편집위원 : 안창원 (ETRI)

빅 데이터 분석 기술동향과 활성화 과제

박종만, 엄태원*, 김하진
한국과학기술정보연구원, *유한대학교

요 약

빅 데이터의 관심이 인프라 및 분석기술 자체에서 가치창조 측면으로 이동하면서 가치정보를 효율적으로 발굴, 분석, 추출, 활용하기 위한 차세대 고급 분석 기술 및 기법이 요구되고 있다. 이에 빅 데이터 활용기반구축을 위한 정부 및 기업의 대응이 시급한 시점이다. 이 연구는 빅 데이터 활용기반 구축과 분석기술 개발에 도움을 주고자 빅 데이터 분석의 핵심기술동향을 분석하고 실천과제를 제시한다.

Ⅰ. 서 론

빅 데이터는 조직 내의 흐르는 전기와 같아서 필요하면 수도꼭지처럼 틀어 사용하는 데이터[1]라는 표현처럼 그 영향과 필요성에 대한 담론은 더 이상 불필요하다. 최근 빅 데이터에서 가치정보를 초고속으로 효율적으로 발굴, 분석, 추출하는 차세대 기술과 아키텍처의 개발 및 산업화가 국내외에서 급속히 진행되고 있다. 이에 글로벌 기술시장에 대한 대처 및 국내 성장기반 구축을 위해 빅 데이터 확보 및 실시간 가치정보 창출기술의 개발과 그 활용기반의 구축이 시급한 시점이다. 이 연구의 목적은 빅 데이터의 활용 및 현황 파악, 융합지식 창출을 위한 빅 데이터 기반 가치정보의 발굴, 분석, 추출 기술 및 동향, 관련특허동향 등을 조사 분석하여, 관련기업의 생존전략 수립과 비즈니스 기회 창출, 정부의 성장산업화 추진과 정책지원 방안의 도출과정에 도움을 주고자 함에 있다. 논문은 1장 서론, 2장 빅 데이터 기술개요, 3장 빅 데이터 기술동향, 4장 빅 데이터 활성화 과제, 5장 결론으로 구성된다.

Ⅱ. 빅 데이터의 기술개요

1. 빅 데이터 현황

1.1 빅 데이터 분석관련 이슈

빅 데이터의 관심이 인프라 및 분석기술 자체에서 가치창조 측면으로 이동하면서 기업, 국가 이상의 인류적 문제해결 수단으로 진화 중[2]이라는 거대인식도 있지만 실질적으로는 구글이나 IBM 등 IT업체의 기술 마케팅적 수순을 갈파하고 산업측면에서 IT기술 로드맵과 전략의 빈번한 변동, 추진과 실천계획의 파편화는 피해야 된다. 진화의 선도나 적응속도도 중요하지만 가치창조를 위한 인프라 및 고급 분석기술의 정착이 더 필요하다. 빅 데이터는 기존에 존재해 왔고 빅 데이터의 분석 개념이나 툴도 전혀 새로운 것은 아니다. 클라우드와 SNS서비스를 통한 비즈니스모델링과 의사결정과정에서 비정형적이고 비구조적인 대규모 데이터에 대한 분석의 필요성이 대두되자, 최적 의사결정을 위한 지능화된 정보획득 방법론으로 새롭게 이슈화되고 있다. 빅 데이터의 생성 및 처리의 필요성이 생태계 수요공급체인의 어디에서부터 촉발되었는지는 불분명하나, 기존의 제한적인 정형적 의사결정이나 비즈니스 서비스수준을 넘어 글로벌 기업경쟁의 경제 원리나 거버넌스 차원에서 데이터주권확보, 국가경쟁력확보의 필수적 과제로까지 확대되는 경향이 있다. 정부차원의 빅 데이터 관련 중심이슈[3]는 국가 안위 및 재난대비, 경제가치 향상을 위한 매크로 측면에서 빅 데이터 추진역량강화를 위한 공공데이터의 연계통합, 정부와 민간의 융합추진, 공공데이터 진단체계 구축, 사회적, 기술적 핵심기반 확보를 위한 법제도 개선, 인력양성, 인프라 구축, 분석 및 운영 기술 확보 등의 추진이다. 산업차원의 빅 데이터 관련 중심이슈는 기존 트랜잭션 데이터를 통한 관계적 정형성데이터 이외에, 소셜미디어 데이터와 같은 비 관계적 비 정형성데이터를 통합 분석하여 경쟁우위를 점유하고자 하는 비즈니스 모델링으로 설명될 수 있다. 공공 및 민간 부문의 소셜 네트워크서비스나 비즈니스 모델링에 대한 이슈도 다양하나 공통적 중심이슈는 아직 빅 데이터의 인프라 구축과 분석처리 및 운영기술의 구현관련 요소들로 압축될 수 있다.

1.2 빅 데이터 기술관련 시장

빅 데이터 관련 서버, 스토리지, 네트워킹, SW, 서비스 기술

표 1. 빅 데이터 세계시장규모($mil)[4]

구분	2012	2013	2015	증가%
서버	803	1032	1657	27.3
스토리지	1224	1968	3479	61.4
네트워킹	242	368	620	42.4
SW	1851	2476	4625	34.2
서비스	2721	3883	6538	39.5
합계	6841	9727	16919	39.4

의 세계시장은 〈표 1〉과 같이 2010년 이후 연 39%이상 성장하여 2015년 169억 달러에 이를 것으로 전망되고 있다. 매출 구성은 서버 9%, 스토리지 21%, 네트워킹 4%, SW 27%, 서비스 39% 순으로 서비스와 SW의 비중이 65%이상으로 SW기반의 솔루션 경쟁이 치열할 것으로 보인다. 비즈니스 분석 SW 시장에서 데이터웨어하우징(DW) 플랫폼부문이 2011년 전년대비 15.2%의 성장률로 가장 빠른 성장세이며 분석 애플리케이션 SW부문은 13.3%, BI 및 분석 SW부문은 13.2% 성장을 보여 DW 플랫폼을 통한 분석이 활성화되고 있음을 보여주고 있다. 인프라측면에서는 스토리지 성장률이 61%, 네트워킹이 42%이상으로 급격히 성장하고 있어 관련 요소 및 응용기술의 경쟁이 예상된다.

1.3 빅 데이터 분석기술의 이해

빅 데이터의 '빅'의 의미는 데이터의 규모나 기술적 처리, 방법 및 기법을 포함하는 포괄적 의미에 가깝다. 빅 데이터 처리를 위한 "21세기 스위스아미 칼"로 표현되기도 하는 기존의 Apache Hadoop은 데이터서비스의 고 가용성을 목적으로 하며 대규모데이터를 분산 저장 및 처리를 위한 확장성을 가진 공개 소프트웨어 프레임워크[5]로 '빅'에 대한 다중적인 의미대상은 아니다. 빅 데이터 분석기술 측면에서 분석에 대한 용어 사 용과 이해는 분석기술의 적용 및 기술개발, 시스템구축의 방향성 결정에 치명적이다.

TDWI의 조사결과 〈표 2〉[6], 빅 데이터 분석의미를 모르거나 기존의 분석, 고급분석, 기타기술의 의미와 혼용하고 있는 것으로 나타났다. 실제 빅 데이터 분석이란 용어보다 기존 BI를 기반으로 사용되어온 고급분석이란 용어에 친밀도가 높다고 한다. 이는 빅 데이터를 분석을 새로운 패러다임이라기보다 기존의 분석 환경에서 처리해야할 다른 유형과 크기의 데이터의 분석으로 보기 때문이다. 빅 데이터분석의 이해도가 다양한 국내 생태계에서도 이해관계자의 적확한 판단이 요구된다.

