The explosion of descriptive and real-time data available to financial services firms has provided new options for internal operations management and fraud detection. The increased regulation on transparency, risk, auditing, reporting, and systems is driving the surge in IT spend year over year. In fact, PricewaterhouseCoopers, in collaboration with CIO Magazine and CSO Magazine, found in its 2014 Global State of Information Security survey of nearly 10,000 corporate executives that compliance was reported as a primary driver of spend among financial services firms 14% more often than in all other industries.

However, a series of recent high-profile decisions by the U.S. Department of Justice against BNP Paribas, JP Morgan Chase, Barclays, and other large, global banks, resulting in multi-billion-dollar fines, has brought anti-money-laundering (AML) to the top of the industry’s priority list for big data. The three most notable sets of regulations that govern money-laundering-related fraud and that represent financial services firms’ incremental focus on AML are:

- The Bank Secrecy Act
- Know Your Customer
- The Foreign Account Tax Compliance Act

**Bank Secrecy Act**

The U.S. Congress started addressing domestic money laundering in 1970 with the passage of the Bank Secrecy Act (BSA), otherwise known as the Currency and Foreign Transaction Reporting Act. The government hoped to identify and address suspicious activity that might suggest money laundering or tax evasion by requiring financial institutions to keep records of and report on large cash transactions and the purchase or sale of suspicious negotiable instruments. The minimum size for these suspicious cash exchanges is set at $10,000 (daily aggregate), and the Financial Crimes Enforcement Network (FinCEN), an agency under the U.S. Department of the Treasury, was established in 1990 to provide oversight.

**Know Your Customer**

During the past 45 years, BSA has been enhanced to include anti-terrorism monitoring and was further strengthened by the U.S.A. Patriot Act, enacted in 2001. The Patriot Act also includes Know Your Customer (KYC) policies requiring financial institutions to perform due diligence to verify the identities of potential clients and keep detailed records of the processes used. In addition to customer identification and identity-matching tactics, KYC programs tend to include user profiling to set a baseline of expected behaviors and transactions, monitoring against the activities of customers with similar profiles, and proactive risk scoring for propensity to commit money laundering, terrorist finance, or identity theft crimes.

More specifically, the financial services investment in big data tools and technology related to fraud has heretofore been targeted at the identification and prevention of nefarious activities that lead to direct costs for banks, payment processors, and their customers, with relatively little focus on money laundering. The former White House Deputy Chief Technology Officer, Daniel Weitzner, recently told The Wall Street Journal, “[Companies have] taken it on themselves to spot fraudulent transactions. [They] have invested billions in incredibly sophisticated Big Data techniques. But the understanding is the government—[and not banks]—will do the analysis to spot money laundering.”

> “The only way as a firm we really can compete—what is our fundamental differentiator—is our intellectual capital.”

Morgan Stanley

---

### Foreign Account Tax Compliance Act

The Internal Revenue Service's (IRS's) Foreign Account Tax Compliance Act (FATCA) of 2010 requires U.S. citizens, including those living abroad, to report on their accounts held with overseas banks and requires foreign financial institutions to report on holdings of and transactions with any U.S. citizen. The primary objective of FATCA is to prevent taxpayer non-compliance and collect federal revenues otherwise concealed as foreign holdings and transactions. Identification and monitoring of these offshore accounts and shell corporations also helps prevent money laundering and funding to terrorist groups.

### Regulations in Financial Services: Anti-Money-Laundering (AML)

<table>
<thead>
<tr>
<th>Regulation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSA: Bank Secrecy Act or Currency and Foreign Transactions Reporting Act (1970)</td>
<td>Requires reports to FinCEN on cash purchases of negotiable instruments of more than $10,000 (daily aggregate), suspicious activity, and tax evasion</td>
</tr>
<tr>
<td>KYC: Know Your Customer (2001)</td>
<td>Compels financial institutions to perform due diligence to verify the identities of potential clients and keep detailed records of the processes used</td>
</tr>
<tr>
<td>FATCA: Foreign Account Tax Compliance Act (2010)</td>
<td>Requires foreign financial institutions to report to the IRS on holdings of and transactions with United States citizens</td>
</tr>
</tbody>
</table>

### Augmenting AML with Hadoop

New cyber-surveillance applications have led to billions of dollars in loss avoidance for the financial services industry via the prevention of consumer fraud, credit card fraud, and internal threats. However, unlike other forms of fraud that are identified with machine learning algorithms that detect anomalies and outliers, money laundering schemes are designed to closely mimic typical banking behaviors—hiding clandestine or illegal activities behind a taxable, auditable mainstream business front—and are, therefore, characteristically less anomalous. The thresholds mandated by reporting policies like BSA and utilized by first- and second-generation AML systems are well known, so criminals have little difficulty modeling the source of their above-board trade and transaction behaviors to be largely imperceptible, even to specialized software.

As a result, these systems must be enriched with much larger and more diverse data sets to isolate signals of possible money laundering. When a signal is detected, human judgment must be applied—a case is opened, which kicks off an inquiry to verify the crime and the extent of the damage. Without big data, the AML indicators are often not sufficiently distinct to be caught by computational models and leave most of the work to a time-consuming and expensive investigation. In fact, respondents to KPMG’s 2014 Global Anti-Money Laundering Survey reported that their top two budget investments for AML are in enhancing and upgrading transaction monitoring systems for real-time fraud scoring and reviewing, updating, and maintaining KYC systems, suggesting they are "increasingly unhappy with their current automated monitoring efforts, [and are] looking for software that can reduce the burden on the compliance department."³

Apache Hadoop is the ideal platform for AML because it augments all of the core functions of a specialized system to better handle big data: data collection, data preparation, automated evaluation, model building, and investigation. The modern AML architecture is fully integrated with an enterprise data hub, with Hadoop initially staging massive complex data for legacy solutions to provide runtimes for the predictive models and perform the actual fraud detection. Beyond the introductory use case of more expansive and affordable storage, Hadoop's natural fit for backtesting against long-term descriptive data is gaining popularity for more advanced AML workloads, as is the use of other components in the Hadoop stack for exploration, discovery, investigation, and forensics.

