TTA Standard

정보통신단체표준(국문표준)

TTAK.KO-

제정일: 2019년 OO월 OO일

데이터 프로파일링 절차

Data Profiling Procedure

한국정보통신기술협회
Telecommunications Technology Association

표준초안 검토 위원회 메타데이터 프로젝트그룹(PG606)

표준안 심의 위원회 소프트웨어/콘텐트 기술위원회(TC6)

	성명	소 속	직위	위원회 및 직위	표준번호
표준(과제) 제안	임성준	한국데이터산업진흥원	팀장	PG606 의장	
표준 초안 작성자	이창한	한국데이터산업진흥원	연구위원	PG606 위원	
사무국 담당	김재웅	TTA	단장	-	
	민선미	TTA	책임연구원	-	

본 표준 발간 이전에 접수된 지식재산권 확약서 정보는 본 표준의 '부록(지식재산권 확약서 정보)'에 명시하고 있으며, 이후 접수된 지식재산권 확약서는 TTA 웹사이트에서 확인할 수 있습니다.
본 표준과 관련하여 접수된 확약서 외의 지식재산권이 존재할 수 있습니다.

발행인 : 한국정보통신기술협회 회장
발행처 : 한국정보통신기술협회
13591, 경기도 성남시 분당구 분당로 47
Tel : 031-724-0114, Fax : 031-724-0109
발행일 : 2019.xx

서 문

1 표준의 목적

이 표준의 목적은 조직이 데이터 값의 관점에서 데이터 품질 평가를 수행할 때에 데이터 프로파일링의 절차를 진행하는 방법을 제시하고, 평가 대상 데이터셋의 컬럼 특성과 구조 특성을 도출하는 방법을 제시한다.

2 주요 내용 요약

데이터 프로파일링이란 데이터의 값과 구조의 특성을 결정하기 위하여 데이터 분석 기술들을 실제 운영되고 있는 데이터에 적용하는 것이다. 데이터 프로파일링은 평가 대상이 되는 데이터셋에 대한 기존의 메타데이터가 불완전하고 부정확하다는 가정에서 출발한다. 데이터 프로파일링을 통하여 실제 데이터에 대한 메타데이터를 역공학적으로 도출하고 이를 기존의 메타데이터와 비교하여 보다 정확하고 완전한 메타데이터를 생성한다. 이러한 과정을 반복해 나가면서 보다 메타데이터의 정확성과 완전성을 높여 나간다. 이렇게 결정된 메타데이터는 데이터 프로파일 형태로 정리되며, 이는 데이터 규칙 작성에 기초가 된다.

3 인용 표준과의 비교

3.1 인용 표준과의 관련성

- 해당사항 없음

3.2 인용 표준과 본 표준의 비교표

- 해당사항 없음

Preface

1 Purpose

The purpose of this standard is to provide the data profiling procedure in which an organization implements in data quality assessment in terms of data values, and the method for deriving the column properties and structural properties of target data sets that form the basis of data rules.

2 Summary

Data profiling means applying data analysis techniques to data in actual operation to determine the values and structural properties of data. Furthermore, it is based on the surmise that the existing metadata of the data set to be evaluated are incomplete and inaccurate. In data profiling, metadata are reverse-engineered for actual data, and compared with the existing metadata to generate more perfect and correct metadata. The repetition of this process improves the accuracy and completeness of the metadata. Furthermore, the metadata determined through this process is arranged in the form of a data profile, which forms the basis of establishing data rules.

3. Relationship to Reference Standards

None

목 차

데이터 프로파일링 절차
(Data Profiling Procedure)

1 적용 범위

이 문서는 데이터 프로파일링의 세부 절차를 명세한다. 조직은 데이터 프로파일링을 통하여 대상 데이터셋에 대한 정보를 추출할 수 있다. 데이터 프로파일링은 컬럼의 정확한 특성들을 도출하며, 데이터셋의 정확한 구조를 파악할 수 있다. 이는 데이터 프로파일과 데이터 품질 이슈로 정리되어 데이터 규칙을 정의하는데 이용되며 데이터 품질을 평가하는데 기초가 된다.

이 문서의 적용 범위는 데이터 품질 이슈와 데이터 프로파일을 작성하기 위한 전반적인 절차, 데이터셋과 관련된 메타데이터의 수집, 데이터셋의 샘플링, 단일 컬럼에 대한 컬럼 특성 분석을 통한 컬럼 프로파일 생성, 그리고 다중 컬럼에 대한 구조 분석을 통한 데이터 프로파일 생성이다. 또한 프로파일링 대상 데이터셋은 관계형 데이터베이스 구조를 지니거나 관계형 데이터베이스 구조로 변환될 수 있어야 한다.

