

모바일 클라우드에서의 빅데이터 분석 전략

이연희, 박혜숙
한국전자통신연구원

요약

모바일 클라우드는 클라우드 컴퓨팅 방식을 이용하여 IT 자원을 모바일 서비스에 활용하는 기술로서 각광받고 있다. 모바일 클라우드는 특성상 정보가 한 곳으로 집중되고, 다양한 모바일 기기를 통한 접근이 가능하므로 이를 빠르게 처리할 수 있는 기술과 처리한 데이터를 구조화하여 저장하고 분석을 통해 새로운 가치를 부여하는 기술이 필요하다. 또한 모바일 단말의 특성상 모바일 사용자를 대변하는 매개로 표현될 수 있으므로, 모바일 디바이스를 통한 개인화된 서비스와 모바일 서비스에 소셜 기능이 통합되어 실시간 데이터-메시지 스트리밍의 처리와 사용자 데이터에 대한 동기화를 통한 접근성에 향상에 대한 요구가 높아졌다. 이러한 이유로 모바일 클라우드 데이터를 분석하기 위한 방편으로 빅데이터 기술을 활용하는 것이 일반적이다. 본고에서는 최근 모바일 클라우드의 정의와 모바일 클라우드에서의 빅데이터의 특징을 살펴보고, 분석을 위한 전략을 알아본다.

은 상이한 형태의 비정형 데이터가 존재하게 된다. 실시간으로 유입되거나 축적된 데이터를 정제하고 분석, 취합하여 서비스 가능한 형태로 가공함으로써 새로운 가치를 창출할 수 있는 기회를 만들 수 있다.

최근의 빅데이터 기술은 플랫폼에 대한 기술에서 분석과 활용에 대한 기술로 진화하였다. 빅데이터 기술은 단순하게는 대규모 분산처리 및 실시간 처리를 위한 플랫폼 기술인 클라우드 컴퓨팅 기술과 NoSQL 기술을 포함하여, 크롤링(Crawling), 자연어 처리 텍스트 마이닝(Text Mining), 기계학습, 시각화에 이르기까지, 데이터의 더미의 수집부터 그 속에서 결정적인 의미를 뽑아내기 위한 복합적이고 다양한 데이터의 처리기술을 총 망라하고 있다.

빅데이터 플랫폼을 이야기할 때 빠지지 않은 이야기가 하둡이다. 하둡은 원래 MapReduce를 기반으로 배치 처리를 위하여 고안되었다. 초기의 Hadoop 생태계의 아키텍처를 살펴보면, 데이터를 저장하기 위한 HDFS와 분석하기 위한 MapReduce와 Apache Pig, Hive를 비롯한 MapReduce 작성과 실행 인터페이스들, 로그를 수집하기 위한 Apache Flume과 같은 수집기, DB로 부터 데이터를 가져오고 저장하기 위한 Apache Sqoop, 그리고 Job을 제어하기 위한 Apache Oozie 등이 포함되어 있다. 여기에서 제시한 분석 방식들은 모두 데이터를 분석하기 위해서 분석할 데이터를 미리 저장해 두는 것을 기본으로 한다. 따라서 분산 방식이라고 하더라도 근본적으로 적게는 수 초에서 수 분, 수 시간까지 지연이 따를 수 밖에 없다. 또한 결과가 분석되었다고 하더라도 이를 서비스까지 연결하기까지 데

I. 서론

클라우드는 모바일 기기와 연결되어 사용자에게 민첩하고 유연한 서비스를 제공하는 사물인터넷 시대에 없어서는 안될 핵심 엔진으로서의 역할을 수행한다. 모바일 기술의 발전과 다양한 사물인터넷 기기의 등장으로 클라우드에는 많은 양의 데이터가 도처로부터 유입되고 축적된다. 따라서 과거에 비해 더 많

표 1. 모바일 환경에서의 클라우드 컴퓨팅 적용방법[1][2]

