CLOUDERA

Three Scenarios only True Hybrid Unlocks

Table of Contents

CLOUDERA

Introduction: The Future is Hybrid

Cloud infrastructure revolutionized the way IT teams support their businesses, from application development and delivery to enterprise security and governance. In particular, it transformed the way companies manage and leverage their data.

The cloud offers many benefits over traditional data infrastructure, including:

- **Separation of compute and storage:** Independent scaling of infrastructure resources means that teams can store data once and provide concurrent access for a variety of workloads in isolation.

- **Managed infrastructure:** Data teams don't have to manage their own infrastructure or refresh their hardware. Additionally, scaling resources up and down takes minutes instead of days. Shared infrastructure often results in lower costs due to the economies of scale.

- **A broad ecosystem:** Data teams have access to many different tools and technologies, and they can rapidly test and adopt new innovations.

However, many organizations have learned, often through trial and error, that not every workload belongs in the public cloud, and not every workload belongs in the same cloud. The reasons vary, from the need to keep data within a certain region or geography for privacy purposes to workloads with zero tolerance for network latency or connectivity issues.

In fact, the vast majority of companies are pursuing a hybrid IT strategy – a mixture of multiple public cloud and on premises deployments, with critical sources of data often residing in distributed data stores.

While the hybrid strategy moving forward should be intentional, the existing architecture is often not. Today, many data teams are managing a complex assortment of legacy and cloud data warehouses and on premises and cloud data lakes. As a result, they spend most of their time responding to support tickets for data access requests or performance issues.

There is a better hybrid solution. Cloudera is a cloud-native platform for data, analytics, and AI that provides a consistent experience across cloud and on premises environments, delivers broad, unified access to data for every analytic workload, and centralizes security and governance.

In this ebook, we will discuss the critical components of a true hybrid platform for data, analytics, and AI, discuss some of popular supporting use cases, and share how companies have transformed their data management and analytics strategies with a hybrid solution.

> Despite the popularity of the cloud, 73% of organizations take a hybrid approach according to Flexera's 2024 State of the Cloud report.[1]

---

[1] Flexera, 2024, *Flexera 2024 State of the Cloud Report*, https://info.flexera.com/CM-REPORT-State-of-the-Cloud

CLOUDERA
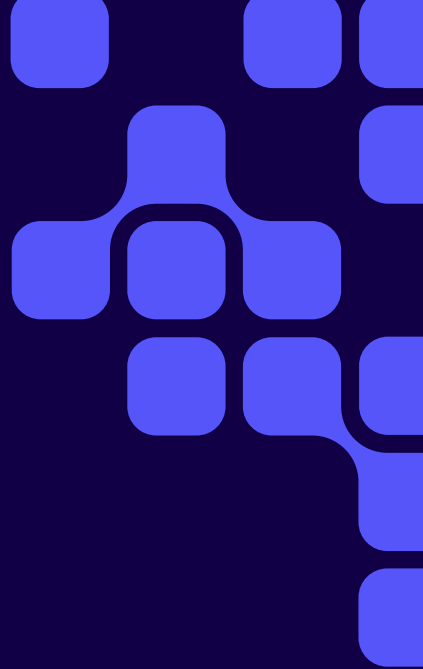
Requirements for a True Hybrid Platform for Data, Analytics, and AI

Cloud-Native Portability

Most companies arrived in their current state of hybridity as a result of a series of architectural decisions where the next Big Data platform provided some benefits, but never fully replaced the previous one.

Data lakes solved a critical storage problem, but they lacked the data management and analytics capabilities of the data warehouse. Cloud data warehouses solved some of the scalability challenges of traditional data warehouses, but they require significant data movement and data copies to meet performance Service-Level Agreements (SLAs), and they are typically only available in the public cloud. Cost and complexity can grow significantly over time.

Data teams are then managing a highly distributed collection of data warehouses and data lakes, with no easy way to provide unified access to all of their data.

A true hybrid platform for data, analytics, and AI provides several advantages over this unintentional architecture:

A consistent platform: Data teams should have a consistent platform with a common set of data services available across on premises and public cloud environments, reducing management complexity and ensuring unified security, governance, and data access.

