МНОГОУРОВНЕВАЯ КОНЦЕПЦИЯ БЕЗОПАСНОСТИ СИСТЕМ УПРАВЛЕНИЯ БОЛЬШИМИ ДАННЫМИ

Полтавцева М. А.[1], Зегжда Д. П.[2], Калинин М.О.[3]

Цель исследования. *Технологии и системы управления большими данными являются основой огромного числа современных цифровых сервисов. С одной стороны, они построены на традиционных решениях, а с другой, включают новые подходы, такие, как полихранилища или аутсорсинг данных. Ключевая роль в технологическом стеке цифровой экономики и новизна определяют как привлекательность таких активов для злоумышленника, так и несовершенство методов защиты. Целью работы является анализ больших данных как объекта защиты и разработка многоуровневой концепции их безопасности на основе консистентного подхода.*

Метод исследования. *В работе используется многоуровневый подход, которому соответствует также архитектура ANSI/SPARC систем управления базами данных. Большие данные рассматриваются на трех уровнях от инфраструктуры до бизнес-логики, выделяются ключевые технологии, уязвимости и методы защиты. Также более детально в рамках ANSI/SPARC определяется уровневая архитектура систем управления большими данными на базе полихранилищ, проводится анализ их безопасности. Задается технологический базис безопасности систем управления большими данными как система распределенного динамического аудита, приводится пример такой системы на основе распределенного реестра.*

Результаты исследования. *В статье выделены три уровня рассмотрения больших данных: инфраструктурный, инженерии данных и бизнес-логики. Авторами сформулированы эволюционные изменения систем больших данных по сравнению с традиционными СУБД с точки зрения информационной безопасности. Дано понятие системы управления большими данными, определены ее собственные архитектурные уровни на основе архитектуры ANIS/SPARK, для каждого из которых выделены проблемы безопасности, причины их появления и направления развития средств защиты. Авторами выделено ключевое требование безопасности систем управления большими данными - консистентное представление на уровне общего монитора безопасности. Для его выполнения, в части сбора данных о системе, предложено использование технологий распределенного динамического аудита. Проведена апробация системы распределенного динамического аудита при управлении большими данными на базе технологии HashGraph.*

Научная новизна. *В работе впервые сформулирована многоуровневая концепция безопасности систем управления большими данными, в рамках которой выделены и систематизированы на различных уровнях ключевые уязвимости систем больших данных, отличные от других классов систем и традиционных СУБД. Впервые предложено применение технологии распределенного реестра для сбора данных о жизненном цикле информации в системе управления большими данными. Проведенные исследования позволяют более комплексно подойти к обеспечению безопасности больших данных и систем управления ими, конкретизировать и согласовать наборы методов и средств защиты, а также закладывают основу построения таких систем в защищенном исполнении.*

Ключевые слова: *информационная безопасность, безопасность больших данных, консистентный подход, архитектура безопасности, модель безопасности, безопасность полихранилищ, распределенный реестр.*

DOI: 10.21681/2311-3456-2023-5-25-36

Введение

Технологии «больших данных» в начале века во многом изменили ландшафт и архитектуру современных информационных систем. Цифровая экономика, цифровое производство, электронное правительство, искусственный интеллект – все современные цифровые системы в большей или меньшей степени осно-

1 Полтавцева Мария Анатольевна, доктор технических наук, доцент, профессор СПбПУ Петра Великого, г. Санкт-Петербург, Россия. E-mail: poltavtseva@ibks.spbstu.ru ORCHID 0000-0001-9659-1244

2 Зегжда Дмитрий Петрович, член–корреспондент РАН, доктор технических наук., профессор, профессор СПбПУ Петра Великого, г. Санкт-Петербург, Россия. E-mail: dmitry@ibks.spbstu.ru ORCHID 0000-0002-2048-6189

3 Калинин Максим Олегович доктор технических наук, профессор, профессор СПбПУ Петра Великого, г. Санкт-Петербург, Россия. E-mail: max@ibks.spbstu.ru ORCHID 0000-0002-9732-0099

Рис.1. Уровни представления больших данных

ваны на данных. Для многих из них подходы на основе данных (data-driven) являются ключевыми. В то же время, как росла зависимость цифровых сервисов от данных, росло и их влияние на жизнь каждого отдельного человека. И также росло число злоумышленников и атак в киберпространстве.

Безопасность современных цифровых сервисов напрямую зависит, в том числе, и от безопасности обрабатываемых в них данных, которые не только используются для предоставления сервиса или получения результата, услуги, но и для создания и функциональности самих цифровых решений. Например, для обучения искусственного интеллекта. Огромные массивы персональных данных в информационных системах, как государственных, так и частных, востребованы злоумышленниками. Результаты утечек информации используются как для простого получения выгоды, рекламы и перепродажи, так и во множестве мошеннических схем, проведении OSINT – исследований с различными целями.

Безопасность современных больших данных, таким образом, является одной из важных современных задач в области кибербезопасности. Несмотря на то, что исследования в этой области перешагнули десятилетней рубеж[4], проблематика сегодня остается прежней [1]. Целью данной работы является, в первую очередь, анализ больших данных и систем управления ими как объекта защиты, выявление основных технологических проблем обеспечения их защищенности на современном этапе а также разработка многоуровневой концепции безопасности на основе консистентного подхода в данной области.

4 Запечников С. В. и др. Проблемы обеспечения информационной безопасности больших данных //Безопасность информационных технологий. 2014. Т. 21.(3). С. 8-17

Большие данные как объект защиты

Несмотря на более чем десятилетнюю историю, устоявшегося и общепринятого понятия больших данных до сих пор не существует [2]. С одной стороны, проблема обработки больших объемов данных, сегодня ассоциированная с big data, была обозначена еще в середине прошлого века [2,3]. Хотя взрывной интерес к этой области и появление современной терминологии отмечается с начала двадцать первого века [4], технологические основы и вызовы, обусловленные проблемами разнородности, объема и скорости данных на уровне систем управления базами данных (СУБД) возникали и ранее, имеют не абсолютно новый, а исторический характер. С другой стороны, сегодня большие данные касаются не только технологических вызовов разработчикам СУБД и инженерам данных, но и формируют новые вызовы для организаций в целом, правовой системы, государственного управления.

