



(22) **Date de dépôt/Filing Date:** 2009/12/02

(41) **Mise à la disp. pub./Open to Public Insp.:** 2010/06/10

(62) **Demande originale/Original Application:** 2 744 463

(30) **Priorité/Priority:** 2008/12/02 (US61/119,201)

(51) **Cl.Int./Int.Cl.** *G06F 17/00* (2006.01),
G06F 3/14 (2006.01)

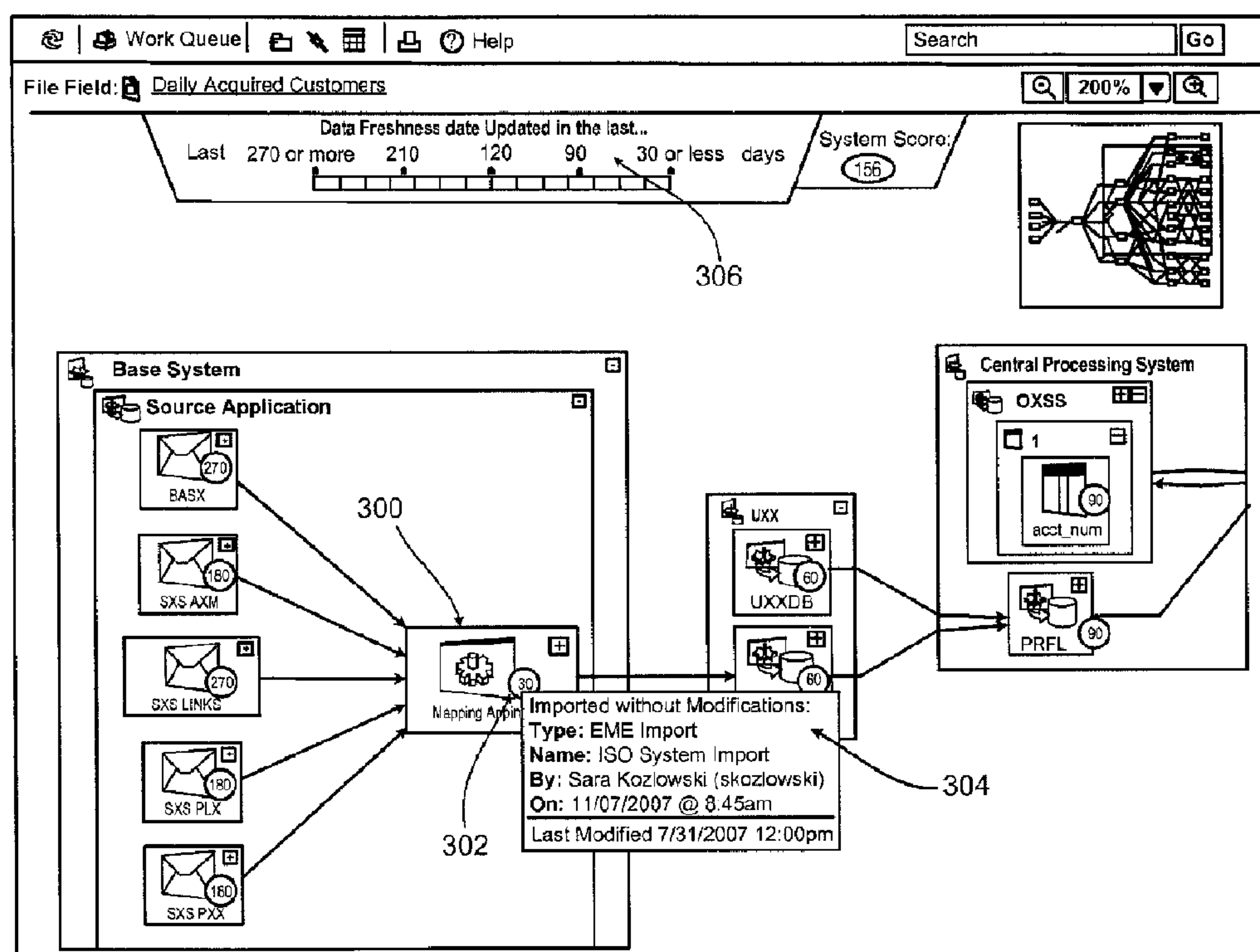
(71) **Demandeur/Applicant:**
AB INITIO TECHNOLOGY LLC, US

(72) **Inventeurs/Inventors:**
BATOR, ERIK, US;
GOULD, JOEL, US;
RADIVOJEVIC, DUSAN, US;
WAKELING, TIM, US

(74) **Agent:** SMART & BIGGAR

(54) **Titre : VISUALISATION DE RELATIONS ENTRE DES ELEMENTS DE DONNEES**

(54) **Title: VISUALIZING RELATIONSHIPS BETWEEN DATA ELEMENTS**



(57) **Abrégé/Abstract:**

In general, a specification of multiple contexts that are related according to a hierarchy is received. Relationships are determined among three or more metadata objects (202B, 204B, 206A), and at least some of the metadata objects are grouped into one or more respective groups. Each of at least some of the groups is based on a selected one of the contexts and is represented by a node (208B, 210B, 212B) in a diagram (200B). Relationships among the nodes are determined based on the relationships among the metadata objects in the groups represented by the nodes, and a visual representation is generated of the diagram including the nodes and the relationships among the nodes.

60412-447D1

ABSTRACT

In general, a specification of multiple contexts that are related according to a hierarchy is received. Relationships are determined among three or more metadata objects (202B, 204B, 206A), and at least some of the metadata objects are grouped into one or more
5 respective groups. Each of at least some of the groups is based on a selected one of the contexts and is represented by a node (208B, 210B, 212B) in a diagram (200B). Relationships among the nodes are determined based on the relationships among the metadata objects in the groups represented by the nodes, and a visual representation is generated of the diagram including the nodes and the relationships among the nodes.

VISUALIZING RELATIONSHIPS BETWEEN DATA ELEMENTS

This application is a divisional of Canadian Patent Application No. 2,744,463 filed on December 2, 2009.

This application claims priority to U.S. Application Serial No. 61/119,201, filed
5 on December 2, 2008.

BACKGROUND

This description relates to visualizing relationships between data elements.

Enterprises use complex data processing systems, such as data warehousing, customer relationship management, and data mining, to manage data. In many data processing systems, data are pulled from many different data sources, such as database files, operational systems, flat files, the Internet, etc, into a central repository. Often, data are transformed before being loaded in the data system. Transformation may include cleansing, integration, and extraction. To keep track of data, its sources, and the transformations that have happened to the data stored in a data system, metadata can be used. Metadata (sometimes called "data about data") are data that describe other data's attributes, format, origins, histories, inter-relationships, etc. Metadata management can play a central role in complex data processing systems.

Sometimes a database user may want to investigate how certain data are derived from different data sources. For example, a database user may want to know how a dataset or data object was generated or from which source a dataset or data object was imported. Tracing a dataset back to sources from which it is derived is called data lineage tracing (or "upstream data lineage tracing"). Sometimes a database user may want to investigate how certain datasets have been used (called "downstream data lineage tracing" or "impact analysis"), for example, which application has read a given dataset. A database user may also be interested in knowing how a dataset is related to other datasets. For example, a user may want to know if a dataset is modified, what tables will be affected.

SUMMARY

In a general aspect, a method includes receiving a specification of multiple contexts that are related according to a hierarchy. Relationships are determined among three or more metadata objects, and at least some of the metadata objects are grouped into one or more respective groups, each of at least some of the groups being based on a selected one of the contexts and being represented by a node in a diagram. Relationships are determined among the nodes based on the relationships among the metadata objects in the groups represented by the nodes, and a visual representation is generated of the diagram including the nodes and the relationships among the nodes.

Aspects can include one or more of the following features. The metadata objects represent data elements or transformations of data. The data elements are grouped in contexts including data sets, applications, or systems. The transformations of data are grouped in contexts including executables, applications, or systems. The nodes are configured to be expanded or collapsed to display different levels of resolution. A user selects the level of resolution to display. A user selects a particular node for which to calculate relationships. Each node not containing the selected data node are collapsed to a least detailed level of resolution according to an algorithm. Nodes meeting a preselected condition are filtered out of the visual representation.

In a general aspect, a system includes means for receiving a specification of multiple contexts that are related according to a hierarchy. A system also includes means for determining relationships among three or more metadata objects, and means for grouping at least some of the metadata objects into one or more respective groups, each of at least some of the groups being based on a selected one of the contexts and being represented by a node in a diagram. A system also includes means for determining relationships among the nodes based on the relationships among the metadata objects in the groups represented by the nodes, and means for generating a visual representation of the diagram including the nodes and the relationships among the nodes.

