Special Issue

03

빅데이터를 활용한 수자원정보화(Hydroinformatiics)

이 주 헌
중부대학교 토목공학과 교수
leejh@joongbu.ac.kr

김 태 화
(주)한국정보문화기술원
수석연구원
th600512@korea.ac.kr

김 재 원
(주)쓰리에스솔루션
책임연구원
jwkim@3ss.co.krr

1. 서론

최근 정보통신기술 동향을 보면 빅데이터 및 IoT(사물인터넷)와 같은 신기술 개발 및 보급 확산이 급속하게 진행되고 있다. 이런 흐름은 제4차 산업혁명 사회로 우리를 이끌 것으로 예견되고 있으며 이제 세상은 초연결 지능사회로 발전할 것이다. 이에 발맞추어 수자원 분야에서도 초연결 지능사회에서의 물이용 및 관리는 어떻게 진화해야 하는가?에 대한 문제해결 대안으로 Smart Water Grid와 같은 R&D 및 Hydro-informatics와 관련된 소사이어티 활동이 활발하게 진행되고 있다.

지난 1세기동안 진행된 지구온난화의 영향으로 기상 및 수문 패턴에 변화가 급격하게 일어나고 있다. 우리나라의 경우 연평균 강우량, 계절별 강우량, 강수의 공간적 분포 등 강수의 시공간적 변동성이 더욱 심해질 것으로 예견되고 있으며 이러한 변화는 수자원 개발 및 이용 측면에서 안전성을 저해하는 심각한 위협으로 받아드려지고 있다.

과거 수자원 개발 및 이용방안은 주로 댐을 건설하고 하천을 정비하며, 지하수를 개발하는 것과 같은 구조적 대안으로 진행되었다. 그러나 오늘날 인구증가 및 사회고도화, 지속가능성에 대한 수자원 패러다임의 변화 등으로 구조적 접근이 어렵게 되었다. 이에 따라 비구조적 수자원 활용 방안에 대한 생산성 제고의 필요성이 높아지고 있다. 기후변화에 의해 촉발되는 위협을 저감시키고, 수자원 이용의 효율성을 제고할 수 있는 방안에 대한 관련 연구가 활발하게 진행되고 있다.

수자원 정보화는 수자원 개발 및 이용의 비구조적 방안으로 주목받고 있으며, 빅데이터는 수자원 정보화에 필수적인 핵심 요소 기술로 인식되고 있다. 빅데이터 기술을 활용한 수자원 정보의 생산 · 관리의 이해와 기술역량은 이제 선택이 아닌 필수 항목으로 받아드려지고 있다.

이에 본고에서는 빅데이터에 대한 기술적 이해 및 수자원 활용 방안에 대해 모색하고자 한다.

2. 빅데이터 기술의 개요

2.1 빅데이터의 정의

빅데이터란 디지털 환경에서 생성되는 정형, 비정형 데이터로서 그 규모가 방대하고, 생성주기가 짧은 대규모 데이터를 말하며 관계형 데이터베이스와 같은 기존의 저장 · 관리 · 분석 · 처리기술을 넘어서는 방대한 규모의 데이터이다. 그리고 단지 데이터 자체뿐만 아니라 데이터를 처리 · 가공 · 분석하여 유용한 정보를 제공하는 기술도 함께 포함하는 광의적 개념이다.

일반적으로 빅데이터의 특징을 Volume(크기), Velocity(실시간 생성), Variety(다양성)이라는 세 단어의 영문 앞 글자를 따서 3V로 정의하기도 한다. Volume(크기)은 개별 시스템으로 관리 및 처리하기 어려운 크기인 수십 테라바이트에서 페타바이트 수준의 데이터를 말하며, Velocity(속도)는 아주 빠른 속도로 데이터가 생성되는 것을 말하고, Variety(다양성)는 비정형, 반정형, 다중구성 데이터 형태와 관련이 있다. 최근에는 빅데이터 특징을 데이터 볼륨, 속도 및 다양성 외 복잡성, 가치를 더하고 이에 빅데이터 분석을 포함해 빅데이터 기술 요소를 "6V"로 확대하여 인식하고 있다.

