# ОБЗОР НАЦИОНАЛЬНОГО СТАНДАРТА РФ ГОСТ Р ИСО/МЭК ТО 20547–1–2021 «ЭТАЛОННАЯ АРХИТЕКТУРА БОЛЬШИХ ДАННЫХ Часть 1. Структура и процесс применения»

**А.А. ЕГОРОВ (Журнал “Автоматизация и IT в нефтегазовой области”)**

АВТОМАТИЗАЦИЯ & IT в нефтегазовой области
отраслевой научно-производственный журнал

В статье рассматривается история разработки национального стандарта РФ ГОСТ Р ИСО/МЭК ТО 20547-1-2021 “Эталонная архитектура больших данных, Часть 1. Структура и процесс применения”. Отмечается, что цифровая трансформация промышленности сфокусировала внимание на необходимости разработки компьютерных систем и программного обеспечения для работы с большими и разнообразными наборами данных. В стандарте введены термины и определения в области архитектуры больших данных. Рассмотрены мотивы и цели стандартизации больших данных. Приведены основные понятия структуры и представления эталонной архитектуры больших данных. Рассмотрен процесс применения эталонной архитектуры больших данных. Определены функциональные компоненты для реализации деятельности по проектированию систем больших данных.

**Ключевые слова:** *ИСО/МЭК 20547; ГОСТ Р ИСО/МЭК ТО 20547–1-2021; большие данные (big data); эталонная архитектура (reference architecture); структура (framework); защищенность (security); конфиденциальность персональных данных (privacy); происхождение (provenance); реляционная модель данных (relational model); жизненный цикл (lifecycle).*

## ИСТОРИЯ РАЗРАБОТКИ СТАНДАРТА

Международная организация по стандартизации и Международная электротехническая комиссия разрабатывают материалы и международные стандарты в сфере информационно-коммуникационных технологий (ИКТ). Это объемный термин для информационных технологий (ИТ), что подчеркивает роль унифицированных коммуникаций и интеграции телекоммуникаций (телефонные линии и беспроводные сигналы) и компьютеров, а также необходимого корпоративного программного обеспечения, промежуточного программного обеспечения, хранения и аудиовизуальных материалов, которые позволяют пользователям получать доступ, хранить, передавать и обрабатывать информацию. Разработки стандартов ИКТ проводятся в рамках Объединенного технического комитета ISO/IEC JTC 1 [1], который, в свою очередь, состоит из подкомитетов, охватывающих 22-е предметные области. В частности, подкомитет SC 42 Artificial Intelligence занимается стандартизацией в области искусственного интеллекта.

Разработанный и опубликованный подкомитетом SC 42 международный стандарт в формате технического отчета ISO/IEC TR 20547-1 описывает структуру и процесс построения архитектуры больших данных. Предлагаемая стандартом структура нацелена на то, чтобы организации могли эффективно и последовательно описать свою архитектуру, ее внедрение с учетом ролей/участников (поставщиков приложений и структуры в сфере больших данных, сервисных партнеров) и проблемных вопросов (технических, операционных, правовых и т.д.) с учетом основополагающей технологии. Затем организации могут сопоставить предлагаемую структуру с собственными мероприятиями и функциональными компонентами, направленными на реализацию архитектуры.

“*Цифровая трансформация промышленности сфокусировала внимание на необходимости компьютерных систем по работе с большими и разнообразными наборами данных, свойства которых (разнообразие, объем, скорость и достоверность) могут значительно отличаться в зависимости от практического применения*, –

сообщил председатель SC 42 **Ваэль Уильям Диаб**. – *Серия стандартов по эталонной архитектуре больших данных [**Big data reference architecture, BDRA**] устанавливает основы для экосистемы больших данных*".

Технический отчет ISO/IEC TR 20547–1 описывает структуру BDRA, позволяет сопоставить конкретные наборы задач/примеров использования с эталонной архитектурой, а также оценить это сопоставление. ISO/IEC TR 20547–1 входит в серию стандартов ISO/IEC 20547-X по эталонной архитектуре больших данных. Эта серия также включает в себя ISO/IEC TR 20547–2 (сценарии использования и производные требования), ISO/IEC 20547–3 (эталонная архитектура), ISO/IEC 20547–4 (безопасность больших данных и конфиденциальность персональных данных), ISO/IEC TR 20547-5 (дорожная карта стандартов). На основе международного стандарта ISO/IEC 20547–3 Российской венчурной компанией в 2019 году была подготовлена первая редакция предварительного национального стандарта "Информационные технологии. Большие данные. Типовая архитектура" (прошла общественное обсуждение летом 2019 года).

