
Оглавление

Предисловие	9
Для кого предназначена эта книга?	11
Обозначения, используемые в этой книге	12
Благодарности	12
Глава 1. Введение в озера данных	15
Зрелость озера данных	18
Лужи данных	20
Пруды данных	21
Создание успешного озера данных	22
Правильная платформа	22
Правильные данные	24
Правильный интерфейс	25
Болото данных	28
Дорожная карта к успешности озера данных	29
Установка озера данных	30
Организация озера данных	30
Настройка озера данных для самообслуживания	32
Архитектуры озера данных	38
Озера данных в открытом облаке	39
Логические озера данных	39
Заключение	44
Глава 2. Исторический ракурс	45
Путь для самообслуживания данных — рождение баз данных	46
Императив аналитики — рождение хранилища данных	49
Экосистема хранилища данных	51
Сохранение и запрашивание данных	53

Загрузка данных — инструменты интеграции данных	59
ETL и ELT	61
Организация и управление данными	64
Использование данных	70
Заключение	72
Глава 3. Введение в большие данные и обработку данных	73
Hadoop возглавил исторический сдвиг к большим данным	74
Файловая система Hadoop	75
Как обработка и хранение взаимодействуют в работе MapReduce	76
Схема на чтение	78
Проекты Hadoop	78
Наука о данных (Data Science)	80
Машинное обучение	86
Заключение	89
Глава 4. Запуск озера данных	91
Что это и почему Hadoop	91
Предотвращение распространения лужи данных	95
Использование преимущества больших данных	96
В лидерах благодаря обработке данных	96
Стратегия 1: «разгрузить» существующую функциональность	100
Стратегия 2: озера данных для новых проектов	102
Стратегия 3: создать центральную точку управления	103
Какой путь подходит вам?	104
Заключение	106
Глава 5. От прудов данных и хранилищ больших данных до озер данных	107
Основные функции хранилища данных	108
Размерное моделирование для аналитики	110
Интеграция данных из отдельных источников	111
Сохранение истории путем использования медленно изменяющихся измерений	111
Ограничения хранилища данных как хранилища для исторических данных	112
Переход к пруду данных	113
Сохранение истории в пруду данных	113
Реализация медленно меняющихся измерений в пруду данных	115
Превращение прудов данных в озеро данных — загрузка данных, которых нет в хранилище	117
Необработанные данные	118

Внешние данные	119
Интернет вещей (IoT) и другие потоковые данные	122
Лямбда-архитектура	126
Преобразования данных	127
Целевые системы	130
Хранилища данных	131
Операционные склады данных	131
Приложения реального времени и продукты данных	131
Заключение	133
Глава 6. Оптимизация для самообслуживания	135
Начало самообслуживания	136
Бизнес-аналитики	139
Поиск и понимание данных — документирование предприятия	140
Установление доверия	143
Обеспечение	152
Подготовка данных для анализа	154
Анализ и визуализация	159
Заключение	165
Глава 7. Архитектура озера данных	167
Организация озера данных	167
Зона загрузки, или необработанная зона	169
Золотая зона	170
Рабочая зона	172
Конфиденциальная зона	173
Множественные озера данных	175
Преимущества разделения озер данных	175
Преимущества объединения озер данных	176
Облачные озера данных	177
Виртуальные озера данных	180
Федерация данных	181
Виртуализация больших данных	182
Устранение избыточности	184
Заключение	186
Глава 8. Каталогизация озера данных	187
Организация данных	188
Технические метаданные	189
Бизнес-метаданные	194
Тегирование	197

Автоматизированная каталогизация	198
Управление логическими данными	200
Управление конфиденциальными данными и контроль доступа	200
Качество данных	202
Соотнесение отдельных данных	204
Установление происхождения	206
Предоставление данных	208
Инструменты для создания каталога	208
Сравнение инструментов	210
Океан данных	211
Заключение	212
Глава 9. Управление доступом к данным	213
Авторизация, или контроль доступа	214
Политики доступа к данным на основе тегов	216
Деидентификация конфиденциальных данных	220
Суверенность данных и соответствие нормативным требованиям	223
Управление доступом на основе самообслуживания	226
Предоставление данных	230
Заключение	238
Глава 10. Перспективы для различных отраслей	239
Об авторе	263
Концовка	264
Предметный указатель	265

Предисловие

В последние годы многие предприятия начали экспериментировать с использованием больших данных и облачных технологий, чтобы строить озера данных и управлять данными культуры и процессом принятия решений. Но проекты часто останавливаются или терпят неудачу, потому что подходы, которые работали в интернет-компаниях, должны быть адаптированы для предприятия — и нет всеобъемлющего практического пособия о том, как успешно это сделать. Я написал эту книгу в надежде предоставить такое пособие.

