Содержание

Об авторах

Шаббир Чаллавала имеет более чем 8-летний богатый опыт в предоставлении решений на основе технологий MySQL и PHP. В настоящее время он работает с KNOWARTH Technologies. Он принимал участие в разработке предназначенных для предприятий различных PHP-решений в области электронной коммерции и обучающих порталов; работал с различными вычислительными платформами на основе PHP, такими как Magento E-commerce, Drupal CMS и Laravel.

Шаббир участвовал в различных корпоративных решениях на разных этапах их реализации, в частности проектировании архитектуры, оптимизации баз данных и настройки производительности, и отлично разбирается в жизненном цикле разработки программного обеспечения. Он работал над интеграцией технологий больших данных, в частности MongoDB и Elasticsearch с платформой PHP.

Я искренне благодарен Чинтан Мехта, который оказал мне доверие, поручив написание этой книги. Я хотел бы поблагодарить KNOWARTH Technologies за предоставленную возможность и поддержку стать частью данной книги. Я также хочу поблагодарить моих соавторов и команду Packt Publishing за замечательную поддержку. Особенно хотел бы поблагодарить моих маму, папу, жену Сакину, прекрасного сына Мохаммада и членов семьи за поддержку на протяжении всего проекта.

Джадип Лакхатария имеет богатый опыт работы в портальных и J2EE-платформах, быстро адаптируется к любой новой технологии и неуклонно работает над совершенствованием своих знаний. В настоящее время Джадип связан с ведущей компанией по внедрению открытого исходного кода на предприятиях, KNOWARTH Technologies (www.knowarth.com), где занимается различными предпринимательскими проектами.

Джадип, являясь полностековым разработчиком, доказал свою универсальность благодаря освоению таких технологий, как Liferay, Java, Spring, Struts, Hadoop, MySQL, Elasticsearch, Cassandra, MongoDB, Jenkins, SCM, PostgreSQL, и многих других.

Он был награжден премиями, в частности за заслуги, приверженность службе, а также в качестве звездного исполнителя. Он любит выступать наставником и проводит курсы по темам порталов и платформ J2EE.

Я искренне благодарен моим прекрасным соавторам, и особенно г-ну Чинтан Мехта, за мотивацию и веру в меня. Хотел бы поблагодарить KNOWARTH за постоянное предоставление новых возможностей самосовершенствования. Хотел бы также поблагодарить всю команду Packt Publishing за замечательную поддержку на протяжении всего проекта.
Наконец, я очень благодарен моим родителям и младшему брату Кейуру за то, что они поддерживали меня на протяжении всего путешествия. Спасибо моим друзьям и коллегам за поддержку.

Чинтан Мехта является соучредителем KNOWARTH Technologies (www.knowarth.com) и возглавляет направление Cloud/RIMS/DevOps. Он имеет богатый про-

грессивный опыт в операционных системах, администрировании серверов Linux, AWS Cloud, DevOps, RIMS и администрировании серверов на технологиях с открытым исходным кодом. Он также является сертифицированным архитектором решений AWS.

На протяжении своей карьеры в области инфраструктуры и операций жизненно важная роль Чинтан также проявлялась в анализе требований, проектировании архитектуры и безопасности, планировании высокой доступности и аварийного восстановления, автоматизированном мониторинге и развертывании, помощи клиентам в процессах сборки, настройки производительности, развертывании и настройки инфраструктуры, а также в настройке и развертывании приложений. Он также отвечал за создание различных офисов в разных местах, с фантастической единоличной ответственностью за достижение оперативной готовности для организаций, с которыми он был связан.

У своего предыдущего работодателя он возглавлял отдел управляемых облачных служб и получил множество премий в знак признания высокоценного вклада в деятельность группы. Он также возглавлял команды внедрения ISO 27001:2005 в качестве представителя совместного руководства компании. Чинтан является автором книги «*Решения для резервного копирования и восстановления Hadoop*» (Hadoop Backup and Recovery Solutions), а также был рецензентом книг «*Рекомендации по повышению производительности портала Liferay*» (Liferay Portal Performance Best Practices) и «*Создание бессерверных веб-приложений*» (Building Serverless Web Applications).

Он обладает дипломом по компьютерному оборудованию и сетям известного в Индии института.

> *При написании этой книги я опирался на многих людей, как прямо, так и косвенно. Прежде всего хотел бы поблагодарить моих соавторов и замечательную команду PacktPub за свои усилия. Я хотел бы особенно поблагодарить мою замечательную жену, Миттал, и моего милого сына, Девам, за то, что они с достоинством вытерпели все те долгие дни, ночи и выходные, когда я ночевал в своем ноутбуке. Многие люди вдохновили и внесли свой вклад в эту книгу и предоставили комментарии, правки, мысли и идеи, в особенности Крупал Кхатри и Чинтан Гаджар. Ничто не могло помешать моей книге. Я также хочу поблагодарить всех рецензентов этой книги.*
>
> *И последнее, но не менее важное: я хочу поблагодарить моих маму и папу, друзей, семью и коллег за поддержку на протяжении всего времени написания данной книги.*

Кандарп Патель возглавляет направление PHP в KNOWARTH Technologies (www.knowarth.com). Он обладает обширным опытом в обеспечении сквозных решений в CMS, LMS, WCM и электронной коммерции, а также во внедренческих проектах, связанных с различными интеграциями, для корпоративных клиентов. Его богатый опыт в предоставлении решений с использованием MySQL, MongoDB и платформ на основе PHP насчитывает более чем 9 лет. Кандарп также является сертифицированным разработчиком MongoDB и Magento.

Кандарп обладает опытом разработки корпоративных приложений на различных этапах жизненного цикла разработки программного обеспечения и играет важную роль в сборе требований, разработке архитектуры, разработке баз данных, разработке приложений, настройке производительности и CD/CI.

Кандарп имеет степень бакалавра технических наук в области информационных технологий из известного в Индии университета.

Хотел бы отметить: Чинтан Мехта вел меня через различные этапы американских горок при написании книги. Хотел бы поблагодарить KNOWARTH Technologies за предоставленную мне возможность стать частью этой книги. Кроме того, хотел бы поблагодарить моих великолепных соавторов и команду PacktPublishing за замечательную поддержку на протяжении всего путешествия.

Последнее, но не маловажное, я хочу поблагодарить моих маму и папу, и мою жену Халапа, за постоянную поддержку и поощрения во время написания этой книги. Я посвящаю свою первую книгу моим прекрасным принцессам, Джейне и Джайсви.

