cloudera
Ask Bigger Questions

INDUSTRY BRIEF

FULL-FIDELITY ANALYTICS AND REGULATORY COMPLIANCE IN FINANCIAL SERVICES

Table of Contents

Introduction

In recent years, federal stress tests have increased the demand for predictability and integrated solutions for capital asset management. New regulatory compliance laws have been put into place to improve operational transparency. Financial services organizations are held much more accountable for their actions and are required to be able to access years of historical data in response to regulators' requests for information at any given time. For example, the Dodd-Frank Act requires firms to maintain records for at least five years; Basel guidelines mandate retention of risk and transaction data for three to five years; and Sarbanes-Oxley requires firms to maintain audit work papers and required information for at least seven years.

Partly because of these pressures, leading financial services firms have realized that the key to optimizing their business operations is maintaining an efficient and comprehensive Big Data infrastructure. However, the size, expense, and complexity of data management at this scale easily overwhelms traditional systems. Six years of publicly available market data amounts to approximately 200 terabytes, and the proprietary data currently collected by individual financial services firms adds up to tens of petabytes altogether.

> **The only way as a firm we really can compete—what is our fundamental differentiator—is our intellectual capital.[1]**
>
> MORGAN STANLEY

Regulations in Financial Services: Risk, Fraud, Misconduct, and Transparency	
Sarbanes-Oxley Act (2002)	Sets stricter penalties to senior management and accounting firms for financial misconduct, stronger board oversight, and greater independence for third-party auditors
Dodd-Frank Wall Street Reform and Consumer Protection Act (2010)	Requires greater transparency and provides consumer and investor risk exposure protections
Volcker Rule within Dodd-Frank (2010)	Prevents speculative investments by banks that don't benefit depositors (e.g., proprietary trading)
Basel III (agreed 2010, enacted 2013)	Limits leverage and requires capital coverage and liquidity to respond to greater stresses
BCBS 239: Basel Committee on Banking Supervision Principles for Effective Risk Data Aggregation and Risk Reporting (2013)	Outlines 14 requirements to strengthen risk data management, calculation, and reporting practices by 2016
EMIR: European Market Infrastructure Regulation (2012)	Requires European Union banks to report on all over-the-counter (OTC) transactions and measure counterparty and operational risk for bilaterally cleared OTC derivatives
Regulations in Financial Services: Auditing and Reporting	
OATS: Order Audit Trail System (1998)	Requires electronic auditing and reporting capabilities on all stock and equity orders, quotes, trades, and cancelations
CAT: Consolidated Audit Trail (TBD)	Will obligate finer-grained order, cancelation, modification, and execution details in a consolidated system governed by the SEC and FINRA
Regulations in Financial Services: Technology Standards	
WORM: Write-Once/Read-Many (1934)	Compels permanent preservation of electronic records without deletion or alteration
PCI DSS: Payment Card Industry Data Secured Standard (2004)	Standardizes credit card transaction security to prevent cardholder exposure and fraud
Regulations in Financial Services: Anti-Money-Laundering (AML)	
BSA: Bank Secrecy Act or Currency and Foreign Transactions Reporting Act (1970)	Compels permanent preservation of electronic records without deletion or alteration
FATCA: Foreign Account Tax Compliance Act (2010)	Requires foreign financial institutions to report to the IRS on holdings of and transactions with United States citizens
KYC: Know Your Customer (2001)	Compels financial institutions to perform due diligence to verify the identities of potential clients and keep detailed records of the processes used

[1] Boulton, Clint. "Morgan Stanley Smith Barney Betting Big on Analytics," *The Wall Street Journal* CIO Journal Blog. 16 September 2012.