Мне как руководителю в IBM и Informatica (крупнейшие поставщики технологий обработки данных), предпринимателю в Menlo Ventures (ведущей венчурной компании), а также основателю и техническому директору Waterline (стартап проекта по большим данным) повезло получить возможность поговорить с сотнями экспертов, стратегов, отраслевых аналитиков и практиков о проблемах построения успешных озер данных и создания культуры, управляемой данными. В этой книге мы обсудим различные темы, поговорим о лучших практиках, с которыми я столкнулся в разных отраслях (от социальных сетей до банковских и правительственных учреждений), а также о ролях (от руководителей по информационным технологиям и других ИТ-руководителей до архитекторов данных, исследователей данных и бизнес-аналитиков).

Большие данные, их обработка и аналитика, поддерживающие принятие решений на основе данных, обещают сделать максимально понятным и эффективным все — от работы с данными и клиентами до поиска лекарства от рака. Но обработка и аналитика зависят от наличия доступа к историческим данным. В знак признания этого компании внедряют озера больших данных, чтобы собрать все свои данные в одном месте и сохранить

историю и чтобы специалисты по работе с данными и аналитики имели доступ к информации, необходимой им для принятия решений. Озера больших корпоративных данных создают связь между свободно распространяющейся культурой современных интернет-компаний, где данные лежат в основе всех практик, каждый сам себе аналитик, и большинство людей могут кодировать и развертывать собственные наборы данных, и хранилищами корпоративных данных, где данные — драгоценный ресурс, предмет тщательной заботы профессионального IT-персонала, предоставляемый в форме тщательно подготовленных отчетов и аналитических наборов данных.

Для успешного применения корпоративные озера данных должны обеспечивать три новые возможности:

- Экономичное эффективное масштабируемое хранилище, чтобы можно было хранить и анализировать большие объемы данных без чрезмерных вычислительных затрат.
- Экономически эффективный доступ к данным и управление ими, чтобы каждый мог найти и использовать нужные данные без дорогостоящих затрат на программирование и ручной сбор специальных данных.
- Многоуровневый управляемый доступ, чтобы сделать разные уровни данных доступными для разных пользователей в зависимости от их потребностей, уровней квалификации и политик управления данными.

Базы данных Hadoop, Spark, NoSQL и гибкие облачные системы — замечательные новые технологии, которые выполняют первый пункт. Хотя они все еще находятся на стадии становления и сталкиваются с некоторыми проблемами, присущими любой новой технологии, они быстро стабилизируются и становятся основными тенденциями. Тем не менее эти мощные технологии не обеспечивают выполнение двух других пунктов, касающихся рентабельного и многоуровневого доступа к данным. Поэтому когда предприятия создают большие кластеры и получают огромные объемы данных, они обнаруживают, что вместо озера у них образовалось болото данных — большое хранилище непригодных наборов данных, которые невозможно найти и понять, и слишком опасно принимать решения на их основе.

Эта книга расскажет читателям о передовых методах построения озер больших данных. В ней рассматриваются различные подходы к запуску и увеличению озера данных, лужи данных (аналитические песочницы) и пруда данных (хранилища больших данных), создание озер данных с нуля. Также рассматриваются преимущества и недостатки различных архитектур озера данных — локальных, облачных и виртуальных, равно как и настройка различных зон для размещения всего: от сырых, необработанных данных до тщательно управляемых и обобщенных данных, — и управление доступом к этим зонам. Объясняется, как сделать возможным самообслуживание, чтобы пользователи могли сами находить, понимать и предоставлять данные, как предоставить разные интерфейсы пользователям с разным уровнем квалификации и как сделать все это в соответствии с политиками предприятия касательно управления данными.

Для кого предназначена эта книга?

Эта книга предназначена для следующих категорий сотрудников крупных традиционных предприятий:

- Службы данных и группы управления: руководители отделов по управлению данными и те, кто обслуживает работу с данными.
- IT-руководители и архитекторы: директора по технологиям и архитекторы больших данных.
- Аналитические команды: исследователи данных, инженеры данных, аналитики данных и руководители по аналитике.
- Команды по соответствию: главные сотрудники по информационной безопасности, сотрудники по защите данных, аналитики по информационной безопасности и руководители по соответствию нормативным требованиям.

В книге использован опыт моей 30-летней карьеры в области создания передовых технологий обработки данных и сотрудничества с крупнейшими мировыми предприятиями в рамках решения их проблем с данными. Книга опирается на передовой опыт ведущих мировых компаний и предприятий в области больших данных, а также на истории успеха практиков и отраслевых экспертов, чтобы предоставить исчерпывающее руководство

по разработке и развертыванию успешного озера больших данных. Если хотите воспользоваться преимуществами, которые дают предприятию захватывающие новые технологии и подходы к большим данным, эта книга — отличное место для начала. Руководители могут прочитать ее один раз и периодически ссылаться на нее, когда на рабочем месте возникают проблемы с большими данными. А для практических специалистов книга может послужить полезным справочным материалом при планировании и выполнении проектов озера больших данных.

