클라우드 기반 빅데이터 기술 동향과 전망

비케이앤씨 ▌김 상 락
울산대학교 ▌강 만 모

1. 서 론

클라우드 컴퓨팅은 온디맨드 네트워크를 통해 구성 가능한 컴퓨팅 자원(예를 들어 네트워크, 서버, 스토리지, 어플리케이션, 그리고 서비스)의 공유 풀로 접속하여 유비쿼터스를 가능하게 하는 모델이다. 그것은 신속하게 프로비저닝(Provisioning)할 수 있으며 또한 서비스 제공자가 손쉽게 서비스를 해제할 수도 있다[1].

IDC에 따르면 2014년에 스마트폰의 판매량이 12% 증가하는데 반해 태블릿의 판매량은 18% 증가할 것으로 전망했다. 또한 이러한 스마트기기들 대부분이 클라우드 방식의 호스팅으로 작동하게 되어 클라우드 호스팅 소비가 25% 증가할 것으로 전망하고 있다. 포레스터 리서치(Forrester Research)의 James Staten은 "2014년에 사물인터넷을 통해 수십억의 데이터가 생산될 것으로 예측되어 지며, 데이터를 캡쳐하고, 분석하고, 그리고 응답하는데 클라우드 서비스를 사용하게 될 것이다."라고 했다.

이렇게 수많은 데이터들로부터 가치있는 정보를 얻기 위해서는 빅데이터 기술을 통한 데이터 수집 및 분석 작업이 필요하다. 이 기술이 필요한 조직이나 기업에서는 프라이빗 또는 퍼블릭 클라우드 서비스 방식의 시스템 도입을 검토하고 있다. 대규모의 기업이나 조직에서는 많은 비용이 들어가고 고급 인력이 필요하더라도 프라이빗 클라우드 서비스 방식의 빅데이터 시스템 도입을 원할 것이다. 그 이유는 안정적인 서비스를 확보할 수 있고 보안 측면에서도 강력하기 때문이다. 하지만 소규모의 기업 또는 조직에서는 빅데이터 기술이 필요한 영역이 그렇게 많지 않기 때문에 빅데이터 기술을 전문 서비스 공급사로부터 빌려서 사용하는 퍼블릭 클라우드 서비스를 이용하게 될 것이다.

본 고에서는 이러한 클라우드 기반 빅데이터 기술의 정의 및 동향을 살펴보고자 한다. 아울러 글로벌 IT 선진 기업들이 발표한 클라우드 기반의 빅데이터 솔루션들에 대한 소개와 활용 사례에 대해서도 알아보고자 한다.

2. 클라우드와 빅데이터 정의

2.1 클라우드의 정의

클라우드(Cloud)는 다양한 클라이언트 디바이스에서 필요한 시점에 인터넷을 이용해서 공유 풀에 있는 서버, 스토리지, 어플리케이션, 서비스 등과 같은 IT 자원에 쉽게 접근하는 것을 가능하게 하는 모델이다. 이것은 서비스를 배치하는 방법에 따라 3가지로 분류할 수 있다. 첫째, 퍼블릭(Public) 클라우드이다. 특정 기업이나 사용자를 위한 서비스가 아닌 인터넷에 접속 가능한 모든 사용자를 위한 클라우드 서비스이다. 모든 사용자를 위한 서비스이지만 서비스 내부에 저장된 데이터나 기능, 서버 같은 자원은 각 서비스에서 사용자별로 권한 관리가 되거나 격리되어 서비스 사용자 간에 전혀 간섭이 일어나지 않는다. 둘째, 프라이빗(Private) 클라우드이다. 퍼블릭 클라우드의 개념 중 일부를 제한된 네트워크 상에서 특정 기업이나 특정 사용자만을 대상으로 하는 클라우드 서비스이다. 컴퓨팅 자원과 클라우드 내에 저장되는 데이터는 기업 내부에 저장되기 때문에 자원의 제어권을 기업 자체에서 갖고 있으며, 데이터 보

표 1 퍼블릭, 프라이빗 클라우드 비교

속성	퍼블릭 클라우드	프라이빗 클라우드	서버 가상화
기민한 탄력성	○	△	△
측정 가능한 서비스	○	○	○
온디멘드 셀프 서비스	○	△	△
네트워크 접근	○	△	△
자원 풀링	서버, 스토리지, 데이터베이스 등 모든 자원	대부분 서버 자원만 필요함	서버 자원만 가능

안 측면에서 퍼블릭 클라우드에 비해 강점이 있다. 셋째, 하이브리드(Hybrid) 클라우드이다. 퍼블릭 클라우드와 프라이빗 클라우드를 병행해서 사용하는 방식이다. 데이터 보안이 중요하거나 컴퓨팅 자원에 대한 제어권을 가져야 하는 서비스나 시스템은 프라이빗 클라우드를 사용하고 그렇지 않은 경우에는 퍼블릭 클라우드를 사용한다[2].

2.2 빅데이터의 정의

데이터의 규모(Volume)가 방대하고, 다양한 종류(Variety)의 정형・비정형 데이터, 분석・예측을 즉시(Velocity)에 해결 등의 특징을 가지는 데이터를 빅데이터라고 한다. 좀 더 포괄적으로 빅데이터를 대용량의 데이터를 저장, 수집, 발굴, 분석, 비즈니스화 하는 일련의 과정을 의미하는 용어로 변화하고 있다. 이러한 빅데이터를 처리하는 핵심 기술이 하둡이다. 하둡은 2005년 루씬(Lucene) 개발자인 더그 커팅과 마이크 카파렐라가 구글의 맵리듀스(MapReduce) 알고리즘을 구현하기 위해 개발하였다. 하둡은 대용량 데이터 처리 분석을 위한 대규모 분산 컴퓨팅 지원 프레임워크로 하둡 분산 파일시스템(HDFS), 분산 처리를 위한 프레임워크인 맵리듀스(MapReduce)가 핵심 기술이며, 이외에도 분산 데이터베이스인 HBase, 검색엔진인 Nutch, SQL을 지원하는 하이브(Hive) 등 빅데이터를 위한 통합 솔루션으로 발전하고 있다[3].

