# 04

# 데이터베이스 기술 동향

CHAPTER 1

# 빅데이터 기술

## 제1절 빅데이터 저장 및 처리

### 1. 맵리듀스 및 하둡

오늘날 우리는 전 세계에서 생성되는 데이터의 양이 2제타바이트(2012년 IDC 기준)가 넘어서는 빅데이터 시대에 살고 있다. 빅데이터로부터 의미있는 소량의 정보와 직관(insight)을 얻는 빅데이터 분석 기술이 매우 중요한 시대가 되었으며, 이러한 빅데이터 분석을 효율적으로 수행할 수 있는 시스템 기술 또한 점점 더 중요해지고 있다. 2004년 구글이 맵리듀스(MapReduce) 프레임워크를 발표한 이후 미국 등 선진국의 IT 산업계를 중심으로 빅데이터를 저장하고 처리할 수 있는 시스템 기술이 앞다투어 개발되었다.

하둡(Hadoop)은 아파치(Apache)가 오픈 소프트웨어로 구현한 맵리듀스 프레임워크이다. 2007년에 최초 공개된 이후 많은 발전을 거듭하여 2013년 가을에 하둡 2.2가 정식으로 발표되었다. 기본적으로 하둡 프레임워크는 많은 수의 범용 컴퓨터들을 비공유(shared-nothing) 방식으로 구축한 클러스터 상에서 데이터를 HDFS(Hadoop Distributed File System)에 분산 저장한 후, 맵(Map)과 리듀스(Reduce)라는 사용자 정의 함수의 실행을 통해 데이터를 배치(batch) 방식으로 처리한다. 하둡 프레임워크는 스케일 아웃(scale out) 방식으로 시스템의 확장이 가능하고 프로그래밍이 비교적 간단하며 내고장성(fault tolerance)을 갖추고 있어서 짧은 기간 안에 빅데이터 분석을 위한 사실상의(de facto) 표준 기술로 자리 잡았다. 그러나 지난 40여 년 동안 발전되어온 관계형 DBMS 기술과 달리, 태동한 지 10년이 채 되지 않은 맵리듀스 또는 하둡 시스템 기술은 기능 및 성능적인 측면에서 아직 부족한 점들이 많이 있으며 미국 등 선진국을 중심으로 기업, 연구소, 대학교에서 이들 기능 및 성능 문제를 개선하기 위한 연구를 매우 활발하게 수행하고 있는 상황이다. 오늘날 대부분의 빅데이터 분석 솔루션들은 이들 빅데이터 시스템 기술들을 기반으로 맞춤형으로 개발되고 있으므

로 주요 시스템 기술들의 특징 및 개발 동향을 이해하는 것은 매우 중요하다.

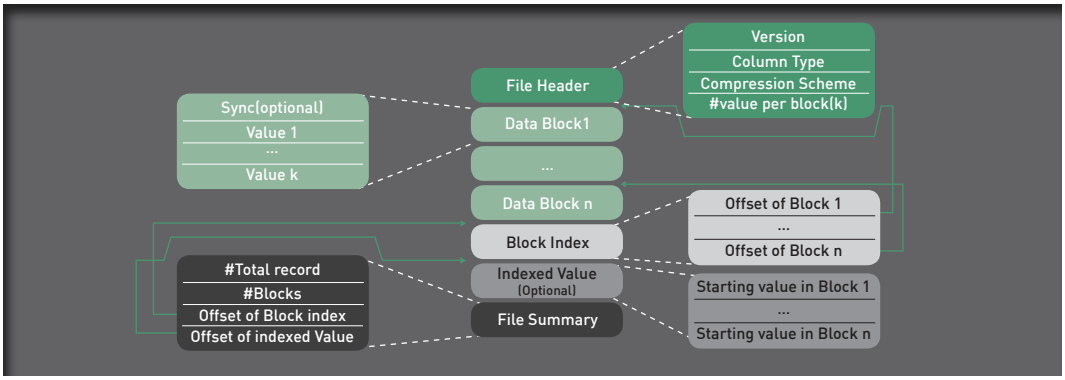
## 2. 빅데이터 저장 기술

### 가. 컬럼 지향 저장 기술

컬럼(column) 지향 저장 기술은 관계형 DB를 저장할 때 레코드 단위로 저장하는 것이 아니라 컬럼 단위로 저장하는 기술로서, 로우(row) 지향 저장 방식에 비해 데이터 압축률이 높고 질의 처리에 필요한 컬럼 데이터만 읽을 수 있기 때문에 OLAP 및 BI(Business Intelligence) 애플리케이션에서 각광받고 있으며, 빅데이터 저장 기술에도 적극적으로 적용되고 있다. 기본적으로 하둡의 HDFS에서 파일들은 64MB의 크기를 가지는 블록들로 구성되는데, 관계형 데이터를 HDFS의 파일로 저장할 때 컬럼 지향 저장 방식으로 저장함으로써 빅데이터 처리 속도를 크게 향상시킬 수 있다. 대표적인 기술들에는 CFile, RCFile, CIF가 있다.

CFile은 데이터를 수직 그룹(vertical group)이라 불리는 컬럼 그룹들로 분할하고, 각 그룹을 (그림 4-1-1)과 같은 포맷으로 저장한다. 각 그룹은 로우 지향 방식으로 다시 분할되어 데이터 블록 단위로 저장되며, 질의 처리 시 필요없는 로우 그룹을 쉽게 건너뛸 수 있도록 각 데이터 블록은 k개의 고정 개수의 컬럼 값들을 가지도록 구성된다. 하나의 레코드를 구성하는 컬럼 값들이 물리적으로 서로 다른 블록에 저장되므로 통신 비용이 발생할 수 있다는 단점이 있다.

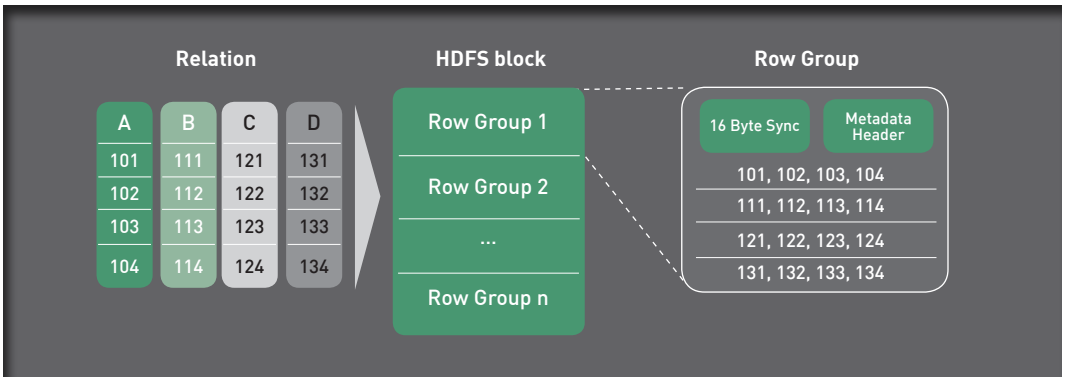
(그림 4-1-1) CFile 포맷 구조




RCFile은 (그림 4-1-2)와 같이 데이터를 논리적인 로우 그룹들로 나누고 HDFS 블록 안에 여러 개의 로우 그룹들을 저장하되, 각 로우 그룹을 컬럼 지향 방식으로 압축해서 저장한다. 하나의 레코드를 구성하는 컬럼 값들이 물리적으로 같은 블록에 저장되어 있어 통신비용이 발생하지 않고, 각 로우 그룹에서 질의에 필

요한 컬럼만 접근할 수 있는 장점이 있다. 그러나 CFile과 RCFile 모두 HDFS 블록구조를 변경해야 하는 단점이 있다.

(그림 4-1-2) RCFile의 포맷 구조




CIF는 HDFS 블록 구조를 변경하지 않고 컬럼 지향 방식으로 데이터를 저장한다. 데이터를 HDFS 블록 크기씩 수평적으로 자른 후 각 파티션을 스플릿(split) 디렉터리라는 개별 디렉터리에 저장하되, 디렉터리 안에서 각 컬럼을 개별 파일로 저장한다. HDFS 블록 구조를 변경하지 않으므로 유연성이 높으나 컬럼 그룹을 저장하는 단위가 RCFile에 비해 커서 처리할 데이터 크기가 작을 때에는 병렬 처리가 충분히 되지 않아 성능이 더 나쁠 수 있다.

## 나. 색인 기술

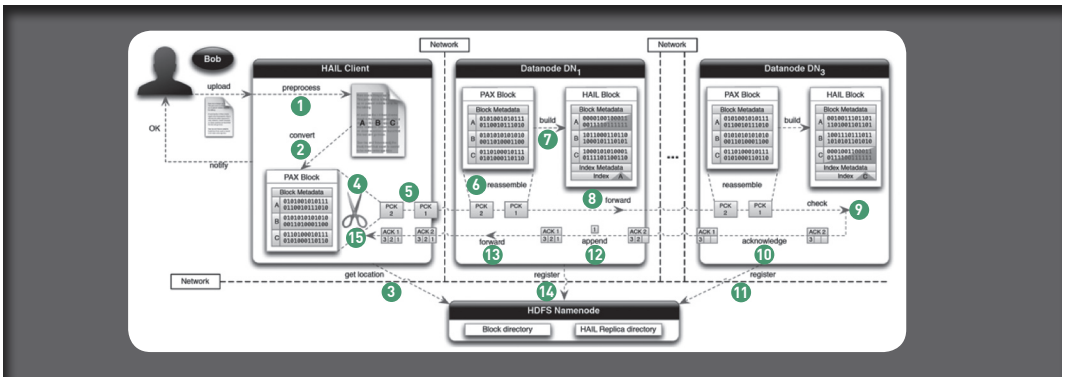
하둡은 파일이 HDFS의 블록(또는 스플릿) 단위로 분산 저장되므로, 네트워크 통신비용을 없애기 위해 색인을 데이터와 함께 단위로 저장하는 기술들이 개발되었다. 대표적으로는 Trojan과 HAIL이 있다.

Trojan 색인 기술은 스플릿이라 불리는 단위로 데이터와 해당 데이터에 대한 색인을 함께 저장한다. 스플릿은 풋터(footer), 헤더, Trojan 색인, 데이터 네 부분으로 구성되는데, 풋터에는 스플릿 경계를 구별하기 위한 정보가 저장되고, 헤더에는 색인 크기 및 데이터 크기 등이 저장된다. 색인은 CSS 트리 형태를 따르며, 트리의 리프(leaf) 노드는 해당 데이터 요소들을 가리킨다. 색인 구축 시간이 오래 걸리는 단점이 있다.

HAIL은 종래의 Trojan 기술의 색인 구축 시간이 너무 오래 걸리는 문제점과 오직 하나의 컬럼에 대한 색인만 구축할 수 있다는 한계를 해결한 기술이다. HAIL은 HDFS의 데이터 로딩 알고리즘을 수정하여 데이터를 하둡 HDFS으로 최초 로딩하는 동안 곧바로 색인을 구축한다.

또한 HDFS 블록이 복제 계수(replication factor)에 따라 복제될 때, 복제 블록마다 서로 다른 색인을 구축한다. 예를 들어, 복제 계수가 3인 경우 서로 다른 3개의 컬럼에 대해 색인을 구축할 수 있다. (그림 4-1-3)은 HAIL의 색인 구축 과정을 보여준다.

(그림 4-1-3) HAIL 색인 구축 과정




※ 출처 : Dittrich, J. et al., Only aggressive elephants are fast elephants, PVLDB 2012

### 다. 데이터 배치 기술

데이터가 분산 저장되는 하둡에서는 조인(join) 질의 처리 시 셔플(shuffle) 단계에서의 과도한 통신비용으로 인해 성능이 크게 저하될 수 있다. 이러한 문제를 해결하기 위해 서로 관련 있는 파일들을 물리적으로 같은 데이터 노드에 저장하는 데이터 배치 기술이 필요하며, 대표적인 기술로는 CoHadoop이 있다. CoHadoop은 HDFS의 데이터 배치 알고리즘을 수정하여 서로 관련 있는 파일들을 복제 블록 수준에서 같은 데이터 노드에 저장하고, 그 정보를 네임 노드(name node)의 로케이터(locator) 테이블에 기록한다. 예를 들어 파일 A가 블록 A1, A2로, 파일 B가 블록 B1, B2, B3로 구성되고, 복제 계수가 3일 때, 블록 집합 {A1, A2, B1, B2, B3}은 3개의 데이터 노드에 모두 함께 배치된다.

## 3. 빅데이터 처리 기술

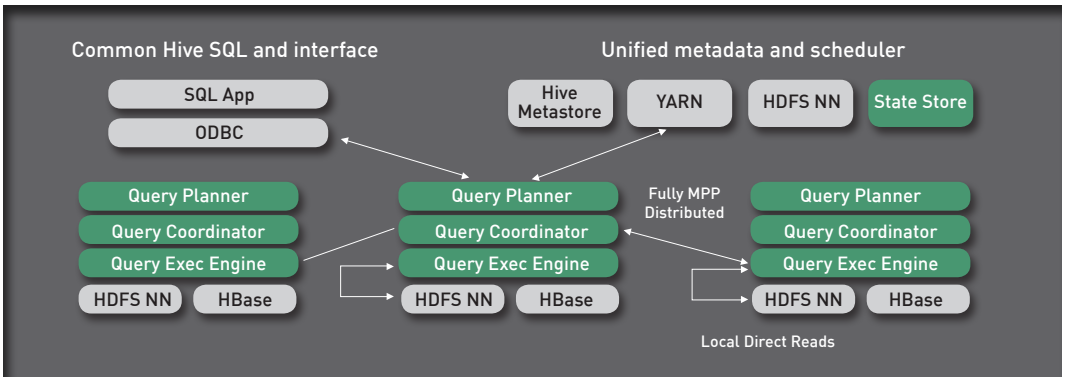
### 가. 하이브리드 시스템 기술

하둡 또는 HDFS를 기존의 관계형 DBMS 기술과 결합함으로써 하둡이 가진 성능 문제를 해결하려는 하이브리드 시스템 기술 개발이 지속적으로 이루어지고 있다. 대표적으로 HadoopDB와 Impala 기술이 있다.

HadoopDB는 각 하둡 슬레이브 노드에 관계형 DBMS를 설치한 후, 데이터 저장 및 질의 처리를 최대한 DBMS에서 수행토록 하고 하둡 프레임워크는 단지 작업 코디네이터(task coordinator) 역할을 수행토록 하는 기술이다. 사용자로부터 받은 질의를 하둡 프레임워크를 이용하여 노드들로 분배하고 각 노드의 DBMS는 질의를 수행한 결과를 키값 형태로 반환한다. 기존의 DB시스템이 가지고 있는 색인 기법 및 우수한 조인 성능 같은 장점들과 하둡 프레임워크가 가지고 있는 내고장성 및 확장 같은 장점들을 모두 얻을 수 있다.

클라우데라(Cloudera)의 Impala는 저장 시스템으로 HDFS 또는 HBase를 사용하지만 질의 처리를 위한 시스템으로 하둡 프레임워크를 사용하지 않고 자체 개발한 질의 처리 엔진을 사용하는 오픈소스 기술이다. 질의 처리 엔진은 기본적으로 종래의 병렬 RDBMS와 유사한 특징들을 가지고 있고 세부적으로 구글의 Dremel 논문에서 공개된 기술적 특징들을 가지고 있다. 하둡의 맵리듀스 함수를 통한 배치 작업 방식을 탈피하고 SELECT, JOIN, 집계 함수들로 구성된 ad hoc 질의를 병렬 DBMS 방식으로 처리함으로써 종래의 하둡에 비해 매우 빠른 속도로 빅데이터 분석 질의를 처리할 수 있다. (그림 4-1-4)는 Impala의 시스템 아키텍처를 보여준다.

**(그림 4-1-4) Cloudera Impala 시스템 아키텍처**




※ 출처 : http://www.cloudera.com

## 나. 성능 최적화 기술

하둡의 느린 성능을 향상시키기 위한 많은 종류의 최적화 기술들이 꾸준히 연구 및 개발되고 있다. 대표적인 기술에는 What-if, SkewTune, YSmart, HaLoop, Restore, MRShare가 있다.

