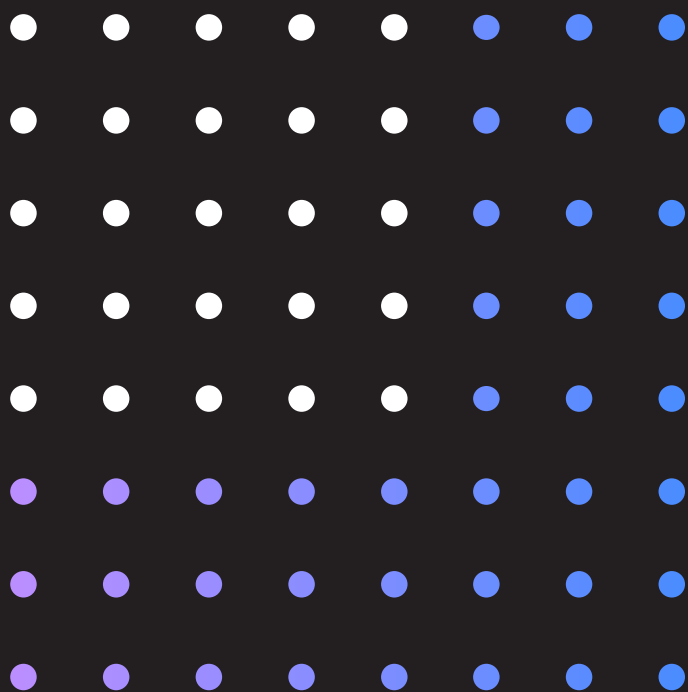


Доставляйте готовые для бизнеса данные с помощью интеллектуальной каталогизации данных и управления озером данных

IBM Watson Knowledge Catalog обеспечивает платформу управления данными, основанную на методах машинного обучения, которая поможет с задачами, связанными с озером данных



Содержание

03

Решение задач, связанных с озером данных, с применением подхода DataOps

03

Задачи, возникающие при использовании корпоративных озер данных

05

IBM Watson Knowledge Catalog

06

Единый источник достоверной информации и единая точка доступа

08

Четыре преимущества построения управляемого озера данных для ИИ

09

Заключение

Ключевые моменты

- Немногие компании получают ожидаемую отдачу от озер данных, которые они построили для хранения и анализа своих данных, чтобы получать достоверную аналитическую информацию.
- Решение DataOps выполняет задачи, с которыми сталкиваются компании в связи с неэффективностью при получении доступа, подготовке, интеграции и передаче данных потребителям, с соблюдением корпоративной и нормативной политики.
- К распространенным задачам, связанным с озером данных, относятся трудности и расходы при импортировании новых источников данных в озеро; невозможность интегрировать внутренние и внешние наборы данных; отсутствие уверенности в управлении данными; отсутствие доступа к инструментам самостоятельной подготовки данных; а также невозможность найти и понять данные, содержащиеся в озере.
- Корпоративная платформа управления данными с каталогизацией, оценкой качества и обнаружением данных способна трансформировать неудачный проект по созданию озера данных в настоящий источник бизнес-преимуществ.
- **IBM Watson® Knowledge Catalog** на базе IBM Cloud Pak™ for Data предоставляет каталог на основе машинного обучения (МО) для обнаружения, каталогизации, обеспечения качества данных и управления ими. Данное решение помогает пользователям быстро обнаруживать, отбирать, категоризировать и распространять ресурсы данных, наборы данных и аналитические модели.
- Если в компании нет глубокого понимания данных, становится труднее доверять такой информации и использовать ее с любыми формами искусственного интеллекта (ИИ), включая МО и глубокое обучение.

Решение задач, связанных с озером данных, с применением подхода DataOps

Десять лет назад начался поиск гибкого универсального подхода к созданию центральной базы данных, где могли бы храниться все корпоративные данные. Решением стало озеро данных — универсальная среда для хранения информации, в которой содержались бы данные практически любого типа. Эта среда также позволила бы бизнес-аналитикам и специалистам по обработке данных применять самые подходящие обработчики и инструменты для аналитики к каждому набору данных в их исходном местоположении.

Обычно такие озера данных строились с помощью Apache Hadoop и распределенной файловой системы Hadoop (Hadoop Distributed File System, HDFS) в комбинации с такими обработчиками, как Apache Hive и Apache Spark. Когда эти озера данных начали расти, стал очевиден ряд проблем. Хотя технология физически поддерживала масштабирование для сбора, хранения и анализа широких и разнообразных наборов структурированных и неструктурированных данных, слишком мало внимания уделялось практическим аспектам того, как внедрить эти возможности в бизнес-процессы.

