Delivering the Enterprise Data Cloud

A dbInsight white paper for Cloudera

Trigger

Data, and the insight it offers, is essential for business to innovate and differentiate. For most enterprises, the data canvas expands beyond traditional back office transaction systems to encompass nonrelational data from a variety of sources, from inside the firewall out to the edge. The declining cost of infrastructure and the availability of cloud computing have made the storage and analysis of large volumes of data affordable and accessible.

Most enterprises expect that cloud computing will play a part in their current or future analytic workloads. Analysts, data scientists, and data engineers working with the data expect the freedom and flexibility to use the tools, frameworks, and languages that are right for the job. They must be able to process the data where it lives, enabling analytic teams to tackle problems with an agile approach that accelerates time to insight while taking advantage of the economies of scale.

Our take

With data anywhere and everywhere, enterprises expect the flexibility and freedom to run the workload(s) of choice in the deployment environment of choice – on-premises, in the public cloud and/or in a hybrid cloud deployment. Enterprises need a zone of safety where everyone is working from the same source of the truth and where they are assured that their data is secured and governed, in compliance with data privacy and protection policies and mandates.

Enterprises expect their providers to deliver a “better cloud than cloud” experience that delivers the best of both worlds. It means providing a consistent experience between deployments that are on-premises or in the cloud, while also enabling them to move or burst their workloads transparently between environments as business needs for cost, performance, and elastic scaling dictate. This empowers customers to keep their options open by allowing them to take advantage of the benefits of cloud-native deployment without being locked into a single cloud platform.

The attraction of the cloud

The draw of the cloud is unmistakable. While early adoption was primarily tactical, now more enterprises adopt cloud-first strategies. This is especially true when it comes to new applications or operational and analytics use cases involving data that originates and naturally lives outside traditional data centers, ranging from IoT to social media, log files, and trading partners.

Initially, the attractiveness of the cloud was about financial and operational flexibility and agility; it allowed costs to be shifted from capital to operating budgets, and it allowed developers to choose exactly the right compute and storage options to fit the use case. But dbInsight believes that access to *managed services* delivers even greater benefits, as it also delivers on the promise of self-service and operational simplification.

However, until now, managed services had a major catch – most came from the cloud providers themselves. Enterprises considering managed cloud services had to choose the tradeoff between operational simplicity vs. the reality of cloud vendor lock-in.

Enterprises seek "a better cloud than cloud"

As enterprises embrace the cloud, they want the advantages of the cloud without the constraints. They want the flexibility of cloud compute, even for data and processes that must continue to remain on-premises. They also want to keep open their options for choosing the right cloud for the right application, in the right geography – implementing multi-cloud and hybrid cloud as a strategy, not as a default. Additionally, they need a single, unified pane of glass for finding where the data is, how it is secured, and how it is governed.

Follow the data

Data gravity and organizational policies drive deployment choices. Today, data lives everywhere. Enterprises can no longer make the right decisions on use cases such as predictive maintenance, customer 360, fleet management, supply chain management, healthcare patient care, or network infrastructure management strictly relying on the data residing inside enterprise IT walled gardens. The data on which enterprises rely comes from a variety of sources, inside and outside the data center, much of it at the edge, where there is a frequent need for local processing either for reducing demand on bandwidth or for the ability for local autonomy.

Most organizations with existing on-premises compute clusters expect flexibility. They are not going to uproot 100% of their data and compute tomorrow, yet they want to leverage the flexibility and agility of cloud operation. They require a range of options – the ability to utilize their existing infrastructure, and the ability to burst to cloud, or launch new applications in the cloud. They want a consistent deployment environment managed under a single pane of glass that supports traditional on-premises bare metal, and public or private cloud – and the flexibility to run in the cloud(s) of their choice. They want the option to lift and shift existing workloads to the cloud and/or take advantage of the simplification that a cloud-native deployment can offer. Yet most of all, they want to be able to do all this with consistent data security and governance across the board.

Use the tools of choice

Enterprises require a platform that provides the freedom, not only to choose where to run their analytic or operational workloads, but also the flexibility to run the analytic and/or operational workloads of their choice.

That's because the types of analytics required for decision support or exploratory analysis will vary – enterprises need access to all the tools, from inquiry and reporting to providing data scientists and developers access to the tools and languages of their choice. There are multiple constituencies, from business analysts to data scientists, and data engineers who expect to work with the right tools for the job. For instance, business analysts, who are accustomed to working with BI query and reporting tools, require support of SQL for data access. Meanwhile, data scientists expect the freedom to work with the languages of choice, from Python to R and Scala; with the libraries and frameworks of choice.

