Amazon Web Services

ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

Технологии Amazon Web Services для обработки больших данных

Компания Gartner Research поместила Amazon Web Services в число лидеров в новом международном отчете «Magic Quadrant for Cloud Infrastructure as a Service». В контексте этого «магического квадранта» облачная модель «инфраструктура как услуга» определяется следующим образом: «стандартизованное высокоавтоматизированное предложение, в котором вычислительные ресурсы, дополненные возможностями хранилища и сетевой конфигурации, принадлежат поставщику услуг и предоставляются пользователям по требованию»

Пятерка лидеров кроет три четверти глобального рынка IaaS

Компания	Выручка, 2018, $ млн	Доля рынка, 2018	Выручка, 2017, $ млн	Доля рынка, 2017	Рост, 2018-2017
Amazon	15 495	47,8%	12 221	49,4%	26,8%
Microsoft	5 038	15,5%	3 130	12,7%	60,9%
Alibaba	2 499	7,7%	1 298	5,3%	92,6%
Google	1 314	4,0%	820	3,3%	60,2%
IBM	577	1,8%	463	1,9%	24,7%
Остальные	7 519	23,2%	6 768	27,4%	11,1%
Всего	32 441	100,0%	24 699	100,0%	31,3%

Этапы обработки больших данных

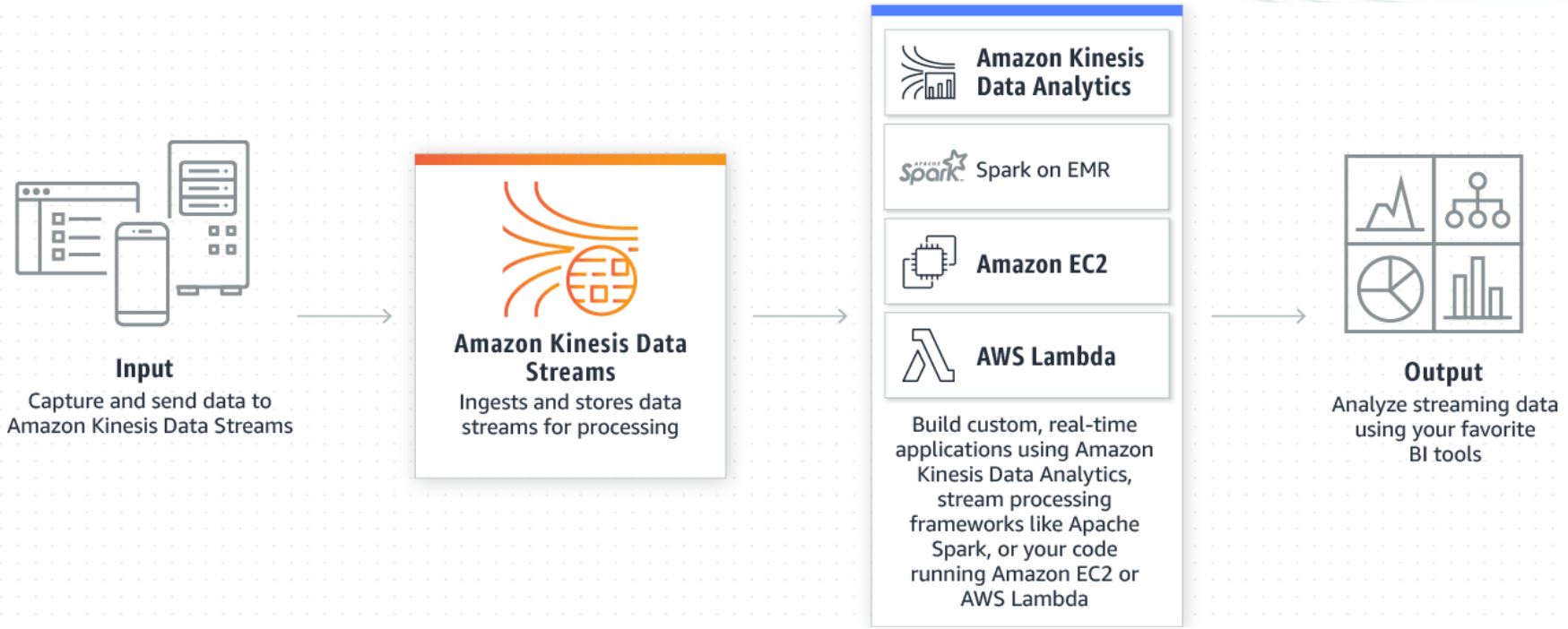
1. Сбор. Сбор необработанных данных (транзакций, записей журналов, событий мобильных устройств и пр.) – это первая проблема, с которой сталкиваются организации при работе с большими данными. Качественная платформа для работы с большими данными упрощает этот этап, предоставляя разработчикам возможность сбора самых разнообразных данных, структурированных и нет, на любой скорости, от режима реального времени до пакетной обработки.

2. Хранение. Любая платформа для работы с большими данными должна включать надежный, безопасный и масштабируемый репозиторий для хранения данных как до обработки, так и после таковой. В зависимости от конкретных требований могут понадобиться и временные хранилища для перемещаемых данных.

3. Обработка и анализ. На этом этапе выполняется преобразование данных из необработанного состояния в пригодный для использования формат. Обычно это достигается за счет сортировки, агрегации, объединения или применения специальных расширенных функций и алгоритмов. После этого итоговые пакеты данных сохраняются для дальнейшей обработки или предоставляются для использования с помощью инструментов бизнес-аналитики и визуализации.

4. Визуализация и использование. Основная цель работы с большими данными – получение на их основании ценных аналитических выводов для практического применения. В идеале большие данные должны становиться доступными для всех заинтересованных сторон, чтобы они получали возможность легко и быстро изучать пакеты данных с помощью инструментов бизнес-аналитики и настраиваемой визуализации, рассчитанных на самостоятельное использование. В зависимости от типа аналитики конечным пользователям могут предоставляться готовые результаты в форме данных статических «прогнозов» (в случае прогнозирующей аналитики) или рекомендованных действий (в случае предписывающей аналитики).

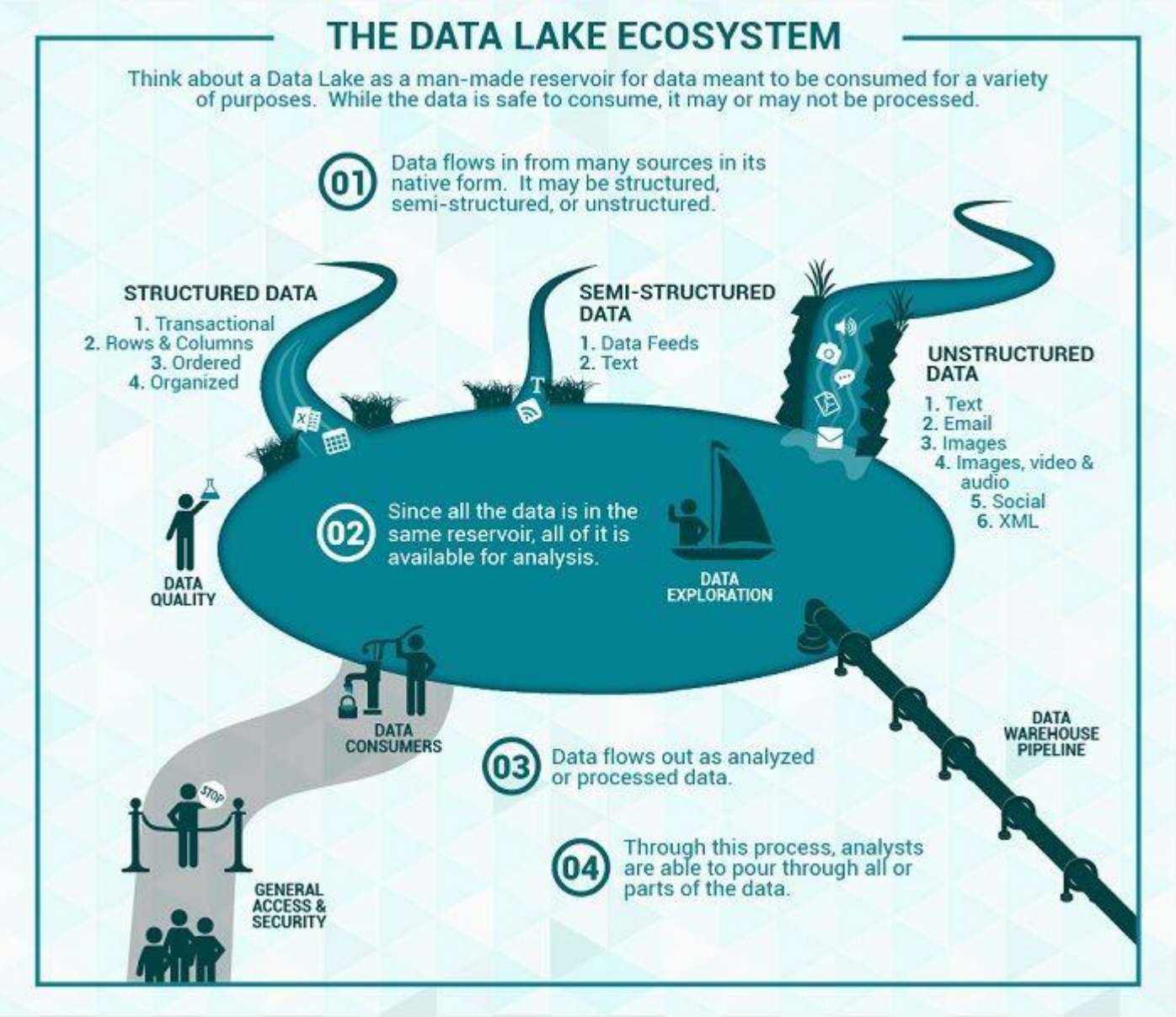
Эволюция обработки больших данных



Изначально инфраструктуры по работе с большими данными, например Hadoop, поддерживали только пакетные рабочие нагрузки. Крупные пакеты данных загружались для обработки сразу, и процесс ожидания результатов растягивался на часы и даже дни.

Но время ожидания результата постепенно стало критическим фактором, и требуемая скорость обработки больших данных послужила толчком к развитию таких новых инфраструктур, как Apache Spark, Apache Kafka, Amazon Kinesis и т. д., способных поддерживать обработку потоковых данных в режиме реального времени.

Что такое озеро данных?



Озеро данных - это централизованное хранилище, которое позволяет хранить все структурированные и неструктурированные данные в любом масштабе. Можно хранить свои данные ***как есть***, без необходимости сначала структурировать данные и запускать различные типы аналитики - от информационных панелей и визуализаций до обработки больших данных, аналитики в реальном времени и машинного обучения для принятия правильных решений.

Озеро данных - это централизованное и защищенное хранилище, в котором хранятся все ваши данные как в исходном виде, так и подготовленные для анализа. Озеро данных позволяет вам разбивать хранилища данных и комбинировать различные виды аналитики, чтобы получать информацию и принимать лучшие бизнес-решения.

Озера данных по сравнению с хранилищами данных - два разных подхода

В зависимости от требований, типичной организации потребуются как **хранилище данных**, так и **озеро данных**, поскольку они служат различным потребностям и вариантам использования.

Хранилище данных - это база данных, оптимизированная для анализа реляционных данных, поступающих из транзакционных систем и бизнес-приложений.
Структура данных и схема определяются ***заранее*** для оптимизации для быстрых запросов SQL, где результаты обычно используются для оперативной отчетности и анализа. Данные очищаются, обогащаются и преобразуются, чтобы они могли действовать как «единый источник правды», которому пользователи могут доверять.

Озеро данных отличается тем, что хранит реляционные данные из бизнес-приложений и нереляционные данные из мобильных приложений, устройств IoT и социальных сетей.
Структура данных или схемы не определяется при загрузке данных. Это означает, что вы можете хранить все свои данные без тщательного проектирования или необходимости знать, на какие вопросы вам могут понадобиться ответы в будущем. Различные типы аналитики ваших данных, такие как запросы SQL, аналитика больших данных, полнотекстовый поиск, аналитика в реальном времени и машинное обучение, могут быть использованы для раскрытия информации.

