cloudera | ZDNet Korea

데이터 관리 신뢰 구축을 위한 데이터 사이언스 윤리강령

디지털 트랜스포메이션 시대를 맞아
데이터 중심의 의사 결정 시대가 오고 있다.

은행, 교육, 병원, 제조, 유통, 운송
그리고 정부까지 모든 산업 전반에서
데이터를 주요 자산으로 활용하고 있다.

하지만 빈번한 개인정보 유출에 기업의
데이터 관리에 대한 신뢰도가 무너진 상황이다.

무너진 데이터 관리 신뢰 구축을 위한
데이터 사이언스 윤리강령에 대해 알아본다.

cloudera | ZDNet Korea

데이터 관리 신뢰 구축을 위한 데이터 사이언스 윤리강령

Index

1. 데이터 사이언스 윤리강령이란?
   - 디지털 트랜스포메이션 시대의 신뢰도
   - 윤리강령과 하둡의 미래
2. 데이터 시대의 도래
   - 데이터 시대의 역사
   - 데이터 시대의 현주소
   - 데이터 사용 가능분야
3. 데이터 시대의 현주소
   - 데이터 시대에 발맞춘 윤리강령
   - 우리가 나아가야 할 길
4. 더그 커팅에게 묻다
   - 페이스북 데이터스캔들 사건
   - 무너진 신뢰 복구 방법
   - 클라우데라의 역할
   - 데이터 윤리의 합의도출과 표준화
   - 윤리강령의 보편성
   - 편견을 배제한 알고리즘

1. 데이터 사이언스 윤리강령이란?

디지털 트랜스포메이션 시대의 신뢰도

우리 사회는 디지털 전환(Digital Transformation) 시대를 향해 나아가고 있다. 디지털 전환의 초기 단계인 지금도 데이터를 활용한 발 빠른 혁신이 일어나고 있다. 기계 학습(Machine learning) 기술로 제품 사용 고객, 환경 등의 변화를 자동으로 학습하고 서비스를 실시간으로 개선할 수 있다. 한 번 개발된 애플리케이션(Application)을 수 십 년 동안 바꾸지 않고 그대로 사용하는 시대는 이제 지났다고 해도 과언이 아니다. **조직 내에서도 마찬가지로 마케팅, 세일즈, 엔지니어링, IT 부서 등 모든 직무에서 데이터를 근거로 업무가 이뤄지고 있다. 사회 발전에 필요한 다양한 가치 창출을 위해서는 데이터 활용이 필수다.**

데이터 중심의 신속하고 정확한 의사 결정을 하려면 다양한 데이터를 수집할 수 있어야 한다. 문제는 데이터 수집에 대한 여론이 부정적이라는 점이다. 미래를 다루는 공상 과학 영화에 등장하는 데이터 수집가가 주로 악당으로 묘사되고 있는 것은 단적인 예다. 데이터를 활용하지 못하면 기업 운영 전반을 개선하기 어렵고 문제 해결의 길도 막힐 수 있다. 기업 가치 창출의 제약은 사회 발전이 지연되는 악순환을 낳는다.

데이터 윤리강령 마련은 신뢰 구축의 지름길이다. 소비자는 데이터 수집 기관이나 담당자를 신뢰할 수 있을 때 정보를 제공하기 마련이다. 물론 현재 암호화, 비식별 등 데이터 수집에 대한 신뢰를 돕는 기술들이 쓰이고 있다. 하지만 이것만으로 충분하지 않다. 굴지의 글로벌 기업의 정보 유출 사건이 심심치 않게 일어나고 있기 때문이다.

윤리강령과 하둡의 미래

훌륭한 데이터 윤리강령이 마련되면 우리 사회는 한 단계 더 발전할 수 있을 것이다. 교육, 의료, 경제 부문에서 상상만으로 그쳐야 했던 혁신적인 서비스가 제공될 것이다. 공정한 사회 구축에도 기여할 것이다. 비윤리적인 방법으로 타인의 데이터를 악용하는 일이 사라질 것이기 때문이다. 지금껏 인류는 구체적인 사회 제도를 구축하고 더 나은 방향으로 개선해 왔다. 데이터 분야도 윤리 강령을 만들어 눈앞에 다가온 디지털 전환 시대의 변화에 적극적으로 대응할 수 있어야 한다.

