KiSTi 한국과학기술정보연구원 Korea Institute of Science and Technology Information	보 도 자 료	http://www.kisti.re.kr
2016.05.17.(화) 조간(온라인은 5.16. 12:00) 이후 보도해주시기 바랍니다.		
대전(본원): 대외협력실 이식 042 - 869 - 0960 / 강동기 0967 문의: 박경석 과학데이터기술연구실장(042-869-1716)		
배포번호 : 2016-8 배포일자 : 2016.05.16.(월)	매수 : 보도자료 6매	배포처 : 대외협력실

70배 빠른 다차원 빅데이터 분석 기술 개발 성공

- KISTI, 해양 생태 변화 연구 등 대규모 데이터 처리를 위한 핵심 기술 '투픽스' 개발 -

- 한국해양과학기술원 및 극지연구소 2개 기관에 무상 기술 이전 -

□ 한국과학기술정보연구원(원장 한선화, 이하 KISTI)이 기존의 상용 빅데이터 분석 플랫폼보다 약 70배* 빠른 다차원 빅데이터 분석 기술 개발에 성공했다.

*동일한 컴퓨팅 환경에서 전통적인 데이터베이스 관리 시스템(DBMS)이나 빅데이터 처리 프레임워크 하둡(Hadoop) 등을 적용한 시스템과 비교한 결과

□ 이번에 개발한 빅데이터 분석 시스템인 '투픽스(TuPiX, Turning Pixels into Knowledge and Science)'는 클러스터와 같은 빅데이터 처리 환경에서 데이터 처리 및 분석에 소요되는 시간을 획기적으로 줄였다.

o 데이터 저장 및 계산 방식의 변화를 통해 대용량 데이터를 병렬분산처리할 때 발생하는 원형 데이터의 전처리·불러오기·재구성 과정 없이 바로 원형 데이터에 접근하는 방식을 취한다.

□ 투픽스의 또다른 장점은 환경설정에서 손쉽게 필요한 만큼 컴퓨팅 노드를 추가·연동할 수 있다는 것이다.

o 고가의 컴퓨팅 자원뿐만 아니라 대중적인 개인용 컴퓨터(PC) 수준의 사양으로도 클러스터를 구성할 수 있어, 고성능 하드웨어를 도입하기 어려운 소규모 조직의 경우 인프라 구축 비용을 크게 절감할 수 있다.

□ KISTI는 이번에 한국해양과학기술원과 극지연구소에 위성영상 및 해양연구에 필요로 하는 기관의 특성을 고려해 맞춤형 분석 및 가시화 기능을 추가한 투픽스 오션 컬러(TuPiX Ocean Color) 시스템을 무상 기술이전했다.

o 그동안 한국해양과학기술원과 극지연구소 연구자들은 데이터 처리와 분석 기술의 한계로 장기간의 고해상도 데이터를 분석하는 데에 한계가 있었으나, 이번 기술이전을 통해 연구기관, 관측범위, 해상도 및 데이터 규모에 큰 제약 없이 연구를 수행할 수 있게 되었다.

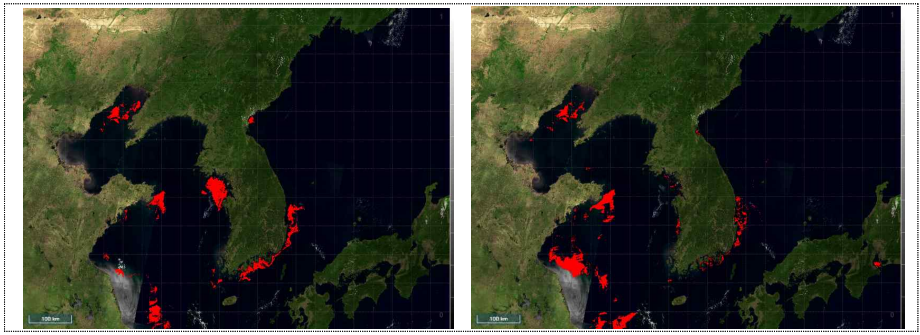
o 현재 KISTI는 두 기관과 함께 투픽스를 기반으로 식물 플랑크톤 번성 패턴 모델과 한반도 연안의 유해 적조종 발생 가능도 모델도 개발해 자연재해로 발생하는 피해를 줄여나갈 수 있는 협업 연구를 진행 중이다.

□ 박경석 KISTI 과학데이터기술연구실장은 "데이터 검색 및 관리에 필요한 비용과 시간을 크게 줄일 수 있어 연구 생산성과 빅데이터 분석 및 관리 효율성 향상에 기여할 것으로 기대"한다며 "빅데이터 플랫폼 확산 및 융합연구 활성화를 위해 투픽스를 기반으로 하는 기관별 특화 시스템을 지속적으로 제공해나갈 계획"이라고 말했다.

