CASE STUDY

Data Architecture at a Leading Investment Management Firm

A top-five global investment management firm needed increased reliability, read/write access, and usability for risk data.

Silicon Valley Data Science designed and tested a more efficient, scalable next-generation architecture to support the needs of future data growth and business demand.

Background and Business Problem

SVDS was engaged by a top-five investment management firm to assist with migrating from MapR to a Cloudera-based platform, designed to enable increased reliability, read/write access, and usability for risk data.

The client generates significant amounts of data each day, which must be stored and made available to a wide variety of tools and downstream systems. Data is being pulled daily into a custom-built in-memory structure that enables users to write user-defined functions (UDFs) and generate their own metrics on demand. Roughly 10,000 additional metrics are generated for users from this system. Legacy workloads are entirely hosted in Oracle, with dedicated specialized hardware. Despite optimized hardware, the legacy systems were no longer scalable.

Prior to our engagement, the client had begun the process of moving data storage and processing to MapR. While initial results were promising, certain design aspects and lack of support for critical components in the Hadoop ecosystem ultimately led them to seek alternatives to MapR.

PHOTO BY SAM WILLARD

The Challenge

Data architecture was unable to scale with rapid growth and analytical demands of daily risk modeling

Modeling needed for portfolio management and risk assessment required an architecture with linear scalability and low latency access

Required a system to store and analyze large scale historical data

Solution

SVDS worked with the client's lead technical architects to identify critical architectural considerations, and develop a vision for how the architecture could be transformed. In order to manage delivery risk and build confidence in the viability of proposed solutions, we prototyped a number of new approaches to existing workloads.

We evaluated and performance tested multiple data storage, processing, and query components to ensure that key performance requirements were met. This involved reviewing not just Hadoop-based query stores, but also memory based stores such as MemSQL, and cloud based tools such as BigQuery, RedShift, and Azure SQL Data Warehouse.

Our efforts resulted in a successful migration of ~400 TB of risk data, and set the foundation for easier, performant, and more consistent usage of risk data throughout the organization.

Our client then used the patterns we established to migrate all workloads from Oracle and MapR to a Cloudera based Oracle Big Data Appliance, utilizing Spark, Impala, Hive, and the other tools in the Hadoop ecosystem.

Our Approach

SVDS created a profile of real world utilization and execution patterns

We designed, prototyped, and benchmarked performance of multiple solutions against execution patterns

We evaluated historical data migration requirements and developed a data migration plan to next-gen data platform

New Capabilities

Better historical data storage and processing with linear scalability

Dramatically increased execute speeds for series based analysis

Roadmap for future tool and architecture development

Key Recommendation

A polyglot data platform using MemSQL, Hadoop, Hive, and Impala will be built to assist with storing, processing, and querying large risk data sets for both current and historical sources in a fast and scalable manner.

Current State

- Scalability limits with existing tools and data models
- Inability to handle large time series queries consistently
- High cost of persisting large amounts of generated data

How to Get There

- ❑ Integrate MemSQL, Hadoop, Hive, Impala, and utilize an optimized schema
- ❑ Integrate reporting and service based data access layers for key
- ❑ Develop a new Risk Data Calculation Engine
- ❑ Replace Oracle for risk data

Future State

- ✓ Provides full SQL access with same interface to current and historical data
- ✓ Provides service-based access patterns to insulate underlying data stores
- ✓ Decreased deployment time for new features based on unified code base
- ✓ Increased transparency in to risk data pipelines

Multi-Phase Risk Data Migration Project