머지않은 시점에 빅데이터 관련 협약이나 표준이 나올 것이다. 앞으로 개발될 기술들은 모두 이 한 가지 룰에 따라 데이터를 수집하고 활용하게 될 것이다. **준비 단계인 지금 빅데이터, 하둡(빅데이터 분석 오픈소스 기술)과 관련한 최상의 정책이 무엇인지를 고민해봐야 한다.** 기술은 사회에 긍정적인 영향을 끼쳤을 때 그 의미가 있다. 데이터 전문가와 관련 기업들의 역할은 향후 수년 내 마련될 빅데이터 정책을 따른 결과가 사회 발전을 견인하도록 하는 것이다. 그래야만 10~20년 후 빅데이터, 하둡에 기술에 대한 자부심을 느낄 수 있다.

“ 데이터 윤리강령 마련은 신뢰 구축의 지름길이다 ”

2. 데이터 시대의 도래

데이터 시대의 역사

데이터는 우리 사회 발전에 매우 중요한 역할을 한다. 기업이 1970~1980년대 수 십 년 동안 사용한 기술은 관계형 데이터 베이스 (Relational Data Base)였다. 관계형 데이터 베이스는 데이터를 단순한 표로 나타내기 때문에 특정 규모의 특정 문제를 해결하는 데 그쳤다. 금융, 재무, 재고관리 등에서 데이터 관리의 요구가 늘어났지만 쓰임에 제약이 많았다. 무어의 법칙(Moor's Law)은 데이터 관리 기술의 발전을 촉진시켰다. 저렴한 하드웨어가 다량 생산되고 디바이스가 늘어남에 따라 점점 더 많은 데이터가 생성됐다. 기존에는 볼 수 없었던 새로운 종류의 데이터가 발생했다. 하지만 **기존의 기술로는 이 방대한 데이터를 다루는 것은 불가능했다.**

빅데이터 시대는 클라우데라가 하둡(HADOOP)을 선보이면서 시작됐다. 2006년 프로젝트가 닻을 올리고 2008년 상용화하면서 새로운 종류의 데이터를 다양한 방식으로 탐구할 수 있는 기본적인 시스템이 갖춰졌다. 또한 사상 처음으로 페타바이트(PB)급 데이터를 저렴한 비용으로 저장하고 처리할 수 있게 됐다. 그러나 하둡만으로는 충분하지 않았다. 웹(Web) 활성화로 인터넷에서 발생하는 무한한 데이터를 관리하는 데 어려움을 겪게 된 것이다. 얼마 지나지 않아 하둡도 사일로(Silo)가 됐다.

오픈소스(Open Source)라는 새로운 소프트웨어가 하둡에 추가되면서 강력하고 유연한 플랫폼이 등장했다. 덕분에 하둡으로 대규모 배치 연산, 대규모 온라인 트랜젝션(Online Transaction), 분산 데이터 베이스(Hbase), 스트리밍 데이터 처리가 가능해졌고 머신 러닝(Machine Learning)도 적용됐다. 마침내 하둡 플랫폼으로 다양한 사업부와 사일로 데이터를 통합할 수 있게 돼 기업은 비즈니스나 사용자에 대한 단일 뷰를 가질 수 있게 됐다.

가장 최근 하둡 플랫폼에 올라간 기술 중 하나가 머신 러닝이다. 머신 러닝으로 언어와 이미지를 학습하고 처리할 수 있게 되면서 앞으로 벌어질 일을 예측할 수 있게 됐다. **기존 애플리케이션을 하둡 플랫폼에 옮기면 최첨단 고급 애널리틱스도 가능하다.** 또한 퍼블릭 클라우드를 사용하기 시작하면서 다양한 조직들이 새로운 시스템으로 문제를 빠른 속도로 해결해 나가고 있다.