	모바일+클라우드	모바일 클라우드
특징	서비스 제공을 위해 서버 대신 클라우드 사용	서비스 제공을 위해 단말 대신 클라우드 사용
클라우드 도입 목적	서비스 인프라 비용 절감 서비스의 빠른 적용과 배포 Pay-as-you-go 모델에 따른 비용절감 안정성 제공	풍부한 사용자 서비스의 제공 다양한 개인의 모바일 및 유선 단말에서 끊임없는 연결성 제공 컨텐츠와 서비스 제공 및 이용관점에 있어 '소유'에서 '접속'으로 유통 및 보안 문제 해결

표 2. 클라우드 서비스 종류에 따른 모바일 애플리케이션 분류[2][6]

분류	정의	사례
mSaaS	모바일 단말 사용자에게 제공하는 소프트웨어 서비스	애플 iCloud, MS Windows Phone, Naver/Daum, 국내 통신사들의 휴대폰 데이터 자동백업 및 데이터 동기화, SNS사진 올리기, 온라인 연락처/ 문자메시지 확인, 휴대폰 위치 찾기 등
mPaaS	모바일 앱/웹 서비스를 생산하기 위한 플랫폼 서비스	Titanium Platform, Titanium Studio, Titanium Mobile, Titanium Desktop으로 구성된 통합 개발 환경(IDE), Parse의 모바일 클라우드 SDK 및 Backend API, Intel의 CloneCloud 등
mIaaS	모바일에 특화된 경량 VM 인프라 제공하는 서비스	VMWare MVP (Mobile Virtualization Platform), 카네기멜론 대학의 Cloudlet 등

이터를 꺼내는 시간에 따른 지연을 피할 수 없다.

최근 Hadoop은 YARN (Yet Another Resource Negotiator) 스케줄러를 통해서 다양한 분석 엔진을 수용함으로써 제2의 도약을 하고 있다. Storm이나 Spark-streaming과 같은 실시간 처리 엔진과, SQL-on-Hadoop 기반의 low-latency 도구들이 소개되고 있다. , Facebook Presto, Apache Spark 그리고, 국산 그루터 Tajo와 같은 좀 더 빠른 분석을 위한 패키지로 전환함으로써 일정 수준의 실시간 처리가 가능하게 되었다.

본 논문에서는 모바일 클라우드에서 빅데이터 분석에 대한 전략을 제시한다. 본론의 1장에서 먼저 모바일 클라우드의 개념과 모바일 애플리케이션의 분류를 살펴보고, 2장에서 빅데이터의 정의와 기술들을 정리한다. 3장에서는 모바일 빅데이터의 특징을 나열하고, 4장에서 분석전략을 논한다.

II. 본론

1. 모바일 클라우드

스마트폰의 보급과 3G와 Wifi 망의 인프라가 발전함에 따라 모바일 단말은 어디서든 네트워크의 접속이 가능하게 되었다. 고객은 PC 뿐만 아니라 스마트TV, 넷북에 이르는 다양한 장비를 보유하고 있으며, 언제 어디서든 자신의 이메일과 문서를 열어볼 수 있게 되었고, 서비스의 원활한 제공을 위하여 구글을 비롯한 애플, 마이크로소프트 등 대표적인 글로벌 IT 기업과 국내에서는 삼성이 모바일 클라우드 플랫폼에 대한 경쟁에 나선지 오래다. 이들은 클라우드 플랫폼을 기반으로 웹을 통한 모바일 애플리케이션 서비스를 제공하는 것을 넘어서, 다양한 서비스를 사용자 중심으로 집약하는 사용자 지향의 모바일 서비스를 중심으로 이동하였다.

