아이리포 지식창고

데이터 - 01 데이터 품질 (Data Quality)

씨에스리 권위수
컴퓨터시스템응용기술사
(delphkws@gmail.com)

데이터 품질 관리중요성 과 품질진단 이야기

Concept	**(데이터 품질 관리란)** '데이터 품질관리'란 조직 내외부의 지식 노동자와 최종 사용자의 기대를 만족시키기 위한 지속적 인 데이터 및 데이터 서비스 개선 활동을 이야기하며, 데이터 수집, 처리, 저장, 활용단계에서 최신성, 정확성, 상호연계성을 확보하여 사용자에게 가치 있는 데이터 전달하는 일련의 활동을 이야기 합니다. **(데이터품질관리 주요내용)** 데이터를 지속적으로 관리하기 위한 데이터 관련조직, 지침 및 가이드, 프로세스가 필요하며, 가이드라인에는 데이터를 효율적으로 관리하기 위한 데이터 관리정책, 원칙, 데이터 관리체계, 데이터 표준, 데이터 구조, 데이터 값 등에 내용을 포함하고 있습니다..
KeyWord	**데이터관리체계, 데이터 값, 데이터 구조관리, 데이터 표준관리**

데이터 품질 관리의 중요성

현대의 기업은 비즈니스 환경이 점점 빠르게 변화하고 있고, 서비스가 다변화 되면서 기업은 서비스 및 마케팅 전략을 수립하기 위하여 데이터 분석 및 통계 자료를 적극적으로 활용하고 있으며, 데이터를 활용해 실시간 의사결정을 하는 데이터 경영 중심으로 빠르게 변화하고 있습니다. 또한 최근 기업들은 빅데이터 활용을 통해 새로운 비즈니스를 모색하고 있으며 빠르게 다가 오는 제 4 차산업혁명 시대를 준비하고 있습니다.

기업은 내부 데이터를 활용하는 단계에서 외부 데이터의 연계를 활용하여 새로운 신규 비즈니스 융합을 시도하고 있습니다. 이렇게 데이터 기반으로 하는 마케팅과 전략을 수립하고 있음에도 정작 국내에 금융을 제외한 나머지 산업에서는 데이터에 관리를 위한 노력이 소홀한 실정이며, 데이터 품질을 관리하기 위한 조직에 인력을 충원하여 투자하는 것이 불필요 하다고 생각하는 경우 많이 있습니다. 이로 인한 **데이터 품질저하로 발생하는 손실 비용은 전체 예산의 10~15%에 달하며, 데이터 재 구축 비용, 전략적 마케팅 저하, 기업 신뢰도 추락, 기업 매출 수익 감소로 직결될 수 있습니다**.

최근 공공에서는 데이터를 민간에게 개방하여 새로운 비즈니스 및 일거리를 창출할 수 있도록 데이터를 제공하고 있습니다. 아직은 데이터 개방 초기 단계로 민간 사용자 중심으로 데이터 관리하기 보다는 기관

업무 데이터에 관리에 초점이 맞추어져 있습니다. 그리고 일회성이 아닌 지속적으로 관리할 수 있는 제도 마련을 통해 고품질의 데이터를 민간에 제공할 수 있도록 정부의 관심이 필요해 보입니다.

국내외 사례

최근 데이터 품질 이슈로 발생하는 국내 및 해외 사례를 통해서 품질의 중요성을 이야기 해 보겠습니다.

국토부의 최근 전국 3,733 만 필지의 토지·임야대장과 707 만동에 대한 건축물 대장 자료를 부동산등기부와 비교 분석한 결과, 토지·임야대장의 자체 오류가 약 560 만건인 것으로 분석되었습니다. 면적이 0 인 경우, 지목코드가 누락되었거나 잘못 입력된 경우, 주요 항목이 누락된 경우가 존재하였으며 이로 인해 국민들이 부동산 거래 시 많은 불편함을 감수해야 합니다.

국내 주식시장에서 코스피 200 을 구성하는 기업의 시가총액이 잘못 적용되어 코스피 200 지수가 실제보다 높게 산출된 사건입니다. 이 사건으로 현물과 선물의 가격차이인 베이시스가 1 포인트가량 왜곡되면서 차익거래 투자자들에게 혼란을 초래 하였으며. 오류가 수정되기 전까지 800 억원 가량의 프로그램 매물이 쏟아져 코스피 지수에도 많은 영향을 끼쳤습니다.

