MARIONA GABARRÓ

BIG DATA

UN MÓN ENORMEMENT NOU

Treball de Recerca
Batxillerat
Tutor: Jordi Rincon

BIG DATA

UN MÓN ENORMEMENT NOU

BIG DATA

UN MÓN ENORMEMENT NOU

MARIONA GABARRÓ ROVIRA

PRESENTACIÓ

Des de ben petits ens fan decidir sobre el nostre futur, cap a on dirigir-lo, què volem estudiar i sobretot, què volem ser de grans. Però la pregunta correcta és, què ens agrada, què volem conèixer, què ens motiva, què ens apassiona, què fem quan ens ho passem bé, amb què gaudim, quan som feliços? I aquesta última part jo sempre l'he tingut clara.

Així que vaig començar el Treball de Recerca preguntant-me: Què m'agrada? La resposta va ser ràpida: aprendre coses noves, conèixer com funciona el món, màrqueting, economia, matemàtiques i psicologia.

No va ser ni és fàcil trobar un tema que ho englobi tot, però va aparèixer davant meu com aquell qui busca ser trobat. Fullejant un llibre vaig veure les dues paraules: BIG DATA. Me n'havien dit alguna cosa però poc. Abans de començar, vaig buscar una mica d'informació per veure si realment encaixava amb el que volia. I sí.

A més a més, era un tema molt extens i podria arribar tan lluny com volgués sense que se'm quedés curt. Com a definició introductòria, Big Data es refereix a tota la informació digitalitzada que existeix sobre cadascun de nosaltres, els que vivim en el planeta Terra.

A poc a poc vaig anar descobrint de quina manera concordava el tema que m'agrada:

Aprendre coses noves? Es queda curt. No és només un tema nou per a mi, és un tema nou per a mi i per a tothom. Poca gent coneix aquesta realitat que ens envolta cada segon que vivim.

Com funciona el món? Conèixer aquesta realitat, aquest món paral·lel que molts de nosaltres sospitem que existeix però preferim ignorar, conèixer com funciona actualment el món gràcies al BIG DATA.

Màrqueting? BIG DATA i les necessitats personals quotidianes es relacionen a través del màrqueting, per tant, el màrqueting és un factor veí al tema.

Matemàtiques? BIG DATA, tal com diu el seu nom, és un grup de dades que es manegen a través de mecanismes matemàtics.

Psicologia? L'aparició del BIG DATA ha afectat directament la nostra societat. És més, l'ha modificat i continua fent-ho a una velocitat exageradament ràpida.

PRÒLEG

Per entendre bé aquest treball, cal entendre bé la seva filosofia. Per començar és important saber que no em va costar començar. És la mítica frase, que tots coneixem: "I ara per on començo?". La pregunta que em vaig fer jo, en canvi, va ser: "Què vull saber?". I és quan van començar a sortir infinites preguntes i finites hipòtesis. Però certament, en el moment de la investigació, tal com es comporten els fractals, anaven apareixent noves i noves preguntes que vaig haver de deixar en blanc, ja que no pots trobar-les totes.

Big Data es refereix a tota la informació digitalitzada que existeix sobre cada un de nosaltres. Però, tota aquesta informació, d'on surt? On és? Com es mou? Quin valor té? Per a què i com s'utilitza? Qui la utilitza? Es pot vendre, és a dir, pot vulnerar la nostra privacitat? I la nostra seguretat?

D'acord amb aquestes preguntes i la informació prèvia que vaig buscar, vaig plantejar-me diverses hipòtesis. En la primera hipòtesi vull contrastar el fet queBig Data tingui una aplicació al màrqueting. Però evidentment vaig proposar-me més hipòtesis sobre Big Data, com ara contrastar els avantatges i desavantatges que pot tenir. També la seva importància en el dia a dia que estem vivint, és a dir, com condiciona el món actual, tan sotmès a la tecnologia i a Internet. Per altra banda, Big Data pot generar beneficis a una empresa, però realment s'utilitza? Si és que no, per què? I si és que sí, de quina manera? Finalment, saber de quina manera Big Data afecta la nostra identitat i la nostra privacitat com a persones socials i individuals.

Totes les hipòtesis esmentades es poden resumir en una de sola: contrastar la seva importància amb el que realment és per al dia a dia de les persones, de les empreses, dels govern i del món.

Un cop assolides totes les respostes a les meves hipòtesis, i molts cops més enllà de les hipòtesis, vaig endinsar-me en el món més científic: l'anàlisi del Big Data, i totes les preguntes i respostes que comporta.

PART TEÒRICA

ÍNDEX

1. ORÍGENS

INTERNET

En el moment que un es compra un mòbil, un ordinador, una televisió,... és a dir, qualsevol aparell electrònic capaç de connectar-se a Internet, i s'hi connecta, comença a deixar tot un rastre de dades que després és utilitzat pels experts. Però tot comença amb uns orígens; què és Internet?

Quan ens referim a Internet ens referim a una xarxa pública i mundial d'ordinadors que estan connectats entre si a través d'un mitjà, com ara la fibra òptica, cable coaxial, radiofreqüència o línies telefòniques.

Una xarxa és un grup d'ordinadors interconnectats que es crea amb la finalitat de compartir recursos i informació a distància.

Quan un es connecta a Internet té accés al World Wide Web.

WORLD WIDE WEB

El concepte de World Wide Web (WWW) neix l'any 1990 a Suïssa gràcies a Tim Berners-Lee, qui va donar lloc al projecte. Té com a funció ordenar i distribuir tot allò que està i existeix a Internet.

Al cap de pocs anys, es va crear el WWW que tots coneixem actualment, una xarxa de pàgines escrites en hipertext* i connectades entre si formant una mateixa unitat. És a dir, metafòricament WWW és una biblioteca formada per moltes (actualment moltíssimes és poc) pàgines d'informació.

*Què és un hipertext? És un sistema d'organització de la informació al qual hi ha relacions entre les diferents paraules claus i altres textos. És a dir, és una manera de relacionar tots els textos o document que, en aquest cas, estan a Internet.

Internet i WWW són, clarament, un concepte i una idea completament nova que provoca una revolució social, política i cultural, ja que té unes grans conseqüències que cada cop són més impactants a la societat, és a dir, a la Globalització.

LLEI DE MOORE

En el moment en què relacionem el concepte d'Internet amb el que entenem per Dades des del punt de vista del futur, apareix la Llei de Moore.

Quan un utilitza Internet, fa servir el buscador, o qualsevol fet relacionat amb la xarxa (WWW), crea tot tipus de dades i informació. Aquesta informació es queda a la xarxa, al *núvol*, però si es vol organitzar o guardar ocupa una certa memòria.

L'any 1965 Gordon Moore, cofundador d'Intel, va pronosticar el que es coneix com la Llei de Moore. Aquesta llei consisteix en el següent: cada dos anys es duplicaria la capacitat de memòria de la tecnologia informàtica. Al mateix temps, com a conseqüència directa, al cap d'un any es reduiria a la meitat el cost d'un producte informàtic, i al cap de dos el producte quedaria obsolet (és a dir, inútil).

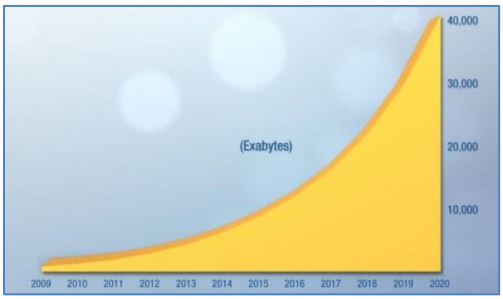
Científicament, la Llei de Moore no és una llei, ja que no es basa en fets científics, sinó que es basa en l'observació de les accions humanes.

Actualment aquest Llei, després de 50 anys, continua confirmant-se. Tot i que s'espera que algun dia s'arribi a un límit.

Des del meu punt de vista, la Llei de Moore és una evidència de l'evolució tecnològica al llarg d'aquests anys i al llarg dels propers anys. És a dir, és una manera de demostrar com ha canviat la investigació i la societat i també que continuarà canviant, sobretot després del naixement d'Internet.

2. UNIVERS DIGITAL

És un fet que l'Univers Digital està creixent de manera enorme i tendeix a créixer cada cop més. Fins ara l'Univers Digital es duplicava cada dos anys, però es preveu que entre el 2013 i el 2020 es multipliqui per deu. És a dir, passarà de 4,4 milions de milions de gigabytes a 44 milions de milions de gigabytes. Tot això és en part conseqüència del que entenem per Internet de les Coses.



El creixement de l'Univers Digital des de principis del 2010 fins a finals del 2020.

INTERNET DE LES COSES

Quan parlem d'Internet de les Coses (IOT, que prové de *Internet of Things*) ens referim al conjunt d'objectes de la vida quotidiana (com ara *smartphones*, cotxes, televisions, ordinadors, ... i fins i tot rellotges) que estan interconnectats tot formant una xarxa. Cada cop existeixen més objectes de la vida quotidiana capaços de connectar-se a Internet (actualment hi ha 14 mil milions de dispositius connectats a Internet) i per tant, creem més Dades, i és per això que afirmem que l'Univers Digital està augmentant enormement. "International Data Corporation" (IDC)* preveu que l'any 2021 hi hagi més de 28 mil milions de dispositius connectats a Internet.

Però la velocitat amb què augmenta la capacitat tecnològica és tan gran que els mecanismes que existeixen d'anàlisi es queden curts i el mercat és incapaç d'absorbir-les. És així com tan sols el 5% de tota la informació que disposem és analitzada.

A més a més, els mecanismes d'anàlisi que disposem no són només insuficients en capacitat quantitativa sinó que també han d'augmentar en la seva capacitat de rebre molts més diferents tipus de format. A diferència de fa uns anys, actualment tota la informació que disposem no

està en un únic format (Base de Dades tradicionals), sinó que arriben amb tot tipus de formats (vídeos, gravacions de veu, textos, pdf, etc.) i és quan parlem que tenim un Llac de Dades.

IOT EN MÀRQUETING

El naixement i creixement d'Internet de les Coses presenta noves maneres d'interactuar amb els clients, i per tant, nous plantejaments de màrqueting.

Segons International Data Corporation (IDC)* i EMC Corporation (EMC)**, hi ha 5 aspectes principals en les noves oportunitats de màrqueting gràcies a IOT:

- **Nous models de negoci.** Amb l'ajuda de les dades que genera IOT, les empreses es podran apropar ràpidament a les necessitats dels clients i del mercat.
- **Informació a temps real.** Gràcies a IOT les empreses recopilen dades a temps real (no necessàriament) que els permet observar els seus processos, com es mouen els clients i per tant millorar l'eficiència operativa i la fidelització dels clients.
- **La diversificació de les fonts d'ingressos.** L'IOT permet a les empreses crear nous productes i serveis més enllà dels productes tradicionals ja que detecten noves necessitats o desitjos a través de les dades.
- **Visibilitat Global.** Quan comptem amb IOT es fa molt més fàcil per a les empreses tenir una visió global de la seva cadena del que passa independentment de la seva ubicació.
- **Operacions eficients i intel·ligents.** El fet de tenir accés a la informació d'autònoms permet a les empreses decidir preus, vendes, etc. a temps real.

IDC*

Quan ens informem o investiguem sobre un tema hem de saber de qui és i d'on surt tota la informació que rebem. És per això que al llarg d'aquest treball he anat comprovant cada font per tal que aquesta sigui fiable. Considero que aquesta font és fiable per dues raons; la primera, és perquè és una empresa important especialitzada en tot el que envolta el Big Data; i la segona, és perquè he anat comparant amb altres pàgines web (articles, opinions, per exemple) i es complementaven i corroboraven entre si. Per tant, m'he trobat davant d'una gran font d'informació: IDC i EMC.

International Data Corporation (IDC) és una empresa mundial d'intel·ligència de mercat, serveis de consultoria i especialitzada en mercats de tecnologia de la informació, telecomunicacions i tecnologia de consum. IDC és una subsidiària d'IDG (International Data Group), els mitjans de comunicació de tecnologia més importants del món.

IDC compta amb més de 1.000 analistes en 110 països arreu del món.

EMC**

EMC² (EMC Corporation) és una empresa global i líder que permet a altres empreses transformar el que fan i oferir un servei en tecnologia de la informació. És a dir, és un fabricant de programari, sistemes i emmagatzemament d'informació per administrar-ho.

EMC també té un gran sector dedicat al Big Data, aplicat al màrqueting, al coneixement i a buscar solucions.

Al llarg del treball, almenys fins ara, EMC ha estat una gran font d'informació, ja que es mostra clara, entenedora i completa.

SEGURETAT I PRIVACITAT

El fet que Big Data no pari de créixer i créixer acceleradament provoca una necessitat de nous sistemes i eines per a la seva protecció (i per tant seguretat) i privacitat.

Entenem per protecció l'acció de protegir, en aquest cas, les dades per tal que no es converteixin en un producte de compra i venda.
Entenem per seguretat el fet que el propietari de les dades no estigui sotmès a cap perill ni risc a causa de la falta de protecció. És a dir, la protecció és la causa i la seguretat la conseqüència. Quan més protecció tenen les dades, més segures estan.
Entenem per privacitat el fet que l'usuari mantingui les seves dades fora dels ulls de l'altre gent. És a dir, que ningú part d'ell conegui les seves dades personals o no tan personals.

S'ha de tenir present que hi ha dades que no necessiten protecció, com ara les fotos del telèfon de la càmera, contingut de la web pública i les dades de codi obert. És un fet que el 43% de les dades de l'Univers Digital necessiten algun tipus de protecció (com ara la propietat in-tel·lectual, dades confidencials, les dades financeres, la informació personal identificable, els registres mèdics, la informació del compte d'usuari, etc.), mentre que tan sols el 48% d'aquestes dades que necessiten protecció estan protegides, és a dir, el 20% del total de les dades. Això ens porta a un problema que la població civil no el viu fins que arriba a un extrem. Tots nosaltres sabem que estem exposats, que qui vol saber qualsevol cosa sobre nosaltres la podrà saber, i que la poca privacitat que tenim no ens priva de res. Però aquells que s'aprofiten de la poca privacitat que tenim sovint són discrets, no els veus. I potser és per això que es converteix en un tema tabú.

Segons IDC, el nivell existent de protecció varia segons la posició geogràfica, en els països emergents* són els que tenen un nivell més baix.
{*Països emergents: Argentina, Brasil, Xina, Colòmbia, República Txeca, Egipte, Hongria, Índia, Indonèsia, Israel, Malàisia, Mèxic, Marroc, Filipines, Perú, Polònia, Rússia, Sud-àfrica, Taiwan, Tailàndia, Turquia.}
Realment sabem que els nostres correus electrònics, les nostres trucades, els nostres estudis, documents electrònics, etc. estan registrats en algun lloc on algú té excés. És cert que aquest algú (Gmail, Android, Hotmail, Dropbox, ...) ens ha promès la nostra privacitat i seguretat, però en el moment que aquest algú ven les nostres dades, tot i que no es fan públiques, són utilitza-

des per crear perfils de consumidors i a partir d'aquí poden fer un millor màrqueting. Continua sent això privacitat?

A continuació s'explicaran els problemes que presenten la privacitat i la seguretat.

El fet de tenir diversos dispositius (des de portes fins a tauletes i cotxes) interconnectats (via Internet) permet un fàcil excés de vigilància il·legal a aquells que en saben suficient.

El fet d'utilitzar una mateixa xarxa per a tots els dispositius de la casa permet que si es connecten a través de l'Smart TV, per exemple, poden arribar al teu ordinador personal. És així com l'existència d'IOT ha dificultat molt la seguretat i privacitat de dades, ja que cada cop és més difícil la protecció d'aquestes.

Stephen Coty, director d'investigació d'amenaces en Alert Logic, proposa una solució al problema de la privacitat i seguretat en els IOT; utilitzar dues xarxes diferents, una per tots els dispositius IOT i l'altra per a tots els dispositius personals. També es recomana una altra xarxa per als dispositius de treball. És cert que és una solució bona i eficaç però no és ni pràctica, ni fàcil ni econòmica.

Però no cal preocupar-se. Més endavant explicaré que no n'hi ha per tant, i que d'una cosa dolenta sempre se'n treu la part bona.

CLOUD SECURITY ALLIANCE (CSA)

La grandesa amb què ha aparegut Big Data ha provocat canvis negatius envers la privacitat i la seguretat. És així com Cloud Security Alliance (CSA), una organització sense ànim de lucre, va anunciar que investigaria per millorar els sistemes de seguretat, no només en Big Data sinó que també en Cloud (pàg. 54).

CSA té un equip de més de 48.000 persones arreu del món. Aquesta organització no només investiga sinó que també ofereix educació, certificació, esdeveniments i productes especialment dedicats a la seguretat del Big Data i del Cloud.

PART BONA DE LA POCA PRIVACITAT

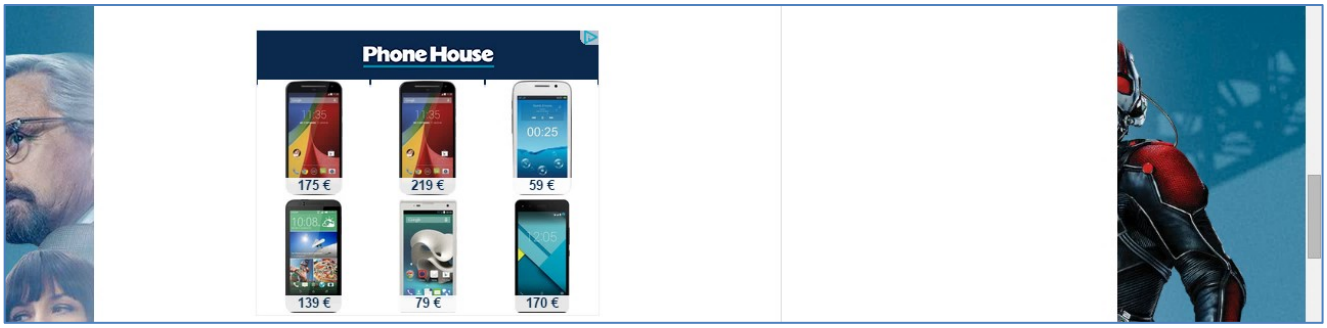
És cert que el fet que ho sàpiguen tot o gairebé tot de nosaltres ens fa una mica de por, ens espanta. Però si per una banda les empreses utilitzen la teva informació en el seu propi benefici, per altra banda aquest mateix benefici també et beneficia a tu.

Per exemple, el fet que Google emmagatzemi el registre de la navegació i de cerca de tots nosaltres permet crear prediccions basades en el que la gent busca, que normalment és el mateix del que els interessa o els preocupa. És així com es produeix la invasió de privacitat en saber tot el que hem fet durant la navegació, però en canvi pot crear un benefici a favor nostre. El primer cas d'això va ser quan l'any 2009 es va descobrir un nou virus de la Grip, el virus H1N1, el qual Google va poder predir per on s'expandia abans que ho fessin el Centre de Control i Prevenció de Malalties d'Estats Units.

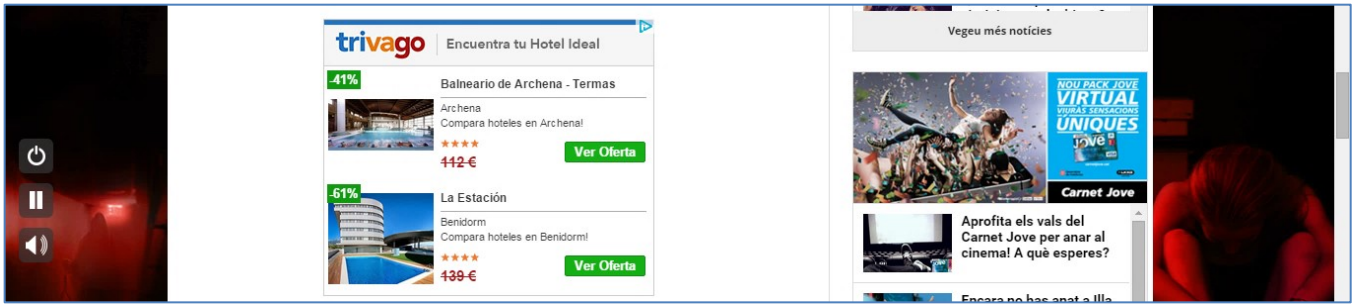
Segons Siva Vaidhyanathan, un historiador de la cultura, "Per cada persona que es queixa d'Street View, milions més els resulta útil.".

No ens agrada que sàpiguen sobre nosaltres però a la vegada volem que ens venguin productes propers a nosaltres. No volem explicar a Google la nostra informació més personal però volem un perfil i un servei personalitzat. No volem que es fiquin dins dels nostres navegadors però agraïm (indirectament) que ens ensenyin publicitat que vagi dirigida 100% cap a nosaltres.

La meva opinió ha quedat reflectida en l'escrit així que el que faré serà posar exemples del que són descaradament els anuncis personalitzats que fa Google. Posaré un exemple personal. Ahir vaig estar mirant mòbils perquè me'l van robar fa poc. Entro en una pàgina web i em surt l'anunci següent:



Recarrego la pàgina web i em surt un anunci del balneari que fa dues setmanes la meva àvia em va demanar que li mirés:



3. COOKIES

Cada cop que entrem a una pàgina web sovint ens avisen de la utilització de *cookies*. El que fa la majoria de gent, per no dir tothom, és "Acceptar" o simplement tancar-ho, però realment, què ens estan dient? Quin permís els estem donant? Qui hi guanya i qui hi perd? Són preguntes que aniran obtenint resposta al llarg de l'apartat.

Comencem amb alguns exemples de Pàgines Web que utilitzen *cookies*:

- https://www.condis.es/
- http://www.seat.es/
- http://www.movistar.es/
- http://www.vilaweb.cat/
- http://www.ara.cat/

Però aquestes cinc pàgines web són tan sols una mínima representació, ja que el 90% de les pàgines que utilitzem al dia a dia utilitzen *cookies*.

QUÈ SÓN?

Les *cookies* per definició són una petita informació o petits arxius que s'envien a través d'una pàgina web i que es guarda al navegador de l'usuari que l'està utilitzant.

Quan un navega per una pàgina web es van creant les *cookies*, que després són utilitzades per fer un seguiment de l'activitat de l'usuari durant la navegació per la web, com ara ajudar-te a recordar el teu inici de sessió (usuari i contrasenya), selecció d'un tema, preferències i altres funcions de personalització.

QUIN ÉS EL SEU ORIGEN?

L'origen del perquè el nom de *cookies* són les tradicionals galetes xineses de la sort per la seva singular semblança. A més a més que no ho podem afirmar al 100% (pot ser un simple rumor), ens fixarem en l'origen de l'ús de les cookies. Quan es van inventar i per a què?

La creació de les cookies s'origina amb la necessitat de Nestcape (empresa informàtica nord-americana i una de les primeres a treballar amb World Wide Web, creadora del navegador Netscape Navigator, una de les bases per crear Mozilla Firefox) de crear un carretó electrònic online fiable. Quan seleccionem un producte per comprar, se'ns guarda al carret de compra. Després pots continuar navegant per la web, mirant altres productes, o fins i tot sortir de la web i entrar-hi més tard, que el carretó electrònic continuarà guardant aquells productes que has seleccionat anteriorment. L'element que permet que aquesta informació es guardi són les cookies, ja que permeten que les dades es guardin en el navegador, que no es perdin, i que la web les pugui utilitzar sempre que vulgui.

Aquesta aplicació que ens fa a la compra online, també s'aplica per exemple a guardar el teu perfil (i no haver-se de connectar cada vegada). A més a més actualment s'aplica al màrqueting online.

PER A QUÈ S'UTILITZEN?

Per començar, les cookies continuen fent-se servir per l'ús original del carretó electrònic i per tant considerem que són essencials en la creació d'un carretó electrònic. Però anant més enllà, ens trobem davant de tres usos principals:

- **Portar el control dels usuaris, ús totalment en benefici de l'usuari.** Quan iniciem sessió en qualsevol pàgina web (Facebook, per exemple), se'ns guarda l'inici de sessió de manera que el següent cop que entrem a la web (Facebook) no haguem de tornar a iniciar sessió (posant l'usuari i la contrasenya). Això passa perquè quan un usuari inicia sessió es guarda una cookie, el següent cop la pàgina web detecta la cookie guardada i llavors és quan ja tenim la sessió oberta.

  De tota manera, la cookie no identifica a cap persona, sinó que identifica una combinació de navegador i ordinador.
- **Espiar la teva navegació, ús en benefici de tercers.** Les cookies que es creen en entrar a una pagina web permeten a aquesta espiar com t'has mogut (el què, quan i com

uses la web) per tal de poder fer una publicitat basada en l'usuari. Algunes cookies també poden espiar quines webs hem visitat, quan, què ens interessa, etc.

- **Personalitzar l'aspecte d'una web segons les preferències de l'usuari.** Per exemple, quan busquem alguna cosa a Google podem triar en quin idioma volem els resultats, quants resultats volem veure a cada pàgina, ... (preferències).

Ara que comencem a entendre més sobre aquest tema, veig que hi ha molts elements que posteriorment s'han utilitzat per a temes de màrqueting (publicitat, per exemple) i que les cookies no són una excepció. Veiem l'evolució en què les cookies es van inventar en benefici de l'usuari i posteriorment s'utilitzen en beneficis de tercers per tal de personalitzar la publicitat.

QUINS TIPUS DE COOKIES HI HA?

SEGONS QUINA IDENTITAT GESTIONA LA PÀGINA WEB

- **Cookies pròpies o d'origen.** Són aquelles que s'envien al navegador de l'usuari directament des de la pàgina web i en la qual s'ofereix el servei que ha sol·licitat l'usuari.
- **Cookies de tercer.** Són aquelles que provenen d'un lloc web que no és el que estàs visitant en aquest moment, sinó a través d'un anunci d'altres webs de la pàgina que estàs visitant. Els llocs web de l'anunci poden utilitzar aquestes cookies per fer un seguiment de l'usuari a Internet. D'aquesta manera, una agència de publicitat o de màrqueting online podrà fer un seguiment de les visites en cada pagina web que hi tenen un anunci, o també adaptar aquest anunci en funció a l'ús anterior de l'usuari a Internet.
  La possibilitat que les webs propietàries de l'anunci puguin crear un perfil dels usuaris posa en perill la nostra privacitat. És per això que cada cop es vigila més i es creen lleis més estrictes.

