - 데이터 통합 전 · 후 수치 · 통계 등 데이터 값들이 일치 할 수 있도록 검증
 - 통합 대상 entity가 통합 이후에 동일한지 여부를 확인하기 위한 동일성 검사 수행
 - 표현 단위(파운드와 kg, inch와 cm, 시간 등) 등 서로 다른 방식에 대해 표현을 일치할 수 있도록 변환

● 데이터 축소

- 분석에 불필요한 데이터를 축소하여 고유한 특성은 손상되지 않도록 하고 분석에 대한 효율성을 증대

〈 데이터 축소 방식 예시 〉

축소 방식	설명
차원 축소	분석에 필요 없거나 중복 항목 제거 - Stepwise forward selection, Stepwise backward elimination 등 활용
데이터 압축	데이터 인코딩이나 변환을 통해 데이터 축소 - lossless(BMP 포맷), lossy(JPEG 포맷) 등 방법 적용
Discrete wavelet transform (DWT)	선형 신호 처리 - 수는 다르지만 길이는 같은 벡터(wavelet coefficients)로 변환 - 여러 개의 벡터 중에서 가장 영향력이 큰 벡터를 선택하여 다른 벡터들을 제거
Principal components analysis (PCA)	데이터를 가장 잘 표현하고 있는 직교상의 데이터 벡터들을 찾아서 압축 - 속성들을 선택하고 다시 조합시켜 다른 작은 집합으로 생성 - 계산하는 과정이 간단하고 정렬되지 않은 속성들도 처리 가능
수량 축소 (Numerosity Reduction)	데이터를 더 작은 형태로 표현해서 데이터의 크기 줄임 - 데이터 파라미터만 저장(예, Log-linear 모델) - 기존의 데이터에서 축소된 데이터를 저장 (예, 히스토그램, 클러스터링, 샘플링 등)

2.1.4 기술 고려사항

데이터 전·후 처리를 위한 데이터 필터링, 변환, 정제, 통합, 변환, 축소 등 기술 활용 시 고려사항

● 데이터 전처리

- 데이터 전처리에 활용되는 데이터 필터링, 유형변환 및 정제 기술에 대한 고려사항

〈 데이터 전처리 기술 고려사항 〉

기능	고려사항
데이터 필터링 (Filtering)	· 데이터 필터링 기준을 정의하고 설정 할 수 있는 기능을 제공해야 한다.
	· 데이터 처리 전후에서 생성된 파일의 중복성을 확인 할 수 있도록 파일명, 확장자 등 필터링 기능을 제공해야 한다.
	· 유의미한 데이터를 선별하기 위하여 사전 정의된 필터링 기준을 비교 검증 할 수 있는 기능이 제공되어야 한다.
	· 데이터 필터링 적용시, 비정형 데이터 처리에서 자연어처리 및 기계학습을 수행하기 전에 사용자가 처리 방식을 선택 할 수 있도록 데이터 파일에 대한 정형화된 사전 저장 기준을 제공 하여야 한다.
	· 수집된 데이터의 품질 기준의 부합 여부 및 오류 등을 확인하고 관리자에게 알릴 수 있는 기능을 구현해야 한다.
데이터 유형 변환 (Transformation)	· 필터링 처리 시 사전 정의된 필터링 기준에 의거하여 데이터 처리에서 오류 발생 후 오류에 대한 이력을 저장 할 수 있는 기능을 제공해야 한다.
	· 수집된 데이터의 유형을 분류 할 경우 분류 기준을 적용 할 수 있는 기능을 제공해야 한다.
	· 데이터의 유형을 분류하고 이에 대한 데이터 변환에 필요한 알고리즘 함수 또는 변환 구조를 정의 할 수 있는 기능이 제공 되어야 한다.
	· 데이터 변환 시 사용자가 지정한 변환 형식에 준하여 변환이 이루어졌는지 확인 할 수 있는 기능이 제공되어야 한다.
	· 데이터 변환 실패 시 데이터 변환 실패 부분에 대하여 재시도 할 수 있는 기능을 제공 하거나 신규 변환 데이터가 생성을 취소 할 수 있는 기능을 제공해야 한다.
	· 데이터 변환이 실패 되었을 경우 이력을 저장하고 사용자에게 전달할 수 있는 기능이 제공되어야 한다.
데이터 정제 (Cleansing)	· 변환된 데이터를 저장 하는 기능을 제공해야 한다.
	· 정제 유형을 사전정의하고 속성 값을 부여하는 기능 및 사용자가 스크립트를 작성 할 수 있는 기능이 제공되어야 한다.
	· 데이터 유형별 정제 시 사용자가 설정한 정제 방법을 사전 정의되어 자동으로 지정 할 수 있는 기능이 제공 되어야 한다.
	· 결측치, 잡음 데이터를 처리하는 경우, 데이터 저장 및 제거 대상에 대하여 삭제, 처리, 확인 할 수 있는 기능이 제공 되어야 한다.
	· 데이터의 불일치성을 교정하기 위하여 단위, 표현형식, 코드체계 등의 불일치성을 교정하거나 자동으로 교정이 되도록 하는 자동 스크립팅 기능이 제공 되어야 한다.

데이터 후처리

- 데이터 후처리를 위하여 필요한 데이터 통합, 변환, 축소 기술 고려사항

〈 데이터 후처리 기술 고려사항 〉

기능	고려사항
데이터 통합 (Integration)	· 데이터의 일관성을 위해 여러 출처(소스)로부터의 데이터들을 결합할 수 있도록 사전에 확인 할 수 있는 기능을 제공해야 한다.
	· 데이터 통합을 위하여 취합된 정보에 대한 상호 관계를 비교 하거나 정보 결합 속성 등의 요건을 체크하는 기능이 제공 되어야 한다.
	· 데이터 통합 시 통합전후의 원시 데이터의 백업을 지원하고 이력을 확인 할 수 있는 기능이 제공 되어야 한다.
	· 데이터 통합을 위해 유일한 키 값을 선정하거나 자동 키(Key) 부여 및 키값(Key Value) 관리 기능이 제공 되어야 한다.
데이터 변환 (Transformation)	· 데이터로부터 잡음을 제거하기 위해 데이터 추세에 벗어나는 데이터(이상치(Outlier) 또는 특이값)를 추세에 맞게 변환 또는 자동 추천 할 수 있는 기능을 제공 하여야 한다.
	· 집계(Aggregation) 시 데이터를 요약 하는 기능이 제공 되어야 한다.
	· 특정 구간에 분포하는 값을 추출 하거나 이를 사용자가 직관적으로 확인 할 수 있도록 하여 데이터 변환 시 발생 할 수 있는 변환, 패턴, 이벤트를 감시 할 수 있는 기능을 제공해야 한다.
	· 데이터 변환 후 사전 저장된 원시 데이터 셋과 변환 후 데이터 간의 변환 로그를 저장 관리 할 수 있는 기능이 제공 되어야 한다.
데이터 축소 (Reduction)	· 데이터 축소를 위한 적용 기준 또는 적용 스크립트를 부여 할 수 있는 기능이 제공되어야 한다.
	· 데이터 크기를 축소하는 경우, 원본 파일의 데이터 축소 범위와 축소가 적용된 속성에 대한 로그를 기록하여 취소 시 재 복구 할 수 있도록 하는 기능이 제공 되어야 한다.

참고

데이터 처리 기술 참조 사이트

- Kafka
 - <http://sites.computer.org/debull/A12june/pipeline.pdf>
 - <https://cwiki.apache.org/confluence/display/KAFKA/Ecosystem>
- Cascading
 - <http://docs.cascading.org/cascading/2.5/userguide/html/ch03s03.html>
- Summingbird
 - <https://blog.twitter.com/2013/streaming-mapreduce-with-summingbird>
- UDF
 - <http://www.slideshare.net/OReillyStrata/large-scale-etl-with-hadoop>
- Tika
 - <http://www.slideshare.net/jukka/content-extraction-with-apache-tika>

2.2 데이터 저장

2.2.1 개요

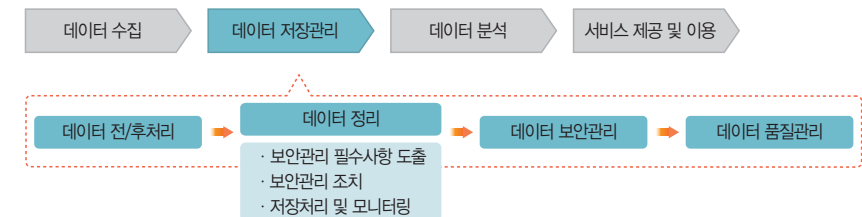
정의

운영자가 수집 · 처리된 데이터를 분석에 활용 할 수 있도록 적합한 방식으로 저장 보관하는 작업

업무 절차

- 데이터 유형에 따라 저장계획을 수립하고 적합한 DB를 구축한 후 데이터 저장 · 관리

〈 데이터 저장처리 절차 〉



- ① (저장 계획 수립) 데이터 유형을 검토하여 저장방식을 선정하고 실행에 필요한 데이터 저장 계획 수립
- ② (DB 구축 및 테스트 수행) 선정된 저장방식에 따라 적합한 DB를 구축한 후 사전 테스트 수행
- ③ (저장처리 및 모니터링) 구축된 DB에 데이터를 저장하고 용량한계 등 수시 모니터링



2.2.2 활용 기술

데이터 유형에 따라서 RDB, NoSQL, 분산파일시스템 등 기술 활용 중

〈 데이터 저장 방식 소개 〉

구분	특징	비고
RDB	- 관계형 데이터를 저장하거나, 수정하고 관리할 수 있게 해주는 데이터베이스 - SQL 문장을 통하여 데이터베이스의 생성, 수정 및 검색 등 서비스를 제공	oracle, mssql, mysql, sybase, MPP DB
NoSQL	- Not-Only SQL의 약자이며, 비관계형 데이터 저장소로, 기존의 전통적인 방식의 관계형 데이터베이스와는 다르게 설계된 데이터베이스 - 테이블 스키마(Table Schema)가 고정되지 않고, 테이블 간 조인(Join) 연산을 지원하지 않으며, 수평적 확장(Horizontal Scalability)이 용이 - key-value, Document key-value, column 기반의 NoSQL이 주로 활용 중	MongoDB, Cassandra, HBase, Redis
분산파일시스템	- 분산된 서버의 로컬 디스크에 파일을 저장하고 파일의 읽기, 쓰기 등과 같은 연산을 운영체제가 아닌 API를 제공하여 처리하는 파일시스템 - 파일 읽기/쓰기 같은 단순 연산을 지원하는 대규모 데이터 저장소 - 범용 x86서버의 CPU, RAM 등을 사용하므로 장비 증가에 따른 성능 향상 용이 - 수 TB~ 수백 PB 이상의 데이터 저장 지원 용이	HDFS

※ MPP(Massively Parallel Processing) DB : 대용량병렬처리 데이터베이스

2.2.3 세부 업무처리 절차

데이터 저장 계획 수립

- (데이터 유형 검토) 저장 할 데이터의 포맷 등 유형을 검토하고 데이터 저장 관리에 유리한 저장방식 선정

– RDB 저장

- RDB 테이블 데이터는 컬럼과 값을 코드 매핑하거나 데이터형을 변환 처리하여 테이블 형태로 저장
- XML, JSON, HTML 등 형식의 파일은 파싱 처리하여 테이블에 저장
- 문서, 이미지, 비디오, 오디오 등 이진파일은 key값을 추출한 후 테이블에 저장

– NoSQL 저장

- 정형데이터(RDB 저장 데이터)는 컬럼과 값을 key와 value로 구분하여 저장
- XML, JSON, HTML 등 형식의 파일은 파싱하여 key-value로 저장 (NoSQL에서 지원하는 데이터 타입으로 변환 저장)

