

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization

International Bureau



(10) International Publication Number

WO 2014/186057 A1

(43) International Publication Date
20 November 2014 (20.11.2014)

WIPO | PCT

(51) International Patent Classification:
G06F 19/00 (2011.01)

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(21) International Application Number:
PCT/US2014/031919

(22) International Filing Date:
26 March 2014 (26.03.2014)

(25) Filing Language:
English

(26) Publication Language:
English

(30) Priority Data:
61/824,544 17 May 2013 (17.05.2013) US
61/843,203 5 July 2013 (05.07.2013) US
14/044,429 2 October 2013 (02.10.2013) US

(71) Applicant: ORACLE INTERNATIONAL CORPORATION [US/US]; 500 Oracle Parkway, Redwood Shores, California 94065 (US).

(72) Inventors: ALLAN, David; 1112 Blackfield Way, Mountain View, California 94040 (US). LAU, Kwok-hung (Thomas); 1070 Mercedes Avenue, #25, Los Altos, California 94022 (US). GONG, Yu (Jeff); 4 Muirfield Road, Half Moon Bay, California 94065 (US).

(74) Agents: PARMENTER, Sean et al.; Kilpatrick Townsend & Stockton LLP, Eighth Floor, Two Embarcadero Center, San Francisco, California 94111 (US).

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) Title: SUPPORTING COMBINATION OF FLOW BASED ETL AND ENTITY RELATIONSHIP BASED ETL

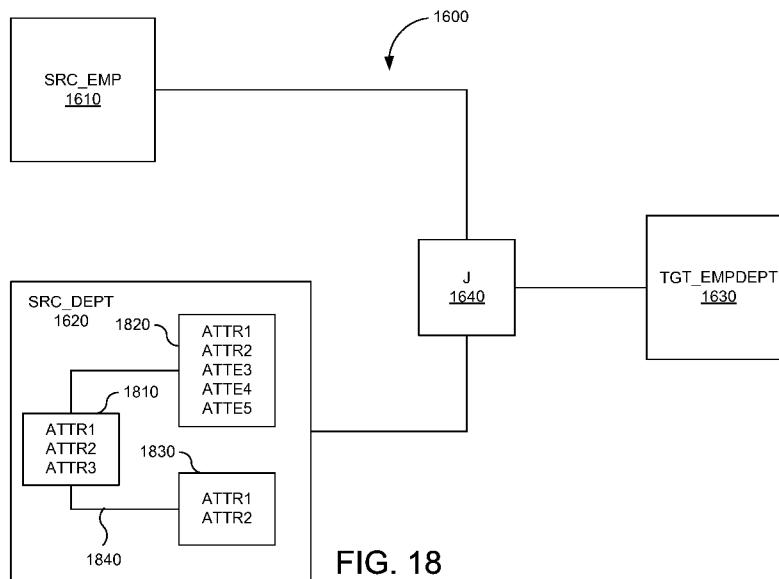


FIG. 18

(57) Abstract: A data integration system is disclosed that incorporates one or more techniques for eases the design and maintenance of a mapping. As components are added to an existing design, the data integration system removes the need to specify all input and output attributes. In one aspect, components types are implement that allow entity relationships to be added and/or edited in a flow view of a logical design. Therefore, attributes of components representing datasets can be added and propagated to downstream components with minimal effort on the part of a map designer.

SUPPORTING COMBINATION OF FLOW BASED ETL AND ENTITY RELATIONSHIP BASED ETL

BACKGROUND OF THE INVENTION

[0001] In today's increasingly fast-paced business environment, organizations need to use more specialized software applications. Additionally, organizations need to ensure the coexistence of these applications on heterogeneous hardware platforms and systems and 5 guarantee the ability to share data between applications and systems.

[0002] Accordingly, what is desired is to solve problems relating to developing data integration scenarios, some of which may be discussed herein. Additionally, what is desired is to reduce drawbacks relating to developing data integration scenarios, some of which may be discussed herein.

10

BRIEF SUMMARY OF THE INVENTION

[0003] The following portion of this disclosure presents a simplified summary of one or more innovations, embodiments, and/or examples found within this disclosure for at least the purpose of providing a basic understanding of the subject matter. This summary does not attempt to 15 provide an extensive overview of any particular embodiment or example. Additionally, this summary is not intended to identify key/critical elements of an embodiment or example or to delineate the scope of the subject matter of this disclosure. Accordingly, one purpose of this summary may be to present some innovations, embodiments, and/or examples found within this disclosure in a simplified form as a prelude to a more detailed description presented later.

20 [0004] In various embodiments, a data integration system enables users to create a logical design which is platform and technology independent. The user can create a logical design that defines, at a high level, how a user wants data to flow between sources and targets. The tool can analyze the logical design, in view of the user's infrastructure, and create a physical design. The logical design can include a plurality of components corresponding to each source and target in

the design, as well as operations such as joins or filters, and access points. Each component when transferred to the physical design generates code to perform operations on the data. Depending on the underlying technology (e.g., SQL Server, Oracle, Hadoop, etc.) and the language used (SQL, pig, etc.) the code generated by each component may be different.

5 **[0005]** In one aspect, a user of data integration system is not required to specify all data attributes at each component in the logical design, from start to end. The data integration system provides a plurality of component types, such as projector and selector types, that avoid the need to fully declare the information that flows through the logical design. The data integration system is able to decide what attributes are needed at operations represented by predetermined component types. This simplifies both the design and maintenance. In various aspects, data transformation and migration is provided that leverages existing RDBMS resources and capabilities to avoid the need for a separate proprietary ETL server to achieve improved performance.

10 **[0006]** In one embodiment, a method facilitating generation of a data mapping includes receiving information specifying a set of entity relationships as a component of the logical design. An equivalent data flow model is determined based on the set of entity relationships. Information is then generated indicative of the equivalent data flow model in the logical flow design. One or more attributes of a dataset representing the set of entity relationships may be derived based on information declaring relationships between attributes of data sources.

15 **[0007]** In further embodiments, information may be received specifying one or more components of the logical design that includes information indicative of an operation that changes shape of the information flowing through the logical design. Information may be received specifying one or more components of the logical design that includes information indicative of an operation that controls the flow of information flowing through the logical design but does not change shape of the information flowing through the logical design.

20 Information may be received specifying one or more components of the logical design that includes information indicative of a target component having one or more attributes of data to be stored in a target datastore.

25 **[0008]** In one aspect, generating the information indicative of the equivalent data flow model in the logical flow design may include exporting a list of attributes to a downstream component.

In another, a change in the logical design may be received through the introduction of one or more relationships. An updated equivalent data flow model may then be determined.

[0009] In one embodiment, a non-transitory computer-readable medium storing computer-executable code for facilitating generation of a data mapping includes code for receiving

5 information specifying a set of entity relationships as a component of the logical design, code for determining an equivalent data flow model based on the set of entity relationships, and code for generating information indicative of the equivalent data flow model in the logical flow design.

[0010] In a further embodiment, a system facilitating generation of a data mapping includes a processor and a memory storing instructions which when executed by the processor configure

10 the processor to receive information specifying a set of entity relationships as a component of the logical design, determine an equivalent data flow model based on the set of entity relationships, and generate information indicative of the equivalent data flow model in the logical flow design.

[0011] In one embodiment, a system facilitating generation of a data mapping includes a receiving unit configured to receive information specifying a set of entity relationships as a

15 component of the logical design; a determining unit configured to determine an equivalent data flow model based on the set of entity relationships; and a generating unit configured to generate information indicative of the equivalent data flow model in the logical flow design.

[0012] In one aspect, the system further may include a deriving unit configured to derive one or more attributes of a dataset representing the set of entity relationships based on information

20 declaring relationships between attributes of data sources. In one aspect, the receiving unit is further configured to receive information specifying one or more components of the logical design that includes information indicative of an operation that changes shape of the information flowing through the logical design. In one aspect, the receiving unit is further configured to receive information specifying one or more components of the logical design that includes information indicative of an operation that controls the flow of information flowing through the logical design but does not change shape of the information flowing through the logical design.

[0013] In one aspect, the receiving unit is further configured to receive information specifying one or more components of the logical design that includes information indicative of a target component having one or more attributes of data to be stored in a target data store. In one

aspect, the generating unit comprises an exporting unit configured to export a list of attributes to a downstream component. In one aspect, the receiving unit is further configured to receive a change in the logical design through the introduction of one or more relationships; and the determining unit is further configured to determine an updated equivalent data flow model.

5 [0014] In one embodiment, a system facilitating generation of a data mapping includes means for receiving information specifying a set of entity relationships as a component of the logical design; means for determining an equivalent data flow model based on the set of entity relationships; and means for generating information indicative of the equivalent data flow model in the logical flow design. In one aspect, the system further comprises means for deriving one or
10 more attributes of a dataset representing the set of entity relationships based on information declaring relationships between attributes of data sources.

[0015] In another aspect, the system further comprises means for receiving information specifying one or more components of the logical design that includes information indicative of an operation that changes shape of the information flowing through the logical design. In
15 another aspect, the system further comprises means for receiving information specifying one or more components of the logical design that includes information indicative of an operation that controls the flow of information flowing through the logical design but does not change shape of the information flowing through the logical design.

[0016] In another aspect, the system further comprises means for receiving information specifying one or more components of the logical design that includes information indicative of a target component having one or more attributes of data to be stored in a target datastore. In another aspect, means for generating the information indicative of the equivalent data flow model in the logical flow design comprises means for exporting a list of attributes to a downstream component. In another aspect, the system further comprises means for receiving a change in the logical design through the introduction of one or more relationships; and means for determining an updated equivalent data flow model.
25

[0017] A further understanding of the nature of and equivalents to the subject matter of this disclosure (as well as any inherent or express advantages and improvements provided) should be realized in addition to the above section by reference to the remaining portions of this disclosure,
30 any accompanying drawings, and the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] In order to reasonably describe and illustrate those innovations, embodiments, and/or examples found within this disclosure, reference may be made to one or more accompanying drawings. The additional details or examples used to describe the one or more accompanying drawings should not be considered as limitations to the scope of any of the claimed inventions, any of the presently described embodiments and/or examples, or the presently understood best mode of any innovations presented within this disclosure.

[0019] FIG. 1 is a simplified illustration of a system that may incorporate an embodiment of the present invention.

[0020] FIG. 2 is a block diagram of a data integration system according to an embodiment of the present invention.

[0021] FIG. 3 is a simplified block diagram of a hardware/software stack that may be used to implement a data integration system according to an embodiment of the present invention.

[0022] FIG. 4 is a block diagram of an environment having various heterogeneous data sources for which data integration scenarios may be created in various embodiments of the present invention.

[0023] FIGS. 5A and 5B depict simplified data flows in conventional data integration processing that may be performed by the data integration system.

[0024] FIGS. 6A and 6B depict simplified data flows in next generation data integration processing that may be performed by the data integration system, in accordance with an embodiment of the present invention.

[0025] FIG. 7 is a simplified block diagram of interactions between an ODI Studio and a repository of the data integration system in one embodiment according to the present invention.

[0026] FIG. 8 depicts a flowchart of a method for creating a data integration scenario in accordance with an embodiment of the present invention.

[0027] FIG. 9 are a screenshot of a user interface for creating data integration scenarios in accordance with an embodiment of the present invention.

[0028] FIG. 10 depicts a flowchart of a method for creating a mapping in accordance with an embodiment of the present invention.

[0029] FIG. 11 is a screenshot of a user interface for providing mapping information in data integration scenarios in accordance with an embodiment of the present invention.

5 **[0030]** FIG. 12 is a screenshot of a user interface for providing flow information in data integration scenarios in accordance with an embodiment of the present invention.

[0031] FIG. 13 depicts a flowchart of a method for creating a package in accordance with an embodiment of the present invention.

10 **[0032]** FIG. 14 is a screenshot of a user interface for providing package sequence information in a data integration scenario in accordance with an embodiment of the present invention.

[0033] FIG. 15 depicts a flowchart of a method for deploying a data integration scenario in accordance with an embodiment of the present invention.

[0034] FIG. 16 is a simplified block diagram of a combined flow-based and entity-based mapping in one embodiment according to the present invention.

15 **[0035]** FIG. 17 depicts a flowchart of a method for generating a combined flow-based and entity-based mapping in accordance with an embodiment of the present invention.

[0036] FIG. 18 is a simplified block diagram of the mapping of FIG. 16 with a dataset view in one embodiment according to the present invention.

20 **[0037]** FIGS. 19A and 19B are simplified block diagrams of logical and physical designs for a combined flow-based and entity-based mapping in one embodiment according to the present invention.

[0038] FIG. 20 depicts a flowchart of a method for generating a physical design of a combined flow-based and entity-based mapping in accordance with an embodiment of the present invention.

25 **[0039]** FIG. 21 is an illustration depicting relationships between static E-R and dynamic ETL models.

[0040] FIG. 22 is an illustration providing a top-level design chart of an automatic conversion system in one embodiment.

[0041] FIGS. 23A and 23B illustrate three-way relationships in two popular E-R notations.

[0042] FIGS. 24A and 24B illustrate an equivalent to the three-way relationships in two popular E-R notations.

[0043] FIG. 25 illustrates an equivalent to the three-way relationship using a series of binary relationships.

[0044] FIG. 26 illustrates three-way relationships using the standard E-R notation.

[0045] FIG. 27 depicts the rows in each table created for the entities in FIG. 26.

[0046] FIGS. 28A and 28B illustrate a three-way relationship in E-R notation and a data flow with data originating from the three entities in one embodiment.

[0047] FIG. 29 depicts a diagram that lays out relationships among various database modeling methods and their semantic contents.

[0048] FIG. 30 is a simplified block diagram of a computer system that may be used to practice embodiments of the present invention.

[0049] FIG. 31 is a simplified block diagram of a system for facilitating generation of a data mapping in accordance with an embodiment of the present invention

DETAILED DESCRIPTION OF THE INVENTION

20 [0050] Introduction

[0051] In various embodiments, a data integration system enables users to create a logical design which is platform and technology independent. The user can create a logical design that defines, at a high level, how a user wants data to flow between sources and targets. The tool can analyze the logical design, in view of the user's infrastructure, and create a physical design. The logical design can include a plurality of components corresponding to each source and target in the design, as well as operations such as joins or filters, and access points. Each component when transferred to the physical design generates code to perform operations on the data.

Depending on the underlying technology (e.g., SQL Server, Oracle, Hadoop, etc.) and the language used (SQL, pig, etc.) the code generated by each component may be different.

[0052] In one aspect, a user of data integration system is not required to specify all data attributes at each component in the logical design, from start to end. The data integration system 5 provides a plurality of component types, such as projector and selector types, that avoid the need to fully declare the information that flows through the logical design. The data integration system is able to decide what attributes are needed at operations represented by predetermined component types. This simplifies both the design and maintenance.

[0053] FIG. 1 is a simplified illustration of system 100 that may incorporate an embodiment or 10 be incorporated into an embodiment of any of the innovations, embodiments, and/or examples found within this disclosure. FIG. 1 is merely illustrative of an embodiment incorporating the present invention and does not limit the scope of the invention as recited in the claims. One of ordinary skill in the art would recognize other variations, modifications, and alternatives.

[0054] In one embodiment, system 100 includes one or more user computers 110 (e.g., 15 computers 110A, 110B, and 110C). User computers 110 can be general purpose personal computers (including, merely by way of example, personal computers and/or laptop computers running any appropriate flavor of Microsoft Corp.'s WindowsTM and/or Apple Corp.'s MacintoshTM operating systems) and/or workstation computers running any of a variety of commercially-available UNIXTM or UNIX-like operating systems. These user computers 110 20 can also have any of a variety of applications, including one or more applications configured to perform methods of the invention, as well as one or more office applications, database client and/or server applications, and web browser applications.

[0055] Alternatively, user computers 110 can be any other electronic device, such as a thin-client computer, Internet-enabled mobile telephone, and/or personal digital assistant, capable of 25 communicating via a network (e.g., communications network 120 described below) and/or displaying and navigating web pages or other types of electronic documents. Although the exemplary system 100 is shown with three user computers, any number of user computers or devices can be supported.

[0056] Certain embodiments of the invention operate in a networked environment, which can include communications network 120. Communications network 120 can be any type of network familiar to those skilled in the art that can support data communications using any of a variety of commercially-available protocols, including without limitation TCP/IP, SNA, IPX,

5 AppleTalk, and the like. Merely by way of example, communications network 120 can be a local area network (“LAN”), including without limitation an Ethernet network, a Token-Ring network and/or the like; a wide-area network; a virtual network, including without limitation a virtual private network (“VPN”); the Internet; an intranet; an extranet; a public switched telephone network (“PSTN”); an infra-red network; a wireless network, including without limitation a network operating under any of the IEEE 802.11 suite of protocols, the Bluetooth™ protocol known in the art, and/or any other wireless protocol; and/or any combination of these and/or other networks.

10

[0057] Embodiments of the invention can include one or more server computers 130 (e.g., computers 130A and 130B). Each of server computers 130 may be configured with an operating system including without limitation any of those discussed above, as well as any commercially-available server operating systems. Each of server computers 130 may also be running one or more applications, which can be configured to provide services to one or more clients (e.g., user computers 110) and/or other servers (e.g., server computers 130).

[0058] Merely by way of example, one of server computers 130 may be a web server, which 20 can be used, merely by way of example, to process requests for web pages or other electronic documents from user computers 110. The web server can also run a variety of server applications, including HTTP servers, FTP servers, CGI servers, database servers, Java servers, and the like. In some embodiments of the invention, the web server may be configured to serve web pages that can be operated within a web browser on one or more of the user computers 110 25 to perform methods of the invention.

[0059] Server computers 130, in some embodiments, might include one or more file and or/application servers, which can include one or more applications accessible by a client running on one or more of user computers 110 and/or other server computers 130. Merely by way of example, one or more of server computers 130 can be one or more general purpose computers 30 capable of executing programs or scripts in response to user computers 110 and/or other server

computers 130, including without limitation web applications (which might, in some cases, be configured to perform methods of the invention).

[0060] Merely by way of example, a web application can be implemented as one or more scripts or programs written in any programming language, such as Java, C, or C++, and/or any scripting language, such as Perl, Python, or TCL, as well as combinations of any 5 programming/scripting languages. The application server(s) can also include database servers, including without limitation those commercially available from Oracle, Microsoft, IBM and the like, which can process requests from database clients running on one of user computers 110 and/or another of server computers 130.

[0061] In some embodiments, an application server can create web pages dynamically for displaying the information in accordance with embodiments of the invention. Data provided by an application server may be formatted as web pages (comprising HTML, XML, Javascript, AJAX, etc., for example) and/or may be forwarded to one of user computers 110 via a web server (as described above, for example). Similarly, a web server might receive web page 15 requests and/or input data from one of user computers 110 and/or forward the web page requests and/or input data to an application server.

[0062] In accordance with further embodiments, one or more of server computers 130 can function as a file server and/or can include one or more of the files necessary to implement methods of the invention incorporated by an application running on one of user computers 110 and/or another of server computers 130. Alternatively, as those skilled in the art will appreciate, 20 a file server can include all necessary files, allowing such an application to be invoked remotely by one or more of user computers 110 and/or server computers 130. It should be noted that the functions described with respect to various servers herein (e.g., application server, database server, web server, file server, etc.) can be performed by a single server and/or a plurality of 25 specialized servers, depending on implementation-specific needs and parameters.

[0063] In certain embodiments, system 100 can include one or more databases 140 (e.g., databases 140A and 140B). The location of the database(s) 140 is discretionary: merely by way of example, database 140A might reside on a storage medium local to (and/or resident in) server computer 130A (and/or one or more of user computers 110). Alternatively, database 140B can 30 be remote from any or all of user computers 110 and server computers 130, so long as it can be

in communication (e.g., via communications network 120) with one or more of these. In a particular set of embodiments, databases 140 can reside in a storage-area network (“SAN”) familiar to those skilled in the art. (Likewise, any necessary files for performing the functions attributed to user computers 110 and server computers 130 can be stored locally on the 5 respective computer and/or remotely, as appropriate). In one set of embodiments, one or more of databases 140 can be a relational database that is adapted to store, update, and retrieve data in response to SQL-formatted commands. Databases 140 might be controlled and/or maintained by a database server, as described above, for example.

[0064] Data Integration Overview

10 **[0065]** FIG. 2 is a simplified block diagram of data integration system 200 according to an embodiment of the present invention. FIG. 2 is a simplified illustration of data integration system 200 that may incorporate various embodiments or implementations of the one or more inventions presented within this disclosure. FIG. 2 is merely illustrative of an embodiment or implementation of an invention disclosed herein should not limit the scope of any invention as 15 recited in the claims. One of ordinary skill in the art may recognize through this disclosure and the teachings presented herein other variations, modifications, and/or alternatives to those embodiments or implementations illustrated in the figures.

20 **[0066]** In this embodiment, data integration system 200 includes information sources 202, information integration 204, and information destinations 206. In general, information flows from information sources 202 to information integration 204 whereby the information may be consumed, made available, or otherwise used by information destinations 206. Data flows may be unidirectional or bidirectional. In some embodiments, one or more data flows may be present 25 in data integration system 200.

25 **[0067]** Information sources 202 are representative of one or more hardware and/or software elements configured to source data. Information sources 202 may provide direct or indirect access to the data. In this embodiment, information sources 202 include one or more applications 208 and one or more repositories 210.

[0068] Applications 208 are representative of traditional applications, such as desktop, hosted, web-based, or cloud-based applications. Applications 208 may be configured to receive,

process, and maintain data for one or more predetermined purposes. Some examples of applications 208 include customer relationship management (CRM) applications, financial services applications, government and risk compliance applications, human capital management (HCM), procurement applications, supply chain management applications, project or portfolio

5 management applications, or the like. Applications 208 may include functionality configured for manipulating and exporting application data in a variety of human-readable and machine-readable formats, as is known in the art. Applications 208 may further access and store data in repositories 210.

[0069] Repositories 210 are representative of hardware and/or software elements configured to 10 provide access to data. Repositories 210 may provide logical and/or physical partitioning of data. Repositories 210 may further provide for reporting and data analysis. Some examples of repositories 210 include databases, data warehouses, cloud storage, or the like. A repository may include a central repository created by integrating data from one or more applications 208. Data stored in repositories 210 may be uploaded from an operational system. The data may pass 15 through additional operations before being made available in a source.

[0070] Information integration 204 is representative of one or more hardware and/or software elements configured to provide data integration services. Direct or indirect data integration services can be provided in information integration 204. In this embodiment, information integration 204 includes data migration 212, data warehousing 214, master data management 20 216, data synchronization 218, federation 220, and real-time messaging 222. It will be understood that information integration 204 can include one or more modules, services, or other additional elements than those shown in here that provide data integration functionality.

[0071] Data migration 212 is representative of one or more hardware and/or software elements configured to provide data migration. In general, data migration 212 provides one or more 25 processes for transferring data between storage types, formats, or systems. Data migration 212 usually provides for manual or programmatic options to achieve a migration. In a data migration procedure, data on or provided by one system is mapped to another system providing a design for data extraction and data loading. A data migration may involve one or more phases, such a design phase where one or more designs are created that relate data formats of a first system to 30 formats and requirements of a second system, a data extraction phase where data is read from the

first system, a data cleansing phase, and a data loading phase where data is written to the second system. In some embodiments, a data migration may include a data verification phases to determine whether data is accurately processed in any of the above phases.

[0072] Data warehousing 214 is representative of one or more hardware and/or software

5 elements configured to provide databases used for reporting and data analysis. A data warehouse is typically viewed as a central repository of data which is created by integrating data from one or more disparate sources. Data warehousing 214 may include the current storage of data as well as storage of historical data. Data warehousing 214 may include typical extract, transform, load (ETL)-based data warehouse whereby staging, data integration, and access layers house key
10 functions. In one example, a staging layer or staging database stores raw data extracted from each of one or more disparate source data systems. An integration layer integrates disparate data sets by transforming the data from the staging layer often storing this transformed data in an operational data store (ODS) database. The integrated data is then moved to yet another database, often called the data warehouse database. The data can be arranged into hierarchical
15 groups (often called dimensions) and into facts and aggregate facts. An access layer may be provided to help users or other systems retrieve data. Data warehouses can be subdivided into data marts whereby each data mart stores subsets of data from a warehouse. In some embodiments, data warehousing 214 may include business intelligence tools, tools to extract, transform and load data into the repository, and tools to manage and retrieve metadata.

