(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) **EP 3 204 858 B9**

(12) **CORRECTED EUROPEAN PATENT SPECIFICATION**

(15) Correction information:
Corrected version no 1 (W1 B1)
Corrections, see
Description Paragraph(s) 37

(48) Corrigendum issued on:
31.03.2021 Bulletin 2021/13

(45) Date of publication and mention of the grant of the patent:
09.09.2020 Bulletin 2020/37

(21) Application number: **15875851.6**

(22) Date of filing: **07.10.2015**

(51) Int Cl.:
G06F 21/62 *(2013.01)* ***G06F 11/30*** *(2006.01)*
G16H 10/60 *(2018.01)*

(86) International application number:
PCT/US2015/054444

(87) International publication number:
WO 2016/108983 (07.07.2016 Gazette 2016/27)

(54) **HIGHLY SECURE NETWORKED SYSTEM AND METHODS FOR STORAGE, PROCESSING, AND TRANSMISSION OF SENSITIVE PERSONAL INFORMATION**

HOCHSICHERES NETZWERKSYSTEM UND VERFAHREN ZUR SPEICHERUNG, VERARBEITUNG UND ÜBERTRAGUNG SENSIBLER PERSÖNLICHER DATEN

SYSTÈME EN RÉSEAU HAUTEMENT SÉCURISÉ ET PROCÉDÉS DE STOCKAGE, DE TRAITEMENT ET DE TRANSMISSION D'INFORMATIONS PERSONNELLES SENSIBLES

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR

(30) Priority: **07.10.2014 US 201462060898 P**

(43) Date of publication of application:
16.08.2017 Bulletin 2017/33

(60) Divisional application:
20189588.5 / 3 767 472

(73) Proprietor: **Optum, Inc.**
Minnetonka, Minnesota 55343 (US)

(72) Inventor: **EDISON, Robin**
Highland Park, Illinois 60035 (US)

(74) Representative: **Banzer, Hans-Jörg**
Kraus & Weisert
Patentanwälte PartGmbB
Thomas-Wimmer-Ring 15
80539 München (DE)

(56) References cited:
US-A1- 2002 073 138
US-A1- 2005 268 094
US-A1- 2007 145 123
US-A1- 2009 300 770
US-A1- 2012 110 469
US-A1- 2012 303 616
US-B1- 7 013 315
US-B1- 7 870 614

EP 3 204 858 B9

Description

FIELD OF TECHNOLOGY

[0001] The present disclosure relates generally to secure networks, and more particularly to securing information within a network of computers.

BACKGROUND

[0002] Various industries collect personal information about individuals. Regardless of the particular industry, personal information is considered highly sensitive and often needs to be secured in an attempt to prevent unauthorized access. For example, in the healthcare context, healthcare providers often use locally maintained computing platforms to manage and secure patient records. These locally maintained computer platforms allow the healthcare providers to satisfy their duties under industry and government standards, such as the Health Insurance Portability and Accountability Act ("HIPAA").

[0003] Often times, researchers and other third parties desire access to personal information. In the healthcare industry, personal information cannot be released to researchers until it is secured. However, modern computer technology frequently makes it possible to reverse engineer the "secure" personal information into its unsecure source data. This is undesirable because it gives third parties unauthorized access to sensitive personal information.

[0004] US 2002/073138 A1 discloses de-identification and linkage of data records used in hospitals and banks, wherein a sever receives from client computers de-identified data records and compares the received de-identified records with master records of database in order to determine which records of the de-identified records and master records are to be linked.

[0005] US 7 870 614 B1 discloses a processing method for using data sets associated with transaction card identifiers, by maintaining an association between alias identifiers and transaction card identifiers for secured access and decoding.

[0006] US 2005/268094 A1 discloses encrypting multi-sourced patient data records to overcome data source variances in individual encryption techniques and in the content of data records, wherein de-identified data records received from multiple data sources are assembled in a longitudinal database for market research and other analysis.

SUMMARY

[0007] The present invention defines a secure computing environment according to claim 1 and a method according to claim 11. Further embodiments are set forth in the dependent claims 2-10 and 12-15.

[0008] The present disclosure generally provides a highly secure networked system and methods for storage, processing, and transmission of sensitive, such as personal/private, information. Data contributor computing environments store sensitive personal information. The personal information is cleansed by a data contributor computing environment using specific rules/logic provided to data contributor computing environments by a secure facility computing environment. The secure facility computing environment transmits the same rules/logic to each data contributor computing environment. Thus, cleansing of data at each data contributor computing environment occurs the same way.

[0009] The cleansed personal information is salted (using a public salt) and hashed using a one-way hashing algorithm. Each data contributor computing environment uses the same public salt and one-way hashing algorithm. This ensures that cleansed personal information for a particular individual or grouping of individuals associated with any given contributor computing environment is hashed into an identically structured hashed field or record across multiple data contributor computing environments (i.e. all contributing computing environments provide data to be secured from that facility structured in the same manner).

[0010] Data (at least partially including the hashed personal information) is communicated from the data contributor computing environments to the secure facility computing environment. The hashed personal information is therein salted and hashed a second time using a private salt and a one-way hashing algorithm. The private salt is maintained solely within the secure facility computing environment, and is not communicated to or otherwise accessible to third parties (i.e., data contributor computing environments and/or third party, e.g. researcher, computing devices). This further secures the private data in a manner that prevents third parties from re-identifying the second hashed personal information into its unhashed source data.

[0011] The second hashed personal data value is linked to previously second hashed data values (when possible) and is assigned a unique ID. This linking of the data values to a unique ID provides a robust secure data platform, e.g. for research or data mining, because unhashed data of one data contributor computing environment is linked and accessible along with related data of other data contributor computing environments via the unique IDs.

[0012] In one embodiment, particular individuals provided access to the highly secure information such as researchers, desiring access to data, identify parameters for a research project. In response, the secure facility computing environment

compiles data from various sources (linked using unique IDs) into a data dictionary. The data dictionary undergoes compliance and statistical analyses related to the potential of a researcher re-identifying the unhashed source data. When the data dictionary satisfies the compliance and statistical analyses, the data dictionary is stored and presented to the researcher as a certified view.

[0013] A researcher may be granted access to one or more certified views via a virtual private network. Each certified view may be presented through a different virtual desktop interface. Access to the certified views, via the virtual private network, by a researcher computer may involve a two factor authentication process. In this manner, a technical solution is provided to permit, among other things, stores of private data to be securely maintained as private data while otherwise being made available in further secure fashion to certified users in a limited and controlled manner.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] Embodiments of devices, systems, and methods are illustrated in the figures of the accompanying drawings, which are meant to be exemplary and non-limiting, in which like references are intended to refer to like or corresponding parts, and in which:

FIG. 1 illustrates an embodiment of a secure networked system for securing personal/private information and providing limited/controlled access to the secure personal information according to the present disclosure;
FIG. 2A is a process flow diagram illustrating a method for transforming personal information into secure personal information in the network according to the present disclosure;
FIG. 2B is a process flow diagram further illustrating the method of FIG. 2A for transforming personal information into secure personal information by two separate and distinct data contributors in the network according to the present disclosure;
FIG. 3 is a process flow diagram illustrating a method for matching and linking hashed data values of an identity vault using unique IDs according to the present disclosure;
FIGS. 4A and 4B are process flow diagrams illustrating a method for rekeying/restating previously generated unique IDs according to the present disclosure;
FIG. 5 illustrates an example data structure representation of unique IDs and associated concatenated hashed values within the identity vault according to the present disclosure;
FIG. 6 illustrates an example of how related claims data and clinical record data may be associated across various data storages using a unique ID in a healthcare data embodiment of a network according to the present disclosure; and
FIG. 7 illustrates a system architecture of the virtual private network (VPN) illustrated in FIG. 1 according to the present disclosure.

DETAILED DESCRIPTION

[0015] The detailed description of the present disclosure set forth herein makes reference to the accompanying drawings, which show various embodiments by way of illustration. While these various embodiments are described in sufficient detail to enable those skilled in the art to practice the disclosure, it should be understood that other embodiments may be realized and that logical and physical changes may be made without departing from the spirit and scope of the disclosure. Thus, the detailed description herein is presented for purposes of illustration only and not of limitation. For example, the steps recited in any of the method or process descriptions may be executed in an order other than as presented and are not limited to the order presented. Moreover, references to a singular embodiment may include plural embodiments, and references to more than one component may include a singular embodiment.

[0016] The present disclosure provides extensive and specialized processes for transforming personal information into secure data. The present disclosure additionally provides particularized databases containing vast amounts of de-identified medical claims, electronic medical records, and other data, such as sociodemographic information collected from consumer research databases structured to provide automated access to vast amounts of information. While the illustrative embodiments described herein may relate to healthcare information, it should be appreciated that secure systems as described may be implemented according to the disclosure for protection and transformation of other types of sensitive information such as financial information, legal information, national security information, or the like (generally referred to herein as "personal" or "private" information). The sophisticated, specialized, and specially configured networked systems and processes described herein, for example, may facilitate data driven healthcare research without compromising privacy with respect to sensitive personal information (again, the information protected, could just as readily be other information requiring high levels of security, privacy and control while facilitating controlled access for manipulation and processing). Specifically, the systems in the illustrated embodiment are configured to provide, among other things, high levels of security of vast amounts of personal/private data while simultaneously providing access to the secure personal/private data in a way that benefits the healthcare industry and its patients.

[0017] Protection and security of personal digital data is a persistent technical problem that arose after the advent of the computer. The problem is exacerbated by the proliferation of use of digital devices and the volumes of data collected, accessible and processed by such devices. An additional technical problem relates to maintaining security and privacy of vast amounts of personal data while providing digital access to the data for research aimed at furthering use and utility of the vast amounts of data, such as for enhancing consumer and patient outcomes and the delivery of healthcare. It will become apparent from reading the below detailed description that these technical problems are solved by the various teachings herein.

[0018] Referring to FIG. 1, a secure networked system 100 for securing personal information and providing limited and controlled access to the secure data is described. The system 100 includes one or more data contributor computing environments 102. Each data contributor computing environment 102 includes a data storage 104 that houses clear text data. For example, the clear text data may include healthcare claims data and clinical data that are each partially made up of personal information. The clear text healthcare claims data may be separate and distinct data from the clear text clinical data. Each data contributor computing environment 102 also includes a processor 106 that is in communication with the data storage 104 and that is configured to sanitize or cleanse and hash the personal information as described herein below with respect to FIGS. 2A and 2B.

[0019] Generally, the cleansed, hashed personal information and associated clear text non-personal information is communicated to a data receipt zone 110 of a secure facility computing environment 108. Transmission of the data may occur actively by the processor 106 (i.e., the data receipt zone 110 passively receives the data). Alternatively, transmission of the data may occur passively with respect to the data contributor computing environment 102 (i.e., a processor of the secure facility computing environment 108 may actively retrieve the data from the data contributor computing environment 102). The transmission of such data may happen in real time or it may happen in batch processes. At the data receipt zone 110, the data is staged and prepared for subsequent processing as disclosed herein below.

[0020] The cleansed data is loaded into and stored within respective storage or databases. For example, healthcare claims data and clinical data may be stored within databases, such as a de-identified claims storage 112 and a de-identified clinical storage 114 of an extract, transform, and load (ETL) zone 116 of the secure facility computing environment 108, respectively. A processor 118 of the ETL zone 114 hashes the previously hashed personal information a second time, as described herein below with reference to FIGS. 2A and 2B.

[0021] The second hashed personal information values are stored within an identity vault 112. The newly hashed and stored personal information values are thereafter "linked" by the processor 118 as described with reference to FIG. 3 below. Once a unique ID is associated with a recently stored personal information value, the same unique ID is associated with the cleansed and second hashed data (within the various storages) to which it relates. For example, when the data is healthcare data, the unique ID is associated within the cleansed, second hashed claims data and clinical data stored within the de-identified claims and clinical storages 116, 118.

