HITACHI
Inspire the Next

Struggling To Unlock Value From Your Hadoop Data Lake?

Modernize Hadoop in Three Steps

E-BOOK

Contents

You are invested in Hadoop, but its complexities and market projections have been difficult.

Is Hadoop still viable for your big data needs?

Current State of Hadoop

Hadoop data lakes are on shaky ground. While it shines at hosting vast amounts of data, Apache Hadoop remains complex and costly to operate. It is difficult to govern on-premises data lakes and cumbersome to extract value from the data. With never-ending demand for faster return on investment (ROI), there is no time to navigate and pull diverse curated information out of these massive data lakes.

Hadoop market consolidation contributes to the uncertainty surrounding its data lakes: MapR Technologies, which distributed a Hadoop-based platform, was not able to keep afloat amid resource-draining Hadoop difficulties. It was acquired by Hewlett-Packard Enterprises (HPE) to complement its container-based approach for managing Hadoop. And the “merger of equals” between Cloudera and Hortonworks is evidence that there was not enough support in the Hadoop market for two very large competitors.

Cloud data lake options, such as those offered by Amazon Elastic Map Reduce (EMR), Microsoft Azure, Google Cloud Dataproc and Databricks, have also caused Hadoop’s on-premises influence to wane. Advantages of these options to enterprises include agile infrastructure, lower engineering costs, thrifty scalability and the ability to refocus from infrastructure administration to business value.

To be fair, it is possible to move data to any cloud at any time, but the cloud is not a problem-free landing zone either. A murky and less-than-useful onsite data lake will translate into the same data lake, full of dark and unstructured data, in the cloud.

.

Exponential Data Growth

Information growth is only expected to increase exponentially: Edge-to-cloud, industrial internet of things (IIOT), AI and machine learning (ML) technologies have become enterprise ready over the past decade. From the edge, Cisco predicts, globally, that the consumer segment's share of total devices and connections will be 74%.[1] Overall, IDC forecasts that the amount of data being stored (utilized storage) in the Global StorageSphere is expected to grow to 8.9ZB by 2024, representing a 2019–2024 CAGR of 20.4%.[2] In addition, once information is collected, data privacy becomes top-of-mind as major businesses have incurred significant fines for noncompliance: For example, the EU's General Data Protection Regulation demanded £183.39 million from British Airways and €50 million from Google[3] for noncompliance, and the U.S. Federal Trade Commission fined Facebook $5 billion for privacy violations.[4]

With Hadoop's massive capacity to ingest and safeguard this structured and unstructured data, is it prudent to move on from this technology?

Overall, IDC forecasts that the amount of data being stored (utilized storage) in the Global StorageSphere is expected to grow to

8.9ZB

by 2024, representing a 2019–2024 CAGR of 20.4%. [2]

[1] Cisco Annual Internet Report (2018-2023, Updated March 9, 2020

[2] IDC Press Release, IDC's Global StorageSphere Forecast Shows Continued Strong Growth in the World's Installed Base of Storage Capacity, May 13, 2020

[3] The Biggest GDPR Fines to Date, March 9, 2020, KYC-Chain

[4] F.T.C. Approves Facebook Fine of About $5, Celia Kang, New York Times, July 12, 2019

Where Do You Stand With Hadoop?

To wrangle this information deluge and drive business outcomes in the midst of uncertain financial markets that are operating in the shadow of COVID-19, you need the business agility and lower costs that come from efficiency and automation. The good news here is that there is still value to gain from your Hadoop investment. You'll want to keep some data on premises, but also unlock the value of the cloud data lakes. The challenge now is to remove complexity, modernize and extend Hadoop's longevity.

Hitachi Vantara offers a three-pronged approach to Hadoop modernization: reduce costs, catalog data, and achieve multicloud agility.

Hitachi Vantara's Approach to Hadoop Modernization

1. **Reduce Costs:** Move cold on-premises data into object storage.
2. **Catalog Data:** Index information with AI-driven tagging.
3. **Achieve Multicloud Agility:** Move relevant data to the cloud for higher business value.

