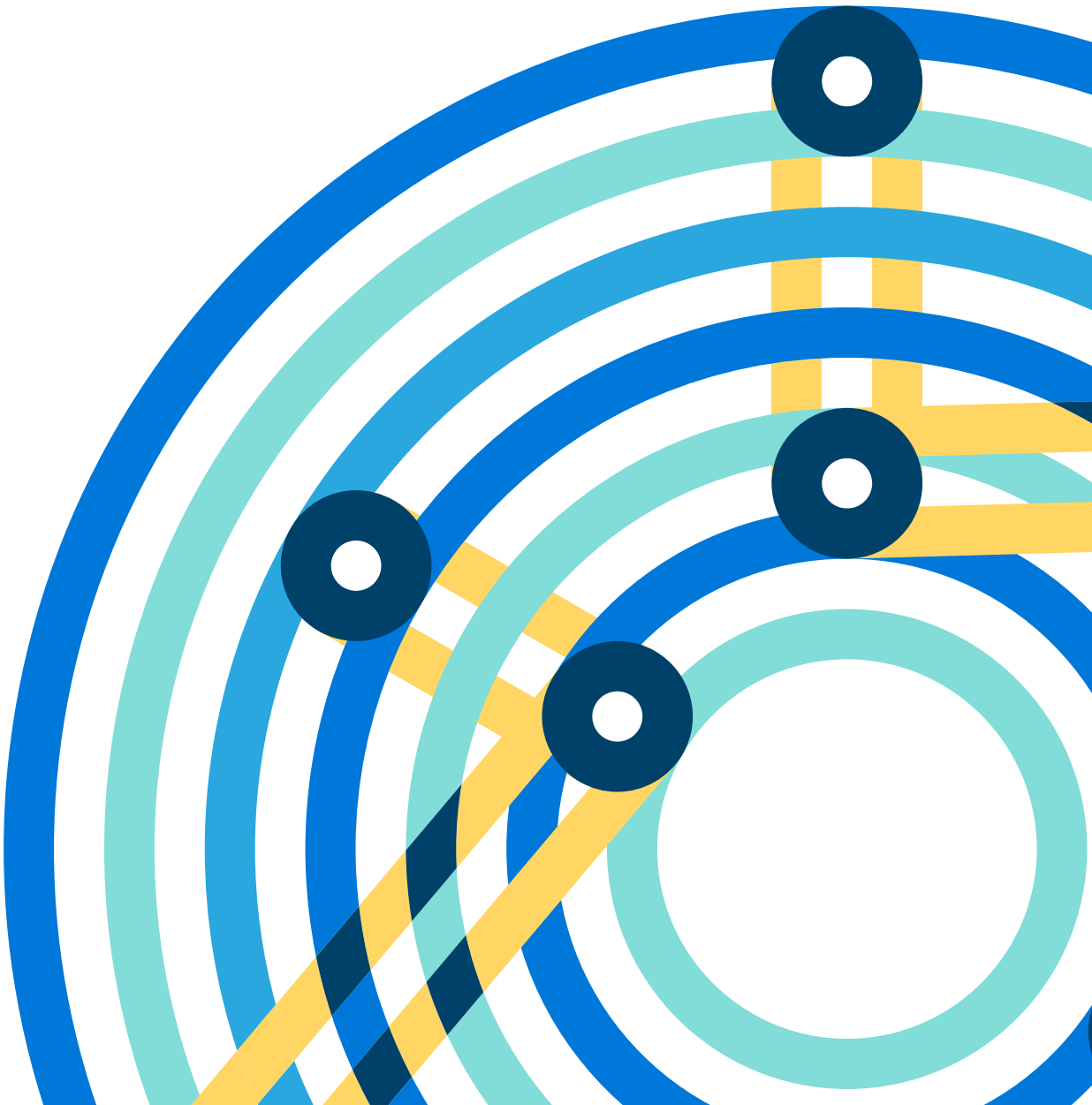
cloudera®

WHITE PAPER

A Modern Data Platform for the Cloud



Version: 101

cloudera®

Introduction

If your company is considering migrating traditional IT workloads to the cloud or instituting a "cloud first" policy, you're not alone. Cloud adoption is surging in the enterprise and far outpacing legacy x86 purchases. IDC predicts cloud infrastructure spending will grow at a compound annual growth rate (CAGR) of 15.1 percent through 2019[1]. And Gartner suggests that by that same year, more than 30 percent of software investments from the 100 largest companies will have shifted from cloud-first to cloud-only.[2]

The following whitepaper will explore common application patterns companies are employing to get more value from their data in the cloud.