[0022] Prior to the cleansed, second hashed data being stored within various storage databases, it may be processed to change encounter IDs, provider IDs, and facility IDs, provided by the data contributor computing environment(s) 102, into sequential alternate IDs. This helps eliminate trace back of the de-identified cleansed data to its source data stored within the data storage 104 of the data contributor computing environment 102.

[0023] When a researcher, via a researcher computing device 122, indicates they want to conduct research, a "data dictionary" is created from the clinical and claims data within the de-identified claims and clinical storages 112, 114. The data dictionary is essentially a description of the data presented in a certified view. The data that makes up a data dictionary may be specific to a geographic area/location, membership (e.g., membership in a particular coverage plan or program), a data domain, data type, or some other interest defined by the researcher. For example, an illustrative data dictionary may contain indicators such as table name, field name, business description, and various parameters including data type, field length, and valid values.

[0024] The generated data dictionary is communicated to a data storage zone 124 where the data dictionary is analyzed for compliance with respect to clarity/proprietary standards. The data storage zone 124 includes an admin schema storage 126 that stores rules/logic relating to clarity/proprietary standards that data must comply with prior to being viewable by a researcher. One illustrative rule identifies fields of data that present re-identification risk and that should be removed from the data dictionary. Another illustrative rule identifies fields that present re-identification risk and contain values that should be consolidated to lessen the risk. An illustrative and non-limiting list of such data values includes bed size, discharge status, and admit channels. A further illustrative rule identifies fields that require standard language to comply with auditing. Yet another illustrative rule may identify fields that are required to be flagged as potentially identifying personal information characteristics.

[0025] Various statistical de-identification rules/logic may also be applied to the data dictionaries while in the data storage zone 124. The statistical de-identification rules/logic use statistical methodology to evaluate data of a data dictionary as well as public records to, through reduction of data and consolidation of values, achieve a statistical determination of a very small risk of re-identification. A threshold risk value of re-identification of de-identified data may be set as a function of, for example, geographic location, number of patients, age, timing of disease occurrence, number

of providers, or related data. The statistical de-identification rules/logic may be stored within the admin schema storage 126 or a separate storage (not illustrated).

[0026] When a data dictionary passes the compliance standards, as set out by the rules/logic of the data storage zone 124, the data dictionary is communicated to a data access zone 128 where it is stored in a certified view storage 130. Each certified view storage 130 may include only one data dictionary. Alternatively, each certified view storage 130 may contain all of the certified views approved for viewing by a single researcher, or certified views may require no storage. A researcher can view different certified views from different desktops in the manner of one and only one view per desktop. This prevents unauthorized access to certified views by unapproved researchers. While there may be overlap (e.g. common data) between two certified views it may be desirable for security purposes that each certified view is wholly independent of another.

[0027] A researcher may indicate they desire a previously approved certified view to be augmented to include more types of data. When this occurs, the other data types are identified within the data storages 112, 114. This data is then compiled into the original data of the certified view. The updated data compilation thereafter undergoes the aforementioned compliance and statistical analyses described with respect to the generation of a new data dictionary. If the augmented certified view is analyzed to have a very small risk of source data re-identification, it is stored as an updated certified view and is published to the researcher(s) authorized to view the original certified view.

[0028] The secure facility computing environment 108 also includes a virtual private network (VPN) 132 that is configured to restrict researcher computer 122 access to certain certified views. The structure of the VPN 132 and its interaction with the data access zone 128 and researcher computers 122 is described in further detail below with respect to FIG. 7.

[0029] Attention is now given to FIG. 2A, which illustrates a method 200 for transforming and securing personal/private/sensitive information (all referred to herein as "personal information"), according to the present disclosure. At block 202 a data contributor computing environment 102 standardizes/sanitizes clear text of personal information. At block 204 the data contributor computing environment 102 salts and hashes the standardized/sanitized personal information. At block 206 the secure facility computing environment 108 thereafter receives the cleansed information from the contributor computing environment and salts and hashes the previously cleansed/hashed personal information values. At block 208 the secure facility computing environment 108 stores the second salted and hashed personal information values within an identity vault. The identity vault contains previously stored second salted and hashed personal information values associated with unique IDs. At block 210 the secure facility computing environment 108 links the stored second salted and hashed personal information values with related previously stored hashed values (if possible as described herein below) and associates the second salted and hashed personal information values with associated unique IDs.

[0030] Each of the blocks illustrated in the method 200 of FIG. 2A involve multiple processing steps, which are more specifically described with reference to FIG. 2B. FIG. 2B illustrates the transformation and securing of personal information data by two separate and distinct data contributor computing environments 102. Corresponding blocks/processes of the data contributor computing environments 102 are collectively described and numbered herein because the processing of personal information data by different data contributor computing environments 102 is exactly the same. This ensures that the processing of personal information data by different data contributor computing environments 102 generates substantially similarly structured output hashed values that can be successfully linked by the secure facility computing environment 108.

[0031] Each data contributor computing environment 102 standardizes/sanitizes the personal information clear text using specific rules/logic (illustrated as 212). An illustrative and non-limiting list of personal information includes social security number (SSN), first name, last name, and address. While not being personal information, contributor IDs are also salted and hashed according to the teachings herein described. In the context of the healthcare industry, personal information may additionally include Medicare ID, Medicaid recipient number, and Medicaid family number, for example. One illustrative sanitization rule may change the clear text (which may include both upper and lower case characters) into purely upper case or purely lower case text. This rule ensures that hashing of the clear text produces identical values because, for example, "E" produces a different hashed value than "e". Another illustrative rule removes punctuation marks/characters from the clear text. Yet another illustrative rule removes spaces within the clear text. A further illustrated rule removes line ends and carriage returns from the clear text.

[0032] The standardized/sanitized text is salted using a "common salt" (illustrated as 214). The term "common salt," as used herein, refers to a single salt used by every data contributor computing environment 102. "Salting" as known to those skilled in the art is a cryptography concept that involves providing additional data, such as a common random number, as an additional input (e.g. along with confidential or private information) to a one-way function that hashes the private information.

[0033] The concatenation of the salt and sanitized clear text of the personal information becomes the message digest (illustrated as 216) that is input to and hashed using a hashing algorithm (illustrated as 218). For example, the hashing algorithm may be a one-way hashing algorithm such as SHA-256. However, one skilled in the art should appreciate that different one-way hashing algorithms may be used, depending upon the data to be hashed and the security level to be

achieved. The resulting personal information hashed values (illustrated as 220) may be represented as 64 character values.

[0034] For illustration, Table 1 below depicts illustrative pre-sanitization clear text personal information and Table 2 depicts the resulting hashed values of the sanitized and salted protected personal information.

Table 1. Illustrative pre-sanitization clear text personal information.

FIRST_NAME	LAST_NAME	SSN
Joseph	Smith	000000000
John	Smith	111111111
Jack	Smith	222222222

Table 2. Illustrative resulting hashed values of the sanitized and salted personal information of Table 1.

FIRST_NAME_E	LAST_NAME_E	SSN_E
12343cd576727e759d144ec497e 19ecem089cdmdb11578ae10a9 456092mabafc1d7d	123m812ae83f17a8488bbm9f5f0 d2m 74c46m94e2f42a6bd76d966 14f71a6fd1efac	123fc7cb8eff142711b4d1e2e7a ead2aa6f4ce605481632a7430e0 m467c79a778b
1239a57m93e4f368d62e7125890 ae6b5230b59024468ff85f7f8cf84 43a1c8e6533	1231dd4m4e87f2486ee1e639a38 b3e5942d68a9118b107354abacb 9me34a37884cb	123503bb674edddd4bcfda8865a d3acfb851555m44ce8325eb47be 39f72f7bef345
123e0ead436f8fem1f52b91621ce 9b524m53c6c5051fbbfc18cfc41 744cb11m742	123c2edcd9b53e3b6966efb2b20 1f52441dd677d2987bcca3d0635 77e882cb6m1d9	123f120bb5698d520c5691b6d60 3am0bfd662d13bf177a04571f9d1 mcm745dfa2a5

[0035] Once the hashed values (illustrated as 220) are communicated to and received by the secure facility computing environment 108, the hashed values are salted using a "private salt" (illustrated as 222). As used herein, the term "private salt" refers to a salt maintained within and used solely by the secure facility computing environment 108. The private salt is not communicated to the data contributor computing environments 102. This provides an additional level of security that prevents reverse engineering of hashed values into their unhashed source data.

[0036] The salted values (illustrated as 222) are hashed a second time using a one-way hashing algorithm (illustrated as 224). The second hashing of the data may produce 64 character values. As with the first hashing of the data (illustrated as 218), the SHA-256 hashing algorithm may also be used by the secure facility computing environment 108. Likewise, one skilled in the art should appreciate that different hashing algorithms may be used depending upon the type of data to be hashed and the security level to be achieved. One skilled in the art should appreciate that the hashing algorithm(s) used to hash the data (illustrated as 218 and 224) may be the same or different depending upon implementation of the present disclosure.

[0037] For illustration, Table 3 below depicts resulting second hashed values of the first hashed values of Table 2. A comparison of Tables 2 and 3 illustrates that the hashed values generated by the data contributor computing environments 102 are different from those generated by the secure facility computing environment 108. This is a direct result of the use of the different salts (i.e., public and private). This diminishes, or eliminates, the potential of the second hashed data being re-identified into its clear text source data.

Table 3. Illustrative resulting second hashed values of the first hashed values of Table 2.

FIRST_NAME_E	LAST_NAME_E	SSN_E
1238ca657102ae567mc8845167 357a78a1cd321342ca09d62a28a m53a7e5830fd89	123m812ae83f17a8488bbm9f5f0 d2m 74c46m94e2f42a6bd76d966 14f71a6fd1efac	123fcee4cf166e0f1d0c11033894 2f6a8cb1049e78c594ef49962aed 348cdda3e63
123bff7m485135c547322b67779 2b39m0f8b08bcd8896m9m3699 a3bmb6e7b6e205a	1231dd4m4e87f2486ee1e639a38 b3e5942d68a9118b107354abacb 9me34a37824cb	1237546m08f4857e29c90f3ffc3e 7509ee072b9234b38a2f3a1a9d5 3em0641d6456

(continued)

FIRST_NAME_E	LAST_NAME_E	SSN_E
1232e6c27e28c72a14a20c07e13 8d45d7fca52333f673f673f560fc4678c 734db8ddaf1	123c2edcd9b53e3b6966efb2b20 1f52441dd677d2987bcca3d0635 77e882cb6m1d9	123b5e932e72f942c27ea570a1 946de34b10463f55ef366c89556a 2a5bf372e634

[0038] The different fields run through the hashing algorithm may be tagged with a hash tag 226 effectively identifying the field of the hashed value prior to storage in an identity vault, or the hashed value output of the hashing algorithm 224 may be the hash tag 226 for direct storage in the identity vault. The second hashed values, in either event, are thereafter stored within the identity vault (such as the identity vault 120) (illustrated as 228).

[0039] Once stored in the identity vault, the second hashed values are matched and linked to previously stored second hashed values, when possible (illustrated as 230). Matching and linking of the data involves the use of unique IDs. Each unique ID is associated with data representative of a specific individual. The recently stored second hashed data either receives a newly generated unique ID (when the individual associated with the recently stored data is not represented by previously stored data within the identity vault) or it receives a unique ID previously generated (when the individual associated with the recently stored data is represented by previously stored data stored within the identity vault).