Hitachi Vantara's Approach to Hadoop Modernization

Reduce Costs:
Move Cold On-Premises Data Into Object Storage

Challenge

Cold data, 60-80% of data by some estimates[5], is sitting unused or infrequently used in data lakes. For standard Hadoop data lakes, because Hadoop co-locates or couples storage with compute in its clusters, you are carrying not only the expense of storage for this cold data, but also the cost of compute. As your Hadoop footprint expands, your costs become uneconomical.

Solution

Use Hitachi's intelligent data tiering solution, Lumada Data Optimizer for Hadoop, to identify cold data within your Hadoop data lake and move it to object storage. This solution:

- **Reduces Hadoop operating costs.** By moving less-frequently accessed data to object storage, you free resources for active data, saving up to 80% on total cost of ownership (TCO). For data lakes of about 1PB, you can expect up to $1 million savings in TCO over a 5-year period.
- **Provides seamless data access.** There is no need to rewrite analytics applications. Simply and transparently tier data between Hadoop Distributed File System (HDFS) data and object storage to enable real-time analytics with Hitachi Content Platform (HCP).
- **Minimizes resource overhead.** Tiering allows you to optimize operations and achieve a lightweight footprint.

[5] Using Komprise to Archive Cold Data to Cloud Storage, Ranjana Bhadoria, Product Manager, Komprise, January 2020, posted on Google Cloud

[6] Veritas press release, Dark data exceeds 50%, creating major security blind spot for most companies, June 4, 2019

[7] DataOps Unlocks the Value of Data, 451 Research, January 2020

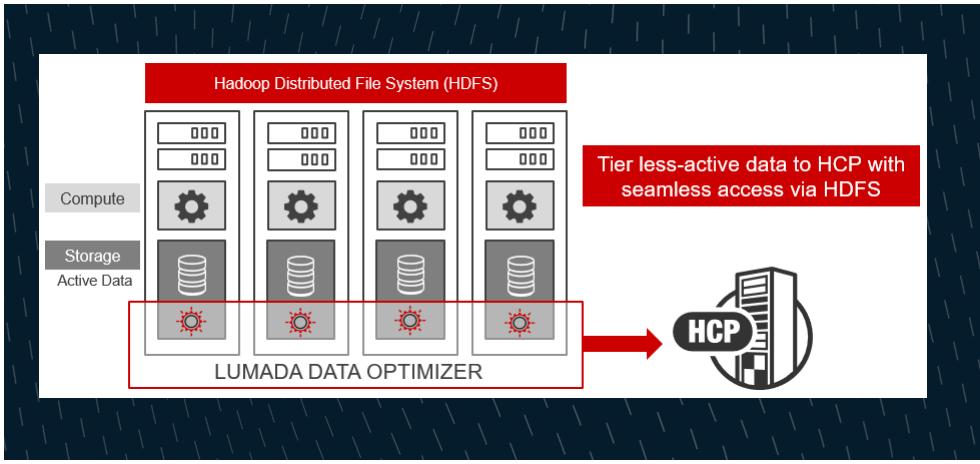


FIGURE 1. HITACHI'S INTELLIGENT DATA TIERING SOLUTION: LUMADA DATA OPTIMIZER FOR HADOOP

Catalog Data:
Index Information With AI-Driven Catalogs

Challenge

Just pulling as much data as possible into the Hadoop data lake from the edge as well as legacy data stores does not provide the knowledge that informs good business decision-making. The Value of Data study, conducted by Vanson Bourne for Veritas, found that 52% of all data within organizations remains unclassified or untagged, with limited or no visibility to the business.[6] And for organizations with some analytics in place, efficiency can be a challenge. In a recent 451 Research survey of North American organizations, more than 50% of respondents said it took more than three days to generate analytics insights, and 30% said it took more than a week.[7] These statistics are exacerbated by organizations with multiple data lakes. To realize value of that data, the next step is to ready it for nimble, predictive and advanced analytics.

