Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Факультет інформатики та обчислювальної техніки

Кафедра автоматизованих систем обробки інформації та управління

Методи і засоби обробки великих даних

Лабораторна робота №3

«Базове знайомство з Spark»

Виконав:

студент групи ІС-12мп

Коноплянка Д. С.

Перевірила:

Тимофєєва Ю. С.

Київ 2021 р.

1. Cтворити кілька своїх DataFrames. Виконати над ними операції об’єднання, додавання та видалення стовпців і рядків.

val spark:SparkSession = SparkSession.builder()

.master("local[1]").appName("SparkByExamples.com")

.getOrCreate()

import spark.implicits.\_

val columns = Seq("language","users\_count")

val data = Seq(("Java", "20000"), ("Python", "100000"), ("Scala", "3000"))

val rdd = spark.sparkContext.parallelize(data)

val dfFromRDD1 = rdd.toDF("language","users\_count")

dfFromRDD1.printSchema()



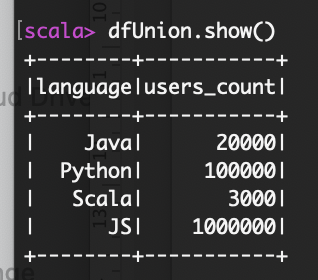
Add new row with union

val newRow = Seq(("JS", "1000000"))

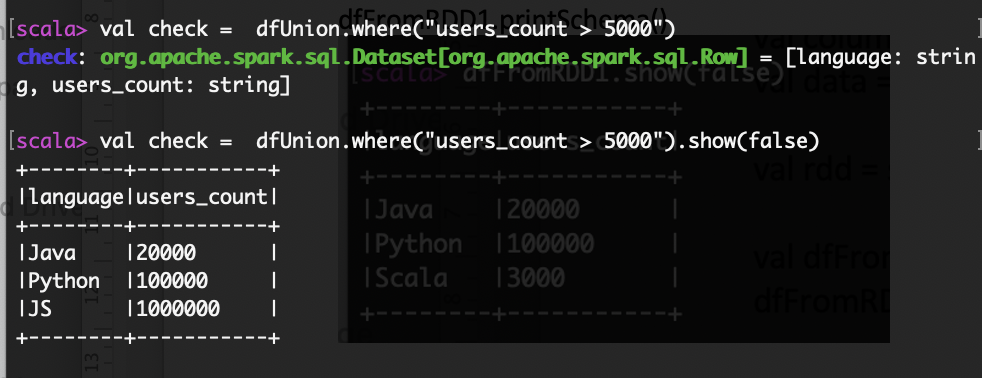
val rdd2 = spark.sparkContext.parallelize(newRow)

val dfFromRDD2 = rdd2.toDF("language","users\_count")

val dfUnion = dfFromRDD1.union(dfFromRDD2)

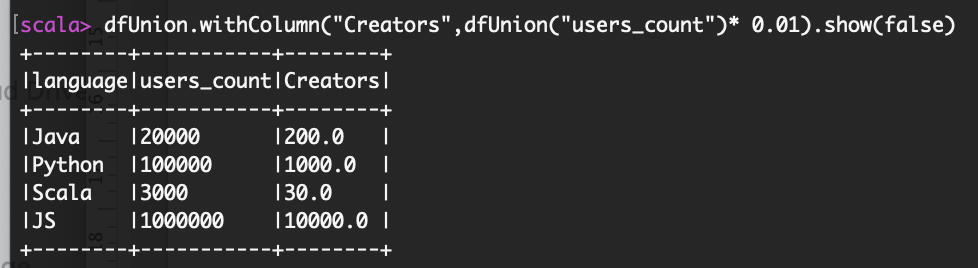


Remove row where users\_count > 5000

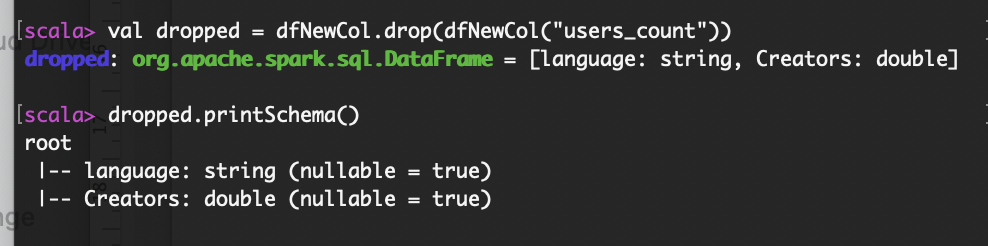


Add new column

dfUnion.withColumn("Creators",dfUnion("users\_count")\* 0.01).show(false)



Drop column



Варіант 6.

