



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

Название: Spark

Дисциплина: Языки программирования для работы с большими
данными

Студент ИУ6-22М

Д. Р. Григорян

Преподаватель

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

Москва, 2023

Цель работы:

Ознакомиться с языком программирования Java, создать проект для написания 10 запросов из выбранного датасета, используя Spark.

Выполнение:

Задача 1:

Использовать Spark, Scala/Java. Выбрать любой датасет на kaggle.com и сделать 10 выборок данных по выбранной предметной области.

Листинг 1 программы main:

```
package org.example;

import org.apache.spark.sql.SparkSession;
import org.apache.spark.sql.Dataset;
import org.apache.spark.sql.Row;

public class Main {
    public static void main(String[] args) {

        SparkSession spark = SparkSession
            .builder()
            .appName("Java Spark SQL basic example")
            .config("spark.master", "local")
            .getOrCreate();

        Dataset<Row> df =
            spark.read().option("header", "true").csv("C:/Users/sonar/IdeaProjects/lab10/NYPD_Shooting_Incident_Data__Year_To_Date_.csv");

        df.createOrReplaceTempView("NYPD");
        df.groupBy("BORO").count().show();
        df.groupBy("VIC_SEX").count().show();
        spark.sql("Select * FROM NYPD").show();
        spark.sql("Select * FROM NYPD WHERE VIC_RACE != 'BLACK'").show();
        spark.sql("Select * FROM NYPD WHERE Police Precincts > 30 AND VIC_SEX = 'M'").show();
        spark.sql("Select * FROM NYPD WHERE STATISTICAL_MURDER_FLAG = 'Y' ORDER BY BORO").show();
        spark.sql("Select * FROM NYPD WHERE LOC_OF_OCCUR_DESC = 'INSIDE' OR PERP_AGE_GROUP NOT LIKE '18-24'").show();
        spark.sql("Select * FROM NYPD WHERE PERP_SEX = '(null)'").show();
        spark.sql("Select * FROM NYPD WHERE City Council Districts > 10 AND Police Precincts < 50").show();
        spark.sql("Select * FROM NYPD WHERE LOC_CLASSFCTN_DESC = 'VEHICLE' ORDER BY PRECINCT DESC").show();
    }
}
```

Результаты запросов показаны на рисунках 1-5:

```
23/04/14 17:29:25 INFO DAGScheduler: Job 10 finished: show at main.java:17, took 0,592977 s
+-----+-----+
|VIC_SEX|count|
+-----+-----+
|      F|  212|
|      M| 1504|
+-----+-----+

23/04/14 17:29:25 INFO FileSourceStrategy: Pruning directories with:
23/04/14 17:29:25 INFO FileSourceStrategy: Post-Scan Filters:
23/04/14 17:29:25 INFO FileSourceStrategy: Output Data Schema: struct<INCIDENT_KEY: string, OCCUR_DATE: string, OCCUR_
23/04/14 17:29:25 INFO FileSourceScanExec: Pushed Filters:
23/04/14 17:29:25 INFO CodeGenerator: Code generated in 57.9883 ms
23/04/14 17:29:25 INFO MemoryStore: Block broadcast_17 stored as values in memory (estimated size 221.4 KB, free 2.5 G
23/04/14 17:29:25 INFO MemoryStore: Block broadcast_17_piece0 stored as bytes in memory (estimated size 20.6 KB, free
```

Рисунок 1 – Результат выполнения запроса

INCIDENT_KEY	OCCUR_DATE	OCCUR_TIME	BORO	LOC_OF_OCCUR_DESC	PRECINCT	JURISDICTION_CODE	LOC_CLASSFCTN_DESC
243566884	04/12/2022	22:08:00	BRONX	OUTSIDE	49	0	STREET
256484816	12/17/2022	04:08:00	BRONX	OUTSIDE	52	0	STREET
250216145	08/27/2022	00:21:00	BRONX	OUTSIDE	44	0	VEHICLE
239207164	01/15/2022	19:50:00	QUEENS	OUTSIDE	113	2	HOUSING MULTI DW
248013313	07/14/2022	01:19:00	BROOKLYN	INSIDE	77	0	DWELLING MULTI DW
247542571	07/04/2022	22:20:00	BRONX	OUTSIDE	48	0	STREET
251813843	09/29/2022	21:32:00	BROOKLYN	OUTSIDE	81	0	STREET
254000625	11/11/2022	15:24:00	BROOKLYN	OUTSIDE	75	0	OTHER MULTI DW
252647537	10/16/2022	20:05:00	BROOKLYN	OUTSIDE	75	0	STREET
249207672	08/08/2022	04:20:00	MANHATTAN	OUTSIDE	20	0	STREET
246104595	06/03/2022	15:41:00	BRONX	OUTSIDE	41	0	STREET
254062679	11/12/2022	22:18:00	MANHATTAN	OUTSIDE	10	0	STREET
247501139	07/02/2022	12:23:00	MANHATTAN	OUTSIDE	7	0	STREET
245766300	05/28/2022	00:41:00	BROOKLYN	OUTSIDE	75	0	DWELLING MULTI DW
247099746	06/23/2022	10:08:00	BRONX	OUTSIDE	48	0	STREET
245824483	05/29/2022	10:41:00	BROOKLYN	OUTSIDE	73	0	HOUSING MULTI DW
245491991	05/22/2022	14:48:00	BRONX	OUTSIDE	46	0	STREET

