УДК 004.032

ПРИМЕНЕНИЕ ДАТАЦЕНТРИЗМА В РАСПРЕДЕЛЁННЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Шатилов А.А., Шмелёв А.В.

ФГБОУ ВО «МИРЭА – Российский технологический университет», Москва, e-mail: tolya870@yandex.ru, opuhla@gmail.com

При проектировании, разработке и развитии программного обеспечения и информационных систем зачастую полученные программные продукты являются обособленными относительно друг друга и систем заказчика или потребителя ввиду отсутствия первоначальных требований в возможности дальнейшей их интеграции с другими сервисами и программами (как существующими, так и будущими). Это приводит к повышению требований к объёмам вычислительных мощностей или/и к дополнительным накладным расходам на развитие информационной инфраструктуры. Это обусловлено либо необходимостью в реализации дополнительного промежуточного программного слоя для оказания информационной совместимости между данными разных видов и форматов, либо необходимостью производить регулярные кардинальные изменения многих уже имеющихся программных компонентов, затраты на модернизацию которых с каждым витком будут увеличиваться вместе с размерами самой системы. Однако с помощью датацентричного подхода эту проблему можно решить. В этой статье описаны преимущества и недостатки датацентричного подхода проектирования архитектуры информационных систем, а также описаны преимущества внедрения датацентризма поверх распределённых архитектур информационных систем, а также какие характеристики получит система, построенная с применением принципов обоих архитектурных подходов.

Ключевые слова: датацентризм, распределённые информационные системы, архитектуры информационных систем

APPLICATION OF DATA-CENTRISM IN DISTRIBUTED INFORMATION SYSTEMS

Shatilov A.A., Shmelev A.V.

MIREA – Russian Technological University, Moscow, e-mail: tolya870@yandex.ru, opuhla@gmail.com

When designing, developing and rising software and information systems, often the resulting program products are isolated from each other and from the customer's or consumer's systems due to the lack of initial requirements for the possibility of their further integration with other services and software (both existing and future ones). This leads to increased requirements for the amount of computing power and/or to additional overhead costs for the development of information infrastructure. This is due either to the need to implement an additional intermediate software layer to provide information compatibility between data of different types and formats, or to make regular cardinal changes to many existing software components, which cost of upgrading will increase with each turn along with the size of the system itself. However, this problem can be solved with using a data-centric approach. This article describes the advantages and disadvantages of the data-centric approach to designing the architecture of information systems, and also describes the advantages of implementing data-centrism on top of distributed architectures of information systems, as well as what characteristics a system will get when using the principles of both architectural approaches.

Keywords: data-centric, distributed information systems, information systems architecture

Развитие ИТ (информационных технологий), а также растущая популярность так называемой «науки о данных» (data science) связана с эффективностью деятельности организаций, которая только возрастает при грамотных сборе, обработке и применении собранного большого количества данных [1] в связи с современными требованиями бизнеса в оперативной реакции на большие потоки разнообразной информации. Большое количество собранных данных стало носить название «большие данные» (big data) [2] и иметь отдельную ценность в зависимости от деятельности организации, собирающей, обрабатывающей и применяющей на практике эти данные [3]. На фоне этого устоявшиеся методы разработки информационных систем становятся неэффективны в рамках развития информационных систем и приводят к громоздким программным решениям.

Цель исследования – провести обзор, характеристику и сравнение традиционного (приложение-ориентированного) и датацентричного подхода при проектировании информационных систем, а также рассмотреть применение датацентризма в рамках распределённых систем и в сфере больших данных.

Материалы и методы исследования

Обзор и анализ открытых источников (в том числе на иностранном языке), сравнение представленных подходов проектирования информационных систем.

Результаты исследования и их обсуждение

На фоне этого одни и те же данные могут иметь разный вид и формат, быть продублированы в разных программных приложениях, выполняющих разные задачи по их обработке. Это приводит к сильному

разделению информации по её разным характеристикам: по характеру выполняемых действий над ними, невзирая на их содержимое, по форме и месту их записи, по типу исходников. Это попутно усложняет ИТ-инфраструктуру организаций, повышая издержки на её обслуживание.

