(19) Евразийское патентное ведомство

(21) 202192295 (13) A1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОЙ ЗАЯВКЕ

(43) Дата публикации заявки
2022.12.26

(22) Дата подачи заявки
2021.09.16

(51) Int. Cl. ***G06F 7/06*** (2006.01)
G06F 16/438 (2019.01)
G06F 21/60 (2013.01)

(54) СПОСОБ И СИСТЕМА ОПРЕДЕЛЕНИЯ НАЛИЧИЯ КРИТИЧЕСКИХ КОРПОРАТИВНЫХ ДАННЫХ В ТЕСТОВОЙ БАЗЕ ДАННЫХ

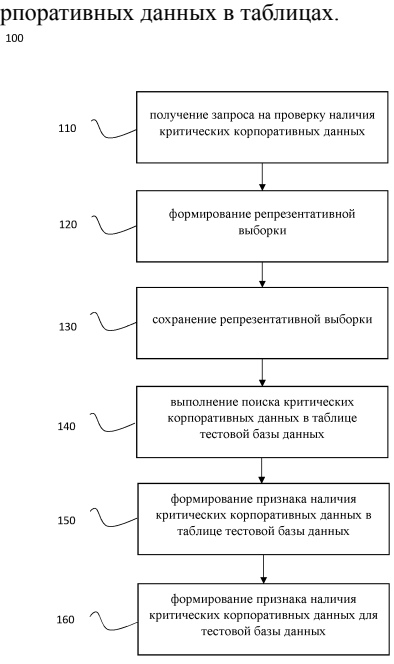
(31) **2021123150**
(32) **2021.08.03**
(33) **RU**

(71) Заявитель:
ПУБЛИЧНОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО "СБЕРБАНК РОССИИ" (ПАО СБЕРБАНК) (RU)

(72) Изобретатель:
Белорыбкин Леонид Юрьевич, Румянцев Алексей Евгеньевич, Теренин Алексей Алексеевич (RU)

(74) Представитель:
Герасин Б.В. (RU)

(57) Изобретение, в общем, относится к базам данных, а более конкретно к определению наличия критических корпоративных данных в базе данных, содержащей тестовые данные. Решаемой технической проблемой при реализации заявленного изобретения является повышение скорости проверки тестовой базы данных на наличие критических корпоративных данных. В предпочтительном варианте реализации заявлен компьютерно-реализуемый способ определения наличия критических корпоративных данных в тестовой базе данных, выполняющийся по меньшей мере одним процессором, и содержащий этапы, на которых получают запрос на проверку наличия критических корпоративных данных, из промышленной базы данных, в тестовой базе данных; получают доступ к промышленной базе данных и формируют по меньшей мере одну репрезентативную выборку; сохраняют по меньшей мере одну сформированную репрезентативную выборку в хранилище данных; получают доступ к тестовой базе данных и производят поиск значений по меньшей мере из одной репрезентативной выборки по меньшей мере в одной таблице тестовой базы данных по полям таблицы репрезентативной выборки в соответствии со списком проверок; формируют признак наличия критических корпоративных данных по меньшей мере для одной таблицы тестовой базы данных на основе результатов поиска, формируют признак наличия критических корпоративных данных для тестовой базы данных на основе признаков наличия критических корпоративных данных в таблицах.



202192295 A1

202192295 A1

МПК: G06F 16/00

СПОСОБ И СИСТЕМА ОПРЕДЕЛЕНИЯ НАЛИЧИЯ КРИТИЧЕСКИХ КОРПОРАТИВНЫХ ДАННЫХ В ТЕСТОВОЙ БАЗЕ ДАННЫХ

ОБЛАСТЬ ТЕХНИКИ

[1] Настоящее техническое решение, в общем, относится к базам данных, а более конкретно к определению наличия критических корпоративных данных в базе данных, содержащей тестовые данные.

УРОВЕНЬ ТЕХНИКИ

[2] В настоящее время процессы тестирования программного обеспечения (ПО) являются неотъемлемым этапом разработки и обновления такого ПО. Большинство компаний и организаций, деятельность которых связана с разработкой программного обеспечения, как правило, привлекают подрядчиков для проведения различных работ, например, для тестирования полученного программного продукта. Привлечение подрядчиков подразумевает доступ подрядчиков к данным, расположенным в базах данных на стендах тестирования и разработки.

[3] Однако, некоторые виды программных продуктов, разрабатываемых компаниями, требуют взаимодействия с данными ограниченного доступа, что делает выполнение работ по тестированию таких продуктов третьими лицами невозможным. К данным ограниченного доступа (критическим корпоративным данным) могут относиться персональные данные клиентов, сотрудников организации, потенциальных клиентов, данные, содержащие сведения, составляющие коммерческую, банковскую тайну и другие сведения ограниченного доступа. В соответствии с законодательными актами (Федеральный закон "О персональных данных", Федеральный закон "О банках и банковской деятельности", Федеральный закон "О коммерческой тайне") передача такого рода информации третьим лицам запрещена. Кроме того, утечки критических корпоративных данных могут повлиять на репутационные и финансовые потери организации.

[4] Для решения указанной задачи могут использоваться тестовые данные на основе обезличенных, маскируемых копий критических корпоративных данных или создания синтетических данных с помощью инструментов генерации данных. Указанные способы применимы в тех случаях, когда требуется создать базу данных, содержащую тестовые данные, на основе существующих критических корпоративных данных. Однако, для

систем, которые изначально разрабатывались и тестировались исключительно сотрудниками компании, базы данных тестовых стендов могут содержать копии критических корпоративных данных. Кроме того, за счет долгого жизненного цикла таких систем, сведения о том, содержатся ли промышленные данные в тестовых базах данных, зачастую утрачены, что не позволяет однозначно определить наличие в таких базах критических данных.

[5] Из уровня техники известны решения, направленные на проверку наличия критических корпоративных данных в тестовых базах данных, основанные на потабличном сравнении базы данных, содержащей тестовые данные, и базы данных, содержащей критические корпоративные данные (промышленные данные). При этом задача сводится к поиску всех записей (строк) из таблицы тестовой базы данных, совпадающих поколоночно с записями (строками) такой же таблицы в промышленной базе данных. Временная сложность такого алгоритма определяется временной сложностью алгоритма упорядочивания двух таблиц и алгоритма перебора записей для их сравнения, и равна O(N) в случае наличия уже упорядоченных данных, например, при наличии уникального индекса, или $O(N^2)$ в случае отсутствия какого-либо упорядочивания.

[6] Также, из уровня техники, известен способ обфускации конфиденциальных данных в промышленной базе данных, раскрытый в патенте США № US8856157 B2 (BUSINESS OBJECTS SOFTWARE LTD), опубл. 07.10.2014. Указанный способ предназначен для формирования тестовой базы данных, содержащей тестовые данные, на основе промышленной базы данных, содержащей конфиденциальные данные. Способ содержит этапы выявления, на основе метаданных столбцов промышленной базы данных, конфиденциальных данных, при экспорте из промышленной базы данных, и обфускации указанных данных. Указанный способ, в частности, также применим и к определению конфиденциальных данных в тестовой базе данных.

[7] К недостаткам указанного способа можно отнести невысокую точность определения конфиденциальных данных, за счет неточных метаданных. Кроме того, при отсутствии метаданных в столбцах тестовой базы данных, выполнение способа будет невозможно. Также, при проверке тестовой базы данных, содержащей большое количество таблиц и записей (тысячи таблиц и миллиарды записей), указанный способ будет обладать очень низкой скоростью проверки.

[8] Общими недостатками существующих решений является отсутствие эффективного и точного способа проверки базы данных на наличие критических корпоративных данных, обеспечивающего возможность проверки больших баз данных, сохраняя при этом высокую скорость проверки.

РАСКРЫТИЕ ИЗОБРЕТЕНИЕ

[9] Данное техническое решение направлено на устранение недостатков, присущих существующим решениям, известным из уровня техники.

[10] Решаемой технической проблемой в данном техническом решении является создание нового и эффективного способа определения наличия критических корпоративных данных в тестовой базе данных.

[11] Основным техническим результатом, проявляющимся при решении вышеуказанной проблемы, является повышение скорости проверки тестовой базы данных на наличие критических корпоративных данных.

[12] Дополнительным техническим результатом, проявляющимся при решении вышеуказанной проблемы, является сокращение времени выполнения проверки на наличие критических корпоративных данных в тестовой базе данных.

[13] Указанные технические результаты достигаются благодаря осуществлению компьютерно-реализуемого способа определения наличия критических корпоративных данных в тестовой базе данных, выполняющегося по меньшей мере одним процессором, и содержащего этапы на которых:

a) получают запрос на проверку наличия критических корпоративных данных, из промышленной базы данных, в тестовой базе данных, причем запрос содержит доверительную вероятность, предельно допустимую ошибку, список проверок, включающий по меньшей мере список полей для каждой проверяемой таблицы тестовой базы данных;

b) получают доступ к промышленной базе данных и формируют по меньшей мере одну репрезентативную выборку, причем указанная выборка формируется по меньшей мере на основе данных промышленной базы данных и списка проверок, доверительной вероятности, предельно допустимой ошибки;

c) сохраняют по меньшей мере одну сформированную репрезентативную выборку в хранилище данных;

d) получают доступ к тестовой базе данных и производят поиск значений из по меньшей мере одной репрезентативной выборки по меньшей мере в одной таблице тестовой базы данных по полям таблицы репрезентативной выборки в соответствии со списком проверок;

e) формируют признак наличия критических корпоративных данных для по меньшей мере одной таблицы тестовой базы данных на основе результатов поиска.

f) формируют признак наличия критических корпоративных данных для тестовой базы данных на основе признаков, полученных на этапе e).