Solution

Use Hitachi's Lumada Data Catalog to speed data discovery and metadata tagging to secure sensitive data, infer hidden relationships, and accelerate data self-service and insights across your Hadoop data lake. Automating discovery and tagging with machine learning and AI increases data visibility across data lakes or silos. Lumada Data Catalog brings you:

- **Fast, accurate insights.** Intuitive interface uses business terms to speed searches and accelerate time to data insights.
- **Governance for sensitive data.** Use data fingerprinting automatically to identify and secure sensitive data for compliance audits.
- **Self-service data discovery.** Leverage best-in-class crowdsourcing to reinforce accurate data recognition and data lineage by documenting the tribal knowledge throughout your organization.

Further shape data to your needs by adding Lumada Data Integration, a Lumada portfolio product. This modern data integration and orchestration platform allows you to access, prepare and blend data from any source.

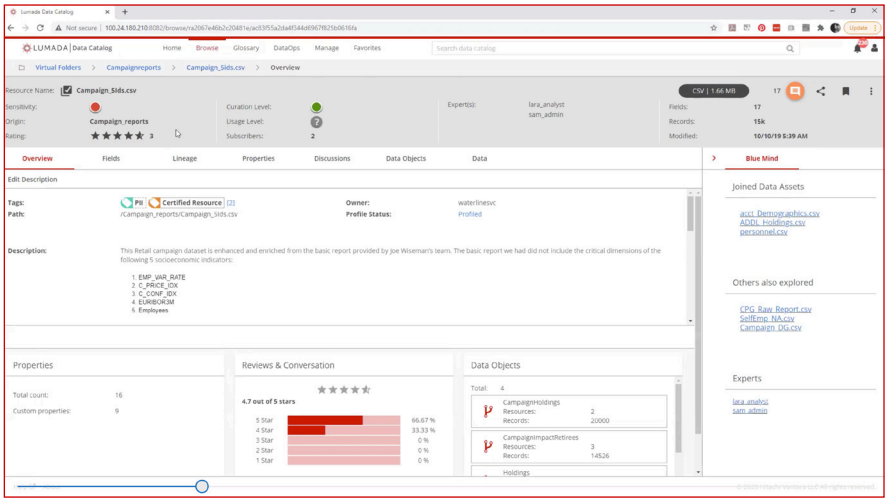


FIGURE 2. LUMADA DATA CATALOG'S INTUITIVE INTERFACE SUPPORTS FAST, EASY SEARCHES.

