(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2017년07월10일
(11) 등록번호 10-1755365
(24) 등록일자 2017년07월03일

(51) 국제특허분류(Int. Cl.)
G06F 17/30 (2006.01)
(52) CPC특허분류
G06F 17/30371 (2013.01)
(21) 출원번호 10-2016-7033559(분할)
(22) 출원일자(국제) 2010년11월12일
심사청구일자 2016년11월30일
(85) 번역문제출일자 2016년11월30일
(65) 공개번호 10-2016-0141872
(43) 공개일자 2016년12월09일
(62) 원출원 특허 10-2012-7013690
원출원일자(국제) 2010년11월12일
심사청구일자 2015년03월02일
(86) 국제출원번호 PCT/US2010/056530
(87) 국제공개번호 WO 2011/060257
국제공개일자 2011년05월19일
(30) 우선권주장
61/260,997 2009년11월13일 미국(US)
(56) 선행기술조사문헌
US20090024639 A1
(뒷면에 계속)
전체 청구항 수 : 총 16 항

(73) 특허권자
아브 이니티오 테크놀로지 엘엘시
미국 02421 매사추세츠주 렉싱턴 스프링 스트리트 201
(72) 발명자
파멘터 다비드 더블유.
미국 02458 매사추세츠주 뉴턴 휴웰 에비뉴 165
구드 조엘
미국 02474 매사추세츠주 알링턴 리 테라스 27
(뒷면에 계속)
(74) 대리인
유미특허법인

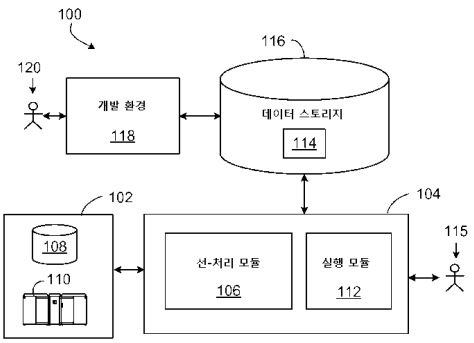
심사관 : 최정권

(54) 발명의 명칭 **레코드 포맷 정보의 관리**

(57) 요 약

데이터는, 포맷 정보를 이용하여 데이터 처리 시스템 내에서 처리를 위해 준비된다. 수신된 데이터는 필드들에 대한 값들을 가지는 레코드들을 포함한다. 데이터를 처리하기 위한 타겟 레코드 포맷이 판정된다. 데이터가 후보 레코드 포맷들에 매치하는지 여부를 판정(810)하기 위한 유효성 테스트에 따라 다수의 레코드들이 분석된다(806). 각 후보 레코드 포맷은 각 필드에 대한 포맷을 지정하고, 각 유효성 테스트는 적어도 하나의 후보 레코드 포맷에 대응한다. 유효성 테스트들의 결과들의 수신에 응하여, 타겟 레코드 포맷이, 적어도 하나의 유효성 테스트에 따라 적어도 일부의 매치가 판정된 후보 레코드 포맷(812), 데이터와 관련된 데이터 타입에 따라 선택된 분석된 레코드 포맷(830, 832, 834, 836, 838), 및 데이터의 특성들의 분석으로부터 생성된 구성된 레코드 포맷(846), 중 적어도 하나에 기초하여 데이터와 관련된다.

대 표 도 - 도1



(72) 발명자

퍼버 제니퍼 엠.

미국 60615 일리노이주 시카고 에스. 캔우드 에비뉴 4815

프레운드리크 로버트

미국 01776 매사추세츠주 서드베리 메이플 에비뉴 55

비뇨 조이스 엘

미국 02180 매사추세츠주 스톤엄 포레스트 스트리트 45

(56) 선행기술조사문헌

US20070276858 A1

US20080215528 A1

US20040139076 A1

JP2001101049 A

명 세 서

청구범위

청구항 1

데이터 저장 시스템 내에서 데이터의 포맷을 설명하는 후보 레코드 포맷을 판정함으로써 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법으로서,

각 필드(field)에 대한 하나 이상의 값을 각각 가지는 레코드들을 포함하는 데이터를 입력 장치 또는 포트를 통해 수신하는 단계;

후보 레코드 포맷들에 액세스하는 단계 - 각각의 후보 레코드 포맷은 하나 이상의 필드의 그룹의 각 필드에 대한 포맷을 정하고 있음 -;

액세스된 둘 이상의 특정 후보 레코드 포맷의 각각에 대하여,

(i) 다수의 레코드 중 각각의 레코드 내의 필드에 대한 값을 (ii) 상기 특정 후보 레코드 포맷에 의해 정해진 필드에 대한 포맷과 비교하고,

상기 특정 후보 레코드 포맷에 의해 정해진 필드에 대한 데이터 타입과 일치하는 값을 가지는 다수의 레코드 내의 필드의 양에 기초하여, 상기 하나 이상의 필드의 그룹의 각 필드에 대한 포맷을 정하고 있는 특정 후보 레코드 포맷에 대한 성공의 척도를 판정하는 단계 - 상기 성공의 척도는 특정 후보 레코드 포맷이 수신된 데이터 내의 각각의 레코드의 포맷을 어느 정도로 정확히 설명하는지를 나타냄 -; 및

상기 둘 이상의 특정 후보 레코드 포맷의 각각에 대한 비교의 수행에 기초하여, 상기 둘 이상의 특정 후보 레코드 포맷에 대한 성공의 척도를 지시하는 정보를 출력하는 단계

를 포함하는,

데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법.

청구항 2

제1항에 있어서,

상기 데이터는 알려진 파일 타입과 관련되는 것인, 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법.

청구항 3

제2항에 있어서,

상기 데이터의 상기 파일 타입은 파일 확장자(file extension)에 대응하는, 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법.

청구항 4

제1항에 있어서,

수신된 데이터는 제1 데이터를 포함하고, 상기 특정 후보 레코드 포맷은 제1 후보 레코드 포맷을 포함하고,

상기 데이터 처리 시스템 내에서 처리하기 위한 데이터를 준비하는 방법은,

상기 입력 장치 또는 포트를 통해 각 필드에 대한 하나 이상의 값을 각각 가지는 레코드들을 포함하는 제2 데이터를 수신하는 단계; 및

상기 제2 데이터에 대한 유효성 테스트의 적용, 상기 후보 레코드 포맷 중 하나 이상에 대한 적어도 일부의 매치를 판정하는 상기 제2 데이터에 적용되는 유효성 테스트가 없는 것, 및 상기 제2 데이터와 관련된 알려진 데이터 타입을 가지지 않는 것에 기초하여, 제2 후보 레코드 포맷을 상기 제2 데이터와 관련시키는 단계를 더 포

함하는, 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법.

청구항 5

제4항에 있어서,

상기 제2 데이터 내의 태그(tag)들을 인식하고, 상기 인식된 태그들에 기초하여 다수의 레코드를 판정하기 위해 상기 제2 데이터를 파싱함으로써, 상기 제2 데이터의 특성들의 분석으로부터 구성된 레코드 포맷을 생성하는 단계를 더 포함하는, 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법.

청구항 6

제4항에 있어서,

상기 제2 데이터 내의 디리미터(delimiter)들을 인식하는 단계, 및 상기 인식된 디리미터들에 기초하여 다수의 레코드를 판정하기 위해 상기 제2 데이터를 파싱하는 단계를 더 포함하는, 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법.

청구항 7

제4항에 있어서,

상기 제2 데이터가 다수의 레코드의 값들을 지시하는 태그들 또는 디리미터들이 없는 바이너리 형태인 것을 인식하고, 사용자 인터페이스로부터 하나 이상의 필드 식별자(identifier)를 수신함으로써, 상기 제2 데이터의 특성들의 분석으로부터 구성된 레코드 포맷을 생성하는 단계를 더 포함하는, 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법.

청구항 8

제1항에 있어서,

각각의 필드에 대한 후보 레코드 포맷에 의해 정해지는 포맷들 내의 각각의 레코드에 대한 값을 판정하기 위해 상기 특정 후보 레코드 포맷을 상기 데이터에 적용하는 단계를 더 포함하는, 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법.

청구항 9

제1항에 있어서,

유효 값들의 수가 미리 정해진 임계치보다 큰지 여부를 판정하기 위하여 유효성 테스트에 따라 상기 다수의 레코드에 대한 값들을 분석함으로써 상기 데이터가 상기 특정 후보 레코드 포맷에 매치하는지 여부를 판정하는 단계를 더 포함하는, 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법.

청구항 10

제9항에 있어서,

상기 유효성 테스트에 따라 상기 다수의 레코드의 제1 레코드에 대해 판정된 값들을 분석하는 것은, 각 필드에 대해 판정된 각 값에 대응하는 필드 테스트를 수행하는 것을 포함하는, 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법.

청구항 11

제10항에 있어서,

제1 필드에 대해 판정된 값에 제1 필드 테스트를 수행하는 것은, 상기 판정된 값 내의 문자 수와 미리 정해진 문자 수를 매치하는 것을 포함하는, 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법.

청구항 12

제10항에 있어서,

제1 필드에 대해 판정된 값에 제1 필드 테스트를 수행하는 것은, 상기 판정된 값을 상기 제1 필드에 대해 미리 정해진 다수의 유효 값들 중 하나와 매치하는 것을 포함하는, 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법.

청구항 13

제10항에 있어서,

상기 유효 값들의 수는, 주어진 필드에 대해 상기 판정된 값이 상기 주어진 필드에 대응하는 상기 필드 테스트를 통과하는 레코드들의 수에 기초하는, 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 방법.

청구항 14

데이터 저장 시스템 내에서 데이터의 포맷을 설명하는 후보 레코드 포맷을 판정함으로써 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 시스템으로서,

각 필드에 대한 하나 이상의 값을 각각 가지는 레코드들을 포함하는 데이터를 입력 장치 또는 포트를 통해 수신하는 수단;

후보 레코드 포맷들에 액세스하는 수단 - 각각의 후보 레코드 포맷은 하나 이상의 필드의 그룹의 각 필드에 대한 포맷을 정하고 있음 -;

액세스된 둘 이상의 특정 후보 레코드 포맷의 각각에 대하여, (i) 다수의 레코드 중 각각의 레코드 내의 필드에 대한 값을 (ii) 상기 특정 후보 레코드 포맷에 의해 정해진 필드에 대한 포맷과 비교하는 수단;

액세스된 둘 이상의 특정 후보 레코드 포맷의 각각에 대하여, 상기 특정 후보 레코드 포맷에 의해 정해진 필드에 대한 데이터 타입과 일치하는 값을 가지는 다수의 레코드 내의 필드의 양에 기초하여, 상기 하나 이상의 필드의 그룹의 각 필드에 대한 포맷을 정하고 있는 특정 후보 레코드 포맷에 대한 성공의 척도를 판정하는 수단 - 상기 성공의 척도는 특정 후보 레코드 포맷이 수신된 데이터 내의 각각의 레코드의 포맷을 어느 정도로 정확히 설명하는지를 나타냄 -; 및

상기 둘 이상의 특정 후보 레코드 포맷의 각각에 대한 비교의 수행에 기초하여, 상기 둘 이상의 특정 후보 레코드 포맷에 대한 성공의 척도를 지시하는 정보를 출력하는 수단

를 포함하는,

데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 시스템.

