

Полная программа

Data Warehouse Analyst

Длительность курса: 126 часов

Модуль 1. ELT: Структура и типы источников данных

Тема 1

Аналитические движки (СУБД) для работы с данными

Цель занятия

рассмотреть принципы работы аналитических СУБД, привести примеры конкретных СУБД, проанализировать сходства и различия.

Краткое содержание

МРР-базы данных и shared-nothing архитектура; колоночное хранение данных и компрессия; сегментация и партиционирование; особенности нагрузки на аналитические СУБД.

Тема 2

Источники данных: классификация и особенности

Цель занятия

рассмотреть особенности различных источников данных: подключение, формат, типы данных, ограничения; классифицировать источники на конкретные реализации: PostgreSQL, S3, Yandex.Metrica, REST API (Exchange rates).

Краткое содержание

структурированные и неструктурированные данные; форматы данных: CSV, JSON, AVRO, PARQUET, ORC; чтение из базы напрямую / logs WAL / REST.

Тема 3

Инструментарий разработки: IDE, Terminal, Docker, Codespaces, Terraform

Цель занятия

организовать удобную современную среду разработки; научиться работы с облачными провайдерами и инфраструктурой; продемонстрировать полученные знания на практическом примере.

Краткое содержание

IDE: VS Code, Command palette, keyboard shortcut extensions; Terminal: Z-shell, shell commands basics; Version control basics (git), Docker essentials; Yandex.Cloud CLI, Terraform, Github Actions, Github Codespaces.

Тема 4

Инструменты для выгрузки данных

Цель занятия

обсудить различия в подходах ETL и ELT; оценить выбор решения с точки зрения критериев: стоимость, сопровождение, развитие, масштабируемость.

Краткое содержание

трансформация из ETL в ELT; обзор Self-managed решений: Nifi, StreamSets; обзор SaaS-решений: Fivetran, Stitch, Hevo.

Домашние задания

Построение простейших пайплайнов переливки данных с помощью DIT

Цель

Реализовать любой из представленных двух вариантов практической работы, поработать с ETL инструментами по переливке данных, отработать навыки развёртывания инструментов и взаимодействия с облаком яндекса и докером. Подготовиться к выполнению более сложных домашних работ в дальнейшем

Модуль 2. DWH Basics

Тема 1

Принципы построения DWH

Цель занятия

сформулировать основные концепции в построении Хранилищ Данных; проследить эволюцию взглядов и концепций на построение DWH.

Краткое содержание

разделение на логические слои: Stage + Intermediate + Detail + Marts + Ad Hoc; Normalization: 3NF, Denormalized, Data Vault, Anchor; тесты данных и качеством данных; Team work & CI; макросы и функции + Maintenance; Security, Access Segregation, WLM.

Домашние задания

Построение архитектуры хранилища данных

Тема 2

Знакомство с Data Build Tool

Цель занятия

познакомиться с Data Build Tool – мультитул для работы с DWH; рассмотреть основные возможности и принципы dbt.

Краткое содержание

Dbt building blocks and principles; Connecting to DWH: profiles yam; Configuration: dbt_project.yam; Launching first project.

Тема 3

DWH powered by Clickhouse and dbt

Цель занятия

рассмотреть направление Analytics Engineering сегодня и место dbt в нем; объяснить, как инструменты подобные dbt могут помочь инженерам и аналитикам.

Краткое содержание

Analytics Engineering; Building complex Data Marts; SQL best practices: Complex SQL transformations + CTE; Analytical functions; Macros + Jinja templates; Code compilation + debugging; Documenting your project; Accessing documentation easily with static website.

Домашние задания

Конфигурирование и запуск проекта dbt

Цель

В этом ДЗ вы: - установите dbt, познакомитесь с cli; - конфигурируете проект и подключите к СУБД; - запустите расчет графа витрин и тестов; - сформируете и рассмотрите веб-сайт с документацией.

Тема 4

Q&A. Сессия вопросов и ответов

Цель занятия

получить ответы на вопросы по ДЗ; получить ответы на вопросы по приложениям.

Краткое содержание

типичные ошибки при выполнении ДЗ; наставники ответят на ваши вопросы.

Модуль 3. DWH Intermediate

Тема 1

Введение к оркестрации

Цель занятия

обсудить, когда нужны инструменты оркестрации; рассмотреть рынок современных решений.

Краткое содержание

когда простого stop становится недостаточным; обзор решений Airflow, Prefect, Dagster; принципы работы DAGs.

Домашние задания

Установка apache-airflow на локальный компьютер

Цель

Установить apache-airflow на локальную машину в контейнере docker

Тема 2

Оркестрация с Apache Airflow

Цель занятия

рассмотреть опции развёртывания и сопровождения решений оркестрации; погрузиться в оптимизацию и конфигурацию DAGs.

