The cloud

Table of Contents

last modified: 2020-11-15

1. Note on the terminology: what is a server?

To understand the cloud precisely, We need first to have a clear vision of what a server is. A server is simply a computer stripped of everything unessential (screen, mouse, graphic card, sound card, keyboard)… To illustrate: when Google is calculating what the best results are for your search "what are cheap and delicious restaurants in Lyon", this calculation must be done on a computer, right? The kind of computer used to do this calculation is **not** this one:

Figure 1. A desktop computer

The reason is, we don't need a screen, mouse, desktop, and not even the big box containing the computer itself. They take too much space, consume energy, and there is no need for them. So when all the unnecessary parts are removed, the computer looks like this, and is called a "server":
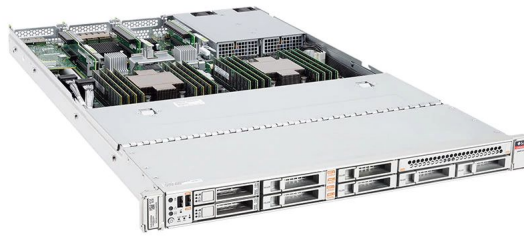
Figure 2. A server

Take a look at the shape: rectangular and very slim. This makes it easy to stack up servers one on the other. Because for Google and other companies crunching data for their business, a lot of servers are needed, so gaining on space is a real issue. When many servers are piled up together and put in a big tall box, this is called a **rack** of servers, and look like this:

Figure 3. A rack of servers

When all the racks of servers are put in the same room, this is called a **data center** and looks like this:

Figure 4. A data center

To get a sens of how "the cloud" is actually a physical space, watch this video showing a tour of a data center at Google:

[] | *https://img.youtube.com/vi/XZmGGAbHqa0/maxresdefault.jpg*

Usually, until 2005 roughly, companies had two options:

- buying their own servers and using them on their premises (at their location).
- paying the services of companies specializing in managing data centers.

Then the "cloud" changed this.

2. The cloud

The term **cloud** was made popular by Amazon with their service "Amazon Elastic Compute Cloud": **Amazon EC2** launched in 2006. This service was new in many ways:

- it is not about renting or buying a particular piece of server anymore. Instead, the user **rents from Amazon a service** of "usage time of a type of server", characterized by a computing power, and a memory size, which you choose. In practice, to deliver this service, Amazon will be able to start 1 large server, or 2 small, or just allocate half of a huge shared server with another client: this is managed transparently for the client. **It is no longer the rental of a "machine" that is billed but the time of use of a service.**
- there is an emphasis on ease of use: no need to know the technical details of these servers (how they are plugged, how they are configured…)
- you are just given a login + password and you can start using these servers for your needs.
- it is **elastic**: if you need more servers, or more powerful servers, this change can be done instantly. No need for signing a new contract, or to worry about stopping the rental of one machine to switch to another. Amazon manages this transparently for you. Let's compare a situation with or without the cloud:

Without the cloud	With the cloud
You make a market study for which server to buy	On Amazon' EC2 website, you click to choose a server among those on offer: it is **on demand**
Get the approval by your finance department to buy it (that's a fixed asset!)	You run your job on it. Costs are metered precisely.
Wait for the server to ship	No maintenance is needed: maintenance is the cloud provider's role, not yours.
Install it and configure it	
Maintain it (security, etc.)	
Without the cloud	**With the cloud**

When the job is over: what do you do with your server? That's a sunk cost. If the job happens to need more computing capacity than your server offers: you are stuck with your too-small-server!	When your job is over, you stop the server with a click and pay the bill. If the job happens to need more computing capacity, you switch to a bigger server with a single click, or it can be done for you automatically: it is **elastic**.
It is a capex	It is an opex

The cloud can fit in the budget as an operational expense instead of a capital expenditure. Opex are not inherently a better or cheaper option than a Capex, but they are easier for a project team or business unit to fit in their budget. See this blog post discussing opex and capex in the context of the cloud, especially the critical comments in response, to continue this discussion.

3. The cloud: what it is for? IaaS, PaaS and Saas

What is the use of the cloud? Companies can externalize operations to the cloud, instead of running these operations on their premises, using their own resources.

Which kind of operations can be externalized to the cloud?

