НАУЧНЫЙ ЖУРНАЛ
НАУКА И МИРОВОЗЗРЕНИЕ

УДК- 004.041

ОСОБЕННОСТИ РАБОТЫ АНАЛИТИКОВ В ПРОЦЕССЕ РАБОТЫ С БОЛЬШИМИ ОБЪЕМАМИ МАЛО ФОРМАЛИЗОВАННОЙ ИНФОРМАЦИИ

Гараджаева Сульгун Атаевна

старший преподаватель Туркменского государственного университета имени Махтумкули
г. Ашхабад Туркменистан

Сегодня тысячи компаний собирают и хранят большие данные о поведении своих клиентов, ассортименте, состоянии производства и других вещах, важных для бизнеса. Но чтобы принимать взвешенные решения на основе данных, просто собирать их недостаточно — нужен еще грамотный анализ. Рассмотрим, что же включает в себя анализ больших данных и какие инструменты для этого можно использовать.

Что такое анализ больших данных

Чёткого определения того, какие данные считать большими, не существует. Нет какого-то предела объёма, после которого обычные данные превращаются в большие. Но обычно речь идёт как минимум о сотнях гигабайт и сотнях тысяч строк в базах данных. Ещё большие данные, как правило, регулярно пополняются, обновляются и изменяются, то есть их не только хранят, но и активно собирают.

Итак, мы собрали большие данные и сохранили их. Но в таком виде это просто набор информации, который не способен принести никакой пользы. Чтобы польза была, необходим анализ больших данных — их структурирование и обработка по специальным алгоритмам с целью сделать определённые выводы.

Например, у нас есть гипермаркет, в котором люди покупают определённые продукты. Большие данные — это сама информация о покупках: какие именно товары берут люди, как часто, в каких количествах. Анализ больших данных — это изучение этой информации, чтобы понять, каких товаров стоит закупать больше, а какие лучше вообще вывести из ассортимента. То есть в данной ситуации анализ больших данных подразумевает изучение информации о товарах с целью получения результатов, которые могут помочь компании в развитии.

Сбор и хранение больших данных
Существует множество источников больших данных для дальнейшей работ. Например:

● **Статистика поведения пользователей на сайте и в приложении.** Какие страницы они посещают, как долго выбирают товар, какие разделы изучают внимательнее всего.

● **Данные о продажах с касс и из CRM.** Что именно и на какую сумму люди покупают.

● **Информация с датчиков на оборудовании.** Как работают станки в цеху, какая температура поддерживается в помещении, какие каналы человек включает на умном телевизоре.

● **Социальные опросы.** Данные о семейном положении, возрасте, предпочтениях в еде и т. п.

● **Данные из медицинских карт.** Информация о состоянии здоровья пациентов.

● **Записи с камер видеонаблюдения.** Возраст и пол людей, их примерный поток в разное время дня, маршруты по торговому залу.

● **Сборная информация из разных баз данных.** Мы берём несколько баз с «маленькими» данными и собираем всё в одном месте, превращая данные в большие.

После сбора данные необходимо где-то хранить для последующего анализа. Есть три группы мест для хранения.

Базы данных (БД). Их используют для хранения как малых, так и больших данных. В базах хранятся чётко структурированные данные, разложенные по полочкам. Данные из баз проще анализировать, но для хранения их нужно предварительно очищать и структурировать. Это отнимает время и может привести к потере данных, которые пока кажутся бессмысленными, но могут стать полезными в будущем.

Для хранения big data обычно используют:

● Классические реляционные БД: MySQL, PostgreSQL, Oracle. Они надёжные, но плохо масштабируются, поэтому не подходят для огромных массивов данных, которые часто обновляются.

● Нереляционные БД: MongoDB, Redis. Такие БД менее надёжные, но гораздо более гибкие.

Хранилище данных. Это сложная система хранения из нескольких баз данных и инструментов для их обработки и структурирования. Часто она также включает в себя сервисы для проведения анализа данных и их визуализации для пользователей.

Для построения хранилищ данных часто используют Greenplum, ClickHouse.

Озеро данных. Это большое хранилище, в котором лежит много «сырой», неструктурированной информации. Туда можно загружать любые данные, чтобы потом их извлекать, анализировать и использовать в бизнесе. Анализировать их потом сложнее, зато при загрузке никакой анализ и структурирование не нужны.

Для построения озёр данных обычно используют Hadoop.

Часто озёра используют вместе с хранилищами или базами данных. Сначала все данные сгружают в озеро, а потом извлекают из него по определённым критериям, структурируют и кладут уже в хранилище или базу.

Технологии анализа и использования больших данных

Главная задача анализа больших данных — помочь бизнесу действовать правильно и автоматизировать отдельные процессы. Для этого есть разные методы использования и работы с большими данными.

Смешение и интеграция данных. Большие данные часто собирают из множества разных источников. При этом их не всегда можно сгружать в единую базу: часто данные разнородные и к общему виду их не привести.

В таком случае применяют технологию интеграции. Это одновременно и обработка, и анализ данных. Для этого всю разнородную информацию приводят к единому формату. Данные дополняют и проверяют: удаляют избыточные, загружают недостающие из других источников. Часто даже после этого по данным уже можно делать определённые выводы.

Традиционно для интеграции данных используют процессы ETL — извлечение, преобразование и загрузку. На базе этих процессов строят ETL-системы.

Статистический анализ. Статистика — это подсчёт данных по определённым критериям с получением на выходе конкретного результата обработки данных в процентах. Лучше всего статистика работает именно на больших данных, поскольку чем крупнее выборка, тем достовернее результат.

При анализе больших данных могут считать:

● Простые проценты, например долю лояльных клиентов.

