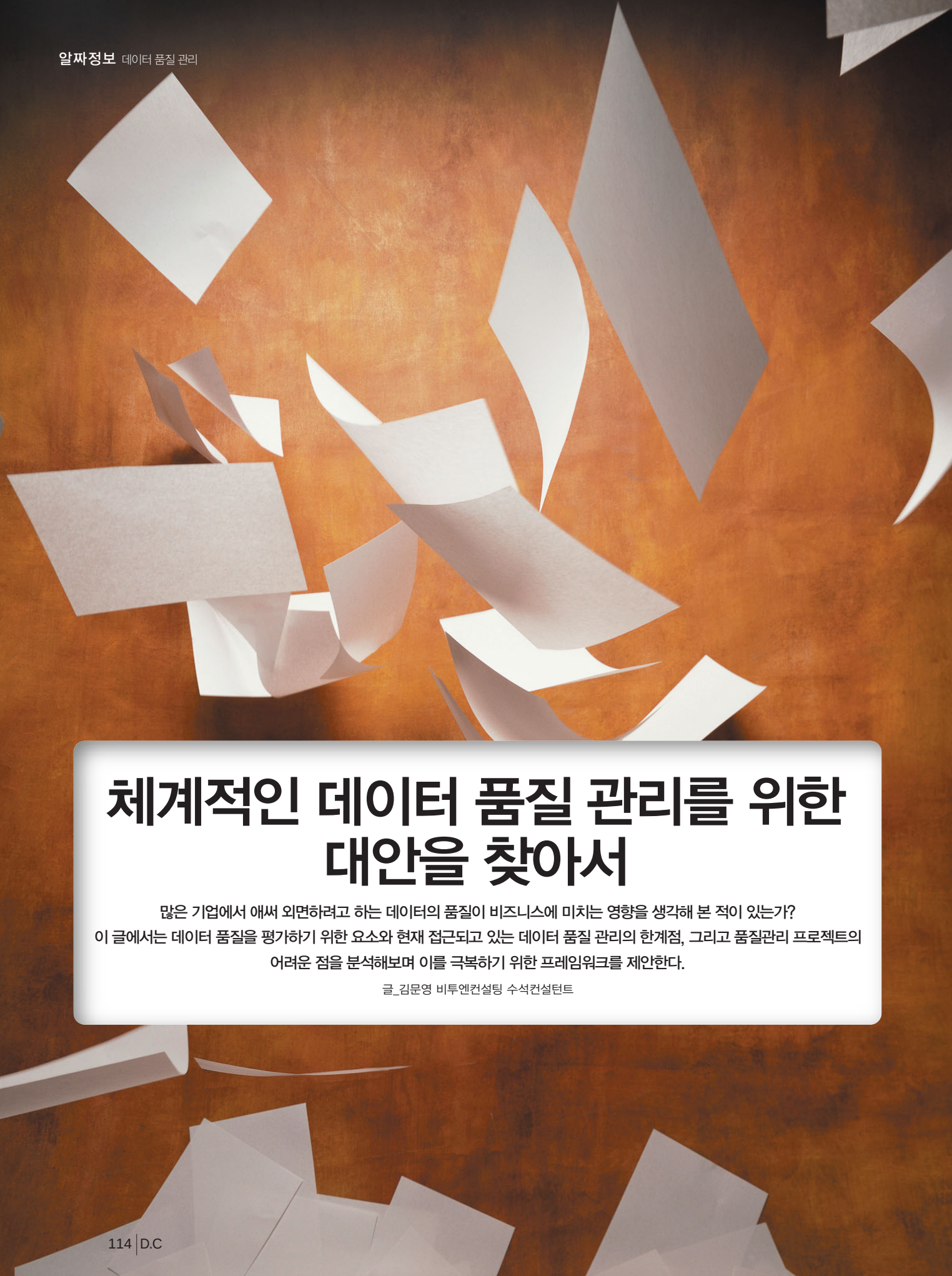


체계적인 데이터 품질 관리를 위한 대안을 찾아서

**많은 기업에서 애써 외면하려고 하는 데이터의 품질이 비즈니스에 미치는 영향을 생각해 본 적이 있는가?
이 글에서는 데이터 품질을 평가하기 위한 요소와 현재 접근되고 있는 데이터 품질 관리의 한계점, 그리고 품질관리 프로젝트의 어려운 점을 분석해보며 이를 극복하기 위한 프레임워크를 제안한다.**

글_김문영 비투엔컨설팅 수석컨설턴트

지금 처럼 경쟁이 치열한 시장 환경에서는 축적된 데이터를 유용한 정보로 변환해 기업의 의사결정에 도움을 줄 수 있는 지식으로 활용하는 것이 기업 생존 경쟁력의 관건이다. 따라서 많은 기업들이 데이터가 가지고 있는 비즈니스적 통찰력을 이해하기 위해 데이터 웨어하우징, CRM 등에 투자해 확보한 대규모 데이터를 필요에 따라 분석 · 활용해 기업 전체가 필요로 하는 지식 정보를 생성하려고 한다.

