Введение в Data Engineering

Innowise

innowise

Немного о данных

Какими 'умными' сервисами вы пользуетесь?

Хочешь узнать, что влияет на то, как люди **делают покупки, путешествуют или общаются в Интернете?**

Планы и цели лаборатории

01

Повышение осведомленности о направлении Data Engineering

Несмотря на прогрессивный рост направления, многие не понимают специфику профессии и отличие от других

Определим в чем же все таки отличие от Python Backend :)

02

Дать базовые знания

В специальность не входят за месяц, наша цель - дать базовые знания и развить навыки для дальнейшего погружения в направление

Строим roadmap развития внутри специальности, сопровождаем на протяжении лабы

Планы и цели лаборатории

03

Развитие бренда в регионе

Innowise растет очень быстро. Хотим формировать сильное комьюнити, растить репутацию, укреплять место на рынке.

Все вместе принимаем участие в создании сильного бренда

04

Тест нового формата

Вечерние лабы для нас эксперимент. Хотим обкатать процесс, собрать фидбек, понять где необходимы улучшения, запускать заново.

Строим новый формат, ждем ваших фидбеков

innowise

Что покроем за месяц?

1

Python

Python помогает автоматизировать задачи и создавать гибкие конвейеры данных.

- Базовые структуры данных (списки, словари, множества, кортежи)
- Многопроцессорная обработка и многопоточность

2

SQL

SQL позволяет задавать вопросы и находить полезную информацию в больших массивах данных.

- Аналитические запросы (оконные функции, CTE)
- Хранимые процедуры и оптимизация

3

Core Data Engineering Theory

Основные знания гарантируют, что вы создадите надежные системы данных.

- ETL и ELT, пакетная обработка и потоковая обработка
- Моделирование данных, форматы файлов, индексирование

innowise

Почему это полезно именно вам?

- **Понять, подходит ли вам это направление:** Это честный тест-драйв профессии, который поможет принять взвешенное решение о дальнейшей карьере

- **Получить прочный фундамент:** Вы получите структурированные знания, с которыми гораздо проще выбрать вектор дальнейшего развития в мире больших данных.

- **Получить уникальные возможности:** Это шанс пообщаться с опытными инженерами, решить реальные кейсы и, для лучших студентов, получить возможность пройти стажировку.

innowise

Немного о расписании

- **4 августа:** Введение в Data Engineering (сегодняшняя тема)
- **8 августа:** Базы данных, Нормализация
- **11 августа:** SQL: Основы
- **14 августа:** SQL: Продвинутые команды (TCL, DML, DDL, DCL)
- **18 августа:** Хранилища данных (Data Warehouse)
- **21 августа:** Введение в Python
- **25 августа:** Python: Циклы и условия
- **29 августа:** Python: Строки, списки, словари
- **1 сентября:** Python: Функции