In a general aspect, a computer system is configured to receive a specification of multiple contexts that are related according to a hierarchy, and determine relationships among three or more metadata objects. At least some of the metadata objects are grouped into one or more respective groups, each of at least some of the groups being based on a

selected one of the contexts and being represented by a node in a diagram. Relationships are determined among the nodes based on the relationships among the metadata objects in the groups represented by the nodes, and a visual representation is generated of the diagram including the nodes and the relationships among the nodes.

5 In a general aspect, a computer-readable medium stores a computer program, and the computer program includes instructions for causing a computer to receive a specification of multiple contexts that are related according to a hierarchy. Relationships are determined among three or more metadata objects. At least some of the metadata objects are grouped into one or more respective groups, each of at least some of the groups being based
10 on a selected one of the contexts and being represented by a node in a diagram. Relationships are determined among the nodes based on the relationships among the metadata objects in the groups represented by the nodes, and a visual representation is generated of the diagram including the nodes and the relationships among the nodes.

 According to an aspect of the present invention, there is provided a method
15 including: receiving a specification of multiple contexts that are related according to a hierarchy; determining relationships among three or more metadata objects, in which for a given metadata object of the metadata objects, the relationships among the metadata objects include one or more of: relationships between the given metadata object and one or more metadata objects on which the given metadata object depends, and relationships between the
20 given metadata object and one or more metadata objects that depend on the given metadata object; grouping at least some of the metadata objects into one or more respective groups, each of at least some of the groups being based on a selected one of the contexts, at least two groups being associated with different contexts; representing the at least some metadata objects by a diagram having nodes, each node representing a metadata object or a group of
25 metadata objects; determining relationships among the nodes based on the relationships among the metadata objects in the groups represented by the nodes; determining groupings of the nodes based on the groupings of the metadata objects; and generating a visual representation of the diagram including the nodes and the relationships among the nodes, the diagram including the at least two nodes or two groups of nodes that are associated with
30 different contexts.

According to another aspect of the present invention, there is provided a system including: a computer including one or more processors; means for receiving a specification of multiple contexts that are related according to a hierarchy; means for determining relationships among three or more metadata objects, in which for a given metadata object of the metadata objects, the relationships among the metadata objects include one or more of: relationships between the given metadata object and one or more metadata objects on which the given metadata object depends, and relationships between the given metadata object and one or more metadata objects that depend on the given metadata object; means for grouping at least some of the metadata objects into one or more respective groups, each of at least some of the groups being based on a selected one of the contexts, at least two groups being associated with different contexts; means for representing the at least some metadata objects by a diagram having nodes, each node representing a metadata object or a group of metadata objects; means for determining relationships among the nodes based on the relationships among the metadata objects in the groups represented by the nodes; means for determining groupings of the nodes based on the groupings of the metadata objects; and means for generating a visual representation of the diagram including the nodes and the relationships among the nodes, the diagram including the at least two nodes or two groups of nodes that are associated with different contexts.

According to still another aspect of the present invention, there is provided a computer system including: one or more computing devices configured to: receive a specification of multiple contexts that are related according to a hierarchy; determine relationships among three or more metadata objects, in which for a given metadata object of the metadata objects, the relationships among the metadata objects include one or more of: relationships between the given metadata object and one or more metadata objects on which the given metadata object depends, and relationships between the given metadata object and one or more metadata objects that depend on the given metadata object; group at least some of the metadata objects into one or more respective groups, each of at least some of the groups being based on a selected one of the contexts, at least two groups being associated with different contexts; represent the at least some metadata objects by a diagram having nodes, each node representing a metadata object or a group of metadata objects; determine

relationships among the nodes based on the relationships among the metadata objects in the groups represented by the nodes; determine groupings of the nodes based on the groupings of the metadata objects; and generate a visual representation of the diagram including the nodes and the relationships among the nodes, the diagram including the at least two nodes or two
5 groups of nodes that are associated with different contexts.

According to yet another aspect of the present invention, there is provided a computer-readable device storing a computer program, the computer program including instructions for causing a computer to: receive a specification of multiple contexts that are related according to a hierarchy; determine relationships among three or more metadata
10 objects, in which for a given metadata object of the metadata objects, the relationships among the metadata objects include one or more of: relationships between the given metadata object and one or more metadata objects on which the given metadata object depends, and relationships between the given metadata object and one or more metadata objects that depend on the given metadata object; group at least some of the metadata objects into one or more
15 respective groups, each of at least some of the groups being based on a selected one of the contexts, at least two groups being associated with different contexts; represent the at least some metadata objects by a diagram having nodes, each node representing a metadata object or a group of metadata objects; determine relationships among the nodes based on the relationships among the metadata objects in the groups represented by the nodes; determine
20 groupings of the nodes based on the groupings of the metadata objects; and generate a visual representation of the diagram including the nodes and the relationships among the nodes, the diagram including the at least two nodes or two groups of nodes that are associated with different contexts.

According to a further aspect of the present invention, there is provided a
25 method including: receiving metadata objects from one or more sources; grouping the metadata objects into one or more first groups associated with a first context; grouping one or more of the first groups into one or more second groups associated with a second context having a higher hierarchy level than the first context; grouping one or more of the second groups into one or more third groups associated with a third context having a higher hierarchy
30 level than the second context; and generating a visual representation of a diagram including

nodes and groups of nodes, each node representing one of the metadata objects or one of the first, second, and third groups, the visual representation showing at least one of the first group of nodes within one of the second groups, and at least one of the second groups within one of the third groups.

5 According to yet a further aspect of the present invention, there is provided a method including: receiving a specification of multiple contexts that are related according to a hierarchy, the contexts including applications context and systems context, the contexts further including at least one of datasets context or executables context; determining relationships among three or more metadata objects; grouping at least some of the metadata
10 objects into one or more respective groups, each of at least some of the groups being based on a selected one of the contexts, at least two groups being associated with different contexts; representing the metadata objects by a diagram having nodes, each node representing a metadata object or a group of metadata objects; determining relationships among the nodes based on the relationships among the metadata objects in the groups; and generating a visual
15 representation of a diagram including the nodes, the diagram also showing the groupings of nodes, the diagram including the at least one group of nodes associated with the datasets context or the executables context, at least one group of nodes associated with the applications context, and at least one group of nodes associated with the systems context.

 According to still a further aspect of the present invention, there is provided a
20 system including: a computer including one or more processors; means for receiving metadata objects from one or more sources; means for grouping the metadata objects into one or more first groups associated with a first context; means for grouping one or more of the first groups into one or more second groups associated with a second context having a higher hierarchy level than the first context; means for grouping one or more of the second groups into one or
25 more third groups associated with a third context having a higher hierarchy level than the second context; and means for generating a visual representation of a diagram including nodes and groups of nodes, each node representing one of the metadata objects or one of the first, second, and third groups, the visual representation showing at least one of the first group of nodes within one of the second groups, and at least one of the second groups within one of the
30 third groups.

According to another aspect of the present invention, there is provided a system including: a computer including one or more processors; means for receiving a specification of multiple contexts that are related according to a hierarchy, the contexts including applications context and systems context, the contexts further including at least one of datasets context or executables context; means for determining relationships among three or more metadata objects; means for grouping at least some of the metadata objects into one or more respective groups, each of at least some of the groups being based on a selected one of the contexts, at least two groups being associated with different contexts; means for representing the metadata objects by a diagram having nodes, each node representing a metadata object or a group of metadata objects; means for determining relationships among the nodes based on the relationships among the metadata objects in the groups; and means for generating a visual representation of a diagram including the nodes, the diagram also showing the groupings of nodes, the diagram including the at least one group of nodes associated with the datasets context or the executables context, at least one group of nodes associated with the applications context, and at least one group of nodes associated with the systems context.

According to yet another aspect of the present invention, there is provided a computer system including: one or more computing devices configured to: receive metadata objects from one or more sources; group the metadata objects into one or more first groups associated with a first context; group one or more of the first groups into one or more second groups associated with a second context having a higher hierarchy level than the first context; group one or more of the second groups into one or more third groups associated with a third context having a higher hierarchy level than the second context; and generate a visual representation of a diagram including nodes and groups of nodes, each node representing one of the metadata objects or one of the first, second, and third groups, the visual representation showing at least one of the first group of nodes within one of the second groups, and at least one of the second groups within one of the third groups.

According to another aspect of the present invention, there is provided a computer system including: one or more computing devices configured to: receive a specification of multiple contexts that are related according to a hierarchy, the contexts including applications context and systems context, the contexts further including at least one

of datasets context or executables context; determine relationships among three or more metadata objects; group at least some of the metadata objects into one or more respective groups, each of at least some of the groups being based on a selected one of the contexts, at least two groups being associated with different contexts; represent the metadata objects by a
5 diagram having nodes, each node representing a metadata object or a group of metadata objects; determine relationships among the nodes based on the relationships among the metadata objects in the groups; and generate a visual representation of a diagram including the nodes, the diagram also showing the groupings of nodes, the diagram including the at least one group of nodes associated with the datasets context or the executables context, at least one
10 group of nodes associated with the applications context, and at least one group of nodes associated with the systems context.

