ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р 70466— 2022/ ISO/IEC TR 20547-1: 2020

Информационные технологии

ЭТАЛОННАЯ АРХИТЕКТУРА БОЛЬШИХ ДАННЫХ

Часть 1

Структура и процесс применения

(ISO/IEC TR 20547-1:2020, IDT)

Издание официальное

Москва
Российский институт стандартизации
2022

Предисловие

1 ПОДГОТОВЛЕН Научно-образовательным центром компетенций в области цифровой экономики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В. Ломоносова» (МГУ имени М.В. Ломоносова) и Автономной некоммерческой организацией «Институт развития информационного общества» (ИРИО) на основе собственного перевода на русский язык англоязычной версии документа, указанного в пункте 4

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164 «Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 8 ноября 2022 г. № 1257-ст

4 Настоящий стандарт идентичен международному документу ISO/IEC TR 20547-1:2020 «Информационные технологии. Эталонная архитектура больших данных. Часть 1. Структура и процесс применения» (ISO/IEC TR 20547-1:2020 Information technology — Big data reference architecture — Part 1: Framework and application process, IDT).

При применении настоящего стандарта рекомендуется использовать вместо ссылочных международных стандартов соответствующие им национальные стандарты, сведения о которых приведены в дополнительном приложении ДА

5 ВВЕДЕН ВПЕРВЫЕ

Правила применения настоящего стандарта установлены в статье 26 Федерального закона от 29 июня 2015 г. № 162-ФЗ «О стандартизации в Российской Федерации». Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок — в ежемесячном информационном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске ежемесячного информационного указателя «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.rst.gov.ru)

Содержание

Введение

Парадигма больших данных относится к быстро развивающейся предметной области со стремительно меняющимися технологиями. Эта динамическая ситуация порождает две существенные проблемы для потенциальных разработчиков технологий. Первая проблема состоит в том, что не хватает стандартных определений терминов, включая ключевое понятие больших данных. Вторая заключается в том, что нет единого подхода к описанию архитектуры больших данных и ее реализации. Первая проблема разрешена в стандарте ИСО/МЭК 20546. Разработка серии стандартов ИСО/МЭК 20547 нацелена на разрешение второй проблемы и представление описания структуры и эталонной архитектуры, которые могут применяться организациями в своих предметных областях для эффективного и последовательного описания архитектуры и ее реализации с учетом лежащих в ее основе технологических решений, а также ролей/исполнителей и связанных с ними интересов (проблем). В настоящем стандарте описаны структура эталонной архитектуры, процесс отображения требований/вариантов использования в эталонной архитектуре, а также оценка этого отображения.

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Информационные технологии

ЭТАЛОННАЯ АРХИТЕКТУРА БОЛЬШИХ ДАННЫХ

Часть 1

Структура и процесс применения

Information technology. Big data reference architecture.
Part 1. Framework and application process

Дата введения — 2023—03—01

1 Область применения

Настоящий стандарт содержит описание структуры эталонной архитектуры больших данных и процесса применения стандарта в рассматриваемой предметной области.

2 Нормативные ссылки

В настоящем стандарте использована нормативная ссылка на следующий стандарт [для датированных ссылок применяют только указанное издание ссылочного стандарта, для недатированных — последнее издание (включая все изменения)]:

ISO/IEC/IEEE 42010:2011, Systems and software engineering — Architecture description (Системная и программная инженерия. Описание архитектуры).

3 Термины и определения

В настоящем стандарте применены следующие термины с соответствующими определениями.

ИСО и МЭК поддерживают терминологические базы данных для использования в стандартизации, расположенные по следующим адресам:

- платформа ИСО для онлайн-просмотра материалов по стандартам (Online Browsing Platform, OBP) доступна по адресу https://www.iso.org/obp/ui;
- Электропедия МЭК (IEC Electropedia) доступна по адресу http://www.electropedia.org/.

3.1 **большие данные** (big data): Большие массивы данных, отличающиеся главным образом такими характеристиками, как объем, разнообразие, скорость обработки и/или вариативность, которые требуют использования технологии масштабирования для эффективного хранения, обработки, управления и анализа.

Примечание — Большие данные повсеместно используются множеством различных способов, например в качестве названия технологии масштабирования, используемой для обработки обширных массивов данных.

[ИСО/МЭК 20546:2019, 3.1.2]

3.2 **эталонная архитектура** (reference architecture): В сфере архитектуры программного обеспечения или архитектуры предприятия определяемое понятие устанавливает проверенное типовое решение для архитектуры определенной предметной области, а также задает словарь общепринятых понятий для обсуждения реализаций этой архитектуры.

[ISO TR 14639-2:2014, 2.65]

Издание официальное

3.3 **структура** (framework): Определенный набор утверждений (концепций, понятий) или идей для описания сценария или решения задачи.

[ИСО/МЭК 15638-6—2014, 4.30]

3.4 **защищенность** (security): защита от преднамеренной подрывной деятельности или отказа. Соединение четырех атрибутов — конфиденциальности, целостности, доступности и подотчетности, и аспектов пятого атрибута — практичности, у каждого из которых имеется соответствующий источник обеспечения.

[ISO/IEC/IEEE 15288:2015, 4.1.39]

3.5 **конфиденциальность персональных данных** (privacy): право отдельных лиц контролировать или влиять на то, какая информация, связанная с ними (персональные данные), подлежит сбору и хранению, а также кем эта информация может быть раскрыта.

[ИСО/МЭК 26927:2011, 3.34]

3.6 **происхождение** (provenance): Сведения о месте и времени появления, извлечения или создания ресурса, записи, доказательства подлинности или принадлежности в прошлом.

[ИСО/МЭК 11179-7:2019, 3.1.10]

3.7 **SQL:** Язык баз данных, описанный в ИСО/МЭК 9075.

Примечание — Аббревиатура SQL иногда расшифровывается как «язык структурированных запросов» (Structured Query Language), но это название не используется в серии стандартов ИСО/МЭК 9075.

3.8 **жизненный цикл** (life cycle): Развитие системы, продукта, услуги, проекта или другой создаваемой человеком сущности от замысла до списания.