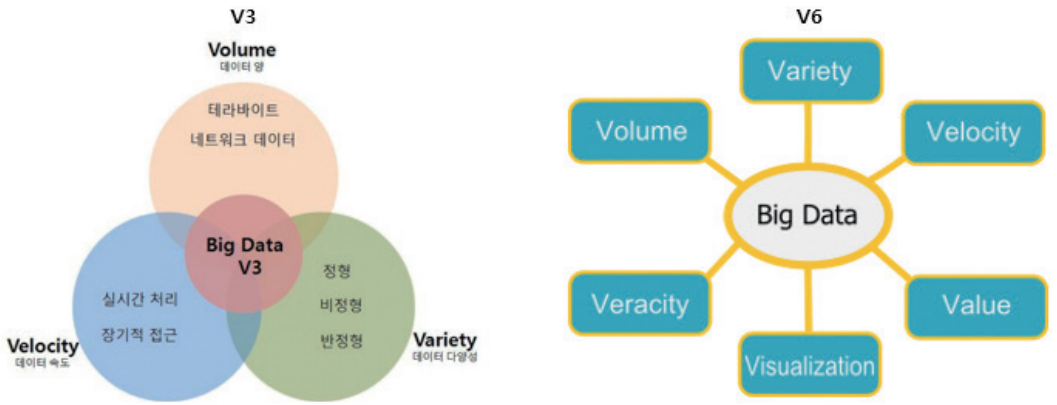

그림 1. 빅데이터의 정의 및 특징 : V3와 V6

빅데이터 기술의 발전은 많이 사람들에 의해 성취되고 있는 사회적 현상이다. 따라서 다양하고 복잡하며 빠르게 진화하고 있다. 이런 빅데이터 기술의 발전 동향은 빅데이터에 대한 인식을 어렵게 할 뿐만 아니라 기술 수용의 진입장벽으로 작용하고 있다. 빅데이터 개념에 대해 간단하게 이해하기 위해서는 Google의 노력과 성과를 주목할 필요가 있다. 빅데이터의 정의에서 등장하는 3V 혹은 6V를 만족하는 시스템을 인류 최초로 만든 회사가 Google이기 때문이다. 빅데이터 기술의 뿌리에 이들의 공헌이 있다. 따라서 필자의 소견이지만 빅데이터하기란 곧 Google처럼 데이터를 모우고, Google처럼 데이터를 처리하며, Google처럼 데이터를 사용하는 것이라고 여겨진다.

우리가 다루는 수자원 데이터는 IoT(사물인터넷) 기술에 힘입어 급속하게 Volume과 Velocity가 증가할 것이며, 대기 · 토양 · 해양 등 물리적 데이터와 인구증가, 물이용패턴 등 사회적 데이터와 같은 여러 종류의 다양한 데이터가 상호 융합되면서 급속하게 빅데이터화 될 것이다. 빅데이터 기술은 수자원

데이터를 효과적으로 처리할 수 있는 방법을 제공하며, 거대하고 복잡하며 빈번하게 변하는 데이터 취급의 피로감으로부터 전문가를 보호하고 올바른 지혜를 갖도록 도와 줄 것이다.

2.2 새로운 가치

빅데이터 트렌드는 과거 버려지던 것을 경제적이고 효과적으로 처리할 수 있게 됨으로써 창출된 새로운 가치 창출 전략이라도 할 수 있다. 20여년전만 해도 전산시스템이 고가여서 수집되는 모든 데이터를 저장하는 것이 경제적으로 합리적인 선택이 되지 못했다. 그래서 꼭 필요한 데이터만 선별하여 관리하고 나머지는 영구보관처리 또는 삭제되었다.

현재는 잉여 전산자원이 풍부하며, 이를 활용할 수 있는 여러 가지 기술도 발달된 상황이다. 이제는 과거와 달리 모든 데이터를 추적하여 새로운 가치를 탐색할 수 있게 되었다. 혹자는 이를 Long Tail 법칙에 의한 새로운 가치 창출이라고도 한다.