Кроме того, в 2019 году был утвержден международный стандарт ISO/IEC 20546, который содержит обзор предметной области работы с большими данными и терминологический словарь (в рамках разработки российским ПК 02 идентичного национального стандарта ГОСТ Р "Информационные технологии. Большие данные. Обзор и словарь" в июле 2020 года завершилась стадия общественного обсуждения документа).

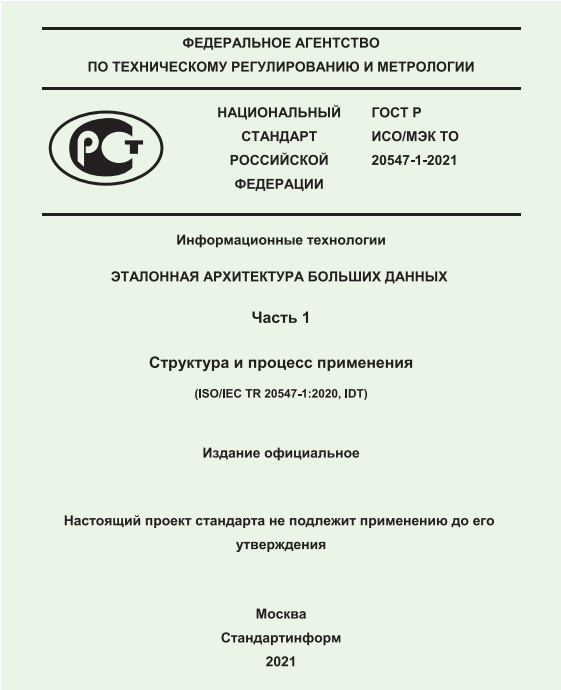
Для того чтобы заинтересованные стороны понимали, какие системы работы с большими данными внедряют, и поддерживали надежную и четкую коммуникацию, необходима однозначная экосистема взаимодействия с потенциальными поставщиками технологий и услуг в области больших данных.

Она включает в себя понимание возможных проблемных вопросов и обязательств, связанных с управлением данными и их контролем, с целью обеспечения их безопасности, качества, соответствия требованиям, соблюдения авторских прав и конфиденциальности.

"*Очень важно, чтобы организации могли определять, устанавливать, формулировать политику безопасности и происхождения данных, политику управления ими, а также внедрять и документировать технические средства контроля для обеспечения соблюдения этих политик. Так они [организации] могут защитить себя от ответственности за нарушения или злоупотребления в сфере данных, которые они [организации] контролируют*", – отметил **Во Чанг**, руководитель рабочей группы по данным подкомитета SC 42, которая разрабатывает стандарты серии ISO/IEC 20547-X.

Актуальность этой тематики, в том числе, связана с тем, что многие организации, работающие с большими данными, получают данные извне. Следовательно, системы, которые собирают и анализируют большие данные, должны безопасно и надежно обмениваться данными и быть совместимыми.

Национальный стандарт РФ ГОСТ Р ИСО/МЭК ТО 20547-1-2021 разработан Федеральным государственным бюджетным образовательным учреждением высшего образования "Московский государственный университет имени М.В. Ломоносова" (МГУ имени М.В. Ломоносова) в лице Научно-образовательного центра компетенций в области цифровой экономики МГУ и Автономной некоммерческой организацией "Институт развития информационного общества" (ИРИО) на основании собственного перевода на русский язык англоязычной версии стандарта, указанного в пункте 4.2 и внесен Техническим комитетом по стандартизации **ТК 164 "Искусственный интеллект"**. На рис. 1 Представлен титульный лист стандарта.


ФЕДЕРАЛЬНОЕ АГЕНТСТВО
ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р ИСО/МЭК ТО 20547-1-2021

Информационные технологии

ЭТАЛОННАЯ АРХИТЕКТУРА БОЛЬШИХ ДАННЫХ

Часть 1

Структура и процесс применения

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

Издание официальное

Настоящий проект стандарта не подлежит применению до его утверждения

Москва
Стандартинформ
2021

*Рис. 1. Титульный лист стандарта ГОСТ Р ИСО/МЭК ТО 20547-1-2021*

Этот стандарт идентичен техническому отчёту ИСО/МЭК ТО 20547–1:2020 “Эталонная архитектура больших данных. Часть 1. Структура и прикладные процессы” (ISO/IEC TR 20547–1:2020 “Information technology – Big data reference architecture – Part 1: Framework and application process”, IDT).

Парадигма больших данных относится к быстро развивающейся предметной области со стремительно меняющимися технологиями. Эта динамическая ситуация порождает две существенные проблемы для потенциальных разработчиков технологий. Во-первых, не хватает стандартных определений терминов, включая ключевое понятие больших данных. Вторая проблема заключается в том, что нет единого подхода к описанию архитектуры больших данных и ее реализации. Первая проблема разрешена в стандарте ГОСТ Р ИСО/МЭК 20546–2020 “Информационные технологии. Большие данные. Обзор и словарь” (ISO/IEC 20546:2019 “Information technology – Big data – Overview and vocabulary”, IDT).

