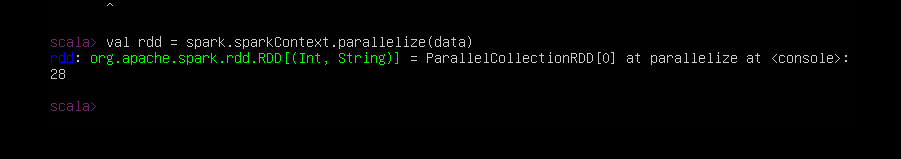
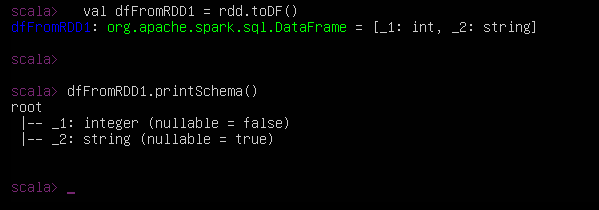
BIGDATA ANALYSIS TUẦN 9 BÀI TẬP SPARK

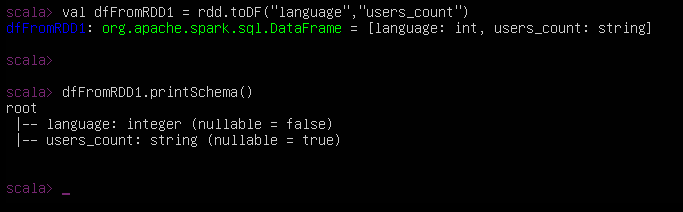
1. **Tạo DataFrame**
2. Sử dụng hàm toDF()

