Customer Case Study

MyFitnessPal

databricks

Customer Case Study

MyFitnessPal

Benefits

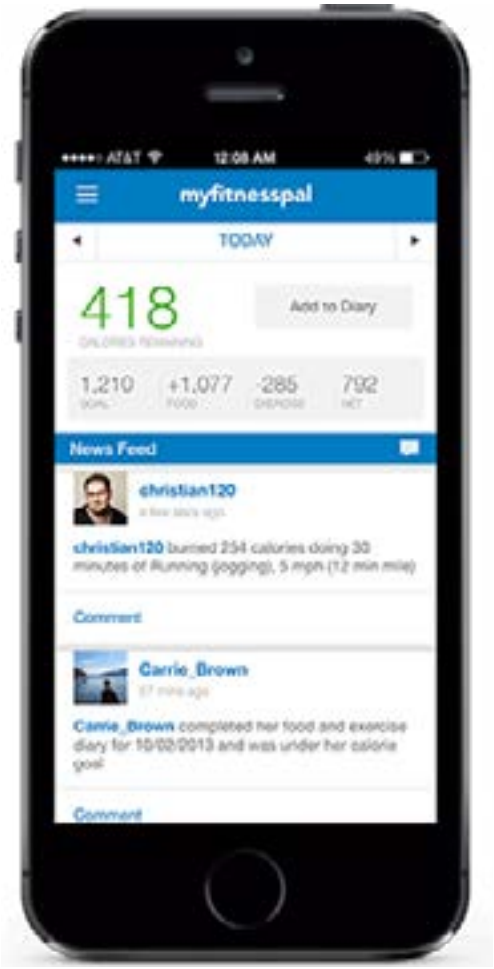
- Implementation of a new feature called “Verified Foods” to provide more accurate nutrition information.
- Ten-fold speed improvement over previous data pipeline implementation.
- Four times more projects completed in the past quarter resulting from an increase in team productivity.
- Improved team efficiency achieved through accessible advanced analytics and better code re-use.

Summary

- MyFitnessPal needed to deliver a new feature called “Verified Foods.”
- The “Verified Foods” feature demanded a faster pipeline to execute a number of highly sophisticated algorithms. Their legacy non-distributed Java-based data pipeline was slow, did not scale, and lacked flexibility.
- MyFitnessPal chose Databricks to harness the power of Apache® Spark™ and to build the data pipeline for “Verified Foods” to successfully deliver the feature to their users while gaining many additional benefits.

Business Background

MyFitnessPal aims to build the largest health and fitness community online by helping people achieve healthier lifestyles through better diet and increased exercise. As of the end of 2014, there are over 80 million members in the MyFitnessPal community worldwide. Members of the community can use the MyFitnessPal website or the smartphone app to track their diet and exercise patterns and use the information to help reach their fitness goals.



The data engineering and science team at MyFitnessPal build data products that facilitate healthy decision-making by end users. One of the most critical data products used by the MyFitnessPal community is the food database which helps people to quickly find and log everything they eat by simply entering the name of a food item, such as a granola bar from a certain brand type. In order to make the database even more simple to navigate, MyFitnessPal created a new feature called "Verified Foods."

> *"Databricks helped us deliver a new feature to market while improving the performance of the data pipeline ten-fold. We would not have been able to fully harness the power of Apache Spark to deliver the feature to market without Databricks."*
>
> – Chul Lee,
> Director of Data Engineering &
> Science at MyFitnessPal

Challenges

To populate the food database, MyFitnessPal relies on its community members to enter nutritional information of food items. While this crowdsourcing approach quickly grew the database to include more than 5 million items, MyFitnessPal needed to harmonize the variations in data caused by human data entry. MyFitnessPal developed a heuristic to accomplish this harmonization by gleaning signals from the data set — signals such as having multiple members enter the same information for a food item. Once sufficient signals have been detected for a food item, its nutritional information is considered "verified." This food item is then labeled as such in the MyFitnessPal application, allowing members to track nutritional information of "Verified Foods" effortlessly.

Developing the verification process for the food database required MyFitnessPal to solve three key challenges:

- **Terabyte-scale data volume:** There are over 5 billion unique items in over 15 billion user-generated food log entries. These datasets are also rapidly increasing with the growth of the community and the member's levels of engagement.
- **Need for fast processing performance:** Nutritional information in the food database must be verified as quickly as possible because members of the MyFitnessPal community would reduce their engagement if the food database cannot be trusted.
- **Diverse and complex analytics algorithm needs:** As part of the verification process, the member-input data needed to be normalized (e.g. removal of stop words, lower-case conversion), de-duplicated, and aggregated by a wide array of machine learning algorithms.

Prior to choosing Databricks to address these challenges, MyFitnessPal attempted an implementation based on a non-distributed Java-based application first, and an Apache Hadoop™ based solution after that. These initial attempts proved to be neither fully scalable nor fast enough: Running the algorithm against the enormous dataset took multiple weeks to complete.

Solution

MyFitnessPal wanted to replace their previous data pipeline with Apache Spark because of Spark's ability to seamlessly integrate different data sources, the availability of data processing libraries within MLlib and GraphX, fast performance to avoid slow table joins, and being able to significantly speed up operations that could be parallelized in a distributed fashion.

However, building a Spark-based data pipeline is not a trivial task. MyFitnessPal chose the Databricks platform, a hosted end-to-end data platform powered by Spark, to seamlessly transition from data ingest through exploration and production.

The data pipeline built with Databricks consists of the following components:

- **Zero management Spark Clusters:** MyFitnessPal team used the simplified management UI in Databricks to instantly create, scale up, and teardown Spark clusters as needed to support its needs, without dedicated DevOps or data scientists to maintain infrastructure.
- **Interactive workspace:** Data scientists used the Databricks interactive workspace to explore data and develop advanced machine learning algorithms in an iterative fashion. The multi-user notebook development environment enabled the entire team to collaborate and document their process seamlessly.
- **Automated production job scheduler:** MyFitnessPal team used the Jobs feature to seamlessly transition from exploration / development to production by simply running the notebooks they developed during exploration with the job scheduler. Jobs are scheduled to run at predetermined intervals, and the job scheduler also automatically monitors the job progress, alerting the team members if anything goes wrong.

MyFitnessPal uses Databricks today to deploy multi-terabyte Spark clusters in production settings to support business-critical data pipelines.

Benefits

MyFitnessPal has been able to deliver a new feature to their member community called "Verified Foods," as a result of the algorithms and the production pipeline developed with Databricks. Because of the high reliability and fast performance of the data pipeline powered by Databricks, the "Verified Foods" database now includes most items whose nutritional information is readily available in the market — such as packaged items.

In addition to powering the "Verified Foods" feature, Databricks also delivered a number of key benefits to the Data Engineering & Science team at MyFitnessPal:

- **Ten-fold speed improvement** to run the "Verified Foods" algorithms against the entire dataset, reducing the processing time from weeks to mere hours.
- **Four times more projects completed in the past quarter** due to higher team productivity. Working on new features is much easier with the interactive workspace and easy cluster management.
- **Improved team efficiency** with more thorough and advanced analysis. This efficiency is driven by the availability of mature libraries in Spark, and the transition speed, because the same codebase could be used for data exploration, production, and BI instead of rewriting code for different needs.

> *"The 'Jobs' feature allowed my team to turn code developed in small experiments into full-scale production pipelines with a few clicks. Today, Databricks powers our entire production pipeline with multi-terabyte Spark clusters."*
>
> – Chul Lee,
> Director of Data Engineering & Science at MyFitnessPal

Evaluate Databricks with a trial account now.

databricks.com/try-databricks