Поскольку организации с хранилищами данных видят преимущества озер данных, они развивают свое хранилище, чтобы включать в себя озера данных и предоставляют разнообразные возможности запросов, примеры использования данных и расширенные возможности для обнаружения новых информационных моделей. Gartner называет это развитие «Решением для управления данными для аналитики» или «DMSA».

Значение озера данных

Возможность использования большего количества данных из большего количества источников за меньшее время и предоставление пользователям возможности для совместной работы и анализа данных различными способами приводит к лучшему и более быстрому принятию решений.

Проблемы Data Lakes

Основная проблема с архитектурой озера данных заключается в том, что необработанные данные хранятся без контроля содержимого. Чтобы озеро данных могло использовать данные, необходимо иметь определенные механизмы для ***каталогизации*** и защиты данных. Без этих элементов данные не могут быть найдены или заслуживают доверия, что приводит к «***болоту в данных***». Для удовлетворения потребностей более широкой аудитории требуется, чтобы озера данных имели управление, семантическую согласованность и контроль доступа.

Развертывание Data Lakes в облаке

Data Lakes - идеальная рабочая нагрузка для развертывания в облаке, поскольку облако обеспечивает производительность, масштабируемость, надежность, доступность, разнообразный набор аналитических движков и значительную экономию за счет масштаба.

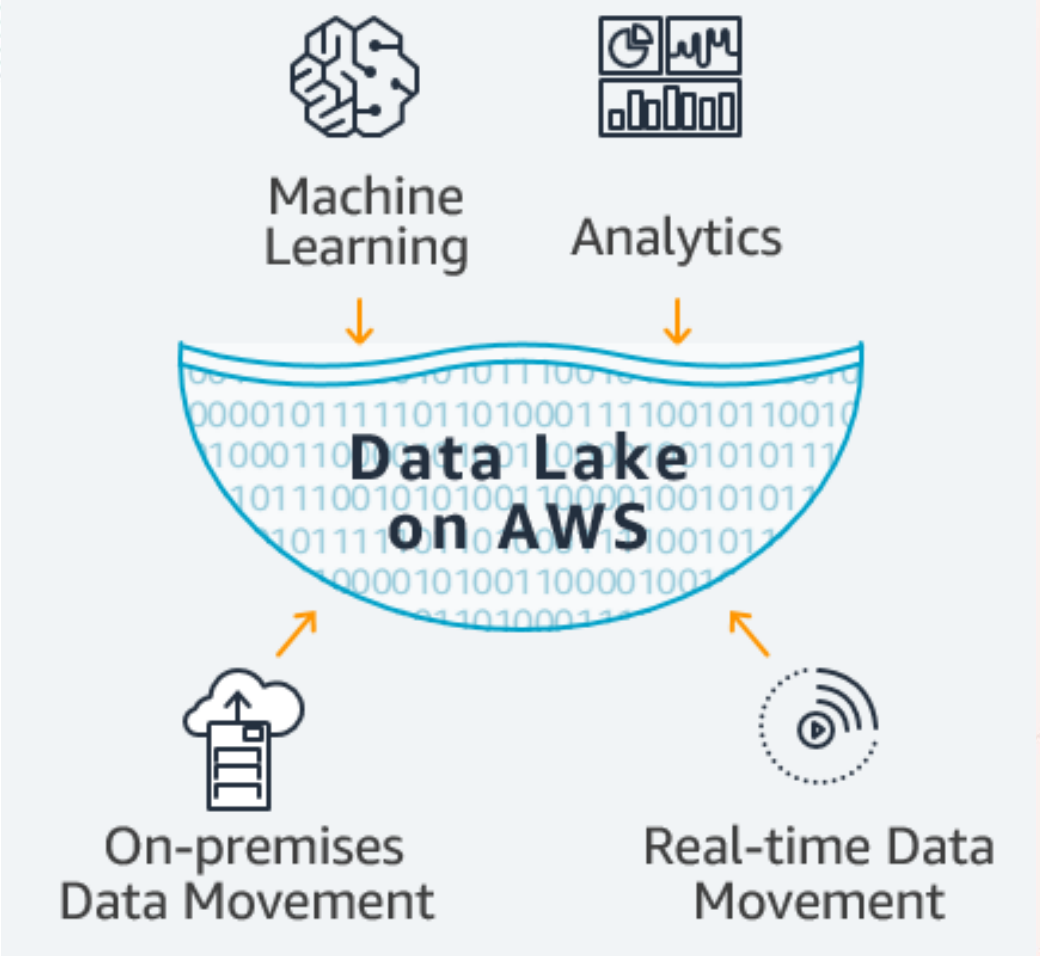
Ключевые возможности Data Lake and Analytics

1. Перемещение данных

Data Lakes позволяют импортировать любое количество данных, которые могут поступать в режиме реального времени. Данные собираются из нескольких источников и перемещаются в озеро данных в ***исходном формате***. Этот процесс позволяет масштабировать данные любого размера, экономя время определения структур данных, схемы и преобразований.

2. Надежно хранить и каталогизировать данные

Data Lakes позволяют хранить реляционные данные, такие как операционные базы данных и данные из бизнес-приложений, а также нереляционные данные, такие как мобильные приложения, устройства IoT и социальные сети. Они также дают вам возможность понять, какие данные находятся в озере, путем сканирования, каталогизации и индексации данных. Наконец, данные должны быть защищены, чтобы обеспечить защиту ваших активов данных.



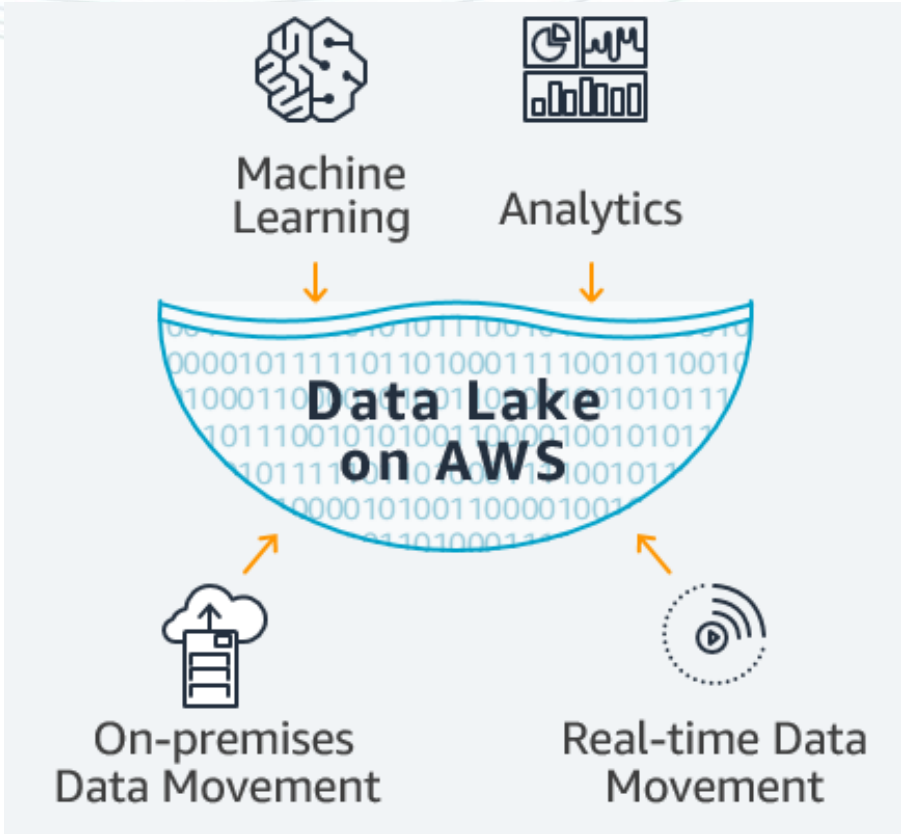
Ключевые возможности Data Lake and Analytics

3. Аналитика

Data Lakes позволяют различным ролям в организации, таким как ученые, разработчики данных и бизнес-аналитики, получать доступ к данным с помощью своих аналитических инструментов и структур. Сюда входят платформы с открытым исходным кодом, такие как Apache Hadoop, Presto и Apache Spark, а также коммерческие предложения от поставщиков хранилищ данных и бизнес-аналитики. Data Lakes позволяют вам запускать аналитику без необходимости переноса данных в отдельную аналитическую систему.

4. Машинное обучение

Data Lakes позволит организациям генерировать различные типы данных, включая отчеты по историческим данным и проведение машинного обучения там, где создаются модели для прогнозирования вероятных результатов, и предлагает ряд предписанных действий для достижения оптимального результата.



Озеро данных на AWS

Объектное хранилище. Amazon S3

Amazon S3 – это надежное и защищенное объектное хранилище с широкими возможностями масштабирования, которое обеспечивает доступ к данным с задержкой на уровне миллисекунд. Сервис S3 предназначен для хранения любых типов данных, поступающих из любых источников: веб-сайтов и мобильных приложений, корпоративных приложений, а также датчиков или устройств IoT.

Хранение данных. Amazon Redshift

Для хранения данных Amazon Redshift предоставляет возможность выполнять комплексные аналитические запросы к массивам структурированных данных в объеме петабайтов, и включает сервис Redshift Spectrum, который исполняет прямые SQL-запросы к массивам из эксабайтов структурированных и неструктурированных данных в S3 без их перемещения.

Озеро данных на AWS

Резервное копирование и архивирование. Amazon Glacier

Amazon Glacier – это надежное, безопасное и очень экономичное хранилище для долговременного хранения архивов и резервных копий, которое обеспечивает доступ к данным в течение считаных минут, а возможность Glacier Select позволяет считывать и извлекать только нужные данные.

Каталог данных. AWS Glue

AWS Glue – это полностью управляемый сервис, который позволяет создавать каталоги для поиска данных в озере. Он обеспечивает возможность выполнять операции по извлечению, преобразованию и загрузке (ETL) данных в целях их подготовки их к анализу. Каталог создается автоматически в форме постоянного хранилища метаданных для всех ресурсов. В результате вести поиск и формировать запросы к данным можно в едином интерфейсе.

Аналитика данных озера на AWS

AWS предлагает самый широкий и экономичный набор сервисов аналитики, которые работают с озерами данных. Каждый аналитический сервис специально спроектирован для широкого спектра примеров использования, таких как интерактивный анализ, обработка больших данных с помощью Apache Spark и Hadoop, хранение данных, анализ в режиме реального времени, операционный анализ, создание информационных панелей и визуализация данных.

Интерактивная аналитика. Amazon Athena

Для задач интерактивной аналитики сервис Amazon Athena упрощает прямой анализ данных в S3 и Glacier с помощью стандартных SQL-запросов.

Обработка больших данных. Amazon EMR

Для обработки больших данных с помощью платформ Apache Spark и Hadoop Amazon EMR предоставляет управляемый сервис, который позволяет быстро, просто и экономично обрабатывать колоссальные объемы данных. Amazon EMR поддерживает 19 различных проектов с открытым исходным кодом, включая Hadoop, Spark, HBase и Presto, с управляемыми блокнотами EMR Notebooks для задач инжиниринга данных, развития науки о данных и организации совместной работы.

Аналитика данных озера на AWS

Аналитика в режиме реального времени. Amazon Kinesis

Для задач аналитики в режиме реального времени Amazon Kinesis позволяет без труда собирать, обрабатывать и анализировать потоковые данные, такие как данные телеметрии с IoT-устройств, журналы приложений и истории навигации по веб-сайтам.

Операционная аналитика. Amazon Elasticsearch Service

Для задач операционной аналитики, таких как мониторинг приложений, анализ журналов и истории навигации по веб-сайтам, сервис Amazon Elasticsearch Service позволяет находить, исследовать, фильтровать, агрегировать и визуализировать данные в режиме реального времени.

