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.**

О Т Ч Е Т

по лабораторной работе № 1 0

Вариант № 17

Название: Работа со Scala и Spark

Дисциплина: Языки программирования для работы с большими данными

Студент ИУ6-23М (Группа) _____ (Подпись, дата) М.О. Усманов (И.О. Фамилия)

Преподаватель _____ (Подпись, дата) П.В. Степанов (И.О. Фамилия)

Москва, 2022

Цель работы

Получение навыков работы с языком программирования Spark .

Ход работы

Задание 1.

– Выбрать любой датасет на kaggle.com

– Сделать 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).

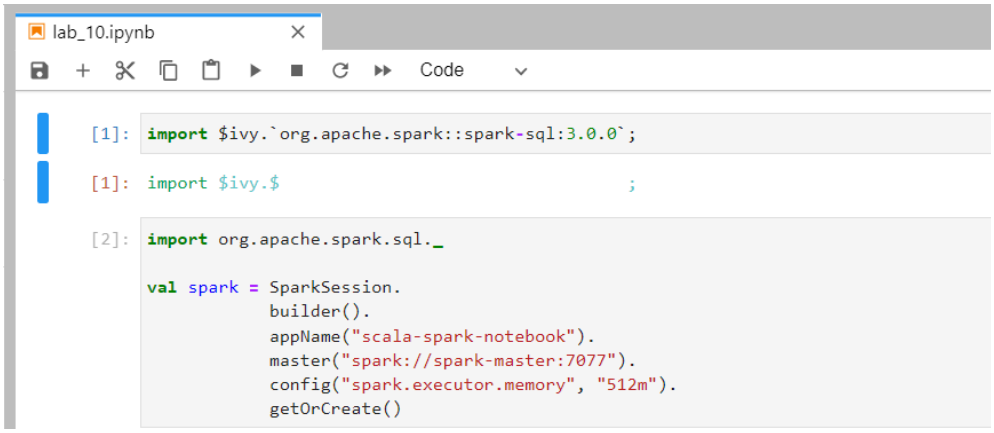

Рисунок 1 – Подключение к Spark из среды JupyterLab

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

```
val airports = data.select("Airport name","Year", "Airport coordinates")
airports.show()
```

```
+--------------------+----+--------------------+
|        Airport name|Year| Airport coordinates|
+--------------------+----+--------------------+
|              Abakan|2020|(Decimal('91.3997...|
|              Aikhal|2020|(Decimal('111.543...|
|               Loss|2020|(Decimal('125.398...|
|            Amderma|2020|(Decimal('61.5774...|
|     Anadyr (Carbon)|2020|(Decimal('177.738...|
|   Anapa (Vitjazevo)|2020|(Decimal('37.3415...|
|   Apatite (Khibiny)|2020|(Decimal('33.5819...|
|Arkhangelsk (Vask...|2020|(Decimal('40.7067...|
|Arkhangelsk (Talagy)|2020|(Decimal('40.7148...|
|Astrakhan (Narima...|2020|(Decimal('47.9998...|
|               Trip|2020|(Decimal('138.042...|
|              Baykit|2020|(Decimal('96.3667...|
|Barnaul (Titov Name)|2020|(Decimal('83.5477...|
|            In Salah|2020|(Decimal('130.399...|
|      white Mountain|2020|(Decimal('146.228...|
|           Belgorod|2020|(Decimal('36.5705...|
|        Novy Urengoy|2020|(Decimal('66.6945...|
|             Belushi|2020|(Decimal('47.6234...|
|              Usinsk|2020|(Decimal('65.0461...|
|        Beringovskiy|2020|(Decimal('179.293...|
+--------------------+----+--------------------+
only showing top 20 rows
```

Рисунок 2 – Запрос Spark SQL Select

```
airports.filter(airports("Airport name") === "Belgorod").show()
```

```
+------------+----+--------------------+
|Airport name|Year| Airport coordinates|
+------------+----+--------------------+
|    Belgorod|2020|(Decimal('36.5705...|
|    Belgorod|2019|(Decimal('36.5705...|
|    Belgorod|2018|(Decimal('36.5705...|
|    Belgorod|2017|(Decimal('36.5705...|
|    Belgorod|2016|(Decimal('36.5705...|
|    Belgorod|2015|(Decimal('36.5705...|
|    Belgorod|2014|(Decimal('36.5705...|
|    Belgorod|2013|(Decimal('36.5705...|
|    Belgorod|2012|(Decimal('36.5705...|
|    Belgorod|2011|(Decimal('36.5705...|
|    Belgorod|2010|(Decimal('36.5705...|
|    Belgorod|2009|(Decimal('36.5705...|
|    Belgorod|2008|(Decimal('36.5705...|
|    Belgorod|2007|(Decimal('36.5705...|
+------------+----+--------------------+
```

Рисунок 3 – Запрос с фильтрацией данных

```
import org.apache.spark.sql.functions._

val columnsToSum = List(col("January"), col("February"), col("December"))
val sum = data.withColumn("Winter", columnsToSum.reduce(_ + _))

val winterly = sum.select("Airport name", "Year", "Winter")
winterly.show()
```

```
+--------------------+----+--------+
|        Airport name|Year|  Winter|
+--------------------+----+--------+
|              Abakan|2020| 28435.0|
|              Aikhal|2020|     0.0|
|               Loss|2020|     0.0|
|             Amderma|2020|     0.0|
|     Anadyr (Carbon)|2020|  8820.0|
|   Anapa (Vitjazevo)|2020| 77012.0|
|   Apatite (Khibiny)|2020|     0.0|
|Arkhangelsk (Vask...|2020|     0.0|
|Arkhangelsk (Talagy)|2020|124106.0|
|Astrakhan (Narima...|2020| 93771.0|
|                Trip|2020|   322.0|
|              Baykit|2020|     0.0|
|Barnaul (Titov Name)|2020| 68026.0|
|            In Salah|2020|     0.0|
|      white Mountain|2020|     0.0|
|            Belgorod|2020| 54244.0|
|        Novy Urengoy|2020|  8504.0|
|             Belushi|2020|     0.0|
|              Usinsk|2020|     0.0|
|        Beringovskiy|2020|   116.0|
+--------------------+----+--------+
only showing top 20 rows
```

Рисунок 4 – Запрос с созданием новых столбцов

Файл ноутбука со всеми запросами доступен в репозитории проекта.

Местоположение репозитория с файлами проекта

Файлы проекта расположены в репозитории веб-платформы для совместной разработки Github. Местоположение в репозитории:

https://github.com/s314/big-data-studies/tree/main/lab_10

Вывод

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