Качественным изменением, по сравнению с проблемами, приведшими к появлению методов оптимизации запросов, параллельным и распределенным системам баз данных, ETL (extract, transfer, loading)-технологии является переход проблематики с уровня технических систем на организационный. Большие данные пытаются определять не только в технических терминах, но и на уровне бизнес-процессов и нормативных документов [6]. В отечественном стандарте определение также достаточно расплывчато[5]. В ито-

5 «Большие данные (big data): большие массивы данных, главным образом, по таким характеристикам данных, как объем, разнообразие, скорость обработки и/или вариативность, – которые требуют использования технологии масштабирования для эффективного хранения, обработки, управления и анализа.» ГОСТ Р ИСО/МЭК 20546–2019 Большие данные. Обзор и словарь

ге сегодня можно говорить о, по крайней мере, трех уровневой архитектуре больших данных, на каждом уровне которой существуют свои понятия, технологии, угрозы, уязвимости и методы защиты (рисунок 1).

На самом нижнем, инфраструктурном уровне сегодня сложился ряд технологий, которые также называют технологиями больших данных. Это облачные и туманные (fog) вычисления [6], центры обработки данных [7,8] и другие связанные технологии [9]. Безусловно, фактически этот уровень представляет собой общую инфраструктуру современных распределенных информационных систем и применим не только для больших данных. Однако, говоря о технологическом стеке и безопасности больших данных нельзя его проигнорировать, так как традиционные угрозы от DDoS атак до программных и аппаратных закладок актуальны для систем больших данных именно на этом уровне.

Следующий уровень относится к области инженерии данных (data engineering) и является продолжением технологий традиционных систем управления базами данных [10]. На этом уровне специалистами рассматриваются вопросы структуризации данных, построения специализированных СУБД, хранилищ, новые архитектуры обработки информации [1]. Здесь же используется и самое устоявшееся определение данного явления: под большими данными понимается информация, которая, в силу своих характеристик (скорости поступления, объема и/или разнообразия), не может быть обработана общераспространенными, “универсальными” средствами. А фактические значения характеристик данных могут различаться от задачи к задаче.

Ключевой новой технологией на этом уровне сегодня становятся полихранилища – системы, объединяющие в себе несколько СУБД на базе различных моделей данных с разными операциями и степенью структуризации информации [11,12]. Технологии безопасности традиционных СУБД сегодня достаточно развиты, хотя и требуют совершенствования, и, к сожалению, далеко не всегда поддерживаются серверами [1,13,14]. Но эти технологии не являются легко переносимыми на полихранилища, так как тесно связаны со структуризацией, грануляцией и операциями с данными внутри каждой конкретной системы управления базами данных.

Третий уровень рассмотрения больших данных во многом определяет уникальность этого явления по сравнению с проблемами роста объема и разнообразия ранее. Большие данные рассматриваются как ценный актив на уровне предприятия [15] и к ним применяются подходы управления, на их основе строятся data-driven (основанные на данных) системы принятия решений [16]. В рамках “цепочки поставок” существует несколько различных способов использования или работы с большими данными [17]. Ключевыми здесь являются как внешние источники данных, так и задачи обмена данными между организациями, передачи данных для анализа на аутсорсинг, совместное использование данных для обучения алгоритмов искусственного интеллекта. Сохранить конфиденциальность данных при этом становится достаточно сложной задачей [18], для решения которой используются криптографические технологии [19], технологии анонимизации [20], федеративное обучение искусственного интеллекта [21]. Тем не менее, все эти технологии сегодня не могут гарантировать безопасность больших данных, а только снижают вероятность их утечки. Попытки формализации комплексной безопасности на этом уровне [22] пока не получили широкого развития.

Систематизация описанных выше уровне с точки зрения кибербезопасности, от ключевых новых технологий больших данных, до эволюции угроз и принятых сегодня методов защиты вместе с ограничениями приведены в табл. 1.

В силу того, что в рамках больших данных можно выделить три технологических уровня, рационально говорить о комплексной безопасности в этой области как о согласованной системе безопасности, охватывающей все уровни представления от инфраструктурного до бизнес-логики. В то же время, так как приведенные уровни практически независимы друг от друга, что подчеркивается, в том числе, разницей в понятийном аппарате, реализация мер защиты на каждом из них также может осуществляться независимо.

В дальнейшей работе будет более детально рассмотрен уровень инженерии данных, как, с одной стороны, высоко технологичный (в отличии от уровня бизнес-логики), а с другой специфический для рассматриваемой области, в отличии от инфраструктурного. К тому же, приведенное выше определение больших данных из ГОСТ Р ИСО/МЭК 20546-2019 также относится в наибольшей степени к уровню инженерии.

Безопасность систем управления базами данных и полихранилищ в экосистеме больших данных

Первым шагом в сторону развития современных технологий и архитектур инженерии больших данных, после появления распределенных СУБД, стало появ-

Таблица 1

Систематизация характеристик уровней представления больших данных с точки зрения кибербезопасности

Параметр	Уровень бизнес-логики	Уровень инженерии данных	Уровень инфра-структуры
Ключевые изменения	Совместное использование данных. Передача данных на аутсорсинг.	Полихранилища (поли-базы данных или гетерогенные базы данных)	Распределенная инфраструктура обработки информации
Эволюциони-ровавшие угрозы	Утечки данных (внутренний нарушитель). Логический вывод над данными (inference attack)	Утечки данных (внутренний и внешний нарушитель, ошибки контроля доступа), искажение и удаление данных.	Не доверенная среда обработки
Технологии защиты	Анонимизация данных. Платформы «безопасной» аналитики. Федеративное обучение искусственного интеллекта.	Контроль доступа. Шифрование. Аудит и журналирование. Обнаружение вторжений и внутреннего нарушителя.	Безопасность облачных технологий (криптография, безопасные вычисления и др). Безопасное ПО.
Ограничения	Нет гарантированных способов защиты от атак логического вывода.	Многие технологии существуют только для отдельных типов СУБД и часто даже не внедрены в промышленные решения.	Сложные цепочки поставок, большие объемы анализируемого кода.

