제 5장

빅데이터 프로젝트 가이드라인

2015.06
조완섭
충북대학교 경영정보학과
대학원 비즈니스데이터융합학과
wscho@chungbuk.ac.kr
043-261-3258
010-2487-3691

- 본 자료는 "빅데이터 업무절차 및 기술활용 매뉴얼 (Ver 1.0), NIA, 2014.03"을 참고하여 정리한 것임

목차

- 배경 및 개요
- 데이터 수집
- 데이터 저장관리
- 보안관리
- 품질관리
- 데이터 분석
  - 가시화
- 분석결과의 활용과 서비스

배경

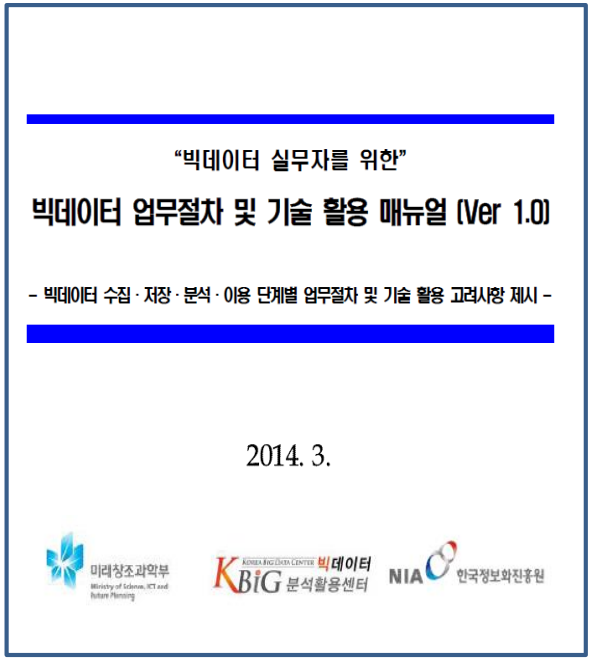
- 빅데이터 시대로의 진입
  - 2011년 맥킨지 보고서
- 빅데이터가 ICT 분야의 새로운 패러다임, 신성장동력
  - 정부3.0으로 공공분야 빅데이터 관심증대
  - IT 기업들은 빅데이터로의 사업확장
  - 비IT기업들도 빅데이터 활용 비즈니스 혁신에 관심
- 선진국, 글로벌 기업 위주로 빅데이터 경쟁심화
  - 미국, 영국, 일본, 싱가폴
  - 중국 핀테크 기업

배경

- 우리나라는 선진국에 비하여 빅데이터 경쟁력 하락
  - 정부 및 공공기관, 지자체 노력에도 불구하고 2015년 OpenData Barometer 국제지표 17위 하락 (2014년 12위)
  - 빅데이터 구축 및 활용경험이 일천하고, 마땅한 지침서나 전문가도 부족한 상황
  - 거버넌스 구축 없이 활용만 강조되는 분위기
- 데이터를 소홀이 하는 문화
  - 데이터 분석 기반의 과학적 의사결정 문화 미흡
  - 조직의 데이터 분석역량이 미흡

