acceldata

Acceldata Data Observability & Support for Hadoop

Contents

Executive Summary **3**

Introduction **3**

Modernization Challenges **4**

Strategy **5**

Cloudera Upgrade Path **5**

- *Modernization Path* **7**
- *Self-support Path* **9**
- *Acceldata Path* **10**
- *Decision Framework* **12**

Acceldata's Data Observability Solution **12**

- Acceldata Multidimensional Data Observability Platform for Hadoop **13**
  - *Incident Prevention: "Predictive Maintenance for Data"* **14**
  - *Performance Analytics* **15**
  - *Resource Efficiency* **16**
  - *Compatibility with the Hadoop ecosystem* **16**
  - *Use Case: Migration Validation* **17**
  - *Deployment Architecture* **17**
- Acceldata Support **18**
  - *Data Observability Platform Support: "Enterprise Support Subscription"* **18**
  - *Hadoop Support: "Enterprise Plus Support Subscription"* **18**
  - *Hadoop Administration: "Enterprise Gold Support Subscription"* **19**
  - *Project-based Support* **20**
  - *Support Model* **20**

Case Studies **21**

Summary **23**

References **23**

Executive Summary

Many organizations have invested heavily in mission-critical solutions built on the Hadoop ecosystem. Cloudera has announced the end-of-support for the legacy Hortonworks Data Platform (HDP) and Cloudera Data Hub (CDH). Organizations must assess the pros and cons of whether to migrate to the Cloudera Data Platform (CDP), modernize on alternative platforms (such as kubernetes or cloud offerings), or go it alone with a self-support model. Fortunately, a fourth option exists, Acceldata provides a Data Observability platform and support services that give organizations running on-premises Hadoop more flexibility with lower risk and cost for their current and future big data environments.

- Risk is eliminated by receiving superior support from Acceldata for legacy Hadoop environments
- Timing is now on the customer's side for deciding if and when migration occurs
- Cost can be significantly lower compared to the alternatives. Massive migration and re-engineering costs can be avoided. Acceldata support can be less than 50% of what Cloudera charges with superior service level agreements. Typically, infrastructure costs can be reduced by 10%-40% through efficiency gains enabled by the Acceldata Data Observability platform.
- Talent requirements are eased as organizations can rely on Acceldata's team of experts and extended support offerings that can include part-time or full-time site reliability engineers

This whitepaper provides insights into how organizations can formulate their data strategy as it relates to Hadoop, describes Acceldata's Data Observability platform and support offerings, and provides case studies of how Acceldata has helped organizations manage large complex Hadoop environments reliably and efficiently.

Introduction

Hadoop brought big data to businesses, enabling cost-effective, high-performance analytics on massive datasets for the first time. Over the last decade, every Fortune 50 company had either deployed a petabyte-scale Hadoop cluster running the open-source technology, or was making plans to do so. Today the technology landscape for big data is quite different. Cloudera stands alone as an on-premises Hadoop vendor, having bought Hortonworks and winning out against HP/MapR, IBM, Oracle, and others. Each of the major Cloud providers offer Hadoop-as-a-Service (AWS EMR, Google Dataproc, and Azure HDInsight). New vendors have emerged to offer commercially-supported alternatives for big data, such as Confluent for Kafka streaming, Databricks for Spark, and Snowflake as an alternative to Hive and Impala for data warehousing. Numerous other technologies and cloud-based solutions provide distributed processing and elastic scale, from S3 to Kubernetes to myriad NoSQL technologies.

Many organizations that have significant investments in on-premises Hadoop need a new strategy going forward, either due to end-of-life for support for their current stack or to take advantage of newer technologies and cloud offerings.

Modernization Challenges

Determining the right strategy for technology and infrastructure can be a complex exercise for organizations with large data environments. Many options exist for adopting new technologies that deliver attractive capabilities, capacity, performance, and pricing options. However, migrations can take a huge amount of time, talent, and money and can also introduce a lot of risk. For this reason, many organizations still use mainframes for mission-critical applications. If the solution meets the business needs, why change?

Cloudera has announced the end-of-support this year for the Hortonworks Data Platform (HDP) and Cloudera Data Hub (CDH) platforms. Organizations running these platforms will be forced into one of the following paths:

1. **Upgrade** to Cloudera Data Platform (CDP)
2. **Modernize** by migrating to another technology such as Amazon EMR
3. **Self-support** the existing platform if the license permits

Each of these options can present numerous challenges for organizations that either don't need or aren't ready to change. Fortunately, Acceldata provides a very attractive alternative:

4. **Acceldata** Data Observability and support for HDP and CDH

Strategy

Choosing the right path requires some analysis. Below is some information as well as insights that organizations have shared with us regarding the upgrade, modernization, self-support, and Acceldata paths.

Cloudera Upgrade Path

We have generally not seen many examples of customers choosing to upgrade to Cloudera CDP but that has increased with the end-of-support for the legacy platforms. We have witnessed tremendous growth and adoption of newer platforms, such as Databricks and Snowflake, with many organizations choosing to migrate from legacy Cloudera to those platforms. Many others have chosen to stick with their existing legacy HDP or CDH platform, having made significant investments in them to meet their business requirements. Some choose a hybrid path, keeping the legacy environment for some workloads and modernizing others. A major factor in choosing the right path forward is whether customers are currently using HDP or CDH.

CDH vs. HDP vs. CDP

The Cloudera Data Hub (CDH) is Cloudera's original distribution, which bundled free open source software, such as Apache Spark, along with its own proprietary software, such as Cloudera Manager. A Cloudera subscription provides OEM support, a valid license key and customer-specific credentials (username and password). These credentials provide access to Cloudera's private software repositories and the license key enables enterprise features on CDH. CDH is available in an express edition that has a light version of Cloudera Manager. CDH Express is made available for free use with certain restrictions (e.g. up to 100 nodes).

CDH competed with the Hortonworks Data Platform (HDP) until Cloudera bought Hortonworks. The Hortonworks business model was based on providing OEM software support by bundling open source packages together into the HDP platform. HDP was released as free open source software (FOSS) that anyone could download and use, with or without a support contract from Hortonworks. HDP users can still use the platform for free today and into the future. After the Hortonworks acquisition, Cloudera continued to support both CDH and HDP product lines, but that is coming to an end.

