Resource Availability

After completing this episode, you should be able to:

- Discuss the topic of resource availability as it relates to modern day cloud technologies

Description: In this episode, you will learn about the topic of resource availability when it comes to cloud architectures. This discussion includes such topics as regions, availability zones, cloud bursting, edge computing, and availability monitoring.

Resource Availability

Regions - the major public cloud providers (Amazon, Google, Microsoft) offer global infrastructures that span the entire globe. They divide these resources into regions. Customers can place resources in specific regions in order to reduce latency on access.

Availability zones - an availability zone is a distinct and isolated location within a geographic region that houses data centers equipped with independent power, cooling, and networking infrastructure. Availability zones are designed to provide redundancy and fault tolerance, enabling applications and services to remain operational even in the event of failures or disruptions affecting one availability zone.

Cloud bursting - cloud bursting is a dynamic and flexible cloud computing strategy that allows organizations to seamlessly scale their workloads from a private cloud to a public cloud environment during periods of peak demand. In essence, it enables businesses to augment their existing on-premises infrastructure with additional computing resources from a public cloud provider when local capacity is exceeded. This burst capacity can accommodate sudden spikes in user activity, seasonal variations, or other unforeseen surges in demand without requiring significant upfront investments in infrastructure expansion. By leveraging cloud bursting, organizations can optimize resource utilization, maintain performance levels, and ensure uninterrupted service delivery, all while minimizing costs and maximizing operational efficiency.

Edge computing - Edge computing is a distributed computing paradigm that brings computational resources closer to the data source or end-user devices, thereby reducing latency and improving real-time processing capabilities. Unlike traditional cloud computing, where data is processed in centralized data centers, edge computing decentralizes computing power to the "edge" of the network, such as IoT devices, routers, or local servers. This approach enables faster data processing, analysis, and response times, making it ideal for applications requiring low latency, high bandwidth, and efficient use of network resources. Edge computing is particularly valuable in scenarios like IoT deployments, autonomous vehicles, and industrial automation, where timely decision-making and local data processing are critical. By pushing computing closer to where data is generated and consumed, edge computing enhances scalability, reliability, and security while unlocking new opportunities for innovation and efficiency in diverse industries.

Availability monitoring - cloud computing makes it easy to provide availability monitoring for both your own cloud workloads, as well as the services of the cloud provider. While availability monitoring for your resources in the cloud requires some level of configuration, the major cloud providers all provide web pages (and even APIs) that provide availability information for their resources and services.

Additional resources

- Cloud Bursting: https://www.vmware.com/topics/glossary/content/cloud-bursting.html (https://www.vmware.com/topics/glossary/content/cloud-bursting.html)