According to still another aspect of the present invention, there is provided a computer-readable device storing a computer program, the computer program including instructions for causing a computer to: receive metadata objects from one or more sources;
15 group the metadata objects into one or more first groups associated with a first context; group one or more of the first groups into one or more second groups associated with a second context having a higher hierarchy level than the first context; group one or more of the second groups into one or more third groups associated with a third context having a higher hierarchy level than the second context; and generate a visual representation of a diagram including
20 nodes and groups of nodes, each node representing one of the metadata objects or one of the first, second, and third groups, the visual representation showing at least one of the first group of nodes within one of the second groups, and at least one of the second groups within one of the third groups.

According to yet another aspect of the present invention, there is provided a
25 computer-readable device storing a computer program, the computer program including instructions for causing a computer to: receive a specification of multiple contexts that are related according to a hierarchy, the contexts including applications context and systems context, the contexts further including at least one of datasets context or executables context; determine relationships among three or more metadata objects; group at least some of the
30 metadata objects into one or more respective groups, each of at least some of the groups being

based on a selected one of the contexts, at least two groups being associated with different contexts; represent the metadata objects by a diagram having nodes, each node representing a metadata object or a group of metadata objects; determine relationships among the nodes based on the relationships among the metadata objects in the groups; and generate a visual
5 representation of a diagram including the nodes, the diagram also showing the groupings of nodes, the diagram including the at least one group of nodes associated with the datasets context or the executables context, at least one group of nodes associated with the applications context, and at least one group of nodes associated with the systems context.

Aspects can have one or more of the following advantages.

10 The system enables users to visualize relationships between objects, and view certain attributes of objects in a contextual setting. When working with metadata, users are able to understand the origins of an object before certain actions are taken. Users can know which objects are affected by manipulation of any particular object. Users are also able to view attributes of certain objects in an environment where relationships between those objects
15 are clearly shown.

DESCRIPTION OF DRAWINGS

FIG. 1 is a block diagram of a computing system.

FIGS. 2A-2E are diagrams showing relationships between nodes of data.

FIGS. 3-5 are diagrams showing graphical overlays superimposed on nodes of
20 data.

DESCRIPTION

The system enables users to visualize relationships between objects, stored in a wide variety of data storage systems. The relationships among the objects can represent various dependencies and/or associations appropriate to the applications for which the data objects are used. As an example of one of the types of systems in which these techniques can be used, a system is described in which the objects represent elements of a graph-based computation environment.

FIG. 1A is a block diagram showing the interrelationship of parts of a computing system 100 for developing, executing and managing graph-based computations. A graph-based computation is implemented using a “data flow graph” that is represented by a directed graph, with vertices in the graph representing components (either data files or processes), and the directed links or “edges” in the graph representing flows of data between components. A graphic development environment (GDE) 102 provides a user interface for specifying executable graphs and defining parameters for the graph components. The GDE may be, for example, the CO>OPERATING SYSTEM® GDE available from Ab Initio. The GDE 102 communicates with a repository 104 and a parallel operating environment 106. Also coupled to the repository 104 and the parallel operating environment 106 are a User Interface module 108 and an executive 110.

In some examples, repository 104 includes both a base data store 105A and an interface data store 105B. A base data store stores technical metadata, and may include applications along with their associated metadata, such as graphs and transforms. In addition to storing technical metadata, the base data store may also perform various kinds of analysis including dependency analysis (e.g., computing data lineage, as described in more detail below), or may receive and store the results of such analysis. In some examples, base data store 105A and interface data store 105B may be combined and implemented as a single data store.

While technical metadata is useful to developers in a variety of functions, there are many instances in which a higher level of metadata needs to be analyzed and manipulated. This higher level metadata, sometimes referred to as “enterprise” or “business” metadata is often useful in data analysis. Some examples of business metadata include data stewardship, which indicates which employee is responsible for the

data, and data dictionaries, which are business definitions for files and fields within files. Business metadata goes beyond technical descriptions of data, and can be stored on a platform that is separate from the base data store 105A, such as an interface data store 105B.

5 The interface data store 105B may be a relational database that primarily serves to store business metadata. The interface data store may communicate with the base data store and extract its metadata, and it can also pull its information from a variety of other sources such as graphs, spreadsheets, logical models, database tables, or additional third party sources of data.

10 In some examples, the base data store 105A is a scalable object-oriented database system designed to support the development and execution of graph-based applications and the interchange of metadata between the graph-based applications and other systems (e.g., other operating systems). The repository 104 is a storage system for all kinds of metadata, including documentation, record formats (e.g., fields and data types of records
15 in a table), transform functions, graphs, jobs, and monitoring information. The repository 104 also stores metadata objects that represent actual data to be processed by the computing system 100 including data stored in an external data store 112. An example of a repository that includes features for importing and managing metadata from various sources is described in co-pending U.S. Provisional Patent Application Serial No.
20 61/119,148, entitled "DATA MAINTENANCE SYSTEM," filed on December 2, 2008. Similar features can be incorporated into the repository 104.

 The parallel operating environment 106 accepts a specification of a data flow graph generated in the GDE 102 and generates computer instructions that correspond to
25 the processing logic and resources defined by the graph. The parallel operating environment 106 then typically executes those instructions on a plurality of processors (which need not be homogeneous). An example of a suitable parallel operating environment is the CO>OPERATING SYSTEM®.

 The User Interface module 108 provides a web-browser-based view of the
30 contents of the repository 104. Using the User Interface module 108, a user 103 may browse objects, create new objects, alter existing objects, specify application parameters,

schedule jobs, etc. The User Interface module 108 generates forms-based browser screens for a user to search for and view objects and information about objects stored in the repository 104.

The repository 104 stores metadata including metadata objects for graph-based applications including graph components and other functional objects for building computation graphs. As stated previously, metadata stored in base data store 105A of repository 104 includes, for example, “technical” metadata (e.g., application-related business rules, record formats, and execution statistics), while the interface data store 105B may include business metadata such as user-defined documentation of job functions, roles, and responsibilities.

The information stored in the repository 104 in the form of metadata objects enables various kinds of analysis about applications and the data processed by those applications. Subsets of this information may be stored in interface data store 105B. For example, as discussed further below, a user can obtain answers to questions about data lineage (e.g., Where did a given value come from? How was the output value computed? Which applications produce and depend on this data?). A developer can understand the consequences of proposed modifications (e.g., If this piece changes, what else will be affected? If this source format changes, which applications will be affected?). A user/developer can also obtain questions to answers involving both technical metadata and business metadata (e.g., Which groups are responsible for producing and using this data? Who changed this application last? What changes did they make?).

The repository 104 is able to track the state of stored metadata objects. Objects stored in the repository 104 are versioned, making it possible to examine the state of things as of last week, last month, or last year, and to compare it with the state of things today. The repository 104 collects job-tracking, or execution information which enables trend analysis (e.g., How fast is our data growing?) and capacity planning (e.g., How long did that application take to run? How much data did it process, and at what rate? What resources did the application consume? When will we need to add another server?).

A user can view (and optionally, edit) information contained in and/or associated with the stored metadata through the User Interface module 108. A metadata viewing environment can represent various kinds of metadata objects using various graphical

representations including icons and groupings of icons presented by the User Interface module 108 on a display. A metadata object can represent different types of data elements (e.g., data used as input or output of an executable program) and/or transformations (e.g., any type of data manipulation associated with a data processing entity, such as data flow graph, that processes or generates data). The viewing environment can show relationships as lines connecting graphical nodes that represent metadata objects or groupings of metadata objects, as described in more detail below. In some cases, the interface data store 105B can extract the relationships (such as lineage information) from the base data store 105A, or from other sources of data. The interface data store 105B may hold a high-level summary of data lineage. The lineage information (or other data dependency analysis) can be computed automatically within the system 100, or can be received from an external system, or from manual input. For example, the system 100 can receive lineage information that has been gathered and prepared by humans analyzing the code. The lineage information can be imported into the repository 104 from files in any of a variety of predetermined formats (e.g., in spreadsheets).

FIG. 2A shows an example of a metadata viewing environment. In some examples, the metadata viewing environment is an interface that runs on top of a browser. In the example of FIG. 2A, the metadata viewing environment displays information related to a data lineage diagram 200A. One example of metadata viewing environment is a web-based application that allows a user to visualize and edit metadata. Using the metadata viewing environment, a user can explore, analyze, and manage metadata using a standard Web browser from anywhere within an enterprise. Each type of metadata object has one or more views or visual representations. The metadata viewing environment of figure 2A illustrates a lineage diagram for target element 206A.

