Scalability and availability are arguably the two most essential buzzwords in the cloud market today. Both clients and providers diligently advocate for highly scalable and available platforms – emphasizing their significance in choosing the right service provider for you.

The three major cloud providers have had runaway success because they have manipulated cloud resources to excel in the critical metrics of scalability, availability, and monitoring. Advancing these capabilities of their platforms, they have been able to allure the biggest names in the world economy and have constructed a cloud empire.

Cloud services are analyzed based on the above-mentioned primary abilities, ensuring customers have a fast, secure, and reliable infrastructure worthy of an investment.

Let's explore what scalability, availability, and monitoring offer clients, shifting to the cloud.

Scalability

Scalability is essentially the measure of the extent to which a piece of infrastructure can be expanded to cater to more users on your application. This ability underlines how efficiently your provider responds to your application's dynamic requirements.

As the load on your application increases, you automatically require increased performance and resources to maintain functionality. How effectively these needs are met determines how scalable your cloud service is.

Scalability is classified into two types, namely, vertical and horizontal.

Vertical scaling, also called scaling-up, means adding more power (CPU, RAM, SSD) to the infrastructure hosting your application. This type of scalability is easier to achieve. Still, it is only a short-term solution, for there is a limit to how much power can be added to a single server until it becomes unfeasible. Moreover, it fails to account for downtime scenarios.

On the other hand, horizontal scaling also referred to as scaling-out, is a must-use technology.

Scaling-out means adding more servers to the hosting infrastructure to spread the load of the hosted application – essentially how the concept of clusters works.

Clusters are a group of interconnected nodes (servers) running in parallel to achieve a unified goal. This way, the increased load on the application can be distributed across multiple host servers, making it scalable.

This type of scalability is more complex due to its distributed nature but is the pith of cloud platforms today.

High Availability

Availability refers to the uptime of any cloud service. No infrastructure is perfectly immune to failure or downtime, whether due to technical issues or scheduled maintenances. Therefore, the ability to continue operating with no stoppages is a vital feature to have in your service provider.

Having many available servers permits the infrastructure to host your application despite the failure or shutting down of specific components within the structure itself.

Up-times are usually estimated with the number of 9's in the figure for up-time percentages. For example, an up-time of 99.9% depicts that the system might only face outages for a combined of 8.76 hours maximum in a year.

Availability incorporates the concepts of redundancy, monitoring, and failover. Through redundancy, the system utilizes additional components mimicking critical elements of the infrastructure, acting as backups. Through monitoring, currently running data is collected to look out for failure. Consequently, failover allows the routing of the application to a redundant component, making sure your program is always running.

Monitoring

Monitoring involves obtaining remote real-time application information to oversee the application's performance and hosting infrastructure for better service. It ensures the efficient running of the program on the cloud.

Monitoring policies are the implementation of strategy between the infrastructure management and configuration management.

Some key attributes of monitoring systems include assessment and evaluation of resource utilization, validating servers and their response times, database management, availability, updates, and security to foresee possible issues.

Let’s delve deeper into the detailed services provided by the ‘big three’ regarding these features.

Scalability

Amazon Web Services (AWS)

AWS has been one of the pioneers of cloud platforms and is the largest cloud provider in the world today. Their infrastructure is highly scalable and provides clients a significant number of options with many scaling features.

Scale-out services	**Description**
Application Load Balancer (ALB)	Distributes load of HTTP and HTTPS traffic. ALBs have default limits set which can be raised on requests. Advanced requests can also be made to route load to specific EC2 instances.

Autoscaling Groups (Amazon EC2 Auto Scaling)	Helps users to automatically add or remove EC2 instances with the help of fleet management to ensure application availability. Fleet management monitors the health of running instances, replaces impaired instances automatically, and efficiently balances load across availability zones based on traffic.
Elastic Beanstalk	Service for building scalable web applications (PHP, Ruby, .NET, etc.) automatically by managing the underlying infrastructure.
Elastic Container Service (ECS)	Cater for containerized applications – packages of application code with dependencies to allow them to run virtually anywhere.

AWS provides users with the greatest number of services coupled with a dynamic array of alternative packages. Their policy of pay-as-you-go plans also allows you to pay for the services you only use. However, AWS's load balancers tend to be flooded with requests and require advance subscriptions for more seamless service due to its demand.

Azure

Microsoft's Azure is second to AWS in popularity and is a challenging competitor in the market. When it comes to scalability, Azure primarily offers built-in autoscaling mechanisms designed for typical scenarios; however, users have the luxury of creating custom scalability implementations. Autoscaling allows adding resources to handle load increases and saves money by removing resources sitting idle.

Autoscaling services	**Description**
Azure Virtual Machines	Azure uses virtual machine scale sets: a system to manage a group of load-balanced VMs to auto-scale based on application load. This is useful for large-scale applications.
Service Fabric	Every node type in a service fabric is set up as a virtual machine scale set – each node type can be scaled in or out independently.
Azure App Service	Also has incorporated autoscaling and can be applied to all applications within.
Azure Cloud Services	Built-in autoscaling at the role level available.
Azure Functions	Automatically allocated appropriate compute power to the running application, scaling in and out as necessary. There is no need even to configure any autoscaling rules.

Azure's scalability services are very enterprise-ready and cater to a variety of user needs. Their discounts are a notable feature for a lot of clients. Adversely, Azure's autoscalers have reported to overprovision resources at times which can prove to be costly.