The most efficient means for addressing these varied constituencies is with a unified data management platform that supports the full spectrum of analytics, from batch processing to interactive and real-time streaming, and the ability to conduct the analytics at the right place, whether in the data center, the cloud, or out at the edge. It is not enough to be restricted to SQL; while data warehouses will remain the workhorses for query and reporting, enterprises also need a platform that supports data

engineering, machine learning, and edge processing. The following use cases show why enterprises need the right tools for the job:

- Fleet maintenance – This requires the ability to process IoT data from sensors on vehicles then manage the movement of aggregated data for processing and real-time operational monitoring, and the ability to train and run ML models that deliver predictive and prescriptive maintenance. This keeps fleets operating with minimal downtime and maximum operational efficiency.
- Location-based marketing – Real-time mobile device data must be anonymized at the edge, with data flow managed in real-time, and the ability to run machine learning that provides the next-best action in real time.
- Cybersecurity – This is another case where real-time streaming analytics are critical. Threats that may be detected along networks and gateways to the data center must be filtered and analyzed in real time. Security professionals require the ability to run ML models to detect and predict how threats are morphing and provide prescriptive solutions that provide recommended actions and/or automatically perform corrective action.

Privacy, Security & Governance

There is also growing awareness of the need to control, secure, and govern retention and access to data. For most organizations, data governance and security have been siloed, tied either to specific databases or applications. Implementation has often been spotty; not surprisingly, "islands of governance" are often the norm.

A perennial issue, the urgency has gotten compounded by the growing sophistication of cyber attackers and growing incidence of data breaches. According to the Identity Resource Center, a US nonprofit group, the number of annual breaches has multiplied from 200 in 2005 to over 1300 last year; another study commissioned by identity intelligence firm 4IQ showed the number of confirmed data breaches quadrupling in 2018 alone with nearly half of compromised identity records striking organizations in the U.S. and China.

Then there is the question of data privacy and protection. It has always been a best practice for enterprises to safeguard the privacy and confidentiality of personally identifiable information, but increasingly voluntary and industry guidelines are being formalized into law. Some of the best known examples, including the 2018 passage of the California Consumer Data Privacy Act, the European Union's General Data Protection

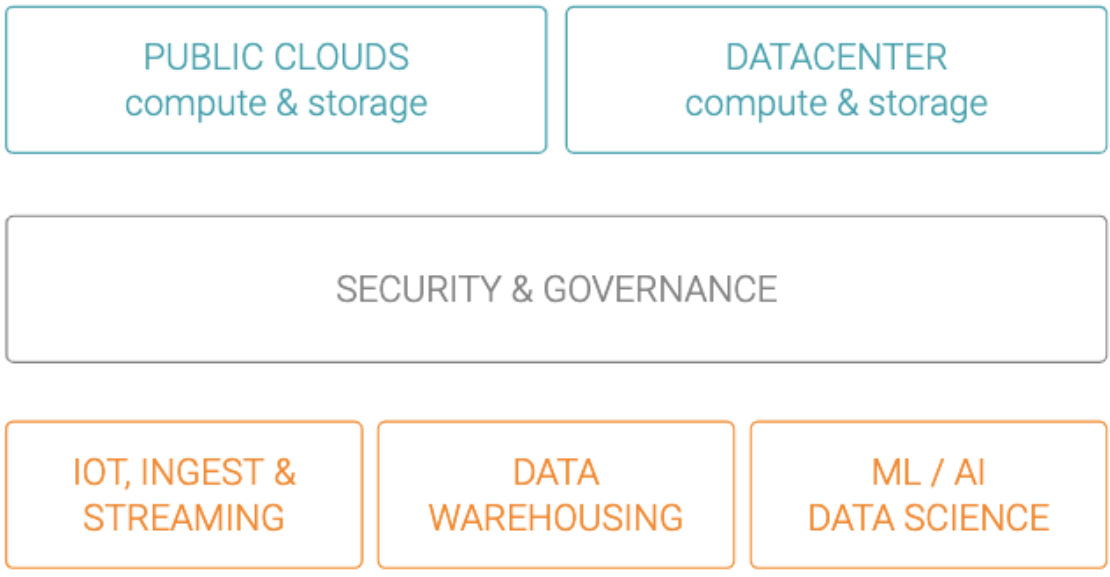
Regulation (GDPR), and China's Cybersecurity Law, may be just the tip of the iceberg when it comes to future regulation. Many of these same laws are also mandating that data stay within the country of origin.

Yet, as enterprises tactically deploy new applications or data sets to the cloud, they risk compounding their issues with islands of data governance. This can be especially problematic for organizations with global presence, who manage compliance under multiple regulatory regimes. They may selectively use different tools for governing and securing their data for individual applications or database on-premises.