20 **[0073]** Master data management 216 is representative of one or more hardware and/or software

elements configured to manage a master copy of data. Master data management 216 may include a set of processes, governance, policies, standards and tools that consistently define and manage master data. Master data management 216 may include functionality for removing duplicates, standardizing data, and incorporating rules to eliminate incorrect data from entering a
25 system in order to create an authoritative source of master data. Master data management 216 may provide processes for collecting, aggregating, matching, consolidating, quality-assuring, persisting and distributing data throughout an organization to ensure consistency and control in the ongoing maintenance and application use of information.

[0074] Data synchronization 218 is representative of one or more hardware and/or software

30 elements configured to synchronize data. Data synchronization 218 may provide for establishing

consistency among data from a source to a target and vice versa. Data synchronization 218 may further provide for the continuous harmonization of the data over time.

[0075] Federation 220 is representative of one or more hardware and/or software elements configured to consolidate a view of data from constituent sources. Federation 220 may

5 transparently map multiple autonomous database systems into a single federated database. The constituent databases maybe interconnected via a computer network and may be geographically decentralized. Federation 220 provides an alternative to merging several disparate databases. A federated database, or virtual database, for example, may provide a composite of all constituent databases. Federation 220 may not provide actual data integration in the constituent disparate

10 databases but only in the view.

[0076] Federation 220 may include functionality that provides a uniform user interface, enabling users and clients to store and retrieve data in multiple noncontiguous databases with a single query -- even if the constituent databases are heterogeneous. Federation 220 may include functionality to decompose a query into subqueries for submission to relevant constituent data sources and composite the result sets of the subqueries. Federation 220 can include one or more wrappers to the subqueries to translate them into appropriate query languages. In some 15 embodiments, federation 220 is a collection of autonomous components that make their data available to other members of the federation through the publication of an export schema and access operations.

20 **[0077]** Real-time messaging 222 is representative of one or more hardware and/or software elements configured to provide messaging services subject to a real-time constraint (e.g., operational deadlines from event to system response). Real-time messaging 222 may include functionality that guarantees an action or response within strict time constraints. In one example, real-time messaging 222 may be tasked with taking some orders and customer data from one 25 database, combining it with some employee data held in a file, and then loading the integrated data into a Microsoft SQL Server 2000 database. Because orders need to be analyzed as they arrive, real-time messaging 222 may pass the orders through to a target database in as close to real time as possible and extract only the new and changed data to keep the workload as small as possible.

[0078] Information destinations 206 are representative of one or more hardware and/or software elements configured to store or consume data. In this embodiment, information destinations 206 may provide direct or indirect access to the data. In this embodiment, information destinations 206 include one or more applications 224 and one or more repositories 226.

[0079] Applications 224 are representative of traditional applications, such as desktop, hosted, web-based, or cloud-based applications. Applications 224 may be configured to receive, process, and maintain data for one or more predetermined purposes. Some examples of applications 224 include customer relationship management (CRM) applications, financial services applications, government and risk compliance applications, human capital management (HCM), procurement applications, supply chain management applications, project or portfolio management applications, or the like. Applications 224 may include functionality configured for manipulating and importing application data in a variety of human-readable and machine-readable formats, as is known in the art. Applications 224 may further access and store data in repositories 226.

[0080] Repositories 226 are representative of hardware and/or software elements configured to provide access to data. Repositories 226 may provide logical and/or physical partitioning of data. Repositories 226 may further provide for reporting and data analysis. Some examples of repositories 226 include databases, data warehouses, cloud storage, or the like. A repository may include a central repository created by integrating data from one or more applications 226. Data stored in repositories 226 may be uploaded or imported through information integration 204. The data may pass through additional operations before being made available at a destination.

[0081] Data Integration System

[0082] FIG. 3 is a simplified block diagram of a hardware/software stack that may be used to implement data integration system 200 according to an embodiment of the present invention. FIG. 3 is merely illustrative of an embodiment or implementation of an invention disclosed herein should not limit the scope of any invention as recited in the claims. One of ordinary skill in the art may recognize through this disclosure and the teachings presented herein other variations, modifications, and/or alternatives to those embodiments or implementations illustrated in the figures. One example of components found within data integration system 200

according to this embodiment may include ORACLE DATA INTEGRATOR, a member of the ORACLE FUSION Middleware family of products provided by Oracle of Redwood Shores, California. ORACLE DATA INTEGRATOR is a Java-based application that uses one or more databases to perform set-based data integration tasks. In addition, ORACLE DATA

5 INTEGRATOR can extract data, provide transformed data through Web services and messages, and create integration processes that respond to and create events in service-oriented architectures. ORACLE DATA INTEGRATOR is based on an ELT [extract-Load and Transform] architecture rather than conventional ETL [extract-transform-load] architectures. A copy of a user manual for ORACLE DATA INTEGRATOR is attached to this disclosure and
10 incorporated herein by reference for all purposes.

[0083] In various embodiments, data integration system 200 provides a new declarative design approach to defining data transformation and integration processes, resulting in faster and simpler development and maintenance. Data integration system 200 thus separates declarative rules from the implementation details. Data integration system 200 further provides a unique E-
15 LT architecture (Extract - Load Transform) for the execution of data transformation and validation processes. This architecture in embodiments eliminates the need for a standalone ETL server and proprietary engine. In some embodiments, data integration system 200 instead leverages the inherent power of RDBMS engines.

[0084] In some embodiments, data integration system 200 integrates in one or more
20 middleware software packages, such as the ORACLE FUSION MIDDLEWARE platform and becomes a component of the middleware stack. As depicted in FIG. 3 data integration system 200 may provide run-time components as Java EE applications.

[0085] In this example, one component of data integration system 200 is repositories 302. Repositories 302 are representative of hardware and/or software elements configured to store
25 configuration information about an IT infrastructure, metadata of all applications, projects, scenarios, and execution logs. In some aspects, multiple instances of repositories 302 can coexist in an IT infrastructure, for example Development, QA, User, Acceptance, and Production. Repositories 302 are configured to allow several separated environments that exchange metadata and scenarios (for example: Development, Test, Maintenance and Production

environments). Repositories 302 further are configured to act as a version control system where objects are archived and assigned a version number.

[0086] In this example, repositories 302 are composed of at least one master repository 304 and one or more work repositories 306. Objects developed or configured for use within data integration system 200 may be stored in one of these repository types. In general, master repository 304 stores the following information: security information including users, profiles and rights, topology information including technologies, server definitions, schemas, contexts, languages and so forth, and versioned and archived objects. The one or more work repositories 306 may contain actual developed objects.

[0087] Several work repositories may coexist in data integration system 200 (for example, to have separate environments or to match a particular versioning life cycle). The one or more work repositories 306 store information for models, including schema definition, data stores structures and metadata, fields and columns definitions, data quality constraints, cross references, data lineage, and so forth. The one or more work repositories 306 may further store projects, including business rules, packages, procedures, folders, knowledge modules, variables and so forth, and scenario execution, including scenarios, scheduling information and logs. In some aspects, the one or more work repositories 306 may contain only execution information (typically for production purposes), and be designated as an execution repository.

[0088] In various embodiments, repositories 302 store one or more ETL projects. An ETL project defines or otherwise specifies one or more data models that model data attributes of data in a source or target. An ETL project further provides for data quality control as well as defining mappings to move and transform data. Data integrity control ensures the overall consistency of the data. Application data is not always valid for the constraints and declarative rules imposed by a particular source or target. For example, orders may be found with no customer, or order lines with no product, and so forth. Data integration system 200 provides a working environment to detect these constraint violations and to store them for recycling or reporting purposes.

[0089] In some embodiments of data integration system 200, there are two different types of controls: Static Control and Flow Control. Static Control implies the existence of rules that are used to verify the integrity of application data. Some of these rules (referred to as constraints)

may already be implemented in data servers (using primary keys, reference constraints, etc.). Data integration system 200 allows for the definition and checking of additional constraints, without declaring them directly in a source. Flow Control relates to targets of transformation and integration processes that implement their own declarative rules. Flow Control verifies an application's incoming data according to these constraints before loading the data into a target. Flow control procedures are general referred to as mappings.

[0090] An ETL project can be automated into a package that can be deployed for execution in a runtime environment. Accordingly, the automation of data integration flows is achieved by sequencing the execution of the different steps (mappings, procedures, and so forth) in a package and by producing a production scenario containing ready-to-use code for each of these steps. A package is typically made up of a sequence of steps organized into an execution diagram.

Packages are the main objects used to generate scenarios for production. They represent the data integration workflow and can perform jobs, such as for example: start a reverse-engineering process on a datastore or a model, send an email to an administrator, download a file and unzip it, define the order in which mappings must be executed, and define loops to iterate over execution commands with changing parameters.

[0091] A scenario is designed to put a source component (mapping, package, procedure, variable) into production. A scenario results from the generation of code (SQL, shell, and so forth) for this component. Once generated, the code of the source component is frozen and the scenario is stored inside repositories 302, such as one or more of work repositories 306. A scenario can be exported and then imported into different production environments.

[0092] In various embodiments, data integration system 200 is organized around repositories 302 in a modular fashion accessed by Java graphical modules and scheduling agents. Graphical modules can be used to design and build one or more integration processes stored in repositories 302. Administrators, Developers and Operators may use a development studio to access repositories 302. Agents can be used to schedule and coordinate a set of integration tasks associated with an integration process stored in repositories 302. For example, at runtime, an agent deployed on a desktop, web services, or otherwise in communication with a source coordinates the execution of one or more integration processes. The agent may retrieve code

stored in master repository 304, connect to various source and target systems, and orchestrate an overall data integration process or scenario.

[0093] In this embodiment, data integration system 200 includes desktop 308 that may include one or more of the above discussed graphical modules and/or agents. Desktop 308 is representative of one or more desktop or workstation computing devices, such as personal computers, laptops, netbooks, tablets, and the like. Desktop 308 includes a Java virtual machine (JVM) 310 and Oracle Data Integrator (ODI) Studio 312. Java virtual machine (JVM) 310 is a virtual machine that can execute Java bytecode. JVM 310 is most often implemented to run on an existing operating system, but can also be implemented to run directly on hardware. JVM 310 provides a run-time environment in which Java bytecode can be executed, enabling features such as runtime web service (WS) 314 and agent 316. JVM 310 may include a Java Class Library, a set of standard class libraries (in Java bytecode) that implement the Java application programming interface (API), and other elements that form a Java Runtime Environment (JRE).

[0094] Agent 316 is configured to schedule and coordinate a set of integration tasks associated with one or more integration processes stored in repositories 302. For example, at runtime, an agent coordinates the execution of integration processes. The agent may retrieve code stored in master repository 304, connect to various source and target systems, and orchestrate an overall data integration process or scenario.

[0095] Referring again to FIG. 3, ODI Studio 312 includes hardware and/or software elements configured to design data integration projects. In this example, ODI Studio 312 includes four graphical modules or navigators that are used to create and manage data integration projects, namely, designer module 318, operator module 320, topology module 322, and security module 324. Designer module 318 is a module configured to define data stores (tables, files, Web services, and so on), data mappings, and packages (sets of integration steps, including mappings). In various embodiments, designer module 318 defines declarative rules for data transformation and data integrity. Accordingly, project development takes place in designer module 318. Additionally, in designer module 318, is where database and application metadata are imported and defined. Designer module 318, in one embodiment, uses metadata and rules to generate data integration scenarios or load plans for production. In general, designer module 318 is used to design data integrity checks and to build transformations such as for example:

automatic reverse-engineering of existing applications or databases, graphical development and maintenance of transformation and integration mappings, visualization of data flows in the mappings, automatic documentation generation, and customization of generated code.

[0096] Operator module 320 is a module configured to view and manage production 5 integration jobs. Operator module 320, thus, manages and monitors data integration processes in production and may show execution logs with error counts, the number of rows processed, execution statistics, the actual code that is executed, and so on. At design time, developers can also use operator module 320 for debugging purposes in connection with designer module 318.

[0097] Topology module 322 is a module configured to create and manage connections to 10 datasources and agents. Topology module 322 defines the physical and logical architecture of the infrastructure. Infrastructure or projects administrators may register servers, database schemas and catalogs, and agents in a master repository through topology module 322. Security module 324 is a module configured to manage users and their repository privileges.

[0098] In general, a user or process interacts with designer module 318 to create a data 15 integration project having one or more data integration processes for sources and targets 326. Each data integration process includes at least one data integration task. In some embodiments, a data integration task is defined by a set of business rules indicative of what bit of data is to be transformed and combined with other bits as well as technical specifics of how the data is actually extracted, loaded, and so on. In preferred embodiments, a data integration task is 20 specified using a declarative approach to build data mappings. A mapping is an object that populates one datastore, called the target, which data coming from one or more other datastores, known as sources. In general, columns in the source datastore are linked to the columns in the target datastore through mapping. A mapping can be added into a package as a package step. As discussed above, a package defines a data integration job. A package is created under a project 25 and is made up of an organized sequence of steps, each of which can be a mapping or a procedure. A package can have one entry point and multiple exit points.

[0099] In some embodiments, when creating a new mapping, a developer or technical business user interacts with designer 318 to first define which data is integrated and which business rules should be used. For example, the developer may specify what tables are to be joined, filters to 30 be applied, and SQL expressions to be used to transform data. The particular dialect of SQL that

is used is determined by the database platform on which the code is to be executed. Then, in a separate step, technical staff can interact with designer 318 to choose the most efficient way to extract, combine, and then integrate this data. For example, the technical staff may use database-specific tools and design techniques such as incremental loads, bulk-loading utilities, slowly changing dimensions, and changed-data capture.

5 [0100] In this embodiment, mappings can be created for sources and targets 326. Sources and targets 326 may include one or more legacy applications 328, one or more files/XML documents 330, one or more applications 332, one or more data warehouses (DW), business intelligence (BI) tools and applications, and enterprise process management (EPM) tools and applications 10 334, and one or more JVMs 336 (including runtime web service 340 and agent 342).

15 [0101] FIG. 4 is a block diagram of environment 400 having various heterogeneous data sources for which data integration scenarios may be created in various embodiments of the present invention. In this example, environment 400 includes ODI Studio 312 and repositories 302. Repositories 302 contain all of the metadata required to generate integration scenarios 400. A user or process interacts with ODI Studio 312 to create integration scenarios 400 using data integrity controls 402 and declarative rules 404.

20 [0102] Orders application 406 is representative of an application for tracking customer orders. An “Orders Application” data model is created to represent data stored in Orders application 406 as well as any data integrity controls or conditions. For example, the “Orders Application” data model may be based on a Hyper Structured Query Language (HSQL) interface and include five datastores, SRC_CITY, SRC_CUSTOMER, SRC_ORDERS, SRC_ORDER_LINES, SRC_PRODUCT, and SRC_REGION.

25 [0103] Parameter file 408 is representative of a flat file (e.g., ASCII) issued from a production system containing a list of sales representatives and the segmentation of ages into age ranges. In this example, a “Parameter” data model is created to represent the data in the flat file. For example, the “Parameter” data model may be based on a file interface and include two datastores, SRC_SALES_PERSON and SRC_AGE_GROUP.

[0104] Sales administration application 410 is representative of an application for tracking sales. The sales administration application 410 may be a data warehouse populated with

transformations of data from orders application 406 and parameter file 408. A “Sales Administration” data model is created to represent data stored in sales administration application 410 as well as any data integrity controls or conditions or transformations. For example, the “Sales Administration” data model may be based on a Hyper Structured Query Language (HSQL) interface and include six datastores, TRG_CITY, TRG_COUNTRY, TRG_CUSTOMER, TRG_PRODUCT, TRG_PROD_FAMILY, TRG_REGION, and TRG_SALE.

5 [0105] FIGS. 5A and 5B depict simplified data flows in conventional data integration processing that may be performed by data integration system 200. In this example, data from 10 orders application 406, parameter file 408, and one or more other optional or additional sources flow through a tradition ETL process targeted to sales administration application 410. Data transforms occur in a separate ETL server 500. The scenario requires dedicated or proprietary resources, results in poorer performance, and incurs high costs.

15 [0106] FIGS. 6A and 6B depict simplified data flows in next generation data integration processing that may be performed by data integration system 200, in accordance with an embodiment of the present invention. In this example, data from orders application 406, parameter file 408, and one or more other optional or additional sources flow through E-LT 20 process targeted to sales administration application 410. Data transforms leverage existing resources resulting in higher performance and efficiency. As described above, prior ETL systems required dedicated and/or proprietary infrastructure to perform data transforms. This was done, in part, to accommodate unknown user infrastructures. For example, without knowing what types of databases are being used, prior ETL systems were unable to anticipate what transform operations would be available in a given system. However, this results in under-utilized resources, such as the user’s existing databases and servers which are capable of 25 executing the appropriate data transforms without any dedicated and/or proprietary infrastructure.

30 [0107] In accordance with an embodiment, the present invention leverages the user’s existing infrastructure by enabling the user to customize a data integration process according to the user’s particular needs. For example, when a data integration plan is designed, it can be divided into discrete portions which are executable by a single system, referred to as execution units. Once a

data integration plan has been divided into a plurality of execution units, the user can be presented with a physical plan based on the user's infrastructure and system resources. This plan can be further customized by the user to change which user systems execute which execution units. For example, a user may be presented with a plan in which a join operation is executed on 5 a first database, and the user may customize the plan by moving the join operation to a second database.

[0108] As shown in FIG. 6B, this results in an extract-load-transform (E-LT) architecture that does not rely on a stand-alone transform server which characterized prior ETL systems. Instead, as described above, data transforms can be performed on the user's existing infrastructure. The 10 E-LT architecture provides users with greater flexibility while reducing costs associated with acquiring and maintaining proprietary transform servers.

[0109] Referring again to FIG. 3, agents can be used to schedule and coordinate a set of integration tasks associated with an integration process. For example, at runtime, an agent coordinates the execution of integration processes. The agent may retrieve code stored in master 15 repository 304, connect to the various source and target systems and orchestrates an overall data integration process or scenario. In various embodiments, there are two types of agents. In one example, a standalone agent is installed on desktop 308, such as agent 316. In another example, an application server agent can be deployed on application server 326 (such as a Java EE Agent deployed on an Oracle WebLogic Server) and can benefit from the application server layer 20 features such as clustering for High Availability requirements. In yet another example, an agent can be deployed on sources and targets 326, such as agent 342.

[0110] In this embodiment, data integration system 200 includes application server 344 that may include one or more of the above discussed agents. Application server 344 is representative of one or more application servers, web-servers, or hosted applications. In this example, 25 application server 344 includes FMW console 346, servlet container 348, web services container 350, and data sources connection pool 352.

[0111] FMW console 346 is representative of one or more hardware and/or software elements configured to manage aspects of application server 344, such as information related to servlet container 348, web services container 350, and data sources connection pool 334. For example, 30 FMW console 346 may be a browser-based, graphical user interface used to manage an Oracle

WebLogic Server domain. FMW console 346 may include functionality to configure, start, and stop WebLogic Server instances, configure WebLogic Server clusters, configure WebLogic Server services, such as database connectivity (JDBC) and messaging (JMS), configure security parameters, including creating and managing users, groups, and roles, configure and deploy Java

5 EE applications, monitor server and application performance, view server and domain log files, view application deployment descriptors, and edit selected run-time application deployment descriptor elements. In some embodiments, FMW console 346 includes ODI plug-in 354 providing FMW console 346 with access to data integration processes in production and may show execution logs with error counts, the number of rows processed, execution statistics, the

10 actual code that is executed, and so forth.

[0112] Servlet container 348 is representative of one or more hardware and/or software elements configured to extend the capabilities of application server 344. Servlets are most often used to process or store data that was submitted from an HTML form, provide dynamic content such as the results of a database query, and manage state information that does not exist in the

15 stateless HTTP protocol, such as filling the articles into the shopping cart of the appropriate customer. A servlet is typically a Java class in Java EE that conforms to the Java Servlet API, a protocol by which a Java class may respond to requests. To deploy and run a servlet, servlet container 348 is used as a component of a web server that interacts with servlets. Accordingly, servlet container 348 may extend functionality provided by public web service 356 and data 20 services 358 of web services container 350 as well as access to data pools provided by data sources connection pool 352. Servlet container 348 is also responsible for managing the lifecycle of servlets, mapping a URL to a particular servlet and ensuring that the URL requester has the correct access rights.

[0113] In this example, servlet container 348 includes Java EE application 360 associated with 25 ODI SDK 362, ODI console 364, and runtime web service 366 associated with Java EE agent 368. ODI SDK 362 provides a software development kit (SDK) for data integration and ETL design. ODI SDK 362 enables automation of work that is common and very repetitive allowing a user to script repetitive tasks.

[0114] ODI console 364 is a Java Enterprise Edition (Java EE) application that provides Web 30 access to repositories 302. ODI console 364 is configured to allow users to browse Design-Time

objects, including projects, models, and execution logs. ODI console 364 may allow users to view flow maps, trace the source of all data, and even drill down to the field level to understand the transformations used to build the data. In addition, end users can launch and monitor scenario execution through ODI console 364. In one aspect, ODI console 364 provides 5 administrators with the ability to view and edit Topology objects such as Data Servers, Physical and Logical Schemas as well as to manage repositories 302.

[0115] Data Scenario Design and Development

[0116] As discussed above, a scenario is designed to put a source component (mapping, package, procedure, variable) into production. A scenario results from the generation of code 10 (SQL, shell, and so forth) for this component. A scenario can be exported and then imported into different production environments.

[0117] FIG. 7 is a simplified block diagram of interactions between an ODI Studio and a repository of the data integration system in one embodiment according to the present invention. In the embodiment shown in FIG. 7, ODI Studio 312 of FIG. 3 uses metadata and rules to 15 generate data integration scenarios 700 for production. In general, designer module 318 is used to design data integrity checks and to build transformations such as for example: automatic reverse-engineering of existing applications or databases, graphical development and maintenance of transformation and integration interfaces, visualization of data flows in the interfaces, automatic documentation generation, and customization of generated code.

[0118] FIG. 8 depicts a flowchart of method 800 for creating a data integration scenario in accordance with an embodiment of the present invention. Implementations of or processing in method 800 depicted in FIG. 8 may be performed by software (e.g., instructions or code modules) when executed by a central processing unit (CPU or processor) of a logic machine, such as a computer system or information processing device, by hardware components of an 20 electronic device or application-specific integrated circuits, or by combinations of software and hardware elements. Method 800 depicted in FIG. 8 begins in step 810.

[0119] In various embodiments, a user may initiate a session with designer module 318 of ODI Studio 312 and connect to repositories 302. The user may interact with one or more user interface features to create a new data integration project or select from existing data integration

projects stored in, for example, master repository 304. In general, designer module 318 is used to manage metadata, to design data integrity checks, and to build transformations. In various embodiments, the main objects handled through designer module 318 are models and projects. Data models contain all of the metadata in a data source or target (e.g., tables, columns, 5 constraints, descriptions, cross-references, etc.). Projects contain all of the loading and transformation rules for a source or target (e.g., mappings, procedures, variables, etc.).

[0120] In step 820, one or more data models are created. In step 830, one or more projects are created. FIG. 9 is a screenshot of a user interface for creating a data integration scenario in accordance with an embodiment of the present invention. In this example, navigation panel 910 10 displays information and includes functionality for interacting with data models. Navigation panel 920 displays information and includes functionality for interacting with projects. As discussed above, the user may not only create the data model, but also develop any data integrity checks for the data in the data models. Additionally, the user may specify interfaces, procedures, variables for projects that provide data integrity and transforms for the data in a flow that loads 15 data from a source into a target. In step 840, one or more data integration scenarios are generated. FIG. 8 ends in step 850.

[0121] FIG. 10 depicts a flowchart of method 1000 for creating a mapping in accordance with an embodiment of the present invention. Implementations of or processing in method 1000 depicted in FIG. 10 may be performed by software (e.g., instructions or code modules) when 20 executed by a central processing unit (CPU or processor) of a logic machine, such as a computer system or information processing device, by hardware components of an electronic device or application-specific integrated circuits, or by combinations of software and hardware elements. Method 1000 depicted in FIG. 10 begins in step 1010.

[0122] In step 1020, target datastore information is received. For example, a user may interact 25 with one or more user interface features of designer module 318 to provide target datastore information. In one embodiment, the user may drag and drop target datastore information comprising one or more data models from navigation panel 910 onto a mapping or flow panel that visually represents aspects of a selected data model and any associated transforms or data integrity checks.

