cloudera

WHITE PAPER

Modernizing the Data Warehouse with Cloudera Enterprise

Version: 103

cloudera

Table of Contents

cloudera

Requirements by Role

Data Architects
Assess the workloads being run today to successfully develop an offload strategy for adopting new technologies like Hadoop

Database Administrators
Easily get usage-based visibility to understand top users, top queries, and more—allowing you to stay on top of changing data requirements

SQL Developers
Have the flexibility to work with the data directly and efficiently—using existing SQL skills, at interactive speeds

Business Analysts
Enable self-service, exploratory access to data using preferred third-party BI tools—all with low-latency performance

Security Teams
Ensure sensitive data is protected and that only the right users can access the right data based on role—all while adhering to regulatory compliance requirements

Introduction

Analytics are the lifeblood of any modern business—providing valuable business insights to best navigate the future. But many enterprises are stuck in a pattern of simply trying to keep up with the status quo, instead of shifting for the future. To become truly data driven, you need to align your data to your key business objectives—whether it's to drive customer insights, improve the efficiency of products and services, or even lower business risk. For many, this ability to successfully align is out of reach without a modern data architecture.

Challenges of Today

Traditional data warehouses are a long-standing part of IT architectures. However, a number of shifts in the industry are testing the limits of these systems – causing frustrations to the business and IT, alike:

- **Tapping into more types of data:** Companies need to be able to effectively store, process, and analyze greater volumes of data than ever before. What's more, they need the flexibility to do so across any type of data—even when it doesn't fit into traditional, rigid schemas.
- **Democratizing data access:** Access to data and the resulting analytics has previously been limited to a small group of users, requiring IT to act as a middleman between the analysts and the data itself. As users across all departments are demanding self-service access to full-fidelity data, businesses can no longer accept these costly delays in time-to-insight.
- **Taking real-time actions:** Streaming data creates new possibilities for understanding and adjusting business decisions in the moment—but only if the organization is prepared to handle it. The right architecture is required—not just to pair data streams with existing data, but also to be able to analyze it as quickly as it comes in.

Paired with additional shifts towards lock-in free, open source technologies and resource-flexible cloud deployments, traditional systems simply aren't designed to keep up with these changing priorities, leading to:

- **Limited Data for Limited Insights:** Rigid structures make it difficult to incorporate new datasets. Additionally, to keep costs under control, data is only limited to what's known to be valuable at the time—preventing complete historical views or new data sources. This leads to an inability to explore and derive new, more valuable insights, and limits analysis to what you already know.
- **Slow Query Performance:** For BI and analytics, users require interactive performance for results to remain relevant. However, with more users running queries and more complex questions being asked, the performance of these systems has dropped. Too much time is wasted waiting on answers and trying to troubleshoot performance—resulting in cyclical frustrations between business users and IT.
- **Taxing Data Movement Across Silos:** Data silos are the norm for many businesses. Not only does this result in separate systems for data preparation, reporting and analytics, and operational use cases, it also results in those systems being spread across different departments with different access privileges. Moving data between these systems takes time, limits volume, and leads to unnecessarily complex management. It can also lead to data security, governance, and version-control issues.

cloudera

A Modern Data Warehouse in Action

Quotient Technologies Increased Sales by 230% through Greater Customer Insights

Quotient Technology (formerly Coupons.com) is a leading digital promotion and media platform, distributing digital coupons, coupon codes, and card-linked offers. The Cloudera Data Warehouse solution provides a scalable, high-performance, and flexible environment that enables Quotient to build sophisticated behavioral targeting and personalization solutions. This helps retailers and CPG companies increase consumer engagement and loyalty, better tailor outreach to consumer needs, and reduce coupon distribution costs.

"The Cloudera platform enables us to move faster and be more nimble. We're able to answer questions that we couldn't answer with our old environment, or if we could answer them, it was a lot more painful. Queries that before ran overnight are now completed in three seconds."