Разработка серии стандартов и технических отчетов 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 стандарте описаны структура эталонной архитектуры, процесс отображения требований/вариантов использования в эталонной архитектуре, а также оценка этого отображения. В стандарте также содержится описание структуры эталонной архитектуры больших данных и процесса применения стандарта в рассматриваемой предметной области.

## ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

Для целей стандарта используются следующие термины и определения. ГОСТ Р ИСО/МЭК ТО 20547-1-2021 2 ISO (ИСО) и IEC (МЭК) поддерживают терминологические базы данных для использования в стандартизации по следующим адресам:

- Онлайн-библиотека стандартов ISO (ИСО): доступна по адресу *https://www.iso.org/obp*
- Международный электротехнический словарь МЭК (IEC Electropedia): доступен по адресу *http://www.electropedia.org/*

**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 – Термин “большие данные” широко применяется в различных значениях, например, в качестве названия технологии масштабирования, используемой для обработки больших массивов данных [Международный стандарт ИСО/МЭК (ISO/IEC) 20546:2019, 3.1.2].

**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 [Технический отчёт ИСО/ТО (ISO/TR) 14639-2:2014, 2.65].

**Структура (*framework*)**: определенный набор утверждений (концепций, понятий) или идей для описания сценария или решения задачи [Международный стандарт ИСО/МЭК (ISO/IEC) 15638-6-2014, 4.30] ГОСТ Р ИСО/МЭК ТО 20547-1-2021.

**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 [Международный стандарт ИСО/МЭК/ИИЭР (ISO/IEC/IEEE) 15288:2015, 4.1.39].

**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 [Международный стандарт ИСО/МЭК ТО (ISO/IEC TR) 26927:2011, 3.34].

**Происхождение (provenance):** сведения о месте и времени появления, извлечения или создания ресурса, записи, доказательства подлинности или принадлежности в прошлом [Международный стандарт ИСО/МЭК (ISO/IEC) 11179-7:2019, 3.1.10].

**Реляционная модель данных (*relational model*):** модель данных, структура которой основана на совокупности отношений [Между-

народный стандарт ИСО/МЭК (ISO/IEC) 2382:2015, 9] Примечание – Примером такой модели является язык SQL.

**Жизненный цикл (*lifecycle*):** развитие системы, продукта, услуги, проекта или другой создаваемой человеком сущности от замысла до списания [Международный стандарт ИСО/МЭК (ISO/IEC) 15288:2015, 4.1.19].

## СОКРАЩЕНИЯ И ОБОЗНАЧЕНИЯ, ПРИМЕНЯЕМЫЕ В СТАНДАРТЕ

**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*);
**DAP** – 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*);
**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*);
**BDRA** – эталонная архитектура больших данных (*Big Data Reference Architecture*);
**BDVP** – сервис-провайдер визуализации больших данных (*Big Data Visualization Provider*);
**GDPR** – общие правила защиты 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 аппаратного и программного обеспечения (*Systems and software Quality Requirements and Evaluation*);
**XML** – расширяемый язык разметки (*Extensible Markup Language*).

## ОБЗОР СТАНДАРТА

Стандарт предназначен для ознакомления с понятиями в сфере эталонной архитектуры больших данных в целях применения других документов серии ИСО/МЭК 20547 “Информационные технологии. Эталонная архитектура больших данных” (ISO/IEC 20547 “Information technology – Big data reference architecture”) к конкретным системам и наборам задач.

В стандарт включены следующие разделы:

- 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 “Информационные технологии. Эталонная архитектура больших данных” (ISO/IEC 20547 “Information technology – Big data reference architecture”) ГОСТ Р ИСО/МЭК ТО 20547–1-2021 будет полезен в следующих случаях:

- для получения общего представления о применении эталонной архитектуры больших данных необходимо использовать содержание параграфов 5, 6 и 7;
- для разработки архитектуры больших данных и приведения ее в соответствие эталонной архитектуре необходимо использовать описание процесса.

