Фабрики больших данных: от общей концепции до реализации

Реализация создания фабрики больших данных. Общая концепция, шаги по реализации, программная реализация

Михаил РОЩИН,
заместитель директора отделения управления проектами и архитектуры IBS

Фабрики данных – относительно новое понятие. Эти системы стали ответом на современные вызовы, в том числе экспоненциальный рост объемов информации, появление новых запросов пользователей и необходимость работы с системами искусственного интеллекта. Что такое фабрика данных и чем она отличается от классических хранилищ и озер данных? Чем может быть полезна бизнесу, и как ее реализовать на практике? Об этом и многом другом рассказывает Михаил Рощин, заместитель директора отделения управления проектами и архитектуры IBS.

Фабрики данных – это хранилища больших объемов разнородной информации, а также механизмы ее сбора, обработки и предоставления вовне. Такие решения стали следующим шагом в эволюции классических хранилищ и озер данных, отличаясь более сложной структурой и расширенными возможностями.

Фабрика работает как со структурированными, так и с неструктурированными данными. Так же, как и хранилище данных, она предназначена для того, чтобы в ней хранилась информация. Но если хранилище представляет собой реляционную базу данных, то фабрика может иметь шардированную архитектуру, что позволяет улучшить производительность и масштабируемость системы.

Помимо этого, фабрика обладает более широким спектром функциональных возможностей для интеграции данных. В ней по-другому реализован подход к управлению данными. В частности, имеются элементы Data Governance для управления качеством, безопасностью и доступностью данных.

В отличие от хранилищ данных в фабрике есть компоненты, позволяющие ей выступать в качестве аналитического сервиса. Она включает в себя нереляционные и реляционные хранилища данных, хранилище баз данных, витрины данных, быстрые СУБД, системы для работы с данными в оперативной памяти (in-memory) и другие составляющие.

Если в классических хранилищах пользователи взаимодействуют с данными через BI-инструменты или выгрузки в Excel, то в фабриках доступны более разнообразные способы работы с ними. Например, виртуализация – единый интерфейс для простого доступа к информации в системе. Кроме того, фабрики больших данных могут предложить пользователям различные сервисы. Например, сервисы машинного обучения, «песочницы» для проведения R&D-экспериментов, предоставление данных вовне и точечные выгрузки по запросам.

Этапы внедрения фабрики данных

Реализация фабрики больших данных во многом похожа

на внедрение классических хранилищ и состоит из аналогичных этапов:

- проектирование системы;
- разработка прототипов;
- тестирование;
- обкатка на большем объеме данных;
- загрузка бизнес-составляющих.

Эти шаги могут быть выполнены с использованием любой методологии – будь то Waterfall или Agile. Главное отличие – в масштабах проекта. Из-за больших объемов данных при создании фабрики весь проект обычно делится на небольшие блоки бизнес-задач, которые можно внедрить в понятные сроки. Кроме того, для успешного внедрения фабрики данных важно активное участие бизнес-пользователей, поэтому задача ИТ-команды – показать, какие выгоды получит бизнес от использования новой системы.

Одним из основных драйверов для подобных проектов стало импортозамещение. Завершив внедрение ключевых систем, таких как ERP и CRM, бизнес готовится к модернизации менее критичных элементов: хранилищ данных, озер данных и BI-компонентов. Появляется возможность устранить накопленные за годы технологические «наследия», обновить базы данных и сделать решения более функциональными.

При реализации фабрики данных возможны два подхода. Первый – поэтапное внедрение небольших функциональных блоков, которые затем дополняются новыми бизнес-областями. В этом случае развертывание базовой архитектуры и подключение первых бизнес-блоков может занять от 6 до 10 месяцев, после чего начинается процесс итеративного расширения системы. Каждая новая итерация обычно длится от одного до двух кварталов, а их количество зависит от масштабов компании и запросов бизнеса.

Второй подход – комплексное внедрение «под ключ» за 1,5–2 года. В этом случае бизнес получает полноценную фабрику данных со всеми необходимыми компонентами. Однако даже после завершения проекта требуется постоянное обновление и расширение системы по мере развития бизнеса.

Одним из главных преимуществ фабрики больших данных является ее блоковая структура. Это похоже на конструктор LEGO: можно собирать необходимые функциональные модули в зависимости от задач бизнеса и исключать ненужные компоненты. Такая гибкость значительно упрощает кастомизацию системы под потребности компании, но она имеет и обратную сторону – для внедрения фабрики данных нужны более квалифицированные специалисты.

В отличие от хранилищ данных в фабрике есть компоненты, позволяющие ей выступать в качестве аналитического сервиса.

Технологический стек

Рассмотрим основные технологические компоненты, применяемые в фабрике данных на каждом этапе работы с информацией.

Интеграция данных

При объединении данных из различных источников применяются два основных подхода – ETL и ELT. При ETL-подходе данные извлекаются, трансформируются и затем загружаются в хранилище. В фабриках данных чаще используется второй подход, когда данные сначала загружаются в систему, а затем трансформируются.

