Развитие технологии «озеро данных»:

идеология, топология, механизмы, преимущества

Озеро данных – горячая тема, в которую вливается все больше компаний разного калибра. Эта технология является одним из наиболее интересных инструментов для хранения и обработки существенных объемов неструктурированных данных. Расскажем, как она устроена, чем может быть полезна бизнесу и как с ней работать.

Что такое озеро данных

Если говорить коротко, термином «озеро данных» (data lake) обозначают ИТ-инфраструктуру или особый тип хранилища, где данные лежат в сыром виде. Это могут быть файлы любых форматов: от медиаконтента и сканов документов до неструктурированной отчетности, например банковских выписок. Содержимое озера обрабатывают, когда появляется определенная задача: сформировать базу данных для документации, визуализировать те или иные метрики либо реализовать алгоритм машинного обучения.

Такой подход получил название schema-on-read – он означает, что данные в озере хранятся без указания типов. Тип определяют лишь в тот момент, когда выполняют запрос. Чаще всего этот процесс начинают с переливания файлов и документов в специальные хранилища, где уже подготовлена структура для систематизации данных. Она может представлять собой и специальные витрины, состоящие из множества баз с информацией, предназначенной для определенных целей, подразделений компании или ее клиентов.

Работать с этими инструментами помогают data lake-провайдеры, которые предоставляют специализированные решения. С помощью таких продуктов компании могут накапливать огромные массивы данных, легко ориентироваться в них и искать разного рода зависимости. Так, в прошлом году компания Amazon собрала в едином репозитории сведения о коронавирусе SARS-CoV-2 – туда включили научные статьи, сводки о течении заболеваний, вызванных вирусом, результаты исследований. К этой базе обращались медики со всего мира, а фармацевтические компании использовали ее при синтезировании вакцин.

Data lake применяют в самых разных сферах – от науки до индустрии видеоигр. Последние обращаются к озерам для анализа статистических показателей. С помощью сервиса для обработки больших данных Data Lake Insight от SberCloud вендоры собирают метрики, отражающие прирост количества игроков, их отток и объем внутриигровых транзакций. Эта информация становится основой для оценки текущего положения дел и восприятия игрового продукта, что помогает установить вектор его развития.

Рынок озерных технологий сегодня серьезно разогрет. Его объем уже превышает 7,6 млрд долл., и аналитики прогнозируют, что эта цифра продолжит увеличиваться вплоть до 2027 г. – при CAGR равном 20,6%. Однако даже на волне ажиотажа вокруг технологии важно не упустить из виду ключевые тонкости, связанные с построением ИТ-инфраструктуры, а также накоплением и обработкой больших данных в озерах. Если их не учитывать, можно впустую потратить время и финансы организации.

Рынок озерных технологий сегодня серьезно разогрет. Его объем уже превышает 7,6 млрд долл., и аналитики прогнозируют, что эта цифра продолжит увеличиваться вплоть до 2027 г. – при CAGR равном 20,6%.

Чтобы озеро не стало болотом

Возможность накапливать данные про запас и впоследствии не собирать их в срочном порядке для решения какой-то определенной задачи – существенное преимущество озера данных. Однако реализация такой системы требует последовательного подхода. Специалисты международной консалтинговой фирмы McKinsey выделяют четыре классических блока, составляющих процесс интеграции полнофункционального озера: а) его развертывание для хранения сырых данных; б) сбор данных и использование по назначению, например для визуализации метрик или алгоритмов машинного обучения; в) интеграция озера с классическими базами и репозиториями компании; г) адаптация опыта работы с data lake для более эффективного использования озера.

Сложности могут возникнуть еще на первом шаге. Если не следить за качеством втекающей в озеро информации, оно быстро превратится в болото, где завязнут последующие аналитические процессы. Чтобы минимизировать риски, стоит обратить внимание на следующие моменты.

- **Настройка фильтров**, которые проверяли бы параметры, например размер поступающих файлов на соответствие предполагаемому. Так, в процессе загрузки в озеро информации, вроде сырых изображений или видеозаписей могут возникать ситуации, когда из того или иного источника начинает поступать поток нестандартной конфигурации. Если не учитывать такие моменты, озеро вскоре потребует дополнительных финансовых вложений на расширение и поддержку. Поэтому необходимо точнее определиться с тем, какие данные вы собираетесь накапливать, и научить озеро видеть существенные отклонения от заданных рамок.
- **Контроль доступа**. Это требование – своеобразное продолжение предыдущего пункта, поскольку наличие механизмов аутентификации и авторизации оказывает значительное влияние на качество данных в озере. Если компания не реализует политики безопасности и решает открыть доступ к озеру данных для неопределенного круга сотрудников или источников, это может привести к появлению нежелательных (либо нерелевантных) файлов в пуле ресурсов и повышению риска утечки корпоративной информации.
- **Предварительная обработка**. Несмотря на тот факт, что озера заточены под работу с неструктурированными данными, к ним придется «притрагиваться», чтобы облегчить себе жизнь при использовании данных из озера в будущем. Имеет смысл подготовить отдельные сценарии обработки под каждый тип файла, документа или источника. Например, если говорить об агрегировании банковских выписок, то их структура отличается для каждой кредитной организации. Поэтому нельзя исключать, что потребуется внедрить технологию OCR (Optical Character Recognition) для распознавания текста со сканов еще до того, как они попадут в озеро.
- **Формирование метаданных**. Это необходимо, чтобы обеспечить возможность навигации по озерам данных. Например, если data lake хранит определенные географические ресурсы, среди них могут быть спутниковые снимки, векторные или неструктурированные данные о пространственном местоположении и даже 3D-модели. Без метаданных будет сложнее вычленять необходимые файлы среди петабайтов информации – при обращении к озеру, в котором нет никаких ориентиров, аналитик будет вынужден каждый раз перебирать всю информацию с нуля. Об этом следует подумать еще на этапе развертки data lake.