3. 클라우드 컴퓨팅과 빅데이터 동향

3.1 클라우드 컴퓨팅 동향

클라우드 컴퓨팅 기술이 21세기 사회 및 산업에 미치는 막대한 영향을 인식하여 미국 등 선진국에서는 새로운 도전적 연구 분야로 선정하여 범국가적인 관심을 가지고 연구를 지원하기 시작했다. 미국은 연방 CIO위원회를 중심으로 연방조달청, 클라우드 컴퓨팅 실행조정위원회, 클라우드 컴퓨팅 자문위원회, 국토안보부, 국립표준기술연구소 등으로 조직을 구성하고 공공부문의 클라우드 관련 추진 전략을 수립하였다. 2008년 미국 국방부 산하 기관에서 클라우드 서비스를 구축하는 정책 수립을 시작으로 2009년에는 연방정부 포털의 클라우드 전환, 2010년에는 공공데이터센터 통합 계획(FDCCI 정책), 2011년에는 Cloud First 정책을 발표하고 공공분야 클라우드 확산체계인 FedRAMP(Rederal Risk Authorization and Management Program)를 개발하여 공공부문에 적용하였다.

일본은 총무성 주도로 가스미가세키 클라우드(전자정부・전자지자체 클라우드)를 추진하여 정부 정보시스템의 통합・집약화 및 각 정보 시스템 데이터를 연계함으로써 2015년까지 디지털기술에 의한 새로운 행정개혁을 추진하는 것을 목표로 하고 있다. 또한 일본 대지진 이후 클라우드의 중요성이 급부상하고 있으며 재난관리, 지자체, 교육, 환경, 안전, 의료분야를 대상으로 적용을 강화하는 정책을 추진 중이며, JCC(Japan Cloud Consortium)과 NICT(National Institute of Infor-

표 2 클라우드 컴퓨팅 분야의 오픈 소스 소프트웨어

클라우드 컴퓨팅 요소 기술	오픈 소스 소프트웨어
가상머신모니터 기술	Xen, KVM, VirtualBox
가상머신모니터 기능 제어 기술	Libvirt, Ovirt, virt-manager
스토리지 가상화 기술	LVM, Hadoop, GFarm, Swift
네트워크 가상화 기술	Neotron, Open Day Light, Mul
가상머신 이미지 관리 기술	Glance, Aeolus
가상머신 시스템 제어 기술	OpenStack, CloudStack, Eucalyptus, Nimbus, OpenNebula, ConVirt, Wakame-VDC, SlapOS
물리시스템 제어 기술	Groundwork, Zabbix, Nagios, Hinemos, Ganglia
응용 설치 자동화 기술	Chef, Puppet, Crowbar, Fuel
사용자 인증 기술	OpenSSO, Higgins, Shibboleth
미터링 기술	Ceilometor
클라우드 인터페이스 기술	Delta Cloud, Simple Cloud, libCloud

표 3 클라우드 컴퓨팅 분야의 표준화 현황

단체명	표준화 내용
DMTF	Open Virtualization Format (OVF) Cloud Infrastructure Management Interface (CIMI)
IEEE	P2302 - Standard for Intercloud Interoperability and Federation (SIIF)
IETF	Cloud Service Broker (DRAFT)
ISO JTC 1/SC 38	Distributed Application Platforms and Services (DAPS)
ITU-T	Future networks including cloud computing, mobile and NGN
NIST	Cloud Computing Reference Architecture
OGF	Open Cloud Computing Interface (OCCI)
Storage Network Industry Association(SNIA)	Cloud Data Management Interface (CDMI)

mation and Communication Technology) 연구소를 중심으로 대규모 연구개발을 추진 중이다.

영국은 2010년 1월 국가 전체를 클라우드화 하는 실질적인 전자정부 구현을 위한 개혁안인 G-Cloud(Government Cloud) 정책을 추진하여 국가 단위의 클라우드 컴퓨팅 프로젝트를 통해 예산 절감(연간 32억 파운드 절감), 공공부문 업무 효율성 증대 등 각종 시너지 효과를 기대하고 있다.

유럽은 ITEA2(IT European Advancement 2) 정책을 통해 클라우드 연동, 클라우드 SLA, 다중 클라우드 테스트베드 구축 등 대규모 연구 투자를 진행하고 있다. 중국은 우시, 상해, 북경, 청도, 항주 등을 중심으로 클라우드 데이터센터 인프라를 구축하였으며, 정부는 클라우드 기반의 소프트웨어 산업 경쟁력 강화를 위해 서비스 모델 발굴, 클라우드 안전관리 규범 제정 및 자국 내의 시범 서비스 사업을 추진하고 있다.

국내의 클라우드 컴퓨팅 연구는 지난 2000년 후반부터 시작되어 지금까지는 주로 단일 클라우드 서비스 기술 분야에 집중하였으나, 향후에는 보다 개방적이고 소비자들을 위한 클라우드 기술과 타 기술과 융합을 위한 연구가 필요하다. 또한 클라우드 컴퓨팅 기술은 오픈 소스 프로젝트 진영에서 매우 활발하게 진행되고 있다. KT 등 많은 기업들이 오픈 소스 소프트웨어를 활용하고 있다. 한국전자통신연구원(ETRI)에서는 대규모 이종 클라우드 인프라 관리 기술, 가상화 기술, 스토리지 기술, SaaS(Software as a Service) 기술 등을 연구하고 있다. 국내외적으로 클라우드 컴퓨팅 분야의 오픈 소스 소프트웨어 기술과 표준화 활동은 매우 활발하게 진행되고 있다[3-6].

3.2 빅데이터 동향

각국 정부의 빅데이터 정책은 크게 R&D, 진흥정책, 그리고 법제도적 개선 정책 등으로 구분할 수 있다. R&D 정책의 세부 내용으로 공공정보 개방, 공공서비스 플랫폼・기술 기준 등이 포함된다. 진흥정책의 핵심 내용은 빅데이터 인력 양성이고 법제도적 개선 정책의 핵심 내용은 개인정보보호, 보안 등이다. 이러한 구분은 영국의 Big Innovation Centre가 영국 정부에 빅데이터 활용을 위한 정책 제언에서 5개의 빅데이터 정책적 도전과제를 선별하여 제언한 기준에 따른 것이다.

미국은 민간부문에서 글로벌 빅데이터 시장을 주도하고 있다. 구글, 페이스북 등 글로벌 인터넷 기업들은 이미 빅데이터를 분석하고 광고하는 마케팅분야에 활용하고 있으며, IBM, Microsoft, Oracle, SAP, SAS 등 글로벌 소프트웨어 전문 기업들과 신생 기업들도 이 분야에서 성장 중에 있다. 민간의 자발적인 성장 뿐만 아니라 정부가 공개하고 있는 공공정보를 활용한 기술 개발 중심으로 구체적인 진흥 정책이 추진되는 등 정책적인 측면에서도 가장 앞서 있다. 미국 정부는 2012년 3월 "빅 데이터 R&D 계획(Big Data R&D Initiative)"을 발표하였다. 이 계획은 정부기관이 공공정보를 개방하고, 빅데이터를 활용하여 공공서비스도 개혁하겠다는 의도를 담고 있다. 정부가 빅데이터를 활용하여 투명하고 효율적이며 혁신적인 서비스를 제공하겠다는 것이다.

