Периодический запуск процедуры очистки датасета мошеннических финансовых транзакций

Домашнее задание No 4

Цель работы. В данном домашнем задании Вы потренируетесь в создании инфраструктуры с помощью инструмента Apache Airflow, организации периодического запуска процедуры очистки данных, познакомитесь с концепцией ориентированных направленных графов (DAGs), с помощью которых организуется последовательность запуска задач по расписанию, научитесь разрабатывать собственные графы с помощью языка Python для Apache Airflow.

Уважаемый слушатель!

Итак, данные очищены и загружены в хранилище. В принципе, можно бы уже начинать их анализ, однако, Вас беспокоит одна проблема. Антифрод-система продолжает работать и накапливать данные, а очистку этих данных Вы выполнили лишь единожды. Мошенники продолжают искать новые уязвимые места в системе защиты, а это означает, что модель, обученная на уже собранных и не обновляемых данных, скоро устареет и будет неспособна к качественному анализу транзакций...

Из сказанного выше вытекает необходимость создания системы, способной периодически получать новые данные из озера компании, проверять их качество, очищать и добавлять к уже существующим в Вашем хранилище. Скрипт для очистки датасета у Вас уже есть; теперь нужно с помощью Apache Airflow обеспечить его периодический запуск на требуемой порции новых данных.

Обратите внимание!

Систему Apache Airflow желательно запустить в облачной среде, что позволит ей создавать необходимые ресурсы и уничтожать их после выполнения задания для экономии ресурсов.

Вам предлагается:

1. Запустить систему Apache Airflow на отдельной виртуальной машине Yandex cloud.

2. Создать DAG для ежедневного автоматизированного создания и удаления Spark-кластера и запуска скрипта очистки датасета и разместить его в директории для DAG’ов, доступной Apache Spark. В графе следует прописать этапы копирования скрипта и необходимых ему файлов на Spark-кластер, а также его запуска на кластере посредством spark-submit.

3. Убедиться, что граф загрузился в систему и отображается в графическом интерфейсе. Файл(-ы) с DAG необходимо разместить в Вашем GitHub- репозитории и предоставить для проверки.

4. Разрешить периодическое исполнение разработанного DAG в Apache AirFlow и протестировать его работоспособность. Требуется дождаться не менее трёх успешных запусков процедуры очистки датасета по расписанию. Снимок экрана, подтверждающий успешную работу системы, необходимо привести в README-файле Вашего GitHub-репозитория.

5. В соответствии с достигнутыми результатами, изменить статус ранее созданных задач на Kanban-доске в GitHub Projects. Возможно, некоторые задачи нужно будет скорректировать, разделить на подзадачи или объединить друг с другом.

6. Полностью удалить созданный кластер, чтобы избежать оплаты ресурсов в период его простаивания.

Для получения положительной оценки за работу необходимо выполнить минимум первые четыре вышеприведенных задания.

Желаем успехов!