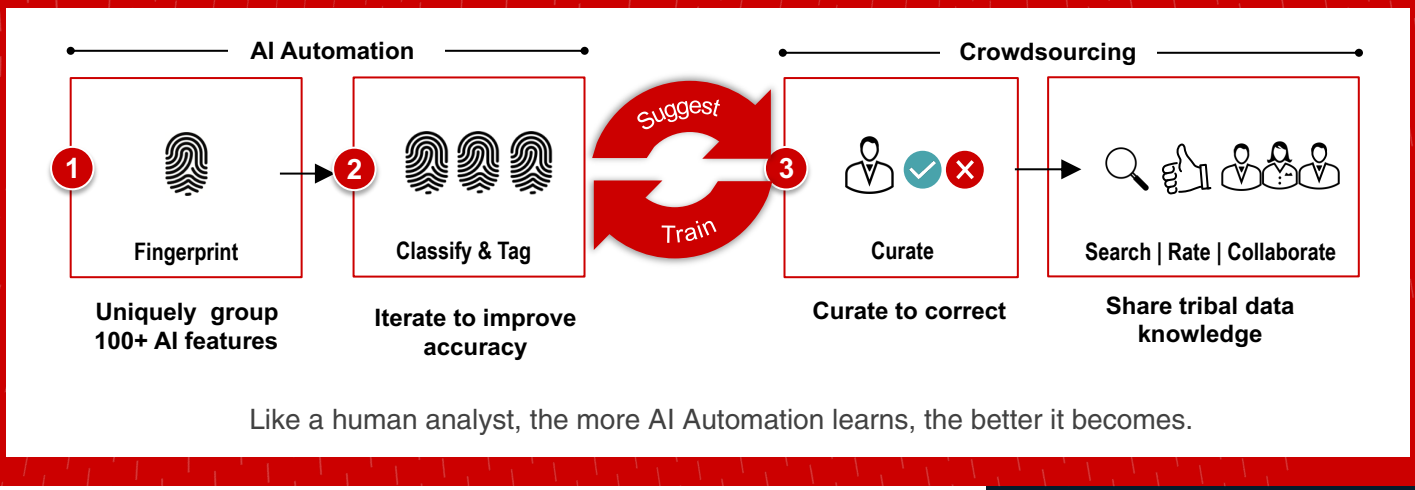


FIGURE 3. AI-DRIVEN DISCOVERY AND TAGGING: LUMADA DATA CATALOG EMPLOYS PATENTED DATA FINGERPRINTING TECHNOLOGY THAT INCORPORATES MACHINE LEARNING AND AI TO AUTOMATE DATA DISCOVERY AND METADATA TAGGING.

Achieve Multicloud Agility: Move Relevant Data to the Cloud for More Business Value

Challenge

Since it began as an on-premises platform, Hadoop and its data lakes have been supported in that rendition by most use cases. Offsetting some data to the cloud offers options that are increasingly popular

The most important new tech development of the passing decade has been the practical success of deep learning (popularly known as "artificial intelligence" or "AI").[8]

[8] 6 Predictions About Data In 2020 And The Coming Decade," Gil Press, Senior Contributor Enterprise and Cloud, Forbes, Jan. 6, 2020

among businesses, such as less setup and maintenance activities, and the choice to pay for server time used if opting for an as-a-service offering. But remaining difficulties are twofold: How do you narrow down the choices of cloud data lake providers? How do you make sure you don't get locked in by a single vendor?

Solution

Modernize your Hadoop data lake. Transform it into an agile, future-ready, multicloud data fabric with no cloud vendor lock-in through asset-based consulting and Hitachi's Lumada DataOps Suite:

- **Increase business agility.** Leverage proven solutions using industry-leading DataOps tools, accelerators and expertise to unlock new business value faster. Choose a solution that allows you to optimize, integrate and govern your data, from edge to core to cloud.
- **Reduce TCO.** Use a pre-integrated, tried and tested solution with better operational efficiency to speed cloud migrations and deploy and manage data lakes at reduced TCO.
- **Lower risk.** Drive transparency via a logical data fabric across data lakes by leveraging an enterprise data catalog.

Move away from complex, ungoverned and expensive Hadoop environments that are difficult to manage, risky and inflexible. Modernize and move toward AI-ready modern data fabrics through a combination of asset-based consulting and the Lumada DataOps Suite.

