

Realizing the Promise of the Cloud with a
**FULLY INTEGRATED
ANALYTICS
PLATFORM**



Authored by



Sponsored by

CLOUDBERA



database
TRENDS AND APPLICATIONS

Realizing the Promise of the Cloud with a
**FULLY INTEGRATED
ANALYTICS
PLATFORM**



TABLE OF CONTENTS

Introduction **1**

Meeting the needs of all types of cloud users in a fast, consistent and controlled way **2**

Giving departments the analytics resources they need so users don't "go rogue" **3**

Broken cloud promises regarding spiraling costs, inconsistent performance, and unmet flexibility **4**

Data lifecycle use cases require flexibility, security, governance, and performance **6**

Conclusion **7**

Realizing the Promise of the Cloud with a Fully Integrated Analytics Platform

Authored by John O'Brien, Principal Advisor and CEO, Radiant Advisors
Sponsored by Cloudera

Cloud platforms offer much more potential for companies to make an impact on business analytics than most realize. The original allure for enterprises migrating to the cloud is the fundamental promise of elastic computing, affordable scalability, and managed platforms and services. However, the real business potential comes from agility and enablement for enterprise data management and analytics.

The cloud facilitates more business analytics to be self-sufficient and agile, which frequently introduces more complexities than bargained for. Efforts for enterprise data infrastructure management, data management, and governance multiply to maintain consistency, efficiency, and cost controls. This is further compounded by the reality that most organizations evolve into a combination of hybrid and multi-cloud environments in which analytics projects are delivered. The challenge for a data and analytics architecture is to find a way to deliver the benefits of centralized management of critical components along with independent agility, scalability, and flexibility. This scenario commonly leads to a point where data and analytics progress becomes paralyzed by operations support, data redundancy, and general confusion.

A data and analytics platform strategy can balance the benefits of centralized management with distributed agility while enabling users with analytic capabilities to tackle business initiatives. This research paper identifies four specific challenges and provides recommendations for evolving the strategy, architecture, and enablement process for data and analytics in the enterprise.



1. Meeting the needs of all types of cloud users in a fast, consistent and controlled way

Enterprise data and analytics platforms are designed with a purpose: to enable many types of users – from business consumers and analysts to data engineers and scientists – to work with data to develop and deploy meaningful analytics. Every business initiative is multi-faceted in its data and analytics needs, which creates a broad spectrum of requirements for both the data (from raw to curated to transformed) and the maturation of analytics (from descriptive to predictive to prescriptive) that are leveraged by people with different roles in the process. Further, data architects, governance and compliance managers, and infrastructure managers each have additional specific needs and roles concerning the data in order to facilitate business analytics. A comprehensive data and analytics platform is able to align users with the appropriate data and analytics capabilities to execute business initiatives.

The first step in the modern analytics lifecycle begins with clearly identifying the business goals and initiatives for alignment to the analytics capabilities that various data and analytics workers need to achieve the business goals. Users then need to discover, access, and sift through dozens or hundreds of data sets and explore integrations and correlations – an iterative, exploratory process that can be quite time-consuming. Understanding the modern analytics lifecycle, the user journey to find, explore, and understand data and the requirements for an analytics platform helps

to optimize the steps to derive useful data and analytics in a self-sufficient manner.

A fully integrated cloud data and analytics platform supports the quest for faster analytics development, consistency in data formats, access, security, and governance by reducing the number of decisions needed with proven configurations. Having a fully integrated analytics platform will include a cloud data lake whereby data is initially ingested, formatted consistently, and curated in zones with appropriate data security and governance applied for downstream applications and data lineage. All ingested data is now part of the enterprise data repository, or data marketplace, from which governed data warehouses and data marts can be built with the analytics platform's optimized database technology. The enterprise data repository can provide self-service data prep for business analysts and data scientists with consistent formats, security, and governance. While cloud-based data lakes are great examples of this already, logical data lakes may need to resolve consistency with sensitive data remaining on-premises, and data ingestion occurring in multiple cloud environments. Consistency can be delivered in ever-increasing distributed environments through a fully integrated analytics platform.