유럽은 미국에 비하여 빅데이터 시장이 제한적으로 형성되어 있다. 금융, 은행, 투자사 등 민간 금융 영역만이 미국과 동등한 수준에서 빅데이터를 활용하는 정도이다[7]. 그러나 공공부문 데이터 공개에 대해서는 적극적인 정책을 진행하였다. 특히, 유럽은 일찍이 공공정보에 대해서는 개인의 접근권리를 인정하고 허용하는 입장을 취해 왔다. 공공기관이 보유한 정보는 결국 국민이 부담한 조세를 통해 축적된 것이므로 납세자인 시민이 공공정보 접근 및 재사용(re-use)할 권리를 가지는 것이 당연하다고 판단하였기 때문이다. 공공부문의 데이터 개방 확대가 진행되면서, 유럽은 글로벌 경제위기, 데이터의 폭증, 데이터 활용기술의 진화 등의 환경 변화로 인해 공공기관 데이터가 가진 경제적 가치에도 관심을 기울이게 되었다.

일본은 민간부문에서 일부 빅데이터 활용 사례가 증가하고 있으나, 전반적인 활용 정도는 미국이나 유럽에 비해 낮은 편이다[8]. 일본의 빅데이터 정책은 2012년 수립된 Active Japan 정책에 반영되어 있다. 일본 총무성은 2000년대 이후 성장 정체, 국제경쟁력 저하, 2011년 동일본 대지진 등 빈번한 자연재해 등 국가위기 상황을 극복하기 위해 2012년 5월, ICT에 기반한 국가 주도의 종합적 진흥 정책인 'Active Japan ICT'계획을 발표하였다. Active Japan 계획은 5개 부문으로 구성되어 있는데, 이 중 'Active Data'부문이 빅데이터 관련 정책에 해당한다. Active Data 정책은 다종, 다량의 빅데이터를 실시간으로 수집, 전송, 해석하여 재난관리를 포함한 정책 과제 해결에 이용함은 물론, 수십조 엔 규모의 데이터 활용 시장 창출을 목표로 하고 있다. 빅데이터를 국가 자산화하여 성장 동력을 육성하겠다는 계획이다[9].

4. 클라우드 기반의 빅데이터 솔루션 현황

본 장에서는 글로벌 IT 선진 기업들이 개발한 클라우드 기반의 빅데이터 주요 솔루션인 구글 빅쿼리, 아마

존 엘라스틱 맵리듀스 플랫폼, IBM 인포스피어 빅인사이트, 페이스북 프레스토, 마이크로소프트 애저 HDInsight 등에 대해서 소개한다.

4.1 구글 빅쿼리

빅쿼리(BigQuery)는 빅데이터를 클라우드 상에서 신속하게 분석해 주는 서비스이다. 빅쿼리 서비스를 사용하면 대규모 데이터셋에 대한 SQL(Structure Query Language) 유사 쿼리를 실행할 수 있다. 빅쿼리는 대규모 데이터셋을 대화식으로 분석하는 데 적합하다. 기존의 관계형 데이터베이스 시나리오에서는 Google Cloud SQL을 대신 사용할 수 있다. 빅쿼리 브라우저 도구라는 웹 사용자 인터페이스나 bq 명령줄 도구를 통해 빅쿼리를 사용할 수도 있다. 또한 자바, 파이썬(Python) 등 여러 언어가 지원되는 다양한 클라이언트 라이브러리를 사용하여 REST(Representational State Transfer) API를 호출하는 방식으로 빅쿼리를 사용할 수도 있다.

표 4 빅쿼리 주요기능

기능	기능 설명
속도	몇 십억 개의 행을 몇 초 만에 분석
규모	조 단위의 레코드를 포함하는 테라바이트 크기의 데이터를 지원
단순성	구글 인프라에서 호스팅되는 SQL 유사 쿼리 언어
공유	구글 계정을 사용하는 강력한 그룹 및 사용자 기반 권한을 제공
보안	보안 SSL 액세스
다양한 액세스 방법	빅쿼리 브라우저, bq 명령줄 도구, REST API 또는 Google Apps 스크립트를 사용하여 빅쿼리에 연결 가능

구글에서 제공되고 있는 BigQuery와 Google Cloud SQL 서비스의 차이점은 표 5와 같다[10].

빅쿼리 서비스는 구글이 2010년 개발자 컨퍼런스 Google I/O에서 공개한 클라우드 분석 서비스로 일부 기업 이용자 및 개발자들에게만 무료 버전으로 제공하고 있다. 빅쿼리 서비스를 이용할 경우 이용자들은 전용 데이터센터에 별도의 자원을 투자하지 않고도 거대한 양의 데이터를 업로드만 하면 분석이 가능하다. 빅쿼리는 사용자에게 그래픽 사용자 인터페이스 기반 SQL 분석 솔루션을 제공하며, 특히 분산된 개별 데이터 분석 결과를 요약이나 통합 과정 없이 모두 제공하므로 사용자가 직접 각 분석 결과를 파악하고 판단할 수가 있다. 기존 빅데이터 분석 시스템은 기업 내부에서 이루어지는 것이 일반적이었으나, 빅쿼리는 클라우드 기반의 분석 플랫폼 제공으로, 기업이 별도의 인프라 투자 없이도 빅데이터 분석 업무를 수행할 수 있다.

구글은 현재 빅쿼리 서비스를 위해 이미 독자적으로 개발한 데이터 분석 툴을 확보하였으며, 이를 클라우드를 통해 서비스 상용화를 검토 중에 있다. 구글 클라우드 플랫폼 팀의 주-카이 퀙(Ju-Kay Kwek) 상품 매니저는 미국에서 발간하는 IT 웹진 기가옴(GigaOM)이 개최한 컨퍼런스에서 '빅쿼리가 급변하는 시장 트렌드를 빠르게 읽는데 유용한 도구가 될 것'이라고 언급했다. 2011년부터 베타 서비스로 제공되고 있는 빅쿼리는 빅데이터에 관심은 많지만 인프라 구축에 부담을 느끼는

