Ovum
TMT intelligence | informa

Harnessing the Cloud to Deliver Big Data Insights

An Ovum white paper for Cloudera and Microsoft

Publication Date: June 15, 2018

Author: Tony Baer

Summary

Catalyst

Becoming data-driven is now a matter of urgency for organizations to stay competitive with challenges ranging from customer engagement to risk management, cybersecurity, and fraud detection, and to operational excellence. The cloud is playing a growing role in helping enterprises benefit from data as it bypasses many of the cost and organizational bottlenecks associated with implementing new capacity in on-premises data centers. The results are borne out with big data. Big data workloads are moving to the cloud. According to Ovum, 27.5% of big data workloads are running in the cloud, with the rate currently growing by over 20% annually. Ovum predicts that the cloud will account for over half of *new* big data workloads by 2019. For most organizations, hybrid deployment spanning the data center and cloud will become the new reality. While few are likely to migrate 100% of their data, applications, or platforms to the cloud, the cloud will play a big role in analytics and business agility – going beyond test/development to quickly on-ramping new workloads. But not all cloud services are alike. How can enterprises choose the right cloud infrastructure service (IaaS) provider *and* the right big data platform to run in that cloud environment? These should be considered separate decisions.

Ovum view

Becoming data-driven requires the ability to serve all stakeholders who work with data. Each of them has varying requirements. Data scientists require an environment that allows them to easily develop models and scale them to run against all the data on the cluster, not just a sample on the laptop. Business analysts require an environment that empowers self-service analytics. Data engineers require a cost-efficient, high-performance environment that allows them to cleanse, integrate, and transform data. However, building dedicated systems to address each of these constituencies will generate new data silos that in the long run will not be sustainable. Enterprises require the flexibility to support each stakeholder group in an environment where not only data, but also governance and security policies are shared and enforced comprehensively and consistently. Many cloud platforms offer the breadth of services, but often only with perimeter security provisions that do not provide integrated governance across data platforms. Ovum believes that a *unified portfolio* of *managed cloud* services that are optimized for a range of workloads, and play well with hybrid, on-premises data centers will play a big role in accelerating how organizations can on-ramp to support each of these constituencies from a common foundation.

Key messages

- Enterprises require fast, cost-efficient on-ramps for addressing the familiar challenges of engaging customers, reducing risk, and improving operational excellence.
- The cloud is playing a key role in accelerating time to benefit for gaining new insights.
- Managed cloud services that automate provisioning, operation, and patching will be critical for enterprises to leverage the full promise of the cloud when it comes to time to value and agility.
- Cloudera and Microsoft's partnership shows the benefits when a managed big data platform-as-a-service (PaaS) is optimized for the underlying cloud platform – in this case, delivering the cost and agility benefits of the cloud with scalable cloud storage that is optimized for the

complex compute problems involved with big data analytics, data engineering, and machine learning.

Managed cloud services deliver on the promise of the cloud

The business and organizational challenges

The business challenges facing enterprises are familiar. They often encompass

- driving customer insights from harvesting their digital footprints
- connecting products and services by leveraging data and compute at the edge with IoT
- guarding businesses from cyberthreats and fraud.

The *challenge* is that addressing all of these problems requires supporting multiple stakeholder groups/roles in the organization. It requires data scientists, who experiment with data and models and who need a productive environment. The environment should provide a full choice of languages, tools, and frameworks for generating models and testing them; the environment should also enable data scientists to scale up their models from laptops to execution on the cluster. Then there is the need to marshal the data. That is where data engineers are needed; they require a cost-efficient, scalable environment to transform data. The circle is not complete without stakeholders from the business, as they are the entities that own these problems and are ultimately responsible for signing off on the solutions. Business analysts expect to be empowered with a productive environment that is designed for self-service, allowing them to use the business intelligence (BI) tools with which they are familiar.