표 2. 빅 데이터 분석의 인식

빅 데이터 분석	18%
고급분석(Advanced Analytic)	12%
분석(BI)	12%
대량 및 대규모 데이터 셋 분석	7%
데이터웨어하우징	4%
데이터마이닝	2%
예측분석	2%
기타	43%

빅 데이터에 대응한 고급분석 실행형태조사[6]에서 빅 데이터 분석의 인식정도와 범위를 그래프〈그림 1〉로 가늠해 볼 수 있다. 조사대상 조직의 대부분(74%)은 분석방법 및 툴의 형태나 빅 데이터 여부에 관계없이 분석을 시행한다. 즉 조사기업의 40%는 빅 데이터 없이 고급분석을 시행하며, 조사기업의 34%는 빅 데이터를 적용하여 고급분석을 시행한다. 본 논문에서는 고급분석의 의미로 빅 데이터 분석의 포괄의미로서 사용한다.

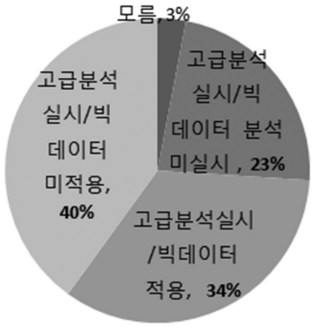

그림 1. 고급분석의 실행 형태

2. 빅 데이터 기술구성

2.1 기술 구성 영역

전문 툴을 이용한 빅 데이터의 분석영역은 〈그림 2〉와 같이 소규모 데이터 셋과 관계형 데이터를 대상으로 했던 기존의 비즈니스 인텔리전스(BI) 영역의 분석과 비교된다.

대규모 셋	전문 툴을 이용 한 빅 데이터분석	확장성 제한
소규모 셋	분석가치 열위	기존 BI
	비 관계형 데이터	관계형 데이터

그림 2. 빅 데이터 분석의 영역[7]

빅 데이터 분석기술의 위상은 데이터 분석을 위한 데이터의 수집, 전처리 및 저장, 분석처리 및 운영, 지능화처리 및 지식관리의 단계에서 서비스 유형(인프라, 플랫폼, 애플리케이션)과

HW/SW 공급관계[8]로 파악될 수 있으며 〈표 3〉과 같이 요약되어 진다.

표 3. 빅 데이터 서비스 및 HW/SW 유형

구분		SW/HW 공급 유형		
		분석 SW	처리/저장 SW	HW
서비스 수요유형	APP	맞춤형 분석 툴 제공	처리 및 저장 툴 제공	서비스인프라 생성 유통
	플랫폼	비종속적 클라우드 서비스	오픈소스 SW플랫폼 서비스	클라우드기반 데이터서비스
	인프라	클라우드 컴퓨팅 인프라기반	분산병렬처리 SW운영	데이터센터 구축, 공급자솔루션 이용

빅 데이터의 기술적 특성[9]에 따른 기술영역은, 첫째, 데이터의 크기(volume)에 따른 저장관련 대응기술은 2020년까지 디지털 데이터의 34% 이상이 클라우드 기반으로 생성 및 유통될 것이라는 전망[10]과 같이 클라우드 간의 연결이나 연합에 의한 복합적 데이터 클라우드 기술로의 진화를 유발할 것으로 보인다. 이는 분산 컴퓨팅에서 다수 클러스터의 연합적 컴퓨팅과 저전력 기반의 마이크로서버로 구성 되는 조합(fabric) 컴퓨팅 기반과 클라우드 플랫폼 형태로의 진화 방향성을 보여준다. 이는 그리드 컴퓨팅[11]을 위한 인프라의 스마트 조합기술로도 설명 가능하다. 둘째, 데이터 다양한 형태(variety)에 따른 인프라 대응기술은 데이터들의 연결 및 결합을 위한 데이터링크와 메타데이터 기술의 필요성과 비 관계형(No-SQL)데이터베이스 기술을 지향한다. 셋째, 데이터 처리속도(velocity)에 대한 대응기술은 다중 소스로부터의 데이터 수집과 스트림데이터의 전처리, 정보의 저장 및 분석처리, 고정도의 지식화와 지능 발굴과정에서 실시간 통찰력을 제시하는 플랫폼의 구성과 인 메모리 컴퓨팅, 멀티코어 컴퓨팅, 병렬 · 분산처리 기술을 지향한다.

빅 데이터 분석기술의 대부분은 개방형 분산 클라우드 컴퓨팅 환경과 BI를 기반으로 한다. BI에서의 분석기술은 통계, 예측, 최적화를 기반으로 향상된 계획과 의사결정 등을 지원하기 위해 정보를 지식수준의 형태로 변환하는 프로세스와 도구로서 현상과 결과적인 관점을 제시하는데 초점이 있다. 고급분석기술은 비즈니스 상황을 예측하고 최적 의사결정을 지원하기 위해 구조 및 비 구조화 된 복잡한 형태의 데이터에서 요인들 간의 상관관계와 의미 데이터의 패턴을 식별하고 예측하기 위한 기법 및 서술관련 기술들로 대용량의 데이터로부터 숨겨진 패턴의 발견과 상황예측에 초점이 있다[12].

서술관련 기술의 분석기법은 분류, 군집, 연관, 추정, 서술 등에서 예측분석을 위한 다양한 모델형태와 알고리즘을 갖고 있다. 이러한 예측분석에서 개방적 도구로의 변화와 비구조적 및 비정형적 데이터 처리 및 분석 지원기능, 최적화 분석기능이 강조되는 경향이 있다. 이외에 고급분석을 위한 콘텐츠분석, 텍스트분석, 실시간분석 기술이 있다. 각 분석기술의 내용을 요약하면 다음 〈표 4〉와 같다.

표 4. 고급 분석 기술

기술 구분	기술 영역
예측분석	분류, 군집, 연계, 추정, 서술 관련 모형
콘텐츠분석	통신 전후의 추론, 통신특성 추론 및 표현
텍스트 분석	언어자료의식별 및 정보재생, 자연어처리, 명명개체 인식, 식별개체 패턴인식, 지시어중복, 관계/사실/사건/추출, 판별분석, 텍스트분석
실시간 분석	인-DB 분석, 분석처리 데이터웨어하우스, 고속 인-메모리, 다중 프로세스 병렬프로그래밍

2.2 분석 플랫폼과 인프라 구성

빅 데이터의 가치는 분석을 통해 의사결정을 대체하거나 지원을 통해 실현된다. 빅 데이터 처리 및 분석 기술은 개별 기술의 집적이 아니라 핵심 기술을 중심으로 구성되는 플랫폼 기술이다[13]. 빅데이터 분석 플랫폼은 빅 데이터 처리 인프라를 기반으로 하며[14], 그 구성 기술은 〈그림 3〉과 같이 데이터의 수집 및 전처리, 데이터 저장 및 관리, 분석 및 가시화 기술관련 SW들로 통합솔루션을 구축한다.

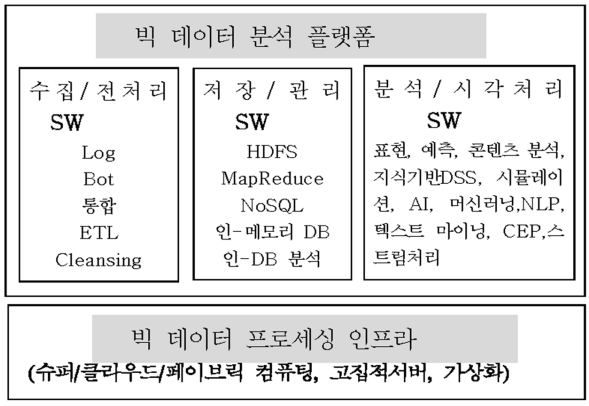

그림 3. 빅 데이터 분석 플랫폼

빅 데이터의 분석기술은 마케팅 측면에서 미래지향적 솔루션과 툴로서 과장되는 면이 있지만 실질적 구축단계에서 과거의 수학/통계적 분석, 의사결정 및 지원시스템, 전문가시스템, 지식관리시스템, AI 및 BI시스템, 고객관계관리 기술에 대한 융합기반을 요구한다. 자연어처리, 시멘틱, 큐레이션 등의 웹 통합 검색기술로 빅 데이터 분석을 지원하기도 하며, 기존의 데이터 수집 및 처리 기능을 강화한 솔루션보다 분석 툴 및 분석전문가, 지능적 의사결정 지원솔루션의 통합을 더 중요시 하는 경

우도 있다. 증대된 비정형데이터의 분석 및 활용 목적성에 따라 기존의 데이터 마이닝, 기계 학습, 자연 언어 처리, 패턴 인식 기법 외에 텍스트마이닝, 오피니언 분석, 소셜네트워크 분석, 클러스터 분석기능의 추가구성이 요구되기도 한다. 산업계에서 사실상의 표준으로 인식되는 Hadoop시스템의 기본구조는 〈그림 4〉[15]와 같이 표현된다.