[0123] In step 1030, source datastore information is received. For example, a user may interact with one or more user interface features of designer module 318 to provide source datastore information. In one embodiment, the user may drag and drop source datastore information comprising one or more data models from navigation panel 910 onto the same mapping or flow panel of the target datastore information that visually represents aspects of a selected data model and any associated transforms or data integrity checks.

5 [0124] In various embodiments, the source datastore information and the target data store information may be composed of one or more data models and optionally operations. Some examples of operations can include one or more data set operations (e.g., unions, joins, 10 intersections, etc.), data transformations, data filter operations, constraints, descriptions, cross-references, integrity checks, or the like. In further embodiments, some of these operations may be preconfigured and visually represented in designer module 318. In other embodiments, custom operations may be provided allowing the user to specify logic, mappings, and the like that implement an operation.

15 [0125] In step 1040, mapping information is received. For example, a user may interact with one or more user interface features of designer module 318 to map the source datastore information to the target datastore information. In one embodiment, the user may visually connect attributes of data elements in the source datastore information with attributes of data elements in the target datastore information. This may be done by matching column names of 20 tables in the source datastore information and the target datastore information. In further embodiments, one or more automatic mapping techniques may be used to provide mapping information.

[0126] FIG. 11 is a screenshot of a user interface for providing mapping information in a data integration scenario in accordance with an embodiment of the present invention. In this 25 example, attributes of source datastore information in panel 1110 are mapped to attributes of target datastore information in panel 1120.

[0127] Referring again to FIG. 10, in step 1050, data loading strategies are received. A data loading strategy includes information on how the actual data from the source datastore information is to be loaded during an extract phase. Data loading strategies can be defined in a

flow tab of designer 318. In some embodiments, a data loading strategy can be automatically computed for a flow depending on a configuration of the mapping.

[0128] For example, one or more knowledge modules may be proposed for the flow. A knowledge module (KM) is a component that implements reusable transformation and ELT (extract, load, and transform) strategies across different technologies. In one aspect, knowledge modules (KMs) are code templates. Each KM can be dedicated to an individual task in an overall data integration process. The code in KMs appears in nearly the form that it will be executed with substitution methods enabling it to be used generically by many different integration jobs. The code that is generated and executed is derived from the declarative rules and metadata defined in the designer module 318. One example of this is extracting data through change data capture from Oracle Database 10g and loading the transformed data into a partitioned fact table in Oracle Database 11g, or creating timestamp-based extracts from a Microsoft SQL Server database and loading this data into a Teradata enterprise data warehouse.

[0129] The power of KMs lies in their reusability and flexibility—for example, a loading strategy can be developed for one fact table and then the loading strategy can be applied to all other fact tables. In one aspect, all mappings that use a given KM inherit any changes made to the KM. In some embodiments, five different types of KMs are provided, each of them covering one phase in a transformation process from source to target, such as an integration knowledge module (IKM), a loading knowledge module (LKM), and a check knowledge module CKM.

[0130] Referring to FIG. 4, a user may define a way to retrieve the data from SRC _ AGE _ GROUP, SRC _ SALES _ PERSON files and from the SRC _ CUSTOMER table in environment 400. To define a loading strategies, a user may select a source set that corresponds to the loading of the SRC _ AGE _ GROUP file and select a LKM File to SQL to implement the flow from a file to SQL. In one aspect, a LKM is in charge of loading source data from a remote server to a staging area.

[0131] In step 1060, data integration strategies are received. After defining the loading phase, the user defines a strategy to adopt for the integration of the loaded data into a target. To define the integration strategies, the user may select a target object and select a IKM SQL Incremental Update. An IKM is in charge of writing the final, transformed data to a target. When an IKM is started, it assumes that all loading phases for remote servers have already carried out their tasks,

such as having all remote source data sets loaded by LKMs into a staging area, or the source datastores are on the same data server as the staging area.

[0132] In step 1070, data control strategies are received. In general, a CKM is in charge of checking that records of a data set are consistent with defined constraints. A CKM may be used to maintain data integrity and participates in overall data quality initiative. A CKM can be used in 2 ways. First, to check the consistency of existing data. This can be done on any datastore or within interfaces. In this case, the data checked is the data currently in the datastore. In a second case, data in the target datastore is checked after it is loaded. In this case, the CKM simulates the constraints of the target datastore on the resulting flow prior to writing to the target.

5 [0133] FIG. 12 is a screenshot of a user interface for providing flow information in a data integration scenario in accordance with an embodiment of the present invention.

[0134] In step 1080, an interface is generated. FIG. 10 ends in step 1090.

[0135] Data Integration Scenario Packages and Deployment

[0136] As discussed above, automation of data integration flows can be achieved in data 15 integration system 200 by sequencing the execution of the different steps (mappings, procedures, and so forth) in a package and by producing a production scenario containing the ready-to-use code for each of these steps. A package is made up of a sequence of steps organized into an execution diagram. Packages are the main objects used to generate scenarios for production. A scenario is designed to put a source component (mapping, package, procedure, variable) into 20 production. A scenario results from the generation of code (SQL, shell, and so forth) for this component. A scenario can be exported and then imported into different production environments.

[0137] FIG. 13 depicts a flowchart of a method for creating a package in accordance with an embodiment of the present invention. Implementations of or processing in method 1300 25 depicted in FIG. 13 may be performed by software (e.g., instructions or code modules) when executed by a central processing unit (CPU or processor) of a logic machine, such as a computer system or information processing device, by hardware components of an electronic device or application-specific integrated circuits, or by combinations of software and hardware elements. Method 1300 depicted in FIG. 13 begins in step 1310.

[0138] In step 1320, package step information is received. Package step information includes information identifying a step, elements, properties, components, and the like. In one example, a user may interact with one or more user interface features of designer module 318 to create, identify, or otherwise specify one or more steps for a package. In one embodiment, one or more 5 components are selected and placed on a diagram. These components appear as steps in the package.

[0139] In step 1330, package step sequence information is received. Package step sequence information includes information identifying an ordering for a step, dependencies, and the like. Once steps are created, the steps are ordered or reordered into a data processing chain. In one 10 example, a user may interact with one or more user interface features of designer module 318 to provide sequencing or ordering for one or more steps of a package. A data processing chain may include a unique step defined as a first step. Generally, each step has one or more termination states, such as success or failure. A step in some states, such as failure or success, can be followed by another step or by the end of the package. In one aspect, in case of some states, 15 such as failure, sequence information may define a number of retries. In another aspect, a package may have but several possible termination steps.

[0140] FIG. 14 is a screenshot of a user interface for providing package sequence information in a data integration scenario in accordance with an embodiment of the present invention.

[0141] In step 1340, a package is generated. FIG. 13 ends in step 1350.

20 [0142] As discussed above, the automation of data integration flows can be achieved by sequencing the execution of different steps (mappings, procedures, and so forth) in a package. The package can then be produced for a production scenario containing the ready-to-use code for each of the package's steps. In various embodiments, the package is deployed to run automatically in a production environment.

25 [0143] FIG. 15 depicts a flowchart of method 1500 for deploying a data integration scenario in accordance with an embodiment of the present invention. Implementations of or processing in method 1500 depicted in FIG. 15 may be performed by software (e.g., instructions or code modules) when executed by a central processing unit (CPU or processor) of a logic machine, such as a computer system or information processing device, by hardware components of an

electronic device or application-specific integrated circuits, or by combinations of software and hardware elements. Method 1500 depicted in FIG. 15 begins in step 1510.

[0144] In step 1520, an integration scenario is retrieved. In one embodiment, a package is retrieved from repositories 302. In step 1530, the integration scenario is deployed to one or more agents. In step 1540, the integration scenario is executed by the one or more agents. In one aspect, the integration scenario can be executed in several ways, such as from ODI Studio 312, from a command line, or from a web service. Scenario execution can be viewed and monitored, for example, via operator module 320 and the like as discussed above. FIG. 15 ends in step 1550.

10 [0145] **Combined Flow-based ETL and Entity Relationship-based ETL**

[0146] In most data integration systems, a mapping requires an explicit definition of all input and output attributes that form part of a map. In typical flow based ETL tools, connectors are made at the attribute level. This results in a very concise mapping model. However, this also generates a huge number of objects and makes constructing and maintaining maps cumbersome

15 due to the number of attribute level connectors.

[0147] In various embodiments, data integration system 200 incorporates one or more techniques for easing the design and maintenance of a mapping. Components can be added to an existing design simple without the need to specify all input and output attributes and allowing component level connectors to be rerouted. In one aspect, a combination of datasets and flow oriented design is provided to handle complexity along with change. Entity relationships can be specified within a logical view of the design, thus allowing datastores, joins, filters and lookups to be added or removed without requiring change to a map in general.

[0148] A dataset as used herein generally represents a data flow coming from a group of datastores. Several datasets can be merged into an interface target datastore using operations, such as set-based operators like Union and Intersect. In various embodiments, datasets can be added, removed, and ordered in the logical view of a design. Accordingly, data integration system 200 enables users to combine flow-based ETL and entity relationship-based ETL in a single view. Therefore, data integration system 200 greatly eases the design and maintenance of

a mapping. Data integration system 200 further makes adding in components to an existing design simple, typically just needing component level connectors to be rerouted.

[0149] FIG. 16 is a simplified block diagram of combined flow-based and entity-based mapping 1600 in one embodiment according to the present invention. In this example, mapping 1600 includes component 1610 representing a data source SRC_EMP, dataset 1620 representing a dataset DATASET, and component 1630 representing data target TGT_EMPDEPT. In order to update data target TGT_EMPDEPT, a join is needed for data source SRC_EMP and DATASET. Component 1640 representing a JOIN is added to mapping 1600 that connects to component 1610 and dataset 1620 as input and to component 1630 as output. Component 1640 is configured to provide a join expression, such as (SRC_EMP.DEPTNO = DATASET.DEPTNO).

[0150] In traditional data integration systems, mapping 1600 requires an explicit definition of all input and output attributes that form part of component 1640 representing the JOIN. In contrast, in various embodiments, a map developer can define entity relationships in dataset 1620 to provide how columns of data target TGT_EMPDEPT are populated directly from attributes of data source SRC_EMP represented by component 1610 and attributes of DATASET represented by dataset 1620 that flow through component 1640 and are thus visible to component 1630.

[0151] FIG. 17 depicts a flowchart of method 1700 for generating a combined flow-based and entity-based mapping in accordance with an embodiment of the present invention.

20 Implementations of or processing in method 1700 depicted in FIG. 17 may be performed by software (e.g., instructions or code modules) when executed by a central processing unit (CPU or processor) of a logic machine, such as a computer system or information processing device, by hardware components of an electronic device or application-specific integrated circuits, or by combinations of software and hardware elements. Method 1700 depicted in FIG. 17 begins in 25 step 1710.

[0152] In step 1720, one or more components are received. As discussed above, some types of components influence the shape of the data that flows through a map while other types of components control the flow of the data but don't fundamentally change the shape of the flow. In step 1730, one or more datasets are received. For example, a map designer may add, edit, or 30 remove datasets from a design. A map designer may interact with a relationship editor to specify

entity relationships between various attributes in a dataset. In one aspect, data integration system 200 is configured to extract defined entity relationships to determine the attributes that will be exposed to downstream components of a design. In step 1740, a map is generated based on the components and the datasets. In various embodiments, logical views and physical views of a 5 design can be updated to reflect changes to the components and datasets. In various aspects, data integration system 200 automatically generates a physical design based on deriving relationships in a flow's dataset view. FIG. 17 ends in step 1750.

[0153] Data integration system 200 further makes adding in components and other datasets to an existing design simple, typically just needing component level connectors to be rerouted. For 10 example, if a filter component were added into a design, changing component level connectors would not require changes in attribute assignments of certain downstream components. In another example, adding another dataset allows a map designer to specify or declare entity relationships directly from within the design view of the map.

[0154] FIG. 18 is a simplified block diagram of mapping 1600 with a dataset view in one 15 embodiment according to the present invention. In this example, component 1620 includes one or more entities 1810, 1820, and 1830. To add entity relationships into mapping 1600, a user only needs to add or defined relationships between entity attributes, such as relationship 1840. In various embodiments, such as change would not require changes to any downstream 20 assignments in mapping 1600 as the output attributes resulting from the one or more entity relationships can be derived directly from the information provided in the design view. In traditional flow tools, everything at the column level would need to be relinked by the introduction of a new dataset.

[0155] FIGS. 19A and 19B are simplified block diagrams of logical and physical designs for a 25 combined flow-based and entity-based mapping in one embodiment according to the present invention. In this example, view 1910 of FIG. 19A includes components A, B, and C representing data sources and component T representing a data target in a flow view of a logical design. Components A, B, and C are represented as a dataset that describes entity relationships in as dataset view of a logical design. Therefore, the dataset has a declared set of attributes as viewed from downstream components, such as component T that are described from entity

relationships defined by a map creator in the dataset view. Components J1 and J2 represent logical operations between attributes of the components in the dataset view.

5 [0156] In this example, view 1920 of FIG. 19B includes components A, B, and C representing data sources and component T representing a data target in a flow view of a physical design. A set of attributes is derived from the entity relationships defined in the dataset view and used to create the physical design.

10 [0157] FIG. 20 depicts a flowchart of method 2000 for generating a physical design of a combined flow-based and entity-based mapping in accordance with an embodiment of the present invention. Implementations of or processing in method 2000 depicted in FIG. 20 may be performed by software (e.g., instructions or code modules) when executed by a central processing unit (CPU or processor) of a logic machine, such as a computer system or information processing device, by hardware components of an electronic device or application-specific integrated circuits, or by combinations of software and hardware elements. Method 2000 depicted in FIG. 20 begins in step 2010.

15 [0158] In step 2020, a component definition is received. For example, a component definition may include rules, operations, procedures, variables, sequences, and the like. In step 2030, a dataset definition is received. For example, a map designer may add or edit entity relationships within a flow view of a logical design. In step 2040, a physical design is generated based on deriving relationship information from the flow design. FIG. 20 ends in step 2050.

20 [0159] Accordingly, data integration system 200 enables users to create a logical design which is platform and technology independent. The user can create a logical design that defines, at a high level, how a user wants data to flow between sources and targets. The tool can analyze the logical design, in view of the user's infrastructure, and create a physical design. The logical design can include a plurality of components corresponding to each source and target in the 25 design, as well as operations such as joins or filters, and access points. Each component when transferred to the physical design generates code to perform operations on the data. Depending on the underlying technology (e.g., SQL Server, Oracle, Hadoop, etc.) and the language used (SQL, pig, etc.) the code generated by each component may be different.

[0160] Thus, a user of data integration system 200 is not required to pre-defined dataset component in the logical design. Data integration system 200 provides tools that allow map designers to declare entity relationships in a dataset view of a logical design. Data integration system 200 is able to decide what attributes are needed at operations represented by 5 predetermined component types. This simplifies both the design and maintenance.

[0161] Entity Relational Modeling

[0162] Relational database design has been based on entity relational modeling, or E-R modeling. Traditionally, an E-R design has been used for describing the static configuration of a problem domain. The more dynamic aspects, such as extracting data out of the data stores and 10 “massaging” them into shape, are generally considered a different problem. Since the mid-1990s, there have been steady efforts toward these so-called “ETL tools.” An ETL tool can help human designers create specifications about dynamic data flows, which are generally called ETL models.

[0163] FIG. 21 is an illustration depicting relationships between static E-R and dynamic ETL 15 models. One interesting question is whether it is possible to eliminate the depicted human factor in the ETL design process. Alternatively, to put the question in another way, does the E-R model contain sufficient operational information so that a dynamic data flow model can be formed automatically without human intervention?

[0164] There are many benefits from automating the ETL design process using the E-R model. 20 One such is the productivity of the ETL designer. The E-R model can be more easily made correct than an ETL process. The E-R model also has a standard notational system understood by database engineers. But the same cannot be said for any ETL tool. Most, if not all, require steep learning curves on the part of designers. Another benefit is the better adaptability to change. Without the “middleman,” when the E-R model is finalized, so is the ETL process.

[0165] In various embodiments, techniques are disclosed to provide an automatic conversion 25 from an E-R model to an ETL model. This is based on the observation that when a database engineer reads an E-R diagram, a data flow model is usually built in his mind. Using this silent data flow model, the engineer can understand the E-R model and is able to communicate with others. The engineer even creates software based on this model. This phenomenon is more

apparent when the E-R model becomes complex. Accordingly, the inventors recognize that there can be one or more hidden data flow models in every E-R model. In one aspect, the equivalent data flow model is provided for the E-R model that has been proven to be accurate in guiding the creation of an automatic conversion system.

5 [0166] FIG. 22 is an illustration providing a top-level design chart of automatic conversion system 2200 in one embodiment. FIG. 22 may merely be illustrative of an embodiment or implementation of an invention disclosed herein should not limit the scope of any invention as recited in the claims. One of ordinary skill in the art may recognize through this disclosure and the teachings presented herein other variations, modifications, and/or alternatives to those 10 embodiments or implementations illustrated in the figures.

15 [0167] As shown in FIG. 22, an E-R model is provided as input to automatic conversion system 2200 together with a set of “user directives.” Automatic conversion system 2200 then creates the equivalent data flow model for ETL purposes. As used herein, the “user directives” 15 are a set of requirements that a user expects the calculation of the data flow model to take into consideration. For example, the user can ask for a specific order on a series of binary relationships, ask for processing a relationship on a designated machine/location due to logical, performance, or security considerations, or the like.

20 [0168] The “equivalent data flow” model as used herein represents a semantic model for the E-R model. A semantic model is used to define unambiguously what a logical model means. The semantic model can be expressed in vastly different ways, such as natural language, set theoretic 25 notations, algebraic equations, mathematical logic, or algorithmic notations (generally known as operational semantic). In various embodiments, the semantic model for the E-R model is an operational semantic model, referred to as a “CFO model.” In one aspect, defining meaning in an operational semantic format provides a twofold benefit, the first being that the operational semantic model is already in a step-by-step form consistent with the data flow model. Second, the operational semantic model is easy for human to understand, compared with other formal semantic models; and it is more rigorous than natural language explanations.

30 [0169] Binary relationships in an E-R model (or diagram) can be trivially mapped to joins in an ETL model. However, a multi-way relationship needs some work, because there is a common misconception around it. FIGS. 23A and 23B illustrate three-way relationships in two popular

E-R notations. Referring to FIG. 23A, model 2310 is drawn or otherwise represented using the standard E-R notation. In this example, model 2310 includes three entities, PET, PET_TYPE, PET_OWNER are being related in a three-way relationship called “Pet-of-Type-and-Owner.” The intuitive understanding of FIG. 23A is that the three entities can interact simultaneously.

5 [0170] In practice, however, standard E-R notation is not used. Instead, it is more common to see the so-called “Crow’s Feet” notation. The difference between the two notational systems is only superficial. Referring to FIG. 23B, model 2320 is drawn or otherwise represented using Crow’s Feet notation. In this example, model 2320 again includes the three entities, PET, PET_TYPE, PET_OWNER being related in a three-way relationship called “Pet-of-Type-and-
10 Owner” illustrated as a box in the middle with corner lines (also called an associative entity). The associative entity is created to tie the three other entities together simultaneously.

[0171] One common misconception is the mistake of equating a multi-way relationship to a series of binary relationships. FIGS. 24A and 24B illustrate an equivalent to the three-way relationships in two popular E-R notations. Referring to FIG. 24A, model 2410 is drawn or
15 otherwise represented as an equivalent to model 2310 of FIG. 23A with two binary relationships in standard E-R notation. Referring to FIG. 24B, model 2420 is drawn or otherwise represented as an equivalent to model 2320 with the equivalent model in the crow’s feet notation.

[0172] Both of these models share the same problem in that they do not require that the two binary relationships must always hold at the same time. For example, an instance in PET, call it
20 “Pet A” can relate to an instance in PET_TYPE, call it “PT A”, but it is not required that “Pet A” must also relate to an instance from PET_OWNER.

[0173] It is, however, possible to model the fact that a pet must participate in two binary relationships at the same time. FIG. 25 illustrates an equivalent to the three-way relationship using a series of binary relationships. In this example, model 2500 represents the PET entity
25 differently than FIG. 24B as an associative entity. Each instance in an associative entity relates to all other connected entities without exception. FIG. 25 may appear like a series of binary relationships, but it is really the three-way relationship in FIG. 23B in disguise – the PET entity absorbs the associative entity shown in FIG. 23B.

[0174] In one aspect, there are two special cases where it is possible for the PET entity to absorb the associative entity. First, one possibility is that each PET instance participates in no more than one relationship instance. Second, another possibility is if PET is a weak entity. (The formal definition of a weak entity is an entity that does not have its own primary key.) Suppose

5 PET was a strong entity, its own primary key would have to be used to only identify pets. It could not also be used to identify the relationship instances. For example, if a pet instance participated in more than one relationship instance, there would be a primary key violation in the strong PET entity. On the other hand, if PET is a weak entity, its partial key (not unique) can be combined with the key (either partial or unique) of the ternary relationship. In this case, PET can

10 absorb the ternary relationship.

[0175] Thus, without making additional assumptions, FIG. 23B cannot be morphed to resemble a series of binary relationships. Therefore, the focus turns to the general ternary relationship depicted in FIG. 23B.

[0176] FIG. 26 illustrates three-way relationships using the standard E-R notation. The schema for the example includes the PET, PET_TYPE, PET_OWNER entities plus one or more ternary relationships on them. Some additional information is also provided. In this example, model 2600 represents PET as an optional participant of the ternary relationship, as indicated by the cardinality range 0..m in the E-R diagram. The other two entities are both full participants of the relationship.

20 **[0177]** Suppose p, t, o are instances of PET, PET_TYPE, and PET_OWNER, respectively. The following are possible instances in the “Pet-of-Type-and-Owner” relationship:

- (p, t, o)
- (<missing>, t, o)

[0178] Here <missing> represents the absence of value from an entity. Whether these 25 candidate tuples are valid relationship instances is determined by a three-way join condition, defined as:

PET.type_id = PET_TYPE.id and PET.owner_id = PET_OWNER.id

[0179] Note: the value <missing> is able to match any other value. So in this example, the tuple (<missing>, t, o) is a valid instance of the relationship because the following condition evaluates to true.

<missing> = PET_TYPE.id and <missing> = PET_OWNER.id

5 **[0180]** The three example tables for the entities are created by the following statements.

create table PET (

id number,

name varchar2(30),

tid number,

10 oid number);

create table PET_TYPE (

id number,

name varchar2(30));

15

create table PET_OWNER (

id number,

name varchar2(30));

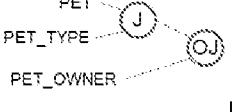
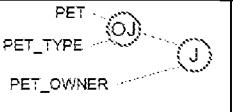
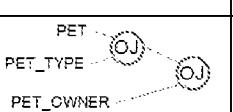
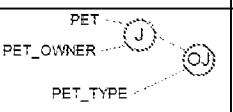
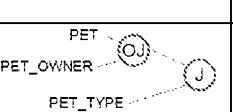
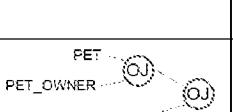
[0181] FIG. 27 depicts the rows in each table. As shown, each entity has only one instance. In 20 one aspect, a user can input the three-way join condition:

PET.type_id = PET_TYPE.id and PET.owner_id = PET_OWNER.id

[0182] The user can also mark entity PET as optional. This is equivalent to having input an E-R model shown in FIG. 26. One challenge is how to generate a SQL statement to best capture the meaning of FIG. 26. In various embodiments, a determination is made as to the syntax that

will provide a good multi-way join implementation. In the following example, ANSI join syntax is used.