innowise

Наши ожидания

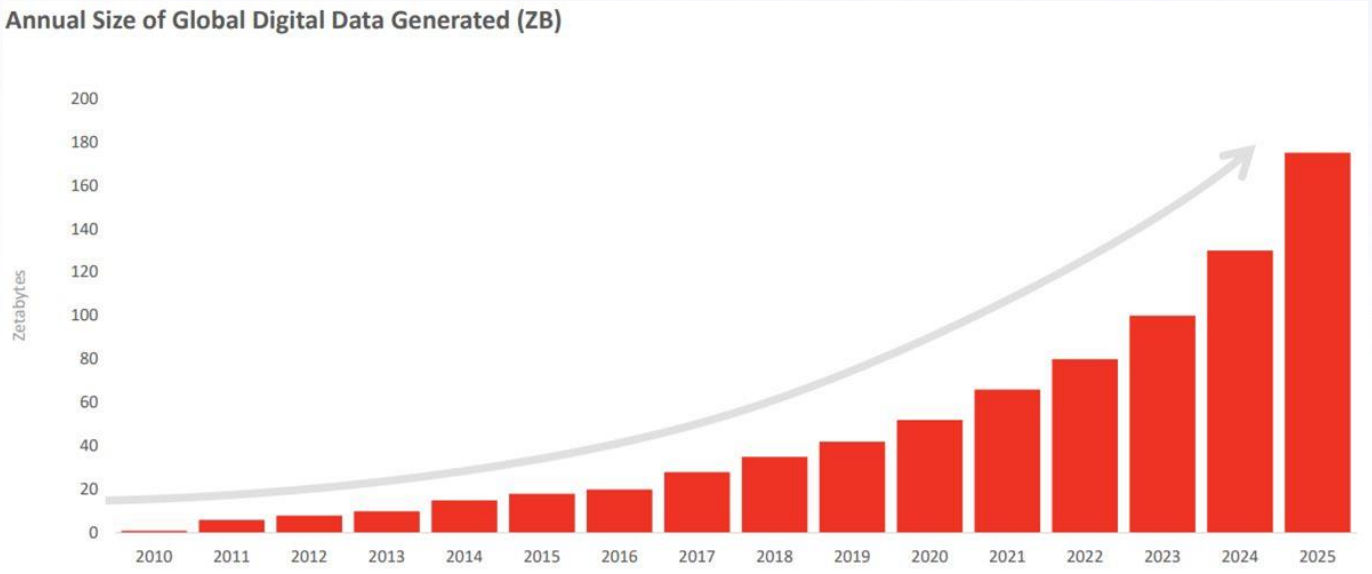
- **Активное участие:** Мы ждем ваших вопросов и вовлеченности в обсуждения.
- **Регулярное посещение:** Старайтесь не пропускать занятия, чтобы не отставать от программы.
- **Соблюдение дедлайнов:** Своевременное выполнение практических заданий — ключ к успеху.
- **Самостоятельное изучение:** Готовность к дополнительному изучению материалов для закрепления тем.

Как мы работали с данными раньше?

Как рос объём информации?

Цифровой потоп: как мы создали океан информации

Перед вами не просто линии, а визуальное доказательство самой быстрой технологической революции в истории. Всего за одно поколение мы перешли от мира, где онлайн были единицы, к миру, где более двух третей человечества имеет доступ к сети. Этот график иллюстрирует не просто рост, а настоящий "Большой взрыв" данных. Если в начале 2010-х мы измеряли весь объем интернета в нескольких зеттабайтах, то сегодня мы генерируем такое количество информации каждый месяц.

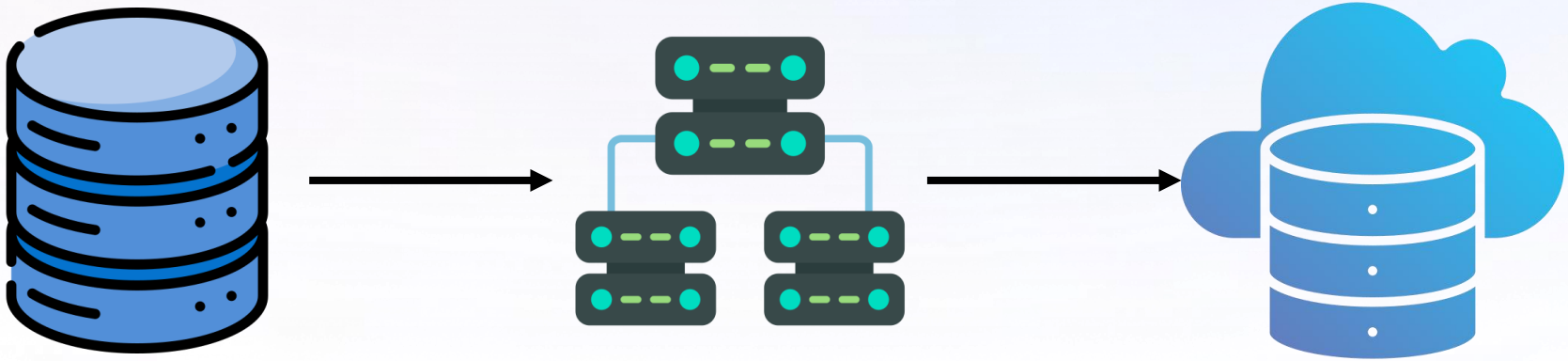


Возили данные в ФУРАХ?!

История о том, что **Amazon Web Services** (AWS, один из самых популярных облачных сервисов) перевозила жёсткие диски фурами. Этот сервис назывался **AWS Snowmobile**, и он представлял собой грузовик, перевозящий 45-футовый (около 14 метров) защищённый морской контейнер, способный транспортировать до **100 петабайт** данных за один рейс. Это эквивалентно 100 000 терабайт.