Cloud benefits on premises: Separation of compute and storage transformed data management and data analytics, providing a single source of truth for all data consumers and independent, dynamic scaling to meet resource needs. A true hybrid platform should deliver separation of compute and storage in any environment.

Open formats: One of the challenges of a distributed, multi-platform approach to data management is that many solutions rely on proprietary formats, resulting in data copies and data movement, and vendor lock-in. Open formats ensure easy access to data for a wide range of use cases with the freedom to choose the best tool for the job.

By providing a consistent platform based on open data standards, formats, and technologies that delivers the benefits of cloud wherever the data resides, data teams can transition to an intentional hybrid strategy where they are free to choose the best infrastructure for each workload.

By achieving cloud-native portability, companies can quickly and easily change the underlying infrastructure as business, market, and technology needs dictate.

Fact-Check: Multi-cloud vs. True Hybrid

While many vendors use "hybrid" and "multi-cloud" interchangeably, enterprise organizations often have very compelling business, operational, and regulatory requirements for data to remain on premises.

A true hybrid solution provides a common platform for data, analytics, and AI that can be deployed in multiple clouds, including private clouds on premises.

CLOUDERA

Full Data Lifecycle Support

The purpose of the data lifecycle is to turn data into insights and value. A true hybrid platform should offer comprehensive tools to manage this lifecycle seamlessly, ensuring secure, governed access to high-quality data products at every stage.

01 Creation and Ingestion

Data sources and Service-Level Agreements (SLAs) for insights have changed significantly, and while batch ingestion is still common, streaming data is crucial for transformative use cases such as cybersecurity, remote monitoring, predictive maintenance, proactive service assurance, and application observability. A hybrid platform should support both batch and streaming ingestion, enabling data teams to handle diverse data sources efficiently.

02 Processing and Transformation

Raw data is rarely ready for immediate use in analytics, machine learning, or AI applications. Transformations are necessary to optimize storage and performance. Standardizing on an open table format like Apache Iceberg can enhance performance and compression while ensuring compatibility with multiple execution engines such as Spark, Flink, and Trino.

03 Analysis and Utilization

High-quality data must be accessible for various analytic purposes, from Business Intelligence (BI) and reporting to AI and machine learning. According to a recent McKinsey survey, 72% of organizations have deployed generative AI in at least one business unit.[2] A hybrid platform should provide the necessary tools for training, tuning, and deploying AI models, supporting advanced analytics across the organization.

04 Archiving and Purging

As data reaches the end of its useful life, it must be archived or purged in compliance with regulatory requirements and organizational policies. A true hybrid platform should facilitate the secure, governed archiving and deletion of data, ensuring a clean and efficient data environment.

By addressing each stage of the data lifecycle, a true hybrid platform empowers organizations to derive maximum value from their data, from creation to deletion.

[2] Richelle Deveau, Sonia Joseph Griffin, Steve Reis, McKinsey & Company, May 2023, *AI-powered marketing and sales reach new heights with generative AI*, https://www.mckinsey.com/~/media/ mckinsey/ business%20functions/marketing%20and%20sales/ our%20insights/ai%20powered%20marketing%20 and%20sales%20reach%20new%20heights%20 with%20generative%20ai/ai-poweredmarketing-and-sales-reach-new-heights-with-generative-ai.pdf

CLOUDERA

Cloudera On Cloud

Management Console | Data Catalog | Replication Manager | Workload Manager

Data Services

DF | DE | DW | OD | ML

Data Services

DE | DW | ML

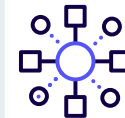**SDX** Metadata | Security | Encryption | Control | Governance

Data Clusters
NiFi*, Kafka, Flink, Spark, Impala, Hive, HBase, Phoenix

Storage
S3 | ADLS | GCS

Cloudera On Cloud

Data Clusters
NiFi*, Kafka, Flink, Spark, Impala, Hive, HBase, Phoenix

Storage
HDFS | Ozone

Cloudera On Premises

*Flow Management rate card

*Standalone SKU

CLOUDERA

Unified Security and Governance

Security and governance are critical components of every data platform. According to the Ponemon Institute, the global average cost of a data breach in 2023 was $4.45 million USD. Data teams must ensure end-to-end security of all of their data while providing broad access to data consumers for a wide range of analytic workloads.