Challenges

Over the past decade, public cloud infrastructure has fundamentally changed the way companies think about IT. What was once rigid and permanent can now be elastic and transient. The up-front cost of data center hardware has been replaced by the on-demand, pay-as-you-consume variability of cloud instances and storage.

There are many advantages of cloud computing, but this approach brings with it myriad questions. In particular, which workloads are best suited for cloud infrastructure and what cloud attributes are important for your business?

76% of companies will embrace hybrid cloud

– Gartner, Market Trends: Cloud Adoption Trends Favor Public Cloud With a Hybrid Twist 2015

82% of enterprises will have a multi-cloud strategy

– RightScale 2016 State of the Cloud Report

Elasticity—Quickly provision, grow, shrink, and de-provision resources as needed to meet peak demand, run off-scheduled jobs, and optimize for cost savings.

Portability—Hybrid and multi-cloud environments are now commonplace in the enterprise. Deploying applications that can be easily ported across heterogeneous architectures mitigates the need for costly migrations and employee retraining, while reducing the risk of vendor lock-in.

Deployment variability—Not all workloads are equal. Batch workloads can take advantage of cloud transience to save on cost while accessing native cloud object storage. Streaming and always-on workloads can benefit from persistent clusters with high availability, disaster recovery, and other enterprise features.

Object store support—As more data is pulled into the cloud, and object stores in particular, the application should be able to read to and write from these data stores without rigid modeling or ETL to another storage location. Furthermore, the applications should emphasize strong security, governance, and performance.

Enterprise-grade operations and security—Security, governance, and operations in the cloud must extend beyond the firewall to the application and data itself.

Today, companies are harnessing cloud data to deliver tremendous insights back to the business, to become data-driven and to digitally transform. With big data processing and analytics on cloud infrastructure, organizations are able to look at massive volumes of data from disparate sources and use this information to drive customer insights, improve products and services efficiency, reduce risk, and modernize their IT environments.

A modern data platform in the cloud

As the pace of business escalates and data growth continues along its wild trajectory, more and more data-driven workloads are migrating to the cloud so companies can take advantage of the increased flexibility and capability as well as reap cost benefits. However, legacy data management systems—even those developed to run in the cloud—are simply too rigid, too slow, and too limited in scope to handle the demands of the modern enterprise. Cloudera Enterprise is a modern data platform, powered by Apache Hadoop, that enables organizations to ingest and store limitless amounts of data and run a variety of processing and analytics workloads against that unified set of data. From powering data engineering and data science workloads to building an operational or analytic database, Cloudera can help companies get the most value from their data whether it lives in the cloud, on-premises, or in a hybrid environment.

[1] IDC Forecasts Worldwide Cloud IT Infrastructure Market to Grow 24% Year Over Year in 2015, Driven by Public Cloud Datacenter Expansion, October 2015

[2] Gartner Says By 2020, a Corporate "No-Cloud" Policy Will Be as Rare as a "No-Internet" Policy Is Today, June 22, 2016

cloudera

Today, organizations are running Cloudera Enterprise on Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform, and a host of other cloud environments, and delivering tremendous results back to the business:

- A camera and lifestyle company is blending user interaction with real-time streaming and batch data from across its online and offline channels to help direct R&D and development spend.
- A retail banking analytics company is crunching hundreds of thousands of business metrics with sub-second responses, to help clients optimize marketing spend by as much as 50 percent.
- One of Europe's largest airports is analyzing sensor data to perform preventive maintenance on their escalators, moving sidewalks, and baggage carousels to reduce operational downtime.
- A media and advertising giant is off-loading data from legacy architecture to the cloud to boost agility and roll out new customer services more quickly.