수자원 전문가 관점에서 빅데이터의 가치는 수문모델을 더욱 정교하게 만들 수 있는 충분조건을 경제적으로 마련할 수 있게 되었다는 것이다. 예나 지금이나 우리는 기상 · 수문시스템이 어떻게 변화할 것이고 이로 인해 우리가 어떤 영향을 받을지에 대해 궁금해 하고 있다. 빅데이터는 미래예측을 더욱 정교하게 만들어 줄 것이다. 뿐만 아니라 과거에는 수십억의 연구비를 투자해야했던 것을 이제는 수백에서 수천만원 수준으로도 연구를 진행할 수 있게 되었다는 것이다. 이제 남은 것은 빅데이터 기술을 도입하고 학습하여 적용하고 성과를 도출하는 것이다.

2.3 빅데이터 기술

빅데이터 기술은 매우 복잡한 생태계를 이루고 있으며 전세계 수많은 기관과 회사가 빅데이터 생태계 구축에 참여하고 있다. 필자의 견해로는 빅데이터 솔루션에 대한 가장 쉽고 경제적인 접근은 오픈소스 솔루션을 활용하는 방법이라고 여겨진다. 오픈소스 솔루션은 인터넷에서 다운로드 받아 설치할 수 있으며 이용 및 수정에 어떤 제약도 받지 않는다. 또한 현재 서적 및 블로그, 유튜브 등에서 관련 정보가 풍부하게 다루어지고 있어 이에 대한 참조도 쉽게 얻을 수 있다는 장점도 가지고 있다.

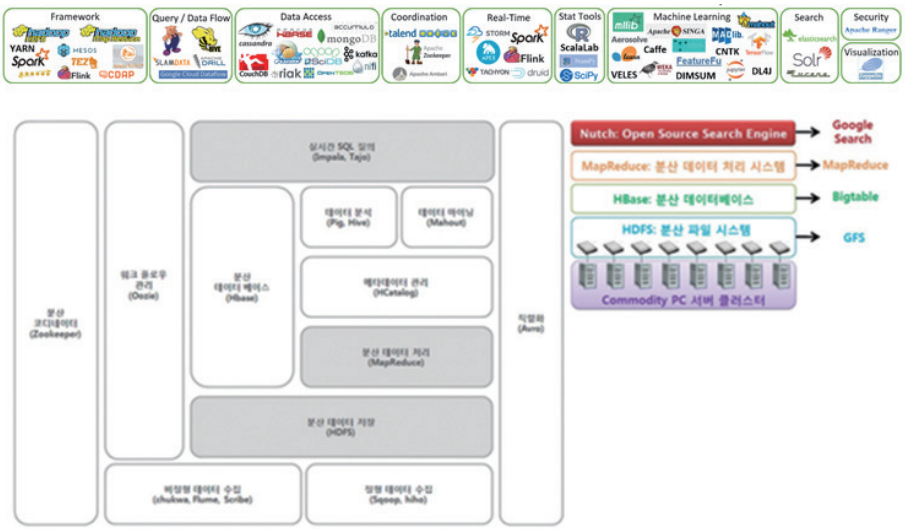

그림 2. 오픈소스 기반 빅데이터 기술 및 하둡 스택

오픈 소스를 활용한 빅데이터 플랫폼은 Hadoop 기반, MongoDB 기반, ElasticSearch 기반, Spark 기반 등 다양한 방식으로 구성할 수 있다.

