제 1 부

빅데이터의 기초

빅데이터는 비즈니스의 성격을 바꿀 수 있는 능력을 갖추고 있다. 실제로, 빅데이터만이 제공할 수 있는 인사이트(insight)를 생성하는 능력을 기반으로 하는 기업이 많다. 제1부에서는 주로 비즈니스 관점에서 빅데이터의 핵심 요소에 대해 다룬다. 기업은 빅데이터를 단지 기술로서만이 아니라 이를 통해서 어떻게 조직을 발전시킬 수 있는지에 대해서 이해해야 한다.

제1부는 다음과 같이 구성된다.

- 제1장에서는 수준 높은 비즈니스 인사이트를 제공하기 위해서 빅데이터의 본질과 가능성을 정의하는 주요 개념과 용어에 관해 설명한다. 분석 기술에 따른 다양한 데이터 유형의 정의와 더불어 빅데이터 데이터 세트를 구별하는 다양한 특성에 관해 설명한다.

- 제2장에서는 시장과 비즈니스 세계의 근본적인 변화 때문에 기업이 빅데이터를 도입해야만 하는 이유에 대한 답을 찾는다. 빅데이터는 비즈니스 변화와 관련된 기술이 아니다. 기업이 빅데이터로부터 얻은 인사이트에 따라 행동하는 경우에 혁신을 가능하게 한다.

- 제3장에서는 빅데이터가 평소와는 완전히 다른 일이라는 것을 보여주고, 빅데이터를 도입하기로 한 경우 고려해야 할 비즈니스 및 기술 사항들을 다룬다. 이는 빅데이터가 적절히 관리되어야 하는 외부 데이터의 영향에 노출되도록 한다는 점을 강조한다. 마찬가지로 빅데이터 분석 수명주기 때문에 독특한 데이터 처리 요구사항이 발생한다.

- 제4장에서는 기업의 데이터 웨어하우스(data warehouse) 및 비즈니스 인텔리전스(Business Intelligence, BI)에 대한 접근법을 살펴본다. 빅데이터 저장소 및 분석 자원은 기업의 분석 기능을 확장하고 비즈니스 인텔리전스에 의해 제공된 인사이트를 강화하기 위해 기업 성능 평가 도구와 함께 사용될 수 있다는 점을 보여주기 위해 이 방법을 확장한다.

올바르게 사용된 빅데이터는 비즈니스 내부 데이터가 모든 해답을 제공하지 않는다는 전제하에 만들어진 전략적 계획의 일부이다. 다시 말해서 빅데이터는 단순히 기술로 해결할 수 있는 데이터 관리 문제가 아니다. 빅데이터는 빅데이터, 분석 기술, 처리 기술의 조합을 통해 해결할 수 있는 비즈니스 문제에 관한 것이다. 이러한 이유로 비즈니스 중심의 제1부는 기술 중심의 제2부를 위한 기초를 제공한다.

제1장

빅데이터의 이해

- 개념과 용어
- 빅데이터 특성
- 다양한 유형의 데이터
- 사례연구 배경

빅데이터는 서로 다른 소스에서 자주 발생하는 대규모 데이터를 분석, 처리, 저장하는 분야이다. 빅데이터 솔루션은 일반적으로 기존의 데이터 분석, 처리, 저장 기술 및 기법들이 충분하지 않을 때 필요하다. 특히, 빅데이터는 서로 관련 없는 여러 데이터 세트의 결합, 대규모 비정형 데이터의 처리 및 숨겨진 정보 수집 등을 주어진 시간 안에 처리하는 것과 같이 요구사항이 뚜렷한 경우를 다룬다.

빅데이터가 새로운 분야로 보일 수 있지만 사실 수년 동안 발전해 왔다. 대규모 데이터 세트의 관리 및 분석은 초기 인구조사의 노동 집약적인 접근에서부터 보험료를 산정하는 보험회계 과학에까지 이르는 오랜 문제였다. 빅데이터 과학은 이러한 뿌리에서부터 진화했다.

빅데이터는 통계를 기반으로 하는 기존의 분석 방식 외에도 컴퓨터 자원과 분석 알고리즘 실행 방법을 활용하는 새로운 기법들을 추가로 사용한다. 이러한 변화는 데이터 세트가 계속해서 커지고, 더욱 다양하고 복잡해지며, 스트리밍 중심적으로 변함에 따라 중요해졌다. 고대로부터 인구를 추정하기 위해 표본을 추출한 통계적 접근법은 계속해서 사용되어 왔지만, 컴퓨터 과학의 진보로 인해 전체 데이터 세트의 처리가 가능해지면서 이러한 표본 추출이 필요 없어졌다.