청구항 15

데이터 저장 시스템 내에서 데이터의 포맷을 설명하는 후보 레코드 포맷을 판정함으로써 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하기 위한 컴퓨터 프로그램이 저장된 컴퓨터 판독가능 저장 매체로서,

상기 컴퓨터 프로그램은 컴퓨터로 하여금,

각 필드에 대한 하나 이상의 값을 각각 가지는 레코드들을 포함하는 데이터를 입력 장치 또는 포트를 통해 수신하고,

후보 레코드 포맷들에 액세스하고 - 각각의 후보 레코드 포맷은 하나 이상의 필드의 그룹의 각 필드에 대한 포맷을 정하고 있음 -,

액세스된 둘 이상의 특정 후보 레코드 포맷의 각각에 대하여,

(i) 다수의 레코드 중 각각의 레코드 내의 필드에 대한 값을 (ii) 상기 특정 후보 레코드 포맷에 의해 정해진 필드에 대한 포맷과 비교하고,

상기 특정 후보 레코드 포맷에 의해 정해진 필드에 대한 데이터 타입과 일치하는 값을 가지는 다수의 레코드 내의 필드의 양에 기초하여, 상기 하나 이상의 필드의 그룹의 각 필드에 대한 포맷을 정하고 있는 특정 후보 레코드 포맷에 대한 성공의 척도를 판정하고 - 상기 성공의 척도는 특정 후보 레코드 포맷이 수신된 데이터 내의 각각의 레코드의 포맷을 어느 정도로 정확히 설명하는지를 지시함 -,

상기 둘 이상의 특정 후보 레코드 포맷의 각각에 대한 비교의 수행에 기초하여, 상기 둘 이상의 특정

후보 레코드 포맷에 대한 성공의 척도를 지시하는 정보를 출력하도록 하는

명령어를 포함하는,

데이터 처리 시스템 내에서 처리하도록 데이터를 준비하기 위한 컴퓨터 프로그램이 저장된 컴퓨터 판독가능 저장 매체.

청구항 16

데이터 저장 시스템 내에서 데이터의 포맷을 설명하는 후보 레코드 포맷을 판정함으로써 데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 컴퓨팅 시스템으로서,

각 필드에 대한 하나 이상의 값을 각각 가지는 레코드들을 포함하는 데이터를 수신하도록 구성된 입력 포트; 및

적어도 하나의 프로세서

를 포함하고,

상기 적어도 하나의 프로세서는,

후보 레코드 포맷들에 액세스하고 - 각각의 후보 레코드 포맷은 하나 이상의 필드의 그룹의 각 필드에 대한 포맷을 정하고 있음 -,

액세스된 둘 이상의 특정 후보 레코드 포맷의 각각에 대하여,

(i) 다수의 레코드 중 각각의 레코드 내의 필드에 대한 값을 (ii) 상기 특정 후보 레코드 포맷에 의해 정해진 필드에 대한 포맷과 비교하고,

상기 특정 후보 레코드 포맷에 의해 정해진 필드에 대한 데이터 타입과 일치하는 값을 가지는 다수의 레코드 내의 필드의 양에 기초하여, 상기 하나 이상의 필드의 그룹의 각 필드에 대한 포맷을 정하고 있는 특정 후보 레코드 포맷에 대한 성공의 척도를 판정하고 - 상기 성공의 척도는 특정 후보 레코드 포맷이 수신된 데이터 내의 각각의 레코드의 포맷을 어느 정도로 정확히 설명하는지를 나타냄 -,

상기 둘 이상의 특정 후보 레코드 포맷의 각각에 대한 비교의 수행에 기초하여, 상기 둘 이상의 특정 후보 레코드 포맷에 대한 성공의 척도를 지시하는 정보를 출력하도록 구성된,

데이터 처리 시스템 내에서 처리하도록 데이터를 준비하는 컴퓨팅 시스템.

발명의 설명

기 술 분 야

[0001] 본 출원은 참조로서 본 명세서에 병합되는 것으로, 2009년 11월 13일에 출원된 미국 특허 출원 61/260,997호에 대해 우선권을 주장한다.

[0002] 본 명세서는 레코드 포맷 정보(record format information)의 관리에 관련된다.

배 경 기 술

[0003] 기관들은 다수의 상이한 시스템들로부터의 데이터를 관리한다. 시스템은 그 시스템에 대한 고유의 형식으로 데이터의 데이터 세트를 생성할 수 있다. 다른 시스템들은, 콤마 독립 파일(Comma-Separated File) 또는 XML 문서와 같은 표준 포맷(standard format)을 이용하여 데이터 세트들을 생성한다. 대체로, 일단 데이터 세트의 포맷이 표준이라면, 데이터 세트 내의 레코드 및 필드(field)들은 시스템에 대해 명확하게 된다.

[0004] 일부 시스템들은 임포트 메커니즘(import mechanism)을 통해 다른 시스템들에 의해 제공된 데이터 세트들을 받아들인다. 임포트는 외부의 데이터 세트를 처리하기 위해 시스템에 대한 고유의 형식으로 변환한다. 다른 시스템들은, 시스템이 필수적으로 요구되는 변환 없이도 외부의 데이터 세트를 처리하도록 하기 위해, 데이터 세트를 충분히 설명하는 레코드 포맷을 생성한다.

발명의 내용

과제의 해결 수단

[0005] 일 측면에서, 전반적으로, 데이터 저장 시스템(data storage system) 내의 포맷 정보에 기초하여 데이터 처리 시스템(data processing system) 내에서 처리하기 위해 데이터를 준비하는 방법이 제시된다. 각각이 입력 장치 또는 포트 상의 각각의 필드들에 대한 하나 이상의 값을 가지는 레코드들을 포함하는 데이터가 수신된다. 데이터 처리 시스템 내에서 데이터를 처리하기 위해 타겟 레코드 포맷(target record format)이 판정된다. 데이터 저장 시스템 내에 저장된 하나 이상의 후보 레코드 포맷(candidate record format)들에 데이터가 매치(match)하는지 여부를 판정하기 위한 복수의 유효성(validation) 테스트들에 따라, 데이터 내의 다수의 레코드 포맷이 분석된다. 각각의 후보 레코드 포맷은 하나 이상의 필드로 된 그룹의 각각의 필드에 대한 포맷을 정하고, 각각의 유효성 테스트는 데이터 저장 시스템 내에 저장된 적어도 하나 이상의 후보 레코드 포맷에 대응한다. 유효성 테스트들의 결과를 수신하는 경우에 응하여, 타겟 레코드 포맷은, 선택된 후보 레코드 포맷 - 선택된 후보 레코드 포맷에 대응하는 적어도 하나의 유효성 테스트에 따라 적어도 일부의 매치가 판정된 것임 -, 데이터와 관련된 알려진 데이터 타입에 따라 파서(parser)에 의해 선택되어 생성된 파싱된 레코드 포맷(parsed record format), 및 데이터의 특성들의 분석으로부터 생성된 구성된 레코드 포맷(constructed record format) 중 적어도 하나에 기초하여 데이터와 관련(associated)된다.

[0006] 본 발명의 태양들은 이하의 특징들 중 하나 이상을 포함할 수 있다.

[0007] 어떠한 유효성 테스트들도 하나 이상의 후보 레코드 포맷에 대해 적어도 일부의 매치를 판정하지 못한 경우에 응하여, 파싱된 레코드 포맷에 기초하여 타겟 레코드 포맷과 데이터를 관련시킨다. 데이터와 관련된 알려진 데이터 타입은 데이터의 파일 타입에 기초하여 인지될 수 있다. 데이터의 파일 타입은 파일 확장자(extension)에 대응할 수 있다. 어떠한 유효성 테스트들도 하나 이상의 후보 레코드 포맷에 대해 적어도 일부의 매치를 판정하지 못하고, 데이터와 관련된 알려진 데이터 타입을 가지지 않는 경우에 응하여, 구성된 레코드 포맷에 기초하여 타겟 레코드 포맷과 데이터를 관련시킨다. 데이터의 특성들의 분석으로부터 구성된 레코드 포맷을 생성하는 것은, 데이터 내의 태그(tag)들을 인식하고, 인식된 태그들에 기초하여 다수의 레코드들을 판정하기 위해 데이터를 파싱(parsing)하는 것을 포함할 수 있다. 데이터의 특성들의 분석으로부터 구성된 레코드 포맷을 생성하는 것은, 데이터 내의 디리미터(delimiter)들을 인식하고, 인식된 경계들에 기초하여 다수의 레코드들을 판정하기 위해 데이터를 파싱하는 것을 포함할 수 있다. 데이터의 특성들의 분석으로부터 구성된 레코드 포맷을 생성하는 것은, 다수의 레코드들의 값을 지시하는 태그나 디리미터 없이도 데이터가 실질적으로 바이너리 형식(binary form)이라는 것을 인식하는 것과 사용자 인터페이스(user interface)로부터 하나 이상의 필드 식별자(identifier)들을 수신하는 것을 포함할 수 있다. 제1 후보 레코드 포맷에 대응하는 복수의 유효성 테스트들 중 제1 유효성 테스트에 따라 데이터 내의 다수의 레코드들을 분석하는 것은, 제1 후보 레코드 포맷을 데이터에 적용하여 각 필드에 대한 제1 후보 레코드 포맷에 의해 지정된(specified) 포맷들 내의 각 레코드에 대한 값들을 판정하는 것을 포함할 수 있다. 제1 후보 레코드 포맷에 데이터가 매치하는지 여부를 판정하는 것은, 제1 유효성 테스트에 따라 다수의 레코드들에 대해 판정된 값들을 분석하여 유효 값(valid value)들의 수가 미리 정해진 임계치(threshold)보다 더 큰지 여부를 판정하는 것을 포함할 수 있다. 제1 유효성 테스트에 따라 다수의 레코드들 중 제1 레코드에 대해 판정된 값들을 분석하는 것은, 각 필드에 대해 각각의 판정된 값에 대응하는 필드 테스트를 수행하는 것을 포함할 수 있다. 제1 필드에 대해 판정된 값에 제1 필드 테스트를 수행하는 것은, 판정된 값 내의 문자(character) 수와 미리 정해진 문자 수를 매치하는 것을 포함할 수 있다. 제1 필드에 대해 판정된 값에 제1 필드 테스트를 수행하는 것은, 판정된 값을 제1 필드에 대해 미리 정해진 다수의 유효 값 중 하나에 매치하는 것을 포함할 수 있다. 유효 값들의 수는, 주어진 필드(given field)에 대해 판정된 값이 주어진 필드에 대응하는 필드 테스트를 통과하는 레코드들의 수에 기초될 수 있다.

[0008] 또 다른 측면에서, 전반적으로, 데이터 저장 시스템 내의 포맷 정보에 기초하여 데이터 처리 시스템 내에서 처리하기 위해 데이터를 준비하는 시스템은: 각각이 입력 장치 또는 포트 상의 각각의 필드들에 대한 하나 이상의 값을 가지는 레코드들을 포함하는 데이터를 수신하기 위한 수단; 및 데이터 처리 시스템 내에서 데이터를 처리하기 위해 타겟 레코드 포맷을 판정하는 수단을 포함한다. 타겟 레코드 포맷을 판정하는 수단은: 데이터 저장 시스템 내에 저장된 하나 이상의 후보 레코드 포맷들에 데이터가 매치하는지 여부를 판정하기 위해 복수의 유효성 테스트들에 따라 데이터 내의 다수의 레코드 포맷, 하나 이상의 필드로 된 그룹의 각각의 필드에 대한 포맷을 정하는 각각의 후보 레코드 포맷, 및 데이터 저장 시스템 내에 저장된 적어도 하나 이상의 후보 레코드 포맷에 대응하는 각각의 유효성 테스트를 분석하고, 유효성 테스트들의 결과를 수신하는 경우에 응하여, 타겟 레코드 포맷을. 선택된 후보 레코드 포맷 - 선택된 후보 레코드 포맷에 대응하는 적어도 하나의 유효성 테스트에 따라 적어도 일부의 매치가 판정된 것임 -, 데이터와 관련된 알려진 데이터 타입에 따라 파서에 의해 선택되어 생성된 파싱된 레코드 포맷, 및 데이터의 특성들의 분석으로부터 생성된 구성된 레코드 포맷 중 적어도 하나에 기

초하여 데이터와 관련시키도록 구성된다.

