



미래창조과학부

NIA

한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY

창조경제 활성화를 위한 빅데이터 역량강화

K-ICT 빅데이터센터 인프라 활용가이드

임동진·김찬수·김이환·신은비

2015. 11. 6(금)

2e 투이컨설팅



미래창조과학부

NIA 한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY

- I. 사업 추진 개요
- II. K-ICT 빅데이터센터 운영현황
- III. K-ICT 빅데이터센터 인프라

Chapter 1.

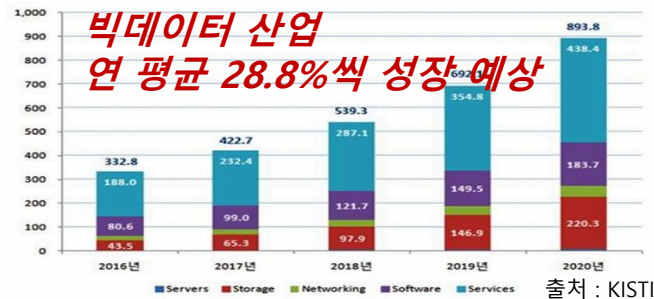
사업 추진 개요



사업 추진 배경 및 목적

'14년부터 분석 및 기술 교육 콘텐츠 개발, 기구축된 분석 실습 교육 콘텐츠 고도화, 빅데이터 분석전문가의 저변확대를 위한 빅데이터 페스티벌을 사업범위로 하고 있습니다.

산업 예측

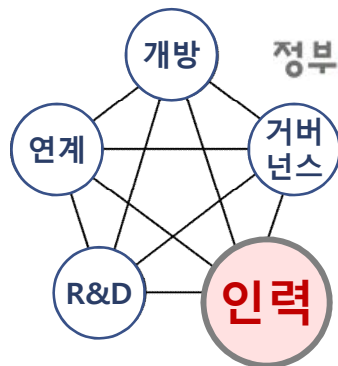


"사업기회가 넘치는데 일할 사람이 없네!!!"

우리 현실

- 산업계에서 요구되는 데이터 과학자급 전문 인력 부족
- 대학(원)에서 활용할 수 있는 빅데이터 교육실습을 위한 실용적 데이터와 분석사례 부족

정책



빅데이터 산업발전 전략

- 미래부 중점추진 사업
- 정부 3.0 주요 핵심 사업
- 데이터 전문인력 양성 및 일자리 연계 필요

✓ 대학(원) 내 실무 중심의
데이터 과학자 양성
전문과정 운영 지원

✓ 분석을 위한 데이터 셋,
분석 모델 및 기법 등
분석 툴킷 개발/제공

✓ 데이터 분석에 대한 사회적
관심과 활성화를 높일 수 있는
빅데이터 페스티벌
활성화

사업 내용

빅데이터 전문인력 양성을 위해 빅데이터 분석 교육·기술 콘텐츠 개발 및 고도화를 통한 대학(원) 실습교육을 지원하고, 저변확대를 위한 빅데이터 페스티벌을 수행 중에 있습니다.

빅데이터 분석 프로세스 기반
교육·기술콘텐츠 개발 및 고도화

대학(원) 대상
빅데이터 실습 교육 지원

빅데이터 저변확대

농산물



소비



소셜



교통



제조



교육·기술 교육
콘텐츠 개발

- 기본/응용 단계 구분
- 데이터 셋 기반 활용 시나리오
- 활용 가능한 실무 기술 가이드

분석 실습 교육
콘텐츠 고도화

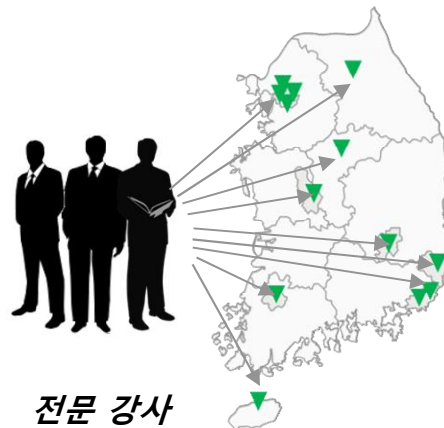
- 데이터 전처리 강화
- 다양한 방법론 적용
- 분석 스토리텔링 강화

실습 매뉴얼 및
웹 매뉴얼 제작



K-ICT의 빅데이터 분석 실습
인프라 및 콘텐츠를 활용하여
빅데이터 전문강사를 통한

전국 대학(원) 대상
약 750명의
분석 실습 교육 실시



전문 강사

✓ **UniBiG 협의체 구성**
- 교육지원 협력을 위한 산학연
공동 협약

✓ **한국통계학회 추계논문 발표회**
- 한국정보화진흥원 기획세션



사단법인 한국통계학회
THE KOREAN STATISTICAL SOCIETY

✓ **강원창조경제혁신센터**
- 빅데이터 분석 전문가 멘토링
(One-Day Job fair 개최)



미래창조과학부

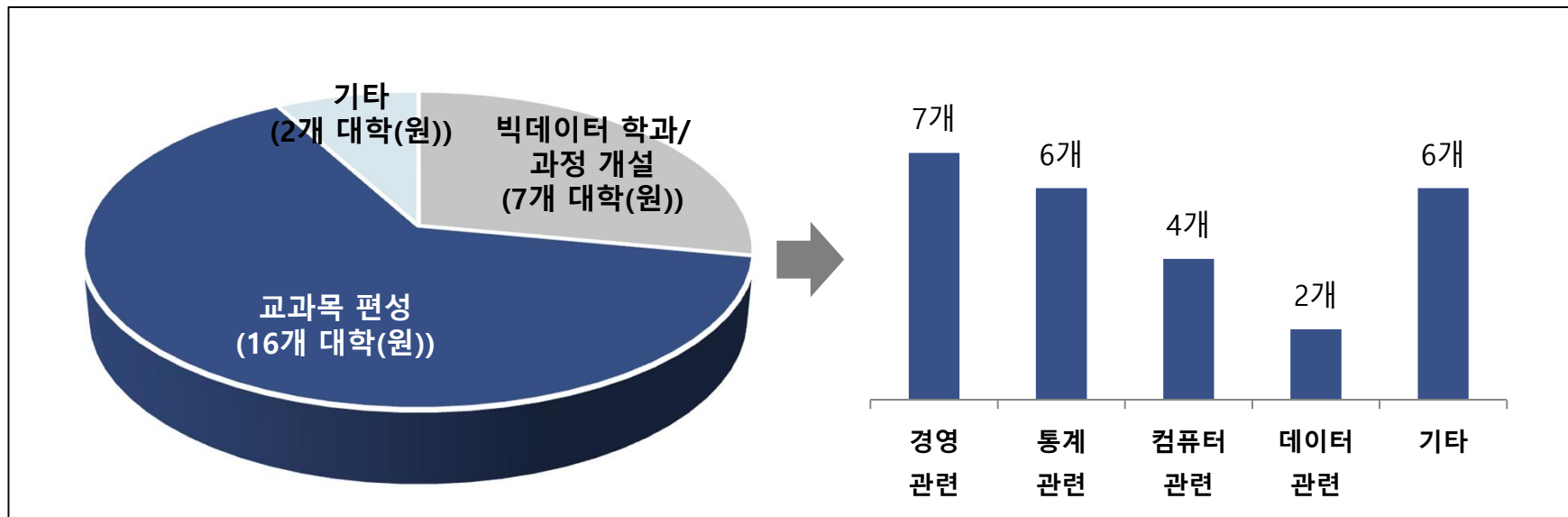
NIA 한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY

2e 투이컨설팅

빅데이터 분석 전문가 교육 수요조사 결과(1/5)

2015년 2학기 중 빅데이터 관련 학과나 교과목을 운영중인 대학(원)은 25개 대학(원)이며, 추가적으로 16개 대학(원)이 개설을 계획 중에 있습니다

빅데이터 관련 학과·과정(과목)의 개설 현황

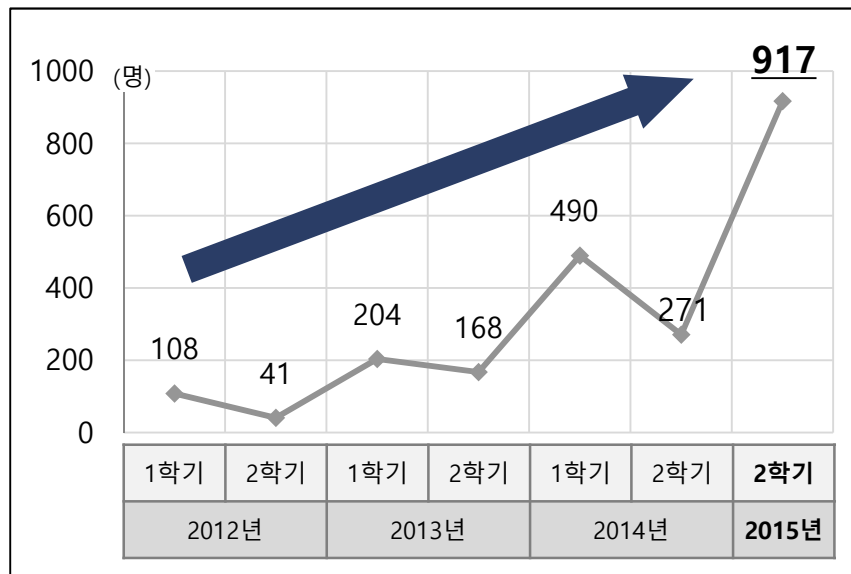


- 총 25개 대학(원)의 학과에서 빅데이터 관련 학과 신설 혹은 교과목 개설을 통해 빅데이터 분석 교육을 진행 중에 있음
- 빅데이터 관련학과 신설 혹은 교과목을 개설한 대학(원) 학과를 살펴보면, 경영관련 학과, 통계관련 학과, 컴퓨터 관련학과에서 관심을 많은 보이고 있음
- 기타 학과로는 문헌정보학과, 멀티미디어학과, 의료IT마케팅학과가 있음

빅데이터 분석 전문가 교육 수요조사 결과(2/5)

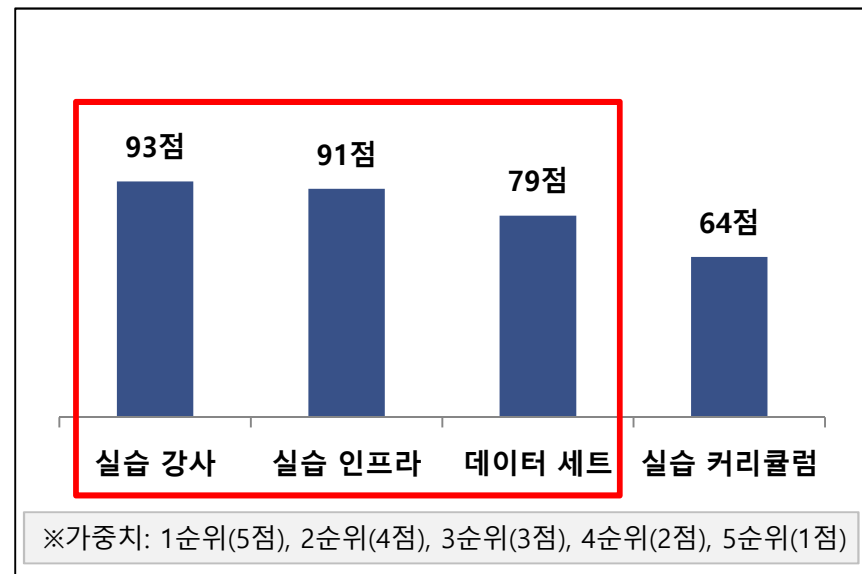
빅데이터 분석교육 수요조사 결과를 보면, 빅데이터 학과 및 과정의 수요는 꾸준히 증가하고 있으며, 실습강사, 실습인프라, 데이터 셋 및 실습 커리큘럼에 대한 니즈가 있습니다.

빅데이터 학과·과정(과목)의 수강생수



- 조사결과 2012년 1학기부터 개설되어 운영되고 있으며, 과정/과목 수강생은 지속적으로 증가하고 있음
- 전년도의 동일학과와 비교하면, **수강생 수가 약 3.4배로 급격히 증가함**

실습 지원 필요 부분

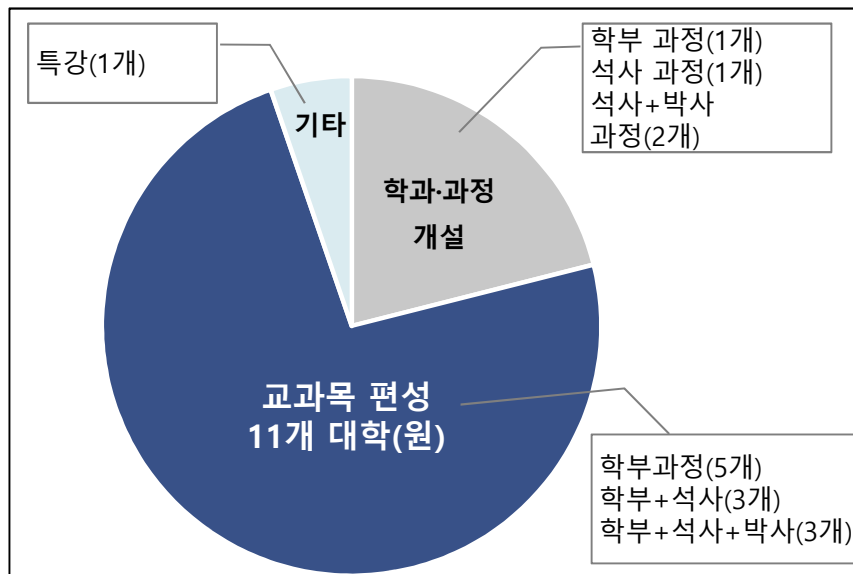


- 실습강사와 실습 인프라에 대한 지원이 우선적으로 필요한 것으로 나타남
- 본 사업을 통해 실습강사, 인프라, 데이터 셋 제공함
- 실습 커리큘럼은 빅데이터 역량모델 개발을 통해 제공 가능

