Как строить озера данных?

Обзор концепции IBM создания озер данных (data lake) с поддержкой технологий AI/ML/DL.

Введение

Объемы данных продолжают расти быстрыми темпами. Но рост данных – это только одна часть истории больших данных. Новый искусственный интеллект (artificial intelligence, AI), Интернет вещей (Internet of Things, IoT), мобильные и другие технологии генерируют не только большие объемы данных, но и широкий спектр данных из многообразных источников. Организации могут фиксировать настроения клиентов, выраженные в социальных сетях; потоковую передачу данных с датчиков; записи врачей о состоянии пациента; данные о погоде; аудиозаписи взаимодействия с кол-центром, переписку по электронной почте, данные о потоках кликов (лайков) и многое другое.

Использование машинного обучения (ML) и глубокого обучения (DL) в средах с большими данными позволяет выявлять исторические закономерности и создавать модели искусственного интеллекта (ИИ), которые могут помочь предприятиям улучшить качество обслуживания клиентов, добавлять услуги и предложения, определять новые потоки доходов или направления бизнеса (LOBs, lines of business), выявлять случаи мошенничества и оптимизировать бизнес или производственные операции.

К сожалению, традиционные хранилища данных (data warehouses) и витрины данных не дают возможность организациям извлекать выгоду из большого обьема и разнообразия данных, доступных сегодня. Созданные для структурированных данных хранилища данных не могут хранить, запрашивать или анализировать полуструктурированные и неструктурированные данные. Это происходит из-за того, что они в основном ориентированы для подготовки заранее определенных отчетов, а также из-за того, что хранилищам данных не хватает гибкости для удовлетворения ad hoc (специальных, целевых) запросов в реальном времени. Опираясь только на хранилища данных (OLAP-хранилища), большинство данных останутся неиспользованными и непроанализированными.

Согласно результатам исследования, проведенного IDC по заказу IBM, время, когда озера данных строились исключительно на однородной архитектуре с использованием стандартных серверов, в основном, на базе hadoop-кластеров, уходит в прошлое. Это обусловлено тем, что все больше компаний развивают инициативы в области искусственного интеллекта (ИИ) на базе технологий глубокого и машинного обучения (DL/ML), а традиционные решения озер данных не ориентированы на их поддержание. В итоге, подготовка данных, тренировка моделей и получение первых результатов анализа могут занимать недели и месяцы, что становится уже совершенно неприемлемым в нашем быстро меняющемся мире с ориентацией на онлайновую аналитику[1)].

Необходимо отметить, что адаптация традиционной архитектуры озер данных под технологии ИИ гораздо более сложная, чем может показаться, которая требует:

- *интеграции новых узлов/компонент – программных и аппаратных – с поддержкой различных ускорителей для нейронных сетей, а также одновременно и интеграции компонент инфраструктуры между собой;*
- *введения новых фреймворков/инструментария, упрощающих(щего) работу с данными и не требующих(щего) от сотрудников специальных глубоких знаний;*
- *низколатентного высокопроизводительного интерконнекта;*
- *поддержки интерфейсов с нереляционными СУБД;*
- *системы управления данными и их безопасностью с распределением прав доступа между различными группами пользователей*

и др.

Многие организации в поисках новых способов использования больших данных и преодоления ограничений традиционных OLAP-хранилищ данных, начали включать озера данных в свою стратегию управления данными[2)].

Озера данных

По своей сути озеро данных – это центральное место, в котором хранятся все данные независимо от их источника или формата. Обычно озеро данных создается с использованием Hadoop или другой масштабируемой архитектуры, которая может хранить экономически значимые объемы данных. Одно из самых распространенных недоразумений – путаница в понятиях озера данных и хранилища данных (OLAP-хранилища). В табл. 1 дано сравнение ключевых атрибутов хранилища данных в отличие от озера данных. Поскольку в озерах данных могут храниться все данные, они являются мощной альтернативой решениям, которые возникают при интеграции данных в традиционном хранилище данных, особенно когда организации обращаются к мобильным и облачным приложениям и интернету вещей (IoT).

Табл. 1. Сравнение особенностей data warehouse и data lake.

Attribute	Data warehouse	Data lake
Schema	Schema-on-write.	Schema-on-read.
Масштабирование	Масштабируется от умеренных до больших объемов при умеренной стоимости.	Масштабируется до огромных объемов по низким ценам.
Методы доступа	Доступ через стандартизированные SQL- и BI-инструменты (Business Intelligence).	Доступ через SQL-подобные системы и программы, созданные разработчиками. Также поддерживаются инструменты анализа больших данных.
Рабочая нагрузка	Поддержка пакетной обработки и тысяч одновременных пользователей, выполняющих интерактивную аналитику.	Поддерживает пакетную и потоковую обработку и обладает улучшенными возможностями по сравнению с хранилищами данных для поддержки запросов больших данных от пользователей
Данные	Очищенные	Сырые и очищенные
Сложность данных	Сложная интеграция	Сложность обработки
Эффективность затрат	Эффективно использует CPU/IO, но имеет высокие затраты на хранение и обработку.	Эффективно использует возможности хранения и обработки при низких затратах.

Озера данных (data lakes) – это гибридные решения для управления данными следующего поколения, которые позволяют решать проблемы с большими данными (в частности, анализировать неструктурированные и полуструктурированные данные, которые не очень хорошо поддерживаются традиционными хранилищами данных) и продвигать новые уровни аналитики в реальном времени. Их высокомасштабируемая среда может поддерживать чрезвычайно большие объемы данных и принимать их в своем родном/нативном формате из широкого спектра источников данных. Озера данных помогают разрушать барьеры, связанные с анализом многочисленных "островков" данных, позволяя организациям получать 360-градусный обзор информации и проводить кросс-департаментную аналитику в организации.