[0183] Since each ANSI join is pair-wise, to join three tables, two joins are needed. Also, since PET is an optional entity, at least one of the two joins must be an outer join. Furthermore, 5 which two tables are joined first is also a factor to consider. Putting all these considerations together, nine permutations are encountered, corresponding to nine possible implementations for the multi-way join using ANSI syntax. These cases are drawn using data flow charts, and shown with their SQL statements and results in Table 1 below.

| | Flow Chart | SQL | Result | | | | | | | | | |
|-----|---|---|---|-----|------|-------|--|--|-------|--|--|------|
| 1 |  | <pre>select p.name "pet", t.name "type", o.name "owner" from PET p join PET_TYPE t on (p.tid = t.id) right outer join PET_OWNER o on (p.oid = o.id)</pre> | <table border="1"> <thead> <tr> <th>pet</th> <th>type</th> <th>owner</th> </tr> </thead> <tbody> <tr> <td></td> <td></td> <td>-----</td> </tr> <tr> <td></td> <td></td> <td>Jeff</td> </tr> </tbody> </table> <p>(Comment: Type is nullified. This is not a good result.)</p> | pet | type | owner | | | ----- | | | Jeff |
| pet | type | owner | | | | | | | | | | |
| | | ----- | | | | | | | | | | |
| | | Jeff | | | | | | | | | | |
| 2 |  | <pre>select p.name "pet", t.name "type", o.name "owner" from PET p right outer join PET_TYPE t on (p.tid = t.id) join PET_OWNER o on (p.oid = o.id)</pre> | <table border="1"> <thead> <tr> <th>pet</th> <th>type</th> <th>owner</th> </tr> </thead> <tbody> <tr> <td></td> <td></td> <td>-----</td> </tr> <tr> <td></td> <td></td> <td></td> </tr> </tbody> </table> <p>no rows selected</p> | pet | type | owner | | | ----- | | | |
| pet | type | owner | | | | | | | | | | |
| | | ----- | | | | | | | | | | |
| | | | | | | | | | | | | |
| 3 |  | <pre>select p.name "pet", t.name "type", o.name "owner" from PET p right outer join PET_TYPE t on (p.tid = t.id) right outer join PET_OWNER o on (p.oid = o.id)</pre> | <table border="1"> <thead> <tr> <th>pet</th> <th>type</th> <th>owner</th> </tr> </thead> <tbody> <tr> <td></td> <td></td> <td>-----</td> </tr> <tr> <td></td> <td></td> <td>Jeff</td> </tr> </tbody> </table> <p>(Comment: Type is nullified. This is not a good result.)</p> | pet | type | owner | | | ----- | | | Jeff |
| pet | type | owner | | | | | | | | | | |
| | | ----- | | | | | | | | | | |
| | | Jeff | | | | | | | | | | |
| 4 |  | <pre>select p.name "pet", t.name "type", o.name "owner" from PET p join PET_OWNER o on (p.oid = o.id) right outer join PET_TYPE t on (p.tid = t.id)</pre> | <table border="1"> <thead> <tr> <th>pet</th> <th>type</th> <th>owner</th> </tr> </thead> <tbody> <tr> <td></td> <td></td> <td>-----</td> </tr> <tr> <td></td> <td></td> <td>Cat</td> </tr> </tbody> </table> <p>(Comment: Owner is nullified. This is not a good result.)</p> | pet | type | owner | | | ----- | | | Cat |
| pet | type | owner | | | | | | | | | | |
| | | ----- | | | | | | | | | | |
| | | Cat | | | | | | | | | | |
| 5 |  | <pre>select p.name "pet", t.name "type", o.name "owner" from PET p right outer join PET_OWNER o on (p.oid = o.id) join PET_TYPE t on (p.tid = t.id)</pre> | <table border="1"> <thead> <tr> <th>pet</th> <th>type</th> <th>owner</th> </tr> </thead> <tbody> <tr> <td></td> <td></td> <td>-----</td> </tr> <tr> <td></td> <td></td> <td></td> </tr> </tbody> </table> <p>no rows selected</p> | pet | type | owner | | | ----- | | | |
| pet | type | owner | | | | | | | | | | |
| | | ----- | | | | | | | | | | |
| | | | | | | | | | | | | |
| 6 |  | <pre>select p.name "pet", t.name "type", o.name "owner" from PET p right outer join PET_OWNER o on (p.oid = o.id) right outer join PET_TYPE t on (p.tid = t.id)</pre> | <table border="1"> <thead> <tr> <th>pet</th> <th>type</th> <th>owner</th> </tr> </thead> <tbody> <tr> <td></td> <td></td> <td>-----</td> </tr> <tr> <td></td> <td></td> <td>Cat</td> </tr> </tbody> </table> <p>(Comment: Owner is nullified. This is not a good result.)</p> | pet | type | owner | | | ----- | | | Cat |
| pet | type | owner | | | | | | | | | | |
| | | ----- | | | | | | | | | | |
| | | Cat | | | | | | | | | | |

| 7 | | <pre>select p.name "pet", t.name "type", o.name "owner" from PET_TYPE t join PET_OWNER o on (1 = 1) left outer join PET p on (p.tid = t.id and p.oid = o.id)</pre> | <table border="1"> <thead> <tr> <th>pet</th> <th>type</th> <th>owner</th> </tr> </thead> <tbody> <tr> <td>-----</td> <td>Cat</td> <td>Jeff</td> </tr> </tbody> </table> <p>(Comment: This is the result we expect. Note that there is no direct join condition between PET_TYPE and PET_OWNER. It defaults to true.)</p> | pet | type | owner | ----- | Cat | Jeff |
|-------|------|--|--|-----|------|-------|-------|-----|------|
| pet | type | owner | | | | | | | |
| ----- | Cat | Jeff | | | | | | | |
| 8 | | <pre>select p.name "pet", t.name "type", o.name "owner" from PET_TYPE t full outer join PET_OWNER o on (1 = 1) join PET p on (p.tid = t.id and p.oid = o.id)</pre> | no rows selected | | | | | | |
| 9 | | <pre>select p.name "pet", t.name "type", o.name "owner" from PET_TYPE t full outer join PET_OWNER o on (1 = 1) left outer join PET p on (p.tid = t.id and p.oid = o.id)</pre> | <table border="1"> <thead> <tr> <th>pet</th> <th>type</th> <th>owner</th> </tr> </thead> <tbody> <tr> <td>-----</td> <td>Cat</td> <td>Jeff</td> </tr> </tbody> </table> <p>(Comment: The generated SQL is equivalent to the one in implementation #7.)</p> | pet | type | owner | ----- | Cat | Jeff |
| pet | type | owner | | | | | | | |
| ----- | Cat | Jeff | | | | | | | |

[0184] From the examination of all the possible implementations, implementation #7 appears to match the expectation for the three-way relationship. Accordingly, in general, multi-way relationship is not equivalent to a series of binary relationships. However, in various 5 embodiments, a multi-way relationship may be implemented using binary joins. Therefore, one aspect, a model is created that is understandable for (casual) human users yet rigorous for use in generating correct data flow implementations.

[0185] As discussed above, the equivalent data flow model fits in the category of “operational semantic models.” Operating semantic model to unambiguously describe the meaning/intention 10 of a system have been created. However, one for equivalently representing the E-R model provides new opportunities as discussed herein.

[0186] FIGS. 28A and 28B illustrate a three-way relationship in E-R notation and a data flow with data originating from the three entities in one embodiment. Using the PET example of FIG. 15 28A, FIG. 28B describes a data flow with data originating from the three entities. Each entity provides a set of tuples. Each tuple is composed of a list of columns/attributes. All the tuples go through three stages: connect, filter, and output phases, defined below.

[0187] The Connect phase: Performs Cartesian product of all the input entities. If an entity is an optional entity (defined shortly), a special tuple with all columns of value <missing> is first added as an extra member of the entity before the Cartesian product is carried out.

5 **[0188]** The Filter phase: In the filter phase, all the tuples coming out of the Connect phase are categorized into three groups:

- Group F includes tuples that fail the relationship condition.
- Group S1 includes tuples that satisfy the relationship condition, e.g.,

PET.tid = PET_TYPE.id and PET.oid = PET_OWNER.id

without comparing any <missing> values.

10

- Group S2 includes all other tuples that satisfy the relationship condition, but the supplemental value <missing> is used in comparison.

[0189] Intuitively, Group S1 includes rows that scored straight success. And group S2 passed the join condition because of ignorable missing values from an optional entity.

[0190] The Output phase: output final result, which is a set of tuples, using the following rule:

15

- All tuples from Group F are discarded.
- All tuples from Group S1 are included in a final result set.
- A tuple from Group S2 is included in the final result set only if it has material contribution to the final result.

20 **[0191]** A tuple is considered having no material contribution to the final result if it matches one of the tuples in the result set. In checking if two tuples match, we assume <missing> value matches any other value. For example the following two tuples match.

(‘ABC’, 123) vs (<missing>, 123)

[0192] Intuitively, the final output phase performs deduplication on tuples in Groups S1 and S2.

[0193] Using the example data in FIG. 27, the result of the Connect phase contains the following two tuples:

(pet_100, pet_type_1, pet_owner_10)

(<missing>, pet_type_1, pet_owner_10)

5 **[0194]** Here, we use pet_100 to represent the row in PET table with id = 100. Notice that the value <missing> is treated as a “valid” pet because PET is an optional entity.

[0195] In the second phase, the multi-way join condition

PET.tid = PET_TYPE.id and PET.oid = PET_OWNER.id

is evaluated. And only the tuple

10 (<missing>, pet_type_1, pet_owner_10)

satisfies the condition.

[0196] The final phase is trivial for this example since there is no need to do any deduplication.

15 **[0197]** The reason for the implementation #7 in Table 1 being able to return the correct result is it does Cartesian product of all tables before starting evaluating the join condition. It is the only implementation that is consistent with the defined operational semantic model defined. By making sure Cartesian product operation is well completed before rows are filtered in the data flow, the simultaneity property inherent in multi-way relationship is protected against the potentially destructive binary joins.

20 **[0198]** In some embodiments, users may be able to visually create models by drawing lines between entities that indicate the need for the Connect phase operation. For example, suppose a user only drew a connection between PET and PET_OWNER, but he entered a relationship condition as follows:

PET.tid = PET_TYPE.id and PET.oid = PET_OWNER.id

25 **[0199]** Upon seeing the above 3-way relationship condition, a connection may be automatically determined and created between PET and PET_TYPE. This is because the

Connect phase operation requires the Cartesian product of all entities involved. The join condition on this derived join is 1 = 1 only for achieving the Cartesian product.

[0200] Supposing also that the user drew an additional line, from PET_OWNER to PET_TYPE, forming a circle among the three entities. In one aspect, the new line that created the circle can be ignored, since all the entities in the relationship have been sufficiently connected. To a human user, he may think a line means “a binary relationship,” but by staying truthful to the operational semantic model FIG. 28B all the time: a line only means to connect entities together using Cartesian product.

[0201] After the entities are connected, the diagram can then be transformed into a tree of binary joins where the join node for PET and PET_TYPE carries the 1=1 condition. And the multi-way join condition is delayed to the last join node. In the whole process, the join condition is not dissembled, but rather, it is maximally delayed to ensure all the rows from all tables have opportunities to interact.

[0202] In contrast, if the join condition is split into two parts (which is syntactically permitted), and assigned the two subconditions to two join nodes, the operational semantic model would be violated because the Filter phase would be started before the Connect phase has completed.

[0203] Accordingly, since the operational semantic model is specified in a detailed, step-by-step manner, it can be easily transformed into a programmatic implementation using any existing programming languages. There is no need to only use SQL to implement it.

[0204] FIG. 29 depicts a diagram that lays out relationships among various database modeling methods and their semantic contents. In this example, the E-R model still needs help from semantic models. The CFO model is one such semantic model for eliminating ambiguities in E-R, particularly for multi-way relationships. As shown in FIG. 29, Object Oriented models can be used for the same purpose. Many patents exist for correctly converting E-R to OO models. But the OO model lacks the capability of integrating with data flow models. A data flow model explicitly spells out step-by-step operations of the data from source to target. An OO model is still too descriptive for that purpose. The CFO model is like an automaton, it is inherently suitable for integrating with data flow models.

[0205] Conclusion

[0206] FIG. 30 is a simplified block diagram of computer system 3000 that may be used to practice embodiments of the present invention. As shown in FIG. 30, computer system 3000 includes processor 3010 that communicates with a number of peripheral devices via bus subsystem 3020. These peripheral devices may include storage subsystem 3030, comprising 5 memory subsystem 3040 and file storage subsystem 3050, input devices 3060, output devices 3070, and network interface subsystem 3080.

[0207] Bus subsystem 3020 provides a mechanism for letting the various components and 10 subsystems of computer system 3000 communicate with each other as intended. Although bus subsystem 3020 is shown schematically as a single bus, alternative embodiments of the bus subsystem may utilize multiple busses.

[0208] Storage subsystem 3030 may be configured to store the basic programming and data constructs that provide the functionality of the present invention. Software (code modules or instructions) that provides the functionality of the present invention may be stored in storage subsystem 3030. These software modules or instructions may be executed by processor(s) 3010. 15 Storage subsystem 3030 may also provide a repository for storing data used in accordance with the present invention. Storage subsystem 3030 may comprise memory subsystem 3040 and file/disk storage subsystem 3050.

[0209] Memory subsystem 3040 may include a number of memories including a main random access memory (RAM) 3042 for storage of instructions and data during program execution and a 20 read only memory (ROM) 3044 in which fixed instructions are stored. File storage subsystem 3050 provides persistent (non-volatile) storage for program and data files, and may include a hard disk drive, a floppy disk drive along with associated removable media, a Compact Disk Read Only Memory (CD-ROM) drive, a DVD, an optical drive, removable media cartridges, and other like storage media.

25 [0210] Input devices 3060 may include a keyboard, pointing devices such as a mouse, trackball, touchpad, or graphics tablet, a scanner, a barcode scanner, a touchscreen incorporated into the display, audio input devices such as voice recognition systems, microphones, and other types of input devices. In general, use of the term “input device” is intended to include all possible types of devices and mechanisms for inputting information to computer system 3000.

[0211] Output devices 3070 may include a display subsystem, a printer, a fax machine, or non-visual displays such as audio output devices, etc. The display subsystem may be a cathode ray tube (CRT), a flat-panel device such as a liquid crystal display (LCD), or a projection device. In general, use of the term “output device” is intended to include all possible types of devices and 5 mechanisms for outputting information from computer system 3000.

[0212] Network interface subsystem 3080 provides an interface to other computer systems, devices, and networks, such as communications network 3090. Network interface subsystem 3080 serves as an interface for receiving data from and transmitting data to other systems from computer system 3000. Some examples of communications network 3090 are private networks, 10 public networks, leased lines, the Internet, Ethernet networks, token ring networks, fiber optic networks, and the like.

[0213] Computer system 3000 can be of various types including a personal computer, a portable computer, a workstation, a network computer, a mainframe, a kiosk, or any other data processing system. Due to the ever-changing nature of computers and networks, the description 15 of computer system 3000 depicted in FIG. 30 is intended only as a specific example for purposes of illustrating the preferred embodiment of the computer system. Many other configurations having more or fewer components than the system depicted in FIG. 30 are possible.

[0214] FIG. 31 is a simplified block diagram of data integration system 3100 for facilitating generation of a data mapping in accordance with an embodiment of the present invention. The 20 blocks of the data integration system 3100 may be implemented by hardware, software, or a combination of hardware and software to carry out the principles of the invention. It is understood by persons of skill in the art that the blocks described in Figure 31 may be combined or separated into sub-blocks to implement the principles of the invention as described above. Therefore, the description herein may support any possible combination or separation or further 25 definition of the functional blocks described herein.

[0215] As shown in FIG. 31, the data integration system 3100 is shown containing a receiving unit 3110, a determining unit 3220, and a generating unit 3130. Optionally, the data integration system 3100 can also comprise a deriving unit 3140 and a exporting unit 3150.

[0216] In one embodiment, the receiving unit 3110 is configured to receive information specifying a set of entity relationships as a component of the logical design. The determining unit 3120 is configured to determine an equivalent data flow model based on the set of entity relationships. The generating unit 3130 is configured to generate information indicative of the equivalent data flow model in the logical flow design.

[0217] In one aspect of the embodiment, the deriving unit 3140 is configured to derive one or more attributes of a dataset representing the set of entity relationships based on information declaring relationships between attributes of data sources. In one aspect of the embodiment, the receiving unit 3110 is further configured to receive information specifying one or more components of the logical design that includes information indicative of an operation that changes shape of the information flowing through the logical design.

[0218] In one aspect of the embodiment, the receiving unit 3110 is further configured to receive information specifying one or more components of the logical design that includes information indicative of an operation that controls the flow of information flowing through the logical design but does not change shape of the information flowing through the logical design. In one aspect of the embodiment, the receiving unit 3110 is further configured to receive information specifying one or more components of the logical design that includes information indicative of a target component having one or more attributes of data to be stored in a target data store.

[0219] In one aspect of the embodiment, the generating unit 3130 comprises exporting unit 3150 configured to export a list of attributes to a downstream component. In one aspect of the embodiment, the receiving unit 3110 is further configured to receive a change in the logical design through the introduction of one or more relationships; and the determining unit 3120 is further configured to determine an updated equivalent data flow model.

[0220] Although specific embodiments of the invention have been described, various modifications, alterations, alternative constructions, and equivalents are also encompassed within the scope of the invention. The described invention is not restricted to operation within certain specific data processing environments, but is free to operate within a plurality of data processing environments. Additionally, although the present invention has been described using a particular

series of transactions and steps, it should be apparent to those skilled in the art that the scope of the present invention is not limited to the described series of transactions and steps.

[0221] Further, while the present invention has been described using a particular combination of hardware and software, it should be recognized that other combinations of hardware and 5 software are also within the scope of the present invention. The present invention may be implemented only in hardware, or only in software, or using combinations thereof.

[0222] The specification and drawings are, accordingly, to be regarded in an illustrative rather than a restrictive sense. It will, however, be evident that additions, subtractions, deletions, and 10 other modifications and changes may be made thereunto without departing from the broader spirit and scope of the invention as set forth in the claims.

[0223] Various embodiments of any of one or more inventions whose teachings may be presented within this disclosure can be implemented in the form of logic in software, firmware, hardware, or a combination thereof. The logic may be stored in or on a machine-accessible memory, a machine-readable article, a tangible computer-readable medium, a computer-readable 15 storage medium, or other computer/machine-readable media as a set of instructions adapted to direct a central processing unit (CPU or processor) of a logic machine to perform a set of steps that may be disclosed in various embodiments of an invention presented within this disclosure. The logic may form part of a software program or computer program product as code modules become operational with a processor of a computer system or an information-processing device 20 when executed to perform a method or process in various embodiments of an invention presented within this disclosure. Based on this disclosure and the teachings provided herein, a person of ordinary skill in the art will appreciate other ways, variations, modifications, alternatives, and/or methods for implementing in software, firmware, hardware, or combinations thereof any of the disclosed operations or functionalities of various embodiments of one or more of the presented 25 inventions.

[0224] The disclosed examples, implementations, and various embodiments of any one of those inventions whose teachings may be presented within this disclosure are merely illustrative to convey with reasonable clarity to those skilled in the art the teachings of this disclosure. As 30 these implementations and embodiments may be described with reference to exemplary illustrations or specific figures, various modifications or adaptations of the methods and/or

specific structures described can become apparent to those skilled in the art. All such modifications, adaptations, or variations that rely upon this disclosure and these teachings found herein, and through which the teachings have advanced the art, are to be considered within the scope of the one or more inventions whose teachings may be presented within this disclosure.

5 Hence, the present descriptions and drawings should not be considered in a limiting sense, as it is understood that an invention presented within a disclosure is in no way limited to those embodiments specifically illustrated.

[0225] Accordingly, the above description and any accompanying drawings, illustrations, and figures are intended to be illustrative but not restrictive. The scope of any invention presented 10 within this disclosure should, therefore, be determined not with simple reference to the above description and those embodiments shown in the figures, but instead should be determined with reference to the pending claims along with their full scope or equivalents.

WHAT IS CLAIMED IS:

- 1 1. A method facilitating generation of a data mapping, the method
2 comprising:
 - 3 receiving, at the one or more computer systems, information specifying a set of
4 entity relationships as a component of the logical design;
 - 5 determining, with one or more processors associated with the one or more
6 computer systems, an equivalent data flow model based on the set of entity relationships; and
 - 7 generating, with the one or more processors associated with the one or more
8 computer systems, information indicative of the equivalent data flow model in the logical flow
9 design.
- 1 2. The method of claim 1 further comprising deriving one or more attributes
2 of a dataset representing the set of entity relationships based on information declaring
3 relationships between attributes of data sources.
- 1 3. The method of claim 1 or 2 further comprising receiving information
2 specifying one or more components of the logical design that includes information indicative of
3 an operation that changes shape of the information flowing through the logical design.
- 1 4. The method of any of claims 1 to 3 further comprising receiving
2 information specifying one or more components of the logical design that includes information
3 indicative of an operation that controls the flow of information flowing through the logical
4 design but does not change shape of the information flowing through the logical design.
- 1 5. The method of any of claims 1 to 4 further comprising receiving
2 information specifying one or more components of the logical design that includes information
3 indicative of a target component having one or more attributes of data to be stored in a target
4 datastore.
- 1 6. The method of any of claims 1 to 5 wherein generating the information
2 indicative of the equivalent data flow model in the logical flow design comprises exporting a list
3 of attributes to a downstream component.

1 7. The method of any of claims 1 to 6 further comprising:

2 receiving, at the one or more computer systems, a change in the logical design

3 through the introduction of one or more relationships; and

4 determining, with the one or more processors associated with the one or more

5 computer systems, an updated equivalent data flow model.

1 8. A non-transitory computer-readable medium storing computer-executable

2 code for facilitating generation of a data mapping, the non-transitory computer-readable medium

3 comprising:

4 code for receiving information specifying a set of entity relationships as a

5 component of the logical design;

6 code for determining an equivalent data flow model based on the set of entity

7 relationships; and

8 code for generating information indicative of the equivalent data flow model in

9 the logical flow design.

1 9. The non-transitory computer-readable medium of claim 8 or 9 further

2 comprising code for deriving one or more attributes of a dataset representing the set of entity

3 relationships based on information declaring relationships between attributes of data sources.

1 10. The non-transitory computer-readable medium of any of claims 8 to 9

2 further comprising code for receiving information specifying one or more components of the

3 logical design that includes information indicative of an operation that changes shape of the

4 information flowing through the logical design.

1 11. The non-transitory computer-readable medium of any of claims 8 to 10

2 further comprising code for receiving information specifying one or more components of the

3 logical design that includes information indicative of an operation that controls the flow of

4 information flowing through the logical design but does not change shape of the information

5 flowing through the logical design.

1 12. The non-transitory computer-readable medium of any of claims 8 to 11

2 further comprising code for receiving information specifying one or more components of the

3 logical design that includes information indicative of a target component having one or more
4 attributes of data to be stored in a target datastore.

1 13. The non-transitory computer-readable medium of any of claims 8 to 12
2 wherein the code for generating the information indicative of the equivalent data flow model in
3 the logical flow design comprises code for exporting a list of attributes to a downstream
4 component.

1 14. The non-transitory computer-readable medium of any of claims 8 to 13
2 further comprising:

3 code for receiving a change in the logical design through the introduction of one
4 or more relationships; and

5 code for determining an updated equivalent data flow model.

1 15. A system facilitating generation of a data mapping, the system
2 comprising:

3 a processor; and

4 a memory storing instructions which when executed by the processor configure
5 the processor to:

6 receive information specifying a set of entity relationships as a component
7 of the logical design;

8 determine an equivalent data flow model based on the set of entity
9 relationships; and

10 generate information indicative of the equivalent data flow model in the
11 logical flow design.

1 16. The system of claim 15 wherein the processor is further configured to
2 derive one or more attributes of a dataset representing the set of entity relationships based on
3 information declaring relationships between attributes of data sources.

1 17. The system of claim 15 or 16 wherein the processor is further configured
2 to receive information specifying one or more components of the logical design that includes

3 information indicative of an operation that changes shape of the information flowing through the
4 logical design.

1 18. The system of any of claims 15 to 17 wherein the processor is further
2 configured to receive information specifying one or more components of the logical design that
3 includes information indicative of an operation that controls the flow of information flowing
4 through the logical design but does not change shape of the information flowing through the
5 logical design.

1 19. The system of any of claims 15 to 18 wherein the processor is further
2 configured to receive information specifying one or more components of the logical design that
3 includes information indicative of a target component having one or more attributes of data to be
4 stored in a target datastore.

1 20. The system of any of claims 15 to 19 wherein to generate the information
2 indicative of the equivalent data flow model in the logical flow design the processor is
3 configured to export a list of attributes to a downstream component.

1 21. The system of any of claims 15 to 20 wherein the processor is further
2 configured to:

3 receive a change in the logical design through the introduction of one or more
4 relationships; and

5 determine an updated equivalent data flow model.

