情報保護學會論文誌
第 22 卷 第 3 號, 2012. 6

데이터베이스 시스템에서 디지털 포렌식 조사를 위한 체계적인 데이터 추출 기법 연구*

이 동 찬,[1†] 이 상 진[2‡]
[1]한국국방연구원, [2]고려대학교

Research of organized data extraction method for digital investigation in relational database system*

Dongchan Lee,[1†] Sangjin Lee[2‡]
[1]Korea Institute for Defense Analyses, [2]Korea University

요 약

기업의 탈법, 비리 등 부정행위를 조사할 경우 인사, 회계, 물류, 생산 등의 업무데이터(Business Data)의 확보가 필요하다. 다수의 기업들은 분산된 업무 데이터를 데이터베이스(Database)화하여 통합적으로 관리하고 있기 때문에 디지털 포렌식 조사를 위하여 데이터베이스에 대한 체계적인 업무데이터 추출기법 연구가 중요하다. 일반적인 정보체계 환경에서 데이터베이스는 상위 어플리케이션 및 대용량 파일 서버와 통합된 정보체계 내의 부분적 형태로 존재한다. 또한 사용자가 입력한 원시 업무 데이터는 정규화 과정을 거친 테이블 설계에 의해 하나 이상의 테이블에 분산되어 저장된다.

기존 데이터베이스 구조 분석에 관한 연구들은 데이터베이스의 최적화와 시각화를 위하여 테이블 간 연관관계 분석이 가장 중요한 연구대상이었다. 그러나 원시 업무데이터를 획득해야 하는 디지털 포렌식 관점의 연구는 테이블 간 연관관계 시각화보다 데이터의 해석이 더 중요한 연구대상이다.

본 논문에서는 데이터베이스 내부에서 미리 정의된 테이블 간 연관관계 분석기술뿐만 아니라 도메인 전문 지식(domain knowledge)을 활용한 체계화된 분석절차를 제시하여 데이터베이스에 저장된 원시 업무 데이터 구조를 분석하고 사건관련 데이터를 추출할 수 있는 분석방안을 제안한다.

ABSTRACT

To investigate the business corruption, the obtainments of the business data such as personnel, manufacture, accounting and distribution etc., is absolutely necessary. Futhermore, the investigator should have the systematic extraction solution from the business data of the enterprise database, because most company manage each business data through the distributed database system, In the general business environment, the database exists in the system with upper layer application and big size file server. Besides, original resource data which input by user are distributed and stored in one or more table following the normalized rule.

The earlier researches of the database structure analysis mainly handled the table relation for database's optimization and visualization. But, in the point of the digital forensic, the data, itself analysis is more important than the table relation.

This paper suggests the extraction technique from the table relation which already defined in the database. Moreover, by the systematic analysis process based on the domain knowledge, analyzes the original business data structure stored in the database and proposes the solution to extract table which is related incident.

Keywords: Database Forensics, Database Reverse Engineering

접수일(2011년 12월 12일), 수정일(1차: 2012년 2월 14일, 2차: 2012년 4월 16일), 게재확정일(2012년 6월 8일)

* 본 연구는 한국연구재단을 통해 교육과학기술부의 바이오 연구개발사업으로부터 지원받아 수행되었습니다. (2011-0027732)

† 주저자, ripil83@kida.re.kr
‡ 교신저자, sangjin@korea.ac.kr

I. 서 론

다수의 기업들은 업무데이터를 관리하기 위하여 정보체계를 사용하고 있다[1]. 전사적 자원관리 시스템(Enterprise Resource Planning System, ERP)과 고객관계관리 시스템(Customer Relationship Management System, CRM) 등과 같은 정보체계에서는 대용량의 자료를 저장하는 데이터베이스가 포함되어 있으며, 기업의 모든 업무 데이터가 데이터베이스에 저장되기 때문에 기업의 부정과 관련된 디지털 포렌식 조사에서 데이터베이스는 중요한 조사 대상이 된다. 전사적 자원관리 시스템 및 고객관계관리 시스템들은 데이터베이스 상위 계층에 어플리케이션이 존재하여 사용자를 대신하여 데이터베이스에 접근하고 데이터를 관리한다. 따라서 디지털 포렌식 조사자는 기업 부정행위 조사에 데이터베이스 시스템뿐만 아니라 상위 계층의 구성도 고려하여 정보체계 전반에 대한 분석을 수행해야 한다.

사용자가 작성하는 원시데이터와 데이터베이스의 데이터 사이에는 이질성이 존재한다. 이는 데이터베이스 설계에 따라 원시데이터가 하나 이상의 테이블에 분할되어 저장되고, 개발자의 개발논리에 의해 원시데이터와는 다른 형태의 자료로 저장될 수 있기 때문이다. 따라서 데이터베이스에서 획득된 데이터는 반드시 해석 과정을 거쳐야 한다.

