빅 데이터의 핵심 플랫폼, 기업용 하둡 동향

김동한
펜타시스템테크놀러지㈜ 고등기술연구소 소장
picollo@penta.co.kr

1. 서론

빅 데이터(Big Data)는 Gartner, IDC, IBM, EMC 등 주요 리서치 기관, IT 벤더들이 2012 년 전략 기술이나 주요 이슈로 꼽은 IT 트렌드 중 하나였다. 작년 한해 IT 시장을 가장 달군 키워드를 꼽으면 단연 빅 데이터일 것이며 2013 년부터는 본격 성장가도로 진입할 것으로 예상된다(<표 1> 참조). 빅 데이터가 언급될 때 자연스럽게 빠지지 않고 같이 등장하는 기술이 하둡(Hadoop)이다. 하둡은 클라우드 컴퓨팅과 빅 데이터가 IT 업계의 핵심 화두로 떠오르면서 가장 '핫'한 오픈 소스 기술 중 하나로 자리매김하고 있다. 2000 년대 중반 Yahoo, Google, Amazon, Facebook 등의 얼리 어답터이자 기타 웹 2.0 개척자들이 시작한 하둡은 이제 대기업, 서비스 공급업체 및 기타 조직의 빅 데이터 전략의 중심으로 자리잡고 있다.

이에 본 고에서 현재 통용되는 있는 빅 데이터의 정의 및 관련 시장 동향, 빅 데이터 구현을 위한 핵심 데이터 처리 플랫폼으로 떠오른 하둡의 개념 및 주목 받게 된 배경을 살펴보고, 실제 일부 기업 환경에 적용되어 사용되고 있거나 기업들이 관심을 가지고 도입을 고려하고 있는 대표적인 하둡 상용 버전들에 대해 살펴보고자 한다. 마지막으로는 하둡 관련 이슈 사항과 향후 전망을 정리하는 것으로 마무리 하고자 한다.

---

* 본 내용과 관련된 사항은 펜타시스템테크놀러지㈜ 고등기술연구소 김동한 소장 (☎ 02-769-9773)에게 문의하시기 바랍니다.

** 본 내용은 필자의 주관적인 의견이며 NIPA 의 공식적인 입장이 아님을 밝힙니다.

<표 1> 2013년 IT 핵심 이슈와 기술 트렌드

NIPA (10 대 이슈)	한국IDC (10 대 예측)	Gartner (10 대 전략 기술)	삼성 SDS (IT Mega Trend)	ETRI (IT 미래 10 대 기술)
빅 데이터 활용	불확실성 확대로 국내 IT 시장 성장세 둔화	모바일 기기 전쟁	빅 데이터를 통한 가치창출	고해상도 홀로그램 디스플레이
특허·지재권 중요도 증대	제3의 플랫폼, 새로운 성장과 변화의 동력	모바일 애플리케이션과 HTML5	클라우드 서비스의 발전	뇌파인지 기반의 인터페이스
클라우드 컴퓨팅 도입 확산	컨슈머라이제이션의 전방위적 확산	퍼스널 클라우드	통합형 IT 비즈니스	인쇄 가능한 태양전지
신정부의 IT 정책방향	스마트 커넥티드 디바이스를 통한 멀티 디바이스 시대 도래	기업용 앱스토어	지능화된 보안 위협	저전력 서버
차세대 반도체·디스플레이	모바일 네트워크 환경 진화 가속	IoT (Internet of Things)	공격적 특허전략	건강 및 복지용 상황인지 로봇
신종 보안 위협	빅 데이터 솔루션 수요 확대	하이브리드 IT·클라우드 컴퓨팅	상황 인지형 기기와 서비스	맞춤의학용 개인 유전체 분석
스마트 홈 가전·서비스	데이터센터의 변모, 기업 경쟁력의 핵심	전략적 빅 데이터	차량의 스마트 기기화	빅 데이터 분석
HTML5	소셜 네트워크 관련 기술, IT 영역 전반으로 확산	실행 가능한 분석	Green IT 의 진보	초고용량 인 메모리 컴퓨팅
소셜 미디어·소셜 엔터프라이즈	영역과 위협의 복잡성 심화에 따른 새로운 보안 의식 대두	인 메모리 컴퓨팅	개방형 생태계를 통한 기업의 급성장	클라우드 컴퓨팅
차별화를 위한 콘텐츠·서비스 경쟁	새로운 환경, 새로운 가치 중심의 IT 마켓플레이스 등장	통합 에코 시스템		감성 교류기반 스마트 러닝

<자료>: [1], [2], [3] 참고 재구성

2. 빅 데이터 트렌드

가. 빅 데이터 정의

빅 데이터의 등장 초기에는 단어의 의미에 충실한 단순히 데이터의 양이 많은 것에 주안점을 둔 정의였었다면, 최근에는 데이터베이스나 아키텍처 등 기존 방법이나 도구로 수집, 저장, 관리, 분석할 수 있는 범위를 초과하는 거대한 규모의 데이터 집합(정형, 비정형), 이를 분석하는 기법, 조직까지 포괄하는 광범위한 의미로 받아들여지고 있다.

클라우드 컴퓨팅의 붐 초창기에 클라우드 컴퓨팅에 대한 다양한 관점의 정의와 시각차가 있었던 것처럼 현재 빅 데이터의 정의와 범위에 대해서도 다양한 시각차가 존재하지만 시간이 지나면서 명확하고 구체화된 분류 기준과 기술, 그리고 서비스 유형에 대해 정리가 이

<표 2> 빅 데이터의 정의

기관	빅 데이터 정의	시사점
McKinsey (2011)	일반적인 데이터베이스 소프트웨어가 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터	데이터 규모에 초점(정량적 측면 강조)
IDC (2011)	다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고 데이터의 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처	데이터 규모가 아닌 업무 수행에 초점, 특징으로 3V(Variety, Velocity, Volume) 또는 4V (3V+ Value)
Wikipedia	기존 데이터베이스 관리도구의 데이터 수집·저장·관리·분석의 역량을 넘어서는 대량의 정형 또는 비정형 데이터 세트 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술	데이터 규모 및 업무 수행의 관점에서 통합된 정의
Gartner	기존 3V(Velocity, Volume, Variety)에 복잡성(Complexity) 추가	구조화되지 않은 데이터, 데이터 저장방식 차이, 중복성 문제 등 데이터 관리 및 처리 복잡성 심화로 복잡성을 특성으로 추가
IBM	기존 3V(Velocity, Volume, Variety)에 데이터의 진실성(Veracity) 추가	진실성이 확보된 데이터를 바탕으로 분석해야 한다는 데이터 품질 측면의 특성 추가
최근 동향	수많은 정형 데이터 혹은 비정형 데이터를 수집하면, 분명한 패턴이 나오게 되며, 이를 통해 수집된 데이터를 기반으로 한 예측 분석	매출 증가, 비용 절감, 고객만족 증대라는 비즈니스 가치를 창출할 수 있는 패턴 발견에 집중

루어 질 것으로 보인다.

