De la informàtica en núvol al *big data*

Visió introductòria per a joves emprenedors

Jordi Torres i Viñals

PID_00194198

UOC
Universitat Oberta de Catalunya

Els textos i imatges publicats en aquesta obra estan subjectes –llevat que s'indiqui el contrari– a una llicència de Reconeixement-NoComercial-SenseObraDerivada (BY-NC-ND) v.3.0 Espanya de Creative Commons. Podeu copiar-los, distribuir-los i transmetre'ls públicament sempre que en citeu l'autor i la font (FUOC. Fundació per a la Universitat Oberta de Catalunya), no en feu un ús comercial i no en feu obra derivada. La llicència completa es pot consultar a http://creativecommons.org/licenses/by-nc-nd/3.0/es/legalcode.ca

Índex

Introducció..... 5

1. Informàtica en núvol..... 7
1.1. La xarxa ha esdevingut l'ordinador 7
1.2. Definició de la informàtica en núvol 9
1.2.1. Infraestructura com a servei 10
1.2.2. Plataforma com a servei 10
1.2.3. Programari com a servei 11
1.3. La flexibilitat de la informàtica en núvol 12
1.4. Consum energètic i dependència de la informació 15
1.5. Els inhibidors de la informàtica en núvol 17
1.6. Un futur mòbil en què tothom i tot estarà connectat 18
1.7. La informàtica en núvol com a motor de canvi 19

2. *Big data*..... 22
2.1. S'acosta una marea d'informació digital 22
2.2. Definició del *big data*..... 22
2.3. Calen noves tecnologies d'emmagatzemament 24
2.4. Les bases de dades relacionals no serveixen per a tot 25
2.5. Es requereixen nous models de programació 26
2.6. La informació no és coneixement 27
2.7. Entrant en la nova era del Data 2.0. 29

Resum..... 31

Introducció

Per als estudiants a qui va dirigit aquest material és un fet natural que els serveis convergeixen i passen del món físic al món digital, i són accessibles des de qualsevol dispositiu electrònic. Formen part d'una generació que ha viscut tota la seva vida adulta en un món físic i digital alhora i que no entenen la tecnologia només com una eina de treball. Es lleven amb tecnologia i se'n van a dormir amb tecnologia. No accedeixen a les xarxes socials únicament des del seu ordinador, sinó que és habitual que també ho facin des del mòbil. Penjar les fotos al moment, enviar un missatge amb el Whatsapp o compartir la ubicació amb la resta d'usuaris, és un fet habitual.

Alhora també el món dels negocis s'està transformant a gran velocitat. La desmaterialització ja ha revolucionat diferents sectors de negoci, i, en alguns, fins i tot ha eliminat un model que funcionava durant anys.

> Un exemple conegut per tots és el món de la música, en què, fins fa poc, les discogràfiques controlaven tota la cadena de producció, des de l'autor fins al consumidor. Internet ha posat fi a aquest negoci amb el sol fet de fer prescindible el suport físic del CD, i ha deixat sense sentit el paper que tenien tots els intermediaris. No és simplement que hagi canviat la manera de vendre cançons, sinó que Internet ha convertit la música en un servei més. Ara, el més important ja no és posseir la música sinó l'accés a les peces que un vol escoltar, oi que sí?

La **informàtica en núvol** (*cloud computing*) és el que fa possible que la tecnologia digital penetri a cada racó de la nostra economia i societat, no solament permetent als usuaris estar connectats a aquest nou món digital per mitjà dels seus dispositius mòbils, sinó que també permetrà en breu que ho estigui qualsevol objecte o dispositiu, el que s'anomena ***Internet of things***. Tot plegat provocarà una marea d'informació digital que requerirà una capacitat d'emmagatzemament i processament de magnituds fins ara mai vistes, fenomen que es comença a anomenar ***big data*** i que sens dubte representa un oceà d'oportunitats professionals per a tots vosaltres.

Aquest mòdul fa un breu repàs a tecnologies entorn de la informàtica en núvol que marcaran el nostre futur proper i també intenta apuntar cap a on evolucionaran, incloent-hi referències a documents que permeten aprofundir en cadascuna i permetre així una formació més personalitzada.

Per això, a més de descriure el fenomen de la informàtica en núvol, que ja és aquí, intentarem visualitzar cap a on va en els propers any, amb l'objectiu d'ajudar els estudiants, als quals va dirigit aquest material, a fer-los partícips d'aquesta profunda transformació que està provocant la informàtica en núvol en l'estructura econòmica, social i cultural, que incideix en gairebé tots els aspectes de la nostra vida: en el món laboral, es manifesta en l'empresa, en l'economia, en la sanitat, en el disseny o en la burocràcia administrativa, i en

el social i cultural, i es fa evident en les nostres relacions personals o en els nostres sistemes d'informació, educació, comunicació i oci. És indubtable que ocupa i ocuparà un paper central en la nostra vida, personal i professional, malgrat que potser no siguem plenament conscients d'aquest fet.

En definitiva, aquest llibre vol afavorir la descoberta d'aquestes noves tecnologies als nostres estudiants, amb el desig d'estimular-los perquè es vulguin endinsar una mica més en aquest apassionant món de la tecnologia i engrescar-se a emprendre.

1. Informàtica en núvol

1.1. La xarxa ha esdevingut l'ordinador

Tot i ser un mòdul que vol parlar de futur, farem una breu referència als orígens de la informàtica en núvol per a ajudar a comprendre per què succeeixen els canvis tecnològics tan ràpidament: les tecnologies de la informació i comunicació[1] són molt joves, i per això encara els queda molt de recorregut. Pensem que els primers ordinadors es remunten no més enllà de la meitat del segle passat, funcionaven amb vàlvules, semblants a unes bombetes, i l'ús era exclusivament per a l'àmbit científic i militar. La programació era feta directament en els circuits mateixos de les màquines. Poc després van aparèixer els transistors i també els primers ordinadors que permetien una certa programació.

(1)A partir d'ara usarem l'acrònim *TIC*.

Però no va ser fins a mitjan anys seixanta que la informàtica va travessar l'hàbitat de la investigació per a entrar progressivament en l'àmbit dels negocis. En aquells moments, la informàtica es basava en grans ordinadors ubicats en centres de procés de dades de les empreses. Uns ordinadors molt costosos que eren compartits per molts usuaris i gestionats pels nous departaments d'informàtica que anaven apareixent en les empreses. En els inicis es programava amb targetes perforades amb unes màquines semblants a les màquines d'escriure, que l'ordinador llegia amb un lector de targetes semblant als comptadors de bitllets que tenen els bancs. El resultat de la computació normalment era impresa sobre paper per mitjà d'unes impressores connectades a aquests ordinadors centrals. Més tard ja van venir el que s'anomenaven *terminals*, amb una pantalla i un teclat per cada lloc de treball, i que estaven connectats a un ordinador central, anomenat *mainframe*. Diversos usuaris alhora ja podien interaccionar amb l'ordinador central per mitjà del teclat i podien visualitzar els resultats per pantalla. Tot un avenç.

Va ser cap a l'any 1980 quan va arribar el que es coneix com a *PC*[2] o **ordinadors personals**. Els ordinadors van deixar de ser un producte extremadament car, i es va fer possible que la informàtica arribés a l'abast de petites i mitjanes empreses, i fins i tot de particulars. Ara, les empreses podien disposar de tants PC com calgués per a suportar les diverses aplicacions, i això va comportar que acabessin tenint una sèrie de màquines distribuïdes per diversos departaments que gestionaven diferents aplicacions.

(2)Acrònim anglès de *personal computer*.

Tot i que podríem considerar que la infraestructura d'Internet, com avui dia la coneixem, va néixer a mitjan anys vuitanta, no va ser fins a mitjan anys noranta quan realment les empreses el van començar a usar. Sens dubte va representar un canvi molt important que va permetre fer créixer tant el nombre d'usuaris com el nombre d'aplicacions disponibles.

A partir d'això és quan van aparèixer webs com Yahoo, Google o Amazon, que donaven servei a un ampli ventall d'usuaris que es connectaven des de qualsevol lloc del món amb el propòsit de cercar informació, comprar o simplement navegar. És en aquest període quan neix el que s'ha anomenat ***Web 2.0***, en què els usuaris ja no solament són consumidors d'informació sinó que també són productors d'una informació que ara poden consumir la resta de persones per mitjà de la xarxa Internet. És el moment dels blogs, els wikis i l'inici de les xarxes socials.

Tot plegat fa aparèixer la necessitat de potents servidors que han de donar servei a una infinitat d'usuaris que s'hi connecten des de qualsevol tipus de dispositiu. Servidors que ja no es troben allotjats en les empreses que han creat els programes informàtics o els serveis. Han desplaçat tota la computació i emmagatzemament dels seus servidors cap a algun lloc exterior on puguin accedir fàcilment per mitjà d'Internet, en grans centres de procés de dades[3]. La Xarxa ha esdevingut el nostre ordinador. És el que es coneix com a **informàtica en el núvol d'Internet**. En els properes apartats aprofundirem una mica més en aquest nou paradigma de computació però n'avancem un breu resum esquemàtic en la taula 1, que compara els tres escenaris de computació que hem presentat en aquest punt.

(3)A partir d'ara usarem l'acrònim *CPD*.

Taula 1. Comparativa dels tres escenaris de computació i emmagatzemament

	Model de computació i emmagatzemament	**Característiques**	**Costos**
Ordinador central	Centralitzat	Sistemes que procuraven aprofitar al màxim els recursos a causa de l'alt cost d'aquests.	Inversió inicial tant de maquinari com de programari.
PC	Distribuït	PC i servidors distribuïts connectats per xarxa (primer local i després Internet).	Inversió inicial per compra de maquinari i costos de llicències de l'SO i les aplicacions programari que s'hi usaven.
Núvol	Centralitzat	Immensos CPD amb recursos TIC de baix cost (*commodity*) que són optimitzats aprofitant l'economia d'escala.	Es paga només pel que es gasta a mesura que es va usant al llarg del temps.