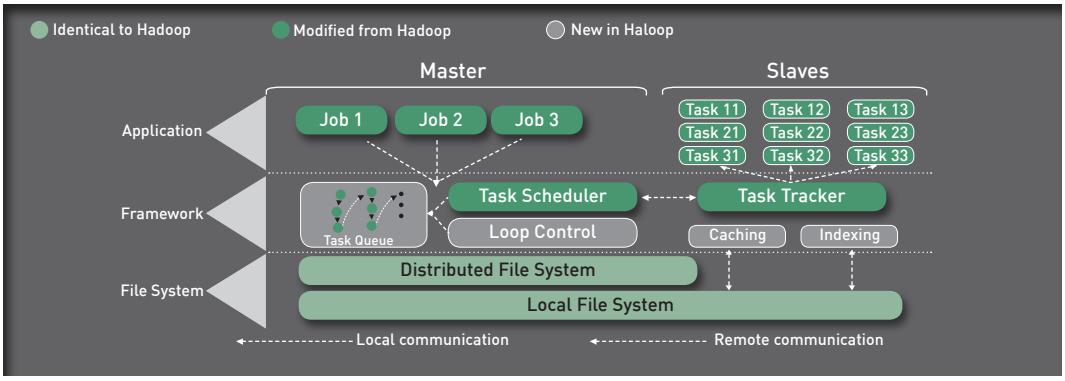
What-If 기술은 하둡의 성능이 파라미터 설정에 크게 좌우됨에 착안하여 하둡 파라미터들을 비용 기반으로 자동으로 최적화한다. 세부적으로 이전 잡(job) 수행에 관한 통계적 정보를 모으는 프로파일러(profiler), 파라미터가 약간 바뀌었을 때의 질의 수행 시간을 예측하는 What-if 엔진, 그리고 파라미터 공간

에서 What-if 엔진과 프로파일러를 이용하여 최적 파라미터를 찾는 비용 기반 최적화기의 세 가지 요소 기술들로 구성된다.

SkewTune 기술은 맵퍼들의 수행 시간의 비대칭도(skewness)가 맵리듀스 성능을 크게 저하시키는 요인임을 착안하여 이 문제를 해결하기 위해 현재 실행중인 태스크(task)를 실시간으로 분석하여 완료까지 남아 있는 시간이 가장 긴 태스크를 낙오자(straggler)로 정의한다. 그리고 낙오자 태스크를 가용한 다른 맵퍼들에게 동적으로 재할당하여 성능을 향상시킨다.

YSmart 기술은 주어진 SQL 질의를 Hive를 통해 맵리듀스 잡들로 변환할 때, 질의 내부의 상관관계(correlation)를 이용하여 잡들의 개수를 줄이고 중간 결과 데이터의 읽기를 줄임으로써 결과적으로 질의 처리 성능을 향상시키는 기술이다. 그러나 YSmart를 통해 변환된 더 적은 개수의 잡들의 수행 시간이 오히려 더 오래 걸릴 수 있다는 단점을 가진다. HaLoop 기술은 머신러닝 알고리즘과 같이 반복적으로 맵리듀스 작업을 수행하는 질의에서 반복 연산 중 변하지 않는 데이터(loop-invariant data)가 성능 저하의 주요 원인임을 착안하고, 입력 데이터들 중 어떤 데이터가 변하지 않는 데이터인지 사용자로부터 명시 받은 후 이 데이터에 대한 처리 결과를 로컬 파일 시스템에 캐싱하여 반복 질의에 대한 성능을 향상시킨다. (그림 4-1-5)는 HaLoop 기술의 아키텍처를 보여준다.

**(그림 4-1-5) HaLoop 기술의 아키텍처**




Restore는 Pig, Hive, Jaql과 같은 상위 수준 데이터 플로우 시스템에서 주어진 질의 또는 작업을 여러 개의 서브 작업들로 나누고, 각 서브 작업들의 결과를 저장한 후, 나중에 다른 질의 또는 작업을 수행할 때 서브 작업 단위로 이전 서브 작업 결과를 재사용하고자 하는 기술이다.

MRShare는 여러 개의 맵리듀스 질의들을 배치 처리할 때, 데이터 스캔 수준에서 공유가 가능하거나, 맵퍼의 입력 수준에서 공유가 가능하거나, 혹은 맵퍼의 출력 수준에서 공유가 가능한 경우를 탐지하여 공유 가능한 연산자들을 그룹핑함으로써 전체 배치 처리 성능을 향상시키는 기술이다.

### 다. 스트리밍 및 인메모리 기반 처리 기술

하둡의 처리 속도를 크게 향상시키기 위해 내고장성 특성을 일부 또는 전부를 포기하고 데이터를 스트리밍 또는 파이프라닝(pipelining) 방식으로 처리하거나, 데이터를 인메모리 방식으로 처리하는 기술들도 꾸준히 연구·개발되고 있다. 대표적으로 HOP, MR-hash/INC-hash, M3R이 있다.

HOP(Hadoop Online Prototype) 기술은 맵퍼(또는 리듀서)의 결과를 디스크에 모두 기록한 후 리듀서(또는 다음 번 잡의 맵퍼)를 실행하는 배치 방식 대신, 일부 결과가 준비되는 대로 리듀서(또는 다음 번 잡의 맵퍼)를 호출하는 파이프라이닝 방식을 사용한다. 그러나 이 기술은 맵리듀스 플로우 상에서 정렬-병합(sort-merge)의 정렬 단계에서는 CPU 오버헤드로 인해, 그리고 병합 단계에서는 I/O 오버헤드로 인해 블록킹 문제가 발생하는 단점이 있다.

MR-hash와 INC-hash 기술은 HOP의 블록킹 문제를 해결하기 위해 정렬-병합 기법 대신 해시 기법을 사용한다. 데이터를 단일 패스(one-pass)로 읽어 들이면서 MR-hash 또는 INC-hash라는 다단계 해시 구조를 사용하여 점진적(incremental)으로 맵리듀스 연산을 수행한다.

M3R(Main Memory MapReduce) 기술은 하둡이 아닌 X10이라는 메모리 기반 분산 프로그래밍 언어를 사용하여 데이터를 처리하며 모든 데이터를 메모리에 상주시킨 채 맵리듀스 연산을 인메모리 방식으로 수행한다. 디스크 입출력 비용 뿐만 아니라 일부 통신비용을 제거함으로써 속도를 크게 향상시키지만 내고장성이 보장되지 않고 처리할 수 있는 데이터의 크기가 메모리 크기에 제한되는 단점을 가진다.

## 제2절 빅데이터 분석 및 시각화

### 1. 빅데이터 기술의 중요성

오바마 행정부는 2012년에 2억 달러 규모의 "Big Data Research and Development Initiative"를 발표하였고 이를 통하여 다양한 분야에서 축적되고 있는 방대하고 복잡한 데이터로부터 지식과 통찰을 추론해 낼 수 있는 빅데이터 기술 개발을 적극 지원하고 있다. 미국이 정부 차원에서 적극 지원하게 된 이유는 빅데이터를 효율적으로 관리하고 그로부터 유용한 정보를 추출해내는 것이 여러 산업에 걸쳐 매우 시급하고, 사회경제적으로 미치는 영향이 아주 크기 때문이다.

최근 몇 년 동안 인간이 만들어낸 데이터가 그 이전에 인간의 모든 역사를 통해서 만들어낸 데이터보다 더 많다고 할 정도로 많은 데이터들이 현재 축적되고 있다. 이렇게 다양한 분야에서 수집된 빅데이터를 분석해서 지식을 찾아내고 이를 기반으로 다양한 서비스에 사용할 수 있다. 또한 빅데이터 분석 기술은 세계의

기업 및 기술 평가 단체로부터 새로운 산업 원동력으로 평가 받고 있으며 이를 필요로 하는 시장은 점점 더 커질 것이라고 전문가들은 전망하고 있다.

## 2. 맵리듀스 프레임워크를 이용한 빅데이터 분석 기술

빅데이터를 처리하기 위해서는 아주 빠른 컴퓨터가 필요한데 여러 가지 물리적인 제약으로 인하여, 약 2년 또는 1년 6개월마다 CPU의 성능이 2배씩 빨라진다는 무어(Moore)의 법칙이 더 이상 실현되기 어려워졌다. 그에 대한 해결책으로 빅데이터의 분석을 위해 저렴한 컴퓨터를 수백 또는 수천 대를 연결해서 만든 클러스터에서 병렬 분산 프로그램을 실현하여 속도를 빠르게 하려는 노력이 시도되고 있다. 이러한 환경에서 분산 병렬 처리 알고리즘을 손쉽게 개발할 수 있는 맵리듀스 프레임워크가 최근 관심을 받고 있다. 구글의 맵리듀스 또는 오픈소스 진영에서 만든 하둡은 그러한 병렬 분산 프로그램을 개발하기 위한 맵리듀스 프레임워크를 제공하는 효율적인 툴이다.

현재 빅데이터의 처리를 필요로 하는 많은 분야에서 전통적인 알고리즘을 맵리듀스 프레임워크를 이용하여 분산 병렬 처리 알고리즘으로 변환하고 있다. 기본적으로 분산 병렬 처리를 위해서는 기존의 알고리즘을 분할 정복(divide and conquer) 형태를 가지도록 변환해야 하고, 가능하면 여러 컴퓨터가 계산을 위해 공유해야 하는 데이터를 줄여야 네트워크 통신비용을 낮추어 효율적인 병렬 분산 알고리즘을 만들 수 있다.

### 가. 맵리듀스 기반 빅데이터 분석

데이터 마이닝 분야의 중요한 기술에는 클러스터링(clustering), 연관규칙(association rules), 시퀀셜 패턴(Sequential patterns), 자동분류(classification), 확률적 모델링(probabilistic modeling) 그리고 그래프 분석(graph analysis)이 있다. 맵리듀스를 이용한 CLIQUE, DBSCAN, OPTICS 그리고 계층적 클러스터링 알고리즘들의 분산병렬 알고리즘이 개발되었다. 또한 행렬에서 비슷한 특성의 원소들을 모아 군집화 하는 코-클러스터링(co-clustering)에 대한 맵리듀스 알고리즘도 개발되었다.

연관규칙에 대한 FP-Growth 알고리즘을 이용한 맵리듀스 알고리즘도 개발되었으며 결정 트리 모델과 아다부스트(Adaboost) 학습 알고리즘도 맵리듀스를 이용하여 개발되었다. 또한 다양한 그래프 마이닝에 대한 맵리듀스 알고리즘들이 개발되었다. 추천시스템에 사용되는 PLSI, LDA, Hidden Markov Model등과 같은 확률 모델의 파라미터를 학습하기 위한 EM-알고리즘들도 맵리듀스를 사용한 병렬 분산 알고리즘들이 이미 개발되었다.

### 나. 맵리듀스를 이용한 데이터베이스 질의 처리

세타-조인(theta join)과 유사도 조인(similarity join)의 맵리듀스 알고리즘들도 개발되었으며 n-way 조인 알고리즘의 수행을 효과적으로 처리하는 맵리듀스 알고리즘도 개발되었다. 유사도 조인에 대해서는 집합(set) 데이터와 벡터 데이터에 대한 유사도 조인 알고리즘이 개발되었다. 또한 최근에는 스카이라인 연산자의 질의 처리를 위해 히스토그램을 활용하여 필요 없는 계산을 줄이고 동시에 리듀스 함수의 부하를 균등하게 분배하는 효율적인 맵리듀스 알고리즘도 개발되었다.

## 3. R을 이용한 빅데이터 분석 기술

R은 통계분석에 널리 이용되고 있는 프로그래밍 언어(혹은 통계 패키지)이다. R은 통계 관련 패키지 뿐만 아니라 연관규칙 마이닝이나 클러스터링과 같은 많은 수의 데이터 마이닝 관련 패키지들도 제공하며, 총 5000개 이상의 패키지를 제공한다. 뿐만 아니라 객체 지향 프로그래밍(object-oriented programming) 기반으로 프로그래밍이 쉽고 시각화(visualization)에 대해 강력한 지원을 하는 등의 장점이 많아 현재 200만 명이 넘는 사용자가 이용하고 있다고 한다. 그러나 R은 기본적으로 메모리상에 모든 데이터를 올려놓고 처리하기 때문에 빅데이터에 대한 처리가 여전히 힘들다는 단점이 있다. 구글에서도 이러한 이유 때문에 R을 상용서비스가 아닌 프로토타입 테스트 등에만 이용하고 있다고 한다.

이러한 문제를 해결하여 R을 빅데이터 분석에도 이용하기 위해 R과 하둡을 연동하려는 움직임이 최근에 일고 있다. R에서 제공하는 다양한 라이브러리 및 시각화 도구와 하둡의 뛰어난 확장성을 결합하려는 것이다. 대표적인 패키지로는 RHadoop이 있다. RHadoop을 이용하면 하둡기반의 맵리듀스 프로그램을 R에서 작성할 수 있어서 비교적 쉽게 하둡 기반의 분석을 할 수 있다. 또한 SQL을 이용해 더욱 쉽게 빅데이터 분석을 하기 위한 패키지인 RHive도 출시되어 이용되고 있다.

## 4. 빅데이터 시각화 기술

데이터 분석을 통한 결과를 보여주기 위해서 표와 그래프가 많이 이용되고 있다. 이러한 표현 방법들을 이용하면 정보를 정확하게 표시할 수는 있지만 사람이 한눈에 그 내용을 파악하기는 힘들다. 사람은 컴퓨터에 비해 숫자나 문자를 빨리 읽을 수는 없지만 색깔이나 크기 등을 이용한 시각적 패턴을 인식하는 능력은 뛰어나다. 따라서 데이터에 대하여 시각적 패턴을 이용해 정보를 전달하는 것을 데이터 시각화라 한다. 데이터 시각화는 소셜 네트워크에서 군집의 분포와 네트워크의 변화양상이나 문서에서의 단어분포를 나타내는

것을 포함해 많은 분야에서 널리 사용되고 있다. 예를 들어 문서에 자주 등장하는 단어들을 시각화할 때 각각의 문서를 따로 분석해 시각화하는 것이 아니라 같은 단어를 비슷한 위치에 비슷한 색깔과 방향으로 표시하면 한 눈에 유사한 문서인지를 판단하기 쉽다. 최근에는 특히 소셜 네트워크 등과 같은 빅데이터의 직관적인 시각화에 대한 필요성이 증가하고 있다. 그러나 빅데이터 분석을 통한 시각화를 위해서는 많은 양의 계산이 필요해 이를 현실적인 시간에 수행해 사용자들에게 제공하기 위해서는 지속적인 연구가 필요하다.

## 제3절 빅데이터 활용 사례

### 1. 해외 기업

#### 가. 자동차 제조사

볼보(Volvo)는 과거 50만대 차가 팔린 뒤에 비로소 알 수 있었던 결함을 이제는 1,000대 판매 시점에서 발견이 가능하다. 이는 자동차 운전 과정에서 발생하는 데이터를 빅데이터 플랫폼으로 전송하고 데이터를 축적 및 분석하여 제품 개발 시에 찾기 어려운 다양한 결함과 고객의 요건을 파악하고 대응할 수 있었다. 운전자의 운전과정에서 수집된 데이터를 분석하여 잠재 수요를 파악하며, 제품 사용데이터 외 콜센터 상담 내용, 사용 후기, 소셜 미디어를 분석하여 R&D, 품질과 서비스 업무 전반의 통합적 관점을 제공한다.

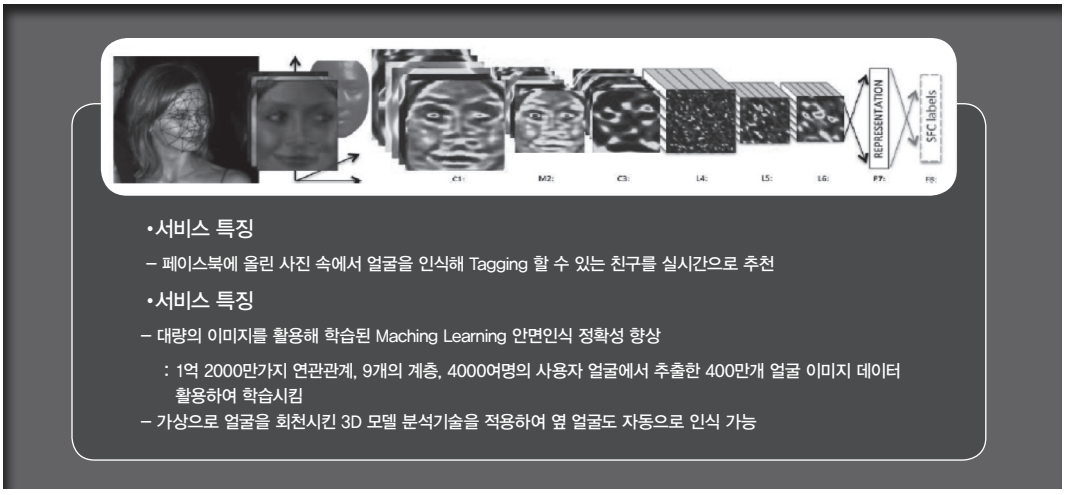
이스라엘에서 르노의 전기자동차 Fluence ZE는 100% 전기 배터리만으로 주행되고 100마일을 달릴 수 있다. 배터리는 충전소에서 교체하는데 5분이 소요되고 집과 사무실에서도 6~8시간 만에 완전 충전할 수 있다. 르노는 이스라엘 전 지역에서 운행되는 자동차의 신호를 실시간으로 수집하고 모니터링 및 분석하여 운전자에게 충전해야 하는 시기와 인근의 가장 가까운 충전소를 알려준다. 뿐만 아니라 전기 자동차 1대는 1가구의 전기를 필요로 하기 때문에 도시 전역의 전력 소비량을 고려한 에너지 관리가 중요한데, 이를 위해 전 지역의 에너지 사용량 데이터를 모아 도시 전체의 에너지 사용량 그리드(grid)를 분석하여 정전 예방 서비스를 개발하고 있다. 도요타는 빅데이터 교통정보 서비스를 개발하여 일본 전국의 지자체나 일반 기업과 개인을 대상으로 서비스를 제공한다. 이 서비스는 텔레매틱스를 통해 수집·축적한 차량의 위치와 속도, 주행 상황 등의 정보를 가공하고 교통 정보나 통계 데이터 등을 지자체나 기업에 제공하여 교통 흐름 개선이나 방재 대책 등에 활용한다. 개인용 smart G-Book은 고객의 음성 정보를 인식하는 센터형 음성 인식 기능을 새롭게 설정해 목적지 검색이나 설정시, 애매한 지시에 대해서 적절한 정보를 인출하는 것이 가능하다.[11]

11) Oracle, Cover Story – Oracle Big Data 사례, 2013; 매일경제, 도요타, '빅데이터' 활용 신개념 정보서비스, 2013.6.7. http://news.mk.co.kr/newsRead.php?no=443993&year=2013; https://www.youtube.com/watch?v=rTAn1bvy8vU 등을 참조하여 재구성.

