BUILD MISSION-CRITICAL APPLICATIONS WITH CLOUDERA OPERATIONAL DATABASE

Cloudera Operational Database

THOMSON REUTERS

Separates real news from fake news on Twitter in 40 milliseconds

- Uncovers ground-breaking events ahead of major news organizations
- Captures and detects news events across millions of tweets in 40 milliseconds
- Frees journalists to focus on higher level reporting

"To assist in evaluating the veracity of an event, we rely on hundreds of features and have trained the platform to look at the history and diversity of sources, the language used in tweets, propagation patterns, and much more, just as an investigative journalist would do."

Director of Research, Thomson Reuters and Lead Scientist, Reuters Tracer

Introduction

Data is driving modern business. Supplied with the right data at the right time, decision makers across industries can guide their organizations toward improved efficiency, new customer insights, better products, better services, and decreased risk.

However, legacy infrastructure and overtaxed analytics teams can stifle the potential of data—a problem that is amplified as data is collected from additional sources and sectors inside and outside the business. Modern businesses require a modern data strategy to provide all employees and functions with the real-time information they need to succeed. To serve such a broad set of stakeholders, a successful data strategy must include the operationalization of data through a flexible, scalable platform that is easily accessed by users and analytic applications alike.

Challenges of Today

As businesses move to a modern data strategy, the operational database is transforming from a straightforward center of day-to-day records to the data engine that powers business decisions. In particular, organizations are now demanding the following from their operational databases:

- **Scalability and flexibility, with performance**: Businesses today want to process, store, and analyze more data—structured and unstructured—than ever before. To fully unlock that ability, they need cost-effective systems that can scale to a future with more data and more users.
- **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. The need to do this for all data requires a system that scales linearly without spikes in cost.
- **Model scoring**: Data is the key to model scoring, which in turn is the key to industries whose offerings rely on the ability to segment and score their customers. To offer better products and pricing via model scoring, companies must store, process, and analyze more data than ever before.
- **Online analytic applications**: Companies looking to realize the value of their data must move beyond single-discovery-oriented analytics and operationalize the delivery of data to the entire organization. Real-time data applications democratize data and enable better decision making at all levels.
- **Need for Hybrid, Multi-Cloud**: Companies need the flexibility to use cloud providers of their choice without single provider lock-in.
- **Lack of Unified Platform**: Fragmented systems lead to increased costs and complexity. IT spends more time fighting fires than adding value

Despite these demands, many organizations still operate within a legacy operational model—one in which systems and reporting were designed to give senior management a very specific view of the business. Today, this model faces considerable stress and breaks down when presented with the demands of modern businesses—particularly the acceleration of data-driven applications—

Saves lives through timely detection of sepsis for successful treatment

- Builds a more complete picture of patients, conditions, and trends
- Has saved 100s of lives already
- Reduces hospital readmissions
- 2PB+ in multi-tenant environment supporting 100s of clients
- Secure yet explorable

"Our clients are reporting that the new system has actually saved hundreds of lives by being able to predict if a patient is septic more effectively than they could before"

Senior Director and Distinguished Engineer, Cerner

resulting in:

- **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 in which data is inaccessible, moves too slowly for real-time decisions, and is duplicated across multiple silos.
- **Limited data**: Legacy operational databases can limit data in multiple ways—the amount of data captured, the types of data captured, and how quickly data is streamed into the database. When unstructured data cannot be stored, or when cost considerations limit the amount of overall data stored, the potential to drive insight is lost. If latency is too high, the opportunity to act on data may be 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.
- **Limited cost control**: Legacy systems that rely on proprietary technologies have cost structures that rise significantly in the face of more data, more users, and varied capabilities. This creates a scenario in which the default decision is to restrict data access and dispose of noncritical data, destroying potential value in the process.

These limits—on access, data, processing capabilities, and cost control—combine to restrict the amount of overall insight an organization can draw from its data. When an organization lacks insight about its market or internal operations, it is put at a competitive disadvantage.