일찌기 구글은 자신들의 모바일 사업전략 기자간담회에서 컴퓨팅 (Computing), 연결성 (Connectivity), 클라우드 (Cloud)의 '3C'가 핵심 기술이 될 것이라고 공언했으며, 이는 기존의 PC용 소프트웨어를 모바일 환경으로 옮기는 것이 아니라 처음부터 모바일 환경에 최적화된 형태로 서비스하겠다는 전략으로 여겨졌다. 여기에서의 클라우드는 단순히 클라우드와 모바일 기술이 합쳐진 것이 아니라, 사용자와 사용자가 이용하고 생성하는 데이터, 콘텐츠 및 서비스에 이동성을 제공하기 위한 모바일 서비스 자체를 클라우드 플랫폼화 해서 제공하는 기술까지를 일컫는다(표 1). 즉, 낮은 컴퓨팅 성능을 가진 모바일 디바이스를 통해 웹을 통하여 다양한 서비스를 제공받을 수 있도록 하는 클라우드를 기반으로 한 서비스와 콘텐츠의 제공 기술까지 포함한 형태를 말한다. 리소스가 제한적인 모바일 디바이스에서 데이터나 애플리케이션을 직접 탑재하는 것은, 적은 리소스를 가진 모바일 장비의 성능 상의 제약을 감안하여 하나의 단말로서의 역할만 수행하도록 제한하고, 모든 콘텐츠와 서비스의 생성은 중앙 클라우드에서 담당하도록 하는 전략을 추구한다.

한편, 모바일 클라우드 서비스는 클라우드 서비스 구조에 따라, 모바일 Infrastructure-as-a-Service (mIaaS), 모바일 Platform-as-a-Service (mPaaS), 모바일 Software-as-a-Service (mSaaS)로 분류하며, 각 서비스 형태에 따라 서비스 제공자가 제공하는 서비스의 정도가 달라진다. 그 중 mSaaS는 스토리지 서비스, 단말간 콘텐츠 동기화, 모바일 오피스와 같은 사용자에게 제공하는 서비스로서 통신사나 포털업체에서 제공하는 실시간 백업과 단말간 동기화와 Google Docs와 같은 서비스가 대표적인 서비스에 속한다. 반면, mPaaS나 mIaaS는 중간 공급자가 사용하는 서비스라고 할 수 있으며, 대표적인 mPaaS사례로는 Titanium platform이나 CloneCloud가, mIaaS로는 VMware의 Mobile Virtual Platform (MVP)나 카네기멜론 대학의 Cloudlet이 있다(표 2).

2. 빅 데이터 기술

빅데이터는 비즈니스의 성패를 좌우하는 핵심기술로 부상하였다. 최근의 빅데이터는 단순히 폭발적인 양의 정형/비정형 데이터 자체를 칭하는 것이 아니라, 이를 분석하여 활용하는 기술까지 포함하여 그 개념을 확장한다. 빅데이터를 정의할 때 규모 (Volume), 다양성 (Variety), 적시성 (Velocity)의 3V를 언급한다. 최근에는 이에 복잡성 (Complexity)를 더해 구조화되지 않는 데이터, 중복성 문제 등 외 관리대상과 처리의 복잡성이 심화됨을 의미한다. 더 넓은 범위의 데이터로 확대하여 다양한 외부 데이터의 활용을 통한 큰 가치를 창출하는 것의 중요성을 강조하고 있다[4].

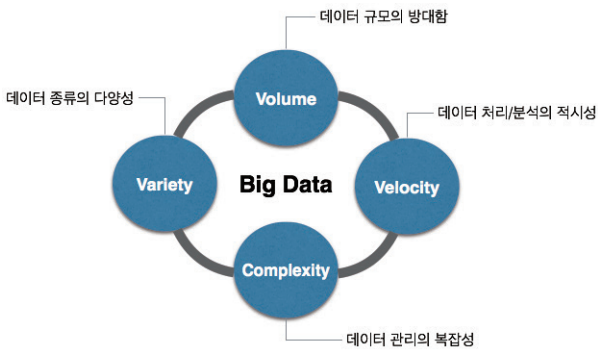


그림 1. 빅데이터의 특성

가장 근간이 되는 데이터를 놓고 보면, 스마트폰과 같은 모바일 기기와 서비스의 증가는 데이터의 양과 이종의 장비로부터의 데이터의 다양성으로 인한 데이터 간의 관계의 복잡성에 기여한 바가 크다. 복잡하고 다양한 데이터로 부터 고객의 구매

