Data Platform Building at an Online Media Company

A global online media company was struggling to get a better understanding of their users within and across applications.

Silicon Valley Data Science designed and prototyped a data platform to support rapid user growth, and launch a new application critical to the client's bottom line.

Background and Business Problem

A global online media company was experiencing rapid user growth, with revenues doubling year over year. Their architecture was designed for an early user base and data volumes, and so struggled with the increased load. It not only had difficulty providing additional insights to marketing, finance, operations, and other internal users, but frequently did not meet current reporting needs with existing SLAs.

Concerns ranged from increasing infrastructure costs to increasing difficulties supporting new, in-application monetization strategies. In addition, limitations of the existing infrastructure prevented the company from integrating with social platforms and in-application SDKs, two data sources that held potential for high value but were essentially untapped. Furthermore, the ability to create unified user profiles across multiple entry points and applications could open new levels of engagement and monetization, if supported by BI and data pipelines.

Our client turned to SVDS to design and build a new data platform that could support their existing reporting needs and meet their desire to move towards real-time analytics. Key design challenges of this platform included:

- improve time-to-insight for information needed to make key business decisions

- increase access to information and data across analytics and BI consumers, including finance, marketing, operations, data science, engineering, and product teams

- enable ad-hoc analysis while turning repeated queries into standardized reports

- increase data quality and information reliability to achieve more confidence with outcomes and decisions

- identify new uses for data, monetization opportunities, and expand third party data services capabilities

The Challenge

Needed a next generation data platform capable of scaling with massive user growth

Wanted to enable both real time and batch streaming

Legacy platform required key engineering personnel to continuously debug to keep the system live

SILICON VALLEY DATA SCIENCE

info@svds.com | www.svds.com

Solution

An SVDS agile engineering team designed and built a high-performance, real time data platform that is capable of processing up to 2TB of event data per day. Features of the delivered capabilities included:

- "lambda" based architecture to manage both high-velocity & low-velocity use cases
- flexible data ingestion and processing to support existing application portfolio and future onboarding
- ability to process all events in real time to support rapid marketing, operations, and automated third-party integrations
- cloud-based architecture with infrastructure and platform automation to enable rapid scalability and dynamic response to user load
- consolidation of business logic to enable simpler support and QA
- increased processing performance and data availability from days to minutes

Our team, in close collaboration with the clients' existing engineering team, designed and prototyped a new data platform to support the launch of a critical new application. Key technical components within this architecture included:

- a data ingestion Spark pipeline that can handle events with frequently-changing payloads
- high volume batch and real-time processing to meet varying use cases across the business, from daily/weekly reporting to near real-time responses in in-application user interactions
- scalable user access to the environment using Impala and Hive for ad-hoc exploratory data science work
- a 360 degree view of the user cross-applications persisted via HBase

Using this architecture, our client was able to demonstrate how the data science, marketing, and operations teams are able to gain much faster, granular insight into event data. Our collaboration ensured their engineers could continue development activity and support once our work completed.

Our Approach

SVDS interviewed all key stakeholder groups to understand the unique requirements of each group

We designed a new platform and prioritized the use cases to understand critical-path development activities

We used Spark streaming to manage real-time eventing and devops to increase overall supportability and design agility.

New Capabilities

More stable platform capable of scaling to PBs of user data

Real-time streaming and eventing capabilities to support live operations and in-application promotions

More robust dev-ops processes that greatly increased supportability

SILICON VALLEY
DATA SCIENCE

info@svds.com | www.svds.com