"하둡 기술은 과거 전통적인 IT 기술로는 불가능했던 문제를 해결할 수 있도록 돕고 있다"

데이터 시대의 현주소

하둡의 등장이 데이터 시대에 혁명적인 변화를 일으켰다고 해도 과언이 아니다. 새로운 스타일의 데이터 시스템이 가능해진 덕분에 더 방대한 데이터를 사용하고, 다양한 소스에서 생성된 데이터를 취합해 전체적인 그림을 그릴 수 있게 됐다. 또한 과거 그 어느 때보다 더 유연하게 데이터를 탐구할 수 있다. 앱 하나를 개발해 수 십 년 동안 쓰는 게 아니라 제품, 고객, 주변 환경 등을 학습해 신속하게 혁신할 수 있게 됐다.

디지털 데이터는 이미 사회 전반에 스며들어 중요한 역할을 하고 있다. 비디오 게임 회사는 하둡 관련 기술을 사용해 플레이어가 어떤 무기를 살지를 미리 예측한다. 운송 업계는 가장 편리한 시점과 장소에서 트럭 정비를 할 수 있도록 한다. 하둡 기술은 과거 전통적인 IT 기술로는 불가능했던 문제를 해결할 수 있도록 돕고 있다.

디지털 전환(Digital Transformation)이 이뤄지는 미래에는 데이터 중심의 의사 결정이 주를 이룰 것이다. 인터넷 기업뿐만 아니라 은행, 통신사, 병원, 보험사, 제조, 유통, 광산, 장비, 교통•운송 업체, 정부 등 모든 산업 전반에서 데이터를 주요 자산으로 활용하고 있다. 조직 내부 부서에서도 데이터는 의사 결정의 핵심이다.

디지털 전환 시대의 초기 단계인 지금 미래의 변화에 능동적으로 대응해야 한다. 디지털 데이터, 하둡 관련 기술이 우리 사회에 미칠 영향을 미리 가늠해 봐야 한다. 데이터는 많은 이점도 있지만 역으로 잘못 사용하면 해를 끼칠 수도 있기 때문이다. 문제 해결을 위한 최적의 시스템과 운영 구조를 고안해내는 데 데이터를 사용할 때 더 좋은 사회를 만들 수 있다.

데이터 사용 가능분야

데이터가 특히 유용하게 사용될 수 있는 분야로 교육과 의료를 꼽을 수 있다. 우선 교육 분야에 데이터를 사용하면 맞춤 교육이 실현이 가능하다. 개별 아동의 행태와 니즈, 학습 속도에 맞게 교육을 할 수 있다. 이 같은 시스템은 아동 데이터 수집을 전제로 하기 때문에 리스크가 따를 수 있다. 의료 역시 데이터를 활용해 혁신이 일어날 것으로 예상되는 분야다. 고령화 시대의 요구에 따라 데이터를 활용한 건강 증진이 가능하다. 기존 의료 시스템 개선뿐만 아니라 유전체학을 통해 세포 구조를 파악하고 맞춤 질병 치유 방법을 개발할 수 있는 길을 열 수 있다. 맞춤 의료를 실현 하려면 의무기록과 개인정보를 수집해야 한다는 문제가 따른다. 교육과 의료 외에도 기후 변화, 빈곤 퇴치, 경제 분야에서 데이터를 활용해 문제를 해결하고 사회 전반을 개선할 수 있다.