Преимущества озер данных

При правильной разработке и реализации данные озера предлагают пять важных преимуществ:

- *упрощенную подготовку данных*. Сохраняя данные в их первоначальном формате, озеро данных может помочь сократить количество времени, затрачиваемое на подготовку данных;
- *упрощенный доступ к данным*. С хорошо построенным озером данных можно расширить доступ к большему количеству пользователей, включая не только специалистов по данным (data scientists), но также пользователей бизнес-направлений и разработчиков приложений. Определяемый пользователем доступ позволяет им работать с данными из нескольких источников в организации, локально или в облаке;
- *повышенную гибкость для пользователей данных*. Озеро данных, оснащенное надлежащими инструментами, может выполнять специальные запросы и анализ в реальном времени, при этом исключая время и затраты, связанные с ИТ-помощью от ИТ-департамента;
- *снижение затрат*. В озерах данных используется аппаратное оборудование, позволяющее экономически эффективно масштабировать их без чрезмерных капитальных затрат – можно использовать озеро данных в качестве хранилища для старых данных, которые в противном случае занимали бы емкость в более дорогих хранилищах. Предоставляя пользователям прямой

1) Rethinking Your Infrastructure for Enterprise AI, Sponsored by: IBM, Peter Rutten, June 2018

2) IBM, Build a better data lake, 16014716-usen-00_16014716USEN.pdf, april, 2018.

доступ к данным, озера данных также позволяют избегать затрат на ИТ-поддержку. Кроме того, реализация надлежащих возможностей управления данными для озера данных позволяет избежать затрат, связанных с исправлением проблем с качеством данных;

– *улучшенное принятие решений*. Анализ данных, взятых из большего количества источников, увеличивает глубину понимания и повышает точность результатов, а функции управления помогают обеспечить актуальность и достоверность данных. Аналитика в реальном времени и технологии ИИ позволяют использовать новые возможности по мере их развития.

Озера данных могут использоваться в многочисленных существующих приложениях в различных отраслях промышленности. Вот несколько примеров:

– *розничная торговля*:

  • определение интересов покупателя в реальном времени при интернет-покупках и предоставление рекомендаций, как умного личного помощника;

  • определение характера корзины покупателя и предложение товаров, которые могли бы дополнить корзину покупателя;

  • предсказание или проактивное выявление мошеннических действий как внутри организации, так и за ее пределами;

– *банковское дело:*

  • предсказание успеха или провала скидок;

  • определение «следующего продукта для покупки» и продвижение этого продукта для клиентов;

  • идентификация клиентов, которые могут уменьшить свою банковскую деятельность, и использование упреждающей маркетинговой активности;

– *гостиничное дело и путешествия:*

  • отслеживание и прогнозирование предпочтений клиентов для продвижения проактивных продаж;

  • улучшение качества обслуживания клиентов и повышение лояльности к бренду за счет индивидуализации и персонализации;

  • проведение и анализ в реальном времени ценовой политики.

Возможные риски

Непродуманно разработанные озера данных, которые не имеют возможностей корпоративного уровня, сделают не намного больше, чем просто хранилище для большой неорганизованной агломерации данных. Эти «болота» собирают данные, не предоставляя простой и безопасный способ для поиска, доступа и анализа необходимой информации.

Ряд факторов может привести к неудаче проекта создания озера данных, среди них:

– *отсутствие бизнес-кейса*. Без четкого формулирования и понимания преимуществ, которые принесет бизнесу озеро данных, проект может не получить одобрения или финансирования;

– *плохая интеграция*. Озеро данных может дополнять или в некоторых случаях заменять хранилище данных (data warehouse). Но если отсутствует план для интегрированного управления данными, достижение полноценного результата, который может обеспечить озеро данных, невозможно;

– *неправильный выбор технологий*. Выбор неправильной платформы или инструментов может значительно усложнить реализацию и увеличить стоимость;

– *неадекватное управление и безопасность*. Стратегии управления и безопасности корпоративного уровня имеют решающее значение для защиты конфиденциальной информации, соблюдения требований и предоставления пользователям возможности в полной мере использовать данные;

– *отсутствие долгосрочного видения*. Озеро данных требует долгосрочного видения в интеграции с планированием для обеспечения непрерывного роста данных.

Требования к проекту создания озера данных

Хотя озеро данных может уменьшить объем начальной работы по подготовке данных, оно не устраняет его. При вводе в озеро данных необходимо добавлять надежные метаданные, которые описывают источник данных. Также необходимо профилировать и проверять данные, чтобы подтвердить их структуру, содержание и качество.

Стратегии озер данных должны обеспечивать:

– масштабируемость озера данных;

– возможность увеличения со временем объема, скорости и разнообразия данных;

– поддержание растущего числа и разнообразия пользователей в масштабах всего предприятия.

Важно четко ставить цели, которые необходимо достичь, анализируя все новые доступные данные. Это поможет определить приоритетные варианты использования озера данных, что, в свою очередь, поможет сузить типы и источники данных, на которых следует сосредоточиться, и позволит точно определить необходимые инструменты и стратегию. Например, если главным приоритетом является мониторинг износа критически важного производственного оборудования, то можно решить проанализировать данные IoT «в облаке» – в облачном озере данных.

Для достижения бизнес-целей важно оценить наличие доступных ресурсов. При выявлении их недостатка, необходимо планировать способы их устранения и учитывать время и затраты на:

– новые технологии, оборудование, программное обеспечение и услуги;

– финансирование квалифицированных специалистов, не состоящих в настоящее время в штате;

– внутреннюю ИТ-поддержку;

– обучение пользователей.

Можно создать озеро данных, которое обрабатывает все данные, но, если данные поступают «как есть» (без обработки, выделения сущностей, повышения качества данных), озеро данных может оказаться менее полезным, чем хранилище данных для анализа традиционных структурированных данных, которое все еще может служить источником извлеченных, проверенных данных для типичной организационной, исторической и финансовой отчетности.