В итоге это приводит к пониженной эффективности систем хранения данных, систем и приложений, а также повышенной стоимости разработки и поддержки программных и аппаратных средств сбора и обработки информации из-за возникновения копий одних и тех же данных в разных формах и/или форматах, направленных на использование в различных программных системах.

Возникает это в первую очередь из-за «традиционного» подхода к проектированию информационных систем, где в центре внимания находится само программное приложение – «приложение центризм» (Application Centric) [4]. Другими словами, проектирование ведётся обособленно от уже имеющихся данных и других информационных систем, начиная с программного кода и заканчивая формированием формата и формы данных, а также организация их хранения для конкретной программы.

Также для работы программ в условиях высокой нагрузки применяются методы проектирования распределённых информационных систем [5], которые способствуют более простому и менее затратному её дальнейшему масштабированию в рамках выполняемых ими задач [6]. Но, когда возникает необходимость в создании дополнительной подсистемы или выполнении специфической работы над собранными данными, полученными исходной информационной системой, возникает необходимость в создании отдельных методов подготовки и переноса данных из существующей системы в новую (другими словами – организовать слой для конвертации данных из системы А в систему Б) посредством отдельного прикладного программного обеспечения, созданного исключительно для этих целей (рис. 1), что требует выделения отдельных ресурсов.

Помимо наличия издержек на расширение функционала за счёт промежуточного прикладного программного обеспечения может возникнуть ситуация, в которой может происходить снижение производительности исходной информационной системы. Это может происходить, если полученные в процессе конвертации данные физически будут находиться и управляться в системе управления базами данных исходной системы, создавая две копии одних и тех же данных, рассчитанные на обработку в разных системах – связано это в первую очередь с разделением аппаратных ресурсов базы данных на обслуживание данных нескольких систем.

Эту проблему можно решить посредством «стандартизации» данных, то есть приведения аккумулированной информации в единый для организации избыточный формат (или приспособленной для масштабирования схемы данных). Благодаря этому компоненты разных информационных систем смогут беспрепятственно оперировать одними и теми же наборами данных, сокращая расходы вычислительных ресурсов за счёт исключения и сокращения процессов копирования, конвертации и переноса (в случае разных хранилищ данных для разных систем) информации для её последующей передачи компонентам систем, а также потенциально даёт возможность сократить издержки, связанные с необходимостью дальнейшего расширения систем хранения данных.

Датацентризм – это архитектурный подход к построению информационных систем, где центральным и главным звеном являются данные, без которых невозможно обеспечить результативность программных систем и отдельных приложений [7]. На уровне архитектуры датацентризм выражается в наличии центральной системы хранения и управления данных в качестве главного компонента информационной системы.

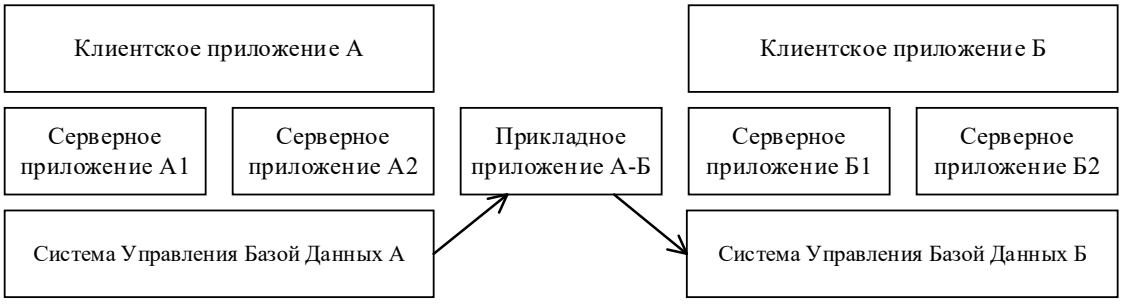

Рис. 1. Появление второстепенной информационной системы на базе собранных данных исходной системы

Сравнение подходов [8]

