특집 03

# 빅데이터 개요와 관련 기술, 그리고 오라클의 지원 전략

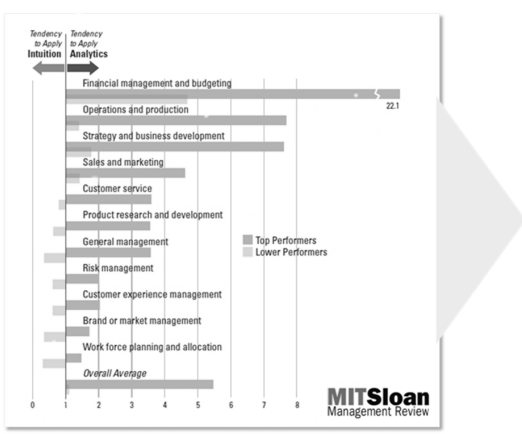
장성우 (한국오라클)

목 차 »



## 1. 서 론

2010년 MIT Sloan Management Review가 전+세계 100여개 국가의 3,000여명의 임원, 관리자 및 분석가를 대상으로 수행한 설문 결과에 따르면 Top-Performing 회사가 그렇지 않은 회사에 비해서 5배 이상의 분석능력을 갖고 있음이 나타났다. 또한, 응답자의 절반 이상이 정보분석시스템의 개선을 가장 중요한 우선과제로 선정하였고 응답자의 60%가 경쟁우위를 위한 혁신을 Top Business Challenge로 답변하였다. 여기서 재미있는 사항은 역시 60% 정도의 응답자가 현재보다 더 많은 가치있는 데이터가 회사내에 존재하고 있으며, 이에 대한 분석이 필요하다고 답변하였다는 점이다. 바로 이 부분이 빅데이터의 분석 및 처리가 중요한 이유가 된다. 최근에 빅데이터에 대한 관심이 커지고 있는 이유는 앞서 언급한 리뷰 결과와 같이 이제 정보 분석 능력이 기업의 경쟁력이 되고 있기 때문이며, 빅데이터의 관리 및 분석을 통한 정보 경쟁력의 향상이 그 해답을 제시해 줄 수 있기 때문이다. 이제 똑똑한 기업들은 분석 능력을 키워서 정보를 통찰력으로 전환시키고 이를 비즈니스 행동으로 바로 연결시키고 있음을 알 수 있다.