3. 데이터 시대의 현주소

데이터 시대에 발맞춘 윤리강령

데이터는 교육, 의료, 빈곤퇴치, 기후 변화 등 우리 사회 발전에 필요한 문제들을 해결하는 데 활용된다. **데이터가 제 역할을 하려면 방대한 양의 데이터 수집이 전제 되어야 하는데 이는 신뢰가 쌓여야 가능하다.** 데이터 시대에 발맞춘 윤리강령이 필요한 이유다. 암호화, 익명처리 등의 기술이 존재하지만 이것만으로는 충분하지 않다. 문화적 기법을 통한 데이터 관련 사회 제도를 구축해야 한다. **빅데이터를 윤리적으로 활용하기 위해 필요한 신뢰 구축의 4요소는 △투명성 △베스트 프랙티스 △경계 설정 △검증 등이다.** 첫째, 투명성은 조직과 개인 간에 데이터가 어떻게 사용될 것인지에 대한 기대치를 조정해야 한다는 의미다. 데이터가 예기치 못한 방식으로 이용되는 것이 드러나면 데이터 수집에 대한 거부감이 생긴다. 따라서 사전에 데이터 사용 절차를 투명하게 공개해야 한다. 둘째, 데이터 관리 모범 사례를 수립해 베스트 프랙티스를 만들어야 한다. 한다. 데이터 관리, 암호화, 익명화 방법과 특정 데이터를 마스킹 해 나타나지 않도록 숨기는 방법 등의 모범 사례를 분명하게 마련해야 한다. 셋째, 허용되는 것과 금지되는 것 사이의 경계를 분명하게 해야 한다. 개인 정보, 사생활 정보는 물론 공개 가능 여부의 기준이 모호한 것 또한 포함이다. 경계를 명확히 해 불신의 싹을 사전에 잘라야 한다. 넷째, 검증이 필요하다. 어떤 기관이나 조직이 데이터 수집을 안전하게 한다고 밝혔다면 공언한 대로 지켜지는 지 스스로 검증할 수 있어야 한다. 정부가 감독 기능을 수행할 수 있겠지만 더 좋은 방법은 산업 별로 자율 규제를 하는 것이다. 미국은 자율규제 기관이 따로 있다.

우리가 나아가야 할 길

훌륭한 데이터 사이언스 윤리강령이 존재한다면 사회 여러 분야의 발전을 촉진할 수 있다. 교육, 의료, 경제 부문에서 많은 기업이 혁신적인 서비스를 제공할 수 있다. 데이터 윤리강령은 공정한 사회 실현에도 기여할 것이다. 비윤리적인 방법으로 데이터를 남용하는 일이 없어질 것이기 때문이다.

기술은 전 세계적으로 공유되고 있다. 데이터 신뢰도 구축은 국지적 문제가 아니다. 전 세계가 함께 고민하고 서로 배우려는 노력이 필요하다. 미국 연방무역위원회는 소비자를 공정하게 대우해야 한다는 측면의 가이드라인을 발표했다. 유럽은 2018년 5월부터 GDPR이라는 소비자 개인데이터 보호에 관한 법률이 시행된다. 민간에서도 이러한 정부 차원의 노력을 넘는 발 빠른 대응이 필요하다.

데이터 신뢰도 문제의 중요성을 전 세계가 인식할 수 있도록 하는 것이 데이터 산업 전체의 책임이자 의무다. 데이터 과학자나 기업 조직의 내부의 노력도 필요하다. 데이터를 책임감 있게 활용하기 위한 첫 걸음은 자기 검열이다. 특정 활동을 해도 괜찮을 지에 대한 질문을 스스로 던져보라는 것이다. 자기 검열을 통과했다면 같은 질문을 다른 사람에게도 똑같이 던져봐야 한다. 사람마다 기대치가 다르고 생각이 다르므로 타인의 기준에서도 같은 답이 나왔을 때 행동하는 것이 바람직하다.