Обозначения, используемые в этой книге

В этой книге используются следующие типографские обозначения:

Курсив

Обозначает новые термины, URL-адреса, адреса электронной почты, названия файлов и расширения файлов.

Моноширинный

Используется для листингов программ, а также для ссылки на элементы кода в тексте, такие как имена переменных или функций, базы и типы данных, переменные среды, операторы и ключевые слова.

Моноширинный курсив

Обозначает код, который должен быть заменен пользовательскими значениями или значениями, заданными контекстом.

Благодарности

Прежде всего я хочу выразить свою глубокую признательность всем экспертам, которые поделились со мной своими историями, опытом и лучшими практиками — эта книга для вас и еще раз для вас!

Большое спасибо всем людям, которые помогли мне работать над проектом. Это моя первая книга, и я действительно не смог бы создать ее без их помощи. Вот их имена:

- Команда O'Reilly: Энди Орам, редактор издательства O'Reilly, который вдохнул новую жизнь в эту книгу, когда я потерял хватку, и помог вывести ее из потока сознания на некоторый уровень согласованности; Тим Макговерн, творческий редактор, который помог

этой книге взлететь; Рэйчел Хэд, литературный редактор, шокировавшая меня тем, сколько еще улучшений можно было бы внести в книгу после более чем двухлетнего написания, редактирования, переписывания, рецензирования, дополнительного переписывания, дополнительного редактирования, усиленного переписывания... и Кристен Браун, которая руководила всем в процессе производства книги.

- Представители отрасли, которые поделились своими мыслями и лучшими практиками: их имена и биографии вы найдете перед их эссе в книге.

Рецензенты, которые внесли огромные улучшения благодаря своему свежему видению, критическому взгляду и отраслевому опыту: Санджив Мохан, Опиндер Бава и Николь Шварц.

Наконец, эта книга не вышла бы без поддержки и любви моей замечательной семьи — моей жены Ирины, моих детей — Ханна, Джейн, Лизы, Джона — и моей мамы Регины, моих друзей и моей замечательной семьи Waterline.

Введение в озера данных

Принятие решений на основе данных меняет то, как мы работаем и живем. Специалистам требуются данные, чтобы принимать решения в различных областях, начиная с обработки данных, компьютерного обучения и расширенной аналитики и заканчивая панелями мониторинга в режиме реального времени. Такие компании, как Google, Amazon и Facebook — это гиганты, управляемые данными и этим замещающие традиционные предприятия. Финансовые сервисные организации и страховые компании всегда руководствовались данными: на первом месте были количественный анализ и автоматизированная торговля. Интернет вещей (IoT) меняет производство, транспорт, сельское хозяйство и здравоохранение. От правительств и корпораций в каждой вертикали до некоммерческих и образовательных учреждений — везде данные рассматриваются как ключевая вещь. Искусственный интеллект и машинное обучение пронизывают все стороны нашей жизни. Мир зависит от данных из-за потенциала, который они представляют. У нас даже есть термин для этого: большие данные, определенные Дугом Лейни из Gartner в терминах трех V¹ (объем, разнообразие и скорость), к которым он позже добавил четвертое и, на мой взгляд, самое важное — достоверность.

С таким большим разнообразием, объемом и скоростью старые системы и процессы уже не могут удовлетворить потребности предприятия в данных. Достоверность — тоже серьезная проблема для продвинутой аналитики и искусственного интеллекта, где принцип «GIGO» (мусор на входе = мусор на выходе) еще более важен. Практически невозможно

¹ Англ. volume, variety, velocity, veracity соответственно. — Прим. ред.

определить, были ли данные плохими и вызвали ли они неправильные решения в моделях статистики и модели машинного обучения, или сама модель оказалась плохой.

Чтобы поддержать предприятия и решить эти проблемы, в управлении данными происходит настоящая революция вокруг хранения, обработки, управления и предоставления данных лицам, принимающим решения. Технология больших данных обеспечивает масштабируемость и экономическую эффективность на порядок выше, чем это возможно в традиционной инфраструктуре управления данными. Самообслуживание приходит на смену трудоемким подходам прошлого, когда армии IT-специалистов создавали хорошо управляемые хранилища и витрины данных, но для внесения каких-либо изменений требовались месяцы.

Озеро данных — это совершенно новый подход, который использует мощь технологии больших данных и объединяет ее с гибкостью самообслуживания. Большинство крупных предприятий сегодня либо уже развернули, разворачивают озера данных.