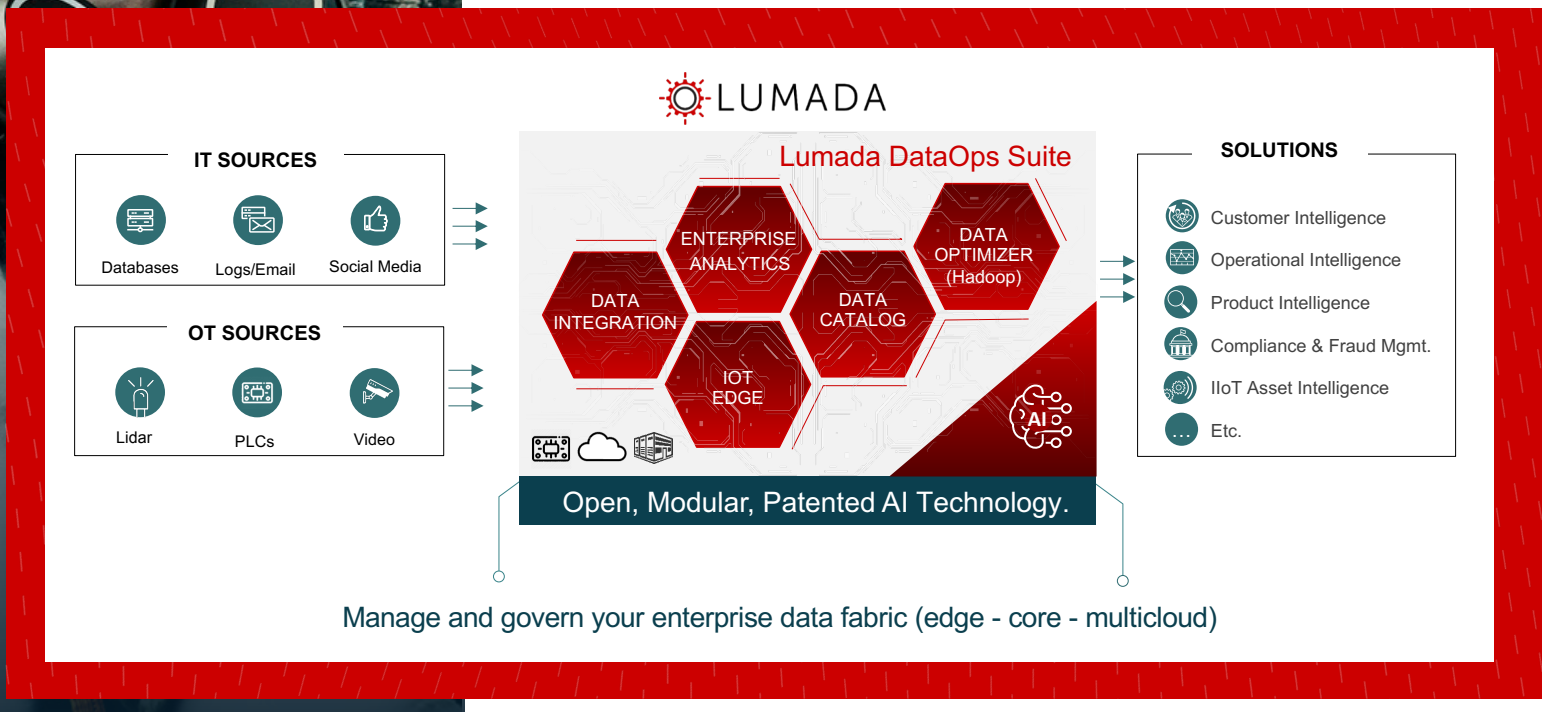
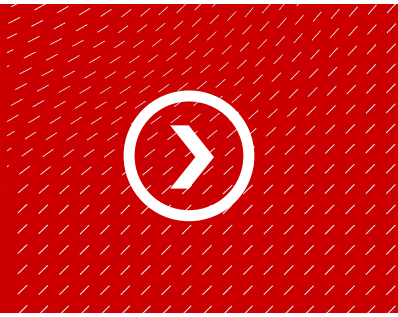


FIGURE 4. HITACHI'S LUMADA SOLUTION SUITE OFFERS AN APPROACH TO DATAOPS THAT IS COMPOSABLE, AUTOMATED AND COLLABORATIVE.

Get Value From Your Hadoop Data Lake: Modernize Now

Although Hadoop's complexity has stifled some of its progress, its open source nature supports continued evolution and flexibility. At Hitachi, we understand the importance of agility and combine that mind frame with the innovative products and services we've been delivering for more than 100 years. With our three-pronged approach to modernization, your Hadoop data lake will be infused with new life and cloud capabilities, delivering cost savings, indexed and usable data, and the agility to increase business value.

For more information about our strategy to modernize Hadoop in three steps, read our blog, contact your Hitachi Vantara representative, or visit www.hitachivantara.com.



We Are Hitachi Vantara

We guide our customers from what's now to what's next by solving their digital challenges. Working alongside each customer, we apply our unmatched industrial and digital capabilities to their data and applications to benefit both business and society.

Hitachi Vantara

Corporate Headquarters
2535 Augustine Drive
Santa Clara, CA 95054 USA
hitachivantara.com | community.hitachivantara.com

Contact Information
USA: 1-800-446-0744
Global: 1-858-547-4526
hitachivantara.com/contact

GEN-62-A BTD February 2021