## 나. 인터넷 기업

페이스북의 얼굴 인식 서비스는 비정형 데이터 분석의 대표적 사례로, 머신 러닝 알고리즘을 활용하여 안면 인식 정확도를 높이고 있다. 머신 러닝 알고리즘은 학습에 필요한 방대한 데이터의 저장 및 오랜 연산처리 시간으로 기존에는 활용에 한계가 있었으나, 빅데이터 기술의 발전에 따라 거대하고 비정형화된 데이터를 저장할 수 있게 되었으며 복잡한 알고리즘을 빠르게 연산 처리 할 수 있게 되었다. 결과적으로 사진 속 안면 인식 정확도를 높이고 태깅(tagging) 대상을 자동으로 추천함으로써 서비스 편의성을 향상시켰다.[12]

**(그림 4-1-6) 페이스북 얼굴 인식 프로세스 설명**




## 다. 의료 사업

미국 Augmedix는 구글 글래스를 활용한 Advanced EMR 환경을 연구하고 있다. 가령 의사가 환자 상태를 음성으로 말하면 구글 글래스가 이를 인지하여 EMR(Electronic Medical Record) 데이터를 자동 입력하고, 의사가 환자의 과거 병력 정보를 쉽게 알 수 있도록 구글 글래스 화면에 표시하여, 의사들이 자신의 환자에게 더욱 집중할 수 있게 하는 기술을 구현하고 있다. Augmedix는 EMR 데이터 입력 시간을 현저하게 줄여줄 뿐만 아니라 더 높은 품질의 데이터를 저장할 수 있게 하는 것이 핵심 기술이며, 미국의 여러 기관에서 채택하여 시범 적용하고 있다.[13]

12) The Huffington POST, 페이스북, 얼굴인식 기능 '딥페이스' 개발, 2014.03.21. http://blog.naver.com/theimc?Redirect=Log&logNo=220001569088 참조
13) 최윤섭의 Healthcare Innovation, Augmedix: 의사들에게 구글 글래스를 !, 2014.3.31. http://www.yoonsupchoi.com/2014/03/31/augmedix_glass_for_doctors 참조

(그림 4-1-7) 구글 글래스를 활용한 환자 진찰


1. 구글 글래스를 통해, 음성을 인식하여 EMR이 서버에 자동으로 기록됨
2, 3. 환자의 진찰 내용들이 자동으로 구글 글래스에 실시간으로 표기 됨

## 라. 미국 보험사

Progressive Insurance는 Snapshot이라는 운전 기록 데이터를 전송하는 기기를 차량에 설치하여 수집한 데이터를 기반으로 운전자의 운전 습관을 분석하여 위험도에 따른 맞춤형 보험을 제공한다.

(그림 4-1-8) Progressive Insurance사의 Snapshot 서비스 프로세스


※ 출처 : Forbes, The Big Data Boom : Is your Company Ready?, 2013.10.30

## 마. Telecom

Telecom Italia는 도시 삶의 질 향상 프로젝트를 위한, '빅데이터 챌린지 2014'를 개최하여 빅데이터 분야에서 최고의 아이디어를 수집했다. 세계 각국의 데이터 분석과 기획 전문가들이 지원하였다. 유럽공과대학 ICT 랩(EIT ICT Laps), 매사추세츠공과대학 미디어 랩(MIT Media Laps)과 트렌토 라이즈(Trento RISE)등이 함께 선정한 우승자는 통신, 공공 및 민간 교통, 에너지 소비, 밀라노의 도시 날씨와 소셜 데이터를 활용한 "Smart City"를 선보였다.[14]

### 바. 기타

사물인터넷(IoT)용 인공 지능 솔루션 Neura는 사람과 주변 환경 상황을 분석하여 필요한 서비스를 예측하고 제시하는 솔루션이다. 예를 들면, 스마트 시계를 착용한 사용자가 조깅 후 돌아오는 시간에 맞춰 특정 지점에서 보일러 온수 버튼을 활성화하여 집에 도착하자마자 샤워를 할 수 있도록 지원한다. 또 뉴욕 야구장에서는 입장과 동시에 스마트폰을 통해 자동으로 안내 메시지를 받는다. 야구장 내 시설·상점의 실시간 위치 안내와 쿠폰을 스마트폰을 통해 제공받는 서비스를 경험할 수 있다. 이는 초절전 무선 통신 기술인 BLE(Bluetooth Low Energy)를 내포한 비콘(Beacon)장비와 스마트폰 단말기와의 통신으로 사용자가 현재 어느 위치에 있으며, 어떤 서비스를 필요로 하는지를 예측하여 사용자에게 적합한 서비스를 제공할 수 있게 하는 것이다.[15]

## 2. 국내 기업

### 가. 금융

IBK기업은행은 소셜 미디어 등 온라인상에서 금융 업계의 이벤트, 서비스, 신상품 등에 대한 정보를 수집하고 분석해 각 부서에서 필요로 하는 외부 빅데이터와 내부 고객관리(CRM) 데이터를 결합하는 마케팅 영역 빅데이터 활용을 확대할 예정이다. 대우증권은 온라인 소셜 데이터를 활용해 사람들의 관심사를 파악하고 금융상품 광고 콘셉트에 어울리는 고객 커뮤니케이션 메시지를 발굴하고 있다.

신한카드의 S-초이스 체크카드는 빅데이터 분석과 고객 패널의 의견 등을 고려해 면밀히 설계된 상품이다. S-초이스 체크카드는 고객이 교통·커피·쇼핑 등 주력 서비스 세 가지 중 하나를 선택하면 여기에 할인 혜택을 집중해 준다. 1년 만에 카드 발급 수 200만 장을 돌파하여 체크카드 시장의 베스트셀러가 되었다.

하나SK카드는 하나은행, SK플래닛, BGF리테일, 이베이코리아 등 은행, 멤버십, 유통, 온라인 쇼핑을 대

14) Telecomltalia, Your chanllenge to Big Data, 2014.4.3.

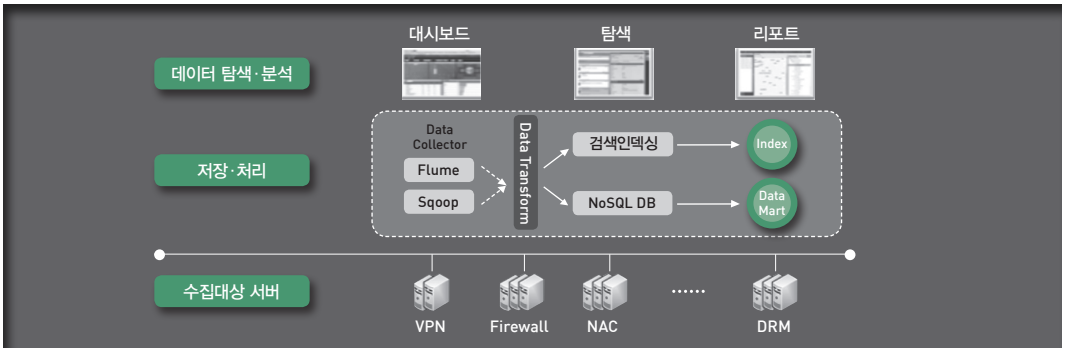
15) KT DigiEco, Trend Spectrum 동향 브리핑, 2014.4.15, http://www.theneura.com/IoT 및 CNET News, MLB tests Apple's iBeacon at Citi Field, 2013.9.27 참조

표하는 기업들이 직접 참여해 OK캐시백, CU편의점, G마켓, 옥션 등 기업의 멤버십 빅데이터를 카드 한 장에 집약했다. 현대증권은 뉴스 데이터로 주가를 예측하는 지표를 개발해 내부 자료로 활용하고 있다. 특정 종목이 언급된 부정적이거나 긍정적인 뉴스가 발생할 경우 해당 뉴스가 주가와 거래량 변화에 어떤 영향을 얼마나 미치는지를 분석한 지표로 리서치센터 내에서 활용한다. 국내외에서 소셜의 특정 텍스트와 감성이 주가 지수에 영향을 미친다는 가정이 입증됨에 따라, 현대증권 외에도 국내 다양한 금융기업에서 소셜 연계 상품 개발 출시를 검토 중이다.[16]

## 나. 전자·제조

건축 자재를 제조하는 A사는 빅데이터 로그처리 기술을 활용하여 200여 대 서버로부터 50여 종의 보안시스템 로그를 수집·분석하고 정보 유출 위험 인자를 도출하여 정보보안 추적성을 강화하였다. 풀 텍스트(full text) 검색 기능을 통해 실시간으로 로그를 탐색하여 위험 인자를 도출하는 등의 활동으로 위험 경보를 통한 보안사고 사전 예방 활동을 수행하고 있으며, 보안 감사, 퇴직자 보안관리 등 컴플라이언스 대응 생산성을 향상시켰다.

**(그림 4-1-9) LG CNS, Smart LAP를 활용한 정보 유출 방지**




LG전자는 한국과 해외 주요국가에서 마케팅을 펼칠 때, 다른 브랜드 경쟁력과 비교해 신제품 출시에 따른 고객 구매 포인트 도출, 광고 콘셉트 개발에 소셜 미디어 분석과 텍스트 마이닝 기술을 활용하고 있다. 기존에 마케팅에 활용되었던 소셜 미디어 분석을 상품 콘셉트 개발 단계에 적용하여 차기 버전 제품의 콘셉트에 포함하는 활동을 진행하고 있으며, 실제로 거실에서의 라이프 스타일과 행동 패턴을 연구해 상품 기능에 접목하고 있다.[17]

16) 매일경제, LG CNS, 금융·제조·통신 '빅데이터' 효과 톡톡, 2013.9.3. 및 LG CNS, 빅데이터에 대한 오해와 금융권의 빅데이터 활용, 2014.1.2
17) 아주경제, 삼성·LG전자, 빅데이터 기반 시장전략 '쪼개고 쪼갠다', 2014.2.16 및 http://bigdata.lgcns.com, http://blog.naver.com/smart_sma/20204418448 참조

## 다. 통신회사

LG유플러스는 스마트폰이 스스로 사용자에게 필요한 맞춤형 정보를 제공하는 능동형 스마트 비서 서비스 'U스푼'을 출시했다. 이는 지능형 감성 서비스로 고객이 자주 이용하는 날씨, 교통, 모닝콜 등의 패턴을 파악해 개인의 상황에 따라 정보를 맞춤형으로 미리 전달해준다. 예를 들어 어제와 비교해 날씨를 알려주거나, 교통으로 인해 제시간에 출근이 어려울 때 더 빨리 깨워주며, 오랜만에 착신된 전화에 '00일 전에 연락한 친구입니다.'라는 메시지를 알려준다. 나아가서 단말기 사용 패턴을 분석하여 출근 시간에 커피를 마시기 좋아하는 직장인에게 커피 쿠폰을 제공하는 등의 인텔리전스 서비스 영역으로 확대할 수 있다.

SK텔레콤은 실내에서 스마트폰 사용자의 위치를 확인시켜 주는 장비인 '블루투스 저전력 비콘'을 출시했다. 비콘은 근거리 위치 인식 기술을 적용시킨 무선 센서로 실내 스마트 디바이스와 통신이 가능하여 소비자에게 각종 정보와 개인 맞춤형 서비스를 제공할 수 있다. 복잡한 실내 건물에서 현재 위치 및 목적지를 스마트폰에 표시해 주고 개인 맞춤형 쿠폰을 발행하기도 하는데, 이러한 서비스는 서울대병원과 잠실학생체육관에서 운행되고 있다. 이외에 SK텔레콤은 데이터 개방과 공유를 위한 플랫폼인 빅데이터 허브(Big Data Hub)를 제공하고 있는데 교통, 기상, 상권 등의 광범위한 사회적 데이터를 이용할 수 있으며, 누구나 데이터를 등록, 전시, 분석할 수 있다.

한편 KT는 연세대 의료원과 손잡고 '후헬스케어(HooHHealthcare)'라는 합작회사를 설립했다. 전자 진료기록부, 의료 영상저장 전송 등 기존에 있던 솔루션 기능은 물론, 근거리 무선통신(NFC), 클라우드, 빅데이터 분석·처리 등의 통신 신기술을 통해 한층 업그레이드된 통합 병원정보 서비스를 제공한다.[18]

## 라. 화장품

LG생활건강은 '개인화 추천 솔루션'을 개발했는데, 고객들의 구매 주기, 품목, 상품 검색 기록, 반품기록, SNS기록 등을 분석해 구매심리를 예측하여 6,000여명의 고객을 대상으로 화장품 재구매 시기를 추천했고, 그 결과 판매량이 3배 증가하였다.한편 로레알은 광고 모델로 배우 공효진을 재선정하였는데, 그 배경에는 SNS 분석의 힘이 있었다. 텍스트 마이닝을 활용한 각종 SNS상의 공효진과 비오템 상품 분석 결과, '따라하고 싶다', '공효진은 이 제품을 직접 쓸 것 같다'는 긍정적 반응이 다른 모델에 비해 압도적으로 높게 나와 '공효진은 패션 스타일은 좋지만, 정형적인 미인 스타일이 아니라서 화장품에 적합하지 않다.'는 업계의 일반적인 인식을 무너뜨렸다.

18) 조선일보, 맞춤형 스마트 비서 서비스 'U스푼', 2014.4.2, 빅데이터허브(http://www.bigdatahub.co.kr), https://www.youtube.com/watch?v=fvcPk_13J8o 참조.

아모레퍼시픽은 '화장대 지수(고객의 화장품 가운데 아모레퍼시픽이 차지하는 비율)'을 개발하고, 구매 패턴과 성향 등을 고려해 11개 그룹으로 고객을 세분화했으며, 특정 소비자 그룹을 대상으로 맞춤 패션 정보를 적극 제안하는 등 빅데이터를 기반으로 맞춤 마케팅 활동을 전개하고 있다.[19]

### 마. 유통

종이 가격표 대신, 전자 가격 표시기가 홈플러스에서 등장했다. 전자가격표시기(Electronic Shelf Label : ESL)은 저 전력 무선 통신 기술인 지그비(Zigbee)를 이용해 상품정보를 전달하여 가격정보와 제품정보를 실시간으로 변경하여 표시할 수 있다. ESL이 도입되면서 대형마트가 문을 닫은 뒤 직원들이 매장을 돌며 수만 개의 바뀐 라벨을 붙이는 모습이 사라져 매장 운영비용 절감 및 효율성을 높였다. 특히 요즘 소비자가 오프라인 매장에서 물건을 보고 온라인으로 구매하는 추세이나 ESL을 사용하면 오프라인 매장 가격을 실시간으로 조정해 온라인에서 소비자를 끌어올 수 있으며, 본사에서 가격을 바꾸겠다고 결정하는 순간 전국 매장의 가격 태그를 동시에 바꿀 수도 있다. 더 나아가 온라인 매장의 가격정보를 수집·분석하고 매장 내 고객들의 실시간 가격정보를 활용하여 원하는 고객에게 적합한 가격과 정보를 제시하거나 프로모션을 시행할 수 있는 인프라가 될 수도 있을 것이다.[20]

## 3. 마무리

빅데이터가 새로운 트렌드 키워드로 떠오른 지 3년이 흘렀다. 산업혁명 시절 미개척지의 금광을 찾는 심정으로 선도적 기업에서 빅데이터의 새로운 가치 창출을 찾아 나섰고, 새롭고 혁신적이며 다양한 시도를 행하였다. 그 결과로 'SNS 정보 분석'이란 울타리에 한정되어 언급되던 빅데이터의 가치는 기업의 필요성과 역동적으로 결합되어 과거에는 생각지도 못했던 분야의 사례들로 발전하였고 산업과 업무를 넘어 다각적으로 전개되고 있음을 여러 사례에서 살펴 볼 수 있었다.