[14] В одном из частных примеров осуществления способа для формирования репрезентативной выборки вычисляется её объем.

[15] В другом частном примере осуществления способа объем репрезентативной выборки вычисляется на основе по меньшей мере:

- алгоритма расчета объема выборки для каждой таблицы промышленной базы данных;
- алгоритма расчета объема выборки по максимально возможному значению объема выборки.

[16] В другом частном примере осуществления способа алгоритм расчета объема выборки для каждой таблицы промышленной базы данных содержит этапы, на которых:

- получают количество записей в таблице промышленной базы данных и присваивают полученное значение размеру генеральной совокупности;
- вычисляют доверительный уровень репрезентативной выборки на основе доверительной вероятности;
- вычисляют объем репрезентативной выборки на основе размера генеральной совокупности, доверительного уровня, вероятности наличия в тестовой базе данных критических корпоративных записей и предельно допустимой ошибки среднего значения выборки.

[17] В другом частном примере осуществления способа алгоритм расчета объема выборки по максимально возможному значению объема выборки содержит этапы, на которых:

- получают количество записей в таблице промышленной базы данных и присваивают полученное значение размеру генеральной совокупности;
- вычисляют доверительный уровень репрезентативной выборки на основе доверительной вероятности;
- вычисляют максимально возможный объем репрезентативной выборки на основе доверительного уровня, вероятности наличия в тестовой базе данных критических корпоративных записей и предельно допустимой ошибки среднего значения выборки;
- присваивают объему репрезентативной выборки наименьшее значение из значений максимального размера репрезентативной выборки и размера генеральной совокупности.

[18] В другом частном примере осуществления способа список проверок содержит по меньшей мере один номер проверки, который включает по меньшей мере одно поле таблицы промышленной базы данных, по значениям которого проводится поиск на совпадение в поле соответствующей таблицы тестовой базы данных.

[19] В другом частном примере осуществления способа по меньшей мере один номер проверки объединяет с помощью булевой операции конъюнкции по меньшей мере два поля таблицы промышленной базы данных, по значениям которых проводится поиск на совпадение в полях соответствующей таблицы тестовой базы данных.

[20] В другом частном примере осуществления способа признак наличия критических корпоративных данных таблицы тестовой базы данных формируется на основе объединения результатов проверки по каждому номеру проверки, согласно булевой операции дизъюнкции.

[21] В другом частном примере осуществления способа признак наличия критических корпоративных данных тестовой базы данных формируется на основе объединения результатов признаков наличия критических корпоративных данных в каждой таблице тестовой базы с помощью булевой операции дизъюнкции.

[22] В другом частном примере осуществления способа все таблицы тестовой базы данных отображаются пользователю с соответствующим признаком наличия критических корпоративных данных.

[23] Кроме того, заявленный технический результат достигается за счет системы определения наличия критических корпоративных данных в тестовой базе данных, содержащей:

- по меньшей мере один процессор;
- по меньшей мере одну память, соединенную с процессором, которая содержит машиночитаемые инструкции, которые при их выполнении по меньшей мере одним процессором обеспечивают выполнение способа определения наличия критических корпоративных данных в тестовой базе данных.

КРАТКОЕ ОПИСАНИЕ ЧЕРТЕЖЕЙ

[24] Признаки и преимущества настоящего технического решения станут очевидными из приводимого ниже подробного описания и прилагаемых чертежей, на которых:

[25] Фиг. 1 иллюстрирует блок-схему выполнения заявленного способа.

[26] Фиг. 2 иллюстрирует бок схему алгоритма расчета объема выборки для каждой таблицы промышленной базы данных.

[27] Фиг. 3 иллюстрирует блок схему алгоритма расчета объема выборки по максимально возможному значению объема выборки.

[28] Фиг. 4 иллюстрирует систему для реализации заявленного способа.

[29] Фиг. 5 иллюстрирует пример системы для реализации заявленного способа.

ОСУЩЕСТВЛЕНИЕ ИЗОБРЕТЕНИЯ

[30] Заявленное техническое решение предлагает новых подход, обеспечивающий определение наличия критических корпоративных данных в тестовой базе данных. Основной особенностью заявленного решения является обеспечение высокой скорости определения наличия критических корпоративных данных в тестовой базе данных любого размера, за счет проверки тестовой базы данных с помощью репрезентативной выборки, полученной из промышленной базы данных. Кроме того, реализация алгоритма расчета объема выборки позволяет уменьшить количество сравниваемых записей, на основе которых выполняются операции поиска данных в базах данных, что также сокращает время, требуемое на выполнение проверки тестовой базы данных на наличие критических корпоративных данных. Также, еще одним дополнительным преимуществом, достигаемым при использовании заявленного способа, является возможность проведения проверок больших тестовых баз данных (миллиарды записей) без нагрузки на промышленную базу данных.

[31] Заявленное техническое решение может выполняться, например, системой, машиночитаемым носителем, сервером и т.д. В данном техническом решении под системой подразумевается, в том числе компьютерная система, ЭВМ (электронно-вычислительная машина), ЧПУ (числовое программное управление), ПЛК (программируемый логический контроллер), компьютеризированные системы управления и любые другие устройства, способные выполнять заданную, четко определенную последовательность операций (действий, инструкций).

[32] Под устройством обработки команд подразумевается электронный блок либо интегральная схема (микропроцессор), исполняющая машинные инструкции (программы).

[33] Устройство обработки команд считывает и выполняет машинные инструкции (программы) с одного или более устройства хранения данных, например, таких устройств, как оперативно запоминающие устройства (ОЗУ) и/или постоянные запоминающие устройства (ПЗУ). В качестве ПЗУ могут выступать, но, не ограничиваясь, жесткие диски (HDD), флеш-память, твердотельные накопители (SSD), оптические носители данных (CD, DVD, BD, MD и т.п.) и др.

[34] Программа - последовательность инструкций, предназначенных для исполнения устройством управления вычислительной машины или устройством обработки команд.

[35] Термин «инструкции», используемый в этой заявке, может относиться, в общем, к программным инструкциям или программным командам, которые написаны на заданном языке программирования для осуществления конкретной функции, такой как, например, получение и обработка данных, формирование профиля пользователя, прием и передача сигналов, анализ принятых данных, идентификация пользователя и т.п. Инструкции могут быть осуществлены множеством способов, включающих в себя, например, объектно-ориентированные методы. Например, инструкции могут быть реализованы, посредством языка программирования C++, Java, Python, различных библиотек (например, “MFC”; Microsoft Foundation Classes) и т. д. Инструкции, осуществляющие процессы, описанные в этом решении, могут передаваться как по проводным, так и по беспроводным каналам передачи данных, например, Wi-Fi, Bluetooth, USB, WLAN, LAN и т. п.

[36] На **Фиг. 1** представлена блок схема способа **100** определения наличия критических корпоративных данных в тестовой базе данных, который раскрыт поэтапно более подробно ниже. Указанный способ **100** заключается в выполнении этапов, направленных на обработку различных цифровых данных. Обработка, как правило, выполняется с помощью системы, которая может представлять, например, сервер, компьютер, мобильное устройство, вычислительное устройство и т. д. Более подробно элементы системы раскрыты на **Фиг. 4**.

[37] На этапе **110** происходит получение запроса на проверку наличия критических корпоративных данных, из промышленной базы данных, в тестовой базе данных.

[38] Под критическими корпоративными данными в настоящем решении понимаются данные объекта, такого как, организация, компания, предприятие и т.д., передача которых третьим лицам является недопустимым и несет в себе существенные риски для такого объекта. Так, критическими корпоративными данными (данные ограниченного доступа, промышленные данные) могут являться, например, персональные данные клиентов, сотрудников организации, потенциальных клиентов, данные, содержащие сведения, составляющие коммерческую, банковскую тайну, врачебную тайну и т.д. не ограничиваясь.

[39] Под промышленной базой данных в данном решении понимается база данных, содержащая критические корпоративные данные.

[40] Под тестовой базой данных в данном решении понимается база данных, содержащая обработанные копии данных промышленной базы данных, удовлетворяющие входным требованиям для выполнения одного или более контрольных примеров, которые могут быть определены в плане тестирования, контрольном примере или процедуре тестирования.

[41] На этапе **110** система определения наличия критических корпоративных данных в тестовой базе данных **400**, раскрытая более подробно на **фиг. 4**, получает запрос на проверку тестовой базы данных на наличие в ней критических корпоративных данных. Указанный запрос, в одном частном варианте осуществления, может быть введен пользователем в графическом интерфейсе пользователя указанной системы **400.** В другом частном варианте осуществления запрос на проверку может быть отправлен внешней системой, например, модулем безопасности, при добавлении/обновлении тестовой базы данных в хранилище данных.