Companies can delegate just the basic IT infrastructure, or IT operations which are closer to the core of their business. These different degrees can be described with the **"pizza model"** (source: source):
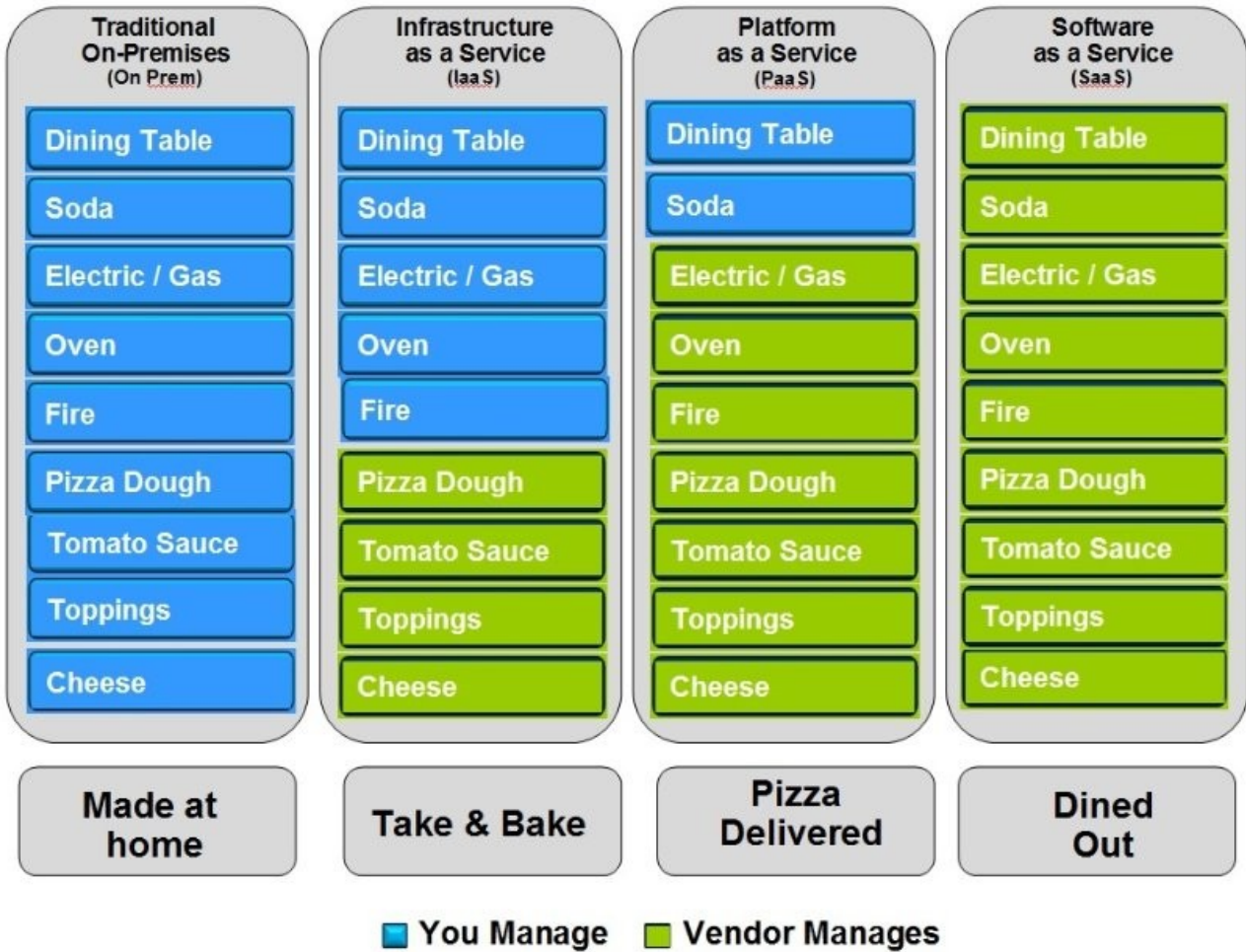
Figure 5. Pizza as a service

This schema illustrates that as a business, you can either run all operations by yourself ("made at home"), or delegate everything ("dining out"). Each of these degrees of externalization has a name:

Infrastructure as a service (IaaS)

The cloud is used to replace the company's local IT infrastructure needs such storing data, or computing operations. For example, instead of storing your data in an on-site database, it is possible to rent a cloud data storage service. It will be charged specifically to the time of use, the size of the data stored, and the volume of data being to and from the cloud. As it is a database service, this type of IaaS can be called a DBaaS: database as a service).

Platform as a Service (Paas)

The cloud is used to run the building blocks of a service: to manage a messaging system, to host apps, …

Software as a Service (Saas)

The cloud is used to host a full software accessible "on demand" through the browser. Popular examples are Google Drive, Brightspace or SalesForce.

4. Private or public cloud? Hybrid cloud?

- Amazon EC2 is an example of a **public cloud**: it is publicly accessible to any customer. Of course, this does not mean that every customer can see what the others are doing on the cloud! Each customer have their private spaces on the cloud.

- Many companies have security requirements which prevent them from accessing public clouds. They need to have their servers on premises. In this case, they can build their own **private cloud**: it is a cloud just like Amazon EC2, except that it is owned, managed and used by the company exclusively - it is not accessible to third parties. But even private, the cloud keeps the basic characteristics of a cloud: on-demand and elastic in particular.

- **Hybrid clouds** are a variety of private clouds: it is a private cloud where some forms of operations can be delegated to a public cloud.

For example, operations which are not security sensitive and which need a capacity of computing in excess of what the private cloud of the company can provide.

The end

Find references for this lesson, and other lessons, here.

This course is made by Clement Levallois.

Discover my other courses in data / tech for business: https://www.clementlevallois.net

Or get in touch via Twitter: @seinecle ter: @seinecle