Home

Cloudera Certified Professional (CCP): Data Scientist

2013 CCP Web Analytics Challenge: Classification, Clustering, and Collaborative Filtering

Time Remaining: 65 days, 15 hours, 39 minutes, 34 seconds

Welcome to the 2013 CCP Web Analytics Challenge! This site will give you everything you need to get started.

Contents

Challenge Background

Cloudera Movies is an up-and-coming video on demand service. The site has recently emerged from obscurity into a steep growth curve. As usage increases, the Cloudera Movies team is scrambling to keep up. The team has now doubled the size of the content delivery farm from 1 modest node to 2 high end nodes, with plans to bring additional servers online very soon. In addition to growing the hardware infrastructure, the Cloudera Movies team is also actively working to improve their software stack. Updates are pushed out frequently to fix bugs and add features, while trying to keep service disruption to a minimum.

In order to better plan how to grow their service, the Cloudera Movies team has brought you in to help them with some burning issues. While the software designers have designed the software store to events in JSON logs, saving you from digging through raw Apache web server logs, there's not much data beyond those application logs. What you have access to is the last four weeks of application log data, as is, for the two nodes in the web farm. The Cloudera Movies team is looking to you for help understanding their customers and how to grow their business. Don't let them down.

Challenge Details

The 2013 CCP Web Analytics Challenge: Classification, Clustering, and Collaborative Filtering is open from noon PDT on July 15th until midnight PDT on September 30th. You may begin the challenge at any time, but submissions received with a timestamp after midnight on September 30th will not be accepted.

This challenge consists of three parts, all using the same data set. The first part is a binary classification problem. The second part is sessionization. The third part is predicting user preferences. The challenge is to build an application or series of applications that can accept raw data and produce the expected results. Successfully completing all three parts of the challenge will require a solid working knowledge of techniques, approaches, and algorithms that are common in big data and data science.

While it is possible to complete the challenge using only local scripts and tools, submissions will also be evaluated against a scaleability criterion. To get the best score, submissions should take advantage of a clustered execution environment.

Submission Guidelines

Each submission should be a complete "data product," meaning that it should be a utility that accepts data in the formats specified for this challenge and produces results that conform to the **challenge results requirements**. (For this challenge, you may have a separate utility for each part of the challenge.) Cloudera must be able to run your utility (or utilities) in the Cloudera execution environment. The Cloudera execution environment will have all of the same components as virtual machine image, installed in the same locations. To facilitate running your data product, please include clear execution instructions.

Customer Portal Feedback

Have comments about this page?

Let us know!

Submissions will be scored against three criteria:

1. Accuracy: for each part of the challenge, each submission will scored against the known correct results:
   - Classification: submissions will get 1 point for each correct classification.
   - Clustering: submissions will get 1 point for each pair of points correctly placed into the same cluster.
   - Recommendations: submissions will be scored by RMSE.
2. Scaleability: each submission will be evaluated on how well the implemented solution performs against very large data sets and in a clustered environment.
3. Robustness: each submission will be evaluated in its tolerance for noisy or bad data inputs.

Submission Results Requirements

The results produced by each submission must conform to the following formats:

1. Classification – the **Task1Solution.csv** file has one line for every user of the format

   ```
   user_id, class
   ```

   where *user_id* is the numeric id for that user taken from the application log files, and *class* is 0 if the user is an adult or 1 if the user is a child.

2. Clustering – the **Task2Solution.csv** file has one line for every session of the format

   ```
   session_id,cluster_id
   ```

   where *session_id* is the alpha-numeric id for that session taken from the application log files, and *cluster_id* is a label for the cluster to which that session is assigned. The cluster labels themselves are unimportant. See the **submission guidelines** above for details on how solutions are scored.