[42] В одном частном варианте осуществления запрос может дополнительно содержать, по меньшей мере, доверительную вероятность, предельно допустимую ошибку, список проверок, включающий по меньшей мере список полей для каждой проверяемой на наличие критических корпоративных данных таблицы тестовой базы данных. Кроме того, необходимо отметить, что в некоторых случаях, запрос на проверку может содержать только доверительную вероятность, предельно допустимую ошибку и название тестовой базы данных для которой осуществляется проверка наличия критических корпоративных данных.

[43] Параметр доверительной вероятности позволяет задавать необходимый уровень точности, формируемой на этапе **120** репрезентативной выборки. На основе доверительной вероятности вычисляется доверительный интервал, т.е. интервал, в пределах которого с заданной вероятностью лежат выборочные оценки статистических характеристик генеральной совокупности. Чем шире доверительный интервал для заданного уровня вероятности (например, 95%), тем ниже уровень «доверия» к выборочным оценкам, и наоборот. Широкий доверительный интервал для выборочного среднего указывает на неточное отражение средней по совокупности. Предельно допустимая ошибка среднего значения выборки, соответственно, показывает допустимое отклонение выборки от заданного уровня точности. Так, если доверительная вероятность 97%, то предельно допустимая ошибка в ± 3% будет отражать допустимый процент отклонения выборки от заданной точности. Более подробно, указанные параметры раскрыты в источнике [1].

[44] В одном частном варианте осуществления, доверительная вероятность и предельно допустимая ошибка среднего значения выборки могут быть заранее установлены в соответствии с предъявляемыми требованиями к определенным типам данных. Так, например, доверительная вероятность для паспортных данных клиентов должна быть выше, чем доверительная вероятность для их номеров телефона. В еще одном частном варианте осуществления доверительная вероятность всей базы данных, проверку которой

требуется осуществить, может быть задана в соответствии с наиболее критичным видом данных. Так, например, если база данных (такая как промышленная база данных) содержит одну таблицу с паспортными данными, а другую таблицу с номерами телефонов, то доверительная вероятность всей базы данных будет задана в соответствии с требованиями доверительной вероятности к паспортным данным, т.е. с максимальной доверительной вероятностью. Для специалиста в данной области техники очевидно, что на основе семантического анализа проверяемых данных могут быть установлены минимальные пороговые значения доверительной вероятности и предельно допустимой ошибки для любого типа данных в соответствии с заданными требованиями.

[45] Кроме того, как указывалось выше, запрос на проверку наличия критических корпоративных данных в тестовой базе данных также может содержать список проверок, включающий список полей для каждой проверяемой на наличие критических корпоративных данных таблицы тестовой базы данных. Указанный список полей содержит колонки таблиц, которые необходимо сравнить для проверки на совпадение значений из промышленной базы данных в тестовой базе данных. Поскольку тестовая база данных содержит обработанные копии данных, например, замаскированные, обезличенные и т.д., из промышленной базы данных, сохраняя при этом формат и структуру, то для ускорения проведения проверки тестовой базы данных, в некоторых случаях, целесообразнее проводить проверку не по всем таблицам тестовой базы данных, а только в представляющих интерес колонках такой базы данных. Так, например, при наличии в тестовой базе данных таблицы, содержащей данные о зарплате, данные о кредите, кредитный рейтинг и ФИО клиента, требуется проверить только ФИО клиентов. Кроме того, за счет выборочной проверки дополнительно сокращается общее время, необходимое на такую проверку, а также снижается нагрузка на промышленную базу данных.

[46] В одном частном варианте осуществления, список полей для каждой проверяемой на наличие критических корпоративных данных таблицы тестовой базы данных может быть определен автоматически на основе метаданных в заголовках таблицы или с помощью, например, семантического и лексического анализа заголовков и т.д. Также, в другом частом варианте осуществления, автоматическое определение требуемых полей для проверки может быть выполнено с помощью нейросети, алгоритма машинного обучения и т.д. Для специалиста в данной области техники очевидно, что могут быть применены любые известные из уровня техники средства для анализа и обработки данных, предназначенные для выявления заданных смысловых категорий данных.

[47] Таким образом, список проверок может содержать номера колонок всех таблиц баз данных, по значениям которых требуется провести проверку. Проверка может выполняется

с помощью поиска значений, полученных из промышленной базы данных в поле (колонке) соответствующей таблицы тестовой базы данных. В одном частном варианте осуществления список проверок содержит по меньшей мере один номер проверки, который включает по меньшей мере одно поле таблицы промышленной базы данных, по значениям которого проводится поиск на совпадение в поле соответствующей таблицы тестовой базы данных. Так как в таблице тестовой базы данных могут находиться несколько колонок, с потенциально критическими корпоративными данными, то в одном частном варианте осуществления по меньшей мере один номер проверки объединяет с помощью булевой операции конъюнкции по меньшей мере два поля таблицы промышленной базы данных, по значениям которых проводится поиск на совпадение в полях соответствующей таблицы тестовой базы данных.

[48] Далее способ **100** переходит к этапу **120**.

[49] На этапе **120** получают доступ к промышленной базе данных и формируют по меньшей мере одну репрезентативную выборку, причем указанная выборка формируется по меньшей мере на основе данных промышленной базы данных и полей, содержащихся в запросе, таком как запрос пользователя, доверительной вероятности, предельно допустимой ошибки. Как указывалось, выше, в одном частном варианте осуществления запрос может содержать доверительную вероятность, предельно допустимую ошибку и адрес расположения тестовой базы данных (например, адрес хранения указанной базы данных в хранилище данных). Для специалиста в данной области техники очевидно, что доступ на основе адреса расположения к базе данных может быть осуществлен, например, посредством элементов системы **400**, например, сетевого модуля и т.д., не ограничиваясь.

[50] На указанном этапе **120**, система **400** подключается к промышленной базе данных, например, с помощью сети Интернет, сети LAN и т.д. В одном частном варианте осуществления, если запрос на проверку тестовой базы данных был отправлен пользователем, то для получения доступа к промышленной базе данных пользователю необходимо пройти процедуру аутентификации. Аутентификация пользователя может осуществляться с помощью ввода пароля на дисплее, произнесения фразы как текст зависимой (произнесения контрольной фразы с экрана), так и текст независимой (произнесения произвольного текста), например, биометрический образец голоса, демонстрации лица пользователя (биометрический образец лица), прикладывания к соответствующему сенсору пальца, ладони и/или ключевого носителя, сканирование сетчатки глаза и т.д. Аутентификация может выполняться с помощью, например, средств ввода-вывода информации в графическом интерфейсе пользователя. Средства отображения графического пользовательского интерфейса могут являться, например, дисплей, экран,

сенсорный дисплей и т.д. средства ввода-вывода информации могут представлять собой, например, микрофонных массив или микрофон, физических и/или сенсорных клавиш (клавиатуры), сенсорный экран, считывателя отпечатка пальца, стерео-камеры, считывателя ключ-карты, динамики и т.д. Для специалиста в данной области техники очевидно, что подключение к промышленной базы данных может осуществляться любым известным из уровня техники методом с использованием сети связи.

[51] После подключения к промышленной базе данных формируется по меньшей мере одна репрезентативная выборка. Репрезентативная выборка представляет собой выборку, в которой все основные признаки генеральной совокупности (промышленной базы данных), из которой извлечена данная выборка, представлены приблизительно в той же пропорции или с той же частотой, с которой данный признак выступает в этой генеральной совокупности.

[52] Как было описано выше, некоторые тестовые базы данных могут иметь длительный срок эксплуатации, и, как следствие, могут иметь очень большую размерность (например, миллиарды записей). Потабличное сравнение данных, содержащихся в промышленной базе данных, с соответствующими данными в тестовой базе данных такого объема является длительной и ресурсозатратной процедурой. Для повышения скорости проверки тестовой базы данных на наличие критических корпоративных данных в настоящем решении предложен новый подход, основанный не на полной проверке по набору всех строк и колонок таблиц тестовой базы данных с соответствующими строками и колонками промышленной базы данных, а по репрезентативной выборке из каждой таблицы промышленной базы данных. При этом такая выборка состоит из малого количества строк по сравнению с полным кол-вом строк (разница может достигать нескольких порядков). Кроме того, в одном частном варианте осуществления, репрезентативная выборка может быть сформирована только из тех колонок, которые участвуют в проверке, например, при получении таких колонок в запросе проверки. Кроме того, применение такого подхода проверки на наличие критических корпоративных данных в тестовой базе данных, также, дает возможность уменьшить кол-во перебираемых записей и как следствие общее время, необходимое на такую проверку.