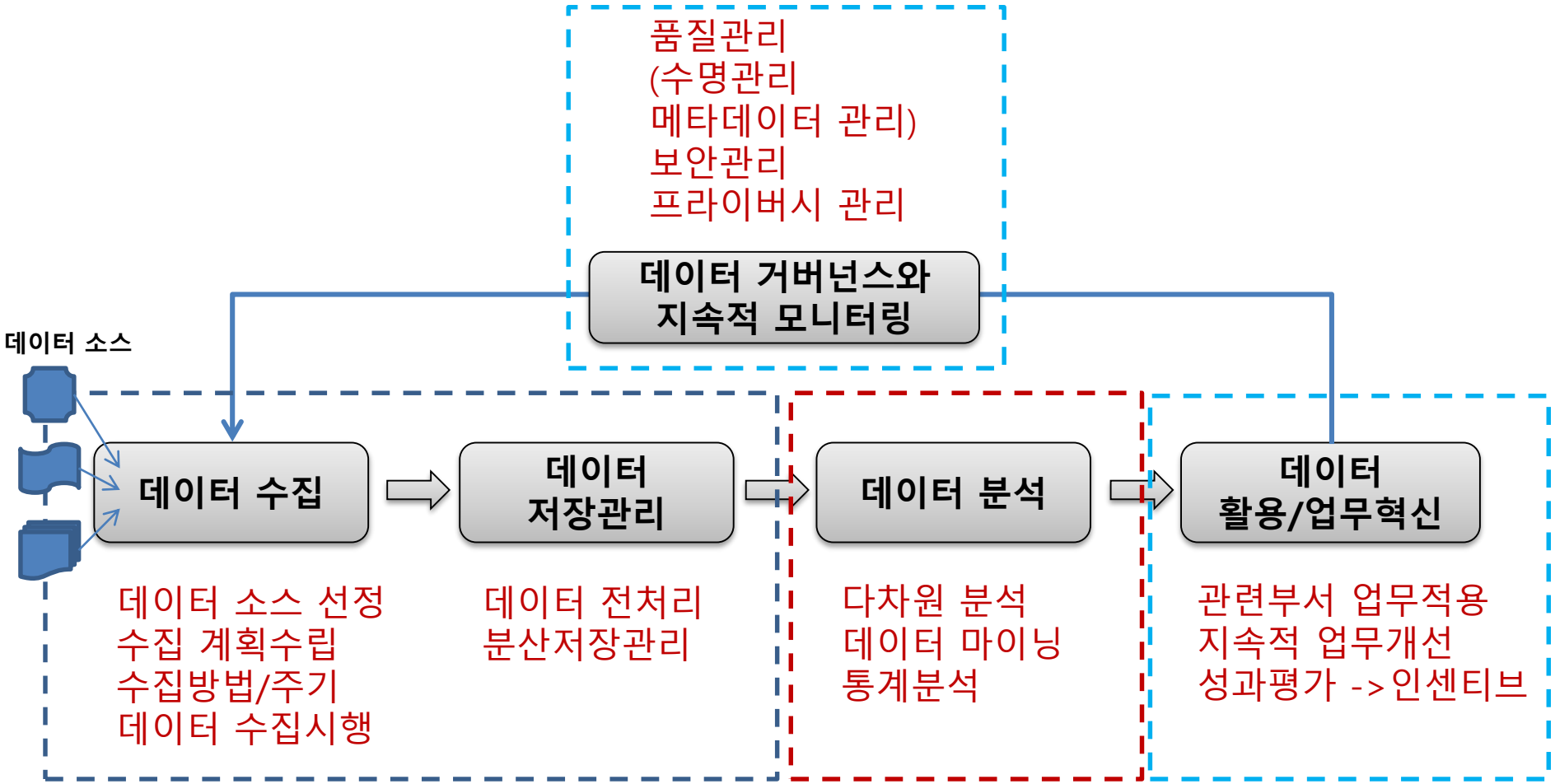
개요

- 빅데이터 활용지침서 (Nia, 2014)
  - 공공과 민간에서 빅데이터를 활용하고자 하는 실무자들이 알아야 할 단계별 업무절차 및 관련기술 소개
  - 빅데이터 프로젝트 수행시 고려사항
    - 빅데이터를 활용한 서비스 기획(rfp 작성)
    - 분석 플랫폼의 구축과 운영
    - 사업관리
    - 데이터 활용 업무혁신 방안