최근 국내 은행들이 앞을 다투어 도입한 리스크 관리 시스템(Risk Management System) 역시 바젤 Ⅱ를 비롯해 체계화된 관리를 위해 선진 회사들의 노하우를 고스란히 담아 놓은 솔루션을 도입한 것이다. 그러나 문제는 전혀 엉뚱한 곳에서 나타났다. 바로 리스크 관리 시스템을 가동해 위험 요인을 제대로 분석하기 위해서는 이를 가능케 해 줄 고품질의 축적된 데이터가 필요했지만 어떤 은행도 이를 충분히 가지고 있지 못했기 때문이다. 결국 가구 하나 없는데 집만 덩그러니 지은 셈이 되었다. 이제는 그동안 많은 투자를 해서 지은 멋진 집, 즉 정보 시스템에 안락하고 편리한 생활을 할 수 있는 가구인 고품질의 데이터를 손볼 시점이다. 이를 위해서는 현재 기업이 보유한 데이터의 상태, 내용 · 구조 · 가치 사슬에 입각한 흐름의 상태를 분석하고 단순히 세척하는 수준의 정제 작업이 아닌 정보 시스템의 웰빙을 위한 전략과 체계를 수립해야만 한다. 이 글에서는 정보 시스템의 웰빙을 위한 데이터 품질관리 전략 수립을 위해 고민하는 조직에게 단초를 제공하고자 한다. 이를 위해 먼저 저품질의 데이터가 가져올 수 있는 비즈니스적인 문제점과 그동안 기업들이 알게 모르게 수행해온 데이터 품질 관리의 형태를 살펴보고자 한다. 그리고 데이터 품질 관리의 개념과 품질 평가요소, 데이터 정제, 메타 데이터에 대한 설명과 한계점을 분석하고 이를 극복하기 위한 대안으로 통합된 데이터 품질관리 프레임워크를 말하고자 한다.

부정확한 데이터로 어떤 손실을 입는가

만약 부정확한 데이터를 기반으로 한 커뮤니케이션이 이루어지고 다시 이 커뮤니케이션을 기반으로 중요한 비즈니스 의사결정이 이루어진다면 어떻게 될까? 혹은 정확한 데이터를 기반으로 한 의사결정과 부정확한 데이터를 기반으로 한 의사결정은 비즈니스에 어떠한 영향을 미칠까? 사실 그 대답은 우리 모두가 이미 잘 알고 있다. 낮은 품질의 데이터가 기업에 미치는 영향은 실로 막대하다. 예를 들어 잘못된 고객 정보 관리로 인한 손실은 엄청난 영향을 가져올 수 있는데 기업이 고객을 한번 잃는 것은 잠재적인 미래 수익의 기반을 잃어버리는 것이기 때문이다. 그 실례들을 보면 다음과 같다.

1. 데이터 품질 저하로 인해 발생된 데이터 손실 비용, 재작업 비용 등이 기관이나 기업의 연간 예산이나 수익의 10~25%에 이름(Data Quality 2001년 9월호)

2. 데이터 품질 저하로 인한 고객 불만, 제품 관련 소송, 재작업 등의 비용이 평균 매출액 25~30%에 이르는 것으로 분석됨(품질 전문가 Joseph M. Juran)

3. "낮은 품질의 데이터로 인한 비즈니스 비용, 예를 들면 회수 불가능한 비용, 제품과 서비스에 대한 추가 작업, 차선책, 수익창출의 기회상실 등은 조직 내 전체 예산이나 수익의 10~25% 이상 차지할 것이다."(데이터 품질 전문가 Larry P. English)

4. 잘못된 고객 정보 관리로 인한 비즈니스 손실 비용이 $611 Billion에 이름(2001년 TDWI 보고서)

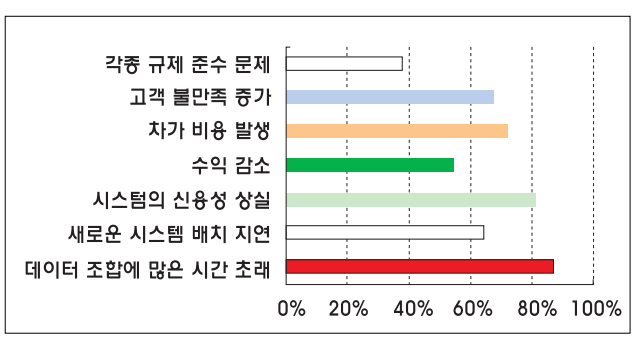
5. 금융계좌 주민번호 398만개 잘못 기재, 이사 소득에 대한 금융과세 못해(조선일보 2004년 1월 7일 기사 일부)

6. 국가 공공 기관 데이터베이스 관리 체계 허술—공공기관 데이터 영역의 품질 수준은 100점 만점에서 58.7점(한국데이터베이스 진행센터 2004 데이터베이스 품질진단 결과)

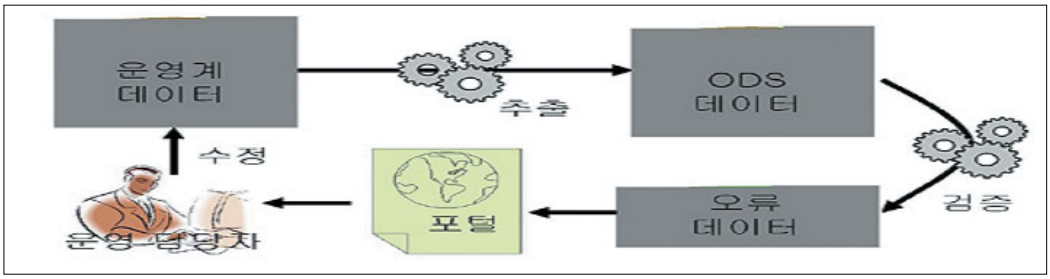
데이터 품질 관리란 무엇인가

앞서 데이터 품질의 중요성은 누구도 의심하지 않는다고 했지만 실제 현실은 그리 낙관적이지는 않다. 대부분의 기업은 데이터 품질을 위한 시스템을 갖추고 있지 않으며, 데이터 오류로 인해 문제가 발생했을 때만 대응하는 임기응변식 해결 방식을 택하고 있다.