Эта книга основана на обсуждениях с более чем сотней организаций, от новых компаний, управляемых данными, таких как Google, LinkedIn и Facebook, до правительств и традиционных корпоративных предприятий, по поводу их инициатив по созданию озера данных, аналитических проектов, опыта и лучших практик. Книга предназначена для руководителей и практиков IT, которые рассматривают возможность создания озера данных, находятся в процессе его создания или унаследовали его, но пытаются сделать его продуктивным и широко распространенным.

Что такое озеро данных? Зачем нам это нужно? Чем оно отличается от того, что у нас уже есть? Пытаясь сохранить краткое изложение, я не буду подробно объяснять и исследовать каждый термин и концепцию прямо сейчас, но оставлю подробное обсуждение для последующих глав.

Принятие решений на основе данных сейчас в тренде. Специалистам в разных областях — от компьютерного обучения и расширенной аналитики до панелей мониторинга в режиме реального времени — требуются данные, которые помогают принимать решения. Этим данным нужен дом, и озеро данных — предпочтительное решение для создания такого дома. Термин был изобретен и впервые описан Джеймсом Диксоном, техническим директором компании Pentaho, который написал в своем блоге²:

² jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

«Если сравнить витрину данных с хранилищем бутилированной воды, очищенной, расфасованной и структурированной для удобства потребления, то озеро данных — это большой водоем в естественном состоянии. Содержимое потока данных поступает из источника, чтобы заполнить озеро, и различные пользователи могут приходить, чтобы исследовать, погружаться в озеро или брать образцы». Я выделил курсивом критические точки, вот они:

- данные в исходном виде и формате (натуральные или сырые данные);
- данные, с которыми работают различные пользователи (то есть доступные для разных пользователей).

В этой книге рассказывается, как создать озеро данных, которое предоставляет сырые (а также обработанные) данные большому сообществу пользователей и бизнес-аналитиков, а не просто применяет их в проектах, ориентированных на IT. Аналитики могут получить сырые данные путем самообслуживания. Оно стало важной мегатенденцией к демократизации данных. Все началось с использования инструментов визуализации самообслуживания, таких как Tableau и Qlik (иногда называемых инструментами обнаружения данных), позволяющих аналитикам изучать данные без необходимости обращаться к специалистам IT. Тенденция самообслуживания подразумевает наличие инструментов подготовки данных, которые помогают аналитикам формировать данные для анализа, инструментов каталогов, которые помогают находить нужные им данные, и инструментов обработки данных, которые помогают выполнять расширенную аналитику. Для еще более продвинутой аналитики, обычно называемой обработкой данных, новый класс пользователей (а именно — специалисты по работе с данными) использует озеро данных в качестве основного источника.

Конечно, большая проблема при самообслуживании — управление и безопасность данных. Все согласны с тем, что данные должны храниться в безопасности. Но во многих регулируемых законом отраслях существуют политики безопасности данных, которые необходимо применять, и таким образом предоставление аналитикам доступа ко всем данным оказывается незаконным. Даже в некоторых нерегулируемых отраслях это не одобряется. Возникает вопрос: как сделать данные доступными для аналитиков, не нарушая правила регулирования доступа к внутренним и внешним

данным? Это иногда называется демократизацией данных и будет подробно обсуждаться в последующих главах.

Зрелость озера данных

Озеро данных — это относительно новая концепция, поэтому полезно определить некоторые этапы зрелости этого феномена, которые вы можете наблюдать, и четко сформулировать различия между ними:

- Лужа данных — это группа данных для определенной цели, построенная с использованием технологии больших данных. Как правило, это первый шаг к внедрению технологии больших данных. Данные в луже используются для отдельного проекта или команды. Обычно это хорошо известно и понятно, и причина, по которой технология больших данных применяется вместо традиционных хранилищ данных, заключается в снижении затрат и обеспечении более высокой производительности.
- Пруд данных — это коллекция луж данных. Это похоже на плохо спроектированное хранилище данных — фактически набор витрин данных; также это может быть выгрузкой существующего хранилища данных. Хотя снижение затрат на технологии и повышение масштабируемости — очевидные преимущества, эти конструкции все еще требуют высокого уровня участия ИТ. Кроме того, пруды данных ограничены только данными, необходимыми для проекта, и используют их только для этой цели. Учитывая высокие затраты на ИТ и ограниченную доступность данных, пруды с данными не помогают нам в демократизации использования данных или стимулирования самообслуживания и принятия решений на основе данных для бизнес-пользователей.
- Озеро данных отличается от пруда данных двумя важными особенностями. Во-первых, оно поддерживает самообслуживание, когда бизнес-пользователи могут находить и использовать нужные им наборы данных, не прибегая к помощи ИТ-отдела. Во-вторых, оно содержит данные, которые могут понадобиться бизнес-пользователям, даже если на тот момент нет ни одного проекта, для которого они требуются.