**Домашнее задание**

| | |
|---|---|
| Дисциплина | Python для инженерии данных |
| Тема | Тема 12. Знакомство со Spark streaming |
| Форма проверки | **Самопроверка. Студент выполняет задание и самостоятельно проверяет его.** |
| Имя преподавателя | Дмитрий Клабуков |
| Время выполнения | 1 час |
| Цель задания | Научиться готовить потоковые данные для аналитиков. Отработать навыки обработки потоковых данных |
| Инструменты для выполнения ДЗ | Spark |
| Правила приема работы | Прикрепите ссылку в LMS на выполненное задание в Google colab или GitHub (если вы использовали Jupyter Notebook)<br>**Важно:** убедитесь в том, что по ссылке есть доступ в Google colab (иногда в колабе нет доступа для другого логина). |
| Критерии оценки | **Задание считается выполненным, если:**<br>- прикреплена ссылка на файл с выполненным заданием<br>- доступ к файлу открыт<br>- код дает правильный ответ к задаче<br><br>**Задание не выполнено, если:**<br>- файл с заданием не прикреплен или отсутствует доступ по ссылке<br>- код выдаёт ошибку или дает неправильный ответ |
| Дедлайн | **16.12.2024** |

**Задание**

1. Создайте топик с транзакционной активностью пользователей банковского приложения. Сгенерируйте данные для потока.

2. Создайте таблицу в любой базе данных с основной информацией о клиенте. Данные можно сгенерировать самостоятельно.

3. Соедините поток с транзакционной активностью с основной информацией. На основе полученных данных сделайте вывод, подозрительная транзакция или нет.

4. Выгрузите данные с подозрительной активностью в новый поток.

**Пример транзакционных данных:**

{"transaction_id": 1001, "user_id": 123, "amount": 1200.50, "timestamp": "2024-12-08T01:30:00", "location": "New York"}

**Пример данных с основной информацией о пользователе**

user_id | name | registration_address | last_known_location

123 | John Doe | New York | Los Angeles

**Условия подозрительности транзакции:**

- Если сумма транзакции превышает пороговое значение (например, 1000 долларов).
- Транзакция происходит в необычное время (например, ночью с 23:00 до 5:00).
- Транзакция происходит в месте, отличном от last_known_location либо от адреса регистрации.

**Чек-лист самопроверки**

| Критерии выполнения задания | Отметка о выполнении |
|---|---|
| Установлен  jupyter notebook либо используется  google colab | |
| Создан профиль на https://github.com (при использовании jupyter notebook) | |
| Создан топик с транзакционной активностью пользователей, сгенерированные данные для потока. Создана БД с основной информацией о клиенте. | |
| Соединены поток с транзакционной активностью и основная информация о клиенте. | |
| Данные с подозрительной активностью выгружены в отдельный поток. | |
| Прикреплена на учебной платформе  ссылка на выполненное задание в Google colab или Github (если вы использовали jupyter notebook) | |
| Если используется  Google colab, то по ссылке есть доступ | |

| | |
|---|---|
| (иногда в колабе нет доступа для другого логина) | |