batch.ly

Social Sentiment Analysis using Hive on Hadoop is 82% cheaper with Batchly

HIGHLIGHTS

Use Case: Periodically calculate Dyadic interactions and Goldstein scale for a particular country on the GDELT dataset in a cost efficient manner

Technology & Dataset Used

Amazon EMR

Hive & HiveSQL

400GB GDELT dataset

Result per Run

- 500 instances
- 96% spot instance usage
- 81.8% savings over on-demand AWS costs
- No IT overhead

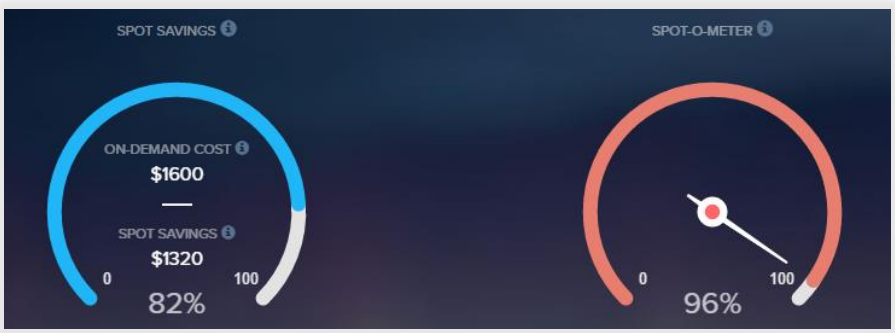
SENTIMENT ANALYSIS

Social Sentiment analysis is a powerful new tool being used by economists, social scientists and data analysts to help support new theories and descriptive understandings of the behaviours and driving forces of global-scale social systems. The GDELT Project monitors the world's news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, emotions, events driving our global society.

CHALLENGE

To accomplish the sentiment analysis goals, a Hive system had to be built to periodically query and monitor the dyadic interactions (i.e. communications between 2 individuals) and the Goldstein scale (i.e. the intensity of conflict or co-operation inherent in different types of international events).

Amazon EMR provides a managed service to run Hadoop clusters. However, the idea was to make this processing extremely cost efficient and repeatable with little to no IT overhead.



SOLUTION

The EMR cluster was initially configured in Amazon EMR management console. This cluster information was then used in Batch.ly, a solution that fully automates AWS infrastructure (including spot instances) provisioning and management to create a Hive job.

The GDELT dataset was loaded from S3 onto EMR and HiveSQL queries were scheduled to run everyday through Batchly to monitor the key social parameters. Batchly also provided autoscaling option for quick turnaround time for ad-hoc queries. Overall, with 96% of the instances running on Spot, Batchly helped save 82% over on-demand AWS costs with no IT overhead.

Batchly - Cost effectively manage your Hadoop, Stream, Batch, Load Testing, Log Processing and Transcoding workloads on AWS