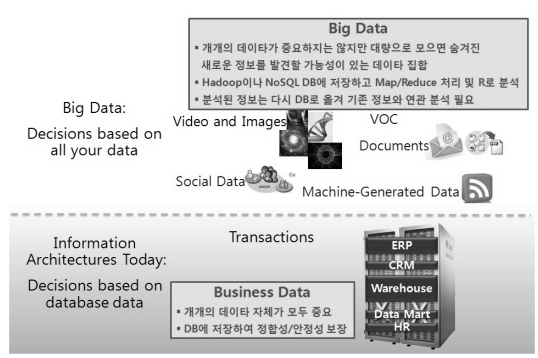

출처: MIT Sloan Management Review, 2010년 12월

(그림 1) 분석 능력 = 기업의 성과

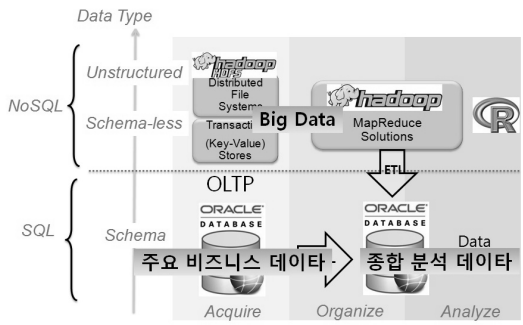
## 2. 빅데이터의 정의 : 오라클의 관점

이 세상에는 다양한 데이터들이 존재한다. 그 중의 일부는 DBMS(Database Management System)에 저장된다. 기업들이 DBMS에 데이터를 저장하는 이유는 '개개의 데이터 자체가 모두 비즈니스적으로 중요하기 때문'이다. 여기서 비즈니스적으로 중요하다는 것은 해당 데이터가 비즈니스 영속성을 위해 정확한 값으로 보존되어야 하며, 이를 위해 기업은 기꺼이 비용을 지불할 의사가 있음을 나타낸다. 이러한 데이터들의 예로는 고객 정보, 직원 정보, 판매/매출 정보 등을 들 수 있다. 이런 데이터들은 하나 하나가 소중하기 때문에 정확한 값을 가져야할 뿐만 아니라 또한 절대 유실되지 않도록 안전하게 보존되어야 한다. 따라서 우리는 이러한 데이터를 Business Data라고 부를 수 있다. Business Data는 업종별로는 달라지는데 제조업의 경우 제품, BOM, 생산계획, 설비, 출하, 물류 등이며, 통신업의 경우는 CDR, Billing, 상품 등, 금융업의 경우는 계좌, 대출, 투자, 자산 등의 데이터들이 Business Data가 된다. 이러한 Business Data는 DBMS에 저장됨으로써 ACID 특성을 지원받는다. ACID는 Atomicity/Consistency/Isolation/Durability를 나타내는데 동일한 데이터를 다수의 사용자가 동시에 사용(생성/수정/삭제)하여도 데이터의 값이 consistent하고 안전하게 보관/관리되는 것을 DBMS가 보증해 주는 특성을 말한다. 따라서, 이렇듯 하나하나가 비즈니스적으로 소중한 Business Data는 아무리 크기가 커도 반드시 DBMS에 저장해야만 한다.

하지만 더 많은 정보들이 DBMS가 아닌 다른 곳에 저장되거나 저장되지 못한채 사용되고 버려지고 있다. 이는 개개의 데이터가 발생은 되지만 비즈니스적으로 DBMS에 저장해서 관리할만큼 개개의 정보가 중요하지는 않다고 판단되거나 그에 따르는 비용을 지불할만큼이 중요성이 높지는 않기 때문이다. 하지만 앞서의 리뷰 결과는 이런 데이터들을 모아서 분석하여야 기업의 경쟁력을 키울 수 있음을 이야기하고 있다. 이런 고찰을 통해서 볼 때 많은 다른 정의 방법이 있겠지만 빅데이터는 '개개의 데이터가 비즈니스적으로 중요하지는 않지만, 대량으로 모으면 그 안에 숨겨진 새로운 정보를 발견할 가능성이 있는 데이터 집합'으로 정의할 수 있다. 예를 들면, 고객이 방문한 웹 페이지의 로그 정보의 경우 하나하나가 중요한 것은 아니지만 이를 대량으로 모아 분석하면 고객 유형별 관심 주제를 파악할 수 있게 된다. 비슷하게 상품의 이동에 따른 시간별 위치 정보를 대량으로 모아 분석하면 상품의 이동 경로의 효율성 파악할 수 있다. 이와 같은 유사한 사례를 많이 찾아볼 수 있는데 고객의 매장내 이동 동선을 모아 분석하여 주요 관심 제품, 결합 상품 파악, 제품/상점 추천 등에 활용할 수 있고, 포탈에서의 주요 검색 키워드를 통한 관심 트렌드 분석, 생산 설비의 상태/센서 정보(온도, 압력, 밀도 등) 분석을 통한 완성품 품질과의 연관성 분석, SNS를 통한 VOC 분석을 통해 고객의 만족도를 분석하는 것 등이 빅데이터를 통한 분석 및 활용의 사




(그림 2) 빅데이터와 Business Data의 구분




(그림 3) 빅데이터의 처리 단계 및 관련 솔루션

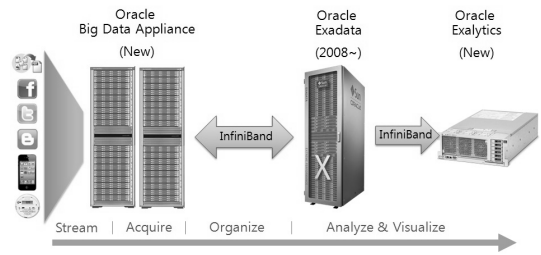
레들이 될 수 있을 것이다.