## 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. [2]. 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.

В стандарте ГОСТ Р ИСО/МЭК ТО 20547-1-2021отмечается что:

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 их достоверности, авторского права и конфиденциальности персональных данных. Кроме того, по оценке компании IDC к 2020 г. более 40 % данных в цифровом пространстве потребует обеспечения надежной защиты, а объем этих данных будет расти быстрее, чем все цифровое пространство [2]. 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 применять аналитические инструменты на уровнях интероперабельности данных, программного обеспечения и приложений. Изучение существующего ландшафта больших данных, рыночных требований к стандартизации области больших данных позволило определить следующие приоритеты:

а) сценарии использования больших данных, определения, словари и эталонные архитектуры (например, система, данные, платформы, онлайн/офлайн и т.д.);
б) спецификации и стандартизация метаданных, включая их источники;
в) прикладные модели (например, пакетной обработки, потоковые и т.д.);
г) 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);
д) предметно-ориентированные языки;
е) семантика конечной согласованности данных (оптимистическая репликация);
ж) расширенные сетевые протоколы для эффективной передачи данных;
з) общие и предметно-ориентированные онтологии и таксономии для описания семантики данных, включая взаимодействие между онтологиями;
и) безопасность больших данных, управление доступом к персональным данным и их конфиденциальность;
к) удаленная, распределенная и федеративная аналитика, включая обнаружение данных, их извлечение и выявление ресурсов обработки;
л) совместное использование данных и обмен ими;
м) хранение данных, например, память, система хранения, распределенная файловая система, хранилище данных и т.д.;
н) использование результатов анализа больших данных (например, визуализация);
о) измерение энергозатрат для обработки больших данных;

п) интерфейс между реляционными (SQL) и не только реляционными (NoSQL) хранилищами данных;
р) качество и достоверность больших данных, описание и управление.

Стандарты ИСО/МЭК 20546 (ISO/IEC 20546) [3] и ИСО/МЭК 20547 (ISO/IEC 20547) были разработаны с учетом указанных приоритетов. В стандарте рассматриваются структура и процесс применения, сценарии использования больших данных и требования к ним (приоритет “а”), эталонные архитектуры (приоритет “а”), безопасность и конфиденциальность персональных данных (приоритет “и”), 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 ИСО/МЭК 20547 (ISO/IEC 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.

Стандарты серии ИСО/МЭК 20547 (ISO/IEC 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 стандарте ИСО/МЭК 20546 (ISO/IEC 20546) [3], 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 (ISO/IEC 20547), была достаточной для описания широкого диапазона потенциальных сценариев использования, реализуемых системами больших данных.

## ЭТАЛОННАЯ АРХИТЕКТУРА. ОСНОВНЫЕ ПОНЯТИЯ

Для понимания того, что включает эталонная архитектура, необходимо определить, что под ней подразумевается. Как указано в стандарте ГОСТ Р 57100–2016 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 стандарте ISO/TR 14639–2:2014. С учетом сказанного эталонная архитектура представляет собой структуру архитектуры, как это описано в стандарте ГОСТ Р 57100–2016, 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.

