Разработка аналитических систем на платформах Big Data и IBM Cloud

номер	преподаватель	тема
1	Красников Антон	основы Linux
2	Красников Антон	Git
3	Ильина Ольга	Python - Основы
4	Ильина Ольга	Python - Углубленные концепции
5	Ильина Ольга	Python - Flask, Swagger, Oauth
6	Бурак Анна	Python - Pandas
7	Бурак Анна	Python - Unit tests
8	Жук Михаил	Docker
9	Жук Михаил	Kubernetes
10	Ильина Ольга	SQL - часть 1 (проектирование)
11	Ильина Ольга	SQL - часть 2 (базовые операции)
12	Ильина Ольга	SQL - часть 3 (продвинутые операции)
13	Бурак Анна	Stored Producedures and UDFs
14	Бурак Анна	Проектирование DWH/ETL
15	Бурак Анна	SQL - оптимизация запросов
16	Бурак Анна	IBM Cloud - Cloud Object Storage
17	Бурак Анна	IBM Cloud - DB2
18	Бурак Анна	IBM Cloud - CloudFoundry and CI/CD - часть 1
19	Бурак Анна	IBM Cloud - CloudFoundry and CI/CD - часть 2
20	Метлович Кристина	IBM Cloud - Functions
21	Жук Михаил	Hadoop (Overview, HDFS, MapReduce, Yarn)
22	Жук Михаил	Hive, Hbase
23	Жук Михаил	Kafka
24	Жук Михаил	NiFi
25	Жук Михаил	Airflow
26	Метлович Кристина	Spark
27	Метлович Кристина	Elasticsearch 1
28	Метлович Кристина	Elasticsearch 2
29	Метлович Кристина	Kibana
30	Бурак Анна	Введение в Scrum

Защита проектов	Через 1 неделю после окончания курсов

Задание:

Раработать решение для анализа открытых данных. В качестве источников могут выступать:
- текстовые данные на веб сайтах
- API социальных сетей и различных веб сайтов
- RSS ленты

Примеры:
- Анализ упоминания климатических тем на новостных сайтах
- Анализ результативности футболистов
- Анализ изменения количества подписчиков/лайков на популярных каналах в соц сетях

Необходимо придумать свою тему.

Техническое описание:

Обязательно:

Разработать Python Flask микросервис, который по запросу вытягивает требуемые данные из открытых источников (web scrapping, rss feeds, APIs,…). Микросервис запустить в Docker контейнере. Возвращаемые данные должны быть в формате JSON со сложной структурой включающей вложенные элементы. Реализовать NiFi процесс, который по расписанию вызывает Python Flask микросервис и записывает полученные данные в Kafka Topic. Информацию о вызове микросервиса необходимо логировать в Elasticsearch.

Реализовать в Kafka парсинг JSON от микросервиса и запись в нормализованные таблицы в PostgreSQL. Реализовать Spark (Python or Scala) скрипт, который читает нормализованные данные из PostgreSQL, производит агрегации/расчет показателей и записывает данные в Star Schema в PostgreSQL. Все dimensions в SCD Type 1.
Реализовать запуск по расписанию Spark скрипта при помощи Airflow. Информацию о запуске Spark скрипта необходимо логировать в Elasticsearch.

Разработать Web приложение для визуализации данных в Star Schema. Визуализация должна включлчать в себя таблицы, фильтры, графики и прочие интерактивные компоненты. Backend - Python Flask, Frontend - React.js, D3.js либо другие js фреймворки. Приложение запустить в Docker контейнере. Информацию о каждом взаимодействии пользователя с приложением необходимо логировать в Elasticsearch.

	Реализовать в Kibana отчеты по статистике работы компонетов, которые логируются в Elasticsearch.
Важно:	**Код проекта необходимо загрузить в локальный Git репозиторий либо в любой приватный репозиторий. До завершения процедуры защиты нельзя выкладывать код проекта в публичных репозиториях.**
Желательно:	Рекомендуется использовать Docker для запуска всех компонентов решения: NiFi, Airflow, Kafka, Elasticsearch, PostgreSQL и т.д. Для хранения персистентных данных использовать Volumes. Однако в случае возникновения сложностей допускается локальная установка ПО кроме случаев, которые явным образом указаны в требованиях.
Усложнение:	Развернуть всё решение на Kubernetes вместо набора отдельных Docker контейнеров. Настроить CI/CD для web приложения при помощи любого инструмента на свой выбор.

	дата	время	
Python	16-июл-2021	19.00	пт
Python	22-июл-2021	19.00	чт
Python	27-июл-2021	9.00	вт
Python	29-июл-2021	9.00	чт
Python	3-авг-2021	9.00	вт
Python	6-авг-2021	9.00	пт
Python	10-авг-2021	9.00	вт
Python	12-авг-2021	9.00	чт
Python	17-авг-2021	9.00	вт
DB/SQL	19-авг-2021	9.00	чт
DB/SQL	24-авг-2021	9.00	вт
DB/SQL	26-авг-2021	9.00	чт
DB/SQL	31-авг-2021	9.00	вт
DB/SQL	3-сен-2021	9.00	пт
DB/SQL	7-сен-2021	9.00	вт
IBM Cloud	10-сен-2021	9.00	пт
IBM Cloud	14-сен-2021	9.00	вт
IBM Cloud	16-сен-2021	9.00	чт
IBM Cloud	17-сен-2021	9.00	пт
IBM Cloud	20-сен-2021	19.00	пн
Big Data	23-сен-2021	9.00	чт
Big Data	28-сен-2021	9.00	вт
Big Data	30-сен-2021	9.00	чт
Big Data	5-окт-2021	9.00	вт
Big Data	7-окт-2021	9.00	чт
Big Data	11-окт-2021	19.00	пн
Big Data	13-окт-2021	19.00	ср
Big Data	18-окт-2021	19.00	пн
Big Data	20-окт-2021	19.00	ср
	22-окт-2021	9.00	пт