SPRi FORUM
소프트웨어정책연구소

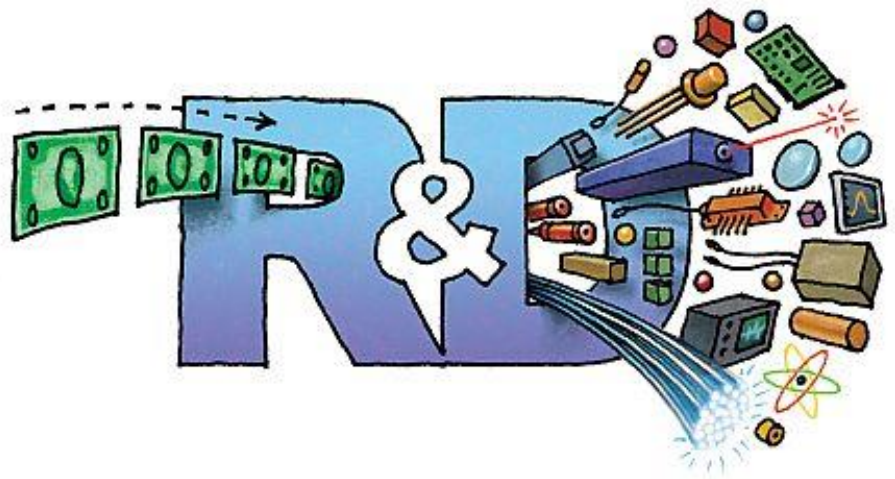
국가 과학기술 빅데이터 공유·활용체제 구축

2014. 10. 14

KISTI 과학기술빅데이터연구실장
이 상 환

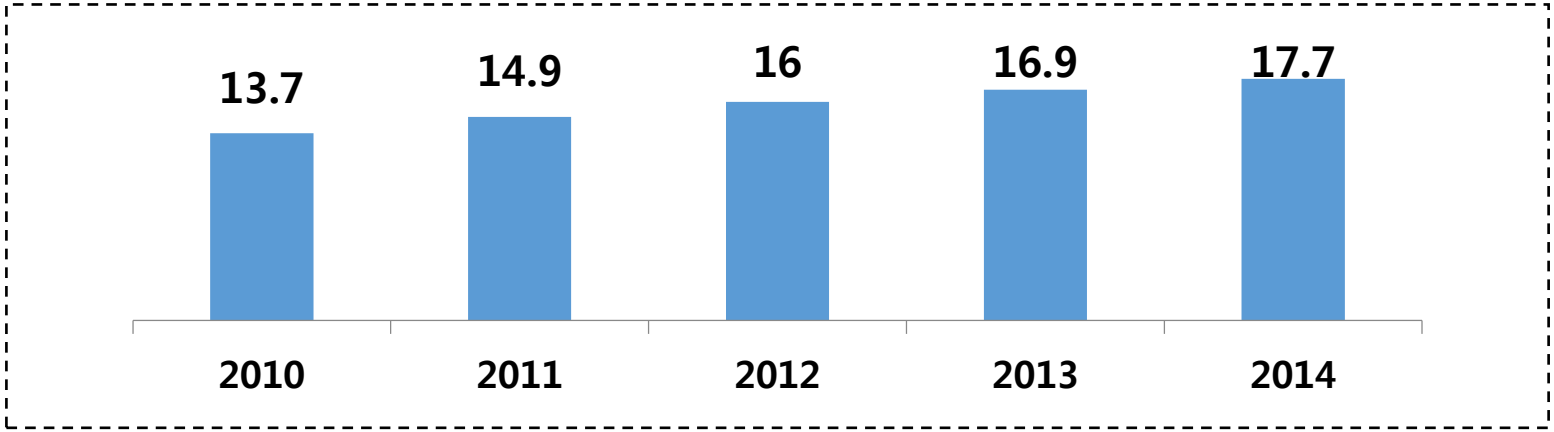
- Contents –

Ⅰ. 국가 R&D 예산 및 성과물

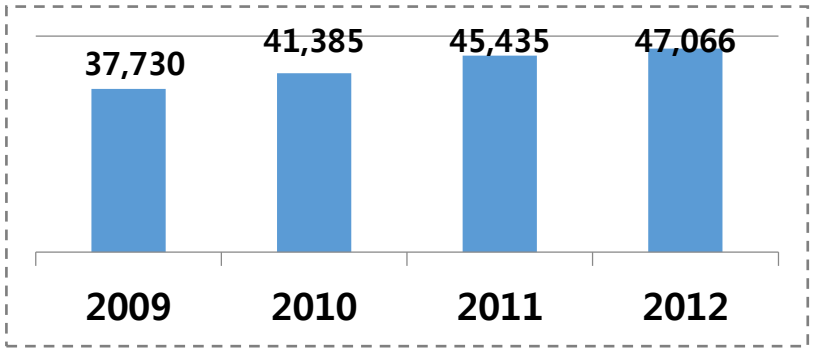


정부 R&D 예산과 성과

- **'14년 정부 R&D 예산은 17.7조원(연평균 7.5%)**



- **이에 반해, 성과 질적 수준 및 지식효과 확산은 상대적으로 미흡**



- **(논문)** 질적수준과 생산성 : 국제 평균대비 낮음
  * 논문당 평균 피인용 수('08-'12) : 4.23(한국)/5.15(세계평균)
  * 연구원 1인당 SCI 논문 수('11) : 0.16(한국)/0.34(OECD)

- **(특허)** 양적 성과는 지속적인 증가 추세
  * PCT 출원건수 : ('10) 9,669 → ('11) 10,447 → ('12) 11,848

- **(지식효과·확산)** 지식창출성과 대비 저조
  *** '12년 글로벌 혁신 지수**
  → 지식창출(3위) : 논문, 특허 건수
  → 지식효과(43위) : 사업화, 표준화 등
  → 지식확산(20위) : 기술료, 컴퓨터통신서비스 수출 등

