perforce
OpenLogic

WHITE PAPER

The Hadoop Service Bundle: An Open Source Alternative to Commercial Big Data Platforms

Executive Summary

The Hadoop Service Bundle from OpenLogic is a strategic open-source alternative to proprietary data platforms offered by vendors like Cloudera. By transitioning to a 100% open-source Hadoop stack — implemented and supported by OpenLogic — organizations can:

- Reduce annual data management costs by up to 60%
- Regain control over their data infrastructure
- Avoid dependence and lock-in from commercial vendors

This guide compares commercial Big Data tools with equivalent OSS, outlines the technical details of this solution and the data migration process, and includes a case study of Lumen Technologies to demonstrate real-world success.

Contents

3..... The Modern Hadoop Ecosystem

4..... Tooling Comparison: Commercial vs. Open Source

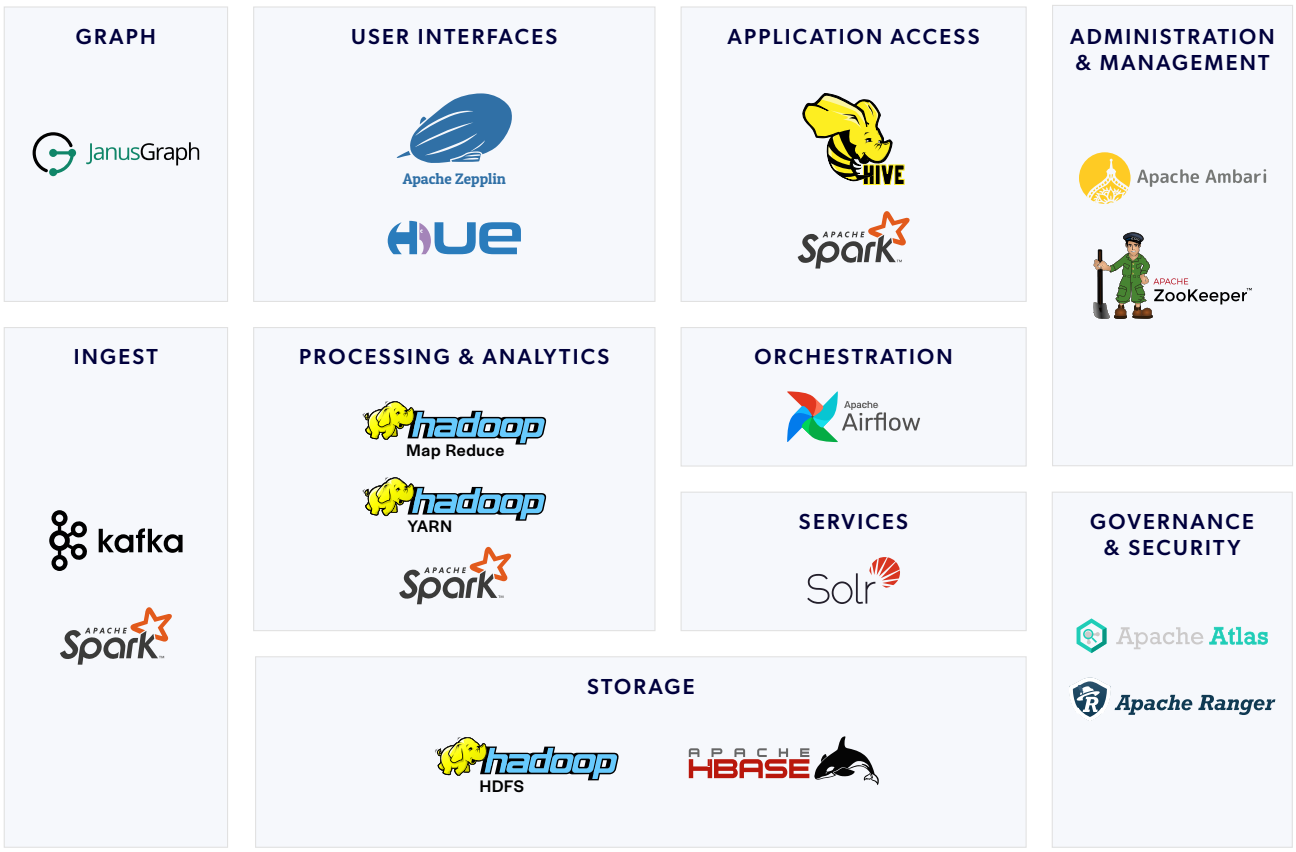
8 ..... Transitioning to the Hadoop Service Bundle

9 ..... Customer Succes Story: Lumen Technologies

9 ..... Frequently Asked Questions

The Modern Hadoop Ecosystem

Two decades have passed since the first incarnation of Apache Hadoop. In that time, the technology has matured dramatically, and the supporting ecosystem has also grown. Additionally, the evolution of established standards (JSON, YAML, RESTful APIs) have made seamless open source integration entirely feasible.