[ISO/IEC/IEEE 15288:2015, 4.1.19]

4 Сокращения

BDA — аудитор больших данных (Big Data Auditor);
BDAcP — сервис-провайдер доступа к большим данным (Big Data Access Provider);
BDAnP — сервис-провайдер аналитики больших данных (Big Data Analytics Provider);
BDAP — сервис-провайдер приложения больших данных (Big Data Application Provider);
BDCP — сервис-провайдер сбора коллекций больших данных (Big Data Collection Provider);
BDFP — сервис-провайдер среды обработки больших данных (Big Data Framework Provider);
BDIP — сервис-провайдер инфраструктуры больших данных (Big Data Infrastructure Provider);
BDPlaP — сервис-провайдер платформы больших данных (Big Data Platform Provider);
BDPreP — сервис-провайдер предобработки больших данных (Big Data Preparation Provider);
BDProP — сервис-провайдер обработки больших данных (Big Data Processing Provider);
BDRA — эталонная архитектура больших данных (Big Data Reference Architecture);
BDSD — разработчик сервиса больших данных (Big Data Service Developer);
BDSO — оркестратор системы больших данных (Big Data System Orchestrator);
BDSP — партнер сервиса больших данных (Big Data Service Partner);
BDVP — сервис-провайдер визуализации больших данных (Big Data Visualization Provider);
GDPR — Общий регламент по защите данных (General Data Protection Regulation);
JSON — обозначение объектов и правил JavaScript (JavaScript Object Notation);
RDF — структура описания ресурсов (Resource Description Framework);
SQuaRE — требования и оценка качества систем и программного обеспечения (Systems and software Quality Requirements and Evaluation);
XML — расширяемый язык разметки (Extensible Markup Language).

5 Обзор документа

Настоящий стандарт предназначен для ознакомления с понятиями в сфере эталонной архитектуры больших данных в целях применения других стандартов серии стандартов ИСО/МЭК 20547 к конкретным системам и наборам задач.

Разделы с 6-го по 9-й включают:

- описание мотивации и целей стандартизации больших данных;
- введение в эталонные архитектуры и сведения об их назначении;

- обзор эталонной архитектуры больших данных и объяснение ее ключевых понятий;
- описание процесса применения эталонной архитектуры больших данных.

При использовании серии стандартов ИСО/МЭК 20547 настоящий стандарт будет полезен в следующих случаях:

- для получения общего представления о применении эталонной архитектуры больших данных необходимо использовать содержание разделов 5—7;
- для разработки архитектуры больших данных и приведения ее в соответствие с эталонной архитектурой необходимо использовать описание процесса, рассмотренного в разделе 8.

6 Стандартизация больших данных: мотивация и цели

В отчете за 2019 г. международная исследовательская и консалтинговая компания International Data Corporation (IDC) прогнозировала мировые доходы от использования и аналитики больших данных в размере 189,1 млрд долл. США, что на 12 % больше, чем в 2018 г., а также совокупный ежегодный рост за 5 лет на 13,2 % с доходами, превышающими 274,3 млрд долл. США в 2022 г. [15].

Покупатели и потенциальные пользователи систем больших данных вынуждены иметь дело со взрывным ростом областей применения новых технологий в условиях, когда определение и понимание термина «большие данные» еще не устоялось. Для того чтобы заинтересованные стороны понимали, что они покупают и внедряют, необходимы четко выстроенные процессы их взаимодействия с потенциальными поставщиками технологий и услуг.

Примечания

1 Понятие «система больших данных» предусматривает использование парадигмы и инженерии больших данных.

2 Понятие «инженерия больших данных» предусматривает перспективные способы использования независимых ресурсов для построения масштабируемых систем данных в тех случаях, когда требуется создание новых архитектур для эффективного хранения, обработки и анализа с учетом характеристик массивов данных.

3 Понятие «парадигма больших данных» предусматривает распределение систем данных по горизонтально связанным независимым ресурсам для обеспечения масштабируемости в целях эффективной обработки больших массивов данных.

Потенциальная ценность результатов анализа больших данных стимулирует внедрение систем больших данных в организациях, поэтому необходимо понимать возможные проблемы и ответственность, связанные с их контролем и управлением. По оценкам компании IDC, предприятия имеют обязательства или несут ответственность почти за 80 % информации в цифровом пространстве и должны быть готовы к решению задач обеспечения ее достоверности, авторского права и конфиденциальности персональных данных. Кроме того, по оценке компании IDC, по состоянию на 2020 г. более 40 % данных в цифровом пространстве требуют обеспечения надежной защиты, а объем этих данных растет быстрее, чем все цифровое пространство [15]. Возникающие риски означают, что организации должны иметь возможность идентифицировать угрозы, определять и формулировать политики безопасности, выявлять источники данных и решать задачи по их управлению, а также внедрять технические средства контроля и документировать их применение для обеспечения реализации этих политик с целью ограничения ответственности организации при неконтролируемом использовании данных, которыми она управляет.

Наконец, очень немногие организации, имеющие дело с большими данными, работают исключительно с собственными данными. Это означает, что системы, с помощью которых решаются задачи сбора и анализа больших данных, должны иметь возможность безопасного обмена данными и надежного взаимодействия. Фактически передача огромного объема больших данных между системами часто становится нецелесообразной, что во многих случаях обуславливает необходимость применения аналитических инструментов на уровнях интероперабельности данных, программного обеспечения и приложений.

Изучение существующего ландшафта больших данных, рыночных требований к стандартизации области больших данных позволило определить следующие приоритеты:

a) сценарии использования больших данных, определения, словари и эталонные архитектуры (например, система, данные, платформы, онлайн/офлайн и т. д.);

b) спецификации и стандартизация метаданных, включая их источники;

c) прикладные модели (например, пакетной обработки, потоковые и т. д.);

d) языки запросов, в том числе к реляционным базам данных, для описания различных типов данных (XML, RDF, JSON, мультимедиа и т. д.) и операций с большими данными (например, матричных операций);

e) предметно-ориентированные языки;

f) семантика конечной согласованности данных (оптимистическая репликация);

g) расширенные сетевые протоколы для эффективной передачи данных;

h) общие и предметно-ориентированные онтологии и таксономии для описания семантики данных, включая взаимодействие между онтологиями;

i) безопасность больших данных, управление доступом к персональным данным и их конфиденциальность;

j) удаленная, распределенная и федеративная аналитика, включая обнаружение данных, их извлечение и выявление ресурсов обработки;

k) совместное использование данных и обмен ими;

l) хранение данных, например память, система хранения, распределенная файловая система, хранилище данных и т. д.;

m) использование результатов анализа больших данных (например, визуализация);

n) измерение энергозатрат для обработки больших данных;

o) интерфейс между реляционными (SQL) и не только реляционными (NoSQL) хранилищами данных;

p) качество и достоверность больших данных, описание и управление.