표 5 BigQuery와 Google Cloud SQL 차이점

BigQuery	Google Cloud SQL
• BigQuery는 대용량 데이터(최대 몇십억 개의 행)에 대한 쿼리를 빠른 속도로 실행하는 데 적합함 • BigQuery는 방대한 양의 데이터를 빠르게 분석하는 데 적합하지만 데이터를 수정하기에는 적합하지 않음. 데이터 분석 측면에서 BigQuery는 OLAP(온라인 분석 처리) 시스템임 • 데이터를 CSV 파일 형식으로 BigQuery에 가져올 수 있으며, 이 경우 데이터는 클라우드에서 상대적으로 적은 수의 테이블에 저장되고 서로 간에 명시적 관계를 갖지 않음 • BigQuery는 데이터베이스 시스템이 아니며 다음과 같은 특징을 가짐 - 테이블 색인 또는 다른 데이터베이스 관리 기능을 지원하지 않음 - BigQuery는 특수 SQL 하위 집합을 지원하지만 업데이트 또는 삭제 요청은 지원하지 않음 - BigQuery는 한쪽 조인이 다른 쪽 조인보다 훨씬 작을 경우에만 조인을 지원함 • BigQuery는 인터넷을 통해 REST 명령을 보낼 수 있는 클라이언트에 사용될 수 있음	• Google Cloud SQL은 Google에서 호스팅하는 MySQL 인스턴스임. MySQL은 전체 SQL 구문 및 테이블 관리 도구를 지원하는 완벽한 관계형 데이터베이스 시스템임 • Google Cloud SQL은 사용자가 자주 쿼리하는 중소형 데이터셋에 적합하며, 쿼리 업데이트, 추가 및 삭제를 지원하므로 일반적으로 데이터를 분석하기 보다는 관리용으로 사용됨. 데이터 분석 측면에서 Google Cloud SQL은 OLTP(온라인 트랜잭션 처리) 시스템임 • 높은 QPS(Query Per Second) 속도로 웹 페이지를 제공하는 데 적합함 • 지연 시간이 짧은 다수의 조회에 적합함(예: 일일 수백만 개의 쿼리로 웹페이지를 채움) • mysqldump를 사용하여 다른 MySQL 데이터베이스에서 데이터베이스를 가져올 수 있음 • Google Cloud SQL은 App Engine 어플리케이션에서만 액세스할 수 있음

업체들, 특히 광고, 의료, 지식정보 업계에서의 도입에 대한 검토가 두드러지고 있다. 전 세계를 대상으로 한 광고 캠페인의 투자수익률(ROI)이나 광고 효율을 분석하는데 빅쿼리가 유용하게 사용될 것이다. 또한 의료 업계에서도 구글 앱 엔진(App Engine)의 인터랙티브 대시보드 형태로 빅쿼리가 제공되고 있는데, 매출관리 어플리케이션, 광고 데이터 통합, 예약 데이터 및 설비・재고품 데이터 관리 등에 광범위하게 활용되고 있다. 비전문가들도 쉽게 각종 데이터에 접근할 수 있으므로 빅쿼리가 제공하는 인터랙티브 대시보드를 통해 통찰력을 얻고 고객과의 대화 방향을 설정하는데 중요하게 사용되어질 것이다[11].

대표적인 빅쿼리 사용 업체로는 Interaction Marketing, Boo-box, redBus 등이 있다. Interaction Marketing에서는 판매 패턴, 구매자 행동 등을 분석하는데 빅쿼리를 사용하고 있다. 또한 Interaction의 모든 데이터가 클라우드로 저장되어 IT인프라 투자 비용을 절감했고, 빅쿼리를 이용하여 거대한 데이터로부터 최소의 투자로 최대의 수익을 올리기 위한 실시간 통찰력을 얻고 있다[12].

Boo-box는 브라질에 있는 최고의 광고 네트워크 사업체 중 하나로 구글 빅쿼리를 사용하여 더욱 효과적으로 광고 전략을 수립하고 있다. 게시자는 자신의 웹사이트와 블로그에 광고를 호스팅하여 수익을 창출한다. 광고주는 웹사이트 상에 표시되는 광고를 통해 새로운 고객에게 접근한다. Boo-box는 빅쿼리를 사용하여 매달 35만개의 블로그와 웹사이트 상에 노출되어 있는 30억개 이상의 광고속에서 거의 실시간으로 통찰력을 얻고 타켓 광고를 하고 있다. 빅쿼리 사용으로 이 회사는 매년 20만 달러를 절감했다. 빅데이터 운영을 위해 추가적인 하드웨어 증설이 필요 없고 하둡을 운영할 훈련된 직원이 필요 없기 때문이다[13].

redBus는 인도에 인터넷 버스 티켓팅을 도입하여 수만 건의 버스 일정을 단일 예약 작업으로 통합한 온라인 여행 대행사이다. 빅쿼리를 사용하여 redBus는 테라바이트 수준의 예약과 미예약 데이터를 단지 수초 만에 분석한다. 또한 빅쿼리는 엔지니어가 신속하게 결함을 수정하여 영업 손실을 최소화하고 고객 서비스를 획기적으로 개선하는데 중요한 도구가 되었다. 빅쿼리는 복잡한 하둡 인프라를 운영하는 비용의 20% 수준에서 실시간에 가까운 데이터 분석 능력을 제공받는다. 이전에 하둡 프레임워크에서 하루가 걸리는 조회 작업들이 구글 웹기반 빅쿼리 서비스를 사용한 후에는 30초보다 적게 걸린다[14].

4.2 아마존 엘라스틱 맵리듀스

엘라스틱 맵리듀스(Elastic MapReduce)는 대용량 데이터를 처리하는 아마존 웹 서비스 플랫폼이다. 클라우데라 등과 달리 아마존 엘라스틱 맵리듀스는 퍼블릭 클라우드 서비스 형태로 제공된다. 하둡 2.2를 반영하였으며 하이브, 피그, HBase 등도 최신 버전이 탑재되어 있다. 이것을 사용하면 실무자, 연구원, 데이터 분석가 및 개발자가 대용량의 데이터를 간편하고 비용 효율적으로 처리할 수 있다. 이 서비스는 호스팅으로 운영되는 하둡 프레임워크를 사용한다. 하둡 프레임워크는 엘라스틱 맵리듀스를 사용해 원하는 만큼의 용량을 즉시 공급하여 웹 인덱싱, 데이터 마이닝, 로그 파일 분석, 데이터 웨어하우징, 시스템 학습, 재무분석, 과학 시뮬레이션, 생물정보학 연구 등과 같은 어플리케이션의 데이터 집약적인 작업을 수행할 수 있다. 엘라스틱 맵리듀스는 클라우드 서비스 형태로 제공되기 때문에 시간이 걸리는 하둡 클러스터에 대한 설정, 튜닝, 그리고 관리하는데 신경을 쓰거나 컴퓨팅 파워에 대해서 걱정할 필요 없이 데이터를 고속으로 처리하고 분석하는데 집중할 수가 있다[15].