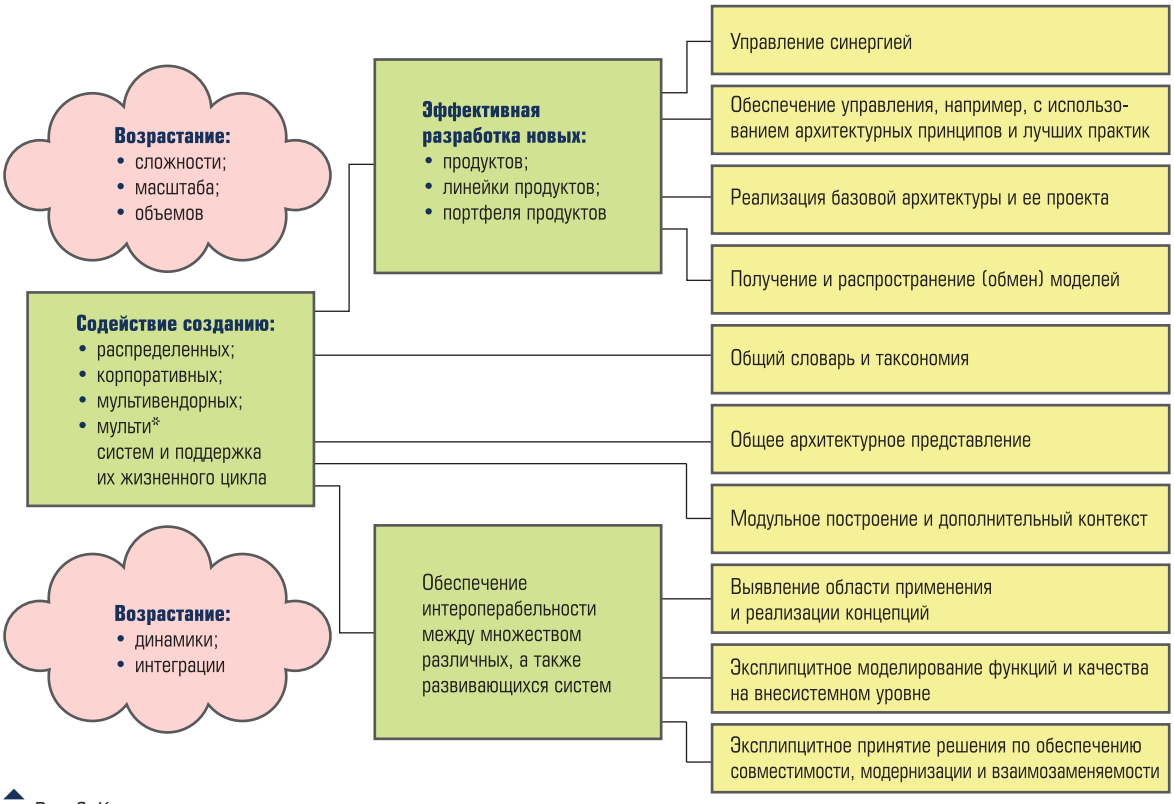



*Рис. 2. Концепция эталонных архитектур*

Эталонные архитектуры, как показано на рис. 2, разрабатываются для решения широкого круга задач [4]. Как утверждается в работе [4], основное предназначение эталонной архитектуры состоит в ориентации на будущее и использовании в качестве основы для будущих реализаций.

## СТРУКТУРА ЭТАЛОННОЙ АРХИТЕКТУРЫ

На рис. 3, в соответствии со стандартом ГОСТ Р 57100–2016, представлено объединение концепции и структуры эталонной архитектуры.

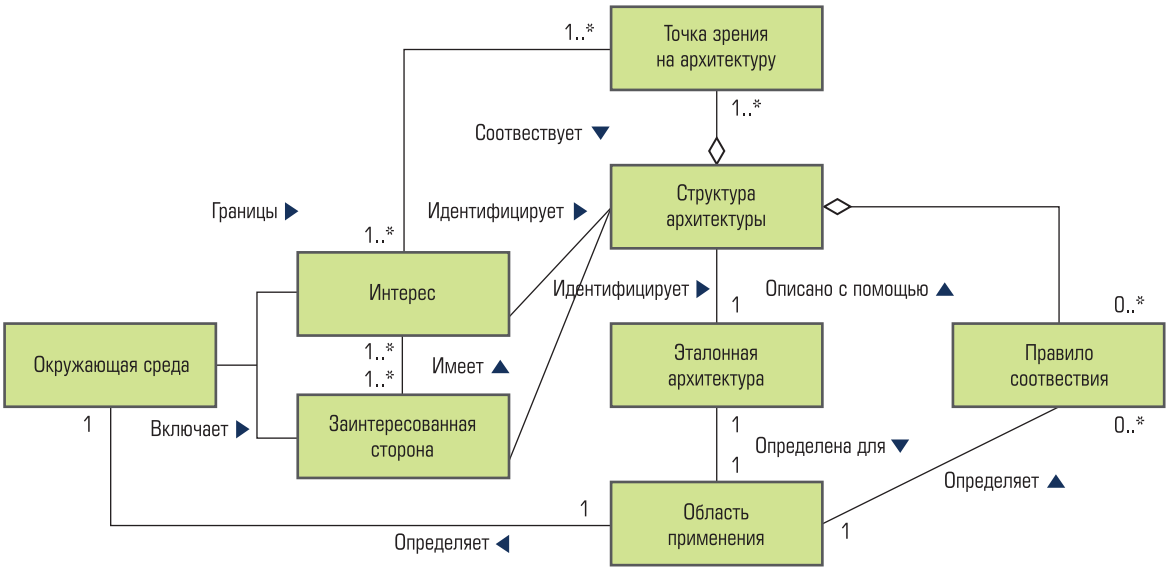



*Рис. 3. Схема взаимосвязей между базовыми понятиями эталонной архитектуры на основе ГОСТ Р 57100–2016*

Эталонная архитектура определяется для каждой конкретной области применения, а 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 [3]. 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 стандарте ИСО/МЭК 20547–3 (ISO/IEC 20547–3) [5] 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 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.

## ОСНОВЫ ЭТАЛОННОЙ АРХИТЕКТУРЫ БОЛЬШИХ ДАННЫХ

Стандарт содержит описание структуры эталонной архитектуры больших данных, предназначенной для использования в окружающей среде системы больших данных, сферы применения стандарта ИСО/МЭК 20547 (ISO/IEC 20547), логической взаимосвязи ГОСТ Р ИСО/МЭК ТО 20547-1-2021 15 между всеми частями стандарта и процесса применения эталонной архитектуры. Ниже представлено описание каждой части стандарта ИСО/МЭК 20547 (ISO/IEC 20547). ИСО/МЭК 20547–1 (ISO/IEC 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 (ISO/IEC 20547–2): “Сценарии использования и порождаемые ими требования” включает примеры описания сценариев использования больших данных в соответствии с областями применения и вытекающими из них техническими аспектами. ИСО/МЭК 20547–3 (ISO/IEC 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 (ISO/IEC 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 (ISO/IEC 20547–5): “Дорожная карта стандартов” включает описание стандартов (как существующих, так и разрабатываемых) в соответствии с анализом приоритетов разработки будущих стандартов, относящихся к большим данным. ГОСТ Р ИСО/МЭК ТО 20547-1-2021 16. На рис. 4 показаны взаимосвязи между частями стандарта ИСО/МЭК 20547 (ISO/IEC 20547).