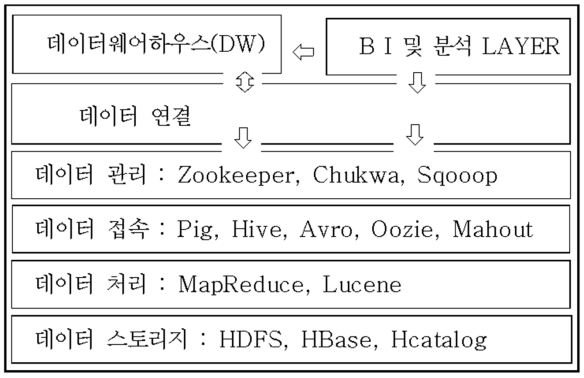

그림 4. Hadoop의 구성

그 구성은 분산파일시스템(HDFS), 대용량 분산병렬처리 프레임워크(MapReduce), SQL기반 데이터웨어하우스(Hive), 스크립트 프로그래밍 언어(Pig), 이기종 데이터교환 프레임워크(Avro), 컬럼구조의 저장소(HBase), 데이터수집 및 HDF S 저장프레임워크(Flume), DBMS자료의 HDFS입출력(Sqoop), 작업흐름 스케줄러(Oozie), 분산 클러스터 시스템관리 코디네이터(Zookeeper), 하둡 자체 로그수집 분석관리(Chukwa) 개방형 소프트웨어 구조(Framework)등의 수십 개 에코시스템으로 구성되며 100개 이상의 아파치재단 프로젝트가 운영되고 있다. 빅 데이터가 2020년 35,200 엑사바이트 규모[16]로 성장하는 전망을 전제로 이에 맞는 적절한 수퍼 컴퓨팅 능력도 필요하다.

2.3 분석 기법과 기술

빅 데이터 분석기법과 기술[17]은 빠른 진화가 진행되고 있으며 IT기술전반의 융합관련성을 갖고 있다. 빅 데이터 분석기법은 시장반응 분석을 위한 그룹 간 비교시험(A/B시험)을 위한 기법, DB속 관계성 룰의 발견 기법, 데이터 카테고리 식별을 위한 분류 및 군집분석 기법(감독 및 비 감독 학습 형태), 웹을 통한 크라우드 소싱 기법, 분석의 정확성과 통찰을 위한 다중 소스데이터분석 및 통합기법(센서신호 처리 및 자연어처리 기법, 판별기법), 통계 및 기계학습의 결합으로 패턴을 추출하는 데이터마이닝 기법(룰 학습, 군집 및 분류, 회귀분석), 다중 예측모델에 의한 학습기법, 유전알고리즘에 의한 최적화 기법, 자연어 기반의 기계학습(판별분석), 패턴인식과 최적화를 위한 비선형적 신경네트워크 기법, 네트워크분석 기법, 최적화 기법(스케줄링, 의사결정, 투자분석, R&D 포트폴리오 등), 분류를 위한 예측모델링 기법, 데이터마이닝을 위한 회귀분석 기법, 판별분석을 위한 자연어처리기법, 시계열분석과 데이터결합을 통한 신호처리 기법, 회귀분석과 시뮬레이션을 이용한 공간분석기법, 통계기법, 시각화기법을 포함한다. 빅 데이터 기술은 분산DB시스템 기술, 데이터 웨어하우스나 마트에 저장된 데이터를 분석, 표현, 보고 하는 BI 소프트웨어기술, 분산 데이터시스템 기반의 개방 데이터베이스 관리시스템 기술, 클라우드 컴퓨팅 기술, 데이터 웨어하우스/마트와 ETL 기술, 분산시스템 기술, 분산데이터 스토리지 기술, 분산파일시스템 기술, 개방 분산시스템 기반의 처리 소프트웨어 공개 프레임워크 기술, 개방 분산 비 관계형 데이터베이스 모델 기술, 다중데이터소스의 Mas h-up 기술, 메타데이터 기술, 통계 및 그래프 개방 소프트웨어 및 언어 기술, RDBMS와 SQL기술, 정형/반 정형/비정형 데이터처리기술, 실시간 처리기술, 시각화처리기술(강조텍스트, 군집분석, 문서진화, 공간정보흐름) 등이다.

Ⅲ. 빅 데이터 기술동향

1. 기술동향 분석

1.1 분석기술 분류

빅 데이터의 분석프로그램은 다양하며 사용자 조직구조, 기술 및 방법에 따라 툴의 형태 및 제원의 선택이 가능하도록 옵션으로 설정된다. 빅 데이터 분석기술을 성장잠재성(x축)과 사용 확약정도(commitment) 즉 사용자의 기술채택 가능성으로 설정한 그래프로 기술을 분류하고 동향을 파악한다. 성장잠재성은 기술 고유의 발전 특성이며 사용 확약정도는 수요기반의 상용화 채택기술의 동향에 대한 특성으로, 2개 특성을 고려하면 기술적 대응방안의 지향점 파악이 가능하다. TDWI의 조사 자료[6]를 토대로 기술동향을 분석 제시한다. 성장잠재성 순위별 빅 데이터 분석 기법과 형태는 ①고급 데이터의 시각화, ②인메모리 데이터베이스, ③실시간 보고서 및 대시보드, ④텍스트 마이닝, ⑤마이닝 및 예측 등의 고급분석, ⑥시각적 발견, ⑦예측분석, ⑧개인 클라우드, ⑨CEP(Complex Event Proce ssing), ⑩데이터마이닝 점수화, ⑪ Hadoop,⑫인-데이터베이스 분석, ⑬HW/SW 가속기, ⑭페쇄루프 분석입출력, ⑮MapReduce, ⑯인-라인분석, ⑰데이터웨어하우스 기기, ⑱No-SQL/비 인덱스 DBMS, ⑲열 기반 저장엔진, ⑳공공 크라우드(Crowd), SaaS, 분석용 Sandbox , Extreme SQL, 데이

터웨어하우스의 혼합 워크로드, EDW안의 분석, 데이터웨어하우스 목적의 DBMS, EDW외부 분석데이터베이스, 통계분석, 중앙 EDW, 분석용 데이터마트, 거래 처리용 DBMS, OLAP 도구, 수작업 코드에 의한 SQL기법 등이다.

1.2 기술군의 동향 분석

가. 첫 번째 기술 군 동향

〈그림 5〉에서 성장잠재성이 급 상향이고 3년 내 사용가능성이 거의 화실시 되는 첫 번째 기술군인 7개 기술 (①고급 데이터의 시각화, ②인-메모리 데이터베이스, ③실시간 보고서 및 대시보드, ④텍스트 마이닝, ⑤마이닝 및 예측 등의 고급분석, ⑥시각적 발견, ⑦예측분석은 현재의 빅 데이터 분석, 데이터웨어하우징, BI의 기술동향을 명백히 반영한다. ⑤항의 마이닝 및 예측 등의 고급분석은 예측분석, 데이터마이닝, 통계분석, 컴플렉스 SQL, 데이터시각화, AI, NLP, 분석지원 데이터베이스 등을 포함하는 기술 집합이다. 현재 BI에서 OLAP를 넘어 고급분석으로의 천이과정을 보여준다.

