instaclustr

Case Study

IBM designed and built a scalable, highly available, globally distributed consumer management, marketing and IoT platform for one of its major customers, a Fortune 500 company

IBM

Overview

IBM is one of the world's leading system integrators and provides consulting services to its customers worldwide. The company commissioned to design and build a scalable, highly available, globally distributed IoT platform for one of its major customers, a Fortune 500 company providing leadership in connected consumer products. The platform supports the management of the consumer base, marketing campaigns, and IoT devices.

Highlights

- Deployment: Two clusters totaling eight data centers/48 nodes
- Purpose: Marketing consumer database
- Technology: Open Source Apache Cassandra®, with Apache Spark™ and the Cassandra Lucene Index plugin, on AWS
- Service: Consulting services and Instaclustr Managed Platform

> ***Instaclustr has done an amazing job helping us design and build the backbone of the platform with Cassandra and Spark. Their team of consultants integrated with our own team and went beyond their core expertise to provide value at all times.***

Mehdi Charafeddine
Senior Industry Architect, Distribution Sector at IBM

Challenge

IBM's customer was operating a suite of local market solutions to manage its products, its consumers and the marketing campaigns associated with those products. The customer had software that handled most consumer interactions with the brand, including social network interactions, marketing offers, voucher generation, and product utilization. Field marketing personnel were also using specific solutions to engage new consumers, through registrations at events and following up with conference delegates.

Working closely with IBM, the customer realized there was an opportunity to merge those segmented functions and solutions into a unified global platform. The vision was that a unified environment would provide superior consumer engagement, global analytics, and a consistent user experience. All of this would result in faster go-to-market for its next-generation products.

The team identified the following technical challenges that needed to be overcome:

- **Data segregation** – For compliance reasons and good data governance, personally identifiable information could not physically leave the geographical region corresponding to the consumer locality. This was critical to meet the strict requirements of the General Data Protection Regulation (GDPR).

- **Data synchronization** – Maintaining multiple databases and keeping them synced was becoming a burden, with complex housekeeping reconciliation rules.

- **Variable demand** – Database solutions used in an earlier iteration of this platform couldn't respond to spike demands, resulting in higher latencies at peak times.

- **Low latency search** – String search at low latency was an important requirement of the platform to help the marketing teams run inexact queries against the data.

- **Cloud-agnostic** – The solution had to be cloud-agnostic to avoid any cloud vendor lock-in.

- **Rapid and controlled scale** – The platform needed to be capable of scaling rapidly to accommodate the significant usage increase anticipated.

- **Operational visibility** – The development team, quality assurance team, and production team required a high level of operational visibility to allow them to understand the application performance at every stage of the release process.

Solution

In developing a solution architecture to fit the brief, the IBM team knew that they needed to develop a unified platform that had truly global reach, performed to requirements, and was scalable and highly reliable. The integrated architecture had to meet the requirement for scalable data storage while also providing a suitable environment for performing deep analytics on the stored data.

IBM determined that the Apache Cassandra NoSQL database would be the platform

architecture's key technology to deliver the requirements of scalability, performance and high availability. Apache Cassandra's datacenter replication capability, which enables truly global data deployment, was instrumental in providing low latency for both the end customers and internal users. Cassandra's advanced keyspace replication options provided the underlying mechanism used to solve the challenge of data locality, as personally identified data could be pinned to a specific Cassandra data center.

Apache Spark was also identified as a key technology. The IBM architects knew that Spark's analytics focus combined with the live production data being stored on Apache Cassandra would provide the required analytics capability to enable real-time housekeeping and business data reconciliation solutions.

The Cassandra Lucene Index plugin for Apache Cassandra was chosen as the most effective solution to provide low latency search string capability while integrating with Cassandra. This technology stack provided the perfect solution for supporting data query, data search and data analytics while relying on a single data source: Cassandra.

The Instaclustr Advantage

The IBM team engaged Instaclustr to help meet these challenges. Initially, Instaclustr was engaged for its consulting services and specific expertise in open-source solutions such as Apache Cassandra and Apache Spark. The Instaclustr console and monitoring dashboard, coupled with Instaclustr's integration with Datadog cloud monitoring, provided IBM with the required levels of operational visibility at all stages.

Specifically, Instaclustr consultants provided the capability for the following key components of the design and build of the unified platform:

- **Apache Cassandra cluster configuration** – Instaclustr's consultants developed suitable cluster and node configurations in the AWS environment. This included:
  - right-sizing of the cluster and determining node type, node count, disk type and volume.
  - developing Cassandra cluster topology through utilization of AWS Availability Zones and multi-data center (AWS cross-region) configuration.
  - configuring Cassandra security, including encryption at rest, encryption of internodes and client nodes, and appropriately configuring security groups.

- **Apache Cassandra data model development** – The consulting team developed the Cassandra data model to address the transition from a relational data model to the NoSQL Apache Cassandra. The data model was developed to support specific queries while the Apache Lucene data model supports search queries using the Cassandra Lucene Index plugin.

- **Co-location of Apache Spark and Apache Cassandra** – The team installed and configured Spark and Cassandra on the same node, with specific tuning to support ad-hoc analytics queries as required in the final solution.

- **Cluster performance tuning** – The Instaclustr consulting team undertook performance testing and continual tuning to meet the specific throughput and latency metrics expected by the end customer.

- **Knowledge transfer** – The Instaclustr consulting team became a small but integral part of the IBM project team, providing a continuous source of specialist knowledge on Apache Cassandra data modeling, driver configuration, query optimization and general DevOps, including the automated deployment of the platform infrastructure on AWS with Terraform (covering Cassandra, Kafka, Spark, Hadoop, and other technologies and microservices).

- **Compliance** – The solution developed by Instaclustr passed both internal IBM audit and the IBM customer's audit. This covered technical architecture, implementation, procedures, and security. The audit concluded that the solution meets both SOC 2 and GDPR requirements.

As the initial consulting engagement continued and the relationship between the Instaclustr consulting team and IBM's team of solution architects matured into a strong working partnership, the decision was made to investigate the use of the Instaclustr Managed Platform as an environment to deploy the Apache Cassandra and Apache Spark clusters on the AWS environment.

Ultimately, the Instaclustr Managed Platform was used to deploy both the performance and production environments. The IBM team saw the value of having Instaclustr manage the complexity of Apache Cassandra and the benefits of de-risking the management and ongoing maintenance of the clusters.

About Instaclustr

Instaclustr delivers reliability at scale through our integrated data platform of open source technologies such as Apache Cassandra®, Apache Kafka®, Apache Spark ™ and Elasticsearch.

Our expertize stems from delivering more than 30+ million node hours under management, allowing us to run the world's most powerful data technologies effortlessly.

We provide a range of managed, consulting and support services to help our customers develop and deploy solutions around open source technologies. Our integrated data platform, built on open source technologies, powers mission critical, highly available applications for our customers and help them achieve scalability, reliability and performance for their applications.