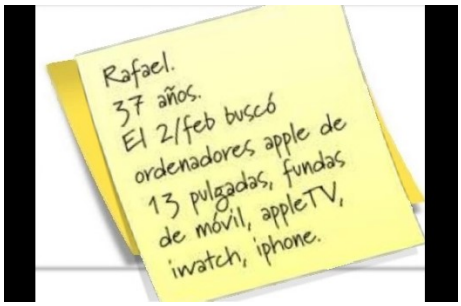
Les cookies són un exemple del nostre desig a tenir-ho tot però sempre en un mínim risc. A tots ens agrada veure anuncis de coses que ens interessen, trobar les coses fàcilment, personalitzar allò nostre, que Internet s'adapti a nosaltres, etc. Però tot això té un preu i una condició, i és la pèrdua de privacitat. Per exemple, si vols que Internet s'adapti al que t'agrada, ha de saber què t'agrada i en el moment en que sap què t'agrada, perds part de la teva privacitat com a persona.

SEGONS EL TERMINI DE TEMPS QUE ES MANTENEN ACTIVADES

- **Cookies de sessió o temporals.** Són aquelles que es creen temporalment al navegador mentre estàs visitant una pàgina web, però quan surts de la pàgina web o del navegador s'eliminen. És a dir, només guarden les dades mentre l'usuari és a la pagina web. S'utilitzen per exemple per guardar la informació dels articles del carretó electrònic.
- **Cookies persistents o permanents**. Són aquelles que continuen guardades al navegador encara que el tanquis. El responsable de la cookie hi pot accedir durant un període de temps que ell mateix defineix que pot estar entre pocs minuts o anys.
  Es poden utilitzar, per exemple, per no haver d'iniciar sessió cada cop que obres de nou el navegador o amb un fons d'anàlisi per utilitzar-ho en màrqueting.

SEGONS LA SEVA FINALITAT

- **Cookies tècniques.** Són aquelles que permeten a l'usuari la navegació a través d'una web i la utilització de diferents serveis (controlar el trànsit, identificar la sessió, etc.)
- **Cookies de personalització.** Són aquelles que permeten a l'usuari accedir a un servei on hi ha unes certes preferències adaptades al mateix usuari, com ara l'idioma, el navegador predeterminats, etc.
- **Cookies publicitàries.** Són aquelles que permeten la gestió d'espais publicitaris que el propietari de la pàgina hagi determinat.
- **Cookies de publicitat comportamental.** Són cookies publicitàries que a més a més poden emmagatzemar informació sobre el comportament de l'usuari en la web i els hàbits de la navegació. Això permet crear un perfil específic de l'usuari i d'aquesta manera escollir una publicitat adaptada a aquell perfil.



El que descriu aquesta *imatge* és el que els propietaris de les cookies acaben sabent i utilitzen per crear anuncis personalitzats. Les cookies que ho permeten són les cookies de tercers i de publicitat comportamental. També persistents ja que parla de dates passades.

Les cookies de publicitat comportamental són les cookies gestionades per tercers, i són aquestes cookies les que fàcilment ens destapen la nostra privacitat, es fiquen dins nostre i ens manipulen amb el que ens ensenyen. Aquestes cookies són les més perilloses però a la vegada són les més interessants.

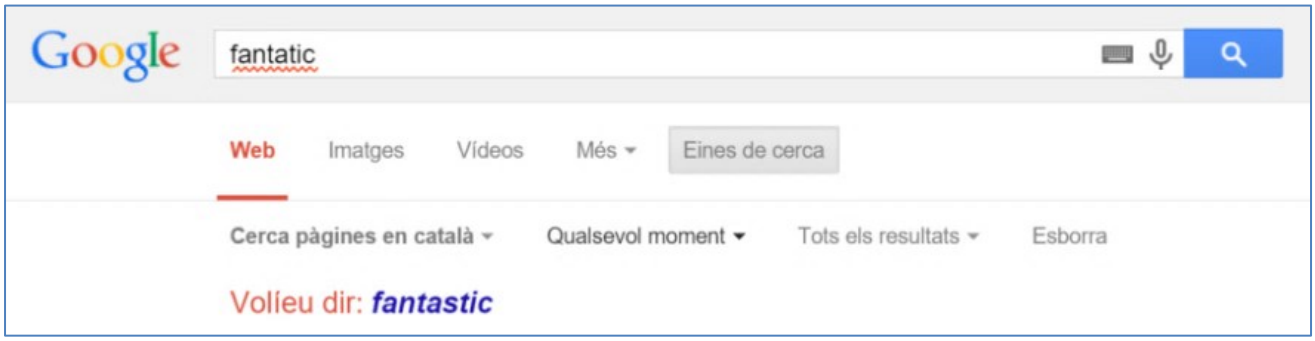
FALSACIONS DE LES COOKIES

Sempre, quan es crea un nou terme a la societat, hi ha qui l'interpreta a la seva manera i això provoca després confusions, malentesos i falsacions. Dedicaré aquest petit apartat a mencionar aquelles afirmacions que no són certes sobre les cookies.

- Les cookies són semblants als cucs i virus que poden esborrar dades dels discos durs.
- Les cookies són un tipus de *spyware* (programa espia) ja que poden llegir informació personal guardada a l'ordinador de l'usuari.
- Les cookies s'utilitzen per generar *spam.*
- Les cookies només s'utilitzen amb un objectiu publicitari

4. BIG DATA

Les aplicacions de les dades massives pot ser una cosa molt simple o, millor dit, natural, tan natural que a vegades pot semblar invisible. En el moment que una persona cerca qualsevol cosa a Google, sap que si comet una errada ortogràfica Google li corregirà. És més, a vegades molts de nosaltres teclegem qualsevol paraula a Google però no per fer una cerca, sinó per saber com s'escriu ortogràficament correcta. Google no és cap persona, i menys una persona que sàpiga com s'escriu cada paraula correctament. Però llavors com s'ho fa? Com sap que quan



busques "fantatic" et refereixes a "fantàstic"? Quan un tecleja una paraula a Google i està mal escrita, l'algoritme de Google, basant-se en elsl milers de vegades que s'ha buscat la paraula "fantàstic" i no pas "fantatic", s'adona del que passa i et suggereix una altra paraula que és aquella paraula que milers de persones han escrit abans que tu, "fantàstic". Sense ser-ne conscients, convivim en una vida quotidiana plena d'aplicacions del Big Data.

Però la informació en si no és tan visual. Amb l'aparició de tot el conjunt d'Internet de les Coses, la nostra societat viu submergida en un món de dades. És cert que el volum de dades que actualment tenim és grandiós, però a més a més la velocitat en què aquesta informació està creixent és també enorme. I és quan ens trobem davant d'un món de dades massives. És quan podem afirmar que aquesta quantitat de dades ha començat a acumular-se fins al punt que està passant alguna cosa nova i especial.

Va ser a la dècada del 2000 quan es va utilitzar per primer cop el terme Big Data per referir-se a "dades massives" i ràpidament es va expandir a totes les àrees de l'activitat humana. Poc després, l'any 2001, Doug Laney va definir "Les 3 V" del Big Data; volum, velocitat i varietat (pàg. 48).

La gran quantitat d'informació va arribar a un punt que l'any 2006 els enginyers necessitaven urgentment modernitzar les eines per poder analitzar la informació, i és quan va aparèixer Hadoop, by Yahoo, (pàg. 91) i MapReduce, by Google, (pàg 94).

La idea abstracta del Big Data està transformant i transformarà el nostre entorn, la nostra societat.

Per tenir una idea de quina gran quantitat de dades estem parlant, nombraré algunes xifres. Segons un estudi de Martin Hilbert, l'any 2007 existien 300 exabytes de dades emmagatzemades (dades analògiques i digitals). Per fer-nos una idea, una pel·lícula (típica del cinema) ocupa aproximadament 1 gigabyte. Un exabytes són mil milions de gigabytes. Si tota la informació fos en forma de pel·lícula, existirien 300 mil milions de pel·lícules. Ho comparem amb la població mundial:

Població mundial:	7 376 471 981
Quantitat de pel·lícules:	300 000 000 000

És a dir, equivaldria a 40,6 pel·lícules per a cada persona. No és una barbaritat?

Però això no és tot. He calculat quants exabytes de dades emmagatzemades existeixen l'any 2015. Com? He seguit la Llei de Moore, és a dir, cada dos anys es duplica la xifra. Però ho he fet fins al 2013.

$$Del\ 2007\ al\ 2009 \rightarrow 300 \cdot 2 = 600$$
$$Del\ 2009\ al\ 2011 \rightarrow 600 \cdot 2 = 1200$$
$$Del\ 2011\ al\ 20013 \rightarrow 1200 \cdot 2 = 2400$$

Es preveu que del 2013 al 2020 es multipliqui la xifra per 10. Per tant, he multiplicat els exabytes del 2013 per deu i així tenim la xifra del 2020. Després li he restat els 2400 per tal que em quedés allò que havia augmentat des del 2013 fins a l'any 2020. El resultat obtingut l'he dividit entre 7 (els anys que van del 2013 a l'any 2020) per tenir una idea del que augmentava cada any entre el 2013 i el 2020, que després l'he multiplicat per dos ja que són els dos anys que em falten per arribar del 2013 al 2015. Al resultat obtingut li he sumat els 2400 que era les dades que ja teníem.

$$Del\ 2007\ al\ 2020 \rightarrow 2400 \cdot 10 = 24000$$
$$Augment\ del\ 2013\ al\ 2020 \rightarrow 24000 - 2400 = 21600$$
$$Augment\ per\ any\ entre\ el\ 2013\ i\ 2020 \rightarrow 21600 : 7 = 3086$$
$$Augment\ del\ 2013\ al\ 2015 \rightarrow 3086 \cdot 2 = 6172$$
$$Xifra\ al\ 2015 \rightarrow 6172 + 2400 = 8572$$

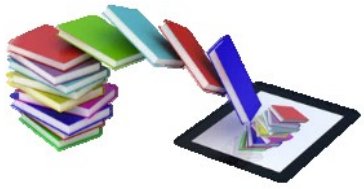
Tot això és molt aproximat ja que del 2013 al 2020 es preveu un creixement exponencial, no pas uniforme (que és el que he utilitzat), però he preferit simplificar-ho i després saber que el resultat obtingut serà aproximat cap amunt.

8572 *aproximat a* 8600 (tot i que és molt possible que estigui entre els 9000 i 10000).

El resultat obtingut ha sigut que actualment existeixen més de 8600 exabytes de dades emmagatzemades. Sorprenent. Per tant, actualment el requadre de comparació quedaria així:

Població mundial:	7 376 471 981
Quantitat de pel·lícules:	8 600 000 000 000

Ara fem el mateix, però passant-ho tot a analògic. Si convertíssim tota la informació a llibres de paper, aquest cobririen tota la superfície d'Estats Units 52 vegades. Si ho passéssim a CD i els apiléssim, farien 5 cops la distància entre la Terra i la Lluna.



Per altra banda, també és un fet que cada cop hi ha més informació digital a la vegada que la informació analògica no creix, és així com l'any 2000 el 75% de tota la informació era analògica, l'any 2007 el 7% era analògica, i es preveu que l'any 2013 menys del 2% sigui analògica.

Les dades massives, principalment, consisteixen a fer prediccions. Normalment provoca confusió i error el fet que Big Data pertanyi a l'àrea d'aprenentatge automàtic. L'ús del Big Data no consisteix a "ensenyar" a un ordinador a "pensar" com un ésser humà, sinó que consisteix a aplicar matemàticament la gran quantitat de dades per tal de deduir i predir certes probabilitats. Quan teclegem la combinació de lletres "fantatic", l'ordinador sap la probabilitat que correspon a "fantàstic". A mesura que un ordinador rep més quantitat de dades, les seves prediccions seran més precises. D'aquesta manera els ordinadors s'actualitzen i es perfeccionen sols a mesura que passa el temps.

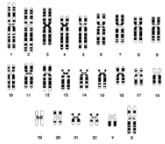
Amazon ens pot recomanar el llibre ideal, Facebook coneix els nostres gustos, Google ens pot dir la pàgina web més rellevant i LinkedIn endevina a qui coneixem. Què ens priva que en poc temps aquestes mateixes tecnologies siguin aplicades a diagnòstics de malalties, recomanació de tractaments o identificació de "delinqüents" quan encara no han comès el delicte?

LES XIFRES

Aviat la gent considerarà Big Data una moda; apareixerà en nombroses revistes i conferències, i moltes empreses emprenedores naixeran a causa de l'entusiasme per les dades. Però després aquesta moda baixarà i totes aquestes empreses se n'aniran en orris. Aquest procés pot provocar confusió en la importància del Big Data i del que està passant.

Per tal de corroborar la importància de la revolució de la informació, cal ser conscients de les xifres de les tendències que envolten la nostra societat.

- L'any 2000 el telescopi de Sloan Digital Sky Survey va recopilar més dades de les que s'havien acumulat en tota la història de l'astronomia en poques setmanes. Però es preveu que el 2016 el Gran Telescopi Sinòptic d'Investigació de Xile acumuli aquestes dades en tan sols 5 dies.



- L'any 2003 els científics van desxifrar el genoma humà en 10 anys de treball intensiu. El 2013 es va poder fer el mateix en tan sols un dia.

- Google processa més de 24 petabytes de dades al dia, milers de vegades més tota la informació guardada a la Biblioteca del Congrés d'EUA.

- A Facebook, amb una dècada d'edat, cada hora es pugen més de 10 milions de fotos. Cada dia els usuaris cliquen "m'agrada" o fan un comentari 3 milions de vegades. Tot això després ho utilitzen per crear i explorar les preferències dels usuaris.

- Cada segon es puja més d'una hora de vídeo a YouTube.

- El nombre de missatges creix un 200% a l'any a Twiter, i el 2012 van superar els 400 milions de piulades diàries.

- La quantitat d'informació emmagatzemada creix quatre vegades més de pressa que l'economia mundial.

Ciències, Finances, Internet, ... Els sectors que engloba són molt diversos, però tots junts desvien cap a una mateixa veritat: la quantitat de dades està creixent tant ràpid que se'n va de les nostres capacitats i de la nostra imaginació.

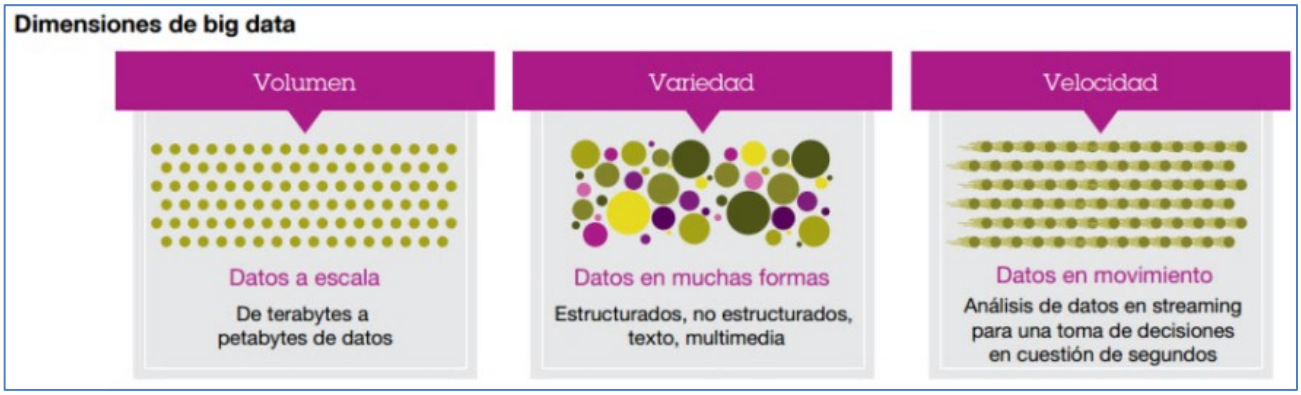
A hora d'ara podem afirmar i reafirmar que Big Data no és una moda. És un fet i un futur que Big Data pot aportar beneficis molt importants a qualsevol sector o indústria. Fins i tot Barcelona (de la mateixa manera que altres ciutats i països) té una Open Data (Big Data obert a l'excés del públic) que recull tota la informació administrativa, de l'entorn urbà, de la població, del territori i de l'economia.

Fins aquí podíem pensar, com moltes altres coses, que Big Data era tan sols paraules, una moda o una notícia del moment. Però un cop veiem les xifres es fa evident que no és això. No és una moda, sinó que és un fet, un procés i un progrés que ens porta a una nova societat i a una nova manera de viure. De la mateixa manera que els telèfons, que les pel·lícules i la televisió, Big Data no és un exemple però és una nova relació entre objectes i persones.

LES 3V

L'any 2001, Doug Laney va publicar un article ("3D Data Management: Controlling Data Volume, Velocity, and Variety") on englobava Big Data en tres conceptes, les tres V: volum, velocitat i varietat.

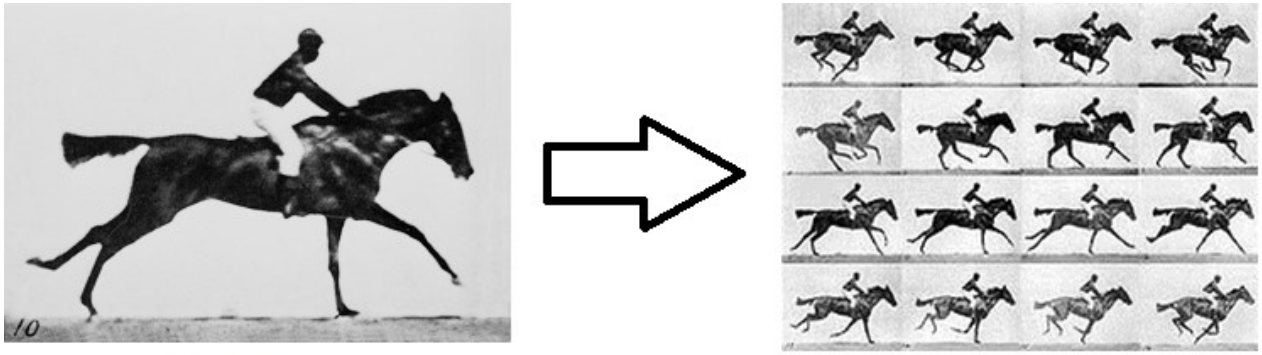
Avui dia, les tres V continuen sent les característiques més comunes acceptades del Big Data.



VOLUM

El volum fa referència a la gran quantitat de dades que s'intenten aprofitar. El volum de les dades creix molt acceleradament i es considera que a partir d'un terabyte ja és Big Data, és a dir, ja té la característica de volum.

Posaré un exemple molt clar que demostra la importància del volum de les dades. Posem per cas que volem la representació d'un cavall. Abans es necessitava molt de temps per dibuixar un cavall. Després de l'invent de les maquines de retratar, podem aconseguir una representació molt més ràpida mitjançant una fotografia. Això descriu un canvi, però no és aquest canvi l'essencial, ja que com a resultat obtenim el mateix; una imatge d'un cavall. Però suposant que en comptes de fer una fotografia, en fem 24 cada segon. Passem d'una simple fotografia a una pel·lícula, és a dir, el canvi quantitatiu provoca un canvi qualitatiu. En aquest exemple també hi entra en joc la velocitat.



Fotografia

Pel·lícula
(16 fotografies)

El fet que el Big Data estigui compost d'una enorme quantitat de volum, al moment d'aprofitar-lo i utilitzar-lo en empreses, ens plantegem algunes preguntes fonamentals.

- Existeix informació que ens ajudi a entendre millor els nostres clients?
- Existeix informació que ens ajudi a entendre les nostres operacions?
- Existeix informació que ens ajudi a entendre millor els nostres competidors?
- Què podríem fer per millorar la posició de la nostra empresa?

VELOCITAT

Una de les conseqüències de l'evolució de l'IOT és el creixement del Big Data. La velocitat de les dades s'incrementa exponencialment i la clau de l'èxit és saber com treure'n profit.

Avui dia, la majoria de les dades no són útils a llarg termini, sinó que són útils quan s'analitzen al mateix moment. I és quan els sistemes tradicionals són incapaços de captar, emmagatzemar i analitzar tota la informació que reben a temps real.

VARIETAT

Abans les dades provenien d'un cercle reduït de llocs i per tant els formats de les dades eren més concrets. Ara les dades provenen de molts llocs diferents i per tant les fonts i el format que arriben de les dades són molt diversos (vídeos, gravacions de veu, textos, pdf, etc). Això porta a diferents tipus de dades (vegeu fotografia gràfica): estructurades, semiestructurades, no estructurades (són la majoria, per exemple quan les dades provenen de les xarxes socials com Facebook) i *streaming*. La varietat de formats també és un dels motius pel qual els experts han d'actualitzar la seva capacitat de rebre dades en diferents formats.

En general, veiem que Big Data són dades a gran volum, varietat i velocitat. És a dir, podríem definir-lo amb les 3V. Són aquests tres factors que provoquen que Big Data sigui una cosa tan enorme (que se'ns escapa de les mans) i que a la vegada fa que sigui tant difícil aprofitar-lo. I és que com més dades hi ha, més difícil es fa aprofitar-les.

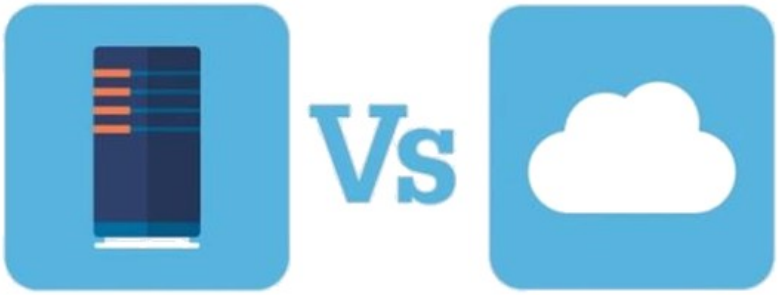
DATA CENTER

Des del primer moment que els éssers humans començàvem a utilitzar la informació (l'escriptura) ja emmagatzemàvem informació. Això sí, el format de la informació ha anat canviant. De fet, 300 anys aC la Biblioteca d'Alexandria era el centre d'emmagatzematge de dades més gran del món, fins que 48 anys dC els romans la van destruir. La informació (incloent el que serien les 3V: velocitat, volum i varietat) ha anat evolucionant al llarg dels anys. És així com l'any 1965 el Govern dels Estats Units tenia previst el primer centre de dades del món per emmagatzemar 742 milions de declaracions d'impostos en una cinta magnètica i 175 milions d'empremtes digitals.

Després de molts anys, el format en què guardem la informació, la quantitat i la varietat han canviat (*ergo* evolucionat) totalment. Vam passar de paper a cintes magnètiques, i de cintes magnètiques a la informàtica actual.
Per tant, actualment definim un centre de dades (Data Center) com un lloc, espai o instal·lació on s'emmagatzemen, es distribueixen i es tracten sistemes informàtics i els components associats com ara telecomunicacions, sistemes d'emmagatzematge, etc. Tots aquests recursos permeten el processament de la informació d'una organització o empresa.

Quan vaig sentir a dir que el conjunt de dades que formen el Big Data es guardava en un lloc físicament, em va sobtar molt. Tenia la idea que totes les dades, tot el que forma Internet, estava al núvol. I després vaig anar descobrint que el Data Center, no només existia sinó que eren espais grans, molt grans. El Data Center és una de les maneres de ser conscient d'on vivim, i fins a quin extrem arriba el Big Data.

DATA CENTER VS CLOUD (CENTRE DE DADES VS NÚVOL)



Molt sovint, el fet que siguin paraules noves i de moda, pot portar a confusions sobre els termes Data Center i Cloud. A afectes pràctics són sinònims. Però s'utilitzen amb un context diferent. Quan parlem de Cloud ens referim metafòricament a Internet i entenen que Internet està a l'aire. Però una persona informada sabrà que Internet no està a l'aire, sinó que està guarda al Data Center. Per tant, tots dos sistemes emmagatzemen dades, però Cloud és metafòric i Data Center és real, tal com realment és.

Per exemple, quan un di que té un document guardat al Cloud, aquell document no està a l'aire, sinó que realment està guardat en algun ordinador d'un Data Center.

NIVELLS (TIERS)

El 2005 es va classificar el Data Center en quatre nivells de fiabilitat, en el qual com més alt és el nivell més fiabilitat té el Data Center.

Nivell	Característiques
1	Sense components redundants* 99,671% de disponibilitat 28,8 hores d'inactivitat anual
2	Parcialment redundant* 99,741% de disponibilitat 22,7 hores d'inactivitat anual
3	Components redundants* (N+1) 99,982% de disponibilitat 94,6 minuts d'inactivitat anual

4	Totalment redundant* (2N+1) 99,995% de disponibilitat 26,3 minuts d'inactivitat anual

*REDUNDANT:

He vist definicions d'aquesta paraula que no s'entenen gens, he vist definicions que no defineixen, així que he decidit fer la meva pròpia definició perquè la pugui entendre jo i tothom que la llegeixi.

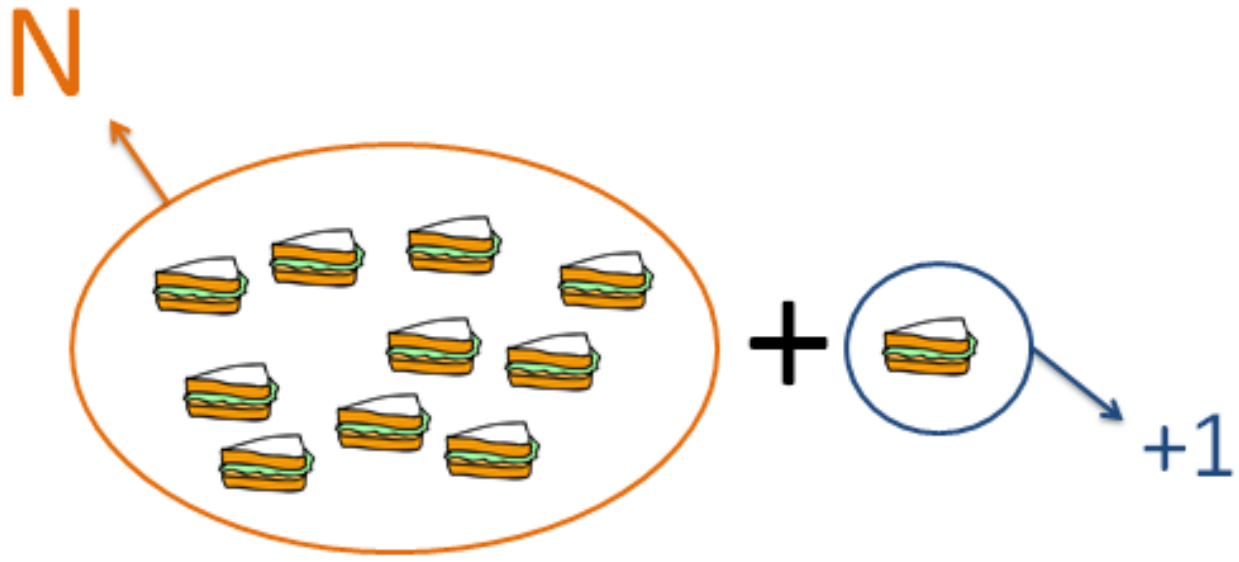
Quan diem que una cosa o un conjunt de coses és redundant ens referim a una cosa que té una còpia de seguretat tant en maquinària com en energia. D'aquesta manera és més difícil que falli i per tant és més fiable.