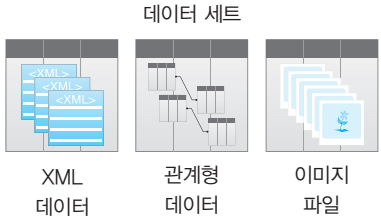
빅데이터 데이터 세트 분석은 수학, 통계학, 컴퓨터과학 및 주제 관련 전문 지식을 결합

한 학제 간의 노력이다. 이러한 기술과 관점의 혼합은 빅데이터와 그 분석 분야를 구성하는 것이 무엇인지를 혼란스럽게 만들었다. 이는 응답자의 관점에 따라 질문에 대한 답이 달라지기 때문이다. 빅데이터 문제는 소프트웨어 및 하드웨어 기술의 끊임없는 변화와 발전으로 인해 변화하고 있다. 이는 빅데이터의 정의가 데이터 특성이 솔루션 환경 설계에 미치는 영향을 고려했기 때문이다. 30년 전에는 1기가바이트의 데이터가 빅데이터 문제가 될 수 있었으며 특수 목적의 컴퓨팅 자원이 필요했다. 이제는 기가바이트의 데이터는 보편적이며 주변에서 쉽게 찾을 수 있는 기기에 의해 쉽게 전송, 처리 및 저장될 수 있다.

빅데이터 환경 내의 데이터는 일반적으로 애플리케이션, 센서 및 외부 소스를 통해 기업 내에 축적된다. 빅데이터 솔루션으로 처리된 데이터는 기업 애플리케이션에서 직접 사용할 수도 있고 기존 데이터를 풍부하게 하기 위해 데이터 웨어하우스(data warehouse)에 공급할 수도 있다. 빅데이터 처리를 통해 얻은 결과는 다음과 같은 광범위한 인사이트와 이점을 끌어낼 수 있다.

- 운영 최적화
- 실행 가능한 지능
- 새로운 시장의 식별
- 정확한 예측
- 장애 및 사기 탐지
- 보다 자세한 기록
- 향상된 의사결정
- 과학적 발견

분명한 건 빅데이터의 응용과 잠재적 이익이 광범위하다는 것이다. 그러나 빅데이터 분석 방법을 채택할 때 고려해야 할 문제가 많다. 정보에 입각한 의사결정 및 계획을 수립하기 위해서 이러한 문제를 이해하고 예상 이익에 비중을 두어야 한다. 이러한 주제는 제2부에서 별도로 논의할 것이다.



▲ **그림 1.1** 데이터 세트는 다양한 형식으로 나타난다.

개념과 용어

시작하기에 앞서 몇 가지 기본 개념과 용어를 정의하고 이해할 필요가 있다.

데이터 세트

관련 데이터의 모음이나 그룹을 일반적으로 데이터 세트라고 한다. 각 그룹 또는 데이터 세트의 구성요소(자료)는 같은 데이터 세트 내의 다른 구성요소들과 같은 속성 또는 성질을 공유한다. 데이터 세트의 몇 가지 예는 다음과 같다.

- 플랫 파일에 저장된 트윗
- 디렉터리에 있는 이미지 파일 모음
- CSV 파일에 저장된 데이터베이스 테이블에서 추출된 행
- XML 파일로 저장된 과거 날씨 관측치

그림 1.1은 3개의 데이터 형식을 기반으로 하는 3개의 데이터 세트를 보여준다.

데이터 분석

▲ **그림 1.2** 데이터 분석을 나타내는 데 사용되는 기호

데이터 분석(Data Analysis)은 사실, 관계, 패턴, 인사이트, 트렌드(trend)를 찾기 위해 데이터를 검토하는 과정이다. 데이터 분석의 전반적인 목표는 더 나은 의사결정을 지원하는 것이다. 데이터 분석의 간단한 예로는 아이스크림 판매 데이터의 분석을 통해 판매된 아이스크림콘의 수가 일일 온도와 어떤 관계가 있는지 알아내는 것을 들 수 있다. 이러한 분석의 결과는 일기예보 정보와 관련하여 아이스크림을 얼마나 주문해야 하는지

에 대한 의사결정을 뒷받침할 것이다. 데이터 분석은 분석 중인 데이터 간의 패턴과 관계를 수립하는 데 도움이 된다. 그림 1.2는 데이터 분석을 나타내는 데 사용되는 기호를 보여준다.