기존의 연구에서는 데이터베이스의 검색 속도 최적화 혹은 테이블 간 관계의 표현 관점에서 데이터베이스 내의 시스템 카탈로그를 분석하거나 테이블 생성 형태를 분석하여 테이블 연관관계를 획득하고 이를 객체화하여 표현하였다[2][3]. 원시 업무데이터는 어플리케이션으로부터 데이터베이스 테이블에 분할되어 저장되기 때문에 단순히 기존 연구들이 수행해 왔던 데이터베이스 내부의 테이블 연관관계 분석 정보만으로는 완벽한 원시 업무데이터를 추출할 수 없다.따라서 본 논문에서는 위에서 제시한 문제점들을 극복하고 정확한 조사 업무데이터의 획득을 위하여 원시 업무데이터 구조 분석 방안과 대용량 객체 데이터 추출 방안을 제안한다. 또한 제안한 방안을 토대로 데이터베이스 조사 도구를 구현하여 가상의 조사환경에서 조사도구 통하여 원시 업무데이터 구조를 분석하고 조사용 업무데이터를 추출한다.

II. 관련연구

기존 연구들은 데이터베이스 최적화 관점에서 설계상의 문제점은 없는지 확인하거나, 테이블 간 연관관계를 ERD(Entity Relation Diagram)로 표현하는 것을 목적으로 하였다. 기존 연구로는 확장된 엔티티 다이어그램의 표현에 관한 연구[2], 테이블 기반의 레거시 데이터베이스에서 엔티티 간 관계의 추출에 관한 연구가 발표되었다[3]. 설계논리에 대한 역 분석 측면에서 데이터베이스 역공학 DBRE(Database Reverse Engineering)이라 칭하고 있다[4].

기존 연구들의 공통적인 과정은 데이터 구조 추출처리(The data structure extraction process)와 데이터 구조 개념화(The data structure conceptualization process)로 나눌 수 있다.

데이터 구조 추출은 세 과정으로 나누어진다. 첫 번째 과정은 속성 추출(Attribute extraction)로써 데이터베이스 테이블에 정의된 각 필드명은 실제 업무에서의 의미와 차이가 있을 수 있기 때문에 데이터베이스 테이블의 필드에 대해서 실제 데이터의 의미를 연결하기 위한 과정이다. 두 번째는 키 추출(Key extraction) 과정으로 데이터베이스 테이블에 대한 기본 키와 외래 키를 획득한 후 테이블 간의 상하 관계를 확인함으로써 테이블 간 연관관계를 확인한다. 세 번째는 제약사항 추출(Constraint extraction)로써 기본 키와 외래 키 사이의 카디널리티를 획득하는 것으로써 관계를 가진 두 테이블의 대응 관계를 확인하여 두 테이블 사이의 관계에 대한 유효성을 검사한다. 이러한 데이터 구조 추출로써 테이블 연관관계를 확인할 수 있다.

데이터 구조 개념화는 데이터 구조 추출에서 나타난 수직, 수평 관계를 ERD(Entity Relation Diagram)로 표현한다. 엔티티(Entity)는 데이터베이스의 개념적인 용어로써 하나 혹은 그 이상의 테이블로 구성되거나, 여러 엔티티가 하나의 테이블로 구성될 수도 있다. 때문에 테이블을 분리, 통합하여 엔티티를 표현한다[4].

하지만 이와 같은 연구들은 대상 범위가 데이터베이스 내부에 국한되며, 분석 관점이 테이블 간 연관관계 유무와 연관되는 대응 관계의 표현에 맞춰졌기 때문에 디지털 포렌식 관점에서 필요한 원시 데이터 구조 분석에는 한계가 존재하였다. 그러므로 정확한 원시 데이터의 구조분석 및 데이터 획득을 위한 추가적인 기술연구가 필요하다.

III. 데이터베이스 조사 환경 및 메타정보

사용자가 입력한 원시 업무데이터가 정보체계 내에서 어떻게 저장되어 있는지 분석하기 위하여 각 단위 시스템별 설계를 역으로 추적해야 한다. 특히 주요 데이터가 저장된 데이터베이스의 설계 및 개발논리의 역추적이 필요하다. 대용량 데이터를 관리하는 정보체계에서는 필수적으로 시스템의 성능을 위하여 가용 자원을 최적화하여 운영한다. 대표적으로 데이터베이스에서 자료의 입출력 시 최적화를 위하여 정해진 키 설정 및 인덱스에 맞게 명령을 입력한다. 최적화되지 않은 명령은 시스템의 부하를 초래하기 때문이다. 이점은 조사자가 데이터베이스에 명령을 내릴 경우 항상 유의해야 할 사항이다.