가장 성장잠재성이 큰 분야는 고급 데이터의 시각화(ADV)기술로 대부분의 툴은 사용자요구에 의거 자가서비스 방식으로 위해 진화되고 있으며, 다수 조직들이 독자적 분석 툴과 일반적 목적의 BI플랫폼으로 ADV와 시각적 분석도구를 채택하고 있다. BI의 공격적 채택은 BI에 실시간 운영기술을 채택하도록 하고 단순 측정 및 보고 이상으로 더욱 분석적인 면이 강조되도록

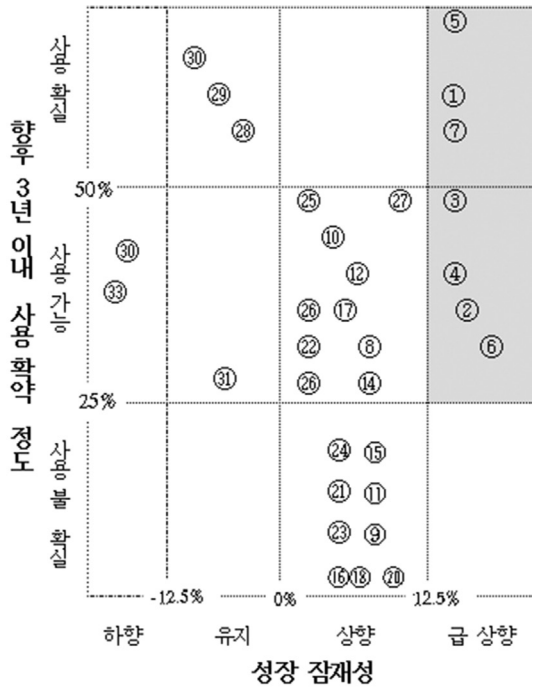

그림 5. 분석기술의 성장성과 상용화 기술

하고 있다. 스트리밍 빅 데이터에 대한 수많은 분석 애플리케이션들이 있으며 아직 실시간 분석 애플리케이션은 소수이나 실시간 처리 대상 및 선택기능이 주요 개발추세이다. 데이터베이스에서 실시간 처리방법의 하나는 서버메모리에서 관리를 통해 디스크 I/O와 속도장애요소를 제거하는 것이다. 이에 인-메모리 데이터베이스가 다양한 목적으로 사용되며, 통상 BI를 위해 실시간 대시보드를 지원하고 측정치와 KP지표, OLAP 큐브를 저장한다. 고급분석을 위한 속도증가 및 분석모델의 점수화를 위해 인-메모리 데이터베이스의 채택이 증가하고 있다. 플래시 메모리와 SSD를 채택 데이터웨어하우스를 공급하는 업체들 역시 인-메모리 데이터베이스로 이동해 가는 추세이다. 비정형 데이터의 정형화처리를 위한 텍스트마이닝 툴과 분석의 사용이 증대되고 있으며, 결과데이터는 고객 판별분석 분야 및 보험과 같은 다양한 분야에 적용 되고 있다.

나. 두 번째 기술 군 동향

〈그림 5〉에서 성장잠재성이 상향이고 3년 내 사용가능을 확약한 지닌 두 번째 기술군인 10개 기술은 기존시스템에 DW 기기, 분석전용 DBMS, 열 데이터 저장, 샌드박스가 추가된 다른 형태의 분석 데이터베이스 플랫폼을 보여준다. 새로운 분석프로그램으로 개조 시 결정요소는 기존 및 기획된 EDW가 OLAP 및 보고 작업의 성능저하 없이 고급분석 및 빅 데이터 처리가 가능한지 여부이다. 최근의 일부 EDW는 데이터베이스 안에서 고급분석 작업이 가능하나 강력한 서버자원을 필요로 한다. 사실 단일 EDW로 모든 데이터를 통합하는 것은 불가능[18]하여 통합 가능 데이터와 불가능 데이터를 분류 처리하는 기술이 필요하다. 다중 작업 시 성능저하에 대한 해결책이나 비용제한적인 이유로 전문 BI전문가들은 EDW외부의 분리된 플랫폼으로 빅 데이터 및 분석 작업을 분리시키길 선호한다. 중앙 EDW에서 분석 작업을 하려는 기업의 3분의 2는 데이터웨어하우스 프로그램에 고정적인 분석전용 데이터베이스를 요구하고 있다. 열 데이터 저장 방식의 데이터베이스는 열 기반 분석쿼리를 가속화시키고 있으며 MapReduce와 No-SQL 인덱싱을 기반으로 하는 데이터웨어하우스이나 분석중심의 분석플랫폼이 출시되고 있다.

다. 세 번째 기술 군 동향

〈그림 5〉에서 성장잠재성이 상향이고 3년 내 사용가능성이 불확실한 세 번째 기술군인 10개 기술은 초기상태의 신기술로 사용에 대한 확신이 없는 기술이다. 하둡의 분산파일시스템(HDFS)의 채택률은 아직 높지 않으나 데이터형태에 상관없이 사용가능한 파일시스템으로 다양한 데이터변환이 가능하여 관심을 끌고 있다. 최근 채택률이 높아지는 이벤트프로세싱은 스트리밍데이터의 통합에 이용되며 분석과 결합되어 사용되는 방향으로 진화하

고 있다. 아직 No-SQL 데이터베이스 채택률은 낮으나 애플리케이션개발자에게 인정받는 경향이 있다. BI전문가들은 데이터보안과 통합관리에 기인하여 BI, DW의 분석에 공공 클라우드 이상의 개인 클라우드를 선호한다. SaaS에 클라우드가 필수적이지는 않으나 대부분의 분석애플리케이션 및 분석데이터베이스 플랫폼 기반의 SaaS는 공공 클라우드를 전제로 하고 있다.

라. 네 번째 기술 군 동향

〈그림 5〉에서 성장잠재성이 유지 및 하향이고 3년 내 사용가능성을 확약한 네 번째 기술군인 6개 기술은 분석용 데이터마트, 중앙 EDW, OLAP 도구, 수작업 SQL, OLTP용 DBMS로 BI, DW, 분석에 가장 일반적 기술이다.OLTP투자와 OLAP기술은 거의 성숙단계에 있으며 향후 DW 및 분석 기반의 분석데이터베이스 플랫폼으로 전이가 예상되나 OLTP용 DW투자도 당분간 지속될 것으로 파악된다.

현재 빅 데이터 분석의 대응방안으로 데이터마트에서 분석 데이터 셋을 취급한다 해도 호스트 OLTP 데이터베이스와 SMP 하드웨어 구조가 구식플랫폼이면 문제의 소지가 있다. 분석 데이터 플랫폼의 개선을 위해 서는 기존 데이터마트를 포기해야 할 것이다. 분석에서 수작업 SQL의 사용은 향후 분석 애플리케이션기반의 SQL생성 도구로 대체될 것이다.

2. 기술 생태계 동향

2.1 솔루션

데이터 중심형태의 개방 혼합형 클라우드 연합으로 진화방향성을 보이며 클라우드 인프라, 플랫폼, 서비스는 공간 및 전력저감과 효율을 위해 분산 가상화에서 단위결집 가상화로 진화하고 있다. 데이터는 연결성 강화와 메타데이터기술, DB는 관계형에서 비관계형 기술, 데이터 처리속도는 인 메모리 채택, OLTP/OL AP 통합기술, 효율적 매니코어 OS와 병렬 분산처리 기술 및 솔루션으로 진화하고 있다[19]. 빅 비정형데이터의 생성환경에 대응하기 위해 구글, 아마존, 야후, 페이스북, 트위터 등 주요 인터넷 플레이어들은 종래 로그 및 트랜잭션 기반의 데이터마이닝 서비스의 양적 성장에서 벗어나 광고플랫폼 구축과 함께 신규 서비스모델에 맞는 서비스엔진과 소프트웨어 구조를 구축하고 있다. 최근 오라클, SAS, IBM, EMC, HP 등 글로벌 빅 데이터 솔루션 업체들이 협력업체들과 국내시장 진출에 적극적이자, 한국 지식경제부는 빅데이터 SW산업육성 및 기술과제 추진 및 중소 SW업체들로 구성된 빅 데이터 솔루션 포럼을 지원[19]하면서 정형성데이터의 처리 및 분석 속도 개선, 메모리기반 검색, 비정형데이터의 언어 처리, 텍스트 마이닝, 시멘틱 분석, SNS 분석, 오픈소스기반의 대용량 데이터 저장관리 솔루션, 데이터 시각화 등 공동솔루션 개발을 독려하고 있다[21]. 구글은 2011년 말 웹서비스 API를 제공하고 DW나 데이터센터의 부가적 투자 없이도 70TB정도의 비 압축데이터를 이용[22]하게 했던 빅 데이터 분석소프트웨어인 빅쿼리(BigQuery)의 제한적 서비스 이후, 2012년 5월 클라우드 시장공략을 위해 빅 데이터 분석지원 서비스와 함께 클라우드 스토리지의 유료서비스를 시작하였다. 페이스 북의 SNS를 기반으로 한 검색알고리즘과 구글의 페이지링크 기반의 검색알고리즘이 상호 강점영역의 부족한 부분이 강화하면서 빅 데이터 서비스 시장은 치열한 경쟁이 예상된다. Hadoop 및 R의 성능강화와 활용증가가 예상된다.