2 인용 표준

해당사항 없음

3 용어 정의

3.1 DBMS(database management system)

데이터베이스를 구성하고 이를 응용하기 위하여 구성된 소프트웨어 시스템(TTA 용어사전)

3.2 관계형 데이터베이스(relational database)

관계들로 구성된 데이터베이스. 관계 데이터베이스의 데이터베이스 관리 시스템은 데이터 원소들을 재결합시켜 새로운 관계를 만들 수 있으며, 이로 인해 데이터의 이용에 많은 다양성이 있다.(TTA 용어사전)

3.3 널(null)

정보의 부재. “0”이나 공백과는 달리 정보가 없음을 나타냄(TTA 용어사전)

3.4 데이터 사전(data dictionary)

자료에 관한 정보를 모아 두는 저장소. 자료의 이름, 표현 방식, 자료의 의미와 사용 방식, 그리고 다른 자료와의 관계를 저장한다. 데이터베이스의 데이터 사전은 그 자신이 하나의 데이터베이스를 이루며 데이터베이스 시스템의 다양한 스키마들, 즉 내부/외부 그리고 개념 스키마 등을 저장하고 있다.(TTA 용어사전)

3.5 데이터 아키텍처(data architecture)

기업의 전사적 아키텍처(EA)의 중요한 하부 구조로, 데이터 측면에서 기업 시스템을 처음부터 끝까지 조망하여 시스템의 본질인 데이터를 체계적·구조적으로 관리하고 설계하는 전 과정. 기업의 핵심 자산인 데이터를 전사적 관점에서 구조적으로 조망하고 리모델링한다는 것을 목표로, 데이터에 관한 모든 계층을 총망라해서 객관적이고 구체적인 접근 방법을 명시한 체계적인 방법론이며, 기존의 데이터 모델링을 포함한 포괄적인 개념이다.(TTA 용어사전)

3.6 데이터 프로파일(data profile)

프로파일링 대상 데이터셋의 특성을 나타내는 메타데이터 객체. 데이터 프로파일링 결과, 즉 데이터셋의 컬럼 특성과 구조 특성에 대한 데이터를 저장한 객체이다.(Data modelling, ETL and data quality guide 11g release 2, Oracle)

3.7 데이터 프로파일링(data profiling)

데이터셋의 실제 구조, 내용 및 품질을 발견하기 위하여 해석적 기술을 사용하는 절차. 데이터셋의 컬럼에 포함된 값들의 특성과 데이터셋의 컬럼들간의 구조 특성을 파악하고 아울러 데이터 품질 이슈를 도출하는 것이다.(Olson, Jack E. Data Quality: The Accuracy Dimension, Elsevier Science. Kindle Edition.)

3.8 데이터베이스(database)

① 주어진 목적이나 주어진 자료 처리 시스템에 사용하기에 충분하도록 적어도 한 개 이상의 파일로 구성한 자료의 집합. ② 여러 사람에 의해 공유되어 사용될 목적으로 통합하여 관리되는 데이터의 집합 또는 여러 응용 시스템들의 통합된 정보들을 저장하여 운영할 수 있는 공용 데이터들의 묶음. 데이터베이스의 특징으로 여러사람의 데이터 동시 공유, 데이터 중복 최소화, 데이터의 특성 및 상호 관계 등을 통한 참조, 데이터 무결성, 보안성 유지 등을 들 수 있다.(TTA 용어사전)

3.9 데이터셋(dataset)

데이터의 집합. 일반적으로 단일 데이터베이스 테이블이나 단일 통계 데이터 매트릭스의 내용에 해당되며, 테이블의 모든 컬럼은 특별한 변수를 표현하며, 각 로우는 데이터셋의

주어진 요소에 해당한다.(What is a database?, DATABASE.GUIDE, retrieved 27 May 2019))

3.10 데이터셋 샘플링(dataset sampling)

품질 평가 대상이 되는 전체 데이터셋을 대표하는 일부 데이터셋을 표본 추출하는 절차.(Olson, Jack E.. Data Quality: The Accuracy Dimension, Elsevier Science. Kindle Edition.)

3.11 도메인(domain)

관계 데이터베이스에서 하나의 속성이 취할 수 있는 값의 집합.(TTA 용어사전)

3.12 로우(row)

관계형 데이터베이스에서 테이블 내에 단일의 암묵적으로 구조화된 행. 데이터베이스의 테이블은 로우들과 컬럼들로 구성된다고 할 수 있다. 테이블의 각 로우는 관련된 데이터의 집합을 나타내고 테이블의 모든 로우는 동일한 구조를 갖는다. 튜플(tuple)이라고도 함.("What is a database row?" Cory Janssen, Techopedia, retrieved 27 May 2019)

3.13 메타데이터(metadata)

일련의 데이터를 정의하고 설명해 주는 데이터. 컴퓨터에서는 데이터 사전의 내용, 스키마 등을 의미하고, 하이퍼텍스트 마크업 언어(HTML) 문서에서는 메타 태그 내의 내용이 메타데이터이다. 방송에서는 방대한 분량의 저작물을 신속하게 검색하기 위해서 프로그램 제작 시 촬영 일시, 장소, 작가, 출연자 등과 음원의 경우 작곡자나 가수명 등을 메타데이터로 처리한다. 메타데이터는 여러 용도로 사용되나 주로 빠른 검색과 내용을 간략하고 체계적으로 하기 위해 많이 사용된다. 엠펙(MPEG)에서는 메타데이터에 대한 표준으로 엠펙-7(MPEG-7) 표준을 제정했다.(TTA 용어사전)

3.14 메타데이터 레포지토리(metadata repository)

메타데이터를 저장하기 위해 생성된 데이터베이스. 레포지토리는 레지스트리와 비교했을 때 부가적인 기능을 갖는다. 메타데이터 레포지토리는 메타데이터 레지스트리와 같이 메타데이터를 저장할 뿐만 아니라 관련된 메타데이터 타입과의 관계를 포함한다.