Краткое содержание

Deployment: Self-managed (пример на Kubernetes) vs Cloud native; Writing dynamic DAGs; Dependency management; Monitoring & Alerting.

Домашние задания

Подготовка и установка на расписание DAG выгрузки данных из источников

Цель

В данном ДЗ мы настроим автоматический data pipeline, который будет получать данные из публичного API и складывать их в БД для дальнейшего анализа.
Конечный продукт:
1) работающий облачный инстанс Apache Airflow;
2) data pipeline, содержащий в себе несколько task-ов и "крутящийся" по расписанию Airflow;
3) работающий облачный инстанс СУБД, куда Airflow заливает данные, получаемые из внешнего API;
4) данные в СУБД.

Тема 3

Оркестрация Dagster и DBT

Цель занятия

получить ответы на вопросы по ДЗ; получить ответы на вопросы по приложениям.

Краткое содержание

типичные ошибки при выполнении ДЗ; наставники ответят на ваши вопросы.

Тема 4

Data Quality

Цель занятия

получить представление о том, что такое качество данных; выяснить как Data Quality влияет на выводы и принимаемые решения; рассмотреть стратегии управления качеством данных.

Краткое содержание

основные метрики качества данных; причины нарушения качества и стратегии реагирования; измерение, мониторинг, исправление; демонстрация: schema, data, freshness tests в DBT; Continuous Integration tests; кросс-проверки источник <-> DWH.

Тема 5

Вопросы оптимизации производительности

Цель занятия

получить представление об источниках проблем с производительностью; рассмотреть лучшие практики в оптимизации производительности.

Краткое содержание

Performance best practices; Execution plan analysis; Compressing data & physical design (DIST, SORT, Materialized views, ...); Incremental updates / building marts by periods; Code refactoring & KISS (Keep it simple, stupid).

Тема 6

Data Vault – 1

Цель занятия

погрузиться в подход к организации детального слоя Data Vault 2.0; рассмотреть пример построения DWH на DV 2.0.

Краткое содержание

формулирование требования к модели DWH; освежение знания о моделировании DWH; нормализация и подход Data Vault 2.0; Building blocks: HUB + LINK + SATELLITE.

Тема 7

Data Vault – 2

Цель занятия

получить представление об архитектурных паттернах и Business Vault; рассмотреть основы автоматизации и генерации кода Data Vault.

Краткое содержание

архитектурные паттерны и Business Vault; оптимизация физической модели; основы кодогенерации и dbtVault.

Домашние задания

Организация детального слоя DWH по методологии Data Vault

Цель

В этом ДЗ вы: - рассмотрите концепции Data Vault и строительных блоков: Hub, Link, Satellite; - сформируете кодогенерацию логической модели данных (ЛМД); - сформируете витрину данных из Data Vault.

Тема 8

Q&A. Сессия вопросов и ответов

Цель занятия

получить ответы на вопросы по ДЗ; получить ответы на вопросы по приложениям.

Краткое содержание

типичные ошибки при выполнении ДЗ; наставники ответят на ваши вопросы.

Модуль 4. Business Intelligence

Тема 1

BI: Обзор

Цель занятия

сформулировать назначение систем класса BI; рассмотреть принципы работы BI-инструментов и решаемые задачи.

Краткое содержание

анализ и сравнение функционала решений; BI building blocks: connecting, modeling, visualising, dashboarding; обзор популярных BI-решений: Looker, PowerBI, Tableau; Open source BI: Superset, Metabase.

Тема 2

BI подготовка данных (AirFlow + pipeline)

Тема 3

BI: Deployment

Цель занятия

рассмотреть опции развёртывания BI-решения; погрузиться в вопросы конфигурации развёртывания BI-решения.

Краткое содержание

Self-hosted vs. Managed; Apache Superset: Docker deployment; Metabase: Deployment with Docker on AWS Elastic Beanstalk; Configuring BI tool: security, metadata, notifications, user access; Software version upgrades.

Тема 4

BI: Modeling & Delivering

Цель занятия

научиться подключаться к источникам данных для BI; создавать метрики, сегменты, фильтры, дашборды для визуальной аналитики.

Краткое содержание

Connecting to data sources; задание метрик, фильтров, сегментов; подготовка визуализаций для представления выводов; сборка аналитических дашбордов: лучшие практики.

Домашние задания

Конфигурация и развёртывание BI-решения

Цель

В этом ДЗ вы научитесь конфигурировать и развёртывать BI-решения.

Тема 5

Разбор ДЗ – Организация детального слоя DWH по методологии Data Vault

Цель занятия

получить ответы на вопросы по ДЗ; получить ответы на вопросы по приложениям.