Cloudera has developed a new distribution, the Cloudera Data Platform (CDP). Ironically, CDP is much closer to HDP than it is to Cloudera's own original CDH. CDP is in many ways a repackaging of HDP but it now includes proprietary software and is no longer free to use. In general, CDH users face a bigger impact from end-of-support than HDP users due to major differences in the underlying technology and the lack of a free open source license. Cloudera's strategy reduces the number of components in the Cloudera portfolio and ensures that all users are paying customers. For organizations wishing to minimize the impact to their existing environment they often consider the following scenarios:

- **HDP without Cloudera -** This requires no change to existing HDP environments and allows organizations to either self-support, purchase support from Acceldata or seek support and services from another 3rd-party, such as a systems integrator (although few options exist).

- **HDP to CDP -** CDP has more in common with HDP than with CDH but there can be significant migration challenges, nonetheless. Some migrations will require a multi-step upgrade and testing process which adds time, complexity and risk. Some configurations and technologies are no longer available in CDP requiring solutions to be refactored or even re-architected.

- **CDH Express -** If they meet the requirements, e.g. less than 100 nodes, for example, they may use this distribution free of charge but support is ending from Cloudera and certain components, such as Cloudera Manager are not part of the open source community which introduces risk.

- **CDH to CDH Express -** Depending on the components used, this migration may introduce the least amount of refactoring for those running CDH.

- **CDH to CDP -** This is a major conversion with risks and costs that can be similar to migrating to other platforms such as Amazon EMR. Organizations that are dependent on Hive or Sentry may require significant refactoring. For example, at present, CDP does not support using the Spark engine for Hive, only Tez.

- **CDH to HDP without Cloudera -** This is similar to a migration to CDP but the result is they will be on a free open source platform that they can continue to run on-premises. Cloudera support is ending for HDP but Acceldata provides support packages as an alternative, with superior SLAs and at a substantially lower cost. Moreover, Acceldata's data observability solution for Hadoop has helped organizations increase reliability, performance, resource utilization, and worker productivity.

Clearly, there are many different scenarios. Below are some key decision factors that have been shared with us by CDH and HDP users:

- **Risk:** Hadoop is a very complex platform and many organizations have spent enormous engineering resources to get to a stable state for their mission-critical data and analytics solutions. The amount of risk in a migration varies depending on the scenario and the organization's environment. Unfortunately, for HDP and CDH customers, every scenario provided by Cloudera requires a migration to a new platform. For many, the potential risk in changing platforms, be it Cloudera or a more modern alternative, is too high. For those organizations, Acceldata technology and support minimizes or, in many cases, completely eliminates this risk.

- **Timing:** Cloudera may offer a limited extension for supporting existing HDP and CDH customers. However, this will almost certainly come with conditions such as an increase in support costs, a purchase agreement for CDP, a services agreement, aggressive timeline for the upgrade; likely all of the above. For those that are unprepared for a modernization initiative or self-support, switching to Acceldata for support may provide a safe and cost effective path that's better aligned to customer timing.

- **Cost:** Similar to risk, the cost of upgrading will vary, depending on the scenario. The cost of re-engineering solutions to use a different technology can be very high. In many cases, paid Cloudera professional services fees are required. There may also be a duplication of infrastructure costs until the upgrade is complete, similar to other migration paths. For some, the total cost to upgrade to CDP can be comparable to migrations to other platforms. Some organizations have reported significant increases in the cost of Cloudera licensing after moving to CDP, making alternatives more attractive. Some have concerns that the cost and risk of migrating off of CDP in the future may be higher as organizations will be locked into proprietary software that may become less compatible with other offerings.

- **Talent:** The talent pool for legacy technologies invariably shrinks over time. Many engineers, particularly new engineers or those with exceptional talent, prefer to focus on the latest technologies and avoid "getting stuck" developing and supporting legacy technologies. This makes the self-support path potentially daunting. On the other hand, migrating to different technologies requires hiring and/or retraining.

Based on the market trends we are observing and engagement with Hadoop users, most organizations prefer to leverage their working HDP and CDH platform until the timing is right to make a change. This is a major driver for customers in choosing Acceldata to provide continued support for legacy Hadoop environments. When organizations choose to make a change, most choose to modernize, typically to cloud services such as Databricks and Snowflake. Deeper analysis often leads some organizations to a hybrid path with some workloads remaining on Hadoop and others migrating. Many Hadoop users state that Cloudera has not kept up with innovation found in the cloud and even the versions available for established technology are more limited, particularly so with the new CDP offering. Some organizations, analysts and engineers question what the future of Cloudera will look like since being acquired by private equity. Many of Acceldata's engineers are former Hortonworks and Cloudera employees who can assist with assessing the Cloudera upgrade path. Next, let's explore the other paths.

Modernization Path

For many organizations, modernizing some or all of their big data environment is either under way or in the planning stages. For large, complex environments the cost of re-engineering solutions to use a different technology can be so high that modernization is restricted to new solutions only, resulting in a hybrid of legacy and modern technology. For those needing to migrate existing solutions, modernization can take different forms:

- **Rehosting** or "Lift-and-shift" approaches aim to move systems to the cloud, largely, as-is. Some technologies, such as Spark, Kafka, and Hive exist as cloud services that can make this possible. This approach can reduce the risk, time and effort of migration but may result in a higher operating cost or poorer performance than the options below.

- **Refactoring** aims to optimize in some areas while keeping others as-is. For example, it is very common to replace the Hadoop filesystem (HDFS) with cloud storage such as S3, which is superior in terms of cost and scalability. Refactoring strikes a balance between the time, risk and cost of migration vs. the longer term cost and capability improvements of alternative technologies.

- **Rearchitecting** solutions requires building new systems and migrating solutions to them. This can involve considerably more time, resources and risk but can pay off in the long run in terms of cost or business benefits from performance or other capabilities enabled by the new architecture.

Two of the most common modernization paths include Hadoop to Databricks and/or Hadoop to Snowflake. Note, Acceldata provides Data Observability for all three technologies to improve migrations and ongoing operations.

- **Databricks** provides a fully-managed cloud environment for big data processing and analytics. It is largely based on Spark which can simplify migration for those using Spark on Hadoop. Databricks brings typical cloud benefits in terms of eliminating hardware and infrastructure management, practically limitless scale, agility to provision resources with the click of a button, and consumption-based pricing. Organizations with complex processing and advanced analytics of unstructured data often choose Databricks. Databricks promotes a combination Data Lake / Data Warehouse architecture, the so-called "Lakehouse". This provides a SQL interface to support more traditional analytics and BI tools. Some organizations determine that a hybrid approach that keeps some or all of their existing workloads on-premises and avoids migration costs and risk makes the most fiscal sense. Moreover, while many platform administration tasks are alleviated by the cloud, the complexities of advanced data engineering and ML at scale mostly remain. If Databricks is part of your current or future architecture, go to https://www.acceldata.io/databricks for more information and to connect with Acceldata to learn how Acceldata's Data Observability Cloud solution for Databricks can support your Databricks journey.