1.2. Definició de la informàtica en núvol

Sembla que el nom anglès ve dels termes *cloud* o núvol, que és el grafisme habitual que es fa servir per a representar Internet, i *computing* o computació, que es podria considerar que reuneix conceptes com informàtica i emmagatzematge. No hi ha una única definició però ens podríem quedar amb la que ha fet el NIST (Institut Americà d'Estàndards i Tecnologia):

> "La informàtica en núvol és un model que permet l'accés sota demanda per mitjà de la Xarxa a un conjunt compartit de recursos de computació configurables (com per exemple, xarxa, servidors, emmagatzematge, aplicacions i serveis) que poden ser ràpidament aprovisionats amb el mínim esforç de gestió o interacció del proveïdor del servei".

Aquest institut identifica diverses característiques essencials de la informàtica en núvol, indispensables per a veure la potència que aquest pot tenir. El primer que cal remarcar és que ha de permetre al consumidor proveir-se **unilateralment** dels serveis que necessiti i sense la interacció dels recursos humans del proveïdor de serveis mateix.

Lectura recomanada

Per a definir els conceptes bàsics que hi ha entorn de la informàtica en núvol s'acostuma a usar com a referència el document *The NIST definition of Cloud Computing* de l'Institut Americà d'Estàndards i Tecnologia (NIST). El podeu baixar en aquesta adreça:

http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf (document en anglès)

Actualment es parla de diferents models o nivells de servei, i els més habituals són tres:

- Infraestructura com a servei (IaaS[4])

(4) Acrònim anglès d'*infraestructure as a service.*

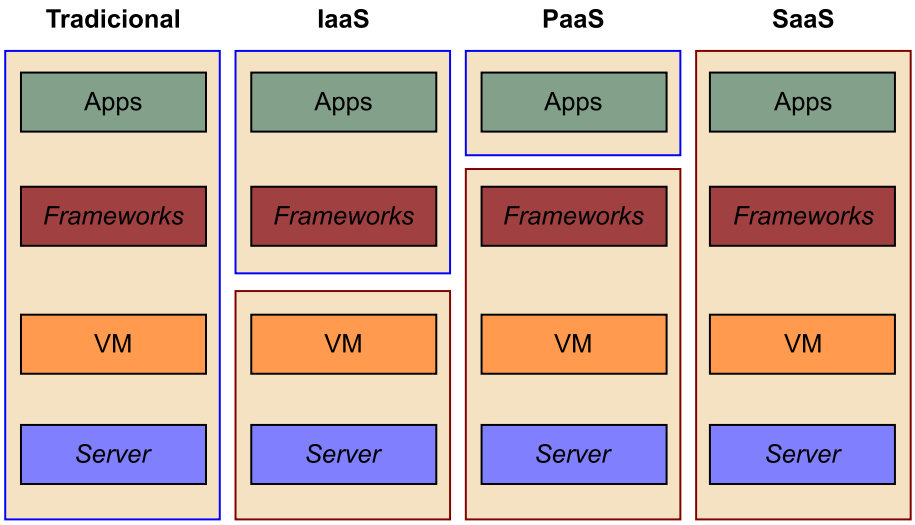
- Plataforma com a servei (PaaS[5])

(5) Acrònim anglès de *platform as a service.*

- Programari com a servei (SaaS[6])

(6) Acrònim anglès de *software as a service.*

Figura 1. Esquema dels tres nivells de servei de la informàtica en núvol habituals respecte al tradicional



1.2.1. Infraestructura com a servei

La **infraestructura com a servei** fa referència a oferir serveis de computació i emmagatzematge de tal manera que podem disposar de recursos com cicles de CPU, memòria, disc o equipaments de xarxa. El consumidor lloga els recursos maquinari en lloc de comprar-los i instal·lar-los en el seu CPD propi, i això li permet anar variant el consum dels recursos en funció de les seves necessitats, cosa que es coneix com a *elasticitat de la infraestructura*.

Imaginem que creem una aplicació web que té diferent nombre de clients per atendre en diferents franges horàries. Si sobtadament es necessiten més recursos, aquests es proveeixen puntualment fins que la necessitat desapareix.

Amazon web service

Un exemple molt citat per a il·lustrar aquest tipus de servei és el d'*Amazon web service* (AWS), amb el seu *elastic cloud computing* (EC2). A més de proporcionar la flexibilitat necessària per a escollir fàcilment el nombre, mida i la configuració de les instàncies de servidors que necessita l'usuari per a la seva aplicació, Amazon EC2 proporciona als clients diferents models que els permeten la flexibilitat necessària per a optimitzar els costos. Per exemple, les instàncies anomenades *sota demanda* permeten pagar una tarifa fixa per hora sense cap tipus de compromís, mentre que les instàncies anomenades *reservades* permeten abonar una tarifa d'entrada i en canvi rebre un descompte cada hora que s'usi.

Exemples europeus

Exemples propers, europeus, poden ser T-Systems, Flexiscale o CloudSigma, que proveeixen instàncies virtuals i una interfície per a engegar, aturar i accedir a la configuració i control del servidor virtual, de la mateixa manera que accedim a qualsevol servidor que tinguem en el nostre CPD.

Amazon

Amazon ofereix un equip d'ús gratuït d'AWS durant un any per a poder-se iniciar amb l'entorn. Està compost de 750 hores/mes d'instàncies de servidors micro, juntament amb diversos recursos d'emmagatzemament i ús de xarxa. Podeu trobar la informació a:

http://aws.amazon.com/es/free/

1.2.2. Plataforma com a servei

Ara bé, a vegades, el que cal és poder tenir una màquina que ja disposi d'un determinat programari que faciliti el desenvolupament i la implantació d'una nova aplicació d'una manera ràpida. Aquest és el cas del que es coneix com a *plataforma com a servei*.

En la **plataforma com a servei** el proveïdor ofereix quelcom més que la infraestructura, ofereix tot el necessari per a construir noves aplicacions i serveis, i facilita el cicle complet de construcció i posada en funcionament de les aplicacions.

Aquest nivell està molt orientat als equips de desenvolupament, i els permet reduir el temps de desenvolupament i implementació de les aplicacions, ja que no es requereix instal·lar eines programari de suport (s'estalvien problemes de versions o llicències), i facilita a la vegada la compartició i la interacció de les aplicacions.

Google app engine

Hi ha diversos exemples de plataforma com a servei, com AZURE, de Microsoft, però probablement el més popular i al vostre abast sigui el *Google app engine*, que ofereix serveis per a crear aplicacions dins de tota la infraestructura de Google. La idea és que no es necessita cap servidor, senzillament es tracta de carregar l'aplicació en l'espai de *Google app engine* perquè sigui accessible als vostres usuaris o clients, i us hi permeti accedir des del vostre domini propi. El *Google app engine* admet aplicacions escrites en diversos llenguatges de programació com Python, Java, JavaScript o Ruby. També segueix un model de pagar pel que s'usa.

Google

Google ofereix un kit gratuït del seu *app engine* de fins a 500 MB d'emmagatzemament amb els cicles de CPU i amplada de banda de la xarxa necessaris per a permetre un nombre molt elevat de visites al més. Podeu trobar la informació a:

https://developers.google.com/appengine/docs/whatisgoogleappengine?hl=es

1.2.3. Programari com a servei

En el cas del **programari com a servei** el proveïdor no solament ofereix la infraestructura maquinari i els entorns d'execució necessaris, sinó també els productes programari que interaccionen amb l'usuari amb un determinat portal o interfície per mitjà d'Internet.

Podríem dir que és l'oposat d'un programari com a producte, i permet convertir el cost d'ús en una variable més del negoci i evita les habituals inversions inicials en llicències. Per exemple, no és necessari comprar el programari o les seves llicències, ni cap tipus de maquinari, i es deixa per al proveïdor el manteniment i el suport, tant del programari, com del maquinari. El preu per ús, en general, resulta molt més baix que si una empresa hagués d'adquirir el mateix servei per si sola, a causa de l'economia d'escala que apliquen els proveïdors.

Ara mateix hi ha un nombre molt gran de serveis d'aquest nivell amb tot tipus d'aplicacions destinades a diferents àrees de l'empresa, com la gestió de recursos humans o la gestió financera. També es poden trobar un conjunt molt reeixit de serveis d'eines d'escriptori que inclouen els habituals fulls de càlcul o processadors de text, entre d'altres.

Gmail i Hotmail

Alguns dels exemples més populars de programari com a servei són el Gmail o el Hotmail que, en aquest cas, fins i tot són gratuïts per a l'usuari final. Aquest és probablement el model de servei més conegut per tots vosaltres.

Ara bé, a banda d'aquests tres tipus de serveis de la informàtica en núvol comentats, caldria esmentar que el mercat de la informàtica en núvol està madurant molt ràpidament i en aquests moments ja s'estan oferint altres productes "com a servei" i, sobretot, s'està preparant el núvol per a oferir-ne molts més.

