CASE STUDY

YOTTAA

BIG DATA WEB ANALYTICS PLATFORM ON AWS FOR YOTTAA

Client Background

Yottaa is a young, innovative company, providing a website acceleration platform to optimize Web and mobile applications and maximize user experience, security, and profitability. It delivers fast, personalized experiences to users on any device, browser, location, or connection, and makes it possible to discover each user's browsing context and tailor the app's content and optimization profile to meet the specific needs of the moment.

softserve

Business Challenge

The client needed to implement an in-house Operation Intelligence platform but the leading tools on the market do not have the ability to quickly detect issues in system performance and security, so they approached SoftServe, renowned for its expertise in complex Big Data solutions and architecture design methods, for help. It was also important that an Agile approach was used as innovation, expansion, continuous improvement and rapid time to market were the key drivers, forcing everyone in the team to constantly focus on results.

Big Data Challenges

The project need to solve several challenges, including:

- High Throughput
  - 1 Billion messages per day
- Large volume
  - Estimated 300 TB
- Near-real time and batch processing at the same time
  - < 1 min event processing latency
  - < 3 sec query response time
- Semi-structured data sources
  - Web Logs

Project Description

Importance of PoC (Proof-of-concept) in Big Data projects

Technical risks are an integral part of any Big Data project. In fact, high complexity and immature technology is the current reality that software architects and engineers face. But building a full-scale prototype is not realistic due to the high cost and time requirements. While architecture analysis alone is insufficient to prove many important system properties such as performance and scalability. Instead, MVP (minimum viable product), throw-away and vertical evolutionary prototypes help in areas that architecture analysis cannot sufficiently address.

For this project a throwaway prototype (also known as rapid prototypes or proof-of-concept) was chosen to quickly evaluate the riskiest technology selections, and an MVP to get early feedback from end users and updating the product roadmap accordingly.

AWS as an Environment for PoC

With no hardware to procure, and no infrastructure to maintain and scale, Amazon Web Service is an extremely effective time-to-market accelerator and was the perfect platform for this project, particularly as, at the early stages, the exact software and hardware requirements were not immediately clear. Particular benefits of the platform were:

1.The opportunity to experiment with software and hardware while looking for the appropriate solution (the environment can be provisioned and unprovisioned in just a few clicks).
2. Tight integration between S3 and EMR enabled a unique Hadoop cluster on demand (Amazon EMR) and resulted in significant cost savings. Hadoop itself has two functions, storing and computing, but with S3 fully covering storage, there was no need to keep the Hadoop cluster up and running when not utilizing the 'computing' function. And as the web logs were stored on S3, it was possible to terminate the Hadoop cluster without data loss and launch the cluster again only when required.

Elasticsearch as a Platform for Dashboarding

During the discovery phase, SoftServe's Big Data experts suggested Elasticsearch as a primary data storage for real-time analytics. The goal was to implement near-real time scenarios with high query performance (< 3 sec) and minimum data latency (< 1 min) in a highly concurrent environment.

The PoC phase was important in order to mitigate performance risks when utilizing Aggregates - a new functionality in Elasticsearch - and focused primarily on three tasks:

Task 1. Quickly populate Elasticsearch with log data, instead of full integration with existing infrastructure which typically takes much longer.

Task 2. Discover the optimal hardware and configuration for Elasticsearch and tune it for the required workload.

Task 3. Create interactive visualization for testing and demos.

The SoftServe team utilized EMR and Elasticsearch-Hadoop driver for Task 1. Pig and Hive was used to parse and load log data from S3 into Elasticsearch (see diagram below) where each Hive external table had been pointed to an Elasticsearch index.

For Task 2 (discover optimal hardware) the team experimented with different EC2 types (general purpose, storage optimized, compute optimized, etc.) and block storages (SSD and HDD drives, instance storages, EBS with provisioned IOPS, etc.). Performance and load tests showed that the CPU was a bottleneck on the required workload, so finally compute optimize instances (c4 model) was selected as a base EC2 instance type for the Elasticsearch cluster.

For Task 3, an initial version of the interactive dashboard using Kibana was created in less than a day.

From PoC to Production

While some of SoftServe's Big Data experts were working on Elasticsearch evaluation using EMR as a processing engine and load data from S3 into Elasticsearch (ETL approach), others were implementing Flume- >Kafka->Logs Consumers processing pipeline for near-real time scenario. Elasticsearch mappings (data model) was agreed at the beginning, allowing work to commence on both Elasticsearch PoC and near real time scenario, in parallel. This enabled the MVP, and the pre-production system, to be created in a short timeframe.

Next Steps

Amazon EMR already supports Apache Spark, a powerful computing engine that can supplement the Operation Intelligence platform by introducing stream processing through Spark Streaming and advanced analytics (Spark MLib), out-of-the-box. But now it fully supports Kafka Direct Approach, Spark has everything required for seamless integration with the system, so the SoftServe team can leverage all advantages of AWS for the future Lambda Architecture.

Value Delivered

AWS proved to be an extremely effective time-to-market accelerator for the Operation Intelligence platform because no time was required for infrastructure deployment and configuration. The pay-as-you-go basis allows costs to be optimized and provides the flexibility to increase capacity over time depending on need.

SoftServe have continued helping the client with implementation and technology advice. The first production version has been successfully released and the system is now evolving with new features.

SoftServe's Big Data accelerators, experienced team and effective cloud technologies along with the client's strong product vision, were crucial to the project's success. The prototyping approach from PoC to MVP and to full-featured production, once again showed its effectiveness and allowed the agreed business and technical goals to be achieved.

ABOUT US

SoftServe is a global digital authority and consulting company, operating at the cutting edge of technology. We reveal, transform, accelerate, and optimise the way large enterprises and software companies do business. With expertise across healthcare, retail, media, financial services, software, and more, we implement end-to-end solutions to deliver the innovation, quality, and speed that our clients' users expect.

SoftServe delivers open innovation – from generating compelling new ideas, to developing and implementing transformational products and services. Our work and client experience is built on a foundation of empathetic, human-focused experience design that ensures continuity from concept to release.

Ultimately, we empower businesses to re-identify their differentiation, accelerate market position, and vigorously compete in today's digital, global economy.

Visit our **website**, **blog**, **Facebook**, **Twitter**, and **LinkedIn** pages.

USA HQ

201 W 5TH STREET, SUITE 1550
AUSTIN, TX 75703
+1 866 687 3588

EUROPEAN HQ

One Canada Square
Canary Wharf
London E14 5AB
+44 (0)800 302 9436

info@softserveinc.com
www.softserveinc.com

softserve