It is essential to recognize that every corporate citizen takes responsibility for working with data. Data and analytics workers want to focus on their work without concerns about risks related to their deliverables. Being able to work confidently on analytics projects will accelerate delivery and increase transparency. When a user trusts that they are working within security protocols that



ensure the data is authorized for them to work with, they stop second-guessing whether they are allowed to incorporate or publish the data. And when data governance provides the additional context for how to properly use the data in analytics and who the data stakeholders are (such as data owners and stewards), users are confident to do their work, collaborate, and share their analyses.

Companies must realize that building analytics for the business involves creating additional new data sets and analytics that will also require the same security and governance that's inherited from its data lineage, at a minimum. A comprehensive data and analytics platform ensures a productive user experience is controlled in a way that accelerates agility and self-sufficiency in analytics development.

2. Giving departments the analytics resources they need so users don't "go rogue"

Business departments are driven to find answers and insights as quickly as possible, and departments are expected to achieve their business goals and provided budgets to do so. Public cloud environments support these needs and position themselves as a partner to empower analytics teams in the business to work the way they want to in order to quickly achieve their goals. Thus, the accessibility of cloud-based resources makes it tempting for individuals and departments to spin up their own "shadow" projects without the involvement of or approval from IT.

An enterprise data and analytics platform must maintain agility and freedom while delivering

even more value than one-off cloud options to be a compelling choice for these empowered users.

Analytics teams require that their most basic data needs are met first with the ability to find and have proper access to data, a place to explore and work with their data, and a place to publish their analytics into production. Commonly, 80% of this analytics work is basic and prioritizes agility. However, the other 20% represents specialized analytics use cases where teams need the flexibility to select the analytics services they want to incorporate. Analytics teams working on these use cases require that they can use any programming or scripting language of choice, have access to data files or SQL interfaces, scalable memory footprint in elastic clusters for highly iterative processing, and access to the preferred analytics models and services – regardless of which cloud they are in. Analytics teams shouldn't be restricted to work in a particular cloud environment because of its available analytics services and shouldn't have to create data flows to move data to that cloud environment. Analytics workbenches address this need. They follow the same paradigm of data science notebooks, where a series of data prep and analytics steps are executed through API calls for analytics models both internally developed or available as a cloud service. Ideally, an analytics team can leverage a fully integrated analytics platform, including analytics workbenches, to work as they choose.

An effective way to address the "rogue" work of analytics teams is not to work against them by presenting barriers and restrictions but rather to understand their needs and provide them with a compelling alternative way to work. When teams realize that they can work faster



and more independently on a fully integrated analytics platform that maintains governance and compliance, they will do so. Once the platform is viewed as a conduit for working quickly and efficiently, with access to their preferred models and services, other teams and departments will adopt the same data and analytics platform.

The goal of the enterprise data and analytics platform is twofold. First, the platform must be the compelling preference for analytics teams. Keep in mind that the user experience must match or beat the options available of the cloud platform alone. This means that the user finds it easier to perform the analytics they want within the analytics platform itself in every step of their process. Second, the analytics platform must transparently deliver the security and governance that the analytics teams also desire to have in their data and analytics work. Data that is distributed across environments with inconsistent security and governance complicates their work and typically leads to duplicating work efforts as well as data.

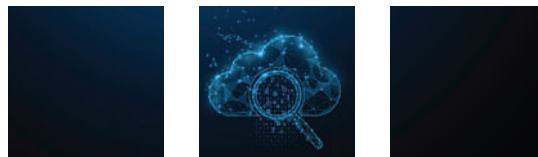
3. Broken cloud promises regarding spiraling costs, inconsistent performance, and unmet flexibility

As companies migrate their systems to a chosen cloud provider, the cloud promised to eliminate the frustrations of their current data centers while doing it at a lower cost to the company. Many cloud success stories featured smaller tech startups that grew exponentially or enterprises that realized benefits starting with dev/test environments, analytic projects, and predictable operational workloads. Yet these stories did

not show that unleashing the potential of cloud benefits to everyone in the enterprise with low barriers to entry would result in unpredictable costs from inexperience operating cloud services, performance inconsistency between on-premises data centers and cloud data centers, and nuances of flexibility with managed services.