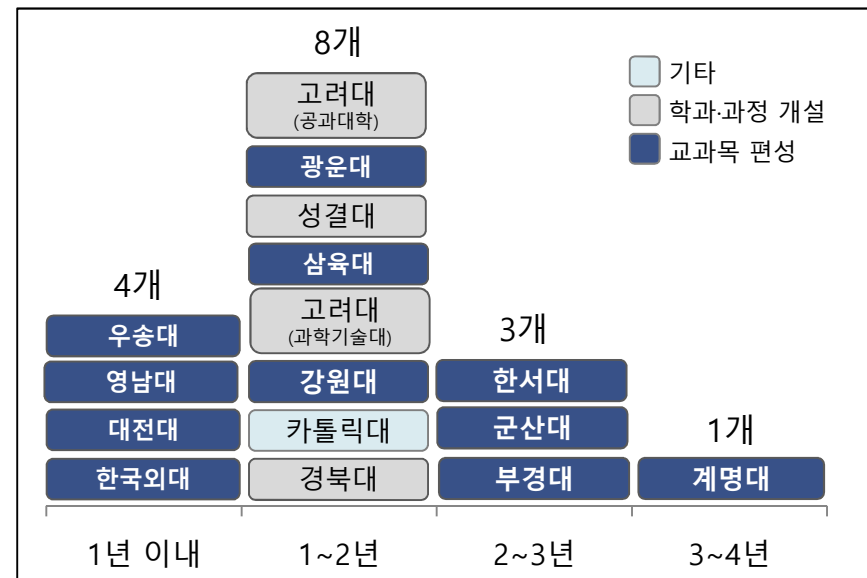
빅데이터 분석 전문가 교육 수요조사 결과(3/5)

2015년 2학기 현재, 빅데이터 관련 학과가 개설되지 않은 16개 대학(원)은 향후 1~3년 이내 개설을 계획 중에 있으며, 이에 대한 지원 방안 수립이 필요할 것으로 보입니다.

교육 과정 개설 희망 형태 및 대상



교육과정 개설 희망 시기



- 교과목 편성 계획 중인 16개 대학(원) 중 10개 대학(원)이 학부 3~4학년 대상으로 교육과정 개설 계획을 가지고 있으며, 여러 과정을 동시에 진행하려는 학교도 다수 있음
- 교과목 편성 계획 중인 대학(원)의 75%는 2년 이내 교육과정 개설을 계획 중임
- 교육과정 개설 희망 시기를 보면, 1년 이내 4개 대학(원), 1~2년 이내 8개 대학(원)으로 조사됨
- 고려대의 경우 단과대학별로 개설을 기획하고 있음

빅데이터 분석 전문가 교육 수요조사 결과(4/5)

1

각 대학(원)의 빅데이터 관련 과정/교과목 개설은 꾸준히 증가

- 2014년 1학기 : 과정 및 교과목 개설 - 14개 대학(원)
2015년 2학기 : 과정 및 교과목 개설 - 25개 대학(원)으로 증가
- 분석 역량 강화 목적으로 빅데이터 분석 이해와 활용 > 빅데이터분석실무 > 빅데이터 구조 설계론 순서로 교과목이 고도화되고 있음.

2

한국정보화진흥원 중심의 빅데이터 분석 실습 교육 지원에 긍정적

- 빅데이터 분석 실습 지원 희망 대학(원) : 25개 대학(원)의 학과(학부:1103, 석박사 : 260)
- 빅데이터 분석 실습 지원 우선 희망 분야 : 실습 강사 지원 > 실습 인프라 지원 > 실습 커리큘럼
- 빅데이터 분석 실습 지원 희망 시기 : 10월 말, 11월 초

3

빅데이터 분석 교과목 중심으로 빅데이터 분석 툴킷 활용 희망

- 빅데이터 분석 툴킷 적용 교육과정은 빅데이터 분석과 빅데이터 이해와 실무 교과목에 집중되어 있음.
- 빅데이터 분석 툴킷 활용한 빅데이터 분석 실습지원 희망 시기는 10월초와 11월초에 집중되어 있음.
- 현재 개설되어 있는 각 대학(원)에서 분석툴킷을 적용하여 실습을 진행하고자 하는 수요가 있음.

빅데이터 분석 전문가 교육 수요조사 결과(5/5)

4

빅데이터 관련 과정/교과목 운영 시 데이터셋 지원 필요

- 현 빅데이터 관련 과정/교과목을 운영 중인 25개 대학(원)은 데이터 셋 확보, 인프라 및 전문가 확보의 어려움이 있음.
- 각 대학(원)은 데이터 분석 및 관리, 플랫폼 개발 및 운영, 실무활용 능력의 중급이상 양성을 목표로 하고 있음.
- 개설 대학(원)의 빅데이터개론 및 이해와 실무, 경영데이터분석, 시스템 분석 및 설계, 데이터마이닝 등의 빅데이터 교과목을 개설 운영 중에 있음.