▲ **그림 1.3** 데이터 애널리틱스를 나타내는 데 사용되는 기호

데이터 애널리틱스

데이터 애널리틱스(Data Analytics)는 데이터 분석을 포괄하는 더 광범위한 용어이다. 데이터 애널리틱스는 수집, 정리, 구성, 저장, 분석 및 데이터 관리를 포함하는 데이터 수명주기 전체를 관리하는 분야이다. 이 용어는 분석 방법, 과학적 기법 및 자동화된 도구의 개발을 포함한다. 빅데이터 환경에서 데이터 애널리틱스는 다양한 소스의 대용량 데이터를 분석할 수 있는 확장성이 뛰어난 분산 기술 및 프레임워크를 사용하여 데이터 분석을 수행할 수 있게 해주는 방법을 개발했다. 그림 1.3은 애널리틱스를 나타내는 데 사용되는 기호를 보여준다.

빅데이터 분석 수명주기는 일반적으로 대량의 미가공 데이터, 비정형 데이터를 식별, 조달, 준비 및 분석하여 패턴을 식별하고, 기존 기업 데이터를 풍부하게 하여 대규모 검색을 할 수 있게끔 하는 의미 있는 정보를 추출한다.

서로 다른 조직은 데이터 애널리틱스 도구와 기술을 다른 방식으로 사용한다. 예를 들어, 다음의 세 분야가 있다.

- 비즈니스 중심 환경에서 데이터 애널리틱스의 결과는 운영 비용을 낮추고 전략적 의사결정을 쉽게 한다.
- 과학적 영역에서 데이터 애널리틱스는 현상의 원인을 파악하여 예측의 정확성을 높일 수 있다.
- 공공 부문 조직과 같은 서비스 기반 환경에서 데이터 애널리틱스는 비용을 줄임으로써 고품질 서비스 제공에 주력할 수 있다.

데이터 애널리틱스는 과학적 뒷받침을 통해 데이터 중심의 의사결정이 가능하게 하므로 과거의 경험이나 직관만으로가 아니라 실제 데이터를 바탕으로 의사결정을 내릴 수 있게 해준다. 생성되는 결과에 따라 애널리틱스에는 4개의 일반적인 분석 범주가 있다.

- 서술(descriptive) 분석
- 진단(diagnostic) 분석
- 예측(predictive) 분석
- 처방(prescriptive) 분석

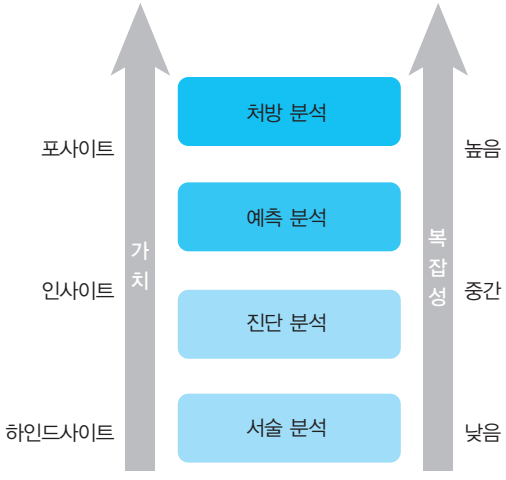
서로 다른 분석 유형은 서로 다른 기법과 분석 알고리즘을 활용한다. 이는 여러 유형의 분석 결과를 쉽게 전달하기 위해서는 서로 다른 데이터, 저장 및 처리 요구조건이 있을 수 있음을 의미한다. 그림 1.4는 가치가 높은 분석 결과를 생성하면 분석 환경의 복잡성과 비용이 증가한다는 사실을 보여주고 있다.

서술 분석

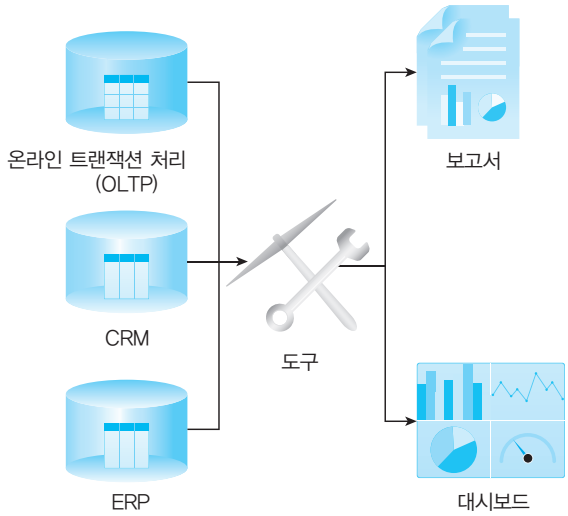
서술 분석은 이미 발생한 사건에 대한 질문에 답하기 위해 수행된다. 이러한 형태의 분석은 정보를 생성하기 위해 데이터를 상황에 맞게 조정한다.

샘플 질문에는 다음과 같은 질문이 포함될 수 있다.