Информационные панели и визуализация. Amazon QuickSight

Для создания информационных панелей и визуализации данных Amazon QuickSight предоставляет быстрый облачный сервис бизнес-аналитики, который позволяет без труда создавать потрясающие визуализации и информационные панели, доступные из любого браузера или с любого мобильного устройства.

Amazon EMR

Amazon EMR

Без труда запускайте и масштабируйте Apache Spark, Hive, Presto и другие платформы для работы с большими данными

Amazon EMR – ведущая в отрасли облачная платформа больших данных для обработки огромных объемов информации с использованием инструментов с открытым исходным кодом, таких как Apache Spark, Apache Hive, Apache HBase, Apache Flink, Apache Hudi и Presto. Amazon EMR упрощает установку, использование и масштабирование сред больших данных посредством автоматизации таких трудоемких задач, как выделение объема и настройка кластеров. EMR позволяет анализировать данные в масштабе петабайтов за вдвое меньшую стоимость по сравнению с традиционными локальными решениями и более чем в три раза быстрее по сравнению со стандартным использованием Apache Spark. Вы можете выполнять рабочие нагрузки на инстансах Amazon Elastic Compute Cloud (Amazon EC2), в кластерах Amazon Elastic Kubernetes Service (EKS) или локально с помощью EMR на AWS Outposts (для запуска инфраструктуры и сервисов AWS в локальной среде для обеспечения действительно согласованной гибридной среды).

Amazon EMR

Преимущества

Простота использования. Аналитики, инженеры по обработке данных и специалисты по работе с данными могут использовать блокноты EMR Notebooks, что позволяет изучать, обрабатывать и визуализировать данные в интерактивном режиме общего доступа. Достаточно просто указать версию приложений EMR и тип вычислений, которые вы хотите использовать. EMR берет на себя выделение, конфигурацию и настройку кластеров, тогда как вы можете сосредоточиться на выполнении аналитических задач.
Низкая стоимость. Стоимость EMR формируется просто и предсказуемо: плата начисляется на основе посекундного тарифа за каждую секунду использования каждого инстанса; минимальный порог оплаты составляет одну минуту. Запустить кластер EMR, содержащий 10 узлов, можно всего за 0,15 USD в час. Кроме того, можно сэкономить 50–80 % от стоимости инстансов, выбрав спотовые инстансы Amazon EC2 для временных рабочих нагрузок и зарезервированные инстансы для продолжительных рабочих нагрузок.
Эластичность. В отличие от жесткой инфраструктуры локальных кластеров, EMR изолирует вычислительные ресурсы и постоянные хранилища, позволяя независимо масштабировать их и воспользоваться преимуществом многоуровневого хранения Amazon S3. EMR позволяет выделить от одного до сотен или тысяч вычислительных инстансов либо контейнеров для обработки данных любого масштаба. Оплате подлежат только реально используемые ресурсы.

Amazon EMR

Преимущества

Надежность. Сэкономьте время на настройке и мониторинге кластера. Сервис EMR оптимизирован для работы в облаке и постоянно контролирует кластер: повторно запускает задания, которые не удалось выполнить, и автоматически заменяет инстансы с низкой производительностью.
Безопасность. EMR автоматически настраивает брандмауэр EC2, управляющий сетевым доступом к инстансам, и запускает кластеры в Amazon Virtual Private Cloud (VPC). Можно применять шифрование на стороне сервера или на стороне клиента с использованием AWS Key Management Service или собственных ключей пользователя. EMR позволяет без труда включать и другие варианты шифрования, например шифрование при передаче и при хранении, и усиленную аутентификацию с помощью Kerberos.
Гибкость. Вы полностью контролируете свои кластеры EMR и отдельные задания EMR. Вы можете запускать кластеры EMR с настраиваемыми AMI Amazon Linux и легко настраивать кластеры с помощью сценариев для установки дополнительных сторонних пакетов ПО. EMR позволяет на лету перенастраивать приложения на работающих кластерах без необходимости их перезапуска.

Amazon EMR

Примеры использования

Машинное обучение. Используйте встроенные в EMR инструменты машинного обучения (в том числе Apache Spark MLlib, TensorFlow и Apache MXNet) для работы с масштабируемыми алгоритмами машинного обучения. С помощью настраиваемых AMI и скриптов при начальной загрузке добавляйте выбранные библиотеки и инструменты, чтобы создать собственный инструментарий для прогнозной аналитики.
Извлечение, преобразование и загрузка данных (ETL). EMR можно использовать для быстрого и экономичного выполнения рабочих нагрузок по трансформации данных (извлечение, преобразование и загрузка данных) – сортировке, агрегированию, слиянию – на больших наборах данных.
Анализ истории посещений. Анализируйте данные о посещениях от Amazon S3, используя Apache Spark и Apache Hive, чтобы разделять пользователей на категории, выяснять их предпочтения и показывать более эффективную рекламу.

Amazon EMR

Примеры использования

Потоковая передача в режиме реального времени. Анализ событий от Apache Kafka, Amazon Kinesis и других потоковых источников данных в режиме реального времени с помощью Apache Spark Streaming и Apache Flink, чтобы создавать долгосрочные и устойчивые к ошибкам конвейеры потоковых данных с обеспечением высокой доступности. Сохранение преобразованных наборов данных в S3 или HDFS, а аналитические выводы – в Amazon Elasticsearch Service.

Интерактивная аналитика. Блокноты EMR Notebooks предоставляют управляемую аналитическую среду на основе решения Jupyter с открытым исходным кодом, с помощью которой специалисты по работе с данными, аналитики и разработчики могут подготавливать и визуализировать данные, совместно работать с коллегами, создавать приложения и выполнять интерактивный анализ.

Геномика. EMR можно использовать для быстрой и эффективной обработки больших объемов данных генома и других больших наборов научных данных. Исследователям предоставляется бесплатный доступ к данным генома, хранящимся в AWS.

Машинное обучение

Для задач прогнозной аналитики AWS предоставляет широкий набор сервисов машинного обучения, а также инструментов, работающих с озерами данных в AWS. Все сервисы построены на базе большого объема знаний, накопленных в компании Amazon, которая использует машинное обучение для систем рекомендаций сайта Amazon.com, организации цепочек поставки, прогнозирования, обеспечения работы центров обработки заказов и управления ресурсами.

Платформы и интерфейсы

Специалисты по машинному обучению и работе с данными могут воспользоваться образами AWS Deep Learning AMI, позволяющими просто создавать модели глубокого обучения и кластеры с помощью инстансов с графическими процессорами, оптимизированных для машинного и глубокого обучения. AWS поддерживает все основные платформы машинного обучения, включая Apache MXNet, TensorFlow и Caffe2, что позволяет клиентам использовать или разрабатывать любые подходящие модели. Эти возможности предлагают непревзойденную мощность, скорость и эффективность, столь необходимую для рабочих нагрузок, использующих машинное и глубокое обучение.

Сервисы платформы

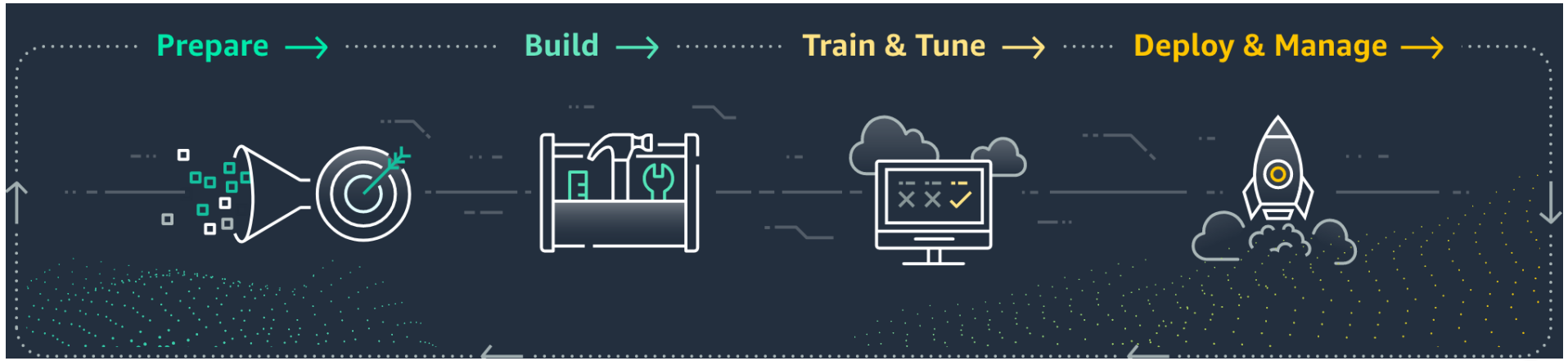
Для разработчиков, которые хотят развиваться в сфере машинного обучения, Amazon SageMaker предлагает платформенный сервис, который упрощает весь процесс создания, обучения и развертывания моделей машинного обучения. Он предоставляет все необходимое для подключения к обучающим данным, выбора и оптимизации лучших алгоритмов и платформ с последующим развертыванием на автомасштабируемых кластерах Amazon EC2. Amazon SageMaker предоставляет размещенные блокноты Jupyter, которые облегчают изучение и визуализацию обучающих данных, хранимых в Amazon S3.

Сервисы приложений

Для разработчиков, которые хотят внедрить в свои приложения готовые функциональные возможности искусственного интеллекта, AWS предлагает ориентированные на использование в таких решениях API машинного зрения и обработки естественного языка. Эти сервисы приложений позволяют разработчикам добавлять в приложения интеллектуальные возможности, не прибегая к разработке и обучению собственных моделей.

Amazon SageMaker

Технологии машинного обучения для любого специалиста по работе с данными и разработчика.
Amazon SageMaker за счет широкого спектра возможностей помогает специалистам по работе с данными и разработчикам в быстрой подготовке, обучении и развертывании высококачественных моделей машинного обучения.
Самый функциональный сервис машинного обучения.
Ускорение инноваций благодаря специализированным средствам для каждого этапа разработки систем машинного обучения, включая разметку и подготовку данных, создание признаков, определение статистического смещения, автоматизированное машинное обучение, обучение, настройка, размещение, анализ, мониторинг и рабочая эксплуатация.



Первая интегрированная среда разработки (IDE) для машинного обучения.
Повышение эффективности работы с помощью Amazon SageMaker Studio, первой полностью интегрированной средой разработки, которая создана специально для машинного обучения и позволяет использовать все необходимое для машинного обучения в едином визуальном пользовательском интерфейсе.
Функциональные возможности, которые сразу создавались с учетом взаимодействия.
Интегрированные возможности Amazon SageMaker для разработки машинного обучения позволяют обойтись без многомесячных трудов по разработке собственного кода для интеграции и значительно снизить затраты.