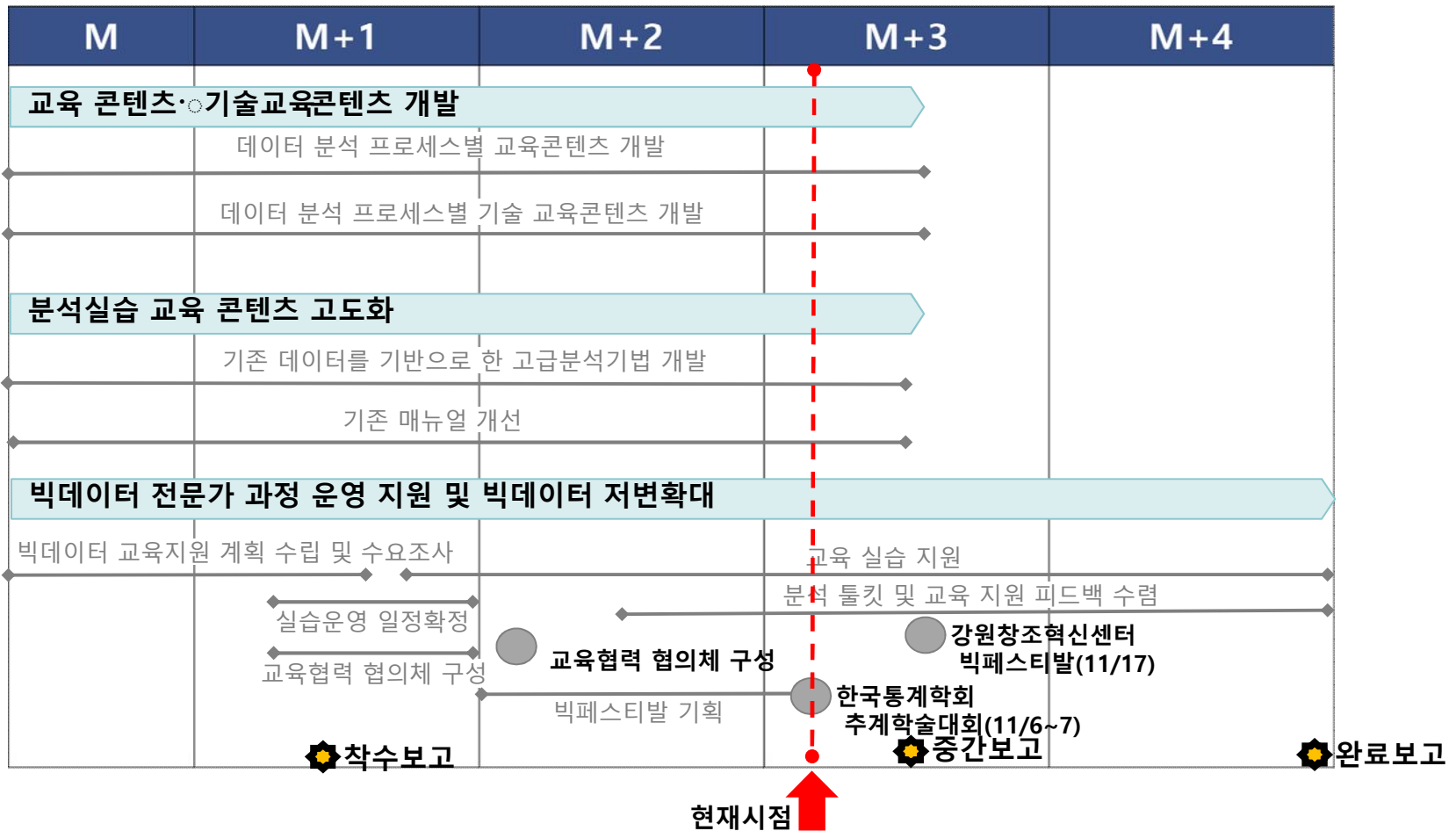
5

빅데이터 관련 과정/교과목 신규 개선을 위한 지속적인 지원 필요

- 향후 1~3년 이내에 16개 대학(원)이 빅데이터 관련 과정/교과목 개선을 계획하고 있음.
- 각 대학(원)은 중급 수준의 기반 역량과 기술 역량 관련 교육 내용을 검토하고 있음.
- 빅데이터 관련 교육과정 개설 계획 시 지원 필요 부문은 인프라 환경 > 예산확보 > 표준 커리큘럼 > 관련 전문가 확보 임.

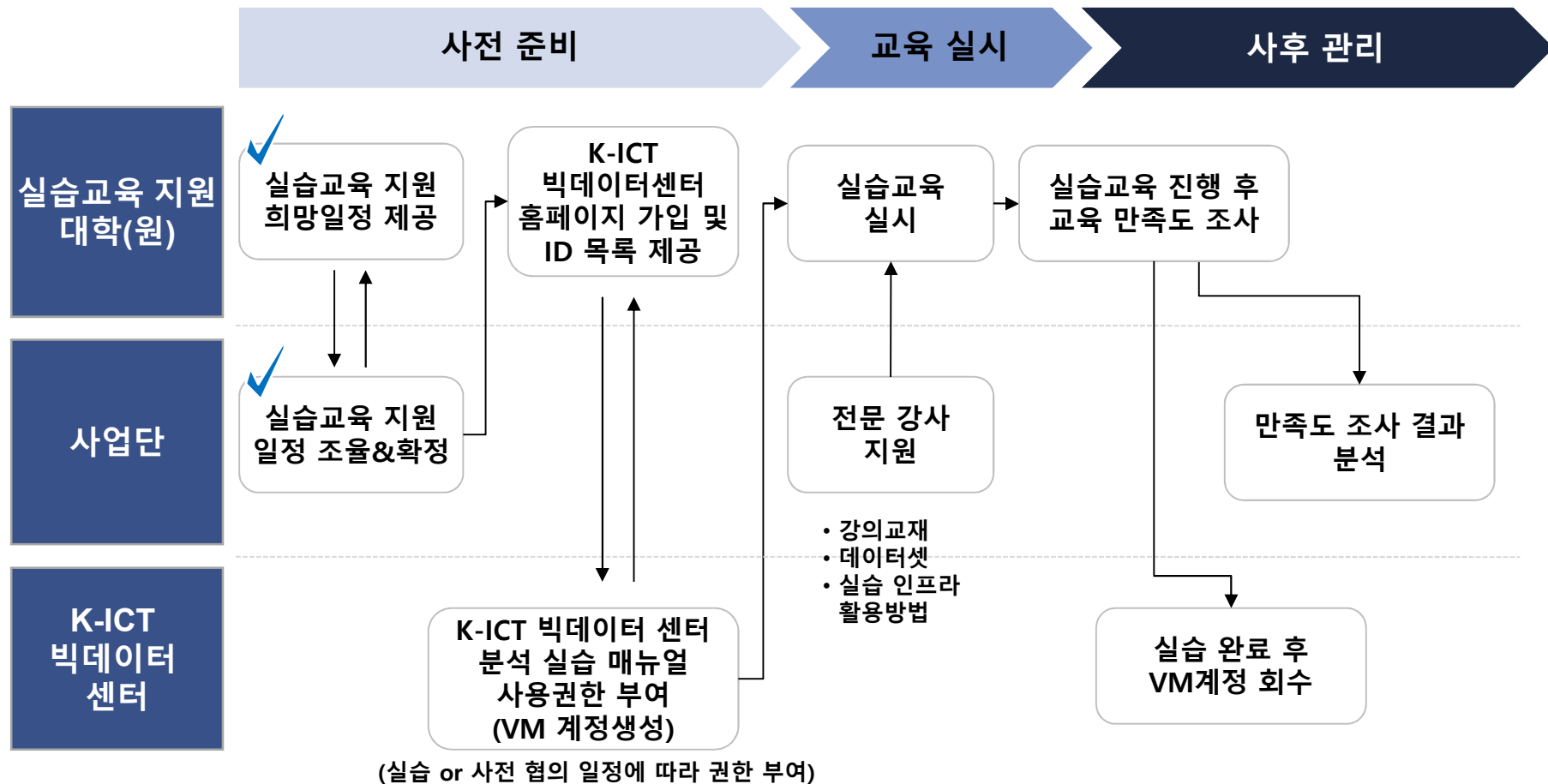
추진 일정

본 사업은 총 5개월의 일정으로 금년 12월 말에 종료 예정입니다.



빅데이터 분석전문가 실습 프로세스

준비, 실시, 사후관리 단계별로 실습 교육이 진행될 예정이며, 정확한 지원 일정은 각 대학의 교수님과 협의 후 확정하였습니다.