Governance is becoming increasingly important as more organizations find themselves managing distributed data environments with important sources of customer and operational data residing on premises and in multiple clouds. Effective metadata management is crucial for ensuring that data consumers have access to a consistent and accurate view of the data.

For hybrid environments, consider the following capabilities for a unified security and governance solution:

Write Once, Apply Everywhere

Writing, synchronizing, and administering multiple access policies across on premises and cloud deployments adds significant management overhead to distributed data environments. Cloudera Shared Data Experience enables data teams to write access control rules once and apply them across deployments, simplifying administration of access controls and ensure consistency across environments.

Provide End-to-End Visibility into Data Pipelines

Track data from ingestion to presentation with a set of governance tools that provide full lineage, auditability, and discoverability of data assets wherever they reside. A unified governance posture ensures that data consumers trust their data.

Streamline Data Movement

For a variety of reasons, data teams must often move between environments. For example, data often originates in a source and needs to be moved elsewhere for organizations to leverage it. Using a tool like Replication Manager in SDX, data can be replicated across environments, and security and governance policies move with the data. This capability ensures a consistent experience whether data teams are migrating data assets, setting up a development and test environment, or executing backup or disaster recovery plan.

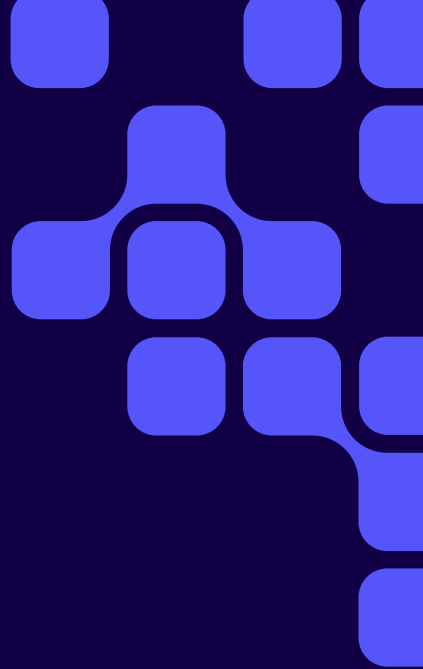The average cost of a data breach in 2023 was $4.45 million USD according to the Ponemon Institute.[3]

Security & Governance

SDX

Metadata | Schema | Migration
Security | Governance

---

[3] IBM, *Cost of a Data Breach Report 2024,* https://www.ibm.com/reports/data-breach

CLOUDERA

Hybrid
Scenarios

Accelerate Cloud Migration

Organizations move to the cloud for a variety of reasons, including increased agility and flexibility, high availability, and an ecosystem of Software-as-a-Service and Platform-as-a-Service technologies.

When choosing to move to the cloud, however, IT organizations must choose the most efficient migration strategy. Broadly, these strategies fit into two categories:

"Lift-and-shift" — This approach is often perceived as the easiest. Organizations simply re-deploy their on premises applications on public cloud infrastructure. However, many organizations encounter operational differences between on premises and cloud deployments, skills gaps, and inevitable migration delays — all of which increase the complexity associated with this strategy.

Replace with SaaS or PaaS — Many organizations choose to move to a SaaS or PaaS platform in the cloud. However, this approach has drawbacks. Cloud data platforms rely on proprietary formats and data copies to achieve performance on large volumes of data. Additionally, organizations often sacrifice control for ease of use, and this trade-off can lead to higher costs for compute and storage.

Both of these traditional approaches are often expensive, they take much longer than expected, they involve some risk of data loss or broken pipelines, and they are usually one-way — meaning future migrations are just as difficult.

A true hybrid platform enables a true lift-and-shift by providing a like-for-like data architecture on premises and in the cloud. This architecture enables data teams to modernize before they migrate, easing the burden of moving analytic workloads to the cloud.

With this approach, companies considering a migration to the cloud can experience several advantages, including:

01 Enabling data teams to adopt cloud-native services like object storage and open table formats before they migrate.

02 Building organizational confidence by giving data teams the opportunity to develop the skill sets necessary to manage data efficiently in the cloud.

