CLOUDERA

EBOOK

Create Greater Business Value With Cloudera On Premises

Enable new business use cases with comprehensive security, performance, and accelerated analytics.

Table of contents

Introduction

Data can either be an anchor or a propeller. Some enterprises are at an infrastructure crossroads, facing the question of how to manage the constant influx of high volumes of data into their systems.

The reality is, data is growing exponentially with no signs of slowing down. This means you're faced with an onslaught of challenges around extracting insights, moving agilely, and preventing breaches and outages.

While legacy systems may feel stable now, they are a heavy anchor to the kind of growth and modernization your business needs to stay competitive and agile.

If you're on Cloudera Distributed Hadoop or Hortonworks Data Platform, Cloudera on premises is not the next version — it's the accelerator.

Cloudera on premises connects data performance to business objectives, propelling enterprises into peak performance and new possibilities without sacrificing stability, speed, or security.

Bottomline, the best thing to do with data is to use it. So let's focus on two things: A look at use cases powered by Cloudera on premises and migration strategies to get the most out of your data.

Want to achieve business critical use cases and unlock new ones on an end-to-end platform?

Cloudera is the answer.

Benefit from world-class customer support?

Cloudera is the answer.

Require improved security and compliance measures?

Cloudera is the answer.

Need to increase efficiency and reduce TOC?

Cloudera is the answer.

What's your first step?

Cloudera on premises is your first step.

Balancing the risks and rewards of migration

We know that migration can be both challenging and risky. There are financial investments tied to migration, the possibility of taking systems offline, and the risk of things not going according to plan.

Complications still exist even when broken architecture isn't the problem. The fact is that even if it isn't structurally damaged, it's a weakness. Legacy IT architectures complicate efforts to deploy technologies that improve decisionmaking, vital business capabilities, and customer experience. Less than optimal performance, speed, and security at the IT level is damaging at the business level.

At the structural level, migration to Cloudera on premises means having a system designed for continual updates that is capable of adapting to data volume and storage scale. In terms of data, migration to Cloudera on premises means more robust security, better and faster performance, real-time insights, costsavings, and the option to adopt a hybrid model to expand to the cloud.

Having the right strategy is essential when it comes down to mitigating the risks of migrating. The right approach and a partner to fill knowledge gaps are even better.

On the next page, we outline the most common migration strategies as you plan to make your move to Cloudera on premises .

Cloudera Support makes the difference. Here's a summary on one case, of many, highlighting a serious incident.

Log4J vulnerability impact:

- Out of 1,800+ log4j cases, 60% were on legacy platforms (CDH and HDP)
- 60 hotfixes released for Cloudera
- Remediation script issued to supported customers within 72 hours

Cloudera on premises by the numbers

2X

Faster analytics to accelerate time to value

Significant increases achieved from integrating the latest versions of Impala, Hive & Spark

20%

Higher resource utilization optimizes costs

Enhanced cluster management and resource scheduling reduces overheads and boosts utilization

50%

Improved density for more cost effective storage

Increased storage density from HDFS erasure coding and augmented with Ozone object storage

2X

Better Infosec compliance helps reduce risk

Dramatic CVE reduction and better encryption of user data through FIPS compliant modules

50%

Stronger governance to ease regulatory compliance

New Cloudera Shared Data Experience (SDX) controls - deny by default, least privilege, policy tags, scalable audit, consistent enforcement

Migration strategies for Cloudera on premises

1 In-place Upgrade

On the same hardware ("in-place"), directly upgrade your existing legacy CDH or HDP clusters to Cloudera. This process is best suited for larger clusters with more significant data footprints.

Choose this strategy if you:

- Want to take the quickest path to Cloudera as possible
- Need to avoid large capital expenditure
- Have many interconnected workloads that would make tenant-by-tenant migration complex

2 Legacy Migration

This method aims to minimize downtime on individual workloads while providing a straightforward rollback mechanism on a per-workload basis. It also allows you to migrate your CDH or HDP deployment to Cloudera on new hardware.

Choose this strategy if you:

- Have tighter service-level agreements that make multi-hour workload downtime impossible
- Want a full hardware refresh

According to Deloitte's survey of IT executives, 47% report that the "cloud is more complex than they expected."[1]

Migration is already a complicated process, so carefully consider the perceived "value" of do-it-yourself (DIY) platform support. The risks of building, maintaining and supporting an enterprise-grade platform — along with handling security and remediation — often dwarf what you anticipate saving in time, resources, and cost.

