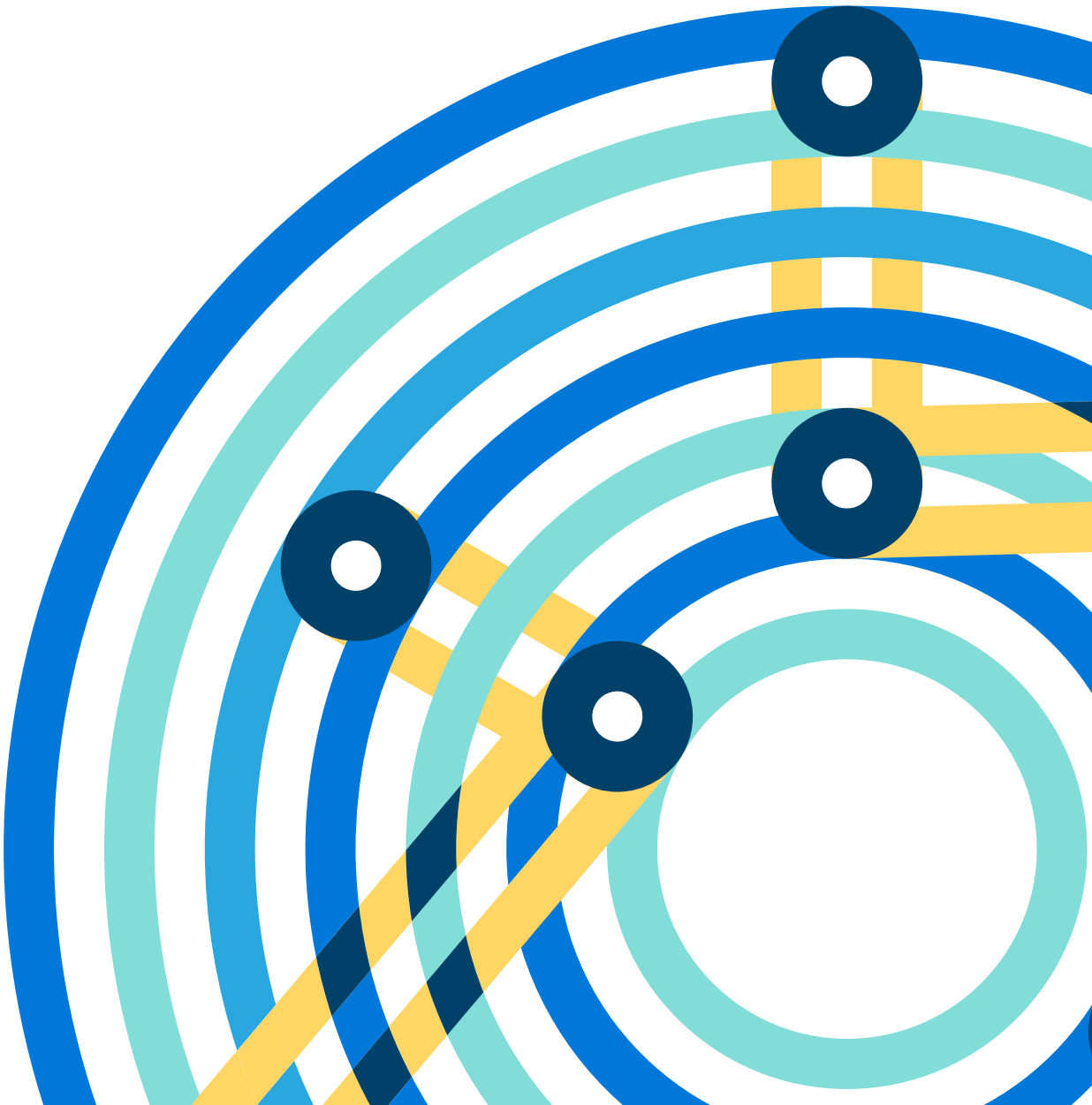
cloudera®

WHITE PAPER

Cloudera Enterprise

The Industry Standard for a Complete Data-in-Motion Solution



Version: 102

cloudera

Table of Contents

cloudera

Data in Motion

Partial solutions yield partial benefits. To derive all the value from your data, you must have an end-to-end solution inclusive of ingestion, processing, and serving.

