TRACE3

Case Study:

ONLINE ADVERTISING SALES – CUSTOMER INTIMACY

How Trace3 helped a leading online marketing company offer deeper customer insights with Hadoop

BENEFITS

- Able to process more data faster and with less human involvement
- Big Data analytics improves accuracy of results and provides more analysis options
- Offering deeper insights into customer behaviors and advertising effectiveness

THE CHALLENGE

Search engine marketing and online display advertising has forever changed the way businesses reach consumers in local markets around the world. The rapid growth in this competitive industry has challenged Internet marketing companies to keep pace with their systems to support new services and offer deeper insights. This Trace3 client, a leading global online marketing firm, was quickly approaching an inflection point where structured database environments were becoming too slow and too limiting to serve the expanding analytical needs of the business.

This particular client develops models of customer behavior based on search engine ad placement, keyword string queries, and customer link-through behavior. More and more data must be analyzed in ever-shrinking time windows to create richer and more valuable reports for its customers. The constantly expanding machine-learning system used a combination of historical reporting in various legacy MySQL databases and other log collection systems that could not provide the scalability and performance that data analytics teams required. To remain competitive, the client needed to expand its analytics capabilities and move from a structured database environment to a Big Data architecture that could return results quickly and scale as needed.

THE SOLUTION

The Trace3 Big Data team showed the client how technology could provide a competitive advantage instead of being a limiting factor, recommending a “Data Lake” architecture based on the Apache™ Hadoop® platform. A Data Lake is a massive, easily accessible repository for storing Big Data that retains all data attributes for maximum flexibility in the nature and scope of subsequent analysis.

> With an infinitely scalable Hadoop deployment based on a Data Lake architecture, the client can now centralize all data collection and processing onto a single platform that can be used by anyone in the company.

Leveraging technical and architectural guidance from Trace3, the client developed a platform for Big Data storage and large in-memory processing in an extremely high-density environment. Trace3 recommended the Cloudera CDH5.x Hadoop distribution to start the client out on a YARN-based architecture, taking advantage of the in-memory capabilities of Hadoop 2.x. Knowing that data center floor space was limited and that future applications would require additional resources, Trace3 architected a 10-node “starter” cluster with each 1U node equipped with dual processors, 256GB RAM, and 10GbE networking. This architecture provided the client with over 100TB of

> The Trace3 Big Data team showed the client how technology could provide a competitive advantage instead of being a limiting factor, recommending a “Data Lake” architecture based on the Apache™ Hadoop® platform.

storage capacity in less than 12U of rack space. The client is using the RabbitMQ messaging broker to support event streaming into its Hadoop cluster, and has added cross-replication capabilities to a second cluster in Europe using Cloudera Manager.

Trace3 and Cloudera both provide extensive technical support on the client’s Hadoop platform. The client’s team developed their understanding of Hadoop through in-person knowledge transfer sessions with Trace3 and also by leveraging Cloudera’s extensive knowledge base. The easy, timely access to relevant and critical information was well worth the cost of support licenses.

With an infinitely scalable Hadoop deployment based on a Data Lake architecture, the client can now centralize all data collection and processing onto a single platform that can be used by anyone in the company. It can support a variety of Big Data activities, from log file aggregation by systems administrators to Deep R analytics by the data science team, and can handle accelerating storage capacity demands without changing the environment. In-memory query processing using next generation Apache Spark solutions will allow up to 10-100x improvements in data analysis compared to existing platforms. With data replication between international boundaries, the client can ingest data and process results closer to its customers, with increased redundancy and visibility. This is a game-changing advancement of Big Data analytics providing deeper insights into making its customers’ digital presence more effective.