В стандарте ИСО/МЭК 20547–2 (ISO/IEC 20547-2) представлены сценарии использования и технические аспекты на основе исследований научных сообществ, оценок экспертов и специалистов предприятий и организаций. В стандарте ИСО/МЭК 20547–3 (ISO/IEC 20547-3) представлена эталонная архитектура больших данных с соответствующими техническими аспектами. В стандарте ИСО/МЭК 20547–4 (ISO/IEC 20547–4) представлены аспекты обеспечения безопасности больших данных и конфиденциальности персональ-

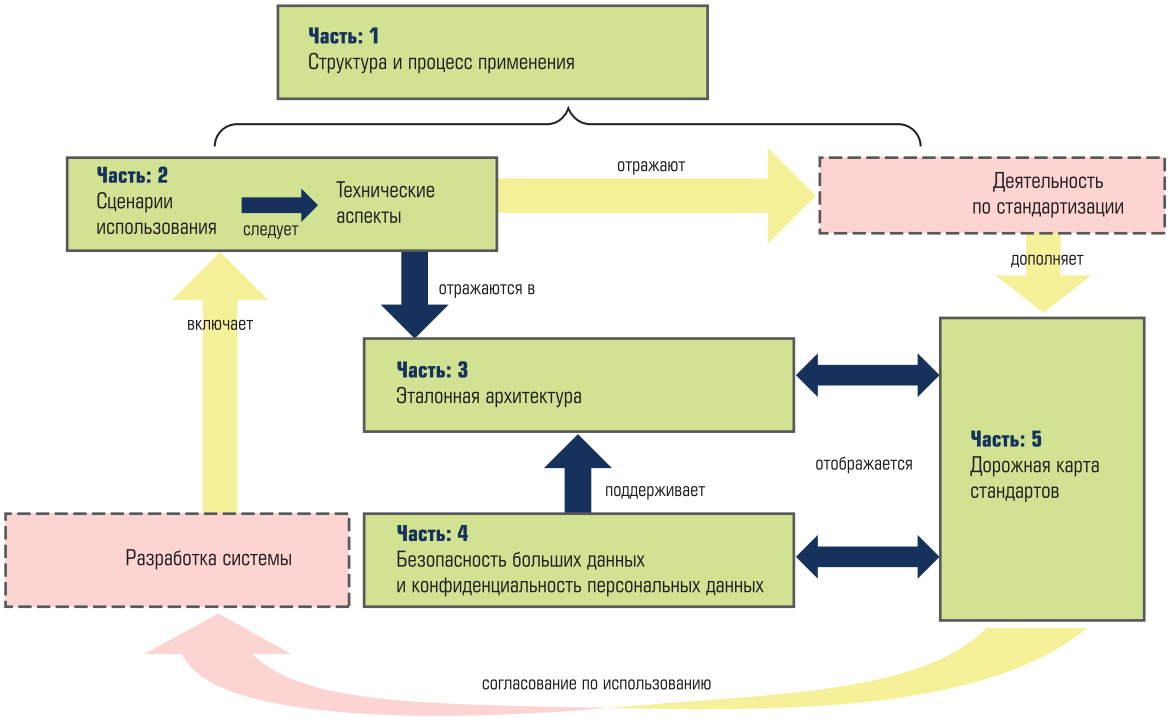



*Рис. 4. Взаимосвязи между частями стандарта ИСО/МЭК 20547*

ных данных. В стандарте ИСО/МЭК 20547-5 (ISO/IEC 20547-5) представлен список применяемых стандартов в рамках эталонной архитектуры больших данных. Для того чтобы применить рассматриваемую структуру в конкретных сценариях, необходимо иметь представление о ГОСТ Р ИСО/МЭК ТО 20547-1-2021 17 окружающей среде, в которой будет реализована система больших данных, заинтересованных сторонах и их интересах (проблемах). В параграфах 8.2-8.4 рассмотрены каждый из перечисленных ключевых аспектов.

## ЗАИНТЕРЕСОВАННАЯ СТОРОНА

Заинтересованная сторона в стандарте ГОСТ Р 57100–2016 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 стандарте ИСО/МЭК 20547–3 (ISO/IEC 20547–3) [5] под заинтересованными сторонами системы больших данных подразумеваются стороны в рамках представления пользователя.