The good news is that we've done this enough to know what migration success looks like and how to achieve it.

[1] Ramachandran, K., & Linthicum, D. (2020, March 5). Why organizations are moving to the cloud. Deloitte Insights. Retrieved June 7, 2022, from https://www2.deloitte.com/us/en/insights/industry/technology/whyorganizations- are-moving-to-the-cloud.html

3 Rolling Side-car Migration

A Rolling Side-car Migration is a modified Side-car Migration. Some of the existing cluster hosts are decommissioned, Cloudera is installed on the decommissioned hosts, and workloads are migrated one at a time. As workloads are migrated, you again decommission more hosts and repeat the process.

This is a good alternative for a customer who needs to shift capacity from legacy systems to a new deployment

Choose this strategy if you:

- Have to shift capacity
- Want less downtime than an In-place Upgrade
- Need a more cost-conscious option than Side-car Migration

2 Migration to Cloudera on cloud

Similar to the Side-car Migration path, this method enables you to build the new Cloudera environment alongside the legacy environment and replicate data using cloud native object storage. You are also able to scale the compute and storage independently

Choose this strategy if you:

- Have an on-premises environment near end-of-life
- Want a more flexible infrastructure model
- Need more adaptability and control in resource allocation

Moving to Cloudera on premises gives you everything you need: the right platform, plan, and people.

Organizations that choose to remain on unsupported legacy CDH or HDP platforms are relinquishing their access to enterprise support from Cloudera. This doesn't mean the platforms stop working. However, the organizations themselves are solely responsible for providing support, which can open the door to risks that affect your critical data workloads. Migrating to Cloudera on premises is the answer.

Use cases that power business

Build an architecture that uses data to power smarter business outcomes.

Keeping up with the competition and growth are foundational business objectives. Data — how it's organized, used, stored — is the key to developing use cases that will spur growth and help develop a competitive edge.

The following pages uncover Cloudera on premises use cases across the data lifecycle:

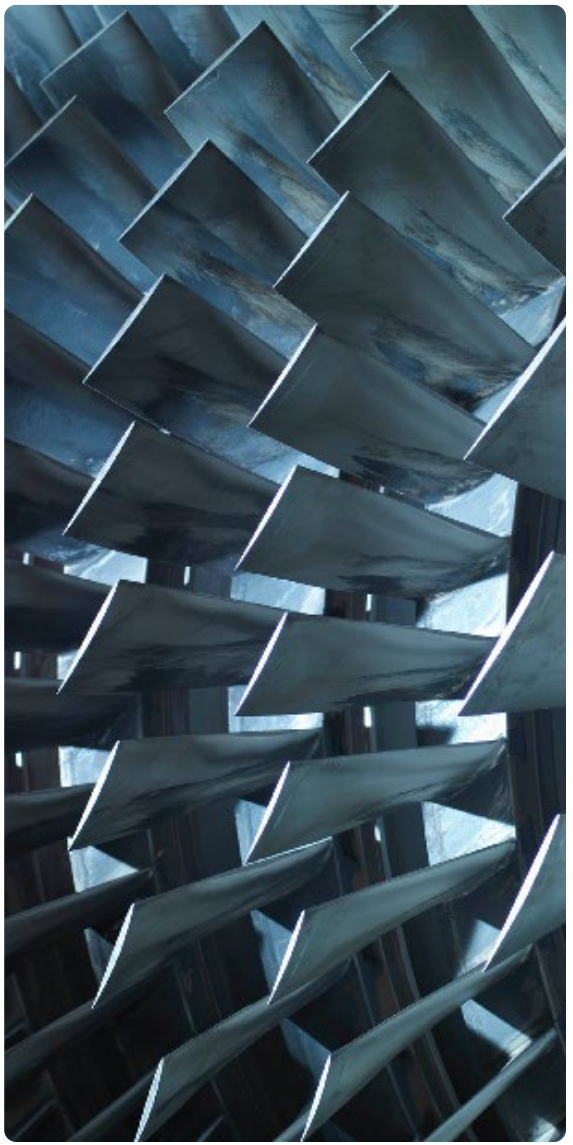
01
Streaming analytics

02
Data engineering

03
Data warehouse

04
Operational database

05
Machine learning and artificial intelligence



USE CASE 1 - STREAMING PROCCESSING

Power business outcomes with streaming analytics

Critical decisions require the right data, however the right data isn't always available at the right time. When data is warehoused instead of instantly processed, decisions are delayed, and the cost of delayed decisions can be high.

