cloudera®

EXECUTIVE BRIEF

Enterprise Data Hub: A New Way to Work with Data

Executive Summary

Across every industry, organizations are on a road to putting data at the center of business transformation, whether the goal is to better understand customers, build new or better products and services, or to manage cost and risk. A recent survey notes that 82% of users believe big data is changing the way they do business[1], and even more believe it will revolutionize operations the same way the Internet did. Analytics are becoming pervasive.

Yet so are the challenges. While existing solutions enable business intelligence teams to build reports and dashboards, analysts to sample data to create models, and development teams to painstakingly convert those models into online applications that help users make decisions in real-time, this is no longer enough. As data volumes grow at exponential rates and new types of data become available every day, users demand more and faster access. It also becomes increasingly burdensome to move all that data around for each new business question or use case. At the same time, IT must simultaneously ensure performance SLAs, control costs, and manage security and compliance.

Many organizations realize their existing systems alone are not sufficient to keep pace with this rate of change, and turn to a new approach to complement their existing investments: an enterprise data hub (EDH). As a unified platform that can economically store unlimited data, and enable diverse access to it at scale, the enterprise data hub is emerging as the architectural center of a modern data strategy.

Data Drives Change

Our relationship with data is changing. Historically a scarce resource, data today is in unprecedented abundance. In order to win and retain customers, deliver better products to market at lower cost, and reduce and manage exposure to risk, today's leaders recognize that data holds the key. Only by embracing data as a strategic resource can organizations begin to capitalize on the opportunity.

Several trends reinforce the importance of having a corporate data strategy:

Instrumentation

The rise of the Internet and connected systems - from mobile phones to instrumented IT infrastructure to Internet of Things (IoT) sensors in cars, medical devices, industrial equipment - have made it possible to capture nearly unlimited information about our world. As a result, we're generating data, and moving it, at a rate that's entirely new. When used appropriately and well, we now have the opportunity to use data can change the world for the better. Unfortunately, according to IDC over 95% of this data is not currently analyzed[2].

Intelligence

In this new online world, expectations of technology have also increased. Across industries, data can help us better understand our customers and channels, build better products and services, or reduce risk and fraud. Retailers delight us when they offer the right product recommendations, at the right time. When we apply for a loan, we want our bank to remember that we are already a loyal customer with credit and checking accounts. We appreciate it when mobile providers offer us service discounts and network performance alerts based on our usage patterns. Yet we also expect that data will be used for good. These new benefits must be balanced with individual privacy. Trust is critical and security matters.

Innovation

With new data and resources now available, the most innovative organizations are embracing agile methods and cultures of experimentation to quickly assess new business opportunities. It is now possible to quickly design and implement data-driven experiments to drive new product and service development, measure the impact of customer service investments, and calculate exposure to risk. New executive roles, such as the Chief Data Officer (CDO), whose responsibility is to develop a strategy for identifying, enriching, managing, and leveraging corporate data assets, are fast emerging.

Big Opportunities. Big Challenges.

Taking advantage of these opportunities presents a challenge for traditional approaches to analytics. It's no longer sufficient to rely solely on a set of predefined dashboards with predetermined data, with multi-month turnaround times for change. We now need the ability to look at any data and ask any question, test our theories, and put the results into action immediately. This process needs to be fast, flexible, and economical, not only for the frequency of questions asked, but for the amount and variety of data and analysis necessary to answer these questions. Experienced IT practitioners understand all too well the challenges in applying existing approaches to the world that runs on data.

[1] Accenture, "Big Success With Big Data." Sept. 2014. www.accenture.com/us-en/Pages/insight-big-data-research.aspx
[2] IDC, "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things." April 2014. http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Limited Data

Faced with the tremendous growth in volume and variety of data, most existing systems cannot keep up, and IT faces poor data processing performance, ballooning storage costs, or both. Scaling these systems to meet growing needs is often expensive and complex.

First, traditional architectures require advanced up-front design thinking to scale, and adding capacity is just as hard. Second, data is expensive to keep online in full fidelity for long periods of time, yet it's often necessary for compliance purposes. With available storage dedicated to existing data, it's hard to think about bringing in new data sources. Most commonly we archive it offline once it's no longer business critical. Unfortunately, once it's in an archive, it's no longer accessible to the business. Third, before data is usable by analysts, it has to be moved and transformed into relational formats, where every transform loses context and creates more copies, at additional cost.

