

1. [Apache Spark Professional Training with Hands On Lab Sessions](#)
2. [Oreilly Databricks Apache Spark Developer Certification Simulator](#)

# SPARK API HANDSON LAB REMOVE HEADER FROM CSV FILE AND MAP TO EACH COLUMN ALSO SAVE OUTPUT TO HDFS AS TEXT FILE

By [www.HadoopExam.com](http://www.HadoopExam.com)

**Note: These instructions should be used with the HadoopExam Apache Spark: Professional Trainings.  
Where it is executed and you can do hands on with trainer.**

1. Hadoop Training
2. Spark Training
3. HBase Training
4. MapR Developer
5. MapR HBase
6. CCA500 Certification
7. Spark Certification
8. EMC Data Science

**Hadoop Specialization offer == 50% + 35% off**

**Hadoop Expert**

~~52000INR ==~~ 16900INR Only  
~~\$1150 ==~~ \$373 Only  
**Hadoop Specialization offer**

\* @ End of the Offer Prices will increase by 25%

**Limited Time Offer (Less Than 5Days Remain)**



**Cloudera CCA175 (Hadoop and Spark Developer Hands-on Certification available with total 75 solved problem scenarios. Click for More Detail)**

**Problem scenario:** You have given a file named spark6/user.csv.

**user.csv**

**id,topic,hits**

Rahul,scala,120

Nikita,spark,80

Mithun,spark,1

myself,cca175,180

Now write a Spark code in Scala which will remove the header part and create RDD of values as below, for all rows. And also if **id** is "myself" then filter out row.

**Map(id -> om, topic -> scala, hits -> 120)**

**Solution:**

**Step 1:** Create file in hdfs (We will do using Hue). However, you can first create in local filesystem and then upload it to hdfs.

**Step 2:** Load user.csv file from hdfs and create PairRDDs

```
val csv = sc.textFile("spark6/user.csv")
```

**Step 3:** split and clean data

```
val headerAndRows = csv.map(line => line.split(",").map(_.trim))
```

**Step 4:** Get header row

```
val header = headerAndRows.first
```

**Step 5:** Filter out header (We need to check if the first val matches the first header name)

```
val data = headerAndRows.filter(_(0) != header(0))
```

**Step 6:** Splits to map (header/value pairs)

```
val maps = data.map(splits => header.zip(splits).toMap)
```

**Step 7:** Filter out the user "myself"

```
val result = maps.filter(map => map("id") != "myself")
```

**Step 8:** Save the output as a Text file.

```
result.saveAsTextFile("spark6/result.txt")
```

1. [Apache Spark Professional Training with Hands On Lab Sessions](#)
2. [Oreilly Databricks Apache Spark Developer Certification Simulator](#)

# Spark Professional Training with Hands on Lab Session

<http://www.HadoopExam.com>

HadoopExam Learning Resource provides the following material for the Advanced Technologies.  
Please visit [www.HadoopExam.com](http://www.HadoopExam.com) for more detail this is just a few products from portfolio.

Price start for training with Just \$79/3500INR



Apache Spark  
Professional  
Training with  
HandsOn Session

+ Certification  
Material



Hadoop Professional  
Training with  
HandsOn Session

+ Certification  
Material



HBase Professional  
Training with  
HandsOn Session

+ Certification  
Material



Certification  
Material



Certification  
Material



Certification  
Material



Certification  
Material



Certification  
Material



Microsoft Azure

Certification  
Material



Certification  
Material