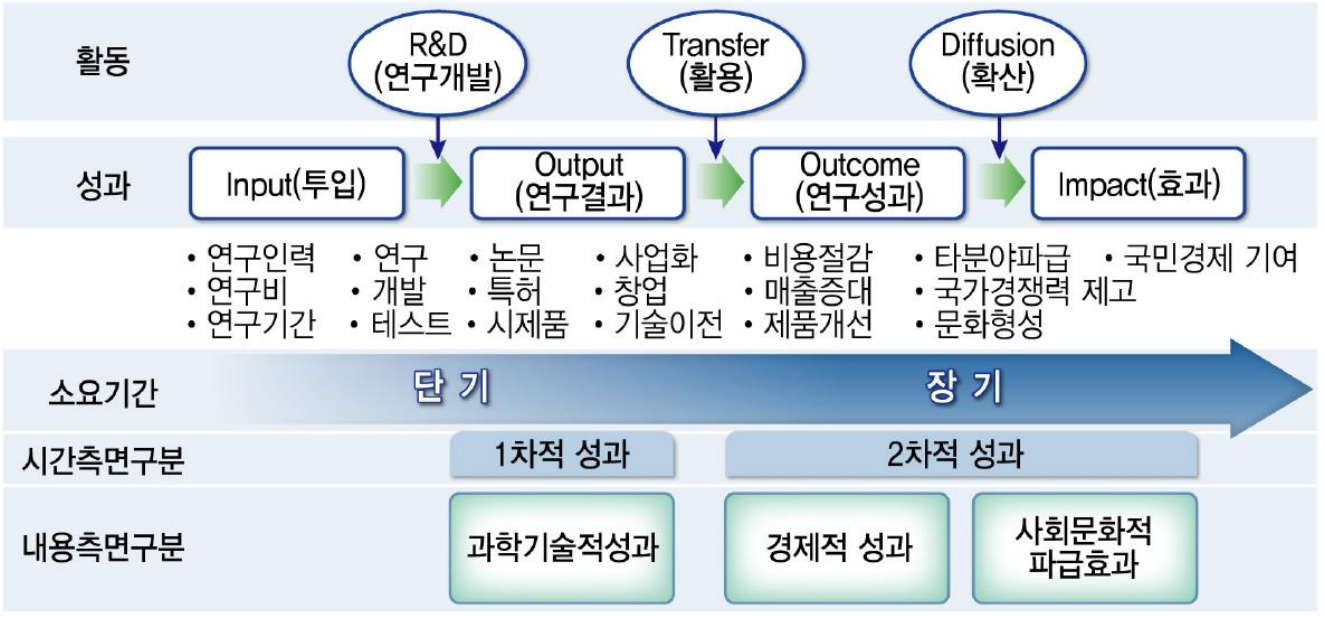
연구성과 개념

정부는 연구개발 활동을 성과 중심으로 평가하고,
연구성과를 효율적으로 관리·활용하기 위해 법률을 제정

* 2005년「국가연구개발사업 등의 성과평가 및 성과관리에 관한 법률」

- '연구성과'는 연구개발을 통해 창출되는 특허·논문 등 과학기술적 성과와 그 밖에 유·무형의 경제·사회·문화적 성과를 포함

- 연구성과는 일반적으로 단기·직접적으로 발생하는 1차적 성과와 장기·간접적으로 발생하는 2차적 성과로 구분



국가 R&D성과물 및 전담기관

국가연구개발사업의 관리 등에 관한 규정 제25조 13항에 따라 연구성과 분야별 관리·유통 전담기관을 두고 있음

구분	성과물	성과관리 전담기관	관리대상(등록 · 기탁 기준)
등록	논 문	한국과학기술정보연구원	· 국내외 학술단체 및 출판사에서 발간하는 학술지 및 학술대회지에 수록된 학술논문(전자원문 포함)
	특 허	R&D 특허센터	· 국내외 출원 또는 등록된 특허정보
	보고서 원 문	한국과학기술정보연구원	· 연구개발 종료 시 제출하는 최종보고서 및 연차보고서(전자원문 포함)
	연 구 기자재	한국기초과학지원연구원	· 국가연구개발사업의 수행 시 취득한 장비 중 가격이 3천만원 이상인 장비 또는 취득가격이 3천만원 미만이라도 공동 활용이 가능한 장비
	기술요약 정 보	한국산업기술진흥원	· 기초·응용·개발단계 등의 최종보고 및 연차보고가 완료된 결과물의 기술정보를 요약하여 공유·활용(기술이전, 사업화 등) 할 수 있도록 작성된 기록정보
	생명자원 (생명정보)	한국생명공학연구원	· 유전체 정보(서열, 발현정보 등) · 단백질정보(서열, 구조, 상호작용 등) · 발현체 정보〈유전자(DNA)칩, 단백질칩 등〉 · 관련정보
	소프트 웨어	한국저작권위원회	· 창작된 소프트웨어 및 등록에 필요한 관련정보
기탁	생명자원 (생물자원)	한국생명공학연구원	· 미생물자원(세균, 곰팡이, 바이러스 등) · 동물자원(사람·동물세포, 수정란 등) · 식물자원(식물세포, 종자 등) · 유전체자원(DNA, RNA, 플라스미드 등) 및 관련정보
	화합물	한국화학연구원	· 합성 또는 천연물에서 추출한 유기화합물 및 관련정보

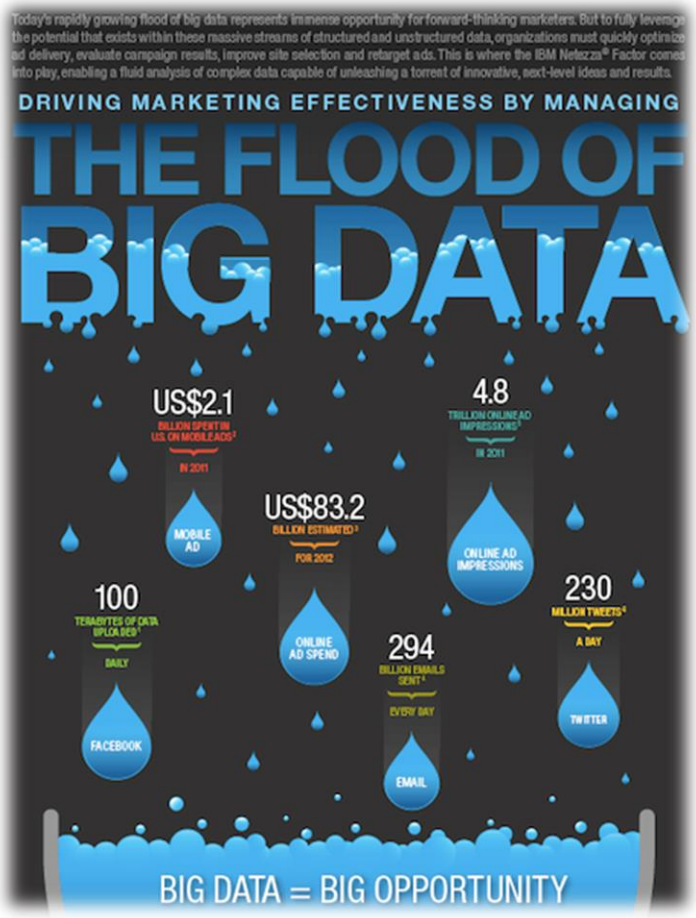
범 부처 차원의 통합적 연구성과 등록관리 및 공동활용은 미흡

*** 전담기관 외 연구기관의 기탁률('09-'12)**

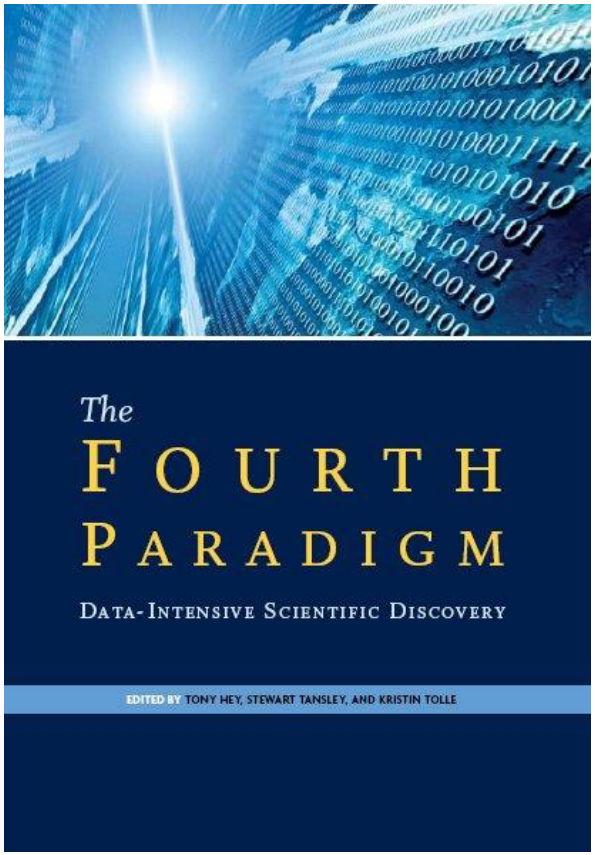
→ 화 합 물 : 3.8%

→ 생물자원 : 0.18%

Ⅱ. 과학기술 빅데이터


Today's rapidly growing flood of big data represents immense opportunity for forward-thinking marketers. But to fully leverage the potential that exists within these massive streams of structured and unstructured data, organizations must quickly optimize ad delivery, evaluate campaign results, improve site selection and retarget ads. This is where the IBM Netezza® Factor comes into play, enabling a fluid analysis of complex data capable of unleashing a torrent of innovative, next-level ideas and results.
DRIVING MARKETING EFFECTIVENESS BY MANAGING
THE FLOOD OF
BIG DATA
US$2.1
BILLION SPENT IN US ON MOBILE ADS[2]
IN 2011
MOBILE AD
US$83.2
BILLION ESTIMATED[3]
FOR 2012
ONLINE AD SPEND
4.8
TRILLION ONLINE AD IMPRESSIONS[3]
IN 2011
ONLINE AD IMPRESSIONS
100
TERABYTES OF DATA UPLOADED[1]
DAILY
FACEBOOK
294
BILLION EMAILS SENT[4]
EVERY DAY
EMAIL
230
MILLION TWEETS[4]
A DAY
TWITTER
BIG DATA = BIG OPPORTUNITY

