Hadoop® Accelerates Earnings Growth in Banking and Insurance

A Modern Data Architecture for Financial Services

A Hortonworks White Paper
AUGUST 2014

Contents

Introduction

Few industries depend as heavily on data as financial services. Banks and insurance companies aggregate, price and distribute capital with the aim of increasing their return on assets with an acceptable level of risk. To stay competitive, they need real-time, actionable data to forecast market movement, understand operations, screen for fraud and comply with regulations. In today's marketplace, this means capturing, storing, and analyzing data from millions of transactions every minute.

For these reasons, financial services companies invest in data solutions earlier and more deeply than other industries.

Despite aggressive investments in data storage, financial services data tends to be fragmented within siloes across many repositories that are expensive to maintain. Data analysts struggle to unify data from multiple platforms, which means that their reports are incomplete or miss narrow time windows for making decisions. This leads to reports of questionable quality that also take a long time to produce.

Because of this inflexibility, legacy data solutions cannot keep pace with the rigorous stress testing and capital adequacy required for regulatory compliance. Nor are traditional solutions swift enough to spot increasingly sophisticated fraud patterns aimed at credit cards and payment networks. Financial executives need prompt answers to their inquiries about trades, fraud exposure, and regulatory risk without waiting weeks and spending millions of dollars to get the information.

Forward-thinking financial services firms have responded to these challenges by embracing Apache™ Hadoop to improve their risk-adjusted return on capital. A survey[1] of C-level Wall Street executives conducted by NewVantage Partners in 2013 found that those leaders were surprised to find big data having had a greater impact on their businesses then they initially imagined.

Hortonworks' founding architects pioneered Apache Hadoop YARN which moved the platform beyond its batch-only roots. Hortonworks invested heavily in YARN and spearheaded its development in the Apache community. Now Hadoop 2 (with YARN) includes the following enhanced capabilities, which nobody understands better than Hortonworks.

Multi-use, Multi-workload Data Processing: Hadoop supports multiple access methods (batch, interactive, and real-time) to a common data set. Analysts can view and transform data in multiple ways at once. This speeds time-to-insight and strengthens confidence in their findings.

New Opportunities for Analytics: Hadoop's schema-on-read architecture lets users store data in its raw format. Analysts then define unique schemas for the data they need for a particular research question or application.

RESULTS WITH HADOOP[1]

Operational data store:

$300,000

instead of $4,000,000 using relational database

Trading warehouse:

$200,000

versus $4,000,000 with a database appliance

Analyzing risk data:

3 hours

versus 3 months

Pricing calculations:

20 minutes

versus 48 hours

Behavioral analytics:

20 minutes

versus 72 hours

Modeling automation:

15,000 models

up from 150 per year

1) Source: http://blogs.wsj.com/cio/2014/01/27/financial-services-companies-firms-see-results-from-big-data-push/

New Efficiencies for Data Architecture: Hadoop runs on low-cost commodity servers and reduces overall cost of storage. This makes it affordable to retain all source data for much longer periods, which provides applications with far deeper historical context.

Data Warehouse Optimization: ETL workloads also benefit from Hadoop's favorable economics. Data with low per-unit value can be extracted and transformed in Hadoop. This data's value grows in the aggregate, when it is joined with other data and stored for longer.

These advantages explain why banks and insurance companies throughout the world adopt Hadoop. In fact, Hortonworks Data Platform (HDP) is currently in use at many of the largest financial services firms in the US and Europe.

This white paper illustrates real-life improvements in business performance achieved by Hadoop in seven common use cases. Banks and insurers extract the most value from Hadoop when they begin by focusing on specific line-of-business challenges, such as those outlined here. This complements the value that Hadoop creates in horizontal functions such as IT, which are not covered specifically in this white paper.