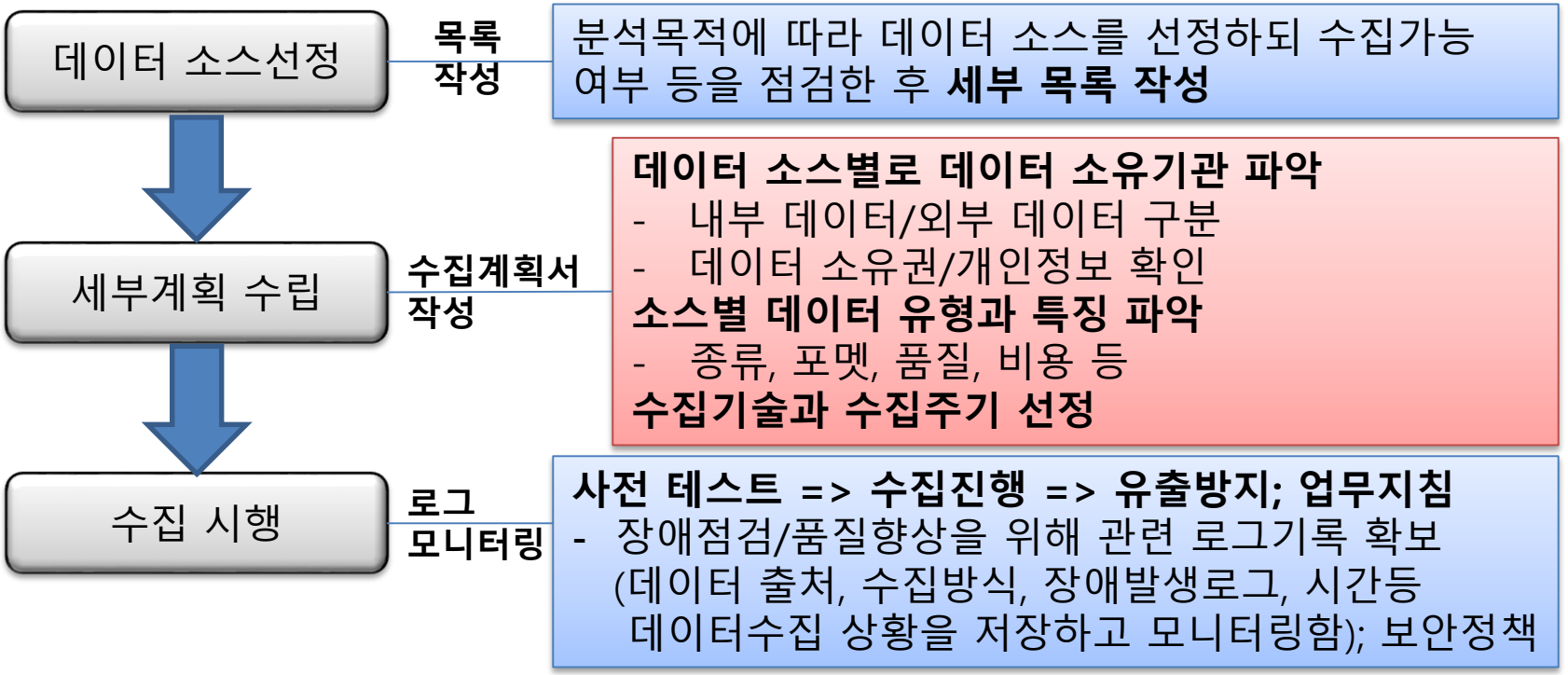
개요

- 빅데이터 사업수행의 수행과 활용 절차



1. 데이터 수집

- 정의
  - 조직 내부·외부의 다양한 데이터를 일괄 · 실시간으로 수집하는 과정 (기술, 업무)
- 절차



1. 데이터 수집

- 데이터 유형

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

담당자 이름	담당자 직위	전화 번호	주소
한석규	영업 사원	(0582)575-5776	가장동 78-3
황영순	대표 이사	(02)681-6889	서초구 방배동 883-11
조자룡	대표 이사	(02)989-9889	강서구 내발산동 318
구재석	영업 사원	(032)76-4568	남구 연수동 208-16
최영희	영업 과장	(042)92-3778	서구 도마동 110-6
손미선	영업 사원	(02)211-1234	서대문구 남가좌 1동 121
장선희	마케팅 2과장	(02)111-2954	영등포구 당산동 3가 16

```
<?xml version="1.0"?>
<quiz>
 <qanda seq="1">
  <question>
   Who was the forty-second
   president of the U.S.A.?
  </question>
  <answer>
   William Jefferson Clinton
  </answer>
 </qanda>
 <!-- Note: We need to add
  more questions later.-->
</quiz>
```

XML

1. 데이터 수집

- 데이터 수집주기
  - 배치 (간격은 ?) 수집과 (준)실시간 수집으로 구분하여 적절한 수집기술 선택
    - 데이터의 종류와 크기, 데이터 발생 빈도주기, 분석주기, 시스템 및 네트워크 부하 정도 등을 고려하여 기술선택
  - 일정기간 샘플 데이터 수집 필요
    - 데이터 량을 점검한 후에 수집주기와 서버 용량 결정
  - 스트림 데이터의 실시간 수집 (IoT)
    - 데이터 폭증에 대비해야 함
    - 중복 데이터 필터링 기술 활용 (예: 방의 온도 센서)
    - 인메모리 처리기술 활용 필요