## ИНТЕРЕС (ПРОБЛЕМА)

Любой интерес, актуальный для одной или нескольких заинтересованных сторон, как это представлено в стандарте ГОСТ Р 57100–2016, является проблемой, которая связана с техническими, функциональными, эксплуатационными, юридическими и даже социальными воздействиями на систему больших данных в окружающей среде. Окружающая среда системы определяется и ограничивается заинтересованными сторонами и их интересами (проблемами и задачами), связанными с этой системой. Некоторые из них могут быть представлены в терминах других международных стандартов.

Например, прозрачность распределения, представленная в эталонной модели открытой распределенной обработки в соответствии со стандартом ГОСТ Р ИСО/МЭК 10746-1-2004 [6], относится к одной из проблем, связанных с функционированием системы больших данных. Ключевыми аспектами таких систем

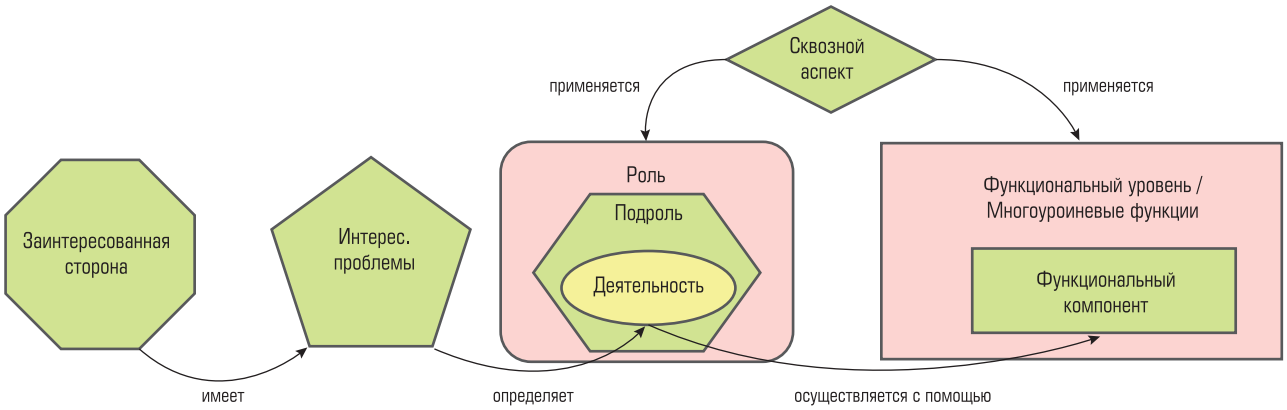



*Рис. 5. Взаимосвязи между элементами представлений эталонной архитектуры больших данных*

являются горизонтальное масштабирование и распределенная обработка.

Свойства программного обеспечения, такие как эффективность, производительность, доверие, риски и их снижение, а также гибкость, представленные в п. 4.2 (требования к оценке качества аппаратного и программного обеспечения) стандарта ГОСТ Р ИСО/МЭК 25010–2015 [7], 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.

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 посвящена отдельная часть стандарта ИСО/МЭК 20547–4 (ISO/IEC 20547–4) [8]. 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.

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

На рис. 5 представлена схема взаимодействия заинтересованных сторон, а также интересы (проблемы), связанные с указанными представлениями.

## 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 группе лиц. Стороны в экосистеме больших данных являются ее заинтересованными сторонами (stakeholder);
- **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. Сквозной аспект учитывается при совместном выполнении конкретных ролей или использовании функциональных компонентов.

Примером сквозных аспектов является безопасность больших данных. 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.

## 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.

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

## ПРОЦЕСС ПРИМЕНЕНИЯ ЭТАЛОННОЙ АРХИТЕКТУРЫ БОЛЬШИХ ДАННЫХ

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, включающие ИСО/МЭК/ИИЭР (ISO/IEC/IEEE) 20547–4 [9] для разработки систем, ГОСТ Р ИСО/МЭК (ISO/IEC) 12207–2010 [10] для описания жизненного цикла разработки систем. Качество процессов оценивается с помощью семейства стандартов ИСО 9000 (ISO 9000) [11-13] 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.

## ИДЕНТИФИКАЦИЯ ЗАИНТЕРЕСОВАННЫХ СТОРОН И ИХ ИНТЕРЕСОВ

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 со стандартом ИСО/МЭК/ИИЭР 15288 (ISO/IEC/IEEE 15288) [9]. В результате первого шага должны быть:

а) указаны необходимые характеристики и контекст использования системных сервисов;
б) определены ограничения на соответствующую систему;
в) определены соответствия требований заинтересованных сторон их потребностям;

г) определены базовые положения для формирования системных требований;
д) определены базовые положения для валидации системных услуг;
е) определены базовые положения для обсуждения и согласования условий поставки системных сервисов или продуктов.