О рецензентах

Анкит Бхавсар является старшим консультантом KNOWARTH Technologies (www.knowarth.com) и возглавляет команду, работающую над решениями для планирования корпоративных ресурсов. Он обладает богатыми познаниями Java, JEE, MySQL, PostgreSQL, Apache Spark и многих других инструментов и технологий с открытым исходным кодом, используемых при создании приложений корпоративного уровня.

В своей карьере Анкит решал разные задачи – как архитектор приложения и программист, архитектор структуры баз данных для портала Astrology, включая объектно-ориентированное программирование, технический архитектурный анализ, проектирование и разработку, а также проектирование, развитие, совершенствование и обработку баз данных, моделирование данных и объектов в широком спектре приложений и сред для обеспечения технических и бизнес-решений для клиентов.

Анкит имеет степень магистра по компьютерным приложениям Университета Северного Гуджарата.

Прежде всего хотел бы поблагодарить моих рецензентов и замечательную команду Packt Publishing за их усилия. Я также хотел бы поблагодарить Субхаш Шаха и Чинтан Мехта. Еще хочу поблагодарить всех авторов этой книги. И последнее, но не менее важное: хочу поблагодарить мою маму, друзей, семью и коллег за поддержку на протяжении всего рецензирования этой книги.

Гаджар Чинтан является консультантом по технологиям в KNOWARTH (www.knowarth.com). Он имеет богатый передовой опыт в JavaScript, NodeJS, BackboneJS, Angularjs, Java и MongoDB, а также предоставляет услуги корпоративного уровня, в частности услуги в области разработки корпоративных порталов, внедрения ERP и интеграции предприятий с привлечением технологий с открытым исходным кодом.

На протяжении всей своей карьеры в сфере корпоративных услуг Чинтан также играл жизненно важную роль в оказании помощи клиентам по вопросам анализа требований, архитектурному дизайну, реализации пользовательского интерфейса и процесса сборки, следуя лучшим практическим рекомендациям по разработке и процессам с доведением порученных проектов до стадии развертывания и реализации идей клиента и организаций, с которыми он был связан.

На протяжении своей карьеры Чинтан активно участвовал в развитии корпоративных решений, связанных с планированием ресурсов, работал над разработкой одностраничного приложения (SPA), а также над мобильным приложением с привлечением NodeJS, MongoDB и AngularJS. Работа Чинтана получила большое признание за весьма ценный вклад, внесенный в команду и компанию в целом. Чинтан участвовал в написании книги «*Решения для резервного копирования и восстановления Hadoop*» (Hadoop Backup and Recovery Solutions). Он имеет степень магистра по компьютерным приложениям (CMA) университета Ганпат.

Хотел бы поблагодарить основных сорецензентов и замечательную команду Packt Publishing за их усилия. Также хотел бы поблагодарить Субхаш Шаха, Чинтан Мехту и Анкита Бхавсара и его коллег за поддержку, оказанную мне в ходе рецензирования этой книги. Благодарю всех авторов этой книги.

Никундж Ранпура обладает богатым передовым опытом в операционных системах и администрировании серверов Linux, AWS Cloud, Devops, RIMS, сетях, системах хранения данных, резервном копировании, обеспечении безопасности и администрирования серверов на основе технологий с открытым исходным кодом. Он быстро адаптируется к любой технологии и имеет острое желание постоянного совершенствования. Он также является сертифицированным архитектором решений AWS.

В своей карьере Никундж выступал в различных ролях, в том числе как системный аналитик, ИТ-менеджер, руководитель направления управляемых облачных служб, архитектор инфраструктуры, разработчик инфраструктуры, архитектор DevOps, архитектор AWS и менеджер поддержки для различных крупных реализаций. Он принимал участие в создании решений и консультировании по созданию служб SaaS, IAAS и PAAS в облаке.

В настоящее время Никундж связан с ведущей компанией по внедрению открытого исходного кода на предприятиях, KNOWARTH Technologies, в качестве ведущего консультанта, где он занимается корпоративными проектами, помогая клиентам в отношении анализа требований, проектирования архитектуры, проектирования безопасности, высокой доступности и планирования аварийного восстановления, а также в руководстве командой.

Никундж окончил Университет Бхавнагара и прошел сертификацию по межсетевым экранам CISCO и UTM. Он был отмечен двумя наградами за свой ценный вклад в компанию. Он также является участником Stack Overflow. С ним можно связаться по адресу ranpura.nikunj@gmail.com.

Хотел бы поблагодарить мою семью за их огромную поддержку и веру в меня на протяжении всей моей стадии обучения. Мои друзья выразили ко мне такое доверие, которое заставляет меня делать лучшее, на что я способен. Я счастлив, что Бог благословил меня такими замечательными людьми, которые меня окружают и без которых мой сегодняшний успех был бы невозможен.

Субхаш Шах является архитектором программного обеспечения с более чем 11-летним опытом разработки веб-ориентированных программных решений на основе различных платформ и языков программирования. Он является энтузиастом объектно-ориентированного программирования и ярым сторонником разработки свободного программного обеспечения с открытым исходным кодом, а также его использования предприятиями для снижения риска, уменьшения затрат и обеспечения большей гибкости. Его карьерные интересы включают проектирование устойчивых программных решений. Лучшие из его технических навыков не ограничиваются анализом требований, проектированием архитектуры, мониторингом доставки проекта, настройкой приложений и инфраструктуры и настройкой процесса выполнения. Он является поклонником написания качественного кода и тестирования.

Субхаш работает главным консультантом в KNOWARTH Technologies Pvt Ltd. и возглавляет ERP-направление. Он получил степень бакалавра в области информационных технологий Университета Северного Гуджарата, Хемчандрачарья.

Приятно выступать рецензентом этой книги. Хотел бы поблагодарить команду Packt Publishing за предоставление такой возможности. Хотел бы поблагодарить мою семью за поддержку на протяжении рецензирования этой книги. Было бы трудно, если бы они не понимали мои приоритеты и не были источником вдохновения. Хочу поблагодарить коллег за постоянную поддержку и помощь. Наконец, хочу поблагодарить авторов за написание такого полезного и подробного контента.

Предисловие

Среди организаций, обрабатывающих крупные объемы данных на регулярной основе, реляционная система управления базами данных MySQL стала популярным решением для работы со *структурированными большими данными*. В этой книге вы познакомитесь с тем, как **Администраторы баз данных** (**АБД**) могут использовать MySQL для обработки миллиардов записей и загрузки и извлечения данных с производительностью, сравнимой или превосходящей коммерческие решения для СУБД с более высокими затратами.