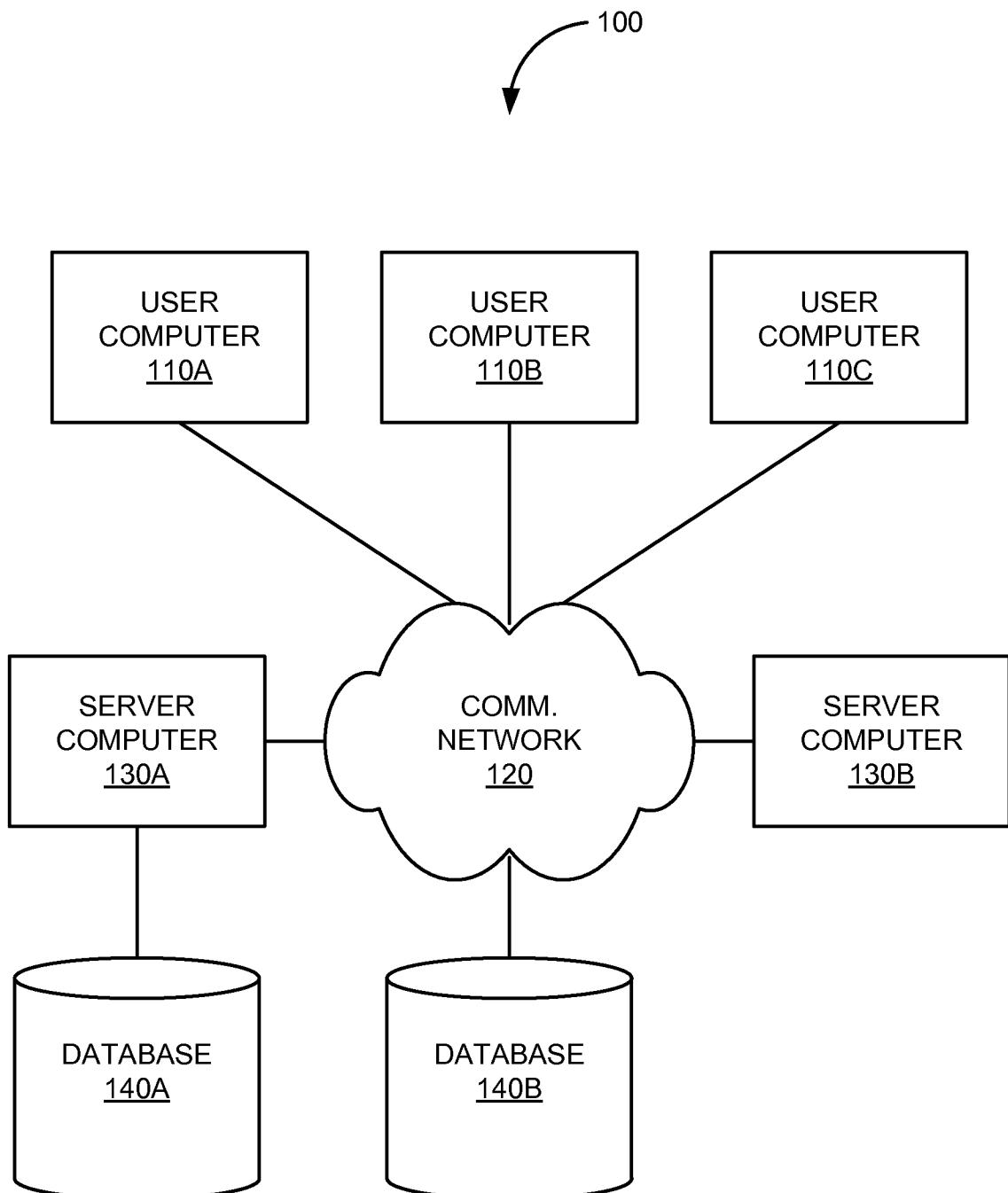


FIG. 1

200 ↘ 2 / 32

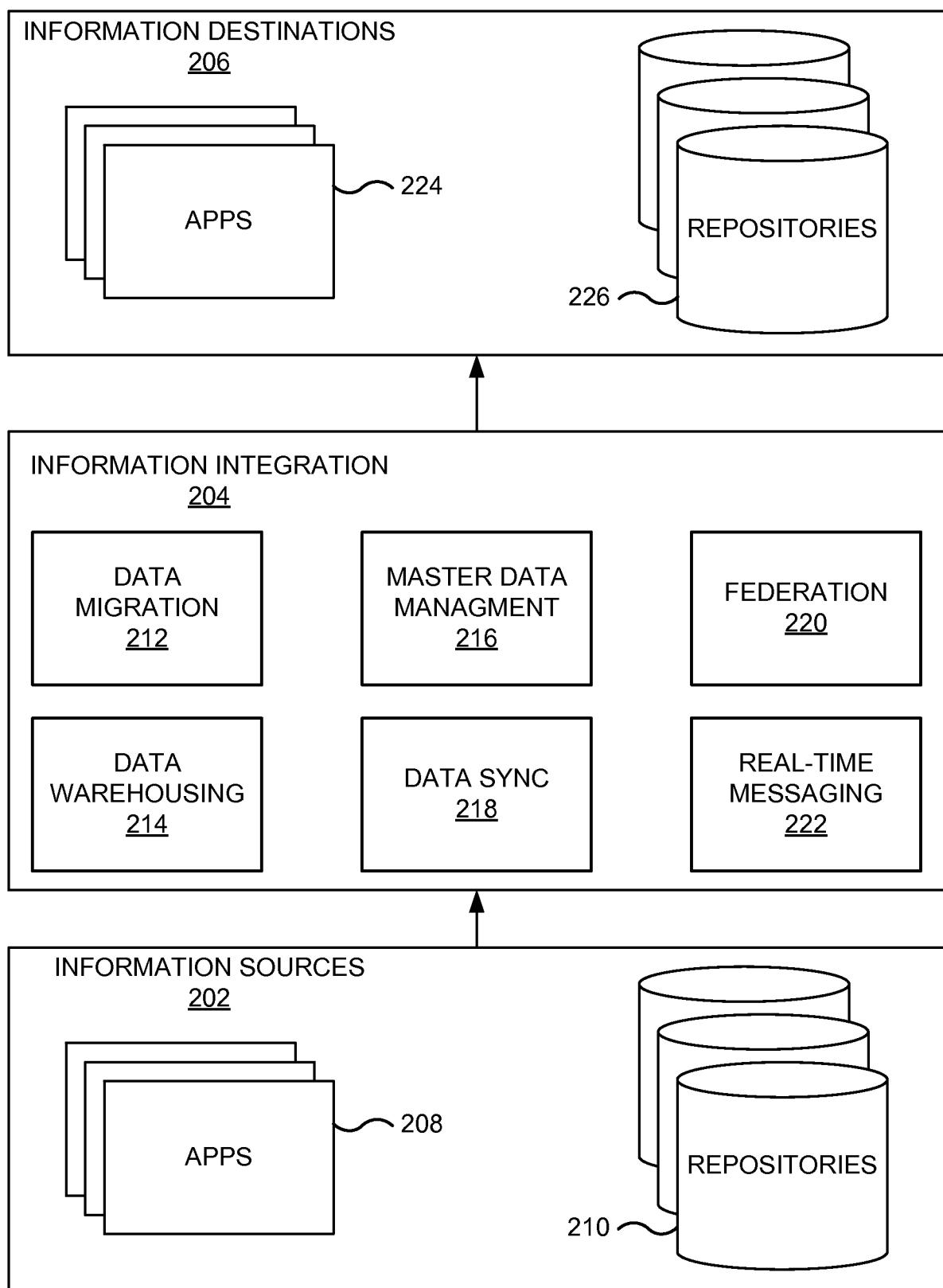


FIG. 2

200 ↗

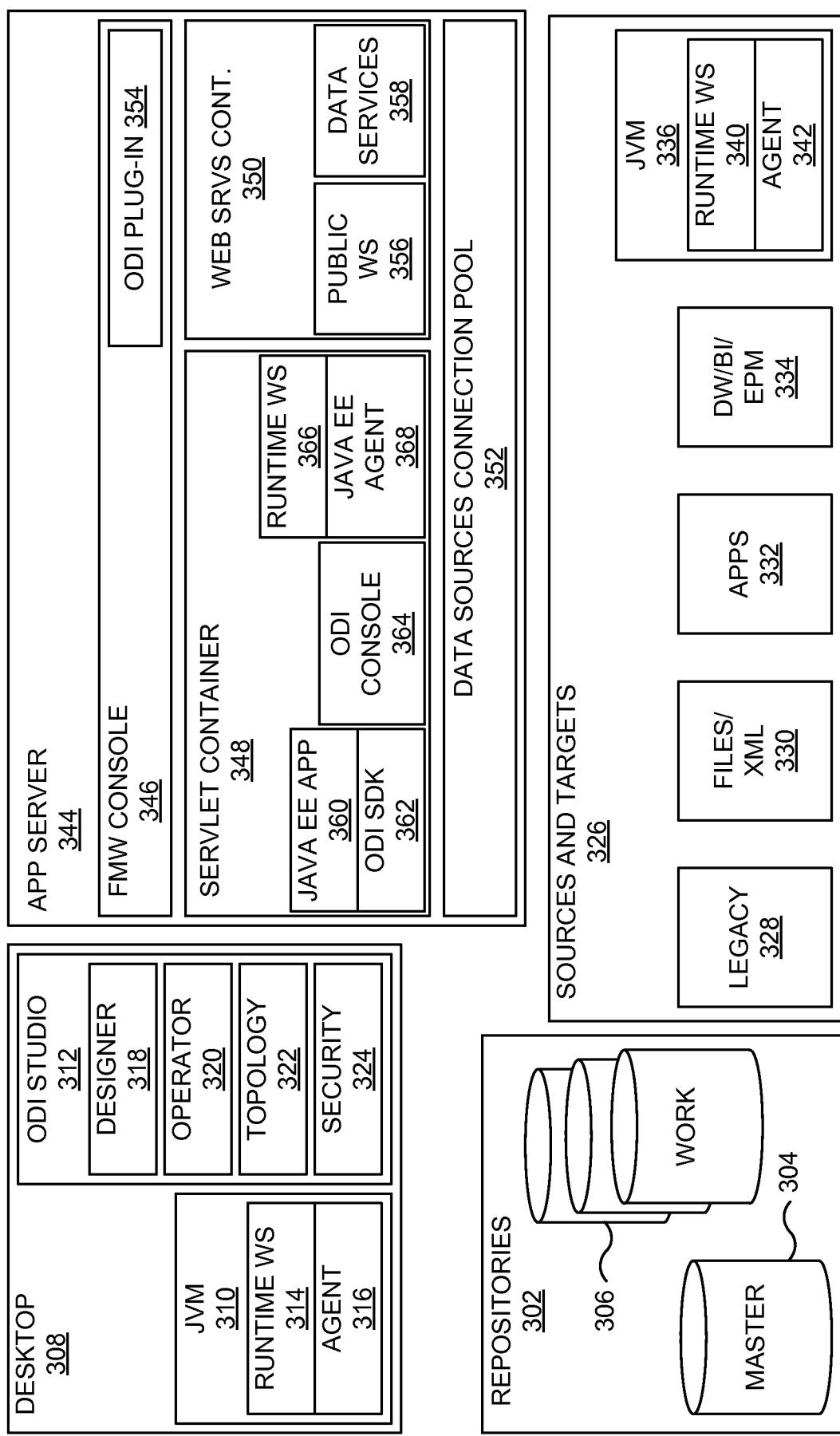


FIG. 3

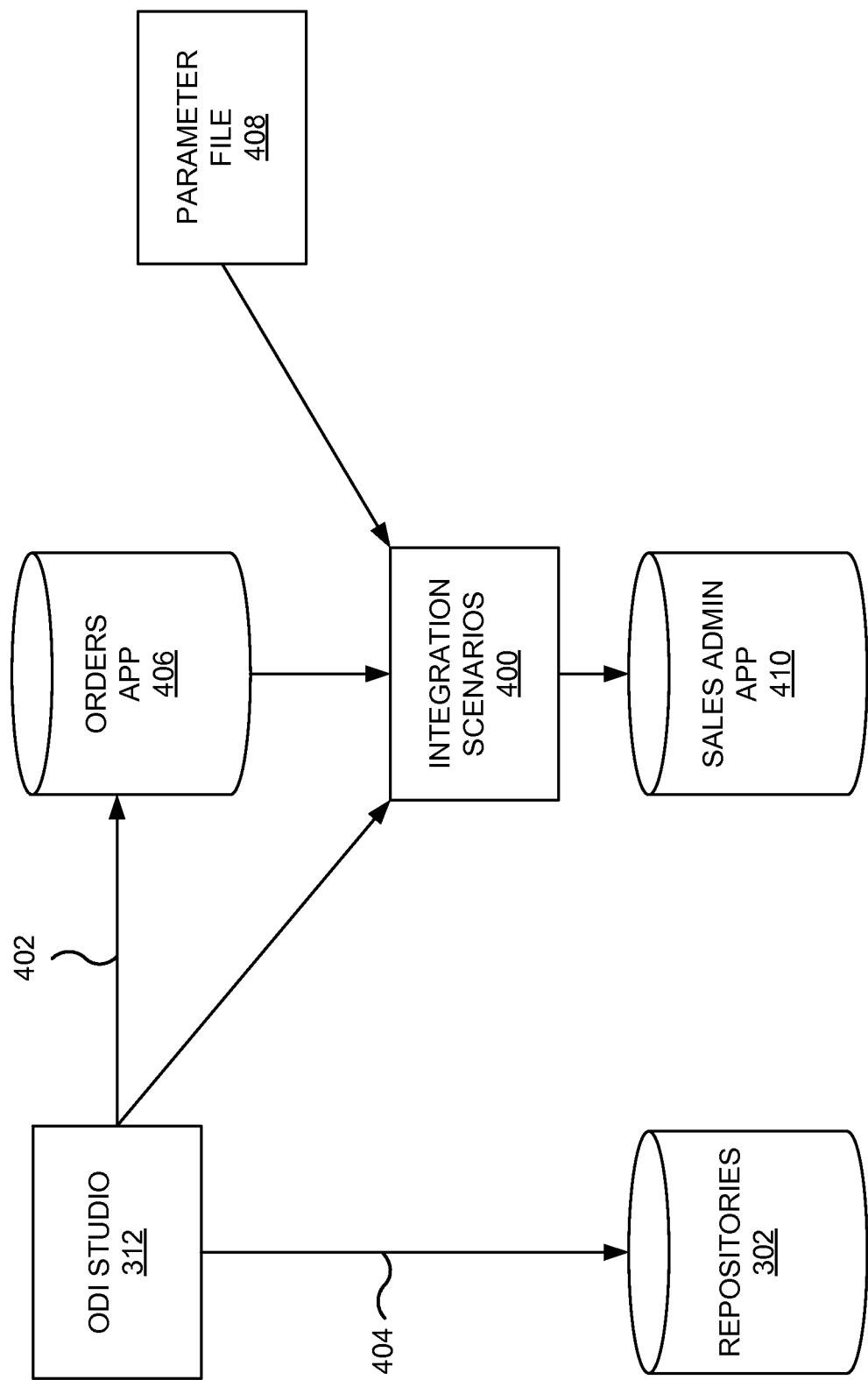


FIG. 4

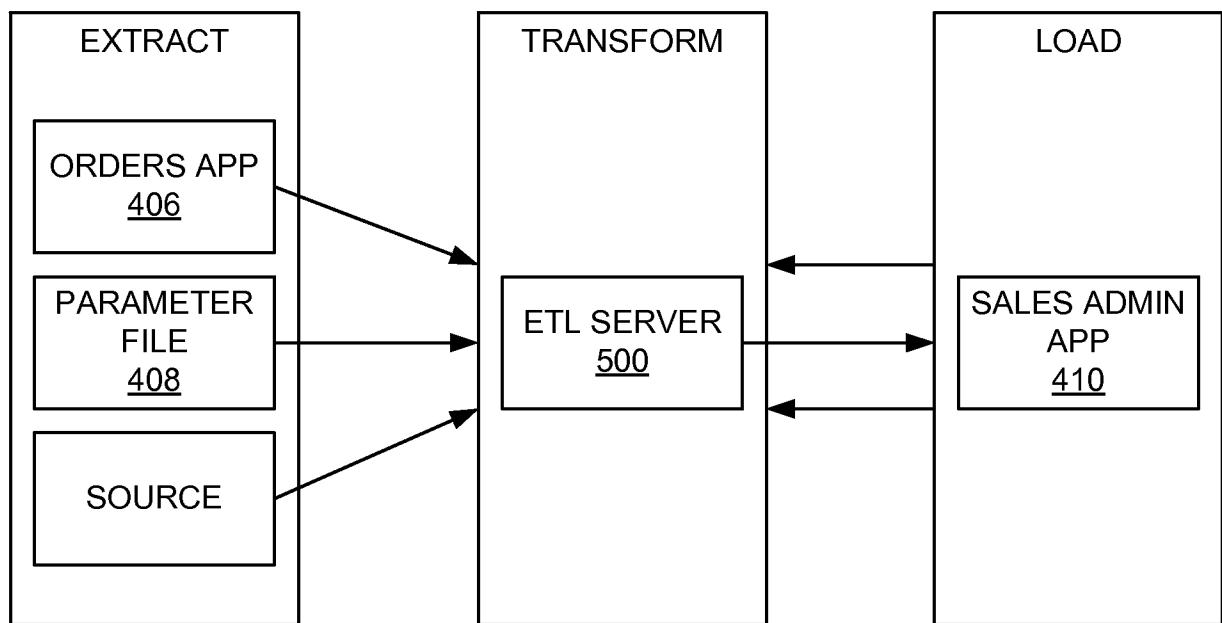


FIG. 5A

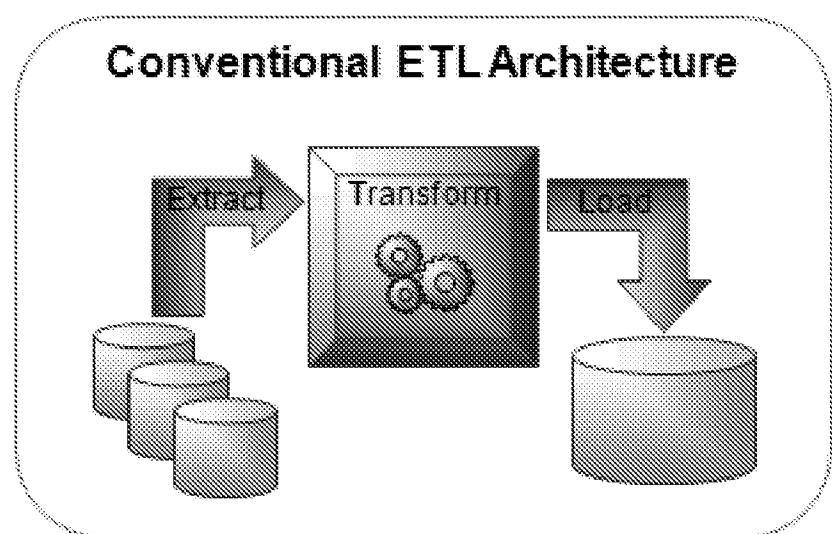


FIG. 5B

6 / 32

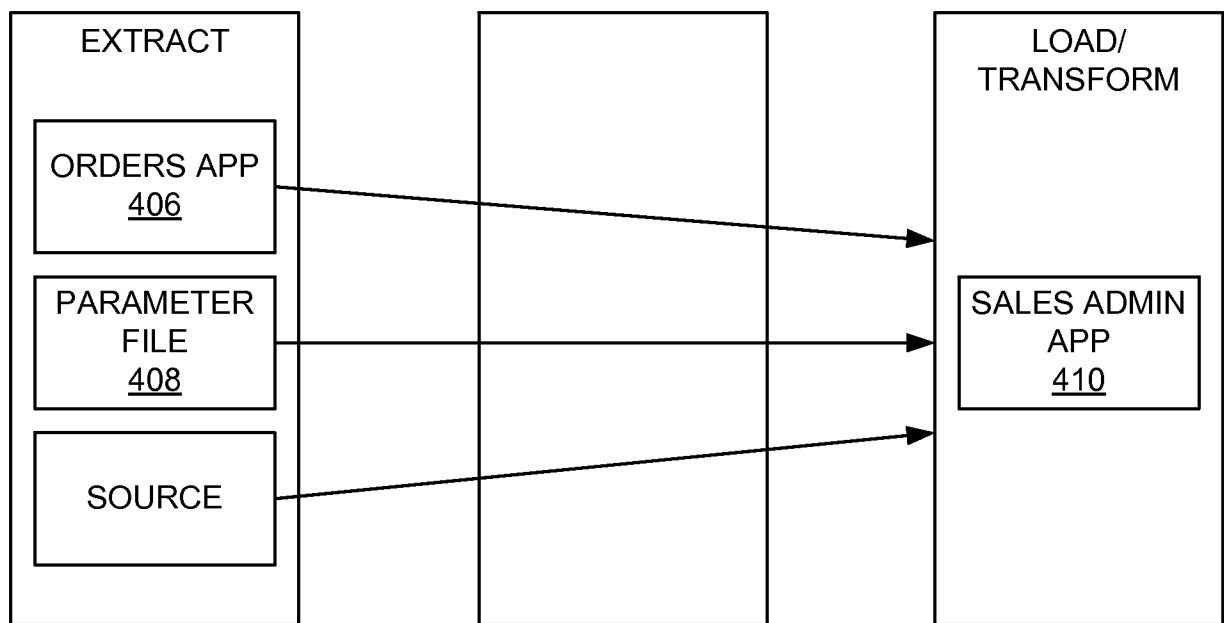


FIG. 6A

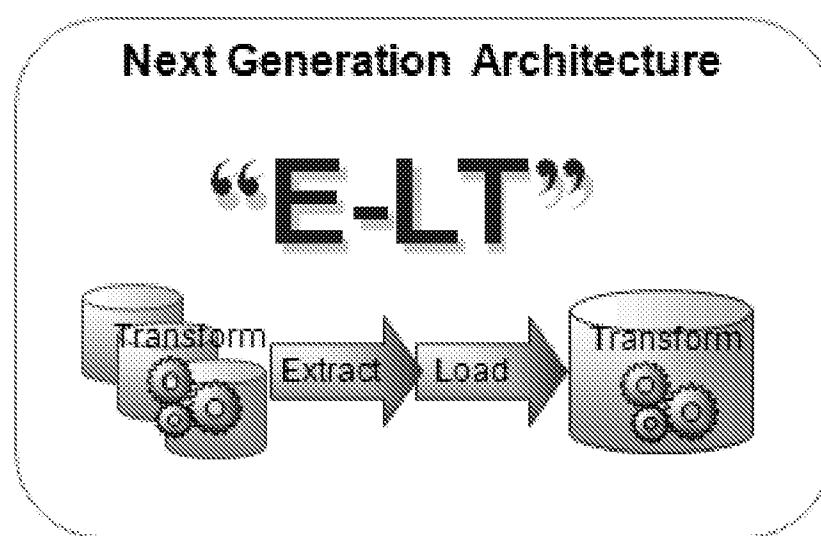


FIG. 6B

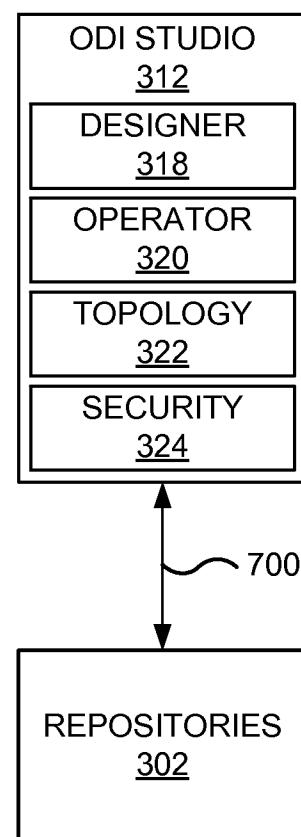


FIG. 7

