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 №4

«Spark SQL»

Виконав:

студент групи ІС-12мп

Коноплянка Д. С.

Перевірила:

Тимофєєва Ю. С.

Київ 2021 р.

Варіант 6.

Файл Titanic.csv.

1. Створити нову таблицю, яка буде містити наступні два стовпці: ціна

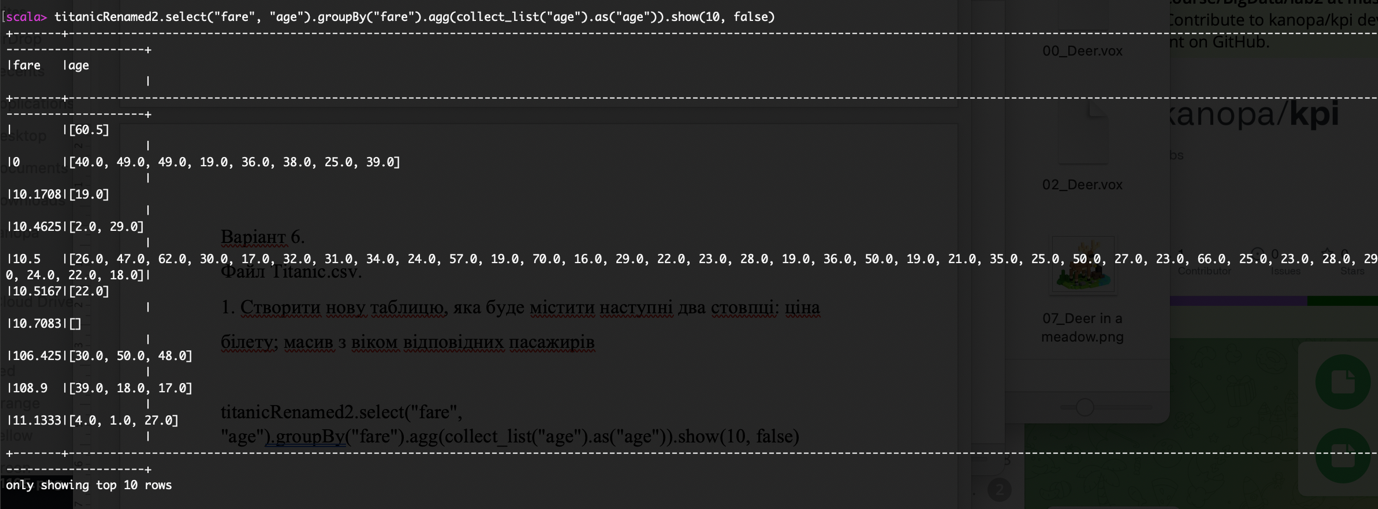
білету; масив з віком відповідних пасажирів

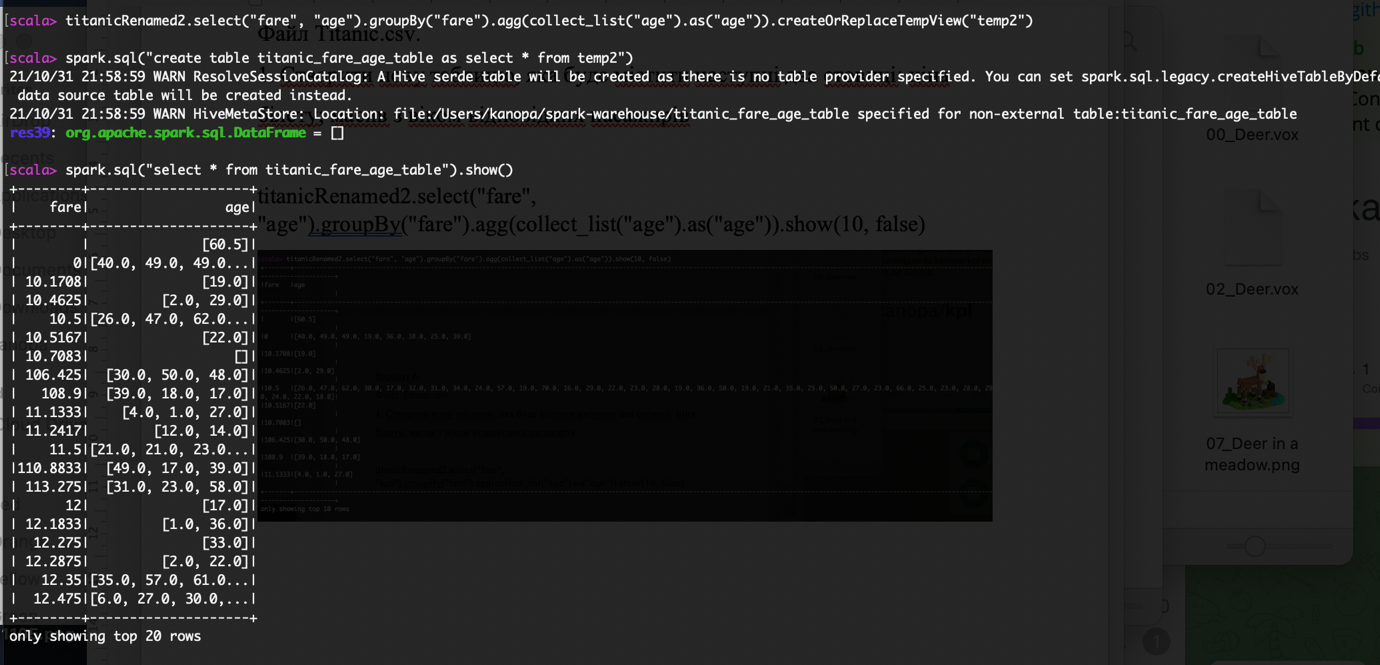
titanicRenamed2.select("fare", "age").groupBy("fare").agg(collect\_list("age").as("age")).show(10, false)