[0009] 또 다른 측면에서, 전반적으로, 컴퓨터 판독가능 매체(computer-readable medium)가, 데이터 저장 시스템 내의 포맷 정보에 기초하여 데이터 처리 시스템 내에서 처리하기 위해 데이터를 준비하는 컴퓨터 프로그램을 저장한다. 컴퓨터 프로그램은, 컴퓨터가, 각각이 입력 장치 또는 포트 상의 각각의 필드들에 대한 하나 이상의 값을 가지는 레코드들을 포함하는 데이터를 수신하고; 데이터 저장 시스템 내에 저장된 하나 이상의 후보 레코드 포맷들에 데이터가 매치하는지 여부를 판정하기 위한 복수의 유효성 테스트들에 따라 데이터 내의 다수의 레코드 포맷, 하나 이상의 필드로 된 그룹의 각각의 필드에 대한 포맷을 정하는 각각의 후보 레코드 포맷, 및 데이터 저장 시스템 내에 저장된 적어도 하나 이상의 후보 레코드 포맷에 대응하는 각각의 유효성 테스트를 분석하고, 유효성 테스트들의 결과를 수신하는 경우에 응하여, 타겟 레코드 포맷을, 선택된 후보 레코드 포맷 - 선택된 후보 레코드 포맷에 대응하는 적어도 하나의 유효성 테스트에 따라 적어도 일부의 매치가 판정된 것임 -, 데이터와 관련된 알려진 데이터 타입에 따라 파서에 의해 선택되어 생성된 파싱된 레코드 포맷, 및 데이터의 특성들의 분석으로부터 생성된 구성된 레코드 포맷 중 적어도 하나에 기초하여 데이터와 관련시키는 것을 포함하여, 데이터 처리 시스템 내에서 데이터를 처리하기 위해 타겟 레코드 포맷(target record format)을 판정하도록 야기시키는 명령어들을 포함한다.

[0010] 본원 발명의 다른 특징들과 장점들은 이하의 상세한 설명과 청구범위로부터 명확해질 것이다.

도면의 간단한 설명

[0011] 도 1은, 그래프 기반 연산들을 실행하기 위한 시스템에 대한 블록 다이어그램이다.

도 2는, 레코드 포맷 정보를 관리하기 위한 예시적인 과정에 대한 플로우 차트이다.

도 3은, 예시적인 선-처리 모듈(pre-processing module)에 대한 블록 다이어그램이다.

도 4는, 샘플 데이터에 기초하여 레코드 포맷을 판정하는 선-처리 모듈의 예시적인 처리를 보여주는 블록 다이어그램이다.

도 5는, 샘플 데이터에 기초하여 레코드 포맷을 유효하게 하는 선-처리 모듈의 예시적인 처리를 보여주는 블록 다이어그램이다.

도 6은, 샘플 데이터에 기초하여 존재하는 레코드 포맷을 식별하는 선-처리 모듈의 예시적인 처리를 보여주는 블록 다이어그램이다.

도 7은, 파서에 기초하여 레코드 포맷을 생성하는 선-처리 모듈의 예시적인 처리를 보여주는 블록 다이어그램이다.

도 8은, 레코드 포맷 정보를 관리하는 예시적인 과정에 대한 플로우 차트이다.

발명을 실시하기 위한 구체적인 내용

[0012] 도 1은 레코드 포맷 관리 기술(record format management technique)들이 이용될 수 있는 예시적인 데이터 처리 시스템(data processing system)(100)을 도시한다. 시스템(100)은, 어떠한 복수의 스토리지 포맷들(예컨대, 데이터 베이스 테이블들, 스프래드시트(spreadsheet) 파일들, 플랫 텍스트 파일, 또는 메인프레임에 의해 이용되는 고유의 포맷)을 가지는 데이터를 저장할 수 있는 각각의 저장 장치들 또는 온라인 데이터 스트림(online data stream)들에 대한 접속들과 같은 하나 이상의 데이터의 소스를 포함할 수 있는 데이터 소스(data source)(102)를 포함한다. 실행 환경(execution environment)(104)은, 선-처리 모듈(pre-processing module)(106)과 실행 모듈(execution module)(112)을 포함한다. 실행 환경(104)은, UNIX 운영 시스템(operation system)과 같은 적합한 운영 시스템의 제어 아래에서 하나 이상의 범용 컴퓨터에서 관리될 수 있다. 예를 들어, 실행 환경(108)은, 근거리에 위치하거나(예컨대, SMP 컴퓨터들과 같은 멀티프로세서 시스템들), 근접하여 분배되거나(예컨대, 클러스터들 또는 MPP들과 같이 연결된 다수의 프로세서들), 원거리 또는 원거리에서 분배되거나(예컨대, 근거리 통신망(local area network, LAN) 및/또는 원거리 통신망(wide area network, WAN)을 통해 연결된 다수의 프로세서들) 그것에 관한 어떠한 조합 중 어느 하나인 다수의 중앙 처리 장치(central processing unit, CPU)들을 이용하여 컴퓨터 시스템의 구성을 포함하는 다점 병렬 연산 환경(multiple-node parallel computing environment)을 포함할 수 있다. 일부의 구현에서는, 실행 모듈(112)은 하나 이상의 프로세서들 상에서 실행되고 있는 병렬 운영 시스템일 수 있는 운영 시스템을 제공하고, 선-처리 모듈(106)은 그 운영 시스템 내에서 실행되고 있는 프로그램으로서 실행된다. 사용자(115)도, 표시되는 출력들을 보는 것 및 사

용자 인터페이스에 입력들을 입력하는 것에 의해 실행 환경(108)과 상호 작용(interact)하는 것이 가능하다.

[0013] 선-처리 모듈(106)은, 데이터 소스(102)로부터 각각의 레코드가 각각의 필드들에 대한 하나 이상의 값을 가지는 레코드들을 포함하는 데이터를 수신하고, 실행 모듈(112)를 이용하여 레코드들을 처리하기 위해 타겟 레코드 포맷(target record format)을 판정한다. 예를 들어, 선-처리 모듈(106)은, 데이터 저장 시스템(data storage system)(116) 내에 적정한 타겟 레코드 포맷(114)이 이미 저장되어 있는지를 판정하고, 저장되어 있지 않다면, 타겟 레코드 포맷(114)를 생성하고 데이터 저장 시스템(116) 내에 생성된 타겟 레코드 포맷(114)을 저장한다. 데이터 소스(102)와 데이터 저장 시스템(116)을 제공하는 저장 장치들은, 예를 들어 실행 환경(104)을 실행하고 있는 컴퓨터와 접속된 저장 매체(예컨대, 하드 드라이브(108))에 저장되어 있는 것처럼, 실행 환경(104)과 근거리에 있을 수 있고, 예를 들어 원격 접속 상에서 실행 환경(104)를 실행하고 있는 컴퓨터와 통신하는 리모트 시스템(remote system)에서 관리되는 것처럼, 실행 환경(104)와 원거리에 있을 수 있다.

[0014] 실행 모듈(112)은, 데이터 소스(102)로부터 수신된 레코드들을 해석하고 처리하기 위해 판정된 타겟 레코드 포맷(114)을 이용한다. 데이터 저장 시스템(116)도, 개발자(developer)(120)가 레코드들을 처리하기 위해 실행 모듈(112)에 의해 실행될 프로그램을 개발할 수 있는 개발 환경(development environment)(118)에 액세스할 수 있다. 개발 환경(118)은, 일부의 구현들에서, 꼭지점들 사이의 다이렉트 링크(link)들(워크 엘리먼트(work element)들의 플로우를 표현하는)에 의해 연결된 꼭지점들(구성요소들 또는 데이터세트들)을 포함하는 데이터 플로우 그래프들과 같은 어플리케이션을 개발하기 위한 시스템이다. 예를 들어, 그러한 것은 본 명세서에 참조로서 병합된 "그래프 기반 어플리케이션의 파라미터들의 관리"로 명명된 미국 특허 출원 2007/0011668호에서 더 상세하게 설명된다.

[0015] 선-처리 모듈(106)은, 상이한 형태의 데이터 베이스 시스템들을 포함하여 다양한 타입의 시스템들로부터 데이터를 수신할 수 있다. 데이터는, 0일 수도 있는 값들을 포함하는 각각의 필드들에 대한 값들("속성(attributes)" 또는 "칼럼(columns)"으로 불리기도 하는)을 가지는 레코드들로서 조직된다. 데이터 소스로부터 데이터를 처음 판독할 때, 그 데이터 소스로부터 레코드들의 레코드 구조를 설명하는 타겟 레코드 포맷이 인지되지 않으면, 비록 일부 환경에서, 선-처리 모듈(106)은 그 데이터 소스 내의 레코드들에 대한 초기 포맷 정보 일부를 가지고 시작한다. 선-처리 모듈(106)은 처리될 레코드들이 저장된 레코드 포맷에 의해 설명되는지 또는 레코드 포맷이 생성될 것인지 여부를 판정하기 위해, 데이터 저장 시스템(116) 내에 저장된 레코드 포맷들의 집합(collection)을 관리한다. 레코드 포맷은, 특별한(distinct)값을 표현하는 다수의 비트들, 레코드 내의 필드들의 순서, 및 비트들에 의해 표현되는 값의 타입(예컨대, 스트링(string), 표시된/표시되지 않은 인티저(integer))과 같은 다양한 특성들을 포함할 수 있다.

[0016] 도 2를 참조하면, 프로세스(220)에 대한 플로우 차트는 레코드 포맷들을 관리하기 위한 선-처리 모듈(106)의 일부 동작들을 포함한다. 다른 가능성들 사이에서, 선-처리 모듈(106)은 데이터를 받아들인다(222). 데이터는, 파일, 데이터 베이스, 사용자 인터페이스, 입력 포트, 또는 다른 어떠한 입력 장치를 통해 수신될 수 있다. 다른 정보들 사이에서, 선-처리 모듈(106)은 데이터 소스로부터의 레코드들에 대한 레코드 포맷을 포함하는 데이터, 데이터 소스(102)로부터의 하나 이상의 레코드들을 포함하는 샘플 데이터, 또는 양자 모두를 수신할 수 있다. 샘플 데이터는, 처리될 모든 레코드들 또는 레코드들의 서브세트(subset)를 포함할 수 있다. 선-처리 모듈(106)은, 선-처리 모듈(106)이 수행되기 위해 어떤 동작들이 요구되는지에 관한 지시(indication)를 수신할 수도 있다.

[0017] 선-처리 모듈(106)의 동작들은 프로세스 경로(process path)의 판정(224)을 포함할 수도 있다. 선-처리 모듈(106)은, 수신된 샘플 데이터의 레코드들을 해석하기 위해 레코드 포맷을 판정하는 다양한 방법들을 가질 수 있다. 선-처리 모듈(106)은, 샘플 데이터에 대한 잠재 레코드 포맷(potential record format)이 입력으로서 제공되는지 여부에 기초하여, 어떠한 프로세스 경로가 적절한지를 판정할 수 있다. 시스템의 일부 구현들에서는, 선-처리 모듈(106)이 어떤 프로세스 경로가 선호되는지를 지시하는 데이터를 받아들인다.

[0018] 하나의 프로세스 경로를 따라, 선-처리 모듈(106)의 동작들은, 이하에서 더욱 상세하게 설명되는 것처럼, 샘플 데이터의 분석에 기초한 샘플 데이터의 타겟 레코드 포맷의 판정(226)을 포함한다.

[0019] 또 다른 프로세스 경로를 따라, 선-처리 모듈(106)의 동작들은, 제공된 레코드 포맷과 제공된 샘플 데이터의 비교(228)에 기초한 샘플 데이터의 타겟 레코드 포맷의 판정을 포함한다. 일부 케이스들에서는, 선-처리 모듈(106)은, 샘플 데이터와, 받아들여진 샘플 데이터에 잠재적으로 대응하는 제공된 레코드 포맷(또는 저장된 레코드 포맷에 대한 식별자(identifier))를 받아들인다. 선-처리 모듈(106)은, 레코드 포맷이 샘플 데이터의 구조(structure)를 표현하고 있는지 여부를 판정하기 위해, 제공된 또는 식별된 레코드 포맷을 샘플 데이터와 비교

한다.

[0020] 또 다른 프로세스 경로를 따라, 선-처리 모듈(106)의 동작들은, 제공된 샘플 데이터에 대한 레코드 포맷의 검색(finding)(230)에 기초한 샘플 데이터의 타겟 레코드 포맷의 판정을 포함한다. 일부 케이스들에서는, 선-처리 모듈(106)은, 어떠한 레코드 포맷들이 정확하게 샘플 데이터의 구조를 표현하고 있는지 여부를 발견하기 위해, 샘플 데이터를 받아들이고, 데이터를 레코드 포맷 집적소(repository)(예컨대, 데이터 저장 시스템(116) 내에서 관리되는) 내에서 기존의 레코드 포맷(existing record format)과 비교한다.