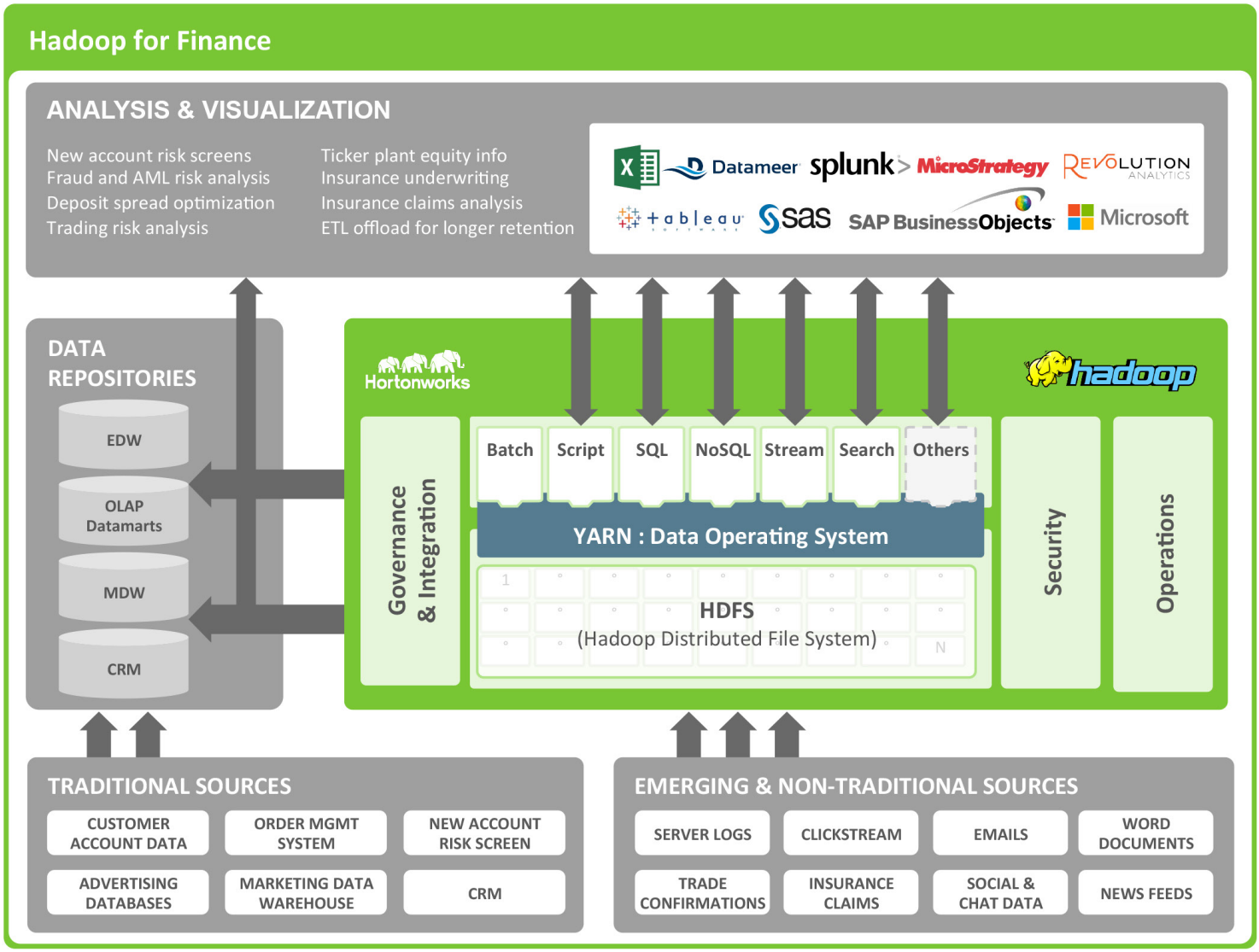

Figure 1: Ingest all the data, store it at scale and analyze it with the tools you already use.

Retail Banking

SCREEN NEW ACCOUNT APPLICATIONS FOR RISK OF DEFAULT

Business challenge

Every day, large retail banks receive thousands of applications for new checking and savings accounts. Bankers that accept these applications consult 3rd party risk scoring services to determine the likelihood that the applicant will mismanage their account or intentionally defraud the bank. These services make "do not open" recommendations, but bankers can (and do) override negative records and open accounts for applicants with poor banking histories.

A disproportionate number of these high-risk accounts overdraw and charge-off due to mismanagement or fraud, costing banks millions of dollars in losses. Banks pass some of this cost on to their entire customer base, including the majority of customers who responsibly manage their accounts.

Solution

Apache Hadoop can store and analyze multiple data streams and help bank managers control new account risk in their branches. They can match banker decisions with the risk information presented at the time of decision and manage risk by sanctioning individuals, updating policies, and identifying patterns of fraud. Over time, the accumulated data informs algorithms that may detect subtle, high-risk behavior patterns unseen by the bank's deposit risk analysts.

Impact

Improved risk management allows the bank to lower its provisions for bad debts and write-offs. Customers with poor account track records can be declined with confidence (and according to regulatory guidelines) or guided into risk-controlled products. Those applicants who intend to defraud the bank research multiple banks and attack those with the weakest safeguards. Use of Hadoop strengthens risk controls and persuades "fraudsters" to target another bank instead.