- **Snowflake** provides a cloud data warehouse that is often a replacement for data warehouses built on Hive or Impala. For organizations working primarily with structured data, Snowflake may be the right option. Similar to Databricks, Snowflake brings many common advantages of the cloud. Generally, users of Snowflake are happy with the functionality, reliability and scalability of the platform. The most common concern we hear is that the cost can be unexpectedly high. Again, many organizations may look to keep their existing on-premises warehouse and take a more cautious and phased approach with Snowflake. If Snowflake is part of your current or future architecture, go to https://www.acceldata.io/snowflake for more information and to connect with Acceldata to learn how Acceldata's Data Observability Cloud solution for Snowflake can support your Snowflake journey.

- **Hybrid** environments with Databricks and Snowflake are also common, where Databricks performs the data processing and lands the results in Snowflake for consumption.

There are many other technology options available to replace different parts of the Hadoop ecosystem, including streaming, NoSQL and others. Identifying the right modernization strategy can be difficult which is why it is often one of the top practice areas for advisory services providers.

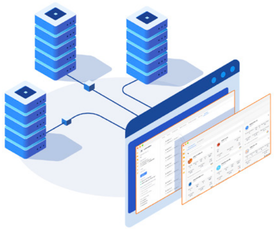
- **Risk:** The risks involved with migrations vary widely depending on the size, complexity and compatibility of the source and target environments. Risk can be a deciding factor in determining whether to keep, rehost, refactor, or re-architect environments. Minimizing risk can be achieved by investing in expertise and technology to validate performance, functionality, data reconciliation and other metrics. Acceldata brings both expertise and technology that help reduce the risk of modernization initiatives (see the Migration Validation use case section below).

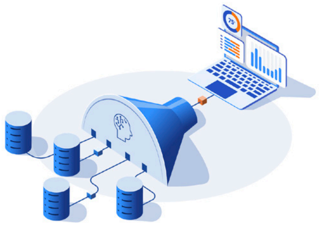
- **Timing:** End of support from Cloudera creates a very tight timeline for modernization initiatives. This can result in organizations having to pursue a simpler, less risky approach to meet the deadline but results in a more expensive or less capable target environment. This may lead to additional modernization phases that introduce more time, risk and cost to get to the ultimate desired end state. A rushed modernization effort introduces additional challenges, including greater risk, additional resources and cost, and potential business impact as other projects are delayed to prioritize the migration.

- **Cost:** If the organization requires the new capabilities, scale, or pricing model of a modernized environment then there may be a return on investment. If the existing HDP or CDH environment already meets business needs then a forced migration may simply be an added expense with a negative return. An accurate apples-to-apples cost comparison between running Hadoop on-premises and in the cloud can be difficult to determine. Refactoring and/or more resources may be consumed to achieve the same level of performance in the cloud, which adds to cost. There is potential for savings with consumption-based pricing that allows you to only pay for resources when needed. However, many organizations need to make resources available 24x7 which essentially means they will need to keep things on all the time. This "always on" requirement may also require excess capacity to be provisioned to accommodate unexpected load. This puts organizations essentially back to a model that's close to what they have on-premises but potentially at a higher operating cost. For organizations with large, mission critical, always-on infrastructure, on-premises infrastructure can be much less expensive than the cloud, particularly when leveraging free open source software. That's why organizations with very large data environments, such as telcos and financial services, are still such big users of open source technology running on-premise. Cost optimization initiatives have shifted many others from an "all cloud" strategy to a hybrid and multi-cloud strategy.

- **Talent:** New skills will be required to operate new technologies and environments. Additional business and technical expertise may be required for strategy, design and execution of a modernization initiative. Organizations that fully transition to the cloud may be able to reduce the skill sets required in-house for operations as capabilities are consumed "as-a-service." It can take considerable time to make a complete transition to the cloud and some organizations never fully transition due to financial, risk or other factors, leading to tech sprawl which can be difficult to manage from a talent perspective.

In summary, timing and cost are the two biggest challenges organizations often cite when taking a modernization path that is forced by the end-of-support for HDP or CDP. Rushed modernization initiatives can lead to suboptimal target environments, unnecessary risk, business disruption, future re-work and cost, and other challenges. Moreover, if the existing HDP or CDP environment already meets the needs of the business, a migration may be an unnecessary expense.





Self-support Path

Many organizations leverage unsupported, free open source software (FOSS) of all types and Hadoop is no different. CDH Express (under certain conditions) and HDP are free to use without a paid subscription from Cloudera as are the underlying components from Apache. Soon all Cloudera offerings will require a paid license. Many organizations choose a modernization path that includes FOSS, such as migrating to Spark on Kubernetes running on-premises.

- **Risk** will obviously increase if dropping commercial support for HDP, CDH or any other product. However, the risk is not the same for all organizations. Some rely on Cloudera support for smooth daily operations. Others only require access to patches and upgrades. Some are running mission-critical streaming data pipelines with very tight service level objectives (SLOs), while others have generous batch windows and SLOs. Some organizations have challenges with acquiring or retaining talent for l egacy systems and need reliable access to subject matter experts. Some have concerns with waning support from the community or from Apache for aging technologies. Some organizations have internal or regulatory requirements mandating commercial support. These and other factors weigh heavily on whether a self-support path contains an acceptable level of risk.

- **Timing** can be a challenge if the organization is required to downgrade or migrate to FOSS. Migrations can take time and talent and introduce additional risk. For organizations that are already running HDP there may not be any changes required. Organizations may need additional time to hire more staff to self-support.
  Cost for the overall system may be lower but not in all cases. Additional support staff may be required to offset risk.

- **Talent** may need to be acquired to mitigate risk. Organizations may want to maintain redundant skill sets to reduce risk if an SME leaves. Organizations may want to upskill to have top experts in house. As technology ages, expertise can be harder to attain and retain. Many engineers prefer to focus on the latest technology. The business may prefer to leverage top talent for revenue generating activities instead of maintenance and support.