Рисунок 2 – Результат выполнения запроса

INCIDENT_KEY	OCCUR_DATE	OCCUR_TIME	BORO	LOC_OF_OCCUR_DESC	PRECINCT	JURISDICTION_CODE	LOC_CLASSFCTN_DESC
256484816	12/17/2022	04:08:00	BRONX	OUTSIDE	52	0	STREET
250216145	08/27/2022	00:21:00	BRONX	OUTSIDE	44	0	VEHICLE
251813843	09/29/2022	21:32:00	BROOKLYN	OUTSIDE	81	0	STREET
249207672	08/08/2022	04:20:00	MANHATTAN	OUTSIDE	20	0	STREET
246104595	06/03/2022	15:41:00	BRONX	OUTSIDE	41	0	STREET
247099746	06/23/2022	10:08:00	BRONX	OUTSIDE	48	0	STREET
245491991	05/22/2022	14:48:00	BRONX	OUTSIDE	46	0	STREET
251276894	09/18/2022	13:35:00	QUEENS	OUTSIDE	110	0	STREET
239583960	01/23/2022	14:44:00	BRONX	OUTSIDE	52	0	STREET
248145612	07/15/2022	21:00:00	QUEENS	OUTSIDE	101	0	STREET
243312638	04/09/2022	02:00:00	BROOKLYN	OUTSIDE	73	0	STREET MULTI DW
243433247	04/08/2022	13:39:00	BRONX	OUTSIDE	40	0	STREET
244102489	04/24/2022	19:29:00	MANHATTAN	OUTSIDE	25	0	STREET
244970464	05/12/2022	21:39:00	QUEENS	OUTSIDE	106	0	STREET
260637443	12/28/2022	18:36:00	BRONX	OUTSIDE	48	0	STREET

Рисунок 3 – Результат выполнения запроса

INCIDENT_KEY	OCCUR_DATE	OCCUR_TIME	BORO	LOC_OF_OCCUR_DESC	PRECINCT	JURISDICTION_CODE	LOC_CLASSFCTN_DESC	LOCATION_DESC	STATISTICAL_MURDER_FLAG	PERP_AGE_GROUP	PERP_S
256484816	12/17/2022	04:08:00	BRONX	OUTSIDE	52	0	STREET	(null)	N	25-44	
239207164	01/15/2022	19:50:00	QUEENS	OUTSIDE	113	2	HOUSING(MULTI DWELL - PUB...		N	(null)	(nul
247542571	07/04/2022	22:20:00	BRONX	OUTSIDE	48	0	STREET	(null)	N	(null)	(nul
251813843	09/29/2022	21:32:00	BROOKLYN	OUTSIDE	81	0	STREET	(null)	N	25-44	
254008625	11/11/2022	15:24:00	BROOKLYN	OUTSIDE	75	0	OTHER(MULTI DWELL - APT...		N	(null)	(nul
245766380	05/28/2022	00:41:00	BROOKLYN	OUTSIDE	75	0	DWELLING(MULTI DWELL - APT...		N	(null)	(nul
247899746	06/23/2022	10:08:00	BRONX	OUTSIDE	48	0	STREET	(null)	N	(null)	(nul
245824483	05/29/2022	10:41:00	BROOKLYN	OUTSIDE	73	0	HOUSING(MULTI DWELL - PUB...		Y	25-44	
251276894	09/18/2022	13:35:00	QUEENS	OUTSIDE	110	0	STREET	(null)	N	25-44	
239583960	01/23/2022	14:44:00	BRONX	OUTSIDE	52	0	STREET	(null)	N	(null)	(nul
243312638	04/09/2022	02:00:00	BROOKLYN	OUTSIDE	73	0	STREET(MULTI DWELL - APT...		N	(null)	(nul
248540227	07/25/2022	22:58:00	BROOKLYN	OUTSIDE	71	0	STREET(MULTI DWELL - APT...		N	(null)	(nul
245824158	05/29/2022	04:41:00	BROOKLYN	OUTSIDE	73	0	STREET	(null)	N	18-24	
247523460	07/04/2022	02:20:00	BROOKLYN	OUTSIDE	75	0	STREET	(null)	N	(null)	(nul
246876573	06/19/2022	21:39:00	BROOKLYN	OUTSIDE	69	0	STREET	PVT HOUSE	N	25-44	
244970464	05/12/2022	21:39:00	QUEENS	OUTSIDE	106	0	STREET	PVT HOUSE	N	18-24	
260637443	12/28/2022	18:36:00	BRONX	OUTSIDE	48	0	STREET	GROCERY/BODEGA	N	25-44	
254110477	11/14/2022	16:19:00	QUEENS	OUTSIDE	107	2	HOUSING(MULTI DWELL - PUB...		N	(null)	(nul
250551052	09/02/2022	20:35:00	BROOKLYN	OUTSIDE	94	0	STREET	(null)	N	(null)	(nul
246172373	06/06/2022	01:25:00	QUEENS	OUTSIDE	104	0	COMMERCIAL	FACTORY/WAREHOUSE	Y	18-24	