빅데이터 처리의 가장 큰 특징은 다수의 컴퓨터를 동시에 사용하는 병렬처리 방식을 취한다는 것이다. 이와 같이 저사양의 컴퓨터를 동시에 다수 활용하여 정보처리 성능을 향상시키는 전략을 Scale-Out 전략이라고 한다. 이와 대조적으로 오라클과 같은 관계형 데이터베이스 기반으로 수십 테라바이트가 넘는 데이터를 처리하기 위해서는 엄청나게 고가의 시스템을 요구하며, 데이터량이 증가할수록 시스템 증설해야하는 문제가 발생하게 된다(Scale-Up전략). Scale-Out 전략은 처리해야 할 일이 많아지면 컴퓨터를 더 많이 투입하는 방식으로 문제 해결에 접근하는 것으로 볼 수 있다. 컴퓨터가 더 투입되면 병렬화 문제가 발생하면서 통합 시스템 운영이 복잡해지게 되는데 빅데이터 기술은 이런 문제를 사용자로부터 은닉하여 투명하게 해준다. 즉, 복잡성 문제를 신경쓰지 않고 업무문제에만 집중할 수 있도록 도와준다. 따라서 시스템 운영 복잡성은 일정하게 유지하면서 정보처리성능은 향상시킬 수 있게 된다.

표 1. 오픈소스 기반 빅데이터 기술

기술		설명
하둡 스택	Hadoop	대용량 데이터를 분산처리 할 수 있는 자바기반의 프레임워크 HDFS 및 Map/Reduce 등으로 구성
	HBase	컬럼 기반 NoSQL 데이터베이스, Google의 Bigtable을 오픈소스화한 데이터베이스
	Zookeeper	빅데이터 서버 시스템 관리, 분산 코디네이터
	Hive	유사 SQL 기반 빅데이터 처리
	Mahout	기계학습 알고리즘 기반 빅데이터 처리
	Flume	비정형 데이터 수집, 전처리
	Sqoop	관계형 DB와 연계
MongoDB		도큐먼트 기반 NoSQL 데이터베이스
ElasticSearch		도큐먼트 기반 NoSQL 데이터베이스
Spark		빅데이터 실시간 분석을 위한 스트리밍 플랫폼
Storm		빅데이터 실시간 분석을 위한 스트리밍 플랫폼
Gnu-R		통계 계산과 시각화을 위한 언어 및 소프트웨어 환경
Grafana		빅데이터 시각화

빅데이터 처리는 여러 가지 빅데이터 기술을 조합하여 수행한다. 사용목적에 따라 여러 가지 기술을 다양하게 조합하게 된다. 데이터 종류, 처리 방식, 분석 결과 이용 등에 따라 여러 가지 방식으로 기술을 조합할 수 있다.

빅데이터 처리과정은 데이터 수집, 데이터 전처리, 정보저장. 관리, 정보처리. 분석, 지식 가시화로 진행되며, 데이터 소스, 지식을 활용하는 서비스 분야가 무엇인지에 따라 일부 단계를 건너뛰거나 반복 수행되기도 한다. 빅데이터를 통해 지식을 활용하기까지 각 단계를 지원하는 데 필요한 공통 소프트웨어를 빅데이터 처리 플랫폼이라 한다.

그림 3. 빅데이터 처리 과정

앞서 설명한 바와 같이 수집, 저장. 관리, 처리, 분석, 지식 시각화 및 공유 등 6가지 요소기술 분류에 따라 요소기술 설명 및 해당기술을 정리하면 아래 표 2와 같이 나타낼 수 있다.

표 2. 빅데이터 처리 과정별 빅데이터 기술

과정	설명	빅데이터 기술
수집	조직내부와 외부의 분산된 여러 데이터 소스로부터 필요한 데이터를 검색하여 자동으로 수집하는 과정과 관련한 기술	ETL, 크롤링 엔진, 로그 수집기, 센싱, RSS, Open API 등
공유	서로 다른 시스템 간의 데이터를 공유하게 하는 기술	멀티 테넌트 데이터공유, 협업 필터링 등
저장	작은 데이터라도 모두 저장하여 실시간으로 저렴하게 데이터를 처리하고, 처리된 데이터를 더 빠르고 쉽게 분석하는 기술	병렬 DBMS, Hadoop, NoSQL 등
처리	엄청난 양의 데이터를 저장 · 수집 · 관리 · 유통 · 분석을 처리하는 일련의 기술	실시간 처리, 분산 병렬 처리, 인-메모리 처리, 인-데이터베이스 처리
분석	데이터를 효율적으로 정확하게 분석하여 비즈니스 등 다양한 영역에 적용하기 위한 기술	통계분석, 데이터/텍스트 마이닝, 예측, 최적화, 소셜 네트워크 분석 등
가시화	기존의 단순 선형적 구조의 방식으로 표현하기 어려운 빅데이터 자료를 시각적으로 묘사하는 기술	편집 기술, 정보 시각화 기술, 시각화도구