> **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

Archiving such massive quantities and varieties of data in traditional storage and staging technologies like a storage area network (SAN) or network-attached storage (NAS) can cost up to ten million dollars per petabyte and does not offer any of the accessibility or compute most firms require of a Big Data strategy. Alternatively, traditional data warehouses built on relational database management systems (RDBMS) offer speed and power by keeping the data active, but were not designed to accommodate such large and diverse data sets. The hardware and software footprint required to consolidate a series of RDBMS as a central archive becomes remarkably complex and expensive—up to ten orders of magnitude over SAN or NAS. Yet, even at higher cost, these legacy systems do not offer the flexibility and centralization that financial services firms seek from a modern Big Data architecture. Converging these specialized systems around an enterprise data hub built on Apache Hadoop™ is the logical next step.

As the requirements for compliance with an increasing variety of risk, conduct, transparency, and technology standards grow to exabyte scale, financial services firms and regulatory agencies are building data infrastructure with Hadoop at its core. Banks, payment processors, and insurance companies are now able to not only fulfill the demands of regulatory bodies at scale without the capital burden of specialized systems, but can also take on more advanced workloads and realize new strategic benefits from the same data that they need to keep on-hand for compliance reporting. By deploying an enterprise data hub, the IT department works across the different business units to build an active archive for multiple users, administrators, and applications to simultaneously access in real time with full fidelity and governance based on role and profile.

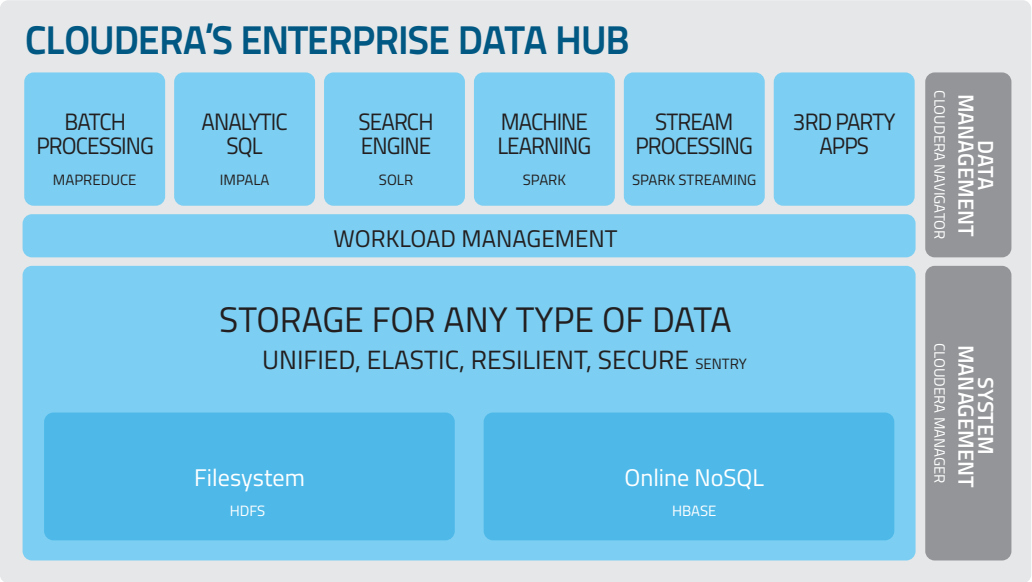
According to the Security Technologies Analysis Center (STAC) in its Intel-sponsored first-quarter 2014 study of 10 of the top global retail and investment banks, almost half of the Big Data projects taken on by the largest financial services firms were driven by regulatory or legal concerns. The research showed that most banks are combining data from far more internal systems than had previously been documented and are designing and deploying systems built on Hadoop that are capable of considerably more powerful analytics than their legacy technologies were. The same survey indicated that another quarter of the banks are adopting Hadoop to offset the costs usually associated with storage and archiving, likely related to compliance reporting requirements. And some of the remaining quarter of respondents were building Big Data projects that took advantage of the data that had already been consolidated for regulatory reporting to enable advanced analytics or improve analytical agility.[2]

Let's consider three financial services business cases tied to federal regulations that have historically required dedicated and specialized technology for compliance:

- Portfolio risk
- Records and reporting
- Data security

[2] Securities Technology Analysis Center. *Big Data Cases in Banking and Securities: A Report from the Front Lines.* June 2014.