В соответствии со стандартом ИСО/МЭК/ИИЭР 15288 (ISO/IEC/IEEE 15288) [9] 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 защиты персональных 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 стандарта ИСО/МЭК 20547–2 (ISO/IEC 20547-2) [14].

Заинтересованные стороны должны рассмотреть полученные результаты и подтвердить их точность, а также обоснованность требований и эффективность процесса проектирования. Отображение в ролях и подролях заинтересованных сторон и их интересов. Целью этого шага является отображение заинтересованных сторон и их проблем 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 стандарте ИСО/МЭК 20547–3 (ISO/IEC 20547–3) [5] определены следующие роли и подроли:

а) провайдер приложения больших данных (**BDAP**):
   1) сервис-провайдер комплектования больших данных (**BDCP**);
   2) сервис-провайдер предобработки больших данных (**BDPreP**);
   3) сервис-провайдер аналитики больших данных (**BDAnP**);
   4) сервис-провайдер визуализации больших данных (**BDVP**);
   5) сервис-провайдер доступа к большим данным (**BDAcP**);
б) провайдер среды обработки больших данных (**BDFP**):
   1) сервис-провайдер инфраструктуры больших данных (**BDIP**);
   2) сервис-провайдер платформы больших данных (**BDPlaP**);
   3) сервис-провайдер обработки больших данных (**BDProP**);
в) партнер сервиса больших данных (**BDSP**):
   1) разработчик сервиса больших данных (**BDSD**);
   2) аудитор больших данных (**BDA**);
   3) оркестратор системы больших данных (**BDSO**);
г) потребитель больших данных;
д) 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). Разработка подробных описаний деятельности и ее соответствие интересам. На данном шаге определяется деятельность системы или архитектуры. На основе эталонной архитектуры больших данных, представленной в категори-

ях 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 стандарта ИСО/МЭК 29148:2011 (ISO/IEC 29148:2011) "Системная и программная инженерия.

Процессы жизненного цикла. Разработка требований" [15]. Этот стандарт поможет трансформировать описанные выше проблемы в конкретные и тестируемые требования. Нормативный стандарт, устанавливающий соответствие операций/действий той или иной роли отсутствует, 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 обозначают субъект, модальность и глагол – предикат, а также объект – предмет, над которым выполняется действие. В 3-й части ИСО/МЭК 20547-3 (ISO/IEC 20547-3) [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.

## 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 но-

вые, 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.

## 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.

## Список литературы

1. *ISO/IEC JTC 1.* (2015) Big Data, Preliminary Report 2014. Geneva, Switzerland: International Organization for Standardization.
2. *IDC (2019)*. Worldwide Semiannual Big Data and Analytics Spending Guide. [Электронный ресурс]: URL: https://www.idc.com/getdoc.jsp?containerId=US4540711 (Дата обращения 25.09.2020).
3. *ISO/IEC 20546:2019* “Information technology – Big data – Overview and vocabulary”.
4. *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 [Электронный ресурс]: URL: doi: https://doi.org/10.1002/sys.20129 (Дата обращения 25.09.2020).
5. *ISO/IEC 20547-3:2020* “Information technology – Big data reference architecture – Part 3: Reference architecture”.
6. *ГОСТ Р ИСО/МЭК 10746-1-2004* Информационная технология. Открытая распределенная обработка. Базовая модель. Часть 1. Основные положения.
7. *ГОСТ Р ИСО/МЭК 25010-2015.* Информационные технологии. Системная и программная инженерия. Требования и оценка качества систем и программного обеспечения (SQuaRE). Модели качества систем и программных продуктов.
8. *ISO/IEC 20547-4* Information technology – Big data reference architecture – Part 4: Security and privacy.
9. *ГОСТ Р 57193-2016.* Системная и программная инженерия. Процессы жизненного цикла систем.
10. *ГОСТ Р ИСО/МЭК 12207–2010.* Информационная технология. Системная и программная инженерия. Процессы жизненного цикла программных средств. ГОСТ Р ИСО/МЭК ТО 20547-1-2021
11. *ISO 9000:2015.* Quality management systems – Fundamentals and vocabulary.
12. *ISO 9001:2015.* Quality management systems – Requirements.
13. *ISO/TS 9002:2016.* Quality management systems – Guidelines for the application of ISO 9001:2015.
14. *ISO/IEC 20547-2.* “Information technology – Big data reference architecture – Part 2: Use cases and derived requirements”.
15. *ISO/IEC 29148:2011.* Systems and software engineering – Life cycle processes – Requirements engineering).

***Егоров Александр Александрович** – канд. техн. наук, профессор АВН РФ, главный редактор журнала “Автоматизация и IT в нефтегазовой области”.*