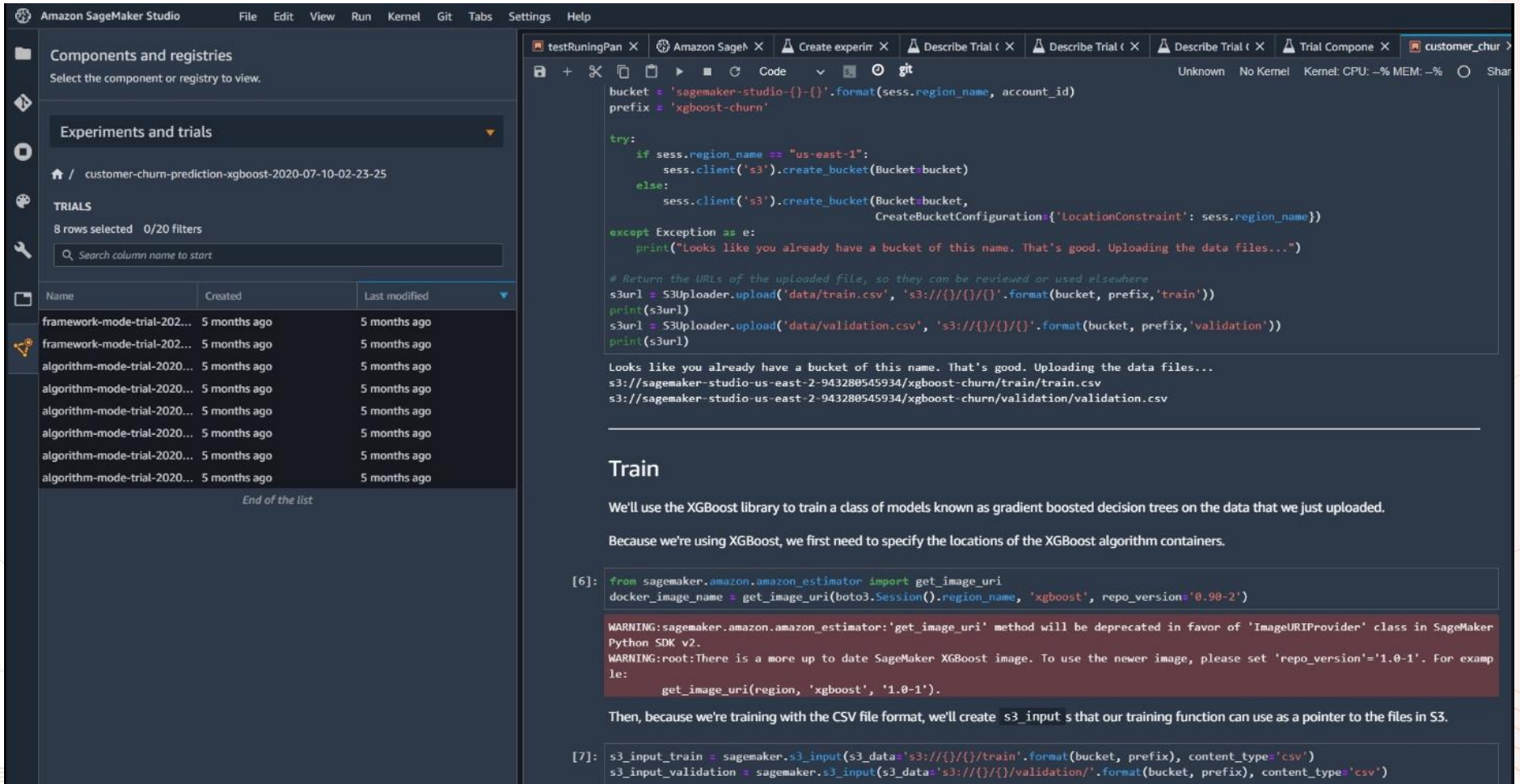
Amazon SageMaker

Amazon SageMaker поддерживает ведущие платформы машинного обучения.

Amazon SageMaker за счет широкого спектра возможностей помогает специалистам по работе с данными и разработчикам в быстрой подготовке, обучении и развертывании высококачественных моделей машинного обучения.

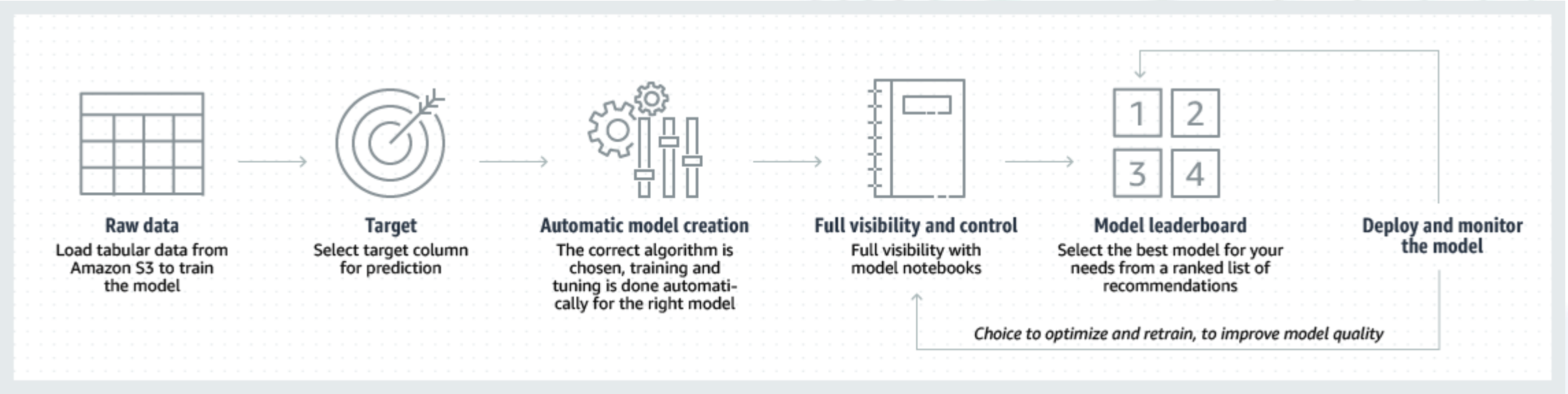
Основные функции для подготовки данных, создания, обучения и развертывания моделей машинного обучения.

Amazon SageMaker Studio предоставляет единый визуальный веб-интерфейс, в котором можно проводить все этапы разработки ML, необходимые для создания, обучения и развертывания моделей.



Amazon SageMaker Autopilot

Автоматическое создание моделей машинного обучения с полной прозрачностью процесса



Основные возможности

Автоматическая предварительная обработка данных и разработка функций.
Вы можете использовать Amazon SageMaker Autopilot даже при отсутствии части данных. SageMaker Autopilot автоматически заполняет недостающие значения, выполняет статистический анализ столбцов в наборе данных и автоматически извлекает информацию из нечисловых столбцов (например, дату и время из меток времени).
Автоматический выбор модели машинного обучения.
Amazon SageMaker Autopilot автоматически определяет тип прогнозов, наиболее подходящий для ваших данных — двоичная классификация, мультиклассовая классификация или регрессия. Затем SageMaker Autopilot изучает высокоэффективные алгоритмы, такие как дерево принятия решений с градиентным бустингом, глубокие нейронные сети с прямой связью и логистическую регрессию, обучает и оптимизирует по этим алгоритмам несколько сотен моделей для поиска оптимального варианта.

Amazon SageMaker Autopilot

Основные возможности

Рейтинг лучших моделей. Amazon SageMaker Autopilot позволяет проверить все модели машинного обучения, автоматически созданные для ваших данных. Вы можете просмотреть список моделей с сортировкой по таким метрикам, как доля правильных ответов, точность, полнота и площадь под кривой, изучить влияние признаков на прогнозы и другие характеристики модели, а также развернуть наиболее подходящую для вашей ситуации модель.

Автоматическое создание блокнота. Для любой модели, созданной в Amazon SageMaker Autopilot, вы можете автоматически создавать блокнот Amazon SageMaker Studio, что позволит изучить подробности создания модели, что-то в них изменить по мере необходимости и восстановить модель из этого блокнота в любой момент в будущем.

Простая интеграция с вашими приложениями. Вы можете использовать интерфейс прикладного программирования (API) Amazon SageMaker Autopilot для упрощения создания моделей и получения выводов прямо из приложения, например в средствах для анализа и хранения данных.

Примеры использования

Прогнозирование цены. Модели прогнозирования цены широко применяются в финансовых сервисах, сфере недвижимости, энергетических и коммунальных компаниях для прогнозирования цен на акции, объекты недвижимости и природные ресурсы. Amazon SageMaker Autopilot позволяет прогнозировать цены и принимать взвешенные решения об инвестициях на основе данных за прошедшие периоды, таких как спрос, сезонные тенденции и стоимость других товаров.

Прогнозирование оттока клиентов. Оттоком клиентов называется уход существующих клиентов, которого стремится избежать любая компания. Модели, автоматически создаваемые в Amazon SageMaker Autopilot, помогают понять тенденции такого оттока. Модели прогнозирования оттока клиентов применяют процессы быстрого обучения к существующим данным и выявляют тенденции в новых наборах данных, позволяя получить прогноз о том, какие клиенты наиболее склонны к уходу.

Оценка риска. Для оценки риска требуется выявление и анализ потенциальных событий, которые могут негативно повлиять на отдельных людей, активы и компанию в целом. Модели, автоматически создаваемые в Amazon SageMaker Autopilot, прогнозируют риски по мере возникновения новых событий. Модели оценки риска обучаются по существующим наборам данных и помогают оптимизировать прогнозы для принятия бизнес-решений.

Сервисы AI

Простое добавление интеллектуальных функций в приложения
Отсутствие необходимости в навыках в сфере машинного обучения

85 %

проектов TensorFlow в облачной среде выполняются в AWS

Предварительно обученные сервисы AI AWS предоставляют готовые интеллектуальные функции, которые можно использовать в своих приложениях и рабочих процессах. Сервисы AI можно без труда интегрировать со своими приложениями для использования в стандартных сценариях использования, например для создания персонализированных рекомендаций, модернизации контакт-центра, повышения безопасности и защищенности и увеличения активности клиентов. Так как используются те же технологии глубокого обучения, которые применяются для сервиса Amazon.com и сервисов ML, вы получаете качество и точность, обеспечиваемые непрерывно обучаемыми API. И самое лучшее заключается в том, что для использования сервисов AI на AWS не требуется опыт в сфере машинного обучения.

Рекомендации

Персонализируйте информацию, предоставляемую вашим клиентам, с помощью той же технологии создания рекомендаций, которая применяется в сервисе Amazon.com.

Прогнозирование

Создавайте модели для точного прогнозирования на основе той же технологии машинного обучения для прогнозирования, которая применяется в сервисе Amazon.com.

Анализ изображений и видео

Добавьте в свои приложения функции анализа изображений и видео, чтобы каталогизировать ресурсы, автоматизировать рабочие процессы обработки мультимедиа и извлекать смысловую информацию.

Сервисы AI

Расширенная текстовая аналитика

Используйте функции обработки естественных языков, чтобы извлекать необходимые сведения и информацию о связях из неструктурированного текста.

Анализ документов

Автоматически извлекайте текст и данные из миллионов документов за считанные часы, сокращая объемы ручного труда.

Голосовая связь

Включите функцию преобразования текста в речь с естественным звучанием для своих приложений.

AMAZON POLLY »

Диалоговые агенты

Легко создавайте диалоговые агенты для повышения качества обслуживания клиентов и увеличения эффективности работы контакт-центров.

Перевод

Расширьте свою аудиторию, включив в нее клиентов, разговаривающих на разных языках, с помощью эффективных и экономически эффективных функций

Транскрибирование

Легко добавляйте высококачественные функции преобразования речи в текст для приложений и рабочих процессов.