Многие организации сегодня зависят от MySQL для своих веб-сайтов и решений по обработке больших данных в плане своих потребностей в архивировании, хранении и анализе данных. Однако их интеграция может оказаться сложной задачей. Эта книга покажет, как реализовывать успешную стратегию больших данных с помощью Apache Hadoop, вычислительной платформы для разработки и выполнения распределенных программ и MySQL 8. В ней будут рассмотрены варианты сценариев использования в режиме реального времени, которые объяснят способы интеграции и достижения решений по обработке больших данных с использованием различных технологий, таких как Apache Hadoop, Apache Sqoop и MySQL Applier.

В книге, в частности, будут рассмотрены такие темы, как особенности MySQL 8, практические рекомендации по использованию MySQL 8 и API NoSQL, предоставляемого этой реляционной СУБД, а также будет приведен пример использования MySQL 8 для управления большими данными. В конце этой книги вы узнаете, как эффективно использовать MySQL 8 в целях управления конкретными данными приложений, предназначенных для обработки больших данных.

О чем эта книга рассказывает

Глава 1 «*Введение в большие данные и MySQL 8*» содержит обзор больших данных и MySQL 8, их важность и жизненный цикл больших данных. В ней рассматривается основной принцип больших данных и их тенденции на текущем рынке. Наряду с этим эта глава также посвящена объяснению преимуществ использования MySQL, она проведет нас в пошаговом режиме по процессу инсталляции MySQL 8 и познакомит с недавно введенными в MySQL 8 функциональными средствами.

Глава 2 «*Методы запроса данных в MySQL 8*» посвящена основам запросов данных в MySQL 8 и тому, как соединять или агрегировать в ней набор данных.

Глава 3 «*Индексирование данных для высокопроизводительных запросов*» подробно объясняет индексирование в MySQL 8, вводит различные типы индексирования, которые имеются в MySQL, и показывает, как выполнять индексирование для более быстрой работы на крупных объемах данных.

Глава 4 «*Использование Memcached с MySQL 8*» предоставляет обзор работы с Memcached в MySQL и информирует о различных преимуществах использования этого плагина. В ней рассматриваются шаги установки Memcached, конфигурирование репликации и различные API Memcached на разных языках программирования.

Глава 5 «*Разделение данных больших объемов*» объясняет, каким образом в MySQL 8 можно разделять крупные объемы данных, используя различные методы разделения. Она охватывает различные типы разделения, которые можно реализовать в MySQL 8, и их использование с большими данными.

Глава 6 «*Репликация для построения высокодоступных решений*» объясняет реализацию групповой репликации в MySQL 8. В главе рассказывается о том, каким образом большие данные можно масштабировать и как репликация данных может становиться быстрее с использованием различных методов репликации.

Глава 7 «*Практические рекомендации по работе с MySQL 8*» посвящена лучшим приемам использования MySQL 8 для больших данных. Она содержит разнообразные советы в отношении того, что можно и чего нельзя делать при использовании MySQL 8.

Глава 8 «*Прикладной программный интерфейс NoSQL для интегрирования с решениями для больших данных*» объясняет интеграцию API NoSQL для получения данных. В ней также даются пояснения в отношении технологии NoSQL и ее различных API на разных языках программирования для соединения NoSQL с MySQL.

Глава 9 «*Практический пример: часть I. Apache Sqoop для обмена данными между MySQL и платформой Hadoop*» объясняет, как массовые данные могут эффективно передаваться между Hadoop и MySQL с помощью Apache Sqoop.

Глава 10 «*Практический пример: часть II. Обработка событий в реальном времени с помощью MySQL Applier*» объясняет интеграцию MySQL в режиме реального времени с Hadoop, чтение событий двоичного журнала сразу после их фиксации и запись их в файл в распределенной файловой системе HDFS.

Что требуется для этой книги

Эта книга послужит для вас гидом по установке всех инструментов, которые вам потребуются, чтобы проследить работу приводимых примеров. Для эффективного выполнения примеров кода, представленных в данной книге, необходимо установить следующее программное обеспечение:

- MySQL 8.0.3;
- Hadoop 2.8.1;
- Apache Sqoop 1.4.6.

Для кого эта книга предназначена

Эта книга предназначена для администраторов баз данных MySQL и профессиональных специалистов по большим данным, которые хотят интегрировать MySQL и Hadoop с целью реализации высокопроизводительного решения по обработке больших данных. Некоторый предыдущий опыт работы с реляционной СУБД MySQL будет полезен.

Условные обозначения

В этой книге вы найдете ряд текстовых стилей, которые выделяют различные виды информации. Вот некоторые примеры этих стилей и объяснение их значения.

Кодовые слова в тексте, имена таблиц баз данных, папок, файлов, расширения файлов, пути, фиктивные URL-адреса, ввод данных пользователем и дескрипторы

социальной сети Twitter показаны следующим образом: «И уже установленные пакеты могут быть обновлены при помощи флага include».

Фрагмент исходного кода оформляется следующим образом:

```
[default]
exten => s,1,Dial(Zap/1|30)
exten => s,2,Voicemail(u100)
exten => s,102,Voicemail(b100)
exten => i,1,Voicemail(s0)
```

Когда мы хотим обратить ваше внимание на определенную часть фрагмента кода, соответствующие строки или элементы выделены жирным шрифтом:

```
[default]
exten => s,1,Dial(Zap/1|30)
exten => s,2,Voicemail(u100)
exten => s,102,Voicemail(b100)
exten => i,1,Voicemail(s0)
```

Любой ввод или вывод командной строки записывается следующим образом:

```
# cp /usr/src/asterisk-addons/configs/cdr_mysql.conf.sample
/etc/asterisk/cdr_mysql.conf
```

Новые термины и важные слова выделены жирным шрифтом. Слова, которые вы видите на экране, например в меню или диалоговых окнах, появляются в тексте следующим образом: «Нажатие кнопки **Next** (Далее) переводит вас на следующий экран».

Предупреждения или важные примечания отображаются в этом поле.

Советы и приемы появляются тут.

Отзывы и пожелания

Мы всегда рады отзывам наших читателей. Расскажите нам, что вы думаете об этой книге – что понравилось или, может быть, не понравилось. Отзывы важны для нас, чтобы выпускать книги, которые будут для вас максимально полезны.

Вы можете написать отзыв прямо на нашем сайте www.dmkpress.com, зайдя на страницу книги, и оставить комментарий в разделе «Отзывы и рецензии». Также можно послать письмо главному редактору по адресу dmkpress@gmail.com, при этом напишите название книги в теме письма.

Если есть тема, в которой вы квалифицированы, и вы заинтересованы в написании новой книги, заполните форму на нашем сайте по адресу http://dmkpress.com/authors/publish_book/ или напишите в издательство по адресу dmkpress@gmail.com.

Скачивание исходного кода примеров

Скачать файлы с дополнительной информацией для книг издательства «ДМК Пресс» можно на сайте www.dmkpress.com или www.дмк.рф на странице с описанием соответствующей книги.