3.15 뷰(view)

하나 이상의 테이블로부터 데이터의 부분집합을 논리적으로 표현하는 논리적인 테이블(TTAK.KO-11.0089 정보검색엔진 품질평가 지침)

3.16 스키마(schema)

데이터베이스를 기술하기 위해 사용하기 시작한 개념. 데이터베이스의 구조에 관해서 이용자가 보았을 때의 논리 구조와 컴퓨터가 보았을 때의 물리 구조에 대해 기술하고 있다. 데이터 전체의 구조를 정의하는 개념 스키마, 실제로 이용자가 취급하는 데이터 구조를 정의하는 외부 스키마 및 데이터 구조의 형식을 구체적으로 정의하는 내부 스키마가 있다.(TTA 용어사전)

3.17 외래 키(foreign key)

관계형 데이터베이스에서 여러 테이블의 내용을 참조할 때 결합에 이용되는 참조하는 테이블의 컬럼. 여러 테이블의 내용을 참조하여 결과를 낼 때는 같은 의미를 가지는 컬럼 값의 연결을 통하게 되는데, 이 때 참조하는 테이블에서의 컬럼을 외래 키라 하며, 이러한 외래 키의 값은 참조되는 테이블에 반드시 존재하는 주 키(primary key) 값이어야 한다. 하지만 외래 키는 널(null) 값일 수도 있다.(TTA 용어사전)

3.18 일차 키(주키, primary key)

테이블 각 로우의 데이터를 유일하게 식별하게 하는 컬럼 또는 속성(TTAK.KO-11.0089 정보검색엔진 품질평가 지침)

3.19 정규형(normal form)

정규화되지 않은 형식에서 좀 더 단순하고 절약화된 형식으로 정규화된 관계 또는 데이터베이스. 관계의 정규형에는 5~6가지가 있는데, 이중에서 가장 흔히 사용되는 것은 제1정규형(1NF: first normal form), 제2정규형(2NF), 제3정규형(3NF)이다. 제1정규형은 가장 단순한 구조로 된 레코드의 집합(예: 직원 명단)으로 각 필드(열)는 식별 번호, 성명 등 고유의 중복되지 않는 정보만을 포함한다. 제2정규형과 제3정규형은 제1정규형을 분해하여 각 필드 간의 상호 관계를 점점 더 세부적으로 규정함으로써 몇 개의 다른 표들을 분리시킨다. 이 밖에 제4정규형(4NF), 보이스 코드 정규형(BCNF), 제5정규형(5NF)인 투영/통합 정규형(PJ/NF: projection-join normal form) 등이 있다.(TTA 용어사전)

3.20 컬럼(column)

관계형 데이터베이스 테이블에서 열을 말한다. 컬럼은 열이 어떻게 구성되어야 할지에 대한 구조를 제공한다. 관계형 데이터베이스 용어에서 컬럼과 같은 의미로 사용되는 것은 속성(attribute)이다. 필드(field)라고도 부름("What is Database Column? - Definition from Techopedia". Techopedia.com. Retrieved 2015-11-05)

3.21 컬럼 프로파일(column profile)

객체의 특성을 나타내는 메타데이터 객체. 컬럼 특성을 분석한 결과, 즉 컬럼 특성에 대한 데이터를 저장한 객체이다.

3.22 테이블(table)

하나 이상의 인수에 의해 애매모호하지 않게 관련 지어진 각 항목이나 자료 배열.(ISO)

4 약어

MD	Metadata
ERD	Entity relationship diagram
DDL	Data definition language
DBMS	Database management system

5 데이터 프로파일링 절차

5.1 목적

데이터 프로파일링은 데이터셋과 문서화된 메타데이터를 활용하여 데이터셋의 컬럼에 포함된 값들의 특성과 데이터셋의 컬럼들간의 구조 특성을 파악하고 아울러 데이터 품질 이슈를 도출하는 것이다. 데이터 컬럼의 특성과 컬럼 구조 특성은 데이터 프로파일의 형태로 정리되어 데이터 품질 평가에 있어서 데이터 오류를 측정하기 위한 데이터 규칙을 도출하는데 기초가 된다.

데이터 프로파일링을 하는 이유는 데이터셋과 관련된 문서화된 메타데이터 혹은 운영중인 데이터셋의 메타데이터가 데이터 품질 측면에서는 불완전하고 부정확하기 때문이다. 다시 말하면, 데이터 프로파일링은 반복적인 과정을 통해 불완전하고 부정확한 메타데이터를 데이터 규칙을 도출하기 위한 완벽하고 정확한 메타데이터로 수정하는 과정이라고 말할 수 있다.