이제 분석된 개개의 정보는 비즈니스적으로 중요한 의미를 가지게 된다. 따라서 분석 정보는 기존의 분석 시스템인 DW로 저장하여 기존 정보들과 상호 연관분석을 수행하여 그 효율성을 높이도록 한다.

## 3. 오라클의 빅데이터 지원 전략

오라클은 이러한 빅데이터 관리를 지원하기 위하여 다양한 솔루션을 제공하고 있다. 오라클의 빅데이터 관련 기본적인 솔루션 전략은 빅데이터를 쉽게 관리할 수 있는 통합된 솔루션 아키텍처를 제공하고, 이를 기존의 DB Architecture와 유연하게 통합시킬 수 있도록 지원하는 것이다. 이를 위해 다음과 같은 제품들을 제공하고 있다 :

- 빅데이터 처리를 위한 새로운 통합된 솔루션 제공 → Big Data Appliance
- 중요한 비즈니스 데이터 처리에 대한 Extreme Performance와 Maximum Availability 제공 → Exadata
- DB 데이터와 빅데이터의 유연한 연결성 제공 → Big Data Connectors
- DB 내에서의 종합 분석 지원 → Advanced Analytics




(그림 4) 빅데이터를 지원하는 오라클의 Engineered Systems

- 메모리 기반으로 생각의 속도의 실시간 분석 제공 → Exalytics

즉, 오라클은 H/W와 S/W가 결합된 빅데이터 Appliance를 통해 빅데이터를 한 곳에서 쉽게 처리할 수 있도록 지원하고 또한 빠른 설치와 쉬운 운영 관리 기능을 제공한다. 그리고, 이를 기존의 Oracle DB 또는 Exadata와 쉽게 연결/통합하게 함으로써 새로운 빅데이터와 기존의 DB Data를 모두 포함하는 전사 데이터 관리 아키텍처를 쉽게 구현할 수 있도록 지원한다. 그리고 단일 벤더의 유지보수 지원을 통해 효율성과 안전성을 제공한다.

## 4. 오라클의 빅데이터 지원 솔루션

### 4.1 빅데이터 처리의 완벽한 플렛폼으로서의 Oracle Big Data Appliance

빅데이터 처리 시스템은 흔히 x86으로 구성된 하드웨어에 아파치 하둡(Hadoop)과 이를 보완하기 위한 서브프로젝트의 프로그램을 설치하여 구성한다. 그렇지만, 이들 서브 프로젝트는 각각 고유의 특성이 있어 때로는 기능이 중복되기도 하고 때로는 상호 연동이 불완전 할 때가 있어 각 특성 별로 이해하여 취사선택하여 효율적으로 시스템을 구성하는 데에는 많은 경험과 기술이 축

적되어 있어야 한다. 또한 오류가 발생할 경우 오픈 소스이다 보니 자체적으로 해결하거나 다른 사람이 해결할 때까지 기다리는 수밖에 없는 것이 현실이다. 그러므로 이를 자체적으로 구성하여 유지 보수 하는 것은 이러한 기술을 바탕으로 본업을 하는 일부 회사를 제외하고는 한계가 있다고 할 수 있다.

이러한 인프라 구성 및 유지 보수에 대한 해결책이 바로 Oracle Big Data Appliance이다. 오라클 Big Data Appliance는 빅데이터 프로그램을 구동하기 위한 하드웨어와 소프트웨어를 최적의 상태로 구성한 상태로 고객에게 전달 함으로서, 고객이 빅데이터 프로젝트를 즉시 수행할 수 있도록 하는 것을 그 목표로 한다. 또한 오류 및 장애 발생시 이에 대한 지원을 수행한다.