Файл Titanic.csv.

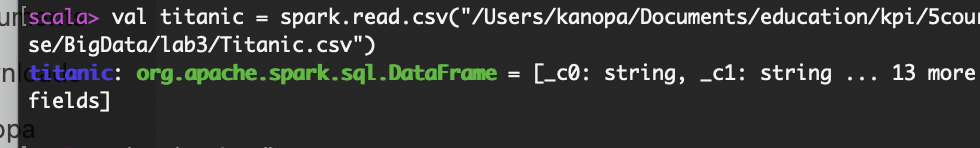
1. Порахувати загальну кількість дітей.

2. Вивести всіх жінок, які подорожували третім класом і вижили.

3. Порахувати, скільки американці витратили на білети.

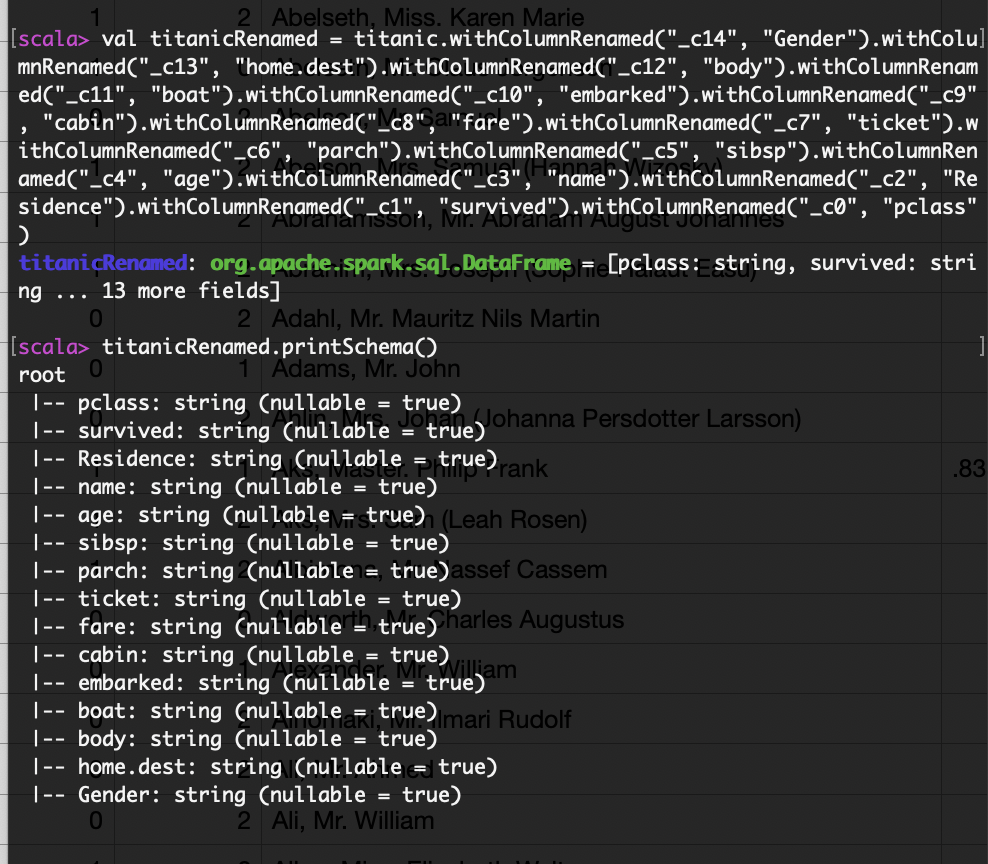
4. Відсортувати дані за віком пасажирів (спадаючий).

val titanic = spark.read.csv("/Users/kanopa/Documents/education/kpi/5course/BigData/lab3/Titanic.csv")

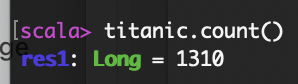


Rename all columns

val titanicRenamed = titanic.withColumnRenamed("\_c14", "Gender").withColumnRenamed("\_c13", "home.dest").withColumnRenamed("\_c12", "body").withColumnRenamed("\_c11", "boat").withColumnRenamed("\_c10", "embarked").withColumnRenamed("\_c9", "cabin").withColumnRenamed("\_c8", "fare").withColumnRenamed("\_c7", "ticket").withColumnRenamed("\_c6", "parch").withColumnRenamed("\_c5", "sibsp").withColumnRenamed("\_c4", "age").withColumnRenamed("\_c3", "name").withColumnRenamed("\_c2", "Residence").withColumnRenamed("\_c1", "survived").withColumnRenamed("\_c0", "pclass")