최근에는 데이터베이스 보안장비의 도입으로 인해 스키마 정보가 은닉되고, 테이블 및 컬럼의 네이밍(Naming)이 별도로 관리되는 등 데이터베이스 분석이 어려워지고 있다. 하지만 합법적인 조사에 대한 필요한 정보의 수집은 가능하기 때문에 데이터 메타정보들의 식별이 필요하다. 기술적 메타정보는 데이터베이스 내부, 외부 스키마 정보를 의미하고 정책적 메타정보는 정보체계 설계표준, 산출물 등의 운영 및 관리상에 필요한 정책적인 정보들이다.

어플리케이션을 통한 분석은 모니터링과 역 공학으로 정보체계 데이터 설계논리를 파악하는 것이 가능하나 기술적, 환경적, 재사용성 측면에서 문제점이 존재한다.

3.1 기술적 메타정보

대용량 DBMS는 일정한 형식의 스키마 데이터가 존재한다. 테이블의 메타정보와 컬럼의 메타정보들을 포함하고 있다. 〔표 1〕은 DBMS별로 메타정보를 확인할 수 있는 시스템 카탈로그 뷰 이다. 테이블연관관계가 완벽하게 스키마에 저장되는 설계로 데이터베이스를 구축하였을 경우 표 1의 정보를 활용하여 테이블의 연관관계를 유추해낼 수 있다. 하지만 스키마에 저장되는 설계는 변경이 어렵다. 때문에 정보체계의 특성상 업무환경변화에 따른 잦은 성능개선, 시스템 부하로 인한 정규화, 유지보수 등의 작업으로 스키마에 모든 연관관계가 저장되어 있지는 않다. 시스템의 설계자와 관리자, 성능개선 및 유지보수 개발자 등이 모두 완벽한 설계를 해야만 모든 연관관계가 스키마에 저장되게 되므로 실제 대용량 데이터베이스 환경에서

〔표 1〕 DBMS 별 시스템 카탈로그 뷰

DBMS의 종류	Oracle	IBM DB/2	MySQL
컬럼 특성정보	dba_tab_cols	syscat.columns	information_schema.columns
제약사항 정보	user_constraints	check_constraints	information_schema.key_column_usage
	all_constraints	refferential_constraints	information_schema.statstics
	dba_constraints	table_constraints	information_schema.table_constraints
테이블 및 뷰 특성 정보	dba_tables	syscat.tables	information_schema.tables
	dba_views	syscat.views	information_schema.views

스키마만으로 연관관계로 모든 연관관계를 분석 하는 것은 어렵다.

3.2 정책적 메타정보

정보체계 개발하는 다수의 업체가 국제 품질표준을 준수하여 품질관리활동을 하고 있다〔7〕〔8〕. 또한 정보시스템감리기준에 의한 정보시스템감리 체계로 인하여 개발과 관련된 감리를 받고 있다〔9〕. 이 같은 활동으로 인해 데이터베이스의 조사와 관련된 정책적 메타정보는 내용과 신뢰도가 증가되었다.

품질 보증활동과 감리활동의 결과물은 데이터베이스 조사를 위한 정책적 메타정보로 활용성이 높다. 개발 및 품질보증 활동으로 발생된 산출 문서는 각 사업별로 테일러링(tailoring)되기 때문에 공통된 정책적 메타정보를 정의할 수는 없다. 다만 정책적 메타정보로 활용 가능한 항목의 예는 다음과 같다.

유스케이스 모델	클래스 모델	실체관계도 (ERD)	엔티티 정의서
컴포넌트 정의서	UI 흐름도	테이블 목록	테이블 정의서
식별자 목록	속성자료 목록	기술표준 정의서	시험 계획서
시험 절차서	감리 보고서		

IV. 제안하는 원시데이터 구조분석 및 추출

대용량의 업무데이터를 데이터베이스에 저장하기

위하여 설계과정에서 정규화를 거친다. 하지만 정규화가 이루어진다 할지라도 기본적인 서비스 가용성을 확보하기 위하여 테이블 분할 시 이점을 고려할 수밖에 없다. 만약, 하나의 업무데이터가 두개의 테이블로 분할될 때 두 테이블 역시 대량의 데이터일 경우 서로간의 연관관계 없이 데이터가 저장된다면 서비스 타임아웃이 발생하며 이는 서비스 가용성을 확보할 수 없는 결과를 가져오게 된다. 업무데이터를 추출하는 과정에서 주 테이블과 보조 테이블의 연관관계가 존재하지 않는 보조테이블의 데이터는 소량일 수밖에 없고, 제원성의 자료만 가능하다. 이와 같은 데이터는 업무지식 데이터 수준에서 추출데이터를 변환하는 것이 가능하다. 따라서 주요 업무데이터는 주 테이블과 연관관계를 맺고 있는 보조테이블들로 저장되므로, 해당 연관관계를 분석하는 것이 중요하다.