The takeaway is that for those organizations that have the talent and can handle the additional risk, self-support can be an attractive path. For the rest, their best option is one of the other paths.

Acceldata Path

Acceldata provides product and support offerings specifically designed for Hadoop to help customers overcome many of the drawbacks of the options above as well as challenges in general with operating their existing Hadoop environments. Acceldata customers leverage offerings in a variety of ways. Common scenarios include:

Acceldata Data Observability Platform + Acceldata Support for Hadoop provides technology and support to successfully operate Hadoop distributions and technologies on-premises or in the cloud without a Cloudera license

Acceldata Data Observability Platform + Cloudera Licenses delivers improved reliability, performance, data quality and resource efficiency for Cloudera customers. Some Cloudera customers also pay for additional Acceldata Support and services.

Acceldata Data Observability Platform Stand-alone delivers improved reliability, performance and resource efficiency for customers wishing to self-support on free versions of Hadoop distributions and technologies on-premises or in the cloud without a Cloudera license.

Note, Acceldata also provides a **Data Observability Cloud** offering that supports many other technologies such as Snowflake, Databricks and others. All of the options are described in further detail in the solutions section below.

Acceldata offerings allow organizations to lower their costs and safely end their contract with Cloudera with continued support and maintenance for their legacy Hadoop environments. Customers cite the following benefits:

- **Risk** is eliminated from self-support or a rushed migration to new or unproven technology. If organizations are leveraging proprietary Cloudera technology, such as CDH, they may need to migrate to free distributions such as CDH Express (note restrictions) or HDP. Acceldata provides services to assist with these migrations and has experience with many successful migrations.

If and when organizations choose to modernize some or all of their stack, Acceldata's data observability platform reduces risk by allowing organizations to benchmark and validate performance parity as well as perform data reconciliation to ensure all data is migrated intact. Performance and data consistency validations can be performed between Hadoop and many other non-Hadoop systems to support a wide array of modernization strategies. See the Migration Validation Use Case below.

- **Timing** is now controlled by the customer, not their vendor. Organizations can take the time needed to properly plan, coordinate resources, and phase modernization and migration efforts or simply continue to leverage a working solution until a modernization strategy makes fiscal sense.

- **Cost** for support from Acceldata can be less than 50% of existing support costs from Cloudera. Most organizations report that the cost for support and licenses for CDP are significantly higher than for HDP and CDH. The cost of a rushed migration can often be much higher than a well-planned and executed migration. Many customers have cited that they do not realize benefits from CDP that justify the cost of upgrading. Migrating to CDP in the short term, only to modernize and migrate to another platform later can unnecessarily double migration costs.

- **Talent** is scarce and most organizations want to apply their talent to revenue-generating business initiatives, not for migrations, support and maintenance. Acceldata maintains a deep bench of expertise in Hadoop so organizations don't have to.

Decision Framework

The following chart compares each option as they relate to time, talent, cost and risk.

⚠ ✓

Decision Criteria	Upgrade to CDP	Modernize	Self-support	Acceldata
Risk	Migration risk due to CDP re-architecture	Migration risk	Unsupported	24x7 support
Timing	Cloudera's schedule	Cloudera's schedule	Cloudera's schedule	Customer's schedule
Cost	Increased price + migration costs	High cost for expedited migration	No software licenses, additional support staff may be needed	Cost-effective support offering
Talent	New platform, deep expertise required to use CDP, less talent/staff needed for CDP cloud	Can be easier to use/operate than CDP (e.g. Snowflake) but new skills may be needed	Deep expertise and redundancy needed to minimize risk	Leverage existing skill sets plus leverage Acceldata exceptional talent

Next, let's explore Acceldata's solution in detail.

Acceldata's Data Observability Solution

Acceldata is a market leader in the emerging "Data Observability" category. Acceldata's Data Observability solution provides monitoring, analytics, and automation to improve key metrics in large-scale data operations and management, including performance, reliability, data quality, productivity, resource efficiency, and cost optimization.

Data Observability Platform for On-premises Hadoop

Acceldata provides an on-premises Data Observability solution specifically for the Hadoop ecosystem. The complexity of managing and operating Hadoop clusters is greatly simplified by Acceldata's platform as it is purpose-built for accelerating root cause analysis of issues and automating the correlation between Hadoop services' configurations, resource consumption and load patterns.

Data Observability Cloud

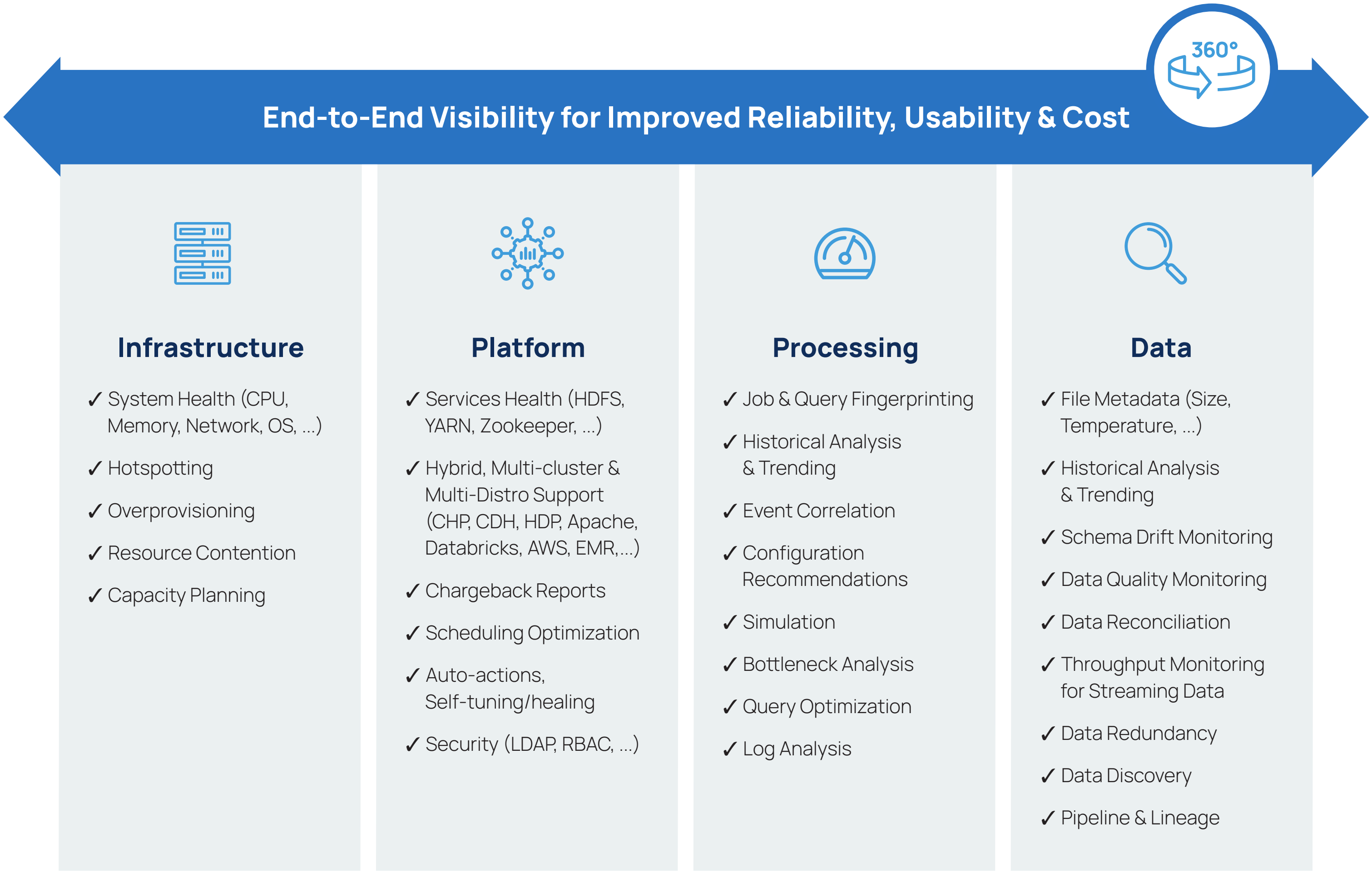
Acceldata also provides a cloud-based Data Observability platform for a wide range of technologies including Snowflake, Databricks, and others.

