Отчет по лабораторной работе №2 «Spark»

Выполнил:

Студент группы М19-502  
Куренков Д. А.

Москва, 2020 г.

1. Бизнес логика:

Реализовать программу, которая агрегирует метрики по заданному интервалу

1. Точка входа - bash
2. Технология хранения - HDFS
3. Технология вычисления – Spark DataFrame

[Ссылка на репозиторий](https://github.com/zpbkda/ds_bda_hw2).



Рисунок 1 - Схема компонентов

Для выполнения работы необходимо установить Apache Spark и переменные окружения. Hadoop уже имеется на виртуальной машине. Для этого используем следующие команды:

wget https://apache-mirror.rbc.ru/pub/apache/spark/spark-2.4.7/spark-2.4.7-bin-hadoop2.7.tgz

tar xvf spark-2.4.7-bin-hadoop2.7.tgz

echo 'export SPARK\_HOME=$HOME/spark-2.4.7-bin-hadoop2.7' >> .bash\_profile

echo 'export PATH=$PATH:$SPARK\_HOME/bin' >> .bash\_profile

Запускаем HDFS и YARN, для этого введем следующие команды в корневой директории Hadoop:

sbin/start-dfs.sh

sbin/start-yarn.sh

Далее необходимо написать скрипт для генерации метрик и запуска программы. Скрипт представлен на рисунке 2.

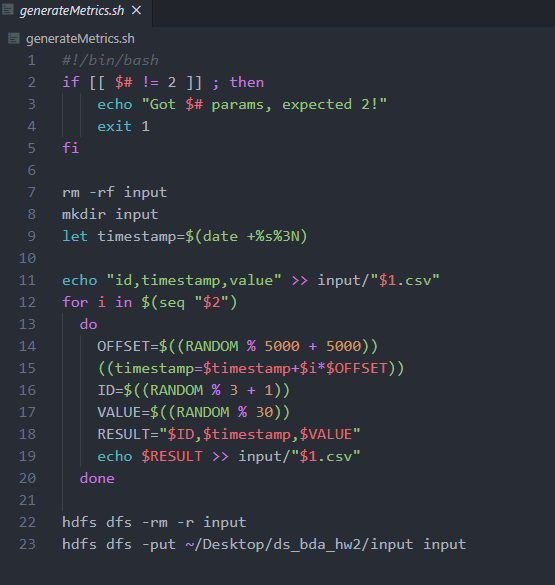


Рисунок 2 – Скриншот скрипта генерации метрик

Скрипт принимает 2 аргумента: первый – имя для создания файла с метриками, второй – количество записей в этом файле. Скрипт создает записи метрик, далее удаляет уже имеющиеся в hdsf папку input и загружает только что сгенерированный файл с метриками.

Далее вторым скриптом запускается программа.

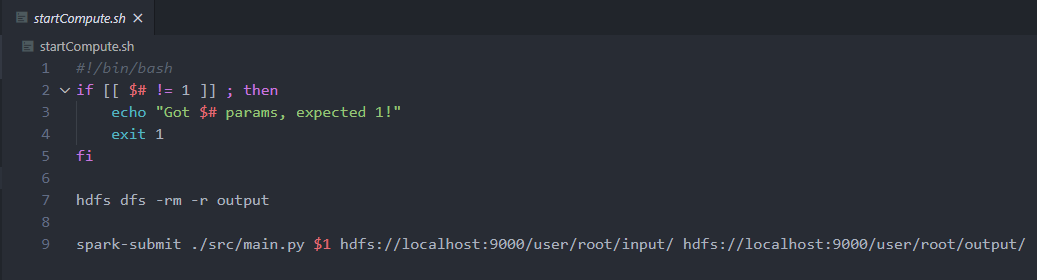


Рисунок 3 – Скриншот скрипта запуска программы

Скрипт принимает один аргумент: интервал агрегации метрик. Далее производит удаление папки HDFS output и запускает программу через команду spark-submit. В качестве аргументов для программы передаются: временной интервал агрегации метрик, папка с файлом метрик в HDFS, папка для записи результата в HDFS.

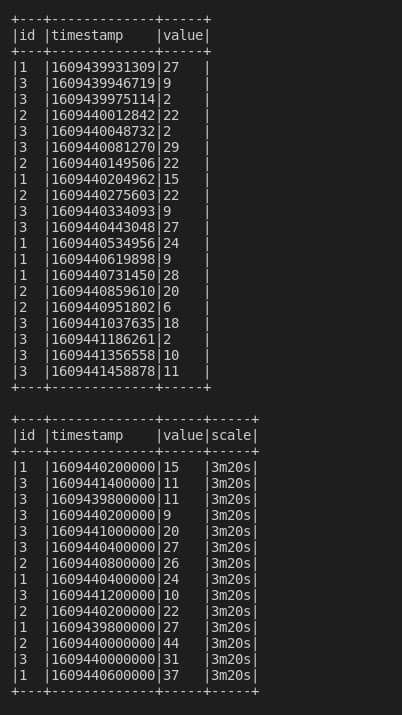


Рисунок 4 – Скриншот успешного выполнения и результатов программы

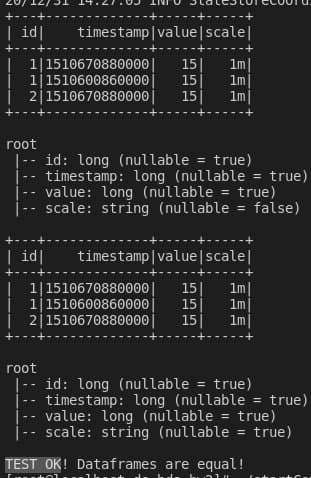


Рисунок 5 – Результаты успешного выполнения тестов

Весь исходный код расположен в репозитории на гитхабе. Ссылка в начале отчета.