cloudera®
CONNECTED

Datameer®

Identifying Fraud, Managing Risk and Improving Compliance in Financial Services

DATAMEER CORPORATION

WEBSITE
www.datameer.com

COMPANY OVERVIEW
Datameer offers the first end-to-end big data analytics solution that enables business users to integrate, analyze and visualize data of any type, size, or source. Founded by Hadoop veterans in 2009, Datameer scales from a laptop to thousands of nodes and is available for all major Hadoop distributions. Datameer is based in San Mateo, CA.

PRODUCT OVERVIEW
Datameer's big data analytics solution built natively on Hadoop enables users to discover opportunities in their business by integrating, analyzing and visualizing data of any type, size, or source. Datameer's pre-built data integration, analytics and Business Infographics functionality requires no pre-built data models and can quickly analyze new data sources as needed with minimal IT involvement so users get insights at the speed of business.

SOLUTION HIGHLIGHTS
- Big Data analytics made easy
- Flexibility and scalability
- End-to-end platform: integration, analytics and visualization

Thanks in large part to the availability of data and the movement from in-person to online banking, today's financial firms look very different from those of yesterday. Because the scale of data is overwhelming traditional systems, firms must look to new technologies to unlock the power of their data. Cloudera helps companies implement and manage Apache Hadoop so they can derive competitive advantage from Big Data. Datameer provides Apache Hadoop-based analytics that help firms increase operational efficiency, minimize risks and meet financial compliance regulations.

On the Brink of Too Much Data

The digital revolution has dramatically changed the financial services industry. Four main factors have driven the need for financial services companies to collect, store and analyze massive volumes of data:

1. **Commoditization and digitization of financial products and services.** Consumers no longer need to visit their local bank to make deposits or complete other banking transactions. Buyers and sellers execute trades online instead of relying on floor traders and brokers. Individuals file their taxes using online tools versus meeting with a tax accountant to prepare and file. As the industry has increasingly moved online, it's become faster, easier, and more affordable for consumers to self-sufficiently handle their own banking and finance transactions.

   The result: financial services and products have become commoditized. Instead of establishing relationships with a local service provider, consumers often choose the most convenient and inexpensive online offering available. Personal connections and customer loyalty have become nearly obsolete. However, every digital action made by a consumer can be captured and analyzed by organizations that seek to understand their customers' behaviors and preferences as they would have traditionally done through face-to-face interactions.

2. **Increased activity.** The ease and affordability of executing financial transactions has led to ever-increasing activity and expansion into new markets. Individuals can make more trades, more often, because they can do so with the click of a button in the comfort of their own homes, or on the go from a mobile device. Individuals anywhere in the world can make trades in the US stock market over the internet. Increased access and ease of use translates into increased activity, which in turn translates into rapidly growing data volumes.

   It is important for banks, trading and investment firms, and other financial services organizations to be able to collect and analyze this information in order to accurately assess risk and market trends. This became painfully evident during the massive market crash of 2007-2008, when banks and brokerage houses scrambled to understand the implications in terms of capital leverage and their ability to model and refine liquidity management. With transaction rates as high as two billion per month, it has become impossible to create models that take into account multi-year data sets using detailed data. Building a model using sampled data sets may use as little as 100GB of data. Even

FOUR MAIN FACTORS ARE DRIVING BIG DATA NEEDS IN FINANCIAL SERVICES:

1. Commoditization and digitization of products and services
2. Increased activity
3. New data sources
4. Increased regulations

with large memory technologies that can accommodate one terabyte (TB) or more, that's a small fraction of the multi-petabytes (and growing) of data that firms have access to. Relying on data samples requires aggregations and assumptions, resulting in large inaccuracies in projections, limited visibility into actual risk exposure, instances of undetected fraud, and poorer performance in the market.

3. **New data sources.** The digital revolution has led to new sources of data that are complex to ingest, such as data from trade execution systems, weblogs, Twitter, blogs and other news feeds. This information, if and when combined with individual financial transactions and history, can help to paint a holistic picture of individuals, families, organizations, and markets.