패턴을 분석하여 필요한 상품을 추천하고 고객의 상담내용을 분석하여 새로운 상품을 기획하거나 서비스를 개선하기 위한 의사결정 도구로 활용하는 것은 이미 많은 기업들의 사례를 통해 알려져 있다. 데이터를 기반으로 한 비즈니스의 성공을 위해서는 단순한 플랫폼 기술이나 데이터를 획득하여 단순한 정보로 전환하는 수준에서 벗어나, 비즈니스에 활용할 수 있는 수준의 지식을 창출할 수 있는 데이터 분석 기술이 필요하다. 이를 위해 사용되어지는 것이 바로 빅데이터 기술이다. 대규모 분산 처리와 실시간 처리를 위한 클라우드 컴퓨팅 기술과 NoSQL 기술 뿐만 아니라, 크롤링 (Crawling), 자연어 처리 텍스트 마이닝 (Text Mining), 기계학습, 시각화에 이르기까지, 데이터의 더미의 수집부터 그 속에서 결정적인 의미를 뽑아내기 위한 복합적이고 다양한 데이터의 처리기술이 필요하다. 대표적인 빅데이터 분석 기술을 요약하면 <표 3>과 같다.

빅데이터 기술 중에서도 그 근간은 저가의 컴퓨팅 능력을 가진 개개의 컴퓨터를 네트워크로 연결하여 수직적인 성능확장이 아닌 수평적 성능확장을 가능하게 하는 대규모 분산처리 기술이라고 할 수 있다. 그 중에서 가장 많이 회자되는 것으로 하둡 (Hadoop)을 빼 놓을 수 없다. 최근에는 가상화를 기반으로 한 클라우드 컴퓨팅 기술과 접목되면서 끊임없는 성능확장을 달성할 수 있게 되었다 [5]. 아마존은 Amazon Elastic MapReduce (EMR)을 웹서비스 형태로 제공하고 있으며 클라우드태라 SQL 분산쿼리 엔진인 임팔라 (Impala)도 함께 제공하고 있다. 구글도 클라우드 스토리지 키벡터를 출시하여 클라우드상의 가상화된 인프라에서 맵리듀스 (MapReduce) 를 위한 데이터 입출력을 가능하도록 제공하고 있다 [6]. 하둡은 분산 소프트웨어 모델인 맵리듀스와 분산파일시스템인 HDFS (Ha-

표 3. 대표적인 빅데이터 기술들[3]

항목	내용
정보수집 (Crawling)	웹으로 부터의 데이터를 수집하기 위한 기술
클라우드 컴퓨팅, NoSQL	대용량의 데이터를 저장, 관리, 운영하기 위한 클라우드 기반의 분산 컴퓨팅 및 질의 기술
정보검색 (IR)	대규모의 데이터로부터 주제와 연관된 데이터를 찾기위한 기술
자연어 처리 (NLP)	형태소 분석, 구문 분석, 개체명 인식과 같이 인간이 사용하는 언어를 컴퓨터를 이용하여 처리하는 기술
기계학습 (Machine Learning)	데이터를 학습하고 모델을 추출하여 대용량 데이터를 자동 분석하는 귀납적 추론 방법
텍스트 마이닝 (Text Mining)	대량의 텍스트의 덩어리로부터 의미있는 정보를 추출, 분석하는 기술, 기계학습을 이용한 통계적 방법과 규칙기반 방법이 존재
시맨틱 (Semantic)	시맨틱 데이터의 자동 추출, 네트워크 생성, 지식 베이스 구축, 온톨로지의 활용, 논리 및 통계적 추론
통계 (Statistics)	통계적 의미를 찾기 위한 기술
시각화 (Visualization)	분석 결과로부터의 직관적 의미 습득을 위한 표현 기술

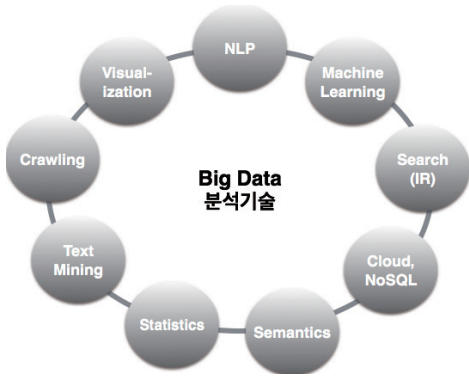


그림 2. 빅데이터 분석 기술

doop Distributed File System) 기술을 중심으로 발생되어, 하둡 2세대라고 불리는 하둡 YARN (Yet Another Resource Negotiator)을 통해 다양한 프로그래밍 모델을 접목할 수 있도록 확장되었다.