From the Edge To AI

Cloudera delivers a unified experience from the Edge to AI through a portfolio of technologies spanning real-time streaming, operational database, data warehouse, and machine learning. Cloudera's Operational Database enables organizations to build real-time, data-driven applications that support high concurrency and robust scale.

Enhances customer experience with network visibility, tailored offers

- Open Network initiative delivers unprecedented visibility and real-time network information to 270 million customers
- Personalized communications and offerings delivered as result of 360-degree customer view, combining data from network, handsets, call centers

"Now, we are able to gain an even greater technical edge, empowering our marketers with intelligent data and analytics to make better decisions and improve the entire customer lifecycle with customized offerings.

Group CIO for Airtel

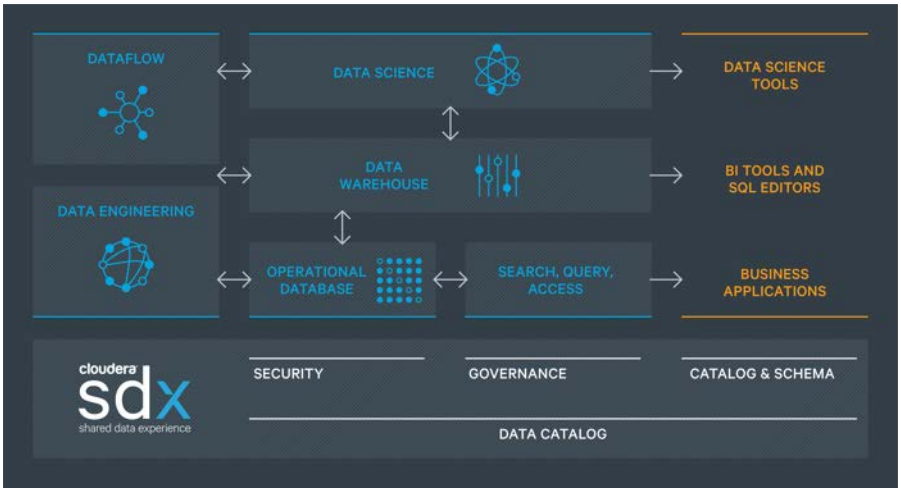

Figure 1 - Cloudera Portfolio

Next Generation Operational Database

Cloudera Operational Database is a real-time, scalable platform with the ability to serve traditional structured data alongside unstructured data within a single open-source platform. These capabilities enable the cost-effective delivery of instant insights to decision makers at all levels of the business, delivering on the promise of data democratization. Cloudera's Operational Database, a vital component of an enterprise data hub, delivers:

- **Unprecedented scale and flexibility**: Bring together and process more data of all types, from more sources (including IoT), and drive new business insights all within a single, scalable platform. Designed for both unstructured and structured functionality, unlimited full-fidelity data can be immediately accessible for processing and analytics, supporting real-time updates. Advanced search features make it easy for nontechnical users to explore, navigate, and correlate data points.
- **Versatile real-time capabilities**: Built with cutting-edge technologies, Cloudera's platform provides real-time database capabilities, even while supporting high user concurrency, so all users across the business can access relevant data.
- **Open Platform**: Cloudera's open platform supports open storage, open compute, and open integration opportunities to avoid vendor lock-in. Cloudera's open-source core, combined with tools that make the platform easy to manage and govern, lowers TCO and eliminates the need to make decisions on what data to keep.
- **Best platform for data applications**: Cloudera Enterprise offers a broad set of processing frameworks within a single platform, including batch, streaming, and interactive processing, that ensure applications deliver information to users efficiently and on time. As data sets, data-driven applications, and data users grow, Cloudera Enterprise offers linear scalability in performance alongside manageable incremental costs.