빅데이터는 단순 분석 단계를 넘어, 문제 해결을 위한 효과적인 수단으로 빠르게 확산되고 있다. 이제 빅데이터는 문제 해결 단계를 넘어, 우리의 생활을 보다 똑똑하게 영위하기 위한 상품 콘셉트 개발과 새로운 서비스 개발로의 혁신적 사고를 이끌어내고 있다.

19) 한국일보, 화장품 광고 모델 선정 뒤엔 빅데이터 입김, 2014.4.14.
20) 식품저널, "종이가격표는 가라" 대형마트에 '전자가격표시기' 등장, 2014.3.11.

CHAPTER 2

# 데이터베이스 기술

## 제1절 메모리 기반 데이터베이스

### 1. 메모리 기반 데이터베이스 최근 기술 동향

DRAM과 플래시메모리 기반 저장 장치 기술 발전에 따라, DB 분야에서 최근에 3가지 기술 동향이 나타나고 있다. 우선 순수 DB 관점에서는 모든 데이터를 DRAM에 상주시켜서 저장 관리하는 메인메모리 DB와 전통적인 디스크 기반의 DB에서 하드디스크 대신에 플래시메모리 SSD(Solid State Disk)를 저장 장치로 대체하는 방식이 주류를 이루고 있다. 한편 구글, 아마존, 페이스북, 애플 등 데이터센터 기반의 클라우드 컴퓨팅 환경에서는 Key-Value 저장소의 저장 장치로 플래시메모리 SSD를 사용하는 이른바 All-flash 데이터센터가 주목할 만한 기술이다.

#### 가. 메인메모리 데이터베이스 활성화

최근 DB 엔진 기술 측면에서 가장 기술적인 진보가 많이 이루어지고 주목을 끄는 분야가 메인메모리 DB 분야이다. 2000년대 초반부터 군소 업체 중심으로 메인메모리 DB 엔진들이 존재해왔으나(예: Altibase, Oracle TimesTen), 최근 ERP 벤더인 SAP에서 자사의 DB 엔진 솔루션으로 Hana를 발표하고, Microsoft로 Hekaton이라는 메인메모리 DB 기술을 발표함으로써 본격적인 시장 진입에 나섰다. 이러한 활성화의 배경에는 극단적으로 빠른 성능을 필요로 하는 통신, 금융 분야 등 애플리케이션들의 요구사항과 기존 하드디스크 기반 저장 장치의 느린 I/O 성능 특성 때문이다. 특히 후자와 관련해서 대형 온라인 트랜잭션 처리 시스템에서는 기존 디스크 기반의 DB 기술로는 CPU와 저장 장치의 균형 상태를 유지하기 위해 저장 장치에

서 초당 수만, 많게는 수십만 IOPS(IO per seconds)를 필요로 하는데, 하드디스크의 경우 대당 빨라야 300 IOPS 정도밖에 제공하지 못한다. 이 경우 수백 또는 수천 대의 하드디스크 구매 비용 및 관리 오버헤드 등을 고려하면 가히 IOPS의 위기 상황이라고 말할 수 있다. 이러한 고비용 저효율의 디스크 기반 DB 기술 대신에 상대적으로 저렴해진 DRAM 가격을 고려해서 모든 데이터를 DRAM에 저장하고 I/O 개념을 최소화함으로써 고성능 DB를 달성하는 메인메모리 DB가 합리적인 대안으로 자리 잡을 수 있었다.

## 나. 플래시메모리 SSD 기반 데이터베이스 활성화

한편 메인메모리 DB 기술이 활성화되는 최근 몇 년 사이 플래시메모리 SSD가 저장 장치 시장에서 급격하게 하드디스크를 대체 또는 경쟁하고 있다. 특히 IOPS 관점에서 최근 SSD는 100만 원 정도의 가격으로 수만 IOPS를 제공하기 때문에, 하드디스크에 비해 단위 가격 당 IOPS 입장에서는 수십 배 이상 저렴해졌다. 따라서 디스크 기반의 DB 기술에 최신 플래시메모리 SSD를 적용할 경우, 저비용 고효율의 IOPS가 달성된다. 플래시메모리 SSD가 DRAM에 비해 1/10 이하의 가격 경쟁력을 가지며 전력 소모, 발열·냉방 등 시스템 유지 관리비용에서도 모든 데이터를 DRAM에 저장, 관리하는 메인메모리 DBMS에 비해 장점을 갖게 된다. 이러한 이유로 Oracle, IBM, Cisco 등의 업체는 TPC-C 벤치마크 성능 평가에 플래시메모리 SSD를 사용하고 있다. 플래시메모리 SSD 기반 DB 분야는 크게 두 가지 움직임, 즉 1) 전체 DB를 하드디스크에 저장하고 자주 사용되는 일부 데이터를 플래시메모리 SSD에 캐싱하는 방식과 2) 모든 데이터를 플래시메모리 SSD에 저장하는 방식이 있다.  첫 번째 방식의 지원을 위해 Oracle과 IBM은 플래시메모리 SSD를 DRAM과 하드디스크 사이의 캐시로 관리하는 소프트웨어 모듈들을 제공하고 있다.

## 다. All-flash 데이터센터의 활성화

클라우드 컴퓨팅을 주도하는 아마존, 애플, 페이스북, 구글, 이베이 등은 자사 데이터 센터의 데이터를 저장하는 저장 장치로 기존 하드디스크 대신에 플래시메모리 SSD로 대체하는 즉 All-flash 데이터 센터로 전환하고 있다. 클라우드 컴퓨팅 환경에서 동작하는 웹 애플리케이션들이 QoS(Quality of Service) 관점에서 빠른 응답 속도를 필요로 하기 때문이다. 특히 빅데이터 기술 동향과 더불어서 전체 데이터를 DRAM에 관리하는 방식은 현실성이 떨어지고, 심지어 자주 사용되지 않는 데이터를 아주 가끔씩 접근하는 경우에도 빠른 응답시간을 보장하기 위해서는 현재 기술로는 플래시메모리 SSD에 데이터를 저장하는 것이 가장 현실적인 대안이기 때문이다.

## 2. 메모리 기반 데이터베이스 기술 발전 방향

향후 메모리 기반 DB의 기술 발전은 산업계 종사자나 연구자들에 따라 의견이 다를 수 있으나, 다음과 같은 일반적인 예측이 가능하다.

### 가. 플래시메모리 SSD 기반 데이터 관리 활성화

All-flash 데이터센터의 활성화와 Oracle, IBM 등 DB 분야 업계 1, 2위 업체가 플래시메모리 SSD 기반의 DB 기술을 적극적으로 드라이브하고, 또한 삼성전자, 인텔 등 플래시메모리 SSD 업체들에서 DB로 대표되는 엔터프라이즈 분야 진입을 위해 가격 및 성능 개선에 노력하고 있기 때문에 당분간 플래시메모리 기반의 DB 기술이 메인스트림 기술로 자리 잡을 것이다. 특히 플래시메모리 SSD가 제공하는 빠른 IOPS로 인해서 적당한 크기의 DRAM만으로도 DB 서버 시스템을 균형상태, 즉 CPU와 저장 장치가 모두 100% 활용되는 상황을 달성 가능하기 때문에 대부분의 DB 시장에서 당분간 가장 비용 대비 효율성 높은 DB솔루션으로 자리 잡을 것이다. 한편 미국의 주요 선도 업체들을 중심으로 기존 하드디스크에 최적화되어 개발된 소프트웨어 기술들을 플래시메모리 SSD의 특성을 반영해서 플래시메모리 SSD에 최적화하는 기술 개발이 적극적으로 이루어지고 있다. 이러한 기술 개발은 전통적인 디스크 기반의 DBMS에서도 이루어질 것이다. 한편 단순하게 읽기, 쓰기만의 기능을 제공하던 하드디스크와 달리 플래시메모리 SSD 벤더들은, 메모리 기반 DB 소프트웨어에서 필요로 하는 기술들인 자주 사용되고 핵심적인 기능들을 잘 정의해서 플래시메모리 SSD 계층에서 지원하려는 움직임도 있을 것이다. 이를 위해서 새로운 저장 장치 인터페이스를 정의하고 표준에 추가하는 움직임도 가속화될 것이다.

### 나. 메인메모리 데이터베이스 기술의 보편화

최근 몇 년 간에 이루어진 메인메모리 DB 분야의 기술 혁신은 단순히 모든 데이터의 DRAM 상주를 통한 관리를 넘어 CPU 기술의 활용을 통한 성능 최적화, 더 높은 수준의 동시성 제어 기술들을 포함한다. 이 기술들은 디스크 기반의 전통적인 DB 기술에도 접목될 것이다. 또한 국내에서 DB 관리의 주된 방식은 디스크 기반의 DB에서 훨씬 더 많은 양의 DRAM을 사용함으로써 하드디스크의 IOPS 문제를 해결하는 방향이다. DB의 양이 상대적으로 크지 않은 경우 DBA 입장에서는 지금까지 검증되고 안정화된 아키텍처를 선호할 것이다. 하지만 빅데이터 시대에 데이터 접근에 대한 전통적인 지역성(locality)이 희석되고 애플리케이션에서 아주 엄격한 응답시간을 요구하는 경우 DB 전체를 DRAM에 상주시키는 전통적인 메인메모리 DB 방식과 상대적으로 큰 DRAM을 사용하는 방식이 더 이상 현실적인 대안이 되지 않을 수도 있다. 이 경우 메인메모

리 DB 기술 기반으로 플래시메모리 SSD를 다양한 형태로 활용하는 하이브리드 기술의 등장도 예상해 볼 수 있다. 또한 DB 입장에서 새로운 메모리 디바이스의 등장도 메인메모리 DB 기술에 영향을 미칠 수 있다. 예를 들어 이미 시장에 출시되고 있는 DRAM용 DIMM 인터페이스를 가지면서 내부적으로는 플래시메모리 칩을 저장매체로 활용하는 장치들은 메인메모리 DB 입장에서 또 다른 기술 혁신의 기회를 제공할 것이다.

## 제2절 비정형 데이터베이스

### 1. 개요

다양한 매체들의 증가, 최신 기술의 발달, 그리고 빠른 인터넷 환경의 제공 등으로 인하여 다양한 종류의 데이터들이 인터넷을 통해 공유되고 있다. 이러한 데이터들은 사용되는 애플리케이션에 따라 정형화되지 않은 형태를 가지므로 비정형 데이터라고 부른다. 비정형 데이터의 증가와 함께 이 데이터를 효율적으로 관리할 수 있는 기술에 대한 관심 또한 크게 증가하고 있다. 비정형 데이터는 종류에 따라 그 특성들이 모두 다르기 때문에 각 데이터 종류마다 적용해야 하는 기술 또한 큰 차이점이 있다. 본 절에서는 비정형 데이터들 중 멀티미디어 데이터(multimedia data), 소셜 네트워크 데이터(social network data), 그리고 시공간 데이터(spatio-temporal data)의 관리 기술에 대하여 논의한다.

### 2. 멀티미디어 데이터

인터넷 상에는 사용자들에 의해 이미지, 동영상, 그리고 음악 등과 같은 수 많은 멀티미디어 데이터들이 생산, 공유되고 있다. 이러한 인터넷 상의 데이터들은 폭발적으로 증가하고 있고 조직화되어 있지 않기 때문에 일반적인 키워드 기반 검색을 이용하여 인터넷 사용자들이 원하는 데이터를 검색하기는 쉽지 않다. 사용자들의 멀티미디어 검색을 지원하기 위해서는 데이터의 특성을 활용한 다양한 관리 기술이 필요하다. 이러한 멀티미디어 데이터 관리 기술은 크게 (1) 데이터 특성 추출 기술(data feature extraction), (2) 내용 기반 데이터 검색 기술(content-based data retrieval), (3) 데이터 마이닝 기술(data mining)의 세 가지로 분류 가능하다. 데이터 특성 추출이란, 멀티미디어 데이터의 종류에 따라 그 데이터 고유의 성질을 잘 표현할 수 있는 특성을 추출하는 기술이다. 예를 들어 전통적으로 사용되는 이미지 데이터 특성 추출 기술에는 색 히스토그램(color histogram) 추출 기술이 있다. 색상 히스토그램 추출 기술이란 이미지를 표현하는 색상들을 n개의 막대(bin)로 구성되어 있는 다차원 히스토그램으로 표현하는 것이다. 만약 5개의 색상(RGB)을 가지는 이미지를 히스토그램으로 표현한다면, 이 이미지는 5개의 막대(bin)를 가지는 히스토그램으로 표현되며 그

이미지 안에서 각 색이 차지하는 비율이 각 막대의 크기로 표현된다. 또한 이 색상들은 RGB 3차원 공간상의 좌표로 표현되기 때문에 히스토그램의 차원 역시 3차원이다.

색상 히스토그램 추출 기술 이외에도 이미지를 구성하는 객체들의 형태에 대한 특성을 추출하는 기술인 SIFT(scale-invariant feature transform) 등 이미지 데이터를 위한 다양한 특성 추출 기술들이 있다. 데이터 검색 및 마이닝 기술을 적용하기 위해서는 데이터 특성 추출 기술이 필수적이며 멀티미디어 데이터의 종류와 사용 목적에 따라 다른 특성 추출 기술들이 요구된다.

멀티미디어 데이터 관리 기술 중 내용 기반 데이터 검색 기술은 멀티미디어 데이터의 특성을 기반으로 하는 검색으로, 사용자가 제시한 질의 데이터의 특성과 유사한 특성을 갖는 데이터를 검색하는 기술이다. 멀티미디어 데이터 자체를 질의로 사용하기 때문에 키워드 기반 검색에 비하여 직관적인 데이터 검색이 가능하다. 내용 기반 데이터 검색을 수행하기 위해서는 먼저 데이터 간의 유사도를 측정할 수 있는 거리 함수가 필요하다. 전통적으로는 두 데이터 간의 유클리드 거리(Euclidean distance)와 맨해튼 거리(Manhattan distance) 등의 거리함수들이 사용되었다. 최근에는 다양한 멀티미디어 데이터 특성에 맞춰 두 데이터 간의 유사도를 좀 더 정확히 계산할 수 있는 Earth Mover's Distance(EMD)와 Signature Quadratic Distance(SQFD)와 같은 기법들이 사용되고 있다.

내용 기반 데이터 검색 기술에서 검색의 정확도 만큼 중요한 것이 검색의 비용이다. 일반적으로 멀티미디어 데이터에서 추출되는 특성들은 고차원이기 때문에 거리 함수를 이용해 질의 데이터와 DB 내 모든 데이터 간의 유사도를 계산한다면 매우 높은 검색 비용이 발생할 수밖에 없다. 이러한 문제를 해결하기 위하여 인덱싱(indexing)과 해싱(hashing) 기법 등이 사용된다. 이러한 기법들을 이용하면 DB 내 일부분의 데이터만을 액세스할 수 있기 때문에 내용 기반 데이터 검색의 비용을 크게 줄일 수 있다. 기존에는 정확한 질의 결과를 도출하는 기법들이 주로 사용되었으나 멀티미디어 DB가 점점 대용량화, 고차원화 되면서 기존 기법들로는 검색 비용을 줄이는데 한계가 발생하였다. 이에 따라 최근에는 정확도를 조금 희생하면서 검색 비용을 크게 줄일 수 있는 근사 인덱싱(approximate indexing) 기법이나 확률적 모델에 기반한 지역민감해싱(Locality Sensitive Hashing : LSH) 기법들이 연구되고 있다.

마지막으로 사용자들의 검색을 지원하기 위한 데이터 마이닝 기술은 데이터 특성을 기반으로 유사한 데이터들끼리 그룹으로 묶어 멀티미디어 DB를 조직한다. 사용자가 원하는 멀티미디어 데이터에 대한 구체적인 정보가 없을 경우 내용 기반 데이터 검색을 수행하기 어렵다. 이 경우 사용자가 원하는 데이터를 발견할 때까지 본인이 직접 DB 내 데이터들을 일일이 탐색해야 하는 부담이 있다. 유사한 멀티미디어 데이터들끼리 그룹화한 후 사용자가 각 그룹별로 데이터를 살펴볼 수 있다면 자신이 원하지 않는 데이터들을 쉽게 탐색 과정에서 배제할 수 있다.