<표 2>와 같이 빅 데이터를 바라보는 관점과 해석은 다양하지만, 빅 데이터의 일반적인 개념 요소로는 정보의 집적, 정보의 결합, 정보의 분석이라 요약하여 정의할 수 있다. 여기서 정보의 집적이란 데이터의 양을 고도화한다는 의미이고, 정보의 결합이란 다양한 목적 또는 형태의 데이터를 연결시키는 것을 의미하며, 정보의 분석이란 거대하고 다양한 데이터를 연결하여 원래 데이터 이상의 효용이나 가치를 창출해 내는 것을 의미한다.

나. 빅 데이터 시장 동향

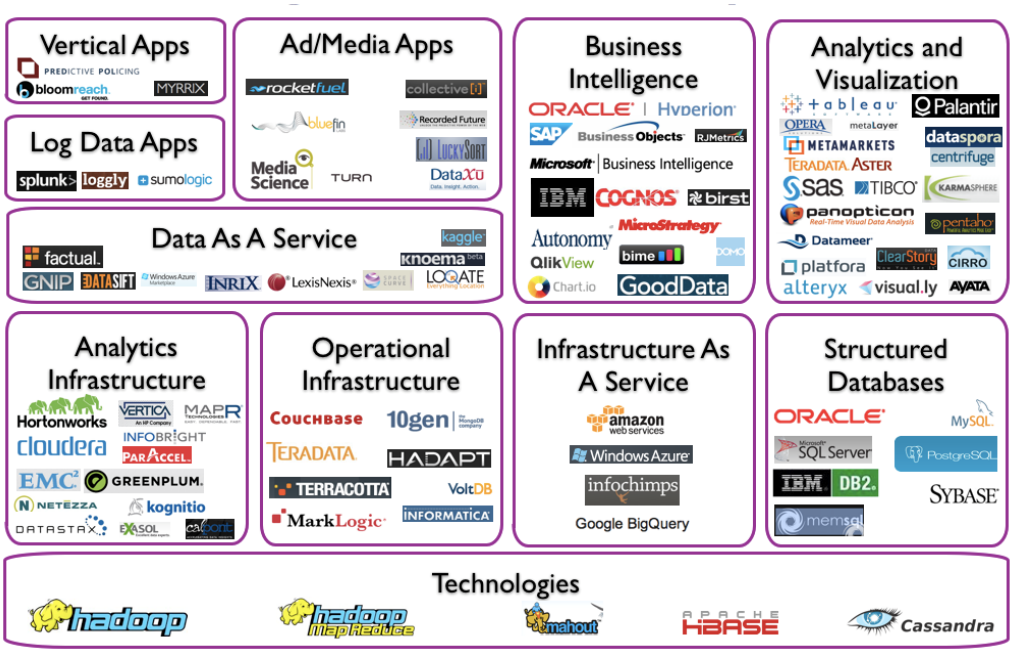
빅 데이터의 등장 초기 '단순한 마케팅 용어'라는 비난을 받기도 했지만, 이제는 IT 의 성장 동력 중 하나로 확실히 자리매김하는 모습이다. 빅 데이터 시장과 관련하여 시장 조사 기관들의 전망을 살펴 보면 다음과 같다.

- 가트너는 빅 데이터 이슈에 따라 형성된 전 세계 IT 시장(지출)이 2012 년 280 억 달러(약 31 조 원)에 이를 것으로 추정하였다. 2013 년에는 340 억 달러(약 37 조 6,000 억 원)로 60 억 달러가 늘어날 것으로 전망하였다[4].
- IDC 는 전 세계 빅 데이터 시장은 2010 년 32 억 달러에서 연평균 39.4% 성장, 2015 년에 169 억 달러 규모에 달할 것으로 전망하였으며, 이는 정보통신기술 전체 시장

성장률의 약 7 배에 이를 것이라는 예측이다. 아시아태평양지역(일본제외)의 빅 데이터 기술과 서비스 시장은 향후 5 년간 연평균 46.8%의 고 성장세를 나타낼 것이라는 전망을 내놓았다[5].

- 위키본(Wikibon)에서는 2011 년 빅 데이터 시장 규모를 52 억 달러로 추산하였다. 분야별로는 하드웨어(HW), 소프트웨어(SW), 서비스 시장 중 서비스 매출이 44%로서 가장 큰 비중을 차지한다고 발표하였다. 기업들의 빅 데이터에 대한 투자가 증가하면서 2013 년에는 이 시장이 102 억 달러, 2017 년에는 530 억 달러 규모로 성장할 것이라고 전망하였다[6].
- 정보통신산업진흥원(NIPA)은 IT 업계 종사자를 대상으로 실시한 2013 년 10 대 이슈 설문 조사에서 IT 업계의 핫 이슈로 떠오른 '빅 데이터'가 1 위로 선정되었다고 밝혔다[4].

이처럼 국내외의 관심을 한 몸에 받고 있는 빅 데이터는 미국의 경우 2013 년 이후에 실험 단계에서 실제 비즈니스에 적용되는 단계로 나아갈 것으로 예상되고, 국내에서도 활발하게 도입 및 적용을 시도하려는 사례가 늘어날 것으로 예상된다.