ление не реляционных (NoSQL) систем управления базами данных и специализация в области СУБД. В итоге для решения задач, выходящих за рамки возможностей промышленного сервера баз данных, требовалось составление комбинаций из разнородных инструментов. Такие системы в разных источниках называются полихранилищами [23], поли-базами данных [24], гетерогенными базами данных[25], гетерогенными системами баз данных[26], системами управления большими данными [27]. Помимо полихранилищ в экосистему инженерии больших данных входят также еще два класса программного обеспечения. Это, во-первых, инструменты потоковой обработки данных и, во-вторых, программы и библиотеки, отвечающие за балансировку нагрузки, преобразования данных и выполняющие другие вспомогательные задачи. Все эти инструменты в совокупности с полихранилищами можно назвать системами управления большими данными.

Рассмотрение полихранилищ с точки зрения теории управления базами данных также позволяет выделить в них уровни обработки информации. Согласно базовой архитектуре всех систем управления базами данных ANSI/SPARC, не потерявшей сегодня своей актуальности [28], выделяется три уровня:

— физический уровень, на котором осуществляется хранение и физические операЗции с данными на диске;
— логический уровень, на котором определяется внутреннее представление данных (модель данных), которое используется для манипулирования информацией в терминах не файловой системы, а семантически значимых фрагментов;
— концептуальный уровень, на котором формируются представления данных для пользователей и внешних программ.

Полихранилища имеют организацию, отличную от систем управления базами данных, так как, фактически, включают в себя несколько СУБД с различной логической (и тем более физической) организацией данных. Для больших данных уровни модели ANSI/SPARC предлагается интерпретировать следующим образом (рисунок 2).

В данной интерпретации на физический уровень выносятся традиционные СУБД, выступающие как «коробочные» решения по управлению поступающей в них информацией. Безусловно, каждый такой инструмент также может (и должен) при построении системы безопасности оцениваться на приведенных выше архитектурных уровнях. Но, с точки зрения полихранилища в целом, манипулирование данными внутри такого инструмента не интерпретируемо.

Логический уровень представляет собой комбинацию структур данных в рамках каждого инструмента. Используемые внутренними СУБД полихранилища

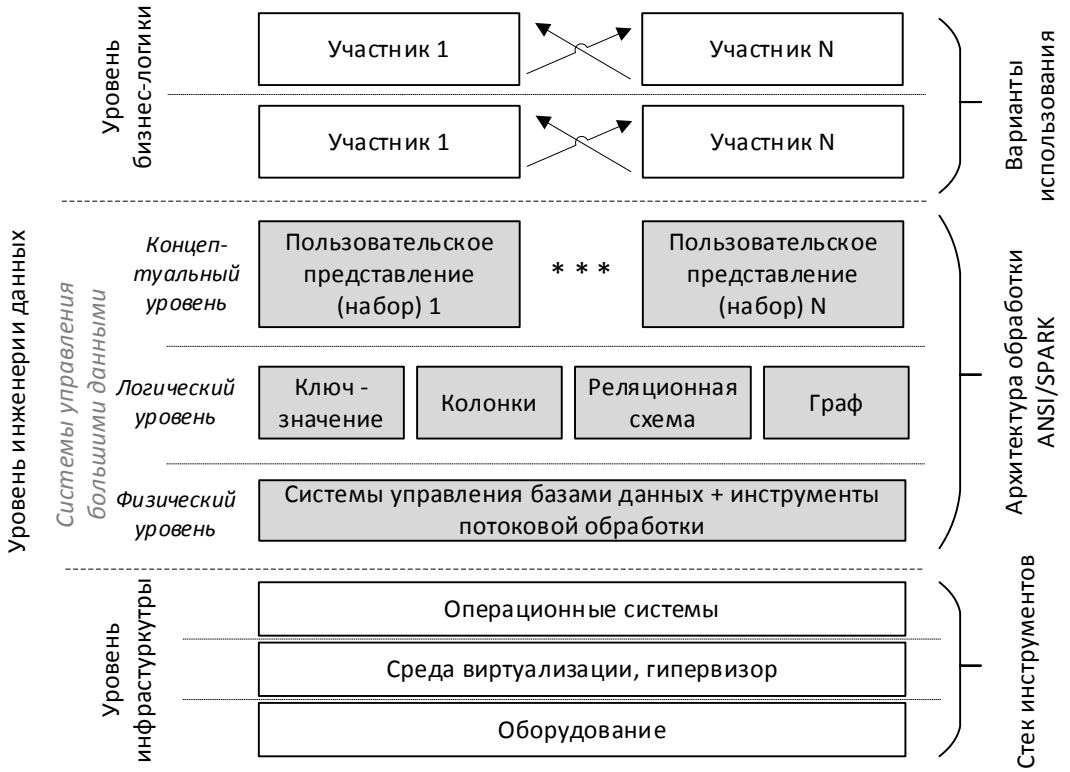

Рис.2. Уровни представления инженерии больших данных в концепции ANSI/SPARC

структуры данных должны быть оценены и согласованы для того, чтобы было возможно управление данными в рамках системы в целом. Такая задача возникла сразу при объединении нескольких систем с разной структуризацией и решается сегодня методами согласования моделей данных [29]. В то же время, на этом уровне, в отличии от СУБД еще рано говорить о едином представлении данных для всего хранилища, так как манипулирование информацией на логическом уровне в каждой модели осуществляется по-разному.

Концептуальным уровнем является сегодня представление больших данных полихранилища и всей системы управления данными для различных пользователей. Причем пользователями в этом случае определяются не только на уровне бизнес-логики, но и, например, ими являются подсистемы обеспечения безопасности верхнего уровня, реализующие политику безопасности на уровне хранилища в целом.

Систематизация объектов различных уровней систем управления большими данными, а также угроз и причин их возникновения приведена в табл. 2. В первую очередь рассматриваются полихранилища, так как именно эта ключевая технология отличает системы управления большими данными от всех других типов распределенных систем.

С появлением полихранилищ ландшафт угроз на уровне инженерии данных практически не изменился [13], как и основные методы защиты [14]. Основной проблемой является согласование реализации политики безопасности между различными инструментами. В чем-то это проблема сродни разработке и реализации единой политики безопасности для разнородных распределенных операционных систем (ОС) и информационных систем [30,31], однако разнородность структуризации данных разных СУБД на логическом уровне в составе одной системы не позволяет напрямую перенести известные практики.

Рассматривая концепцию безопасности систем управления большими данными как многоуровневую архитектуру безопасности, определим функции безопасности каждого уровня.