3. 모바일 빅데이터의 특징

가. 모바일 클라우드 데이터

모바일 서비스가 확대됨에 따라 유통되는 모바일 트래픽이 폭증하고 있다. 따라서 이를 위한 요소 기술도 급속히 발전하는 추세이다. 또한 모바일 단말에서 실시간으로 획득되는 센싱 정보들은 밀려드는 속도가 점차 가속화되는 추세이다. 데이터의 질적으로도 현재 약 80%가 비정형 데이터가 차지하고 가운데, 미래에는 이 비율이 최대 90% 까지 늘어날 것으로 보고 있다. 고객의 시대로 진입함에 따라, 개인 기반의 원활한 서비스를 위하여 데이터들은 정제되고 개인을 중심으로 재배열되어 서비스에 활용될 수 있는 모습으로 전환되어 의사결정에 활용할 수 있는 형태로 보존되어야 한다.

나. 모바일 빅데이터 분석의 요구사항

모바일 디바이스를 통한 개인화된 서비스와 모바일 서비스에 소셜 기능이 통합되어 실시간 데이터-메시지 스트리밍의 처리와 사용자 데이터에 대한 동기화를 통한 접근성에 대한 요구가 높아졌다. 모바일 애플리케이션의 발전과 효과적이고 안정적인 서비스를 위해서 새로운 서비스를 창출하기 위한 기술과 안정적인 서비스 질을 위한 기술, 안전한 서비스를 위한 기술이 필요하다. 모바일 애플리케이션의 새로운 가치를 창출하기 위해서는 사용자 중심으로 데이터를 확립하고, 폭발적인 사용자 데이터의 증가에 대한 관리와 사용자의 데이터의 보안성 확보, 그리고 사용자 데이터와 콘텐츠의 서비스를 위한 충분한 속도가 제공이 되어야 한다. 모바일 빅데이터의 특징에 따른 분석을 위한 요구사항들은 요약하면 다음과 같다.

- 애플리케이션들이 쏟아내는 데이터 양의 폭발적인 증가에도 관리가 가능해야 한다.
- 모바일 유저와 단말의 증가에도 안정적인 서비스의 성능을 보장해야 한다.
- 사용자의 모바일 클라우드 애플리케이션 의 사용하기에 문제가 없는 수준의 보안을 제공해야 한다.

클라우드 컴퓨팅과 모바일 단말의 조합으로 이동 중에도 어려움없이 자신의 생활에 필수적인 정보와 애플리케이션을 언제 어디서나 액세스할 수 있는 서비스가 일반화됨에 따라 서비스 질이나 보안 등 다양한 문제에 직면하게 되었다. 또한 새로운 서비스를 창출할 수 있는 기회도 함께 맞게 되었다. 애플리케이션 개발자는 다양한 문제를 고려하여 새로운 서비스를 위한 소프트웨어를 개발하여야 한다.

4. 빅데이터 분석 전략

모바일 기기는 이전의 PC 환경과는 달리 개개인을 대신하는 대체물이라고 할 수 있다. 모바일 단말을 통해 사용자의 위치, 장치 정보, 행동 패턴 등 개인에 관한 모든 정보를 획득하는 것이 상대적으로 용이하다. 따라서 모바일 데이터를 분석하여 개인의 특성을 파악하고 개인별로 구체적인 대응 전략을 수립하여 일대일 타겟 마케팅과 같은 새로운 부가가치를 창출에 활용하는 것이 가능하다. 이를 실현하기 위한 기술로서 다양한 스스로부터 유입되는 데이터를 즉각적으로 처리하기 위한 실시간 스트리밍 처리 기술과 처리된 데이터를 입맛에 맞는 형태로 만들어 주기 위한 세그먼테이션 과정, 분석에 따른 실시간 질의 기술들이 요구된다.