4.3 IBM 인포스피어 빅인사이트

인포스피어 빅인사이트(InfoSphere BigInsights)는 IBM의 대용량 정보분석 솔루션이다. 빅인사이트는 대용량 정보를 분석한다는 의미와 통찰력을 높인다는 뜻이 동시에 내포되어 있다. IBM은 빅 인사이트 제품 이전에도 대용량 정보처리 및 저장을 위한 하둡 기술을 확보하고 있었다. 하지만 하둡은 정보처리 측면에서는 우수한 성능을 보였지만 분석 기능에서는 취약점이 있었다. 빅인사이트는 이러한 기존 하둡의 약점을 보완하는 기술이다. 빅인사이트는 음성, 영상 등과 같이 구조화 되어 있지 않은 비정형 데이터에 대한 분석 성능을 향상시켰다.

빅인사이트는 많은 기업들에게 하둡의 힘을 제공한다. 아파치 하둡(Apache Hadoop)은 오픈 소스 소프트웨어 프레임워크로, 대용량의 구조화된 데이터 및 구조화되어 있지 않은 데이터에 대한 신뢰성 있는 관리 작업에 사용된다. 빅인사이트는 분석 기능과 함께 관리와 워크플로우, 프로비저닝, 보안 기능을 추가하여 기존 하둡 기술을 보다 향상시켰다[17].

대용량 분석 시장을 타겟으로 하는 빅인사이트는 자사 분석 솔루션인 코그노스(Cognos), SPSS, 인포스피어 서버 등과 결합할 경우 분석 솔루션 포트폴리오의 완성도가 더욱더 높아질 전망이다. 또한 빅인사이트는 클라우드 컴퓨팅 환경에서의 운영이 가능하도록 최적

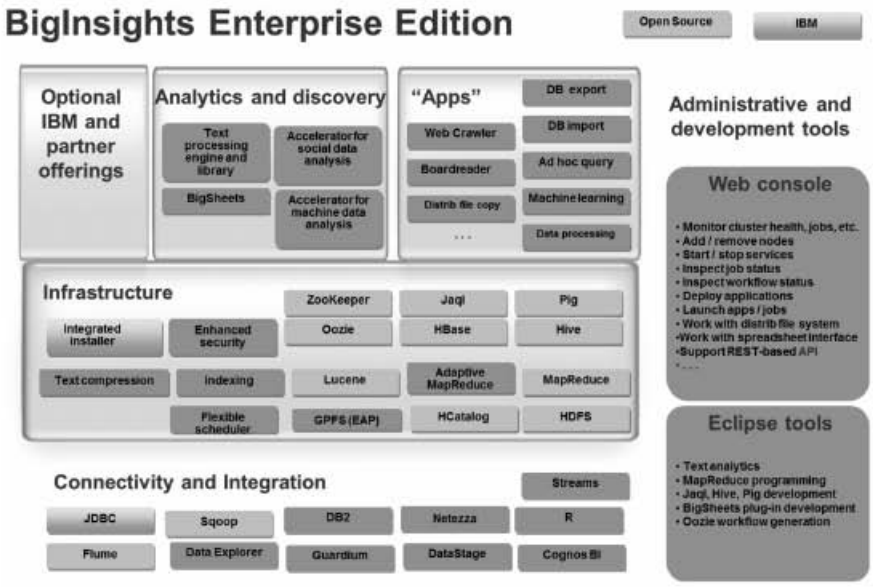


그림 1 BigInsights Enterprise Edition[16]

화되었다. 최대 140여대에 이르는 여러 컴퓨터들을 서로 연결하여 스토리지 사용과 분석 기능을 최적화시킴으로써 한 컴퓨터에 문제가 생기더라도 전체 시스템에 영향을 미치지 않고 업무를 수행할 수 있다.

인포스트림(InfoSphere Streams)으로 구현되는 실시간 분석과 이력 데이터 분석을 지원하는 빅인사이트는 매우 효과적으로 상호 시너지를 낼 수 있는 솔루션이다. 예를 들어 실시간 분석은 경우의 수에 적용되는 일종의 패턴들에 따라 분석을 수행하게 되는데, 빅인사이트가 분석 대상으로 하는 엄청난 양의 축적된 이력 데이터를 분석한 결과의 패턴은 더욱 정확하기 때문이다[18].

빅인사이트를 사용하면 내장된 처리 엔진과 어노테이터 라이브러리를 통해 대량의 텍스트 기반 정보를 쉽게 분석할 수 있다. 개발자들은 문서와 메시지에서 사람, 이메일 주소, 주소, 전화번호, URL, 공동투자업체, 협력업체 등 관심 있는 항목을 빠르게 조회하고 식별할 수 있다. 이러한 기능으로 기업들은 비구조화된 텍스트 데이터에 숨겨진 연관된 비즈니스 정보의 컨텍스트와 내용을 보다 쉽게 이해할 수 있다. 또한 프로그래머는 빅인사이트의 이클립스 기반 플러그인을 사용하여 자신만의 텍스트 분석 함수를 작성할 수 있다. 패턴 검색, 표현식 빌더, 테스트 환경이 내장되어 빠른 프로토타이핑이 가능하며, 테스트를 통해 특정 어플리케이션의 요구사항에 맞는 복합적인 텍스트 분석 함수를 작성할 수 있다. 사용자들은 스프레드시트형 데이터 검색 및 시각화 도구를 통해 많은 양의 데이터를 분석할 수 있다. 이 도구는 데이터를 수집 및 통합한 후 내장된 함수와 매크로로 데이터를 탐색하고 조작하는 기능과 차트를 작성하고 결과를 내보낼 수 있는 기능 등을 제공하는 단순한 그래픽 사용자 인터페이스 도구이다.

빅인사이트와 빅데이터에 대한 분석을 기존의 전사적 소프트웨어에 통합하는 것은 IBM의 핵심 이니셔티브 중 하나이다. 이를 위해 빅인사이트는 DB2, 네티자(Netezza) 등과 같은 유명 데이터 웨어하우스 플랫폼에 대한 연결을 제공한다. 특히, 빅인사이트는 네티자와 DB2에 대한 JDBC(Java Database Connectivity) 연결을 제공하기 때문에 개발자들이 데이터베이스의 기본적인 병렬 처리 기능을 사용하는 방법으로 각 데이터 소스와 데이터를 주고 받을 수 있어 효율성과 확장성이 보장된다. IBM의 접근 방법은 기업의 빅데이터와 기존 데이터를 서로 격리시키지 않는다. 오히려 기존의 전사적 소프트웨어 플랫폼의 가치를 최대화하고, 기존 플랫폼에 맞지 않거나 쓸모 없는 분석 워크로드에는 빅인사이트를 활용하여 통합적인 방식으로 기업의 비즈니스 분석 기능을 확장시킬 수 있도록 지원한다.