Però que vol dir això que N+1, 2N i 2N+1? Com que definir-ho seria complicat i poc entenedor, proposo un exemple senzill però complex que difícilment no us deixarà cap dubte.

N+1

Imaginem que tenim una festa d'aniversari (d'aquestes infantils) i volem preparar alguna cosa de berenar, com ara entrepans. Imaginem que tenim 10 convidats i per tant necessitarem 10 entrepans. Però sempre pot haver-hi algun convidat inesperat, i per tant farem 11 entrepans.

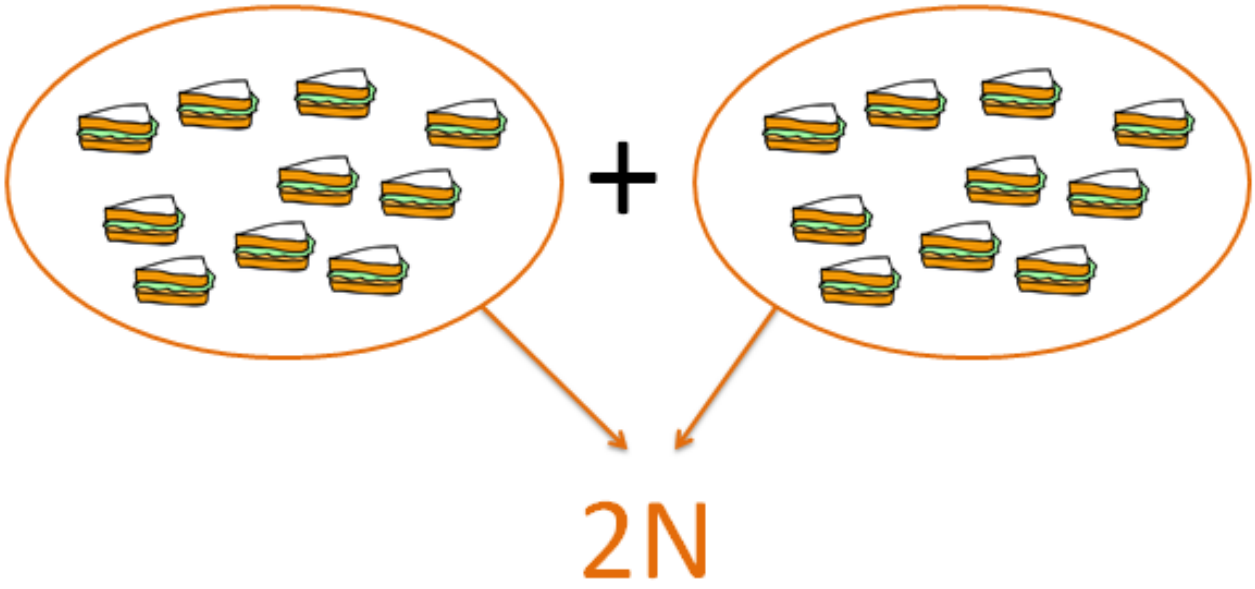
"N" representa la quantitat exacta d'entrepans que necessitem i el "+1" representa l'entrepà addicional. Per tant, tenim N+1 entrepans fets per a la festa.



Quan traspassem això a un centre de dades, N+1 sistemes és una garantia de disponibilitat i menys temps d'inactivitat. Però no és un sistema del tot redundant ja que pot fallar alguna cosa, ja que funcionen amb circuits comuns.

2N

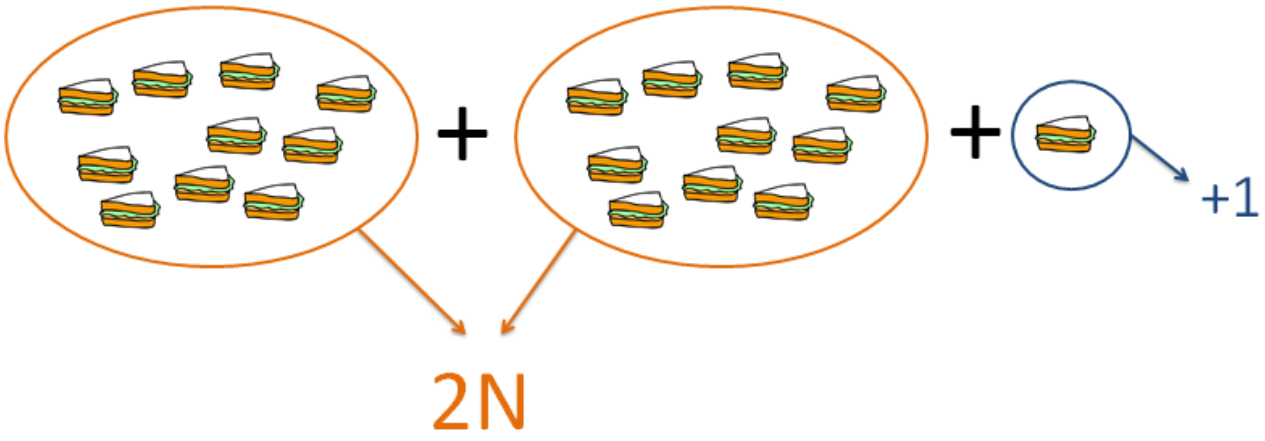
Si en comptes de voler 10 entrepans per als 10 convidats, volguéssim 10 entrepans més, és a dir, 20 entrepans. "2N" simplement representaria dues vegades (el doble de) la quantitat necessària.



Això traspassat a un centre de dades, voldria dir que tenim el doble de la quantitat de maquinària necessària, sense circuits comuns. Un sistema 2N és molt més fiable que un de N+1.

2N+1

Aquest últim sistema és simplement el resultat de sumar els dos anteriors. Tenim el doble de la quantitat de maquinària necessària, més un tros d'un equip. Així que, passat a l'exemple dels entrepans, tindríem 21 entrepans per a 10 persones; dos per persona i 3 per a tu!



GOOGLE DATA CENTER

Si he posat a Google com a exemple no és perquè tingui els centres de dades més grans, que no és el cas, sinó perquè és una empresa experta en l'àmbit del Big Data. Google ensenya al seu web els seus centres de dades i els ven positivament, però realment no ens explica gaire cosa. Això és així perquè Google considera que les seves operacions del centre de dades podrien donar avantatges competitius. Tampoc se sap del cert quants centres de dades té Google, però de moment Google n'ensenya 14 arreu del món a la seva web i ofereix fotografies i l'explicació de cadascuna.

Mapa d'ubicació dels 14 centres de dades de Google:

DEFINICIÓ

No existeix una definició universal de Big Data que cobreixi tots els aspectes i que tothom hi estigui d'acord. En aquest cas, cada organització o persona defineix el Big Data amb una definició pròpia segons el punt de vista en el qual l'enfoca. És per això que recolliré diferents definicions de diferents llocs per tal de contrastar i també, a la vegada, tenir clara la idea del Big Data.

Segons IBM (Institute for Business Value) i la Universitat d'Oxford: "Big data, un concepte que significa moltes coses per a moltes persones, ha deixat d'estar limitat al món de la tecnologia. Avui en dia es tracta d'una prioritat empresarial donada la seva capacitat per influir profundament en el comerç d'una economia integrada a escala global. A més de proporcionar solucions a antics reptes empresarials, Big Data inspira noves formes de transformar processos, empreses, sectors sencers i fins i tot la pròpia societat. Tot i així, l'àmplia cobertura mediàtica que està rebent no ens permet distingir clarament el mite de la realitat: què està passant realment? Després de la nostra última investigació hem descobert que les empreses utilitzen Big Data per obtenir resultats centrats en el client, aprofitar les dades internes i crear un millor ecosistema d'informació."

Segons marketingtechblog.com: "Big Data és un terme usat per descriure la recollida, el tractament i la disponibilitat de grans volums de dades de *streaming* en temps real. Format per les tres V que són el volum, la velocitat i la varietat, que neix amb Doug Laney. Les empreses estan

combinant màrqueting, vendes, dades de clients, dades transaccionals, converses socials i fins i tot les dades externes, com preus de les accions, clima i notícies per identificar la correlació i causalitat estadística de models vàlids per ajudar-los a prendre decisions més precises."

Segons OnERP (Solución de Gestión Empresarial Online): "Definim Big Data a la gestió i anàlisi d'enormes volums de dades que no poden ser tractats de manera tradicional, ja que superen els límits i capacitats de les eines de programari habitualment utilitzades per a la captura, gestió i processament de dades.
Aquest concepte engloba infraestructures, tecnologies i serveis que han estat creats per donar solució al processament d'enormes conjunts de dades estructurades, no estructurats o semiestructurats (missatges en xarxes socials, senyals de mòbil, arxius d'àudio, sensors, imatges digitals, dades de formularis, correus electrònics, dades d'enquestes, etc.) que poden provenir de sensors, micròfons, càmeres, escàners mèdics, imatges."

Segons la Viquipèdia: "El Big Data o dades massives és un concepte que fa referència a l'acumulació massiva de dades i als procediments usats per identificar patrons recurrents dins d'aquestes dades."

Segons Viktor Schönberger al llibre *Big data: La revolución de los datos masivos*: "Big Data (o les dades massives) es refereixen a coses que es poden fer a gran escala, però no a una escala inferior, per extraure noves percepcions o crear noves formes de valor, de tal forma que transforma els mercats, les organitzacions, les relacions entre ciutadans i els governs, etc."

En general, totes les definicions fan un recull o petit resum de tot el que he explicat aquí anteriorment, és a dir, tots els aspectes que envolten el Big Data. Les petites diferències es basen que en cada cas, segons com convingui, es fixen més en unes característiques o unes altres. És important veure que no es contradiuen, sinó que es complementen entre si. Per tant, ajuntant totes cinc definicions aconseguim una definició global, general i completa.

Al principi em va sorprendre molt que el concepte de Big Data no tingués cap definició establerta i oficial. Però a mesura que vaig anar-me documentant i avançant aquest treball vaig veure per què no hi havia cap definició concreta. Big Data és un terme i un concepte que comporta tantes coses, que té tantes característiques, tants punts de vista, tantes utili-

tats, tantes xifres... que és impossible que tothom que en faci ús, el faci de la mateixa manera, aprofitant les mateixes característiques i el mateix punt de vista.

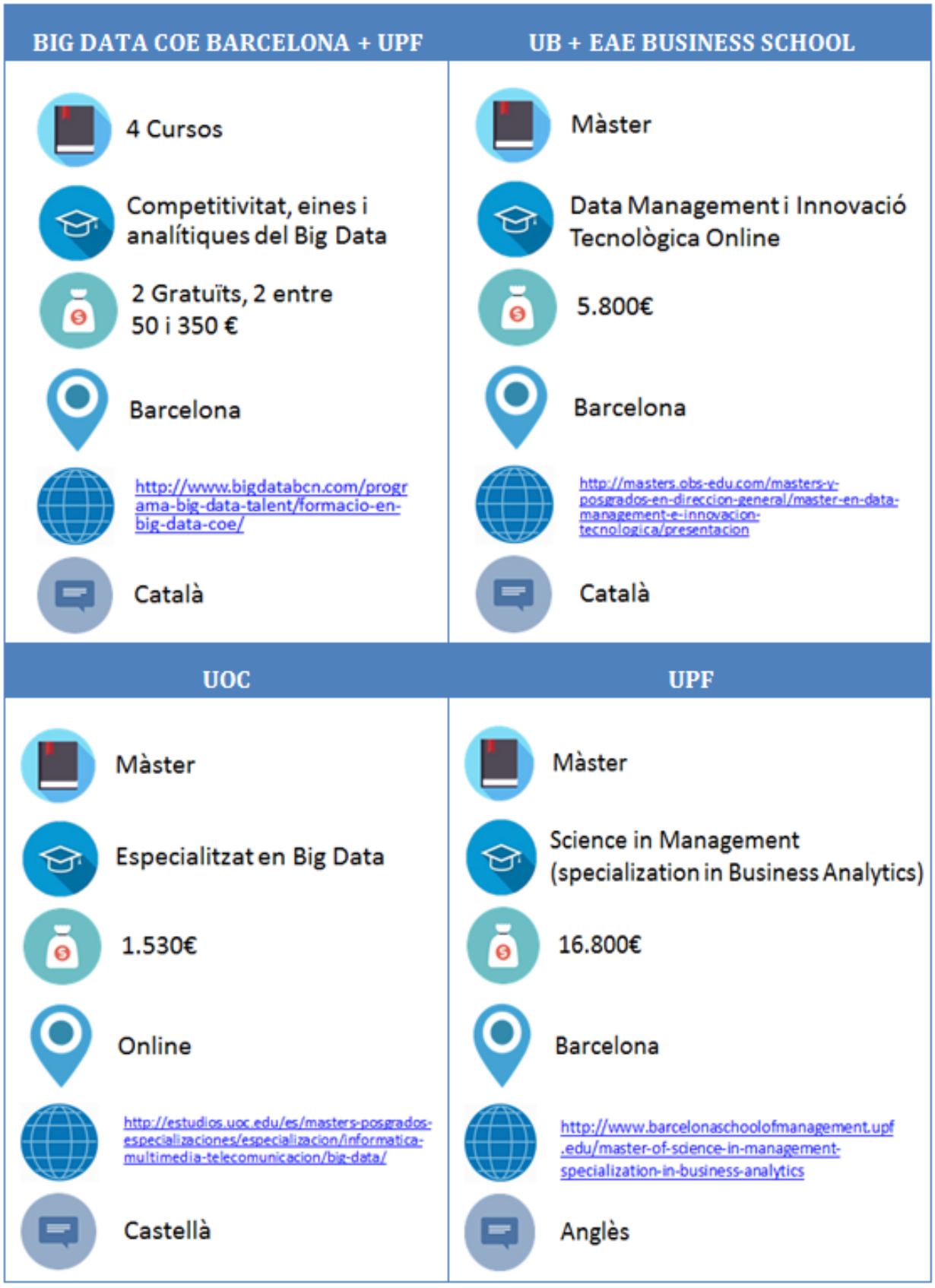
5. FORMACIÓ

Llegint atentament aquest Treball de Recerca podem aprendre molt i més sobre el Big Data, però sempre hi ha persones que els agrada anar més enllà. I també en el cas de ser un empresari o una empresària et pot interessar saber el màxim possible sobre aquest tema. És per això que he decidit fer un apartat dedicat a presentar un seguit de llocs on rebre formació de tot tipus sobre el tema.

He dividit la formació segons el lloc on es fa, ja que (menys en el cas de ser online) és important la situació geogràfica del lloc on fan la formació.

He recollit la formació que es dóna en els llocs més importants i representatius.

CATALUNYA


BIG DATA COE BARCELONA + UPF
4 Cursos
Competitivitat, eines i analítiques del Big Data
2 Gratuïts, 2 entre 50 i 350 €
Barcelona
http://www.bigdatabcn.com/programa-big-data-talent/formacio-en-big-data-coe/
Català
UB + EAE BUSINESS SCHOOL
Màster
Data Management i Innovació Tecnològica Online
5.800€
Barcelona
http://masters.obs-edu.com/masters-y-posgrados-en-direccion-general/master-en-data-management-e-innovacion-tecnologica/presentacion
Català
UOC
Màster
Especialitzat en Big Data
1.530€
Online
http://estudios.uoc.edu/es/masters-posgrados-especializaciones/especializacion/informatica-multimedia-telecomunicacion/big-data/
Castellà
UPF
Màster
Science in Management (specialization in Business Analytics)
16.800€
Barcelona
http://www.barcelonaschoolofmanagement.upf.edu/master-of-science-in-management-specialization-in-business-analytics
Anglès

ESPANYA

CIFF (UNIVERSIDAD DE ALCALÁ)	UNIVERSIDAD CARLOS III DE MADRID
Màster	Màster Universitari
Big Data y Business Analytics	Métodos analíticos para datos masivos: Big Data
10.200€	Entre 5.500 i 8.500€
Madrid	Madrid
http://www.ciff.net/master-en-big-data-y-business-analytics.html	http://www.uc3m.es/ss/Satellite/Postgrado/es/Detalle/Estudio_C/1371210340413/1371211096495/Master_Universitario_en_Metodos_Analiticos_para_Datos_Masivos:_Big_Data
Castellà	Anglès

EOI (ESCUELA DE ORGANIZACIÓN INDUSTRIAL)	KSCHOOL
Programa Superior	Màster
Big Data & Business Analytics	Data Science
6.500€	6.500€
Madrid	Madrid
http://www.eoi.es/es/cursos/17010/programa-superior-en-big-data-business-analytics-madridç	http://kschool.com/cursos/master-en-data-science/
Castellà	Castellà

PAÏSOS DE PARLA ANGLESA (UK I EUA)

BIG DATA UNIVERSITY	CITY UNIVERSITY LONDON
Curs	Master
Big Data Fundamentals	Data Science
Gratuït	Entre £4.500 i £14.500
Online	London
http://bigdatauniversity.com/bdu-wp/bdu-course/big-data-fundamentals/	https://www.city.ac.uk/courses/postgraduate/data-science-msc
Anglès	Anglès

QUEEN MARY (UNIVERSITY OF LONDON)	BRUNEL UNIVERSITY LONDON
Master	Màster
Big Data Science	Data Science and Analytics
Entre £4.000 i £8.00	Entre £4.500 i £16.500
London	London
http://www.qmul.ac.uk/postgraduate/coursefinder/courses/121386.html	http://www.brunel.ac.uk/courses/postgraduate/data-science-and-analytics-msc
Anglès	Anglès

En el cas d'Estats Units, existeix un mapa virtual on és molt fàcil i ràpid de trobar qualsevol màster arreu del país relacionat amb el Big Data. Està al següent enllaç: http://data-informed.com/bigdata_university_map/

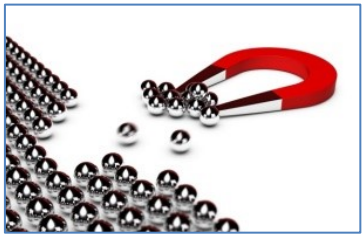
He pogut comprovar que arreu del món hi ha formació per saber sobre el Big Data, sobretot enfocat a les empreses. També que n'hi ha de tot tipus, de qualsevol preu i que sobretot són màsters. Això ens podria dir que el Big Data és una qüestió complexa, que requereix un nivell de formació superior.

6. ÚS DEL BIG DATA EN MÀRQUETING

MÀRQUETING: QUÈ ÉS?

Ha arribat el moment en què combinem el món del Big Data amb el màrqueting, és a dir, per a què ens serveix i com utilitzar el Big Data en màrqueting.

Hem descrit prou bé què és el Big Data i què comporta, però en canvi encara no en sabem res, del màrqueting. Com que el màrqueting no és una cosa tan nova ni molt menys, i més o menys tothom en té alguna idea, aquest apartat serà més curt, senzill i clar.



El màrqueting és un concepte que cadascú defineix segons com un creu i segons els seus interessos. Però generalment, el màrqueting és un conjunt d'activitats amb l'objectiu de satisfer les necessitats dels consumidors mitjançant un producte. Per dur a terme el propòsit, el màrqueting disposa d'un gran ventall d'eines.

El que s'ha de tenir clar és que màrqueting no és igual a publicitat. En canvi, la publicitat és una de les eines del màrqueting.

AVANTATGES I BENEFICIS

És una evidència que el Big Data té avantatges, i molt importants, ja que és un fenomen que està revolucionant el món empresarial. Però la pregunta és: Quins avantatges?

DECISIONS MÉS RÀPIDES I MILLORS

Les empreses sempre s'han mogut per trobar la satisfacció del client, mitjançant decisions sobre el seu producte. El Big Data potencia enormement això, ja que contínuament es recullen dades a temps real. Si l'empresa aconsegueix aprofitar-les, aconseguirà avançar-se i adaptar-se

a les necessitats del client, abans que aquest es mostri insatisfet i per tant crear fidelitat. A més a més, si bases les teves decisions en el Big Data, hi ha menys risc ja que et bases en l'experiència del client.
Aquestes decisions més ràpides i millors redueixen els costos i augmenten l'eficàcia de les vendes.

NOUS PRODUCTES I SERVEIS

La part més creativa i interessant en l'aplicació del Big Data és quan, a partir de grans dades, sabem la necessitat del client i per tant podem crear nous productes i serveis per al client que hem analitzat. Per exemple, segons el que vol el teu client, pots crear ofertes, o altres productes a partir d'un altre.

MILLORES DINS L'EMPRESA

El Big Data és en part conseqüència de la digitalització de les dades. Un cop tens les dades digitalitzades, comptes amb eines que faciliten la cerca de la informació i això ajuda a fer una feina més dinàmica, ràpida i eficaç.
També, en tenir les dades digitalitzades, es poden fer anàlisis de com funciona l'empresa i d'aquesta manera veure quins són els punts forts i els punts dèbils, i quines necessitats té l'empresa.
S'ha fet una enquesta i s'ha trobat que, segons *Ivey Business Journal*, els treballadors gasten un 25% de l'esforç en la cerca de les dades. En reduir això digitalitzant-ho tot, reduïm un esforç que era ineficient. A més a més, la informació digitalitzada pot ser més precisa i detallada.
Per tant, podem dir que el Big Data presenta una millora en l'accessibilitat i fluïdesa de la informació.

MILLORAR L'EFICIÈNCIA DELS PRODUCTES

Quan una empresa analitza una gran quantitat de dades sobre un producte, obté una informació molt valuosa que l'ajuda a evolucionar i a fer créixer més ràpidament el producte, sobretot si parlem de Big Data a temps real.
També ajuda a veure que si un producte no va pel bon camí, pot crear-ne un altre de nou o a redissenyar-lo.

A més a més, abans de llençar el producte, gràcies al Big Data l'empresa pot fer simulacions d'aquest producte, i d'aquesta manera reduir costos fins a un 30 o 50%.

MILLORAR LA COMPETITIVITAT

Poder modificar el producte segons el comportament del client per tal de guanyar fidelitat o atreure clients, gràcies a la digitalització de les dades és el que permet a l'empresa tenir un avantatge competitiu envers altres empreses que no utilitzin el Big Data.

Dit d'una altra manera, el Big Data ajuda les empreses a apropar-se als seus clients, comparant el que aquests volen i el que realment reben.

Però està clar que cal esperar que les empreses inverteixin més en Big Data per tal de veure clarament l'avantatge competitiu que dóna.

Segons *Ivey Business Journal* "Totes les empreses han de prendre Big Data i el seu potencial per crear valor seriosament si volen competir".

Els avantatges que mostra el Big Data engloba tot allò que gira al voltant dels objectius d'una empresa. És una demostració de com el Big Data és un concepte molt ampli. En els avantatges es mostra com les dades, depèn de com les analitzem i manegem, poden tenir múltiples funcions.

DESAVANTATGES

Tot allò que té uns avantatges, per regla general, també comporta uns desavantatges. I el Big Data no n'és cap excepció.

En aquest cas, com en molts d'altres, com que el Big Data és una "cosa" gran i potent, genera molts dubtes ja que suposa un risc per al fracàs. És per això que en aquest cas els desavantatges són importants.

Per començar, el desavantatge més important, com acabem de dir, és el risc que l'empresa fracassi.

Tot seguit exposem diferents desavantatges més concrets i específics, però també molt importants. Alguns desavantatges són més aviat reptes que ha d'afrontar l'empresa per tal de gaudir de l'ús del Big Data.

BON ENFOCAMENT EN L'ÚS DEL BIG DATA

Hi ha moltíssimes dades. Tantes, que s'ha de saber quines dades concretes es busquen i analitzar-les correctament, enfocant-les en allò que interessa. Si no és així, l'empresa es troba davant de moltíssimes dades que no li "diuen res" i és quan es desaprofita el Big Data.

L'empresa ha de saber què és el que vol saber i com ho vol saber. A partir d'aquí, ha de seleccionar aquelles dades que li facin servei i analitzar-les correctament. (No és que hi hagi una manera correcta general, sinó correctament per tal d'aconseguir aquella informació que l'empresa té com a objectiu aconseguir.)

Això també inclou trobar persones capaces de fer-ho, és a dir, experts en dades. Aquesta classe de persones són escasses i per tant, difícils de trobar.

PRIVACITAT ATACADA

Un dels desavantatges i alhora problema més estès i reconegut a causa del Big Data és la privacitat i la vigilància. És per això que anteriorment hi he dedicat un apartat exclusiu sobre el tema (pàg. 27). Però cal esmentar-ho aquí també per deixar clar que, des d'un punt de vista global i personal, el Big Data pot perjudicar la nostra privacitat.

LEGALITAT

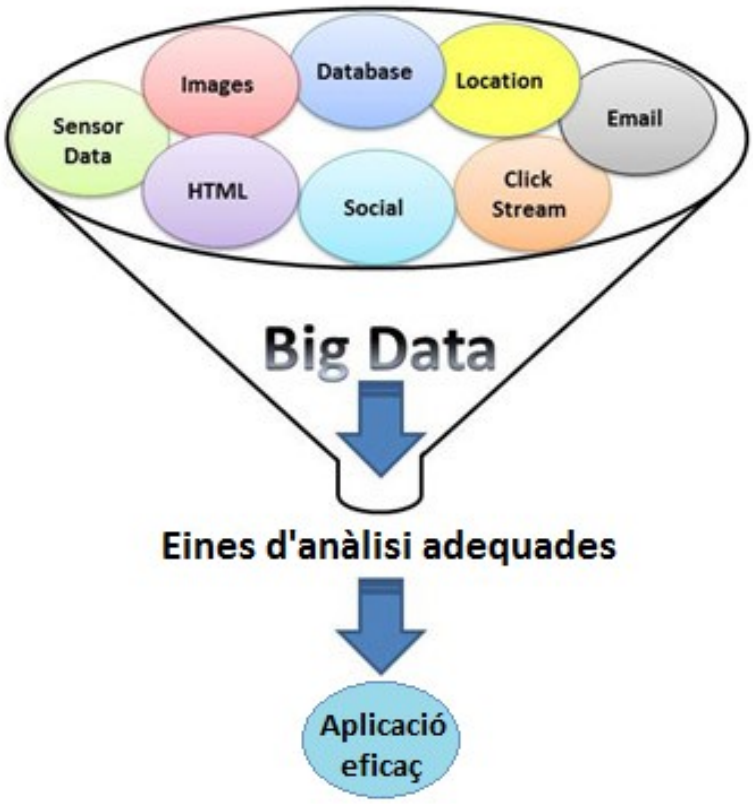
També és un problema actual per a les empreses la legalitat sobre la protecció de dades, ja que sovint poden traspassar els límits i ficar-se en problemes. De fet, Google mateix ha rebut un nombre elevat de denúncies pel tema de la protecció de dades.