가. 사용자 식별 정보의 획득과 익명화

개인화된 서비스를 제공하기 위해서는 개인을 식별할 수 있는 정보가 필수적이다. 모바일 환경에서는 이름, 주민등록번호와 같은 개인에 할당된 신상정보 외에 기기마다 할당된 MAC 주소나 전화번호 등이 개인의 식별정보로 활용될 수 있다. 그 외에도 IP 주소도 어느정도의 개인 식별 기능을 가질 수 있으나, 모바일 환경의 특성상 완벽하진 않다. 이런 정보들은 사용자 중심의 서비스를 제공하기 위한 식별키로 사용하는 것이 용이하다는 점이 장점인 반면 개인정보보호에 대한 논란의 여지를 안고 있다. 이러한 논란에서 자유롭기 위해서는 가능한 필요한 경우에 한해서만 정보를 수집하도록 해야 한다. 부득이하게 수집하여야 하는 식별 정보는 반드시 보안이 부가된 채널을 통해서 전송하여야 하며, 이를 보존할 때도 해쉬와 같은 익명화 알고리즘을 통해서 가공하여 보존해야 한다.

나. 분석을 위한 효과적인 세그먼테이션

획득한 데이터는 사용자 분석에 용이한 형태로 미리 가공하여 뒤으로써 추후에 있을 분석과정에서의 복잡성을 덜어줄 수 있다. 데이터를 어떻게 구조화 할 것인지는 정책에 따라 상이할 수 있으나, 시간과 데이터의 유입경로, 사용자 유형에 따라 분류하는 것이 일반적이다.

- 시간에 따른 세그먼테이션

특정 기간 동안의 광고들의 성과의 성과를 분석하는데 유용하며 적시마케팅 정책을 수립하는데 활용할 수 있다.

- 유입경로에 따른 세그먼트

사용자가 광고에 어떤 소스로부터 유입이 되었는지 한눈에 파악하는 것을 가능하게 한다. 따라서 온라인 광고의 성과를 파악하여 의사결정에 활용할 수 있다.

- 사용자 유형별 세그먼트

나이, 성별과 같은 사용자의 유형을 인지하여 특정 그룹이 광고에 어떻게 반응하고 어떤 행동특성을 갖는지 파악하는데 도움이 된다. 이를 통해 타겟 고객층을 공략하기 위한 전략을 수립을 하는데 의사결정 수단으로 유용하게 사용할 수 있다.

다. 실시간 스트리밍 처리 기술

사물인터넷 시대의 도래는 기업이 실시간으로 수집하고 관리하며 분석해야 하는 데이터의 양, 다양성, 속도를 증가시킨다. 또한 빅데이터를 지향하는 주요 동력이자 정보 분석을 활용해 비즈니스 성과를 최적화하고, 고객 경험을 향상시키며 경쟁력 있는 가치를 이끌어 내는 원동력이 된다. 기존의 방식은 엄청난 양의 데이터를 실시간으로 처리하는데 한계를 보이기 시작한지 오래이므로 일찌기 빅데이터 스트리밍 처리와 같은 기술이 도입되었다. 이는 끊임없이 들어오는 데이터를 유입 시점부터 분석해 미리 원하는 데이터 뷰로 만드는 방식으로, 데이터의 세그먼테이션 정책에 따라 다양한 형태로 구조화하여 태그하여 저장하는 것이 가능하다. 최근 실시간 분산 쿼리나 스트리밍 처리 기법에 대해 많은 연구가 이루어지고 있다.