1. 데이터 수집

- 데이터 수집기술

구 분	특징	비고
Crawling Web Robot	- SNS, 뉴스, 웹 정보 등 인터넷상에서 제공되는 웹문서·정보 수집 (URL List => 데이터 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

1. 데이터 수집

- 데이터 수집기술

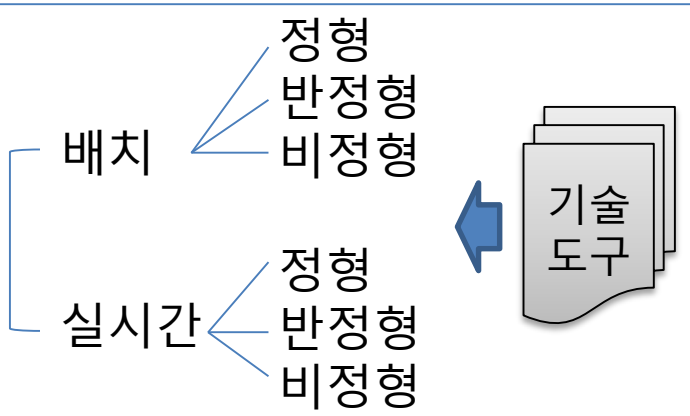
Arriving machine data is processed at rates of up to 1 million records/second/CPU core

Streaming	- 인터넷에서 음성, 오디오, 비디오 데이터를 실시간으로 수집할 수 있는 기술 (종류: SQLstream, ETL for IMDG)	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 , **TeraStream for Hadoop**	RDB 기반 데이터 수집