- **Ingest** - Continuous streaming of real-time data from anywhere in the network, including from lightweight edge nodes powered by open-source software. Best-in-class partners provide a rich visual user interface.
- **Process** - Apache Spark Streaming provides the strongest processing engine available in open-source software. Cloudera's unique experience integrating Spark into the Hadoop ecosystem provides a seamless experience.
- **Serve** - A broad set of choices means your data is served efficiently and quickly to suit your use case: Apache Solr's natural language search, Apache Kakfa's high-throughput distributed messaging, Apache HBase's fast NoSQL access, or a combination of batch analysis and real-time streaming through Apache Kudu.

Introduction

Data in motion refers to data that is in the midst of a process meant to transform it into information an organization can act upon—including ingestion, processing, and serving. Data is the lifeblood of any modern business, and presenting relevant information at decision points helps make those decisions better. Whether information is used by an automated process, delivered broadly via a data application, or leveraged by executives to make strategic decisions, it has the potential to make a dramatic impact on the organization.

The acceleration of the "3 V's of Data"—variety, volume, and particularly, velocity—is making data in motion more relevant than ever. In the past, we collected standard business data and ran basic analyses on that limited data. Today, data is generated not only from standard business processes, but also social media outlets, IoT sensors, markets, and further internal/external sources. Unstructured data will account for more than 80% of the data collected by organizations within a decade[1], something that Moore's law alone can not overcome. The businesses that have a modern, software-based data strategy capable of driving real-time insights out of this torrent of data will have a distinct competitive advantage over those who cannot.

The need to drive real-time insights means that each step of the data-in-motion process—ingestion, processing, and serving—is critical. Like a chain, it can only be as strong as its weakest link. IoT data has driven a focus on ingestion, but without the ability to process this data into something useful, there is no insight. Similarly, without the ability to serve processed data to the right place at the right time, opportunities are lost and poor decisions can be made.

Cloudera offers a fast, easy, and secure data-in-motion solution that encompasses best-in-class software for ingestion, processing, and serving of your data. Our expertise, gained over hundreds of real-time use cases, ensures we can get your data-in-motion solution into production quickly, yielding a rapid return on investment.

Challenges of Today

As businesses move to a modern data strategy, they know they need to be more nimble than their competitors. The ability to ingest, process, and serve data quickly yields more informed decisions based on the latest information. As businesses modernize their data-in-motion strategy, they are looking for:

- **Real-time decisions:** Real-time action requires the guidance of real-time data. Companies need a platform that can make information available to business users as it's created by streaming, indexing, processing, and serving data as quickly as it arrives. If the insights this data drives arrive late, the opportunity and value are lost. The need to do this for all data requires a system that scales linearly without spikes in cost.
- **Scalability and flexibility, with performance:** Businesses today want to ingest, process, and serve more data—structured and unstructured—than ever before. To fully unlock that , they need cost-effective systems that can scale to a future with more data and more users. With data, cost-effective solutions are about more than saving money; savings enable more data to be stored, which will enable models to become much more accurate.
- **Data integration:** Up to 40% of the value of data comes through combining data sets[2]. A focus on ingesting data must be matched by the technical ability to merge data sets in real-time while maintaining data security.

Despite the demand for a next-generation platform that can simplify complex data-in-motion issues, many organizations are bogged down by legacy systems. These legacy systems yield:

- **Limited access:** Legacy data silos were typically purpose-built with a specific use and set of stakeholders in mind. To minimize costs, these systems weren't built for cross-organizational access. The result is an environment where data is inaccessible, moves too slowly for real-time decisions, and is duplicated across multiple silos.

[1] Source: Human-Computer Interaction & Knowledge Discovery in Complex Unstructured Big Data

[2] McKinsey Global Institute Analysis

- **Limited data:** Legacy databases can limit data in two ways: the amount of data captured and the types of data captured. When unstructured data cannot be stored, and when cost considerations limit the amount of overall data stored, potential insight is lost.
- **Limited processing capabilities:** Developers need access to a broad set of tools to process data, including batch, streaming, and interactive processing. Many legacy systems offer only one method, which can result in higher costs and an inability to efficiently deliver information to users on time.
- **Higher Administrative Cost:** A natural product of the complexity involved in managing and maintaining separate systems is an increase in administrative costs. IT teams will be less responsive to new requests, and security becomes difficult to manage across these systems.
- **Partial solutions:** Organizations that recognize the need to respond to the real-time demands of today's information environment can over-rotate around the first issue they identify (often ingestion). A full solution must tackle ingestion, processing, and serving of data. Any breakdown in that chain will bleed value.