Можно с уверенностью сказать, что для построения с нуля

собственной ИТ-инфраструктуры, включающей озера, нужна экспертиза. Извлечь пользу из массы неструктурированной информации способны только высококвалифицированные аналитики и специалисты по обработке big data. Однако прикладной инструментарий облачных платформ в полной мере закрывает вопросы, связанные с оптимизацией сбора и анализа данных в data lake. Этот факт открывает больше возможностей для рядовых сотрудников, которые могут проводить анализ и извлекать инсайты из корпоративных данных в озерах.

Чем поможет облако

Если раньше специалисты, не обученные работе с большими данными, рисковали утонуть в многообразии неструктурированных файлов, то теперь облачные решения, подобные Data Lake Insight, упрощают процесс обработки информации. В частности, наша система позволяет извлекать данные из озер с помощью обыкновенных SQL-запросов. Кроме того, дополнительную поддержку аналитикам оказывают сервисы, представляющие собой функциональные надстройки над озерами данных. Они помогают следить за качеством поступающей в озеро информации и способны на лету конвертировать файлы из одного формата в другой, а также подготавливать их для быстрого доступа извне.

Наша облачная платформа для работы с data lake сохраняет информацию в колоночно-ориентированном формате Parquet. Он был создан для открытого фреймворка Hadoop, заточенного под работу с большими данными. Колоночный подход к хранению позволяет вычленять и читать только те значения в файлах, которые необходимы в рамках текущего запроса, за счет чего значительно возрастает скорость его выполнения. Кроме того, аналитические запросы можно автоматизировать – для этого достаточно составить алгоритм действий для преобразования файлов. К примеру, можно дать облачной платформе команду записывать поступившие данные во временную таблицу, проводить их дедупликацию (так называют процесс удаления повторяющихся строк или элементов) и, наконец, копировать результат в целевую таблицу.

Стоит заметить, что облачные сервисы способны облегчить работу не только аналитиков, но и системных администраторов, настраивающих ИТ-инфраструктуру для озер. Облачные провайдеры предлагают инструментарий для ее защиты, а также автоматического масштабирования. Наше решение Data Lake Insight для каждой задачи по записи, сбору или обработке данных автоматически выделяет нужное количество вычислительных юнитов. Это кусочки виртуальных машин, имеющие в своем распоряжении необходимые вычислительные ресурсы. Когда задача решена, система высвобождает их в течение часа, и компаниям не приходится переплачивать за неиспользуемые мощности.

В то же время сисадминам не нужно закупать и обслуживать накопители, на которых строится озеро данных, если емкости начинает не хватать. В облаке дополнительные хранилища подключаются по необходимости, так что специалисты компании-клиента могут сфокусироваться на решении профильных задач, которые помогут бизнесу расти и не думать о том, как поддерживать озеро.

В заключение еще раз отметим, что грамотная работа с data lake помогает компаниям из самых разных сфер – маркетинга, логистики, фармацевтики – накопить большие данные и извлечь из них инсайты.

Что в итоге

В заключение еще раз отметим, что грамотная работа с data lake помогает компаниям из самых разных сфер – маркетинга, логистики, фармацевтики – накопить большие данные и извлечь из них инсайты.

Настроить такую систему бизнес может самостоятельно, если помнить о мерах предосторожности – внедрять фильтрацию данных и отсеивать нерелевантные или слишком объемные файлы, форматировать их для дальнейшего использования и применять метаданные для навигации. Эти шаги предотвратят превращение озера в болото бесполезной информации.

Решить большую часть проблем и упростить работу для аналитиков позволяют облачные платформы. Они же помогают снизить риски и взять на себя отдельные задачи по настройке инфраструктуры. Их инструменты делают data lake доступным решением как для крупных корпораций, так и для небольших стартапов, не располагающих серьезными ресурсами. ■

Благодарим за помощь в подготовке статьи Сергея МАЛЫШЕВА, продуктового менеджера компании SberCloud, и архитектора облачных решений Huawei Павла КУТАКОВА