К 2022 году более 80 % проектов, связанных с озерами данных, не смогут работать эффективно, поскольку окажется, что поиск, инвентаризация и отбор данных являются наибольшими препятствиями для аналитики и успешной обработки данных¹. В результате такие вопросы как «какие данные нам следует поместить в озеро?», «кто будет их использовать?», «как нам облегчить поиск данных?», «откуда берутся эти данные?» и «как нам предотвратить злоупотребление данными?», часто остаются без ответа. Эти важнейшие ограничения при решении проблем с персоналом, процессами и технологиями на практике приводили к безуспешному внедрению озер данных.

Сегодня многие организации признали свои ошибки, трансформировали руководящие группы по внедрению озер данных и начинают вторую, третью или даже четвертую попытку успешного внедрения озера данных — на это раз руководствуясь обработкой данных [DataOps](#).

В данной брошюре дается оценка распространенных проблем, связанных с озерами данных, и предлагается новый подход (например, DataOps), который может помочь превратить их из трясины данных в важнейший элемент магистрали готовых для бизнеса данных.

DataOps представляет собой практику совместного управления данными, сосредоточенную на улучшении коммуникации, интеграции и автоматизации потоков данных между администраторами и потребителями данных в масштабе всей организации.

Представляем DataOps!

DataOps объединяет лучшие методики DevOps (Developer Operations — интеграция разработки и эксплуатации программного обеспечения), управления и контроля данных в общую интегральную среду для совместной разработки и поддержки потоков данных среди нескольких заинтересованных сторон. Решение DataOps разработано для выполнения задач, с которыми сталкиваются компании в связи с неэффективностью при получении доступа, подготовке,

интеграции и передаче данных потребителям, с соблюдением корпоративной и нормативной политики. Отмечается результативность в рамках подразделения компании, аналитической группы или даже операционного процесса.

Чтобы следовать данной методологии, необходимо решать проблемы на уровне персонала, процессов и технологий, в чем и заключается разница между успешным и неудачным внедрением озера данных. С технологической точки зрения DataOps подчеркивает важность применения полностью интегрированной сквозной платформы для усвоения и интеграции, обеспечения качества, управления и использования данных, чтобы создать управляемое озеро данных. Правила валидации качества данных должны работать автоматически в рамках процесса усвоения данных, чтобы поддерживать непрерывную магистраль данных на уровне всего предприятия. Процесс усвоения данных должен быть полностью интегрирован в каталог данных, который становится центром вашей магистрали. Потребители данных должны иметь доступ к оценке качества и результатам профилирования данных из каталога и верить, что организация работает с теми же данными с учетом контекста.

Рост данных превосходит способность организации получать от них отдачу. Когда у компаний спросили, с какими наиболее серьезными проблемами они столкнулись при использовании аналитических систем, они ответили следующее: 1) 40 % занимает объединение существующих бизнес-процессов с источником данных, чтобы проанализировать их, и 2) 39 % — поиск источников, сбор, управление и контроль данных в процессе их роста². Сегодня дело не только в сохранении большого количества времени и ресурсов, которые уже были вложены в технологии озера данных, а в том, что альтернативных вариантов не существует. При внедрении ИИ или даже выполнении сравнительного анализа важно иметь полное представление о как можно большем объеме данных, что означает, что вам нужна архитектура, способная хранить, анализировать и управлять всеми этими данными централизованно. Во многих случаях управляемое озеро данных — единственный реалистичный вариант выполнения этих требований.

Сегодня компании могут — и должны — найти способ получать отдачу от озера данных, убедившись, что оно поддерживает магистраль готовых для бизнеса данных для DataOps.

Задачи, возникающие при использовании корпоративных озер данных

Обмен данными

Когда какая-либо группа в компании приобретает или создает новый набор данных, по-видимому, она осознает значимость и конфиденциальность этих данных. Если в них, к примеру, содержится информация, представляющая коммерческую тайну, персональные идентифицирующие данные или информация о потребителях, группа будет знать, как можно использовать такую информацию, а как нельзя, и примет меры предосторожности, чтобы убедиться, что никто не станет ей злоупотреблять.