※ 문서, 이미지, 비디오, 오디오 파일의 저장은 별도 처리 방안 필요

– 분산파일시스템 저장

- 문서(XML · JSON · HTML, 텍스트 등), 이미지, 비디오, 오디오 등 텍스트 및 이진파일을 분산파일시스템에서 지원하는 파일형식으로 저장

- **(저장 공간 용량 설계)** 수집 할 데이터 크기 및 최대 저장기간 등을 고려하여 용량 설계

- RDB는 레코드 최대 크기 등을 고려하여 저장 공간 용량 설정
 - RDB는 솔루션별 가용 공간이 다르므로 “솔루션 업체별 설계 가이드” 등을 참고하고 scale-up, scale-out 방식으로 최대 TeraByte씩 확장 가능
- NoSQL과 분산파일시스템은 보편적으로 3개의 동일한 복제파일을 운영해야 하므로 이를 감안하여 저장 공간 설정
 - NoSQL 및 분산파일시스템은 scale out 방식으로 Peta Byte 이상 확장 가능

〈 확장 기술 비교 〉

구분	Scale up	Scale out
개요	CPU, 메모리, 하드디스크 등 서버 자원을 추가하여 처리 능력을 향상시키는 방식	서버의 대수(노드)를 추가하여 처리 능력을 향상시키는 방식
비용	컨트롤러나 네트워크 인프라 비용은 발생하지 않고 디스크만 추가	추가된 노드들이 하나의 시스템으로 운영되기 위한 NW장비 필요
용량	하나의 스토리지 컨트롤러가 지원 가능한 Device 수가 한정되어 있어 용량확장 시 제약	스토리지 용량 확장성이 매우 좋음

- **(계획서 작성)** 데이터 유형에 따른 수집주기, 저장방식, 보관주기 및 백업방식, 저장 공간 확대방안 등 업무처리 내용을 포함한 저장계획서 작성

DB 구축 및 테스트 수행

- 데이터 저장계획서에 따라서 확장성 등을 고려하여 DB를 구축하고 운영에 필요한 주요기능에 대한 사전테스트 진행

저장처리 및 모니터링 수행

- 시 · 일 · 주 · 월 주기적으로 데이터 운영관련 에러 및 여유공간 등을 실시간 모니터링하고 문제 발생 시 신속한 대응체계 마련
 - 오류가 발생하였을 경우, 내역을 분석하여 수집 · 처리 · 저장 단계 관련 담당자와 협의하여 해결

※ 클라우드 및 IDC 자원을 활용할 경우, 저장 공간 사용관련 SLA Level에 따라 운영 및 확장

Tip 데이터 저장에서의 주요 사항

- ① 빅데이터 저장 환경에서는 적절한 저장 공간의 확보가 중요하며 또한 주기적인 모니터링을 수행해야 함
 - RDB는 인덱스 및 여유 공간을 계산하여 확보
 - NoSQL, 분산파일시스템은 복제파일 운영 여유 공간을 계산하여 확보
- ② 데이터 수집 및 저장이 장기간, 실시간, 대용량으로 수행될 경우 관련 사전 점검이 선행되어야 하고 일정 저장 공간 이상 사용되었을 때 scale-out 될 수 있는 방안이 마련되어야 함

2.2.4 기술 고려사항

데이터 저장을 위한 RDB, NoSQL, 분산파일시스템 저장방식에 대한 기술 고려사항

RDB 저장방식

- 데이터를 RDB 테이블에 컬럼으로 저장하기 위한 기술 고려사항

〈 RDB 테이블 저장 기술 고려사항 〉

기능	고려사항
Text 데이터 저장	· 저장 테이블과 컬럼 관계를 매핑하여 저장하는 기능을 제공하여야 한다.
	· 데이터 형이 다른 컬럼은 데이터 형을 변환하여 저장하는 기능을 제공하여야 한다.
	· 코드가 다른 컬럼은 코드 변환하여 저장하는 기능을 제공하여야 한다.
	· 컬럼 중 NULL 데이터가 있는 경우 NULL 변환 처리하여 저장하는 기능을 제공하여야 한다.
	· 수집한 웹문서 내용으로부터 스키마에 저장된 각 속성들의 값을 추출한 후 DB 로딩 형식에 맞추어 변환하는 기능을 제공해야 한다.
XML, JSON, HTML 데이터 저장	· 저장 테이블의 컬럼에 맞도록 데이터를 파싱 처리하여 저장하는 기능을 제공하여야 한다.
	· 저장 시 자동 집계 또는 후처리 집계 처리가 수행 될 수 있도록 기능을 제공해야 한다.
	· 저장 테이블의 문자열 데이터형에 데이터 전체를 저장하는 기능을 제공하여야 한다.
문서, 이미지, 오디오, 비디오 데이터 저장	· 이진 데이터는 key값을 추출하여 저장하는 기능을 제공하여야 한다.
	· 저장 테이블의 이진 데이터형에 데이터 전체를 저장하는 기능을 제공하여야 한다.
RDB 성능관리	· 서버중설, 튜닝, 데이터 관리주기를 적용하여야 한다.
	· RDB 연계에 따라 분산된 저장소간의 저장 상황 및 자원 활용 상황을 모니터링 할 수 있는 기능이 제공되어야 한다.
	· 검색 속도를 높이기 위한 다양한 색인 기능을 제공하여야 한다.

NoSQL 저장방식

- key/value, 컬럼, document 등 특성에 따라서 NoSQL에 저장하기 위해 필요한 기술 고려사항

〈 NoSQL 저장 기술 고려사항 〉

기능	고려사항
RDB 테이블 데이터를 NoSQL에 저장	· key 값을 추출하고 컬럼 관계를 매핑하여 저장하는 기능을 제공해야 한다.(컬럼기반)
	· 테이블 데이터를 document 기반으로 변환하여 저장하는 기능을 제공해야 한다.(document 기반)
	· key 값을 추출하고 전체 데이터는 value에 저장하는 기능을 제공하여야 한다.(key/value)
	· Key 값과 추출값 간의 릴레이션을 파악할 수 있는 변환용 검증기를 부여 할 수 있는 기능이 제공 되어야 한다.
XML,JSON,HTML 문자열/문서파일을 NoSQL에 저장	· key 값을 추출하고 컬럼 관계를 매핑하여 저장하는 기능을 제공하여야 한다.(컬럼기반)
	· 데이터 내용을 document 기반으로 변환하여 저장하는 기능을 제공하여야 한다.(document 기반)
	· key 값을 추출하고 데이터 내용은 value에 저장하는 기능을 제공하여야 한다.(key/value)
NoSQL 성능관리	· 저장소 및 성능 향상을 위한 scale out 기능을 제공하여야 한다.
	· 데이터 처리 성능을 높이기 위한 분산병렬처리 기능을 제공하여야 한다.

분산파일시스템 저장방식

- 분산파일시스템에서 제공하는 파일 형식으로 저장하기 위한 기술 고려사항

〈 분산파일시스템 저장 기술 고려사항 〉

기능	고려사항
RDB 테이블 데이터를 분산파일시스템에 저장	· 저장 테이블과 컬럼 관계를 매핑하여 저장하는 기능을 제공하여야 한다.
	· 데이터 형이 다른 컬럼은 데이터 형을 변환하여 저장하는 기능을 제공하여야 한다.
	· 코드가 다른 컬럼은 코드 변환하여 저장하는 기능을 제공하여야 한다.
	· 컬럼 중 NULL 데이터가 있는 경우 NULL 변환 처리하여 저장하는 기능을 제공하여야 한다.
	· 맵리듀스 작업을 통해 중복 체크 후 분산파일시스템에 저장하는 기능을 제공해야 한다.
XML, JSON, HTML 문자열/문서파일을 분산파일시스템에 저장	· 저장 테이블의 컬럼에 맞도록 데이터를 파싱 처리하여 저장하는 기능을 제공하여야 한다.
	· 저장 테이블의 문자열 데이터형에 데이터 전체를 저장하는 기능을 제공하여야 한다. ※문자열 데이터형 참조
	· 여러 웹로부터 수집한 파일들을 분산파일시스템에서 제공하는 저장파일로 통합하는 기능을 제공해야 한다.
	· 분산파일시스템의 저장단위로 통합된 파일을 나누어 저장하는 기능을 제공해야 한다.
	· 나누어 저장된 통합 파일을 분산파일시스템에 중복되지 않도록 순차처리 파일로 저장하는 기능을 제공해야 한다.
	· 여러 에이전트로부터 수집한 로그파일들을 DFS(분산파일시스템)에서 제공하는 저장파일로 통합하는 기능을 제공해야 한다.
	· 매핑과 변환 처리된 데이터를 맵리듀스 작업에서 중복 제거 후 분산파일시스템에 저장하는 기능을 제공해야 한다.
문서, 이미지, 오디오, 비디오 이진파일을 분산파일시스템에 저장	· 이진 데이터를 key값을 추출하여 저장하는 기능을 제공하여야 한다.
	· DFS(분산파일시스템)의 이진 데이터형에 데이터 전체를 저장하는 기능을 제공하여야 한다.
	· 여러 에이전트로부터 수집한 파일들을 DFS(분산파일시스템)에서 제공하는 저장파일로 통합하는 기능을 제공해야 한다.
	· DFS(분산파일시스템) 저장단위로 통합된 파일을 나누어 저장하는 기능을 제공해야 한다.
	· 나누어 저장된 통합 파일을 DFS(분산파일시스템)에 중복되지 않도록 순차처리 파일로 저장하는 기능을 제공해야 한다.
분산파일시스템 성능관리	· 저장소 성능 향상을 위한 scale out 기능을 제공하여야 한다.
	· 데이터 처리 성능을 높이기 위한 분산병렬 처리 기능을 제공하여야 한다.

참고

데이터 저장 기술 참조 사이트

• RDB 제품

- Oracle : 분산처리 지원 데이터베이스
 - <https://www.oracle.com/>
- MS SQL : 윈도우 환경을 제공하는 데이터베이스
 - <https://www.microsoft.com/en-us/sqlserver/default.aspx>
- My SQL : 보편적으로 사용중인 데이터베이스
 - <http://www.mysql.com/>
- Sybase : 엔터프라이즈 정보시스템 구축을 위한 client/Server 환경에서 적용되는 데이터베이스
 - <http://www.sybase.com/>

• No SQL 제품

- Redis : key/value 기반
 - <http://redis.io/>
- Apache HBase : 컬럼 기반
 - <http://hbase.apache.org/>
- Apache Cassandra : 컬럼 기반
 - <http://cassandra.apache.org/>
- MongoDB : document 기반
 - <http://www.mongodb.org/>
- Couchbase
 - <http://www.couchbase.com/>

• 분산파일시스템 제품

- Hadoop HDFS
 - <http://hadoop.apache.org/>

2.3 데이터 보안관리

2.3.1 개요

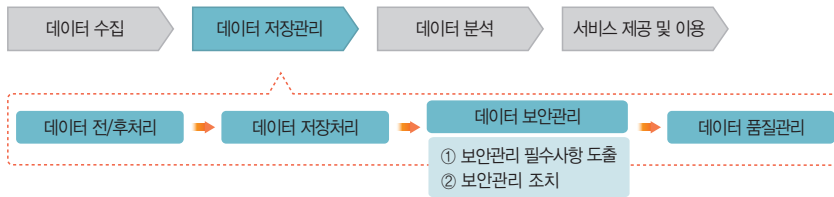
정의

데이터의 유출을 방지하고 안전한 활용을 위한 개인정보 처리 등 관련 조치하는 작업

업무 절차

- 데이터 활용 관련 보안관리 필수사항을 도출하고 관련 기준에 따라 보안 관리 조치

〈 데이터 보안관리 업무 절차 〉



- ① (보안관리 필수사항 도출) 데이터 수집, 저장처리, 분석활용 단계에서 발생 가능한 보안 침해 가능성 및 개인정보처리 관련 사항 검토
- ② (보안관리 조치) 관련 법제도 및 지침서 등을 활용하여 데이터 보안관리 조치