2007 년 일본에서는 5000 만여명이 낸 국민연금 기록이 유실된 적이 있습니다. 25 년간 연금을 내야 받을 수 있는 연금 명세서가 정부 데이터 관리부실로 사라졌습니다. 이 여파로 '연금기록 확인이 불가능하다'는 정부측 답변에 국민은 공황에 빠졌었다. 당시 정부와 여당에 대한 국민의 불신을 증폭시켰고, 결국 자민당은 선거에서 지는 요인 중 하나가 됐었다고 합니다.

미 최신예 순양함인 포트로얄호가 호놀룰루 앞바다에 좌초 되었습니다. 미 해군 안전국의 발표에 의하면 항법시스템의 오류를 추정 원인으로 손 꼽았습니다. 비행기 혹은 선박에는 자동항법시스템(Auto Pilot System)이 있습니다. 이는 원거리 운항시 미리 설정된 데이터에 따라 자동으로 방향각을 조절하여 운항되도록 만든 시스템 입니다. 시스템에 초기 입력된 좌표 데이터에 오류가 있다면 어떠한 일이 발생될 것인가? 말할 나위 없이 잘못된 위치를 향해하거나 비행하여 대형 사고가 발생할 가능성이 있게 됩니다. 포트로얄호의 좌초로 인해 선박의 수리 비용 견적만 4,000 만 달러가 나왔으며 수리기간 7 개월이라는 손실을 입었습니다. 이 사건으로 미 해군은 국제적 망신을 당했습니다

미국발 금융위기를 겪은지 어느덧 수해가 지나가고 있습니다. 앨런 그린스펀 전 미연방준비제도이사회(FRB) 의장은 금융위기 원인중의 하나가 "부정확한 데이터" 때문이라고 했습니다. 당시 글로벌 금융기관들은 최첨단 슈퍼컴퓨터로 무장한 위험예측시스템을 운영하고 있었으며 이들을 통해서 정부규제 없이 위험관리가 가능하리라 믿습니다. 그러나 입력되는 원천데이터가 잘못돼 제대로 위험을 예측하고 관리할 수 없었습니다.

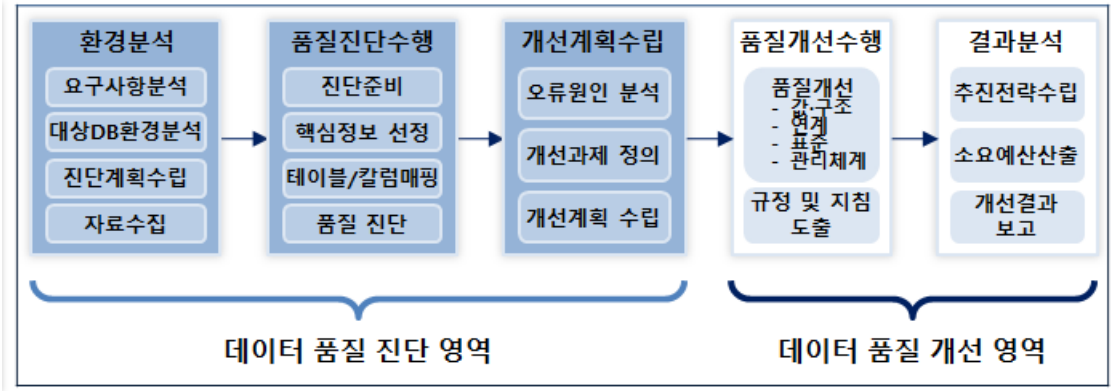
위 사례 통해 데이터 품질의 관리가 공공 및 기업에 필요한 상황이며, 향후 데이터는 융·복잡화 되면서 데이터 품질관리가 IT 의 핵심 요소로 자리 매김될 것으로 예상 됩니다..

국내에서는 행자부 소속기관인 정보화진흥원에서 공공데이터 품질 향상을 위해 매년 공공기관에 품질컨설팅 사업을 적극적으로 추진하여 공공데이터 품질을 개선 하려고 노력하고 있습니다. 기업은 데이터베이스진흥원에서 데이터베이스 인증심사제도 마련 및 기업의 품질 중요성을 홍보하는 활동을 꾸준히 하고 있습니다.

실무 기반 데이터 품질 컨설팅 경험 및 품질진단 방법 이야기

데이터 품질 컨설팅 수행 시 보통 아래와 같이 컨설팅을 수행하게 됩니다. 진단 영역의 환경분석 단계에서는 현행 DB 환경과 함께 관련 산출물 수집, 업무의 전반적인 이해를 위해서 담당자와 미팅을 수행하며, 품질진단 수행 단계에서는 진단대상 선정 및 품질진단 수행, 개선계획수립 단계에서는 오류에 대한 원인을 분석을 통해 개선 계획을 수립하게 되며, 품질개선단계에서 계획 일정에 따른 품질 개선을 수행 합니다.