5.2 구성

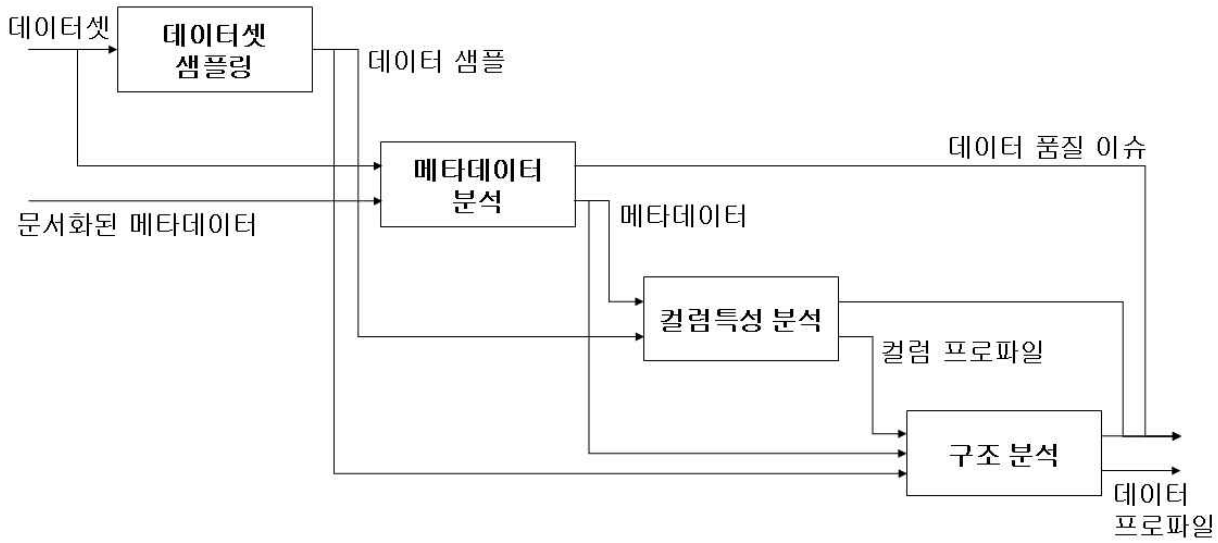
데이터 프로파일링 과정은 데이터셋 샘플링, 메타데이터 분석, 컬럼 특성 분석 및 구조 분석으로 구분된다.

데이터셋 샘플링은 식별된 데이터셋에 접속하거나 데이터셋을 로딩하여 프로파일링될 데이터 샘플을 결정하기 위한 도출 과정이다.

메타데이터 분석은 식별된 데이터셋과 관련된 문서화된 메타데이터를 수집하고, 운영중인 데이터셋의 메타데이터를 추출하며, 수집된 메타데이터 사이에 불일치가 존재하면 이를 데이터 품질 이슈로 등록한다. 이 과정에서 가능한 많은 메타데이터를 여러 소스로부터 수집하면, 다음의 컬럼 특성 분석과 구조 분석이 용이하게 수행된다.

컬럼 특성 분석은 결정된 프로파일링 대상 데이터셋의 개별 컬럼에 대하여 컬럼 특성들을 도출하는 것이다. 도출된 컬럼 특성들과 수집된 문서화된 메타데이터를 비교하여 정확한 컬럼 특성들을 결정한다. 이후 컬럼명을 중심으로 결정된 컬럼 특성들을 포함한 컬럼 프로파일을 생성한다. 또한 수정된 메타데이터를 정리하여 데이터 품질 이슈로 등록한다.

구조 분석은 단일 테이블내 혹은 여러 테이블에 걸쳐 다중 컬럼 사이의 구조 특성을 도출하는 것이다. 도출된 구조 특성은 수집된 구조 관련 메타데이터와 비교하여 정확한 구조 특성들을 결정한다. 이후 컬럼 프로파일에 구조 특성들을 포함하여 데이터 프로파일을 생성한다. 또한 비교 결과로부터 수정된 메타데이터를 정리하여 데이터 품질 이슈로 정리한다.



(그림 5-1) 데이터 프로파일링의 절차

6 데이터셋 샘플링

6.1 개요

데이터 샘플링은 데이터 프로파일링의 목적에 따라 프로파일하고자 하는 대상의 범위를 선정하는 것이다. 대상 테이블과 컬럼을 선정하여 목록으로 작성한다. 이 목록은 데이터 품질 평가의 대상이 되고 지속적으로 관리해야 할 항목이다.

6.2 입력

데이터셋 샘플링의 입력은 데이터 품질 평가 대상이 되는 전체 데이터셋이다. 가능한 모든 데이터셋을 프로파일하는 것이 바람직하지만 데이터 프로파일링은 계산량이 크기 때문에 운영시스템에 부담으로 작용한다. 또한 효과적인 데이터 프로파일링을 위해 정규화된 형식으로 변환될 필요가 있다.