Cloudera Streaming Processing (CSP) is a real-time analytics solution that helps you respond to the critical events that drive business outcomes. Powered by Apache Flink and Apache Kafka, CSP processes high volumes of streaming data at scale, enabling you to analyze streaming data for complex patterns that lead to insight-driven action.

Capabilities of CSA:

- Low latency
- Actionable intelligence in real-time
- Unified security, governance, and management
- Manage millions of data points from multiple sources

Data is important, but it's not usable if it's not accessible or available in real-time. Today's businesses need capabilities such as fraud detection, click-stream analysis, network monitoring for telecommunications, and the application monitoring that CSP provides.

Opportunities:

- Execute sophisticated analytics to enable actionable insights in real-time
- Process streaming and batch data with unified SQL
- Streamline data processing to remove wasteful, error-prone steps
- Join, clean, and transform data in motion
- Democratize the analysis of streaming data with SQL
- Integrate with enterprise lineage, security, and governance standards

Featured Use Case

Real-time fraud detection

Finding a way to make banking more secure in a country that ranks among the lowest for financial literacy is no small task. For PT Bank Rakyat Indonesia (BRI), that meant developing a real-time fraud detection service to address the country's rising concern over the security of financial data. The end result: Speed to insight, enabling BRI to cut fraud by 40% — a record low compared to competing banks.

Read the full story

Why is data streaming so valuable?

The biggest benefit of data streaming — or the continuous flow of data from a multitude of sources — is the real-time insight it provides. Applied to all aspects of a business, data streaming enables an organization to read and react to events in the moment as well as apply that insight to long-term strategy. How could data streaming help you improve efficiency, better address customer needs and mitigate business risk?

USE CASE 2 - DATA ENGINEERING

Accelerate data engineering with Cloudera and NVIDIA GPUs

Rapid data growth causes bottlenecks and stressed resources. Data engineering teams and workflows need speed to deliver data pipelines or use cases and meet service level agreements and objectives (SLAs and SLOs).

The Cloudera and NVIDIA GPU integration pushes the performance boundaries and powers more use cases faster with seamless integration and reduced data engineering costs.

Capabilities of GPU accelerated data engineering:

- 5×+ performance at 50% the cost of an equivalent CPU-based system
- Total Cost of Ownership (TCO) increases by 3× with NVIDIA GPUs on Cloudera

As data analytics demands increase, so does cost of ownership. Cloudera powered by NVIDIA GPUs helps relieve pressure on budgets and resources, enabling you to take advantage of use cases that will positively impact your business.

Opportunities:

- Accelerate end-to-end data preparation and machine learning (ML)
- Generate models that produce highly accurate data and insights
- Expand Artificial Intelligence (AI) use cases with a complete production ML toolkit
- Operate a fully secure ML environment that can meet evolving requirements
- Reduces ML training time and the frequency of model deployment from days to minutes

Featured Use Case

Accelerate data preparation for ETL workloads

If you needed to comb a 3+ terabyte dataset at the IRS for patterns that might help uncover fraud, how long do you think it would take? With a large bank of CPU servers, it would certainly take all day or night — at a minimum. However, this story doesn't end that way. After implementing Apache Spark 3.0, accelerated by NVIDIA GPUs on Cloudera, this team saw a 5× increase in speed.

Read the full story

Why is acceleration important?

The proliferation of data today can cause bottlenecks and spread resources thin. The ability to accelerate data preparation for ETL workloads will enable you to maximize the value of your data. After all, what is data without analysis? How effective is analysis without the speed and accuracy?

USE CASE 3 - DATA WAREHOUSE

Storage for fast analytics on fast data

When you need to act on data quickly rather than analyze day-old data, you need the Kudu integration on Cloudera on premises.

Kudu is a storage system that performs analytics at the edge and stores data for analysis using historical data at the same time. Designed and implemented from the ground up, Kudu fills this gap between high-throughput, sequential-access storage systems (HDFS) and lowlatency, random-access systems (HBase).

