

Data Science Stack Exchange is a question and answer site for Data science professionals, Machine Learning specialists, and those interested in learning more about the field. Join them; it only takes a minute:

Sign up

### Here's how it works:

Anybody can ask a question

Anybody can answer

The best answers are voted up and rise to the top

## How to select particular column in Spark(pyspark)?

```
testPassengerId = test.select('PassengerId').map(lambda x: x.PassengerId)
```

I want to select PassengerId column and make RDD of it. But .select is not working. It says 'RDD' object has no attribute 'select'

apache-spark

pyspark

edited May 10 at 13:56



Sean Owen ♦  
2,773 2 12 30

asked Jan 3 at 2:10



dsl1990  
1

You can access columns pandas-style using DataFrame notation. – Emre Jan 3 at 4:34

## 2 Answers

'RDD' object has no attribute 'select'

This means that `test` is in fact an RDD and not a dataframe (which you are assuming it to be). Either you convert it to a dataframe and then apply `select` or do a `map` operation over the RDD.

Please let me know if you need any help around this.

answered May 18 at 9:52



Shagun

523 3 20

---

Assuming you have an RDD each row of which is of the form `(passenger_ID, passenger_name)`, you can do `rdd.map(lambda x: x[0])`. This is for a basic RDD

If you use Spark `sqlcontext` there are functions to select by column name.

answered May 18 at 11:11



Hrishikesh Ganu

337 9