오픈소스 스트림 처리 엔진으로는 Twitter의 Storm과 Apache Spark Streaming이 대표적이다. Storm은 분산처리를 위한 노드들의 topology를 구성하여 일련의 데이터가 프로세싱 노드를 순회하며 최종적인 결과를 얻어내는 방식이다. 각 노드는 데이터의 조건에 따른 필터링, 그루핑, 정렬, 집계, 통합 등의 작업을 수행하며 혈통 추적 (lineage tracking) 방식을 통해 처리 과정 중에 특정 노드가 실패하면 이를 이전 노드로 거슬러 올라가 다시 처리하는 방식으로 오류를 복구한다. Storm은 정형화된 이벤트를 추출하고 만드는 전처리기로 사용하여 대용량 이벤트 처리를 효과적으로 달성할 수 있다.

한편으로, 원래 태생부터 실시간 처리를 위해 생겨난 엔진들이 많이 있다. 일반적인 CEP (Complex Event Processing) 엔진들이 이에 해당한다. CEP 엔진은 잘 정리된 대용량의 이벤트에서 실시간 통계, 계산, 패턴 매칭 등의 처리를 효과적으로 수행한다. 대표적인 오픈소스 CEP 엔진인 Esper는 EPL이라고 하는 SQL 친화적인 문장을 통해 이벤트 처리를 정의할 수 있으며, 다양한 형태의 결과를 계산해 낼 수 있다.

이와는 조금 다르게 데이터를 로그의 관점에서 실시간으로 수집하고 인덱스를 생성하여 사용자에게 뷰를 제공하는 솔루션들로 Splunk와 Elasticsearch의 ELK(Elasticsearch, Logstash, Kibana) Stack이 있다. Splunk는 사용자가 원하는 데이터를 그 자리에서 즉시 분석하고 보고하는 기능을 제공하는데 사용자는 쿼리를 이용하여 검색이 가능하다. 하지만 사용 솔루션으로서 소스가 오픈되어 있지 않고, 종량제 형태의 과금방식을 적용하고 있다. ELK Stack은 Elasticsearch, Logstash, Kibana의 약자로서 Elasticsearch는 Apache의 Lucene을 바탕으로 개발한 실시간 분산 검색 엔진이며, Logstash는 각종 로그를 가져와 JSON형태로 만들어 Elasticsearch로 전송하고 Kibana는 Elasticsearch에 저장된 Data를 사용자에게 Dashboard 형태로 보여주는 오픈소스 실시간 처리 솔루션이다.

표 5. 대표적인 스트리밍 처리 엔진들

엔진	특징
Esper	Source CEP 엔진 (Enterprise Edition 제공) 어느 정도 정형화된 데이터에 대한 패턴 매칭 기반의 이벤트 처리 Java와 .NET 사용이 가능하며 SQL과 흡사한 Language도 제공함
Apache S4	Yahoo 에서 공개한 오픈소스 기반의 실시간 프로세싱 프레임워크, 아파치 인큐베이터 프로젝트 실시간 개인화 검색광고를 구현하기 위해서 개발
Backtype Storm	트위터의 오픈소스 프로젝트, 분산 스트림 컴퓨팅 토폴로지의 구성에 따라 다양한 형태의 처리 가능
Hstreaming	Pig 의 Load / Store 에 지정하는 데이터 소스와 싱크를 HDFS01 외에 실시간 Data Streaming을 받아들일 수 있도록 I/O 를 확장

라. 실시간 분산쿼리

최근에도 데이터베이스는 데이터의 분석을 위해 가장 많이 사용하는 도구이다. 데이터베이스를 도구로 하여 데이터베이스 내의 테이블 간의 관계를 정의하여 서로다른 튜플간의 연관관계를 유지한다. 모바일 사용의 행동 패턴을 분석하고 SNS 정보와 같은 다른 정보와의 연계분석 등을 분석하는 것이 가능하다. 하지만 전통적인 관계형 데이터베이스와 데이터 관리 기법

은 구조화된 트랜잭션 (Transaction) 데이터를 위해 필요하겠지만, 구조화되지 않은 데이터를 처리하기 위한 기술을 포함하지 않고 있으며 처리할 수 있는 능력도 한계에 달했다. 따라서 빅데이터 기술과 원활히 공존하고 통합되어야 한다.