Краткое содержание

типичные ошибки при выполнении ДЗ; наставники ответят на ваши вопросы.

Тема 6

Analytics: Базовые аналитические витрины

Цель занятия

получить представления о типах аналитических витрин и их особенности; рассмотреть практики применения аналитики для поиска ответов на бизнес-проблемы.

Краткое содержание

сегментация – Segments; ключевые показатели и метрики – KPI; анализ временных рядов – Timeseries analytics + Period-by-period; когортный анализ – Cohort analysis.

Тема 7

Analytics: Сквозная аналитика

Цель занятия

познакомиться с организацией Сквозной Аналитики в маркетинге; сделать первые шаги в построении собственного решения на практике.

Краткое содержание

требования бизнеса и ожидаемые результаты; эволюция подходов, используемых инструментов, практик; рейтинг проблем и узких мест; знакомство с датасетом и постановкой домашнего задания.

Домашние задания

Сквозная аналитика – Performance Marketing Analytics

Цель

В этом ДЗ вы получите понимание задач и результатов Сквозной аналитики в Маркетинге.

Тема 8

Разбор ДЗ – Конфигурация и развёртывание BI-решения

Цель занятия

получить ответы на вопросы по ДЗ; получить ответы на вопросы по приложениям.

Краткое содержание

типичные ошибки при выполнении ДЗ; наставники ответят на ваши вопросы.

Тема 9

Analytics: Продвинутые аналитические витрины

Цель занятия

получить представления о продвинутых аналитических витринах; рассмотреть практики применения аналитики для поиска ответов на бизнес-проблемы.

Краткое содержание

разбиение событий на сессии – Sessionization; построение воронок и расчет конверсий – Funnels and conversions; привлечение, вовлечение, удержание – Acquisition, Engagement, retention analysis; recency, Frequency, Monetary Value – RFM analysis.

Модуль 5. DWH Advanced topics

Тема 1

DWH: Advanced topics

Цель занятия

обсудить продвинутые функциональные возможности Хранилищ Данных; разобраться в основных трендах и развивающихся фишках.

Краткое содержание

DWH: Extending with UDF; Complex analytics SQL: Geospatial + Sessionizing + Pattern Matching; DBT: Advanced macros + Jinja; Enabling Slim CI; CI and deployment with Github Actions.

Домашние задания

Advanced DWH: Configuring CI, dbt modules, External tables

Цель

В этом ДЗ вы научитесь настраивать собственный модуль.

Тема 2

DBT: Extending with modules

Цель занятия

научиться работать с модулями dbt – импорт, версионирование, использование; разобраться с написанием собственного модуля или расширением существующего.

Краткое содержание

Importing modules (libraries); Overview of modules: dbt_utils, calendar, logging; Creating your own module; Testing your newly created module.

Тема 3

Разбор кейса: end-to-end solution

Цель занятия

повторить все пройденные материалы курса; разобрать несколько кейсов применения полученных знаний в решении бизнес-проблем.

Краткое содержание

Put everything in place, собираем воедино все части; где могут возникнуть проблемы и как их решить; разбор реальных кейсов компаний; коммуникация – понимание, чего хочет заказчик, и делаем чуть больше; поставка дата-сервисов и результатов – Deliver results.

Тема 4

Дальнейшее развитие навыков

Цель занятия

получить обзорную картину полезных навыков и умений, сферы их применения; пополнить багаж полезных знаний для совершенствования.

Краткое содержание

развитие Твёрдых (Hard) навыков: 20 часов практики на навык, увеличить производительность работы, ресурсы; развитие Мягких (Soft) навыков: правила подготовки CV, прохождения интервью, общение с ментором, карьерное развитие.

Домашние задания

Soft skills checklist: CV, LinkedIn, достижимые цели обучения

Цель

В этом ДЗ вы научитесь делать резюме, создавать личные страницы в LinkedIn.

Модуль 6. Проектная работа

Тема 1

Выбор темы и организация проектной работы

Цель занятия

выбрать и обсудить тему проектной работы; организовать работу над проектом; ознакомиться с регламентом работы над проектом.

Краткое содержание

правила работы над проектом и специфика проведения итоговой защиты; требования к результату проекта и итоговой документации.

Домашние задания

Проектная работа

Цель

В этом ДЗ необходимо выбрать и утвердить в чате по ДЗ тему проекта, разработать и презентовать проект.

Тема 2

Консультация по проектам и ДЗ

Тема 3

Предзащита

Тема 4

Защита проектных работ

Цель занятия

защитить проект и получить рекомендации экспертов; узнать, как получить сертификат об окончании курса, как взаимодействовать после окончания курса с OTUS и преподавателями, какие вакансии и позиции есть для выпускников (опционально – в России и за рубежом)