사실 많은 기업에서 불량 데이터는 알려지지 않은 문제로 남아있다. 기본적으로 여러 시스템들이 불량 데이터 문제를 숨긴 채 오랫동안 그럭저럭 운영되어 왔으며, 이로 인해 각 부서는 정보가 일치하지 않는 보고서들을 만들어 내거나 정확한 대답을 찾기 위해 사람들이 직접 데이터를 손보는 작업을 해야만 했다. 즉 데이터 품질 문제 자체를 부정하며 비밀로 남겨두는 것으로 의사 결정자들의 눈에는 이 문제가 보이지 않았다. 비즈니스 의사 결정자들이 데이터 품질 문제를 직시하지 못하는 이유는 일반적으로 데이터가 그들에게 보고서 형태로 제출되기 전에 얼마간의 수작업을 거치는데, 이 과정에서 불량 데이터가 제거되면서 올바른(?) 데이터만 보기 때문이다. 이런 기업의 경우, 큰 사건이 터지지 않는 한 비즈니스에 불량 데이터가 미치는 영향을 공식적으로 평가 · 분석하기 위한 시간



〈그림 1〉 저품질 데이터로 인한 문제들



〈그림 2〉 데이터 정제

이나 노력을 들이지 않을 것이다.

다행히도 최근 들어 낮은 품질의 데이터의 문제점을 인식하기 시작한 기업들은 스스로에게 "우리 회사가 보유하고 있는 고객 및 비즈니스 정보와 데이터는 과연 어느 정도 가치가 있을까"라는 고민을 던지면서 자신이 보유한 다양한 데이터(고객 및 비즈니스 데이터)를 무엇인가 가치 있는 자산으로 변환시키는 방안을 모색하기 시작했다. 그렇다면 데이터 품질 관리란 무엇이며 사용자가 원하는 품질을 가진 데이터는 과연 무엇인가? 먼저 품질 관리의 대상인 데이터의 의미부터 살펴볼 필요가 있다.

데이터의 의미를 찾아서

기업에서의 데이터란 조직의 전략 및 목적을 달성하기 위해 전략, 구현, 운영 등의 정보 시스템 가치 사슬을 통해 생성된 산출물을 의미한다. 또한 데이터베이스 내에 저장되어 있는 데이터 값뿐만 아니라 기업이 요구하는 정보 체계를 형상화한 데이터 모델 및 데이터에 관한 데이터를 관리하는 메타 데이터 등의 구조적 요인까지 포함하는 개념이다. 따라서 데이터 품질이란 다음의 데이터 요건에서 볼 수 있는 바와 같이 조직의 전략과 목적을 달성하기 위해 필요한 데이터를 요구하는 조직 구성원 또는 이해 관계자의 만족도를 충족시킬 수 있는 수준으로 지속적으로 제공할 수 있는 것을 의미한다. 이러한 데이터 품질을 유지하고 개선하기 위해 수행하는 모든 활동을 '데이터 품질 관리' 라고 한다.

- ◆ 사용자가 요구하는 품질을 가진 데이터 요건
- ◆ 사용자는 업무와 관계되는 데이터를 원한다.
- ◆ 사용자는 정확한 데이터를 원한다.
- ◆ 사용자는 데이터의 불일치가 가능한 최소이기를 원한다.
- ◆ 사용자는 가능한 가장 최신의 데이터를 원한다.
- ◆ 사용자는 그들이 사용하는 도구에 가장 적합한 데이터를 원한다.
- ◆ 사용자는 필요한 데이터를 쉽게 제어하기를 원한다.
- ◆ 사용자는 그들의 데이터가 보안성과 비공개성이 지켜지기를 원한다.

과거에 수행된 데이터 품질 관리 활동

그렇다면 과연 기업은 데이터 관리를 전혀 하지 않았는가? 데이터에 기반해 정보시스템이 움직이고 있으니 데이터 관리를 안 해왔다고 말할 수도 없는 노릇이지만 그렇다고 체계화된 방법이나 특별한 담당자가 있는 것도 아니니 관리해왔다고 말하기도 어렵다. 즉 알게 모르게 이루어진 데이터 관리인 셈이다. 그러나 '알게 모르게' 이루어지던 데이터 관리는 작업자의 시각에 따라 하나의 비즈니스 사실에 대해 정확하지 않은 리포트들을 생산해 의사결정에 피해를 입히는 중요한 원인이 되었으며, 작업자의 오류로 인해 발생하는 사소한 문제도 많았다. 최근에 데이터 품질 관리가 새로운 이슈가 되고 있는 것은 알게 모르게 이루어지던 일을 체계화하는 것이라고도 볼 수 있다.

국내 기업들이 그 동안 데이터 품질 관리 활동으로 대표적으로 수행한 것은 대형 SI 프로젝트에서 필수적으로 행해지는 데이터 이관 작업에서 수행해야 하는 데이터에 관한 작업들일 것이다.

신규 시스템의 신규 데이터 모델로 데이터를 이관하기 위해서는 기존에 운영 중인 시스템의 소스 데이터가 깨끗해야만 신규 시스템도 제대로 운영될 수 있으므로, 현 시스템의 데이터를 정제하는 작업을 제일 먼저 실시하게 될 것이다. 데이터 정제 작업은 오래된 기업일수록 매우 힘들 수밖에 없는데 이는 오랜 세월 동안 시스템이 변경되면서 운영에 필수적인 데이터의 구조로만 유지되었고 정확한 비즈니스 룰을 지키기보다는 단순히 만들어진 프로그램에서 운영상 오류가 발생하지 않을 정도로만 데이터가 관리됐으니 앞뒤가 맞지 않는 데이터가 발생할 수밖에 없다. 특히 이력을 관리하는 데이터의 경우는 더욱 심각해 반드시 순서대로 일어나야 할 이벤트가 뒤죽박죽 일어나기도 한다. 이는 비즈니스 룰이 바뀐 것이 아니고 정보 시스템 관리상의 문제가 대부분이다. 이러한 운영상에 발생되는 데이터의 오류를 개선하기 위한 노력이 현재 고객들이 생각하는 데이터 품질 관리의 한 축이기도 하다. 이러한 운영 데이터의 품질 개선은 뒤에서 다시 상세히 다루어 보기로 하겠다.