Application-Centric (приложение-ориентированный подход)	Data-Centric (датацентричный подход)
Высокая стоимость изменений	Адекватная стоимость изменений
Данные завязаны на приложении, так как оно владеет ими	Данные – это открытый ресурс, который переживет любое данное приложение
Каждый новый проект сопровождается проектом преобразования больших данных	Каждый новый проект использует существующие хранилища данных
Данные существуют в большом разнообразии разнородных форматов, структур, значений и терминологии	Глобально интегрированные данные имеют общий смысл, экспортируются из общего источника в любой подходящий формат
На интеграцию данных уходит 35–65% ИТ-бюджета	Внедрение данных будет практически бесплатным
Трудно или невозможно интегрировать внешние данные с внутренними данными	Внутренние и внешние данные легко интегрируются

На начальных стадиях проектирования ИС, когда ещё нет чётких требований к конечной системе, необходимо определить, по какому принципу следует стремиться производить разработку. В таблице приведено сравнение приложение-ориентированного и дата-ориентированного подходов – исходя из них можно сделать вывод, что датацентризм актуален в случаях, когда информационная система вынуждена оперировать данными (или они потенциально имеют перспективу перехода в категорию «больших данных») и имеет потенциал к дальнейшему развитию и расширению, что потенциально сокращает издержки на внедрение и поддержку новых версий компонентов, информационных сервисов и отдельных подсистем и приложений.

Датацентричный подход позволяет одновременно стандартизировать, упорядочить и снизить избыточность хранимой информации за счёт её централизованного хранения [9] – благодаря этому расход ресурсов на хранение данных может существенно сократиться, позволяя сократить издержки на развитие существующих и создание новых программных компонентов без необходимости в создании дополнительных промежуточных слоёв, а также организовать взаимосвязанность подсистем между собой.

Однако на фоне растущей популярности направления «науки о данных» в связи с накоплением больших массивов информации и, как следствие, переходом от простых массивов данных к «big data» (большим данным) в датацентричном подходе возникает проблема производительности – при растущей нагрузке на единую систему хранения данных страдать от низкой производительности начинают все приложения, которые опираются на один и тот же большой набор данных.

Эта проблема приводит к необходимости использования методов построения распределённых информационных систем в рамках системы хранения данных, что может привести к отклонению от датацентризма и последующей фрагментации, избыточности и усложнения потоков данных.

Поэтому имеет смысл одновременное совместное применение методов проектирования распределённых информационных систем и датацентрического архитектурного подхода – это одновременно поможет поддержать возможности масштабирования информационных систем и их компонентов, обеспечить низкую избыточность и расхождение данных между разными программными приложениями, а также не отойти от самого датацентризма во время масштабирования отдельных частей системы.

Таким образом, предлагается объединить положительные аспекты датацентризма и распределённых систем.

Принципы построения систем на базе датацентризма таковы [10]:

− основа архитектуры – данные, а не приложения;

− каждый объект данных должен быть представлен только один раз и быть уникальным;

− представление каждого объекта должно содержать все возможные точки зрения на него, в явном виде выделяя как его общие признаки, так и уникальные;

− структура данных должна следовать структуре концептуальных представлений о предметной области;

− приложения и отдельные компоненты информационных систем не имеют собственных хранилищ данных и представляют собой не монолитные решения, а сервисы, предназначенные для решения конкретных задач над данными;

– приложения должны быть готовы к изменению структуры данных и не зависеть от неё;

– как можно больше логики алгоритмов должно быть внесено в онтологическую модель данных, что позволит настраивать алгоритмы обработки данных одновременно с изменением их структуры.

Основные принципы построения распределённых информационных систем [11, 12]:

– Прозрачность – способность системы скрыть от пользователя факт того, что система является распределённой.

– Открытость – определяется как полнота и ясность описания интерфейсов работы с системой и службами, которые она предоставляет через эти интерфейсы, давая возможность переноса системы на другое аппаратное обеспечение, а также расширения функциональности за счёт добавления новых компонентов.