Acceldata Support and Services for Hadoop

Acceldata's founders and many of its engineers are data pioneers who have worked at Hortonworks and Cloudera and helped build and support their innovative data platforms. Acceldata extends its pool of talent with flexible support and services offerings to help organizations manage and optimize their Hadoop environments.

Acceldata Multidimensional Data Observability Platform for Hadoop

Even the most competent Hadoop shops will readily admit that the platform is complex. This complexity comes in multiple forms. The classic big data challenges of volume, velocity, variety, and other "V's" still ring true today. Beyond the data itself, complexity also exists in the infrastructure, platform, and data processing. Moreover, challenges arise not just within these four pillars, but also in the interplay between them. That's why Acceldata built the first and only multidimensional data observability solution that addresses the entire Hadoop ecosystem, including infrastructure, platform services, processing and data. The table below provides a high-level overview of capabilities within these pillars.



End-to-End Visibility for Improved Reliability, Usability & Cost

Infrastructure

- ✓ System Health (CPU, Memory, Network, OS, ...)
- ✓ Hotspotting
- ✓ Overprovisioning
- ✓ Resource Contention
- ✓ Capacity Planning

Platform

- ✓ Services Health (HDFS, YARN, Zookeeper, ...)
- ✓ Hybrid, Multi-cluster & Multi-Distro Support (CHP, CDH, HDP, Apache, Databricks, AWS, EMR,...)
- ✓ Chargeback Reports
- ✓ Scheduling Optimization
- ✓ Auto-actions, Self-tuning/healing
- ✓ Security (LDAP, RBAC, ...)

Processing

- ✓ Job & Query Fingerprinting
- ✓ Historical Analysis & Trending
- ✓ Event Correlation
- ✓ Configuration Recommendations
- ✓ Simulation
- ✓ Bottleneck Analysis
- ✓ Query Optimization
- ✓ Log Analysis

Data

- ✓ File Metadata (Size, Temperature, ...)
- ✓ Historical Analysis & Trending
- ✓ Schema Drift Monitoring
- ✓ Data Quality Monitoring
- ✓ Data Reconciliation
- ✓ Throughput Monitoring for Streaming Data
- ✓ Data Redundancy
- ✓ Data Discovery
- ✓ Pipeline & Lineage

Acceldata customers have measured significant improvements in their data operations by leveraging the data observability platform. Common results for customers include:

90%
reduction in mean-time-to-resolution

95%
reduction in severity 1 incidents

30%
improvement in capacity and throughput

300%
improvement in worker productivity

Achieving these results comes not just from the breadth of telemetry that's monitored, but the analytics and automation applied to the telemetry. This is what distinguishes observability from monitoring. Here are some examples:

Incident Prevention: "Predictive Maintenance for Data"

True reliability does not come from identifying and fixing issues. Reliability comes from avoiding incidents to begin with. Acceldata takes a Predictive Maintenance approach, common to manufacturing and other industries and applies it to data operations. Three key capabilities are involved:

Performance Trending Analysis: Acceldata automatically calculates a variance score and other metrics that measure trends in runtime, resource consumption, data volume and other telemetry. This allows customers to identify which areas are likely to exceed SLOs or fail in the future, even when everything shows green at present. Some customers have transitioned from weekly, reactive incident response to 12 months and counting without a single severity 1 incident.

Data Reliability: Poor data quality is the number one obstacle for organizations to generate actionable business insights, according to 42 percent of executives surveyed by the Harvard Business Review. To improve quality coverage, Acceldata scans data and automatically generates data quality rules to quickly and easily cover the majority of potential data quality issues. Acceldata also provides an easy way to cover a wide range of data quality concerns that are lacking in most data quality solutions in the market. Schema drift, data reconciliation, data drift and anomaly detection are capabilities that customers cite as being blind spots that Acceldata addresses well.