In the cloud, service providers have effective perimeter blanket security, but typically use a combination of their own frameworks and reliance on third party tools for securing data, governing access, and complying with retention policies or mandates drilling down to the data set or even record level.

The bottom line is that enterprises require a consistent, shared experience for managing and securing their data.

Figure 1. Enterprise data cloud architecture



Source: Cloudera

Cloudera's strategy: Delivering the enterprise data cloud

Addressing agility, flexibility, under common governance

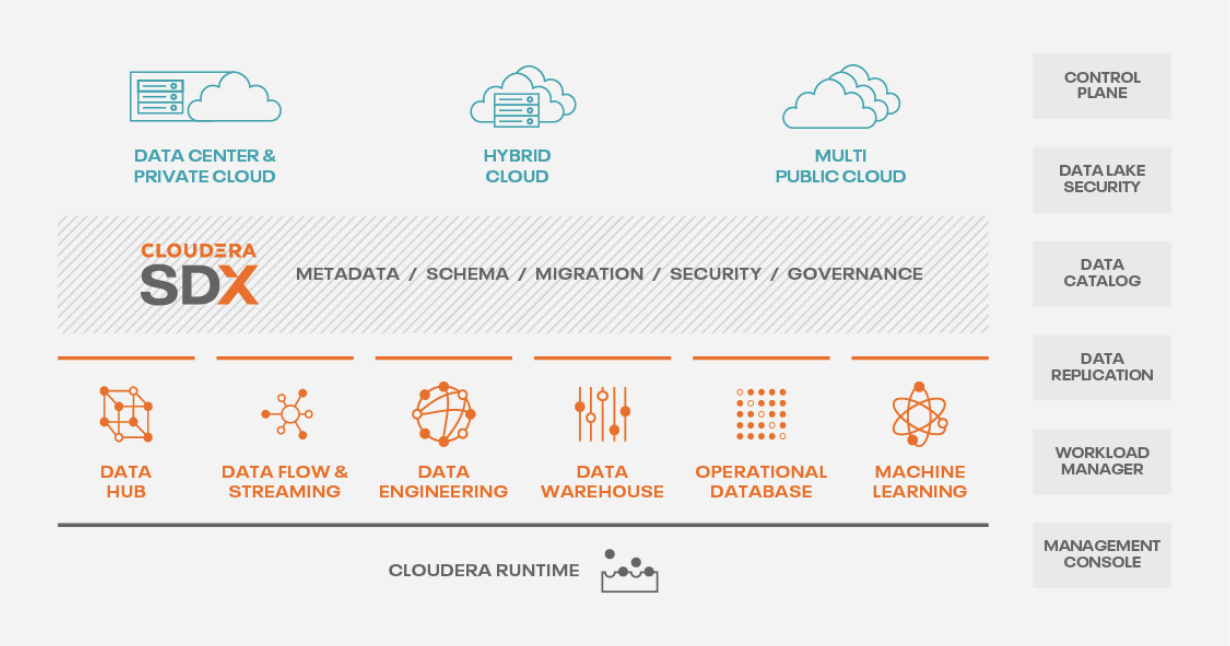
An enterprise data cloud levels the playing field between on-premises and hybrid deployment across all major public clouds and private clouds (see Figure 1). the same platform can perform ingest, processing, data warehousing, real-time, and ML with common, consistent governance and security. And it delivers on the promise of a "better cloud than cloud" experience through several key design principles that leverage:

- Flexibility and economics of cloud-native deployment without getting locked in to running on any single public cloud
- Open source and open APIs to prevent vendor lock-in, maintaining interoperability with the vast ecosystem of data service providers
- Tenant isolation that prevents noisy neighbors from draining resources and delivers on the promise of cloud agility and scale

How the Cloudera Data Platform delivers an enterprise data cloud

As noted earlier, enterprises rely on data from a variety of sources inside and outside the data center, much of it at the edge. Critical business initiatives such as customer retention, supply-chain optimization, and cybersecurity must span all of the places enterprise generate and consume data and expose their business processes. Siloed point products, such as using a dedicated IoT service and a standalone deep learning service, that are also secured and governed in their own manner, may deliver inconsistent results and create security and governance gaps. A key requirement of the enterprise data cloud is that it spans, as Cloudera terms it, "from edge to AI."