(그림 1) 빅데이터 업계 지도(Big Data Landscape)[7]

<표 3> 국내 빅 데이터 시장 플레이어([8] 참고 재구성)

구분	시장 접근	대표 기업	동향
기존 데이터 분석 시장을 중심으로 한 솔루션 벤더	기존 시장을 보호하기 위해 포장, 데이터 분석 쪽에 초점을 맞춤	IBM, EMC, SAP, Oracle, 테라데이터, SAS, HP 등	BI, 데이터 웨어하우스(DW), 데이터베이스 벤더들이 기존의 강점과 전문성을 빅 데이터 영역으로 확장, 새로운 서비스와 솔루션 개발
오픈 소스를 중심으로 한 기술 중심 업체	오픈 소스 기술을 중심으로 데이터 플랫폼 구현, 빅 데이터 수집/저장/분석/표현의 전체 과정을 통합적으로 처리할 수 있는 빅 데이터 플랫폼 구축 및 토털 솔루션 서비스 제공	KT Cloudware(구 NexR), 그루터, 클라우다인 등	국내 빅 데이터 관련 실적의 대부분을 수행, 대외적인 움직임은 두드러지지 않으나 주요 산업군에 POC 및 레퍼런스 확보에 주력 중
IT 서비스 업체	기존에 소위 '솔루션'으로 구현됐던 시스템을 하둡 기반으로 마이그레이션, 특정 서비스를 구현하기 위해 기술 개발	삼성 SDS(하둡 활용, 바이오 인포메틱스 플랫폼 구축), LG CNS('스마트 빅 데이터 플랫폼(SBP)': 하둡 표준 배포판 '빅팩(BigPack)' 포함), SK C&C(비즈니스 분석 솔루션 사업) 등	업체 특성상 기술 내재화보다 외주 용역을 중심으로 움직임. 사업성을 중심으로 움직일 것으로 예상

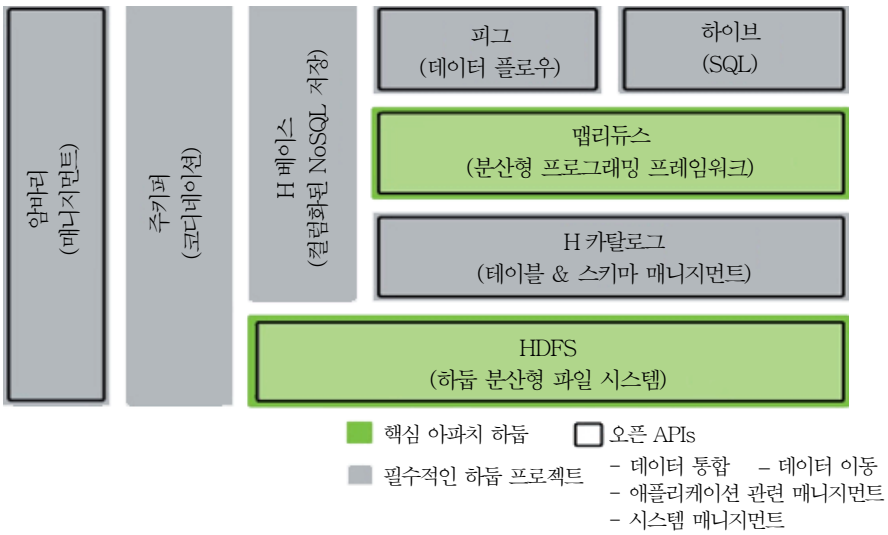
빅 데이터 시장을 선점하기 위해 관련 기업들은 다양한 빅 데이터 솔루션과 서비스를 쏟아내고 있으며((그림 1) 참조), 현재 국내 빅 데이터 시장을 주도하는 기업들을 구분해 보면 <표 3>과 같이 정리해 볼 수 있다. 여기서 주목할 만한 사실은 빅 데이터 솔루션 업체들이 어떤 방식으로든 자신들의 빅 데이터 솔루션에 데이터 저장 및 처리를 위해 하둡을 내장하거나 연동하는 방식으로 채택, 적용하고 있다는 점이다.

3. 기업용 하둡 동향

가. 하둡의 이해

앞의 빅 데이터 정의에서 언급한 빅 데이터 구성 요소 중 정보의 집적 부분에 대응되는 기술이 분석 인프라 기술이며, 이는 분석과 표현을 수행할 수 있도록 해주는 기반 기술과 플랫폼에 해당된다고 할 수 있다.

하둡은 빅 데이터 분석 인프라에 속하는 기술로서 대용량의 정형(구조적) 데이터, 반-구조적 데이터, 비정형(비구조적) 데이터를 분석하고 저장하는 데 많이 활용되어 빅 데이터라는 용어가 등장하자마자 그 옆자리를 꿰차면서 현재 시장에서 빅 데이터를 처리하는 사실상(de facto) 표준 플랫폼으로 자리매김하였다. 아마존 웹서비스(AWS)를 비롯하여 페이스북과 구글 같은 기업들은 하둡을 도입해 실시간 데이터 처리와 고객 분석을 통한 서비스



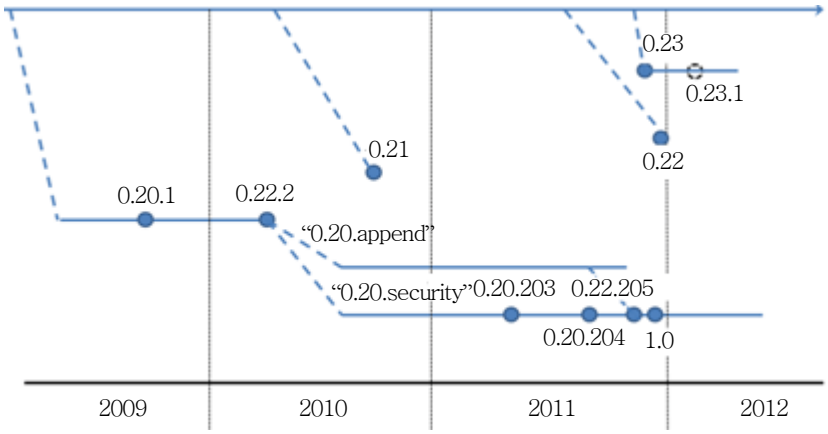
(그림 2) 하둡 프로젝트 구성[9]

연구에 한창이다. IBM, EMC, Oracle, SAP 같은 업체들도 빅 데이터 어플라이언스를 선보이며 하둡 끌어안기에 나섰다. 하둡은 그만큼 성능과 안정성이 검증된 기술이다.

