Анализ качества и очистка датасета мошеннических финансовых операций

Домашнее задание No 3

Цель работы. В данном домашнем задании Вы познакомитесь с основными проблемами, которые могут встречаться в данных, потренируетесь определять их наличие в датасете, обрабатывать и очищать набор данных, разрабатывать скрипты очистки данных с использованием Apache Spark.

Уважаемый слушатель!

Вы создали свое объектное хранилище, в котором разместили фрагмент данных о совершенных финансовых транзакциях компании, предоставленный системным администратором. Казалось бы, можно постепенно приступать к их анализу, однако, Вас смущает один момент. Структуру данных в датасете Вы знаете, но вопрос их качества остается открытым. Неясно, что конкретно находится в тех гигабайтах данных, которые Вы получили. К тому же, во время очередного обеденного перерыва Вы пообщались с коллегами, занимающимися обслуживанием системы транзакций, и они с улыбкой рассказали Вам, как занимались ее восстановлением после очередного отказа...

Скорее всего, сбор данных тоже в эти моменты тоже нарушался, поэтому нужно проанализировать содержимое датасета на отсутствие в нем ошибок. Поскольку процесс оценки качества и очистки датасета придется выполнять периодически, необходим скрипт, выполняющий эти действия.

Обратите внимание!

В целях экономии ресурсов в Yandex cloud все манипуляции необходимо произвести с использованием terraform скриптов, которые являются частью домашнего задания и должны быть представлены в вашем github репозитории.

Вам предлагается:

1. Создать служебный аккаунт в Yandex Cloud для работы с кластером Data Proc и предоставить ему необходимые роли (необходимые роли подробно указаны в официальной документации).

2. Создать новый bucket в Yandex Cloud Object Storage и предоставить созданному выше системному аккаунту право на запись к нему. Для проверки преподавателем данный bucket необходимо сделать общедоступным на чтение, а точку доступа к нему привести в README-файле Вашего GitHub- репозитория.

3. Создать Spark-кластер в Data Proc, указав в настройках созданный выше bucket, с двумя подкластерами со следующими характеристиками:

а) Мастер-подкластер: класс хоста s3-c2-m8, размер хранилища 40 ГБ.

б) Compute-подкластер: класс хоста s3-c4-m16, от 3 хостов, размер хранилища – от 128 ГБ. Требуемый объем локальных дисков зависит от объема кэшированных DataFrames в процессе работы Вашего скрипта.

Обратите внимание!

Для использования кластера в учебных целях проще всего организовать публичный доступ к мастер-узлу. Данное действие потребует создания группы безопасности, ограничивающей возможные соединения. Кроме разрешений входящих и исходящих соединений, указанных в документации, для удобства работы с кластером рекомендуется разрешить:

а) входящие TCP-соединения на порт 22 для SSH;

б) входящие TCP-соединения на порт 8888 для Jupyter Notebook;

в) исходящие TCP-соединения со всех портов мастер-узла.

4. Проанализировать датасет мошеннических транзакций на наличие в нем ошибочных данных. Данное действие рекомендуется выполнять с помощью среды Jupyter Notebook, запущенной на мастер-узле кластера. Нужно оценить, какие из основных проблем с данными могут иметь место в рассматриваемом датасете, и постараться выявить факт их наличия, колонки, которые они затрагивают, объем некорректных данных и т.д.

Обратите внимание!

Основные проблемы, которые могут встречаться в данных, описаны в: https://en.wikipedia.org/wiki/Data_cleansing#Data_quality

5. На основе проведенного анализа качества создать скрипт, который должен выполнять очистку данных с использованием Apache Spark. Скрипт должен иметь возможность автоматического запуска внешней системой.

6. Выполнить очистку датасета с использованием созданного скрипта и сохранить его в созданном выше bucket’е в формате parquet, подходящем для хранения большого объема структурированных данных.

7. 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.

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

Для получения положительной оценки за работу необходимо выполнить минимум первые шесть вышеприведенных заданий, обнаружив не менее трех типов некорректных данных.

Желаем успехов!