제1장 빅데이터의 정의

1. 빅데이터 시대

1.1. 데이터가 넘쳐 나는 세상

급격히 증가하고 있는 현재의 데이터는 그 규모나 형태 때문에 데이터베이스를 포함한 기존 정보기술로는 보관, 처리, 분석하기 어려운 문제가 발생하고 있다.

우리는 지금까지 정보는 데이터로부터 파생되고, 지식은 정보로부터 얻어진다고 보면서 데이터·정보·지식을 구분했지만 빅데이터에서는 데이터와 정보를 구분하기가 어려워진다. 왜냐하면 데이터라는 단어의 원래 의미가 우리에게 '주어진 것(things given)' 이라는 뜻이지만 빅데이터 시대에는 주어진 것 외에도 특정 의도로 만들어 내는 데이터도 상당히 많기 때문이다. 의도를 가지고 있는 데이터는 정보라고 할 수 있기 때문에 빅데이터 시대에는 데이터와 정보의 구분이 모호해진다.

1.2 기업 데이터 관리의 어려움도 증가

기업들이 통제할 수 없는 속도로 데이터가 증가하면서 더 이상 자체적으로 모든 데이터를 보관하고 관리하는 일이 어려워졌다.

스마트기기의 보급으로 단순한 문서로만 존재하던 데이터보다 동영상, 음악, 사진 등의 비정형 데이터의 비중이 늘어났으며 데이터의 특성 또한 다양하고 복잡해졌기 때문에 기존의 시스템이나 방식으로 기업들이 대응하는 것이 어려워졌다.

데이터가 증가하고, 공유되고, 데이터 소스도 다양해지면서 보안이나 개인정보 보호문제도 함께 심화된다.

2. 빅데이터의 출현 배경

데이터가 넘쳐 나는 데는 몇 가지 이유가 있을 것이다. 가장 분명한 것은 디지털화의 진전, 즉 디지털 기술의 발전과 기기의 보급이다.

2.1 스마트폰과 인스턴트 메신저 등 SNS의 보급 확대

불과 2~3년 사이에 SNS와 스마트폰 이용으로 쏟아지는 대화나 글, 사진 및 동영상 데이터가 낯설지 않은 현상이 바로 빅데이터의 한 축을 형성하고 있다.

2.2 사물인터넷 및 센서의 보급 확대

도로·자동차·건물 등에 센서 부착이 증가하는 것은 물론 이들 간에 사람의 개입 없이 스스로 데이터를 주고받게 되는 사물인터넷(Internet of Things) 세상이 되면서 데이터는 더욱 크게 증가하고 있다.

2.3 데이터 처리를 위한 하드웨어의 발전과 비용 하락

데이터와 관련된 처리, 저장, 네트워크 기술의 발전도 빅데이터의 활용을 가능하게 하고 있다.

- 네트워크 발달 속도는 매년 25% 씩 증가
- 반도체 메모리의 용량이 1년마다 2배씩 증가
- CPU의 연산능력은 매년 59% 씩 성장

2.4 새로운 데이터 유형을 다루는 신기술의 등장

비정형 데이터를 저장 분류하는 하둡(Hadoop)의 출현에서부터 비정형 데이터를 분석할 수 있는 텍스트 애널리틱스 등 여러 가지 분석기법(analytics)의 등장, 그리고 실시간 데이터 처리능력의 증가로 기업들이 실제로 빅데이터를 다룰 수 있는 기회가 늘어나고 있다.

2.5 빅데이터의 부각과 기업활동

세상에 데이터가 많아지고 다양해지면서 기업들 또한 고객들이 물건을 구매할 때 여러

가지 외부 요인들에 영향을 받는다는 사실을 이해하기 시작했다.

정확하고 적시성 있는 데이터를 파악하면 자사 제품의 수익을 늘릴 수 있다는 사실로 인해 빅데이터를 활용하려는 기업들이 더욱 늘고 있다.

3. 빅데이터의 정의

3.1 규모

분석하는 데이터의 크기가 일정 수준 이상이어야 의미 있는 데이터로 취급한다.

엄밀한 정의는 없지만, 적게는 수 테라바이트에서 많게는 수 페타바이트 정도 크기의 데이터 집합을 지칭하는 것이 일반적이다.

