cloudera®

WHITE PAPER

Information-Driven Financial Services, Big Data, and the Enterprise Data Hub

Version: Q115-102

cloudera®

Table of Contents

cloudera®

> "The only way as a firm we really can compete—what is our fundamental differentiator—is our intellectual capital." [1]
>
> Morgan Stanley

Introduction

As Walter Wriston, the former chairman and chief executive of Citibank and a technology visionary in his own right, said in 1984: "Information about money has become almost as important as money itself."[2]

Today, financial services firms rely on data as the basis of their industry. In the absence of the means of production for physical goods, data is the raw material used to create value for and capture value from the market. Perhaps even more important than the size and availability of data, however, are the efficiency and effectiveness with which firms are able to scale big data across multiple use cases to drive opportunity—from a 360-degree view of the customer and product personalization to fraud detection using machine learning and next-best-offer models built on recommendation engines.

As recent trends towards increased competition and shrinking margins become pervasive, an information advantage has emerged as one of the keys to maintaining and growing profit in the global financial services industry. Simultaneously, firms must also balance the massive proliferation of data with the challenge of increased compliance requirements meant to aid recovery from the Great Recession.

Every sector of the industry faces the tremendous challenge on a daily basis to neutralize inbound pressures and threats: regulation, risk exposure, fraud, competition, customer acquisition cost, and negative sentiment. From commercial banking and credit card processing to insurance companies and broker-dealers, the changing nature and role of data within the larger financial services industry is both straightforward and staggering:

- **Retail and Commercial Banking:** more products, more marketing, more channels, more transactions, more regulation
- **Credit Cards and Payment Processing:** increasing card/holder ratio, more specialization by user profile, competitive rewards programs, increased fraud
- **Investment and Wholesale Banking:** more complex trading models and tools, more research, more rapid and incremental trade execution, more over-the-counter channels, more regulation
- **Insurance:** proliferation of sensor and surveillance data, availability of environmental and climate data, new policy types, diverse actuarial requirements
- **Consulting, Services, and Regulatory Agencies:** growth of the financial information industry, changing compliance requirements, expanded role of and demands on regulatory agencies

Three Factors Entrenching Big Data in Financial Services

The definition of big data is rapidly changing within the financial services industry—pivoting from a measure of volume, variety, and velocity to an evaluation of systems-enabled strategy. Whereas most of the discussion has revolved around the challenges of managing petabytes of unstructured data-in-motion, in practical terms, the most important questions relate to the potential of analyzing full data sets spanning multiple silos: the option to combine many data formats and structures in a single project, delivering much faster speed to insight and greater iterative flexibility.

[1] Boulton, Clint. "Morgan Stanley Smith Barney Betting Big on Analytics," The Wall Street Journal CIO Journal Blog. 16 September 2012.
[2] Wriston, Walter B. "The Citi of Tomorrow: Today (March 7, 1984)," The Walter B. Wriston Papers: 1918 – 2006. Tufts University Digital Library, 2007.

cloudera®

The most successful information-driven firms are using a modern infrastructure to go beyond future-proofing the data center against larger operational efficiency demands. They are also advancing the business case for big data to attack line-of-business questions tied to the largest corporate objectives:

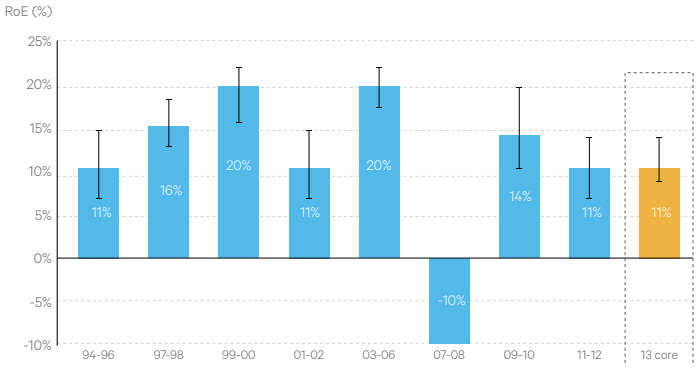
- Prioritization and competitive advantage
- Regulatory compliance and recession recovery
- Customer profiling and segmentation

Towards Competitive Advantage: Consolidation Around High-Return Opportunities

According to a 2013 report by Morgan Stanley Research and Oliver Wyman, technology and competition are forcing wholesale banks to prioritize core competencies and market fit in order to build on their strengths for scope and scale. New models are emerging and, more than ever, operating leverage and the ability to deal with regulation are impacting the delta between winners and losers. In the short term, consolidation is likely and will lead to greater divergence in strategy and positioning.[3]