데이터 적합성 검사가 품질 관리의 시발점

다시 원점으로 돌아와서 데이터 이관 작업에서 필수적인 데이터 정제 작업과 데이터를 신규 시스템으로 옮기려면 새로운 데이터 구조(신규 데이터 모델)에 맞추어 데이터를 가공해야 하므로, 이때 발생할 수 있는 데이터 가공

시의 오류를 찾아내고 바로 잡아야 하는 데이터 적합성 검사가 데이터 품질 관리의 시발점이라고 생각하는 것이다. 또 한 가지 많이 수행되는 데이터 품질 관리 활동은 정보계에서 운영계 데이터의 오류를 찾아내고 이를 근거로 운영계 데이터에 오류를 수정하는 사후 조치적인 데이터 품질 관리 활동이다.

예를 들어 K사의 경우 데이터 웨어하우스 시스템의 ODS(Operational Data Store)에 운영계 데이터가 모이면 사전에 정의한 중요한 비즈니스 룰에 따라 데이터가 정확하게 발생되었는지 확인하기 위한 프로그램들이 수행되고, 발견된 오류는 데이터 웨어하우스의 포탈 시스템 또는 별도의 응용 프로그램을 통해 운영계 담당자에게 전달되는 체계로 운영하고 있다. 데이터 품질을 검사하기 위한 중요한 비즈니스 룰이 변경되거나 늘어남에 따라 품질 측정 프로그램의 유지보수가 또 다른 이슈가 될 수 있다. 왜냐하면 엄청나게 많은 양의 데이터를 대상으로 복잡한 로직의 프로그램이 여러 번 수행되어야 하므로 막대한 전산 자원이 필요하게 되고 고급 유지보수 인력이 필요하므로 각 기업들로서는 여러 가지 난관에 봉착하게 된다.

이와 같이 데이터 품질 관리는 일반적으로 전사적 데이터 웨어하우스(Enterprise Data Warehouse) 구축 작업의 일환으로서, 특히 ETL(Extraction, Transformation, Loading) 작업의 데이터 검증 차원에서 수행되어 데이터 정제(Cleansing) 작업을 통한 데이터 내용의 품질 관리 활동이 주를 이루었다.

시장에서 논의되는 데이터 품질 관리

최근에는 ETL을 중심으로 한 데이터 정제 중심의 품질 관리 활동에서 데이터 품질 평가 요소를 통한 현황 분석, 표준화를 통한 데이터의 정제, 메타 데이터 관리 등 보다 체계적인 접근법이 제시되고 있다. 여기에서는 데이터 품질 관리 활동의 기본 요소인 데이터 품질 평가 요소와 메타 데이터에 대해 살펴보고자 한다.

먼저 데이터 품질의 평가 요소는 크게 데이터의 내용, 데이터의 구조, 데이터의 이동과 흐름의 관점에서 파악할 수 있다.

첫 번째 데이터 품질 관리 솔루션들이 주목하고 있는 데이터의 내용의 품질 평가 요소를 살펴보자. 데이터의 내용은 데이터베이스에 저장되어 있는 데이터 값을 의미한다. 조직이 기능을 수행하는데 반드시 필요한 데이터로 작업 처리상 일시적으로 필요한 임시 데이터는 제외하는데 데이터 내용의 주요 품질 기준은 완전성, 유효성, 정확성, 일관성을 꼽을 수 있다.

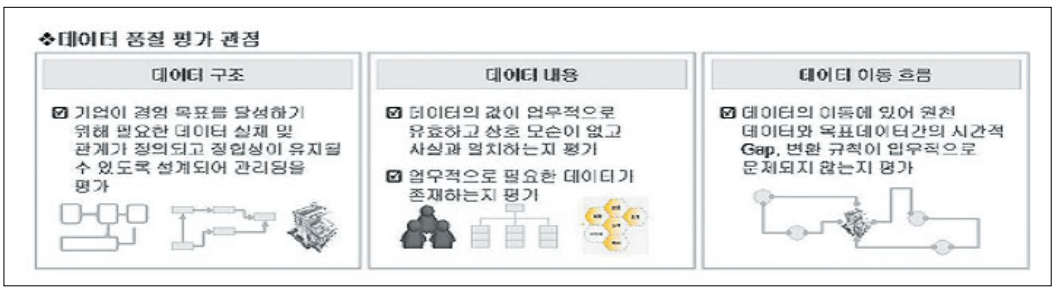
데이터 내용의 완전성(Completeness)이란 작게는 저장된 데이터 값이 NULL 값을 가지는 것인가를 의미하기도 하지만 진정한 의미의 완전성이란 비즈니스 요건을 만족시키기 위해 요구되는 데이터의 값을 보유하고 있는지를 의미한다. 데이터 내용의 유효성(Validity)은 존재하는 데이터의 값이 업무적으로 유효한 의미를 지니는 것을 의미하는데 도메인의 유효성, 속성 값의 유효성이 해당된다.

데이터 내용의 정확성(Accuracy)은 수록되어 있는 데이터가 오류 없이 데이터 원천(source)들이 가지고 있는 값과 동일한가를 의미하는 것으로, 외부 데이터가 대량 유입되는 경우나 조직의 인수 · 합병 또는 신규 시스템의 구축으로 인해 대량의 데이터가 이전되는 경우 문제가 발생할 소지가 있다. 따라서 데이터의 입력 후 원천 데이터를 이용해 비교 검증하는 과정이 필요하다.

데이터 내용의 일관성(Consistency)은 데이터베이스의 관련 있는 데이터 값들이 상호 모순이 없이 일관되어야 함을 의미한다. 예를 들어 고객 테이블의 고객명과 카드 계약 테이블의 고객명은 동일한 의미와 값을 가져야 함을 말하는 것으로 용어, 속성 정의의 표준 규칙이 불분명할 경우 많은 문제가 발생한다.