DB 조직화를 위해 사용되는 데이터 마이닝 기술 중 가장 대표적인 기술로 클러스터링(clustering)이 있다. 클러스터링은 각 데이터 간의 유사도를 이용해 유사한 데이터들을 하나의 클러스터(cluster)로 그룹화 한

다. 기존의 전통적인 클러스터링 기법들은 각 데이터 간의 유사도 만을 따지기 때문에 이상치(outlier)에 민감하며, 특정 클러스터에 너무 많은 데이터들이 포함될 수 있다. 이러한 문제들을 해결하기 위하여 최근에는 클러스터의 밀도(density)와 데이터 분포(distribution) 등을 고려한 다양한 클러스터링 기법들이 연구되고 있다. 이 이외에도 좀 더 효율적인 탐색을 위하여 DB를 트리 형태로 구조화하는 계층적 클러스터링(hierarchical clustering) 기법들이 연구되고 있다.

이 밖에도 키워드 기반 검색을 지원하기 위한 태깅(tagging) 기술, 사용자들의 취향을 고려한 멀티미디어 데이터 추천 기술 등 사용자들의 효율적인 멀티미디어 데이터 검색을 지원하기 위해 다양한 기술들이 연구되고 있는 추세이다.

## 3. 소셜 네트워크 데이터

인터넷 사용자들은 소셜 네트워크 서비스(Social Network Service : SNS)를 이용해 인터넷 상에서 자신들의 인맥 관계를 유지하며 다양한 정보를 공유한다. 최근 들어 트위터, 페이스북, 인스타그램과 같은 소셜 네트워크 서비스를 이용하는 사용자들이 폭발적으로 증가하면서 소셜 네트워크 데이터를 분석하는 기술들이 연구되고 있다. 이러한 분석 기술들은 마케팅, 물품 추천, 그리고 소셜 네트워크 사이트의 유지보수 전략 등의 다양한 애플리케이션에 적용될 수 있다. 소셜 네트워크 데이터 분석 기술은 크게 (1) 네트워크 구조 분석 기술(network structure analysis), (2) 콘텐츠 기반 분석 기술(content-based analysis), (3) 신뢰 관리 기술(trust management)의 세 가지로 분류 가능하다.

네트워크 구조 분석 기술이란, 네트워크의 위상 구조(topology structure)나 통계적인 분석(statistical analysis)을 이용하여 소셜 네트워크의 특성을 분석하는 것이다. 일반적으로 소셜 네트워크에서 사용자들은 그래프의 정점(node)으로 표현되며 사용자 간의 친구 관계, 정보 공유, 또는 평가와 같은 상호 간 행동(interaction)은 그래프의 간선(edge)으로 표현한다. 네트워크 구조 분석은 이러한 사용자 간의 상호 간 행동으로 이루어진 그래프 구조를 분석하여 네트워크 안의 커뮤니티를 탐지(community detection)하거나 소셜 네트워크의 진화(evolution in social networks) 단계를 분석할 수 있다.

또한 통계적인 분석을 더함으로써 많은 사용자들에게 영향을 끼치는 주요 정점이나 간선이 무엇인지 찾아낼 수 있다. 이 이외에도 네트워크 구조 분석 기술을 이용함으로써 특정 사용자나 정보가 소셜 네트워크에 어떤 영향을 끼치는지 분석하는 사회적 영향력 분석(social influence analysis)을 수행할 수 있다. 이러한 기술들은 주요 사용자를 이용한 광고, 바이럴 마케팅(viral marketing), 그리고 소셜 네트워크 서비스의 고객 유치를 위한 향후 전략 등에 이용된다.

콘텐츠 기반 분석 기술은 사용자들이 생성하는 콘텐츠들을 기반으로 소셜 네트워크의 특성과 사용자들의 취향 등을 분석하는 기술이다. 소셜 네트워크 서비스 사용자들은 자신들이 관심을 가지고 있는 주제, 뉴

스, 그리고 소비물품 등에 관련된 다양한 콘텐츠들을 생성한다. 콘텐츠 기반 분석 기술은 데이터 마이닝 기술을 기반으로 이러한 콘텐츠들을 분석함으로써 사용자들의 취향에 맞는 다른 콘텐츠나 사용자들을 추천하는 서비스에 적용할 수 있다. 예를 들어 옥션 사이트[21] 인 아마존과 이베이에서는 사용자들이 구매하거나 클릭한 물품 이력을 기반으로 사용자들이 관심을 가질 만한 다른 물품을 추천하는 서비스를 제공한다.

또한 콘텐츠 기반 분석 기술은 네트워크 구조 분석의 정확도를 높이는데 사용된다. 사용자들이 작성한 콘텐츠들은 소셜 네트워크 구조와 밀접한 관계를 맺고 있기 때문이다. 예를 들어 커뮤니티 탐지를 수행할 때, 콘텐츠 기반 분석 기술을 통해 사용자들이 공유하고 있는 콘텐츠나 주제를 파악한다면 네트워크 구조 분석 기술만을 수행했을 때보다 더 정확히 사용자들의 커뮤니티를 탐지할 수 있다. 또한 사회적 영향력을 분석할 때에도 콘텐츠 기반 분석 기술을 통해 콘텐츠의 유행과 정보의 흐름 등을 파악할 수 있다.

마지막으로 신뢰 관리 기술은 소셜 네트워크 내 각 사용자의 신뢰도를 평가하고 관리하는 기술이다. 소셜 네트워크 서비스의 폭발적인 증가와 함께 그 안에 확인되지 않은 루머와 거짓 정보를 생산하는 거짓 사용자(fraudster)들이 크게 증가하고 있는 추세이다. 이러한 거짓 사용자들은 소셜 네트워크 서비스의 질을 떨어뜨리고 일반 사용자들의 소셜 활동을 방해한다. 이러한 문제를 해결하기 위하여 신뢰 관리 기술은 각 사용자의 신뢰도를 평가하여 믿을 수 있는 사용자와 거짓 사용자들을 구분한다. 각 사용자의 신뢰도를 평가하기 위하여 콘텐츠 기반 분석 기술과 네트워크 구조 분석 기술이 모두 이용된다. 거짓 사용자들은 공범자(accomplice)들을 이용하여 자기 자신의 정체를 위장하기 때문에 정확한 신뢰도 평가를 위해서는 위의 두 가지 기술이 모두 적용되어야 한다.

먼저 콘텐츠 기반 분석 기술을 이용해 특정 사용자의 행동 패턴을 분석함으로써 그 사용자의 초기 신뢰도를 파악할 수 있다. 예를 들어 사용자가 생성한 콘텐츠 정보, 그 콘텐츠에 대한 다른 사용자들의 직접적인 평가, 그리고 그 사용자가 다른 사용자들에게 준 평가 등을 분석하면 그 사용자의 초기 신뢰도를 파악할 수 있다. 그러나 전술한 바와 같이 거짓 사용자들은 공범자를 이용해 이러한 정보들을 조작함으로써 초기 신뢰도를 위장할 수 있다. 만약 공범자와 거짓 사용자가 서로 좋은 평가를 주고 받는다면 거짓 사용자의 초기 신뢰도가 높게 측정된다. 이에 따라 네트워크 구조 분석을 통해 그 사용자가 현재 구성하고 있는 이웃 관계를 살펴봐야 한다. 그 사용자와 이웃들이 맺고 있는 관계와 다른 사용자들에게 받은 직·간접적인 간선을 이용한다면 그 사용자의 실제 신뢰도를 측정할 수 있다. 이렇게 측정된 신뢰도는 거짓 사용자 판별, 친구 추천 서비스 등 다양한 애플리케이션에 활용될 수 있다. 이 밖에도 각 사용자의 사생활을 보장하면서도 소셜 활동을 유지할 수 있도록 하는 프라이버시 관리(privacy management) 기술, 소셜 네트워크 구조를 한 눈에 확인할 수 있는 시각화(visualization) 기술 등이 활발히 연구되고 있다.

21) 옥션 사이트들 또한 소셜 네트워크로 표현 가능하다. 예를 들어 이를 판매자와 구매자, 두 가지 타입의 정점으로 이루어진 다종 네트워크(heterogeneous network)로 표현할 수 있다.

## 4. 시공간 데이터

시공간 데이터(spatio-temporal data)란 시간의 흐름에 따라 공간적인 위치가 변화하는 데이터를 의미한다. 대표적인 시공간 데이터로는 이동 객체(moving object)가 있다. 최근 들어, 위성 항법 시스템과 이동 통신망의 발달로 인하여 선박, 자동차와 같은 이동 객체를 대상으로 다양한 위치 기반 서비스(location-based services)가 이루어지고 있다. 이러한 위치 기반 서비스들의 예로는 비상 구난, 차량 보안, 차량 항법, 그리고 교통 정보 안내 등이 있다. 이러한 서비스들을 제공하기 위해서는 위치정보를 관리하고 사용자의 이동객체 관련 질의를 효과적으로 처리할 수 있는 DB 기술들이 요구된다.

이동 객체 관련 질의를 처리하기 위해서는 우선 객체들이 위치하고 있는 공간의 특성을 파악해야 한다. 여기서, 공간은 이동 객체 움직임의 제약 유무에 따라 이동의 제약이 전혀 없는 공간과 이동이 제약된 공간으로 나눌 수 있다. 제약이 전혀 없는 공간의 대표적인 예로 유클리드 공간(euclidean space)이 있고, 제약이 있는 공간의 대표적인 예로는 네트워크 공간(network space)이 있다. 과거에는 주로 유클리드 공간을 대상으로 위치정보를 관리하는 기술들이 연구되었으나 근래에는 도로 위를 움직이는 자동차나 선로를 따라 움직이는 기차 등의 위치정보를 관리하기 위해 네트워크 공간을 대상으로 많은 기술들이 연구되고 있다. 네트워크 공간에서의 이동 객체 질의처리 기술은 질의 종류에 따라 (1) 정적 객체 질의처리 기술, (2) 이동 객체 질의처리 기술의 두 가지로 나눌 수 있다. 정적 객체 질의는 사용자가 관심이 있는 정적 객체를 대상으로 하는 질의이다. 예를 들어 현재 사용자의 위치로부터 가장 가까운 주유소를 찾는 질의가 이에 해당된다. 정적 객체 질의를 세분화하면 질의 점으로부터 일정 거리 안에 존재하는 영역 질의(range query), 질의 점으로부터 가까운 K개의 정적 객체들을 검색하는 인접 이웃 질의(nearest neighbor query), 그리고 두 정적 객체들 사의 거리가 가장 가까운 쌍을 검색하는 공간 조인 질의(spatial join query) 등이 있다.

도로 네트워크 공간에서 정적 객체 질의 처리에 대한 많은 연구들은 이동 객체의 현재 위치에서 가장 가까운 정적 객체를 검색하는 인접 이웃 질의에 초점을 맞추고 있다. 이러한 연구들은 인접 이웃 질의를 처리하기 위해서 먼저 도로 네트워크 안에서 질의를 하는 사용자의 위치를 파악한 후 사용자가 원하는 정적 객체가 가장 가까이 존재하는 도로를 검색한다. 이때 고려해야 될 점은 질의처리 비용과 저장 공간의 오버헤드이다. 질의처리의 비용을 감소하기 위해서 쓰이는 방법 중 하나는 정적 객체 간의 거리를 사전 계산(pre-computation)하여 저장하는 것이다. 정적 객체 간의 거리를 미리 계산하여 둔다면 질의 처리에 들어가는 비용을 크게 줄일 수 있다. 그러나 이러한 방식은 정적 객체의 수에 비례하여 사전 계산양이 크게 증가하기 때문에 질의 처리의 비용과 저장 공간 간의 트레이드-오프(trade-off)를 고려해야 한다.

이동 객체 질의는 차량, 항공, 선박 등의 이동 객체를 대상으로 하는 질의이다. 예를 들어 질의 영역과 시간이 주어졌을 때, 주어진 시간에 그 영역을 통과하는 자동차를 검색하는 질의가 이에 해당된다. 이동 객체 질의는 처리 대상이 되는 시간에 따라 과거 질의와 미래 질의로 구분할 수 있다. 현재 시간에 대한 질의는 과

거 질의 혹은 미래 질의를 위한 기술들과 같은 방식으로 처리할 수 있기 때문에 일반적으로 과거와 미래로 이동 객체 질의를 구분한다. 과거 질의는 특정 시간에 이동 객체의 위치를 묻는 질의이다. 과거 질의에는 다양한 세부 질의가 존재하며 각 질의에 따라 질의 처리 기술 또한 달라진다. 과거 질의에 관련된 이전 연구들은 특정 세부 질의 타입만을 대상으로 했기 때문에 다른 세부 질의 타입에는 상대적으로 떨어지는 성능을 보여줬다. 최근에는 다양한 타입의 과거 질의를 처리하기 위한 많은 연구들이 진행되고 있으며 이러한 연구들은 주로 여러 인덱싱 기법을 함께 사용하는 복합 인덱스 구조(complex index structure)를 사용한다.

미래 질의는 특정 시간대에 이동 객체의 위치를 예측하는 것이다. 과거 질의는 유클리드 공간과 네트워크 공간 양쪽에서 모두 활발히 연구되고 있다. 그러나 이와 달리 미래 질의에 대한 연구는 여전히 유클리드 공간을 위주로 진행되고 있다. 일방통행, 좌회전 금지, 유턴 가능 여부 같은 도로 네트워크의 다양한 제약 조건으로 인하여 이동 객체의 미래 위치를 예측하기가 어렵기 때문이다. 유클리드 공간에서의 미래 질의 처리 기술들은 이동 객체의 현재의 위치·속도·방향 등을 기반으로 이동 객체의 미래 위치를 예측한다. 최근에는 과거·현재·미래 질의를 모두 처리할 수 있는 질의 처리 기법들이 연구되고 있다. 그러나 이 연구들 역시 유클리드 공간을 기반으로 진행되고 있으며 도로 네트워크상에서의 미래 질의 처리에 대한 연구들은 여전히 미미한 상황이다.

이 밖에도 실내에서의 이동 객체의 위치를 관리하기 위한 RFID 기반 기술, 야생 동물들의 관리 및 보호를 위한 위치 추적 기술 등 다양한 시공간 DB 관련 기술들이 연구되고 있는 추세이다.

CHAPTER 3

# 데이터베이스 관련 기술

## 제1절 클라우드 컴퓨팅

### 1. 클라우드 컴퓨팅 기술

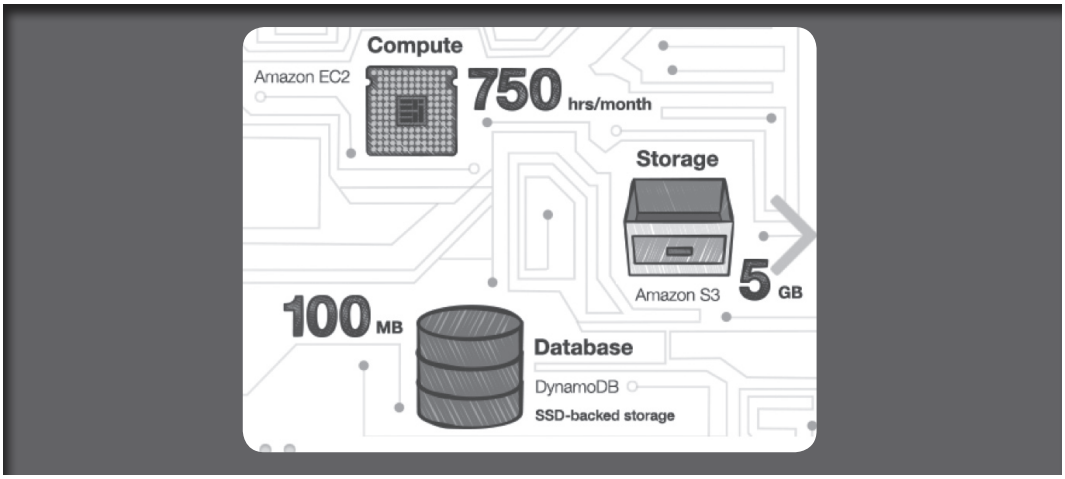
#### 가. 클라우드 컴퓨팅 개요

페타바이트(PB), 엑사바이트(EB)로 구성된 초 대용량 데이터를 처리해야 한다고 가정하면 고민에 빠지지 않을 수 없다. 수억 원에 달하는 고가 서버 여러 대를 구매하고 아마도 소형 데이터센터를 구축해야 할지도 모르며 이를 처리하기 위한 상용 소프트웨어, DB 엔진들도 다수 구입해야 할 것이다. 이를 위해서는 수 억, 수십억 원의 비용이 들어갈 것이므로 상당한 여유가 있는 기관이 아니라면 엄두를 낼 수 없을 것이다. 이에 대한 해법의 하나가 클라우드 컴퓨팅이다.