[53] Репрезентативная выборка формируется с помощью системы **400**, например, с помощью по меньшей мере одного процессора указанной системы, на основе данных, содержащихся в промышленной базе данных, доверительной вероятности и предельно допустимой ошибки. Формирование выборки происходит посредством выбора неповторяющихся случайных записей из таблиц промышленной базы данных в автоматическом режиме. Объем выборки зависит от доверительной вероятности и

предельно допустимой ошибки, заданных для такой выборки. Так, объем репрезентативной выборки может вычисляться с использованием, по меньшей мере, следующих алгоритмов: алгоритма расчета объема выборки для каждой таблицы промышленной базы данных **200**; алгоритма расчета объема выборки по максимально возможному значению объема выборки **300**.

[54] Рассмотрим более подробно каждый из указанных выше алгоритмов определения объема выборки. Указанные алгоритмы могут применяться в зависимости от количества записей и таблиц в промышленной базе данных, по которым требуется провести сравнение с тестовыми данными из тестовой базы данных.

[55] Алгоритм расчета объема выборки для каждой таблицы промышленной базы данных **200** показан на **фиг. 2**. Указанный алгоритм **200** может быть применен в тех случаях, когда в проверке используется большое количество таблиц. При этом критичными показателями являются нагрузка на систему и/или скорость обработки таблиц. Например, если при проверке тестовой базы данных на наличие критических корпоративных данных, требуется провести проверку 1000 таблиц, каждая из которых содержит менее 1000 записей, то в таком случае целесообразнее применять алгоритм **200**. Как было описано выше, при проведении проверки данные (записи) из промышленной базы данных, сравнивают с соответствующими данными (записями) из тестовой базы данных. Соответственно, чем больше таблиц требуется проверить, тем больше обращений будет к промышленной базе данных, а объем записей в таблице будет непосредственно влиять на нагрузку такой базы (чем больше записей будет извлекаться, тем большую нагрузку будут оказывать указанные операции на промышленную базу данных).

[56] На этапе **210** получают количество записей в таблице промышленной базы данных и присваивают полученное значение размеру генеральной совокупности.

[57] На указанном этапе **210** для каждой таблицы промышленной базы данных, содержащий записи, наличие которых необходимо проверить в тестовой базе данных, определяется размерность такой таблицы. Как указывалось выше, определение требуемой таблицы промышленной базы данных, на основе которой формируется репрезентативная выборка может быть получено, например, в запросе пользователя, с помощью семантического анализа заголовков таблицы или в соответствии с полями, содержащимися в списке проверок т.д. Кроме того, если в запросе содержится только адрес проверяемой базы данных, то все таблицы и поля указанных таблиц могут считаться полями, которые необходимо проверить. Общее количество записей может быть получено, например, посредством SQL функции Count (*), посредством системных функций СУБД и т.д., не ограничиваясь. Для специалиста в данной области техники очевидно, что может быть

применен любой известный из уровня техники метод, выполняющий подсчет количества записей в таблице базы данных.

[58] После определения количества записей в таблице, указанное количество присваивается значению объема генеральной совокупности. Под генеральной совокупностью, в данном решении понимается совокупность всех объектов (записей таблицы), относительно которых предполагается делать выводы при изучении конкретной задачи (проверки их наличия в тестовой базе данных).

[59] Далее, на этапе **220,** вычисляют доверительный уровень репрезентативной выборки на основе доверительной вероятности. Как было описано выше, доверительный уровень или доверительный интервал, определяет интервал, в пределах которого с заданной вероятностью лежат выборочные оценки статистических характеристик генеральной совокупности. Вычисление доверительного уровня может выполняться, например, с помощью процессора. Расчет доверительного уровня может быть произведен разными способами, например, методом итераций (последовательных приближений) или применением таблицы коэффициентов Стьюдента. Как указывалось, выше, расчет может выполняться посредством элементов системы **400,** например, посредством процессора. Для расчета доверительного уровня, в одном частном варианте осуществления может применяться следующая формула:

$$t = F^{-1}\left(\frac{1+p}{2}\right), \text{ где} \qquad (1)$$

p – доверительная вероятность

F^{-1} – функция, возвращающая обратное значение стандартного нормального распределения.

[60] Далее, на этапе **230** вычисляют объем репрезентативной выборки на основе размера генеральной совокупности, доверительного уровня, вероятности наличия в тестовой базе данных критических корпоративных записей и предельно допустимой ошибки среднего значения выборки. Формула для расчета объема репрезентативной выборки:

$$n = \left\lceil \frac{t^2\omega(1-\omega)N}{N\Delta^2 + t^2\omega(1-\omega)} \right\rceil, \text{ где} \qquad (2)$$

n – объем выборки

t – доверительный уровень

ω – вероятность события

N – размер генеральной выборки

Δ – предельная допустимая ошибка среднего значения выборки

[61] Стоит отметить, что в одном частном варианте осуществления, значение параметра ω - вероятности наличия в тестовой базе данных промышленной записи равно 0.5. Указанное значение подсчитано исходя из наихудшего сценария, когда вероятность наличия промышленной записи равна вероятности отсутствия промышленной записи. Значение предельно допустимой ошибки, в свою очередь, берется из данных входного запроса.

[62] Таким образом, благодаря использованию данного алгоритма **200**, обеспечивается дополнительное уменьшение нагрузки на промышленную базу данных, а также уменьшение количества обращений к промышленной базе данных.

[63] Теперь рассмотрим более подробно алгоритм расчета объема выборки по максимально возможному значению объема выборки **300**.

[64] Указанный алгоритм **300** может быть применен в тех случаях, когда требуется провести проверку таблиц тестовой базы данных, когда допускается повышение доверительной вероятности для таблиц за счет дополнительной нагрузки на систему и/или за счет скорости обработки таблиц. Преимуществами такого алгоритма является сокращение до нуля время на расчет размера выборки для каждой таблицы и повышение доверительной вероятности. Особенность указанного алгоритма заключается в использовании максимально возможного объема выборки в качестве присваиваемого значения объему репрезентативной выборки вне зависимости от количества записей в генеральной совокупности промышленной базы данных. При таком подходе все таблицы, в которых кол-во записей меньше чем максимально возможный объем выборки проверяются целиком, что увеличивает нагрузку на систему и уменьшает скорость обработки, но при этом доверительная вероятность по таким таблицам достигает 100%. Кроме того, расчет выборки для каждой таблицы в таком алгоритме происходит быстрее, чем в алгоритме **200**. Это обусловлено тем, что исключается необходимость в вычислении размера репрезентативной выборки (он ограничен двумя значениями, а именно либо максимальное значение размера выборки, либо значение размера проверяемой таблицы).

[65] На этапе **310** получают количество записей в таблице промышленной базы данных и присваивают полученное значение размеру генеральной совокупности.

[66] Указанный этап **310** аналогичен этапу **210**, который раскрыт более подробно выше. На указанном этапе происходит определение общего количества записей в таблице и присвоение полученного значения значению объема генеральной совокупности.

[67] Далее алгоритм переходит к этапу **320.**

[68] На этапе **320** вычисляют доверительный уровень репрезентативной выборки на основе доверительной вероятности. Указанный этап, также аналогичен этапу **220**, который описан более подробно выше. Для расчета доверительного уровня, также может быть применена формула (1).

[69] На этапе **330** вычисляют максимально возможный объем репрезентативной выборки на основе доверительного уровня, вероятности наличия в тестовой базе данных критических корпоративных записей и предельно допустимой ошибки среднего значения выборки.

[70] Для вычисления максимально возможного объема репрезентативной выборки может быть применена, например, следующая формула:

$$n_{max} = \left\lceil \frac{t^2 \omega(1-\omega)}{\Delta^2} \right\rceil, \text{ где} \qquad (3)$$

n_{max} – максимально возможный объем выборки

t – доверительный уровень

ω – вероятность события

Δ – предельная допустимая ошибка среднего значения выборки

[71] Стоит отметить, что в одном частном варианте осуществления, значение параметра ω - вероятности наличия в тестовой базе данных промышленной записи равно 0.5. Указанное значение подсчитано исходя из наихудшего сценария, когда вероятность наличия промышленной записи равна вероятности отсутствия промышленной записи. Значение предельно допустимой ошибки, в свою очередь, берется из данных входного запроса.

[72] На этапе **340** присваивают значению объема репрезентативной выборки наименьшее значение из значений максимального размера репрезентативной выборки и размера генеральной совокупности. Указанный этап может быть выполнен также с использованием SQL функции Count (*), посредством системных функций СУБД и т.д., не ограничиваясь.

[73] При одновременной проверке множества таблиц, содержащих большое количество записей могут возникать ситуации, когда в одной конкретной таблице общее количество записей будет меньше, чем максимально возможный объем выборки, вычисленный для заданного доверительного уровня. В такой случае, в качестве значения объема репрезентативной выборки будет принято значение всей генеральной совокупности.

[74] Рассмотрим пример реализации данного алгоритма **300**. Так, при доверительной вероятности 97% и предельно допустимой ошибке в 3% (математическая запись 97±3%) необходимо проверить лишь 1292 случайные записи при 100 тыс. записей в таблице, 1307 - при 1 млн. записей, 1309 - при кол-ве записей от 100 тыс. до 1 млрд. В последнем примере количество проверяемых записей сокращается в ~1 млн. раз, что повышает скорость проверки, а также сокращает время проверки.