DISCRIMINACIÓ

El fet de tenir poca privacitat, ens porta a la capacitat de discriminar més fàcilment. A causa del Big Data, hi ha un cert neguit que la discriminació pugui ser fins i tot automàtica.

EINES ADEQUADES

A la vegada que el Big Data es fa més gran, més potent i més eficaç, també ho han de fer les eines amb les quals es tracta. Aquest és un dels problemes més bàsics, ja que és difícil triar i disposar de l'eina adequada, i és més difícil encara saber-la utilitzar correctament. Tot i això, existeixen aquestes eines, però escasses i justes.



M'he adonat que els desavantatges que comporta el Big Data no són desavantatges en si, com ara els afectes secundaris de les medicines, que són inevitables. Aquests desavantatges són més aviat dificultats que es troben per tal de poder arribar als avantatges.

7. SEGMENTACIÓ

L'avantatge més important que proposa el Big Data és la segmentació dels clients.
L'objectiu essencial del màrqueting en una empresa és satisfer les necessitats del seu consumidors. Però sovint els clients o consumidors no s'assemblen entre ells, tenen necessitats diferents o les necessiten de diferents maneres.... És a dir, cada client té unes característiques diferents com a consumidor. És per això que es creen segments (grups) de consumidors amb característiques i comportament semblant.

QUÈ ÉS?

El mercat és un conjunt de persones o empreses amb característiques, necessitats i preferències heterogènies. Majoritàriament no es pot considerar el mercat com una unitat ja que dins d'un mercat cada consumidor té un comportament diferent envers un producte.
En el moment que les empreses coneixen les diferències i semblances entre consumidors, pot crear grups segons unes necessitats i preferències semblants, és a dir, grups amb característiques homogènies.

Aquests grups homogenis ajuden l'empresa a adaptar l'oferta del seu producte en tots els àmbits i per tant oferir un millor servei i cobrir millor les necessitats dels clients.
La segmentació de mercat, per definició, és un procés que divideix un mercat en grups homogenis de consumidors amb característiques i necessitats semblants.

Poques vegades ens plantegem perquè hi ha anuncis que ens interessen i d'altres que no. Quan de petita mirava la televisió i feien algun anunci que no li veia ni cap ni peus exclamava: "Però a qui li importa aquest anunci!". I la meva mare sempre em deia el mateix: "Això és perquè no va dirigit a tu". I ens parem a pensar, què vol dir que un anunci no va dirigit a mi? És a dir, quan una persona adulta mira anuncis que fan als canals infantils, notarà que difícilment hi hagi un anunci que li interessi.

Això té una resposta molt senzilla: segmentació. Els anuncis són un gran exemple de la segmentació. Cada anunci va dirigit a un públic específic i és per això que hi ha alguns anuncis que ens interessen més i d'altres que menys.

AVANTATGES

La segmentació presenta un seguit d'avantatges que són claus en el màrqueting.

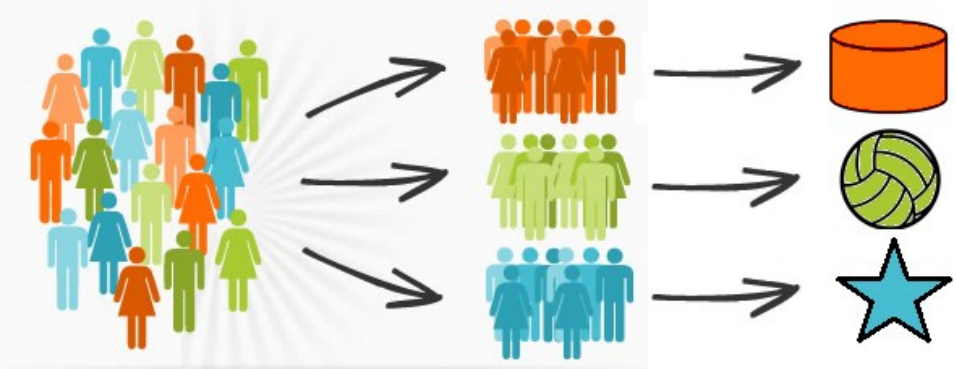
- **Permet identificar els segments de mercat més atractius.** Durant el procés de segmentació s'identifiquen els diversos segments i és quan l'empresa es planteja quins grups seleccionarà. L'empresa es fixarà en el grup que tingui una necessitat que cal satisfer, que tingui potencial de comprar, de fàcil accés i la capacitat d'adaptar-se a les seves necessitats.
- **Facilita l'anàlisi de les competències.** Si ens trobem davant d'una competència que ha segmentat, podem conèixer millor el seu producte i les seves estratègies de màrqueting per tal de trobar-hi forats, és a dir, les necessitats que no estiguin cobertes.
- **Permet adaptar el producte** a les necessitats del consumidors per tal de satisfer-lo millor. Si adaptem el producte als gustos, necessitats,... provocarà una millora de posicionament envers la competència. Com millor sigui el posicionament del producte més possibilitats tindrà de créixer.

- **Permet adaptar les estratègies de màrqueting** a cada segment. El fet de tenir analitzats cada un dels segments permet la creació d'estratègies de màrqueting específiques per a cada segment.
- **Permet d'identificar oportunitats de negoci.** En segmentar el mercat un es pot trobar amb necessitats que encara no s'han satisfet i per tant, l'empresa té una oportunitat nova de negoci que pot explotar.
- **Permet d'adaptar-se millor al client.** Gràcies a una millora de coneixement dels segments, l'empresa pot adaptar la seva oferta a les necessitats i comportaments corresponents.

En general, l'avantatge que mostra la segmentació és que ens permet personalitzar quin producte, com mostrar-lo, quan ensenyar-lo, què explicar, etc.

OBJECTIU DE LA SEGMENTACIÓ

L'objectiu és bàsic i clar, i és poder aplicar una estratègia de màrqueting diferenciada i especialitzada per a cada segment. D'aquesta manera serà més eficaç i de més qualitat.



CRITERIS DE SEGMENTACIÓ

Hi ha establerts de forma general una sèrie de criteris que ens ajuda a tenir una idea de com segmentar, i en què es fixen o es poden fixar les empreses de nosaltres per tal de segmentar-nos.

Els criteris a seguir més comuns són els criteris segons les característiques del consumidor.

- **Variables geogràfiques;** determinades pel país i la seva mida, la comunitat, la regió, el municipi, el barri, l'hàbitat rural o urbà, i la climatologia.

Aquesta variable, a causa de la globalització, pot ser que s'estigui desfent i perdent importància, ja que cada vegada és menys important la situació geogràfica on visquis.

- **Variables sociodemogràfiques;** són les més fàcils d'identificar. Són l'edat, el sexe, l'estat civil i el tipus (nombre) de família, els ingressos, la professió, el nivell educatiu, la religió, la nacionalitat, etc. Aquestes variables sovint estan relacionades amb els desitjos i les preferències dels clients.

- **Variables psicogràfiques;** que permeten la segmentació segons la personalitat i l'estil de vida. Però són les variables més difícils de mesurar ja que són subjectives. Per identificar la personalitat ens fixem en l'actitud, hàbits i preferències. Mentre que per identificar l'estil de vida ens fixem en les aficions, els interessos, les ideologies, etc. També són molt importants els valors.

A més a més dels criteris segons les característiques del consumidor, també podem establir criteris segons el comportament del consumidor.

- **Variables de comportament;** són aquells que segmenten els consumidors segons l'ús que fan del producte, la fidelitat amb la marca, quins beneficis busquen en fer la compra,

l'actitud envers el producte (entusiasme, indiferència o hostilitat), el moment i la freqüència d'ús.

SEGMENTACIÓ + BIG DATA

Per poder dur a terme la segmentació, s'ha de tenir un cert coneixement del consumidor. Proporcionalment, com més informació tinguem dels consumidors, més bona i precisa serà la segmentació i per tant els beneficis i els avantatges que comporta la segmentació s'accentuen i milloren de forma clara.

Però, de quina manera podem aconseguir molta més informació en menys temps i més valuosa? Utilitzant el Big Data. Però, exactament en què ajuda el Big Data?

Gràcies al Big Data, la informació que sabem és més específica i detallada i això permet dirigir els productes molt més adaptats i específics a cada consumidor. És a dir, el Big Data permet personalitzar el producte per a cada consumidor. Això s'anomena la microsegmentació, s'utilitza sobretot en anuncis personalitzats i funciona a temps real.

Ha quedat clar que la segmentació és una eina molt útil i important. En el moment que sumem aquesta eina amb Big Data, en part també una eina que ens amplia informació de tot allò que vulguem saber, ens surt una súper eina molt important i molt útil.

La segmentació és la part més important a l'hora de fer màrqueting, i el Big Data l'ajuda a ser-ho i a fer-ho millor.

8. ANÀLISI

Moltes empreses tenen grans, per no dir enormes, quantitats de dades. Però tota aquestes dades en realitat no signifiquen gaire si no es poden emmagatzemar, i després analitzar per tal de descobrir una nova o millor visió sobre els clients, productes, operacions, etc. de l'empresa.
Moltíssimes empreses, per no dir gairebé totes (sobretot les que tenen un espai a Internet), tenen a disposició grans quantitats de dades sobre el funcionament de la seva empresa entorn als clients i productes, per exemple. D'aquesta majoria, una part, és a dir un nombre elevat d'empreses del total, és capaç d'emmagatzemar-les. Però només una petita minoria és analitzat i per tant, només aquesta petita minoria aprofita el que és realment el Big Data. Això és degut als desavantatges que comportava (costos, dedicació, riscos, etc.).
L'anàlisi del Big Data és el procés d'examinar-lo per tal de descobrir patrons, tendències i correlacions desconegudes i treure'n informació útil. Aquesta informació és la que ens permetrà gaudir dels avantatges del Big Data (pàg. 70).
La clau està a convertir les dades en informació, la informació en coneixement i, finalment, el coneixement en una estratègia per a l'empresa.

OBJECTIU DE L'ANÀLISI

L'anàlisi de les dades té com objectiu principal ajudar les empreses a millorar les decisions de negoci.

TÈCNIQUES D'ANÀLISI

Existeixen diverses tècniques d'anàlisis de dades, de les quals explicaré quatre; les principals.

ASSOCIACIÓ

Permet trobar relacions entre diferents variables. Entenent que els fets són causals, aquesta tècnica pretén trobar una predicció del comportament d'altres variables. Això ens portaria a les relacions sistemàtiques de "vendes creuades" (pàg. 87).
Un exemple molt clar és quan Facebook ens prediu quines persones ens interessen (hi ha més possibilitats de) per fer-nos amics. Un altre exemple d'ús és quan s'analitza el que compra la

gent, que permet determinar quins productes es compren sovint junts, com ara el descobriment, segons BigData-MadeSimple.com (un centre d'informació bona i completa, oberta format per un banc d'activistes de Big Data construint una comunitat Big Data arreu del mon), que els compradors de bolquers tendeixen a comprar també cervesa.
Per a més informació detallada, l'article "Relaciones Causales en Reglas de Asociación", de M. Amparo Vila, Daniel Sánchez y Luis Escobar.

VENDES CREUADES

És una tàctica en la qual el venedor promociona la venda de productes complementaris als que consumeix (o té intenció de consumir) el client. D'aquesta manera, el client comprarà més i l'empresa incrementarà les vendes entre un 10 i un 30%.
Un exemple molt clar i que tothom coneix és el moment en què vas a comprar un *smartphone*, et recomanen la compra d'una funda, protectors de pantalla o altres accessoris
Aquesta tàctica és molt usada a Internet [en part gràcies al "poder" de les *cookies* (pàg. 33)], com per exemple a Amazon:

MINERIA DE DADES (DATA MINING)

Permet trobar comportaments que es puguin predir. És un conjunt de tècniques que estan directament relacionades amb els models que s'utilitzen per crear i descobrir patrons en Big Data. La mineria de dades és una combinació dels mètodes d'estadística i de la màquina d'aprenentatge (Machine learning; pàg. 89). Per tant, l'objectiu principal de la Mineria de Dades és la predicció.
El concepte de "mineria de dades" cada cop és més popular en els negocis com a eina de gestió de la informació.
El procés de mineria de dades consta de tres etapes:

1. EXPLORACIÓ INICIAL

Aquesta etapa comença amb la preparació de les dades, és a dir, neteja de les dades, transformacions, seleccionar subconjunts.

2. IDENTIFICACIÓ DE LA CONSTRUCCIÓ DE MODEL O PATRÓ I VALIDAR-LO

En aquesta etapa es consideren diversos models i en funció del rendiment predictiu es tria el millor. Pot semblar la fase més senzilla, però és on s'utilitzen més varietat de tècniques, ja que pot arribar a ser un procés molt elaborat.

3. APLICACIÓ DEL MODEL AMB LES NOVES DADES I AIXÍ GENERAR PREDICCIONS

L'etapa final comporta utilitzar el model seleccionat anteriorment. S'aplica a noves dades, es generen prediccions i es compara el resultat esperat amb l'obtingut.

AGRUPACIÓ (CLUSTERING)

És una de les tècniques que engloba la minaria de dades. Permet dividir grups en subgrups segons semblances que eren desconegudes abans de fer l'anàlisi. L'objectiu és trobar semblances entre grups.

Per exemple, ens trobem davant d'una noia de 18 anys de València i un home de 40 anys de Madrid. Què tenen en comú? Doncs que als dos els agrada fer puzzles en el seu temps lliure.

MACHINE LEARNING (MÀQUINA D'APRENENTATGE)

Aquesta tècnica es basa a evolucionar un comportament basat en dades empíriques (és a dir, reals) mitjançant algoritmes. Fent això, aconsegueix un aprenentatge automàtic (part de la intel·ligència artificial), és a dir, aprendre de forma automàtica per tal de reconèixer patrons i prendre decisions intel·ligents basades en dades.

Hi ha dos tipus d'aprenentatge: supervisat i no supervisat.

SUPERVISAT

"És el conjunt de tècniques d'aprenentatge automàtic que infereixen una funció o relació d'un conjunt de dades d'entrenament."

Un exemple és l'anàlisi de sentiments en xarxes socials, que es classifiquen segons les opinions (de favorable a desfavorable).

NO SUPERVISAT

"És un conjunt de tècniques d'aprenentatge automàtic que es troba ocult en l'estructura de dades no etiquetats."

Un exemple és la segmentació dels clients segons les seves característiques, sense necessitat que el client faci una classificació prèvia.

9. HADOOP

Es necessiten eines per tal de tractar el Big Data. N'hi ha diverses, però la més important i utilitzada és Hadoop.

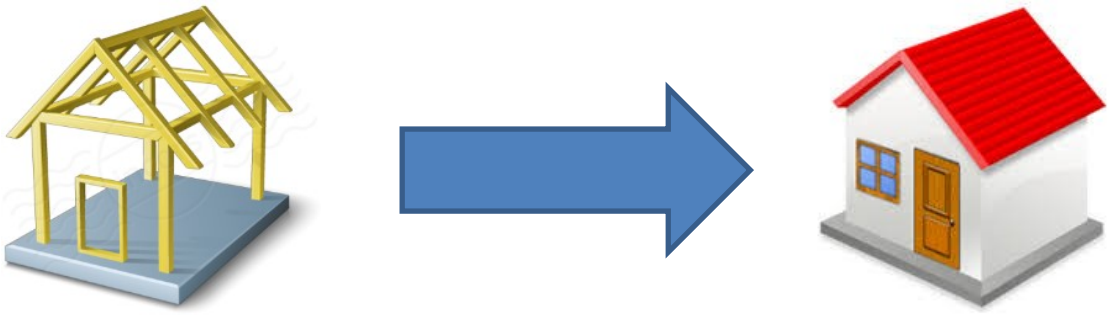
Hadoop és un *framework** que s'utilitza per emmagatzemar, processar i analitzar grans volums de dades, és a dir, el Big Data. Ho fa a través de clústers* mitjançant un model simple de programació.

FRAMEWORK

Un *framework* és una estructural conceptual (o conjunt de peces, metafòricament) i tecnològica que serveix de base per utilitzar i desenvolupar aplicacions (software).

Però un *framework* no només existeix en l'àmbit tecnològic. Té sentit en tots els àmbits, sobretot en l'àmbit de treball, de construcció i d'art.

Un exemple molt senzill és el següent:



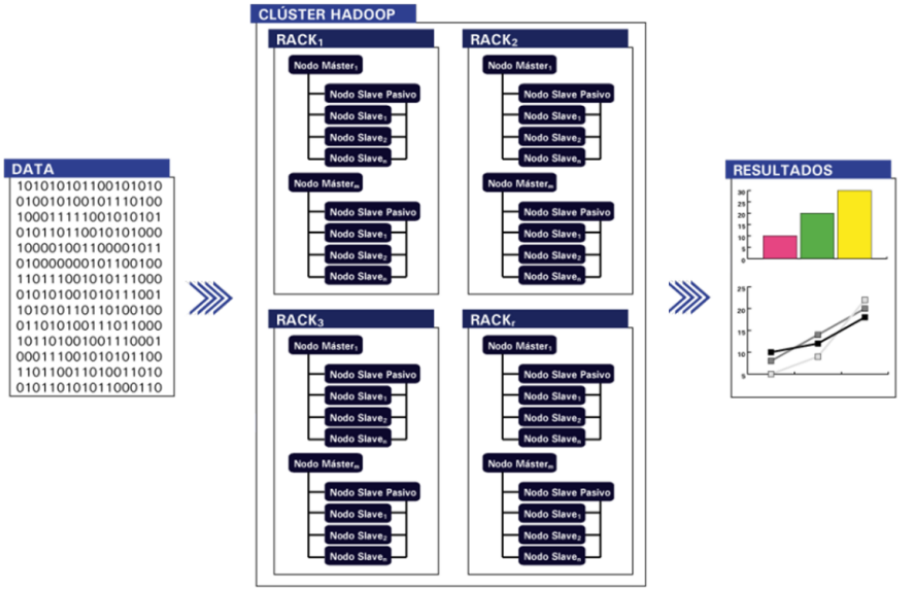
Estructura
(Framework)

Casa
(software)

CLÚSTERS

La definició de clúster varia molt segons l'àmbit. Però tots tenen una mateixa base, i és que un clúster és una agrupació de X coses (Y) que tenen alguna cosa en comú.

Per tant, quan diem que Hadoop funciona mitjançant clústers és el mateix que dir que Hadoop funciona mitjançant agrupacions de dades que tenen alguna cosa en comú. D'aquesta manera, es creen sectors de mercat (segmentació) i tenim tot una sèrie de perfils extrets de les dades que estan agrupats segons les seves semblances.



Node: és un registre que conté una dada d'interès i que s'enllaça amb almenys un altre node. És així com es creen llistes, arbres, jerarquies, etc. segons el nombre de punters que tenen.

Node Màster: és el responsable de mantenir l'estatus dels seus *nodes slave*. Un d'ells s'estableix com a *node slave* passiu, que es convertirà en *node màster* en cas que aquest es bloquegi.

Node Slave: és el node encarregat d'emmagatzemar la informació que s'està processant pels *node màster*.

Rack: és el nom que rep una combinació de nodes específica. Pot tenir un màxim de 40 *nodes màster*. Cada *rack* té la capacitat de comunicar-se amb els altres *racks*.

Un avantatge que tenen els clústers de Hadoop és que poden treballar amb enormes quantitats de dades, que poden ser tant estructurades com no estructurades.

Els clústers de Hadoop augmenten considerablement la velocitat en l'anàlisi de les dades i també creen seguretat, ja que creen còpies contínuament.

L'any 2013, Facebook va ser reconegut per tenir el clúster de Hadoop més gran del món.

COMPONENTS DE HADOOP

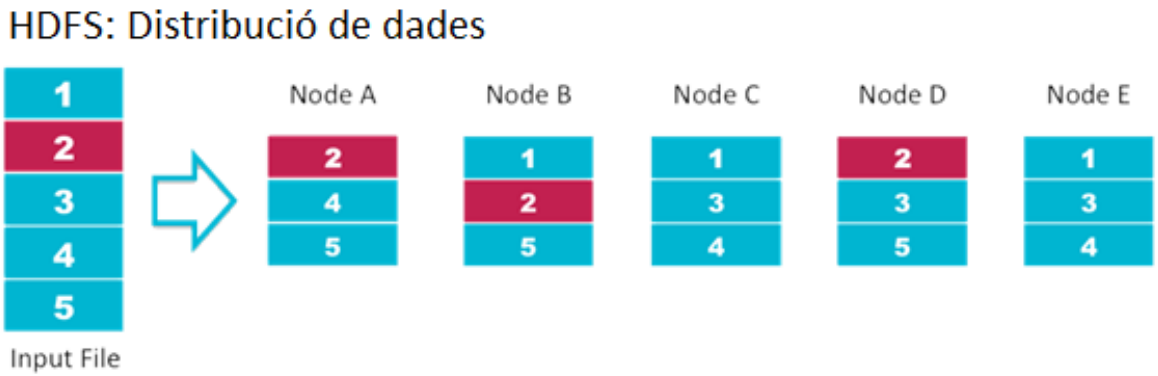
Hadoop consisteix en els següents dos components:

- Emmagatzematge de dades de codi obert, anomenat **HDFS** (Hadoop Distributed File System)
- Processament API, anomenat **MapReduce.**

El més comú és que Hadoop també inclogui altres projectes o biblioteques. Actualment n'hi ha més de 25 de diferents, entre els quals els més comuns són **HBase**, Hive i Pig.

(HADOOP DISTRIBUTED FILE SYSTEM) HDFS

HDFS és un sistema d'arxius que està dissenyat per emmagatzemar i gestionar Big Data de forma fiable. HDFS té una capacitat de 200 petabytes d'emmagatzematge. Funciona a través de clústers. Normalment té fallades al servidor a causa que sempre s'utilitza en maquinària de baix cost (Linux). Però el sistema està dissenyat per ser tolerant a les fallades.

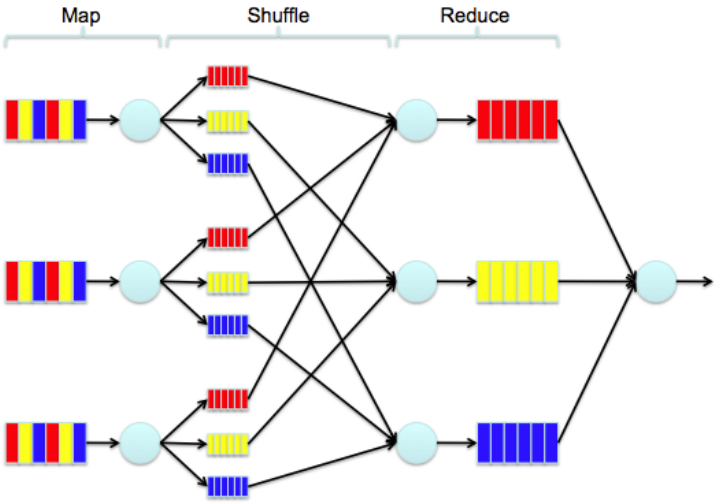


MAPREDUCE

MapReduce és un model de programació i també el component més important que forma Hadoop (és el cor de Hadoop, metafòricament).

MapReduce està format per dues parts.

- Procediment Map, que filtre i ordena les dades. És a dir, disposem d'un conjunt de dades, que les converteix en un altre conjunt de dades, però on cada element se separa (es desglossa) en columnes.
- Procediment Reduce, que analitza les dades (freqüències, recompte, mitjanes, etc.). És a dir, fa la feina de reducció de les dades.



MapReduce, de la mateixa manera que HDFS també està preparat per a una gran redundància (pàg. 55) i una tolerància als errors.

La funció principal de MapReduce és organitzar i reduir el resultats (de cada Node) per tal que en surti una resposta coherent a una consulta.

Originalment el nom de MapReduce es referia únicament a la tecnologia de Google, però avui dia s'ha convertit en un concepte genèric.

CAS EXEMPLAR

MapReduce se'ns pot quedar com una cosa abstracta, és a dir, un concepte que entenem de què va però que no sabem realment com funciona. Podríem buscar i buscar i trobaríem moltíssims llocs on explica el funcionament (a fons) de MapReduce. Però hi ha un problema important, i és que les explicacions exigeixen un nivell (tant de llenguatge com de coneixement) informàtic i tecnològic professional, que no tenim ni l'autor ni els lectors.

És per això que he trobat una altra sortida; explicar-ho a través d'un exemple senzill. Sembla utòpic, però a través d'aquest exemple es pot entendre fàcilment el mateix que moltes paraules específiques d'informàtica et dirien i no entendries.

Suposem que tenim cinc arxius i cada arxiu conté dues columnes (una clau i un valor en termes de Hadoop) que representen una ciutat i la temperatura corresponent registrat en aquesta ciutat per als diferents dies de mesurament.

En aquest exemple, la ciutat és la clau i temperatura és el valor.

