batch.ly

Flight Delay Impact Analysis using Pig on EMR is 79% cheaper with Batchly

HIGHLIGHTS

Use Case: Periodically parse airline dataset for flight delay analysis.

Technology & Dataset Used

Amazon EMR

Apache Pig

40GB TranStats Dataset

Result per Run

- 50 instances
- 94.5% spot instance usage
- 79% savings over on-demand AWS costs
- No IT overhead

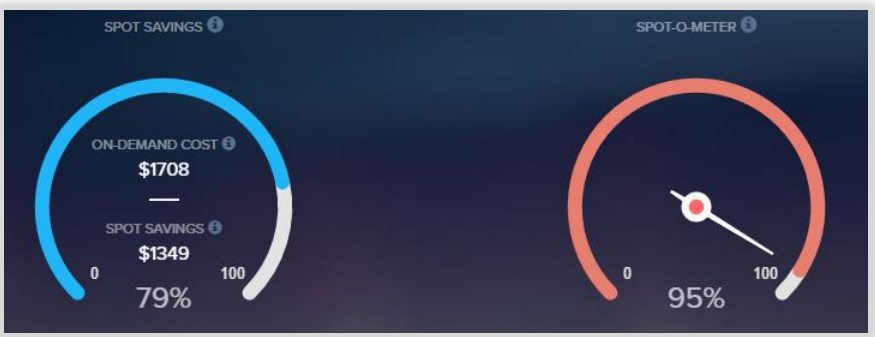
FLIGHT DELAY ANALYSIS

Flight delays is a challenging problem for all airline companies and may lead to financial losses and negative impact on their brand and business reputation. Every year approximately 20% of airline flights are delayed or cancelled, resulting in significant costs to both travellers and airlines. The annual cost of delays in 2007 for US carriers was estimated to be $31 billion.

CHALLENGE

The Flight delay impact analysis was performed using Apache Pig. Pig, a high level scripting language used with Apache Hadoop, enables data workers to write complex data transformations.

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 configure a Pig job.

The airline dataset was loaded from S3 onto HDFS in EMR and Pig scripts were scheduled to run through Batchly to compute parameters related to flight delays. Batchly also provided autoscaling option for quick turnaround time on ad-hoc reporting. Overall, with 94.5% of the instances running on Spot, Batchly helped save 79% 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