2.4 빅데이터 시스템 아키텍처

빅데이터 시스템은 빅데이터를 처리하기 위해 빅데이터 기술을 조합하여 구축한 플랫폼으로 마치 공장이나 플랜트와 같이 구성된다. 그림 4는 하둡을 기반으로 한 빅데이터 시스템 아키텍처이며 시스템 아키텍처란 일종의 설계도와 같은 것으로 이해할 수 있다.

3. 빅데이터기반 수자원 정보화

3.1 수자원 정보화의 현황

우리나라는 1999년에 “물관리 업무의 효율화”, “물관리 업무 DB화에 의한 국가 수자원 경쟁력 강화”, “물 정보 공개를 통한 국민의 알 권리 충족 및 열린 정부 지향”, “물 정보의 대국민 서비스를 통한

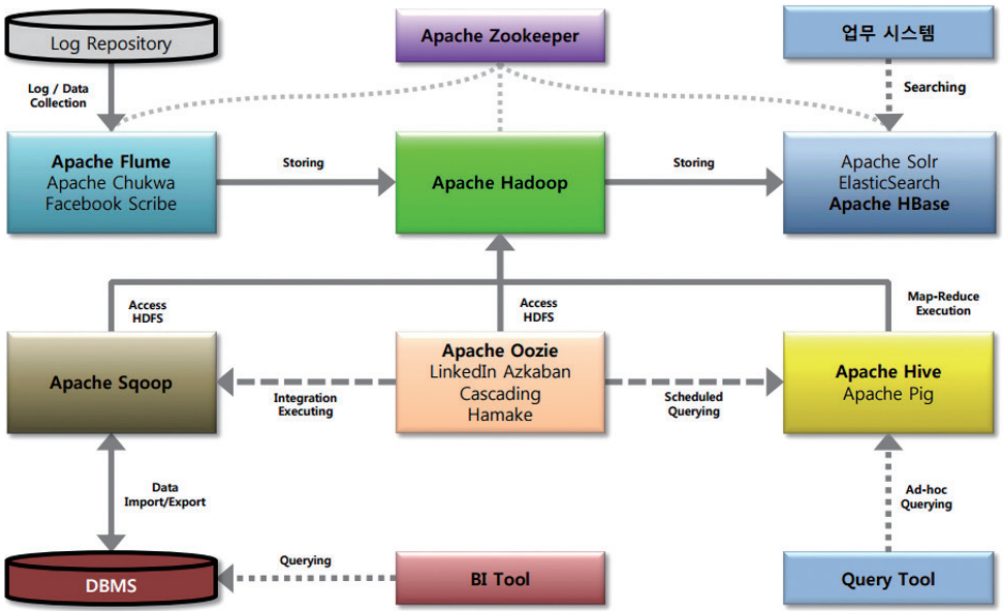

그림 4. 하둡기반 빅데이터 아키텍처

물에 대한 인식 제고" 라는 네 가지 목표를 달성하기 위해 국가수자원관리종합정보시스템 구축 기본계획이 수립되었고, 2002년에는 국가 차원의 지속적인 물관리정보화 추진을 위해 물관리 정보 표준화 기본전략이 수립 · 제시되었다.