Arxiu 1:

Toronto, 20

De Whitby, 25

Nova York, 22
Roma, 32
Toronto, 4
Roma, 33
De Nova York, 18

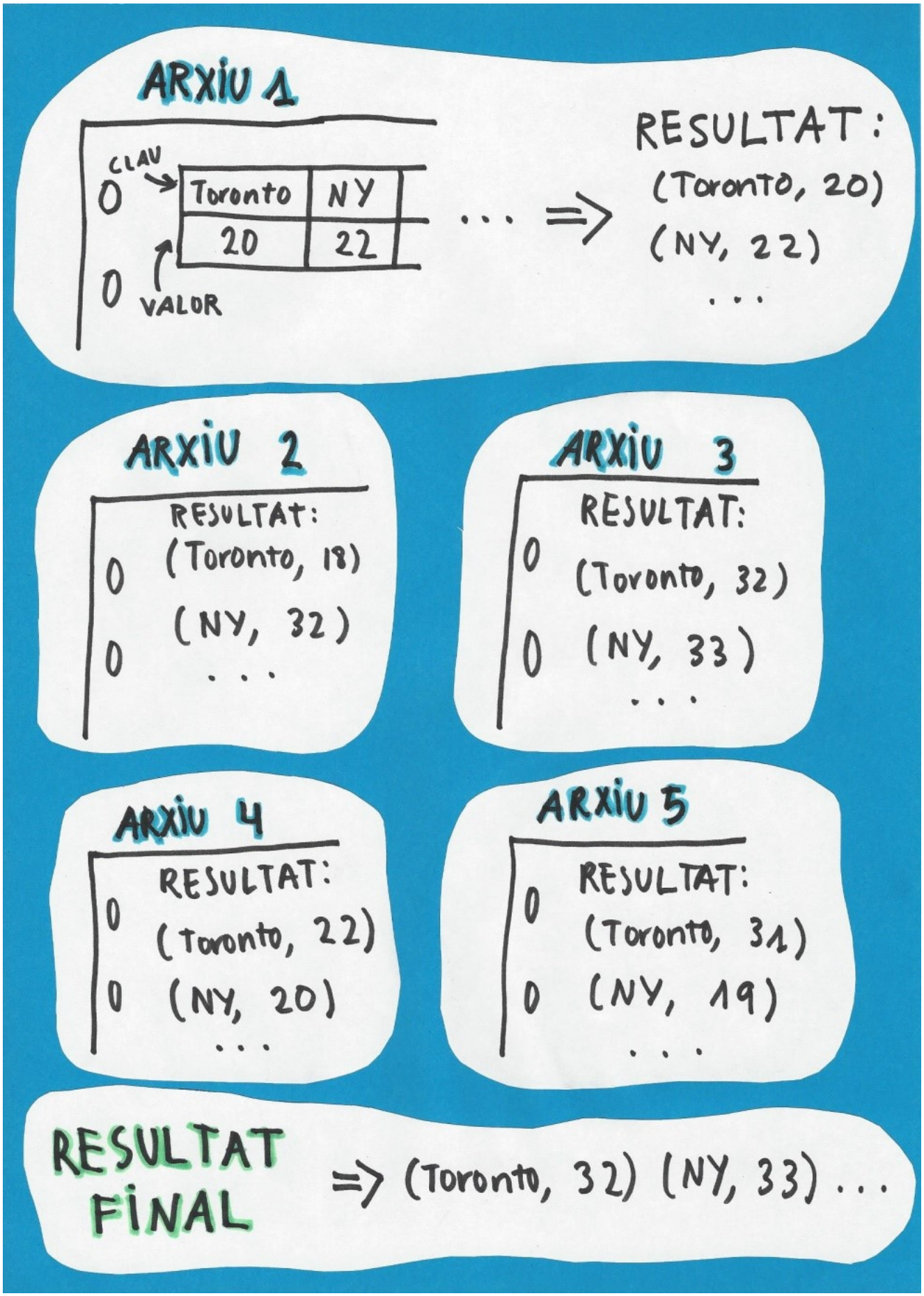
Volem trobar la temperatura màxima per a cada ciutat a través de tots els arxius de dades (hem de tenir en compte que cada arxiu pot tenir la mateixa ciutat representada diverses vegades). Utilitzant el marc MapReduce, podem descompondre'l en cinc tasques, on cada *mapejador* treballa en un dels cinc arxius i la tasca assignada passa per les dades i retorna la temperatura màxima per a cada ciutat. Per exemple, els resultats produïts de les dades anteriors es veuria així:
(Toronto, 20) (Whitby, 25) (Nova York, 22) (Roma, 33)
Suposem que els altres quatre *mapejadors* (que treballen en els altres quatre arxius restants) produeixen els següents resultats:

Arxiu 2:
(Toronto, 18) (Whitby, 27) (Nova York, 32) (Roma, 37)
Arxiu 3:
(Toronto, 32) (Whitby, 20) (Nova York, 33) (Roma, 38)
Arxiu 4:
(Toronto, 22) (Whitby, 19) (Nova York, 20) (Roma, 31)
Arxiu 5:
(Toronto, 31) (Whitby, 22) (Nova York, 19) (Roma, 30)

Els cinc arxius arriben a les tasques de reducció, que combinen els resultats d'entrada i sortida d'un únic valor per a cada ciutat, produint un resultat final fixat de la següent manera:

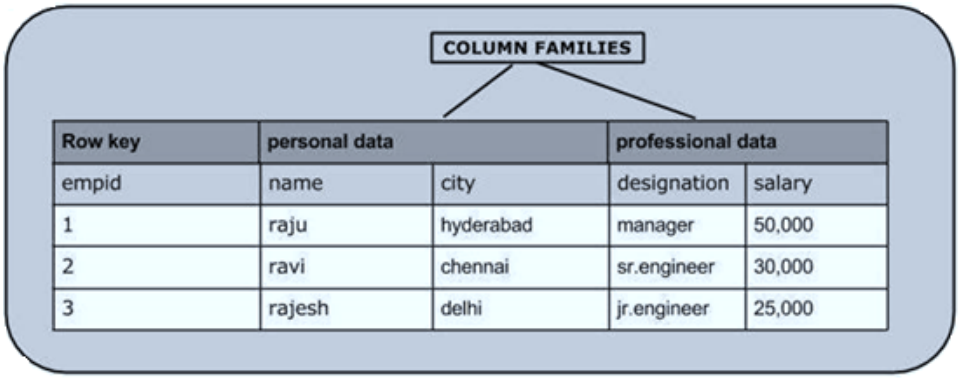
(Toronto, 32) (Whitby, 27) (Nova York, 33) (Roma, 38)

Com que no he trobat la manera d'explicar-ho més clarament en format redacció, he decidirt decantar-me per fer calcar l'exemple de forma esquemàtica i més visual (a la pàgina següent).


ARXIU 1
CLAU
0
Toronto
NY
20
22
0
VALOR
...
RESULTAT:
(Toronto, 20)
(NY, 22)
...
ARXIU 2
RESULTAT:
0
(Toronto, 18)
(NY, 32)
0
...
ARXIU 3
RESULTAT:
0
(Toronto, 32)
0
(NY, 33)
...
ARXIU 4
RESULTAT:
0
(Toronto, 22)
0
(NY, 20)
...
ARXIU 5
RESULTAT:
0
(Toronto, 31)
0
(NY, 19)
...
RESULTAT FINAL
=> (Toronto, 32) (NY, 33) ...

HBASE

HBase és un sistema de gestió de dades que forma part de HDFS. Funciona mitjançant taules amb files i columnes. Una columna representa un valor o una clau (vist a l'exemple de MapReduce). HBase permet la lectura i l'escriptura de les dades, creant taules de mida il·limitada. A diferencia de HDFS, HBase permet fer cerques ràpides.



Row key	personal data		professional data	
empid	name	city	designation	salary
1	raju	hyderabad	manager	50,000
2	ravi	chennai	sr.engineer	30,000
3	rajesh	delhi	jr.engineer	25,000

Exemple de taula de HBase

DISTIBUCIONS DE HADOOP

Hadoop és distribuït per diferents organitzacions, agrupades en tres conjunts:
El primer conjunt és de codi obert al 100%, i el forma la Fundació Apache. Però diverses empreses troben massa "infantil" i que no arriba al seu nivell professional. És així com sorgeix el segon conjunt, que són distribucions comercials, on s'afegeixen altres utilitats. Els més populars que el formen són Cloudera, Hortonworks i MapR.
A més a més, també és molt comú que les empreses utilitzin Hadoop en el núvol (pàg. 54), i aquest és el tercer conjunt. Els més populars i comuns que formen aquest conjunt són Amazon Web Services (AWS) i Windows Azure HDInsight mitjançant Microsoft. Quan utilitzes una distribució del núvol, pot utilitzar, per exemple, una distribució d'Amazon (AWS) per tal d'utilitzar una versió de Hadoop [ja sigui Apache Hadoop (de codi obert), com MapR (comercial, etc.)], però no totes les versions comercials estan disponibles a tots els núvols.

PER QUÈ UTILITZAR HADOOP?

- **Econòmic**→ utilitza escales enormes de dades (petabytes o més).
- **Ràpid**→ va a la mateixa velocitat que el Big Data gràcies al MapReduce.
- **Millor**→ capaç d'emmagatzemar i tractar tot tipus de Big Data (estructurat, no estructurat, etc.)

ORGANITZACIONS QUE UTILITZEN HADOOP

- Facebook, és l'usuari més gran que utilitza Hadoop.
- Yahoo, els seus ex-treballadors del qual van fundar Hortonworks.
- Amazon, un expert a utilitzar-lo per a la recomanació.
- eBay, semblant als anteriors.
- Aerolínies americanes, que recullen dades del comportament en el seu vol.

I més de centenars d'empreses, com ara New York Times, IBM, Junta de la Reserva Federal americana, etc., que utilitzen Hadoop amb la finalitat de prendre millors decisions.

ALTERNATIVES A HADOOP

És cert que Hadoop és l'eina més utilitzada, més popular i més completa a l'hora de tractar amb Big Data. Però això no vol dir que no hi hagi alternatives.

Alguns exemple són, tot i que a alguns d'ells estan sovint relacionats amb Hadoop d'alguna manera o altra: Tableau, Cloudera Impala, IBM Big SQL, HP Haven, Spark, Disco, Python, HPCC Systems, Pentaho, Pervasive Software, etc.

UNITATS DEL BYTE

Bytes (B)

Quilobytes (KB) = 2^{10} bytes

Megabytes (MB) = 2^{10} bytes

Gigabytes (GB) = 2^{10} bytes

Terabytes (TB) = 2^{10} bytes

Petabytes (PB) = 2^{10} bytes

Exabytes (EB) = 2^{10} bytes

Zettabytes (ZB) = 2^{10} bytes

PART PRÀCTICA

ÍNDEX

1. FER FRONT AL COMENÇAMENT

Moltes vegades una part teòrica requereix una part pràctica. En aquest cas no és essencial per entendre el concepte tema però sí que és interessant per entrar en aquest món i entendre, de primera mà, més enllà del concepte de Big Data.

D'entrada, ens trobem amb dos problemes indispensables:

1. Aconseguir unes dades.
2. Saber gestionar les dades i treballar-les.

Sí que és cert que hi ha diversos Open Data de diversos temes. Però cap d'ells em motivava. Jo volia unes dades relacionades amb la psicologia i el màrqueting.

Primer de tot vaig pensar en les dades d'un supermercat. Són les adients, les que més m'agraden i les més exemplars. Però a la vegada són les més difícils, per no dir impossible, d'aconseguir ja que són 100% comercials i hi ha un risc (que les empreses no assumiran, clarament) a que aquestes dades puguin acabar a la competència, que suposaria un avantatge per l'adversari important.

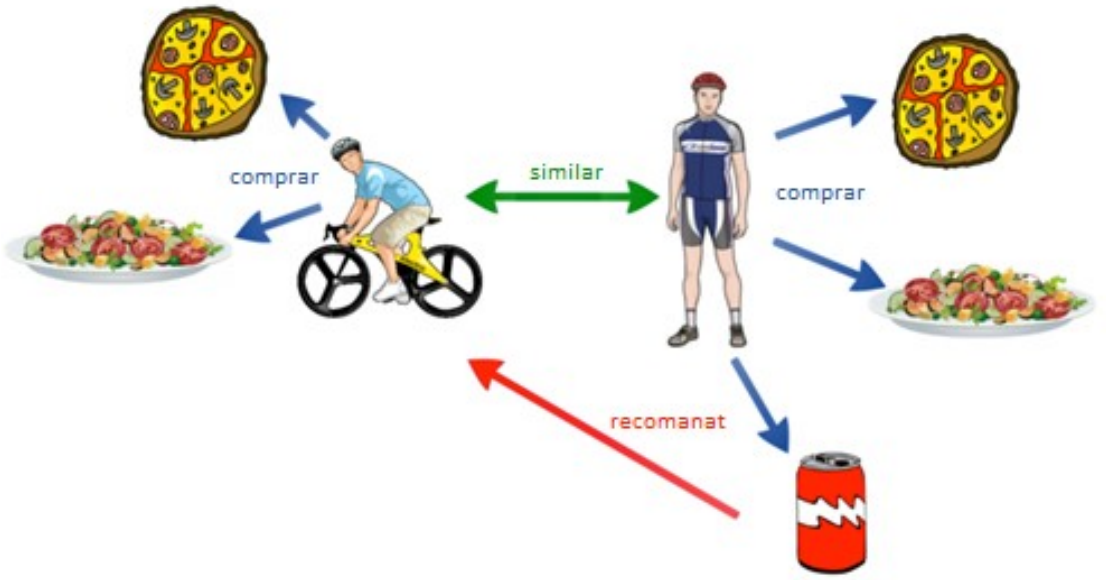
És per això que em vaig obrir a més opcions. I vaig trigar dos mesos fins a aconseguir unes dades que complissin els meus objectius.

2. DADES

Les dades que vaig aconseguir eren de la pàgina web IMDb. Les dades ens explicaven les pel·lícules que havia vist cada usuari i la puntuació (de l'1 al 5) que li havia posat.

Es tracta d'unes dades amb 943 usuaris, 1682 pel·lícules i 100000 valoracions.

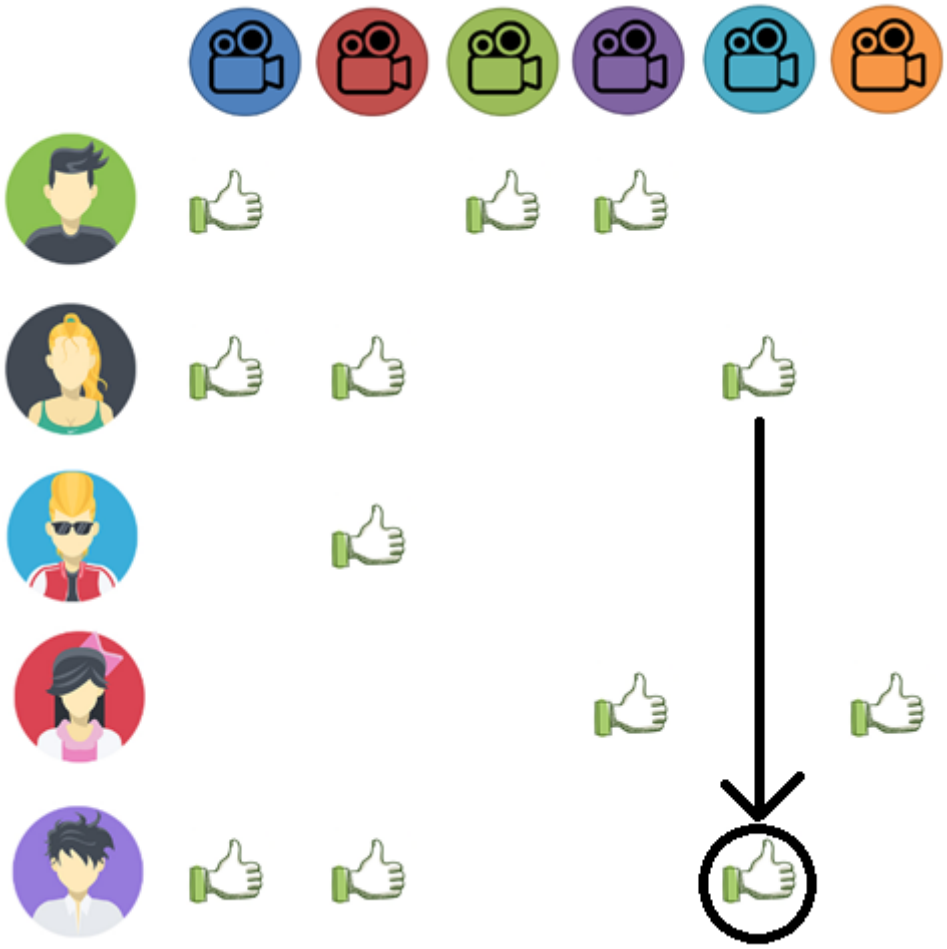
Amb aquestes dades es podia treballar la recomanació. Hi ha diversos models de recomanació, i el que vaig seguir va ser el User-based recommendation.



Cas exemplar en el que un usuari li agrada la pizza, l'amanida i una beguda. Un usuari semblant que també li agrada la pizza i l'amanida, li recomanem la beguda.

3. USER-BASED RECOMMENDATION

Aquest model pretén obtenir recomanacions a partir de les valoracions que hagin introduït altres usuaris. Posem per cas que cinc usuaris han donat les seves valoracions respecte de sis pel·lícules, que estan a cada fila del gràfic. Podem observar que la 2a usuària i el 5è usuari han coincidit en dues pel·lícules, però la usuària 2 ha vist la cinquena pel·lícula i el 5è usuari no. Aleshores, a partir de la informació de la usuària 2, es pot recomanar la pel·lícula 5 a l'usuari 5.



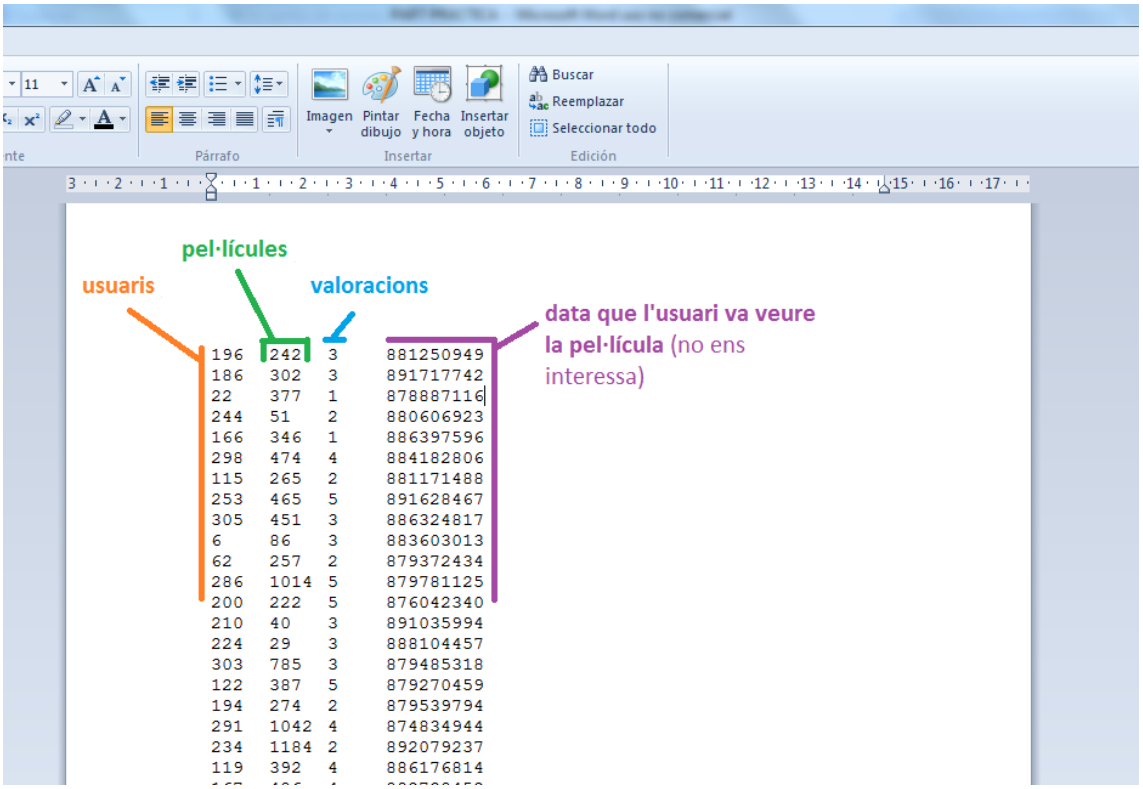
COM FER-HO: EINES

El món del Big Data es treballa amb programació, com ara Python, o programes experts, com ara Hadoop. Però des d'un principi vaig descartar aquestes dues opcions. Els programes experts, com bé diu la paraula, són pensats per a gent professional del tema, és a dir, es requereix un nivell molt avançat, també en informàtica, i jo aquest nivell no el tinc. Per altra banda, mai havia treballat amb programació així que vaig descartar-ho.

Vaig pensar en l'*excel*, una eina relativament fàcil d'utilitzar i dins de les meves capacitats i coneixements.

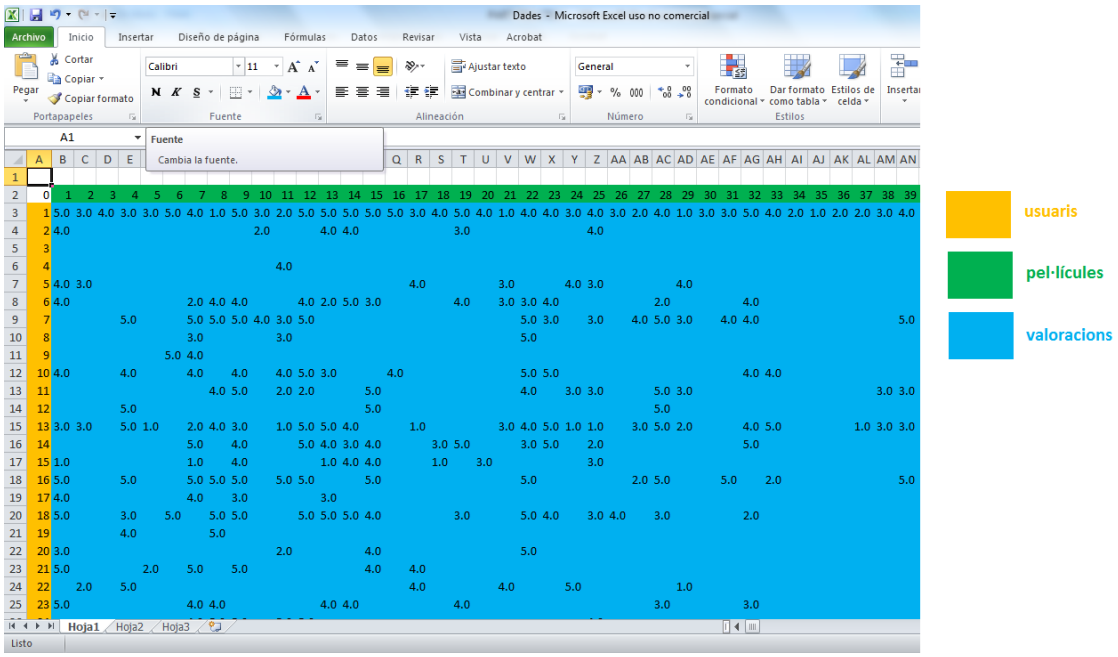
FORMAT DE LES DADES

En un primer moment tenia les dades en un format de text, que es mostrava de la següent manera.



En aquest format no hi havia cap ordre ni cap relació. Cada usuari apareixia a la columna per a cada pel·lícula que havia vist. Com que volia analitzar les dades mitjançant l'excel, les vaig passar a un format d'excel. Però un cop a l'excel, les vaig pivotar. És a dir, col·locar ordenades i formant una taula de tipus X on es relacionava clarament cada usuari quines pel·lícules havia vist i quina valoració els havia ficat.

Em va quedar de la següent manera.



En el moment d'analitzar les dades em vaig trobar amb un problema definitiu. Tant definitiu que em van impedir continuar per aquest camí.

PROBLEMES AMB L'EXCEL

L'excel és un programa molt correcte per a analitzar dades. Pots trobar mitjanes, medianes i modes, desviacions, correlacions, i moltíssimes coses més. Fins i tot era possible plantejar-me analitzar amb el model User-based recomendation. Però em deixava un fet tant important que era bàsic i imprescindible. No estem parlant d'unes simples dades, no estem parlant que si els companys de classe els agrada més el pernil dolç o salat, estem davant d'una grandària de dades significativa. Havia oblidat que estava fent Big Data.

Davant de la gran quantitat de dades, l'excel no és capaç d'automatitzar el que fa a petita escala per fer-ho a gran escala. És a dir, analitzaria pas a pas, manualment, els 943 usuaris, les 1682 pel·lícules i les 100000 valoracions.

I va ser d'aquesta manera que vaig decidir abandonar l'excel i endinsar-me, per primer cop, en el món de la programació.

4. PYHTON

Python segurament és un dels programes més estàndards a l'hora de fer programació i és per això que és el més comú i popular entre els que coneixen el tema. És per això que vaig decidir utilitzar Phyton i no cap altre. Això sí amb qualsevol altre programa hagués funcionat igual de bé.

Python, per definició, és un llenguatge de programació àmpliament usat per a propòsits generals. Va ser creat el 1991 per Guido van Rossum.

Comencem creant un exemplar de dades reduït per analitzar i després traspassarem la feina feta a les dades que reals i amb grans quantitats. Les següents dades són totalment inventades. A la columna de l'esquerra representen els diversos usuaris que han valorat les pel·lícules, situades a la primera fila. Al centre, repartides per tot l'espai, hi ha les valoracions corresponents a la fila de l'usuari que les ha fet i a la columna de la pel·lícula valorada.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
1		3	2		5	5		5		
2		2	3		3	2		3	3	3
3	1		4	2	3		2	5	4	4
4	1						1			5
5	1	4	3	4		2		4		3
6		3	1	2				1	1	1
7	2				4					2
8		3		1	2	1	4	4		

Les dades exemplars són aleatòries amb un 50% de possibilitats de no haver vist la pel·lícula i un 50% d'haver vist la pel·lícula, amb un 1/5 part de possibilitats per a cada valoració de la pel·lícula vista, valorades del 1 al 5.

Aquestes dades les introduïm al Python de la següent manera, on 'u1' significa l'Usuari 1 i 'p1' la pel·lícula 1, i així respectivament:

```
# Diccionari de les Dades Exemplars
DadesExemplars={'u1': {'p2': 3.0, 'p3': 2.0, 'p5': 5.0, 'p6': 5.0, 'p8': 5.0},
         'u2': {'p2': 2.0, 'p3': 3.0, 'p5': 3.0, 'p6': 2.0, 'p8': 3.0, 'p9': 3.0, 'p10': 3.0},
         'u3': {'p1': 1.0, 'p3': 4.0, 'p4':2.0, 'p5':3.0, 'p7':2.0, 'p8':5.0, 'p9':4.0, 'p10':4.0},
         'u4': {'p1': 1.0, 'p7': 1.0, 'p10':5.0},
         'u5': {'p1': 1.0, 'p2': 4.0, 'p3':3.0, 'p4':4.0, 'p6':2.0, 'p8':4.0, 'p10':3.0},
         'u6': {'p2': 3.0, 'p3': 1.0, 'p4': 2.0, 'p8': 1.0, 'p9': 1.0, 'p10':1.0},
         'u7': {'p1': 2.0, 'p5': 4.0, 'p10': 2.0},
         'u8': {'p2': 3.0, 'p4': 1.0, 'p5': 2.0, 'p6': 1.0, 'p7': 4.0, 'p8':4.0}}

from math import sqrt
```

TIPUS DE CRITERIS

Un cop tenim les dades exemplars, tenim dos criteris per mesurar la similaritat i dur a terme el model User-Based Recomendation.

