Qubole®

2018 Cloud Big Data Platforms Buyers Guide

Who is this guide for?

The cloud and big data are really a perfect match as the dynamic scalability and affordability of the cloud match well with the massive scale and “bursty” nature of resources that are required for big data analytics. Whether you are just evaluating cloud big data platforms or already running big data in the cloud – this guide is for you. It will give you the fast track to considerations about big data in the cloud based on insights from companies already succeeding.

STEP ONE

BIG DATA IN THE CLOUD

The cloud has become the new norm in enterprise software infrastructure and applications across almost every category. And with each software category that moves to the cloud, a new vendor who designed with cloud-first thinking, moves to the forefront and displaces the past leaders (e.g. Salesforce rewriting CRM). This happens because the cloud is different and requires new architectural thinking.

How is the Cloud Better Supporting The Growth of Big Data?

The shift to the cloud is well underway. In fact, 51% of companies say they are actively increasing their investment in supporting big data in the cloud according to Forrester Research. The majority of new big data programs are being built in the cloud today (vs. on-premises) on Infrastructure as a Service Providers (IaaS) like Amazon Web Services (AWS).

What is Behind The Shift of Big Data to The Cloud?

Managing big data today creates several challenges for data infrastructure teams that the cloud architecture solves:

1. Managing "bursty" and unpredictable workloads
2. Coordinating ad hoc and batch workloads
3. Scaling rapidly growing data stores
4. Integrating data generated at the edge
5. Managing storage and compute costs

Cloud Architectural Advantages

Separation of Compute and Storage

Unlike an on-premises solution, which combines compute and storage in HDFS, the cloud separates compute and storage. This allows for the expansion of compute and storage separately. Given that storage is cheap and compute is expensive, this allows businesses to use compute selectively. Rather than provisioning storage and compute to peak capacity, resources can be scaled as needed.

Learn More

Separation of Logical and Physical

Cloud architecture also creates a separation of the logical and physical, allowing data teams to provision workloads without using physical resources at all times. For example, one logical cluster can be configured for Spark and another for Hive workloads, and the physical cluster would be provisioned only when there is a workload. This of course saves money by not paying for unused dedicated resources for each workload.

Tiered Storage

The cloud offers greater flexibility via tiered storage. Data teams can select and adjust how data is stored depending on how accessible the data needs to be. Amazon Glacier, for example, is a low-cost storage service built for long-term storage of data that is accessed infrequently.

Service Orientation

In the cloud, data teams manage APIs rather than hardware, allowing teams to focus more on applications that impact the bottom line. A service orientation is also a low latency way to provision capacity.

Pooled Resources

The pooled resources of a cloud environment allow for practically unlimited access to resources enabling greater elasticity and multitenancy. In the cloud, workloads can expand and be shifted between different nodes and engines depending on the user's need.

Connectivity

Since the cloud is not limited by hardware capacity, it offers greater connectivity between data sources and enables high-throughput data pipelines.

Geography

Since major cloud services are distributed around the world, businesses have greater flexibility over where data is stored. This provides more options for businesses collecting data in countries with legal restrictions or needing to ingest data from 3rd parties. Geographic distribution also supports hub and spoke collection, as the cloud store acts as a centralized hub that can be accessed by offices scattered throughout the world.

BENEFITS

The cloud's architecture leads to six main benefits for big data workloads:

Adaptability

Cloud infrastructure can adapt seamlessly to changing workloads and business use cases. The elasticity of the cloud allows data teams to focus more on managing data and less time managing the data platform. Data teams can scale clusters up and down as needed or rely on advanced cloud data platforms, like Qubole, that offer complete cluster lifecycle management to scale clusters up and down automatically as needed to match query workload.

The cloud also allows you to select the instance type that is best suited for a given workload, and gives you access to an assortment of different engines — Hive, Pig, Presto, and more —depending on the use case.

Agility

The cloud also creates greater agility. While on-premises solutions frequently require 6 to 9 months to implement, we have found that Qubole customers can begin querying their data in an average of 2.8 days in the cloud. With such a low startup time, business teams are able to allocate time and resources to building applications, running queries and extracting actual business value.

The cloud also allows teams to iteratively determine the best performance and cost and adjust it as needs change. By using the cloud, teams can adjust and optimize the configuration, such as the machine type or cluster size. On-premises solutions do not give teams this option, meaning they're stuck with what they bought and deployed.

Geographic Reach

As mentioned earlier, the cloud allows organizations a choice in where they can store their data. This decision can be based on factors such as overall convenience, where the data originates from, and any legal issues with the data being used.

Lower Overall Cost

Big Data is expensive and becoming more so as the volumes in data lakes continue to expand year over year. The cloud makes affording the growth in big data possible vs. on-premises solutions. For peak capacity to handle bursty workloads, the cloud is able to scale as needed, allowing businesses to only pay for compute space when it is needed. This takes advantage of the elasticity of the cloud, meaning you only have to pay for what you're actually using.