Google Cloud Platform (GCP)

GCP, a quickly rising giant, also offers its clients state-of-the-art scalability features. GCP provides autoscaling to add or remove Virtual Machine instances through a managed instance group, similar to Azure. It simply requires users to define the autoscaling policy, and the scaling is self-regulated depending on the load.

Autoscaling signals	Description
CPU utilization	The most basic form of scaling. It enables the autoscalers to monitor the average CPU utilization of a group of instances and scale accordingly.
Load Balancing Capacity	The autoscalers watch the serving capacity of an instance group and scale when the VM instances are over or under capacity.
Monitoring metrics	Autoscaling to collect data of a specific metric and perform scaling based on your desired utilization level. Allows for custom metrics as well.
Schedules	Improve the availability of workloads by scheduling capacity ahead of predicted load based on previous patterns.

GCP's billing by the minute policy takes it to an extreme of pay-as-you-go policy. Google Compute Engine's load balancers are reported to be the most efficient, and GCP allows scaling of massive data applications in swift amounts of time. However, it lacks various services available compared to AWS and Azure, making it slightly less enterprise-ready.

High Availability

Amazon Web Services (AWS)

AWS being a household name provides its users with a highly available cloud infrastructure. Amazon instills the capabilities of Elastic Load Balancing, a massive network of availability zones, and auto-scaling features to ensure minimal downtime of services.

AWS is prevalent in 81 availability zones (each consisting of self-sufficient datacenters) spread across 25 geographic regions globally, making it the most widespread cloud service provider.

This, coupled with the ability to balance load requests between instances contained in separate availability zones, ensures that clients receive continuous uptime of guaranteed 99.99% monthly.

Amazon also guarantees a staggering 99.999% per-year availability on their S3 storage service with SoftNAS, a state-of-the-art software-defined network to assign storage.

Azure

Dominating with massive infrastructure, Azure currently has 39 regions worldwide, each having a minimum of three availability zones – accounting for over 200 physical datacenters.

In addition to a large number of availability zones, Azure allows customers to set up Availability Sets to ensure minimal downtime in case of maintenance or hardware failure.

Availability sets allow users to run a virtual machine (VM) with more than one replicated copy on different hardware in the same availability zone.

Essentially, availability zones protect customers from entirely compromised datacenters, while availability sets offer to safeguard from hardware failures within a datacenter. Azure offers 99.99% of uptime per month.

Azure also offers the service of Region Pairs which enables a VM to run on different availability zones with one or more replicated copies. However, these availability zones will always be within the same geographical region.

Google Cloud Platform (GCP)

Racing to expand its infrastructure, GCP has managed to cover around 27 regions worldwide and have a total of 82 availability zones for the time being. Each zone consists of multiple clusters of physical machines, and Google's Compute Engine preserves an abstraction layer between its availability zones and clusters.

GCP provides its users with managed instance groups, a collection of virtual machines grouped to serve a singular purpose. These instance groups can distribute the load over multiple virtual machines through a load balancer and permit the creation of a group spread over various zones in the same region, similar to Azure's model.

GCP offers a monthly uptime of 99.99% as well.

You can get to know more about the global infrastructures in one of our recently published articles here.

Monitoring

While several third-party monitoring services are available for all cloud service platforms, below is a summary of dedicated monitoring services and features of the three cloud providers.

Amazon Web Services (AWS)

Service	Description
AWS CloudWatch	Captures monitoring and operational data in various forms such as metrics, logs, events, AWS resources, and services rendered on AWS servers.
AWS CloudTrail	Tracks occurred events and records them in S3 storage – includes IP addresses, user identities, timestamps, etc.
Amazon EC2 Dashboard	Manages the states and health of EC2 instances across the infrastructure.

AWS Certificate Manager	Authenticates consumer’s services and devices to provide a secure connection.

Azure

Service	Description
Azure Monitor	Provides detailed metrics and logs of any Azure resource regarding health, performance, views, etc., and can provide alerts on predetermined conditions.
Azure Advisor	Examines currently used resources and proposes services and plans for optimized performance, availability, security, and feasibility.
Azure Cost Management	Analyzes resources incurring costs and manages them conservatively.
Azure Automation	Automates management tasks and organizes tasks across external systems.

Google Cloud Platform (GCP)

Features	Description
SLO monitoring	Consistently monitors service-level objectives (service’s desired objectives) and notifies upon violations.
Custom metrics	Generate application and business-related metrics.
Google Cloud Integration	Allows users to track and examine all Google Cloud resources without any additional orchestration.

Review

All three service providers have set up a very competitive market and comprehensively cater to all the key capabilities required for reliable and efficient cloud infrastructure.

AWS tends to own the bragging rights in the industry because of its role as a pioneer in the technology. However, this is far from being a decisive factor for the services provided by all three do not differ significantly.

Talking about scalability, AWS and Azure seem to provide packages better suited for enterprises, but Google, with its leaps in big data analytics, provides data-extensive services like no other. GCP's extreme pay-as-you-go packages for small-scale users tend to be slightly cost-effective than Amazon’s and Microsoft’s.

AWS boasts the vastest physical infrastructure to date, with Azure a very close second and GCP catching up rapidly. Still, in practicality, this tends to have little effect on the availability of services.

Finally, when it comes to monitoring and assisting capabilities, AWS has a large pool of associated third-party services, with GCP being the second most popular platform for businesses. Azure tends to rely on its monitoring service as it is favored for enterprise relations.

Conclusively, trade-offs are a critical factor in deciding your cloud provider and depend focally on the features required by individual businesses.

For further comparisons regarding price models, you can find more information here.