03 Ensuring business continuity by providing access to a unified view of the data before, during, and after migration, with no disruptions.

04 Eliminating the need to rewrite and refactor workloads in the future when they need to move between cloud and on premises environments.

Ultimately, a hybrid approach ensures that data teams are well prepared for a successful cloud migration.
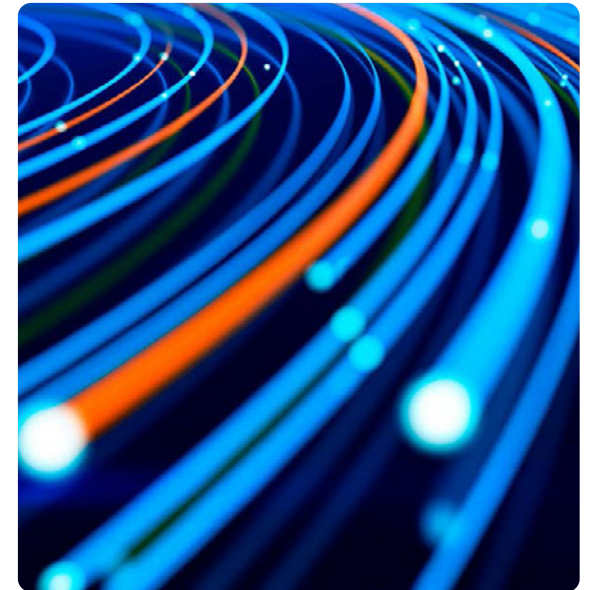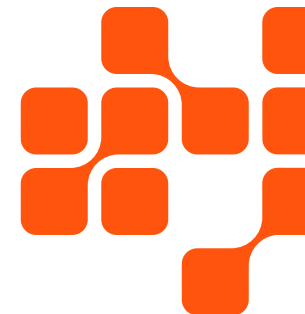
CLOUDERA

An IT Service Provider Accelerates Cloud Migrations with Cloudera

Ukrainian IT Services company's customers faced an increasing need for operational resilience during the war. Cloud migration quickly became a business imperative.

The services company was tasked with quickly migrating workloads while helping customers navigate data volume and regulatory concerns.

The company chose to build a modern data platform, leveraging Shared Data Experience (SDX) for unified security and governance, and for data replication. By powering their data platform with Cloudera, the company successfully reduced the time to migrate a customer to the cloud by 50%. Moreover, by moving to the cloud, they managed to reduce storage costs by 30%.

50%

50% reduction in time to securely migrate data.

"

The [data platform], powered by Cloudera, is redefining business resilience in Ukraine. It ensures data security, compliance, and cost efficiency at a time when many companies need it."

CLOUDERA

Adaptive Use Case Deployment

One of the benefits of true hybrid deployments is that workloads can be deployed using the optimal resource profile for that particular workload. Many use cases, however, have different optimal resource profiles for different parts of the workflow.

Consider the machine learning (ML) workflow depicted on the right. At a high level, training in ML consists of data preparation, model development and training, and deployment. The data may live on premises, where teams can accurately predict the resource require-ments. However, model development and training is an intermittent workload and better suited for the cloud, where data teams can spin up resources when needed and spin them down afterward to save on costs.

Using a true hybrid platform, customers can deploy parts of each workload using the optimal resource profile, while still achieving consistency through unified security and governance.

This deployment model delivers several advantages:

Cost Efficiency

By leveraging the optimal resource profile for each part of a workflow, data teams can significantly reduce costs. Intermittent and resource-intensive tasks can move to the cloud, enabling on-demand scaling and avoiding the need to maintain expensive, idle infrastructure on premises. This flexibility ensures that organizations only pay for the resources they need when they need them.

Performance Optimization

This capability enables customers to harness the best infrastructure for each stage of their workflow. On premises resources can be used for predictable, data-intensive tasks, while the cloud can handle computationally demanding processes. This approach maxi-mizes performance and ensures that each part of the workflow operates at peak efficiency, resulting in faster and more reliable outcomes.