Физический уровень является технологической основной остальных операций с данными. Основными требованием к нему является безопасность и доверие к среде обработки данных, отсутствие программных закладок, несанкционированного доступа,

Таблица 2

Систематизация объектов различных уровней систем управления большими данными (полихранилищ), угроз безопасности и причин их возникновения

Уровни представления	Физический	Логический	Концептуальный
Объекты	Отдельные СУБД и инструменты обработки	В соответствии с моделью данных	В соответствии с моделью представления
Угрозы	Эксплуатация уязвимостей отдельных инструментов	Ошибки разграничения и контроля доступа, отсутствие аудита и оценки защищенности	Несогласованность политики безопасности верхнего уровня и реализованных ниже политик
Причины возникновения угроз	Аналогичны «классическим» СУБД	Разнородность структуризации и грануляции данных	Динамичность данных, сложный жизненный цикл данных

уязвимостей программных компонентов. Также на этом уровне для защиты от внутреннего нарушителя, которым может быть администратор системы, применима концепция нулевого доверия и безопасности операций. Аудит операций внутри инструментов обработки данных также относится к этому уровню.

Логический уровень определяет грануляцию данных и операции манипулирования с ними на более высоком уровне. Основное требование к безопасности логического уровня это согласованный контроль доступа между инструментами обработки данных и аудит операций с данными между инструментами обработки данных.

Концептуальный уровень определяет безопасность в отношении каждого набора данных, предоставляемых пользователям. С точки зрения защиты информации на этом уровне применимы технологии, использующиеся для технической защиты наборов данных с точки зрения бизнес-логики в системе больших данных в целом. Это защита данных при передаче их на аутсорсинг, в частности, методы анонимизации данных и защита от логического вывода (inferrence attack [32]).

На уровне систем управления большими данными уровни уже являются менее независимыми, чем в более высокоуровневом рассмотрении, и согласованность методов и средств безопасности между ними является ключевым требованием. В данной концепции логический уровень является связующим между представлениями данных для пользователей и физическим манипулированием ими, аналогично уровню инженерии данных в общей многоуровневой концепции больших данных и логическому уровню архитектуры ANSI/SPARC в СУБД. Сегодня основной проблемой, обуславливающей угрозы логического и концептуального уровней в таблице 2, является несогласованность данных в рамках инструментов физического уровня (и их логических моделей), и, как следствие:

- отсутствие единого представления данных в рамках системы в целом (в том числе, для реализации согласованного контроля доступа);
- отсутствие единой системы аудита данных не только внутри инструментов обработки, но и между узлами на которых они функционируют (для распределенных инструментов) а также между инструментами;
- отсутствие единой системы контроля реализации политики безопасности, включая защиту от внутреннего нарушителя.

Решение этих проблем является ключевым шагом для достижения безопасности в классе систем управления большими данными и в рамках безопасности больших данных в боле широкой интерпретации.

Технологический базис безопасности систем управления большими данными

Таким образом, ключевым требованием безопасности систем управления большими данными является консистентное представление на уровне общего монитора безопасности. Для этого необходимо решить целый ряд задач, таких, как построение общей концептуальной модели данных над всеми логическими моделями инструментов обработки данных (по крайней мере, для решения задач безопасности) и обеспечение согласованного аудита операций на физическом и логическом уровне. Решение этих задач позволит подойти к решению и проблемы в целом, включая контроль реализации политики безопасности, оценку защищенности и другие вопросы.

Построение обшей концептуальной модели данных является сложной научной задачей и заслуживает

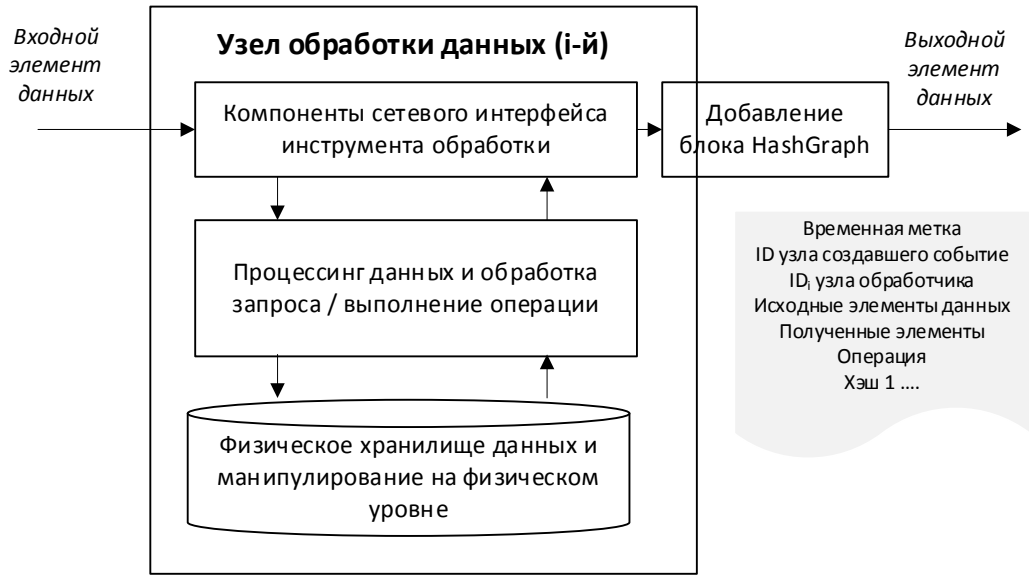

Рис.3. Введение распределенного реестра в систему управления большими данными

отдельного рассмотрения. Задача аудита представляется более простой, так как для большинства отдельных СУБД с точки зрения внутренних операций она уже решена [33], и даже для тех СУБД, в которых нет встроенных модулей аудита, такое решение несложно построить [14]. Основной проблемой остается аудит между инструментами данных и узлами обработки информации. Решение уже этой задачи сегодня может быть построено с использованием технологий распределенного реестра, в частности HashGraph [34]. В таком случае распределенный реестр используется для ведения информации об операциях с данными между узлами (рисунок 3). Под элементом данных понимается не отдельный пакет, а высокоуровневый семантически значимый фрагмент: банковская транзакция, отчет о продажах, запись файла лога и т.д.