[0040] A method 300 for generating a unique ID and matching and linking hashed data values to the identity vault is described with reference to FIG. 3. The individual hashed values (e.g., contributor ID, SSN, first name, last name, address, DOB, etc.) are concatenated for each record/file uploaded into the identity vault (illustrated as 302). In an illustrative healthcare industry example, the individual hashed values may additionally include a Medicare ID. A processor determines whether an exact match exists between the recently uploaded/stored hashed contributor ID value and a previously stored hashed contributor ID value (illustrated as 304). The processor may also determine whether an exact match exists between the recently stored hashed SSN value and associated clear text DOB value, and previously stored hashed SSN values and their associated clear text DOB values (illustrated as 306). The processor may further determine whether an exact match exists between the recently stored hashed Medicare ID value and its associated clear text DOB value, and previously stored hashed Medicare ID values and their associated clear text DOB values (illustrated as 308). The processor may additionally determine whether an exact match exists between the recently stored hashed first name value, hashed last name value, and their associated clear text DOB value, state value, and zip code value, and previously stored hashed first name values, hashed last name values, and their associated clear text DOB values, state values, and zip code values (illustrated as 310). If any matches are determined to exist, the unique ID associated with the matching previously stored data is associated with the recently stored data (illustrated as 312). Linking of the data ensures that stored hashed personal information of an individual is associated within the identity vault. This provides a robust data set for research purposes. Alternatively, if no matches are determined to exist, a new, non-previously generated unique ID is generated and associated with the recently stored hashed personal information (illustrated as 314). If the recently stored hashed data is the first batch of data received from a specific data contributor computing environment 102, decision point 304 may be omitted because no matching hashed contributor ID for that contributor will exist in the identity vault.

[0041] If an exact match is determined between the recently stored hashed data value and previously stored data, the processor may compare all demographic values of the recently stored data (including both hashed and clear text values) to analogous values of the matching previously stored data (not illustrated). Each generated unique ID may be an alpha, numeric, or alphanumeric value.

[0042] Referring now to FIGS. 4A and 4B, a method 400 for rekeying/restating previously generated unique IDs is described. A unique ID associated with specific hashed data may need to be restated/altered when additional data is subsequently stored within the identity vault. Restating of unique IDs may involve a rekey process. For example, when subsequently stored hashed data includes an additional data element/field value not present in the previously stored data, the unique ID of the previously stored data may be altered/rekeyed and applied to both the previously and subsequently stored and related hashed data. According to an example, stored hashed data may be a bridge record that associates multiple hashed data files. If the bridge record gets deleted/corrupted (e.g., during maintenance of the secure facility computing environment 108), the unique ID associated with the bridge record may need to be restated for the bridge record and/or any associated records/data files.

[0043] Rekeying may involve determining whether an exact match exists between a recently uploaded/stored hashed contributor ID value and previously stored hashed contributor ID values (illustrated as 402). Rekeying may also include determining whether an exact match exists between a recently stored hashed SSN value and its associated clear text DOB value, and previously stored hashed SSN values and their associated clear text DOB values (illustrated as 404). Rekeying may further involve determining whether an exact match exists between a recently stored hashed Medicare ID value and its associated clear text DOB value, and previously stored hashed Medicare ID values and their associated

clear text DOB values (illustrated as 406). Additionally, rekeying may include determining whether an exact match exists between recently stored hashed first and last name values, and their associated clear text DOB value, state value, and zip code value, and previously stored hashed first and last name values, and their associated clear text DOB values, state values, and zip code values (illustrated as 408). While the described rekeying processes are identical to those used for matching/linking data described herein with respect to FIG. 3, one skilled in the art should appreciate that the logics/processes used in the rekeying and matching/linking processes may be different.

[0044] If it is determined that no matches exist, a new unique ID is generated for and associated with the recently stored hashed data (illustrated as 410). Conversely, if it is determined that a match exists, all data fields of the currently stored and matching previously stored data are compared (illustrated as 412). If there is an exact match across all of the data fields, the unique ID associated with the previously stored data remains unchanged, and is also associated with the recently stored data (illustrated as 414). Alternatively, if the hashed contributor ID values match and at least one demographic data field (e.g., DOB, state, zip code, etc.) of the recently stored data is not present in the previously stored data, a new unique ID is generated and associated with both the previously and recently stored data (illustrated as 416). When step 416 occurs, the previously stored data is no longer associated with its previous unique ID.

[0045] FIG. 5 illustrates how unique IDs and their associated concatenated hashed values may be represented within the identity vault. As stated above with respect to FIG. 1, once a unique ID is associated with the recently loaded hashed personal information, the same unique ID is also associated with the clear text data (e.g., within the de-identified claims and clinical storages 116, 118) to which it relates. FIG. 6 illustrates how related healthcare claims data and clinical record data may be associated across data storages via a unique ID. The data format illustrated in FIG. 6 may be how data is represented within a data dictionary.

[0046] Attention is now given to FIG. 7, which illustrates the system architecture of the VPN 132 illustrated in FIG. 1. A researcher, via a computing device 122, accesses certified views for which the researcher has been granted access. Each researcher may be granted access to a single certified view or multiple certified views. When a researcher properly authenticates a session (as described in detail below) the VPN 132 accesses certified views of the researcher from the certified view storage 130 of the data access zone 128 (illustrated in FIG. 1) and places them in a sandbox 702. The sandbox 702 provides a researcher with access to all of its authorized views during a single session. Each certified view may correspond with one virtual desktop interface (VDI). For example, a researcher that has access to two (2) certified views may be presented with two (2) VDIs upon accessing the virtual private network (VPN) 132. This separation of the certified views eliminates the possibility of a researcher with access to multiple certified view combining data from two or more certified views, thereby effectively limiting the risk of personal information re-identification.

[0047] Access by the researcher computer 122 to the certified view via the VPN 132 may be restricted by a two factor authentication. An illustrative two factor authentication includes (1) a username and password factor and (2) a phone factor. Each of the authentication factors may use secure protocols and may be stored in an encrypted manner. The username and password factor may involve the researcher inputting a username and password specific to a particular directory on their computing device 122. Once the username and password credentials are supplied, the researcher is prompted to perform the second (i.e., phone factor) authentication. This authentication may be performed a variety a ways. For example, the VPN 132 may cause the researcher's phone to receive a call. To authenticate, the researcher answers the phone call and provides a personal identification number (PIN). In another example, the VPN 132 causes the researcher's phone to receive a text message. To authenticate, the researcher responds to the text message using the PIN. In another example, the VPN 132 may cause an app on the researcher's phone to receive a notification. To authenticate, the researcher accesses the app, inputs the PIN, and selects "authenticate."

[0048] The two factor authentication information (i.e., username, password, and PIN) are communicated from the researcher computer 122 through a firewall 704 to an analytics server 706. For example, the analytics server 706 may be the NETEZZA MAKO® server produced by International Business Machines Corporation (IBM). To authenticate the researcher within the VPN 132, the analytics server 706 communicates with an authentication server 708. The authentication server may be a lightweight directory access protocol (LDAP) server.

[0049] Although aspects of the present disclosure are described with respect to examples in a health care data environment, it should be understood that various disclosed techniques can be used in numerous other fields of technology in which sensitive information needs to be transformed, secured, and provided to individuals on a limited and controlled access basis. Various applications of the disclosed techniques provide substantial improvements to the functioning of the computer apparatus and the technical environments in which the various applications are implemented.

Claims

1. A secure computing environment (108) for transforming, securing, and transmitting personal information, comprising:

   a processor (118); and

a memory device including instructions operable to be executed by the processor (118) to perform a set of actions, and configuring the processor (118) to:

receive sanitized, first hashed personal information values (220) and corresponding clear text data from data contributor computing environments (102) being in communication with the secure computing environment (108);
transform (206) the sanitized, first hashed personal information values (220) into second hashed personal information values (226);
link (210), using unique IDs, the second hashed personal information values (226) to further second hashed data values (226) previously stored in an identity vault (120, 228), wherein each of the unique IDs is associated with one of the previously stored second hashed data values;
generate a data dictionary in response to receiving research project parameters, the data dictionary including the clear text data from multiple data storages, wherein each of the data contributor computing environments includes one of the multiple data storages that house the clear text data, at least some of the clear text data being associated by the unique IDs;
apply compliance and statistical rules to the data dictionary clear text data;
store the data dictionary as a certified view when the data dictionary clear text data is determined compliant with the compliance and statistical rules; and
provide restricted access to the data dictionary clear text data via secure virtual desktop interfaces.

2. The secure computing environment (108) of claim 1, wherein the sanitized, first hashed personal information values (220) and corresponding clear text data are received from healthcare provider computing devices,
wherein the sanitized, first hashed personal information values (220) are hashed by the healthcare provider computing devices using a single common salt (214), and
wherein the second hashed personal information values (226) are generated using a private salt (222) inaccessible to third party devices.

3. The secure computing environment (108) of claim 1 or claim 2, wherein providing restricted access to the clear text data includes compiling certified views approved for a single researcher into a single storage and providing a virtual desktop interface for each certified view, data accessible through one virtual desktop interface being uncommunicable with data accessible through another virtual desktop interface.

4. The secure computing environment (108) of any one of claims 1 to 3, wherein the processor is further configured to generate a new unique ID for the second hashed personal information values (226) when the second hashed personal information values (226) do not link with the previously stored hashed data values (226).

5. The secure computing environment (108) of any one of claims 1 to 4, wherein the sanitized, first hashed personal information values (220) are represented as 64 character values.

6. The secure computing environment (108) of any one of claims 1 to 5, wherein the second hashed personal information values (226) associated with a single individual record are concatenated prior to being linked using the unique IDs.

7. A highly secure system (100) for transforming and securing personal information, comprising:

a data contributor computing environment (102) including a processor (106) configured to:

sanitize clear text personal information; and
transform, using a common salt (214) and a first hashing algorithm (218), the sanitized clear text personal information into first hashed values (220); and

a secure facility computing environment (108) of any one of the preceding claims,
wherein the processor (118) is further configured to:

store the second hashed personal information values (226) within the identity vault (120, 228), the identity vault (120, 228) including previously stored hashed personal information values (226) associated with unique IDs, each unique ID corresponding to a single individual;
associate in the multiple data storages (112, 114) each stored clear text non-personal information with a unique ID associated with a corresponding second hashed personal information value (226);

generate the certified view for a researcher, the certified view comprised of the data dictionary, the data dictionary including data from the multiple data storages (112, 114); and
securely control, using a virtual private network (132), access to the certified view by a researcher computer (112).

8. The system (100) of claim 7, wherein the clear text personal information is sanitized by at least one of changing the clear text into purely upper case or purely lower case text, removing punctuation marks from the clear text, removing spaces from the clear text, or removing line ends from the clear text.

9. The system (100) of claim 7 or claim 8, wherein the second hashed personal information values (226) are linked to previously stored hashed personal information values (226) by determining the satisfaction of at least one of the following:

(i) an exact match between hashed contributor ID values;
(ii) an exact match between hashed social security number values and clear text date of birth values;
(iii) an exact match between hashed Medicare ID values and clear text date of birth values; and
(iv) an exact match between hashed first name values, hashed last name values, clear text date of birth values, clear text state address values, and clear text zip code values.

10. The system (100) of any one of claims 7 to 9,
wherein the virtual private network (132) is configured to present access to multiple certified views to a single researcher computing device (122), each of the certified views being viewable through its own corresponding virtual desktop interface, the virtual private network (132) further configured to prevent the researcher computing device (122) from communicating data from one virtual desktop interface to another virtual desktop interface,
wherein the first hashing algorithm (218) and a second hashing algorithm (224) used to transform the sanitized, first hashed personal information values (220) into second hashed personal information values (226) are the same one-way hashing algorithm; and
wherein the virtual private network (132) requires a researcher perform a two factor authentication process in order to gain access to the certified view.