빅인사이트는 내장 웹 콘솔에 LDAP(Lightweight Directory Access Protocol) 인증을 적용하여 강력한 보안을 제공한다. 관리자는 LDAP 및 역방향 프록시 지원을 활용하여 승인되지 않은 사용자의 접근을 제한할 수 있다[19].

4.4 페이스북 프레스토

프레스토(Presto)는 아파치 오픈소스 진영을 둘러싼 여러 업체가 개발 중인 'SQL온하둡' 기술 가운데 하나

이다. 프레스토는 클라우드 기반 서비스로 기가바이트에서부터 페타바이트에 이르기까지 모든 크기의 데이터 소스에 대하여 상호 분석 쿼리를 실행하기 위한 오픈 소스 분산 SQL이다. 프레스토는 상호작용 분석을 위해 설계 및 개발되었으며 페이스북과 같은 조직에 적용되면서 기술의 수준이 상용 데이터 웨어하우스의 처리 속도에 접근하고 있다.

프레스토는 하이브, HBase, 관계형 데이터베이스 또는 전용 데이터 저장소의 데이터를 조회할 수 있다. 단일 프레스토 쿼리는 다수의 데이터 소스로부터 데이터를 결합할 수 있다. 페이스북은 300페타바이트 데이터 웨어하우스를 포함한 여러 내부 데이터 저장소에 대한 대화형 쿼리를 위해 프레스토를 사용한다. 천명 이상의 페이스북 직원들은 매일 3만 쿼리 이상을 실행하기 위해 프레스토를 사용한다. 하루에 전체 검색 데이터 량이 페타바이트 이상이다. Airbnb와 Dropbox 등의 선도적인 인터넷 기업들이 프레스토를 사용하고 있다.

프레스토는 기계의 클러스터에서 실행되는 분산 시스템이다. 전체 구조는 조정자(Coordinator)와 작업자(Worker)를 포함한다. 쿼리는 프레스토 CLI(Command Line Interface)와 같은 클라이언트로부터 조정자로 제출된다. 조정자는 구문을 파싱, 분석, 그리고 쿼리 실행 계획을 세운 후 작업자로 처리를 분배한다.

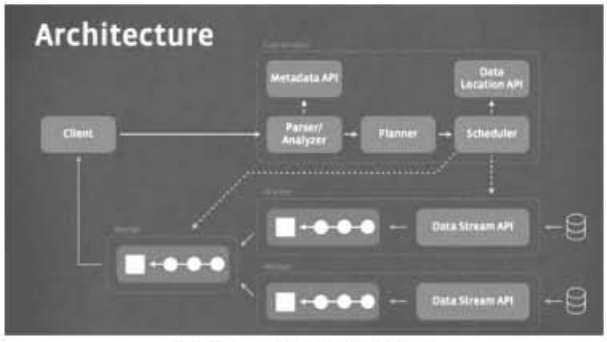


그림 2 프레스토 아키텍처

SQL온하둡 기술은 하둡에서 빠른 속도와 대화형으로 쿼리할 수 있는 환경을 제공한다. 프레스토처럼 맵리듀스를 사용하지 않고 SQL을 지원하는 빠른 쿼리 엔진으로는 맵알 '드릴(Drill)', 클라우데라 '임팔라(Impala)', 스토리지업체 EMC의 자회사 피보탈(Pivotal)에서 개발한 '호크(HAWK)', 호튼웍스의 '스팅거(Stinger)' 등이 있다. 또한 국내 업체인 그루터에서 만든 '타조'가 있다[20,21].

4.5 마이크로소프트 HDInsight

HDInsight는 100% 아파치 하둡 솔루션을 클라우드 방식으로 제공하는 마이크로소프트의 하둡 기반 서비스이다. 구조화 여부 및 크기에 관계없이 모든 형식의 데이터를 관리하는 현대적인 클라우드 기반 데이터 플랫폼인 HDInsight는 빅데이터의 완전한 가치를 얻을 수 있게 한다. HDInsight를 사용하면 단순함, 관리의 용이성 및 클라우드에서 실행되는 모든 개방형 엔터프라이즈 지원 하둡 서비스를 제공하는 마이크로소프트의 현대적인 데이터 플랫폼을 통해 모든 형식의 데이터를 원활하게 처리할 수 있다. 마이크로소프트 데이

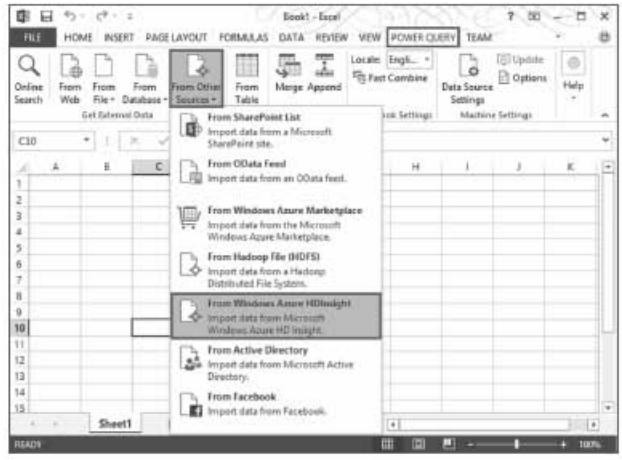


그림 4 엑셀에서 HDInsight 데이터 가져오기

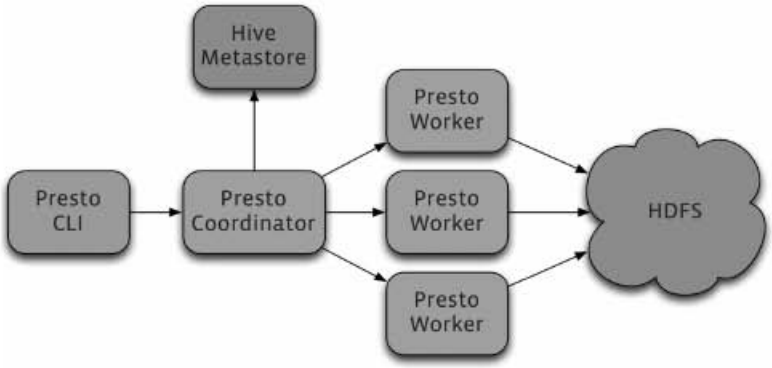


그림 3 프레스토 명령어 처리 흐름

터 플랫폼과의 통합으로 파워피봇(PowerPivot), 파워뷰(Power View) 및 기타 마이크로소프트 BI 도구를 사용하여 하둡 데이터를 분석할 수 있다.