클라우드 컴퓨팅은 네트워크, 서버, 저장 장치, 애플리케이션, 서비스 등의 컴퓨팅 자원을 공유할 수 있도록 미리 준비해 두고 언제 어디서나 편리하게 필요한 만큼 네트워크를 통해 접근할 수 있도록 해주는 컴퓨팅 방식이다. 이를 테면 IT 장비 대여소라고 할 수 있다(그림 4-3-1). 앞에서 언급한 문제는 퍼블릭 클라우드로부터 서버, 스토리지, DBMS, 맵리듀스 솔루션을 임대해서 해결할 수 있다. 클라우드 컴퓨팅의 전신을 유틸리티 컴퓨팅(utility computing)이라고도 하였다. 말 그대로 전기, 수도, 가스처럼 컴퓨팅 파워를 가정이나 직장에서 필요한 만큼 인터넷을 통해 접속해 사용하고 사용한 만큼 요금을 지불한다고 본 것이다.

클라우드 컴퓨팅은 주문형 자가 서비스, 광역 네트워크 접근, 자원 풀링(pooling), 신속한 탄력성, 측정된 서비스를 특징으로 한다.

(그림 4-3-1) IT자원의 렌탈샵: 클라우드 컴퓨팅




※ 출처 : http://aws.amazon.com

클라우드는 특정 기관만을 위해서 운영되는 프라이빗 클라우드(private cloud), 다수의 기관에 의해서 공유가 되고 특정 커뮤니티를 위해 운영되는 커뮤니티 클라우드(community cloud), 클라우드 서비스를 판매하는 기관이 소유한 클라우드 인프라를 일반 대중이나 대형 산업 그룹에게 제공하는 퍼블릭 클라우드(public cloud), 이러한 형태가 혼재된 하이브리드 클라우드(hybrid cloud)로 구분된다.

## 나. 클라우드 컴퓨팅 도입 동기

클라우드 컴퓨팅의 가장 큰 장점은 컴퓨팅 자원 수요에 대한 유연성과 신축성(확장성)을 확보할 수 있다는 것이다. 또한 클라우드 컴퓨팅은 컴퓨팅 전력과 운영비용을 절감하는 효과가 있다. 퍼블릭 클라우드의 경우 컴퓨팅 자원을 직접 구매하는 대신 필요할 때 필요한 만큼의 컴퓨팅 자원을 빌려서 사용하고 사용료를 지불하는 방식으로 사용되므로 CAPEX(Capital Expenditure)를 OPEX(Operational Expenditure)로 전환하여 초기 투자 부담을 대폭 경감할 수 있다. 또한 예를 들어 4대의 서버를 가상화 기술을 이용하여 1대의 서버로 통합하는 경우 전력 소모량을 이론적으로 71%까지 절감할 수 있다.

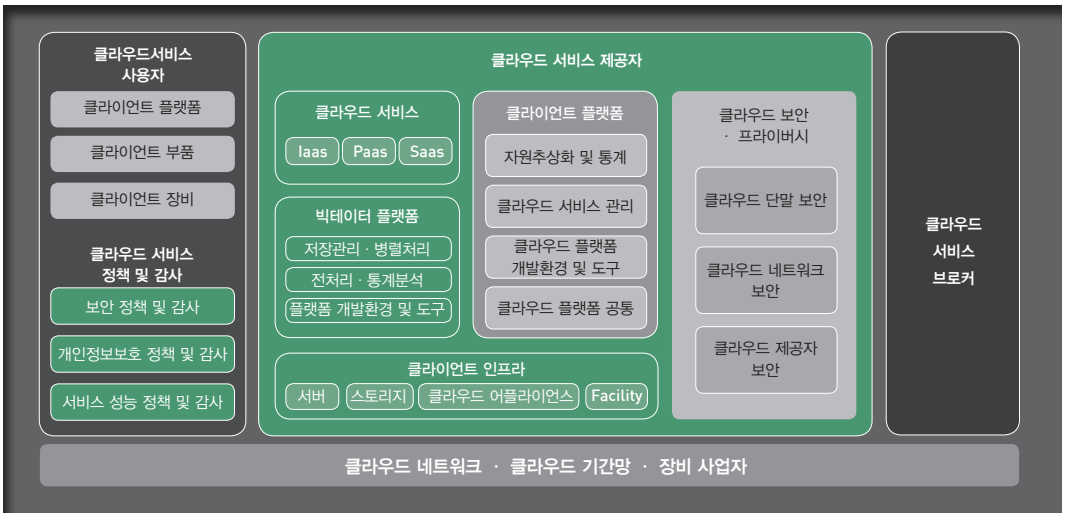
## 다. 클라우드 컴퓨팅 기술 스택

클라우드 컴퓨팅 환경은 클라우드 제공자, 클라우드 네트워크, 클라우드 단말, 클라우드 보안, 클라우드 서비스 브로커로 구성된다. 클라우드 단말(cloud client)은 클라우드 서비스를 이용하는 수단으로 스마트폰, 태블릿 PC, PC와 노트북, 씬 클라이언트, 제로 클라이언트 등을 예로 들 수 있다. 클라우드 서비스 브로커(cloud

service broker)는 이종의 클라우드를 연결 또는 중계해서 통합적으로 사용할 수 있도록 한다. (그림 4-3-2)는 NIST의 클라우드 컴퓨팅 참조 구조를 일부 수정한 클라우드 컴퓨팅 기술 스택을 보이고 있다.

클라우드 제공자(cloud provider)는 서비스, 플랫폼(클라우드 플랫폼과 빅데이터 플랫폼), 인프라로 구성된다. 클라우드 서비스(cloud service)는 클라우드 서비스 모델에 따라 SaaS, PaaS, IaaS 등으로 구분이 된다. 클라우드 플랫폼(cloud platform)은 가상화 등 자원 추상화 및 통제, 과금 및 운영을 위한 클라우드 서비스 관리, 클라우드 개발 환경 및 도구, 가용성 및 편의성 등을 제공하는 공통부분으로 구성된다. 빅데이터 플랫폼은 대용량 데이터를 실시간으로 처리 분석하기 위한 저장 및 병렬 처리, 전처리 및 통계 분석, 개발 환경 및 도구로 구성된다. 클라우드 인프라는 서버, 스토리지, 네트워크, 클라우드 어플라이언스, 이러한 클라우드 장비에 전력 및 냉방을 공급하는 설비(facility) 등으로 구성된다. 클라우드 어플라이언스는 서버, 스토리지, 네트워크, 가상화 소프트웨어 등을 통합 설치한 일체형 제품을 의미하며 VCE의 VBlock을 예로 들 수 있다.

**(그림 4-3-2) 클라우드 컴퓨팅 기술 스택**




※ 출처 : 한국산업기술평가관리원 차세대컴퓨팅PD실, NIST Cloud Reference Architecture 2011, 2012

## 라. 클라우드 컴퓨팅 솔루션

클라우드의 핵심 기술의 하나인 가상화(virtualization)는 하나의 물리적 서버에서 여러 대의 가상 서버, 즉 가상 머신(virtual machine : VM)을 실행시키는 기술이다. 이러한 가상 서버를 관리하는 소프트웨어를 하이퍼바이저(hypervisor)라고 한다. 가상화 솔루션으로는 VMware의 vSphere, Citrix의 XenServer, 마이

크로소프트 의 Hyper-V, 오라클의 VirtualBox, RedHat의 KVM 등이 대표적이다. 클라우드 운영체제는 수천, 수백만 대의 서버를 클러스터로 연결하여 고성능 컴퓨팅을 제공하는 기술로 오픈소스인 OpenStack 과 구글의 독자적 기술인 Googleware가 대표적이다. 국내의 경우 KT와 삼성SDS 등에서 오픈소스인 OpenStack을 이용하여 클라우드 구축을 진행한 바 있다. 클라우드를 기반으로 다수의 컴퓨터를 연결하여 수퍼컴퓨팅 또는 초고성능컴퓨팅(HPC) 파워를 제공하는 솔루션 기업도 등장하고 있다. 클루닉스의 RNTier는 공학용 설계 및 시뮬레이션을 위한 수퍼컴퓨팅 파워를 클라우드를 기반으로 제공하고 있다.

## 2. 클라우드 서비스

### 가. 클라우드 서비스 모델

클라우드 컴퓨팅은 3개의 서비스 모델을 가지고 있다. SaaS(Software as a Service)는 고객이 클라우드 인프라 상에서 실행되는 제공자의 애플리케이션을 사용하는 방식이다. PaaS(Platform as a Service)는 고객이 클라우드 인프라 상에서 제공자가 지원하는 프로그래밍 언어와 도구를 이용하여 고객이 생성하거나 획득한 애플리케이션을 사용하는 방식이다. IaaS(Infrastructure as a Service)는 고객에게 처리기, 저장 장치, 네트워크 등의 컴퓨팅 자원이 제공되며, 고객이 이를 이용해 운영 체제와 애플리케이션과 같은 임의의 소프트웨어를 사용하고 실행하는 방식이다.

전통적인 SaaS·PaaS·IaaS의 서비스 모델 외에 고객의 데스크톱, 즉 PC를 클라우드를 통해 제공하는 VDI(Virtual Desktop Infrastructure)와 유사한 서비스를 DaaS(Desktop as a Service)라고 하여 구분하기도 한다. DaaS는 인터넷 접속이 가능한 단말기만 있으면 언제 어디서나 클라우드 안에 존재하는 내 PC에 접속할 수 있게 되므로 DaaS 형태의 서비스를 '인터넷 안의 내 컴퓨터'라고 표현하기도 한다.

### 나. 상용 클라우드 서비스

상용 클라우드 서비스 업체로는 구글, 마이크로소프트, 아마존, Salesforce.com, 애플이 대표적이며, 국내는 SKT, KT, LG U+ 등의 통신업체와 네이버, 다음 등의 포털을 중심으로 개인용 스토리지 서비스 등을 제공 중이다. SaaS의 경우 GoogleApps, MS Office Live 등이 대표적이며, 국내 기업으로는 더존이 클라우드 기반의 회계관리 ERP로 1천억 이상의 매출을 올리고 있다. PaaS의 경우 Google AppEngine, Salesforce의 Force.com, MS Azure가 대표적이다. IaaS의 경우 아마존의 컴퓨트(compute) 서비스인 EC2(Elastic Compute Cloud), EMR(Elastic MapReduce), 스토리지 서비스인 S3(Simple Storage Service), EBS(Elastic Block Store), DB 서비스인 DynamoDB, RDS(Relational Database Service)가 대표적이며, 국

내에서도 통신업체를 중심으로 KT의 uCloud Pro, SKT의 T cloud biz 등의 IaaS 서비스가 출시되고 있다. 국내에서는 또한 애플의 iCloud와 유사하게 KT U 클라우드, SKT T 클라우드, LG U+ 박스, 네이버 N드라이브, 다음 클라우드와 같은 개인 스토리지 서비스 및 N-스크린 동기화 서비스가 비교적 활발하게 출시되고 있다. 빅데이터 분석의 경우 넥스알은 하둡을 이용한 대용량 데이터 처리, 다음소프트, 그루터는 하둡을 이용한 소셜 데이터 분석 서비스를 제공하고 있다.

아마존의 RDS를 이용하면 클라우드 방식으로 MySQL, Oracle, SQL Server, PostgreSQL DB 엔진을 이용할 수 있다. 빅데이터 분석을 위해 컴퓨팅 자원이 필요한 경우 큰 돈을 들이지 않고서도 아마존의 EC2에서 서버를 임대하고, S3에서 스토리지를 임대하고, EMR을 이용해 맵리듀스 프로그램을 실행할 수 있다.

## 3. 클라우드 보안

### 가. 클라우드 보안 이슈

클라우드 보안과 관련해서는 서비스 제공자의 전문적인 보안관리 및 통제 하에 있어서 더 안전하다는 주장과 자원공유, 가상화 등 클라우드 특성으로 인한 보안 위협이 증가한다는 주장이 상반적으로 존재한다. 클라우드 보안의 한 가지 문제는 가상화 취약성이다. 가상화 기술(하이퍼바이저)을 통해 이용자의 가상 머신들이 상호 연결되어 다양한 공격 경로가 존재한다. 하이퍼바이저 해킹으로 인한 통제권 상실, 가상 머신의 악성코드 감염 및 확산 가능성이 있다. 이러한 이유로 공용 클라우드 서비스 이용 약관에 가상 머신의 해킹 등에 대한 문제는 서비스 제공자가 아닌 고객이 책임진다는 형태가 많다.

IT 자원의 공유, 정보 집중화 등으로 인한 문제도 있다. IT 자원 공유, 멀티테넌시 환경에서 해킹 및 관리자 실수 등에 의해 이용자 정보의 유출이 가능할 수 있다. 이용자의 정보가 클라우드 서버 내 어디에 저장되고, 백업되고, 누가 접근하는지에 대해 알기도 어렵다. 클라우드 서버에 고객 정보가 집적되어 저장되기 때문에 해킹, DDoS 공격의 표적이 되기 쉽고, 사고 발생시 전체 이용자 서비스의 연쇄 중단 및 대규모 피해를 야기할 가능성이 있다. 대량 고객정보 집적, 타 고객사 정보가 혼재되어 발생하는 설정 오류, 취약한 패스워드 사용 등으로 인가되지 않는 외부 이용자의 정보 접근이 발생할 가능성이 있다.

사업자의 내부직원에 의한 권한 외 정보 접근 및 유출 가능성도 있다. 최근의 개인정보 유출 사고의 대부분은 권한을 가진 내부자에 의한 경우가 많은 편이다. 클라우드 내부 관리자에 의한 고의적 불법적 접근 위험은 주요 정보에 대한 접근의 로깅(logging), 관리자에 대한 최소한의 접근 권한 부여, 기밀 정보 누출 사고 등에 대한 책임을 묻는 명확한 계약 조건 설정 등을 통해 완화가 가능하다.

### 나. 클라우드 보안 기술

클라우드 관련 보안 기술은 소프트웨어·시스템·서비스 보증, 접근 및 식별 라이프사이클 관리, 데이터와 정보 방어, 거버넌스, 보안 권리, 데이터 정책 실현 등으로 구성된다. Trend Micro, Boxcryptor, McAfee, Symantec, VMware, Juniper, HP, Intel 등이 클라우드 보안 관련 제품을 출시하고 있으며, 안랩, 파수닷컴, 이글루시큐리티 등 국내 기업도 클라우드 보안 기술에 대한 연구개발을 진행 중이다.

## 제2절 데이터베이스 보안과 개인정보 보호

### 1. 개인정보 침해 현황 및 문제점

오늘날 사회가 지식 정보 사회로 빠르게 변화함에 따라 디지털화된 데이터의 양은 폭발적으로 증가하고 있으며, 이에 따른 개인정보 침해 사례 또한 함께 늘어나고 있다. 국내의 경우, 지난 2014년 1월 KB카드, 롯데카드, NH카드에서 총 1억 5백만 건에 달하는 개인정보가 유출된 사건이 발생한 데 이어 같은 해 3월 KT 홈페이지가 해킹당해 1천 2백만 명의 개인정보가 유출되어 사회적으로 큰 물의를 빚은 바 있다. 국외에서도 유통업체인 'Target'과 백화점 'Neiman Marcus'에서 각각 개인정보 1억 1,000만 건과 110만 건의 개인정보가 유출되어 해당 기업들의 신용도에 치명적인 피해를 입혔다.