Oracle Big Data Appliance는 Full rack 1대 기준으로 864 GB의 메인 메모리와 648TB의 스토리지로 구성되어 있다. 주요 하드웨어 구성은 다음과 같다

- 18대의 노드로 구성, 각 노드의 서버구성은 다음과 같다
  - 2 CPUs (6-core Intel Processors)
  - 48 GB의 메인 메모리(96 GB 또는 144GB로 업그레이드 가능)
  - 12 X 3TB 디스크
- 인피니밴드 네트워킹
- 10 Gb 이더넷 연결

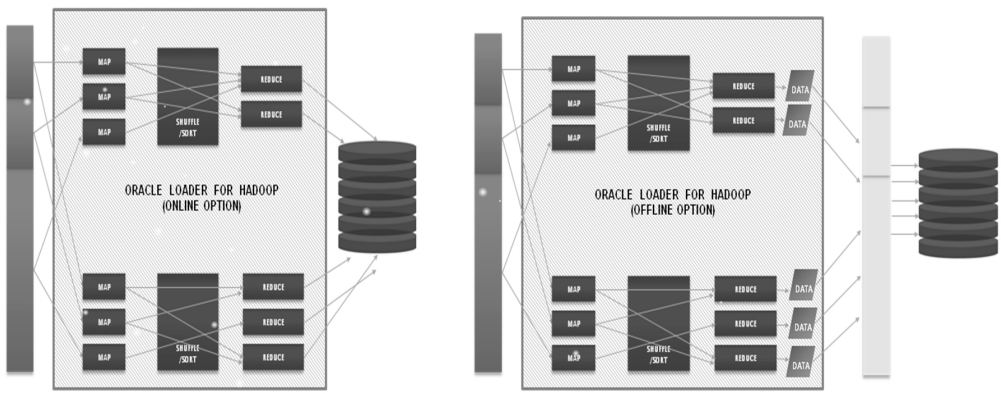
Oracle Big Data Appliance는 오픈 소스를 조합하여 구성된 시스템 소프트웨어와 오라클에서 개발한 빅데이터 Connectors를 포함한다. 주요 구성은 다음과 같다.

- Cloudera CDH(Cloudera's Distribution Including Apache Hadoop)
  - Hadoop Core
  - HDFS
  - Hive
  - HBase
  - Zookeeper
  - Oozie
  - Mahout
  - Sqoop
  - Cloudera Manager
- Oracle Linux 5.6
- Java HotSpot Virtual Machine
- Open Source R Distribution
- Oracle NoSQL Database CE
- Oracle Big Data Connectors

오라클은 클라우데라와 파트너쉽을 맺고 Oracle Big Data Appliance에 클라우데라의 아파치 하둡 배포판 CDH(Cloudera's Distribution Including Apache Hadoop)과 클라우데라 매니저(Cloudera manager) 솔루션을 포함하여 구성하였다. 이 CDH는 기업환경에서 적용 가능한 아파치 하둡의 100% 오픈 소스 배포판으로 안정성과 확장성 측면에서 뛰어난 평가를 받고 있다. Oracle Big Data Appliance는 이러한 클라우데라의 하둡 배포판에 오라클 리눅스, 오라클 Java Hotspot VM, Open Source R Distribution, Oracle NoSQL Database 및 Oracle Big Data Connectors를 설치하여 철저한 테스트와 검증을 통해 최적의 환경으로 구성한 시스템이다.