하둡은 몇 가지 구성요소들로 이루어져 있으며((그림 2) 참조), 그 가운데 핵심 요소(부 프로젝트)는 맵리듀스(MapReduce) 데이터처리 프레임워크와 하둡 분산형 파일시스템(Hadoop Distributed File System: HDFS)이다. HDFS 란 하둡 네트워크에 연결된 아무 기기에나 데이터를 밀어 넣는 분산형 파일시스템으로 애플리케이션 데이터에 대한 높은 처리량의 액세스를 제공하고, 맵리듀스는 컴퓨팅 클러스터 상의 대용량 데이터에 대한 분산 처리를 위한 소프트웨어 프레임워크이다. 한마디로 말하자면 하둡은 대용량 데이터 처리 분석을 위한 대규모 분산 컴퓨팅 지원 프레임워크로 하둡의 가장 큰 특성은 바로 분산이다. 즉 분산 처리, 분산 저장이라는 하둡의 핵심 특성을 기반으로 여러 개의 컴퓨터를 마치 하나인 것처럼 묶어 저장 공간과 계산 능력을 늘릴 수 있다.

하둡은 기업들이 이전에는 비용과 복잡성 그리고 도구가 없어 폐기했던 데이터(예; 로그 파일)를 저장하고 처리할 수 있게 해준다. 모든 데이터의 가치가 동등하고, 각각이 비슷한 수준으로 이용될 수 있기 때문에, 비즈니스 시나리오들은 언제라도 아무런 제약 없이 원 데이터를 가지고 추진될 수 있다.

하둡은 확장성이 있을 뿐만 아니라 유연하기도 하다. 하둡에는 스키마가 없어(Schema-



(그림 3) Hadoop 연혁[11]

less) 사용자가 더 복잡한 분석을 하기 위해 서로 전혀 다른 여러 소스로부터 데이터를 추가하거나 모을 수 있다. 필요에 따라 새로운 노드를 추가할 수 있으며, 하둡의 내장된 장애 대처 기능은 해당 노드에 장애가 발생하면 시스템이 작업을 다른 지역으로 전송할 수 있게 해준다.

이렇게 주목 받는 하둡에 관한 놀라운 사실 하나는 현재 버전이 1.0 라는 것이다((그림 3) 참조). 아파치소프트웨어재단(ASF)은 하둡 1.0 버전에 커베로스(Kerberos)를 통한 강화된 인증, 웹 HDFS, H 베이스 트랜잭션 로깅과 로컬 파일 접근 시 성능 향상 등 빅 데이터를 다루기 위한 클라우드 컴퓨팅 기술을 모두 담았다고 밝히고 있다.

하둡 정식 버전 발표에 대한 의미를 미국 컨설팅업체 레드몽크의 공동설립자 제임스 가버너 애널리스트는 "웹에서 태어난 하둡이 정식 공개되면서 엔터프라이즈 기술로 변신"했다며, "데이터 제공자와 기술 사용자들에게 비용 효율적인 오픈 소스 클라우드 컴퓨팅 플랫폼을 조직 내 인프라에 적용시킬 수 있게 해줄 것"이라고 평했다[10].

나. 하둡 상용 버전 동향

정식 버전이 출시되면서 엔터프라이즈 플랫폼으로서 하둡의 부상은 여러 가지 면에서 리눅스의 등장과 흡사하다. 폭 넓은 규모로 채택하기에 앞서 소프트웨어의 이점을 시험하기 위해 그늘에 가려진 IT 프로젝트나 비밀 실험용으로 배포가 진행되었고, 하둡은 빅 데이터를 저렴한 방식으로 대응할 수 있는 오픈 소스 분석처리 기술로 알려지게 되었다. 오픈 소스 기술 특성상 업계의 관심이 높을수록 자발적으로 지원하려는 개발자들로 커뮤니티

가 발달하고 빠른 발전을 이룰 수 있다는 장점이 있는 반면, 이것이 곧 기업에서 원하는 안정적이고 편리한 제품의 형태로 쉽게 구현으로 연결되는 것은 아니라는 것이다. 바로 이 부분을 수익 창출의 기회로 간파한 신생 벤처 기업들이 기업용 하둡 제품으로 발전시켜 하둡을 낯설어하는 기업 환경에 도입하고 적용할 수 있도록 기술과 서비스를 공급하고 있다.

이들은 주로 오픈 소스로 하둡 기반의 빅 데이터 처리 플랫폼을 구현하여 제공하며, 이를 다루기 위한 기술뿐 아니라 현업에 적용하기 위한 컨설팅, 실무자를 위한 교육, 기술적 문제에 대응하기 위한 대응 서비스를 프로그램으로 구성하여 사업을 운영한다. 이러한 엔터프라이즈 지향적인 하둡 상용 버전(상용 배포판, 기업용 하둡이라고도 함) –기술 지원, 관리 도구 그리고 구성 지원 포함– 공급업체의 부상이 기업 영역에서의 채택을 더욱 가속화시키고 있다. 이들 업체들은 모두 하둡 데이터 플랫폼 운영 노하우를 근간으로 활동한다는 점은 동일하지만 각자 상이한 지원 환경, 플랫폼 관리도구, 솔루션을 제공해 눈길을 끌고 있다.

상용 하둡 버전에 대한 옵션 증가와 함께 이 오픈 소스 플랫폼이 힘을 받아가고 있다는 다른 징후도 있다. 벤처캐피털이 자금을 대고 있으며, 관리 부가 기능과 분석 애플리케이션을 갖춘 신생 업체들이 무서운 속도로 등장하고 있다. 한몫 잡으려는 기존 데이터 관리/분석 업체(IBM, Oracle, EMC, SAP, Teradata, 마이크로소프트 등)들의 관심도 더 커져가고 있다.

현재 하둡 상용 버전 및 관련 솔루션을 소개하고 활발히 활동하고 있는 기업들을 사업자 유형별로 살펴보도록 하자.

(1) 글로벌 SW 벤더

기존의 기존 데이터베이스(DB), 데이터 웨어하우스(DW) 및 비즈니스인텔리전스(BI) 전문 업체들이 이 유형으로 분류된다. 하둡의 등장 초기에 이들 기업들은 “하둡을 통한 빅 데이터 처리는 안정성이 떨어진다”라고 비난하고 폄하했지만, 지금은 전세가 역전되어 글로벌 SW 벤더들은 하나 같이 자사의 솔루션에 하둡을 적용했다는 점을 강조하고 있다.