Modern Data-in-Motion: Moving Business Forward

The growing interest in data in motion tracks in direct correlation with the increased focus on real-time insights. By definition, real-time insights are drawn from the latest data, which means the full value chain that drives data from creation all the way to a decision point must function flawlessly. Ingestion, processing, and serving must occur in near real-time while still providing enterprise-level security; if they do not, the opportunity to take action is lost, or worse, data is compromised. Cloudera Enterprise offers the leading solution for data in motion:

Ingest: Collecting the Data

Today's data-in-motion conversation, like the data journey itself, starts with ingestion. The increase in sensor-generated data associated with IoT, combined with the demands for social media data collection, has created a deluge of unstructured data that is difficult for organizations to contend with. As a common initial bottleneck in the data-in-motion journey, organizations often reach for a robust ingestion solution. However, it's important to understand ingestion as part of a broader real-time data context; it's a critical component, but only the first of three.

Cloudera takes an open-source approach to ingestion, as it does with all three stages of the data-in-motion journey. Identifying the need for a streaming data capture system, Cloudera led the development of Apache Flume, the open standard for collecting and moving a vast amount of log data. The subsequent integration of Flume with Apache Kafka created an ingest architecture that has been replicated across Cloudera's customer base in a variety of use cases.

With Flume and Kafka, Cloudera deploys the leading streaming ingest platform. Flume can provide light weight agents deployed on edge nodes that number in the hundreds or thousands, each of which can be tiered to enable efficient ingest topologies. The integration between Kafka and Flume is bidirectional, meaning either component can be a producer or consumer of data depending on the specifics of your use case.

A rising trend in data ingestion is the use of a rich visual interface that enables a user to interact with their ingestion architecture in an easy-to-use manner. While Cloudera delivers all the functionality underneath, we partner with best-in-class partners such as Streamsets, Cask, and others to deliver rich visualization. This enables Cloudera to focus on our core competency of data management, while enabling vendors with large engineering teams dedicated to visualization to focus on theirs. Portability, neutrality, and history of success for companies like Informatica,Talend, and others in similar spaces creates the best experience for our customers.

Processing: Unlock the Value at Speed

Cloudera relies on Spark Streaming to process data once it is ingested. As the leading open-source processing framework for real-time use cases, Spark Streaming is an open standard and one of the most easily-recognizable components of the broader Apache Hadoop™ ecosystem. Cloudera has a the broadest base of Hadoop-adjacent experience with Apache Spark™ and Spark Streaming; this is a product of early adoption and integration of these projects into Cloudera Enterprise.

cloudera®

Spark Streaming provides the strongest processing solution for data-in-motion use cases as a result of:

- Best-in-class performance:
  - High throughput ensures that jobs will not bottleneck at the processing stage
  - Sub-second latency enables real-time capabilities
- Best-in-class API and Features:
  - Easy-to-use SQL based API's for authoring streaming jobs help expand the number of use cases and value of data in motion
  - "Exactly once" stream processing semantics help ensure accuracy
  - Sliding window computations enable fast insights into time period data slices
  - Built-in API's for maintaining and updating in-memory information
- Best-in-class ecosystem:
  - Largest set of vendors working with and around Spark among available processing engines, enabling access to latest innovations
  - Broadest and deepest machine learning library (MLib) is seamlessly integrated

Spark Streaming from Cloudera, in particular, benefits users through the most robust integration into the ingestion and serving phases that bookend the data-in-motion story. This integration ensures a fast, easy, and secure delivery of processed data to the serving stage of data in motion.