Insurance

IMPROVE UNDERWRITING EFFICIENCY FOR USAGE-BASED AUTO INSURANCE

Business challenge

Traditional auto insurance attempts to identify and reward safe drivers for their historical driving records based on accidents and speeding tickets that have already happened. But underwriters cannot tell whether a good driving record signifies prudence behind the wheel, or just good luck so far. Newer usage-based insurance (also called Pay as You Drive or PAYD) aligns premiums with empirical risk, based on how policyholders actually drive.

One insurance company launched PAYD products.. The growing volume, velocity and variety of data required for PAYD insurance taxed their existing systems and processes. The high cost of storing PAYD data on an RDBMS platform forced the company to discard 75% of the available data. Processing the remaining 25% took one working week. Risk analysis was too slow. And the results were based on sample data, with inherent risk of error.

Solution

After adopting Hadoop, the company retains 100% of policyholders' PAYD geo-location data and processes that quadrupled data stream in three days or less. More data and faster processing enables the insurer to price risk with much more confidence. It can retain low-risk drivers that might have churned because other insurers were offering better rates. And it can re-price high-risk drivers so that their revenue justifies the risk.

Impact

The insurer is able to acquire certain segments—the prudent drivers—with more affordable rates. It can also charge higher rates for riskier drivers, based on empirical data about their behavior. Apache Hadoop stores driver data far more economically than other alternatives, which lowers the cost to assess risk for all applicants. Lower storage costs combined with more predictive power means better margins across the entire pool of policyholders.

ANALYZE INSURANCE CLAIMS WITH A SHARED DATA LAKE

Business challenge

Processing insurance claims is still a fairly high-touch, document-intensive, manual process. Adjusters need to distinguish fraudulent from valid claims, without easy access to all the data.

The experience of one major property, casualty, life and mortgage insurance company illustrates the potential of Hadoop for claims processing.

The company already had systems in place for analyzing structured data at scale. Underwriters did use less-structured claims notes or social media analysis on a claim-by-claim basis, but inclusion of these types of data did not scale easily. Combining all textual and social data with all structured data was not economically viable, yet the union of the disparate data sets had the potential to add valuable information to claims analysis and reduce servicing costs.

Solution

Apache Hadoop changes those economics. Its "schema on read" architecture permits ingest of a much wider range of data types. Data puddles that were previously scattered about the provider are now unified in a data lake, to process claims more efficiently. This deep data reservoir can still be analyzed using existing business intelligence tools and employee skills, thanks to close integration between HDP and Hortonworks partners SAS, Tableau and QlikView.

Impact

Hadoop allows the insurer to blend and correlate data from various sources using a variety of processing engines and analytical applications. This is of crucial importance when combatting claims fraud, because what may appear a legitimate claim in one system can be quickly exposed as fraudulent when additional structured and unstructured data from different sources are brought into the analysis.

Capital Markets and Investments

MAINTAINING SLAs FOR EQUITY TRADING INFORMATION

Business challenge

A leading provider of financial market data offers real-time information for trading equities and other financial instruments. This "ticker plant" collects and processes massive data streams, displaying prices for traders and feeding computerized trading systems fast enough to capture opportunities in fractions of a second.

Every day, the provider ingests about 50GB of server log data from 10,000 different feeds. Applications query the data at a rate of 35,000 times per second. While 70% of queries are for data less than 1 year old, 30% are for data older than one year. The existing architecture was only able to hold 10 years of trading data, and the growing volume of data was degrading performance, with a risk of missing the ticker plant's 12 millisecond SLA.

Solution

The provider re-architected its ticker plant with Hadoop as its cornerstone. The team offloaded ETL jobs to Hadoop for affordable long-term data retention. Now Apache HBase enables low latency responses to queries, meeting rigorous SLA requirements. Years of historical data are also available for long-term analysis of market trends. Now the company can store more data for longer and make it available more quickly, all at a lower cost than before.

Impact

The company realized a more than ten times improvement in price-performance for this particular area of its business. Based on this outcome, the company recently expanded its Hadoop cluster by an order of magnitude, to aggressively extend its ETL offload capabilities. The company is also investigating new data services to offer its customers—opportunities previously unavailable when data was less available and more expensive.

TRADE SURVEILLANCE AND COMPLIANCE ANALYSIS

Business challenge

An investment services firm with thousands of financial advisors who serve millions of individual clients. It processes fifteen million transactions every day.

As the company grew, it's existing architecture could only make data highly accessible for a limited period of time. Analysis of older historical data required a cumbersome extraction of archived data.

Most importantly, each day's trading data was not available for risk analysis until after the close of business. This created an unacceptable window of several hours, during which time the firm was exposed to risks from rogue trading, without a timely way to intervene to block improper activity. The company needed to strengthen its limited ability to do intraday risk analysis.