Les dues vies ens porten a un resultat similar però amb mitjançant un raonament diferent.

DISTANCE-BASED SIMILARITY (SIM_DISTANCE)

Distance-Based Similarity (Similitud Basada en la Distància) es basa en la distància entre les valoracions de diferents pel·lícules que fan dos o més usuaris. Per fer-ho, primer hem de calcular la distància matemàtica entre les diverses valoracions entre dos usuaris.

Agafarem la valoració de l'Usuari 1 i la valoració de l'Usuari 2 d'una mateixa pel·lícula i les restarem. Un cop feta la resta, elevarem el resultat al quadrat. Això mateix ho farem amb totes les pel·lícules. Finalment sumarem tots els resultats obtinguts per a cada pel·lícula. Ens podem trobar tres casos de valoracions:

- Si tots ambdós usuaris han valorat la pel·lícula es fan les operacions tal com les he explicat.
- Si un usuari ha valorat la pel·lícula i l'altre no, s'elimina aquella pel·lícula, és a dir, es valora amb un 0.
- Si cap dels dos usuaris han valorat la pel·lícula, no existeix la pel·lícula, és a dir, es valora amb un 0.

Comparem l'usuari 1 amb l'usuari 2 de l'exemple i calcularem la distància entre ells. Per fer-ho seguim els càlculs següents:

$$Distància = (3-2)^2 + (2-3)^2 + (5-3)^2 + (5-2)^2 + (5-3)^2$$

$$Distància = 1^2 + (-1)^2 + 2^2 + 3^2 + 2^2$$

$$Distància = 1 + 1 + 4 + 9 + 4$$

$$Distància = 19$$

Ara sabem que com més gran sigui el nombre que obtinguem, més llunyans seran els usuaris entre ells, ja que com més gran sigui el la nombre obtingut més distància hi ha entre els usuaris. Però és poc visual i precís. Així que decidim normalitzar la distància fent-ne el recíproc i sumant-li una unitat al denominador. Al valor calculat li direm Similitud.

$$S = \frac{1}{1+19} = 0{,}05$$

D'aquesta manera, com més a prop estigui a 1, més semblants seran els usuaris, on 1 és la igualtat entre els usuaris i 0 és la oposita entre els usuaris. Veiem que en aquest cas està molt lluny de 1 i per tant l'Usuari 1 i l'Usuari 2 són llunyans.

Finalment, només ens falta passar tots aquests càlculs a Python. Seguint la seva sintaxi, la Distance-Based Similarity es defineix com a "sim_distance". Ha quedat de la següent manera:

```
# Càlcul de la distància entre Usuari 1 i Usuari 2 a partir de la Similitud Basada en la Distància
def sim_distance(prefs,person1,person2):
    # Get the list of shared_items
    si={}
    for item in prefs[person1]:
        ##print item
        if item in prefs[person2]:
            si[item]=1
        ##else: print "no trobat"
    # if they have no ratings in common, return 0
    if len(si)==0: return 0
    # Add up the squares of all the differences
    sum_of_squares=sum([pow(prefs[person1][item]-prefs[person2][item],2)
        for item in prefs[person1] if item in prefs[person2]])
    return 1/(1+sum_of_squares)
    # return sum_of_squares/2
```

SIM_PEARSON (DISTRIBUCIÓ)

Pearson Correlation Coefficient (Coeficient de Correlació de la Persona) es basa en la correlació de la valoració de diverses pel·lícules entre dos usuaris. Per fer-ho, l'Usuari 1 representarà l'eix X i l'Usuari 2 representarà l'eix Y. A més a més, repre-

sentarem les pel·lícules amb punts. La valoració que ha fet l'Usuari 1 de una determinada pel·lícula serà el coeficient X del punt, mentre que la valoració feta per l'Usuari 2 en la mateixa pel·lícula serà el coeficient Y del punt.

Agafant l'exemple, l'Usuari 1 i l'Usuari 2 formarien els següents punts:

$$P2 = (3,2)$$

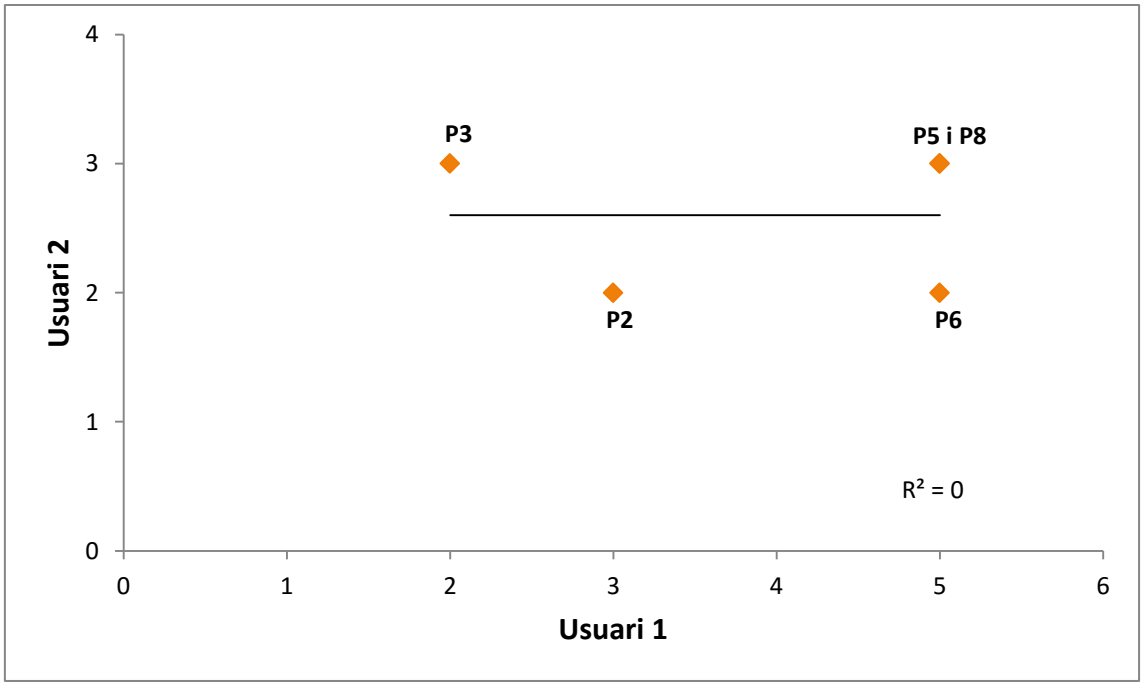
$$P3 = (2,3)$$

$$P5 = (5,3)$$

$$P6 = (5,2)$$

$$P8 = (5,3)$$

Els diferents punts que s'aniran formant en una gràfica, on es podrà veure la correlació que tenen. Per calcular la correlació, creem un gràfic a l'excel mitjançant un gràfic.



Veiem que la Correlació és igual a 0, per tant, que són usuaris llunyans. Quan la correlació és igual a -1, ens diu que els usuaris són oposats. Quan la correlació és igual a 1, els usuaris comparteixen una màxima correlació, és a dir, són màxima-

ment propers. Per tant, com més a prop de 1 més propers i com més a prop de -1 més llunyans.

Observem que en les dues maneres hem arribat a la mateixa conclusió respecte la relació Usuari1-Usuari2.

Finalment, només ens falta passar tots aquests càlculs a Python, de la mateixa manera que hem fet anteriorment. Queda de la següent forma:

```
# Càlcul de la correlació entre Usuari 1 i Usuari 2
def sim_pearson(prefs,person1,person2):
    # Get the list of mutually rated items
    si={}
    for item in prefs[person1]:
        if item in prefs[person2]: si[item]=1
    # Find the number of elements
    n=len(si)
    # if they are no ratings in common, return 0
    if n==0: return 0
    # Add up all the preferences
    sum1=sum([prefs[person1][it] for it in si])
    sum2=sum([prefs[person2][it] for it in si])
    # Sum up the squares
    sum1Sq=sum([pow(prefs[person1][it],2) for it in si])
    sum2Sq=sum([pow(prefs[person2][it],2) for it in si])
    # Sum up the products
    pSum=sum([prefs[person1][it]*prefs[person2][it] for it in si])
    # Calculate Pearson score
    num=pSum-(sum1*sum2/n)
    den=sqrt((sum1Sq-pow(sum1,2)/n)*(sum2Sq-pow(sum2,2)/n))
    if den==0: return 0
    r=num/den
    return r
```

TOP MATCHES

Un cop fets tots aquests càlculs passem al pas següent. Quins usuari s'assembla més a l'Usuari 1? Hem calculat la semblança entre l'Usuari 1 i l'Usuari 2. Ara hauríem de comparar l'Usuari 1 amb els altres 6 usuaris. Un cop calculades les 7 similituds, ordenaríem de més a menys proper. D'aquesta manera, podem saber l'usuari que s'assembla més a l'Usuari 1, o per exemple, els tres que s'hi assemblin més, etc.

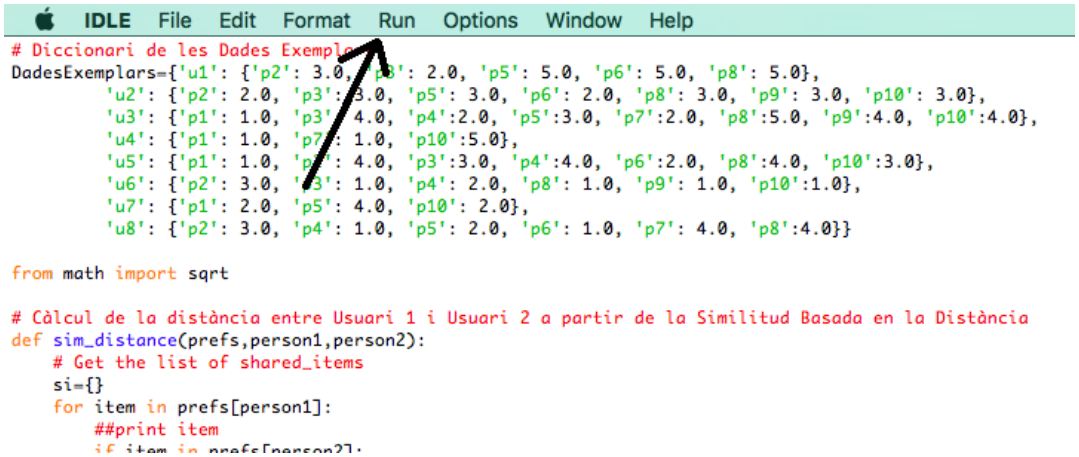
Per fer aquesta funció, podem utilitzar tant el "sim_distance" com el "sim_pearson".

Això es pot automatitzar amb el Pyhton amb la funció topMatches, que introduïda queda de la següent manera.

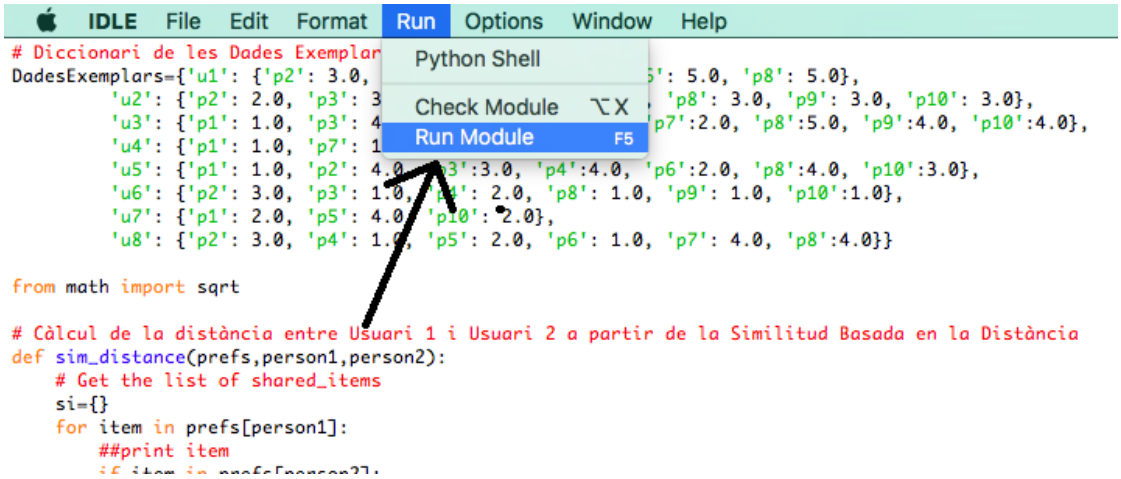
```
# Funció de similitud entre l'Usuari 1 i la resta.
def topMatches (prefs,person,n=5,similarity=sim_distance):
    scores=[(similarity(prefs,person,other),other)
        for other in prefs if other!=person]
    # Sort the list so the highest scores appear at the top
    scores.sort( )
    scores.reverse( )
    return scores[0:n]
```

APLICAR-HO

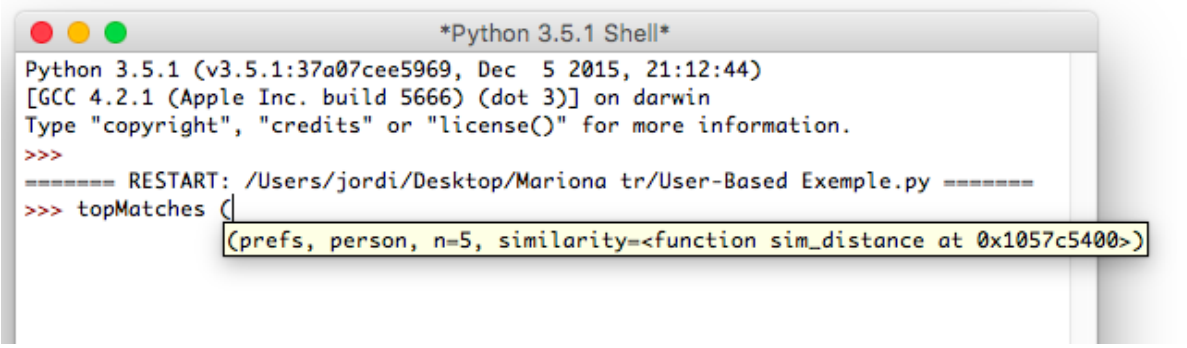
Ara cal que ho apliquem utilitzant les Dades Exemplars. Per començar, cliquem a Run.



Un cop a Run, cliquem Run Module (F5):



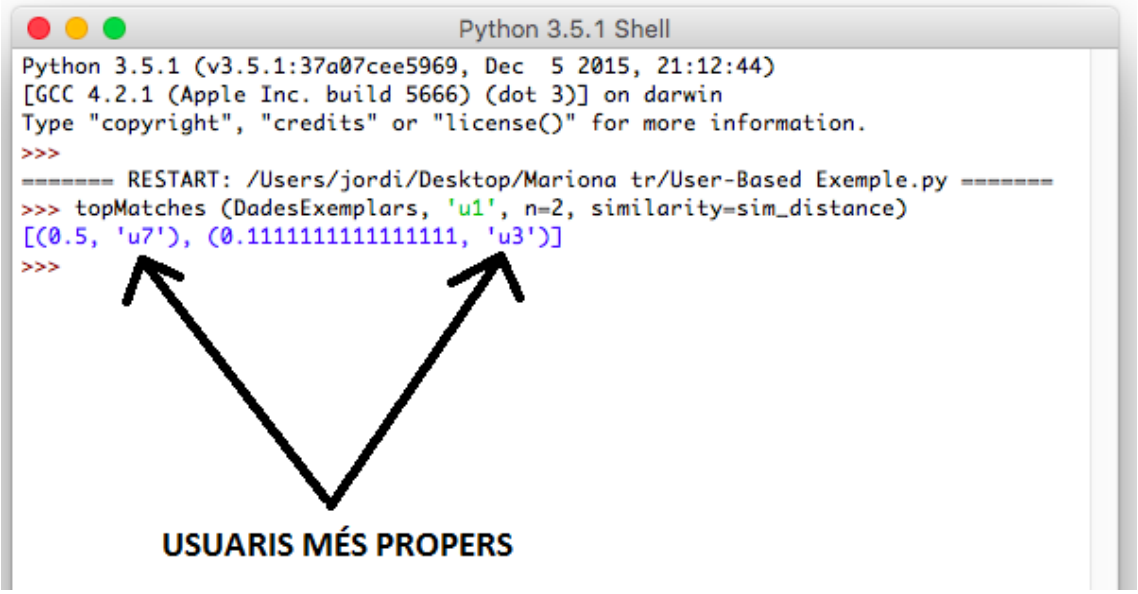
Llavors s'obra una nova finestra de Python, on podrem inserir el que volem fer. Per tant, escrivim topMatches ja que és la funció que volem utilitzar.



Un cop hem escrit topMatches, obrim un parèntesis i veiem que s'obre tot una sèrie de condicions que s'han d'anar omplint de la següent manera:

- Prefers: es refereix al nom que té la taula de dades, en aquest cas, DadesExemplars.
- Person: es refereix a l'usuari el qual volem comparar amb la resta d'usuaris, al llarg del treball l'anomenarem Usuari Base.
- n=5 significa el mateix que: "busca'm els 5 usuaris més propers". És un exemple, és a dir, enlloc de 5 es pot ficar el nombre que et vagi millor. Jo posaré n=2.
- Similarity= : ens demana el tipus de càlcul que volem utilitzar. En aquest cas podem triar entre sim_distance o sim_pearson. Triaré sim_distance ja que m'agrada més.

De manera que queda així:



Veiem que ens ha sortit que els usuaris 7 i 3 són els més propers a l'Usuari 1, respectivament. Al costat de l'usuari ens apareix el valor de la similitud.

[Ep! Compte; si hi ha un error mínim ortogràfic Python no ho entén i surt ERROR.]

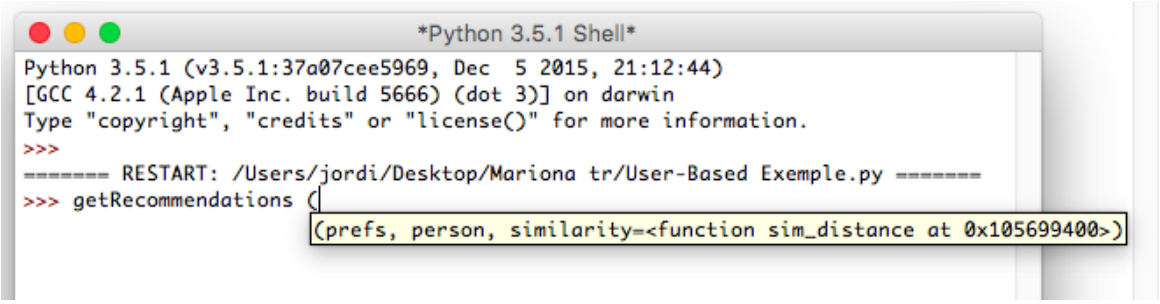
GET RECOMMENDATIONS

És interessant saber quins usuaris són més propers a l'Usuari Base però no hem d'oblidar el principal objectiu: recomanar pel·lícules.

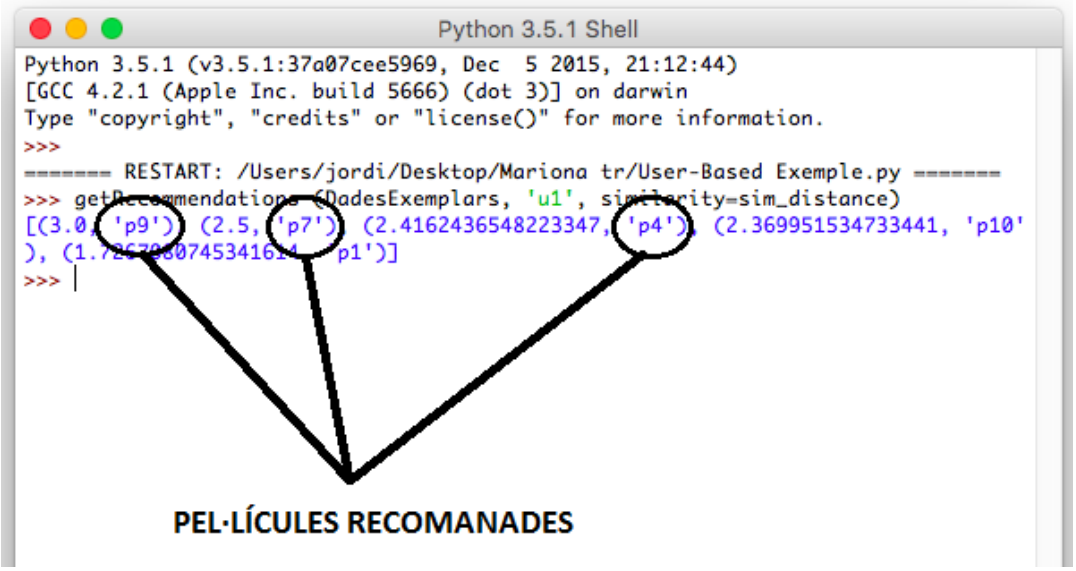
Per saber quina pel·lícula hem de recomanar, és a dir, quina pel·lícula que un usuari proper a l'Usuari Base hagi vist però que aquest últim no, utilitzem la funció getRecommendations. Aquest funció s'introdueix de la següent manera:

```
# Gets recommendations for a person by using a weighted average of every other user's rankings
def getRecommendations(prefs,person,similarity=sim_distance):
    totals={}
    simSums={}
    for other in prefs:
        # don't compare me to myself
        if other==person: continue
        sim=similarity(prefs,person,other)
        # ignore scores of zero or lower
        if sim<=0: continue
        for item in prefs[other]:
            # only score movies I haven't seen yet
            if item not in prefs[person] or prefs[person][item]==0:
                # Similarity * Score
                totals.setdefault(item,0)
                totals[item]+=prefs[other][item]*sim
                # Sum of similarities
                simSums.setdefault(item,0)
                simSums[item]+=sim
    # Create the normalized list
    rankings=[(total/simSums[item],item) for item,total in totals.items( )]
    # Return the sorted list
    rankings.sort( )
    rankings.reverse( )
    return rankings
```

Ara cal que ho apliquem utilitzant les Dades Inventades. Fem el mateix procediment que hem fet anteriorment amb topMatches (pàg 16). Cliquem Run, Run Module. Però aquest cop escrivim getRecommendations, obrim un parèntesis i ens demanarà el mateix que en el cas de topMatches.



I col·loquem els elements necessaris ficant l'Usuari 1 com a Usuari Base. Ens sortirà el següent:



Veiem que ens apareixen cinc pel·lícules per recomanar a l'Usuari 1: la 9, la 7, la 4, la 10 o la 1.

5. DADES REALS 1

Fins ara he fet tot el procés amb unes Dades Inventades, però era un pas per poder fer-ho amb les dades reals que em van facilitar d' IMDb.

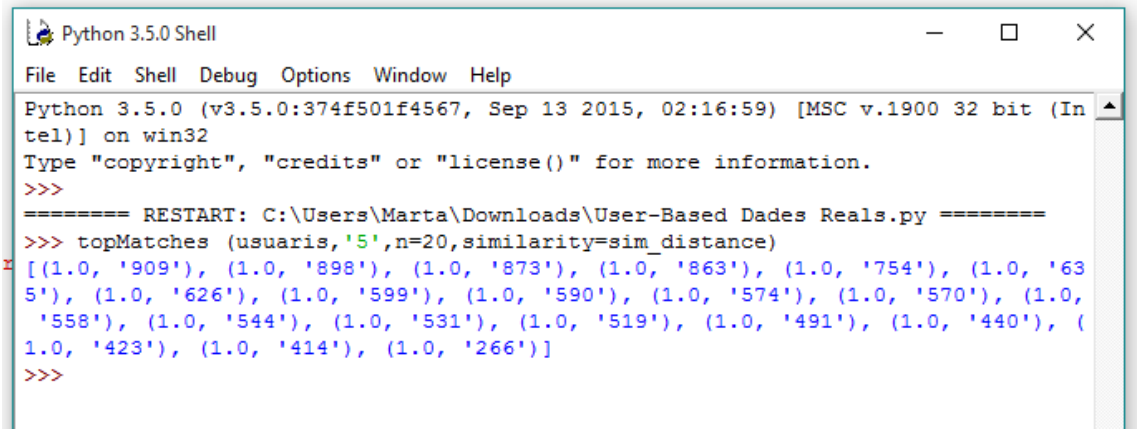
El primer pas és insertar les dades en el Python, cosa que no es pot fer de la mateixa manera que hem fet anteriorment (a mà i un per un) ja que ens trobem davant d'una enorme quantitat de dades. Com que jo no tinc un nivell suficientment elevat de programació, vaig demanar ajuda a l'Aleix Rodríguez per poder insertar-les. Per fer-ho va utilitzar el programa Sublime, després ho vam passar a Python i va quedar de la següent manera:

```
# Diccionari de les Dades Reals
usaris = {}
f = open('D:\Dropbox\MARIONAA\TR\ml-100k\ml-100k\u.data').readlines()
for i in f:
        x = i.split()[:3]
        if x[0] not in usaris:
                actual = 'p' + str(x[1])
                usaris[str(x[0])] = {}
                usaris[str(x[0])][actual] = int(x[2])
        else:
                actual = 'p' + str(x[1])
                usaris[str(x[0])][actual] = int(x[2])

from math import sqrt
```

La part de ('D:\Dropbox\MARIONAA\TR\ml-100k\ml-100k\u.data') pot variar segons on està ubicat el fitxer u.data, que és el document que conté les dades reals.