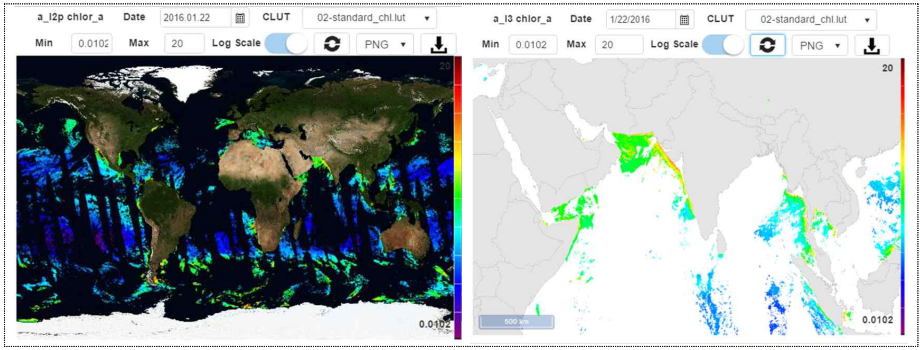
o 향후 KISTI는 핵융합, 유전체 분석, 에너지, 지진 및 해일 등의 과학연구뿐만 아니라 계산 금융, 지리정보, 사회연결망 등 대규모의 계산과 데이터 관리가 필요한 다양한 응용 분야를 발굴해 빅데이터 분석 플랫폼을 확산시켜 나갈 예정이다.(끝)(이어서 참고자료)

[참고자료: 투픽스(TuPiX) 시스템 적용 및 활용 결과]

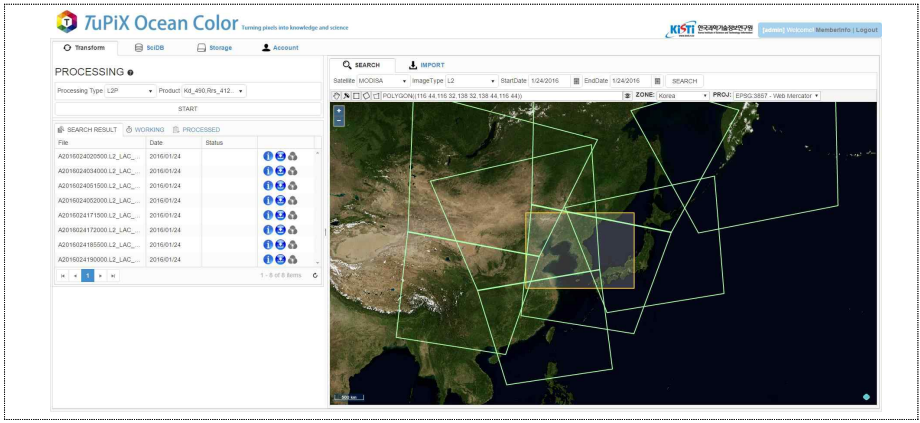
< TuPiX를 활용한 한반도 연안 유해 적조 발생 가능도 분석 >



< TuPiX를 활용한 전 지구 식물 플랑크톤 밀도 분석 >



< TuPiX 기반 대용량 고해상도 위성데이터 처리 >



[참고자료: 투픽스(TuPiX) 시스템 개요]

1. 연구개발 배경/목적

○ 전통적인 데이터베이스 관리 시스템(DBMS)이나 빅 데이터 처리를 위해 개발된 Hadoop 프레임워크 등은 Array 형태의 다차원 데이터 처리에 비효율적이며, 대용량 다차원 데이터를 효율적으로 관리하고 수학적 모델 및 알고리즘 등을 빠르게 분석할 수 있는 기술 필요.

○ 기존 대용량 다차원 데이터 분석은 데이터 적재 시간이 많이 걸려 전체 분석 시간이 크게 지연됨에 따라 이를 혁신적으로 개선할 수 있는 기술 필요.

○ 제한된 컴퓨팅 자원을 효율적으로 활용하여 대용량 다차원 데이터 분석이 가능한 클러스터 기반 플랫폼 필요.

2. 연구개발 수행 내용

○ 과학 기술 데이터 병합 및 배치 기술 개발

- 분산된 노드에서 Raw 데이터 파일들을 접근하기 위해서는 다수의 파일에 대한 접근이 필요하며 데이터 I/O에 대한 오버헤드(특정한 처리를 위해 소요되는 간접적인 처리 시간 및 메모리 등을 의미)가 큼.
- Raw 데이터 파일들을 병합(merge)하여 분산 노드에 배치함으로써 데이터 처리 성능을 향상.
  - 과학 기술 데이터 파일을 노드 개수에 맞게 병합
  - 병합된 각 파일은 분산 노드에 알맞게 배치되며, 사용자의 질의가 오면 해당 파일만을 접근하여 데이터 파일 접근에 따른 I/O 시간 오버헤드를 최소화함

○ Raw 데이터를 직접 분석 가능한 In-Situ 분석 엔진 개발

- 원본 데이터를 분석 시스템에서 처리 가능한 포맷으로 변환하여 적재할 때까지 분석을 할 수 없는 기존 In-Database 기반 기술의 한계를 극복.
  - In-Database 분석 방식에서는,
    - 데이터 분석을 위해 원본 데이터 파일 전체를 데이터베이스에 적재해야 함.
    - 데이터 적재를 위해서는 Raw 데이터 파일을 데이터 적재에 알맞은 데이터 형식으로 변환하는 전처리 과정, 변환된 데이터를 1차원Array 데이터 형식으로 데이터베이스에 로딩하는 과정, 그리고 1차원 데이터를 질의 처리에 적합한

