Здравоохранение в информационном обществе

ЦИФРОВИЗАЦИЯ ЗДРАВООХРАНЕНИЯ: ЧТО МОЖНО СДЕЛАТЬ УЖЕ СЕЙЧАС

Богданов Александр Владимирович

Доктор физико-математических наук, профессор
Санкт-Петербургский государственный университет, кафедра фундаментальной информатики и распределенных систем, профессор
Научно-аналитический журнал «Информационное общество», член Редакционного совета журнала
Санкт-Петербург, Россия
a.v.bogdanov@spbu.ru

Залуцкая Наталья Михайловна

Кандидат медицинских наук, доцент
Федеральное государственное бюджетное учреждение «Национальный медицинский исследовательский центр психиатрии и неврологии им. В.М. Бехтерева» Министерства здравоохранения Российской Федерации, ведущий научный сотрудник
Санкт-Петербург, Россия
nzalutskaya@yandex.ru

Щеголева Надежда Львовна

Доктор технических наук, профессор
Санкт-Петербургский государственный университет, кафедра компьютерного моделирования и многопроцессорных систем, профессор
Санкт-Петербург, Россия
n.shchegoleva@spbu.ru

Зайналов Нодир Расулович

Кандидат физико-математических наук, доцент
Самаркандский филиал Ташкентского университета информационных технологий, заведующий кафедрой информационной безопасности
Самарканд, Узбекистан
nodirz@mail.ru

Киямов Жасур Уткирович

Санкт-Петербургский государственный университет, кафедра фундаментальной информатики и распределенных систем, аспирант
Санкт-Петербург, Россия
st080634@student.spbu.ru

Дик Александр Геннадьевич

Санкт-Петербургский государственный университет, кафедра фундаментальной информатики и распределенных систем, аспирант
Санкт-Петербург, Россия
st087383@student.spbu.ru

https://doi.org/10.52605/16059921_2022_05_58

Аннотация

Медико-биологические исследования являются прекрасным примером генерации сверхбольших объемов данных разных типов, с которыми невозможно справиться простыми средствами. Тот факт, что существует более 2000 программ для работы с различными типами данных, в том числе с большими данными, делает задачу их обработки чрезвычайно сложной даже для больших федеральных центров. Корень проблем лежит в необходимости предварительной работы с данными и проведения двух операций – консолидации и виртуализации. Хранилища могут быть разных типов, в том числе порталы, архивы, витрины, базы данных разных видов, облака данных и сети. Они могут иметь синхронные или асинхронные компьютерные соединения. Поскольку тип данных часто заранее неизвестен, существует необходимость в очень гибкой системе хранения, которая позволила бы легко переключаться между различными источниками и системами. Сочетание концепции виртуального персонального суперкомпьютера с классификацией больших данных, учитывающей различные схемы хранения, позволяет решить эту проблему.

Ключевые слова

большие данные, виртуализация данных, виртуальный персональный суперкомпьютер, сеть передачи данных, рынки данных

Введение

В настоящее время необходимость создания больших отраслевых информационных систем осознается на всех уровнях, от руководителей ведомств до практикующих врачей. Биомедицинские технологии занимают особое место в парадигме информатизации из-за гигантских объемов обрабатываемых данных и большого разнообразия их типов, а, следовательно, протоколов работы с ними. Практика показала, что для эффективной работы с такими данными необходимо реализовать возможность доступа к ним через единый шлюз и в рамках единого программного продукта. Такой подход носит название консолидация [1] и наиболее эффективно может осуществляться в рамках федеративной распределенной Базы Данных. Такой подход можно непосредственно реализовать для реляционных данных, однако в последнее время в медицине все больше используются потоковые данные (ЭКГ, ФМРТ и др.). Все это приводит к необходимости составлять для консолидации стэки программ, оптимизация которых представляет отдельную проблему.

Исследования, проводимые в данном направлении, показали, что, хотя для каждого отдельного случая такая оптимизация возможна, количество ситуаций, а, значит, и количество необходимых программ, столь велико, что даже специалисты-информационщики не всегда могут с ними оптимальным образом работать. Выходом из этой ситуации может стать виртуализация [2]. Виртуальная машина — это комплекс программ, который эмулирует реальный процессор с требуемыми функциональностями. Аналогичным образом можно построить виртуальную память, виртуальную сеть и виртуальную файловую систему. Достоинства виртуализации состоят в том, что в полученной системе используются только нужные приложения, параметры ее адаптированы под параметры решаемой задачи, а если такая система не используется, она не требует ресурсов. Такой подход позволяет всем вычислительным объектам, таким как приложения, компьютеры, машины, сети, данные и даже услуги, преодолевать физические ограничения с помощью широкого спектра технологий, инструментов и методов, а также обеспечивает значительные операционные преимущества для всей инфраструктуры. Таким образом, во все более виртуализирующемся мире наиболее эффективным подходом к данным являются структуры, позволяющие виртуализировать их. Хорошо известный подход к виртуализации данных, предложенный в [3–5], дополняет упомянутую выше парадигму. В его основе лежит создание универсальной и гибкой системы хранения и обработки данных, позволяющей реализовать все достоинства виртуализированных систем.