ИСО/МЭК 20546 и серия стандартов ИСО/МЭК 20547 разработаны с учетом указанных приоритетов.

В настоящем стандарте рассматриваются структура и процесс применения, сценарии использования больших данных и требования к ним (приоритет «a»), эталонные архитектуры (приоритет «a»), безопасность и конфиденциальность персональных данных (приоритет «i»), а также дорожная карта стандартов. Кроме того, организации, имеющие потребности в анализе больших данных, не могут ждать разработки конкретных стандартов, решая задачи внедрения своих систем. Поскольку большие данные — это, по сути, подмножество всевозможных данных, а почти каждый стандарт в области информационных технологий связан с данными, сегодня существует большое число разработанных или разрабатываемых стандартов, которые затрагивают вопросы, связанные с большими данными. Поэтому последняя часть серии стандартов ИСО/МЭК 20547 представляет собой дорожную карту разработки стандартов, в которой существующие стандарты приведены в соответствие с эталонной архитектурой больших данных, что может быть использовано заинтересованными сторонами в качестве руководства при решении текущих задач. В разделе 7 описаны все остальные части указанной серии.

7 Концептуальные основы

7.1 Общие сведения

Стандарты серии ИСО/МЭК 20547 призваны обеспечить широкому кругу заинтересованных сторон основу для однозначного описания и эффективного обмена сведениями о характеристиках и атрибутах конкретной системы больших данных. В соответствии с терминами и определениями, представленными в ИСО/МЭК 20546, система больших данных позволяет:

- обрабатывать большие массивы данных, отличающиеся объемом, разнообразием, скоростью обработки и/или вариативностью, с помощью масштабируемой архитектуры для эффективного хранения, обработки, управления и анализа;
- применять передовые методики построения масштабируемых систем данных на основе независимых ресурсов в ситуациях, когда характеристики массивов данных требуют разработки новых архитектур для эффективного хранения, обработки, управления и анализа;
- реализовывать парадигму распределения систем данных по горизонтально соединенным независимым ресурсам с целью достижения масштабируемости, необходимой для эффективной обработки больших массивов данных.

Разнообразная природа систем больших данных определяет необходимость того, чтобы эталонная архитектура, которая представлена в серии стандартов ИСО/МЭК 20547, была достаточной для описания широкого диапазона потенциальных сценариев использования, реализуемых системами больших данных.

7.2 Эталонная архитектура. Основные понятия

Для понимания того, что включает эталонная архитектура, необходимо определить, что под ней подразумевается. Как указано в ISO/IEC/IEEE 42010:2011, 3.2, эталонная архитектура неизбежно обладает всеми характеристиками архитектуры, представляющей основные понятия или свойства системы в окружающей среде, которые воплощены в ее элементах, отношениях и конкретных принципах разработки и развития. В данном случае эталонная архитектура больших данных должна быть достаточно общей, охватывать многообразие потенциальных архитектур систем больших данных.

С объектно-ориентированной точки зрения эталонная архитектура представляется как абстрактный класс, определяющий структуру и атрибуты конкретных вариантов архитектур.

Определение эталонной архитектуры в области архитектуры программного обеспечения или архитектуры предприятия, типовое решение для архитектуры в конкретной области применения, а также общий словарь, который позволяет выявлять общие черты и обсуждать варианты реализации, даются в ISO TR 14639-2.

С учетом сказанного эталонная архитектура представляет собой структуру архитектуры, как это описано в ISO/IEC/IEEE 42010:2011, включающую строение и взаимосвязь компонентов, правила и ограничения, общие для всех систем больших данных, а также ряд соглашений, принципов и практик для описания архитектур систем больших данных.

Эталонные архитектуры, как показано на рисунке 1, разрабатываются для решения широкого круга задач [14]. Как утверждается в работе [14], основное предназначение эталонной архитектуры состоит в ориентации на будущее и использовании в качестве основы для будущих реализаций.