Lectura complementària

Un document en què es descriuen en detall les diferents modalitats de núvol i alhora un bon punt de partida per a conèixer l'estratègia europea sobre la informàtica en núvol és l'informe de la Comissió Europea *The Future of Cloud Computing*, que podeu baixar de l'adreça següent:

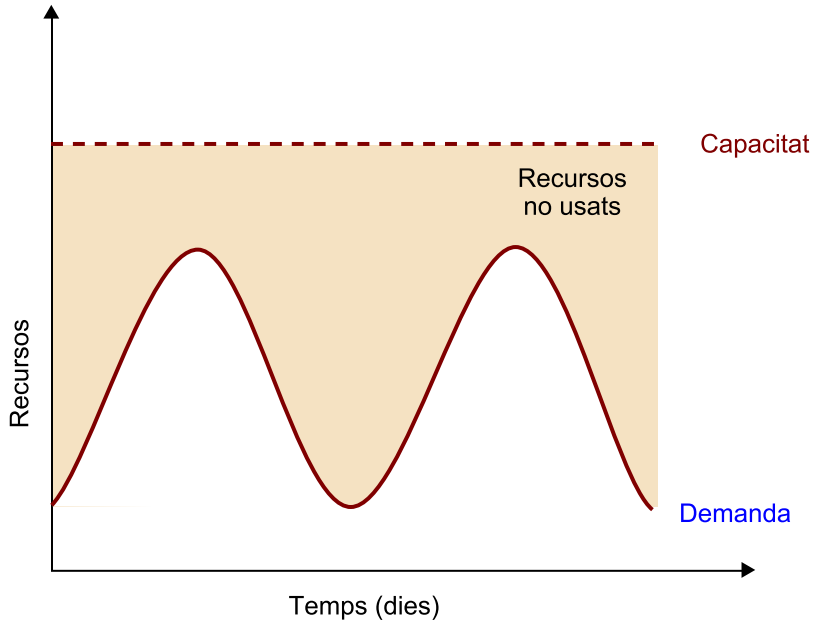
http://cordis.europa.eu/fp7/ict/ssai/docs/cloud-report-final.pdf (document en anglès)

1.3. La flexibilitat de la informàtica en núvol

Amb la informàtica en núvol apareixen un ventall de nous serveis, alhora que permet convertir despeses fixes en variables, coneixent millor els costos reals de cada producte i minimitzant alhora els riscos financers en el llançament de nous productes o serveis que depenguin de les TIC. Una de les raons principals per a adoptar-lo té a veure amb l'**agilitat** amb què ens podem autoproveir dels serveis oferts per mitjà de la informàtica en núvol. Això comporta que els terminis d'implantació d'una nova necessitat TIC (la gestió dinàmica dels serveis, dels recursos de càlcul, d'emmagatzematge o d'accés a la xarxa) es pugui veure reduïda significativament i passi a minuts o hores, en lloc de setmanes o mesos (que és el que ens representaria en un CPD propi). És a dir, les infraestructures o serveis TIC deixaran de ser el coll d'ampolla en la posada en marxa de nous productes o serveis i ens permetran centrar-nos en l'aplicació de negoci concreta que hi ha davant de qualsevol nou projecte.

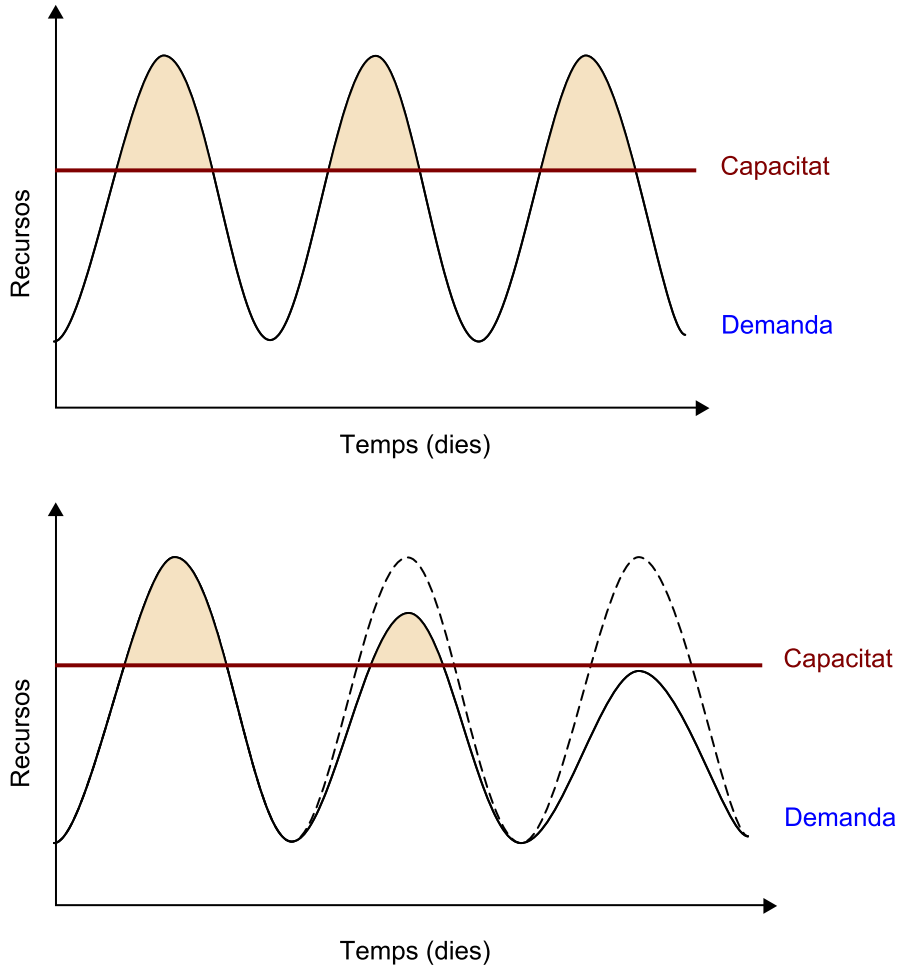
Un altre aspecte rellevant que presenta la informàtica en núvol és el que s'anomena l'***escalabilitat***. Fins ara, si l'empresa volia créixer en oferta, havia d'invertir primer en els seus recursos informàtics propis (tant maquinari com programari), i en cas contrari es podria arribar a col·lapsar el sistema. Però, per exemple, en el supòsit que les vendes tinguin un cicle de temporada o basat en promocions especials, cosa que implica pics i valls enormes en l'ús dels recursos TIC, pren molt d'interès poder disposar d'un servei flexible. Sense la informàtica en núvol la dificultat de la previsió de la demanda fa que s'aprovisionin uns sistemes TIC en excés per tal de garantir que sempre es puguin atendre totes les sol·licituds.

Figura 2. Esquema de demanda de càrrega variable coberta per suficients recursos que puguin absorbir les puntes de feina.



En la figura 2 es poden observar de manera esquemàtica les necessitats de recursos que una empresa requereix per tal de suportar una determinada càrrega de clients. La demanda de recursos és molt variable i proporcional a les peticions dels clients, i fem la suposició que està relacionada amb l'hora del dia en què es produeixen. La gràfica representa esquemàticament l'arribada de peticions en el sistema (i el requisit de recursos equivalent) en funció del moment del dia, i és alta durant el dia i baixa durant la nit (una visió simplificada del comportament habitual en els portals web). Per tal de cobrir una demanda variable es capacita el sistema amb un nombre de recursos que permeti absorbir les puntes de feina, no fàcilment previsibles. Això provoca una pèrdua de recursos en el sentit que no són usats, representada amb la zona ombrejada. Pensem que els servidors, a més del cost de compra, tenen un cost de consum d'energia pel sol fet d'estar engegats, encara que no estiguin atenent peticions.

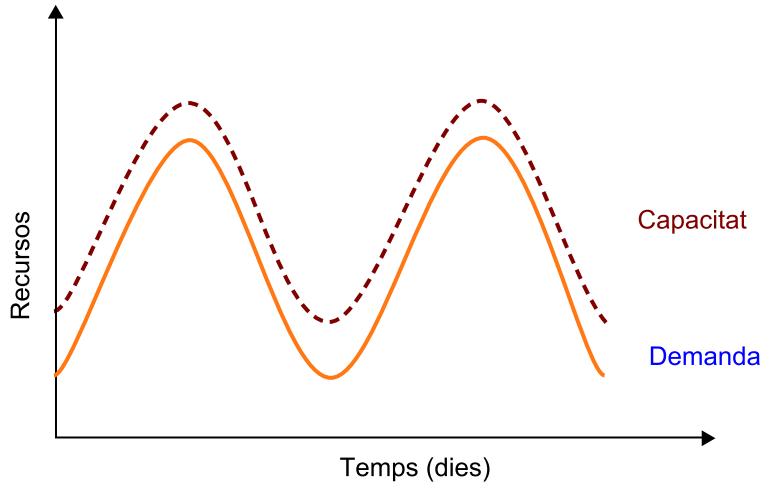
Figura 3. Comportament dels clients davant d'una reducció de recursos disponibles per a atendre'ls.



Si s'optés per reduir la provisió de recursos per tal de minimitzar aquest malbaratament, tal com es representa en la primera part de la figura 3, aleshores s'estarien sacrificant guanys potencials provinents de clients no atesos (tots aquells que no poden ser atesos per la falta de recursos en les puntes de feina). Es podria plantejar que la pèrdua de no atendre aquesta petita porció de clients compensa la reducció del cost de malbaratament. Però, habitualment, els usuaris de serveis per mitjà d'Internet abandonen un lloc de manera permanent després d'experimentar un mal servei. És a dir, la demanda s'adapta ràpidament a la capacitat del sistema, i redueix de manera definitiva els nostres clients i, per tant, una part del flux d'ingressos, com es mostra esquemàticament en la segona part de la figura 3.

La **elasticitat** o **escalabilitat** que ofereix la informàtica en núvol en la disponibilitat de recursos permet adaptar-se a la demanda real, i redueix el risc que aquests quedin infrautilitzats durant els períodes menys productius, i alhora és capaç de suportar pics de treball imprevisibles, com es mostra esquemàticament en la figura 4.

Figura 4. Esquema de com l'elasticitat de la informàtica en núvol permet adaptar-se a la demanda.



És per tot això que la propietat d'elasticitat o escalabilitat de la informàtica en núvol és clarament una oportunitat per a adaptar-se a la demanda i eliminar aquest sobrecost derivat de mantenir els sistemes infrautilitzats fora de les puntes de demanda.