Также она будет знать, что вне группы другие потенциальные пользователи данных могут не разделять понимание значимости данных или рисков, связанных со злоупотреблением ими. Естественно, эти риски заставят группу быть чрезвычайно осторожной при распространении данных или хранении в местах, не подконтрольных ей.

Это плохие новости для озер данных. Если компании рассматривают озеро данных как обычную неконтролируемую «свалку» данных, они крайне неохотно отнесутся к идее доверить свои ценные данные озеру. В итоге другие подразделения компании не смогут получить отдачу от этих данных, и вся концепция использования озера данных в качестве хранилища с самообслуживанием, предназначенного для распространения корпоративных данных, развалится.

Интеграция данных

Даже если группа согласна интегрировать свои данные в озеро, процесс может быть мучительным. Исходная концепция озера данных — импортировать данные в необработанном формате, не требующая комплексных процессов извлечения, трансформации и загрузки (extract, transform and load, ETL), происходящих в традиционном хранилище данных. Однако на практике почти для всех источников данных требуется предварительная обработка в той или иной мере, прежде чем информация станет пригодной к какому-либо осмысленному анализу.

В результате интеграция нового источника данных в озеро часто занимает месяцы. И поскольку большая часть этих данных ранее хранилась в небольших разрозненных операционных базах данных, а не в корпоративных системах, в целом для интеграции могут накопиться десятки или даже сотни источников.

Это означает, что во многих случаях информация, необходимая бизнес-аналитикам и специалистам по обработке данных, еще не добавлена в озеро и поступит туда только через несколько месяцев или даже лет. Что, в свою очередь, может стать значительным препятствием при внедрении.

Хранение данных

Несмотря на то что стоимость аппаратного хранилища и вычислительных ресурсов за последние несколько лет резко снизилась, кластеры Hadoop не предоставляются бесплатно. Хранение массивных объемов данных в озере намного экономичнее, чем в высокоэффективных устройствах для хранения данных, но стоимость все еще может быть значительной.

Кроме того, в отличие от данных, традиционно содержащихся в хранилищах, отношение стоимости к объему больших данных, хранимых в озере, сравнительно невелико. Вам может потребоваться очень большой стог, чтобы разместить горсть особенно ценных иголок.

Если вы не знаете, какие наборы данных действительно полезны и ценны для ваших специалистов по обработке данных, то рискуете вложить значительные суммы в интеграцию и хранение данных, которые утонут на дне вашего озера и никогда не будут использоваться.

Поиск данных

При условии что вы определили наиболее ценные наборы данных для хранения, убедили заинтересованных лиц распространить их и успешно интегрировали наборы данных в свое озеро данных, вам все равно необходимо

Задачи, возникающие при использовании корпоративных озер данных

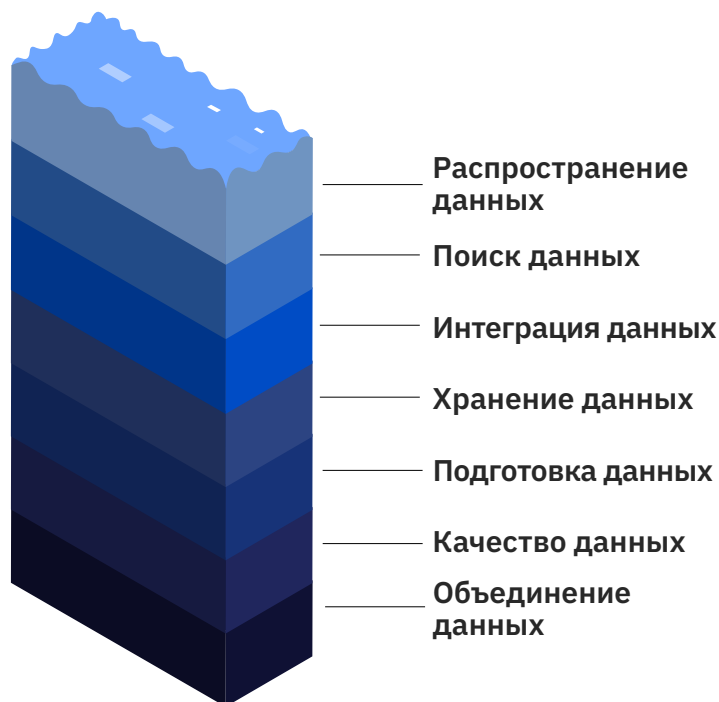


Рисунок 1. Предприятия, внедрившие технологии озера данных, могут столкнуться с одной или несколькими распространенными проблемами.