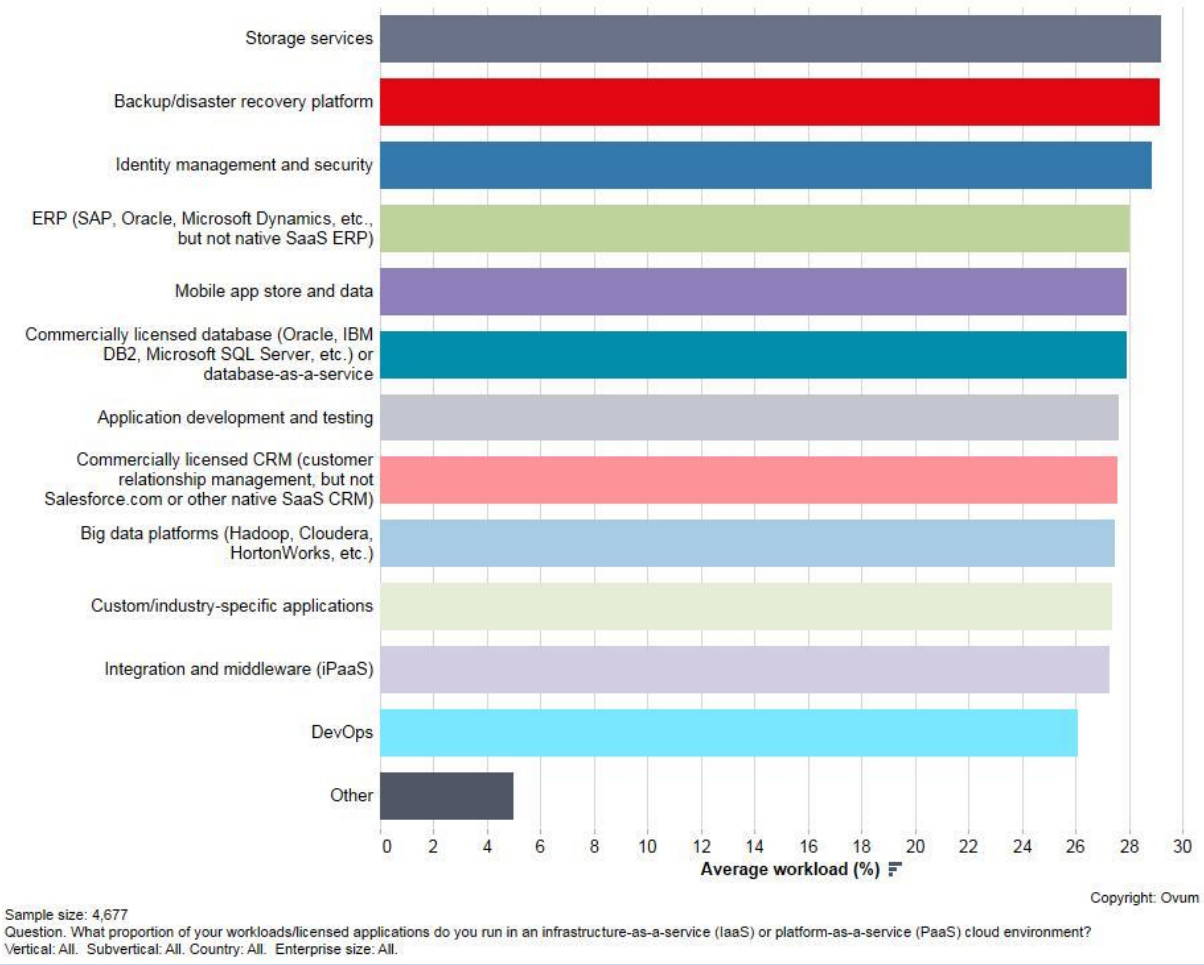
The *opportunity* is serving all of these constituencies. But getting there is often easier said than done because the workloads that data scientists, data engineers, and business analysts run are likely to be different. They will require different configurations of clusters. For instance:

- Data scientists expect a friction-free experience, allowing them to design experiments through notebooks, and automation for scaling up and deploying their models from laptop to cluster. They may also require clusters with heavy or dense compute and, in some cases, specialized hardware such as GPUs to process artificial intelligence models using machine learning or deep learning.
- Data engineers are seeking to modernize the process of transforming data, leveraging commodity infrastructure, open source software, and optimized open source compute engines such as Spark. Their compute needs are very IOPS-driven.
- Business analysts require an efficient environment for interactive query to enable self-service on big data. For organizations where the business end-user population is sizable, they may require clusters optimized for higher concurrency.

With all the varying needs by stakeholder group, there are several important common threads. IT and the business require a unified, secure, consistently governed environment that protects against cyberthreats, enables them to safeguard data privacy, and meets emerging regulatory mandates such as the General Data Protection Regulation (GDPR) that has just gone into force for organizations that

operate in the EU community. They cannot afford to spend their time managing and governing multiple data silos each with its own varied capabilities (and limitations).

Figure 1: Proportion of workloads running in the cloud



Source: Ovum ICT Enterprise Insights

Why use the cloud?

Enterprises are embracing the cloud. According to Ovum's enterprise insights research, roughly 25–30% of enterprise workloads are currently running in the cloud; of that, big data is in the middle at 27.5% (see Figure 1). For *new* big data workloads, Ovum expects that by next year, cloud will become the consensus choice, accounting for over half of them. The cloud helps enterprises overcome many of the hurdles associated with big data deployment, starting with investment and lead time for IT teams to evaluate, procure, install, and integrate new computing capacity. By enabling organizations to bypass the hurdles of on-premises procurement and deployment, the cloud allows them to become more agile. For customers, adding new capabilities – such as for performing machine learning, modernizing ETL, and opening self-service – is a matter of reserving cloud capacity.

In turn, deployments that use a cloud-native architecture that abstracts storage from compute provide additional advantages. Customers with workloads that are highly variable and/or are prone to sudden spikes can take advantage of the cloud's *elasticity*, ramping compute up and down on an on-demand

basis. This allows considerable flexibility: instead of buying capacity to handle anticipated workloads, customers only pay for the compute capacity that they need.

It will be a hybrid world

At most organizations, on-premises deployment will not disappear. For instance, long-running/persistent operational data warehouses that deliver scheduled reporting often continue to run on premises, while new applications or transient workloads may be more easily started in the cloud.

There will not be any single silver bullet formula or recipe regarding which workloads run on premises or in the cloud. Each organization will have different criteria for what runs where, such as internal policies or external mandates regarding where data is physically stored; availability of a specific application or system in the cloud; available capacity for running a specific workload; government or industry regulations and best practices; and/or other criteria. Enterprises should keep their options open regarding where they run their workloads, and their cloud provider should support such hybrid operation to run transparently and uniformly across environments.

Types of cloud services

There is no such thing as a generic cloud service – the cloud provides a wide array of options.

Infrastructure-as-a-service (IaaS) and software-as-a-service (SaaS)

IaaS is the most basic service. With IaaS, your organization subscribes to raw compute and storage infrastructure that can be paid for either through long-term contracts or on a pay-as-you-go basis. With IaaS, IT organizations still have the same responsibilities for supervising operations as they would with their own data centers – selecting, provisioning, and managing infrastructure, along with installing, updating, and patching software.

Conversely, SaaS is where an application is hosted and managed by the solution provider in the cloud. The customer subscribes to the application and does not manage infrastructure or deployment – the application provider handles that.

Managed platform-as-a-service (PaaS)

PaaS is the other major category of cloud service. Ovum classifies PaaS services as "managed" cloud services, because here, the *platform* provider actively manages the cloud service with a prescriptive approach. This prescriptive approach delivers preconfigured packaged cloud services that reduce or eliminate the need for customers to specify and manage infrastructure and software. Platform providers offering managed cloud services can include databases, application development environments, integration services, or other tools.