With an enterprise data hub, firms have begun to scale out their multi-tenant data architecture both to accommodate the demands of these processes at a fraction of the cost—but with comparable or better performance—and to ensure that the large data sets required for any given application are also available in full fidelity and with full governance for any number of additional tools and tasks.



Source: Cloudera

Portfolio Risk

To comply with the Basel III regulations implemented in January 2014, FDIC-supervised institutions with holdings of $500 million or more, including but not limited to investment banks, must be able to report their risk profiles against set capital adequacy and leverage ratios on an on-demand basis. As a result, financial services firms, particularly the largest banks with multiple holding companies, are compelled to stress test against scenarios that would require them to pay out potentially catastrophic losses with liquid capital (covering total net cash outflows over 30 days and available stable funding accounting for a one-year period of extended stress).

Given these new transparency measures, firms need risk management systems that are, ideally, flexible enough to both incorporate current market and credit risk regulations and respond to future rules and calculation standards that may be introduced in the medium term. Basel III requires banks to build complex models that perform ongoing stress tests and scenario analyses across all their portfolios in order to adequately plan for the possibility of an economic downturn in multiple time horizons throughout the life of each exposure. Accordingly, monitoring and reporting systems must be fully interactive to evaluate the new loss-mitigation strategies and hedges banks intend to put in place as a result of future analyses.

Unfortunately, many current systems are not capable of evaluating positions that are synchronously valued on large sets of historic data across multiple market factors like volatility, foreign exchange rates, and interest rates. Today's trading desks typically model scenarios using Microsoft Excel™ spreadsheets, which means they are only able to consider snapshots of data—insufficient to fulfill new requirements. Conversely, specialized architecture for risk and capital adequacy is complex and expensive, with separate systems for model authoring, extract-transform-load (ETL) processes, grid computation, and data warehousing. Dedicated systems also may not be only able to deal with the rapid iteration required to test new models that may at first be error-prone and inconsistent before they are coded for firm-wide reporting on the entire portfolio.

An enterprise data hub built on Hadoop enables risk managers to model tens of thousands of opportunities per second and the trading desk to perform intra-day calculations by running scenarios against a real-time events database—or tick store—as well as against massive historic data, all accessed centrally with full fidelity in a scalable, governed, unified, and open architecture.

A strategy to address the steps in the market-risk data processing chain of storage, ETL, analysis, and reporting may have historically required several purpose-built technologies. However, an enterprise data hub offers Impala—Hadoop's massively-parallel-processing structured query language (SQL) engine—and Apache Spark—the next-generation, open-source processing engine that combines batch, streaming, and interactive analytics on all the data in HDFS (the distributed file system and primary storage layer for Hadoop) via in-memory capabilities—fully integrated with the storage and applications layers of existing data infrastructure to provide fast, complete transformation, calculation, and reporting at a fraction of the cost. Apache HBase—Hadoop's distributed, scalable, NoSQL database for Big Data—provides real-time storage of massive tick data and more descriptive data to enable analysis of intra-day risk at much greater scale. For the first time, the Hadoop stack creates the option to affordably and scalably analyze custom scenarios on an ad hoc basis prior to trade execution, facilitating Basel III compliance by extending the capabilities of tools within the data center, rather than requiring expensive, new, dedicated systems.

Records and Reporting

Since 1934, the Securities and Exchange Commission (SEC) has mandated that broker-dealers preserve a wide range of records as part of its Rule 17a-4, originally intending a two-year retention period for all transaction and financial documentation related to the trade of stocks, bonds, and futures. During the next 80 years, the SEC expanded its requirements to all communication sent and received, later including all electronic correspondence—whether by e-mail, instant message, phone, text message, or other channel—as well as company reports, website interactions, and other information. Records must also be easily searchable and retrievable, and they now must be maintained for at least three years, with some types of information requiring longer shelf life. Today, the most common storage and compliance method is write-once/read-many (WORM) technology that captures and stores this data in a non-rewritable, non-erasable format, such as tape, hard disk, or redundant array of independent disks (RAID).