Озеро данных может дополнять это хранилище данных, позволяя хранить, запрашивать и анализировать дополнительные типы данных в более экономичной среде. Кроме того, озеро данных может предоставить более дешевое хранилище для старых данных, необходимости хранить которых нет или которые не должны храниться в своем хранилище данных.

Самодостаточность для всех типов пользователей при работе с озером данных – одно из ключевых требований при его создании, обеспечивающее его эффективность. Предоставляя простой и быстрый доступ к данным, а также подходящие инструменты для запроса данных, можно поддерживать ad hoc анализ (анализ по запросу) и стимулировать инновации – и все это при одновременном снижении потребности в ИТ-помощи. В качестве примера приведем 3 категории пользователей, для которых озеро данных может быть полезно.

Коммерческие (Line-of-business, LOB) пользователи, работающие в сфере маркетинга, могут воспользоваться озером данных для разработки целевых маркетинговых кампаний, в то время как финансисты могут определить способы повышения внутренней эффективности. Пользователи LOB отвечают за создание сводок и аналитических отчетов. Им нужны простые инструменты для доступа и анализа данных, относящихся к их проектам.

Специалисты по обработке данных (Data scientists) могут использовать аналитику для выявления новых тенденций в бизнесе или использовать прогнозную аналитику, чтобы помочь отделам продаж определить следующее наилучшее действие для клиентов. В целом, исследователи данных – это те, кто строят модели и алгоритмы, создают визуализации данных и сотрудничают с бизнес-командами для создания новых идей из больших наборов данных.

Разработчики приложений (App developers) могут использовать озеро данных в качестве тестовой среды для новых мобильных приложений. Разработчикам приложений необходимо выполнять ad hoc и запросы в режиме реального времени, а также интегрировать несколько источников данных по всей организации. Им важен контроль доступа к данным с минимальным использованием ИТ-персонала.

Технологии для создания озера данных

Hadoop стал предпочтительной платформой для создания озер данных. Эта хорошо масштабируемая структура позволяет обрабатывать очень большие наборы данных на сотнях или тысячах вычислительных узлов, причем все они работают параллельно. Как технология с открытым исходным кодом Hadoop создается сообществом и поддерживается сообществом. Используя стандартное оборудование, Hadoop может помочь снизить расходы.

Решения Hadoop корпоративного уровня, предлагаемые Hortonworks и другими поставщиками, могут учитывать общие сложности и ограничения управления, а

также добавлять ключевые функции безопасности. С правильным решением можно ускорить прием (ingestion) данных и использовать инструменты управления и функции безопасности, которые помогут соблюдать строгие политики и правила, используя при этом обработку данных в реальном времени и потоковую аналитику.

Если необходимо интегрировать озеро данных с существующими средами хранилища данных (data warehouse), потребуется выбор правильных инструментов и стратегия. Традиционные корпоративные сервисные шины (enterprise service bus, ESB) и инструменты ETL предназначены для работы с пакетными процессами, а не с процессами реального времени, поэтому они не могут справиться с требованиями низкой задержки для озера данных. Традиционные методы могут привести к проблемам с контекстом, связыванием и визуализацией данных – все это необходимо учитывать при анализе больших данных.

Традиционные инструменты также требуют от персонала понимания инструментов, источника данных и целевого хранилища данных, но это требует глубоких знаний и опыта, что может быть дорогостоящим. Правильные инструменты облегчают взаимодействие с существующими средами и сокращают время и усилия, необходимые для интеграции. Использование возможностей автоматизации может помочь контролировать расходы и сосредоточить усилия специалистов по обработке данных на решении других задач.

Как управлять данными и безопасностью?

Озеро данных является общей платформой, к которой могут обращаться многие пользователи в различных ролях. Чтобы защитить данные и обеспечить соответствие нормативным требованиям, необходимы достаточные возможности управления, безопасности и аудита. Сильная стратегия управления имеет решающее значение для соблюдения правил и обеспечения того, чтобы пользователи могли легко находить, понимать и доверять данным. Использование инструментов управления метаданными в рамках этой стратегии поможет в полной мере использовать свое озеро данных. По оценкам IDC (*IDC for Seagate, "Data Age 2025: The Evolution of Data to Life-Critical," April 2017*), к 2025 году почти 90% всех данных, создаваемых в глобальной сфере данных, будут требовать некоторого уровня безопасности, но защита будет обеспечена только для менее половины данных (42%).

Правильные инструменты управления метаданными позволят создать индекс активов данных, добавить метаданные, чтобы классифицировать контент и отслеживать происхождение данных. Это даст возможность пользователям легче находить то, что им нужно, и приобрести уверенность в том, что полученные ими сведения являются точными.

Если необходимо хранить данные кредитных карт, информацию о состоянии здоровья пациентов, корпоративную финансовую информацию, интеллектуальную собственность или другие конфиденциальные данные в озере данных, то следует обеспечить защиту данных от потери или кражи, а также реализовать возможности управления, которые помогут соблюдать регуляторные правила.

Важно правильно разграничивать права доступа к данным. Например, можно разрешить членам маркетинговой команды запрашивать и анализировать широкий спектр данных о клиентах, ограничивая при этом их доступ к финансовой информации компании. В дополнение к внедрению инструментов для защиты от несанкционированного доступа к определенной информации необходимо планировать обучение пользователей озера данных тому, как соблюдать политики и регулятивные нормы.

Инфраструктура ИИ IBM для построения озер данных

Чтобы упростить использование компаниями всех преимуществ от использования ИИ, IBM представила концепцию под названием «Инфраструктура ИИ» ("AI Infrastructure"), которая состоит из единой платформы для конвейеров данных, служб/сервисов и ИИ. По сути, это комплексная платформа для серверов, хранилищ и программного обеспечения. На рис. 1 представлены основные программные и аппаратные ее компоненты, а также структурная схема их взаимодействия для построения озер данных с поддержкой AI в соответствии с концепцией IBM.

