

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

# SPARK API HANDSON LAB USE JOIN, GROUP BY, SWAP FUNCTION AND 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 files as below.

spark5/EmployeeName.csv (id,name)

spark5/EmployeeSalary.csv (id,salary)

<u>EmployeeName.csv</u>	<u>EmployeeSalary.csv</u>
E01,Lokesh	E01,50000
E02,Bhupesh	E02,50000
E03,Amit	E03,45000
E04,Ratan	E04,45000
E05,Dinesh	E05,50000
E06,Pavan	E06,45000
E07,Tejas	E07,50000
E08,Sheela	E08,10000
E09,Kumar	E09,10000
E10,Venkat	E10,10000

Now write a Spark code in Scala which will load these two files from hdfs and join the same, to produce the (name, salary) values.

And save the data in multiple file group by salary (Means each file will have name of employees with same salary). Make sure file name include salary as well.

---

**Solution:**

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

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

```
val name = sc.textFile("spark5/EmployeeName.csv")  
val namePairRDD = name.map(x=> (x.split(",")(0),x.split(",")(1)))
```

**Step 3:** Load EmployeeSalary.csv file from hdfs and create PairRDDs

```
val salary = sc.textFile("spark5/EmployeeSalary.csv")  
val salaryPairRDD = salary.map(x=> (x.split(",")(0),x.split(",")(1)))
```

**Step 4:** Join all pairRDDs

```
val joined = namePairRDD.join(salaryPairRDD)
```

**Step 5:** Remove key from RDD and have Salary as a Key.

```
val keyRemoved = joined.values
```

**Step 6:** Now swap filtered RDD.

```
val swapped = keyRemoved.map(item => item.swap)
```

**Step 7:** Now groupBy keys (It will generate key and value array)

```
val grpByKey = swapped.groupByKey().collect()
```

**Step 8:** Now create RDD for values collection

```
val rddByKey = grpByKey.map{case (k,v) => k->sc.makeRDD(v.toSeq)}
```

**Step 9:** Save the output as a Text file.

```
rddByKey.foreach{ case (k,rdd) => rdd.saveAsTextFile("spark5/Employee"+k)}
```

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