Лабораторная работа №5

МОДЕЛЬ КЛАСТЕРИЗАЦИИ НА PYSPARK

Целью этой лабораторной является получение навыков разработки и настройки Spark приложения и реализация кластеризации.

Для реализации в будущем следующих лабораторных работ был использован образ spark приложения который позволяет запускать узлы в docker. Ссылка на репозиторий: <https://github.com/big-data-europe/docker-spark>

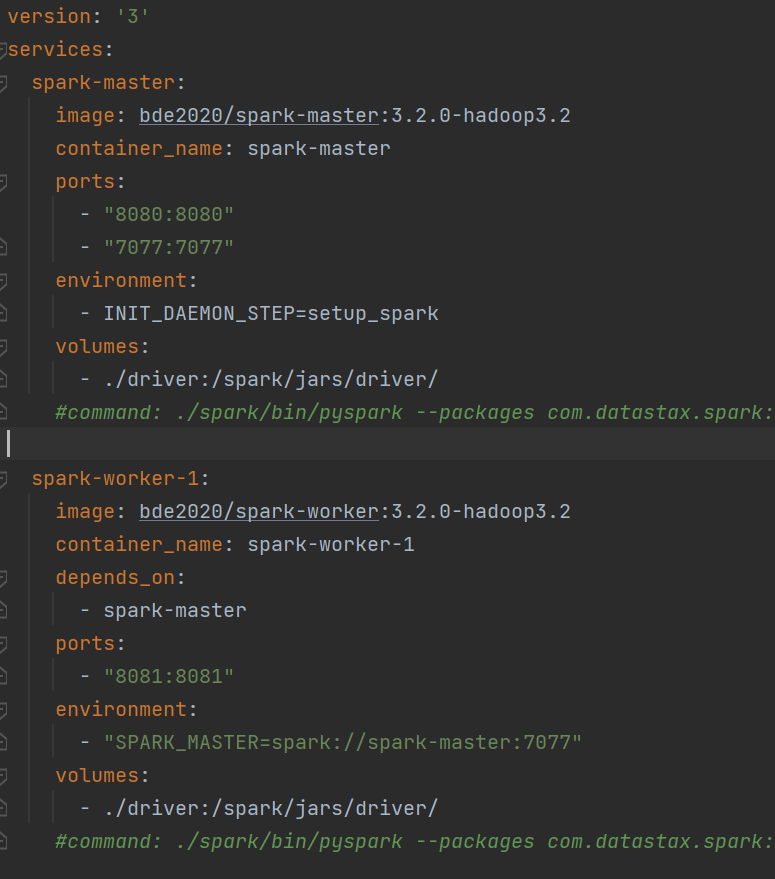
Был разработан скрипт который отделяет от основного файла данные размером около 300мб. Также разработан отдельный контейнер который служит для записи данных в базу данных Cassandra которая развернута также в докер контейнере. Репозиторий с кодом доступен ниже.

Ссылка на репозиторий git: https://github.com/Saivaks/big\_data\_lab\_5

В качестве модели был выбран алгоритм k средних. Были выбраны некоторые данные из файла а именно значения 'creator', 'pnns\_groups\_1', 'pnns\_groups\_2' для проведения кластеризации. Результат кластеризации был также зафиксирован в базе cassandra.

Также в отчете есть приложения в виде скринов обновленного docker compose и CI/CD

Docker compose:



Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

CI/CL:

Изображение выглядит как текст

Автоматически созданное описание