2.3.2 활용 기술

- 데이터 보안을 위하여 사용자 인증, 접근 제어, 데이터 암호화, 개인정보 비식별화 및 개인정보 암호화 등 다양한 기술들이 활용중

〈 데이터 보안 적용 기술 설명 〉

구분	설명
사용자 인증	- 누가 어떤 데이터에 어떤 조치를 취할 수 있는가를 미리 정한 바에 따라, 데이터나 데이터 관리 시스템에 접근하는 사람의 접근 자격을 확인하는 것 - 세부 기술 : ID/Password 방식 / 일회용 패스워드(OTP) 방식, 전자인증, 통합사용자 인증
접근 제어	- 어떤 주체가 어떤 객체를 읽고자 하거나, 객체에 기록하고자 하거나, 객체를 실행(객체가 실행 파일일 경우)시키고자 할 때 마다 그 주체가 그 객체에 대한 권한을 가지고 있는지를 체크하고 통제하는 것 - 세부 기술 : 강제 접근제어, 임의 접근제어, 역할기반 접근제어
암호화	- 평문을 해독 불가능한 형태로 변형하거나 또는 암호화된 암호문을 해독 가능한 형태로 변형하기 위한 원리, 수단, 방법 등을 취급하는 기술 - 암호화는 복호화 가능 여부에 따라 양방향 암호화와 단방향 암호화로 분류되며, 양방향 암호화는 암호·복호화 시 사용되는 키의 대칭여부에 따라 대칭키 암호화와 비대칭키 암호화로 분류 - 암호화 알고리즘 : DES, AES, RSA, MD5, SHA 등
개인정보 비식별화	- 수집된 데이터에 포함된 개인정보의 일부 또는 전부를 삭제, 다른 정보로 대체 또는 다른 정보와 결합(Mash-up)하여도 특정 개인을 식별하기 어렵도록 하는 일련의 조치
개인정보 암호화	- 데이터베이스 전체에 대한 보호가 아닌 개인정보가 포함된 특정 필드에 대한 보호기술 - 개인정보의 암호화 저장 후 데이터베이스에 저장된 개인정보의 정상적인 이용을 위해 데이터베이스를 안전하고 효율적으로 인덱싱하는 기술

2.3.3 업무 처리 절차

● 보안관리 필수 사항 도출

- 데이터를 수집하여 활용하는 단계에서 발생할 수 있는 데이터 보안 침해 가능성을 검토하고,
 - 이에 대한 예방과 문제 발생 시 신속한 대처를 위한 데이터 보안 정책 및 적용 기술 등 기능요건 도출

〈 데이터 보안 기능 요건 및 세부 기술과 적용 단계 〉

기능 요건		단계		
분류	세부기술	수집	저장관리	분석활용
사용자 인증	지식 기반 인증 시스템		◎	◎
	전자인증	◎	◎	◎
	통합사용자인증	◎	◎	◎
접근 제어	강제 접근제어		◎	◎
	임의 접근제어		◎	◎
	역할기반 접근제어		◎	◎
암호화	PKI(공개키 기반구조)	◎		◎
	PMI(권한관리기반구조)	◎		◎
	암호화 알고리즘	◎	◎	◎
비식별화	개인정보 비식별화		◎	◎
	개인정보 암호화		◎	
	사후 모니터링	◎	◎	◎

● (수집 단계) 수집되는 데이터에 대한 사전 동의사항 확인

– 데이터를 생성하는 주체로부터 사전 동의 받은 데이터만 수집

※ SNS, 웹 정보 등 자동화된 시스템(크롤러)을 통하여 데이터를 수집하는 경우 데이터 소유자에게 수집 관련 동의를 별도로 받기 어려운 상황으로 소유자가 데이터의 공개/비공개 여부를 설정하였을 경우 공개 설정된 데이터에 한하여 수집

– 수집되는 데이터에 대한 ‘접근 통제’ 기능 적용

- 내부 시스템 로그와 같이 외부 유출을 차단해야 하는 데이터들은 해당 담당자만 접근할 수 있도록 통제
- 특히, 웹 로봇과 같은 자동 수집기로부터 데이터 유출을 방지하고자 하는 경우 웹 로봇이 데이터를 수집하는 행위를 원천적으로 차단할 수 있도록 해당 기술 적용 필요

● (저장 단계) 데이터의 안전한 저장 및 관리 방안 마련

– 저장되는 데이터가 외부의 시스템 침입 등에 의하여 불법적으로 유출되지 않도록 조치

※ 데이터 저장 시스템에 대한 "접근권한"을 설정하고 필요한 항목에 대하여 "암호화" 조치 수행

– 개인정보의 포함 여부를 파악하여 비식별화 등 필요한 조치 수행

※ 일정 부분의 데이터를 마스킹 처리하여 식별 불가능하도록 저장하는 등 "개인정보에 대한 비식별화" 처리 후 데이터 저장

● (분석·활용 단계) 원칙적으로 개인정보 관련 데이터는 비식별화 등 보안조치한 후 분석에 활용

● (폐기) 수집된 데이터는 목적에 맞게 안전하게 활용한 후 폐기 조치할 것을 권장

● 보안관리 조치

- (접근통제, 차단, 인증, 암호화 등) ‘개인정보 암호화 조치 안내서 1.0(행정안전부/KISA, 2012.10)’ 등 참조하여 접근통제, 차단, 인증, 암호화 등 필요한 조치 수행
- (개인정보 처리) 개인정보보호법, 안행부 ‘공공정보 개방·공유에 따른 개인정보 보호 지침(13.9)’ 등을 참고하여 처리
 - (비식별화 방안 수립) 개인정보보호법 이행을 준용하여 개인정보 비식별화 원칙 수립

데이터 유형별 비식별화 조치 방안 (예시)

- ① 개인정보보호법 및 공공정보 개방 공유에 따른 개인정보보호 지침 등을 활용하여 데이터 유형별로 비식별화가 필요한 키워드 생성
 - 그 자체로 개인을 식별할 수 있는 정보 등을 포함하여 데이터 도메인별로 노출에 대한 민감성 등을 고려하여 비식별화 키워드 도출

〈 그 자체로 개인을 식별할 수 있는 정보 (예시) 〉

- ① 쉽게 개인을 식별할 수 있는 정보(이름, 전화번호, 주소, 생년월일 등)
- ② 고유식별정보(주민등록번호, 운전면허번호, 외국인등록번호, 여권번호)
- ③ 생체정보(지문, 홍채, DNA 정보 등)
- ④ 기관, 단체 등의 이용자 계정(등록번호, 계좌번호, 이메일 주소 등)

② 비식별화 키워드에 대한 비식별화 조치 수행

- 개인정보 비식별화 적용 시 식별 및 비식별화 처리가 용이하도록 Keyword 단위의 데이터로 분할 적용
 - ※ 문서파일(한글, 워드 등), HTML(웹페이지), 센서, 영상, 음성 등 다양한 형태로 존재하는 데이터를 Keyword 형태로 분할/변환
 - ※ Keyword 기반의 패턴을 활용하여 개인정보 검사·식별
 - ※ Keyword 형태의 개인정보 예 : 주민등록번호, 여권번호, 의료보험번호 등
- 데이터가 Mash-up되어도 개인정보 관련 문제가 발생하지 않도록 비식별화 조치

– (비식별화) 가급적 자동 필터링 기법을 적용하여 자동 비식별성 추가

〈 비식별화 처리 기술 〉

기능	설명 및 예시
· 가명처리	· 개인정보 중 주요 식별요소를 다른 값으로 대체하여 개인식별을 곤란하게 함 (예) 홍길동, 35세, 서울 거주, 한국대 재학 => 임꺽정, 30대, 서울 거주, 국제대 재학
· 초계처리 또는 평균 값 대체	· 데이터의 총합 값을 보임으로서 개별 데이터의 값을 보이지 않도록 함 (예) 임꺽정180cm, 홍길동170cm, 이궁쥬160cm, => 물리학과 학생 키 합 : 660cm, 평균키 165cm
· 데이터값 삭제	· 데이터 공유 개방 목적에 따라 데이터 셋에 구성된 값 중에 필요 없는 값 또는 개인식별에 중요한 값을 삭제(예) 홍길동, 35세, 서울 거주, 한국대 졸업 => 35세, 서울 거주 (예) 주민등록번호 901206-1234567 => 90년대 생, 남자
· 범주화	· 데이터의 값을 범주의 값으로 변환하여 명확한 값을 감춤 (예) 홍길동, 35세 홍씨, 30~40세
· 데이터 마스킹	· 공개된 정보 등과 결합하여 개인을 식별하는데 기여할 확률이 높은 주요 개인식별자가 보이지 않도록 처리하여 개인을 식별하지 못하도록 함(예) 홍길동, 35세, 서울 거주, 한국대 학 재학 => 홍**, 35세, 서울 거주, **대학 재학

– (데이터 검증) 개인정보 비식별화 여부 등을 검증 할 수 있는 시스템을 갖출 것을 권고

※ 검증 시스템 구축 : 비식별화 알고리즘 역추적, 패턴검출, 개인정보 유사성 탐지 등을 적용한 개인식별정보 검증 시스템 구축

– (데이터 인증) 개인정보가 정상적으로 처리된 데이터는 사후관리를 위해 가능한 범위 내에서 인증표기하여 관리 할 것을 권고

※ 인증 시스템 구축 : 검증작업이 완료된 데이터 셋에 인증 Tag를 부여하고, 필요한 경우에 한하여 원본데이터는 태깅 및 암호화하여 보관(백업)

– (사후 모니터링) 관리가 필요한 데이터의 경우, 데이터 목적에 맞게 활용하고 폐기되는 단계까지 사후모니터링 할 것을 권고

2.3.4 기술 고려사항

데이터 보안관리

- 사용자 인증, 접근제어, 암호화 등 관련 기술 활용 시 기술 고려사항

〈 데이터 보안 기술 고려사항 〉

단계	고려사항
데이터 수집	· 데이터에 대한 사용자별 접근제어를 지원해야 한다.
	· 사용자 인증 및 인증 방법을 제공해야 한다.
	· 인증된 사용자의 수집 데이터 접근 제어 기능을 지원해야 한다.
	· 중요 데이터의 암호화 저장 여부와 암호화 방식을 지원해야 한다.
데이터 관리	· 사용자 접근 로그 관리 기능을 제공해야 한다.
	· 사용자 접근 제어를 지원해야 한다.
	· 사용자 데이터 처리 권한 설정 기능을 제공해야 한다.
	· 암호화 대상 영역의 성능 이슈를 관리 할 수 있는 모니터링 기능이 제공 되어야 한다.
데이터 분석 활용	· 사용자의 데이터 분석 가능 여부, 인증 방법을 제공해야 한다.
	· 인증된 사용자의 분석 기능 사용 권한 설정 기능을 제공해야 한다.
	· 분석된 결과에 대한 보안성 검토 기능을 제공해야 한다.
	· 사용자의 분석결과 사용 인증과 인증 방법을 제공해야 한다.
	· 인증된 사용자의 활용 기능 사용 권한 설정 기능을 제공해야 한다.
	· 데이터 전송 채널에서 암호화 처리 여부를 지원해야 한다.