Lectura recomanada

Un dels documents tècnics molt citats sobre la informàtica en núvol és l'informe que va fer un grup de recerca de Berkeley, *Above the Clouds: A Berkeley View of Cloud*; encara que del 2009, continua essent de gran valor per les seves explicacions didàctiques del nou paradigma de computació. El document es pot baixar en l'adreça següent:

http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf (document en anglès)

1.4. Consum energètic i dependència de la informació

El veritable cor de la informàtica en núvol es troba en els grans **centres de procés de dades (CPD)**, on l'energia s'ha convertit avui dia en el principal cost per a fer-los funcionar. Afortunadament els avenços en maquinari i programari permeten millorar el consum i gestió en aquests CPD. Però malgrat aquestes millores que han reduït l'electricitat requerida per aquests CPD, aquests consumiran cada vegada més megawatts de potència dels que consumeixen actualment. Pensem que els ciutadans cada vegada van emmagatzemant més informació en aquest núvol, suportat per grans CPD dispersos per tot el món, molts de la mida de diversos camps de futbol. Us imagineu una nau industrial d'aquesta mida plena de maquinari i programari?

Internet no té fronteres com els països, cosa que implica que la nostra informació pot acabar essent deslocalitzada a l'altre costat del planeta en un d'aquests CPD, les factories d'informació del segle XXI, en què el cost de l'energia pot ser molt més baix (pensem que hi ha llocs on el cost de l'energia pot ser 5 vegades més petit que el que costa a Barcelona). Sorgeixen llavors diverses preguntes: està garantida la disponibilitat de la nostra informació amb aquesta deslocalització? Ens podríem trobar algun dia que no tinguem accés a les nostres dades?

La resposta la tenim amb l'exemple del tancament per part del Departament de Justícia dels EUA i l'FBI del popular servei d'intercanvi d'arxius Megaupload al principi del 2012, a causa de la descàrrega dels seus continguts, dels quals una

part important pertanyien a usuaris que no eren els titulars dels drets. Però no oblidem que també una part dels seus continguts emmagatzemats eren lícits, d'empreses i usuaris que tenien tots els drets sobre les seves dades.

Què va passar amb aquestes empreses que no van poder accedir als seus continguts de Megaupload? Us imagineu el que us passaria si els vostres comptes de Dropbox deixessin de ser accessibles? O que el servei de Gmail us deixés de funcionar? Us heu parat a pensar on són les vostres dades i a qui us hauríeu de dirigir per a poder recuperar-les?

En la societat actual, la informació és essencial i per això hem d'evitar córrer el risc de perdre-la per esdeveniments que es troben fora del nostre control. El motiu en el cas de Megaupload va ser perseguir uns continguts il·lícits d'altres usuaris, però en l'escenari actual es pot perdre l'accés a la nostra informació per moltes altres causes. Per a garantir l'accessibilitat de la informació, hem de desar la nostra part d'aquest núvol dins d'un territori, on es pugui tenir un control sobre aquestes factories de producció d'informació, que ja comencen a ser un servei públic d'interès general, com pot ser l'aigua o l'electricitat.

Tanmateix, aquestes fàbriques d'informació requereixen molts milers de quilowatts d'energia per a funcionar, de manera que si aproximadament el 75% del sistema energètic global a Espanya és importat, ens trobem igualment davant una situació de dependència. Per això, la nostra posició és que no ens queda un altre camí que augmentar la producció d'energia renovable (en realitat l'única energia que és autòctona) que es genera dins de l'Estat per a assegurar el manteniment d'aquestes noves fàbriques d'informació de manera que no es depengui de l'energia que ve de fora i que ningú no ens garanteix que sempre sigui així.

Una possibilitat que s'està valorant a escala de recerca és moure aquestes fàbriques d'informació als parcs d'energia renovable, de manera que aquests parcs no solament seran proveïdors d'energia, sinó també proveïdors d'informació, que no deixa de ser l'energia que necessita la nostra societat de la informació en què estem ja immersos. Aquesta proposta és molt eficient en diversos aspectes; per exemple, la transferència de dades a través de fibra òptica des d'aquests parcs té un cost energètic mínim en comparació de les pèrdues que provoca el transport d'energia, que pot arribar al 10%. Encara que, a causa de la discontinuïtat de l'energia renovable, es requereix ara mateix energia de suport. Les investigacions que s'estan duent a terme en aquests moments en centres de recerca com el Barcelona Supercomputing Center (BSC) i la Universitat Politècnica de Catalunya (UPC), en col·laboració amb altres centres, ens fan ser optimistes i pensar que es podrà mantenir la informació en gran part només amb l'energia renovable.

Per exemple, basant-se en la predicció del temps, és possible anar movent càrregues de computació entre diferents parcs, en funció de l'energia que cada parc tingui disponible en aquell moment, o retardar automàticament l'execució de certs treballs fins que es torni a tenir energia disponible.

No obstant això, no podem ignorar que una part considerable de l'energia consumida per les TIC és en realitat a causa d'un gran nombre de petits CPD en petites i mitjanes empreses. Aquestes instal·lacions van des d'una dotzena de servidors allotjats en una sala de màquines a diversos centenars de servidors en el cas d'empreses més grans. Per a aquest tipus d'entorns avui dia, les investigacions van en la línia de construir contenidors (com els d'ús per a transportar mercaderies marítimes) però de mida menor, que serien acompanyats en les instal·lacions per plaques fotovoltaiques o petits aerogeneradors per a proveir-se d'energia, i alhora estarien connectats a la xarxa elèctrica com a suport quan l'energia renovable no estigui disponible.

El repte de la recerca actual de proveir els CPD amb energia solar i eòlica és que no sempre aquesta es troba disponible a causa de la seva discontinuïtat, si bé se'n poden usar d'altres com la biomassa. Un tema al qual encara queda molt de recorregut i en què encara hi ha moltíssimes oportunitats en recerca i innovació.

Lectura complementària

Més detall de propostes tècniques des del nostre grup de recerca en aquest àmbit es poden trobar en l'article "Towards Sustainable Solutions For European Cloud Computing" de la revista *UPGRADE*, que publica l'Associació Europea de Professionals Informàtics. L'article es pot baixar en l'adreça següent:

http://www.cepis.org/upgrade/media/le_2011_41.pdf (document en anglès)

Lectura recomanada

El capítol espanyol de la Cloud Security Alliance va publicar l'informe Cloud Compliance Report en el qual es tracten amb una orientació pràctica qüestions com la protecció de dades o els contractes relacionats amb el núvol. L'informe es pot baixar a:

http://cloudsecurityalliance.org/csa-news/csa-issues-first-cloud-compliance-report-for-spain/ (document en anglès)

1.5. Els inhibidors de la informàtica en núvol

Com hem vist en el subapartat anterior, Internet no té fronteres i per això la nostra informació pot acabar deslocalitzada a l'altre extrem del món i provocar **problemes de disponibilitat**. Però no solament és aquest un dels temes que amoïnen els usuaris del núvol, sinó que també hi ha altres problemes que preocupen i que han frenat fins ara el desenvolupament del núvol. Entre aquests, els dubtes sobre la **seguretat**, **privacitat** i **integritat** de les dades al núvol figuren entre les preocupacions més freqüents. També s'alerta la manca d'estàndards que ens permetin un canvi fàcil de proveïdor. O la maduresa dels proveïdors en aquest àmbit no sempre està prou acreditada. També la legislació encara no del tot homogènia en els diferents països, alhora que es requereix més definició en aspectes com el control i la recuperació de les dades.

Per a prevenir-se davant aquests problemes i, fins i tot, anar superant les seves pors pròpies no sempre fonamentades (com pensar que les dades estan més segures en un servidor en l'empresa mateixa que en el servidor del proveïdor), moltes empreses estan optant pel núvol híbrid, que combina el núvol (entès com a núvol públic) i el núvol privat per a la distribució dels recursos d'una empresa.

S'entén per **núvol privat** quan s'habiliten els CPD propis de l'empresa amb les darreres tècniques de gestió que permeten fer un ús eficient dels seus recursos per a oferir serveis a l'empresa mateixa.

Malgrat que no es disposa d'escalabilitat quasi infinita, d'aquesta manera es disposa d'uns alts nivells de seguretat i confidencialitat, i s'assegura el compliment de la legislació pròpia (com pot ser la llei LOPD espanyola).

Una solució òptima és portar al núvol únicament els pics d'activitat, per no haver de sobredimensionar les màquines (i recursos en general) ni provocar col·lapses en el servei. Deixem aquí aquest tema per motius d'espai i l'enfocament d'aquest material, però sense cap dubte és un tema molt extens i de gran importància per a la implantació del núvol.

1.6. Un futur mòbil en què tothom i tot estarà connectat

El fenomen núvol coincideix en el temps amb la difusió massiva dels mòbils.

La informàtica en núvol fa possible un futur amb extrema mobilitat en el qual tothom estigui connectat.

Sense núvol no seria possible, per exemple, emmagatzemar i manejar les dades dels més de 900 milions d'usuaris de Facebook, dels quals una part important es connecten per mitjà dels seus mòbils.

S'estima que el 2012 seran més de 15 milions els usuaris de la xarxes socials a Espanya. És clara la tendència que això marca, malgrat els seus problemes de seguretat i privacitat, i per això les empreses ja consideren les xarxes socials un mitjà eficaç per a la comercialització i la comunicació dels seus productes. No hi ha dubte que a partir d'ara les empreses només poden augmentar el seu posicionament i visibilitat en aquestes xarxes. I encara que ara, amb la diversitat de xarxes socials, aquesta presència no és fàcil de gestionar, segur que s'imposaran nous sistemes i eines integrades que permetran gestionar de manera més fàcil l'administració de les dades i processos d'una empresa en tot aquest ampli entramat de les xarxes socials.