   But bringing together large volumes of data from many sources and in a variety of formats — including both structured and unstructured — is cumbersome and impractical using traditional relational database and data warehousing technologies. The sheer scope and costs associated with bringing this data together and making it usable are often too complex for organizations to accomplish successfully.

4. **Increased regulations.** In recent years, federal stress tests have increased the demand for predictability and integrated solutions for capital asset management. Stringent regulatory compliance laws have been put in 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 companies have realized that the key to optimizing their business operations is in maintaining an efficient and large-scale data management infrastructure. But this is very expensive and complex to accommodate using traditional systems.

   Due to these four main factors, the scale of data that financial services companies need to manage today is overwhelming traditional systems. It's truly a Big Data problem: six years of publicly available market data can easily exceed 200TB, and the proprietary data collected by individual firms today often exceeds 5PB altogether. Firms must adapt to new technologies in order to store and analyze their data. Financial services companies' ability to access this data and turn it into action is driving competition; those that put all of their data to use have a significant competitive advantage over those who don't.

SOLUTION COMPONENTS

- Cloudera Enterprise Core: CDH + Cloudera Manager + Cloudera Support
- Datameer: Analytics application native to Hadoop

Today's Systems Can't Keep Up

Financial services companies struggle to integrate cross-functional data across their business. Traditional systems, limited to finite storage and compute resources, require complex extract, transform, load (ETL) processes and rigorous data schemas that must be revised each time a new question or data source is needed. These tools are restricted to structured data only; unstructured data must be given structure before it can be analyzed. Further, traditional systems become expensive as data volumes grow.

Bringing data in from different silos creates integration problems that traditional tools struggle with. Broker interaction data from application logs need to be combined with transactional data from the data warehouse and customer relationship management (CRM) tools. Call center log data needs to be combined with trading actions across systems and account types. Combining structured data from schema-driven relational database management systems (RDBMS) with unstructured data is not easily done using traditional tools. Companies need to be able to correlate these sources quickly and easily with tools that business analysts can use so they can quickly adapt to changing requirements.

The goal: to combine any number of relational and unstructured data about customer interactions and transactions into a single system. Financial services firms are looking for flexibility in combing and mapping very large data sets and using security to restrict who has accessed the most sensitive data.

A New Data Infrastructure

The world of finance set out to find new technologies that would allow them to manage and take advantage of larger data sets. They discovered the technologies deployed by large web and online advertising companies like eBay and Facebook. In short, they found Hadoop: open source software that enables distributed parallel processing of huge amounts of multi-structured data. The foundation for this infrastructure is industry standard, low-cost compute and storage capabilities. These include either off the shelf hardware managed in-house or public cloud resources offered as infrastructure-as-a- service. With Hadoop, no data is too big.

This infrastructure runs Cloudera Enterprise Core and Datameer as a single cluster. Cloudera Enterprise Core combines CDH, the 100% open source Apache Hadoop distribution, with management software and support that makes Hadoop a stable and reliable platform trusted by financial organizations. Datameer, an analytics application native to Hadoop, provides the next generation data management, analytics and discovery platform for Big Data.

The cluster servers are configured with required CPU, RAM and storage components, networked together with standard gigabyte Ethernet. There are two sets of roles for these servers: one set provides management and interaction with analysts and operators; the other set handles storage and processing for all of the data.

The management servers include Datameer, which provides analysts with capabilities to analyze data regardless of any variety, size or source. Datameer can reside on its own physical or virtualized server, either on-premise or in the cloud. Datameer integrates with existing LDAP and Active Directory infrastructures, as well as monitoring tools like Nagios and JMX-connected tools. Datameer runs natively on the Hadoop cluster and generates optimized MapReduce code, in the form of jobs, directly on the Hadoop nodes.

Similar to the way social networking companies find relationship links that are complicated to identify, anti-fraud teams search for connections that are implied by detailed trace data.

The other management servers handle administration capabilities using Cloudera Manager. Cloudera Manager is the management platform in Cloudera Enterprise Core, delivering granular visibility into and control over every part of CDH. Cloudera Manager empowers operators to improve cluster performance, enhance quality of service, increase compliance and reduce administrative costs.