HDInsight를 비롯한 여러 원본의 데이터를 파워 쿼리(Power Query)와 원활하게 결합할 수 있으며, Excel 2013의 3D 매핑 도구인 새로운 파워 맵(Power Map)을 사용하여 데이터를 쉽게 매핑할 수 있다. 앞의 그림 4는 엑셀에서 HDInsight 데이터를 가져오기 위해 메뉴를 선택하는 화면이다. 그림 5는 엑셀로 가져올 데이터를 선택하는 화면이다. 그림 6은 HDInsight 데이터를 엑셀로 가져온 결과 데이터를 조회한 화면이다[22].

HDInsight는 변화하는 조직의 요구 사항에 신속하게 대응할 수 있는 민첩성을 제공한다. 다양한 파워셸(PowerShell) 스크립트 라이브러리를 사용하여 몇 시간 또는 며칠이 아닌 몇 분 안에 하둡 클러스터를 배포하고 프로비저닝할 수 있다. 보다 큰 클러스터가 필요한 경우 데이터 손실 없이 몇 분 안에 클러스터를 삭제하고 더 큰 클러스터를 만들 수 있다.

HDInsight는 엔터프라이즈급 보안, 확장성 및 관리의 용이성을 제공한다. 전용 보안 노드 덕분에 HDInsight는 하둡 클러스터를 안전하게 보호할 수 있다. 윈도우 애저의 탄력적인 확장성도 모두 이용할 수 있다. 또한 파워셸 스크립팅을 위한 광범위한 지원을 통해 하둡 클러스터의 관리 용이성이 단순화되었다. HDInsight는 닷넷, 자바 및 기타 언어를 비롯한 원하는 언어를 사용하는 강력한 프로그래밍 기능을 제공한다. 닷넷 개발자는 LINK to Hive와 함께 전체 통합 언어 쿼리 기능을 사용할 수 있고, 데이터베이스 개발자는 데이터를 쿼리하고 변환하는데 하이브를 사용할 수 있다[23].

5. 결 론

본 고에서는 클라우드와 빅데이터 기술에 대한 정의 및 동향에 대해서 살펴보았다. 또한 클라우드 방식으로 빅데이터 서비스를 제공하고 있는 글로벌 IT 선진

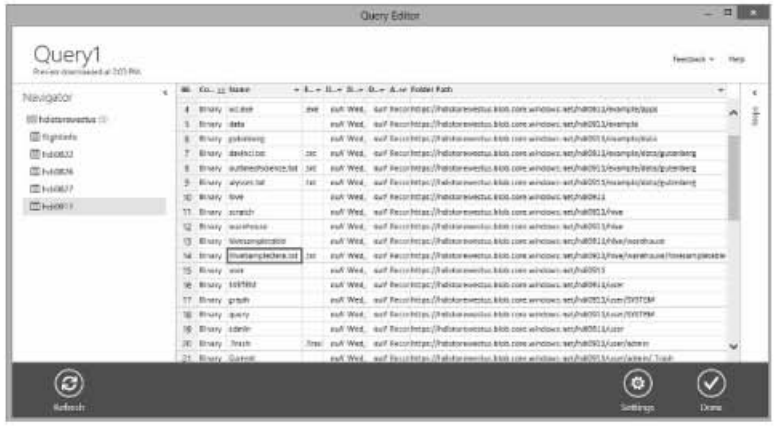
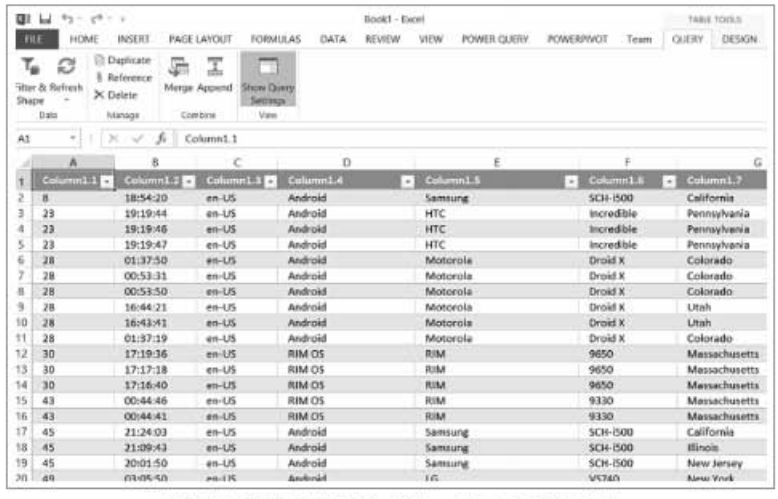


그림 5 엑셀로 가져올 데이터 선택



	A	B	C	D	E	F	G
1	Column1.1	Column1.2	Column1.3	Column1.4	Column1.5	Column1.6	Column1.7
2	8	18:54:20	en-US	Android	Samsung	SCH-I500	California
3	23	19:19:44	en-US	Android	HTC	Incredible	Pennsylvania
4	23	19:19:46	en-US	Android	HTC	Incredible	Pennsylvania
5	23	19:19:47	en-US	Android	HTC	Incredible	Pennsylvania
6	28	01:37:50	en-US	Android	Motorola	Droid X	Colorado
7	28	00:53:31	en-US	Android	Motorola	Droid X	Colorado
8	28	00:53:50	en-US	Android	Motorola	Droid X	Colorado
9	28	16:44:21	en-US	Android	Motorola	Droid X	Utah
10	28	16:43:41	en-US	Android	Motorola	Droid X	Utah
11	28	01:37:19	en-US	Android	Motorola	Droid X	Colorado
12	30	17:19:36	en-US	RIM OS	RIM	9650	Massachusetts
13	30	17:17:18	en-US	RIM OS	RIM	9650	Massachusetts
14	30	17:16:40	en-US	RIM OS	RIM	9650	Massachusetts
15	43	00:44:46	en-US	RIM OS	RIM	9330	Massachusetts
16	43	00:44:41	en-US	RIM OS	RIM	9330	Massachusetts
17	45	21:24:03	en-US	Android	Samsung	SCH-I500	California
18	45	21:09:43	en-US	Android	Samsung	SCH-I500	Illinois
19	45	20:01:50	en-US	Android	Samsung	SCH-I500	New Jersey

그림 6 엑셀로 가져온 HDInsight 조회데이터

기업들의 제품들에 대해서도 소개하였다. 국제분석협회(IIA)는 클라우드에서 언제든 간편하고 빠르게 실행할 수 있는 서비스형 분석 서비스를 채택하는 기업이 늘어날 것으로 예상하고 있다. 컨설팅 업체 캡제미니도 사물인터넷 데이터가 증가하면서 클라우드 기반 서비스형 분석이 인기를 끌 것으로 전망했다. 대체적으로 자체 빅데이터 인프라를 구축하는 것보다 외부 서비스를 선호하는 기업이 증가할 것으로 보고 있다[24].