3.2 다양성

빅데이터의 중심에 서 있는 데이터는 바로 비구조적 또는 비정형 데이터(unstructured data)이다. 비정형 데이터는 기존의 데이터베이스에 담을 수 없는 형태의 데이터이다.

3.3 속도

정보의 생성-유통-소비의 전 주기는 그야말로 눈 깜짝할 사이에 이루어지고 있다.

3.4 정의

협의의 빅데이터: 보통 수십에서 수천 테라바이트 정도의 거대한 크기를 갖고, 여러가지 다양한 비정형 데이터를 포함하고 있으며, 생성-유통-소비(이용)가 몇 초에서 몇 시간 단위로 일어나 기존의 방식으로는 관리와 분석이 매우 어려운 데이터 집합을 의미한다.

광의의 빅데이터: 기존의 방식으로는 관리와 분석이 매우 어려운 데이터 집합, 그리고 이를 관리·분석하기 위해 필요한 인력과 조직 및 관련 기술까지 포함하는 용어

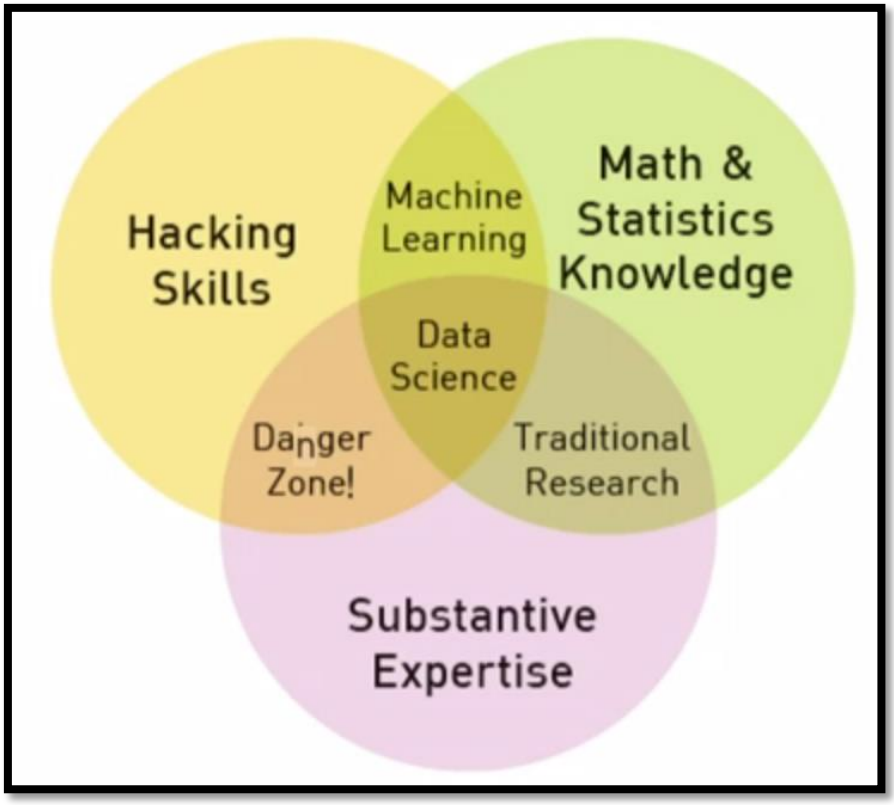
4. 데이터 과학과 데이터 과학자

4.1 데이터 과학

데이터 과학: 빅데이터를 효과적으로 수집·분석하여 가치 있는 통찰을 얻고 이를 최종 사용자에게 간단한 비기술적 언어로 표현하는 것과 관련된 학문

데이터 과학은 통계, 수학, 프로그래밍, 컴퓨터 과학 등의 기술과 이론이 결합된 융합 학문이다.

[데이터 과학을 구성하는 분야]



4.2 데이터 과학자

데이터 과학자는 일반적인 데이터 분석에 머물지 않고 다양한 원천의 데이터를 결합·분석하여 새로운 가치를 창출하는 일을 한다.

데이터 과학자는 해당 분야의 비즈니스에 대해 우선 이해하고, 컴퓨터 프로그래밍, 통계학은 물론 결과를 설득할 수 있는 대화능력을 가져야 한다.

[데이터 과학자에게 요구되는 지식·기술]