적확한 윤리적인 접근이 무엇인지 파악하는 것은 쉽지 않다. 윤리적인 것은 흑백으로 명확하게 구분되는 것이 아니기 때문이다. 보통 사람의 기대치에서 벗어나지 않고, 시스템 사용자의 관점에서 스스로 묻고 윤리적이라고 판단되는 기술을 개발하는 것이 바람직한 자세다. 신뢰를 구축하기 까지는 많은 시간이 소요되고 지난한 과정을 거쳐야 한다. 향후 10년 후 빅데이터 관련 윤리 협약이나 표준이 마련될 것으로 예상되는데 한 발 앞서 데이터의 올바른 활용을 위한 최상의 정책을 고안해 둔다면 자부심을 느낄 수 있게 될 것이다.

“ 교육, 의료, 경제 부문에서 많은 기업이 혁신적인 서비스를 제공할 수 있다 ”

4. 더그 커팅에게 묻다

클라우데라의 수석 아키텍트 더그 커팅에게 직접 데이터 사이언스 윤리강령에 대해 물었다.

페이스북 데이터스캔들 사건

페이스북 데이터 스캔들 사건에 대한 직접적인 답변은 어렵다. 개발자 윤리에 대해 오랫동안 고수해 온 입장으로 대신하겠다. 데이터에 대해 아주 신중한 거버넌스를 확립해야 한다는 것이 우리의 생각이다. 이에 따라 클라우데라는 데이터 윤리와 관련한 다양한 기술을 제공하고 있다. **암호화 상태를 유지하는 것이 가장 중요하고, 보안이 강화된 방식으로 데이터를 관리해야 한다.** 특히 유저의 데이터를 보호할 수 있는 정책을 수립하고 지키는 것이 중요하다고 생각한다.

무너진 신뢰 복구 방법

무너진 신뢰를 재구축 하는 것은 매우 어려운 일이다. **신뢰를 구축하기 위해서는 제3자가 데이터와 개인정보 사용에 대한 직접 검증을 하는 것이 중요하다.** 클라우데라는 오래 전부터 데이터 액션 오디팅이 중요하다는 점을 역설해 왔다. 데이터 정책을 외부의 제3자가 감시할 수 있도록 한다면 개인정보보호를 철저히 한다는 것을 증명할 수 있을 것이다. 현재 제3자에 의한 데이터 액션 오디팅이 이뤄지지 않고 있는 것이 현실이다. 더 많은 기업이 외부 감사 정책을 채택해야 한다. 일종의 인증 같은 개념으로 바라보면 된다. 금융권이 투자자의 신뢰를 얻기 위해 감사를 받는 것처럼 데이터 감사에 대한 요구가 앞으로 늘어날 것이다. 로널드 레이건 전 대통령은 '신뢰하되 검증하라' 고 말했다. 핵무기 개발 프로그램에 대한 언급인데 특정 국가의 핵무기 감축 주장을 어떻게 믿을 수 있냐는 질문에 레이건 대통령은 '방법은 사찰단을 파견하는 것' 이라고 답했다. 즉, 데이터 오디팅은 핵무기 사찰과 같은 역할이라고 보면 된다.

클라우데라의 역할

급증하는 AI에 대한 관심이 높아지고 있다. 클라우데라는 이에 부응해 관련 기술 지원에 많은 노력을 기울이고 있다. 향후 많은 기업이나 조직이 다양한 라이브러리를 선보일 것이고, 유즈 케이스에 적합한 라이브러리를 채택할 것이다. 글의 텐서플로나 스파크, R, 파이썬 등과 관련해 다양한 노력이 AI 영역에서 이뤄지고 있다. **시장에 다양한 솔루션이 나와 있지만 고객은 오디팅과 제어가 가능하고 사용자의 접근권을 효과적으로 제한할 수 있는 머신 러닝 소프트웨어를 선택할 것이라고 본다.** 선택은 고객의 몫이다.

“ 암호화 상태를 유지하는 것이 가장 중요하고, 보안이 강화된 방식으로 데이터를 관리해야 한다 ”

데이터 윤리의 합의도출과 표준화

합의 도출은 어려운 문제다. 모든 이해 당사자가 산업 발전을 위해서는 반드시 합의가 필요하다는 것에 대해 동의해야 가능하다. 산업발전이 하나의 인센티브 역할을 하는 것이다.