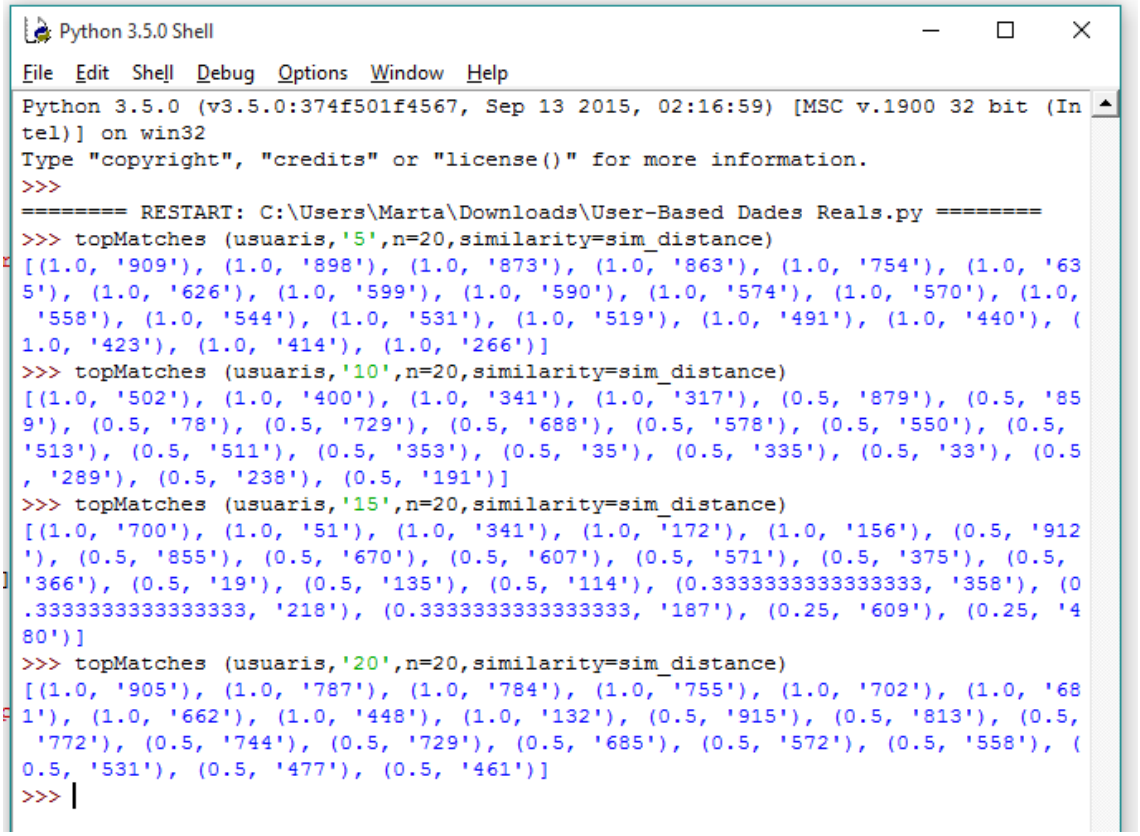
A partir d'aquí seguim el mateix procés que amb les dades exemplars. Utilitzem la funció topMathces. Aquest cop, per omplir el parèntesis, hi escriurem diferents coses.

- Prefers: En aquest cas escriurem *usuaris*.
- Person: Aquest cop faré que l'Usuari Base sigui l'Usuari 5.
- n=5: Posaré n=20 ja que ara tenim moltes més dades.
- Similarity= : Continuo utilitzant sim_distance ja que m'agrada més.



Ens surten diversos usuaris propers a l'Usuari 5. Però és llavors quan me n'adono d'alguna cosa que no va bé. El valor de similitud és sempre 1, cosa que és gairebé impossible que passi amb un usuari ja que voldria dir que de totes les pel·lícules que han vist els dos han puntuat amb la mateixa nota. I menys probable és encara que això mateix passi amb els 20 usuaris que em mostra.

Vaig sospitar que passava alguna cosa i vaig provar això mateix canviant l'Usuari Base.



Veiem que continua havent-hi masses usuaris amb valoració de similitud 1 però n'hi ha alguns que ja baixen de l'1.

Decideixo investigar què passa amb els usuaris que són el màxim de propers (és a dir, surt 1). Ho faig a través del full de càlcul on tinc totes les dades i comparo l'Usuari Base amb un dels usuaris que té propers. Ho faig amb uns quants i trec una conclusió.

En tots els casos l'Usuari Base i l'altre usuari només tenien una pel·lícula en comú i la valoració d'aquesta era la mateixa. De manera que quan calculem el valor de similitud ens surt 1 ja que només agafem aquelles pel·lícules que tenen en comú. Però això evidentment no és viable ja que no podem basar les nostres recomanacions en una sola pel·lícula. Per fer-ho fiable, els dos usuaris haurien de tenir un mínim de 10 pel·lícules en comú.

Per tant, ens està sortint una informació inservible. Per fer-la útil hauríem de demanar que elimines tots aquells usuaris de valor de similitud 1, però se m'escapa de les mans ja que no tinc capacitat de manipulació de programació.

Davant d'aquest problema, vaig decidir no continuar amb les Dades Reals de IMDb. Però continuar el treball amb unes dades exemplars sense cap mena de sentit no tenia gaire interès, així que vaig decidir crear les meves pròpies dades adequades però reals.

6. DADES REALS 2

Per crear unes dades reals i interessants vaig preguntar a un seguit d'amics i amigues quines pel·lícules d'una llista concreta havien vist i quina valoració els hi ficava. Tots els amics i amigues estaven dins d'una franja d'edat concreta (16-17 anys). Havien de valorar entre el 1 al 5 (de pitjor a millor).

Per fer-ho, creo un formulari que envio a persones concretes. El formulari va ser el següent.

goo.gl/1lnvel

Per poder-me referir a les dades utilitzant un llenguatge adequat vaig crear la següent llegenda:

PEL·LÍCULA	REFERENT
Furious 7	p1
Los juegos del hambre	p2
Jurassic World	p3
Minions	p4
Ocho apellidos vascos	p5
Cincuenta sombras de Grey	p6
Marte	p7
Inside Out	p8
Barcelona, nit d'hivern	p9
Ahora o nunca	p10
Interstellar	p11
Toy Story	p12
En busca de la felicidad	p13
Los chicos del coro	p14

USUARI	REFERENT
Mariona Gabarró	u1
Xavi Orti	u2
Paula Garcia	u3
Arnau Palos	u4
Laura Felip	u5
Andrea Quintana	u6
Berta Plandolit	u7
Guillem Ferrer	u8
Maria De la Fuente	u9
Marina Ramos	u10
Aleix Rodriguez	u11
Ignasi Andreu	u12

Un cop recollides les dades i passem del nom real a un referent convencional queda de la següent manera.

	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14
u1		5	4		5	4		4		4		2	5	5
u2	4	4	4	1	3		5	2			4	2		2
u3		4		3	4	4				4		3	5	4
u4					5		4				5			4
u5		5	4		3	4			4		1	4		5
u6					4	4		5			1	5	5	5
u7		4		4	4			5				4	4	5
u8	4	4	4	4	5	3		4				5		3
u9		3	4	4	4							4		5
u10	1	4		4	4			5				5		5
u11		4	4	4	4		4	5		4	4	5		
u12	4	3	4	2	3		4				4	3	4	2

Ara passem les dades al Python tal com vam fer amb les Dades Inventades. Queda de la següent manera:

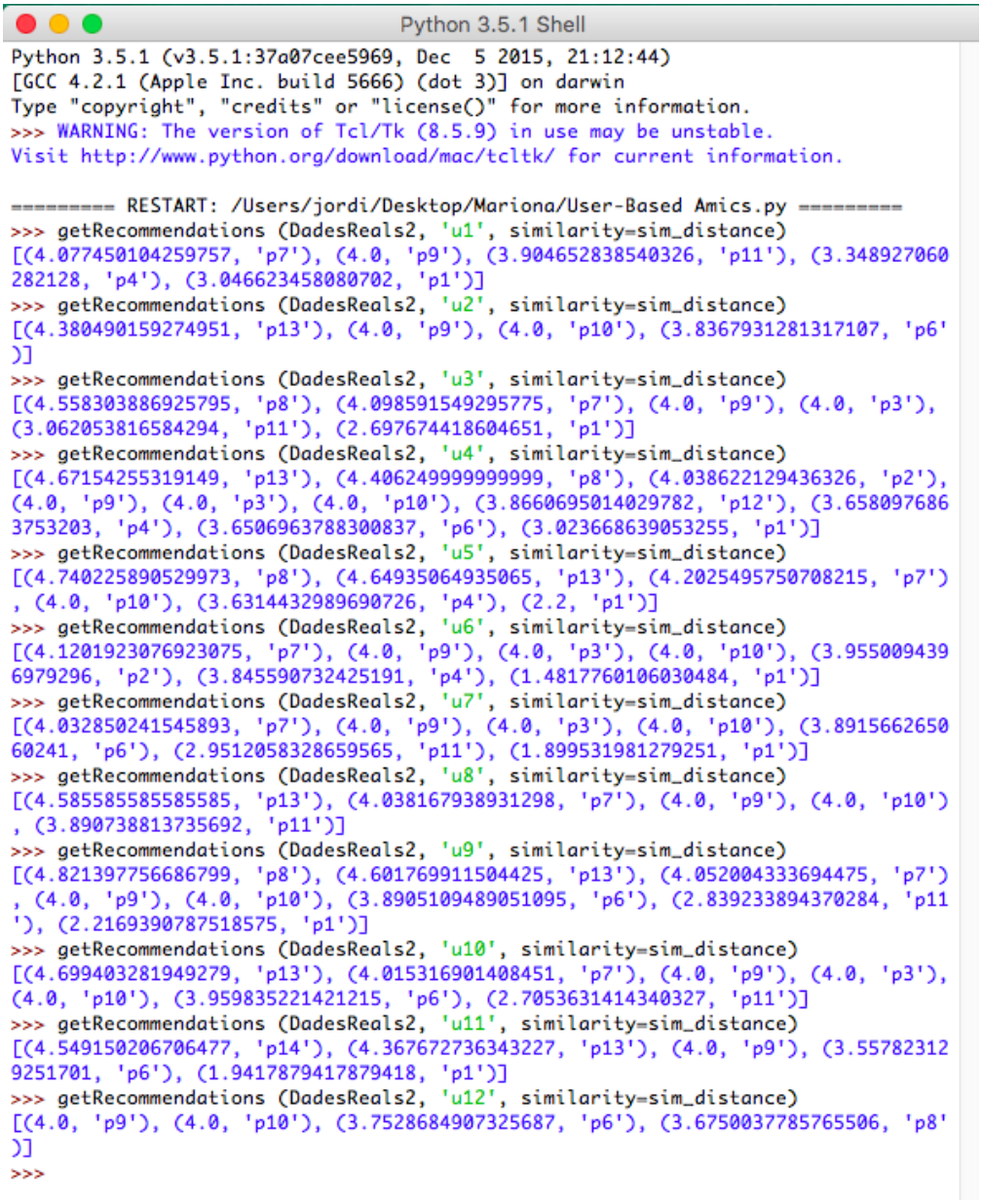
```
# Diccionari de les Dades Reals 2
DadesReals2={'u1': {'p2': 5.0, 'p3': 4.0, 'p5': 5.0, 'p6': 4.0, 'p8': 4.0, 'p10': 4.0, 'p12': 2.0, 'p13': 5.0, 'p14': 5.0},
         'u2': {'p1': 4.0, 'p2': 4.0, 'p3': 4.0, 'p4': 1.0, 'p5': 3.0, 'p7': 5.0, 'p8': 2.0, 'p11': 4.0, 'p12': 2.0, 'p14': 2.0},
         'u3': {'p2': 4.0, 'p4': 3.0, 'p5': 4.0, 'p6': 4.0, 'p10': 4.0, 'p12': 3.0, 'p13': 5.0, 'p14': 5.0},
         'u4': {'p5': 5.0, 'p7': 4.0, 'p11': 5.0, 'p14':4.0},
         'u5': {'p2': 5.0, 'p3': 4.0, 'p5': 3.0, 'p6': 4.0, 'p9': 4.0, 'p11': 1.0, 'p12': 4.0, 'p14': 5.0},
         'u6': {'p5': 4.0, 'p6': 4.0, 'p8': 5.0, 'p11': 1.0, 'p12': 5.0, 'p13': 5.0, 'p14': 5.0},
         'u7': {'p2': 4.0, 'p4': 4.0, 'p5': 4.0, 'p8': 5.0, 'p12': 4.0, 'p13': 4.0, 'p14': 5.0},
         'u8': {'p1': 4.0, 'p2': 4.0, 'p3': 4.0, 'p4': 4.0, 'p5': 5.0, 'p6': 3.0, 'p8': 4.0, 'p12': 5.0, 'p14': 3.0},
         'u9': {'p2': 3.0, 'p3': 4.0, 'p4': 4.0, 'p5': 4.0, 'p12': 4.0, 'p14': 5.0},
         'u10': {'p1': 1.0, 'p2': 4.0, 'p4': 4.0, 'p5': 4.0, 'p8': 5.0, 'p12': 5.0, 'p14': 5.0},
         'u11': {'p2': 4.0, 'p3': 4.0, 'p4': 4.0, 'p5': 4.0, 'p7': 4.0, 'p8': 5.0, 'p10': 4.0, 'p11': 4.0, 'p12': 5.0},
         'u12': {'p1': 4.0, 'p2': 3.0, 'p3': 4.0, 'p4': 2.0, 'p5': 3.0, 'p7': 4.0, 'p11': 4.0, 'p12': 3.0, 'p13': 4.0, 'p14': 2.0}}

from math import sqrt
```

7. RESULTATS

Per calcular els resultats utilitzaré la funció getRecommendations ja que em diu directament quines pel·lícules recomana a l'Usuari Base.

Com que ens trobem de tant poca quantitat de dades (14 pel·lícules, 12 usuaris i 94 valoracions) els resultats seran poc fiables però l'objectiu és mostrar com funciona.

Faré les recomanacions amb tots els usuaris com a Usuari Base incloent-me a mi mateixa. Els resultats són els següents:



8. RECOMANACIONS

Un cop tenim els resultats creem les recomanacions. Només recomanaré aquelles pel·lícules que es predigui una valoració de 4 o més punts (sobre 5), amb un màxim de 2 pel·lícules recomanades. El següent pas és seleccionar per a cada usuari les pel·lícules recomanades i passar del codi convencional al nom real per a poder fer ja les recomanacions i presentar-les als clients.

La pel·lícula de l'esquerra equival a la pel·lícula més recomanada i la de la dreta a la segona més recomanada.

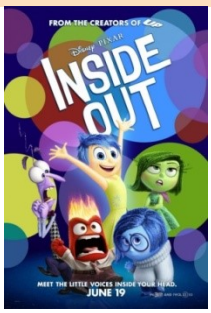
MARIONA GABARRÓ

	Marte	BCN, nit d'hivern
		

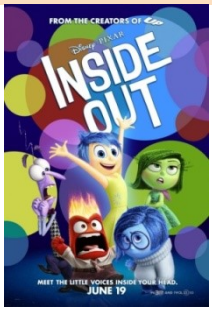
XAVI ORTI

	En busca de la felicidad	BCN, nit d'hivern
		

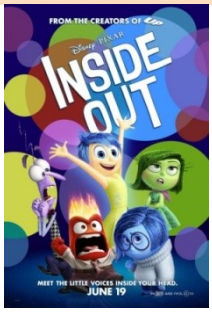
PAULA GARCIA

Inside Out	Marte
	

ARNAU PALOS

En busca de la felicidad	Inside Out
	

LAURA FELIP

Inside Out	En busca de la felicidad
	

ANDREA QUINTANA

Marte	BCN, nit d'hivern
	

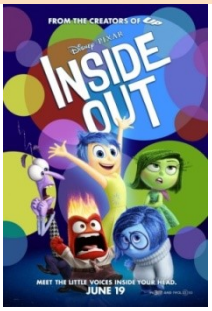
BERTA PLANDOLIT

Marte	BCN, nit d'hivern
	

GUILLEM FERRER

En busca de la felicidad	Marte
	

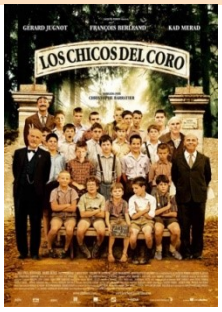
MARIA DE LA FUENTE

	Inside Out	En busca de la felicidad
		

MARINA RAMOS

	En busca de la felicidad	Ahora o nunca
		

ALEIX RODRIGUEZ

	Los chicos del coro	En busca de la felicidad
		

IGNASI ANDREU

	BCN, nit d'hivern	Ahora o nunca
		

9. MÉS ENLLÀ

Sovint les recomanacions tenen una segona intenció que no és la de recomanar. Aquesta segona intenció sempre amagada té un objectiu de màrqueting i vendes.

Molts cops ens recomanen ítems (en aquest cas pel·lícules) que no serien les primeres de la llista, és a dir, les quals es prediu una millor valoració, sinó que, dins d'un valor mínim de valoració, són ítems que els interessa vendre a priori. Normalment es fa amb pel·lícules o productes que s'estrenen i l'empresa recomanadora proposa abans que altres productes.

No he trobat cap espai on expliques de quina manera les empreses recomanadores varien les seves recomanacions segons els interessos del moment. Així que he decidit explicar dues maneres que jo veig possibles i no molt lluny d'on deu anar a la realitat.

CAS PARTICULAR 1

Posem per cas que som la productora que hem fet "Barcelona, nit d'hivern" i que, amb l'estrena de la pel·lícula, ens interessa que les empreses com IMDb la recomanin als seus usuaris per tal d'augmentar l'audiència.

Situem el valor mínim de recomanació a 3.5 només per aquesta pel·lícula i només pel fet de ser aquesta pel·lícula l'elegirem encara que no sigui a la segona de la llista amb més valoració.

Seguint aquest procés les recomanacions afectades per l'interès amb la pel·lícula "Barcelona, nit d'hivern" quedaria de la següent manera. (S'ha de tenir present que es tracta de molt poques dades i per tant és un exemple pobre però suficient).

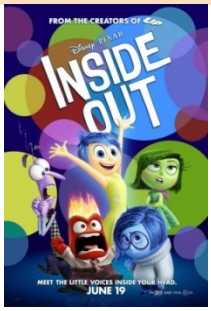
MARIONA GABARRÓ

	Marte	Barcelona, nit d'hivern
		

XAVI ORTI

	En busca de la felicidad	Barcelona, nit d'hivern
		

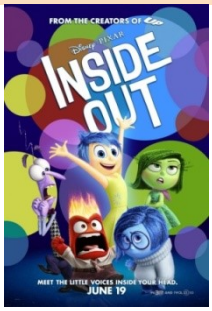
PAULA GARCIA

	Inside Out	Barcelona, nit d'hivern
		

ARNAU PALOS

	En busca de la felicidad	Barcelona, nit d'hivern
		

LAURA FELIP

	Inside Out	En busca de la felicidad
		

ANDREA QUINTANA

	Marte	Barcelona, nit d'hivern
		

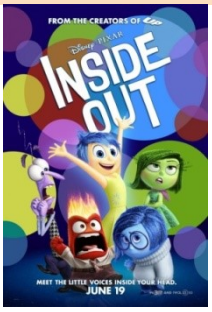
BERTA PLANDOLIT

	Marte	Barcelona, nit d'hivern
		

GUILLEM FERRER

	En busca de la felicidad	Barcelona, nit d'hivern
		

MARIA DE LA FUENTE

	Inside Out	Barcelona, nit d'hivern
		

MARINA RAMOS

	En busca de la felicidad	Barcelona, nit d'hivern
		

ALEIX RODRIGUEZ

	Los chicos del coro	Barcelona, nit d'hivern
		

IGNASI ANDREU

	Barcelona, nit d'hivern	Ahora o nunca
		

Veiem que d'aquesta manera la recomanació de la pel·lícula "Barcelona, nit d'hivern" augmenta considerablement. En aquest cas apareix en tots els usuaris menys en el cas de la Laura ja que és ja l'ha vist i per tant no la hi podem recomanar.

Per tant, és un mètode eficaç ja que, sempre assegurant que a la persona li agradarà, recomanem aquella pel·lícula que ens interessa recomanar per motius diversos (normalment econòmics).

CAS PARTICULAR 2

Però hi ha altres possibles maneres de fer-ho. Podem sumar-hi punts amb relació al que paga la producció per la situació de la pel·lícula.

Posem per cas que per cada 10 euros (evidentment no parlem de xifres reals ja que segurament estaríem parlant de milions d'euros) la valoració de recomanació de la pel·lícula puja 1 punt i que, proporcionalment, si la productora paga 7 euros la valoració pujarà 0,7.

La pel·lícula "Barcelona, nit d'hivern" paga 10 euros a l'empresa recomanadora i la pel·lícula "Marte" en paga 5. Per tant, la valoració de "Barcelona, nit d'hivern" pujarà 1 punt per a cada usuari i la de "Marte" en pujarà 0,5. Totes aquests valors es pujaran respecte els resultats obtinguts a la pàg 24.

Seguint aquest procés les recomanacions queden de la següent manera.

MARIONA GABARRÓ

Barcelona, nit d'hivern	Marte
	

XAVI ORTI

Barcelona, nit d'hivern	En busca de la felicidad
	

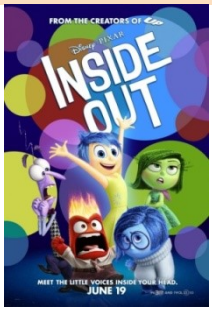
PAULA GARCIA

Barcelona, nit d'hivern	Marte
	

ARNAU PALOS

En busca de la felicidad	Barcelona, nit d'hivern
	

LAURA FELIP

Inside Out	Barcelona, nit d'hivern
	

ANDREA QUINTANA

Barcelona, nit d'hivern	Marte
	

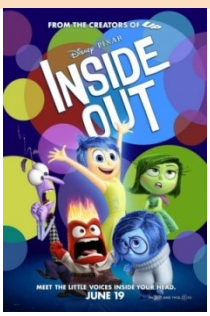
BERTA PLANDOLIT

	Barcelona, nit d'hivern	Marte
		

GUILLEM FERRER

	Barcelona, nit d'hivern	En busca de la felicidad
		

MARIA DE LA FUENTE

	Barcelona, nit d'hivern	Inside Out
		

MARINA RAMOS

	Barcelona, nit d'hivern	En busca de la felicidad
		

ALEIX RODRIGUEZ

	Los chicos del coro	Barcelona, nit d'hivern
		

IGNASI ANDREU

	Barcelona, nit d'hivern	Ahora o nunca
		

En aquest cas “Barcelona, nit d’hivern” tenia molta preferència i “Marte” en tenia poca. Veiem com el nombre de recomanacions de la pel·lícula “Barcelona, nit d’hivern” augmenta considerablement, de la mateixa manera que la anterior, ja que fer la suma d’1 punt és molt. En canvi, el nombre de recomanacions de “Marte” no augmenta. El fet que no augmenti és a causa a que la pel·lícula “Barcelona, nit d’hivern” la supera en molts casos i li treu la posició que tindria “Marte”, deixant-lo fora de la llista. Però en un cas més àmpli, amb una llista més llarga de recomanacions, “Marte” augmentaria el nombre de vegades que es recomana però en una posició gairebé sempre més baixa que “Barcelona, nit d’hivern”. De manera que com més punts sumes, proporcionalment, més cops surt recomanada.

En els dos casos hem vist una clara modificació de les recomanacions, tot i que sempre respectant les altres pel·lícules.

10. FI DE L'EXPERIMENT

I fins aquí la meva part pràctica.

M'agrada més anomenar-la experiment o tastet ja que ha sigut una introducció des d'un nivell zero al món del Big Data i de les dades en general. Com a experiment valoro el resultat i el procés molt positivament ja que em trobava davant d'un món totalment nou i, no només he sabut afrontar-lo i encarar-lo, sinó que a més a més he pogut moure'm amb comoditat (limitada, és clar) per poder dirigir-me cap allà on jo volia que anés el meu treball.

Més enllà dels meus limitats coneixements, provant, investigant i treballant he aconseguit crear recomanacions reals i experimentals tal com les creen les grans empreses d'avui en dia. També he aconseguit conèixer com funciona el Big Data mentre s'analitza i darrera del que veiem i se'ns mostra com a clients o consumidors.

CONCLUSIONS

Al llarg del treball he anat fent conclusions més concretes del tema que es tractava en cada moment. N'he fet un recull de les més significatives:

"Big Data no és una moda, sinó que un fet, un procés i un progrés que ens porta a una nova societat i a una nova manera de viure."

"Els desavantatges que comporta el Big Data no són desavantatges en si, com ara els afectes secundaris de les medicines, que són inevitables. Aquests desavantatges són més aviat dificultats que es troben per tal de poder arribar als avantatges."

"La clau està en convertir les dades en informació, la informació en coneixement i, finalment, el coneixement en una estratègia per a l'empresa."

"El Data Center és una de les maneres de ser conscient d'on vivim, i fins a quin extrem arriba el Big Data."

"Tan sols el 5% de tota la informació que disposem és analitzada."

Ara és el moment de fer una conclusió general de Big Data i una valoració del treball a nivell personal.

Big Data realment ha canviat la nostra societat, la nostra vida quotidiana, el món de les empreses, el tipus de seguretat, i moltes altres coses més. Tothom del nostre entorn no coneix el concepte Big Data però en canvi sí que coneix els seus efectes i els canvis que ha provocat, com ara que quan obri Youtube em recomani uns certs vídeos. Però no n'hi ha prou amb saber els seus afectes, cal ser conscients del món on vivim, què podem fer i que no, què ens poden fer i què no.

La situació òptima seria ser conscient del tema i a més a més controlar-lo, dominar-lo. Un cop domines Big Data i tot el que comporta entens com funcionen moltes coses, i el perquè i el com de moltes empreses, etc. Jo, després de vuit mesos d'investigació, començo a familiaritzar-me amb el que seria, segurament, la punta d'un iceberg.

Sempre hi ha l'altre punt de vista. La innocència fa la felicitat. En el moment que segueixes aquesta via, en converteixes en un peó de la societat i per tant, aquest tema deixa de ser interessant. Què hi ha darrera de cada pàgina web comercial, què hi ha darrera de cada recomanació, i de cada aplicació de Play Store, per exemple. Tot això són preguntes que es responen amb el Big Data. Per tant, saber sobre Big Data és un coneixement que t'obre els ulls a aquesta nova societat, molts cops amagada, i de la mateixa manera és contrària a la filosofia de la innocència.