It is important to note that not every organization will need or want to deploy all of these technologies. A major advantage of open source over out-of-the-box solutions is the ability to customize your stack and only deploy the software you need.

Tooling Comparison: Commercial vs. Open Source

The foundation of Big Data solutions like the Cloudera Data Platform and Amazon EMR is open source Hadoop, with minimal proprietary tooling layered in to justify subscription costs.

For example, when you compare a Cloudera stack to a 100% open source stack, you can see there is considerable overlap — and that OSS components predominantly comprise Cloudera's platform:

Function	Sample Hadoop Stack	Cloudera
Cluster Administration	**Ambari**	**Cloudera Manager**
Cluster Data Services	**HDFS, MapReduce, Hive, HBase, Hue, Yarn, Spark**	**HDFS, MapReduce, Hive, HBase, Hue, Yarn, Spark**
Metadata Management and Data Governance	**Atlas**	**Cloudera Navigator**
Cluster Execution Services	**Airflow**	**Oozie**
Cluster Security Services	**Atlas, Ranger**	**Cloudera Navigator, Sentry**
Cluster Coordination	**ZooKeeper**	**ZooKeeper**

OSS Proprietary

Note:
Cloudera Navigator and Sentry are deprecated and have been replaced with Atlas and Ranger in greenfield implementations. However, since Cloudera customers with legacy infrastructures may still be deploying Navigator and Sentry, we have included them here.

Now let's compare some of the technologies (proprietary and open source) found in Cloudera's data platform vs. a modern Hadoop stack.

Cluster Administration: Cloudera Manager vs. Ambari

Key Differences

Feature	Cloudera Manager	Ambari
Multi-Cluster Management	Single interface manages multiple clusters (multi-tenancy)	Separate administrative server per cluster
Risk Profile	Higher risk of cross-environment errors (dev vs. prod) due to shared control plane	Strong environment isolation; reduced cross-cluster risk
Deployment Format	Parcels (proprietary binary distribution format)	Apache BigTop RPMs + Ambari Management Packs (Mpacks)

Cloudera Manager is a proprietary tool for Hadoop cluster administration. The open source equivalent is Apache Ambari, a project started by Hortonworks, which was acquired by Cloudera in 2019. Both have a web-based user interface and programmatic APIs that allow organizations to provision, configure, manage, and monitor Hadoop clusters and associated services.

While both Ambari and Cloudera Manager offer comparable feature sets for maintaining data clusters, distinct architectural differences affect implementation and future flexibility.

Multi-Tenancy and Risk Management

Cloudera Manager supports multi-tenancy, allowing the management of multiple clusters within a single interface. While this "one-stop shop" model offers convenience for navigating data collections, it introduces significant risks, such as human error where actions intended for a development cluster are accidentally applied to production.

Apache Ambari, on the other hand, requires a separate administrative server instance for each cluster. This approach aligns better with modern DevSecOps practices, utilizing virtualization and containerization to create smaller, controlled units. This separation enhances security through traditional boundaries like firewalls and improves performance management via horizontal scaling.

Deployment Flexibility and Standardization

In terms of deployment, Apache Ambari utilizes Apache BigTop RPMs and Ambari Management Packs (Mpacks) to standardize service deployments. This decouples the administrative interface from the service stack, granting organizations the freedom to create, update, and upgrade their own curated service stacks based on specific requirements. This open standard also facilitates the sharing of service binaries across other Big Data tools.

Cloudera Manager uses a proprietary binary distribution format called Parcels. This restricts organizations to vendor-curated stacks and timelines, limiting the ability to define custom stacks or share binaries across different platforms.

Metadata Management and Data Governance: Cloudera Navigator vs. Atlas

Key Differences

Feature	Cloudera Navigator	Atlas
Extensibility	Limited customization outside Cloudera ecosystem	Highly extensible via APIs, hooks, and custom type systems
Classification and Tagging	Basic governance capabilities	Advanced classification framework with custom tags and policies
Integration with Security	Separate from authorization tools	Deep integration with Apache Ranger for tag-based access control

Before it was deprecated in favor of open source Atlas, Cloudera Navigator handled data governance. It offered a wide range of features for auditing and compliance, from organization policy creation and tracking to regulatory requirements like GDPR and HIPPA. It also included data lineage tracking to look back upon data transformation and evolution, as well as metadata management for tagging and categorizing data to assist in searching and filtering.