Amazon QuickSight

Масштабируемый, бессерверный, встроенный облачный сервис бизнес-аналитики на основе машинного обучения

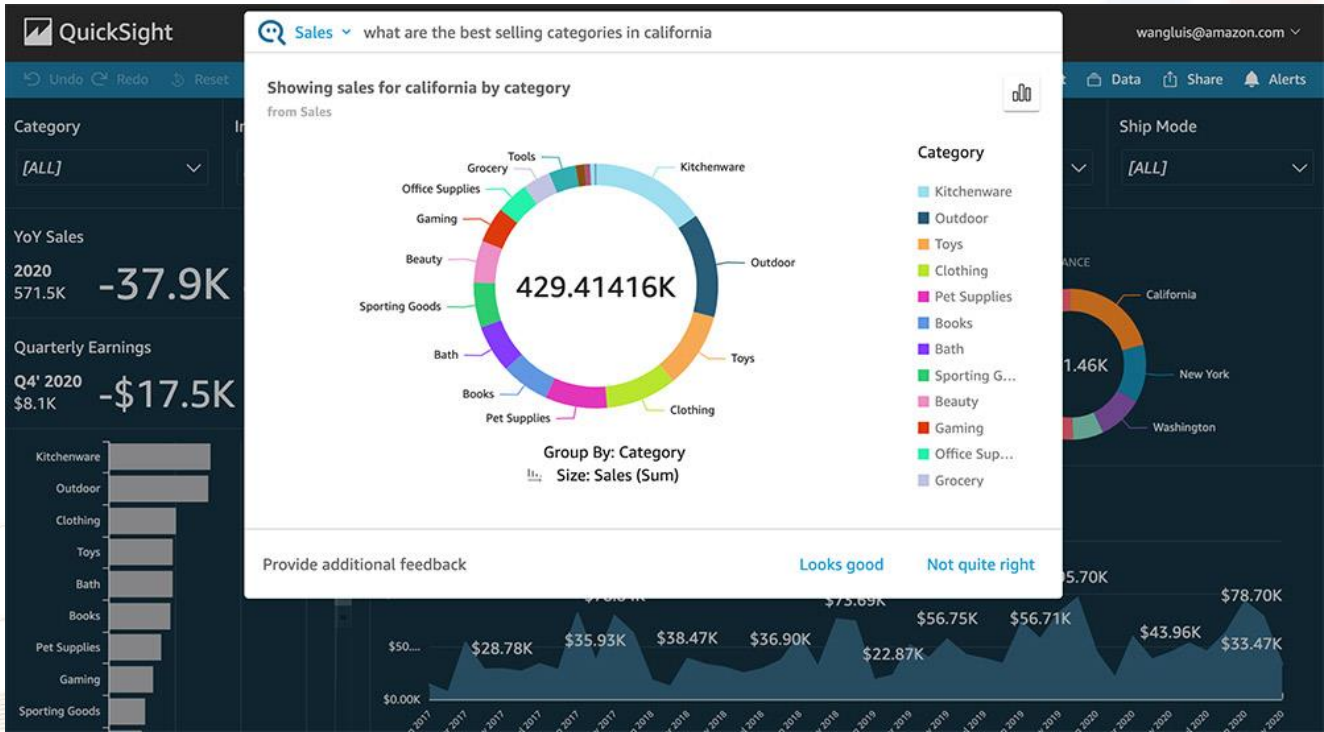
Преимущества

Масштабирование от десятков до десятков тысяч пользователей.

Благодаря бессерверной архитектуре Amazon QuickSight позволяет автоматически выполнять масштабирование до десятков тысяч без необходимости в установке и настройке собственных серверов, а также управлении ими. При этом благодаря оплате по количеству сеансов плата взимается только за доступ пользователей к информационным панелям или отчетам.

Встраивание информационных панелей бизнес-аналитики в приложение.

Благодаря QuickSight вы можете быстро встраивать интерактивные информационные панели в приложения, веб-сайты и порталы. QuickSight предлагает широкий выбор API-интерфейсов и SDK, которые позволяют легко настраивать внешний вид информационных панелей в соответствии с приложениями.



Amazon QuickSight

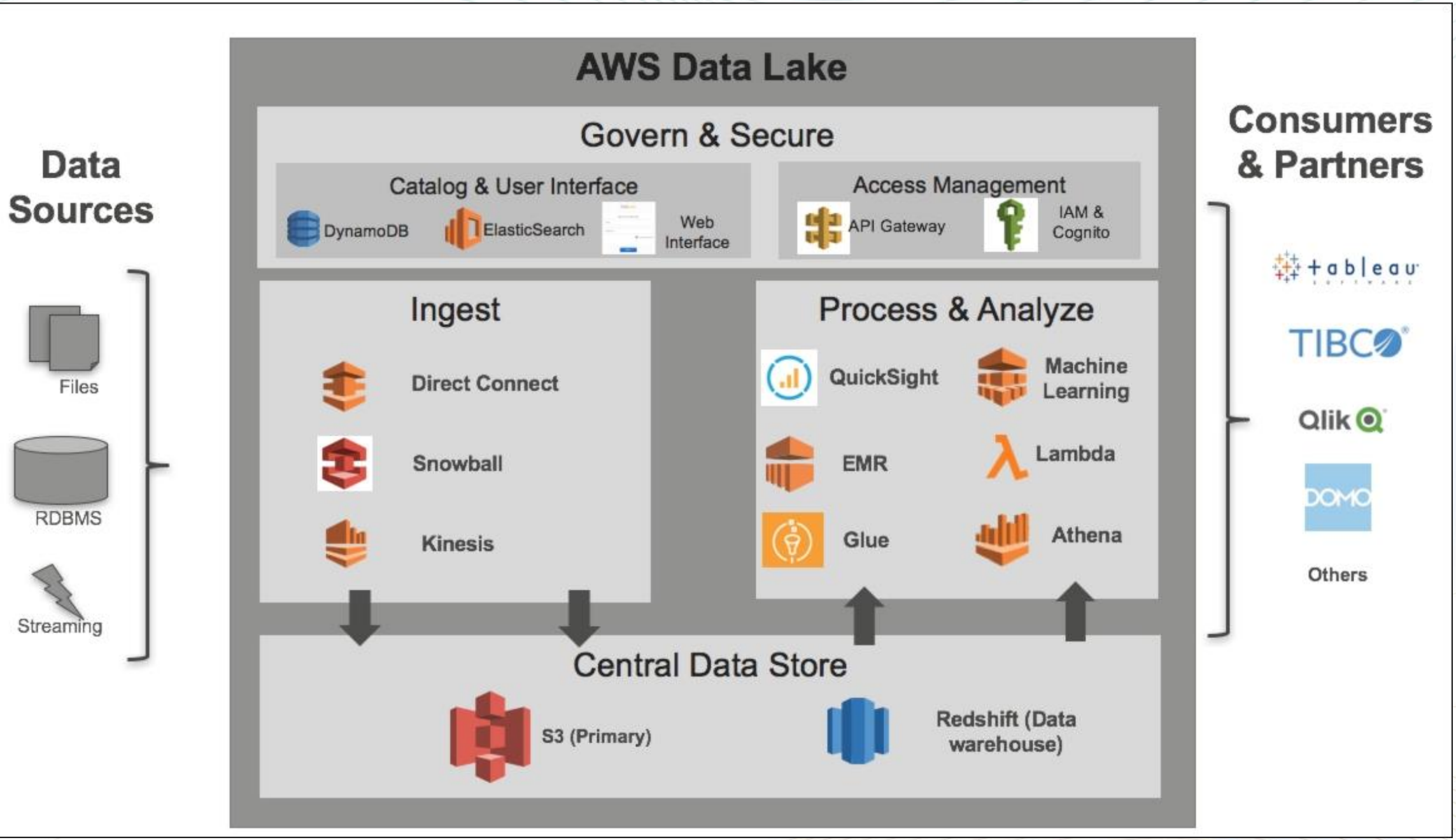
Преимущества

Углубленная аналитика на основе Machine Learning. QuickSight использует проверенные возможности машинного обучения (Machine Learning) AWS, упрощая процесс расширенной аналитики (например, анализ возможных вариантов, обнаружение аномалий, прогнозирование на основе машинного обучения, прогнозирование оттока клиентов и т. д.) для команд бизнес-аналитиков без опыта работы с данными. Вы можете использовать предварительно созданные модели QuickSight или применять собственные модели машинного обучения из Amazon SageMaker, которые интегрируются с QuickSight с помощью нескольких щелчков мыши. На основе машинного обучения QuickSight также автоматически генерирует краткие сведения об информационной панели на простом языке. Такие сведения включают в себя интерпретацию и описание ключевых данных, тем самым обеспечивая последовательность и единообразие информации для всех пользователей.

Задавайте вопросы о данных и получайте ответы. Благодаря QuickSight вы быстро получите ответы на деловые вопросы, заданные на естественном языке. Раньше, чтобы получить ответы на подобные вопросы, бизнес-пользователям приходилось ждать несколько недель, пока команды бизнес-аналитиков обновят модели данных и информационные панели. Но поскольку новая возможность QuickSight Q на основе машинного обучения позволяет создавать запросы на естественном языке, вы можете ввести вопрос в строке поиска на простом английском языке [например, what is the year-over-year sales trend? (Какова тенденция сбыта в годовом исчислении?)] и получить ответы за считанные секунды. Функция Q выделяет в вопросах пользователей бизнес-терминологию и цель и после извлечения соответствующих данных из источника выдает ответ в виде числа, диаграммы или таблицы.

Встроенная аналитика. Встраивайте в свои приложения возможность просматривать и создавать информационные панели без труда.

Аналитика на основе машинного обучения. Создавайте обзоры бизнес-метрик на простом языке или используйте машинное обучение для прогнозирования результатов, например для обнаружения аномалий или прогнозирования, без опыта в области работы с данными.


Data Sources
Files
RDBMS
Streaming
AWS Data Lake
Govern & Secure
Catalog & User Interface
DynamoDB
ElasticSearch
Web Interface
Access Management
API Gateway
IAM & Cognito
Ingest
Direct Connect
Snowball
Kinesis
Process & Analyze
QuickSight
Machine Learning
EMR
Lambda
Glue
Athena
Central Data Store
S3 (Primary)
Redshift (Data warehouse)
Consumers & Partners
tableau
TIBCO
Qlik
DOMO
Others

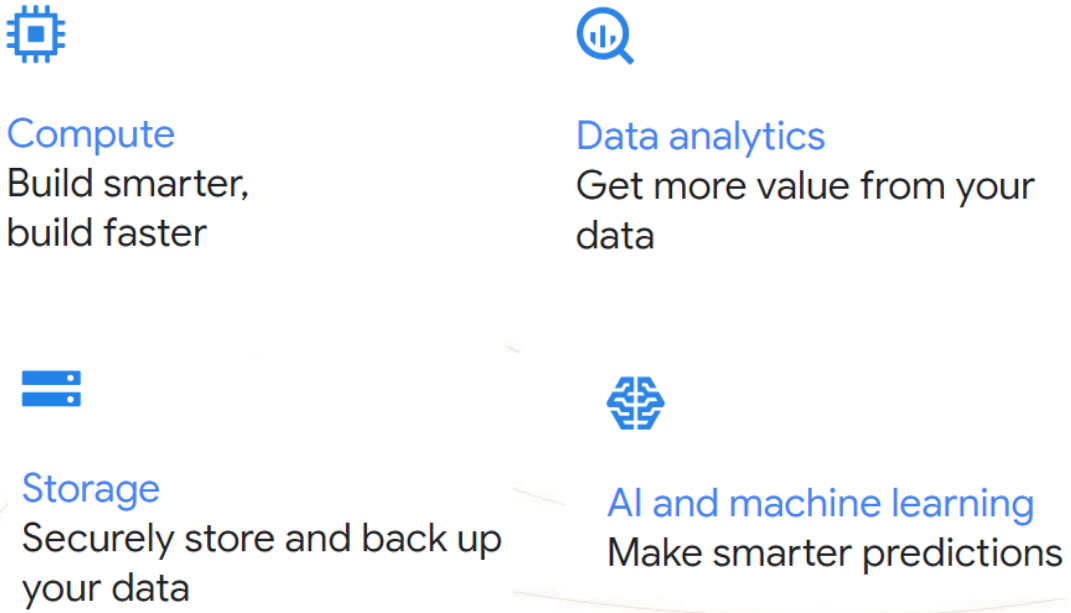
Google Cloud Platform

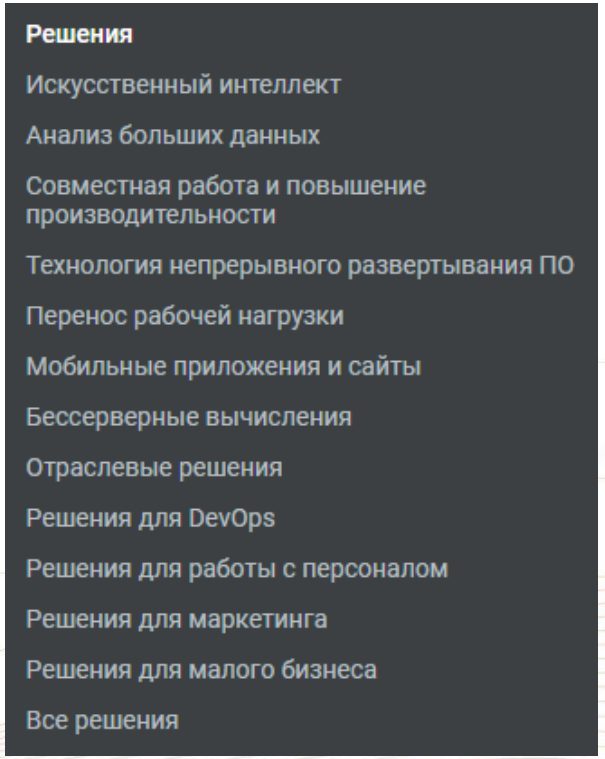
Google Cloud Platform

Google Cloud Platform

Облачная платформа Google , предлагаемая Google , представляет собой набор служб облачных вычислений, которые работают на той же инфраструктуре, которую Google использует для своих конечных пользователей, таких как Google Search и YouTube . Наряду с набором инструментов управления, она обеспечивает ряд модульных облачных сервисов , включая вычисление, хранение данных, анализ данных и машинное обучение .