개인정보보호 처리

- 개인정보 식별 및 비식별화 처리를 위한 기술개발 활용 시 고려사항

〈 개인정보 처리 시 고려사항 〉

기능	고려사항
개인정보 식별	· 파일 내용을 토크로 추출하는 형태소 분석 기능을 지원해야 한다.
	· 추출된 토큰과 룰(Rule) 파일의 개인정보 식별 규칙과 비교하여 개인정보 포함 여부를 식별하는 기능을 제공해야 한다.
	· 룰(Rule)을 사용자 지정 방식으로 추가 또는 수정 할 수 있는 룰 관리 기능이 제공 되어야 한다.
	· 개인정보 식별을 추론 하거나 예측 패턴을 적용 할 수 있는 식별 패턴 사전 적용 할 수 있는 기능이 제공 되어야 한다.
개인정보 비식별화	· 개인정보를 룰 파일에 정의된 개인정보 비식별화 기법을 적용하는 기능을 지원해야 한다.
	· 비식별화 이력 관리 기능을 제공해야 한다.
	· 비식별화 변환을 위한 적용 값 관리 기능이 제공되어야 한다.
	· 자동/반자동/수동 변환 등의 사용자 지정에 따른 변환 프로세스가 적용될 수 있는 기능이 제공되어야 한다.
	· 개인정보 비식별화 변환 후 변환 값의 복원 또는 원본 대조용 Key 값의 관리 기능이 제공 되어야 한다.
	· 변환 후 개인식별 변환에 따른 변환검증을 확인 하기 위한 변환 로그를 저장 하는 기능이 제공되어야 한다.
정책 관리	· 정책 설정에 개인정보 식별 규칙 정의 기능을 제공해야 한다.
	· 룰 파일에 개인정보 비식별화 기법(가명처리 등)을 정의할 수 있도록 지원해야 한다.
검증	· 개인 식별 정보를 탐지 및 검증할 수 있는 개인 식별 정보 DB 구성기능을 지원해야 한다.
	· DB로부터 검증 대상에 따른 검증 기준 설정 기능을 제공해야 한다.
	· 검증 대상 데이터 셋에 해당하는 인증 Tag를 구성할 있도록 지원해야 한다.
	· 개인 식별정보 검증을 위한 판별 알고리즘 등의 검증 set을 지정할 수 있도록 지원해야 한다.
	· 검증 대상문서 분석, 개인식별 정보 패턴 식별, 판단 및 기록을 수행할 수 있도록 지원해야 한다.
	· 검증 대상문서에 대한 개인 식별 결과를 시각적으로 표현할수 있도록 지원해야 한다.

빅데이터 품질관리 고려사항

- 각 기관은 빅데이터 결과의 정확성 및 신뢰성 향상을 위해 빅데이터 품질 관리 체계 확보 필요
 - 아래 업무 지침 등을 활용 또는 준하는 자체 품질관리 지침서를 마련하여 데이터의 정확성, 완전성, 적시성, 일관성 등 관리

지침 및 매뉴얼	주요내용
공공기관의 데이터베이스 품질관리 지침	· “국가정보화 기본법”에 따라 공공기관이 국가정보화를 추진하는 과정에서 생산, 유통 또는 활용되는 데이터베이스 품질을 관리하기 위한 세부 사항 규정 · 공공기관의 데이터베이스 품질관리를 위한 데이터베이스 품질관리계획을 수립하고, 품질 오류 신고 접수 및 처리, 데이터베이스 표준화, 연계 데이터 품질관리, 품질관리 지원센터 등에 대한 지침을 규정
공공정보 품질관리 매뉴얼	· 공공기관이 정보화사업을 계획하거나 실제 구축사업을 추진하는데 있어 데이터 품질관점에서 무엇을 고려하여야 하고 현재 운영중인 정보시스템의 데이터 품질을 확보하기 위해 어떻게 해야 하는지에 대한 방안 제시 · 데이터 품질관리에서는 공공기관에서 수행해야 할 데이터 품질관리 전반에 대한 구성체계를 제시하고 각 구성 요소 별 의미와 주요 내용을 설명 · 데이터 품질진단 및 개선에서는 기관차원에서 가시적인 데이터 품질을 향상시키기 위한 품질 진단 및 개선 단계별 절차를 설명하고 공공기관 담당자가 각 단계별 활동에서 고려해야 할 사항과 주의할 점 등을 설명
데이터 품질 관리 지침	· 데이터 품질관리 실무 담당자들이 데이터 품질관리의 개념과 방법을 이해하고 업무에 활용할 수 있도록하기 위해 개발 · 내용은 데이터 품질관리 프레임워크, 데이터/데이터 구조/데이터 관리 프로세스로 품질관리 대상을 구분하여 각 대상 별 관리 목적, 관리 기준 및 관리 방법 및 절차에 대하여 설명

- 빅데이터는 기존 데이터 품질 요소는 물론 분석을 위한 충분한 데이터량의 확보도 고려되어야 함

- 빅데이터 활용에 있어 수집, 저장관리, 분석 활용 단계별로 데이터 품질 관리를 위해 아래 사항 참고

단계	점검 사항
데이터 수집	- 수집 기준의 타당성(근거, 통계적 유의성)이 확보 되었는가? - 추출조건에 맞는 정보의 관련 항목 모두가 추출 되었는가? · 필수 정보항목 누락 여부 등 - 약의적 유포데이터 제거방법을 확보하였는가? (예, 아크바이트 댓글, 고의적 업무방해)
데이터 저장관리	- 저장 시 누락된 데이터는 없는가? - 저장을 위한 키 구성이 적절한가? - 품질검증 방법 · File 검증방법 · 저장 layout에 따른 검증
데이터 분석 활용	- 최신 데이터인가? - 충분한 정보가 제공되고 있는가? - 사용자가 원하는 정보가 제공되었는가? - 품질관리 확인 방법 · 내부 사용자 설문조사 · 피드백 · 만족도 조사 · 사후 고객만족도 분석

3 데이터 분석

3.1 개요

정의

빅데이터 분석 플랫폼을 활용하여 수집 및 저장된 데이터를 분석하여 서비스를 개발하고 통찰력(Insight)을 끌어내는 작업

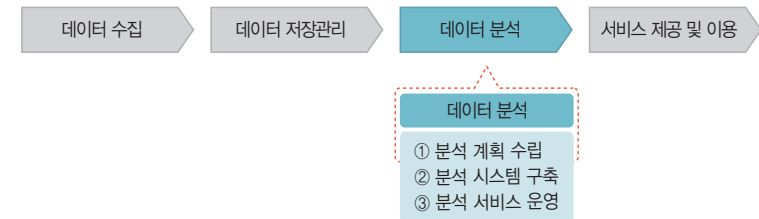
〈참고 사항〉

- 분석결과를 통해 해결하고자하는 문제를 명확히 정의하고 분석절차, 기법 등을 포함한 세부 시나리오 마련
- 외부 전문가 자문 등을 통하여 최적화된 분석 플랫폼을 구축하고 분석에 필요한 전문가 확보

업무 절차

- 분석 계획을 수립하고, 분석 환경을 구축(자체구축 또는 타 분석인프라 활용) 한 후 분석 서비스 운영

〈데이터 분석 절차〉



- ① (분석계획 수립) 분석 목적을 정의하고 분석을 위한 시스템 환경 및 방법론 등 세부 분석계획을 수립
- ② (분석시스템 구축) 분석 시스템 구축 비용 및 운영방식 등을 고려하여 자체 구축 및 타 분석시스템 활용
- ③ (분석 서비스 운영) 분석 전문가를 포함한 전담 인력을 확보하고 서비스 운영

3.2 활용 기술

- 데이터 분석은 분석 용도에 따라서 통계, 데이터 마이닝, 텍스트 마이닝, 소셜 네트워크 분석 등 기술 활용중

〈 빅데이터 분석 기술 소개 〉

구 분	설 명
통계적 분석	– 전통적인 분석 방법으로 주로 수치형 데이터에 대하여 확률을 기반으로 어떤 현상의 추정, 예측을 결정하는 기법
	기술통계량 대표적으로 평균(산술평균, 중앙값, 최빈값) 분산, 표준편차 등.
	상관분석 (Correlation analysis) 두 변수간에 어떤 선형적 관계를 갖고 있는 지를 분석하는 방법. 두변수는 서로 독립적인 관계로부터 서로 상관된 관계일 수 있으며 이때 두 변수간의 관계의 강도를 상관관계(Correlation, Correlation coefficient)
	회귀분석 (Regression analysis) 연속형 변수들에 대해 독립변수와 종속변수 사이의 상관관계에 따른 수학적 모델인 선형적 관계식을 구하여 어떤 독립변수가 주어졌을 때 이에 따른 종속변수를 예측, 또한 이 수학적 모델이 얼마나 잘 설명하고 있는지를 판별하기 위한 적합도를 측정하는 분석 방법.
	분산분석 (Analysis of variance : ANOVA) 두 개 이상 다수의 집단을 비교하고자 할 때 집단 내의 분산, 총평균과 각 집단의 평균의 차이에 의해 생긴 집단 간 분산의 비교를 통해 만들어진 F분포를 이용하여 가설검정을 하는 방법
	주성분분석 (Principal component analysis) 다양한 변수들에 대해 분석하는 다변량(multivariate) 분석으로 많은 변수들로부터 몇 개의 주성분들을 추출하는 방법. 즉, 주성분분석은 차원축소(dimension reduction)를 위한 것임
데이터 마이닝	– 대용량의 데이터로부터 패턴인식, 인공지능 기법 등을 이용하여 숨겨져 있는 데이터간의 상호 관련성 및 유용한 정보를 추출하는 기술
	예측 (Forecasting) 대용량 데이터집합내의 패턴을 기반으로 미래를 예측 (예: 수요예측)
	분류 (Classification) 일정한 집단에 대한 특정 정의를 통해 분류 및 구분을 추론 (예: 이탈한 고객)
	군집화 (Clustering) 구체적인 특성을 공유하는 자료들을 분류, 미리 정의된 특성에 대한 정보를 가지지 않는다는 점에서 분류와 다름 (예 : 유사 행동 집단의 구분)
	패턴분석 (Association) 동시에 발생한 사건간의 상호연관성을 탐색 (예: 장바구니의 상품들의 관계)
	순차패턴분석 (Sequencing) 연관 규칙에 시간(time)의 개념을 반영하여 시계열(time series)에 따른 패턴들의 상호연관성을 탐색 (예: 금융상품 사용에 대한 반복 방문)
텍스트 마이닝	– 텍스트 마이닝은 텍스트 기반의 데이터로부터 새로운 정보를 발견할 수 있도록 정보 검색, 추출, 체계화, 분석을 모두 포함하는 Text-processing 기술 및 처리 과정
소셜 네트워크 분석	– 대용량 소셜 미디어를 언어분석 기반 정보추출을 통해 이슈를 탐지하고, 시간의 경과에 따라 유통되는 이슈의 전체과정을 모니터링하고 향후 추이를 분석하는 기술

3.3 세부 업무처리 절차

● 분석 계획 수립

● (분석 시나리오 작성) 데이터 분석 목적을 정의하고 시나리오 작성

– 데이터 분석을 통해 해결하고자 하는 문제를 명확히 정의하고 분석 절차, 분석기법 등 세부 시나리오 작성

● (분석 환경 조사) 분석에 필요한 인프라 구축 조건 등을 분석하여 자체 구축 및 외부 인프라 활용여부 결정

〈 분석 환경 조사 시 고려사항 〉

구 분	설 명	고려사항
자체 구축	– 빅데이터 분석과 활용을 위해 분석 시스템과 운영환경을 기관 내에 구축하는 방식	– 내부 데이터의 관리 정책과 보안문제로 외부 서비스를 활용하기 어려운 경우 – 분석 요구사항을 외부 서비스 기관에서 지원하지 못하는 경우
외부 활용	– 외부 분석업체의 분석 서비스를 활용하는 방식 – 기관의 시스템을 외부 분석 시스템과 연계하여 활용하는 방식	– 외부 분석 시스템의 기능과 분석 품질이 활용 목표 수준에 부합할 경우 – 초기 투자비용이 적고, 초기 활용이 필요한 경우