Список опечаток

Хотя мы приняли все возможные меры для того, чтобы удостовериться в качестве наших текстов, ошибки все равно случаются. Если вы найдете ошибку в одной из наших книг — возможно, ошибку в тексте или в коде, — мы будем очень благодарны, если вы сообщите нам о ней. Сделав это, вы избавите других читателей от расстройств и поможете нам улучшить последующие версии этой книги.

Если вы найдете какие-либо ошибки в коде, пожалуйста, сообщите о них главному редактору по адресу dmkpress@gmail.com, и мы исправим это в следующих тиражах.

Нарушение авторских прав

Пиратство в Интернете по-прежнему остается насущной проблемой. Издательства «ДМК Пресс» и Packt очень серьезно относятся к вопросам защиты авторских прав и лицензирования. Если вы столкнетесь в Интернете с незаконно выполненной копией любой нашей книги, пожалуйста, сообщите нам адрес копии или вебсайта, чтобы мы могли применить санкции.

Пожалуйста, свяжитесь с нами по адресу электронной почты dmkpress@gmail.com со ссылкой на подозрительные материалы.

Мы высоко ценим любую помощь по защите наших авторов, помогающую нам предоставлять вам качественные материалы.

Глава 1

Введение в большие данные и MySQL 8

Сегодня мы живем в эпоху цифровизации. Мы производим огромное количество данных по многим направлениям: социальные сети, покупки в продуктовых магазинах, транзакции по банковским/кредитным картам, электронные письма, хранение данных в облаках и т. д. Один из первых вопросов, который приходит на ум: получаете ли вы максимальную отдачу от собранных данных? Для этого цунами данных мы должны иметь соответствующие инструменты, чтобы можно было получать эти данные в организованном порядке, который мог бы использоваться в различных областях, таких как научные исследования, передача данных в режиме реального времени, борьба с преступностью, обнаружение мошенничества, цифровая персонализация и т. д. Все эти данные должны фиксироваться, храниться, отыскиваться, совместно использоваться, передаваться, анализироваться и визуализироваться.

Анализ структурированных, неструктурированных или полуструктурированных данных помогает нам обнаруживать скрытые закономерности, тенденции рынка, корреляции, личные предпочтения и т. д. С помощью правильных инструментов обработки и анализа организация данных может привести к гораздо лучшим маркетинговым планам, дополнительным возможностям получения дохода, улучшению обслуживания клиентов, более здоровой операционной эффективности, конкурентным преимуществам и многому другому.

Каждая компания собирает данные и их использует; однако, чтобы обеспечить потенциальный успех, компания должна использовать данные эффективнее. Каждая компания должна создавать прямые ссылки на выведенные данные, которые могут улучшить бизнес прямо или косвенно.

Хорошо, теперь у вас есть большие данные, под которыми обычно подразумевается крупный объем данных, и вы выполняете анализ – разве это все, что вам нужно? Держитесь! Другим наиболее важным фактором является успешная монетизация данных. Итак, приготовьтесь и пристегните ремни – сейчас мы объясним важность больших данных!

В этой главе мы рассмотрим приведенные ниже разделы, чтобы выяснить роль больших данных в сегодняшней жизни и основные шаги установки реляционной СУБД MySQL 8:

- важность больших данных;
- жизненный цикл больших данных;

- что такое структурированная база данных;
- основы MySQL;
- новые функциональные средства, введенные в MySQL 8;
- преимущества использования MySQL 8;
- как установить MySQL 8;
- эволюция MySQL для больших данных.

Важность больших данных

Важность больших данных проистекает не только из того, сколько данных у вас есть, а, скорее, из того, что именно вы собираетесь с данными делать. Данные могут быть получены и проанализированы из непредсказуемых источников и могут быть использованы для решения многих вопросов. Давайте посмотрим на примеры использования, представляющие важность для реальной жизни, построенные на известных сценариях с помощью больших данных.

Следующий ниже рисунок помогает нам понять характер решения по обработке больших данных, обслуживающего различные отрасли промышленности. Хотя это не обширный список отраслей, в которых большие данные играют важную роль в бизнес-решениях, давайте обсудим эти несколько отраслей:

Социальные медиа

Контент социальных медиа – это информация, равно как и взаимодействия, такие как просмотры, лайки, демография, акции, читатели, уникальные посетите-

ли, комментарии и скачивания. Поэтому если рассматривать социальные медиа и большие данные, то они взаимосвязаны. В конце концов, важно то, как ваши усилия, связанные с социальными медиа, способствуют бизнесу.

Я наткнулся на одно замечательное высказывание: нет такого понятия, как доход от инвестиций в социальные медиа, – это во всех смыслах доход от инвестиций в бизнес.

Одним из примечательных примеров возможностей больших данных в Facebook является предоставление информации об образе жизни потребителей, шаблонах поиска, симпатиях, демографии, покупательских привычках и т. д. Facebook хранит около 100 Пб данных и накапливает 500 Тб данных почти ежедневно. С учетом количества подписчиков и собранных данных ожидается, что в ближайшие три года этот объем составит более 60 зеттабайт. Чем больше данных у вас есть, тем больше аналитической обработки вы можете проводить с помощью сложных прецизионных подходов для получения лучшей **отдачи от инвестиций** (ROI). Информация, полученная из социальных медиа, также используется при отслеживании целевой аудитории на предмет привлекательности и прибыльности рекламы.

Facebook имеет умный сервис под названием **Graph Search**, который помогает вам выполнять расширенный поиск по нескольким критериям. Например, вы можете искать *людей мужского пола, живущих в Ахмадабаде, которые работают в KNOWARTH Technologies*. Google также помогает вам уточнять поиск. Такие виды поиска и фильтров ими не ограничиваются; поисковый запрос также может содержать образование, политические взгляды, возраст и имя. Таким же образом вы еще можете запрашивать отели, фотографии, песни и многое другое. Вот здесь как раз и находится отдача от бизнес-инвестиций компании Facebook, которая предоставляет свои рекламные услуги, опирающиеся на конкретные критерии, такие как регионы, интересы или другие специфические особенности пользовательских данных. Google также предоставляет аналогичную платформу под названием **Google AdWords**.

Политика

Эра больших данных играет важную роль в политике; политические партии используют различные источники данных для отслеживания своих избирателей и улучшения своих избирательных кампаний. Аналитическая обработка больших данных также внесла значительный вклад в переизбрание Барака Обамы в 2012 году благодаря повышению вовлеченности и освещению именно тех тем, которые были важны для избирателей.