Для понимания основных проблем систем хранения больших данных, достаточно заметить, что они возникают, когда нужно сопоставить конкретный тип Больших данных с типом подключения хранилища к серверам данных и организации работы с данными. Самый естественный способ классификации больших данных следует из теоремы Брюера и приводит к шести различным типам данных [3]. Использование облачного хранилища в дополнение к традиционному асинхронному и синхронному подключению позволяет реализовать различные способы подключения хранилища к серверам данных, чаще всего гибридные [4]. И, наконец, работа с данными может осуществляться в рамках электронных архивов, регистров, баз данных, баз

знаний, потоковых библиотек, озер данных, сеток данных и т. д. При этом в распределенной системе, как правило, реализуется несколько способов организации работы с данными.

Поэтому сейчас центральной проблемой систем хранения является их гибкость. Теперь уже очевидно, что для достижения гибкости наиболее эффективно использовать виртуализацию. В отчете Gartner [5] дано такое определение виртуализации данных - объединение запросов к различным источникам данных в виртуальные образы, которые затем используются приложениями или промежуточным программным обеспечением для создания аналитических выводов. Однако с таким непосредственным пониманием виртуализации пользователю потребуется достаточно высокая квалификация и много технических усилий для достижения эффективности. Выходом из сложившейся ситуации является интеллектуальная виртуализация [6]. Основная идея такого подхода состоит в выполнении основной части вычислений на удаленных ресурсах, которые сгруппированы для увеличения скорости обработки по типам данных и используемым инструментам. Несмотря на привлекательность такого подхода, его все же довольно сложно реализовать, а кроме того, на распределенных системах существует проблема снижения скорости обработки из-за необходимости контролировать ошибки данных, объединенных в один пул (ситуация очень похожа на проблему снижения скорости обработки консенсуса в распределенных реестрах) [7, 8].

Мы полагаем, что значительная часть проблем может быть решена, если будет использована парадигма виртуального персонального суперкомпьютера [9], которая была разработана для вычислений, однако использовалась и для построения структуры для распределенных реестров [10]. Идея этого подхода заключается в виртуализации не только самой обработки, но и всего поля, в котором выполняется обработка, а именно сети, файловой системы и разделяемой памяти. Это позволяет создать единый образ операционной среды, что упрощает работу пользователя и увеличивает скорость обработки. В этой статье мы покажем, как предлагаемый подход позволяет создать экосистему, сочетающую в себе функции федеративных баз данных, озер данных и сетей данных.

1 Подходы к хранению сверхбольших объемов данных

Данные — это необработанная информация без контекста. Концепция информации гарантирует, что люди имеют дело с данными, вплетенными в контекст. Однако данные и информация — это всего лишь материал, пригодный для отчетности. Данные, которые ориентированы на бизнес-контекст или несут некоторую функциональность, — это знания.

Поэтому самое пристальное внимание нужно обратить на источники самих данных. Различие в объеме данных также играет огромную роль. Влияние этого фактора особенно заметно для научных проектов. Помимо самой структуры данных, в идеале следует также обратить внимание на физическую доступность этих данных и на то, сколько ресурсов будет доступно разработчикам (при этом стоит учитывать, что переход в облако может в значительной степени упростить развертывание и дальнейшую поддержку).

Кроме того, стоит учитывать, что данные можно разделить на такие категории, как структурированные, неструктурированные и слабо структурированные (полу структурированные данные). Структурированные данные — это данные с определенной моделью и структурой, например, базы данных. Неструктурированные данные не имеют структуры и часто хранятся в двоичном формате, например, в виде изображения. Слабо структурированные данные — это текстовые данные, хранящиеся по некоторому шаблону. Примерами являются файлы с расширениями .log,. json, .xml. Как показывает практика, неструктурированных данных во много раз больше, чем полу структурированных и структурированных данных. В то же время неструктурированная информация быстро накапливается и несет в себе много потенциально важной информации.

Отметим основные характеристики современных платформ данных: централизованная, монолитная, с тесно связанной конвейерной архитектурой, управляемая группой высококвалифицированных инженеров по данным.

В настоящее время существует три способа организации хранения данных:

1) Собственные корпоративные хранилища данных и платформы бизнес-аналитики, которые представляют собой чрезвычайно дорогие решения, понятные лишь небольшой группе специалистов, что приводит к недооценке положительного влияния такого хранилища на бизнес.

2) Экосистема больших данных с озером данных, управляемая командой высококлассных инженеров по данным - Data Marketplace.

3) Существующие решения в той или иной степени похожи на предыдущее поколение, с уклоном в сторону потоковой передачи для обеспечения доступности данных в реальном времени с такими архитектурами, как Kappa (рис. 1), сочетающими пакетную и потоковую обработку для преобразования данных с такими платформами, как Apache. Beam, а также полностью управляемые облачные сервисы хранения, механизмы конвейера данных и платформы машинного обучения. Очевидно, что такая платформа данных устраняет некоторые проблемы предыдущих, такие как анализ данных в реальном времени, но также снижает затраты на управление инфраструктурой больших данных. Однако они сохраняют часть проблем предыдущих решений.

