HW2:
Basic Spark

Описание работы и критериев оценивания

Домашнее задание основано на данных
https://www.kaggle.com/datasets/bahramjannesarr/goodreads-book-datasets-10m

ДЗ предлагается сделать в ноутбуке и в формате ipynb загрузить в репозиторий на Github
При разработке решений следует обратить особое внимание на эффективность производимых преобразований

Бонусы и штрафы:

- **100%** за плагиат
- **30%** за посылку решения в течение недели после deadline

Блок 1. Standalone Spark

1) Развернуть standalone cluster Spark: master + 2 workers. Приложить скрипт и/или алгоритм + скрин webui (10 баллов)
2) Подключиться к кластеру с помощью Jupyter и/или Zeppelin. Приложить скрипт и/или алгоритм + скрин рабочей сессии из инструмента (20 баллов)

+ 10 дополнительных баллов за развертывание и подключение к HDFS

Блок 2. Работа с данными на Spark

1) Преобразовать данные исходного датасета в parquet объединяя все таблицы. Оценить разницу в скорости чтения / занимаемом объеме. Сделать выводы. (15 баллов)
2) Используя весь набор данных с помощью Spark вывести (5 баллов за каждое задание)
   a) Топ-10 книг с наибольшим числом ревью
   b) Топ-10 издателей с наибольшим средним числом страниц в книгах
   c) Десять наиболее активных по числу изданных книг лет
   d) Топ-10 книг имеющих наибольший разброс в оценках среди книг имеющих больше 500 оценок
   e) Любой интересный инсайт из данных

Блок 3. Spark Streaming

В задании предлагается реализовать расчет среднего рейтинга книги на Spark Streaming со следующими условиями (30 баллов):

- использовать данные user_rating как file source
- использовать file sink в формате parquet

+ 10 дополнительных баллов за использование Kafka как source и sink с описанием развертывания и процессом первичной доставки данных