CASE STUDY

Building a Next-Generation Analytical Platform

A leading television entertainment company wanted to build a new central data platform that could scale and flex as they ask new questions of an ever- growing data set.

Silicon Valley Data Science designed and developed a next-generation analytical data platform that could serve as the foundation for product and service development.

Background and Business Problem

Our client is a leader in advanced television entertainment, providing innovative solutions that navigate the "content chaos" challenge of the market. To enhance this leadership position, the client wanted to create a next-generation data analytics platform to serve as the foundation for new products and services and to help fuel business growth.

The client recognized that a centralized new data platform held the potential to unlock benefits across several critical business and product design functions. Their data science teams could develop sharper models and a deeper understanding of their customers and viewing habits. Their product teams could build more personalized services to reflect the dynamic needs of a growing customer base. Their sales and operations teams could gain real-time insight into market trends and create proactive strategies.

Building such a platform required the design and implementation of an architecture optimized for scalability and flexibility. New data storage, processing, and analytics capabilities were needed to support a rapidly growing data set, including data collected from set-top boxes across the United States. The client also needed to manage the transition from existing legacy systems that were designed to answer a specific set of analytic questions, and weren't built with the flexibility needed to scale with the loads and processing requirements of new data sets.

Further compounding the challenge, multiple divisions of the company had built separate pipelines for performing similar data processing activities. And there was no historical archive of the raw data received, so data analysts had difficulty answering new questions.

The Challenge

Needed to scale from approximately 350k users' data to millions.

Wanted microsecond granularity while scaling.

Existing systems were difficult to modify.

Historical data needed to be reprocessed to be analyzed in different ways.

Solution

SVDS worked hand-in-hand with the client to design and develop a next-generation analytical data platform that could serve as the foundation for product and service development. Where there had previously been multiple platforms and pipelines, we created a single common platform that now allows multiple business divisions within the company to consume data from a centralized location. Furthermore, that data has been transformed in a repeatable manner, which allows new lines of inquiry to be run on historical data.

PHOTO BY SAM WILLARD

The solution featured an end-to-end data pipeline built with flexible and scalable data processing and storage technologies such as the Hadoop Distributed Filesystem (HDFS) and Apache Spark. With these in place, the platform was able to ingest hundreds of gigabytes of compressed data on a daily basis, while allowing for longitudinal queries of historical data. The result was a highly scalable, extensible architecture that enabled the the client's team to modernize and transform their data processing and analytics workflows.

Through our agile development approach, we also identified valuable intermediate capabilities that could be exposed to end users during the development process. This allowed the team to learn from real-life interactions with the data and adapt. Additionally, we chose technologies that allow the team to transform and experiment with the data in a query layer first, and then select valuable analytics out of that experimentation layer to be productionalized for efficiency.

The client is now able to gather and analyze data from their comprehensive deployment of set-top boxes rather than working from a representative sample, allowing them to compound their industry leadership.

Our Approach

We built a centralized data platform and ingestion pipeline that gave the client's data science team full population-level data

We built a framework that enables powerful analytics in a customizable query environment

We introduced Spark to transform, process, store, and query vast quantities of historical data in a cost-effective manner

New Capabilities

Analytics across data from millions of set-top boxes for the first time

Turnaround time on new questions measured in days instead of weeks

Raw historical data now available for query

Process for creating a data product to answer a specific business question