8 / 32

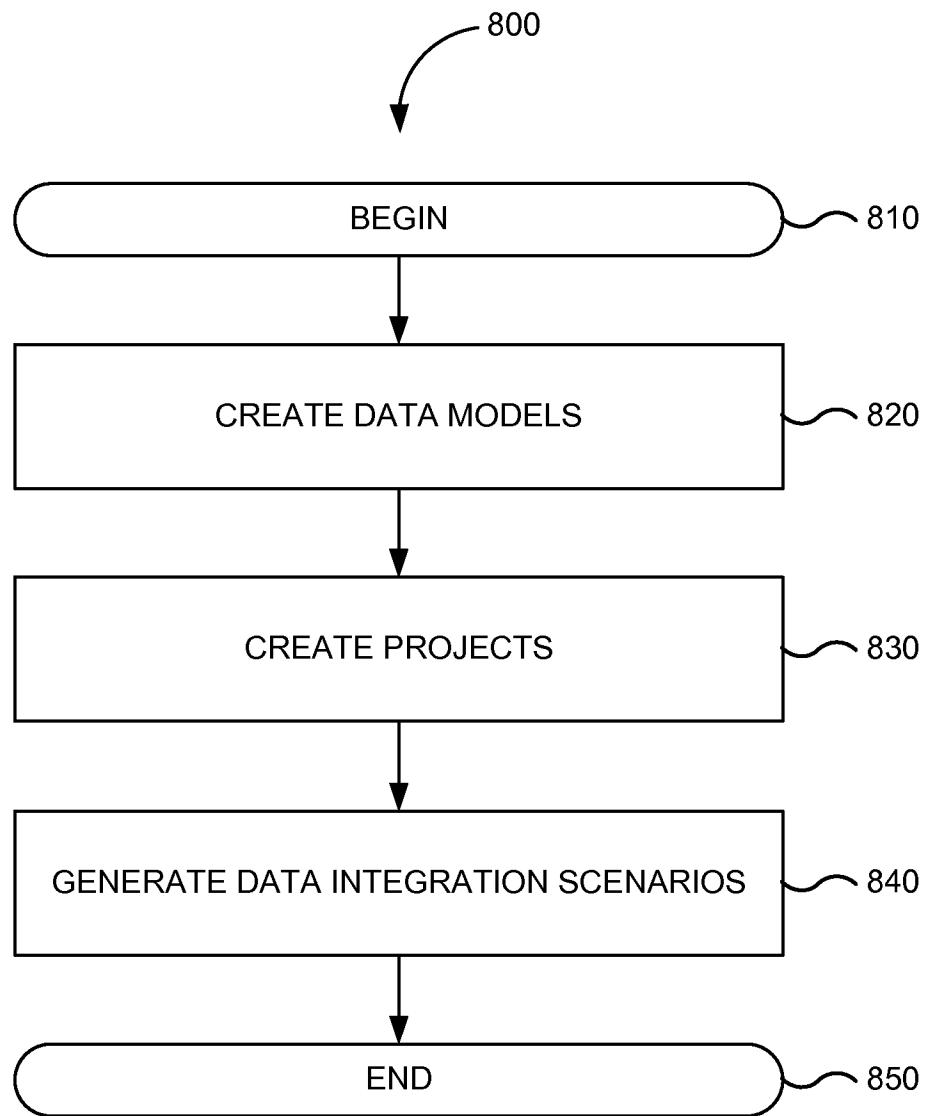


FIG. 8

9 / 32

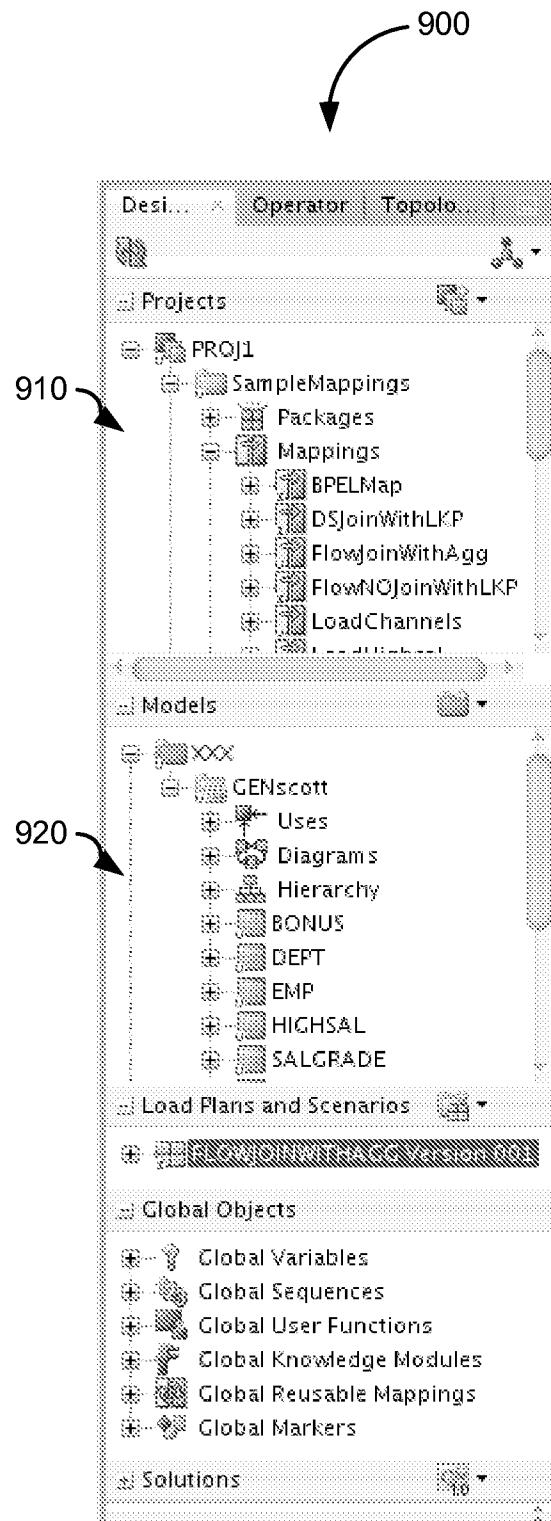


FIG. 9

10 / 32

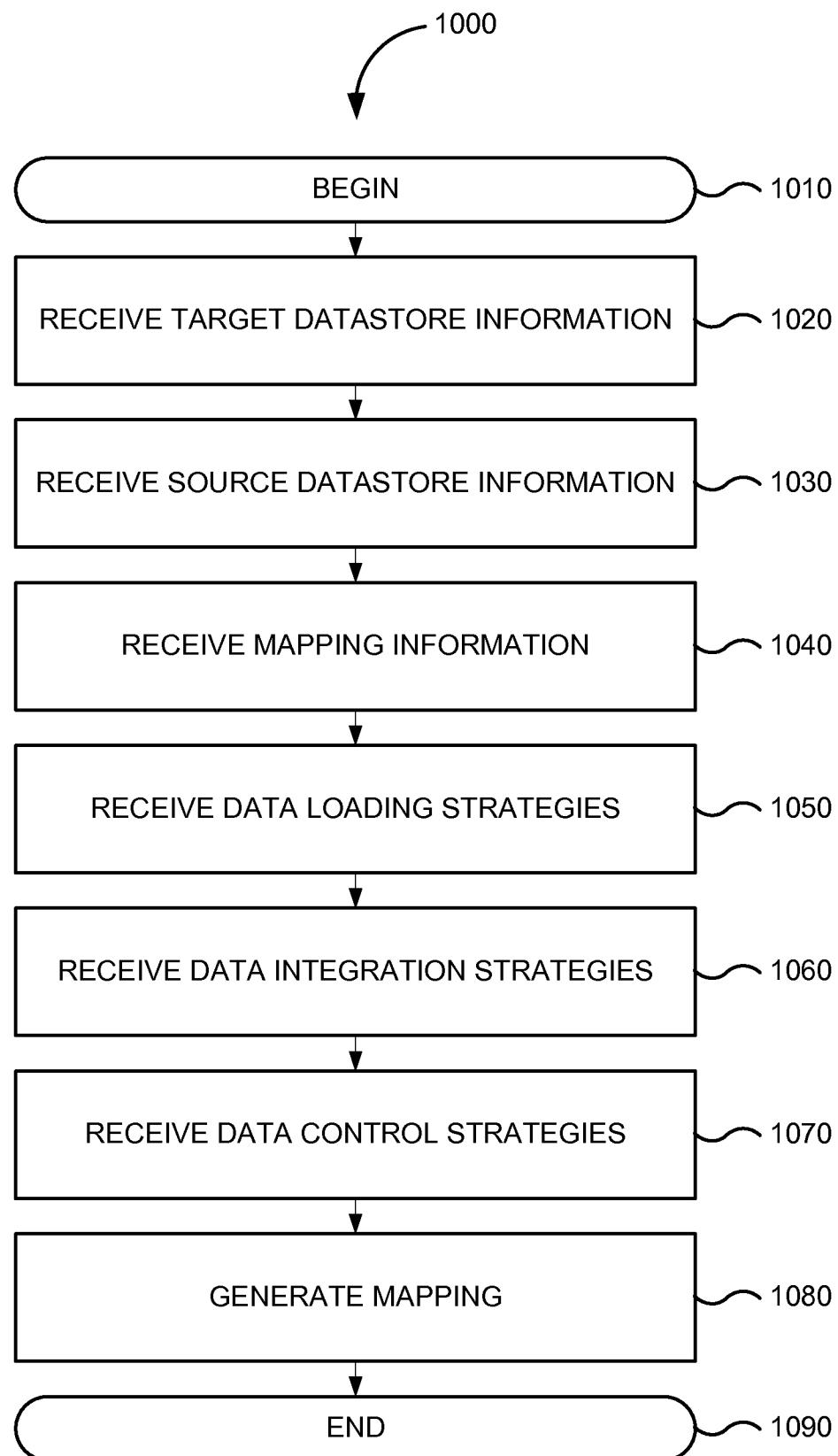


FIG. 10

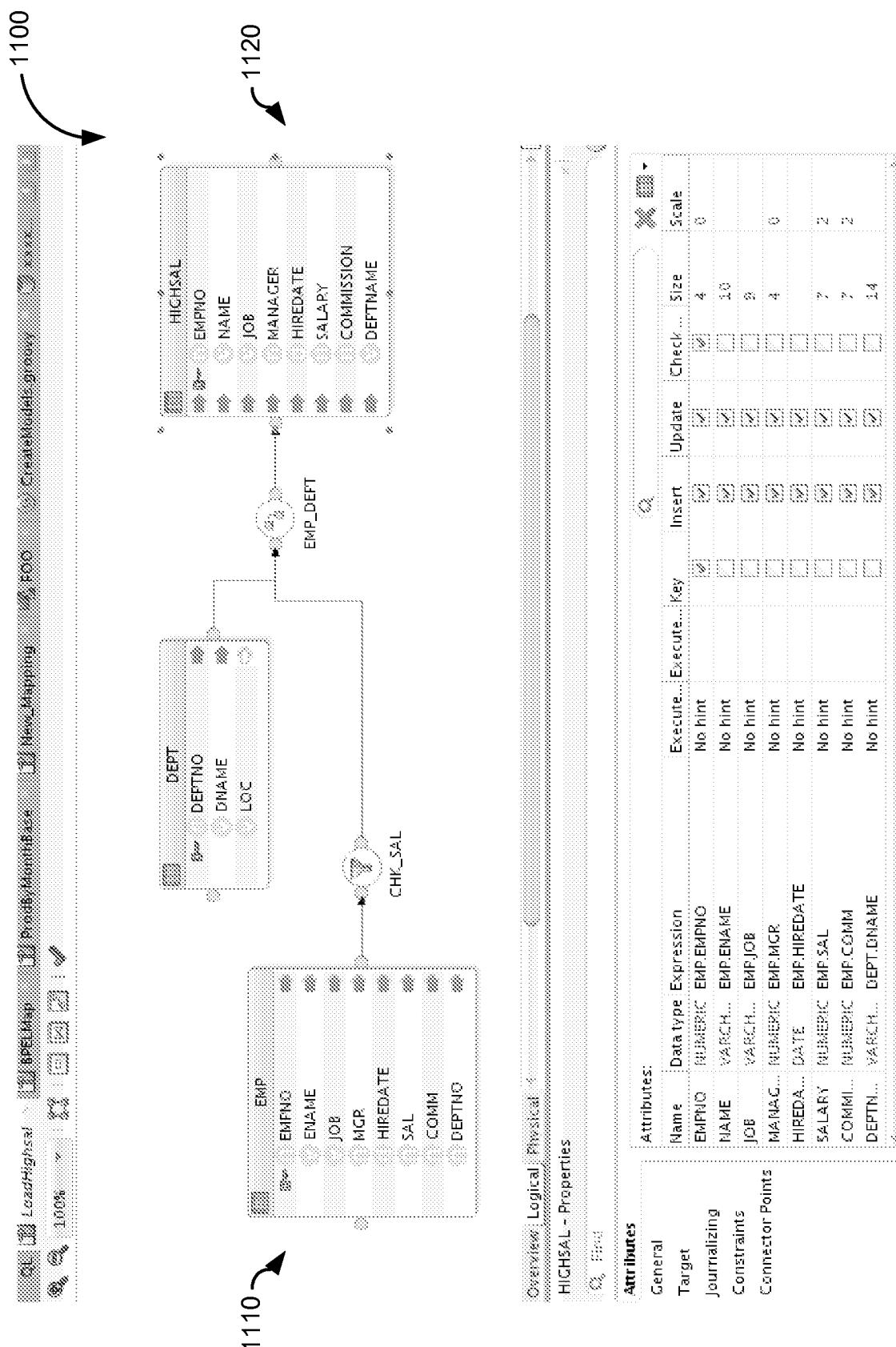


FIG. 11

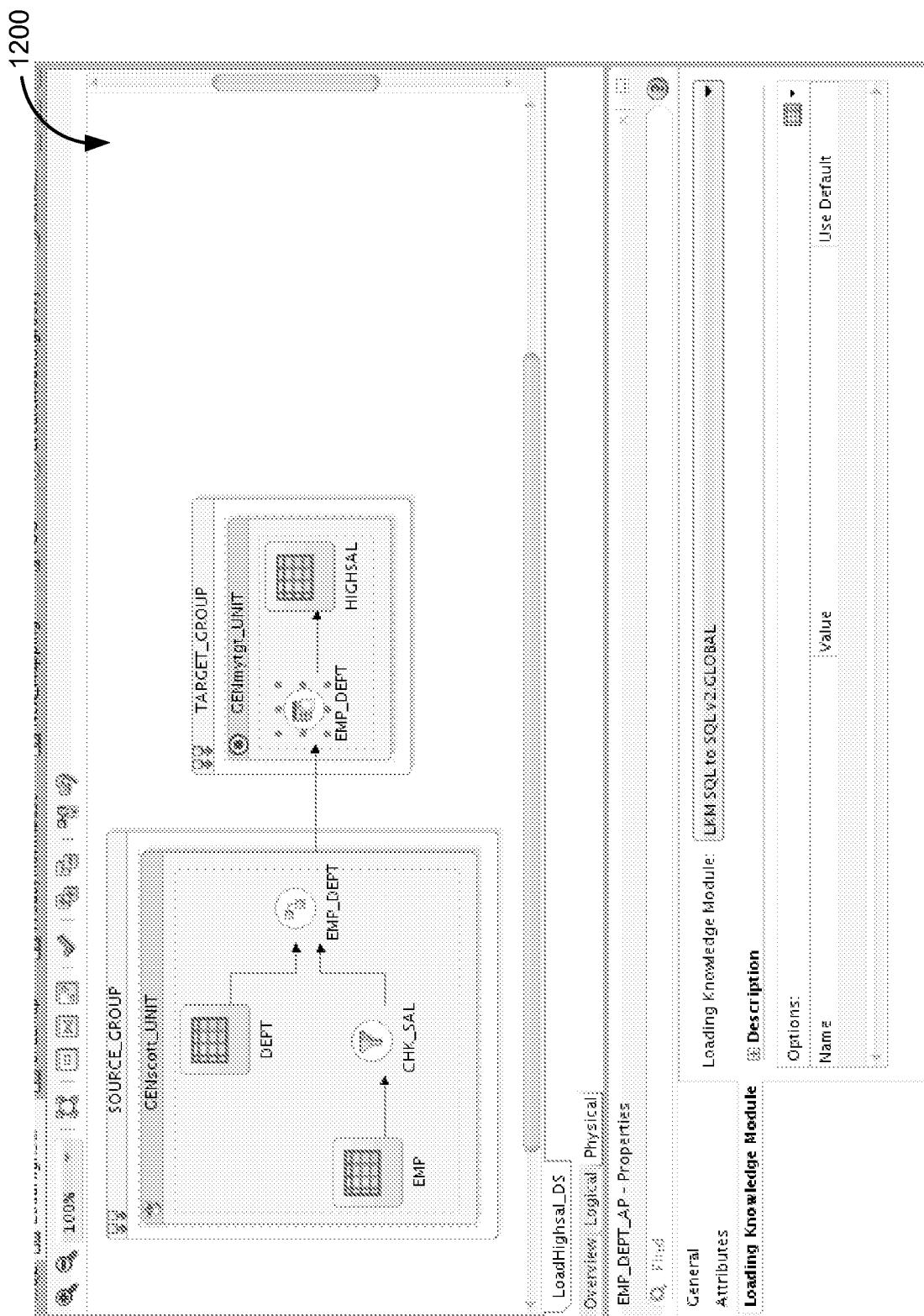


FIG. 12

13 / 32

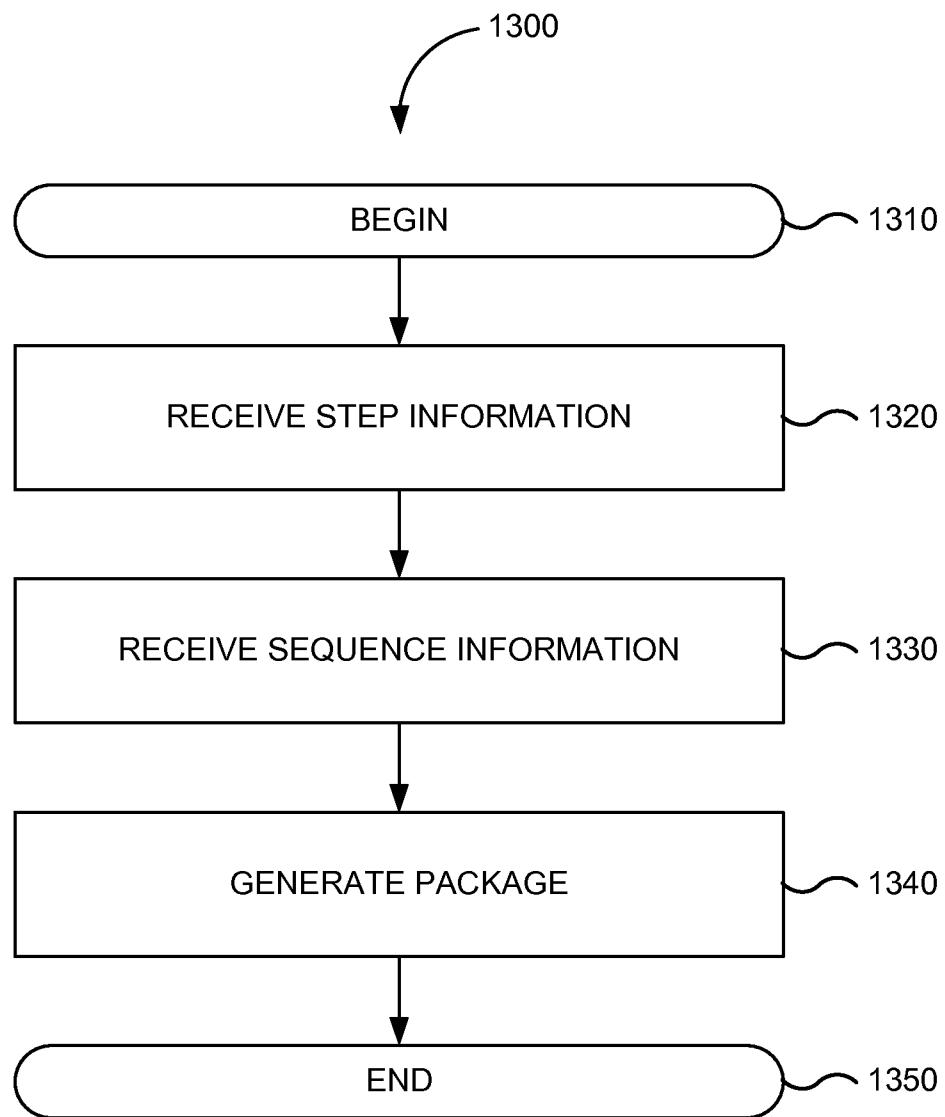


FIG. 13

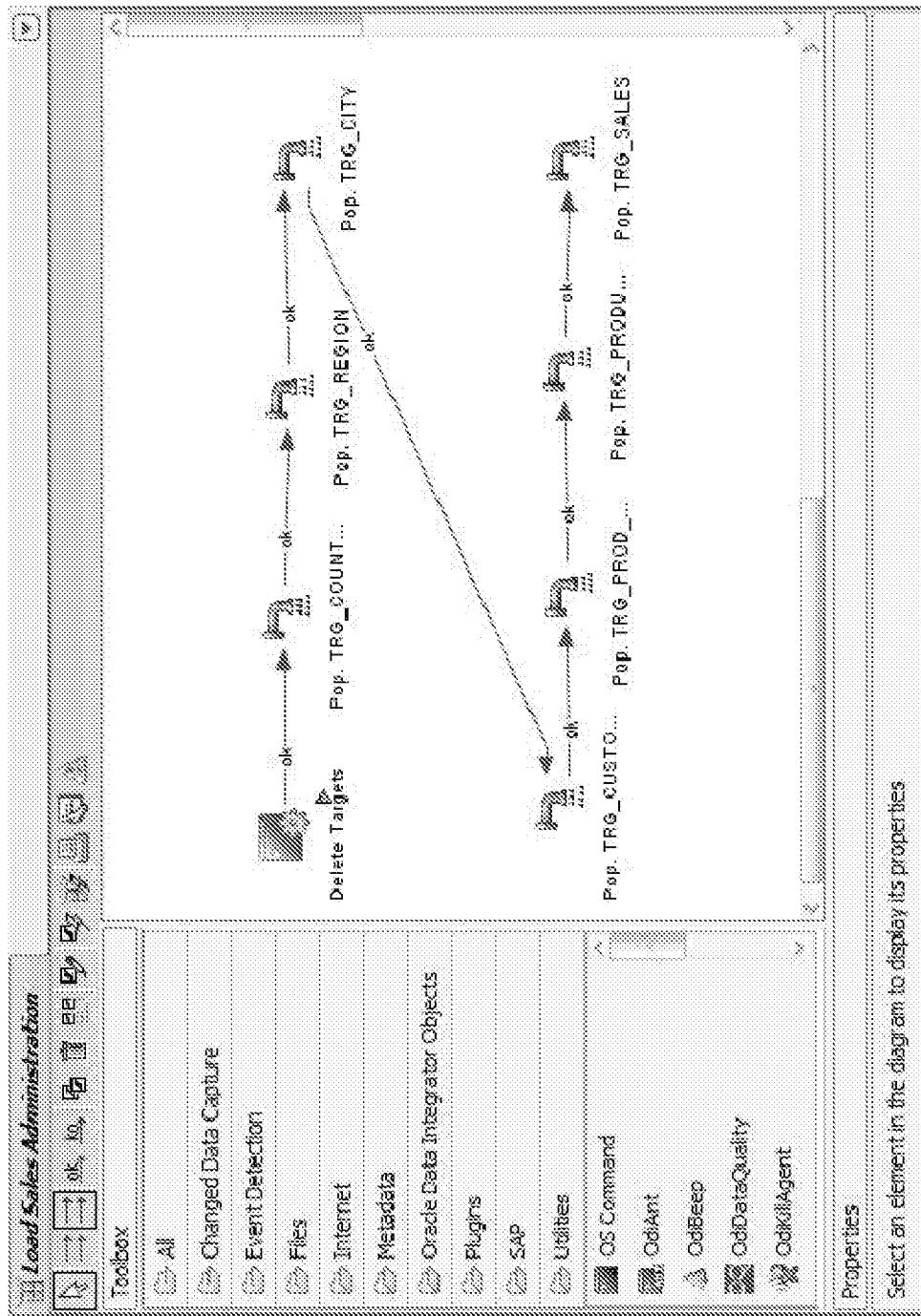


FIG. 14

15 / 32

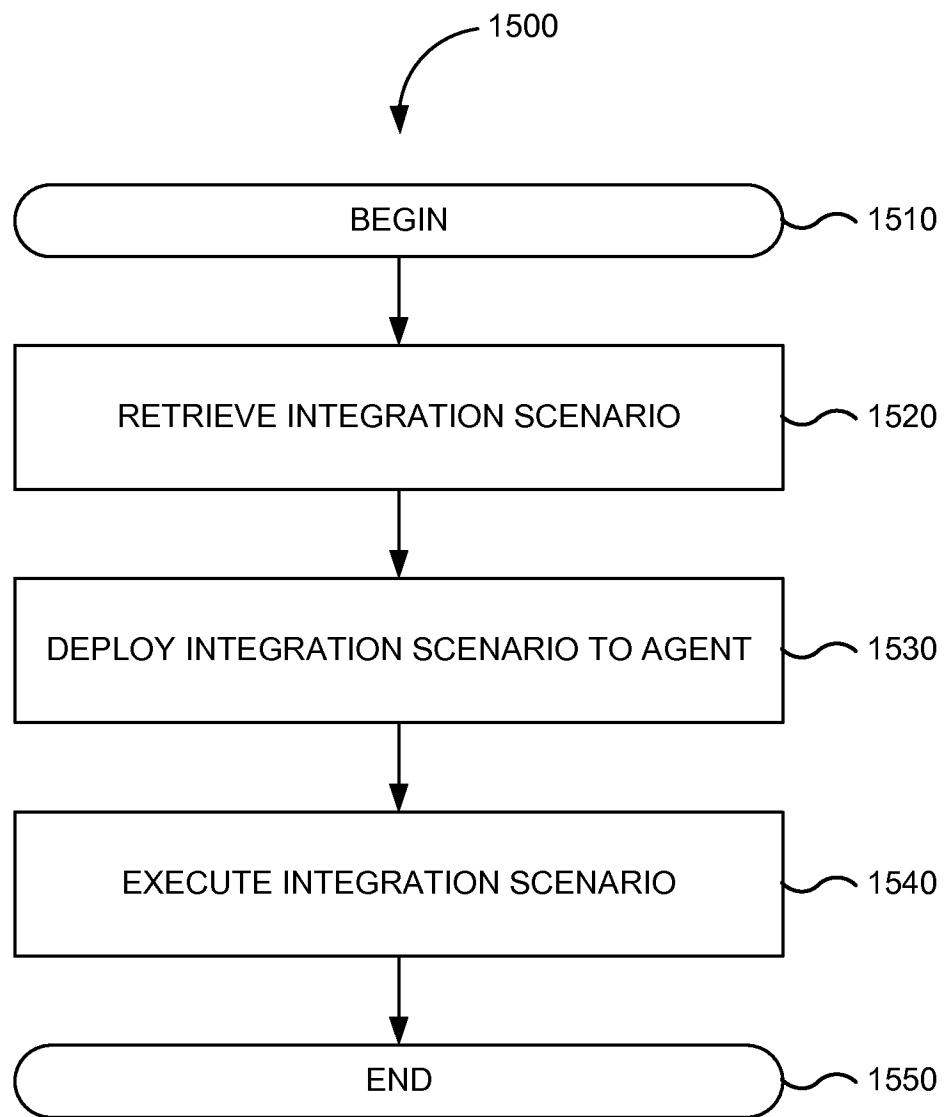
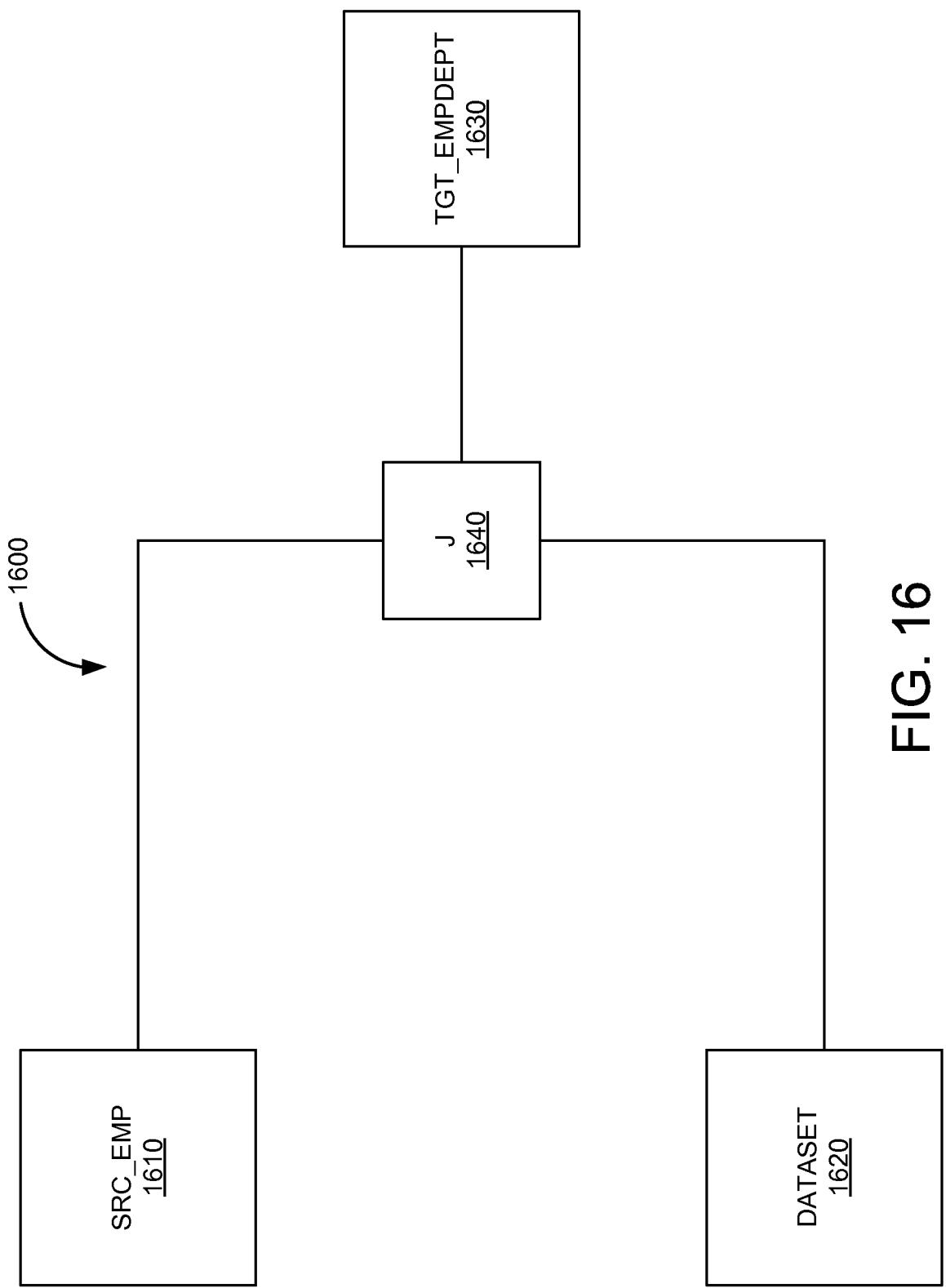


