Типичный конвейер больших данных

В мире больших данных главная задача — превратить сырые данные в понимание, которое можно использовать для принятия бизнес-решений. Путь от сбора данных до этого понимания состоит из нескольких ключевых этапов.

1. **Сбор данных**: Первый шаг — это собрать данные с помощью специализированных инструментов. Здесь важно знать, откуда и какие данные вам нужны.

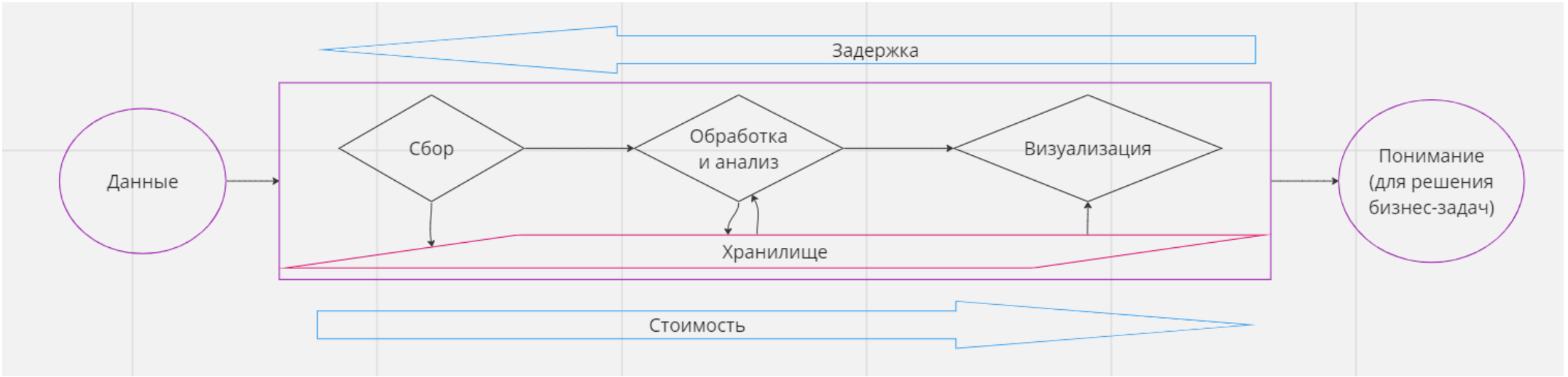
2. **Хранение**: После сбора данные сохраняются в хранилище. Это может быть как облачное хранилище, так и локальный сервер.

3. **Обработка и Анализ**: Затем данные извлекаются из хранилища для обработки или анализа. Здесь можно применять различные методы, начиная от простого фильтрования до сложных алгоритмов машинного обучения.

4. **Дальнейший Анализ**: После первичного анализа данные могут быть повторно использованы для более глубокого исследования или комбинированы с другими данными.

5. **Визуализация и Интерпретация**: Результаты анализа затем превращаются в удобочитаемый формат — диаграммы, графики и так далее — чтобы сделать их доступными для бизнес-аналитиков или других заинтересованных лиц.

6. **Принятие Решений**: Наконец, интерпретированные данные используются для принятия бизнес-решений, запуска новых инициатив или коррекции стратегии.



Весь этот процесс зависит от выбранных инструментов и методов, и тут важно найти баланс между **скоростью (задержкой)** и **стоимостью**. Более быстрые и мощные решения обычно стоят дороже, поэтому необходимо тщательно взвешивать, какие ресурсы вы готовы вложить в анализ данных.

На самом деле часто внедряются процессы ETL (Extract, Transform, Load — Извлечение, Трансформация, Загрузка)

Извлечение (Extract)

На этом этапе происходит сбор данных из различных источников. Это могут быть базы данных, потоки данных в реальном времени, файлы, логи и так далее. Цель этого этапа — подготовить данные к дальнейшей обработке, собрав их в единую временную локацию.

Трансформация (Transform)

Здесь данные преобразуются в формат, который подходит для аналитики. Это может включать в себя очистку данных (удаление дубликатов, исправление ошибок), изменение структуры данных (нормализация, агрегация), а также обогащение данных путем комбинирования с другими источниками информации.