Respecte al desenvolupament del treball, a més a més dels àmbits els qual jo volia enfocar el treball, em vaig trobar davant de l'àmbit tecnològic.. Era inevitable. És més, no tindria sentit investigar sobre Big Data sense trobar-te amb un punt de tecnologia. Em de ser conscients, i jo al principi no n'era, que Big Data és la conseqüència directe de l'era de la tecnologia. Poc a poc em vaig anar familiaritzant amb aquest àmbit fins a sentir-m'hi còmode. Per altra banda, també em vaig haver d'endinsar al món de la programació, que en aquest cas començava des de zero. Amb paciència, ganes d'aprendre i ajuda me'n vaig poder sortir i fins i tot entendre gran part de les accions que feia al programar.

Personalment opino que he fet un bon treball. El tema ajudava una mica, ja que tot el que investigava era nou per a mi i per tant també interessant. Això e motivava a fer-ho millor i amb més ganes. Però per altra banda, també vaig veure l'oportunitat de fer un bon treball. Tenia temps, projecte ampli per tirar endavant, motivació, capacitats, condicions i sobretot ganes.

Com és normal, vaig tenir alguns problemes durant l'elaboració del treball. Pel que fa a la part teòrica, més que un problema vaig tenir una dificultat afegida. Al ser un tema extens i treballat gairebé només a Estats Units em vaig trobar amb moltíssimes pàgines web, en anglès, que parlaven del tema i cada un ho explicava des del seu punt de vista. Per això vaig haver de contrastar diverses fonts contínuament, i crear el meu propi punt de vista.

Pel que fa a la part pràctica, vaig estar gairebé dos mesos buscant i buscant dades que em permetessin fer el que em proposava. Però finalment, quan ja estava a punt d'abandonar i plantejar-me un canvi de part pràctica, vaig trobar les dades que necessitava. Un cop les vaig tenir se'm va complicar encara més, ja que jo mai havia

treballat amb Big Data. Big Data es treballa amb eines com Hadoop, però si no ets un professional pots treballar amb eines com Python. El cas és que jo era encara menys experta que els inexperts i en un principi volia fer-ho amb Excel, l'eina que jo dominava. Però no va ser possible i em vaig trobar amb la dificultat el Python. Un cop ja ho tenia gairebé tot enllestit, em vaig trobar amb un últim problema que m'impedia treure resultats de les dades que havia obtingut. Però tots els problemes els vaig solucionar sense aturar el projecte així que el s'han convertit en obstacles superats.

Tot i els vuit mesos d'investigació, en cap moment se m'ha fet llarg ni pesat, cosa que considero una bona senyal. Això envolta el treball amb una energia positiva que fa que faci, moltes vegades, més del que se m'ha demanat.

AGRAÏMENTS

Tot treball es forma gràcies a un conjunt de persones que hi ha dedicat el seu temps, el seu esforç i la seva confiança. I aquest no n'és cap excepció.

En primer lloc, agraeixo a Jordi Rincón, el meu tutor de treball i professor d'Economia de l'empresa, per donar-me la confiança i dirigir-me el treball. En el mateix nivell, agraeixo al meu pare, Jordi Gabarró, pel seu temps i les seves opinions.

Agraeixo també a la meva mare, Marta Rovira, per passar-me articles constantment i per mostrar el seu interès i la seva ajuda en part de la correcció. Respecte pel que fa a la correcció final i decisiva, agreixo a la meva professora de llengua catalana, Eva Sánchez Flamerich, la seva dedicació.

Per últim, agraeixo a la meva germana, Laia Gabarró, la seva ajuda, el seu temps i el seu punt de vista que sovint va ser decisiu en el treball.

Respecte a la part pràctica del treball, agraeixo a Jaume Serra, el meu cotutor i professor de Matemàtiques, per la seva dedicació i ajuda necessària en el treball. Sense aquesta ajuda difícilment hauria estat possible una part pràctica.

També agraïments a Aleix Rodríguez per aportar sempre la seva part informàtica i la seva confiança en el treball.

Infinits agraïments a Oriol Pujol, professor de la Universitat de Barcelona del departament de Matemàtica Aplicada i Anàlisi, per la seva imprescindible col·laboració en la part pràctica.

Per últim, mencionar l'Institut Vilatzara de Vilassar de Mar, on s'ha fet possible i correspon aquest treball de recerca.

BIG DATA,

UN MÓN ENORMEMENT NOU

MARIONA GABARRÓ

- BIBLIOGRAFIA -

Aquest treball de recerca s'ha format majoritàriament a través d'investigació via internet. Gran part de la informació exposada s'ha format a partir de recollir molta i molta informació, buscant i buscant, contraposant diferents punts de vista, etcètera.

Per tant, la bibliografia està composa de moltíssimes pàgines web i algun llibre, que són els següents.

ÍNDEX

A més a més també hi ha la bibliografia en format digital per fer-ho més fàcil i més lleuger a l'hora de consultar-la. L'enllaç és el següent.

goo.gl/TY8sEr

PART TEÒRICA

INTERNET

- http://www.revistadyna.com/noticias-de-ingenieria/big-data-que-es-y-de-donde-viene
- https://ca.wikipedia.org/wiki/Xarxa_inform%C3%A0tica
- http://definicion.de/internet/
- http://www.cad.com.mx/que_es_internet.htm
- https://ca.wikipedia.org/wiki/Internet

WWW

- https://ca.wikipedia.org/wiki/World_Wide_Web
- http://www.fotonostra.com/digital/paginasweb.htm
- http://www.educoas.org/portal/bdigital/contenido/valzacchi/ValzacchiCapitulo-2New.pdf

LLEI DE MOORE

- http://www.ara.cat/premium/opinio/Moore-turisme-xines-balanca-comercial_0_522547756.html
- https://ca.wikipedia.org/wiki/Llei_de_Moore
- http://www.muyinteresante.es/tecnologia/preguntas-respuestas/ique-es-la-ley-de-moore
- http://curiosidades.batanga.com/4933/que-es-la-ley-de-moore-y-para-que-sirve

UNIVERS DIGITAL (IOT)

- http://www.ara.cat/premium/opinio/Moore-turisme-xines-balanca-comercial_0_522547756.html
- http://www.computerworld.es/tendencias/el-universo-digital-se-expande-acelerado-por-el-crecimiento-de-los-datos
- http://mexico.emc.com/leadership/digital-universe/index.htm
- https://ca.wikipedia.org/wiki/Internet_de_les_coses
- http://www.muycomputerpro.com/2012/12/12/estudio-idc-big-data
- https://en.wikipedia.org/wiki/International_Data_Corporation
- https://www.youtube.com/watch?v=FkClKZODdgk
- http://www.docpath.com/es/art-big-data-document-technology-software.aspx

- https://infocus.emc.com/william_schmarzo/5-ways-the-internet-of-things-drives-new-opportunities/
- http://mexico.emc.com/infographics/digital-universe-2014.htm
- http://mexico.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf
- http://www.computerworld.es/sociedad-de-la-informacion/tres-motivos-para-estar-preocupado-por-el-internet-de-las-cosas
- http://www.computerworld.es/negocio/ibm-invierte-3000-millones-de-dolares-en-una-nueva-unidad-de-negocio-de-internet-de-las-cosas
- https://www.idc.com/about/about.jsp
- http://spain.emc.com/corporate/emc-at-a-glance/index.htm
- https://en.wikipedia.org/wiki/EMC_Corporation

SEGURETAT I PRIVACITAT

- http://www.computerworld.es/sociedad-de-la-informacion/tres-motivos-para-estar-preocupado-por-el-internet-de-las-cosas
- http://www.computerworld.es/tendencias/el-universo-digital-se-expande-acelerado-por-el-crecimiento-de-los-datos
- http://mexico.emc.com/leadership/digital-universe/index.htm
- http://mexico.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf
- http://mexico.emc.com/infographics/digital-universe-2014.htm
- http://www.muycomputerpro.com/2012/12/12/estudio-idc-big-data
- https://www.youtube.com/watch?v=FkClKZODdgk
- http://www.docpath.com/es/art-big-data-document-technology-software.aspx
- http://www.computerworld.es/tendencias/la-seguridad-el-reto-de-la-tecnologia-emergente-del-internet-de-las-cosas
- http://www.ericrettberg.com/datacultj1/reading-response-blog-posts/google-not-the-next-evil-villain/
- http://www.slideshare.net/joyanes/ioi-bi-tema6-bigdata
- http://www-01.ibm.com/software/data/bigdata/
- https://www.youtube.com/watch?v=3dVz9hMpbOM
- http://definanzas.com/paises-emergentes-lista-completa/
- https://es.wikipedia.org/wiki/Mercados_emergentes
- https://cloudsecurityalliance.org/about/
- https://en.wikipedia.org/wiki/Cloud_Security_Alliance
- http://searchcloudsecurity.techtarget.com/definition/Cloud-Security-Alliance-CSA
-

COOKIES

- https://es.wikipedia.org/wiki/Cookie_(inform%C3%A1tica)
- http://www.allaboutcookies.org/es/galletas/
- http://www.pcactual.com/articulo/actualidad/noticias/13535/que_son_las_cookies_por_que_aparecen_tantos_avisos_las_web.html
- http://faqoff.es/que-son-las-cookies/
- http://windows.microsoft.com/es-es/windows/cookies-faq#1TC=windows-7
- http://www.meetic.es/misc/cookie.php
- http://comohacerpara.com/el-curioso-origen-de-algunos-terminos-informaticos_7337n.html
- http://www.agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/common/Guias/Guia_Cookies.pdf
- http://jesusredondo.es/%C2%BFque-son-las-cookies-de-terceros

PRESENTACIÓ DEL BIG DATA

- http://www.eldiario.es/turing/Big-data_0_161334397.html
- https://www.linkedin.com/pulse/brief-history-big-data-everyone-should-read-bernard-marr
- http://elprofejose.com/2015/03/04/una-breve-historia-sobre-big-data-que-todo-el-mundo-deberia-leer/
- http://www.winshuttle.es/big-data-historia-cronologica/
- http://www-05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf
- https://es.wikipedia.org/wiki/Poblaci%C3%B3n_mundial

Llibre: Big Data: A Revolution that Will Transform how We Live

XIFRES

- http://www.nae.es/ca/big-data-y-analytics-un-tandem-imparable/
- http://opendata.bcn.cat/opendata/ca/what-is-open-data

LES 3V

- http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
- http://elprofejose.com/2015/03/04/una-breve-historia-sobre-big-data-que-todo-el-mundo-deberia-leer/

- http://www.winshuttle.es/big-data-historia-cronologica/
- http://www-05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf
- http://www.sap.com/bin/sapcom/es_co/downloadasset.2014-10-oct-28-06.big-data-3-v-s-what-do-they-mean-for-financial-planners-pdf.bypassReg.html
- https://news.microsoft.com/es-xl/big-data-volumen-variabilidad-y-velocidad/
- http://www.lantares.com/blog/las-tres-v-de-de-big-data-aplicadas-al-marketing
- http://www.brandchats.com/7-tipos-de-datos-que-comprende-el-big-data/
- http://www.dataprix.com/blog-it/big-data/big-data-gestion-datos-no-estructurados

DATA CENTER

- https://en.wikipedia.org/wiki/Data_center
- http://www.eldiario.es/turing/Big-data_0_161334397.html
- http://elprofejose.com/2015/03/04/una-breve-historia-sobre-big-data-que-todo-el-mundo-deberia-leer/
- http://www.winshuttle.es/big-data-historia-cronologica/
- http://es.slideshare.net/redondoj/big-data-desarrollando-soluciones-efectivas
- http://www.datacenterknowledge.com/archives/2013/09/05/big-data-what-it-means-for-data-center-infrastructure/
- http://www.networkcomputing.com/data-centers/big-data-analytics-and-why-datacenter-infrastructure-matters/a/d-id/1234647?
- http://www.datacenterknowledge.com/archives/2012/05/15/google-data-center-faq/
- http://www.google.com/about/datacenters/
- http://www.acens.com/blog/que-es-un-data-center.html
- http://www.acens.com/blog/lo-que-las-empresas-buscan-en-un-data-center.html
- http://www.businessnewsdaily.com/4982-cloud-vs-data-center.html
- https://ca.wikipedia.org/wiki/Centre_de_c%C3%A0lcul
- http://www.cyberciti.biz/faq/data-center-standard-overview/
- http://www.datacenters.com/news/featured/redundancy-n1-vs-2n/

MÀRQUETING

- http://www.ciberconta.unizar.es/leccion/marketing/100.HTM
- http://www.elblogsalmon.com/marketing-y-publicidad/que-es-el-marketing

- http://www.marketing-free.com/marketing/definicion-marketing.html
- http://www.significados.com/marketing/
- https://www.youtube.com/watch?v=dE4UdsxZnVE

AVANTATGES

- https://www.sas.com/fr_fr/news/sascom/2014q3/Big-data-davenport.html
- http://iveybusinessjournal.com/publication/why-big-data-is-the-new-competitive-advantage/
- http://www.cio.com/article/2392067/data-management/big-data-analytics-a-big-benefit-for-marketing-departments.html
- https://www.marketingtechblog.com/benefits-of-big-data/
- http://www.lantares.com/blog/big-data-ventajas-de-la-revolucion-de-los-datos-masivos
- http://www.siliconweek.es/workspace/las-cinco-ventajas-competitivas-que-aporta-el-big-data-49286
- http://onerp.es/beneficios-big-data-empresa/
- http://www.compromisorse.com/rse/2014/06/17/como-aplicar-las-ventajas-del-big-data-a-la-vida-real/
- https://books.google.es/books?id=apjBAgAAQBAJ&printsec=frontcover&dq=big+data+at+work&hl=en&sa=X&redir_esc=y#v=onepage&q=big%20data%20at%20work&f=false

FORMACIÓ

- http://www.bigdatabcn.com/programa-big-data-talent/formacio-en-big-data-coe/
- http://masters.obs-edu.com/masters-y-posgrados-en-direccion-general/master-en-data-management-e-innovacion-tecnologica/presentacion
- http://www.ciff.net/master-en-big-data-y-business-analytics.html?gclid=CjwKEAjw__atBRDfqe-9voylym8SJAAOBeQ_RPzJ8El1rV8jYOMUxcLnBDJsH5sKegP3vtoP9uKVwBoCfizw_wcB
- http://estudios.uoc.edu/es/masters-posgrados-especializaciones/especializacion/informatica-multimedia-telecomunicacion/big-data/presentacion?utm_medium=cpc&utm_source=google&utm_campaign=20151_pg_es_mktope_3wpoliedric_producte_notbranded&utm_content=area_immt_2052&utm_term=%2Bformacion%20%2Bbig%20%2Bdata
- http://fundacionbigdata.org/formacion/

- http://www.unir.net/ingenieria/master-visual-analytics-big-data/549200001429/
- https://www.google.es/search?q=where+to+learn+abut+big+data&oq=where+to+learn+abut+big+data&aqs=chrome..69i57.6989j0j7&sourceid=chrome&es_sm=93&ie=UTF-8#q=emc+big+data+master
- http://bigdatauniversity.com/
- https://josejuliolopezsantos.wordpress.com/2014/06/14/los-recursos-mas-destacados-para-la-formacion-en-big-data-en-espana-y-online/
- http://estudios.unir.net/programa/master-online-big-data/539000017206/?utm_source=google&utm_medium=bus&utm_content=texto&utm_campaign=googleunireu_estextoig_vabd_bus&gclid=CjwKEAjw__atBRDfqe-9voylym8SJAAOBeQ_rLaKLyMbUS5Pae12ZKL-2hmLWdE--m4w1RCjiExjDhoC88_w_wcB
- http://kschool.com/?utm_campaign=KS&utm_source=Ideas-project-manager&utm_medium=post
- https://www.u-tad.com/estudios/experto-en-big-data/
- http://www.mbitschool.com/
- http://www.barcelonaschoolofmanagement.upf.edu/master-of-science-in-management-specialization-in-business-analytics
- http://www.uc3m.es/ss/Satellite/Postgrado/es/Detalle/Estudio_C/1371210340413/1371211096495/Master_Universitario_en_Metodos_Analiticos_para_Datos_Masivos:_Big_Data
- http://citius.usc.es/masterbigdata/
- https://www.city.ac.uk/courses/postgraduate/data-science-msc
- http://www.qmul.ac.uk/postgraduate/coursefinder/courses/121386.html
- http://www.brunel.ac.uk/courses/postgraduate/data-science-and-analytics-msc
- http://data-informed.com/bigdata_university_map/
- http://data-informed.com/saint-peters-u-unveils-plans-data-science-masters-program/
- http://data-informed.com/marketing-analytics-focus-of-new-program-at-iit-stuart-school/
- http://www.usfca.edu/artsci/msan/

DEFINICIÓ

- https://www.marketingtechblog.com/benefits-of-big-data/
- http://onerp.es/beneficios-big-data-empresa/
- http://www-05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf

- https://es.wikipedia.org/wiki/Big_data

SEGMENTACIÓ

- http://iveybusinessjournal.com/publication/why-big-data-is-the-new-competitive-advantage/
- http://www.cio.com/article/2392067/data-management/big-data-analytics-a-big-benefit-for-marketing-departments.html
- http://www.lantares.com/blog/big-data-ventajas-de-la-revolucion-de-los-datos-masivos
- http://www.siliconweek.es/workspace/las-cinco-ventajas-competitivas-que-aporta-el-big-data-49286
- http://onerp.es/beneficios-big-data-empresa/
- http://www.ice.udl.es/udv/demoassig/recursos/dima/fitxer/unitat_3.pdf
- https://www.google.es/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0CDIQFjACahUKEwjp1YiOra3HAhWFORoKHSz3AW0&url=http%3A%2F%2Fwww.xtec.cat%2F~tsoler24%2Fm5%2Fm5exprojecte%2Fic%2FApunts%2FSegmentacio%2520de%2520mercat.doc&ei=UmDQVenkG4XzaKzuh-gG&usg=AFQjCNFj-zJlY_FXR1rK7Fuagfj0Aytmzw&sig2=vA6eHYhdEO3Q53cxD58gfQ&bvm=bv.99804247,d.d2s
- https://ca.wikipedia.org/wiki/Segmentaci%C3%B3_de_mercat
- http://emprenedoria.barcelonactiva.cat/emprenedoria/cat/edit.do?id=672716&id_activitat_mestre=672716&codiIdioma=1
- http://retos-directivos.eae.es/variables-y-beneficios-de-la-segmentacion-de-mercado/
- https://marketingpertu.wordpress.com/2013/11/21/quins-criteris-podem-utilitzar-per-segmentar/

DESAVANTATGES

- http://www.informationbuilders.com/blog/rado-kotorov/11432
- http://spotfire.tibco.com/blog/?p=10941
- http://www.information-management.com/gallery/6-problems-big-data-will-make-worse-10023544-1.html
- http://bigdata-madesimple.com/5-advantages-and-disadvantages-of-cloud-storage/
- https://www.aclu.org/blog/eight-problems-big-data
- http://siliconangle.com/blog/2014/01/03/3-big-data-problems-facing-businesses-in-2014/
- https://datafloq.com/read/the-power-of-real-time-big-data/225

ALTRES

- http://www.ericrettberg.com/datacultj1/tag/big-data/
- http://www.slideshare.net/joyanes/ioi-bi-tema6-bigdata

PART D'ANÀLISI

ANÀLISI

- http://searchdatacenter.techtarget.com/es/definicion/Analisis-de-big-data
- https://www.accenture.com/es-es/insight-building-foundation-big-data.aspx
- http://spain.emc.com/collateral/emc-perspective/h8668-ep-cloud-big-data-analytics.pdf
- http://w27.bcn.cat/porta22/images/es/Barcelona_Treball_Capsula_BigData_nov2012_es_tcm24-33137.pd
- http://www.territoriocreativo.es/etc/2015/05/datos-analizados-datos-aprovechados-big-data-innoturismo.html
- http://www-03.ibm.com/systems/infrastructure/us/en/big-data-analytics/index.html

TÈCNIQUES

- https://es.wikipedia.org/wiki/Big_data#An.C3.A1lisis_de_datos
- http://rocreguant.com/tecnicas-de-analisis-de-big-data/306/
- https://josejuliolopezsantos.wordpress.com/2014/07/11/las-principales-tecnicas-big-data-y-sus-aplicaciones/
- http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html
- http://www.sas.com/en_us/insights/articles/big-data/real-world-big-data-tips.html
- http://bigdata-madesimple.com/26-popular-techniques-for-analysing-big-data/
- http://documents.software.dell.com/statistics/textbook/data-mining-techniques
- file:///C:/Users/Mariona/Desktop/Music/MGI_big_data_full_report.pdf
- file:///C:/Users/Mariona/Desktop/Music/p44%20(1).pdf
- https://es.wikipedia.org/wiki/Venta_cruzada
- http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

EINES

- http://searchdatacenter.techtarget.com/es/definicion/Analisis-de-big-data
- http://www.tableau.com/es-es/solutions/big-data-analysis
- http://www.informationweek.com/big-data/big-data-analytics/16-top-big-data-analytics-platforms/d/d-id/1113609

- http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-time-for-new-tools/a/d-id/1318106
- http://searchbusinessanalytics.techtarget.com/essentialguide/Guide-to-big-data-analytics-tools-trends-and-best-practices
- http://www.datamation.com/data-center/50-top-open-source-tools-for-big-data-1.html
- http://www.computerworld.com/article/2690856/big-data/8-big-trends-in-big-data-analytics.html
- http://www.ibm.com/developerworks/ssa/local/im/utilizando-jaql-para-analizar-big-data/
- http://www-03.ibm.com/software/products/es/category/bigdata
- http://spanish.peopledaily.com.cn/n/2015/0821/c92121-8939601.html
- http://www8.hp.com/es/es/software-solutions/big-data-platform-haven/
- http://spain.emc.com/collateral/emc-perspective/h8668-ep-cloud-big-data-analytics.pdf
- http://fundacionbigdata.org/glosario-big-data/
- http://www.actian.com/products/analytics-platform/
- http://www.pentaho.com/product/big-data-analytics
- http://www.ibm.com/software/products/en/category/bigdata

HADOOP

- https://es.wikipedia.org/wiki/Big_data
- https://es.wikipedia.org/wiki/Hadoop
- http://www.lynda.com/Hadoop-tutorials/Hadoop-Fundamentals/191942-2.html
- http://momentotic.com/2013/05/16/que-es-hadoop/
- http://www.ticout.com/blog/2013/04/02/introduccion-a-hadoop-y-su-ecosistema/
- http://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/328879/C%C3%B3mo-se-relacionan-Big-Data-y-Hadoop
- https://www.youtube.com/watch?v=vuaxrIUae60
- http://www.nubelo.com/blog/que-son-los-frameworks/
- http://jordisan.net/blog/2006/que-es-un-framework/
- https://es.wikipedia.org/wiki/Framework
- http://clusterfoodmasi.es/cluster/que-son-los-clusters/
- http://www.chw.net/2004/01/clusters-que-son-y-para-que-sirven/
- http://searchbusinessanalytics.techtarget.com/definition/Hadoop-cluster
- https://es.wikipedia.org/wiki/Nodo_(inform%C3%A1tica)

- http://www.monografias.com/trabajos24/libro-sistemas-operativos/libro-sistemas-operativos.shtml
- https://es.wikipedia.org/wiki/Replicaci%C3%B3n_(inform%C3%A1tica)
- https://ca.wikipedia.org/wiki/Node_(inform%C3%A0tica)
- https://ca.wikipedia.org/wiki/Rack
- http://momentotic.com/2013/05/16/que-es-hadoop/
- http://www.monografias.com/trabajos24/libro-sistemas-operativos/libro-sistemas-operativos.shtml
- https://es.wikipedia.org/wiki/Replicaci%C3%B3n_(inform%C3%A1tica)

HDFS

- http://www-01.ibm.com/software/data/infosphere/hadoop/hdfs/
- http://www.aosabook.org/en/hdfs.html
- http://hortonworks.com/hadoop/hdfs/
- http://zoo.cs.yale.edu/classes/cs422/2014fa/readings/papers/shvachko10hdfs.pdf
- http://searchbusinessanalytics.techtarget.com/definition/Hadoop-Distributed-File-System-HDFS
- http://www.webopedia.com/TERM/H/hadoop_distributed_file_system_hdfs.html
- http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/hdfs-and-mapreduce.html

MAPREDUCE

- https://en.wikipedia.org/wiki/Apache_Hadoop
- http://www.lynda.com/Hadoop-tutorials/Hadoop-Fundamentals/191942-2.html
- https://en.wikipedia.org/wiki/MapReduce
- http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
- http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf
- http://searchcloudcomputing.techtarget.com/definition/MapReduce
- http://www.webopedia.com/TERM/H/hadoop_mapreduce.html
- http://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/
- http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html#Overview

HBASE

- http://www.lynda.com/Hadoop-tutorials/Hadoop-Fundamentals/191942-2.html
- http://www-01.ibm.com/software/data/infosphere/hadoop/hbase/
- https://en.wikipedia.org/wiki/Apache_HBase
- https://unpocodejava.wordpress.com/2012/08/28/un-poco-de-hbase/
- http://www.tutorialspoint.com/es/hbase/hbase_overview.htm
- http://hbase.apache.org/0.94/book/mapreduce.example.html
- http://www.pragsis.com/sites/default/files/documentation/comprenderhbaseybigtable.pdf
- https://sites.google.com/site/estrategiasdepersistencia/material-teorico/bigtable/bigtable---prctica-con-hbase
- http://www.tutorialspoint.com/es/hbase/hbase_create_data.htm

EXCEL

- http://www.excel-easy.com/data-analysis.html
- http://www.excel-easy.com/data-analysis/analysis-toolpak.html
- https://www.youtube.com/watch?v=DrTTNwo-bjc
- https://www.youtube.com/watch?v=YoDJa8ZSHik
- https://www.youtube.com/watch?v=5MFjwM6K5Sg
- https://technet.microsoft.com/en-us/magazine/ff969363.aspx
- http://people.umass.edu/evagold/excel.html
- https://www.edx.org/course/excel-data-analysis-visualization-microsoft-dat206x
- https://www.youtube.com/watch?v=EbVazVpQMnc

LLIBRES

Viktor Mayer-Schönberger, Kenneth Cukier. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think.*

PART PRÀCTICA

- https://en.wikipedia.org/wiki/Collaborative_filtering
- https://upload.wikimedia.org/wikipedia/commons/5/52/Collaborative_filtering.gif
- http://www.analyticsvidhya.com/blog/2015/08/beginners-guide-learn-content-based-recommender-systems/

LLIBRES

Toby Segaran. (2007) . *Programming Collective Intelligence.*

Gràcies per la vostra lectura,
pel vostre interès i pel vostre temps.

Mariona Gabarró Rovira
Desembre de 2015