1. 데이터 수집

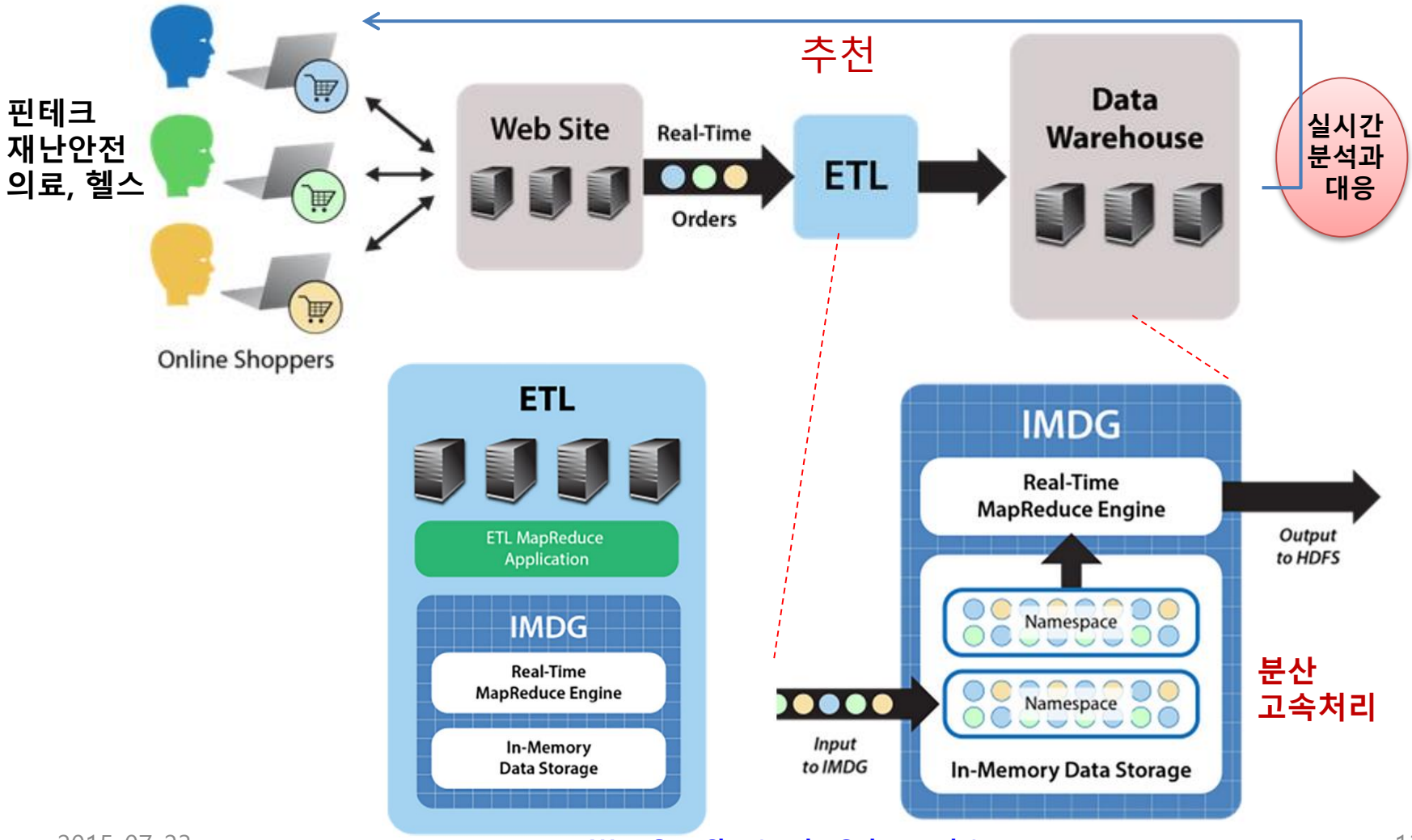
- 빅데이터 유형에 따른 수집기술

<데이터 유형에 따른 수집 기술>

데이터 유형	데이터 종류	수집 기술의 선택
정형 데이터	RDB, 스프레드 시트	
반정형 데이터	HTML, XML, JSON, 웹문서, 웹 로그, 센서 데이터	
비정형 데이터	소셜 데이터, 문서(워드, 한글), 이미지, 오디오, 비디오	

1. 데이터 수집

- 실시간 데이터 수집의 필요성 증대



1. 데이터 수집

- 60초 동안에 발생하는 events

출처 : http://gizmodo.com/how-much-happens-on-the-internet-every-60-seconds-950463150

1. 데이터 수집

- 사전 테스트
  - 수집 계획에 따라 수집주기와 기술을 적용, 사전 테스트 진행
  - 네트워크 트래픽 문제, 데이터 누락여부, 정확성 (원본과 수집된 데이터 비교), 보안성 등을 점검하여 필요시 수집방법 보완 변경
- 데이터 수집 시행
  - 수집을 진행하되 향후 장애 점검 등을 위해 관련 로그 기록을 확보함
  - 수집당시 상황을 정보 : 데이터의 출처, 수집방식, 장애발생 여부와 시스템 로그, 시간 등의 정보
- 데이터의 수집 후 처리
  - 데이터 수집 후 저장된 데이터에 대한 외부인 접근방지 및 유출시 대처방안 등과 관련된 업무지침 마련