Apache Atlas, which was also developed by Hortonworks, is what is now used for data governance and metadata management. While Cloudera Navigator was only applicable to Cloudera's data platform, Apache Atlas works across a broad range of Hadoop distributions and data ecosystems. It is extensible and integrates with other packages, like Apache Hive and Apache HBase.

Apache Atlas logs creation, modification, access, and lineage information about each data asset. It tracks who has accessed or modified data to provide an audit trail for compliance and monitoring purposes. Policies can be defined in Atlas to manage role-based access control (RBAC), attribute-based access control (ABAC), and data masking. To enforce these policies, Atlas integrates with Apache Ranger (another open source package in the Hadoop ecosystem).

Cluster Execution Services: Oozie vs. Airflow

Key Differences

Feature	Oozie	Airflow
Ecosystem Scope	Designed specifically for Hadoop jobs (MapReduce, Hive, Spark)	Orchestrates workflows across cloud, on-prem, Hadoop, databases, APIs, and SaaS tools
Workflow Definition	XML-based workflow definitions	Python-based DAGs (code-driven)
Error Handling and Retries	Basic retry and control flow	Advanced retry logic, branching, dynamic task generation

Note: Cloudera now supports Airflow, but only for customers using Cloudera Data Engineering (CDE) on public or private clouds.

At a time when more modern organizations are moving toward Apache Airflow for workflow, the base Cloudera product is still shipping with, and relying on, Apache Oozie. Apache Oozie workflows are tied to the Hadoop ecosystem and require unwieldy XML-based definitions.

In contrast, Apache Airflow is a more modern, flexible, and scalable workflow and data pipeline management tool that integrates well with cloud services and various systems beyond Hadoop. It has a friendly user interface, a strong community, and advanced error handling.

Cluster Security Services: Sentry vs. Ranger

Key Differences

Feature	Sentry	Ranger
Policy Management and Service Coverage	Service-specific policies (Hive, Impala) managed via service interfaces (e.g. Hue); limited ecosystem coverage (primarily SQL engines)	Centralized policy admin across services and dedicated web UI for unified policy management; broad coverage – HDFS, Hive, HBase, Kafka, YARN, Spark, and more
Access Control Model and Fine-Grained Controls	Role-Based Access Control (RBAC); basic database/table-level permissions	RBAC + Attribute-Based Access Control (ABAC); Column-level masking, row-level filtering, dynamic policies
Audit capabilities	Basic auditing	Centralized, detailed audit logging and reporting

Modern Hadoop and new Cloudera implementations use a combination of Apache Atlas and Apache Ranger for cluster security. Both products achieve significant improvements over the legacy Navigator and Sentry found in older Cloudera distributions.

Apache Ranger has a more user-friendly web-based interface that makes it easier to create and manage security policies. Unlike Sentry, Ranger includes built-in robust auditing capabilities for tracking events and activities across the platform, even outside of Hadoop proper.

Transitioning to the Hadoop Service Bundle

In this section, we'll walk through the implementation steps and explain what services are included post-installation.

Implementation Process

1. **Initial Assessment and Scope of Work**
   In the first phase of this remote professional services engagement, we collaborate with your team to analyze and document the current state of your environment and business requirements in order to develop a clear statement of work that details:

   - Schedule of work and cutover strategy
   - Change controls, change agents, and paths to approval
   - Responsibilities of the OpenLogic team and your internal team
   - Methodology (Agile vs. Waterfall)
   - Frequency of reporting progress and team meeting cadence

2. **Base Installation and Testing**
   This phase entails the side-by-side installation of Hadoop — on-prem, cloud, or hybrid deployment — plus up to 11 ecosystem components selected by you. This will be followed by testing and validation to ensure stability before cutover from other systems.

3. **Data Migration**
   During this phase, OpenLogic will oversee the cutover from your existing data platform to your custom Hadoop implementation and troubleshoot any issues that arise.

4. **Reference Architecture**
   After the migration, we will configure a reference installation to make it easier for you to scale or add new integrations in the future.