Managed cloud services promise the simplest and fastest on-ramps

Ovum believes that PaaS cloud services will become the default choice for the next wave of big data adopters. Our belief is grounded in the fact that these services allow the customer to focus strictly on data and analytics, thereby speeding time to benefit. This will be especially true for new customers who have modest (or no) experience with big data analytics.

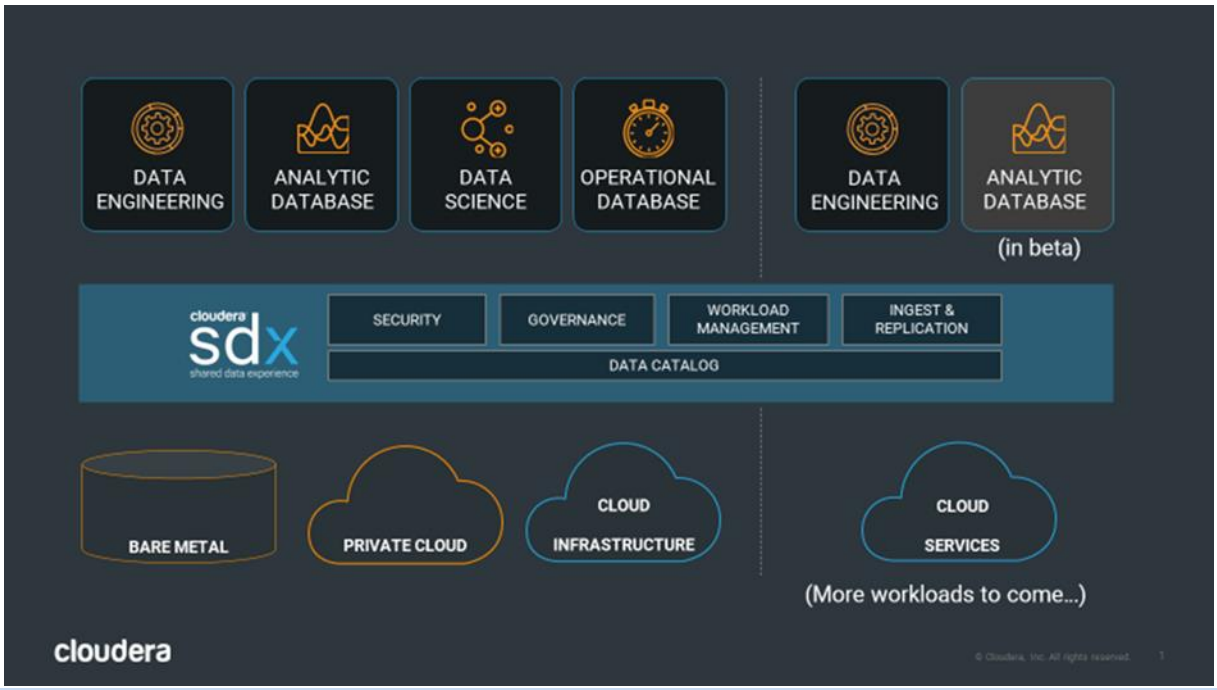
Increasingly, PaaS services are being packaged around the workloads of choice, providing services that are optimized for the user/role and workload type. This eliminates the need to configure cloud instances for the varying processing and storage requirements for use cases such as self-service analytics, AI and data science modeling, and data transformation.

How Cloudera and Microsoft are partnering to meet customer cloud requirements

The Cloudera platform is designed around choice

Designed as a machine learning and analytics platform, Cloudera Enterprise is designed to provide customers full choice regarding functionality and deployment. It can be deployed on premises as single-tenant “bare metal,” in a private cloud, or in the public cloud of choice. And with a choice of editions, Cloudera customers can get the configuration that best suits their business/use case requirements. It is available as an enterprise edition designed to support multiple types of workloads, and there are specially configured editions tailored to data engineering, data science, and operational database. All editions of Cloudera Enterprise are managed and governed through a common Shared Data Experience (SDX) that ensures consistent security, governance, workload management, and ingest and replication. All of the functions and services governed by SDX are driven through a shared data catalog.