Serving: Inject Data into Real-Time Decisions

Whereas ingestion and processing have a relatively consistent flow irrespective of use case, the serving phase of a data-in-motion solution requires a variety of options in order to deliver the right data, to the right place, at the right time. Without this ability to quickly serve data to decision points, a solution loses its real-time capability and ceases to become a data-in-motion solution. Cloudera has a variety of options that help serve the diverse needs of individual use cases:

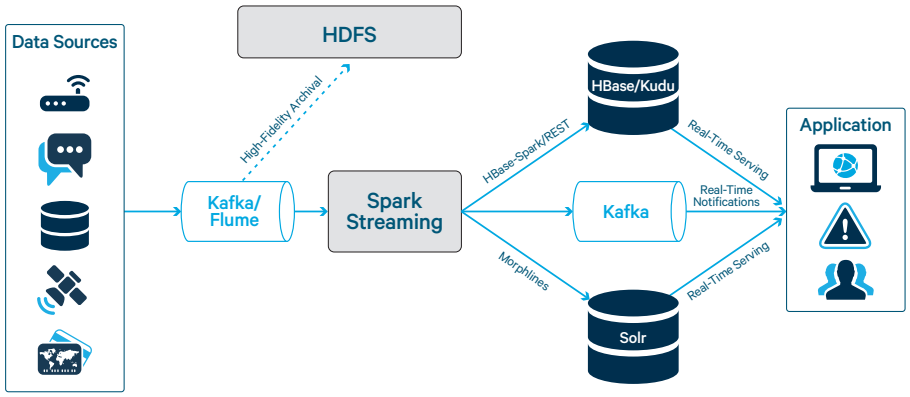
- **Apache Kudu™:** A new, Cloudera-initiated Apache project, Kudu offers the unique ability to do fast scans on fast data. With an overwhelming number of data-in-motion use cases requiring analysis or visualization of streaming data, Kudu can enable the required batch analysis and real-time serving within the same storage layer.
- **Apache HBase™:** HBase offers the best random read/write performance of any component within the Hadoop ecosystem. This capability, combined with high levels of concurrent access, enables online applications and operational needs that require the ability to query the latest data.
- **Cloudera Search:** Powered by Apache Solr™, Cloudera Search democratizes data by enabling non-technical users to perform SQL-like, faceted search in natural language. Solr's native integration into Cloudera Enterprise generates faster and more secure results.
- **Apache Kafka:** Kafka's fast, scalable, and durable design enables hundreds of megabytes of reads and writes per second, from thousands of clients.In addition to playing a role in ingestion, Kafka can be used to serve data to applications and users.

This "last mile" step in the data-in-motion story is arguably the most critical step, which is why this breadth of options is necessary. Each use case, including the tendencies and workflows of the expected users, requires a different set of data access capabilities. Cloudera can meet any requirement through these tools, and can do so as the final step in an end-to-end data-in-motion story.

cloudera

A Canonical Architecture

Cloudera's position as the leader in the Apache Hadoop ecosystem allows us to leverage the experience of answering hundreds of use cases across a variety of industries. Though each use case is unique, a standard architecture for handling data in motion has become clear:



Stream Processing Integration

Cloudera's data-in-motion architecture is open-source, easy-to-use, secure, and integrated. Intensive interoperability testing between these standard Cloudera components means it is easy to assemble this architecture, which decreases time-to-value. The performance of this architecture is proven, enabling the customer to collect, process, and serve millions of messages per second with modest hardware, all with sub-second latency.

"Our remote diagnostics IoT platform built on Cloudera's data hub has helped reduce fleets' maintenance costs by an average of 30 to 40 percent."

– Terry Kline, CTO, Navistar

Navistar Improves Fleet Efficiency with Data in Motion Powered by Cloudera

Real-Time Vehicle Telematics with Cloudera Transforms Fleet Efficiency

Navistar is a leading manufacturer of commercial trucks, buses, defense vehicles, and engines. With traditional data warehouse technologies, Navistar struggled to implement telematics—the integrated use of telecommunications and informatics to inform decisions on and around the vehicle. With a data-in-motion solution from Cloudera, Navistar now has the ingest and processing ability to stream vehicle sensor data and marry it with meteorological, geographical, engineering, traffic, warranty, historical, part, and service data. Over 180,000 vehicles are now tracked—with the data served to multiple divisions in a variety of means. The solution has enabled predictive analytics, vehicle management, vehicle diagnostics, route optimization, and more.