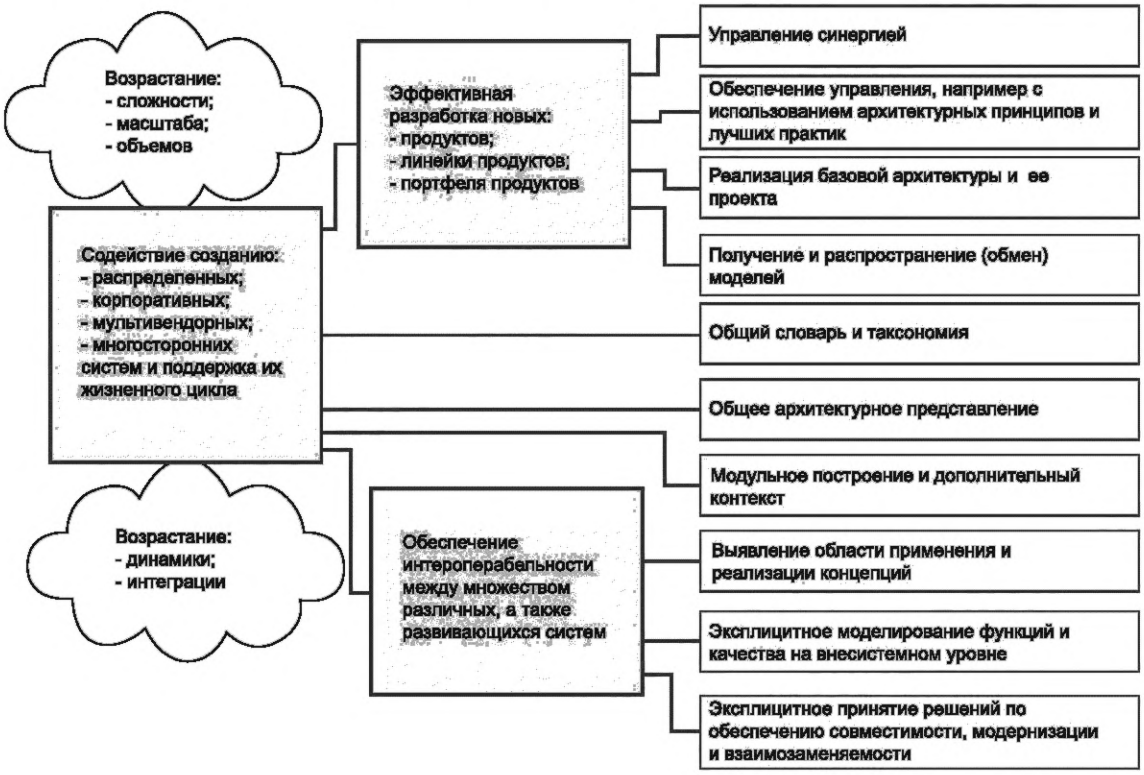

Рисунок 1 — Концепция эталонных архитектур

7.3 Структура эталонной архитектуры

На рисунке 2 в соответствии с ISO/IEC/IEEE 42010:2011 представлено объединение концепции и структуры эталонной архитектуры.

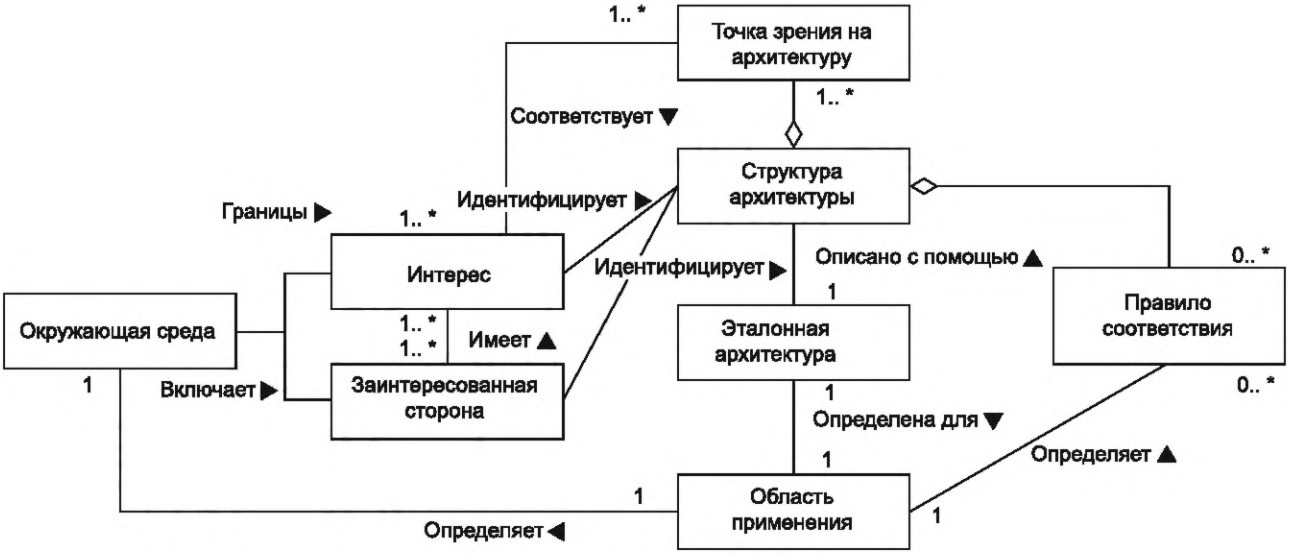

Рисунок 2 — Схема взаимосвязей между базовыми понятиями эталонной архитектуры на основе ISO/IEC/IEEE 42010:2011

Эталонная архитектура определяется для конкретной области применения. Областью применения для рассматриваемой эталонной архитектуры являются большие данные.

Область применения, в свою очередь, определяет окружающую среду. В случае больших данных окружающая среда в первую очередь определяется основными характеристиками больших данных — объемом, скоростью обработки, разнообразием, вариативностью [7].

Под заинтересованными сторонами в окружающей среде подразумеваются все пользователи, владельцы, архитекторы и другие субъекты любой системы, а также те, у кого имеется интерес, связанный с данными и их свойствами.

Интересы ограничиваются окружающей средой. Поскольку в данном случае окружающая среда определяется характеристиками больших данных, интересы ограничиваются этими характеристиками, а каждый из них должен соотноситься с одной или несколькими характеристиками, а также с заинтересованной стороной.

Эталонная архитектура описывается на основе ее структуры, которая представлена в ИСО/МЭК 20547-3 и рассматривается с двух точек зрения:

- пользовательское представление — роли и деятельность;
- функциональное представление — функциональные компоненты.

Каждая из этих точек зрения, в свою очередь, затрагивает один или несколько интересов.

В рамках указанных архитектурных представлений эти интересы могут быть воплощены в одной или нескольких ролях, действиях и функциональных компонентах.

Пример — В системе кредитного мониторинга каждое лицо, имеющее запись, является заинтересованным лицом. Для большинства заинтересованных лиц существует необходимость обеспечения безопасности и конфиденциальности персональных данных.

Как в пользовательском, так и в функциональном представлениях эталонной архитектуры больших данных существует сквозной аспект безопасности и конфиденциальности персональных данных. Этот сквозной аспект связан с видом деятельности «Проведение аудита» и функциональным компонентом «Структура аудита», которые позволяют разрешить проблему.

8 Основы эталонной архитектуры больших данных

8.1 Общие положения

Настоящий документ содержит описание структуры эталонной архитектуры больших данных, предназначенной для использования в окружающей среде системы больших данных, сферы применения серии стандартов ИСО/МЭК 20547, логической взаимосвязи между всеми частями данной серии

стандартов и процесса применения эталонной архитектуры. Ниже представлено описание каждой части серии стандартов ИСО/МЭК 20547:

- ИСО/МЭК 20547-1: «Структура и процесс применения» включает описание структуры эталонной архитектуры больших данных и процесса применения стандарта в конкретной предметной области;
- ИСО/МЭК 20547-2: «Примеры использования и производные требования» включает примеры описания сценариев использования больших данных в соответствии с областями применения и вытекающими из них техническими аспектами;
- ИСО/МЭК 20547-3: «Эталонная архитектура» содержит описание эталонной архитектуры больших данных (BDRA), включающей в себя понятия и архитектурные представления (пользовательское и функциональное);
- ИСО/МЭК 20547-4: «Безопасность больших данных и конфиденциальность персональных данных» содержит описание аспектов безопасности и конфиденциальности применительно к эталонной архитектуре больших данных (BDRA), включая роли, действия, функциональные компоненты, а также руководство по обеспечению безопасности и конфиденциальности при операциях с большими данными;
- ИСО/МЭК 20547-5: «Направления стандартизации» включает описание стандартов (как существующих, так и разрабатываемых) в соответствии с анализом приоритетов разработки будущих стандартов, относящихся к большим данным.