2.2 DB

TDWI의 조사결과[6] 빅 데이터 분석의 기술적 측면에서 가장 큰 장애요소(46%)는 기존 BI나 DW기반 개발자들의 보유기술 능력이 빅 데이터 분석에 적절하지 않은 면이 있는데, 이는 빅 데이터 분석시스템의 구조화(33%)와 사용자 맞춤형 빅 데이터 생성(22%)에 대한 기술적 대응의 어려움이 이유인 것으로 나타났다. 특히 데이터베이스 소프트웨어의 장애요소는 인-데이터베이스 분석능력 부족(32%), 빅 데이터로의 확장성문제(23%), 데이터 로딩속도(21%), 쿼리 분석속도(22%), 기존 DW의 보고 및 OLAP의 빅 데이터 관리한계(22%), 기타 다중시스템에서의 빅 데이터 수집 및 선별, 제어 및 시험능력 부족 등으로 조사되었다. 그러나 기술적 장애요소를 문제로 인식(30%)하는 것보다 기회(70%)로 인식하는 경향이 크다.

고급분석을 위한 빅 데이터 관리 및 운영 동향도 〈표 5〉와 같이 현재 기업 데이터웨어하우스(EDW)의 활용현황(64%)과 선호도(63%)가 높다. 빅 데이터의 고급분석을 위해 기업들은 원래 OLAP, 성능관리, 보고위주의 기존 EDW에 확장성과 쿼리 성능을 추가하기도 하는데 이는 공급영역이 아니라 사용자 디자인 영역이다. 분산 및 집중식 EDW나 분산 및 관계형 데이터베이스의 어디에서 빅 데이터 분석 프로그램을 시작 및 확장할지는 사용자의 기본적 선택 구조이다.

표 5. 고급분석을 위한 빅 데이터 관리장소

빅 데이터 관리장소	현황(선호도)%
기업 데이터 웨어하우스(EDW)	64(63)
데이터마트, 운영데이터저장(ODS)	38(20)
빅데이터분석용 상용 데이터베이스	28(30)
분산 파일시스템(HADOOP)	24(30)
클라우드 기반 분석플랫폼	12(30)
일반적 파일모음	17(5)
기타	7(5)

〈표 5〉에서 종래 증식되어온 데이터마트와 ODS방식이 감소하고 분석용 상용데이터베이스(28%)사용이 증가하고 있다. 실제 인터뷰결과, 빅 데이터의 3V특성에도 불구 구식의 데이터마트와 유사한 상용 분석 데이터베이스를 사용하는 것으로 조사되었지만 데이터마트와 연관성이 있다고는 할 수 없다. 현재 소수 사용자들이 클라우드 기반 분석플랫폼을 사용하지만 향후 증대될 것이며, 다양한 분석플랫폼이 일반화 될 것으로 예상된다. Hadoop은 이미 24%정도의 사용비중을 보이나 실험용도로 다운로드하는 경우도 있으며 과장되었다고 인식될 경우 실제로 사용될지는 미지이다. 기타 분석플랫폼을 EDW의 샌드박스, 데이터웨어하우스 레이어, 별도 분석데이터 웨어하우스, 데이터웨어하우스 생성 애플리케이션으로 관리하려는 소수의 동향도 파악된다. 저 바용 분산 환경의 솔루션기술을 요구는 기존의 정형데이터 및 관계형 DB와 비정형데이터의 효율적 연계분석과 NoSQL DB의 접목, 데이터 분산 및 트랜잭션 기술, 인 메모리 분석기술을 필요로 한다. 아마존, 구글, 페이스북, 이베이를 비롯한 다수 업체들은 비정형 데이터 관리에 NoSQL DB를 사용하고 있으며, Hadoop 기반 Hbase, 구글의 Bigtable, 페이스북의 cassand ra, 아마존의 Dynamo DB 등을 주로 사용하고 있다. Oracle은 RAC(Real Application Clus ter)분산 그리드를 기반으로 분산 데이터들에 대한 통합적 접근성을 제공하며, IBM은 DB2pur eScale을 통해 분산 환경에서의 OLTP를 위한 클러스터링 기술을 지원한다. SAP은 인 메모리 데이터 분석기술인 HANA(High performance ANalytics Appliance)를 통해 분산환경 성능및 가용성을 지원한다. 국내경우는 Hadoop중심적이나 Hadoop이 지원하지 않는 표준SQL을 지원하고 혼합형 아키텍처로 실시간 모니터링과 통계분석을 강화한 제품도 출시되고 있다.

2.3 인 메모리

빅 데이터의 고속처리를 위한 인 메모리 기술이 발전되고 있다. 2016년 IMC(In-memory computing)채택률이 35%, 2020년 65%에 이를 것이라고 전망한다[23]. 인 메모리를 이용한 분석기술은 메인 메모리에 데이터를 저장하고 분석하는 기술로, 멀티스레드 기술이 적용되어 병렬 처리되므로 디스크기반의 기존 데이터베이스보다 검색과 분석이 빠르다. 특히 필요정보를 메모리인덱스를 통해 빠르게 검색하여 데이터 검색시간을 감소시킬 수 있다. SAS는 빅 데이터의 분석, 시각화, 예측, 데이터 마이닝,고급 통계분석작업 등을 메모리상에서 처리해주는 인 메모리 분석기술을 고급분석 및 위험관리, 스트레스 테스트, 시각적 분석 등의 솔루션에 적용하고 있다. SAP은 데이터저장 및 압축 파티셔닝 기술에 차별성을 가진 인 메모리 기기를 채택하고 OLTP와 OLAP를 동일플랫폼에서 수행하는 제품을 개발 출시하고 있다. 인 메모리 분석 솔루션 및 장치 등은 금융, 유통분야에서의 우선적 도입ㆍ적용이 예상된다. 국내의 DB MS업체들도 인 메모리 기반 제품로드맵을 공개하고 비정형데이터를 메모리 테이블에 그대로 올리는 기술과 실시간분석기술, 고성능 DBMS를 개발하고 있다. 최근 실시간 분석능력 향상을 위해 IMC SW를 서버, 스트리지, 네트워크가 일체화 기기에 탑재하고 다중 코어프로세서를 채택하여 OLTP/OLAP를 단일플랫폼에서 처리하려는 통합플랫폼 형태가 급격히 증가하고 있다. 한편, 비정형 데이터에 대응한 스토리지 가상화, 중복성 제거, 압축 신기술 등의 기술 확산에도 여전히 외장형 스토리지 시장이 2011년 전년대비 10.6% 성장하는 추세도 분석할 필요가 있다.