전 테이블을 대상으로 데이터를 찾는 방식은 테이블의 수나 데이터의 양이 적은 경우에는 가능하지만, 대량의 업무데이터를 저장하는 데이터베이스에서 모든 테이블의 자료를 검색하는 것은 서비스 부하 및 락(Rock)을 발생시킬 수 있다. 따라서 제안하는 원시데이터 구조분석 및 추출에서는 주 테이블의 선정과 이와 관련된 연관관계를 중심으로 분석하는 절차를 제안한다.

4.1 사전준비단계

기술적 메타정보와 정책적 메타정보를 획득하여 연관관계 구조분석에 활용할 수 있도록 한다. 스키마를 통한 기본적인 테이블, 컬럼 정보와 산출물 분석에 의한 추가정보, 그리고 샘플링 데이터를 구축한다. 특성화 정보는 다음과 같다.

테이블스키마	테이블 네임	컬럼 네임
널 허용여부	데이터 타입	문자열 최대길이
숫자 범위	도메인 지식	테이블인덱스

샘플링데이터는 “반드시 있어야 하는 데이터”와 “있을만한 데이터”로 구분하여 분석에 활용하도록 한다. “있을만한 데이터”는 참조 연관관계 분석에서 활용하고, “반드시 있어야 하는 데이터”는 연관관계 검증에서 활용한다. 이는 샘플링 데이터가 잘못 설정되어 있을 경우 다시 연관관계 분석절차를 거치는 것이 시간상으로 비효율적이기 때문이다. 참조하고자 하는 테이블 인덱스를 활용하여 참조조건을 설정하고, 인덱스 정보를 자료화하여 최적화된 명령을 사용하도록 한다.

본 논문에서 도메인 지식이란 예를 들어 경제용어, 법률용어, 의학용어 등 각 업무별 데이터 표준자료와 과학 공식, 수학 공식, 회계 산법, 알고리즘 등의 지식정보를 의미한다. 도메인 지식정보의 경우 정책적 메타정보와 해당 업무분야 전문가의 지식을 바탕으로 구축한다. 도메인 지식정보는 프로젝션 및 검증 과정에서 활용한다.

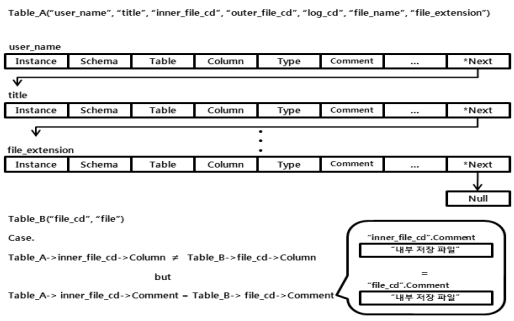
4.2 특성정보 객체화 및 참조 연관관계 분석

분석 대상의 식별을 위하여 추출 정보의 종류, 자료의 활용빈도, 도메인에 따라 주 테이블을 선정해야 한다. 특히 자료의 활용빈도를 통한 주 테이블선정은 데이터베이스 워크로드 식별 및 테이블별 데이터의 양과 테이블 구성형식을 조사함으로써 주 테이블 선정이 가능하다.

선정된 주 테이블의 특성 정보를 객체화 한 후, 주 테이블에 대한 모든 참조 테이블을 확인한다. 이때 두 가지 검색방법이 존재한다.

첫 번째는 기존의 DBRE 기술과 동일하게 시스템 카탈로그 내에 주 테이블에 대한 참조테이블 관계정보가 존재할 경우 이를 바탕으로 참조테이블을 검색하는 방법이다. 데이터베이스 설계 및 관리상 키, 인덱스 설정정보 및 동일한 컬럼명을 활용하여 주 테이블의 각 컬럼별로 검색을 수행하여 참조 테이블을 검색한다.

두 번째는 객체화한 주 테이블의 각 컬럼 별 특성정보와 동일한 컬럼을 검색하고 검색된 컬럼을 포함하는 테이블을 검색하는 방법이다. 이 경우 테이블 목록은 첫 번째 방법에 비해 확실한 참조테이블 목록은 아니지만 어플리케이션에 의한 테이블 간 연관관계가 가능한 참조테이블 목록이며, 원시 데이터에 대한 지식을 소유한 조사자에 의해 연관관계를 확인할 수 있는 목록이 된다. 〔그림1〕은 주 테이블인 Table_A 테이블을 각 컬럼 별로 특성정보를 객체화 하고 참조 테이블 검색에 어떻게 활용하는지 보여주는 예이다. Table_A의 inner_file_cd 컬럼은 Table_B의 file_cd 컬럼과 키에 의한 연관관계가 존재하지 않으며 컬럼명 또한 동일하지 않다. 하지만 실제 어플리케이션에서는 ″inner_file_cd = file_cd″ 와 같은 조건을 두어 테이블에 연관관계를 맺고 있다. 먼저 첫 번째 검색 방법인 시스템 카탈로그를 활용한 검색에서는 키에 의한 연관관계 및 컬럼명을 기준으로 참조 테이블 검색을 수행할 경우 Table_B 테이블을 검색하지