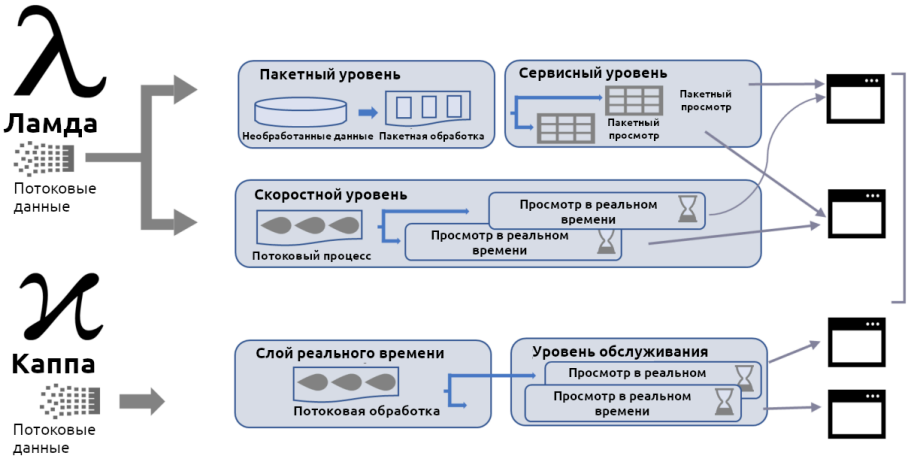

Рис. 1. Архитектуры Лямбда и Каппа.

При реализации архитектуры централизованной платформы данных можно выделить следующие основные проблемы, которые часто приводят к ее отказу:

1. Постоянное появление новых источников данных. По мере того, как становится доступным больше данных, возможность использовать и координировать их все в одном месте под контролем одной платформы уменьшается.

2. Потребности организаций в новых комбинациях данных приводят к постоянно растущему количеству их преобразований – агрегирование, построение проекций и срезов, что увеличивает время отклика. Это всегда было проблемой и остается проблемой в сегодняшней архитектуре платформы данных.

Учитывая влияние предыдущих поколений на архитектуру платформ данных, специалисты по их внедрению выделяют несколько этапов обработки данных. Именно это относится к проблеме структурирования команд, которые создают платформу и управляют ею. Как правило, некоторые из них являются высококлассными инженерами данных, которые понимают источники происхождения данных и принципы их использования для принятия решений. Другая часть — это специалисты, имеющие большой опыт технической работы с инструментами для работы с Big Data. Последние, однако, часто не обладают знаниями в сфере бизнеса и предметной области.

Формирование новой корпоративной архитектуры платформы данных в виде распределенной сети передачи данных — это новая парадигма в этой области, позволяющая решить указанные выше проблемы.

Чтобы децентрализовать платформу монолитных данных, необходимо изменить наше представление о данных, их местонахождении и владении. Вместо передачи данных из доменов в озеро или платформу, находящуюся в центральном владении, домены должны размещать и поддерживать свои наборы данных в удобной для использования форме. Это означает, что мы можем дублировать данные в разных доменах, поскольку мы преобразуем их в форму, подходящую для использования в этом конкретном домене.

В этом случае наборы данных исходного домена должны быть отделены от внутренних наборов данных исходных систем. Природа наборов данных предметной области сильно отличается от внутренних данных, которые операционные системы используют для своей работы. Они имеют гораздо больший объем, являются неизменно синхронизированными и меняются реже,

чем системы их обработки. По этой причине фактическое базовое хранилище должно подходить для больших данных и быть отделено от существующих операционных баз данных. Наборы данных исходного домена являются наиболее фундаментальными наборами данных и меняются очень редко. При этом наборы данных исходного домена представляют собой необработанные данные на момент создания и не настраиваются или не моделируются для конкретного потребителя.

Платформа данных для конкретной предметной области должна иметь возможность легко восстанавливать эти наборы пользовательских данных из источника.

В этом случае владение наборами данных делегируется с центральной платформы доменам, которые должны обеспечивать очистку, подготовку, агрегацию и обслуживание данных, а также использование конвейера данных. Команды, управляющие доменами, предоставляют возможность обрабатывать свои данные другим специалистам в организации в форме API.

Для этого должен быть реализован безопасный и управляемый глобальный контроль доступа к наборам данных. Это требование является обязательным независимо от того, является ли архитектура централизованной или нет.

Предлагаемая распределенная сеть передачи данных [11] в качестве платформы ориентирована на корпоративные сети, принадлежащие независимым группам, в которых есть инженеры по обработке данных и владельцы данных, использующие общую инфраструктуру данных в качестве платформы для размещения, подготовки и обслуживания своих данных.

2 Типы больших данных

В [3] показано, что решение этой проблемы должно основываться на новой спецификации типов больших данных. В статье предложена методика определения типов Big Data, формирования экосистем (программных стеков) для разных типов данных и обоснована концепция Data Lake.

Рассмотрим подробнее сами данные. Согласно теореме CAP, их можно разделить на 6 классов (рис. 2), однако возможны только 5 классов из 6 потенциальных, потому что PA-класс не может существовать сам по себе, а современные корпоративные архитектуры распределены по умолчанию. Таким образом получаем следующие классы данных.

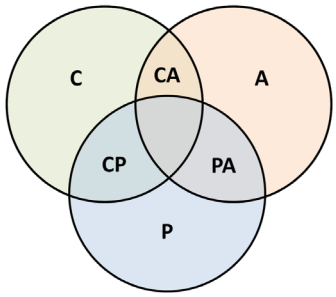

Рис. 2. Графическое представление классов Big Data.

C-класс (согласованность). Характеризуется данными, которые: согласованы - это гарантия того, что одновременное чтение из разных мест предоставит одно и то же значение; то есть система не поддерживает устаревшие или противоречивые данные; хранятся в одном месте (обычно); могут не иметь резервных копий (слишком много данных, чтобы сделать резервную копию); часто это аналитические данные с коротким сроком службы.

A-класс (доступность) Характеризуется данными, которые: всегда должны быть доступны; могут храниться в разных местах; имеют хотя бы одну резервную копию или хотя бы одно другое хранилище; являются важными данными, но не требуют значительного масштабирования.