- **Compliance-ready security and governance**: The democratization of data through applications and broad direct access cannot come at the expense of security—you need to ensure that all data is protected. Cloudera provides compliance-ready data platform, with built-in security and governance at the core. Cloudera SDX provides a central location for governing, securing, and auditing all data usage across multiple workloads regardless of where they are deployed. No matter how users, across roles, access data or how regulated the data itself is, you can continue to enable new users and new insights without compromising your most valuable asset.
- **More value across more workloads, in any environment**: Cloudera's single, unified platform comes complete with the best-in-breed technologies for a wide range of critical workloads and the ability to extend for new workloads. For operations, this means delivering real-time business information blended with additional data sets that lend context for decision makers. The platform can also power data engineering, data science, and analytical workloads across the same data. All powered by the top open source technologies, and portable across any environment—be it on-premises, cloud, or a hybrid deployment—equating to complete freedom from lock-in for your business.

Key Use Cases

The operational database has been a critical component of enterprise technology for decades. These systems have incrementally expanded to accommodate additional data and reporting requirements over the years. However, today's combination of increased data volumes and increased demand for data-driven insights requires a technological leap forward. New use cases that provide instant insights, and even traditional use cases that now deliver more data to more users, require the power of Cloudera Operational Database. These new use cases include:

- **Self-service/embedded data applications**: Cloudera's unlimited storage, high concurrency, and real-time capabilities are uniquely equipped to enable broadly distributed, self-service BI and analytic capabilities via data applications. In addition to operationalizing data usage, applications help to make data delivery fast, data consumption easy, and data access secure. With data in hand, better decisions are made throughout the organization, both manually and automatically.
- **Monitoring and detection**: As the amount of data created and managed continues to grow, so too does the incentive for bad actors to attack. By deploying a robust monitoring and detection data platform, companies can protect their data with data. Outliers are flagged based on modeled expected behavior, and Cloudera's real-time capabilities including Apache Spark, Cloudera DataFlow (CDF) allow action to be taken immediately.
- **Model Scoring and Serving**: Real-time capabilities, combined with historical data sets, enable predictive modeling that allows businesses to take action using the latest information in the brief window that can influence business outcomes. This can include decisions where the outcome is in the distant future, such as credit approvals, or in the immediate future, such as IoT decisions on assembly lines or oil rigs. Cloudera Operational Database, in

conjunction with Cloudera Data Science Workbench (CDSW) and expertise from Fast Forward Labs, helps customers rapidly accelerate model building, scoring, and serving.

- **Recommendation engines**: Recommendation engines can help retailers build multi-channel strategies and increase the cross-sell and upsell of products. However, customer decisions on whether or not to buy are made in an instant. Cloudera's Operational Database provides the real-time capabilities needed to analyze a potential purchase and provide instant relevant offers to increase sales.
- **Customer 360**: As companies mature, it becomes crucial for them to segment their customers and provide more personalization via targeted marketing and specialized offers. They need to react fast and deliver time-relevant offers to customers based on a myriad of data. Cloudera Operational Database helps customers develop a holistic solution by combining data points from multiple sources to help create a holistic 360-degree view of customers, products, devices, and systems.
- **IoT – Operational & Monetization**: As companies collect an eclectic mix of new and old data and aim to become more data-driven, they are challenged in their attempts to utilize this data strategically to improve operations and introduce new business models for a sustainable advantage. Cloudera Operational Database enables customers to improve asset utilization and product design, reduce operational expenses, and introduce new services and monetization models.

Cloudera Enterprise, fueled by Cloudera's leadership in the open source software community, provides the most flexible operational database platform. Cloudera offers both batch and stream processing frameworks, both relational- and NoSQL-styled storage layers, and the ability to store an unlimited amount of structured and unstructured data. Furthermore, advanced search capabilities can help any users access all relevant data. When combined with Cloudera's management, security, and governance tools, only Cloudera can provide users with an operational database that works today and has the flexibility for tomorrow.