Capabilities of real-time analytics with Kudu on Cloudera on premises:

- Fast analytics on fast data
- Simplified architecture
- Ecosystem integration

So, where does Kudu fit exactly? Data can either be landed quickly but not analyzed in HDFS or served really fast once landed with HBase. Kudu lets you do both.

Opportunities:

- Fast, real-time fraud detection and prevention
- IoT and time series data
- Machine data analytics (network security, health, etc.)
- Online reporting
- Create an updateable, open-source data warehouse

Featured Use Case

Fast, real-time fraud detection and protections

With over 1,300 bank locations and approximately 2,000 ATMs, fraud protection and prevention is critical for Regions Bank. Transaction data includes billions of records, so processing this volume for fraud models is a significant challenge. By modernizing its data environment, Regions now utilizes ML models to lead to fraud prevention and detection.

Read the full story

Why is fraud prevention and detection valuable?

Financial institutions can shift the dynamic in fraud detection from reactive to proactive. In the past, fraudulent transactions were detected after-the-fact and then remediated. Now, there is the ability to detect suspicious activity as it occurs and block potentially fraudulent activity or transactions from even occurring.

USE CASE 4 - OPERATIONAL DATABASE

Empower developers to scale, build, and automate faster

Cloudera's operational database (OpDB) in Cloudera serves traditional structured data alongside new unstructured data within a unified open-source platform.

Part of Cloudera on premises's OpDB capability is the enablement of transactional applications. Enterprises use applications that leverage transactions for many reasons but run into issues when the traditional relational database management system (RDBMS) can't scale.

Transaction support combined with ANSI SQL support and the scalable nature of Apache HBase is a combination now provided by OpDB to help reduce the operational complexity of transactions.

Capabilities of OpDB:

- Real-time
- Always available
- Secure and scalable
- Lower database TCO

Opportunities:

- Faster development and deployment of mission-critical applications
- Build enterprise-class applications faster
- Automatically improve database performance based on application requirements
- Future-proof applications
- Enable transactional applications

Featured Use Case

Predictive maintenance

In the transportation industry, if a truck isn't moving, no one is making money. As a leading manufacturer of commercial trucks, buses, defense vehicles, and engines, Navistar needed to help fleet and vehicle owners move from a reactive approach to a more predictive model for vehicle maintenance and be able to analyze a wider range of data in real-time from vehicle sensors. The solution was an IoTenabled remote diagnostics platform, called OnCommand® Connection, on Cloudera Enterprise with SDX.

Read the full story

Why is predictive maintenance valuable?

Predictive maintenance benefits any industry relying on infrastructure that requires regular maintenance. Being able to predict maintenance or anticipate issues reduces maintenance costs, saves on infrastructure costs, and helps to eliminate downtime or financial loss from inoperable infrastructures.

USE CASE 5 - MACHINE LEARNING AND AI

Exploratory data science with ML and AI

There is a gap between getting data results and being data-driven.

In a recent survey by NewVantage Partners, 92.1% of organizations said they achieved measurable results from their data and AI investments. But only 27% said they'd created a data-driven organization.[2]

Disconnected data initiatives may reveal value, but the first steps in actualizing a data-driven enterprise are exploratory data science and equipping your teams with the right tools.

Capabilities of Cloudera AI, Cloudera Data Engineering, and Cloudera Data Warehouse for exploratory data science:

- Unified workflows
- Trusted sources of data
- Flexibility in choice of tools
- Containerized environments
- Data pipeline tools and environments
- Enterprise orchestration tools

Data exploration is far more effective and beneficial from a single platform over a patchwork of tools. Unifying capabilities like machine learning, data engineering, and data warehousing helps data science practitioners work efficiently, deliver results quickly, and drive business outcomes

Opportunities:

- Turn data into action with data visualization
- Solve problems in the early stages of the data science process
- Uncover new opportunities to leverage the data you have

Featured Use Case

Increase accuracy and accelerate discovery

As a global provider of advanced analytics, technology solutions, and contract research services to the life sciences industry, IQVIA challenges the status quo using data and analytics. In order to reduce healthcare costs, improve access to care, and deliver new treatments, IQVIA needed to find a way to explore and iterate on data much more quickly. By creating a modern data platform with Cloudera, IQVIA can now generate insights from data for its customers in seconds, rather than in days, weeks, and months. This massive boost in performance enables life sciences companies to innovate faster and save lives by speeding up new drug development lifecycles.