S'ha de tenir en compte que des del final del 2010 ja es fabriquen més dispositius mòbils que PC. El centre de gravetat s'està desplaçant ràpidament del PC a un altre tipus de tecnologies mòbils. Aquests dispositius, connectats al núvol, crearan una nova relació entre els ciutadans i els seus dispositius. Per a millorar la qualitat de la interacció amb l'usuari es podrà utilitzar informació sobre el context i les preferències de l'usuari com serveis de localització, realitat augmentada en dispositius mòbils o comerç mòbil. Les empreses es podran anticipar a les necessitats de l'usuari, i de manera proactiva servir productes o serveis més apropiats i personalitzats (per exemple, el màrqueting i comercialització basats en la geolocalització per a trobar ofertes en temps real o el basat en codis promocionals, ofertes o descomptes, etc.).

Tot plegat està comportant una mutació dels continguts per tal d'arribar a l'usuari. Per exemple, per a molts usuaris les aplicacions dels mòbils estan resultant més senzilles d'entendre que una web, i els permeten accedir a una funcionalitat de manera còmoda, ràpida i fàcil un cop instal·lades al dispositiu. Per això l'imparable creixement d'aquestes "*apps*" representa un retrocés de l'ús del Web des dels mòbils, que, com hem vist, cada vegada són més. En conseqüència, el lloc web tradicional anirà deixant de ser el nucli de l'estratègia de visibilitat i eix de negoci de les empreses. El futur implica apostar pel contingut digital amb altres mètodes i dispositius. A més, més enllà de la modernització del Web amb nous estàndards com l'HTML5 o el CSS3, la informació mateixa que genera l'usuari està prenent altres formats més dinàmics. Per exemple, cada minut es generen 50 hores de contingut a YouTube, una mostra del creixement del trànsit de reproducció en temps real davant d'altres tipus de continguts.

Però encara és més, en un futur no gaire llunyà fins i tot els mòbils deixaran de ser la manera habitual de consumir la informació continguda en la informàtica en núvol i es farà per mitjà d'altres dispositius, com per exemple les ulleres de realitat augmentada amb connexió a Internet que Google va anunciar fa uns mesos que estava dissenyant. Unes ulleres que a partir de la veu de l'usuari permeten fer fotos, iniciar videoxats, mostrar adreces, demanar previsions meteorològiques, estar connectat a la nostra xarxa social, entre desenes d'opcions més.

Aviat el núvol podrà posar en contacte, a més dels milions de persones que l'utilitzen per mitjà del seu ordinador o dispositiu mòbil, també els milers de milions d'objectes que actualment encara no ho estan. Qualsevol objecte es podrà transformar en un objecte amb capacitat de comunicar-se en qualsevol moment i en qualsevol lloc. És el que s'anomena ***Internet of things*** (Internet de les coses), que implica que tot objecte pugui ser una font de dades i que qualsevol cosa pugui ser monitorada a través del temps i de l'espai. Es tracta de xips de baix cost incrustats en gairebé qualsevol objecte de consum (objectes domèstics com la nevera o els cotxes mateixos). Tot això generarà noves àrees de negocis, com ara la de les *smart cities*. Però sobretot representarà que els productes es converteixin en serveis. Us imagineu que un fabricant de motors d'avions posi sensors en els seus equips per a permetre, en lloc de vendre un motor, llogar-los per hores de vol? Tot un canvi de panorama per als serveis que empreses com Roll Royce, fabricant de motors d'avió, ja està fent.

Lectura complementària

La Fundación de la Innovación Bankinter ha publicat l'informe *Internet de las cosas*, en què es descriu l'estat de l'art en aquesta tecnologia emergent. El document pot ser baixat en l'adreça següent:

http://www.fundacionbankinter.org/system/documents/8168/original/XV_FTF_El_internet_de_las_cosas.pdf (document en castellà)

1.7. La informàtica en núvol com a motor de canvi

El núvol és aquí i ha vingut per a quedar-se. El que veníem intuint en els últims anys es veu confirmat amb l'aparició diària de nous serveis en el núvol o les dades d'estudis com els de CISCO, que ens diuen que en un parell d'anys l'origen de tot el trànsit de la Xarxa tindrà el seu origen en la informàtica en núvol.

Com a fórmula eficient per al subministrament de serveis, el model d'informàtica en núvol (especialment la seva pedra angular, la infraestructura com a servei) ofereix un gran estalvi de costos, redueix dràsticament l'espai requerit en els CPD i ofereix més simplicitat de les operacions TIC. El que ha permès la informàtica en núvol és bàsicament la industrialització de les TIC, en un procés similar al que es va viure amb la generació elèctrica. Al principi del segle passat, les empreses generaven la seva pròpia energia per a autoabastir-se, fins que van sorgir les companyies elèctriques i les economies d'escala del nou model de subministrament van superar els avantatges de la producció pròpia.

Malgrat alguns temes de seguretat o privacitat, entre molts altres, que encara han de millorar en la informàtica en núvol, no hi ha dubte que aquest obre un nou model de TIC flexibles, que s'adapten a cada empresa i, una cosa molt important, en què els costos són predictibles. Entre els avantatges que ofereix, destaca la possibilitat d'accés des de tot arreu, la fórmula de pagament per ús, l'autoservei a la carta o la ràpida elasticitat.

Ara bé, si una empresa es limita a adoptar el model d'informàtica en núvol només per estalviar diners, estarà errant l'estratègia probablement. L'avantatge diferencial de la informàtica en núvol és que permet fer realitat nous models de negoci, provar processos de negoci i efectivitat operacional en el moment que ho necessitem i amb la potència que requereixin els nostres projectes. Pensem que val el mateix llogar una màquina 1.000 hores que 1.000 màquines una hora, per exemple. Això obre un gran ventall a les empreses per a provar i assajar conceptes o processos. La mateixa oportunitat s'estén als emprenedors i empreses més petites, alliberades de grans inversions de capital gràcies al model de pagament per ús descrit.

La informàtica en núvol a més, ha comportat que telèfons intel·ligents i altres dispositius de consum massiu hagin desplaçat el PC del centre del món digital. La informàtica en núvol està donant pas a una nova era que ofereix als usuaris un nivell de flexibilitat amb els dispositius per a la realització de les activitats diàries. Que, a més, amb noves maneres d'interactuar-hi, com és el reconeixement tàctil i gestual, juntament amb la identificació de la parla i el coneixement contextual, permeten una rica interacció entre l'usuari i els dispositius.

Tot plegat acompanyat del que es coneix com a *Internet of things*, per mitjà de moltes empreses que ja hi estan treballant. I és que gairebé qualsevol cosa es pot convertir en digital, interconnectant i integrant les seves dades amb les de la resta que captin els altres dispositius, per a poder, per tant, fer un tractament a temps real. No hi ha dubte que a mesura que més objectes tinguin sensors incrustats i més possibilitat de comunicar, es generarà una xarxa de connexions que farà que hi hagi més informació disponible ara que en tota la història de la humanitat. L'ús intel·ligent d'aquesta informació tindrà la

Lectura complementària

El llibre que de manera divulgativa explica tot el fenomen de la informàtica en núvol és:

Jordi Torres i Viñals (2011). *Empreses en el núvol, claus per entendre la Internet Global*. Barcelona: Libros de Cabecera. Disponible a:

http://www.librosdecabecera.com/empreses-en-el-nuvol

capacitat de generar una transformació de la societat mai vista fins ara, que, en el pla dels negocis, representarà la creació de nous models, i millorar els existents amb la reducció de costos i riscos de tots.

2. *Big data*

2.1. S'acosta una marea d'informació digital

Ja hem explicat que la *Internet of things* generarà una ingent quantitat de dades des de qualsevol tipus de dispositiu o sensor.

Per fer-nos una idea de com de fàcil és produir dades, fixem-nos en les xarxes socials: per exemple, Twitter, que gestiona més de 90 milions de piulades per dia, que representen un total de 8 terabytes de dades per dia. O les dades transaccionals de les aplicacions web, que han crescut a una gran velocitat. Avui, Wal-Mart, el detallista més gran del món, gestiona un milió de transaccions dels clients per hora, que alimenta una base de dades que s'estima de 2,5 petabytes. I no oblidem el món científic, en què, per exemple, el col·lisionador de partícules del CERN pot arribar a generar 40 terabytes de dades per segon durant els experiments.

Boeing

Per exemple, els motors a reacció de Boeing que hem comentat abans poden produir 10 terabytes d'informació cada 30 minuts. És a dir, un jumbo de 4 motors pot crear 640 terabytes de dades en un vol transoceànic.

Nota

1 terabyte (TB) = 1.000 gigabytes (GB)

1 petabyte (PB) = 1.000.000 gigabytes (GB)

1 exabyte (EB) = 1.000.000.000 gigabytes (GB)

1 zettabyte (ZB) = 1.000.000.000.000 gigabytes (GB)

Tot plegat comporta un creixement exponencial de la informació disponible i ens porta a la gran tendència següent, que dominarà la indústria TIC en els propers anys, que es coneix com a ***big data***. Segons alguns estudis el 2015 circularan pel planeta 7.910 exabytes d'informació (el 2005 eren 130 exabytes). En definitiva, és molt fàcil generar terabytes de dades. Una altra cosa és processar-les i transformar-les en informació de valor. El problema ja no és trobar dades, sinó saber què se'n fa. Per això a continuació avancem els reptes i oportunitats que presenta el *big data*.

Lectura complementària

Un bon punt de partida per a saber l'abast del fenomen *big data* el podem trobar en l'informe de McKinsey, titulat *Big data: The next frontier for innovation, competition, and productivity*, que podeu baixar de l'adreça següent:

http://www.mckinsey.com/Insights/MGI/Research/Technology_and_ Innovation/Big_data_The_next_frontier_for_innovation (document en anglès)

2.2. Definició del *big data*

Igual que passava amb la informàtica en núvol, no hi ha una única definició de consens per a aquest nou fenomen que anomenem *big data*.