〈표 4-3-1〉 국내외 개인정보 유출 사례

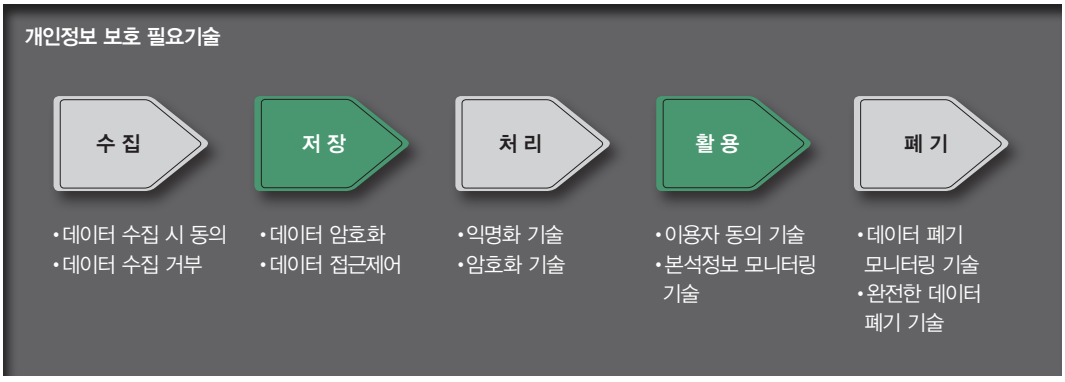
| 구분 | 기업명 | 발생일자 | 유출 내용 |
|---|---|---|---|
| 국내 | 옥션 | 2008년 2월 | 해킹을 통해 약 1,081만 명의 개인정보 유출 |
| | 네이트 | 2011년 7월 | 해킹으로 네이트온 가입자 개인정보 3,500만 건 유출 |
| | 국민,NH, 롯데카드 | 2014년 1월 | 1억 580만 건의 개인정보와 결제계좌 유출 |
| | KT | 2014년 3월 | 1,200만 명의 개인정보가 해킹으로 유출 |
| 국외 | 카드시스템즈 (CardSystems) | 2005년 6월 | 해킹으로 인해 4천만 명의 개인정보가 유출되었으며 이 여파로 2008년 퇴출 |
| | 티제이엑스 컴패니즈 (TJX Companies) | 2007년 1월 | 약 9,400만 건의 이름과 신용카드 정보 유출 |
| | 타겟(Target) | 2013년 11월 | 4천만 명의 신용카드와 체크카드 정보 유출 |
| | 니만 마커스 (Neima Marcus) | 2013년 12월 | 110만 명의 신용카드와 거래 정보 유출 |

이와 같은 개인정보 유출 위험은 스마트폰과 태블릿 PC의 대중화, IoT(Internet of Thing)의 보급이 진행됨에 따라 더욱 증가할 것으로 예상된다. 또한 데이터 수집 방법의 다양화와 본인이 인식하지 못하는 자동화된 데이터 수집의 증가, 여러 데이터 소스로부터 취합된 정보를 사용한 데이터 분석 능력의 향상에 따라 기존의 개인정보 보호 기술로는 제어하지 못하는 형태의 개인정보 침해 사례들이 등장하고 있다. 한 예로, ETRI는 페이스북 657만 개 계정과 트위터 277만 개 계정 등 한국인 SNS 계정 934만 개를 대상으로 개인정보 노출 현황을 분석한 결과, 트위터와 페이스북에 노출된 ID, 이름 등의 정보를 사용하여 17만 개 이상의 트위터 계정과 페이스북 계정을 연결 가능하다는 것을 보였다. 이처럼 서로 다른 SNS 계정을 연결할 경우 광범위한 개인정보 추론이 가능하나, 기존의 개인정보 보호 기법으로는 이를 방지할 수 없다. 따라서 변화하는 데이터 처리 환경에서의 개인정보 보호 요구사항을 만족시킬 수 있는 기술에 대한 필요성이 더욱 강조되고 있다.

## 2. 개인정보 보호 기술

데이터는 일반적으로 데이터 수집, 저장, 처리, 분석, 활용, 폐기의 생애 주기를 지닌다. Agrawal은 2002년 개인정보 노출을 방지하기 위하여 히포크라틱 DB(hippocratic database)라는 개념을 제시하였다. 히포크라틱 DB 개인정보 추론과 DB보안 문제를 해결하기 위하여 기존의 관계형 DB시스템에 데이터 생애주기를 반영한 개인정보 보호 기술을 통합하였다. 오늘날의 데이터 처리 기술에 히포크라틱 DB의 개념을 적용한다면 (그림 4-3-3)과 같다.

(그림 4-3-3) 데이터 생애주기에 따른 개인정보 보호 기술



## 가. 데이터 수집 단계

데이터는 데이터를 생성하는 주체가 데이터를 수집하는 주체에게 직접 데이터를 전달하는 능동적 데이터 수집과 데이터를 수집하는 주체가 크롤링 등의 방법으로 데이터를 수집하는 수동적 데이터 수집이 있다.

### 1) 데이터 수집 시 동의 기술

데이터 수집 시 개인정보가 포함되어 있는 데이터에 대해 해당 개인에 대한 동의를 받아야 한다. 특히 개인이 인지하기 어려운 수동적 데이터 수집의 경우 이를 지원할 수 있는 기술이 요구된다.

### 2) 데이터 수집 거부 기술

개인이 자신의 개인정보가 포함된 데이터가 수집되는 것을 원하지 않을 경우, 수집을 거부할 수 있는 기술이 필요하다. 현재 웹상에서 자신의 사이트 정보가 수집되는 것을 방지하기 위하여 robot.txt 파일에 수집 거부를 명시할 수 있으나, 권고안일 뿐 강제사항이 아니며 IoT의 발달은 일상적인 데이터 수집을 야기한다. 따라서 개인의 의사를 반영하는 데이터 수집 거부 기술에 대한 연구가 필요하다.

## 나. 데이터 저장 단계

데이터 저장 단계는 수집된 데이터를 저장 및 관리한다. 이 단계에서의 개인정보 보호를 위해서는 데이터의 암호화 및 사용자에게 부여된 권한에 따라 정보에 대한 접근을 제어하는 접근제어 기술이 필요하다.

### 1) 데이터 암호화

데이터 암호화는 데이터의 기밀성과 무결성을 보장해준다. 암호화 기술은 DB의 성능을 저하시키고 사용 가능한 질의의 범위를 제한하는 단점을 지니고 있어 이를 개선하기 위한 연구가 수행중이다. 암호화는 방식에 따라 공개키 암호화와 대칭키 암호화로 분류되며, 구현 방식에 따라 API 방식, 하드웨어 방식, 플러그인 방식으로 분류된다.

### 2) 데이터 접근제어

현재 데이터에 대한 접근제어는 임의적(discretionary) 접근제어, 강제적(mandatory) 접근제어, 역할 기반(role based) 접근제어가 주로 사용되며, 업무 구조를 반영할 수 있는 역할기반 접근제어가 기업 및 조직에서 일반적으로 사용되고 있다. 나아가 상황 변화를 인지하여 동적으로 접근 권한을 변경, 위임할 수 있는 상황 인식(context-aware) 접근제어 기술에 대한 연구도 진행 중이다.

## 다. 데이터 처리 단계

### 1) 익명화 기술

기존의 익명화는 개인 식별 정보(이름, 주민등록번호 등)를 제거하여 데이터로부터 특정 개인이 식별되는 것을 방지하고자 하였다. 그러나 개인 식별 정보 이외의 정보로부터 개인을 추론하는 것이 가능하게 되자, k개 이상의 동일한 데이터를 유지하여 특정인이 추론될 확률을 1/k 이하로 낮추는 k-익명화(k-anonymity) 기술이 제안되었다. k-익명화 이후로 민감한 데이터의 종류를 l개 이상 유지하는 l-diversity, k개에 속하는 민감한 데이터 간의 거리를 일정 수준 이상으로 유지하는 t-밀접성(t-closeness)와 같은 다양한 요구사항을 충족시키는 익명화 기술이 제안되었다.

### 2) 암호화 기술

데이터 처리 단계에서의 암호화는 복호화를 하지 않고 연산이 가능한 검색 가능 암호화(searchable encryption)와 동형 암호화(homomorphic encryption)에 대한 연구가 진행 중이다. 동형 암호화는 복호화로 인해 발생하는 데이터 노출을 근본적으로 차단할 수 있는 기술이나, 아직 해당 기술이 실제로 적용되기 위해서 많은 연구가 필요하다.

## 라. 데이터 활용 단계

### 1) 이용자 동의 기술

수집된 데이터 분석 결과는 개인정보를 침해할 수 있다. 예로 페이스북은 자체 데이터 분석을 통해 성적 취향의 차이에 따른 마케팅 서비스를 제공하여 개인이 공개를 원하지 않는 성적 취향을 노출시켰다. 따라서 데이터 분석 결과 사용 시 개인에게 해당 데이터 이용 여부에 대한 동의를 받아야하며, 데이터 분석에 따른 결과의 영향을 예측할 수 있는 기술이 제공되어야 한다.

### 2) 분석정보 모니터링 기술

이용자의 동의 하에 분석된 결과는 사용 내역에 대한 고지와 공개가 해당 개인에게 제공되어야 한다. 그러나 모든 정보에 대한 고지와 공개가 불가능할 수 있으므로 개인의 정보보호 요구 수준에 따른 단계별 모니터링을 제공할 수 있어야 한다.

### 마. 데이터 폐기 단계

#### 1) 완전한 데이터 폐기 기술

디지털 데이터는 한 번 저장된 후 사라지지 않는 영속적인 특성을 지니므로 데이터를 폐기하지 않고 보관하거나 복구 가능한 형태로 폐기할 경우, 추가 유출의 위험이 있다.

#### 2) 데이터 폐기 모니터링 기술

수집된 데이터는 수집 시 명시한 기간과 목적이 달성된 후 파기되어야 하므로 데이터 폐기 시 이에 대한 고지와 공개를 통한 모니터링이 이루어져야 한다.

## 3. 개인의 자기 정보 제어권

데이터 생애주기에 따른 개인정보 보호 기술은 기업에 의한 개인정보 침해를 방지할 수 있으나 지나치게 높은 수준의 개인정보 보호는 오히려 서비스 품질을 저하시킴으로써 개인과 서비스 제공자 모두에게 불이익을 가져온다. 따라서 개인정보 보호는 무조건적인 개인정보 보호 대신 기업에 의한 무분별한 정보 남용을 방지하면서 자신의 개인정보에 대한 변경, 삭제 등의 제어가 가능한 자기 정보 제어권을 보장하는 방향으로 나아가야 한다. 하버드 대학에서 처음 제안된 VRM(Vender Relationship Management)의 경우, 기업이 보관하는 고객 데이터를 개인이 직접 관리하는 아이디어에 기반하여 필요에 따라 기업이나 친구 등의 제3자에게 해당 내용을 공유할 수 있는 기술로 사용자 중심의 안전한 정보 공유를 지향하고 있다. 앞으로도 개인정보 보호 기술은 이와 같은 개인의 자기 정보 제어권을 보장하기 위한 방향으로 더욱 다양한 연구들이 수행될 것이며 이를 통해 보다 안전한 개인 데이터 활용이 가능해질 것으로 예상된다.

# 제3절 사물인터넷과 감성형 단말

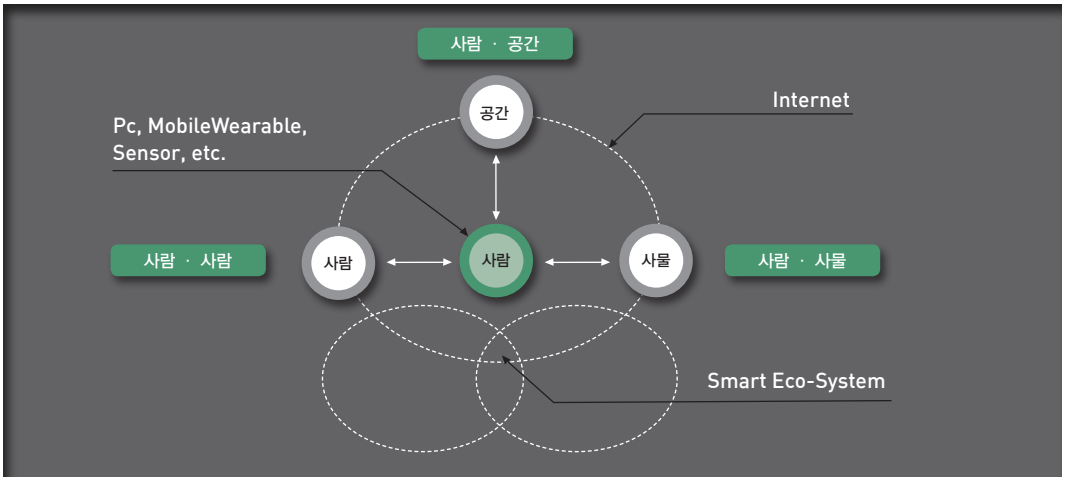
## 1. 개요

### 가. 정의 및 주요 특징

모든 사물들이 네트워크에 항상 연결되고 누구나 쉽게 사용할 수 있으며 상황에 따라 적절한 서비스가 이루어져야 한다는 유비쿼터스 컴퓨팅은 사물인터넷, 웨어러블 컴퓨팅, 빅데이터, 인공지능, 클라우드 컴퓨팅

등의 기술을 통해 현실화되고 있다. 사물인터넷(Internet of Things : IoT)은 1세대 PC 인터넷, 2세대 모바일 인터넷을 잇는 '3세대 미래 인터넷' 개념으로, 모든 사물이 인터넷으로 연결되는 환경을 의미한다.[22] 인간과 사물, 서비스 세 가지 분산된 환경 요소에 대해 인간의 명시적 개입 없이 상호 협력적으로 센싱, 네트워킹, 정보 처리 등 지능적 관계를 형성하는 사물 공간 연결망으로서 유무선 네트워크 디바이스, 인간, 차량, 교량, 문화재, 자연 환경을 구성하는 물리적인 사물은 물론 현실과 가상세계의 모든 정보와 상호작용하는 만물 인터넷(Internet of Eveything : IoE) 개념으로 확장되고 있다.[23] 사물인터넷을 통해 센싱된 정보를 기반으로 빅데이터 분석과 인공지능 기술을 활용하여 사용자에게 직접 적시(Just-in-time)에 서비스를 해야 하므로 보다 쉽고 편리하게 사용이 가능하고 안전한 인간중심의 단말 기술의 중요성도 함께 커지고 있다.

**(그림 4-3-4) 초연결사회에서의 웨어러블 디바이스**




※ 출처 : 김지은, 손용기, 손종무, 정현태, ICT 변화 흐름에 대응하는 웨어러블 디바이스의 역할, NIPA ICT 기획시리즈, 2014.4

사용자의 오감을 활용한 상황인지형 단말 사용 기술을 감성형 단말기술이라 한다.[24] 웨어러블 컴퓨터는 어디서나 사용할 수 있고 언제나 서비스가 가능하므로, 사물인터넷 환경에서 인간 중심의 서비스를 제공하는 감성형 단말의 대표적인 핵심 디바이스로 부상하고 있다. 기존 안경, 시계, 의류, 신발 등에 적용되어 특수한 목적과 서비스를 위한 제품들이 나오면서 이제는 '웨어러블 디바이스(wearable device)'라는 용어로 널리 쓰이고 있다. 웨어러블 디바이스란 안경, 시계, 의복 등과 같이 착용할 수 있는 형태로 된 컴퓨팅 기기로

22) 전현철, 2014년 IT산업 7대 메가트렌드 제2장 사물인터넷, 2014. 2.
23) 민경식, 사물 인터넷(Internet of Things), NET Term, 한국인터넷진흥원, 2012년, 심수민, 2014 웨어러블 디바이스 산업백서 비즈니스 수익 모델을 중심으로, 디지에코 보고서, 2014. 1.
24) 미래창조과학부, ICT R&D 중장기 전략, 2013. 10.

서, 궁극적으로는 사용자가 거부감 없이 신체의 일부처럼 항상 착용하고 사용할 수 있으며 인간의 능력을 보완하거나 배가시키는 것이 목표인 모든 기기를 말한다.[25] 웨어러블 디바이스는 사용자의 요구에 즉각적으로 반응해야 하고 기기 사용에 따른 안정성을 보장해야 하며 보기에도 좋아서 사회·문화적인 수용성을 가져야 하고, 보다 쉽고 직관적인 인터랙션 방법을 지원해야 하는 특징을 가지고 있다.[26]

〈표 4-3-2〉 웨어러블 컴퓨터의 기본 특성

| 기능 | 요구 내용 |
|---|---|
| 착용감 | 일상생활에서 사용하는 의복, 액세서리와 같이 착용을 의식하지 않을 정도의 무게감과 자연스러운 착용감 제공 |
| 항시성 | 사용자 요구에 즉각적인 반응을 제공하기 위하여 컴퓨터와 사용자 간 끊임없는 통신을 지원할 수 있는 채널 존재 |
| 사용자 인터페이스 | 인간의 신체적, 지적 능력의 연장선상에 있어야 하므로 사용자와의 자연스러운 일체감과 통합감 제공 |
| 안정성 | 장시간 착용에 따른 불쾌감과 신체적 피로감을 최소화하고 전원 및 전자파 등에 대한 안정성 보장 |
| 사회성 | 착용에 따른 문화적 이질감을 배제하고 사회 문화적 통념에 부합되는 형태와 개인의 프라이버시 보호 |

※ 출처 : 손용기, 김지은, 조일연, 웨어러블 컴퓨터 기술 및 개발 동향, 전자통신동향분석, 2008.10

사물인터넷 인프라 환경 속에서 웨어러블 디바이스는 인간의 생체 정보와 환경 정보를 바탕으로 적시에 적합한 서비스를 제공함으로써 삶의 질을 향상시키고 현실 문제를 해결하여 바로 대응하게 해 주는 기술로서 큰 의미를 갖는다.

