Тема 1. Введение в анализ больших данных. Основные понятия и характеристики

Лекция 1.1. Феномен больших данных: происхождение и причины роста

Термин Big Data появился в научном обиходе в конце 1990-х годов, когда рост вычислительных мощностей и распространение цифровых устройств привели к лавинообразному накоплению данных. Впервые проблему обозначил Роджер Маугли в 1997 году, описав ситуацию, когда объём визуализируемых данных превышает ёмкость оперативной памяти. В 2001 году аналитик Gartner Дуг Лейни систематизировал наблюдения и ввёл модель трёх «V» — Volume, Velocity, Variety, — которая стала первым формальным определением явления.

К началу 2010-х годов стало очевидно, что три характеристики описывают явление не в полной мере. Исследователи добавили Veracity (достоверность) и Value (ценность), сформировав модель 5V, которая сегодня является стандартом при описании больших данных.

Три глобальных процесса, обусловивших взрывной рост данных

— Цифровизация всех сфер деятельности. Транзакции, медицинские записи, навигационные треки, производственные показатели — каждая область жизни генерирует цифровой след. По данным IDC, к 2025 году мировой объём данных достиг 120 зеттабайт.

— Распространение интернета и мобильных устройств. Каждую минуту пользователи загружают около 500 часов видео на YouTube, отправляют свыше 200 миллионов электронных писем, выполняют более 4 миллионов поисковых запросов в Google.

— Интернет вещей (IoT). Промышленные датчики, умные счётчики, носимые устройства и системы видеонаблюдения непрерывно генерируют потоки данных, которые в совокупности формируют значительную часть мирового объёма информации.

Эволюция инструментов работы с данными шла параллельно с ростом их объёма. В 1970-е годы основу составляли реляционные базы данных (IBM System R, Oracle). В 1990-е появились хранилища данных и OLAP-системы. В 2004 году Google опубликовал статью о MapReduce, а в 2006 году Apache выпустил открытую реализацию — Hadoop. В 2010-е годы Apache Spark заменил MapReduce как основная платформа обработки, а облачные провайдеры (AWS, Azure, GCP) сделали инфраструктуру Big Data доступной для компаний любого размера.

Лекция 1.2. Модель 5V: детальная характеристика больших данных

Модель 5V описывает большие данные через пять измерений. Каждое из них порождает конкретные технические и организационные задачи, решение которых составляет суть работы с большими данными.

Volume — объём

Объём данных давно вышел за пределы возможностей одного сервера. Большой адронный коллайдер генерирует около 15 петабайт в год. Социальные сети хранят сотни петабайт пользовательского контента. Геномика: полная последовательность генома одного человека занимает около 3 гигабайт сырых данных, а для полноценного анализа требуется многократно больше. Единицы измерения: терабайт (10 в 12 степени байт), петабайт (10 в 15 степени), экзабайт (10 в 18 степени), зеттабайт (10 в 21 степени). Объём данных определяет выбор инфраструктуры хранения — распределённые файловые системы, сжатие, партиционирование.

Velocity — скорость

Скорость описывает как темп генерации, так и требования к скорости обработки. Система обнаружения мошенничества Visa анализирует более 65 000 транзакций в секунду и должна вынести решение до завершения транзакции — то есть за доли секунды. Биржевые алгоритмы принимают решения за микросекунды. Производственные системы мониторинга требуют обработки показателей датчиков в режиме реального времени для предотвращения аварий. Различают три режима обработки: пакетный (batch) — данные накапливаются и обрабатываются периодически; потоковый (streaming) — данные обрабатываются по мере поступления; микро-пакетный (micro-batch) — компромиссный вариант с задержкой в несколько секунд.

Variety — разнообразие

— Структурированные данные: таблицы с чёткой схемой в реляционных базах данных. Составляют около 20% всех данных в мире.

— Полуструктурированные данные: JSON, XML, логи, данные IoT-сенсоров. Имеют теги или метки, но схема гибкая или отсутствует.

— Неструктурированные данные: текст, изображения, аудио, видео, PDF-документы. Составляют около 80% мирового объёма данных и плохо поддаются традиционным методам обработки.

Veracity — достоверность

Реальные данные содержат ошибки, пропуски, дубликаты и противоречия. Принцип GIGO (Garbage In — Garbage Out) означает: некачественные данные на входе дают ненадёжные результаты на выходе независимо от сложности модели. По данным Harvard Business Review, аналитики тратят до 50% рабочего времени на поиск,

исправление и согласование данных. Работа с достоверностью включает валидацию, дедупликацию, импутацию пропусков и управление мастер-данными (MDM).

Value — ценность

Ценность — конечная цель всей работы с данными. Сами по себе данные не имеют ценности: она возникает в процессе анализа и принятия решений на его основе. Netflix сокращает отток подписчиков, точно предсказывая интересы пользователей. Amazon генерирует около 35% выручки через рекомендательные системы. DeepMind сократил энергопотребление систем охлаждения дата-центров Google на 40% благодаря анализу данных.

Лекция 1.3. Сравнение Big Data с традиционным анализом данных

Традиционный подход к анализу данных сложился в 1970–1990-е годы и ориентирован на структурированные данные ограниченного объёма. Его инструменты — реляционные СУБД, SQL, OLAP-кубы, BI-системы — хорошо работают при объёмах до нескольких терабайт и строгой схеме данных. Big Data-подход возник как ответ на ограничения классических инструментов.