Этот, казалось бы, низкотехнологичный на первый взгляд метод на самом деле был самым быстрым способом перемещения **эксабайтных** (1 000 000 терабайт) объёмов данных в облако. Передача такого же объёма информации по высокоскоростному интернет-соединению в 1 Гбит/с заняла бы **более 20 лет**.

Как компании адаптировались?



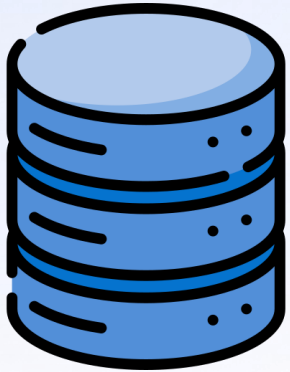
Локальные базы данных (PostgreSQL, Oracle)

Распределенные хранилища (Hadoop, Greenplum)

Облачные хранилища (AWS, GCP)

Локальные Базы Данных

Парадигма целостности данных и вертикального роста



В основе этого этапа лежит одна главная аксиома: **целостность данных превыше всего**. Для бизнеса, работающего с финансами, заказами или складскими остатками, любая ошибка или потеря данных недопустима.

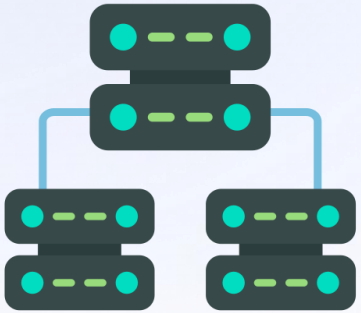
Фундаментальный принцип — **ACID**. Он гарантирует максимальную надежность: данные не теряются, не повреждаются, а операции всегда завершаются корректно.

Архитектурная модель: Scale-Up (Вертикальное масштабирование). Теория была проста: если вам нужна большая производительность, вы покупаете более мощный сервер. Этот подход упирался в два теоретических предела: **физический** (нельзя бесконечно наращивать мощность одного сервера) и **экономический** (каждое следующее удвоение мощности стоит экспоненциально дороже).

Эволюционный тупик: Эта парадигма идеально работала для структурированных данных и предсказуемых нагрузок, но оказалась совершенно неготовой к взрывному росту интернета, который принес с собой хаос неструктурированной информации и непредсказуемый трафик. Модель "один большой сейф" исчерпала себя.

Распределённые Хранилища

Парадигма масштабируемости и гибкости



В основе этого этапа лежит новая аксиома: **масштаб и скорость** важнее идеального порядка. Когда интернет породил лавину неструктурированных данных (Big Data), стало ясно, что пытаться уместить их в один "сейф" — бессмысленно.
Фундаментальный принцип — **CAP-теорема**. Она показала, что нельзя одновременно быть и на 100% согласованным, и всегда доступным при сбоях сети. Пришлось пойти на компромисс: пожертвовать мгновенной согласованностью ради того, чтобы система работала без перебоев.

Архитектурная модель: Scale-Out (Горизонтальное масштабирование). Теория изменилась кардинально: вместо одного дорогого супер-сервера система строится из сотен или тысяч дешевых, стандартных машин, работающих как единое целое. Это позволило достичь практически безграничного масштаба.

Эволюционный тупик: Эта парадигма решила проблему масштаба, но породила колоссальную операционную сложность. Управление таким "флотом" серверов требовало огромных вложений и целой армии высококлассных инженеров, что было недоступно большинству компаний.

Облачные Хранилища

Парадигма абстракции и эластичности

В основе этой эры лежит идея, изменившая всю индустрию: **инфраструктура — это сервис, а не собственность**. Цель — сделать мощнейшие технологии доступными для всех, скрыв их сложность. **Фундаментальный принцип** — абстракция и оплата по факту (**Pay-as-you-go**). Пользователь больше не думает о "железе". Он арендует необходимые ресурсы — от хранилища до вычислительной мощности — и платит только за то, что реально использует, как за электричество из розетки.

Архитектурная модель: Эластичная инфраструктура как сервис (**IaaS**). Ключевым стало полное разделение хранения данных и их обработки. Это дало невероятную гибкость: можно хранить петабайты информации за копейки и подключать для их анализа мощные вычислительные ресурсы лишь на время, когда они нужны.