[75] Таким образом, за счет определения максимально возможного объема выборки обеспечивается возможность присвоения каждой таблице, содержащий количество записей больше вычисленного объема выборки, заранее вычисленного значения объема, без проведения дополнительных расчётов.

[76] Таким образом, на этапе **120** происходит формирование репрезентативной выборки.

[77] Далее способ **100** переходит к этапу **130**.

[78] На этапе **130** сохраняют по меньшей мере одну сформированную репрезентативную выборку в хранилище данных.

[79] На указанном этапе выборку, сформированную на предыдущем шаге, выгружают из промышленной базы данных в промежуточное надежное хранилище, с ограниченными правами доступа. В качестве хранилища **130** может быть использовано, например, репозиторий данных, который может представлять собой средство хранения данных **403**, которое более подробно раскрыто на **фиг. 4**. Сохранение выборки в отдельном хранилище позволяет уменьшить нагрузку на промышленную базу данных и выполнить проверку тестовой базы данных без дальнейшего обращения к промышленной базе данных.

[80] Стоит отметить, что в одном частном варианте осуществления, сформированная репрезентативная выборка также может быть сохранена в промышленной базе данных, однако, при дальнейшей проверки тестовой базы данных, к промышленной базе данных будут совершаться обращения для извлечения записей репрезентативной выборки, что, как было указано выше, приведет к дополнительной нагрузке и, как следствие, может замедлить параллельные процессы, выполняющиеся в промышленной базе данных.

[81] На этапе **140** получают доступ к тестовой базе данных и производят поиск значений из по меньшей мере одной репрезентативной выборки по меньшей мере в одной таблице тестовой базы данных по полям таблицы репрезентативной выборки в соответствии со списком проверок.

[82] На основе сформированной репрезентативной выборки выполняется поиск записей, содержащийся в указанной выборке, в соответствующих таблицах тестовой базы данных. На указанном этапе **140** по каждому значению из репрезентативной выборки проводится поиск соответствия указанного значения в полном наборе записей тестовой базы данных.

Указанный поиск может быть выполнен, процессором системы, например, при помощи, SQL запросов, направленных в СУБД.

[83] Так, например, если объем репрезентативной выборки будет составлять 100 записей, а объем таблицы тестовой базы данных будет составлять 1500 записей. То по каждому значению выборки будет выполнен поиск соответствия по всем 1500 записях таблицы тестовой базы данных.

[84] Подключение к тестовой базе данных, может осуществляться, например, с помощью сети Интернет, сети LAN и т.д.

[85] Кроме того, в одном частном варианте осуществления, поиск может быть выполнен в соответствующих колонках таблицы тестовой базы данных на основе номеров проверки, которые содержатся в списке проверок. Как указывалось выше, поиск может быть выполнен и по всем таблицам тестовой базы данных (в случаях указания только адреса базы данных).

[86] Так, если требуется провести проверку на наличие критических корпоративных данных не во всей таблице тестовой базы данных, а только в определенных колонках, то список проверок будет содержать список полей (колонок) таблиц базы данных, по значениям которых проводится проверка на совпадение в строках тестовой и промышленной базы данных.

[87] Например, при необходимости проверки только первой колонки тестовой базы данных на наличие в ней корпоративных данных, поиск на совпадение значений из репрезентативной выборки будет проведен только по указанной колонке. Указанная особенность, также позволяет сократить общее время проверки и/или исключить колонки тестовой базы данных, о которых присутствует информация об отсутствии критических корпоративных данных.

[88] В еще одном частном варианте осуществления, по меньшей мере один номер проверки может объединять с помощью булевой операции конъюнкции по меньшей мере два поля таблицы промышленной базы данных, по значениям которых проводится поиск на совпадение в полях соответствующей таблицы тестовой базы данных.

[89] Так, например, если в номере проверки указаны сразу два поля таблицы тестовой базы данных, значение которых необходимо сопоставить с соответствующими значениями идентичных полей в промышленной базе данных, то при проведении поиска требуется учитывать зависимость одного поля от другого, а следовательно, требуется проводить проверку по двум полям одновременно и учитывать результаты наличия критических корпоративных данных только в тех случаях, когда совпадение происходит и в первой и во второй колонках одновременно в соответствующей строке. Для решения указанной

проблемы результаты проверки полей объединяются с помощью булевой операции конъюнкции, что обеспечивает возможность проведения поиска по нескольким полям.

[90] Так, если в таблице тестовой базы данных содержатся колонки (столбцы), включающие персональные данные клиентов, например, колонка 1 содержит фамилии клиентов, колонка 2 содержит имена клиентов, колонка 3 содержит отчество клиентов, колонка 4 содержит номер паспорта клиентов, колонка 5 содержит адрес клиента, то для исключения утечки персональных данных, первый номер проверки будет содержать проверку совпадения по первым четырём колонкам (колонка 1 – колонка 4), а не по совпадению значений каждой отдельной колонки с колонкой из репрезентативной выборки. Это связано с тем, что по отдельности фамилии имена и отчества могут совпадать, но одно лишь совпадения в какой-либо из этих колонок не позволит идентифицировать личность клиента, однако если совпадают все колонки, то в таком случае указанное лицо уже можно идентифицировать, а, следовательно, завладеть его персональными данными и использовать их в мошеннических целях.

[91] Далее способ **100** переходит на этап **150.**

[92] На этапе **150** формируют признак наличия критических корпоративных данных для по меньшей мере одной таблицы тестовой базы данных на основе результатов поиска.

[93] По результатам проведение поиска на этапе **140**, для каждой таблицы формируется признак наличия критических корпоративных данных. Указанный признак формируется на основе по меньшей мере одного совпадения, при проведении поиска, значения из репрезентативной выборки со значением в тестовой базе данных.

[94] Кроме того, если поиск по таблице проводился по нескольким номерам проверки, то признак наличия критических корпоративных данных таблицы тестовой базы данных формируется на основе объединения результатов проверки по каждому номеру проверки, согласно булевой операции дизъюнкции.

[95] Для специалиста в данной области техники очевидно, что если хотя бы в одном номере проверки выявилось совпадение значений из репрезентативной выборки со значениями в таблице тестовой базы данных, то признак наличия критических корпоративных данных должен быть присвоен всей таблице, независимо от результатов других номеров проверки. Для выполнения этого условия и используется булева операция дизъюнкции, на основе которой обеспечивается возможность присвоения признака наличия критических корпоративных данных всей таблице на основе выявления хотя бы одного совпадения.

[96] На этапе **160** формируют признак наличия критических корпоративных данных для тестовой базы данных на основе признаков наличия критических корпоративных данных в таблицах.

[97] Для каждой тестовой базы данных, проверяемой на наличие критических корпоративных данных, формируется признак их наличия. Т.к. каждая тестовая база данных может содержать несколько таблиц, то на основе наличия критических корпоративных данных по меньшей мере в одной таблице, формируется признак для всей базы данных.

[98] Так, в одном частном варианте осуществления, признак наличия критических корпоративных данных тестовой базы данных формируется на основе объединения результатов признаков наличия критических корпоративных данных в каждой таблице тестовой базы с помощью булевой операции дизъюнкции. Указанный принцип аналогичен принципу, изложенному на этапе **150**. Очевидно, что если хотя бы в одной таблице выявилось наличие критических корпоративных данных, то всей тестовой базе данных должен быть присвоен признак наличия корпоративных данных. Для выполнения этого условия и используется булева операция дизъюнкции, на основе которой обеспечивается возможность присвоения признака наличия критических корпоративных данных всей базе данных на основе выявления хотя бы одного совпадения.

[99] Кроме того, в одном частном варианте осуществления по результатам проверки формируется отчет, содержащий список таблиц с указанием наличия в них корпоративных данных. Так, например, в графическом интерфейсе пользователя по результатам проверки отображаются все таблицы со статусом наличия в указанной таблице критических корпоративных данных по результатам проверки.

[100] На основе отображенных результатов проверки принимаются дальнейшие решения по тестовой базе данных. Так, если по результатам проверки выявляются таблицы с критическими корпоративными данными, то для деперсонализации данных и защиты других данных ограниченного доступа могут быть совершены необходимые для этого процедуры, например, маскирование, обезличивание, анонимизация и т.д. Указанные процедуры широко известны из уровня техники и могут применяться любые из известных методов для деперсонализации данных и защиты других данных ограниченного доступа.

[101] Соответственно, если тестовая база данных не содержит критических корпоративных данных, то к указанной базе может быть предоставлен доступ третьим лицам.

[102] Таким образом, в вышеприведенных материалах был описан способ определения критических корпоративных данных в тестовой базе данных, обеспечивающий высокую скорость проверки не зависимо от размера тестовой базы данных.

[103] На **Фиг. 4** представлена система (400), реализующая этапы заявленного способа (100).

[104] В общем случае система (400) содержит такие компоненты, как: один или более процессоров (401), по меньшей мере одну память (402), средство хранения данных (403), интерфейсы ввода/вывода (404), средство В/В (405), средство сетевого взаимодействия (406), которые объединяются посредством универсальной шины.