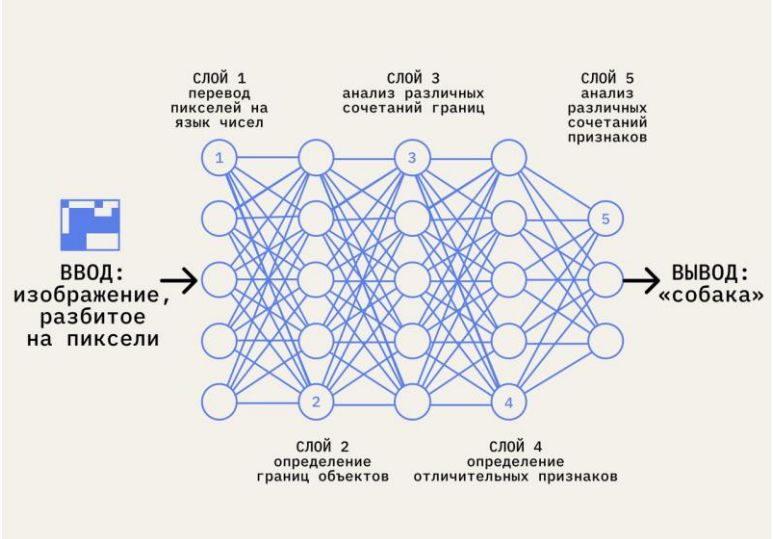
● Средние значения данных из разных групп, например средний чек у разных категорий покупателей.

● Корреляцию, чтобы вычислить, как изменение одних данных влияет на другие. Например, как возраст клиента влияет на его покупательную способность.

Машинное обучение и нейронные сети. Большие данные можно использовать для того, чтобы составлять автоматизированные системы, способные самостоятельно принимать решения. В самом простом виде это чат-боты, которые умеют распознавать ответы пользователей. В сложном — большие распределённые системы управления закупками или производством.

Чтобы такие системы работали, им нужны наработанные паттерны поведения. Эти паттерны извлекаются как раз из работы с большими данными. Система смотрит, как данные изменялись в прошлом, и на основе этого действует в настоящем. Такие системы называют нейронными сетями.

В процессе обучения нейронные сети можно научить анализировать большие данные. Например, нейросети можно «скормить» тысячи фотографий женщин и мужчин. И потом она научится определять пол по фото или видео, что даёт возможность использовать её для классификации поведения покупателей.



Предиктивная аналитика. Это составление прогнозов на основе данных. Например, мы смотрим на поведение покупателей за прошлый год и можем предположить, какой будет спрос на конкретные товары в конкретный день. Или определить, какие именно параметры влияют на поведение клиентов.

Предиктивную аналитику используют, чтобы предсказать колебания валют, поведение покупателей, время доставки грузов в логистике, финансовые показатели компаний.

Для предиктивной аналитики большие данные тщательно изучают, а затем вычисляют корреляции и строят графики, чтобы предугадать, как ситуация повернётся в будущем.

Имитационное моделирование. Предиктивная аналитика помогает предсказать, что будет, если ничего не изменится и система будет существовать в тех же данных. Моделирование же помогает ответить на вопрос: «А что, если?..» Чтобы это сделать, мы строим на базе больших данных максимально точную модель ситуации, а потом меняем в ней параметры: повышаем цену товара, увеличиваем поток клиентов, изменяем размер изготавливаемой на станке детали. Модель реагирует на это и показывает, что будет: как изменится прибыль, что произойдёт с лояльностью клиентов, снизится ли скорость производства.

Инструменты для анализа больших данных

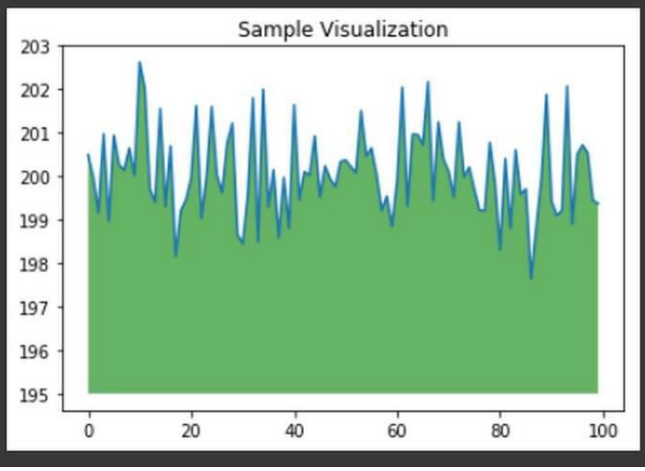
Чаще всего для анализа больших данных используют скрипты и программы, написанные на языке Python. Чтобы работать совместно и эффективно, эти скрипты и программы пишут в специальных интерактивных средах — Jupiter Notebook, Kaggle и Google Collab. Эти среды позволяют выгружать данные, использовать машинное обучение и нейронные сети, собирать статистику.

```python
import numpy as np
from matplotlib import pyplot as plt

ys = 200 + np.random.randn(100)
x = [x for x in range(len(ys))]

plt.plot(x, ys, '-')
plt.fill_between(x, ys, 195, where=(ys > 195), facecolor='g', alpha=0.6)

plt.title("Sample Visualization")
plt.show()
```



Для визуализации результатов анализа данных используют Power BI и Tableau. Они позволяют строить наглядные диаграммы, графики и таблицы для демонстрации результатов аналитики тем, кто недостаточно глубоко разбирается в анализе данных.

Также существуют специальные инструменты и фреймворки для обработки больших данных по разным технологиям: Hadoop, Caffe и другие. Ими пользуются для машинного обучения и сложного анализа данных, выбирая инструмент в зависимости от используемых в компании технологий и бизнес-задач.