Organizations running on AWS can also save a lot of money by incorporating and using Spot instances (spare EC2 computing instances offered at a discount compared to On-Demand pricing.) Using Spot instances can significantly reduce the cost of running big data applications or increase vs increasing existing application's compute capacity.

Greater Fault Tolerance, Resilience, Disaster Recovery

The cloud is more fault tolerant and resilient than on-premises solutions. It allows enterprises to recover more quickly in the event of a disaster. If there's a node failure or issue with a cluster, teams can seamlessly provision a new node or spin up a cluster in an alternate location. By automating this process, data teams spend less time on maintenance and can focus more on business needs.

Up-to-Date Security

It's a commonly held myth that the cloud is less secure than an on-premises solution. However, the cloud is often the more secure option. Cloud providers typically dedicate much more time and resources to security and are able to adopt best practices faster.

STEP TWO

MAKE YOUR BIG MOVE TO THE CLOUD

When companies migrate workloads from on-premise to the cloud they get to take advantage of all that the cloud has to offer without suffering any of the traditional limitations of an on-premise solution. This newfound freedom allows data teams to get to the real work of expanding the number of active users thereby enhancing the analytic value of the data.

Data Storage - Cluster Cost Containment Strategies

Before we can begin our discussion of the architectural approaches companies take to move big data to the cloud, we need to better understand cloud data storage and approaches. Reducing data storage costs and elastic computing are critical reasons for moving big data to the cloud, so make sure your platform will provide you the scalability you seek.

1. Storage

One common storage pattern is to store HDFS data blocks in Hadoop clusters using local instance storage. The issue with using local instance storage is that it's ephemeral. If a server goes down, whether it is stopped or due to failure, data on instance storage is lost. This can be metadata, schemas and result sets and can really set back your job completion schedule and create risk. Your big data platform should protect against this loss.

2. Autoscaling

is impacted by the ephemeral nature of data as well when clusters are changing on long running jobs (metadata and result sets again will not be available if the cluster is not running). Autoscaling in a big data environment is different than autoscaling in a transactional, short running job web server environment which is what the IaaS platforms were designed for. So look closely at how auto-scaling will work if you have long running, bursty nature jobs.

3. Automated Support of Spot Instances

AWS Spot instances represent excess capacity and are priced at up to 80% discount from on-demand instance prices. By setting a simple policy, such as "bid up to 100% of the on-demand price and maintain a 50/50 on-demand to Spot ratio", some platforms will automatically manage the composition and scaling of the cluster while making bids for Spot instances automatic instead of manual. One way to advance the cost savings potential of Spot instances further is to look for heterogeneous cluster support enabling the inclusion of multiple instance types for nodes within a cluster. By casting a wider net in instance types, you can take greater advantage of the broader Spot market and pricing efficiencies, for example substituting one extralarge node for 2 large nodes if it costs less. By taking advantage of these efficiencies as Qubole does, customers can save up to 90% from On-Demand instance pricing.

Big Data Architectural Approaches

Typically, big data migrations to the cloud fit into one of these 3 architectural styles:

Lift and Shift

in this style of migration, the business simply replicates their on-premise big data clusters into the cloud and continues to own their software stack. The cloud is used to achieve OPEX vs. CAPEX financial advantages of rental vs. purchase and to relieve the business from purchasing, operating and supporting hardware. None of the agility advantages of cloud computing are achieved.

Lift and shift strategies can make sense for always-on workloads as long as the organization owning the big data is tracking the cloud resource consumption and ensuring that it is provisioning the right amount of resources.

However, many businesses utilize multiple engines on top of Hadoop for different use cases of data science, ETL or BI (Hive, MapR, Spark etc). With a lift and shift architecture, each software must be run on its own cluster sized for peak capacity making sharing of resources impossible and making this a very high cost option. Further, with Lift and Shift, clusters are not optimized for BI query environments which 75% of big data organizations now support.

Sophisticated scheduling is not available, opening up issues of individual users or queries consuming the clusters resources without regard to service agreements.

Lift and Reshape

in this style true generic cloud computing is adopted and this is the minimum “right approach” architecture for most, The full benefits of the cloud begin to materialize when an organization adopts a workload driven approach rather than a capacity driven approach that fully takes advantage of the cloud’s elasticity. With lift and reshape, IT can move from the role of provisioning expensive “what if” capacity and become a facilitator of business impact.

With lift and reshape, the organization migrates their big data to Amazon, Azure, Oracle, or another IaaS provider. They achieve the scalability and cost benefits of separating compute from storage options, they can control and manage cluster cost and take advantage of the wide range of managed compute and storage options available, cloud provider rules-based auto-scaling is available based on CPU utilization and other form entered metrics and spot bidding for clusters can be performed but it is not optimized and automated.

