BATCHLY

OCR Processing For AWS Customer

HIGHLIGHTS

Use Case: Digitize physical documents for a legal firm as part of e-discovery.

Specs:

Input

- Source – Amazon S3
- File size - 1.5 mil PNG
- Binary – Tessaract Engine from Google
- SLA – 3 hours

Result

- 4XL EC2 instances used at a price lower than single XL EC2 instance
- 59% savings over on-demand AWS costs
- 214 EC2 instances
- 99.53% SPOT instance usage
- Successful adherence to SLA
- No IT overheads

CHALLENGE

Customer is a large commercial law firm in Australia and provides commercial advice for top 100 companies in Australia.

As part of the litigation process, Customer has to "discover" the facts and evidence relevant to the legal issues. The major part of the discovery will include contracts, note and other evidence in physical copies which need to be converted to a digital format for searching, indexing and document management purposes. A typical large lawsuit would involve over 2 million pages, amounting to 5TB of image data. Though they were already using Amazon Web Services, they didn't have a scalable solution which would be simple to use and cost effective.

SOLUTION

Batchly, the cloud driven enterprise data processing solution, was considered to run this process. Since Customer is already on AWS, Batchly needs a delegated trust user permission to start and stop instances and access to Amazon CloudWatch to monitor the progress. Once this is done, Customer can specify the Amazon S3 bucket which contains the input image files, specify the output location and add-in the processing logic (OCR processing using Tessaract).

With this setup, Customer ran a workload on Batchly to process 1.5 million PNG images with an SLA of 3 hours. Batchly not only managed to adhere to the SLA, but was able to bring in 59% saving over on-demand AWS costs. In addition, during the run, Batchly's algorithm noticed and used 4XL instances that were available for 40 cents per hour while a regular large instance on-demand goes for 48 cents per hour. This meant

Batchly completely automated the activities around instance provisioning, monitoring and failure handling. This meant further savings for the customer with respect to IT overheads apart from the direct savings on cost & time.

ABOUT BATCHLY

Batchly is an enterprise data processing solution; abstracts cloud complexity, uses existing code and runs the processing job in AWS at user defined cost and time.