# Using Spark-Scala complete following tasks

## Preparation

1. Use [IntelliJ IDEA + sbt](https://www.jetbrains.com/help/idea/getting-started-with-sbt.html) guide to set up an sbt project.
2. Use appropriate sbt configuration:

*scalaVersion*:= "2.10.5"​  
​  
libraryDependencies ++= *Seq*(​  
   "org.apache.spark" %% "spark-core" % "1.6.0",​  
 "org.apache.spark" %% "spark-sql" % "1.6.0",  
 "org.apache.spark" %% "spark-hive" % "1.6.0",

"com.databricks" %% "spark-csv" % "1.5.0"  
**)**​

Guide on reading csv files here: <https://github.com/databricks/spark-csv/blob/master/README.md>

Execution time measurement topic here: <http://biercoff.com/easily-measuring-code-execution-time-in-scala/>

Provided data files:

* SacramentocrimeJanuary2006.csv
* ucr\_ncic\_codes.tsv

## Task:

Answer following question using RDD and DataFrame (2 answers on each question).

Measure time for each solution (this might be done in standalone mode).

For DataFrame solutions write data into Hive (this requires creating jar and uploading it to VM).

## Tasks submission:

1. IDEA project with all code
2. Screenshot(s) of the execution on Virtual Machine. (you can use SELECT \* FROM <hive\_table> for DataFrame solutions to print results on screen)
3. Document with Conclusions.

**Expected output format:**

Task 1.1. Result: …

Task 1.2. Result: …

**To create logs-free output you can either:**

1. [Reduce Spark logging](https://stackoverflow.com/questions/27781187/how-to-stop-messages-displaying-on-spark-console)
2. Save results and write them only when all calculations are done.

## Question 1:

Print unique districts sorted by name (contents of a district column).

RDD:

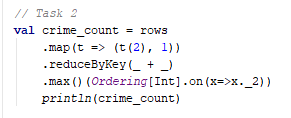




## Question 2:

In which district most crimes occurred?

RDD:





## Question 3:

What is most dangerous time of day?

NOTE:

Morning 6:00-12:00

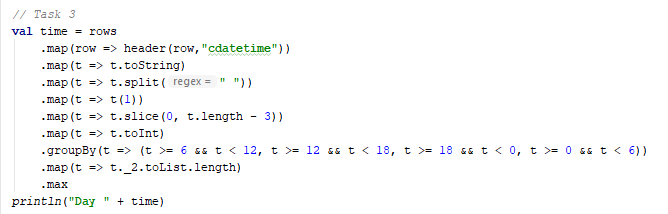
Day 12:00-18:00

Evening 18:00-00:00

Night 00:00-6:00

You might want to split cdatetime column into 2 different for this task

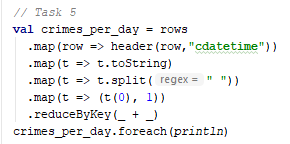
RDD:

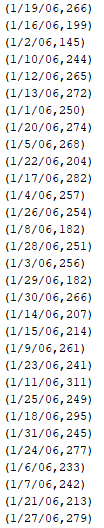




## Question 5:

Count crimes per day.





## Question 6:

Count crimes per day per Level of crime.

NOTE: Crime levels can be found in ucr\_ncic\_codes.tsv

## Question 7:

What are 5 most occurring Crime Category Codes?

NOTE: Crime Category Codes can be found in ucr\_ncic\_codes.tsv

## Question 8:

What are 3 drugs most involved in crimes? (names and count)

NOTE: Drugs can be found in ucr\_ncic\_codes.tsv

## Conclusion:

Speculate on results and provided solutions. You might want to touch these topics:

1. Was there execution time difference? Why? Why not?
2. Ease of use, readability. Why?
3. Difficulties, problems e.t.c