Cost Management

Data center cost management and strategy is a centralized IT function. The transition to pay-as-you-go cloud infrastructure is inherently different from the on-premises investment infrastructure model that pays fixed costs up-front for new hardware on a three to five-year depreciation schedule and accounts for fixed monthly costs in power and real estate. Companies and their IT departments need to embrace the new cost management paradigm; however, they still have the responsibility for planning and budgeting on a cloud platform where costs are going to be hindsight and reactive. When every application or analytics team has the ability to launch the services, cost management and optimization becomes their responsibility – which they may or may not have the skills and experience to handle because they previously relied on IT. When centralized usage management and, therefore, cost management is available, cost models are developed from similar applications or projects and optimizations can be learned and implemented. A centralized operating plane, or platform, is ideal for overall cost management and optimization. A cloud service provider will have their own paradigm and the company's data center may have their own (albeit a



different) paradigm, but cost management and optimization over multiple data centers and cloud vendors is still a challenge.

System Performance and SLAs

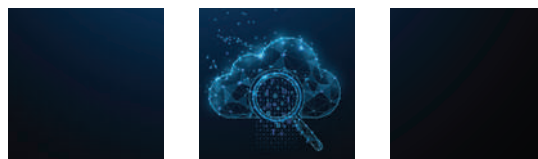
Managing system performance and service level agreements to the business is another challenge that companies face in understanding how to work in the cloud. Analytics require data processing, database and analytics engines, and data visualization that operate with dedicated servers in the on-premises data centers or in virtualized data center environments. In this scenario, the virtual machines abstracted the CPUs, memory, and storage at a sacrifice for the hypervisor's own compute and storage needs. In some cases, there were low-level compatibility issues translating the operations of the hosted operating system through the base operating system or bare-metal hypervisor to communicate with the hardware. In the cloud, there are similar options for Infrastructure as a Service (IaaS), but the cloud can further abstract the infrastructure management with Platform as a Service (PaaS), and Software as a Service (managed service). With further abstraction of virtual CPUs (vCPUs), memory, storage, and networking configurations, no two physical data centers are alike. A company's on-premises data center was engineered, and applications were "burn-in" tested for setting performance benchmarks. Cloud vendors provide many instances of configurations with combinations of compute and storage to simplify this process as much as possible for those choosing to replicate their on-premises virtual machines. Keep in mind

that IaaS may appear identical, but its real-world performance is likely to vary from one physical data center to another. For analytics that further leverage cloud benefits with PaaS and SaaS configurations on the cloud, the performance is likely to be more inconsistent due to the transition from one product to another, such as an installed mature RDBMS to a PaaS database as a service or SQL as a service.

Ideally, analytics teams will achieve more consistent performance with a common technology platform that's both on-premises and across their various cloud data centers. This removes the biggest variables in maintaining consistent performance, leaving only slight variations due to physical infrastructure. Cloud instance configurations are likely to be consistent with those available as virtual machines in the on-premises data centers. While some cloud PaaS databases have promised elastic scalability, the reality has proven to deliver quite mixed results. Unless the underlying PaaS database technology is based on cloud-native architecture with clear separation of compute and storage, when adding virtual instances it is possible but not easy to redistribute data. Cloud-native architecture and its similarity with the Hadoop/YARN architecture will increase the possibility of smoothly scaling compute and storage independently in the cloud.

Flexibility

Flexibility is another promise of the cloud where reality reveals many pros and cons. Public cloud platforms offer hundreds of services to choose from, spanning every area of data and analytics needs. Sometimes there are even



multiple cloud services to choose from that can perform the same function; architecture patterns or solution architectures then play a role. Further, the IaaS option to BYOL (bring your own license) allows even more options for analytics teams. There's clearly flexibility with these options, but this manifests as a challenge as too much flexibility is left in the hands of self-sufficient analytics teams without overall guidance for consistency and manageability.

Flexibility in the analytics platform should allow teams to leverage the cloud analytics services and data they need regardless of which cloud environment it's in. Thus, they are able to capitalize on the technologies and services provided across multiple cloud providers without hesitation or complexity. Naturally, cloud providers will focus on their own services and integration, but enterprise data and analytics platforms need to have strength as a logical architecture that is not restricted by physical implementations. This is not easy to accomplish, but it delivers data and analytics to the entire enterprise rather than dealing with a combination of pointed solution architectures.

