Success Story

Data Indexing using Elasticsearch Engine for an eCommerce Price Aggregator

Business:
Data Indexing in a Big Data Environment

Domain:
eCommerce

Tools:
Elasticsearch, Java/ Python (ETL Process),RDS (Data Base)
AWS

Key Highlights

Key Success:
» Efficient Data Storage implementation using batch processing which reduced process time from 25 hours to 7 hours.
» Data refresh and response time to near 1 second providing real time price information with no latency.

Engagement:
» Offshore engagement with:
- 1 Big Data Architect
- 2 Data Engineers
- 1 Database Engineer

» 1+ years - Ongoing

Client
The client is one of the internet based price comparison website helping consumers with quick comparison information of e-commerce products to avail best deals for their choice of products.

Overview
As the world of retailers move online, the increasing number of online platforms- across e-commerce sites, second hand online trading boards, retailer owned sites, online catalogues etc.- has made it increasingly difficult for a customer to compare prices across all the online options and find the best bargain. Consequently, the solution Indium Software developed enabled the client's customers real time visibility of products across sites amid dynamically changing prices. The functionality was built upon an efficient data handling system that could process real time data requests/responses with minimal lag and effectively crawl dozens of ecommerce sites to return only the most relevant information in the correct taxonomy.

1 Status Quo

The client is one of the internet based price comparison website helping consumers with quick comparison information of e-commerce products to avail best deals for their choice of products. The platform combines data collection from a range of eCommerce websites of over 1500+ retailer websites.

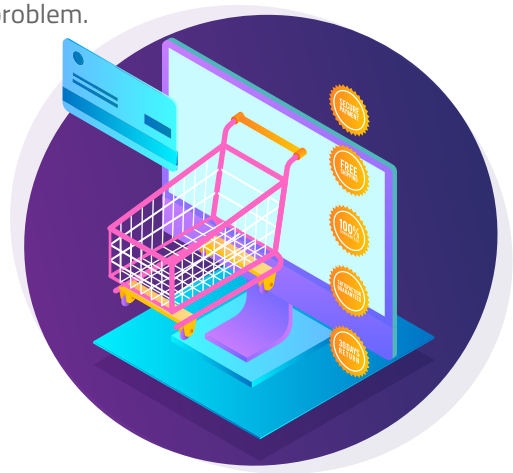
The key challenge in the workflow is to enable end users with real time visibility of products and their dynamic changing prices. This drives the need for an efficient data handling system that can process real time data requests and responses.

2 Business Challenges

» Data gathering from multiple eCommerce sources imposes a product taxonomy challenge. The data should therefore be categorized, redefined into a standard format before storing
» Typically, an end user waiting time for price updates is 2 seconds. This defines the response time for processing the data refresh requests to a ballpark 1 second or less.
» Data storage loads are huge owing to data collection from 1500+ retailer websites and storing them record by record – Single Line Indexing problem. An average data storage process takes about 25 hours.

3 Indium Software's Approach and Implementation

To overcome the mentioned challenges and make real time price information available for users on the website, Indium Software has initiated the following solution approach:
» Implemented a powerful Elasticsearch engine in the Big Data environment for Data indexing.
» Implemented a ML Algorithm using Google Taxonomy to standardize the product data.
» Processed billions of data points using the Big Data ecosystem lending a response time of less than 1 second for n number of query requests fired.
» Our solution leveraged the ETL Batch Processing capabilities of Elasticsearch with batch counts of 10GB data processed in ~1 second time, solving the single line indexing problem.



4 Business Impact

» Efficient Data Storage implementation using batch processing which reduced process time from 25 hours to 7 hours.
» Powerful data load carrying capacity.
» Data refresh and response time to near 1 second providing real time price information with no latency.

INDIUM
Digital | QA | Gaming

INDIA
Chennai
+91 44 6606 9100
Bengaluru
+91 80 4645 7777
Mumbai
+91 022 6215 4028

USA
Cupertino | Princeton
Toll-free: 1 888 207 5969

SINGAPORE
+65 9630 7959

UK
London
+44 773 653 9098

General Inquiries : info@indiumsoftware.com
Sales Inquiries : sales@indiumsoftware.com