Solution

Hadoop accelerates the firm's speed-to-analytics and also extends its data retention timeline. Now a shared data lake across multiple lines of business provides more visibility into all trading activities. The trading risk group accesses this shared data lake to process more data on positions, trade executions and balances. They can do this analysis on data from the current workday, and it is available for at least five years—much longer than before.

Impact

The new Hadoop data lake accelerates time-to-insight and extends data retention. On any given day, operational data is available to risk analysts while markets are still open, enabling them to reduce risk of that day's trading activities. The company's compliance officers appreciate the greater visibility and longer data retention made possible by Hadoop. Now they are able to answer regulators' questions with more confidence and less effort.

MINING DATA ASSETS WITH AN ENTERPRISE-WIDE DATA LAKE

Business challenge

A leading global investment company wanted to derive revenue from its disparate data assets. However, those datasets were fragmented and difficult to combine for mining across the organization.

The existing enterprise data warehouse solutions were appropriate for some data workloads, but prohibitively expensive for others. For example, financial log data was difficult to aggregate and analyze at scale. Log schema was highly variable, which made it technically challenging to transform it for storage in an RDBMS platform. The relatively high cost of legacy storage meant that data was discarded before the end of its useful life. This meant that analyses of lifetime customer value or total cost to serve yielded incomplete results with sampling error.

Solution

The company deployed a multi-tenant Hadoop cluster to merge data across groups. For instance, server log data merged with structured data uncovered previously invisible trends across assets, traders and customers. For sensitive customer-specific data, the team used Apache Accumulo, a component included in Hortonworks Data Platform which enforces read permissions down to the level of individual data cells.

Impact

The company started mining its data assets with this enterprise-wide data lake and quickly identified more than one hundred use cases, as more departments came to understand Hadoop's potential. Already, the project has more than covered its cost through offloading workloads not suited for the enterprise data warehouse. Accumulo permits broader access for more users, with confidence that company policies are systematically enforced.

New and Adjacent Businesses

MONETIZE CONSUMER FINANCE DATA FOR INVESTORS, ADVERTISERS, AND MERCHANTS

Business challenge

A large financial institution was looking for a way to monetize aggregate, anonymous consumer financial data. Banks possess massive amounts of operational, transactional and balance data that holds information about larger macro-economic trends. This economic information can be valuable for investors, advertisers and merchants.

The firm faced two technical challenges in generating revenue from this data. Data protection regulations and policies require that the privacy of bank customers be strictly protected. This required valuable banking data to be aggregated and served in a way that did not contain personally identifiable information. Secondly, the company stored data across isolated silos controlled by different lines of business. The firm needed a way to bring all the data together.

Solution

The bank turned to Hadoop for a common cross-department data lake to serve different lines of business including mortgage, consumer checking, personal credit, and wholesale and treasury banking. A single point of security and privacy enforcement allows the bank to operationalize security and privacy measures such as de-identification, masking, encryption, user authentication and access control.

Impact

Now both internal bank executives and clients in the secondary market derive value from the data. Mortgage bankers, consumer bankers, the credit card group and treasury bankers have access to the same dataset. As a result, the bank is able to improve operational decisions and also generate revenue from the sale of actionable intelligence to investors, advertisers and merchants.

Any financial services business cares about minimizing risk and maximizing opportunity. Banks weigh the risk of opening accounts and extending credit against the opportunity to earn interest and service revenue. Insurance companies manage the risk that claim liabilities will outpace revenue from premiums. Investment companies pursue long-term portfolio appreciation knowing that some securities will lose value.

Storing and processing all data in Hadoop provides better insight into the optimal balance of these risks and opportunities. Financial services firms that integrate Hadoop into their operations build competitive advantage by improving risk management, reducing fraud, driving customer upsell and improving investment decisions.

The preceding examples from the financial services industry show what enterprises are learning in other industries: Hadoop presents superior economics compared to legacy data warehousing and storage technologies and it also uncovers exciting new capabilities for growing the business.

Build a Modern Financial Services Architecture with Enterprise Hadoop

To realize the value of your investment in big data, use the blueprint for Enterprise Hadoop to integrate with your EDW and related data systems. Building a modern data architecture enables your organization to store and analyze the data most important to your business at massive scale, extract critical business insights from all types of data from any source, and ultimately improve your competitive position in the market and maximize customer loyalty and revenues. Read more at: http://hortonworks.com/hdp

Hortonworks Data Platform provides Enterprise Hadoop

Hortonworks Data Platform (HDP) is powered by 100% open source Apache Hadoop. HDP provides all of the Apache Hadoop related projects necessary to integrate Hadoop alongside an EDW as part of a Modern Data Architecture. It ships with efficient Data Management and versatile Data Access capabilities, along with three capabilities enterprises require for widespread adoption: Data Governance & Integration, Security and Operations.