For example, the lineage diagram displays the end-to-end lineage for the data and/or processing nodes that represent the metadata objects stored in the repository 104; that is, the objects a given starting object depends on (its sources) and the objects that a given starting object affects (its targets). In this example, connections are shown between data elements 202A and transformations 204A, two examples of metadata objects. The metadata objects are represented by nodes in the diagram. Data elements 202A can

represent datasets, tables within datasets, columns in tables, and fields in files, messages, and reports, for example. An example of a transformation 204A is an element of an executable that describes how a single output of a data element is produced. The connections between the nodes are based on relationships among the metadata objects.

5 FIG. 2B illustrates a corresponding lineage diagram 200B for the same target element 206A shown in FIG. 2A except each element 202B is grouped and shown in a group based on a context. For example, data elements 202B are grouped in datasets 208B (e.g., tables, files, messages, and reports), applications 210B (that contain executables such as graphs and plans and programs, plus the datasets that they operate
10 on), and systems 212B. Systems 212B are functional groupings of data and the applications that process the data; systems consist of applications and data groups (e.g., databases, file groups, messaging systems, and groups of datasets). Transformations 204B are grouped in executables 214B, applications 210B, and systems 212B. Executables such as graphs, plans or programs, read and write datasets. Parameters can
15 set what groups are expanded and what groups are collapsed by default. This allows users to see the details for only the groups that are important to them by removing unnecessary levels of details.

Using the metadata viewing environment to perform data lineage calculations is useful for a number of reasons. For example, calculating and illustrating relationships
20 between data elements and transformations can help a user determine how a reported value was computed for a given field report. A user may also view which datasets store a particular type of data, and which executables read and write to that dataset. In the case of business terms, the data lineage diagram may illustrate which data elements (such as columns and fields) are associated with certain business terms (definitions in an
25 enterprise).

Data lineage diagrams shown within the metadata viewing environment can also aid a user in impact analysis. Specifically, a user may want to know which downstream executables are affected if a column or field is added to a dataset, and who needs to be notified. Impact analysis may determine where a given data element is used, and can also
30 determine the ramifications of changing that data element. Similarly, a user may view

what datasets are affected by a change in an executable, or whether it safe to remove a certain database table from production.

Using the metadata viewing environment to perform data lineage calculations for generating data lineage diagrams is useful for business term management. For instance, it is often desirable for employees within an enterprise to agree on the meanings of business terms across that enterprise, the relationships between those terms, and the data to which the terms refer. The consistent use of business terms may enhance the transparency of enterprise data and facilitates communication of business requirements. Thus, it is important to know where the physical data underlying a business term can be found, and what business logic is used in computations.

Viewing relationships between data nodes can also be helpful in managing and maintaining metadata. For instance, a user may wish to know who changed a piece of metadata, what the source (or “source of record”) is for a piece of metadata, or what changes were made when loading or reloading metadata from an external source. In maintaining metadata, it may be desirable to allow designated users to be able to create metadata objects (such as business terms), edit properties of metadata objects (such as descriptions and relationships of objects to other objects), or delete obsolete metadata objects.

The metadata viewing environment provides a number of graphical views of objects, allowing a user to explore and analyze metadata. For example, a user may view the contents of systems and applications and explore the details of any object, and can also view relationships between objects using the data lineage views, which allows a user to easily perform various types of dependency analysis such as the data lineage analysis and impact analysis described above. Hierarchies of objects can also be viewed, and the hierarchies can be searched for specific objects. Once the object is found bookmarks can be created for objects allowing a user to easily return to them.

With the proper permissions, a user can edit the metadata in the metadata viewing environment. For example, a user can update descriptions of objects, create business terms, define relationships between objects (such as linking a business term to a field in a report or column in a table), move objects (for instance, moving a dataset from one application to another) or delete objects.

In FIG. 2C a corresponding lineage diagram 200C for target element 206A is shown, but the level of resolution is set to applications that are participating in the calculation for the target data element 206A. Specifically, applications 202C, 204C, 206C, 208C, and 210C are shown, as only those applications directly participate in the calculation for the target data element 206A. If a user wishes to view any part of the lineage diagram in a different level of resolution (e.g., to display more or less detail in the diagram), the user may activate the corresponding expand/collapse button 212C.

FIG. 2D shows a corresponding lineage diagram 200D at a different level of resolution. In this example, an expand/collapse button 212C has been activated by a user, and the metadata viewing environment now displays the same lineage diagram, but application 202C has been expanded to show the datasets 214D and executables 216D within application 202C.

FIG. 2E shows a corresponding lineage diagram 200E at a different level of resolution. In this example, a user has selected to show everything expanded by a custom expansion. Any field or column which is an ultimate source of data (e.g., it has no upstream systems) is expanded. In addition, fields that have a specific flag set are also expanded. In this example, the specific flags are set on datasets and fields at a key intermediate point in the lineage, and one column is the column for which the lineage is being shown. The User Interface module 108 determines which nodes need to be collapsed and which nodes need to be excluded from the diagram entirely.

Users can also configure their own diagrams. For example, diagrams can be configured so that they follow primary/foreign key relationships in the metadata. Filters may also be applied to the dependency analysis to exclude information from the lineage diagram. For example, if a user desires to exclude datasets from the lineage diagram that are reject files, the user could toggle the display of reject files in the lineage diagram on and off.

Viewing elements and relationships in the metadata viewing environment can be made more useful by adding information relevant to each of the nodes that represent them. One exemplary way to add relevant information to the nodes is to graphically overlay information on top of certain nodes. These graphics may show some value or characteristic of the data represented by the node, and can be any property in the

metadata database. This approach has the advantage of combining two or more normally disparate pieces of information (relationships between nodes of data and characteristics of the data represented by the nodes) and endeavors to put useful information “in context.” For example, characteristics such as metadata quality, metadata freshness, or source of record information can be displayed in conjunction with a visual representation of relationships between data nodes. While some of this information may be accessible in tabular form, it may be more helpful for a user to view characteristics of the data along with the relationships between different nodes of data. A user can select which characteristic of the data will be shown on top of the data element and/or transformation nodes within the metadata viewing environment. Which characteristic is shown can also be set according to default system settings.

In the example of FIG. 3, node 300 also displays a graphical overlay 302 that contains information pertaining to the freshness of the metadata represented by the node. The “metadata freshness” refers to how recently the metadata has been updated or modified from an external source. By “hovering” a cursor over graphical overlay 302, a window 304 can be called up that contains more detail about the characteristic currently displayed by the graphical overlay 302. The graphical overlays may be color-coded, with the different colors of the graphics mapping to different meanings via legend 306.

In the example of FIG. 4, graphical overlays representing levels of metadata quality are superimposed on top of data element nodes including overlay 402 on node 400. Measures of metadata quality can be used by a business, for example, to profile a periodic (e.g., monthly) data feed sent from a business partner before importing or processing the data. This would enable the business to detect “bad” data (e.g., data with a percentage of invalid values higher than a threshold) so it doesn’t pollute an existing data store by actions that may be difficult to undo. Like the previous example, by hovering a cursor over graphical overlay 402, a window 404 can be called up that contains more detail about the characteristic currently displayed by the graphical overlay 402.

In the example of FIG. 5, graphical overlays representing the type of the source of record are superimposed on top of data element and transformation nodes. Node 500 has an overlay 502 that indicates the source of record is an “Active Import Source.” This means that the metadata was imported automatically from a source such as a spreadsheet

file. By hovering a cursor over graphical overlay 502, a window 504 can be called up that contains details such as the type of file used for the import (an Excel spreadsheet in this example), the name of the file, the owner of the file, and the date of the import.

Node 506 has an overlay 508 that indicates the source of record is “Manually Maintained.” This means that the metadata was modified manually by a user (e.g., using the User Interface Module 108). By hovering a cursor over the graphical overlay 508, a window 510 can be called up that contains details such as the name of the user that modified the metadata and the date of the modification.

The record storage and retrieval approach described above, including the modules of the system 100 and the procedures performed by the system 100, can be implemented using software for execution on a computer. For instance, the software forms procedures in one or more computer programs that execute on one or more programmed or programmable computer systems (which may be of various architectures such as distributed, client/server, or grid) each including at least one processor, at least one data storage system (including volatile and non-volatile memory and/or storage elements), at least one input device or port, and at least one output device or port. The software may form one or more modules of a larger program, for example, that provides other services related to the design and configuration of computation graphs. The nodes and elements of the graph can be implemented as data structures stored in a computer readable medium or other organized data conforming to a data model stored in a data repository.

The approaches described above can be implemented using software for execution on a computer. For instance, the software forms procedures in one or more computer programs that execute on one or more programmed or programmable computer systems (which may be of various architectures such as distributed, client/server, or grid) each including at least one processor, at least one data storage system (including volatile and non-volatile memory and/or storage elements), at least one input device or port, and at least one output device or port. The software may form one or more modules of a larger program, for example, that provides other services related to the design and configuration of computation graphs. The nodes and elements of the graph can be implemented as data structures stored in a computer readable medium or other organized data conforming to a data model stored in a data repository.