– Rumman Chowdhury, Analytics Scientist, Quotient

A Modern Solution

Cloudera Enterprise, powered by Apache Hadoop™, is a modern database warehouse designed to tap into the full value of your data. As an adaptive, high-performance, data warehouse, it opens up BI and exploratory analytics over more data—using the skills analysts already rely on—to derive instant value. It's a complete solution all built within a single, adaptable platform—ensuring you'll not only better address the business needs of today, but that you can quickly evolve to address the needs of tomorrow. With Cloudera's data warehouse solution, you get:

- **Unprecedented scale and flexibility:** Cloudera delivers a single, scalable platform that joins and processes more data of all types, from any source, to drive new business insights. Designed for schema-on-read functionality, unlimited full-fidelity data can be immediately accessible for processing and analytics, even supporting real-time updates.
- **High-performance, high concurrency:** Built with cutting-edge technologies, Cloudera's platform provides high-performance analytic SQL—even while supporting high user concurrency—so all analysts across the business have exploratory access to data. And, through integrations with the leading BI tools, your business has immediate access using the tools and skills they already know.
- **Compliance-ready security and governance:** Opening up data for analytics can never come at the expense of security—even the most sensitive data must remain protected. Cloudera is the only compliance-ready Hadoop platform, with built-in security and governance at the core. No matter how users access data, what permissions they have, or how regulated the data 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 with best-in-breed technologies for a wide range of critical workloads, and the ability to extend for new workloads. For analytics, this means the fastest data processing engine and the leading high-performance SQL engine with no data movement. Cloudera's platform can also power data engineering, data science, and operational workloads across the same data. All powered by the top open source technologies and portable across any environment—whether on-premises, in the cloud, or a hybrid deployment - equating to freedom from lock-in for your business.

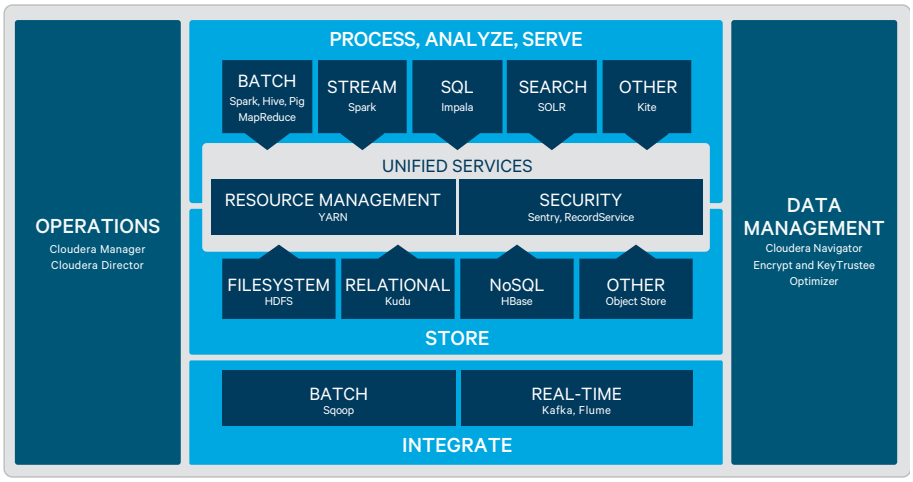
Key Use Cases

Enterprise data warehouses (EDWs) play a key architectural role in many businesses today. However, the previously discussed challenges have put more and more pressure on these systems. ETL pipelines are running slowly or breaking altogether, business intelligence (BI) reports are coming in past their deadline, and query performance altogether is taking a hit as more users try and run ad hoc queries. A modern data warehouse, powered by Cloudera Enterprise, is a perfect complement to these EDWs—able to relieve the pressure from these systems to focus on the complex reporting that they're built to handle. Cloudera Enterprise is a great solution for key use cases, including:

- **ETL Offload:** Free up processing power from your EDW by migrating ETL processing to Cloudera Enterprise. Supported by Hadoop, this platform is designed to handle large-scale, batch processing workloads over flexible data types. This means workloads will run orders of magnitude faster and scale to support more pipelines—so data is always available for reporting or other workload needs, while missed SLAs become a thing of the past.
- **Self-Service BI and Exploratory Analytics:** Open up direct access to your data for more users—without straining the system or your IT department. With schema-on-read flexibility and the compute power to support highly concurrent user access, your developers and analysts can explore data on their own to answer new questions with the fastest time-to-insights.
- **EDW Optimization:** By augmenting your solution with Cloudera Enterprise, you can focus your system on more complex reporting and processing over hot data—without having to make compromises on what data is stored or used for other analyses. An architecture supported by both an EDW and Cloudera Enterprise can reduce data storage costs, improve processing and analytic speed, and unlock new value and use cases across the business.

cloudera

A Deeper Look at the Technologies



Only Cloudera's data warehouse solution provides the tools for success among all key users—from identifying the best workloads to supporting sub-second latencies across millions of queries and thousands of users, and everything in between.

