QuickStart

To get started and execute the data application, follow these simple steps:

```
# cd ~
# unzip ds-shorsman.zip
```

<copy heckle.tar.gz and jeckle.tar.gz logfiles into local ds-shorsman/data directory>

```
# cd ds-shorsman/bin
# ./task0_workflow.sh (select Y to install R libraries)
# ./task1_workflow.sh
# ./task2_workflow.sh
# ./task3_workflow.sh
```

Results are saved into ds-shorsman/output directory:

```
# cd ../output
# ls -l
total 17748
-rw-rw-r-- 1 stuart stuart   272748 Sep 19 19:47 Task1Solution.csv
-rw-rw-r-- 1 stuart stuart 17575311 Sep 19 19:47 Task2Solution.csv
-rw-r--r-- 1 stuart stuart   322391 Sep 19 19:47 Task3Solution.csv
```

R scripts for all analysis can be found in ~/ds-shorsman/Rscripts. These scripts require certain CRAN libraries which will require manual installation by the user.

```
# cd ../Rscripts
# ls -l
total 52
-rw-rw-r-- 1 certification certification  8240 Sep 24 13:14
psFunctions.R
-rw-rw-r-- 1 certification certification  1193 Sep 24 13:14
Task0Analysis.R
-rw-rw-r-- 1 certification certification  6184 Sep 24 13:14
Task1Analysis.R
-rw-rw-r-- 1 certification certification  6893 Sep 24 13:14
Task1Solution.R
-rw-rw-r-- 1 certification certification  5818 Sep 24 13:14
Task2Analysis.R
-rw-rw-r-- 1 certification certification 10592 Sep 24 13:14
Task3Analysis.R
```

CCP Data Product

The data product is made up of the following:

./bin – contains the binaries required for running the data product.

./output – contains the output created by the binary files.

./docs – contains instructions, abstracts and general information about the data product.

./data – contains configuration files and initial heckle and jeckle logfiles.

./lib – contains the libraries required by the binaries for running the data product.

./libjars – contains 3rd party libraries required by ../lib/task*.jar files.

./Rlibs – contains R libraries required by the data product.

./Rscripts – contains R scripts and analysis from the data product.

./hql – contains Hive HQL scripts required by the data product.

./src – contains all the src and test files used for the data product.

./client – contains the Cloudera Machine Learning binaries and libraries.

./mahout – contains the Mahout distribution.