### 4.2 기존 인프라 스트럭쳐로의 연결을 위한 Oracle Big Data Connectors

하둡에 데이터를 저장하고 이를 MapReduce를




(그림 5) Oracle Loader For Hadoop(Online vs Offline)

이용해 분석하려고 하면 이를 Java와 같은 프로그래밍 언어로 개발해야 하는데, 이 경우 개발에 필요한 인력 및 시간이 많이 필요할 뿐만 아니라 분석가들이 프로그램 언어를 배워야 한다. 현재 대부분의 분석가는 R언어나 SQL을 구사할 수 있고 많은 분석 툴은 SQL을 기반으로 구성되어 있다. 그러므로 어느 정도 구조화된 데이터의 분석은 하둡 시스템에서 수행하는 것보다는 기존 오라클과 같은 데이터베이스와 연동하여 처리하는 것이 더 효율적일 것이다.

Oracle Big Data Connectors는 하둡에서 생성된 데이터에 접근하기 위한 다양한 방법을 제공하며 다음과 같은 구성요소를 가진다.

- Oracle Loader for Hadoop
- Oracle Direct Connector for Hadoop Distributed File System (HDFS)
- Oracle Data Integrator Application Adapter for Hadoop
- Oracle R Connector for Hadoop

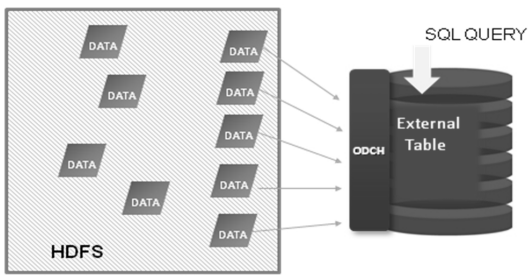
#### 4.2.1 Oracle Loader for Hadoop

하둡의 데이터를 오라클 데이터베이스로의 적재를 위해 최적화된 MapReduce 유틸리티로 이전 단계에서 생성된 Delimiter로 구분된 Text나 Hive 상의 테이블들을 Input 데이터로 하여 데이터를 읽은 후 MapReduce를 이용하여 오라클 데이터베이스로 적재시킬 수 있다. MapReduce작업이 시작되면 오라클 데이터베이스로 접속하여 테이블의 구조에 대한 정보를 얻은 후 각각의 테이블과 대응되는 파티션 단위로 Reduce작업을 완료 한다. 이때 정렬도 함께 수행하며 데이터의 형 변환이 필요한 경우 하둡 시스템의 리소스를 활용하여 미리 형 변환을 완료한 후 적재 작업으로 들어간다. MapReduce작업 시 오라클에 직접 적재하는 Online 방식과 적재 가능한 형태로 파일을 생성해 놓는 Offline방식을 지원하며 파일로 저장 시 datapump 포맷의 파일로도 저장 가능하다.

Oracle Loader for Hadoop의 기능은 Oracle Big Data Appliance 상에서 만 수행되는 것이 아니라 일반 하둡으로 구성된 시스템에서도 오라클과 연동하고 싶다면 설치하여 구동 할 수 있다.

#### 4.2.2 Oracle Direct Connector for Hadoop Distributed File System (HDFS)

HDFS 상의 파일을 오라클 데이터베이스로의 External table을 이용하여 직접 읽는 방식이다.




(그림 6) Oracle Direct Connector for HDFS

이 방식은 하둡 시스템에서 생성된 결과를 업무상 데이터베이스로 적재할 필요가 없거나 데이터베이스 내에 공간이 없을 경우에 사용되며 필요한 데이터만 선별하여 적재할 수 있다. 데이터는 SQL을 이용해 질의할 수 있어 오라클 내의 다른 테이블과 조인도 가능하다.

#### 4.2.3 Oracle Data Integrator Application Adapter for Hadoop

하둡 클러스터로부터 오라클 데이터베이스로 데이터를 추출, 변경, 적재를 수행한다. 이에 대한 정의는 GUI(Graphical User Interface)를 이용한다.