우리나라의 물정보 관리체계는 국토교통부 산하 한강홍수통제소를 중심으로 관계부처와의 시스템 연계 및 정보 제공을 통하여 관리되고 있으며, 수질은 환경부, 재해관리는 국민안전처가 관리하고 있다. 물정보 관리기관에서 생성되는 다양한 물정보가 국토교통부의 수자원관리종합정보시스템(WAMIS), 환경부의 물환경정보시스템(WIS), 농림축산식품부의 농촌용수종합정보시스템(RAWRIS) 및 국민안전처의 재난정보 공동활용시스템 등을 통하여 정부기관, 물관련 기관 등에 제공되고 있다. 이와 같이 물정보 관리기관은 기관별 또는 목적별로 다양한 물정보를 생산하여 관리하고 있다.

그간 수자원정보화의 중요한 기능을 유지했던 물관리정보유통시스템(WINS)의 활용성을 제고하고 그간의 운영결과를 통하여 도출된 문제점을 개선하기 위해서는 기존유통기관의 신규 항목(품질검토 완료자료)의 지속적인 발굴과 확대가 필요하고 파악된 물정보와 관련된 기관과 MOU 체결 등을 통하여 필요한 자료를 즉각적으로 획득하기 위한 다양한 방안을 강구하는 등 운영방안의 전환이 필요한 시점이라고 본다.

3.2 수자원 정보화의 정체와 도약

우리나라의 수문/기상/수질/재난 등 수자원정보화시스템은 현재까지 각 부처별로 개별 시스템을 구축하고 운영하며 이를 더욱 전문화하는 방향으로 발전시키고 있으며 이는 수자원정보화의 기반 인프라를 구축하는데 이바지하였다. 하지만 현재에 이르러서는 기후변화 및 사회 고도화로 효율적인 물관리에 대한 이슈가 제기되고 있으며, 이에 따라 수자원정보의 통합관리에 대한 필요성이 지속적으로 증대되고 있다. 지금까지의 수자원 정보화 수준만으로

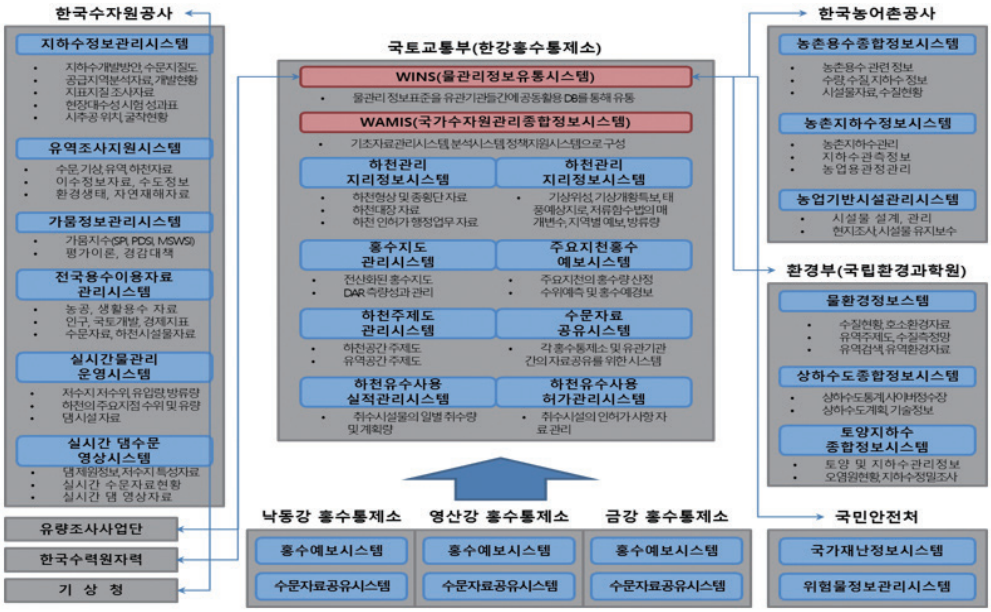

그림 5. WINS–WAMIS 관계도

는 수자원 개발 및 효율적 이용에 관한 차세대 요구를 만족하는데 분명한 한계가 있다. 이 수준을 뛰어넘는 더 높은 수준의 수자원 정보화가 이루어져야만 한다고 생각된다.