적용 방법은 하둡과 통합(자체 하둡 상용 배포판 개발 또는 전문 업체 제휴) 및 연계, 어플라이언스 형태로 내재화(전문업체와 전략적 제휴 및 M&A) 등의 다양한 형태를 취하여 정형 분석은 물론 하둡으로 비정형까지 커버하는 진정한 빅 데이터 분석에 무게 중심을 두고 빅 데이터 관련 시장을 공략하고 있다.

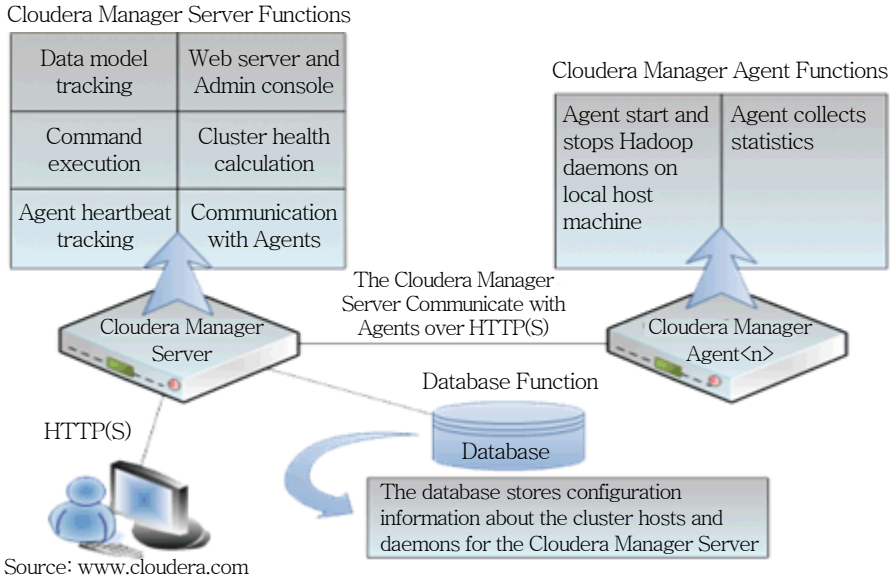
<표 4> 글로벌 SW 벤더 하둡 상용 버전

기업	제품명 (하둡 상용 버전)	특징	비고
IBM	InfoSphere BigInsight (자체 하둡, CDH)	자체 하둡 에디션 공급(베이직, 엔터프라이즈 에디션) - 베이직 에디션: 하둡, 하이브, 마홋, 우지, 주키퍼, 휴와 다른 오픈 소스 도구 포함, IBM 인스톨러 베이직 버전과 데이터 액세스 도구 제공, - 엔터프라이즈 에디션: 정교한 작업관리 도구, 주요 데이터 소스와 통합되는 데이터 액세스 계층, 클러스터에서 데이터 조작을 위한 스프레드시트 같은 빅시트(BigSheets) 추가	데이터 분석 패키지로 하둡 통합, 완전한 Hadoop 어플라이언스로는 부족, 비주얼 Map Reduce 모델링 도구 제공하지 않음, 자사 CEP 기술을 Hadoop 포트폴리오에 완전히 통합하지 않은 상태, 클라우데라와 파트너십 체결(CDH; Cloudera's Distribution Including Apache Hadoop)
Oracle	Oracle Big Data Appliance(CDH)	자체 어플라이언스에 통계 패키지 R, Oracle Enterprise Linux 5.6 운영체제 및 CDH 포함	클라우데라와 협력, 어플라이언스에 CDH 탑재
EMC	Greenplum (Greenplum HD, Greenplum MR)	정형 데이터 분석을 위한 DB 모듈과 비정형 데이터 분석을 위한 하둡 모듈을 어플라이언스로 지원(Apache Hadoop, MapR R5)	MapR 과 제휴(OEM), Greenplum MR(MapR M5, SW-only 제품)
SAP	SAP HANA (자체 하둡 없음)	자체 하둡 버전 없으며 하둡 상용 버전 공급사 제휴 예상	SAP Integrator 를 통해 Hadoop 연계 지원
Teradata	Teradata Aster Big Analytics Appliance (HDP + Aster Data)	애스터(맵리듀스와 SQL 을 결합한 분석 플랫폼)와 호튼웍스와 협력을 통해 애스터 SQL-H 를 개발, 어플라이언스에 통합	자사 데이터 처리 방식에 하둡 기술을 담음, 호튼웍스와 협력

<자료>: [12], [13], [15] 참고 재구성

나. 외국 하둡 전문 벤더

단순히 하둡을 지원한다고 해서 빅 데이터를 감당할 수 있게 되는 건 아니다. 기존 데이



(그림 4) 클라우데라 매니저 아키텍처[16]

터와 하둡을 이어질 수 있는 플랫폼이 있어야 한다. 같은 하둡을 도입해도 데이터 플랫폼에 따라 데이터 처리와 분석 과정이 얼마든지 달라질 수 있다. 본류인 미국에서 하둡을 이용한 데이터 플랫폼이 대거 등장하고 이에 대한 투자가 발 빠르게 이어지고 있는 이유가 여기에 있다.

오픈 소스 하둡을 전문적으로 개발해 상용 솔루션으로 배포하는 전문 기업들로는 클라우데라와 호튼웍스, 맵R이 대표적이다. 이들 업체들은 하둡을 기반으로 한 자체 플랫폼을 만들고 이를 기존 DB, DW, BI 솔루션 업체와의 전략적 제휴를 통해 배포하고 있다(<표 5> 참조).