데이터의 구조 품질 요건

다음은 데이터의 구조 품질 요건에 대해 살펴보자. 데이터의 구조 품질평가는 기업의 정보 구조를 체계적으로 나타내기 위해 데이터를 사용자 및 전사적 관점에서 인식 · 분석하고 정의하였는가를 평가하는 것이다. 데이터 구조 품질은 단순히 참조 무결성이나 유일성, 관계 검증, 정규화 정도를 의미하는 것이 아니라 분석 작업을 통해 해석된 업무 관계 정의의 품질을 의미하는 것으로 모델링 전문가는 물론 비즈니스 담당자에 의한 데이터 구조 정의가 필수적이다. 데이터 구조의 품질은 품질 관리 솔루션을



〈그림 3〉 데이터 품질 관리

〈그림 4〉 일바적인 데이터 정제 프로세스

통해서는 매우 제한적인 효과만을 거둘 수 있으며 표준화된 데이터 모델 관리 정책, 전문 인력의 확보, 케이스 툴과 리포지토리를 이용한 모델 및 비즈니스 규칙의 중앙 집중화된 관리 등이 결합되어야 한다.

마지막으로 데이터의 이동과 흐름의 평가 요소를 살펴보자. 데이터의 이동과 흐름은 데이터를 최종 사용자에게 의미 있는 정보나 지식으로 변화하기 위한 과정으로, 이 단계의 평가 요소는 적시성(Timeliness)과 편리성의 관점에서 파악할 수 있다.

데이터 이동과 흐름의 적시성은 데이터의 이동에 있어 원천 데이터와 목표 데이터간의 시간적 차이(gap)가 업무적으로 문제되지 않아야 함을 의미하며, 최근의 실시간 기업 환경(RTE)에서 더욱 중요한 품질 평가 요소이다. 편리성은 데이터를 활용함에 있어 사용자 커뮤니케이션의 지원성이나 속도의 적절성, 이동과 흐름 과정의 안정성을 의미한다.

데이터 정제

데이터 품질 향상을 위한 대표적인 활동이 데이터 정제(Cleansing)이다. 데이터 정제 작업은 데이터 웨어하우스(Enterprise Data Warehouse) 구축 작업의 일환으로서, 특히 ETL 작업의 데이터 검증 차원에서 계속적으로 수행되어 왔다. 이 부분은 규칙 기반의 자동화된 방법으로 지원하는 많은 솔루션이 출시되었으며 최근에는 정형 데이터의 정제뿐만 아니라 이름, 주소, 전화번호 등 비정형성 텍스트 데이터의 정제도 지원한다. 〈그림 4〉는 데이터 품질 관리 툴이 지원하는 일반적인 데이터 정제 프로세스이다.

데이터 정제 작업은 매우 힘든 세척과 반복되는 작업이며 기업 데이터 품질의 방화벽 역할을 수행하는 중요한 단계이다. 그러나 이런 고된 세척과 반복 작업으로 인해 데이터 품질을 둘러싸고 가장 위험하고 잘못된 인식 중 하나가 크게 유포되어 왔다. 그것은 불량 데이터란 잘못 입력된 이름, 불완전한 주소, 데이터 필드의 누락 등과 같이 단순히 '부정확성' 에 관한 것뿐이란 인식이다. 이런 생각을 통해 유추해 나가면, 결국 인력을 투입해 데이터 정제 작업을 하면 그 문제는 사라진다는 것이다. 그러나 데이터 정제는 데이터 품질 문제에 있어 첫 번째 단계일 뿐이며 더 중요한 것은 비즈니스에서 요구하는 정보가 어떻게 표현되는가에 대한 표준을 확립하는 것이다. 이렇게 해야만 데이터가 청구서에 사용되든, 우편 판매 캠페인에 사용되든 일관성이 유지될 수 있다는 것이다.

메타데이터의 관리

마지막 데이터 품질 관리 활동은 메타 데이터 관리다. 메타 데이터란 데이터에 대한 구조적인 데이터로서 관리할 정보 자원을 기술하는 것으로 실제 정보를 가지고 있는 콘텐츠가 아니라 콘텐츠에 대한 정보를 가지고 있는 데이터를 의미한다.

메타 데이터 관리는 데이터 관리 효율성 증진을 위해 제기됐으며 전사적으로 운용되는 정보 자원에 대한 데이터 관리 및 운용하기 위한 요소가 제대로 적용되었는지가 품질 관리의 주요 대상이다. 수집된 메타데이터와 정의된 표준 데이터를 효율적으로 공유 · 관리하기 위해 메타 데이터 리포지토리(Metadata Repository) 구축이 필요하다.

메타 데이터 저장소인 메타 데이터 리포지토리는 기업의 다양한 메타 데이터를 자동으로 수집하기 위한 기능과 요청, 승인, 확인 등 데이터 관리 업무 프로세스를 효율적으로 수행하기 위한 워크플로우 기능이 제공되어야 한다.