предоставить возможность другим пользователям найти, понять и использовать их надлежащим образом. Качество данных в озере представляет еще одну проблему. Вы не уверены в качестве данных, но их помещают в озеро.

К сожалению, в большинстве озер данных достичь качества не так просто. Часто данные хранятся без какого-либо контекста, из-за чего новому пользователю трудно или невозможно декодировать их, не обращаясь к исходному владельцу. Терминология бывает настолько связана с предметной областью, что метрика, используемая в одной сфере бизнеса, может быть известна совершенно под другим названием (или описываться немного по-другому) в другой. Потенциальная путаница и неправильная интерпретация может быть настолько существенной, что многие наборы данных фактически оказываются бесполезными или даже опасными для аналитика, незнакомого с ними.

Объединение внутренних и внешних данных

Наконец, даже в крупнейшем озере данных не стоит пытаться сохранить все возможные наборы данных, которые специалисты компании захотят использовать. Например, бессмысленно импортировать полные копии Google Maps, Weather.com® или Bloomberg в ваше озеро данных только потому, что один из ваших специалистов по обработке данных хочет провести геопространственный анализ или интегрировать информацию о погоде или котировки акций в алгоритм.

Поскольку ваше озеро данных не может вместить все данные, необходимые вашим бизнес-аналитикам для анализа, им придется тратить время на поиск в различных приложениях. Учитывая, что значительная часть полезного анализа, вероятно, включает комбинацию

внутренних и внешних наборов данных, снова возникает барьер для входа и, с точки зрения пользователей, снижается ожидаемая отдача от озера данных.

Подготовка данных

Существует множество факторов, затрудняющих [подготовку данных](#) — от понимания того, где искать данные, до последующего форматирования. Подготовка данных к использованию во время анализа — наиболее неэффективная и отнимающая много времени задача для пользователей данных. Пользователи тратят большую часть времени на поиск, очистку и форматирование информации вместо того, чтобы сосредоточиться на анализе данных, моделировании и предоставлении аналитической информации, влияющей на бизнес.

Ограничение доступа к управляемым наборам данных также вызвало излишнюю зависимость от ИТ на стадии подготовки. Ограниченный доступ сигнализирует о необходимости улучшить возможности для самообслуживания и навыки грамотного использования данных на уровне всего предприятия, чтобы сгладить это препятствие.

Качество данных

Сброс данных в озеро может сделать его непригодным к использованию. Поскольку нет правил оценки качества и валидации данных, применяемых к ним перед загрузкой в озеро, оно не предоставляет проверенную информацию, которую можно использовать. Высокое качество данных — важная характеристика, определяющая их надежность для принятия решений. Данные — ценный актив, которым необходимо управлять при перемещении по организации в целом. С ростом количества и разнообразия источников информации и большей целенаправленностью инициатив соответствия нормативным требованиям потребность в интеграции и получении доступа к информации из разрозненных источников в виде последовательных, надежных и многократно используемых данных становится решающей.

Целостный подход к построению управляемых озер данных

В большинстве озер данных используется Apache Hadoop с широкой экосистемой проектов с открытым исходным кодом для уровней хранения данных и обработчиков для аналитики. Неудивительно, что сообщество разработчиков ПО с открытым исходным кодом, созданное вокруг Hadoop, признаёт проблемы, с которыми сталкиваются в настоящее время при внедрении озер данных, поэтому недавно начали возникать многочисленные проекты, целью которых является решение различных проблем по отдельности. Также на рынке существует ряд проприетарных инструментов, предназначенных для решения этих проблем.

Так что возникает желание решать проблемы с озером данных по отдельности, при их возникновении. Если количество наборов данных становится слишком большим для управления, — добавьте инструмент для каталогизации. Если пользователи жалуются, что они не могут найти нужные данные — «прикрепите» пользовательский интерфейс с функцией поиска. А если ваши операторы базы данных больше не могут следить, откуда берутся данные или кто их использует, — разверните инструменты для определения происхождения данных и интегральную среду для управления данными.

Звучит просто, но на практике такой разрозненный подход часто реализуется за счет значительно возрастающей сложности и ухудшения поддержки, особенно из-за увеличения масштаба и объема озера данных. Подобно тому, как добавление новых источников в озеро данных увеличивает сложность требований к ETL, внедрение

новых инструментов приводит к увеличению сложности нефункциональных требований озера данных.

