clearscale

aws partner network
Premier Consulting Partner

The Big Data Infrastructure that Powers The Globe and Mail's Article Recommendation Engine

THE GLOBE AND MAIL

Executive Summary

The Globe and Mail is Canada's #1 weekday and weekend newspaper. In print for 170 years, *The Globe* delivers lively and authoritative coverage of national, international, business, technology, arts, entertainment, and lifestyle news. Millions of weekly readers access *The Globe* on the web or with their mobile device.

The Globe has been innovative in its adoption of technology, enabling them to better understand their readership, offer engaging content, and expand advertising opportunities. Recently, *The Globe* launched a new encrypted system called SecureDrop to offer protection to whistleblowers and anonymous sources. *The Globe* was also the first Canadian publisher to offer MOAT analytics to measure ad viewability, exposure, and interactivity, which enables advertisers to spend and track their advertising dollars more effectively.

> "I was impressed with ClearScale's ability to deliver such a complex yet robust, custom system, within our tight timelines. They are extremely capable — we chose the right vendor."
>
> **Michael O'Neill,** Director of Data Science, *The Globe and Mail*

The Challenge

In May 2015, *The Globe* was launching a new mobile application for readers to access the newspaper's stories and breaking news on their mobile devices. They decided to take this opportunity to introduce a recommendation system, which would present a personalized selection of articles to readers and update in real time. Their Data Science team had been prototyping a number of recommendation algorithms in Amazon, but given that *The Globe* had not built such a production system in the cloud before, they sought outside firms that had expertise in this area. They needed a partner that could architect a robust real-time production infrastructure in AWS, and do so within tight timelines.

Their existing data center did not have the infrastructure and services options that were required to implement real-time processing. They decided that Amazon was the right cloud provider because AWS has the technology and Big Data services that would enable them to rapidly process a massive amount of data and immediately deliver back to the reader a set of personalized article recommendations. They created a PoC in Amazon to test out the idea.

The ClearScale Solution

ClearScale partnered with *The Globe* to build out the AWS infrastructure that powers the recommendation engine. ClearScale worked with *The Globe* to understand their requirements and to identify the right AWS services for the project.

The cloud infrastructure relies on Big Data tools from AWS including Kinesis, Elastic MapReduce, DynamoDB, and S3; automation from Chef, CloudFormation, and OpsWorks; as well as a custom data transfer application built by ClearScale.

The recommendation engine works at scale such that recommendations for the next set of articles are ready by the time the reader has loaded the current article. And depending on which article the reader subsequently clicks on, a new unique set of recommendations is generated, and so on. This helps readers stay engaged with the content they find most relevant.

At a high level, this is the basic technical flow:

1. Data is constantly aggregated from several sources, including user clicks and engagement with articles, article promotion, and article metadata;
2. That data is then processed, stored, and analyzed using AWS Big Data tools including Kinesis, DynamoDB, and Elastic MapReduce, as well as a custom data transfer app built by ClearScale, and custom algorithms that *The Globe* Data Science team produced;
3. Customized recommendations are delivered back to the reader within the article they just loaded.

Let's dig into each step and discuss how it works.

Collecting Real-Time, Crawl, and Batch Data

If you want to deliver dynamic, targeted content to your users, you must understand what your users are interested in and what content you have available.

With reader consent, *The Globe* collects data from a wide variety of sources to identify reader's interests and behavior. Real-time and batch data sources and feeds include:

- Clickstream data batch feeds from Adobe Omniture
- Clickstream data real-time feeds delivered via Kinesis
- Real-time data from MOAT delivered via Kinesis
- Batch advertising data from DoubleClick
- User registration and subscription information including demographics and stated interests saved in an Oracle database in their corporate data center

Real-time user behavior data is captured using custom web beacon applications that *The Globe* developed in-house. Real-time data is generated at several hundred records per second.

To discover what content is available, *The Globe* and Mail's Content Management System is synced with DynamoDB to discover and classify new content, typically at the rate of several hundred new articles per day.

Batch data is collected hourly throughout the day from Google Drive and the Adobe FTP server.

Data forking is used to split raw data up for production, development, and staging environments to create copies of the data without having to split it at the origin.

Processing the Data at Scale

Both Amazon Kinesis as well as a custom data transfer application built by ClearScale are used to collect and process the data at scale.

Kinesis producers read data from the web beacon log files, do preliminary cleaning and extraction, and send to the Kinesis stream. Kinesis consumers then pick up records from the stream for processing.