#### 4.2.4 Oracle R Connector for Hadoop

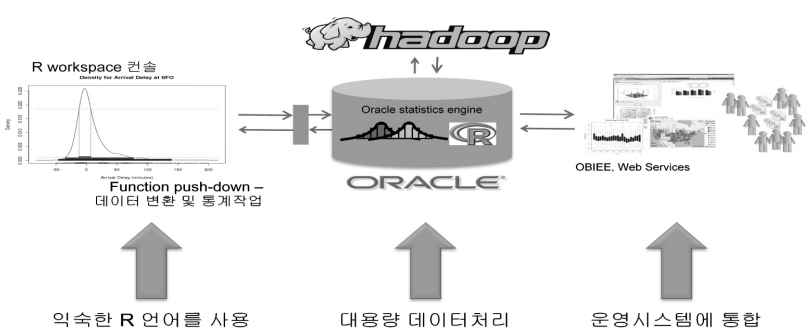
로컬의 R 환경에 오라클 데이터베이스, 하둡 간에 인터페이스를 제공함으로 이들 세가지 플랫폼 상의 데이터를 이용하여 분석 할 수 있다. 이를 이용하면 R 사용자는 하둡 상의 데이터를 이용하기 위해 하둡 환경이나 새로운 프로그램 언어를 배울 필요 없이 R언어를 이용 해 분석 할 수 있다.

### 4.3 빅데이터 분석을 지원하는 Oracle Advanced Analytics

Oracle Advanced Analytics는 오라클 데이터베이스 옵션으로 통계나 데이터 마이닝에 필요한 기능을 제공하며 Oracle R Enterprise와 Oracle Data mining으로 구성되어 있다.

#### 4.3.1 Oracle R Enterprise

오픈 소스인 R의 환경과 언어를 오라클 데이터베이스 11g에 통합 함으로서 분석가나 통계학자들이 기존의 R 로 작성된 어플리케이션을 재사용 가능하다. 오픈 소스 R이 데스크 탑에서 수행됨으로써 존재하였던 메모리크기 한계나 CPU의 처리능력의 한계도 분석 작업을 오라클이 설치된 엔터프라이즈 서버에서 수행하게 함으로서 극복되었다. 뿐만 아니라 데이터가 있는 서버에서 직접 수행함으로 불필요하게 데이터를 전송할 필요가 없고 오라클의 병렬처리 기능을 이용할 수 있게 되어 빠른 성능으로 데이터를 처리할 수 있게 되었다.



#### 4.3.2 Oracle Data Mining (ODM)

Oracle Data Mining(ODM)을 이용하면 예측 분석이나 통찰력이 필요한 차세대 응용프로그램을 편리하게 개발이 가능하다. 응용프로그램 개발자는 오라클 데이터베이스 내 데이터들에 대해 자동적으로 발굴하는 ODM의 SQL API를 이용할 수 있고 이 과정에서 사용되는 데이터나 모델 및 결과를 오라클 데이터베이스 내부에서 저장함으로서 불필요하게 데이터를 이동시키거나 중요 데이터가 유출되는 보안상의 문제를 방지 할 수 있다. 데이터 분석가는 Oracle Data Miner 11g Release 2의 그래픽 사용자 인터페이스를 이용하여 데이터의 패턴이나 관계 및 숨겨진 내용에 대한 통찰을 할 수 있다.

#### 4.3.3 Oracle R Distribution

Oracle R Distribution은 오픈 소스 R의 배포판에 x86 하드웨어상에서 고성능의 수치 계산을 위한 인텔의 MKL라이브러리를 확장한 것이다. 오라클은 R 소프트웨어를 지지하고 있고 오픈 소스 R에 기업수준의 지원을 제공할 계획이다.