Ключевой особенностью инфраструктуры ИИ IBM является глубокая интегрированность и оптимизация корпоративного класса всех компонент для поддержки технологий AI/ML/DL, за счет чего сокращается время подготовки данных (при одновременном упрощении работы с ними), тренировки моделей и получение конечных результатов анализа.

Аппаратные компоненты

IBM Power System AC922

IBM – одна из первых компаний в отрасли (декабрь 2017 г.) стала продвигать на рынке интегрированные бандлы (ускоритель нейронных сетей; обзор типов ускорителей – см. отдельную публ. в данном SN, *прим. ред.*) для машинного обучения на базе процессоров Power и GPU от NVIDIA, интегировав их в архитектуру инфраструктуры ИИ IBM. В настоящее время предлагается 2 типа ускорителей: на базе IBM Power System AC922 серверов (POWER9 CPU; от 2 до 6 NVIDIA Tesla V100 GPUs; ОП – до 2 ТВ) и на базе IBM Power System S822LC для HPC-нагрузок.

GPU Tesla V100 с 640 тензорными ядрами является первым в мире графическим процессором, преодолевшим барьер в 100 TFLOPS DL-производительности. Это новое поколение NVIDIA NVLink GPU соединяет несколько графических процессоров V100 со скоростью до 300 Гбит/с для создания самых мощных в мире вычислительных серверов. Благодаря новым GPU, модели ИИ, которые потребляли недели вычислительных ресурсов в предыдущих системах, теперь могут быть обучены за несколько дней.

В основе этой платформы лежит современное озеро данных с улучшенным хранилищем, Hadoop и Spark корпоративного уровня и расширенными возможностями управления данными. Этот фундамент поддерживает платформы данных, которые охватывают несколько схем (RDBMS, NoSQL, графы) и несколько архитектур, поддерживающих вычисления (CPU, GPU,

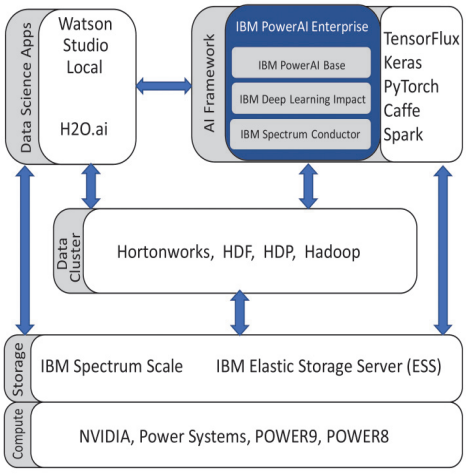

Рис. 1. Ключевые компоненты "Инфраструктуры ИИ" IBM и их взаимосвязи.

in memory обработка), а также высокую производительность системы. Вместе озеро данных и платформы данных обеспечивают гибкую ИТ-среду с динамическим хранением, вводом-выводом и памятью, которые предоставляют сервисы/услуги для анализа в реальном времени, для таких применений, как: CRM, IoT, обнаружение мошенничества и многие другие. Эти строительные блоки становятся основой для углубленного обучения, позволяющего интегрировать существующие приложения с ИИ и разрабатывать новые приложения на основе ИИ.

Платформа данных Hortonworks

HDP (Hortonworks Data Platform) – это дистрибутив Hadoop и Spark, который позволяет предприятиям безопасно хранить, обрабатывать и анализировать большие объемы данных в состоянии покоя. Он включает в себя HDFS, которая может масштабироваться до тысяч узлов.

Хранилище для HDFS можно развернуть на локальных дисках или на IBM Elastic Storage Server (IBM ESS) через IBM Spectrum Scale™ HDFS Transparency connector. HDP включает в себя Apache Spark – платформу распределенной обработки, включающей MLlib, которая является библиотекой Spark ML.

HDP 3.0 с Hadoop 3.1 предоставляют новые возможности, которые включают более быструю разработку приложений за счет контейнеризации, помехоустойчивого кодирования (erasure coding), объединения/федерацию namenode и улучшенную поддержку приложений DL за счет управления ресурсами графического процессора в качестве ресурса.

Основные преимущества использования графических процессоров в узлах кластера HDP:

- отсутствие необходимости настраивать отдельные кластеры;
- обнаружение и настройка GPU: администратор может настроить ресурсы и архитектуру графического процессора на каждом узле;
- возможность изоляции и мониторинга GPU. Если несколько процессов используют один графический процессор, процесс должен дождаться завершения текущего, прежде чем он сможет запуститься, что может вызвать задержки. Другой проблемой могут быть фреймворки, такие как TensorFlow,

которые активно пытаются использовать ресурсы GPU. Если назначить два приложения TensorFlow для одного GPU, оба они не смогут работать, потому что GPU не хватает памяти.

На производительность приложения не должно влиять отсутствие ресурсов графического процессора. Степень детализации для графических процессоров зависит от устройства для каждого графического процессора, что достигается с использованием контрольных групп (cgroups) или Docker для обеспечения изоляции;

– простое использование GPU (как самостоятельного ресурса такого же, как CPU и ОП) приложениями в рамках YARN без использования меток узлов.

HDF (Hortonworks DataFlow) обеспечивает потоковую аналитику в реальном времени для данных в движении, которую дополняет HDP. Это интегрированное решение, в состав которого входят Apache Nifi и MiNifi, Kafka, Storm, Streaming Analytics Manager и Schema Registry для обеспечения платформы анализа потоковых данных в реальном времени. HDF управляется установкой Ambari и координируется Zookeeper.