R&D 패러다임과 빅데이터

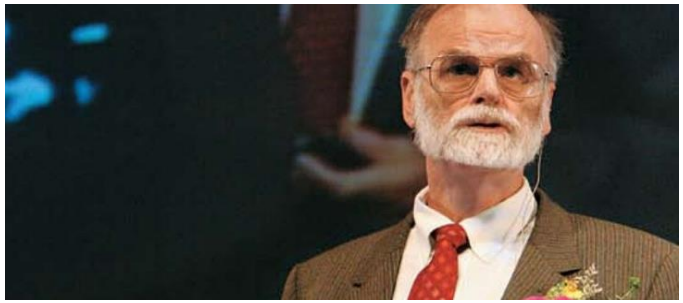


1세대 연구
실험 중심

2세대 연구
이론중심

3세대 연구
컴퓨팅자원중심

4세대 연구
데이터중심

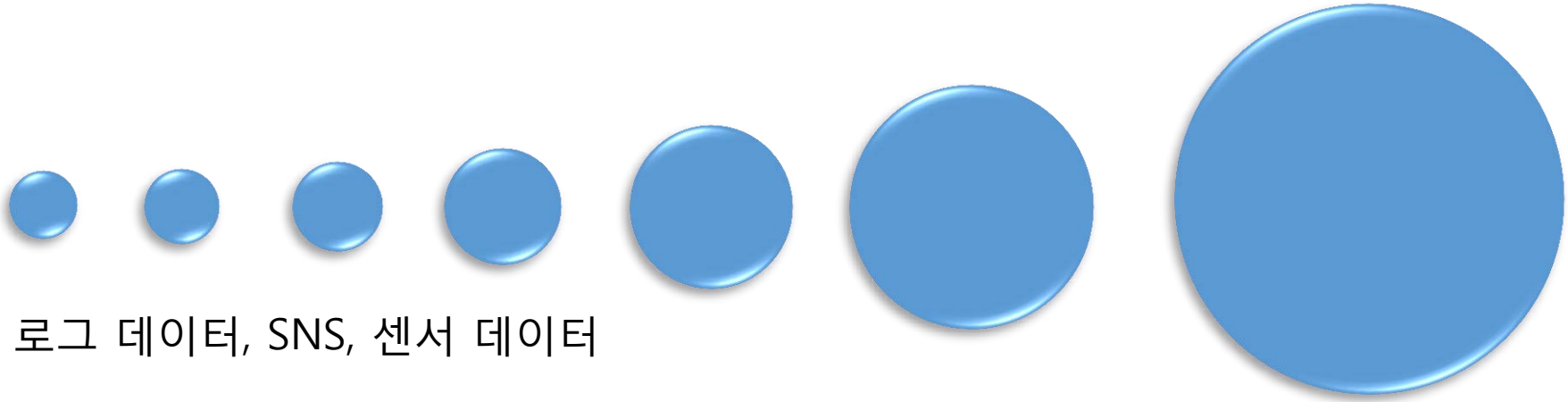


Jim Gray

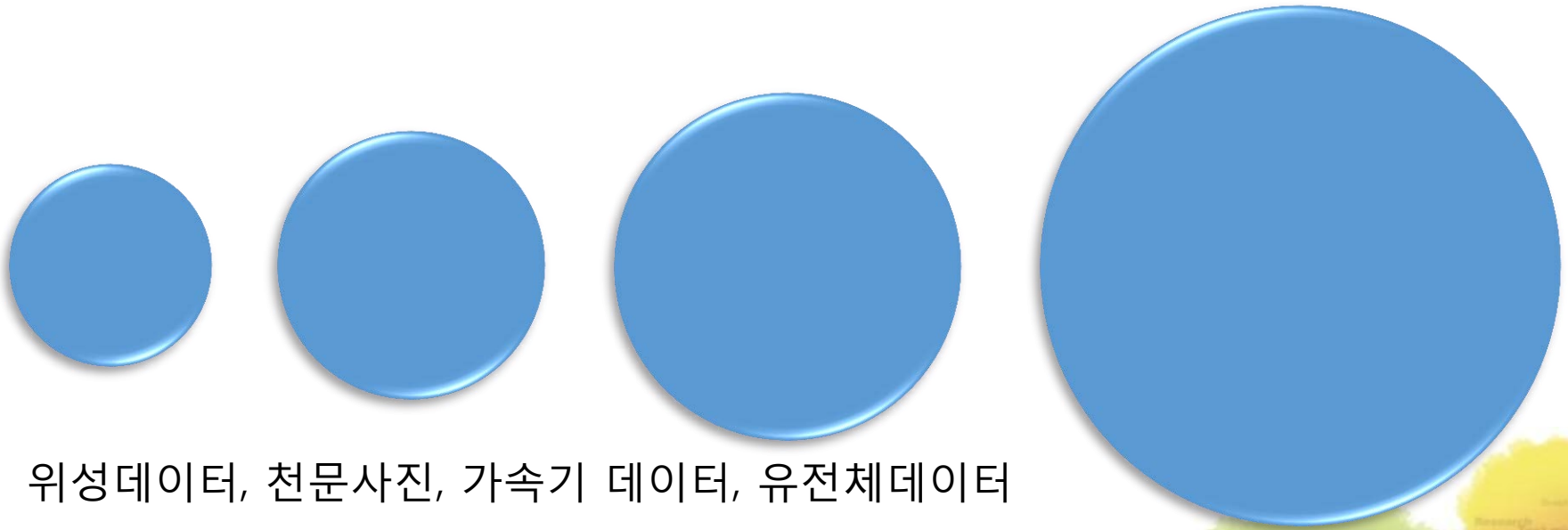
- 미래의 연구방식은 데이터 분석 기법이 다양한 과학연구의 일반적인 현상
- 데이터 분석 소프트웨어 기술을 더욱 발전시켜야 함을 강조

2007년 "Computer Science and Telecommunications Board"에서

우리가 알고 있는 빅데이터는?

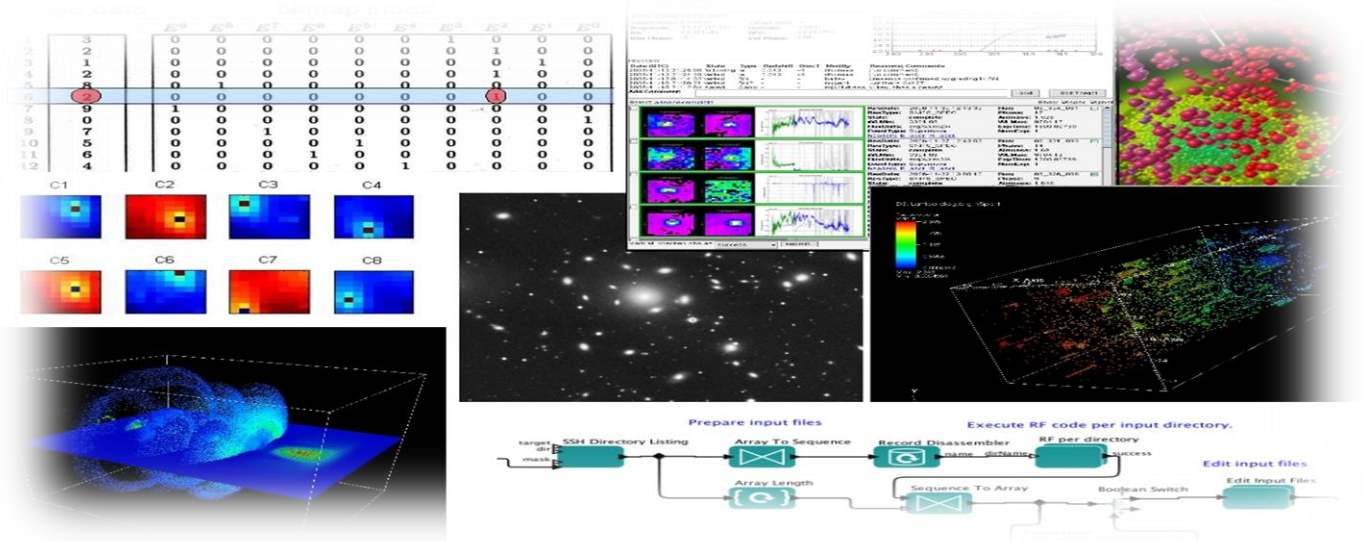


로그 데이터, SNS, 센서 데이터



위성데이터, 천문사진, 가속기 데이터, 유전체데이터

과학기술 빅데이터(1/2)



형태(type)

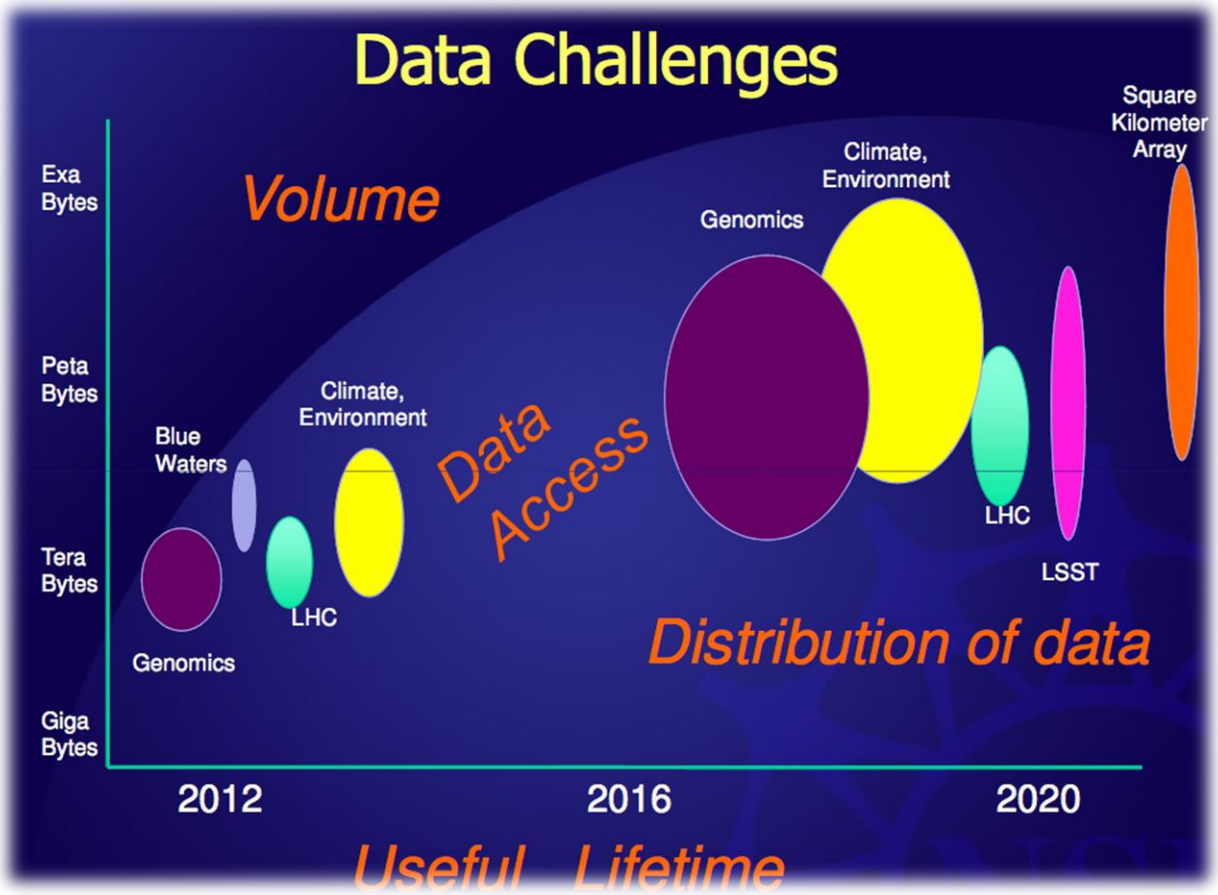
- 수치(numerical)
- 공간(spatial)
- 도표(graphical)
- 문서(text) 등

