|  |  |
| --- | --- |
| 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 |

**Название:** Spark

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | ИУ6-22М |  |  | Н.Б. Гендина | |
|  | (Группа) |  | (Подпись, дата) | | (И.О. Фамилия) |
|  |  |  |  | |  |
| Преподаватель |  |  |  | | П.В. Степанов |
|  |  |  | (Подпись, дата) | | (И.О. Фамилия) |

Москва, 2024

**Цель работы:** освоить базовые принципы работы со Spark на языке Java.

**Задание 1**

Используемый датасет Netflix Movies and TV Shows - все фильмы и шоу Netflix . Пример записей в формате csv.

|  |
| --- |
| show\_id,type,title,director,cast,country,date\_added,release\_year,rating,duration,listed\_in,description  s1,Movie,Dick Johnson Is Dead,Kirsten Johnson,,United States,"September 25, 2021",2020,PG-13,90 min,Documentaries,"As her father nears the end of his life, filmmaker Kirsten Johnson stages his death in inventive and comical ways to help them both face the inevitable."  s2,TV Show,Blood & Water,,"Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Molaba, Dillon Windvogel, Natasha Thahane, Arno Greeff, Xolile Tshabalala, Getmore Sithole, Cindy Mahlangu, Ryle De Morny, Greteli Fincham, Sello Maake Ka-Ncube, Odwa Gwanya, Mekaila Mathys, Sandi Schultz, Duane Williams, Shamilla Miller, Patrick Mofokeng",South Africa,"September 24, 2021",2021,TV-MA,2 Seasons,"International TV Shows, TV Dramas, TV Mysteries","After crossing paths at a party, a Cape Town teen sets out to prove whether a private-school swimming star is her sister who was abducted at birth." |

Выбрать любой датасет на kaggle.com. Cделать 10 выборок данных по выбранной предметной области

|  |
| --- |
| package net.alchim31.maven;  import org.apache.spark.sql.\*;  public class Main {  public static void main(String[] args) {  System.out.println("Инициализация spark");  SparkSession sparkSession = SparkSession.builder().appName("spark\_test.com").master("local[\*]").getOrCreate();  sparkSession.sparkContext().setLogLevel("ERROR");  String path = System.getProperty("user.dir") + "/netflix\_titles.csv";  Dataset<Row> dataset = sparkSession.read().format("csv").option("header", "true").load(path);  dataset.createOrReplaceTempView("netflix");  System.out.println("\n\nРезультаты запросов:");  System.out.println("\n1. Пример записей");  sparkSession.sql("select \* from netflix limit 10").show();  System.out.println("\n2. Пропорция фильмов и передач");  sparkSession.sql("select sum(case when type = 'Movie' then 1 end) / count(\*) \* 100 as `Процент фильмов`, sum(case when type = 'TV Show' then 1 end) / count(\*) \* 100 as `Процент передач` from netflix").show();  System.out.println("\n3. Самые популярные режиссеры");  sparkSession.sql("select explode(SPLIT(director, ', ')) as director, count(\*) as `Количество работ` from netflix where director is not null group by director order by 2 desc limit 10").show();  System.out.println("\n4. Страны, выпускавшие фильмы в разные годы");  sparkSession.sql("select release\_year, explode(SPLIT(country, ', ')) as country, count(\*) from netflix where release\_year > 1980 group by 1, 2 order by 1, 3 desc limit 10").show();  System.out.println("\n5. Средняя продолжительность фильмов");  sparkSession.sql("select AVG(TRIM(SUBSTRING(duration, 1, 3))) as `Средня продолжительность` from netflix where type='Movie'").show();  System.out.println("\n6. Самые часто снимающиеся актеры");  sparkSession.sql("select explode(SPLIT(cast, ', ')) as `Актер`, count(\*) as `Количество` from netflix group by 1 order by 2 desc").show();  System.out.println("\n7. Фильмы, содержащие в описаниии journey");  sparkSession.sql("select title from netflix where description LIKE '%journey%'").show();  System.out.println("\n8. Возрастное деление");  sparkSession.sql("select rating, count(\*) from netflix group by rating order by 2 desc limit 10").show();  System.out.println("\n9. Самые частые жанры");  sparkSession.sql("select explode(SPLIT(listed\_in, ', ')) as `Жанр`, count(\*) as `Количество` from netflix group by 1 order by 2 desc").show();  System.out.println("\n10. Жанры, которые снимают режиссеры");  sparkSession.sql("select director as `Режиссер`, concat\_ws(', ', collect\_set(listed\_in)) as `Жанры` from netflix group by 1").show();  sparkSession.stop();  }  } |

**Вывод:** в ходе лабораторной работы были освоены базовые принципы работы со Spark на языке Java.