Итог эволюции: Облако не стало "тупиком", а новой ступенью развития. Оно не отменило предыдущие подходы, а предложило их в виде удобных сервисов. Это демократизировало технологии, позволив любой компании получить доступ к инфраструктуре мирового класса без капитальных вложений.

innowise

Что такое Big Data?

Big Data — это большие, сложные наборы данных, которые **традиционные инструменты обработки данных не могут эффективно обрабатывать.** Обычно они характеризуются 5 V:

Volume

Огромный объем данных, генерируемый каждую секунду.

Velocity

Скорость, с которой создаются, обрабатываются и анализируются данные.

Variety

Различные типы и форматы данных (структурированные, неструктурированные, полуструктурированные).

Veracity

Достоверность и качество данных.

Value

Инсайты и бизнес-преимущества, извлеченные из данных.

innowise

Volume, 1 петабайт - это...

200 000 HD-фильмов
Если бы вы начали смотреть их без остановки, это заняло бы у вас **больше 2,5 лет**

200 миллионов фотографий с разрешением 5 Мб.
Это в **5000 раз больше**, чем может вместить самый большой iPhone

innowise

Velocity, 100 миллисекунд - это...

В 3-4 раза быстрее, чем вы моргаете
За это время рекламный аукцион успевает пройти, и победитель — показать вам свой баннер.

За это же время система **Fraud Detection** в банке должна проанализировать вашу транзакцию, сравнить ее с тысячами предыдущих и принять решение — одобрить или отклонить платеж.

innowise

Variety, разнообразия данных — это...

Одна покупка в интернет-магазине — это

- **Запись в базе данных** (ID транзакции, сумма, время).
- **Поток JSON-событий** (каждый ваш клик и скролл на сайте).
- **JPEG-изображение** самого товара.

Один трек в Spotify — это

- **MP3/OGG** аудиофайл
- **JPEG-изображение** обложки альбома.
- **Текстовые метаданные** (имя артиста, название альбома, год).
- **Числовые данные** (темп, громкость, длительность).

innowise

Veracity, можно ли доверять этим данным?

Плохое качество данных обходится компаниям в среднем в **$15 миллионов** в год из-за неверных решений, упущенных возможностей и прямых потерь.

Аналитики и Data Scientist'ы тратят **до 80%** своего рабочего времени не на анализ, а на очистку и подготовку "грязных" данных, прежде чем их можно будет использовать.

innowise

Value: Как данные создают ценность

Проект **Геном человека** занял **13 лет**. Сегодня, благодаря **Big Data**, секвенирование одного генома занимает **часы**, что позволяет врачам подбирать **персонализированное лечение от рака**.

Логистическая компания UPS ежедневно анализирует **петабайты** данных о маршрутах, погоде и трафике. Их система **ORION** строит оптимальный маршрут для каждого из десятков тысяч водителей, что позволяет компании экономить около **$400 миллионов в год**.

innowise

Где используется Big Data?
Буквально везде!

Healthcare
Прогностическая диагностика, мониторинг состояния пациентов, разработка лекарств.`

Finance
Выявление мошенничества, алгоритмическая торговля, кредитный скоринг.

Retail & eCommerce
Сегментация клиентов, системы рекомендаций, прогнозирование запасов.

Media & Entertainment
Персонализированный контент (например, Netflix), аналитика аудитории.

Manufacturing & IoT
Профилактическое техобслуживание, аналитика датчиков, оптимизация процессов.

Transportation & Logistics
Оптимизация маршрутов, отслеживание автопарка, автономные системы.

Smart Cities
Анализ транспортных потоков, прогнозирование энергопотребления, аналитика в области общественной безопасности.

Energy & Utilities
Оптимизация умных сетей, прогнозирование потребления, прогнозирование отказов оборудования.

innowise

Как мы справляемся с этим хаосом?

Три фундаментальных принципа

1

Распределенное хранение

Данные делятся на части и хранятся на множестве серверов, что обеспечивает отказоустойчивость и практически бесконечную масштабируемость.

- Шардирование
- Партиционирование
- Отказоустойчивость

2

Распределенные вычисления

Задача по обработке данных разбивается на части и выполняется параллельно на множестве серверов (кластере), что многократно ускоряет получение результата.

