Cloud computing, also referred to as “the cloud,” is the delivery of on-demand computing resources—everything from applications to data centers—over the internet on a pay-for-use basis. To get a common understanding of cloud computing, let’s start with the US National Institute of Standards and Technology (NIST’s) definition of cloud computing. NIST defines cloud computing as a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. Examples of computing resources include networks, servers, storage, applications, and services. This cloud model is composed of five essential characteristics, three deployment models, and three service models. Let’s start with understanding the five essential characteristics of the cloud—which include on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. On-demand Self-service, the 1st characteristic, means that you get access to cloud resources such as the processing power, storage, and network you need, using a simple interface, without requiring human interaction with each service provider. The 2nd characteristic, Broad Network Access, means that cloud computing resources can be accessed via the network through standard mechanisms and platforms such as mobile phones, tablets, laptops, and workstations. The 3rd characteristic, Resource Pooling, is what gives cloud providers economies of scale, which they pass on to their customers, making cloud cost-efficient. Using a multi-tenant model, computing resources are pooled to serve multiple consumers; cloud resources are dynamically assigned and reassigned, according to demand, without customers needing to concern themselves with the physical location of these resources. Rapid Elasticity, the 4th characteristic, implies that you can access more resources when you need them, and scale back when you don’t—because resources are elastically provisioned and released. And the 5th characteristic, Measured Service, means that you only pay for what you use or reserve as you go; if you’re not using resources, you’re not paying. Resource usage is monitored, measured, and reported transparently based on utilization. As we see, cloud computing is really about utilizing technology “as a service”—leveraging remote systems on-demand over the open internet, scaling up and scaling back, and paying for what you use. It is a revolution in that it has changed the way the world consumes compute services by making them more cost-efficient while also making organizations more agile in responding to changes in their markets. As I mentioned earlier in this talk, the cloud model is composed of five essential characteristics, three deployment models, and three service models. We just went over the five essential characteristics. While we will go into greater depth of the deployment and the service models in the later videos, let me leave you with a brief overview of these models. There are three types of cloud deployment models—Public, Private, and Hybrid. Public cloud is when you leverage cloud services over the open internet on hardware owned by the cloud provider, but its usage is shared by other companies. Private cloud means that the cloud infrastructure is provisioned for exclusive use by a single organization. It could run on-premises or it could be owned, managed, and operated by a service provider. And when you use a mix of both public and private clouds, working together seamlessly, that is classified as the Hybrid model. Now, let’s look at the three service models that are based on the three layers in a computing stack - Infrastructure, Platform, and Applications. These cloud computing models are aptly referred to as Infrastructure as a Service (Iaas), Platform as a Service (PaaS), and Software as a Service (SaaS). In an Infrastructure as a Service model, you get access to infrastructure and physical computing resources such as servers, networking, storage, and data center space - without the need to manage or operate them. In a Platform as a Service model, you get access to the platform, that is the hardware and software tools, usually those needed to develop and deploy applications to users over the Internet. Software as a Service is a software licensing and delivery model in which software and applications are centrally hosted and licensed on a subscription basis, and sometimes also referred to as "on-demand software." In the next training, we will go over some of the key evolutions that brings us to what cloud computing is today.