Неструктурированные данные (Unstructured data)

> Синонимы: Неструктурированная информация, Unstructured information

Неструктурированными называют данные, не соответствующие заранее определенной модели их представления, которая позволяет выполнять машинную обработку без предварительных процедур. Понятие «неструктурированные данные» определяется не само по себе, а скорее как противоположность структурированным данным, которые организованы в табличную структуру, состоящую из типизированных столбцов (полей) и строк (записей), и могут обрабатываться машинными алгоритмами напрямую.

Традиционно к неструктурированным относят данные, не имеющие табличной структуры, то есть тексты, изображения, аудио и видеозаписи. Такие данные могут содержать полезную информацию, которая служит источником новых знаний, используемых для поддержки принятия решений. Однако применять к ним аналитические алгоритмы без предварительной структуризации нельзя.

Проблема неструктурированных данных в том, что их сложно включать в рабочие процессы бизнес-аналитики вместе со структурированными, поскольку для них требуются специфические системы хранения и алгоритмы обработки. Так, для неструктурированных данных невозможно применять традиционные хранилища и витрины данных, использующие реляционные или многомерные модели, поэтому их приходится накапливать в озерах данных и системах NoSQL.

При этом около 90% данных, генерируемых компаниями, являются именно неструктурированными. Но игнорировать их нельзя, поскольку они также несут полезные знания. Можно выделить два вида источников неструктурированной информации:

- генерируемые человеком — текстовые документы, электронные письма, посты в соцсетях, изображения, видео и т.д.;
- машиногенерируемые — автоматически создаются различными устройствами: журналы логов, данные позиционирования мобильных устройств, результаты работы Интернета вещей и т.д.

Большая часть неструктурированных данных представлена в виде текста: сообщения электронной почты, документы, блоги и публикации в социальных сетях, расшифровки звонков и др.

Независимо от происхождения, аналитическая обработка неструктурированных данных более сложна и затратна. Несмотря на трудности, они являются ценным ресурсом, который при правильном использовании может служить источником знаний и обеспечивать конкурентные преимущества.

Кроме того, выделяют слабоструктурированные (semi-structured) данные, которые занимают промежуточное положение между структурированными и неструктурированными и обладают частичными свойствами как тех, так и других. Слабоструктурированные данные также называют полуструктурированными, плохо структурированными или частично структурированными.

Так же, как и неструктурированные, слабоструктурированные данные не имеют заранее определенной модели или схемы представления. Тем не менее они содержат теги или другие маркеры для разделения семантических элементов и обеспечения иерархии записей и полей. Примерами полуструктурированных данных могут служить файлы JavaScript Object Notation (JSON), CSV и XML.

В слабоструктурированных данных объекты, принадлежащие к одному классу, могут иметь разные атрибуты, даже если они сгруппированы вместе, при этом порядок атрибутов не имеет значения. Иными словами, схема данных может быть инвариантной для разных записей. Например, адреса или наименования изделий, которые в каждой строке могут быть представлены в разном формате.

Характеристики	**Структурированные данные**	**Неструктурированные данные**	**Слабоструктиро данные**
Формат	Табличный (поля, записи)	Текст, изображения, аудио, видео	XML, CSV, JSON, ED
Модель данных	Реляционная, многомерная	Отсутствует	Иерархическая, инвариантная
Системы хранения	Реляционные и многомерные СУБД, хранилища и витрины данных	Озера данных, NoSQL СУБД	NoSQL СУБД
Схема данных	Определенная, фиксированная	Неопределенная	Переменная
Методы анализа	Интеллектуальный анализ данных, машинное обучение	Алгоритмы анализа соответствующего вида контента	Интеллектуальны анализ данных

До появления технологий больших данных, позволяющих работать как со структурированными, так и с неструктурированными данными, последние считались темными данными, поскольку сложности, связанные с их хранением и обработкой, делали их малопригодными для использования во многих задачах бизнес-аналитики. Однако в настоящее время компании, располагающие большими объемами неструктурированных

данных, обладают значительным стратегическим активом: сочетание структурированной и неструктурированной информации дает наиболее полное представление о бизнес-процессах предприятия.