Загрузка (Load)

После трансформации обработанные данные загружаются в хранилище данных или озеро данных для дальнейшего анализа и визуализации. Загрузка может быть как пакетной, так и почти в реальном времени, в зависимости от требований бизнеса.

То есть сначала идёт обработка данных и только потом загрузка (в озеро данных или другое хранилище). Но, можно настраивать и процесс ELT (Извлечение, Загрузка, Трансформация) - например, озеро данных не нуждается в "правильных и чистых" входных данных.

- **ETL**: Сначала извлекаются данные, затем трансформируются (например, очищаются, агрегируются и т.д.), и после этого загружаются в хранилище (чаще всего Data Warehouse). Идеален, когда бизнес-логика трансформации сложна и хранилище оптимизировано для быстрого чтения.
- **ELT**: Данные сначала извлекаются и загружаются в хранилище (чаще всего Data Lake), а трансформация происходит уже в хранилище. Этот подход подходит для сценариев, где важна скорость загрузки данных и их последующая обработка не требует сложной логики.

Применение и ограничения:

- **ETL**: Используется, когда необходима предварительная обработка данных перед анализом. Не подходит для сценариев с реальным временем из-за задержек на этапе трансформации.
- **ELT**: Подходит для сценариев, где нужно быстро загрузить большие объемы данных для последующего анализа. Не идеален для случаев, когда требуется сложная трансформация данных.

Различия в Хранилище данных и Озере данных

Data Lake — это озеро данных, которое может хранить любые типы данных в их исходном формате. Оно подходит для хранения неструктурированных данных и способно масштабироваться горизонтально. Данные в Data Lake хранятся в "сыром" виде и могут быть трансформированы или структурированы при необходимости. Оно работает с ELT.

Применение:

- Хранение больших объемов неструктурированных данных.
- Исследовательские задачи и Data Science.
- Реальное время и потоковая обработка данных.

Data Warehouse — это хранилище данных, предназначенное для быстрого анализа и отчетности. В отличие от Data Lake, оно хранит только структурированные и трансформированные данные. Data Warehouse оптимизирован для быстрого выполнения запросов и хранения данных в формате, удобном для анализа. Оно работает с ETL.

Применение:

- Отчеты и бизнес-аналитика.
- Исторический анализ данных.
- Интеграция данных из различных источников.

Далее по тексту мы не будем привязываться к этой разнице, а в общем рассмотрим инструменты.

Проектирование конвейера для обработки больших данных — сложная и ответственная задача.

Ошибка, которую часто совершают проектировщики, — это попытка использовать один инструмент для решения всех задач, от сбора до анализа и визуализации данных. Хотя такой подход может казаться простым и экономичным, он часто приводит к увеличению рисков, снижению эффективности и высокой стоимости обслуживания.

Декомпозиция конвейера: Лучший подход состоит в том, чтобы "разделить" конвейер на отдельные этапы: сбор, хранение и обработка данных. Такой подход улучшает отказоустойчивость системы. Если, например, на этапе анализа происходит сбой, не нужно начинать всё с начала — можно просто перезапустить этот конкретный этап.

Множественные хранилища: Разделение данных между разными хранилищами позволяет использовать различные хранилища для разных этапов, что облегчает операции чтения и записи.

Выбор инструментов и факторы, которые необходимо учитывать:

1. **Структура данных**: Ваши инструменты должны быть совместимы с форматами данных, с которыми вы работаете.
2. **Задержка и пропускная способность**: В зависимости от того, каковы ваши требования к скорости обработки, выбирайте инструменты, которые могут обеспечить нужный уровень производительности.
3. **Паттерны использования**: Как конечные пользователи будут использовать данные? Нужны ли им сложные запросы, объединение таблиц или просто доступ к определенным записям? Знание этих аспектов поможет вам выбрать правильные инструменты и методы хранения.

Хорошо спроектированный конвейер больших данных должен быть гибким, масштабируемым, отказоустойчивым и, что немаловажно, экономически эффективным. Соблюдение этих принципов позволит создать систему, которая будет эффективно служить вашим бизнес-целям.