● (추진계획 수립) 분석 목적, 분석 방법론, 분석 시나리오 작성, 분석 인프라 구축 방식 및 운영예산 등 세부계획 수립

※ 분석인프라 구축 비용 등을 감안하여 인프라 설계 시 관련 분야의 전문가 활용 권고

● 분석 플랫폼 구축 · 운영

- (HW 구축) 빅데이터 수용 용량 및 분석작업에 대한 부하 등을 감안하여 HW 인프라 구축

– 수집 데이터 저장 서버, 데이터 처리서버(하둡기반 분석, 정형데이터 분석 DW 등), 기타 보안, 통신장비 등 구축

- (SW 구축) 분석에 필요한 수집, 관리, 분석, 이용자 환경 등 관련 SW 구축

〈 SW 구축 시 필요 기능(예시) 〉

기능	구성요소 (예)	주요 내용
빅데이터 수집	Flume, Sqoop, 크롤러, Open API	외부 데이터 추출, 변환, 적재
분산파일 관리	분산파일시스템 (HDFS 등)	MapReduce 지원 가능 분산파일 시스템
빅데이터 분석	MapReduce	대용량 로그 파일 처리 프레임워크
	Pig	HDFS 대용량 로그 파일을 처리하는 스크립트 언어
	Hive	SQL 기반 대용량 로그 파일의 집계기능을 제공하는 SQL 실행 엔진
	Mahout	알고리즘 패키지
	R	오픈소스 통계 패키지
사용자 및 인프라 관리	사용자 인증 관리	사용자 등록, 인증, 권한 관리 등
	인프라 자원 관리	사용자별 인프라 자원 할당 및 이용 스케줄 관리 등
	빅데이터 인프라 상태 모니터링	빅데이터 분석 장비 상태 정보(이벤트, 장애 등), HDFS/MapReduce Job 모니터링 등
	빅데이터 인프라 설정 관리	하둡 등 오픈소스 패키지 자동인스톨, 설정파일 작성/변경/운영 등
빅데이터 분석 이용 환경	파일 시스템 브라우저	파일 업로드/다운로드, 파일 목록 관리 등
	작성된 분석 프로그램 저장관리	사전에 작성된 Hive, Pig 등 쿼리 및 스크립트 파일 목록관리
	통합분석 환경	Pig, Hive, R 등 분석 스크립트/쿼리 작성 및 관리 Editor
	분산처리 자원 통계분석 기능(R)	R을 위한 MapReduce 연동기능
	워크플로우 관리	다수의 분석 작업 설계, 저장, 로딩, 실행 등 스케줄링 관리
	분석 결과 레포팅	분석 결과를 다양한 방식의 차트(시각화) 및 파일(csv 등)로 제공

외부 시스템 활용 시 고려사항

- 외부 전문 분석 서비스들을 조사하고 분석 목표에 부합되는 대상 선정
- 외부 분석 업체의 데이터 보안 및 개인정보보호 등 데이터 보안 정책과 책임에 대해 확인
- 참조 데이터와 분석 시나리오에 기반 한 시험분석을 사전 실시/컨설팅
- 시험 분석 후 분석 항목, 품질 및 리포트 등이 요구사항과 분석 목표에 부합하는지 평가
- 데이터 분석을 위한 분석 업체(기관)와의 데이터 교환 방식 설계 및 구현
- 분석 업체의 서비스 시스템과의 연계 필요성 검토 및 설계
- 데이터 전송이 필요한 경우 접근권한, 보안체계 등에 대한 설계
- 사용자 인터페이스와 접근 권한 관리 등의 기능을 고려한 연계 시스템 구현
- 연계 시스템의 통합 분석 시험을 통해 서비스가 정상적으로 작동되는지 확인

● 분석 서비스 운영

- 자체 분석 시스템을 운영하는 경우, 분석 전문가를 확보하고 지속적인 교육·훈련이 필요하며 보안 담당자 지정 필요
 - 외부 시스템을 이용할 경우 데이터 분석 결과에 대한 품질, 데이터 운영 관련 보안 사항 등 필요한 사항에 대한 SLA 협약 권고

Tip 데이터 분석에서의 주요 사항

- ① 데이터 분석을 통해 해결하고자 하는 문제를 명확히 정의하고, 기술적 해결책 위주의 접근방법이 아닌 사용자 관점의 분석 시나리오 작성이 필요하다.
- ② 분석 지원을 위한 조직 환경(인력, 데이터 관리 정책, 보안, 인프라 등)에 따라 자체 분석 시스템 구축 또는 외부 분석 서비스를 활용한다.
 - 자체 분석 시스템 구축 시 내외부 시스템 간 연동 체계 및 분석 플랫폼 아키텍처 정의, 분석 기술 및 알고리즘 선정과 배치 등을 상세하게 설계하여 구축
 - 외부 분석 서비스 활용 시 분석업체(기관)를 조사하여 선정하고, 해당 업체(기관)와의 통합 분석 시험을 통해 분석 목적을 달성하는지 확인한다.
- ③ 분석을 통해 고 부가가치 정보를 얻기 위해서는 분석 모델링, 분석 결과 해석 등을 수행할 수 있는 분석 전문가 확보 및 인력 양성이 필요하다.
- ④ 무엇보다도 빅데이터 분석 시 개인정보처리에 대한 보안 담당자를 지정하여 관리토록 해야한다.

3.4 기술 고려사항

통계적 분석 및 데이터마이닝, 텍스트마이닝, 소셜네트워크 분석, CEP, ESP 등 분석기법 구현 시 기술 고려사항

통계 분석 및 데이터마이닝

- 데이터 내에 존재하는 관계, 패턴, 규칙 등을 탐색하고 모형화 하여 지식을 추출하기 위해 필요한 기술 고려사항

〈 통계적 분석 및 데이터 마이닝 고려사항 〉

기능	고려사항
샘플링(추출) Sampling or Selection	· 데이터 선정 시 각각의 개체가 모두 무작위 방법 선정 기능을 제공해야 한다.
	· 데이터 선정 시 몇 개의 층으로 나누어 각 층으로부터 동등하게 무작위 방법 선정 기능을 제공해야 한다.
	· 데이터 선정 시 몇 개의 층으로 나누어 각 층의 비례에 따라 무작위 방법 선정 기능을 제공해야 한다.
	· 데이터 선정 시 몇 개의 층으로 나누고 그 중 하나의 층에서만 무작위 방법 선정 기능을 제공해야 한다.
데이터 사전 처리 및 변환	· 평균형태의 데이터 분포에서 혼자 떨어져 있는 값(이상치, 특이값)을 관측하여 제거하는 기능을 제공해야 한다.
	· 데이터 범위를 일치시키거나 유사하게 만들어 주는 기능(표준화 : normalization)을 제공해야 한다.
	· 가장 크게 영향을 줄 것으로 판단되는 변수부터 하나씩 선택하여 판단하는 기능을 제공해야 한다.
	· 가장 적게 영향을 줄 것으로 판단되는 변수부터 하나씩 제거하여 판단하는 기능을 제공해야 한다.
	· 중요한 변수를 하나씩 선택하면서 단계별로 변수의 중요성을 판단하여 제거유무 판단 기능을 해야 한다.
	· 데이터 기존 수치 적용 또는 추론 / 패턴 적용을 위한 데이터 사전 기능이 제공 되어야 한다
	· 데이터 사건의 추가/수정을 위한 관리 기능이 제공 되어야 한다.
분석모형 생성	· 연속형 데이터의 독립변수와 종속변수의 상관관계에 따른 선형적 관계식을 구하여 예측하는 기능을 제공해야 한다.
	· 종속변수의 이진 데이터화를 통해 독립변수와와의 관계를 발생가능성으로 예측하는 기능을 제공해야 한다.
	· 반응변수의 형태에 따라 연결함수 자동 지정 기능을 제공해야 한다.
	· 전체 데이터 셋을 반복해서 두 개의 노드로 분리하여 예측변수를 사용하여 의사결정 트리 생성 방법 지원 기능을 제공해야 한다.
	· 범주형 데이터의 예측 변수와 결과변수와의 관계를 찾는 방법 지원 기능을 제공해야 한다.
시각화	· 데이터의 분석결과를 시각화하여 시각적 분석기능을 제공해야 한다. 기반의 문서를 처리하여 시각화 하는 기능을 제공해야 한다.
	· 사용자 UI 환경에 호환되도록 시각화 기능을 제공해야 한다.
	· 지도 및 각종 차트를 이용한 데이터 시각화 기능을 제공해야 한다.
	· 사용자 상호작용을 통하여 분석된 결과를 탐색하거나 검색할 수 있는 기능을 제공해야 한다.

텍스트마이닝

● 텍스트 기반의 데이터 분석 구성, 분해, 예측을 위한 기술 고려사항

< 텍스트마이닝 기술 고려사항 >

기능	고려사항
분석데이터 구성	- 수집된 다양한 비정형 데이터(오피스 문서, 소셜 미디어 데이터 등)로부터 텍스트 데이터를 분리해 내는 텍스트 필터링 기능이 제공되어야 한다. - 이 과정에서 원본 문서의 메타데이터나 문서 구조 정보를 별도로 추출 및 수집하는 기능이 필요한 경우도 있다.
형태소 분석	- 텍스트 데이터의 각 문장들에 포함된 단어(또는 어절)를 구성하는 각 형태소 분리하고 분리된 형태소의 기본형 및 품사 정보 추출한다. - 정보로서 가치가 없는 용어인 불용어를 제거하기 위한 기능을 제공해야 한다. - 유사어, 합성어 등의 처리를 위해 준비된 사전에 따라 분석을 수행할 수 있어야 한다. - 불용어, 유사어, 합성어 등의 언어 사전은 분석 도메인에 따라 특화될 수 있으므로 분석시 언어 사전을 조정 및 관리할 수 있어야 한다.
개체명 분석	- 텍스트 데이터로부터 인명, 지명, 기관명 등과 같은 개체명을 인식하여 해당 개체명에 개체분류를 위한 태그를 부착하여 주는 기능을 수행해야 한다. - 분석 도메인에 따라 특화된 다양한 개체를 식별할 수 있도록 개체명 사전을 등록, 수정 및 관리할 수 있어야 한다.
텍스트 군집과 분류	- 주어진 텍스트 데이터로부터 특성 데이터를 추출할 수 있어야 한다. - 추출된 특성을 중심으로 그들 간의 유사도를 계산하고 유사한 문서들끼리 군집화할 수 있어야 한다. - 정의된 분류체계를 기준에 합당한 텍스트 데이터를 이용하여 분류체계에 대한 기계학습을 수행할 수 있어야 한다. - 기계학습된 분류체계에 따라 새로운 텍스트 데이터를 자동으로 분류할 수 있어야 한다. - 기계학습으로 분류할 수 없는 적은 양의 텍스트 데이터를 분류하기 위해 규칙 기반의 분류 기능을 제공해야 한다. - 규칙 기반 자동 분류를 위해 분류 규칙을 추가, 수정 및 관리할 수 있어야 한다.
자동 요약	주어진 텍스트 데이터 활용도를 높이기 위해 텍스트 데이터의 특성을 유지시키면서 텍스트 데이터의 길이와 복잡도를 줄일 수 있어야 한다.
토픽 분석	- 일련의 텍스트 데이터로부터 주요 토픽(키워드 혹은 주제어)을 추출하고 토픽들이 시계열적인 트렌드와 공간적인 분포를 측정할 수 있어야 한다. - 주어진 텍스트 데이터로부터 주요 토픽(키워드 혹은 주제어)을 추출하고 토픽들 간의 연관성을 측정하여 연관성 네트워크를 구성할 수 있어야 한다.