CA-класс: данные должны быть согласованными и доступными; потенциально монолитная система без возможности масштабирования или масштабирования при условии мгновенного обмена информацией об измененных данных между узлами ведущий-ведомый; нет ограничений по распространению, если для ветвей предусмотрено масштабирование, то каждая ветка работает с относительно независимой базой данных.

Учитывая сказанное, класс CA делится на 3 подкласса:

1. Большие данные больших размеров, которые невозможно представить в структурированном виде или они слишком велики (хранятся в Data Lake или Data Warehouse): данные имеют любой формат и расширение (текст, видео, аудио, изображения, архивы, документы, карты и т. д.); собраны данные целиком, так называемые «сырые данные»; большие данные, которые нецелесообразно помещать в базу данных (неструктурированные данные в случае хранилищ данных); многомерные данные.

Медицинские данные, которые невозможно сохранить в табличной форме (рентген, МРТ, ДНК и т. Д.), являются примером данных этого типа.

2. Данные определенного формата, которые могут быть представлены в структурированной форме (биологические данные, последовательности ДНК и белков, данные о трехмерной структуре, полные геномы и т. д.). Типично это многомерные данные; данные должны быть проанализированы, и их размеры достигают гигантских значений. Примером этого типа являются данные полученные в рамках биоинформатики и медицины, которые необходимо хранить в реляционной таблице с расширениями xml, json и т. д.

3. Другие данные, которые можно хранить в реляционных базах данных, которые: имеют четкую структуру или могут быть представлены в концепции реляционной базы данных; размер хранимых данных не имеет значения (при условии, что в хранилище хранятся объекты небольшого объема или ссылки на область, где хранятся большие объекты).

Примером этого типа являются «сырые» данные, данные клиентов, журналы, клики, статистика погоды или бизнес-аналитика, личные данные, которые обновляются редко, база клиентов и т. д.

CP-класс характеризуется данными, которые: должны быть непротиворечивыми и в то же время есть поддержка распределенного состояния системы, имеющего потенциал масштабирования; структурированы, но легко меняют свою структуру; должны быть представлены в несколько ином формате (график, документ), то есть данные для социальных сетей, географические данные и любые другие данные, которые могут быть представлены в виде графика; имеют сложную структуру, из-за чего существует потенциальная необходимость хранения файлов в документ-ориентированном формате; они очень быстро накапливаются, поэтому необходим распределительный механизм; нет требований к постоянной доступности.

Примером этого типа являются часто записываемая, редко читаемая статистика, а также временные данные (веб-сеансы, блокировки или краткосрочная статистика), хранящиеся во временном хранилище данных или кеше.

PA-класс характеризуется данными, которые: должны быть доступны и в то же время имеется высокая поддержка распределенного состояния системы, имеющего потенциал масштабирования; имеют сложную структуру, потенциальная необходимость хранения файлов в другом формате с возможностью изменения схемы без необходимости переноса всех данных в новую схем, быстро накапливаются.

Этот класс подходит для данных исторического характера. Основная задача здесь - хранение больших объемов данных с потенциальным ростом этой информации каждый день, статистическая и другая обработка информации онлайн и офлайн с целью получения определенной информации (например, об интересах пользователей, настроении в разговорах, для выявления тенденций и т. д.).

Однако прежде чем определять тип системы, мы должны оценить общие параметры системы (максимальное количество пользователей для одновременной работы, возможность масштабирования услуг, наличие персонализированного доступа), оценить проект (наличие собственной мощности сервера, сравнение затрат со стоимостью построения аренды услуг), оценить время доступа к данным, оценить производительность запросов для облачных инфраструктур, построить систему автоматического распределения и отправить запросы в распределенной базе данных.

3 Виртуализация данных

3.1 Общие подходы

В последнее время наблюдается значительный рост спроса на оперативный доступ к данным. Термин «аналитика по запросу» означает очень быстрый процесс принятия бизнес-решений на

основе данных. Между тем, на традиционные процессы преобразования и загрузки данных в результате этого процесса практически не выделяется времени. Ситуация усложняется из-за объема и скорости новых данных, которые появляются со скоростью, превышающей возможности типичных современных корпоративных инфраструктур.

Самый эффективный способ преодолеть эти ограничения - иметь «виртуализированный доступ к данным». Виртуализация данных появилась давно, ее примерами могут служить наборы данных из реляционных баз данных, баз данных NoSQL, платформ больших данных и даже корпоративных приложений, что позволяло создавать логические хранилища данных, к которым можно получить доступ через SQL, REST и т. д. (рис. 3). Такая организация обеспечивает доступ к данным из большого количества распределенных источников и различных форматов, при этом пользователям не требуется знать, где они хранятся. Это избавляет от необходимости перемещать данные или выделять ресурсы для их хранения. Помимо большей эффективности и более быстрого доступа к данным, виртуализация данных может дать необходимую основу для выполнения требований управления данными.

Рис. 3. Виртуализация данных.

Виртуализация реализует 3 функции, которые поддерживают масштабируемость и операционную эффективность, необходимые для сред больших данных:

- разделение: совместное использование ресурсов и переход к потоковой передаче данных;
- изоляция: переход к объектному представлению данных со ссылкой на модель предметной области;
- инкапсуляция: логическое хранилище как единое целое.

Это решение меняет общий подход к данным в абстракции доступа к данным, семантическом хранении, доступе к данным в реальном времени и децентрализованной безопасности (рис. 4).