〔그림 1〕 특성정보 객체화 및 검색

못한다. 반면 두 번째 방법인 특성정보 "Table_A→inner_file_cd→comment"와 동일한 값을 갖는 테이블 및 컬럼을 검색할 경우 "Table_B→file_cd→commnet"가 동일함을 확인하여 참조 테이블을 검색할 수 있다. 이와 같이 키 및 컬럼명에 의한 연관관계 검색에 참조테이블이 도출되지 않을 경우 객체화한 특성정보를 활용하여 참조테이블을 검색할 수 있다.

4.3 주 테이블과 참조테이블의 연관관계 검증

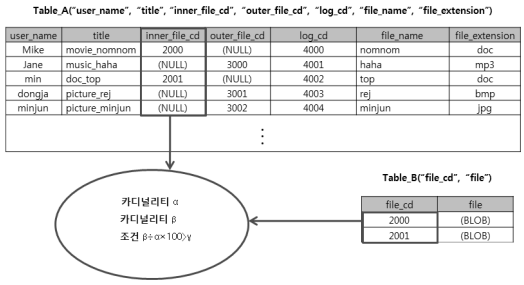
참조 테이블 검색방법을 통해 주 테이블의 각 컬럼별로 참조테이블 리스트를 구할 수 있다. 하지만 전사적 자원관리 시스템이나 고객관계관리 시스템에 포함된 데이터베이스는 기업의 업무 데이터를 총괄하기 때문에 다수의 테이블과 컬럼들로 이루어져 있다. 그러므로 조사자가 조사 데이터의 항목을 선택하기 전 참조 테이블에 대한 검증을 수행하여 불필요하게 확인해야 할 참조테이블의 수를 줄여야 한다. 검증 방법에는 카디널리티 조사를 통한 검증방법과 조사자 지식을 통한 검증방법으로 나눌 수 있다.

카디널리티 조사를 통한 검증방법은 〔그림 2〕와 같이 주 테이블인 Table_A의 컬럼 inner_file_cd와 참조테이블 Table_B의 컬럼 file_cd의 데이터를 비교하여 검증한다. 이때 주 테이블인 Table_A의 inner_file_cd의 유일성을 가지는 데이터가 반드시 하나 이상이어야 하며 NULL이 아니어야 한다. 또한 유일성을 만족하는 데이터 튜플(Tuple)의 카디널리티(cardinality)가 클수록 검증에 유리하다. 검증방법은 다음과 같다.

- inner_file_cd의 유일성을 가지는 데이터 튜플의 카디널리티(α)와 데이터를 조사한다.
- 앞에서 구한 Table_A->inner_file_cd의 데이터와 동일한 Table_B->file_cd의 데이터 튜플의 카디널리티(β)를 조사한다.
- 첫 번째에서 구한 카디널리티(α)와 두 번째에서 구한 카디널리티(β)의 값이 동일한지 확인하여 검증한다.

하지만 실제 연관관계가 존재하는 주 테이블과 참조 테이블에서 일치하지 않는 데이터가 일부 존재할 경우 검증에 실패하게 된다. 그러므로 카디널리티의 동일여부를 확인한 결과, 참조테이블을 도출하지 못할 경우 혹은 조사자가 원하는 조사데이터 항목이 도출되지 못할 경우 동일 카디널리티의 최소 퍼센트(ɣ)를 입력 받아 참조 테이블 구한다(β÷α×100〉 ɣ).

카디널리티를 활용한 검증 이외에 샘플링 데이터의 "반드시 있어야 할 데이터"로 해당 참조관계를 검증하도록 한다. 이후 두 연관관계의 1 대 N, N 대 N 관계를 검증한다.



〔그림 2〕 테이블 간 연관관계 검증 예

4.4 데이터 정합성 판단 및 세부 선별조건 설정

데이터 정합성을 판단하는 과정은 주 테이블과 참조테이블의 조인 및 원시 데이터 항목에 맞는 프로젝션의 수행을 통하여 획득할 수 있는 데이터가 원시 데이터와 모순되는 점은 없는지 판단하는 것을 의미한다. 주 테이블과 참조테이블의 조인 전·후 카디널리티를 비교함으로써 만약 카디널리티가 변경되지 않아야 할 조인일 경우, 카디널리티가 조인 후 변경된다면 정합성이 훼손되었다고 판단하여 참조 테이블과의 조인을 취소한다. 뿐만 아니라 조사데이터의 항목이 완벽하게 주 테이블에 참조되지 못했을 경우, 지속적으로 참조 테이블의 조인과 프로젝션 과정을 반복 수행하여 조사 데이터를 추출한다. 또한 도메인 지식정보를 활용하여 프로젝션 과정까지 끝난 데이터를 검증하도록 한다.