1. 데이터 수집

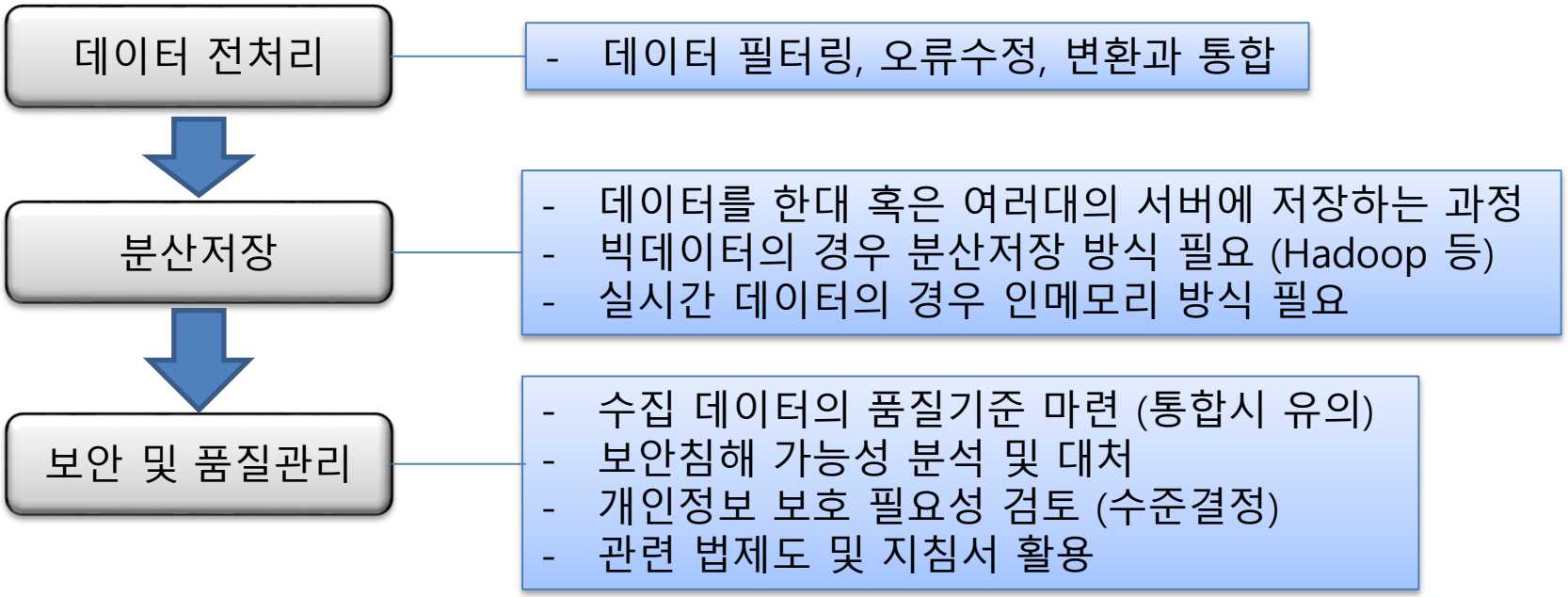
- 데이터 수집기술 활용시 고려 사항
  - Crawling, FTP, OpenAPI, 실시간 streaming, Log aggregator, RDB aggregator 등
  - 빅데이터 업무절차 및 기술활용 매뉴얼 (NIA)
    - 정보설정 기능
    - 수집 에이전트 기능
    - Collector 기능
    - 기타 기능
  - 매뉴얼의 주요 내용
    - 각 수집기술별로 고려할 사항을 정리함
    - Page 11 ~ 17

1. 데이터 수집

- 빅데이터 수집시 주의사항
  - 빅데이터 수집시에는 데이터의 질, 수집 기술, 데이터 보안 및 개인정보보호 문제 등 다양한 부분을 고려해야 함; 전문가의 조언 필요
  - 데이터 수집 활동은 분석 결과의 질을 좌우하는 중요한 과정임; 분석에 필요한 데이터 항목들을 반드시 포함해야 하고, 품질도 원하는 수준으로 확보하는 것이 중요함
  - 수집기술은 다양한 데이터 소스로부터 다양한 유형의 데이터를 수집하기 위해 확장성, 안정성, 실시간성 및 유연성을 확보해야 함 (도구사용으로 체계화)

2. 데이터 저장관리

- 정의
  - 데이터 전처리, 분산저장, 보안 및 품질관리 등을 수행하는 단계
- 업무절차



2. 데이터 저장관리 - 전처리

- 전처리 기술

방식	설 명
데이터 여과 (Filtering)	- 오류 발견, 보정, 삭제 및 중복성 확인 등의 과정을 통해 데이터 품질을 향상 시키는 기술 (예: 센서의 경우 동일한 값 출력=>압축)
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

* 평활화 : 데이터에 포함된 잡음제거를 위해 추세를 벗어나는 데이터를 적절한 값으로 변환함

2. 데이터 저장 - 정제

- 결측치 처리방법

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

2. 데이터 저장 - 정제

- 잡음의 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

잡음발생 원인 : 센서의 작동실패, 데이터 입력오류, 데이터 전송문제, 기술적인 한계, 데이터 속성값의 부정확성 등

2. 데이터 저장 - 축소

- 불필요한 데이터 축소=>분석효율성 제고(고유 특성은 유지)