Figure 2. Cloudera Data Platform



Source: Cloudera

The Cloudera Data Platform (CDP) was designed to meet these requirements and more. CDP delivers a unified experience that:

- Runs all analytic workloads, ranging from IoT to real-time streaming, analytic query, operational database, and data science/machine learning, enabling Cloudera's vision of "edge to AI" on a unified platform
- Governs through a Shared Data Experience (SDX) common access security and control of data
- Provides a consistent experience and control plane, extending from on-premises to private, hybrid, or public cloud
- Is 100% open source, based on the highly familiar OSI-approved the Apache License, Version 2 and the GNU Affero General Public License, Version 3 ("AGPL").
- Is delivered in multiple form factors: as a suite of cloud-native services in the public cloud, a software-based platform or as a private cloud platform in a customer's on-premises data center

Supports all analytic workloads

In its initial release, CDP public cloud supports the following services:

- Data Hub – A service that creates custom-configured virtual private clusters in the public cloud(s) such as AWS, Azure, and/or GCP. The developer chooses the specific runtime services (e.g., HBase, Impala, Kafka, Hive, Spark, and so on) and then, through the management console, launches the configured cluster.
- Data Warehousing – This is a self-service cloud-based data warehouse offering that can be used for several use cases such as enterprise data warehouse (EDW) optimization that complements your existing analytic environment; as an operational database that allows analytics on large volumes of variable structured data; and as a research and discovery data warehouse targeted at exploratory analytics.
- Machine Learning - This service creates self-service machine learning workspaces and the underlying compute clusters for teams of data scientists.

On the horizon, Cloudera will be adding more services to CDP that will support use cases such as data flow and streaming (for IoT), data engineering, and operational database.

Provides a consistent experience on-premises and in the cloud

CDP can accommodate existing Cloudera customers with workloads on-premises while providing a clear path to the cloud that minimizes risk and disruption. Using a common administration console, CDP allows you to manage your deployment – regardless of whether it is running inside the data center and/or in the cloud from the same pane of glass. From this console, you can choose specific services and monitor their operation. The home page provides direct access to control functions for managing operation; a data catalog for locating data; a workload manager that manages resources; and a replication manager for moving data.

A key design principle of CDP is fostering the sharing of data. The Shared Data Experience (SDX) is a superset of capabilities designed to enable data sharing, involving several supporting services including the:

- Data Lake Service for creating and securing data lakes that are stored in HDFS on-premises or object storage in the cloud

- Data Catalog, enabling data stewards to curate data with search, versioning, and organizational capabilities
- Replication Manager, which can move (and will soon support snapshots) data between different data center clusters and/or clouds
- Workload Manager, which monitors and optimizes all workloads through improving query performance, and manages data migration and cloud bursting

These services operate on multiple form factors, from traditional "bare metal" clusters in the data center to private clouds that are managed by customers and public clouds that are managed through Cloudera's Platform-as-a-Service (PaaS). Initially the CDP public cloud service supports AWS public cloud, with Microsoft Azure and Google Cloud Platform on the near-term roadmap. Additionally, CDP supports hybrid cloud deployment through the virtual private clusters described above that can be deployed on the private or public cloud of choice.

Takeaways

With the expansion of the data canvas, the pull of the cloud, and heightened awareness and need for security and governance, enterprises are seeking data platforms that keep their deployment options open with the assurance that they are operating in a secure environment that also is in compliance with current and emerging regulatory mandates for data protection and data privacy. Cloudera has broadened its vision to focus on the enterprise data cloud that is designed to enable its customers to run a full set of analytics to get insights from large scale, complex data, regardless of where it lives, and where it should be processed.

Cloudera supports the vision of running "modern analytic workloads" on a 100% open source platform. CDP provides a consistent runtime, management, and control plane that supports enterprises as they keep their options open for deploying on-premises and/or in public clouds. It supports a diverse range of analytic workloads on data wherever it lives, from the edge to the data center and cloud, and is offering a range of preconfigured services enabling users to quickly and seamlessly spin up data warehousing and machine learning workloads on demand. It addresses the need from customers to take advantage of the elasticity, resilience, and scale of the cloud environment without cloud platform lock-in. In a landscape where there are numerous alternatives to running analytic workloads on-premises or in the cloud, Cloudera delivers a unified experience that supports the broad spectrum of analytic and operational workloads that enterprises require.

Author

Tony Baer, Principal, dbInsight

tony@dbinsight.io

Twitter @TonyBaer

About dbInsight

dbInsight LLC provides an independent view on the database and analytics technology ecosystem. dbInsight publishes independent research, and from our research, distills insights to help data and analytics technology providers understand their competitive positioning and sharpen their message.

Tony Baer, the founder and principal of dbInsight, is a recognized industry expert on data-driven transformation. *Analytics Insight* named him one of the 2019 Top 100 Artificial Intelligence and Big Data Influencers. His combined expertise in both legacy database technologies and emerging cloud and analytics technologies shapes how technology providers go to market in an industry undergoing significant transformation. His regular ZDnet *"Big on Data"* posts are read 25,000 – 30,000 times monthly.