### 4.4 데이터 관리를 위한 Oracle Engineered System 제품군

빅데이터라는 신개념과 이에 수반되는 technology에 접하게 되면 마치 모든 데이터 처리를 이 새로운 개념의 시스템이 대체하여 처리하는 것으로 생각하는 오류를 범하게 된다. 그렇지만 빅데이터 시스템은 기존의 인프라를 대체하는 것이 아니라 기존 시스템으로 저장 및 처리가 힘들었던 데이터를 처리하여 의미 있는 데이터를 추출하는 시스템으로 기존 인프라에 추가되는 것이다. 이를 그림으로 나타내면 (그림 4)와 같이 인프라 영역의 앞 단에 빅데이터 처리 시스템이 추가 되어 빅 데이터의 취득/저장, 구조화 및 분석하는 일련의 과정을 수행하게 된다. 정제된 데이터는 빅데이터 시스템에서 분석되거나 기존의 데이터웨어하우징 시스템에 적재되어 분석 시스템을 이용하여 분석할 수 있다.

#### 4.4.1 Oracle Exadata

Oracle Exadata는 스토리지, 서버, DBMS를 통합하여 최적화함으로서 DW와 OLTP 업무 구분 없이 통합 가능한 데이터베이스 전용 시스템이다. 대규모 병렬아키텍처를 사용해 데이터베이스 서버와 스토리지 간의 데이터 대역폭을 높였고 지능형 스토리지 소프트웨어를 이용하여 오라클의 질의의 일부를 스토리지 노드에서 수행 하여 유효한 데이터만을 접근하며 한번 읽은 정보는 최대한 캐쉬함으로서 처리 속도를 획기적으로 향상시킬 수 있는 엔지니어된 시스템이다. 이밖에 선형 확장성과 미션 크리티컬한 안정성을 제공할 수 있다.

#### 4.4.2 Oracle Exalytics

엑사리틱스는 비즈니스인텔리전스(BI) 어플라이언스로 디스크 없이 메모리에서 데이터를 분석하는 인메모리 기술이 적용된 제품이다. 여기에 시각화 기능 및 성능 최적화를 제공하는 오라클의 BI 파운데이션과 분석 성능이 확장된 타임스텐 인메모리 데이터베이스 최적화 버전, 에스베이스(ESSBASE) 등의 소프트웨어로 구성돼 있다.

## 5. 결 론

우리는 빅데이터 활용을 논하기 위해 먼저 빅데이터와 Business Data를 명확히 구분할 수 있어야 한다. Business Data는 개개의 데이터 모두가 비지니스적으로 소중한 데이터이며, 이런 데이터

들은 아무리 크기가 커도 DBMS에서 저장해서 ACID 특성을 지원받아야만 한다. 반면 빅데이터는 개개의 데이터가 비즈니스적으로 중요하지는 않지만 대량으로 모으면 그 안에서 의사결정에 도움이 될 수 있는 어떤 분석 정보를 얻을 수 있는 데이터를 말한다. 하나하나가 중요하지는 않기 때문에 DBMS에 저장하지는 않지만 대량으로 모아 분석해야만 하므로 Hadoop과 같은 분산화일 처리 시스템에 저장하고 R과 같은 통계 언어를 이용하여 분석한다.

오라클은 이러한 빅데이터 관리를 지원하기 위하여 빅데이터 처리를 위한 새로운 통합된 솔루션으로서 빅데이터 Appliance를, 중요한 비즈니스 데이터 처리에 대한 Extreme Performance와 Maximum Availability를 제공하기 위해 Exadata를, DB 데이터와 빅데이터의 유연한 연결성을 제공하기 위해 빅데이터 Connectors를, DB 내에서의 종합 분석 지원을 위해 Advanced Analytics를, 그리고 생각의 속도의 실시간 분석 제공을 위해 Exalytics를 통합 솔루션으로 제공한다.

## 저 자 약 력

**장 성 우**

이메일 : Sungwoo.chang@oracle.com

- 1990년 서강대학교 전자계산학과(학사)
- 1992년 서울대학교 컴퓨터공학과(석사)
- 1994년 서울대학교 컴퓨터공학과(박사 수료)
- 2005년 서강대학교 MBA
- 1997년~현재 한국오라클 근무중 (기술컨설팅 상무)
- 관심분야 : 빅 데이터, BI/DW, 경영과학, 정보분석