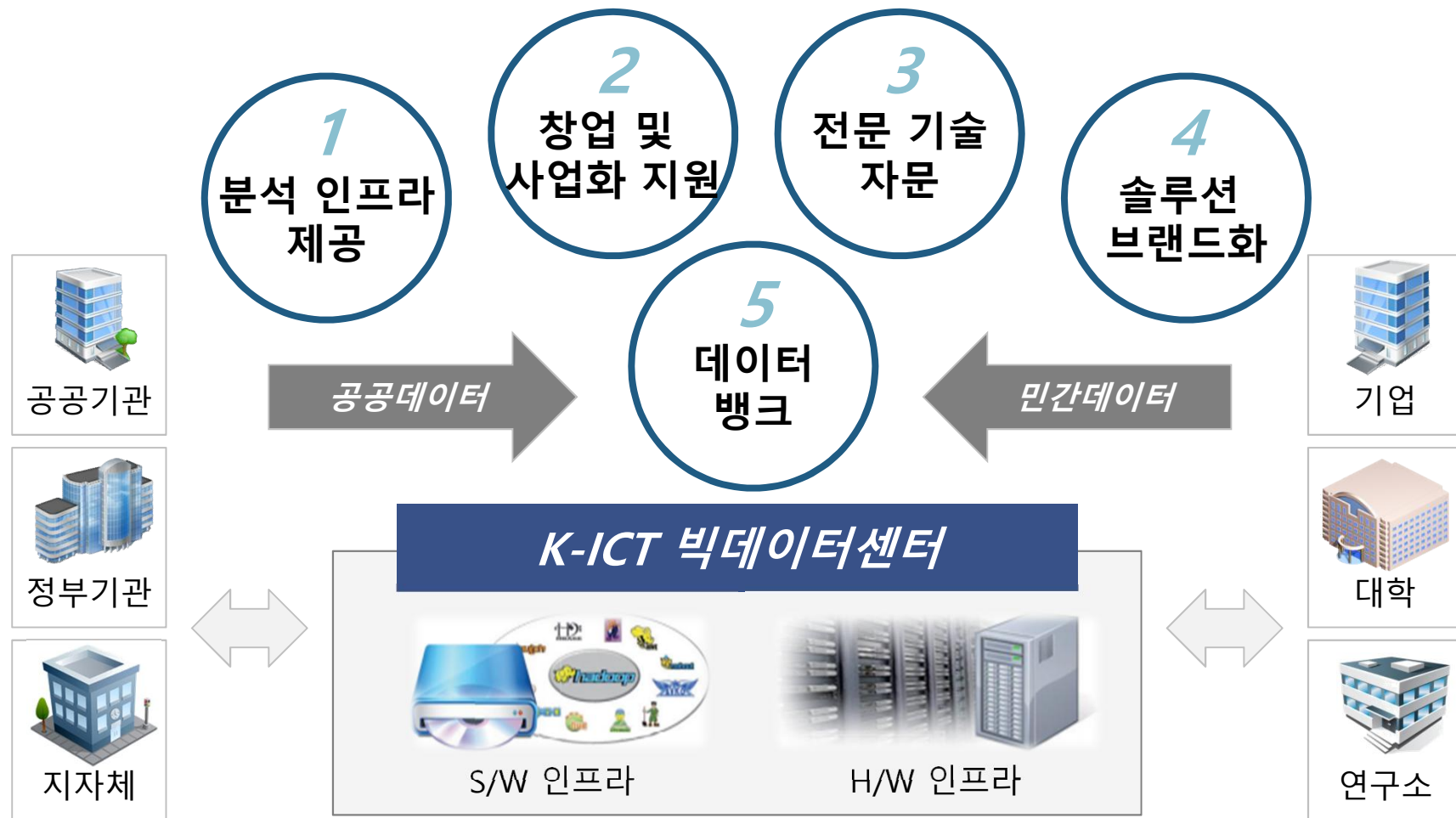


Chapter II. K-ICT 빅데이터센터 운영현황



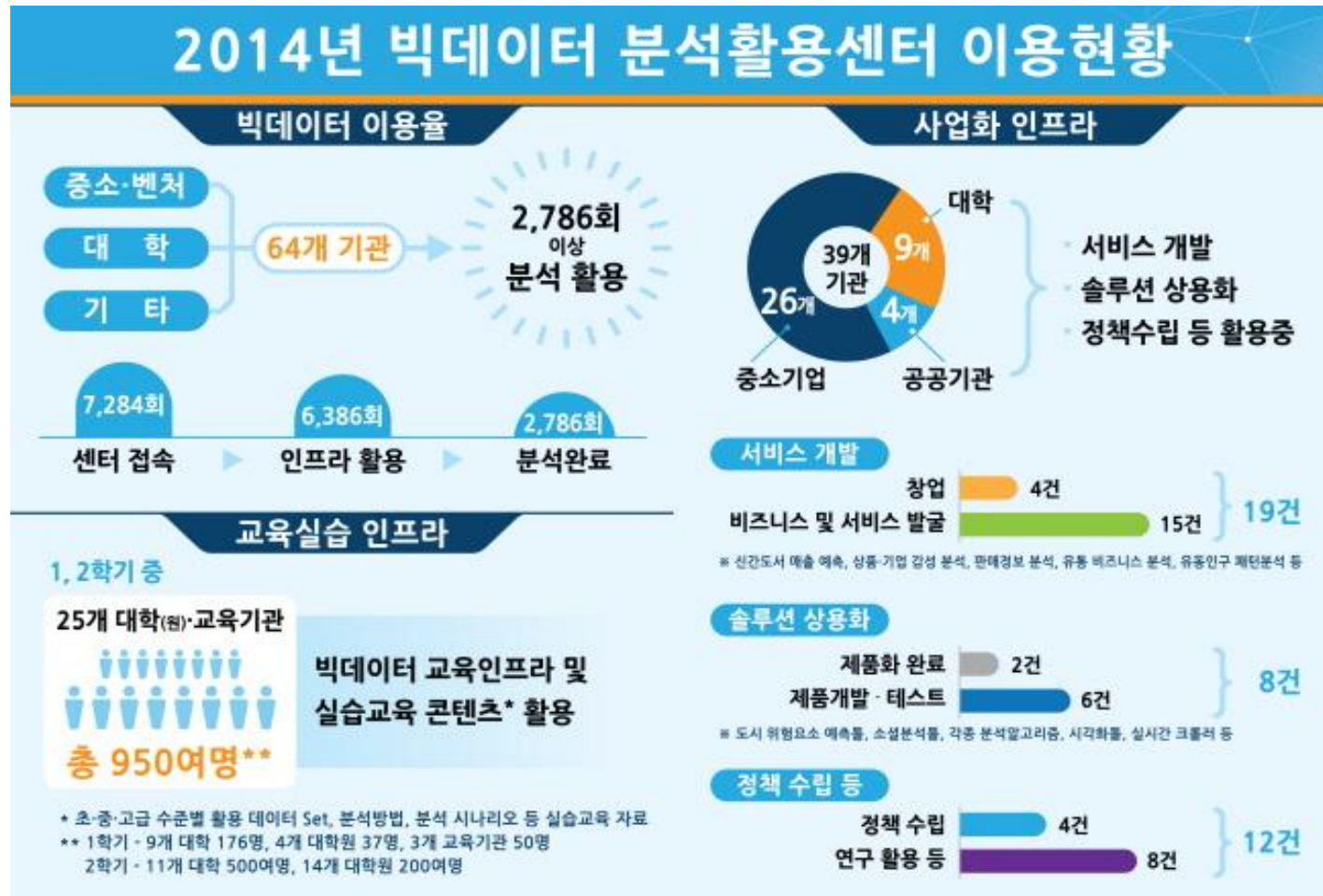
추진 목표 및 방향

창업자, 중소벤처, 대학 등에 Shared Service를 제공하여 빅데이터 사업화 및 인력양성 지원 등 산업활성화를 촉진하는데 목표가 있습니다.