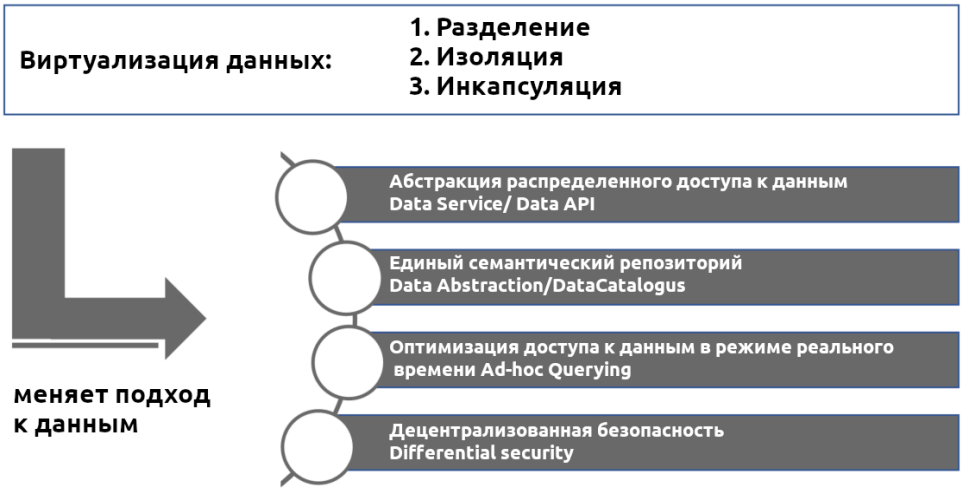

Рис. 4. Особенности виртуализации данных.

Поиск и обучение инженеров по обработке данных может оказаться медленным процессом. Между тем, результаты могут быть неоптимальными, если специалист по данным не понимает требований бизнес-пользователей или не знает, какие методы следует использовать для достижения поставленных целей. Поэтому поставщики часто разрабатывают аналитические продукты, которые позволяют пользователям решать эти проблемы самостоятельно.

Основная проблема заключается в том, что у компаний есть множество различных типов данных в разных форматах, которые расположены на разных системах и серверах. Часть из них находится в облаке, часть может быть расположена в локальных сервисах, и доступ ко всем из них определяется различными политиками и практиками безопасности.

Для обеспечения эффективной работы должен быть способ собрать разрозненные данные со всех ресурсов организации в одном месте и представить их одним точным образом. Чтобы обеспечить унифицированный доступ к данным, организации обычно выполняют процесс, известный как преобразование облачных данных.

Преобразование облачных данных делает данные всех форматов и источников (как облачных, так и локальных) удобочитаемыми и доступными. Однако существует ряд проблем, которые делают преобразование данных долгой, сложной и зачастую дорогостоящей задачей.

Многие поставщики технологий виртуализации данных вынуждают клиентов преобразовывать данные в свой собственный формат, прежде чем их можно будет прочитать и использовать. Однако этот процесс преобразования данных может привести к искажению или потере данных во время этого преобразования. Более того, проприетарные форматы многих поставщиков несовместимы с другими технологиями. Следовательно, вы сталкиваетесь с новыми проблемами непрерывной интеграции из-за привязанности к определенному поставщику. По мере увеличения размеров данных увеличивается объем инженерных работ, необходимых для управления различными источниками данных для быстрого выполнения запросов.

Решением этих проблем является виртуализация данных, которая создаст полную независимость от формата источника данных. Это означает, что данные не нужно каким-либо образом реплицировать или преобразовывать. Вместо того, чтобы полагаться на сложные и трудоемкие процессы преобразования и передачи данных, было бы более эффективно использовать некоторый бизнес-язык, который позволил бы пользователям легко работать с данными.

Следовательно, существует потребность в решении, которое интеллектуально виртуализировало бы все разрозненные данные из различных источников в единое унифицированное представление. Отсюда различные инструменты бизнес-аналитики могут получать быстрые и единые ответы для принятия бизнес-решений. Данные запрашиваются «как есть», но пользователи воспринимают их как единое хранилище данных.

Таким образом, интеллектуальная виртуализация данных – это новая парадигма управления данными. «Умная» виртуализация данных решает проблемы масштабируемости и производительности. Платформы интеллектуальной виртуализации данных позволяют пользователям избежать больших объемов трафика из-за федеративных подключений, которые создают распределенный кеш, оптимизированный для платформы данных. Избегая ненужной передачи данных, интеллектуальная виртуализация данных обеспечивает более стабильную производительность запросов при гораздо меньших требуемых ресурсах.

Между тем, необходимо реализовать работу с разными источниками данных: базами данных отношений (например, Oracle, Teradata, Snowflake), файловыми (CSV, JSON, XML, HDFS, S3), основанными на API (REST, HTML.) и прикладные (Salesforce, Workday, ServiceNow). Это позволит использовать практически любые данные: локальные, облачные, структурированные и неструктурированные, без использования ETL или передачи данных вручную.

В то же время платформа виртуализации должна обеспечивать даже лучшую производительность, чем собственные платформы, с которыми они работают, поскольку уровень виртуализации должен соответствовать или превосходить текущие решения, которые они заменяют.

Мы должны отметить, что, поскольку платформы виртуализации данных являются промежуточным программным обеспечением для аналитических запросов, необходимо, чтобы платформа была интегрирована со структурой безопасности предприятия.

Все вышеперечисленные задачи можно было решить с помощью виртуального суперкомпьютера.