[0021] 동작들은, 사용자에 대한 하나 이상의 잠재적 타겟 레코드 포맷들의 표시(232)를 포함할 수도 있다. 일단 하나 이상의 레코드 포맷들이 판정되면, 레코드 포맷들은 사용자에 대해 표시될 수 있다. 사용자는 복수개의 레코드 포맷들로부터 하나의 레코드 포맷을 선택할 수 있다. 사용자는 레코드 포맷을 수정할 수도 있다.

[0022] 동작들은 타겟 레코드 포맷의 유효화(234)를 포함할 수도 있다. 레코드 포맷이 선-처리 모듈(106)에 의해 받아들여지기 전에, 선-처리 모듈(106)은 제공된 샘플 데이터에 대한 레코드 포맷들 유효화할 수 있다.

[0023] 동작들은 타겟 레코드 포맷에 대해 조정들을 제안(236)하는 것도 포함한다. 레코드 포맷이 제공된 샘플 데이터를 파싱할 수 없다면, 선-처리 모듈(106)은 레코드 포맷과 샘플 데이터 사이의 불일치들(inconsistencies)을 식별한다. 불일치들은, 샘플 데이터를 파싱할 때 발생하는 에러들의 분석에 의해 식별될 수 있다. 불일치들은 샘플 데이터와 레코드 포맷을 분석하는 것에 의해서도 식별될 수 있다. 프로세스(220)는, 불일치를 수정하는 제안들을 생성한다. 일 실시예에서는, 프로세스(220)가 샘플 데이터에 기초한 레코드 포맷의 수정을 제시할 수 있다. 예를 들어, 레코드 포맷이 필드가 인티저(integer)의 표현(예컨대, 1, 2, 3, 4, 등과 같은 인티저 값의 바이너리 표현)이 되길 기대하고, 샘플 데이터 내의 필드가 포맷된 날짜(formatted date)의 표시(예컨대, 1/21/2008, 21/1/2008, 01-JAN-2008 등)를 포함한다면, 프로세스(220)는 조정을 제안할 수 있다. 인티저 필드가 포맷된 날짜를 유지할 수 없고 날짜 필드가 인티저를 유지할 수 없기 때문에, 프로세스(220)는 필드를, 날짜 또는 인티저 중 어느 하나를 포함하는 스트링(string)으로 수정하도록 제안할 수 있다. 또 다른 예에서는, 프로세스(220)는 레코드 포맷에 의해 받아들여진 유효 값들의 범위를 확장하도록 제안할 수 있다.

[0024] 동작들은, 타겟 레코드 포맷을 저장(238)하는 것도 포함한다. 타겟 레코드 포맷은 레코드 포맷 집적소 내에 저장될 수 있다.

[0025] 도 3을 참조하면, 데이터 처리 시스템 내에서 처리되기 위해 데이터를 준비하는 선-처리 모듈이 데이터를 받아들이는 메커니즘(300)을 포함한다. 일례에서, 입력 데이터는 데이터 베이스(310)일 수 있다. 데이터 베이스(310)는 시스템(100)에 의해 처리될 데이터를 포함할 수 있다. 다른 예에서, 데이터 베이스(310)는 시스템(100)에 의해 처리될 데이터의 더 큰 세트(larger set)를 대표하는 데이터의 샘플 세트를 포함할 수 있다. 다른 예에서는, 데이터 베이스는 데이터의 레코드 포맷에 대한 설명을 포함할 수 있다. 다른 예에서는, 입력 데이터가 샘플 데이터와 레코드 포맷의 조합을 포함할 수 있다. 입력 데이터는, 관계된 데이터 베이스, 플랫 파일(flat file), 또는 포트나 또 다른 입력 장치를 통해 수신된 데이터와 같이 레코드 포맷 프로세스(302) 내로 입력을 제공하기 위한 또 다른 메커니즘을 통해 레코드 포맷 프로세스(302)와 통신될 수 있다.

[0026] 레코드 포맷 프로세스(302)는, 입력 데이터(310)를 받아들이고 타겟 레코드 포맷을 판정한다. 일례에서, 입력 데이터는, 각 레코드가 다수의 필드들에 대한 값들을 포함하고 있는 다수의 레코드들을 위해 생성된, 샘플 데이터를 포함한다. 샘플 데이터가 레코드 포맷을 판정하기 위해 분석된다. 다른 예에서는 샘플 데이터가 제공된 레코드 포맷과 비교된다. 다른 예에서는, 샘플 데이터가, 최상의 피트(fit)를 판정하기 위해, 레코드 포맷 집적소(304) 내의 기존의 레코드 포맷들과 비교된다.

[0027] 일례에서, 레코드 포맷 프로세스(302)는, 타겟 레코드 포맷을 판정함에 있어서 어떠한 기존의 파서가 입력 데이터(310)의 파싱이 가능한지 여부를 판정하기 위해 파서들을 포함하는 파서 카탈로그(306)를 검사한다. 입력 데이터(310)를 처리하기 위한 어떠한 파서도 존재하지 않는다면, 레코드 포맷 프로세스(302)는 타겟 레코드 포맷을 판정하기 위해 새로운 파서들의 건조(construction)를 가능하게 하는 커스텀 파서 빌더 모듈(custom parser builder module)(308)을 액세스(access)할 수 있다.

[0028] 사용자는 레코드 포맷에 의해 표시되고 레코드 포맷을 조정하는 것이 허용될 수 있다. 조정된 레코드 포맷은, 레코드 포맷이 샘플 데이터와 양립할 수 있는 것으로 남아있음을 보증하도록, 샘플 데이터에 대하여 체크(check)될 수 있다.

[0029] 도 4를 참조하면, 일 실시예에서는, 시스템이 몇몇의 샘플 레코드들을 포함하는 샘플 데이터를 받아들인다. 선-처리 모듈(106)은 데이터의 레코드 포맷의 식별을 시도한다. 일 실시예에서는, 기존의 저장된 레코드 포맷에

매치되는 것이 없다면, 어떻게 인코드되어야(encoded) 하는지를 판정하기 위해 데이터가 분석된다. 예를 들어, 데이터는 ASCII 또는 EBCDIC 특성 인코딩에 기초하거나 바이너리 포맷에 의해 인코드될 수 있다. 일 실시예에서는, 시스템이 데이터의 파싱에 이용 가능한 파서를 가지고 있는지 여부를 판정할 수 있다. 시스템은 샘플 데이터에 대한 레코드 포맷을 판정하기 위해 샘플 데이터를 검사할 수 있다. 예를 들어, 샘플 데이터에 기초한 텍스트(text)가 디리미터가 판정된 필드들 및 레코드들, 고정된 길이 필드들, Extensible Markup Language(XML) 또는 Standard Generalized Markup Language(SGML)와 같은 태그된 데이터를 이용하여 포맷될 수 있다. 데이터는 레코드 포맷의 판정을 지원하기 위해, 태그들이나 디리미터들이 없는 바이너리 폼일 수도 있다. 바이너리 데이터는, 데이터 베이스, 스프래드 시트, 워드 프로세싱 문서, 이미지 또는 다른 바이너리 데이터일 수 있다. 일 실시예에서는, 바이너리 데이터의 데이터 타입이, 데이터의 자체 검사에 기초하여 얻어질 수 있다. 다른 실시예에서는, 바이너리 데이터의 데이터 타입이, 파일 확장자와 같은 파일의 이름에 기초하여 추정될 수도 있다. 시스템은, 샘플 포맷의 파싱에 기초하여 필드들과 레코드들을 판정할 수 있다. 예를 들어, 시스템이 디리미터가 판정된 필드들 및 레코드들을 인지한다면, 시스템은 디리미터들에 기초하여 필드들 및 레코드들 내의 데이터를 분리한다. 시스템이 태그된 데이터를 인지한다면, 시스템은 태그들에 기초하여 파일을 파싱한다.

[0030] 일례에서, 도 4를 참조하면, 시스템은 샘플 데이터 파일(402)을 수신한다. 본 예시에서의 데이터는 ASCII 텍스트를 이용하여 인코드되고, 상이한 레코드들을 분리하고 돌아온 운반체(carriage)를 가지는 콤마 독립 필드들(comma separated fields)을 이용하여 조직된다.

[0031] 프로세스 화살표(404)에 의해, 시스템이 샘플 데이터에 대한 레코드 포맷(406)을 판정하기 위해 샘플 데이터의 다수의 레코드들을 분석하는 것이, 표현된다. 본 예시에서, 시스템은 5개의 필드들을 식별한다: 스트링(String), 스트링, 룩업 값(Lookup value), 전화 번호(Phone Number), 및 날짜(Date). 인티저, 플로팅 포인트(floating point) 번호, 고정된 길이 텍스트 필드들, 및 고정된 길이 10진법 번호들과 같이, 다른 데이터 타입들이 검출되고 식별될 수도 있다. 일 실시예에서는, 룩업 필드들에 이용 가능한 값들이, 샘플 데이터에 의해 제공된 값들을 프로파일링(profiling)하는 것에 의해 식별될 수 있다. 일 실시예에서는, 샘플 데이터의 레코드 포맷이 얻어지면, 시스템은 각각의 데이터 필드들에 대한 값들을 판정하기 위해 샘플 데이터를 파싱할 수 있다. 본 정보는, 예를 들면, 유효 값들 중에서 상대적으로 작은 번호를 포함할 뿐인 필드들을 식별하기 위해 이용될 수 있다. 일 실시예에서는, 샘플 데이터의 레코드 포맷이, 데이터의 발견적 학습(heuristics)의 분석에 기초하여 판정될 수 있다. 예를 들어, 고정된 길이 레코드들의 세트의 길이는, 레코드들의 수에 의해 평등하게 나누어질 것이다.

[0032] 일 실시예에서, 샘플 데이터에 대한 레코드 포맷이 판정되면, 레코드 포맷은 데이터와 관련된다. 또 다른 실시예에서, 레코드 포맷이 사용자에게 표시될 수 있고, 레코드 포맷을 수정하도록 허용될 수 있다. 프로세스 화살표(414)에 의해, 수정된 레코드 포맷이 샘플 데이터에 대해서 여전히 샘플 데이터와 양립할 수 있음을 확정하기 위해 테스트되는 것이 표현된다. 사용자가 샘플 데이터를 파싱하는 것을 불가능하게 하는 레코드 포맷을 야기하는 데이터 타입을 입력할 때, 세스템은 사용자에게 오류들을 표시하고, 위 문제를 바로 잡는 레코드 포맷으로 변경할 수 있다. 본 실시예에서는, 레코드 포맷들이 마무리되면, 레코드 포맷이 데이터와 관련된다.

[0033] 도 5를 참조하면, 일 실시예에서, 시스템은, 사용자에 의해 제공될 수 있었거나 본 명세서에 설명된 검색 기술들을 이용하여 식별될 수 있었던, 가능 레코드 포맷(possible record format)(504)를 샘플 데이터(502)에 따라 수신한다. 여기에는, 가능 레코드 포맷이 정확하게 샘플 데이터(502) 내의 레코드들의 포맷을 설명하는지에 대한 어느 정도의 불확정성이 있을 수 있다. 가능 레코드 포맷(504)은, XML 문서 타입 정의(XML document type definition) 또는 마스터로부터 복사되고, COBOL 카피북(copybook)과 데이터 조작 처리 언어(Data Manipulation Language, DML) 레코드 포맷과 같은 몇몇 상이한 프로그램들에 삽입될 수 있는 프로그램 데이터의 물리적 레이아웃(layout)을 정의하는 코드의 섹션이 될 수 있다.

[0034] 프로세스 화살표(508)에 의해, 시스템이 처리 도중에 발생한 어떠한 오류들을 기록한 가능 레코드 포맷(504)을 이용하여 샘플 데이터의 파싱을 시도하는 것이 표현된다. 본 예시에서, 제1 필드는 가능 레코드 포맷 내에서 숫자로서 정의되는 반면에, 샘플 데이터(502) 내의 제1 필드는 가변의 길이 특성 필드(variable length character field)이다. 시스템이 레코드 포맷을 이용하여 데이터의 파싱을 시도할 때, 오류 로그(506)가 생성되고 사용자에게 표시된다. 사용자는, 충돌의 해결책(suggestions for resolving the conflict)과 함께 제공된다. 예를 들어, 사용자는 필드 1의 데이터 타입을 가변의 길이 특성 필드로 변경하기 위한 제안과 함께 표시될 수 있다.