다양한 이해를 중재하고 주도할 수 있는 조직이나 단체가 조직된다면 합의 과정을 촉진할 수 있을 수도 있겠다. 이런 조직이 데이터 사이언스 윤리 강령 기준을 설정하고, 준수 여부를 감독한다면 좋을 것 같다. 금융, 법률, 의료 분야에서는 이러한 조직이 존재하고 제도도 마련돼 있지만 데이터 사이언스 분야는 전무하다. 데이터 분야에도 반드시 필요하다. **사람을 공정하게 처우하고 데이터 윤리를 제대로 지키는 지 체크하는 시스템이 있어야 한다고 본다.**

올해 5월 중 발효되는 유럽의 GDPR이 데이터 윤리 관련 규제의 표준안이 될 수 있을지 판단하는 것은 시기상조다. 어떤 효과가 나타날지 예측하기 어렵다. 너무 엄격해서 데이터를 유용하게 사용하지 못하는 역효과가 나타날 수도 있다. 다른 한편으로는 너무 루즈해서 개인이 자기 데이터 권리를 제대로 확보하지 못할 가능성도 있다. 하지만 유럽에서 이러한 법을 시작하는 건 전 세계가 주목해야 할 좋은 실험이라고 생각한다.

신뢰를 구축하는 데는 시간이 오래 걸린다. 브랜드 인지도를 쌓고 가치를 높이는 것은 오랜 시간이 걸리지만 브랜드 가치를 상실하는 것은 한 순간이다. 앞서 밝힌 △투명성 △베스트 프랙티스 △경계설정 △검증은 기업과 브랜드가 신뢰를 확보하기 위한 하나의 원칙이라고 말할 수 있다.

윤리강령의 보편성

나라마다 윤리에 차이가 있다. 그렇다 하더라도 보편적인 가치는 있다. 유엔에서 선언한 보편적 인권을 바탕으로 데이터 윤리강령을 시작해도 좋을 것 같다. 보편적인 인권을 보장하는데 데이터 시스템을 사용하도록 하자는 것이다. **성별, 종교, 빈부격차에 관계없이 동일하게 사람을 대해야 하듯이 보편적인 인권 원칙부터 시작해 과학 윤리나 원칙, 강령을 만들어 갈 수 있을 것이라고 본다.** 물론 나라마다 차이는 있겠지만 살인은 어느 곳에서나 죄악이다. 이러한 큰 문제, 보편적인 것부터 시작해 볼 수 있다. 유럽의 GDPR을 보면 '잊혀질 권리'가 명시돼 있다. 언론의 자유를 중시하는 미국에선 이 권리를 절대로 인정하지 않을 것이다. 유럽은 개인 사생활을 중시하기 때문에 이 권리를 보장하려 한다. 보편적인 윤리 강령을 마련해도 문화에 따른 차이는 있을 것이다.

방법론적 접근과 편견을 배제한 알고리즘

데이터 윤리와 관련해 오픈소스 방법을 도입하는 것은 좋은 접근이라고 생각한다. 하지만 충분할 지에 판단은 유보해야 할 것 같다. 적은 숫자의 개발자나 이해당사자가 설득을 한다고 할 때 큰 기업이나 조직이 응하지 않을 수 있기 때문이다. 인공지능이나 머신 러닝 기술 발전으로 인한 윤리적 문제도 나타날 수 있다. 중요한 건 해석 가능성의 이슈다. 단순히 머신 러닝이 내린 결정이 좋은 결정이었는지를 평가할 것이 아니라 어떤 과정을 거쳤는지를 투명하게 알 수 있도록 해야 한다. 알고리즘 자체에 바이어스가 내재돼 있는 것은 아닌 지를 체크할 수 있어야 한다는 말이다. 특정 인종, 성별을 기준으로 내린 결정은 불법이고 비윤리적이라고 할 수 있다. 알고리즘 자체에서 편견을 배제하는 것이 중요하다.