A Kinesis consumer is a Java application built with the Amazon Kinesis Client Library. Kinesis consumer applications run on EC2 instances and are scalable. The quantity of instances is optimized depending on the Kinesis provisioned throughput. Consumers read data from the stream, parse the data, transform it to the standardized JSON format, and asynchronously store the data to DynamoDB. ClearScale built custom Kinesis consumer apps for this project; there are several Kinesis consumers, one for each data type.

Crawl data is transferred and processed using SQS and a custom data transfer application built by ClearScale. The data transfer app then saves that file to S3, transforms that data to the standardized JSON format, and copies the updated file from S3 to DynamoDB.

Batch data from Google Drive and the FTP server is also transferred and processed by the custom data transfer app. There are checks in place to ensure that files are uploaded completely and that partially-uploaded files can't be transferred. The data transfer app follows the same process for batch data as described above for crawl data, but then copies the updated file from S3 directly to the Hadoop Distributed File System (HDFS). We'll talk more about HDFS in a second.

Storing the Data

DynamoDB is the hub of the Big Data infrastructure and is used as central storage for real time data:

- Kinesis consumer applications push real-time data into DynamoDB
- Custom data transfer application pushes crawler, user profile, and content metadata
- Elastic MapReduce reads that processed data and writes recommendations
- API tier reads information and delivers that to the user on the mobile application
- API tier writes the recommendations that were delivered and from which algorithm, so that the system's built-in A/B testing can be evaluated and algorithms can continually be enhanced by *The Globe* Data Science team

DynamoDB is a NoSQL database that can store and retrieve any amount of data with seamless scalability. DynamoDB automatically spreads the data and traffic for the table over a sufficient number of servers to handle the request capacity specified by the customer and the amount of data stored, while maintaining consistent and fast performance.

Crunching the Data to Deliver Intelligent Recommendations

With Amazon Elastic MapReduce, you can analyze and process vast amounts of data. EMR does this by distributing the computational work across a cluster of virtual servers running in the Amazon Cloud. The cluster is managed using Hadoop, a distributed processing architecture in which a task is mapped to a set of servers for processing. EMR is integrated with the Hadoop Distributed File System (HDFS) to help store large amounts of data over a distributed network with redundancy to protect against data loss.

EMR works with all the real time data in DynamoDB as well as the batch data sent to HDFS. EMR crunches the data using *The Globe*'s recommendation algorithms, which have the logic defining which articles to recommend and how to match available content with user interests. *The Globe*'s algorithms rely primarily on Spark and Mahout for machine learning, and the results are augmented in various ways that have proven to drive substantially increased relevance of the recommendations to readers.

EMR then sends the result back to DynamoDB, and the recommendations are available to be served at the next user interaction.

Adopting a DevOps Approach for Infrastructure Automation

ClearScale created templates for deploying each component of the infrastructure, which makes it more flexible, scalable, and redundant. Management of *The Globe*'s production, staging, and development environments on AWS is made easier with automation from Chef, Amazon CloudFormation, and AWS OpsWorks.

AWS OpsWorks and Chef are used to automate systems management, server configuration, system settings, and code deployment. There is a Chef cookbook for each instance type in the infrastructure, including cookbooks that define Kinesis producers and consumers, the front end app servers, and the article crawl, section crawl, and data transfer apps.

New server creation is automated with Amazon CloudFormation templates that create instances that are part of an auto-scaling array, including API, real-time data, crawl data, and batch data servers. The array is configured with autoscaling policies that monitor conditions such as CPU and memory utilization. CloudFormation eliminates the need to manually create servers.

The entire infrastructure is running in Amazon VPC for maximum security and control.

The Benefits

A number of off-the-shelf recommender systems were evaluated by *The Globe*, some of which they already have access to as part of their Digital toolsets subscriptions. They feel that none of the off-the-shelf recommender systems offered them anything near the potential they now have with their custom solution. Almost anything that *The Globe* discovers might boost performance of the recommender can easily be added to the recommender, and the lift in performance quickly measured.

In a time when online audiences have ever shrinking attention spans given the constantly evolving online choices for news and information, *The Globe* and Mail is now well positioned to make it easy for their online audience to discover a wealth of great *The Globe* and Mail content relevant to them without having to look very hard.

“I was impressed with ClearScale’s ability to deliver such a complex yet robust, custom system, within our tight timelines,” remarks Michael O’Neill, their Director of Data Science. “They are extremely capable — we chose the right vendor.”

While the recommender is currently only available in *The Globe*’s new mobile app, it is coming soon to their web and mobile web platforms, which is requiring zero changes to the system ClearScale built for them, apart from increasing some of the throughput settings and cluster sizes.