[0035] 도 6을 참조하면, 일 실시예에서, 시스템은 샘플 데이터(602)를 수신하고, 기존의 레코드 포맷이 시스템이 데이터를 처리할 수 있게끔 데이터 내의 레코드들의 포맷을 정확하게 설명할 수 있는지 여부를 판정할 것이 요구된다. 시스템은 샘플 데이터가 레코드 포맷 집적소(604) 내의 어떠한 후보 레코드 포맷들(606a-g)에 매치하는지 여부를 판정하도록 샘플 데이터 내의 다수의 레코드들을 분석할 수 있다. 일 실시예에서, 이러한 분석은 레코드 포맷 집적소(604) 내에 저장된 각각의 후보 레코드 포맷들(606a-g)을 이용하여 샘플 데이터를 파싱하기 위한 시도를 포함할 수 있다. 일 실시예에서, 데이터를 파싱하는 것은 각 레코드 내의 각 필드의 샘플 값들을 판정하도록 샘플 데이터에 후보 레코드 포맷을 적용하는 것을 포함한다. 샘플 값들은, 샘플 값들이 그 후보 레코드 포맷과 일관되는지 여부를 판정하도록 후보 레코드 포맷과 비교될 수 있다. 일 실시예에서, 분석은, 후보 레코드 포맷에 의해 필드에 대해 수립된 유효 값들 또는 유효 값들의 범위를 정의하는 유효성 테스트에 대하여 샘플 데이터 내의 값들의 유효화를 포함할 수 있다. 예를 들어, 필드는 한정된 숫자의 유효 값들(50 states, 2 genders, 등)을 허용할 수 있다.

[0036] 각 레코드 포맷에 대해 시스템은, 유효성 테스트로 불리는 파싱의 성공 척도를 판정한다. 예를 들어, 하나의 예시적인 유효성 테스트에서, 시스템은 성공적으로 파싱되지 않은 레코드들의 숫자에 대한 카운트를 보관한다. 또 다른 예시적인 유효성 테스트에서, 시스템은 처리될 수 없었던 필드들의 표시뿐만 아니라 성공적으로 파싱되지 않은 필드들의 숫자에 대한 카운트를 보관한다. 시스템은 후보 레코드 포맷들(606e, 606f, 606g)의 세트로 레코드 포맷들을 좁히고, 이를 사용자에게 표시한다. 일 실시예에서, 레코드 포맷은 샘플 데이터와 관련된 레코드 포맷에 확실한 매치를 제공하지 않을 수 있다. 예를 들어 후보 레코드 포맷(606e)이 스트링 필드로 끝나는 반면에, 다른 후보 레코드 포맷들은 데이터 필드로 끝난다; 그러나, 스트링은 날짜 값들과 함께일 수 있기 때문에, 레코드 포맷은 여전히 샘플 데이터와 양립할 수 있다. 다른 파싱 불일치들도 허용될 수 있다. 예를 들어, 하나의 테스트에 대해, 미리 정해진 유효 값들의 범위 밖의 값들은 후보 레코드 포맷을 여전히 생성할 수 있는데, 예를 들어, 잠재 레코드 포맷(606g)은 유효 값 “M”과 “S”를 가지는 “군복무 상태(marital status)” 필드를 포함한다. 샘플 데이터 세트는 “M” 또는 “F” 중 어느 하나를 포함하는 필드를 포함한다. 시스템은, 파싱 오류를 기록하는 동안에 잠재 데이터 레코드(potential data record)(606g)를 포함할 수 있다. 일 테스트에서는, 파싱 오류들의 숫자가 주어진 임계치 이하이면, 잠재 데이터 레코드가 포함된다. 다른 테스트에서는, 유효한 파싱된 값들의 숫자가 주어진 임계치를 초과하면, 잠재 데이터 레코드가 포함된다.

[0037] 일 실시예에서, 시스템은 사용자에게 후보 레코드 포맷들을 표시하고, 사용자가 데이터에 맞는 레코드 포맷을 선택하도록 허용할 수 있다. 본 예시에서, 사용자는 최적의 피트로서 후보 레코드 포맷(606f)을 선택할 수 있다. 일 실시예에서, 시스템은 양립 가능한 레코드 포맷들을 검사하고, 샘플 데이터와 후보 레코드 포맷의 프로파일(profile)에 기초하여 어떤 레코드 포맷이 최적인지를 판정할 수 있다. 일 실시예에서, 사용자는 레코드 포맷을 수정할 수 있다. 잠재 레코드 포맷들의 리스트가 단일 타겟 레코드 포맷으로 좁혀지면, 시스템은 제공된 샘플 데이터(602)를 파싱하는 것에 의해 선택된 타겟 레코드 포맷을 유효화한다. 유효화가 완료된 이후에, 시스템은 샘플 데이터를 선택된 타겟 레코드 포맷과 관련시키고 선택된 타겟 레코드 포맷을 저장 및/또는 사용자에게 선택된 타겟 레코드 포맷을 제공한다. 일 실시예에서는, 샘플 데이터가 레코드 포맷 내에 제공된 데이터 타입들에 일치하지 않은 때에, 사용자는 이를 샘플 데이터에 일치되도록 레코드 포맷을 수정하기 위한 옵션(option)과 함께 표시될 수 있다.

[0038] 일 실시예에서, 도 7을 참조하면, 시스템은, 샘플 데이터에 맞는 레코드 포맷 집적소(604) 내의 기존의 레코드 포맷을 식별하지 못할 수 있다. 이러한 조건 아래에서, 시스템은 기존의 파서가 제공된 샘플 데이터를 파싱할 수 있는지 여부를 판정한다. 예를 들어, 샘플 데이터 세트(702)가 XML 포맷으로 보여진다. 본 예시에서, 레코드 포맷 집적소(604)는 샘플 데이터에 매치하는 어떠한 레코드 포맷도 포함하지 않는다. 프로세스 화살표(704)에 의해, 시스템이 레코드 포맷이 XML 포맷의 ASCII 파일이라는 것을 식별하는 것이 표현된다. 프로세스 화살표(708)에 의해, 시스템이 기존의 파서(예컨대, XML 파서(710))가 데이터를 해독하는 것이 가능한지를 판정하는 것이 표현된다. 파서와 샘플 데이터에 기초하여, 시스템은 샘플 데이터(714)의 레코드 포맷을 이끌어낸다. 상기에서 논의된 바와 같이, 시스템은 파서가 샘플 데이터를 해독할 수 있음을 검증(verify)하고, 파서에 의해 생성된 결과 타겟 레코드 포맷(resulting target record format)을 샘플 데이터(714)와 관련시키며, 레코드 포맷 집적소 내에 결과 타겟 레코드 포맷을 저장한다. 일 실시예에서, 시스템은 레코드 포맷 집적소 내에 타겟 레코드 포맷이 저장되기 전에 허가를 위해 사용자에게 새로이 생성된 타겟 레코드 포맷을 표시한다.

[0039] 도 8은, 선-처리 모듈(106)이 타겟 레코드 포맷을 판정하기 위해 이용할 수 있는 또 다른 예시적인 프로세스(800)에 대한 플로우 차트를 보여준다. 선-처리 모듈의 동작들은 공급된 입력 데이터가 샘플 데이터를 포함하는지 여부를 판정(802)하는 것을 포함한다.

[0040] 동작들은, 입력 데이터가 샘플 데이터를 포함하고 있다면, 샘플 데이터를 업로드 및/또는 배치(804)하는 것도 포함한다. 선-처리 모듈은 입력 데이터에 의해 정의된 위치로부터 샘플을 액세스할 수 있다. 일 실시예에서는, 선-처리 모듈은 액세스 포트를 통해 또 다른 서버로부터 샘플 데이터를 업로드 또는 액세스할 수 있다. 또 다른 실시예에서는, 선-처리 모듈은 샘플 데이터를 포함하는 파일 또는 다른 데이터 스토리지 메커니즘을 액세스할 수 있다.

[0041] 동작들은, 샘플 데이터의 분석(806)도 포함하며, 선택적으로 분석의 결과들을 저장하는 것도 포함한다. 샘플 데이터는, 캐릭터 세트(character set), 메타 데이터(metadata), 레코드 포맷 타입 및/또는 레코드 포맷 자체를 판정하기 위해 분석될 수 있다. 일 실시예에서는, 시스템은 레코드 포맷 집적소 내에 저장된 하나 이상의 알려진 레코드 포맷들에 대한 검색을 수행하는지 여부를 판정하기 위해, 샘플 데이터를 분석한다. 예를 들어, 선-처리 모듈은 샘플 데이터가 제1 타입(예컨대, 콤마 독립 파일)이면서 샘플 데이터가 제2 타입(예컨대, XML)으로 판정되지 않을 때, 잠재 레코드 포맷을 판정하기 위해 검색을 수행할 수 있다. 다른 실시예에서는, 샘플 데이터는, 레코드 포맷의 생성과 유효화를 도울 수 있는 메타 데이터를 찾기 위해 분석된다. 일 실시예에서는, 선-처리 모듈은 필드 세퍼레이터들(field separators), 탈출 캐릭터들(escape characters), 및 파일 명들을 포함하는 헤더를 식별한다. 분석의 결과들은 이후의 판정 프로세스에서 이용되도록 계속 유지된다.

[0042] 본 실시예에서는, 동작들은 샘플 데이터를 포함하는 문서의 타입이 XML인지 여부를 판정(808)하는 것도 포함한다. 일 실시예에서는, 본 예시에서의 XML 포맷과 같은 하나 이상의 미리 정해진 포맷들의 문서들이, 다른 포맷의 문서들로부터 독립되어 취급된다. 본 실시예에서는, 샘플 XML 문서들은 XML 파서에 의해 처리된다(826).

[0043] 동작들은, 샘플 데이터가 레코드 포맷 집적소 내에 저장된 하나 이상의 알려진 레코드 포맷들에 매치하는지 여부를 판정(810)하는 것도 포함한다. 이는, 상기에서 논의된 것처럼, 유효성 테스트들을 이용하여 각 레코드 포맷에 대하여 샘플 데이터 내의 하나 이상의 레코드를 유효화하고 유효화 오류들의 숫자를 판정하는 것에 의해 성취될 수 있다. 다른 실시예에서는, 샘플 데이터을 분석(806)하는 동안 얻어진 정보들이, 어떤 샘플 데이터가 유효화되는 것에 대해 데이터 포맷들의 숫자를 감소하도록 이용될 수 있다.

[0044] 동작들은 매칭 레코드 포맷들(matching record formats)을 사용자에게 보여주는 것(812)도 포함한다. 상기에서 논의된 것처럼, 선-처리 모듈은 잠재적인 매칭 레코드 포맷들의 리스트를 사용자에게 표시할 수 있다.

[0045] 동작들은 사용자가 잠재적인 매칭 레코드 포맷들의 리스트로부터 매칭 레코드 포맷을 선택하는지 여부를 판정하는 것(814)도 포함한다.

[0046] 동작들은, 저장된 레코드 포맷에 대해 어떠한 매치도 사용자에 의해 발견 및/또는 선택되지 않았다면, 샘플 데이터가 파서가 이용 가능한 알려진 고유의 포맷을 가지는 파일 내에 포함되어 있는 것과 같은 알려진 데이터 타입들을 가지는지 여부를 판정(816)하는 것도 포함한다. 고유의 포맷은 어플리케이션 또는 시스템에 의해 이용되는 외부의 알려진 포맷이다.

[0047] 동작들은, 고유의 포맷이 알려진 것이라면, 적절한 이용 가능한 파서에 데이터 매치를 판정하는 것도 포함한다. 예를 들어, 샘플 데이터는 알려진 파서에 의해 처리될 수 있는 태그된 레코드들을 포함할 수 있다(820).

[0048] 동작들은, 태그된 샘플 데이터에 대한 파서를 식별하는 것(830)도 포함한다.

[0049] 본 실시예에서, 이용 가능한 파서에 대한 매치를 판정하는 것은, 샘플 데이터가 COBOL 내에 있는지 여부를 판정하는 것(822)을 포함한다. 일 실시예에서는, 동작들은 샘플 데이터가 이용 가능한 파서에 의해 파싱될 수 있는 표준 데이터 레코드 포맷 구조(standard data record format structure)를 활용하는 또 다른 프로그래밍 언어인지 여부를 판정하는 것도 포함할 수 있다.