Active Data Optimization

Cloudera's platform is a perfect complement to traditional enterprise data warehouses (EDW)—able to relieve the pressure from growing numbers of ETL jobs and BI analytics so the EDW can focus on the complex reporting that it's designed for. However, this means your big data architects need a plan to offload these workloads the right way. Cloudera Navigator Optimizer ensures you start off on the right foot with Hadoop for analytics. This tool analyzes existing SQL workloads—giving your architects the necessary insights and risk-assessments into workloads across the entire business in order to build out a comprehensive offloading strategy based on risk and development costs.

Additionally, once the appropriate ETL and BI workloads have been migrated, Navigator Optimizer provides continued usage visibility so database administrators can actively manage and recommend appropriate data models for peak performance with Hadoop technologies, such as Apache Hive and Apache Impala.

Finally, SQL developers can quickly master these new technologies with query design and performance tuning assistance based on best practices and compatibilities for Hive and Impala.

High-Performance ETL and Analytic SQL

Cloudera's platform opens up the power of big data to SQL users across a company. When it comes to data preparation, Hive is the de facto technology—with large-scale, flexible data processing and familiar SQL support. Hive-on-Spark now brings even faster speeds to these workloads, leveraging Apache Spark's in-memory processing engine. Once data is prepared, it is immediately available to Impala, the leading analytic SQL engine. Impala enables the interactive performance that SQL developers and analysts require, and, as a massively parallel processing (MPP) engine, it provides this speed over all data while supporting high user concurrency.

With Hue as the out-of-the-box SQL editor, developers can seamlessly discover, design, troubleshoot, and publish queries for other user groups—including better preparation and serving for BI end-users. For these end-users, their favorite BI tools are uniquely integrated with Impala and the platform for uninterrupted analysis.

Finally, Apache Kudu brings this fast analytic performance to real-time updating data. When paired with Impala, this new storage engine lets your business enhance self-service BI and exploratory analytics with streaming or constantly changing data, for near real-time insights to better drive business decisions.

cloudera®

Fine-Grained Permissions and Data Protection

More users accessing more data with more tools can often mean a security nightmare, especially for highly regulated or sensitive data. That is why Cloudera has built multi-layered security into the core of its platform, so businesses can embrace the flexibility and accessibility of Hadoop without the risk to their data and reputation. Apache Sentry and RecordService (available in beta) play a key part in this. Sentry allows security administrators to easily set access permissions for users based on their role once, and the permissions automatically persist across the entirety of the platform. RecordService complements this capability by uniformly enforcing these permissions at a fine-grained level across all access paths in Hadoop, including extending the security to multiple storage engines and third-party tools. Together, this makes it easy for users across the business to have access to the data and platform, without putting the data at risk.

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.

Comprehensive Governance

Governance is not only critical for compliance purposes, it also plays a key role in making the right data available and ensuring results can be validated and trusted. Only Cloudera provides comprehensive governance across the platform—enabling full auditing, column-level lineage, and data lifecycle management. For your security team, this means complete visibility into who is accessing what data and what they're doing with it. For your data stewards, this means they can automatically manage data from ingest to purge, in order to best classify and make the right data available to end-users. Finally, for business users, they can better explore data on their own, understand what is most relevant, and validate any results.

Easy Multi-Tenant Management

With Cloudera Manager, administrators can ensure that every department has the right resources they need to meet their SLAs and achieve peak performance. Administrators can easily monitor and tune resources based on needs and usage, and even troubleshoot problematic queries. Paired with Cloudera Director, resourcing can even be extended across environments with support for cloud deployments, and the ability to elastically scale based on changing demand.

Integrated with Your Ecosystem

Cloudera has a broad partner ecosystem in the thousands, ensuring that you can continue to use your favorite third-party technologies—whether within the Hadoop platform or not. Industry-leading certifications means 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.

cloudera®

Conclusion

A modern data warehouse, powered by Cloudera Enterprise, helps you bring your tools and teams closer to the data in an agile, adaptable environment. It's time to remove the barriers of traditional systems so you can get the full value of all your data. Contact us for more information on how you can become truly data-driven.

About Cloudera

Cloudera delivers the modern platform for data management and analytics. 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 built on Apache Hadoop. Our customers can 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. Learn more at cloudera.com.

cloudera.com

1-888-789-1488 or 1-650-362-0488
Cloudera, Inc. 395 Page Mill Road, Palo Alto, CA 94306, USA