В результате анализа цепочек блоков распределенного реестра, полученных в результате анализа прохождения элементов данных в системе, возможен полноценный аудит данных на протяжении цикла жизненного цикла. К тому же, защищенный от подделок со стороны внутреннего нарушителя – администратора системы. За счет высокого уровня абстрагирования от отдельных пакетов на концептуальный уровень семантически значимой информации при таком подходе можно получить достаточно хорошие показатели производительности. Это критически важно, так как именно баланс между производительностью и защищенностью – классическая дилемма систем управления данными (и базами данных).

Пример задержки при внедрении распределенного реестра для системы управления большими данными с достаточно маленькими отдельными элементами, анализирующей трафик для решения задач сетевой безопасности, приведен на рисунке 4. Для систем с семантически значимыми фрагментами большего размера задержка будет еще меньше.

При интеграции этой технологии с существующими журналами и системами аудита отдельных инструментов обработки данных и СУБД формируется комплексная система аудита, позволяющая проследить жизненный цикл каждого фрагмента данных. На основе графов жизненного цикла при наличии общей математической модели описания данных уже может быть согласована политика доступа, настроены параметры разграничения доступа и потом транслированы на уровень инструментов, а также – может проводиться мониторинг, анализ и оценка всей системы в целом.

Заключение

Использование многоуровневого подхода к безопасности сложных современных систем больших данных, как и их компонентов – систем управления большими данными, полихранилищ, позволяет системно взглянуть на задачу обеспечения их защищенности, выделить сходные технологии и методологии в смежных областях для каждого узкого круга задач, выявить проблемные области и искать пути решения проблем.

Сложность обеспечения безопасности больших данных во многом определяется широтой и комплексностью этого понятия. На каждом из трех выделенных в работе основных уровней представления: ин-

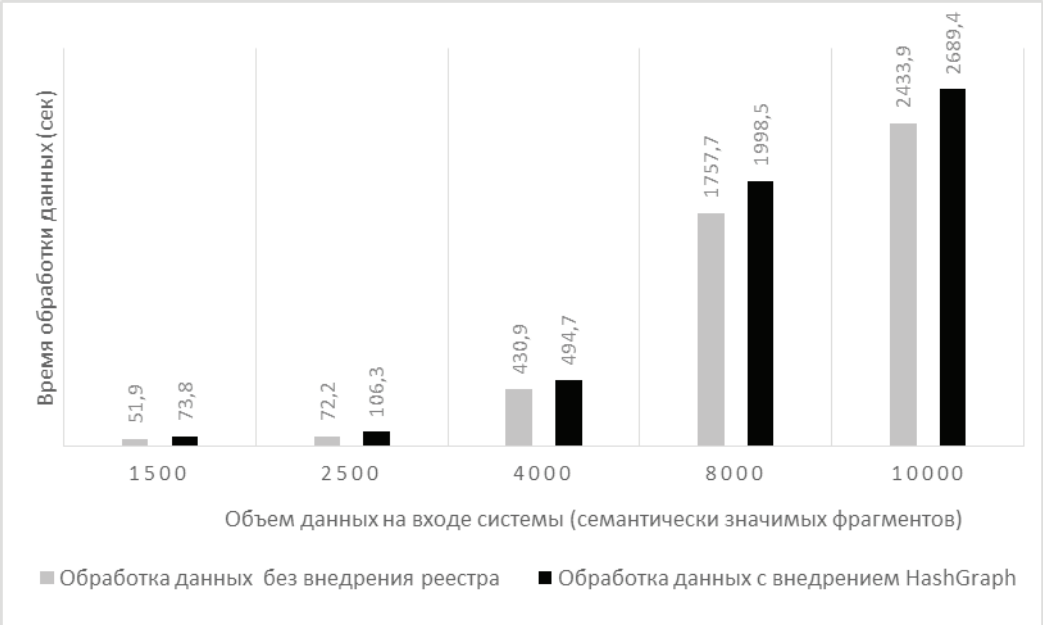

Рис.4. Время обработки данных в системе до введения распределенного реестра и после

фраструктурном, инженерии данных и бизнес-логики не только отличаются технологии, методы и средства работы с большими данными, но и даже само понятие. Каждому уровню присущи свои уязвимости, угрозы и развивающиеся сегодня направления защиты.

Уровень инженерии данных является ключевым технологическим уровнем, отличающим системы управления большими данными от других классов систем. Основными компонентами этого уровня являются разнородные системы управления базами данных, или полихранилища, дополняемые инструментами других классов. Экстраполируя на него принципы построения систем управления данными архитектуры ANSI/SPARC, можно говорить также о трех уровнях: физическом, логическом и концептуальном. Эти уровни коррелируют аналогичными с уровнями традиционных СУБД. Основным отличием является тот факт, что физический уровень включает в себя полностью реализации отдельных инструментов работы с данными.

Ключевые проблемы безопасности систем управления большими данными связаны как раз с разнородностью компонентов полихранилищ, а точнее, с отсутствием единого представления данных на общем логическом уровне, вместо гетерогенных структур данных различных СУБД. С другой стороны, существует проблема организации распределенного аудита на уровне системы управления большими данными в целом, а не, опять же, отдельных инструментов. Сочетание этих двух технологий позволит подойти к задаче разработки и практической реализации средств защиты, таких как компоненты контроля и разграничения доступа, анализа и мониторинга, аудита, оценки защищенности, форензики для систем управления большими данными. Использование технологии распределенного реестра, в частности – HashGraph, позволяет обеспечить эффективный аудит данных между инструментами системы и сделать шаг к решению всей комплексной проблемы.

Исследование выполнено за счет гранта Российского научного фонда № 23-11-20003, https://rscf.ru/project/23-11-20003/, грант Санкт-Петербургского научного фонда (Соглашение №23-11-20003 о предоставлении регионального гранта).