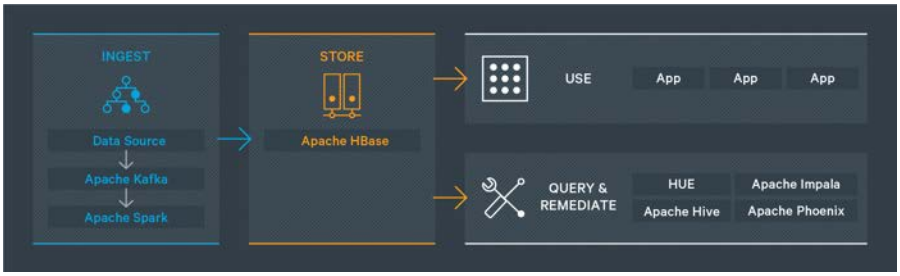

Figure 2 – Build Mission-Critical Applications With Cloudera Operational Database

Processing Framework Flexibility

Developers need flexibility in processing frameworks in order to manage costs,

Cloudera SDX

Tackling complex data-driven problems requires analytics working in concert, not isolation. Cloudera Shared Data Experience, or SDX, combines enterprise-grade centralized security, governance, and management capabilities with a shared data catalog, eliminating costly data silos, preventing lock-in to proprietary formats, and eradicating resource contention.

SLAs, and additional features. With Cloudera's operational database, users have access to both batch processing and stream processing, as well as search functionalities.

Batch processing is an effective approach for use cases that can tolerate some latency in order to save computing resources. In any data management tool, the objective is to give users the ability to efficiently meet the SLAs the business requires. For use cases where the output is needed only periodically (payroll systems, bank statements, membership lists, etc.), this is an excellent option.

Stream processing, powered by Spark and Cloudera DataFlow (CDF), provides users with the real-time capabilities that many next-generation use cases rely on. Processing data upon ingest makes data available immediately to users. Without this real-time information, any use case that demands immediate action —next best offer, predictive maintenance, etc.—couldn't be acted on in time to change the outcome.

Data Store and Serving Flexibility

The storage layer of a database system must align with the use case, or users and developers alike will be frustrated as complex hybrid architectures struggle to meet the needs of the business. Cloudera's operational database offers multiple storage options to enable customers to choose the option that works best for their use case.

Cloudera's NoSQL database is powered by Apache HBase. HBase is a high-performance, strictly consistent, distributed data store that provides flexible, integrated data storage and real-time data access. This is ideal for use cases in which it's important to quickly find and write to individual rows, a common requirement for operational databases. The tight integration with the broader ecosystem enables HBase to take full advantage of the broad suite of tools, while integration with Cloudera's management, security, and governance tools such as Workload XM, Shared Data Experience (SDX) makes it fast, easy, and secure.

Search functionality driven by Apache Solr and SQL access facilitated by query engines such as Phoenix, Impala, and Hive, make Cloudera's operational database accessible to authorized users, delivering data democratization. Interactive full-text search and faceted navigation let users explore and analyze data in real time to find what's relevant and gain new insights. Solr supports batch, real-time, and on-demand indexing (and re-indexing) of data of any type, so more users can get value from data, faster.

Apache Phoenix

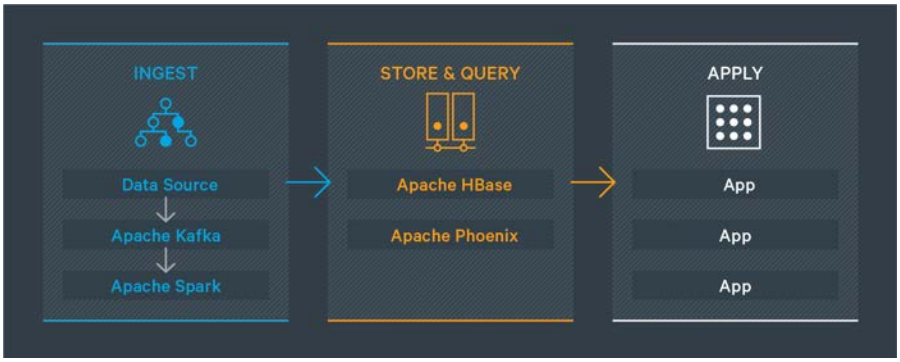

Figure 3 - Hbase and Phoenix

Apache Phoenix abstracts the underlying data store and facilitates data access with a familiar SQL interface.