6.2 절차

데이터 샘플링을 통해 데이터 프로파일링 대상 데이터 샘플을 선정할 때 다음의 사항을 고려한다.
- 테이블의 건수
- 컬럼의 종류
- 컬럼의 레코드 건수
- 컬럼의 데이터 타입
- 컬럼의 필수 여부 및 기본값
- 컬럼의 도메인 등

이외에 선정된 데이터 샘플에 중복된 컬럼, 상이한 외부 뷰에 의한 재정의 및 반복된 로우가 포함되었는지 여부를 파악한다. 중복된 컬럼은 복수의 컬럼으로 분리되어야 한다. 상이한 외부 뷰에 의한 재정의는 그것이 동일 사실인지 혹은 다른 컬럼에 있는 값에 기초한 다른 사실을 대표하는지를 분석 과정을 통해 결정해야 한다. 반복된 로우는 동일 비즈니스 사실의 다중 발생을 대표하는지 별개의 사실을 대표하는지를 분석 과정을 통해 결정해야 한다.

6.3 출력

데이터셋 샘플링의 출력은 프로파일링 대상 데이터 샘플이다. 이 데이터 샘플에는 중복된 필드, 상이한 외부 뷰에 의한 재정의 혹은 반복된 로우가 제거되어 있다. 또한 데이터 샘플은 제 3 정규형이 바람직하지만, 경우에 따라 제 1 정규형 혹은 제 2 정규형 중 하나이며 어떤 상태인지 파악되어야 한다.

7 메타데이터 분석

7.1 개요

메타데이터 분석은 식별된 데이터셋과 관련된 문서화된 메타데이터를 수집하고, 운영중인 데이터셋의 메타데이터를 추출하여 이들 간의 불일치를 찾아내어 메타데이터를 더욱 정확하게 하는 과정이다. 이후의 프로파일링 절차에서 보다 정확하고 완벽한 메타데이터

를 제공하기 위한 것이다.

7.2 입력

메타데이터 분석의 입력은 문서화된 메타데이터와 데이터 프로파일링 대상 데이터셋 이다. 예를 들어 문서화된 메타데이터는 데이터 사전, 메타데이터 레포지토리, 테이블 정의서, 컬럼 정의서, ERD, 데이터 아키텍처 등이다. 프로파일링 대상 데이터셋은 현재 운영중인 데이터셋의 테이블과 컬럼의 스키마 정보를 추출하기 위한 것이다.

7.3 절차

메타데이터 분석은 메타데이터 수집 및 비교 과정으로 다음의 절차로 구성된다.

- 절차1: 운영 중인 데이터베이스에 접속하여 테이블 및 컬럼의 스키마를 추출하고 일반적인 컬럼 스키마 외에 관계 정의에 관한 정보를 추출한다.

- 절차2: 문서화된 데이터 명세를 수집한다.

- 절차3: 운영 중인 데이터셋의 메타데이터, 문서화된 데이터 명세 및 전사적인 메타데이터 표준과 비교하여 누락된 메타데이터와 불일치 유형을 파악한다(예, 테이블명 누락 혹은 불일치, 컬럼명 누락 혹은 불일치, 데이터 타입 불일치).

실제 운영중인 데이터셋의 메타데이터는 DBMS의 데이터 사전, 데이터 입력 프로그램의 인터페이스 정의, 데이터 관리 응용 소스 등에서 추출된다.

7.4 출력

메타데이터 분석의 출력은 메타데이터와 데이터 품질 이슈이다. 이 과정으로부터 출력된 메타데이터는 문서화된 메타데이터와 운영중인 데이터셋으로부터 출력된 메타데이터의 집합이며 상호 불일치하거나 표준 위반 사항은 정확하게 수정된 상태이다. 메타데이터의 불일치, 누락 및 표준 위반 사항은 데이터 품질 이슈로 등록된다.

이 단계의 메타데이터는 수정된 상태이지만 여전히 이후의 절차를 통해 수정될 메타데이터에 비하면 부정확하거나 불완전하다고 간주된다. 그럼에도 불구하고 이 메타데이터는 다음 절차에서 주요 비교 기준이 된다. 이 메타데이터는 데이터 프로파일링이 반복되면서 정확성과 완전성이 제고된다.

8 컬럼 특성 분석

8.1 개요

데이터베이스에서 가장 기본적인 요소가 컬럼이다. 이는 데이터베이스에 저장된 단일 비즈니스 객체에 대한 단일 사실을 수용하는 장소를 가리킨다. 컬럼 특성 분석은 메타데이터 분석 결과의 메타데이터와 실제 데이터 샘플의 컬럼에 수록된 값들로부터 도출된 컬럼 특성을 비교하여 보다 정확하고 완벽한 컬럼 특성을 결정하고 관련한 데이터 품질 이슈를 도출하는 것이다.

기본적으로 데이터베이스 시스템에서 데이터의 생성, 질의, 검색, 응용 개발 등을 위하여 컬럼을 정의한다. 또한 데이터 사전이나 메타데이터 저장소에도 보다 제한적인 관점에서 컬럼을 정의한다. 데이터 입력 폼 등의 매뉴얼에서 컬럼을 정의한다. 그러나 이러한 정의는 데이터셋에 대한 나름의 목적을 위한 것이므로 데이터셋의 품질 평가 측면에서 보면 부족하고 부정확할 가능성이 많다. 데이터 프로파일링은 데이터셋에 대한 기존의 컬럼 특성의 완벽성과 정확성에 대하여 비판적인 입장에서 출발한다.