En aquest material considerarem ***big data*** tot allò que fa referència al fet que les dades han esdevingut tan grans que no es poden processar, emmagatzemar i analitzar mitjançant mètodes convencionals.

Una manera de caracteritzar aquestes dades que es fa servir és usant el que en diuen les *3 V*, en referència a *volum*, *varietat* i *velocitat*:

- **Volum**: l'univers digital continua expandint les seves fronteres i s'estima que ja hem superat la barrera del zettabyte.

- **Velocitat**: la velocitat a la qual generem dades és molt elevada i la proliferació de sensors n'és un bon exemple. A més, les dades en trànsit –dades amb vida efímera però amb alt valor per al negoci– creixen més ràpid que la resta de l'univers digital.

- **Varietat**: les dades no solament creixen sinó que també canvien el seu patró de creixement, i ara és el contingut desestructurat.

A vegades també s'afegeix una altra *V*, la de **valor**. Extreure valor de tota aquesta informació marcarà la pròxima dècada. El valor el podrem trobar de diferents maneres: millora en el rendiment del negoci, noves fonts de segmentació de clients, automatització de decisions tàctiques ...

Com ja hem vist, l'origen de les dades per a una empresa pot ser divers. Per exemple, li poden venir dels seus propis sistemes d'informació de suport a les vendes o d'interacció amb els seus clients. O de dades originades per màquines o sensors encastats en qualsevol tipus de dispositiu o producte de l'empresa. I no oblidem la informació que circula per les xarxes socials i que fa referència a una determinada empresa, informació sens dubte molt valuosa.

Però un altre origen molt important de les dades són les plataformes de dades que estan obrint diversos governs. Aquestes dades públiques poden ser informes, mapes, estadístiques, estudis, anàlisis, que crea i gestiona l'administració a tots els nivells (sanitat, economia, educació, població, etc.) i que són de gran interès públic.

És el que es coneix com a ***open data***, que persegueix que les institucions públiques exposin les dades públiques que estan en el seu poder de manera reutilitzable, per tal que tercers en puguin crear serveis derivats. Com a conseqüència, els conjunts de dades exposades s'ofereixen sota llicències de propietat obertes, que en permeten la redistribució, reutilització i aprofitament amb fins comercials.

L'*open data* representa també el desig que les empreses alliberin dades per a permetre el desenvolupament de noves aplicacions i usos. Però, en el moviment *open data*, el paper protagonista és el de ciutadà. L'usuari connectat a la Xarxa és una gran font d'informació. En compartir el que veiem (informació de trànsit, opinions, accidents, etc.), on som (geolocalització) o les imatges que captem, estem construint, entre tots, una gran intel·ligència col·lectiva.

Referència web

Hi ha projectes d'*open data* que s'estan portant a terme a Euskadi, Catalunya, Astúries o Navarra, i des d'ajuntaments com el de Barcelona, Saragossa, Lleida, Badalona, Gijón o Còrdova. Recomanem visitar:

http://datos.fundacionctic.org/sandbox/catalog/faceted

2.3. Calen noves tecnologies d'emmagatzemament

L'escenari que fins ara hem tingut en l'emmagatzemament i accés a dades de les aplicacions és que els servidors usen memòria RAM[7] per a la informació que necessita sovint, i usa l'emmagatzemament en disc per a la resta de dades que l'aplicació no usa en aquell instant. El fet de no tenir-ho tot a la memòria és perquè és cara i, per tant, habitualment n'hi ha poca. En canvi, les unitats de disc dur (HDD[8]) presenten una alta capacitat molt més assequible, malgrat presentar un rendiment molt més baix que la memòria. Per això habitualment en un servidor només hi ha memòria per a contenir la fracció de dades que el processador necessita per a treballar. A més de ser cara la memòria, també és volàtil, cosa que significa que si un servidor cau o es queda sense electricitat, la informació emmagatzemada en la memòria es perd. A causa d'aquest fet tota la informació es manté en disc durant la seva vida útil, ja que aquests mantenen la informació encara que apaguem l'ordinador o el desendollem de la corrent.

(7)Acrònim anglès de *random access memory*.

(8)Acrònim anglès de *hard disk drives*.

En aquests moments l'emmagatzemament d'informació en HDD és 100 vegades més barat que fer-ho en la memòria RAM, però 1.000 vegades més lent d'accés. Per a solucionar-ho s'estan començant a usar com a sistema d'emmagatzemament tecnologies de memòria no volàtil, com les usades fins ara en dispositius mòbils, que s'anomenen *discos d'estat sòlid* (SSD[9]).

(9)Acrònim anglès de *solid-state drive*.

SSD

Els **SSD** ja es troben en el mercat i són una opció per a emmagatzemar dades que requereixen un temps d'accés molt ràpid. Un bon punt de partida per a conèixer-ne més detalls pot ser la Wikipedia:

http://en.wikipedia.org/wiki/Solid-state_drive (document en anglès)

Memòria no volàtil

Un exemple en són els xips de memòria flaix que, de la mateixa manera que els HDD, són no volàtils i mantenen la informació quan es queden sense alimentació elèctrica.

Els **SSD** semblen una opció amb futur si tenim en compte que el preu de la memòria flaix està baixant i que a més, en ser bàsicament un xip de semiconductors, té un consum energètic molt més petit que els discos HDD, que usen components mecànics per a llegir els discos. Malgrat això, per a certes aplicacions que requereixin reescriure moltes vegades una mateixa dada, cal tenir en compte que aquests dispositius encara envelleixen més ràpid que els HDD i, per tant, no són ara mateix una solució per a qualsevol tipus d'aplicació.

L'aparició d'aquesta solució a la persistència de l'emmagatzemament de les dades per mitjà de tecnologies de memòria no volàtil portarà en breu moltes millores d'emmagatzemament. Com a memòries que són, s'està fent recerca per a connectar-les al bus intern dels ordinadors que connecta amb el processador. D'aquesta manera es permetrà al processador accedir a les dades persistents per mitjà de les instruccions de memòria *load/store*, per a oferir un accés simple

i ràpid, i es reduirà encara molt més el temps d'accés a l'emmagatzemament persistent. És una recerca que es coneix com a *storage class memory*, que tracta de dissenyar un sistema d'emmagatzematge de capacitat i economia similars als HDD, però amb unes prestacions de rendiment equivalents a la memòria. Actualment és una de les àrees de recerca més actives, que sens dubte tindrà un impacte molt important en el món del *big data*, i permetrà accelerar l'execució de les aplicacions intensives en dades.

2.4. Les bases de dades relacionals no serveixen per a tot

Fins ara l'escenari de les bases de dades ha estat marcat pel que es coneix com a *SQL*[10] (llenguatge de consulta estructurat). En el mercat trobem moltes opcions consolidades com DB2, Oracle, MySQL, Informix, Microsoft SQL Server, PostgreSQL, etc. Aquestes bases de dades, que s'acostumen a anomenar *relacionals*, segueixen les regles ACID, fet que els permet garantir que una dada sigui emmagatzemada de manera inequívoca i amb una relació definida sobre una estructura basada en taules que contenen files i columnes.

(10)Acrònim anglès de *structured query language*.

Regles ACID

En el context de bases de dades, **ACID** (acrònim anglès d'*atomicity, consistency, isolation, durability*) són tot un seguit de propietats que ha de complir tot sistema de gestió de bases de dades per tal de garantir que les transaccions siguin fiables. Per a entrar en una mica més de detall podeu consultar la Wikipedia:

http://ca.wikipedia.org/wiki/ACID

Però en el món del *big data* apareix el problema que les bases de dades relacionals no poden manejar la mida, la complexitat dels formats o la velocitat de lliurament de les dades que requereixen algunes de les aplicacions d'avui dia, com per exemple aplicacions en línia amb milers d'usuaris concurrents i milions de consultes al dia.

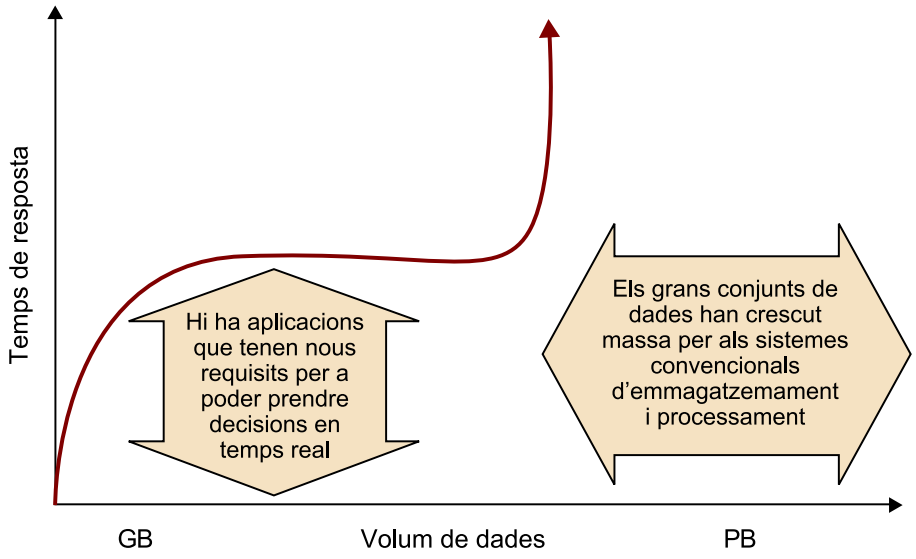
En la figura 5 es presenta esquemàticament el comportament dels sistemes d'emmagatzemament actuals. Podríem resumir que tenen dues grans limitacions:

1) Per una banda, no poden atendre els requisits que tenen avui dia algunes aplicacions, com les de tenir un temps de resposta molt baix. En aquests casos ara mateix s'estan posant les dades en memòria (que es coneix com a *in-memory*). Aquest tipus d'aplicacions es veuran beneficiades de les millores que es preveuen en emmagatzemament comentades en l'apartat anterior.