Литература

1. Naeem M. et al. Trends and future perspective challenges in big data //Advances in Intelligent Data Analysis and Applications: Proceeding of the Sixth Euro-China Conference on Intelligent Data Analysis and Applications, 15–18 October 2019, Arad, Romania, Springer Singapore, 2022. pp. 309–325.
2. Корнев М. С. История понятия “большие данные” (Big Data): словари, научная и деловая периодика // Вестник РГГУ. Серия: Литературоведение. Языкознание. Культурология. 2018. №1 (34). С. 81-85
3. Otto B. The evolution of data spaces // Designing Data Spaces: The Ecosystem Approach to Competitive Advantage. Cham: Springer International Publishing. 2022. pp. 3-15.
4. Gupta D., Rani R. A study of big data evolution and research challenges //Journal of information science. 2019. Vol. 45. Is 3. pp. 322-340
5. Антипова К. Г. Способы определения больших данных: Российский и зарубежный опыт // Юридические исследования. 2021. № 9. С. 143–157. doi: 10.25136/2409-7136.2021.9.36591
6. Badidi E., Mahrez Z., Sabir E. Fog computing for smart cities’ big data management and analytics: A review // Future Internet. 2020. Vol. 12. No. 11. pp.1-28. doi: 10.3390/fi12110190
7. Wang J. et al. Big data service architecture: a survey // Journal of Internet Technology. 2020. Vol. 21. No 2. pp. 393-405.
8. Bhattarai B. P. et al. Big data analytics in smart grids: state of the art, challenges, opportunities, and future directions //IET Smart Grid. 2019. Vol. 2. No. 2. pp. 141-154. doi: 10.1049/iet-stg.2018.0261
9. Ndikumana A. et al. Joint communication, computation, caching, and control in big data multi-access edge computing //IEEE Transactions on Mobile Computing. 2019. Vol. 19. No 6. pp. 1359-1374. doi: 10.1109/TMC.2019.2908403.
10. Vogt M. et al. Polystore Systems and DBMSs: Love Marriage Marriage of Convenience? //Heterogeneous Data Management, Polystores, and Analytics for Healthcare: VLDB Workshops, Poly 2021 and DMAH 2021, Virtual Event, August 20, 2021, Revised Selected Papers 7. – Springer International Publishing, 2021. pp. 65-69.
11. Lu J., Holubová I., Cautis B. Multi-model databases and tightly integrated polystores: Current practices, comparisons, and open challenges //Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018. pp. 2301-2302.
12. Gobert M. Design, Manipulation and Evolution of Hybrid Polystores. [электронный ресурс] 2023. https://pure.unamur.be/ws/portalfiles/portal/74437131/2023_GobertM_these.pdf (дата доступа 01.08.2023)
13. Poltavtseva, M. A. Evolution of Data Management Systems and Their Security // Proceedings - 2019 International Conference on Engineering Technologies and Computer Science: Innovation and Application, EnT 2019, Moscow, 26–27 March 2019. – Moscow, 2019. pp. 25-29. doi: 10.1109/EnT.2019.00010. EDN AGZBGD.
14. Полтавцева, М. А. Безопасность баз данных / Санкт-Петербург: Федеральное государственное автономное образовательное учреждение высшего образования “Санкт-Петербургский политехнический университет Петра Великого”, 2023. 143 с. EDN RICQEN.
15. Титаренко, Д. В., Исмайлов Э. И. Безопасность больших данных // Проблемы информационной безопасности социально-экономических систем: VII Всероссийская с международным участием научно-практическая конференция, Гурзуф, 18–20 февраля 2021 года. – Симферополь: Крымский федеральный университет им. В. И. Вернадского, 2021. С. 121–122. EDN NFQQHT.
16. Скворцов Н. Константинов А., Кузнецов С. Ценность ваших данных / Москва: Альпина ПРО, 2022. – 750 с.
17. Ogbuke N. J. et al. Big data supply chain analytics: ethical, privacy and security challenges posed to business, industries and society // Production Planning & Control. 2022. vol. 33. №. 2-3. С. 123–137. doi: 10.1080/09537287.2020.1810764
18. Binjubeir M. et al. Comprehensive survey on big data privacy protection //IEEE Access. 2019. vol. 8. pp. 20067-20079. doi: 10.1109/ACCESS.2019.2962368
19. Madan S., Bhardwaj K., Gupta S. Critical analysis of big data privacy preservation techniques and challenges //International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Vol. 3. – Springer Singapore, 2022. – pp. 267-278. doi: 10.1007/978-981-16-3071-2_23
20. Mehta B. B., Rao U. P. Improved l-diversity: scalable anonymization approach for privacy preserving big data publishing //Journal of King Saud University-Computer and Information Sciences. 2022. vol. 34. Is. 4. pp. 1423-1430. doi: 10.1016/j.jksuci.2019.08.006
21. Dhiman G. et al. Federated learning approach to protect healthcare data over big data scenario //Sustainability. 2022. vol. 14. Is 5. pp. 1-14. doi: 10.3390/su14052500
22. Статьев В. Ю., Докучаев В. А., Маклачкова В. В. Информационная безопасность на пространстве” Больших данных” //T-Comm-Телекоммуникации и Транспорт. 2022. Т. 16. №. 4. С. 21-28.
23. Poudel M. et al. Development of a polystore data management system for an evolving big scientific data archive //Heterogeneous Data Management, Polystores, and Analytics for Healthcare: VLDB 2019 Workshops, Poly and DMAH, Los Angeles, CA, USA, August 30, 2019, Revised Selected Papers 5. – Springer International Publishing, 2019. pp. 167-182. doi: 10.1007/978-3-030-33752-0_12
24. Ecy M.T, Вальдуриес П. Принципы организации распределенных бах данных. – М.: ДМК Пресс. – 2021. 678с.
25. Fong J. S. P. et al. Heterogeneous Database Connectivity //Information Systems Reengineering, Integration and Normalization: Heterogeneous Database Connectivity. 2021. pp. 317-367. doi: 10.1007/978-3-030-79584-9_9
26. Abdennebi A. et al. Machine learning based load distribution and balancing in heterogeneous database management systems // Concurrency and Computation: Practice and Experience. 2022. vol. 34. Is 4. pp. 1-13. doi: 10.1002/cpe.6641
27. Kim T. et al. Similarity query support in big data management systems //Information Systems. 2020. vol. 88.pp. 1-61. doi: 10.1016/j.is.2019.101455
28. van Gils B. Data Storage and Operations //Data in Context: Models as Enablers for Managing and Using Data. – Cham: Springer Nature Switzerland, 2023. pp. 105-114.
29. Dziedzic A., Elmore A. J., Stonebraker M. Data transformation and migration in polystores //2016 IEEE High Performance Extreme Computing Conference (HPEC). IEEE, 2016. pp. 1-6. doi: 10.1109/HPEC.2016.7761594
30. Зегжда, Д. П. Особенности обеспечения информационной безопасности вычислительных систем // Безопасность информационных технологий. 2021. Т. 28, № 1. С. 42–61. doi: 10.26583/bit.2021.1.04. EDN ETQPVN
31. Белим С. В., Белим С. Ю. Проблемы построения политики безопасности при объединении информационных систем //Математические структуры и моделирование. 2018. №. 3 (47). С. 126–131.

