Темные данные (Dark data)

Темные данные — это обобщающий термин для информации, которая собирается, накапливается и хранится, но никак не используется. Несмотря на их наличие, данные не принимаются в расчет, в результате чего могут быть сделаны неправильные выводы и приняты неверные решения.

Такие данные обычно составляют большую часть информационных активов организаций. Можно сказать, что они являются «подводной частью айсберга» всех данных, имеющихся у компании, но из-за сложностей в обработке не используются для улучшения бизнес-процессов или извлечения прибыли.

Объем темных данных в организациях постоянно растет, что происходит по следующим причинам:

- **Устаревание**. Большинство информации имеет ценность только в течение какого-то времени, например, данные о клиентах. Зачастую они не удаляются даже после их использования и когда теряют актуальность. Особенно данная проблема актуальна для автоматически собираемых данных: логи, телеметрия, записи разговоров и т.п.
- **Конфиденциальность**. Существуют стандарты, требующие от компаний соблюдения правил, связанных с защитой информации. Примером могут быть персональные

данные, которые многие организации хранят в большом объеме, но пользоваться ими произвольным образом не могут.
- **Разрозненность**. В ходе работы над каким-либо проектом сотрудники собирают и локально сохраняют информацию для последующего анализа. В будущем она могла бы пригодиться, но из-за разрозненности коллеги не догадываются о ее существовании, к тому же часто такие сведения дублируются.
- **Трудность обработки и анализа**. Данные в исходном виде, такие как изображения, видео или аудиофайлы не удобны для анализа. Необходимо приложить много усилий, чтобы преобразовать их к нужному представлению.

Основным недостатком в сборе и хранении темных данных является то, что это пустая трата денежных ресурсов и времени. Также стоит отметить риски утечек данных, которые сложно обнаружить в связи с тем, что информация не контролируется и чаще всего о ней ничего не известно службам безопасности компании.

Составить исчерпывающую классификацию темных данных практически невозможно, так как существует огромное количество их типов. Но можно выделить две основные группы:

1. Не используемые данные, о существовании которых известно. Обычно это автоматически собираемые, устаревшие или конфиденциальные данные.
2. Данные, которые собираются, но о них знает ограниченное число лиц. В основном это информация, собираемая и хранящаяся локально в департаментах компаний или на рабочих станциях сотрудников.

Как правило, Dark data — это неструктурированные данные, работа с которыми затруднена. Отчасти эту задачу можно решать с внедрением искусственного интеллекта. Нейросети и машинное обучение позволяют упростить данный процесс.

Для эффективной работы с темными данными необходимо:

- выполнять периодическую проверку и очистку ненужных и устаревших данных;
- установить четкие правила для хранения и использования информации;
- внедрять современные инструменты для управления данными;
- использовать технологии работы с неструктурированными данными.

Необходимо помнить о безопасности и правилах хранения информации. Темные данные могут быть как эффективным средством повышения качества продукта или решения каких-либо задач, так и являться фактором риска, в зависимости от того, насколько компания способна их использовать.