The software may be provided on a storage medium, such as a CD-ROM, readable by a general or special purpose programmable computer or delivered (encoded in a propagated signal) over a communication medium of a network to the computer where it is executed. All of the functions may be performed on a special purpose
5 computer, or using special-purpose hardware, such as coprocessors. The software may be implemented in a distributed manner in which different parts of the computation specified by the software are performed by different computers. Each such computer program is preferably stored on or downloaded to a storage media or device (e.g., solid state memory or media, or magnetic or optical media) readable by a general or special
10 purpose programmable computer, for configuring and operating the computer when the storage media or device is read by the computer system to perform the procedures described herein. The inventive system may also be considered to be implemented as a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer system to operate in a specific and
15 predefined manner to perform the functions described herein.

A number of embodiments of the invention have been described. Nevertheless, it will be understood that various modifications may be made without departing from the scope of the invention. For example, some of the steps described above may be order independent, and thus can be performed in an order different from that described.

20 It is to be understood that the foregoing description is intended to illustrate and not to limit the scope of the invention, which is defined by the scope of the appended claims. For example, a number of the function steps described above may be performed in a different order without substantially affecting overall processing. Other embodiments are within the scope of the following claims.

60412-447D1

CLAIMS:

1. A method including:

receiving metadata objects from one or more sources;

grouping the metadata objects into one or more first groups associated with a
5 first context;

grouping one or more of the first groups into one or more second groups
associated with a second context having a higher hierarchy level than the first context;

grouping one or more of the second groups into one or more third groups
associated with a third context having a higher hierarchy level than the second context; and
10 generating a visual representation of a diagram including nodes and groups of
nodes, each node representing one of the metadata objects or one of the first, second, and third
groups, the visual representation showing at least one of the first group of nodes within one of
the second groups, and at least one of the second groups within one of the third groups.
2. The method of claim 1 in which the first context includes datasets context, the
15 second context includes applications context, and the third context includes systems context.
3. The method of claim 1 in which the first context includes executables context,
the second context includes applications context, and the third context includes systems
context.
4. The method of claim 1 in which the diagram is a data lineage diagram
20 showing, for a given object, other objects that depend on the given object and other objects
that affect the given object.
5. A system including:

a computer including one or more processors;

60412-447D1

means for receiving metadata objects from one or more sources;

means for grouping the metadata objects into one or more first groups associated with a first context;

5 means for grouping one or more of the first groups into one or more second groups associated with a second context having a higher hierarchy level than the first context;

means for grouping one or more of the second groups into one or more third groups associated with a third context having a higher hierarchy level than the second context; and

10 means for generating a visual representation of a diagram including nodes and groups of nodes, each node representing one of the metadata objects or one of the first, second, and third groups, the visual representation showing at least one of the first group of nodes within one of the second groups, and at least one of the second groups within one of the third groups.

6. The system of claim 5 in which the first context includes datasets context, the
15 second context includes applications context, and the third context includes systems context.

7. The system of claim 5 in which the first context includes executables context, the second context includes applications context, and the third context includes systems context.

8. The system of claim 5 in which the diagram is a data lineage diagram showing,
20 for a given object, other objects that depend on the given object and other objects that affect the given object.

9. A computer system including:

one or more computing devices configured to:

receive metadata objects from one or more sources;

60412-447D1

group the metadata objects into one or more first groups associated with a first context;

group one or more of the first groups into one or more second groups associated with a second context having a higher hierarchy level than the first context;

5 group one or more of the second groups into one or more third groups associated with a third context having a higher hierarchy level than the second context; and

generate a visual representation of a diagram including nodes and groups of nodes, each node representing one of the metadata objects or one of the first, second, and third groups, the visual representation showing at least one of the first group of nodes within one of
10 the second groups, and at least one of the second groups within one of the third groups.

10. The computer system of claim 9 in which the first context includes datasets context, the second context includes applications context, and the third context includes systems context.

11. The computer system of claim 9 in which the first context includes executables
15 context, the second context includes applications context, and the third context includes systems context.

12. The computer system of claim 9 in which the diagram is a data lineage diagram showing, for a given object, other objects that depend on the given object and other objects that affect the given object.

20 13. A computer-readable device storing a computer program, the computer program including instructions for causing a computer to:

receive metadata objects from one or more sources;

group the metadata objects into one or more first groups associated with a first context;

60412-447D1

group one or more of the first groups into one or more second groups associated with a second context having a higher hierarchy level than the first context;

group one or more of the second groups into one or more third groups associated with a third context having a higher hierarchy level than the second context; and

- 5 generate a visual representation of a diagram including nodes and groups of nodes, each node representing one of the metadata objects or one of the first, second, and third groups, the visual representation showing at least one of the first group of nodes within one of the second groups, and at least one of the second groups within one of the third groups.

14. The computer-readable device of claim 13 in which the first context includes
10 datasets context, the second context includes applications context, and the third context includes systems context.

15. The computer-readable device of claim 13 in which the first context includes executables context, the second context includes applications context, and the third context includes systems context.

- 15 16. The computer-readable device of claim 13 in which the diagram is a data lineage diagram showing, for a given object, other objects that depend on the given object and other objects that affect the given object.