3.2 Виртуализация больших данных

Виртуализация больших данных через логические конструкции и доступ к объектам (сами данные могут храниться в разных источниках, собираться по запросу и / или быть доступны (интерпретированы) в различных триггерных точках (интеграция на основе событий)):

- логическое хранение данных по функциям аналогично традиционному хранению данных, за некоторыми исключениями; для начала, в логическом хранилище данных (LDW) данные не хранятся, в отличие от хранилищ данных, где данные подготавливаются, фильтруются и размещаются;
- логическая абстракция и разделение: разнородные источники данных теперь могут легко взаимодействовать посредством виртуализации данных;
- дифференциальная конфиденциальность (пересекающиеся уровни доступа).

Тот факт, что существует более 2000 программ для работы с различными типами данных, в том числе с большими данными, делает вопрос гибкого хранения очень важным. Хранилища могут быть разных типов, включая порталы, архивы, витрины, базы данных разных типов, облака данных и сети. Они могут иметь синхронные или асинхронные компьютерные соединения. Поскольку тип данных часто заранее неизвестен, существует необходимость в очень гибкой системе хранения, которая позволила бы легко переключаться между различными источниками и системами.

Сочетание виртуального персонального суперкомпьютера с классификацией больших данных, учитывающей разные хранилища, решило бы эту проблему.

3.3 Сети передачи данных и маркетплейсы

Мы рассмотрели несколько важных характеристик современных платформ данных: централизованные, монолитные и с жесткой конвейерной архитектурой, контролируемые группой высокоспециализированных инженеров по данным.

При классификации хранилищ данных следует отметить три основных подхода.

Первый — это проприетарное хранилище данных. Эти хранилища и платформы бизнес-аналитики – очень негибкие и очень дорогие решения. В их использовании участвует небольшая группа специалистов, что приводит к потере потенциала, который это хранилище могло иметь для бизнес-операций.

Вторая – экосистема больших данных. Он содержит озеро данных, управляемое централизованной командой высокоспециализированных инженеров по данным.

Наконец, третий тип – это маркетплейсы (рис. 5). Это решение аналогично первым двум, но ориентировано на потоковую передачу данных и доступ к аналитическим данным в реальном времени. Процессы пакетного и потокового преобразования данных объединяются с помощью таких платформ, как Apache Beam, используются архитектуры Kappa, а также полностью управляемые облачные службы хранения, механизмы конвейера данных и платформы машинного обучения.

Анализ в реальном времени и дорогостоящие инфраструктуры больших данных являются проблемами для первых двух подходов, но не для последнего. Если мы рассмотрим основные проблемы использования централизованной архитектуры платформы данных, следует отметить следующее:

Постоянное появление новых источников данных. Объем доступных данных растет с экспоненциальной скоростью, а способность использовать и согласовывать эти данные под контролем одной платформы пропорционально уменьшается.

Организации стремятся объединить данные по-разному, чтобы отразить изменчивую бизнес-среду и потребности. Это приводит к увеличению числа преобразований, агрегатов, проекций и срезов данных. Время отклика превышает допустимый уровень, что является проблемой даже для современных архитектур платформ данных.

При реализации архитектур платформы данных при определении этапов обработки данных специалисты опираются на прошлые поколения архитектуры. В частности, это видно при структурировании команд, создающих платформу и управляющих ею. В большинстве своем это

высококлассные инженеры по работе с данными, которые разбираются в источниках данных и принципах использования данных для принятия решений. Некоторые специалисты обладают большим техническим опытом, но часто не знают бизнеса и областей применения.

Новой парадигмой архитектур корпоративных платформ данных является децентрализованная сеть передачи данных, поскольку она позволяет успешно решать вышеупомянутые проблемы. Эта парадигма требует изменения понимания данных, их местонахождения и принадлежности.

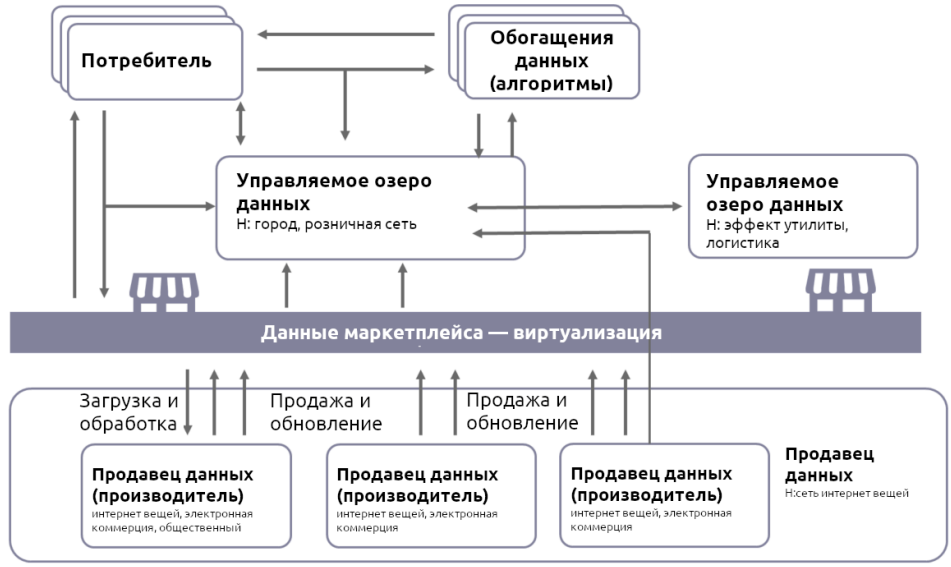

Рис. 5. Маркетплейс данных.