### Data Collection

Having more data generally improves the chances of identifying fraudulent activity. However bank data tends to be segregated into silos. There are typically several account databases, while transactions and other notable events can be sourced from a variety of channels, including mobile, online, point-of-sale, and ATM. The high cost of storage causes even more fragmentation, as data is offloaded onto secondary or tertiary systems that trade off accessibility for lower cost. As a result, the data made available for investigation and modeling is usually limited to a few weeks or months.

---

In contrast, the cost of storing data on Hadoop is typically orders of magnitude lower than every alternative, meaning data spanning decades can easily and affordably be retained and queried. Hadoop’s flexibility to data of any format or structure and its inherent scalability simplify the consolidation of data ingested from any source—relational databases, mainframes, message queues, networked storage—in batch or as real-time streams into a central, multi-tenant data hub.

**Data Preparation**

Data undergoes enrichment, transformation, and vectorization prior to being scored for fraud. Event data often needs to be combined with other data, such as location, account details, or transaction data from other systems prior to being used for security intelligence or other projects. In some cases, an AML application requires a heuristic matching algorithm to prepare certain types of data, as with a street address. Hadoop excels at this type of data processing, both in real-time and as a batch job, and integrates with familiar tools for extract-transform-load (ETL), which govern the business rules while Hadoop handles the heavy data collection, transformation, and preparation prior to the data being addressed by an existing solution like Oracle Mantas or SAS AML.

**Fraud Scoring**

Predictive models are used in fraud detection as an early warning system to flag suspicious signals. Hadoop’s support for multiple frameworks can bring a variety of computational techniques to bear on the AML problem, including static rules engines, state machines, graph algorithms, natural language processing, and machine learning. Having a variety of solutions at their disposal enables developers and data scientists to improve the accuracy of their fraud detection models.

**Model Development**

Criminal methods evolve to evade detection, requiring predictive models to be developed and improved over time. Some models are relatively static: known patterns that have been hardcoded or encoded using a rules engine. Some use machine learning techniques like linear regression or clustering, which require training from a historical data set. Model development is ultimately a human activity, whereby data analysts, statisticians, or data scientists test combinations of algorithms and parameters in a drive towards more accurate outcomes.

Hadoop aids in model development in two key ways. First, the availability of interactive query tools such as Cloudera Search—the full-text, interactive search and scalable, flexible indexing component of an enterprise data hub—and Impala—Hadoop’s massively-parallel-processing structured query language (SQL) engine—facilitate the discovery of new patterns and associations. Second, the availability of more data and processing power allow models to incorporate more parameters, train on longer historical perspective, and iterate more rapidly when backtesting new variations.

**Investigation**

When a predictive model flags a suspicious signal for potential money laundering, an investigator follows up to verify the occurrence of a crime, the extent of the damage, and possible ways to remediate the situation and mitigate future incidents. Investigation is a costly activity because of the human element required to prosecute an incident. Improving model accuracy to eliminate false positives, thereby reducing the caseload, is a major way Hadoop decreases the cost of AML. As part of an enterprise data hub, ad hoc interactive query tools such as Cloudera Search and Impala also reduce the burden of investigation by providing fast answers to arbitrary questions over large data sets.

---

"For our advanced analytics projects [using Cloudera], we’ve been able to look back at more historical files and achieve more accurate and more detailed predictive modeling while identifying more salient variables... For certain projects across all 50 states plus Canada and other territories, we’ve achieved a 500-time speedup on reports, and we see even faster times with Impala."

Allstate

---

**Source:** Cloudera

"For our advanced analytics projects [using Cloudera], we’ve been able to look back at more historical files and achieve more accurate and more detailed predictive modeling while identifying more salient variables... For certain projects across all 50 states plus Canada and other territories, we’ve achieved a 500-time speedup on reports, and we see even faster times with Impala."

Allstate
Hadoop is rapidly becoming the de facto solution for AML workloads at banks around the world because it is capable of capturing and retaining petabytes of data; it has the power to combine and enrich large data sets, yet remains agile to account for future changes in processes and techniques; it can process event data individually, as a stream, or as large batches to support different fraud types; it supports a number of different processing and storage frameworks that, when combined, enable significantly better detection; and low-latency, interactive query engines allow for exploration and discovery, which benefit model-building and investigation.

As part of an enterprise data hub, Hadoop’s flexibility, scalability, and affordability are extending existing investments in dedicated fraud-detection solutions like Oracle Mantas or SAS AML by increasing the volume, age, and variety of data that can be examined while speeding up data transformation for faster time to insight. Once such massive data is consolidated, Hadoop can increasingly take on more advanced AML workloads such as entity matching—the process by which records from disparate systems are mapped to individuals or legal identities—while associated tools in the data hub like Cloudera Search and Impala remove the complexity of model development, process automation, and case investigation.

**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 22,000 individuals worldwide. Over 1,200 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.