2.4 플랫폼

일반적인 분석플랫폼은 분석모델을 형성하는 툴이나 DBMS, 혹은 둘 다를 지칭하는 것으로 정확한 지향점의 정의 없이 분석플랫폼을 대체하려는 측면도 있고 체계적 대응 없이 기존의 플랫폼으로 분석플랫폼을 커버하려는 측면도 있다. 아직 강력하고 설득력 있는 빅 데이터 분석플랫폼이 거의 없지만 기술개발을 구체화 할 수 있는 배경을 통해 동향예측이 가능하다. 플랫폼 성능 향상 및 확장을 필요로 하는 배경[6]은 데이터크기 감당불가(42%), 저속로드(29%), 저속반응(24%), 분석처리능력(17%), 혼합작업의 동시처리 불가(11%), 분석플랫폼에 필요한 분석모델의 부재(32%), OLAP 단일기능(28%), 비주얼분석 및 자가 서비스기능 부재(11%)가 있으며, 실시간 분석능력 부재(24%), DB의 분석지원불가(20%), 웹 서비스 한계(10%), 인 메모리 프로세싱 지원제한(7%) 등으로 이의 융합상품화 개발 가속화가 예상된다. IDC 보고서[24]에서 2011년 세계 비즈니스 분석 소프트웨어 시장의 매출액이 전년대비 14.1% 성장한 317억 달러이며 2016년까지 연평균 9.8%로 성장하여 507억 달러 규모에 달할 것으로 전망했다. 데이터 웨어하우징(DW) 플랫폼 부문이 전년대비 15.2%의 성장률로 가장 빠른 성장세를 보였다. 분석 애플리케이션 부문은 13.3%, BI 및 분석도구 부문은 13.2% 성장한 것으로 조사됐다. 한국지식경제부가 추진하고 있는 유망기술의 세부 기술개발 로드맵[25]상의 빅데이터 SW/HW관련 대응기술개발전략은 매니코아, 스토리지-메모리 장치 등 차세대 하드웨어 기반의 분산병렬 컴퓨팅 플랫폼 기술 확보로 대비한다는 요지이다. 빅 데이터 처리플랫폼 기술과 시스템을 구축하기 위한 페타 및 엑사 스케일 엔터프라이즈 시스템기술이 핵심이다. 〈표 6〉과 같이 빅 데이터 요소기술의 구성을 통해 기술개발 지원 분야와 기술 발전동향을 판단할 수 있다. 이미 2011~2016년 동안 349억 원 규모의 유전체분석용 수퍼 컴퓨팅 시스템 과제가 진행 중이다. 산업일반측면에서 분석플랫폼은 빅 데이터 전문가시스템 개

발을 위한 빅 데이터 수집 및 필터링기술, 자연어인식기반 지식 구조화기술, 기계학습 기반 실시간 추론기술과 지능형 마이닝을 위한 미래예측기술, 생활로그 마이닝, 소셜관계 추적 및 분석기술의 통합을 요구하고 있다.

표 6. 빅 데이터 요소기술

목표기술	요소 기술
빅 데이터 처리플랫폼	매니코어 스케쥴링, 차세대메모리 OS, NW 트래픽처리, 병렬파일시스템, 클라우드 분산파일시스템, 메모리기반 분산데이터관리, 분산병렬OLAP, 데이터 분산처리 기술

3. 특허동향

빅 데이터 분석기술 관련 특허 조사기간은 10년(2002년~2012년 7월 기준)으로 하며 조사대상은 PCT(세계특허)를 기준으로 한다. 빅 데이터 분석기술 특허조사는 소셜 네트워크 기반 4개 관련기술과 기존 기술의 발전성과 사용가능성이 가장 높은 최상위 기술을 구성하는 7개의 핵심 요소기술을 키워드로 한다. 조사결과 표현은 최근 3개년 출원추이, 주요 출원인, 주요 출원기술을 대상으로 한다. 빅 데이터분석의 주요대상이 되고 있는 소셜 네트워크 기반 분석기술은 여론 탐색 및 검토, 브랜드모니터링, 소문모니터링, 온라인인류학, 시장영향분석, 대화분석, 온라인 고객지능 등과 같이 다양하게 사용되어 왔으며 비표준화 상태이다. 이들의 포괄적 의미는 주로 소셜 미디어 모니터링 및 분석의 맥락에서 텍스트기반 주관, 오피니언 및 감성에 대한 컴퓨팅처리를 의미하는 오피니언 마이닝 및 판별분석[26]으로 주로 사용되고 있다.

3.1 소셜 네트워크 기반 분석기술 특허

가. Text Mining

최근 5년간 연도별 출원추이는 증가세에 있으며 주요출원인은 〈표 7〉과 같이 NEC와 국내의 KISTI가주요 플레이어다. 주요 출원기술 분야는 정보검색을 위한 데이터베이스 구조분야의 38건(75%)이다.

표 7. 주요 출원인별 출원건수

주요 출원인	출원건수(전체 51건)
NEC COR.	9
NEC Cor.	6
KISTI(한국과학기술정보연구원)	4
IBM COR.	3
BOEING Com.	2

나. Opinion Mining

텍스트 마이닝의 한분야로 표현되기도 하는 오피니언 마이닝은 'opinion mining과 analyt ic'의 키워드로 검색되지 않아 기존의 관련분석 개념인 'log ,reputation, sentiment, semantic, opinion analysis'의 키워드로 조사했고 결과〈표 8〉에서 뚜렷한 경향이 발견되지 않으나 로그분석이 증가추세에 있다. 이밖에 대화분석, 브랜드모니터링, 시장영향분석 관련 기술출원이 소수지만 계속되고 있음은 비 정형성데이터의 분석이 기대만큼의 활성화속도는 아니라고 본다. 학문적, 마케팅적 이슈에도 불구 특허 출원속도가 완만해진 이유가 신규 알고리즘 개발, 기 개발 알고리즘 및 기술의 상용화 개선 및 개발, 빅 데이터 분석에 대한 기존 소셜미디어 분석 중심의 기술의 패러다임 변경에 기인하는지는 상세 조사할 필요가 있다.

표 8. Opinion Mining 관련기술 출원건수

유사분석기술	중점기술 (10년간)	최근추세 (10년간)	주요 출원인
Log analysis	통계 (47)	증가 (139)	Fujitsu, MS, IBM, Recourse Tec, Go ldman Sachs
Reputation analysis	데이터처리(3)	감소(8)	Visible Tech. We b sense, Symant ec
Sentiment analysis	연산구조 (13)	증감 소 (20)	J.D Power, IPC sy s. IBM
Semantic analysis	DB검색구조 (83)	증감 소 (121)	Alcatel, MS, BI T, Invention Machin e Co. Siemens
Opinion analysis	처리장치방법 (6)	감소 (10)	Sony, Fujitsu, Vi vecon Co. Swiss qual License
Opinion search	데이터처리(6)	소수 (8)	Buzzi, Thomson Global, 스위스재보 험, Fujitsu
Opinion review	데이터처리(3)	'08후출원무 (5)	MS, Jaxr R., First Opinion, CNET

다. Social Network Analysis

출원건수가 2008년 이후 연간 6,7건에서 2012년 현재 1건으로, 소강상태로 판단되며 주요 출원인은 〈표 9〉와 같다. 주요 출원기술 분야는 GO6F 정보검색을 위한 데이터베이스 구조분야 18건(51%)과 GO6Q 응용시스템과 방법 10건(29%)이다.

표 9. 주요 출원인별 출원건수

주요 출원인	출원수(총35건)
Yahoo	3
SMART Link Medical, Inc	3
OWAVE Media Co.	2
21st Century Tech. Inc.	2
Icosystem, HuWEi, HP, Alcatel 등	1

라. Cluster Analysis

군집분석관련 총 출원건수(184건)는 타 분석 방법보다는 상대적으로 많으며 2008년 이후 증가추세에서 2011년 감소한다. 주요출원인은 〈표 10〉과 같다. 주요 출원기술 분야는 G06F 정보검색을 위한 데이터베이스 구조분야 51건(28%)과 G01N 클래스색인과 실행방법 18건(10%)이다.