На рисунке 3 показаны взаимосвязи между частями серии стандартов ИСО/МЭК 20547.

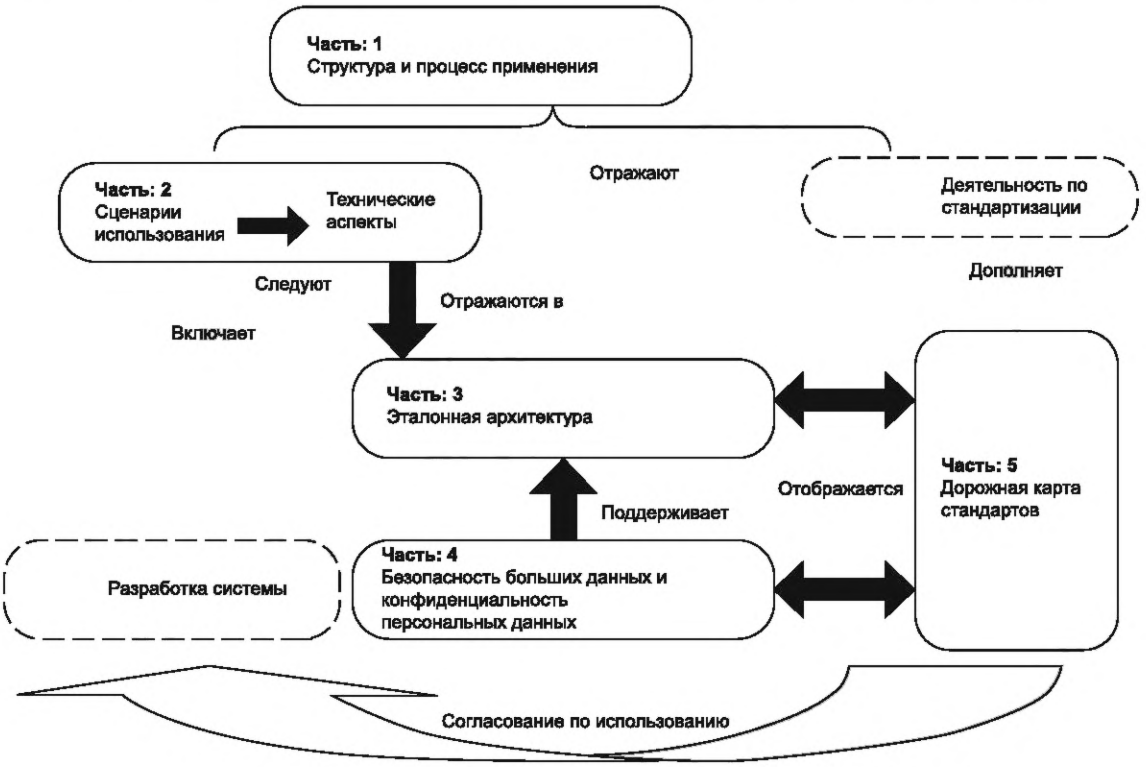

Рисунок 3 — Взаимосвязи между частями стандарта ИСО/МЭК 20547

В ИСО/МЭК 20547-2 представлены сценарии использования и технические аспекты на основе исследований научных сообществ, оценок экспертов и специалистов предприятий и организаций. В ИСО/МЭК 20547-3 представлена эталонная архитектура больших данных с соответствующими техническими аспектами. В ИСО/МЭК 20547-4 представлены аспекты обеспечения безопасности больших данных и конфиденциальности персональных данных. В ИСО/МЭК 20547-5 представлен список применяемых стандартов в рамках эталонной архитектуры больших данных.

Для того чтобы применить рассматриваемую структуру в конкретных сценариях, необходимо иметь представление об окружающей среде, в которой будет реализована система больших данных, заинтересованных сторонах и их интересах (потребностях). В 8.2—8.4 рассмотрен каждый из перечисленных ключевых аспектов.

8.2 Заинтересованные стороны

Заинтересованная сторона в ISO/IEC/IEEE 42010:2011 определяется как индивидуум, команда, организация или их группы, имеющие интерес к системе. В общем случае заинтересованные стороны включают в себя владельцев системы, клиентов, специалистов по внедрению и др. К заинтересованным сторонам также относятся лица и организации, которых интересуют данные, обрабатываемые системой. К ним относятся владельцы данных, которые могут поставлять их в систему, потребители, принимающие решения на основе данных, поступающих из этой системы, а также лица или организации, сведения о которых представлены в системе. Определение заинтересованных сторон и их интересов (проблем) является первым шагом в разработке архитектуры больших данных. В ИСО/МЭК 20547-3 под заинтересованными сторонами системы больших данных подразумеваются стороны в рамках представления пользователя.

8.3 Потребности сторон

Любая заинтересованность в системе, актуальная для одной или нескольких заинтересованных сторон, представляет собой потребность, которая может быть связана с любым аспектом системы больших данных, включая технические, функциональные, эксплуатационные, юридические и даже социальные воздействия на систему больших данных в ее среде, как это представлено в ISO/IEC/IEEE 42010:2011.

Примечания

1 Например, прозрачность распределения, представленная в эталонной модели открытой распределенной обработки в соответствии с ИСО/МЭК 10746-1, относится к одной из проблем, связанных с функционированием системы больших данных. Ключевыми аспектами таких систем являются горизонтальное масштабирование и распределенная обработка.

2 Свойства программного обеспечения, такие как эффективность, производительность, доверие, риски и их снижение, а также гибкость, представленные в ИСО/МЭК 25010:2011, 4.2 (требования к оценке качества аппаратного и программного обеспечения), обозначают проблемы, связанные с качеством программного обеспечения. При работе с большими данными возникают дополнительные проблемы, связанные с такими характеристиками, как объем, скорость обработки и разнообразие больших данных.

