cloudera®

intel®

The Enterprise Data Hub in Financial Services

Three Customer Case Studies

The typical financial services adoption cycle for Apache Hadoop usually begins with one of the two most prominent operational efficiency and cost reduction use cases: data consolidation and multi-tenancy or full-fidelity analytics and regulatory compliance with a centralized data hub. However, an October 2013 study by Sand Hill Group found that only 11% of respondents had progressed beyond their first Hadoop project, and only 9% were using Hadoop for advanced analytics, despite the fact that 62% indicated that they anticipated advanced analytics becoming a top use case during the next 12 to 18 months.[1] With so many organizations seeking a reliable, real-time, and affordable big data solution, what is the barrier to full adoption and production?

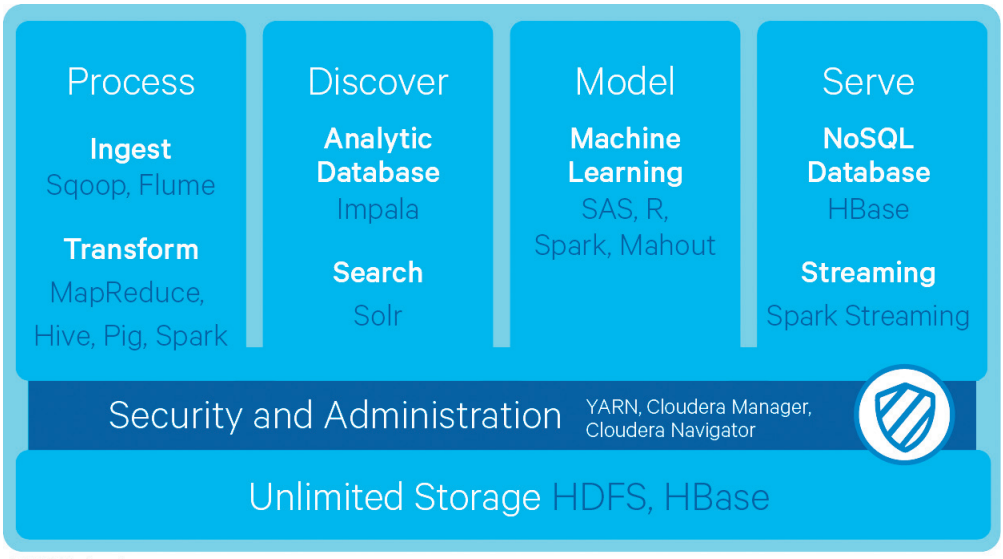
Unlike traditional data management and analytics platforms that are usually deployed as specialized systems with specific objectives, the central, open, and scalable nature of an enterprise data hub makes it more akin to a solutions engine for the financial services industry. By minimizing opportunity cost and emphasizing integration with a vast and growing ecosystem of relevant technologies and familiar applications, Cloudera is helping firms address their big data challenges today and maximize the option value of their data infrastructure for more advanced business objectives downstream. Bringing compute to all your data in service of an introductory use case actually enables, facilitates, and affords the opportunity for firms to quickly take advantage of new information-driven business competencies that were previously too expensive or complex for most enterprises: machine learning models for more effective and automatic fraud detection and prevention, recommendation engines to personalize the customer experience for up-sell and cross-sell opportunities, and a 360-degree view of the business for ad hoc exploration, experimental analysis, and advanced risk modeling.

A LEADING PAYMENT PROCESSING COMPANY AND FRAUD DETECTION

With the movement from in-person to online financial transaction processing, the number of daily transactions processed by a leading global credit card company has ballooned, causing increased susceptibility to fraud. By definition, fraud is an unexpected or rare event that causes significant financial or other damage—the effective response to which can be categorized, from the enterprise perspective, by detection, prevention, and reduction. In the financial services industry, anomalies usually occur because a fraudster has some prior information about how the current system works, including previous fraud cases and the fraud detection mechanisms, which makes building a reliable statistical model for detection very difficult.

In the case of this large credit card processor, despite an annual $1 billion budget for data warehousing, statisticians were limited to fairly simple queries on relatively small samples of data because anything more extensive would consume too many compute resources. In particular, data scientists within the

[1] Graham, Bradley, and Rangaswami, M.R. *Do You Hadoop? A Survey of Big Data Practitioners.* Sand Hill group, October 2013.



© 2015 Cloudera, Inc.

global information security group wanted faster query response and unconstrained access to better mine and analyze data in the relational database management system (RDBMS).

By deploying Hadoop as part of Cloudera Enterprise, this firm not only streamlined its data processing workflows and significantly reduced its anticipated costs by integrating all the jobs usually assigned to separate storage area network (SAN), extract-transform-load (ETL) grid, and data warehousing systems, but also immediately began examining data from a longer period of time and a greater variety of sources to identify more and different potentially anomalous events. To overcome latency, Apache Flume—Hadoop's service for efficiently collecting, aggregating, and moving large amounts of log data—can load billions of events into HDFS—Hadoop's distributed file system and primary storage layer—within a few seconds and analyze them using Cloudera Impala—Hadoop's massively-parallel-processing structured query language (SQL) engine—or even run models on streaming data using the in-memory capabilities of Apache Spark—the next-generation, open-source processing engine that combines batch, streaming, and interactive analytics on all the data in HDFS.