11. A method for transforming, securing, and transmitting personal information by a secure computing environment (108), comprising the steps of:

receiving sanitized, first hashed personal information values (220) and corresponding clear text data from data contributor computing environments (102) being in communication with the secure computing environment (108);
transforming (206) the sanitized, first hashed personal information values (220) into second hashed personal information values (226);
linking (210) the second hashed personal information values (226) to further second hashed data values previously stored in an identity vault (120, 228), linking of the values including the use of unique IDs, wherein each of the unique IDs is associated with one of the previously stored second hashed data values;
generating a data dictionary in response to receiving research project parameters, the data dictionary including the clear text data from multiple data storages, wherein each of the data contributor computing environments includes one of the multiple data storages that house the clear text data, at least some of the clear text data being associated by the unique IDs;
applying compliance and statistical rules to the data dictionary clear text data;
storing the data dictionary as a certified view when the data dictionary clear text data is determined compliant with the compliance and statistical rules; and
providing restricted access to the data dictionary clear text data via secure virtual desktop interfaces.

12. The method of claim 11, wherein the sanitized, first hashed personal information values (220) are generated using text manipulation processes, a common salt (214), and a one-way hashing algorithm, and
wherein the text manipulation processes, the common salt (214), and the one-way hashing algorithm are identically used by the various data contributor computing devices.

13. The method of claim 11 or claim 12,
wherein each of the virtual desktop interfaces provides access to one certified view.

14. The method of any one of claims 11 to 13, further comprising the step of:

updating the unique ID associated with particular second hashed personal information values (226) when subsequently linked second hashed personal information values (226) associated with the particular second hashed personal information values (226) includes at least one data field not present in the particular second hashed personal information values (226).

15. The method of any one of claims 11 to 14, wherein:

the sanitized, first hashed personal information values (220) are generated using a common salt (214) used by all of the data contributor computing devices; and
the second hashed personal information values (226) are generated using a private salt (222) inaccessible to the data contributor computing devices and researcher computing devices (122).

Patentansprüche

1. Sichere Computerumgebung (108) zum Umwandeln, Sichern und Übertragen persönlicher Informationen, umfassend:

einen Prozessor (118); und
eine Speichervorrichtung, welche Befehle aufweist, welche geeignet sind, um von dem Prozessor (118) ausgeführt zu werden, um eine Gruppe von Aktionen auszuführen, und welche den Prozessor (118) konfigurieren, um:

bereinigte erste Hash-basierte persönliche Informationswerte (220) und entsprechende Klartextdaten von Datenbeitragscomputerumgebungen (102) zu empfangen, welche mit der sicheren Computerumgebung (108) kommunizieren;
die bereinigten ersten Hash-basierten persönlichen Informationswerte (220) in zweite Hash-basierte persönliche Informationswerte (226) umzuwandeln (206);
die zweiten Hash-basierten persönlichen Informationswerte (226) unter Verwendung eindeutiger IDs mit weiteren zweiten Hash-basierten Datenwerten (226) zu verbinden (210), welche zuvor in einem Identitätsdatenspeicher (120, 228) gespeichert wurden, wobei jede der eindeutigen IDs einem der vorher gespeicherten zweiten Hash-basierten Datenwerte zugeordnet ist;
ein Datenverzeichnis als Reaktion auf ein Empfangen von Rechercheprojektparameter zu erzeugen, wobei das Datenverzeichnis die Klartextdaten aus mehreren Datenspeichern aufweist, wobei jede der Datenbeitragscomputerumgebungen einen der mehreren Datenspeicher aufweist, welche die Klartextdaten beherbergen; wobei zumindest einige der Klartextdaten durch die eindeutigen IDs zugeordnet sind;
Konformitäts- und Statistikregeln auf die Datenverzeichnisklartextdaten anzuwenden;
das Datenverzeichnis als eine zertifizierte Ansicht zu speichern, wenn festgestellt wird, dass die Datenverzeichnisklartextdaten mit den Konformitäts- und Statistikregeln übereinstimmen; und
einen eingeschränkten Zugriff auf die Datenverzeichnisklartextdaten über sichere virtuelle Desktop-Schnittstellen bereitzustellen.

2. Sichere Computerumgebung (108) nach Anspruch 1, wobei die bereinigten ersten Hash-basierten persönlichen Informationswerte (220) und entsprechenden Klartextdaten von Gesundheitsanbieter-Computervorrichtungen empfangen werden,
wobei die bereinigten ersten Hash-basierten persönlichen Informationswerte (220) von den Gesundheitsanbieter-Computervorrichtungen unter Verwendung eines einzelnen gemeinsamen Salt (214) gehashed werden, und
wobei die zweiten Hash-basierten persönlichen Informationswerte (226) unter Verwendung eines privaten Salt (222), welcher für Vorrichtungen einer dritten Partei nicht zugreifbar sind, erzeugt werden.

3. Sichere Computerumgebung (108) nach Anspruch 1 oder Anspruch 2, wobei das Bereitstellen des eingeschränkten Zugriffs auf die Klartextdaten ein Zusammenstellen von zertifizierten Ansichten, welche für einen einzelnen Rechercheur zugelassen sind, in einem einzelnen Speicher und ein Bereitstellen einer virtuellen Desktop-Schnittstelle für jede zertifizierte Ansicht aufweist, wobei Daten, welche über eine virtuelle Desktop-Schnittstelle zugreifbar sind, nicht mit Daten in Verbindung setzbar sind, welche über eine andere virtuelle Desktop-Schnittstelle zugreifbar sind.

4. Sichere Computerumgebung (108) nach einem der Ansprüche 1 bis 3, wobei der Prozessor ferner ausgestaltet ist, eine neue eindeutige ID für die zweiten Hash-basierten persönlichen Informationswerte (226) zu erzeugen, wenn

die zweiten Hash-basierten persönlichen Informationswerte (226) nicht mit den zuvor gespeicherten Hash-basierten Datenwerten (226) verbunden sind.

5. Sichere Computerumgebung (108) nach einem der Ansprüche 1 bis 4, wobei die bereinigten ersten Hash-basierten persönlichen Informationswerte (220) als 64 Zeichenwerte dargestellt werden.

6. Sichere Computerumgebung (108) nach einem der Ansprüche 1 bis 5, wobei die zweiten Hash-basierten persönlichen Informationswerte (226), welche einem einzelnen individuellen Datensatz zugeordnet sind, verknüpft werden bevor sie unter Verwendung der eindeutigen IDs verbunden werden.

7. Hochsicheres System (100) zum Umwandeln und Sichern persönlicher Informationen, umfassend:

eine Datenbeitragscomputerumgebung (102), welche einen Prozessor (106) aufweist, welcher ausgestaltet ist,

persönliche Klartextinformationen zu bereinigen; und
die bereinigten persönlichen Klartextinformationen unter Verwendung eines gemeinsamen Salt (214) und eines ersten Hash-Algorithmus (218) in erste Hash-basierte Werte (220) umzuwandeln; und

eine sichere Anlagen-Computerumgebung (108) nach einem der vorhergehenden Ansprüche, wobei der Prozessor (118) ferner ausgestaltet ist,

die zweiten Hash-basierten persönlichen Informationswerte (226) innerhalb des Identitätsdatenspeichers (120, 228) zu speichern, wobei der Identitätsdatenspeicher (120, 228) zuvor gespeicherte Hash-basierte persönliche Informationswerte (226), welchen eindeutige IDs zugeordnet sind, aufweist, wobei jede eindeutige ID einem einzelnen Individuum entspricht;
in den mehreren Datenspeichern (112, 114) jeder gespeicherten nicht-persönlichen Klartextinformation eine eindeutige ID zuzuordnen, welche einem entsprechenden zweiten Hash-basierten persönlichen Informationswert (226) zugeordnet ist;
die zertifizierte Ansicht für einen Rechercheur zu erzeugen, wobei die zertifizierte Ansicht aus dem Datenverzeichnis gebildet wird, wobei das Datenverzeichnis Daten von den mehreren Datenspeichern (112, 114) aufweist; und
einen Zugriff auf die zertifizierte Ansicht durch den Recherchecomputer (112) unter Verwendung eines virtuellen privaten Netzes (132) sicher zu steuern.

8. System (100) nach Anspruch 7, wobei die persönliche Klartextinformation bereinigt wird durch Ändern des Klartexts in reinen Großbuchstaben- oder reinen Kleinbuchstabentext, Entfernen von Satzzeichen aus dem Klartext, Entfernen von Leerzeichen aus dem Klartext und/oder Entfernen von Zeilenenden aus dem Klartext.

9. System (100) nach Anspruch 7 oder Anspruch 8, wobei die zweiten Hash-basierten persönlichen Informationswerte (226) mit zuvor gespeicherten Hash-basierten persönlichen Informationswerten (226) verbunden werden durch Bestimmen der Erfüllung von mindestens einer der Folgenden:

(i) eine genaue Übereinstimmung zwischen Hash-basierten Beitrags-ID-Werten;
(ii) eine genaue Übereinstimmung zwischen Hash-basierten Sozialversicherungsnummerwerten und Klartextgeburtsdatumswerten;
(iii) eine genaue Übereinstimmung zwischen Hash-basierten Krankenversicherung-ID-Werten und Klartextgeburtsdatumswerten;
(iv) eine genaue Übereinstimmung zwischen Hash-basierten Vornamenwerten, Hash-basierten Nachnamenwerten, Klartextgeburtsdatumswerten, Klartextadresswerten und Klartextpostleitzahlwerten.

10. System (100) nach einem der Ansprüche 7 bis 9,
wobei das virtuelle private Netz (132) ausgestaltet ist, einen Zugriff auf mehrere zertifizierte Ansichten für eine einzelne Recherchecomputervorrichtung (122) zu bieten, wobei jede der zertifizierten Ansichten über seine eigene entsprechende virtuelle Desktop-Schnittstelle sichtbar ist, wobei das virtuelle private Netz (132) ferner ausgestaltet ist, zu verhindern, dass die Recherchecomputervorrichtung (122) Daten von einer virtuellen Desktop-Schnittstelle zu einer anderen virtuellen Desktop-Schnittstelle kommuniziert,
wobei der erste Hash-Algorithmus (218) und ein zweiter Hash-Algorithmus (224), welcher verwendet wird, um die bereinigten ersten Hash-basierten persönlichen Informationswerte (220) in zweite Hash-basierte persönliche Infor-

mationswerte (226) umzuwandeln, der gleiche Einweg-Hash-Algorithmus sind; und
wobei das virtuelle private Netz (132) erfordert, dass ein Rechercheur ein Zwei-Faktor-Authentifizierungsverfahren ausführt, um einen Zugriff auf die zertifizierte Ansicht zu erlangen.

11. Verfahren zum Umwandeln, Sichern und Übertragen persönlicher Informationen mittels einer sicheren Computerumgebung (108), umfassend die Schritte:

Empfangen von bereinigten ersten Hash-basierten persönlichen Informationswerten (220) und entsprechenden Klartextdaten von Datenbeitragscomputerumgebungen (102), welche mit der sicheren Computerumgebung (108) kommunizieren;
Umwandeln (206) der bereinigten ersten Hash-basierten persönlichen Informationswerte (220) in zweite Hash-basierte persönliche Informationswerte (226);
Verbinden (210) der zweiten Hash-basierten persönlichen Informationswerte (226) mit weiteren zweiten Hash-basierten Datenwerten, welche zuvor in einem Identitätsdatenspeicher (120, 228) gespeichert wurden, wobei das Verbinden der Werte die Verwendung von eindeutigen IDs aufweist, wobei jede der eindeutigen IDs einem der vorher gespeicherten zweiten Hash-basierten Datenwerte zugeordnet ist;
Erzeugen eines Datenverzeichnisses als Reaktion auf ein Empfangen von Rechercheprojektparametern, wobei das Datenverzeichnis die Klartextdaten aus mehreren Datenspeichern aufweist, wobei jede der Datenbeitragscomputerumgebungen einen der mehreren Datenspeicher aufweist, welche die Klartextdaten beherbergen; wobei zumindest einige der Klartextdaten durch die eindeutigen IDs zugeordnet werden;
Anwenden von Konformitäts- und Statistikregeln auf die Datenverzeichnisklartextdaten ;
Speichern des Datenverzeichnisses als zertifizierte Ansicht, wenn festgestellt wird, dass die Datenverzeichnisklartextdaten mit den Konformitäts- und Statistikregeln übereinstimmen; und
Bereitstellen eines eingeschränkten Zugriffs auf die Datenverzeichnisklartextdaten über sichere virtuelle Desktop-Schnittstellen.