Sử dụng đối tượng rdd

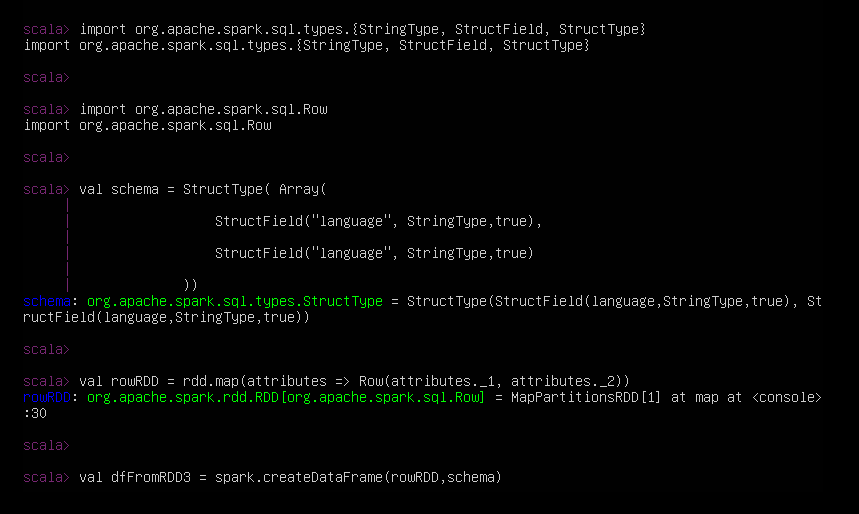


sử dụng toDF()để tạo DataFrame trong Spark



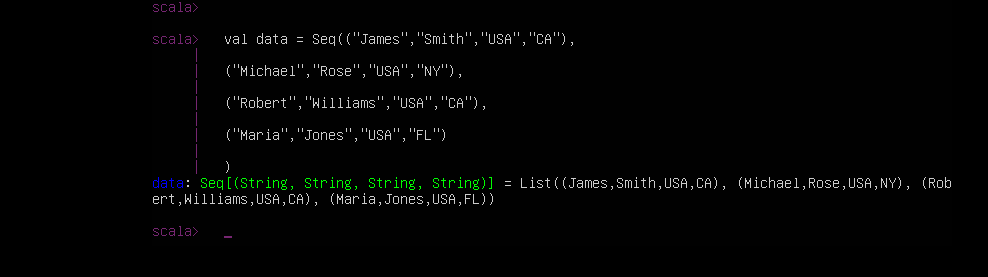


1. Sử dụng createDataFrame() với kiểu hàng

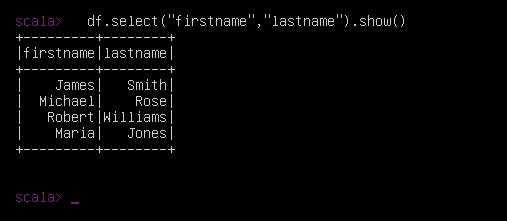


1. Chọn cột

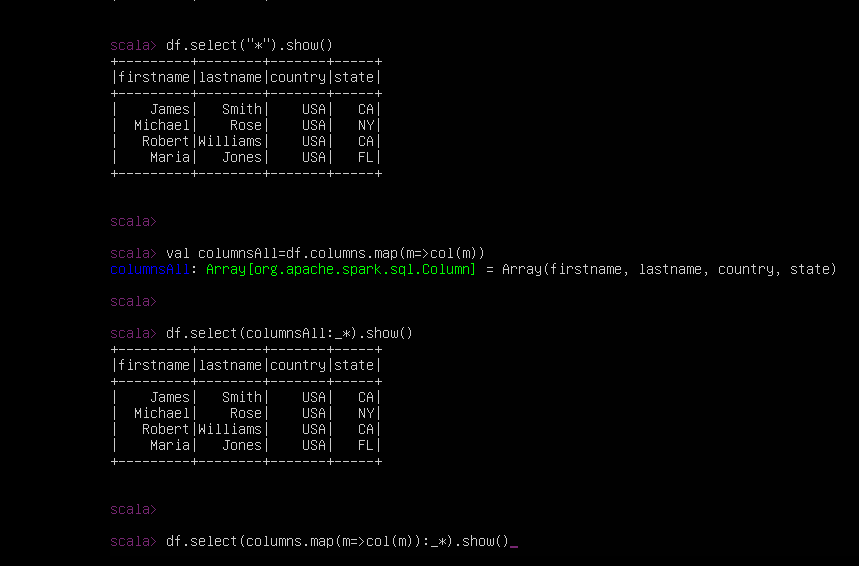
Tạo các dối tượng



Đưa ra kết quả khi tìm FirstName và LastName



Chọn tất cả các cột thì sẽ show ra hết các cột



Chọn các cột từ danh sách



1. Thêm và cập nhật cột
2. import org.apache.spark.sql.{Row, SparkSession}
3. import org.apache.spark.sql.types.{StringType, StructType}
4. val data = Seq(Row(Row("James;","","Smith"),"36636","M","3000"),
5. Row(Row("Michael","Rose",""),"40288","M","4000"),
6. Row(Row("Robert","","Williams"),"42114","M","4000"),
7. Row(Row("Maria","Anne","Jones"),"39192","F","4000"),
8. Row(Row("Jen","Mary","Brown"),"","F","-1")
9. )
10. val schema = new StructType()
11. .add("name",new StructType()
12. .add("firstname",StringType)
13. .add("middlename",StringType)
14. .add("lastname",StringType))
15. .add("dob",StringType)
16. .add("gender",StringType)
17. .add("salary",StringType)
18. val df = spark.createDataFrame(spark.sparkContext.parallelize(data),schema)



1. Thêm cột mới vào DataFrame

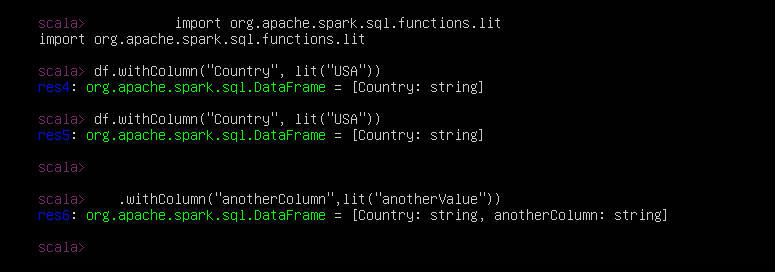
import org.apache.spark.sql.functions.lit

df.withColumn("Country", lit("USA"))

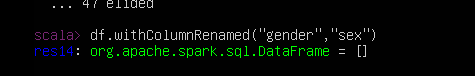
//chaining to operate on multiple columns

df.withColumn("Country", lit("USA"))

withColumn("anotherColumn",lit("anotherValue"))