Облачная платформа Google предоставляет инфраструктуру как сервис , платформу как сервис и бессерверные вычислительные среды.





Продукты для обработки больших данных

Эффективно собирайте, обрабатывайте и анализируйте данные с помощью продуктов Google Cloud для анализа данных.

Полностью управляемый, безсерверный подход
Легко откройте для себя бизнес-идеи с помощью полностью управляемых, проверенных, комплексных продуктов для анализа данных Google Cloud Platform.

BigQuery

Бессерверное, масштабируемое и экономичное многооблачное хранилище данных, предназначенное для обеспечения гибкости бизнеса.

BigQuery ML. BigQuery ML позволяет специалистам по обработке данных и аналитикам данных создавать и вводить в действие модели машинного обучения на основе структурированных или полуструктурированных данных планетарного масштаба непосредственно внутри BigQuery, используя простой SQL, - в кратчайшие сроки. Экспортируйте модели BigQuery ML для онлайн-прогнозирования в Cloud AI Platform или на свой собственный уровень обслуживания.

BigQuery GIS. BigQuery GIS уникальным образом сочетает в себе бессерверную архитектуру BigQuery с встроенной поддержкой геопространственного анализа, поэтому вы можете расширить свои рабочие процессы аналитики с помощью анализа местоположения. Упростите анализ, просматривайте пространственные данные по-новому и откройте для себя совершенно новые направления бизнеса с поддержкой произвольных точек, линий, многоугольников и многополигонов в распространенных форматах геопространственных данных.

BigQuery BI Engine. BigQuery BI Engine - это молниеносная служба анализа в памяти для BigQuery, которая позволяет пользователям анализировать большие и сложные наборы данных в интерактивном режиме со временем ответа на запрос менее секунды и высокой степенью параллелизма. BigQuery BI Engine легко интегрируется со знакомыми инструментами, такими как Data Studio, и поможет ускорить изучение и анализ данных для Looker , Sheets и др.

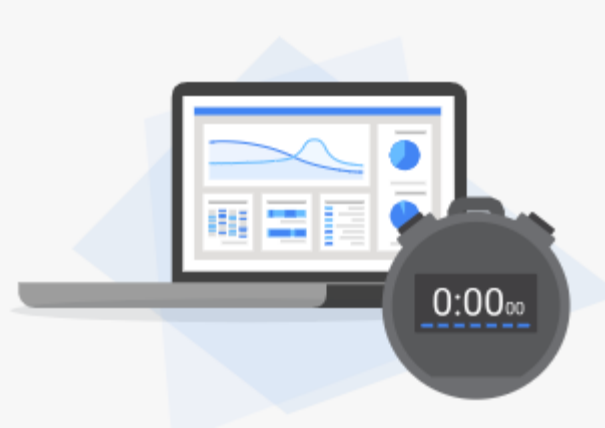
Подключенные таблицы.

Connected Sheets позволяет пользователям анализировать миллиарды строк живых данных BigQuery в Google Sheets, не требуя знания SQL. Пользователи могут применять знакомые инструменты, такие как сводные таблицы, диаграммы и формулы, чтобы легко извлекать информацию из больших данных.

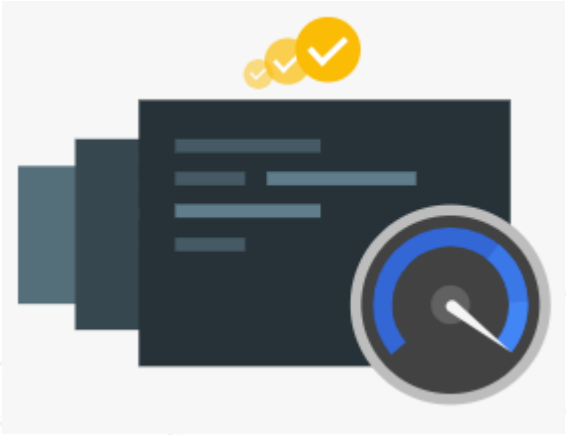
BigQuery ML

Сборка моделей с использованием SQL
BigQuery ML позволяет ученым и аналитикам данных создавать и внедрять модели ML на основе структурированных или полуструктурированных данных в масштабе планеты непосредственно в BigQuery с использованием простого SQL - за короткий промежуток времени.

BigQuery - это полностью управляемое, недорогое, бессерверное хранилище данных Google, которое масштабируется с учетом потребностей в хранении и вычислительной мощности.



Настройте хранилище данных за считанные секунды и немедленно начните запрашивать данные. BigQuery выполняет молниеносные запросы SQL от гигабайтов до петабайтов данных и позволяет легко объединять общедоступные или коммерческие наборы данных с данными.



Масштаб без проблем

Устраните головную боль планирования емкости хранилища данных и достигните бесконечности с гибким масштабированием емкости, которое не имеет ограничений. BigQuery решает задачи аналитики в реальном времени, используя безсерверную инфраструктуру Google, которая использует автоматическое масштабирование и высокопроизводительную потоковую передачу для загрузки данных. Управляемое столбчатое хранилище BigQuery, массовое параллельное выполнение и автоматическая оптимизация производительности позволяют всем пользователям быстро и одновременно анализировать данные независимо от количества пользователей или размера данных.

Cloud Dataflow

Пакетная и потоковая обработка данных

Cloud Dataflow - *это полностью управляемый сервис для преобразования и обогащения данных в потоковом (в реальном времени) и пакетном (историческом) режимах с равной надежностью и выразительностью.* Безсерверный подход Cloud Dataflow освобождает от оперативных задач, таких как планирование емкости, управление ресурсами и оптимизация производительности, при этом платя только за то, что вы используете. Кроме того , Cloud Dataflow работает не только с инструментами Google, но и сторонних инструментами, такими как Apache Spark и Apache Beam.

(Apache Beam - это унифицированная модель для определения как пакетных, так и потоковых конвейеров для параллельной обработки данных, а также набор специфических для языка SDK (Software Development Kit) для построения конвейеров и вычислительных процессов в back-end для их выполнения на серверах распределенной обработки, включая Apache Apex, Apache Flink, Apache Spark, и Google Cloud Dataflow.)

Cloud Dataproc

Управляемый **Apache Spark и Apache Hadoop**

Cloud Dataproc - *это быстрый, простой в использовании, полностью управляемый облачный сервис для запуска кластеров Apache Spark и Apache Hadoop более простым и экономичным способом.* Операции, которые раньше занимали часы или дни, вместо этого занимали секунды или минуты - и платите только за ресурсы, которые вы используете с посекундной тарификацией. Cloud Dataproc интегрируется со службами хранения, вычислений и мониторинга в продуктах Google Cloud, предоставляя вам мощную и полную платформу обработки данных.

Cloud Dataprep

Интеллектуальная подготовка данных

Cloud Dataprep от Trifacta - *это интеллектуальная служба данных для визуального исследования, очистки и подготовки структурированных и неструктурированных данных для анализа.* Поскольку Cloud Dataprep не имеет сервера и работает в любом масштабе, нет инфраструктуры для развертывания или управления. Следующее преобразование данных предлагается и прогнозируется с каждым новым вводом пользовательского интерфейса, поэтому не нужно писать код. А благодаря автоматической схеме, определению типов данных, возможным объединениям и обнаружению аномалий можно пропустить трудоемкое профилирование данных и сосредоточиться на анализе данных.

Google Data Studio

Google Data Studio *позволяет раскрыть всю мощь ваших данных с помощью интерактивных информационных панелей и привлекательных отчетов, которые помогают принимать более взвешенные бизнес-решения.* Это позволяет легко читать, делиться и настраивать информацию. Интеграция между BigQuery , Data Studio и Sheets и популярными инструментами BI и ETL обеспечивает гибкость при приеме и представлении данных. Data Studio также использует модель совместного использования Google Drive, позволяющую группам сотрудничать в режиме реального времени. А с помощью готовых шаблонов отчетов команды могут сосредоточиться на рассказывании убедительных историй, а не на обработке данных.

Cloud Bigtable

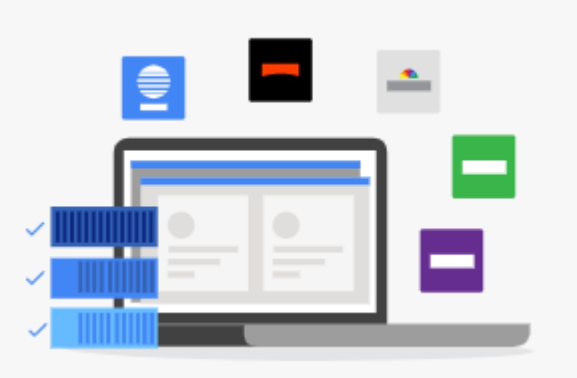
Полностью управляемый сервис базы данных NoSQL

Cloud Bigtable *предоставляет масштабируемую базу данных NoSQL, подходящую для рабочих нагрузок с низкой задержкой и высокой пропускной способностью.* Cloud Bigtable легко интегрируется с популярными инструментами для работы с большими данными, такими как Hadoop и Spark, и поддерживает HBase API с открытым исходным кодом. Cloud Bigtable - отличный выбор как для операционных, так и для аналитических приложений, включая IoT, пользовательская аналитика и анализ финансовых данных.

Cloud Datalab

Мощное исследование данных

Cloud Datalab - *это мощный интерактивный инструмент, созданный для исследования, анализа, преобразования, визуализации данных и построения моделей машинного обучения на Google Cloud Platform.* Это интерактивная записная книжка на основе Jupyter , интегрированная с BigQuery и Cloud Machine Learning Engine, чтобы обеспечить легкий доступ к ключевым службам обработки данных. А с TensorFlow или Cloud Machine Learning Engine можно легко превратить данные в развернутые модели машинного обучения, готовые к прогнозированию.



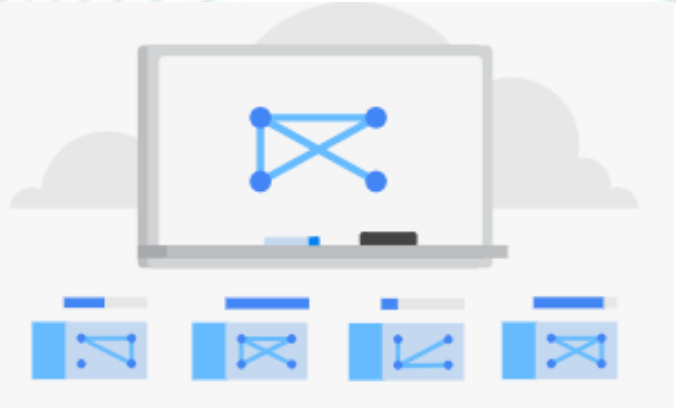
KUBERNETES ENGINE

Kubernetes Engine (GKE) - это управляемая, готовая к работе среда для развертывания контейнерных приложений.
Kubernetes Engine *обеспечивает быструю разработку и итерацию приложений, упрощая развертывание, обновление и управление вашими приложениями и сервисами.* Просто опишите вычислительные ресурсы, память и ресурсы хранения, необходимые для контейнеров приложений, и Kubernetes Engine обеспечит и автоматически управляет базовыми облачными ресурсами. Поддержка аппаратных ускорителей упрощает запуск машинного обучения, графических процессоров общего назначения, высокопроизводительных вычислений и других рабочих нагрузок, в которых используются специализированные аппаратные ускорители.