Phoenix is typically used in two ways: to gain access to HBase for developers and analysts, or as the interface to support operational workloads, with the operational data residing on HBase.

Apache Phoenix has its own query engine, metadata repository and JDBC driver. It provides DDL like schemas analogous to the relational world on top of data residing in HBase. It innovates by providing read-only views and ability to dynamically add columns at runtime for schema extension. Phoenix supports all of the SQL join syntax with equality constraints and correlated subquery support over HBase. The latest release of Phoenix has support for writing UDFs as well as support for extensive out-of-the-box functions and transactions. Apache Phoenix compiles SQL queries to HBase scans and orchestrates the execution of these scans

Fine-Grained Permissions and Data Protection

More users accessing more data using more tools can often mean a security nightmare, especially for highly regulated or sensitive data. That is why Cloudera has built multilayered security into the core of its platform, so businesses can embrace the flexibility and accessibility without the risk to their data and reputation. Cloudera SDX ensures security administrators have a single, central location from which to set access permissions for users and their roles once, and for these permissions to apply across the entirety of the platform. Organizations can be confident they comply with regulations and safeguard their data to which entitled users have easy access.

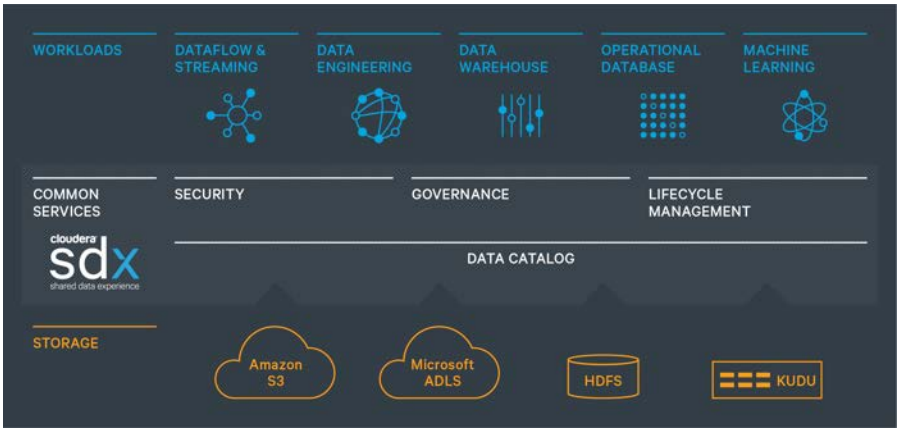

Figure 4 - Cloudera SDX

For the data itself, only Cloudera provides enterprise-grade encryption and key management. With chip-level optimizations, Cloudera Navigator Encrypt lets you encrypt all data, including metadata, logs, and more, without impacting the performance of end analytics. And Navigator Key Trustee ensures your encryption keys are secured and separate.

Data Anywhere and Deployed Anywhere

Cloudera Operational Database can be deployed in any location: on-premises, public cloud, private cloud or hybrid; you choose where and how to deploy. With Cloudera's flexible and portable distribution, you can move between your data center and the cloud, or even between public clouds, elastically and cost-effectively.

Integrated with Your Ecosystem
Cloudera has a broad partner ecosystem of thousands, ensuring that you can continue to use your favorite third-party technologies, whether they're within the Cloudera platform or outside of it. Industry- leading certifications mean these technologies are deeply integrated for a seamless end-to-end experience, so you can get started with Cloudera's platform without disrupting your business.

Conclusion

An operational database powered by Cloudera Enterprise enables businesses to find instant insights from their data. 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, Inc. 395 Page Mill Road Palo Alto, CA 94306 USA cloudera.com