Нарендра Моди считается одним из самых технологичных и социально-медийно подкованных политиков в мире! У него почти 500 миллионов просмотров в Google+, 30 миллионов подписчиков в Twitter и 35 миллионов лайков в Facebook! Нарендра Моди состоит в партии **Бхартия Джанта** (BJP); анализ больших данных несет главную ответственность за успешные для партии BJP и ее партнеров Индийские всеобщие выборы в 2014 году с использованием инструментов с открытым исходным кодом, которые помогли им войти в прямой контакт со своими избирателями. BJP дотянулась до своих колеблющихся избирателей и даже до тех, кто не был настроен идти голосовать, поскольку они проводили мониторинг разговоров

в социальных сетях, отправляли соответствующие сообщения и использовали тактические приемы, чтобы улучшить свое видение избирательной кампании.

За семь месяцев загодя Нарендра Моди сделал заявление о приоритете туалетов перед храмами, после чего цифровая команда внимательно следила за разговорами в социальных сетях вокруг этого заявления. Было отмечено, что, по крайней мере, 50% пользователей были согласны с этим заявлением. Это был именно тот случай, когда возможность завоевать сердца избирателей была преобразована в миссию Swacch Bharat, что означает гигиеническую Индию. Результаты были ошеломляющими; поддержка партии BJP выросла почти до 30% всего за 50 часов.

Наука и исследование

Знаете ли вы, что с помощью больших данных расшифровка генома человека, которая на самом деле заняла 10 лет, теперь происходит едва ли не за день, а это почти в 100 раз меньше стоимости, предсказываемой законом Мура? Еще в 2000 году, когда **Sloan Digital Sky Survey** (SDSS, Слоановский цифровой небесный обзор), проект широкомасштабного исследования многоспектральных изображений и спектров красного смещения звезд и галактик, начал собирать астрономические данные, этот сбор происходил со скоростью около 200 Гб за ночь, что в то время было намного выше, чем данные, собранные за всю историю астрономии.

Национальное управление по аэронавтике и исследованию космического пространства (NASA) широко использует большие данные, учитывая огромный объем научных исследований. NASA собирает данные со всей Солнечной системы, чтобы раскрыть неизвестную информацию о Вселенной; его массивная коллекция данных является важным активом для науки и исследований и принесла пользу человечеству по различным направлениям. NASA получает данные, хранит их и эффективно использует самыми разными способами. Случаев использования данных агентства NASA так много, что их было бы трудно здесь перечислить!

Энергетика

Одна из ведущих энергетических компаний в Индии помогает улучшать потребление энергии с помощью предсказательного анализа больших данных, помогающего строить более прочные отношения с клиентами. Эта компания подключается к более чем 150 коммунальным услугам и обслуживает более 35 миллионов домашних хозяйств для улучшения использования энергии и снижения затрат и выбросов углерода. Она также предоставляет аналитические отчеты поставщикам коммунальных услуг, состоящих из более чем 10 миллионов точек данных каждый день, с целью целостного обзора использования для аналитических целей. Бытовые клиенты получают эти отчеты в счетах, которые показывают, где потребление энергии может быть уменьшено, и тем самым непосредственно помогают потребителям оптимизировать затраты на энергию.

Обнаружение мошенничества

Когда дело доходит до безопасности, обнаружения мошенничества или соблюдения законодательных требований, большие данные – это ваша вторая половинка, и именно если ваша вторая половинка помогает вам в выявлении и предотвращении проблем, прежде чем они ударят, то это становится «золотой серединой» для

бизнеса. Большую часть времени мошенничество обнаруживается уже после того, как оно произошло, когда вам, возможно, уже был нанесен ущерб. Последующие шаги, очевидно, будут минимизировать воздействие и улучшать области, которые могут помочь вам предотвратить его повторение.

Многие компании, которые вовлечены в самые разные типы обработки транзакций или претензий, широко используют методы обнаружения мошенничества. Платформы больших данных помогают им анализировать транзакции, претензии и т. д. в режиме реального времени, а также тенденции или аномальное поведение для предотвращения мошеннических действий.

Агентство национальной безопасности (АНБ) также занимается аналитической обработкой больших данных, чтобы срывать планы террористов. С помощью передовых методов обнаружения мошенничества на основе больших данных многие органы безопасности используют инструменты больших данных для прогнозирования преступной деятельности, мошенничества с кредитными картами, отлавливают преступников и предотвращают кибератаки. С каждым днем, по мере того как изменяются схемы мошенничества, несоблюдения законодательных требований и нарушения систем безопасности, соответственно, становятся все богаче методы противоборства мошенническим транзакциям со стороны органов безопасности, чтобы идти на шаг впереди для таких нежелательных сценариев.

Здравоохранение

В настоящее время трекер здоровья на запястье – это весьма обыденная вещь; однако с помощью больших данных он не только показывает вашу личную панель мониторинга или изменения показаний с течением времени, но и дает вам соответствующие предложения, основанные на медицинских данных, которые он собирает для улучшения вашего рациона, а также аналитические факты о таких людях, как вы. Таким образом, из простых наручных трекеров здоровья можно получить много сведений, которые могут улучшить здоровье пациента. Компании, предоставляющие такие услуги, также анализируют влияние на здоровье, прослеживая тенденции. Постепенно такие носимые устройства также начинают использоваться в отделениях неотложной медицинской помощи для быстрого анализа характера неотложных восстановительных мер со стороны врача.

Используя данные, накопленные государственными учреждениями, файлы социальных служб, отчеты о травматизме и клинические данные, больницы могут помочь оценить потребности в здравоохранении. Географические статистические данные, основанные на многочисленных факторах, от роста численности населения и заболеваемости до повышения качества жизни людей, сопоставляются для определения наличия медицинских услуг, машин скорой помощи, служб экстренной помощи, планов борьбы с пандемией и других соответствующих служб здравоохранения. Все эти меры, которые осуществляются несколькими учреждениями на регулярной основе для прогнозирования эпидемий гриппа, помогают свести к нулю вероятные экологические опасности, риски для здоровья и негативные тенденции.

Бизнес-картирование

Компания Netflix имеет миллионы подписчиков; она использует большие данные и аналитические данные о привычках подписчиков на основе возраста, пола и географического положения, чтобы индивидуально настраивать те аспекты, ко-

торые, как оказалось, генерируют большую деловую активность в соответствии с ожиданиями компании.

Компания Amazon еще в 2011 году начала присуждать $5 своим клиентам, которые используют мобильное приложение Amazon Price Check для сканирования продуктов в магазине, захвата изображения и поиска самых низких цен. Оно также имело возможность проставлять внутреннюю цену магазина на товары. Роль больших данных заключалась в том, чтобы всю информацию о товарах можно было увязать с товарами Amazon для сравнения цен и тенденций клиентов и, соответственно, планировать маркетинговые кампании и предложения на основе ценных данных, которые аккумулировались, чтобы доминировать на быстро развивающемся конкурентном рынке электронной коммерции.

