МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ

ФЕДЕРАЦИИ

Федеральное государственное автономное

образовательное учреждение высшего образования

«Самарский национальный исследовательский

университет имени академика С.П. Королёва»

(Самарский университет)

Институт информатики и кибернетики

Кафедра технической кибернетики

Лабораторная работа № 2

«REPORTS WITH APACHE SPARK»

Выполнил студент: Ермаков Д.Д.

Гр. 6132 - 010204D

Проверил преподаватель: Попов С.Б.

Самара 2022

**ЗАДАНИЕ**

1. Преобразовать любой файл набора данных stackoverflow в parquet формат с помощью Apache Spark.
2. Сформировать отчёт с информацией о частоте обсуждения 10 наиболее популярных языков программирования в каждом году с 2010 года по сегодняшний день. Используйте теги входящие в список языков перечисленных в википедии https://en.wikipedia.org/wiki/List\_of\_programming\_languages.

**ХОД РАБОТЫ**

Подготовим необходимые файлы для лабораторной работы, для этого загрузим файлы programming-languages.csv и posts\_sample.xml. Напишем программу, которая формирует отчет о частоте обсуждения популярных языков программирования с 2010 по 2022.

Используемый язык для программы — Scala.

Сформируем список годов с 2010 по 2022 и список языков программирования. В файле post\_sample оставляем год публикации и тэг, состоящий из названия языка программирования. Произведем фильтрацию RDD по году и тэгу. Вычислим количество постов каждого языка программирования, отсортируем по времени, выберем первые 10 за каждый год и получим топ 10 языков программирования.

Выведем результат на экран, сохранив датафрейм в формате parquet при помощи команды df.write.format("parquet").save("name.parquet").

Ссылка на репозиторий: https://github.com/ShadowOfDen/BigData/tree/main/Lab2.

Внутри репозитория в файле README.md содержится таблица по популярности языков за каждый год.

**ЗАКЛЮЧЕНИЕ**

В результате лабораторной работы с помощью Apache Spark был преобразован файл набора данных stackoverflow в parquet формат. Сформирован отчёт с информацией о частоте обсуждения 10 популярных языков программирования в каждом году с 2010 года по 2022.