1/9

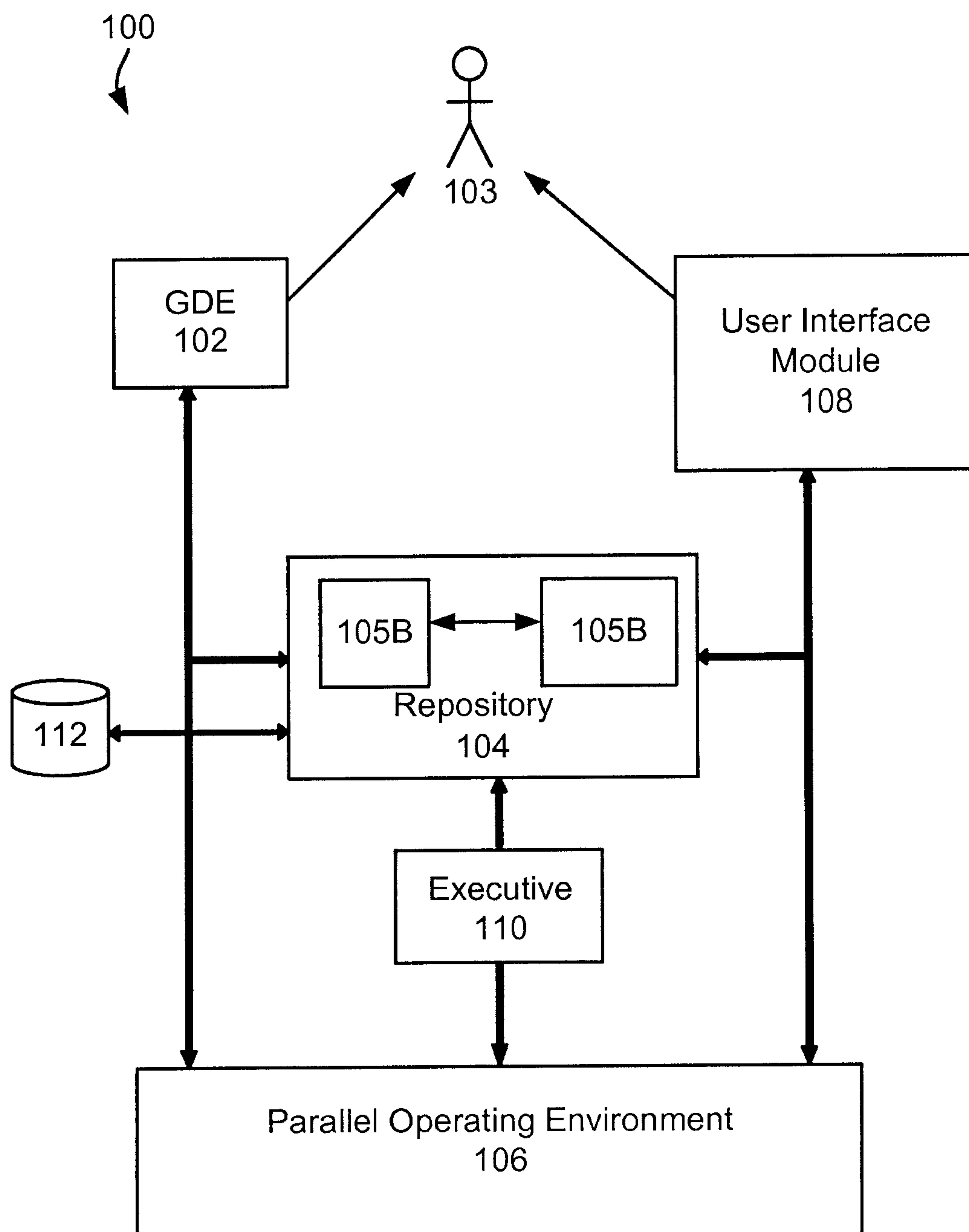
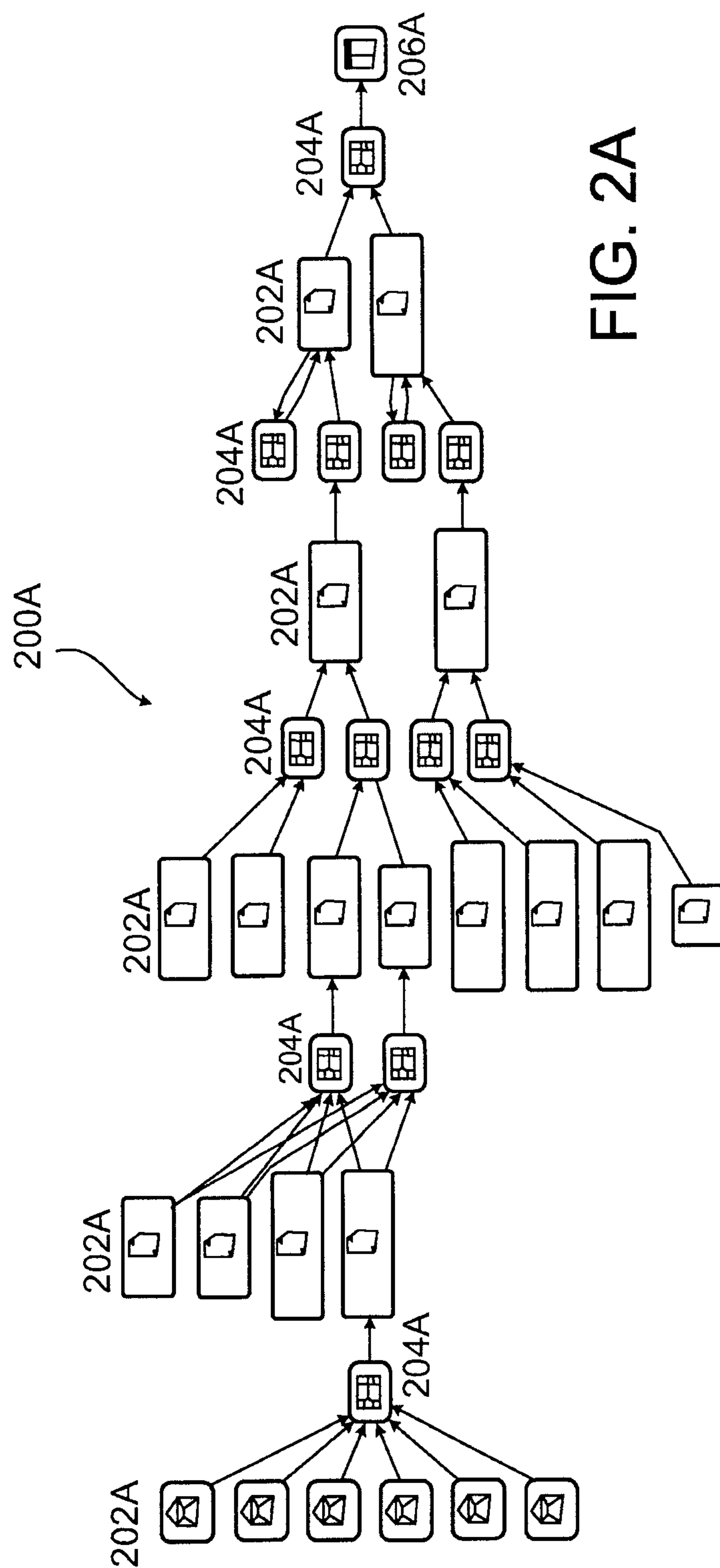
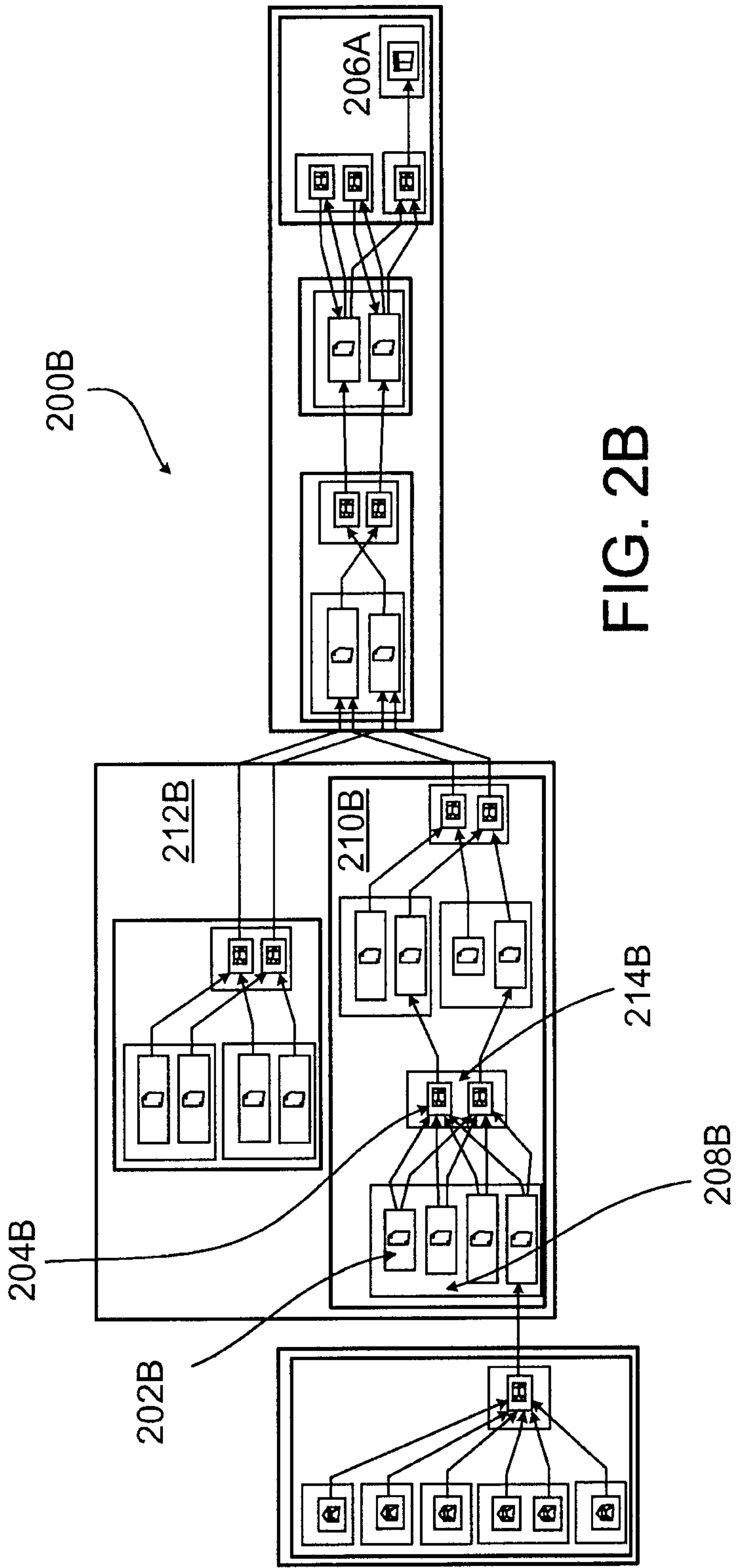


