Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). The name comes from the use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts remote services with a user's data, software and computation.

The business model, using software as a service, users also rent application software and databases. The cloud providers manage the infrastructure and platforms on which the applications run.

End users access cloud-based applications through a web browser or a light-weight desktop or mobile app while the business software and user's data are stored on servers at a remote location. Proponents claim that cloud computing allows enterprises to get their applications up and running faster, with improved manageability and less maintenance, and enables IT to more rapidly adjust resources to meet fluctuating and unpredictable business demand.[2][3]
Cloud computing relies on sharing of resources to achieve coherence and economies of scale similar to a utility (like the electricity grid) over a network.

At the foundation of cloud computing is the broader concept of converged infrastructure and shared services.

The origin of the term cloud computing is obscure, but it appears to derive from the practice of using drawings of stylized clouds to denote networks in diagrams of computing and communications systems. The word cloud is used as a metaphor for the Internet, based on the standardized use of a cloud-like shape to denote a network on telephony schematics and later to depict the Internet in computer network diagrams as an abstraction of the underlying infrastructure it represents. The cloud symbol was used to represent the Internet as early as 1994.

In the 1990s, telecommunications companies who previously offered primarily dedicated point-to-point data circuits, began offering virtual private network (VPN) services with comparable quality of service but at a much lower cost.

By switching traffic to balance utilization as they saw fit, they were able to utilize their overall network bandwidth more effectively. The cloud symbol was used to denote the demarcation point between that which was the responsibility of the provider and that which was the responsibility of the users. Cloud computing extends this boundary to cover servers as well as the network infrastructure.
The underlying concept of cloud computing dates back to the 1950s; when large-scale mainframe became available in academia and corporations, accessible via thin clients / terminal computers. Because it was costly to buy a mainframe, it became important to find ways to get the greatest return on the investment in them, allowing multiple users to share both the physical access to the computer from multiple terminals as well as to share the CPU time, eliminating periods of inactivity, which became known in the industry as time-sharing.

As computers became more prevalent, scientists and technologists explored ways to make large-scale computing power available to more users through time sharing, experimenting with algorithms to provide the optimal use of the infrastructure, platform and applications with prioritized access to the CPU and efficiency for the end users.

John McCarthy opined in the 1960s that "computation may someday be organized as a public utility."

Almost all the modern-day characteristics of cloud computing (elastic provision, provided as a utility, online, illusion of infinite supply), the comparison to the electricity industry and the use of public, private, government, and community forms, were thoroughly explored in Douglas Parkhill's 1966 book, The Challenge of the Computer Utility. Other scholars have shown that cloud computing's roots go all the way back to the 1950s when scientist Herb Grosch (the author of Grosch's law) postulated that the entire world would operate on dumb terminals powered by about 15 large data centers.
Due to the expense of these powerful computers, many corporations and other entities could avail themselves of computing capability through time sharing and several organizations, such as GE's GEISCO, IBM subsidiary The Service Bureau Corporation (SBC, founded in 1957), Tymshare (founded in 1966), National CSS (founded in 1967 and bought by Dun & Bradstreet in 1979), Dial Data (bought by Tymshare in 1968), and Bolt, Beranek and Newman (BBN) marketed time sharing as a commercial venture.

The development of the Internet from being document centric via semantic data towards more and more services was described as "Dynamic Web". This contribution focused in particular in the need for better meta-data able to describe not only implementation details but also conceptual details of model-based applications.

After the dot-com bubble, Amazon played a key role in the development of cloud computing by modernizing their data centers, which, like most computer networks, were using as little as 10% of their capacity at any one time, just to leave room for occasional spikes.

The ubiquitous availability of high-capacity networks,
low-cost computers and storage devices as well as the widespread adoption of hardware virtualization, service-oriented architecture, autonomic, and utility computing have led to a tremendous growth in cloud computing.

Having found that the new cloud architecture resulted in significant internal efficiency improvements whereby small, fast-moving "two-pizza teams" (teams small enough to be fed with two pizzas) could add new features faster and more easily, Amazon initiated a new product development effort to provide cloud computing to external customers, and launched Amazon Web Service (AWS) on a utility computing basis in 2006. In early 2008, Eucalyptus became the first open-source, AWS API-compatible platform for deploying private clouds. In early 2008, Open Nebula, enhanced in the RESERVOIR European Commission-funded project, became the first open-source software for deploying private and hybrid clouds, and for the federation of clouds. In the same year, efforts were focused on providing quality of service guarantees (as required by real-time interactive applications) to cloud-based infrastructures, in the framework of the IRMOS European Commission-funded project, resulting to a real-time cloud environment. By mid-2008, Gartner saw an opportunity for cloud computing "to shape the relationship among consumers of IT services, those who use IT services and those who sell them" and observed that "organizations are switching from company-owned hardware and software assets to per-use service-based models" so that the "projected shift to computing... will result in dramatic growth in IT products in some areas and significant reductions in other areas."