Data Warehouse Analyst

Длительность курса: 124 академических часа

1 ELT: Структура и типы источников данных

1 Источники данных: классификация и особенности

Цели занятия:

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

Краткое содержание:

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

---

2 Инструменты для выгрузки данных

Цели занятия:

научиться делать выгрузки своими силами: RDBMS, S3/HDFS, REST API, Webhooks;
найти баланс в использовании внешних сервисов и написании своих pipelines.

Краткое содержание:

организация регулярных выгрузок с помощью облачных сервисов GCP / AWS;
репликация базы данных при помощи Debezium и чтения бинарного лога.

Домашние задания

1 Выгрузка данных веб-счетчика

Цель: В этом ДЗ вы:
- изучите документации API сервиса (Яндекс.Метрика)
- программно сформируете запросы и ответы от сервиса

Подробная инструкция к ДЗ:
https://docs.google.com/document/d/1lUCzK9r9h1r9pj

OpGzRqL0iWJ_HCuUUB_utTKGEPCL8/edit?usp=sharing

Необходимо:

1. Изучить визуальный интерфейс Яндекс.Метрика Live Demo: https://metrika.yandex.ru/dashboard?group=day&period=week&id=44147844

2. Сформировать несколько запросов на выгрузку данных за последний месяц, аналогично веб-интерфейсу:
- Traffic: https://metrika.yandex.ru/stat/traffic?period=week&accuracy=1&id=44147844&stateHash=5f60b5fbd470f0000ed40699
- Conversions: https://metrika.yandex.ru/stat/conversion_rate?period=week&accuracy=1&id=44147844
- Sources, summary: https://metrika.yandex.ru/stat/sources?metric=ym:s:pageDepth&sort=-ym:s:pageDepth&chart_type=pie&period=week&attribution=Last&accuracy=1&id=44147844&stateHash=60fea065e6a4b9001463439b

Использовать curl или любой язык программирования

3. Сохранить результаты запросов в хранилище Yandex Object Storage (S3).
Делать обращения из Virtual Machine / Cloud Function в Яндекс.Облаке."

---

3 **Инструменты для выгрузки данных**

Цели 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.

2 DWH Basics

1 Принципы построения DWH

Цели 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 качество данных;
Team work & CI;
макросы и функции + Maintenance;
Security, Access Segregation, WLM.

---

2 Аналитические движки (СУБД) для работы с данными

Цели занятия:

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

Краткое содержание:

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

---

3 **Знакомство с Data Build Tool**

Цели 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.yaml;
Configuration: dbt_project.yaml;
Launching first project.

---

4 **DBT: Analytics Engineering**

Цели 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

1 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

https://gist.github.com/kzzzr/8d50126079df1a8e5646342f6247df22

На проверку прислать ссылку на свой репоизторий с проектом dbt, в README.md вложить скриншоты результатов запросов из задания Q2.1, Q3.3, Q4.2

Push Git repo with dbt project to Github:

external tables to S3
sources.yml
base tabes
wide table
tests and docs for models

https://clickhouse.tech/docs/en/getting-started/example-datasets/star-schema/

Send results of queries:

Q2.1
Q3.3
Q4.2

---

5 **Разбор ДЗ – Выгрузка данных веб-счетчика**

Цели 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.

3 DWH Intermediate

1 Оркестрация скриптов и задач – 1

Цели занятия:

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

Краткое содержание:

когда простого cron становится недостаточно;
обзор решений Airflow, Prefect, Dagster;
принципы работы DAGs.

---

2 Оркестрация скриптов и задач – 2

Цели занятия:

рассмотреть опции развертывания 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

1 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

Часть из операций мы разбирали на занятии. При необходимости - можно пересмотреть запись.

1) Создаем Виртуальную машину с Apache Airflow 2.0 в YandexCloud

2) Создаем "managed instance" PostgreSQL/ClickHouse/MySQL - по выбору в YandexCloud.
Создаем БД "analytics".

3) Добавляем наш psql в Connections через UI Airflow.