Вместо того чтобы получить интегрированную комплексную платформу, способную интегрировать данные, выполнять операции по обеспечению качества и каталогизацию, чтобы ваши бизнес-аналитики могли эффективно ими пользоваться, вы, как правило, обнаруживаете, что каждый инструмент исправляет отказы по-своему и имеет собственный подход к их регистрации. В итоге устранение неисправностей и решение проблем может отнимать очень много времени.

Другой, более существенный недостаток разрозненного подхода становится очевидным, когда вы смотрите на проблемы, с которыми обычно сталкиваетесь в связи с озером данных, не столько с технической, сколько с концептуальной точки зрения. Ключевая идея заключается в том, что масштабируемость, доступность для поиска, интеграция, качество и управление данными — не отдельные проблемы, они неразрывно связаны между собой. Для их решения потребуется значительно более целостный подход.

Масштабируемость, доступность для поиска, интеграция, качество и управление данными — не отдельные проблемы, они неразрывно связаны между собой. Для их решения потребуется целостный подход к управлению информацией.

IBM Watson Knowledge Catalog Обнаружение, каталогизация и обеспечение качества данных

[IBM Watson Knowledge Catalog](#) на базе IBM Cloud Pak for Data помогает пользователям данных быстро находить, отбирать, категоризировать и распространять активы и наборы данных, аналитические модели и их соотношения совместно с другими сотрудниками организации. Данное решение помогает группам, управляющим данными, устанавливать бизнес-гlossарий, политику и правила, а также обеспечивает усовершенствованный автоматизированный документооборот для управления. Каталог служит единым источником достоверной информации для инженеров, операторов, специалистов по обработке данных и бизнес-аналитиков для получения самостоятельного доступа к данным, которым они могут доверять и использовать с уверенностью.

Такие решения, как IBM Watson Knowledge Catalog на базе IBM Cloud Pak for Data, способны предоставить все возможности, необходимые для решения основных проблем современных озер данных на одной комплексной платформе. Каталог помогает справиться с первопричиной возникновения взаимосвязанных проблем — широко распространенной неспособностью озер данных предоставить эффективные инструменты для сбора, хранения и управления метаданными, а также отслеживания происхождения данных.

Отдача от озера данных во многом зависит от метаданных, которые оно содержит, в той же степени, что и от самих данных. Без метаданных, объясняющих, откуда взялся набор данных, кто его создал, что он содержит, кому разрешается им пользоваться и как он используется, сами по себе данные практически бесполезны. Пользователи не смогут их найти, а даже если и смогут, то не поймут, что эти данные означают, не смогут уверенно им доверять и не будут знать, как их использовать.

Watson Knowledge Catalog

Доставка надежных и значимых данных

Организируйте ваши данные



Знание

Данные должны быть полными, применимыми и доступными везде. Найдите, классифицируйте и понимайте данные любого типа.

Управляйте вашими данными



Доверие

Данные должны быть безопасными, чистыми и легко находимыми, что способствует надежному доступу с самообслуживанием. Узнайте о происхождении и качестве данных.

Распространяйте ваши данные



Потребление

Возможность предоставить самостоятельное обнаружение и автоматизированное принятие решений для развития бизнеса. Обеспечьте обзор всей информации для тех, кому она нужна, и позвольте им получить к ней доступ.

Рисунок 2. IBM Watson Knowledge Catalog обеспечивает широкий спектр возможностей для обнаружения, каталогизации и управления данными.

Единый источник достоверной информации и единая точка доступа

Решение IBM Watson Knowledge Catalog на базе IBM Cloud Pak for Data справляется с проблемами, сделав метаданные ключевым приоритетом. В его основе заложен мощный обработчик для каталогизации, индексирующий все наборы данных и аналитические активы, к которым ваша компания имеет доступ, вне зависимости от того, где хранятся данные, например в вашем озере данных, хранилище данных, транзакционной системе или даже наборе электронных таблиц. Не важно, структурированы ваши данные или нет, хранятся ли они локально или размещены в облаке. Кроме того, каталог также может включать внешние наборы данных и источники, например проприетарные службы данных, на которые подписана ваша компания, или API (application programme interface — прикладной программный интерфейс) с открытыми данными.