[105] Процессор (401) выполняет основные вычислительные операции, необходимые для обработки данных при выполнении способа (100). Процессор (401) исполняет необходимые машиночитаемые команды, содержащиеся в оперативной памяти (402).

[106] Память (402), как правило, выполнена в виде ОЗУ и содержит необходимую программную логику, обеспечивающую требуемый функционал.

[107] Средство хранения данных (403) может выполняться в виде HDD, SSD дисков, рейд массива, флэш-памяти, оптических накопителей информации (CD, DVD, MD, Blue-Ray дисков) и т.п. Средства (403) позволяют выполнять долгосрочное хранение различного вида информации, например, истории обработки транзакционных запросов (логов), идентификаторов пользователей и т.п.

[108] Для организации работы компонентов системы (400) и организации работы внешних подключаемых устройств применяются различные виды интерфейсов В/В (404). Выбор соответствующих интерфейсов зависит от конкретного исполнения вычислительного устройства, которые могут представлять собой, не ограничиваясь: PCI, AGP, PS/2, IrDa, FireWire, LPT, COM, SATA, IDE, Lightning, USB (2.0, 3.0, 3.1, micro, mini, type C), TRS/Audio jack (2.5, 3.5, 6.35), HDMI, DVI, VGA, Display Port, RJ45, RS232 и т.п.

[109] Выбор интерфейсов (404) зависит от конкретного исполнения системы (400), которая может быть реализована на базе широко класса устройств, например, персональный компьютер, мейнфрейм, ноутбук, серверный кластер, тонкий клиент, смартфон, сервер и т.п.

[110] В качестве средств В/В данных (405) может использоваться: клавиатура, джойстик, дисплей (сенсорный дисплей), монитор, сенсорный дисплей, тач-пад, манипулятор мышь, световое перо, стилус, сенсорная панель, трекбол, динамики, микрофон, средства дополненной реальности, оптические сенсоры, планшет, световые индикаторы, проектор, камера, средства биометрической идентификации (сканер сетчатки глаза, сканер отпечатков пальцев, модуль распознавания голоса) и т.п.

[111] Средства сетевого взаимодействия (406) выбираются из устройств, обеспечивающий сетевой прием и передачу данных, например, Ethernet карту, WLAN/Wi-Fi модуль, Bluetooth модуль, BLE модуль, NFC модуль, IrDa, RFID модуль, GSM модем и т.п. С помощью средств (405) обеспечивается организация обмена данными между, например, системой (400), представленной в виде сервера и вычислительным устройством пользователя, на котором могут отображаться полученные данные (результаты проверки тестовой базы данных на наличие критических корпоративных данных) по проводному или беспроводному каналу передачи данных, например, WAN, PAN, ЛВС (LAN), Интранет, Интернет, WLAN, WMAN или GSM.

[112] На **фиг. 5** показан частный случай реализации системы **400.**

[113] Указанная система состоит из промышленной базы данных **501**, тестовой базы данных **502**, графического интерфейса пользователя **503**, модуля управления **504**, модуля получения выборки **505** модуля проверки базы данных **506**, репозитория данных **507**.

[114] Модули **504-506** могут являться программно-аппаратными средствами и могут представлять собой, по меньшей мере, сервер, компьютер и т.д., выполняющий предписанную ему функцию. Кроме того, для специалиста очевидно, что указанные модули могут быть реализованы с помощью по меньшей мере одного процессора и могут являться логическими модулями.

[115] Представленные материалы заявки раскрывают предпочтительные примеры реализации технического решения и не должны трактоваться как ограничивающие иные, частные примеры его воплощения, не выходящие за пределы испрашиваемой правовой охраны, которые являются очевидными для специалистов соответствующей области техники.

Источники информации:

1. Лекции по теории вероятностей и математической статистике, И.Н. Володин, Казанский государственный университет, Казань – 2006, с. 226-239. Найдено в Интернет 20.03.2021 по адресу: https://kpfu.ru/docs/F1021260618/TViMS.pdf

ФОРМУЛА

1. Компьютерно-реализуемый способ определения наличия критических корпоративных данных в тестовой базе данных, выполняющийся по меньшей мере одним процессором, и содержащий этапы на которых:
    a) получают запрос на проверку наличия критических корпоративных данных, из промышленной базы данных, в тестовой базе данных, причем запрос содержит доверительную вероятность, предельно допустимую ошибку, список проверок, включающий по меньшей мере список полей для каждой проверяемой таблицы тестовой базы данных;
    b) получают доступ к промышленной базе данных и формируют по меньшей мере одну репрезентативную выборку, причем указанная выборка формируется по меньшей мере на основе данных промышленной базы данных и списка проверок, доверительной вероятности, предельно допустимой ошибки;
    c) сохраняют по меньшей мере одну сформированную репрезентативную выборку в хранилище данных;
    d) получают доступ к тестовой базе данных и производят поиск значений из по меньшей мере одной репрезентативной выборки по меньшей мере в одной таблице тестовой базы данных по полям таблицы репрезентативной выборки в соответствии со списком проверок;
    e) формируют признак наличия критических корпоративных данных для по меньшей мере одной таблицы тестовой базы данных на основе результатов поиска.
    f) формируют признак наличия критических корпоративных данных для тестовой базы данных на основе признаков, полученных на этапе e).
2. Способ по п.1, характеризующийся тем, что для формирования репрезентативной выборки вычисляется её объем.
3. Способ по п.2, характеризующийся тем, что объем репрезентативной выборки вычисляется на основе по меньшей мере:
    - алгоритма расчета объема выборки для каждой таблицы промышленной базы данных;
    - алгоритма расчета объема выборки по максимально возможному значению объема выборки.
4. Способ по пп. 2-3, характеризующийся тем, что алгоритм расчета объема выборки для каждой таблицы промышленной базы данных содержит этапы, на которых:

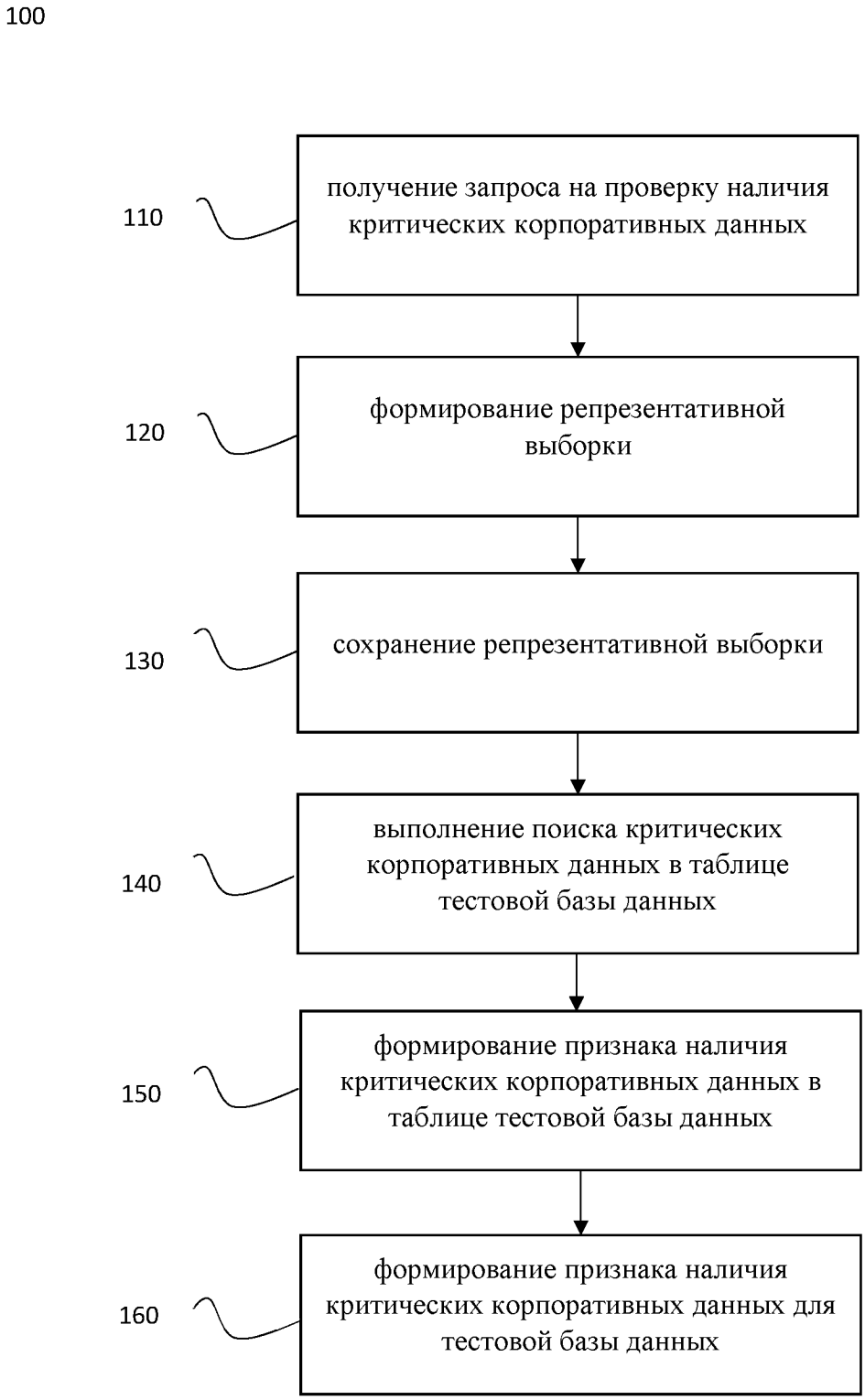
- получают количество записей в таблице промышленной базы данных и присваивают полученное значение размеру генеральной совокупности;
- вычисляют доверительный уровень репрезентативной выборки на основе доверительной вероятности;
- вычисляют объем репрезентативной выборки на основе размера генеральной совокупности, доверительного уровня, вероятности наличия в тестовой базе данных критических корпоративных записей и предельно допустимой ошибки среднего значения выборки.