과학기술 활동의 과정·결과를 통해 얻은 데이터
연구 결과 검증을 위해 필수적인 데이터

출처:
OECD Principles and Guidelines for Access to Research Data from Public Funding (2007)

- 관측(observation) by 망원경, 전자현미경, 인공위성 등
- 감시(monitoring) by 센서 등
- 조사(investigation) by 설문조사, 기술/시장조사/ 기술가치평가 등
- 실험(experiment) by 가속기, 화학/바이오 실험장비 등
- 연구 분석(research analysis) by 분석도구 등
- 계산(computation) by 슈퍼컴퓨터 등

과학기술 빅데이터(2/2)

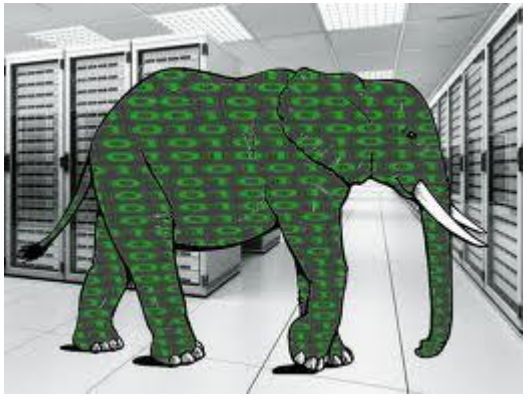


LHC : 입자 가속기 / LSST: 천체망원경

빅데이터, 오늘날의 문제인가?

- 1997년 10월, Michael Cox와 David Ellsworth가 IEEE 8th conference on Visualization에 게재한 논문에서 'big data' 용어를 최초 사용 (*Application-controlled demand paging for out-of-core visualization*)

- Big Data는 Data 자체가 아닌 problem.



"Visualization provides an interesting challenge for computer systems: ***data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk***. **We call this the *problem of big data***. When data sets do not fit in main memory (*in core*), or when they do not fit even on local disk, the most common solution is to acquire more resources."

- Data-Intensive Scalable Computing(DISC) 기술의 필요성 제기

빅데이터의 중요성

- **데이터 없이 힉스입자 발견 가능할까?**

- 아마존닷컴 CEO 제프 베조스

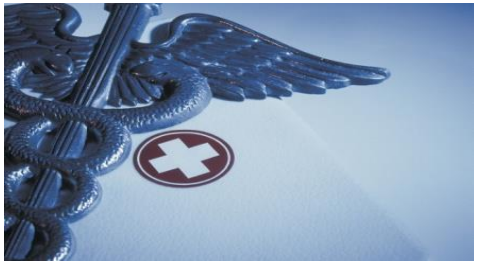
  "우리는 절대로 데이터를 내다버리지 않는다"

- IBM CEO 버지니아 로메티

  "앞으로 모든 산업에서 데이터가
  승자와 패자를 가를 것이다"

빅데이터 가치 - 경제적인 관점

Europe public sector administration
- 연 €2,500억 가치 창출
- 연 ~0.5% 생산성 향상



US health care
- 연 $3,000억 절감
- 연 0.7% 생산성 향상

Global personal location data
- 서비스 제공업체들에게는 $1,000억 이상의 가치 창출
- end user들에게는 최대 $7,000억 가치 제공

Manufacturing
- 개발 비용 50% 감소
- 노농 자본 투자 7% 감소

US retail
- 60% 추가 이익 발생
- 년 0.5~1.0% 생산성 향상

빅데이터 가치 - R&D 관점

"데이터를 활용한 시뮬레이션에
1달러 투자하면
3~9달러의 ROI 얻을 수 있음"

※ IDC 백서

"전문적인 데이터 관리와
큐레이션 서비스를 활용하여
연구시간의 **1%를 절약하면**
1년에 1,000만 유로를 절약"

※ Winkler, S., "Research Data - A Funder's Perspective", DataCite Summer Meeting, 8 June 2010.

Ⅲ. 국가 과학기술 빅데이터 거버넌스 체제 구축

미래창조과학부
Ministry of Science, ICT and Future Planning

추진배경

창의적 서비스를 발굴하고, 신성장동력 창출을 위해서는
결과 중심의 과학데이터 관리에서 연구과정까지 포괄할 필요

R&D 환경변화

- 시공간에 제약 없는 다양한 채널의 R&D서비스 요구
- 연구결과 중심에서 연구과정 데이터 공유 필요
- 산업 및 디지털 기기간 융합 가속화

국가현안해결

- 빅데이터 R&D투자가 지속적으로 증가하고 있으나, 성과는 아직 미흡
- 천문 등 분야별 현안 해결을 위해 과학기술 빅데이터 활용 움직임 증가
- 데이터 기반의 분석 정보 제공 필요

범 국가차원 에서 과학기술 R&D를 수행하고 있는 31개 부처에서 생성된
과학기술 빅데이터 공동활용 생태계 조성

현황 및 문제점

데이터 개별관리

- 연구자 단위로 폐쇄적으로 관리·활용

연구자가 직접 관리	개인적으로 관리	실험실내 관리자	공동관리
31.5%	24.3%	15.4%	11.3%

관리체계 미흡

- 범부처 차원에서 관리대상 및 범위 미정의

- 국가차원의 종합계획 부재
- 전담조직체계 미흡
- 공동활용을 위한 플랫폼 부재
- 활용에 관한 법·제도 부재

비전 및 목표

"데이터 활용을 통한 과학기술 강국 도약"

구축 중심이 아닌 「활용 중심 접근」

「연구현장 맞춤형」 정보제공

시범사업을 통한 「성공사례 발굴」

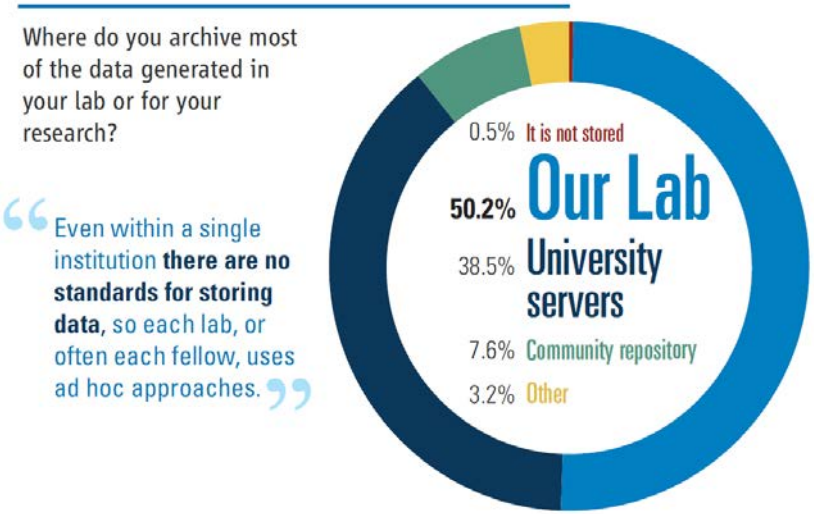
중점 추진과제

5대 분야	15개 중점추진과제
관리정책	1.1 과학기술 빅데이터 서비스 정책 수립 1.2 공유·공동활용 문화조성 1.3 인력양성 및 조직체계 강화 1.4 빅데이터 품질관리체계 확립
법·제도	2.1 과학기술 데이터 보존·관리·활용 법적 근거 마련 2.2 과학기술 운영지침 마련 2.3 과학기술 데이터 가치보호를 위한 법·제도 정비
표준화	3.1 과학기술 빅데이터 표준화 추진 3.2 분야별 과학기술 빅데이터 연계 추진 3.3 과학기술 빅데이터 기술맵 제시
오픈 플랫폼	4.1 공통 인프라 조성 4.2 오픈 사이언스 랩 구축 4.3 사이언스 데이터 맵 구축
시범사업	5.1 다부처 융합 사업 기획 및 추진 5.2 분야별 과학기술 핵심 데이터 구축

IV. KISTI 차원의 빅데이터 대응

연구현장의 목소리(1/3)

해외



※ "Challenges and Opportunities", Science, Vol. 33. 2011.

국내

"우리나라의 경우, 응답자의 약 69%가 연구 수행자가 소유와 관리 주체임"

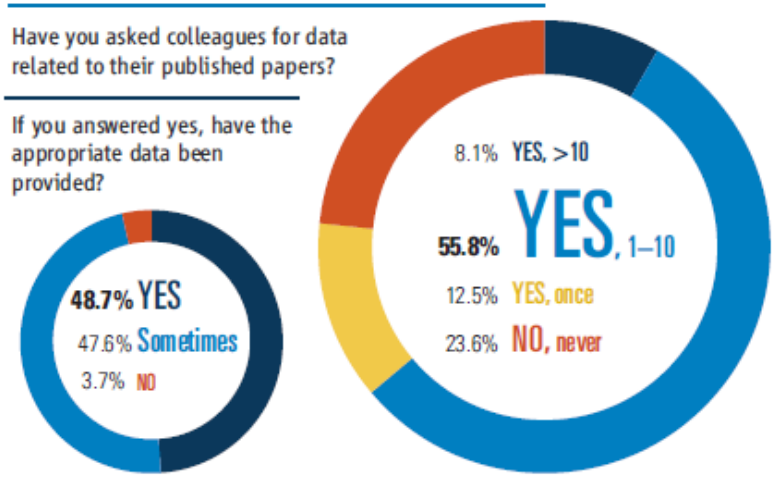
※ 국가과학데이터의 효율적 관리 및 활용을 위한 법제도 기본연구(2012)

"응답자 33% 데이터 분실 경험"

데이터 관리 미흡과 유실

연구현장의 목소리(2/3)

해외



※ "Challenges and Opportunities", Science, Vol. 33. 2011.