All of the storage and processing nodes run agents or daemons that handle data placement, replication and analysis job execution. Each of the servers in this set is identically configured. In case of any failures, other nodes in the set take over responsibility for the data and processing of the failed nodes and continue with the work in progress. Servers can be added and removed in order to accommodate growth in data and processing requirements.

Today's top financial services firms have already deployed Cloudera's Platform for Big Data with Datameer to store, access and analyze large volumes of diverse and detailed data, increasing operational efficiency in several key business applications such as identifying internal fraud and reducing the cost of risk related regulatory compliance.

Use Case: Fraud Detection and Risk Management

Challenges in fraud detection have increased dramatically with the introduction of new access points to financial services offerings and increased sophistication of perpetrators. Furthermore, as companies expand into new markets, they face new risks that must be modeled.

With each new access point, firms are more susceptible to new methodologies and ever more complex cross-channel fraud. For example, customers commonly create separate accounts for different service offerings without linking them; they may create a checking account for a relative without also giving the relative access to their personal savings. With multiple accounts sharing access to the same financial products, a perpetrator has the opportunity to execute complex trades and transfers between independent accounts to mask fraudulent behavior.

The demand for more data driven risk modeling is increasing at a feverish pace. It is no longer sufficient to rely on brilliant quant staff with complex algorithms driven by data samples, which leave firms susceptible to hidden deficiencies and irreconcilable predictions. The most valuable tools available help firms combat fraud by providing detailed traces across all operational systems. Since perpetrators work hard to exploit gaps in financial systems, firms must be vigilant and self-aware of every place where they are exposed. By collecting detailed behaviors from online channels and automated systems, fraud detection teams can recombine logically linked accounts by looking for common patterns of money movement and related transactions. Similar to the way social networking companies find relationship links that are complicated to identify, anti-fraud teams search for connections that are implied by detailed trace data. Collecting detailed information on both customer and internal interactions leads to new models that help identify patterns of normal and suspect behavior.

The Cloudera and Datameer Difference

Together, Cloudera and Datameer foster better fraud identification and tracking by analyzing user interactions across systems to discover emergent patterns. The combined technologies also help firms perform detailed risk analysis for compliance purposes, saving significant costs and time while delivering a clearer picture of actual risk exposure across the organization. By identifying fraud more quickly and assessing risk more accurately, firms can avoid costly breaches and keep pace with regulations more cost effectively.

Customer Examples

- One major global financial services conglomerate uses Cloudera and Datameer to help identify rogue trading activity. Teams within the firm's asset management group are performing ad hoc analysis on daily feeds of price, position, and order information. Having ad hoc analysis to all of the detailed data allows the group to detect anomalies across certain asset classes and identify suspicious behavior. Users previously relied solely on desktop spreadsheet tools. Now, with Datameer and Cloudera, users have a powerful platform that allows them to sift through more data more quickly and avert potential losses before they begin.

- A leading retail bank is using Cloudera and Datameer to validate data accuracy and quality as required by the Dodd-Frank Act and other regulations. Integrating loan and branch data as well as wealth management data, the bank's data quality initiative is responsible for ensuring that every record is accurate. The process includes subjectingthe data to over 50 data sanity and quality checks. The results of those checks are trended over time to ensure that the tolerances for data corruption and data domains aren't changing adversely and that the risk profiles being reported to investors and regulatory agencies are prudent and in compliance with regulatory requirements. The results are reported through a data quality dashboard to the Chief Risk Officer and Chief Financial Officer, who are ultimately responsible for ensuring the accuracy of regulatory compliance reporting as well as earnings forecasts to investors.

A Closer Look at the Technology

Data Collection

With the majority of financial consumer interactions happening online, modern data sources for consumer interactions primarily involve web based platforms together with ATM and in store data. Additionally, new data sources are emerging based on mobile interactions. These data sources generate streams that include log files (also known as clickstreams) as well as transactions. Often, customer profile data and product information is integrated in order to make sense of user interactions. Datameer's data integration functionality makes it simple to access and integrate each of these sources into CDH.