Решение IBM и Hortonworks обеспечивает инфраструктуру Hadoop на основе озера данных с улучшенными возможностями исследования, обнаружения, тестирования и расширенного запроса данных. Оно также предлагает масштабируемость, безопасность и управление с возможностью объединения/федерации как данных в состоянии покоя, так и данных в движении по всей организации. Пользователи могут легко делать запросы как к реляционным базам данных, так и к Hadoop, развернутым локально или в облаке, получая преимущества от доступа к данным в режиме самообслуживания и возможности выполнять специальные (ad hoc) запросы и запросы в режиме реального времени. HDP является предпочтительным озером данных для серверов IBM Power Systems и хорошо интегрируется с IBM Watson Studio Local.

IBM Spectrum Scale (ранее – IBM General Parallel File System, IBM GPFS™) – это кластерная файловая система, которая обеспечивает одновременный доступ к одной файловой системе или набору файловых систем с нескольких узлов. Узлы могут быть подключены через SAN-сеть, IP-сеть, в смешанном варианте (SAN+IP) или в конфигурации кластера без разделения ресурсов. Эти настройки обеспечивают высокопроизводительный доступ к этому общему набору данных для поддержки решения с горизонтальным масштабированием или для обеспечения платформы высокой доступности. IBM Spectrum Scale – это параллельная файловая система в основе IBM ESS.

IBM Spectrum Scale позволяет объединить варианты виртуализации, аналитики и использования файлов и объектов в едином решении для горизонтального хранения и обеспечивает единое пространство имен для всех данных, что дает единую точку управления. Благодаря поддержке широкого спектра протоколов обмена данными, таких как файловые системы POSIX, сетевая файловая система (NFS), блок сообщений сервера (SMB, Server Message Block), объектное хранилище (например, Swift и S3) и блочное хранилище (например, iSCSI и HDFS) клиенты могут выполнять свои рабочие аналитические работы на месте без необходимости дублировать наборы данных. Помимо этого, обеспечивается эффективная консолидация различных источников данных в одном глобальном пространстве имен.

Используя политики хранения, прозрачные для пользователей, данные могут быть сжатыми или многоуровневыми, что сокращает расходы. Данные также могут быть привязаны к высокопроизводительным носителям, включая кэш-память сервера, на основе тепловой карты данных для снижения задержки и повышения производительности.

Основные преимущества использования IBM Spectrum Scale с HDP:

– высокая масштабируемость с параллельной архитектурой файловой системы. При параллельной архитектуре ни один узел метаданных не может стать узким местом. Каждый узел в кластере может обслуживать как данные, так и метаданные, что позволяет одной файловой системе IBM Spectrum Scale хранить миллиарды файлов. Эта архитектура позволяет клиентам беспрепятственно расширять свою среду HDP по мере роста данных;

– глобальное пространство имен, которое может охватывать несколько кластеров Hadoop и географических областей. Используя глобальное пространство имен IBM Spectrum Scale, клиенты могут создавать активные и удаленные копии данных и обеспечивать глобальное сотрудничество в реальном времени. Это пространство имен позволяет глобальным организациям формировать озера данных по всему миру и размещать свои распределенные данные в одном пространстве имен. IBM Spectrum Scale также позволяет нескольким кластерам Hadoop получать доступ к одной файловой системе, сохраняя при этом всю необходимую семантику изоляции данных. Функция прозрачного облачного многоуровневого хранения IBM Spectrum Scale может архивировать данные в S3/SWIFT-совместимую систему хранения облачных объектов, такую как IBM Cloud™ Object Storage или Amazon S3, с помощью мощных политик управления жизненным циклом информации (ILM) IBM Spectrum Scale;

– IBM Spectrum Scale обеспечивает наиболее полную поддержку протоколов доступа к данным с использованием NFS, SMB, Object, POSIX и HDFS API. Эта функция устраняет необходимость поддерживать отдельные копии одних и тех же данных для традиционных приложений и для аналитики;

– возможность (за счет того, что это программно-определяемое хранилище) развернуть IBM Spectrum Scale как ПО непосредственно на обычных серверах с большим объемом хранилища, работающих со стеком HDP, или развернуть его как часть предварительно подготовленной системы с использованием IBM ESS. Клиенты могут использовать программные опции, чтобы начать с малого, но при этом использовать преимущества корпоративного хранилища. С помощью IBM ESS клиенты могут контролировать разрастание кластера и наращивать объем хранилища независимо от вычислительной инфраструктуры. IBM ESS использует помехоустойчивое кодирование (erasure coding), чтобы исключить необходимость трехсторонней репликации для защиты данных.

IBM ESS – это SDS-реализация, которая сочетает в себе IBM Spectrum Scale, работающую на серверах на базе процессора POWER8®. Основным преимуществом IBM ESS является его способность снижать требования к емкости. IBM ESS требуется только 30% дополнительной емкости, чтобы обеспечить преимущества защиты данных, аналогичные тройной репликации HDFS.

IBM ESS с HDP позволяет перемещать данные между различными системами для анализа данных. Это движение происходит быстрее и легче, что устраняет необходимость копировать данные из файловой системы на основе POSIX в HDFS благодаря единому пространству имен, предлагаемому IBM Spectrum Scale.

Решение IBM ESS поддерживает огромный объем данных, что всегда является требованием для реального озера данных. Одна ESS может обеспечить пропускную способность до 40 Гбит/с, а несколько ESS могут масштабироваться до эксабайт хранилища для поддержки масштабного расширения бизнеса. Этот рост данных может управляться и распределяться практически по любому виду хранения, дисковым технологиям, облаку и ленте.

Ядром IBM ESS является IBM Spectrum Scale, который управляет всеми задачами, связанными с дисками, данными и файловой системой, в полностью поддерживаемом POSIX-режиме. К файловым системам может обращаться любая операционная система на основе UNIX и Linux. IBM Spectrum Scale также поддерживает протокол SMB, который обеспечивает увеличение числа различных серверов, которые могут получать доступ к данным внутри пространства имен (файловой системы) без необходимости использования специальной клиентской программы.