표 4. 수자원 정보화 현황 및 문제점

구분	설명
데이터 수집 및 저장측면	수문정보(국토교통부), 기상(기상청), 수질(환경부), 재난(국민안전처), 중소하천(지방자치단체), 다목적댐(한국수자원공사), 농업용저수지(한국농어촌공사) 등이 각 기관의 목적에 맞게 각자 데이터베이스를 구축하여 운영
	각 기관은 타 기관과 협정을 통하여 데이터를 수집하여 각자 구축한 데이터베이스에 저장하고 있으나 데이터 중복성 및 데이터 품질관리에 어려움
데이터 공개 및 공동 활용측면	수요자의 needs 분석을 통한 정보의 공개 및 포털 구축 필요
	다자간 데이터 연계는 일부 기능을 위한 데이터 연계에 불과하여 국가적인 차원에서 데이터 연계 및 정보공동 활용을 통한 장기적인 대책 시급
데이터 예측 및 운영 측면	타 기관에서 받은 데이터의 정확성이 확보되지 않은 상태에서 각 기관마다 확보한 데이터를 이용하여 기관의 역할에 맞는 예측정보를 생산하기에는 어려움
	빅데이터 기반 수문기상관리체계 구축 및 장단기 예측을 통한 적극적인 운영이 필요함에 따라 국가차원의 통합 수문기상자료 관리 시급

효율적인 물관리를 위해서는 유역기반 물정보의 통합관리가 필요하며 물정보 관리기관 간 정보공유가 필수적이다. 지금까지 매우 다양한 시도가 있었으나 형식적인 연계 및 관심 부재로 활용도가 미흡한 수준에 있으며 이는 기술적, 관리적, 정책적 원인으로부터 기인하는 문제일 수 있다.

하지만, 현재는 기술적으로 분산정보처리를 가능하게 하는 빅데이터 기술이 개발되어 보급되고 있으며 이제는 분산환경에서 데이터를 공유하고 처리하는 것에 대한 기술적 제약사항이 사라지고 있다. 정책적 측면에서는 수문기상협력센터와 같은 조직이 만들어져 각기 다른 이해 집단간의 협력을 도모하고 있다. 국내 · 외 다수 기관에서 이제 빅데이터 기술의 수용을 통한 도약을 진행하고 있으며 기존의 한계를 극복하고 새로운 가능성을 탐색하고 가치를 창출하는 활동을 진행하고 있다. 우리나라의 경우 빅데이터에 기반한 수자원 정보화 사례로 기상청에서 추진하고 있는 기상기후 빅데이터 활용 기상융합 서비스를 들 수 있다.

3.3 빅데이터 기술 도입을 위한 준비

수자원 관리에 빅데이터 기술을 활용하면 선행 적용 사례와 같이 훌륭한 성과를 기대할 수도 있을 것이다. 그러나 이를 위해서는 많은 준비가 필요하다는 것을 알아야 한다.

• 인재 교육

많은 빅데이터 기술이 프로그래밍 기술과 밀접하게 결합되어 있다. 일예로 빅데이터 분석에 자주 사용되는 NoSQL 플랫폼은 SQL언어를 사용할 수 없다는 것을 의미하며, 데이터 해석을 프로그래밍 기법으로 접근해야 한다는 것을 의미한다. Java, Python, Scalar 등의 언어가 여기에 사용되는 도구이다. 또한 빅데이터 플랫폼은 리눅스 환경 친화적이며 우리가 익숙한 윈도우에서도 사용할 수 있지만 좋은 성능이 나지 않는다. 과제의 규모가 커지면 결국에는 리눅스로 갈 수 밖에 없으며 이들은 하나같이 숙달 시간을 요구하는 것이다.

수자원 분야에 빅데이터를 적용하기 위해서는 새로운 정보 분석 프로세스에 훈련된 인재가 절대적으로 필요한 것이며 교육 프로세스를 변경하고 이에 맞는 교육을 실시하여 미래형 빅데인터 인재를 키워야 할 것이다.