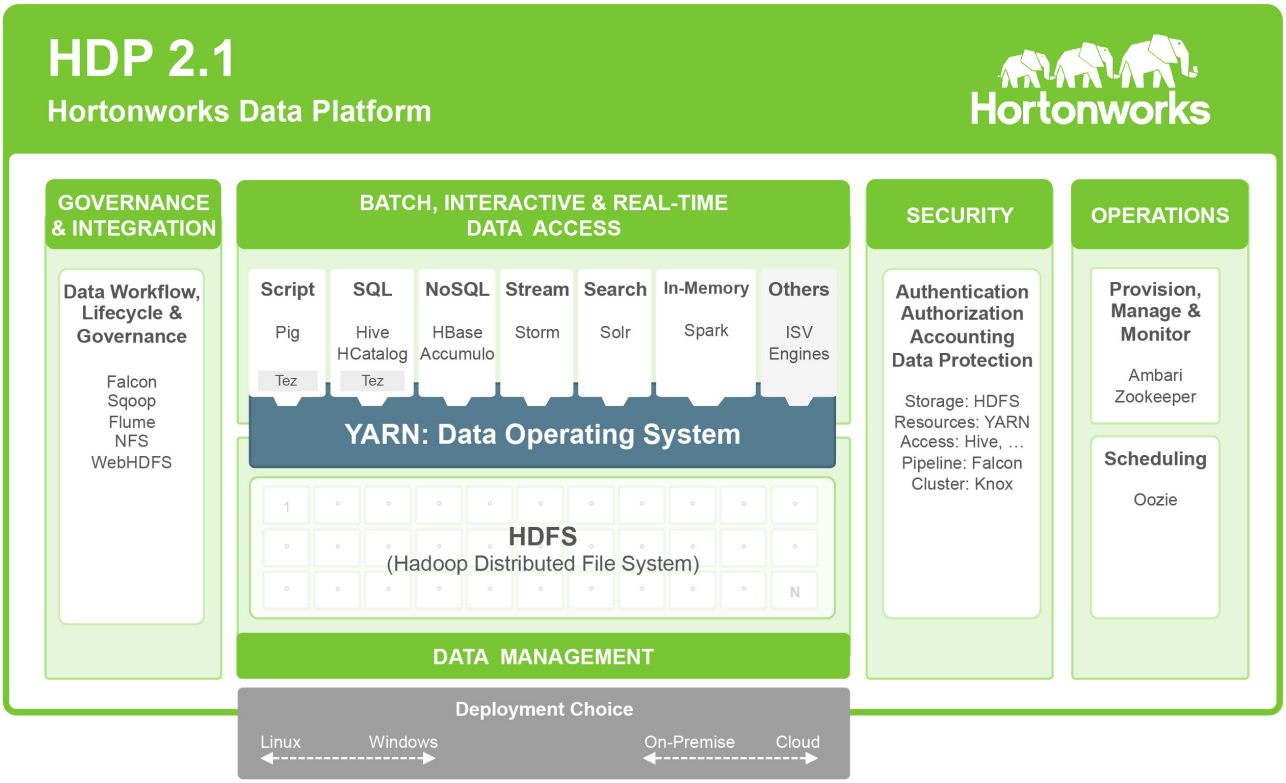

Figure 2: Five core capabilities—data governance & integration, data management, data access, security, and operations.

Why Hortonworks for Hadoop?

Founded in 2011 by 24 engineers from the original Yahoo! Hadoop development and operations team, Hortonworks has amassed more Hadoop experience under one roof than any other organization. Our team members are active participants and leaders in Hadoop development, designing, building and testing the core of the Hadoop platform. We have years of experience in Hadoop operations and are best suited to support your mission-critical Hadoop project.

For an independent analysis of Hortonworks Data Platform and its leadership among Apache Hadoop vendors, you can download the Forrester Wave™: Big Data Hadoop Solutions, Q1 2014 report from Forrester Research.

About Hortonworks

Hortonworks develops, distributes and supports the only 100% open source Apache Hadoop data platform. Our team comprises the largest contingent of builders and architects within the Hadoop ecosystem who represent and lead the broader enterprise requirements within these communities. Hortonworks Data Platform deeply integrates with existing IT investments upon which enterprises can build and deploy Hadoop-based applications. Hortonworks has deep relationships with the key strategic data center partners that enable our customers to unlock the broadest opportunities from Hadoop. For more information, visit www.hortonworks.com.