[0050] 동작들은, 샘플 데이터가 COBOL 내에 있다면, COBOL 카피북을 업로드하고 파싱하는 것(832)도 포함한다.

[0051] 또 다른 이용 가능한 파서에 알려진 고유의 포맷을 매치하는 것은, 샘플 데이터가 데이터 베이스 내에 저장되는 것을 판정하고 선-처리 모듈이 데이터 베이스에 액세스할 수 있음을 인증하는 것(824)을 포함한다. 데이터 베이스에 대한 액세스는, 선-처리 모듈이 유효 증명서, 예를 들어 사용자 이름 및 비밀 번호에 액세스하는 것에 대한 검증을 포함할 수 있다. 데이터 베이스에 대한 액세스는, 증명서가 샘플 데이터에 대한 액세스를 제공하는 것을 판정하는 것도 포함한다.

[0052] 동작들은, 데이터 베이스 내에 저장된 샘플 데이터를 분석하고 분석으로부터 레코드 포맷을 판정하는 것(834)도 포함한다(예컨대, SQL 에디터 내에서). 일 실시예에서는, 선-처리 모듈이 레코드 포맷을 이끌어내도록 데이터

베이스의 테이블 구조를 분석한다.

[0053] 또 다른 이용 가능한 파서에 알려진 고유의 포맷을 매치하는 것은, 샘플 데이터가 XML 포맷인지 및 샘플 데이터가 문서 타입 정의 또는 XML 개요 정의(XML Schema Definition, XSD)를 포함하는지 여부를 판정하는 것(826)을 포함한다.

[0054] 동작들은 레코드 포맷 내로 XML 문서의 구조를 번역하는 것(836)을 포함한다(예컨대, XML 경로 에디터 내에서).

[0055] 또 다른 이용 가능한 파서에 알려진 고유의 포맷을 매치하는 것은, 데이터가 SAP 포맷인지 여부를 판정하는 것(828)도 포함한다. 일 실시예에서는, 다른 엔터프라이즈 솔루션 소프트웨어 패키지들(enterprise solution software packages)이 검출될 수 있는데, 예를 들면, 오라클 파이낸셜(Oracle Financials)의 샘플 데이터이다.

[0056] 동작들은 기업 소프트웨어 패키지에 대한 임포트 포듈(import module)을 이용하여 레코드 포맷을 판정하는 것(838)도 포함한다.

[0057] 샘플 데이터의 데이터 타입이 알려지지 않았거나 데이터 타입에 대해 이용 가능한 파서가 없다면, 동작들은 샘플 데이터의 특성들을 판정하고 샘플 데이터의 특성들의 분석으로부터 구성된 레코드 포맷을 생성하는 것도 포함한다. 예를 들어, 본 실시예에서는, 동작들은 샘플 데이터가 대부분 태그되었는지 여부를 판정하는 것(840)을 포함한다. 대부분 태그된 데이터는, 예를 들어, 태그된 데이터 구조들을 우세하게 포함하도록 나타나는 데이터이나, 필수적으로 태그된 구조를 따르지 않는 일부 데이터도 포함한다.

[0058] 동작들은, 데이터가 대부분 태그된 것으로 판정되면, 태그된 데이터와 같이 데이터를 처리하기 위해 시도하는 것(842)도 포함한다(예컨대, 태그 에디터를 이용하여). XML에 부가하여, 다른 태그된 포맷들도 다루어질 수 있는데, 예를 들면 전세계 은행 금융 텔레커뮤니케이션 협회 포맷들(Society for Worldwide Interbank Financial Telecommunication formats, SWIFT)이 있다.

[0059] 동작들은, 일반적인 태그된 데이터 파서, 또는 알려진 파서가 샘플 데이터의 처리가 가능한지 여부를 판정하는 것(844)도 포함한다.

[0060] 동작들은 샘플 데이터를 파서 빌더로 보내는 것(848)도 포함한다.

[0061] 동작들은 샘플 데이터가 대부분 텍스트임을 판정하는 것(852)도 포함한다. 대부분의 텍스트 데이터는, 예를 들어, 잘 알려진 텍스트 포맷, 예를 들어 ASCII 또는 EBCDIC를 이용하여 우선적으로 인코딩된 데이터이다.

[0062] 동작들은, 데이터의 구조의 판정을 시도하는 것(854)도 포함한다. 일 실시예에서는, 데이터의 구조가 레코드 및 필드 디리미터들을 식별하는 것에 의해 판정될 수 있다. 레코드 디리미터들은 샘플 데이터 내의 마지막 문자(character)를 검사하는 것에 의해 식별될 수 있다. 디리미터들은 프린트되지 않거나 알파벳순으로 나열되지 않는 문자들에 대한 데이터를 검사하는 것에 의해 식별될 수도 있다. 샘플 데이터 내에서 프린트되지 않는 문자들 또는 알파벳순으로 나열되지 않는 문자들이 발생되는 곳은, 흔하게는 필드 디리미터일 수 있고, 흔하지 않게는 레코드 디리미터일 수 있다. 디리미터가 아닌 프린트되지 않는 문자의 존재는, 샘플 데이터가 디리미터가 판정되지 않았음을 지시할 수 있다. 디리미터들의 식별 이후에, 선-처리 모듈은 샘플 데이터에 디리미터들을 적용할 수 있고, 불일치들을 체크할 수 있다. 예를 들어, 시스템은 각 레코드가 동일한 수의 필드들을 포함하는지를 체크할 수 있다. 시스템은 각 레코드 내의 동일한 필드가 유사 또는 양립 가능한 데이터 타입을 포함하는지를 체크할 수 있다. 일 실시예에서, 선-처리 모듈은 데이터의 분석(806) 동안에 데이터에 대해 판정된 정보에 의존한다.

[0063] 동작들은, 데이터가 대부분 바이너리임을 판정하는 것(856)도 포함한다. 바이너리 데이터는, 예를 들어, 잘 알려진 텍스트 포맷들, 예를 들어 ASCII 및 EBCDIC을 이용하여 인코드되지 않는 데이터이다.

[0064] 동작들은 적절하다면(예컨대, 데이터가 대부분 바이너리임을 판정하는 것(856)에 응하여) 샘플 데이터 내로 필드 이름들을 삽입하는 것(858)도 포함한다. 일 실시예에서는, 사용자는 삽입될 필드 이름들을 입력(예컨대, 붙여넣기 또는 키 입력)할 수 있다.

[0065] 동작들은 결과들을 검증하는 것(850)도 포함한다. 레코드 포맷을 검증하는 것은 레코드 포맷을 이용하고 샘플 데이터를 파싱하도록 시도하는 것을 포함할 수 있다.

[0066] 동작들은 사용자가 레코드 포맷을 구성 또는 에디트하도록 허용하는 것(846)을 포함한다. 일 실시예에서는, 사

용자는 레코드 포맷을 에디트 및/또는 샘플 데이터의 타입, 이름들, 및 구조를 변경할 수 있다.

[0067] 동작들은 레코드 포맷 집적소 내에 레코드 포맷을 저장하는 것(860)도 포함한다. 일 실시예에서는, 선-처리 모듈이 데이터 포맷을 샘플 데이터와 관련시키고, 다른 실시예에서는 선-처리 모듈이 데이터 포맷의 복사본을 생성하고 데이터와 복사본을 관련시킨다.

[0068] 상기 설명된 레코드 포맷 발견 접근(record format discovery approach)은 컴퓨터 상에서의 실행을 위한 소프트웨어를 이용하여 구현될 수 있다. 예를 들어, 소프트웨어는, 각각이 적어도 하나의 프로세서, 적어도 하나의 데이터 저장 시스템(휘발성 및 비휘발성 메모리 및/또는 저장 요소를 포함하는), 적어도 하나의 입력 장치 또는 포트, 및 적어도 하나의 출력 장치 또는 포트를 포함하는 하나 이상의 프로그램된 또는 프로그래머블 컴퓨터 시스템(분배된, 클라이언트/서버 또는 그리드와 같이 다양한 아키텍처일 수 있는) 상에서 실행하는 하나 이상의 컴퓨터 프로그램들 내에 과정들을 형성한다. 소프트웨어는, 예를 들어, 연산 그래프들의 디자인과 구성과 관련된 다른 서비스들을 제공하는 더 큰 프로그램의 하나 이상의 모듈을 형성할 수 있다. 그래프의 노드들 및 엘리먼트들은 컴퓨터 판독가능 매체 내에 저장된 데이터 구조들로서 또는 데이터 집적소 내에 저장된 데이터 모델을 따르는 다른 조직된 데이터로서 구현될 수 있다.

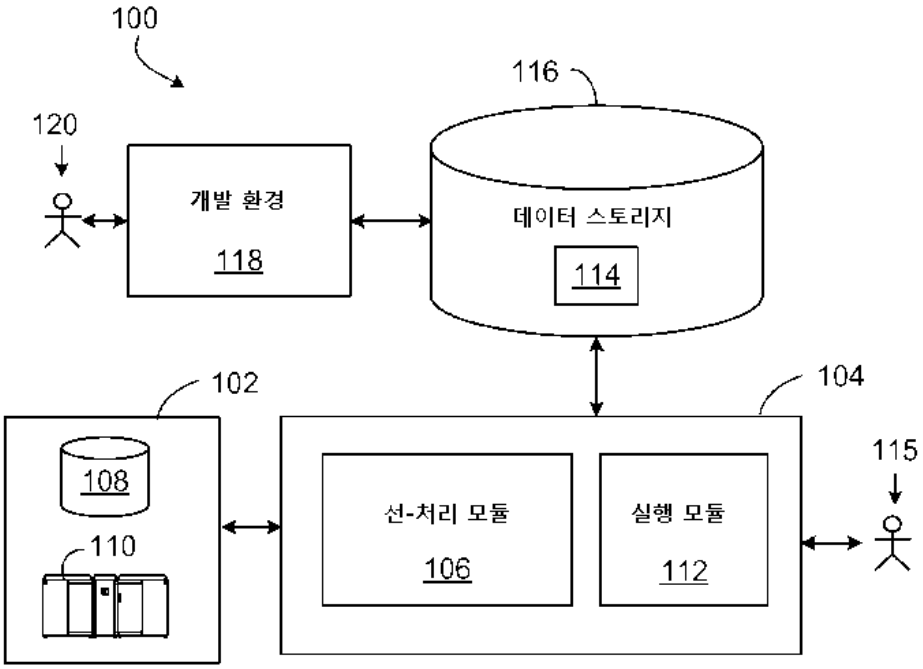
[0069] 소프트웨어는, CD-ROM과 같은, 범용 또는 특수 목적의 프로그래머블 컴퓨터에 의해 판독가능 하거나 실행되는 컴퓨터에 대해 네트워크의 통신 매체를 넘어 전송되는(전파된 신호로 인코딩된) 저장 매체로 제공될 수 있다. 모든 기능들은, 코-프로세서들과 같은, 특수 목적 컴퓨터 또는 특수 목적 하드웨어를 이용하여 수행될 수 있다. 소프트웨어는, 소프트웨어에 의해 지정된 연산의 상이한 부분들이 상이한 컴퓨터들에 의해 수행되는 분배된 방법으로 수행될 수 있다. 각각의 그러한 컴퓨터 프로그램은, 저장 매체 또는 장치가 본 명세서에 설명된 과정들을 수행하기 위해 컴퓨터 시스템에 의해 판독되는 때에 컴퓨터를 구성하고 동작하기 위한, 범용 또는 특수 목적의 프로그래머블 컴퓨터에 의해 판독 가능한 저장 매체 또는 장치(예컨대, 고체 상태 메모리 또는 매체, 또는 자기 또는 광학 매체)에 우선적으로 저장 또는 다운로드된다. 본 발명의 시스템은, 본 명세서에 설명된 기능들을 수행하도록 특정 및 미리 정해진 방법으로 컴퓨터 시스템이 동작하도록 야기시키도록 구성된, 컴퓨터 프로그램과 함께 구성된 컴퓨터 판독가능 저장 매체로서 구현되는 것으로 고려될 수도 있다.