Программные компоненты

IBM для инфраструктуры ИИ предлагает 9 решений, в полной мере обеспечивающих весь набор инструментария для подготовки и анализа данных с использованием технологий AI/ML/DL для всех возможных вариантов совместной работы: IBM Watson® Machine Learning Accelerator (ранее – IBM PowerAI Enterprise); IBM Watson Studio Local; IBM Power Systems™; IBM Spectrum™ Scale; IBM Data Science Experience (IBM DSX); IBM Elastic Storage™ Server (IBM ESS); Hortonworks[3] Data Platform (HDP); Hortonworks[3] DataFlow (HDF); H2O Driverless AI.

IBM PowerAI base

IBM PowerAI base представляет собой загружаемый дистрибутив ПО корпоратив-

3) Компания Hortonworks объединилась с Cloudera в январе 2019 г. Новая компания называется Cloudera. Ссылки на Hortonworks как на коммерческое предприятие в этой публикации теперь относятся к объединенной компании. Названия продуктов, начиная с Hortonworks, продолжают продаваться и продаваться под их оригинальными названиями.

ного класса, который включает в себя фреймворки с открытым исходным: ML-фрейворки (такие как scikit-learn) и DL-фреймворки (такие как TensorFlow, PyTorch и Caffe) и доступный бесплатно. IBM PowerAI, интегрированный с IBM Watson Studio Local, является одной из доступных сред для ноутбуков. IBM PowerAI также включен в IBM Watson Machine Learning Accelerator для крупномасштабных развертываний.

IBM PowerAI включает в себя следующие функции для повышения производительности и сокращения времени обучения:

– распределенную библиотеку глубокого обучения (DDL, distributed deep learning), которая масштабирует DL до сотен серверов Power Systems и использует преимущества серверов на базе процессоров IBM POWER9™ и графических процессоров NVIDIA. DDL ограничен четырьмя узлами в IBM PowerAI, поэтому для больших конфигураций используется IBM Watson Machine Learning Accelerator;
– библиотеку Snap ML, которая поддерживает распределенные GPU-акселераторы для алгоритмов ML, таких как scikit-learn;
– поддержку больших моделей (LMS, Large model support), которая позволяет увеличивать размер моделей (то есть, например, включать изображения с более высоким разрешением) за счет интеграции с фреймворками DL. LMS использует быстрые интерфейсы NVLinks на процессорах POWER9 для обмена частями модели с основной памятью и из нее в различные моменты во время обучения. Производительность TensorFlow LMS улучшена в IBM PowerAI V1.5.4, а LMS для PyTorch предоставляется в качестве предварительной версии.

IBM Watson Machine Learning Accelerator (ранее – IBM PowerAI Enterprise)

IBM Watson Machine Learning Accelerator – это готовое к применению решение для ИИ, которое включает в себя IBM PowerAI и IBM Spectrum Conductor™. Оно предоставляет законченную DL-платформу для групп исследователей данных и может создавать и управлять кластерами Spark, которые называются Spark Instance Groups (SIG) и совместно используют базовые ресурсы серверов IBM Watson Machine Learning Accelerator, включая память, процессоры и графические процессоры. Совместное использование ресурсов контролируется политикой планирования, которую можно динамически настраивать для ускорения завершения задания.

IBM Watson Machine Learning Accelerator включает DDL и LMS для ускорения получения результатов. Инструменты разработки модели включают визуализацию обучения в режиме реального времени, мониторинг точности во время выполнения, а также поиск и оптимизацию гиперпараметров. DDL может масштабироваться до тысяч узлов.

IBM Watson Machine Learning Accelerator сочетает в себе популярные DL-структуры с открытым исходным кодом, эффективные инструменты разработки AI и ускорители на базе серверов Power Systems.

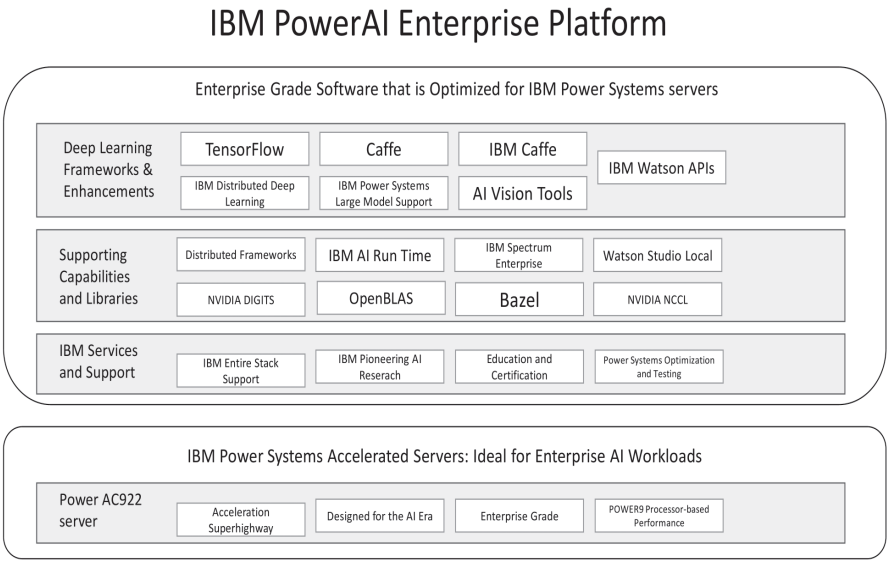

Рис. 2. Поддерживаемые фреймворки и библиотеки IBM Watson Machine Learning Accelerator.