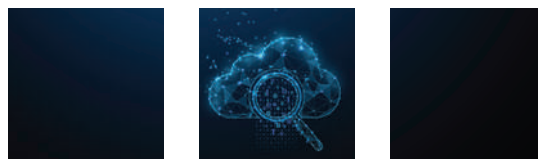
Avoiding the broken promises of cloud computing comes down to seeing beyond the allure of low barriers to pay-as-you-go, agility of on-demand services, and their elastic scalability. When you understand what analytics teams need, you are able to recognize the responsibility that managing the enterprise analytics platform carries. A fully integrated analytics platform that is independent of cloud environments will centralize operations and optimizations to allow

for forecasted costs, deliver more consistent performance, and guide options for flexibility.

4. Data lifecycle use cases require flexibility, security, governance, and performance

The data lifecycle truly does encompass data from its birth or origin to its death or purging, and there are clear business use cases for the data throughout. Within moments after data is created, automated reactions are carried out from human-based rules or machine learning algorithms – such as fraud detection or next-best actions. That data can be integrated and aggregated to drive near real-time streaming applications and dashboards throughout the company to watch for trends and thresholds in operations. The data in its raw format can be curated into data lakes where business analysts and data scientists explore the data through data prep for new insights and testing their own business hypotheses.

The value of data shifts throughout its lifecycle from individual events and change data capture records to aggregate data sets for business value in seeking patterns, clusters, and trends over time that feed management reporting, OLAP, and machine learning training data. Eventually, data reaches the end of its useful life through compliance rules that require full deletion and purging (or the opposite: archived forever). More often than realized, the world around us changes and the data context is no longer relevant for any use case throughout the entire data lifecycle. Even more likely is that the application that



created the data is evolving and the data drifts in structure and context.

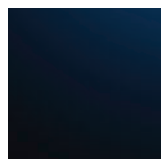
The challenge for companies comes from maintaining data management functions for proper security and governance across analytic use cases that require different technologies. Even though there are tools that can abstract the underlying data engines and locations to centralize the data catalog, data prep, or federated query engines, each of those tools must also share consistent data security and governance. The same is true for the tools that can abstract performance monitoring and alerts. These tools must share common data management functions for consistent security, authentication (single sign-on), and data governance regardless of the use case.

A fully integrated data and analytics platform that works with multiple cloud platforms address not only data lifecycle use cases but also how IT is required to facilitate data management functions and operations. The ability to operate in multiple cloud environments requires a unified security and governance experience for users that is compatible with each cloud's implementation. Enterprise analytics platforms that are fully integrated can operate on a level abstracted above the on premises, private and public cloud data centers, providing the use case flexibility while avoiding serious "cloud vendor lock-in" with an open and portable architecture.

The measure of completeness of an analytics platform is its ability to deliver as many analytics use cases, data management functions, and platform management with a common user experience of terminology and



The measure of completeness of an analytics platform is its ability to deliver as many analytics use cases, data management functions, and platform management with a common user experience of terminology and usability regardless of the underlying hybrid or multi-cloud configuration.




usability regardless of the underlying hybrid or multi-cloud configuration.

The complexity of integrating IT ecosystems to perform reliably in hybrid cloud environments has fueled the next generation of technologies. Cloud-native architecture is designed to allow applications, computing, and storage to operate ubiquitously. Technologies, such as Kubernetes, Kafka, Spark, and Presto, also give rise to an enterprise data and analytics platform that is optimized to perform across distributed cloud platforms. All aspects of performance and scalability also require a consistent user experience to avoid the inefficiencies and complexity risks that inherently exist with each cloud providers' service management and terminology.

Conclusion

The benefits of cloud computing remain despite the initial realities as companies trying to achieve them as part of their data and analytics evolution. In actuality, companies are trying to achieve the bigger promise of business analytics that requires cloud computing to be interwoven for a fully integrated data and analytics platform. Companies should look beyond the functionalities of cloud computing to focus on delivering the analytics capabilities needed by all types of users with improved cost, performance management, and unified enterprise security and governance. Learning from the past, it's become clear that platforms and standards that can operate ubiquitously across on-premises and cloud environments are the future.



About Cloudera:

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises. Offering analytics for the full data lifecycle, Cloudera Data Platform, Cloudera's fully integrated analytics platform, powers data-driven decision making by easily, quickly, and safely connecting and securing the entire data lifecycle on all clouds and data centers. Discover Cloudera Data Platform and take it for a test drive.

About Radiant Advisors:

Radiant Advisors is a strategic research and advisory firm that delivers innovative, experience-based research and thought-leadership to transform today's organizations into tomorrow's data-driven industry leaders. To learn more, visit www.RadiantAdvisors.com