3. Recommendations – the **Task3Solution.csv** file has one line for each predicted rating of the format

   ```
   user_id,item_id,rating
   ```

   where *user_id* is the numeric id for that user taken from the application log files, *item_id* is the numeric id for that item taken from the application log files, and rating is the predicted rating (between 1 and 5). **rateme.csv** contains the list of user-show combinations whose ratings will be scored.

What to Submit

Each submission should be a zip file or gzipped tar file. The file must the following:

- The three solution files called **Task1Solution.csv**, **Task2Solution.csv**, and **Task3Solution.csv**
- The source files for your data product
- Any binary files for running your data product
- Instructions for running your data product
- An abstract in **PDF format** that describes your approach to solving each of the three parts of the challenge. Please limit the length of the abstract to 5 pages.

How to Submit

Each submission should be uploaded to the **FTP drop box** using the HTML submission form. There is no automatic confirmation email, but will get a confirmation that your upload was received by the next business day.

Challenge File Downloads

- Challenge Data Files – these archive contain the application log files for four weeks of content serving from the two servers, *heckle* and *jeckle*. Logs are rotated daily and on restart of the web application servers. We know that the dataset is not completely clean, just as most datasets are not. As a data scientist, part of your job in this challenge is to handle issues such as this; we expect you to clean the data set, and you should feel free to install any additional tools on the VM which you feel will help you to complete the challenge.
  - **Data file archive 1**
  - **Data file archive 2**
- **Predicted Ratings** – this file lists the users and movies for which you should predict ratings.
- **Challenge Virtual Machine** – This virtual machine gives you a basic environment for the challenge. If you want to add other tools, or modify the environment to fit your way of working, you are completely free to do so. We do not provide support for customizing the environment, however, as we consider the ability to setup, customize, and maintain your working environment one of the tools of the data scientist's trade.
  - Virtual machine contents:
    - 64-bit CentOS 6.4 with 1GB RAM and a maximum hard disk capacity of 128GB

- CHD4.3, Impala, Hive, Pig, Apache Crunch, ClouderaML, R, Octave, Scala, Python, NumPy, SciPy
- Impala is installed and running on the VM when you boot. If you wish to stop Impala and improve the VM performance for other tasks, execute:

```
$ sudo service impala-state-store stop
$ sudo service impala-server stop
```

Useful Links

- **Download Cloudera Manager and Cloudera's Distribution including Apache Hadoop**
- **Apache Crunch**
- **Cloudera ML**
- **Data-Intensive Text Processing with MapReduce**
- **Mining of Massive Datasets**
- Challenge Forum *Launches July 28th*

Hints & Tips

1. Get started early! This challenge will take a significant amount of time to complete. Especially if you're doing it in your spare time, get started as early as you can.
2. Make sure your data product works against the raw, uncleaned data. If your product expects pre-cleaned data as input, it will not score well against the robustness criterion.
3. Make sure your data product is able to take full advantage of a clustered environment to score well against the scaleability criterion.
4. You can submit as many times as you like, but only your last submission will be scored.
5. The virtual machine is a great way to get started and to test your data product before submission. Cloudera's testing environment will be very similar to the VM's environment.
6. If you're having trouble understanding any part of the challenge, try searching the **challenge forum** or pasting a question there. *Launches July 28th*

Why Cloudera?
Hadoop & Big Data
Cloudera & Hadoop
Our Customers
FAQs
Blog

Products
Cloudera Enterprise Core
Cloudera Enterprise Free
Cloudera Manager
CDH
Latest Releases
All Downloads
Professional Services
Training

Solutions
For Your Industry
For Your Use Case
Partner Solutions

Partners
Resource Library
Support

About
Team
Events
Awards
Press Center
Careers
Contact Us

Follow us: Share:

Cloudera, Inc.
220 Portage Avenue
Palo Alto, CA 94306

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

©2012 Cloudera, Inc. All rights reserved | Terms & Conditions | Site Map
Hadoop and the Hadoop elephant logo are trademarks of the Apache Software Foundation.