32. Полтавцев А. А., Хабаров А. Р., Селянкин А. О. Атаки логического вывода и защита информации в базах данных // Проблемы информационной безопасности. Компьютерные системы. 2019. № 4. pp. 20–25. EDN NTRSDO
33. Полтавцева М. А. и др. Модели форензики и расследование инцидентов в СУБД // Защита информации. Инсайд. 2021. № 3(99). С. 18–23. EDN OMKIRU.
34. Полтавцева М. А., Торгов В. А. Применение технологий распределенного реестра для аудита и расследования инцидентов в системах обработки больших данных // Проблемы информационной безопасности. Компьютерные системы. 2021. № 4. С. 144–156. doi: 10.48612/jisp/r3x6-ea4a-aaxn. EDN YRRVXG.

MULTI-LEVEL SECURITY CONCEPT FOR BIG DATA MANAGEMENT SYSTEMS

Poltavtseva M.A.[6], Zegzhda D.P.[7], Kalinin M. O.[8]

The purpose *of the study. Big data management technologies and systems are the basis for a huge number of modern digital services. On the one hand, they are built on traditional solutions, and on the other hand, they incorporate new approaches such as polystores or data outsourcing. The key role in the technology stack of the digital economy and novelty determine both the attractiveness of such assets for an attacker and the imperfection of protection methods. The aim of the paper is to analyze big data as an object of protection and to develop a multilevel concept of their security based on the consistency approach.*

Methods *of the study. The paper uses a layered approach, which also corresponds to the ANSI/SPARC architecture of database management systems. Big data is considered at three levels from infrastructure to business logic, key technologies, vulnerabilities and protection methods are highlighted. The ANSI/SPARC also defines in more detail the level architecture of big data management systems based on polystores and analyzes their security. The technological basis for the security of big data management systems as a system of distributed dynamic auditing is defined, an example of such a system based on a distributed registry is given.*

Results *of the study. The article identifies three levels of big data consideration: infrastructure, data engineering and business logic. The authors formulate the evolutionary changes of big data systems in comparison with traditional DBMS from the point of view of information security. The concept of big data management system is given, its own architectural levels based on ANIS/SPARK architecture are defined, for each of them security problems, reasons for their appearance and directions of protection development means are highlighted. The authors highlighted the key security requirement of big data management systems - consistent representation at the level of a global security monitor. For its implementation, in terms of collecting data about the system, the use of distributed dynamic ledger technologies is proposed. The system of distributed dynamic auditing for big data management based on HashGraph technology has been tested.*

Scientific novelty. *The paper is the first to formulate a multilevel security concept for big data management systems, within the framework of which the key vulnerabilities of big data systems, different from other classes of systems and traditional DBMSs, are identified and systematized at different levels. For the first time the application of distributed ledger technology for collecting data on the life cycle of information in the big data management system was proposed. The conducted research allows for a more comprehensive approach to ensuring the security of big data and big data management systems, specifies and coordinates the sets of protection methods and means, and lays the foundation for the construction of such systems in a secure design.*

6 Maria A. Poltavtseva, Dr.Sc., Associate Professor, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia. E-mail: poltavtseva@ibks.spbstu.ru

7 Dmitri P. Zegzhda, Corresponding member of RAS, Dr.Sc., Professor, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia. E-mail: dmitry@ibks.spbstu.ru

8 Maxim O. Kalinin, Dr.Sc., Professor, Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia. E-mail: max@ibks.spbstu.ru

Keywords: *information security, big data security, consistency approach, security architecture, security model, polystore security, distributed registry.*

The study was supported by the grant of Russian Science Foundation No.23-11-20003, https://rscf.ru/project/23-11-20003/; grant of St.Petersburg Science Foundation (Agreement No.23-11-20003 on the regional grant).