<표 5> 외국 전문 벤더 하둡 상용 버전

기업명	하둡 상용 버전	특징	비고
클라우데라 (Cloudera)	CDH, Cloudera Manager	CDH: 하둡, 하이브, 마춧, 우지, 피그, 주키퍼, 휴와 다른 오픈 소스 도구 포함, 고유 제품 포함하지 않음 클라우데라 매니저: CDH 환경 관리 도구(CDH 배포 및 모니터링을 지원, 무료와 엔터프라이즈 버전) - 프리 에디션: CDH 포함, 최대 50 개 노드 클러스터 지원, 하둡 인프라 서비스 및 설정 관리 외 부가 기능 제한 - 엔터프라이즈 에디션: CDH 포함, 무제한의 노드 클러스터 지원, 능동적 모니터, 추가 데이터 분석 도구 결합	하둡 기본 소프트웨어는 무료, 클라우데라 매니저 엔터프라이즈 에디션에 대한 라이선스료(서브스크립션 방식)와 지원을 판매, Hadoop 모델링 도구 제공하지 않으며 실시간/대기 시간을 단축하는 데이터 통합도 제공하지 않음
호튼웍스 (Hortonworks)	호튼웍스 데이터 플랫폼 (HDP)	하둡, 하이브, 마춧, 우지, 피그, 주키퍼, 휴와 다른 오픈 소스 도구 포함, 업체 고유 제품 포함하지 않음, 모든 소프트웨어 무료 제공, 교육과 지원 프로그램 통해 수익	Hadoop 모델링/개발 도구, Hadoop 비즈니스 애플리케이션 또는 MapReduce 모델 라이브러리 제공하지 않음, Hadoop 데이터베이스 옵션과 함께 작동하지 않음
맵알 테크놀로지스 (MapR Technologies)	M3, M5, M7	하둡, 하이브, 마춧, 우지, 주키퍼, 휴와 다른 오픈 소스 도구 포함 - M3: 무료 버전, NFS access, 통합 관리 UI, 향상된 확장성 등 제공 - M5: 유료 버전(서브스크립션), no single points of failure, mirroring, snapshots, NFS HA, data placement control 등의 기능 제공 - M7: HBase 개선, 속도, 확장성과 안정성 향상	ODBC 지원하는 자사 기술이 가장 개방적인 하둡 배포판이라 주장, 기본적으로 HDFS 를 탑재하지 않음, HDFS API 지원, 다양한 EDW 와 통합, 탄탄한 하둡 모델링 도구와 파트너십과 OEM 파트너 보유

<자료>: [13], [14], [15] 참고 재구성

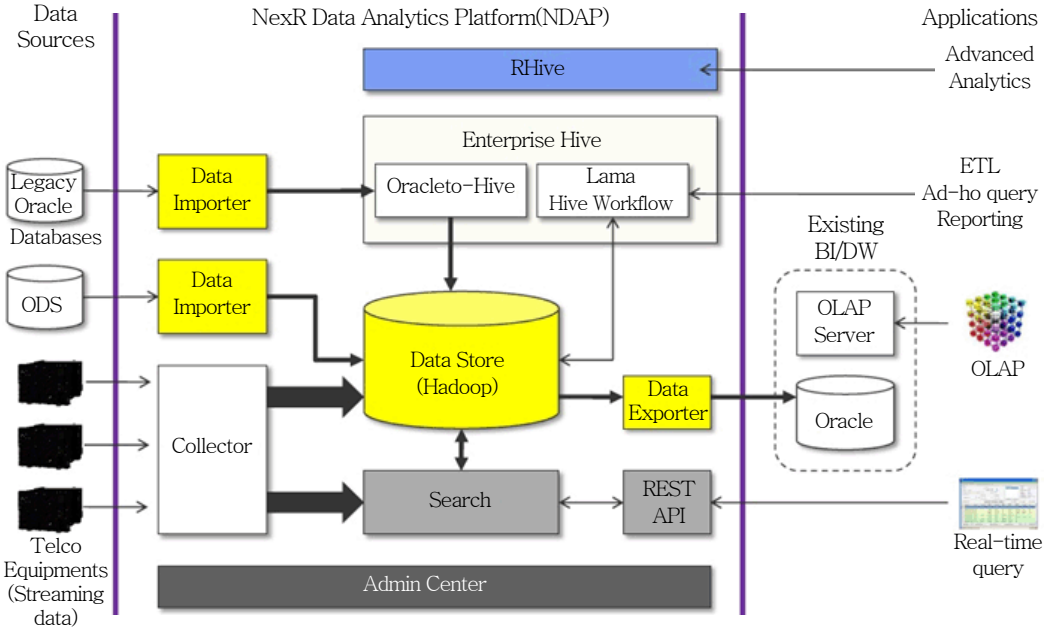
다. 국내 전문 벤더

국내에서 하둡깨나 다루어봤다는 KT Cloudware, 그루터, 클라우다인 등이 자체적으로 준비한 하둡 기반의 상용 버전을 시장에 출시하고 외산 솔루션과 주도권 싸움과 시장 선점 경쟁에 뛰어들었다.