빅데이터 실시간 분산 쿼리는 클러스터를 구성하는 각 노드가 분산하여 쿼리를 처리하는데, 처리되는 데이터를 작은 단위로 나누어 가능한 한 분산된 메모리 내에서 처리를 하는 기법을 통해 쿼리의 응답시간을 실시간 수준으로 높이는 방식이다. Dremel의 논문을 기반으로 한 Cloudera의 Impala와 Apache Tez, 그리고 최근 공개된 Facebook의 Presto가 이 방식에 속하며 흔히 SQL on Hadoop 이라고 불리는 처리 엔진들로서 <표 6>과 같다.

표 6. 대표적인 실시간 분산쿼리 처리 엔진들

엔진	특징
Dremel	2011에 공개된 구글 빅쿼리 서비스에 사용된 분석기술 위키피디아를 수초만에 조회
Impala	Dremel을 참조하여 구성된 고유의 분산 쿼리 엔진 Impala Daemon, Statestore, 실행 모델(execution model)을 포함 하둡 기반으로 동작, Hadoop Core Component, HDFS, Hive, 그리고 HBase까지 모두 지원
Apache Tajo	하둡 기반의 대용량 데이터웨어 하우스 시스템 SQL 표준 지원
Spark/ Shark	데이터를 읽어들여 메모리에서 처리하여 쿼리성능 향상 데이터 마이닝을 위해 UC 버클리 AMP 랩에서 개발

III. 결론

본 고에서는 클라우드 컴퓨팅을 방식을 이용하여 IT 자원을 모바일 서비스에 활용하는 기술로 각광받고 있는 모바일 클라우드 서비스에 대해 알아보고 클라우드에서의 빅데이터의 특징과 이를 위한 분석을 위한 전략을 알아보았다. 모바일 클라우드는 클라우드 컴퓨팅 방식을 이용하여 IT 자원을 모바일 서비스에 활용하는 기술로서 각광받고 있으며, 모바일 클라우드는 특성상 정보가 한 곳으로 집중되고, 다양한 모바일 기기를 통한 접근이 가능하므로 이를 빠르게 처리할 수 있는 기술과 처리한 데이터를 구조화하여 저장하고 분석을 통해 새로운 가치를 부여할 수 있는 빅데이터 분석 기술과 전략이 반드시 필요하다.

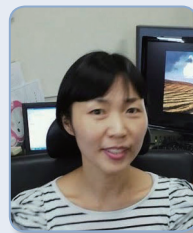
Acknowledgement

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신·방송 연구개발사업의 일환으로 수행하였음. [10044556, 고품질 융합서비스를 위한 국방 기반체계 연동 기술 개발]

참고문헌

- [1] “Mobile Cloud 동향 : 모바일 클라우드의 개요,” http://kwonhanseung.blogspot.kr/2012/10/blog-post_7981.html
- [2] “개방형 모바일 클라우드 기술 및 시장 동향,” KCA 방송통신기술 이슈&전망, 제 20호, 2013
- [3] “빅데이터 분석 개요,” <http://www.saltlux.com/bigdata-analysis-overview/>
- [4] “새로운 미래를 여는 빅데이터,” NIA 정보화진흥원, 2013. 2
- [5] Apache projects, Hadoop, <http://Hadoop.apache.org>, 2009.
- [6] J. Dean and S. Ghemawat, “MapReduce: simplified data processing on large clusters,” Communications of ACM, vol. 51, no. 1, 2008, pp. 107-113.

약 력



이 연 희

2010년 충남대학교 공학학사
 2012년 충남대학교 공학석사
 2015년 충남대학교 공학박사
 2014년~현재 한국전자통신연구원
 클라우드네트워크킹연구실 선임연구원
 관심분야: 빅데이터, 인터넷 트래픽 측정, 네트워크 모니터링, SDN, NFV



박 혜 속

1992년 경성대학교 이학학사
 1994년 부산대학교 이학석사
 2005년 충남대학교 이학박사
 1994년~현재 한국전자통신연구원
 클라우드네트워크킹연구실장
 관심분야: 클라우드네트워킹, 고신뢰네트워킹