기능	고려사항
감성 분석	- 주어진 텍스트 문장으로부터 감성 표현을 식별하고, 주어진 감성 분류 기준에 따라 자동으로 감성 분류를 수행 할 수 있어야 한다. - 감성 표현과 감성 분류 기준은 분석 대상이 되는 시기, 지역, 도메인에 따라 지속적으로 변화하므로 필요에 따라 감성 표현 사전을 구성하고 관리하는 기능을 제공해야 한다.

소셜네트워크 분석

● 데이터 간 상호 연관성을 추출하고 분석하기 위한 기술 고려사항

< 소셜네트워크 분석 기술 고려사항 >

기능	고려사항
데이터 추출 및 변환	- 네트워크 데이터 추출 - 대상 데이터로부터 객체간(node 혹은 entity)의 관계 데이터를 자동으로 추출하여 전체 네트워크 데이터를 구성하는 기능에 제공되어야 한다. - 데이터 변환 - 네트워크 데이터를 분석 설계에 따라 분석 가능한 형태로 자유로운 변형이 가능하여야 한다.
소셜네트워크 분석	- 네트워크 통계 - 전체 네트워크 데이터의 특성을 알 수 있는 기본적인 지표를 제공해야 한다. - 중심성 분석 - 전체 혹은 일부 네트워크 데이터 내의 중심적인 객체를 파악하고 중심성을 계량적 수치로 제공해야 한다. - 응집성 분석 - 네트워크 데이터 내의 객체들 간의 최적의 커뮤니티(응집그룹)를 분석하는 기능을 제공해야 한다. - 연결구조 분석 - 특정 객체로부터 다른 객체들로 연결된 네트워크 상 최단경로 혹은 전파경로를 분석할 수 있는 기능이 제공되어야 한다. - 역할구조 분석 - 특정 응집그룹(커뮤니티) 내의 객체들 간의 역할을 분석하고 이를 계량적 수치로 제공해야 한다. - 영향력 및 파급력 평가 - 네트워크 내의 객체들간의 영향력과 파급력을 계량적 수치로 제공해야 한다. - 네트워크 시각화 - Graph DB 혹은 RDB로부터 즉각적인 네트워크 시각화 인터페이스를 제공해야 한다. - Flex, HTML5, SVG 등 이용자 환경과 호환기능을 제공해야 한다. - Spring map, Concentric map 등 다양한 네트워크 시각화 레이아웃을 제공해야 한다. - 사용자 인터페이스 구성을 위해 API 지원기능을 제공해야 한다.

🌐 CEP(Complex Event Processing)

- 여러 소스로부터 발생한 이벤트를 대상으로 실시간 의미 있는 데이터를 추출하기 위한 기술 고려사항

〈 CEP 기술 고려사항 〉

기능	고려사항
이벤트 필터링	· 조건문을 이용하여 원하지 않는 데이터를 걸러내는 기능을 제공해야 한다.
이벤트 추출	· 일정 간격으로 발생한 데이터의 중복 내용을 제외한 새로운 이벤트 데이터만 추출하는 기능을 제공해야 한다.
	· 데이터에 여러 개가 섞여 입력될 경우 분리를 통한 추출 기능을 제공해야 한다.

🌐 ESP(Event Stream Processing)

- 지속적인 이벤트 소스로부터 발생한 스트림 데이터를 실시간으로 분석 처리하기 위한 기술 고려사항

〈 ESP 기술 고려사항 〉

기능	고려사항
이벤트 추출	· 지속적으로 발생하는 스트림 데이터를 실시간 질의하여 데이터를 추출하는 기능을 제공해야 한다.
	· 실시간 ETL 기능을 제공해야 한다.
모델 생성	· 실시간 분석 모델과 예측 모델 생성 기능을 제공해야 한다.

참고

데이터 분석 기술 참조 사이트

- 샘플링/통계적 분석
 - R : <http://www.r-project.org/>
- 시각화
 - D3 : <http://d3js.org/>
 - Leaflet : <http://leafletjs.com/>
 - arbor.js : <http://arborjs.org/>
 - Envision : <http://www.humblesoftware.com/envision/documentation>
 - Kartograph : <http://kartograph.org/>
 - Paper.js : <http://paperjs.org/>
 - Polymaps : <http://polymaps.org/>
 - Processing.js : <http://processingjs.org/>
 - Sigma.js : <http://sigmajs.org/>
- 추천/분류/군집/예측
 - Mahout
 - <http://mahout.apache.org/>
- Esper : CEP를 활용
 - <http://esper.codehaus.org/>
- Storm : ESP를 활용
 - <http://storm-project.net/>

4 데이터 제공 · 이용

4.1 개요

정의

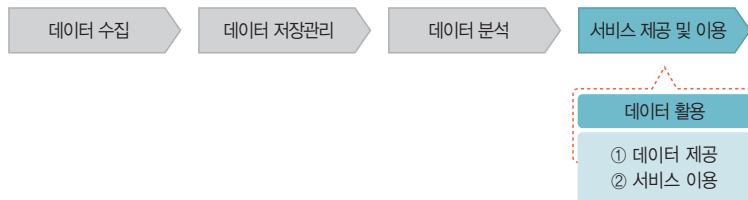
원시 데이터 또는 분석 결과를 서비스로 제공 또는 이용하는 과정

〈 참고 사항 〉

- 분석된 결과 및 데이터 자원을 효율적으로 이용 할 수 있도록 표준화 된 체계(open API 등) 마련
- 서비스 개발자, 데이터 이용자들이 분석 플랫폼에 접근하여 서비스를 이용할 수 있는 절차 등 이용체계 마련

업무 절차

〈 데이터 활용 절차 〉



① (데이터 제공) 데이터 제공 기준, 이용자 인증, 비용 산정 등 데이터 제공 절차 마련

② (서비스 이용) 빅데이터 오픈 플랫폼에 접속하여 제공 데이터 및 분석결과 등 이용

4.2 활용 기술

- 빅데이터 활용 기술로 Open API와 사용자 인증 기술 등이 적용중

〈 데이터 교환 기술 소개 〉

구분	설명
데이터 교환 (Open API)	- 응용프로그램 및 서비스를 개발하는데 필요한 운영체제나 라이브러리 등의 특정 기능을 추상화하여 사용하기 쉽도록 만든 인터페이스
데이터 사용 인증 (OAuth 2.0)	- 어플리케이션이 안전하게 해당 데이터에 접근할 수 있도록 제어하는 기술

4.3 세부 업무처리 절차

① 데이터 제공

데이터 제공 대상 선정

- (데이터 제공기준 마련) 데이터 제공과 관련된 법령을 참고하여 데이터 제공 대상 선정 기준 마련

※ 「공공기관의 정보공개에 관한 법률」, 「공공데이터의 제공 및 이용 활성화에 관한 법률」, 「개인정보보호법」등과 관련 시행령 및 규칙 등을 참조

- 비공개 대상 정보 정의 : 「공공기관의 정보공개에 관한 법률」 제9조, 「저작권법」 및 그 밖의 다른 법령 등 준수
 - 또는 필요한 경우, 기관별 내부 규정 등에 따라 보유 데이터의 비공개 대상 정보 선정 기준 마련

● **(제공 대상 데이터 목록 작성)** 제공 및 비공개 데이터 전체 목록 작성

– 제공 대상 및 비공개 데이터 목록 작성

- 「공공데이터의 제공 및 이용 활성화에 관한 법률」 제17조제1항 각 호 및 기관별 내부 규정 등에 따라 제공대상이 아닌 정보가 포함되지 않은 경우 제공대상 목록으로 작성

※ 제공 대상 데이터 목록 작성 시 ‘공공데이터의 제공 및 이용 활성화에 관한 법률 시행규칙’(안전행정부령 제23호, 2013.10.31.) [서식4] 제공 대상 데이터 등록서 참고

- 데이터에 제공 대상이 아닌 정보가 포함된 경우 해당 정보의 기술적 분리 가능여부 확인 후 부분제공, 미 제공 목록 작성

● **(제공 대상 데이터 목록 공개)** 제공 대상 목록을 쉽고 편리하게 조회·검색할 수 있도록 기관별 홈페이지 등에 서비스 제공

– 제공 데이터 목록을 조회·검색할 수 있도록 기관 홈페이지 또는 자체 데이터 포털 운영

– 공공기관의 경우, 기관별 홈페이지를 통해 접속하는 국민이 제공목록을 조회하고 이용할 수 있도록 기관별 개방정보를 공공데이터포털에 등록하고 연계

※ 공공데이터포털(www.data.go.kr)에서 제공하는 I-frame을 이용하여 연결

 데이터 제공(개방) 서비스 마련

● **(제공 방식 결정)** 이용자가 제공 데이터에 쉽게 접근하고 이용할 수 있도록 데이터 제공방식을 선정

– 빅데이터 제공서비스 구현 시 이용자의 사용 편리성 및 데이터 안전성 등 고려

〈 데이터 제공 서비스 고려사항 〉

구분	고려 사항
사용 편리성	· 데이터 사용자가 자유롭게 원하는 서비스를 만들어 낼 수 있어야 함 · 사용자들이 특정 서비스만 개발할 수 있도록 제공자는 제공 서비스를 확정하거나 제한하지 않아야 함 · 사용자가 적절한 인증 절차를 거친 후에 제공된 데이터를 자유롭게 사용할 수 있어야 함
데이터 안전성	· 데이터 전송 중 데이터의 노출 및 변경이 없어야 함 · 사용자는 접근하는 데이터에 대한 사용권한이 있음을 인증 받아야 함

– Open API 활용 시 다수기관 사용, 실시간 전달, 멀티채널, 멀티 플랫폼별 적용 방안 고려

〈 Open API 활용 시 적용 방안 〉

구분	적용 방안
다수기관 사용	· 다수 사용자가 데이터를 공통적으로 활용하는 경우 · 다수 사용자가 활용 할 수 있도록 기능을 개발하여 배포하는 경우 공통 컴포넌트로 구현 · 특정 시스템에 접속하여 서비스를 호출하고 결과를 가져가는 경우 Open API로 구현
실시간 전달	· 제공 데이터에 실시간으로 접근하여 요청·전달 받아 사용하는 경우 · 데이터베이스 동기화 등 대량의 데이터를 다루는 경우에는 성능 문제가 발생 할 수 있어 동기화 수행을 위한 이벤트나 제어만 처리 · 직접적인 데이터 교환은 다른 기술을 활용(ESB 연계 등)
멀티채널	· 웹, 모바일 앱, 키오스크 등 다양한 채널에서 제공·활용되는 경우 · 서비스가 제공하는 SOAP, REST 메시지를 채널별로 사용자 인터페이스 기능을 통해서 가공하여 전달 · 특정 데이터를 웹, 모바일 웹, 앱 등의 다양한 채널로 제공해야 하는 경우 채널별로 중복하여 구현하는 것이 아니라, 제공서비스 형태로 식별·구축
멀티플랫폼	· 구축된 서비스가 J2EE, NET, PHP 등 다양한 플랫폼에서 연계하여 사용되는 경우 · 모든 플랫폼에서 지원하는 SOAP, REST 서비스 활용 · 플랫폼 별로 활용되고 있는 언어, 기술표준 등에 독립적으로 동작하여 이기종 환경 간에서 사용되는 경우, 공유서비스를 식별하여 구축