4) Выбираем один из 2-х API, с которым будем работать:
- Положение Международной Космической Станции на текущий момент времени (timestamp-latitude-longitude). Source: http://api.open-notify.org/iss-now.json
- Курс BTC: https://docs.coincap.io/#2a87f3d4-f61f-42d3-97e0-3a9afa41c73b
Тут нас интересует следующий endpoint: "api.coincap.io/v2/rates/bitcoin"

5) Создаем схему данных (таблицу в бд analytics) с названиями и типами полей, релевантными тому, что будем забирать из API.

6) Описываем DAG (программируем на Python) для получения данных с периодичностью 30 min.
Сам оператор для обращения к API можно выбрать любой.
Для простоты рекомендуется слать GET-запросы через python-библиотеку requests (`PythonOperator`), либо через bash (`BashOperator`) с помощью curl.

Концептуальная схема Dag-а: отправить запрос в API->распарсить пришедший результат->положить данные в БД (сделать

insert в таблицу)

7) Кладем .py файл с DAG-ом в нужную директорию виртуалки с airflow.

8) Запускаем DAG тумблером в UI Airflow

9) Отлаживаем DAG до работоспособного состояния.

В интерфейсе Airflow (облачный инстанс) есть информация о наборе успешно завершившихся Dag runs (темно-зеленые кружки) + в БД есть данные за >5 периодов времени.

В качестве проверки мы зайдем в вам airflow webserver и отправим sql-запрос в таблицу с данными.

10) Проверяем, что данные появились в БД.

11) Поздравляю, Вы завершили ДЗ!

---

3 **Data Quality**

Цели 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.

---

4 **Вопросы оптимизации производительности**

Цели занятия:

получить представление об источниках проблем с производительности
изучить 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).

---

5 **Разбор ДЗ – Конфигурирование и запуск проекта dbt**

Цели 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.

---

6 **Data Vault – 1**

Цели 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;
освежаем знания о моделировании DWH;
нормализация и подход Data Vault 2.0;
Building blocks: HUB + LINK + SATELLITE.

7	**Airflow 3**	
8	**Разбор ДЗ – Подготовка и установка на расписание DAG выгрузки данных из источников**	**Цели 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.

9 **Data Vault – 2**

Цели 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

1 Организация детального слоя DWH по методологии Data Vault

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

Вариант ДЗ №1 (легче):
https://docs.google.com/document/d/1t_P0Cww9MgHYeGkIC6p-V4ZXddFaXj31_PrE1vLKPdU/edit?usp=sharing

Вариант ДЗ №2 (сложнее):

1. Fork demo repo:
https://github.com/kzzzr/dbtvault_greenplum_demo
2. Add README.md with answers to the following questions:
https://gist.github.com/kzzzr/4ab36bec6897e48e44e792dc2e706de9

10 **Data Vault – 3**

Цели 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.

4 Business Intelligence

1 BI: Обзор

Цели занятия:

сформулировать назначение систем класса BI;
рассмотреть принципы работы BI-инструментов и решаемые задачи7

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.

---

2 BI: Deployment

Цели занятия:

рассмотреть опции развертывания BI-решений;
погрузиться в вопросы конфигурация развертывания 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.

---

3 **BI: Modeling & Delivering**

Цели занятия:

научиться подключаться к источникам данных для BI;
создать 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

1 Конфигурация и развертывание BI-решения

Цель: В этом ДЗ вы научитесь конфигурировать и развертывать BI-решения.

Необходимо:
- развернуть и конфигурировать BI-решения
- настроить подключение к источникам данных
- рассмотреть интерактивный визуальный анализ данных

---

4 **Разбор ДЗ – Организация детального слоя DWH по методологии Data Vault**

Цели 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.

5	**Analytics: Базовые аналитические витрины**	**Цели занятия:** получить представления о типах аналитических витрин и их особенностях; 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.

6 **Analytics: Сквозная аналитика**

Цели 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

1 Сквозная аналитика – Performance Marketing Analytics

Цель: Получить понимание задач и результатов Сквозной аналитики в Маркетинге;
Поключиться к СУБД с исходными данными (read-only) и провести разведочный анализ (EDA);
Реплицировать данные в целевую базу для моделирования (на выбор Postgres /Clickhouse / Greenplum);
Смоделировать витрины данных с заданным набором измерений и метрик;
Визуализировать данные на графиках и дашбордах (BI-инструмент на выбор).

https://gist.github.com/kzzzr/3f9b167291e32544004fa0ed75d1f4b2

7	**Разбор ДЗ – Конфигурация и развертывание BI-решения**	**Цели 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.
8	**Analytics: Продвинутые аналитические витрины**	**Цели 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.