Information is at a premium as investment banks optimize along different paths and narrow their focus on strategic growth opportunities. The median return for investment banks is leveling off and the range of returns is narrowing. Banks will need to have more and clearer data on their capabilities and concentrations, in addition to diverse market information, to identify and capture competitive advantage.[4]

Historical spread of wholesale bank returns are around the average



Source: Morgan Stanley Research & Oliver Wyman

Simultaneously, competition around data is increasing as third-party research and analytics organizations proliferate and technology firms intermediate between consumers and traditional financial services institutions. Incumbents have an early lead on data collection, but investment and focus are required to advance the strategy and maintain control. Where the technology required to manage, secure, and access petabyte-scale and complex data sets was historically seen as a luxury afforded by only the largest banks, big data infrastructure has become table stakes throughout the industry. Competition, strategic prioritization, and differentiation rely on a complete view of the market, no matter where you fit into the financial services landscape.

Providing evidence of the rate at which profits have shrunk due to competition and the high demand for information infrastructure, Oliver Wyman finds that, during the past 20 years, the margins on deposits and cash equities have declined by 33% to 50% while the need for computing power in the financial services industry has grown 200% to 500% faster than revenue.[5]

[3] Morgan Stanley Research and Oliver Wyman. *Wholesale & Investment Banking Outlook 2013: Global Banking Fractures: The Implications*. 11 April 2013.

[4] Morgan Stanley Research and Oliver Wyman. *Wholesale & Investment Banking Outlook 2013: Mis-allocated Resources: Why Banks Need to Optimise Now*. 20 March 2014.

[5] Wyman, Oliver. *The State of the Financial Services Industry 2013: A Money and Information Business*. January 2013.

Ultimately, the ability to cost-effectively manage and scale data systems will not only enable the speed at which trades are executed and the premium services offered to different customer types, but will also inform the entirety of banking strategy. Competitive advantage in financial services will be derived from the amount of work done by—the compute workloads deployed on—a single, central, massive data store that is fully governed and can accommodate many different users and profiles in real time.

Recovering from 2008: Growth in a Stringent Regulatory Environment

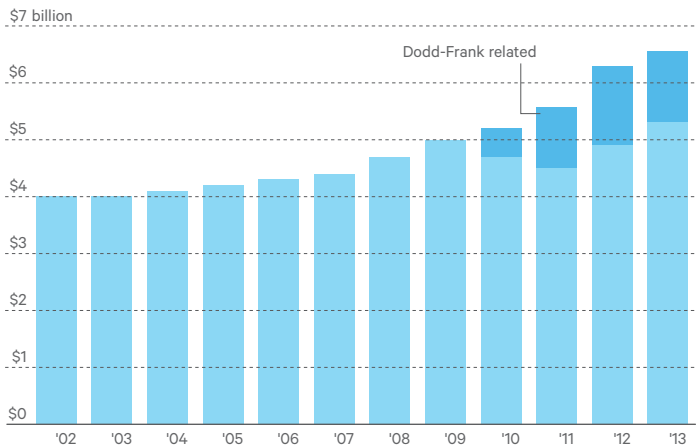
Although the industry policies and technology regulations that have emerged in the aftermath of the most recent global economic downturn carry significant new costs for implementation and compliance, they were developed to serve the interests of the customer, the economy, and, eventually, the firms themselves. Safeguards and increased transparency both require more complete data management, analysis, and reporting, but help curtail risk and prevent the trends that led to the last financial crisis in hopes of avoiding future crises.

After a half-decade leveling-off period following the first Wall Street crash of the early-2000s, the cost to the United States wholesale banking sector for technology to comply with government regulations has increased by more than 40%. According to estimates by the CEB TowerGroup published in *The New York Times*, more than 85% of that increase (not the year-over-year increase, but the total lift over the 2009 baseline) can be accounted for by spending related to the Dodd-Frank Wall Street Reform and Consumer Protection Act that became law in mid-2010.[6]

The mandatory use of data also adds rigor to tests of new financial instruments and compels banks to plan for scale as their models take advantage of and become more reliant on larger data sets. Over time, the most proactive financial institutions driving strategy from big data will consider IT and infrastructure spend for regulatory compliance and reporting an option-value play, enabling downstream capabilities like predictive analytics and anomaly detection, rather than a source of opportunity cost.