[표-1] 데이터 품질 진단 및 개선 절차



기업 및 공공의 데이터 품질 컨설팅을 수행하기 전 데이터 품질컨설팅 계획을 수립하고 보유하고 있는 데이터에 대한 현황분석을 수행합니다. 수행하는 단계에서 담당자와 업무 미팅을 수행하면서 데이터 관련 산출물을 요청 합니다. 그러나 기업 및 공공기관이 데이터 관련 산출물을 관리하지 않고 통제하는 조직이 없는 실정이며 시스템을 유지보수 담당자 수준에서 관리하고 있고, 만약 있더라고 현행화가 되지 않고 있는 실정 인 경우가 대부분 입니다. **시스템 담당자는 데이터 품질의 중요성은 알고 있으나 데이터 품질을 관리하는 조직 및 인원 충원에 대한 투자는 어려운 상황을 이야기 하면서 너무 이상적인 관리라고 하시는 분들도 많이 있습니다**. 데이터를 관리할 수 있는 체계 구축이 없이 데이터를 관리한다는 것은 부실한 기초 위에 아름다운 성을 쌓은 것과 같다고 볼 수 있습니다. 진정으로 기업 및 공공기관의 고품질 데이터를 관리하고 싶다면 먼저 품질을 관리하는 인원을 충원해 담당자의 역할과 프로세스가 정의 하여야 합니다. 예산 부족으로 데이터 품질이 어려운 상황에서는 기업 및 공공기관에 맞는 최소한의 데이터 품질 관리를 할 수 있도록 또한 가이드라인 마련도 필요 합니다. 그러나 **기업이나 기관에 수준에 맞는 가이드가 아닌 획일적이고 복잡한 수준에서 가이드라인 제공하고 있거나, 문서 수준에서 관리되고 있는 경우도 많이**

존재합니다. 이를 개선하기 위한 수준별 데이터 관리 가이드라인을 만들어 기업 및 공공에 확산할 필요가 있습니다.

그리고 데이터 품질을 관리하는 활동이 대부분 시스템 구축이 되고 난 후부터 수행하는 것으로 오인되는 하는 경우가 많은데 시스템 구축, 운영, 활용, 폐기 단계 까지 전반적인 품질관리가 필요합니다.,

보유 데이터를 잘 관리하기 위해서는 기업 및 기관에 맞는 데이터 품질 가이드 라인을 정의하고 연차별 개선방향을 수립하여 조금씩 품질관리활동 수행할 수 있는 실질적 계획 수립이 필요합니다. 데이터를 관리하기 위한 **기본적인 데이터관리지침 및 품질가이드 정의가 필요하며, 구체인 데이터 값, 데이터 구조, 데이터 표준관리, 연계관리 등의 세부내용을 작성하여 배포 및 교육하고 통제 관리해야 합니다. 대부분 데이터 품질관리 활동은 아래와 같이 정의하여 수행 합니다**.

[표-2] 데이터 품질 진단 및 개선 내용

구분	데이터 진단	데이터 개선
데이터 값	- 데이터 프로파일링 기반의 품질 진단 - 필수 값, 유효 값, 패턴, 날짜, 코드 등 - 업무규칙 기반의 품질 진단	- 오류 추정 데이터 원인 분석 및 정비 - 모델 재설계 및 변경 후 정제 데이터 이관
데이터 표준	- DB의 데이터 요소에 대한 명칭, 정의, 형식, 규칙에 대한 수립을 통한 기관 차원의 적용 및 지속적 관리를 위한 제반 활동	- 데이터 요소 표준화 - 전사표준코드 및 용어 준용 - 코드체계 표준화 및 준용
데이터 구조	- 데이터 중복, 미사용, 산출물 현황, 식별자, 정규화 등의 관점을 진단	- 데이터 모델 재설계 및 변경 - 데이터 모델 설계 표준 적용 - 데이터 이력 관리항목 적용 - 미사용 테이블 정비
데이터 관리체계	- 전사 및 DB차원의 조직, 지침·가이드, 프로세스, 도구(인프라) 관점에서 품질진단	- 데이터품질 규정 마련 - 데이터 연계방식 절차 및 관리정책 수립 - 근거 규정 및 업무규칙 준수 - 데이터 오너쉽 조직 및 절차 수립