titanicRenamed2.select("fare", "age").groupBy("fare").agg(collect\_list("age").as("age")).createOrReplaceTempView("temp2")

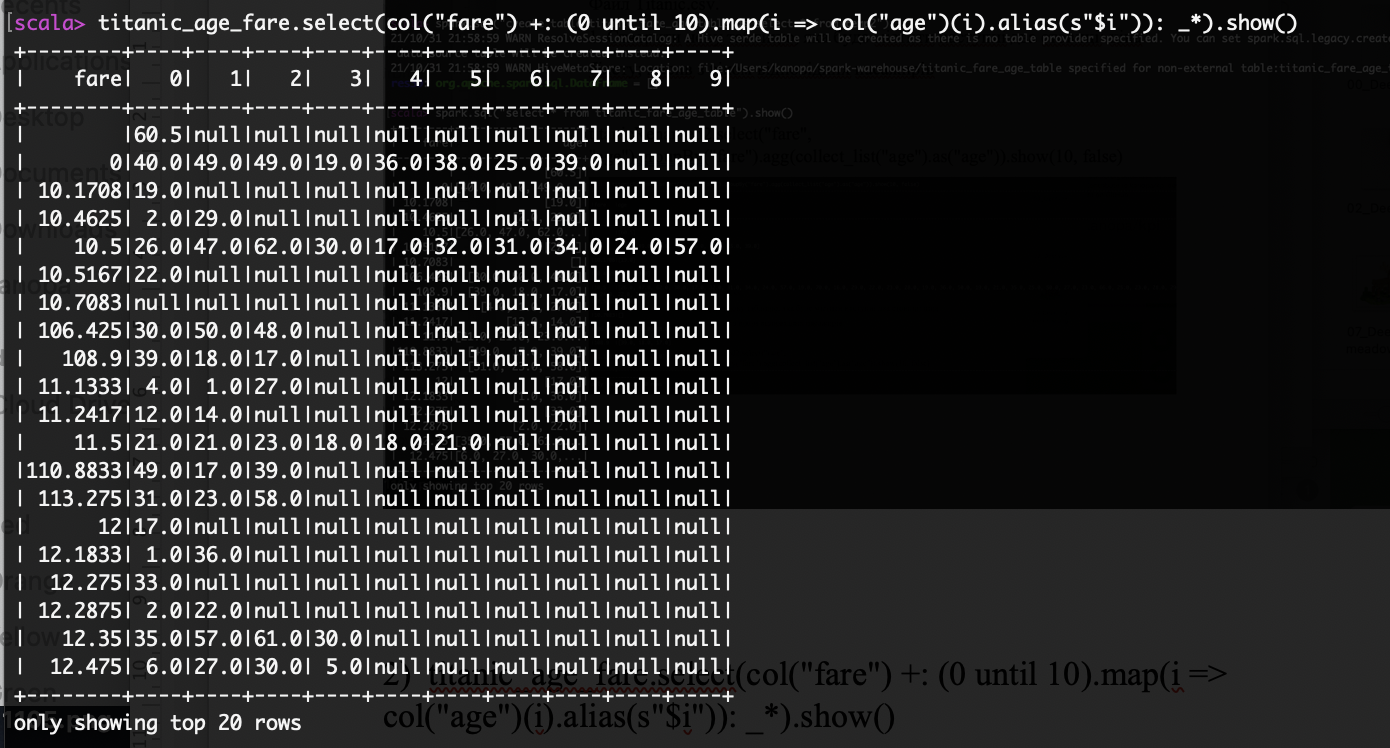
spark.sql("create table titanic\_fare\_age\_table as select \* from temp2")

spark.sql("select \* from titanic\_fare\_age\_table").show()