Key Use Cases

Data-in-motion solutions have a vast range of use cases across all industries. If you're looking to inform decisions in real time, you have a data-in-motion use case; if you have a data-in-motion use case, you need to consider the complete process of ingestion, processing, and serving. A broad look at industry use cases includes:

- **Credit cards:** Identify fraudulent transactions as soon as they occur
- **Transportation:** Dynamic re-routing of traffic or vehicle fleet
- **Retail:** Dynamic inventory management; real-time in-store offers and recommendations
- **Consumer Internet & Mobile:** optimize user engagement based on user's current behavior
- **Healthcare:** Continuously monitor patient vital stats and proactively identify at-risk patterns
- **Manufacturing:** Identify equipment failures and react instantly; perform proactive maintenance
- **Physical Surveillance:** Identify threats and instructions in real-time
- **Digital Advertising & Marketing:** Optimize and personalize content based on real-time information

The aforementioned use cases, as well as many others, provide a specific view into the industry use cases Cloudera can deliver via a data-in-motion solution. These industry use cases can be grouped into two broader categories that describe the two primary applications of data in motion:

- **Real-Time Stream Processing:** Processing data as it is produced to make it available and actionable at a decision point.
- **Streaming Data Integration:** The ingestion of continuously produced data from a variety of disparate sources, which is brought together to create new insights.

These two use cases are collectively exhaustive, but not mutually exclusive. There will be use cases where data needs to be simultaneously streamed in real-time and appended with data from multiple sources to truly drive insight. An example of this would be GPS guidance that pushes a route to a driver based on where the driver is, what the traffic conditions are, and what the weather is.

Security: Protecting the Data

Data security is no longer a checkbox in the IT or operations departments, but is a top business priority for most enterprises. As compliance requirements like HIPAA and PCI-DSS continue to expand in scope, and as unstructured data—often sensitive information relating to customers and corporate IP—proliferates inside an organization, the requirements and restrictions on data and data access have come under increased scrutiny.

Cloudera Enterprise is the only Hadoop platform to provide out-of-the-box encryption for both data in motion as well as data at rest as it persists on disk or other storage mediums. As part of Cloudera Navigator, Cloudera customers can now leverage industry standard AES-256 encryption for all HDFS files, HBase records, Apache Hive™ metadata, and audit logs. The encryption runs at the filesystem level and is completely transparent to the applications reading and writing to disk; so performance overhead is minimal, and deployment is quick and painless.

Cloudera's security features include:

- **Massive Scalability**: Node-based encryption for the fastest, most scalable security on Hadoop
- **Powerful, Flexible Key Management:** Software-based key management adds multiple layers of policy and protection for any and all Hadoop encryption keys
- **Secure Sensitive Artifacts:** Store, secure, and manage any sensitive, security-related object generated by Hadoop (i.e. keys, truststores, passphrases, and more) in a "virtual safe-deposit box"
- **Quick And Easy Deployment:** Software, available through Cloudera Navigator, can be deployed in hours, rather than days or weeks
- **HSM Integration:** Master encryption keys can be stored in environments with existing HSM-centric compliance and security policies
- **At-Rest Encryption:** High-performance data-at-rest encryption for Hadoop that meets standards for compliance and safeguarding PII

Conclusion

A data-in-motion solution provided by Cloudera enterprise gives architects, developers, and users the easiest, fastest, and most secure experience. Cloudera's unparalleled ability across the combined architecture of ingestion, processing, and serving makes it the best choice for real-time use cases.

These insights enable better decisions throughout the organization, which can reduce risk, advance new products or services, and lead to better customer insights. Contact us for more information on how you can become truly data-driven.

cloudera®

About Cloudera

Cloudera delivers the modern data management and analytics platform built on Apache Hadoop and the latest open source technologies. The world's leading organizations trust Cloudera to help solve their most challenging business problems with Cloudera Enterprise, the fastest, easiest and most secure data platform available for the modern world. Our customers efficiently capture, store, process and analyze vast amounts of data, empowering them to use advanced analytics to drive business decisions quickly, flexibly, and at lower cost than has been possible before. To ensure our customers are successful, we offer comprehensive support, training, and professional services.

cloudera.com

1-888-789-1488 or 1-650-362-0488
Cloudera, Inc. 1001 Page Mill Road, Palo Alto, CA 94304, USA