12. Verfahren nach Anspruch 11, wobei die bereinigten ersten Hash-basierten persönlichen Informationswerte (220) unter Verwendung eines Textmanipulationsverfahrens, einem gemeinsamen Salt (214) und eines Einweg-Hash-Algorithmus erzeugt werden, und
wobei das Textmanipulationsverfahren, der gemeinsame Salt (214) und der Einweg-Hash-Algorithmus von den verschiedenen Datenbeitragscomputervorrichtungen identisch verwendet werden.

13. Verfahren nach Anspruch 11 oder Anspruch 12,
wobei jede der virtuellen Desktop-Schnittstellen einen Zugriff auf eine zertifizierte Ansicht bereitstellt.

14. Verfahren nach einem der Ansprüche 11 bis 13, ferner umfassend die Schritte:
Aktualisieren der eindeutigen ID, welche speziellen zweiten Hash-basierten persönlichen Informationswerten (226) zugeordnet ist, wenn nachfolgend verbundene zweite Hash-basierte persönliche Informationswerte (226), welche den speziellen zweiten Hash-basierten persönlichen Informationswerten (226) zugeordnet sind, zumindest ein Datenfeld aufweisen, welches in den speziellen zweiten Hash-basierten persönlichen Informationswerten (226) nicht vorhanden ist.

15. Verfahren nach einem der Ansprüche 11 bis 14, wobei:

die bereinigten ersten Hash-basierten persönlichen Informationswerte (220) unter Verwendung eines gemeinsamen Salt (214) erzeugt werden, welcher von allen der Datenbeitragscomputervorrichtungen verwendet wird; und
die zweiten Hash-basierten persönlichen Informationswerte (226) unter Verwendung eines privaten Salt (222) erzeugt werden, welcher für die Datenbeitragscomputervorrichtungen und Recherchecomputervorrichtungen (122) nicht zugreifbar ist.

Revendications

1. Environnement informatique sécurisé (108) pour transformer, sécuriser, et transmettre des informations personnelles, comprenant :

un processeur (118) ; et

un dispositif de mémoire incluant des instructions opérationnelles pour être exécutées par le processeur (118) pour effectuer un ensemble d'actions, et configurant le processeur (118) pour :

recevoir des premières valeurs d'informations personnelles hachées et nettoyées (220) et des données en texte clair correspondantes provenant d'environnements informatiques contributeurs de données (102) qui est en communication avec l'environnement informatique sécurisé (108) ;
transformer (206) les premières valeurs d'informations personnelles hachées et nettoyées (220) en secondes valeurs d'informations personnelles hachées (226) ;
lier (210), en utilisant des ID uniques, les secondes valeurs d'informations personnelles hachées (226) à d'autres secondes valeurs de données hachées (226) précédemment stockées dans un coffre-fort d'identités (120, 228), dans lequel chacun des ID uniques est associé à l'une des secondes valeurs de données hachées précédemment stockées ;
générer un dictionnaire de données en réponse à la réception de paramètres de projet de recherche, le dictionnaire de données incluant les données en texte clair provenant de multiples stockages de données, dans lequel chacun des environnements informatiques contributeurs de données inclut l'un des multiples stockages de données qui referment les données en texte clair, au moins certaines des données en texte clair étant associées par les ID uniques ;
appliquer des règles de conformité et statistiques aux données en texte clair du dictionnaire de données ;
stocker le dictionnaire de données en tant que vue certifiée lorsque les données en texte clair du dictionnaire de données sont déterminées conformes aux règles de conformité et statistiques ; et
fournir un accès restreint aux données en texte clair du dictionnaire de données via des interfaces de bureau virtuelles sécurisées.

2. Environnement informatique sécurisé (108) selon la revendication 1, dans lequel les premières valeurs d'informations personnelles hachées et nettoyées (220) et les données en texte clair correspondantes sont reçues en provenance de dispositifs informatiques de prestataires de soins,
dans lequel les premières valeurs d'informations personnelles hachées et nettoyées (220) sont hachées par les dispositifs informatiques de prestataires de soins en utilisant un seul sel commun (214), et
dans lequel les secondes valeurs d'informations personnelles hachées (226) sont générées en utilisant un sel privé (222) inaccessible à de tiers dispositifs.

3. Environnement informatique sécurisé (108) selon la revendication 1 ou la revendication 2, dans lequel la fourniture d'un accès restreint aux données en texte clair inclut la compilation de vues certifiées approuvées pour un chercheur unique dans un stockage unique et la fourniture d'une interface de bureau virtuelle pour chaque vue certifiée, des données accessibles via une interface de bureau virtuelle ne pouvant pas communiquer avec des données accessibles via une autre interface de bureau virtuelle.

4. Environnement informatique sécurisé (108) selon l'une quelconque des revendications 1 à 3, dans lequel le processeur est en outre configuré pour générer un nouvel ID unique pour les secondes valeurs d'informations personnelles hachées (226) lorsque les secondes valeurs d'informations personnelles hachées (226) ne sont pas liées avec les valeurs de données hachées précédemment stockées (226).

5. Environnement informatique sécurisé (108) selon l'une quelconque des revendications 1 à 4, dans lequel les premières valeurs d'informations personnelles hachées et nettoyées (220) sont représentées sous la forme de valeurs à 64 caractères.

6. Environnement informatique sécurisé (108) selon l'une quelconque des revendications 1 à 5, dans lequel les secondes valeurs d'informations personnelles hachées (226) associées à un enregistrement individuel unique sont concaténées avant d'être liées en utilisant les ID uniques.

7. Système hautement sécurisé (100) pour transformer et sécuriser des informations personnelles, comprenant :

un environnement informatique contributeur de données (102) incluant un processeur (106) configuré pour :

nettoyer des informations personnelles en texte clair ; et
transformer, en utilisant un sel commun (214) et un premier algorithme de hachage (218), les informations personnelles en texte clair nettoyées en premières valeurs hachées (220) ; et

un environnement informatique de service sécurisé (108) selon l'une quelconque des revendications précédentes, dans lequel le processeur (118) est en outre configuré pour :

stocker les secondes valeurs d'informations personnelles hachées (226) dans le coffre-fort d'identités (120, 228), le coffre-fort d'identités (120, 228) incluant des valeurs d'informations personnelles hachées précédemment stockées (226) associées à des ID uniques, chaque ID unique correspondant à un individu unique ;
associer dans les multiples stockages de données (112, 114) chacune des informations non personnelles en texte clair stockées avec un ID unique associé à une seconde valeur d'informations personnelles hachées correspondante (226) ;
générer la vue certifiée pour un chercheur, la vue certifiée étant composée du dictionnaire de données, le dictionnaire de données incluant des données provenant des multiples stockages de données (112, 114) ; et
commander de manière sécurisée, en utilisant un réseau privé virtuel (132), l'accès à la vue certifiée par un ordinateur de chercheur (112).

8. Système (100) selon la revendication 7, dans lequel les informations personnelles en texte clair sont nettoyées par au moins l'un parmi un changement du texte clair en texte purement en majuscule ou purement en minuscule, une suppression de signes de ponctuation du texte clair, une suppression d'espaces du texte clair, ou une suppression de fins de ligne du texte clair.

9. Système (100) selon la revendication 7 ou la revendication 8, dans lequel les secondes valeurs d'informations personnelles hachées (226) sont liées à des valeurs d'informations personnelles hachées précédemment stockées (226) en déterminant si au moins l'une des correspondances suivantes est satisfaite :

(i) une correspondance exacte entre des valeurs ID de contributeur hachées ;
(ii) une correspondance exacte entre des valeurs de numéro de sécurité sociale hachées et des valeurs de date de naissance en texte clair ;
(iii) une correspondance exacte entre des valeurs ID d'assurance-maladie (Medicare) hachées et des valeurs de date de naissance en texte clair ; et
(iv) une correspondance exacte entre des valeurs de prénom hachées, des valeurs de nom de famille hachées, des valeurs de date de naissance en texte clair, des valeurs d'adresse indiquée en texte clair, et des valeurs de code postal en texte clair.

10. Système (100) selon l'une quelconque des revendications 7 à 9,
dans lequel le réseau privé virtuel (132) est configuré pour présenter un accès à de multiples vues certifiées à un dispositif informatique de chercheur unique (122), chacune des vues certifiées étant visible via sa propre interface de bureau virtuelle correspondante, le réseau privé virtuel (132) étant en outre configuré pour empêcher que le dispositif informatique de chercheur (122) communique des données d'une interface de bureau virtuelle à une autre interface de bureau virtuelle,
dans lequel le premier algorithme de hachage (218) et un second algorithme de hachage (224) utilisés pour transformer les premières valeurs d'informations personnelles hachées et nettoyées (220) en secondes valeurs d'informations personnelles hachées (226) sont les mêmes algorithmes de hachage unidirectionnel ; et
dans lequel le réseau privé virtuel (132) nécessite qu'un chercheur effectue un processus d'authentification à deux facteurs afin de donner un accès à la vue certifiée.

11. Procédé pour transformer, sécuriser, et transmettre des informations personnelles par un environnement informatique sécurisé (108), comprenant les étapes consistant à :

recevoir des premières valeurs d'informations personnelles hachées et nettoyées (220) et des données en texte clair correspondantes provenant d'environnements informatiques contributeurs de données (102) qui est en communication avec l'environnement informatique sécurisé (108) ;
transformer (206) les premières valeurs d'informations personnelles hachées et nettoyées (220) en secondes valeurs d'informations personnelles hachées (226) ;
lier (210) les secondes valeurs d'informations personnelles hachées (226) à d'autres secondes valeurs de données hachées précédemment stockées dans un coffre-fort d'identités (120, 228), la liaison des valeurs incluant l'utilisation d'ID uniques, dans lequel chacun des ID uniques est associé à l'une des secondes valeurs de données hachées précédemment stockées ;
générer un dictionnaire de données en réponse à la réception de paramètres de projet de recherche, le dictionnaire de données incluant les données en texte clair provenant de multiples stockages de données, dans

lequel chacun des environnements informatiques contributeurs de données inclut l'un des multiples stockages de données qui referment les données en texte clair, au moins certaines des données en texte clair étant associées par les ID uniques ;
appliquer des règles de conformité et statistiques aux données en texte clair du dictionnaire de données ;
stocker le dictionnaire de données en tant que vue certifiée lorsque les données en texte clair du dictionnaire de données sont déterminées conformes aux règles de conformité et statistiques ; et
fournir un accès restreint aux données en texte clair du dictionnaire de données via des interfaces de bureau virtuelles sécurisées.

12. Procédé selon la revendication 11, dans lequel les premières valeurs d'informations personnelles hachées et nettoyées (220) sont générées en utilisant des processus de manipulation de texte, un sel commun (214), et un algorithme de hachage unidirectionnel, et
dans lequel les processus de manipulation de texte, le sel commun (214), et l'algorithme de hachage unidirectionnel sont utilisés de manière identique par les divers dispositifs informatiques contributeurs de données.

13. Procédé selon la revendication 11 ou la revendication 12,
dans lequel chacune des interfaces de bureau virtuelles donne accès à une vue certifiée.