2) titanic\_age\_fare.select(col("fare") +: (0 until 10).map(i => col("age")(i).alias(s"$i")): \_\*).show()



3) Знайти середній вік для усіх можливих поєднань параметрів survived

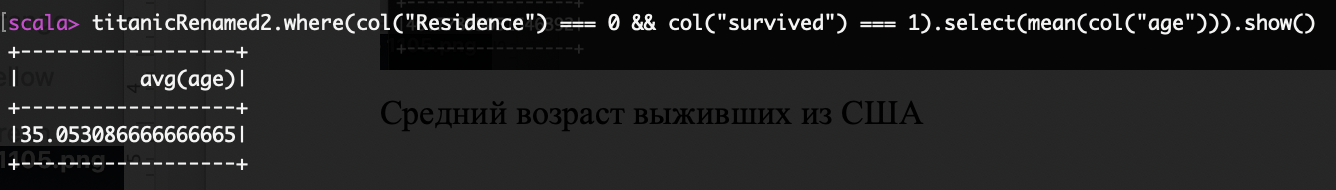
та Residence.

Средний возраст умерших из США

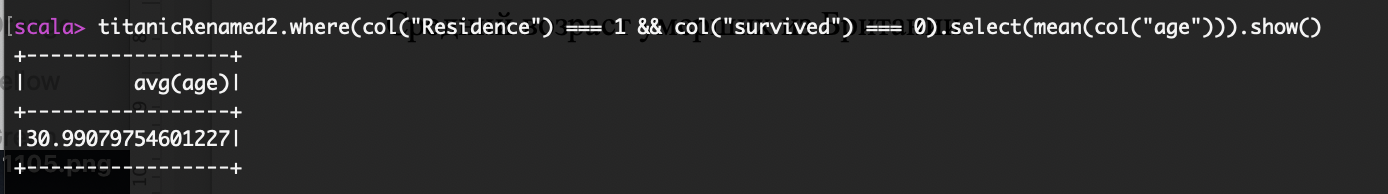
titanicRenamed2.where(col("Residence") === 0 && col("survived") === 0).select(mean(col("age"))).show()



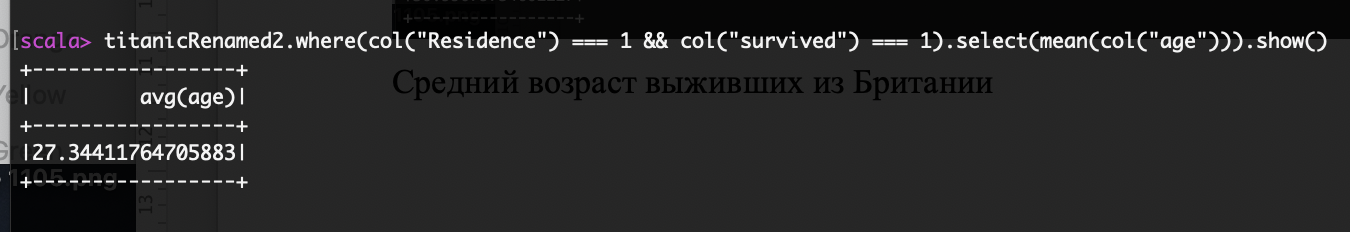
Средний возраст выживших из США



Средний возраст умерших из Британии



Средний возраст выживших из Британии



Средний возраст умерших с других стран

Средний возраст выживших с других стран



4) Додати в таблицю Titanic новий стовпець, що містить різницю між

максимальним та поточним віком в залежності від класу.

Наприклад, якщо максимальний рік пасажирів 2-го класу 50, а в

першому рядку вік 30 і другий клас, то значення в додатковому

стовпчику = 50-30=20. Зберегти новий файл.

titanicRenamed2.withColumn("strangeAge", (col("age") \* -1) + when(col("pclass") === 1, titanicRenamed2.where(col("pclass") === 1).agg(max(col("age"))).first.get(0)).when(col("pclass") === 2, titanicRenamed2.where(col("pclass") === 2).agg(max(col("age"))).first.get(0)).when(col("pclass") === 3, titanicRenamed2.where(col("pclass") === 3).agg(max(col("age"))).first.get(0))).show(100, false)



newFile.repartition(1).write.format("com.databricks.spark.csv").option("header", "true").save("data.csv")



