FLYING PHASE

Leveraging Data Infrastructure and Governance to Speed Model Development and Execution

CAPABILITIES COVERED

- Data Management
- Data Platforming
- Cloud Technologies
- Compliance and Governance

FEATURED CONSULTANTS

- David Buckler

THE SITUATION

A CONSUMER CREDIT TEAM FOR A TOP 10 BANK was responsible for maintaining the accuracy of annualized loss predictions for a loan portfolio exceeding $100 billion. But the cycle time to continually rebuild and deploy new loan-level models was taking over a year, requiring more than 20 analyst resources. The primary causes of the problem were scattered data, poor data quality and difficulties in tracing data to its source. Additionally, new privacy laws required all customer data to be maintained in production-controlled systems with close monitoring of every data element used.

CHALLENGE

Streamline the loss forecasting model development process while ensuring that data processes and platforms are compliant and well-managed.

MEASURABLE RESULTS

- Created a single, authoritative data source for the development and execution of all loss forecasting models, **reducing development cycle times by 50% and maintenance resources by 70%**
- Implemented data management practices that **satisfied all compliance and regulatory requirements** while improving data quality and providing additional knowledge to data users
- Automated robust **change point detection and data quality checks** for completeness, consistency and correctness to ensure accurate model execution
- **Utilized powerful distributed computing** technologies to enable analysts to derive even deeper insights from bigger data

OUR APPROACH

We began by conducting an inventory to understand data elements and sources for every model. To ensure reliability and compliance with standards, we cataloged metadata, data quality rules and lineage to source for all critical data elements.

In parallel, we assessed data processes and platforms relative to the enterprise data and modeling infrastructure standards. We collaborated with the loss forecasting team and their technology partners to collect data directly from the disparate sources and consolidate everything into cloud-based data lake. Data was organized into optimized columnar file formats for faster read/write times, and every element was monitored for quality. Datasets were then merged into analytical warehouses to provide account-level data to analysts in a navigable way.

We created a dynamic, distributed computing pipeline to move massive volumes of data, validate quality in flight and optimize each data element for use in its respective model(s). Dynamic cluster sizing allowed for maximum performance while minimizing costs. In total, the system became the authoritative source for retrieving, validating, consolidating and transforming all data elements needed for critical loss forecasting processes. We built a new capability that would allow users to quickly integrate new data sources into the platform. We also enabled "model-ready" views of data to keep model development teams focused on using the data, not pulling and preparing it.