|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ **Информатика и системы управления**

КАФЕДРА **Компьютерные системы и сети (ИУ6)**

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.04.01 Информатика и вычислительная техника**

МАГИСТЕРСКАЯ ПРОГРАММА **09.04.01/07 Интеллектуальные системы анализа,**

**обработки и интерпретации больших данных.**

**Отчет**

|  |  |
| --- | --- |
| **по лабораторной работе №** | 10 |
| **Вариант №** 8 |  |

**Название:**

Работа со Scala и Spark

**Дисциплина:** Языки программирования для работы с большими данными

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | ИУ6-23М | |  |  | Н.М. Иванюк |
|  | | (Группа) |  | (Подпись, дата) | (И.О. Фамилия) |
|  | |  |  |  |  |
| Преподаватель | |  |  |  | П.В. Степанов |
|  | |  |  | (Подпись, дата) | (И.О. Фамилия) |

Москва, 2022

## Цель работы

Получение навыков работы с языком программирования Spark .

## Ход работы

*Задание 1.*

Выбрать любой датасет на kaggle.com

Cделать 10 выборок данных на ваше усмотрение

В качестве среды установки для Spark и Hadoop была выбрана система контейнерной виртуализации Docker. Была использована следующая конфигурация docker-compose:

Листинг 1 – Код конфигурации системы контейнеров для запуска Spark

version: "3.6"

volumes:

shared-workspace:

name: "hadoop-distributed-file-system"

driver: local

services:

jupyterlab:

image: andreper/jupyterlab:3.0.0-spark-3.0.0

container\_name: jupyterlab

ports:

- 8888:8888

- 4040:4040

volumes:

- shared-workspace:/opt/workspace

spark-master:

image: andreper/spark-master:3.0.0

container\_name: spark-master

ports:

- 8080:8080

- 7077:7077

volumes:

- shared-workspace:/opt/workspace

spark-worker-1:

image: andreper/spark-worker:3.0.0

container\_name: spark-worker-1

environment:

- SPARK\_WORKER\_CORES=1

- SPARK\_WORKER\_MEMORY=512m

ports:

- 8081:8081

volumes:

- shared-workspace:/opt/workspace

depends\_on:

- spark-master

spark-worker-2:

image: andreper/spark-worker:3.0.0

container\_name: spark-worker-2

environment:

- SPARK\_WORKER\_CORES=1

- SPARK\_WORKER\_MEMORY=512m

ports:

- 8082:8081

volumes:

- shared-workspace:/opt/workspace

depends\_on:

- spark-master

...

В качестве датасета для работы был выбран набор данных <https://www.kaggle.com/vitaliymalcev/russian-passenger-air-service-20072020>. (Russian Passenger Air Service).

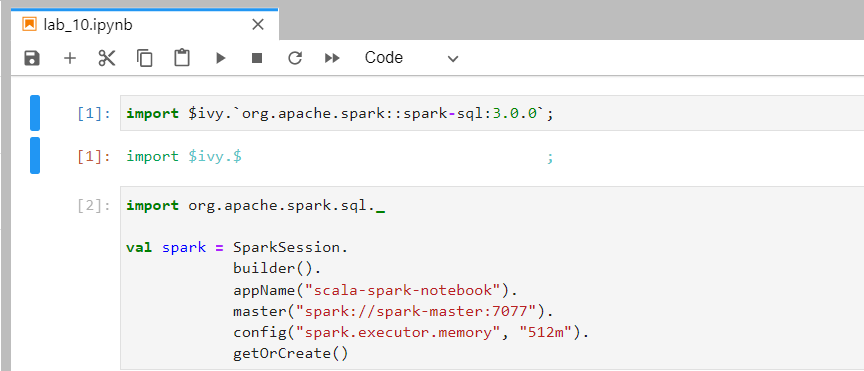


Рисунок 1 – Подключение к Spark из среды JupyterLab

В среде JupyterLab были отработаны запросы для поиска данных в датасете. Приведем некоторые из них.



Рисунок 2 – Запрос Spark SQL Select

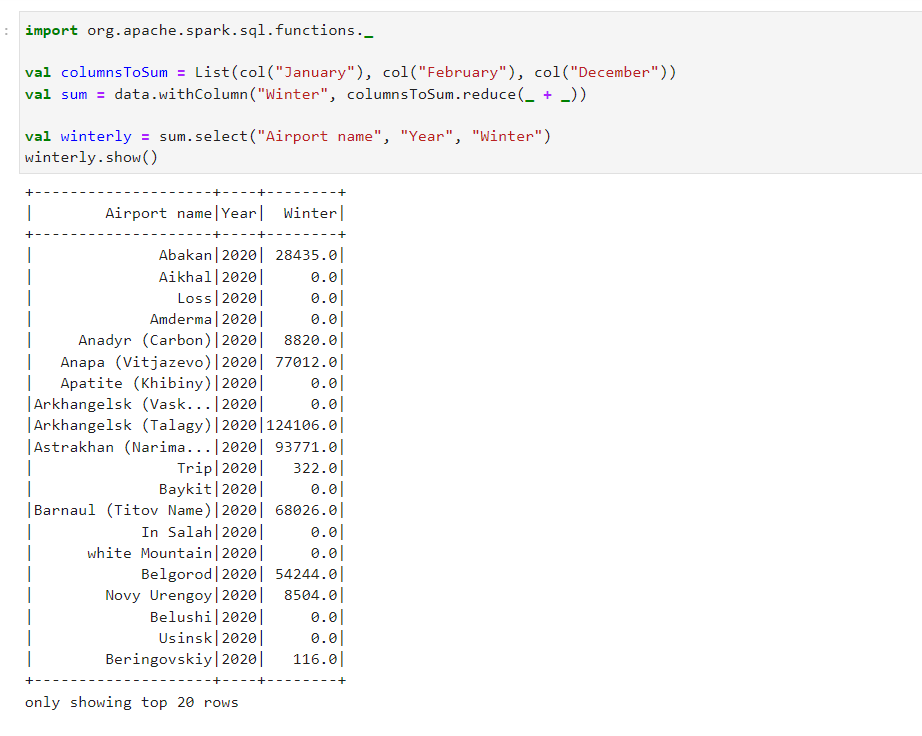


Рисунок 3 – Запрос с созданием новых столбцов

## Вывод

По итогам выполнения лабораторной работы были получены навыки программирования на языке Scala, а также получен опыт выполнения запросов к системе Spark.