14. Procédé selon l'une quelconque des revendications 11 à 13, comprenant en outre l'étape consistant à :
mettre à jour l'ID unique associé à des secondes valeurs d'informations personnelles hachées particulières (226) lorsque des secondes valeurs d'informations personnelles hachées liées par la suite (226) associées aux secondes valeurs d'informations personnelles hachées particulières (226) incluent au moins un champ de données non présent dans les secondes valeurs d'informations personnelles hachées particulières (226).

15. Procédé selon l'une quelconque des revendications 11 à 14, dans lequel :

les premières valeurs d'informations personnelles hachées et nettoyées (220) sont générées en utilisant un sel commun (214) utilisé par tous les dispositifs informatiques contributeurs de données ; et
les secondes valeurs d'informations personnelles hachées (226) sont générées en utilisant un sel privé (222) inaccessible aux dispositifs informatiques contributeurs de données et aux dispositifs informatiques de chercheur (122).

FIG. 1

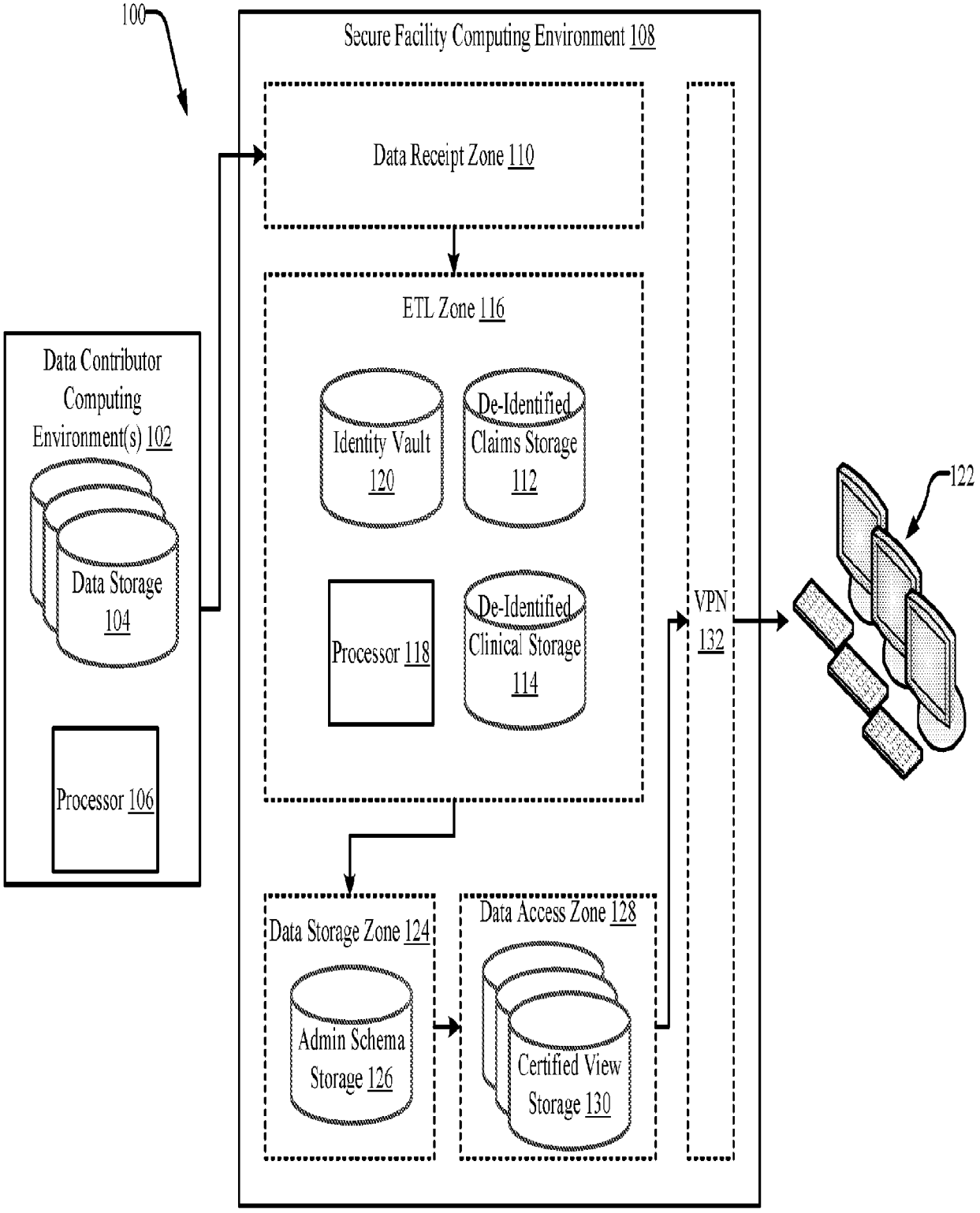

100
Secure Facility Computing Environment 108
Data Receipt Zone 110
ETL Zone 116
Identity Vault 120
De-Identified Claims Storage 112
Processor 118
De-Identified Clinical Storage 114
Data Contributor Computing Environment(s) 102
Data Storage 104
Processor 106
Data Storage Zone 124
Admin Schema Storage 126
Data Access Zone 128
Certified View Storage 130
VPN 132
122

FIG. 2A

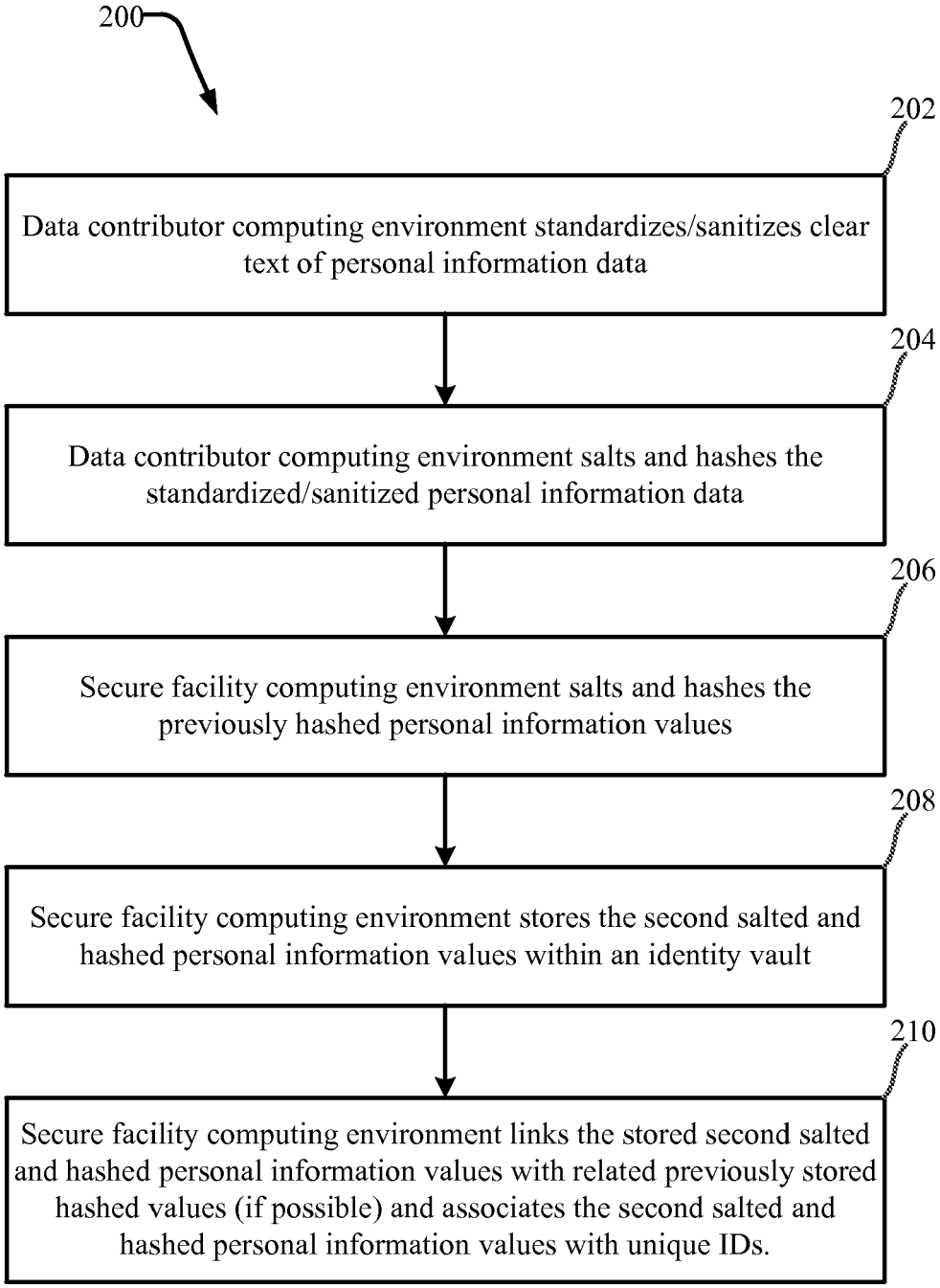

200
202
Data contributor computing environment standardizes/sanitizes clear text of personal information data
204
Data contributor computing environment salts and hashes the standardized/sanitized personal information data
206
Secure facility computing environment salts and hashes the previously hashed personal information values
208
Secure facility computing environment stores the second salted and hashed personal information values within an identity vault
210
Secure facility computing environment links the stored second salted and hashed personal information values with related previously stored hashed values (if possible) and associates the second salted and hashed personal information values with unique IDs.