2) I per altra banda apareix el problema que determinades aplicacions requereixen un volum de dades que ja no poden ser emmagatzemades i processades usant bases de dades tradicionals.

Per això han sorgit noves varietats de bases de dades (anomenades *NoSQL*) que permeten resoldre els problemes d'escalabilitat i rendiment que presenta el *big data*. NoSQL aglutina les diferents solucions de bases de dades centrades a ser no relacionals, distribuïdes i escalables de manera horitzontal. Hi ha nombrosos productes disponibles, molts de codi obert, com el Cassandra, que pertany al projecte de codi obert Hadoop. Aquest és un camp que durant el 2012 serà molt actiu, ja que les principals empreses del sector ja han entrat en aquest nou mercat.

Figura 5. Comportament dels sistemes tradicionals d'emmagatzematge de la informació



Les bases de dades NoSQL no impliquen que l'SQL no s'hagi d'usar, simplement que per a determinats problemes i aplicacions hi ha solucions millor. Per això NoSQL també ho podem llegir com *"not only SQL"*. La idea és que per a determinats escenaris cal relaxar moltes de les limitacions inherents a les bases de dades convencionals i la manera d'accedir-hi. Per exemple, podem pensar en col·leccions de documents amb camps definits de manera laxa que, fins i tot en el temps, poden anar canviant, en lloc de taules amb files i columnes prefixades. D'alguna manera, fins i tot podríem arribar a pensar que un sistema d'aquest tipus no és ni tan sols una base de dades entesa com a tal, sinó un sistema d'emmagatzematge distribuït per a gestionar dades dotades d'una certa estructura que pot ser enormement flexible.

Referència web

Si es té interès a conèixer la llarga llista de bases de dades NoSQL i saber-ne més detalls, la pàgina web següent és un bon punt de partida:

http://nosql-database.org/ (pàgina web en anglès)

2.5. Es requereixen nous models de programació

No hi ha dubte que les dades creixeran encara més i que moltes organitzacions voldran fer-les servir i no solament recol·lectar-les i emmagatzemar-les, ja que les dades només són útils per a les organitzacions si és possible fer-ne alguna cosa. Però processar aquestes grans quantitats de dades emmagatzemades distribuïdament de manera poc estructurada és un gran repte i novament no serveixen els mètodes tradicionals de processament.

Per a aconseguir el repte de processar grans conjunts de dades, Google va crear **MapReduce**, motor que està al darrere dels processaments de dades actuals de Google. Però ha estat el desenvolupament **Hadoop MapReduce**, per part de Yahoo, el que ha propiciat un ecosistema d'eines de codi obert de processament de grans volums de dades que està essent usat per la majoria d'implementacions d'altres empreses.

La innovació clau de MapReduce és la capacitat de fer una consulta, dividint-la i executant-la en paral·lel per mitjà de molts servidors alhora sobre un conjunt de dades immens.

D'aquesta manera es resol el problema de les dades quan són massa grans perquè càpiguen en una sola màquina. Aquest model té dues fases:

1) Fase *map*, en la qual les dades d'entrada són processades, una per una, i transformades en un conjunt intermedi de dades.

2) Fase *reduce*, en què aquests resultats intermedis es redueixen a un conjunt de dades resumides, que és el resultat final desitjat. Ara mateix és un procés en lot, i pot requerir minuts o hores per a processar-se.

El problema del MapReduce i l'emmagatzemament amb sistemes NoSQL és que no segueix la manera de processar i emmagatzemar que fa anys que fem i ensenyem en les universitats, i per això es fa difícil "pensar" en aquest nou paradigma, que en cert sentit, per entendre-s'hi, s'ha de "desaprendre". Davant d'això han sorgit projectes com el **Hive**. De manera senzilla podríem dir que és un sistema de magatzem de dades basat en Hadoop que va ser desenvolupat per Facebook, i ara és un projecte de codi obert dins de l'ecosistema de Hadoop. L'interès principal del Hive és que permet als usuaris escriure consultes en SQL, que després es converteixen en MapReduce de manera transparent al programador. Això permet als programadors d'SQL que no tenen experiència en MapReduce usar-lo i integrar-lo en els seus entorns habituals. Per altra banda, el **Pig Latin** és un paquet també en l'ecosistema Hadoop desenvolupat originalment per Yahoo, que ofereix un llenguatge d'alt nivell per a descriure i executar tasques de MapReduce. El seu objectiu també és fer accessible Hadoop als desenvolupadors familiaritzats amb la manipulació de dades amb SQL, i proporcionar una interfície interactiva i també una interfície per poder usar-lo des de Java.

Referència web

Un bon punt de partida per a entendre en més detall Hadoop i tot el seu entorn és la seva pàgina de la Wikipedia, a partir de la qual s'hi pot anar aprofundint a partir dels enllaços:

http://es.wikipedia.org/wiki/Hadoop (pàgina web en castellà)

2.6. La informació no és coneixement

Sembla que Albert Einstein va dir que "la informació no és coneixement". Quanta raó tenia, ja que les dades necessiten ser analitzades per a poder extreure'n el valor que contenen.

És fonamental l'habilitat per a transformar les dades en informació, i la informació en coneixement, de manera que es puguin, amb aquest coneixement, optimitzar processos en els negocis.

Pensem que aquesta **intel·ligència de negoci** actua com un factor estratègic per a una empresa o organització, i genera un avantatge competitiu potencial, que no és cap altre que proporcionar informació privilegiada per a respondre als problemes de negoci.

Posem pel cas una d'aquestes tecnologies, com la de l'aprenentatge (*machine learning*), que es troba al darrere de moltes d'aquestes anàlisis que comentàvem. Certament estan ja en un moment de molta maduresa. Per exemple, al febrer de l'any 2012, i durant tres nits consecutives, la computadora Whatson d'IBM va jugar contra dos dels més brillants campions de Jeopardy[11]. La computadora era capaç de respondre en menys de tres segons executant milers de tasques complexes simultàniament que usaven l'aprenentatge. Ara bé, l'aprenentatge en *big data* és un gran repte encara, ja que la majoria d'algorismes simples s'executen bé en milers de registres, però són ara mateix impracticables en milers de milions.

(11)Concurs cultural de la televisió nord-americana en què el concursant ha de respondre una pregunta amb una sèrie de pistes.

Per això els mètodes d'anàlisi existents fins ara, en general, encara s'han d'adaptar a la mida i complexitat dels formats del *big data*, i també les bases de dades que, en ser en general relacionals, hem vist que no permeten emmagatzemar aquestes ingents quantitats de dades. Sense cap mena de dubte aquest serà un dels camps que més avançarà durant aquests propers anys, atesa la importància que tindrà per a les empreses.

En aquesta línia en el món del codi obert, i dins de l'ecosistema de Hadoop que hem comentat abans, podem trobar iniciatives com **Mahout**. L'objectiu és produir una implementació lliure d'un paquet que inclogui els principals algorismes d'aprenentatge escalables sobre la plataforma de Hadoop. El projecte és molt actiu però encara queden molts algoritmes pendents d'incorporar. Malgrat això, probablement no és un reflex dels avenços que s'estan fent en aquest camp, ja que segurament molta recerca s'està desenvolupant de manera silenciosa dins de les grans corporacions, atesa la importància que pot tenir per als seus negocis. Bé, això és una opinió personal, però estic convençut que està passant. Posem per cas Google, que disposa del seu sistema d'emmagatzematge GFS i MapReduce per al desenvolupament dels seus sistemes d'aprenentatge a gran escala.

De Google, per exemple, sabem que els seus investigadors d'aprenentatge desenvolupen algorismes de computació que milloren el rendiment de manera automàtica per mitjà de l'experiència. Podem trobar que s'aplica per a resoldre molts problemes com recomanar etiquetes per a vídeos de YouTube o la identificació de diferents aspectes en les opinions en línia dels usuaris.

Lectura complementària

En el llibre *Understanding Big Data*, publicat per IBM, és descriu el *big data* des del punt de vista de negoci. El llibre es pot baixar de l'adreça següent:

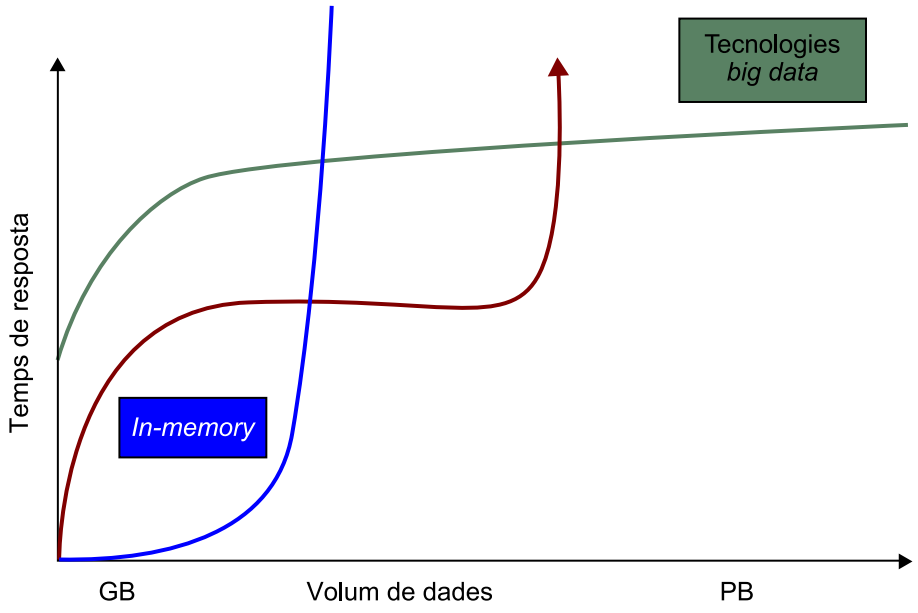
http://www-01.ibm.com/software/data/bigdata/ (libre en anglès)

2.7. Entrant en la nova era del Data 2.0.

La mida, la complexitat dels formats i la velocitat de lliurament superen les capacitats de les tecnologies de gestió de dades tradicionals. Per això es requereix l'ús de noves tecnologies que permetin la gestió de grans volums de dades. Noves tecnologies sorgiran amb el potencial de tenir gran impacte, com ara gestors de bases de dades *in-memory* per a atendre els nous requisits de dades.