Read the full story

Why is speed and accuracy valuable for exploratory data science?

Many organizations that are on the cutting edge of innovation are also havens for enormous quantities of data. To really maximize the impact of data on their new solutions, it must be easily accessible, fast, and accurate.

[2] NewVantage Partners, "2022 Big Data and AI Executive Survey"

Make the move and migrate to Cloudera on premises

Across industries, organizations must deal with a dual reality that requires them to strike a difficult balance: On one hand, keep up with the volume of data that continues growing at a rapid rate. On the other, ensure the enterprise moves at the accelerated pace necessary to meet demands. Relying on legacy systems like CDH and HDP to meet both of those objectives — in the name of maintaining a competitive edge and improving customer experiences — places a substantial burden on the business.

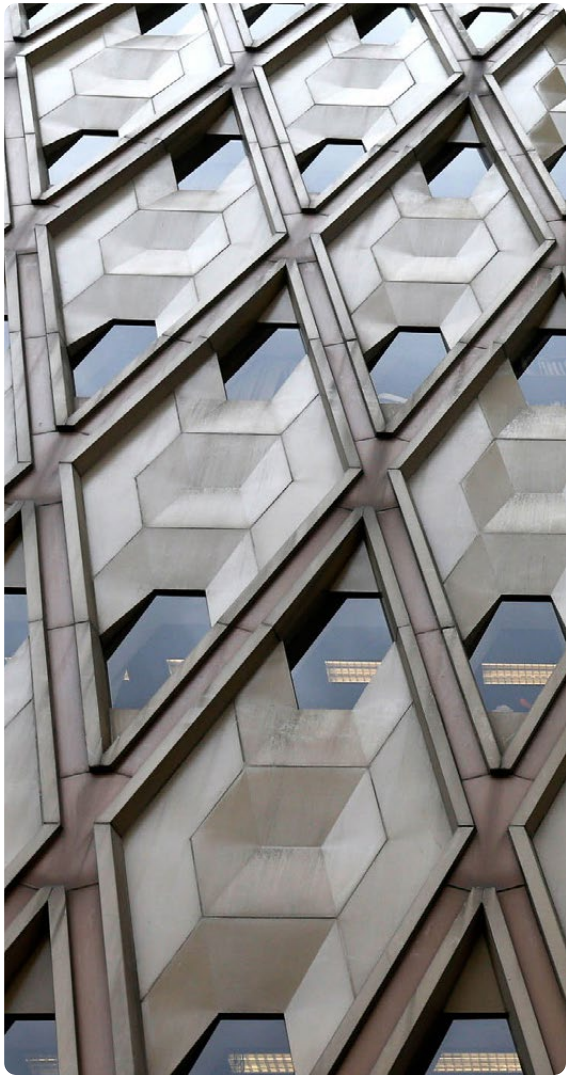
Furthermore, CDH and HDP are now end-of-life, requiring organizations to act with a sense of urgency when it comes to migration.

That's why we've made modernizing to Cloudera on premises quick and easy.

With Cloudera on premises and its world-class customer support, your data platform team can ensure the organization turns insights into action faster, operates in an environment with reduced risk, and ultimately derives greater value from its investment.

Cloudera on premises provides stability, compliance, agility, and innovation — all in one platform.

Ready to take the first step and migrate to Cloudera on premises? Let's get started.



About Cloudera

Cloudera is the only true hybrid platform for data, analytics, and AI. With 100× more data under management than other cloud-only vendors, Cloudera empowers global enterprises to transform data of all types, on any public or private cloud, into valuable, trusted insights. Our open data lakehouse delivers scalable and secure data management with portable cloud-native analytics, enabling customers to bring GenAI models to their data while maintaining privacy and ensuring responsible, reliable AI deployments. The world's largest brands in financial services, insurance, media, manufacturing, and government rely on Cloudera to be able to use their data to solve the impossible — today and in the future.

To learn more, visit **Cloudera.com** and follow us on **LinkedIn** and **X**. Cloudera and associated marks are trademarks or registered trademarks of Cloudera, Inc. All other company and product names may be trademarks of their respective owners.

CLOUDERA

Cloudera, Inc. | 5470 Great America Pkwy, Santa Clara, CA 95054 USA | cloudera.com