※ SOAP(Simple Object Access Protocol): HTTP 등을 사용하여 XML 기반의 메시지를 네트워크 상에서 교환하는 형태의 프로토콜

※ REST(REpresentational State Transfer): 웹 환경에서와 같이 정보자원을 정의하고 자원에 대한 주소를 유일한 식별자인 URI로 지정하여 접근하는 네트워크 메커니즘 기반

● **(사용자 인증방안 수립)** 제공되는 데이터의 안전한 사용을 위하여 데이터 제공 관련 인증방안 수립

- 데이터가 공개된 경로를 통해서 전달되는 경우 'Confidentiality(데이터 노출이 없음)'과 'Integrity(데이터 변경이 없음)' 두 가지 원칙을 활용하여 데이터 인증
 - 전송 시에 데이터의 노출이 없도록 암호화 적용
 - 전송 시에 데이터의 변경이 없는 것을 확인
 - 데이터 전송 구간 전체에 대해서 데이터 보호 보장
 - 전송구간 암호화가 사용되는 보안 프로토콜은 가급적 최신버전 활용
 - 인증할 수 있는 클라이언트와 인증할 수 없는 클라이언트는 구별하여 처리
- 데이터 사용 인증을 위한 접근권한 부여
 - 데이터 제공자와 사용자의 권한 부여 방식(구조) 구성
 - 사용자 등록과 별도로 사용자 자격 증명(아이디 혹은 토큰, 패스워드 혹은 암호화)은 필요한 범위에서 발급
 - 권한 부여 시 데이터 제공자에게 충분한 정보 전달 필요(등록된 사용자의 정보, 부여되는 권한, 권한 부여가 유지되는 기간 등)
 - 권한 부여 유효기간이 연장될 때마다 다른 자격 증명 사용
- 그 외 아래 사항들을 참고하여 데이터 활용에 대한 신뢰도 확보
 - 사용자는 사전 등록된 경우만 접속 허용
 - 데이터 URL을 호출할 수 있는 권한은 사용자 등록 시 같이 설정
 - 발급된 자격증명이 짧은 유지 시간을 갖도록 설정
 - 발급된 자격증명이 특정 사용자, 데이터에 한정 되도록 관리

● **(데이터 비용 산정)** 데이터의 제공에 드는 비용을 이용자에게 부담시킬지 여부는 기관별 자율적으로 판단 · 결정

- 필요한 경우, '공공기관의 정보공개에 관한 법률 시행규칙'에 명시된 수수료의 금액 등 참조

〈 공공기관 데이터 제공 비용 산정범위 〉

- 공공데이터의 제공에 소요되는 전자기록매체 비용 등 일반 경비
- 공공데이터 제공을 위해 추가로 필요한 정보시스템 및 네트워크 증설 · 유지보수 비용
- 제3자 권리 정보가 포함된 공공데이터 제공에 대한 정당한 권리이용 비용

● **데이터 제공**

● **(제공 절차 정의)** 기관별 데이터 제공 절차를 공개(개방) 데이터 제공과 이용 승인 데이터 제공으로 구분 가능

- 공개(개방) 데이터는 별도 신청 · 검토 없이 온라인 또는 오프라인을 통해 이용 가능하도록 제공

공개(개방) 데이터 제공절차 예시

- 온라인 : 이용자 → 홈페이지 또는 포털 접속 · 신청 → 다운로드 및 이용
- 오프라인 : 이용자 → 직접 방문 또는 우편 신청 → 데이터 수령 및 이용

- 이용 승인 데이터는 이용자에게 신청인, 활용 목적 혹은 법적 근거, 신청내용 등을 작성하여 제출하도록 신청서를 정의하여 제공
- 이용자는 이용 승인 데이터를 제공받기 위해 제공 데이터 신청서를 작성하여 제공자에게 온라인 또는 오프라인 방법으로 제출

이용 승인이 필요한 데이터 제공절차 예시

- (신청서 접수) 홈페이지 또는 포털 접속 · 신청(제공 데이터 신청서 입력 및 등록) 또는 직접 방문하여 신청서 접수
- (검토) 제공여부 검토 및 결정
- (통보) “제공결정” 혹은 “제공거부 결정” 통보

● (데이터 제공 신청 검토 및 통보) 데이터 제공 신청에 대한 제공 여부 검토 기준 수립

- 데이터 제공 신청서를 접수한 후 10일 이내 신청 내용을 검토하여 “제공 결정” 혹은 “제공거부 결정”

데이터 제공 신청 검토 항목 예시

- 요청한 법적 근거와 활용 목적의 정당성 검토
- 제공 요청한 목적에 따른 제공 항목 및 범위의 적정성 검토
- 제공 요청한 데이터의 보유 및 이용기간의 적정성 검토

- 검토 결과, 제공 결정 시 신청인에게 제공 방법 및 절차를 온라인 또는 문서전달 방법으로 통보
- 검토 결과 제공거부 결정 시 신청인에게 거부결정 사유 통보

② 서비스 이용

- (서비스 정의) 플랫폼 제공자는 운영 정책에 따라서 단순한 데이터 제공에서 부터 데이터 수집 · 분석 · 공유 기능 등 서비스 제공

빅데이터 서비스 유형 (예시)

- ① (이용자 보유 데이터 공유) 이용자가 보유한 데이터를 가지고 플랫폼 제공 기능을 이용하여 다른 사람과 공유하고자 하는 자
- ② (이용자 보유 데이터 분석) 이용자가 보유한 데이터를 가지고 플랫폼 제공 기능을 이용하여 분석업무를 수행하고자 하는 자
- ③ (플랫폼 제공 데이터 분석) 플랫폼에서 제공하는 데이터를 가지고 플랫폼 제공 기능을 이용하여 새로운 가치의 정보를 생성하고자 하는 자
- ④ (다수 사이트 데이터 수집 · 분석) 타 정보 제공 사이트의 데이터를 플랫폼에 수집해서 데이터를 공유 또는 새로운 가치의 정보를 생성하고자 하는 자

- (서비스 이용 절차) 이용자는 플랫폼을 통하여 제공서비스 유형을 확인하고 서비스 이용 절차에 따라 활용

- 플랫폼을 통하여 제공되는 서비스 유형을 확인 (매뉴얼 등 확인)
 - 데이터 수집에 관한 기능을 제공 또는 활용 가능한지 여부
 - 이용자에게 필요한 데이터가 플랫폼에서 제공중이거나 이용자가 보유한 데이터를 플랫폼에 저장할 수 있는지 여부
 - 데이터 분석을 위한 기술과 이용자 화면을 제공하는지 여부
 - 분석 결과를 제공 받을 수 있는 방법 및 기술 제공 여부
- 플랫폼 이용을 위한 이용 신청 및 승인 확인
 - 플랫폼 이용 목적, 범위, 데이터, 이용자 정보 등을 작성하여 플랫폼 운영자에게 제출
 - 이용 신청에 대한 플랫폼 운영자의 승인 후 플랫폼에 접속

- 플랫폼에서 제공하는 데이터를 이용 또는 이용자가 보유하고 있는 데이터를 플랫폼에 등록
 - 플랫폼에 저장되어 있는 데이터 검색을 통해 필요한 데이터 선정
 - 이용자가 보유하고 있는 데이터를 등록할 경우, 플랫폼에서 제공하는 등록 기능을 활용하여 데이터 등록
 - 필요한 데이터의 수집이 필요한 경우 플랫폼의 수집 기능 이용
- 데이터 분석 환경을 활용하여 빅데이터 분석 및 오픈 API, 분석기법, APP 등 개발 · 등록 · 실행
 - 데이터 전처리, 후처리, 분석 작업을 위해 플랫폼에서 제공하는 기능을 이용하여 작업 설정 후 실행
 - 필요한 경우, 플랫폼에서 제공하는 분석 프로그램 생성 · 등록 기능을 이용하여 프로그램 등록 및 실행

4.4 기술 고려사항

데이터 이용(교환) 및 서비스 사용자 인증에 대한 기술 고려사항

데이터 이용(교환)

- 데이터 서비스 이용에 활용중인 RESTful Open API, Open API 제공 Web 방식에 대한 기술 고려사항

※ REST는 SOAP 등과는 다르게 표준규약 및 강제적인 제약은 아니나, 제공될 API가 RESTful API가 되게 하려면 모든 항목 적용 필요

〈 RESTful Open API 고려사항 〉

기능	고려사항
실제로 제공하는 대상 (리소스) 설계	· 리소스를 기능이나 동작이 아니라 그 대상을 중심으로 두고 설계해야 한다.
	· 기능과 동작은 HTTP의 Method로 정의해야 한다.
	· 리소스간의 구조는 Set과 같이 나열해 놓은 형태로 구성해야 한다.
	· 필요한 경우 부모-자식 관계를 표현하는 Tree 형태로 구성한다.
리소스를 웹에서 표현(URI(Uniform Resource Identifier))	· 이외의 다른 리소스간의 구조는 표현하지 않아야 한다.
	· 리소스는 대응하는 URI를 가지고 있어야 하며, 이 URI는 하나의 리소스를 명확하게 지정하고 있어야 한다.
	· 리소스에 대응하는 URI는 여러 개 있을 수 있다.
	· URI에 프로그래밍 언어에 의존적인 확장자는 사용하지 않아야 한다.
	· URI에 프로그래밍 언어에 의존적인 경로명은 사용하지 않아야 한다.
	· URI에 Session ID를 사용하지 않아야 한다.
	· URI에는 소문자만 사용한다.
	· 하나의 리소스가 여러 개의 Representation을 가질 때는 확장자로 구분한다.
	· 리소스가 계층구조가 아닌 동등구조를 표현할 때는 Forward Slash(/_ 대신 Comma(,) 혹은 Semi-Colon(:)을 사용해야 한다.
	· 항목들의 순서가 의미를 가질 경우에는 Comma(,)를, 순서가 의미를 가지지 않는 경우 Semi-Colon(:)을 사용한다.
	· URI의 버전이 변경되더라도 이전에 사용하던 클라이언트들은 중단 없이 계속해서 기능을 사용해야 한다.

기능	고려사항
	· URI에 버전을 명시한다.
	· 각각의 URI에 대한 호출은 이전의 다른 호출의 영향을 받지 않아야 한다.
	· 실제 호출되는 uri는 그 자체로 필요한 모든 정보를 포함해야 한다.
리소스에 대한 동작 기술 (UI(Uniform interface))	· 리소스의 기능은 HTTP의 Method를 이용해서 구현해야 한다.
	· 생성 기능은 HTTP POST Method를 이용해서 구현해야 한다.
	· 조회 기능은 HTTP GET Method를 이용해서 구현해야 한다.
	· 갱신 기능은 HTTP PUT Method를 이용해서 구현해야 한다.
	· 삭제 기능은 HTTP DELETE Method를 이용해서 구현해야 한다.
	· 메타 정보 검색 기능은 HTTP HEAD Method를 이용해서 구현해야 한다.
	· 기능 목록 조회 기능은 HTTP OPTIONS Method를 이용해서 구현해야 한다.
	· Custom HTTP Method는 사용하지 않아야 한다.
	· Custom HTTP Header는 사용하지 않아야 한다.
	· 하나의 리소스에 대해서 반복적으로 호출해도 동일한 상태여야 한다.
	· 하나의 리소스를 생성할 때에 POST Method가 한번만 호출되는 것을 보장해야 한다.
	· 리소스에 대해서 Last-Modified와 ETag를 관리 한다.
	· If-Match, If-None-Match Header를 처리할 수 있어야 한다.
	· If-Modified-Since, If-Unmodified-Since Header를 처리 한다.
리소스의 현재 상태 표시 (Representation)	· 리소스에 대해서 GET으로 조회한 내용과 POST나 PUT으로 생성하거나 수정하는 내용의 형식이 일치해야 한다.
	· 응답에는 Representation에 대한 메타 정보를 Header에 포함한다.
에러 코드와 처리 방식(Error Condition)	· 클라이언트에서 잘못된 요청이 왔을 경우 Error 처리를 하고, Error Condition 상태 코드를 포함하여 응답해야 한다.
	· 클라이언트에서 요청한 요청에 대하여 서버에서 문제가 발생했을 경우 올바른 응답을 송신해야 한다.
	· 클라이언트에서 요청한 요청에 대하여 성공적으로 작업을 수행했을 경우, 올바른 응답을 해야 한다.
	· Header의 상태 코드와 Body의 메시지는 동일한 의미의 내용이어야 한다.
	· 에러 코드와 처리하는 방식은 HTTP에 정의된 규격에 맞추어 예외 처리 프로세스를 작성 한다.