3 Например, проблема может быть связана с потерей данных из-за скорости их обработки.

Кроме того, существует ряд вопросов, связанных с самими данными, включая их происхождение и защищенность. Проблемы, связанные с безопасностью и конфиденциальностью больших данных, являются настолько существенными, что им посвящен ИСО/МЭК 20547-4. Например, возможность слияния данных из нескольких источников с применением технологий больших данных для их деанонимизации представляет собой особую проблему конфиденциальности.

Проблемы, выявленные для системы больших данных, в свою очередь, обуславливают особенности функционирования системы и ее компонентов.

8.4 Представления эталонной архитектуры больших данных

Как отмечено выше, архитектура больших данных может быть определена в терминах представлений. В эталонной архитектуре больших данных определяются два основных варианта представления:

- пользовательское представление, которое включает роли, подроли, действия и сквозные аспекты, обеспечивающие удовлетворение потребностей заинтересованных сторон;
- функциональное представление, которое включает функциональные уровни, компоненты и многоуровневые функции, обеспечивающие реализацию действий и сквозных аспектов, указанных в представлении пользователя.

На рисунке 4 представлена схема взаимодействия заинтересованных сторон, а также интересы (проблемы), связанные с указанными представлениями.

8.4.1 Пользовательское представление

Связь пользовательского представления с экосистемой больших данных описывается с помощью следующих понятий:

- сторона: физическое или юридическое лицо, вне зависимости от принадлежности к организации, или группа лиц. Стороны в экосистеме больших данных являются ее заинтересованными сторонами (stakeholder);

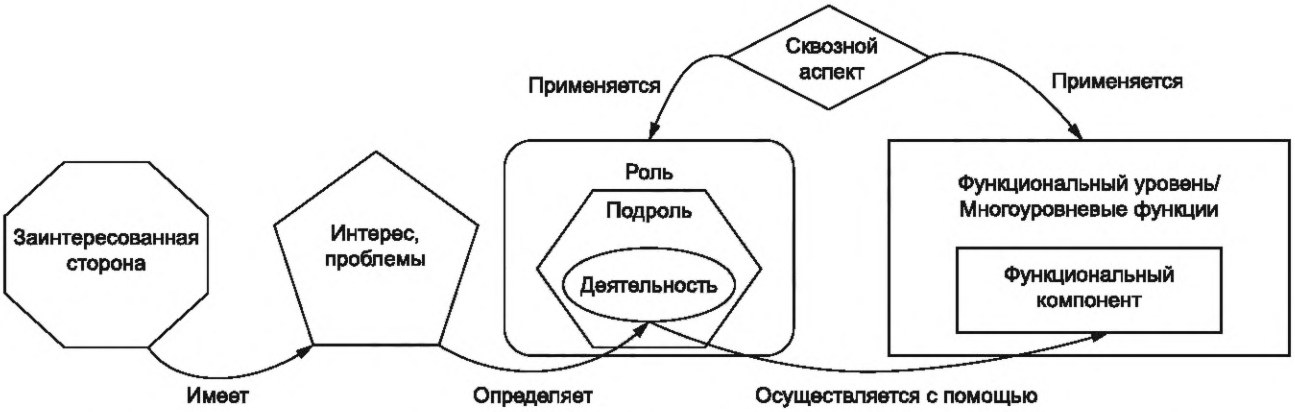

Рисунок 4 — Взаимосвязи между элементами представлений эталонной архитектуры больших данных

- роль и подроль: роль — множество действий с большими данными, которые служат общей цели; подроль — это подмножество действий с большими данными для конкретной роли, различные подроли могут совместно использовать действия с большими данными, связанные с данной ролью;

- деятельность: определенная последовательность действий или набор задач. Действия с большими данными, которые выполняются с использованием функциональных компонентов, должны иметь цель и обеспечивать один или несколько результатов;

- сквозной аспект: поведение или возможности, которые должны координироваться между ролями и последовательно реализовываться в экосистеме больших данных. Сквозной аспект влияет на выполнение нескольких ролей, на действия с большими данными и функциональные компоненты и учитывается при совместном выполнении конкретных ролей или использовании функциональных компонентов.

Примечания

1 Примером сквозных аспектов является безопасность больших данных.

2 Сторона может взять на себя более чем одну роль в любой момент времени и участвовать в определенном подмножестве действий этой роли. Примерами сторон являются крупные корпорации, малые и средние предприятия, правительственные департаменты, академические институты и частные лица.

8.4.2 Функциональное представление

При формировании системы больших данных функциональное представление является технологически нейтральным и позволяет описывать распределение функций, необходимых для поддержки действий с большими данными. Взаимосвязь между функциями определяется функциональной архитектурой.

Функциональное представление соответствует следующим концептам больших данных:

- функциональный компонент: функциональный строительный блок, необходимый для обеспечения деятельности и поддержки реализации;

- функциональный уровень: набор функциональных компонентов, предоставляющих одинаковые возможности или обеспечивающих достижение общей цели. Функциональная архитектура является частично многоуровневой (т. е. имеет уровни и набор многоуровневых функций);

- многоуровневые функции: включают в себя сгруппированные в подмножества функциональные компоненты, которые обеспечивают возможность их сквозного использования на нескольких функциональных уровнях.

Примечание — В конкретной системе больших данных не всегда содержатся все уровни или функциональные компоненты.

9 Процесс применения эталонной архитектуры больших данных

9.1 Общие положения

В данном разделе представлен пошаговый процесс применения эталонной архитектуры для разработки архитектуры конкретной системы больших данных. Эталонная архитектура больших данных является достаточно общей и предназначена для применения в широком диапазоне систем, однако

из-за потенциального разнообразия систем больших данных и их компонентов процесс применения дает возможность расширения эталонной архитектуры с учетом конкретных требований. Основной особенностью этого расширения является идентификация дополнительных действий, связанных с ролями, и/или назначение действий различным ролям/подролям. На данном шаге рекомендуется использовать соответствующие стандарты ИСО, включая ISO/IEC/IEEE 15288 для разработки систем, ИСО/МЭК 12207 для описания жизненного цикла разработки систем. Качество процессов оценивается с помощью серии стандартов ИСО 9000 для подтверждения того, что разработанная архитектура действительно охватывает и учитывает весь спектр проблем.

Прежде чем приступить к рассматриваемому процессу, необходимо определить инструменты, которые будут использоваться для сбора и управления сгенерированными данными.