Technical Support

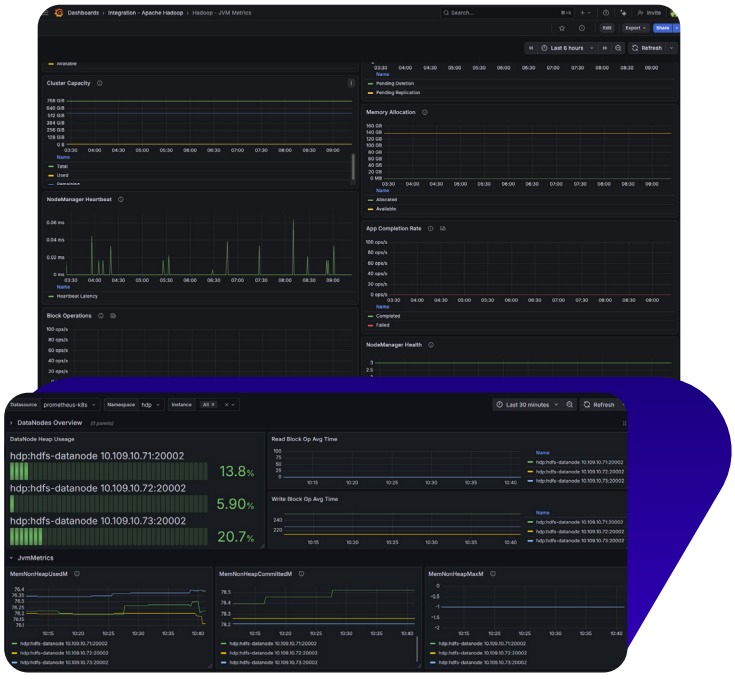
The Hadoop Service Bundle comes with OpenLogic Gold-Level Technical Support:

- 24/7/365 access to Enterprise Architects with 15+ years of experience
- Iron-clad SLAs for response times and workarounds/ solutions
- Support tickets can be submitted via phone, email, or online portal

Administration as a Service

OpenLogic will remotely monitor and remediate issues in your Hadoop stack. This includes:

- Supervising security patch updates of Apache Hadoop and configured services
- Conducting regular Health Checks and application performance monitoring
- Sharing insights into resource utilization within your Hadoop instance and offer recommendations on when to scale up or scale down



Customer Success Story: Lumen Technologies

"OpenLogic worked with our team effortlessly, ensuring our transition off our previous commercial platform was efficient and smooth. Their attention to detail, communication, and perseverance were instrumental in making this migration a success."

– Cindy Powell, Senior Manager of Software Engineering, Lumen Technologies

Lumen Technologies, a leading global communications services provider, successfully navigated the challenges of spiraling Big Data costs by switching from a commercial platform to the Hadoop Service Bundle. Lumen leveraged OpenLogic's deep technical expertise to migrate to a more cost-effective, open source stack that gives them more options for the future.

Key Outcomes:

Substantial Cost Reduction: Lumen achieved massive savings by avoiding an expensive commercial renewal when they migrated to the Hadoop Service Bundle.

Seamless In-Place Migration: Guided by OpenLogic, the transition to an open source Hadoop implementation was completed with minimal downtime.

Elimination of Vendor Lock-In: The move mitigated the risk of vendor dependency, providing Lumen with greater flexibility for future cloud and hybrid infrastructure strategies.

Expert Support and Collaboration: OpenLogic acted as an extension of the Lumen team, filling critical knowledge gaps and ensuring a well-documented transition.

Frequently Asked Questions

Q: Can we keep our data on-premises?

A: Yes. Organizations can host their data on-premises, in a private cloud, or in a hybrid architecture with no pressure to migrate to the public cloud.

Q: How does the Hadoop Service Bundle reduce costs?

A: By eliminating subscription licensing fees and deploying only what they need, organizations can reduce their annual Big Data management spend by up to 60% while maintaining enterprise-grade support and services.

Q: How far in advance should we engage your team to avoid paying for two solutions?

A: The sooner the better, but ideally six months to a year before your subscription renewal.

Q: How does the Hadoop Service Bundle prevent vendor lock-in?

A: Because the stack is 100% open source, you retain full control over your infrastructure. There are no proprietary extensions or licensing traps that limit portability or future flexibility.

Q: How does your pricing work?

A: Our pricing is straightforward — implementation and migration is a one-time fee, and support and admin-as-a-service are required in year one, with the option to renew in subsequent years.

Q: Is this solution only relevant if I have Cloudera?

A: While to date we have primarily migrated customers off of Cloudera, we are happy to explore the feasibility of migrating from any data platform built off of Hadoop. OpenLogic also provides technical support for teams already deploying open source Hadoop.

Rethinking your Big Data strategy and curious if the Hadoop Service Bundle might be a good fit?

Contact Us Today ▶

openlogic.com/talk-to-expert