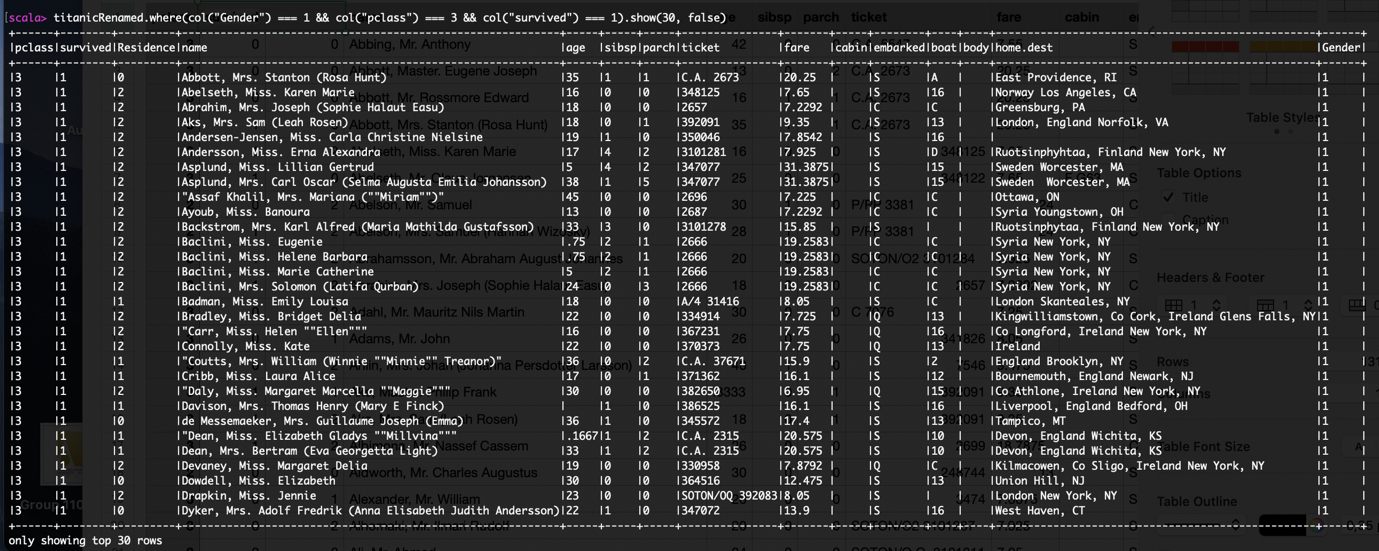


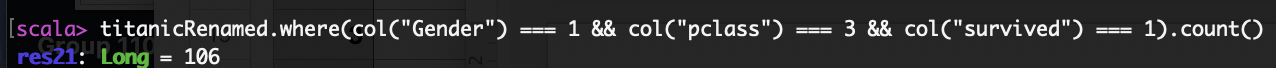
1) Порахувати загальну кількість дітей.



2) Вивести всіх жінок, які подорожували третім класом і вижили.

titanicRenamed.where(col("Gender") === 1 && col("pclass") === 3 && col("survived") === 1).count()

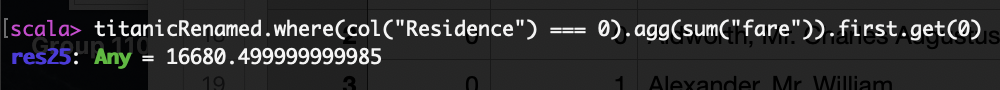




3) Порахувати, скільки американці витратили на білети.

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

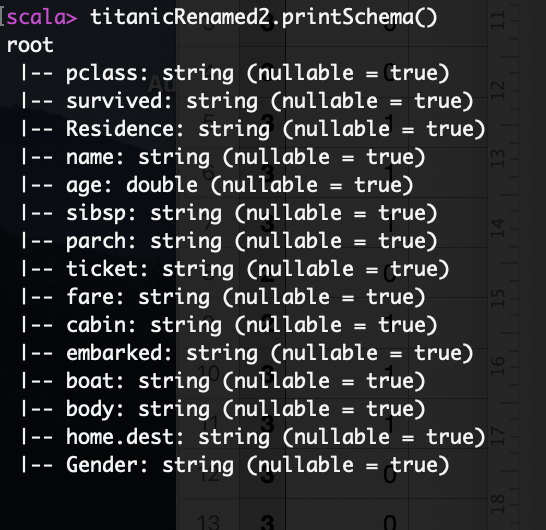
titanicRenamed.where(col("Residence") === 0).agg(sum("fare")).first.get(0)



4) Відсортувати дані за віком пасажирів (спадаючий).

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

val titanicRenamed2 = titanicRenamed.withColumn("age", col("age").cast("Double"))



titanicRenamed2.orderBy(desc("age")).show(300 ,false)