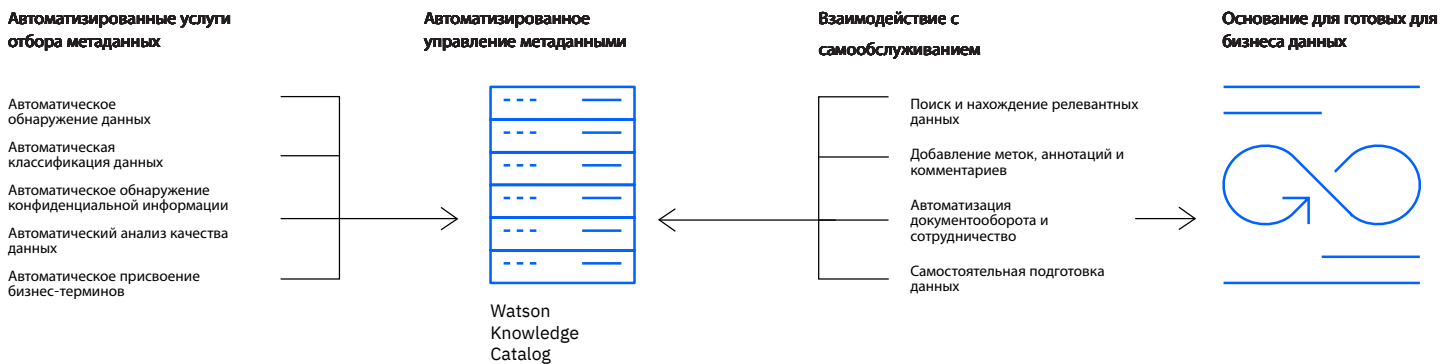
Наряду с предоставлением единого источника достоверной информации обо всех ваших наборах данных, каталог данных также обеспечивает единую точку доступа. Возможности поиска и предложения на базе ИИ помогают бизнес-аналитикам, специалистам по обработке данных, инженерам по обеспечению качества данных и группам по управлению данными легче находить активы и предоставляют доступные метаданные, помогающие пользователям понять, что именно они нашли, и оценить, полезна ли эта информация для них.

Встроенные возможности самостоятельной подготовки данных сокращают время, необходимое для трансформации данных перед продуктивным использованием в приложениях для аналитики и ИИ. Бизнес-аналитикам и специалистам по обработке данных не нужно тратить время на подготовку и анализ данных. Интеграция с решением для подготовки данных в масштабах предприятия, например [IBM® InfoSphere® Advanced Data Preparation](#), помогает обеспечить создание управляемых наборов данных через поверхность каталога, наиболее подходящих по контексту, чтобы получить важные для бизнеса знания и действия для бизнес-пользователей. Такая интеграция способствует взаимодействию по магистрали данных.

Масштабируемость, доступность для поиска, интеграция, качество и управление данными — не отдельные проблемы, они неразрывно связаны между собой. Для их решения потребуется целостный подход к управлению информацией.

Каталог также помогает операторам базы данных, работающим в отделе директора по обработке данных (chief data officer, CDO), благодаря меткам и классификации наборов данных и автоматическому отслеживанию их происхождения и использования, а также применению встроенного бизнес-гlossария для стандартизации бизнес-терминологии по всем данным. В результате операторам проще понять, что содержит каждый набор данных, где находится конфиденциальная или персональная информация и кому следует разрешить доступ к ней.

Единый каталог для многочисленных источников данных внутри организации и вне ее



Автоматизированные услуги по управлению ключевыми и эталонными данными

Отслеживание происхождения данных	Управление политикой и принудительное соблюдение требований	Непрерывное управление	Управление бизнес-гlossарием
Архивация и удаление данных	Управление моделями и сообщения об отклонениях	Управление организацией и решение проблем	Управление качеством данных

Машинное обучение и автоматизация

Локально	IBM Cloud	Amazon Web Services	OpenStack
Частное облако	Red Hat OpenShift	Azure	Google Cloud

Рисунок 3. С интеллектуальным указателем метаданных IBM Watson Knowledge Catalog данные — структурированные и неструктурированные — можно хранить в исходных системах, но пользователи смогут быстро их найти для более взвешенной аналитики.

В IBM Watson Knowledge Catalog метаданные — ключевой приоритет, благодаря чему обеспечивается единый источник достоверной информации и единая точка доступа ко всем наборам данных вашей компании.