FIG. 1





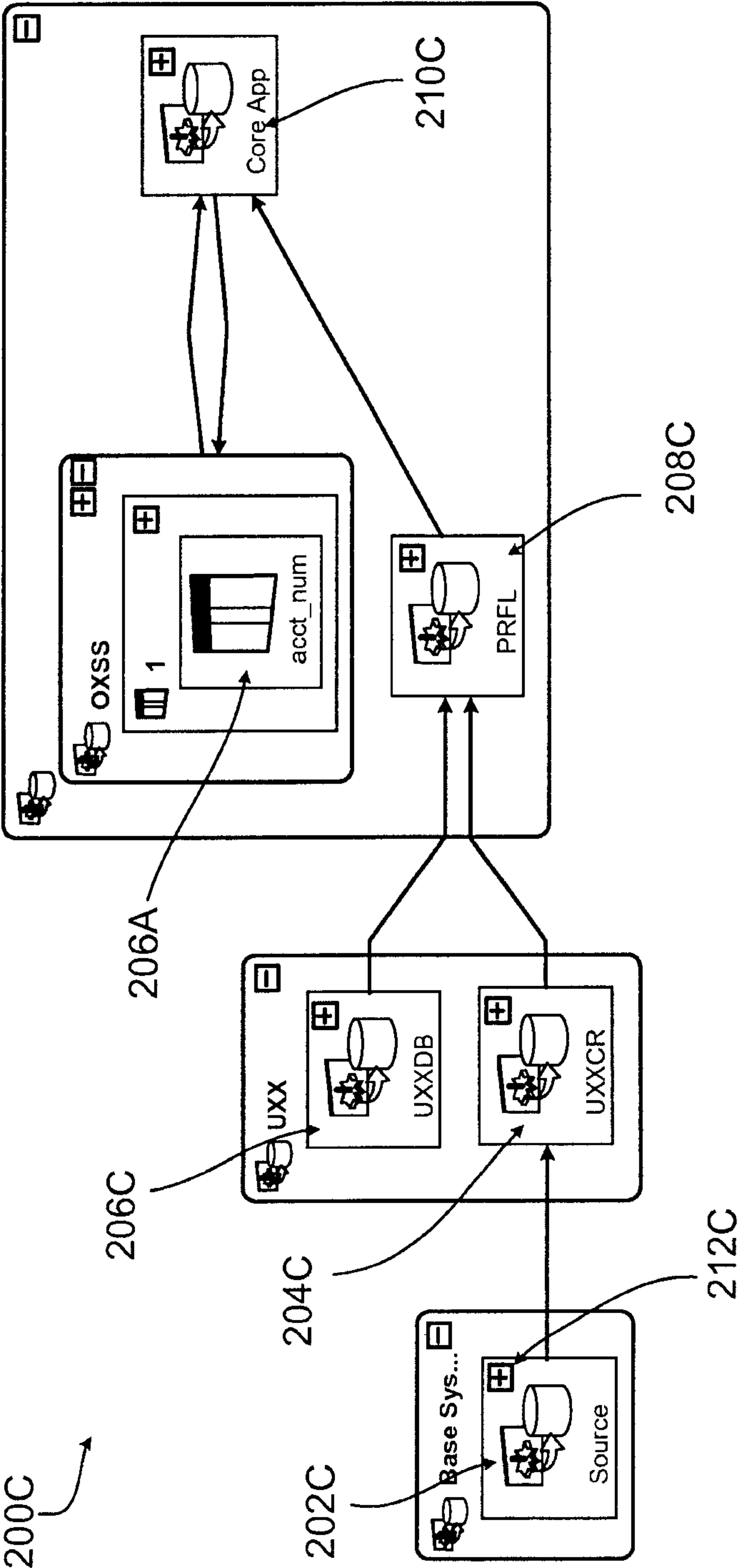


FIG. 2C

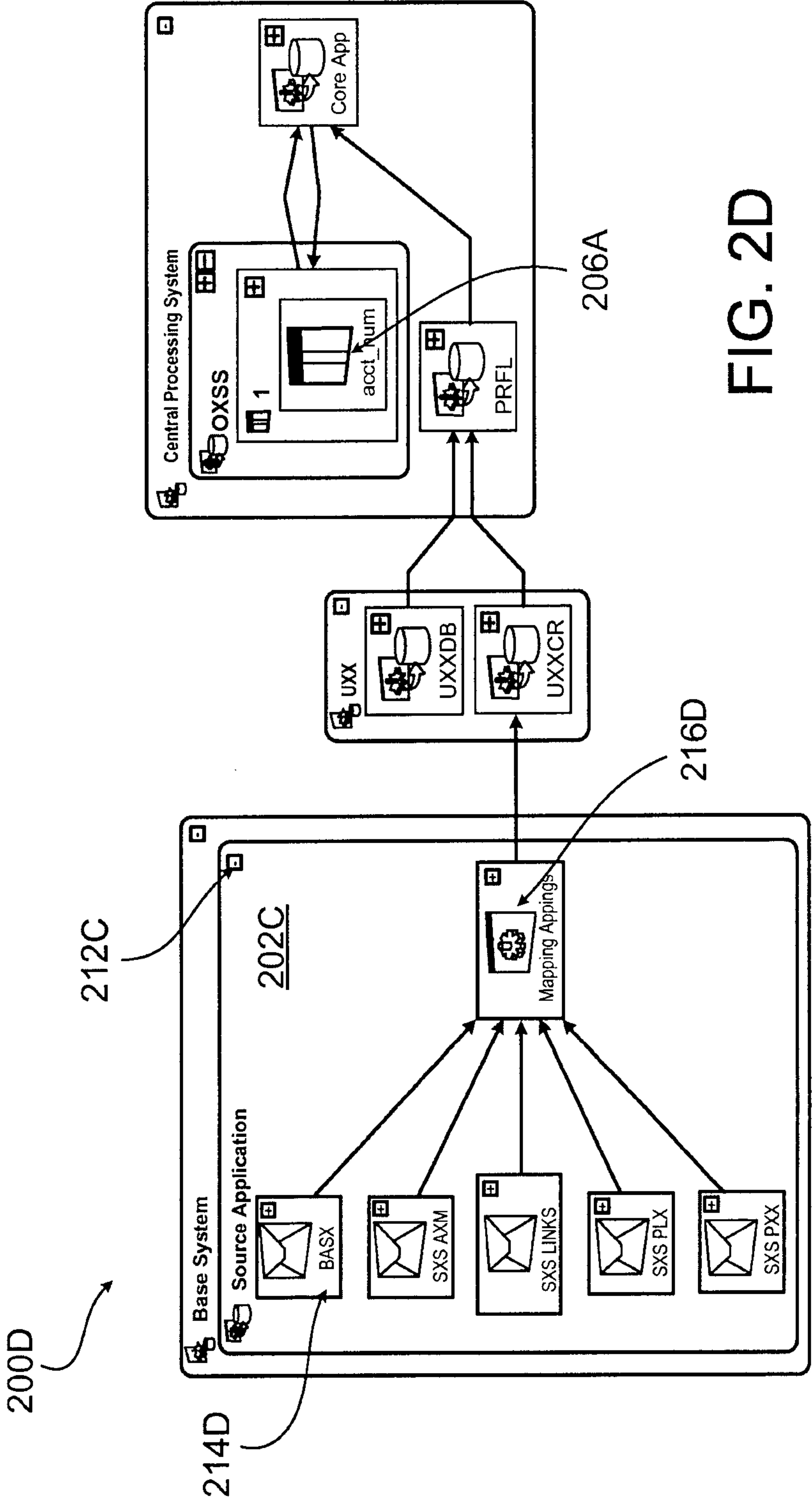


FIG. 2D