Cloud Composer

Оркестровка рабочего процесса

Cloud Composer - *это полностью управляемая служба управления рабочими процессами, которая позволяет создавать, планировать и отслеживать конвейеры, охватывающие облака и локальные центры обработки данных.* Cloud Composer, созданный на основе популярного проекта Apache Airflow с открытым исходным кодом и использующий язык программирования Python и прост в использовании. Кроме того, благодаря комплексной интеграции рабочих нагрузок GCP вы можете организовать полный конвейер со всеми продуктами Google Cloud для обработки больших данных.



Облако AutoML

Обучить пользовательские модели машинного обучения
Cloud AutoML - это набор продуктов для машинного обучения, который позволяет разработчикам с ограниченным опытом в области машинного обучения обучать высококачественным моделям, соответствующим их бизнес-потребностям, используя передовые технологии обучения Google и технологию поиска нейронной архитектуры.

Продукты AutoML

Создайте свое собственное видение, естественный язык и модели перевода с минимальными необходимыми навыками машинного обучения

AutoML Vision

Получите информацию из изображений в облаке или на краю.

AutoML Translation

Динамически обнаруживать и переводить между языками.

Таблицы AutoML

Автоматическое создание и развертывание современных моделей машинного обучения на структурированных данных.

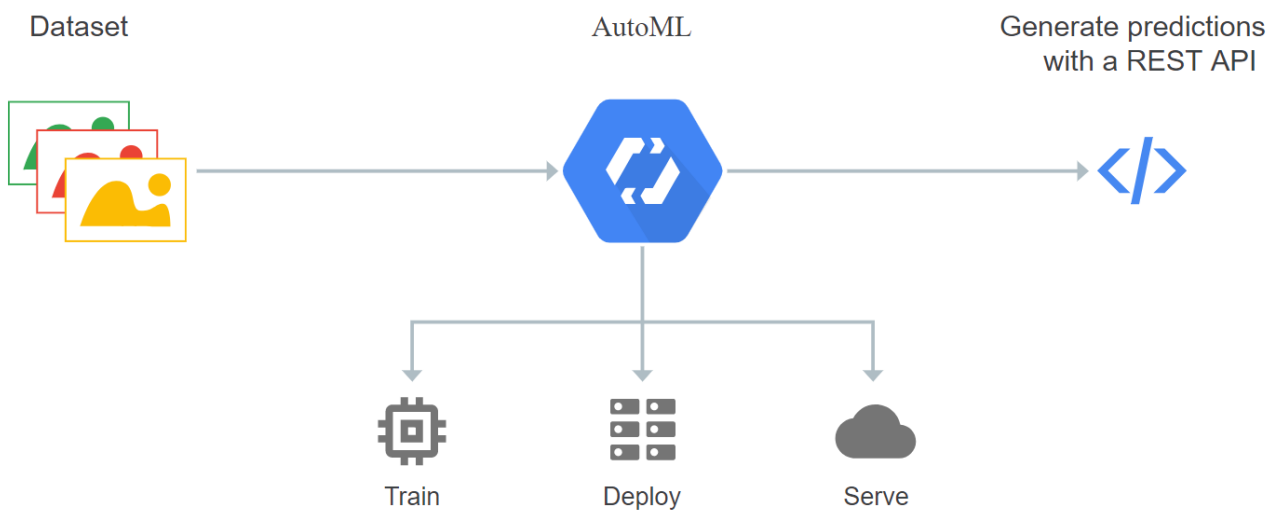
AutoML Video Intelligence

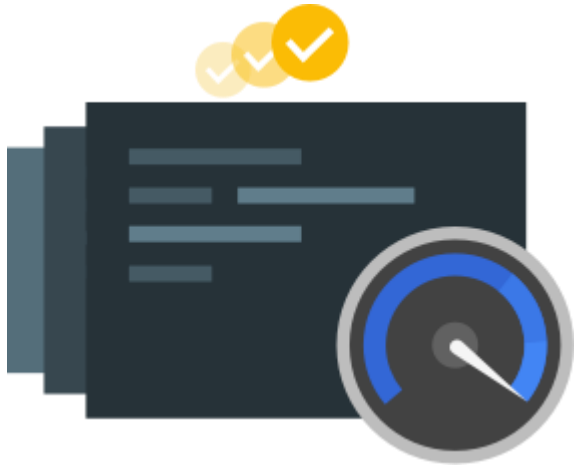
Обеспечьте мощное обнаружение контента и увлекательное видео.

AutoML Natural Language

Раскройте структуру и значение текста с помощью машинного обучения.

Как работает AutoML

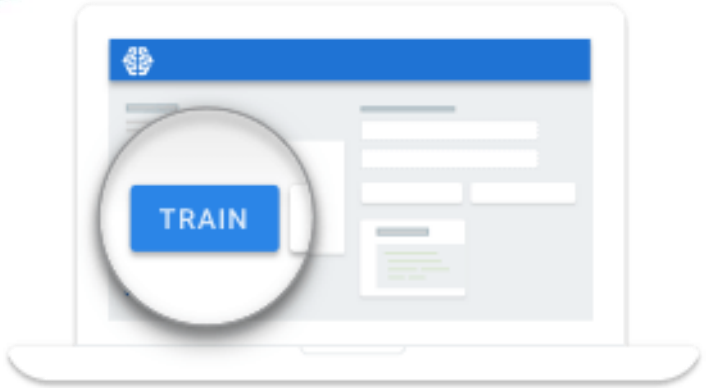




Cloud TPUs

Оборудование, предназначенное для производительности

Облачные TPU - это семейство аппаратных ускорителей, которые Google специально разработала и оптимизировала для ускорения и увеличения рабочих нагрузок ML для обучения и применения нейросетей, запрограммированных с помощью TensorFlow. Облачные TPU предназначены для обеспечения наилучшего соотношения производительность/стоимость для выполнения приложений TensorFlow и позволяют инженерам и исследователям быстрее выполнять машинное обучение.



Движок облачного машинного обучения

Легко построить и еще проще масштабировать

Cloud Machine Learning Engine позволяет вам создавать сложные, масштабные модели машинного обучения, которые охватывают широкий спектр сценариев, от построения сложных моделей регрессии до классификации изображений. Он переносим, полностью управляем и интегрирован с другими продуктами платформы данных Google Cloud, такими как Cloud Storage , Cloud Dataflow и Cloud Datalab, так что вы можете легко обучать свои модели.

Облачные продукты ИИ

Быстрые, масштабные и простые в использовании продукты и услуги AI.

- **AI Hub** , наш размещенный репозиторий подключаемых ИИ компонентов AI, поощряет эксперименты и сотрудничество внутри вашей организации.
- **Строительные блоки искусственного интеллекта** позволяют разработчикам легко добавлять визуальные, языковые, разговорные и структурированные данные в свои приложения.
- **AI Platform** , наша среда разработки данных на основе кода, позволяет разработчикам ML и специалистам по данным быстро переходить от проектов к идеям и внедрять их.

Инновационные AI-решения на надежной платформе
Облачный ИИ предоставляет современные сервисы машинного обучения, с предварительно обученными моделями и сервисом для создания собственных пользовательских моделей. Служба ML на основе нейронных сетей имеет лучшую производительность обучения и более высокую точность по сравнению с другими системами глубокого обучения. Основные приложения Google используют облачное машинное обучение, в том числе севисы «Фотографии» (поиск изображений), «Переводчик», «Входящие» (Smart Reply) и приложение Google (голосовой поиск). Платформа теперь доступна как облачный сервис, чтобы обеспечить беспрецедентный масштаб и скорость для бизнес-приложений.

Строительные блоки искусственного интеллекта позволяют легко добавлять визуальные, языковые, разговорные и структурированные данные в ваши приложения

SIGHT

видение
Получите информацию из изображений в облаке или на краю.

видео
Обеспечьте мощное обнаружение контента и увлекательное видео.

LANGUAGE

Перевод
Динамически обнаруживать и переводить между языками.

Естественный язык
Раскройте структуру и значение текста с помощью машинного обучения.

РАЗГОВОР

Dialogflow
Создавайте виртуальные агенты и другие возможности общения.

Облачный текст в речь API
Преобразуйте текст в речь, похожую на человеческую, используя голоса WaveNet.

Cloud Speech-to-Text API
Преобразуйте речь в текст автоматически с выдающейся точностью.

СТРУКТУРИРОВАННЫЕ ДАННЫЕ

Таблицы AutoML
Автоматическое создание и развертывание современных моделей машинного обучения на структурированных данных.

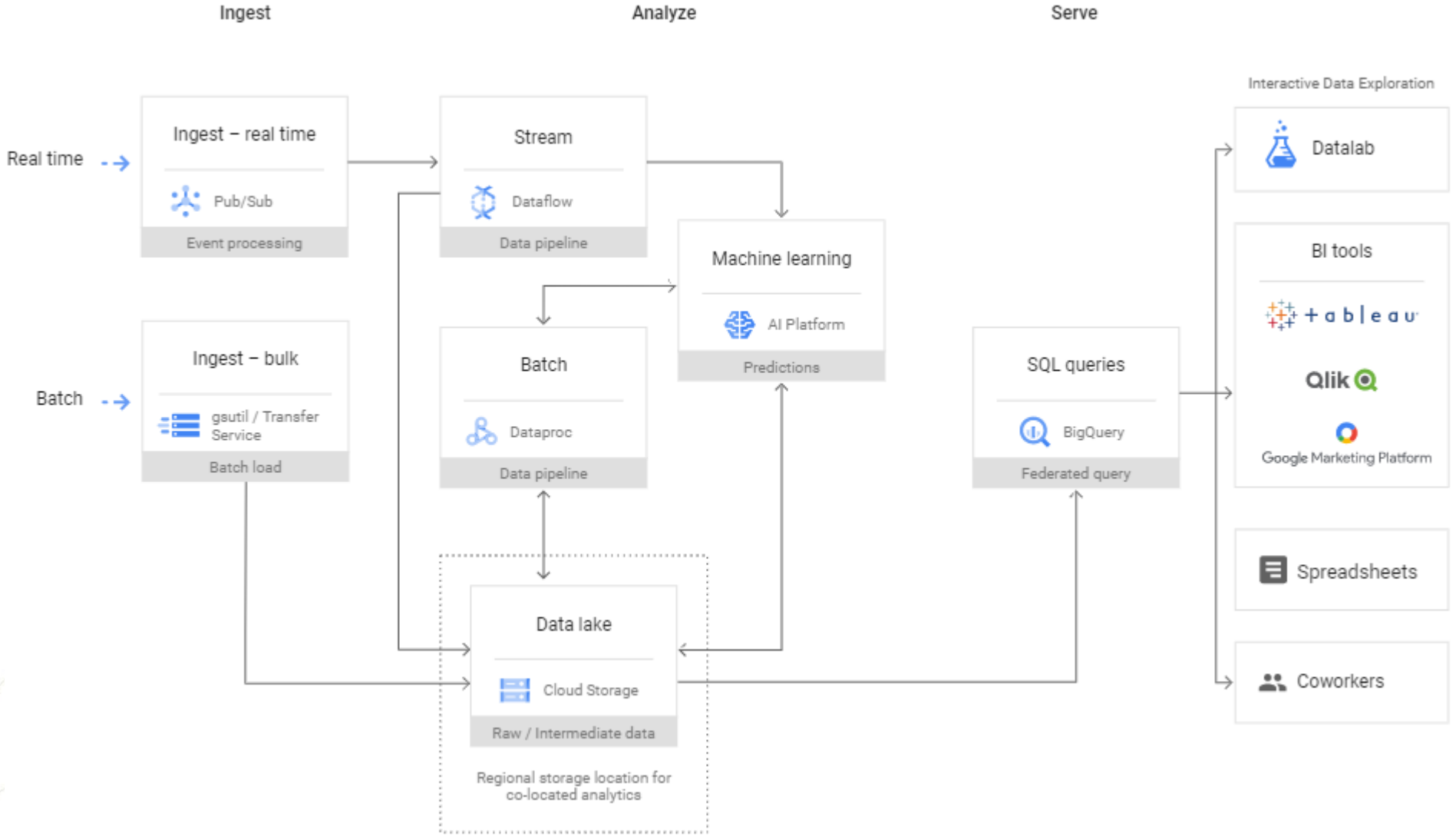
Рекомендации AI
Предоставляйте высоко персонализированные рекомендации по продукту в масштабе.