조인 및 프로젝션 수행 이후에 컬럼의 각 항목 별 조건 설정에 의해 조사데이터의 선별적인 획득이 가능

하다. 각 항목 별 데이터 선별 조건 설정은 특정 문자, 숫자, 날짜 등 특정 조건에 해당하는 데이터만을 선별하여 조사할 경우 문자비교, 숫자 대소비교, 날짜 및 시간 범위 설정 등을 통하여 조건에 맞는 데이터만을 선별하여 추출한다.

4.5 데이터 정합성 판단 및 세부 선별조건 설정

조사자가 대상 테이터베이스 연관관계를 분석하는 도중에도 서비스가 계속적으로 제공되기 때문에 데이터 테이블스페이스의 무결성을 확보하기 어렵다. 서비스 가용성 및 조사자료 신뢰성 확보 방안을 위하여 다음과 같은 조치사항을 제안한다.

조사자의 행위로 인하여 데이터 테이블스페이스(Data Tablespace)가 변경되지 않았다는 것을 증명하기 위하여 조사행위에 대한 로그를 기록한다. 딕셔너리 테이블스페이스(Dictionary Tablespace)와 같은 메타정보들이 더 이상 변경될 수 없도록 전체 사용자 권한을 분석기간 내에 일시적으로 제거하여 수정을 제어한다. 언두 테이블스페이스의 용량이 부족할 경우 서비스의 가용성을 해칠 염려가 있기 때문에 가용한 용량을 모니터링 하고, 분석행위에 의한 락(Rock)의 발생여부를 체크하여 조치한다. 조사의 분석 방식, 추출방식 등을 피 조사자에게 확인 받고, 이후 추출된 조사 데이터, 딕셔너리 테이블스페이스 데이터, 조사행위 로그, 모니터링 데이터 등의 해쉬 값을 기록하여 피 조사자로 하여금 확인 및 서명하도록 한다.

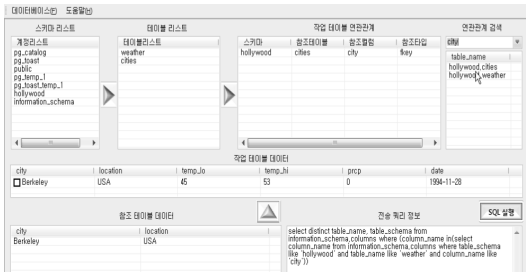
시스템 가용성이 확보된 상태에서 분석 및 추출 방식에 대한 동의, 조사행위에 대한 기록, 딕셔너리 테이블스페이스에 대한 무결성 확보 등으로 분석 및 추출과정 중에 데이터 테이블스페이스 영역에 수정을 가하지 않았음을 증명하여 추출한 조사데이터의 신뢰성을 확보하는 것이다.

V. 도구구현 및 실험

5.1 도구구현 목적

4장의 방안을 토대로 도구를 〔그림 3〕과 같이 구현하였다. 디지털 포렌식 측면에서 조사도구는 목적은 다음과 같다.

○ 분석 및 추출 프로세스를 도구에 적용하여 조사의 신속성과 편의성 증대

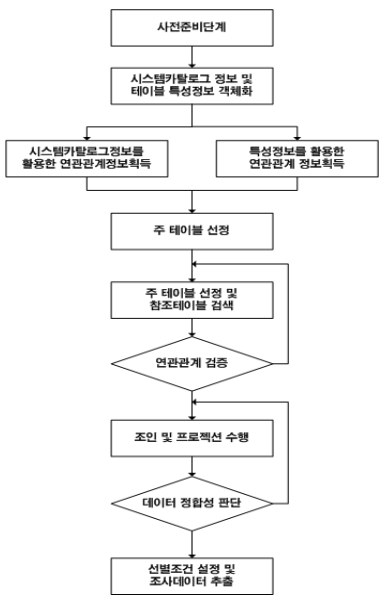


〔그림 3〕 데이터베이스 구조분석 및 추출 도구화면

- 조사자에 의한 임의적인 행위제어
- 조사행위에 대한 로그를 기록
- 연관관계 분석 중 서비스 가용성을 침해할 수 있는 조회를 제어
- 샘플링 데이터를 통한 데이터 정합성 자동 검증

5.2 조사 진행 절차

조사자가 수행하는 전반의 프로세스 과정은 다음 〔그림 4〕와 같다. 조사자는 데이터베이스 모든 테이블에 접속할 수 있어야 하므로 시스템 관리자 계정 권한으로 접속해야 하며, 테이블스페이스와 주 테이블을 선택해야 한다. 목표로 하는 원시데이터의 입력 및 수정 빈도, 기록성 자료여부, 데이터의 양 등에 따라 워크로드와 테이블 메타정보 등을 통하여 테이블스페이스와 주 테이블을 결정할 수 있다.