Компания Макдональдс имеет более чем 35 000 местных ресторанов, которые обслуживают около 75 миллионов клиентов в более чем 120 странах. Она использует большие данные, чтобы получать представление о том, как улучшить опыт клиентов, и предлагает ключевые факторы компании, такие как состав меню, продолжительность ожидания в очереди, размер заказа и модель заказов клиентов, что в совокупности помогает им оптимизировать эффективность своих операций и индивидуальных настроек на основе географических местоположений для обеспечения прибыльности бизнеса.

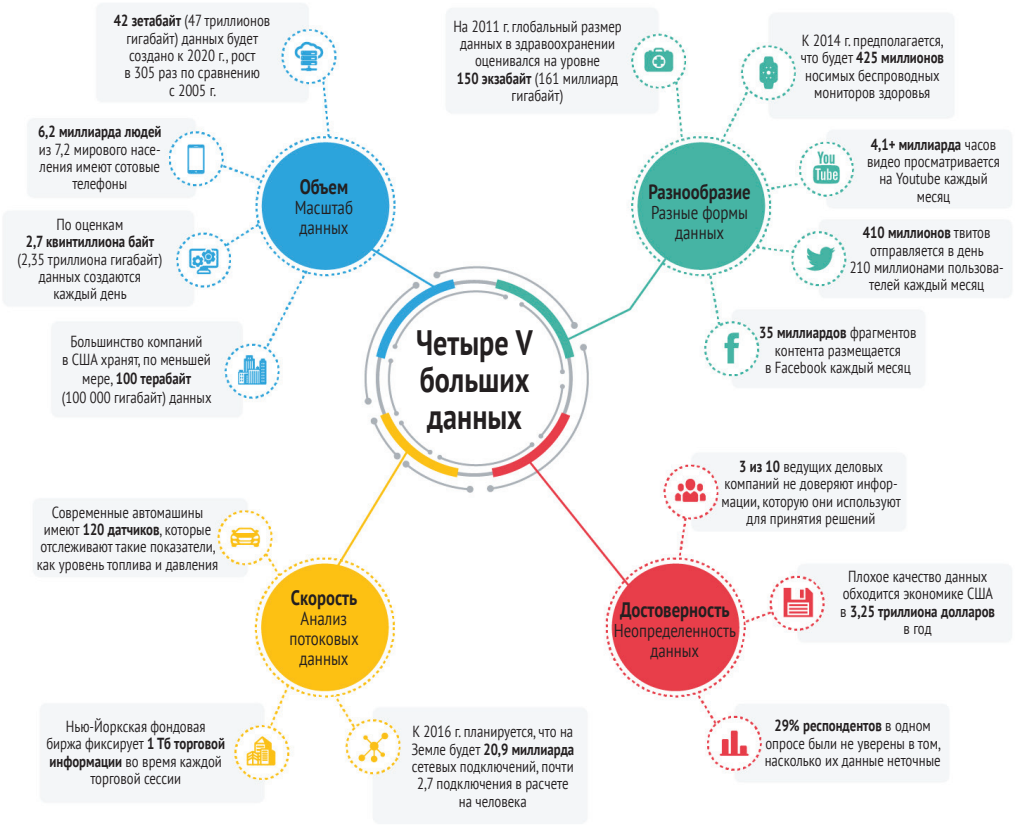
Есть много реальных случаев использования больших данных, которые изменили человечество, технологии, предсказания, здоровье, науку и исследования, закон и порядок, спорт, опыт клиентов, энергетику, финансовую торговлю, робототехнику и многие другие области. Большие данные – это неотъемлемая часть нашей повседневной жизни, которая не всегда очевидна, но, без всяких сомнений, во многих отношениях играет значительную роль в том, что мы делаем. Пришло время подробно рассмотреть то, как структурирован жизненный цикл больших данных. Это даст более четкую картину многих областей, которые играют значительную роль в размещении данных в том месте, где они могут быть использованы для обработки.

Жизненный цикл больших данных

Многие организации рассматривают большие данные не только как модное слово, но и как интеллектуальную систему для улучшения бизнеса и получения соответствующей информации и идей. Большие данные – это термин, который относится к управлению огромным объемом сложных необработанных данных из различных источников, таких как базы данных, социальные медиа, изображения, сенсорное оборудование, файлы журналов регистрации событий, мнения людей и т. д. Эти данные могут быть структурированными, полуструктурированными или неструктурированными. Поэтому для обработки таких данных, которая в условиях традиционных процедур обработки представляла бы собой сложный и трудоемкий процесс, используются специальные инструменты больших данных.

Жизненный цикл больших данных может быть сегментирован на **объем**, **разнообразие**, **скорость** и **достоверность** – обычно известные как **ЧЕТЫРЕ V** (Volume, Variety, Velocity и Veracity) **БОЛЬШИХ ДАННЫХ**. Давайте кратко их рассмотрим, а затем перейдем к четырем фазам жизненного цикла больших данных, то есть сбору, хранению, анализу и управлению данными.

Ниже показано несколько реальных сценариев, которые дают нам гораздо лучшее понимание четырех V и определения больших данных:



Объем

Объем подразумевает огромное количество данных, генерируемых и хранимых каждую секунду. Счет размера данных на предприятиях ведется уже не в терабайтах – он наращивается в зеттабайтах или бронтобайтах. Новые инструменты для работы с большими данными в настоящее время, как правило, используют распределенные системы, которые иногда могут быть диверсифицированы по всему миру.

Ожидается, что объем данных, полученных по всему миру к 2008 году, будет к 2020 году генерироваться всего за минуту.

Разнообразие

Разнообразие относится к нескольким типам и характеру данных, таких как потоки нажатий на веб-страницах, текст, датчики, изображения, голос, видео, файлы журналов регистрации событий, беседы в социальных сетях и многое другое. Это помогает людям, которые тщательно их изучают, эффективно их использовать для углубленного понимания.

70% данных в мире не структурированы, в частности текст, изображения, голос и т. д. Однако ранее структурированные данные были популярны из-за их доступности для анализа, поскольку они могут храниться в файлах, базах данных или поддаются традиционным процедурам хранения данных.

Скорость

Скорость подразумевает скорость генерирования, усвоения и обработки данных для удовлетворения потребностей и решения задач, которые возникают на пути эволюции и расширения.

Каналы связи нового поколения, такие как социальные сети, электронные письма и мобильные телефоны, прибавили скорости данным в больших данных. Процесс ежедневного отслеживания около 1 Тб информации о торговых событиях для выявления мошенничества чувствителен ко времени, когда иногда каждая минута имеет важное значение для предотвращения мошенничества. Просто представьте разговоры в социальных сетях, которые в считанные секунды могут становиться вирусными; на таких платформах анализ помогает нам выявлять тенденции.