국내

"폐쇄된 커뮤니티 내에서 공유 59.4%,
공유하지 않는 경우 30%"

"연구용으로 활용 33.6%,
제공할 의무가 있을 때만 30.2%
적절한 대가를 지불한다면 24.9%"

제한적인 데이터 공유 및 활용

※ 국가과학데이터 공유·융합체제 구축에 관한 연구(2011)

연구현장의 목소리(3/3)

연구 현장에서의 빠른 계산 서비스 필요	저렴한 비용의 Scalable Computing 기술 개발·제공
데이터 관리의 어려움	데이터 관리를 위한 Repository 개발·제공
타 연구 그룹의 데이터 활용 필요	플랫폼 기반 융합 연구 지원
융합 연구를 통한 새로운 발견 필요	분산·연계형 계산·공유 체제 구현
데이터 공유는 부정적	

KISTI 중점 추진사업

과학기술 빅데이터 기반의 Data-Intensive Science 연구 환경 구축

과학기술 빅데이터 거버넌스 체제 구축 지원

- 과학기술 빅데이터 거버넌스 체제 구축 지원
- 국내외 협력 네트워크 구축

과학기술 빅데이터 공유 플랫폼개발

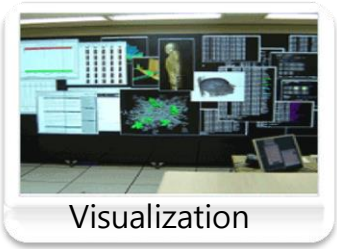
- 과학기술 공동 활용 플랫폼 개발 및 보급 (대용량 분산 저장/관리 기능)

과학기술 빅데이터 핵심/활용기술 개발

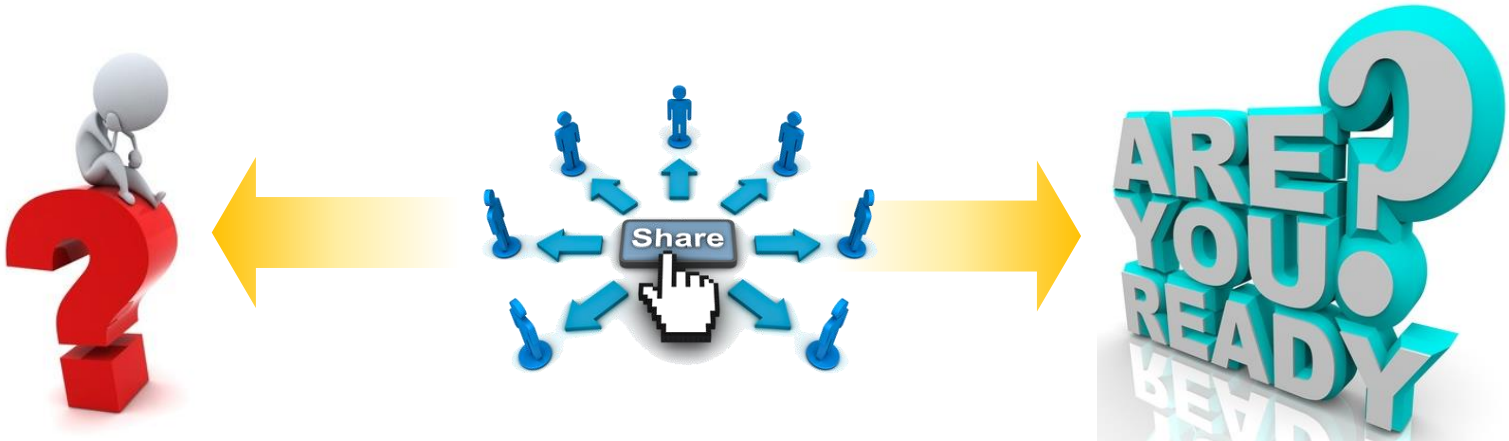
- 데이터 집중형 문제 해결을 위한 병렬 처리 기술 개발 (위성 데이터 분산 병렬 처리)


Supercomputer


Storage


Visualization

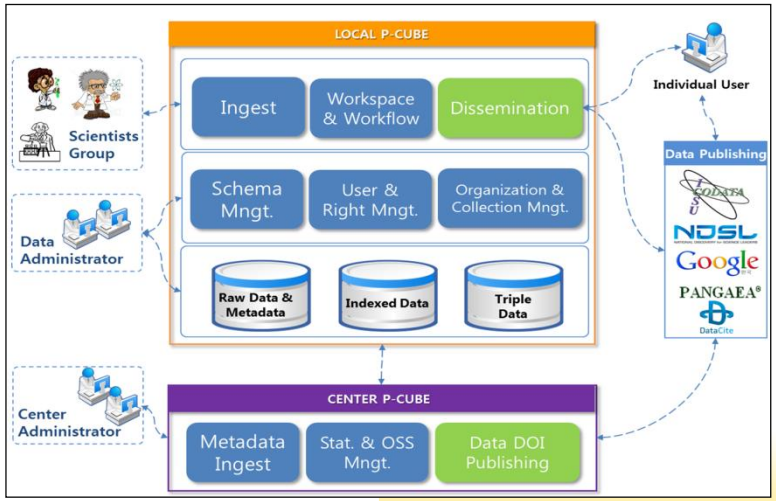
2. 과학기술 빅데이터 공유 시스템(1/2)

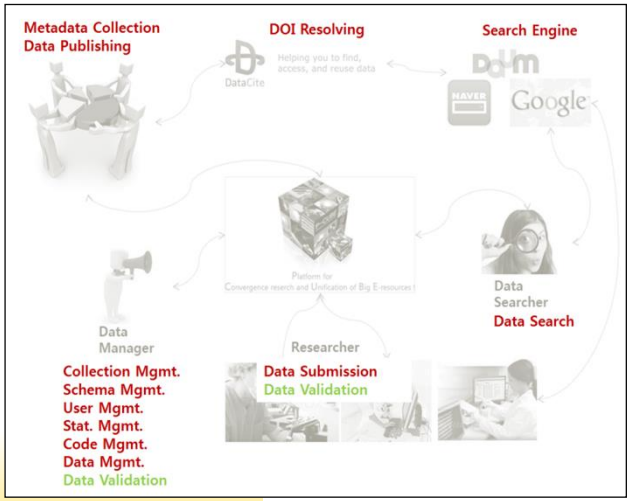


※ 우리나라 연구자의 **약 69%**가 **연구책임자 또는 연구자 개인**이 관리

※ 국내 연구자의 **약 37%**가 실험·관측을 통해 확보한 **데이터 유실 경험**

과학기술 빅데이터 공유 시스템(2/2)





"데이터 관리,
접근을
쉽고 편하게"

KOPRI 극지연구소

- 극지분야 1종

KRISS 한국표준과학연구원

- 금속 5종
- 물리화학 16종
- 보건 7종
- 생명과학 3종
- 에너지 2종
- 재료 4종

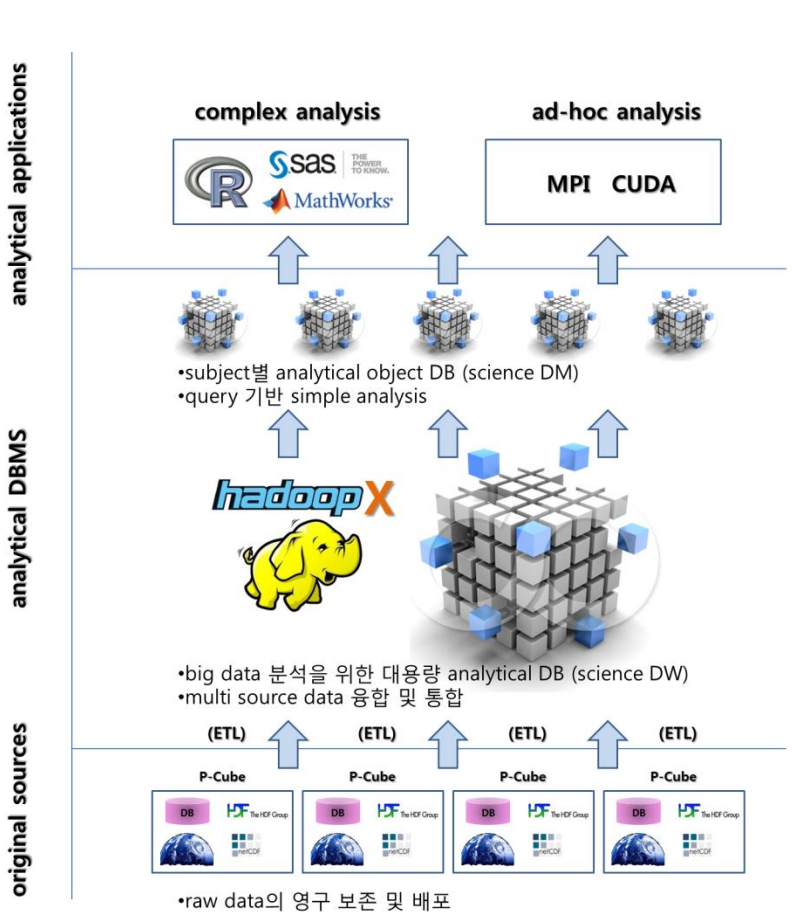
NFRI 국가핵융합연구소 National Fusion Research Institute