Flexibility & Agility

Data teams gain the ability to dynamically allocate resources based on current needs as well as changing demands. This flexibility allows them to quickly adapt to new require-ments or unexpected changes in a workload, ensuring continuous optimal deployment.
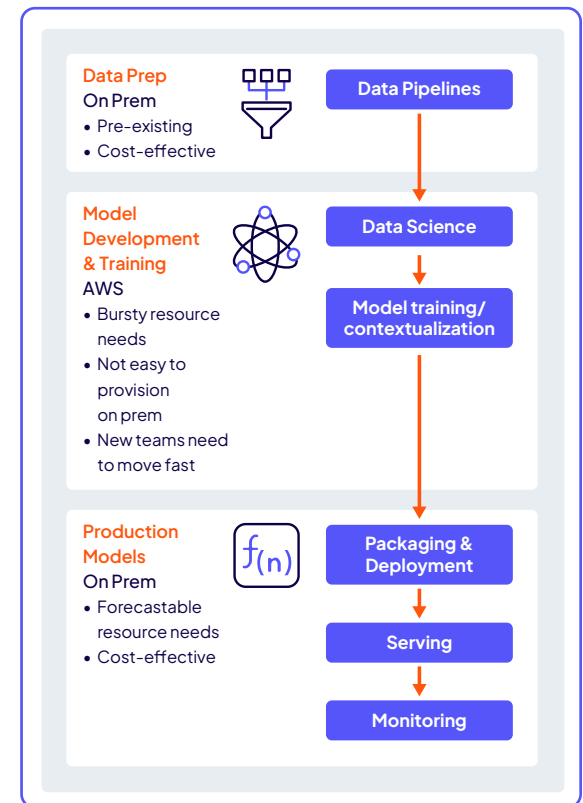
A Multinational Oil & Gas Company Improves Cybersecurity with Cloudera
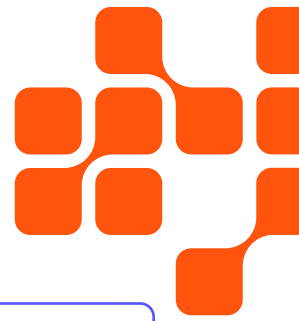
A multinational oil and gas company headquartered in Houston, Texas, needed to improve its ability to detect and respond to cybersecurity threats and protect critical business intellectual property. This process involved rapidly collecting and processing alerts from thousands of connected devices in real-time, and ingesting and analyzing log data from a range of environments, including Personal Computers (PCs), multiple clouds, and on premises data centers.

To solve this challenge, the company replaced Amazon EMR with Cloudera on AWS. They used a combination of Apache Kafka, NiFi, and Flink to collect, process, and store data from a variety of sources and environments in real-time.

Finally, they extended their cloud environment to Microsoft Azure and their on premises private cloud to run parts of their machine learning workflow in different environments.

With this hybrid solution, they could easily ingest, process, and analyze data from 130,000 PCs globally, reduce log data by 60%, and reduce log data software costs by $2 million over five years.

More importantly, they reduced the time to detect a threat from 70 minutes to 7 minutes — a 90% improvement.