As evidence, Accenture found in a 2013 survey of financial services and resources industry executives from North America and Europe that, despite half of respondents anticipating spending at least $50 million for compliance, 83% agreed that Dodd-Frank regulations will benefit their own company's customers, and 64% believed the spending on technology to comply with the new requirements will ultimately strengthen their competitive positioning.[7]

Technology spending by Wall Street banks and asset managers for compliance with government regulations



Source: New York Times

[6] Dash, Eric. "Feasting on Paperwork," *The New York Times*. 8 September 2011.
[7] Accenture. *Coming to Terms with Dodd-Frank*. 15 January 2013.

cloudera®

Mass Personalization: Tailoring Products and Services Across the Value Chain

Retail and commercial banking contribute the largest portion of revenue to the financial services industry, but weakened loan demand, widespread skepticism over fee structures, and scrutiny of new products have squeezed margins and increased competition to acquire and retain customers.

Due to a series of widely publicized missteps in recent years by some of the largest retail banks as they cope with new industry policies—most notably Bank of America's maligned five-dollar debit card fee in 2011—Edelman PR finds that financial services is the least-trusted industry in the world, rebounding to only 46% trust among U.S. consumers (47% globally, 29% in Europe) in 2013, following an all-time low of 25% in 2011 and a high of 68% in 2007.[8,9,10]

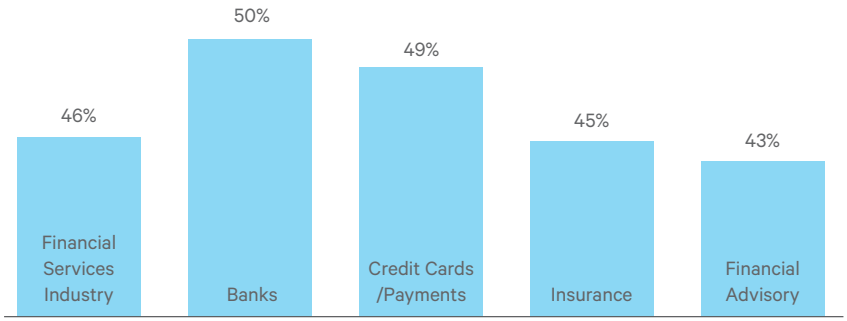
This distrust has not only affected banks' ability to do business, but has also bottomed out entrenchment and forced down switching costs, requiring banks to absorb more of the price of gaining and keeping new customers. According to a CorePROFIT research study, average customer acquisition costs U.S. retail banks more than $350 and requires each customer to carry an average balance of nearly $10,000 for the bank to just break even.[11] In an atmosphere of lower margins, higher costs, and less loyalty, banks seek methods to build a more complete picture of the customer and restore profits.

The consumerization of the retail banking industry through online and multimedia channels has increased the quantity of customer data that can be used to better segment the market, tailor products to target profiles, and create more marketing occasions. Mobile, app-based, and remote banking generate much more user information that, even when anonymized, can be paired with unconventional data sets, including ethnographic research, social media trends, and public utility usage, to build a rich platform for advanced analytics that lead to deeper insights and more sales opportunities.

The Deloitte Center for Financial Services summed it up in its *2014 Banking Industry Outlook*: "Delivering high-quality differentiated customer experiences will likely be critical in driving revenue growth… Banks that better leverage advanced analytics to translate big data into valuable insights could be better positioned to meet customer needs, offer a superior customer experience, and simultaneously deepen their product relationships with better cross-selling."[12]

No Sector of the Financial Services Industry is Trusted

In the U.S., insurance sector is most trusted; credit cards least trusted



Source: Edelman PR

[8] Edelman PR. *Edelman Trust Barometer 2013: Financial Services Industry Findings*. April 2013.
[9] Edelman PR. *Edelman Trust Barometer 2012: U.S. Financial Services and Banking Industries*. March 2012.
[10] Edelman PR. *Edelman Trust Barometer 2007*. February 2007.
[11] Andera and CorePROFIT. *The Future of Account Opening 2011*. June 2011.
[12] Deloitte Center for Financial Services. *2014 Banking Industry Outlook: Repositioning for Growth*. February 2014.

cloudera®

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?**

Data Consolidation and Multi-Tenancy

One of the most compelling benefits of applying a data hub strategy to the enterprise architecture of a financial services firm is the consolidation of many different data types, from many different sources, into a single, central, active repository. Accessibility, continuity, and scalability of the data management system are key to integrating Hadoop into your existing infrastructure because driving utilization increases the value of the data itself and the return on investment for the systems, freeing up cycles for more advanced analytics and budget for investments that grow the business.