– Масштабируемость – это зависимость изменения характеристик системы от количества ее пользователей и подключенных дополнительных ресурсов, а также от степени географической распределённости системы.

На рис. 2 показана схема результата совмещения методов построения распределённых систем с датацентричным архитектурным подходом – уровень данных представляет собой горизонтально распределенную подсистему, обеспечивающую сбор, записи изменений и предоставления одинаковых или смежных данных из единого массива данных разным приложениям и компонентам информационной системы, которые могут выполнять различные функции по работе с данными, но записывать результаты работы в одно распределённое хранилище.

В зависимости от назначения системы аккумулируемые в ней данные могут перерасти в так называемые «большие данные», что потребует дополнительных мер по поддержанию быстродействия системы и её работоспособности в целом.

Так, для аккумулирования результатов обработки данных следует иметь достаточно ресурсов [13] в системе хранения данных, чтобы обеспечить:

– место хранения обработанной информации;

– высокоскоростной доступ к исходным данным.

Сами большие данные измеряются следующими свойствами [14]:

– *Объём*: это могут быть данные разных источников – каналы в социальных сетях, данные посещаемости веб-ресурсов, данные мобильных приложений, сетевой трафик, данные датчиков и т.д. В некоторые компании могут поступать десятки терабайт данных, в другие – сотни петабайт.

– *Скорость*: это скорость приёма данных и, возможно, действий на их основе. Некоторые «умные» продукты, функционирующие на основе интернета, работают в режиме реального или практически реального времени. Соответственно, такие данные требуют моментальной оценки и действий.

– *Разнообразие*: разнообразие означает, что доступные данные принадлежат к разным типам. Традиционные типы данных структурированы и могут быть сразу сохранены в реляционной базе данных. С появлением больших данных данные стали поступать в неструктурированном виде. Такие неструктурированные и полуструктурированные типы данных, как текст, аудио и видео, требуют дополнительной обработки для определения их значения и поддержки метаданных.

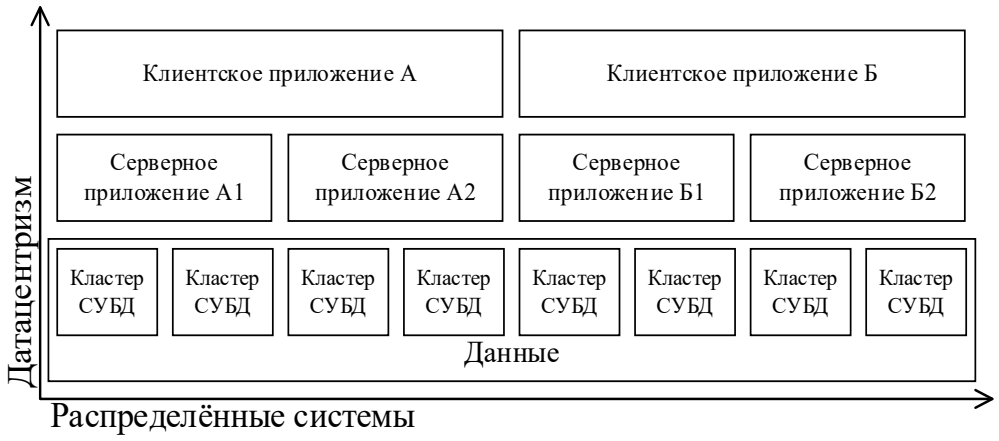

Рис. 2. Схема одновременно распределённой и датацентричной архитектуры информационной системы

Исходя из возможных свойств больших данных, системы хранения данных в датацентризме требуют не просто большого объёма памяти, но и много вычислительных ресурсов, чтобы успешно обеспечивать сбор, обработку, выдачу и конвертацию данных в режиме реального или почти реального времени, а также обеспечивать работоспособность уровня данных, который помимо обеспечения сбора и доступа также выполняет функции разграничения доступа приложениям к отдельным сегментам накопленной информации.

Датацентричный подход к проектированию информационных систем актуален, когда применяются большие динамически изменяющиеся массивы данных или/и есть потенциал для развития системы, расширяя её возможности в рамках работы с данными (включая расширения самих данных).