IBM Watson Machine Learning Accelerator – это полноценная среда для обработки данных как сервиса (DSaaS), позволяющая внедрять новые прикладные приложения ИИ в производство.

В комплект ПО IBM Watson Machine Learning Accelerator входят библиотеки Anaconda, IBM PowerAI base и IBM Spectrum Conductor Deep Learning Impact (рис. 2).

Anaconda – это дистрибутив с открытым исходным кодом языков программирования Python и R для специалистов, занимающихся обработкой данных (data science) и ML, который упрощает управление и развертывание пакетов. IBM Spectrum Conductor обеспечивает многопользовательскую работу, представляя бизнес-направления (LOB) в качестве потребителей, каждый из которых имеет свой собственный набор SIG, которые могут совместно использовать пулы ресурсов и управляться динамическими политиками. IBM Spectrum Conductor Deep Learning Impact включает автоматическую настройку гиперпараметров, которая может создавать десятки обучающих заданий с разными гиперпараметрами, а затем отслеживать задания и удалять те, которые кажутся менее перспективными.

IBM Watson Studio Local

IBM Watson Studio Local (ранее – IBM Data Science Experience (IBM DSX) Local) – это локальная версия решения продукта IBM Watson Studio и корпоративное решение для ученых и инженеров данных, которая может работать локально и в облаке, и не только в IBM Power, но и на других процессорных архитектурах. IBM DSX Local предоставляет такие инструменты, как RStudio, Spark, Jupyter Notebooks и Zeppelin Notebooks, для исследователей данных, инженеров данных, разработчиков приложений и экспертов по предметам (SME, subject matter experts) для совместной и простой работы с данными, для построения и обучения моделей в масштабе. Это дает гибкость для построения моделей, в которых находятся данные, и развертывания их в любом месте в гибридной среде. DSX Local работает не только в IBM Power, но и на других процессорных архитектурах.

Пользователи имеют возможность связывать свои инструменты, такие как Jupyter Notebooks с ядрами R, Python и Scala, со средами исполнения Spark. Среды могут быть настроены и сохранены. К активам данных относятся файлы и подключения к существующим источникам данных, таким как рас-

		Deep Learning			ML and DL	Machine Learning
		Power AI Base	Power AI Enterprise	AI Vision	Watson Studio Local	H2O Driverless AI
Offering	Description	Deep Learning	Deep Learning for the Enterprise	Deep Learning with Video tools	Notebook oriented development environment for ML and DL	Automated Machine learning
Applications	Text & Numeric	Yes	Yes	No	Yes	Yes
	Images	Yes	Yes	Yes	Yes	No
	Video	-	Optional add-on	Yes		No
	Primary Persona	Data Scientist	Data Scientist	Line of Business	Data Scientist	Data Scientist
	Second persona	IT	IT	IT	IT	Line of Business
	User Skill Level	High	Medium to high	Low	Medium to high	Low to Medium
	Strengths	Rapid deployment, high performance, scale	Medium to high enterprise grade, High performance, rapid Deployment	Rapid deployment, simple GUI high performance	Notebook based development environment, strong collaboration, model management	Simplified deployment, intuitive user interface, automatic pipelines, "explainability" for models, end to end automation
Platform	Distributed DL (DDL)	1-4 nodes	1-thousands of nodes	-	-	-
	Large Model Support	Yes	Yes	-	-	-
	Server(s)	S822LC or AC922	S822LC or AC922	S822LC or AC922	S822LC or AC922, LC922	S822LC, AC922, LC921/922
IBM Products	Spectrum MPI (DDL)	Limited to 4 nodes	Included	-		Optional add-on
	Spectrum Conductor DLI	Optional add-on	Included	-	Optional Add On	Optional add-on
	IBM Watson Studio Local	Optional add-on	Optional add-on	No		Optional add-on
Cloud	IBM Cloud Public	Yes	No	Trial only	Watson Studio	-
	IBM Cloud Private	Yes	Yes	Yes	Yes	-

Рис. 3. Ключевые отличия AI-решений для серверов Power Systems.

пределенные файловые системы Hadoop (HDFS) и базы данных. Пользователи могут преобразовывать и формировать данные, используя компонент Data Refinery.

Есть несколько ролей, которые поддерживаются для соавторов проекта, включая администраторов проекта, редакторов и зрителей. Раздел сообщества с примерами записных книжек предоставлен, чтобы помочь ускорить время разработки. Консоль администрирования предназначена для управления и мониторинга оборудования, пользователей и служб.

IBM Watson Studio Local работает на кластере серверов Kubernetes, который выполняет роли главного (уровня управления) хранилища и вычислений. Архитектура требует минимум четырех узлов. Роль развертывания необязательна при установке, но необходима при развертывании моделей. IBM Watson Studio Local может управлять своими пользователями для аутентификации или интеграции с существующим корпоративным сервером каталогов LDAP.

H2O Driverless AI

H20 Driverless AI – это решение H20.ai для автоматического машинного обучения (ML), которое упрощает многие data science ML-задачи. Данные могут быть получены из облака, Hadoop или настольных систем. Визуализация данных выполняется автоматически, показывая форму данных, выбросы и пропущенные значения. Автоматизированная система ML использует рецепты лучших практик и базовую мощь машины для итерации по тысячам возможных моделей, включая разработку функций и настройку параметров. Автоматический конвейер скоринга включает в себя преобразования функций и модели для быстрого развертывания в производство.

H2O Driver AI поддерживается на серверах IBM Power, использует графические процессоры для ускоренного ML и может устанавливаться и управляться как приложение WLM-A.

IBM PowerAI Vision

IBM PowerAI Vision облегчает задачу маркировки объектов на изображениях и видео через интерфейс «наведи и щелкни», который позволяет не-экспертам участвовать в проекте по исследованию данных.