- Кластер
- Массово-параллельная обработка (MPP)

3

Моделирование данных

Это проектирование структуры данных для нужд бизнеса. Сырые данные преобразуются через несколько логических слоев, чтобы в итоге стать удобными для анализа "витринами данных" (Data Marts).

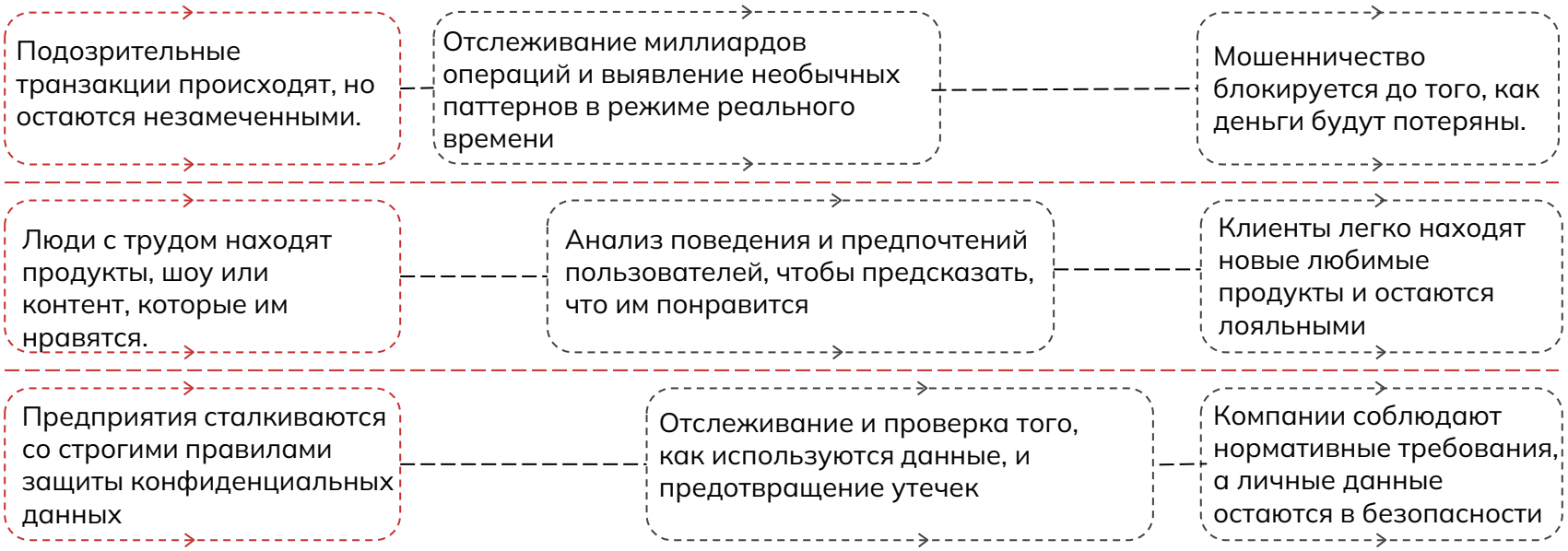
- Бизнес-процессы
- Витрины данных
- Трансформация

innowise

Реальное влияние, каждый день

Большие данные помогают решать реальные проблемы, которые затрагивают миллионы людей:

Если ты любишь решать головоломки и хочешь, чтобы твоя работа приносила реальную пользу, то Big Data — это то, что тебе нужно.



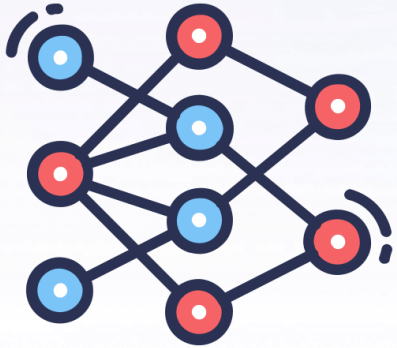
innowise

Три ключевые роли в индустрии данных

Data Analyst

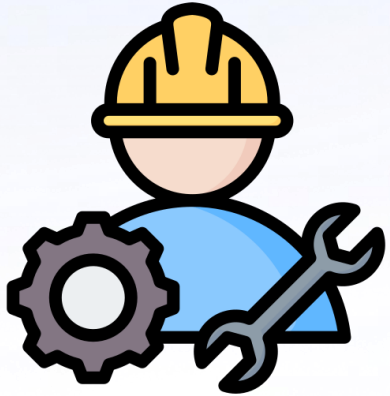
"Что произошло и почему?"