솔루션 중심 데이터 품질 관리 활동의 한계점

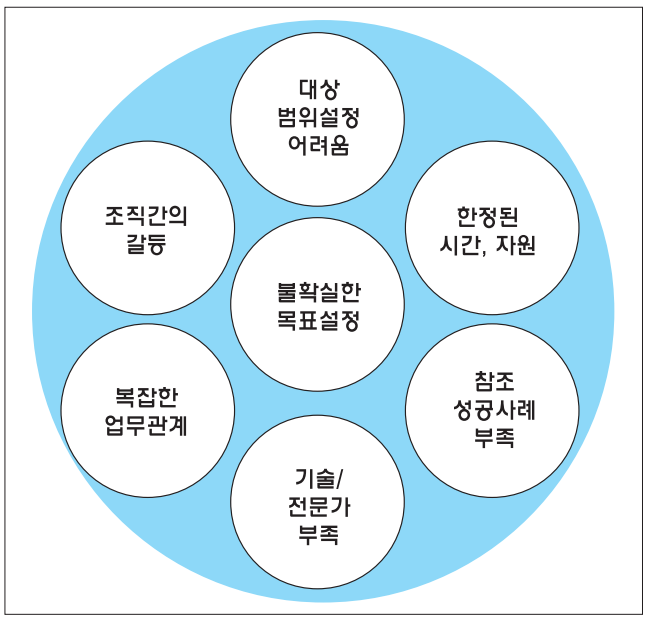
지적 자산과 노하우(Know-how) 정보가 물질적 인프라보다 훨씬 더 중요한 자산으로 인식하고, 정보 품질 관리 작업의 중요성을 인식한 기업이 증가하고 있음은 매우 다행스러운 일이다. 특히 최근에 시작되고 있는 데이터 품질 관리는 데이터에 대해 단순 정제나 재작업을 하면 된다는 사고방식에서 '데이터 품질 개선이란 지속적인 과정' 이라는 새로운 사고방식으로 서서히 대체되고 있으며, 이에 따라 직원들에게 데이터 품질과 관련해 다양한 단계에서 더 많은 책임과 권한을 부여하고 있다. 그러나 이 과정에서 솔루션 중심의 접근으로 인해 데이터 품질 관리가 솔루션만 도입하면 해결되는 단기적이고 기술적인 문제로 인식되는 경향 또한 나타나고 있다. 솔루션 중심의 접근은 기존 국내의 정보 시스템들이 선진 사례의 도입이라는 명분 아래 고가의 솔루션을 도입하고도 뚜렷한 효과를 보지 못한 것과 같은 결과를 초래할 수 있다. 앞에서 설명한 데이터 품질 관리 활동 요소별로 발생 가능한 문제

점에는 어떤 것이 있을까?

단순한 통계치 중심의 접근

먼저 데이터의 내용 · 구조 · 이동 · 흐름에 대한 품질 평가에서 솔루션 지향적인 평가의 경우 단순히 통계치 중심의 접근 방법을 채택함으로써 비즈니스 관점의 품질 정의 및 평가 부족을 초래할 수 있다. 대부분의 데이터 품질 관리 솔루션은 데이터의 품질 평가를 위해 데이터 프로파일링과 데이터 점검(Auditing) 기법을 이용한다. 데이터 프로파일링이란 데이터 자체로부터 지식을 획득하는 프로세스라고 정의할 수 있는데, 데이터 값의 누락, 도메인의 일관성, 관련 항목간의 상관성 등을 보기 쉽게 나타내준다. 데이터 프로파일링을 위해 통계적 접근 방법 및 인공 지능에 의한 추론 방법 등을 사용하며 불완전한 데이터 정의와 관계없이 데이터 자체로부터 의미 있는 품질 평가치를 도출한다는 점에서 의의를 찾을 수 있다. 데이터 점검은 비즈니스 규칙과 데이터 간의 정합성을 검증하는 단계로 주로 사용자에 의해 업무 규칙을 정의하고 이것을 툴이 구현해 품질을 검증하는 형태이다.

데이터 프로파일링의 경우 실제 데이터 내용과 구조 정의 간의 상관관계에 따른 품질 평가가 지원이 부족해 실용적으로 도움이 되지 않는 결과가 도출될 수 있다. 예를 들어 고객 테이블의 생년월일과 카드 계약 테이블의 계약일자 실제 데이터 값은 상당히 유사할 수 있지만 업무적으로 아무런 연관 관계를 가지지 않는다. 따라서 이와 같은 비교는 많은 시간과 자원을 들임에도 불구하고 실용적으로 도움이 되지 않는 평가 결과를 도출한 셈이 된다. 데이터 점검의 경우는 복잡한 업무 규칙의 점검 기능 구현에 제한이 있다는 문제점을 안고 있다.



〈그림 5〉 데이터 품질 관리 접근 방법의 한계

비즈니스 요건이 전사적 관점에서 체계적으로 구조화가 되었는가를 평가하는 데이터 구조 평가에서는 솔루션 중심의 접근 방법은 보다 많은 한계점을 드러낸다. 데이터 품질 관리 솔루션들은 단순히 관계형 데이터베이스의 정규화 기능을 제공하는(때론 아예 지원하지 않거나) 것 이상의 데이터 구조의 품질을 개선 방안을 도출해내기 어렵다. 사실 비즈니스 요건의 체계적 구조화란 객관식 문제와 같은 정답이 존재하는 것이 아니며 수많은 결정 요인과 물리적인 환경을 감안해 도출되는 것이므로 자동화된 품질 개선 효과를 바라는 것 자체가 무리일 수 있다.

데이터의 이동과 흐름 부분은 품질 관리가 이슈가 되기 이전부터 툴에 의한 접근 방법이 꾸준히 활용된 분야다. 현재 데이터 품질 관리 솔루션을 제공하는 업체들 역시 ETL 툴을 기반으로 데이터 품질 관리 또는 통합 솔루션으로 확대 · 발전시킨 경우가 많다. 데이터의 이동과 흐름 부분에서 솔루션의 문제점은 오랜 기간 제기되어 왔음에도 불구하고 지속적으로 제기되는 것으로 성능과 복잡한 이동 규칙의 구현 문제로 요약될 수 있다.

데이터를 정보 활용의 다음 단계로 이동하거나 변환하는 과정은 자동화된 툴을 사용할 경우 많은 장점을 얻을 수 있다. 우선 사람의 수작업으로 인한 작업 오류를 최소화할 수 있으며 스케쥴링 기능을 이용해 작업 간의 배치나 조정을 효율적으로 할 수 있고, 데이터 변환이나 정제 작업을 통합해 진행할 수도 있다. 이런 장점에도 불구하고 여전히 제한 요건이 되고 있는 것은 성능 문제다. 특히 대용량 데이터(수십 테라의 데이터의 이동, 변환, 정제 작업에 단지 34시간만이 제공되는 경우도 많다)의 이동에 있어서는 성능 문제는 매우 민감한 요건이라고 할 수 있다. 다만 병렬 처리 엔진이나 하드웨어의 급속한 발달로 인해 많은 개선 효과가 나타나고 있으므로 해결이 가능한 부분이라는 것이 위안으로 삼을 수 있겠다.