Модель может быть обучена с начальным набором помеченных данных и использована для маркировки новых изображений. Новые маркированные изображения можно просматривать, исправлять и использовать для обновления модели с целью повышения ее точности. Процесс повторяется до достижения необходимой точности модели.

Ключевые отличия среди решений AI представлены на рис. 3. И IBM PowerAI, и IBM Watson Machine Learning Accelerator поддерживают DDL, но IBM Watson Machine Learning Accelerator может масштабироваться до тысяч узлов. IBM Watson Studio Local – основанная на ноутбуках среда – хороший выбор для команд по анализу данных, которым комфортно работать с Jupyter Notebooks и R. H2O Driverless AI и AI Vision позволяют людям с более низкими навыками работать над проектами AI.

Заключение

Дальнейшее развитие озер данных будет осуществляться в нескольких направлениях. Во-первых, будет развиваться и совершенствоваться поддержка множества платформ данных (включая и поддержку неструктурированных данных традиционными реляционными СУБД) – например, NoSQL – для IoT, графовые – для обнаружения мошенничества. Во-вторых, будет происходить дальнейшее наполнение "Инфраструктуры ИИ" специализированными аппаратными компонентами (на базе специализированных чипов), что позволит расширить диапазон анализируемых данных, увеличить глубину и точность моделей, сократить время их обучения. Все это даст возможность создавать расширенные озера данных с использованием баз данных с открытым исходным кодом, таких как: Redis, MongoDB, EDB Postgres, Neo4j (графовая база данных) и получить качественно новый уровень ИИ. В-третьих, будет усиливаться интеграция озер данных с OLTP- и OLAP-хранилищами, что позволит выводить на рынок новые "умные" онлайновые бизнес-сервисы.

Публикация подготовлена на основе материалов из открытых источников

Финансовый сектор: 81% одобряют внедрение ИИ

Март 2019 г. — Компания SAS и Глобальная Ассоциация специалистов по управлению рисками (GARP) опубликовали результаты исследования «Искусственный интеллект в банковской сфере и управлении рисками» (*https://www.sas.com/en/white-papers/artificial-intelligence-banking-risk-management-110277.html*) . Согласно полученным данным, искусственный интеллект (ИИ) уже оказывает влияние на все отрасли, в том числе финансовую. Так, 81% специалистов по рискам в сфере финансовых услуг уже успели оценить эффект от внедрения технологий ИИ.

Основные области, в которых респонденты отмечают положительное влияние ИИ: автоматизация процессов – 52%; кредитный скоринг – 45%; подготовка данных – 43%. Почти треть респондентов сообщили об ускорении и повышении гибкости таких процессов, как валидация, калибровка и подбор моделей расчета риска.

Из результатов опроса видно, что из тех специалистов по рискам и финансам, кто еще не использует ИИ, 84% планируют внедрить эти технологии в ближайшие 3 года.

Почти все респонденты ожидают, что в ближайшие три года ИИ поможет им в работе. Детальнее об ожиданиях:

- технологии ИИ приведут к повышению производительности (96%);
- ускорят время получения информации из данных (95%);
- увеличат объем информации и упростят ее обработку для быстрого принятия эффективных решений (95%).

Препятствиями для внедрения ИИ остаются социальное напряжение, например, тревога из-за изменений на рынке труда, и нехватка специалистов для работы с системами ИИ. Больше половины опрошенных – 52% – обеспокоены недостатком квалифицированных кадров. Тем не менее, респонденты уверены, что в их организациях продолжат внедрять ИИ.

Также среди проблем, препятствующих внедрению ИИ, респонденты отметили низкую доступность и качество данных (59%), недостаточное понимание ИИ ключевыми заинтересованными сторонами (54%) и сложности интерпретируемости математических моделей, лежащих в основе работы ИИ (47%).

Pure Storage: DirectFlash™ Fabric

Март 2019 г. — Pure Storage анонсировала DirectFlash™ Fabric, новую технологию сквозной поддержки протокола NVMe и NVMe-oF в операционной среде Purity 5.2, программно-определяемом сердце линейки FlashArray//X. Нововведение поднимает использование гибридного облака с унифицированной инфраструктурой на новый уровень, позволяя запускать приложения где угодно и защищать данные повсюду.

Технология DirectFlash Fabric позволяет клиентам Pure повысить производительность критически важных приложений предприятия, а также новых веб-приложений, которые обычно используют СХД с прямым подключением. Таким образом, Pure становится первым производителем корпоративных систем хранения данных, который полноценно поддерживает транспортные опции NVMe-oF RoCE, что позволяет компаниям размещать флэш-носители ближе к приложениям для улучшения доступа в реальном времени и возможностей консолидации.

DirectFlash Fabric оптимизирует обмен данными между контроллерами массива и хостами по быстрой сети, что делает Ethernet предпочтительным транспортом для СХД в дата-центре. Аналогичные решения сегодня могут предлагать неполный набор корпоративных функций или использовать опцию NVMe over Fabrics с протоколом Fibre Channel, а не с RDMA over converged Ethernet (RoCE), в то время как последний обеспечивает самый большой потенциальный скачок производительности с уменьшением задержки передачи данных на 50% по сравнению с iSCSI. Благодаря нововведению, Pure расширяет возможности технологии DirectFlash для протокола доступа Non-Volatile Memory Express (NVMe) over Fabrics и повышает эффективность работы с ПО по локальной сети. В частности, с Red Hat Enterprise Linux и облачными веб-приложениями, включая MongoDB, Cassandra и MariaDB. Последние получают преимущества и эффективность единой СХД корпоративного класса.

Массивы FlashArray//X поддерживают сквозной протокол NVMe с подключением к Ethernet со скоростью 25G и 50G. Совместимые сетевые карты с поддержкой NVMe-oF уже доступны или планируются к выпуску таких производителей, как: Broadcom, Cisco, Marvell и Mellanox.