Today, the credit card processor ingests an average of four terabytes of data into its Hadoop cluster every day and is able to maintain thousands more across hundreds of low-footprint nodes for its fraud modeling. Shortly after deploying its enterprise data hub, the company was notified by a partner of a small incidence of fraud that had reportedly only been occurring for two weeks before detection. In response, the global information security group was able to run an ad hoc descriptive analytics model on its long-term detailed data in Hadoop—a task that would have been virtually impossible with traditional data infrastructure alone. By searching through the broader data set, the company found a pattern of the fraudulent activity over a period of months. This became the sector's largest detection of fraud ever, resulting in at least $30 million in savings.

Additionally, the company is using the data from its Hadoop cluster to create revenue-driving reports for merchants. Historically, certain monthly reports took two days to complete and required a large amount of processing power managed by a technical team. Now, the credit card processor is building a billion-dollar business by selling reports generated by combining much larger transaction data with purchase data from banks. The reports can be run in a matter of hours and overcome a latency issue merchants had faced when collecting data for customer segmentation and cross-sell analytics.

A TOP INVESTMENT BANK AND THE 360-DEGREE VIEW OF THE BUSINESS

With growing data volume and variety available for portfolio analysis, many investment banks struggle to figure out the best way to process, gain visibility into, and derive value from more data. Most rely on data sampling, which reduces the accuracy of their models and prohibits exploration.

The concept of a 360-degree view is usually associated with retail banks that want to use more data from more sources across multiple business units combined with on- and offline behavior trends to understand how to effectively and efficiently engage customers for greater loyalty and new selling opportunities. However, a broad, informed, real-time view of the business is not necessarily limited to customer happiness and marketing metrics. Combining related or even disparate data sets can reveal patterns, correlations, or causal relationships that, when translated ▶

into opportunity or risk, can provide investment banks with a valuable head start over other firms.

At a leading wholesale bank, competitive advantage is directly related to not only the quantity and quality of its data but, perhaps more importantly, the flexibility to investigate the relevance and relationship of insights to outcomes. The firm, which reported client assets under management in the trillions of dollars in 2013, balances not only its own market and investment data, but also relies on custom algorithms to draw actionable insights from public and policy information, macroeconomic data, client profiles and transaction records, and even web data—essentially always seeking to go one click down on any individual observation.

The investment bank's data scientists wanted to put very large data volumes to use for portfolio analysis, but the traditional databases and grid computing technologies they had in-house would not scale. In the past, IT would create a custom data structure, source the data, conform it to the table, and enable analysts to write SQL queries. This process was extremely precise and time-consuming. Often, when the application was handed off to the business, the analyst would indicate that the project did not deliver on the original request, and the application would go unused and be abandoned.

As a first big data proof-of-concept with Cloudera, the bank's IT department strung together 15 end-of-life servers and installed CDH, Cloudera's open-source distribution of Apache Hadoop, loaded with all the company's logs, including a variety of web and database logs set up for time-based correlations. With so much data online and available in Hadoop, the bank was able to explore its investment operations at petabyte scale from all angles for the first time. Because Hadoop stores everything in a schema-less structure, IT was able to flexibly carve up a record or an output from whatever combination of inputs the business wanted, and results could be delivered to the business on demand.

As a Cloudera Enterprise customer, the investment bank no longer relies on sampling, meaning its portfolio analysis is run at a much larger scale, delivering better results. Hadoop can search through huge volumes of data and run pattern-matching for every single imaginable attribute. A user does not have to know what he or she is looking for—just let the software and models detect patterns and then follow up with further investigation.

The time-based correlations over log data that are powered by an enterprise data hub allow the bank to see market events and how they correlate with web issues and database read-write problems with an unprecedented level of completeness and clarity. For instance, the company has access to an event's entire traceability in real time, in terms of who did what, when, and how, what caused the issue, and what kind of data was being transacted. The bank can tie front-office activities with what is going on in the back office and which data is causing unexpected results. In the past, figuring out what caused a system to perform incorrectly would take months and could cost the business plenty.

With Cloudera, the company can now figure out and solve problems as they happen, or even prevent them before they happen. Furthermore, advanced analytics tools deployed as part of the enterprise data hub also provide the bank's financial advisers with customized recommendations for clients to sell or buy stocks based on information gathered in real time on current positions and market conditions—essentially monetizing Hadoop's capabilities delivered and supported by Cloudera Enterprise: Data Hub Edition.

A LARGE INSURER AND FINANCIAL PRODUCT PERSONALIZATION

With the proliferation of sensors, mobile devices, nanotechnology, and social apps, individuals are more inclined than ever to monitor and passively or actively share data about their day-to-day behaviors. Insurers, who have historically competed on general pricing or via broad, expensive marketing campaigns, want to differentiate their coverage options by customizing plans based on information collected about the individual's lifestyle, health patterns, habits, and preferences. However, traditional databases cannot scale to the volume and velocity of real-time, multi-structured data required for policy personalization. An enterprise data hub enables real-time storage and stream processing for a competitive pay-for-use insurance model.