보다 본질적인 데이터 이동과 흐름 부분의 문제는 자동화된 툴로 복잡한 업무 이동과 변환 규칙을 구현하는데 한계를 가진다는 것이다. 따라서 많은 기업들이 복잡한 업무 규칙의 경우 별도의 SQL 로직을 구사하고 있으며 이의 유지 보수로 인력과 시간을 소비하고 있는 것이 현실이다.

〈그림 5〉는 기존의 데이터 품질 관리 접근 방법의 한계를 도식화한 것으로 요약하면 기술 중심, 솔루션 중심의 접근 방법이 가지는 한계를 의미한다고 볼 수 있다.

데이터 품질 관리 프로젝트의 어려움

데이터 품질 관리의 중요성을 인식하고 전략적으로 데이터 품질 관리 프로젝트를

추진하고자 하는 조직은 곧 어려움에 부딪히기 쉽다. 데이터 품질 관리 프로젝트의 어려움 중 몇 가지만 살펴보기로 하자.

복잡한 업무 관계

더러운 데이터에서 깨끗한 데이터를 구분해 주는 엄격하고도 신속한 규칙 같은 것은 존재하지 않는다. 단순히 사용하는 데이터 항목의 단위가 다르거나 도메인을 검증하는 것으로 데이터의 품질이 획기적으로 향상되지는 않는다는 의미이다. 데이터 품질 관리 소프트웨어들이 비정상적인 데이터 포맷을 알려주는 것과 같이 도움을 주기는 하지만 기술은 여기까지만 그 힘을 발휘한다. 즉 데이터가 가지고 있는 복잡한 업무 관계를 도출하고 품질을 향상시키는 것은 때로는 온전히 프로젝트 인력의 몫이 될 가능성이 높다.

조직 간의 갈등

더러운 데이터를 정제하는 일보다 더 좋은 것은 처음부터 깨끗한 데이터를 확보하는 것이다. 정확한 정보에 크게 의존하는 업무일수록 데이터가 생성될 때 실시간으로 그 정확성을 검증하는 프로세스를 구축하는 데 앞장설 필요가 있다. 그러나 데이터를 생성하거나 입력하는 조직이나 사람들의 입장과 활용하는 조직이나 사람들의 입장이 상이한 경우가 다반사이다.

예를 들어 통신 회사에서 신규 고객이 서비스를 가입하고자 하는 경우 이를 처리하는 직원은 "우리의 정보는 원래 고객이 통신 회선 가입을 하고 서비스를 이용할 수 있도록 하기 위해 만들어지고 있는 것이지, 보고서를 제출할 수 있도록 하기 위해 만들어지는 것은 아니다"라고 말한다.

그 결과 데이터에는 서비스를 이용하기 위한 최소한의 정보만이 입력되고 품질이 고려될 뿐 잘못된 고객 정보나 정보간의 불일치에는 별 관심이 없다.

이것은 서비스의 가입이나 변경을 막을 만큼 심각한 것은 아니라고 생각할 수도 있지만 기업이 고객 정보를 활용해 수익성을 높이거나 매출을 확대하기 위해 분석 작업을 할 경우는 상당히 심각한 문제가 발생할 수 있는 소지를 남기는 것이다.

또한 현행 데이터의 품질 평가 결과에 따라 조직 간의 마찰과 불신이 심해질 가능성도 존재한다. 데이터 소유주나 관리자에 대한 명확한 정의가 사전에 없는 경우 데이터 품질에 대한 결과는 현업 부서와 IT 부서 간의 불신감과 책임 공방만을 불러일으키고 현실적인 품질 관리를 위한 프로젝트 진행에 어려움을 가중시키는 요인이다.

대상 범위 설정의 어려움, 불확실한 목표 설정

실패하는 프로젝트의 대부분은 범위나 목표의 설정이 불확실한 경우가 많다. 데이터 품질 관리 프로젝트 역시 데이터 가치와 품질 관리 비용 분석, 목표 데이터 품질에 대한 명확한 정의가 없을 경우 많은 어려움을 겪게 될 것이다. 따라서 데이터 품질 관리 프로젝트를 수행하기 전에 측정 가능하고, 현업과 IT의 기대를 맞춘 목표를 설정하는 것이 무엇보다 중요하다.

기술 전문가 및 참조 성공 사례 부족

데이터 품질 관리는 최근에 본격적으로 이슈가 되고 있는 부분이다. 데이터 품질 관리에 대한 종합적인 기술과 프로젝트 수행 경험을 가진 전문가가 매우 부족한 것 또한 현실이다.

특히 전사적인 데이터 품질 관리를 성공적으로 수행한 사례 역시 부족해 프로젝트 수행에 따른 경험 지식을 공유할 기회가 매우 제한적이다. 따라서 데이터 아키텍처에 대한 이해와 모델링 수행 능력, ETL과 같은 데이터 활용 능력은 물론 비즈니스 도메인에 대한 이해가 풍부한 외부 전문가와 내부의 업무 전문가 및 IT 조직 구성원들의 협업이 반드시 요구된다.

〈그림 6〉 데이터 품질 관리 조직

프레임워크 기반의 데이터 품질 관리 활동

데이터 품질 관리 활동이 한계를 극복하고 나아갈 방향을 살펴보기 전에 데이터 품질 관리의 목적을 다시 보면 데이터 소스, 수집, 변환과 통합, 데이터 웨어하우스, 분석에 이르는 기업의 데이터 가치 사슬의 품질을 관리함으로써 조직 목표를 달성하고 데이터 자산의 보호 및 활용도를 제고하는 것이다.

