Background

The customer wanted to develop a social networking site which connects users to like-minded businesses and activities. Understanding the current options in the marketplace, the customer believed the addition of a recommender engine would give their application significant competitive advantage.

However, with a launch date six months out, it would be hard to create a recommender engine because there was no customer data to analyze.

How could a data science company provide a solution with no data?

Mosaic Approach

With no real data on which to develop and test the recommender, the recommender models would need to be able to dynamically learn as data entered the system and to be easily tuned by analysts to better model observed patterns.

Mosaic determined that a hybrid *collaborative* and *content-based filtering* machine learning model would provide the necessary performance and flexibility. Content-based filtering makes use of known data about site users and "items" – businesses, advertisements, in-site product offers, etc. – to identify patterns in user preferences. For example, users with a specific set of demographic traits may tend to accept offers from businesses of a given type. These models are intuitive and can quickly begin generating relevant recommendations even for users that are new to the site.

Collaborative filtering uses behavioral patterns among the community of site users to generate recommendations. *User-based* collaborative filtering finds users with similar behavioral patterns and bases recommendations for one user on the observed preferences of similar users. For example, if user A tends to display the same preferences as users B, C, and D then any items that have not been seen by A and that have been favorably rated by B, C, and D are good candidates for recommendation to user A. *Item-based* collaborative filtering looks instead for similarities between how items tend to be rated by users. If item 1 tends to be rated similarly to items 2, 3, and 4 then item 1 is a good candidate to recommend to any user who has shown preference for items 2, 3, and 4 but has not seen item 1.

Mosaic's patent-pending machine learning model blended the three model types – content-based filtering and both user-based and item-based collaborative filtering – to create a recommender that is able to leverage the strengths of each approach while maintaining model stability. This allows the recommender to gradually adjust its recommendations as a user moves from being newly registered with only demographic data available to being an experienced user with a long history of interactions on the site. Mosaic built in a number of "dials" allowing site administrators and analysts to tune various model parameters: the relative weighting of different user behaviors in determining preferences, how the content-based and collaborative filtering models are blended as more user data becomes available, how user and item similarity are determined, etc. This will allow site administrators to refine the recommender engine's performance as more data is collected and to set up A/B tests comparing the impact of different parameterizations on the site's business objectives.

Mosaic Solution

Technical requirements from our customer dictated that our solution should be able to model the social

connections of 1 million users every six hours. We provisioned a cluster of four cloud-based Linux servers from RackSpace to run our model.

The model was implemented in a version of Mahout that we optimized to leverage specific aspects of our model and to meet the customer's specific performance objectives. We implemented an improved disk-caching scheme and maximized in-memory processing for improved performance. We also prepared our data to support sequential iteration of data elements for critical calculations rather than performing map lookups.

We were able to leverage the geographic nature of the problem based on the knowledge that user interactions were only computed within discrete geographic bounds. This allowed us to segment data into relatively small sets and distribute the data across processing nodes according to these geographic boundaries, thereby limited the amount of data processed by any single node in any one execution of our model.

For data integration between the customer data storage and our data modeling system, we designed database schemas for staged data and modeling results. We then implemented simple scripts allowing the customer to initiate our modeling process. This approach provided a simple, minimal contact approach to integration which proved to be highly reliable and fault tolerant.

The initial Mahout implementation of our model required more than four days to run on the Mahout cluster. The optimizations that we performed reduced the run time to less than two hours. While some data science firms might have resorted to increasing the number of servers in the cluster to process the data more quickly, drastically increasing costs to the customer, Mosaic Data Science used optimization to reduce the processing time, thus keeping recurring costs much lower without sacrificing the integrity of the modeling results.

It is hard to quantify the economic benefit as the site is just starting to go live. Mosaic plans to continue to monitor and tune the models as needed, allowing the site to recommend with even better accuracy, ultimately helping the application build a large user base, connect users to others with similar interests, connect users to business they would like, keep users active on the site, and use the large amount of data gathered on each user to target in-site advertising and increase the value to advertisers.

For More Information
Contact Mosaic today to learn how we can help your organization achieve equally dramatic returns on your data science investment!