단계적 회귀분석 (stepwise regression)
- 독립변수를 하나씩 추가/삭제하면서 최적의 모형을 만들어 나감

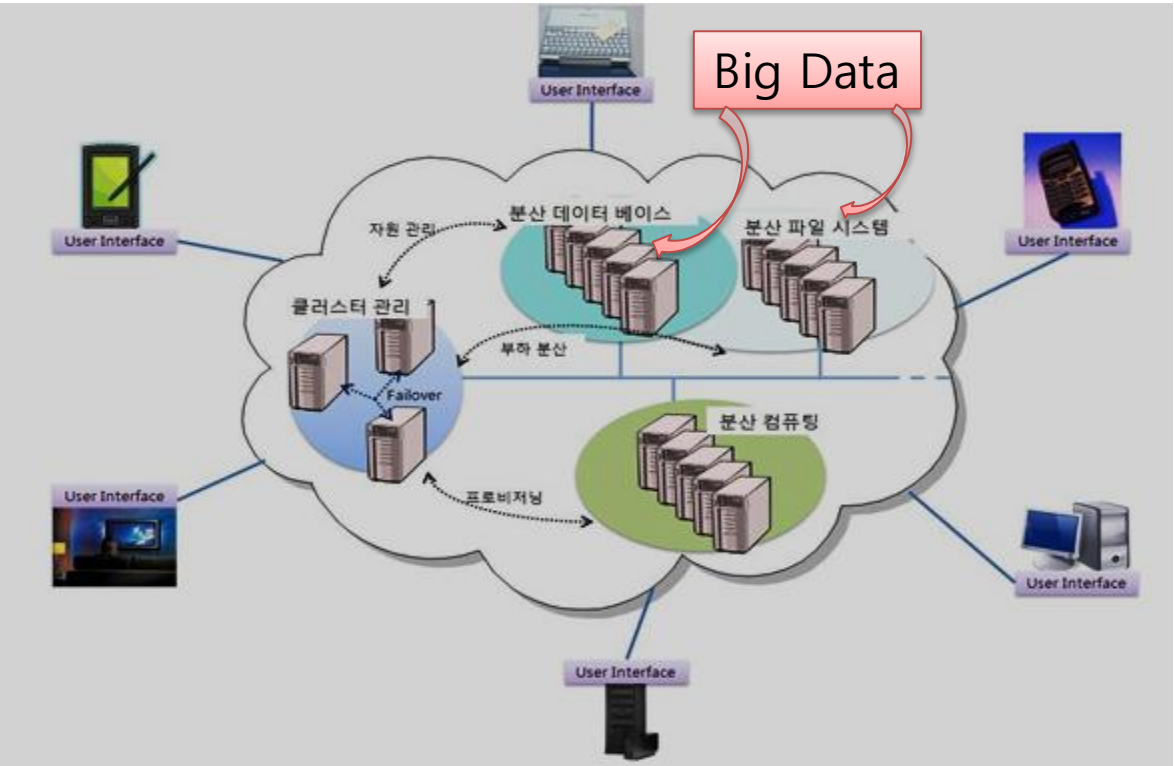
축소 방식		설명
차원 축소	분석에 필요 없거나 중복 항목 제거	- Stepwise forward selection, Stepwise backward elimination 등 활용
데이터 압축	데이터 인코딩이나 변환을 통해 데이터 축소	- lossless(BMP 포맷), lossy(JPEG 포맷) 등 방법 적용
Discrete wavelet transform (DWT)	선형 신호 처리	- 수는 다르지만 길이는 같은 벡터(wavelet coefficients)로 변환 - 여러 개의 벡터 중에서 가장 영향력이 큰 벡터를 선택해서 다른 벡터들을 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. 데이터 저장 – 전처리/후처리

- 데이터 전처리 관련 기술 활용시 고려사항
  - 데이터 전처리
    - 데이터 필터링 기술 활용시 고려사항
    - 데이터 유형 변환시
    - 데이터 정제시
  - 데이터 후처리
    - 데이터 통합시
    - 데이터 변환시
    - 데이터 축소시
  - Page 23~26 참고

2. 데이터 저장 - 분산저장

- 빅데이터 저장
  - 수집된 데이터는 한대의 컴퓨터에 저장하거나 (작은 경우) 혹은 여러대의 컴퓨터 (클라우드)에 분산저장함
  - 실시간 처리가 필요한 경우에는 메인 메모리에 저장함



2. 데이터 저장 - 분산저장

- 데이터 저장계획 수립
  - 데이터 유형에 따른 저장방식 선정
    - RDB, NoSQL, 분산 파일시스템, IMDG 등
  - 데이터 수집량에 따라 저장공간 산정
    - RDB는 제조업체 문의; scale-up / scale-out 확장성 확인
    - NoSQL은 scale-out 방식으로 peta-byte 이상까지 확장 (복제고려)
  - 계획서에는 데이터 유형에 따른 수집주기, 저장방식, 보관주기, 벡업 방식, 저장공간 확장방안 등을 세부적으로 명시