즉 데이터 품질은 단순히 기술적인 이슈가 아니며 프로세스 및 조직적인 이슈 등 보다 포괄적인 원인으로 발생하므로 해결 방법 역시 기술 · 프로세스 · 조직의 전 방위적 관점에서 효과적 접근이 필요하다. 또한 IT 중심이 아닌 비즈니스 중심의 데이터 품질 관리 접근을 강조해 데이터 관리의 전 범위를 망라하는 프레임워크 중심의 포괄적 접근이 필요하다.

품질 관리 원칙 및 목표 정의

데이터 품질 관리 목표 정의 및 전략적 방향 설정을 하는 것으로 조직의 목표를 달성하기 위해서 데이터 품질 관리 활동에 관한 전사 규정을 수립하는 활동을 의미한다.

품질 관리 조직

데이터 품질 관리의 목표를 달성하기 위한 조직 구조, 역할, 필요 기술 및 활동을 정의하는 것으로 전략, 구현, 운영이라는 정보 시스템 사이클과 유기적으로 연계되어야 하며 시스템별 데이터 담당자와 함께 전사적 관점에서 데이터를 관리하기 위한 데이터 관리 담당자(스튜어드)를 정의해 이를 기반으로 전사적 데이터 정합성을 보장할 수 있어야 한다.

아키텍처

데이터 품질의 개선을 위한 데이터, 애플리케이션, 기술 아키텍처의 확립을 의미하는 것으로 사실상 데이터 품질 관리란 데이터 아키텍처를 비롯한 각 아키텍처의 구성요소 품질을 높이는 것과 밀접한 상관관계를 가진다. 데이터 아키텍처의 프레임워크 구성 요소인 데이터 구조, 데이터 관리, 데이터 이동과 변환의 단계별 품질을 높이는 것이 성숙도 높은 데이터 품질 관리의 모습이 된다.

데이터 관리 프로세스

메타데이터 관리, 표준 데이터 관리, 데이터 품질 관리 및 데이터 품질 관리 운용 등 전사 데이터 관리 기능을 정의하고 기존의 IT 시스템 전략 · 구현 · 운용 프로세스와 연계해 데이터 품질 관리가 효율적이고 효과적으로 수행될 수 있는 절차를 규정하는 것이다. 예를 들어 데이터 품질 관리 활동은 일정한 프로세스로 수행될 수 있으며, 이외에도 프레임워크에서 정의되는 활동들의 기능과 절차가 규정되어야 한다.

데이터 및 업무 규칙 관리를 위한 메타 시스템

전사적으로 일관된 데이터 정의 및 비즈니스 규칙의 유지/관리를 위한 표준 데이터를 포함한 시스템에 대한 메타 데이터를 수집하고 표준 및 원천 데이터에 대한 매핑 기능을 통해 변경 관리 및 영향 관리를 수행해야 한다. 메타 시스템은 메타 데이터 및 표준 데이터를 관리하기 위한 리파지토리를 기반으로 구축되어야 하며 데이터 관리 프로세스의 지원을 위해 워크플로우 기능을 포함할 수 있어야 한다.

그 밖에도 데이터 품질의 표준 및 평가 지표의 정의, 품질 관리 과정의 지속적인 모니터링 및 효과적 수정 조치, 데이터 품질 관리 활동에 대한 평가 및 보상 체계, 데이터의 오용이나 남용을 막기 위한 보안 및 권한 등이 정의되어야 전 방위적인 데이터 품질 관리를 수행할 수 있다

성공적인 데이터 품질 관리를 위해

지금까지 저품질의 데이터가 비즈니스에 미치는 영향, 데이터 품질 관리의 개념, 데이터 품질 관리 활동의 한계점 및 이를 극복하기 위한 프레임워크 중심의 데이터 품질 관리 활동을 살펴봤다. 데이터 품질 관리는 무엇보다 페이퍼 작업으로 끝나지 않고 실제 비즈니스 효과를 가져와야 한다. 데이터 품질에 대해 고민하고 있는 기업의 대부분은 겉으로 말하지는 않지만 마음속으로는 한 가지 공통점을 갖고 있다. 그것은 데이터에 대해 더 자세한 초상화를 그림으로써, 더 나은 서비스를 제공해 매출을 증대시킨다는 목적이다. 이를 위해서는 데이터 품질 문제가 더 이상 후방 지원 시스템의 문제가 아니라 데이터를 다루는 모든 직원이 회사나 조직이 결정한 데이터 지침에 따라야만 하는 문제로 이동해야 함을 의미한다.

마지막으로 성공적인 데이터 품질 관리를 위해 고려할 사항을 제시하며 글을 맺고자 한다.

1. 데이터 기여도와 품질 관리 난이도를 분석해 확실한 비즈니스 효과를 볼 수 있는 부분을 선정해 성공 사례를 만든다.

2. 데이터 품질 관리는 시간과 자원, 즉 비용이 드는 프로젝트임을 공감한다. 이를 위해 저품질 데이터로 인해 야기되는 비용을 구체적으로 제시하기 위한 노력이 필요하다.

3. IT 부서와 비즈니스 부서를 함께 프로젝트에 참여시키고 CIO 이상의 고위 관계자의 지원을 획득한다.

4. 데이터 품질은 물론 데이터 아키텍처, 데이터베이스 전문 인력과 함께 프로젝트를 수행하며 Coaching 기법을 통해 내부 조직원의 역량을 강화한다.

5. 검증된 방법론과 솔루션 활용을 적극 검토하되 자체 개발 비용과의 비교를 통해 자신의 조직에 적합한 품질 관리 방법을 적용한다.

6. 데이터 품질 관리를 위한 공식적인 조직과 절차를 구현하고 지속적으로 품질 관리 활동을 수행한다.

7. 데이터 품질 관리 성과지표를 설정하고 ROI를 측정한다.