• 데이터 공유

빅데이터 분석에 꼭 필요한 것이 바로 데이터이다. 그런데 우리나라의 경우 데이터를 획득하기가 너무 어렵다. 게다가 요즘에는 공공기관 망분리 정책으로 대용량의 수문기상데이터를 구하는 것이 정말 어려워졌다.

좋은 열매를 맺기 위해서는 우선 좋은 토양이 준비되어야 하며 수문기상에 관한 데이터는 최대한 일반에게 공개되어 연구에 활용되어야 한다. 많이 활용되면 될수록 데이터 품질이 높아지고 해석 기법이 고도화 될 것이다. 따라서 수문기상협력센터와 같은 다기관 협력기관을 통해 데이터를 공급할 필요가 있으며 실시간 데이터 공급이 보안상 문제가 있다면, 아카이브 데이터라도 공급하는 것이 적극적으로 추진되어야한다.

4. 결론

기후변화, 인구증가, 1인당 물수요 증가 등으로 인해 세계 각국은 지속 가능한 수자원 확보의 불확실성이 증대됨에 따라 수자원 관리체계를 공급위주에서 수요관리로 패러다임이 전환하고 있다. 특히, 우리나라는 최근 몇 년간 경험했던 것과 같이 기후변화로 인한 수자원 변동량이 더욱 심해지고 있어 기존의 공급위주에서 수요관리체계로의 전환이 절실히 요구되고 있다. 현재 우리나라는 시기별, 연도별, 지역별로 강우량 편차가 심해서 물관리 여건이 매우 불리하며 물이용과 홍수 및 가뭄대비 측면에서

모두 취약한 실정이다.

따라서 현재 보유하고 있는 수자원을 효용성, 지속 가능성, 공평성 측면에서 잘 활용하여 물이용 효율을 높이기 위해서는 유역 내 한정된 수자원을 기관 간, 지역 간 장벽 없이 효율적으로 활용하여 가뭄, 홍수, 수질환경 등의 문제를 예방 및 해결하도록 방안을 강구할 필요가 있다.

향후 안정적인 수자원 확보 기반을 조성하고 홍수, 가뭄 등 피해를 저감하고, 장기적인 필요 수자원의 안정적인 공급을 위해서는 수문 및 기상 자료 등을 연동한 빅데이터 기반의 국가 물관리 체계로 조속히 전환하여, 다양한 이해집단 간 상호조율을 통한 수자원의 통합 관리 체계를 구축할 필요가 있다. 또한 기후변화에 따른 강수변동성 증가로 물관련 재해에 대한 예측성 저하 및 재해 가능성이 증가함에 따라 국가 경제 및 산업뿐만 아니라 지역 경제에 악영향을 미칠 수 있어 기상자료 및 수문자료 등의 빅데이터 인프라 조성을 통한 수문기상 기반의 분석 및 예측기법 향상을 기하여 정확한 예측에 기반한 선행적 위험관리 대응체계를 구축할 필요가 있다.

더욱이 조만간 출현할 것으로 예견되는 인터넷을 기반으로 하는 초거대 지능체를 이용하여 물에 관한 지능화를 수행할 수 있다면 수자원 개발과 관리 분야에서 많은 해결할 수 있을 것으로 기대된다.

감사의 글

본 원고는 2015-2016년도 Kwater의 재원으로 수문기상협력센터의 지원(2015-WR-RR-270-1097)을 받아 수행된 연구 성과의 일부로서 지원에 감사드리며 연구보고서 내용의 일부와 동일한 내용이 있을 수 있습니다.

Hadoop Ecosystem을 활용한 Hybrid DW 구축사례,(2013)
한국정보화진흥원 (2013), 빅데이터전략센터, 빅데이터 기술 분류 및 현황
수문기상협력센터 (2015), 빅데이터 기반 수문기상정보 활용체제 구축 및 가뭄정보 생산기술 개발 (1차년도), 한국수자원공사
http://mattturck.com/2016/02/01/big-data-landscape