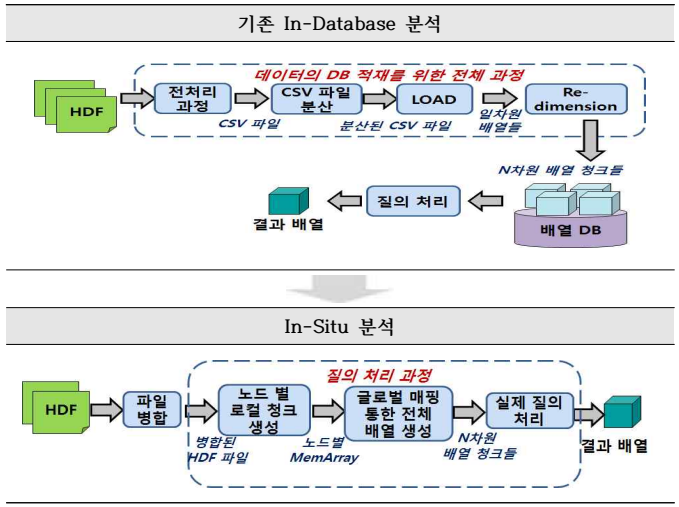
N-차원 Array 데이터로 재구성(re-dimension)하는 과정이 필요함

◦ 대용량의 과학 기술 데이터에 대한 분석을 위해 데이터베이스 적재의 전 과정은 질의 처리 시간에 비해 오버헤드가 너무 큼

• In-Situ 분석 방식에서는,

◦ 전처리 과정, 로딩 과정, 재구성 과정 없이 사용자의 질의를 처리하도록 개선

◦ 대용량의 과학 기술 데이터를 데이터베이스에 적재하지 않으므로 즉시 질의 처리가 가능하며 전체 처리 시간 단축



<In-Database 분석과 In-Situ 분석 비교>

- 데이터의 원형을 보존하고 저장소 간 대규모 데이터 전송을 하지 않고 원본 데이터에 대한 직접 고속 질의, 수학 모델 및 고급 분석 알고리즘에 대한 병렬연산이 가능한 기술 개발

• In-Situ 분석 레이어를 개발하여 다차원 데이터 분석 플랫폼에 적용

• 효율적인 질의 처리를 위해 분산 배치된 과학 기술 데이터에 대한 In-Situ 분석 경로를 제공하여 물리적인 데이터 흐름을 제어함.

• 사용자가 질의를 하면 In-Situ 분석 경로를 통해 데이터 포맷 변환 없이 데이터를 바로 접근하며, In-Situ 분석 레이어의 필터링 기능을 통해 불필요한 데이터를 미리 제거하여 질의 처리 성능을 향상시킴.

- In-Situ 분석을 위한 질의 계획 변경 및 In-Situ Scan operator 적용 기술 개발

• 사용자에 의해 주어진 질의는 질의 계획(Query Plan)을 통해 관련 operator들을 수행하며, Scan operator를 통해 스토리지에 저장된 데이터를 접근함.

• Raw 데이터를 읽고 처리하기 위해 기존의 scan operator를 개발된 In-Situ scan operator로 변경하여 처리할 수 있도록 질의 계획 변경이 필요.

• 사용자가 Join 질의를 수행할 때, In-Situ scan operator를 사용하도록 질의 계획을 변경하여, 과학 기술 데이터를 직접 접근하여 처리함.

[참고자료2: 투픽스(TuPiX) 시스템의 특징]

□ 클러스터 기반으로 다차원 빅데이터의 효율적 분산 병렬 처리

- 단순히 컴퓨팅 노드를 추가하는 scale-out 방식의 확장을 통해 기존에 처리가 불가능했던 대용량 다차원 빅데이터 분석 가능

□ 다양한 분석 알고리즘 지원 및 클러스터 기반 병렬 처리

- 통계모형, 수리모형, 기계학습 등 분석 알고리즘의 병렬 처리

- 다양한 계산 블록을 활용하여 신규 분석 알고리즘의 신속한 개발

□ 분석 결과에 대한 효과적인 시각화

- 분석 결과를 고해상도로 시각화하고 다양한 방식으로 표현

- 시각화에 요구되는 데이터만 선택 처리하여 질의 처리 성능 향상

□ 분석 라이프사이클 전 과정의 효율적 관리와 사용자 협력 지원

- 데이터 획득부터 분석, 시각화까지 전 과정을 단일 UI로 수행

- 대용량 데이터의 체계적 관리와 사용자 간 분석 과정 협력 지원

□ 이용자 분석 환경에 특화된 맞춤형 분석 기능 지원

- 반복적으로 발생하는 분석 워크플로우 및 시각화를 이용자 분석 요구에 맞춰 제공