## 나. 발전 추이

웨어러블 디바이스는 1960년대부터 구체화되어 여러 형태의 제품과 기술로 선보여져 왔지만 핵심 애플리케이션의 부재, 불편한 사용자 인터페이스, 기술적인 장벽에 부딪쳐 한 동안 크게 활성화되지 못한 채 침체되어 있었다. 2000년대에 접어들면서 스마트폰의 발전과 함께 부품의 소형화·저 전력화가 이루어지고, 음성인식, 상황 인지, 클라우드 컴퓨팅 등 주변 기술의 발전과 더불어 2013년 구글 글래스의 발표 이후 폭발적인 관심 속에 다양한 기술과 제품들로 새롭게 선보이고 있다. 웨어러블 기술은 초기 손목시계, 안경 형태의 액세서리형에서 의복에 전자기기가 일체화된 의류 일체형, 피부에 부착할 수 있는 패치형태의 전자회로인 신체 부착형, 그리고 생체에 삽입할 수 있는 생체 이식형으로 발전해 나가고 있다.[27]

25) 나연묵, 정현태, 최재훈, 웨어러블 컴퓨터의 현황과 전망, KEIT PD 이슈리포트, 2013. 6.
26), 27) 정현태, 2014년 IT산업 7대 메가트렌드 제1장 웨어러블 컴퓨터, 2014. 2.

(그림 4-3-5) 웨어러블 컴퓨터의 현황과 전망


※ 출처 : 나연묵, 정현태, 최재훈, 웨어러블 컴퓨터의 현황과 전망, KEIT PD 이슈리포트, 2013.6

건강한 삶에 대한 대중의 욕구에 따라 헬스케어, 피트니스 분야의 서비스를 목표로 한 스마트워치, 스마트 밴드와 같은 제품들이 이제야 상용화되고 있는 상황이며 더불어 몸에 붙이는 신체부착형 시스템의 관심도 높아지고 있다. 향후 디자인과 개성 표현에 대한 요구가 커지면 의류일체형 웨어러블 기술의 발전도 가속화 될 것으로 예상된다.

## 2. 기술 동향

과거에는 웨어러블 디바이스의 소형화와 완성도를 높이기 위한 하드웨어 기술이 주요 관심사였다면, 현재는 어떤 서비스를 제공하고 얼마나 편리하게 사용할 수 있으며 적시에 정보를 요구하고 제공받을 수 있는지에 관한 인터랙션 기술, 플랫폼 기술이 중요시되고 있다. 최근 십여 년간 ISWC(International Symphosium on Wearable Computers) R&D 사례를 분석한 결과를 보면 인지과학적 접근의 연구 논문이 중요시되고 있으며, 단순한 사용자 인터페이스(user interface : UI) 연구보다는 사용성(usability)을 향상시키고 평가 프로세스를 모델화하는 사례가 점점 많아짐을 알 수 있다.[28]

최근의 웨어러블 디바이스 제품들은 아직 해결해야 할 문제들이 여전히 남아 있음에도 하드웨어의 소형화, 집적화 기술은 상당 수준에 이르러 기업이 경쟁적으로 제품을 출시하고 있는 상태이다.

이런 가운데 초연결 사회에서 차별화되고 가치 있는 서비스를 위해 웨어러블 환경에 최적의 UI 솔루션을

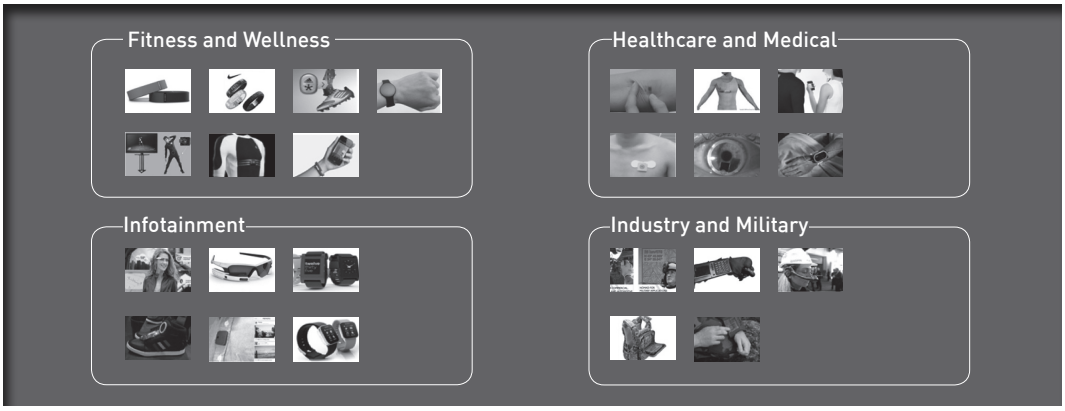
28) 이상국, 웨어러블 디바이스 산업 엔진을 위한 R&BD 7대 전략, KESSIA Issue Report, 2014. 3.

제공하는 소프트웨어 플랫폼 기술, 상황인지에 기반하여 언제 어디에서나 쉽고 빠르게 서비스를 가능하게 하는 기술 등이 새롭게 주목받고 있다. 특히 스마트폰에서 멀티터치 UI가 기기의 사용성과 콘텐츠의 소비에 큰 영향을 미친 것처럼, 웨어러블 디바이스에서도 쉽고, 직관적이고, 상황에 적합한 UI를 제공하는 핵심 기술의 개발이 지속된다면 웨어러블 디바이스 시장 영향력의 빠른 확장에 도움이 될 것으로 기대된다.[29]

## 3. 시장 동향

시장에는 이미 다양한 형태의 웨어러블 디바이스 제품들이 출시된 가운데, CES 2014와 MWC 2014를 통해 스마트 안경, 스마트 워치, 피트니스 밴드·이어폰·신발 등 다양한 기업이 다수의 제품들을 선보임으로써 시장 활성화에 대한 긍정적인 전망이 늘어나고 있다. 웨어러블 시장은 기존 시장을 유지·견인해 온 산업·군사 분야, 헬스케어·의료 분야와, 새로운 아이디어와 제품으로 신규 시장을 확대하고 있는 인포테인먼트 분야, 피트니스·웰니스 분야로 크게 나뉜다.[30] 이중 건강한 삶을 유지하고자 하는 대중의 욕구가 반영되어 헬스케어·의료 분야와 피트니스·웰니스 분야를 중심으로 한 성장이 당분간 지속될 것으로 예상된다.

**(그림 4-3-6) 시장에 출시된 웨어러블 디바이스 제품**




※ 출처 : 손용기, 웨어러블 컴퓨터 제품 및 기술 개발 현황, 광학세계, 2013.11

웨어러블 디바이스 시장의 규모는 매출액 기준으로 2016년에는 60억 달러로 성장할 것으로 IMS 리서치는 전망하고 있으며, 크레디트스위스는 2018년 300억~500억 달러까지 증가할 것으로 전망하고 있다.[31] 웨

29) 한국방송통신전파진흥원, 웨어러블 디바이스 동향과 전망, 방송통신기술 이슈&전망 제29호, 2013.
30) 정현태, 2014년 IT산업 7대 메가트렌드 제1장 웨어러블 컴퓨터, 2014. 2. 및 손용기, 웨어러블 컴퓨터 제품 및 기술 개발 현황, 광학세계, 2013. 11.
31) 전황수, 차세대 PC 웨어러블 디바이스 시장 및 개방 동향, NIPA ICT 기획시리즈, 2014. 3.

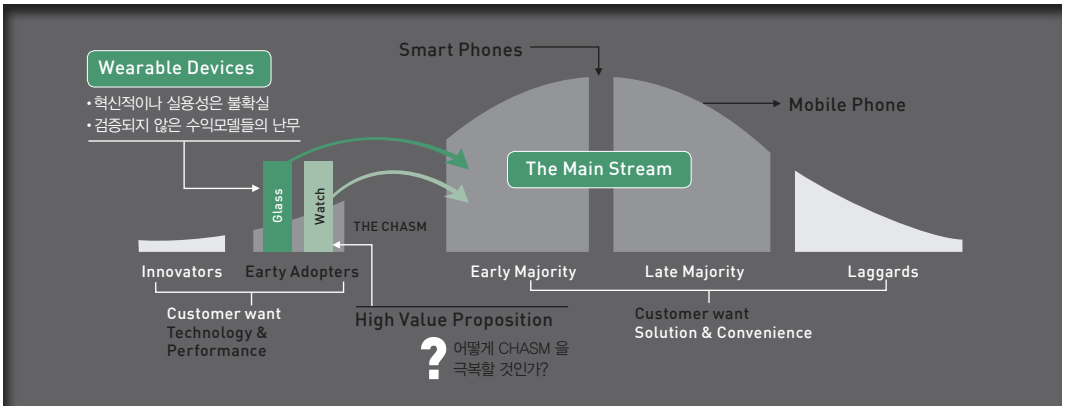
어러블 디바이스 출하량 기준으로도 ABI 리서치, 비즈니스 인사이더 인텔리전스는 2018년까지 각각 4억 8,500만대, 3억대를 기록할 것으로 전망하고 있다. 이처럼 조사 기관마다 큰 차이를 보이는 이유는, 웨어러블 디바이스 시장의 성장에 대해서는 이견이 없는 가운데 시장의 상황을 주시하면서 시장의 방향성과 크기의 결정을 2016년 이후로 보려는 조심스러운 예측 때문인 것으로 분석된다.

## 4. 향후 전망

사물인터넷의 연결 대상과 범위가 '사물과 사물'에서 '사람과 사물 그리고 공간(Internet of Everything : IoE)'을 넘어 '가상세계와 융합된 지능화된 만물 인터넷 세상(Intelligent IoE : IIoE)'으로 진화할 것이라는 예상 속에, 사용자와 항상 연결되어 있는 웨어러블 디바이스의 중요성이 그 어느 때보다 강조되고 있다.

다만, 웨어러블 디바이스 제품이 혁신성으로 환영을 받고 있지만 대중에게 확산되기 위해 위해서 넘어서야 할 캐즘(chasm)에 직면한 상황이기 때문에 이를 극복하기 위한 전략적인 접근이 중요한 시점이다.[32] 따라서 참신한 아이디어에 기반한 새로운 혁신 기술 개발을 통해 미래 시장의 경쟁력 확보에 힘씀과 동시에 기존 기술을 활용한 혁신적인 제품으로 대중의 인식 변화를 유도하는 노력을 병행해야 한다. 향후 웨어러블 디바이스는 생리적 욕구와 안전 욕구 중심의 신체보조 기능에서 개성 표출 수단으로 진화할 것으로 전망된다.[33]

(그림 4-3-7) 웨어러블 디바이스의 CHASM을 극복하기 위한 Value Proposition의 필요성




※ 출처 : 심수민, 2014 웨어러블 디바이스 산업백서 비즈니스 수익 모델을 중심으로, 디지에코 보고서, 2014.1

32) 심수민, 2014 웨어러블 디바이스 산업백서 비즈니스 수익 모델을 중심으로, 디지에코 보고서, 2014. 1.

33) DMC Media, 웨어러블 디바이스 시장의 현황과 전망, DMC Report, 2014. 4.

웨어러블 디바이스에서 여전히 부품, 배터리, 통신, 센서 기술 등의 동반 성장과 협력이 중요하지만, 웨어러블 디바이스 자체의 경쟁력을 갖추고, 보조 기기로서가 아닌 꼭 필요한 개인 단말로서 가치를 부여할 수 있는 기술과 서비스를 발굴하는 것이 무엇보다 중요하다. 따라서 혁신적인 기술 개발을 지속적으로 추진하고 서비스와 사용자의 특성을 반영한 디바이스, 킬러 콘텐츠와 디자인으로 경쟁력을 높이는 데 힘을 모아야 하며 이런 노력이 뒷받침될 때 웨어러블 디바이스는 사물인터넷, 빅데이터, 클라우드 컴퓨팅 등을 하나로 묶을 수 있는 최적의 연결 고리로서 역할을 할 것이다.

# 매력적인 데이터베이스 산업과 함께하기

**안 동 혁**

2013 DB솔루션 이노베이터 수상
(주)위세아이텍 상무

각종 통계 자료나 전문가들의 예측을 들어보면 IT 산업, 특히 데이터베이스 산업은 앞으로도 계속 성장하고 중요시될 전망이라고 한다. 단기적인 침체는 있을지라도 산업 자체는 계속 일할 사람과 기업을 필요로 하니 꽤 괜찮은 분야가 아닌가!

문제는 이 산업이 너무 매력적이라 경쟁이 치열한데다 변화무쌍하기까지 해서 어떻게 준비하고, 경쟁자들과 어떻게 싸워야 할지 모르겠다는 것인데, 이 매력적인 산업과 오랫동안 함께 하기 위한 방법을 나름대로 생각해 보고자 한다.

첫째, 산업이 주는 위험을 감수하되 사용하는 입장에서 충분히 검토하자. IT 산업은 모바일, 클라우드, 빅데이터, 사물인터넷과 같은 아이템을 계속해서 던져 주고 있고 이를 잘 이용해야 이 산업에서 살아남을 수 있다는 것을 알고 있다. 과거 개인화 추천·웹 로그 분석 솔루션 기업들이 반짝 나타난 후 대부분이 없어졌는데, 당시에도 웹 로그 분석을 통해 얻는 페이지뷰 수, 방문자 수 정보가 사실 고객에게 큰 의미가 없다는 인식이 있었다. 최근의 많은 소셜 분석 솔루션 기업들도 소수의 경쟁력 있는 기업을 제외하고는 대부분 사라지고 있다. 내가 잘하고 있는 분야를 기반으로, 가지고 있는 기술을 활용하지 않고 단순히 유행을 좇다가는 같은 신세가 될 것이다. 남들과 비슷한 기능에, 몇 번 보다보면 항상 비슷한 소셜 경향 분석 정보를(게다가 그리 정확하지도 않다) 제공하는 정도라면 과연 돈 주고 살 사람이 있을까?

둘째, 유행을 좇기 위해 새로운 기술을 끊임없이 살펴보되, 공짜 기술에 조심하자. 빅데이터를 준비하기 위해 하둡을 꼭 해야만 할까? SQL을 하둡 환경에서도 쓸 수 있도록 한 Impala, 맵리듀스를 사용하지 않는

Presto 분석 엔진 등이 나오고 있는데, 아직까지도 빅데이터 전문가 교육과정에서 Hive, Pig, MapReduce를 중요하게 배우고 있다. 어려운 기술을 힘들게 익히지 말고 더 쉽고 더 성능이 뛰어난 기술이 있는지 찾아보자. 이 바닥에서는 우수한 기술이라도 쉽게 이용되지 못하면 금방 없어져 버리곤 한다. 기술을 찾다보면 간혹 괜찮은, 심지어 공짜인 오픈소스라는 선물을 만날 수 있다. 하지만 진짜 공짜는 세상에 없다. 기업 솔루션 용도로 사용할 경우의 라이선스 비용, 각종 제약 사항을 따져보면 결국 오픈소스는 기회와 지식을 줄 뿐이지 쉽게 이용만 해서는 돈을 벌 수 있게 해주지는 않는다.

셋째, 내 영역을 기술이 아닌 영업으로 지키지 말자. 특정 업종에 대한 노하우가 다른 업종에도 응용되어 사업을 할 수 있다면 기술이 있는 것이지만 특정 고객, 특정 업종에만 한정적이면 기술보다는 영업에 의존하는 것일 수 있다. 최근 금융권이 어려워지면서 많은 IT 기업들도 덩달아 힘들어하고 있다. 지금은 그나마 큰 공공 IT 시장 규모가 앞으로 만약 줄어든다면 어떻게 될까?

넷째, 똑똑한 고객과 같이 가자. 까다로운 고객은 솔직히 싫다. 모든 환경을 다 지원해야 한다거나 자기 개인 취향에 맞게 버튼 모양이나 색상을 바꿔달라는 요구는 도움이 안 된다. 비즈니스 활용 측면에서 의견을 내는 똑똑한 고객의 요구는 어렵더라도 우리 솔루션을 발전시키는데 도움이 된다.

기상예보 데이터가 있을 때 분석에 어떻게 사용할까? 까다롭기만 한 고객은 정확하지 않다고 아예 쓰지 않거나 실상 보지도 않을 수십 개의 기상 분석 리포트를 만들어 달라고 요구할지도 모른다. 똑똑한 고객은 부정확한 예보 데이터로도 의미 있는 분석을 한다. 맑음으로 예보가 되었는데 실제 '맑았을 경우'와 '비가 올 경우'로 나누어 예상 방문객 증감률을 분석하면, 예보와 실제 차이에 따른 마케팅 시나리오를 만들어서 활용할 수 있다(우리는 시나리오 분석 기능을 추가한 솔루션으로 업그레이드하고 레퍼런스도 확보할 수 있다).

만일 모두가 정확한 예보 데이터를 가지고 있다면 분석에 따른 경쟁력을 어떻게 확보할 수 있을까? 부정확하고 부족한 상황에서의 분석이 경쟁력을 가져온다. 마찬가지로 제품 개발 인력도 충분하고, 만들면 잘 팔아주는 영업조직이 있다면 좋은 제품이 나올 수 있을까? 그전에, 그런 환경이라면 연구소장인 나를 필요로 할 것 같지 않다. 좀 부족한 환경에서 조금 더 노력해서 결과를 쥐어 짜내는 과정이 이 매력적인 산업과 오랫동안 함께 하기 위한 나만의 마지막 방법이다.