These may seem like disparate use cases, but in fact, they all stem from the following common application patterns:

- **Data Engineering and Data Science**—Data processing, development, and serving of predictive models
- **Analytic Database**—ELT, reporting, exploratory business intelligence
- **Operational Database**—Low-latency storage and analytics, real-time serving at web scale, event processing, and model scoring

Cloudera makes it easy, cost-effective, and convenient to deploy these workloads on cloud infrastructure, taking advantage of cloud elasticity, low-cost storage and compute options, and rapid provisioning to deliver a modern data platform that can tackle even the most challenging business problems.

Cloud object store support

Cloud object stores are becoming increasingly popular for their resiliency, scalability, and relatively low cost. The core execution engines within Cloudera Enterprise—including Apache Spark, Hive, Hive on Spark, and Impala (incubating)—can natively read from and write to Amazon S3. This means users don't have to run time-consuming data transformations, make data copies, or deal with the latency inherent in moving data from cloud storage into a proprietary file system so that it can be processed or analyzed.

Data science can also be done directly on cloud native data, and machine learning algorithms can be developed, trained, and saved for additional flexibility. Additionally, Cloudera Navigator supports Amazon S3 with auto-classification, lineage, and audit trails, so cloud users can get a more complete picture of their data in the cloud.

Dynamic data processing and data science in the cloud

Cloudera Enterprise is a comprehensive platform for data science and engineering in the public cloud. ETL and other batch processes can run reliably as single-tenant clusters, leveraging cloud transience. Jobs can run at scheduled times and the cluster can be terminated upon completion, so companies don't have to dedicate 24x7 compute resources to processes that may only run a few hours a day a handful of times a week. In addition to the flexibility you get with the cloud, you gain the performance of access engines like Apache Spark and the ability to address batch and real-time processing in a single system.

For data science workloads, users can explore and model data that lives natively on cloud object storage with familiar native-language APIs. Organizations can leverage compute resources to perform ad hoc data discovery and then terminate those resources while the data persists.

Beyond right-sizing resource utilization, organizations can also recognize significant cost savings by paying only for the time spent on compute and processing. No more hefty up-front investments in hardware that spends most of its time sitting idle. Users can achieve even deeper savings by utilizing infrastructure at its cheapest via spot instances.

cloudera

Cloudera Enterprise allows users to elastically scale compute and storage separately and select the right instances for the specific workload, so you can deploy a single-purpose cluster with the optimal configurations. If greater performance is needed to speed processing, a user can simply add more nodes to the cluster without having to expand storage, another significant cost savings when compared to tightly coupled storage and compute offerings in the cloud.

Deploying Cloudera on the cloud

Cloudera Director is the easiest way for customers to deploy and manage Cloudera clusters on transient environments in AWS, Azure, and Google Cloud. Benefits of Cloudera Director include the following:

- Dynamic cluster lifecycle management (automated provision, grow, shrink)
- Rapid cluster spin up in 10 minutes
- Single pane of glass with multi-cluster and multi-cloud views
- Support for pay-as-you-go pricing and spot instance provisioning
- Cluster cloning and creating and saving templates
- Integration with Cloudera Manager

Cloud-native BI and analytics

Analytic database workloads are also well suited to take advantage of the benefits of the cloud. Cloudera's analytic database is designed with decoupled storage and compute layers, so customers can run high-performance SQL analytics on data wherever it may live—including the HDFS file system, the updatable Apache Kudu columnar store, and object stores like Amazon S3. This also enables greater flexibility to access and query data of any type, at a greater scale, all while avoiding the need for rigid data modeling or a lengthy data load typical of traditional monolithic analytic database architectures. Raw or prepared data is immediately available for reporting, BI, and exploratory analytics. And as part of a cloud-native, shared data platform, the value of the data and results can be extended beyond SQL to data science teams for model development or operational workloads.

