Learning Paths
Training Courses
Certification
CCP: Data Scientist
Hadoop Developer CCDH
Hadoop Admin CCAH
HBase Specialist CCSHB
Online Resources
Private Training
Training Partners

CCP: Data Scientist Challenge One Solution Kit

Table of Contents

Welcome to the CCP: Data Scientist Challenge One Solution Kit. These tutorials walk you step-by-step through a solution based on the 2013 CCP: Data Scientist Web Analytics Challenge: Classification, Clustering, and Collaborative Filtering.

Each section is designed to take you start to finish through a data science project and introduce you to the type of projects you'll encounter on your way to CCP: Data Scientist.

There are five sections that begin with exploring, then cleaning, and finally analyzing web log data. You will work through some of the common issues a data scientist encounters with log data and with data in JSON format. You will develop an alternate approach to the problem of building a classifier that takes advantage of the structure of the data to create a more accurate classifier. You will learn how to make use of a tool like Cloudera ML to discover clusters within a data set. Finally, you will learn to select an optimal recommender algorithm and extract ratings predictions from the recommender algorithm using Apache Mahout.

These tools and techniques are typical skills employed in data science: cleaning, modeling, selecting an algorithm, and tuning the parameters.

For these tutorials, we have kept the data set small enough to work with on a laptop and move through them quickly. The data set is specifically selected to have the same characteristics as the original data set used in the CCP: Data Scientist Web Analytics Challenge: Classification, Clustering, and Collaborative Filtering. The approaches and techniques presented in these tutorials are designed to perform well at scale, and for those who did participate in that challenge, you will find the process applies equally to the challenge data set.

CCP: Data Scientist Challenge One Solution Kit components

- A Hadoop cluster running in pseudo-distributed mode on a virtual machine image (VM), providing you all the software tools you need to complete the tutorials.
- a data set (on the VM).
- solution code (all of the code used in the tutorials).

Working with the VM

The solution kit comes with a virtual machine that is preconfigured and preloaded with everything you need to be able to complete the exercises. The VM is available for both VirutalBox and VMware.

- Download the VirtualBox version (3GB)
- Download the VMware version (3GB)

VM notes:

- The VM is a 64-bit VM. It requires a 64-bit host OS and a virtualization product that can support a 64-bit guest OS.
- The VM uses 4 GB of total RAM. The total system memory required varies depending on the size of your data set and on the other processes that are running.
- To use the VMware VM, you must use a player compatible with WorkStation 8.x or higher: Player 4.x or higher, ESXi 5.x or higher, or Fusion 4.x or higher. Older versions of WorkStation can be used to create a new VM using the same virtual disk (VMDK file), but some features in VMware Tools won't be available.

To install and configure the virtual machine from the Solution Kit, please follow these steps:

1. If you do not already have VirtualBox or VMware installed, download and install one of them.

- VirtualBox: http://virtualbox.org/wiki/Downloads
- VMware Player (not for MacOS X): https://my.vmware.com/web/vmware/free#desktop_end_user_computing/vmware_player/6_0
- VMware Fusion (free trial only for MacOS X or paid): https://my.vmware.com/web/vmware/info/slug/desktop_end_user_computing/vmware_fusion/6_0

2. Download the virtual machine.
3. Unpack the virtual machine. If you do not have a utility installed that can unpack a 7-Zip file, you can download a utility here: http://www.7-zip.org/
4. Import the virtual machine:
   - VirtualBox instructions: https://www.virtualbox.org/manual/ch01.html#ovf
   - VMware Player instructions: http://pubs.vmware.com/workstation–10/index.jsp?topic=%2Fcom.vmware.ws.using.doc%2FGUID–79B88EBD-DFA4–4C09-B33D-E011AABCA8D4.html
   - VMware Fusion instructions: http://pubs.vmware.com/fusion–5/index.jsp#com.vmware.fusion.help.doc/GUID-C5B1CDE2–5E09–4E8E-AD2E–15B62380379B.html
5. It is recommended that you configure the virtual machine to have at least 2GB of RAM and 2 processors. The more memory and processors you can make available to the virtual machine, the better it will perform.
   - VirtualBox instructions: https://www.virtualbox.org/manual/ch03.html#settings-system
   - VMware Player instructions: http://www.vmware.com/pdf/vmware_player40.pdf
   - VMware Fusion instructions: http://mylearn.vmware.com/mgrreg/courses.cfm?ui=www&a=det&id_course=56833
6. Start the virtual machine. After the virtual machine starts, you will see a desktop screen with a web browser displayed.