데이터는 입력, 저장, 변경, 흐름을 가지면서 움직이는 데이터로 지속적인 데이터 관리가 필요합니다. 보통 데이터 값에 대한 시스템에 의한 원천적 데이터 진단도 더불어 **현실 세계의 정보와 입력된 데이터의 차이가 나는지도 관리 되어야 합니다. 이러한 데이터는 데이터 관리체계를 통해서 데이터에 지속적 현행화가 필수적으로 필요합니다.** 공공인 경우 법규 및 제도적 개선을 통해서 데이터를 지속적으로 관리할 수 있도록 개선하는 것도 방법입니다. 또한 시멘틱한 품질관리도 매우 중요하며, 단순 TEXT 는 데이터 품질 도구로 측정 하는데 한계가 있으며 사용자의 교육 및 관리를 통해서만 지속적으로 개선을 할 수 있는 품질활동 입니다.

필자가 경험한 2015 년도 인허가 데이터 품질 진단한 경험을 이야기 해보면 인허가 데이터는 시·군·구 지방자체단체에서 입력하는 자료로 인허가 업종에 대한 휴폐업 정보를 새올 행정정보시스템에서 관리하는 시스템 입니다. 휴폐업에 대한 현행화는 지속적으로 관리되어야 하나 휴폐업 신고에 대한 법규에 의무 사항이 아닌 이유로 데이터 관리가 어려운 상황입니다. 이로 인해 단순 시스템을 통한 진단을 하는 경우에는 오류가 없다고 판단되지만, 현실세계와 비교한다면 상당히 많은 데이터 오류가 존재할 수 있습니다. 인허가 데이터는 소상공인들이 상권데이터 분석시스템에 활용되고 있습니다. 현행화되지 않은 데이터는 상권에 대한 분석 시 잘못된 결과를 제공할 수 있고 소상공인의 사업 실패로 이어질 수 있어 데이터 관리가 정말 중요한 데이터 라고 판단 됩니다.

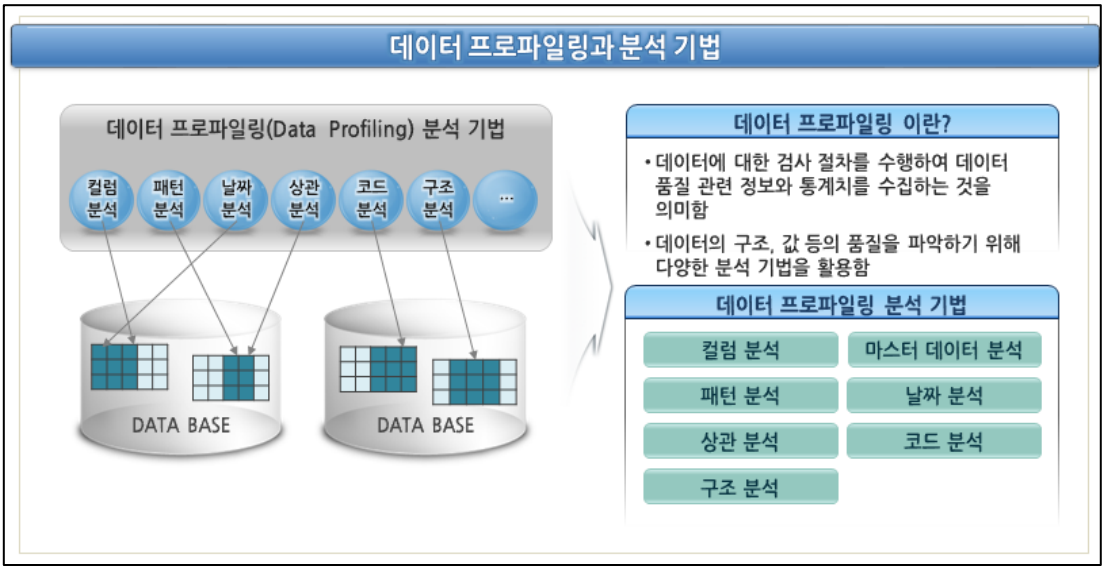
다시 본론으로 돌아와서 데이터 품질진단 절차 와 데이터 값 진단이 기본이 되는 데이터 프로파일링에 대해서 알아 보겠습니다.

데이터 품질진단 절차는 데이터 품질진단의 목적, 범위, 계획수립, 품질진단 대상을 선정, 품질의 이슈사항을 점검 및 품질진단 핵심항목을 도출하여 데이터 프로파일링 및 데이터 업무규칙 품질진단을 수행하고 진단된 결과 값에 데이터 원인을 분석하여 결과를 보고하는 순으로 진행 합니다.

여기에서 데이터 프로파일링이란 데이터에 대한 검사를 수행하여 데이터 품질 관련 정보와 통계치를 수집하는 것을 의미합니다.