Критерий	Традиционный подход	Big Data подход
Объём	Гигабайты — терабайты	Терабайты — зеттабайты
Скорость обработки	Пакетная (часы, дни)	Реальное время / потоковая
Тип данных	Структурированные	Все типы, включая неструктурированные
Хранилище	RDBMS, Data Warehouse	HDFS, Data Lake, NoSQL
Схема данных	Schema-on-write (заранее)	Schema-on-read (при чтении)
Масштабирование	Вертикальное (мощнее сервер)	Горизонтальное (больше узлов)
Инструменты	SQL, Excel, SAS, SPSS	Hadoop, Spark, Kafka, HBase

Ключевое технологическое отличие — горизонтальное масштабирование. Вместо замены одного дорогого сервера на более мощный, Big Data-системы добавляют новые стандартные узлы в кластер. Тысяча серверов по тысяче долларов дешевле и надёжнее одного суперсервера за миллион: выход из строя одного узла не останавливает работу всей системы, данные автоматически восстанавливаются из реплик.

Лекция 1.4. Области применения больших данных

Здравоохранение

— Геномика и персонализированная медицина: сравнение тысяч геномов позволяет выявлять генетические маркеры заболеваний и подбирать терапию индивидуально. Проект «Геном человека» создал базу данных объёмом около 200 гигабайт на одного человека.

— Электронные медицинские карты: анализ истории болезней, лабораторных данных и показателей мониторинга позволяет прогнозировать риски инфаркта, диабета, сепсиса.

— Эпидемиология: во время пандемии COVID-19 данные о перемещениях людей использовались для моделирования распространения вируса и оценки эффективности ограничительных мер.

Финансы

— Обнаружение мошенничества в реальном времени: каждая транзакция проверяется на соответствие поведенческому профилю клиента и общим паттернам мошенничества.

— Алгоритмическая торговля: системы High-Frequency Trading принимают инвестиционные решения за микросекунды на основе потоков биржевых данных.

— Кредитный скоринг нового поколения: помимо кредитной истории, модели учитывают поведение в интернете, историю платежей за коммунальные услуги и геолокационные данные.

Промышленность и транспорт

— Предиктивное техническое обслуживание: датчики на оборудовании генерируют непрерывный поток показателей. Анализ отклонений позволяет предсказать поломку за несколько дней и провести обслуживание в плановом порядке.

— Цифровые двойники: виртуальная копия физического объекта (завода, двигателя, здания), питаемая данными с датчиков. Позволяет тестировать сценарии без риска для реального объекта.

— Оптимизация логистики: Amazon и UPS используют большие данные для построения оптимальных маршрутов доставки с учётом трафика, загрузки складов и прогноза спроса.

Лекция 1.5. Роли специалистов в экосистеме больших данных

Работа с большими данными — многопрофильная деятельность, требующая команды специалистов с разными компетенциями. Понимание ролей необходимо для правильного выстраивания процессов внутри организации.

— Data Engineer — строит и поддерживает инфраструктуру: пайплайны сбора данных, хранилища, ETL-процессы. Основные инструменты: Hadoop, Spark, Kafka, Airflow, SQL.

— Data Scientist — исследует данные, строит статистические и машинного-обучения модели, формулирует и проверяет гипотезы. Инструменты: Python, R, Jupyter, Spark MLlib.

— Data Analyst — отвечает на конкретные бизнес-вопросы с помощью анализа данных, создаёт дашборды и отчёты для нетехнических аудиторий. Инструменты: SQL, Power BI, Tableau.

— ML Engineer — переводит модели Data Scientist в производственную среду, обеспечивает их масштабируемость, мониторинг и обновление (MLOps).

— Chief Data Officer (CDO) — отвечает за стратегию работы с данными на уровне организации: governance, качество, доступность, соответствие регуляторным требованиям.

Лекция 1.6. Этика и правовое регулирование в области больших данных

Персональные данные и конфиденциальность

Анализ больших данных часто затрагивает персональную информацию граждан. В Европейском союзе с 2018 года действует GDPR (General Data Protection Regulation) — один из наиболее строгих регуляторных актов в мире. Штрафы за нарушение достигают 4% годовой выручки компании или 20 миллионов евро. В Казахстане действует Закон «О персональных данных и их защите», устанавливающий аналогичные требования к сбору, хранению и обработке данных граждан.

Ключевые принципы: минимизация данных (собирать только необходимое), ограничение цели (использовать только в заявленных целях), право на удаление («право на забвение»), прозрачность обработки. Технические меры: анонимизация, псевдонимизация, шифрование, контроль доступа.

Алгоритмическая предвзятость

Если обучающие данные содержат системную предвзятость, модель её воспроизводит и нередко усиливает. Система COMPAS, применявшаяся в США для оценки риска рецидива преступлений, по результатам независимых исследований демонстрировала статистически значимые различия в оценках для разных демографических групп. Специалист по данным обязан анализировать данные на предвзятость, выбирать соответствующие метрики качества и документировать ограничения модели. Концепция Responsible AI включает принципы объяснимости (explainability), справедливости (fairness) и подотчётности (accountability).