DATA CONSOLIDATION AND MULTI-TENANCY IN FINANCIAL SERVICES
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td>Security and Governance</td>
<td>4</td>
</tr>
<tr>
<td>Resource Isolation and Management</td>
<td>4</td>
</tr>
<tr>
<td>Chargeback and Showback</td>
<td>5</td>
</tr>
<tr>
<td>Big Data and an Enterprise Data Hub</td>
<td>5</td>
</tr>
<tr>
<td>About Cloudera</td>
<td>6</td>
</tr>
</tbody>
</table>
Introduction

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 Apache 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.

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.

### 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 Linux 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.

**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.

**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.