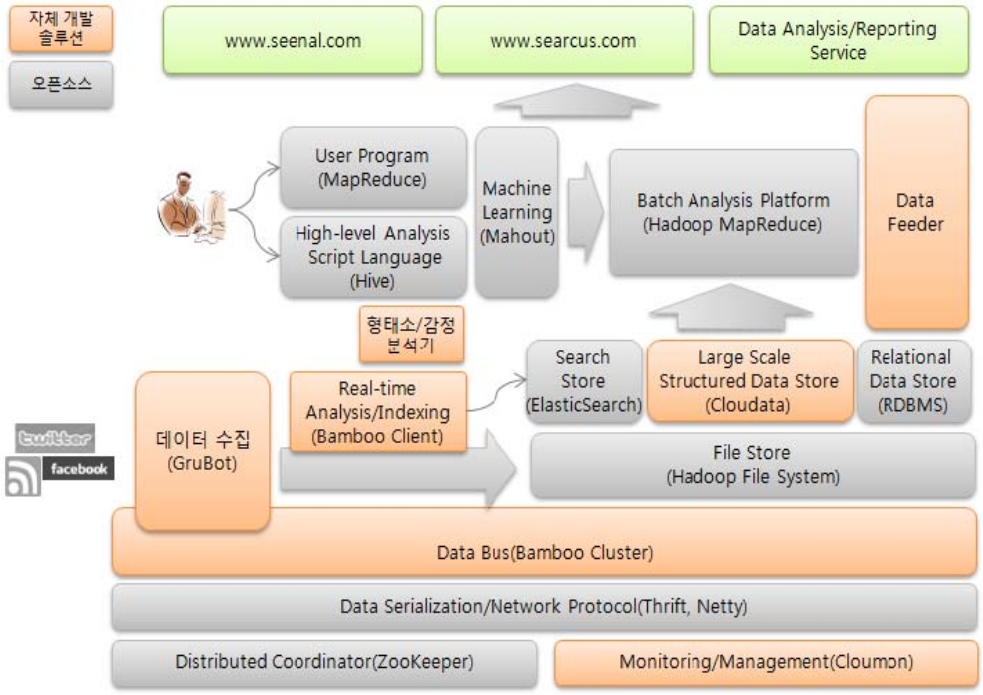
<표 6> 국내 전문 벤더 하둡 상용 버전

기업명	하둡 상용 버전	특징	비고
KT Cloudware (구 넥스알)	NDAP, RHive	NDAP(NexR Data Analytics Platform): 빅 데이터 분석을 위한 모든 작업(수집/저장/분석/검색/관리 등) 및 실시간 데이터 질의를 처리할 수 있는 소프트웨어 플랫폼(All-In-One 솔루션) RHive: 빅 데이터 분석 플랫폼, 가장 대중적인 분석 도구인 R 과 검증된 대용량 분산 DW 시스템인 Hive 를 결합, 군집 분석/회귀 분석/기계 학습/이상 징후 예측 분석/시계열 분석 등의 고급 분석 가능	낮은 TCO 와 손쉬운 확장, 빠른 분석 성능 등이 장점, 국내 실제 적용 사례 보유. 빅 데이터 플랫폼 구축 및 컨설팅 서비스 제공
그루터 (Gruter)	Qoobah, Cloumon	쿠바: 자체 하둡, 수집, 실시간 분석, 저장, 배치에 이르는 데이터 처리 과정을 관리하는 소프트웨어 스택 솔루션(HDFS, Hive, HBase, 카산드라를 활용해 PB 이상의 원본 데이터와 수백억 건 이상의 실시간 트랜잭션 처리) 클라우몬: 데이터를 손쉽게 관리, 하둡 생태계를 이루는 각 개별 요소에 대한 모니터링 기능뿐만 아니라 하둡의 파일/작업관리, 주키퍼의 노드 관리, 플럼의 데이터 플로우 관리, 하이브 쿼리 워크벤치 등과 같은 관리 기능을 제공하는 플랫폼(하둡과 하둡 에코 시스템 관리도구, 쿠바와 연동)	빅 데이터 플랫폼 구축 및 컨설팅 서비스, 빅 데이터 분석 및 데이터 제공 서비스, 빅 데이터 분석 플랫폼 제공 서비스 구축
클라우다인 (Cloudine)	플라밍고 하둡 매니저 (Flamingo Hadoop Manager)	빅 데이터 처리와 분석을 위한 웹 관리 도구로 누구나 빅 데이터 기술을 활용해 데이터를 가공 수 있는 개발·운영 환경을 제공	커뮤니티 버전과 커머셜 버전 제공, Hadoop 관련 다양한 오픈 소스를 지원하는 Ajax Rich Web Interface + Workflow Engine+ Data Source Engine 를 조합

<자료>: [13], [15], [17] 참고 재구성



(그림 5) NDAP 아키텍처[18]



(그림 6) Qoobah 아키텍처[19]

<표 6>과 같이 KT Cloudware 와 그루터가 데이터 처리 과정을 핵심으로 한 하둡 플랫폼을 선보이고 있다면, 클라우다인은 하둡을 몰라도 하둡으로 분석할 수 있는 환경을 만드는 데 집중하고 있다.

4. 결론

하둡의 잠재성에 대해 매우 열광적이지만 극복해야 할 이슈들도 많다. 이들 이슈들을 정리해보면 다음과 같다.

- 하둡은 확장성과 유연성의 장점을 갖고 있기는 하지만, 각기 다른 구성요소들로 이루어진 복잡한 프레임워크이기 때문에 관리 측면에서 복잡성을 띠는 단점이 있다. 서로 다른 구성요소를 동시에 관리해야 한다는 부담은 향후 하둡을 도입하는 기업들이 반드시 고려해야 할 부분이다. 이에 대응하기 위해서는 지속적인 활용을 위해 필요한 인력의 확보 및 내부 육성 프로그램을 통해 기반을 다지는 노력이 병행되어야 한다.
- 하둡의 상대적인 미성숙, 사용 애플리케이션의 부재, 기술의 부족 등으로 인해 기업들

이 문제를 겪게 될 가능성이 있다. 하둡에 대한 투자가 증가하고 하둡 개발자 확보를 위한 치열한 경쟁이 벌어지고 있지만 하둡 생태계에 있는 대부분의 기술은 아직 정식 1.0 버전도 배포되지 않은 상태이다.

- 어떤 애플리케이션과 어떤 작업이 하둡으로 배치되어야 하고, 어떤 것들이 기존의 RDBMS 나 엔터프라이즈 데이터 웨어하우스(EDW)에 남아 있어야 하는지에 대해서도 아직 명확한 구분이 이루어지지 않은 상태이다. 우선 기업이 데이터에서 초 단위 이하(sub-second)로 상호 간에 보고하거나, 혹은 데이터를 다단계로 복잡한 트랜잭션에서 이용하고 있다면, RDBMS 를 그대로 사용하는 편이 좋다. 하둡은 이런 영역에서는 그리 강점을 살리지 못하기 때문이다. 또한 데이터가 삽입 및 삭제를 통해 자주 갱신되고 바뀌는 경우에도 역시 하둡을 사용하지 않는 편이 좋다. 따라서 하둡과 RDBMS, 빅 데이터와 데이터 웨어하우스를 비교의 관점보다는 통합의 관점에서 조망하려는 접근방식이 필요하다.
- 오픈 소스를 상용처럼 해달라는 기업의 인식 전환이 필요하다. 여태까지 데이터 처리는 상용 솔루션으로 처리하였으므로 하둡이 상용 솔루션처럼 적용될 것이라고 인식하고 도입을 하는 것은 곤란하다.
- 하둡은 기존 오픈 소스와는 다른 유형의 오픈 소스라는 인식이 필요하다. 일반적인 오픈 소스는 프레임워크, 라이브러리 정도 수준이지만 하둡은 인프라 성격을 가진 오픈 소스라는 점을 인식해야 한다. 적용하기까지 많은 테스트와 최적화가 필요하므로 도입에 대한 구체적인 계획과 상당한 인내가 필요하다는 점을 주지하고 있어야 한다.

하둡 관련 시장에 대한 전망은 어떨까? 시장조사기관 IDC 에서는 “하둡과 맵리듀스 생태계 소프트웨어 풍경 2012”라는 보고서를 통해 2011 년 7,700 만 달러 수준인 하둡과 맵리듀스 관련 시장이 2016 년이 되면 8 억 1,280 만 달러에 이를 것으로 보인다고 분석하였다[20]. 매년 60% 넘게 성장하는 셈이다. 가히 폭발적인 증가를 예상하고 있다.