One of the largest personal insurance companies in the United States was initially founded as part of a national department store chain in the early-1930s. Over its more than 80 years in operation, the company has collected massive quantities of data, much of which was never digitized, and most of which was unstructured document content. As the insurer began to transition its historical and current policy data into online records and attempt to run programs that correlated such external data as traffic patterns, socioeconomic studies, and weather information, the IT department found that the systems would not scale to accommodate such variety of formats and diversity of sources.

A primary example of the challenge faced by business analysts was graph link analysis. For instance, they could look at data from a single U.S. state at a time—with each state's analysis requiring about a day to process—but could not run analytics on multiple states, no less all 50 states, at once. Although new data systems were being put in place to capture and prepare data for reporting and business intelligence, they were primarily aligned to marginally improve on old approaches to data management, which separated data types and workloads into distinct silos.

With a first objective of speeding up processing times and consolidating

its disparate data sets to achieve more scalable analytics, this leading insurance company built an enterprise data hub with Cloudera Enterprise. Its centralized Hadoop implementation spans every system across the entire company to break down data silos and provide a single, comprehensive view of all its data. The three main technical cases for adopting Hadoop were flexible and active data storage, integrated and efficient ETL, and applied statistics and computation.

The insurer brought together customer account information, public economic and social studies, and telemetric sensor data in its initial Hadoop cluster. Some of these data sources had never been brought together before, and much of the historical data, which was newly digitized, could not be analyzed in tandem with external sources prior to landing in Hadoop. Today, the company's enterprise data hub is integrated with its incumbent mainframes and data warehouses—it was designed specifically to complement, not replace, existing infrastructure.

Now that it can run descriptive models across historical data from all 50 states using Apache Hive—open-source software that makes transformation and analysis of complex, multi-structured data scalable in Hadoop—the insurer is experiencing an average 7500% speed-up on analytics and seeing even better results with Impala. Unburdened by data silos, its analysts and data scientists are building predictive models that help the business customize products that are better aligned to the individual behaviors and risks of each customer, tune pricing of insurance plans more precisely to maximize lifetime value, and develop differenti¬ated marketing offers that communicate value for the most appropriate cross-sell and up-sell opportunities without diminishing margins.

BIG DATA AND AN ENTERPRISE DATA HUB

When information is freed from silos, secured, and made available to the data analysts, engineers, and scientists who answer key questions about the market—as they need it, in its original form, and accessed via familiar tools—everyone in the C-suite can rest assured that they have a complete view of the business, perhaps for the first time. For financial services firms, overcoming the frictions related to multi-tenancy on compliant and secure systems is the gateway to advanced big data processes: machine learning, recommendation engines, security information and event management, graph analytics, and other capabilities that monetize data without the costs typically associated with specialized tools.

Today, the introduction of an enterprise data hub built on Apache Hadoop at the core of your information architecture promotes the centralization of all data, in all formats, available to all business users, with full fidelity and security at up to 99% lower capital expenditure per terabyte compared to traditional data management technologies.

The enterprise data hub serves as a flexible repository to land all of an organization's unknown-value data, whether for compliance purposes, for advancement of core business processes like customer segmentation and investment modeling, or for more sophisticated applications such as real-time anomaly detection. It speeds up business intelligence reporting and analytics to deliver markedly better throughput on key service-level agreements. And it increases the availability and accessibility of data for the activities that support business growth and provide a full picture of a financial services firm's operations to enable process innovation—all completely integrated with existing infrastructure and applications to extend the value of, rather than replace, past investments.

However, the greatest promise of the information-driven enterprise resides in the business-relevant questions financial services firms have historically been unable or afraid to ask, whether because of a lack of coherency in their data or the prohibitively high cost of specialized tools. An enterprise data hub encourages more exploration and discovery with an eye towards helping decision-makers bring the future of their industries to the present:

How do we use several decades worth of customer data to detect fraud without having to build out dedicated systems or limit our view to a small sample size?

What does a 360-degree view of the customer across various distinct lines of business tell us about downstream opportunity and risk?

Can we store massive data on each customer and prospect to comply with regulatory requirements, secure it to assure customer privacy, and make it available to various business users, all from a single, central point?

ABOUT CLOUDERA

Cloudera is revolutionizing enterprise data management by offering the first unified Platform for big data, an enterprise data hub built on Apache Hadoop. Cloudera offers enterprises one place to store, access, process, secure, and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. Cloudera's open source big data platform is the most widely adopted in the world, and Cloudera is the most prolific contributor to the open source Hadoop ecosystem. As the leading educator of Hadoop professionals, Cloudera has trained over 40,000 individuals worldwide. Over 1,600 partners and a seasoned professional services team help deliver greater time to value. Finally, only Cloudera provides proactive and predictive support to run an enterprise data hub with confidence. Leading organizations in every industry plus top public sector organizations globally run Cloudera in production.

CLOUDERA
www.cloudera.com