Challenge Project Introduction

Machine-generated data is one of the primary data sources classically labeled as big data, and the log files generated by web servers and web applications are a significant source of modern machine-generated data. Locked within these log files is a wealth of information on user behavior and preferences. For an online retailer, unlocking that information is often a significant competitive advantage. For companies that sell online services, like online game developers or video on demand providers, the data in those log files represent their lifeblood. Understanding their users and predicting their users' needs and actions can be the difference between success and failure.

Cloudera Movies is a promising internet on-demand streaming video service. The site has recently emerged from obscurity and is experiencing rapid growth. As site use increases, the Cloudera Movies team is scrambling to keep up. The team has now doubled the size of the site's content delivery server farm from one modest node to two 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. The development team pushes updates 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 critical issues. The company has tasked you with creating a picture of the user base and building a recommendation engine that accurately models their customers' watching preferences. You must build a data product that solves for three data science problems presented to you by Cloudera Movies. First, based on only the log data provided, the Cloudera Movies legal team wants to understand which user accounts are used most often by younger viewers. Second, the product team wants you to segment sessions based on the actions that users take in order to improve the site's usability. Third, the product team wants a recommendation engine they can deploy to their site to help drive users to the content they will like, in an effort to increase time on site and reduce churn.

For your data, you have access to Cloudera Movies' raw application log files. Cloudera Movies has designed their system to store events in JSON logs. You have no other data to work with besides these application logs. You have access to the last four weeks of application log data for the two nodes in the Cloudera Movies web server 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.

Project Description

The project consists of three parts, all using the same data set.

1. A binary classification problem requiring you to label each user account as including to an adult or child.
2. A sessionization problem requiring you to group user sessions based on behavior.
3. A user prediction problem requiring you to build a recommender that will predict ratings for a series of user-item pairs.

Data Description

As mentioned in the introduction, 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.) The log files rotate at least once a day, as evidenced by the timestamps in the files. From an example line of the log data, you can see that the structure of the data is fairly simple:

```
{"created_at": "2013-05-08T08:00:00Z", "payload": {"item_id": "11086", "marker": 3540}, "session_id":
"b549de69-a0dc-4b8a-8ee1-01f1a1f5a66e", "type": "Play", "user": 81729334, "user_agent": "Mozilla/5.0
(iPad; CPU OS 5_0_1 like Mac OS X) AppleWebKit/534.46 (KHTML, like Gecko) Mobile/9A405"}
```

Data Source

The log files are generated by the Cloudera Movies's video on demand service. Users pay a subscription fee to access the Cloudera Movies content library. Once logged into the site, users are offered recommendations and several popular choices. Users can also play content from their queues or find content by searching (which they can play immediately or add to their queues). The site also includes ratings and reviews. While logged in, users can also perform account operations, like bill payment, password management, etc.

Data Details

In addition to the data files, some details about the data are also provided:

- The log files include parental control events which can be used to identify whether some accounts are being used by adults or kids. If an account enables or disables parental controls, you should label it according to the most recent state. When parental controls are enabled, other account management controls are disabled, and content access is restricted.
- In events that relate to content playback the *marker* field indicates the player app's position in the content.
- Ratings are from 1 to 5, with 5 being the highest.
- Content IDs with an 'e' in them are television shows. The 'e' indicates episode; the number following the 'e' indicates the episode number.
- When the content playback reaches the end, a 'stop' event is logged. If the user leaves the page or moves on to other content before the end, the 'stop' event may not get logged.
- All timestamps are local to the Cloudera Movies' servers. Users are mostly located in the United States.

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