Most organizations currently employ a variety of legacy systems to support their diverse data-driven goals, thus lack a unified view of their information. They have data warehouses for operational reporting, storage systems to keep data available and safe, specialized massively parallel databases for large-scale analytics, archives for cost-effective backup, and systems for finding and exploring information with the ease of a web search engine.

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."[1]
>
> Allstate

An open, integrated approach to unifying these systems around a central data hub allows each to share data more easily among the others for analyses that span formats, structures, and lines of business. Hadoop helps serve up more data than previously possible and for a much wider variety of business objectives while offsetting the overload more specialized architectures encounter as data volumes explode and new users, workloads, and applications are introduced. An enterprise data hub scales to accommodate size and diversity of data so that the existing systems on which the business relies can better fulfill the jobs for which they were conceived and implemented. Hadoop complements traditional architecture with a high-visibility, un-siloed, single view of all data that drives insights and facilitates more complete and consistent information-driven conclusions.

Over time, the ability to bring more users to where the data lives also drives up data quality because there is less need for duplication and transformation, which prohibit analytical consistency and alter the original data such that important information can be irretrievably lost. Furthermore, landing more applications and workloads on all your data in a unified system spanning existing architectures and tools drives much faster speed to insight, since it is inefficient and expensive to transform and relocate data between silos simply to answer new business questions.

When the data and workload mix accommodate a multi-tenant environment, IT teams need to address three critical facets:

- Security and governance
- Resource isolation and management
- Chargeback and showback capabilities

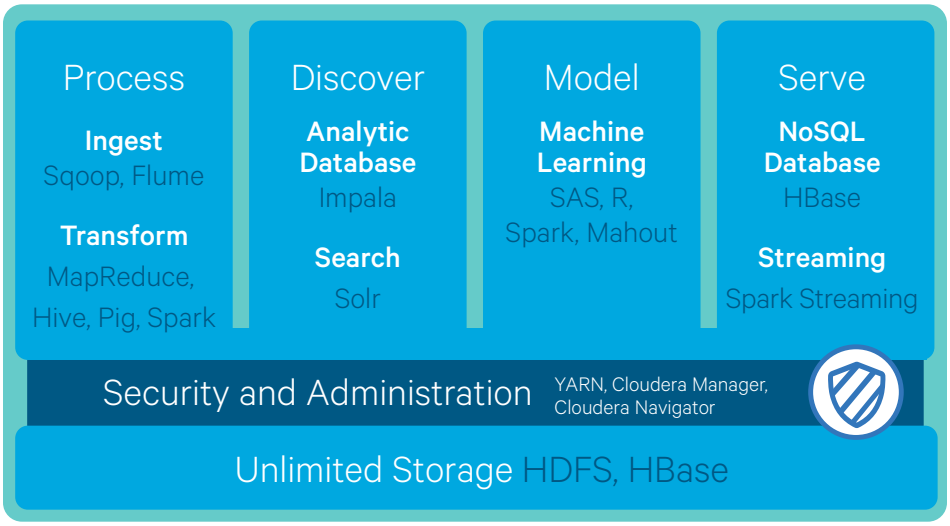
In order for a single data environment to support the operations of multiple users, administrators, and applications, a shared business situation requires manageability of power hogs and noisy neighbors, prevention against malfeasance and malpractice, and on-demand operations reporting to enable effective planning and maintenance. An enterprise data hub built on Apache Hadoop is designed with multi-tenancy in mind.

Security and Governance

Security Management Delegation. Organizations can use Apache Sentry (incubating), the open-source, role-based access control system for Hadoop, to delegate permissions management for given data sets. Using this approach, local data administrators are responsible for assigning access for those data sets to the appropriate individuals and teams.

Auditor Access. Cloudera Navigator is the first fully integrated data security and governance application for Hadoop-based systems. It provides a data auditor role that partitions the management rights to the cluster so that administrators can grant the audit team access to only the data needed and, thus, mitigate the impact to operations and security.

Data Visibility. Cloudera Navigator provides the only data encryption and key management natively integrated with Hadoop, both on-disk and in-use, such that only users with the correct access can view data, and even administrators without proper access cannot view stored data.



Source: Cloudera

Resource Isolation and Management

Resource Management. MapReduce, the batch processing engine in Hadoop, provides a scheduler framework that administrators can configure to ensure multiple, simultaneous user jobs share physical resources. More specifically, environments using Fair Scheduler provide maximum utilization while enforcing service-level agreements through assigned resource minimums.