5. Способ по пп. 2-3, характеризующийся тем, что алгоритм расчета объема выборки по максимально возможному значению объема выборки содержит этапы, на которых:
   - получают количество записей в таблице промышленной базы данных и присваивают полученное значение размеру генеральной совокупности;
   - вычисляют доверительный уровень репрезентативной выборки на основе доверительной вероятности;
   - вычисляют максимально возможный объем репрезентативной выборки на основе доверительного уровня, вероятности наличия в тестовой базе данных критических корпоративных записей и предельно допустимой ошибки среднего значения выборки;
   - присваивают значению объема репрезентативной выборки наименьшее значение из значений максимального размера репрезентативной выборки и размера генеральной совокупности.
6. Способ по п.1, характеризующийся тем, что список проверок содержит по меньшей мере один номер проверки, который включает по меньшей мере одно поле таблицы промышленной базы данных, по значениям которого проводится поиск на совпадение в поле соответствующей таблицы тестовой базы данных.
7. Способ по п.6, характеризующийся тем, что по меньшей мере один номер проверки объединяет с помощью булевой операции конъюнкции по меньшей мере два поля таблицы промышленной базы данных, по значениям которых проводится поиск на совпадение в полях соответствующей таблицы тестовой базы данных.
8. Способ по любому из п. 7, характеризующийся тем, что признак наличия критических корпоративных данных таблицы тестовой базы данных формируется на основе объединения результатов проверки по каждому номеру проверки, согласно булевой операции дизъюнкции.
9. Способ по п. 8, характеризующийся тем, что признак наличия критических корпоративных данных тестовой базы данных формируется на основе объединения

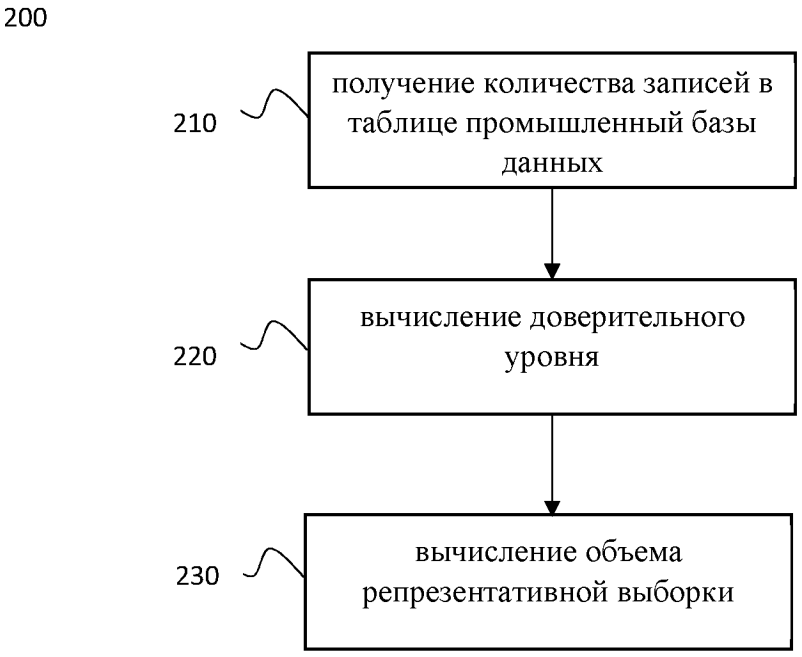
результатов признаков наличия критических корпоративных данных в каждой таблице тестовой базы с помощью булевой операции дизъюнкции.

10. Способ по п.1, характеризующийся тем, что все таблицы тестовой базы данных отображаются пользователю с соответствующим признаком наличия критических корпоративных данных.
11. Система определения наличия критических корпоративных данных в тестовой базе данных, содержащая:
    - по меньшей мере один процессор;
    - по меньшей мере одну память, соединенную с процессором, которая содержит машиночитаемые инструкции, которые при их выполнении по меньшей мере одним процессором обеспечивают выполнение способа по любому из п.п. 1-10.

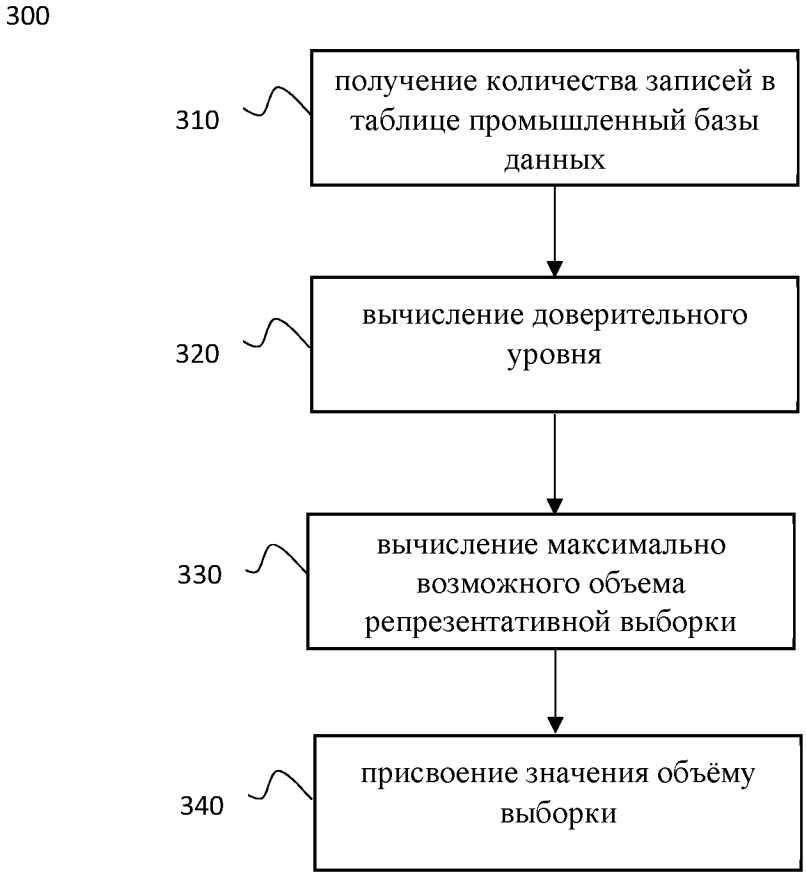
ЧЕРТЕЖИ К ОПИСАНИЮ



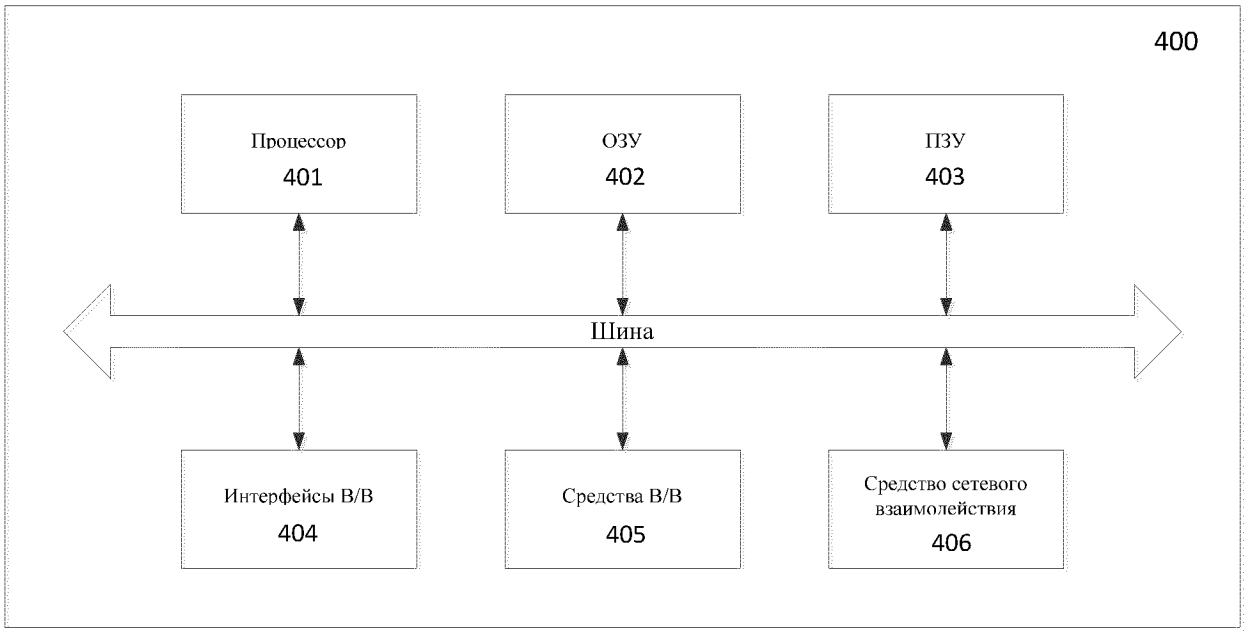
Фиг. 1


200
210
получение количества записей в таблице промышленный базы данных
220
вычисление доверительного уровня
230
вычисление объема репрезентативной выборки