9.2 Идентификация заинтересованных сторон и их требований

Первым шагом в процессе применения эталонной архитектуры является идентификация заинтересованных сторон и их интересов, связанных с разработкой системы больших данных. Анализ требований заинтересованных сторон проводится в соответствии с ISO/IEC/IEEE 15288. В результате первого шага должны быть:

a) указаны необходимые характеристики и контекст использования системных сервисов;

b) определены ограничения на соответствующую систему;

c) определены соответствия требований заинтересованных сторон их потребностям;

d) определены базовые положения для формирования системных требований;

e) определены базовые положения для валидации системных услуг;

f) определены базовые положения для обсуждения и согласования условий поставки системных сервисов или продуктов.

В соответствии с ISO/IEC/IEEE 15288 интересы (требования) заинтересованных сторон выражаются в потребностях, желаниях, ожиданиях и предполагаемых ограничениях. Они представляются в виде вербальной или формальной модели, сфокусированной на цели и поведении системы, и описываются в контексте операционной среды и условий функционирования.

Интересы заинтересованных сторон должны включать в себя потребности и требования, налагаемые обществом (например, в отношении защиты персональных данных) и нормативным регулированием (например, Общий регламент по защите данных в Европейском Союзе).

Заинтересованные стороны и их интересы должны быть отражены в модели, которая впоследствии может быть использована с целью отслеживания действий системы и ее компонентов для поддержки верификации процесса.

В частности, все заинтересованные стороны и их интересы должны быть однозначно идентифицированы для прослеживаемости на следующих шагах. Общие интересы заинтересованных сторон следует, по возможности, объединить в один интерес и соотнести его с каждой заинтересованной стороной.

На данном шаге при определении заинтересованных сторон и их интересов могут применяться результаты анализа других сценариев использования больших данных и соответствующих требований ИСО/МЭК 20547-2.

Заинтересованные стороны должны рассмотреть полученные результаты и подтвердить их точность, а также обоснованность требований и эффективность процесса проектирования.

9.3 Отображение в ролях и подролях заинтересованных сторон и их требований

Целью этого шага является отображение заинтересованных сторон и их требований в общей структуре понятий и представлений о системе больших данных. Этот шаг критически важен для систем больших данных, поскольку во многих случаях интересы должны реализовываться через представление о системе систем (то есть, нескольких систем, совместно координируемых для удовлетворения требований).

В ИСО/МЭК 20547-3 определены следующие роли и подроли:

- провайдер приложения больших данных (BDAP):
  - сервис-провайдер комплектования больших данных (BDCP);
  - сервис-провайдер предобработки больших данных (BDPreP);
  - сервис-провайдер аналитики больших данных (BDAnP);
  - сервис-провайдер визуализации больших данных (BDVP);
  - сервис-провайдер доступа к большим данным (BDAcP);

- провайдер среды обработки больших данных (BDFP):
  - сервис-провайдер инфраструктуры больших данных (BDIP);
  - сервис-провайдер платформы больших данных (BDPlaP);
  - сервис-провайдер обработки больших данных (BDProP);
- партнер сервиса больших данных (BDSP):
  - разработчик сервиса больших данных (BDSD);
  - аудитор больших данных (BDA);
  - оркестратор системы больших данных (BDSO);
- потребитель сервиса данных;
- сервис-провайдер больших данных.

В этом случае полезным инструментом для отображения соответствия интересов ролям/подролям является матрица перекрестных ссылок. Действия ролей/подролей обеспечивают разрешение проблем. В зависимости от точности соответствия требований архитектуре матрица может быть представлена диаграммой, которая имеет индикатор соответствия (точка пересечения) между интересом и ролью, или содержать конкретное подтверждение/обоснование причины этого соответствия, и/или включать экспертные заключения о соответствии определенного аспекта интереса той или иной роли. В любом случае следует позаботиться о том, чтобы не описывать в этом отображении действия и решения. Этот шаг представляет собой начальное распределение требований (интересов) по классам высокоуровневых процессов (ролям/подролям).

9.4 Разработка подробных описаний деятельности и ее соответствие интересам

На данном шаге определяется деятельность системы или архитектуры. На основе эталонной архитектуры больших данных, представленной в категориях ролей и подролей, создаются механизмы для сбора и организации результатов реализации процессов системной и программной инженерии. При выполнении этого шага группа разработчиков должна опираться на положения ISO/IEC/IEEE 29148. Этот стандарт поможет трансформировать описанные выше проблемы в конкретные и тестируемые требования.

Стандарта, устанавливающего соответствие операций/действий той или иной роли, не существует, поэтому один из подходов заключается в том, чтобы зафиксировать это соответствие в формате утверждения требований. Эти утверждения должны иметь следующую общую форму:

<Роль или подроль> должна <глагол><объект>

Роль или подроль соответствуют субъекту, модальность и глагол — предикату, а объект — предмету, над которым выполняется действие. В ИСО/МЭК 20547-3 подробно описываются классы действий, которые должны выполняться системами.

Как правило, объект обозначает некоторый фрагмент данных, подлежащий обработке, а глагол — операцию/действие, подлежащее выполнению.

Такие утверждения могут рассматриваться как бизнес-правила, реализуемые в системе, для которой разрабатывается архитектура. Ниже приведены примеры возможных бизнес-правил, создаваемых на данном шаге:

- сервис-провайдер комплектования больших данных должен подтверждать соответствие данных XXX стандарту YYYY;
- сервис-провайдер визуализации больших данных должен представлять результаты в виде ориентированного графа;
- сервис-провайдер доступа к большим данным должен фиксировать доступ пользователя к системе.

В целом, рассмотренные утверждения должны соответствовать требованиям по полноте и непротиворечивости. На данном шаге крайне важно заранее не предписывать выполнение действий. Для крупномасштабных систем может оказаться полезной организация деятельности в соответствии с классами, представленными в эталонной архитектуре больших данных. Иногда подобная дополнительная структурированность может быть полезной для определения функциональных компонентов, необходимых для реализации.

Как и в случае с заинтересованными сторонами и их интересами, для идентификации каждого действия должна использоваться уникальная нумерация/номенклатура, которая имеет важное значение при отображениях. Один из подходов заключается в присвоении префикса (указателя) роли/подроли перед уникальным номером для каждого действия.