En tot cas, el futur segur que serà híbrid, amb no solament una tecnologia d'emmagatzemament, sinó que serà una barreja de totes. Cal tenir en compte que no totes les aplicacions tenen el seu coll d'ampolla en les dades només. D'altra banda, per més que s'abarateixin els costos d'emmagatzemament en memòria, no sempre sortirà a compte l'emmagatzematge d'unes dades determinades a uns costos més elevats. Per això un dels temes de recerca importants és considerar de manera global la gestió de recursos d'emmagatzemament en els CPD, considerant els requisits de les aplicacions a l'hora de decidir on s'executen i on hi ha les dades. En la figura 6 es representa esquemàticament el comportament de les noves propostes enfront dels sistemes tradicionals d'emmagatzematge de la informació que hem presentat abans.

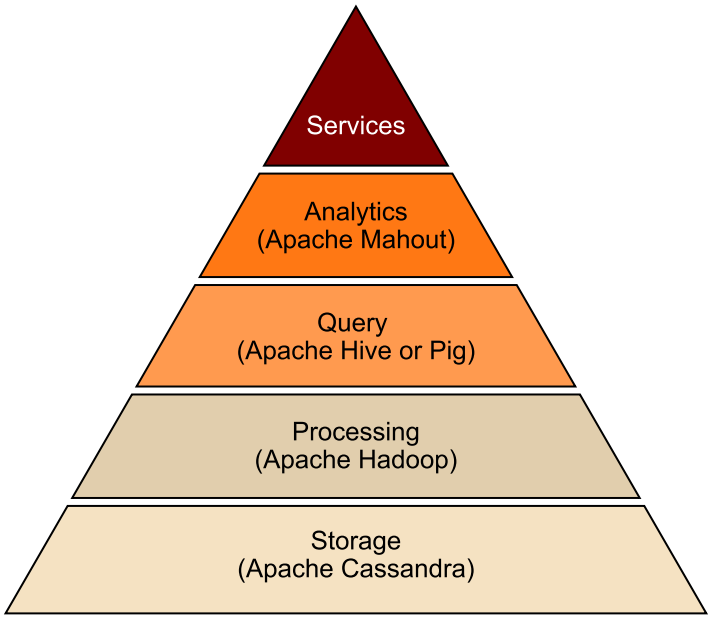
Figura 6. Comportament de les noves tecnologies per a superar les limitacions de la gestió de dades tradicional.



Un fet molt important és que aquestes noves tecnologies estan a l'abast de tothom, no solament de corporacions que poden fer grans inversions. De la mateixa manera que quan va madurar el Web va arribar l'era que va ser anomenada *Web 2.0*, gràcies a paquets de programari obert com els anomenats *LAMP* (Linux, Apache, MySQL i PHP), ara estem entrant en una nova era per a les dades, la **Data 2.0**, gràcies a una nova generació de tecnologies i arquitectures, dissenyades per a extreure valor a un gran nombre de dades de diversos

tipus i en alguns casos en temps real. En la figura 7 es presenten algunes de les tecnologies possibles que estan a l'abast de tothom per a poder entrar en aquest món del *big data*.

Figura 7. *Stack data* 2.0 d'eines *big data* de codi obert



Per acabar, cal apuntar que les aplicacions analítiques que s'usen en el procés de presa de decisions en les empreses començaran a mirar la informàtica en núvol usant com a base les eines que hem comentat en l'apartat "Informàtica en el núvol". Per a diversos sectors (*smart cities*, detallistes, financer, salut, etc.) apareixeran solucions verticals, és a dir que integraran tots els nivells per a processos específics que permetran obtenir diferenciació i creació d'avantatges competitius.

Lectura complementària

L'organització O'Really ha publicat el llibre *Big Data Now, Current perspectives from O´Reilly Media*, que es pot baixar de la pàgina web següent:

http://shop.oreilly.com/product/0636920022640.do (llibre en anglès)

Resum

Conclusió: un món d'oportunitats per a vosaltres

Tots els estudis apunten que la informàtica en núvol transformarà el món laboral, generarà nous llocs de treball, i òbviament en farà obsolets d'altres. Per això no tinc dubte que adquirir habilitats en el món de la informàtica en núvol és una aposta segura per a tots vosaltres. I el mateix passa amb el *big data*, ja que les companyies que vulguin donar sentit a tota la informació de què disposen hauran de contractar persones amb aquests coneixements.

Però també podeu pensar a emprendre, amb la vostra pròpia iniciativa. En aquest cas recordeu que la informàtica en núvol és el vostre gran aliat. La informàtica, tant si és en forma d'un paquet de programari com d'un servidor maquinari, ja es pot consumir com un servei més, com ho són l'electricitat o l'aigua. Només cal connectar-se a la xarxa d'Internet i, de la mateixa manera que passa amb l'electricitat, el consumidor paga segons el que consumeix, la qual cosa permet ajustar el consum a les necessitats de cada moment i evita grans inversions inicials en TIC. A més, recordeu també que és molt important l'agilitat que ofereix aquest nou paradigma a l'hora d'obtenir nous recursos maquinari o programari, enfront del temps que es requereix per a crear una instal·lació informàtica pròpia, sobretot si es té en compte que la pressió per a innovar dels emprenedors obliga a haver de disposar de recursos amb la rapidesa i quantitat necessaris.

Pel que fa al *big data*, ja hem vist que hi ha a la vostra disposició tot un conjunt de paquets de codi obert que us poden permetre grans aventures. Com ja s'ha comentat, el conjunt de paquets de programari Linux, Apache, MySQL i PHP (LAMP) va donar poder als desenvolupadors innovadors i els va permetre programar potents servidors web a partir de codi obert. Tot plegat va portar a la creació de moltes noves empreses en el sector TIC, la base del que es coneix com a *Web 2.0*. Estic segur que a partir d'ara veurem noves empreses basades en la pila de programari Data 2.0 que hem presentat anteriorment.

Recordeu, a més, que per a obtenir dades a partir de les quals generar-ne valor, disposeu del moviment *open data* que hem presentat. És una gran oportunitat, i voldria fer aquí molt èmfasi en el fet que els emprenedors han de veure en aquest moviment la possibilitat de generar noves aplicacions a partir d'aquestes dades que ens ofereix l'*open data*. I no oblidem el món dels mòbils i l'*Internet of things*. Hi ha tot un filó en el desenvolupament de programari més enllà de les simples "*apps*", que recol·lectin, emmagatzemin, organitzin, analitzin i integrin la informació amb altres programes o persones per mitjà dels seus dispositius.

En resum, la meva visió és que la informàtica en núvol i les noves tecnologies que l'acompanyen ja són aquí i ja no hi ha marxa enrere. Estic convençut que els inhibidors de la informàtica en núvol que hem comentat abans s'aniran solucionant. Per a expressar el poder transformador que crec que tindrà la informàtica en núvol i les tecnologies que l'acompanyen, em permeto usar una analogia amb l'adveniment del cotxe al començament del segle XX descrita per Rolf Harms i Michael Yamartino, de Microsoft, en l'article "The Economics of Cloud":

> "Quan els cotxes van aparèixer al principi del segle XX, van ser anomenats inicialment *carruatges sense cavalls*. Era comprensible que les persones es mostressin escèptiques al principi, i que veiessin l'invent a través de la lent del paradigma que havia prevalgut durant segles: el cavall i el carro. Per això els primers cotxes buscaven una imatge molt similar a la del cavall i el carro, ja que els enginyers mateixos no arribaven a comprendre les noves possibilitats del nou paradigma, com més velocitat o seguretat. Increïblement, els enginyers van mantenir una reproducció del cap d'un cavall al davant en els primers models de cotxe abans d'adonar-se que no tenia cap sentit. Inicialment hi va haver una incomprensió total del nou paradigma, s'afirmava que el cavall és aquí per a quedar-se, però l'automòbil és només una novetat, una moda passatgera. Fins i tot els pioners del cotxe mateixos no acabaven de comprendre l'impacte potencial que el seu treball podria tenir en el món. Quan Daimler, possiblement l'inventor de l'automòbil, va intentar estimar el mercat potencial de l'automòbil a llarg termini, va arribar a la conclusió que mai no hi podria haver més d'1 milió de cotxes, a causa de l'alt cost i l'escassetat de xofers capacitats."
>
> **Rolf Harms; Michael Yamartino** (2010). "The Economics of Cloud". Microsoft. Disponible a: http://www.microsoft.com/presspass/presskits/cloud/docs/The-Economics-of-the-Cloud.pdf

Us encoratjo a avançar i aprofitar aquestes apassionants oportunitats que teniu al davant, malgrat no sapigueu del cert què us espera en cada pas a causa de la rapidesa amb què evoluciona la tecnologia. Sobretot confieu en vosaltres mateixos i perdeu la por a equivocar-vos, ja que equivocar-se és inseparable d'actuar.

Espero que aquest material us hagi ajudat en la descoberta d'aquestes noves tecnologies i que sobretot us hagi engrescat a endinsar-vos una mica més en aquest apassionant món, per a comprendre la transformació que està patint la manera com interactuem i, particularment, la manera com fem negocis. I per què no, animar-vos a emprendre aprofitant totes aquestes eines que estan al vostre abans. Sort!

Referència web

Com a divulgador mantinc un blog on vaig escrivint els avenços en aquestes tecnologies. Si crieu que us pot ser útil em podeu trobar a:

www.JordiTorres.org/blog