FIG. 15

**FIG. 16**

17 / 32

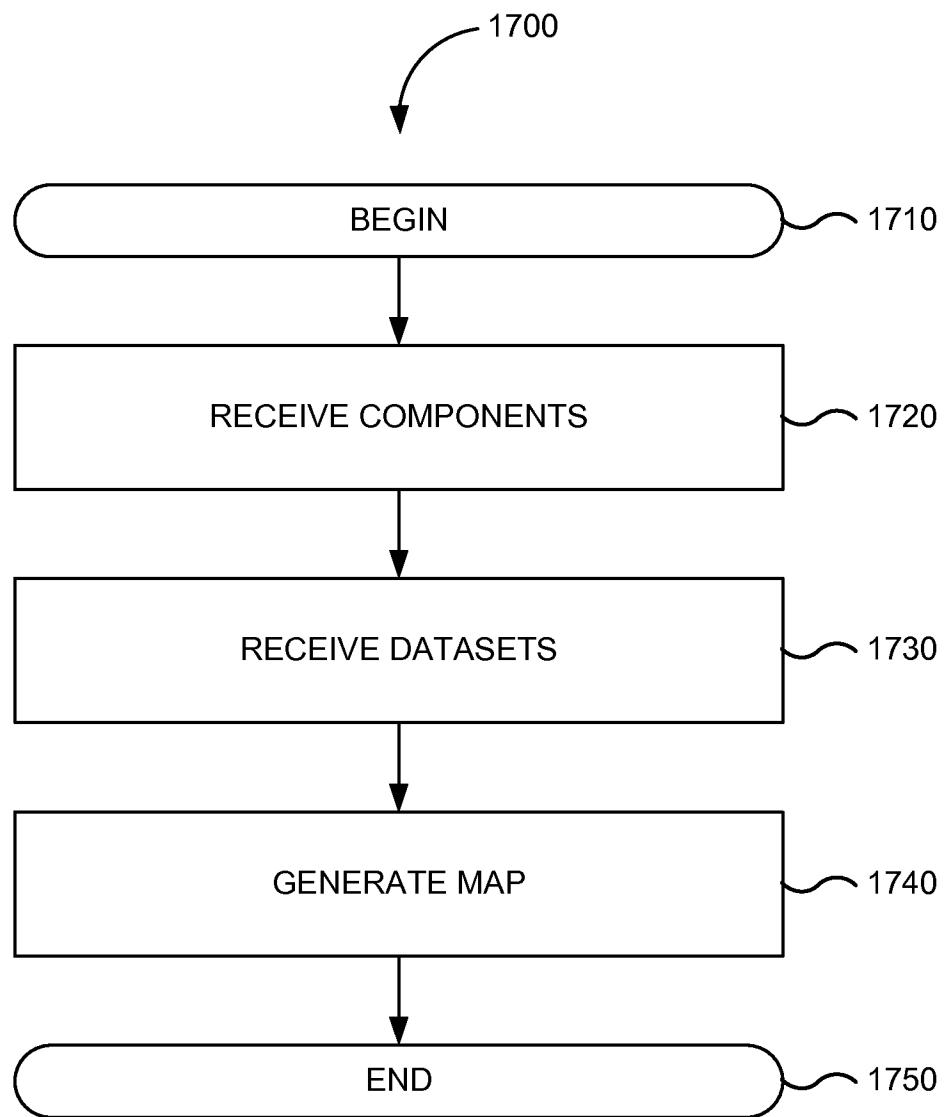


FIG. 17

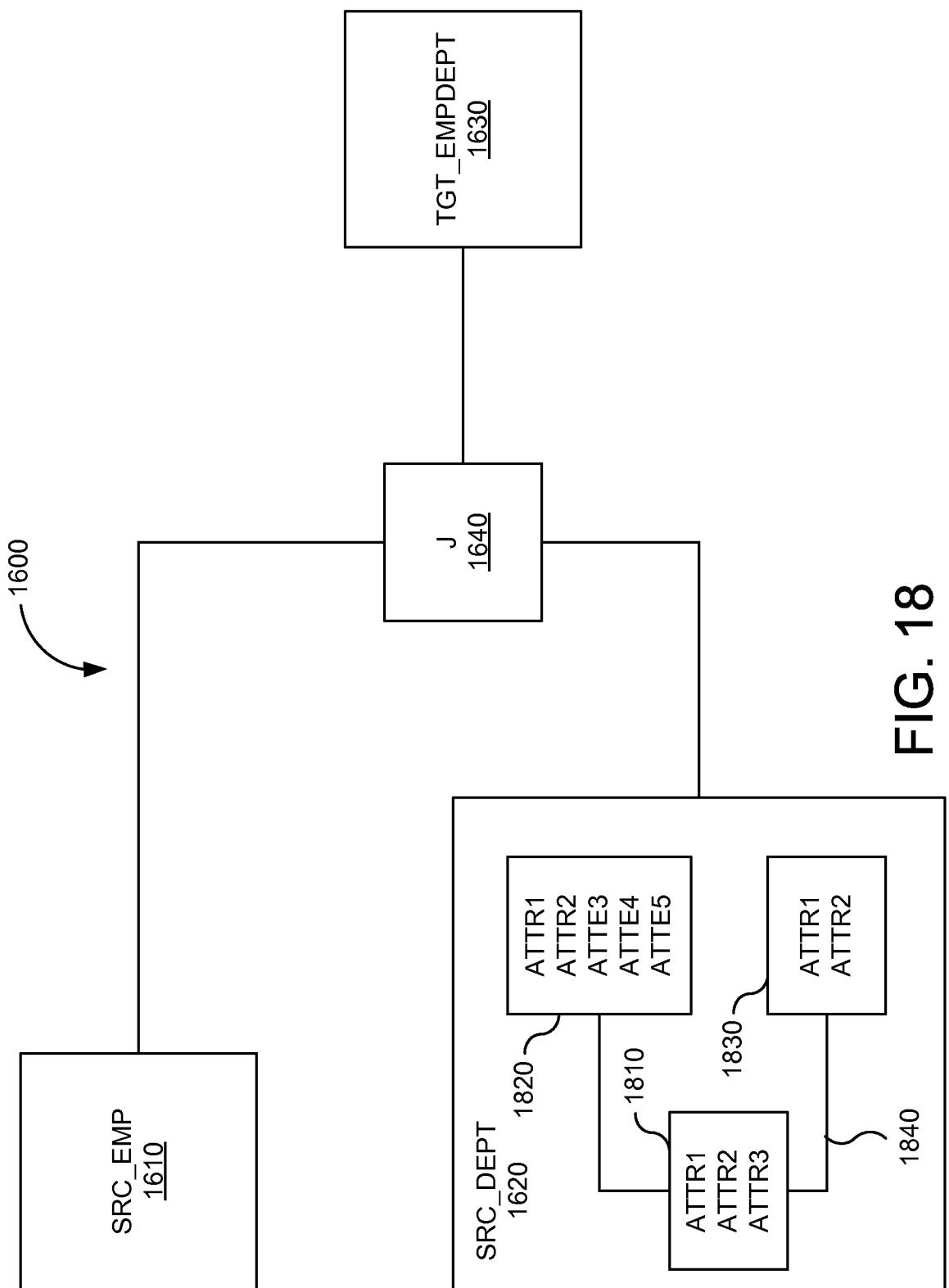


FIG. 18

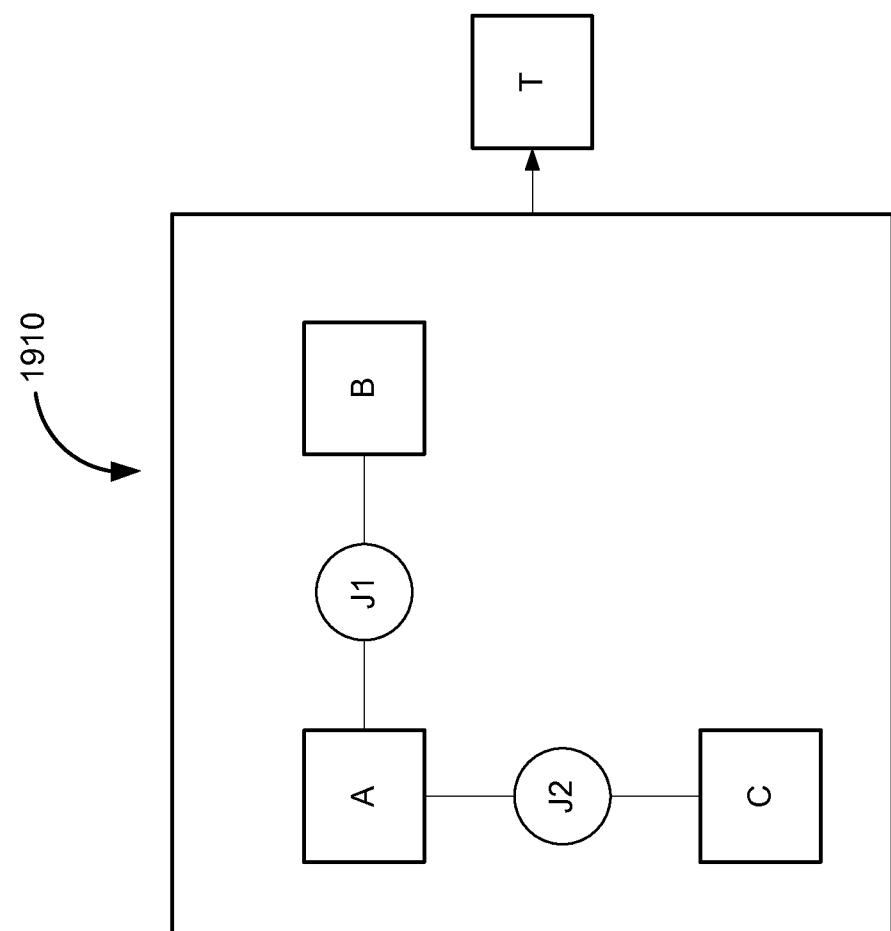


FIG. 19A

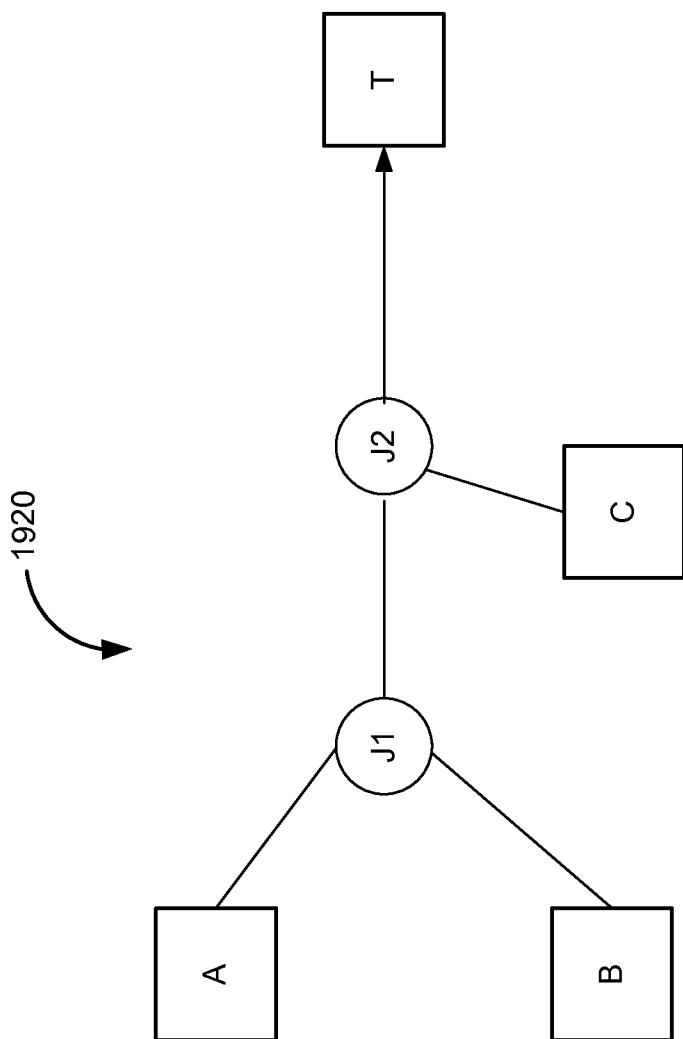


FIG. 19B

21 / 32

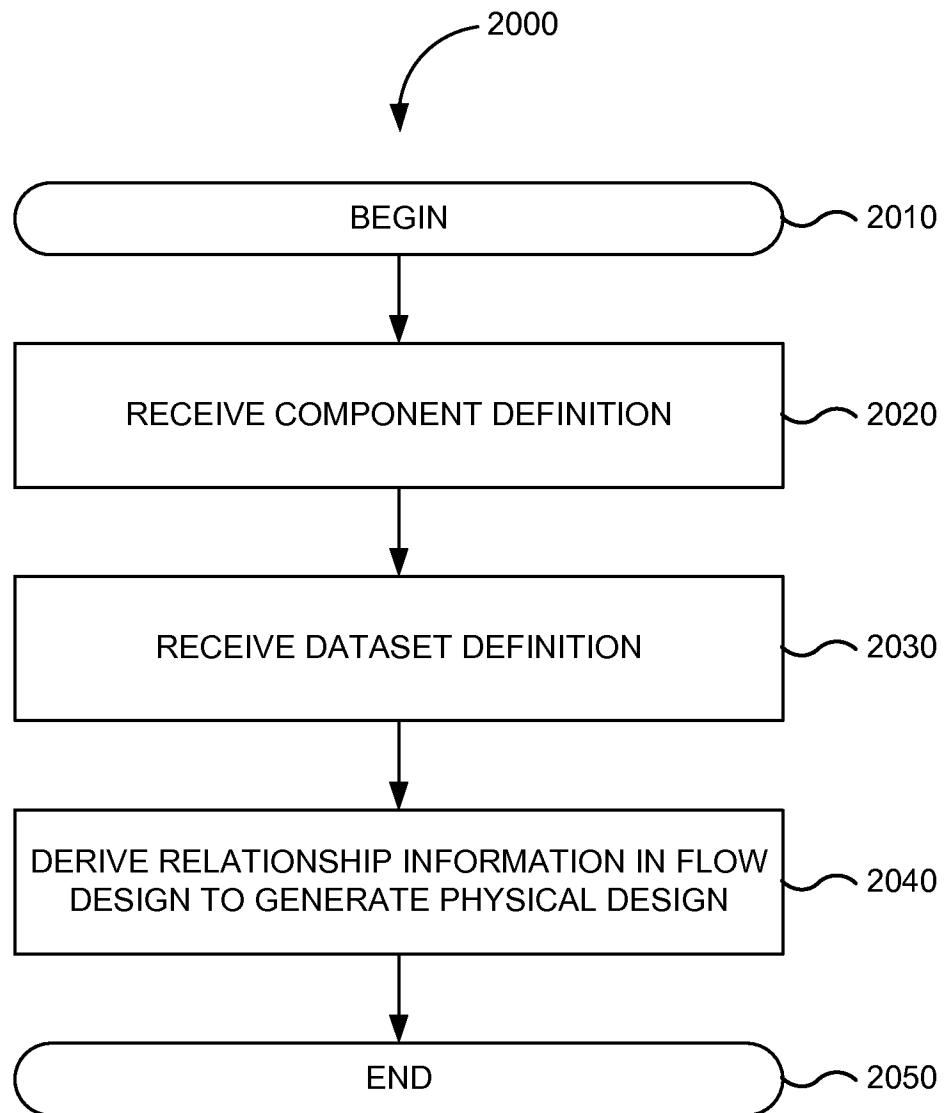


FIG. 20

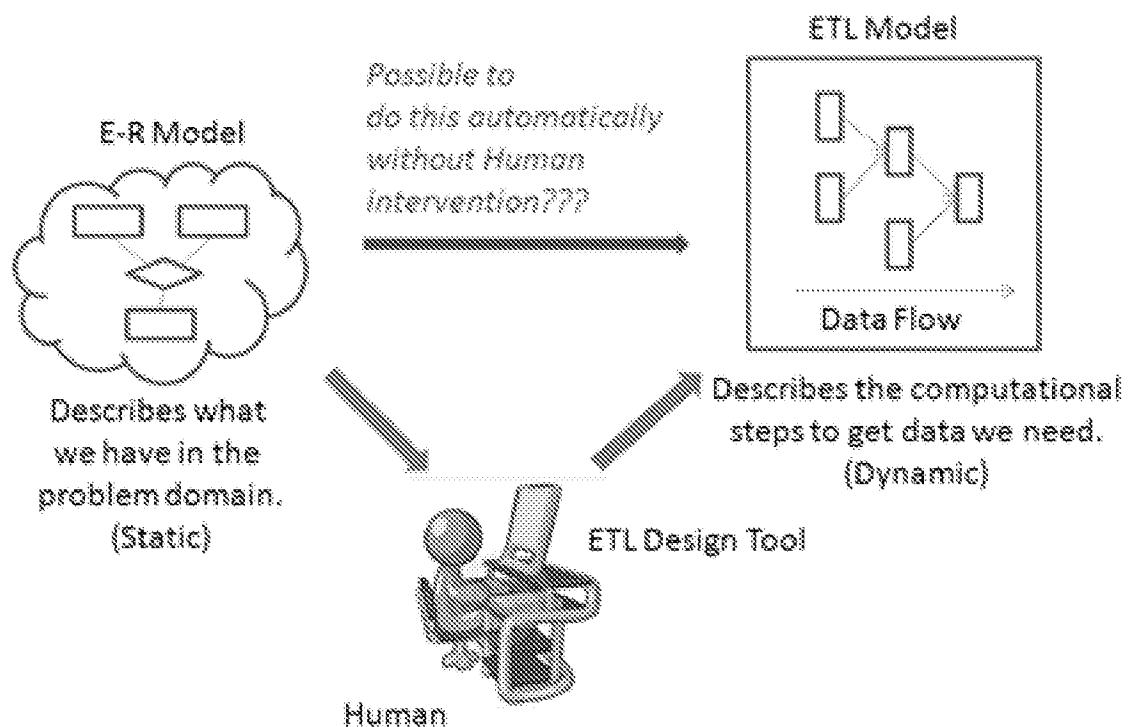


FIG. 21

23 / 32

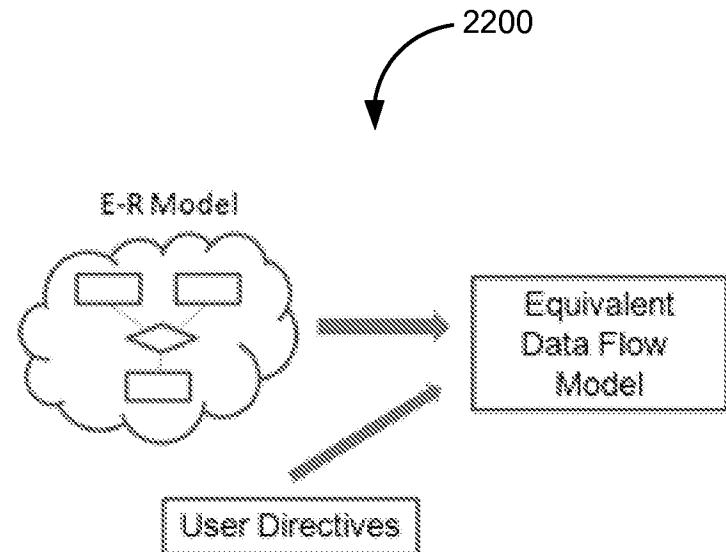
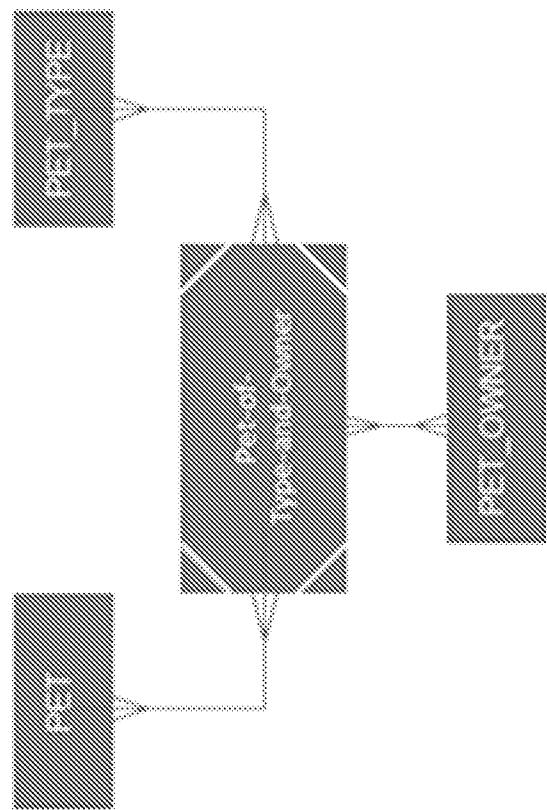
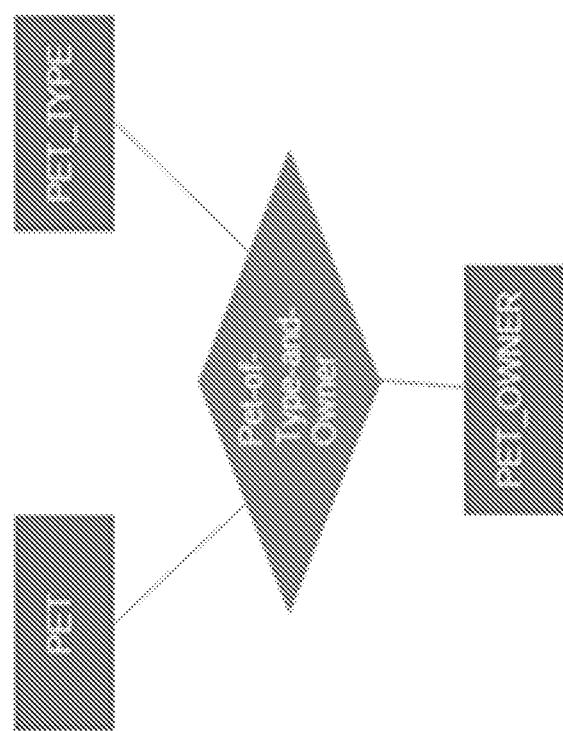


FIG. 22



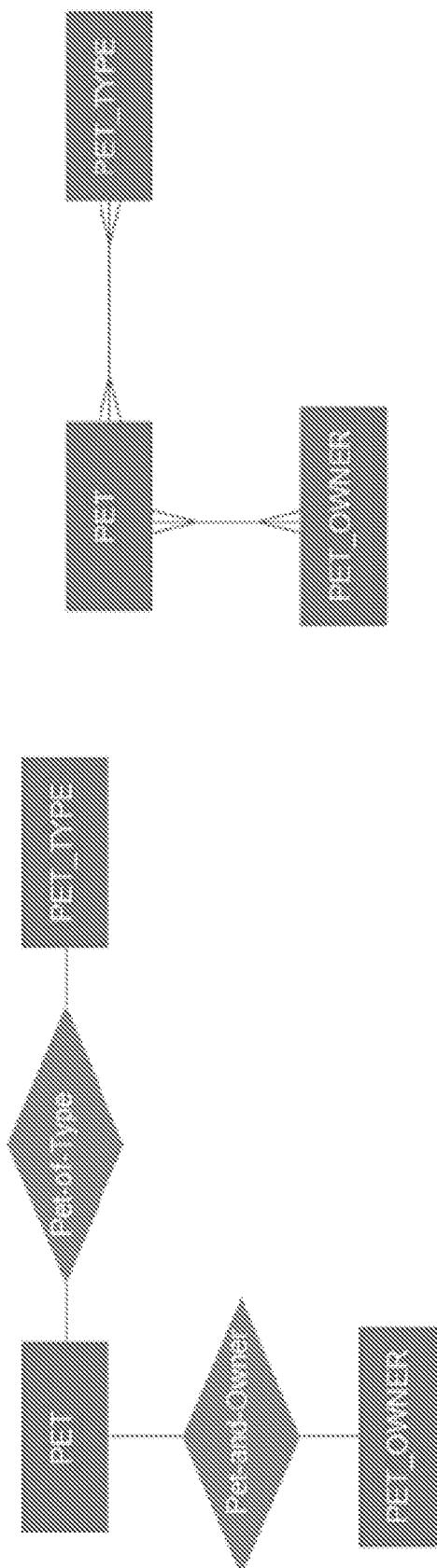
The Crow's Feet Notation

FIG. 23B



Standard ER Notation

FIG. 23A



Standard E-R Notation

The Crow's Foot Notation

FIG. 24A

FIG. 24B

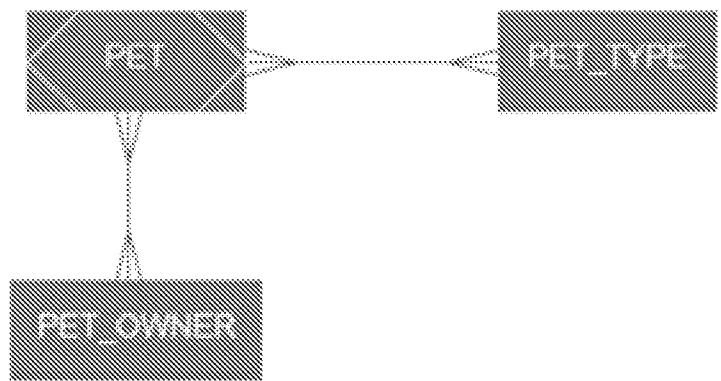


FIG. 25

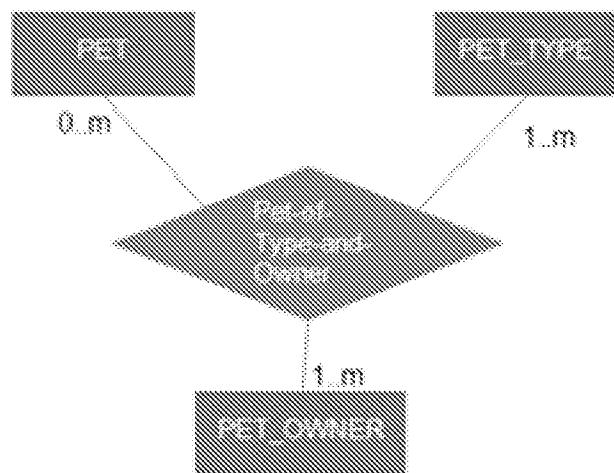


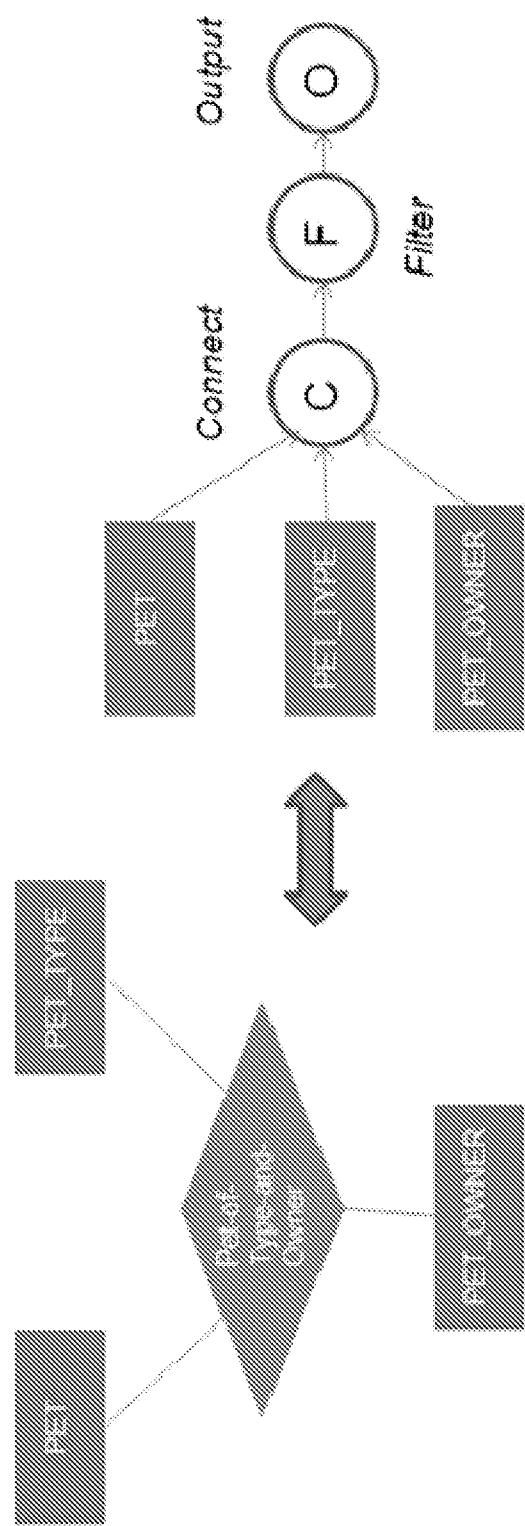
FIG. 26

| ID | NAME | TYPE |
|-----|-----------|------|
| 100 | Stray Dog | 10 |
| 101 | Cat | 1 |

SQL> select * from pet;

SQL> select * from pet_owner;

FIG. 27



Operational Semantic Model

FIG. 28B

Entity Relationship

FIG. 28A

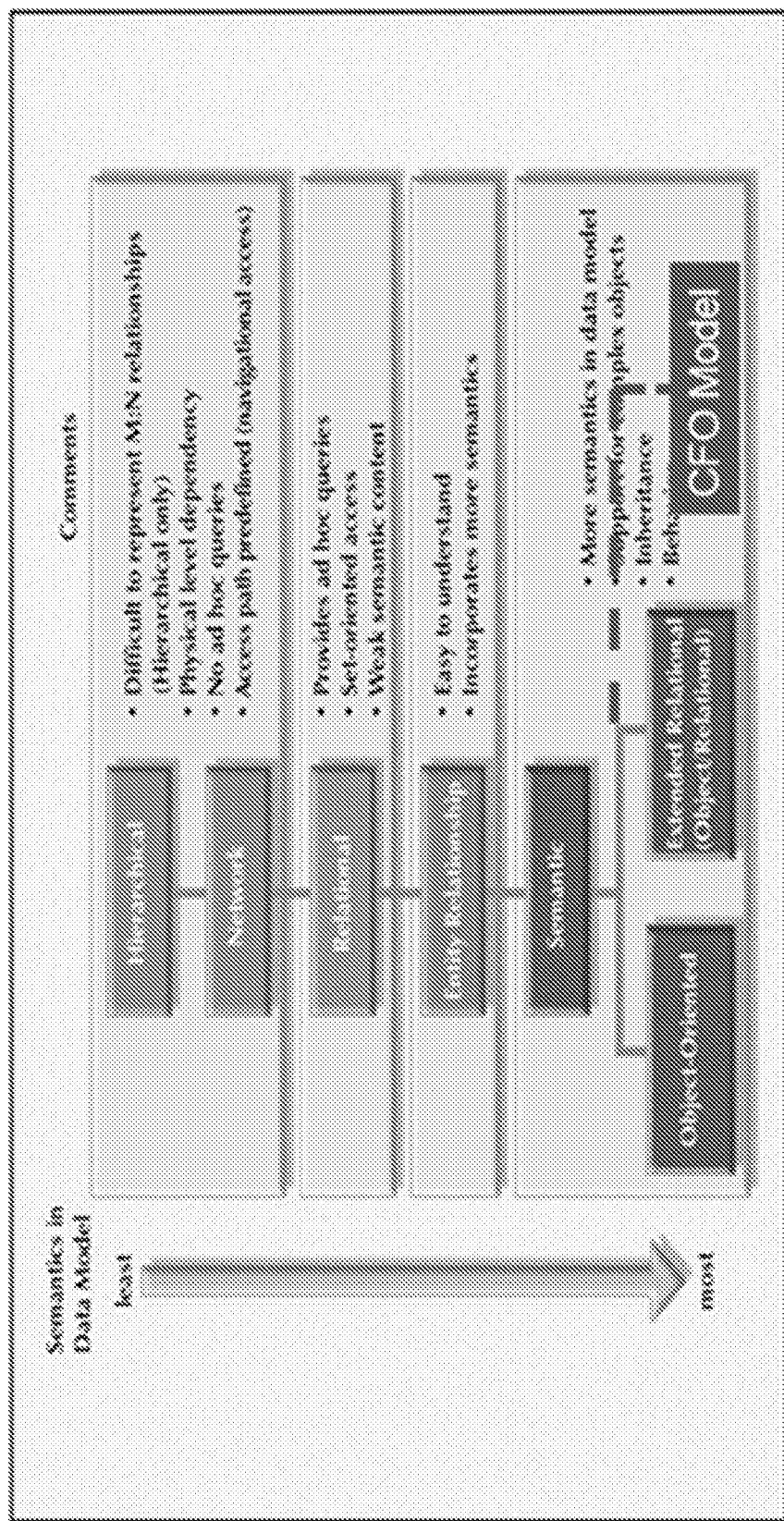


FIG. 29

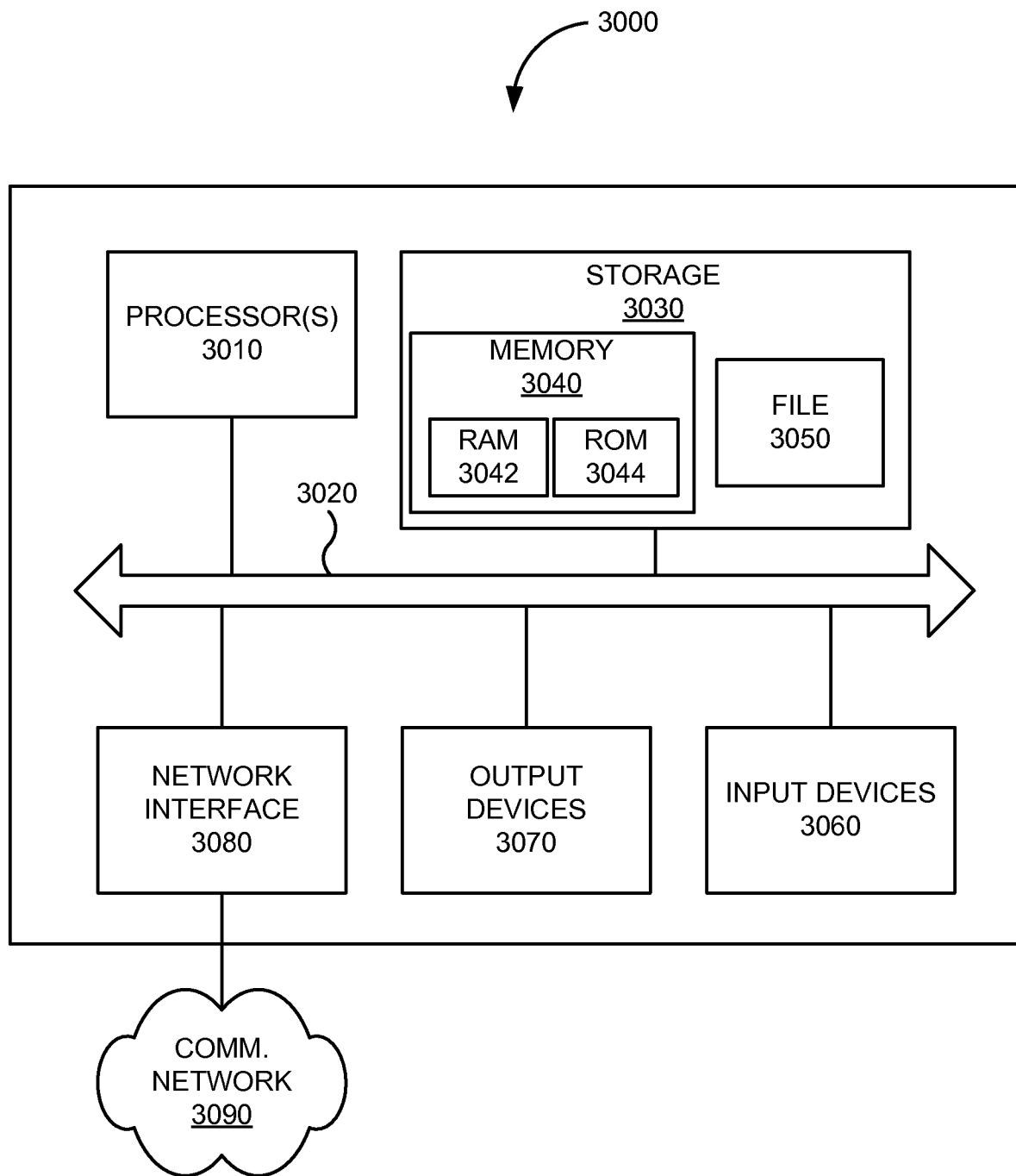


FIG. 30

32 / 32

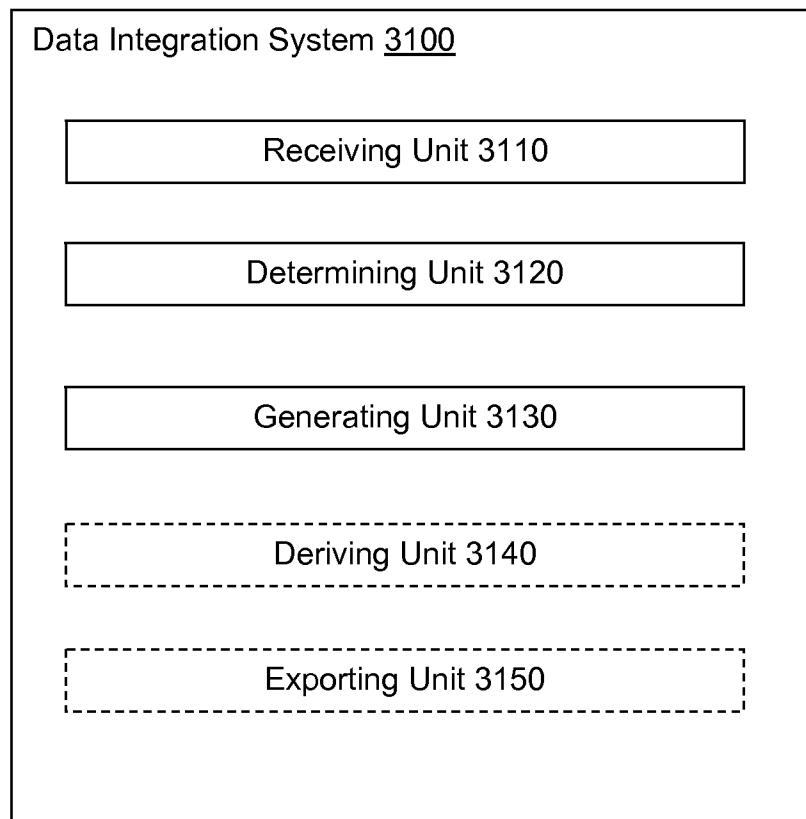


FIG. 31

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/031919

A. CLASSIFICATION OF SUBJECT MATTER

G06F 19/00(2011.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F 19/00; G06F 17/00; G06F 9/44; G06F 17/30Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean utility models and applications for utility models
Japanese utility models and applications for utility modelsElectronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & keywords: data mapping, specifying, logical design, model

C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|-----------|---|-----------------------|
| X | US 2013-0103705 A1 (SUSAN MARIE THOMAS) 25 April 2013 See paragraphs [0075]–[0087]; and figures 1, 3A. | 1-3, 8-10, 15-17 |
| A | US 2008-0281849 A1 (KAZUO MINENO) 13 November 2008 See paragraphs [0094]–[0096]; and figure 12. | 1-3, 8-10, 15-17 |
| A | US 2005-0050068 A1 (ALEXANDER VASCHILLO et al.) 3 March 2005 See paragraphs [0037]–[0044]; and figures 1-2. | 1-3, 8-10, 15-17 |
| A | US 2011-0295792 A1 (ALEXTAIR MASCARENHAS et al.) 1 December 2011 See paragraphs [0049]–[0052]; and figure 5. | 1-3, 8-10, 15-17 |