〔그림 4〕 조사 진행 순서도

원시데이터 분석 도구는 원시데이터에 대한 배경지식이 있는 조사자에 의해 수행되는데, 절차에 따른 분석을 위해 분석 중 원시 데이터 배경 지식에 따른 선택사항과 정보입력이 필요하다. 사용자 선택과 정보입력에 의한 프로세스는 크게 세 가지로 분류할 수 있다.

첫째, 분석 도구에서 데이터베이스로 접속이 성공하면 목표로 하는 원시데이터의 구조 분석을 위하여 목표 테이블스페이스, 주 테이블 이름을 선택한다. 데이터베이스는 테이블스페이스 별로 접근 가능 테이블이 한정되기 때문에 목표로 하는 원시데이터의 획득을 위하여 주 테이블을 선택하는 과정에서 테이블스페이스, 주 테이블 이름 선택은 반드시 필요한 절차이다.

둘째, 주 테이블에서 각 컬럼별 참조 테이블 목록을 조사한다. 4장의 원시데이터 구조분석 절차와 같이 미리 객체화된 특성정보에서 찾을 수 있는 참조테이블은 자동으로 검색되며, 추가적으로 도메인 지식 및 샘플링 데이터를 통하여 검색한다. 이때 조사자는 어떠한 특성정보, 도메인 지식, 샘플링 데이터로 연관관계 테이블을 검색할 것인지 선택하도록 한다.

셋째, 검색된 참조테이블 리스트에서 참조테이블을 선택하여 샘플링 데이터를 열람한다. 참조테이블의 데이터 중에서 원시 데이터 항목에 포함되어야 하는 데이터가 존재한다면 참조테이블의 조인 조건 선정과 조인 대상 선정 과정이 필요하다. 조인 조건 선택은 주 테이블에 참조 테이블을 조인하기 위한 조건 C에 참여하는 컬럼을 참조테이블에서 선택하는 것이고, 조인 대상은 원시 데이터에 포함되는 참조 테이블의 컬럼을 선택하는 것이다. 이때 조사도구는 테이블 인덱스 조건, 파티셔닝 등을 자동으로 적용하여 최적화된 조인으로 데이터베이스 서비스 가용성을 해치지 않도록 한다. 세 번째 과정은 원시데이터의 항목을 모두 포함할 때까지 계속해서 수행한다.

5.3 실험

조사도구 실험을 위하여 대용량 데이터가 저장된 주 테이블과 연관관계 테이블 샘플을 테스트 서버에 구축하였다. 테이블 연관관계는 아래의 표 2의 A~E 조건과 같다. 대용량 데이터베이스는 데이터 조회 서비스 가용성을 확보하기 위하여 키, 인덱스를 필수적으로 구성해야만 하기 때문에 이외의 조건들은 실험대상에서 제외하였다. 연관관계 검색 조건 당 타임아웃은 30초 이며, 메모리 사용률 CPU 사용률이 동일한 조건에서 조사도구를 테스트하기 위하여 사용자 접근을 차단하였다. 데이터 모델링 혹은 데이터베이스 디자인 소프트웨어에서 레거시 데이터베이스의 연관관계를 분석하는 DBRE 기능과 본 논문에서 제시한 조사도구를 비교하였다. 모델링 도구는 연관관계 분석 외에도 엔티티 다이어그램으로 표현하는 시간이 추가적으로 소요되기 때문에 시간비교는 제외하였다. 모델링 도구의 경우 데이터베이스 내 모든 테이블에 대한 연관관계 분석을 수행하지만 본 논문의 조사도구는 목표로 하는 연관관계만을 선별적으로 분석한다. 따라서 실험대상 데이터베이스의 연관관계 범위는 주 테이블과의 연관관계만을 대상으로 하였다. 모델링 도구 'E'는 ER-WIN 4.1의 Database Engineer 기능이고, 데이터베이스 설계도구 'M'은 Microsoft Visio 2007의 Database Engineer 기능이다.

〔표 2〕 모델링 및 설계 도구와의 조건 별 비교

	A	B	C	D	E
모델링 도구 'E'	○	○	×	×	×
데이터베이스 설계도구 'M'	○	×	×	×	×
제안하는 조사도구	○	○	○	○	○

A : 키 연관관계
B : 인덱스로만 설정, 컬럼명이 동일한 연관관계
C : 키, 인덱스 설정, 두 컬럼의 결합에 의한 연관관계
D : 키, 인덱스 미 설정, 컬럼명이 다른 연관관계
E : 키, 인덱스 데이터 영역에 연관 컬럼명이 저장

완벽하게 키, 인덱스로의 연관관계를 맺는 데이터베이스에서는 모델링 도구를 통해 연관관계를 검색할 수 있지만, 실험조건에 따라 연관관계를 검색하지 못하였다. 제안한 조사도구는 연관관계를 정확하게 검색하였다. 하지만 조사자의 판단이나 데이터가 제공되어야만 정확한 검색이 가능하였다.