[0070] 본원 발명에 대한 특정 수의 실시예들이 설명되었다. 그럼에도 불구하고, 본원 발명의 핵심 및 범위로부터 벗어남이 없이도 다양한 수정사항들이 가능할 수 있는 것으로 이해될 것이다. 예를 들어, 상기 설명된 단계들 중 일부는 순서에 독립적이고, 그러므로 설명된 순서와 상이한 순서로 수행될 수 있다.

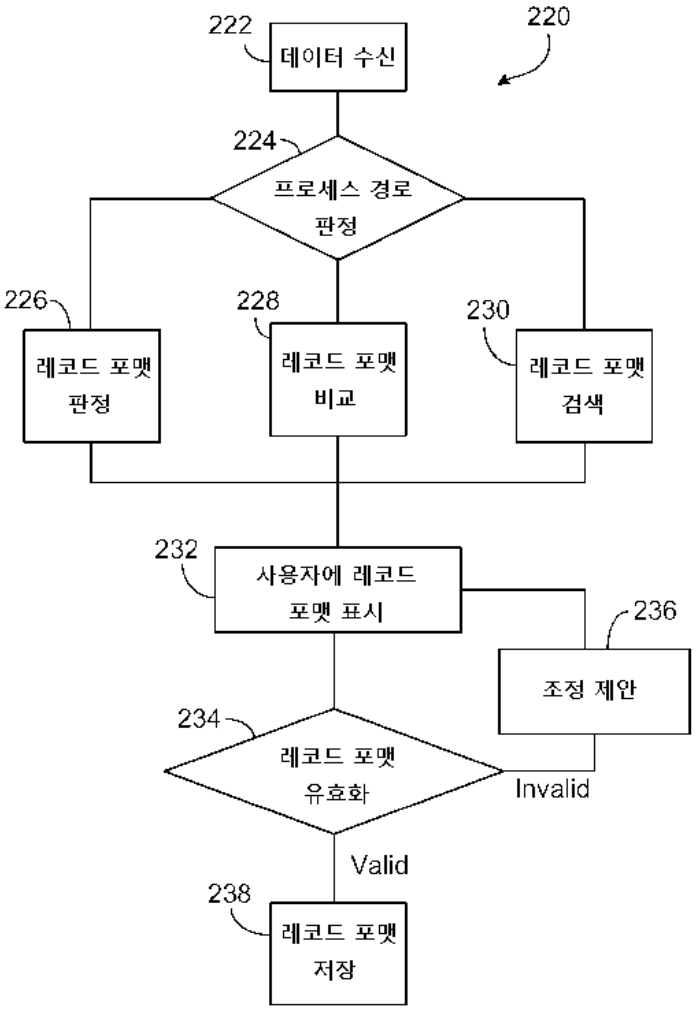
[0071] 선행하는 설명은 본원 발명의 설명을 위해 의도된 것일 뿐이고, 첨부된 청구범위의 범위에 의해 정의되는 본원 발명의 범위를 한정하는 것으로 이해될 수 없다. 예를 들어, 상기에서 설명된 기능의 단계들은 실질적으로 전체 처리에 영향을 미침이 없이도 상이한 순서로 수행될 수 있다. 다른 실시예들은 이하의 청구범위들의 범위에 포함된다.

도면

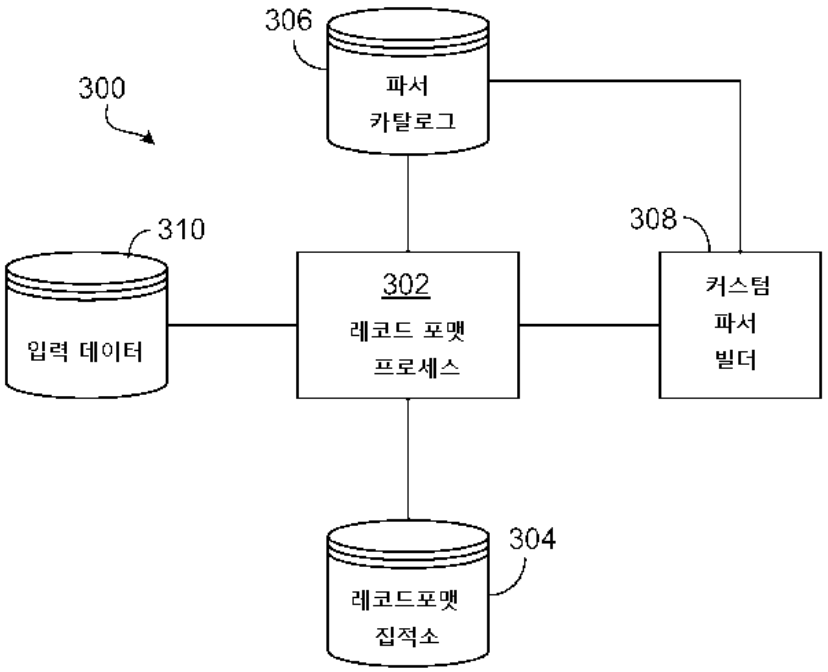
도면1


100
116
120
개발 환경
118
데이터 스토리지
114
102
104
115
108
110
선-처리 모듈
106
실행 모듈
112

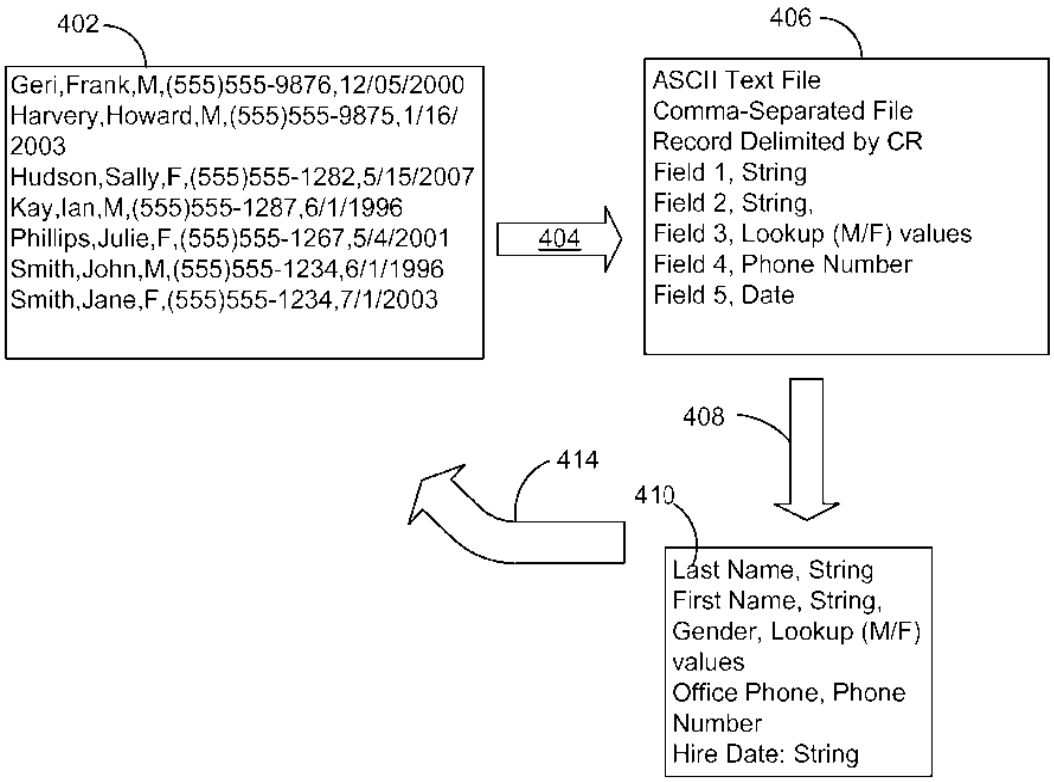
도면2


220
222
데이터 수신
224
프로세스 경로 판정
226
레코드 포맷 판정
228
레코드 포맷 비교
230
레코드 포맷 검색
232
사용자에 레코드 포맷 표시
236
조정 제안
234
레코드 포맷 유효화
Invalid
Valid
238
레코드 포맷 저장

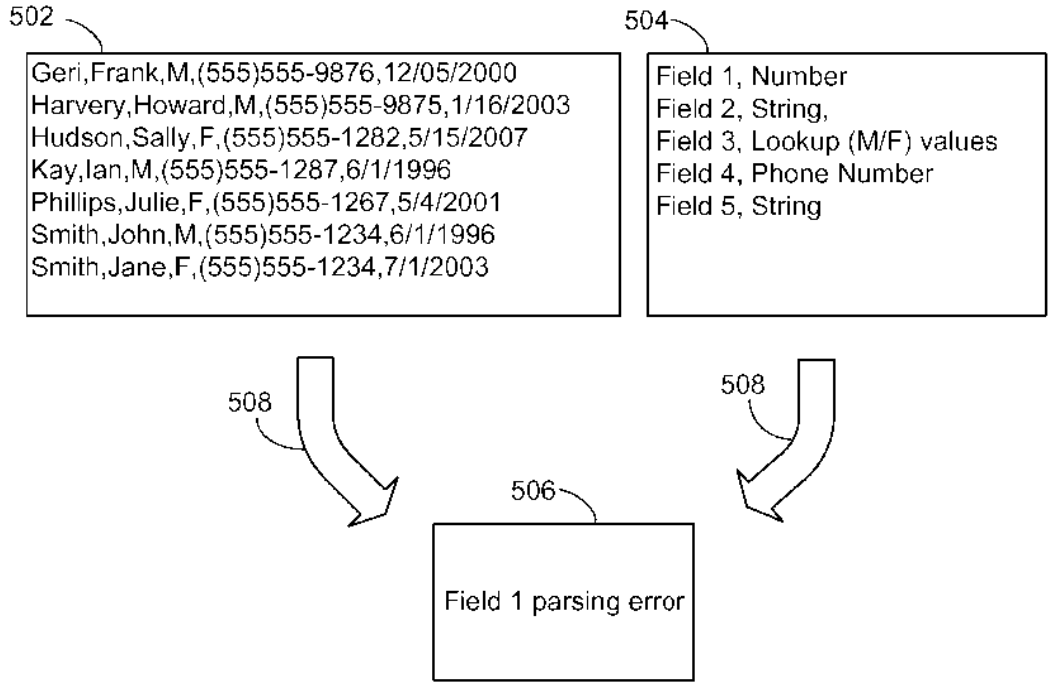
도면3


306
파서
카탈로그
300
310
308
302
레코드 포맷
프로세스
입력 데이터
커스텀
파서
빌더
304
레코드포맷
집적소

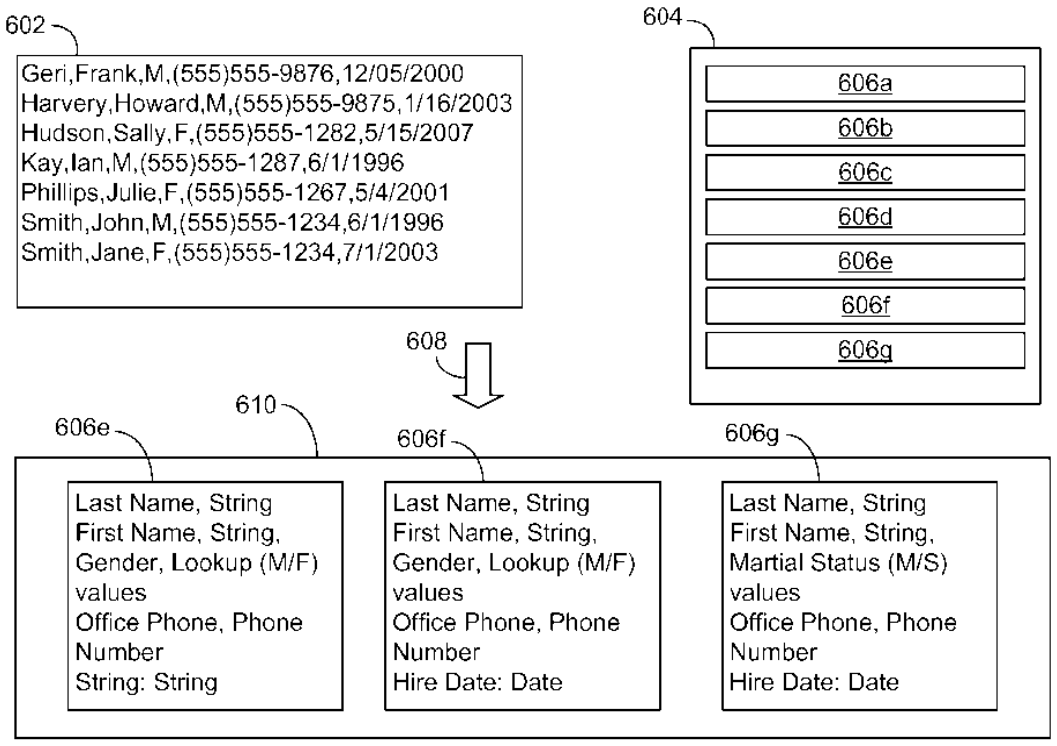
도면4


402
Geri,Frank,M,(555)555-9876,12/05/2000
Harvery,Howard,M,(555)555-9875,1/16/2003
Hudson,Sally,F,(555)555-1282,5/15/2007
Kay,Ian,M,(555)555-1287,6/1/1996
Phillips,Julie,F,(555)555-1267,5/4/2001
Smith,John,M,(555)555-1234,6/1/1996
Smith,Jane,F,(555)555-1234,7/1/2003
404
406
ASCII Text File
Comma-Separated File
Record Delimited by CR
Field 1, String
Field 2, String,
Field 3, Lookup (M/F) values
Field 4, Phone Number
Field 5, Date
408
414
410
Last Name, String
First Name, String,
Gender, Lookup (M/F) values
Office Phone, Phone Number
Hire Date: String