The Order Audit Trail System (OATS) SEC regulation requires electronic auditing and reporting capabilities on all stock and equity orders, quotes, trades, and cancelations. Audits are complex and costly because they require data to be found, collected, transformed, stored, and reported on-demand from a variety of sources and data formats with relatively short timelines in order to avoid fines (or worse). Once the data is brought together, it typically sits in storage and is no longer easily available to the business. Soon, the Consolidated Audit Trail (CAT) will obligate finer-grained order, cancelation, modification, and execution details in a system governed by the Financial Industry Regulatory Authority (FINRA).

Records and reporting requirements have long been a challenge for the financial services industry and are the original definition of the sector's Big Data problem. The dual objectives of managing historical data to comply with federal requirements and being able to retrieve and query more data on an ad hoc basis can be both disruptive to the business and prohibitively expensive. The head of enterprise architecture at NYSE Euronext described the problem in 2012: "When you're talking about billions of transactions per day, building systems that can take unfriendly data and turn it into regulation-friendly, analysis-ready information is a key, ongoing struggle… We are obligated to maintain data [for seven years]. There [was] not one system out there that could actually store that data and have it online."[3]

[3] Hemsoth, Nicole. "Big Data Engenders New Opportunities and Challenges on Wall Street." *HPCwire.com*. 27 September 2012.

Expanding reporting requirements—for both industry firms and regulatory agencies—are overwhelming systems that were originally built in traditional data warehouses and duplicated and archived for WORM on tape or RAID. On the reporting side, the RDBMS breaks down because of increasing volume and variety of data required for OATS (and, eventually, CAT) compliance. The diversity of data makes reporting expensive due to the variety of workloads required—ETL, warehousing, reporting—while SQL, which is used primarily for business intelligence and analysis, is not an adequate tool for order linkage. Although tape is inexpensive, it does not ease retrieval of data and is subject to depletion or deletion over time. Ultimately, recordkeeping and auditing for regulatory compliance are costly exercises because they have historically not served core business objectives or scaled with the growth and complexity of industry data.

By building an active archive with Hadoop, the data required for reporting becomes less disparate and requires less movement to staging and compute. HDFS and MapReduce (the batch processing engine in Hadoop) offer significant cost savings over all other online WORM-compliant storage technologies and are far more format-tolerant and business-amenable than tape storage. The industry-standard servers on which Hadoop clusters are built also provide the benefit of latent compute alongside storage, which can easily be applied to ETL jobs to speed transformation and cut reporting timelines. All data is searchable and retrievable with Cloudera Search, the full-text, interactive search and scalable, flexible indexing component of an enterprise data hub. Impala provides in-cluster reporting and investigation capabilities to keep the data required for auditing accessible in its original format and fidelity for business intelligence and other workloads, while Spark provides significantly faster and more robust order linkage.

When used in conjunction with traditional storage and data warehousing, an enterprise data hub is a solution for both the companies building reports and agencies, such as FINRA (a Cloudera Enterprise customer), that receive, store, and scrutinize them due to Hadoop's relatively low cost, scalability, and ease of integration. In fact, Cloudera Enterprise customers in the retail and wholesale banking industries, such as JPMorgan Chase, Citigroup, and Capital One, have reported completing natural-language-processing jobs that are required for SEC record-keeping in only two hours, compared to at least two weeks to run the same jobs on specialized systems with much larger hardware footprints.

Data Security

The Payment Card Industry Data Security Standard (PCI DSS) originated as separate data security standards established by the five major credit card companies: Visa, MasterCard, Discover, American Express, and the Japan Credit Bureau (some of whom are Cloudera Enterprise customers). The goal of ensuring that cardholder data is properly secured and protected and that merchants meet minimum security levels when storing, processing, and transmitting this data was formalized as an industry-wide standard in 2004 by the Payment Card Industry Security Standards Council.