5 DWH Advanced topics

1 **DWH: Advanced topics**

Цели занятия:

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

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.

---

2 **DBT: Extending with modules**

Цели занятия:

научиться работать с модулями dbt – импорт, версионирование, использование;
разобраться с написание 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.

3	**DWH: Monitoring + Workload management**	**Цели занятия:** научиться измерять и отслеживать основные метрики DWH; визуализировать метрики Хранилища Данных для анализа. **Краткое содержание:** Database performance metrics: CPU, RAM, Disk, Network, Errors and restarts; DWH deployement metrics: frequency, time to build, bottlenecks, problem models; Logging dbt runs: 1. via pre- & post-hooks (logging module) 2. via parsing dbt artifacts; Visualizing and analysing metrics on a dashboard; Segregating access and resources to users with roles, groups and WLM.
4	**Разбор ДЗ – Визуализация и дашбординг для аналитических витрин**	**Цели 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.

5 **DWH: External + Semi-structured data**

Цели занятия:

рассмотреть принципы работы с внешними для Хранилища Данными;
познакомиться с инструментами и возможностями обращения к External Data.

Краткое содержание:

S3 based Data Lake;
Accessing Semi-structured data;
External Data: Parquet, ORC;
Hive, Presto, Athena;
Dbt module: external data.

Домашние задания

1 Advanced DWH: Configuring CI, dbt modules, External tables

Цель: В этом ДЗ вы научитесь настраивать собственный модуль.

Необходимо:
- разработать собственный модуль dbt
- настроить Continuous Integration & Testing
- обратиться к внешним данными (External Tables in S3)

---

6 **DWH: Reverse-ETL**

Цели занятия:

получить представление о процессах выгрузки данных из DWH;
рассмотреть возможные сценарии использования reverse-ETL.

Краткое содержание:

push data from DWH to Operational systems;
обзор фреймворков Grouparoo, Census;
что выгружаем: LTV, customer labels, scoring;
примеры и демонстрация: Braze, Hubspot, Mailchimp.

---

7 **DWH: Machine Learning capabilities**

Цели занятия:

выяснить место DWH по отношению к Data Science и Machine Learning;
обсудить какими возможностями в области ML обладают современные аналитические движки7

Краткое содержание:

Feature engineering with dbt;
In-database ML (BigQuery, Vertica);
Redshift + SageMaker.

6 Recap

1 **Разбор кейса: end-to-end solution**

Цели 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 - понимаем, чего хочет заказчик и делаем чуть больше;
поставка дата-сервисов и результатов – Deliver results.

---

2 **Разбор ДЗ – Advanced DWH: Configuring CI, dbt modules, External tables**

Цели занятия:

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

Краткое содержание:

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

---

3 **Дальнейшее развитие навыков**

Цели занятия:

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

Краткое содержание:

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

Домашние задания

1 Soft skills checklist: CV, LinkedIn, достигнутые цели обучения

Цель: В этом ДЗ вы научитесь делать ревью, создавать личные страницы в LinkedIn.

Необходимо:
- провести ретроспективный анализ полученных навыков
- подготовить и сделаете ревью CV
- создать личные странички в LinkedIn и расширить сети контактов

7 Проектная работа

1 Выбор темы и организация проектной работы

Цели занятия:

выбрать и обсудить тему проектной работы;
спланировать 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

1 Проектная работа

Цель: В этом дз необходимо выбрать и утвердите в чате по ДЗ темы проекта, разработать и презентовать проект.

Необходимо:
- Definition of Done
- Сформулировать идею аналитического приложения, business value
- Представить архитектуру ресурсов и компоненты системы
- Классифицировать источники данных, регулярность загрузки, форматы данных + структуру
- Сформировать Хранилище Данных
- Развернуть BI-инструмент
- Представить данные в наглядном и понятном виде на дашбордах
- Сделать несколько интересных выводов
- Опубликовать результаты работы на Хабр/…, выложить код на Github

2	**Защита проектных работ**	**Цели занятия:** защитить проект и получить рекомендации экспертов. **Краткое содержание:** презентация проектов перед комиссией; вопросы и комментарии по проектам.