only showing top 20 rows

Рисунок 4 – Результат выполнения запроса

INCIDENT_KEY	OCCUR_DATE	OCCUR_TIME	BORO	LOC_OF_OCCUR_DESC	PRECINCT	JURISDICTION_CODE	LOC_CLASSFCTN_DESC	LOCATION_DESC	STATISTICAL_MURDER_FLAG	PERP_AGE_GROUP	PERP_SEX
239552853	01/22/2022	06:00:00	BRONX	OUTSIDE	40	0	STREET	(null)	Y	18-24	M
240254036	02/07/2022	16:10:00	BRONX	OUTSIDE	40	0	STREET	(null)	Y	25-44	M
250109555	08/24/2022	21:30:00	BRONX	INSIDE	44	2	HOUSING(MULTI DWELL - PUB...		Y	(null)	(null)
250503799	09/01/2022	16:05:00	BRONX	INSIDE	46	0	DWELLING(MULTI DWELL - APT...		Y	18-24	M
242517445	03/16/2022	13:15:00	BRONX	INSIDE	48	0	DWELLING(MULTI DWELL - APT...		Y	25-44	M
238885819	01/09/2022	22:54:00	BRONX	OUTSIDE	44	0	STREET	(null)	Y	18-24	M
246359110	06/08/2022	11:28:00	BRONX	OUTSIDE	46	0	STREET	(null)	Y	<18	M
242158428	03/15/2022	19:23:00	BRONX	OUTSIDE	44	0	STREET(MULTI DWELL - APT...		Y	25-44	M
247492010	07/02/2022	13:30:00	BRONX	OUTSIDE	40	0	STREET	(null)	Y	18-24	M
253001794	10/22/2022	05:08:00	BRONX	OUTSIDE	47	0	DWELLING	PVT HOUSE	Y	45-64	M
246876580	06/18/2022	00:20:00	BRONX	OUTSIDE	44	0	STREET	GROCERY/BODEGA	Y	25-44	M
251582994	09/24/2022	23:57:00	BRONX	OUTSIDE	41	0	STREET	SMALL MERCHANT	Y	<18	M
252648397	10/15/2022	08:09:00	BRONX	OUTSIDE	41	0	STREET	(null)	Y	18-24	M
243115069	04/04/2022	19:06:00	BRONX	OUTSIDE	46	0	STREET	GROCERY/BODEGA	Y	25-44	M
245384614	05/16/2022	04:19:00	BRONX	OUTSIDE	45	0	STREET	(null)	Y	25-44	M
244513744	05/03/2022	00:46:00	BRONX	INSIDE	46	0	COMMERCIAL	GROCERY/BODEGA	Y	25-44	M
247943981	07/12/2022	22:46:00	BRONX	OUTSIDE	47	0	STREET	(null)	Y	25-44	M
243566884	04/12/2022	22:08:00	BRONX	OUTSIDE	49	0	STREET	(null)	Y	18-24	M
250858391	09/09/2022	00:12:00	BRONX	INSIDE	49	2	HOUSING(MULTI DWELL - PUB...		Y	25-44	M
238779470	01/06/2022	20:20:00	BRONX	INSIDE	48	2	HOUSING(MULTI DWELL - PUB...		Y	(null)	(null)

only showing top 20 rows

Рисунок 5 – Результат выполнения запроса

Вывод: в ходе выполнения лабораторной работы была написана программа, которая использует Spark Session для подключения к определенному датасету и с помощью Spark Sql выполняет из него 10 выборки.