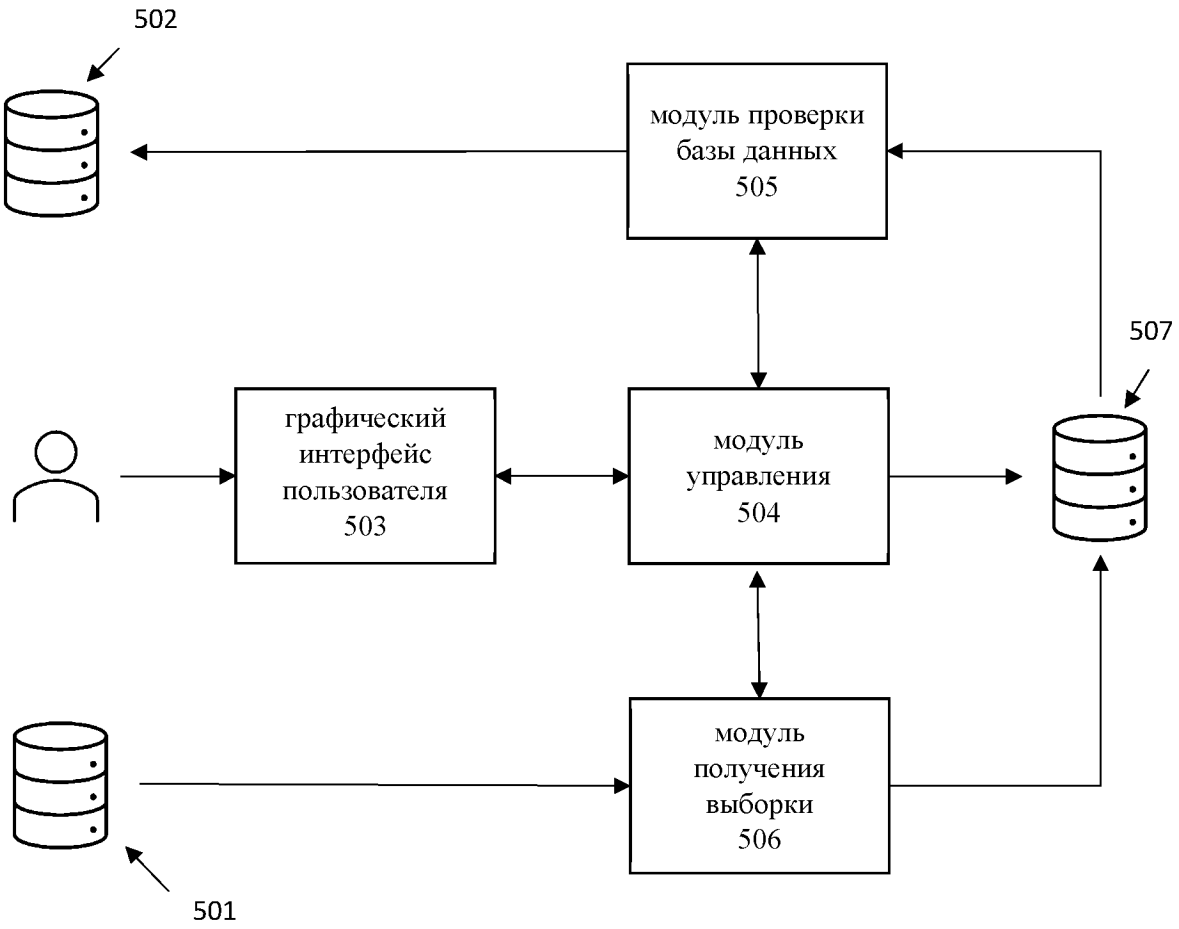
Фиг. 2


300
310
получение количества записей в таблице промышленный базы данных
320
вычисление доверительного уровня
330
вычисление максимально возможного объема репрезентативной выборки
340
присвоение значения объёму выборки

Фиг. 3


400
Процессор
401
ОЗУ
402
ПЗУ
403
Шина
Интерфейсы В/В
404
Средства В/В
405
Средство сетевого взаимолействия
406

Фиг. 4


502
модуль проверки базы данных 505
507
графический интерфейс пользователя 503
модуль управления 504
модуль получения выборки 506
501

Фиг. 5

ОТЧЕТ О ПАТЕНТНОМ ПОИСКЕ

(статья 15(3) ЕАПК и правило 42 Патентной инструкции к ЕАПК)

Номер евразийской заявки:

202192295

А. КЛАССИФИКАЦИЯ ПРЕДМЕТА ИЗОБРЕТЕНИЯ:

G06F 7/06 *(2006.01)*
G06F 16/438 *(2006.01)*
G06F21/60 *(2006.01)*

Согласно Международной патентной классификации (МПК)

Б. ОБЛАСТЬ ПОИСКА:

Просмотренная документация (система классификации и индексы МПК)
G06F 7/00-7/06, 17/00, 17/30, 21/60-21/62, 3/482-3/484; H04L 29/06; G06Q 10/06, 99/00; H04L 29/06

Электронная база данных, использовавшаяся при поиске (название базы и, если, возможно, используемые поисковые термины)
ESPACENET; GOOGLEPATENTS; YANDEX;

В. ДОКУМЕНТЫ, СЧИТАЮЩИЕСЯ РЕЛЕВАНТНЫМИ

Категория*	Ссылки на документы с указанием, где это возможно, релевантных частей	Относится к пункту №
A	US20150324606 A1 (INFORMATICA LLC) 12-11-2015 абзацы [0005]-[0011]; [0054]; [0056-0058]; [0061]; [0062]; [0065]; [0066]; [0068]; [0075-0079]; [0081]; [0083-089]; [0101]; [0103]	1-11
A	US20180096021 A1 (SWISSCOM AG) 05-04-2018 [0018]; [0019]; [0023-0025]; [0029]; [0036]; [0038]; [0043]; [0047]; [0048]; [0050]; [0051]; [0053-0055]; [0057];	1-11
A	US20100088305 A1 (AB INITIO TECHNOLOGY LLC) 08-04-2010 абзацы [0003]; [0007]; [0009-0011]; [0027]; [0029]; [0031-0039]; [0043]; [0047-0052];	1-11
A	US20190124119 A1 (ONETRUST LLC) 25-04-2019 реферат; абзацы [0003]; [0009-0011]; [0031]; [0052]; [0054]; [0056]; [0061-0063]; [0070]; [0086];	1-11
A	US20100121773 A1 (INTERNATIONAL BUSINESS MACHINES CORP) 13-05-2010 [0006]; [0017]; [0018]; [0034]; [0050]; [0051];	1-11
A	US20100042583 A1 (HARTFORD FIRE INSURANCE CO) 18-02-2010 [0021-0024]; [0028]; [0029]; [0032]; [0035]; [0038][0040-0043]; [0046]; [0054]	1-11
A	US20150161397 A1 (MICROSOFT TECHNOLOGY LICENSING LLC) 11-06-2015 [0002]; [0004]; [0014]; [0022]; [0024-0029]; [0040]; [0042];	1-11
A	US20190012476 A1 (KASPERSKY LAB AO) 10-01-2019 [0003]; [0007-0010];[0012]; [0023-0034]; [0040-0046];	1-11

☐ последующие документы указаны в продолжении

* Особые категории ссылочных документов:
«A» - документ, определяющий общий уровень техники
«D» - документ, приведенный в евразийской заявке
«E» - более ранний документ, но опубликованный на дату подачи евразийской заявки или после нее
«O» - документ, относящийся к устному раскрытию, экспонированию и т.д.
"P" - документ, опубликованный до даты подачи евразийской заявки, но после даты испрашиваемого приоритета"

«T» - более поздний документ, опубликованный после даты приоритета и приведенный для понимания изобретения
«X» - документ, имеющий наиболее близкое отношение к предмету поиска, порочащий новизну или изобретательский уровень, взятый в отдельности
«Y» - документ, имеющий наиболее близкое отношение к предмету поиска, порочащий изобретательский уровень в сочетании с другими документами той же категории
«&» - документ, являющийся патентом-аналогом
«L» - документ, приведенный в других целях

Дата проведения патентного поиска: **31/03/2022**

Уполномоченное лицо:
Начальник отдела механики,
физики и электротехники

Д.Ф. Крылов