*Министерство науки и высшего образования Российской Федерации  
Федеральное государственное бюджетное образовательное учреждение высшего образования*

|  |  |
| --- | --- |
| **Gerb-BMSTU_01** | ***«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

ДИСЦИПЛИНА Языки программирования для работы с большими данными

**Отчет**

**по лабораторной работе № 10**

**Вариант 5**

Студент гр. ИУ6-21М **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_**Войтов Д.Ю.\_\_\_\_

(Подпись, дата) (И.О. Фамилия)

Преподаватель  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О. Фамилия)

Москва, 2020

*Условие*

Исследовать предложенный набор данных при помощи языка Scala и фреймворка Apache Spark. Для изучения был выбран набор Russian\_demography с ресурса Kaggle. Вид набора данных приведен на рисунке 1.

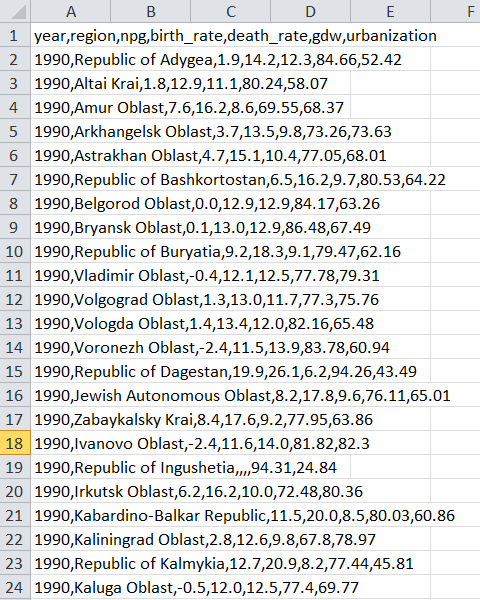


Рисунок 1 – Набор данных

*Решение*

Код главного объекта Scala приведен ниже:

|  |
| --- |
| import org.apache.spark.{SparkConf, SparkContext}  import org.apache.spark.sql.SparkSession  object MO\_Analyzer {  def main(args: Array[String]): Unit = {  //val logFile = "C:/Spark/readme.md"  val conf = new SparkConf().setAppName("CounterDemo").setMaster("local[\*]")  val sc = new SparkContext(conf)  val spark = SparkSession.builder().appName(name = "TestApp").getOrCreate()  val dataFile = spark.read.format("com.databricks.spark.csv")  .option("header", true).load("E:\\IdeaProjects\\Lab10\\src\\main\\scalarussian\_demography.csv")  dataFile.createOrReplaceTempView("Demography")  // Рождаемости/смертности по МО  spark.sql("SELECT year, birth\_rate, death\_rate FROM Demography WHERE region='Moscow Oblast'").show()  // Средняя рождаемость в МО в первую декаду 21 века  spark.sql("SELECT AVG(birth\_rate) FROM Demography WHERE region='Moscow Oblast' AND year<2010 AND year>=2000").show()  spark.stop()  }  } |

*Результаты работы*

Результат работы первого запроса приведен на рисунке 2. Здесь выведена статистика по рождаемости и смертности в МО

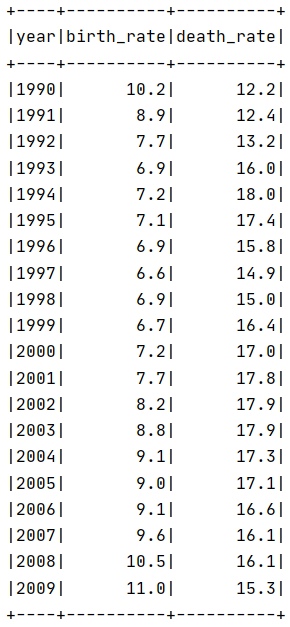


Рисунок 2 – Рождаемость\Смертность в МО

Второй запрос (рисунок 3) выводит статистику средней рождаемости в МО в период между 2000 и 2010 годами.

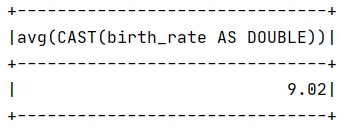


Рисунок 3 – Средняя рождаемость в МО на период первой декады 21 века

*Вывод*

В ходе выполнения данной работы было настроено взаимодействие Apache Spark и языка Scala на базе среды IntelliJ IDEA. Полученные результаты выполнения SQL-запросов свидетельствуют о верной настройке системы.