Правдивость

Под достоверностью понимается несогласованность данных, которую можно обнаружить; она может повлиять на эффективное управление и обработку данных. Управление такими данными и придание им ценности – вот где могут помочь большие данные.

Когда мы говорим о больших данных, качество и точность остаются главной задачей. Разве не ради этого все крутится? Количество Twitter-каналов является подходящим вариантом применения, где изобилуют хештеги, опечатки, неофициальный текст и аббревиатуры; вместе с тем мы ежедневно сталкиваемся со сценариями, где большие данные отлично справляются со своей работой в серверной части и позволяют нам работать с таким типом данных.

Фазы жизненного цикла больших данных

Эффективное использование больших данных с экспоненциальным ростом в типах и объемах данных имеет огромный потенциал для преобразования экономической, деловой и маркетинговой информации и наращивания клиентской базы. Большие данные стали ключевой мантрой успеха для текущих конкурентных рынков, для существующих компаний и фактором, меняющим правила игры в конкурентной борьбе для новых компаний. Все это может оказаться истиной, только если задействуется **ЦЕННОСТЬ ДАННЫХ**. Давайте посмотрим на следующий ниже рисунок:

Как показано на рисунке, жизненный цикл больших данных можно разделить на четыре этапа. Рассмотрим их подробнее.

Сбор

Этот раздел является ключевым в жизненном цикле больших данных. Он определяет, какой тип данных фиксируется в источнике. В качестве примеров можно привести сбор журналов регистрации событий с сервера, извлечение профилей пользователей, автоматический обзор организаций для анализа мнений и сведений о заказах. Примеры, которые мы упомянули, могут включать в себя работу с локальным языком, текстом, неструктурированными данными и изображениями, в которых мы будем заинтересованы по мере продвижения вперед в жизненном цикле больших данных.

С повышением уровня автоматизации потоков коллекций данных меняются и организации, которые классически тратят много усилий на сбор структурированных данных для анализа и оценки ключевых точек успеха бизнеса. Зрелые организации теперь используют данные, в обычном случае игнорируемые из-за их размера или формата, которые в терминологии больших данных часто называются неструктурированными данными. Эти организации всегда стараются использовать максимальный объем информации, будь то структурированная или неструктурированная, так как для них данные представляют ценность как таковую.

Данные можно передать в платформу больших данных, такую как **HDFS** (Hadoop Distributed File System), и их там консолидировать. После того как данные обработаны с помощью таких инструментов, как Apache Spark, их можно загрузить обратно в базу данных MySQL, которая поможет заполнить соответствующими данными, чтобы показать, из каких составляющих MySQL состоит.

С ростом объемов данных и увеличением скорости теперь Oracle имеет интерфейс NoSQL, предназначенный для подсистем хранения данных InnoDB и MySQL Cluster. Подсистема MySQL Cluster дополнительно полностью обходит слой SQL. Без синтаксического анализа и оптимизации SQL данные в формате ключ-значение могут вставляться прямо в таблицы MySQL в девять раз быстрее.

Хранение

В этом разделе мы обсудим хранение данных, собранных из различных источников. Рассмотрим пример автоматических обзоров организаций для анализа мнений, где каждый обзор собирает данные с разных сайтов и по каждому отображаются уникальные данные.

Традиционно данные обрабатывались с помощью процедуры **ETL** (извлечение, трансформация и загрузка), которая использовалась для сбора данных из различных источников, их изменения в соответствии с требованиями и загрузки в хранилище для дальнейшей обработки или отображения. Для подобных сценариев очень часто использовались такие инструменты, как электронные таблицы, реляционные СУБД, инструменты бизнес-аналитики и т. д., а иногда и вручную.

Наиболее распространенным хранилищем, используемым в платформе больших данных, является HDFS. HDFS также предоставляет в распоряжение язык запросов **HQL** (Hive Query Language), который помогает выполнять многие аналитические задачи, которые традиционно выполняются инструментами бизнес-аналитики. Можно рассмотреть несколько других вариантов хранения – это Apache Spark, Redis и MongoDB. Каждый вариант хранения имеет свой собственный способ работы в серверной части; однако большинство поставщиков хранения предоставляет прикладные программные интерфейсы SQL, которые могут использоваться для дальнейшего анализа данных.

Также может возникнуть ситуация, когда нам нужно собирать и демонстрировать данные в реальном времени, которая практически не требует хранения данных для будущих целей и позволяет выполнять аналитическую обработку в реальном времени для получения результатов на основе запросов.

Анализ

В этом разделе мы обсудим, как эти различные типы данных анализируются на основе универсального вопроса, который начинается со слов «что, если…?». Эволюция организаций вместе с данными также повлияла на новые стандарты метаданных, организующие их с целью первичного обнаружения и переработки для структурных подходов, вызревающих на основе ценности создаваемых данных.

Большинство зрелых организаций надежно обеспечивают доступность, превосходство и ценность для бизнес-подразделений с постоянным автоматизированным процессом структурирования метаданных и результатов, которые будут обрабатываться для анализа. В зрелой организации, управляемой данными, механизм анализа обычно работает с несколькими источниками данных и типами данных, которые также включают в себя данные, поступающие в режиме реального времени.

В фазе анализа обрабатываются сырые данные, для которых СУБД MySQL имеет задания MapReduce в Hadoop, которые проводят анализ и выводят результат. Когда данные MySQL расположены в HDFS, к ним в целях дальнейшего анализа может обращаться остальная часть экосистемы инструментов, связанных с платформой больших данных.

Управление

Невозможно извлечь ценность данных для бизнеса без сформулированной политики управления данными на практике. В отсутствии продуманной политики управления данными предприятия могут столкнуться с неправильной интерпретацией информации, что в конечном итоге может привести к непредсказуемому ущербу для бизнеса. С помощью управления большими данными организация может достигнуть последовательной, точной и действенной осведомленности в данных.

Управление данными осуществляется с целью соблюдения законодательных требований, конфиденциальности, нормативно-юридических актов и всего, что является обязательным в соответствии с требованиями бизнеса. В целях управления данными непрерывный мониторинг, изучение, пересмотр и оптимизация качества процесса также должны учитывать потребности в безопасности данных. До сих пор, когда речь шла о больших данных, управление данными принималось с легкостью; однако, с ростом объемов данных и их повсеместным использованием управление данными стало привлекать к себе все больше внимания. Оно постепенно становится обязательным фактором для любого проекта в области больших данных.