< 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

2. 데이터 저장 - 분산저장

- 데이터 저장기술

구 분	특징	비고
RDB	- 관계형 데이터를 저장하거나, 수정하고 관리할 수 있게 해주는 프로그램 - SQL 문장을 통하여 데이터베이스의 생성, 수정 및 검색 등 서비스를 제공 최대 Terabyte씩 확장가능	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

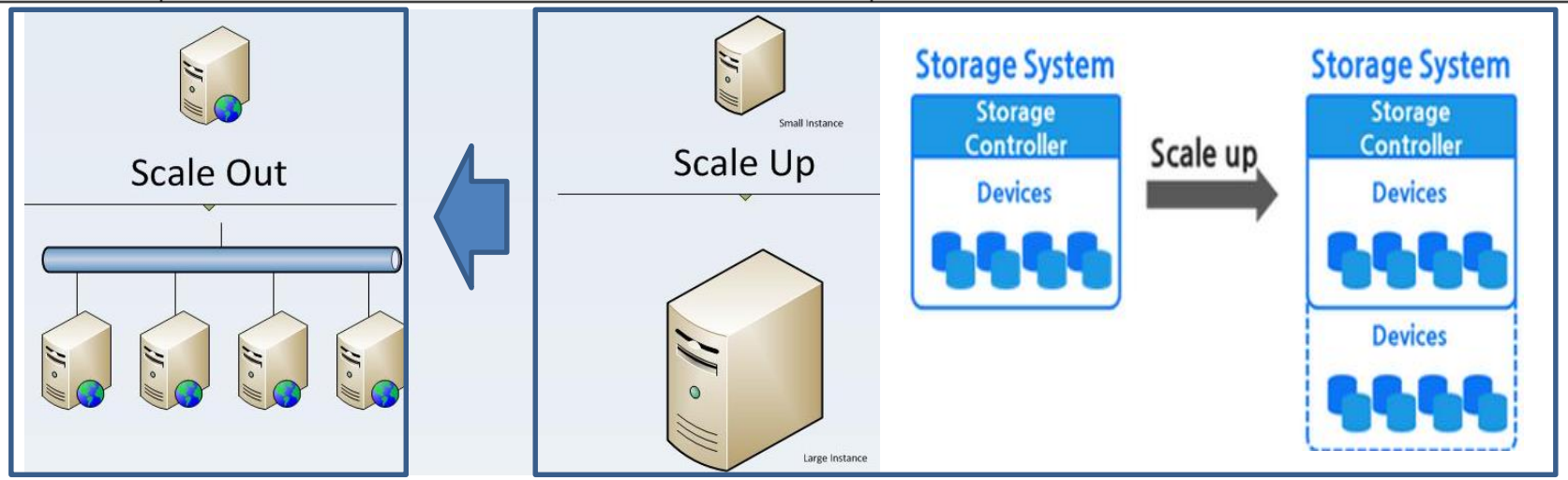
2. 데이터 저장 - 분산저장

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 지원 - 범용 x86서버의 CPU, RAM 등을 사용하므로 장비 증가에 따른 성능 향상 - 수 TB~ 수백 PB 이상의 데이터 저장 지원	HDFS
인메모리 데이터 그리드	- 분산된 서버의 메인 메모리에 데이터 저장 - 다수의 컴퓨터로 고속 병렬 처리 (고성능 실시간 처리) - 필요한 경우 하드 디스크 DB와 연동 및 동기화	IMDG

2. 데이터 저장 - 분산저장

- 저장공간의 확장방식

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



2. 데이터 저장 – 시험운영 및 모니터링

- 구축 및 시험운영
  - 계획에 따라 DB를 구축하고 운영에 필요한 주요 기능을 테스트함
- 시행 및 모니터링
  - 주기적으로 데이터 저장관련 에러, 여유공간 등을 실시간으로 모니터링하고 문제발생시 대응체계 마련
    - RDB의 경우 인덱스 공간을 감안하여 여유공간 확보
    - NoSQL, ,Hadoop의 경우 복제파일 운영 고려
  - 저장공간이 일정수준 이상 사용된 경우 미리 scale-out 방안 강구