- 개발된 Open API를 사용자들에게 Web 방식으로 활용 할 수 있도록 제공 (등록, 검색, 활용, 사용자 관리 기능 등)

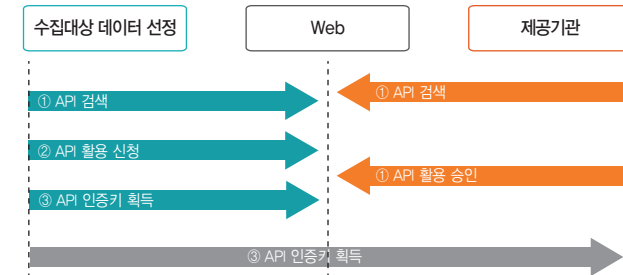
〈 Open API 웹 방식 관련 기술 고려사항 〉

기능	고려사항
등록	Open API 등록 · 데이터 제공을 위해 개발된 Open API는 홈페이지 또는 포털에 등록하여 사용자에게 제공해야 한다. · Open API 등록 시 API명, 유형, 설명, 공유범위, 분류체계, 저작권 등 상세 내용을 입력한다.
	등록 관리 · 관리자가 등록된 Open API에 대한 상세 등록 내용을 조회 및 확인해야 한다. · 관리자가 등록된 Open API를 확인한 후 승인 과정을 통해 서비스를 개시해야 한다.
검색	디렉토리 검색 · 제공기관별 디렉토리 서비스를 제공해야 한다. – 공공기관은 정부부처 조직체계를 기준으로 디렉토리 구성 – 민간기관은 산업분류 체계를 기준으로 디렉토리 구성 · 사용자별 디렉토리 서비스를 제공한다. – 사용자가 공공 및 민간 디렉토리에서 자주 사용하거나 관심 있는 디렉토리를 선택하여 별도 디렉토리를 구성
	인기 검색 · 일간/주간/월간 Open API 검색 순위를 제공한다. · 검색어별 관련 정보는 통합검색과 연계하여 세부 매뉴얼, 디렉토리(분류)별 등으로 제공한다.
	통합 검색 · 해당 검색어와 관련된 정보를 매뉴얼, 디렉토리 (분류)별로 제공해야 한다. · 연관 검색 기능을 제공한다. · 상세 검색어 설정 기능을 제공해야 한다.
	검색 결과 · 사용자 선택에 따라 검색 결과를 정확도순, 날짜순으로 제공한다. – 검색 결과는 원문 데이터의 제목을 기준으로 제공 – 원문 데이터 분류, 제공기관, 매체유형, 상세 정보 등 제공 · 검색된 Open API를 선택할 경우 결재 또는 활용신청 페이지로 연결해야 한다. – 제공되는 Open API가 유료일 경우 결재 페이지로 연결 – 활용신청이 필요한 경우 활용신청 페이지로 연결
	활용 활용 신청 · 활용 신청서 작성 및 신청 기능을 제공해야 한다. – 해당 Open API의 활용 허락 범위 안내 – 사용자 정보, 활용 유형, 활용목적 등을 등록하여 신청하는 양식 제공 · 신청자가 활용 신청에 대한 처리결과를 확인할 수 있도록 기능을 제공해야 한다.

기능		고려사항
활용	활용 승인 심의	· 제약없이 공개된 Open API의 경우 제공기관 담당자의 별도 확인 없이 신청과 동시에 활용할 수 있도록 자동 승인 기능을 제공해야 한다.
		· 제공기관의 활용 승인이 필요한 경우 활용신청에 대해 제공기관 담당자가 확인 후 승인 처리 기능을 제공해야 한다.
		· 활용 신청에 대한 승인 거부 시 사유를 입력하여 신청자에게 전달 기능을 제공한다.
	활용 관리	· 등록된 Open API의 활용 내역을 조회할 수 있는 기능을 제공해야 한다.
		· 유료로 제공되는 Open API에 대한 결제 내역 조회기능을 제공해야 한다.
		· Open API 활용 내역 및 결제 내역에 대한 통계 기능을 제공한다.
사용자 관리	사용자 등록	· 사용자는 크게 개인과 기관으로 구분한다. - 개인은 14세 미만 어린이, 일반인, 외국인으로 구분
		· 사용자는 입력된 개인정보를 기준으로 가입되어 있는지를 확인해야 한다.
		· 본인 및 기관 확인을 위한 실명인증 및 기관인증을 해야 한다. - 실명인증은 가상식별코드 실명확인서비스, 아이핀 인증 등 개인정보보호 방안 제공
		· 개인정보 수집 · 활용 및 사이트 이용에 대한 약관 동의 및 정보 입력 기능을 제공해야 한다.
	사용자 승인	· 회원 가입한 개인에 대한 확인 및 변경 관리 기능을 제공해야 한다.
		· 기관 담당자에 대한 승인 및 변경 관리 기능을 제공해야 한다.
		· 개인 및 기관에 대한 접근 권한 부여 관리기능을 제공해야 한다.
	마이 페이지	· 사용자가 개인 정보 변경 시 등록된 정보를 변경 및 삭제 기능을 제공해야 한다.
		· 사용자가 관심 있는 디렉토리 및 메뉴를 등록하고 관리할 수 있는 기능을 제공한다.
		· 사용자가 등록한 Open API에 대한 목록, 활용 및 결제 현황에 대한 조회 기능을 제공해야 한다.

참고

Open API 이용 예시



- ① API 등록: 제공기관에서 데이터 제공을 위한 API를 개발하여 Web에 등록
- ② API 검색: 이용자는 Web에서 필요한 데이터를 제공하는 API 검색
- ③ API 활용 신청: 이용자는 API 검색 결과 필요한 데이터를 제공하는 API에 대해 Web을 통해 활용 신청
- ④ API 활용 승인: 제공기관은 사용자 API 활용 신청에 대해 검토 후 Web을 통해 승인
- ⑤ API 인증키 취득: API 활용 승인을 받은 사용자는 해당 API에 대한 인증키를 Web에서 취득
- ⑥ API 호출: 사용자는 인증키를 포함한 API 호출을 통해 제공기관의 제공 데이터 수집

데이터 사용 인증

데이터의 안전한 사용을 위하여 적용중인 인증방식(Oauth)에 대한 기술 고려사항

※ OAuth는 어플리케이션끼리 별도의 인증이 필요 없이 여러 어플리케이션을 통합하여 사용 가능하도록 표준화된 인증방식임

〈 사용자 인증 기술 고려사항 〉

기능	고려사항
Request Token 발급	· 클라이언트를 서비스 제공자에게 등록하는 기능을 제공해야 한다.
	· 클라이언트가 서비스 제공자에게 Request Token 요청 기능을 제공한다.
	· 서비스 제공자는 인증되지 않은 Request Token 발급 기능을 제공해야 한다.
사용자 인증	· 클라이언트가 사용자를 서비스 제공자 인증화면으로 안내하는 기능을 제공한다.
	· 사용자가 Request Token을 인증하는 기능을 제공해야 한다.
Access Token 요청 및 발급	· 클라이언트가 인증된 Request Token을 가지고 권한서버에게 Access Token을 요청하는 기능을 제공해야 한다.
	· 권한서버는 Access Token을 클라이언트에게 발급하는 기능을 제공해야 한다.
데이터 접근	· 클라이언트는 Access Token을 사용하여 자원서버에 접근하는 기능을 제공해야 한다.

※ 자원서버(Resource Server): OAuth가 보호하는 데이터를 호스팅하는 서버

※ 자원 소유자(Resource Owner): 자원서버에서 호스팅하는 데이터에 대한 접근을 허가하는 서비스 제공자(Service Provider)

※ 클라이언트(Client): API를 이용하여 보호된 자원에 작업을 요청하는 어플리케이션 또는 웹서비스

※ 권한서버: 클라이언트에게 액세스 토큰을 발행하는 서버

참고

데이터 활용 기술 참조 사이트

데이터 교환(Open API)

- 공공데이터 포털 : Open API 활용사례, 개발가이드, 실제 데이터 확인
□ <https://www.data.go.kr/#/L3B1YnMvcG90L0lyb3NPcGVuQXBP5W50cm8kQF4wNTFhNQ==>
- 다음 : Open API 안내, 참조 및 활용 자료 제공
□ <http://dna.daum.net/apis>
- 네이버 : Open API 이용 가이드 및 참고자료 제공
□ <http://developer.naver.com/wiki/pages/OpenAPI>

데이터 사용 인증(OAuth)

- OAuth 2 공식 홈페이지
□ <http://oauth.net/2/>
- Facebook : JavaScript, Android, iOS, PHP를 위한 페이스북 SDK
□ <http://developers.facebook.com/docs/reference/javascript/>
□ <http://developers.facebook.com/docs/reference/androidsdk>
□ <http://developers.facebook.com/docs/reference/iossdk>
□ <http://developers.facebook.com/docs/reference/php>
- Google : Java, Objective-C, PHP, Python, Ruby, JavaScript를 위한 Google API 클라이언트 라이브러리
□ <http://code.google.com/p/google-api-java-client/>
□ <http://code.google.com/p/google-api-objectivec-client/>
□ <http://code.google.com/p/google-api-php-client/>
□ <http://code.google.com/p/google-api-python-client/>
□ <http://code.google.com/p/google-api-ruby-client/>
□ <http://code.google.com/p/google-api-javascript-client/>
- MicroSoft : MS Developer Network의 OAuth 2.0 소개
□ <http://msdn.microsoft.com/en-us/library/hh243647.aspx>

– 다음 : 다음 개발자네트워크의 OAuth 안내, 활용하기, 참고자료 제공

□ <http://dna.daum.net/apis/oauth>

– 네이버 : 네이버 개발자센터의 OAuth API 명세, 튜토리얼 제공

□ <http://dev.naver.com/openapi/apis/oauth/guide>

※ 개발자들이 OAuth를 활용하여 쉽게 사용자 인증을 구현할 수 있도록 구글, 페이스북, 다음, 네이버 등에서 제공하는 OAuth 지원 라이브러리를 참조하여 개발



빅데이터 활용 단계별 업무절차 및 기술 활용 매뉴얼 (Version 1.0)

– 빅데이터 수집 · 저장 · 분석 · 이용 단계별 업무절차 및 기술 활용 시 고려사항 제공 –

발행일 | 2014년 5월 20일

발행처 |



미래창조과학부

427-140 경기도 과천시 관문로 47, 4동

www.msip.go.kr



한국정보화진흥원

무교청사 : 100-775 서울특별시 중구 청계천로 (무교동 77번지) 14 NIA빌딩

등촌청사 : 157-715 서울특별시 강서구 공항대로 489 (등촌동)

www.nia.or.kr

편집 · 디자인 | (주)아이디어스토리지