Figure 2: Cloudera deployment options



Source: Cloudera

PaaS services on Azure that are tailored to your workloads

Cloudera and Microsoft have partnered to optimize deployment of all Cloudera editions on the Azure cloud through an elastic architecture that separates compute from storage. This allows Cloudera customers to fully take advantage of the economics of cloud compute.

Cloudera Enterprise on Azure: The broad-based platform for spanning multiple workloads

Cloudera Enterprise is available on the Azure cloud. It provides over 100 services, a broad selection of end-to-end tools, and financially backed service-level agreements (SLAs) with over 99% uptime guarantees. Available in the Azure Marketplace, Cloudera Enterprise on Azure can be deployed through a single mouse click. This eliminates the lead time associated with procuring and installing new nodes in an on-premises data center.

Cloudera Enterprise on Azure supports a wide variety of workload types such as Impala, HBase, Spark, and Solr. It integrates with Microsoft data and analytic platforms, querying structured and unstructured data from SQL Server via PolyBase and leveraging Impala to open access to business analysts through Power BI.

Cloudera Altus on Azure: The jointly engineered solution tailored for specific workloads

The Cloudera/Microsoft partnership has taken the next step through support of Cloudera Altus, Cloudera's tight-knit family of managed PaaS services. The Altus release has been jointly engineered by Cloudera and Microsoft to be optimized for Azure. It provides a cloud-native managed service that keeps computing elastic, which maximizes the value for customers with transient or highly variable workload patterns.

Under the hood, Cloudera Altus on Azure utilizes Azure Data Lake Store (ADLS), an enterprise-wide hyper-scale repository for big data analytics workloads. Azure Data Lake enables you to capture data of any size, type, and ingestion speed in one single place for operational and exploratory analytics. It provides very strong durability, which is especially critical for complex, iterative compute workloads – such as data engineering – that use Spark.

Cloudera Altus for Data Engineering is the first Altus service to be released on the Microsoft Azure cloud. Cloudera Data Engineering is a foundational workload for the development and operation of data pipelines for transforming data and training machine learning models. As such, data engineering can be considered the launchpad for machine learning and big data analytics. Operating in single-tenant mode, Altus leverages the SDX governance, including the shared data catalog that is core to the Cloudera platform. Ovum expects that Cloudera will expand the Altus portfolio in the Azure cloud, and as they do, they will all leverage the common SDX governance later.

Takeaways

Enterprises are embracing the cloud – and especially so for big data analytics, data science, and machine learning workloads. The cloud is playing a key role in accelerating time to benefit for gaining new insights. Enterprises require fast, cost-efficient on-ramps for addressing the critical business challenges. Managed cloud services will be critical for enterprises to leverage the full promise of the cloud when it comes to time to value and agility. Cloudera and Microsoft's partnership in delivering

Altus services on Azure, optimizing for Azure infrastructure, and delivering integrations to Microsoft's data platforms and BI analytics stack provides a good 1 + 1 = 3 case in point for delivering time to value for customers

Appendix

Author

Tony Baer, Principal Analyst, Information Management

tony.baer@ovum.com

Ovum Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Ovum's consulting team may be able to help you. For more information about Ovum's consulting capabilities, please contact us directly at consulting@ovum.com.

Copyright notice and disclaimer

CONTACT US

www.ovum.com
analystsupport@ovum.com

INTERNATIONAL OFFICES

Beijing
Dubai
Hong Kong
Hyderabad
Johannesburg
London
Melbourne
New York
San Francisco
Sao Paulo
Tokyo