K-ICT 빅데이터센터 이용현황

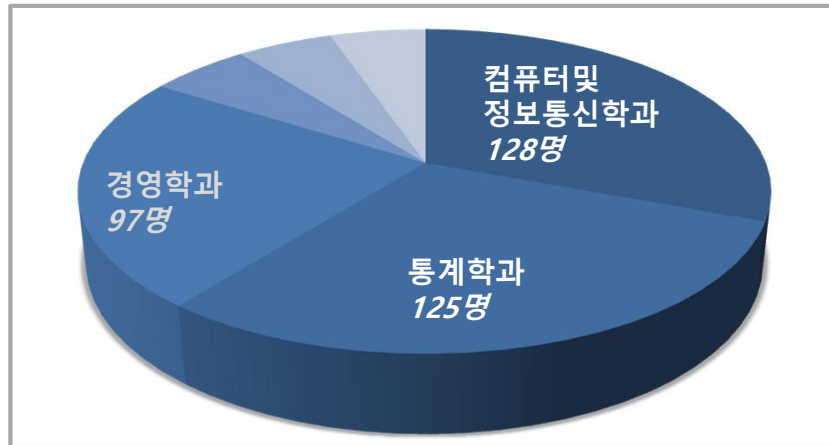
'14년도 중소벤처, 대학 등 64개 기관이 2,786회 이상 분석에 활용 되었습니다.
(서비스 개발 19건, 솔루션 상용화 8건, 정책 수립 12건)



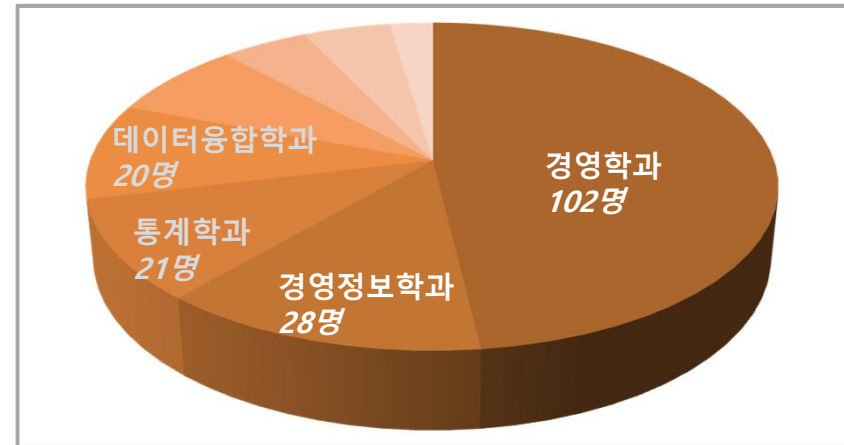
K-ICT 빅데이터센터 교육인프라 활용 현황

'14년 교육 인프라를 통해 경영학, 통계학, 컴퓨터 및 정보통신학 등의 전공 학생 629명을 대상으로 실습 교육을 진행하였습니다. '15년에는 20개 대학(원)을 지원할 예정입니다.

학부 학생 - 417명



대학원 과정 학생 - 212명



학과	학부	대학원	합계
경영학과	97	102	199
통계학과	125	21	146
컴퓨터 및 정보통신학과	128	10	138
경영정보학과	23	28	51
언론정보학과	22	5	27
시각영상디자인학과	22	-	22
데이터융합학과	-	20	20
문헌정보학과	-	16	16
산업공학과	-	10	10
합계	417	212	629

Chapter III. K-ICT 빅데이터센터 인프라

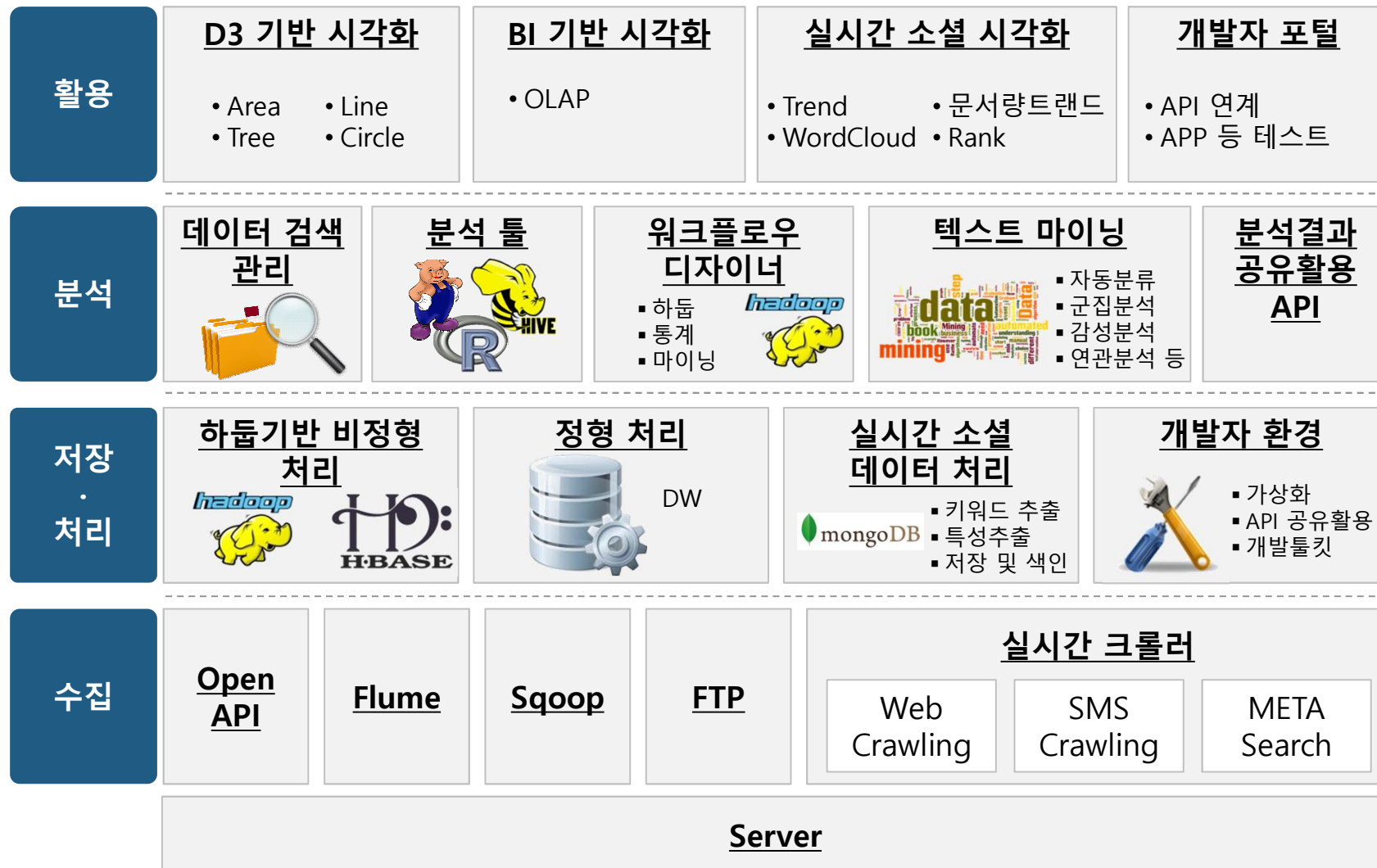
1. 창업 및 사업화 분석 인프라
2. 교육 실습 인프라
3. 개발인프라



1. 빅데이터 분석 인프라

III. K-ICT 빅데이터센터 인프라

빅데이터 분석 인프라는 분석활용, 분석실행, 저장·처리, 수집 기능이 있습니다.



