arm TREASURE DATA

Vertical Study: Dating Services

Integrating every data source helps dating services improve rentention

Dating services live and die on the quality of their recommendations. Since beauty is in the eye of the beholder, a great way to improve match quality is by performing cohort analysis according to acquisition channels; even down to specific campaigns. This allows dating apps to serve up pre-qualified sets of active profiles to new users based on the campaign they responded to, dramatically improving both the initial experience and long term retention. But it's only possible when all your data is integrated and accessible in a single source of truth.

The data lake design paradigm has many other benefits for dating apps, like enabling easier bot detection, or determining API rate limit thresholds to protect against DDoS attacks without affecting power users. For example, unifying conversation logs alongside user tables and server IP lists allows dating apps to train models on typical message latencies, using previously identified bot accounts to match messaging patterns in more sophisticated agents. On a more fundamental level, it allows these apps to quickly search across all conversations to detect common phone numbers or URLs being distributed by multiple accounts and flag these accounts for review. Here again, the accessibility of multiple data sources to a common query engine is key.

Dating services live and die on the quality of their recommendations.

This is a non-trivial endeavor for production scale apps. There are messaging logs, API server logs, Mobile and Web clickstreams, HTTP server logs, MySQL or Postgres user tables, and marketing automation databases which all must be ingested and harmonized before the first query can be run.This usually calls for countless ETL scripts to be written and monitored, NoSQL clusters to be provisioned and loaded, virtual query engine layers to instantiate, and a lot of trial-and-error schema management.

Treasure Data Speeds Time to Market

Treasure Data has already helped several dating services achieve all of the above in less than 14 days with our hosted data lake environment. Integrating our extremely popular open source data collection tools into a fully supported end-to-end architecture, we provide the infrastructure to quickly ingest data from any source, store it in schema-on-read form, and immediately begin building queries in the popular Hive, Presto or Pig engines. We've also pioneered the open source Hivemall machine learning library, to enable teams to build features and train models on their entire dataset using familiar SQL statements. The results of any query or training batch can be easily exported to visualization or BI tools, Amazon S3 storage, Google Sheets, or almost any schemaful Data Warehouse for compute intensive analysis, with automatic type conversion provided seamlessly by our output connectors.

In addition to bulk and streaming data ingestion agents, we provide a wealth of API connectors to automatically pull raw data from services like Marketo, Salesforce, and Mixpanel. We also provide hosted Postgres Data Marts, to quickly surface the results of regularly scheduled queries in the world's most popular production SQL language. This allows dating services to maintain near real-time spam lists based on their detection workflows, providing a blacklist table that can be queried by their production apps at the beginning of each session. Our columnar storage technique ensures these lookups will return 50% faster than standard Postgres, providing minimal latency impact to real users.

Building on all of these amazing collection and analysis features, our dating service customers have successfully productionized the following workflows:

Recommendations Tailoring

By enabling performant analysis of thousands of likes and dislikes alongside marketing automation data and profile features, dating services are able to create micro-segmented user cohorts to improve recommendations. Several platforms that allow for profile "favoriting" continuously monitor the mean geographical distance between a user and the profiles they are favoriting to determine their intent in dating. The underlying assumption is that a user who frequently favorites profiles far away is strongly biased towards attractiveness and is likely to prioritize hookups, while another user who stays closer to home is more likely to be seeking long term companionship. Setting these flags on their user models allows these services to more frequently match users with similar dating priorities, increasing overall satisfaction with the service. It also allows them the increase engagement with push notifications to a user when a far-away favorite travels to their area.

Custom Analytics

With complete access to all their raw data, dating services are able to build robust analytics that make sense for their business. Many prefer to combine clickstreams with API data to measure latency, and segment results by geographic region to determine which areas to focus marketing spend and retention efforts. They also closely monitor the average number of profile blocks per user on a regional, cohort, and system wide level, as this is a leading indicator of general dissatisfaction with the service. Higher than average number of blocks in specific geographical regions can also alert them to a heavy bot presence in these areas, allowing them to focus manual attention on reviewing those accounts. Joining their conversation logs with user tables allows their in-house moderators to quickly pull and review chat transcripts of offending users and manually identify bots or swiftly enforce code of conduct rules on the platform.

Spam Prevention

Bot detection is the number one issue facing dating services today, and many turn to big data to tackle it. Some of our active users in this vertical came to us with as many as 5 bot accounts to each active user. After integrating all their data on our platform, they were able to quickly remove as much as a third of these accounts using the techniques described above of searching for phone numbers and URLs distributed by more than one account. Converting these message logs into time-series event data has allowed them to train pattern detection models and remove a large number of more sophisticated bots as well. After fine-tuning these queries, they now use Treasure Data's batch scheduling tools to automate these workflows and deactivate new malicious accounts on a 24 hr basis.

About Treasure Data

By simplifying the process of collection, storage and analysis of massive quantities of data, businesses using Treasure Data get value from big data in days, not months. Now anyone can easily and economically process enormous amounts of data in near real-time with no infrastructure setup or management. Our hosted data lake supports data of any type, including multi-structured web, application, mobile and sensor data. It integrates seamlessly with your existing data management and BI environments through our many output connectors and plug-and-play relationships with popular tools, along with our ODBC/JDBC driver support for more custom integrations. Similar to MUJI, many companies use Treasure Data to quickly and economically add new capabilities to their existing infrastructure. Our hundreds of satisfied customers include Pebble, Pioneer, Gree and several other global Fortune 500s.

Learn more at TreasureData.com

arm TREASURE DATA

+1.866.899.5386
info@treasuredata.com
2565 Leghorn St.
Mountain View, CA 94040