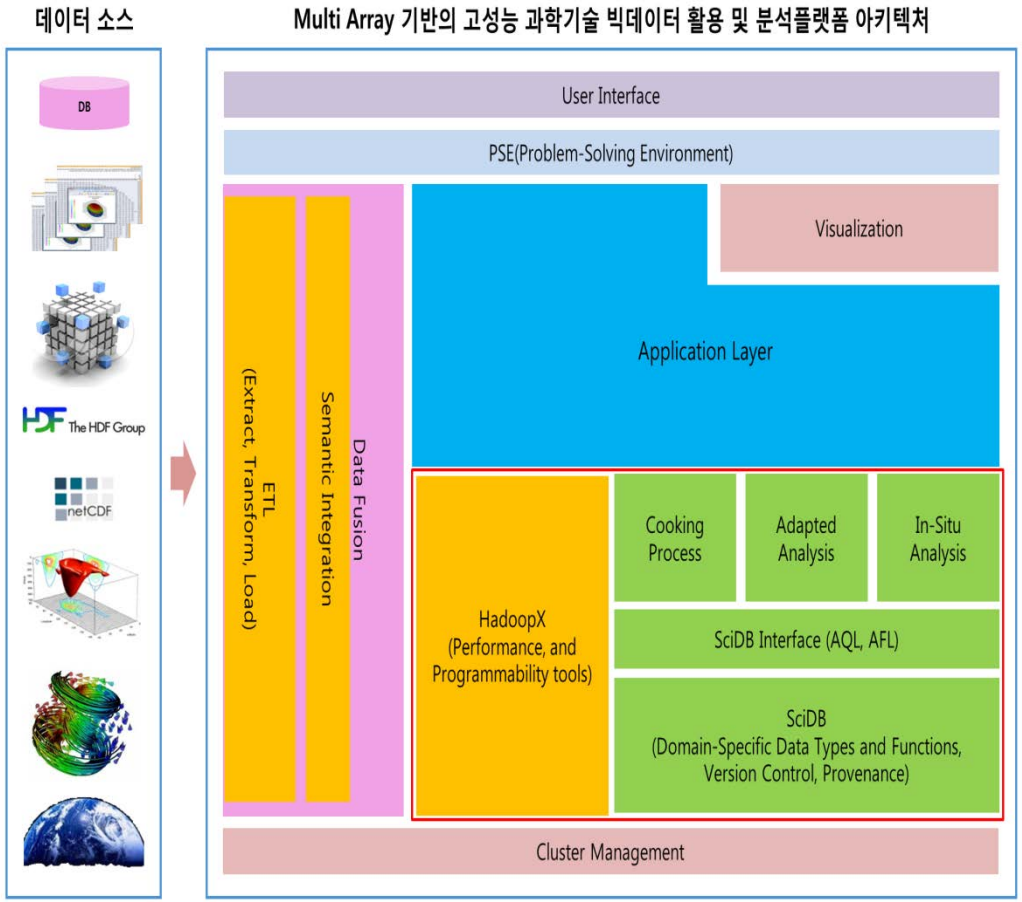
- 플라즈마 1종

KISTI www.kisti.re.kr 한국과학기술정보연구원 Korea Institute of Science and Technology Information

- 인체영상 1종

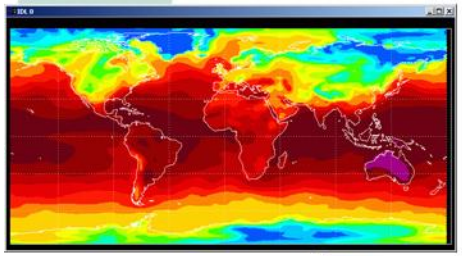
Data-Intensive Science 지원(1/7)



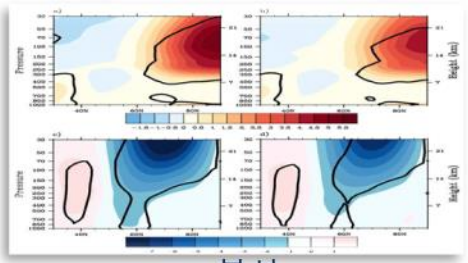


Data-Intensive Science 지원(2/7)

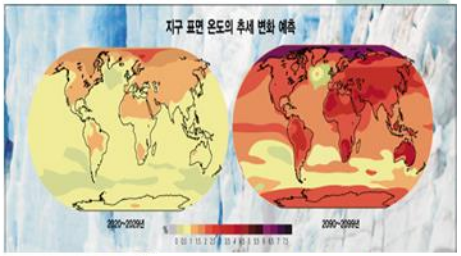
인공위성원격탐사 빅데이터의 자료처리 고성능화를 통한 기후변화 연구



해색인공위성 자료



분석



기후변화 요인 분석

KiSTi
www.kisti.re.kr

- 빅데이터 수집/저장 플랫폼 개발
- 빅데이터 처리/분석 플랫폼 개발

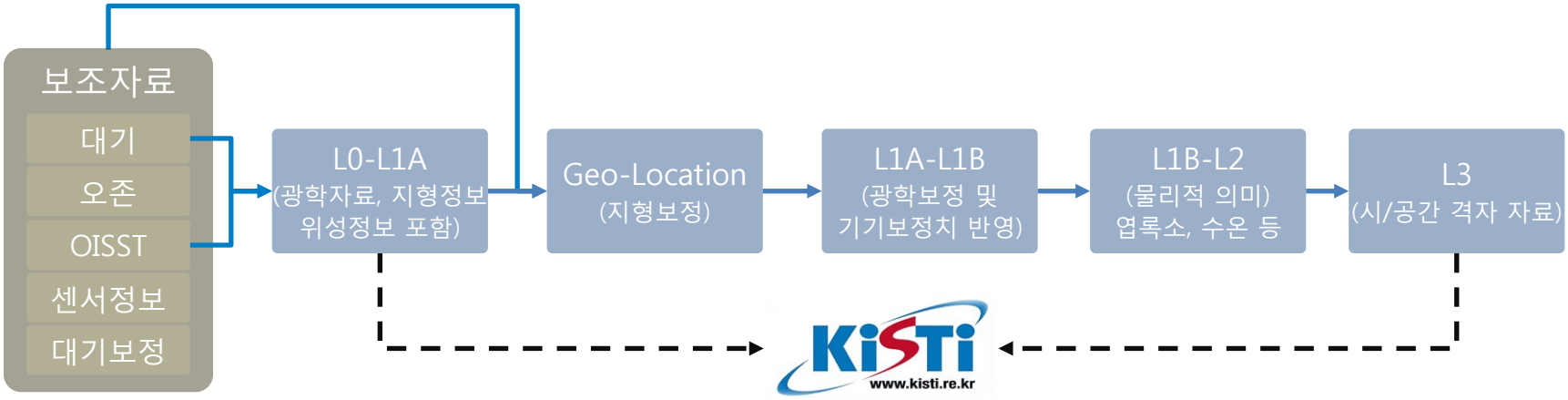
KOPRI
극지연구소

- 한반도-극지역 주변 해역 기후양상 상관성 분석
- 전지구 규모 기후 양상 연구

KIOST
한국해양과학기술원

- 한반도 주변 녹적조 발생 경향 연구

Data-Intensive Science 지원(3/7)



소규모 자료

- 특정 지역
- **약 10,000장** 정도 사용
- **약 53일 소요**
- 개별 연구자 처리 가능

대규모 자료

- 다수 지역 또는 전지구
- **약 500만장** 정도 사용
- **약 72년 소요**
- 개별 연구자 처리 불가능

** 1개 영상당 평균 7분 소요*

고성능 Big Data
S/W 플랫폼 활용

고정밀·대용량 위성자료를
사용한 연구 효율성 제고

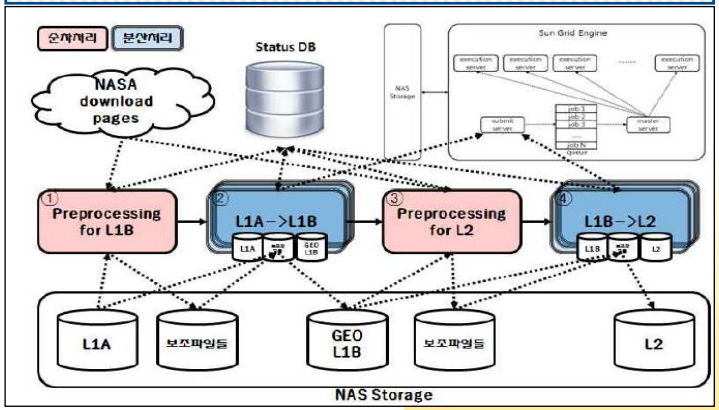
Data-Intensive Science 지원(4/7)

```
==========================
SeaDAS v6.4 MODIS PROCESSING TIMES
==========================
CPU TIMES (user + sys):
======================
L0 to L1A       : 4.19s
L1A to GEO      : 9.40s
L1A to L1B      : 34.93s
L1B to L2       : 267.93s
space binning   : 1.32s
time binning    : 1.17s
L3bin to SMI    : .50s
======================
TOTAL CPU TIME  : 319.44s
```