표 10. 주요 출원인별 출원건수

주요 출원인	출원건수(전체 184건)
NEC	3
PHILIPS Electronics N.V.	3
HITACHI, LTD	3
EASTMAN KODAK Com	3
BURSTEIN Tech, INC	3

3.2 성장성 최상위 기술 군 특허

성장성과 3년 내 사용가능성이 높은 최상위 기술 군〈그림 7〉의 주용 구성기술을 대상으로 조사한 결과는 〈표 11〉과 같다.

표 11. 기술 분야별 중점기술 및 추세

기술 분야	중점기술 (10년 건수)	최근추세 (10년건수)	주요출원인
Data Mining	데이터처리(189), DB 구조 (61)	4년 증가 이후 둔화 (477)	Minnesota Mining, IBM, MS, SIEMENS
Statistical Analysis	데이터처리 (429), 색인처리 (163)DB구조(116)	2009년 이후 둔화 (773)	IBM, PHLIPS, ETRI, ALCAT EL, SAMSUNG ,MS, KT, SIEM ENS, GE
Complex SQL	DB구조(6)	극히 소수 2011(1)	ZTE, MS, SAP
Data Visualization	데이터처리 (168), 이미 지처리(112)	증감 둔화 (531)	PHILIPS, SIE MENS, BI SOL., PHILIPS, MS
Artificial Intelligence	데이터처리 (5 4), 응용처 리 (27),보안(22)	증감 둔화 (234)	ACCENTURE, BOEING, MS, BASF,
Natural Language Process	데이터처리 (371)	증감 둔화 (493)	MS, PHILIPS, INVENTION M ACHINE, BT, VOICEBO X, IBM, SIEME NS
Analytic Database	데이터처리 (13)	증감둔화 (23)	NCR, INFOMA TICA, EDSA , IBM, HIT ACHI

기술성장성과 사용자 채택 예상가능성이 높은 기술의 7개 요소기술의 출원추이가 최근 4년간 대체적으로 증감이 둔화되거나 완만한 상태이다. 이는 빅 데이터 버블개념의 탈피나 추진을 위한 실무적 스위칭과정에서 재원 조달문제, 기술적 갭으로 인한 일시적 케즘(chasm), 인프라 및 분석의 기술적 완성도 증가나 포화, 공개 플랫폼이나 툴(Hadoop이나 R등)의 채택확대 및 사실상의 표준화에 의한 영향으로 판단될 수도 있어 향후 상세조사 및 분석이 필요하다. 7개 기술의 특허추이는 빅 데이터 분석기술의 진화나 천이가 완성된 신기술로서의 특이점은 없으며 기존 기술의 진화 및 천이과정 추이와 유사함도 연구대상이다.

Ⅳ. 빅 데이터 활성화 과제

1. 활성화 환경 인식

최근 거버넌스 차원에서, 빅 데이터를 활용한 스마트 정부 구현 안[3]을 기반으로 빅 데이터 서비스 활성화[27] 방안이 제시되고 있다. 국정운영과 전략 수립을 위한 빅데이터 국가전략포럼[28]과 빅 데이터 산업 경쟁력 제고와 시장 확대를 목표로 하는 빅 데이터 포럼[29]이 생성되었다. 행정안전부는 빅데이터 마스터플랜 수립을 위한 태스크포스(TF)를 가동하고 2012년 9월 각 부처 데이터 보유현황과 활용 수요 조사를 완료하고 빅데이터 마스터플랜[30]을 진행 중이다. 한국과학기술정보연구원(KISTI)은 국가적 과학기술패러다임에서 빅 데이터의 수집과 공유를 위해 과학기술용 빅데이터 플랫폼 개발과 고성능 슈퍼컴퓨팅체계 구축(2013년~ 2021년)을 추진중이다. 대기업도 빅데이터 추진 및 활성화를 위한 전사적 인프라와 플랫폼 구축을 추진 중이다.

빅데이터 활성화는 빅데이터 추진환경의 세밀한 인식과 분석을 기반으로 해야 한다. 신 성장기술의 등장시마다 정부의 활성화 및 지원확대 정책에 익숙한 중소기업들은 자신들의 직접수혜가 없으면 정책 및 활성화 전개가 달라진 것이 없다고 불평하는 경우가 있다. 거버넌스 차원의 빅 데이터 처리 및 분석의 편향적 추진이나 비즈니스모델의 부재와 수익성으로 연계되지 않는 상황에서 초기의 소모적 기술개발 투자는 기술버블의 위험을 알면서 기꺼이 감수하려는 기술 테크노크라트(technocrat)만 양성할 수도 있다는 우려도 있다. 수익성 고려 없이 기술적 측면만 부각된 분석기법의 적용과 개방적 분산처리를 위한 인프라와 플랫폼 형성이 기술적 패션의 일부로 전락할 수 있다는 우려도 있으나, 기술이나 정책 패션도 발전의 일부다. 빅 데이터 활용을 위한 다양한 정책의 방향성과 발전과제가 발표되고 있는 현재, 중소기업의 관심은 빅 데이터 생태계의 선행수요 유발을 위한 개발의 정책지원 확대와 시범과제의 참여이다. 정부

가 소수의 국내 빅 데이터의 구축 IT 대기업을 파트너로 할 수 밖에 없으나, 대기업과의 연관관계나 확정수요처가 미흡한 중소기업군을 대상으로 적극적 클러스터 구성과 프로젝트 추진을 지원해야 한다. 과제나 프로젝트 추진시 기술패션이나 구축의욕 과다에 의한 가시적 의사결정의 위험이나 기회손실이 적어지도록 계획단계부터 Fail-safe 개념의 비즈니스와 ROI 가치를 추구하도록 유도할 시기이다.

2. 활성화를 위한 과제

빅 데이터 활성화를 위한 정책적, 기술적 과제가 있다. 기술적 과제의 성공은 전방위적 정책과제의 실천과 성공을 병행기반으로 한다. 정책과제의 방향성으로, ①빅 데이터 관련 소프트웨어 개발을 위한 중소기업들의 기술 습득과 적용경험에 의한 자체 솔루션 능력을 배양토록 인큐베이터 역할과 지원범위 확대, ②빅 데이터 분석을 활성화하기 위한 데이터 분석가와 데이터 과학자의 양성과 함께 SW레퍼런스와 응용분야별 템플릿, 활용가이드의 개발 보급, ③국내 SW 사업자들이 빅 데이터 분석 공개SW 기술기반의 생태계 형성을 위한 연구개발과제와 지원자원의 확대, ④빅 데이터 활용정책이나 가치추구에 앞서 기존 데이터 활용의 득과 실에 위한 기술적 로드맵의 개선과 시나리오별 실천계획 수정, 성과평가지표 개선 및 평가 실적관리의 신뢰성 강화, ⑤중소기업들의 데이터 취득 및 관리 채널형성과 빅 데이터 관련 규제 및 법령 개정의 검토, ⑥개인정보보호법에 의한 정보수집 이용에 대한 보완책 및 프라이버시 문제 해결과 데이터의 거래활성화 방안 등의 전개가 필요하다.

기술과제의 방향성으로, ①데이터 자체보호기술, 디바이스 인증기술, 클라우드 서비스 인증기술, 스마트 IT용 블록암호 설계기술, 빅 데이터 관리 암호기술의 연구 개발, ②빅 데이터를 활용한 지능적 미래 예측과 자동화 대응을 위한 BI와 고급 분석 SW 및 모델, 저가형 보급형 모델의 연구 개발, ③오픈소스 분석도구를 활용한 저가의 분석엔진으로 분석 라이브러리의 생성과 지원, 대응 및 예측 분석 툴의 결합이 쉬운 소프트웨어 플랫폼의 연구 개발, ④데이터 분석에 대한 템플릿 및 알고리즘, 모델의 적합도 검증 가이드, 가치정보의 실시간 검증 및 보정 체계의 연구 개발, ⑤이종 및 불일치 데이터의 통합, 비 상관관계 데이터의 여과, 비 정형데이터의 목표 지향적 변환, 분석구조, 저장, 고장방지, 자가 관리, 확장성, RDB MS와 NoSQL의 결합, Query 최적화, 스트리밍 처리기술 등의 심층 연구 개발 등의 전개가 필요하다.

