Power Query и большие данные 3/17/25

Что такое большие данные? Обычно речь о больших данных заходит в тех случаях, когда традиционные методы обработки данных уже не позволяют справиться с объемом, скоростью, многообразием и достоверностью информации. Одного конкретного объема данных, который можно было бы назвать большими данными, не существует – это, как правило, наборы данных, выходящие за пределы возможностей традиционных баз данных и инструментов обработки данных. Что делать, если необходимо исследовать такой гигантский набор данных, а мгновенный доступ к полноценной платформе больших данных отсутствует? В статье рассмотрим, как с помощью доступных инструментов получить некоторое представление о таких больших данных, если специализированного решения еще нет.

В качестве одного из таких инструментов можно использовать Power Query. Это мощный инструмент трансформации данных, встроенный в Excel и Power BI, который предлагает эффективный способ создать некоторое представление о больших данных, позволяя провести первичное исследование и подготовку данных перед выполнением более сложного анализа. Power Query сам по себе не является полным решением для больших данных, однако помогает разобраться и подготовить данные для дальнейшей обработки.

Power Query играет важную роль на начальных этапах рабочего процесса перед обработкой данных. Инструмент выполняет функцию моста между разными источниками данных. Где бы ни находились ваши данные – в больших CSV-файлах, обширных базах данных или в других местах, – Power Query может подключиться и извлечь необходимую информацию. Однако прямая загрузка миллионов или миллиардов строк в программу Excel зачастую непрактична. Поэтому способность Power Query отбирать образцы разных наборов данных приобретает особую важность.

Можно получить меньшие, репрезентативные образцы больших данных для анализа, используя несколько методов. Получение первых N строк обеспечивает быстрое представление, а отбор случайных образцов предлагает статистически более обоснованный набор. Фильтрация на основании особых критериев, например сделок из конкретного региона или временного отрезка, позволяет сосредоточиться на соответствующих подгруппах данных. Когда управляемый образец данных отобран, можно воспользоваться функцией подготовки данных, присущей Power Query. Обычно выполняемые задачи включают обработку недостающих значений, стандартизацию форматов данных (дат, валют и т.д.), обобщение данных на более высоком уровне детализации и создание рассчитанных столбцов для получения новых выводов. К примеру, можно подтвердить удельный вес выручки по клиентам, исходя из истории сделок, или категоризировать продукты, опираясь на объемы продаж.

Однако важно упомянуть и ограничения Power Query при работе с наборами данных, которые можно считать большими данными. Производительность может существенно ухудшиться при увеличении объемов данных, а при использовании Excel в качестве конечного пункта загрузки данных препятствием может стать ограничение количества строк. Поэтому крайне важно оптимизировать этапы обработки данных и запросы, влияющие на скорость обработки. Данные нужно фильтровать в начале процесса, чтобы уменьшить объем загружаемых данных. Желательно уменьшить количество этапов трансформации и использовать соответствующие конвертации типов данных во избежание ненужного использования обрабатывающей мощности. Такой подход подойдет для первичного исследования и анализа меньшего масштаба. Предприятиям, которым требуется обрабатывать большой объем данных, необходимы специализированные платформы

больших данных, например Hadoop, Spark или облачные хранилища данных.

Один из примеров, когда нужно провести первичное исследование данных, – анализ данных о сделках клиентов с платформы э-коммерции. Используя Power Query, можно создать подключение к базе данных, где находятся данные о сделках. Из них можно отобрать образец случайных 10 000 сделок, а затем использовать Power Query для расчета средней стоимости сделки одного клиента, идентификации наиболее продаваемых видов продуктов и сегментации клиентов на основании истории сделок. Такой первичный анализ может обеспечить ценные выводы и указать, какие наборы данных стоит проанализировать углубленно с помощью специализированных инструментов.

Резюмируя вышеизложенное, отметим, что Power Query – практичный и доступный инструмент для получения некоторого представления о больших данных. Он обеспечивает отбор и обработку образцов данных, предоставляя пользователям возможность исследовать гигантские наборы данных и создать первичное представление без полной инфраструктуры больших данных. Даже если платформа обработки больших данных недоступна, с помощью Power Query можно провести первичное исследование данных. Для пополнения знаний о больших данных приглашаем изучить ресурсы, посвященные хранилищам данных, облачным вычислениям и методам раздельной обработки.