| A | US 2012-0096426 A1 (SERGHEI SARAFUDINOV) 19 April 2012 See paragraphs [0028]–[0042]; and figures 1-8. | 1-3, 8-10, 15-17 |

 Further documents are listed in the continuation of Box C. See patent family annex.

- * Special categories of cited documents:
- "A" document defining the general state of the art which is not considered to be of particular relevance
- "E" earlier application or patent but published on or after the international filing date
- "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- "O" document referring to an oral disclosure, use, exhibition or other means
- "P" document published prior to the international filing date but later than the priority date claimed

- "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
- "&" document member of the same patent family

Date of the actual completion of the international search
31 July 2014 (31.07.2014)

Date of mailing of the international search report

31 July 2014 (31.07.2014)

Name and mailing address of the ISA/KR



International Application Division
Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan City, 302-701,
Republic of Korea

Facsimile No. +82-42-472-7140

Authorized officer

JI, Jeong Hoon

Telephone No. +82-42-481-5688



INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2014/031919**Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)**

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. Claims Nos.: 4-7,11-14,18-21
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of any additional fees.
3. As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2014/031919

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|--|------------------|--|--|
| US 2013-0103705 A1 | 25/04/2013 | US 2007-203923 A1 US 8307012 B2 | 30/08/2007 06/11/2012 |
| US 2008-0281849 A1 | 13/11/2008 | US 2012-323841 A1 US 8280840 B2 US 8639670 B2 WO 2007-083371 A1 | 20/12/2012 02/10/2012 28/01/2014 26/07/2007 |
| US 2005-0050068 A1 | 03/03/2005 | CN 100468396 C0 CN 1604082 A EP 1519266 A2 EP 1519266 A3 JP 04847689 B2 JP 2005-327232 A KR 10-1159311 B1 KR 10-2005-0022272 A US 7739223 B2 | 11/03/2009 06/04/2005 30/03/2005 26/12/2007 28/12/2011 24/11/2005 22/06/2012 07/03/2005 15/06/2010 |
| US 2011-0295792 A1 | 01/12/2011 | CN 102918530 A EP 2577507 A2 JP 2013-531844 A WO 2011-149666 A2 WO 2011-149666 A3 | 06/02/2013 10/04/2013 08/08/2013 01/12/2011 19/01/2012 |
| US 2012-0096426 A1 | 19/04/2012 | None | |