П. Е. Новоженцев
(ГГУ имени Ф. Скорины, Гомель)
Науч. рук. **Е. В. Рафалова**, ст. преподаватель

СРАВНИТЕЛЬНЫЙ АНАЛИЗ СОВРЕМЕННЫХ ПОДХОДОВ К ОБРАБОТКЕ И АНАЛИЗУ ГЕТЕРОГЕННЫХ ДАННЫХ

Объем данных, доступных для анализа, растет с каждым днем. Они поступают из различных источников: веб-сайтов, систем мониторинга, журналов логов, датчиков, корпоративных баз данных. Однако сами по себе данные бесполезны, если их не обработать и не представить в удобном виде. Прежде чем извлекать из них ценную информацию, необходимо пройти несколько ключевых этапов: очистку, трансформацию, загрузку в хранилища и последующую визуализацию [1].

На первом этапе работы с данными возникает одна из главных проблем − разрозненность источников и отсутствие единого формата. Например, одни данные могут храниться в виде таблиц Excel, другие − в файлах JSON, третьи − поступать из веб-API или логов серверов. Более того, внутри одного источника могут быть разночтения: даты записаны в разных форматах, в числовых значениях используются разделители, а в некоторых ячейках встречаются пропущенные значения.

Перед тем как данные попадут в хранилище или систему аналитики, их необходимо очистить. Этот процесс включает несколько этапов:

– обнаружение и удаление дубликатов. Если данные поступают из разных источников, возможны ситуации, когда одни и те же записи повторяются. Дубликаты искажают анализ, поэтому их необходимо удалять;

– обработка пропущенных значений. Например, если в таблице измерений уровней шума отсутствуют данные за определенные дни, можно либо заполнить их средними значениями, либо отметить их как «неопределенные» (NaN);

– приведение форматов к единому виду. Это может касаться дат (например, 01.02.2024 и 2024-02-01), числовых значений (1000,50 и 1 000.50), единиц измерения (кг и граммы).

Автоматизация этих процессов возможна с помощью языков программирования (Python, SQL) или специализированных ETL-инструментов, таких как Apache NiFi и Airflow.

Одним из ключевых этапов подготовки данных является процессы ETL (Extract, Transform, Load – извлечение, трансформация, загрузка). Это последовательность действий, которая помогает превратить сырые данные в пригодные для анализа.

Извлечение данных. Данные загружаются из различных источников: API, файловых систем, баз данных. Например, в случае мониторинга уровня шума можно получать данные от датчиков, установленных в разных точках города.

Трансформация. Данные очищаются, нормализуются и объединяются. Например, можно привести все измерения к единому формату (децибелы), удалить выбросы и добавить дополнительные параметры (например, уровень загруженности дорог, если шум измеряется в городской среде).

Загрузка. Подготовленные данные отправляются в хранилище, откуда их можно будет извлекать для аналитики.

После очистки и трансформации данные должны храниться в одном из вариантов хранилища. Существует несколько вариантов: реляционные базы данных (PostgreSQL, MySQL, SQLite), используются для структурированных данных; хранилища данных (Data Warehouse, DWH), такие как Snowflake или Google BigQuery, предназначены для анализа больших объемов информации, позволяют быстро выполнять сложные запросы, агрегировать данные и строить отчеты; озера данных (Data Lake) используются, если необходимо хранить сырые, неструктурированные данные, которые могут быть полезны в будущем (логи работы сети или неочищенные данные с датчиков).

Выбор хранилища зависит от целей проекта: если важна скорость аналитики − предпочтителен DWH, если нужно просто сохранить большой объем данных для возможного использования в будущем — Data Lake. При регулярном обновлении данных важно минимизировать ручной труд. Для этого можно использовать автоматизированные пайплайны:

– Apache Airflow позволяет настраивать сложные процессы обработки данных с возможностью планирования. Например, можно автоматически загружать новые данные каждое утро и обрабатывать их перед загрузкой в хранилище;

– Data Build Tool (dbt) помогает создавать модели данных, упрощая аналитику;

– Power Automate и скрипты на Python позволяют автоматизировать мелкие рутинные задачи, например, загрузку Excel-файлов в базу.

После того как данные подготовлены, их можно анализировать и визуализировать. Необходимо учитывать, что графики должны быть понятными, без перегрузки лишними деталями. Также информация должна быть структурирована, группировка данных должна подчеркивать основные закономерности. Значительный вес имеет выбор правильного типа визуализации. Для временных рядов лучше всего подходят линейные графики, для сравнения категорий − столбчатые диаграммы, для представления состава − круговые диаграммы.

Инструменты для визуализации:

– Power BI − инструмент для построения интерактивных дашбордов;

– Tableau − альтернатива Power BI с расширенными возможностями аналитики;

– Python (Matplotlib, Seaborn, Plotly) − для кастомных графиков и интерактивных визуализаций.

Обработка данных – сложный, но важный процесс, позволяющий превращать хаотичные информационные потоки в ценные инсайты. Современные методы ETL, автоматизированные пайплайны и удобные инструменты визуализации помогают упростить этот процесс, обеспечивая высокую точность и доступность информации.

Литература

1. Новоженцев, П. Е. Разработка программы для сбора и систематизации информации с последующей визуализацией средствами POWER BI / П. Е. Новоженцев, А. С. Руденков // Сборник материалов XII Республиканской научной конференция студентов, магистрантов и аспирантов, посвященной 80-летию со дня рождения профессора Максименко Николая Васильевича – «Актуальные вопросы физики и техники», Гомель. – 2023. – Часть № 1. – С. 430–433.