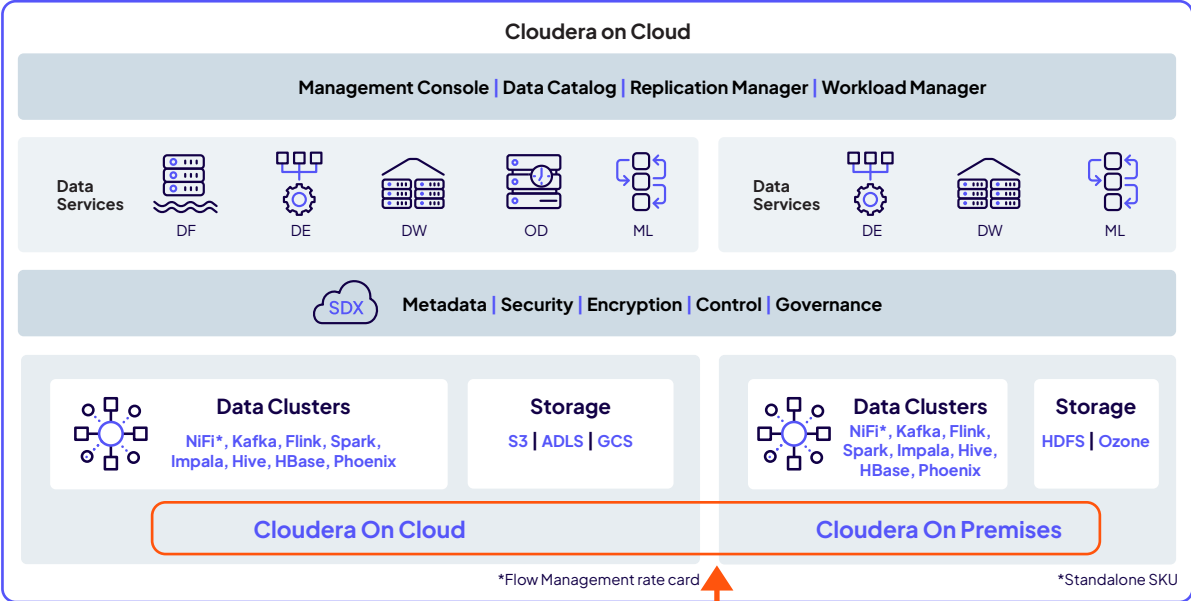Improved detection time of cybersecurity threat by
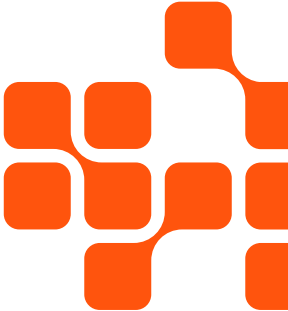
90%

CLOUDERA

Continuous Optimal Deployment

Continuous optimal deployment is the natural evolution of the previous two use cases. By leveraging a "Develop once, deploy anywhere" methodology and adopting a consistent set of data services that are available in any cloud or on premises environment, data teams can choose the best infrastructure for each workload and easily lift-and-shift to a more ideal infrastructure if business and/ or technology requirements change.

This deployment strategy not only ensures that workloads always run at peak performance and cost efficiency, it also ensures that the architecture is future-proof. Organizations are not locked in to a single infrastructure solution, and they can be agile and quickly respond to market conditions.

Cloudera on Cloud

Management Console | Data Catalog | Replication Manager | Workload Manager

Data Services — DF | DE | DW | OD | ML

Data Services — DE | DW | ML

SDX — Metadata | Security | Encryption | Control | Governance

Data Clusters — NiFi*, Kafka, Flink, Spark, Impala, Hive, HBase, Phoenix

Storage — S3 | ADLS | GCS

Data Clusters — NiFi*, Kafka, Flink, Spark, Impala, Hive, HBase, Phoenix

Storage — HDFS | Ozone

Cloudera On Cloud

Cloudera On Premises

*Flow Management rate card

*Standalone SKU

Business Imperatives

| Economy	Carbon Footprint
Resilience	Compliance
Performance	Etc.

CLOUDERA

State Government Builds an Innovation Platform with Cloudera to Deliver Citizen Services while Reducing Total Cost of Ownership

A top 10 US state by population needed to build a flexible platform for delivering citizen services while reducing the Total Cost of Ownership (TCO) of maintaining that platform. Using their modern data architecture, they sought to quickly roll out new services, and make those services easily accessible by state citizens.

To achieve this, they built a hybrid data architecture with an on premises environment and a public cloud environment on AWS. They adopted Cloudera services to address the full data lifecycle. With the hybrid portability of Cloudera, they could choose the best infrastructure for each workload, based primarily on cost-efficiency.

With SDX, the data team could ensure security and governance across all of their data sources while also improving data sharing and the delivery of shared services. Finally, they achieved cost savings and lowered the TCO of their innovation platform through consolidation of data services and elimination of third-party point solutions.
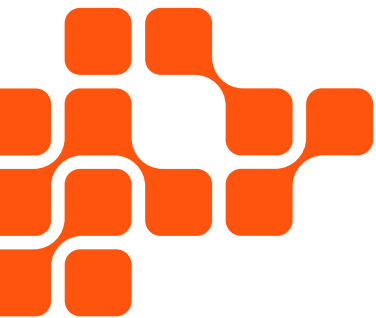
CLOUDERA

About Cloudera

Cloudera is the only true hybrid platform for data, analytics, and AI. With 100x 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