Встроенное интеллектуальное обнаружение данных

Чтобы еще более улучшить обнаружение данных, каталог позволяет пользователям оставлять метки и комментарии в наборах данных и аналитических активах, обогащая метаданные и добавляя дополнительный контекст, помогающий сотрудникам, работающим вместе, найти то, что им нужно. Решение также содержит встроенные алгоритмы обнаружения данных, которые используют МО для автоматической классификации содержимого каждого набора данных. Установив общие типы полей, например имена, адреса, почтовые индексы и номера социального страхования, решение сокращает потребность в создании аннотаций к данным вручную. Внедряется автоматизация и МО для автоматического отбора данных и управления метаданными. Благодаря встроенным функциям оценки качества данных, решение поддерживает глубокое профилирование данных, обеспечение качества и правила валидации.

Автоматизированная обработка данных создает магистраль отобранных данных с обеспечением качества и управлением информацией, а также помогает поддерживать непрерывный поток высококачественных контролируемых данных в озеро.

Подобным образом, добавление интеллектуальной модели метаданных для ваших активов обеспечивает уникальную методику автоматического соблюдения требований, например Общего регламента защиты персональных данных (General Data Protection Regulation, GDPR) и Закона штата Калифорния о защите конфиденциальности потребителей (California Consumer Privacy Act, CCPA).

IBM Watson Knowledge Catalog на базе IBM Cloud Pak for Data помогает доставлять достоверные, высококачественные, готовые для бизнеса данные фактически всем пользователям.

Все компоненты решения разработаны как микросервисы с единым набором принципов проектирования и общим подходом к нефункциональным требованиям, например масштабируемости, устранению ошибок, безопасности и регистрации.

Решение IBM Watson Knowledge Catalog обеспечивает платформу корпоративного управления с МО, готовую для внедрения масштабируемого ИИ.

Вместо сбивающих с толку ошибок и «узких мест» в работе, к которым, вероятно, приведет разрозненный подход в стиле «сделай сам», решение IBM Watson Knowledge Catalog обеспечивает платформу корпоративного управления с МО, чтобы подготовиться к внедрению масштабируемого ИИ.

Доступно три варианта IBM Watson Knowledge Catalog:

- В виде программного обеспечения как услуги (software as a service, SaaS) в IBM Cloud™
- В [IBM Cloud Pak for Data](#)
- Интегрированный в [IBM Watson Studio](#)

Решения, подобные IBM Watson Knowledge Catalog, помогут получить отдачу, которая ожидалась при создании озер данных. Watson Knowledge Catalog с возможностями интеллектуальной каталогизации и управления помогает построить надежное и управляемое озеро данных для ИИ.

Четыре преимущества построения управляемого озера данных для ИИ

1. Укрепление надежности и уверенности в данных благодаря обеспечению качества и управлению

- Возможности обеспечения качества данных помогают вам повысить качество информации и сделать доступными высококачественные данные в вашем озере.
- Политика управления настроена автоматически и принудительно соблюдается, так что если вы находите набор данных, то знаете, разрешается ли вам его использовать и как именно.
- Вы можете отбирать данные, поскольку пользователи добавляют оценки, комментарии и другую информацию, помогающую другим определить, будет ли набор данных полезен для них.

2. Расширение возможностей ваших пользователей

- Ваши подразделения по разным направлениям бизнеса (line-of-business, LOB) свободно делятся своими данными, поскольку они уверены, что информация надлежащим образом контролируется и защищена от злоупотребления.
- Вы можете поощрять сотрудничество и трансформировать данные в надежные активы предприятия посредством динамичной политики управления данными и принудительного соблюдения требований.
- Со временем ваши данные будет легче найти и повторно использовать, поскольку пользователи будут добавлять соответствующие метки и метаданные, которые помогут другим получить отдачу от них.
- Единый интерфейс обеспечивает вам доступ к любому набору данных, которым владеет ваша компания, вне зависимости от того, где он хранится.

3. Верните себе время

- Автоматическое обнаружение данных сокращает время и усилия, которые вы тратите на добавление метаданных для новых наборов данных.
- Автоматический отбор данных и управление метаданными сокращает время, необходимое для обнаружения метаданных и назначения терминов, а также уменьшает время создания бизнес-гlossария.

- С помощью простых и интуитивно понятных инструментов самостоятельной подготовки данных ваши пользователи потратят меньше времени на подготовку данных и больше времени на получение аналитической информации.