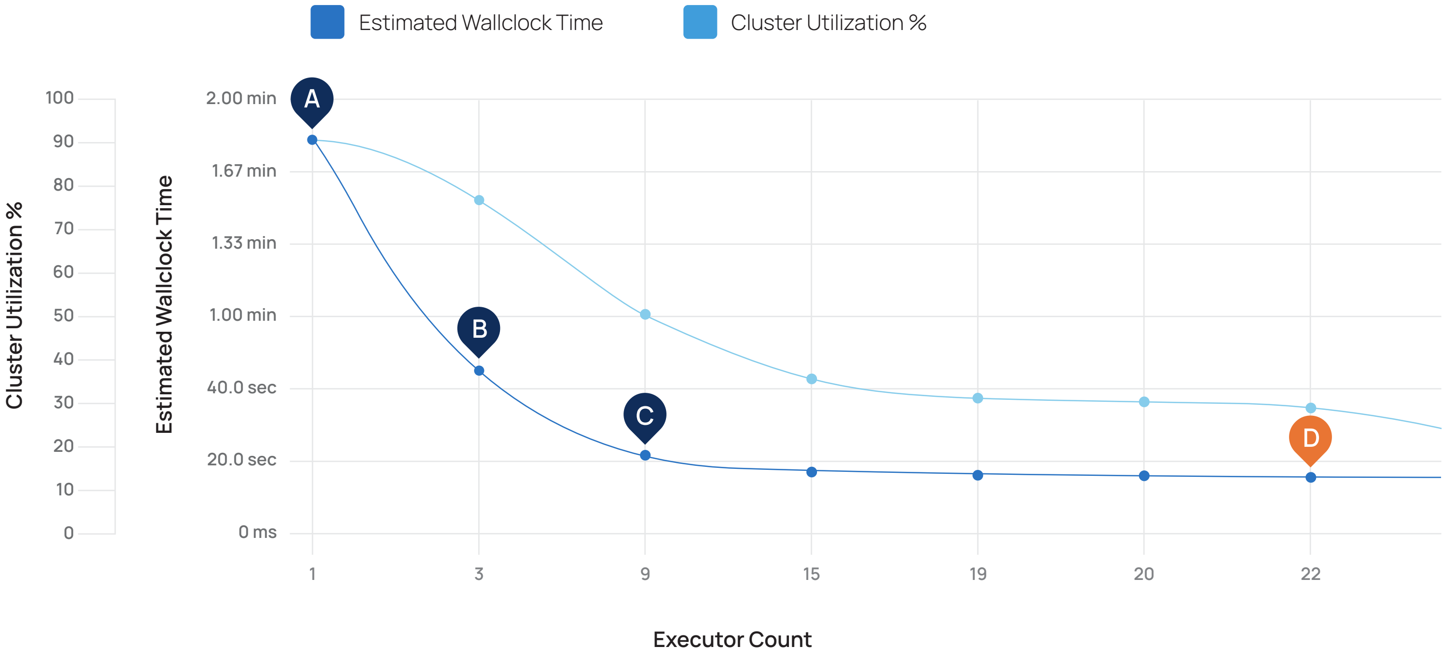
Auto-remediation: Acceldata provides an auto-action framework based on Ansible. The solution includes over 20 runbooks out-of-the-box with the ability to develop new custom runbooks. This not only eliminates manual effort but also provides near instantaneous self-tuning and self-healing. A rich set of APIs plus alerts, notifications and triggers allow flexible orchestration between Acceldata, the Hadoop platform, ticketing systems and other external systems and processes.

Performance Analytics

A chain is only as strong as its weakest link. Similarly, a single bottleneck will weaken the performance of a query, job or other workload. Workloads can be quite complex. It's one thing to monitor performance, it's another to identify the root cause and how to improve performance. Acceldata provides performance analytics across three broad categories:

Recommendations: Acceldata automatically analyzes workloads and provides recommendations for query optimization and job configuration.

Simulation: Acceldata simplifies the process of right-sizing job configuration to meet a specific SLO. For example, in the chart below, an engineer can see on a curve, the runtime for a Spark job with minimal executors **(A)**, the recommended executor count to get below a specific runtime **(B)**, the executor count recommended for high performance **(C)**, and where price/performance drops off with high resource consumption for small performance gains **(D)**. This takes out the guesswork and cumbersome trial and error to right-size job configurations.



Workload Analysis: Acceldata provides a rich suite of analytic tools to identify performance bottlenecks, correlate events, and optimize jobs, queries and configuration. For example, identify which stages within a Spark job are single-threaded, which parts of a Hive query perform large scans, or where overhead is high due to many sub-tasks. Event correlation can show where resources are constrained, what metrics have changed from one execution to another and how they relate to each other (e.g. data volume vs. runtime vs. memory, etc.).

Resource Efficiency

Unlike the cloud, adding capacity to on-premises infrastructure takes more than the click of a button. Purchasing excess capacity insures against unexpected surges and the need to go through procurement ahead of schedule to meet increased demand. Even with excess capacity, organizations often find themselves running out. Improving resource efficiency can not only help avoid capacity issues but it can also save a lot of money and enable new use cases to be onboarded with a greater return on investment. Here are three areas where Acceldata helps improve efficiency:

Capacity Optimization: Utilization analytics can help organizations optimize the scheduling of workloads to even out resource utilization over time. Chargeback reports align resource cost to business benefits to ensure the highest priority workloads are served. Short-lived workloads can be identified for execution in the cloud for the optimum hybrid cloud strategy.

Data Processing Optimization: Acceldata automatically profiles and flags jobs that could be run more efficiently. Data engineers can then drill down into those jobs and leverage the performance analytics tools to achieve the same or better performance with fewer resources.

Data Engineering Optimization: HDFS analytics identifies “cold data” that is infrequently updated or accessed for potential archiving. “Small files” reports identify opportunities to improve data processing efficiency. Hot spot visualizations identify how data can be reorganized to balance load across a cluster.

Acceldata provides a suite of tools to support each of the use cases above and many others. This has enabled customers to predict and prevent incidents, scale performance by orders of magnitude, and reduce their infrastructure costs 20%-50%. Acceldata’s data observability solution is the foundation by which Acceldata can provide extended support offerings for Hadoop with the best SLAs in the industry.

Compatibility with the Hadoop ecosystem

The Acceldata Data Observability platform supports the following Distributions:

- **HDP >= 2.6.x & 3.x**
- **CDH >= 5.10.x & 6.x**
- **CDP 7.x**
- **Apache Open Source**

Technology integrations include:

- **Kafka**
- **MapReduce**
- **Druid**
- **YARN**
- **NIFI**
- **HIVE**
- **HBase**
- **SPARK**
- **Impala**
- **HDFS**

Acceldata services may be available to advise and assist with migrations to supported distributions and technologies (see the Project-based support section on page 18).

Use Case: Migration Validation

Over time, all systems get updated and eventually replaced. Data Observability can significantly reduce the time, risk and cost associated with upgrades and migrations by automating many aspects of testing.

Data Reconciliation: Acceldata's platform makes it easy to reconcile data between two data stores. Both the structure and the data can be compared across diverse technologies. Data can be reconciled in complex scenarios where data has been integrated, aggregated or otherwise transformed.