В результате выполнения этого шага будут сформированы конкретные тестируемые требования, упорядоченные по ролям, которые позволят создать архитектуру системы для удовлетворения потребностей заинтересованных сторон. При этом каждый интерес должен соответствовать одному или нескольким действиями, а каждое действие должно соответствовать одному или нескольким интересам.

Графическое представление/документирование информации зависит от степени детализации конкретных действий и может оказаться невозможным. В этом случае архитектору следует учитывать потребности пользователей при подготовке информации и форматировании данных для их представления. Если применяется описанный выше подход, желательна дальнейшая декомпозиция подролей на классы, которые, в свою очередь, соответствуют отдельным утверждениям о действиях. Этот процесс может быть очень полезным на следующем шаге, поскольку позволяет облегчить выбор функциональных компонентов, поддерживающих несколько видов действий.

9.5 Определение функциональных компонентов для реализации деятельности

Если предыдущий шаг связан с частью анализа требований к разработке процессов жизненного цикла системы/программного обеспечения, то данный шаг представляет собой фазу высокоуровневого проектирования системы больших данных. В то же время функциональные уровни и классы функциональных компонентов в функциональном представлении эталонной архитектуры формируют организационную структуру для конфигурационных элементов (программных или аппаратных), которые обеспечивают построение архитектуры системы больших данных.

В начале данного шага должны быть выбраны соответствующие компоненты, с помощью которых будут выполняться действия, определенные на предыдущем шаге. Компонентами могут быть инструменты, продукты или программные средства — существующие или новые, которые следует разработать с учетом необходимого набора действий. Уровень детализации, установленный на данном шаге, зависит от архитектора и потребностей системы. При разработке крупномасштабных систем рекомендуется иерархически структурировать компоненты, организуя их в подсистемы внутри уровней.

Во всех случаях компоненты должны быть соотнесены с одним или несколькими действиями. Необходимо обратить внимание на отсутствие требования соответствия действия одному компоненту. В зависимости от степени детализации задокументированных действий, возможно, для некоторых видов деятельности потребуется несколько компонентов.

Для упрощения трассировки и сквозного процесса разработки, которые рассмотрены выше, должна быть принята стандартная номенклатура компонентов.

Интерфейсы между функциональными компонентами выходят за рамки рассматриваемого архитектурного представления и будут определены и специфицированы в процессе внедрения как составная часть детализированной разработки.

В то же время в зависимости от используемого уровня детализации графическое представление информации на одной диаграмме не всегда возможно, хотя для каждого уровня/многоуровневой функции может быть разработан вариант графического представления. При этом в результате взаимодействия уровней некоторая часть контекста может быть потеряна.

9.6 Соответствие сквозных действий/функциональных компонентов интересам

Это последний шаг процесса разработки, который включает валидацию того, что посредством деятельности может быть выполнена трассировка любого требования до функционального компонента и наоборот, трассировка каждого функционального компонента до интереса, при этом любая деятельность фактически присутствует в этих взаимосвязях.

Именно на этом шаге с целью эффективной валидации высокоуровневой архитектуры оказывается важным выбор базы данных или других инструментов трассировки требований для сбора необходимой информации.

Приложение ДА
(справочное)

Сведения о соответствии ссылочных международных стандартов национальным стандартам

Таблица ДА.1

Обозначение ссылочного международного стандарта	Степень соответствия	Обозначение и наименование соответствующего национального стандарта
ISO/IEC/IEEE 42010:2011	IDT	ГОСТ Р 57100—2016/ISO/IEC/IEEE 42010:2011 «Системная и программная инженерия. Описание архитектуры»

Примечание — В настоящей таблице использовано следующее условное обозначение степени соответствия стандарта:

- IDT — идентичный стандарт.

Библиография

[1] ISO 9000:2015 Quality management systems — Fundamentals and vocabulary

[2] ISO 9001:2015 Quality management systems — Requirements

[3] ISO/TS 9002:2016 Quality management systems — Guidelines for the application of ISO 9001:2015

[4] ISO/IEC/IEEE 12207:2017 Systems and software engineering — Software Life Cycle Processes

[5] ISO/TR 14639-2:2014 Health informatics — Capacity-based eHealth architecture roadmap — Part 2: Architectural components and maturity model

[6] ISO/IEC/IEEE 15288:2015 Systems and software engineering — System life cycle processes

[7] ISO/IEC 20546:2019 Information technology — Big data — Overview and vocabulary

[8] ISO/IEC/TR 20547-2:2018 Information technology — Big data reference architecture — Part 2: Use cases and derived requirements

[9] ISO/IEC 20547-3:2020 Information technology — Big data reference architecture — Part 3: Reference architecture

[10] ISO/IEC 20547-4 Information technology — Big data reference architecture — Part 4: Security and privacy

[11] ISO/IEC/TR 20547-5:2018 Information technology — Big data reference architecture — Part 5: Standards roadmap

[12] ISO/IEC/IEEE 29148:2018 Systems and software engineering — Life cycle processes — Requirement's engineering

[13] ISO/IEC JTC 1 (2015). Big Data, Preliminary Report 2014

[14] Cloutier R., Muller G., Verma D., Nilchiani R., Hole E., Bone M. (2009). The Concept of Reference Architecture. Systems Engineering, 13, 14-27. doi: https://doi.org/10.1002/sys.20129

[15] IDC (2019). Worldwide Semiannual Big Data and Analytics Spending Guide. IDC

УДК 004.01:006.354 ОКС 35.020

Ключевые слова: информационные технологии, большие данные, архитектура, структура архитектуры, архитектурное представление, точка зрения на архитектуру, эталонная архитектура, интерес, заинтересованная сторона, окружающая среда, исполнитель, защищенность, конфиденциальность персональных данных, происхождение, безопасность больших данных, репозиторий, жизненный цикл, пользовательское представление, функциональное представление

Редактор *В.Н. Шмельков*
Технический редактор *В.Н. Прусакова*
Корректор *О.В. Лазарева*
Компьютерная верстка *И.А. Налейкиной*

Сдано в набор 05.12.2022. Подписано в печать 06.12.2022. Формат 60×84⅛. Гарнитура Ариал.
Усл. печ. л. 2,32. Уч.-изд. л. 2,12.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «Институт стандартизации»
для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.
www.gostinfo.ru info@gostinfo.ru