Data Engineering Assignment

Requirements

- Maven 3.x
- Java >= 1.8 or Scala 2.11

Task Description

You are given a multi-module maven project consisting of two modules, *hadoop-mr* and *spark*.

Additionally you are given **data files in csv format** which are located within the *datasrc* folder under the project's root directory, each row contains one record, separated by a newline (\n), the fields are separated by comma (,).

c5dfa197-09b9-4606-8d8f-c4732bc2d8d7.csv
consists of data from our measurement scripts and has the following schema:

```
00  MainDomainCode,
01  AdFormatCode,
02  AdDetectionFlag,
03  ProjectID,
04  AdCampaign,
05  AdPlacement,
06  AdCreative,
07  AdZone,
08  AdSite,
09  IframeFlag,
10  RequestDayMonthYear,
11  AdSize,
12  RequestTimestampMin,
13  RequestTimestampMax,
14  ImageWidth,
15  ImageHeight,
16  ImageArea,
17  AdAverageVisibleArea,
18  TotalAdAreaSize,
19  VisibleAdAreaSize,
20  TotalPageAreaSize,
21  VisiblePageAreaSize
```

domain_codes.csv contains a mapping between MainDomainCode and MainDomainValue:

```
00 MainDomainCode
01 MainDomainValue
```

(columns from left to right)

Your task is to calculate the **Average AdSize per MainDomainValue**, therefore the two data files have to be joined somehow.

You can choose whether you want to write a Hadoop MapReduce Job (using ToolRunner) in Java or implement a Spark Job in Scala.
You do not need to have hadoop or spark installed on your machine but Java 8 / Scala 2.11 is required to be able to run the code.

For the hadoop-mr task you can extend net.meetrics.assignments.dataengineer.mapreduce.AbstractTool, as it includes some basic Logging configuration.
After running maven (clean package) you can excute your program via

```
java -cp target/hadoop-mr.jar package.name.yourClass
```

For the spark task you can implement the trait net.meetrics.assignments.dataengineer.MeetricsSparkApp, which already defines a SparkConf and SparkContext, and excute your programm via

```
scala -J-Xmx1g -cp target/spark.jar package.name.yourClass
```

Results

Please hand in your code **and** your results (and necessary documentation), compressed with a common compression format (e.g. tar.gz).