In January 2014, PCI DSS Version 3.0 went into effect, requiring organizations to mitigate payment card risks posed by third parties such as cloud computing and storage providers and payment processors. The new version also stresses that businesses and organizations that accept and/or process cards are responsible for ensuring that the third parties on whom they rely for outsourced solutions and services use appropriate security measures. In the event of a security breach resulting from non-compliance, the breached organization could be subject to stiff penalties and fines.

The simplest way to comply with the PCI DSS requirement to protect stored cardholder data is to encrypt all data-at-rest and store the encryption keys away from the protected data. An enterprise data hub featuring Cloudera Navigator—the first fully integrated data security and governance application for Hadoop-based systems— is the only Hadoop platform offering out-of-the-box encryption for data-in-motion between processes and systems, as well as for data-at-rest as it persists on disk or other storage media.

Within the tool, the Navigator Encrypt feature is a transparent data encryption solution that enables organizations to secure data-at-rest in Linux. This includes primary account numbers, 16-digit credit card numbers, and other personally identifiable information. The cryptographic keys are managed by the Navigator Key Trustee feature, a software-based universal key server that stores, manages, and enforces policies for Cloudera and other cryptographic keys. Navigator Key Trustee offers robust key management policies that prevent cloud and operating system administrators, hackers, and other unauthorized personnel from accessing cryptographic keys and sensitive data.

Navigator Key Trustee can also help organizations meet the PCI DSS encryption requirements across public networks by managing the keys and certificates used to safeguard sensitive data during transmission. Navigator Key Trustee provides robust security policies—including multifactor authentication—governing access to sensitive secure socket layer (SSL) and secure shell (SSH) keys. Storing these keys in a Navigator Key Trustee server will prevent unauthorized access in the event that a device is stolen or a file is breached. Even if a hacker were able to access SSH login credentials and sign in as a trusted user, the Navigator Key Trustee key release policy is pre-set to automatically trigger a notification to designated trustees requiring them to approve a key release. If a trustee denies the key release, SSH access is denied, and an audit log showing the denial request is created.

With Navigator Encrypt, only the authorized database accounts with assigned database rights connecting from applications on approved network clients can access cardholder data stored on a server. Operating system users without access to Navigator Encrypt keys cannot read the encrypted data. Providing an additional layer of security, Navigator Key Trustee allows organizations to set a variety of key release policies that factor in who is requesting the key, where the request originated, the time of day, and the number of times a key can be retrieved, among others.

OUT-OF-THE-BOX DATA PROTECTION FOR THE ENTERPRISE DATA HUB		
At-Rest Encryption	**Key Management**	**Access Controls**
> High-performance transparent data encryption for HDFS, Hive, HBase and more > Rapid deployment and configuration through Cloudera Navigator, requiring no changes to any Hadoop applications	> Software-based key and certificate management with strong configurable management policies and lifecycle > Any security-related object can be stored in a secure vault that allows for true separation of keys and objects from encrypted data	> Fine-grained access controls to data and metadata in hadoop and role-based authorization through Sentry > Supports process-based access controls to prevent unauthorized users and systems from accessing sensitive data

Source: Cloudera

Regulatory compliance, data security, and systems governance should be seen as table stakes for any Big Data platform. The enterprise data hub was designed specifically as an open and cost-effective means to respond to stricter regulations while removing the opportunity cost to more advanced capabilities. As you comply with the stringent regulations governing data for the financial services industry, that data remains available and active with full management and security so that it never has to be archived, siloed, or duplicated, and it can be integrated with your preferred analytics tools, at scale and without friction.

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, process 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. Only Cloudera offers everything needed on a journey to an enterprise data hub, including software for business critical data challenges such as storage, access, management, analysis, security and search. As the leading educator of Hadoop professionals, Cloudera has trained over 40,000 individuals worldwide. Over 800 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. www.cloudera.com.

cloudera® Ask Bigger Questions

Cloudera, Inc. 1001 Page Mill Road, Palo Alto, CA 94304 | 1-888-789-1488 or 1-650-362-0488 | cloudera.com