2. 창업 및 사업화 분석 인프라

III. K-ICT 빅데이터센터 인프라

빅데이터 분석 인프라를 보유하기 어려운 중소기업, 1인 창조기업, 대학, 연구소 등이 대용량 데이터 분석 및 기술개발·검증에 사용할 수 있는 테스트베드를 제공합니다.

제공목적	창업자·중소벤처의 서비스 개발, 대용량 데이터 분석 및 컨설팅 등 사업화 지원을 위한 인프라
이용대상	창업자, 중소벤처, 개발자 커뮤니티, 대학/연구소 등
운영방식	Job 스케줄링 방식(Fair Scheduler)을 적용하여 개발자 및 분석전문가들이 동시에 분석 인프라 활용
제공기능	수집서버, 분석서버, 이용자환경, 분석 툴, 개발자 환경 및 웹 포털 등

기능	세부 내역	
하둡 기반 서버	<ul style="list-style-type: none"> 서버 사양 : 2.4GHz, 32GB Mem, 2TB HD HD 용량 : 총 400TB 	
이용자 자원 할당	<ul style="list-style-type: none"> 최대 400TB까지 이용 가능 * 신청 스케줄, 동시 이용기관 수, 자원 활용의 적절성 등을 검토하여 자원할당 	
RDB	<ul style="list-style-type: none"> HD용량 : 32TB 	BI 분석용
개발자 포털	<ul style="list-style-type: none"> 저장용량 : 8TB 제공기능 : 프로그램·App, 알고리즘 등록 활용 시험 등 	

3. 교육 실습 인프라

III. K-ICT 빅데이터센터 인프라

빅데이터 전문인력 양성을 위해 원격으로 교육실습 인프라 접속하여 시스템 구축 및 데이터 분석, 프로그램 개발 등 다양한 실습을 수행할 수 있는 환경을 제공합니다.

제공목적	대학, 전문교육 기관 등 인력 양성 지원을 위한 실습 인프라
이용대상	대학(원), 커뮤니티, 협단체, 공공/민간 교육기관 등
운영방식	실습서버(가상머신 기반), 이용자환경, 분석툴, 교육용 데이터 set 등
제공기능	가상화 방식을 적용하여 실습환경에 따라 맞춤형 환경설정이 가능하여 여러 교육과정 동시 지원

기능	세부 내역
교육용 데이터 셋	▪ 쇼핑, 농산물, 제조, 관광 등 10종의 비즈니스 분석 실습 데이터 제공
물리적 자원	▪ 서버사양 : 2.4GHz, 32GB Mem, 2TB HD
이용자 자원할당	▪ 교육 용도에 따라서 최대 10GB 용량 범위 내에서 할당 * 단 교육용도에 따라서 필요한 자원에 대해 센터와 협의 후 신청 가능
가상화 이미지	▪ 리눅수, 하둡, 하둡-Hbase, 하둡-Hive, 하둡-Eco(Pig, Hive, Mahout), 하둡-분석(Pig, Hive, R) 등

빅데이터 분석 교육 콘텐츠


III. K-ICT 빅데이터센터 인프라
3. 교육실습 인프라

2014년 10개 산업군의 데이터를 활용하여 수준별 총 30종의 실습 교육 콘텐츠를 제작 · 활용하였습니다.

10개 산업군 데이터

1차	2차	3차
 농산물 농축산 식품 유통 정보 데이터	 쇼핑 온라인 쇼핑몰 정보 데이터	 관광 제주도 내외국인 관광 데이터
 소비 신용카드(가계부) 정보 데이터	 교통 지하철/버스 교통 정보 데이터	 제조 자동차 부품 제조 정보 데이터
 소셜 SNS(트위터, 뉴스) 정보 데이터	 유통 대한상공회의소 유통 정보 데이터	 패션 패션 정보 데이터
		 글로벌 해외 의료/헬스 정보 데이터

실습 교육 데이터



원시데이터

수집 → 가공 → 저장 → 분석 → 시각화

초급 중급 고급

산업/수준 별 실습 교육 데이터

매뉴얼



활용 매뉴얼



교육 동영상








웹 매뉴얼

빅데이터 분석 교육 콘텐츠 개발 경과

III. K-ICT 빅데이터센터 인프라
2. 교육실습 인프라

총 5개 분야에 대해 수준별로 기술·교육 콘텐츠의 개발 및 고도화 작업이 진행 중입니다.

	초 급	중 급	고 급
농산물 	<ul style="list-style-type: none"> 탐색적 자료 분석 및 변환 기초 자료 분석을 통한 Topic 발견 (상관분석, 공적분 분석) 데이터 시각화 가격 연관성 있는 농산물의 시계열(+/-) 	<ul style="list-style-type: none"> 농축산물 가격 연관성 탐색(clustering) 지역별 돈육 가격 연관성 탐색(clustering) 데이터 시각화 <ul style="list-style-type: none"> 지역별 돈육 가격 시계열 날씨 및 뉴스 자료의 가격 변화 요인 	<p>고급 수준의 분석 문제 제공</p>
소비 	<ul style="list-style-type: none"> 데이터 탐색 군집 분석 적용 분석 시각화(기초 통계량 요약 및 시각화) 날씨 데이터와 결합 <ul style="list-style-type: none"> 계절별 소비 패턴 변화 파악 	<ul style="list-style-type: none"> 대응 일치 분석 <ul style="list-style-type: none"> 분석 결과를 통한 카드사 포지셔닝 분석 범주형 자료분석 기법 <ul style="list-style-type: none"> 카이제곱 검정, 피셔의 exact검정 등 모자이크 플롯 시각화 	
소셜 	<ul style="list-style-type: none"> 문제의 발견(corpus 구축) 검색을 통한 Topic 발견 특정 단어 데이터 색출(string R package) Word Cloud 구축 실제 기사 확인 	<ul style="list-style-type: none"> 사전 만들기 및 구성 개발 사전 적용 후 Term Document Matrix 구성 각 뉴스별 어떤 Topic 언급했는지 파악 <ul style="list-style-type: none"> clustering, 시각화 	
교통 	<ul style="list-style-type: none"> 자료 탐색 및 정제 수치요약 및 경향 파악(정량화) Aggregate(), xtabs() 등 관련함수 사용 수치결과의 시각화 <ul style="list-style-type: none"> barplot, pie, matplotlib, high-level plot 	<ul style="list-style-type: none"> 지도 맵핑 특정 질문에 대응하는 지하철역 및 수치를 지도상에 시각화 <ul style="list-style-type: none"> Google Map, 공공데이터 활용 등 	
제조 	<ul style="list-style-type: none"> 자료 탐색(자료성격 및 변수 속성 파악) 각 변수별 특성 파악(분포, 요약통계량) 이상치 및 결측치 처리 변수간 연관성 파악 <ul style="list-style-type: none"> 2차적 산점도, 상관분석 	<ul style="list-style-type: none"> 불량률 분석 <ul style="list-style-type: none"> 로지스틱 회귀, 의사결정나무 적합 분석결과 시각화 분류 분석 <ul style="list-style-type: none"> 로지스틱 회귀, 랜덤포레스트 ROC 커브 적용 모형 평가 	