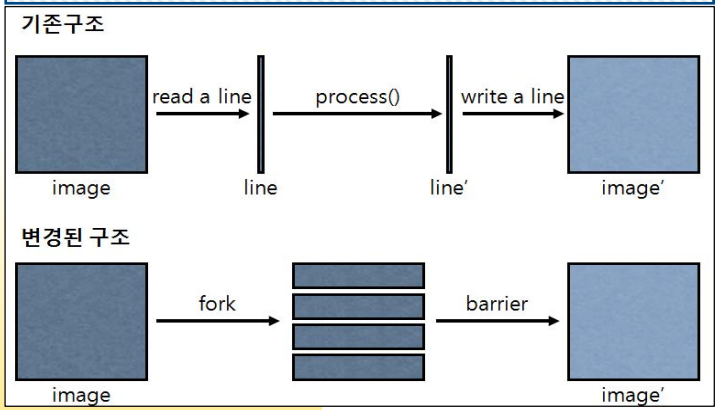
개선 구간

- 한반도 주변 위성영상
- 1,203 파일, 285GB

프로세스 개선



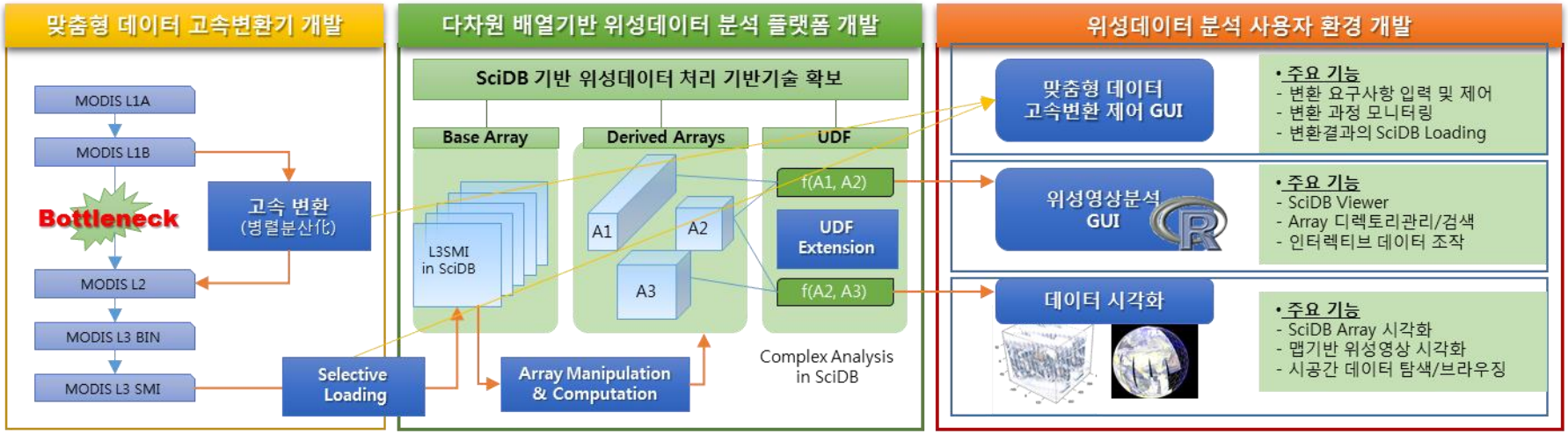
데이터 및 프로세스 병렬화

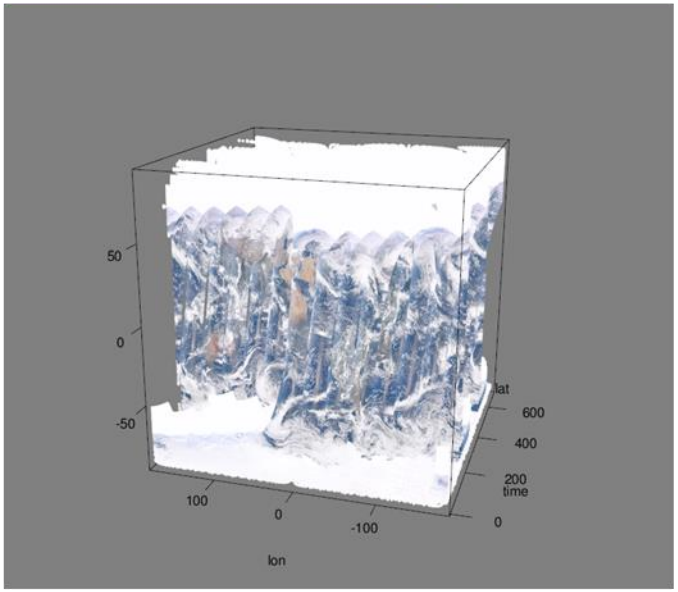


34시간 30분 ⇒ 2시간 53분

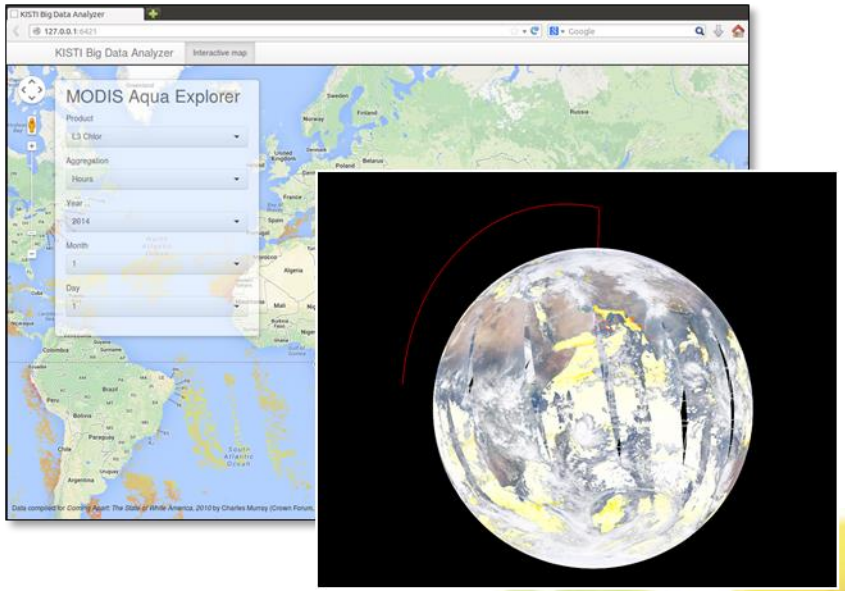
(약 1,193% 성능 향상)

Data-Intensive Science 지원(5/7)