It's no wonder that the common solution is to simply leave data out of the analytic process, especially new data with uncertain business value. Having to make these tradeoffs is not a good foundation for a data strategy.

Limited Impact

All the data in the world has no value unless it's accessible and, ultimately, actionable. Putting modern data into action, though, requires more agility than traditional approaches can support.

Data warehousing methodologies dictate that first you need a model, and then you gather data into that model. This works well when both model and data are well-understood and predictable. But what if they aren't? The leads to painful cycles between business and IT, who can't build without direction from the business, but the business doesn't know what they want until IT produces it. By then it's too late, and there are always new questions to ask.

Given the variety of data of data available today to support analysis, it follows that different techniques are required to work with that data: SQL, exploratory navigation, text search and analytics, and machine learning, for example. Most database management systems stop at SQL, or provide limited extensions. Working with data in other ways requires setting up new analytics silos. Or more likely, most users get no access to analytics.

Finally, even with the most advanced analytic techniques, turning insight into action means getting in front of regular people - customers, employees, patients, students - who are the ultimate beneficiaries of your data strategy. This typically involves anything from publishing better reports to building real-time online or mobile applications that embed models of customer behavior to drive recommendations or detect fraud. Traditional approaches to analytics stop at the analytics; building the application requires completely separate technology and costly reimplementation of the models.

Trust and Compliance

Increased regulatory oversight and individual privacy concerns mean security and governance have never been more crucial. In 2014, 43% of companies experienced a data breach[3] and the accompanying financial and brand damage, and it's only getting worse.

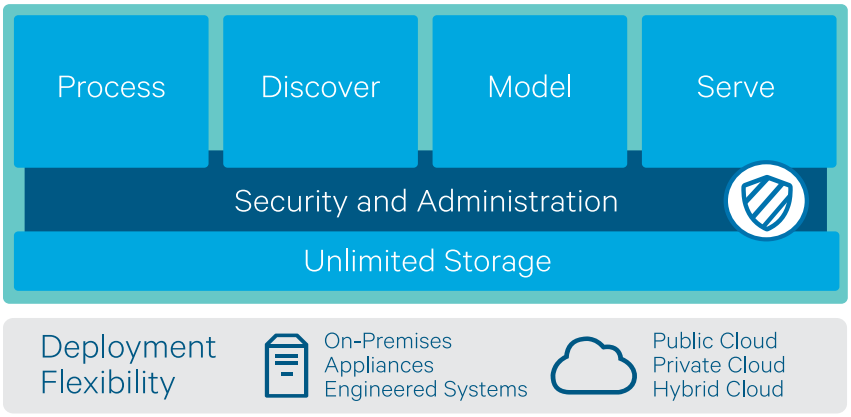
The availability of more data, and more users who want access to data, dramatically increase the complexity of providing a comprehensive strategy. Self-service BI means a change in the control IT has traditionally exercised over analytics. New business models around data services mean more drive to make data available externally, not less. IT and information security teams face the daunting challenge of balancing these concerns with the desire to support business agility.

[3] Ponemon Institute, "Is Your Company Ready for a Big Data Breach? The Second Annual Study on Data Breach Preparedness." Sept. 2014.

Introducing the Enterprise Data Hub

In response to these challenges, a new approach to working with data is required; one that overcomes technological, economical, and team cultural barriers. Leading data-driven organizations have turned to a new architecture to complement and extend their existing analytic investments: the enterprise data hub (EDH).

An enterprise data hub is a powerfully simple idea: A unified platform that can collect and store unlimited data, cost-effectively and reliably, and enable diverse users to quickly gain value from that data through a collection of frameworks that span data processing, interactive analytics, and real-time serving applications. With an enterprise data hub, it is now possible to deliver integrated analytic solutions for less cost and effort than ever before.



Cloudera's enterprise data hub, powered by Apache Hadoop.

An enterprise data hub offers the following key benefits:

Managed Unlimited Data

An enterprise data hub can store unlimited data, in its original formats and fidelity, for as long as you need. As a data staging area, it can not only prepare data quickly and cost-effectively for use in downstream systems, but also serves as an automatic compliance archive to satisfy internal and external regulatory demands. Unlike traditional archival storage solutions, an enterprise data hub is an online system: all data is available for query.