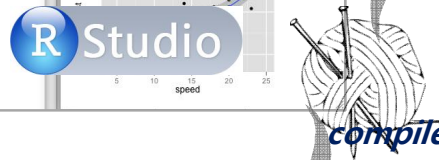
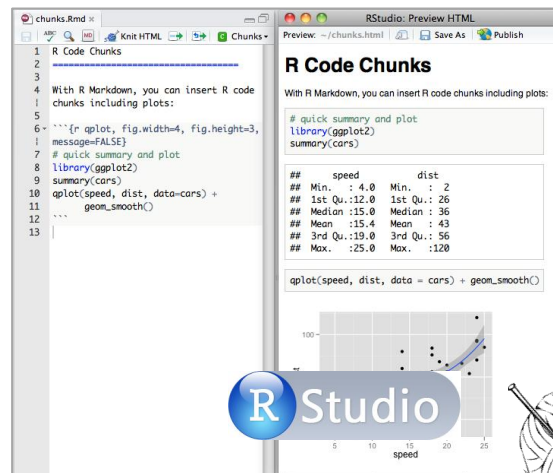
교육 콘텐츠 강의교재 set 구성

III. K-ICT 빅데이터센터 인프라
2. 교육실습 인프라

R의 사용자 환경인 Rstudio를 활용한 reproducible research 개념을 근간으로 데이터 분석 프로세스별 교육 콘텐츠를 작성 배포할 예정입니다.

R코드 및 마크다운 텍스트

HTML5 기반의 문서생성



MS Word 기반의 문서생성



MS PowerPoint로 일부 가공 생성

교통 분야 콘텐츠 sample

III. K-ICT 빅데이터센터 인프라
2. 교육실습 인프라

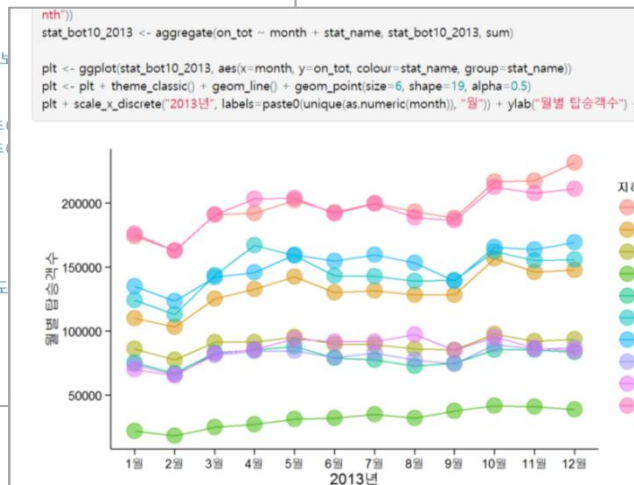
교육 콘텐츠는 스토리 기반의 분석 시나리오를 기반으로 데이터의 가공, 시각화 기반의 기초 분석 및 각종 방법의 적용 등 다양하게 구성되어 있습니다.

*“ 다양한 데이터에 대하여 시나리오를 구성하여
Story 기반으로 데이터를 분석하도록 콘텐츠를 작성함 ”*

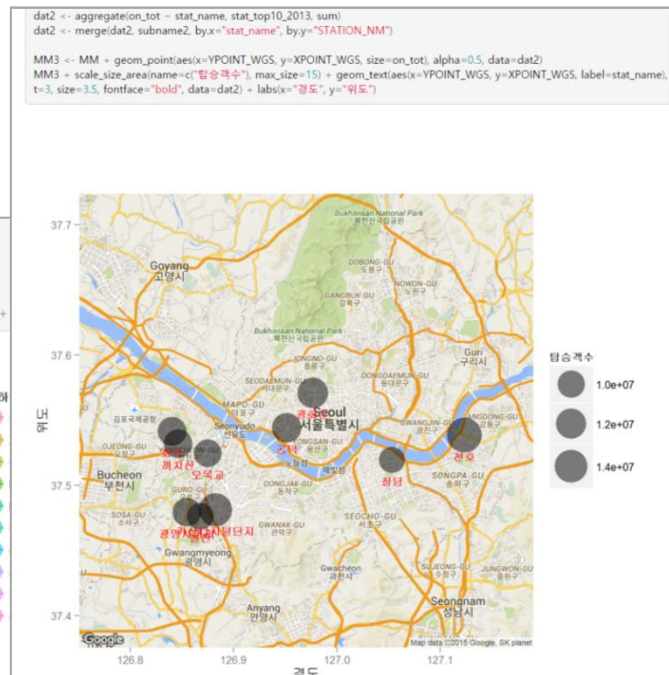
교통 (데이터 분석 콘텐츠 활용 매뉴얼)

- 학습목표
 - 활용 데이터 소개
- 1. 데이터 로딩 및 전처리 과정
 - 1.1 필요 패키지 불러오기
 - 1.2 subway.csv 파일 로딩 및 자료 정리
 - 1.3 subway_latlong.csv 파일 로딩 및 자료 정리
- 2. 지하철역별 수치요약 및 시각화
 - 2.1 연도별, 역별 지하철 탑승객 수의 계산
 - 2.2 탑승객수 기준 상위/하위 10개역 추출 및 정보
 - 2.2.1 상위 10개역 추출
 - 2.2.2 하위 10개역 추출
 - 2.2.3 2013년 상위 10개역 월별 탑승객 추
 - 2.2.4 2013년 하위 10개역 월별 탑승객 추
- 3. 지하철노선별 수치요약 및 시각화
 - 3.1 노선별 역당 평균 탑승객수의 계산 및 비교
 - 3.2 노선별 누적 승객수의 상대비교
- 4. 구글맵을 활용한 지도 맵핑
 - 4.1 구글맵을 활용한 지도맵핑의 예시
 - 4.2 2013년도 탑승객수 상위 10개역에 대한 지도
- 5. 뉴스데이터의 활용
 - 5.1 회귀분석
 - 5.2 잔차분석을 통한 이상점 탐지
 - 5.3 뉴스자료의 로딩 및 뉴스검색

콘텐츠의 인덱스 도입



- 하위 10개 역의 2013년 자료를 stat_bot10_2013 에 저장하여 이를 선그래프로 시각화
- 상위 10개 역과 비교하여 월별 탑승객 수의 패턴이 상이함을 알 수 있음



상세한 설명

R코드와 해당 결과물 공개 및 연결

3. 개발 인프라

개발자들이 만든 API, 앱, 알고리즘에 대한 정보를 제공하며, API는 정보 보기 페이지에서 API 호출을 테스트해 볼 수 있습니다.

제공목적	창업 및 중소벤처의 신규 서비스 개발 및 테스트 환경
이용대상	창업자, 중소벤처 등
제공기능	<p>분석 결과를 API로 연계하여 APP 등 개발서비스 등록 및 테스트가 가능하도록 인프라 구성</p> <p>- 개발 툴킷, 분석결과 API 활용 기능, APP 등록 및 서비스 활용 테스트 기능 등 제공</p>

기 능	세부 내역	
서버	<ul style="list-style-type: none"> ▪ 서버 사양 : 2.2GHz, 48GB Mem, 12TB HD ▪ HD 용량 : 총 32TB 	
콘솔접속 환경	<ul style="list-style-type: none"> ▪ 웹 서버, 개발 툴킷(JAVA 등), 데이터 연계 API, 개발자 포털(App 등록, 변경, 삭제 등) 	

Q&A



미래창조과학부

NIA

한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY

2e 투이컨설팅