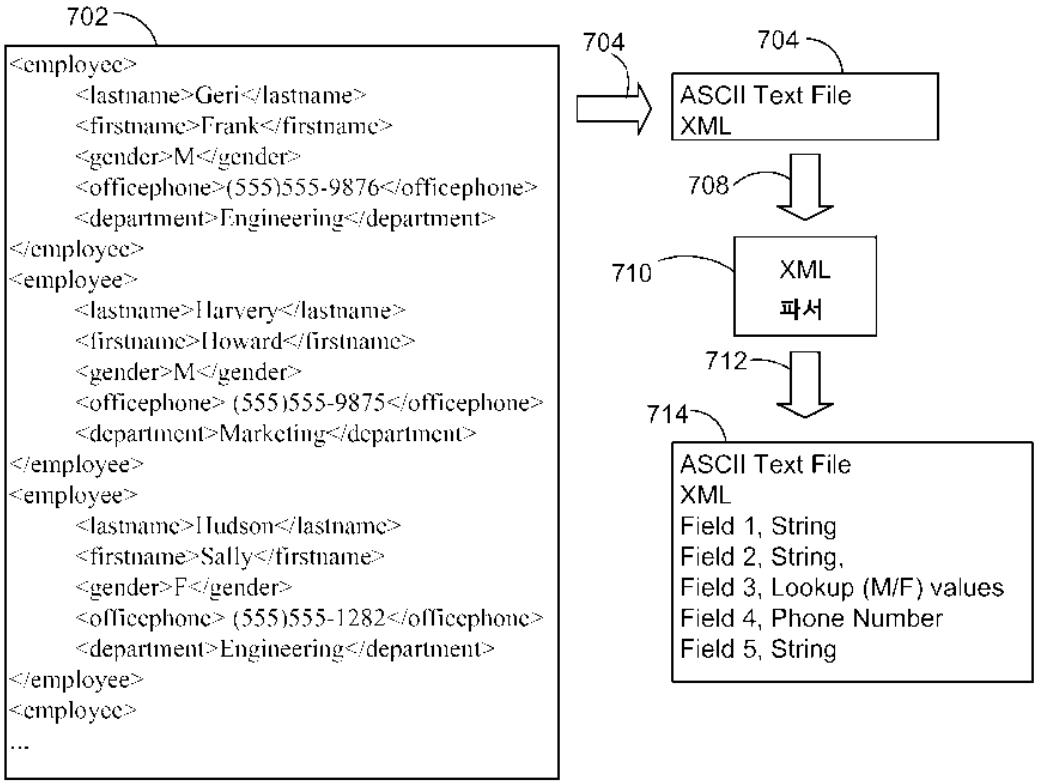
도면5


502
Geri,Frank,M,(555)555-9876,12/05/2000
Harvery,Howard,M,(555)555-9875,1/16/2003
Hudson,Sally,F,(555)555-1282,5/15/2007
Kay,Ian,M,(555)555-1287,6/1/1996
Phillips,Julie,F,(555)555-1267,5/4/2001
Smith,John,M,(555)555-1234,6/1/1996
Smith,Jane,F,(555)555-1234,7/1/2003
504
Field 1, Number
Field 2, String,
Field 3, Lookup (M/F) values
Field 4, Phone Number
Field 5, String
508
508
506
Field 1 parsing error

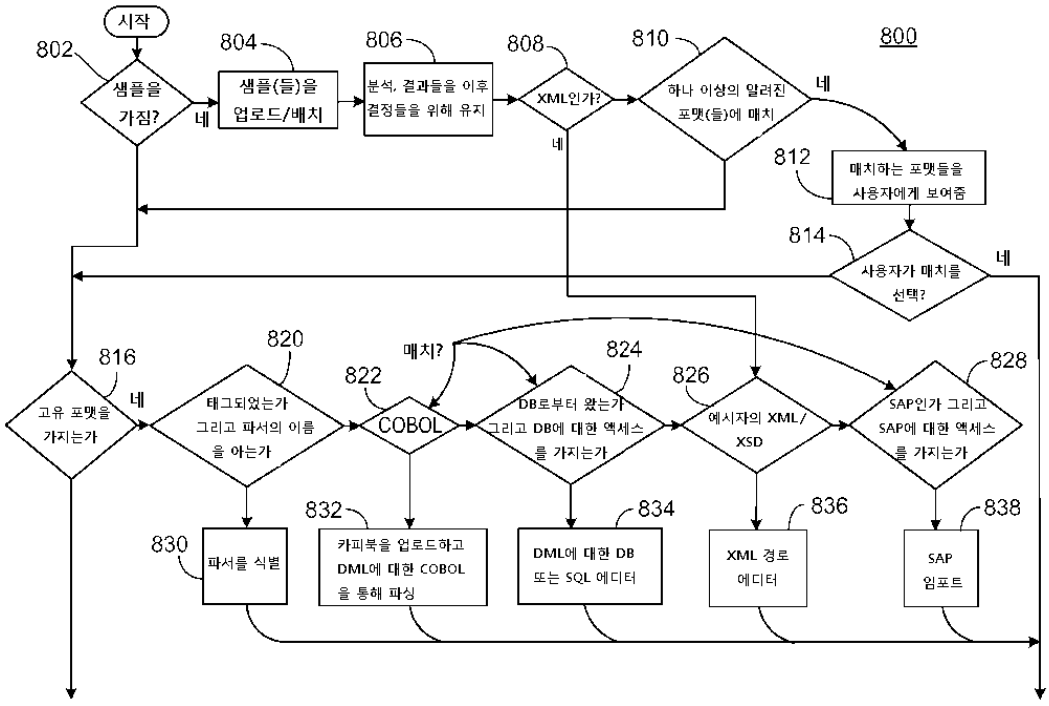
도면6


602
Geri,Frank,M,(555)555-9876,12/05/2000
Harvery,Howard,M,(555)555-9875,1/16/2003
Hudson,Sally,F,(555)555-1282,5/15/2007
Kay,Ian,M,(555)555-1287,6/1/1996
Phillips,Julie,F,(555)555-1267,5/4/2001
Smith,John,M,(555)555-1234,6/1/1996
Smith,Jane,F,(555)555-1234,7/1/2003
604
606a
606b
606c
606d
606e
606f
606g
608
610
606e
606f
606g
Last Name, String
First Name, String,
Gender, Lookup (M/F) values
Office Phone, Phone Number
String: String
Last Name, String
First Name, String,
Gender, Lookup (M/F) values
Office Phone, Phone Number
Hire Date: Date
Last Name, String
First Name, String,
Martial Status (M/S) values
Office Phone, Phone Number
Hire Date: Date

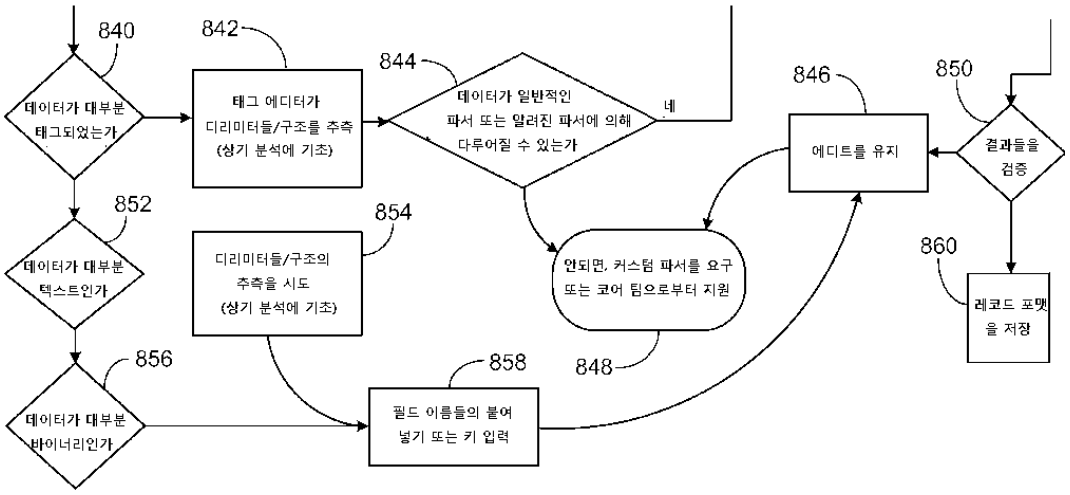
도면7


702
<employee>
<lastname>Geri</lastname>
<firstname>Frank</firstname>
<gender>M</gender>
<officephone>(555)555-9876</officephone>
<department>Engineering</department>
</employee>
<employee>
<lastname>Harvery</lastname>
<firstname>Howard</firstname>
<gender>M</gender>
<officephone> (555)555-9875</officephone>
<department>Marketing</department>
</employee>
<employee>
<lastname>Hudson</lastname>
<firstname>Sally</firstname>
<gender>F</gender>
<officephone> (555)555-1282</officephone>
<department>Engineering</department>
</employee>
<employee>
...
704
704
ASCII Text File
XML
708
710
XML
파서
712
714
ASCII Text File
XML
Field 1, String
Field 2, String,
Field 3, Lookup (M/F) values
Field 4, Phone Number
Field 5, String

도면8a


시작
802
샘플을 가짐?
네
804
샘플(들)을 업로드/배치
806
분석, 결과들을 이후 결정들을 위해 유지
808
XML인가?
네
810
하나 이상의 알려진 포맷(들)에 매치
네
800
812
매치하는 포맷들을 사용자에게 보여줌
814
사용자가 매치를 선택?
네
816
고유 포맷을 가지는가
네
820
태그되었는가 그리고 파서의 이름을 아는가
822
COBOL
매치?
824
DB로부터 왔는가 그리고 DB에 대한 엑세스를 가지는가
826
예시자의 XML/XSD
828
SAP인가 그리고 SAP에 대한 엑세스를 가지는가
830
파서를 식별
832
카피북을 업로드하고 DML에 대한 COBOL을 통해 파싱
834
DML에 대한 DB 또는 SQL 에디터
836
XML 경로 에디터
838
SAP 임포트

도면8b


840
842
844
846
850
852
854
856
858
860
848
데이터가 대부분 태그되었는가
태그 에디터가 디리미터들/구조를 추측 (상기 분석에 기초)
데이터가 일반적인 파서 또는 알려진 파서에 의해 다루어질 수 있는가
네
에디트를 유지
결과들을 검증
데이터가 대부분 텍스트인가
디리미터들/구조의 추측을 시도 (상기 분석에 기초)
안되면, 커스텀 파서를 요구 또는 코어 팀으로부터 지원
레코드 포맷을 저장
데이터가 대부분 바이너리인가
필드 이름들의 붙여 넣기 또는 키 입력