This modern analytic database also uniquely supports elastic scalability, allowing customers to tap into the cost savings of the cloud. Cloudera Director makes it easy to grow or shrink the cluster to support peak access or new use cases, as needed. For instance, the cluster can scale to support higher user concurrency during prime business hours and then scale back down over the weekend for pay-per-use savings. In addition, the platform supports transient clusters to help further optimize costs. Clusters can easily be spun up to run a specific report or other analytic job and then be terminated after the job completes, all using Cloudera Director.

Operational database in the cloud

Companies that can deliver the right data at the right time enable better decisions across both manual and automated processes. Cloudera's real-time operational database is designed as a low-latency, web-scale platform for rapidly changing data. These capabilities are enhanced by the public cloud, which provides further opportunities to increase convenience, decrease costs, and grow or shrink on demand.

Cloud-enabled capabilities can make a dramatic difference in the exploration and execution of customer use cases. Fast provisioning in the cloud enables rapid testing of new software, new upgrades, and new applications; this yields faster time-to-value. Low-cost cloud storage can bring more data under management, yielding better data-driven results. Elastic growth within the public cloud can help to scale a cluster as needed, and then shrink as business demands or SLAs change, saving resources.

For "always-on" operational applications, the cloud offers the ability for you to take frequent snapshots and provides reliable, low-cost storage for backup and disaster recovery. Cloudera's platform can leverage this stored data as an active archive to ensure your operational database applications always have access to business-critical data.

cloudera®

Whether the intended use of the operational database is event processing, model scoring/serving, delivering data applications, or any other myriad of real-time uses, the public cloud can increase the versatility of your data.

50% of customers "repatriate" at least one public cloud workload back to private cloud or fixed infrastructure for cost reasons

– 451 Research: AWS Lambda: new and exciting, old and rehashed, more vendor lock-in (or all the above)?, November 22, 2016

Running in hybrid environments

While Cloudera continues to see an increase in cloud-only deployments among its customer-base, we believe the majority of customers are more likely to deploy in a hybrid IT environment where data exists both in the cloud (and increasingly often in multiple clouds) and on-premises. In these architectures, it's important for companies to ensure the applications that power their modern data strategy can work across environments.

Solutions that are inextricably tied to a specific cloud environment can lead to vendor lock-in, making it extremely difficult to pull data from the environment should you choose to do so. It also reduces the ability for a customer to negotiate infrastructure discounts or shop around for cloud vendors that may provide benefits that are more aligned with your business needs.

Cloudera Enterprise can be deployed on any cloud and any industry-standard x86 server, so customers can expect a consistent experience regardless of the environment. Customers can choose whether to pay-per-usage as they often do on ephemeral clouds or pay-per-node as is the case for fixed and on-premises environments.

Enterprise-grade security and governance

Cloudera complements and extends cloud infrastructure security, enabling customers to work with production data with confidence in any environment. In the case of fixed HDFS-backed clusters, Cloudera recommends using the same security best practices used in on-premises clusters. That includes authentication with Kerberos, integration with a directory service through LDAP, authorization and permission controls through Sentry and RecordService, end-to-end encryption with HDFS encryption and Navigator Encrypt, and key management with Navigator Key Trustee.

For details on securing transient object-store-backed Cloudera clusters, we recommend reviewing the Cloudera Reference Architecture guides available here: http://www.cloudera.com/documentation/other/reference-architecture.html

Conclusion

Whether you're all in on the cloud already, migrating certain workloads, or simply exploring how the cloud might impact your business, it's important to think about how architecture choices can impact your data strategy. Cloudera Enterprise is a modern data platform that can run on any environment and is flexible enough to deliver the applications you need to empower IT, data engineers, business analysts, and more to get more value from their data. For more information, visit http://www.cloudera.com/products/cloud.html

cloudera.com

1-888-789-1488 or 1-650-362-0488
Cloudera, Inc. 1001 Page Mill Road, Palo Alto, CA 94304, USA