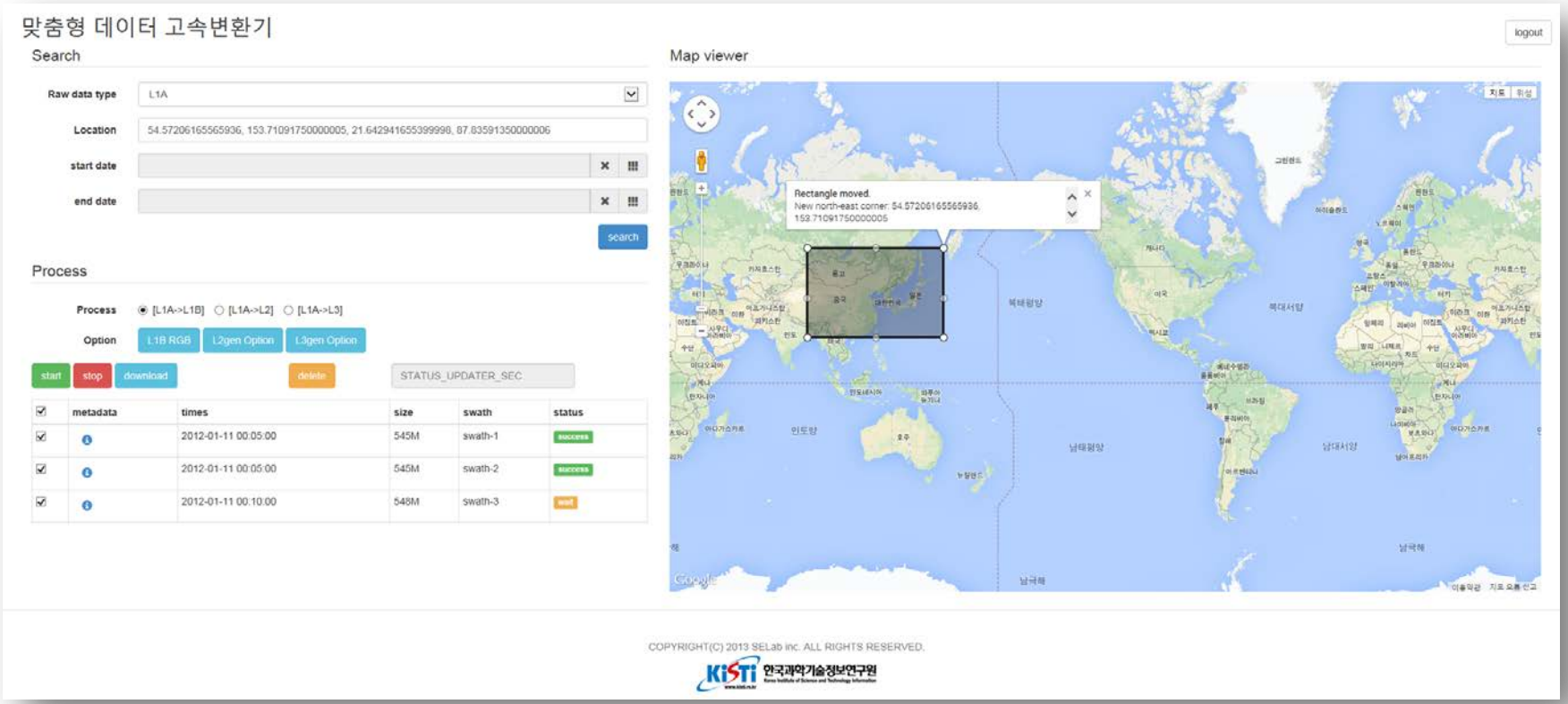


<시공간 차원의 1B RGB 영상 관리>



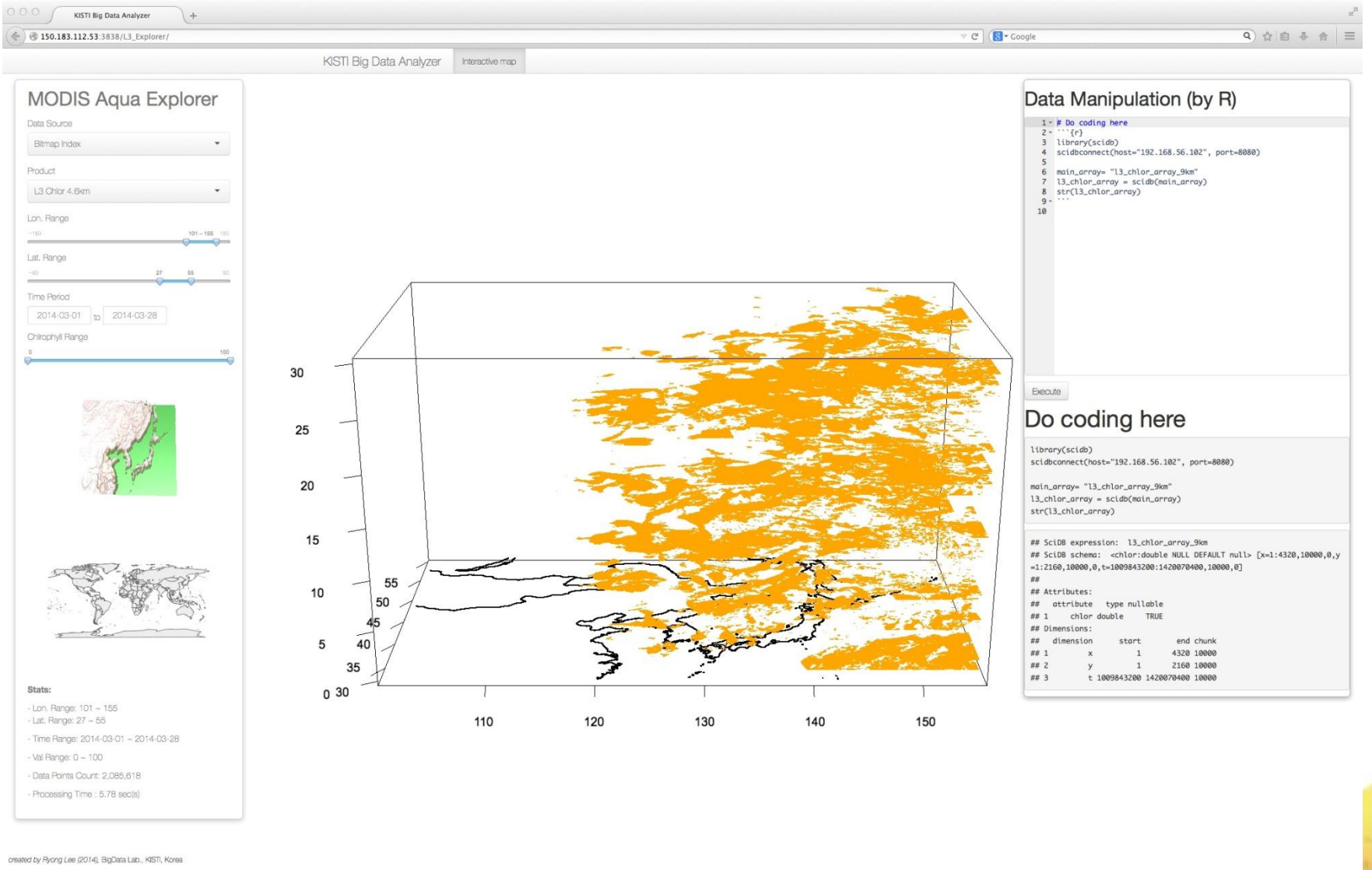
<L2 Chlorophyll 3D Projection>

Data-Intensive Science 지원(6/7)

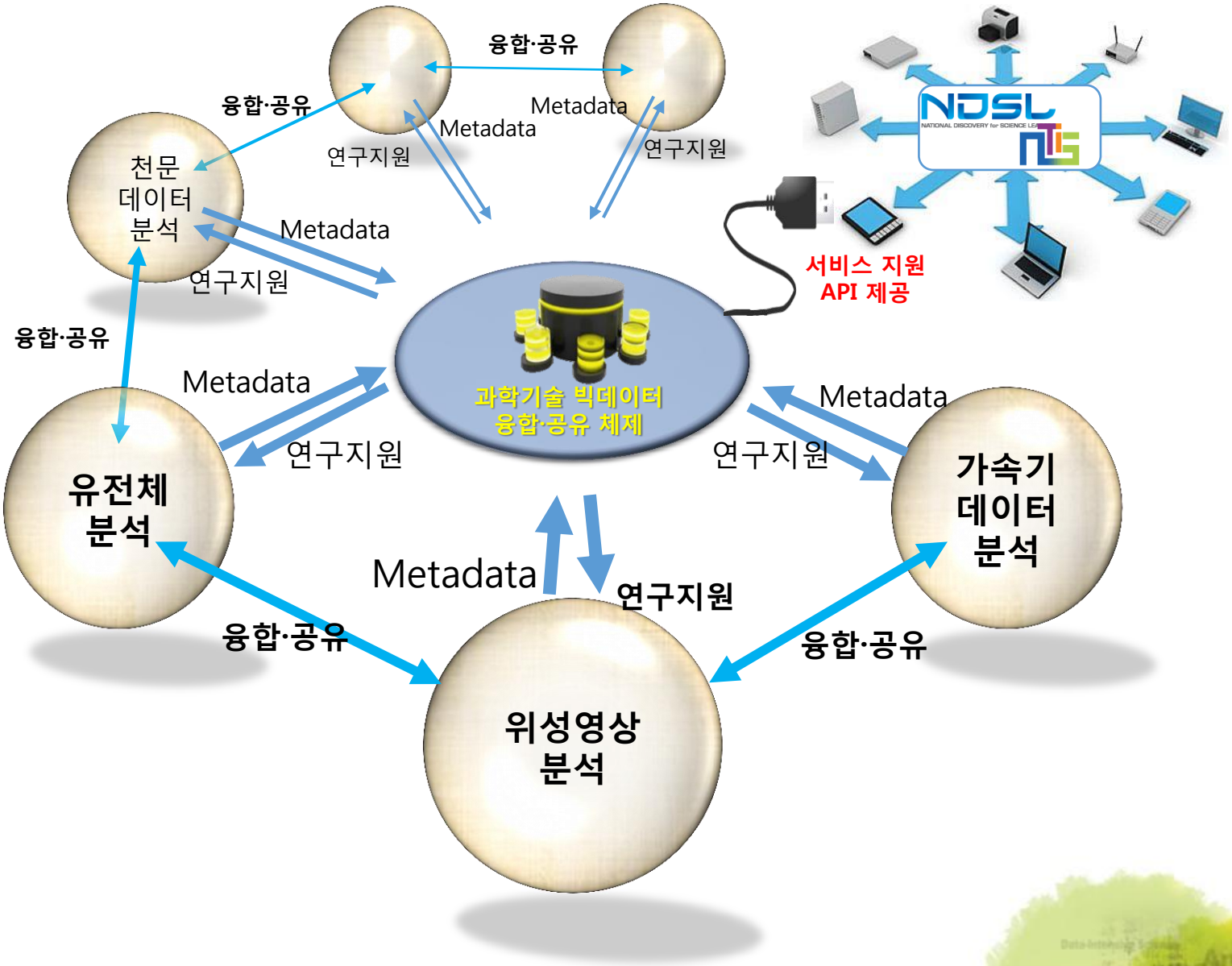


- 이용자 맞춤형 위성영상 빅데이터 변환
- 단계별 위성영상 빅데이터 변환 과정 모니터링
- 이용자별 변환 위성영상 빅데이터 관리

Data-Intensive Science 지원(7/7)



향후 방향


융합·공유
융합·공유
Metadata
Metadata
연구지원
연구지원
천문
데이터
분석
Metadata
연구지원
융합·공유
서비스 지원
API 제공
과학기술 빅데이터
융합·공유 체제
Metadata
연구지원
Metadata
연구지원
유전체
분석
가속기
데이터
분석
Metadata
연구지원
융합·공유
융합·공유
위성영상
분석

감사합니다.