8.2 입력

컬럼 특성 분석의 입력은 프로파일링 대상 데이터 샘플과 메타데이터이다. 데이터 샘플에 수록된 값으로부터 실제 컬럼의 특성이 도출된다. 메타데이터는 이 과정에서 도출되는 실제 컬럼 특성과 비교되는 기준이 된다.

8.3 절차

컬럼 특성 분석은 다음과 같은 세부 절차를 수행한다.

- 절차1: 입력된 메타데이터 중에서 프로그래밍 명세, DDL과 같은 관계데이터베이스 명세, 데이터 입력 스크린 명세, 데이터 입력 절차 매뉴얼, 데이터 사전 및 메타데이터 저장소로부터 문서화된 컬럼 특성을 정리한다.

- 절차2: 정리된 컬럼 특성과 무관하게 실제 데이터 샘플에 대한 컬럼 특성을 도출한다. 이때 프로파일링 도구와 비즈니스 전문가의 전문지식을 이용하여 반복적으로 수행한다. 컬럼 특성 도출의 방법은 발견(discovery), 규칙 시험(assertion testing) 및 시각적 점검(visual inspection) 등이며, 소프트웨어의 지원을 받는다.

- 절차3: 발견된 컬럼 특성과 문서화된 컬럼 특성을 비교한다. 양자간의 차이가 부정확한 데이터 혹은 부정확한 메타데이터에서 비롯되었는지 파악한다. 이 작업은 프로파일링 분석가, 비즈니스 분석가, 고유 분야 전문가, 데이터 설계자, DBA 및 응용 프로그램 개발자들 간의 협업으로 진행된다.

8.4 출력

8.4.1 분석 결과

컬럼 특성 분석의 출력은 각 컬럼에 대한 보다 정확하고 완벽한 컬럼 특성 리스트, 즉 컬럼 프로파일과 향후 데이터를 사용할 때 발생되는 데이터 품질 이슈이다.

컬럼 특성은 단일 컬럼에 허용되는 값들을 위한 규칙들을 가리킨다. 이들은 컬럼 프로파일로 정리된다. 컬럼 프로파일은 구조 분석과 데이터 규칙 도출에 기초가 된다.

컬럼 특성은 비즈니스 의미, 기수 특성, 저장 특성 및 유효값 특성 등으로 분류할 수 있다.

8.4.2 비즈니스 의미

비즈니스 의미에서 가장 대표적인 컬럼 특성은 컬럼명이다.

컬럼명

- 설명: 의미적으로 하나의 컬럼에 저장되어야 할 것을 정의, 컬럼의 내용을 정확하게 지시할 수 있는 서술적이고 대표성 있는 이름을 가져야 함
- 역할: 비즈니스 의미를 파악하는데 중요한 특성
- 사례: ORDER_NAME, ORDER_NUMBER, ORDER_DATE 등

8.4.3 기수 특성(Cardinalities)

기수 특성은 컬럼 전반적으로 값들이 어떻게 변화하는지를 나타낸다. 값들이 합리적으로 변화하는지 발견하는데 기초적인 역할을 한다.

로우의 개수

- 설명: 로우의 개수
- 역할: 컬럼 내의 어떤 값이 차지하는 비율을 구할 때 분모가 됨
- 사례: 30, 1000, 10000 등

값의 크기

- 설명: 최대값, 최소값, 중간값 및 평균값
- 역할: 컬럼 내의 값들의 통계적인 특성을 나타냄
- 사례: 최대값 (10000), 최소값 (-2000), 중간값 (300), 평균값 (200)

널(null)

- 설명: 널의 개수는 전체 로우에 대하여 값이 null인 로우의 개수이고 널의 %는 전

체 로우 개수 대비 널 값을 갖는 로우 개수의 차지 비율
- 역할: 컬럼 내 값들의 필수(mandatory), 선택(optional) 및 조건(conditional) 특성을 도출하는데 활용
- 사례: 널의 개수(3000), 널의 %(1%, 99%)

개별값의 개수
- 설명: 컬럼 내의 값들 중에서 개별 값들의 개수
- 역할: 컬럼의 도메인 발견에 도움
- 사례: 3(컬럼 값이 100, 100, 200, 200, 300일 경우) 등

유일성
- 설명: 컬럼 내의 값들 중에서 유일한 정도
- 역할: 컬럼의 키 발견에 도움
- 사례: 98%, 99%, 1% 등

8.4.4 저장 특성(Storage properties)

저장 특성은 컬럼에 값들의 외형을 지배하는 기본적인 규칙으로 일반적으로 시스템에서 강제적으로 관리되어 대개 근본적인 위반 방지책이 강구되나 실제 자주 위반되는 특성이다.