- Вы можете помочь специалистам по обработке данных и бизнес-аналитикам предоставлять более качественную аналитическую информацию за более короткий промежуток времени.
- Интеллектуальный поиск на базе ИИ помогает за секунды находить нужные данные вместо того, чтобы ждать недели, пока другая команда предоставит их вам.

4. Управляйте растущими данными и затратами

- Вы можете оптимизировать затраты на хранение, избегая расходов на внедрение малоценных наборов данных в озеро.
- Вы также можете просматривать все внешние наборы данных, на которые подписана ваша организация, чтобы снизить риск оплаты большего количества подписок, чем вам необходимо.
- Вы можете расставить приоритеты при внедрении новых источников данных в озеро на основании спроса пользователей на данные, что позволит вам интегрировать наиболее ценные источники в первую очередь.

Начните получать отдачу от ваших данных

Работая в отделе CDO, ИТ, специалистом по обработке данных или аналитиком бизнес-направления, вы с коллегами разделяете общую цель. Если вы сможете построить озеро данных, которое действительно предоставляет обещанные преимущества, вы не только облегчите свою работу и сделаете ее более продуктивной. Вы способны также сыграть ключевую роль при получении конкурентного преимущества вашей компанией, которое на сегодняшний день есть лишь у небольшого количества организаций.

Если сможете очистить воды вашего озера данных, пока конкуренты все еще вязнут в трясине, и откроете возможности, о которых они могут только мечтать. Истинное преимущество первопроходца ожидает того, кто первым начнет получать отдачу от ранее неосвоенных данных.

Заключение

Узнайте, где хранятся все ваши данные и кто их использует, получите отдачу для вашей бизнес-аналитики.

Каталоги данных — важнейший элемент для инициатив DataOps, поскольку они способствуют предоставлению автоматизированного открытого управления метаданными путем интеграции управления данными, обеспечения качества и активного управления политикой.

Решение IBM Watson Knowledge Catalog с возможностями интеллектуальной каталогизации и управления помогает построить надежное и управляемое озеро данных для ИИ. Каталог включает интеграцию, обеспечение качества и управление данными в среде вашего озера, помогая предоставлять готовые для бизнеса данные для DataOps, а также служит единым источником достоверной информации.

Более подробная информация

Для получения дополнительной информации посетите веб-сайт

ibm.com/cloud/watson-knowledge-catalog

© IBM Corporation, 2019.

IBM Восточная Европа/Азия

123112, Россия, Москва,
Пресненская наб., 10
Тел.: +7 (495) 775-8800,
Факс: +7 (495) 940-2070

Выпущено в США, октябрь 2019 г. IBM, логотип IBM, **ibm.com**, IBM Cloud, IBM Cloud Pak, IBM Watson и InfoSphere являются товарными знаками International Business Machines Corp., зарегистрированными во многих юрисдикциях по всему миру.

Red Hat и OpenShift являются товарными знаками или зарегистрированными товарными знаками корпорации Red Hat, Inc. или ее дочерних компаний в США и других странах. Названия других продуктов и услуг могут являться товарными знаками IBM или других компаний. Актуальный перечень товарных знаков IBM см. на веб-сайте в разделе Copyright and trademark information («Сведения об авторском праве и товарных знаках») по адресу www.ibm.com/legal/copytrade.shtml.

Этот документ является актуальным по состоянию на дату первоначальной публикации и может быть изменен компанией IBM в любое время. В некоторых странах, где работает компания IBM, некоторые предложения недоступны. Информация в данном документе приводится «как есть», без каких-либо гарантий, явных или подразумеваемых, в том числе без каких-либо гарантий товарной пригодности или соответствия конкретному намерению использования, а также без каких-либо гарантий или условий нарушения прав интеллектуальной собственности. На продукты IBM распространяется гарантия в соответствии с положениями и условиями соглашений, согласно которым они предоставляются. Клиент несет ответственность за обеспечение соответствия законам и нормативным постановлениям, применимым к ним. Корпорация IBM не предоставляет юридические консультации, а также не заявляет и не гарантирует, что предоставляемые ею услуги или продукты обеспечивают соблюдение клиентом всех законодательных и нормативных требований.

1. Augmented Data Catalogs: Now an Enterprise Must-Have for Data and Analytics Leaders—Gartner, Sept 2019
2. The Forrester Wave: Machine Learning Data Catalogs, Q2 2018

ASW12449-RURU-03