- 지난 12개월 동안의 판매량은 얼마인가?
- 심각성 및 지리적 위치별 문의 전화는 몇 건인가?
- 각 판매원이 받은 월간 수수료는 얼마인가?



▲ **그림 1.4** 서술 분석에서 처방 분석으로 갈수록 가치와 복잡성이 증가한다.



▲ **그림 1.5** 왼쪽에 묘사된 운영 시스템은, 오른쪽에 묘사된 보고서 또는 대시보드를 생성하기 위해 설명 분석 도구를 통해 쿼리된다.

생성된 분석 결과의 80%는 본질적으로 설명 가능한 것으로 추산된다. 서술 분석은 가장 가치가 떨어지며 상대적으로 기본적인 기술을 필요로 한다.

서술 분석은 그림 1.5에서 볼 수 있듯이 주로 애드혹(ad-hoc) 보고나 대시보드를 통해 수행된다. 보고서는 일반적으로 정적이며 데이터 그리드 또는 차트 형식으로 표시되는 과거 데이터를 보여준다. 쿼리(query)는 고객 관계 관리(Customer Relationship Management, CRM) 또는 전사적 자원 관리(Enterprise Resource Planning, ERP) 시스템과 같은 기업 내에서 작동하는 데이터 저장소에서 실행된다.

진단 분석

진단 분석은 사건의 원인에 초점을 둔 질문을 이용하여 과거에 발생한 현상의 원인을 파악하는 것을 목표로 한다. 이러한 형태의 분석의 목표는 어떤 일이 왜 발생했는지 판단하려고 하는 질문에 대답할 수 있도록 현상과 관련된 정보가 무엇인지 결정하는 것이다.

샘플 질문에는 다음과 같은 질문이 포함될 수 있다.

- 2/4분기 판매가 1/4분기 판매보다 적은 이유는 무엇인가?
- 왜 동부 지역이 서부 지역보다 더 많은 문의 전화량이 있었는가?

- 왜 지난 3개월 동안 환자 재입원율이 증가했는가?

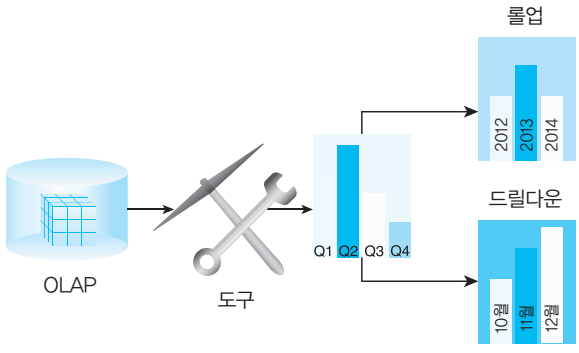
진단 분석은 서술 분석보다 더 많은 가치를 제공하지만 더 고급 기술을 필요로 한다. 진단 분석은 대개 여러 소스의 데이터를 수집하고 이를 그림 1.6과 같이 드릴다운(drill-down) 및 롤업(roll-up) 분석을 수행할 수 있는 구조에 저장해야 한다. 진단 분석 결과는 사용자가 추세 및 패턴을 식별할 수 있게 해주는 대화형 시각화 도구를 통해 볼 수 있다. 실행된 쿼리는 서술 분석의 쿼리보다 더 복잡하며 분석 처리 시스템에 보관된 다차원 데이터에 대해 수행된다.

예측 분석

예측 분석을 수행하면 미래에 발생할 수 있는 사건의 결과를 알게 된다. 예측 분석은 사건의 결과를 예측하려고 하며, 예측은 과거 및 현재 데이터에서 발견된 패턴, 추세 및 예외를 기반으로 이루어진다. 이는 위험과 기회에 대한 파악으로 이어질 수 있다. 이러한 형태의 분석에는 내부 및 외부 데이터와 다양한 데이터 분석 기법으로 구성된 대규모 데이터 세트가 사용된다. 예측 분석에 사용되는 모델은 과거 사건이 발생한 상황에 대해서 암묵적으로 의존성을 가지고 있음을 이해하는 것이 중요하다. 이러한 기본 조건이 변경되면, 예측을 수행하는 모델을 업데이트해야 한다.

질문은 대개 다음과 같이 what-if 방식을 이용하여 생성된다.

- 특정 고객이 대출에 대해 채무 불이행할 가능성은 어느 정도인가?
- 이러한 고객의 특징은 무엇인가? 예를 들어, 월세를 제대로 내지 못하는가?



▲ **그림 1.6** 진단 분석을 통해 드릴다운 및 롤업 분석을 수행하는 데 적합한 데이터를 생성할 수 있다.