Helpful Tips:

- The username for the primary account in *cloudera*, and the password for that account is *cloudera*.
- The *cloudera* user has permission to run the sudo command, so separate *root* account credentials are not needed.
- To open a terminal window, right-click on the desktop (not in the browser) and select *Open in Terminal* or click on the *Applications* menu at the top of the desktop and select *System Tools* > *Terminal*.
- To open a file editor, click on the *Applications* menu at the top of the desktop and select either *Accessories* > *gedit Text Editor* for a simple text editor or *Programming* > *Geany* for a simple IDE environment.
- Many commands use the `$STREAMING` environment variable rather than long paths. The variable represents the path to the streaming jar file, which is usually located at `/usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-*.jar`. In the VM, the `$STREAMING` environment variable has been automatically set for you.

Working with the data

The sample data files are located in the cloudera user's home directory: `/home/cloudera/data`.

The data for this solution kit is provided in the form of a 7MB compressed archive that expands into 200MB of JSON log data spread across 20 files. (The original challenge used two 1.6GB archives, one for each of the Cloudera Movies server nodes, that expanded into 17GB of JSON log data spread across 68 files. This lab uses a smaller data set to reduce the time required to run the models.)

There is a more information about the data in the project description below.

Working with the command-line interface

The instructions below assume basic familiarity with the Linux command-line. If you want more information of any of the commands used and the options they take, please refer to the command's `man` pages. You access a man page for a command by running `man <command>` in the terminal.

Working with the solution code

As a convenience, all of the code used in this solution kit is provided for you in the `/home/cloudera/scripts` directory. Within that directory are subdirectories, one for each major section of the solution kit: exploring the data, cleaning the data, classifying the user, clustering the session, and predicting ratings. In each subdirectory you will find files containing the code from that section.

In the cases where a section calls for reopening a file and making modifications, the subdirectory will contain the original code and the results of applying the edits in separate files. The edited files will have `_editN` appended to the file name, where `N` is a number indicating which set of edits was applied. For example, if a section says to create a file called `foo.py`, then later says to modify it, and then later says to modify it again, the subdirectory for that section will contain `foo.py`, `foo_edit1.py`, and `foo_edut2.py`.

For the final section, predicting user ratings, the full Maven project is included. Rather than using the above strategy to include multiple edits within the source files directory, the edit naming scheme is applied to the full Maven project, i.e., you'll find a project in `recommend`, another in `recommend_edit1`, and another in `recommend_edit2`. This was required

to prevent build errors as Java has strict requirements regarding file names.

If you would like to make sure of the code in the scripts directory, it is recommended that you copy the code to your working directory and run it from there, rather than using the files in the scripts directory. This approach will be particularly helpful when working with files with multiple edits, as the command lines given in the solution kit assume that the file name does not change between edits. To copy a file from the scripts directory to your current working directory, use the cp command from the terminal. For example:

```
$ cp -f scripts/exploring/summary_map_edit3.py ./summary_map.py
```

That command will copy the fourth version (after three rounds of edits are applied to the original) of the file into the current working directory and rename it to the original file name. Note that this command will overwrite any `summary_map.py` file that already exists in your working directory.

These exercises build on one another, so data created in one might be used as input in those that follow. It's important that you finish each exercise correctly before continuing to the next one. If you make mistakes, or find yourself in trouble and cannot debug your work, you can simply start over and run the final script in each section to catch up.

Updating the provided code

The code provided in the scripts directory is part of a Github repository so that updates and changes can be pushed out as needed. The VM is configured to update periodically the local copy of the files.

If it becomes necessary, you can manually force a refresh of your local copy of the scripts directory. You can manually refresh the scripts directory very simply by opening a terminal, changing to the scripts directory, and issuing the `git pull origin master` command. For example:

```
$ cd ~/scripts
$ git pull origin master
From github.com:clouderacertifiedprofessional/datascientistchallenge1
  * branch            master     -> FETCH_HEAD
Already up-to-date.
```

Updates that are found will automatically be downloaded and applied. In the example above, no updates were found.
Note: It is strongly recommended that you do not make edits to the files in the scripts directory as doing so can result in conflicts with future updates.

Community Forum

This material is self-paced and intended for self-study. If you have questions or issues, visit the forum at community.cloudera.com in forum titled CCP: Data Scientist Challenge 1 Solution Kit.

Disclaimer

The material contained within these pages is instructional in nature and does not guarantee a passing score on any CCP: Data Scientist challenge project. Cloudera recommends that a candidate thoroughly understand the objectives for each challenge and utilize the resources and training courses recommended on these pages to gain a thorough understand of the domain of knowledge related to the role the challenge evaluates.

Navigation
Table of Contents
Solution Kit Introduction
Project Introduction **(next)**
Exploring the Data
Cleaning the Data
Classifying Users
Clustering Sessions
Predicting User Ratings (Building a Recommender)
Conclusion

Cloudera, Inc.
1001 Page Mill Road Bldg 2
Palo Alto, CA 94304

www.cloudera.com
US: 1-888-789-1488
Intl: 1-650-362-0488