References

1. Naeem M. et al. Trends and future perspective challenges in big data //Advances in Intelligent Data Analysis and Applications: Proceeding of the Sixth Euro-China Conference on Intelligent Data Analysis and Applications, 15–18 October 2019, Arad, Romania, Springer Singapore, 2022. pp. 309-325.
2. Kornev M. S. Istorija ponjatija “bol’shie dannye” (Big Data): slovari, nauchnaja i delovaja periodika // Vestnik RGGU. Serija: Literaturovedenie. Jazykoznanie. Kul’turologija. 2018. №1 (34). pp. 81-85
3. Otto B. The evolution of data spaces // Designing Data Spaces: The Ecosystem Approach to Competitive Advantage. Cham: Springer International Publishing. 2022. pp. 3-15.
4. Gupta D., Rani R. A study of big data evolution and research challenges //Journal of information science. 2019. Vol. 45. Is 3. pp. 322-340
5. Antipova K.G. Sposoby opredelenija bol’shih dannyh: Rossijskij i zarubezhnyj opyt // Juridicheskie issledovanija. 2021. № 9. pp. 143 - 157. doi: 10.25136/2409-7136.2021.9.36591
6. Badidi E., Mahrez Z., Sabir E. Fog computing for smart cities’ big data management and analytics: A review // Future Internet. 2020. Vol. 12. No. 11. pp.1-28. doi: 10.3390/fi12110190.
7. Wang J. et al. Big data service architecture: a survey // Journal of Internet Technology. 2020. Vol. 21. No 2. pp. 393-405.
8. Bhattarai B. P. et al. Big data analytics in smart grids: state of the art, challenges, opportunities, and future directions //IET Smart Grid. 2019. Vol. 2. No. 2. pp. 141-154. doi: 10.1049/iet-stg.2018.0261.
9. Ndikumana A. et al. Joint communication, computation, caching, and control in big data multi-access edge computing //IEEE Transactions on Mobile Computing. 2019. Vol. 19. No 6. pp. 1359-1374. doi: 10.1109/TMC.2019.2908403
10. Vogt M. et al. Polystore Systems and DBMSs: Love Marriage or Marriage of Convenience? //Heterogeneous Data Management, Polystores, and Analytics for Healthcare: VLDB Workshops, Poly 2021 and DMAH 2021, Virtual Event, August 20, 2021, Revised Selected Papers 7. – Springer International Publishing, 2021. pp. 65-69.
11. Lu J., Holubová I., Cautis B. Multi-model databases and tightly integrated polystores: Current practices, comparisons, and open challenges //Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018. pp. 2301-2302.
12. Gobert M. Design, Manipulation and Evolution of Hybrid Polystores. 2023. https://pure.unamur.be/ws/portalfiles/portal/74437131/2023_GobertM_these.pdf (access date 01.08.2023)
13. Poltavtseva, M. A. Evolution of Data Management Systems and Their Security // Proceedings - 2019 International Conference on Engineering Technologies and Computer Science: Innovation and Application, EnT 2019, Moscow, 26–27 march 2019. – Moscow, 2019. pp. 25-29. doi: 10.1109/EnT.2019.00010. EDN AGZBGD.
14. Poltavceva, M. A. Bezopasnost’ baz dannyh / Sankt-Peterburg : Federal’noe gosudarstvennoe avtonomnoe obrazovatel’noe uchrezhdenie vysshego obrazovanija “Sankt-Peterburgskij politehnicheskij universitet Petra Velikogo”, 2023. 143 p. EDN RICQEN.
15. Titarenko, D. V., Ismajlov Je. I. Bezopasnost’ bol’shih dannyh // Problemy informacionnoj bezopasnosti social’no-jekonomicheskih sistem : VII Vserossijskaja s mezhdunarodnym uchastiem nauchno-prakticheskaja konferencija, Gurzuf, 18–20 fevralja 2021 goda. – Simferopol’: Krymskij federal’nyj universitet im. V.I. Vernadskogo, 2021. pp. 121-122. EDN NFQQHT.
16. Skvorcov N. Konstantinov A., Kuznecov S. Cennost’ vashih dannyh / Moskva: Al’pina PRO, 2022. – 750 s.
17. Ogbuke N. J. et al. Big data supply chain analytics: ethical, privacy and security challenges posed to business, industries and society // Production Planning & Control. 2022. vol. 33. №. 2-3. pp. 123-137. doi: 10.1080/09537287.2020.1810764
18. Binjubeir M. et al. Comprehensive survey on big data privacy protection //IEEE Access. 2019. vol. 8. pp. 20067-20079. doi: 10.1109/ACCESS.2019.2962368
19. Madan S., Bhardwaj K., Gupta S. Critical analysis of big data privacy preservation techniques and challenges //International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Vol. 3. – Springer Singapore, 2022. – pp. 267-278. doi: 10.1007/978-981-16-3071-2_23
20. Mehta B. B., Rao U. P. Improved l-diversity: scalable anonymization approach for privacy preserving big data publishing //Journal of King Saud University-Computer and Information Sciences. 2022. vol. 34. Is. 4. pp. 1423-1430. doi: 10.1016/j.jksuci.2019.08.006
21. Dhiman G. et al. Federated learning approach to protect healthcare data over big data scenario //Sustainability. 2022. vol. 14. Is 5. pp. 1-14. doi: 10.3390/su14052500
22. Stat’ev V. Ju., Dokuchaev V. A., Maklachkova V. V. Informacionnaja bezopasnost’ na prostranstve” Bol’shih dannyh” //T-Comm-Telekommunikacii i Transport. 2022. vol. 16. №. 4. pp. 21-28.
23. Poudel M. et al. Development of a polystore data management system for an evolving big scientific data archive //Heterogeneous Data Management, Polystores, and Analytics for Healthcare: VLDB 2019 Workshops, Poly and DMAH, Los Angeles, CA, USA, August 30, 2019, Revised Selected Papers 5. – Springer International Publishing, 2019. pp. 167-182. doi: 10.1007/978-3-030-33752-0_12
24. Esu M.T, Val’duries P. Principy organizacii raspredelennyh bah dannyh. – M.: DMK Press. – 2021. 678p.
25. Fong J. S. P. et al. Heterogeneous Database Connectivity //Information Systems Reengineering, Integration and Normalization: Heterogeneous Database Connectivity. 2021. pp. 317-367. doi: 10.1007/978-3-030-79584-9_9

26. Abdennebi A. et al. Machine learning based load distribution and balancing in heterogeneous database management systems // Concurrency and Computation: Practice and Experience. 2022. vol. 34. Is 4. pp. 1-13. doi: 10.1002/cpe.6641
27. Kim T. et al. Similarity query support in big data management systems //Information Systems. 2020. vol. 88.pp. 1-61. doi: 10.1016/j.is.2019.101455
28. van Gils B. Data Storage and Operations //Data in Context: Models as Enablers for Managing and Using Data. Cham : Springer Nature Switzerland, 2023. pp. 105-114.
29. Dziedzic A., Elmore A. J., Stonebraker M. Data transformation and migration in polystores //2016 IEEE High Performance Extreme Computing Conference (HPEC). IEEE, 2016. pp. 1-6. doi: 10.1109/HPEC.2016.7761594
30. Zegzhda, D. P. Osobennosti obespecheniya informacionnoj bezopasnosti vy`chislitel`ny`x sistem // Bezopasnost` informacionny`x texnologij. 2021. Vol. 28, № 1. pp. 42-61. doi: 10.26583/bit.2021.1.04. EDN ETQPVN.
31. Belim S. V., Belim S. Ju. Problemy postroenija politiki bezopasnosti pri ob#edinenii informacionnyh sistem //Matematicheskie struktury i modelirovanie. 2018. №. 3 (47). pp. 126-131.
32. Poltavcev A. A., Habarov A. R., Seljankin A. O. Ataki logicheskogo vyvoda i zashhita informacii v bazah dannyh // Problemy informacionnoj bezopasnosti. Komp'juternye sistemy. 2019. № 4. pp. 20-25. EDN NTRSDO
33. Poltavceva M. A. i dr. Modeli forenziki i rassledovanie incidentov v SUBD // Zashhita informacii. Insajd. 2021. № 3(99). pp. 18-23. EDN OMKIRU.
34. Poltavceva M. A., Torgov V. A. Primenenie tehnologij raspredelennogo reestra dlja audita i rassledovanija incidentov v sistemah obrabotki bol'shih dannyh // Problemy informacionnoj bezopasnosti. Komp'juternye sistemy. 2021. № 4. pp. 144-156. doi: 10.48612/jisp/r3x6-ea4a-aaxn. EDN YRRVXG.