Data Scientist



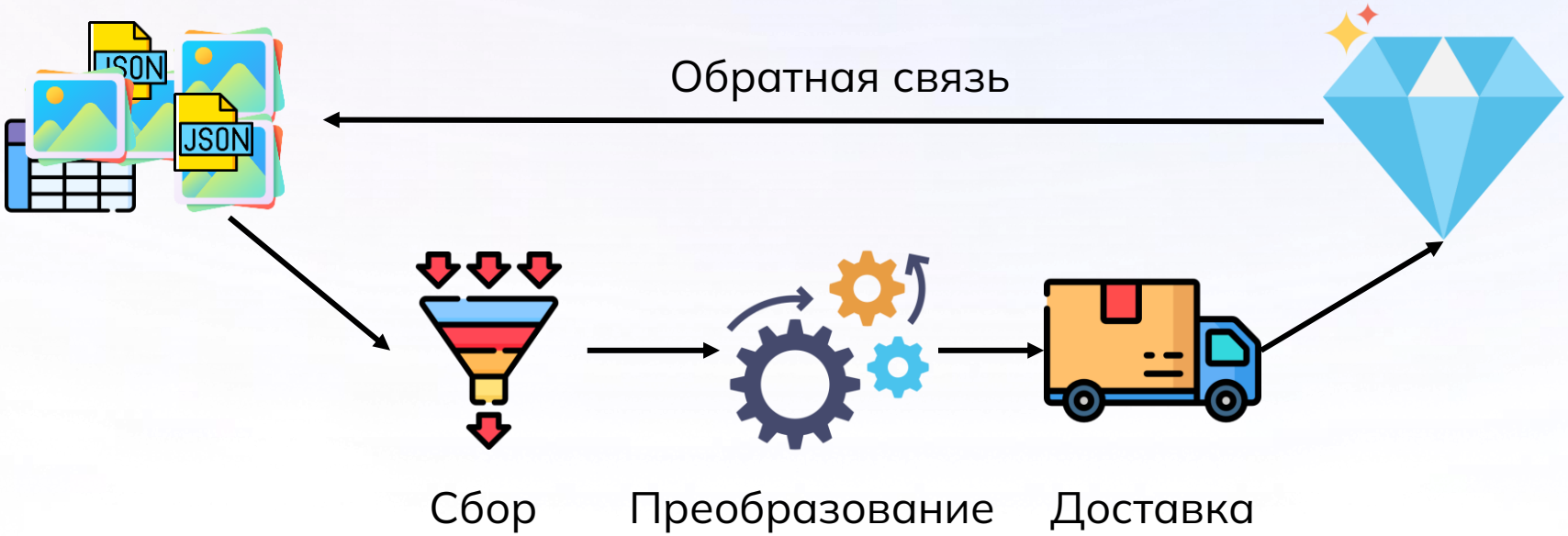
"Что может произойти?"

Data Engineer



"Как заставить всё это работать?"

От Хаоса к Порядку: Непрерывный цикл



innowise

Проектируем вместе: плейлист 'Открытия недели'

Проектируем вместе: плейлист 'Открытия недели'

Проектируем вместе: плейлист 'Открытия недели'

Лайки, время прослушивания, пропуски

Проектируем вместе: плейлист 'Открытия недели'

Лайки, время прослушивания, пропуски

Сбор

Проектируем вместе: плейлист 'Открытия недели'

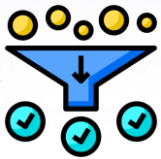
Лайки, время прослушивания, пропуски

Сбор

Очистка

Проектируем вместе: плейлист 'Открытия недели'