FIG. 2B

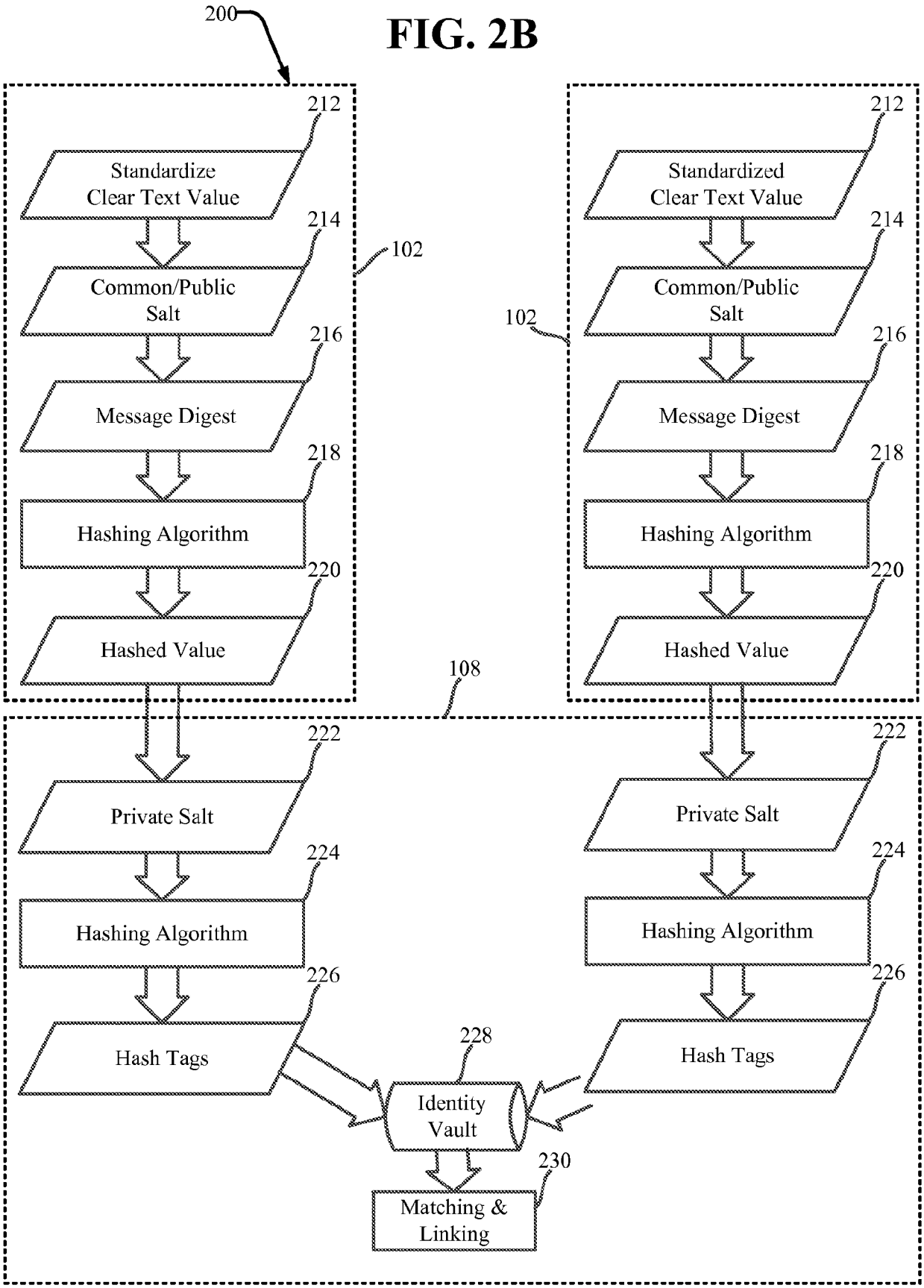

200
212
Standardize Clear Text Value
214
Common/Public Salt
216
Message Digest
218
Hashing Algorithm
220
Hashed Value
102
212
Standardized Clear Text Value
214
Common/Public Salt
216
Message Digest
218
Hashing Algorithm
220
Hashed Value
102
108
222
Private Salt
224
Hashing Algorithm
226
Hash Tags
222
Private Salt
224
Hashing Algorithm
226
Hash Tags
228
Identity Vault
230
Matching & Linking

FIG. 3

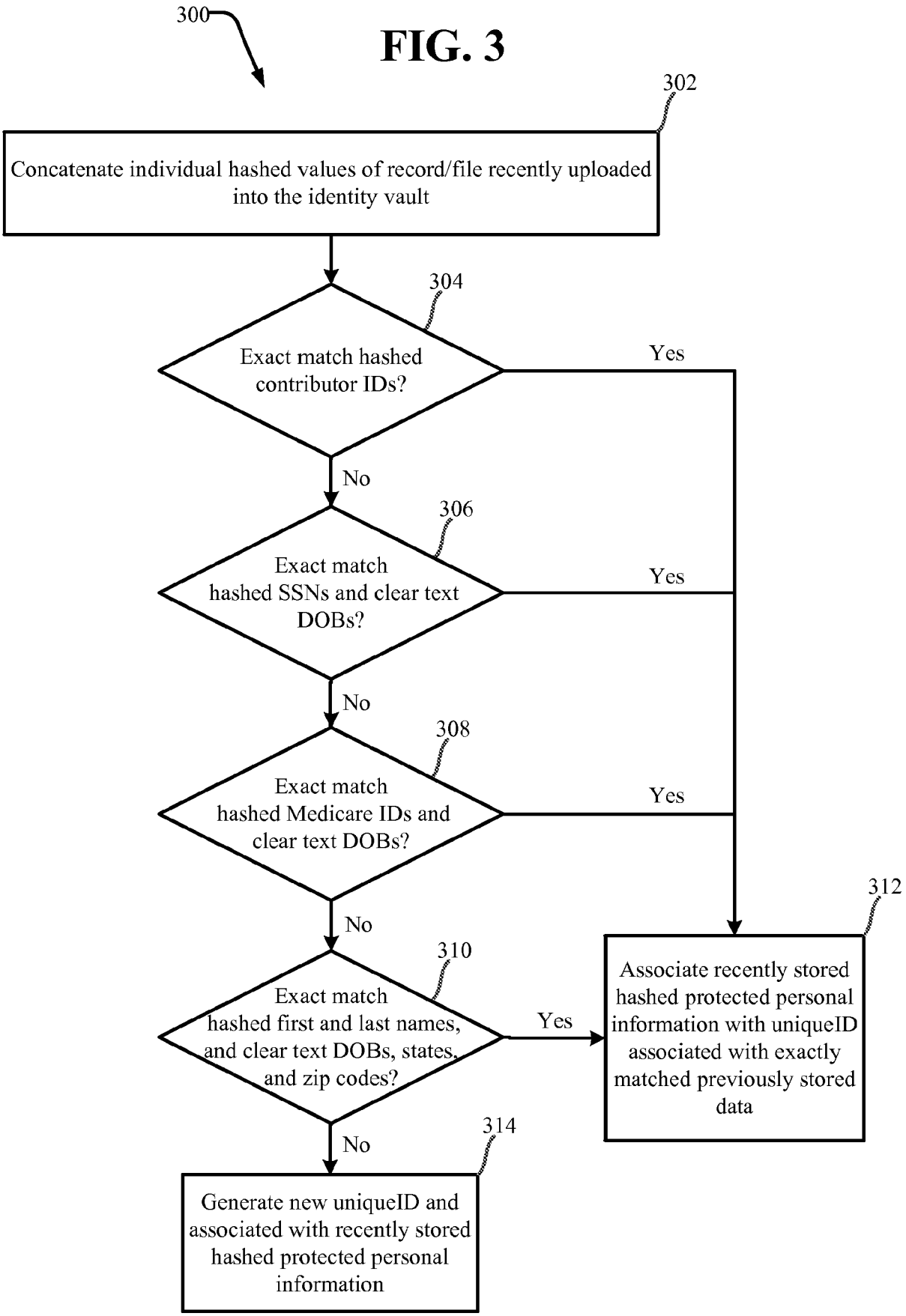

300
302
Concatenate individual hashed values of record/file recently uploaded into the identity vault
304
Exact match hashed contributor IDs?
Yes
No
306
Exact match hashed SSNs and clear text DOBs?
Yes
No
308
Exact match hashed Medicare IDs and clear text DOBs?
Yes
No
310
Exact match hashed first and last names, and clear text DOBs, states, and zip codes?
Yes
No
312
Associate recently stored hashed protected personal information with uniqueID associated with exactly matched previously stored data
314
Generate new uniqueID and associated with recently stored hashed protected personal information

FIG. 4A

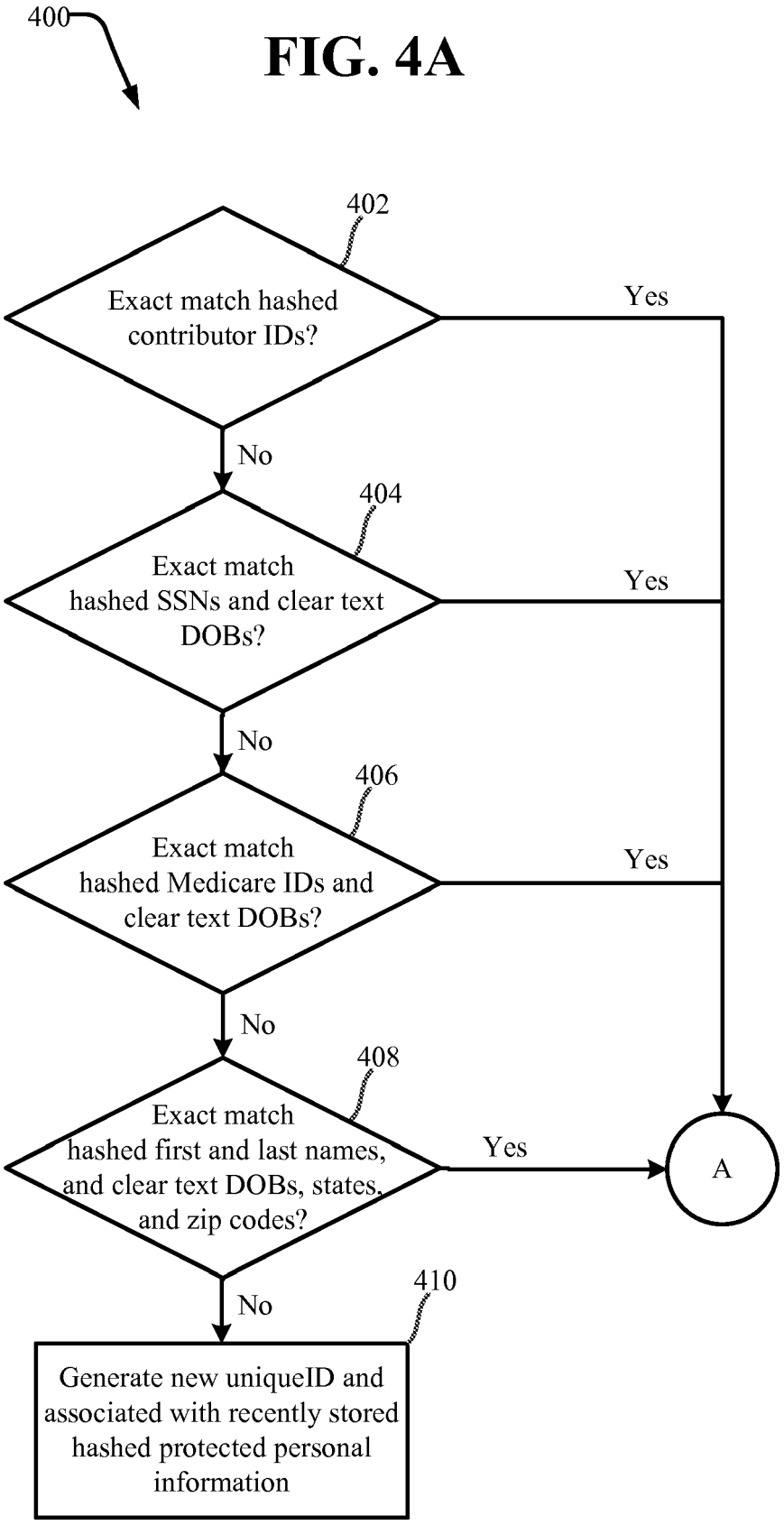

400
402
Exact match hashed contributor IDs?
Yes
No
404
Exact match hashed SSNs and clear text DOBs?
Yes
No
406
Exact match hashed Medicare IDs and clear text DOBs?
Yes
No
408
Exact match hashed first and last names, and clear text DOBs, states, and zip codes?
Yes
A
No
410
Generate new uniqueID and associated with recently stored hashed protected personal information