Поскольку у нас есть хорошее понимание жизненного цикла больших данных, давайте теперь подробнее рассмотрим основы MySQL, ее преимущества и несколько превосходных функциональных средств этой реляционной СУБД.

Структурированные базы данных

Многие организации используют структурированную базу данных для хранения своих данных в организованном виде в отформатированном хранилище. В сущности, данные в структурированной базе данных имеют фиксированное поле, предопределенную длину данных, и в них установлено, какие данные там должны храниться, в частности числа, дата, время, адрес, валюта и т. д. Короче говоря, структура уже определена до того, как данные будут вставляться, что дает более четкое представление о том, какие данные могут там находиться. Ключевым преимуществом использования структурированной базы данных является простота хранения, запросов и анализа данных.

Неструктурированная база данных – это полная противоположность; она не имеет идентифицируемой внутренней структуры. Она может иметь массивный неорганизованный агломерат или различные объекты. Источник структурированных данных главным образом генерируется машиной, имея в виду, что информация поступает от машины и без вмешательства человека, в то время как неструктурированные данные генерируются человеком. Организации используют структурированные базы данных для таких данных, как транзакции банкоматов, бронирование авиабилетов, системы инвентаризации и т. д. Точно так же некоторые организации используют неструктурированные данные, такие как электронные письма, мультимедийный контент, текстовые документы, веб-страницы, бизнес-документы и т. д.

Структурированные базы данных – это традиционные базы данных, которые используются многими предприятиями уже более 40 лет. Тем не менее в современном мире объем данных становится все больше и больше, и возникла общая

потребность – аналитическая обработка данных. Со структурированными базами данных аналитика становится все сложнее, поскольку объем и скорость цифровых данных растут быстрее с каждым днем; нам нужно найти способ удовлетворения таких потребностей эффективным и действенным образом. Наиболее распространенной в мире СУБД, которая используется в качестве структурированной базы данных с открытым исходным кодом, является MySQL. Вы познакомитесь с тем, как сделать эту структурированную СУБД пригодной для обработки больших данных, что в итоге приведет к упрощению комплексного анализа. Прежде всего давайте в следующем разделе рассмотрим некоторые идеи MySQL.

Основы MySQL

MySQL – это структурированная реляционная система управления базами данных с открытым исходным кодом, хорошо известная в силу ее производительности, простоты в использовании и надежности. Это наиболее распространенный вариант для веб-приложений на основе реляционной базы данных. На текущем рынке тысячи веб-приложений опираются на MySQL, включая такие гиганты отрасли, как Facebook, Twitter и Wikipedia. Она также зарекомендовала себя в качестве хорошего варианта для **SaaS** (программное обеспечение как служба) на основе таких приложений, как SugarCRM, Supply Dynamics, Workday, RightNow, Omniture и Zimbra. MySQL была разработана шведской компанией MySQL AB, и теперь она распространяется и поддерживается корпорацией Oracle.

MySQL как реляционная система управления базами данных

Данные в реляционной базе данных хранятся в организованном формате, который позволяет легко извлекать информацию. Данные хранятся в различных таблицах, состоящих из строк и столбцов. Вместе с тем также может быть настроена связь между различными таблицами, в которых эффективно хранятся огромные объемы данных и из которых эффективно извлекаются отобранные данные. Это позволяет выполнять операции базы данных с огромной скоростью и гибкостью.

Как реляционная СУБД, MySQL имеет возможности устанавливать связи с различными таблицами по схеме один-ко-многим, многие-к-одному и один-к-одному, предоставляя первичные ключи, внешние ключи и индексы. Для получения точной информации мы также можем выполнять соединения между таблицами, такие как внутренние соединения и внешние соединения.

В MySQL для взаимодействия с реляционными данными в качестве интерфейса используется **язык структурированных запросов SQL** (Structured Query Language). SQL является стандартным, согласно **ANSI** (Американскому национальному институту стандартов), языком, с помощью которого мы можем оперировать данными, выполняя такие операции, как создание, удаление, обновление и извлечение.

Лицензирование

Многие отрасли предпочитают технологии с открытым исходным кодом в силу их гибкости и экономии средств, тогда как MySQL оставила свой след на рынке, став самой популярной реляционной СУБД для веб-приложений. Открытый исходный код означает, что вы можете просматривать исходный код MySQL и настраивать

его под свои потребности без каких-либо затрат. Вы можете скачать исходные или двоичные файлы с сайта MySQL и использовать их по своему усмотрению.

Сервер MySQL подпадает под действие лицензии **GNU** (General Public License, универсальная общедоступная лицензия), что означает, что мы можем свободно его использовать для веб-приложений, соответствующим образом изучать и изменять его исходный код. Он также имеет корпоративную версию с расширенными функциональными возможностями. Многие предприятия приобретают у MySQL корпоративную поддержку, чтобы получать помощь по различным вопросам.

Надежность и масштабируемость

СУБД MySQL работает очень надежно, не требуя широкомасштабного устранения проблем из-за узких мест или других замедлений. Она также включает в себя ряд улучшающих производительность механизмов, таких как поддержка индексов, утилиты загрузки и кеши памяти. MySQL использует InnoDB как подсистему хранения данных, которая обеспечивает очень эффективные ACID-совместимые (с поддержкой транзакционной семантики) транзакционные возможности, гарантирующие высокую производительность и масштабируемость. Для обработки быстро растущей базы данных масштабировать ее помогают подсистемы MySQL Replication и MySQL Cluster.

Совместимость платформ

СУБД MySQL имеет большую кросс-платформенную доступность, что делает ее популярнее. Она гибко работает на основных платформах, таких как Red Hat, Fedora, Ubuntu, Debian, Solaris, Microsoft Windows и Apple macOS. Она также предоставляет **прикладной программный интерфейс** (API) для взаимодействия с различными языками программирования, такими как C, C++, C#, PHP, Java, Ruby, Python и Perl.

Выпуски

Вот список главных версий MySQL, выпущенных до настоящего времени:

- версия 5.0 GA была выпущена 19 октября 2005 г.;
- версия 5.1 GA была выпущена 14 ноября 2008 г.;
- версия 5.5 GA была выпущена 3 декабря 2010 г.;
- версия 5.6 GA была выпущена 5 февраля 2013 г.;
- версия 5.7 GA была выпущена 21 октября 2015 г.

Теперь пришло время для выпуска основной версии – MySQL 8, которая была анонсирована 12 сентября 2016 г. и пока еще находится в режиме разработки. Давайте посмотрим, что нового появилось в последней версии.

Новые возможности в MySQL 8

Команда разработчиков СУБД MySQL недавно анонсировала свой основной релиз как MySQL 8 **Development Milestone Release** (DMR) со значительными обновлениями и исправлениями проблем, которые были очень необходимы в изменениях больших данных.