VI. 결론

제안한 원시데이터의 구조분석 및 데이터 추출 방안은 다음과 같은 이점을 갖는다.

- 다양한 데이터베이스 환경에서 원시 데이터 구조분석방안 적용이 가능하다.
- 연관 테이블 검색 방법으로 기존의 시스템 카탈로그 정보를 포함한 연관관계 검색에 도메인 지식 정보를 활용한 검색 및 검증 방안을 제안함으

로써 원시데이터의 구조 분석에 도움을 준다.

- 설계된 도구를 통하여 질의문을 자동으로 생성한다. 이와 같은 기능은 조사자가 직접 작업해야 하는 질의문 작성 작업을 도구에서 대신 수행함으로써 조사의 편의성을 향상시킨다.
- 원시 데이터 구조 분석 및 추출 도구로부터 데이터베이스로 전송되는 명령 중 테이블 데이터의 변경과 관련된 명령은 모두 필터링 되며, 전송된 모든 명령은 로그로 기록된다. 이와 같은 기능은 조사자가 데이터베이스 조사 데이터에 변경을 가하지 않았음을 증명할 수 있다.

본 논문에서는 기존의 연구 및 기술로 해결할 수 없었던 테이블 간 연관관계분석의 문제점을 보완하였다. 연관관계 검증 및 해석에 있어 기술적 메타정보와 정책적 메타정보를 활용하여 연관관계 정확성을 향상시켰다. 데이터 테이블스페이스에 변경을 가하지 않았음을 증명함으로써 조사 자료의 신뢰성을 향상시켰다. 대용량 데이터베이스 환경에서 최적화된 명령을 사용함으로써 정보체계 자체의 가용성을 해치지 않았고, 데이터 변경 또한 가하지 않았다. 마지막으로 제안한 원시데이터 구조분석을 설계 및 구현된 도구를 통해 실현 가능하도록 하였다.

참고문헌

〔1〕 한국데이터베이스진흥센터, “2007년 데이터베이스 산업 현황 및 전망 보고서”, Jan. 2007

〔2〕 Alhajj, R., “Extracting the extended entity-relationship model from a legacy relational database.”, Information Systems 28, pp.597 - 618, 29 May 2002.

〔3〕 Dowming Yeh, Yuwen Li, William Chu, ″Extracting entity-relationship diagram from a table-based legacy database.″, The Journal of Systems and Software 81, pp. 764 - 771, 26 July 2007.

〔4〕 Jean-Luc Hainaut, Introduction to Database Reverse Engineering, LIBD - Laboratory of Database Application Engineering Institut d′Informatique - University of Namur, 24 Sep. 2002.

〔5〕 ISO/IEC 9126, ″Information Technology - Software Quality Characteristics and metrics“

〔6〕 ISO/IEC 25000, “Software Engineering - Software Quality Requirements and Evaluation (SQuaRE) - Guide to SQuaRE”

〔7〕 ISO/IEC 15504, “Information Technology-Software Process Assessment”

〔8〕 CMMI, “Capability Maturity Model Integration for Development”

〔9〕 행정안전부고시 제2010-85호 정보시스템 감리기준

〔10〕 Oracle, “Advanced Replication Management API Reference 10g Release 1 (10.1)”, Part No. B10733-01, Oracle® Database, Dec. 2003

〔11〕 IBM, “IBM DB2 Database for Linux, UNIX, and Windows”, IBM DB2 Information Center, Nov. 2011

〔12〕 Oracle, “Concepts 10g Release 2 (10.2)”, Part No. B14220-02, Oracle® Database, Oct. 2005

〔13〕 Oracle, “SecureFiles and Large Objects Developer′s Guide 11g Release 2 (11.2)”, Part No. E18294-01, Oracle® Database, Aug. 2010

〔14〕 Oracle, “Programmer′s Guide 10g Release 1 (10.1)”, Part No. B10778-01, Oracle® C++ Call Interface, Dec. 2003

〈著者紹介〉

이 동 찬 (Dongchan Lee) 정회원
2010년 5월~현재: 한국국방연구원 국방정보체계관리단 전문연구원
〈관심분야〉 데이터베이스, 정보보호

이 상 진 (Sangjin Lee) 정회원
1989년~1999년: ETRI 연구원
1999년~현재 : 고려대학교 교수
〈관심분야〉 디지털포렌식, 모바일포렌식, 심층 암호, 해쉬 함수