데이터 타입
- 설명: 하나의 컬럼에 저장될 수 있는 데이터의 형태를 정의함
- 역할: 컬럼내의 값들의 형태에 대한 제한으로 작용
- 사례: CHARACTER, INTEGER, DECIMAL, DATE, TIME, TIMESTAMP, BINARY, DOUBLEBYTE 등

값의 길이
- 설명: 숫자의 자리수 혹은 문자의 길이
- 역할: 컬럼내의 값들의 길이에 대한 제한으로 작용
- 사례: VARIABLE, FIXED 5, NUMERIC 5 등

소수점 자리
- 설명: 숫자 값들의 소수점 이하의 자리수
- 역할: 컬럼 내 숫자들의 정밀도를 나타냄
- 사례: DECIMAL 2, DECIMAL 3 등

8.4.5 유효값 특성(Valid value properties)

유효값 특성은 컬럼에 허용되는 값에 더욱 상세한 제한을 가한다. 이 특성은 정의가 상

세할수록 허용 가능한 값들의 범위가 제한된다.

이산 값 리스트
- 설명: 소규모 특정 값들의 리스트
- 역할: 컬럼내의 값들의 범위에 대한 제한으로 작용
- 사례: COLOR, STATE 등

값의 범위
- 설명: 값의 최소값과 최대값 사이의 범위
- 역할: 컬럼내의 값들의 range에 대한 제한으로 작용
- 사례: 100~200 포함 등

스킵 오버(skip-over) 규칙
- 설명: 일정 범위내에서 특정 값을 배제
- 역할: 컬럼내의 값들의 범위에 대한 제한으로 작용
- 사례: 날자 및 시간 컬럼에서 평일(weekdays), 휴일(holidays) 등

패턴
- 설명: 값들이 공통된 format을 갖음
- 역할: 컬럼내의 값들의 range에 대한 제한으로 작용
- 사례: 전화번호 (zzz-zzzz-zzzz), 주민등록번호, 이메일 등

도메인
- 설명: 컬럼이 갖을 수 있는 값들의 리스트
- 역할: 컬럼내의 값들의 range에 대한 제한으로 작용
- 사례: 성별 (M or F), 신용카드 등

9. 구조 분석

9.1 개요

구조 분석은 컬럼 특성 분석의 결과인 컬럼 프로파일을 기초로 테이블 내부 컬럼들 사이 혹은 다른 테이블의 컬럼과의 관계(relationships)를 도출하는 것이다.

구조 분석을 수행하는 이유는 컬럼의 값이 비록 개별 컬럼 상에서는 유효하나 데이터셋의 구조적인 문제, 즉 컬럼들의 비유효한 조합으로 인하여 컬럼의 값이 비유효해 질 수 있기 때문이다.

9.2 입력

구조 분석의 입력은 메타데이터와 컬럼 프로파일들이다. 메타데이터는 메타데이터 분석의 결과이고 컬럼 프로파일은 컬럼 특성 분석에서 도출되어 결정된 컬럼 특성들이다.

9.3 절차

구조 분석은 다음과 같이 메타데이터 중에서 문서화된 구조 정보 수집, 실제 데이터에 기반한 후보 구조 정보 도출, 수집 구조 정보와 도출 구조 정보의 비교, 정확한 구조 정보의 결정 등으로 구성된다.

- 절차1: 구조 정보 수집 과정은 다양한 방식으로 구조 정보를 획득될 수 있다. 메타데이터 분석의 결과 메타데이터 중에서 기존 데이터 구조 관련 문서, 메타데이터 저장소, ERD와 같은 데이터 모델, 데이터 정의 및 데이터 컬럼의 상식적인 점검 등으로부터 구조 정보를 수집한다.

- 절차2: 후보 구조 정보 도출의 첫째 방법은 데이터에 존재하는 모든 함수적 종속 특성과 동의 특성으로 보이는 모든 컬럼 페어를 찾아낸다. 둘째 방법은 가능성이 있는 구조 규칙을 데이터에 적용하여 적합도(degree of conformance)를 찾는 것이다. 첫째 방법은 계산량이 너무 크기 때문에 비현실적이고, 두 번째 방법은 구조 관계 누락의 가능성이 매우 크다. 따라서 첫째 방법을 채택하되 전문가의 경험을 통하여 구조 발견 전문 툴의 계산 범위를 한정하는 방식이 권고된다.

- 절차3: 수집 구조 정보와 도출 구조 정보의 비교를 통하여 구조 정보를 결정한다. 데이터 구조의 이슈는 데이터 구조와 관련된 기본적인 비즈니스 정의로 귀결된다. 테이블들은 실세계 객체와 상관관계가 있고, 테이블 간의 관계는 비즈니즈 지식이 필요하다. 구조 이슈는 의미적 해석이 반드시 필요하므로 비즈니스 전문가와 프로파일 전문가의 협업을 통하여 정확한 구조 정보를 결정한다.

9.4 출력

9.4.1 분석 결과

구조 분석의 출력은 프로파일링 대상 데이터셋의 구조 특성과 구조 정보의 부정확성으로 인해 향후 데이터를 사용할 때 발생될 수 있는 데이터 품질 이슈이다.