Для автоматизированной интеграции данных обычно используют Apache Airflow и DBT (Data Build Tool). Востребованность этих инструментов объясняется их открытой архитектурой и возможностью быстрого тестирования в реальных проектах. Также может применяться Real-time-интеграция, когда данные поступают в систему в ответ на события в бизнес-процессах. Например, после заключения договора или совершения платежа. Для этого используются шины данных и такие технологии, как Kafka, WSO2 и DATAREON. Кроме того, популярны системы захвата изменений (CDC). Например, Oracle GoldenGate или Open-source альтернатива Debezium, которая успешно интегрируется с шинами данных, включая Kafka.

Хранение

Системы хранения в фабриках данных обычно базируются на «температурном подходе», когда данные делятся на горячие, холодные и основные. Это позволяет распределять их по разным контурам. Горячие данные – те, к которым обращаются часто, поэтому они требуют быстрого доступа и высокой производительности. Холодные данные менее актуальны, но могут быть востребованы в будущем.

Для холодных данных, как правило, используется Hadoop. Этот набор инструментов хорошо подходит для хранения неструктурированных данных, а также создания «песочниц» для их анализа. Hadoop обеспечивает относительно низкую стоимость хранения и эффективно управляет ресурсами, снижая риски того, что некорректный запрос выведет из строя кластер или всю систему. Виртуализация данных в Hadoop также удобна с точки зрения доступа. Его слабой стороной является скорость обработки данных: простые запросы могут выполняться значительно медленнее по сравнению с традиционными реляционными базами данных.

Основные данные обычно размещаются в реляционных базах данных, таких как PostgreSQL и Greenplum. По нашим замерам, PostgreSQL подходит для хранилищ объемом примерно до 15 терабайт, а для больших объемов лучше использовать Greenplum.

Для хранения горячих данных и создания быстрых витрин сейчас чаще всего используется ClickHouse, который опережает другие решения по простоте внедрения и работы с ним, а также благодаря относительно невысокой стоимости.

Управление данными и их качеством

Раньше для управления данными могли использоваться коробочные решения крупных зарубежных вендоров, таких как SAP и Oracle, которые закрывали все потребности в этом направлении. Однако такие системы часто были избыточными и дорогими. Сейчас обычно применяется модульный подход. Например, для каталога данных бизнес может использовать собственные разработки или решения других компаний на базе OpenMetadata.

Одним из главных преимуществ фабрики больших данных является ее блоковая структура. Это похоже на конструктор LEGO: можно собирать необходимые функциональные модули в зависимости от задач бизнеса и исключать ненужные компоненты.

Взаимодействие между сегментами осуществляется либо через интеграционный компонент, либо с помощью модуля XF в Greenplum.

Виртуализация

Для удобного доступа к данным используется виртуализация. Точкой доступа обычно является Greenplum и его язык SQL. Если такой вариант не подходит, то можно вывести точку доступа на отдельный класс систем.

На практике в качестве инструмента виртуализации часто используется Trino. Эта система позволяет создавать единую точку доступа к данным, даже если они хранятся в разных форматах и системах. Trino можно использовать как в Open-source варианте, так и в поддерживаемых коммерческих сборках.

Что касается инструментов контроля качества, то их основная ценность в наличии проверок. Никто лучше самого бизнеса не знает, какие именно проверки нужны и в каком объеме. Поэтому компании часто разрабатывают свои компоненты для контроля качества данных. Как правило, они берут готовые интеграционные инструменты и добавляют к ним необходимые проверки, контрольные процедуры и отчеты по метрикам качества.

Другая важная задача – построение взаимосвязей между данными. Здесь компании часто используют решения на базе DBT.

Предоставление данных

На уровне предоставления данных может применяться Trino или компоненты Greenplum. Эти решения помогают пользователям легко получать доступ к информации.

Сервисы в автоматическом режиме берут данные из фабрики, обрабатывают информацию и предоставляют уже готовые рекомендации. Также стоит отметить так называемые «песочницы данных». Это отделенные друг от друга области с дата-сетами, которые могут использоваться для проведения исследований и тестирования бизнес-гипотез. Их значение возрастает с развитием искусственного интеллекта и машинного обучения.

Еще один важный аспект – предоставление данных вовне. Компании, владеющие большими наборами данных, начинают понимать, что они могут представлять ценность не только для внутренних, но и для внешних пользователей. Первыми в этом направлении стали работать поисковые системы и телеком-компании, которые начали продавать данные о поведении пользователей и их покупательских предпочтениях. Сейчас этот тренд подхватывают компании из других отраслей, хотя многие пока не готовы делиться всей информацией.

Фабрики больших данных на практике

Фабрики данных помогают решать задачи, с которыми не справится ни один другой класс систем, особенно если речь идет об очень больших данных.

В начале 2000-ых весь интернет занимал около 6 петабайт. Современные фабрики уже способны оперировать сопоставимыми объемами данными. Например, у нас есть проекты, где создается фабрика на 4 петабайта. Причем эти данные не просто хранятся в защищенных хранилищах, они используются для проведения сложных вычислений, построения рекомендательных систем, формирования отчетности и передачи информации внешним пользователям. ■