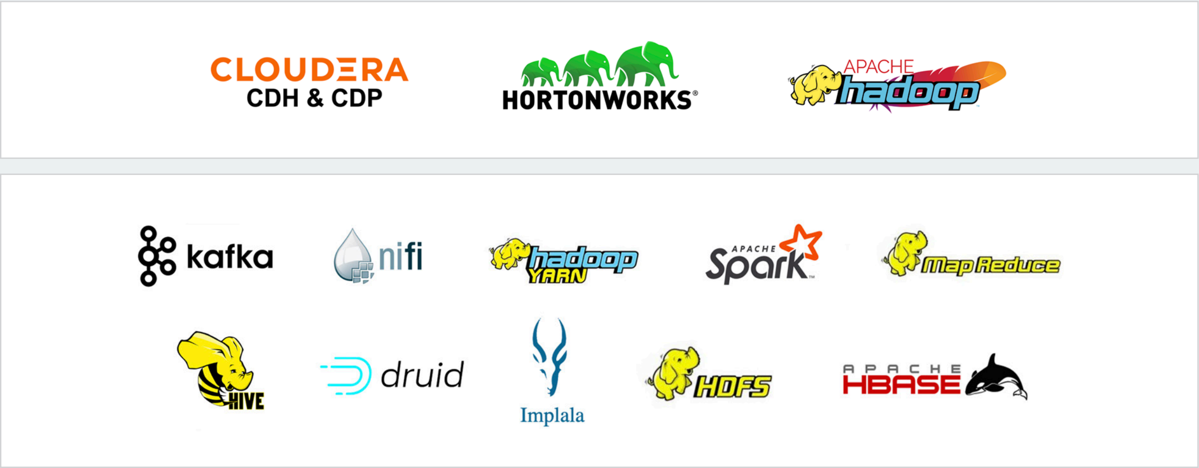
Performance Testing: Acceldata allows complex data pipelines to be compared and analyzed at high-level with the ability to drill down and compare the individual processing.

Resource Utilization: Acceldata allows detailed analysis of resource consumption between different executions in different environments. This assists with comparing and assessing cost estimates before and after migration.

These same capabilities serve post-migration to ensure performance, data quality and cost remain consistent with requirements and expectations. Data Observability's value throughout the lifecycle is one of the benefits of engaging Acceldata early, even at the strategy and design phase.

Deployment Architecture

The Acceldata platform is deployed alongside the Hadoop environment as a set of docker containers, connectors, repositories, and lightweight agents that work in concert to collect, store, and analyze telemetry data. The diagram below provides a high-level overview of the architecture.



Acceldata Pulse provides the most comprehensive observability platform for the Hadoop ecosystem

Acceldata Support

Acceldata provides several support packages to meet the needs of its customers:

1. **Data Observability Platform Support** provides support for Acceldata products
2. **Hadoop Support** provides support for the Hadoop platform
3. **Hadoop Administration** provides named site reliability engineering (SRE) resources on a part-time and full-time basis
4. **Project-based Support** provides assistance with upgrades, migrations and other one-time events on a per-project basis.

Note, all support packages require an active software license subscription for Acceldata's Data Observability platform. This provides Acceldata and customers the operational insight and tools necessary for continuous, trouble-free operations.

1. Data Observability Platform Support: "Enterprise Support Subscription"

This package provides support for Acceldata products and is included with software subscriptions. Support for Hadoop and other 3rd-party technologies are not included.

2. Hadoop Support: "Enterprise Plus Support Subscription"

This package extends the Enterprise Support Subscription to provide support for compatible Hadoop platforms and components. This package includes support from experienced subject matter experts, who are well versed in the operational aspects of running open source components on large scale deployments. The combination of Data Observability technology and Hadoop expertise enables Acceldata to offer superior service level agreements (SLAs). As of this writing, only Acceldata goes beyond initial response SLAs to provide SLAs for ongoing response times. Furthermore, only Acceldata goes beyond response SLAs to provide SLAs for the resolution of incidents.

Acceldata experts will work closely with your team to support the following activities:

- **Incident Resolution:** Support for production outages
- **Patches:** Acceldata provides patch creation and implementation of community driven and custom patches. These patches will include CVE fixes and security fixes.

Note, delivery of patches in a break-fix scenario are limited to HDP (2.6.5 - 3.1.4), Ambari 2.6.x, and Apache Open source distributions. Not included are patches for HDP 3.1.5, CDH, and CDP distributions.

Technical Support: Implementation of best practices for maintenance-related activities for the Hadoop platform

Architecture: Guidance for architectural design and prototyping new technologies
Upgrades: Support for upgrades in software versions

The following are not included in this support package:

- All data engineering/DevOps activities
- Technical support on any third-party components beyond those specified in the support package.
- New Feature requests for the Hadoop ecosystem.
- Patches that do not exist in the Open source community. See Patches above.
- Hadoop Cluster upgrades (see Project-based support below).

**** Break Fix Scenario for Hadoop ecosystem -** If a Customer encounters a Critical BUG on the Hadoop Ecosystem, for which a patch is required. The delivery of said Hadoop patches is under the sole discretion of the Acceldata product management team, as it depends on the complexity of the issue and the availability of patches in the open-source community.

**** Software version & Source Code for Hadoop ecosystem:** The client will provide the major, minor version and the corresponding open-source Git links for Hadoop distribution code in production for the clusters covered in support, prior to the commencement of support.

3. Hadoop Administration: “Enterprise Gold Support Subscription”

This package extends the Enterprise Plus Support Subscription to provide additional administrative services with named site reliability engineer(s) (SREs) on a part-time or full-time basis. Acceldata SREs become an extended part of the customer’s team, assisting with administrative tasks, sharing expertise, applying best practices, and tapping into the technology and expertise of the broader Acceldata team.

The Enterprise Gold subscription has the following features:

Scope of Work: Standard BAU Hadoop admin activities such as platform maintenance, service restarts, and cluster expansion.

Location: All Acceldata SREs are based remotely

Working Hours: 9x5 - as per customer preference on workdays. The named SRE(s) would be available as per customers’ standard working hours only. For any technical support outside of standard working hours, the customer can reach out to Acceldata pool-based technical support team.

Access: VPN/VDI access with appropriate privileges to the Hadoop ecosystem components is required.