Automated Cloud

This style builds on top of Lift and Reshape big data cloud adoption and adds advanced features built specifically to optimize costs and cloud computing for big data operations. Using a combination of heuristics and machine learning, big data cloud automation ensures:

- Workload continuity
- High performance
- Low reliance on cloud resources
- Greater cost savings

By automating lower-level, repetitive tasks engineering teams can be less reactive to problems and more focused on directing better business outcomes. Autonomous clouds for big data constantly analyze metadata about infrastructure (cluster, nodes, CPU, memory, disk), platforms (data models and compute engines) and applications (SQL, reporting, ETL, machine learning) so you can better understand performance, usage patterns and cloud spend.

Three primary areas impacted by the automated cloud – cluster lifecycle management, auto-scaling clusters and the automated optimization of spot bidding.

1. With cluster lifecycle management, automation is used to automatically manage the entire lifecycle of Hadoop, Spark, and Presto clusters. This simplifies both the user and admin experiences. Users such as data analysts and data scientists can simply submit jobs to a cluster label and an automated cloud platform like Qubole Data Service (QDS) will automatically bring up clusters. There is no dependency on an admin to ensure cluster resources. Similarly, admins no longer need to spend time manually deploying clusters or developing scripts or templates to automate this action. Further admins do not need to worry about shutting down clusters to avoid charges when jobs are complete as this occurs automatically.

2. Auto-scaling in autonomous cloud goes beyond generic cloud provider auto-scaling to optimize for price and availability across available node types while ensuring data integrity and required compute resources are applied to meet service agreements. Using auto- scaling optimized for big data vs. generic approaches has been shown to save on compute costs as much as 33% and lower the risks of data loss described earlier in the data storage section.

3. With automated Spot bidding an agent 'shops' for the best combination of price and performance, based on the policy you provide. It achieves this by shopping across different instance types, by dynamically rebalancing Spot and On Demand nodes and by considering different Availability Zones and time shifting work. In addition, replicas of HDFS blocks are stored on stable nodes to prevent job failures if AWS reclaims Spot nodes. Automated Spot bidding for big data has been shown to achieve costs 90% less than Spot bidding with On-Demand clusters.
   Learn more

STEP THREE

CLOUD DATA PLATFORM FEATURES OVERVIEW

	Hortonworks	Cloudera Altus	HDInsight	Databricks	Amazon EMR	Qubole
Product Summary	Big Data IaaS in AWS Cloud	Big Data IaaS in AWS Cloud	Big Data IaaS in Azure Cloud	Standalone Spark Service	Big Data IaaS in Azure Cloud	Cloud agnostic Big Data IaaS
Out-of-The Box Data Processing Engines	Spark, Hive	Spark	Hive, Spark, HBase, Storm, Kafka, RServer	Spark	MapReduce, Hive, Pig, Cascading, Impala, Spark, Presto, HBase	Spark, MapReduce, Presto, Hive, Pig, Cascading
Supported Clouds	AWS	AWS, Azure (Beta)	Azure	AWS	AWS	AWS, Azure, Oracle
Big Data Architectural Styles	Lift and Shift, Lift and Replace	Lift and Shift, Lift and Replace	Lift and Shift, Lift and Replace	Lift and Shift, Lift and Replace	Lift and Shift, Lift and Replace	Lift and Shift, Lift and Replace, Automated

Easing the Migration From On-Premise to the Cloud

Companies choosing to migrate workloads from on-premise to the cloud should note that tools and vendor programs are available to speed up the migration and lower the costs. One vendor tool for instantaneous workload movement to the cloud is WANdisco's Fusion Active Migration platform which can result in real time data availability in the cloud on AWS. Through the Qubole Cloudera Migration Program, Qubole and WANDisco allow customers to synchronize their on-premise and cloud data and enable one-time migration, workload bursting to the cloud and parallel environments.

Summary

We hope this buyer's guide to big data in the cloud has been useful at providing valuable information about why big data is moving to the cloud, how savings and agility are achieved and which data platforms offer which advanced cloud capabilities.

Qubole was founded by the creators of Hive and the leaders of the big data team at Facebook who used automation to achieve mass adoption across usage types and technologies of big data enterprise-wide. We welcome you to learn more about what they have built to make the high levels of big data automation available to others at www.qubole.com

READY TO GIVE QUBOLE A TEST DRIVE?

Try QDS for Free

About Qubole

Qubole is passionate about making data-driven insights easily accessible to anyone. Qubole customers currently process nearly an exabyte of data every month, making us the leading cloud-agnostic big-data-as-a-service provider. Customers have chosen Qubole because we created the industry's first autonomous data platform. This cloud-based data platform self-manages, self-optimizes and learns to improve automatically and as a result delivers unbeatable agility, flexibility, and TCO. Qubole customers focus on their data, not their data platform. Qubole investors include CRV, Lightspeed Venture Partners, Norwest Venture Partners and IVP. For more information visit **www.qubole.com**

For more information:

Contact:
sales@qubole.com

Try QDS for Free:
https://www.qubole.com/products/pricing/

469 El Camino Real, Suite 205
Santa Clara, CA 95050
(855) 423-6674 | info@qubole.com

WWW.QUBOLE.COM