# **Практическая работа (экзамен)**

|  |  |
| --- | --- |
| Дисциплина | ETL-процессы |
| Тема | Полная реализация ETL-процесса |
| Форма проверки | **Задание с индивидуальной проверкой преподавателем.** |
| Имя преподавателя | Артём Озерков |
| Время выполнения | 10 часов |
| Цель задания | * Освоить работу с Yandex DataTransfer * Научиться выполнять автоматизацию работы с Yandex Data Processing (Hadoop + Spark) * Освоить работу с топиками Apache Kafka с помощью PySpark-заданий (Yandex Data Processing) * Построить дашборды в DataLens для визуализации результатов |
| Инструменты для выполнения ДЗ | * Облачные сервисы: Yandex Cloud (Data Processing, DataLens, Managed Service for Apache Airflow, Managed Service for Apache Kafka) * Базы данных: Hive/Spark SQL, YDB * Язык запросов: SQL (HiveQL, Spark SQL, YQL) * Инструменты разработки: GitHub (для хранения SQL-скриптов), Yandex Object Storage * Оркестрация: Apache Airflow |
| Правила приёма работы | Прикрепить в ЛМС ссылки на репозиторий GitHub и опубликованные образы Docker Hub. Настройки репозитория должны быть public. |
| Критерии оценки | **Выполнение задания 1** — 2 балла  **Выполнение** **задания 2** — 3 балла  **Выполнение задания 3** — 3 балла  **Выполнение** **задания 4** (дополнительное) — до 2-х баллов (зависит от полноты решения и качества реализации) |
| Дедлайн | 19.06.2025 |

## **Описание задания**

## В существующую аналитическую платформу, которую вы создали, выполняя проверяемое задания по дисциплине «Семинар наставника», требуется добавить новые источники для улучшения точности аналитики.

## **Задание 1: Работа с Yandex DataTransfer**

[*Ссылка на документацию*](https://yandex.cloud/ru/docs/data-transfer/tutorials/ydb-to-object-storage)

Требуется перенести данные из Managed Service for YDB в объектное хранилище Object Storage. Выполнить необходимо с использованием сервиса Data Transfer*.*

1. **Создать БД Yandex DataBase**.
2. **Подготовить данные**:  
   * transactions\_v2

Данные можно взять [здесь](https://www.kaggle.com/datasets) или подготовить их самостоятельно.

1. **Создать трансфер в Object Storage**
2. **Проверить работоспособность трансфера**

**Все SQL-скрипты** (YQL) необходимо сохранить в репозитории домашней работы (GitHub).

* Консультация по этой части работы на вебинаре 29 мая

## **Задание 2: Автоматизация работы с Yandex Data Processing при помощи Apache AirFlow**

[Ссылка на документацию](https://yandex.cloud/ru/docs/managed-airflow/tutorials/data-processing-automation)

Требуется обрабатывать файлы (parquet или CSV) из внешнего источника. Размер входящих файлов меняется в различные дни месяца.

Данные можно взять [здесь](https://www.kaggle.com/datasets) или подготовить их самостоятельно.

1. **Подготовить инфраструктуру**
2. **Подготовить PySpark-задание**
   * Создать кластер Yandex Data Processing
   * Создать и запустить задание PySpark
   * Удалить кластер Yandex Data Processing.
3. **Подготовить DAG-файл, запустить и проверить результат**

* Консультация по этой части работы на вебинаре 29 мая

# **Задание 3: Работа с топиками Apache Kafka® с помощью PySpark-заданий в Yandex Data Processing**

[*Ссылка на документацию*](https://yandex.cloud/ru/docs/managed-kafka/tutorials/data-processing)

Требуется настроить чтение топиков kafka для реализации потоковой аналитики.

1. **Подготовить архитектуру**
2. **Создать задания PySpark**

* Консультация по этой части работы на вебинаре 5 июня

По завершении выполнения трёх заданий подготовьте отчёт в свободной форме с описанием проделанных действий.

## **Задание 4 (дополнительное): Визуализация в DataLens**

**С помощью Yandex DataLens** построить **дашборды** для визуализации загруженных данных.

* Консультация по этой части работы на вебинаре 5 июня

**Обратите внимание:**

* Рекомендуем выполнять работу по частям к вебинарам с консультациями, чтобы иметь возможность задать вопросы преподавателю и прояснить непонятные детали/сложности задания
* Не забывайте останавливать сервисы в Яндекс облаке, чтобы вам хватило бесплатного ресурса на выполнение заданий