기술적인 측면의 전망으로는 규모와 혁신에 초점을 맞춘 하둡 2.0 을 발표할 예정이다[21]. 2012 년 초 알파 버전이 출시된 2.0 버전은 맵리듀스 계층을 완벽하게 다시 쓰기를 할 수 있고 스토리지 논리 계층과 HDFS 계층도 마찬가지다. 알려진 바에 의하면 하둡 2.0 은 규모와 혁신에 초점을 맞추고 있으며, 차세대 맵리듀스인 얀(Yarn)과의 연합(Federation) 기능도 추가될 예정이라 한다. 얀은 사용자가 독자적인 컴퓨터 모델을 추가하여 맵리듀스에 대

한 전적인 의존도를 없애도록 한 기능이다. 2.0 버전의 이런 기능들은 다운타임이 없는 클러스터의 구현을 가능하게 할 것이다. 확장 가능한 스토리지 또한 계획되고 있다.

여태까지 하둡 상용 버전들의 특징, 이슈, 전망 등을 살펴 보았듯이 빅 데이터의 핵심 구현 기술로서 하둡이 많이 알려진 것은 부정할 수 없는 사실이다. 하지만 분명하게 알아두어야 할 점이 있다. 하둡은 빅 데이터에만 배타적으로 사용되는 것이 아니라는 것과 하둡만으로 빅 데이터 전체를 구현할 수 있는 것은 아니라는 것이다. 하둡이 빅 데이터를 훨씬 쉽게 처리할 수 있게 해주지만, 만병통치약은 아니라는 것이다.

세계적인 상황도 그렇지만 국내 환경에서 하둡 적용에 가장 큰 장애물은 관련 전문 인력을 구할래야 구할 수 없다는 것이다. 하둡 또는 하둡과 유사한 형태의 소프트웨어 플랫폼을 다루어 본 엔지니어들이 국내에는 전무하다. 국내 데이터베이스 혹은 분석 분야에 일하고 있는 기업들은 하둡 분야에 투자를 단행하지 않아 내부 인력 확보도 거의 안되어 있다. 외산 IT 벤더들도 해외 시장에서 클라우데라, 호튼웍스, 맵R 과 같은 하둡 전문 회사들과 협력하여 고객들의 빅 데이터 처리 요구에 대응하고 있지만 정작 하둡 전문 외산 벤더들은 국내에 지사를 설립하지도 않았다. 또 국내 대기업들은 외산 상용 벤더들의 제품 사용에 익숙해 오픈 소스 기반의 하둡 전문 국내 기업들에 대한 인식이 매우 낮은 편이고, 그 가치에 대해서 여전히 제대로 평가를 해주지 않고 있다. 현재 하둡 관련 어플라이언스를 국내에 소개한 외산 IT 업체 중 전문 파트너를 확보한 곳은 거의 없다. 이러한 하둡에 대한 수요와 인력 보충 간의 격차를 메우기 위해서는 산・학・연・관 합동으로 장기 인력 양성 차원의 전문 인력 육성 전략도 수립되어 추진되어야겠지만, 한편으로는 하둡 컨설팅, 소프트웨어 개발, 교육 서비스 등을 모두 제공할 수 있는 경험과 기술 역량 및 독자적인 플랫폼을 갖춘 국내 전문 기업에 과감히 눈을 돌려 이들 기업과의 수평적 파트너십을 통해 인력 문제 해결과 관련 기술 및 경험의 이전으로 기업 내부 역량을 확보하는 것이 가장 현실적이고 현명한 방안이 될 수 있을 것이다.

<참 고 문 헌>

[1] 강동식, “빅 데이터・클라우드, 본격 성장가도 진입한다”, 디지털타임스, 2012. 12. 18.
[2] 장순환, 삼성 SDS, “2013 년 IT 핵심 트랜드 9 개 발표”, 뉴스핌, 2012. 9. 26.
[3] 김광현, “ETRI, 10 대 기술 선정…이런 IT 기술이 미래 세상을 바꾼다”, 한국경제, 2012. 7. 8.

[4] 김민숙, "빅 데이터 쟁탈전", ITDaily, 2012. 10. 31.
[5] 김지선, "IDC, 아태지역 빅 데이터 향후 5 년간 연평균 46%대 고성장", 디지털타임스, 2012. 11. 27.
[6] http://wikibon.org/wiki/v/Big_Data_Market_Size_and_Vendor_Revenues
[7] http://www.forbes.com/sites/davefeinleib/2012/06/19/the-big-data-landscape/
[8] 김병곤, "아파치 하둡으로 구현하는 빅 데이터", 한국데이터베이스진흥원, 2012. 9. 13.
[9] IT World 편집부, "빅 데이터를 위한 개방형 DB 프레임워크 '하둡'의 이해", IDG Korea, 2012. 1.
[10] 임민철, 아파치재단, "빅 데이터 기술 하둡 정식판 공개", ZDNet, 2012. 1. 6.
[11] http://www.cloudera.com/blog/2012/01/an-update-on-apache-hadoop-1-0/
[12] 오현식, "'빅 데이터 시장 패권 경쟁' 불꽃", 데이터넷, 2012. 12. 17.
[13] "빅 데이터 기업의 솔루션 및 서비스 추진 현황", NIA, 빅 데이터 전략연구센터, 2012. 9.
[14] Peter Wayne, "엔터프라이즈 하둡, 더 쉬워진 빅 데이터 처리", IDG, 2012. 4.
[15] James G. Kobielus, "The Forrester Wave™: 엔터프라이즈 Hadoop 솔루션 (2012 년 1 분기)", Forrester Research, Inc, 2012. 2. 2.
[16] https://ccp.cloudera.com/display/FREE374/Introducing+ Cloudera+ Manager+ Free+ Edition
[17] 이지영, "국내 하둡 3 인방이 준비하는 '플랫폼'은", 블로터닷넷, 2012. 3. 23.
[18] http://www.nexr.com/hw11/ndap.pdf
[19] 김형준, "Gruter's 빅데이터 플랫폼 아키텍처 및 솔루션 소개", 2012. 10.
[20] 이지영. IDC "하둡 시장, 매년 60%씩 성장", 블로터닷넷, 2012. 5. 8.
[21] Paul Krill, "빅 데이터의 핵심으로 자리잡은 하둡, IDG Tech Focus '하둡의 효과와 사례'", 2012. 11, pp.3-4.