Датацентричный подход не является актуальным, когда информационная система не подразумевает под собой работы с динамически изменяющимися данными.

Заключение

Таким образом, построенная распределённая информационная система с применением датацентризма позволит сократить издержки на аппаратное обеспечение, снизить стоимость масштабирования и расширения функционала систем, а также, при внедрении датацентризма в уже существующую систему, позволит стандартизировать и упорядочить уже собранные наборы больших данных для последующей их обработки посредством функционала исходной информационной системы, дополнительными сервисами и новыми функциями дополненной исходной системы.

Помимо этого также можно повысить эффективность работы существующих систем для обработки больших данных (а также потенциально подготовить те системы, которые пока ещё не оперируют большим объёмом информации, которые можно назвать «большими данными»), тем самым повысив эффективность деятельности в сфере науки о данных (которая занимается вопросами больших данных, их методами обработки и достижения максимальной эффективности их применения) [1], а также в целом достичь высокой эффективности применения собранной информации в рамках деятельности организаций и их партнёров.

Список литературы

1. Идигова Л.М., Абубакаров А.Х. Datascience как новый тренд. Исследование методов работы с большим объемом данных в организации // Влияние новой геополитической реальности на государственное управление и развитие Российской Федерации: материалы II Всероссийской научно-практической конференции (Грозный, 20–21 сентября 2019 г.) / Под ред. З.А. Саидова. Грозный: Чеченский государственный университет, 2019. С. 275–280.

2. Big Data Explained [Электронный ресурс]. URL: https://www.mongodb.com/big-data-explained (дата обращения: 01.05.2022).

3. Величко Н.А., Митрейкин И.П. Технология Big Data. Анализ рынка Big Data // Синергия Наук. 2018. № 30. С. 937–943.

4. The Data-Centric Revolution [Электронный ресурс]. URL: https://tdan.com/the-data-centric-revolution/18780 (дата обращения: 01.05.2022).

5. Обзор современных распределенных систем [Электронный ресурс]. URL: https://scienceforum.ru/2020/article/2018023457 (дата обращения: 01.05.2022).

6. Страх и ненависть в распределённых системах [Электронный ресурс]. URL: https://habr.com/ru/post/322876/ (дата обращения: 01.05.2022).

7. Шатилов А.А. Разработка программной архитектуры мобильного приложения «Социальный будильник» // Всероссийская научная конференция молодых исследователей с международным участием «Инновационное развитие техники и технологий в промышленности (ИНТЕКС-2021)». М.: ФГБОУ ВО «РГУ им. А.Н. Косыгина», 2021. С. 111–115.

8. The Data-Centric Manifesto [Электронный ресурс]. URL: http://datacentricmanifesto.org/principles (дата обращения: 30.05.2022).

9. Data-centric Architecture – A Different Way of Thinking [Электронный ресурс]. URL: https://www.vistaprojects.com/blog/data-centric-architecture (дата обращения: 01.05.2022).

10. Три шага к датацентричной архитектуре [Электронный ресурс]. URL: https://www.osp.ru/os/2019/04/13055224 (дата обращения: 01.05.2022).

11. Лекция 12: Компонентные технологии и разработка распределенного ПО [Электронный ресурс]. URL: https://intuit.ru/studies/higher_education/3406/courses/64/lecture/1888 (дата обращения: 01.05.2022).

12. Построение распределенных систем обработки информации [Электронный ресурс]. URL: https://spravochnaya.com/7688_postroenie-raspredelennyh-sistem-obrabotki-informacii.html (дата обращения: 01.05.2022).

13. VK Cloud Solutions. Что такое big data: зачем нужны большие данные, как их собирают и обрабатывают. [Электронный ресурс]. URL: https://mcs.mail.ru/blog/big-data-vse-govoryat-no-malo-kto-shchupal (дата обращения: 31.05.2022).

14. Oracle Cloud Infrastructure. Что такое большие данные? [Электронный ресурс]. URL: (дата обращения: 31.05.2022).