Dynamic Partitioning. YARN is a Hadoop sub-project that allows resource dynamism across multiple processing frameworks. It becomes a building block for computing engines like MapReduce and Cloudera Impala— Hadoop's massively-parallel-processing structured query language (SQL) engine—to coordinate consumption and usage reservations and ensure fair allocation. Hadoop and YARN also support Access Control Lists for the various resource schedulers, thus ensuring that a user, application, or group may only access a specified resource pool at a given time.

Static Partitioning. Cloudera Manager, the first and most sophisticated management application for Hadoop and the enterprise data hub, supports a technology available on modern Linuxv operating systems called container groups, also known as cgroups. IT administrators specify policies within the host operating system to restrict a particular service or application to a given allocation of cluster resources.

Quota Management. HDFS, the distributed file system and primary storage layer for Hadoop, supports two quota mechanisms that administrators can tune to manage space usage by cluster tenants:

- **Disk Space Quotas.** Administrators can set disk space limits on a per-directory basis.
- **Name Quotas.** Administrators limit the number of files or subdirectories within a particular directory to optimize the metadata subsystem—the NameNode—within the Hadoop cluster.

Monitoring and Alerting. Cloudera Manager identifies dangerous situations like low disk space conditions and can send alerts to a network operations center dashboard or an on-call resource via pager for immediate response.

cloudera®

Chargeback and Showback

Cloudera Manager offers historical and trending disk and CPU usage. This information—which can be exported in common formats, such as Microsoft Excel, to financial modeling applications—can provide a strong foundation for an internal chargeback model or showback to illustrate compliance. These metering capabilities can also facilitate capacity planning and accurate budgeting for growth of the shared platform, thus ensuring that IT teams allocate sufficient resources in line with cluster demand.

Full-Fidelity Analytics and Regulatory Compliance

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.

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 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.

cloudera®

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 Security 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)	Requires reports to FinCEN on cash purchases of negotiable instruments of more than $10,000 (daily aggregate), suspicious activity, and tax evasion
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

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.[13]

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

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.

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 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.

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

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 and Apache Spark—the next-generation, open-source processing engine that combines batch, streaming, and interactive analytics on all the data in HDFS 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 cancellations. 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."[14]

[14] 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 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 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.

cloudera®

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.

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.

Out-of-the-Box Data Protections 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

cloudera®

Case Studies in Financial Services

The typical financial services adoption cycle for Hadoop begins with the operational efficiency and cost reduction use cases discussed herein: 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.[15] With so many organizations seeking a reliable, real-time, and affordable big data solution, what is the barrier to full adoption and production?

Top Current Use Cases (October 2013)	Top Future Use Cases (12-18 months)
1. Basic Analytics (59%)	1. Advanced Analytics (62%)
2. Business Intelligence (48%)	2. Business Intelligence (46%)
3. Data Preparation (46%)	3. Data Preparation (41%)

Source: Sand Hill Group

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 global information security group wanted faster query response and unconstrained access to better mine and analyze data in the RDBMS.

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

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 SAN, 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 within a few seconds and analyze them using Impala or even run models on streaming data using Spark's in-memory capabilities.

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.

Snapshot of Featured Technologies Included in an Enterprise Data Hub	
Apache Hadoop	An open-source software framework for storage and large-scale processing of large data sets on clusters of industry-standard hardware
HDFS	The distributed file system and primary storage layer for Hadoop
MapReduce	The batch processing engine in Hadoop
Cloudera Manager	The first and most sophisticated management application for Hadoop and the enterprise data hub
Cloudera Navigator	The first fully integrated data security and governance application for Hadoop-based systems, providing full discoverability, lineage, and encryption with key management
Cloudera Impala	Hadoop's massively-parallel-processing SQL query engine
Apache Spark	The next-generation, open-source processing engine that combines batch, streaming, and interactive analytics on all the data in HDFS via in-memory capabilities
Cloudera Search	The full-text, interactive search and scalable, flexible indexing component of Hadoop
YARN	A Hadoop sub-project that allows resource dynamism across multiple processing frameworks
Apache Sentry	The open-source, role-based access control system for Hadoop
Apache HBase	Hadoop's distributed, scalable, NoSQL database for big data
Apache Flume	Hadoop's service for efficiently collecting, aggregating, and moving large amounts of log data
Apache Hive	Open-source software that makes transformation and analysis of complex, multi-structured data scalable in Hadoop

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 differentiated marketing offers that communicate value for the most appropriate cross-sell and up-sell opportunities without diminishing margins.

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. www.cloudera.com.

cloudera.com

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