Заголовки разделов третьего уровня нумеруются вручную или автоматически, точка после третьей цифры номера заголовка не ставится.

Вместо переноса данных из доменов в озера или с платформ, находящихся в центральном владении, должен быть более простой способ хранения и обслуживания данных, включая дублирование данных в разных доменах, чтобы обеспечить большую гибкость в их преобразовании.

Недавним примером такой децентрализованной платформы для хранения данных является сеть DGT [11]. Она создает виртуальную сеть данных, соединяя различные источники данных через границы корпоративной информации в единую аналитическую систему, к которой имеют доступ авторизованные пользователи, таким образом, чтобы обеспечить различную конфиденциальность.

Заключение

Таким образом, в настоящее время разработан гибкий и эффективный набор инструментов для цифровизации медицинских организаций. При этом, центральным моментом является использование современных подходов для систем хранения и обработки данных. Их основные особенности рассмотрены в данной статье.

Виртуализация данных — это метод организации доступа к данным, не требующий информации о ее структуре или месте в какой-либо конкретной информационной системе.

Основная цель состоит в том, чтобы упростить доступ и использование данных, превратив их в службу, по существу, сместив парадигму с хранения на эффективное использование.

Ранее задача использования данных решалась за счет интеграции в промежуточную систему хранения. В нем уже были некоторые элементы виртуализации через витрины данных, созданные производителями данных. Теперь в центре внимания оказывается потребитель данных.

Существует три основных характеристики виртуализации, которые поддерживают масштабируемость и операционную эффективность, необходимые для сред больших данных. К ним относятся: разделение на части, то есть разделение ресурсов и переход к потоковым данным; изоляция, которая представляет собой объектно-ориентированный подход к данным с учетом

приложения предметной области; и инкапсуляция, сохраняющая логическое хранилище как единый объект.

Сервисы данных и API изменяют способ доступа к распределенной информации. Абстракция данных формируется из единого семантического репозитория. Оптимизирован доступ к данным в режиме реального времени и специальные запросы. Наконец, достигается дифференцированная безопасность и конфиденциальность. Виртуализация данных — это больше, чем просто современный подход, это совершенно новый способ использования данных.

«Исследования выполнены при финансовой поддержке Минобрнауки России в рамках реализации программы Научного центра мирового уровня по направлению «Передовые цифровые технологии» (соглашение от 16.11.2020 № 075-15-2020-903)

Литература

1. Alexander Bogdanov, Alexander Degtyarev, Vladimir Korkhov, Vladimir Gaiduchok, Ivan Gankevich. Virtual Supercomputer as basis of Scientific Computing // Horizons in Computer Science Research. Volume 11, ch. 5, p. 159 – 198, NOVA Science Publishers, 2015.
2. Alexander Bogdanov, Private cloud vs Personal supercomputer. Distributed computing and GRID technologies in science and education, JINR, Dubna,2012, pp. 57 – 59.
3. Bogdanov, A. V., Shchegoleva, N. L. & Ulitina, I. V., Database Ecosystem Is The Way To Data Lakes, Proceedings of the 27th Symposium on Nuclear Electronics and Computing (NEC 2019). Korenkov, V., Strizh, T., Nechaevskiy, A. & Zaikina, T. (ред.). RWTH Aahen University, pp. 147-152 (CEUR Workshop Proceedings, vol. 2507).
4. Rinku Nemade, Apoorva Nitsure, Poorwa Hirve, Sunil B. Mane R. Nemade, A. Nitsure, P. Hirve and S. B. Mane. Detection of Forgery in Art Paintings using Machine Learning International Journal of Innovative Research in Science, Engineering and Technology, vol. 6, no. 5, 2017.
5. Menon S., Beyer M. Zaidi E., Jain A. Market Guide for Data Virtualization. Published: 16 November 2018, ID: G00340606 https://www.gartner.com/en/documents/3893219/market-guide-for-data-virtualization
6. Ivan Gankevich, Vladimir Korkhov, Serob Balyan, Vladimir Gaiduchok, Dmitry Gushchanskiy, Yuri Tipikin, Alexander Degtyarev, Alexander Bogdanov Constructing Virtual Private Supercomputer Using Virtualization and Cloud Technologies // Lecture Notes in Computer Science, 2014. Vol. 8584, p. 341-354
7. Alexander Bogdanov, Alexander Degtyarev, and Vladimir Korkhov. New Approach to the Simulation of Complex Systems. EPJ Web of Conferences, vol. 108, 01002, 2016, pp. 1 – 12.
8. Alexander Bogdanov, Alexander Degtyarev, and Vladimir Korkhov. Desktop Supercomputer: What Can It Do? - ISSN 1547-4771, Physics of Particles and Nuclei Letters, 2017, Vol. 14, No. 7, pp. 985–992. © Pleiades Publishing, Ltd., 2017.
9. Vladimir Korkhov, Sergey Kobyshev, Alexander Degtyarev, Alexander Bogdanov. Light-Weight Cloud-Based Virtual Computing Infrastructure for Distributed Applications and Hadoop Clusters. - ICCSA: International Conference on Computational Science and Its Applications, 17th International Conference, Trieste, Italy, July 3-6, 2017, Proceedings, Part V, pp. 399 – 411.
10. Zhamak D. How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh/ URL: http:// https://martinfowler.com/articles/data-monolith-to-mesh.html
11. DGT, the Decentralized Enterprise Platform. URL: http://dgt.world/