클라우드 방식의 빅데이터 서비스를 고려할 경우 데이터 보안과 서비스 공급업체가 파산했을 때 발생할 수 있는 문제에 대한 대책을 강구해야 한다. 안정적인 서비스 수준 확보와 데이터 손실을 방지하기 위해 서비스 공급업체와의 긴밀한 협조가 있어야 한다. 클라우드 서비스 사용을 결정한 기업의 경우 퍼블릭 클라우드에 배치할 어플리케이션과 데이터를 분류하고, 중요성에 따라 다른 클라우드나 기업 내부에 복제본을 만들 수 있도록 백업 장치를 두어야 한다[25].

데이터 중심의 증거기반 의사결정을 위해 빅데이터 기술은 앞으로 없어서는 안될 중요한 기술이다. 하지만 시스템이 매우 복잡하고 구축과 운영에 많은 비용이 발생하기 때문에 개별적으로 도입하기에는 많은 제약이 따른다. 따라서 전산 자원을 효율적으로 활용할 수 있고, 에너지를 절감할 수 있으며, 환경보호 측면에서도 장점이 있는 클라우드 서비스 방식으로 가는 것이 최선의 길이다. 다만 클라우드 서비스 방식에서 발생할 수 있는 문제들을 충분히 검토하여 적절한 대응방안을 수립한다는 전제가 선행되어야 한다.

대부분의 사람들은 아직까지 이러한 빅데이터 기술이 쉽게 사용할 수 있는 것이 아닌 매우 어려운 기술로 생각하고 있다. 그 이유는 기업이나 조직이 비즈니스 활용 측면보다는 빅데이터 관련 기술에만 너무 치중하기 때문이다. 빅데이터 기술을 하나의 클라우드 서비스 형태로 제공받을 수 있게 된다면 기술적인 측면보다는 활용적인 측면에 더욱더 집중할 수 있게 된다. 그러면 지금보다 다양한 분야에 빅데이터 기술이 활용되어 많은 성과를 낼 수 있을 것으로 기대한다.

참고문헌

[1] Mell, P. and Grance, T., "The NIST Definition of Cloud Computing", Tech Report National Institute of Standards and Technology, USA, 2009.

[2] 김형준, 조준호, 안성화, 김병준, "클라우드 컴퓨팅 구현 기술", 에이콘, 2011.

[3] 안창원, 황승구, "빅데이터 기술과 주요 이슈", 정보과학회지 제30권, 제6호, pp. 10-17, 2012.06.

[4] 한국정보진흥원, "클라우드 컴퓨팅 활성화 전략", 2009.

[5] 임용재 외, "스마트인터넷 서비스를 위한 클라우드와 빅데이터", 한국방송통신전파진흥원 R&D기획본부, 방송통신 PM Issue Report 제3권, 2013.09.

[6] Selam Abrham Gebregiorgis and Jorn Altmann, "IT Service Platforms: Their Value Creation Model and the Impact of their Level of Openness on their Adoption", TEMEP Discussion Papers 201297, Seoul National University; Technology Management, Economics, and Policy Program (TEMEP), revised Nov 2012.

[7] McKinsey Global Institute, "Big data: The next frontier for innovation, competition, and productivity", 2011.06.

[8] 전승수, "초연결 사회의 빅데이터 생태계 분석과 시사점", KISTEP ISSUE PAPER, 2012.10.

[9] 배동민 외, "빅데이터 동향 및 정책 시사점", 정보통신정책연구원 방송통신정책, 제25권 10호, 통권 555호, pp.37-74, 2013.06.

[10] Google BigQuery, https://developers.google.com/bigquery/docs/overview?hl=ko

[11] 한국인터넷진흥원, "클라우드 기반 빅데이터 분석 플랫폼 '구글 빅쿼리(Big Query)' 적극 행보", KISA 주간인터넷동향 4월 1주, 2012.04.

[12] Interactions Marketing digs into storm data to analyze consumer behavior using Google BigQuery, https://cloud.google.com/files/interactions_marketing.pdf.

[13] Ad Network Sharpens Targeting with Google BigQuery, https://cloud.google.com/files/BooBox.pdf.

[14] Travel Agency Masters Big Data with Google BigQuery, https://cloud.google.com/files/Redbus.pdf.

[15] 클라우드 빅데이터 서비스가 몰려온다, http://www.zdnet.co.kr/news/news_view.asp?artice_id=20131113161620 Amazon Elastic MapReduce, http://aws.amazon.com/ko/elasticmapreduce/

[16] IBM InfoSphere BigInsights, http://www.ndm.net/datawarehouse/IBM/ibm-infosphere-biginsights

[17] InfoSphere BigInsights, http://www-01.ibm.com/software/kr/data/infosphere/biginsights/

[18] 신무기 '빅 인사이트'...IBM, 분석시장 공략에 날개, http://www.ddaily.co.kr/news/article.html?no=69894

[19] InfoSphere BigInsights의 가치, http://public.dhe.ibm.com/software/kr/201112/InfoSphere_BigInsights_oct_2011.pdf.

[20] 페이스북, SQL온하둡 '프레스토' 공개, http://www.zdnet.co.kr/news/news_view.asp?artice_id=2013

1107085722
[21] Presto|Distributed SQL Query Engine for Big Data, http://prestodb.io/
[22] Connect Excel to Windows Azure HDInsight with Power Query, http://www.windowsazure.com/en-us/documentation/articles/hdinsight-connect-excel-power-query/?fb=ko-kr
[23] HDInsight 서비스, http://www.windowsazure.com/ko-kr/services/hdinsight/
[24] 새해 빅데이터 시장 규모 17조원…서비스형 분석 증가, http://www.etnews.com/news/international/2885288_1496.html
[25] 클라우드 최악의 상황: 서비스 업체 파산에 대응하는 방법, http://www.itworld.co.kr/news/85528

약 력

김 상 락

1992 울산대학교 중어중문학과 학사
2010 울산대학교 자동차선박기술대학원 석사
2012 울산대학교 정보통신공학 박사
2010~현재 울산대학교 경영정보학과 강사, 비케이앤씨 대표
2000~2009 (주)아이티스타 연구소장
관심분야 : 서비스수준관리, 인포그래픽스, 빅데이터, 사물인터넷
E-mail : shem0304@gmail.com

강 만 모

1998 울산대학교 전자계산학과 학사
2000 울산대학교 전자계산학과 석사
2011 울산대학교 정보통신공학 박사
2006~2009 울산대학교 객원교수
2009~현재 대광산업 기술연구소 책임연구원, 울산대학교 전기공학부 강사
관심분야 : 전자상거래, 멀티에이전트, 소프트웨어공학, 빅데이터 분석
E-mail : manmoakng@ulsan.ac.kr