CASE STUDY

Agile Data Engineering at a UK Retail and Banking Company

A UK retail and banking company needed an enterprise data platform to support their top priority programs: transforming their loyalty program, improving store optimization, and meeting new regulatory requirements.

Silicon Valley Data Science built and launched a scalable and extensible data platform to meet current and future needs.

Background and Business Problem

Our client operates more than 2000 stores, and is comprised of multiple entities. They had an existing data warehouse solution built on Teradata in an on-premise location, that was struggling to meet their Business as Usual (BAU) needs. Any development work for data science or future analytical capabilities was deprioritized or cancelled as there was no additional storage or processing capacity available. In addition, the existing platform was not able to provide access to data in near real-time, as the data warehouse processed its batch jobs in a 24-hour cycle.

Our client launched a transformational program that focuses on increasing revenue by creating stronger relevancy with their customers and therefore larger basket sizes and stronger loyalty. Their immediate technical challenges to realise their vision were lack of access to data, lack of a common platform to access that data, and lack of an environment that provides cost-effective compute power to do exploratory or development work.

The Challenge

Business was not able to meet their top priority use cases

Unable to understand their customers across different areas of the business

Wanted to expand their data science and analytical capabilities

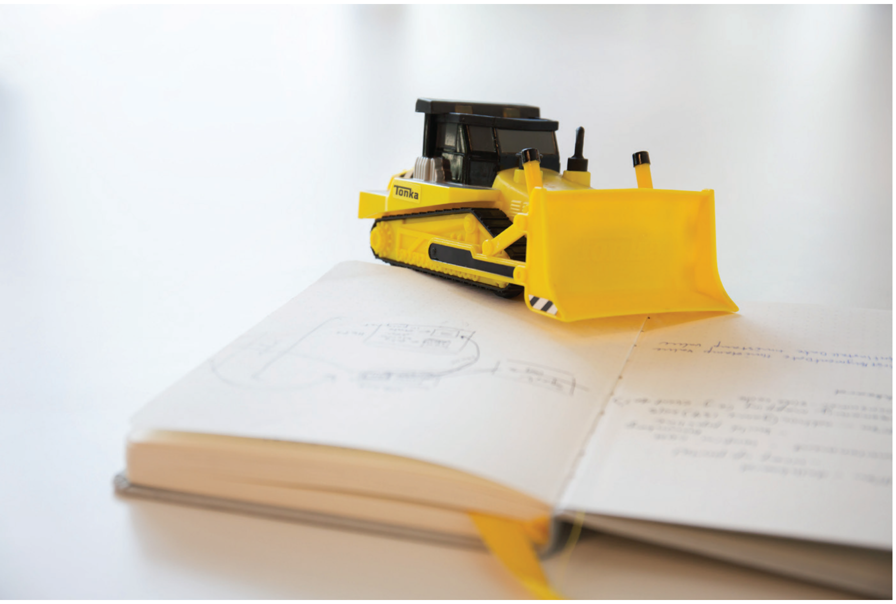

PHOTO BY SAM WILLARD

Solution

SVDS was engaged to design, build, test, and launch a data platform program for the client's top priority use cases across the business. Through an agile process, our team delivered a secure, Cloudera-based Hadoop platform on AWS. The design was built from concepts and prototype code from our push-button project. The platform included the following pieces:

- automation of infrastructure and platform: Terraform, Ansible
- cloud provider: Amazon Web Services, S3
- Hadoop distribution: Cloudera, Cloudera's Kafka, HDFS, Cloudera Navigator
- query engines: Hive, Impala, SparkSQL
- data science toolkits: Jupyter, RStudio, Hue
- continuous integration: Jenkins, Nexus, Consul

The codebase can create ephemeral and/or permanent environments within hours. For data science or engineering development, these ephemeral environments, or "DataLabs," are instantiated to allow for specific project needs. As the environments are easily built or torn down, they live only when needed and are shut down or destroyed during off-peak periods, allowing the company to manage costs effectively. As this process is driven from configuration and repeatable, the codebase is re-used across multiple lines of businesses within the company.

Together with the client team, we launched a new data pipeline that ingests both real-time and batch data. As the system will contain Personally Identifiable Information (PII) data, the platform is also secured at the infrastructure, OS, platform, network, and web layers. Data feeds include, among others, real-time transaction data from cash registers, store location, product taxonomies, and customer data from third parties. To ensure continued success, we assisted with the recruiting of engineering and product owner talent to transition and backfill SVDS and non-SVDS positions, and trained the team on maintaining and extending the system in the future.

Our Approach

SVDS built a Cloudera-based Hadoop platform that is secure, reproducible, and extensible

We worked closely with the client to identify and build an end-to-end MVP

We drove engineering leadership and established best practice processes for build, CI, testing, and agile methods

New Capabilities

Extensible enterprise-wide data platform that serves up use cases across LoBs

Ability to further data science capabilities with cost-effective and easily reproducible environments

Target operating model for the run-state operations