데이터 값 진단에 가장 기본적인 진단을 수행하는 방법으로 컬럼, 날짜, 패턴, 코드, 상관분석 등이 존재 합니다.

[표-3] 데이터 프로파일링 분석 방법



컬럼분석	통계적 기법을 통한 데이터 분석
	- NULL 개수, Space 개수,Min Max, 평균, 표준편차, 분산
날짜분석	Data Type 은 Character 이나 의미상 날짜/시간 유형 데이터에 대한 유효성 분석
	- 고객의 생년월일에 대한 YYYYMMDD 날짜유형 오류 분석

패턴분석	문자, 숫자 등으로 구성된 특정 패턴을 갖는 값에 일관된 패턴 여부 점검
	- 주민등록번호, 사업자번호, 법인번호 자리 수 체크
코드분석	개별코드/통합코드 마스터와 트랜잭션 코드와의 유효성 분석
	- 고객의 직종 세부코드
상관분석	부모 자식 관계 데이터의 참조 무결성 분석
	- 상품 주문시 고객번호는 고객의 고객번호를 참조

프로파일링은 데이터간에 연관된 업무 규칙 보다는 하나의 단일 테이블로 구성된 컬럼에 대한 단순 진단이라고 할 수 있습니다. 기준이 되는 마스터 데이터 진단 시 용이하게 사용되며, 데이터 프로파일링에 **가장 어려운 점이라고 생각되는 것은 유지보수 담당자의 협조라고 생각됩니다. 컬럼분석 및 코드분석등은 유지보수 담당자만이 구체적으로 알고 있어 제공하지 못하는 경우 대략적인 수준에서 진단을 하는 경우도 존재합니다.** 담당자가 없는 경우 관련 매뉴얼이나 산출물 기반으로 진단을 수행하나 데이터 진단에 대한 한계가 존재합니다. 단순 컬럼을 진단 하지만 정확도를 높이면 관련 데이터의 통계 신뢰도 및 마스터 데이터 정확도를 높여 기업의 전략적 마케팅 활용에 큰 도움을 줄 수 있습니다. 진단 수행 시 마스터 데이터 및 핵심이 되는 데이터를 최우선 대상으로 수행하고 나머지를 순차적으로 진단하는 것이 진단의 효율적인 방법이라고 할 수 있습니다.

데이터 업무 규칙 진단이란 데이터 사용자가 요구하는 수준을 만족시키기 위하여 업무적으로 규정된 기준에 맞도록 데이터 값을 관리하기 위한 조건에 대한 일반적 표현 이라고 할 수 있습니다. 데이터 업무 규칙은 해당 서비스를 제공함에서 업무상 규정된 절차에 대한 규칙 이라고 할 수 있으며. 공공에서는 법규에 의해서 업무 규칙이 만들어 지는 경우가 상당히 많이 존재 합니다.

예를 들면 자동차운전면허는 1 종 대형, 1 종 특수(대형견인+구난), 2 종 소형면허 3 개를 취득하면 대한민국에 존재하는 모든 자동차를 운전할 수 있습니다. 정의 되어 있다면 2 종 면허 취득자가 1 종 대형 자동차에 등록되어 있다면 데이터 오류로 볼 수 있습니다. 업무규칙은 단순한 한 개의 테이블로도 만들어 지나 보통 여러 개의 복합적 관계에서 만들어 지는 경우가 많이 존재 합니다. 업무 특성을 많이 이해하고, 관련 담당자와 업무 미팅 및 관련된 메뉴얼 및 규정서를 통해서 업무 규칙을 도출하여 진단을 수행할 수 있습니다.

이상으로 데이터 품질이 무엇이며, 데이터 품질관리의 중요성과 데이터 품질 진단 및 개선 절차에 대해 간단하게 알아보았습니다.

다음에는 구체적인 진단 및 개선에 대한 컨설팅을 어떻게 수행하는지 상세 하게 컨설팅 보고서 기반으로 알아 보겠습니다.

"끝"

Contents connect communications!!

아이리포에 오시면 더 많은 지식을 가져가실 수 있습니다.

아이리포 온라인 : http://www.ilifo.co.kr
아이리포 지덤시리즈 : http://www.jidum.com
아이리포 IT지식창고 : https://www.ilifo.co.kr/boards/knowledge
아이리포 기술사/감리사 카페 : http://cafe.naver.com/itlf

서울시 마포구 상암동 1610번지, DDMC 3층 아이리포 교육센터
TEL: 02-303-9997 | MAIL: edu@ilifo.co.kr