API Cloud Inference
Быстро запустите крупномасштабные корреляции для типизированных наборов данных временных рядов.

ПРИМЕР ИСПОЛЬЗОВАНИЯ

Интегрированный репозиторий для аналитики и машинного обучения

Высочайший уровень доступности и производительности в пределах одного региона идеально подходит для рабочих нагрузок вычислений, аналитики и машинного обучения в конкретном регионе. Облачное хранилище также полностью согласовано, что дает вам уверенность и точность при выполнении аналитических задач.



Платформа AI

Полностью управляемая комплексная платформа для анализа данных и машинного обучения.

Сквозной жизненный цикл машинного обучения

Подготовка.
Подготовьте и сохраните свои наборы данных с помощью BigQuery и Cloud Storage , а затем используйте встроенную службу маркировки данных, чтобы пометить свои обучающие данные для классификации, обнаружения объектов, извлечения сущностей и других целей для изображений, видео, табличных и текстовых данных.

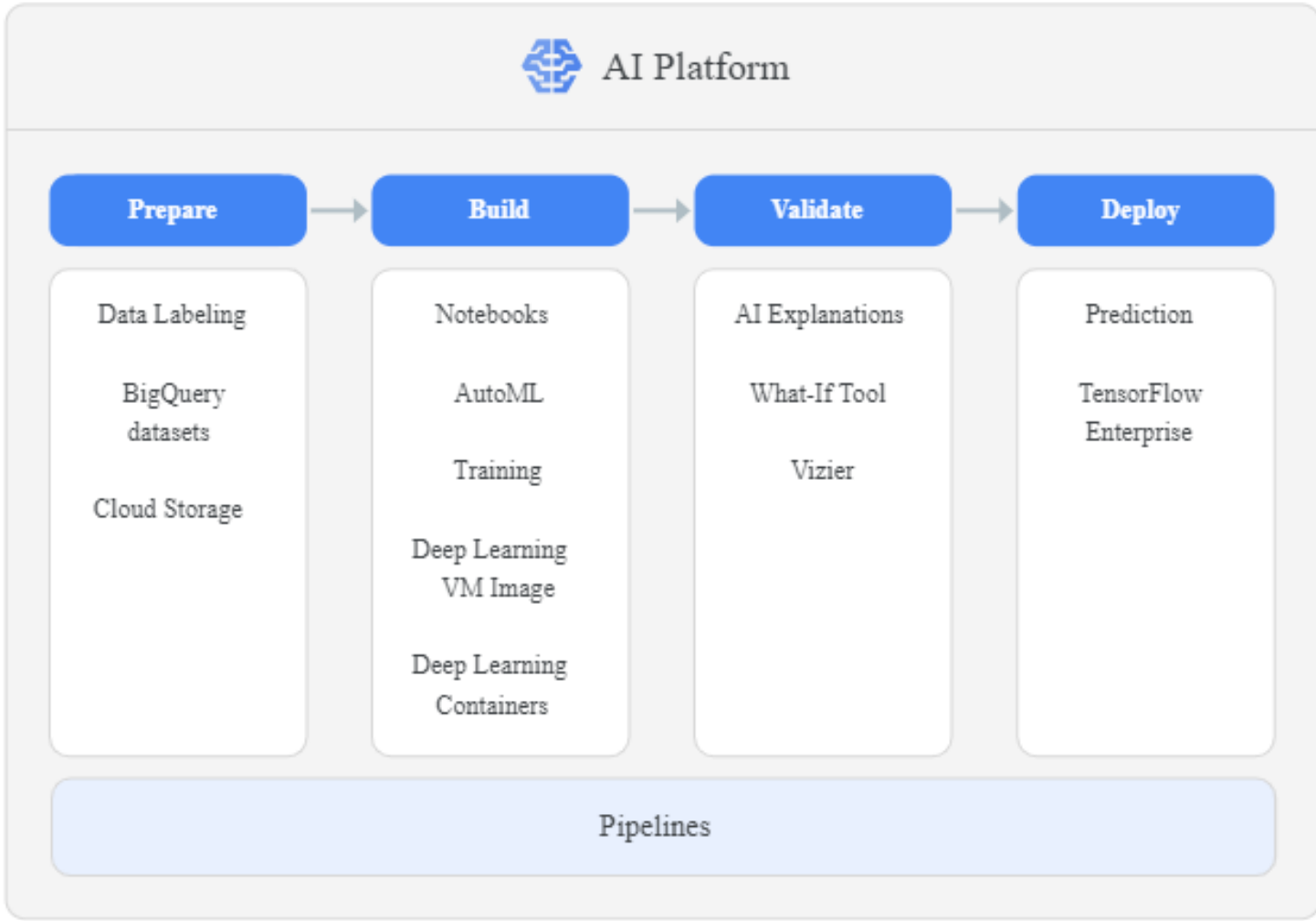
Разработка.
Создавайте лучшие в своем классе модели машинного обучения без написания кода с помощью простого в использовании пользовательского интерфейса AutoML или использования собственного кода, написанного в Notebooks , управляемой службе Jupyter Notebook. Используйте новейшие платформы глубокого обучения с открытым исходным кодом для Deep Learning VM Image или Deep Learning Containers . Затем обучите свои модели с помощью нашей полностью управляемой службы обучения .

Подтвердить.
Подтвердите свою модель с помощью AI Explanations и What-If Tool , которые помогут вам понять результаты вашей модели, проверить поведение модели, выявить предвзятость и найти способы улучшить вашу модель и данные обучения. Сделайте шаг вперед в настройке модели с помощью Vizier , службы оптимизации черного ящика, чтобы настроить гиперпараметры и оптимизировать производительность вашей модели.

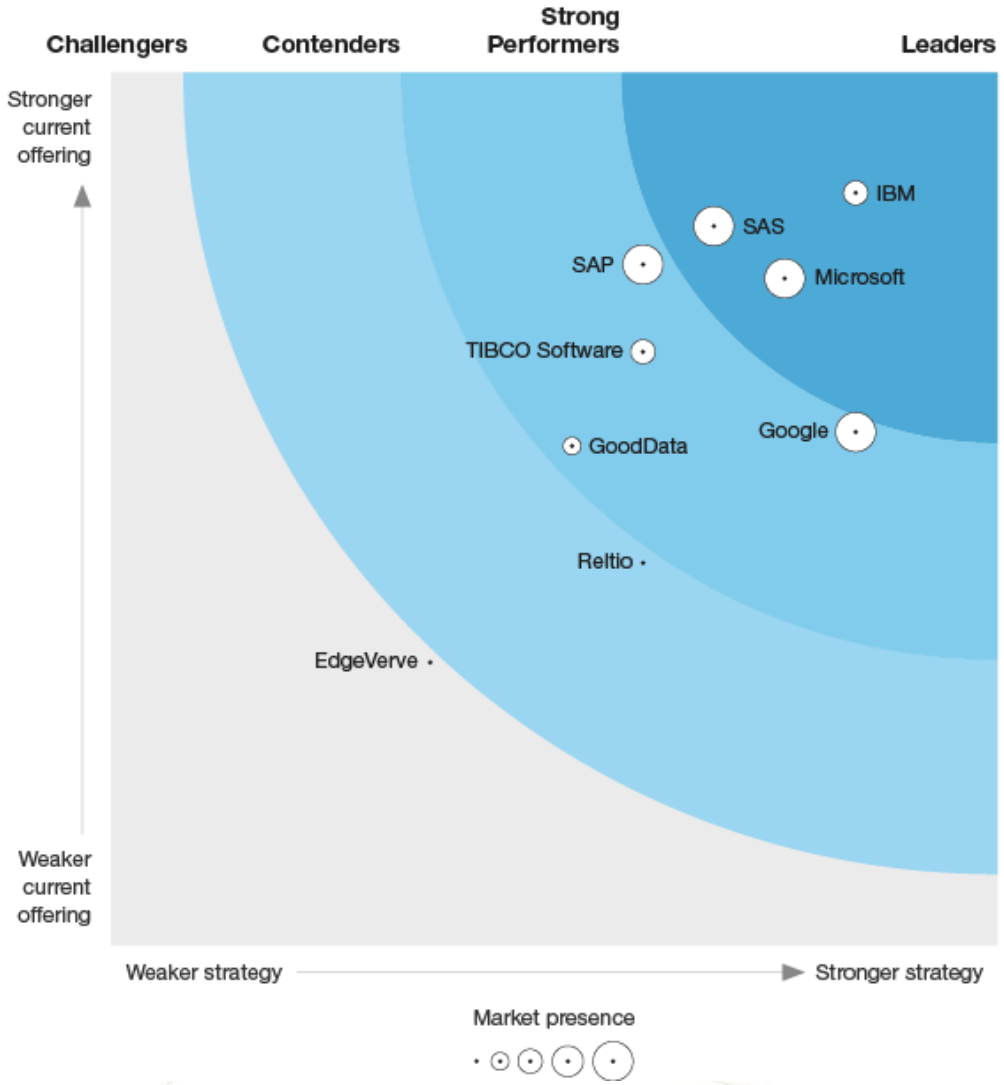
Платформа AI

Развернуть. Разверните свои модели в масштабе, чтобы получать прогнозы в облаке с помощью Prediction , в котором размещается ваша модель для онлайн-запросов и запросов на пакетное прогнозирование. Вы также можете использовать AutoML Vision Edge для развертывания ваших моделей на границе и запуска действий в реальном времени на основе локальных данных. TensorFlow Enterprise предлагает поддержку корпоративного уровня для экземпляров TensorFlow.

MLOps. Управляйте своими моделями, экспериментами и сквозными рабочими процессами с помощью конвейеров , применяя передовые методы MLOps с надежными, повторяемыми конвейерами. Непрерывная оценка помогает вам контролировать производительность ваших моделей и обеспечивает постоянную обратную связь с течением времени.



Forrester Wave: платформы корпоративной аналитики, первый квартал 2019 года



Рынок платформ корпоративной аналитики, которые сочетают в себе инструменты управления данными, аналитики и разработки аналитических приложений, вырос на двузначные цифры в 2019 году. Клиенты потратили **50** миллиардов долларов на широкий рынок бизнес-аналитики и управления информацией в 2019 году.

Google остается приверженцем бессерверных инноваций, искусственного интеллекта и открытого исходного кода. У Google есть беспрецедентная стратегия предоставления бессерверных данных и аналитических услуг, таких как **AutoML; BigQueryML; и API "строительных блоков" ИИ**, которые работают в любом масштабе как часть своей ведущей облачной платформы.
Google продолжает внедрять гибридные облачные инновации с новыми сервисами, такими как Cloud Composer (построенный на Apache Airflow) и KubeFlow.
Google - единственный поставщик в этой Forrester Wave, который обрабатывает статичные и потоковые данные с той же парадигмой программирования. Облачная платформа Google - хороший выбор для компаний, имеющих опыт в облачной разработке и данных.