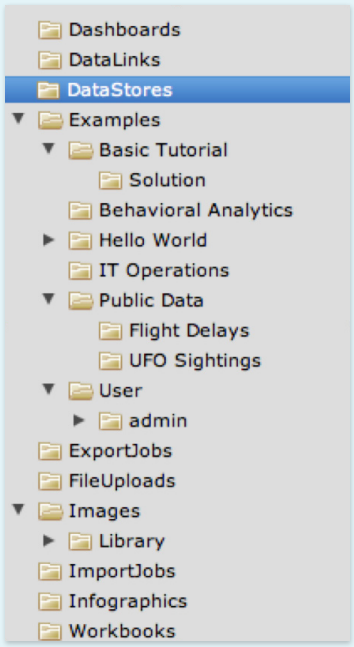
Web servers and application servers generate log files or clickstream data via standard logging mechanisms. These include configurable Java logging frameworks such as log4j and system level logging such as syslog. On Microsoft Windows, there are file and event system based logging systems. Each of these logging systems can be ingested into the Hadoop Distributed File System (HDFS) via Flume, a robust distributed log collection framework, or the data can be loaded directly by Datameer. Once in CDH, all data can then be accessed, joined, transformed and analyzed by Datameer.

Transactional data, customer profile data and product data are stored in relational databases. Each of these is updated regularly with changes and new information regarding customer transactions, changes to profiles or updates to products. Using Sqoop or Datameer's integration capabilities, this data is copied or streamed into HDFS on a regular basis. The data is synchronized and integrated with other data on a schedule configured and managed by Datameer.

Information Architecture and Collaboration

Ingested data is stored in HDFS and maintained in its raw form. The data may be partitioned by time and can then be joined, transformed or analyzed as needed. Regardless of analysis, the original raw data, by source, is maintained. Once imported into Datameer, it is also tracked using basic data lineage. All data operations in Datameer are captured so that users can determine what actions have been performed on the data and the corresponding data

FIGURE 1: SAMPLE USER DIRECTORY



results. Determining data lineage is an important component of regulatory requirements such as SOX, Dodd-Frank and Basel III.

Each user may store copies of their analysis in their home directory, directly under the top-level user directory. Users can download and export data from Datameer to external systems as well. For production jobs, the output data is stored under a system directory that the system administrator specifies. (See Figure 1.)

Data Analysis and Visualization

Datameer provides end users with data integration, analysis and visualization that directly leverages CDH. Providing hundreds of built-in analytic functions, the Datameer spreadsheet interface makes analyzing large complex datasets very straightforward. Built-in functions such as joins and transforms as well as complex math and statistical analytics give users powerful data transformation and analysis capabilities. Analysts use the spreadsheet metaphor instead of writing code and Datameer generates optimized MapReduce code directly on the CDH cluster behind the scenes.

The visualization functionality of Datameer enables end users to create powerful data visualizations ranging from simple dashboards to Business Infographics. Analysts design visualizations within a web-based canvas and configure features visually. Visualizations can be published to any URL and integrated with the existing portal infrastructure. Datameer utilizes HTML5, allowing end users to access visualizations from any device including smart phones and tablets.

FIGURE 2: DATAMEER END-TO-END FUNCTIONALITY

- > 40+ enterprise connectors
- > Metadata mgt., parsing
- > Job scheduling
- > 200+ analytic functions
- > Joins, transforms, math, etc.
- > Data lineage
- > WYSIWYG Designer
- > Dashboards and infographics
- > Any device

Conclusion

Using Cloudera and Datameer, financial services firms have access to a single, powerful platform providing big data analytics. Users are now able to analyze increasingly complex problems by unlocking the power of their data with business insights. The capability to understand and act upon these insights facilitates regulatory compliance while helping firms lower their costs, understand risk exposure and reduce incidents of fraud. Cloudera and Datameer are your partners in implementing world-class analytics to drive competitive advantage.