****

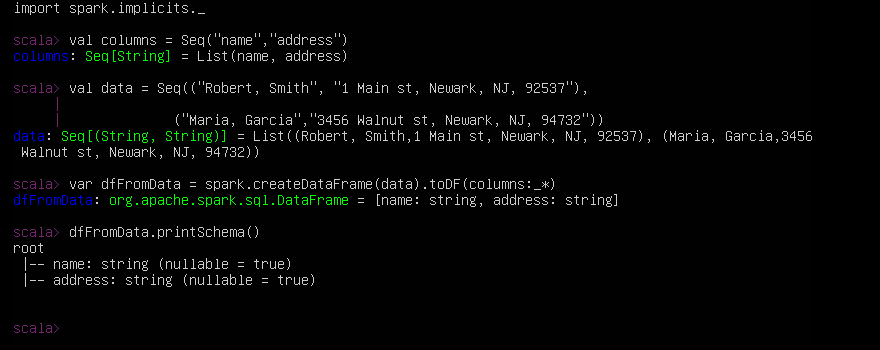
1. Đổi tên tên cột

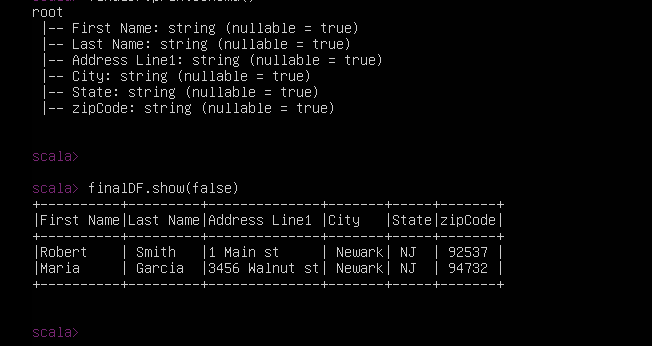


1. Thả một cột

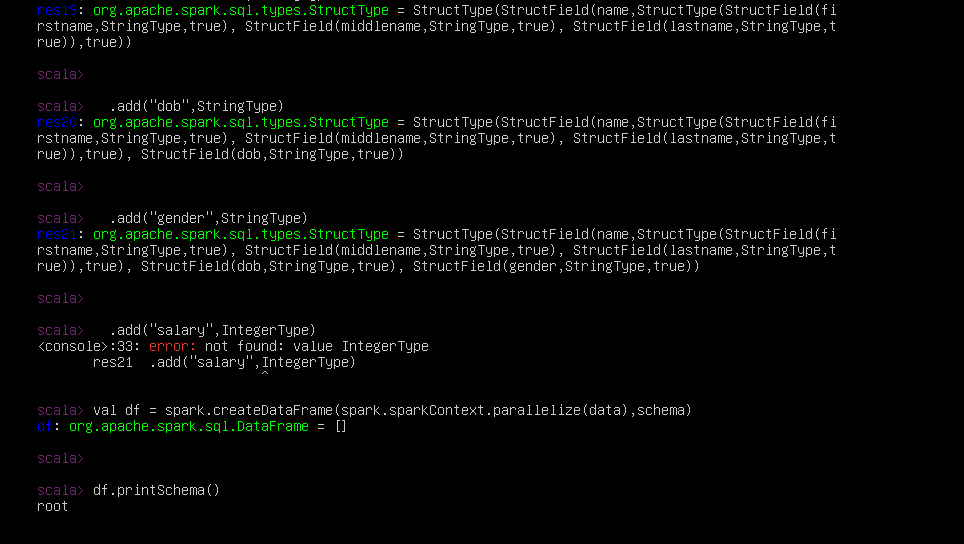


1. Chia cột thành nhiều cột

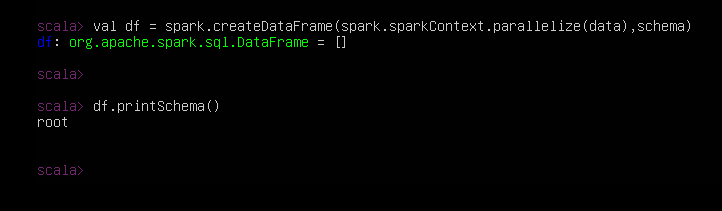




4. Đổi tên cột lồng nhau



Đỏi tên 1 cột



Đổi tên nhiều cột