V. 결 론

빅 데이터의 수집과 축적, 관리 기술보다 중요한 것은 어떤 목적성과 분석방법으로 의미와 가치가 있는 데이터를 해석, 추출해 낼 것인가에 대한 통찰력과 의사결정능력이 더욱 중요하다. 비전문가인 최종사용자의 빅 데이터 분석능력 고급화요구에 따른 맞춤형 개발 수요가 기존 공급기술의 능력을 능가하고 전문가가 배제될 수도 있는 상황에 대응하여, 분석 툴 개발 시 분석의 오남용과 데이터누락에 의한 의사결정오류 방지 및 적정수준으로 분석이 자동 보정되도록 하는 통찰능력까지도 고려하자는 빅 플레이어들의 개발동향도 있음을 주시하고 국내 기술개발 범위를 확장해야 할 것이다. 이와 같은 빅 데이터의 추세와 기술동향에 적극적 대응이 중요하지만 해외 빅 플레이어들의 특화 서비스 및 솔루션 시장의 생성과 마케팅논리에 말려들 필요는 없다고 본다. 특화기술의 벤치마킹은 필요하다. 빅 데이터 활용기반 구축 시 본고에서 제시된 기술동향을 고찰하고 빅 데이터 분석체계 구축 전 기존 정보시스템에서의 분석과 BI의 구현수준, 클라우드 컴퓨팅 및 SNS에 의한 비정형데이터기반의 분석과 BI 구현수준, 기존 및 미래형 정보시스템의 수익성과 가치에 대한 정확한 진단과 예측이 필요하다. 보급형 빅 데이터 고급분석기술 및 BI 알고리즘 개발, 빅 데이터 활용기반 구축과 활용성 증대를 위한 과제의 실천이 시급하다. 정부의 활성화 투자는 동일과제의 중복성과 병렬경쟁을 철저히 배제해야 할 것이다.

Acknowledgement

본 연구는 한국과학기술정보연구원 ReSEAT Program의 기술동향 연구과제로 수행되었습니다.

참 고 문 헌

[1] Dan wood,"Tableau Software's PatHanraha n on 'What Is a Data Scientist?'" Forbes Data Review, 11.30.2011 (http://www.forb es.com/sites/danwoods/2011/11/30/tableau -softwares-pat-hanrahan-on-what-is-a-dat a-scientist/)

[2] 김승윤, " Big Data 최근 글로벌 동향과 이슈 ",KT경제경영연구소 ISSUE&TREND,2012.7

[3] 이각범, "빅 데이터를 활용한 스마트 정부 구현 (안)", 국가정보화전략위원회 위원장 보고사항 146호, 3, 3-4p

2011.10.26
[4] IDC," Worldwide Big Data Technology and Service 2012-2015 Forecast", IDC 2012
[5] Apache Hadoop,"What is Apache had oop?", http://hadoop.apache.org/
[6] Philip Russom," Big Data Analytic s", TDWI Best Practices report, TDWI Reserarch, 4th Quarter 2011
[7] Galen Gruman, "Tapping into the power of Big Data", PWC Technology Forecast A Quarterly Journal 2010. issue 3,(http: //www.pwc.com/us/en/technology-forecast/assets/PwC-Tech-Forecast-Issue3-2010.pdf)
[8] 백인수, 빅데이터시대:에코시스템을 둘러싼 시장 경쟁과 전략분석, IT & Future Strategy, NIA 제 4호, 2012,4,26
[9] http://tdwi.org/Blogs/Philip-Russom/2011 /06/Three-Vs-of-Big-Data-Analytics-3-Da ta-Velocity.aspx
[10] IDC, Digital Universe Study 2010.5
[11] Ranjan, Rajiv, Harwood. Aaron, Buy ya. Rakumal, "Coordinated load manage ment in Peer-to-Peer coupled federated grid systems", Journal of Supercom putting, Aug. 2012, Vol. 61 Issue 2, p.292-316,
[12] 이명진, "빅 데이터 환경의 고급분석 기법과 지 원 기술동향, 연세대학교지식정보화연구소, 2012(htt p://wwwslidesharenet/onlyjiny/ advanced-analytics-and-technologies-fo r-big-data)
[13] 안창원. 황승구,"빅데이터 기술과 주요이 슈",한국전자통신연구원 특집원고, 정보과학회 지 2012.6
[14] 황승구, " Big Data 기술현황 및 전망", 발표 자료, 한국전자통신연구원, 20120416,
[15] http://hadoop.apache.org
[16] KB, "빅데이터(Big Data)의 이해와 금융업에 대한 시사점",KB금융지주경영연구소, KB daily 12-87호, 2012.6
[17] Danyel Fisher, Rob Deline, Mary Czerwin ski, steven Drucker, "Interactions with Big Data Analytics",Interactions May+Ju ne 2012, pp. 50-59
(18)Colin White,"MapReduce와 데이터 사이언티스 트", BI research, TERADATA(한국), 2012.1, 18p.
[19] McKinsey Global Institute, "Big Data :The next frontier for innovation, compet ition and Productivity", McKinsey and Company, May 2011.
[20] 지경부 소프트웨어 진흥팀,"빅데이터 SW산업 육성 전략회의",13th SW Quality Insight 컨 퍼런스 보도자료, 2012.7.11
[21] 중소소프트웨어 업체포럼, "빅데이터 솔루션 공동 구축 및 시장대응을 위한 포럼발족", 아이 뉴스, 2012.5.22
[22] 이지영,"구글빅데이터분석서비스공개"(htt p://www.bloternet/archives/83805,2011111 5)
[23] Massimo Pezzini, "The Next Gener ation Architecture: In-Memory Computing", Gart ner by SAP Innovation forum in Netherlan s, 2012.3.8
[24] IDC,"Worldwide Business Analytics Software 2012(2016 Forecast and 2011 Vendor Shares Report)
[25] KIAT 미래기술기획팀, "2011 산업기술로드맵: 정보통신 SW 최종보고서", 한국산업기술진흥원, 지식경제부, 2012.3
[26] Bo Pang and Lillian lee,"Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1-35(pre-version)
[27] 보도자료," 빅데이터 서비스 활성화방안", 방송통신위원회 네트워크정책국 스마트네트워크 서비스팀, 2012.6.21
[28] 한국정보화진흥원 빅데이터 전략지원센터(www.bigdataforum.or.kr),"빅데이터 국가전략포 럼 발족", 2012.4.27
[29] 한국정보통신진흥협회(KAIT), "2012.8.16 빅 데이터포럼 창립 발표", 2012.7.19
[30] 맹형규, "빅데이터 마스터플랜 추진현황 및 향 후계획", 행정안전부 보고 221호, 2012.7.6

약 력

박 종 만

1978년 인하대학교 공학사
1983년 연세대학교 경영학석사
1987년 미국 Lehigh 대학교 공학석사(박사수학)
1997년 인하대학교 공학박사
1977년~1983년 ADD/KIDA 연구원
1988년~1998년 대우그룹 회사원
1999년~현재 유한대학교 초빙 및 겸임교수
2008년~현재 한국과학기술정보연구원 ReSEAT(IT부문) 전문연구위원
관심분야: 빅 데이터분석, 의사결정시스템, RFID/USN, 정보통신 응용

엄 태 원

1978년 인하대학교 공학사
1981년 인하대학교 공학석사
1994년 인하대학교 공학박사
1981년~현재 유한대학교 산업경영공학과 교수
관심분야: IT서비스기술, U-Service

김 하 진

1962년 서울대학교 수학과 이학사
1977년 불란서 Grenoble1대학교 이학석사
1980년 불란서 Saint-Etienne대학교 이학박사
1974년~2004년 아주대학교 정보통신대학 컴퓨터공학과 교수
2004년~현재 아주대학교 정보통신대학 명예교수
2006년~현재 한국과학기술정보원 ReSEAT 프로그램 전문연구위원
2004년~2015년 ISO/IEC JTC 1/SC 24 chairman
관심분야: 컴퓨터그래픽스, 영상표현기술 국제표준화