Лайки, время прослушивания, пропуски



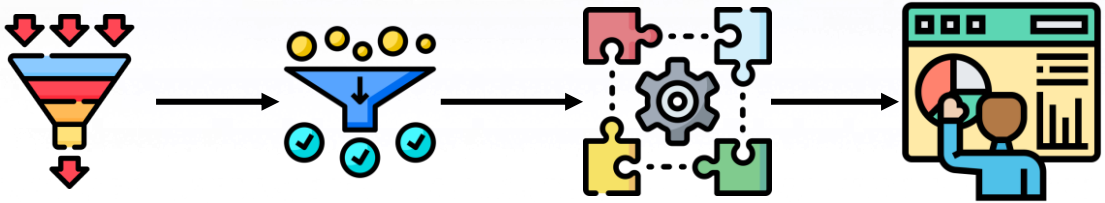
Сбор

Очистка

Обогащение

Проектируем вместе: плейлист 'Открытия недели'

Лайки, время прослушивания, пропуски



Сбор | Очистка | Обогащение | Витрина данных

Проектируем вместе: плейлист 'Открытия недели'

Лайки, время прослушивания, пропуски

Сбор

Очистка

Обогащение

Витрина данных

Проектируем вместе: плейлист 'Открытия недели'

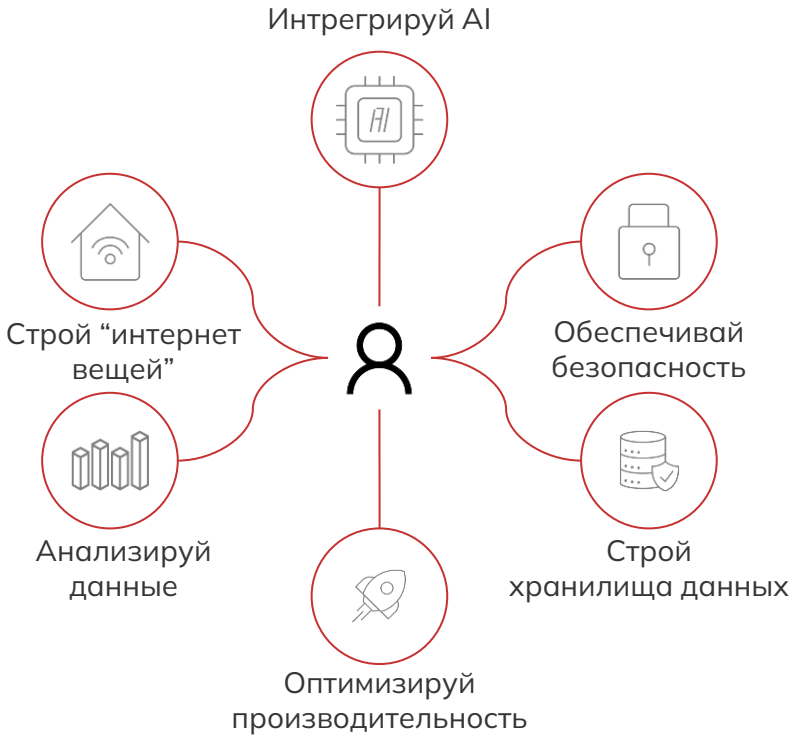
innowise

“Но где я могу все это освоить?”

Здесь мы собрали документ со всеми темами и ресурсами, которые понадобятся, чтобы стать хорошим стажером в Big Data.

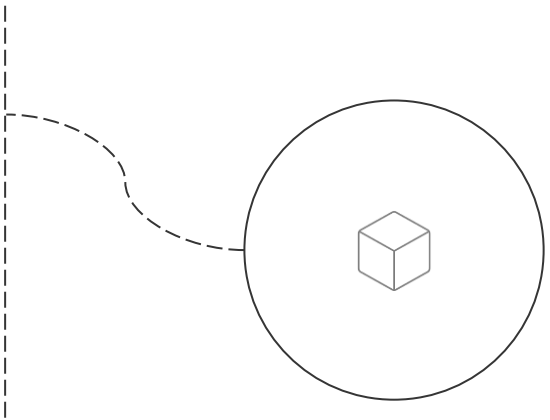
innowise

Профессия будущего



Специалисты по большим данным всегда будут востребованы, потому что мир с каждой секундой генерирует все больше и больше данных.

Есть так много всего, что можно исследовать - от технологий до отраслей - и нет двух одинаковых путей развития инженеров по большим данным. Это область, в которой можно продолжать расти в течение многих лет.



Формируй свой личный путь

innowise

Карьерные пути, которыми ты можешь пойти

Сформируй собственный путь роста и развития, и знай, что мы будем поддерживать тебя на каждом этапе этого пути.

Спасибо за внимание!

Innowise