

Пакетная аналитика (Batch analytics)

Разделы: [Бизнес-задачи](#)

Пакетная аналитика — технология [анализа данных](#), в которой информация, поступающая из множества источников, сначала собираются в блоки, называемые пакетами, и сохраняются на постоянных носителях, а затем подвергаются аналитической обработке по мере необходимости. Пакеты обрабатываются по порядку, а результаты сохраняются перед переходом к следующему элементу. В итоге появляется возможность автоматизировать процесс, сделав его более эффективным и [масштабируемым](#).

Пакетная аналитика является альтернативой [поточковой](#), которая предполагает, что данные обрабатываются в режиме реального времени по мере их формирования и в процессе перемещения в [хранилище](#) (т.е. в потоке).

Данные обрабатываются пакетами с определенной периодичностью в соответствии с заданным регламентом. Например, каждый час, день, неделю и т.д. Пакетная аналитика позволяет выполнять более сложную обработку данных, которая невозможна при потоковой.

Пакетная аналитика является менее затратной, чем потоковая, так как не требует применения специальных аппаратных и программных платформ. Кроме этого, поскольку анализ пакетов выполняется не «на лету», а по «вызову», обработку [больших массивов данных](#) можно производить, не перегружая информационную систему компании.

При внедрении пакетной аналитики, как правило, требуется определить следующие параметры:

- **Размер пакета** — количество данных, обрабатываемых за один раз. Он выбирается исходя из производительности системы и требуемого объема вычислений, так, чтобы время ожидания результатов было в пределах разумного, а сами результаты были своевременными и не потеряли актуальность. Размер пакета может задаваться в единицах объема данных или времени их накопления (например, за час, день и т.д.).
- **Частота обработки** — периодичность, с которой выполняется анализ пакетов. Она выбирается в зависимости от потребности компании в результатах анализа, а также от интенсивности потоков данных.
- **Время обработки** — это время, которое требуется для обработки одного пакета. Его следует оптимизировать, чтобы процесс завершался вовремя и не создавал узких мест в системе.
- **Отказоустойчивость** — способность системы восстанавливаться после сбоев во время пакетной обработки. Ее уровень должен быть таков, чтобы данные

гарантировано не были потеряны или повреждены в процессе обработки.

- **Интеграция** — это процесс объединения данных из нескольких источников для формирования репрезентативного набора.

К недостаткам пакетной аналитики можно отнести:

- **Отложенная обработка** — анализ данных производится «по требованию» или с некоторой периодичностью. Это приводит к задержке получения результатов анализа относительно событий, отражаемых данными, и делает невозможным принятие упреждающих решений в проблемных ситуациях.
- **Ограниченная интерактивность** — прежде чем данные станут доступными для анализа, они должны быть перемещены в систему хранения, что может привести к определенной задержке результатов. Кроме этого, может потребоваться предобработка и очистка данных, что потребует дополнительного времени.

Таким образом, внедряя решения в области BI, необходимо тщательно оценивать преимущества и недостатки пакетной и потоковой аналитики, учитывая бизнес-цели и сферу деятельности компании, интенсивность и объем потоков данных в ней, а также сложность вычислений, применяемых в ходе анализа.