Capacity: The number of named Acceldata SREs will be determined by Acceldata based on the evaluation of existing ecosystems and workloads. A higher head-count may be made available to the customer as a professional services engagement if desired (see Project-based Support below).

The following are not included in this support package:

- All data engineering activities.
- Customer-end Level 1 support (e.g. Provisioning of User, Permission & Cluster-wide Basic Admin Activity).
- Data and cluster upgrade and migration initiatives (see Project-based Support below)
- Any PoCs on the Hadoop ecosystem.

4. Project-based Support

Acceldata may be able to assist with project-based activities such as upgrades, downgrades, assessments, and others. Acceldata may also be able to assist with connecting with official Acceldata partners such as system integrators that can offer a wider range of services. Please consult with your Acceldata account executive to explore further.

Support Model

All subscription-based support packages include the following:

- **Training:** Initial purchase of the platform includes customer training to simplify and accelerate success with data observability.

- **Onboarding:** The Customer team will be provided access to the Acceldata Support Portal to create and track tickets.

- **Acceldata Product Support:** This includes all Acceldata upgrades, feature releases, hotfixes, and patches.

- **Technical Support and Guidance:** Provided through the Acceldata Support Portal.

- **Pool-based Support Model and Working Hours:** Skilled support engineers work around the clock, 24x7x365 to provide the highest degree of technical support for severity 1 incidents. Technical support working hours for severity 2 and below are 24x5 IST following the standard India calendar.

- **Easy Access:** No VPN/VDI access is required for Enterprise Support and Enterprise Plus Support. Virtual Assistance sessions will be conducted over Zoom, Webex, or MS Teams by the Acceldata Support Team.

Case Studies

The following are four case studies of organizations with large, mission-critical Hadoop environments that are leveraging Acceldata for improved reliability, performance, efficiency and support.

PubMatic

PubMatic is one of the largest AdTech companies in the United States. Using Acceldata, PubMatic manages 4,000+ nodes scattered over 60+ on-premises clusters. By running everything on HDP 3.1.0, a free open source Hadoop distribution, they do not require a Cloudera subscription, saving them millions in licensing costs. Furthermore, Acceldata helped PubMatic optimize performance, significantly reducing their infrastructure size and spend, saving millions more. Acceldata also helped PubMatic eliminate frequent outages and bottlenecks with analytics and automation to predict and prevent incidents and reduce mean time to resolution (MTTR) if incidents occur.

> *Acceldata "helped us optimize HDFS performance, consolidate Kafka clusters, and reduce cost per ad impression, which is one of our most critical performance metrics," says Ashwin Prakash, engineering leader at Pubmatic. "Acceldata's data observability saved us millions of dollars for software licenses that we no longer need. Now we can focus on scaling to meet the needs of rapidly growing business."*

truedigital

True Digital is a leading Southeast Asian telecom provider. With Acceldata, True Digital solved pervasive system performance and scalability issues in its 35 PB HDP-based data lake. Acceldata enabled True Digital to double data processing throughput while also reducing infrastructure by 25%, saving millions. Moreover, Acceldata transitioned them from frequent unplanned outages to eight months and counting without any unplanned outages or severity 1 incidents.

> *"Acceldata's tools fixed our analytics pipeline issues, improved visibility into our data systems and recommended ways to scale and optimize our systems to meet future requirements," according to Wanlapa Linlawan, True Digital's Analytics Head.*

PhonePe

PhonePE is the leading e-payment services company in India, supporting over 350 million consumers. Acceldata helped them achieve this position by smoothly scaling its on-premises Hadoop (HDP 2.6.5 and 3.1.4) clusters by over 2,000 percent — from 70 to 1,500+ nodes. The Walmart subsidiary also delivered 99.97% availability while freeing up its data engineers from daily troubleshooting. By leveraging free open source Hadoop they avoid $5+ million in annual software license costs.

> *"Acceldata supports our hyper-growth and helps us manage one of the world's largest instant payment systems. PhonePe's biggest-ever data infrastructure initiative would never have been possible without Acceldata." - Burzin Engineer, Founder & Chief Reliability Engineer*

ORACLE

Oracle, the enterprise software giant, now beats all of its performance and reliability SLAs for its 170+ HDP 2.6.5 nodes with the help of Acceldata. Hadoop/Hive queries now run twice as fast, while its engineering team is three times more productive.

Case Studies Takeaways

A key takeaway is that these companies didn't just choose Acceldata over Cloudera and other Hadoop commercial support providers to save millions on licensing and support costs. They are relying on Acceldata's Data Observability platform and team of experts to gain insight into their environment to predict and prevent incidents, improve resource efficiency and worker productivity.

Summary

Organizations that have invested in on-premises Hadoop may be facing a difficult set of choices due to the end of Cloudera support for HDP and CDH. Fortunately, Acceldata provides technology and support offerings that enable organizations to continue to receive the benefits of their legacy investments without having to "go it alone". Moreover, many Acceldata customers made the switch long ago and attest that the support, service, expertise, and technology they receive from Acceldata surpass their previous experience with Cloudera and at a much lower cost.

Go to https://www.acceldata.io/cloudera to learn more and to schedule a free assessment to help you identify the best path forward for your organization.

References

Data Observability for HDP/CDH Customers
https://www.acceldata.io/cloudera

PhonePe case study
https://global-uploads.webflow.com/60ddb7e2e50eaef5bec9595c/6297f39f82aede55c5faef19_PhonePe-Case-Study-53122.pdf

True Digital case study
https://global-uploads.webflow.com/60ddb7e2e50eaef5bec9595c/6297f3bd36e6731ca70ec902_True-Digital-Case-Study-53122.pdf

PubMatic case study
https://global-uploads.webflow.com/60ddb7e2e50eaef5bec9595c/6297f36abcb979cf69dd5790_PubMatic-Case-Study-53122.pdf

As Cloudera 6.3 Goes End of Life, How Can You Minimize Your Hadoop Risks?
https://www.acceldata.io/blog/as-cloudera-6-3-goes-end-of-life

Got Hortonworks or Cloudera? How to Avoid A Disastrous, Costly Forced Migration
https://www.acceldata.io/blog/got-hortonworks-or-cloudera-how-to-avoid-a-disastrous-costly-forced-migration

Cloudera Manager FAQ
https://docs.cloudera.com/documentation/enterprise/6/6.3/topics/cm_faqs.html

CDH Data Sheet
https://www.cloudera.com/content/dam/www/marketing/resources/datasheets/cloudera-enterprise-datasheet.pdf?daqp=true