DIGITALIZATION OF HEALTH CARE: WHAT CAN BE DONE NOW

Bogdanov, Alexander Vladimirovich

D. Sc., professor
St. Petersburg State University, Department of fundamental informatics and distributed systems, professor
Research and analytical journal "Information Society", member of the Editorial board
Saint-Petersburg, Russiam Federation
Author's e-mail (используйте стиль «Адрес автора»)

Zalutskaya, Natalia Mikhailowna

Candidate of medical sciences, associate professor
Federal state budgetary institution "Bekhterev National Medical Research Psychiatry and Neurology Center", Ministry of Health of the Russian Federation, leading researcher
Saint-Petersburg, Russiam Federation
nzalutskaya@yandex.ru

Schegoleva, Nadezhda Lvovna

D. Sc., Professor
St. Petersburg State University, Department of fundamental informatics and distributed systems, professor
Saint-Petersburg, Russiam Federation
n.shchegoleva@spbu.ru

Zaynalov, Nodir Rasulovich

Candidate of physics and mathematics, associate professor
Samarkand branch of Tashkent University of Information Technologies, head of Department of information security
Samarkand, Uzbekistan
nodirz@mail.ru

Kiyamov, Jasur Utkirovich

St. Petersburg State University, Department of fundamental informatics and distributed systems, Ph.D. student
Saint-Petersburg, Russiam Federation
st080634@student.spbu.ru

Dik, Aleksander Gennadievich

St. Petersburg State University, Department of fundamental informatics and distributed systems, Ph.D. student
Saint-Petersburg, Russiam Federation
st087383@student.spbu.ru

Abstract

Life sciences research is an excellent example of the generation of very large amounts of data of various types, which cannot be handled by simple means. The fact that there are more than 2,000 programs for working with various types of data, including big data, makes the task of processing them extremely difficult even for large federal centers. The root of the problem lies in the need for preliminary work with data and two operations – consolidation and virtualization. Repositories can be of various types, including portals, archives, storefronts, databases of various kinds, data clouds, and networks. They may have synchronous or asynchronous computer connections. Since the data type is often not known in advance, there is a need for a very flexible storage system that would allow easy switching between different sources and systems. The combination of the concept of a virtual personal supercomputer with the classification of big data, which takes into account various storage schemes, allows us to solve this problem.

Keywords

big data, data virtualization, virtual personal supercomputer, data network, data markets

References

1. Alexander Bogdanov, Alexander Degtyarev, Vladimir Korkhov, Vladimir Gaiduchok, Ivan Gankevich. Virtual Supercomputer as basis of Scientific Computing // Horizons in Computer Science Research. Volume 11, ch. 5, p. 159 – 198, NOVA Science Publishers, 2015.
2. Alexander Bogdanov, Private cloud vs Personal supercomputer. Distributed computing and GRID technologies in science and education, JINR, Dubna,2012, pp. 57 – 59.
3. Bogdanov, A. V., Shchegoleva, N. L. & Ulitina, I. V., Database Ecosystem Is The Way To Data Lakes, Proceedings of the 27th Symposium on Nuclear Electronics and Computing (NEC 2019). Korenkov, V., Strizh, T., Nechaevskiy, A. & Zaikina, T. (ed.). RWTH Aahen University, pp. 147-152 (CEUR Workshop Proceedings, vol. 2507).
4. Rinku Nemade, Apoorva Nitsure, Poorwa Hirve, Sunil B. Mane R. Nemade, A. Nitsure, P. Hirve and S. B. Mane. Detection of Forgery in Art Paintings using Machine Learning International Journal of Innovative Research in Science, Engineering and Technology, vol. 6, no. 5, 2017.
5. Menon S., Beyer M. Zaidi E., Jain A. Market Guide for Data Virtualization. Published: 16 November 2018, ID: G00340606 https://www.gartner.com/en/documents/3893219/market-guide-for-data-virtualization
6. Ivan Gankevich, Vladimir Korkhov, Serob Balyan, Vladimir Gaiduchok, Dmitry Gushchanskiy, Yuri Tipikin, Alexander Degtyarev, Alexander Bogdanov Constructing Virtual Private Supercomputer Using Virtualization and Cloud Technologies // Lecture Notes in Computer Science, 2014. Vol. 8584, p. 341-354
7. Alexander Bogdanov, Alexander Degtyarev, and Vladimir Korkhov. New Approach to the Simulation of Complex Systems. EPJ Web of Conferences, vol. 108, 01002, 2016, pp. 1–12.
8. Alexander Bogdanov, Alexander Degtyarev, and Vladimir Korkhov. Desktop Supercomputer: What Can It Do? - ISSN 1547-4771, Physics of Particles and Nuclei Letters, 2017, Vol. 14, No. 7, pp. 985–992. © Pleiades Publishing, Ltd., 2017.
9. Vladimir Korkhov, Sergey Kobyshev, Alexander Degtyarev, Alexander Bogdanov. Light-Weight Cloud-Based Virtual Computing Infrastructure for Distributed Applications and Hadoop Clusters. - ICCSA: International Conference on Computational Science and Its Applications, 17th International Conference, Trieste, Italy, July 3-6, 2017, Proceedings, Part V, pp. 399 – 411.
10. Zhamak D. How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh/ URL: http:// https://martinfowler.com/articles/data-monolith-to-mesh.html
11. DGT, the Decentralized Enterprise Platform. URL: http://dgt.world/