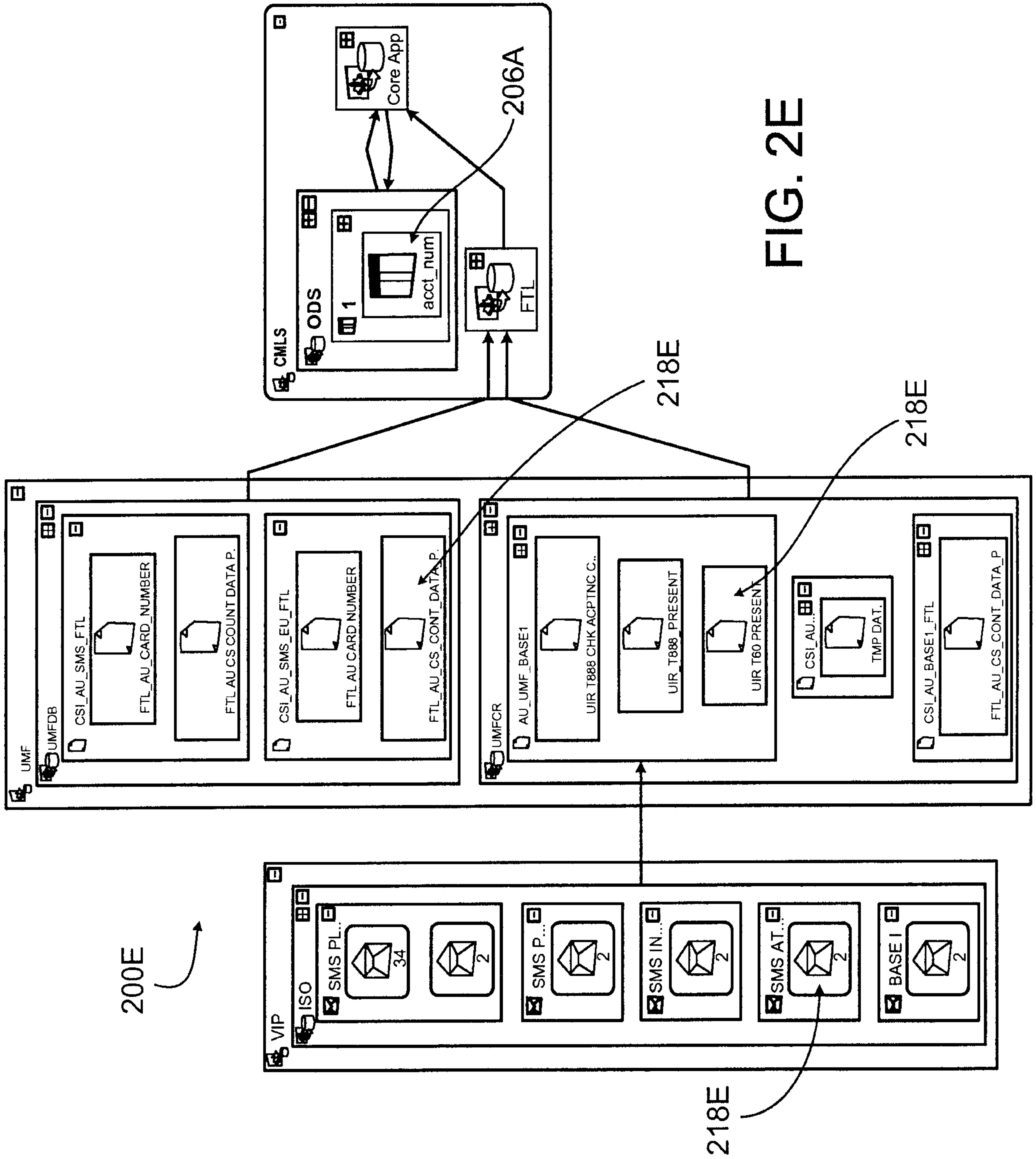


FIG. 2E

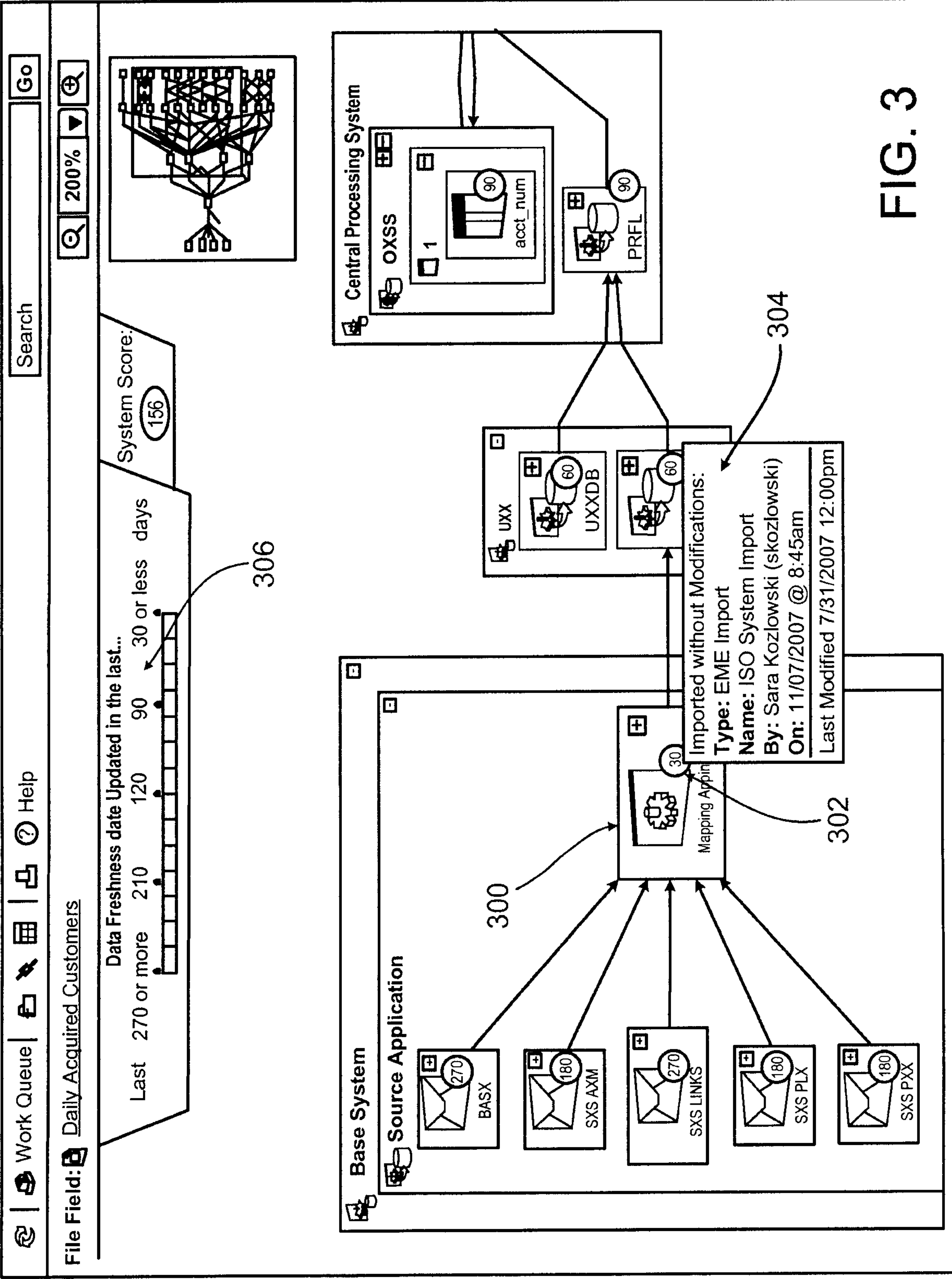


FIG. 3

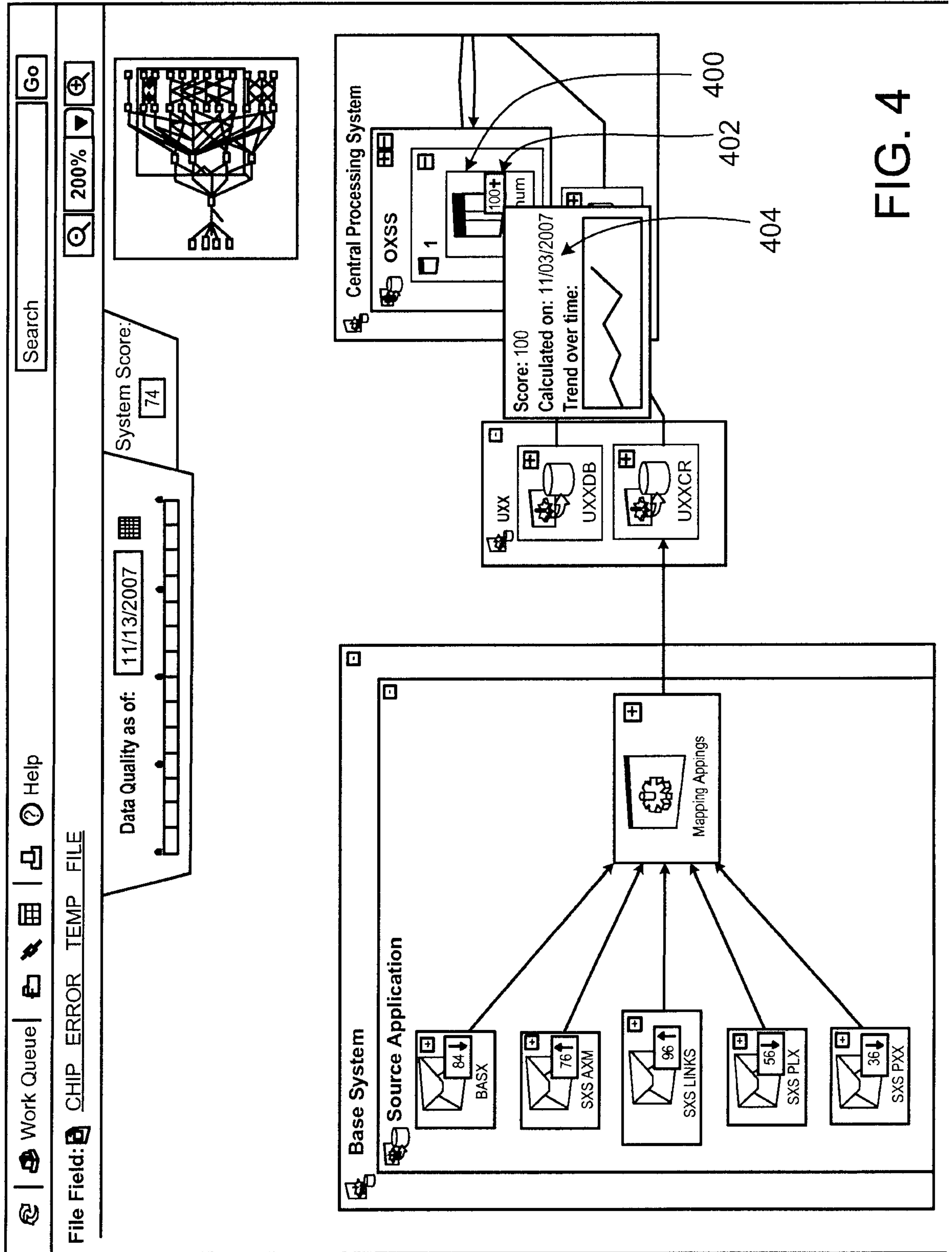


FIG. 4