데이터셋의 구조 특성은 데이터 프로파일에 정리된다. 구조 특성은 종속 특성과 동의 특성으로 분류할 수 있다.

9.4.2 종속 특성(dependency properties)

종속 특성은 컬럼들간의 종속 관계를 설명하는 구조 특성이다. 종속 특성에는 일차키, 외래키, 함수적 종속 특성 및 유도 컬럼으로 구분된다.

일차키(primary key)

- 설명: 컬럼의 각 로우를 유일하게 정의하는 컬럼의 집합
- 역할: 테이블의 각 로우를 식별
- 사례: SOCIAL-SECURITY_NUMBER, PERSON_ID 등

외래키(foreign key)

- 설명: 부모 테이블에서 로우를 식별하는 종속 테이블의 하나 혹은 그 이상의 컬럼
- 역할: 두 테이블간의 부모/종속 관계를 부여
- 사례: 부모 테이블 PERSONNEL의 primary key PERSON_ID를 참조하는 종속 테이블 DEPARTMENT의 컬럼 DEPT_MANAGER_ID

함수적 종속 특성(functional dependency)

- 설명: 테이블의 한 컬럼이 동일 테이블의 다른 컬럼들의 특정 값 집합에 대한 유일한 값을 가질 경우 그 컬럼들의 관계
- 역할: 테이블 내에 컬럼들 사이의 기능적 종속 관계를 나타냄
- 사례: EMPLOYER_ID (좌측 컬럼) -> EMPLOYER_NAME (우측 컬럼)

유도 컬럼(derived column)

- 설명: 기능적 종속 관계의 LHS의 값들에 대한 하나의 규칙에 의해 값이 결정되는 RHS 테이블의 컬럼
- 역할: 테이블 내에 컬럼들 사이의 유도 관계를 나타냄
- 사례: COLUMN_A + COLUMN_B (좌측 컬럼)이 COLUMN_C (우측 컬럼)과 동일하면, COLUMN_C는 유도 컬럼이라고 함

9.4.3 동의 특성(synonyms)

두 개의 컬럼이 동일한 비즈니스 사실을 포함하고 있으면, 두 개의 컬럼은 서로 동의 구조를 갖고 있다. 동의 특성은 일차키/외래키 동의 특성, 잉여 데이터 동의 특성 및 도메인 동의 특성으로 구분된다.

일차키/외래키 동의 특성(primary/foreign key synonym)

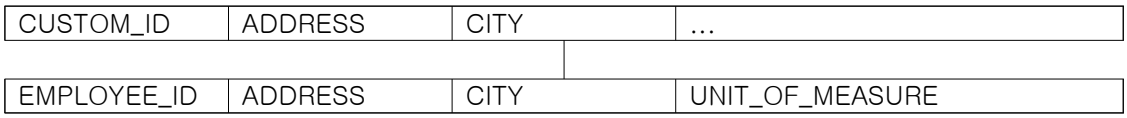
- 설명: 한 컬럼이 일차키이고 다른 컬럼이 외래키일 경우, 두 컬럼은 일차키/외래키 동의 특성 관계에 있음
- 역할: 두 테이블간의 부모/종속 관계를 부여
- 사례: 부모 테이블 PERSONNEL의 일차키 PERSON_ID와 이를 참조하는 종속 테이블 DEPARTMENT의 컬럼 DEPT_MANAGER_ID

잉여 데이터 동의 특성(redundant data synonym)

- 설명: 하나의 컬럼이 다른 테이블의 컬럼과 동의 관계에 있고, 그 컬럼이 키에 종속적이며 다른 테이블의 해당 외래 키에 종속적이면 그 두 컬럼은 잉여 데이터 동의 특성 관계에 있음
- 역할: 두 컬럼중 하나는 정보의 손실 없이 삭제될 수 있으나, 삭제할 경우 종속 테이블에 대한 질의문에 JOIN 명령어를 추가해야 함
- 사례:

도메인 동의 특성(domain synonym)

- 설명: 한 컬럼의 도메인과 다른 컬럼의 도메인이 동일할 경우 두 컬럼은 도메인 동의 관계에 있음
- 역할: 구조적인 관계는 없고 동일한 비즈니스 사실을 갖고 있음. 특정 비즈니스 엔티티를 식별하는 도메인들의 일관된 동의 관계는 질의어에서 데이터를 연결하는데 도움이 됨
- 사례:



부 록 Ⅰ-1

지식재산권 확약서 정보

해당 사항 없음

부 록 Ⅰ-2

시험인증 관련 사항

해당 사항 없음

부 록 Ⅰ-3

본 표준의 연계(family) 표준

해당 사항 없음

부 록 Ⅰ-4

참고 문헌

- 해당사항 없음

부 록 Ⅰ-5

영문표준 해설서

해당 사항 없음

부 록 Ⅰ-6

표준의 이력

판수	채택일	표준번호	내용	담당 위원회
제1판	2019.xx.xx	제정 TTAK.KO-xx.xxxx	-	메타데이터 PG(PG606)