FIG. 4B

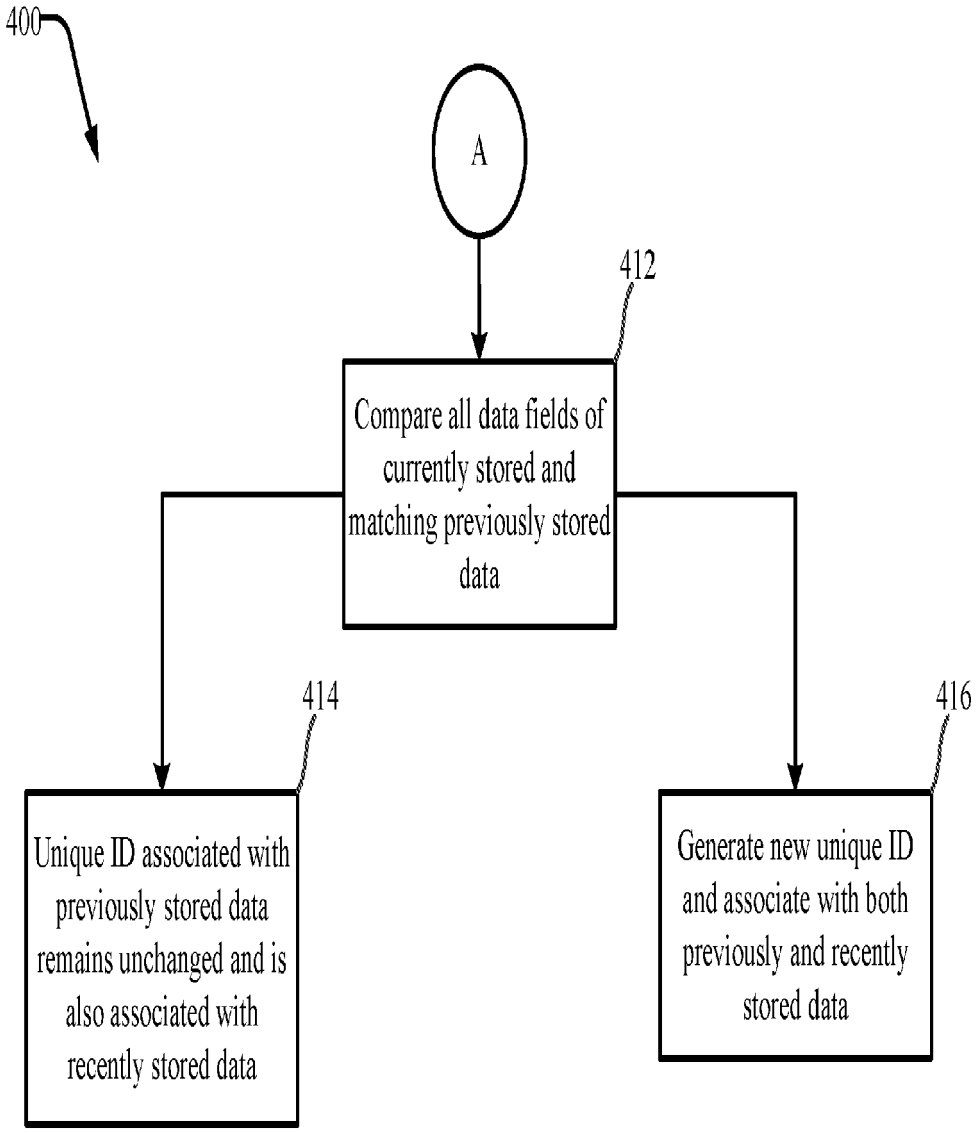

400
A
412
Compare all data fields of currently stored and matching previously stored data
414
Unique ID associated with previously stored data remains unchanged and is also associated with recently stored data
416
Generate new unique ID and associate with both previously and recently stored data

FIG. 5

UNIQUE_ID	FIRST_NAME_E|LAST_NAME_E|SSN_E
94425100	1238ca657102ae567mc8845167957a78a1cd321342ca09d82a28am53a7e5830fd89|123m812ae83f17a8488bbm9f5f0d2m74c46m94e2f42a6bd76d96614f71a6fd1efac|123fcee4cf166e0f1d0c110338942f6a8cb1049e78c594ef49952aed348cdda3e63
94425101	123bff7m485135c547322b677792b39m0d8b08bcd8896m9m3699a3bmb6e7b6e205a|1231dd4m4e87f2486ee1e639a38b3e5942d68a9118b107354abacb9me34a7509ee072b9234b38a2f3a1a9d537824cb|1237546m08f4857e29c90f3ffc3e3em0641d6456
94425102	1232e6c27e28c72a14a20c07e138d45d7fca52333f673f560fc4678c734db8ddaf1|123c2edcd9b53e3b6966efb2b201f52441dd677d2987bcca3d0635577e882cb6m1d9|123b5e9329e72f942c27ea570a1946de34b10463f55ef366c89556a2a5bf372e634

FIG. 6

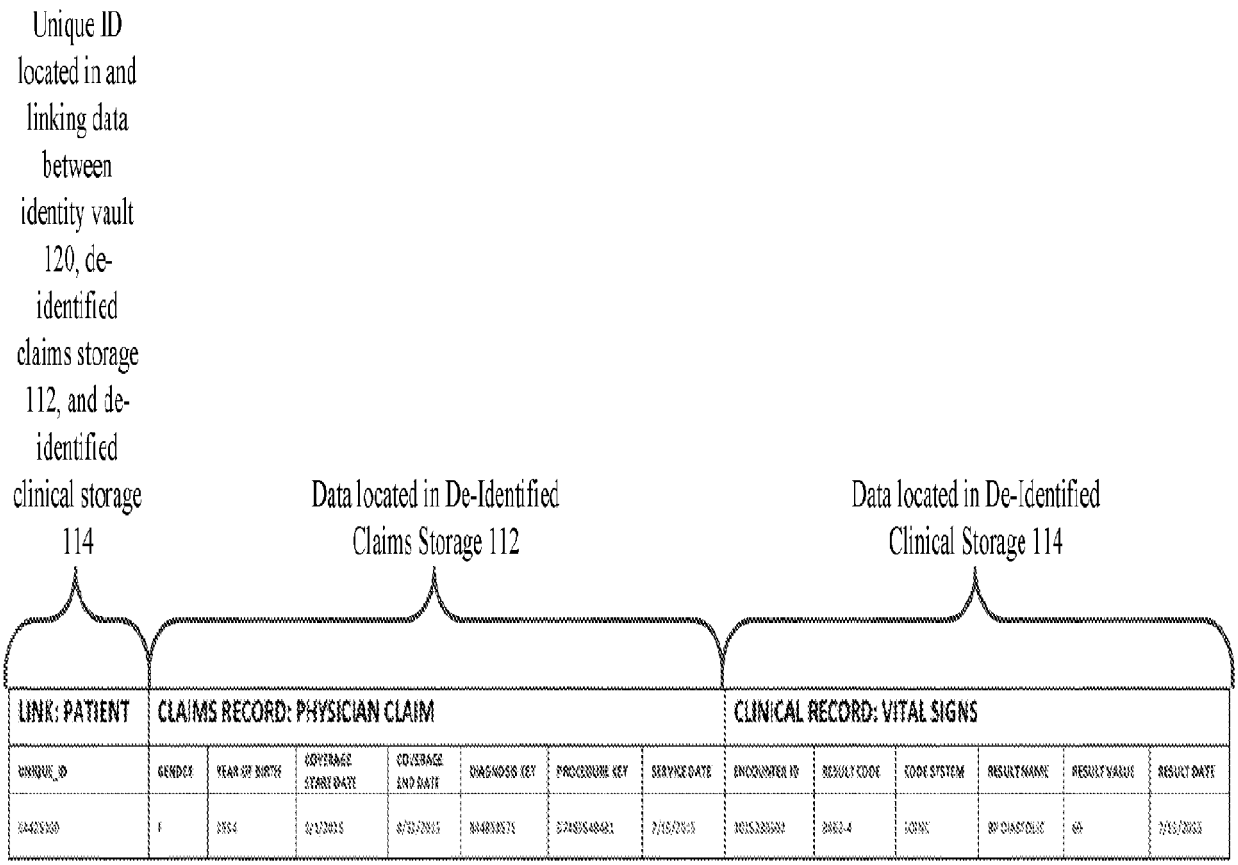
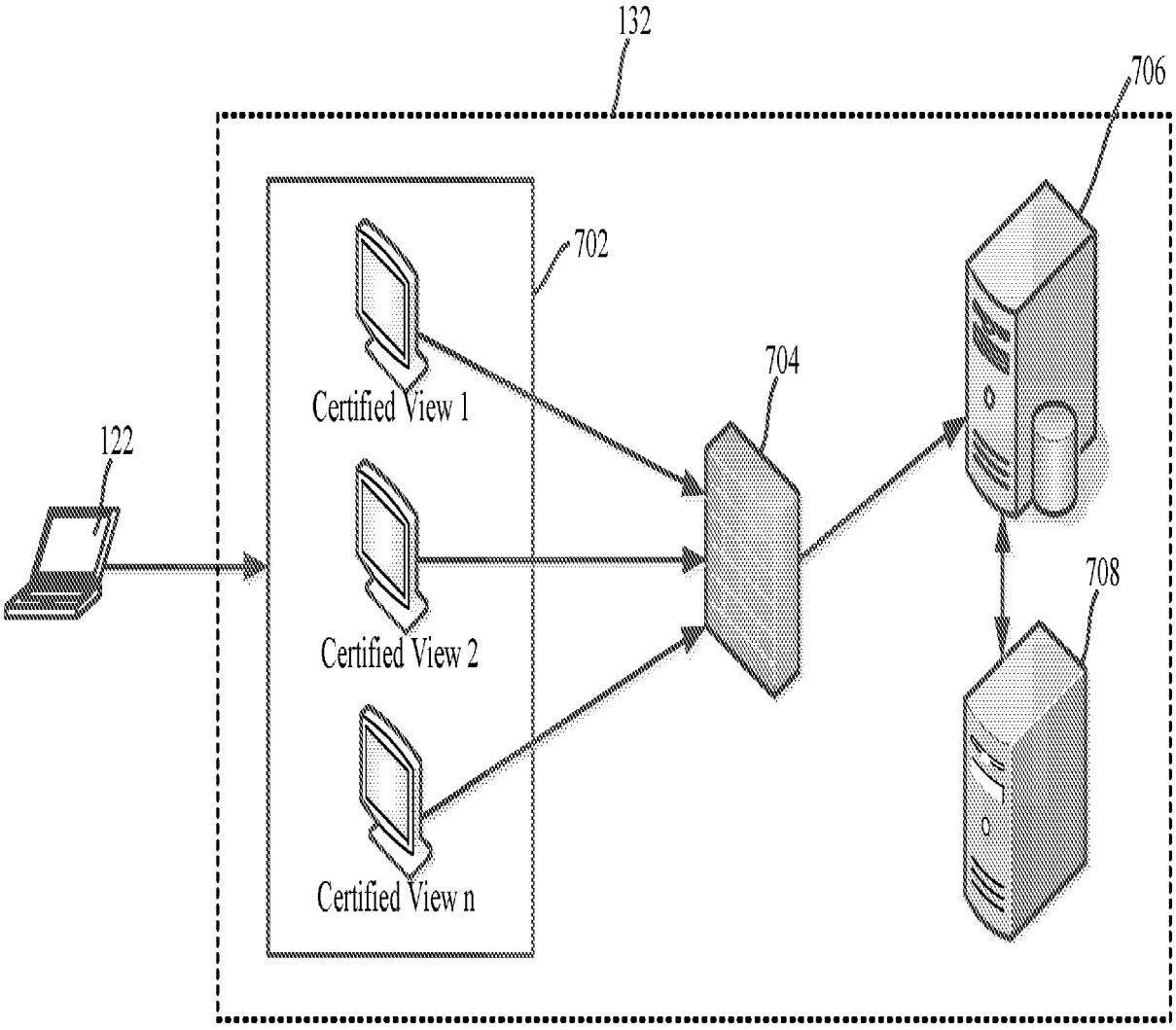

Unique ID located in and linking data between identity vault 120, de-identified claims storage 112, and de-identified clinical storage 114
Data located in De-Identified Claims Storage 112
Data located in De-Identified Clinical Storage 114
LINK: PATIENT
CLAIMS RECORD: PHYSICIAN CLAIM
CLINICAL RECORD: VITAL SIGNS
UNIQUE_ID
GENDER
YEAR OF BIRTH
COVERAGE START DATE
COVERAGE END DATE
DIAGNOSIS KEY
PROCEDURE KEY
SERVICE DATE
ENCOUNTER ID
RESULT CODE
CODE SYSTEM
RESULT NAME
RESULT VALUE
RESULT DATE

FIG. 7


132
706
702
704
122
Certified View 1
Certified View 2
Certified View n
708

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- US 2002073138 A1 **[0004]**
- US 7870614 B1 **[0005]**
- US 2005268094 A1 **[0006]**