Accelerate Data Preparation and Reduce Costs

Increasingly, data processing workloads that previously had to run on expensive systems can migrate to an enterprise data hub, where they run at very low cost, in parallel, much faster than before. Optimizing the placement of these workloads and the data on which they operate frees capacity on high-end data warehouses, making them more valuable by allowing them to concentrate on the business-critical OLAP and other applications for which they were designed.

Explore and Analyze, Fast

Above all else, an enterprise data hub enables analytic agility. IT can provide analysts and data scientists with a self-service environment to ask new questions and rapidly integrate, combine, and explore any data they need. Structure can be applied incrementally, at the right time, rather than necessarily up front. Not limited to standard SQL, an enterprise data hub offers options for full-text search, machine learning, scripting, and connectivity to existing business intelligence, data discovery, and analytic platforms. An enterprise data hub finally makes it cost-effective to run data-driven experiments and analysis over unlimited data.

Powered by Cloudera

Built on the transformative Apache Hadoop open source software project, Cloudera Enterprise is designed for the demanding requirements of enterprise customers. Cloudera is the leading contributor to the Hadoop ecosystem, and has created a rich suite of complementary open source projects that are included in Cloudera Enterprise.

Hadoop has evolved into a stable, scalable, flexible core for next-generation data management. Yet, along it lacks several critical capabilities necessary for deploying it as the center of an enterprise data hub. It lacks a comprehensive security model across the entire ecosystem of projects. It was built for batch-mode data processing workloads, which limits Hadoop to an ancillary position in the datacenter - a central enterprise data hub must be real-time. And Hadoop doesn't support the range of industry-standard interfaces for query and search applications, among others, that business users require. Cloudera has addressed all of these challenges and more with its solution.

Cloudera offers a single platform from which organizations tackle diverse critical business problems:

- Automatically archiving the complete set of enterprise data to meet compliance requirements with immediate online access;
- Complementing exisitng enterprise data warehouses to offload data and workloads to improve performance while managing costs, and enabling the deliver of high value data sets for operational reporting;
- Supporting self-service business intelligence, through familiar tools, on more data and more kinds of data than ever before possible;
- Enabling and consolidating enterprise search on data and documents in-place within the single environment; and
- Accelerating advanced analytics solutions, such as recommendation engines, fraus detection, or image processing.

Build Data Applications

Once developed and tested, analytical models can be instantly deployed at web scale to create real-time data-driven applications. Naturally, these applications themselves generate rich usage data that can be used to better understand user behavior and optimize the models. Creating automated feedback loops is easy in an enterprise data hub because the architecture provides both analytical and operational capabilities; instrumentation data simply lands right back in the hub where, as with other data, it's immediately available for human or machine analysis.

Conclusion

The modern era of data abundance presents new opportunities and challenges alike for those seeking to gain value from data. For decades the only architectural options have been based on traditional systems, such as storage or data warehousing, which necessarily leave data behind or fail to expand use to all the users who need it. True unified analytic data management has remained out of reach.

By extending existing solutions with an enterprise data hub, it is finally possible to manage rapidly increasing data volume and variety, turning that data into opportunity. Cloudera, the leading contributor to and provider of enterprise Apache Hadoop, has helped hundreds of organizations implement an enterprise data hub on the Cloudera Enterprise platform, which adds to Hadoop the security, governance, and operational capabilities required to succeed in the most demanding environments.

Whether improving performance and reducing costs on current systems, delivering new self-service analytical capabilities and data to analysts and business owners, or enabling the next generation of online data-driven applications, an enterprise data hub has emerged as a powerful new platform at the center of modern data strategy.

About Cloudera

Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data, an enterprise data hub built on Apache Hadoop. Cloudera offers enterprises one place to store, access, process, secure, and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. Cloudera's open source Big Data platform is the most widely adopted in the world, and Cloudera is the most prolific contributor to the open source Hadoop ecosystem. As the leading educator of Hadoop professionals, Cloudera has trained over 22,000 individuals worldwide. Over 1,200 partners and a seasoned professional services team help deliver greater time to value. Finally, only Cloudera provides proactive and predictive support to run an enterprise data hub with confidence. Leading organizations in every industry plus top public sector organizations globally run Cloudera in production.

cloudera.com

1-888-789-1488 or 1-650-362-0488

Cloudera, Inc. 1001 Page Mill Road, Palo Alto, CA 94304, USA